Natural Language Processing for Precision Medicine

Hoifung Poon, Chris Quirk, Kristina Toutanova, Scott Wen-tau Yih

First Half

Precision medicine

Annotation bottleneck

Extract complex structured information

Beyond sentence boundary

Second Half

Reasoning

Applications to precision medicine

Resources

Open problems

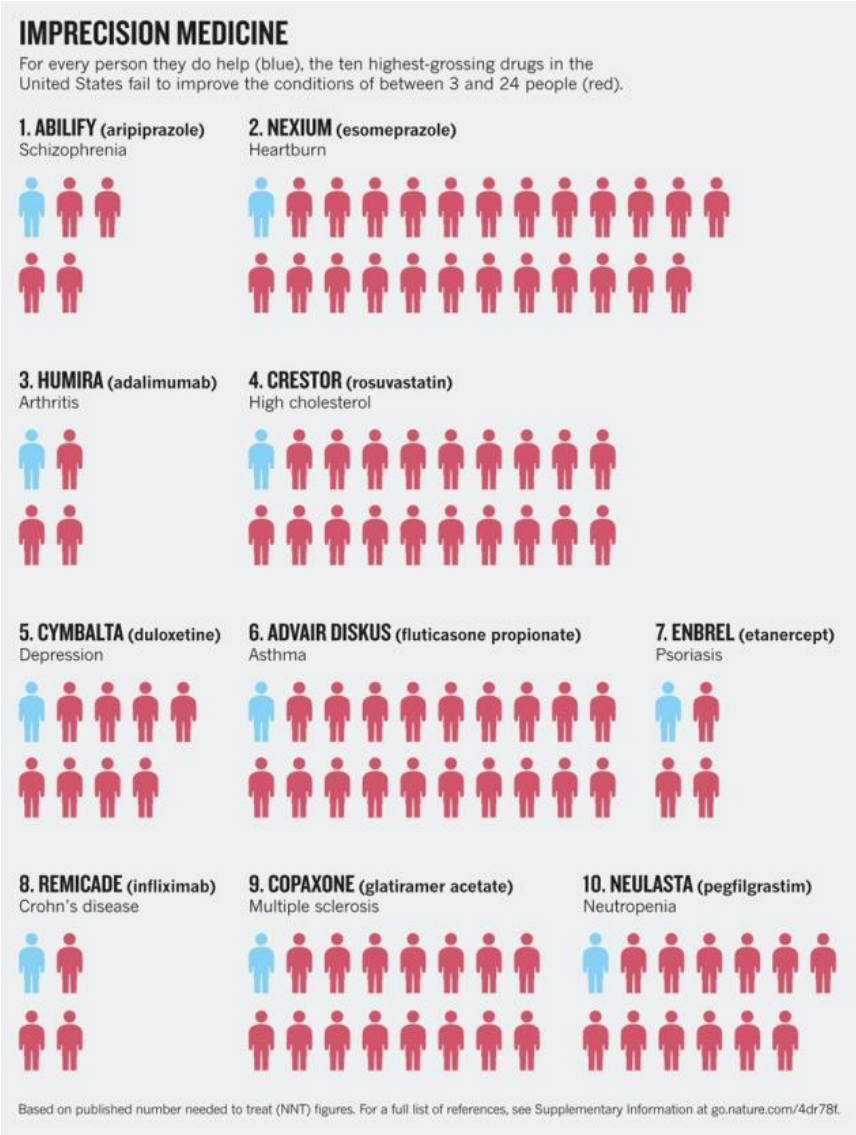
Part 1: Precision Medicine

What is precision medicine

Why it’s an exciting time to have impact

How can NLP help

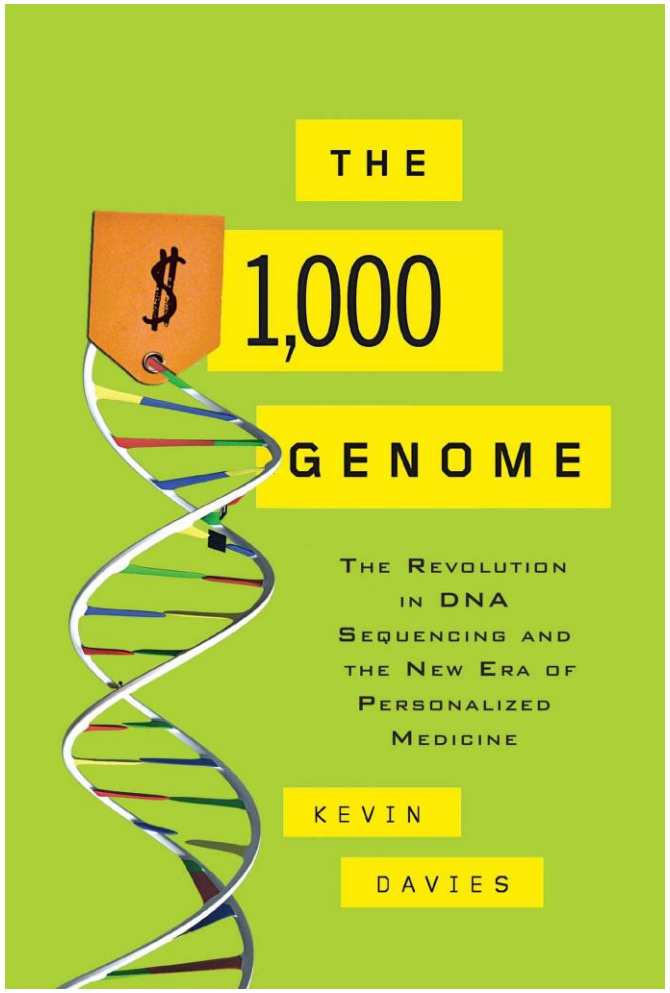
Medicine Today Is Imprecise



Top 20 drugs
80% non-responders

Wasted
1/3 health spending
$1 Trillion / year

Disruption: Big Data

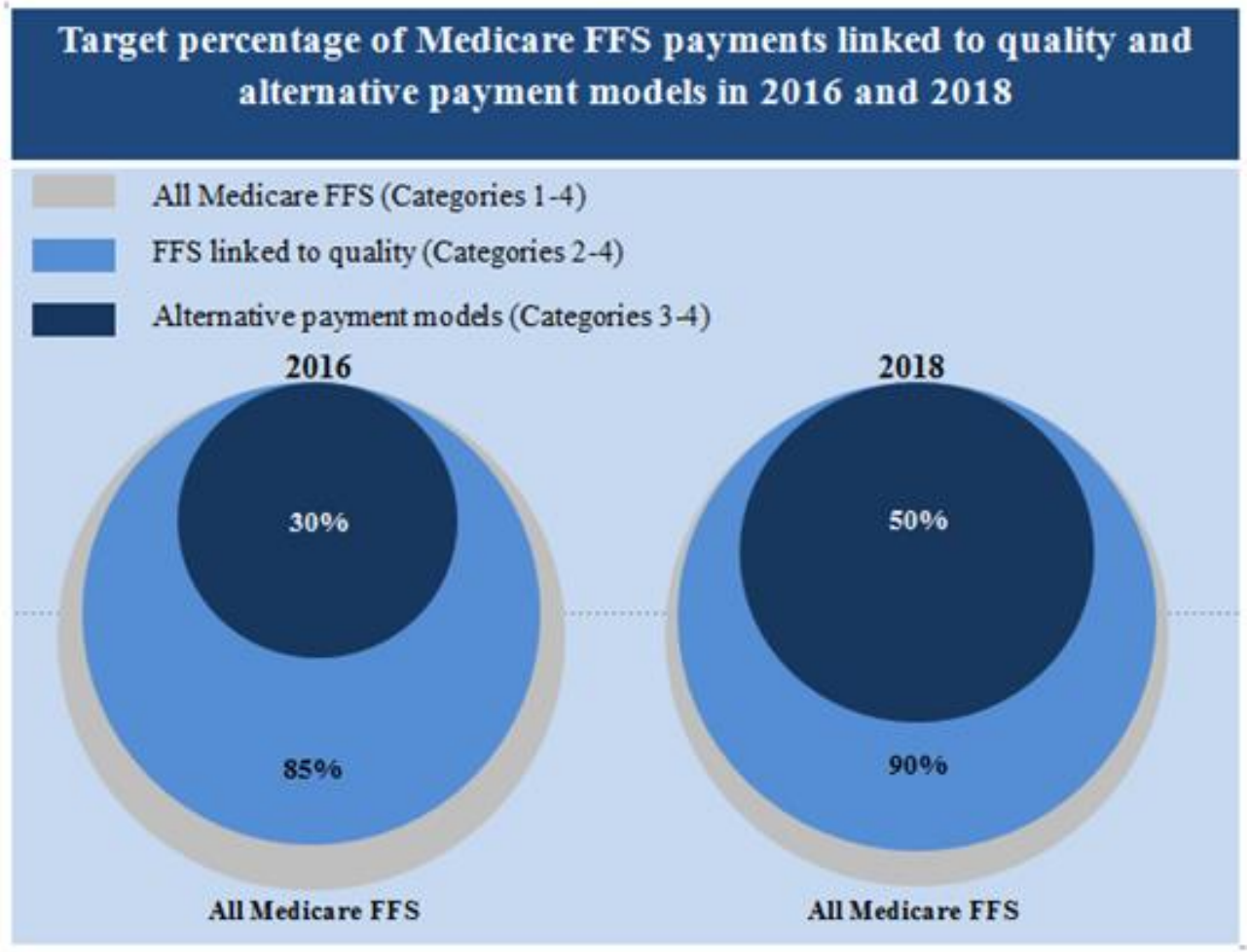


Accenture study: 93% of US doctors using EMRs

May 14, 2013 | IHQRE informatics, IHQRE Journal Club | EHR, EMR, Meaningful Use

2009 – 2013: 40% → 93%

Disruption: Pay-for-Performance

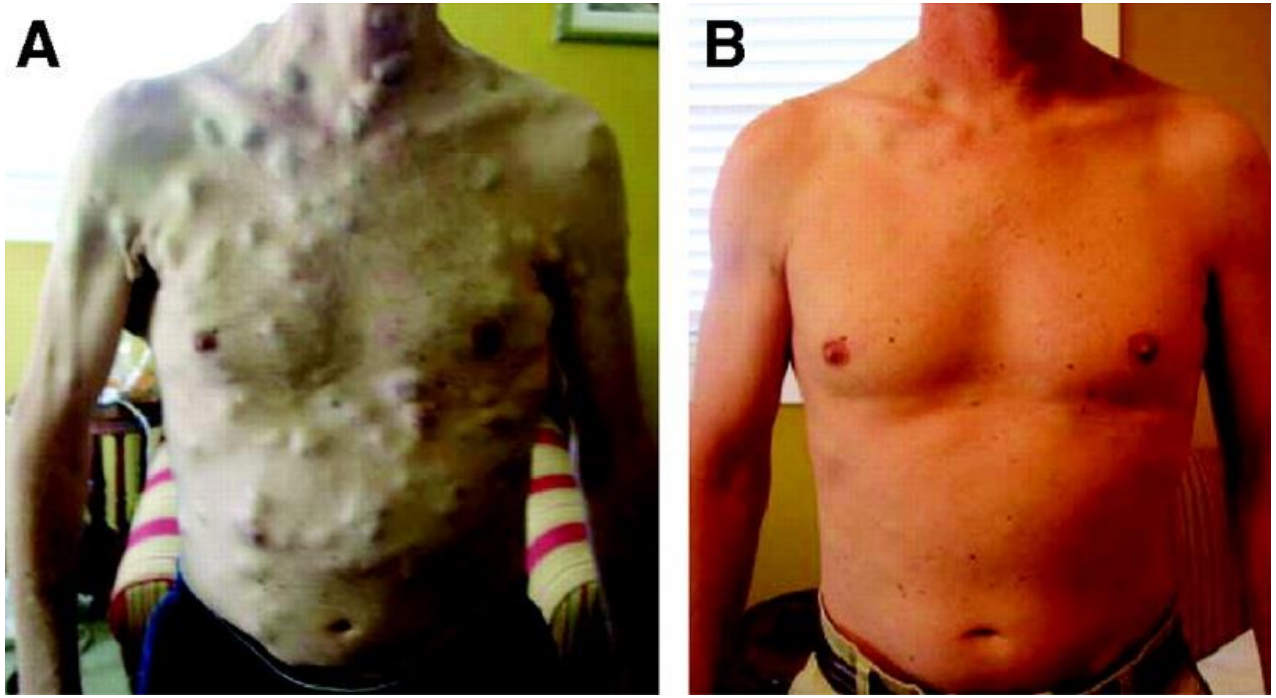


Aetna®

BlueCross®

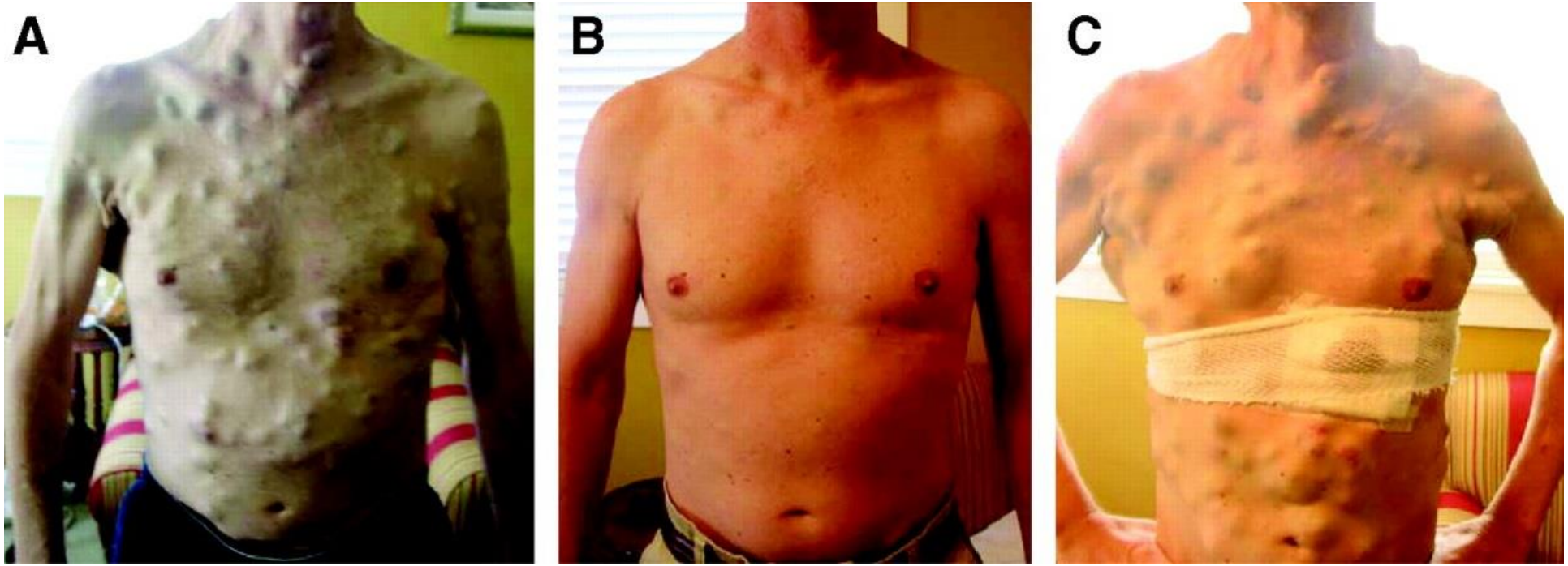
Goal: 75% by 2020

Vemurafenib on BRAF-V600 Melanoma



Before Treatment **15 Weeks**

Vemurafenib on BRAF-V600 Melanoma



Before Treatment **15 Weeks** **23 Weeks**

Why Curing Cancer Is Hard?

Cancer stems from normal biology

Cancer is not a single disease

Cancer naturally resists treatment

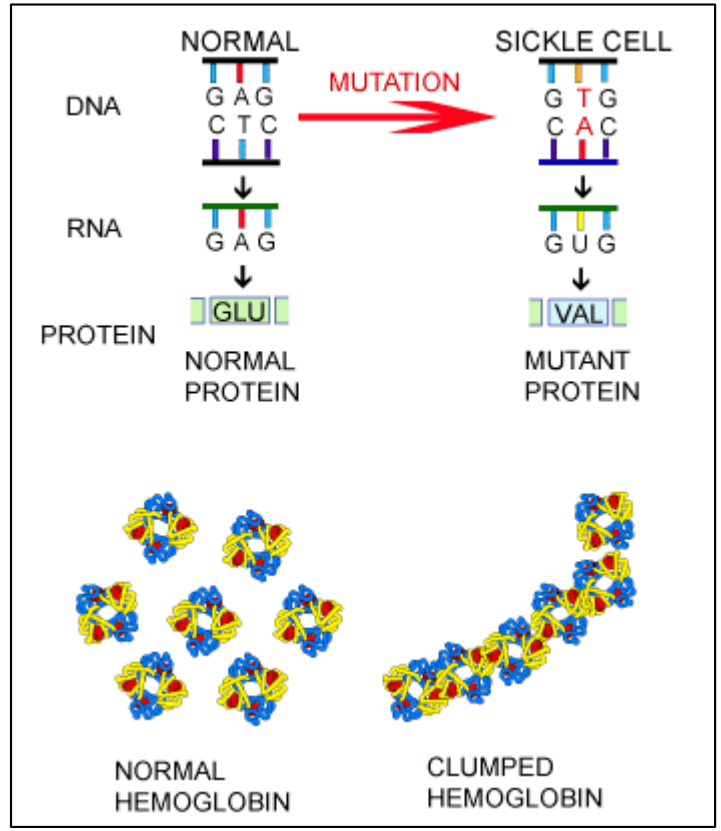
Cancer Stems from Normal Biology

Cancer is caused by genetic mutations

Cells divide billions of times everyday

Each division generates a few mutations

Inevitable: Enough of right mutations



Understanding Evolution
your one-stop source for information on evolution

Cancer Is “Thousands of Diseases”

Traditionally classified by originating organ

“Similar” tumors might have few common mutations

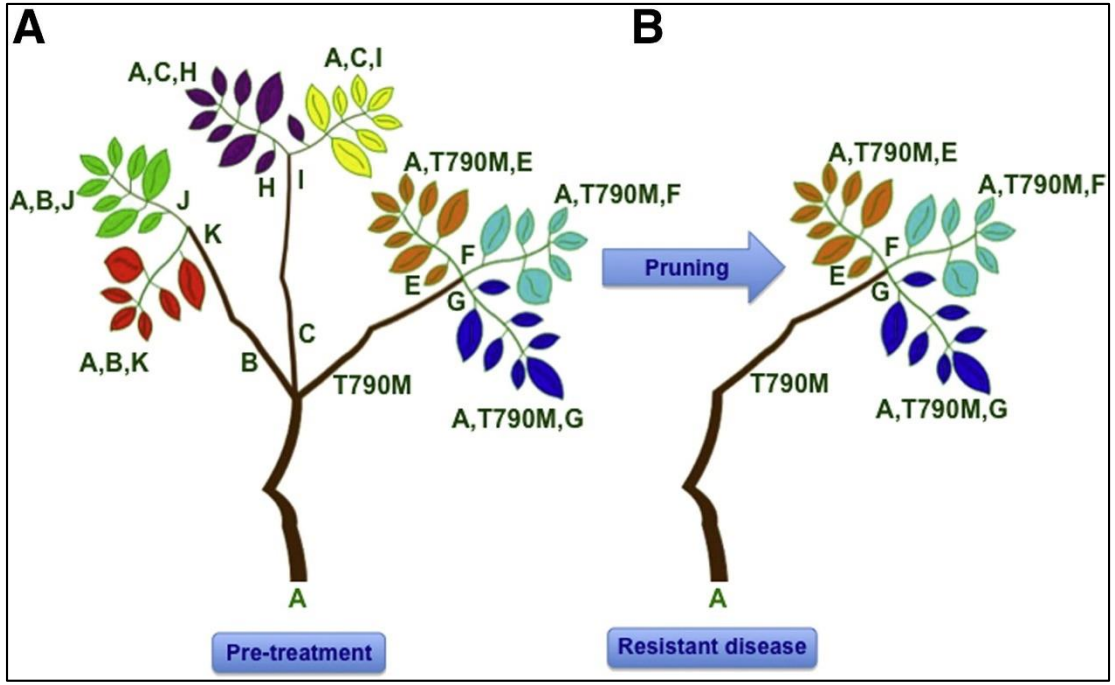
“20-80 rule”: Treatments often fail for most patients

Cancer Has Evolution on Its Side

Over a billion cells upon detection

Many “clones” w/ different characteristics

Killing primary clone liberates resistant subclones



Adapting Clinical Paradigms to the Challenges of Cancer Clonal Evolution. Mrurgaesu et al., Am. J. Pathology 2013.

The New Hope

Think HIV

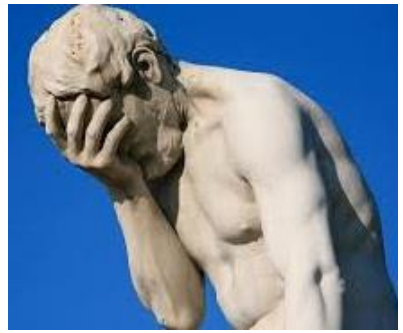
Example: Gleevec for CML

Cancer → Chronic disease

Why We Haven't Solved Precision Medicine?

… ATTCGG**A**TATTTAAG**G**C …
… ATTCGGGTATTTAAGCC …
… ATTCGG**A**TATTTAAG**G**C …
… ATTCGGGTATTTAAGCC …
… ATTCGG**A**TATTTAAG**G**C …
… ATTCGGGTATTTAAGCC …

High-Throughput Data



Discovery

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

AI is the key to overcome these bottlenecks

Molecular Tumor Board

www.ucsf.edu/news/2014/11/120451/bridging-gap-precision-medicine

Key Scenario: Molecular Tumor Board

Problem: Hard to scale

U.S. 2016: 1.7 million new cases, 600K deaths

902 cancer hospitals

Memorial Sloan Kettering

- Sequence: Tens of thousands
- Board can review: A few hundred

Wanted: Decision support for precision medicine

First-Generation Molecular Tumor Board

Knowledge bottleneck

E.g., given a tumor sequence, determine:

- What genes and mutations are important
- What drugs might be applicable

Can do manually but hard to scale

Next-Generation Molecular Tumor Board

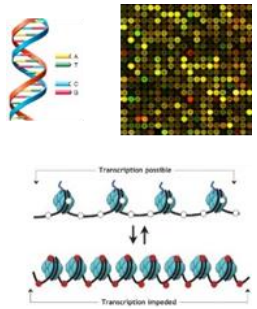
Reasoning bottleneck

E.g., personalize drug combinations

Can’t do manually, ever

How Can We Help?

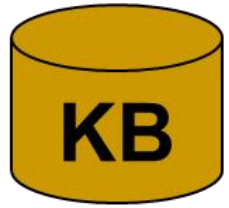
Big Medical Data

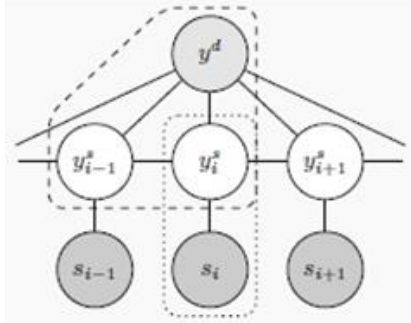


Decision Support

Precision Medicine

Machine Reading





Predictive Analytics

Example: Tumor Board KB Curation

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

Gefitinib can treat tumors w. EGFR-L858E mutation

OncoKB Team

OncoKB is developed and maintained by the Knowledge Systems group in the Marie Josée and Henry R. Kravis Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center.

Design & Development

Debyani Chakravarty, PhD
Jianjiong Gao, PhD
Sarah Phillips, PhD
Hongxin Zhang, MSc
Ritika Kundra, MSc
Jiaojiao Wang, MSc
Ederlinda Paraiso, MPA
Julia Rudolph, MPA
David Solit, MD
Paul Sabbatini, MD
Nikolaus Schultz, PhD

Clinical Genomics Annotation Committee

Shrujal Baxi, MD, MPH
Margaret Callahan, MD, PhD
Sarat Chandarlapaty, MD, PhD
Alexandra Charen-Snyder, MD
Ping Chi, MD, PhD
Daniel Danila, MD
Mrinal Gounder, MD
James Harding, MD
Matthew Hellman, MD
Alan Ho, MD, PhD
Gopa Iyer, MD
Yelena Janjigian, MD
Thomas Kaley, MD
Maeve Lowery, MD
Antonio Omuro, MD
Paul Paik, MD
Michael Postow, MD
Dana Rathkopf, MD
Alexander Shoushtari, MD
Neerav Shukla, MD
Tiffany Traina, MD
Martin Voss, MD
Rona Yaeger, MD

Core Curators

Moriah Nissan, PhD
Lindsay Saunders, PhD
Tara Soumerai, MD
Fiona Brown, PhD
Tripti Shrestha Bhattarai, PhD
Kinisha Gala, BSc
Aphrothiti Hanrahan, PhD
Anton Henssen, MD
Phillip Jonsson, PhD
Iñigo Landa-Lopez, PhD
Eneda Toska, PhD

Quest Diagnostics

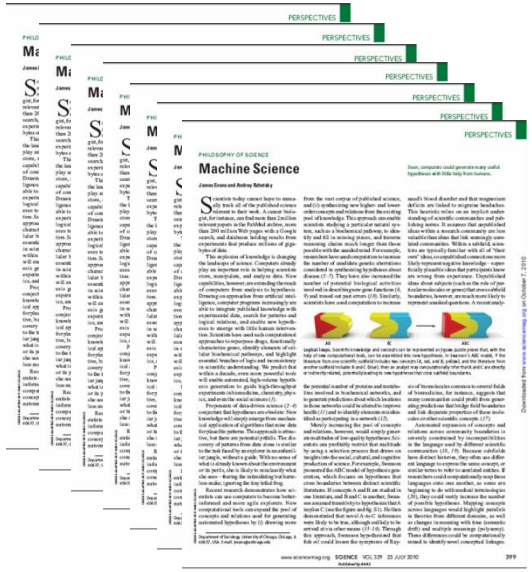
Feras M Abu Hantash, PhD
Andrew Grupe, PhD
Matthew Beer, BSc

PubMed

27 million abstracts

Two new abstracts every minute

Adds over one million every year



OncoKB Home About Team Levels of Evidence Actionable Genes News

Memorial Sloan Kettering Cancer Center

Can we help increase curation speed by 100X?

Design & Development

Debyani Chakravarty, PhD
Jianjiong Gao, PhD
Sarah Phillips, PhD
Hongxin Zhang, MSc
Ritika Kundra, MSc
Jiaojiao Wang, MSc
Ederlinda Paraiso, MPA
Julia Rudolph, MPA
David Solit, MD
Paul Sabbatini, MD
Nikolaus Schultz, PhD

Clinical Genomics Annotation Committee

Shrujal Baxi, MD, MPH
Margaret Callahan, MD, PhD
Sarat Chandarlapaty, MD, PhD
Alexandra Charen-Snyder, MD
Ping Chi, MD, PhD
Daniel Danila, MD
Mrinal Gounder, MD
James Harding, MD
Matthew Hellman, MD
Alan Ho, MD, PhD
Gopa Iyer, MD
Yelena Janjigian, MD
Thomas Kaley, MD
Maeve Lowery, MD
Antonio Omuro, MD
Paul Paik, MD
Michael Postow, MD
Dana Rathkopf, MD
Alexander Shoushtari, MD
Neerav Shukla, MD
Tiffany Traina, MD
Martin Voss, MD
Rona Yaeger, MD

Core Curators

Moriah Nissan, PhD
Lindsay Saunders, PhD
Tara Soumerai, MD
Fiona Brown, PhD
Tripti Shrestha Bhattarai, PhD
Kinisha Gala, BSc
Aphrothiti Hanrahan, PhD
Anton Henssen, MD
Phillip Jonsson, PhD
Iñigo Landa-Lopez, PhD
Eneda Toska, PhD

Quest Diagnostics

Feras M Abu Hantash, PhD
Andrew Grupe, PhD
Matthew Beer, BSc

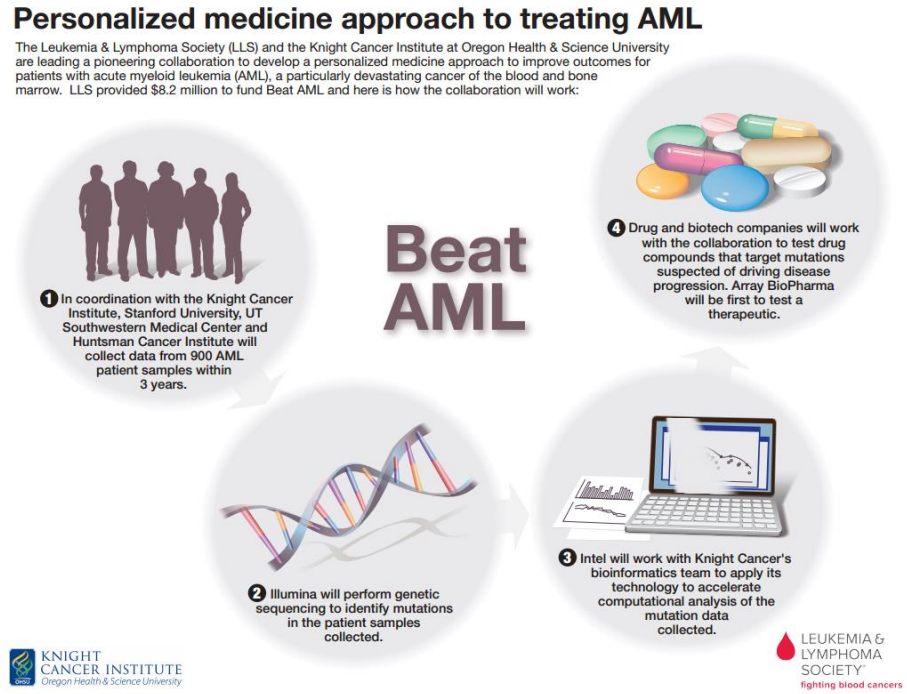
Example: Personalize Drug Combos

Targeted drugs: 149

Pairs: 11,026

Tested: 102 (in two years)

Unknown: 10,924



Can we find good combos in months, not centuries?

What Can We Achieve?

Cancer → Solved

Chronic diseases → Predict / prevent

Healthcare → Save trillions

NLP Challenges

Train machine reader w. little labeled data

Understand complex semantics

Reason beyond explicitly stated in text

Part 2: Annotation Bottleneck

Machine reading

Annotation bottleneck

Distant supervision

Grounded learning

Machine Reading

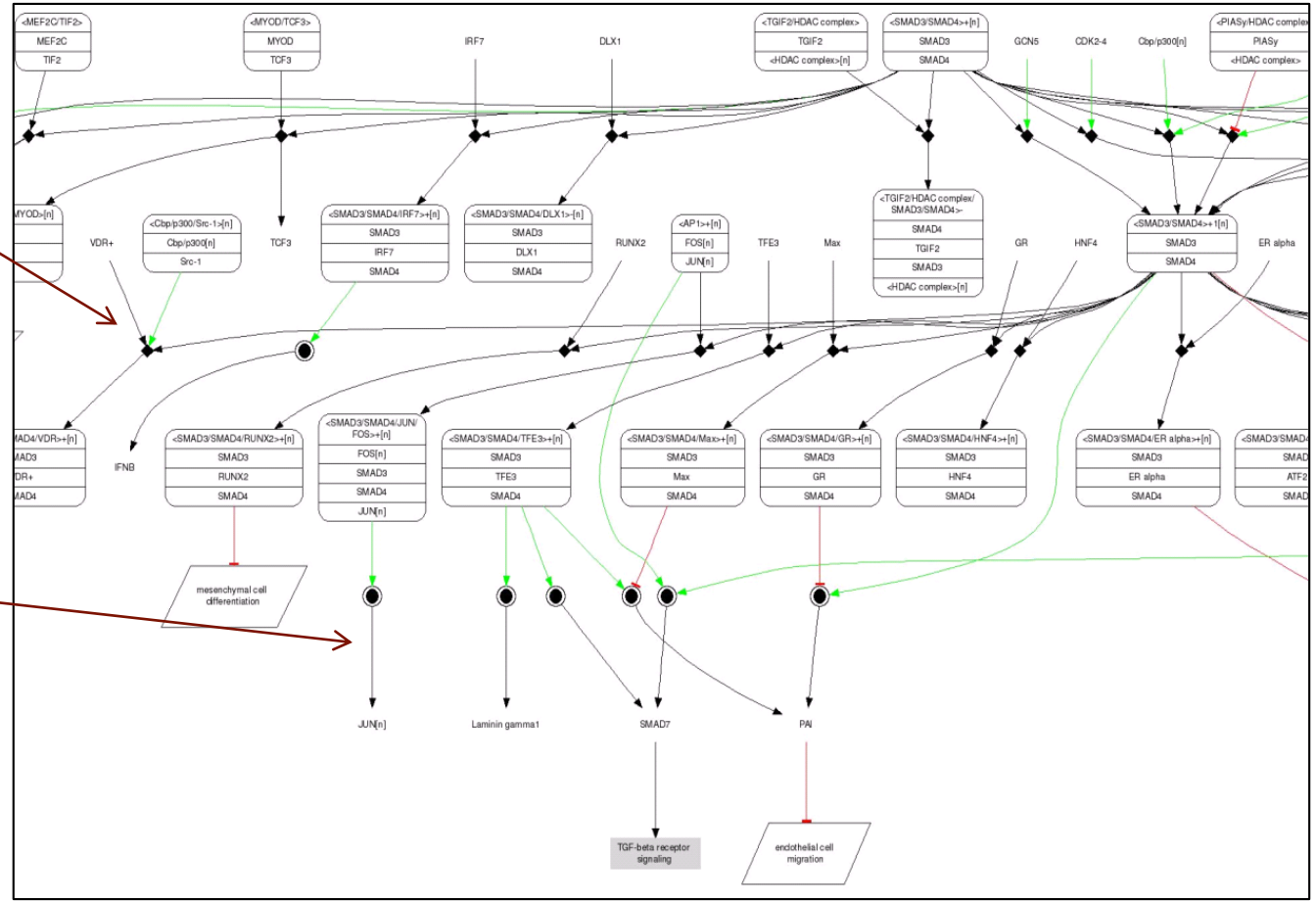
PMID: 123

…
VDR+ binds to SMAD3 to form
…

PMID: 456

…
JUN expression is induced by SMAD3/4
…

⋮



Knowledge Base

Complex Semantics

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

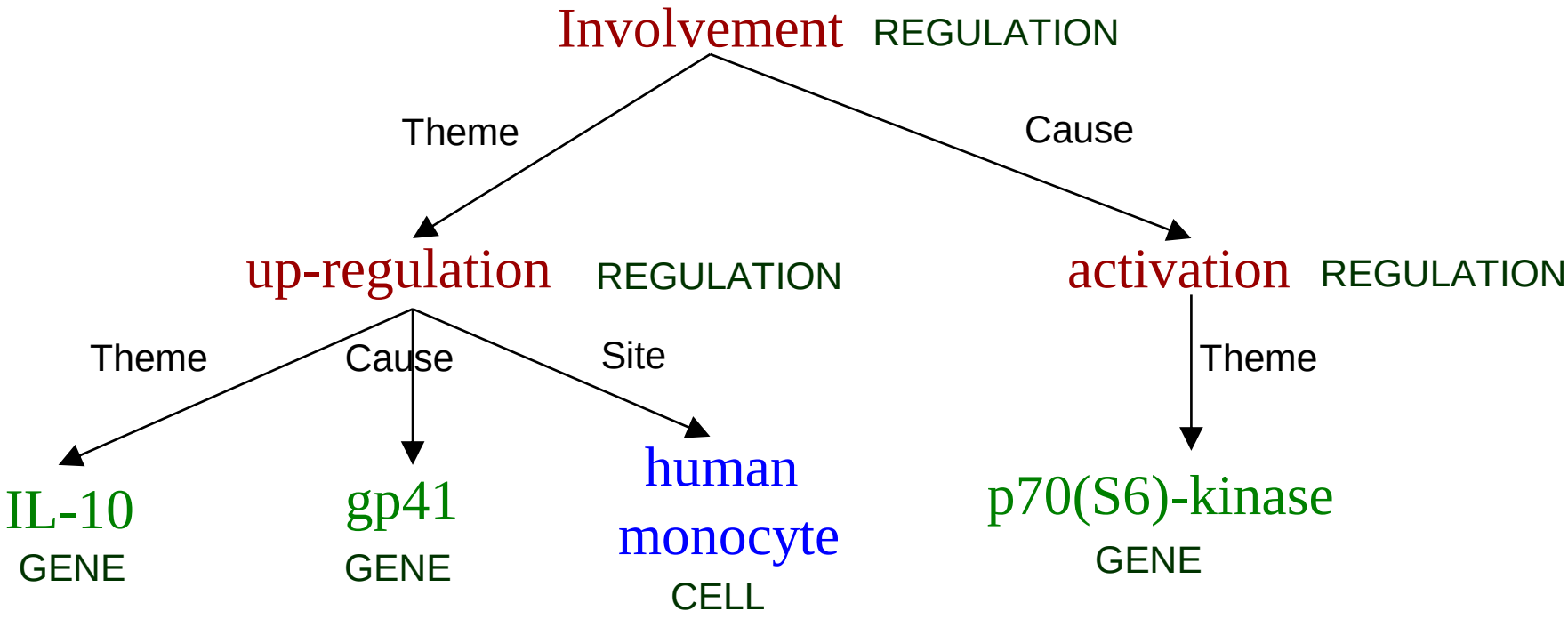
Complex Semantics

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

IL-10	gp41	human monocyte	p70(S6)-kinase
GENE	GENE	CELL	GENE

Complex Semantics

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



Long Tail of Variations

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

negative regulation

532 inhibited, 252 inhibition, 218 inhibit, 207 blocked, 175 inhibits, 157 decreased, 156 reduced, 112 suppressed, 108 decrease, 86 inhibitor, 81 Inhibition, 68 inhibitors, 67 abolished, 66 suppress, 65 block, 63 prevented, 48 suppression, 47 blocks, 44 inhibiting, 42 loss, 39 impaired, 38 reduction, 32 down-regulated, 29 abrogated, 27 prevents, 27 attenuated, 26 repression, 26 decreases, 26 down-regulation, 25 diminished, 25 downregulated, 25 suppresses, 22 interfere, 21 absence, 21 repress ……

Problem Formulation

Entity: Recognition, linking

Simple relation classification: binary, n-ary

Complex event extraction

Entity Recognition (a.k.a. Tagging)

BioCreative II

Task 1A: Gene Mention Tagging [2006-04-01]

Gene Mention Tagging task is concerned with the named entity extraction of gene and gene product mentions in text.

Premise

Systems will be required to return the start and end indices corresponding to all the genes and gene products mentioned in a given MEDLINE sentence. This named entity task is a crucial first step for information extraction of relationships between genes and gene products.

System Input

The input file will consist of ascii sentences, one per line. Each sentence will be preceded on the same line by a sentence identifier.

System Output

Each system must output an ascii list of reported gene name mentions, one per line, and formatted as:

```
sentence-identifier-1|start-offset-1 end-offset-1|optional text...
sentence-identifier-1|start-offset-2 end-offset-2|optional text...
sentence-identifier-1|start-offset-3 end-offset-3|optional text...
sentence-identifier-2|start-offset-1 end-offset-1|optional text...
sentence-identifier-3|start-offset-1 end-offset-1|optional text...
```

Entity Recognition (a.k.a. Tagging)

BioCreative II

Task 1A: Gene Mention Tagging [2006-04-01]

Gene Mention Tagging task is concerned with the named entity extraction of gene and gene product mentions in text.

Premise

Systems will be required to retur
MEDLINE sentence. This named
products.

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

System Input

The input file will consist of ascii sentences, one per line. Each sentence will be preceded on the same line by a sentence identifier.

System Output

Each system must output an ascii list of reported gene name mentions, one per line, and formatted as:

```
sentence-identifier-1|start-offset-1 end-offset-1|optional text...
sentence-identifier-1|start-offset-2 end-offset-2|optional text...
sentence-identifier-1|start-offset-3 end-offset-3|optional text...
sentence-identifier-2|start-offset-1 end-offset-1|optional text...
sentence-identifier-3|start-offset-1 end-offset-1|optional text...
```

Entity Recognition (a.k.a. Tagging)

Introduction to the Bio-Entity Recognition Task at JNLPBA

Jin-Dong KIM, Tomoko OHTA, Yoshimasa TSURUOKA, Yuka TATEISI
CREST, Japan Science and Technology Agency, and
Department of Computer Science, University of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

Nigel COLLIER
National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan†

Protein, DNA, RNA, cell line, cell type

Abstract

We describe here the JNLPBA shared task of bio-entity recognition using an extended version of the GENIA version 3 named entity corpus of MEDLINE abstracts. We provide background information on the task and present a general discussion of the approaches taken by participating systems.

1 Introduction

Bio-entity recognition aims to identify and clas-

We have shown that **<cons sem="G#protein">***interleukin-1***</cons>** *(*<cons sem="G#protein">*IL-1*</cons>*) and* <cons sem="G#protein">*IL-2*</cons> *control* <cons sem="G#DNA">*IL-2 receptor alpha (IL-2R alpha) gene*</cons> *transcription in* <cons sem="G#cell_line">*CD4-CD8-murine T lymphocyte precursors*</cons>.

Figure 1: Example MEDLINE sentence marked up in XML for molecular biology named-entities.

Entity Recognition (a.k.a. Tagging)

Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets

Burr Settles
Department of Computer Sciences
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison
Madison, WI, USA
bsettles@cs.wisc.edu

Entity Recognition (a.k.a. Tagging)

Even biologists hard to determine

Rich ontologies available

- HUGO: Human genes
- MeSH: Diseases, drugs, …
- dbSNP: point mutations

Lessons learned

What we need is entity linking (a.k.a. normalization)

Entity Linking (a.k.a. Normalization)

In eubacteria and eukaryotic organelles the product of this gene, peptide deformylase (PDF), removes the formyl group from the initiating methionine of nascent peptides. …… The discovery that a natural inhibitor of PDF, actinonin, acts as an antimicrobial agent in some bacteria has spurred intensive research into the design of bacterial-specific PDF inhibitors. …… In humans, PDF function may therefore be restricted to rapidly growing cells.

GeneCards®
HUMAN GENE DATABASE

Aliases for PDF Gene

- Peptide Deformylase (Mitochondrial) [2 3 5]
- Polypeptide Deformylase [4]
- EC 3.5.1.88 [4]
- PDF1A [4]

PDF Gene (Protein Coding) ★

Peptide Deformylase (Mitochondrial)

GCID: GC16M069328 ?
GIFtS: 44 ?

Relation: Classification

The p56Lck inhibitor Dasatinib was shown to enhance apoptosis induction by dexamethasone in otherwise GC-resistant CLL cells.

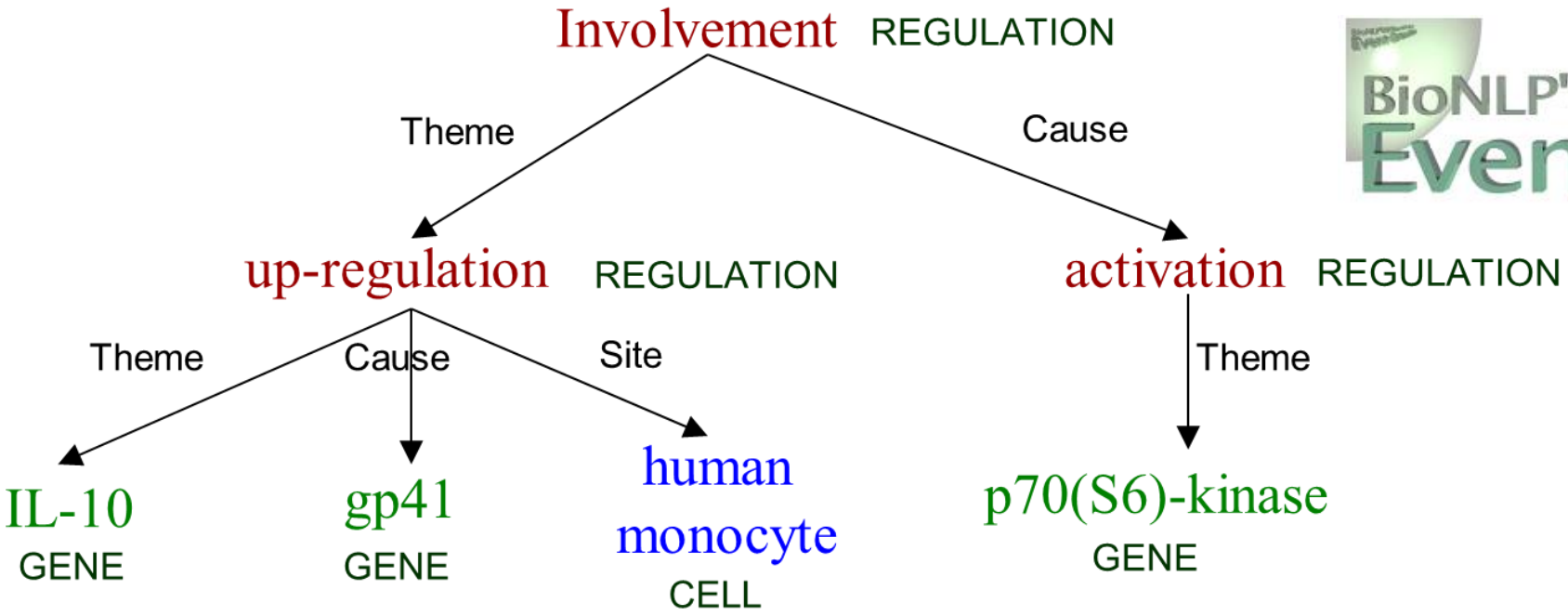
This finding concurs with the observation by Sade showing that Notch-mediated resistance of a mouse lymphoma cell line could be overcome by inhibiting p56Lck.

Dasatinib could be used to treat Notch-mutated tumors.

TREAT(Dasatinib, Notch)

Relation: Complex Event Extraction

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



BioNLP'09 Shared Task on Event Extraction

Machine Reading

Prior work

- Focused on Newswire / Web
- Popular entities and facts
- Redundancy → Simple methods often suffice

High-value verticals

- Healthcare, finance, law, etc.
- Little redundancy: Rare entities and facts
- Novel challenges require sophisticated NLP

Annotation Bottleneck

Hire experts to label examples: Scalable?

Crowdsource: “Are these English?”

Learning with Indirect Supervision

Unsupervised learning

Statistical relational learning

Distant supervision

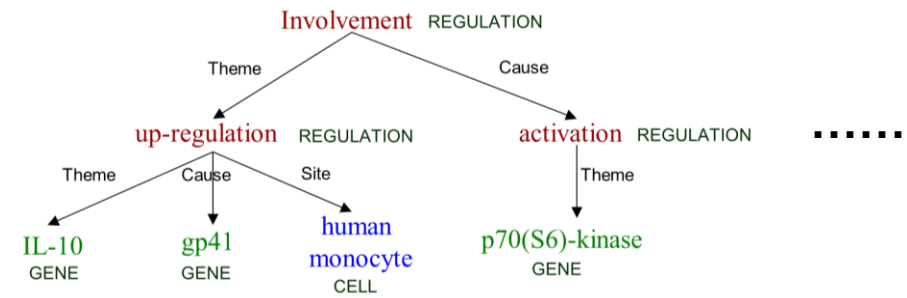
Incidental learning

Situated learning

Grounded language learning

Grounded Learning

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…



Grounding Takes Many Forms

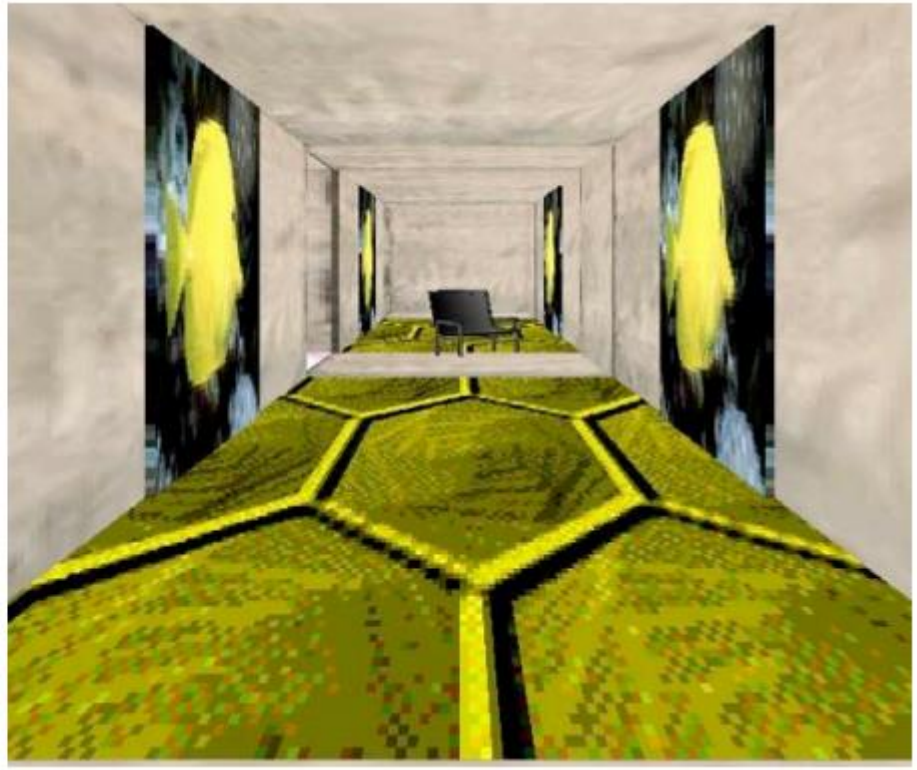
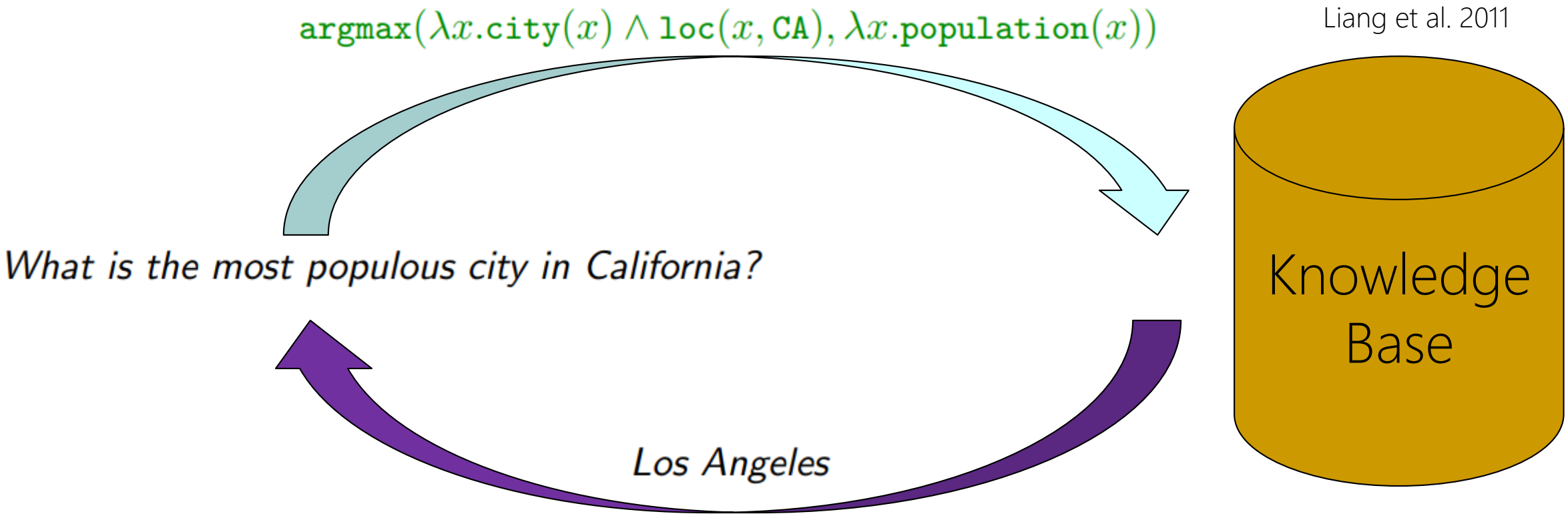


Image from
Artzi & Zettlemoyer 2013

[MacMahon et al. 2006; Chen & Mooney 2011; Artzi & Zettlemoyer 2013; ……]

Grounding Takes Many Forms

Example from
Liang et al. 2011



[Clark et al. 2010; Liang et al. 2011; .....]

Free Lunch: Existing KB

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

Free Lunch: Existing KB

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

Free Lunch: Existing KB

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Free Lunch: Existing KB

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

BCL2 transcription is suppressed by P53 expression.

The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …

……

Free Lunch: Existing KB

NCI Pathway KB

Regulation	Theme	Cause
Positive	A2M	FOXO1
Positive	ABCB1	TP53
Negative	BCL2	TP53
…	…	…

TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Distant Supervision

Distant Supervision

[Craven & Kumlien 1999, Mintz et al. 2009]

Use KB to annotate examples in unlabeled text

Binary relation classification

Assume entity linking is done

Recipe

Identify co-occurring entity pairs in text

Construct training data

- Positive: Pairs w/ known relation in KB
- Negative: Randomly sampled

Train your favorite classifier

Evaluation

Sample precision

Absolute recall

Examples in Newswire/Web

WordNet hypernym [Snow et al 2005]

Wikipedia infobox [Fei & Weld 2007]

Freebase [Mintz 2009]

Examples in Biomedicine

Protein localization [Craven & Kumlien 1999]

Genetic pathway [Poon et al. 2015, Mallory et al 2016]

Drug adverse effect [Bing et al. 2015]

MicroRNA-gene interaction [Lamurias et al. 2017]

The Literome Project

Welcome 24.16.142.216
change to user id

Microsoft Research

Search for directed genic interactions: tp53 bcl2 Search

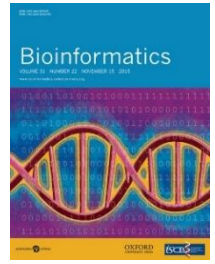
Direct Search
- BCL2 → TP53 (18)
- BCL2 ⊣ TP53 (15)
- BCL2 → TP53 (5)
- TP53 → BCL2 (25)
- TP53 ⊣ BCL2 (13)
- TP53 → BCL2 (10)

Possible intermediates for BCL2 → TP53
- BCL2 → AGTR1 → TP53
- BCL2 → AKT1 → TP53
- BCL2 → ANGPT2 → TP53
- BCL2 → ANXA1 → TP53
- BCL2 → ANXA6 → TP53
- BCL2 → APAF1 → TP53
- BCL2 → ATG5 → TP53
- BCL2 → ATM → TP53
- BCL2 → ATRAID → TP53
- BCL2 → BAX → TP53
- BCL2 → BCL10 → TP53
- BCL2 → BCR → TP53
- BCL2 → BECN1 → TP53
- BCL2 → BNIP3 → TP53
- BCL2 → BRCA1 → TP53

BCL2 ⊣ TP53 (1 - 15 of 15)

PMID: 10037739 Inhibition of p53 transcriptional activity by Bcl-2 requires its membrane-anchoring domain.	... protein Bcl-2 potently inhibits p53 ... various p53-responsive promoters ... (details)
PMID: 10866313 Mitochondrial amplification of death signals determines thymidine kinase/ganciclovir-triggered activation of apoptosis.	... since Bcl-2 overexpression ... strongly reduced TK/GCV ... wild-type p53 protein ... (details)
PMID: 10888647 The chicken anemia virus-derived protein apoptin requires activation of caspases for induction of apoptosis in human tumor cells.	... functional p53 and are inhibited by Bcl-2. ... (details)
PMID: 17036395 Expression of p53, Bax and Bcl-2 proteins in hepatocytes in non-alcoholic fatty liver disease.	... NAFLD induces proapoptotic protein p53 with ... antiapoptotic Bcl-2. (details)
PMID: 17201158 Curcumin-induced apoptosis of human colon cancer colo 205 cells through the production of ROS, Ca2+ and the activation of caspase-3.	... p53 and ... but inhibited the ... of Bcl-2. (details)
PMID: 18201729 Resveratrol induces apoptosis involving mitochondrial pathways in mouse skin tumorigenesis.	... application induces the ... the p53 and ... protein Bcl-2. (details)
PMID: 19227007 Inhibition of progression of erythroleukemia induced by Friend virus in BALB/c mice by natural products--berberine, curcumin and picroliv.	... of Bcl-2,... induce the ... of p53. (details)

Poon et al. "Literome: PubMed-scale genomic knowledge base in the cloud", *Bioinformatics-14*.



The Literome Project

Welcome 24.16.142.216

change to | user id

Microsoft Research

BCL2 ⊣ TP53

Type: negative regulation

Is this interaction correct?

- Yes
- No

clear feedback

PMID: 10037739

Inhibition of p53 transcriptional activity by Bcl-2 requires its membrane-anchoring domain.

Source

The Journal of biological chemistry (3/5/1999)

Abstract

Inhibition of p53 transcriptional activity by Bcl-2 requires its membrane-anchoring domain. We show here that the anti-apoptosis protein Bcl-2 potently inhibits p53-dependent transcriptional activation of various p53-responsive promoters in reporter gene co-transfection assays in human embryonic kidney 293 and MCF7 cells, without affecting nuclear accumulation of p53 protein. In contrast, Bcl-2 (Deltatransmembrane (TM)), which lacks a hydrophobic membrane-anchoring domain, had no effect on p53 activity. Similarly, in MCF7 cells stably expressing either Bcl-2 or Bcl-2 (DeltaTM), nuclear levels of p53 protein were up-regulated upon treatment with the DNA-damaging agents doxorubicin and UV radiation, whereas p53-responsive promoter activity and expression of p21 (CIP1/WAF1) were strongly reduced in MCF7-Bcl-2 cells but not in MCF7-Bcl-2 (DeltaTM) or control MCF7 cells. The issue of membrane anchoring was further explored by testing the effects of Bcl-2 chimeric proteins that contained heterologous transmembrane domains from the mitochondrial protein ActA or the endoplasmic reticulum protein cytochrome b5. Both Bcl-2 (ActA) and Bcl-2 (Cytob5) suppressed p53-mediated transactivation of reporter gene plasmids with efficiencies comparable to wild-type Bcl-2. These results suggest that (a) Bcl-2 not only suppresses p53-mediated apoptosis but also interferes with the transcriptional activation of p53 target genes at least in some cell lines, and (b) membrane anchoring is required for this function of Bcl-2. We speculate that membrane-anchored Bcl-2 may sequester an unknown factor necessary for p53 transcriptional activity.

Direct Search

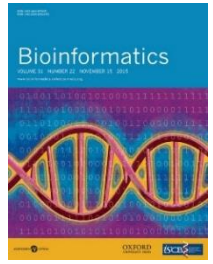
BCL2 → TP53
BCL2 ⊣ TP53
BCL2 → TP53
TP53 → BCL2
TP53 ⊣ BCL2
TP53 → BCL2

Possible interm

BCL2 → TP53
BCL2 → AGTR
BCL2 → AKT1
BCL2 → ANGP
BCL2 → ANXA
BCL2 → ANXA
BCL2 → APAF
BCL2 → ATG5
BCL2 → ATM
BCL2 → ATRA
BCL2 → BAX
BCL2 → BCL1
BCL2 → BCR
BCL2 → BECN
BCL2 → BNIP3 → TP53
BCL2 → BRCA1 → TP53

natural products--berberine, curcumin and picroliv.

Poon et al. "Literome: PubMed-scale genomic knowledge base in the cloud", *Bioinformatics-14*.



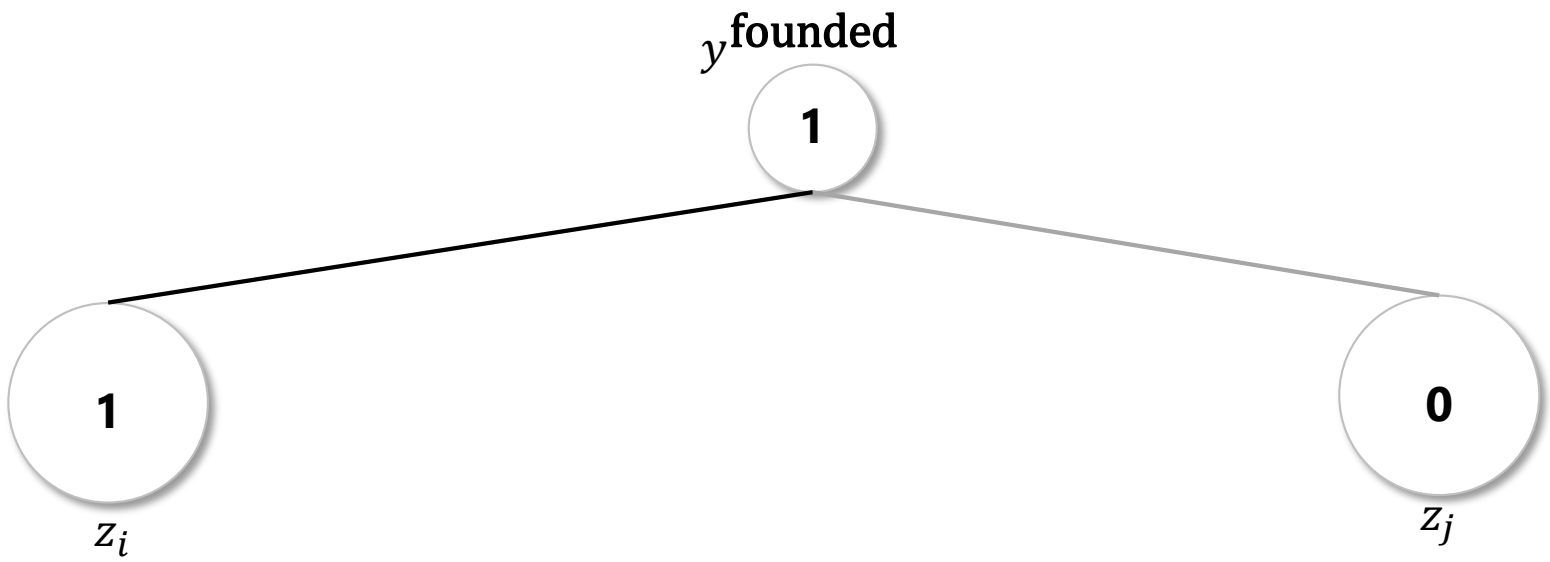
Combatting Noise

Introduce latent variables

Case study: Riedel, Hoffman, Betteridge

Mentioned at least once [Reidel et al. 2010]

Roger McNamee × Elevation Partners

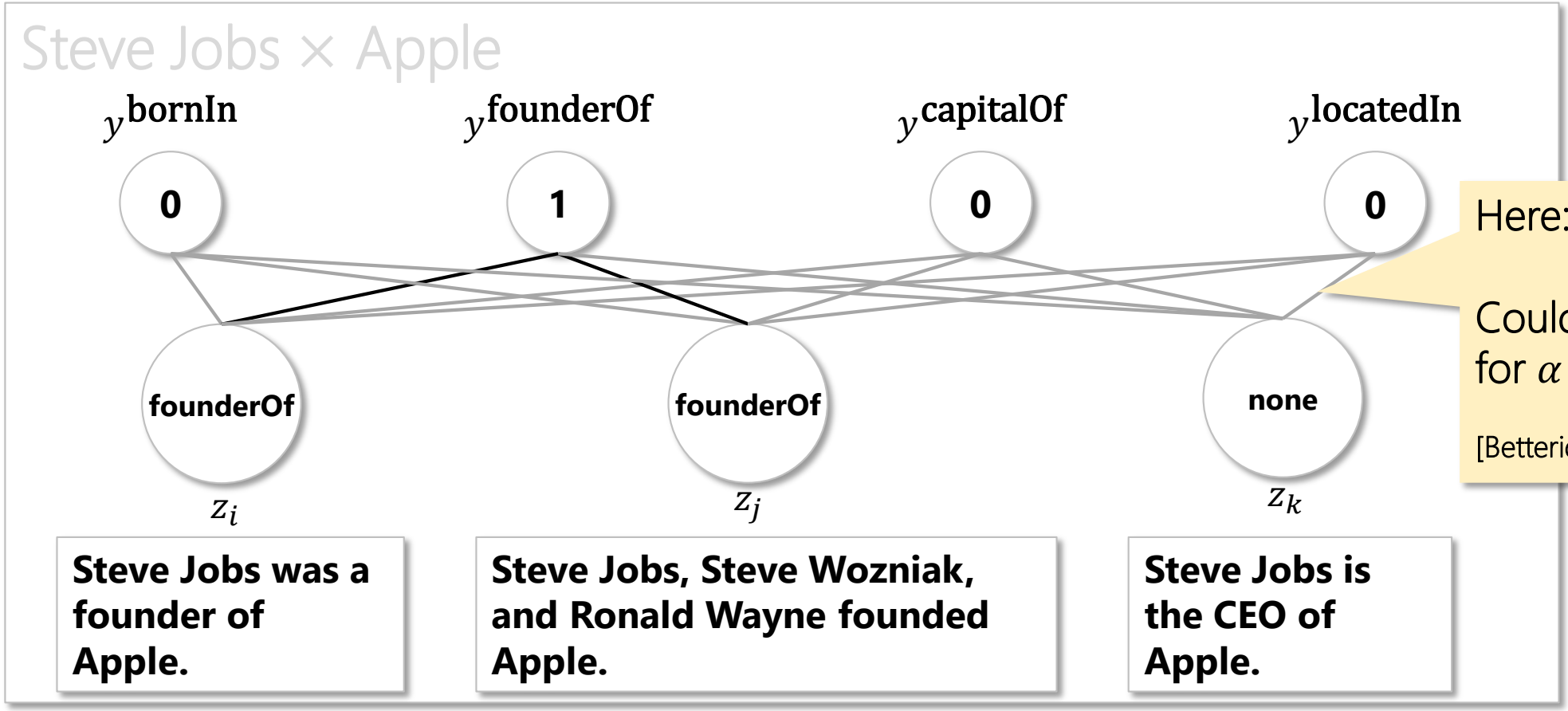


Elevation Partners, the $1.9 billion private equity firm that was founded by Roger McNamee …

Roger McNamee, a managing director at Elevation Partners …

MultiR: multi-instance learning with overlapping relations [Hoffmann 2011]

For each entity pair, construct a graph with one node for each mention, and one for each relation



Here: exists ≥ 0

Could say: true $\geq \alpha$, for $\alpha \in (0,1]$

[Betteridge, Ritter, and Mitchell 2013]

Beyond Classification

Complex semantic structures

Semantic parse → Latent variables

Part 3: Extract Complex Structured Info

Web: Question answering

Biomedicine: Nested event extraction

Recipe

Semantic parse = latent variables

Grounding = Inductive bias

Expectation maximization

Web: Question Answering

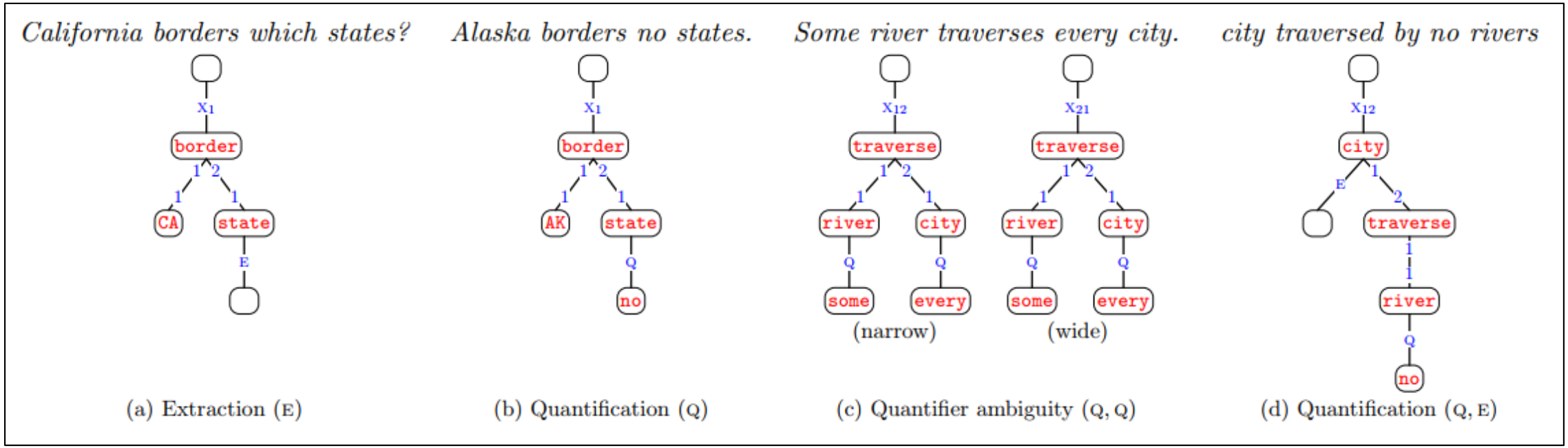
Supervision: Example QA pairs + KB

Grounding: Semantic parse + KB $\rightarrow$ correct answer

E.g., Clarke et al. [2010], Liang et al. [2011].

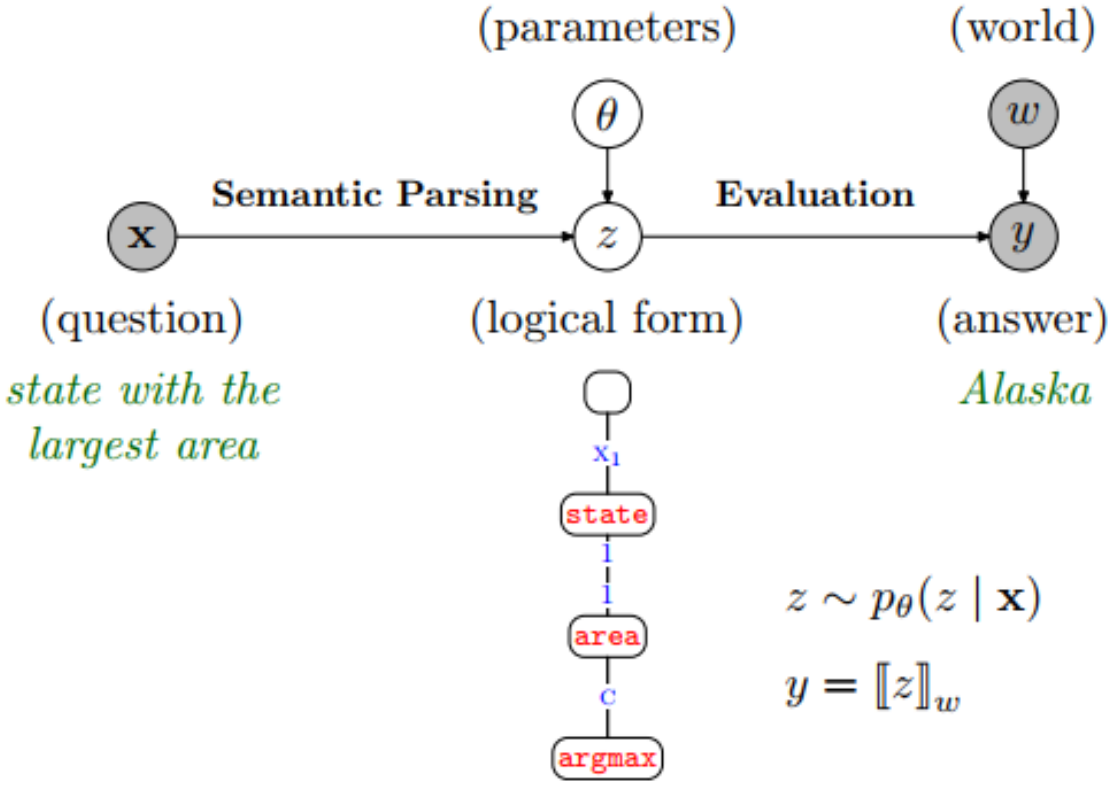
Example: Liang et al. 2011

Grammar: Dependency-based compositional semantics (DCS)



Example: Liang et al. 2011

Grounding: KB query yields correct answer



$$z \sim p_\theta(z \mid \mathbf{x})$$

$$y = [\![z]\!]_w$$

Example: Liang et al. 2011

Discriminative training w/ log-linear model

Problem: Exponential number of semantic parses

Solution: K-best by beam search

Challenge: No correct answer in K-best

Strategy: Constrain Search Space

Krishnamurphy & Mitchell [2012]: Sentences of length ≤ 10

Berant & Liang [2014]: Use manual parse templates

Reddy et al. [2014]: Entities directly connected & known

Yih et al. [2015]: Assume conjunction of binary relations

Work reasonably well for simple factoid questions

Semantic Grammars

Logical form ~ Semantic graph

Relation algebra: Liang et al. [2001], Berant & Liang [2004], …

Combinatory categorial grammar (CCG): Kwiatkowski et al. [2013], Reddy et al. [2014], …

Supervision Signals

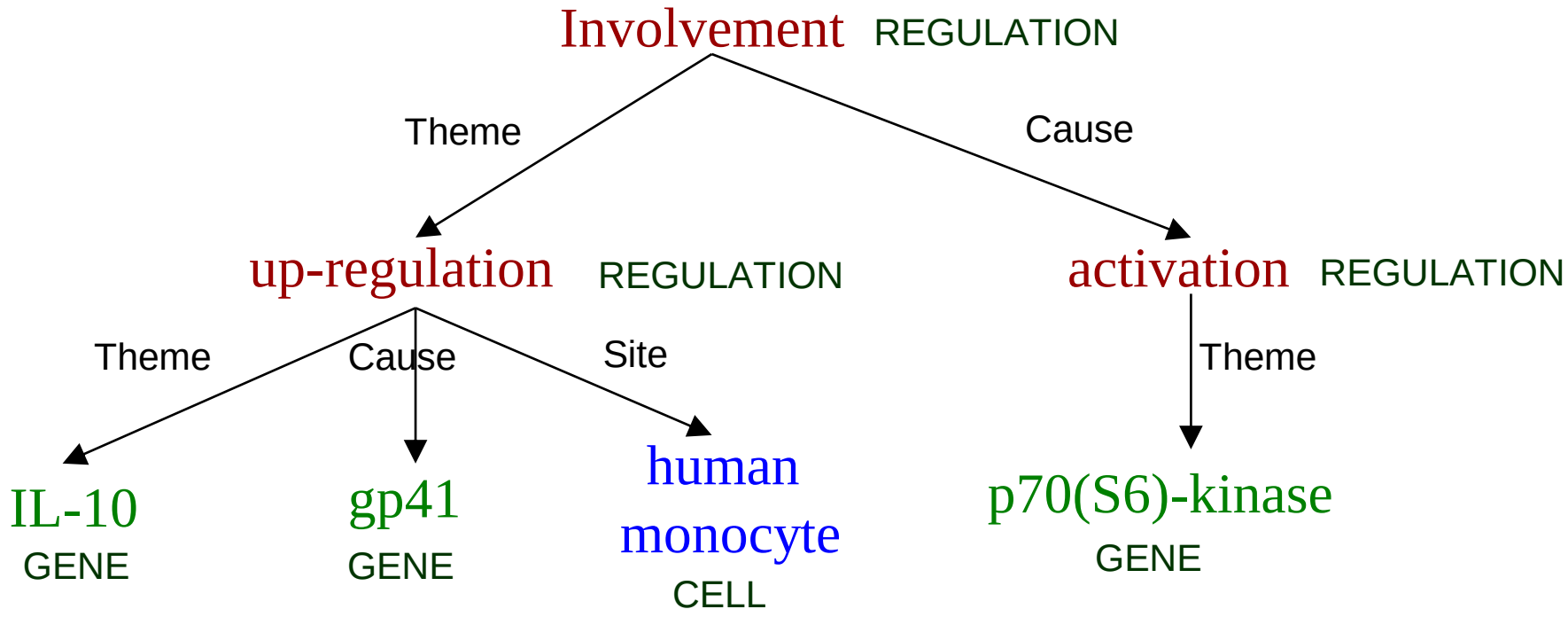
Example question-answer pairs

Relational tuples in KB

Paraphrases

Biomedicine: Nested Event Extraction

Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...



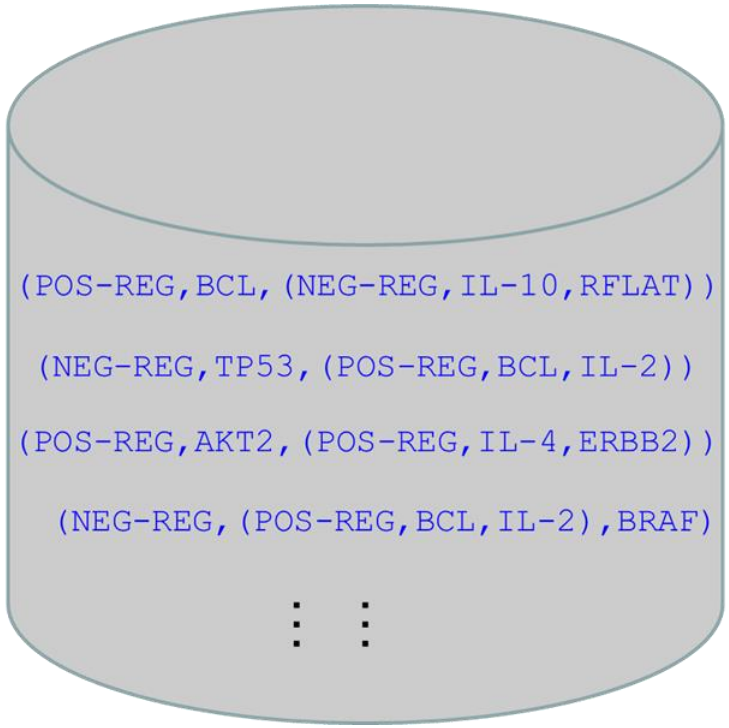
Example: GUSPEE

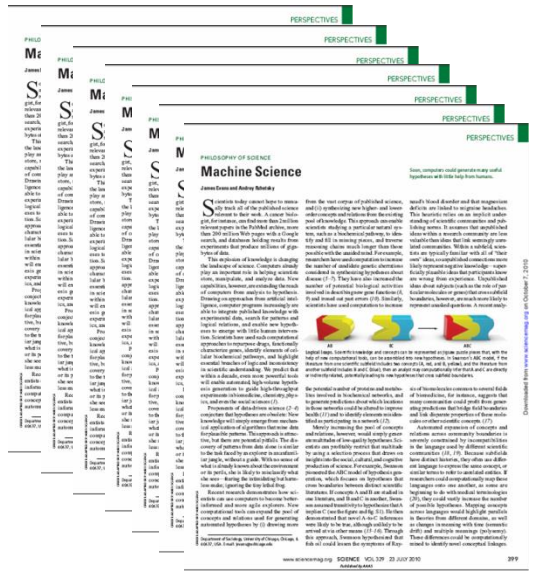
Generalize distant supervision to nested events

Prior: Favor semantic parses grounded in KB

Outperformed 19 out of 24 participants in GENIA Shared Task [Kim et al. 2009]

Parikh et al.. "Grounded Semantic Parsing for Complex Knowledge Extraction", *NAACL-15*.

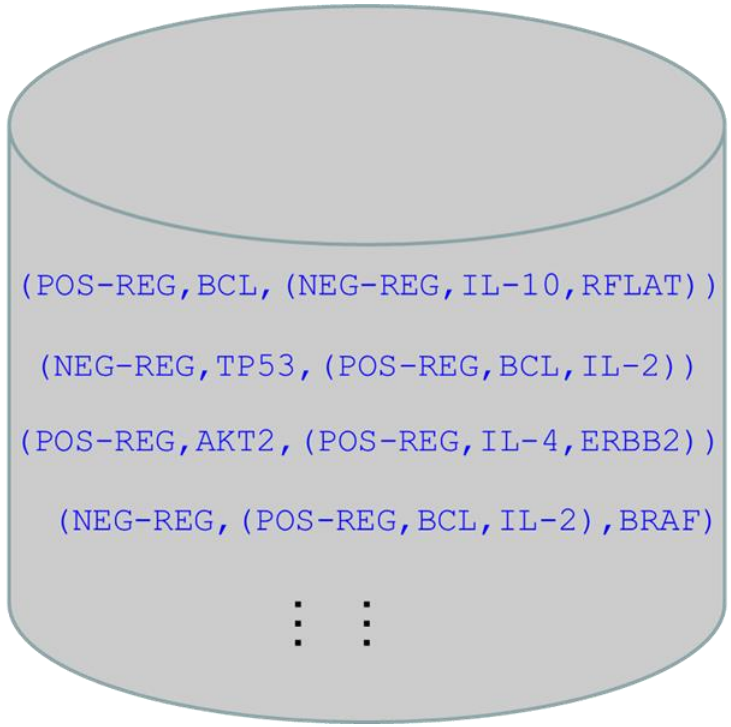


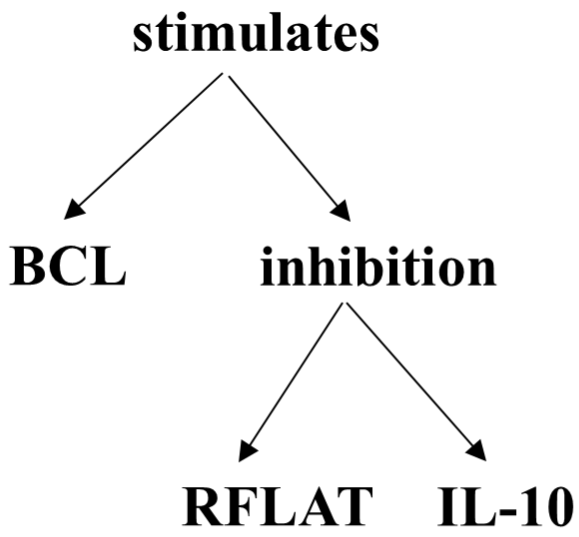


GUSPEE

Semantic parser for event extraction

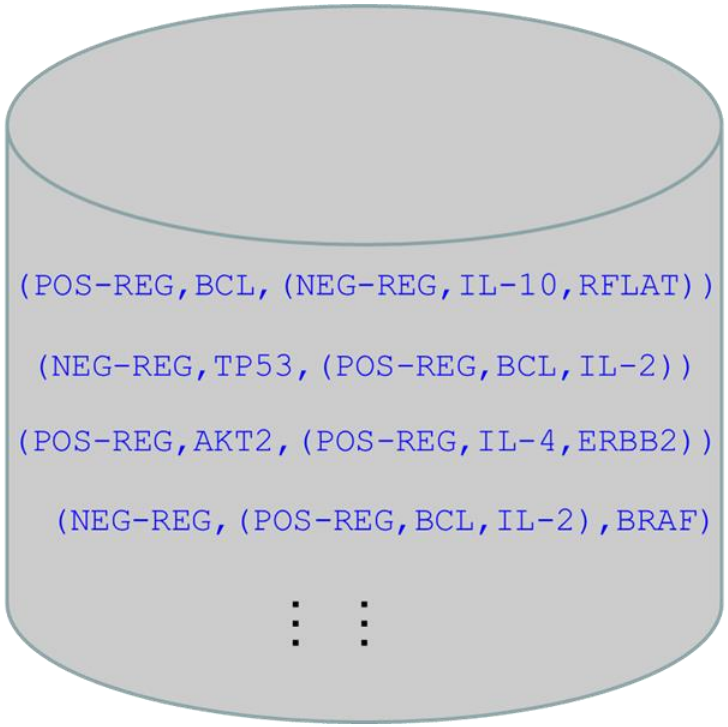
Tree HMM

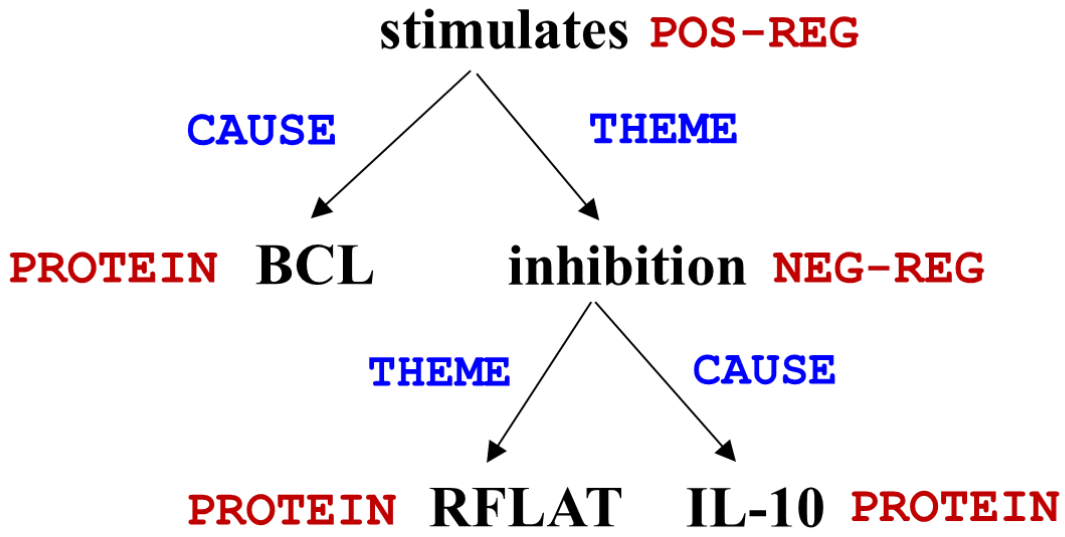




BCL stimulates inhibition of RFLAT by IL-10.

Tree HMM





BCL stimulates inhibition of RFLAT by IL-10.

$$P_\theta(z,t) = \prod_m P_{\text{EMIT}}(t_m | z_m, \theta) \cdot P_{\text{TRANS}}(z_m | z_{\pi(m)}, \theta)$$

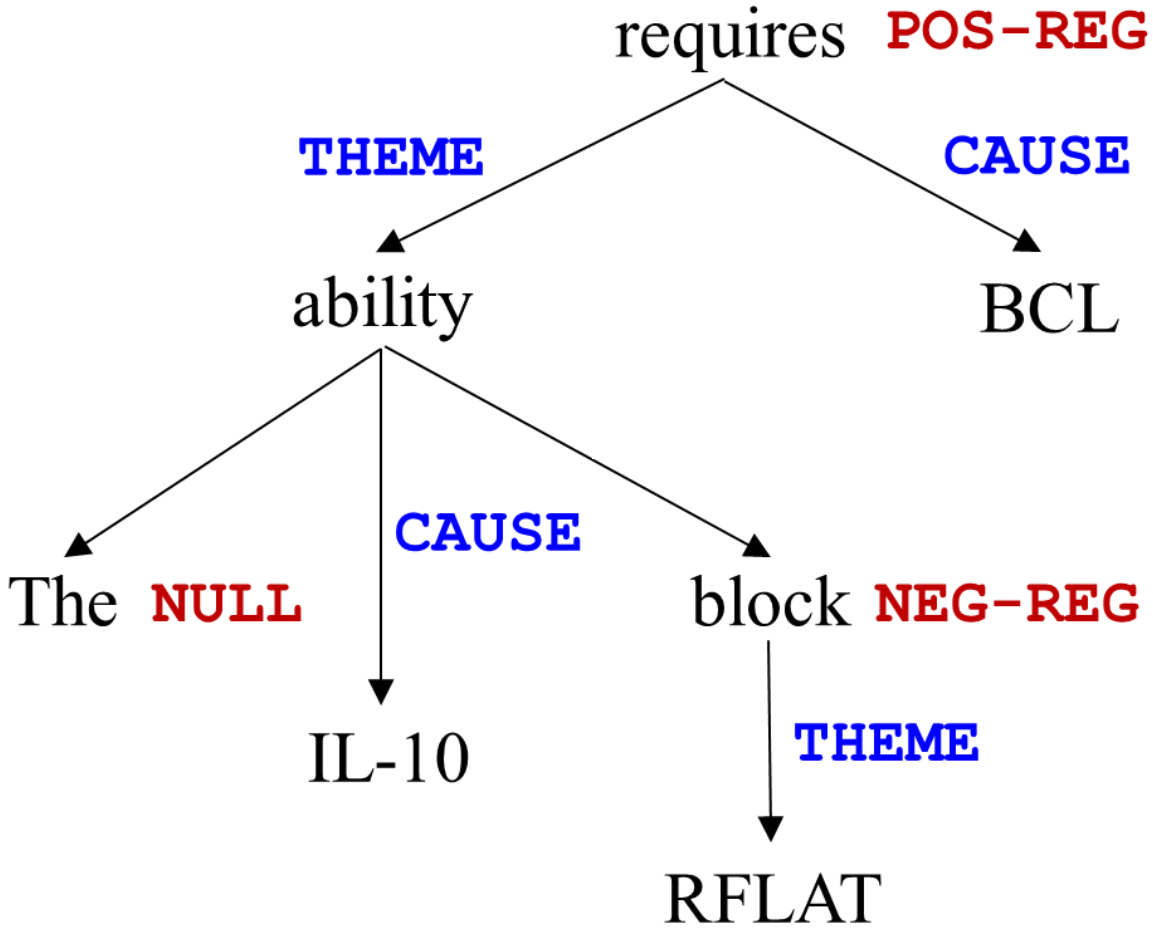
Expectation Maximization

$$\theta^* = \arg\max_{\theta} \quad \log P_\theta(T|K)$$

$$= \arg\max_{\theta} \quad \sum_{t \in T} \log \sum_{z} P_\theta(z,t) \cdot \phi_K(z)$$

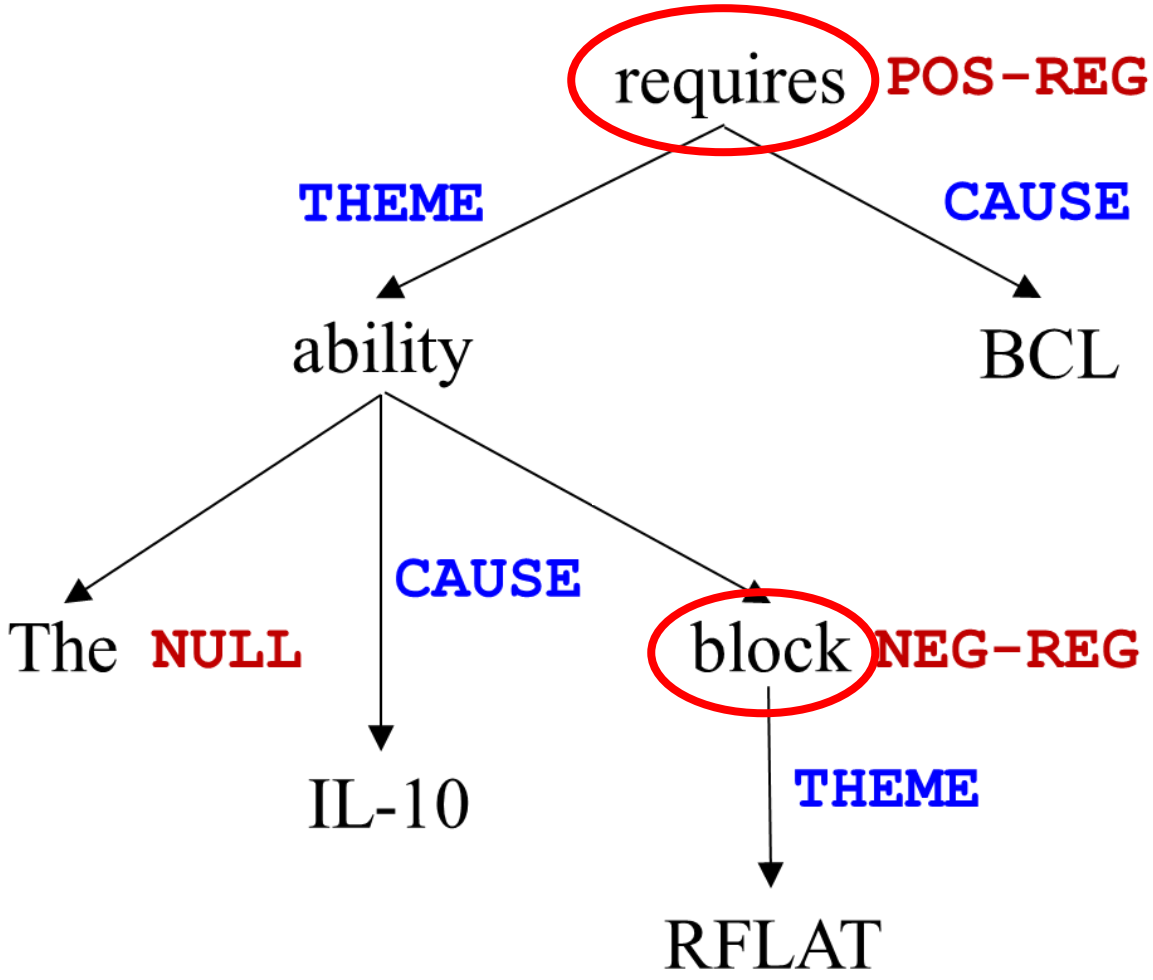
Virtual Evidence

Syntax-Semantics Mismatch



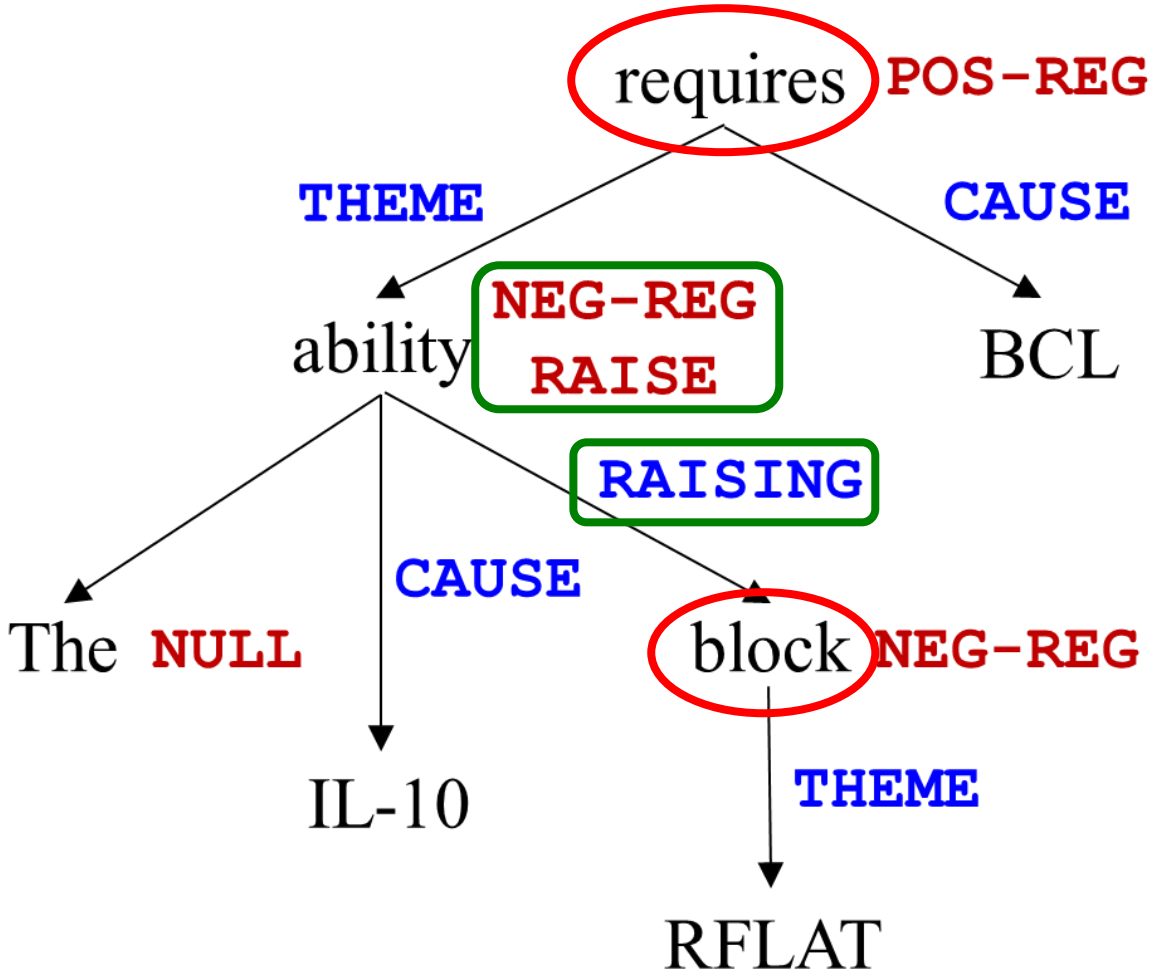
The ability of IL-10 to block RFLAT requires BCL.

Syntax-Semantics Mismatch



The ability of IL-10 to block RFLAT requires BCL.

Syntax-Semantics Mismatch



The ability of IL-10 to block RFLAT requires BCL.

Best Supervised System

Event Type	Rec.	Prec.	F1
Expression	76.4	81.5	78.8
Transcription	49.4	73.6	59.1
Catabolism	65.6	80.0	74.4
Phosphorylation	73.9	84.5	78.9
Localization	74.6	75.8	75.2
Binding	48.0	50.9	49.4
Regulation	32.5	47.1	38.6
Positive_regulation	38.7	51.7	44.3
Negative_regulation	35.9	54.9	43.9
Total Event F1	50.2	62.6	55.7

Preliminary Results

Event Type	Rec.	Prec.	F1
Expression	50.8	41.9	45.9
Transcription	18.3	14.0	15.9
Catabolism	0	0	0
Phosphorylation	36.2	43.6	39.5
Localization	0	0	0
Binding	24.0	42.6	30.7
Regulation	2.5	5.0	3.3
Positive_regulation	11.4	21.4	14.9
Negative_regulation	4.4	16.4	6.9
Total Event F1	19.1	29.4	23.2

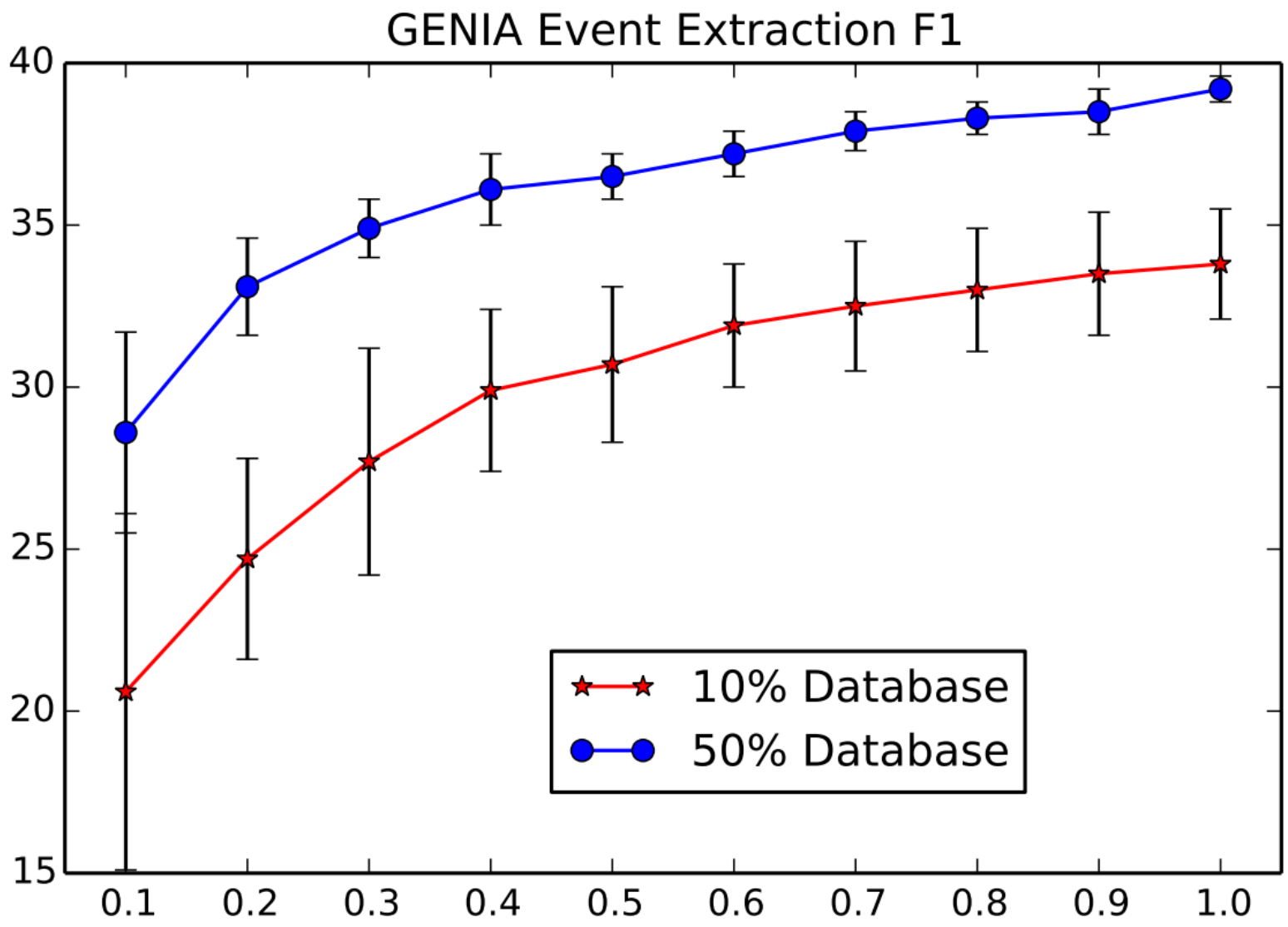
Prototype-Driven Learning

Event Type	Rec.	Prec.	F1
Expression	55.3	88.3	68.0
Transcription	50.0	39.1	43.9
Catabolism	52.4	100.0	68.9
Phosphorylation	61.7	82.9	70.7
Localization	52.8	100.0	69.1
Binding	20.2	92.7	33.2
Regulation	24.1	64.0	35.0
Positive_regulation	17.4	63.8	27.4
Negative_regulation	8.4	52.8	14.5
Total Event F1	27.9	72.2	40.2

Outperformed 19 out of 24 supervised participants

Event Type	Rec.	Prec.	F1
Expression	55.3	88.3	68.0
Transcription	50.0	39.1	43.9
Catabolism	52.4	100.0	68.9
Phosphorylation	61.7	82.9	70.7
Localization	52.8	100.0	69.1
Binding	20.2	92.7	33.2
Regulation	24.1	64.0	35.0
Positive_regulation	17.4	63.8	27.4
Negative_regulation	8.4	52.8	14.5
Total Event F1	27.9	72.2	40.2

Incomplete KB



Next: Improve Semantic Learning

Syntax-semantics mismatch

Ontology matching

Leverage relation interdependencies

Next: More Semantic Complexities

Cellular context

Experimental settings

Relations to diseases, drugs, mutations, …

Scope: Paragraph, document, literature

Part 4: Beyond Sentence Boundary

Why cross sentence

Prior work

Generalize distant supervision

Graph LSTM

Challenge: Cross-Sentence Relation Extraction

The p56Lck inhibitor Dasatinib was shown to enhance apoptosis induction by dexamethasone in otherwise GC-resistant CLL cells.

This finding concurs with the observation by Sade showing that Notch-mediated resistance of a mouse lymphoma cell line could be overcome by inhibiting p56Lck.

Dasatinib could be used to treat Notch-mutated tumors.

TREAT(Dasatinib, Notch)

Challenge: Cross-Sentence Relation Extraction

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

Gefitinib could be used to treat tumors w. EGFR mutation L858E.

TREAT(Gefitinib, EGFR, L858E)

Related Work

Cross-sentence: Received little attention

- Supervised [Swampillai & Stevenson 2011]
- Newswire/Web: Single sentences often suffice

Distant supervision: Focused on single-sentence

- Entity-centric attributes [Wu & Weld 2007; TAC KBP]
- Coreference [Koch et al. 2014; Augenstein et al. 2016]

DISCREX: Distant Supervision → Cross-Sentence

Document graph: Unified representation

Linguistic analysis: Syntax, discourse, coreference, etc.

Features: Multiple dependency paths

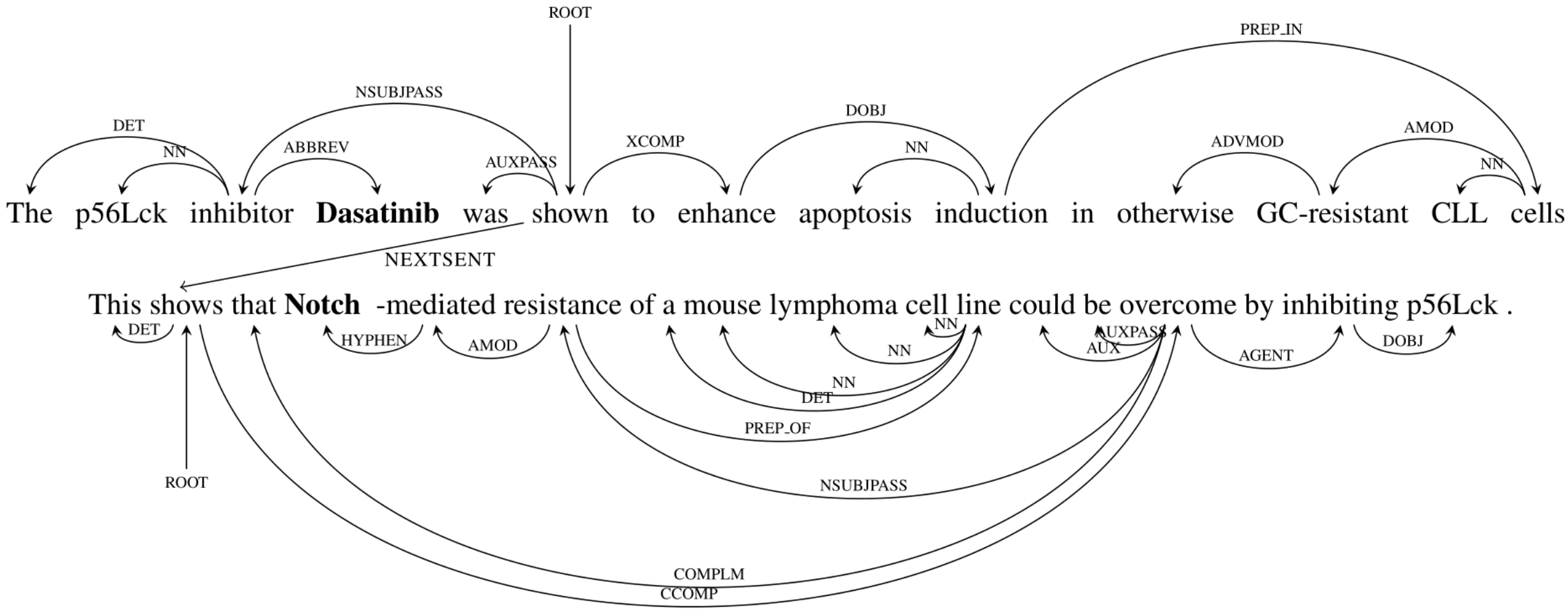
Candidate selection: Minimal-span

Quirk & Poon. "Distant Supervision for Relation Extraction beyond the Sentence Boundary", *EACL-17*.

EACL 2017

Document Graph

Sequence, syntax, discourse



Features

Prior work: Used single shortest path

DISCREX: Multiple paths help

Templates

- Nodes: Token, lemma, POS
- Whole paths
- Path n-grams

Distant Supervision: Minimal-Span Candidates

Imatinib could be used to treat KIT-mutated tumors.

Since amuvatinib inhibits KIT, we validated MET kinase inhibition as the primary cause of cell death.

Additionally, imatinib is known to inhibit KIT.

Distant Supervision: Minimal-Span Candidates

Imatinib could be used to treat KIT-mutated tumors.

Since amuvatinib inhibits KIT, we validated MET kinase inhibition as the primary cause of cell death.

Additionally, imatinib is known to inhibit KIT.

Not minimal-span

Experiments: Molecular Tumor Board

Drug-gene interaction

Distant supervision

- Knowledge bases: GDKD
- Text: PubMed Central (~ 1 million full-text articles)

GDKD

Gene-Drug Knowledge Database [Dienstmann et al. 2015]

Disease	Gene	Variant	Description	Effect	Association_1	Therapeutic context_1
ALL	ABL1	T315A	missense mutation	gain-of-function	response	nilotinib, ponatinib
ALL	ABL1	T315I	missense mutation	gain-of-function	response	ponatinib
ALL	ABL1	F317L/V/I/C	missense mutation	gain-of-function	response	nilotinib, ponatinib
ALL	ABL1	F359V/C/I	missense mutation	gain-of-function	response	dasatinib, ponatinib
CML	ABL1	T315A	missense mutation	gain-of-function	response	nilotinib, bosutinib, ponatinib
CML	ABL1	T315I	missense mutation	gain-of-function	response	ponatinib
CML	ABL1	F317L/V/I/C	missense mutation	gain-of-function	response	nilotinib, bosutinib, ponatinib
CML	ABL1	F359V/C/I	missense mutation	gain-of-function	response	dasatinib, bosutinib, ponatinib
ALL	ABL1	Y253H	missense mutation	gain-of-function	response	dasatinib, ponatinib
ALL	ABL1	E255K/V	missense mutation	gain-of-function	response	dasatinib, ponatinib

PubMed-Scale Extraction

Relations	Single-Sent.	Cross-Sent.
Candidates	169,168	332,969
$p \geq 0.5$	32,028	64,828
$p \geq 0.9$	17,349	32,775
GDKD	162	

PubMed-Scale Extraction

Relations	Single-Sent.	Cross-Sent.
Candidates	169,168	332,969
$p \geq 0.5$	32,028	64,828
$p \geq 0.9$	17,349	32,775
GDKD	162	

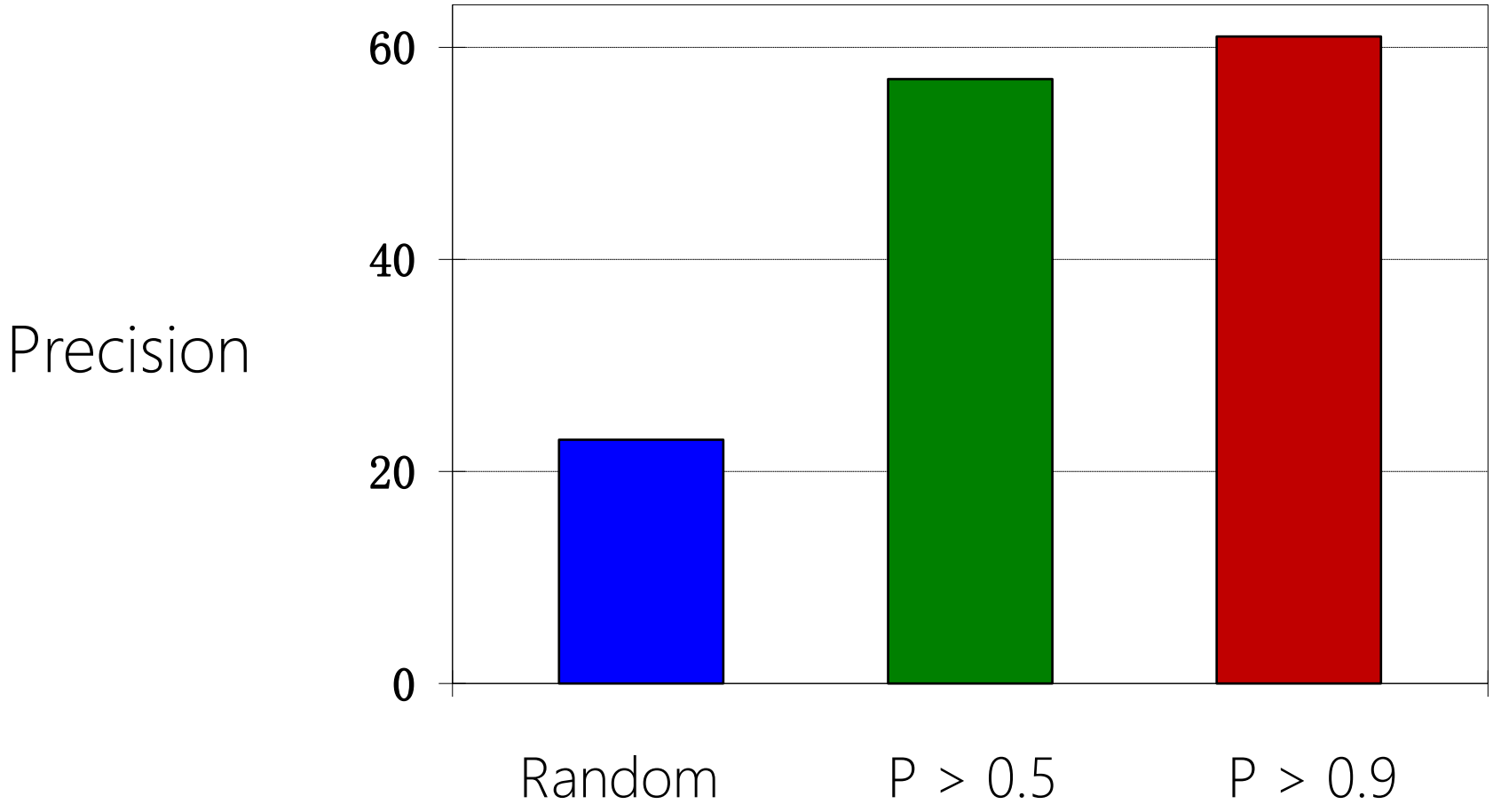
Cross-sentence extraction doubles the yield

PubMed-Scale Extraction

Relations	Single-Sent.	Cross-Sent.
Candidates	169,168	332,969
$p \geq 0.5$	32,028	64,828
$p \geq 0.9$	17,349	32,775
GDKD	162	

Orders of magnitude more knowledge by machine reading

Manual Evaluation



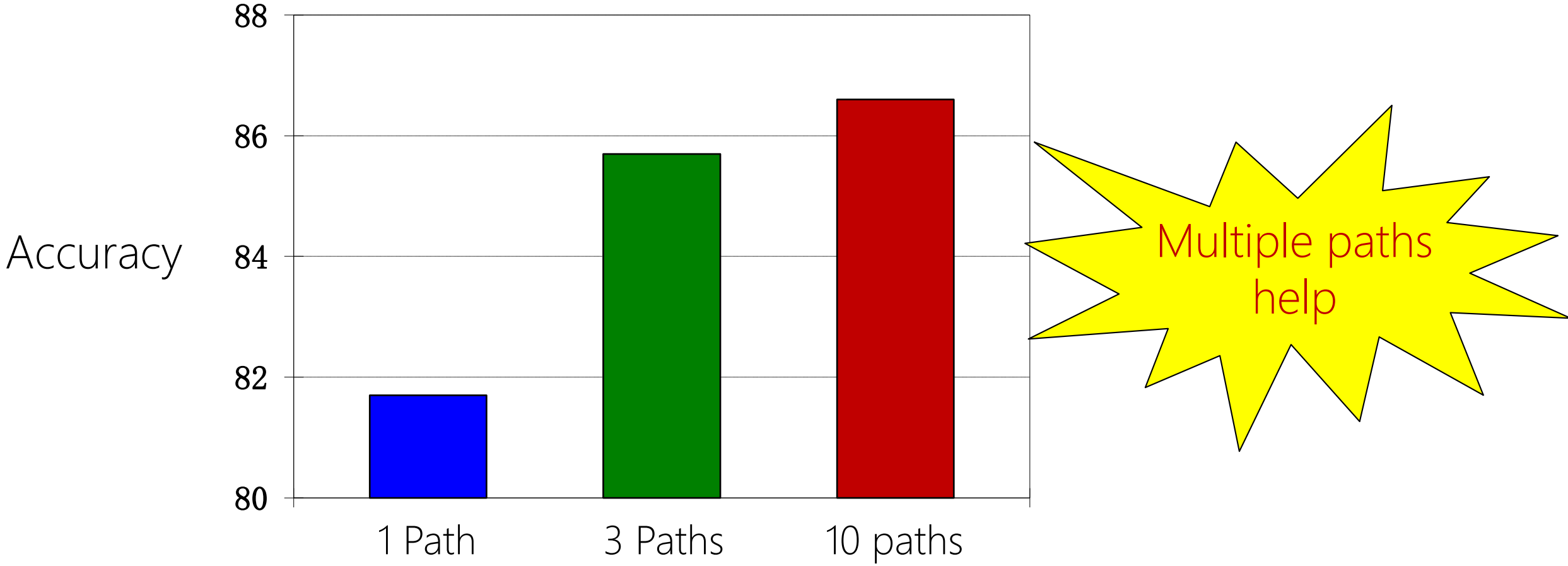
Automatic Evaluation

Distant-supervision: Treat labels as gold

Five-fold cross-validation

Balanced dataset → Report average accuracy

Shortest Paths → Features



Other Take-Aways

Prioritizing dependency edges helps

Discourse / coreference no impact yet

Generalize to N-ary Relations

The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.

All patients were treated with gefitinib and showed a partial response.

Peng et al. "Cross-Sentence N-ary Relation Extraction with Graph LSTM", *TACL-17*.

TACL 2017

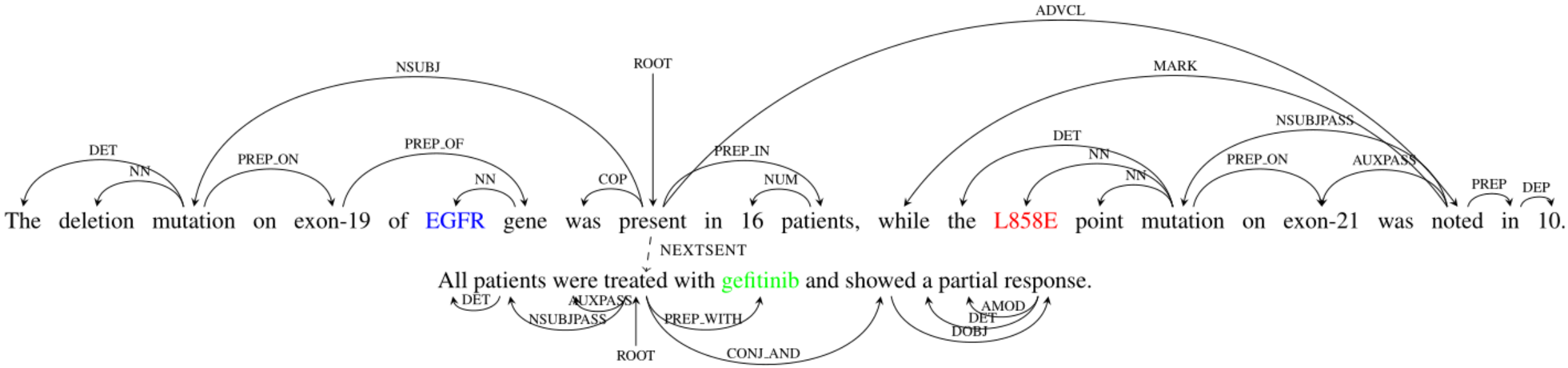
Why LSTM?

Cross-sentence → Features become much sparser

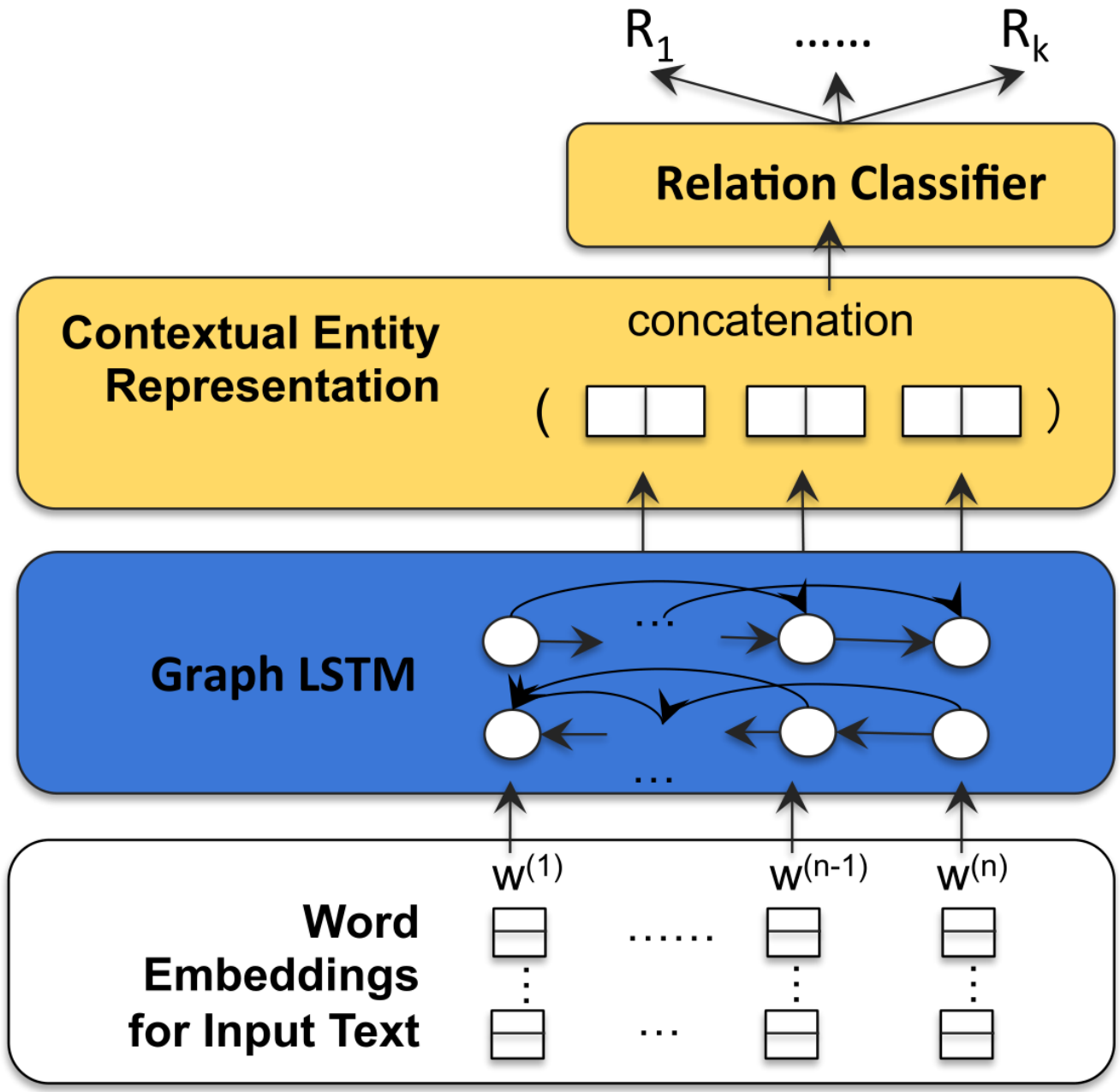
N-ary → Want to scale to arbitrary n

Multi-task learning: Easy

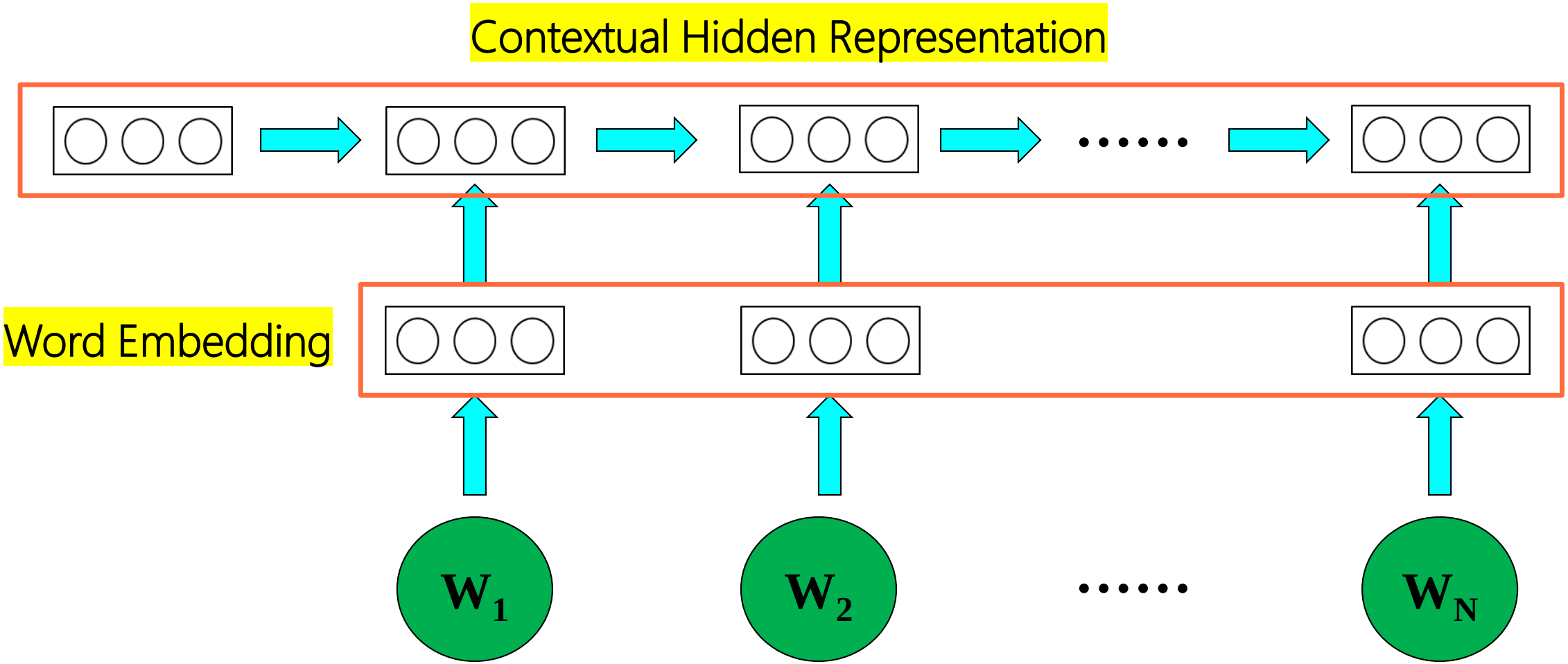
Why Graph?


ADVCL
NSUBJ
ROOT
MARK
NSUBJPASS
DET
PREP_ON
PREP_OF
PREP_IN
DET
NN
PREP_ON
AUXPASS
NN
NN
COP
NUM
NN
PREP
DEP
The deletion mutation on exon-19 of EGFR gene was present in 16 patients, while the L858E point mutation on exon-21 was noted in 10.
NEXTSENT
All patients were treated with gefitinib and showed a partial response.
DET
AUXPASS
NSUBJPASS
PREP_WITH
AMOD
DET
DOBJ
ROOT
CONJ_AND

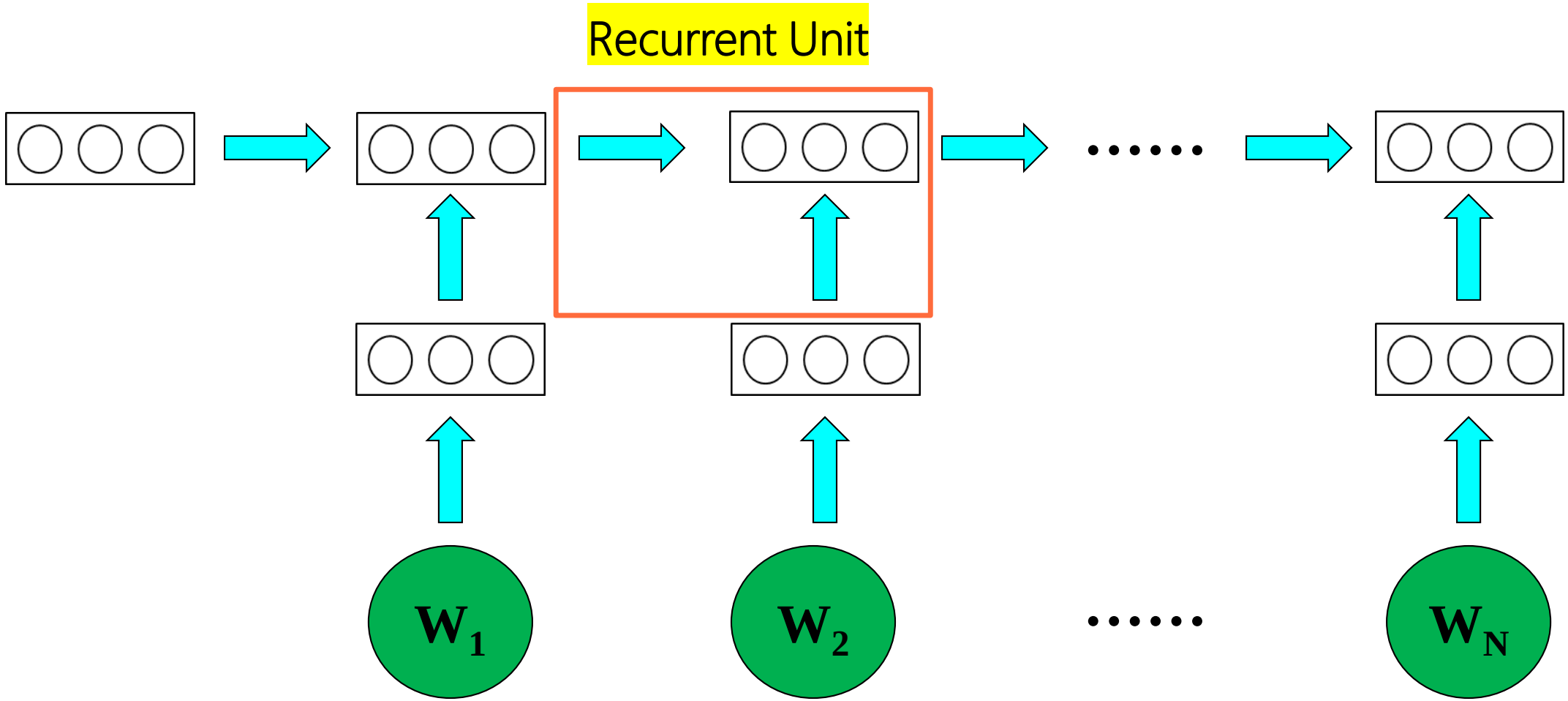
Graph LSTM


R1
.....
Rk
Relation Classifier
Contextual Entity Representation
concatenation
Graph LSTM
w(1)
w(n-1)
w(n)
Word Embeddings for Input Text

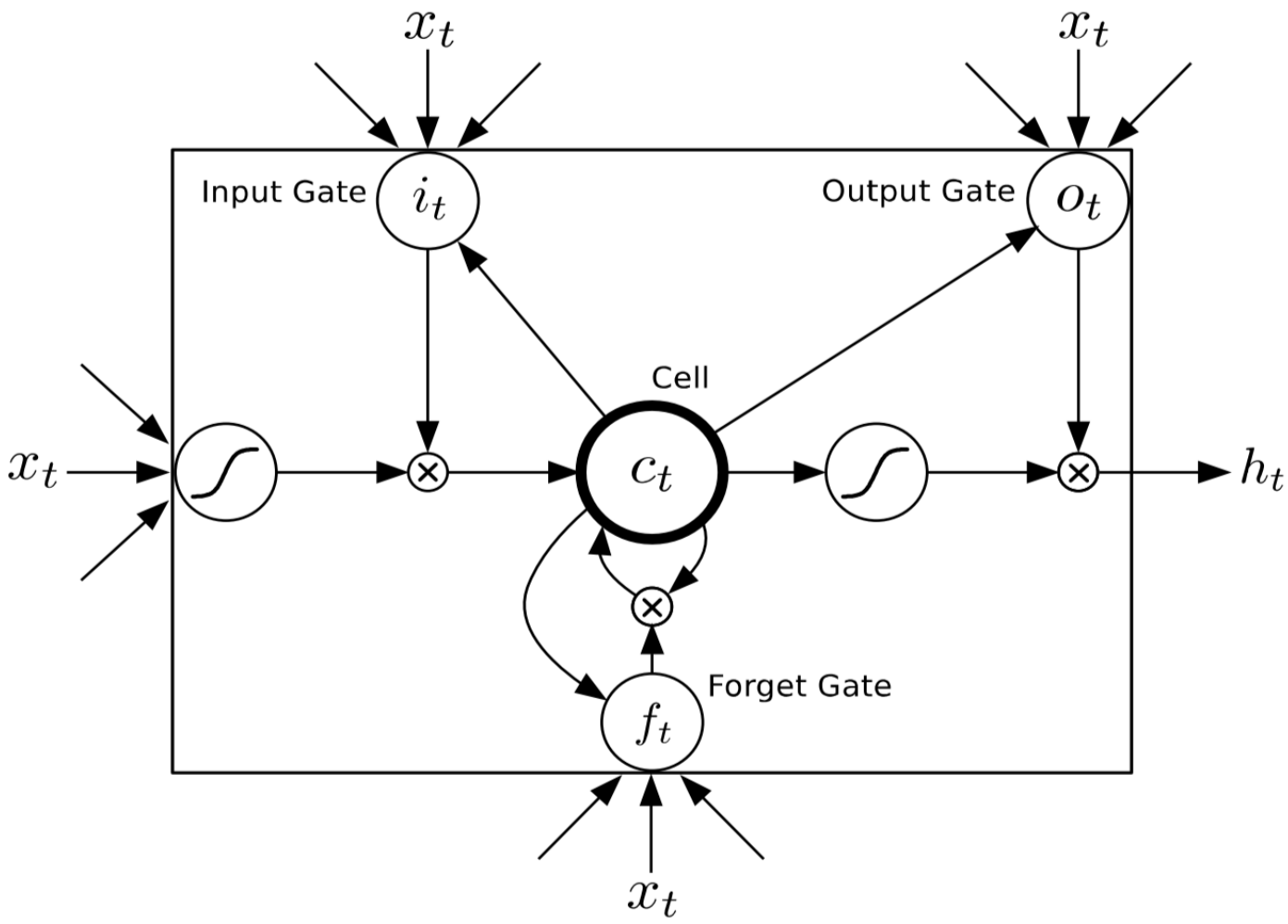
Recurrent Neural Network



Recurrent Neural Network



Long Short-Term Memory (LSTM)



Little Work beyond Linear-Chain

NLP: Tree LSTM

Programming verification: Graph Neural Network

Challenge in Backpropagation

Standard approach

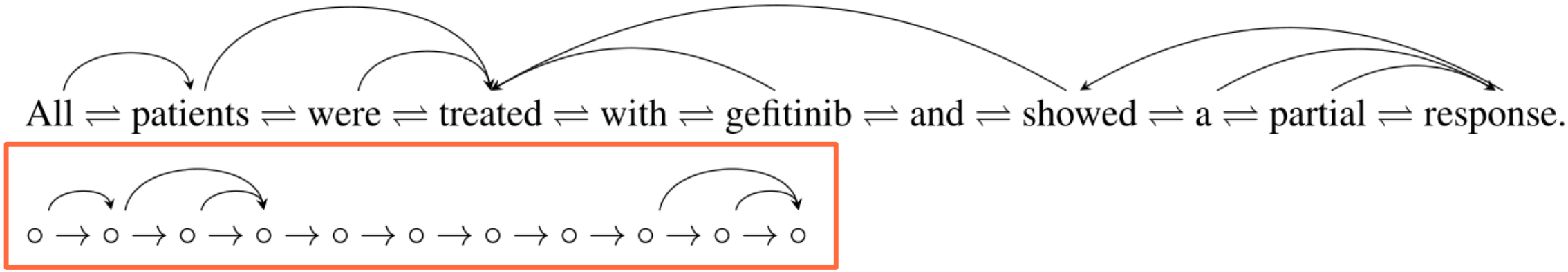
- Unroll recurrence for a number of steps
- Analogous to loopy belief propagation (LBP)

Problems

- Expensive: Many steps per iteration
- Similar to LBP: Oscillation, failure to converge

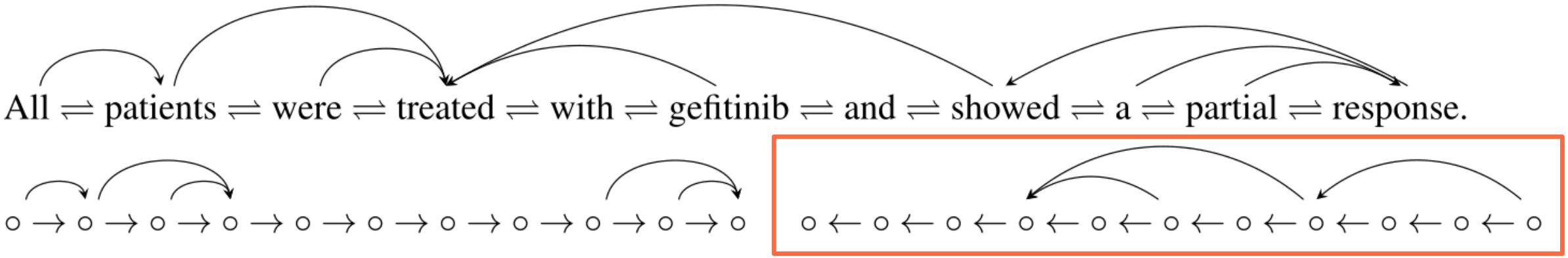
Asynchronous Update

Asynchronous Update



Forward Pass

Asynchronous Update



Backward Pass

Domain: Molecular Tumor Board

Ternary interaction: (drug, gene, mutation)

Distant supervision

- Knowledge bases: GDKD + CIVIC
- Text: PubMed Central articles (~ 1 million full-text articles)

PubMed-Scale Extraction

	Single-Sent.	Cross-Sent.
Candidates	10,873	57,033
$p \geq 0.5$	1,408	4,279
$p \geq 0.9$	530	1,461
GDKD + CIVIC	59	

PubMed-Scale Extraction

	Single-Sent.	Cross-Sent.
Candidates	10,873	57,033
$p \geq 0.5$	1,408	4,279
$p \geq 0.9$	530	1,461
GDKD + CIVIC	59	

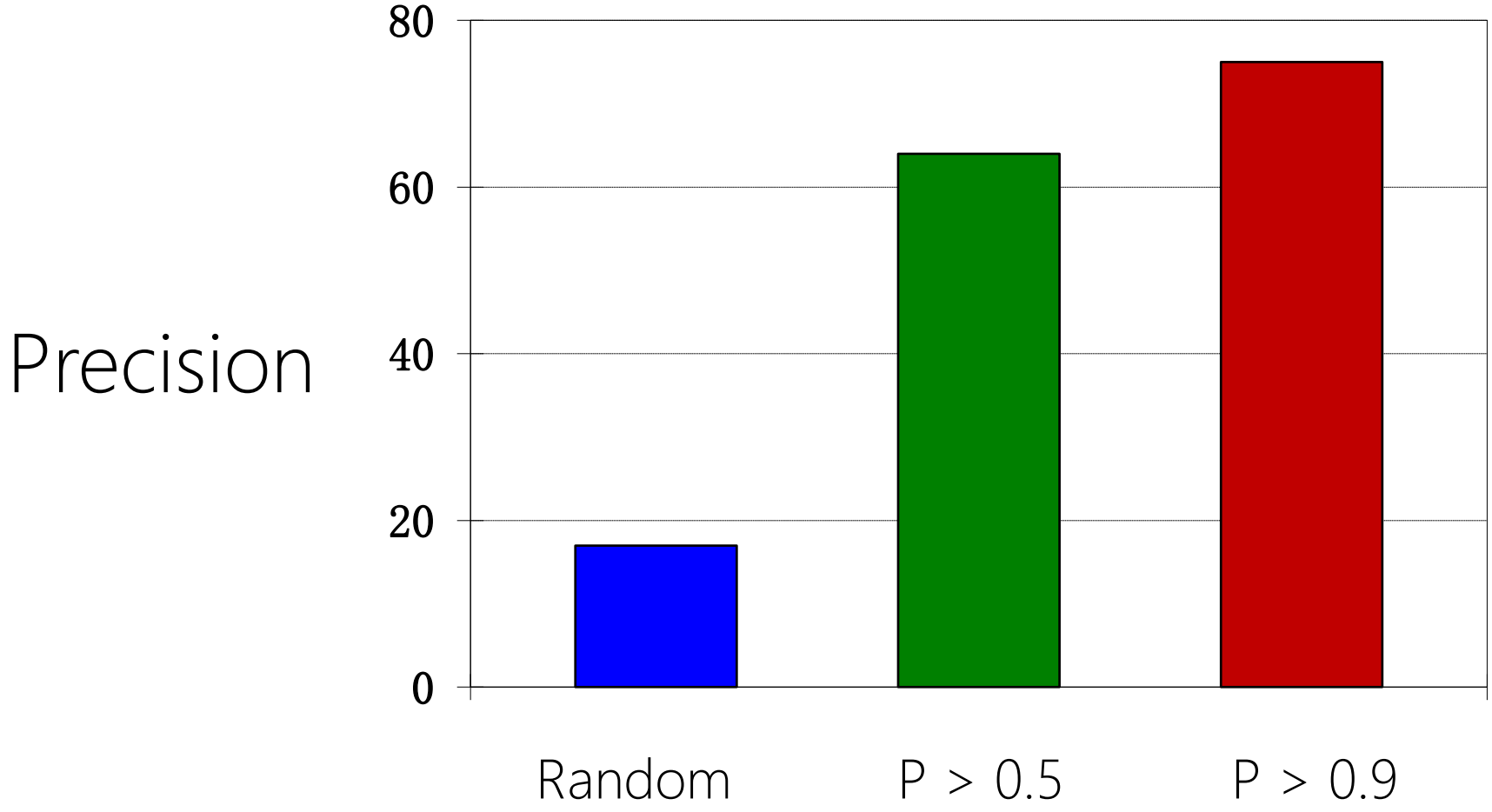
Cross-sentence extraction triples the yield

PubMed-Scale Extraction

	Single-Sent.	Cross-Sent.
Candidates	10,873	57,033
$p \geq 0.5$	1,408	4,279
$p \geq 0.9$	530	1,461
GDKD + CIVIC	59	

Machine reading extracted orders of magnitudes more knowledge

Manual Evaluation

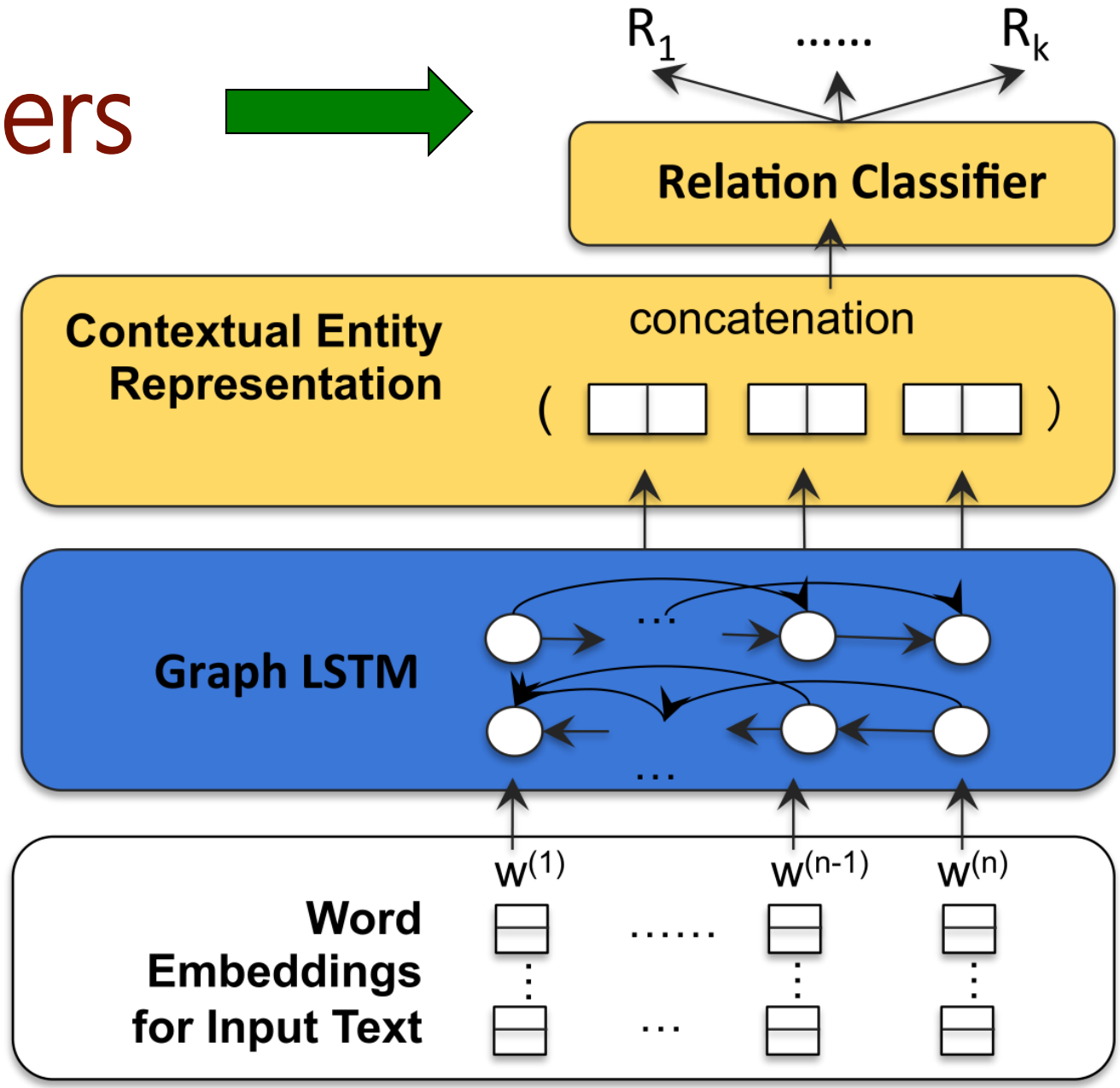


Multi-Task Learning

Leverage related tasks w/ more supervision

E.g., binary sub-relations

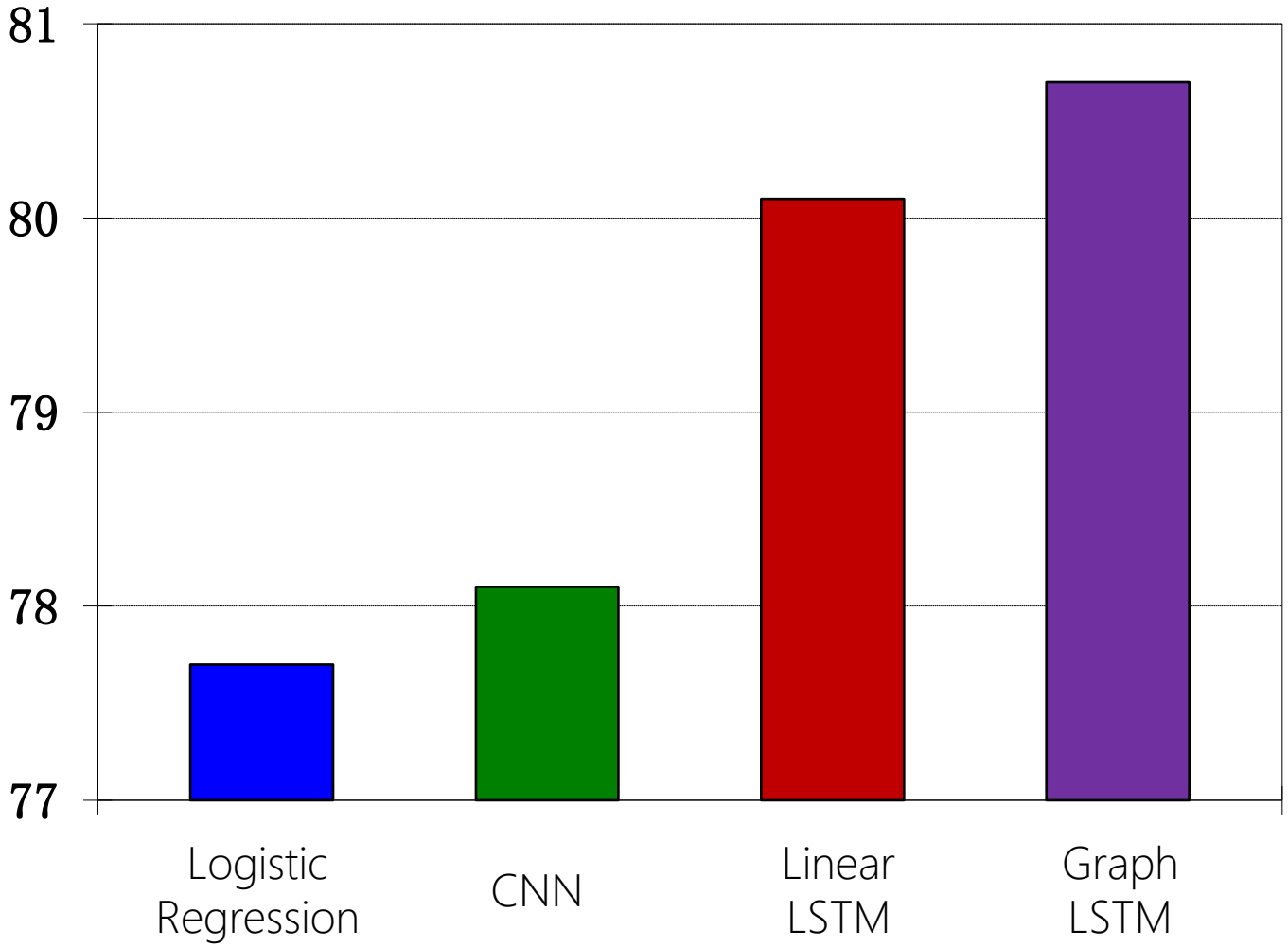
Just add top classifiers



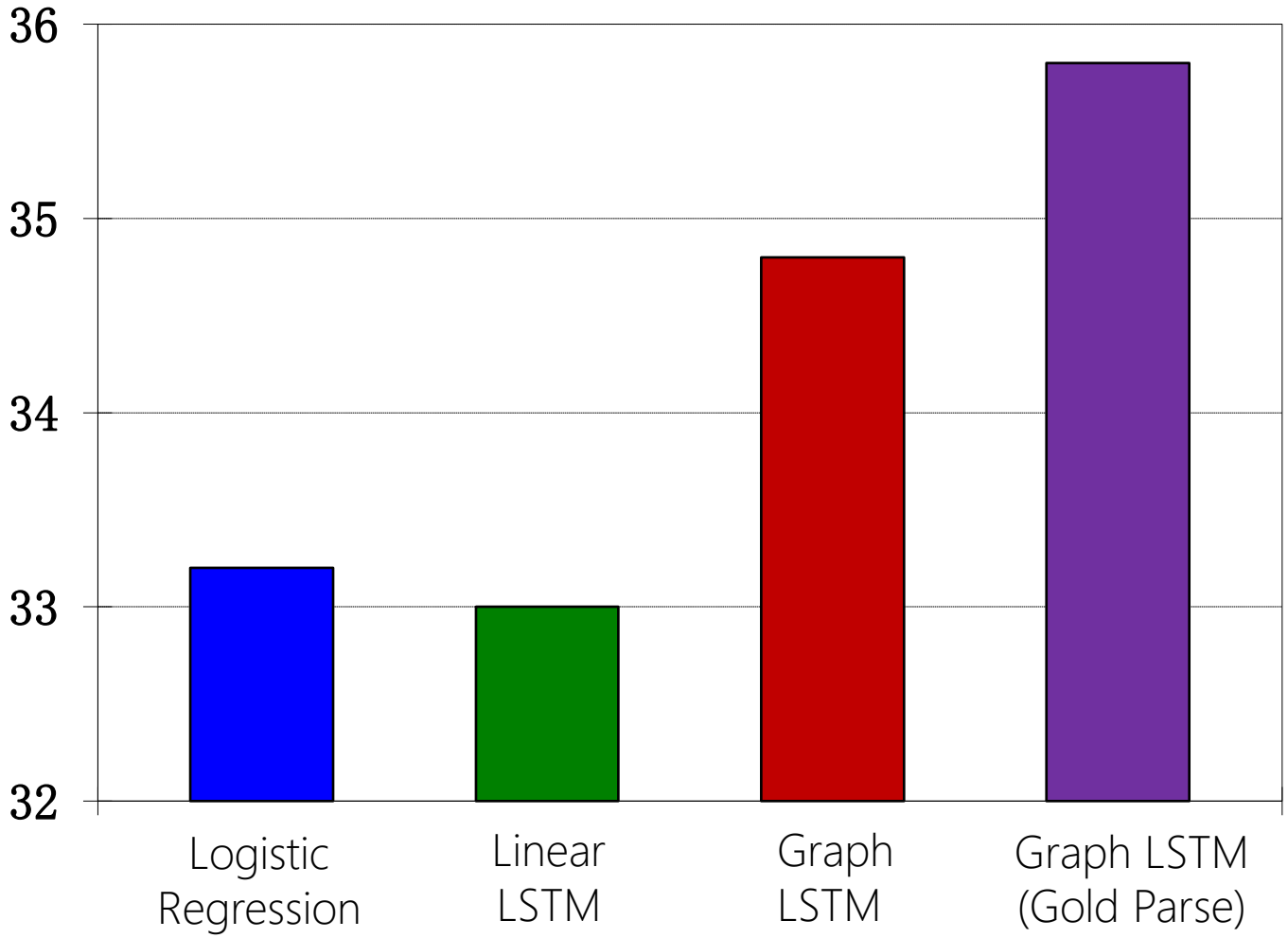
Multi-Task Learning

	Drug-Gene-Mutation	Drug-Mutation
Single-Task	80.7	76.7
Multi-Task	**82.1**	**78.4**

System Comparison



GENIA: Impact of Syntactic Parses



Take-Aways

Linear: Capture some long-ranged dependencies

Graph: Quality of linguistic analysis matters

What's Next?

Parametrization

Joint syntax & semantics

Multi-task learning: Imbalance

Discourse modeling

Part 5: Reasoning

Reasoning with embeddings of entities and relations

- Representing texts

Reasoning with relation paths (PRA)

A hybrid method embedding triples, text, and relation paths

So far: Relationships Directly Expressed in Text

Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.

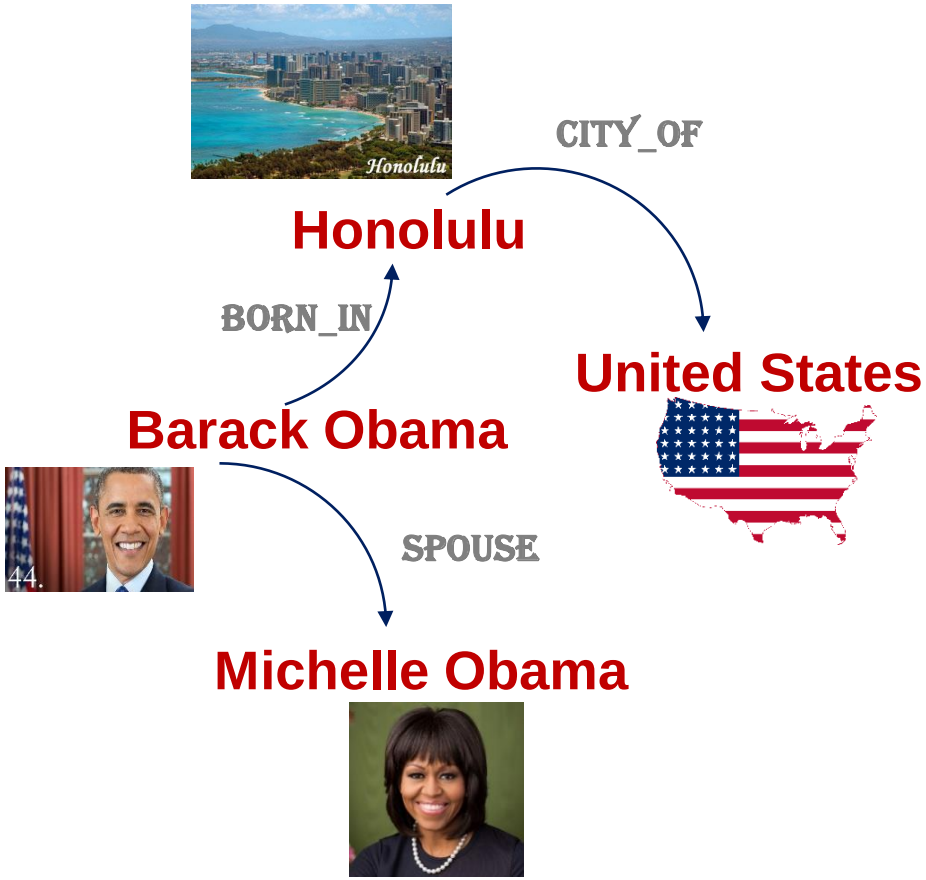
↓

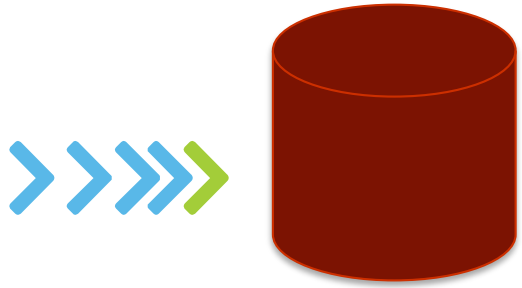
negative_regulation(P53,BCL-2)

Reasoning: combining several pieces of relevant information.

General Domain Knowledge Base

Captures world knowledge by storing properties of millions of entities, as well as relations among them





Freebase

DBpedia

...

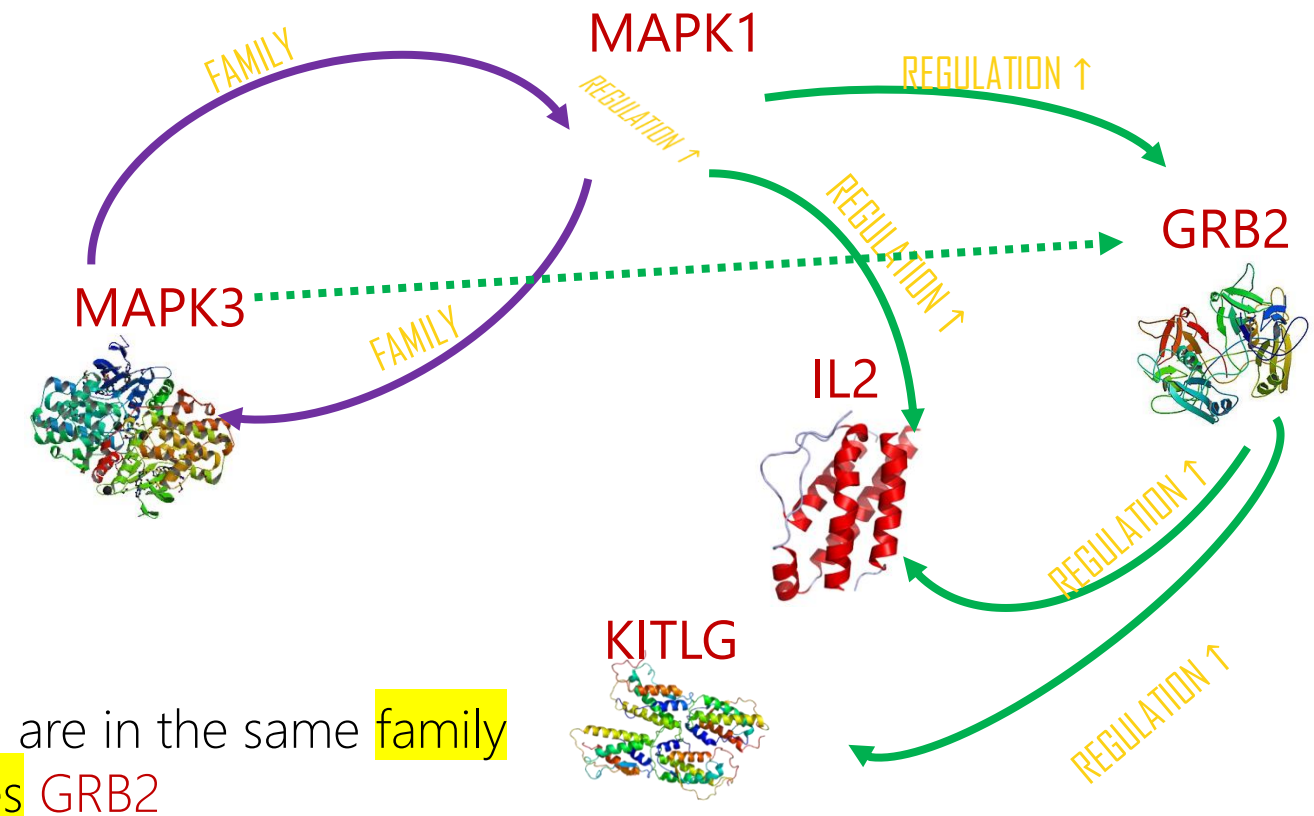
OpenIE/ReVerb

Reasoning

Barack Obama born-in Honolulu
Honolulu city-of Unites States

Likely that Barack Obama nationality USA

Genomics Knowledge Base (Network)



MAPK3 and MAPK1 are in the same family
MAPK1 up-regulates GRB2

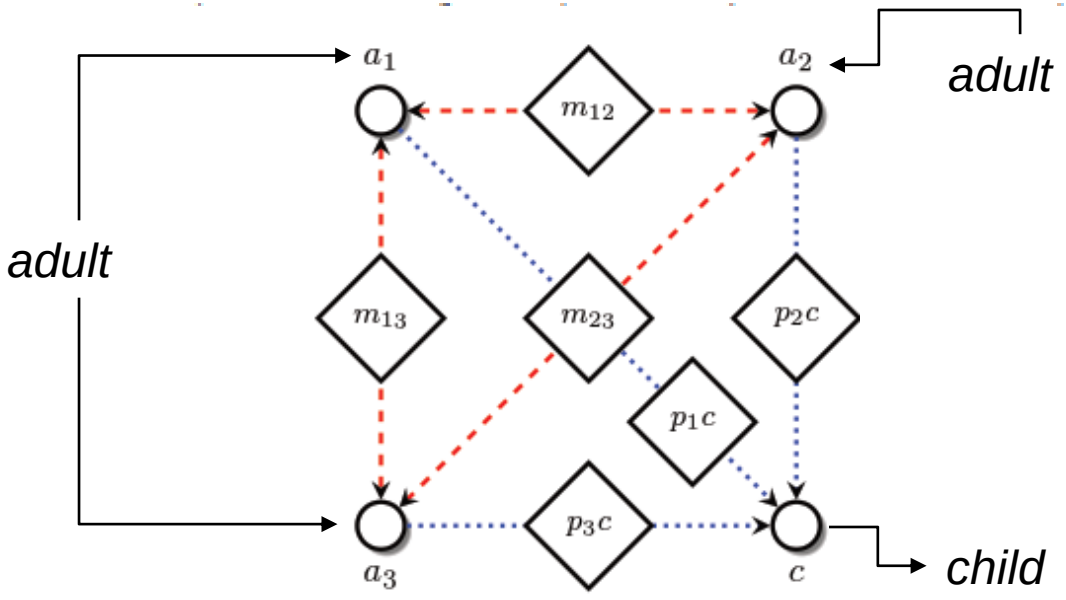
Likely that MAPK3 up-regulates GRB2

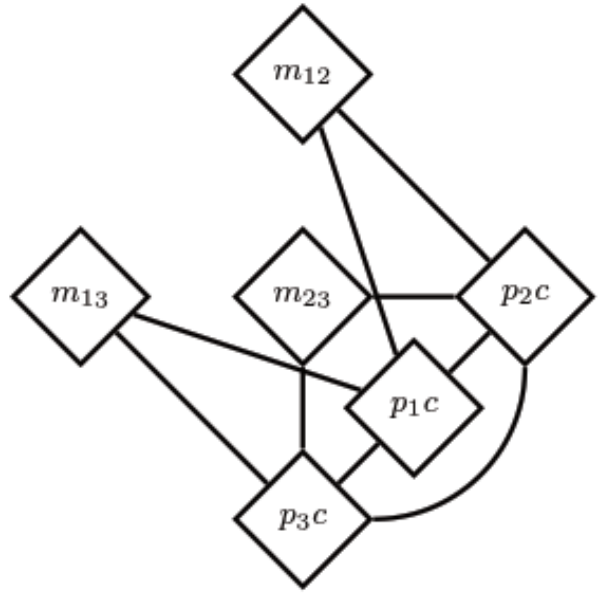
Reasoning with Knowledge Bases -I

Statistical relational learning [Getoor & Taskar, 2007]

- Modeling dependencies among the truth values of multiple possible relations

$$F_1 : (x, parentOf, z) \wedge (y, parentOf, z) \Rightarrow (x, marriedTo, y)$$





- Can be prohibitively expensive (e.g. marginal inference is exponential in the treewidth for Markov Random Fields)

Reasoning with Knowledge Bases - II

Knowledge base embedding

- Assumes truth values of facts are independent given latent features (embeddings) of entities and relations
- Can be very efficient (e.g. matrix multiplication for prediction)
- Has difficulty generalizing when graph has many small cliques

Path ranking methods (e.g., random walk) [e.g., Lao+ 2011]

- Assumes truth values of unknown facts are independent given observed facts
- Difficulty capturing dependencies through long relation paths
- Sparsity when number of relation types is large

Hybrid of path ranking and embedding methods

Overview of Part 5

Reasoning with embeddings of entities and relations

- Representing texts

Reasoning with relation paths (PRA)

A hybrid method embedding triples, text, and relation paths

Basic Approach: Continuous Representations (Embeddings)

Michelle Obama

Chicago

LIVED_IN

Entity embeddings

Encoding relevant properties of the entities, predictive of their relationships.

Relation embeddings

Encoding relevant properties of the relations that help define the set of entity pairs for which the relation holds.

Properties: can capture similarities among entities and relations, can encode relevant information from the graph and achieve high accuracy on KB completion [e.g. Nickel et al. 2011, 2016, Bordes et al. 2011, 2013]

Scoring Functions

Models assign scores to triples (candidate directed labeled links in KB):

$$s, t \in E, r \in R_{kb}$$

$$T = (s, r, t)$$

Scores

$$f(s, r, t|\Theta)$$

Θ : Embeddings of entities and relations

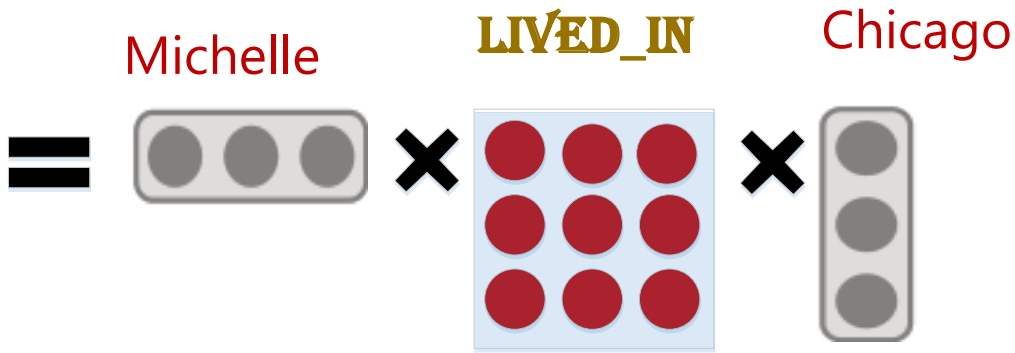
Used to predict the existence of triples:

$$y_T \in \{0,1\}$$

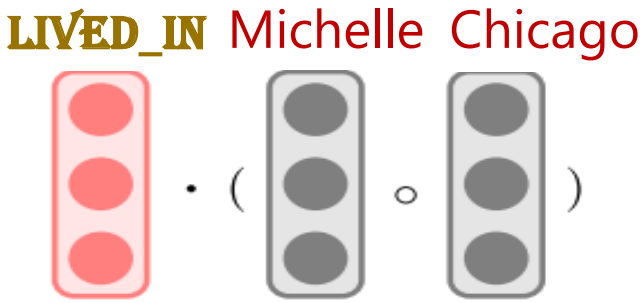
Scoring Functions

Bilinear Model [Nickel et al. 2011]

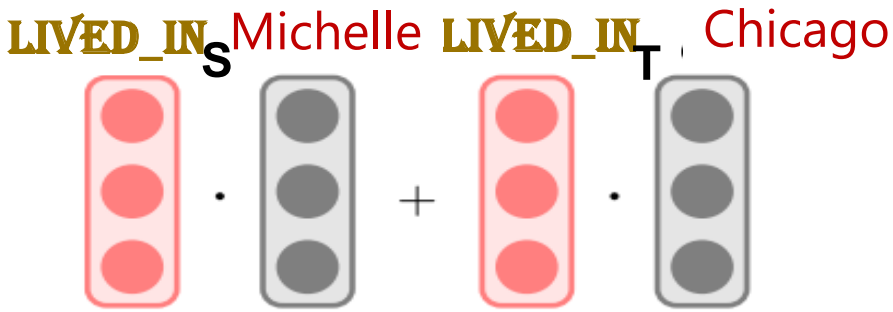
f(Michelle Obama, lived_in, Chicago)



Bilinear-diag Model [Yang et al. 2015]



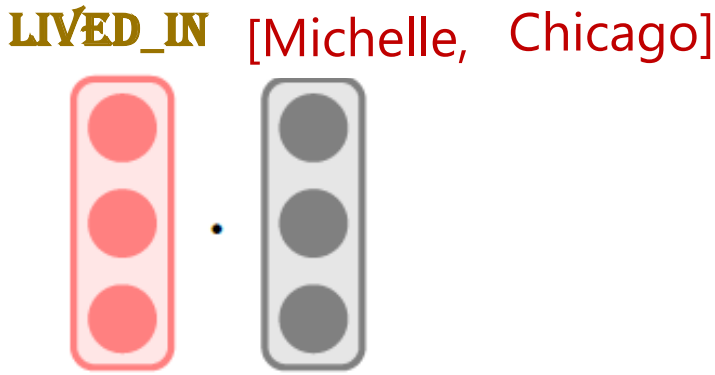
Model E [Riedel at al. 2013]



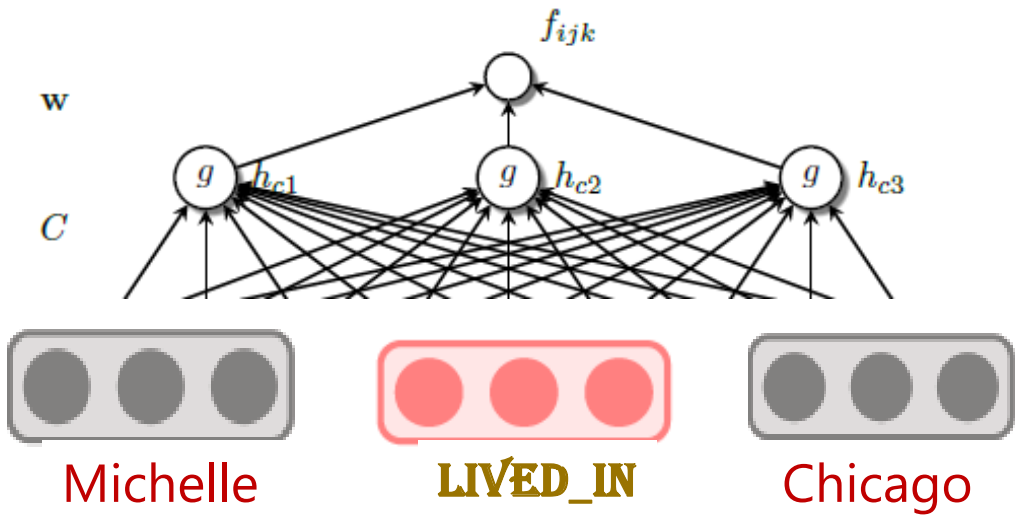
Scoring Functions

Model F [Riedel et al. 2013]

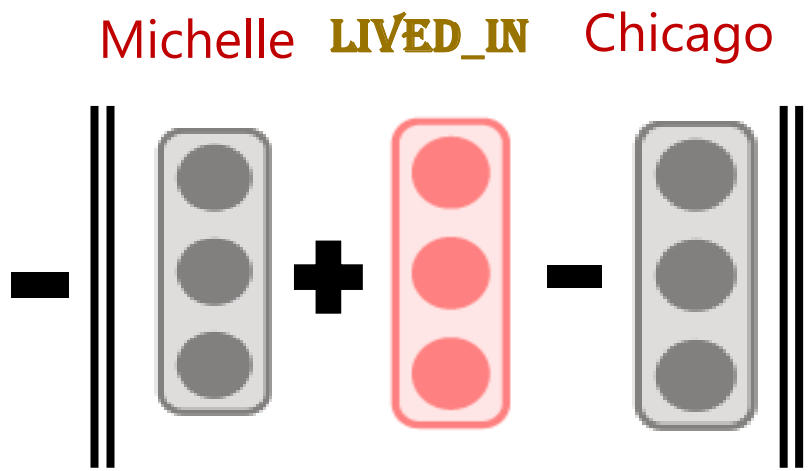
f(Michelle Obama, lived_in, Chicago) =



ER-MLP [Dong et al. 2014]



TransE [Bordes et al. 2013]



Loss functions for training model parameters

Learning $\boldsymbol{\theta}$: maximize conditional probability of correct answer for training queries $(s, r, ?)$ and $(?, r, o)$ e.g. (Barack Obama, nationality, ?)

Loss function in our prior work:

$$\mathrm{P}(t|s,r) = \frac{e^{f(s,r,t|\theta)}}{\sum_{t' \in Neg(s,r,?) \cup t} e^{f(s,r,t'|\theta)}}$$

$$L(\theta) = \lambda||\theta||^2 - \left(\sum_i \log P(t_i|s_i, r_i) + \log P(s_i|r_i, t_i)\right)$$

Loss functions for training model parameters

Learning θ: minimize a margin-based loss-function: the score for observed training triples $(s, r, t) = x^+$ should be higher than the score of negative triples $(s', r', t') = x^-$

Pair-wise margin loss:

$$\min_{\Theta} \sum_{x^+ \in \mathcal{D}^+} \sum_{x^- \in \mathcal{D}^-} \mathcal{L}(f(x^+; \Theta), f(x^-; \Theta)) + \lambda \operatorname{reg}(\Theta)$$

$$\mathcal{L}(f, f') = \max(1 + f' - f, 0).$$

Other losses: survey [Nickel et al. 2016] tutorial [Bouchard et al. 2015]

Overview of Part 5

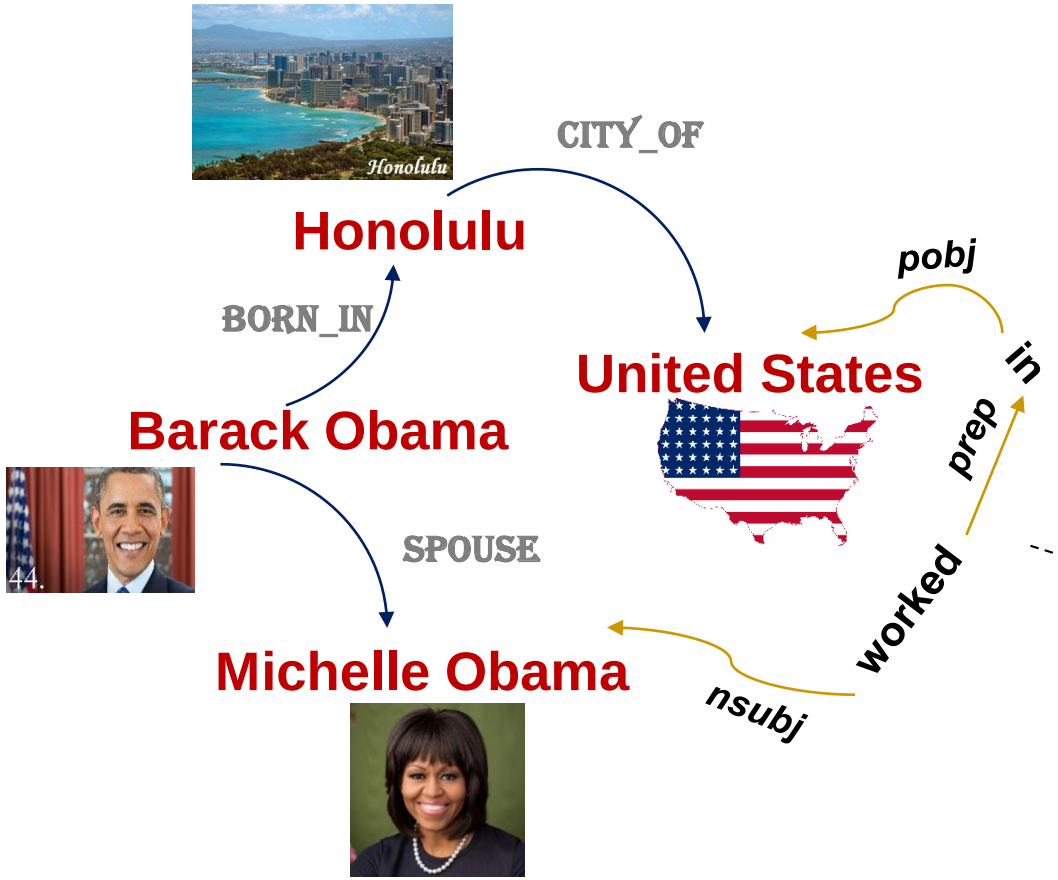
Reasoning with embeddings of entities and relations

- Representing texts

Reasoning with relation paths (PRA)

A hybrid method embedding triples, text, and relation paths

Knowledge Bases Augmented with Textual Relations



[Lao et al. 2012] [Riedel et al. 2013]

Facts stated in text often directly or indirectly support knowledge base facts.

Can treat textual mentions as another type of relations.

Michelle Obama worked in the United States.

Models for graphs including text

$\text{SUBJECT} \xrightarrow{\text{dep}} \text{co-founder} \xrightarrow{\text{prep}} \text{of} \xrightarrow{\text{pobj}} \text{OBJECT}$

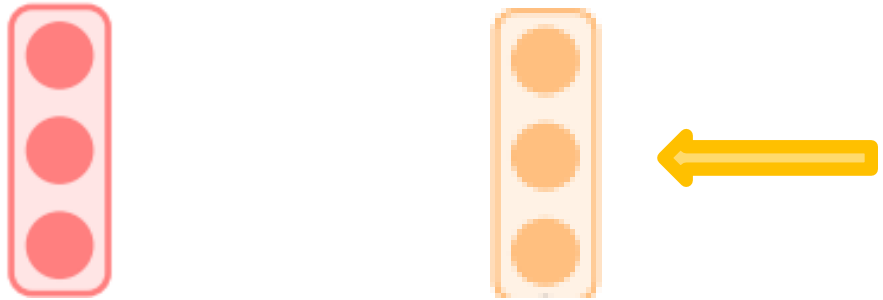
Basic

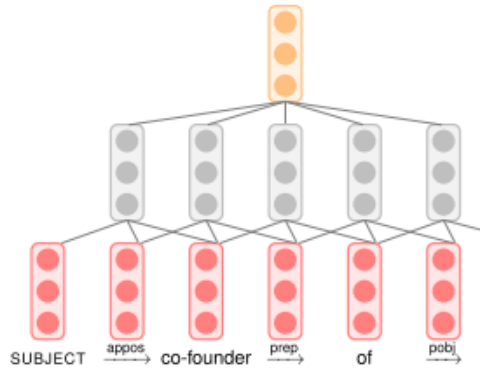
KB relations

Textual relations

[Toutanova et al. 2015]

Conv





KB relations

Textual relations

Bi-LSTM and cross-lingual [Verga et al. 2016]

Overview of Part 5

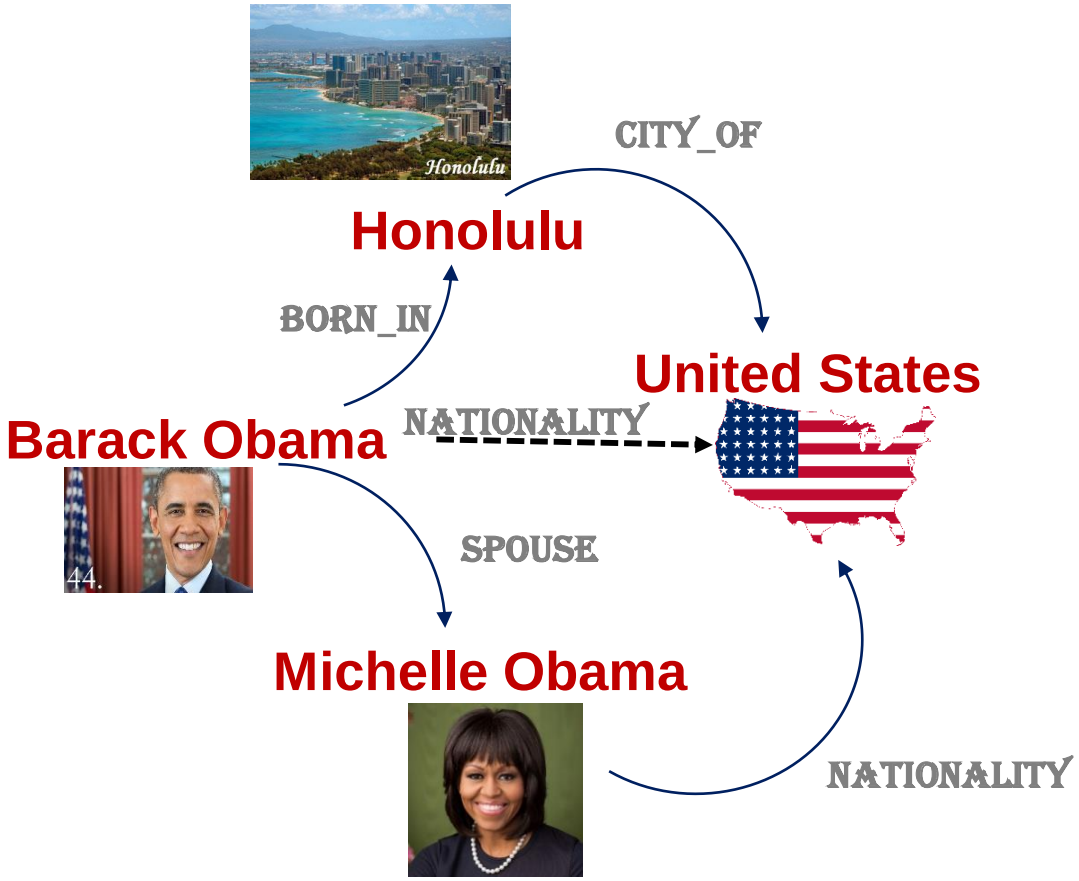
Reasoning with embeddings of entities and relations

- Representing texts

Reasoning with relation paths (PRA)

A hybrid method embedding triples, text, and relation paths

Path Ranking Algorithm [Lao et al. 11]



To score (*s*, *r*, *t*), collect the path types of paths connecting s and t

$\boldsymbol{\pi_1}$: BORN_IN CITY_OF $p = 1$

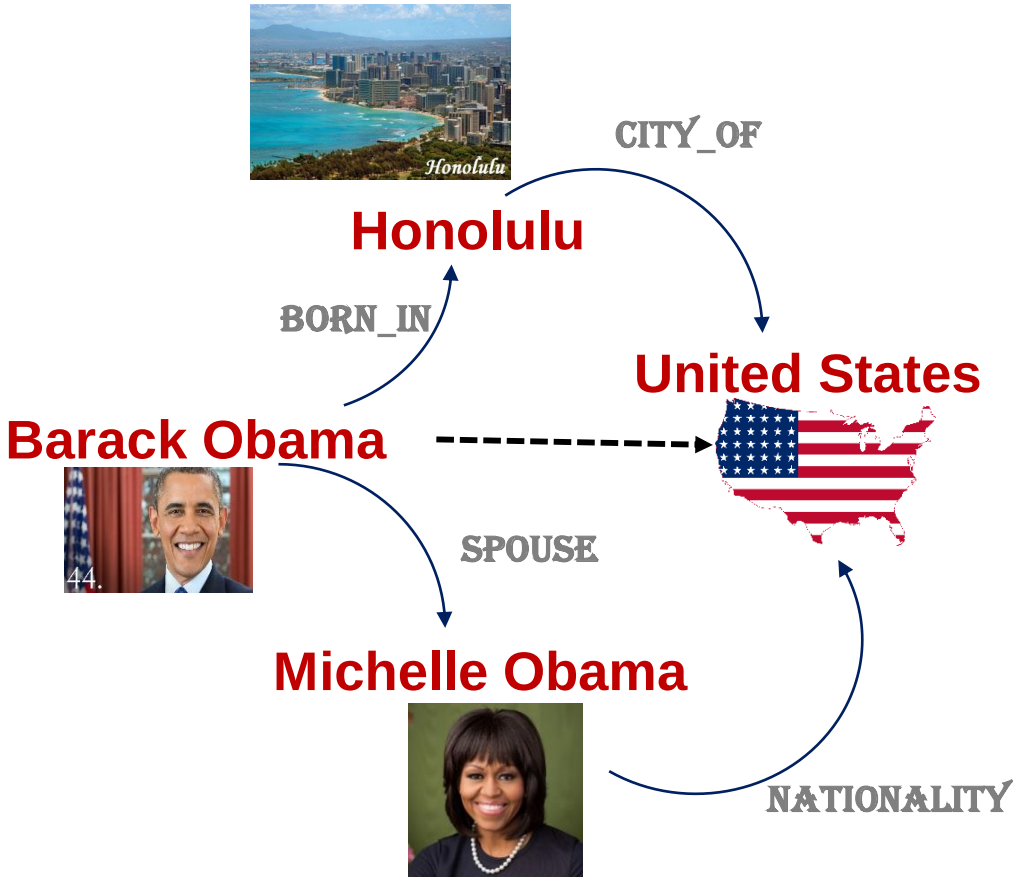
$\boldsymbol{\pi_2}$: SPOUSE NATIONALITY $p = 1$

Each path type is a feature with value the path-constrained random walk probability.

Scoring function: linear in the given feature values

$$f = \boldsymbol{w_1} \times 1 + \boldsymbol{w_2} \times 1$$

Path Ranking Algorithm [Lao et al. 11]



Computationally expensive and data-sparse if many relation types and long paths allowed

For 3000 relation types:

L=1	*L=2*	*L=3*	*L=4*
3000	**9 million**	**27 billion**	**81 trillion**

Grows exponentially as $|R|^L$
$|R|$ increases when textual links are considered.

Approach: pruning or sampling of path types, other approximation.

Overview of Part 5

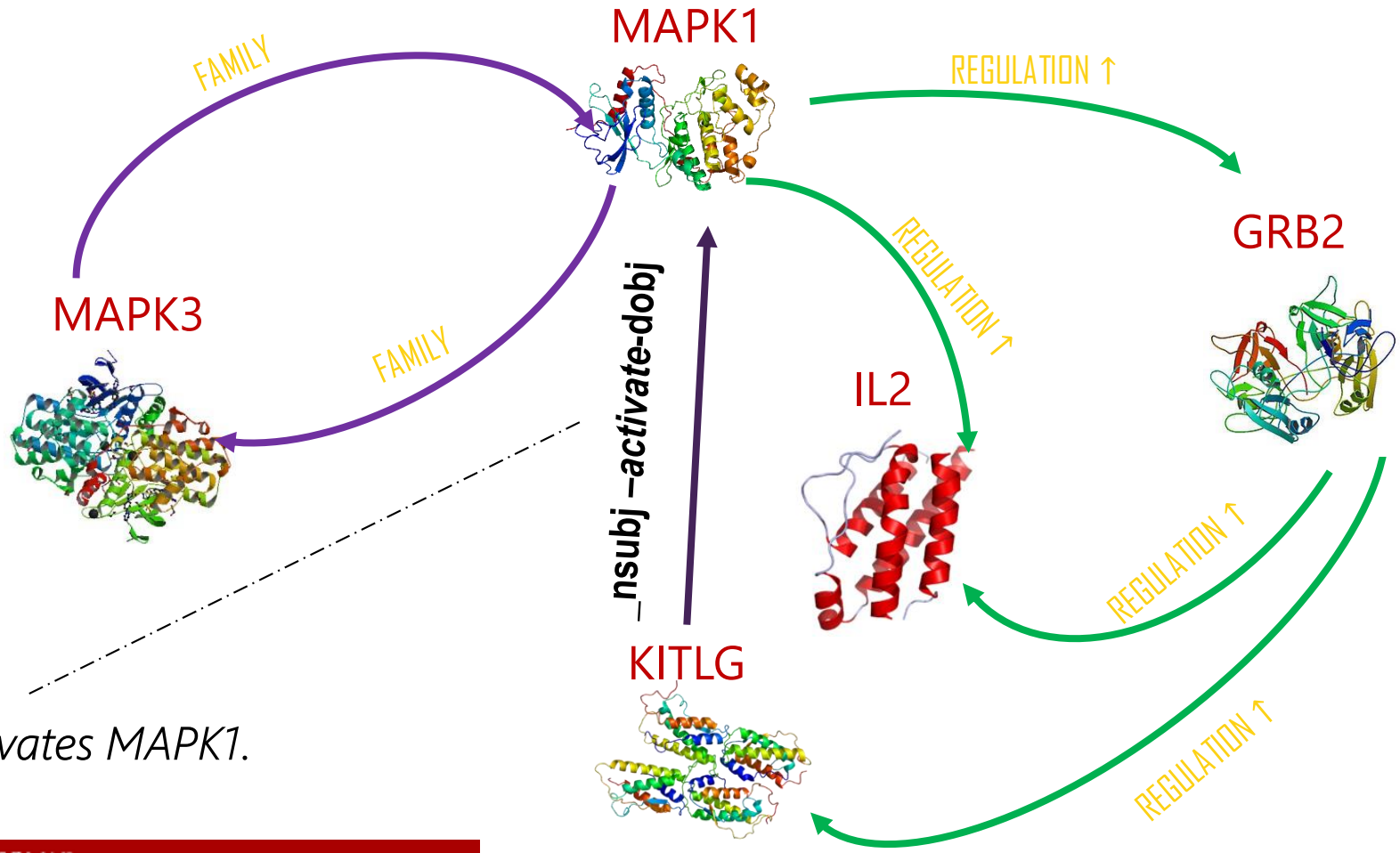
Reasoning with embeddings of entities and relations

- Representing texts

Reasoning with relation paths (PRA)

A hybrid method embedding triples, text, and relation paths

Network with KB relations and text



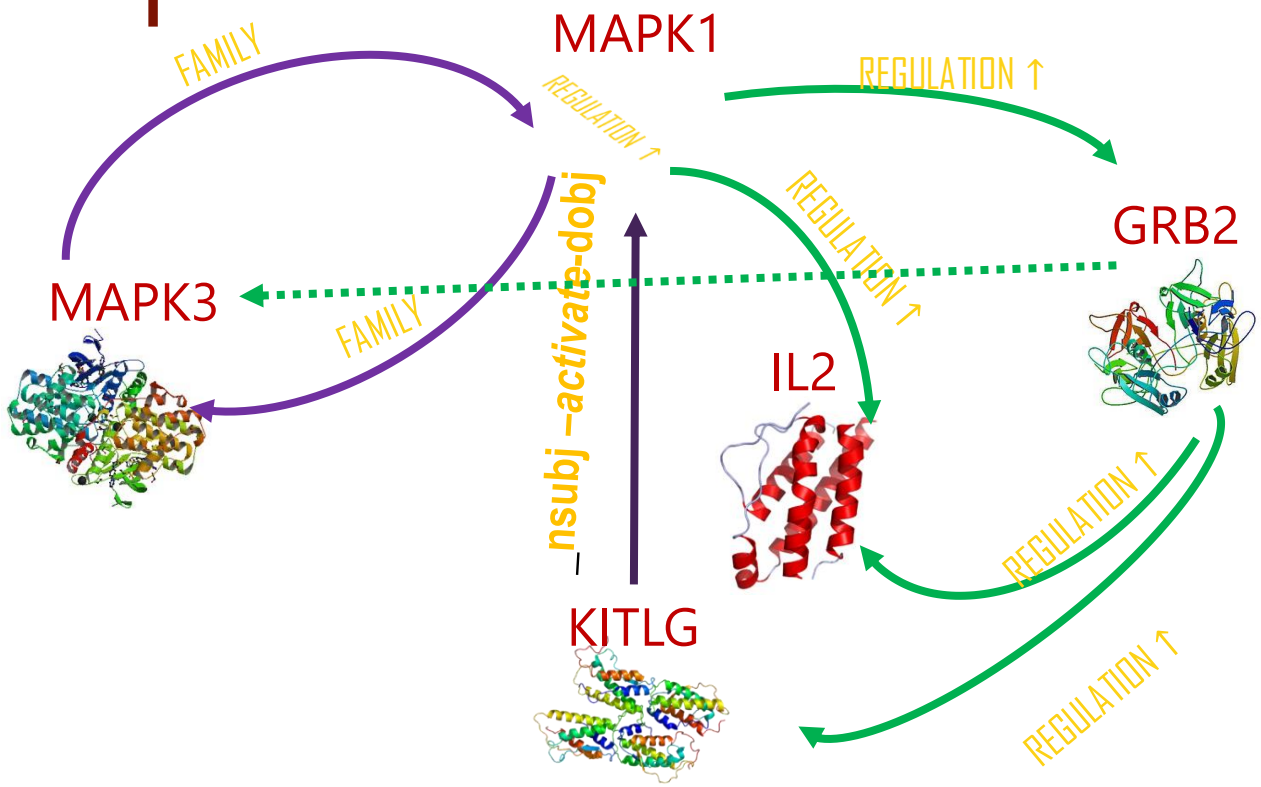
PubMed

KITLG activates MAPK1.

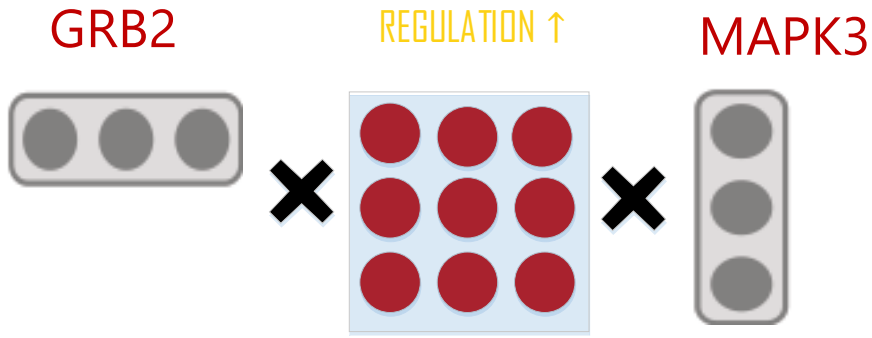
NATIONAL CANCER INSTITUTE nature PathwayInteractionDatabase

NCI-PID-PubMed Genomics Knowledge Base Completion Dataset
http://aka.ms/NCI-PID-PubMed

Reasoning with embeddings and relation paths



Triple-based Embedding Model



Paths from GRB2 to MAPK3

π_1: REGULATION↑ IL2 _REGULATION↑ MAPK1 FAMILY

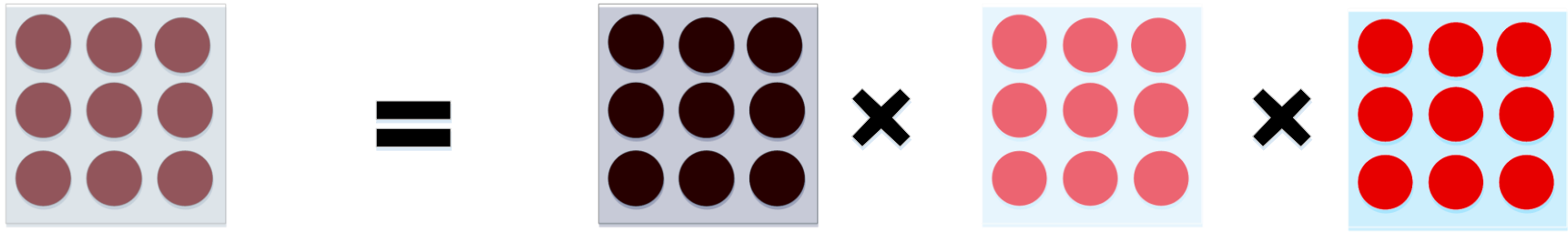
π_2: REGULATION↑ KITLG _nsubj-activate-dobj MAPK1 FAMILY

π_3: _REGULATION↑ MAPK1 FAMILY

Problems when using relation paths: sparsity → compositional representations

π_1: REGULATION↑ _nsubj-activate-dobj FAMILY

Compositional representations of path types: vector or matrix compositional embeddings $\Phi(\pi)$.



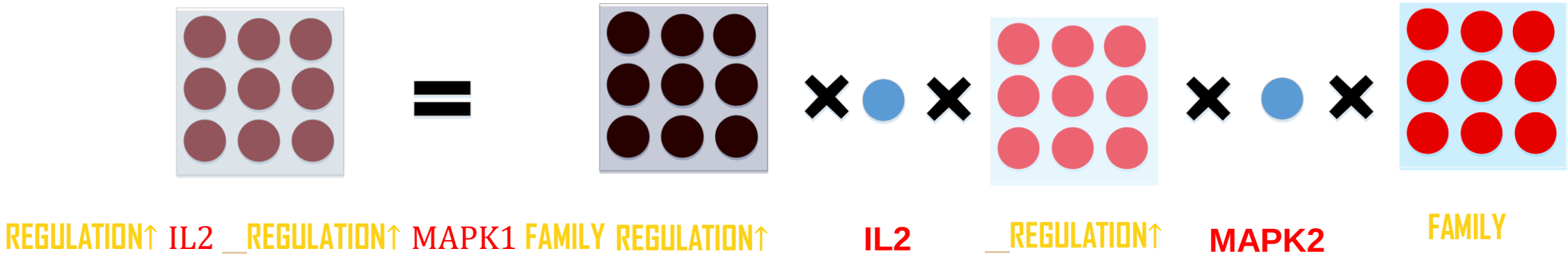
REGULATION↑ _nsubj-activate-dobj FAMILY REGULATION↑ _nsubj-activate-dobj FAMILY

[Guu et al. 2015]

Also: RNN [Neelakantan et al. 2015], or sum of vectors [Lin et al. 2015]

See [Gardner et al. 2013, 2014] for different methods to combat sparsity.

Compositional representations of paths including nodes



What nodes does a path pass through?

Compositional representations enable path representations to depend on intermediate nodes.

- In a first implementation, a scalar weight for each node [Toutanova et al. 2016]
- [Das et al. 2016] also shows gains from intermediate nodes as vectors.

We can derive even more power from compositional representations!

[Toutanova, Lin, Yih, Poon, Quirk, 16]

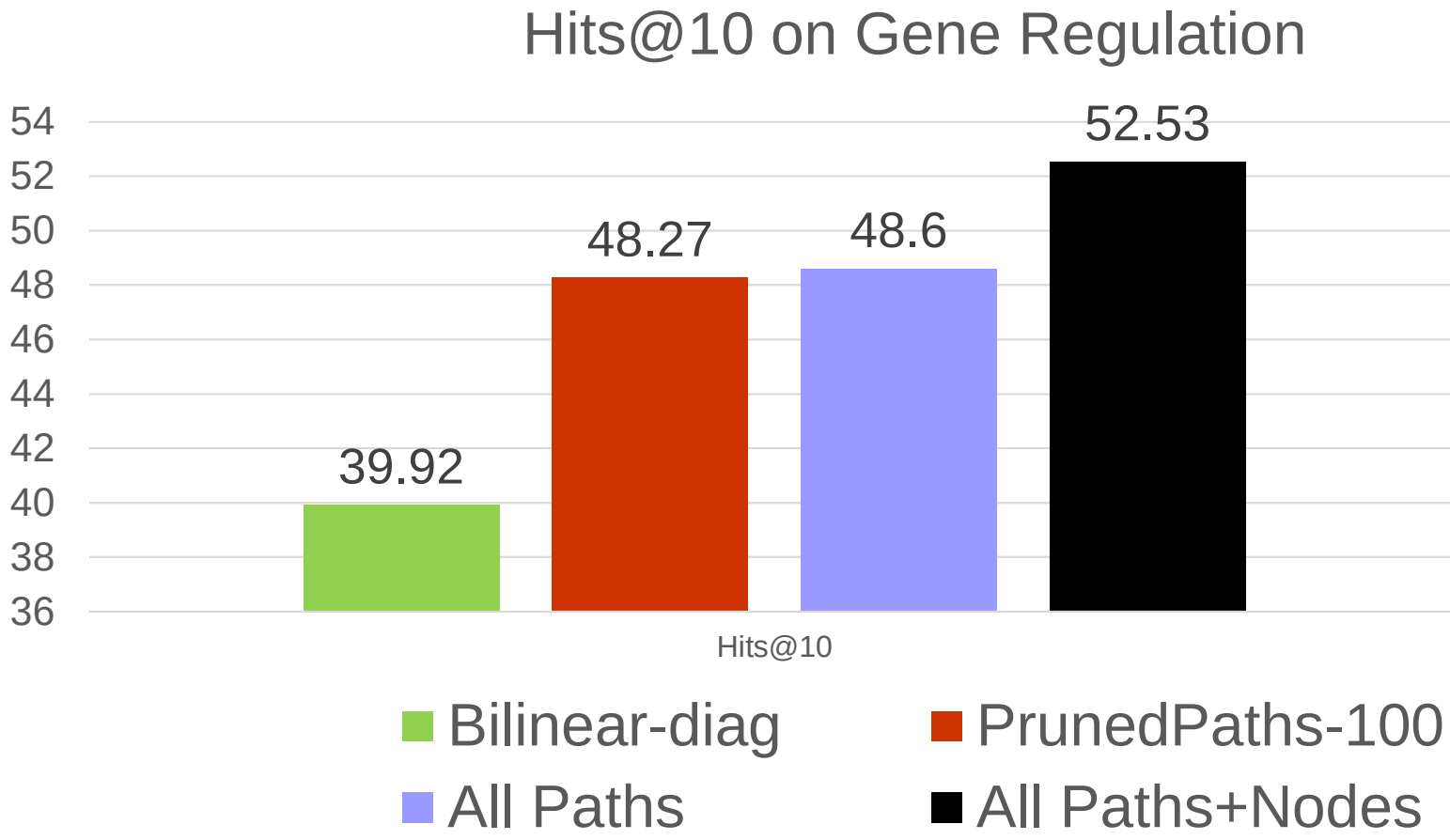
The bilinear compositional model of paths permits *exact inference* with all relation paths of bounded length, using dynamic programming.

Polynomial in graph size and maximum path length

This model also allows finer-grained modeling of relation paths by distinguishing paths according to their specific intermediate nodes.

No increase in asymptotic complexity

Results: using compositional representations of relation paths from KB and text relations



NCI-PID database + textual mentions from Pubmed

d=100, L=3 (no gain from longer paths)

Other Applications of Embeddings of Networks

In neural network models pre-trained embeddings of inputs can often provide strong improvements

Can train network embedding models to encode network knowledge

- Gene embeddings
- Relation embeddings
- Textual mention embeddings

Part 6: Applications to Precision Medicine

Knowledge curation for tumor board

Personalize cancer drug combinations

Disease modeling from electronic medical records

NLP for open science

OncoKB Team

OncoKB is developed and maintained by the Knowledge Systems group in the Marie Josée and Henry R. Kravis Center for Molecular Oncology at Memorial Sloan Kettering Cancer Center.

Design & Development

Debyani Chakravarty, PhD
Jianjiong Gao, PhD
Sarah Phillips, PhD
Hongxin Zhang, MSc
Ritika Kundra, MSc
Jiaojiao Wang, MSc
Ederlinda Paraiso, MPA
Julia Rudolph, MPA
David Solit, MD
Paul Sabbatini, MD
Nikolaus Schultz, PhD

Clinical Genomics Annotation Committee

Shrujal Baxi, MD, MPH
Margaret Callahan, MD, PhD
Sarat Chandarlapaty, MD, PhD
Alexandra Charen-Snyder, MD
Ping Chi, MD, PhD
Daniel Danila, MD
Mrinal Gounder, MD
James Harding, MD
Matthew Hellman, MD
Alan Ho, MD, PhD
Gopa Iyer, MD
Yelena Janjigian, MD
Thomas Kaley, MD
Maeve Lowery, MD
Antonio Omuro, MD
Paul Paik, MD
Michael Postow, MD
Dana Rathkopf, MD
Alexander Shoushtari, MD
Neerav Shukla, MD
Tiffany Traina, MD
Martin Voss, MD
Rona Yaeger, MD

Core Curators

Moriah Nissan, PhD
Lindsay Saunders, PhD
Tara Soumerai, MD
Fiona Brown, PhD
Tripti Shrestha Bhattarai, PhD
Kinisha Gala, BSc
Aphrothiti Hanrahan, PhD
Anton Henssen, MD
Phillip Jonsson, PhD
Iñigo Landa-Lopez, PhD
Eneda Toska, PhD

Quest Diagnostics

Feras M Abu Hantash, PhD
Andrew Grupe, PhD
Matthew Beer, BSc

Knowledge Curation for Tumor Board

Everyday: 4000 new papers

Manual: GDKD, CIVIC, OncoKB, …

Wanted: Machine reading assisted curation

PROJECT HANOVER

The Jackson Laboratory

Leading the search for tomorrow's cures

Hanover Gene search Drugs JAX

- [x] Clinical
- [ ] Preclinical Patient Derived Xenograft
- [ ] Preclinical Patient Derived Cell Culture
- [ ] Preclinical Cell Line Xenograft
- [ ] Preclinical Cell Line Culture
- [ ] Unknown

Drugs (59)

Filter

bortezomib
bosutinib
cabozantinib
capecitabine
carboplatin
cetuximab
chlorambucil
ci-1033
cisplatin
colchicine
conjugated estrogens
crizotinib
dasatinib
docetaxel
erlotinib
ethanol

cetuximab

Genes (7)

BRAF
EGF
EGFR
ERBB2
KRAS
YWHAB
ZFP36

Variant **PubMed ID** **Level of evidence**

Then there is a plan for a test of the new Braf inhibitor Vemurafenib, shown to be effective in melanoma patients whose tumours display a mutation in BRAF V600E , but in colorectal patients instead of melanoma. It seems that bowel tumours treated with an inhibitor of this mutated gene switch on EGFR which is the target for a number of agents including cetuximab , so a combination of the two agents is logical to trial.

V600E 19738166 Clinical

Disease type: Colorectal Neoplasms

Di Nicolantonio et al. () also demonstrated that introduction of the BRAF V600E allele could confer resistance to either cetuximab or panitumumab in wild-type BRAF colorectal cancer cells.

V600E 20972475 Clinical

Disease type: Unknown

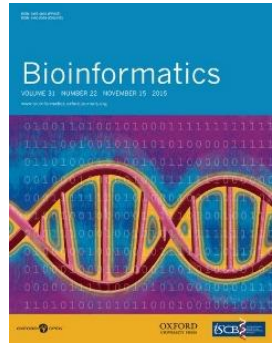
Also available is a second assay, the BRAF (V600E Sequencing) (V6S), which uses sequencing to detect the BRAF p.Val600Glu sequence variant. Public Health Importance Available evidence indicates that the clinical benefit from treatment with

Personalize Cancer Drug Combos

Kurtz et al. “Identifying Combinations of Targeted Agents for Hematologic Malignancies”. *PNAS, to appear*.

PNAS

Fried et al. “Learning to Prioritize Cancer Drug Combinations”. *In preparation*.



Drug Combination

Problem: What combos to try?

- Cancer drug: 250+ approved, 1200+ developing
- Pairwise: 719,400; three-way: 287,280,400

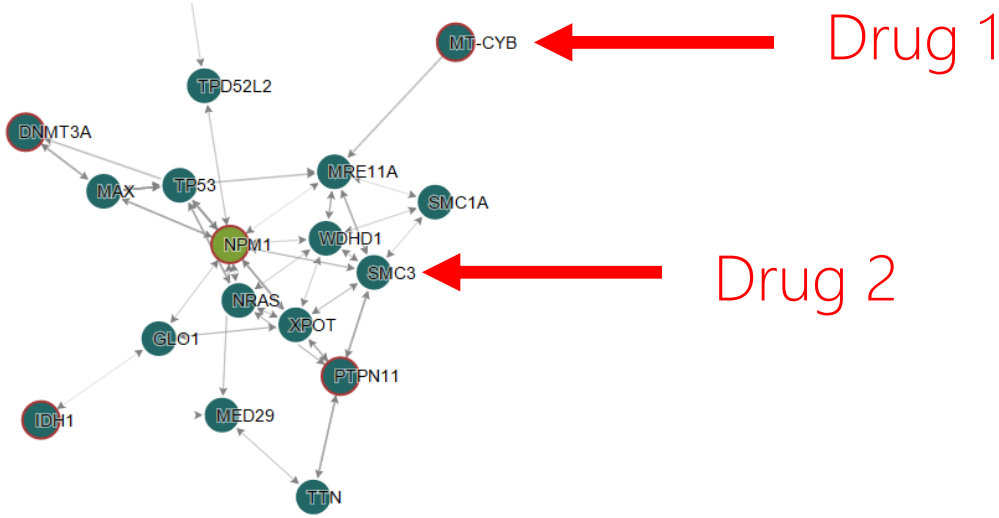
Wanted: Prioritize drug combos

Drug Combination

Problem: What combos to try?

- Cancer drug: 250+ approved, 1200+ developing
- Pairwise: 719,400; three-way: 287,280,400

Wanted: Prioritize drug combos



Drug 1

Drug 2

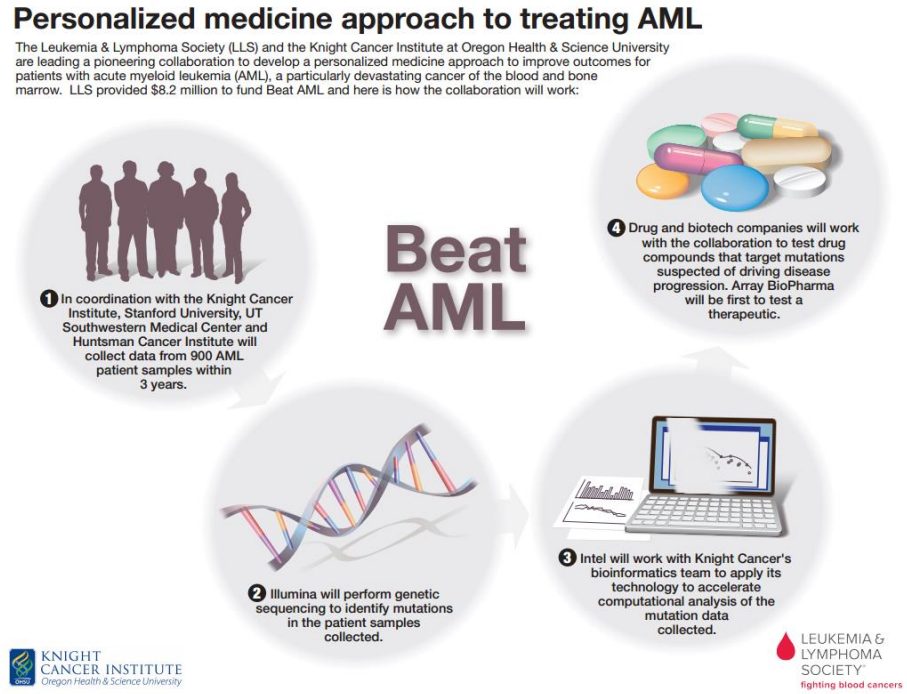
Personalize Drug Combos

Targeted drugs: 149

Pairs: 11,026

Tested: 102 (in two years)

Unknown: 10,924



Machine Learning

Patient: Transcriptome (RNA expression level)

Drug: Gene targets

Machine-read gene network → key features

Personalized medicine approach to treating AML

The Leukemia & Lymphoma Society (LLS) and the Knight Cancer Institute at Oregon Health & Science University are leading a pioneering collaboration to develop a personalized medicine approach to improve outcomes for patients with acute myeloid leukemia (AML), a particularly devastating cancer of the blood and bone marrow. LLS provided $8.2 million to fund Beat AML and here is how the collaboration will work:

1 In coordination with the Knight Cancer Institute, Stanford University, UT Southwestern Medical Center and Huntsman Cancer Institute will collect data from 900 AML patient samples within 3 years.

Beat AML

4 Drug and biotech companies will work with the collaboration to test drug compounds that target mutations suspected of driving disease progression. Array will be first t therapeu

2 Illumina will perform genetic sequencing to identify mutations in the patient samples collected.

3 Intel will work with Knight Cancer's bioinformatics team to apply its technology to accelerate computational analysis of the mutation data collected.

Ongoing:
Cell line experiments on Hanover predictions

KNIGHT CANCER INSTITUTE
Oregon Health & Science University

LEUKEMIA & LYMPHOMA SOCIETY
fighting blood cancers

Modeling Disease Progression

Wanted: Predict onset, complication, treatment

Electronic medical records (EMRs)

Clinical notes contains rich patient information

Modeling Disease Progression

MIMIC

```
1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date:  [**
7**]

Date of Birth

Service: SURG

Allergies:
Patient recor

Attending:[**
Chief Complai
headache and

Major Surgica
central line
```

```
History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike. She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16.  She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV.  The patient was evaluated by Neuro in the
ED.
```

Example: Classifying Breast Diseases

Breast pathology report; 20 categories (e.g., atypia)

Supervised learning; n-gram features

On par w/ rule-based accuracy (>90%)

Follow-up: Category transfer learning

Yala et al. "Using machine learning to parse breast pathology reports". *Breast Cancer Research and Treatment, 2017.*

Example: Classifying Heart Failure

Hospitalization: Did heart failure occur?

Supervised learning

Structured + Clinical notes → Best accuracy

Blecker et al. "Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data". *JAMA Cardiology, 2016*.

Example: Learning Patient Embedding

Representation learning: Denoising autoencoder

Evaluation: Predict new disease onset

Outperformed standard dimension reduction

NLP: Negation, family history, entity linking

Miotto et al. "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". *Scientific Reports, 2016*.

NLP for Open Science

Explosive growth in public data

Discovery hindered by lack of access & annotation

WideOpen: “Make public data public”

EZLearn: Extreme zero-shot learning

Big Data for Precision Medicine

NCBI Resources How To Sign in to NCBI

GEO Home | Documentation | Query & Browse | Email GEO

Gene Expression Omnibus

GEO Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession | Search

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

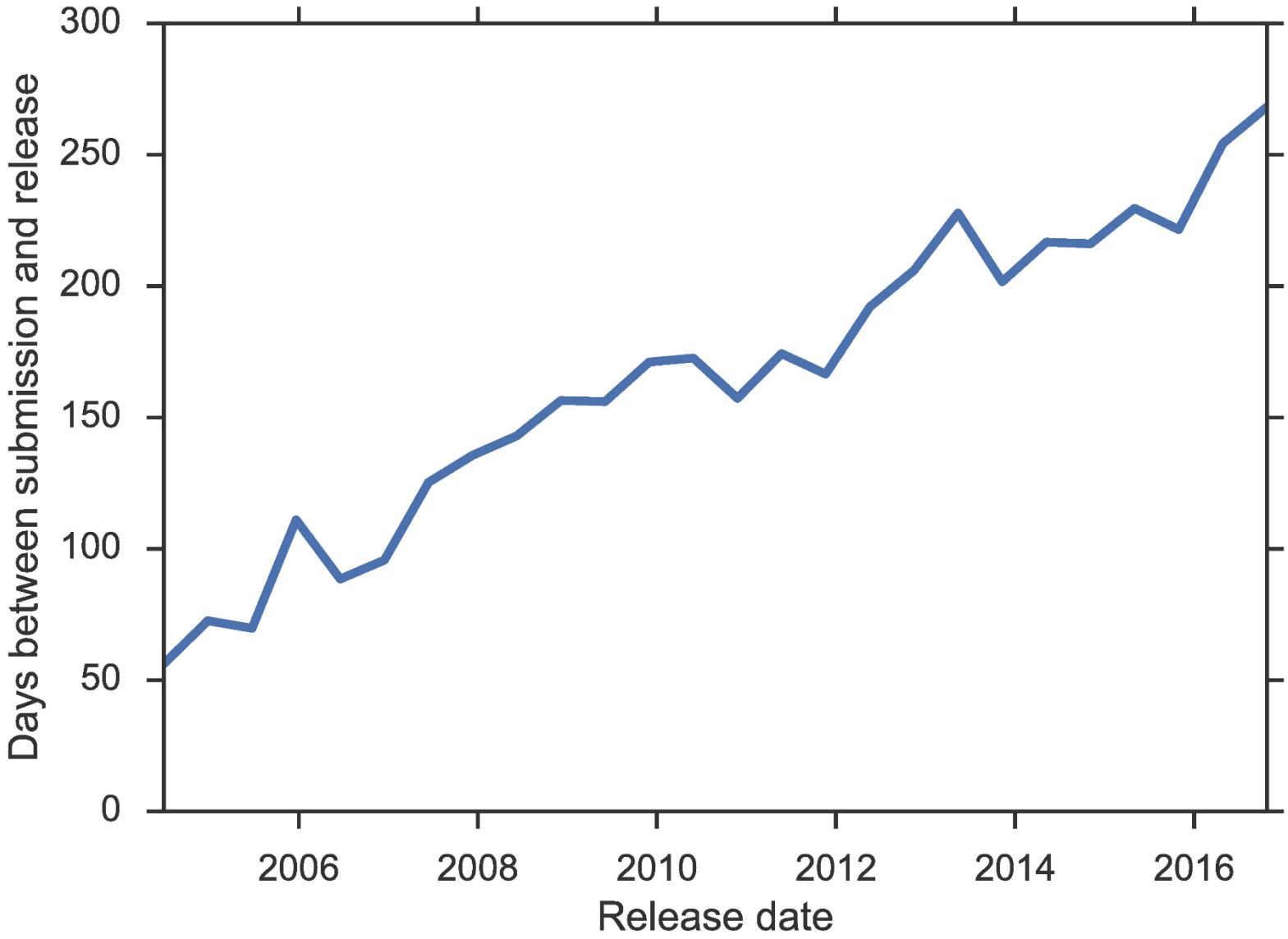
- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- GEO BLAST
- Programmatic Access
- FTP Site

Browse Content

Repository Browser	
DataSets:	4348
Series:	86086
Platforms:	17402
Samples:	2119205

Billions of data points

Public Data Is Not Public



Gene Expression Omnibus

WideOpen: "Make Public Data Public"

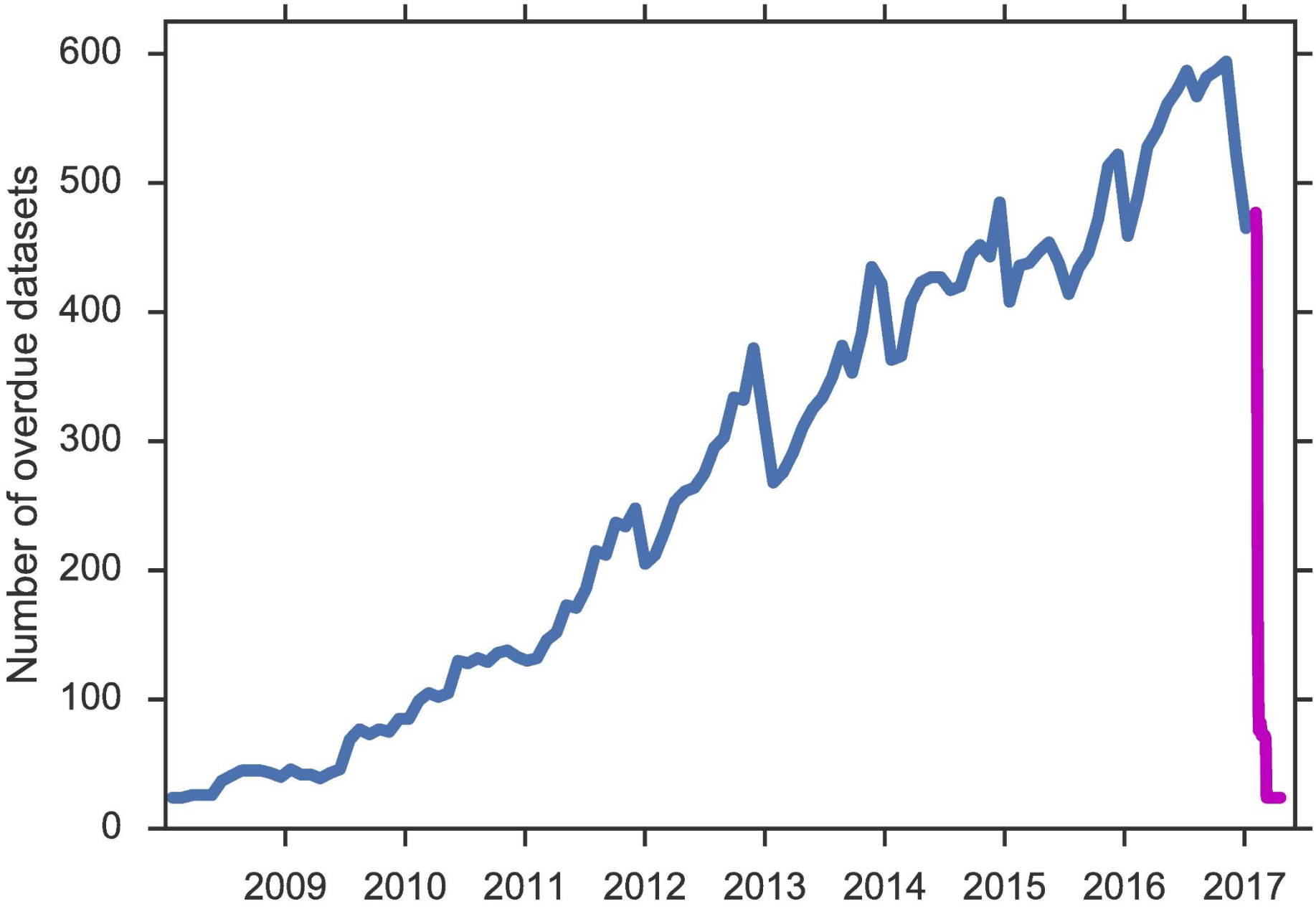
NLP: Automate detection of overdue datasets

PubMed: Identify dataset mentions

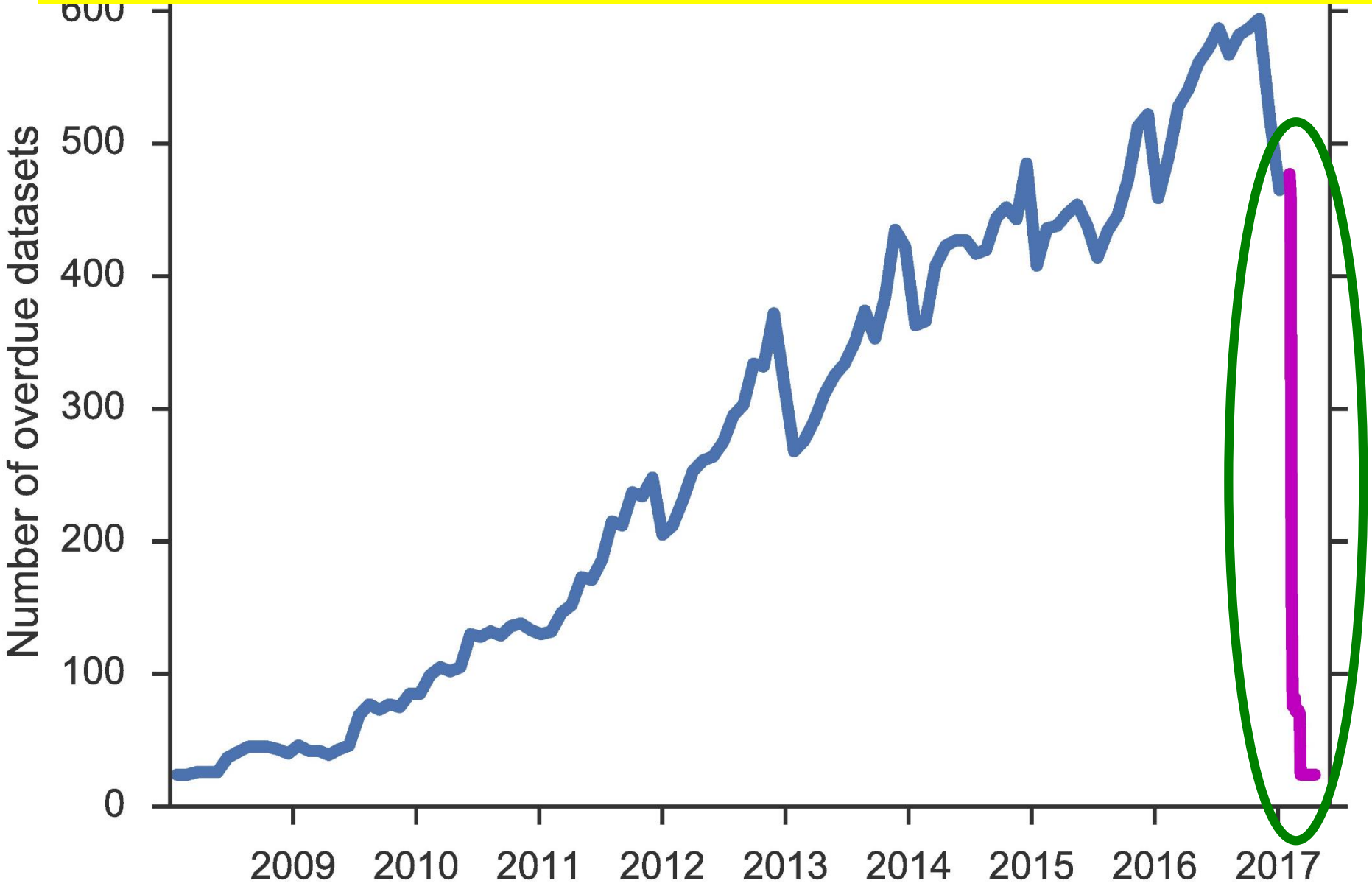
Repo: Parse query output to determine if overdue

Grechkin et al. "Wide-Open: accelerating public data release by automating detection of overdue datasets". *PLOS Biology, 2017*.

Microsoft Research


Number of overdue datasets
600
500
400
300
200
100
0
2009
2010
2011
2012
2013
2014
2015
2016
2017

Enabled GEO to release 400 datasets in a week



WideOpen: “Make Public Data Public”

nature International weekly journal of science

NATURE | NEWS

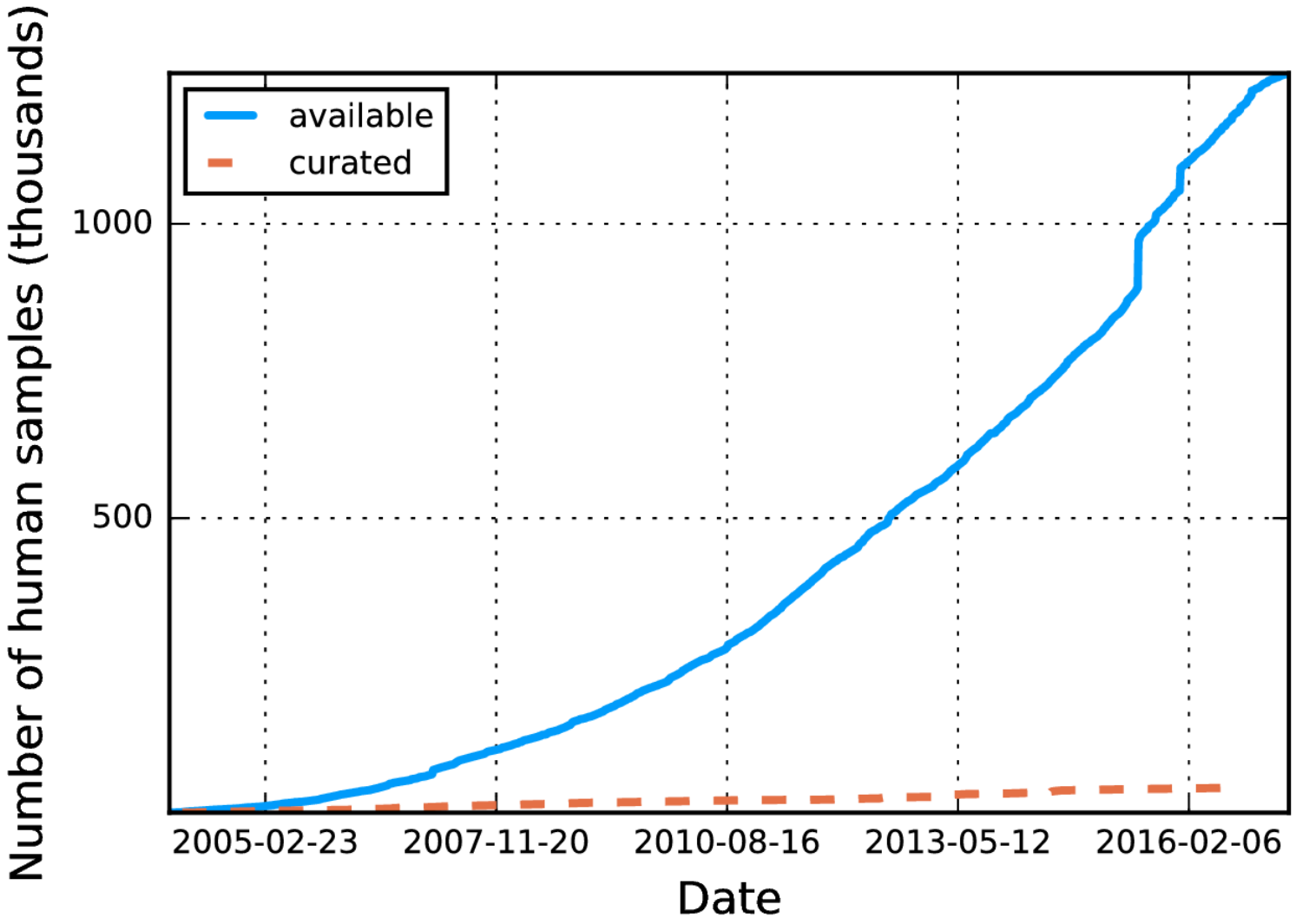
Text-mining tool seeks out ‘hidden data’

Wide-Open checks that the data sets underlying published studies are made freely available.

Dalmeet Singh Chawla

08 June 2017

Public Data Is Not Annotated



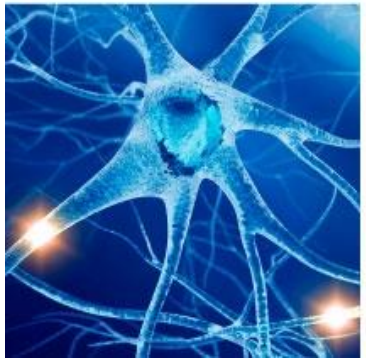
Gene Expression Omnibus

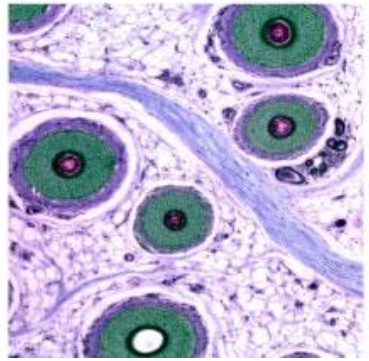
Key Annotation: Cell Type

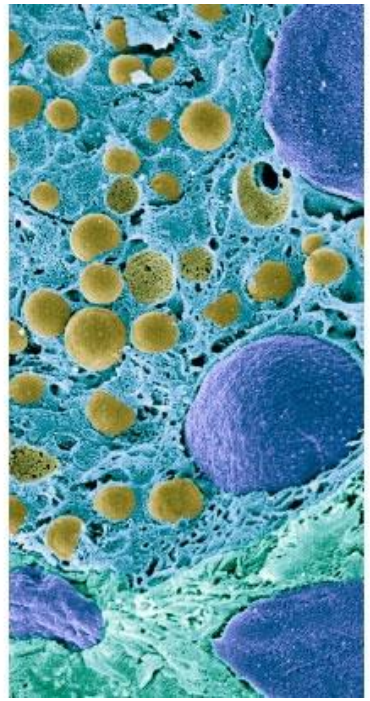
Same DNA, different expression, different functions

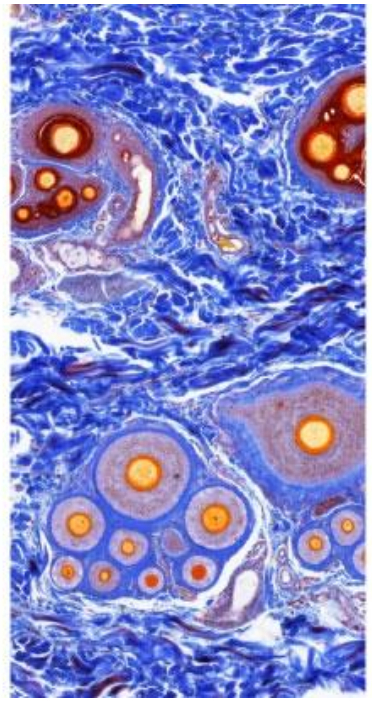
Crucial for understanding development & cancer

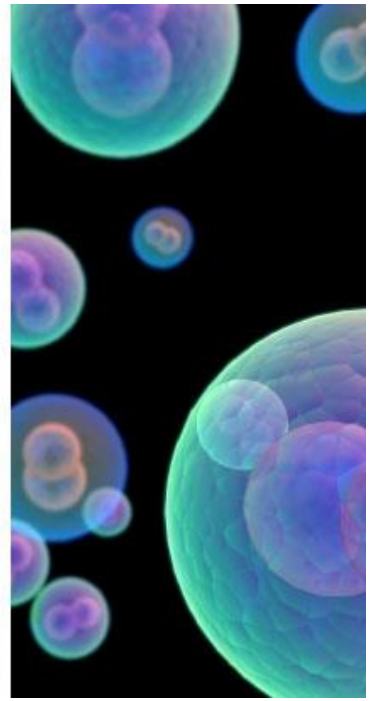
HUMAN CELL ATLAS







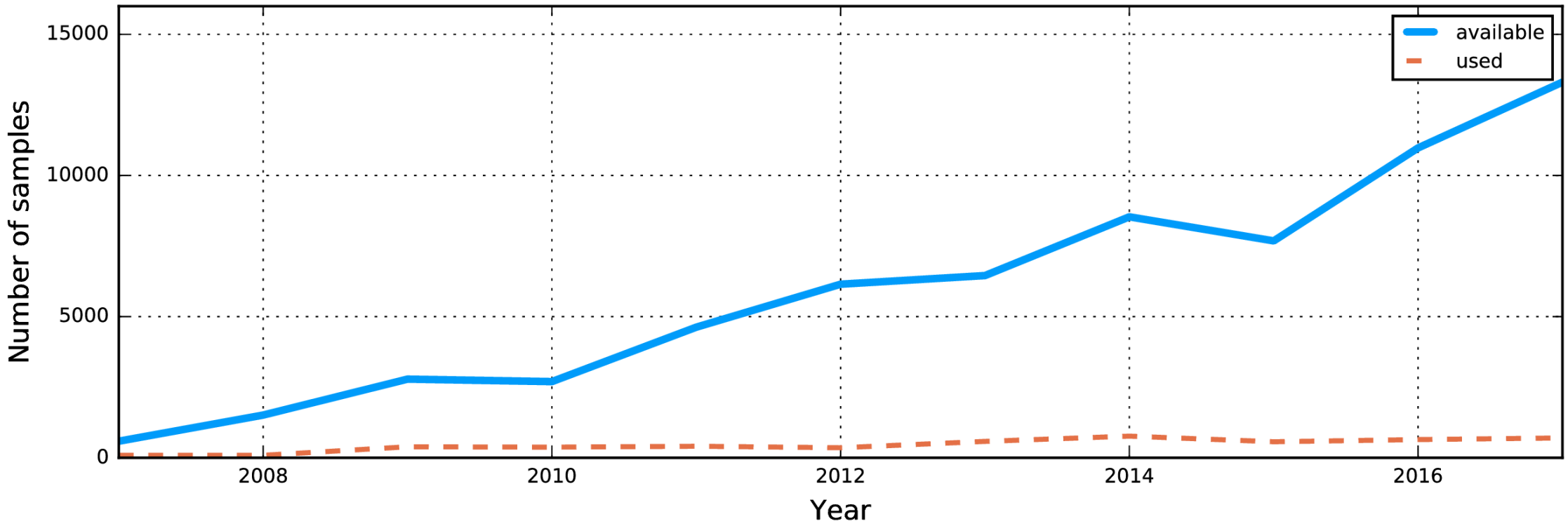




BROAD INSTITUTE

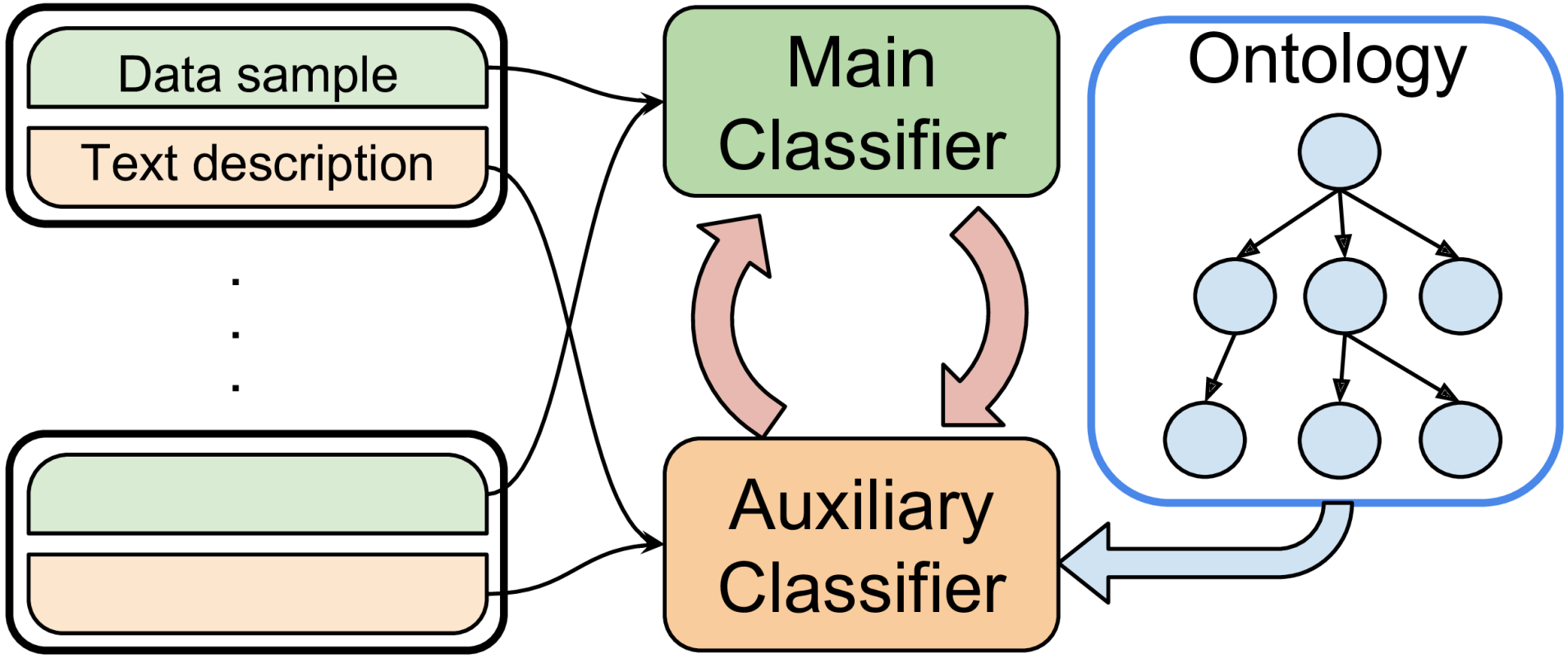
CHAN ZUCKERBERG INITIATIVE

Integrative Studies Remain Small Scale



Gene Expression Omnibus

EZLearn: Extreme Zero-Shot Learning



BRENDA
The Comprehensive Enzyme Information System

4931 types

Grechkin et al. "EZLearn: Extreme Zero-Shot Learning for Unsupervised Data Annotation". *In submission*.

Part 7: Resources

Text

Ontology

Databases

Shared tasks

Project Hanover

Text

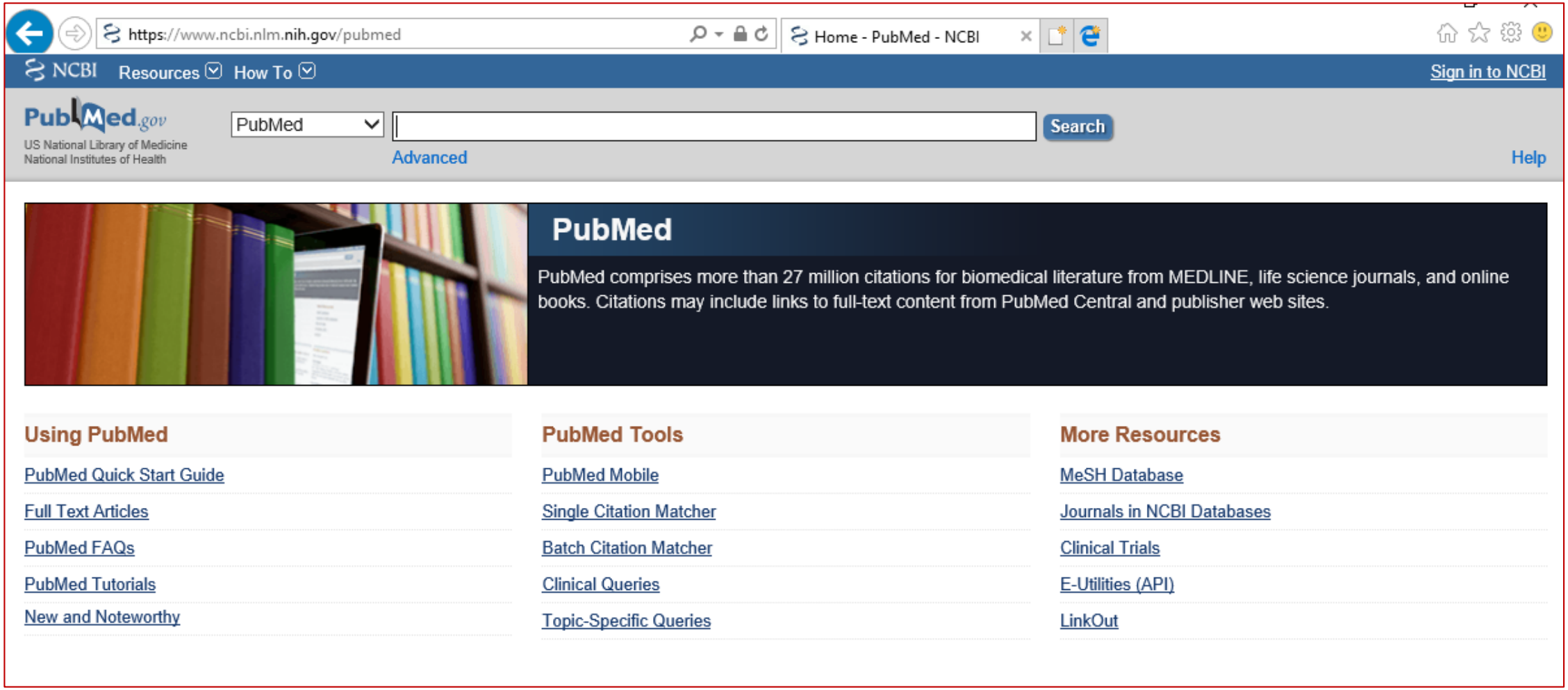
PubMed

Electronic medical record (EMR)

Clinical trial

Pathology report

PubMed


https://www.ncbi.nlm.nih.gov/pubmed
Home - PubMed - NCBI
NCBI
Resources
How To
Sign in to NCBI
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
Search
Advanced
Help
PubMed
PubMed comprises more than 27 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.
Using PubMed
PubMed Quick Start Guide
Full Text Articles
PubMed FAQs
PubMed Tutorials
New and Noteworthy
PubMed Tools
PubMed Mobile
Single Citation Matcher
Batch Citation Matcher
Clinical Queries
Topic-Specific Queries
More Resources
MeSH Database
Journals in NCBI Databases
Clinical Trials
E-Utilities (API)
LinkOut

PubMed

https://www.ncbi.nlm.nih.gov/pubm

NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Using PubMed

- PubMed Quick Start Guide
- Full Text Articles
- PubMed FAQs
- PubMed Tutorials
- New and Noteworthy

https://www.ncbi.nlm.nih.gov/pmc/ | Home - PMC - NCBI

NCBI Resources How To | Sign in to NCBI

PMC
US National Library of Medicine
National Institutes of Health

PMC | Search

Journal List Advanced | Help

PMC

PubMed Central® (PMC) is a free full-text archive of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

PubReader

A whole new way to read scientific literature at PubMed Central

Get Started

- PMC Overview
- Users' Guide
- Journal List
- PMC FAQs
- PMC Copyright Notice

Participate

- Information for Publishers
- Add a Journal to PMC
- Participation Agreements
- File Submission Specifications
- File Validation Tools

Keep Up to Date

- New in PMC | RSS
- PMC Announce Mail List
- Utilities Announce Mail List
- Tagging Guidelines Mail List

Other Resources

- PMC International
- Text Mining Collections
- Developer Resources
- NLM LitArch
- PMC Citation Search
- PMC Accessibility

4.3 MILLION Articles

are archived in PMC.

Content provided in part by:

2011	326	4259
Full Participation Journals	*NIH Portfolio* Journals	*Selective Deposit* Journals

Public Access

- Funders and PMC
- How Papers Get Into PMC
- NIH Manuscript Submission System
- My Bibliography
- PMCID/PMID/NIHMSID Converter

PubMed

Abstracts: 27 millions

Full text: 4.3 millions

Open-access: 1.5 million

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 483 > Issue 7388 > News > Article

NATURE | NEWS

Trouble at the text mine

Computers can rapidly scan through thousands of research papers to make useful connections, but work is being slowed by publishers' unease.

Richard Van Noorden

07 March 2012 | Corrected: 08 March 2012

Electronic Medical Record (EMR)

A.k.a. electronic health record (EHR)

Structured: Billing (ICD), lab test, …

Semi-structured or free text:

- Discharge summary
- Medical history
- Family history
- ……

Electronic Medical Record (EMR)

Collaborative research

MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with ~40,000 critical care patients. It includes demographics, vital signs, laboratory tests, medications, and more.

```
1,23224,174680,2147-12-05,,,"Discharge summary","Report",,"","Admissi
on Date:  [**2823-9-29**]              Discharge Date:   [**2823-10-1
7**]

Date of Birth:  [**2768-10-11**]             Sex:   F

Service: SURGERY

Allergies:
Patient recorded as having No Known Allergies to Drugs

Attending:[**First Name3 (LF) 1**]
Chief Complaint:
headache and neck stiffness

Major Surgical or Invasive Procedure:
central line placed, arterial line placed

History of Present Illness:
54 year old female with recent diagnosis of ulcerative colitis
on 6-mercaptopurine, prednisone 40-60 mg daily, who presents
with a new onset of headache and neck stiffness. The patient is
in distress, rigoring and has aphasia and only limited history
is obtained. She reports that she was awaken 1AM the morning of
[**2823-9-28**] with a headache which she describes as bandlike. She
states that headaches are unusual for her. She denies photo- or
phonophobia. She did have neck stiffness. On arrival to the ED
at 5:33PM, she was afebrile with a temp of 96.5, however she
later spiked with temp to 104.4 (rectal), HR 91, BP 112/54, RR
24, O2 sat 100 %. Head CT was done and relealved attenuation
within the subcortical white matter of the right medial frontal
lobe. LP was performed showing opening pressure 24 cm H2O WBC of
316, Protein 152, glucose 16.  She was given Vancomycin 1 gm IV,
Ceftriaxone 2 gm IV, Acyclovir 800 mg IV, Ambesone 183 IV,
Ampicillin 2 gm IV q 4, Morphine 2-4 mg Q 4-6, Tylenol 1 gm ,
Decadron 10 mg IV.  The patient was evaluated by Neuro in the
ED.
.
Of note, the patient was recently diagnosed with UC and was
started on 6MP and a prednisone taper along with steroid enemas
for UC treatment. She was on Bactrim in past but stopped taking
it for unclear reasons and unclear how long ago.
.


Past Medical History:
chronic back pain, MRI negative
osteopenia - fosamax d/c by PcP
leg pain/parasthesias
h/o hiatal hernia

Social History:
No tob, Etoh. Patient lives alone in a 2 family home w/ a
friend. She is an administrative assistant


Family History:
brother w/ ulcerative proctitis, mother w/ severe arthritis,
father w/ h/o colon polyps and GERD
```

Clinical Trial

ClinicalTrials.gov

Try our beta test site

ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. Learn more about clinical studies and about this site, including relevant history, policies, and laws.

IMPORTANT: Listing of a study on this site does not reflect endorsement by the National Institutes of Health. Talk with a trusted healthcare professional before volunteering for a study. Read more...

Find Studies | About Clinical Studies | Submit Studies | Resources | About This Site

ClinicalTrials.gov currently lists 246,107 studies with locations in all 50 States and in 200 countries.

Text Size

Search for Studies

Example: "Heart attack" AND "Los Angeles"

Search

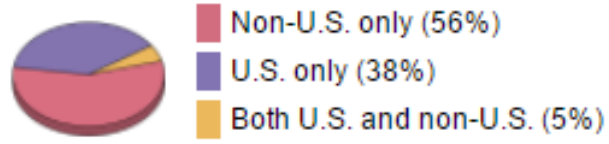
Advanced Search | See Studies by Topic

See Studies on Map

Search Help

- How to search
- How to find results of studies
- How to read a study record

Locations of Recruiting Studies



Total N = 42,836 studies
(Data as of May 31, 2017)

Clinical Trial

Eligibility

Ages Eligible for Study: 18 Years and older (Adult, Senior)
Sexes Eligible for Study: Female
Accepts Healthy Volunteers: No

Criteria

Inclusion Criteria: A subject will be eligible for inclusion in this study only if all of the following criteria are met:

1. Female subjects, age ≥ 18 years at the time informed consent is signed
2. Pathologically confirmed adenocarcinoma of the breast
3. Pathologically confirmed as triple negative, source documented, defined as both of the following
   a. Estrogen Receptor (ER) and Progesterone Receptor (PgR) negative: < 1% of tumor cell nuclei are immunoreactive in the presence of evidence that the sample can express ER or PgR (positive intrinsic controls)
   b. Human Epidermal Growth Factor Receptor 2 (HER2) negative as per American Society of Clinical Oncology - College of American Pathologists (ASCO/CAP) guidelines i. Immunohistochemistry (IHC) 0 or 1 Fluorescence In Situ Hybridization (FISH) negative (or equivalent negative test). Subjects with IHC 2 must have a negative by Fluorescence In Situ Hybridization (FISH),, (or equivalent negative test).
4. Subjects with prior breast cancer history of different phenotypes (ie, ER/PgR/HER2 positive) must have pathologic confirmation of triple negative disease in at least one of the current sites of metastasis
5. Subjects must have received prior adjuvant or neoadjuvant anthracycline therapy; unless (a) anthracycline treatment was not indicated or was not the best treatment option for the subject in the opinion of the treating physician; and (b) anthracycline treatment remains not indicated or, in the opinion of the treating physician, is not the best treatment option for the subject's metastatic disease. a. Newly diagnosed subjects presenting with TNMBC are eligible for the study if anthracycline treatment is not indicated or is not the best treatment option for the subject in the opinion of the treating physician.
6. Subjects with measurable metastatic disease, defined by Response Evaluation Criteria in Solid Tumors 1.1 (RECIST 1.1) guidelines
7. Life expectancy ≥ 16 weeks from randomization
8. No prior cytotoxic chemotherapy for metastatic breast cancer. Prior immunotherapy and/or monoclonal antibody therapy are acceptable. Prior treatments must have been discontinued at least 30 days prior to start of study treatment and all related toxicities must have resolved to Grade 1 or less.
9. Prior neoadjuvant or adjuvant chemotherapy, if given, must have been completed at least 6 months before randomization with all related toxicities resolved, and documented evidence of disease progression per RECIST 1.1 guidelines is required. a. If prior neoadjuvant or adjuvant chemotherapy contained taxane, gemcitabine, or platinum agents, the treatment must have completed at least 12 months before randomization
10. Prior radiotherapy must have completed before randomization, with full recovery from acute radiation side effects. At least one measurable lesion must be completely outside the radiation portal or there must be unequivocal radiologic or clinical exam proof of progressive disease within the radiation portal, in accordance with RECIST 1.1 guidelines
11. At least 30 days from major surgery before randomization, with full recovery
12. Eastern Cooperative Oncology Group (ECOG) performance status 0-1
13. Subject has the following blood counts at screening:
    - Absolute Neutrophil Count (ANC) ≥ 1500/mm^2 ;
    - Platelets ≥ 100,000/mm^2 ;
    - Hemoglobin (Hgb) ≥ 9 g/dL

Ontology

HUGO

MeSH

DrugBank

UMLS

ICD

HGNC

HUGO Gene Nomenclature Committee

Search everything | Search symbols, keywords or IDs

Use * to search with a root symbol (eg ZNF*)

Home | Downloads | Gene Families | Tools | Useful links | About | Newsletters | Contact Us | Help | VGNC | Request Symbol

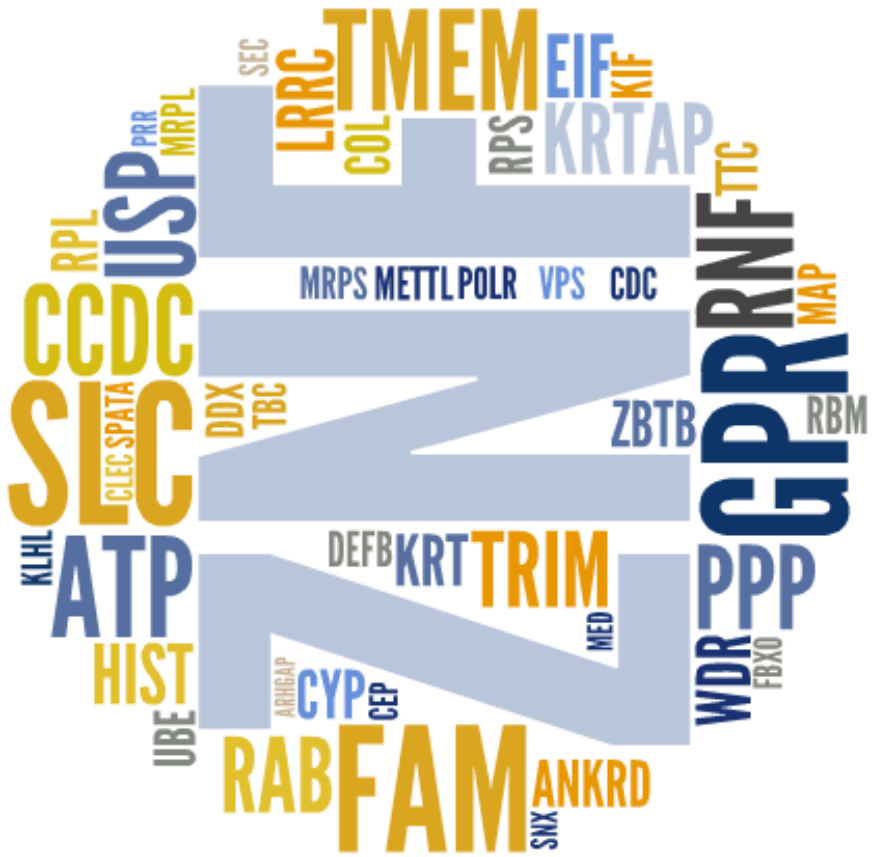
HGNC is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication.

genenames.org is a curated online repository of HGNC-approved gene nomenclature, gene families and associated resources including links to genomic, proteomic and phenotypic information.

Search our catalogue of more than 40,000 symbol reports using our improved search engine (see Search help), search lists of symbols using our Multi-symbol checker and identify possible orthologs using our HCOP tool.

Download our ready-made data files from our Statistics and Downloads page, create your own datasets using either our Custom Downloads tool or BioMart service, or write a script/program utilising our REST service.

Submit your gene symbol and name proposals to us to be accredited with HGNC approved nomenclature for use in publications, databases and presentations.



GeneCardsSuite | GeneCards | MalaCards | LifeMap Discovery | PathCards | TGex | VarElect | GeneAnalytics | GeneALaCart | GenesLikeMe

GeneCards®
HUMAN GENE DATABASE

Free for academic non-profit institutions. Other users need a Commercial license

WEIZMANN INSTITUTE OF SCIENCE

LifeMap SCIENCES

Keywords | Search Term | Advanced

Home | User Guide | Analysis Tools | News And Views | About | My Genes | Log In / Sign Up

GeneCards®: The Human Gene Database

GeneCards is a searchable, integrative database that provides comprehensive, user-friendly information on all annotated and predicted human genes. It automatically integrates gene-centric data from ~125 web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

Explore a Gene

BTK GO

Jump to section for this gene:

Aliases, Disorders, Domains, Drugs, Expression, Function, Genomics, Localization, Orthologs, Paralogs, Pathways, Products, Proteins, Publications, Sources, Summaries, Transcripts, Variants

GeneCardsSuite

NGS Analysis Tools

TGex, VarElect

Affiliated Databases

MalaCards, LifeMap, PathCards, GeneLoc

Analysis Tools

GeneAnalytics, GeneALaCart, GenesLikeMe

NIH U.S. National Library of Medicine

MeSH

Search | Tree View | MeSH on Demand | MeSH 2016 | MeSH Suggestions

- Anatomy [A]
- Organisms [B]
- Diseases [C]
- Chemicals and Drugs [D]
- Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
- Psychiatry and Psychology [F]
- Phenomena and Processes [G]
- Disciplines and Occupations [H]
- Anthropology, Education, Sociology, and Social Phenomena [I]
- Technology, Industry, and Agriculture [J]
- Humanities [K]
- Information Science [L]
- Named Groups [M]
- Health Care [N]
- Publication Characteristics [V]
- Geographicals [Z]

- Neoplasms [C04]
  - Cysts [C04.182]
  - Hamartoma [C04.445]
  - Neoplasms by Histologic Type [C04.557]
  - Neoplasms by Site [C04.588]
    - Abdominal Neoplasms [C04.588.033]
    - Anal Gland Neoplasms [C04.588.083]
    - Bone Neoplasms [C04.588.149]
    - Breast Neoplasms [C04.588.180]
    - Digestive System Neoplasms [C04.588.274]
    - Endocrine Gland Neoplasms [C04.588.322]
    - Eye Neoplasms [C04.588.364]
    - Head and Neck Neoplasms [C04.588.443]
    - Hematologic Neoplasms [C04.588.448]
    - Mammary Neoplasms, Animal [C04.588.531]
    - Nervous System Neoplasms [C04.588.614]
    - Pelvic Neoplasms [C04.588.699]
    - Skin Neoplasms [C04.588.805]
    - Soft Tissue Neoplasms [C04.588.839]
    - Splenic Neoplasms [C04.588.842]
    - Thoracic Neoplasms [C04.588.894]
    - Urogenital Neoplasms [C04.588.945]
  - Neoplasms, Experimental [C04.619]
  - Neoplasms, Hormone-Dependent [C04.626]
  - Neoplasms, Multiple Primary [C04.651]
  - Neoplasms, Post-Traumatic [C04.666]
  - Neoplasms, Radiation-Induced [C04.682]
  - Neoplasms, Second Primary [C04.692]
  - Neoplastic Processes [C04.697]
  - Neoplastic Syndromes, Hereditary [C04.700]
  - Paraneoplastic Syndromes [C04.730]
  - Precancerous Conditions [C04.834]
  - Pregnancy Complications, Neoplastic [C04.850]

DRUGBANK Browse Search Downloads About Help Contact Us Search Drugs

Get DrugBank to go! The DrugBank app for iOS and Android is coming soon.

Sign up to get early access

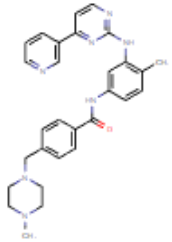
DRUGBANK

DrugBank Version 5.0

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information. The database contains 8261 drug entries including 2021 FDA-approved small molecule drugs, 233 FDA-approved biotech (protein/peptide) drugs, 94 nutraceuticals and over 6000 experimental drugs. Additionally, 4338 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. Each DrugCard entry contains more than 200 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

More about DrugBank

Search DrugBank Drugs Search

Identification	
Name	**Imatinib**
Accession Number	**DB00619** (APRD01028, EXPT02967, DB03261)
Type	Small Molecule
Groups	Approved
Description	Imatinib is a small molecule kinase inhibitor used to treat certain types of cancer. It is currently marketed by Novartis as Gleevec (USA) or Glivec (Europe/Australia) as its mesylate salt, imatinib mesilate (INN). It is occasionally referred to as CGP57148B or STI571 (especially in older publications). It is used in treating chronic myelogenous leukemia (CML), gastrointestinal stromal tumors (GISTs) and a number of other malignancies. It is the first member of a new class of agents that act by inhibiting particular tyrosine kinase enzymes, instead of non-specifically inhibiting rapidly dividing cells.
Structure	 MOL | SDF | 3D-SDF | PDB | SMILES | InChI | View 3D Structure
Synonyms	4-(4-METHYL-piperazin-1-ylmethyl)-N-[4-methyl-3-(4-pyridin-3-yl-pyrimidin-2-ylamino)-phenyl]-benzamide alpha-(4-Methyl-1-piperazinyl)-3'-((4-(3-pyridyl)-2-pyrimidinyl)amino)-P-toluidide Imatinib Imatinib Methansulfonate Imatinibum STI 571
External IDs	CGP-57148B / STI-571

Product Ingredients	Ingredient	UNII	CAS	InChI Key	Details
	Imatinib Mesylate	8A1O1M485B	220127-57-1	YLMAHDNUQAMNNX-UHFFFAOYSA-N	Details

Unified Medical Language System® (UMLS®)

Unified Medical Language System (UMLS)

The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. More information...

UMLS News RSS Feed

Access the UMLS

Sign up for a license, download files, and browse UMLS data.

Sign Up | Downloads | Browser | Knowledge Sources | API

Training and Documentation

View release information and learn how to use the UMLS.

Tutorials | Technical Documentation | Local Installation

UMLS Community

Learn about UMLS use across diverse organizations.

Applications | Listserv | User Contributions

Related Terminology Resources

Read more about the UMLS and other NLM products.

Vocabularies & Mappings | Publications | Health Information Technology

What is the UMLS?

The UMLS, or Unified Medical Language System, is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

You can use the UMLS to enhance or develop applications, such as electronic health records, classification tools, dictionaries and language translators.

UMLS in Use

One powerful use of the UMLS is linking health information, medical terms, drug names, and billing codes across different computer systems. Some examples of this are:

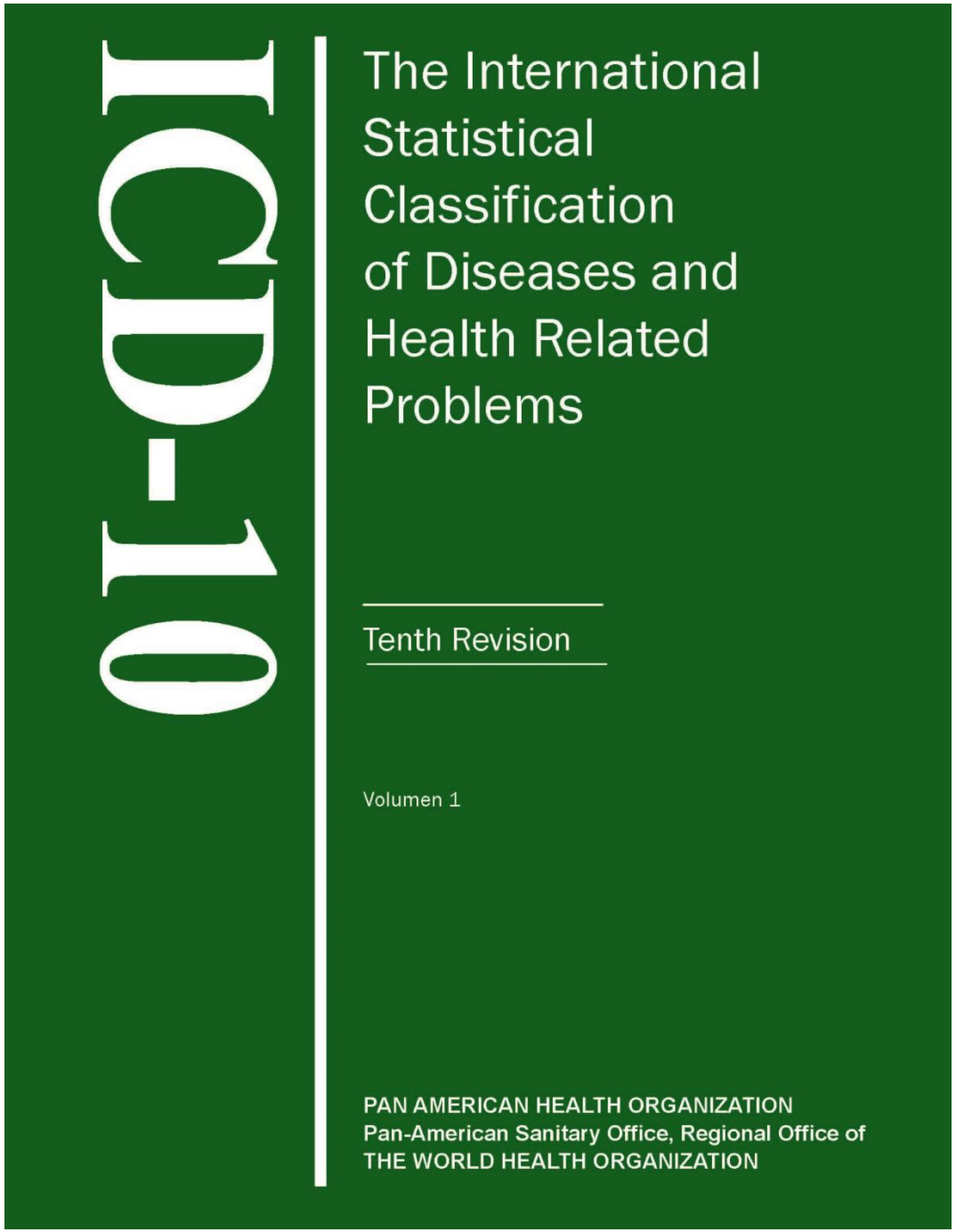
- Linking terms and codes between your doctor, your pharmacy, and your insurance company
- Patient care coordination among several departments within a hospital

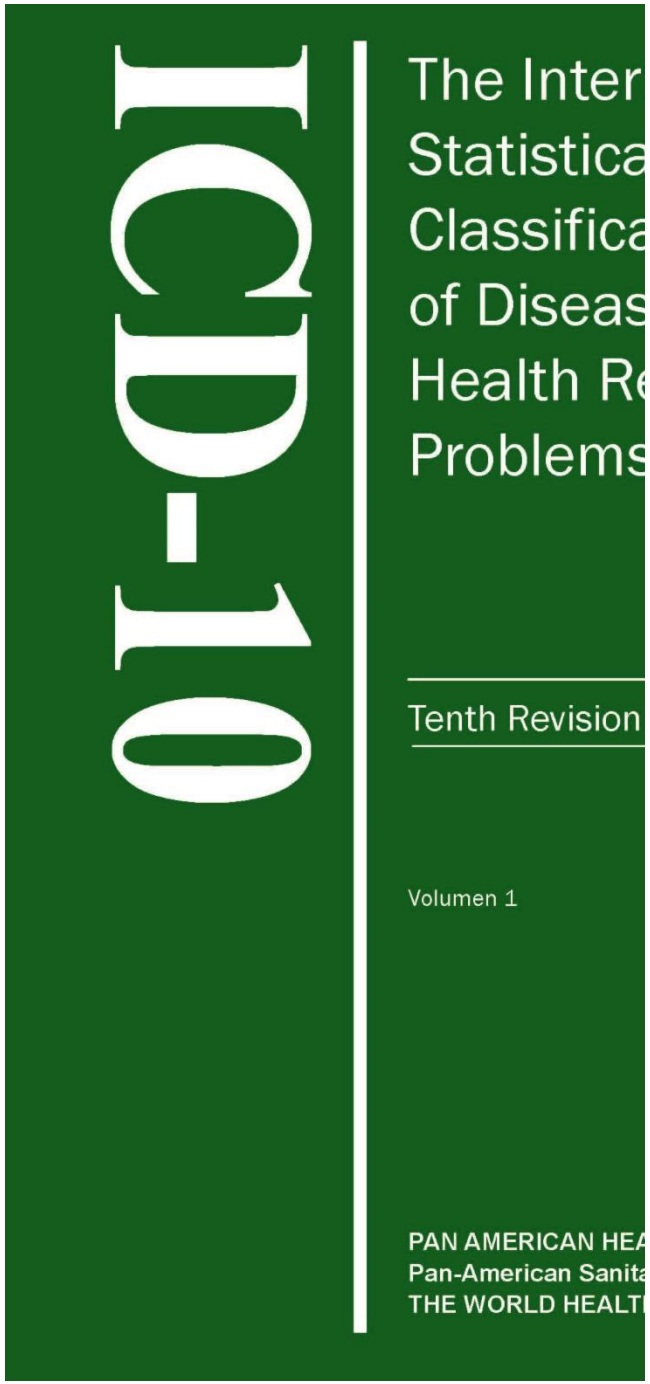
The UMLS has many other uses, including search engine retrieval, data mining, public health statistics reporting, and terminology research.

The Three UMLS Tools

The UMLS has three tools, which we call the Knowledge Sources:

- **Metathesaurus**: Terms and codes from many vocabularies, including CPT®, ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®
- **Semantic Network**: Broad categories (semantic types) and their relationships (semantic relations)
- **SPECIALIST Lexicon and Lexical Tools**: Natural language processing tools


ICD-10
The International Statistical Classification of Diseases and Health Related Problems
Tenth Revision
Volumen 1
PAN AMERICAN HEALTH ORGANIZATION
Pan-American Sanitary Office, Regional Office of
THE WORLD HEALTH ORGANIZATION



International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)-WHO Version for ;2016

Chapter II
Neoplasms
(C00-D48)

This chapter contains the following blocks:

- C00-C97 Malignant neoplasms
  - C00-C75 Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue
    - C00-C14 Malignant neoplasms of lip, oral cavity and pharynx
    - C15-C26 Malignant neoplasms of digestive organs
    - C30-C39 Malignant neoplasms of respiratory and intrathoracic organs
    - C40-C41 Malignant neoplasms of bone and articular cartilage
    - C43-C44 Melanoma and other malignant neoplasms of skin
    - C45-C49 Malignant neoplasms of mesothelial and soft tissue
    - C50-C50 Malignant neoplasm of breast
    - C51-C58 Malignant neoplasms of female genital organs
    - C60-C63 Malignant neoplasms of male genital organs
    - C64-C68 Malignant neoplasms of urinary tract
    - C69-C72 Malignant neoplasms of eye, brain and other parts of central nervous system
    - C73-C75 Malignant neoplasms of thyroid and other endocrine glands
  - C76-C80 Malignant neoplasms of ill-defined, secondary and unspecified sites
  - C81-C96 Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue
  - C97-C97 Malignant neoplasms of independent (primary) multiple sites
- D00-D09 In situ neoplasms
- D10-D36 Benign neoplasms
- D37-D48 Neoplasms of uncertain or unknown behaviour

Notes

1. Primary, ill-defined, secondary and unspecified sites of malignant neoplasm

Categories C76–C80 include malignant neoplasms for which there is no clear indication of the original site of the cancer or the cancer is stated to be 'disseminated', 'scattered' or 'spread' without mention of the primary site. In both cases the primary site is considered to be unknown.

2. Functional activity

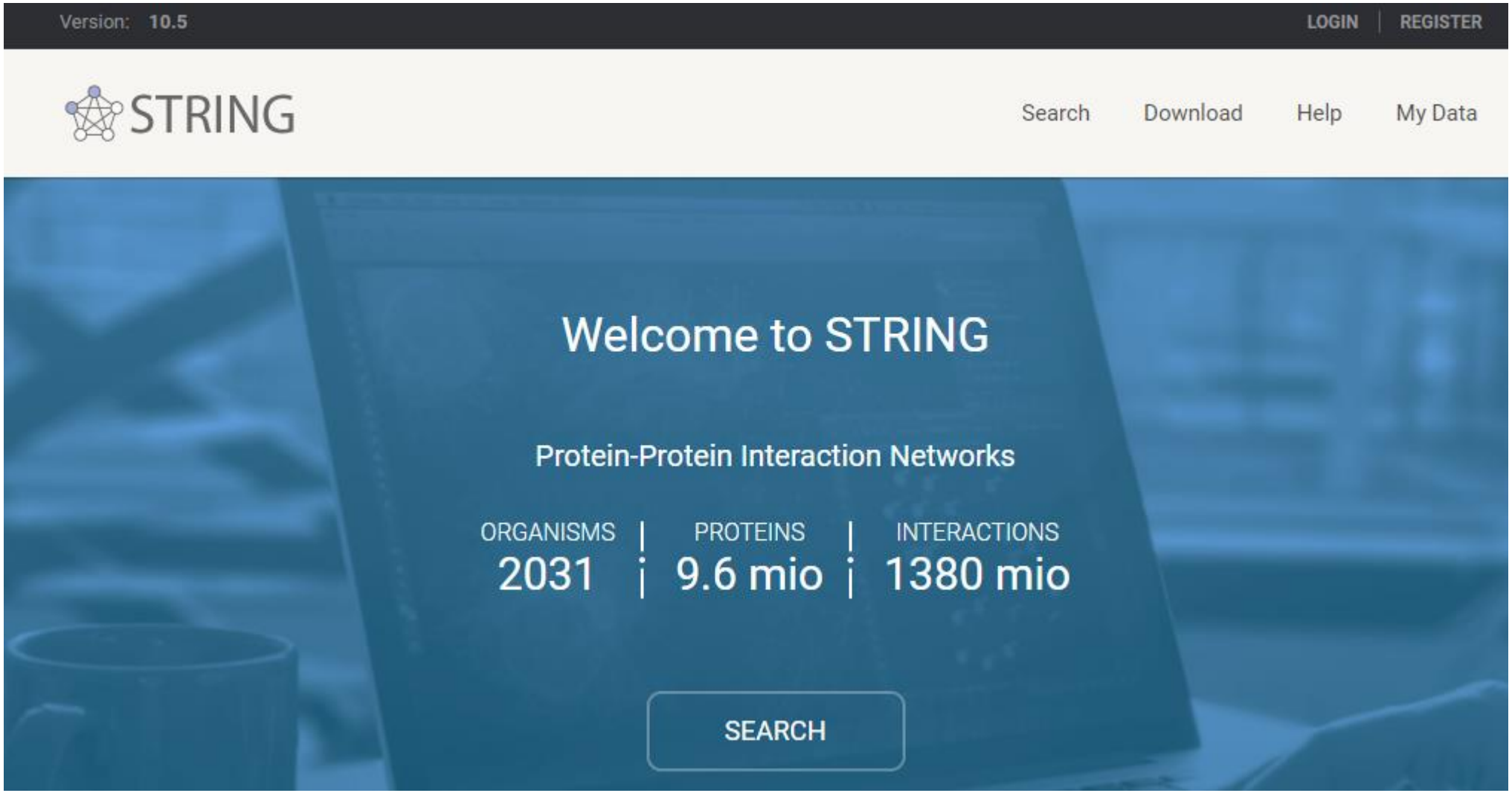
All neoplasms are classified in this chapter, whether they are functionally active or not. An additional code from Chapter IV may be used, if desired, to identify functional activity associated with any neoplasm. For example, catecholamine-producing malignant phaeochromocytoma of adrenal gland should be coded to C74 with additional code E27.5; basophil adenoma of pituitary gland with Cushing syndrome should be coded to D35.2 with additional code E24.0.

Databases

Anything of import → Manual KBs exist

Problem: Unsubstainable by manual effort

Free lunches abound for machine learning


Version: 10.5
LOGIN
REGISTER
STRING
Search
Download
Help
My Data
Welcome to STRING
Protein-Protein Interaction Networks
ORGANISMS
2031
PROTEINS
9.6 mio
INTERACTIONS
1380 mio
SEARCH

Search Gene | Downloads | Apps | FAQ | Publications | Contact

Version 8: Over 42 000 Pathways and 1 350 000 Interactions from 22 Data Sources

Pathway Commons

Pathway information. Single point of access.

Pathway Commons aims to store and disseminate knowledge about biological pathways. Information is sourced from public pathway databases and is readily searched, visualized, and downloaded. The data is freely available under the license terms of each contributing database.

Pathway Commons, a web resource for biological pathway data. Cerami E et al. Nucleic Acids Research (2011).

OncoKB Home About Team Levels of Evidence Actionable Genes Data Access News

OncoKB

Precision Oncology Knowledge Base

476	3701	65	97
Genes	Variants	Tumor Types	Drugs

Search Gene

Level 1	Level 2	Level 3	Level 4
FDA-approved	Standard care	Clinical evidence	Biological evidence
12 Genes	11 Genes	26 Genes	20 Genes

When using OncoKB, please cite: Chakravarty et al., JCO PO 2017

Shared Tasks

BioCreative

BioNLP

TREC

I2b2

SemEval

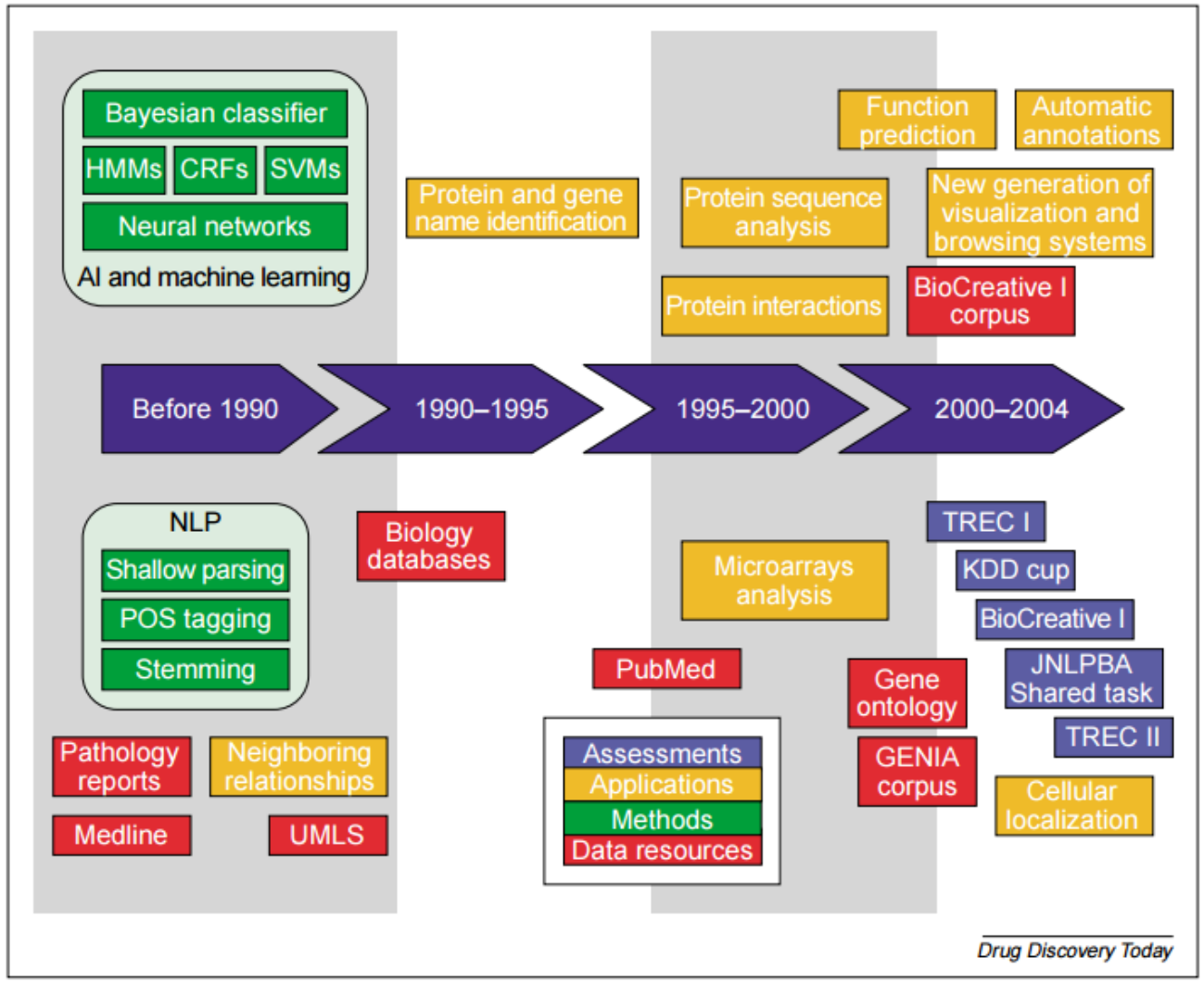


FIGURE 1

Historical perspective of the use of NLP in biomedicine and molecular biology. The hits are divided into different categories: dark-blue boxes show the different community-wide evaluations, whereas orange boxes refer to applications of text-mining strategies in biomedicine and molecular biology. Methods used for text mining and information extraction, such as artificial intelligence (AI), ML and statistical NLP techniques, are shown in green boxes, whereas relevant data resources are depicted in red boxes. Abbreviation: CRF, conditional random fields.

"Text-mining approaches in molecular biology and biomedicine". Martin Krallinger, Ramon Alonso-Allende Erhardt and Alfonso Valencia. Drug Discovery Today.

By Topic
News
Events
Tasks
Resources

By Chapter
BC Workshop '12
BCBioCuration2014
BioCreative 2016
BioCreative I
BioCreative I
BioCreative II
BioCreative II.5
BioCreative III
BioCreative IV
BioCreative V
BioCreative V.5

BioCreative VI

PM-task-trainingdata (Tasks) [2017-05-31]

Please download the training data for the Precision Medicine Task.

- Triage training dataset consists of 4082 annotated PubMed documents as relevant or not relevant. The file is prepared in BioC collection where each document contains two passages: the title and abstract. Each document in the collection has a document id corresponding to the article's PubMed ID, and an infon tag marking the document as relevant or not.

Feel free to contact task organizers for questions:

Task organizers:

Rezarta Islamaj Dogan (NCBI)
Andrew Chatr-aryamontri (BioGrid)
Sun Kim (NCBI)
Don Comeau (NCBI)
Zhiyong Lu (NCBI)

Downloads

- PMtask_Triage_TrainingSet

BioCreAtIvE evaluation

Organism (evaluation)	ProMiner®	Best performance
Mouse (BioCreAtIvE04)	0,79	0,79
Fly (BioCreAtIvE04)	0,82	0,82
Yeast (BioCreAtIvE04)	0,90	0,92
Human (BioCreAtIvE07)	0,80	0,81

Results in the international "critical assessment of text mining in biology" (BioCreAtIvE I and II).

BMC Bioinformatics. 2005;6 Suppl 1:S14. Epub 2005 May 24.

ProMiner: rule-based protein and gene entity recognition.

Hanisch D[1], Fundel K, Mevissen HT, Zimmer R, Fluck J.

Author information

Abstract

BACKGROUND: Identification of gene and protein names in biomedical text is a challenging task as the corresponding nomenclature has evolved over time. This has led to multiple synonyms for individual genes and proteins, as well as names that may be ambiguous with other gene names or with general English words. The Gene List Task of the BioCreAtIvE challenge evaluation enables comparison of systems addressing the problem of protein and gene name identification on common benchmark data.

METHODS: The ProMiner system uses a pre-processed synonym dictionary to identify potential name occurrences in the biomedical text and associate protein and gene database identifiers with the detected matches. It follows a rule-based approach and its search algorithm is geared towards recognition of multi-word names. To account for the large number of ambiguous synonyms in the considered organisms, the system has been extended to use specific variants of the detection procedure for highly ambiguous and case-sensitive synonyms. Based on all detected synonyms for one abstract, the most plausible database identifiers are associated with the text. Organism specificity is addressed by a simple procedure based on additionally detected organism names in an abstract.

RESULTS: The extended ProMiner system has been applied to the test cases of the BioCreAtIvE competition with highly encouraging results. In blind predictions, the system achieved an F-measure of approximately 0.8 for the organisms mouse and fly and about 0.9 for the organism yeast.

BioNLP-ST 2016

Navigation

- Home
- Tasks
  - BB3
  - GE4
  - SeeDev
- BioNLP–ST 2016 Workshop
  - BioASQ / BioNLP–ST Workshop Program
  - Student travel grants
- Previous Editions
  - 2009
  - 2011
  - 2013
- About Us

Home

The 4th BioNLP Shared Task in 2016

The BioNLP Shared Task (BioNLP-ST) series represents a community-wide trend in text-mining for biology toward fine-grained information extraction (IE). BioNLP-ST 2016 follows the general outline and goals of the previous tasks in 2011 and 2013. It identifies biologically relevant extraction targets and proposes a linguistically motivated approach to event representation.
As in previous events, manually annotated data is provided for training, development and evaluation of information extraction methods. According to their relevance for biological studies, the annotations are either bound to specific expressions in the text or represented as structured knowledge. Many tools for the detailed evaluation and graphical visualization of annotations and system outputs will be available for participants. Support in performing linguistic processing will be provided to the participants in the form of analyses created by various state-of-the art tools on the dataset texts.
Participation to the task is open to the academia, industry, and all other interested parties. The access to the on-line evaluation services remains open on each individual task page after the end of the official test period.

- The results of BioNLP-ST'16 will been presented at the BioNLP-ST workshop, organized jointly by BioNLP and BioASQ. It is collocated with ACL BioNLP workshop in Berlin in 2016. The proceedings are available as ACL archive.
- Note that the workshop will be two folds. The joint shared tasks workshop will be held on 13th, which is right "after" ACL conference, and it will be dedicated to the BioASQ and BioNLP-ST sessions. The BioNLP workshop will be held on 12th, and it will accommodate posters of shared task presentations.

BioNLP'09 Shared Task on Event Extraction

in conjunction with BioNLP, a NAACL-HLT 2009 workshop, June 4-5 2009, Boulder, Colorado

NOTICE:

- Call For Papers for the special issue of Computational Intelligence.
- Online evaluation service for the test data set available.
- Shared task data sets released to public.
- Evaluation tools released to public.

Contents

Introduction

The BioNLP'09 Shared Task concerns the recognition of bio-molecular events (bio-events) that appear in biomedical literature.

The definition of bio-event is broadly described as a change on the state of a bio-molecule or bio-molecules, e.g. phosphorylation of IkB involves a change on the protein IkB.

The goal of the shared task is to provide common and consistent task definitions, datasets and evaluation for bio-IE systems based on rich semantics and a forum for the presentation of varying but focused efforts on their development.

Task definition

The BioNLP'09 Shared Task focuses on extraction of bio-events particularly on proteins or genes. (Proteins and gene are not distinguished.)

To concentrate efforts on the novel aspects of the extraction task, it is assumed that the protein recognition has been already performed, and the shared task begins with a gold standard set of proteins annotations.

The shared task is designed to address a semantically rich IE problem as a whole, but divided into three subtasks to allow separate evaluation of the performance for different aspects of the problem.

Event Annotation

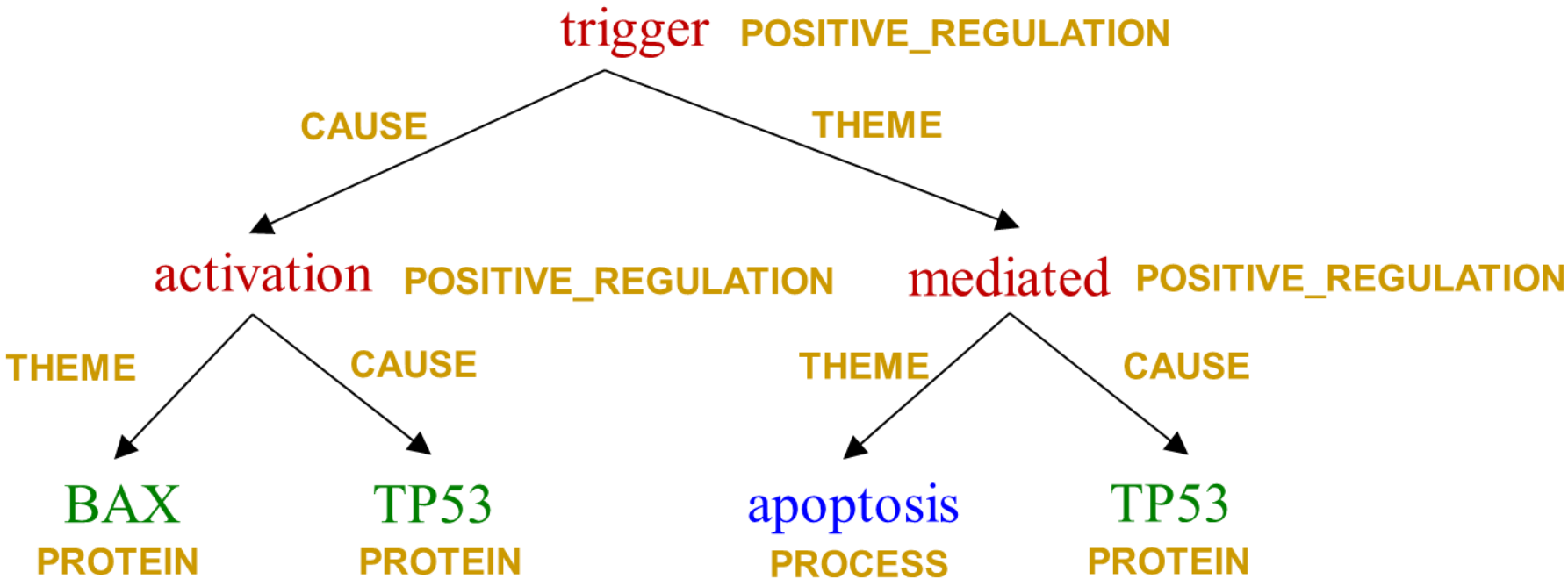
The activation of Bax by the tumor suppressor protein p53 is known to trigger the p53-mediated apoptosis …

T1	PROTEIN	BAX	19	22	Bax
T2	PROTEIN	TP53	30	58	tumor suppressor protein p53
T3	PROTEIN	TP53	83	86	p53
T4	PROCESS	apoptosis	96	105	apoptosis
T5	POSITIVE_REGULATION		5	15	activation
T6	POSITIVE_REGULATION		71	78	trigger
T7	POSITIVE_REGULATION		87	95	mediated

E1 T5 Theme:T1 Cause:T2
E2 T6 Theme:E3 Cause:E1
E3 T7 Theme:T4 Cause:T3

Event Annotation

The activation of Bax by the tumor suppressor protein p53 is known to trigger the p53-mediated apoptosis …



TREC Precision Medicine / Clinical Decision Support Track

Overview

Most work on precision medicine focuses on developing new treatments based on an individual's genetic, environmental, and lifestyle profile. The result is a data-driven approach investigating the best treatment for an individual patient. This promising approach has led to significant advances, including an explosion of scientific research, as embodied by the Precision Medicine Initiative (PMI). This presents an information problem for clinicians, however, as the vast literature available for precision medicine can make it difficult to find the most appropriate treatment for the clinician's current patient. The ability to quickly locate relevant information for a current patient using information retrieval (IR) has the potential to be an important tool for helping clinicians find the most up-to-date evidence-based treatment for their patients.

The TREC Precision Medicine track is a specialization of the previous TREC Clinical Decision Support track. Specifically, the 2017 Precision Medicine track focuses on the case of providing clinical decision support to cancer patients with genetic variations that might impact the choice of treatment. The track uses synthetic patients developed by precision oncologists at the world-famous MD Anderson Cancer Center in Houston, TX. For each patient, participants are challenged with retrieving relevant scientific literature articles discussing potential treatments, as well as potential clinical trials for which the patient may be eligible.

2017 Coordinators

Kirk Roberts, University of Texas Health Science Center at Houston (UTHealth)
William Hersh, Oregon Health and Science University (OHSU)
Dina Demner-Fushman, U.S. National Library of Medicine (NLM)
Ellen Voorhees, National Institute of Standards and Technology (NIST)
Alexander Lazar, University of Texas MD Anderson Cancer Center (MDACC)
Shubham Pant, University of Texas MD Anderson Cancer Center (MDACC)

Mailing List

http://groups.google.com/d/forum/trec-cds

Relational databases

NLP Tools Required

Diagnosis codes

Fake ID	ENTRY_DAT	CODE
34068	5/13/2001	41.85
37660	8/6/2002	79.99
140680	8/31/2003	79.99
23315	5/14/2003	112
75936	7/9/2004	117.9

Lab tests

Fake ID	TEST	ENTRY_DAT	VALU
3536	pO2	1/23/1996	314
72921	LDL	2/5/1996	34
102460	pCO2	1/26/1996	45
135043	HDL	1/25/1996	35
135432	MonAb	1/24/1999	0.16

Structured

Problem lists:

---- Medications known to be prescribed:
Keppra 750 mg 1/2 tab q am and pm
Dexilant 60 mg by mouth daily
aspirin 325 mg 1 tablet by mouth daily
clopidogrel 75 mg tablet 1 tablet by mouth daily

---- Known adverse and allergic drug reactions:
Sulfa Drugs

---- known significant medical diagnoses:
Seizure disorder
Aneurysm
Heartburn

---- Known significant operative and invasive procedures:
2003 Appendectomy
2005 Stents put in **DATE [Aug 29 05]

Semi-structured

Clinical notes

EXAM: BILATERAL DIGITAL SCREENING MAMMOGRAM WITH CAD, **DATE[Mar 16 01]: COMPARISON: **DATE[Jul 01 01] TECHNIQUE: Standard CC and MLO views of both breasts were obtained. FINDINGS: The breast parenchyma is heterogeneously dense. The pattern is extremely complex with postsurgical change seen in the right upper outer quadrant and scattered benign-appearing calcification seen bilaterally. A possible asymmetry is seen in the superior aspect of the left breast. The parenchymal pattern otherwise remains stable bilaterally, with no new distortion or suspicious calcifications. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy.. LEFT: Possible developing asymmetry superior aspect left breast for which further evaluation by true lateral and spot compression views recommended. Ultrasound may also be needed.. RECOMMENDATION: Left diagnostic mammogram with additional imaging as outlined above.. A left breast ultrasound may also be needed. BI-RADS Category 0: Incomplete Assessment - Need additional imaging evaluation. IMPRESSION: RIGHT: No interval change. No current evidence of malignancy….

Unstructured

"Extracting research-quality phenotypes from electronic health records to support precision medicine". Wei-Qi Wei and Joshua Denny. *Genome Medicine* 2015.

Table 1

Efforts and incentives to leverage clinical data for genomics research

Projects	Region	Start year	Website	Aims
eMERGE	United States	2007	http://emerge-network.org [152]	To develop methods and best practices for the utilization of EHRs for genetic research
i2b2	United States	2004	http://www.i2b2.org [153]	To provide researchers with useful tools to leverage EHRs for clinical and genetic research
PGPop	United States	2010	http://pgpop.mc.vanderbilt.edu [59]	To understand how a person's genes affect his or her response to medicines
deCODE genetics	Iceland	1996	http://www.decode.com [60]	To leverage population-based and EHR-linked biosamples to investigate inherited causes of common diseases
UK Biobank	United Kingdom	2007	http://www.ukbiobank.ac.uk [61]	To improve the prevention, diagnosis and treatment of a wide range of serious and life-threatening illnesses through a collection of around 500,000 volunteers' biosamples and clinical information
MVP	United States	2011	http://www.research.va.gov/mvp [52]	To enroll one million volunteers and use their clinical and genetic data to improve health care for veterans
KP RPGEH	United States	2009	http://www.rpgeh.kaiser.org [53]	To examine the genetic and environmental factors that influence common diseases
CKB	China	2004	http://www.ckbiobank.org [154]	To explore the complex interplay between genes and environmental factors on the risks of common chronic diseases

"Extracting research-quality phenotypes from electronic health records to support precision medicine". Wei-Qi Wei and Joshua Denny. *Genome Medicine* 2015.

i2b2
Informatics for Integrating Biology & the Bedside
A National Center for Biomedical Computing

About Us | Driving Biology Projects | Software | Resources | Events | Training | News | Collaborations | Publications

MISSION

i2b2 (Informatics for Integrating Biology and the Bedside) is an NIH-funded National Center for Biomedical Computing based at Partners HealthCare System. The i2b2 Center is developing a scalable informatics framework that will enable clinical researchers to use existing clinical data for discovery research and, when combined with IRB-approved genomic data, facilitate the design of targeted therapies for individual patients with diseases having genetic origins. This platform currently enjoys wide international adoption by the CTSA network, academic health centers, and industry. i2b2 is funded as a cooperative agreement with the National Institutes of Health.

i2b2 Foundation Community Wiki
Current i2b2 community projects sponsored here

DRIVING BIOLOGY PROJECTS

- Overview
- Current DBPs
  - Autoimmune/CV Diseases
  - Diabetes/CV Diseases
- Past DBPs
  - Airways Diseases
  - Hypertension
  - Type 2 Diabetes Mellitus
  - Huntington's Disease
  - Major Depressive Disorder
  - Rheumatoid Arthritis
  - Obesity

RESOURCES

- Overview
- Computational Tools
- Software
- NLP Research Data Sets
- NLP Shared Tasks
- Academic Users' Group
- Publication Data

SOFTWARE

i2b2 Downloadables | i2b2 Documentation | i2b2 Community Wiki

HIGHLIGHTS

*** i2b2 Installations Worldwide ***

i2b2 AUG AND SHRINE ANNUAL CONFERENCE

Tuesday, June 21, 2016
Agenda, Registration, and Lodging info here.

*** Precision Medicine 2015 Conference: Patient Driven ***

Wednesday, June 22, 2015
Agenda, Registration, and Lodging info here.

*** i2b2 NLP DATA SET #5 (from 2011 Challenge) ***

A complete set of annotated and unannotated, deidentified patient discharge summaries from the First, Second (Obesity),Third (Medication) and Fourth Shared Tasks for Challenges in NLP for Clincial Data are now available to the community for research purposes. Check it out at our NLP Data Sets page. Please note you must register **AND submit a DUA for access.**

Contact | Sitemap | Search

PARTNERS HEALTHCARE ©2017 Partners Healthcare

N-GRID | HARVARD MEDICAL SCHOOL | Department of Biomedical Informatics

UNIVERSITY AT ALBANY
State University of New York

2016 NLP Shared Task

NLP email password

CSAIL

- Home
- *User*
  - Login
- Announcements
- Documentation
- Workshop Schedule
- Workshop Registration
- About
- Previous Challenges

Announcement of Data Release and Call for Participation

2016 CEGS N-GRID Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data

Tentative Timeline
Registration: begins May, 2016
Data Release for Sight Unseen Track: 6th June 2016
System Outputs Due for Sight Unseen Track: 10th June 2016
Training Data Release: 11th June 2016
Test Data Release: 10th August 2016 (12am Eastern Time)
System Outputs Due: 12th August 2016 (11:59pm Eastern Time)
Abstract Submission: 1st September 2016
Workshop: 11th November 2016, Chicago, IL, USA
Journal Submissions: TBD

Registration

The 2016 Centers of Excellence in Genomic Science (CEGS) Neuropsychiatric Genome-Scale and RDOC Individualized Domains (N-GRID) challenge, a.k.a., RDoC for Psychiatry challenge, aims to extract symptom severity from neuropsychiatric clinical records. Research Domain Criteria (RDoC) is a framework developed under the aegis of the National Institute of Mental Health (NIMH) that facilitates the study of human behavior from normal to abnormal in various domains. The challenge goal is to classify symptom severity in a domain for a patient, based on information included in their initial psychiatric evaluation.

This challenge will be conducted on initial psychiatric evaluations (1 per patient), which have been fully de-identified and scored by clinical experts in a symptom domain. The data for this task is provided by Partners Healthcare and the Neuropsychiatric Genome-Scale and RDoC Individualized Domains (N-GRID) project (HMS PI: Kohane; MGH PI: Perlis) of Harvard Medical School, and will be released under a Rules of Conduct and Data Use Agreement. Obtaining the data requires completing the registration, which will start in May 2016.

All data are fully de-identified and manually annotated for RDoC.

The tracks
The 2016 CEGS N-GRID challenge consists of three NLP tracks:

Track 1: De-identification: Removing protected health information (PHI) is a critical step in making medical records accessible to more people, yet it is a very difficult and nuanced task. This track addresses the problem of de-identifying medical records over a new set of ~1000 initial psychiatric evaluation records, with surrogate PHI for participants to identify. We intend to run two versions of the de-id track.

1. **Sight unseen track**: this track involves running existing home-grown de-id systems on the RDoC data without any training and modification to the systems, as a way of measuring how well the existing systems generalize to brand new data. The RDoC data will be provided for this track without any gold standard training annotations and system outputs will be collected within 3 days of data release.
2. **Regular track**: this track will allow the development and training of de-id systems on the RDoC training data. Evaluation will be on the RDoC test data.

Track 2: RDoC classification: The goal of RDoC classification is to determine symptom severity in a domain for a patient, based on information included in their initial psychiatric evaluation. The domain has been rated on an ordinal scale of 0-3 as follows: 0 (absent), 1 (mild=modest significance), 2 (moderate=requires treatment), 3 (severe=causes substantial impairment) by experts. There is one judgment per document, and one document per patient.

Track 3: Novel Data Use: The data released for this 2016 challenge are the first set of mental health records released to the research community. These data can be used for mental health-related research questions that go beyond what is posed by the challenge organizers. This Track is for participants who want to build on their existing systems, or the systems developed for Tracks 1 and 2, with the aim of addressing new research questions.

Home | Task Description | Data and Tools | Results | Relevant Publications | Important Dates

FAQs

SemEval-2015 Task 14

SemEval-2015 Task 14: Analysis of Clinical Text

The purpose of this task is to enhance current research in natural language processing methods used in the clinical domain. The second aim of the task is to introduce clinical text processing to the broader NLP community. The task aims to combine supervised methods for text analysis with unsupervised approaches. More specifically, the task aims to combine supervised methods for entity/acronym/abbreviation recognition and mapping to UMLS CUIs (Concept Unique Identifiers) with access to larger clinical corpus for utilizing unsupervised techniques. It also comprises the task of identifying various attributes of the disorders and normalizing their values. We refer to this as the template filling task.

Contact Info

Organizers (in alphabetical order)

- Wendy W. Chapman, University of Utah
- Noemie Elhadad, Columbia University
- Suresh Manandhar, University of York, UK
- Sameer S. Pradhan, Harvard University
- Guergana K. Savova, Harvard University

Contact:

- Guergana.Savova@childrens.harvard.edu
- Noemie.Elhadad@columbia.edu

Other Info

Home | Important Dates | Software

SemEval-2017 Task 12
Clinical TempEval

Clinical TempEval

Clinical TempEval 2017 follows in the footsteps of the i2b2 2012 shared task, Clinical TempEval 2015, and Clinical TempEval 2016 in bringing timeline extraction to the clinical domain. As in past Clinical TempEvals, data will be drawn from clinical notes and pathology reports for cancer patients at the Mayo Clinic.

New in 2017

This year, Clinical TempEval will focus on domain adaptation: systems will be trained on data from colon cancer patients, but will be asked to make predictions on brain cancer patients. Adapting to the many differences between the two domains will be a key challenge for the task.

Participants

For more details, including what tasks are included, where to obtain the data, and how to submit your system output, visit the Clinical TempEval 2017 competition on CodaLab.

Please also sign up on the mailing list: clinical-tempeval@googlegroups.com.

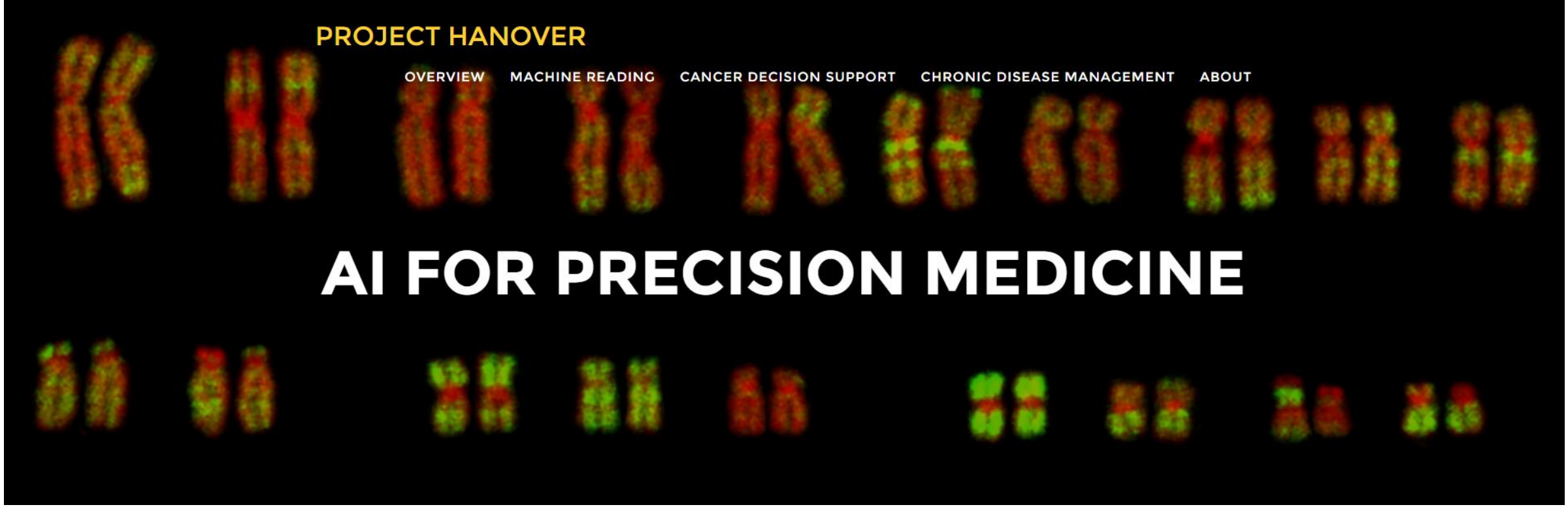
Contact Info

Organizers:
- Steven Bethard
- Guergana Savova
- Martha Palmer
- James Pustejovsky

Mailing List:
clinical-tempeval@googlegroups.com

Other Info

Announcements
- 1 Dec 2016 - The CodaLab competition site is available.
- 11 Sep 2016 - Source domain training data is available. See the Data page for details.

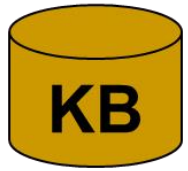


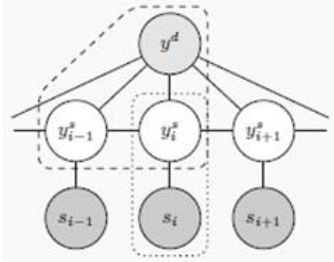
Knowledge

Machine Reading

Can be done manually,
need automation to scale

E.g., PubMed search





http://hanover.azurewebsites.net

Reasoning

Predictive Analytics

Can't be done manually,
need automation to enable

E.g., personalize drug combinations

Community Portal for Precision Medicine

Tasks

Datasets

Source codes

Leader board

Part 8: Open Problems

Grand challenges

How to maximize impact

How to measure progress

Where to find applications

Reality check

Grand Challenge: Solve Cancer

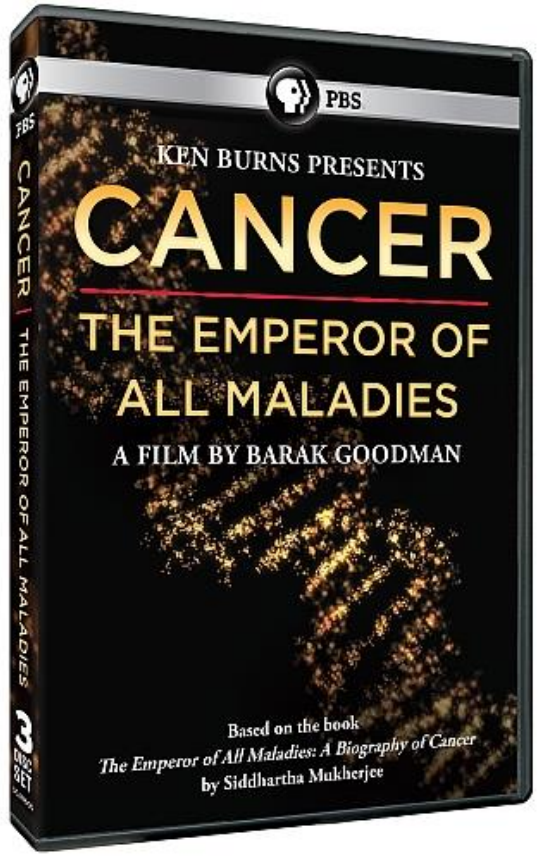
Goal: Turn cancer into a non-fatal disease

Prevention, detection, treatment

Tailor to individuals

NLP can play a key role

- Knowledge: Machine reading
- Reasoning: Knowledge-rich ML



Grand Challenge: Precision Healthcare

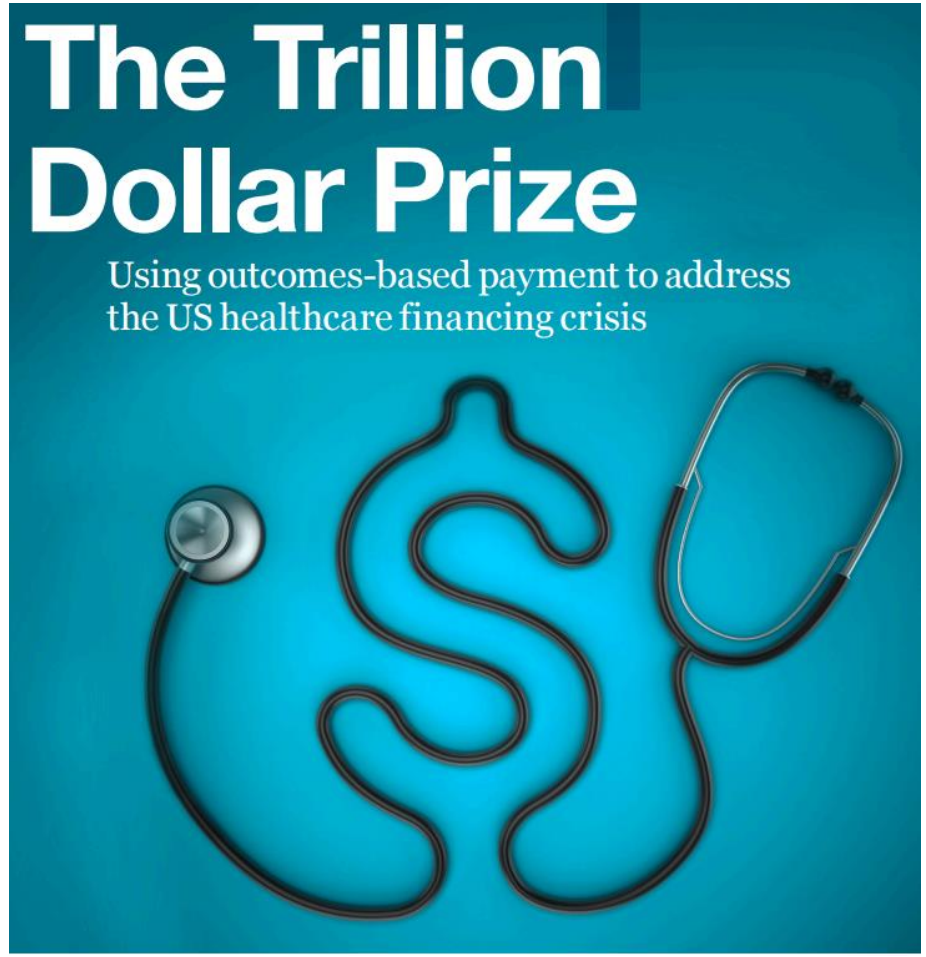
Annual spending: $3 trillion

Chronic diseases = 86% cost

Genomics less important

EMR; 24 x 7 sensor data

Wanted: Predict & prevent



How to Maximize Impact

Think end-to-end scenarios

"What difference can it make if we get 100%"

Case in point: Alignment for machine translation

How to Measure Progress

"What accuracy to be usefully deployed?"

Human-machine symbiosis

E.g.: machine reading → curation candidates

Feedback loop

High-recall, reasonable precision

Where to find applications

Follow the text: Literature, EMR notes, clinical trials, radiology reports, tumor board meetings, …

What to do with my hammer?

Syntactic Parsing

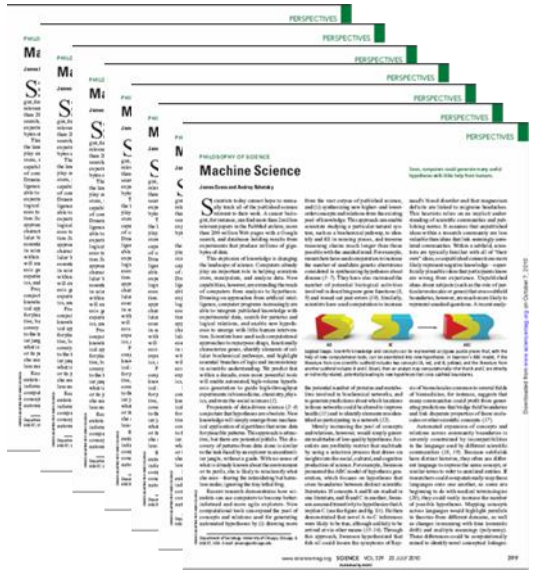
Key to many downstream tasks

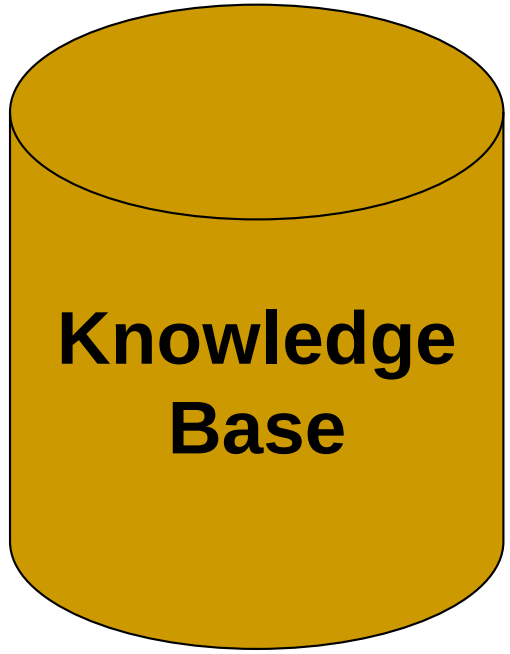
Challenge: Adapt to biomed text

Semantics

Prior work focuses on parsing questions

Priority = Extract structured information





Discourse

Prior work focuses on newswire/web

Adapt to biomed domains

Connect to end tasks

E.g.: Cross-sentence machine reading

Dialog

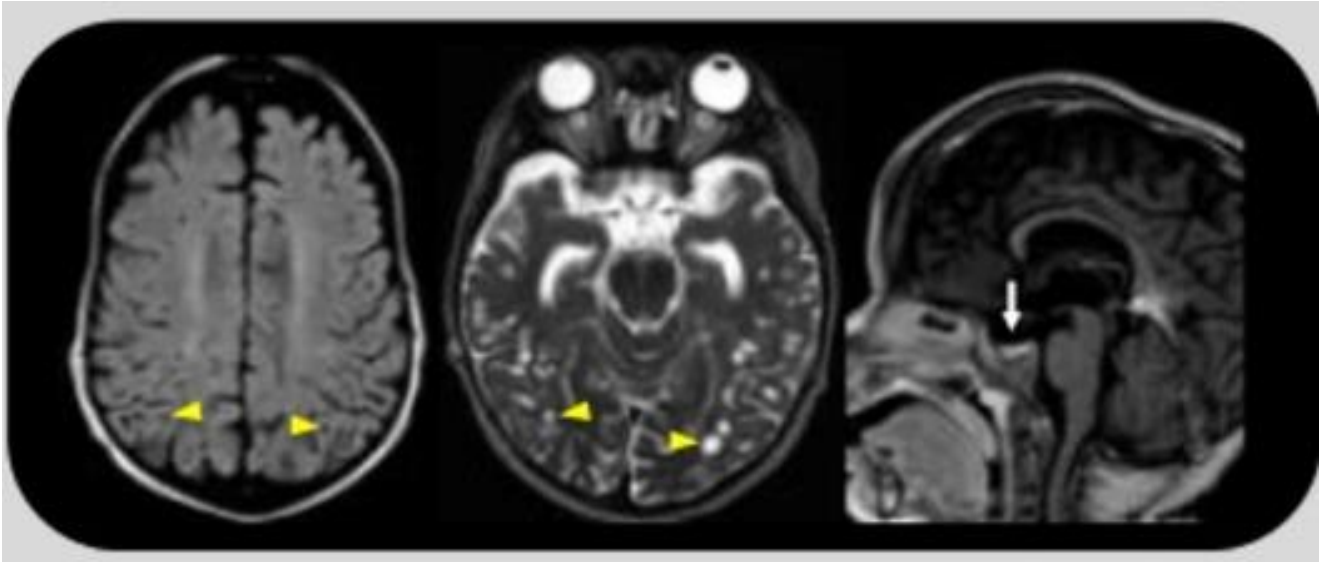
AI bot for
molecular tumor board

Language-Vision

It is fun …

Five cows graze on a grass land

Language-Vision



"Step up to bat and practice dictating complex cases"
Mamlouk & Sonnenberg

It is fun …

Findings:
There are numerous perivascular spaces bilaterally that follow CSF signal. The sella is J-shaped.

Impression:
Findings suggestive of a mucopolysaccharidosis (Hurler disease, in this case)

and might save life!

Medical Image Net

A petabyte-scale, cloud-based, multi-institutional, searchable, open repository of diagnostic imaging studies for developing intelligent image analysis systems.

Featured Goals

- Data migration/federation/honest broker
- Linkage to EMR and multi-omics
- Cohort discovery tools
- Image viewing software
- Governance
- Image classification and annotation
  - Natural language processing, research data sets, crowd source

It is fun ...

Findings:
There are numerous perivascular spaces bilaterally that follow CSF signal. The sella is J-shaped.

Impression:
Findings suggestive of a mucopolysaccharidosis (Hurler disease, in this case)

and might save life!

Summarization

Medical error = Third top killer

Imagine an ICU nurse in a new shift:

Read 20 pages of notes in 2 mins …

Not your traditional summarization

Contextual, knowledge-rich

Reality Check

Entry barrier

Data access

Engagement

“Biomedicine is an ocean that’s one meter deep”

Data Access

Literature: Publishers against text mining

Medical records: Privacy

Successes can help turn the tide

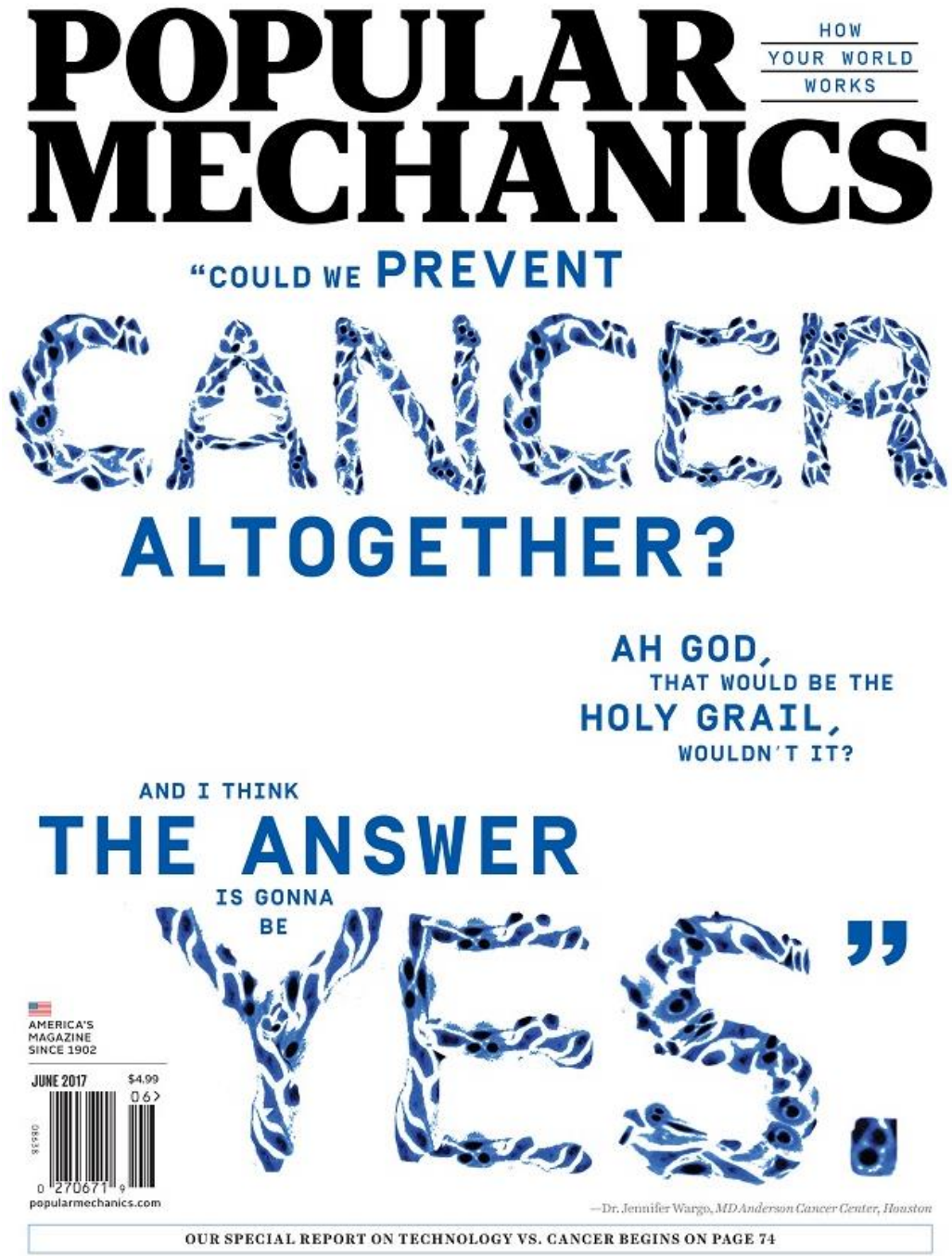
Engagement

Deep partnership is rewarding

Need to bridge disciplines

Patience, patience, patience

E.g.: BeatAML – started in 2014



Helping some cancer patients, the luckiest of the unlucky, live in relative normalcy for years is not just possible. It is happening.

Breaking News: The emperor of all maladies abdicates

Summary

AI for Precision medicine

Machine reading: Text → KB

Predictive analytics: Data + Knowledge → Decision

Machine learning: Annotation bottleneck

Many nails for your NLP hammer

References: Distant Supervision

Constructing biological knowledge bases by extracting information from text sources. Mark Craven and Johan Kumlien. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, 1999.

Distant supervision for relation extraction without labeled data. Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. ACL 2009.

Modeling relations and their mentions without labeled text. Sebastian Riedel, Limin Yao, and Andrew McCallum. In Proceedings of the Sixteen European Conference on Machine Learning, 2010.

Knowledge-based weak supervision for information extraction of overlapping relations. Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. ACL 2011.

Distant Supervision for Cancer Pathway Extraction from Text. Hoifung Poon, Kristina Toutanova, and Chris Quirk. In Proceedings of the Pacific Symposium on Biocomputing, 2015.

Incidental Supervision: Moving beyond Supervised Learning. *Dan Roth. Senior Member Summary Track, AAAI 2017.*

References: Complex Semantics

Driving semantic parsing from world's response. James Clarke, Dan Goldwasser, Ming-Wei Chang, and Dan Roth. CoNLL 2010.

Learning dependency-based compositional semantics. Percy Liang, Michael I. Jordan, Dan Klein. ACL 2011.

Weakly supervised training of semantic parsers. Jayant Krishnamurthy and Tom M. Mitchell. EMNLP 2012.

Scaling semantic parsers with on-the-fly ontology matching. T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer. EMNLP 2013.

Semantic parsing via paraphrasing. Jonathan Berant, Percy Liang. Association for Computational Linguistics (ACL), 2014.

Large-scale semantic parsing without question-answer pairs. Siva Reddy, Mirella Lapata, and Mark Steedman. TACL 2014.

Grounded Semantic Parsing for Complex Knowledge Extraction. Ankur Parikh, Hoifung Poon, and Kristina Toutanova. NAACL 2015.

Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, Jianfeng Gao. ACL 2015.

References: Cross-Sentence Extraction

Automatically semantifying wikipedia. Fei Wu and Daniel S. Weld. CIKM 2007.

Extracting relations within and across sentences. Kumutha Swampillai and Mark Stevenson. RANLP 2011.

Type-aware distantly supervised relation extraction with linked arguments. Mitchell Koch, John Gilmer, Stephen Soderland, and Daniel S. Weld. EMNLP 2014.

Distantly supervised web relation extraction for knowledge base population. Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Semantic Web 2016.

Distant Supervision for Relation Extraction beyond the Sentence Boundary. Chris Quirk and Hoifung Poon. EACL 2017.

Cross-Sentence N-ary Relation Extraction with Graph LSTMs. Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Scott Yih. TACL 2017.

References: Reasoning (1)

Translating embeddings for modeling multi-relational data. Antoine Bordes, Nicolas Usunier, Alberto GarciaDuran, Jason Weston, and Oksana Yakhnenko. In Advances in Neural Information Processing Systems (NIPS), 2013.

Embedding entities and relations for learning and inference in knowledge bases. Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. In International Conference on Learning Representations (ICLR), 2015.

Representing Text for Joint Embedding of Text and Knowledge Bases. Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. EMNLP 2015.

Compositional Learning of Embeddings for Relation Paths in Knowledge Bases and Text. Kristina Toutanova, Xi Victoria Lin, Wen-Tau Yih, Hoifung Poon, and Chris Quirk. ACL 2016.

Introduction to Statistical Relational Learning. Lise Getoor and Ben Taskar. (Eds). MIT press, 2007.

Random walk inference and learning in a large scale knowledge base. Ni Lao, Tom Mitchell, William Cohen. EMNLP 2011.

Reading the web with learned syntactic-semantic inference rules. Ni Lao, Amarnag Subramanya, Fernando Pereira, and William W. Cohen. EMNLP 2012.

References: Reasoning (2)

A three-way model for collective learning on multi-relational data. Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. ICML 2011.

A review of relational machine learning for knowledge graphs. Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. arXiv preprint arXiv:1503.00759 (2015).

Learning Structured Embeddings of Knowledge Bases. Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. AAAI 2011.

Relation Extraction with Matrix Factorization and Universal Schemas. Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. HLT-NAACL. 2013.

Knowledge vault: A web-scale approach to probabilistic knowledge fusion. Dong, Xin, et al. KDD 2014.

Matrix and Tensor Factorization Methods for Natural Language Processing. Bouchard, Guillaume, et al.. ACL (Tutorial Abstracts). 2015

Multilingual relation extraction .using compositional universal schema. Verga et al. NAACL-HLT 2016.

Traversing knowledge graphs in vector space. Guu et al. EMNLP 2015.

References: Reasoning (3)

Compositional Vector Space Models for Knowledge Base Completion. Neelakantan et al. ACL 2015.

Modeling relation paths for representation learning of knowledge bases. Lin et al. EMNLP 2015.

Improving learning and inference in a large knowledge-base using latent syntactic cues. Gardner et al. EMNLP 2013.

Incorporating vector space similarity in random walk inference over knowledge bases. Gardner et al. EMNLP 2014.

Chains of reasoning over entities, relations, and text using recurrent neural networks. Das et al. arXiv preprint arXiv:1607.01426, 2016.

References: Applications

Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records". Miotto et al. *Scientific Reports 2016*.

Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. Blecker et al. *JAMA Cardiology 2016*.

Using machine learning to parse breast pathology reports. Yala et al. *Breast Cancer Research and Treatment 2017*.

Identifying Combinations of Targeted Agents for Hematologic Malignancies. Kurtz et al. *PNAS, to appear*.