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

IMPRECISION MEDICINE
For every person they do help (blue), the ten highest-grossing drugs in the United States fail to improve the conditions of between 3 and 24 people (red).

1. ABILIFY (aripiprazole)
Schizophrenia

2. NEXUM (esomeprazole)
Heartburn

3. HUMIRA (adalimumab)
Arthritis

4. Crestor (rosuvastatin)
High cholesterol

5. CYMBALTA (duloxetine)
Depression

6. ADVAIR DISKUS (fluticasone propionate)
Asthma

7. ENBREL (etanercept)
Psoriasis

8. REMICADE (infliximab)
Crohn's disease

9. COPAXONE (glatiramer acetate)
Multiple sclerosis

10. NEULASTA (pegfilgrastim)
Neutropenia

Based on published number needed to treat (NNT) figures. For a full list of references, see Supplementary information at genetics.nature.com/484/NM
Disruption: Big Data

Accenture study: 93% of US doctors using EMRs

2009 – 2013: 40% → 93%
Disruption: Pay-for-Performance

Target percentage of Medicare FFS payments linked to quality and alternative payment models in 2016 and 2018

- All Medicare FFS (Categories 1-4)
- FFS linked to quality (Categories 2-4)
- Alternative payment models (Categories 3-4)

2016:
- 30%
- 85%

2018:
- 50%
- 90%

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
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

The New Hope

Think HIV

Example: Gleevec for CML

Cancer → Chronic disease
Why We Haven’t Solved Precision Medicine?

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

AI is the key to overcome these bottlenecks
Molecular Tumor Board
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
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 Joséé 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
Jiaoqiao 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
Minal Gounder, MD
James Harding, MD
Matthew Hellman, MD
Alan Ho, MD, PhD
Gopa Iyer, MD
Yelena Janjigian, MD
Thomas Kaloy, MD
Maeve Lowery, MD
Antonio Omuro, MD
Paul Paik, MD
Michael Postow, MD
Dana Rathkopf, MD
Alexander Shousharti, 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
Kinala Gala, BSc

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
Can we help increase curation speed by 100X?
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
VDR+ binds to SMAD3 to form...

JUN expression is induced by SMAD3/4...

PMID: 123

PMID: 456
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...
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**
  - Theme: up-regulation
  - Cause: activation

- **Theme**
  - IL-10
  - gp41

- **Cause**
  - REGULATION
  - Site: human monocyte

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

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 partici-
pating systems.

1 Introduction
Bio-entity recognition aims to identify and clas-

Protein, DNA, RNA, cell line, cell type
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
What we need is entity linking (a.k.a. normalization)

Lessons learned
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.
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 ...
Machine Reading

Prior work
- Focused on Newswire / Web
- Popular entities and facts
- Redundancy $\rightarrow$ 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?
Crowdsourcing: “Are these English?”
Learning with Indirect Supervision

Unsupervised learning
Statistical relational learning
Distant supervision
Incidental learning
Situated learning
Grounded language learning
Grounded Learning

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Involvement

up-regulation

Theme

Cause

Site

human monocyte

CELL

p70(S6)-kinase

GENE

Context
Grounding Takes Many Forms

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

Image from Artzi & Zettlemoyer 2013
Grounding Takes Many Forms

Example from Liang et al. 2011

What is the most populous city in California?

Los Angeles

[argmax(\(\lambda x.\text{city}(x) \land \text{loc}(x, \text{CA}), \lambda x.\text{population}(x))\)]

[Clark et al. 2010; Liang et al. 2011; ......]
Free Lunch: Existing KB

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

NCI Pathway KB
## Free Lunch: Existing KB

**NCI Pathway KB**

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
### NCI Pathway KB

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Regulation</th>
<th>Theme</th>
<th>Cause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>A2M</td>
<td>FOXO1</td>
</tr>
<tr>
<td>Positive</td>
<td>ABCB1</td>
<td>TP53</td>
</tr>
<tr>
<td>Negative</td>
<td>BCL2</td>
<td>TP53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*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 ...
TP53 inhibits BCL2.

Tumor suppressor P53 down-regulates the activity of BCL2. BCL2 transcription is suppressed by P53 expression.

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

......
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]
Combatting Noise

Introduce latent variables
Case study: Riedel, Hoffman, Betteridge
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 ...

[y founded = 1]

[y founded = 1]

[y founded = 0]
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

Steve Jobs × Apple

Steve Jobs was a founder of Apple.
Steve Jobs, Steve Wozniak, and Ronald Wayne founded Apple.
Steve Jobs is the CEO of Apple.

Here: exists ≥ 0
Could say: true ≥ α, for α ∈ (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)

(a) Extraction (E)  (b) Quantification (Q)  (c) Quantifier ambiguity (Q, Q)  (d) Quantification (Q, E)
Example: Liang et al. 2011

Grounding: KB query yields correct answer
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 $\leq 10$
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
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]

Tree HMM

(PAS-REG, BCL, (NEG-REG, IL-10, RFLAT))
(NEG-REG, TP53, (POS-REG, BCL, IL-2))
(POS-REG, AKT2, (POS-REG, IL-4, ERBB2))
(NEG-REG, (POS-REG, BCL, IL-2), BRAF)

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

\[ \theta^* = \arg \max_\theta \log P_\theta(T|K) \]

\[ = \arg \max_\theta \sum_{t \in T} \log \sum_z P_\theta(z, t) \cdot \phi_K(z) \]

Virtual Evidence
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

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>76.4</td>
<td>81.5</td>
<td>78.8</td>
</tr>
<tr>
<td>Transcription</td>
<td>49.4</td>
<td>73.6</td>
<td>59.1</td>
</tr>
<tr>
<td>Catabolism</td>
<td>65.6</td>
<td>80.0</td>
<td>74.4</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>73.9</td>
<td>84.5</td>
<td>78.9</td>
</tr>
<tr>
<td>Localization</td>
<td>74.6</td>
<td>75.8</td>
<td>75.2</td>
</tr>
<tr>
<td>Binding</td>
<td>48.0</td>
<td>50.9</td>
<td>49.4</td>
</tr>
<tr>
<td>Regulation</td>
<td>32.5</td>
<td>47.1</td>
<td>38.6</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>38.7</td>
<td>51.7</td>
<td>44.3</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>35.9</td>
<td>54.9</td>
<td>43.9</td>
</tr>
<tr>
<td><strong>Total Event F1</strong></td>
<td>50.2</td>
<td>62.6</td>
<td>55.7</td>
</tr>
</tbody>
</table>
## Preliminary Results

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>50.8</td>
<td>41.9</td>
<td>45.9</td>
</tr>
<tr>
<td>Transcription</td>
<td>18.3</td>
<td>14.0</td>
<td>15.9</td>
</tr>
<tr>
<td>Catabolism</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>36.2</td>
<td>43.6</td>
<td>39.5</td>
</tr>
<tr>
<td>Localization</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Binding</td>
<td>24.0</td>
<td>42.6</td>
<td>30.7</td>
</tr>
<tr>
<td>Regulation</td>
<td>2.5</td>
<td>5.0</td>
<td>3.3</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>11.4</td>
<td>21.4</td>
<td>14.9</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>4.4</td>
<td>16.4</td>
<td>6.9</td>
</tr>
<tr>
<td><strong>Total Event F1</strong></td>
<td>19.1</td>
<td>29.4</td>
<td>23.2</td>
</tr>
</tbody>
</table>
Prototype-Driven Learning

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

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Rec.</th>
<th>Prec.</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expression</td>
<td>55.3</td>
<td>88.3</td>
<td>68.0</td>
</tr>
<tr>
<td>Transcription</td>
<td>50.0</td>
<td>39.1</td>
<td>43.9</td>
</tr>
<tr>
<td>Catabolism</td>
<td>52.4</td>
<td>100.0</td>
<td>68.9</td>
</tr>
<tr>
<td>Phosphorylation</td>
<td>61.7</td>
<td>82.9</td>
<td>70.7</td>
</tr>
<tr>
<td>Localization</td>
<td>52.8</td>
<td>100.0</td>
<td>69.1</td>
</tr>
<tr>
<td>Binding</td>
<td>20.2</td>
<td>92.7</td>
<td>33.2</td>
</tr>
<tr>
<td>Regulation</td>
<td>24.1</td>
<td>64.0</td>
<td>35.0</td>
</tr>
<tr>
<td>Positive_regulation</td>
<td>17.4</td>
<td>63.8</td>
<td>27.4</td>
</tr>
<tr>
<td>Negative_regulation</td>
<td>8.4</td>
<td>52.8</td>
<td>14.5</td>
</tr>
<tr>
<td><strong>Total Event F1</strong></td>
<td>27.9</td>
<td>72.2</td>
<td>40.2</td>
</tr>
</tbody>
</table>
Incomplete KB

GENIA Event Extraction F1

10% Database
50% Database
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)
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

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

This shows that **Notch**-mediated resistance of a mouse lymphoma cell line could be overcome by inhibiting p56Lck.
Features

Prior work: Used single shortest path
DISCREX: Multiple paths help

Templates
- Nodes: Token, lemma, POS
- Whole paths
- Path n-grams
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)
### Gene-Drug Knowledge Database [Dienstmann et al. 2015]

<table>
<thead>
<tr>
<th>Disease</th>
<th>Gene</th>
<th>Variant</th>
<th>Description</th>
<th>Effect</th>
<th>Association</th>
<th>Therapeutic context</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>T315A</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>nilotinib, ponatinib</td>
</tr>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>T315I</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>ponatinib</td>
</tr>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>F317L/V/I/C</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>nilotinib, ponatinib</td>
</tr>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>F359V/C/I</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>dasatinib, ponatinib</td>
</tr>
<tr>
<td>CML</td>
<td>ABL1</td>
<td>T315A</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>nilotinib, bosutinib, ponatinib</td>
</tr>
<tr>
<td>CML</td>
<td>ABL1</td>
<td>T315I</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>ponatinib</td>
</tr>
<tr>
<td>CML</td>
<td>ABL1</td>
<td>F317L/V/I/C</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>nilotinib, bosutinib, ponatinib</td>
</tr>
<tr>
<td>CML</td>
<td>ABL1</td>
<td>F359V/C/I</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>dasatinib, bosutinib, ponatinib</td>
</tr>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>Y253H</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>dasatinib, ponatinib</td>
</tr>
<tr>
<td>ALL</td>
<td>ABL1</td>
<td>E255K/V</td>
<td>missense mutation</td>
<td>gain-of-function</td>
<td>response</td>
<td>dasatinib, ponatinib</td>
</tr>
</tbody>
</table>
## PubMed-Scale Extraction

<table>
<thead>
<tr>
<th>Relations</th>
<th>Single-Sent.</th>
<th>Cross-Sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>169,168</td>
<td>332,969</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>32,028</td>
<td>64,828</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>17,349</td>
<td>32,775</td>
</tr>
<tr>
<td>GDKD</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>
## PubMed-Scale Extraction

<table>
<thead>
<tr>
<th>Relations</th>
<th>Single-Sent.</th>
<th>Cross-Sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>169,168</td>
<td>332,969</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>32,028</td>
<td>64,828</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>17,349</td>
<td>32,775</td>
</tr>
<tr>
<td>GDKD</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>

**Cross-sentence extraction doubles the yield**
## PubMed-Scale Extraction

<table>
<thead>
<tr>
<th>Relations</th>
<th>Single-Sent.</th>
<th>Cross-Sent.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>169,168</td>
<td>332,969</td>
</tr>
<tr>
<td>( p \geq 0.5 )</td>
<td>32,028</td>
<td>64,828</td>
</tr>
<tr>
<td>( p \geq 0.9 )</td>
<td>17,349</td>
<td>32,775</td>
</tr>
<tr>
<td>GDKD</td>
<td>162</td>
<td></td>
</tr>
</tbody>
</table>

Orders of magnitude more knowledge by machine reading
Manual Evaluation

Precision

- Random
- \(P > 0.5\)
- \(P > 0.9\)
Automatic Evaluation

Distant-supervision: Treat labels as gold
Five-fold cross-validation
Balanced dataset → Report average accuracy
Shortest Paths → Features

Accuracy

1 Path | 3 Paths | 10 paths
---|---|---
82 | 86 | 88

Multiple paths help
Other Take-Aways

Prioritizing dependency edges helps
Discourse / coreference no impact yet
The deletion mutation on exon-19 of \textit{EGFR} gene was present in 16 patients, while the \textit{L858E} point mutation on exon-21 was noted in 10. All patients were treated with \textit{gefitinib} and showed a partial response.

\begin{quote}
\textbf{Peng et al.} “Cross-Sentence N-ary Relation Extraction with Graph LSTM”, \textit{TACL-17}.
\end{quote}
Why LSTM?

Cross-sentence → Features become much sparser
N-ary → Want to scale to arbitrary n
Multi-task learning: Easy
Why Graph?

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.
Graph LSTM
Recurrent Neural Network

Contextual Hidden Representation

Word Embedding

W₁  W₂  ......  Wₙ
Recurrent Neural Network

Recurrent Unit

W1 \rightarrow W2 \rightarrow \ldots \rightarrow WN
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

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

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

Forward Pass
Asynchronous Update

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>10,873</td>
<td>57,033</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>1,408</td>
<td>4,279</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>530</td>
<td>1,461</td>
</tr>
<tr>
<td>GDKD + CIVIC</td>
<td></td>
<td>59</td>
</tr>
</tbody>
</table>
## PubMed-Scale Extraction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>10,873</td>
<td>57,033</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>1,408</td>
<td>4,279</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>530</td>
<td>1,461</td>
</tr>
<tr>
<td>GDKD + CIVIC</td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

Cross-sentence extraction triples the yield
# PubMed-Scale Extraction

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>10,873</td>
<td>57,033</td>
</tr>
<tr>
<td>$p \geq 0.5$</td>
<td>1,408</td>
<td>4,279</td>
</tr>
<tr>
<td>$p \geq 0.9$</td>
<td>530</td>
<td>1,461</td>
</tr>
<tr>
<td>GDKD + CIVIC</td>
<td>59</td>
<td></td>
</tr>
</tbody>
</table>

Machine reading extracted orders of magnitudes more knowledge
Manual Evaluation

Precision

Random | $P > 0.5$ | $P > 0.9$

- Random: 0
- $P > 0.5$: 60
- $P > 0.9$: 80
Multi-Task Learning

Leverage related tasks w/ more supervision

E.g., binary sub-relations
Just add top classifiers
## Multi-Task Learning

<table>
<thead>
<tr>
<th></th>
<th>Drug-Gene-Mutation</th>
<th>Drug-Mutation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-Task</td>
<td>80.7</td>
<td>76.7</td>
</tr>
<tr>
<td>Multi-Task</td>
<td>82.1</td>
<td>78.4</td>
</tr>
</tbody>
</table>
System Comparison

- Logistic Regression
- CNN
- Linear LSTM
- Graph LSTM

The diagram shows a comparison of different system performances.
GENIA: Impact of Syntactic Parses

- Logistic Regression
- Linear LSTM
- Graph LSTM
- Graph LSTM (Gold Parse)
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.

Barack Obama born in Honolulu, Honolulu city-of United States.

Michelle Obama spouse of Barack Obama.

Reasoning:
- Barack Obama born in Honolulu
- Honolulu city-of United 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

Statistical relational learning [Getoor & Taskar, 2007]

- Modeling dependencies among the truth values of multiple possible relations

\[ F_1 : (x, \text{parentOf}, z) \land (y, \text{parentOf}, z) \Rightarrow (x, \text{marriedTo}, y) \]

- 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
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

\[ f(s, r, t|\Theta) \]

Used to predict the existence of triples:

\[ y_T \in \{0,1\} \]
Scoring Functions

Bilinear Model [Nickel et al. 2011]

\[ f(\text{Michelle Obama}, \text{lived_in}, \text{Chicago}) \]

Bilinear-diag Model [Yang et al. 2015]

\[ \text{lived_in} \quad \text{Michelle} \quad \text{Chicago} \]

Model E [Riedel et al. 2013]

\[ \text{lived_in}_S \quad \text{Michelle} \quad \text{lived_in}_T \quad \text{Chicago} \]
Scoring Functions

Model F [Riedel et al. 2013]

\[ f(\text{Michelle Obama, lived\_in, Chicago}) = \]

ER-MLP [Dong et al. 2014]

TransE [Bordes et al. 2013]
Loss functions for training model parameters

Learning $\theta$: 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:

$$P(t | s, r) = \frac{ef(s,r,t|\theta)}{\sum_{t' \in \text{Neg}(s,r,?)} ef(s,r,t'|\theta)}$$

$$L(\theta) = \lambda ||\theta||^2 - (\sum_i \log P(t_i | s_i, r_i) + \log P(s_i | r_i, t_i))$$
Loss functions for training model parameters

Learning $\theta$: 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:

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

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

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

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.

[Lao et al. 2012] [Riedel et al. 2013]
Models for graphs including text

- **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 \(t\):

\[
\begin{align*}
\pi_1 & : \text{BORN_IN} \quad \text{CITY_OF} \quad p = 1 \\
\pi_2 & : \text{SPOUSE} \quad \text{NATIONALITY} \quad p = 1
\end{align*}
\]

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

Scoring function: linear in the given feature values

\[
f = w_1 \times 1 + w_2 \times 1
\]
Path Ranking Algorithm [Lao et al. 11]

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

For 3000 relation types:

\[
\begin{align*}
L=1 & \quad 3000 \\
L=2 & \quad 9 \text{ million} \\
L=3 & \quad 27 \text{ billion} \\
L=4 & \quad 81 \text{ trillion}
\end{align*}
\]

Grows exponentially as \(|R|^L\) 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

KITLG activates MAPK1.

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

\[ \pi_1: \text{REGULATION} \uparrow \text{IL2} \quad \text{REGULATION} \uparrow \text{MAPK1} \quad \text{FAMILY} \]

\[ \pi_2: \text{REGULATION} \uparrow \text{KITLG} \quad _{\text{nsubj-activate-dobj}} \text{MAPK1} \quad \text{FAMILY} \]

\[ \pi_3: \text{REGULATION} \uparrow \text{MAPK1} \quad \text{FAMILY} \]
Problems when using relation paths: sparsity → compositional representations

$\pi_1 : \text{REGULATION}^\uparrow \_\text{nsubj-activate-dobj} \text{FAMILY}$

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

Also: RNN [Neelakantam et al. 2015], or sum of vectors [Lin et al. 2015]
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 \textit{exact inference} with all relation paths of bounded length, using dynamic programming.

\textbf{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.

\textbf{No increase in asymptotic complexity}
Results: using compositional representations of relation paths from KB and text relations

Hits@10 on Gene Regulation

NCI-PID database + textual mentions from Pubmed

- Bilinear-diag
- PrunedPaths-100
- All Paths
- All Paths+Nodes

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
- Jianjiang Gao, PhD
- Sarah Philips, PhD
- Hongxin Zhang, MSc
- Ritika Kundra, MSc
- Jiaojiao Wang, MSc
- Ederlinda Paraíso, MPA
- Julia Rudolph, MPA
- David Solit, MD
- Paul Sabbatini, MD
- Nikolaus Schultz, PhD

**Clinical Genomics Annotation Committee**
- Shrutij 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 Kaloy, 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

**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
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 \textit{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 \textit{cetuximab}, so a combination of the two agents is logical to trial.

\begin{verbatim}
V600E    19738166   Clinical
\textbf{Disease type: Colorectal Neoplasms}
\end{verbatim}

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

\begin{verbatim}
V600E    20972475   Clinical
\textbf{Disease type: Unknown}
\end{verbatim}

Also available is a second assay, the \textit{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


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
Ongoing: Cell line experiments on Hanover predictions
Modeling Disease Progression

Wanted: Predict onset, complication, treatment

Electronic medical records (EMRs)

Clinical notes contains rich patient information
Modeling Disease Progression

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 1 AM 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, 02 sat 100 %. Head CT was done and released attenuation within the subcortical white matter of the right medial frontal lobe. LP was performed showing opening pressure 24 cm H20 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

Example: Classifying Heart Failure

Hospitalization: Did heart failure occur?
Supervised learning
Structured + Clinical notes $\rightarrow$ Best accuracy

Example: Learning Patient Embedding

Representation learning: Denoising autoencoder
Evaluation: Predict new disease onset
Outperformed standard dimension reduction
NLP: Negation, family history, entity linking

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

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.

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

Days between submission and release

Release date

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

Enabled GEO to release 400 datasets in a week
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
Key Annotation: Cell Type

Same DNA, different expression, different functions
Crucial for understanding development & cancer
Integrative Studies Remain Small Scale
Grechkin et al. "EZLearn: Extreme Zero-Shot Learning for Unsupervised Data Annotation”. In submission.
Part 7: Resources

Text
Ontology
Databases
Shared tasks
Project Hanover
PubMed
Electronic medical record (EMR)
Clinical trial
Pathology report
PubMed

Abstracts: 27 millions
Full text: 4.3 millions
Open-access: 1.5 million
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

......
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.
Clinical Trial

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...

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

Search for Studies
Example: "Heart attack" AND "Los Angeles"

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

Locations of Recruiting Studies
- Non-U.S. only (56%)
- U.S. only (38%)
- Both U.S. and non-U.S. (5%)

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 (i.e., 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 TNM0C 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 been 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^3
   - Platelets ≥ 100,000/mm^3
   - Hemoglobin (Hgb) ≥ 9 g/dL
Ontology

HUGO
MeSH
DrugBank
UMLS
ICD
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 BiMart 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.
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

Jump to section for this gene:

- Aliases
- Paralogs
- Disorders
- Pathways
- Domains
- Products
- Drugs
- Proteins
- Expression
- Publications
- Function
- Sources
- Genomics
- Summaries
- Localization
- Transcripts
- Orthologs
- Variants
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
**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 mesylate (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**

![Chemical Structure](image)

**Synonyms:**

- 4-(4-METHYL-piperazin-1-ylmethyl)-N-[4-methyl-3-(4-pyridin-3-yI-pyrimidin-2-yIamino)-phenyl]-benzamide
- alpha-[4-Methyl-1-piperazinyl]-3-[(4-(3-pyridyl)-2-pyrimidinyl)amino]-P-toluidide
- Imatinib
- Imatinib Methanesulfonate
- Imatinibum
- STI 571

**External IDs:**

- CGP-57148B / STI-571

**Product Ingredients:**

- **Ingredient** | **UNII** | **CAS** | **InChI Key** | **Details**
  - Imatinib Mesylate | 8A101M48SB | 220127-57-1 | YLMAHDNUQAMNNX-UHFFFAOYSA-N | Details
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
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
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
Chapter II
Neoplasms (C00-D48)

This chapter contains the following blocks:

- C00-C07 Malignant neoplasms
  - C00-C07 Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue

- C08-C79 Malignant neoplasms of unspecified site

- C80-C96 Malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissue

- C97-C99 Malignant neoplasms of ill-defined, secondary and unspecified sites

- C99-D48 Malignant neoplasms of independent (primary) multiple sites

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 pheochromocytoma of adrenal gland should be coded to C74 with additional code E27.5; basophil adenoma of pituitary gland with Cushings syndrome should be coded to D35.2 with additional code E24.6.
Databases

Anything of import → Manual KBs exist
Problem: Unsubstantiable by manual effort
Free lunches abound for machine learning
Welcome to STRING

Protein-Protein Interaction Networks

ORGANISMS | PROTEINS | INTERACTIONS
--- | --- | ---
2031 | 9.6 mio | 1380 mio

SEARCH
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.

Shared Tasks

BioCreative
BioNLP
TREC
I2b2
SemEval
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-aryamonti (BioGrid)
Sun Kim (NCBI)
Don Comeau (NCBI)
Zhiyong Lu (NCBI)

**Downloads**

- PMtask_Triage_TrainingSet
ProMiner: rule-based protein and gene entity recognition.

Hanisch D, Fundel K, Mevissen HT, Zimmer R, Fluck J

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 BioCreAtlV 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 BioCreAtlV 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.
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 be 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.
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.
The activation of Bax by the tumor suppressor protein p53 is known to trigger the p53-mediated apoptosis …
Event Annotation

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

### Diagnosis codes

<table>
<thead>
<tr>
<th>Fake ID</th>
<th>ENTRY_DAT</th>
<th>CODE</th>
</tr>
</thead>
<tbody>
<tr>
<td>34068</td>
<td>5/13/2001</td>
<td>41.85</td>
</tr>
<tr>
<td>37660</td>
<td>8/6/2002</td>
<td>79.99</td>
</tr>
<tr>
<td>140680</td>
<td>8/31/2003</td>
<td>79.99</td>
</tr>
<tr>
<td>23315</td>
<td>5/14/2003</td>
<td>112</td>
</tr>
<tr>
<td>75936</td>
<td>7/9/2004</td>
<td>117.9</td>
</tr>
</tbody>
</table>

### Lab tests

<table>
<thead>
<tr>
<th>Fake ID</th>
<th>TEST</th>
<th>ENTRY_DAT</th>
<th>VALU</th>
</tr>
</thead>
<tbody>
<tr>
<td>3536</td>
<td>pO2</td>
<td>1/23/1996</td>
<td>314</td>
</tr>
<tr>
<td>72021</td>
<td>LDL</td>
<td>2/5/1996</td>
<td>34</td>
</tr>
<tr>
<td>102460</td>
<td>pCC2</td>
<td>1/26/1996</td>
<td>45</td>
</tr>
<tr>
<td>135043</td>
<td>HDL</td>
<td>1/25/1996</td>
<td>35</td>
</tr>
<tr>
<td>135432</td>
<td>MonAb</td>
<td>1/24/1999</td>
<td>0.16</td>
</tr>
</tbody>
</table>
Table 1
Efforts and incentives to leverage clinical data for genomics research

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

“Extracting research-quality phenotypes from electronic health records to support precision medicine”. Wei-Qi Wei and Joshua Denny. Genome Medicine 2015.
**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.
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.
Knowledge
Machine Reading

Can be done manually, need automation to scale
E.g., PubMed search

Reasoning
Predictive Analytics

Can’t be done manually, need automation to enable
E.g., personalize drug combinations

http://hanover.azurewebsites.net
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
Five cows graze on a grass land.

It is fun ...
“Step up to bat and practice dictating complex cases”
Mamlouk & Sonnenberg
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

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)

It is fun ...

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 $\rightarrow$ KB

Predictive analytics: Data + Knowledge $\rightarrow$ Decision

Machine learning: Annotation bottleneck

Many nails for your NLP hammer
References: Distant Supervision


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


References: Complex Semantics

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


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


References: Cross-Sentence Extraction


References: Reasoning (1)


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


Traversing knowledge graphs in vector space. Guu et al. EMNLP 2015.
References: Reasoning (3)


