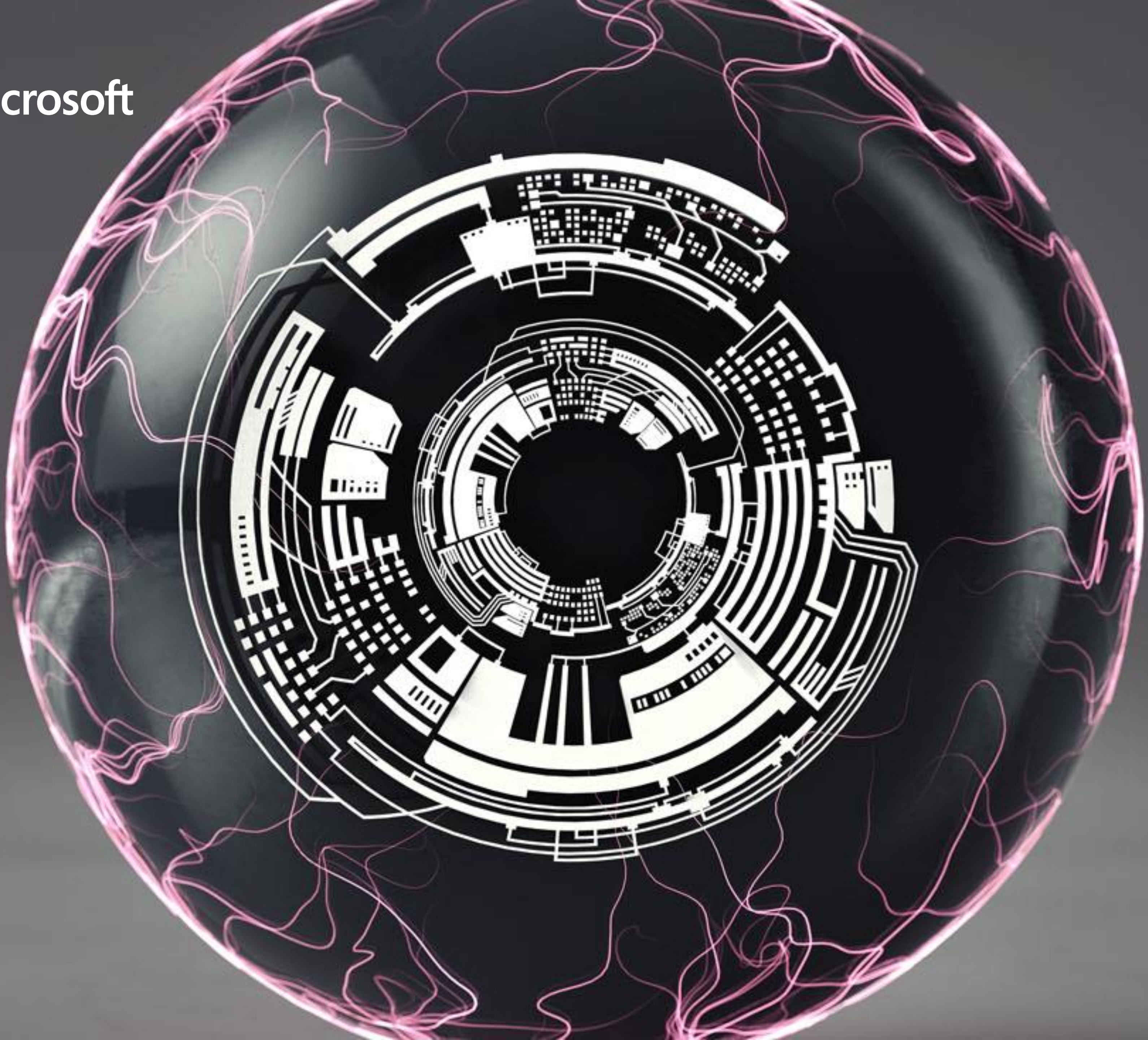
Microsoft



# Microsoft Research
# Faculty Summit
# 2016

# Machine Reading for Precision Medicine

*Hoifung Poon*

# Machine Reading

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

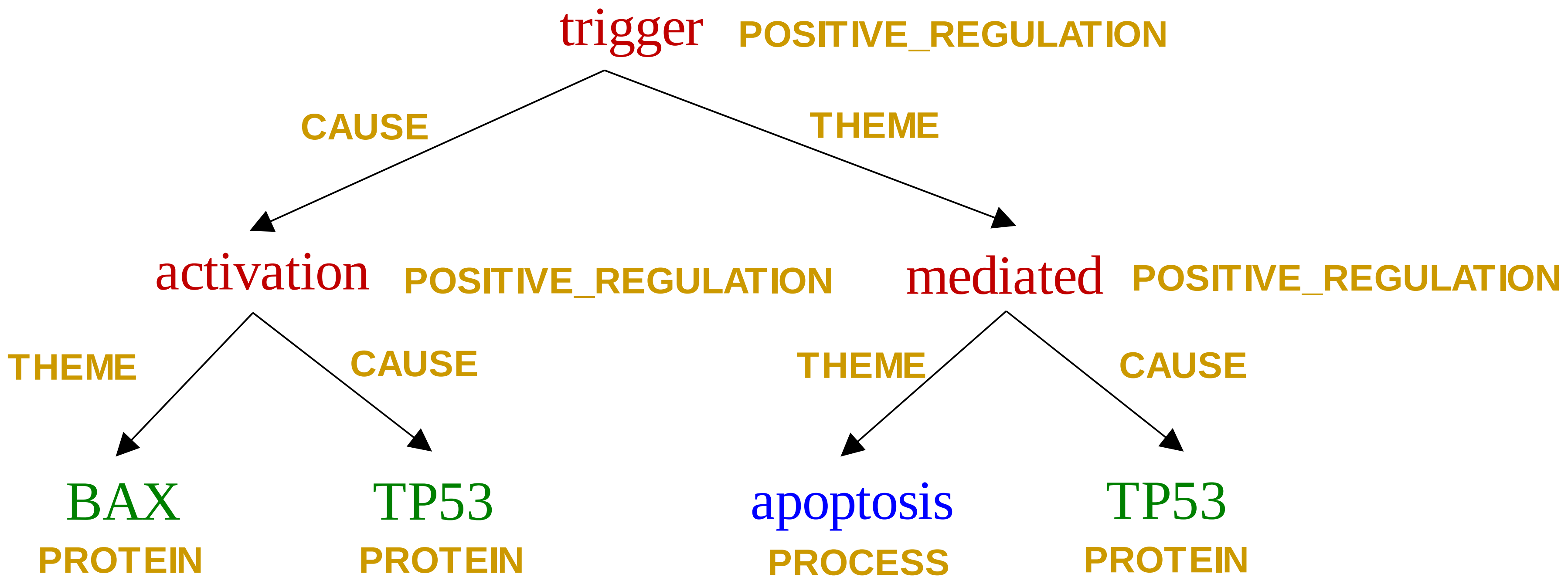
# Machine Reading

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

| BAX | TP53 | apoptosis | TP53 |
|---|---|---|---|
| **PROTEIN** | **PROTEIN** | **PROCESS** | **PROTEIN** |

# Machine Reading

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

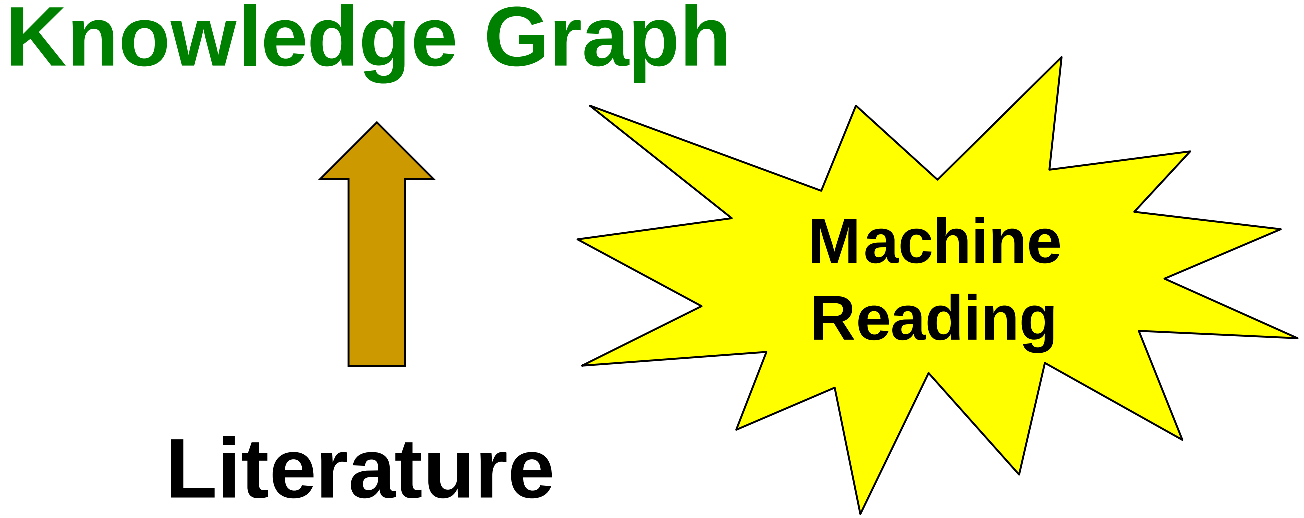


# Machine Reading

- 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

| **Precision Medicine Challenges** | | **Machine Reading Advances** |
|---|---|---|
| Annotation Bottleneck | → | Distant Supervision |
| Complex Knowledge | → | Grounded Semantic Parsing |
| Reasoning | → | Neural Embedding |
| Beyond Sentence Boundary | → | Distant Supervision with Discourse Modeling |

# Literome



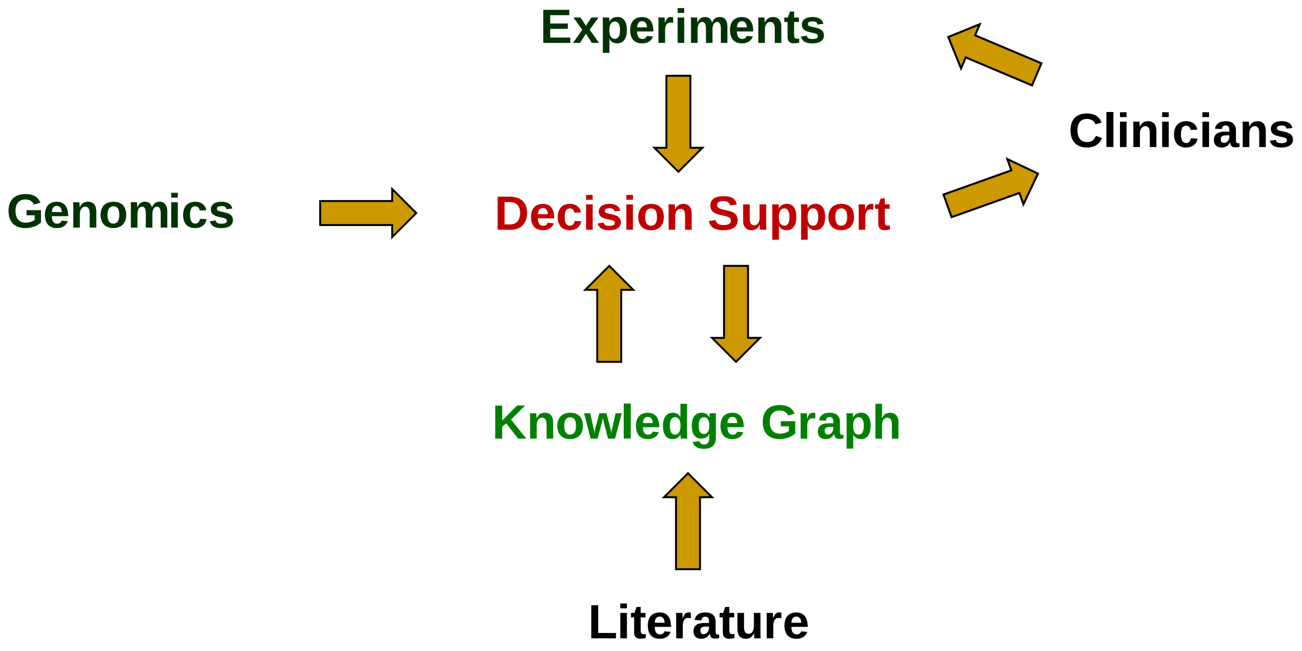
**Knowledge Graph**

**Machine Reading**

**Literature**

Orders of magnitude more extracted knowledge
E.g., 10,000 → 1.5 million gene regulations

# Cancer Precision Medicine



**Experiments**

**Clinicians**

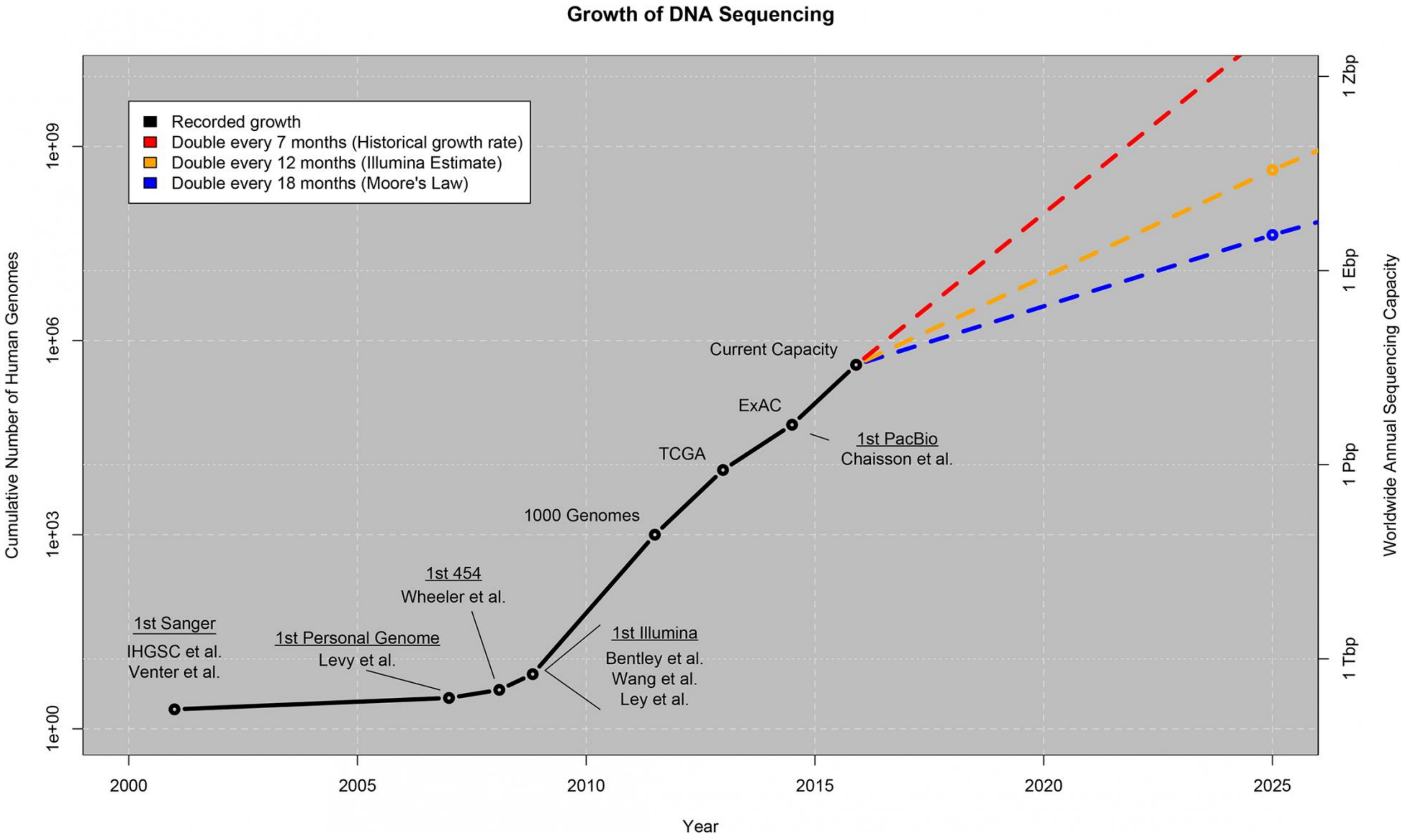
**Genomics**

**Decision Support**

**Knowledge Graph**

**Literature**

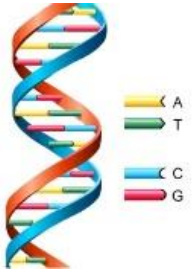
Literome doubles accuracy in cancer drug prediction

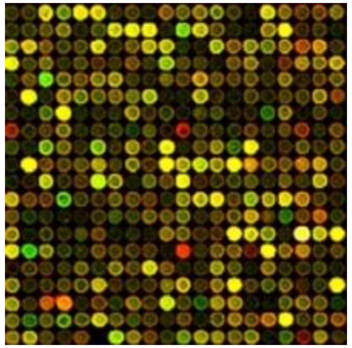


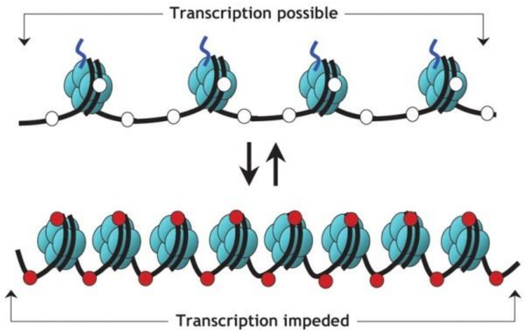
The Information Within Ourselves, Coughlin, Forbes 2015

# Panomics

… ATTCGG**A**TATTTAAG**G**C …



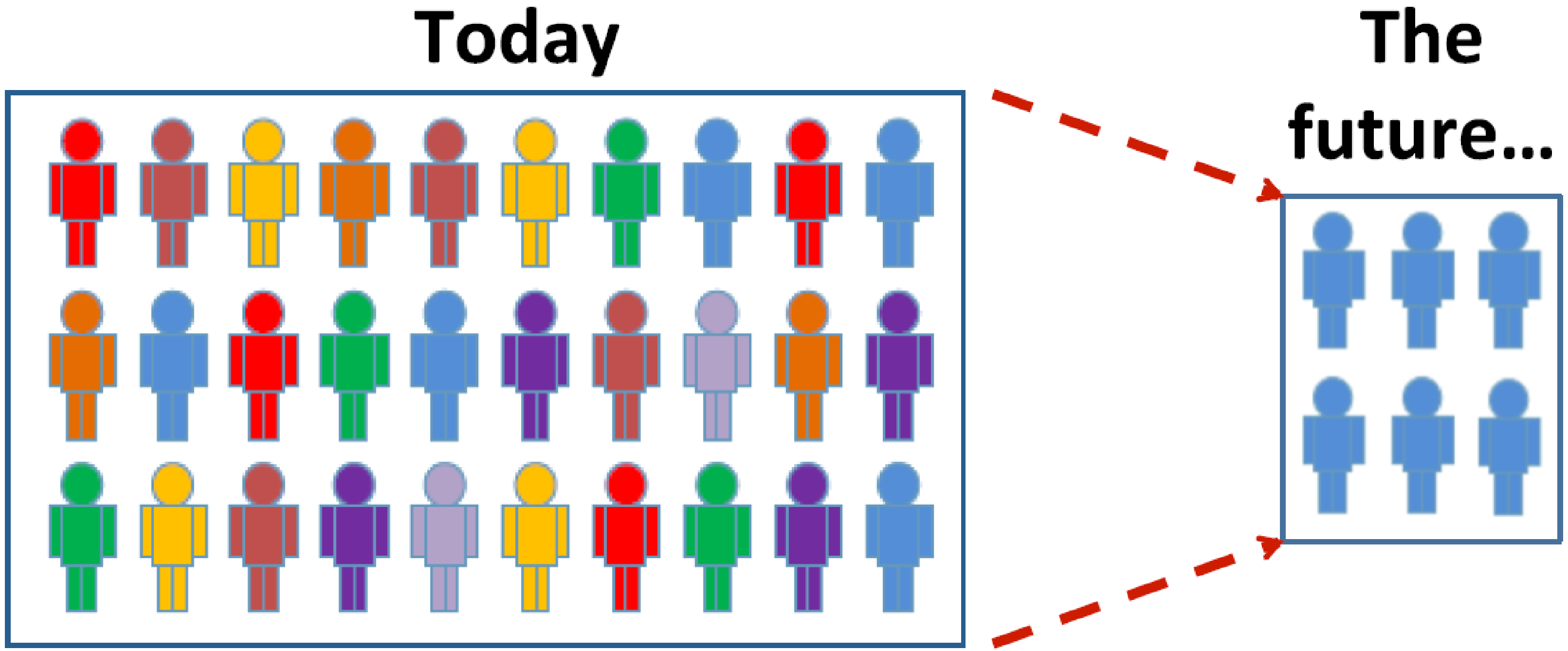




**Genome** **Transcriptome** **Epigenome**

……

# Precision Medicine



**Today**

**The future…**

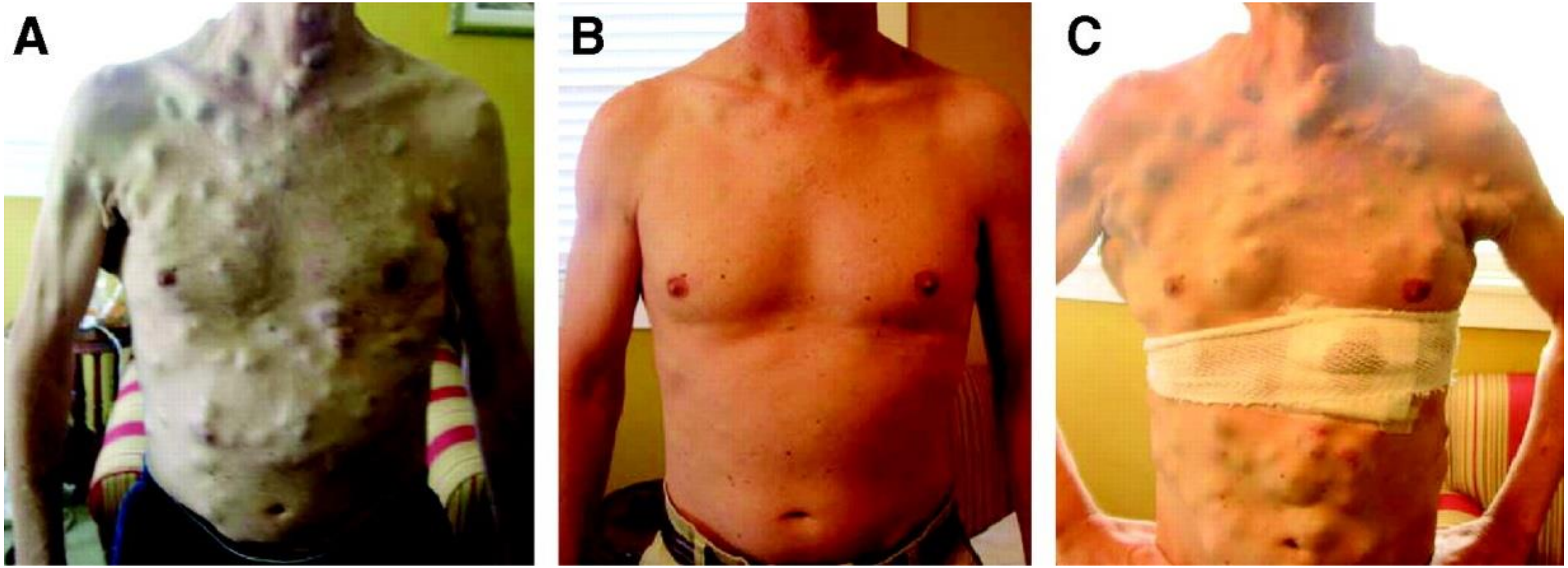
# Cancer

- **Caused by genetic mutations**
- 22.5% deaths: Top killer for age under 85
- Expected in 2030:
  - Increase by 45%
  - 2.3 million new cases annually
  - No. 1 killer for all

# Vemurafenib on BRAF-V600 Melanoma

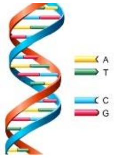
**Before Treatment** **15 Weeks**

# Vemurafenib on BRAF-V600 Melanoma



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

# Precision Medicine



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

**High-Throughput Measurements**

**Discovery**

Bottleneck #1: Knowledge

Bottleneck #2: Reasoning

# Example: Molecular Tumor Board

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

# Example: Molecular Tumor Board

- Problem: Hard to scale

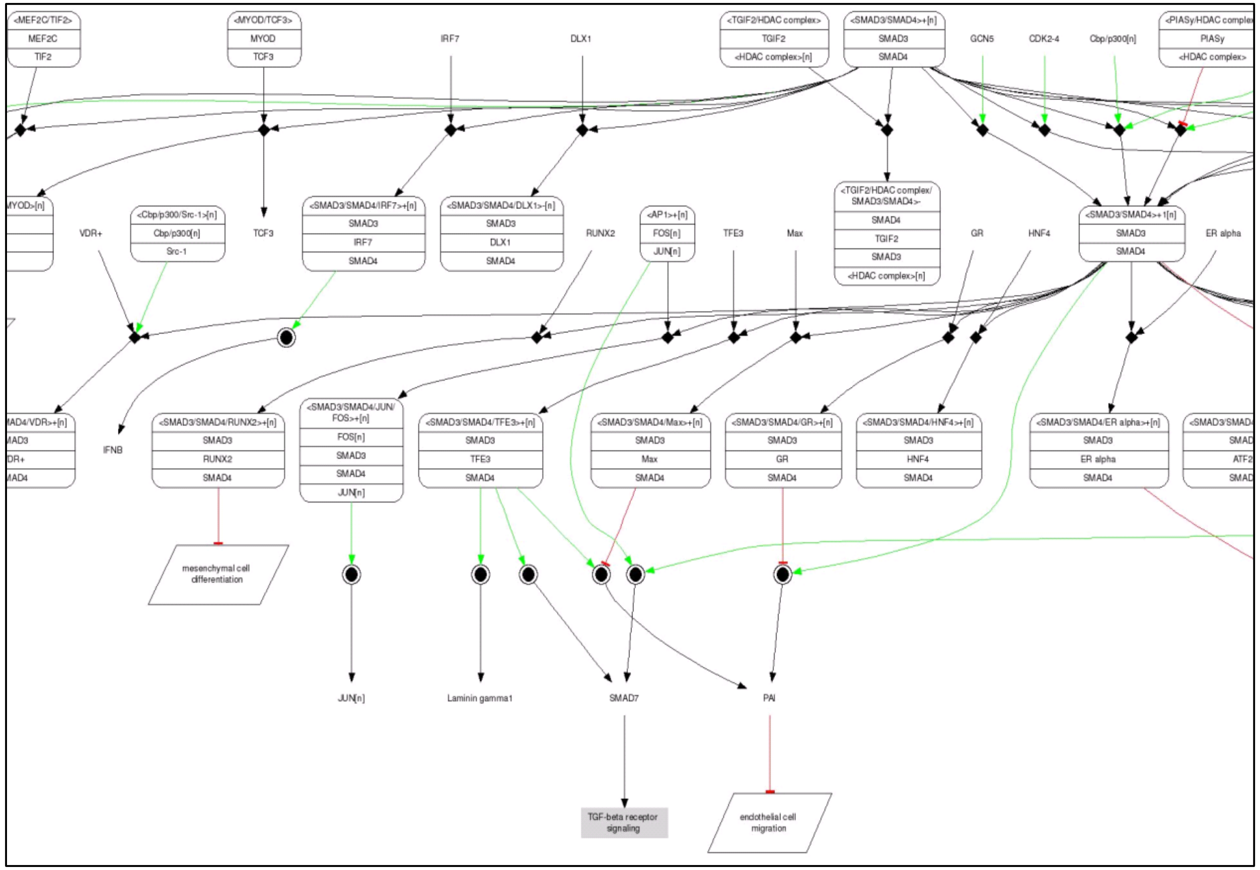
  U.S. 2015: 1.6 million new cases, 600K deaths

- 902 cancer hospitals
- Memorial Sloan Kettering 2016:
  - Sequence ~ 60,000
  - Board can review ~ 200

**Wanted: Decision support for cancer precision medicine**

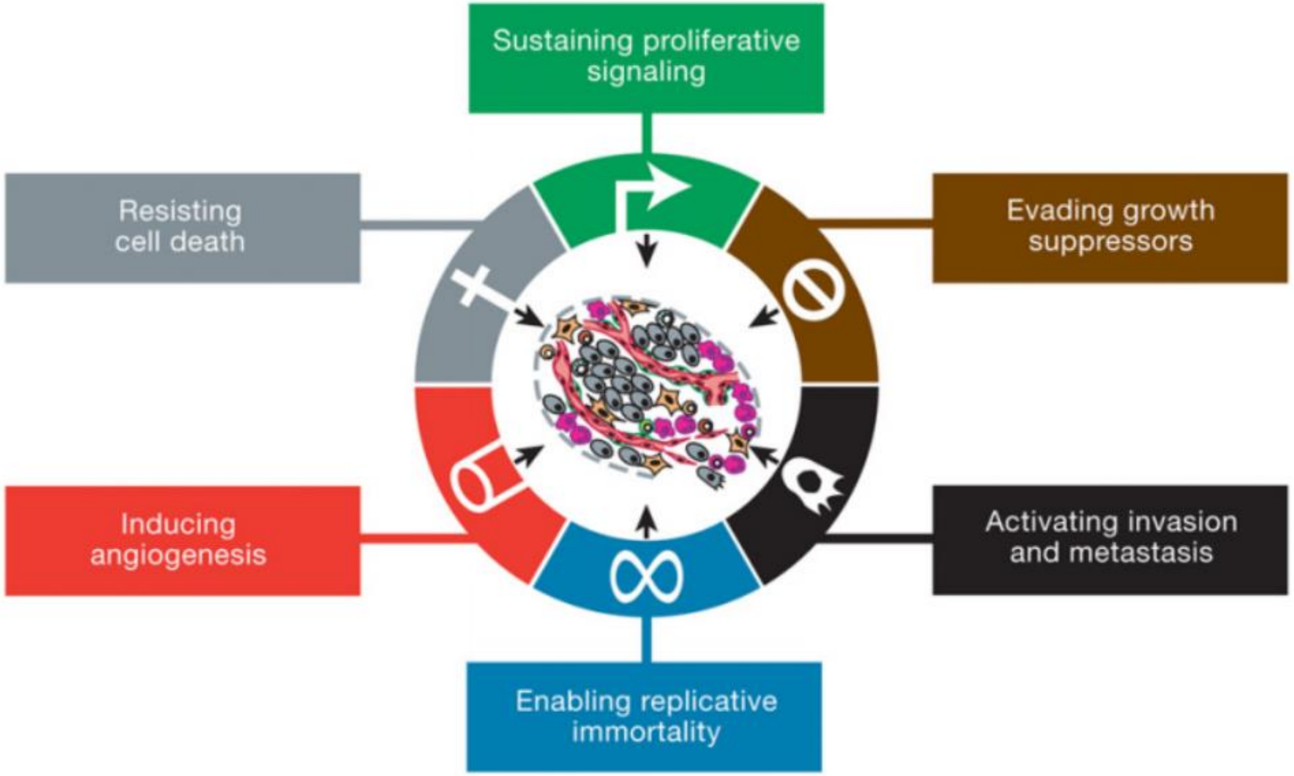
# Network Knowledge

## Genes work synergistically in pathways



# Why Hard to Identify Drivers?

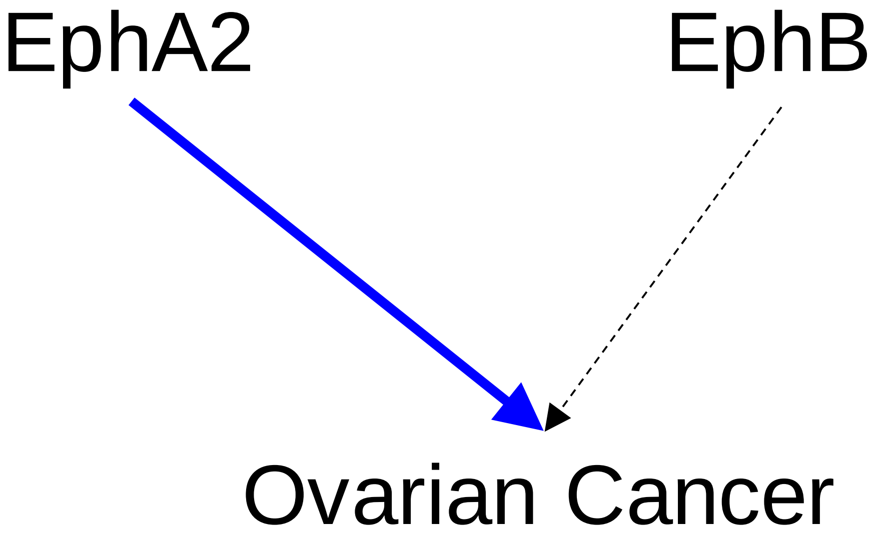
Complex diseases ← Perturb multiple pathways



Hanahan & Weinberg [Cell 2011]

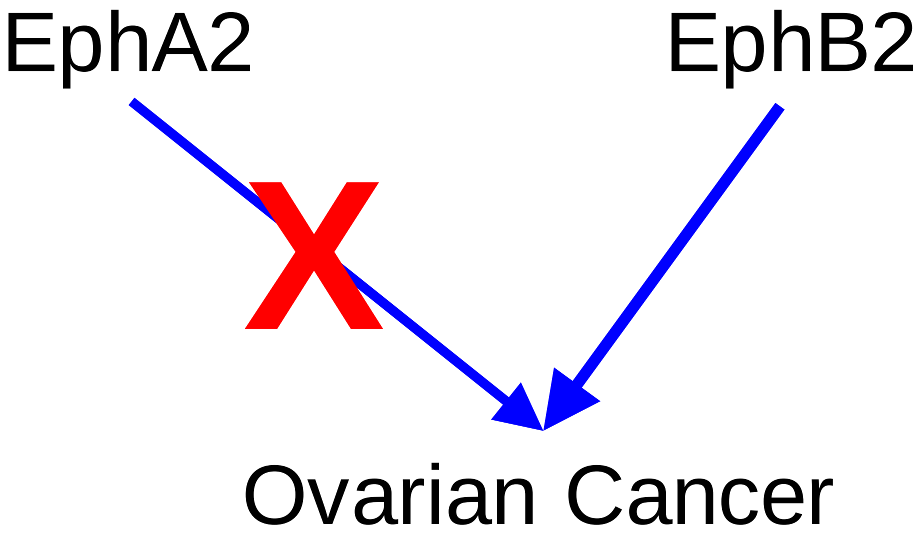
# Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated



# Why Cancer Comes Back?

- Subtypes with alternative pathway profile
- Compensatory pathways can be activated

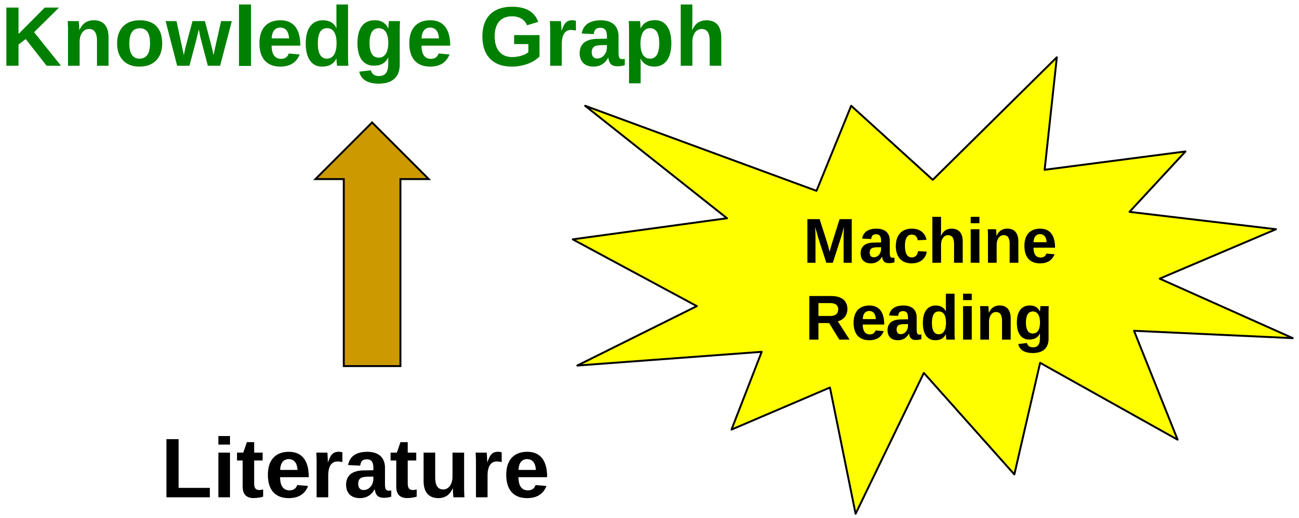


EphA2

EphB2

Ovarian Cancer

# Literome



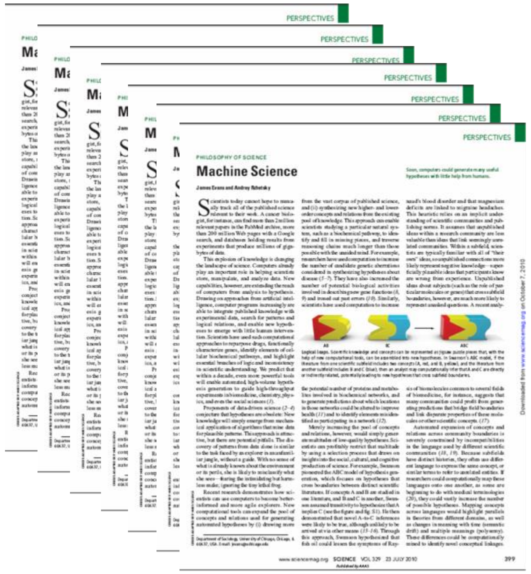
Knowledge Graph

Machine Reading

Literature

# PubMed

- 24 millions abstracts
- Two new abstracts every minute
- Adds over one million every year



PERSPECTIVES

PHILOSOPHY OF SCIENCE

Machine Science

James Evans and Andrey Rzhetsky

# Machine Reading

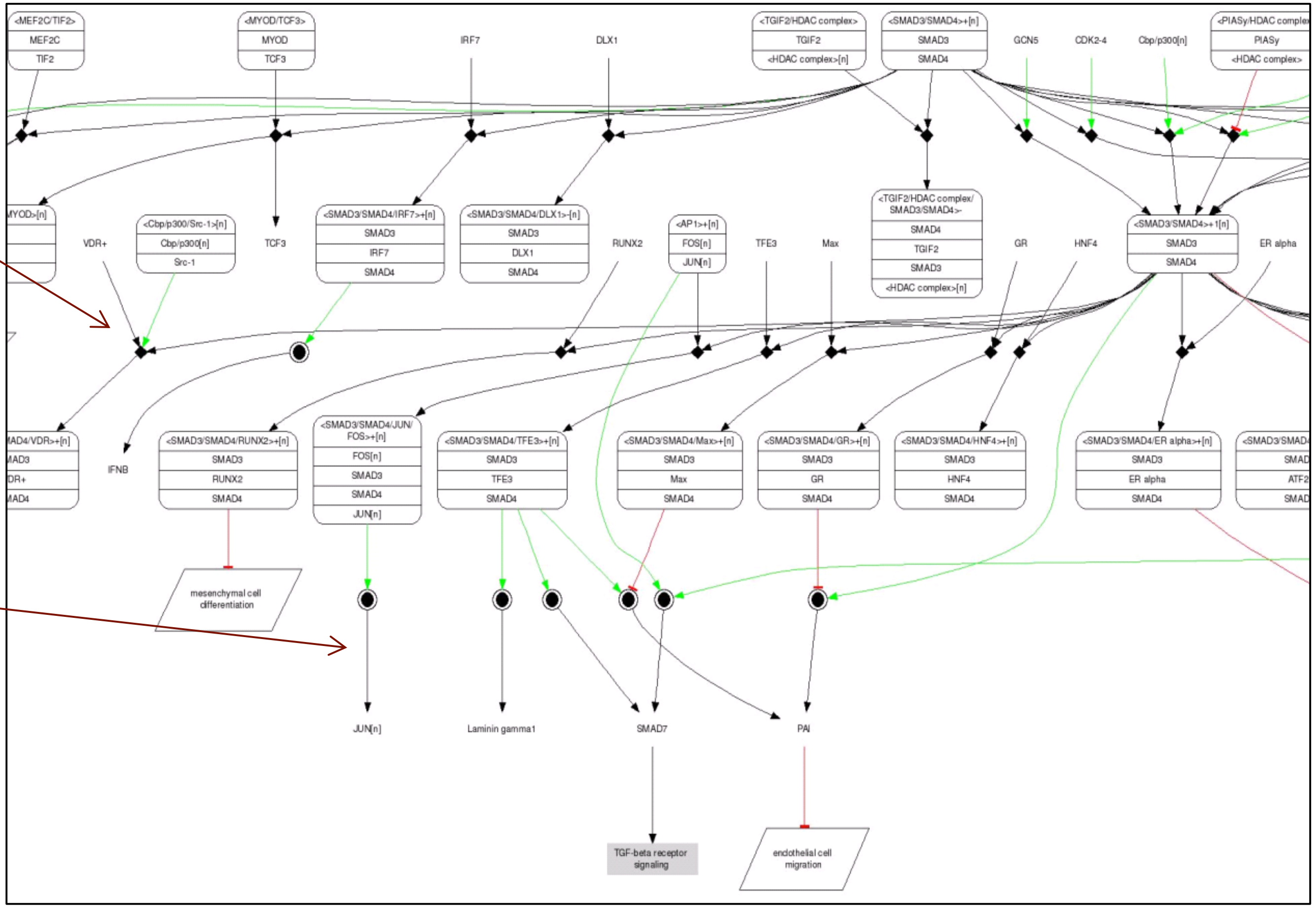
**PMID: 123**

…
VDR+ binds to SMAD3 to form
…

**PMID: 456**

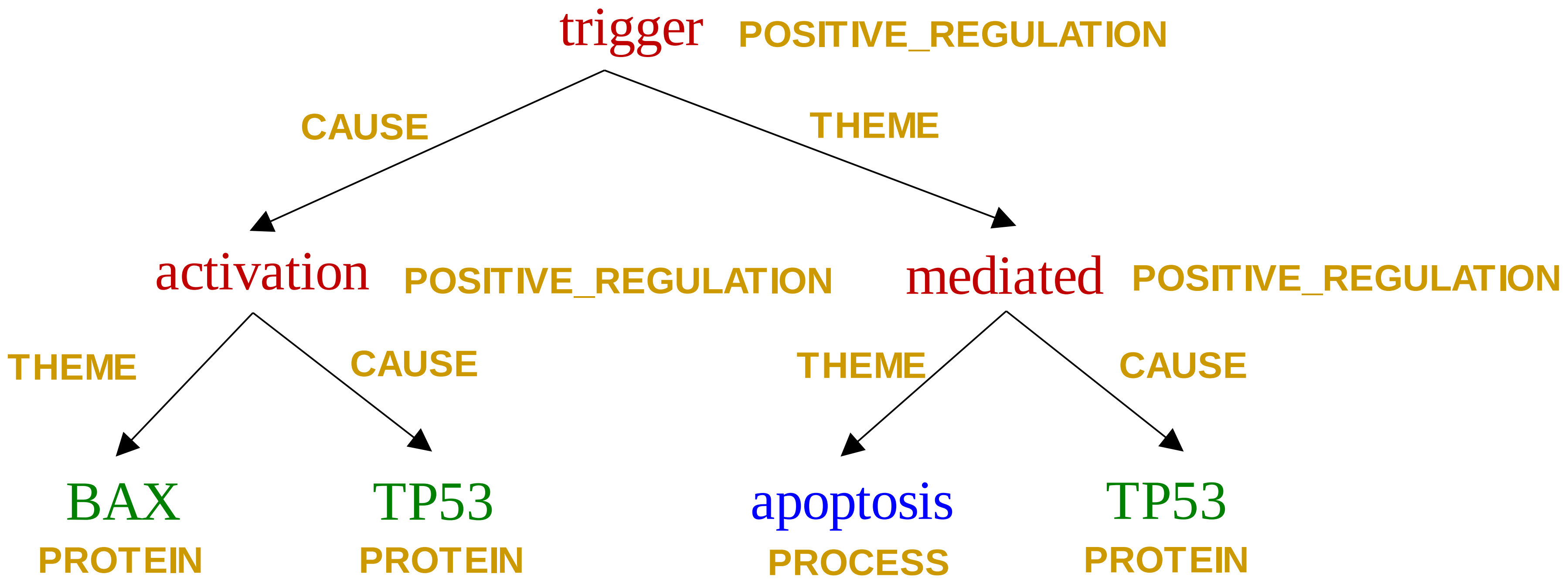
…
JUN expression is induced by SMAD3/4
…

⋮



# Machine Reading

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



# 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 ……

# 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
```

# Bottleneck: Annotated Examples

- Challenge for supervised learning
- Hard to crowd-source
- Complex semantics
- Can we breach this bottleneck?

# Free Lunch: Indirect Supervision

Poon & Domingos, “Unsupervised Semantic Parsing”. *EMNLP-09*. **Best Paper Award**.

Poon, Toutanova, Quirk, “Distant Supervision for Cancer Pathway Extraction from Text”. *PSB-15*.

Parikh, Poon, Toutanova. “Grounded Semantic Parsing for Complex Knowledge Extraction”, *NAACL-15*.

# Free Lunch: Existing KB

**NCI-PID**
**Pathway KB**

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

# Free Lunch: Existing KB

**NCI-PID Pathway KB**

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

# Free Lunch: Existing KB

**NCI-PID 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-PID 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-PID Pathway KB**

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

*TP53 inhibits BCL2.*

*Tumor suppressor P53 dow* [illegible] *s.*

*BCL2 transcription is suppre* [illegible]

*The inhibition of B-cell CLL/Lymph* [illegible] *2* [illegible] *ssi* [illegible] *TP53* [illegible]

*……*

**Distant Supervision**

# PubMed-Scale Extraction

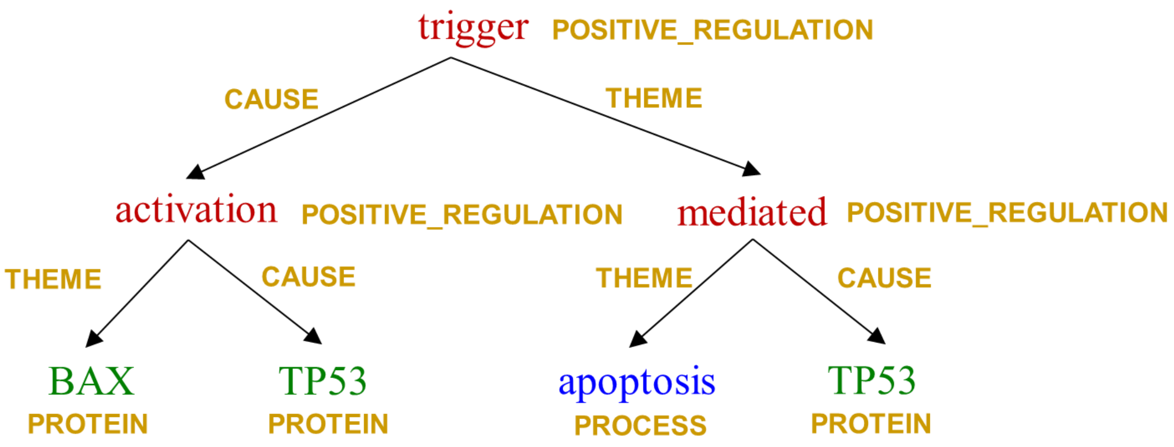
- Preliminary pass:
  - 3 million instances
  - 15,000 genes, 1.5 million unique regulations
- Applications:
  - UCSC Genome Browser, MSR Interactions Track
  - Validate *de novo* pathway prediction
  - Breast cancer study
  - Etc.

Poon, Toutanova, Quirk, “Distant Supervision for Cancer Pathway Extraction from Text”. *PSB-15*.

# Extract Complex Knowledge

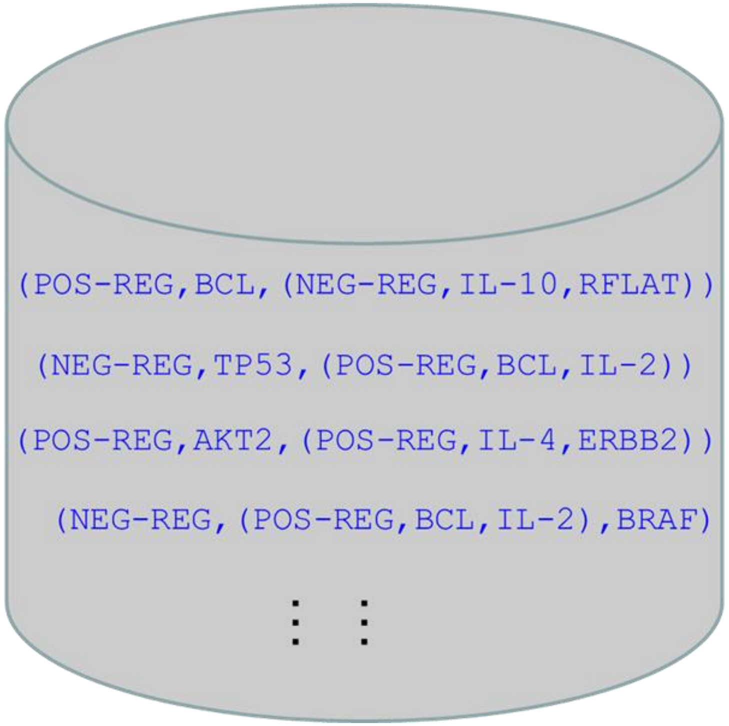
**Outperform 19 out of 24 supervised participants in GENIA Shared Task**

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



Parikh, Poon, Toutanova. “Grounded Semantic Parsing for Complex Knowledge Extraction”, *NAACL-15*.

# Tree HMM



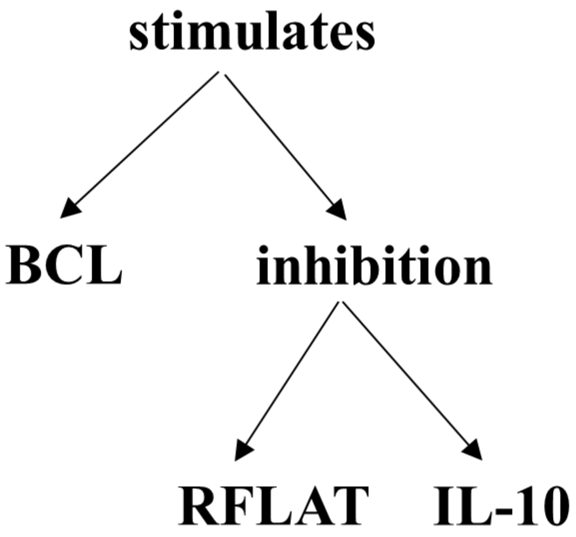
(POS-REG,BCL,(NEG-REG,IL-10,RFLAT))

(NEG-REG,TP53,(POS-REG,BCL,IL-2))

(POS-REG,AKT2,(POS-REG,IL-4,ERBB2))

(NEG-REG,(POS-REG,BCL,IL-2),BRAF)

⋮ ⋮



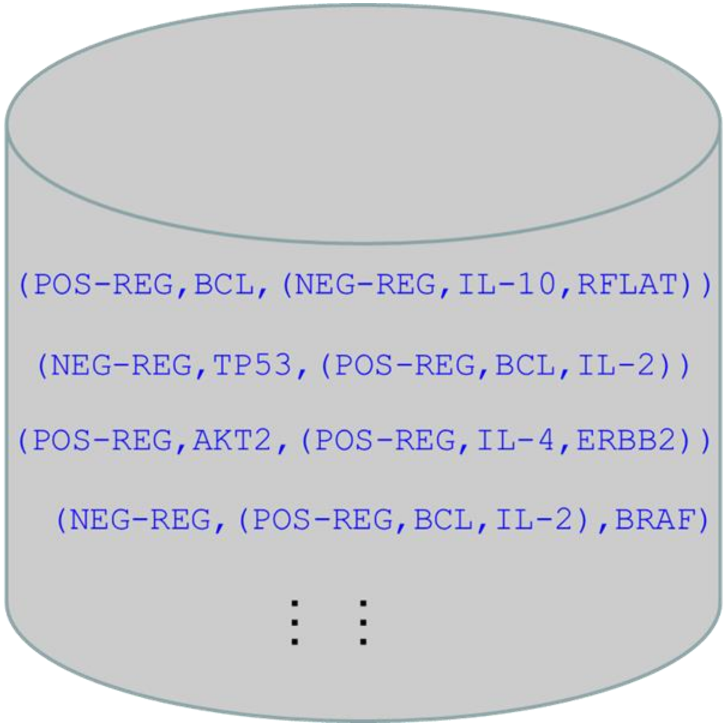
stimulates

BCL inhibition

RFLAT IL-10

*BCL stimulates inhibition of RFLAT by IL-10.*

# Tree HMM



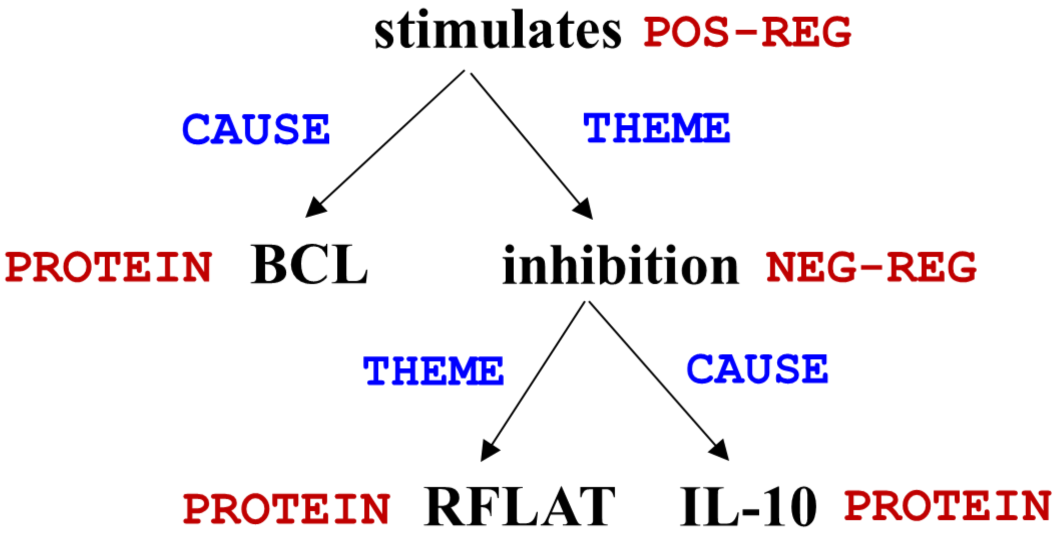
(POS-REG,BCL,(NEG-REG,IL-10,RFLAT))

(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.*

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

# Expectation Maximization

$$\begin{aligned}\theta^* &= \arg\max_{\theta} \quad \log P_\theta(T|K) \\ &= \arg\max_{\theta} \quad \sum_{t \in T} \log \sum_{z} P_\theta(z, t) \cdot \phi_K(z)\end{aligned}$$

**Virtual Evidence**

# Reasoning

- Extraction: Limited to explicitly stated facts
- Wanted: Reasoning
- Challenge: Inference is expensive

# Neural Embedding

- Entity / Relation → $(v_1, v_2, \ldots, v_k)$
- Relation candidate $<\text{subj, rel, obj}>$ → Score
- Distant sup.: Known relations score higher
- Increased recall by 20 points on held-out

Toutanova *et al.*, “Representing Text for Joint Embedding of Text and Knowledge Bases”, *EMNLP-15*.

Toutanova *et al.*, “Compositional Learning of Embeddings for Relation Paths in Knowledge Bases and Text”, *ACL-16*.

# Drug-Gene Interaction

- Sensitivity or resistance to drugs
  - ALK (L1196M): Resistant to crizotinib
  - ALK (R1275Q): Sensitive to crizotinib
- State of the art: GDKD
  - Oncologists read reviews for years
  - 162 drug-gene interactions
- Can we automate extraction?

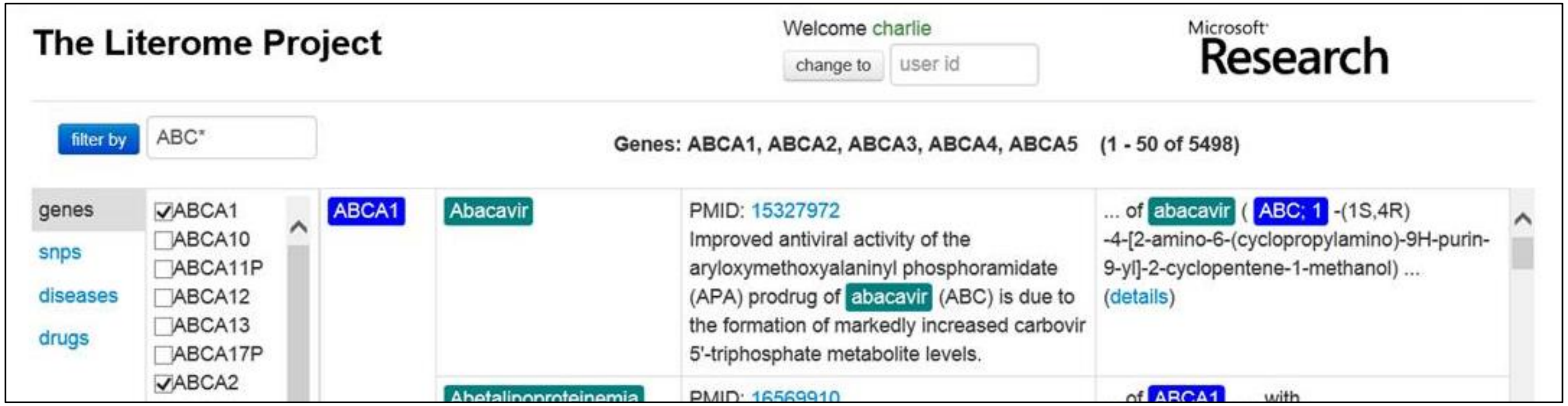
# Drug-Gene Interaction

- Problem: Need discourse modeling
  - Usually appear in full text
  - Often span multiple sentences

*We studied the effect of crizotinib on lung cancer cells.*
*Cells with ALK variant L1196M are not sensitive.*
*However, we observed a strong response on cells with variant R1275Q.*

- Distant supervision w. discourse modeling
- Orders of magnitude increase: 162 → 79,952

# Literome: Genomic KB on Azure



Genes: ABCA1, ABCA2, ABCA3, ABCA4, ABCA5 (1 - 50 of 5498)

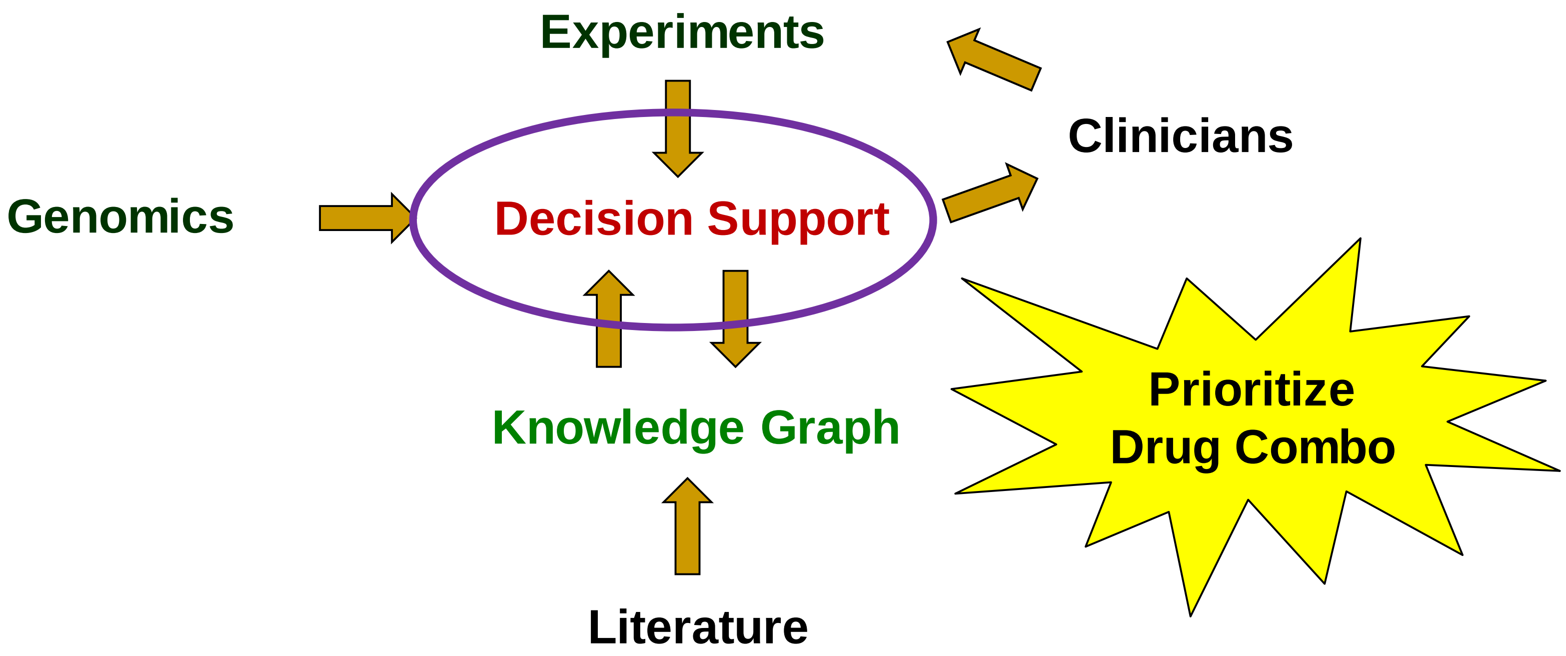
- ☑ ABCA1
- ☐ ABCA10
- ☐ ABCA11P
- ☐ ABCA12
- ☐ ABCA13
- ☐ ABCA17P
- ☑ ABCA2

| | | | |
|---|---|---|---|
| ABCA1 | Abacavir | PMID: 15327972<br>Improved antiviral activity of the aryloxymethoxyalaninyl phosphoramidate (APA) prodrug of abacavir (ABC) is due to the formation of markedly increased carbovir 5'-triphosphate metabolite levels. | ... of abacavir ( ABC; 1 -(1S,4R) -4-[2-amino-6-(cyclopropylamino)-9H-purin-9-yl]-2-cyclopentene-1-methanol) ... (details) |
| | Abetalipoproteinemia | PMID: 16569910 | ... of ABCA1 ... with |

Poon *et al.*, “Literome: PubMed-Scale Genomic Knowledge Base in the Cloud”, *Bioinformatics-14*.

http://literome.azurewebsites.net

# Cancer Precision Medicine



Experiments

Clinicians

Genomics

Decision Support

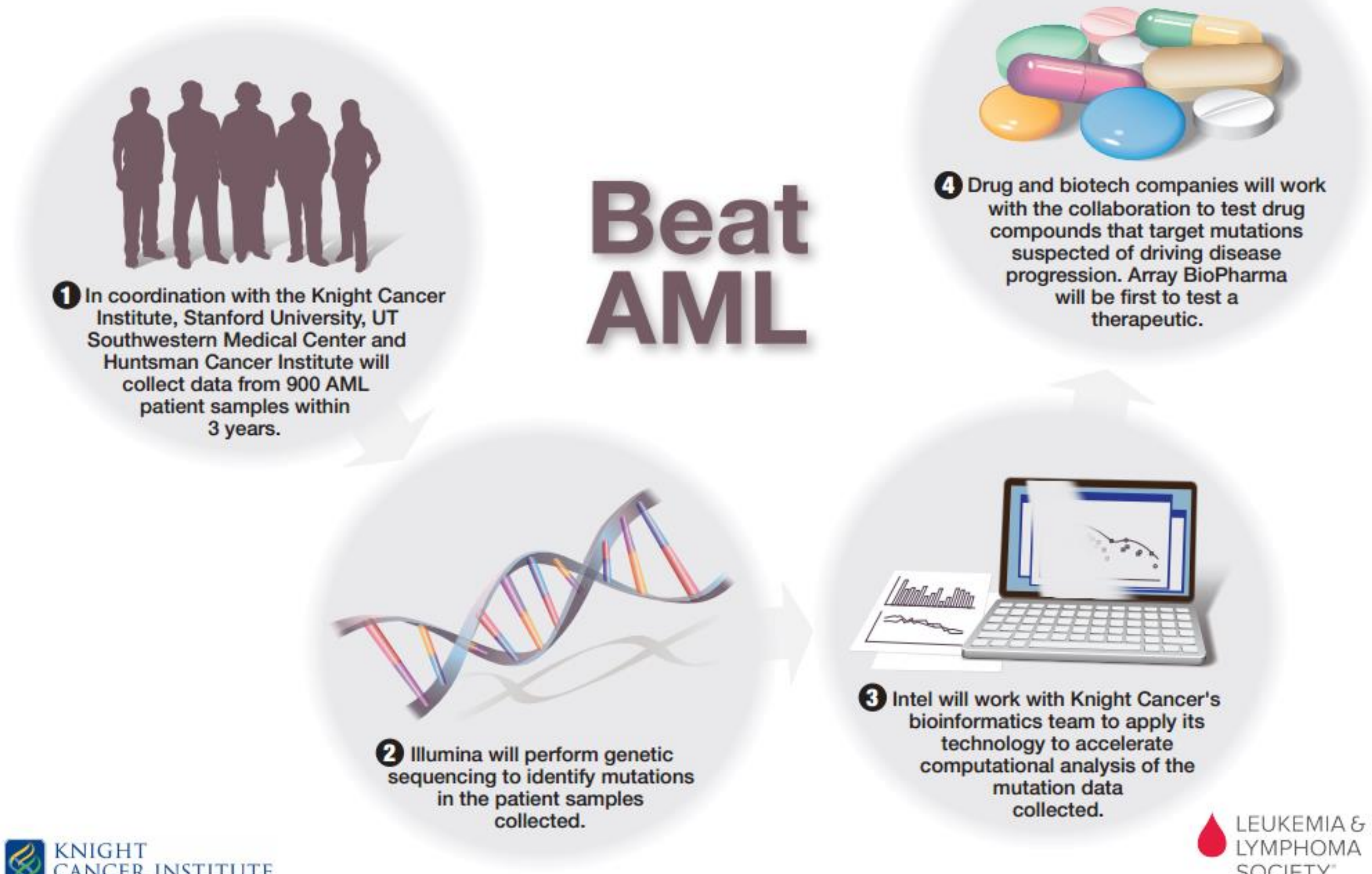
Knowledge Graph

**Prioritize Drug Combo**

Literature

# 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:



**Beat AML**

❶ 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.

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

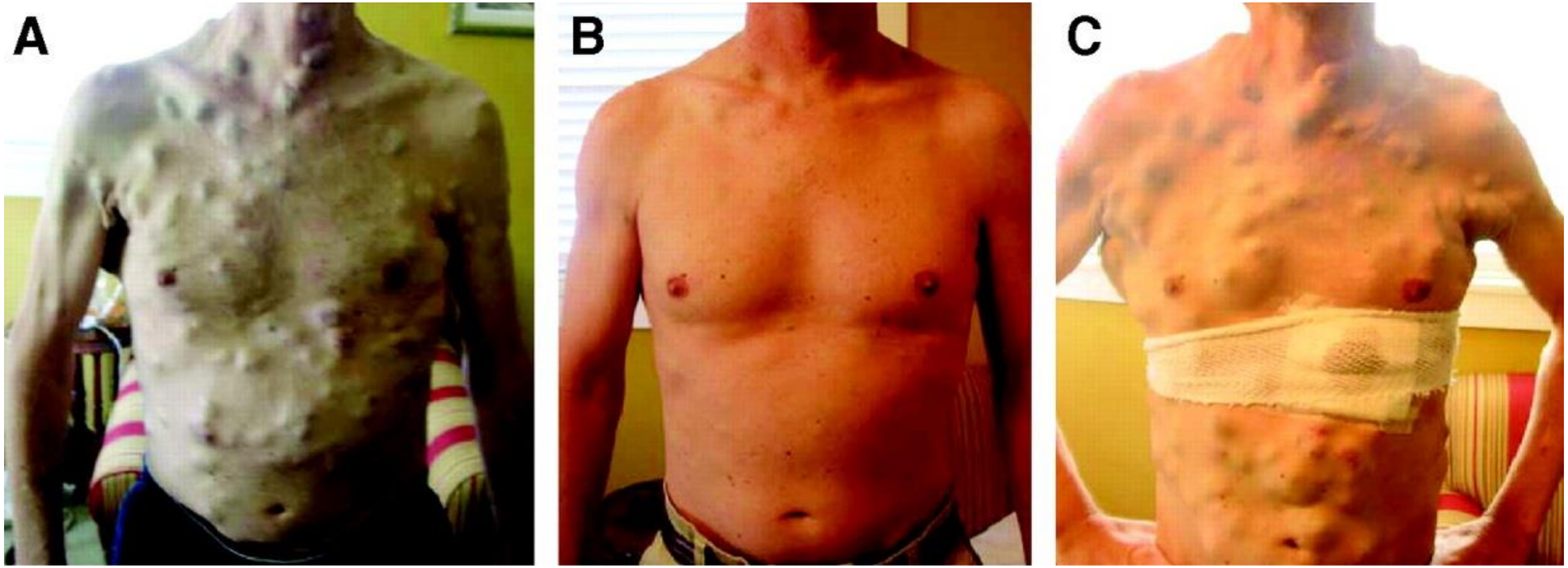
❸ Intel will work with Knight Cancer's bioinformatics team to apply its technology to accelerate computational analysis of the mutation data collected.

❹ Drug and biotech companies will work with the collaboration to test drug compounds that target mutations suspected of driving disease progression. Array BioPharma will be first to test a therapeutic.

KNIGHT
CANCER INSTITUTE
Oregon Health & Science University

LEUKEMIA &
LYMPHOMA
SOCIETY®
fighting blood cancers

# Vemurafenib on BRAF-V600 Melanoma



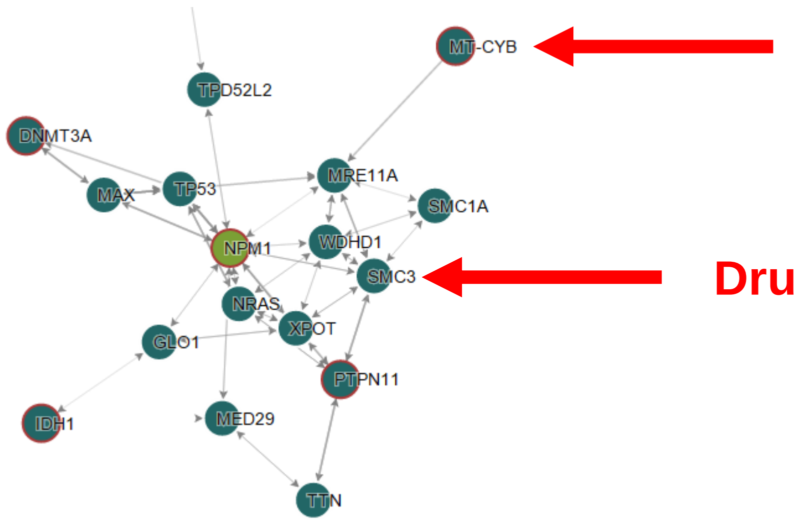
**Before Treatment** **15 Weeks** **23 Weeks**

# 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 *in silico*

# 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 *in silico*



**Drug 1**

**Drug 2**

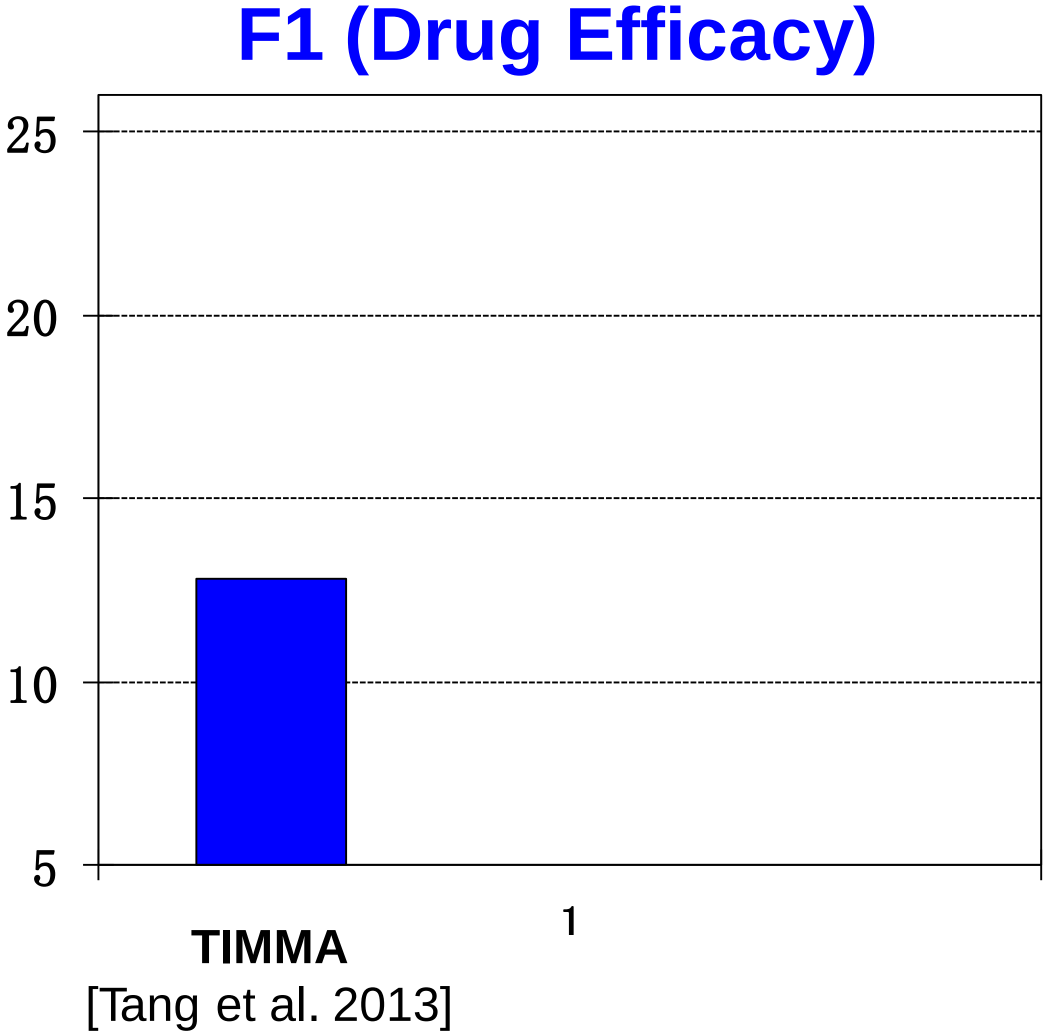
# Related Work

- Drug → Rank patients
- Drug / target / patient → Find similar ones
- Our focus:
  - Patient → Rank drugs
  - Target combination
  - Network knowledge

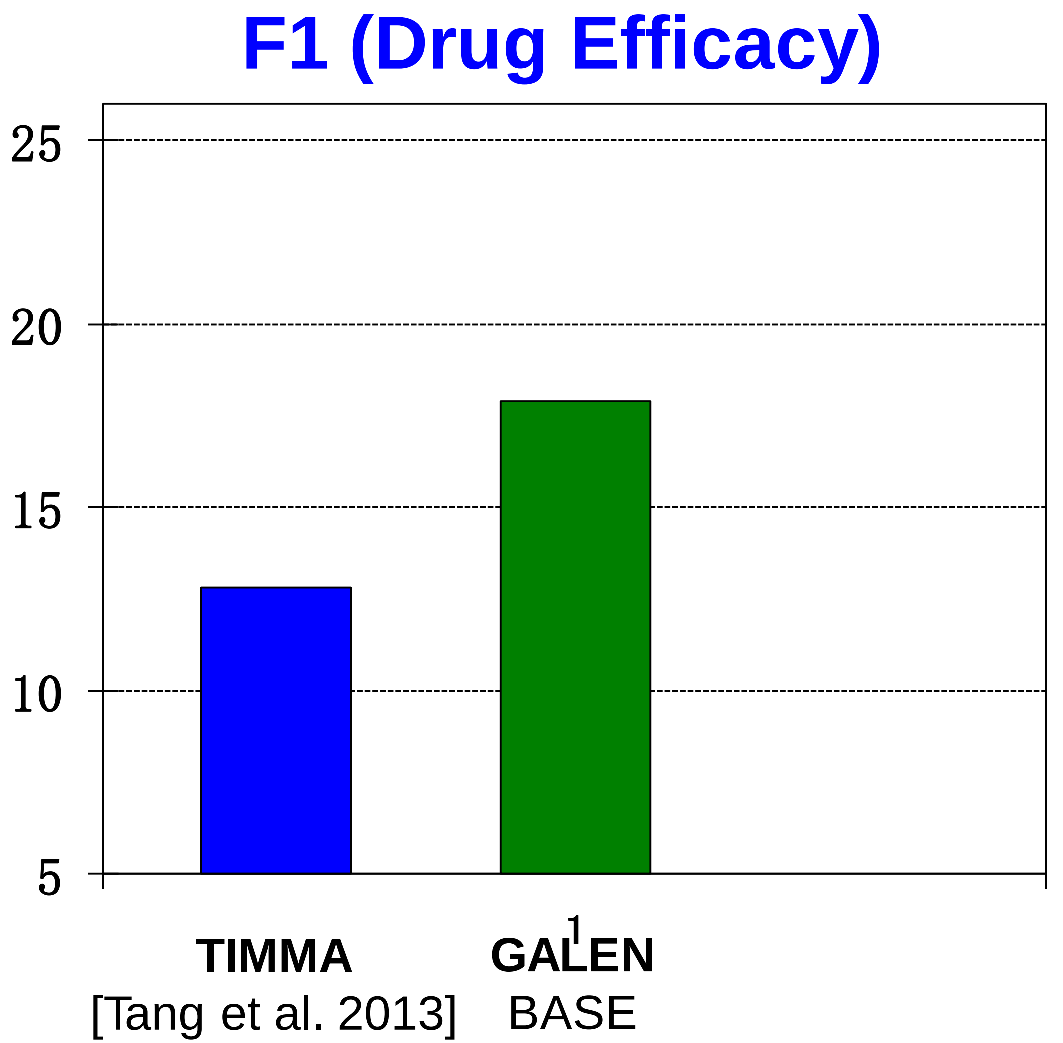
# Machine Learning

- **Prediction:** Classify top efficacy
- **Features**
  - Panomics: Gene expression, …
  - Pharmacology: Drug targets
  - Network knowledge: TP53 inhibits BCL2, …
- Start with single-drug experiments

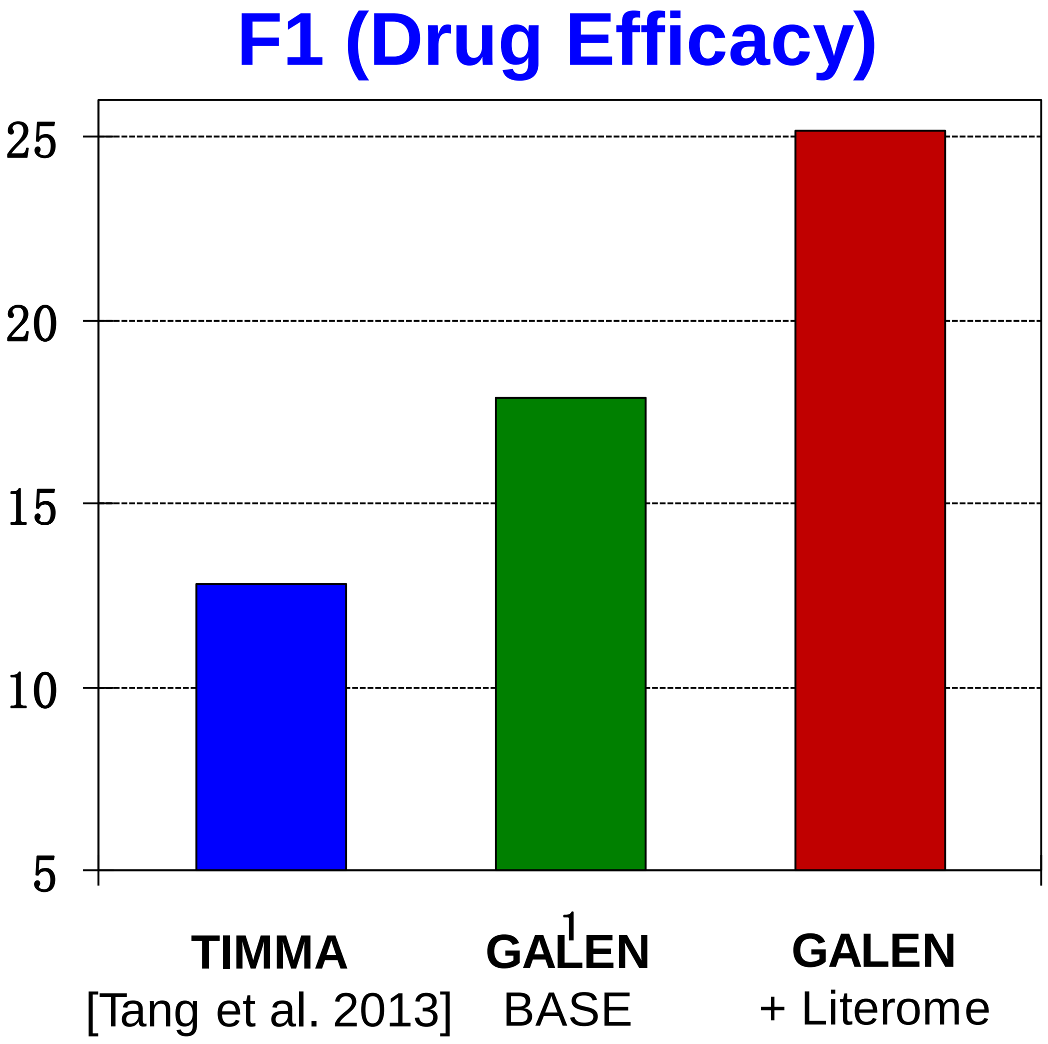
# Cancer Cell Line Encyclopedia



# Cancer Cell Line Encyclopedia



# Cancer Cell Line Encyclopedia



# Next: BeatAML

- Goal: Identify promising drug combos
- Symbiosis: Oncologist + Computer
  - Interpretable model
  - Learn from feedback
  - Exploitation + Exploration

# Big Mechanism

- 42-million program
  - Reading, Assembly, Explanation
  - Domain: Cancer signaling pathways
- We are in
  - PI: Andrey Rzhetsky
  - Co-PI w. James Evans, Ross King

# Emerging Frontier:
# Electronic Medical Record (EMR)

**Cost to develop a drug > $2.5 billion**

**Virtual clinical trial: “EMR as a lab”**

**Disruption: Pay-by-quality**

**Patientomics**

000000011 Age: 62; Gender: Female
Week 27: 250.00, 272.4, 739.3
Week 32: 250.00, 715.90
Week 33: 250.00, 715.95
…….

**Predict Progression**

Patient was admitted with hypernatremia and acute mental status changes and right sided paralysis. Her corrected sodium on admission was about 170s and her free water deficit was 5.5L. She was volume repleted with normal saline. She also recieved D51/4 NS for free water repletion initially and this was changed to free water boluses through nasogastric tube. Her sodium gradually trended down with free water repletion. She also was in pre-renal renal failure and her creatinine trended down with hydration. With regards to the acute mental changes, this is partially explained by the hypernatrmic state. ……

**Machine Reading**

# Summary

- Genomics disruption: Precision medicine
- Challenge: Knowledge + Reasoning
- Machine reading: Literome
- Prioritize cancer drug combos

# Collaborators

- **Knight:** Brian Drucker, Jeff Tyner
- **U. Chicago:** Andrey Rzhetsky
- **Wisconsin:** Mark Craven, Anthony Gitter
- **UCSC:** Max Haeussler, David Haussler
- **Microsoft Research:** Chris Quirk, Kristina Toutanova, David Heckerman, Scott Yih, Lucy Vanderwende, Bill Bolosky, Ravi Pandya
- **Interns**: Ankur Parikh, Maxim Grechkin, Victoria Lin, Sheng Wang, Stephen Mayhew