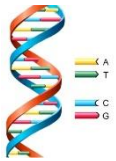
Microsoft Research

Machine Reading for Cancer Panomics

Hoifung Poon

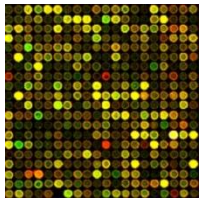
Overview



… ATTCGGATATTTAAGGC …

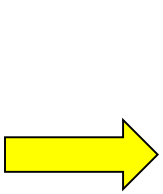
… ATTCGGGTATTTAAGCC …

……



……

High-Throughput Data



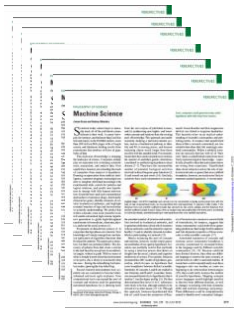
KB

Disease Genes

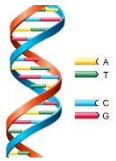
Drug Targets

……

Cancer Systems Modeling



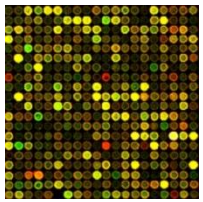
Overview



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

……



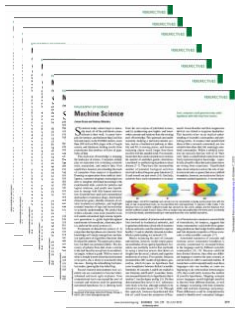
……

High-Throughput Data

Disease Genes

Drug Targets

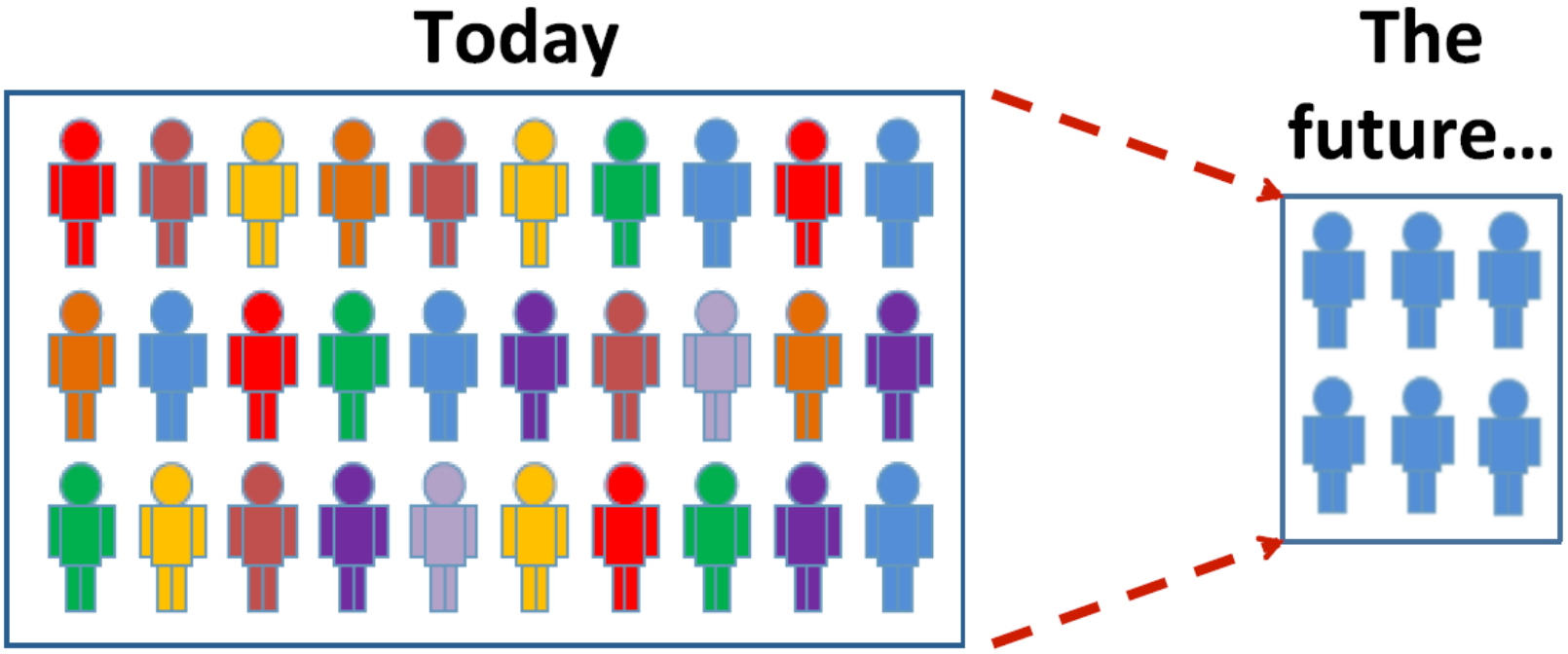
…



Grounded Semantic Parsing

Extract Pathways from PubMed

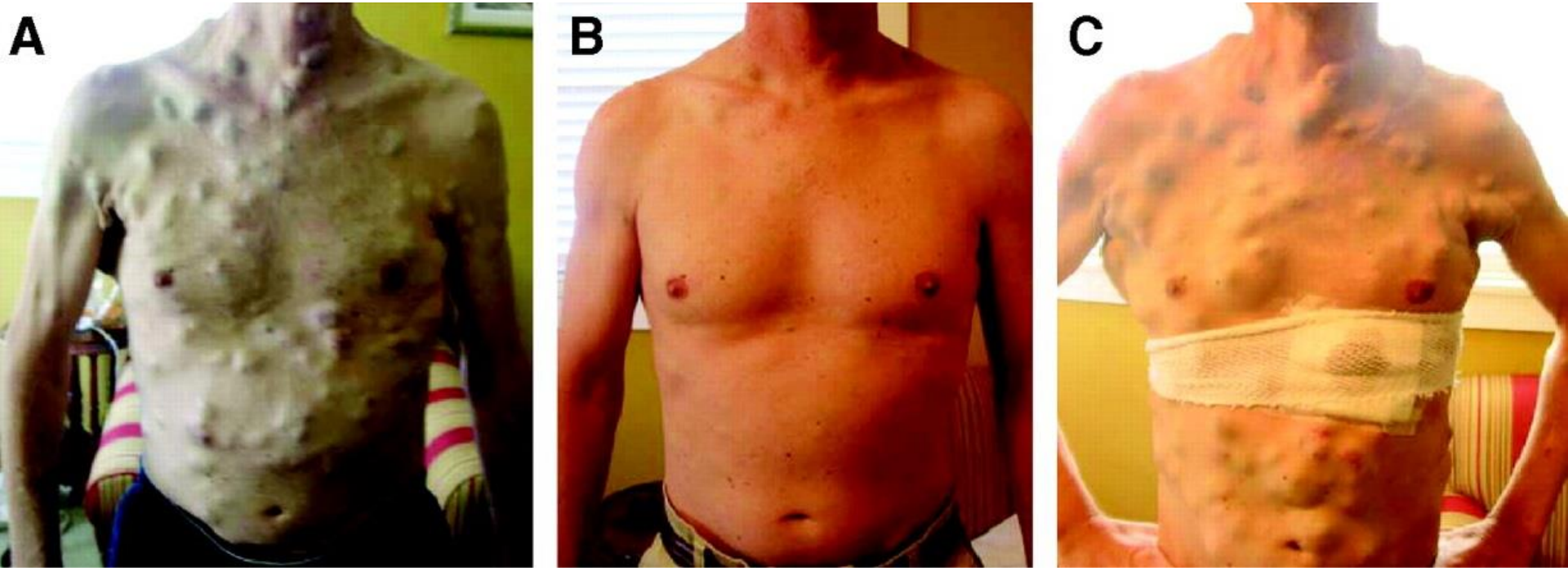
Precision Medicine



Vemurafenib on BRAF-V600 Melanoma

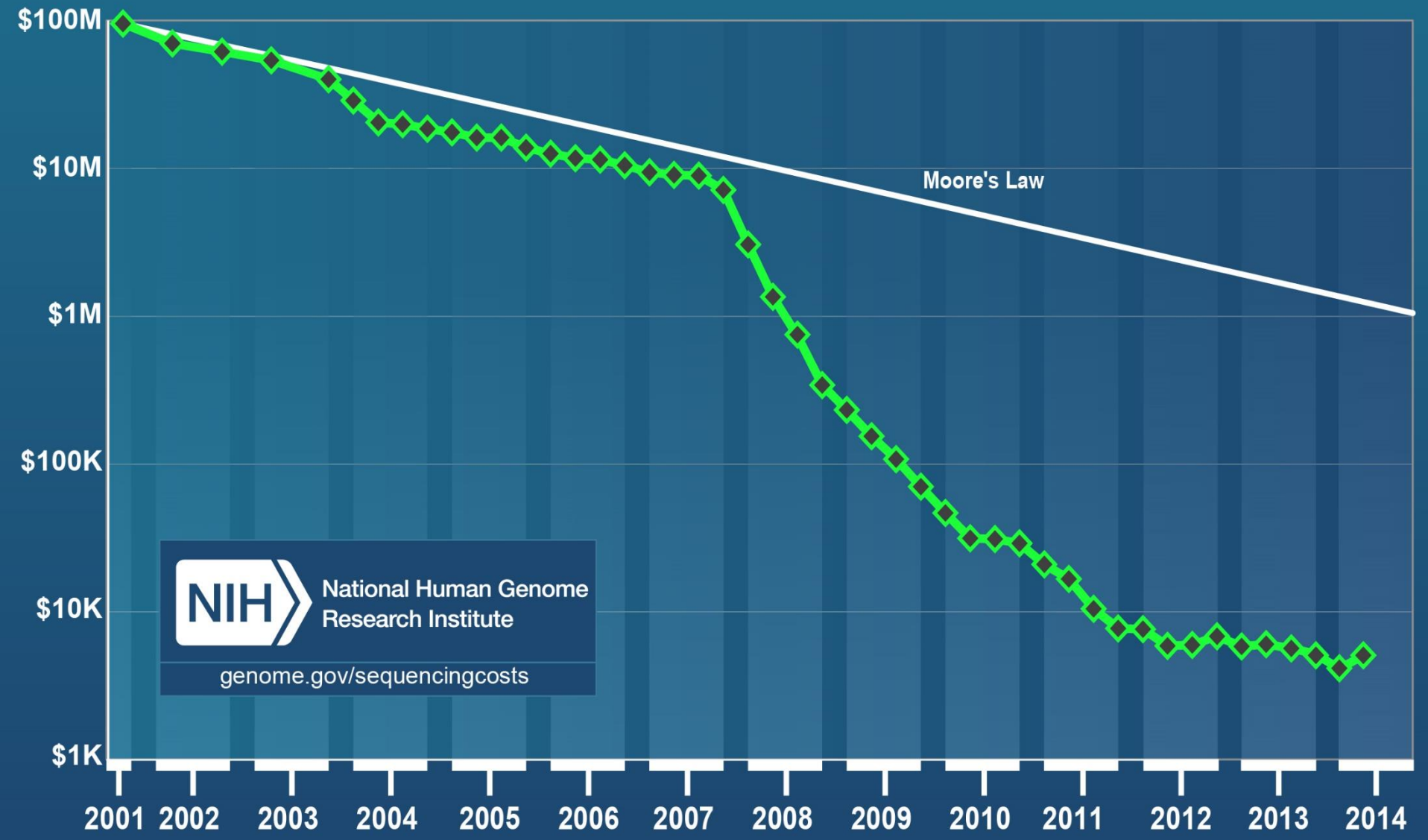
Before Treatment **15 Weeks**

Vemurafenib on BRAF-V600 Melanoma

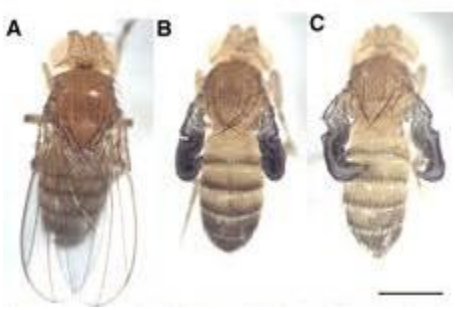


Before Treatment **15 Weeks** **23 Weeks**

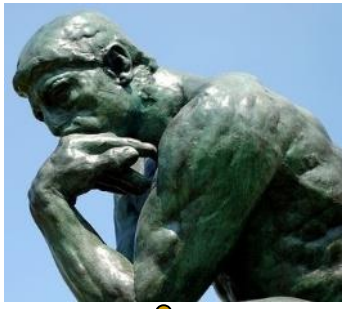
Cost per Genome



Traditional Biology



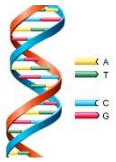
Targeted Experiments



One hypothesis

Discovery

Genomics



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

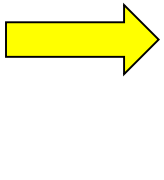
… ATTCGGATATTTAAGGC …

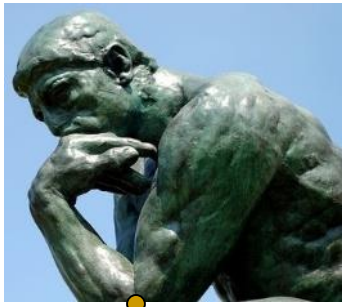
… ATTCGGGTATTTAAGCC …

… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

High-Throughput Experiments

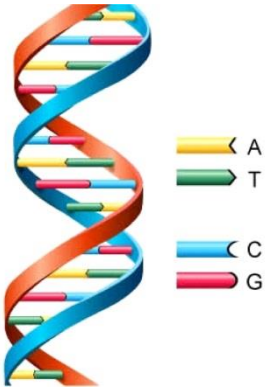




Discovery

Many hypotheses

Genome-Wide Association Studies (GWAS)



… ATTCGG**A**TATTTAAG**G**C …

Disease
(e.g., Alzheimer, Cancer)

… ATTCGGGTATTTAAGCC …

Healthy

2000

"Genetic diagnosis of diseases would be accomplished **in 10 years** and that treatments would start to roll out perhaps five years after that."

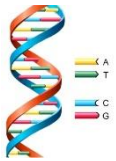
2010

"**A Decade Later, Genetic Maps Yield Few New Cures**"
New York Times, June 2010.

Key Challenges

- Human genome: 3 billion base pairs
- Potential variations: > 10 million variants
- Combination: > $10^{1000000}$ (1 million zeros)
- **Machine learning problem**
  - Atomic features: > 10 million
  - Feature combination: Too many to enumerate

Genomics



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

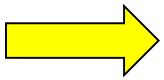
… ATTCGGATATTTAAGGC …

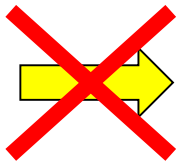
… ATTCGGGTATTTAAGCC …

… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

High-Throughput Experiments

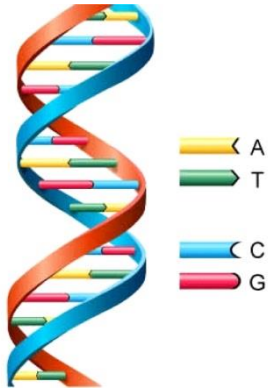




Discovery

How to Scale Discovery?

Cancer



… ATTCGG**A**TATTTAAG**G**C …

Tumor cells

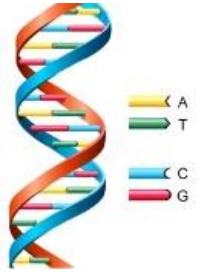
… ATTCGGGTATTTAAGCC …

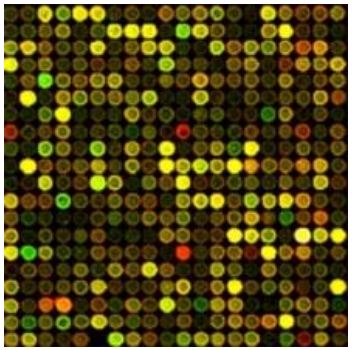
Normal cells

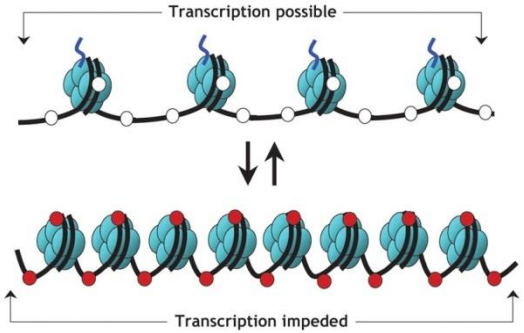
- Hundreds of mutations
- Most are “passenger”, not driver
- Can we identify likely drivers?

Panomics

… ATTCGG**A**TATTTAAG**G**C …





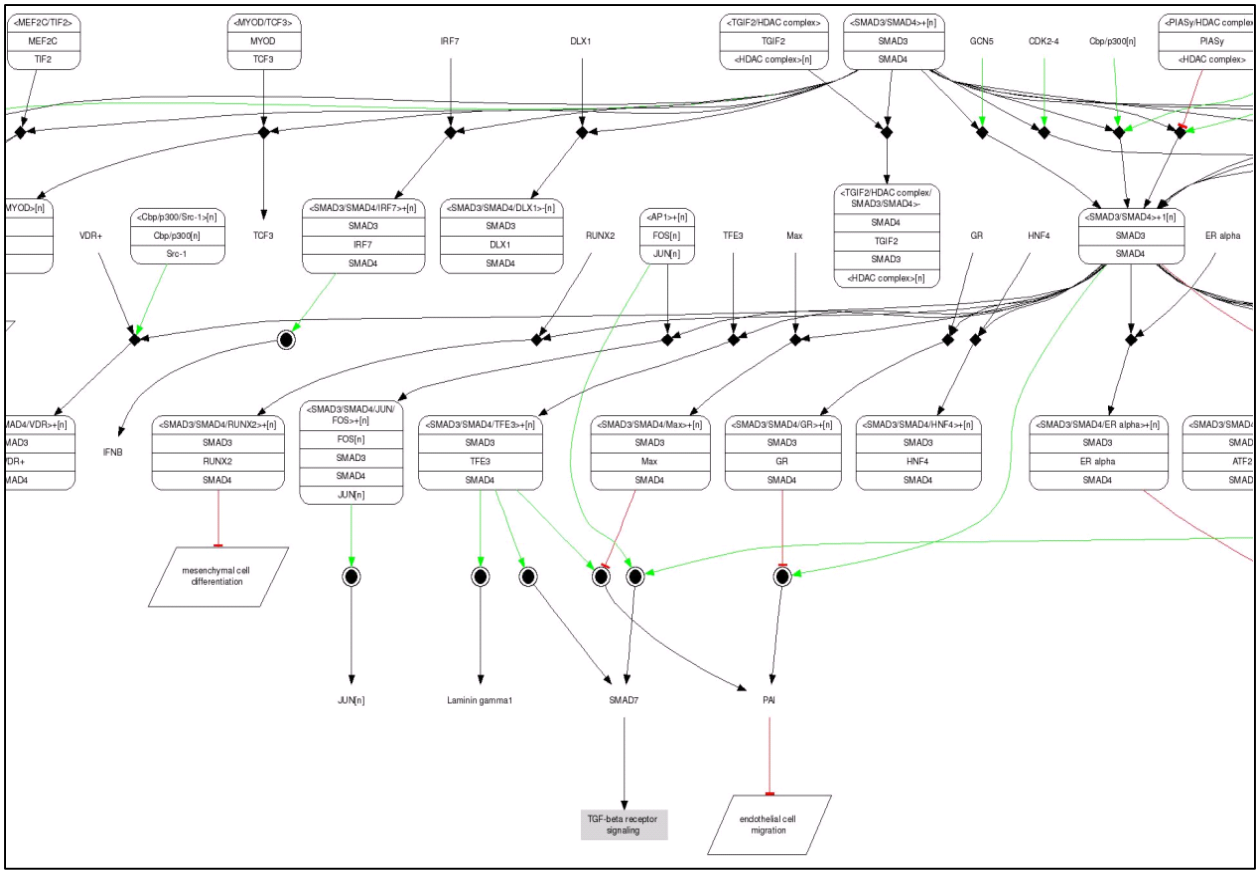


Genome **Transcriptome** **Epigenome**

……

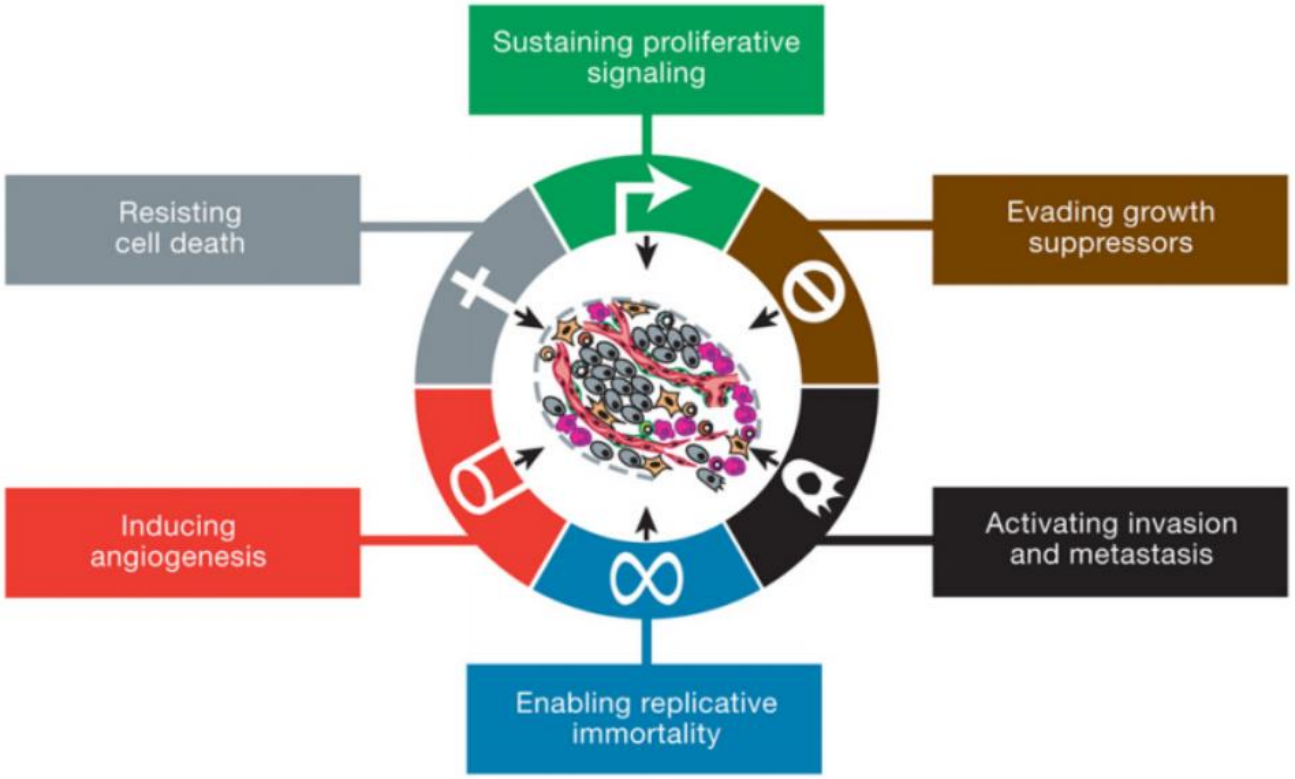
Pathway Knowledge

Genes work synergistically in pathways



Why Hard to Identify Drivers?

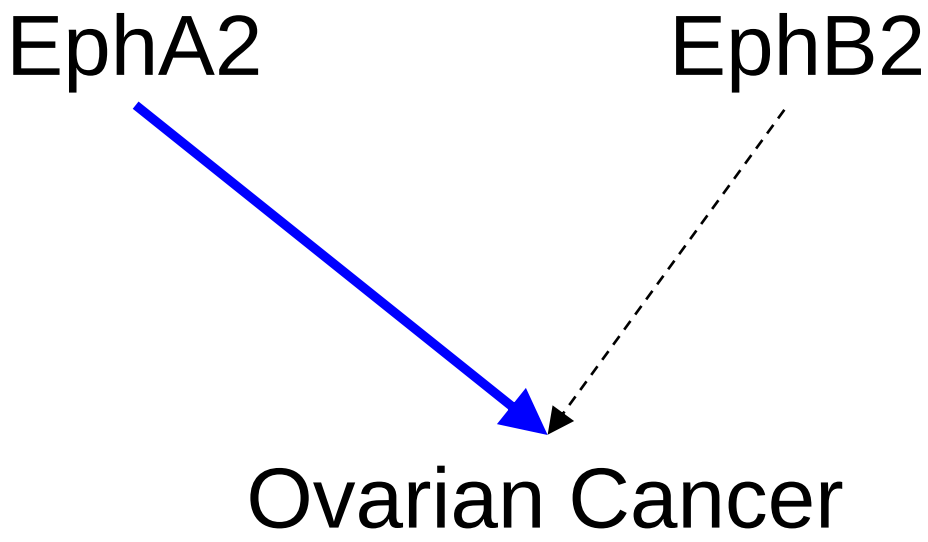
Complex diseases ← Perturb multiple pathways



Hanahan & Weinberg [Cell 2011]

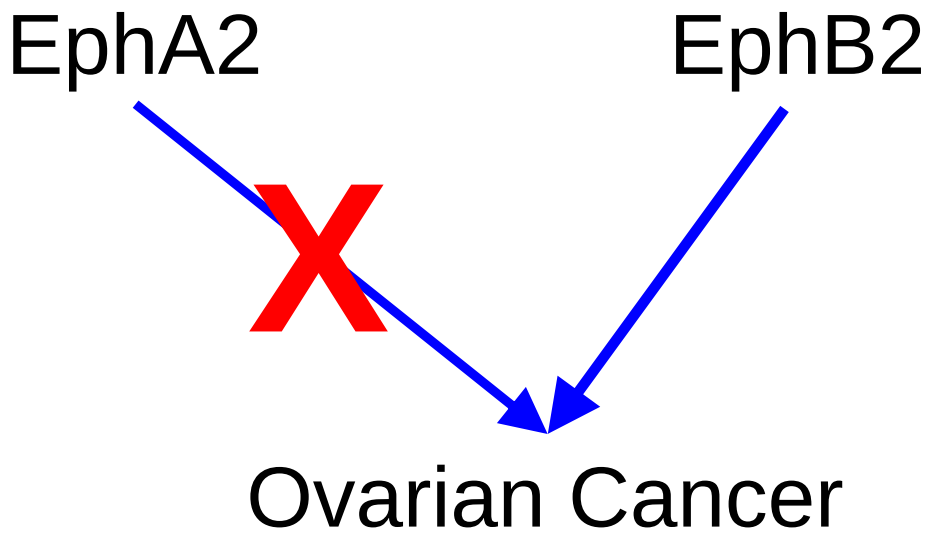
Why Cancer Comes Back?

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

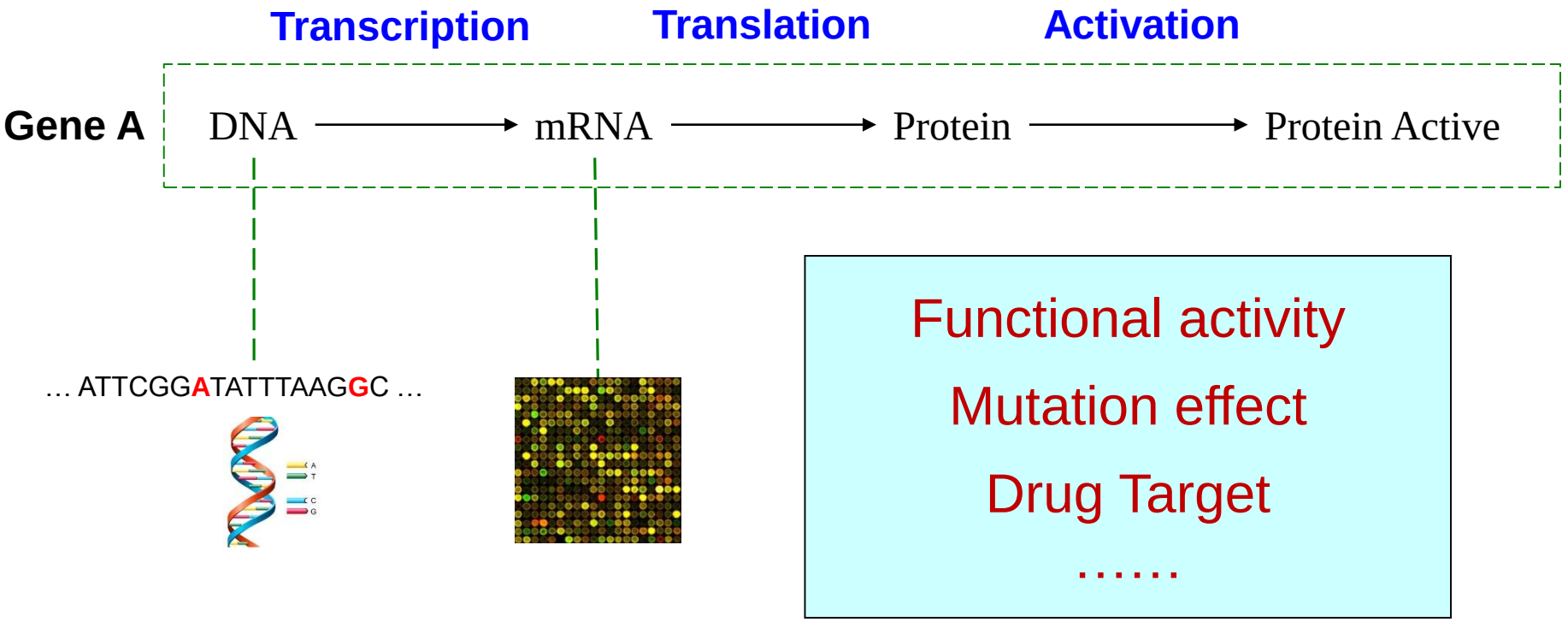


Why Cancer Comes Back?

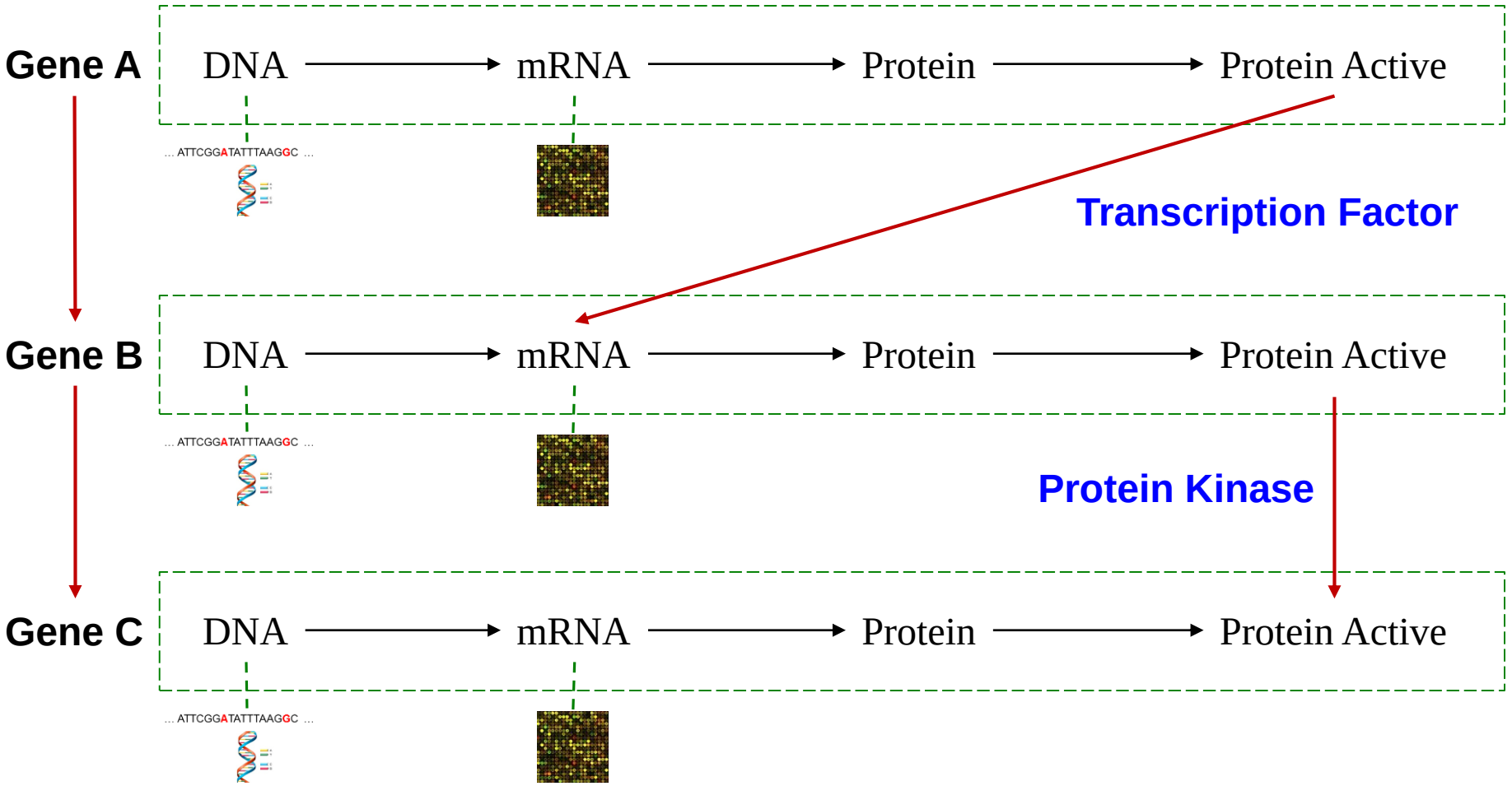
- Subtypes with alternative pathway profile
- Compensatory pathways can be activated



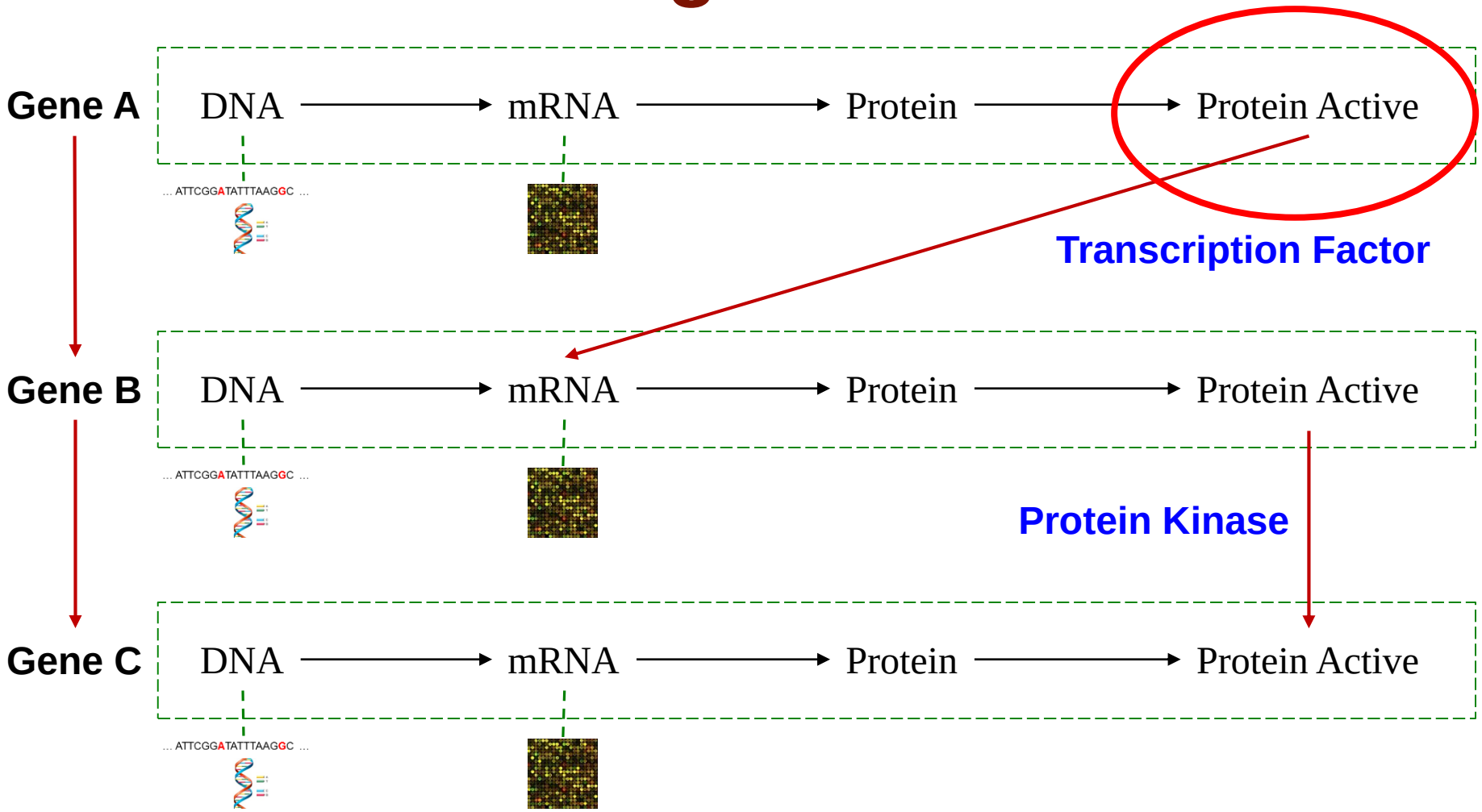
Cancer Systems Modeling



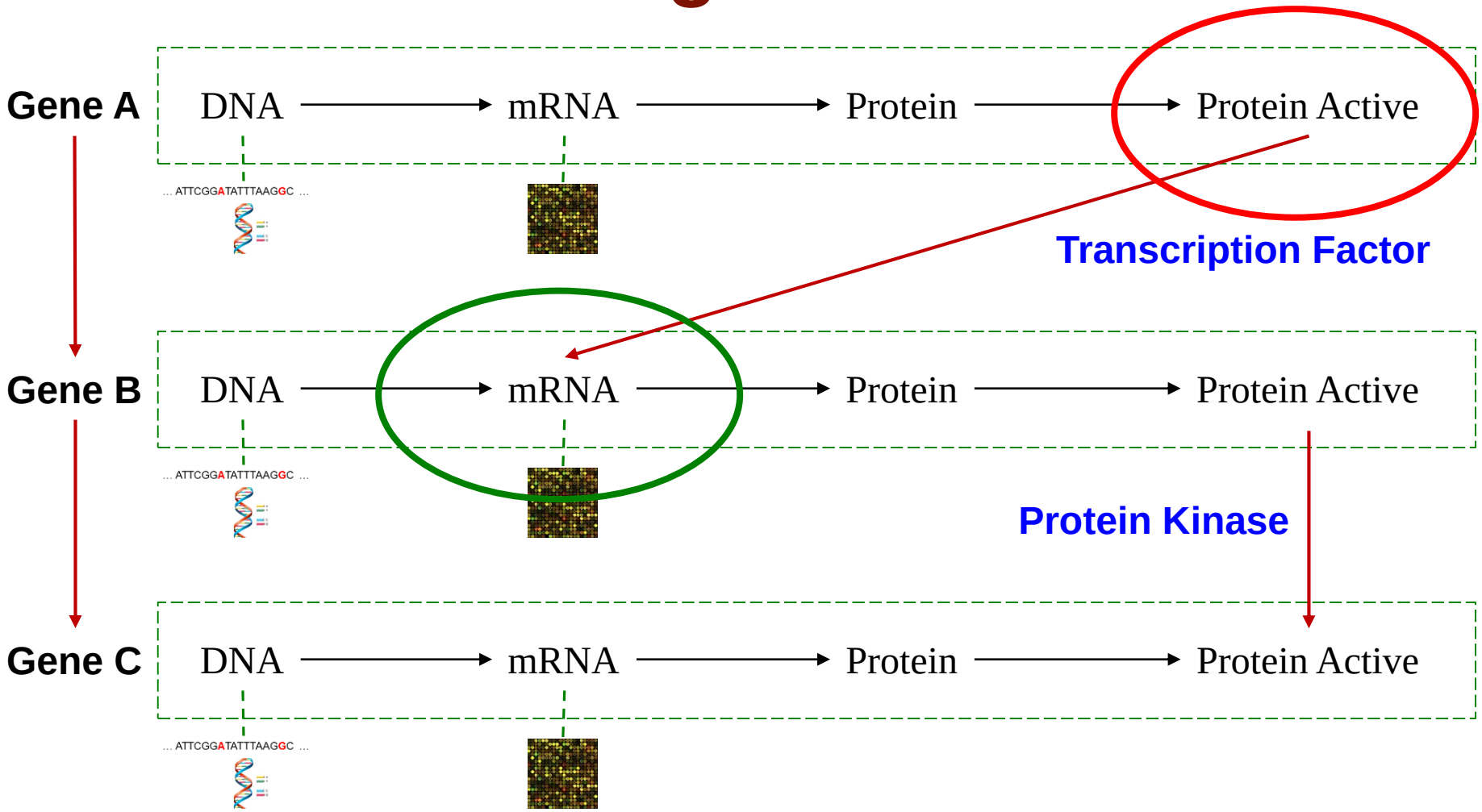
Knowledge → Model



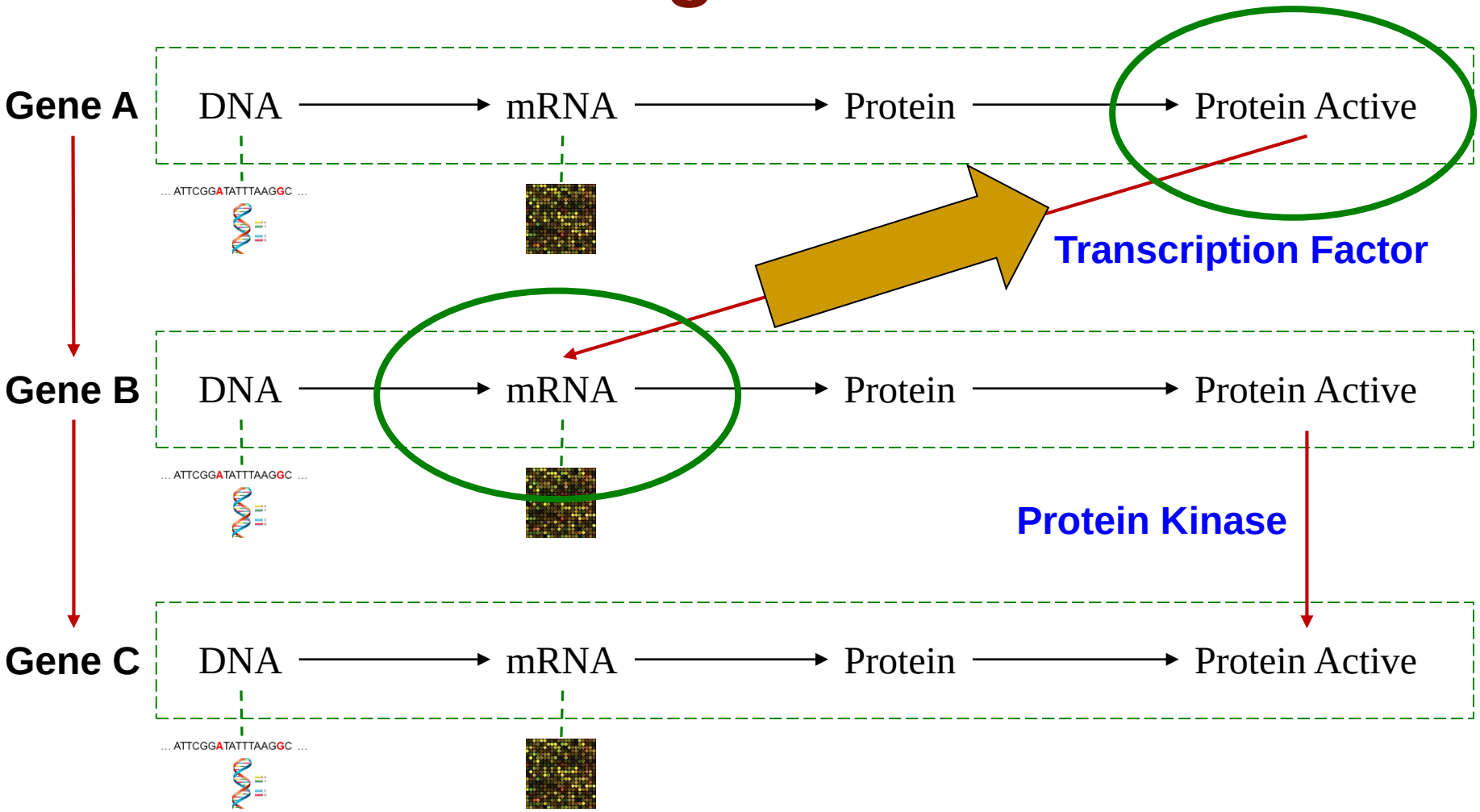
Knowledge → Model ?



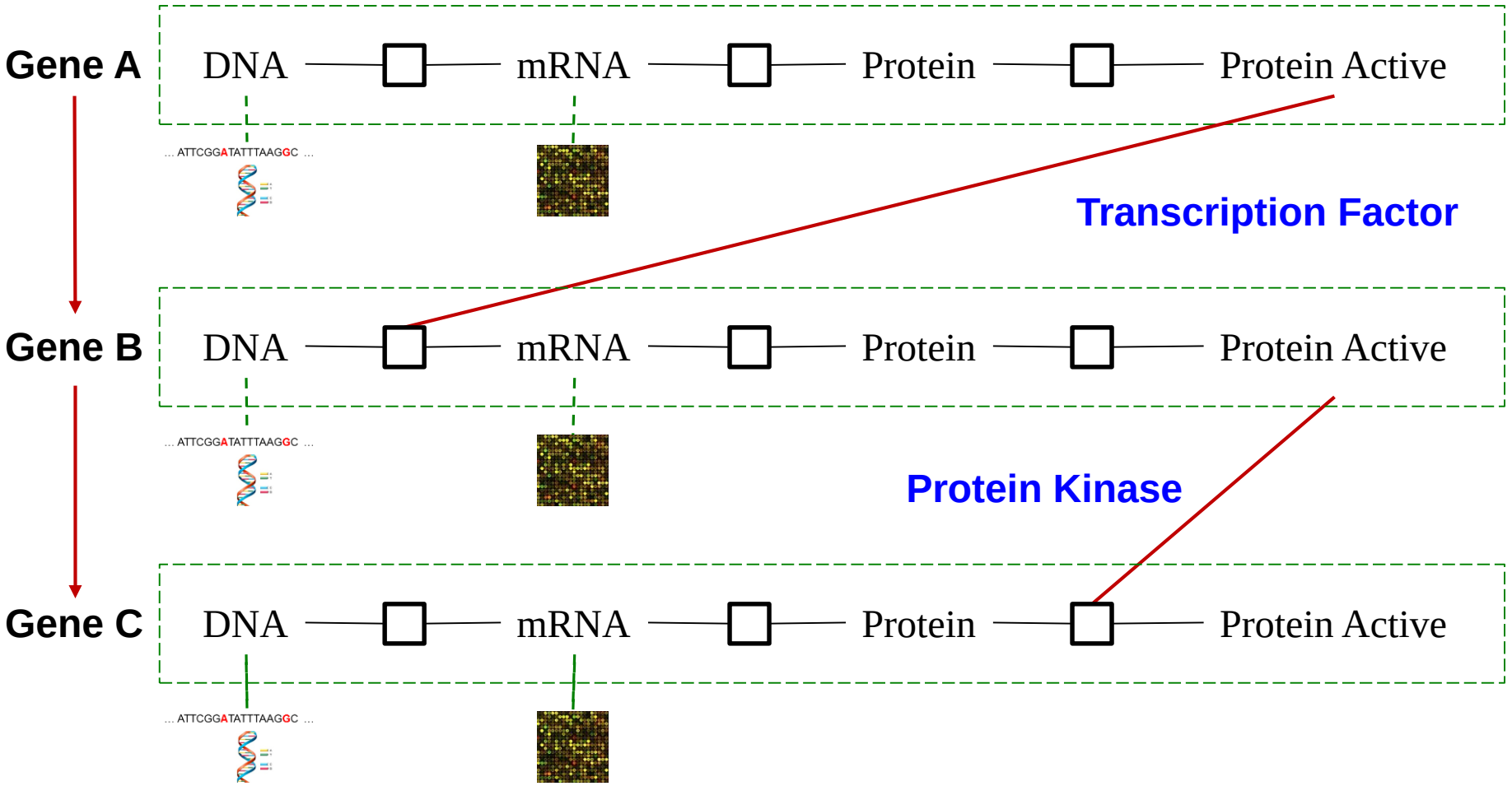
Knowledge → Model ?



Knowledge → Model !



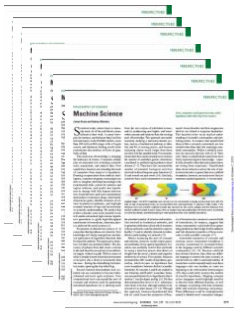
Approach: Graph HMM



Extract Pathways from PubMed

High-Throughput Data

KB



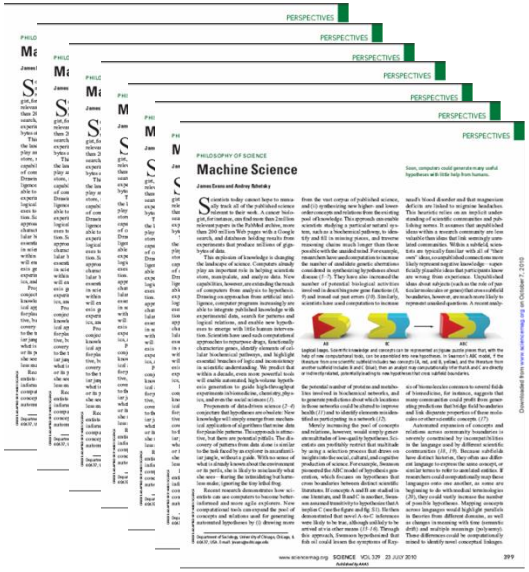
Disease Genes

Drug Targets

……

PubMed

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



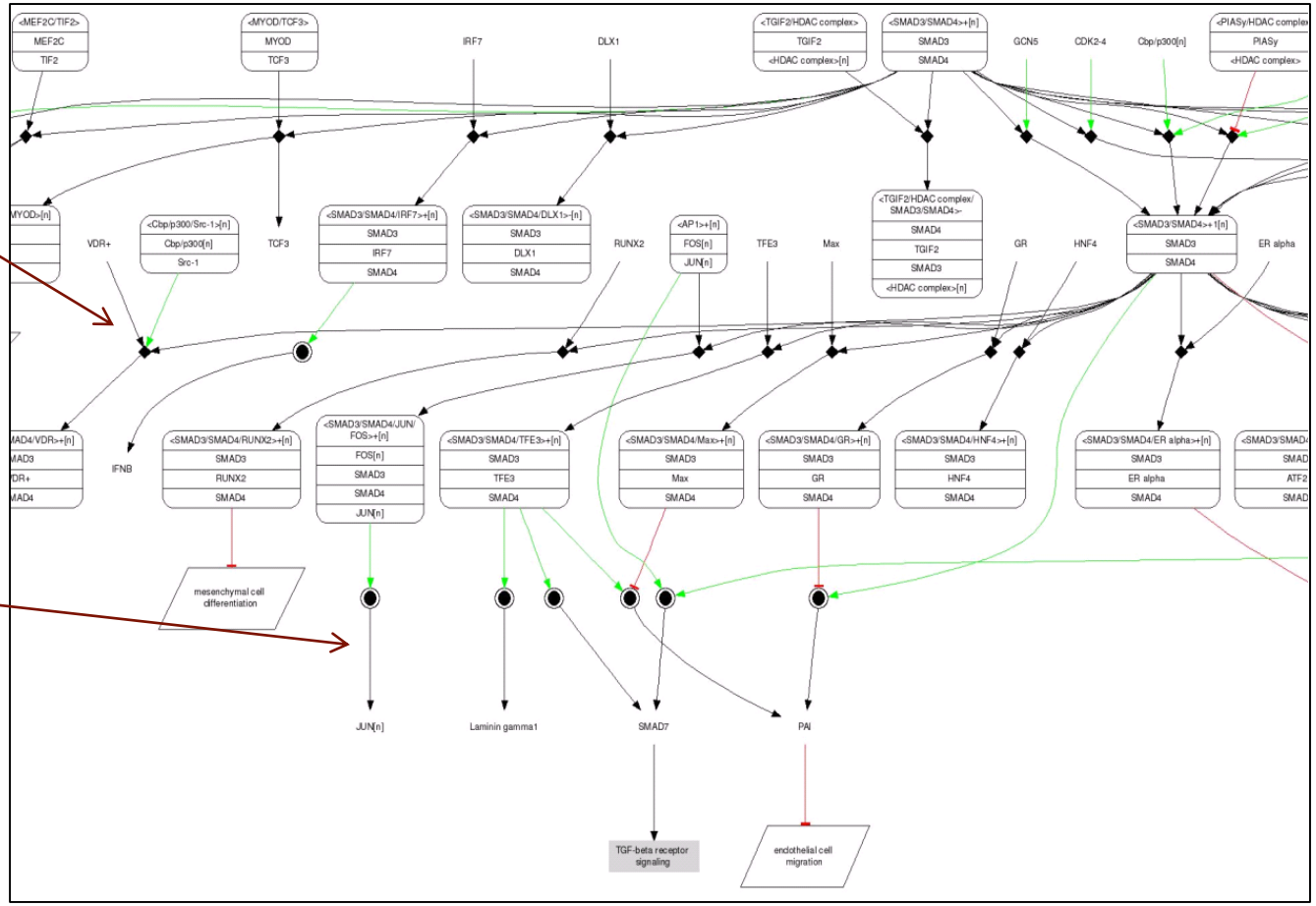
Machine Reading

PMID: 123

…
VDR+ binds to SMAD3 to form
…

PMID: 456

…
JUN expression is induced by SMAD3/4
…



Machine Reading

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

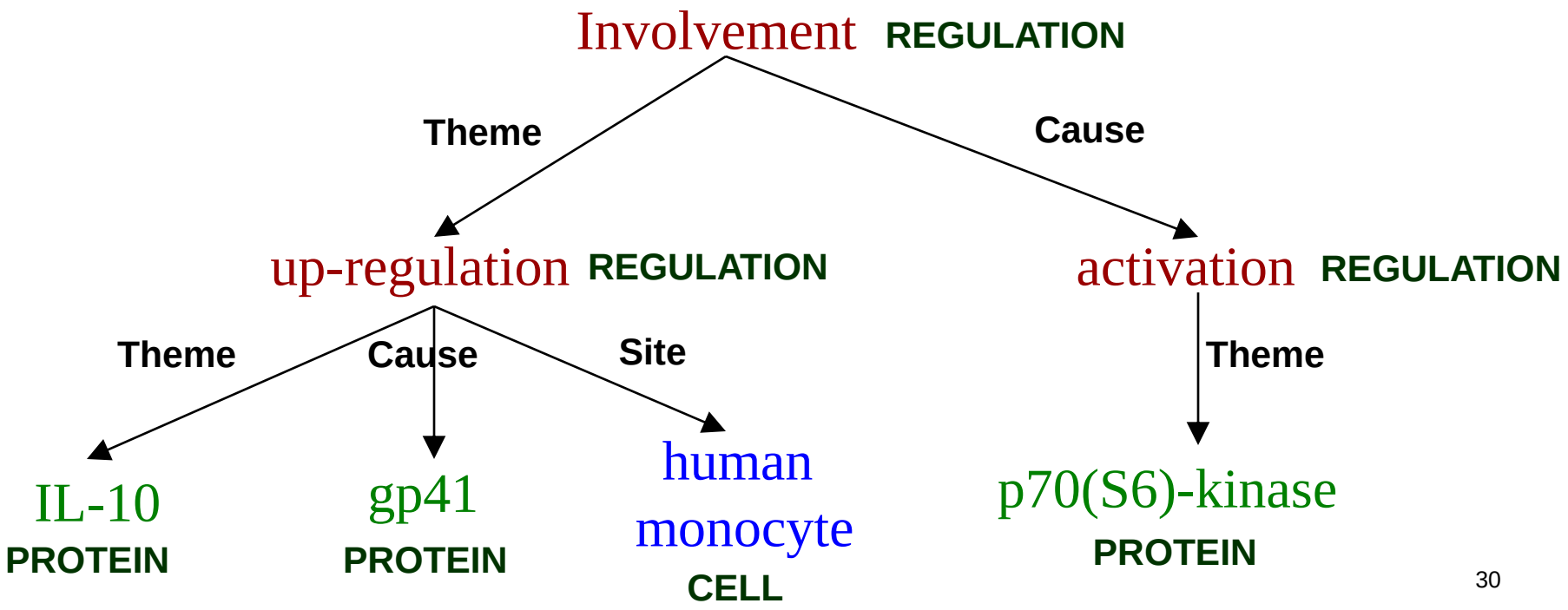
Machine Reading

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

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

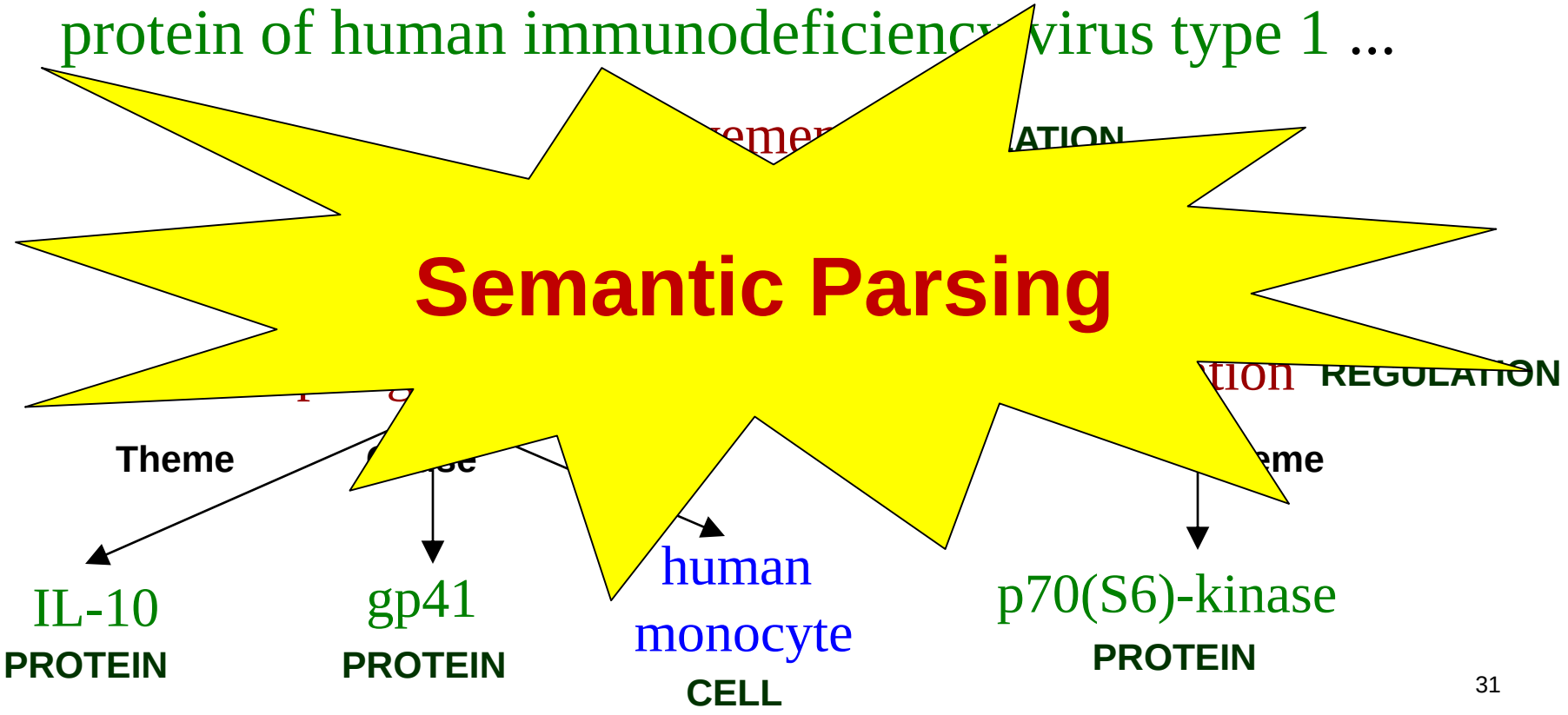
Machine Reading

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

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



Long Tail of Variations

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

Bottleneck: Annotated Examples

- GENIA (BioNLP Shared Task 2009-2013)
  - 1999 abstracts
  - MeSH: human, blood cell, transcription factor
- Challenge for “supervised” machine learning
- Can we breach this bottleneck?

Free Lunch #1: Distributional Similarity

- Similar context → Probably similar meaning
- Annotation as latent variables

  Textual expression → Recursive clusters
- Unsupervised semantic parsing

Poon & Domingos, "Unsupervised Semantic Parsing". EMNLP 2009. **Best Paper Award**.

Recursive Clustering

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 …
……

Recursive Clustering

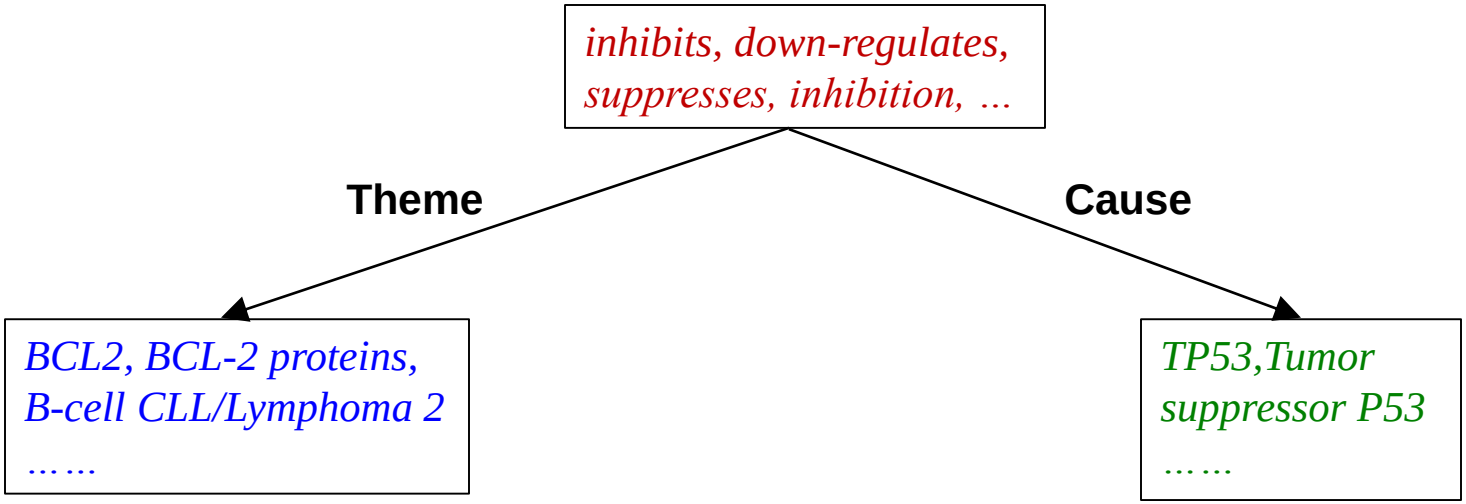
TP53 inhibits BCL2.
Tumor suppressor P53 down-regulates the activity of BCL-2 proteins.
BCL2 transcription is suppressed by P53 expression.
The inhibition of B-cell CLL/Lymphoma 2 expression by TP53 …
……

Recursive Clustering

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 …
……

Recursive Clustering

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 #2: Existing KBs

- Many KBs available
  - Gene/Protein: GeneBank, UniProt, …
  - Pathways: NCI, Reactome, KEGG, BioCarta, …
- Annotation as latent variables

  Textual expression $\rightarrow$ Table, column, join, …
- Grounded semantic parsing

Entity Extraction

HGNC

ID	Symbol	Alias
990	BCL2	B-cell CLL/Lymphoma 2, …
11998	TP53	Tumor suppressor P53, …
…	…	…

Entity Extraction

HGNC

ID	Symbol	Alias
990	BCL2	B-cell CLL/Lymphoma 2, …
11998	TP53	Tumor suppressor P53, …
…	…	…

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 …
……

Relation Extraction

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 …
……

Relation Extraction

NCI-PID Pathway KB

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

TP53 inhibits BCL2.

Tumor suppressor P53 dow…

BCL2 transcription is suppre…

The inhibition of B-cell CLL/Lymph… TP53…

……

Grounded Learning

Question Answering w.r.t. KB

System	Accuracy
ZC07	84.6
FUBL	82.8
GUSP	83.5

Supervised

Unsupervised

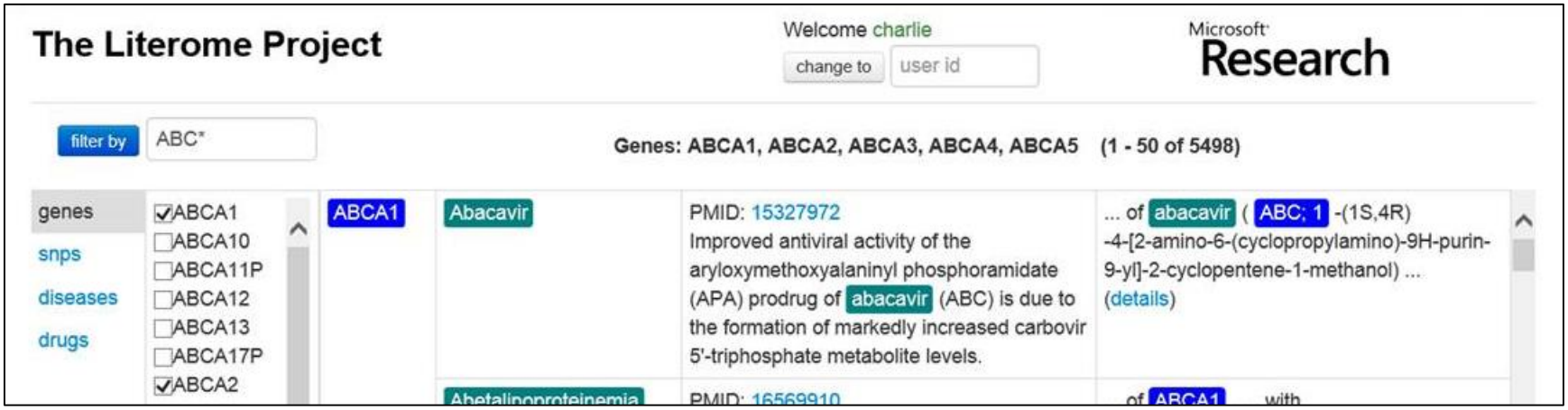
Poon, “Grounded Unsupervised Semantic Parsing”. ACL 2013.

Pathway Extraction

- **Generalize distant supervision**:
  Nested events in KB likely occur in semantic parse of some sentence
- **Prior:** Favor semantic parse grounded in KB
- Outperformed the majority of participants in original GENIA Event Shared Task

Parikh, Poon, Toutanova. *In Progress*.

Literome



Poon *et al.*, "Literome: PubMed-Scale Genomic Knowledge Base in the Cloud", *Bioinformatics* 2014.

http://literome.azurewebsites.net

PubMed-Scale Extraction

- Preliminary pass:
  - 2 million instances
  - 13,000 genes, 870,000 unique regulations
- Applications:
  - UCSC Genome Browser, MSR Interactions Track
  - Expression profile modeling
  - Validate *de novo* pathway prediction
  - Etc.

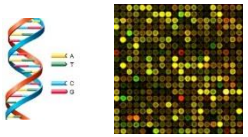
Poon, Toutanova, Quirk, “Distant Supervision for Cancer Pathway Extraction from Text”. PSB 2015. *To appear*.

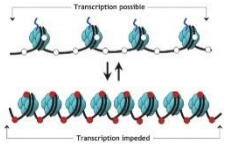
Machine Science

Evans & Rzhetsky, “Machine Science”. Science, Vol. 329, 2010.

Machine Science

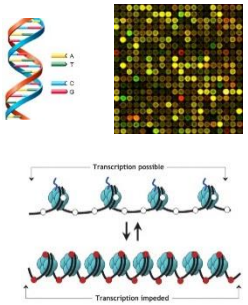
Big Data



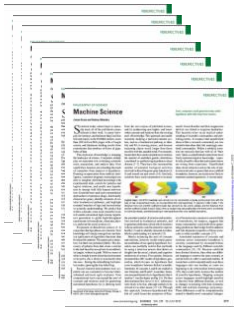


Machine Science

Big Data



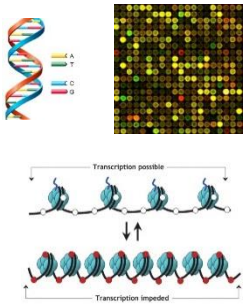
Rich Knowledge

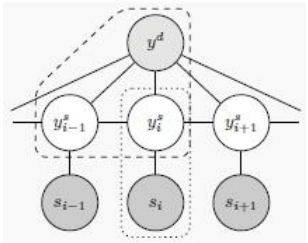


Machine Science

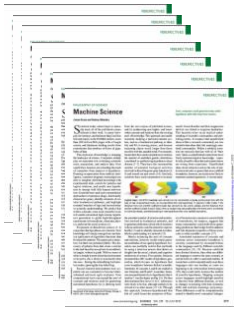
Deep Model

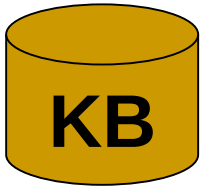
Big Data





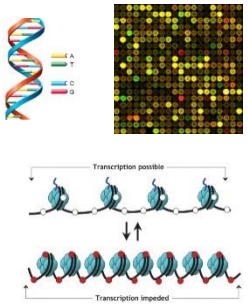
Rich Knowledge



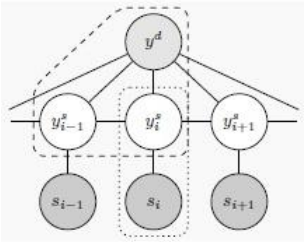


Machine Science

Big Data

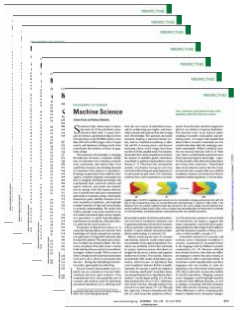


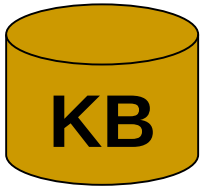
Deep Model



Hypotheses

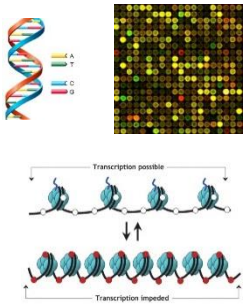
Rich Knowledge



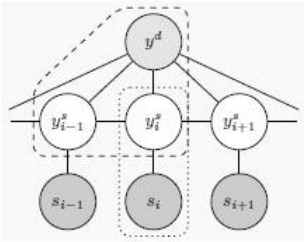


Machine Science

Big Data



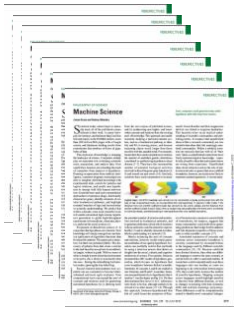
Deep Model



Hypotheses

Experiments

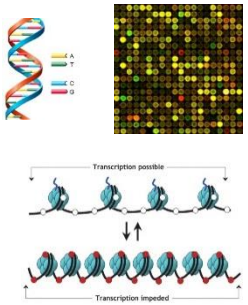
Rich Knowledge



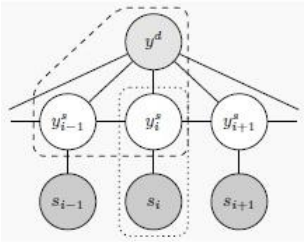
KB

Machine Science

Big Data



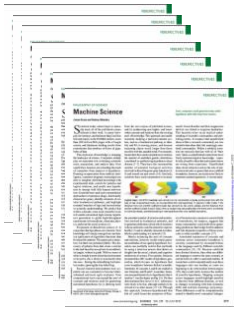
Deep Model

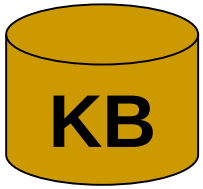


Hypotheses

Experiments

Rich Knowledge





Roadmap

- **Extract richer knowledge:**
  - Cell type, experimental condition, …
  - Hedging, negation, …
- **Formulate coherent models:**
  - Supporting evidence, contradiction, …
  - Intellectual gaps, hypotheses, …
- **Integrate w. data & experiments:**
  - Cancer panomics → Driver genes / pathways
  - Single-drug response → Drug combo prioritization

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

Personalized medicine approach to treating AML

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

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

Beat AML

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

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

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

KNIGHT CANCER INSTITUTE
Oregon Health & Science University

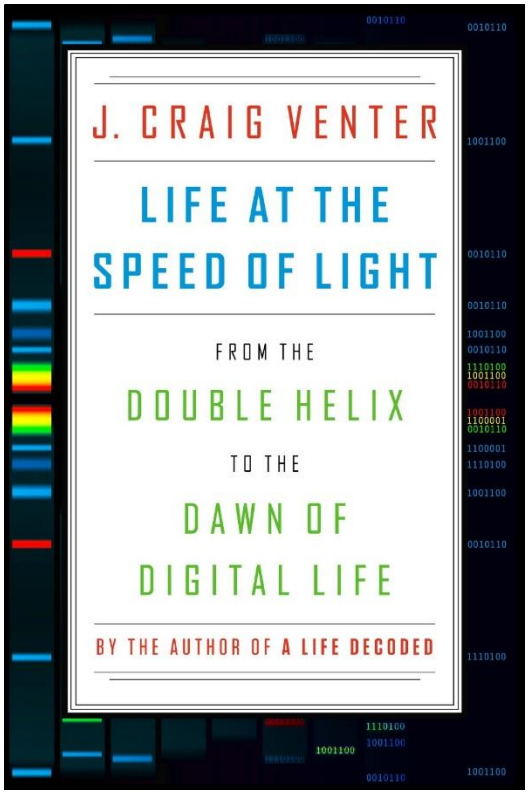
LEUKEMIA & LYMPHOMA SOCIETY
fighting blood cancers

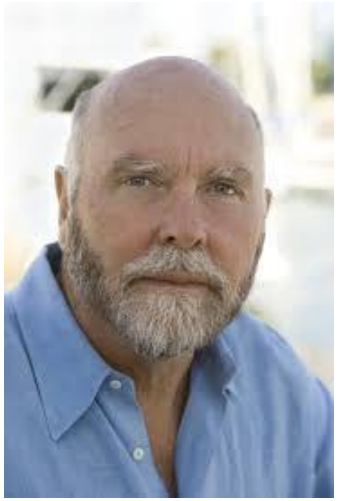
Berkeley AMP Lab

OHSU

Microsoft Research

We Have Digitized Life





Next: Digitize Medicine

PERSPECTIVE

CANCER

RNAi Therapies: Drugging the Undruggable

Sherry Y. Wu,[1] Gabriel Lopez-Berestein,[2,3] George A. Calin,[2,3] Anil K. Sood[1,3,4]*

RNA interference (RNAi) therapy is a rapidly emerging platform for personalized cancer treatment. Recent advances in small interfering RNA delivery and target selection provide unprecedented opportunities for clinical translation. Here, we discuss these advances and present strategies for making RNAi-based therapy a viable part of cancer management.

Knock down genes A, B, C → Cure

Summary

- Precision medicine is the future
- **Cancer systems modeling**

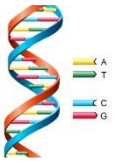
  Graphical model: Pathways + Panomics data
- **Extract pathways from PubMed**

  Machine reading by grounded semantic parsing
- **Literome**: KB for genomic medicine

Acknowledgments

- **U. Chicago:** Andrey Rzhetsky, Kevin White
- **OHSU:** Brian Drucker, Jeff Tyner
- **Berkeley AMP Lab:** David Patterson
- **U. Wisconsin:** Anthony Gitter
- **Microsoft Research:** Chris Quirk, Kristina Toutanova, David Heckerman, Ankur Parikh, Lucy Vanderwende, Bill Bolosky, Ravi Pandya

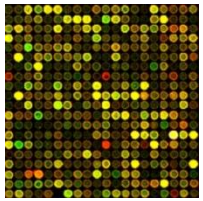
Summary



… ATTCGGATATTTAAGGC …

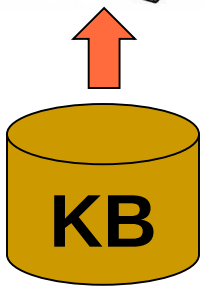
… ATTCGGGTATTTAAGCC …

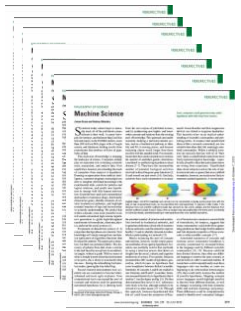
……



……

High-Throughput Data





Disease Genes

Drug Targets

……