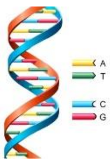
Semantic Parsing for Cancer Panomics

Hoifung Poon

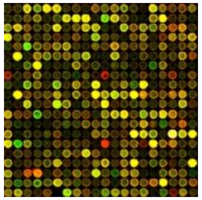
Overview



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

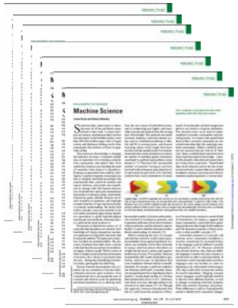
……



……

High-Throughput Data

KB

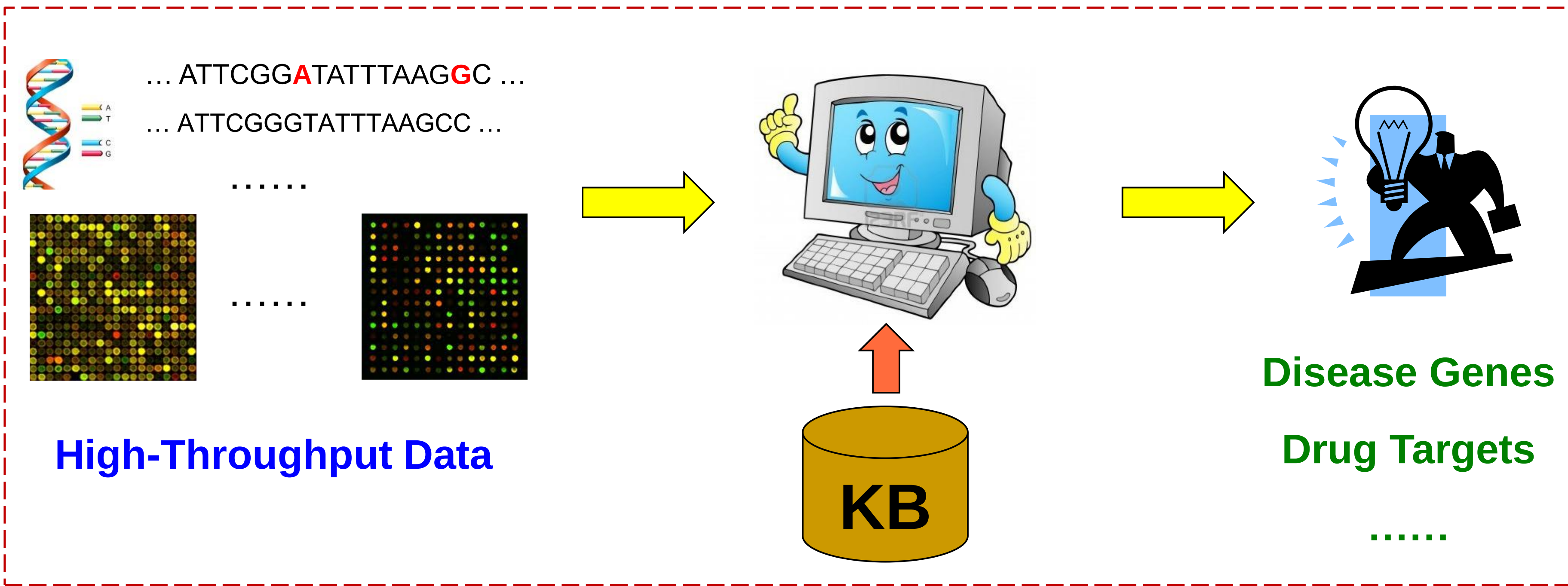


Disease Genes

Drug Targets

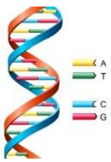
……

Overview



Infer cancer driver mutations

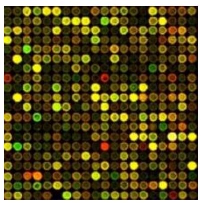
Overview



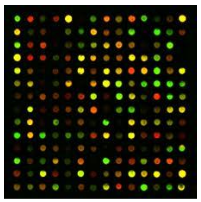
... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...

......

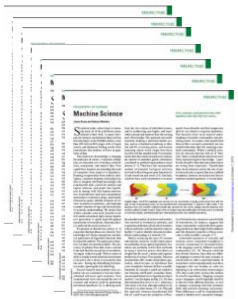


......



High-Throughput Data

KB



Grounded Unsupervised Semantic Parsing

Disease Genes

Drug Targets

...

Extract Pathways from Pubmed

Collaborators

David Heckerman

Kristina Toutanova

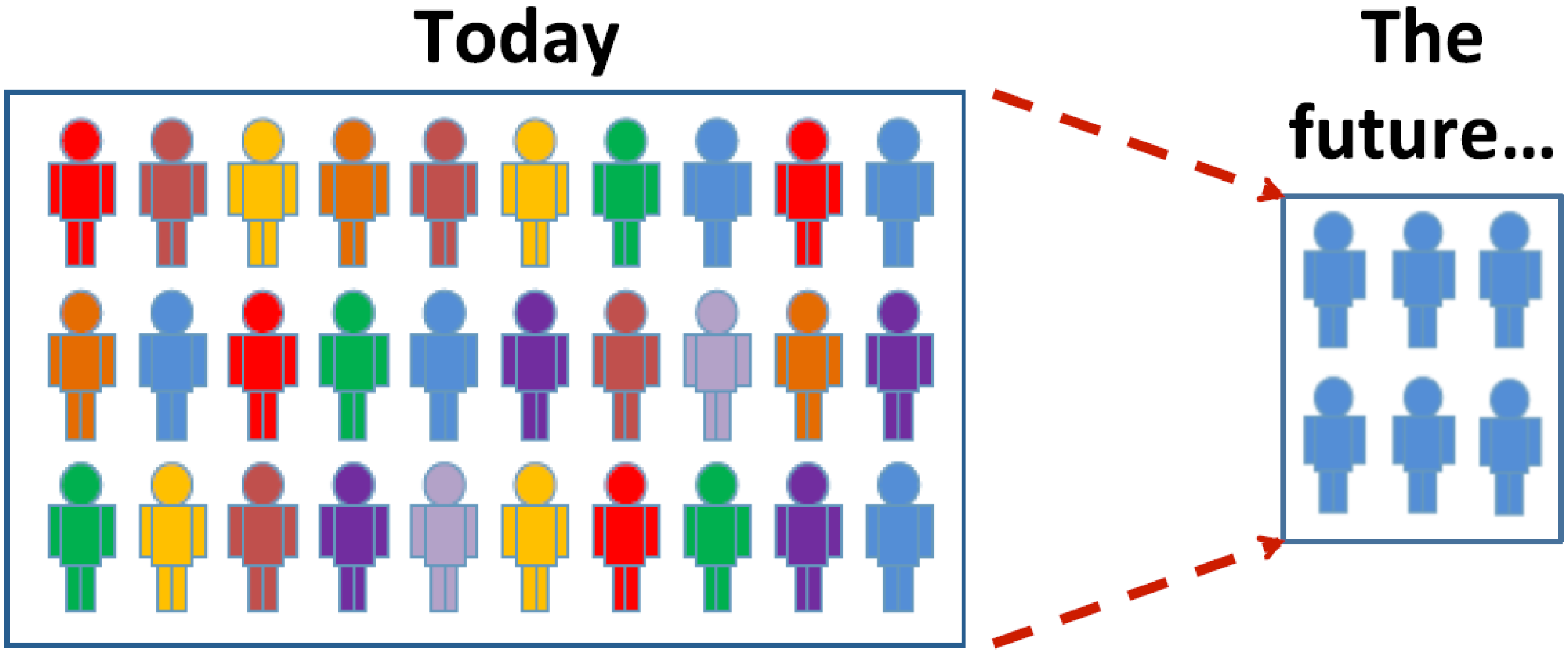
Chris Quirk

Tony Gitter

Ankur Parikh

Lucy Vanderwende

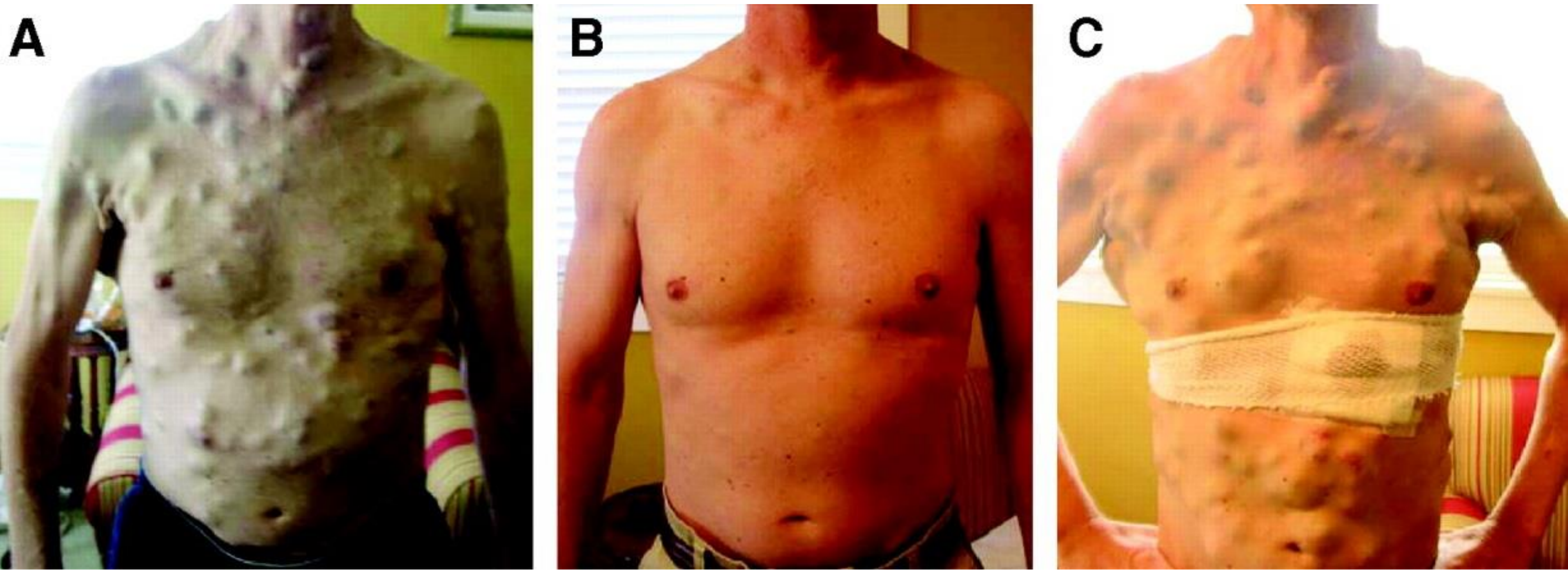
Precision Medicine



Vemurafenib on BRAF-V600 Melanoma

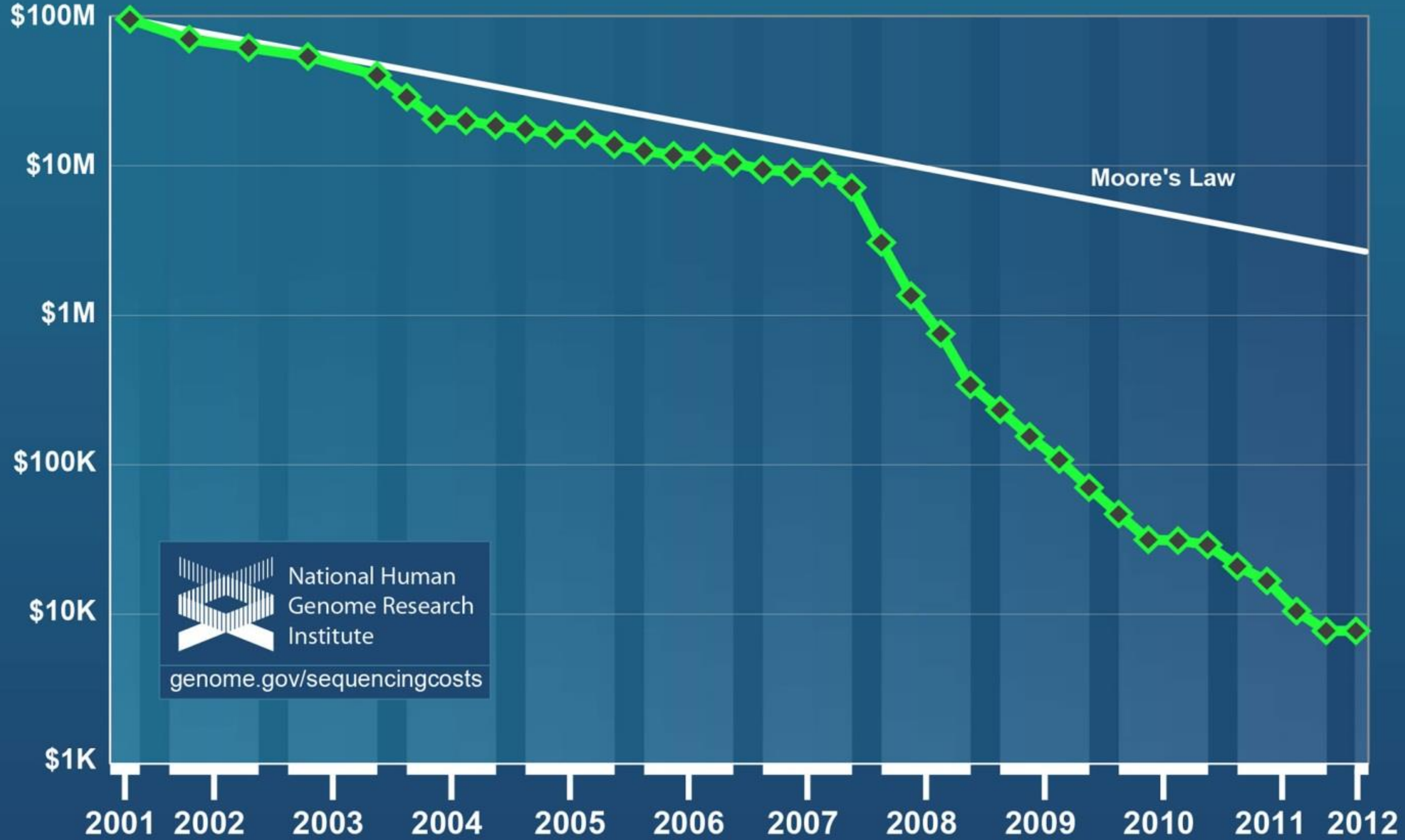
Before Treatment **15 Weeks**

Vemurafenib on BRAF-V600 Melanoma

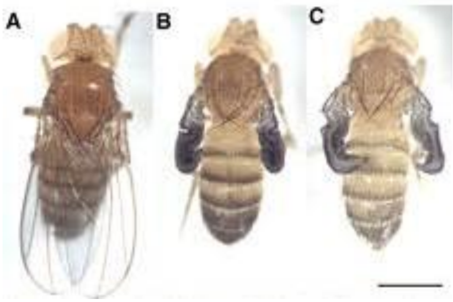


Before Treatment **15 Weeks** **23 Weeks**

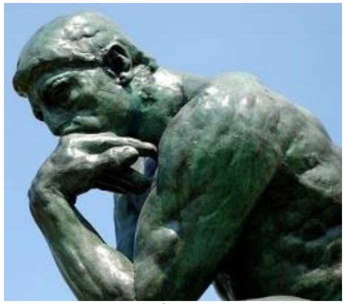
Cost per Genome



Traditional Biology



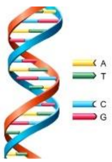
Targeted Experiments



One hypothesis

Discovery

Genomics



… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …
… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …
… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …

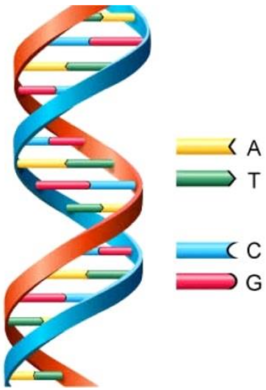
High-Throughput Experiments

?

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.”

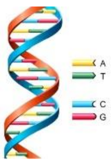
2010

“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 mutations
- 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

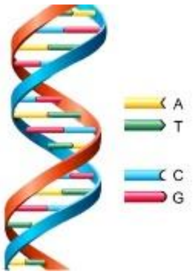
… ATTCGG**A**TATTTAAG**G**C …  Tumor cells

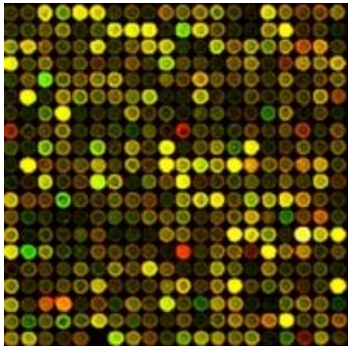
… ATTCGGGTATTTAAGCC …  Normal cells

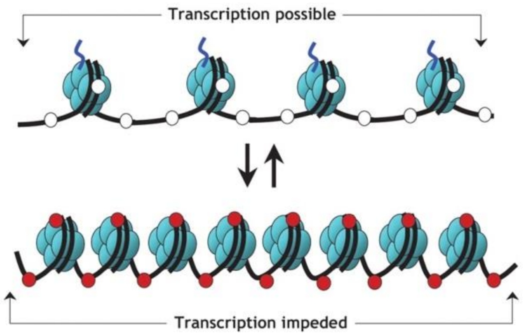
- Hundreds of mutations
- Most are “passenger”, not driver
- Can we identify likely drivers?

Panomics

… ATTCGG**A**TATTTAAG**G**C …





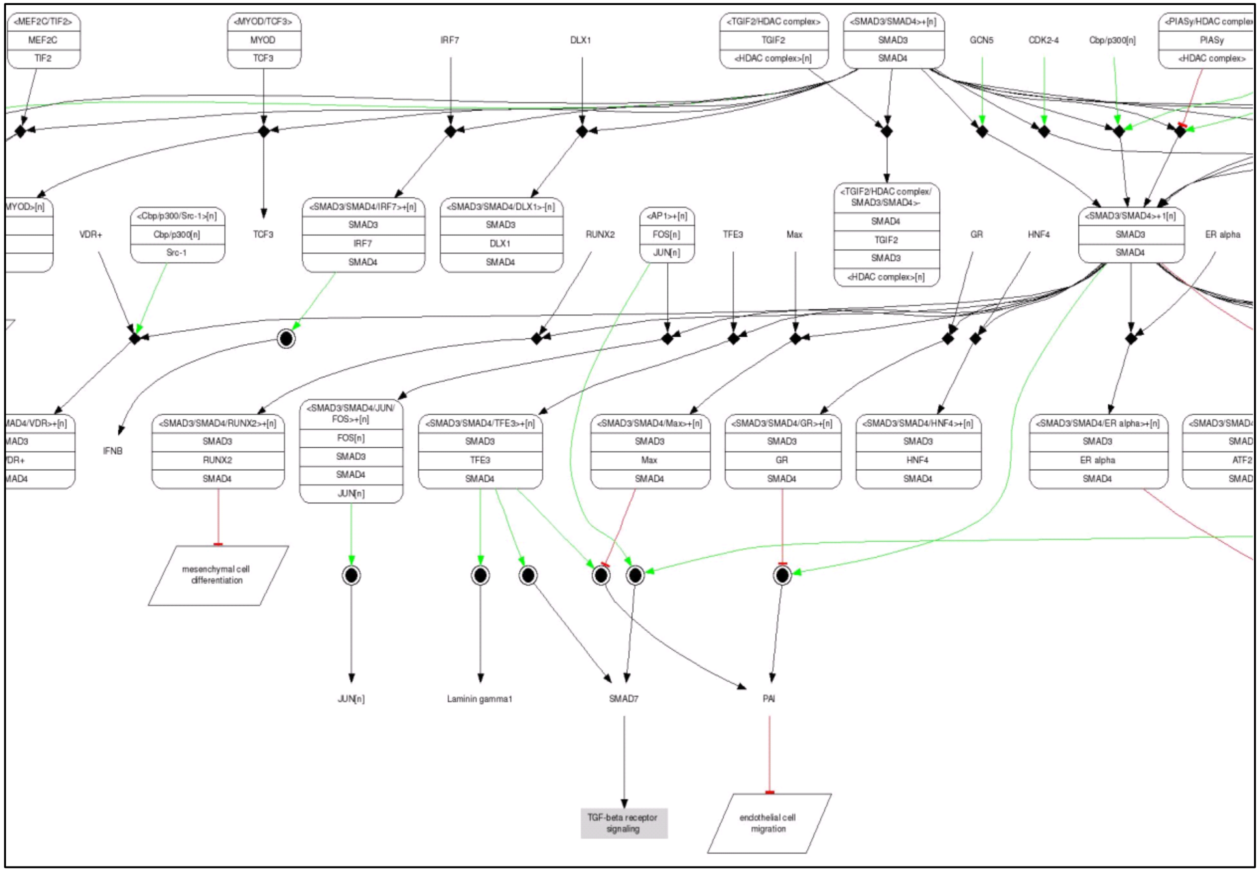


Genome **Transcriptome** **Epigenome**

……

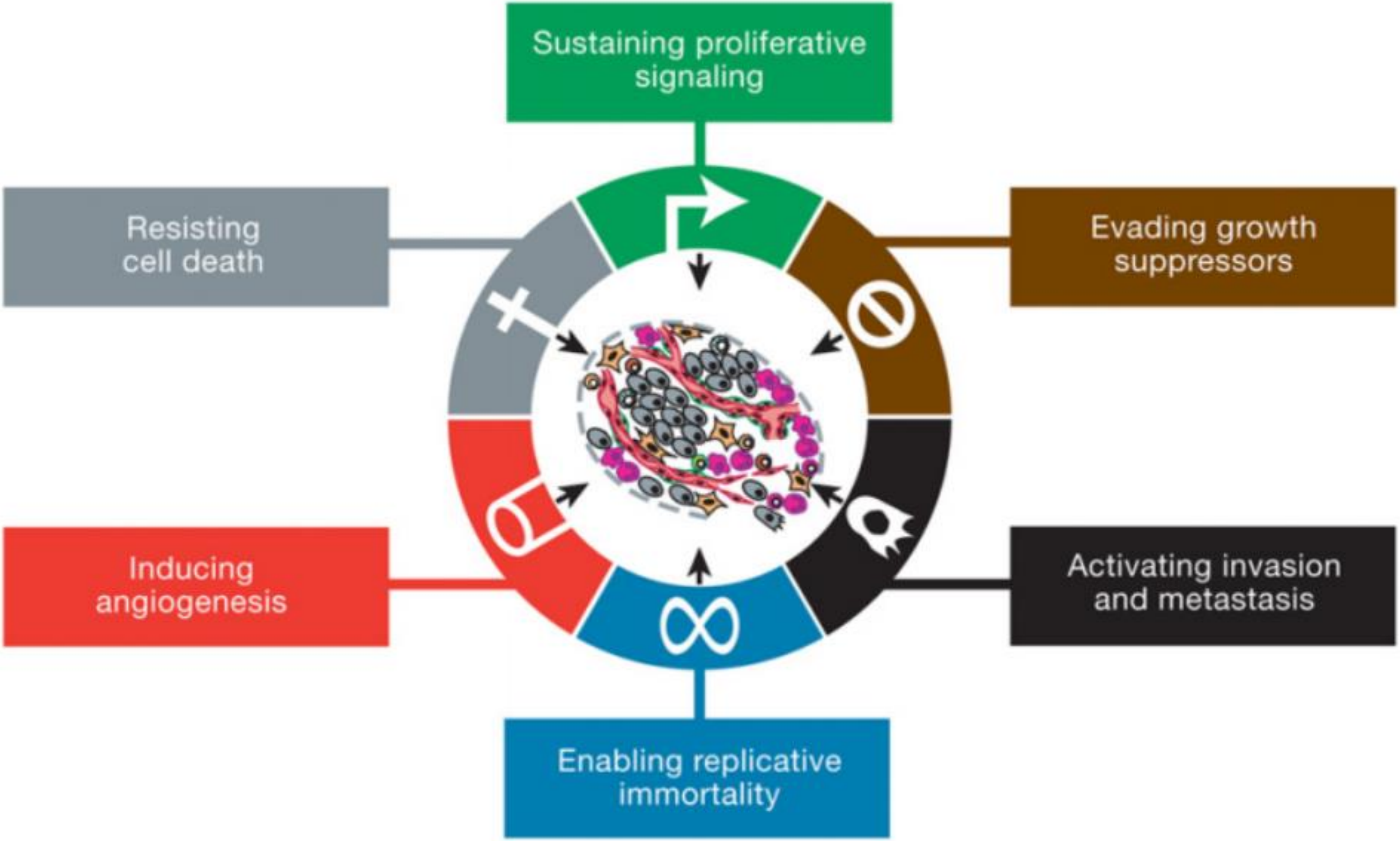
Pathway Knowledge

Genes work synergistically in pathways



Why Hard to Identify Drivers?

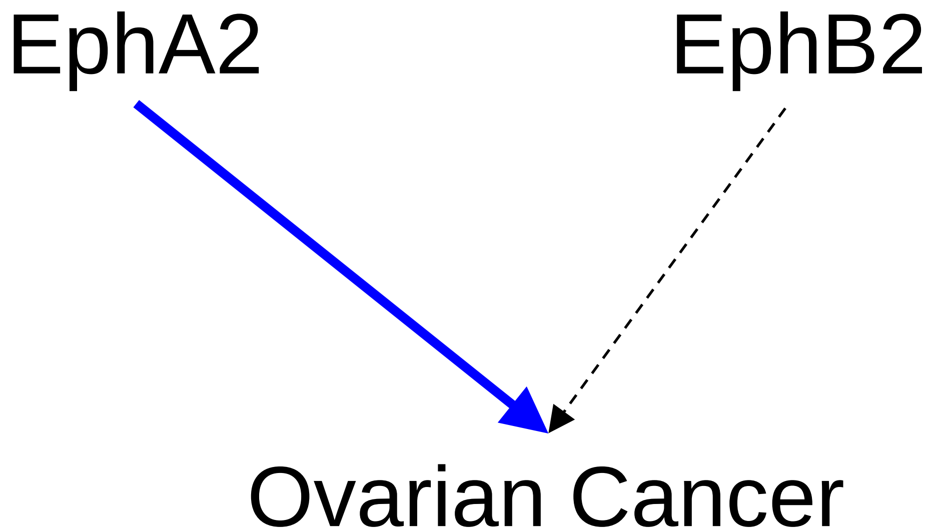
- Complex diseases $\leftarrow$ Synergistic perturbation of multiple pathways
- Cancer: 6 – 8 “hallmarks”
  - Promote growth
  - Avoid suicide
  - Evade immune attack
  - Induce blood vessels
  - Invade neighboring tissues
  - …



Hanahan & Weinberg [Cell 2011]

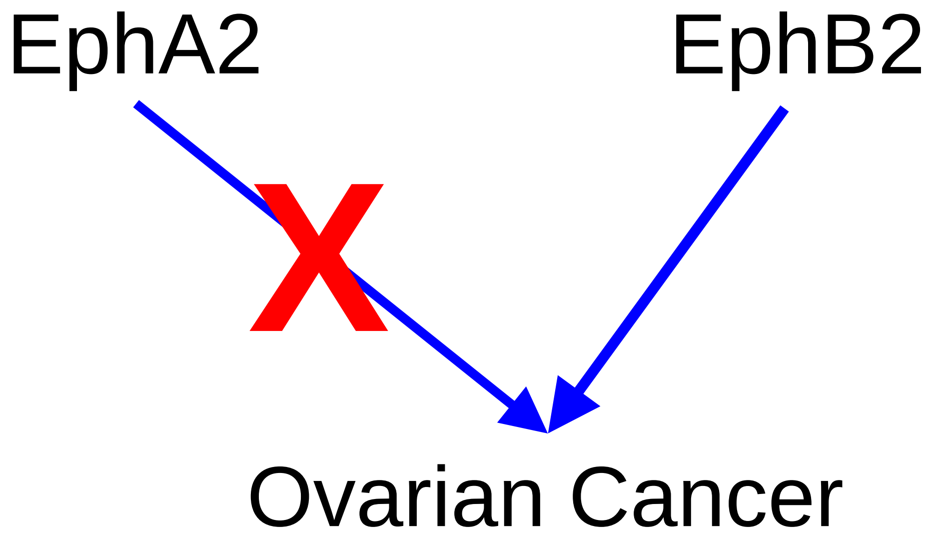
Why Cancer Comes Back?

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



Why Cancer Comes Back?

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



A Grammar of Cancer?

Cancer $\rightarrow$ Anti-Apoptosis & ProGrowth & …

Anti-Apoptosis $\rightarrow$ Deactivate TP53

Anti-Apoptosis $\rightarrow$ Activate BCL-2

…

Infer Cancer Driver Mutations

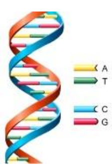
Transcription **Translation** **Activation**

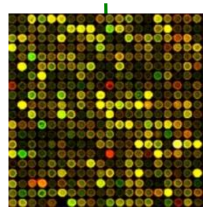
Gene A DNA → mRNA → Protein → Protein Active

What's the level of activity?

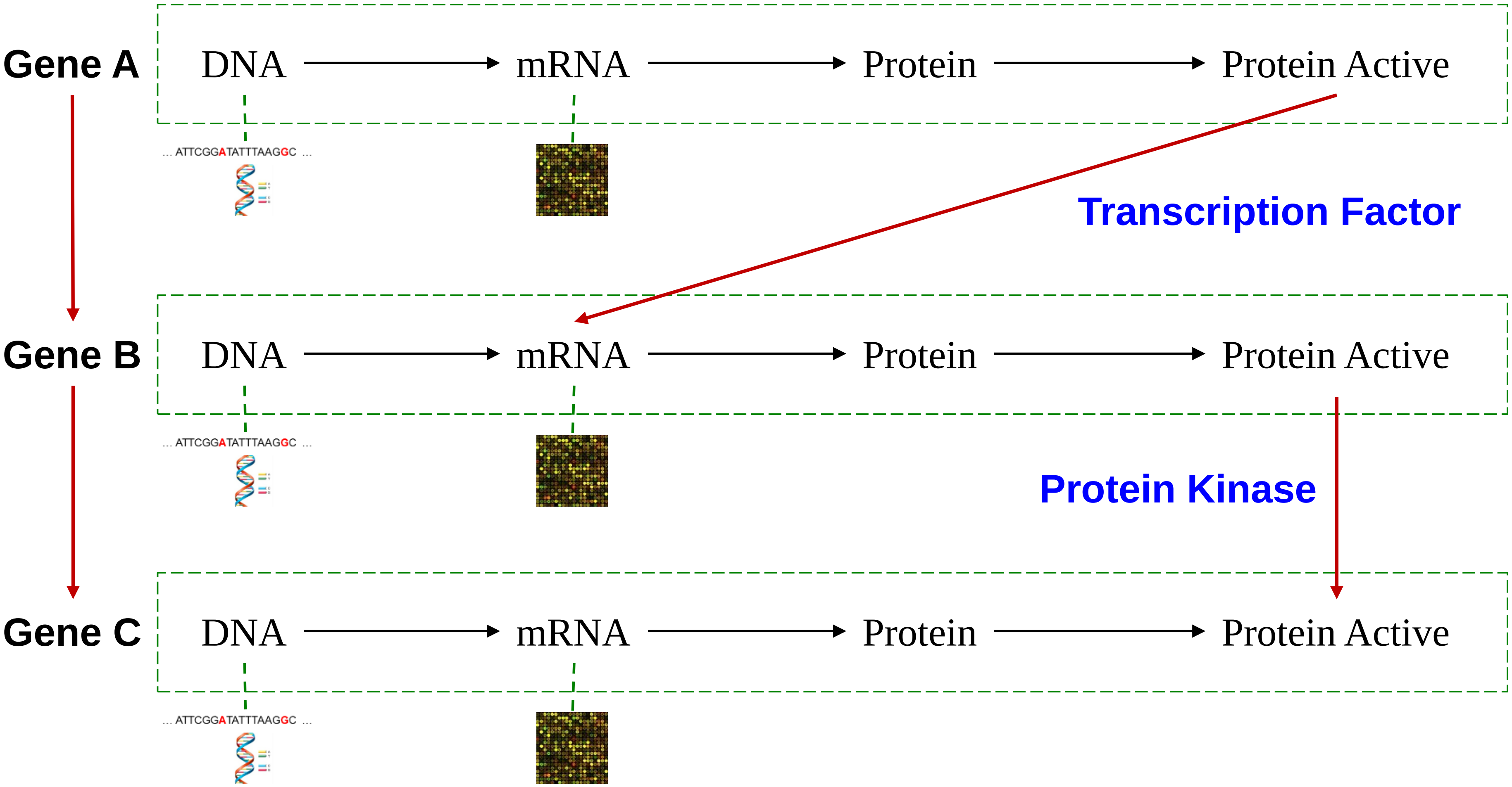
Is change caused by mutation?

… ATTCGG**A**TATTTAAG**G**C …



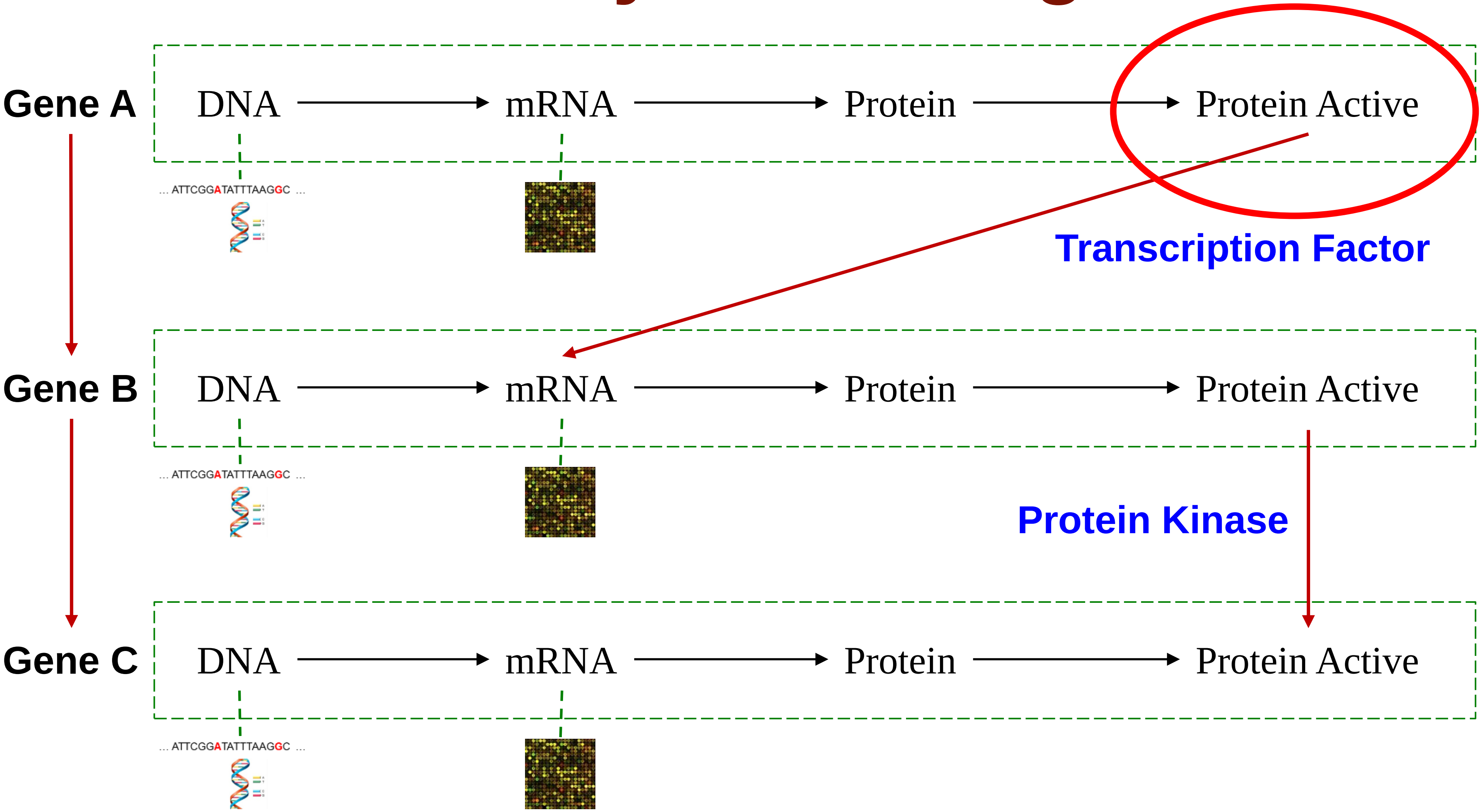


Pathway Knowledge

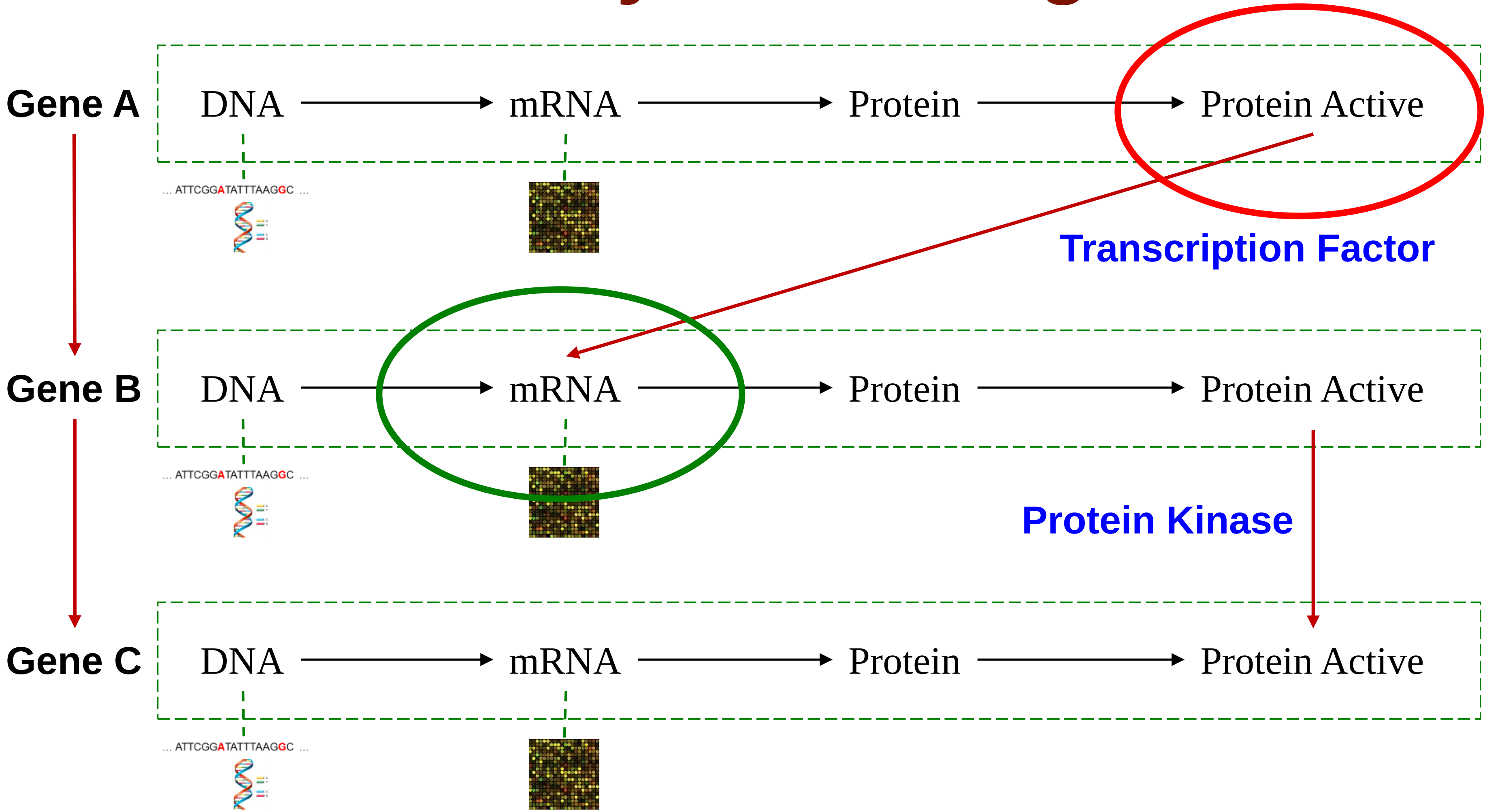


Pathway Knowledge

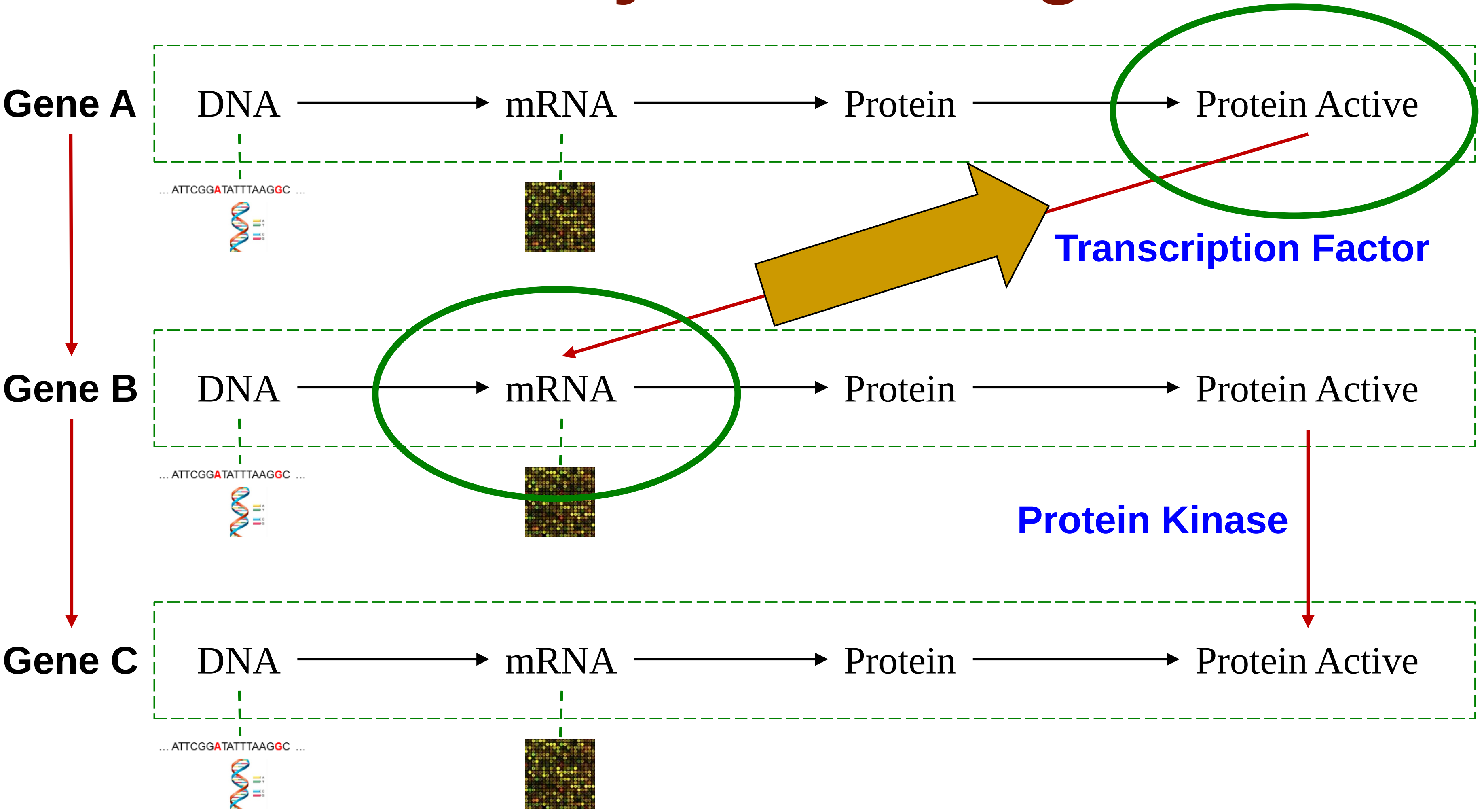
?



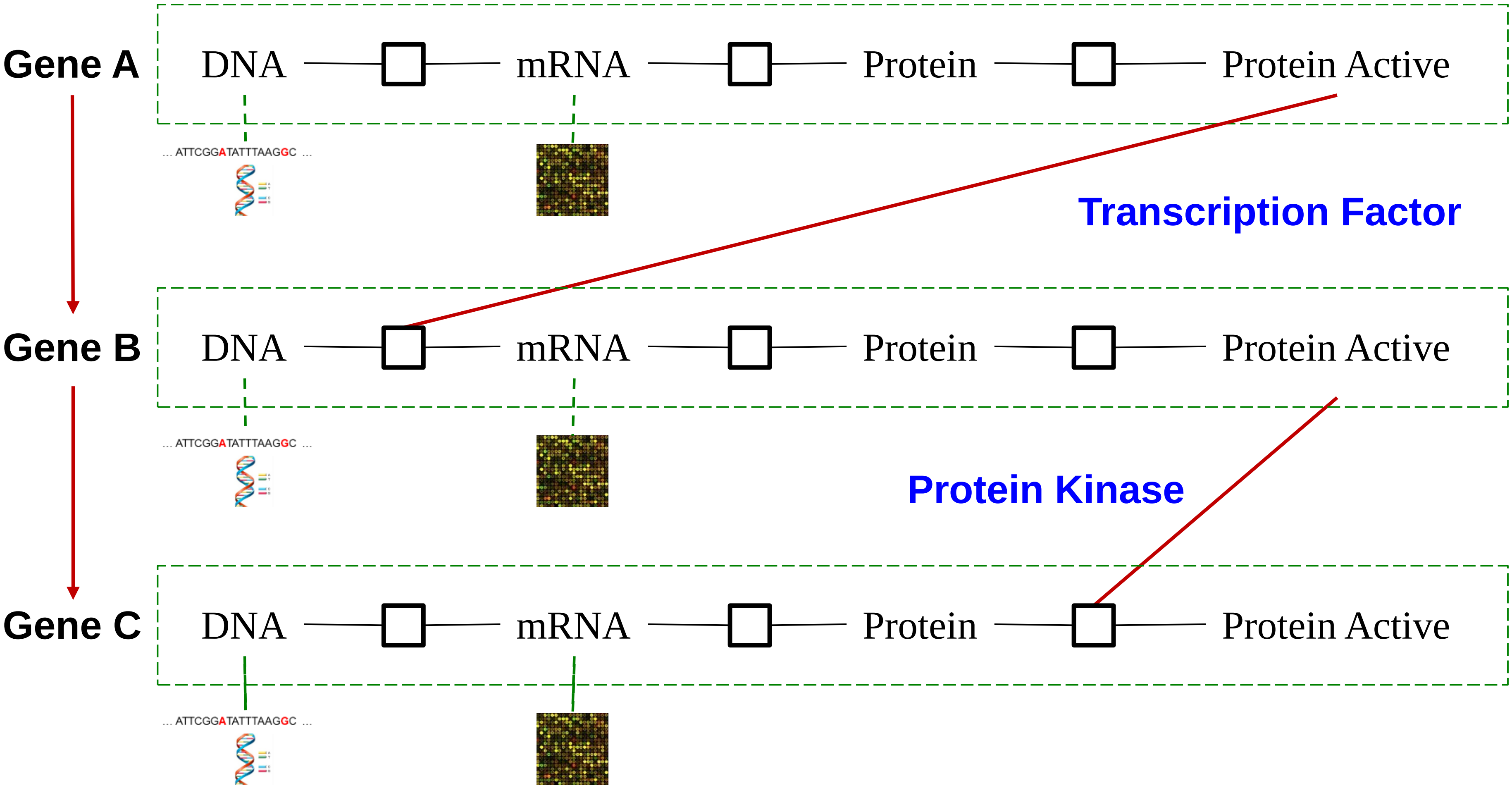
Pathway Knowledge ?



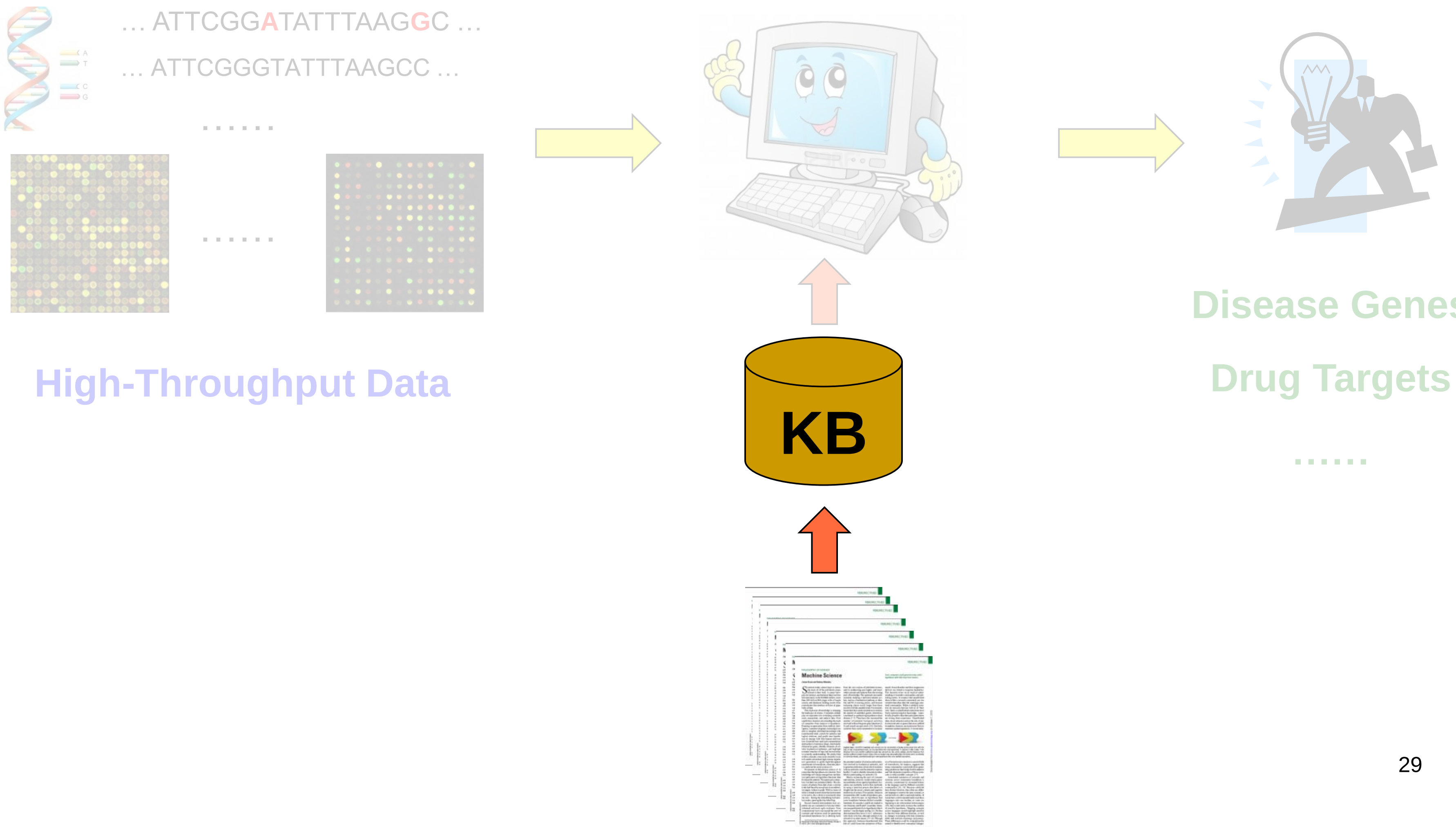
Pathway Knowledge !



Approach: Graph HMM

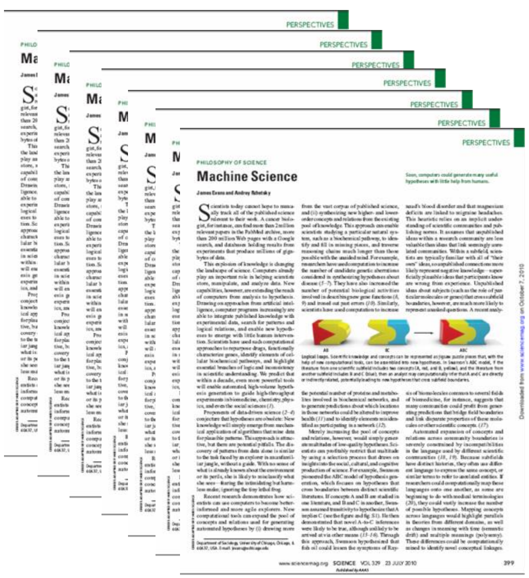


Extract Pathways from Pubmed



PubMed

- 22 millions abstracts
- Two new abstracts every minute
- Adds 2000-4000 every day


PERSPECTIVES

PHILOSOPHY OF SCIENCE

Machine Science

James Evans and Andrey Rzhetsky

Extract Pathways from Pubmed

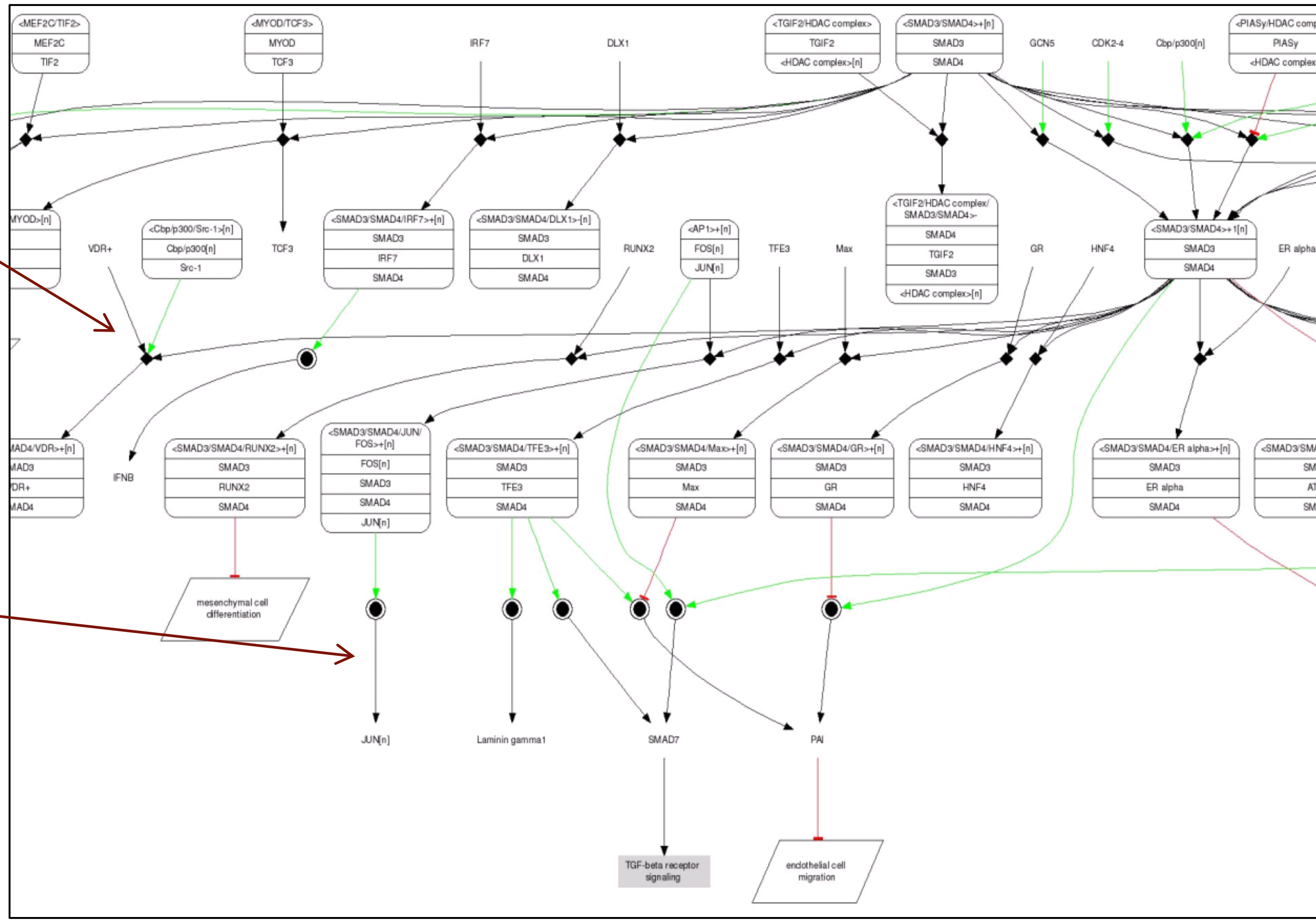
PMID: 123

…
VDR+ binds to SMAD3 to form
…

PMID: 456

…
JUN expression is induced by SMAD3/4
…

⋮



Extract Complex Knowledge

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

up-regulation

activation

IL-10

gp41

human monocyte

p70(S6)-kinase

Extract Complex Knowledge

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 **REGULATION**

up-regulation **REGULATION**

activation **REGULATION**

IL-10
PROTEIN

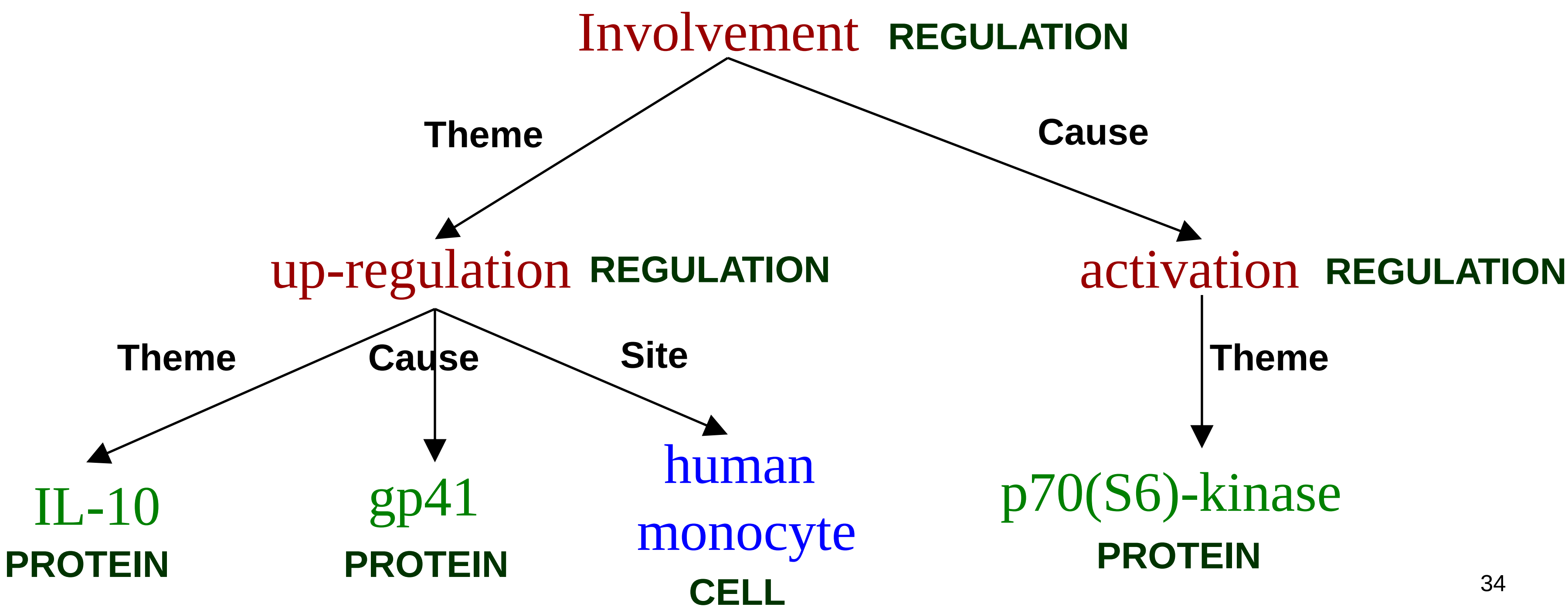
gp41
PROTEIN

human monocyte
CELL

p70(S6)-kinase
PROTEIN

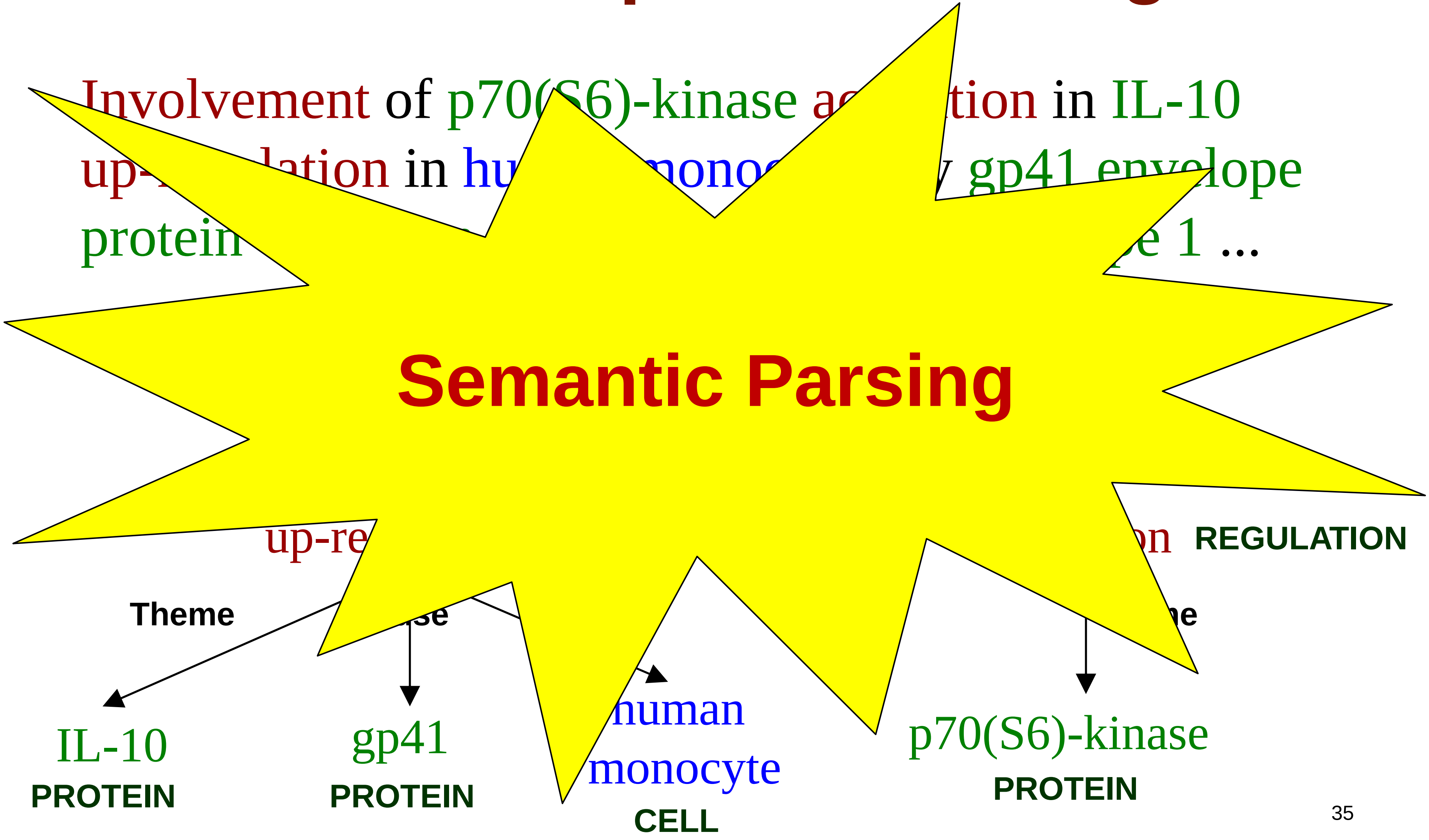
Extract Complex Knowledge

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



Extract Complex Knowledge

Involvement of p70(S6)-kinase ac tion in IL-10 up- lation in hu mono y gp41 envelope protein pe 1 ...



Bottleneck: Annotated Examples

- GENIA (BioNLP Shared Task 2009-2013)
  - 1999 abstracts
  - MeSH: human, blood cell, transcription factor
- Can we breach the annotation bottleneck?

Free Lunch #1: Distributional Similarity

- Similar context $\rightarrow$ Probably similar meaning
- Annotation as latent variables

  Textual expression $\rightarrow$ Recursive clusters
- Unsupervised semantic parsing

Poon & Domingos, “Unsupervised Semantic Parsing”. EMNLP-2009 (Best Paper Award).

Problem Formulation

Dependency tree d Semantic parse z

Probability $P_\theta(d, z)$

Parsing $z^* = \arg\max_z \log P_\theta(d, z)$

Learning $\theta^* = \arg\max_\theta \sum_d \log \sum_z P_\theta(d, z)$

Prior: Favor fewer parameters

Free Lunch #2: Existing KBs

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

  Textual expression → Table, column, join, …
- Grounded unsupervised semantic parsing

Poon, "Grounded Unsupervised Semantic Parsing". ACL-13.

Natural-Language Interface to Database

Get flight from Toronto to San Diego stopping at DTW

SELECT flight.flight_id
FROM flight, city, city c2, flight_stop, airport_service, airport_service as2
WHERE flight.from_airport = airport_service.airport_code AND flight.to_airport = as2.airport_code AND airport_service.city_code = city.city_code AND as2.city_code = city2.city_code AND city.city_name = ‘toronto’ AND city2.city_name = ‘san diego’ AND flight_stop.flight_id = flight.flight_id AND flight_stop.stop_airport = ‘dtw’

Answers

Clusters = KB Elements

- Entity: Table, Column, Cell
- Relation: Relational join
- **Priors:**
  - Favor lexical similarity
  - Favor short relational joins

GUSP: Key Ideas

- **Leverage target database**

JOB

Job ID	Company	System
001	IBM	Unix
002	Roche	IBM
003	Microsoft	Windows

⋮

Bootstrap learning with lexical prior

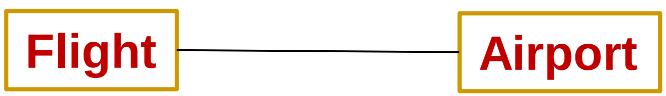
Prior: Favor Unix → System

GUSP: Key Ideas

- **Leverage target database**

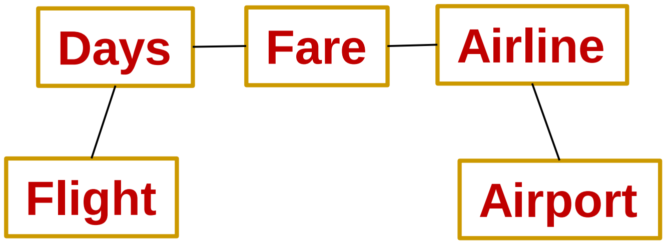
GUSP: Key Ideas

- **Leverage target database**



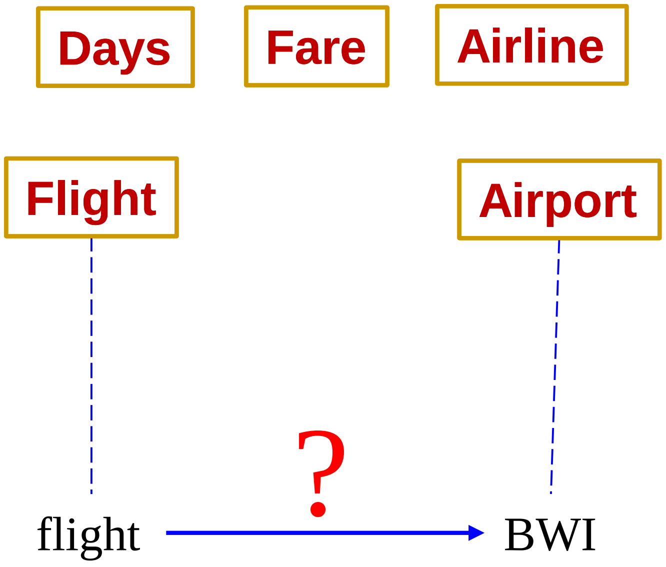
GUSP: Key Ideas

- **Leverage target database**



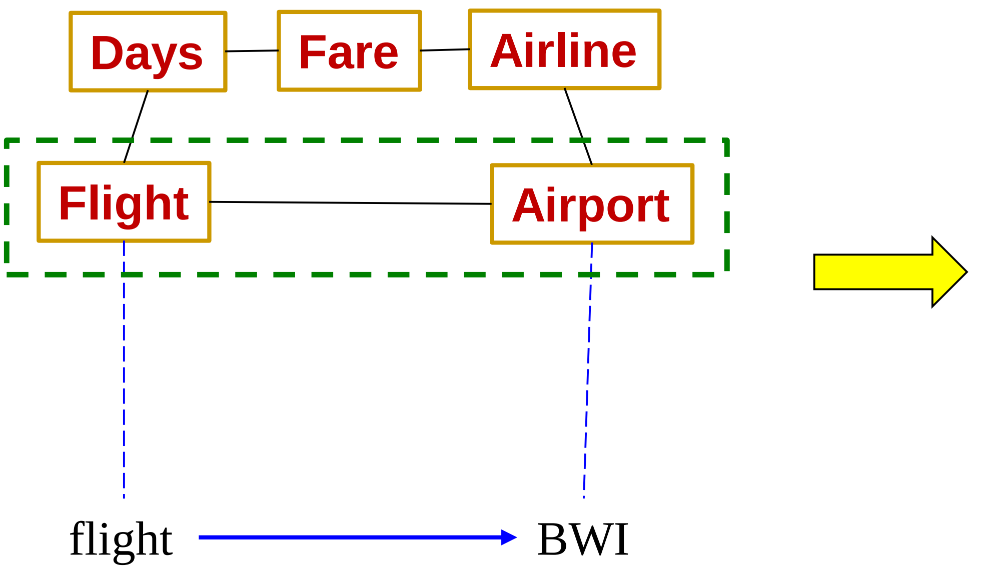
GUSP: Key Ideas

- **Leverage target database**



GUSP: Key Ideas

- **Leverage target database**



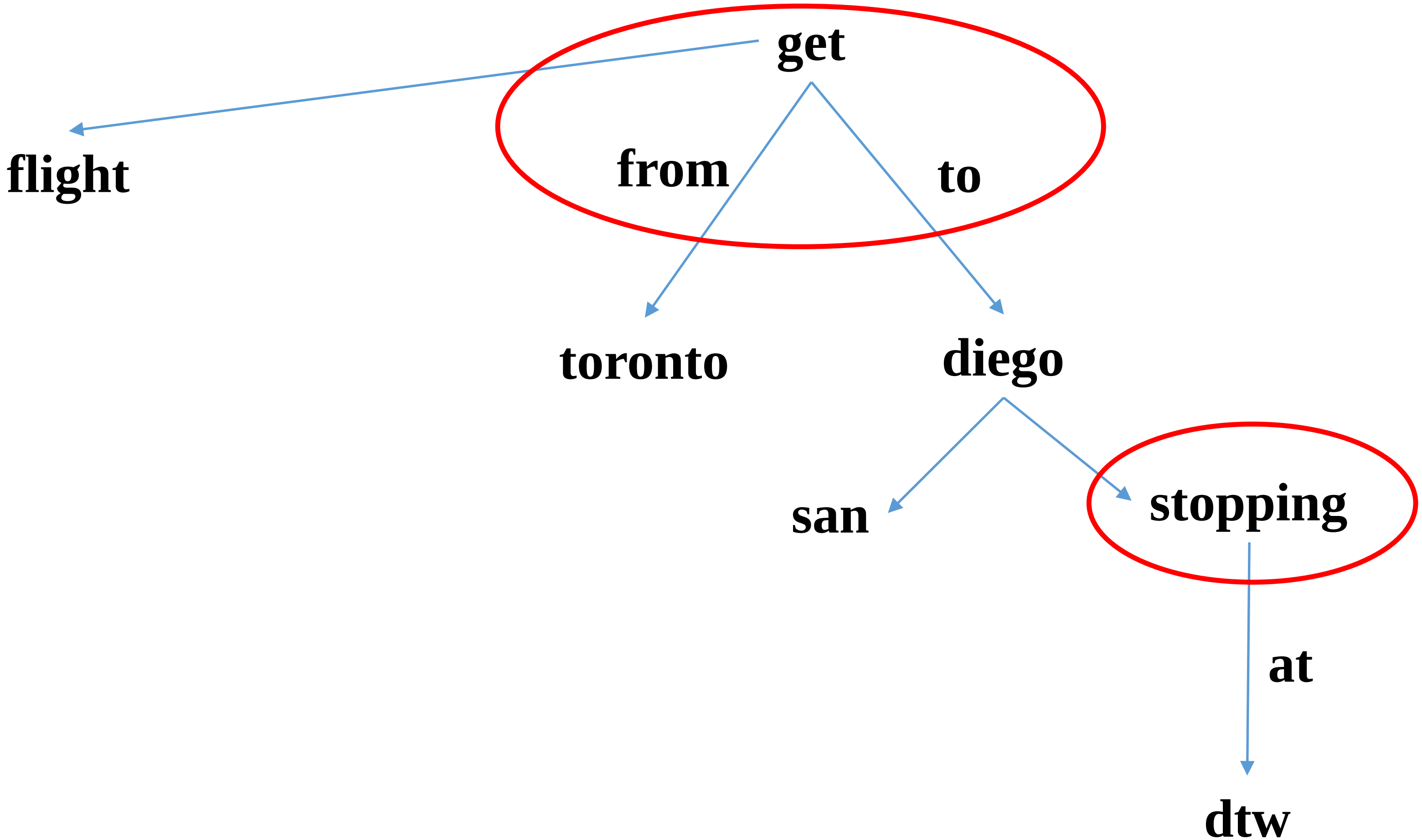
Leverage schema to guide learning

Prior: Favor shorter join

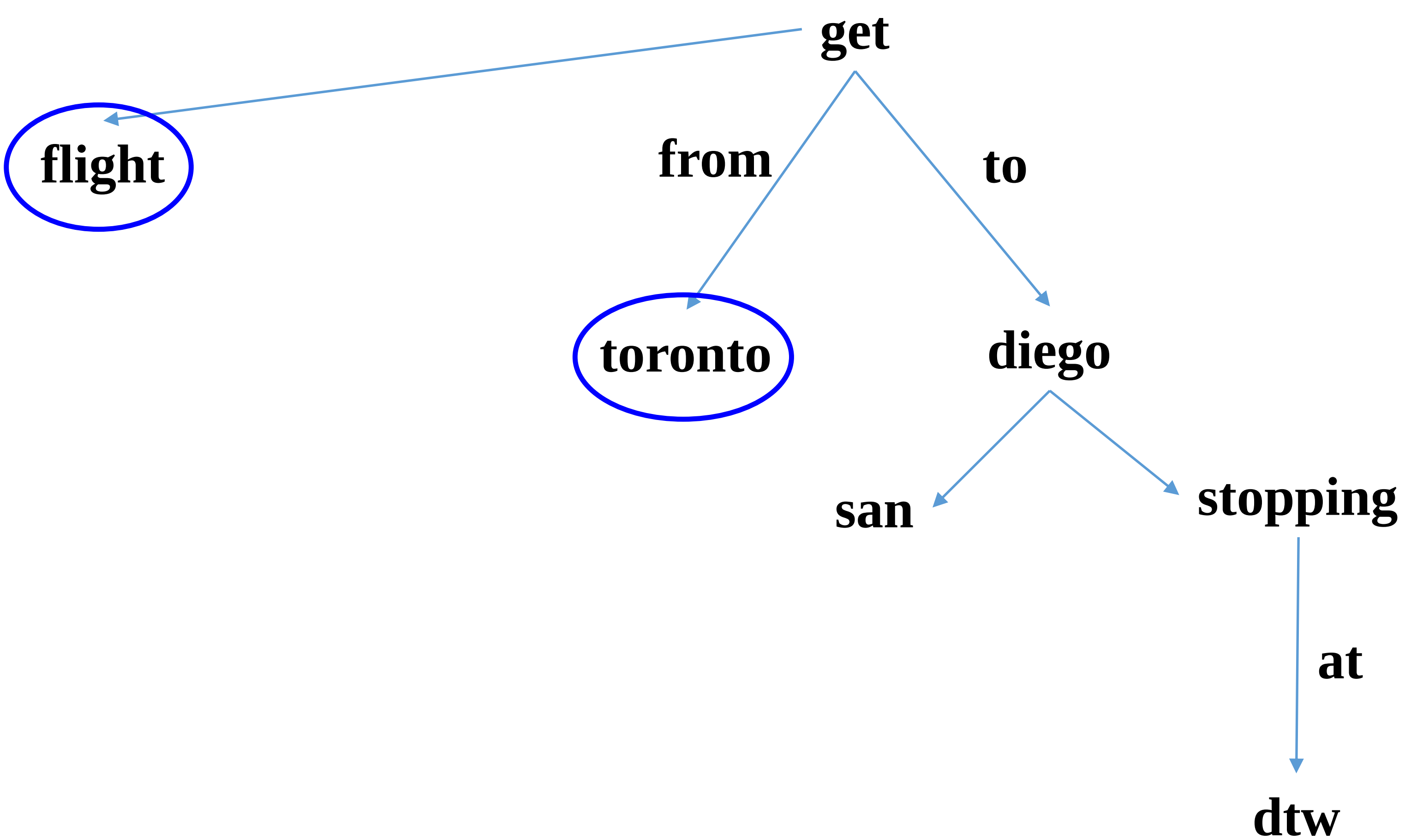
Free Lunch #3: Dependency Parses

- **Start from syntactic parse**
- Rich resources and available parsers
- Intractable structure learning $\rightarrow$ Tree HMM
- Exact inference is linear-time
- Need to handle syntax-semantics mismatch

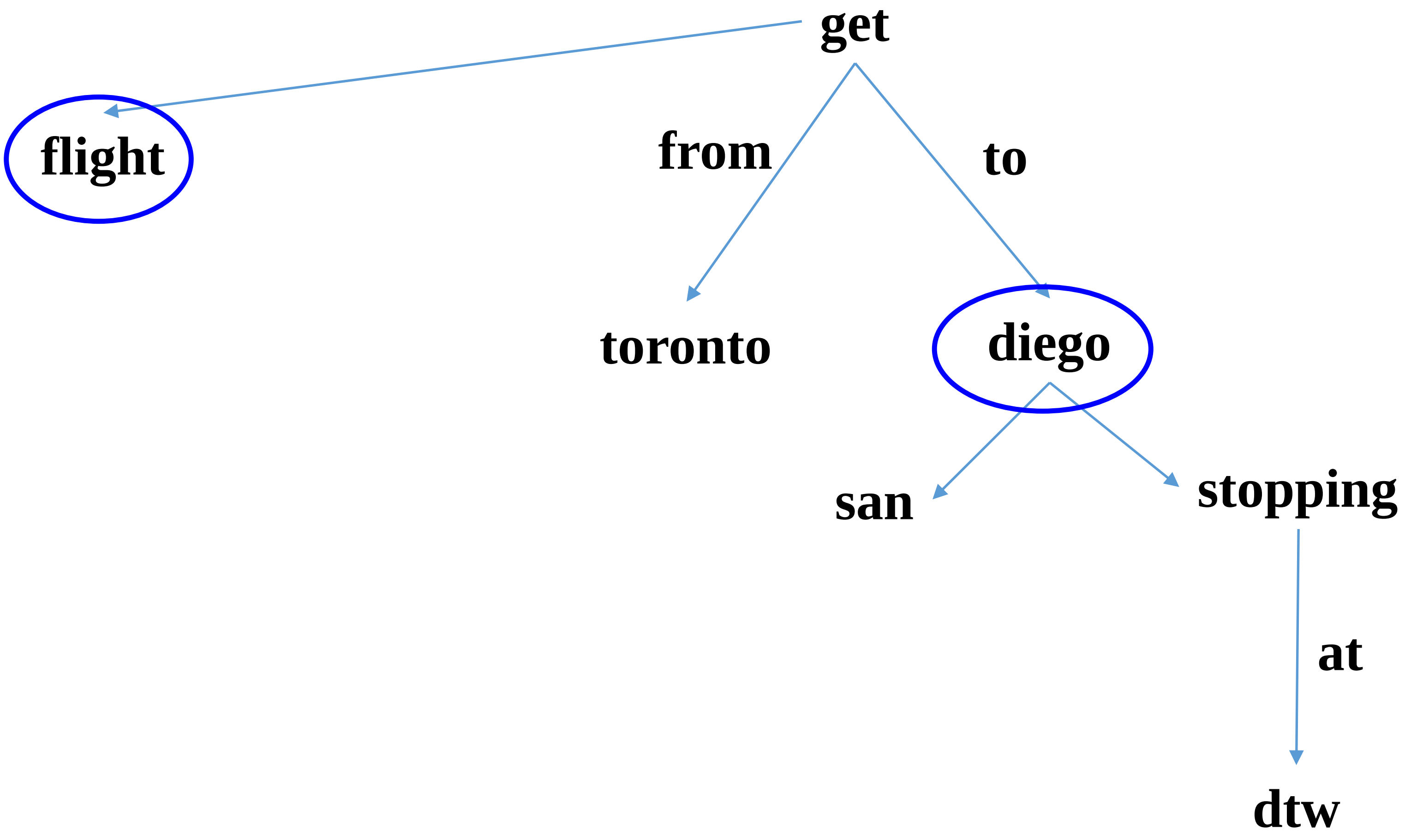
Syntax-Semantics Mismatch



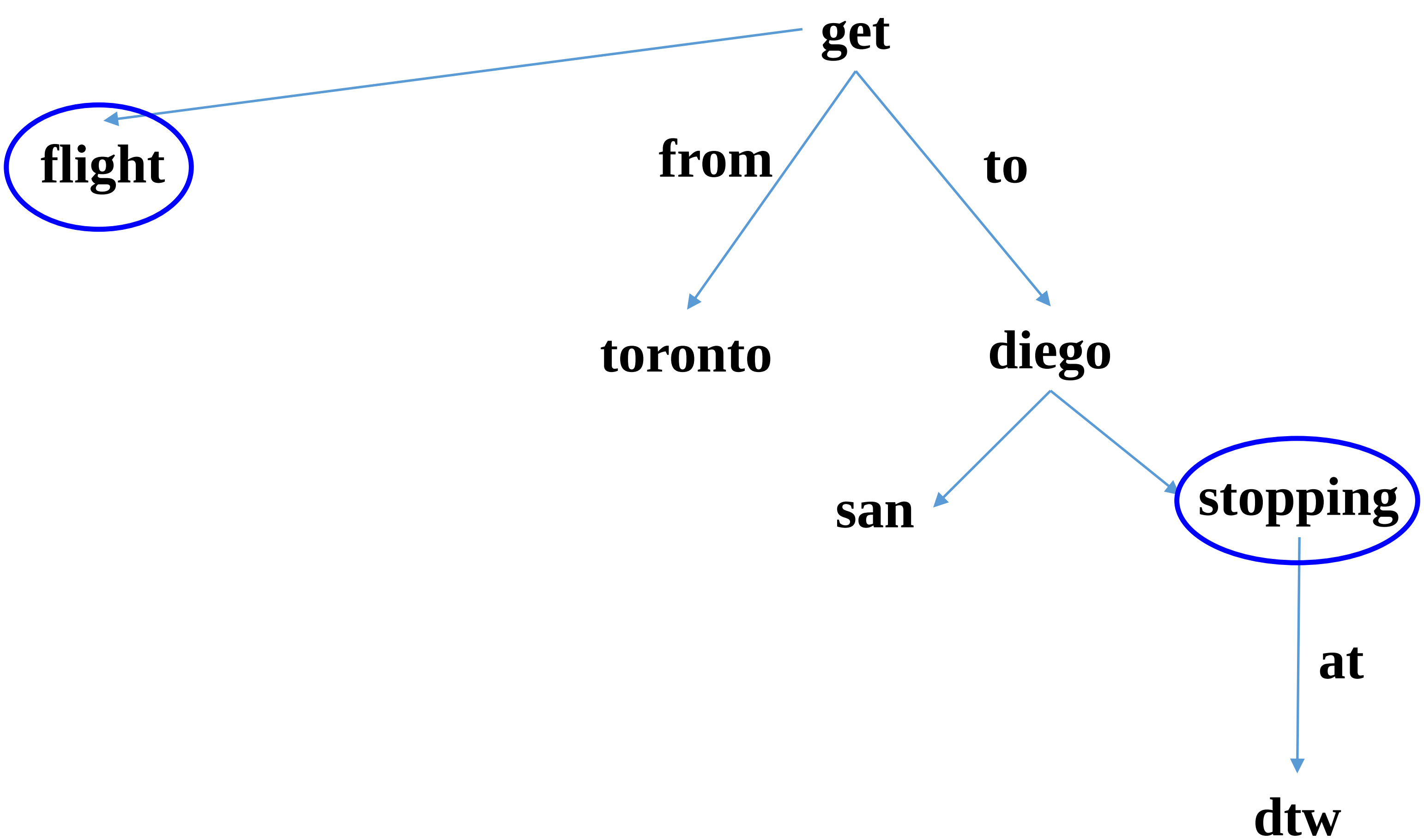
Syntax-Semantics Mismatch



Syntax-Semantics Mismatch



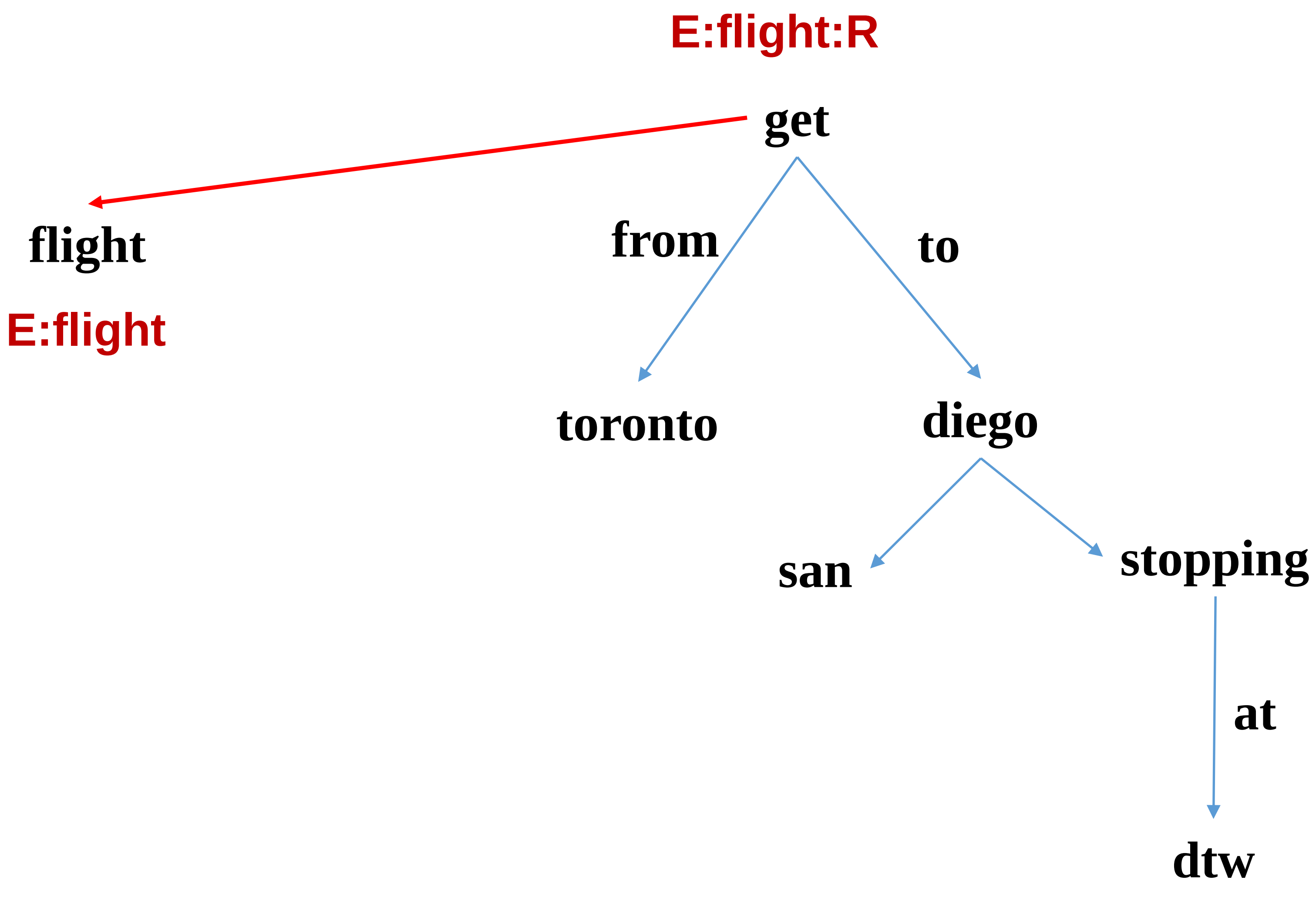
Syntax-Semantics Mismatch



Introduce Complex States

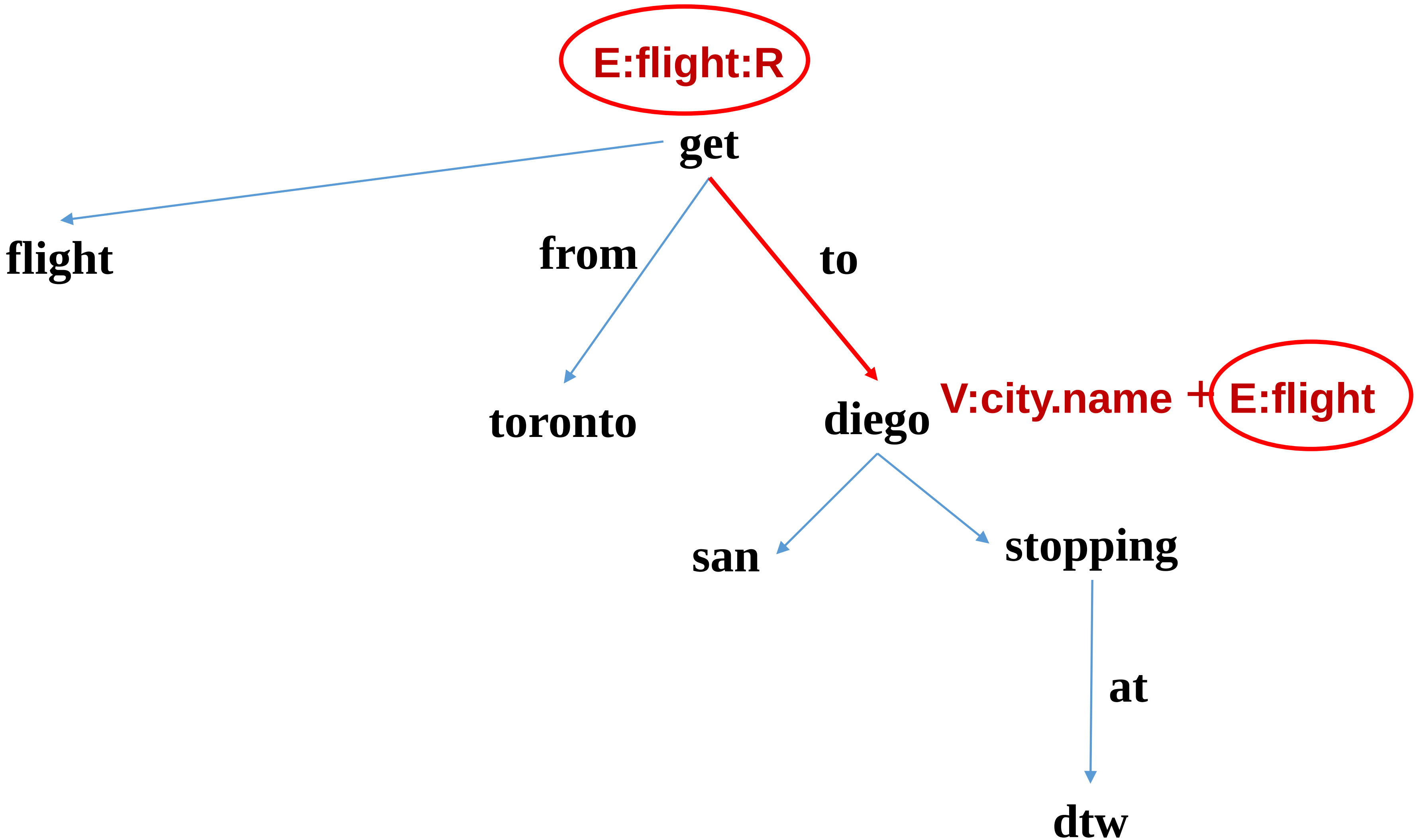
- Raising
- Sinking
- Implicit

Raising



E:flight:R
get
flight
E:flight
from
to
toronto
diego
san
stopping
at
dtw

Sinking



E:flight:R
get
flight
from
to
toronto
diego
V:city.name +
E:flight
san
stopping
at
dtw

Implicit

Give me the fare (of the flight) from Seattle to Boston

fare ⟹ fare

E:fare ⟹ **E:fare + E:flight**

Experiment: Dataset

- ATIS
  - Questions and ATIS database
  - Dev. / Test: Follow ZC07 [Zettlemoyer & Collins 2007]
  - Gold SQLs: Use at evaluation only
  - Gold logical forms in ZC07: Not used
- Evaluate on question-answering accuracy

Experiment: Systems

- **LEXICAL**: Lexical-trigger prior only
- Supervised learning
  - **ZC07**: Zettlemoyer & Collins [2007]
  - **FUBL**: Kwiatkowski et al. [2011]
- **GUSP–SIMPLE**: Simple states only
- **GUSP**++: All states

Results

System	Accuracy
ZC07	84.6
FUBL	82.8
GUSP++	83.5

Ablation

System Variant	Accuracy
LEXICAL	33.9
GUSP–SIMPLE	66.5
GUSP++	83.5
– Raising	75.7
– Sinking	77.5
– Implicit	76.2

Pathway Extraction

- More to leverage from KB:

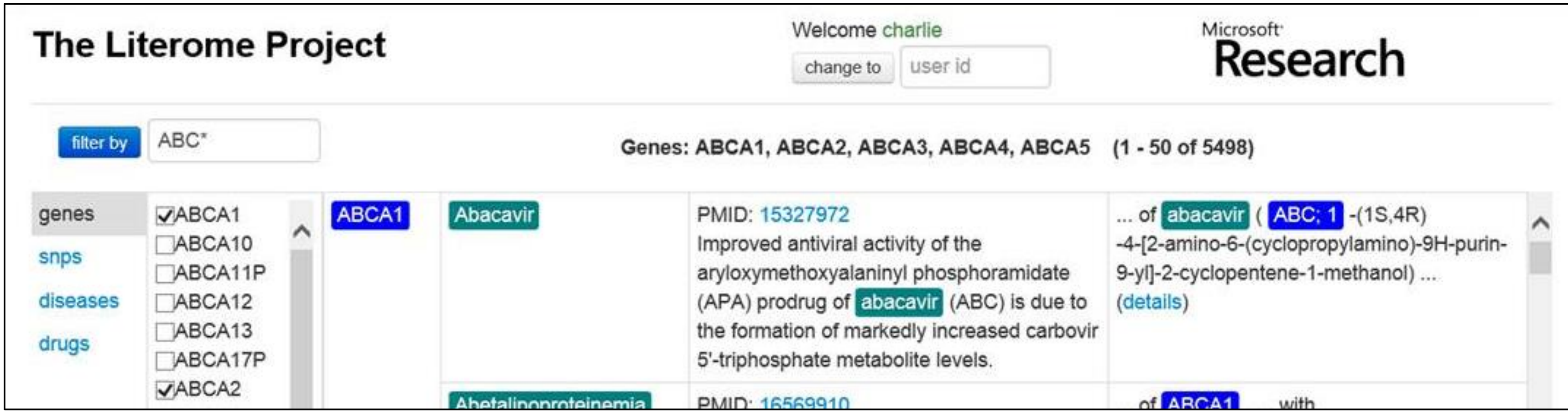
  Semantic relations in KB likely occur in semantic parse of some sentence
- **Priors:**
  - Favor a parse w. relations in KB
  - Penalize a parse w. relations not in KB

Distant-Supervision

- Existing work: Binary relation, classification
  - Mintz et al. [2009]
  - Riedel et al. [2010]
  - Hoffmann et al. [2011]
  - Krishnamurphy & Mitchell [2012]
  - Etc.
- Our approach: Generalize distant supervision to semantic parsing

Parikh, Poon, Toutanova. In progress.

Literome



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

http://literome.azurewebsites.net

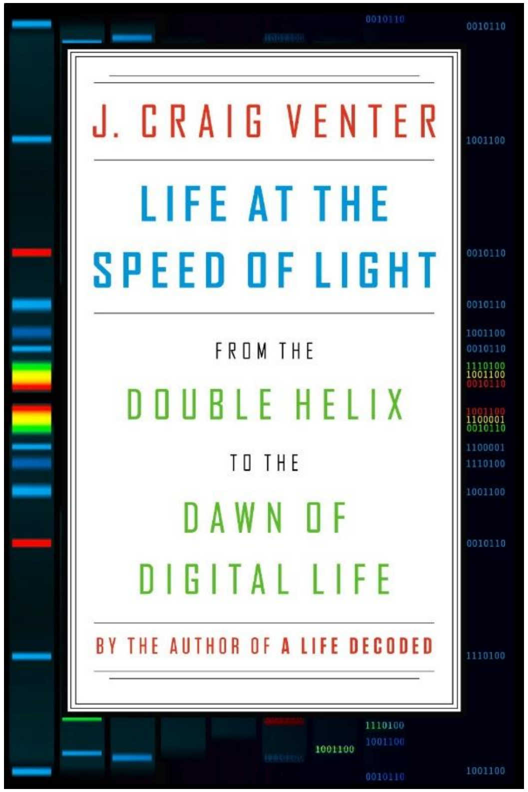
PubMed-Scale Extraction

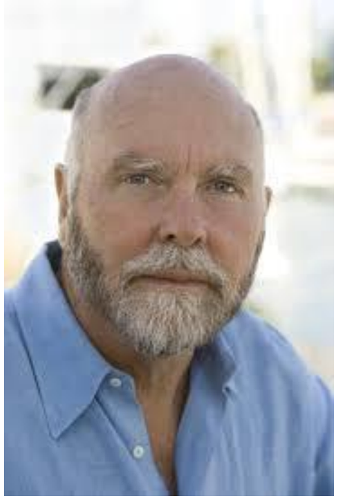
- Preliminary pass:
  - 2 million instances
  - 13,000 genes, 870,000 unique interactions
- Applications:
  - UCSC Genome Browser, MSR Interactions Track
  - Cancer expression profile modeling
  - Validate *de novo* pathway prediction
  - Etc.

Big Mechanism

- 42-million program for 12 teams
  - Reading, Assembly, Explanation
  - Domain: Cancer signaling pathways
- We are funded
  - PI: Andrey Rzhetsky
  - Co-PI w. James Evans, Ross King

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

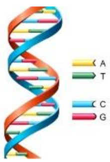
Summary

- Precision medicine is the future
- **Infer cancer driver mutations**

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

  Semantic parsing grounded in KBs
- **Literome**: KB for genomic medicine

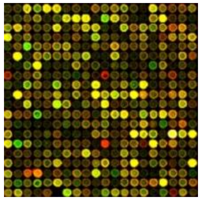
Summary



… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

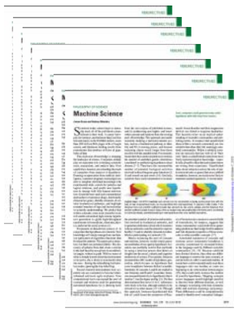
……



……

High-Throughput Data

KB



Disease Genes

Drug Targets

……