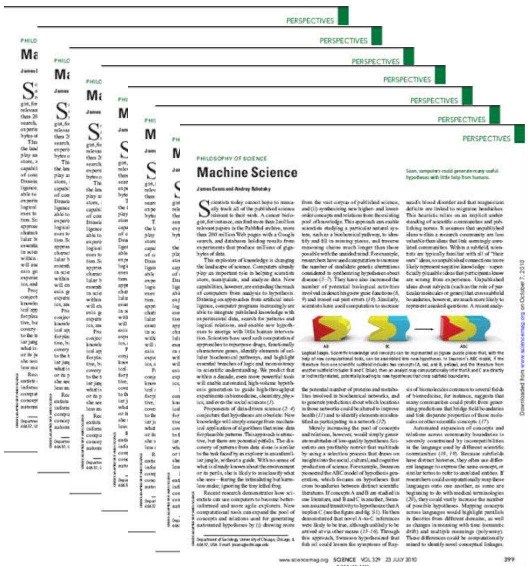
Machine Reading for Cancer Panomics

Hoifung Poon

Machine Reading



INDUCE(e1) ∧ IL-4(e2) ∧ CD11B(e3) ∧ INDUCER(e1,e2) ∧ INDUCED(e1,e3)

REGULATE

regulate, control, govern, modulate

ISA ISA ISA

induce, enhance, trigger, augment, up-regulate

INDUCE

inhibit, block, suppress, prevent, abolish, abrogate, down-regulate

INHIBIT

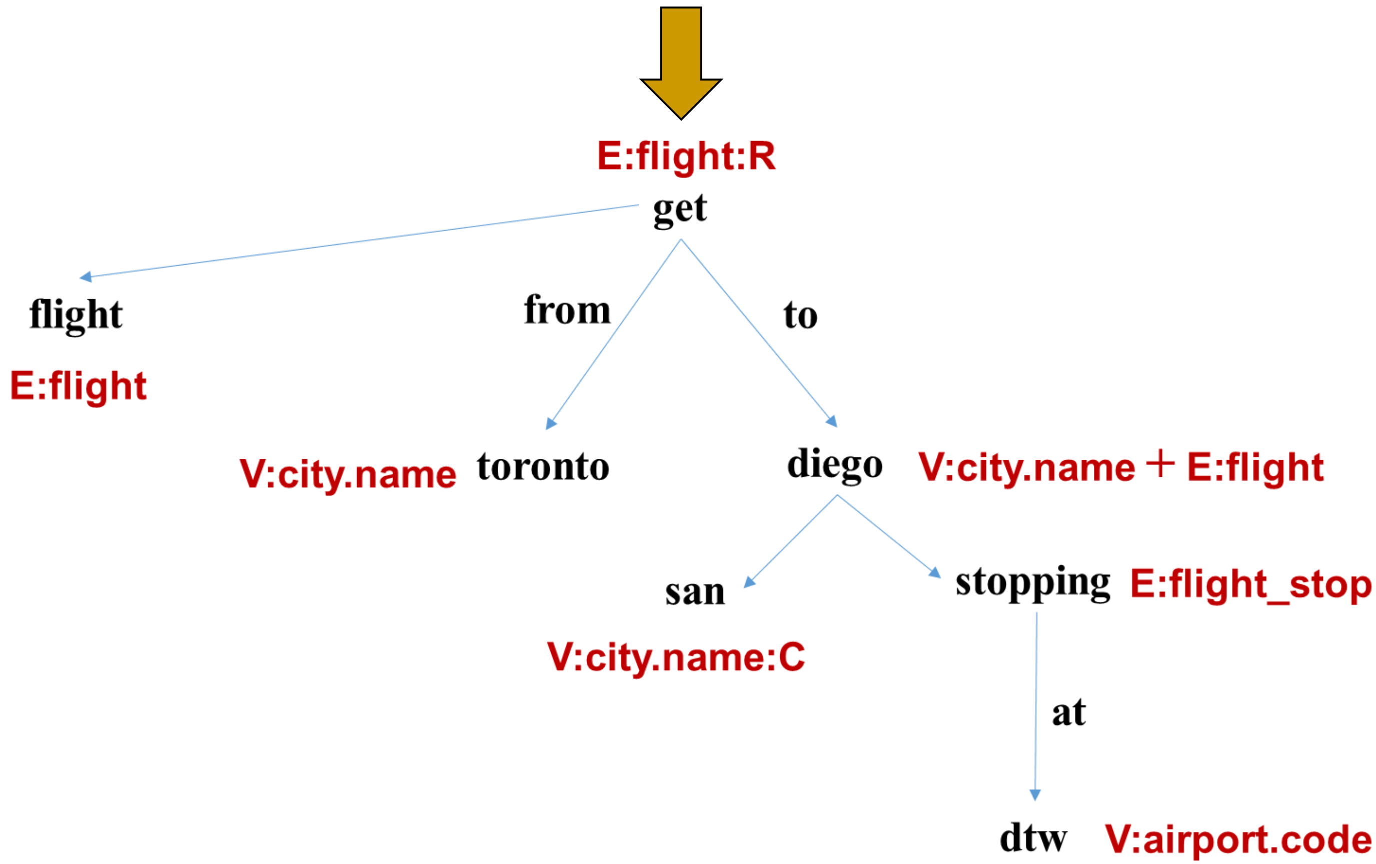
activate

ACTIVATE

⋮

Natural-Language Interface to Database

Get flight from Toronto to San Diego stopping at DTW



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’

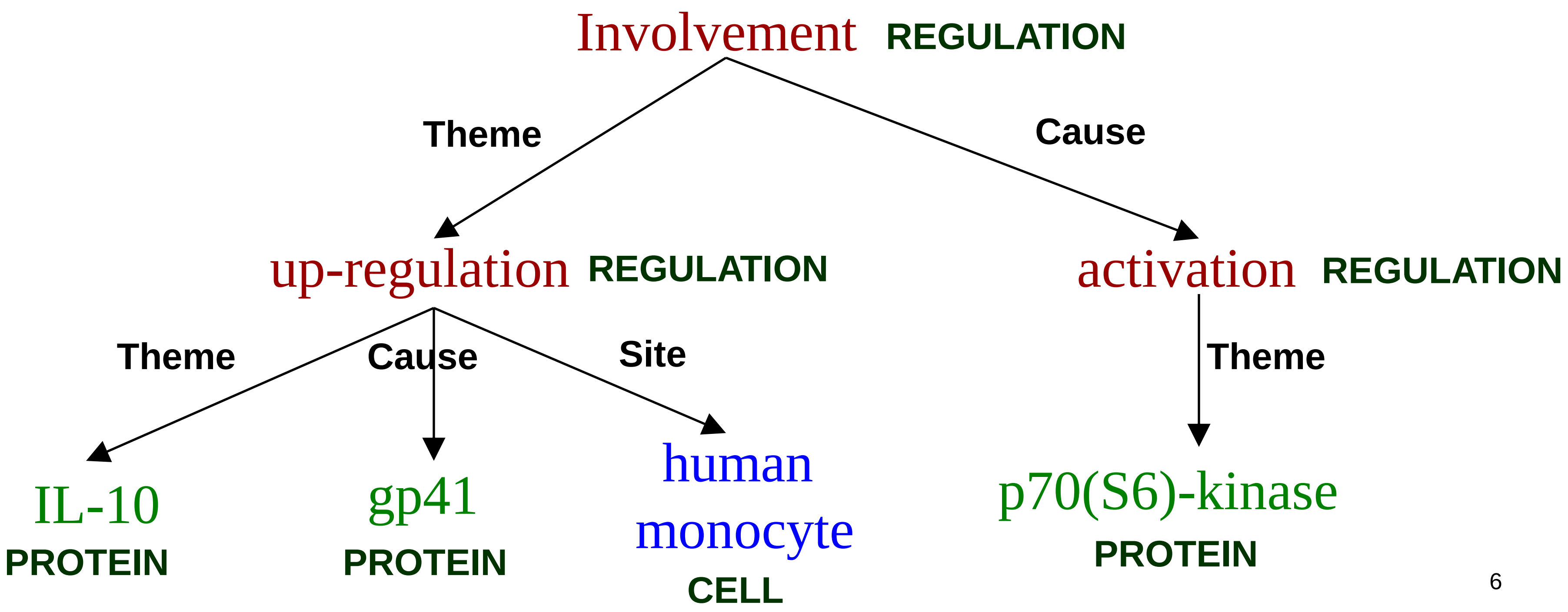
Answers

Extract Complex Knowledge

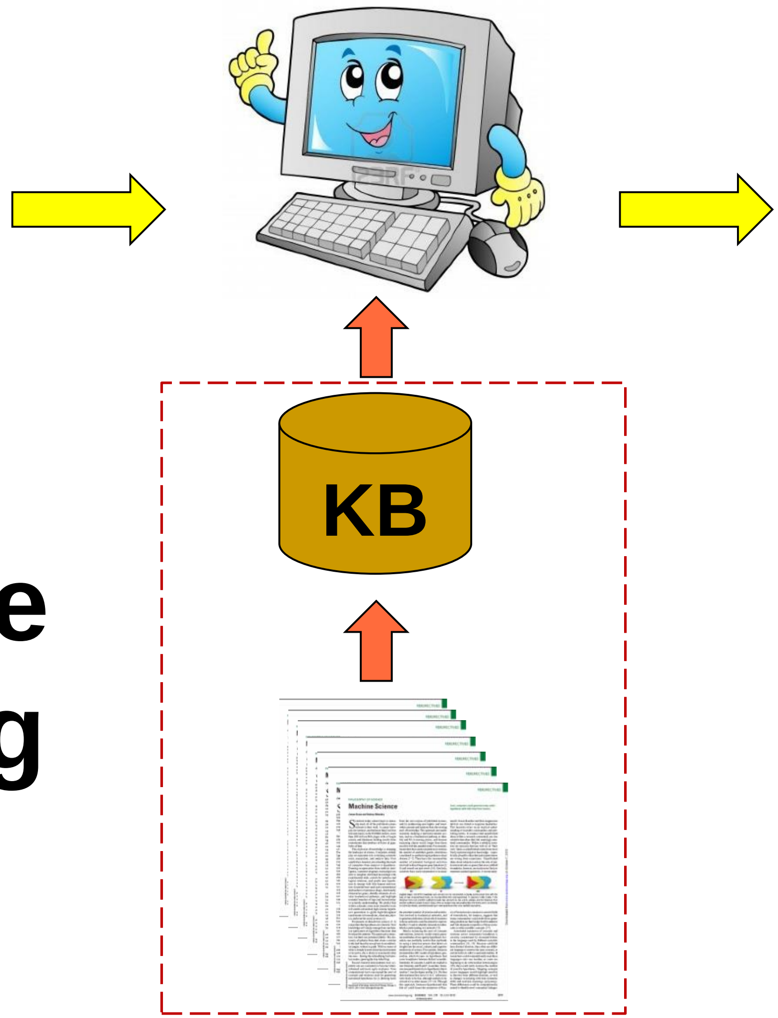
Involvement of p70(S6)-kinase activation in IL-10 up-regulation in human monocytes by gp41 envelope protein of human immunodeficiency virus type 1 ...

Extract Complex Knowledge

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



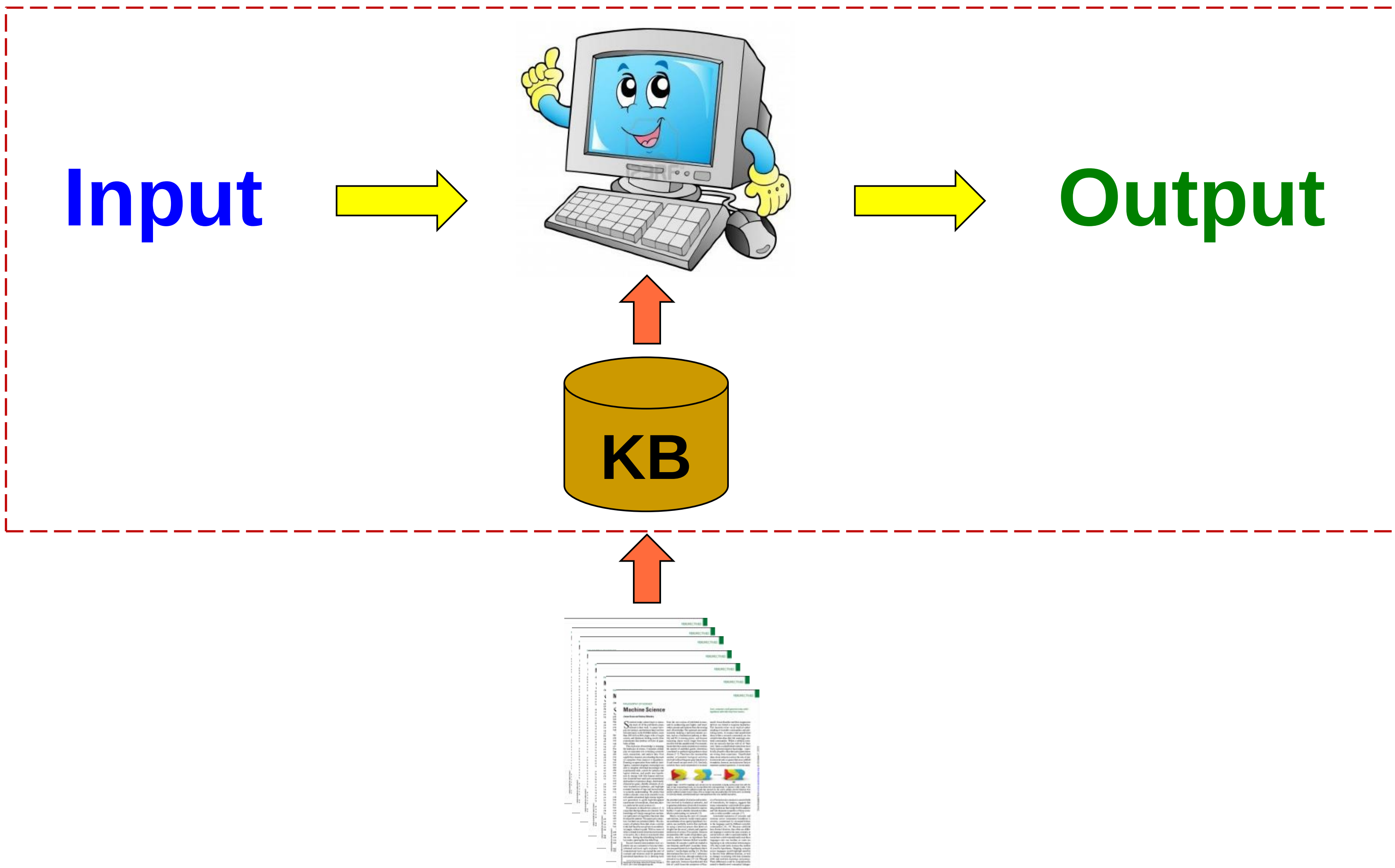
Input



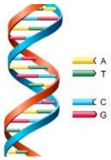
Output

Machine Reading

Knowledge-Rich Machine Learning



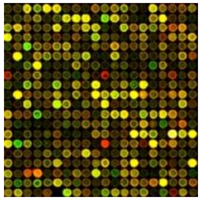
Overview



... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...

......



......

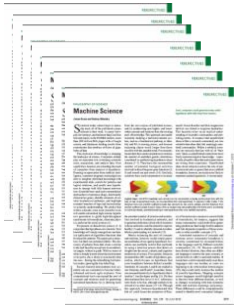
High-Throughput Data

Disease Genes

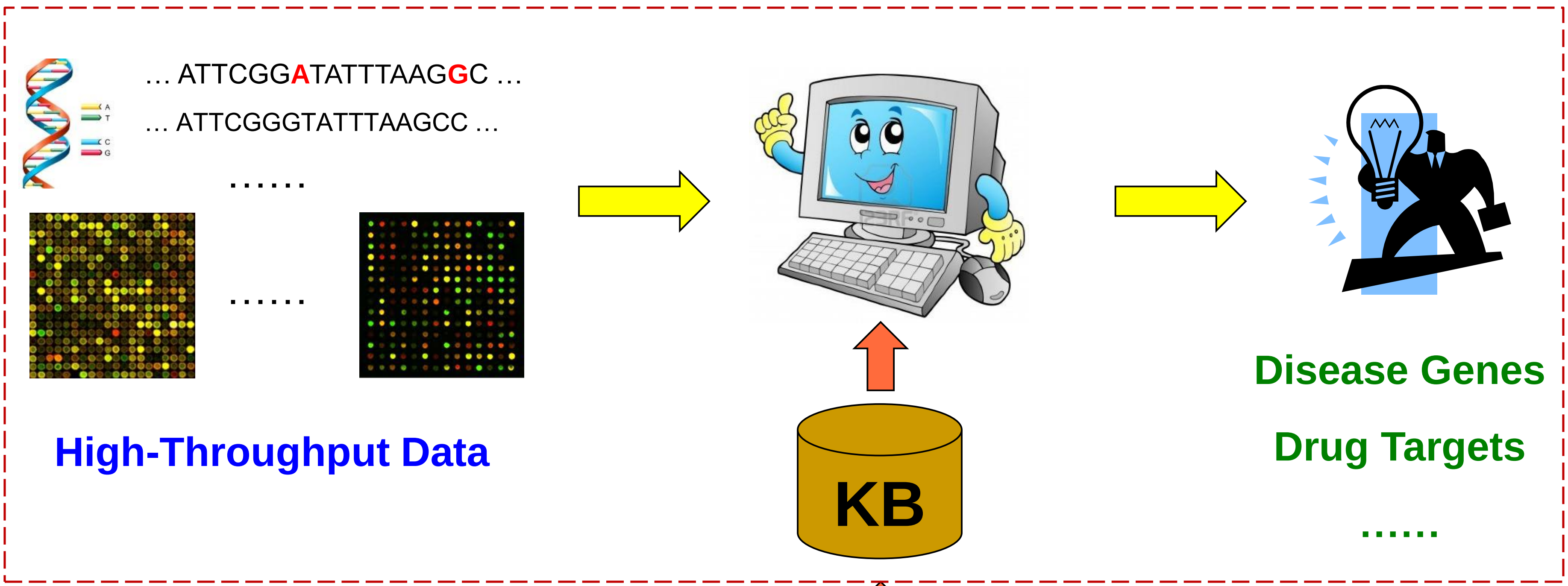
Drug Targets

......

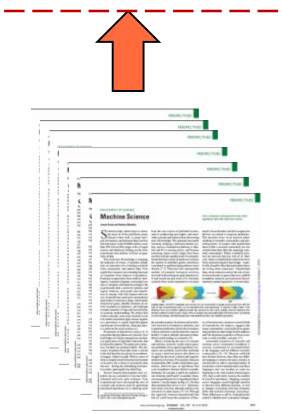
KB



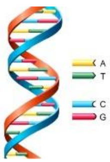
Overview



Infer cancer driver mutations



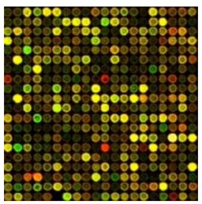
Overview



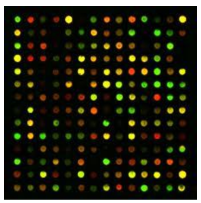
… ATTCGGATATTTAAGGC …

… ATTCGGGTATTTAAGCC …

……



……



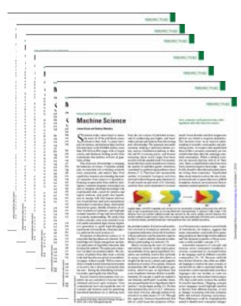
High-Throughput Data

Disease Genes

Drug Targets

…

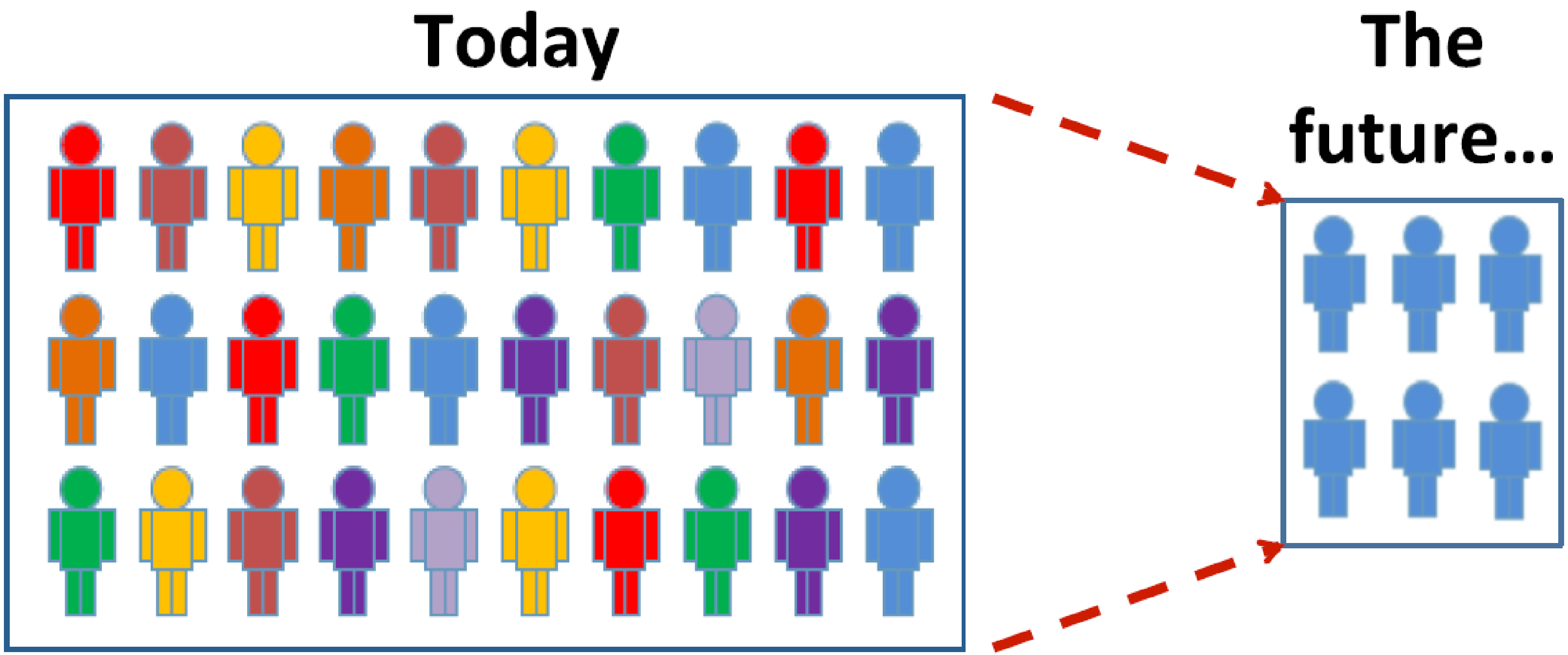
KB



Grounded Unsupervised Semantic Parsing

Extract Pathways from Pubmed

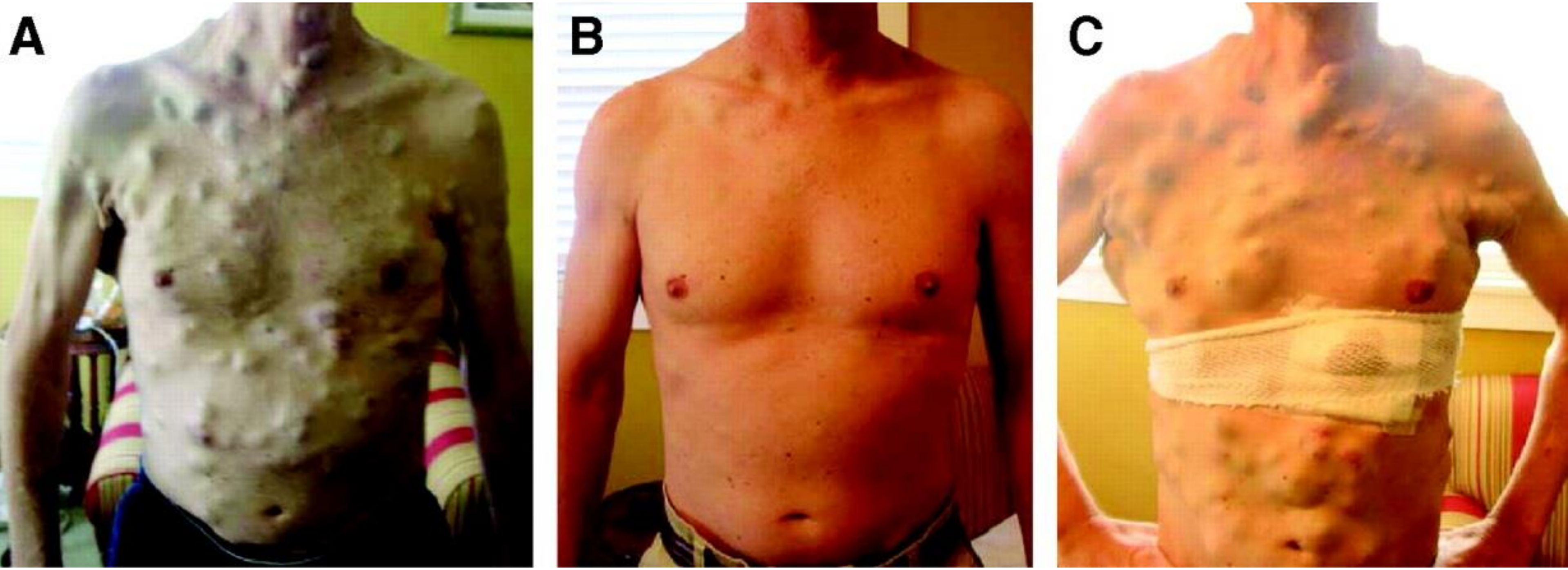
Precision Medicine



Vemurafenib on BRAF-V600 Melanoma

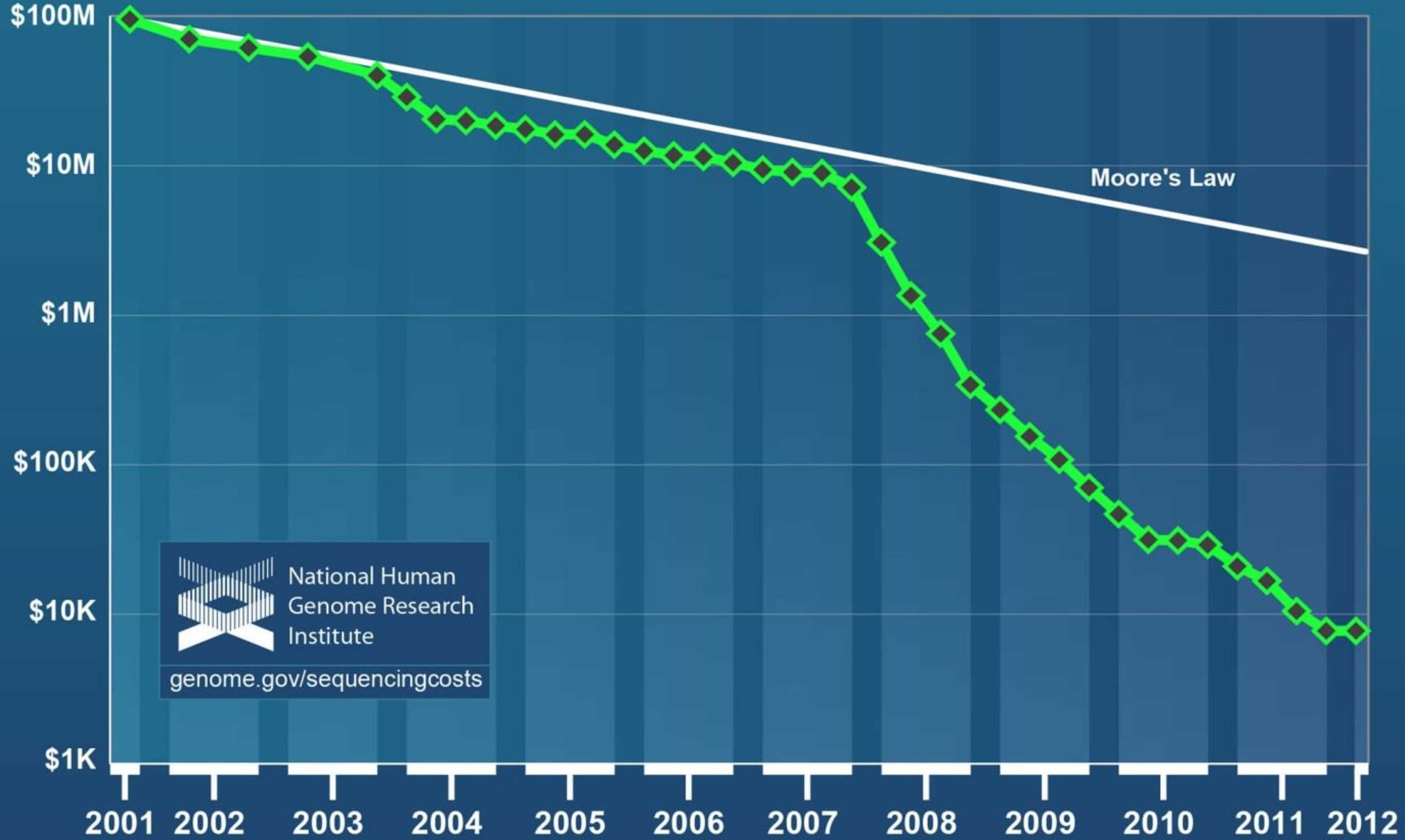
Before Treatment **15 Weeks**

Vemurafenib on BRAF-V600 Melanoma

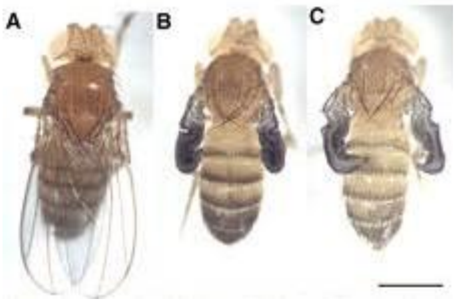


Before Treatment 15 Weeks 23 Weeks

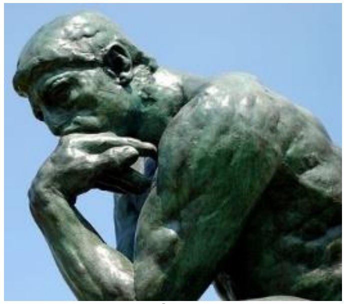
Cost per Genome



Traditional Biology



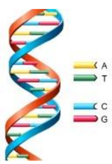
Targeted Experiments



One hypothesis

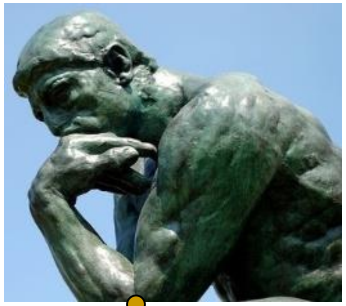
Discovery

Genomics



... ATTCGGATATTTAAGGC ...
... ATTCGGGTATTTAAGCC ...
... ATTCGGATATTTAAGGC ...
... ATTCGGGTATTTAAGCC ...
... ATTCGGATATTTAAGGC ...
... ATTCGGGTATTTAAGCC ...

High-Throughput Experiments

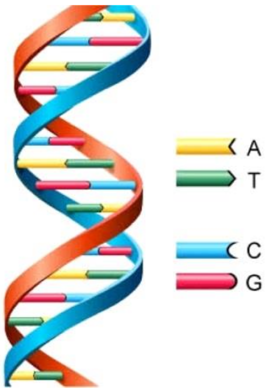


Many hypotheses

?

Discovery

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."

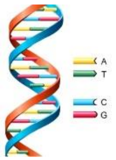
2010

"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 ...

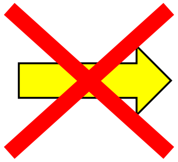
... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...

... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...

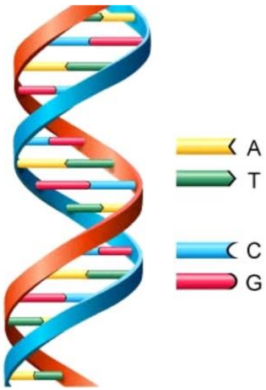
High-Throughput Experiments



Discovery

How to Scale Discovery?

Cancer



… ATTCGG**A**TATTTAAG**G**C …

Tumor cells

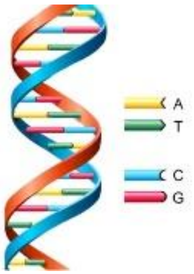
… ATTCGGGTATTTAAGCC …

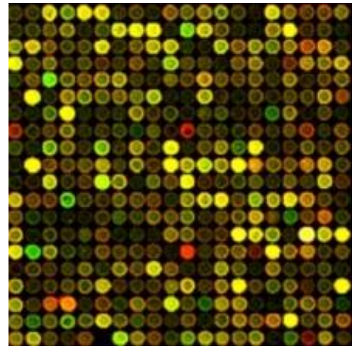
Normal cells

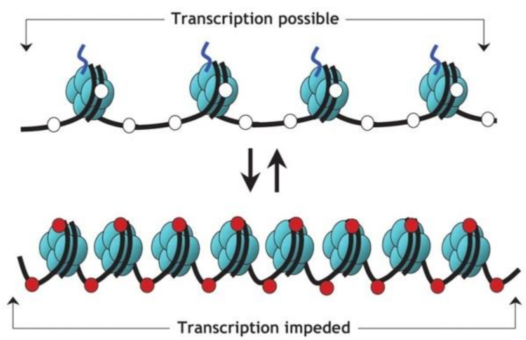
- Hundreds of mutations
- Most are “passenger”, not driver
- Can we identify likely drivers?

Panomics

… ATTCGG**A**TATTTAAG**G**C …





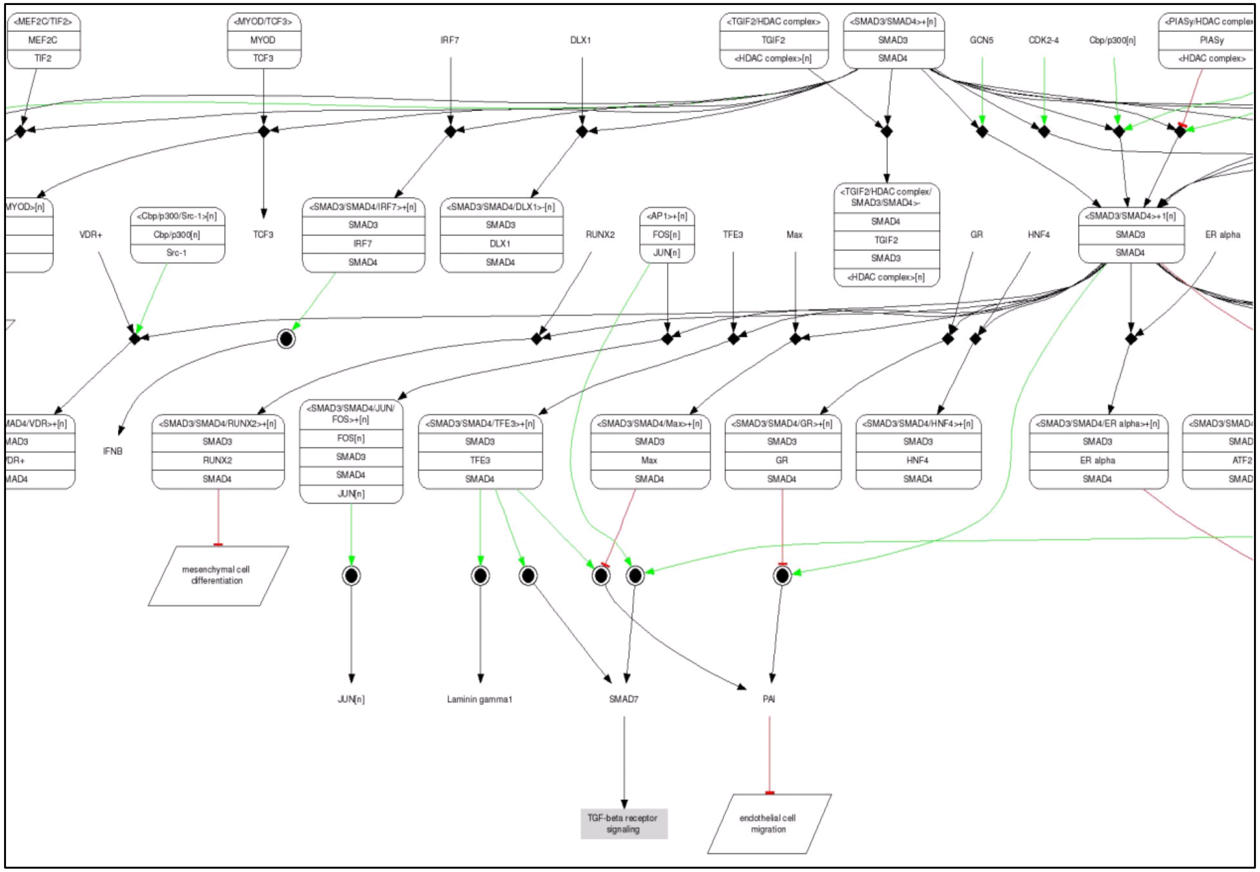


Genome **Transcriptome** **Epigenome**

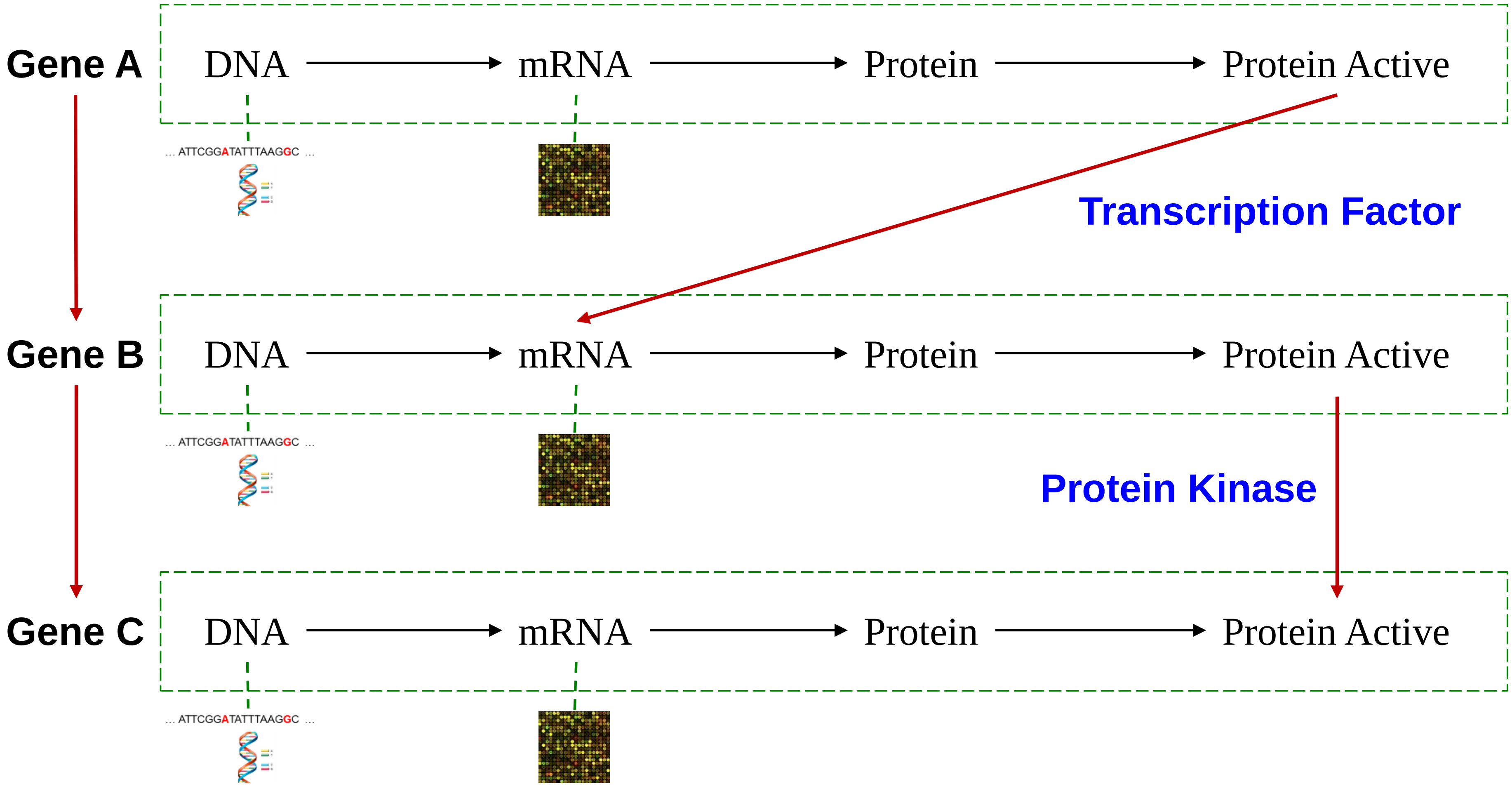
……

Pathway Knowledge

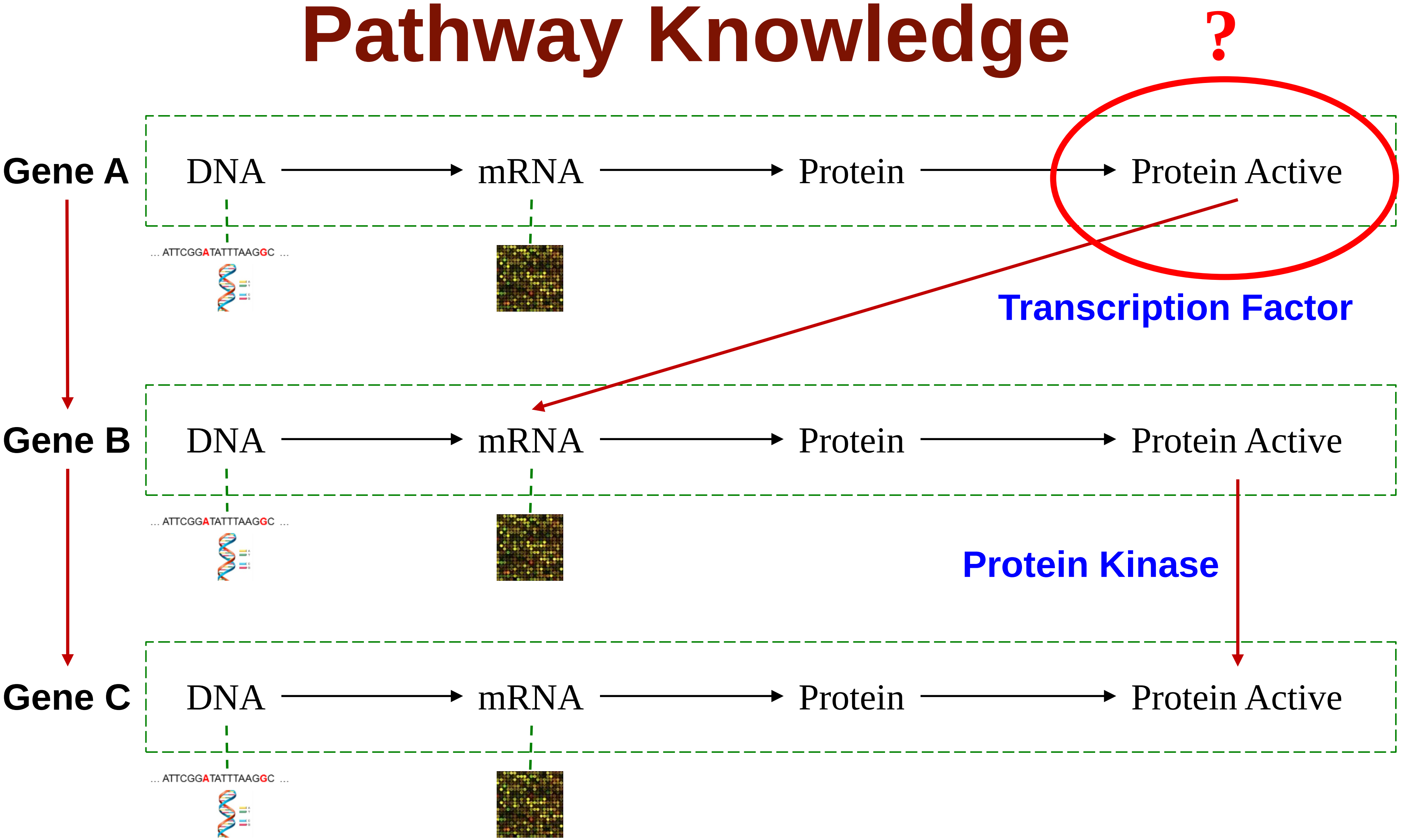
Genes work synergistically in pathways



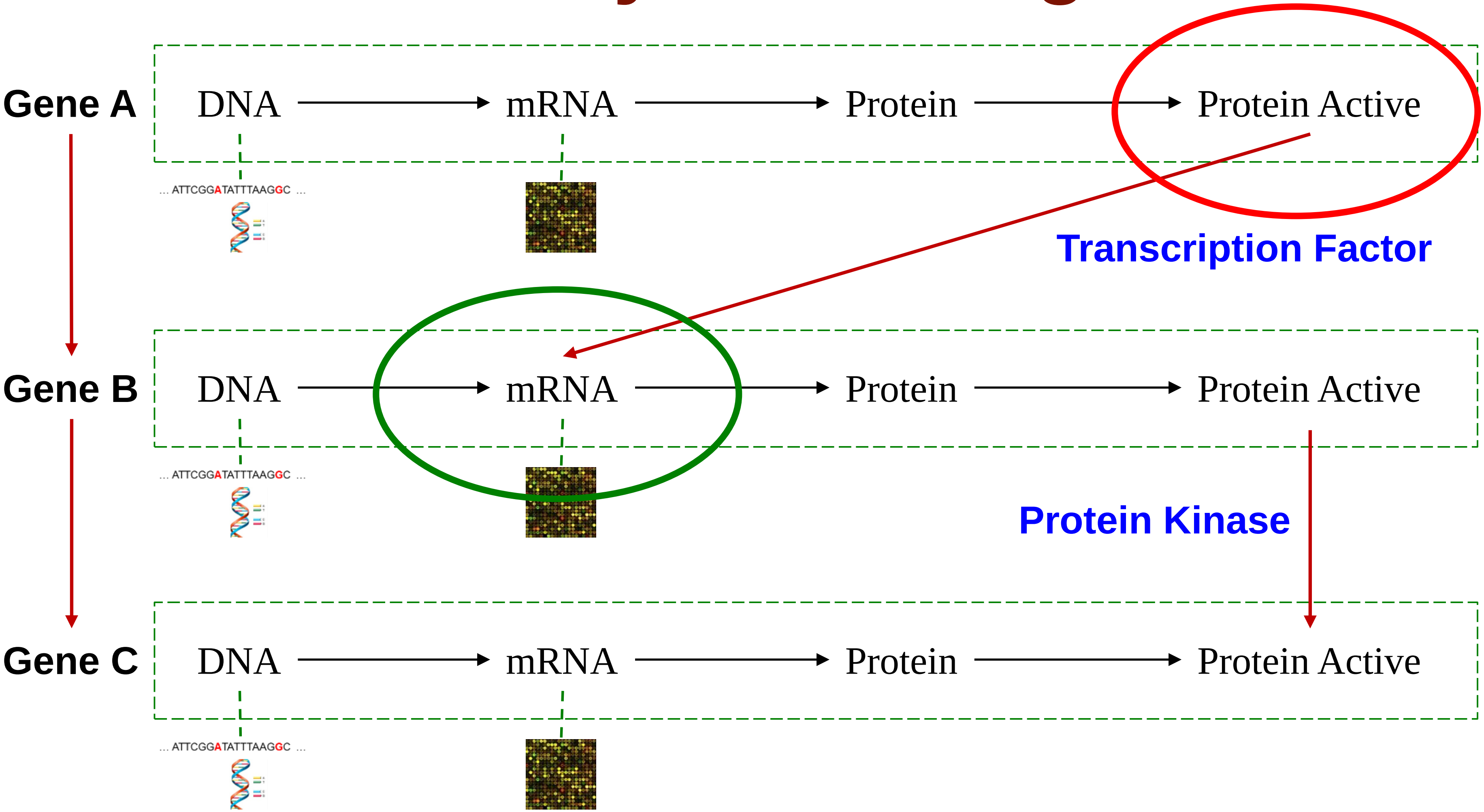
Pathway Knowledge



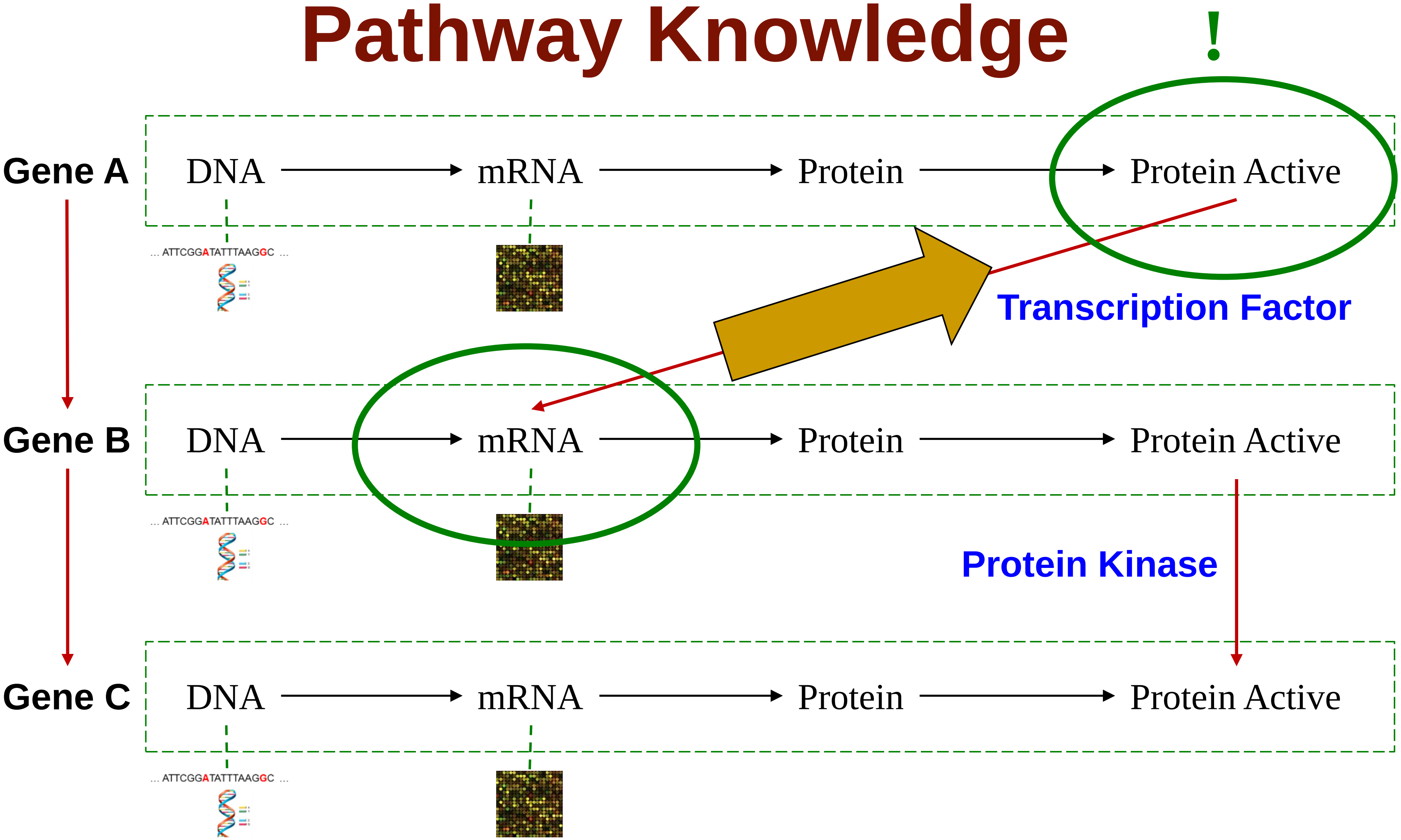
Pathway Knowledge



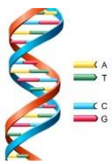
Pathway Knowledge ?



Pathway Knowledge



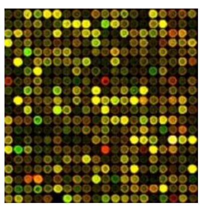
Infer Cancer Driver Mutations



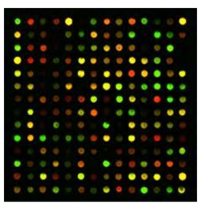
... ATTCGGATATTTAAGGC ...

... ATTCGGGTATTTAAGCC ...

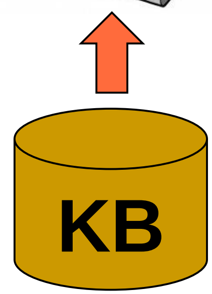
......



......



High-Throughput Data



Disease Genes

Drug Targets

......

Infer Cancer Driver Mutations

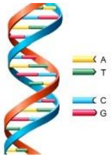
Transcription Translation Activation

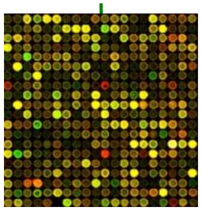
Gene A DNA → mRNA → Protein → Protein Active

What's the level of activity?

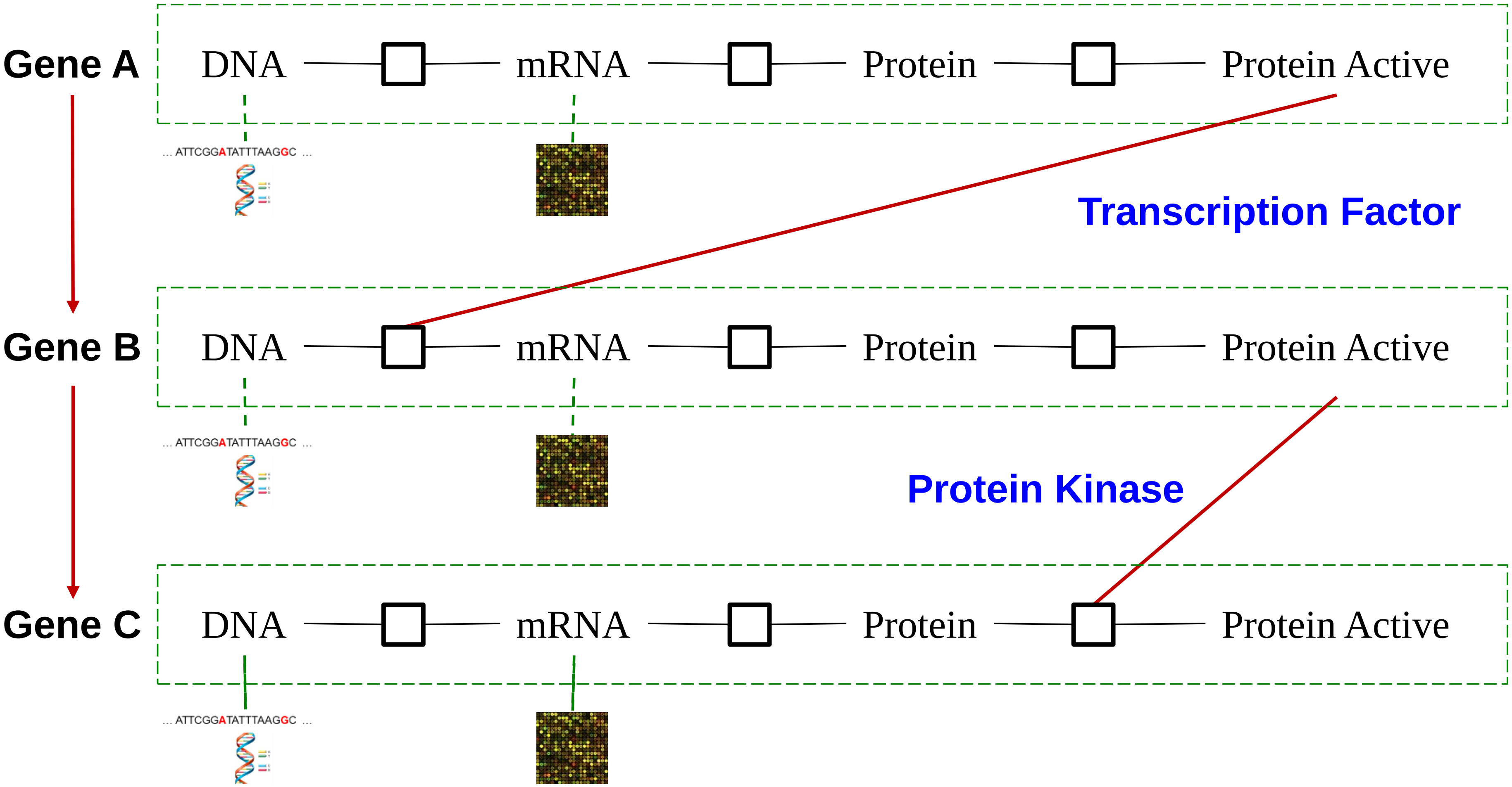
Is change caused by mutation?

… ATTCGGATATTTAAGGC …

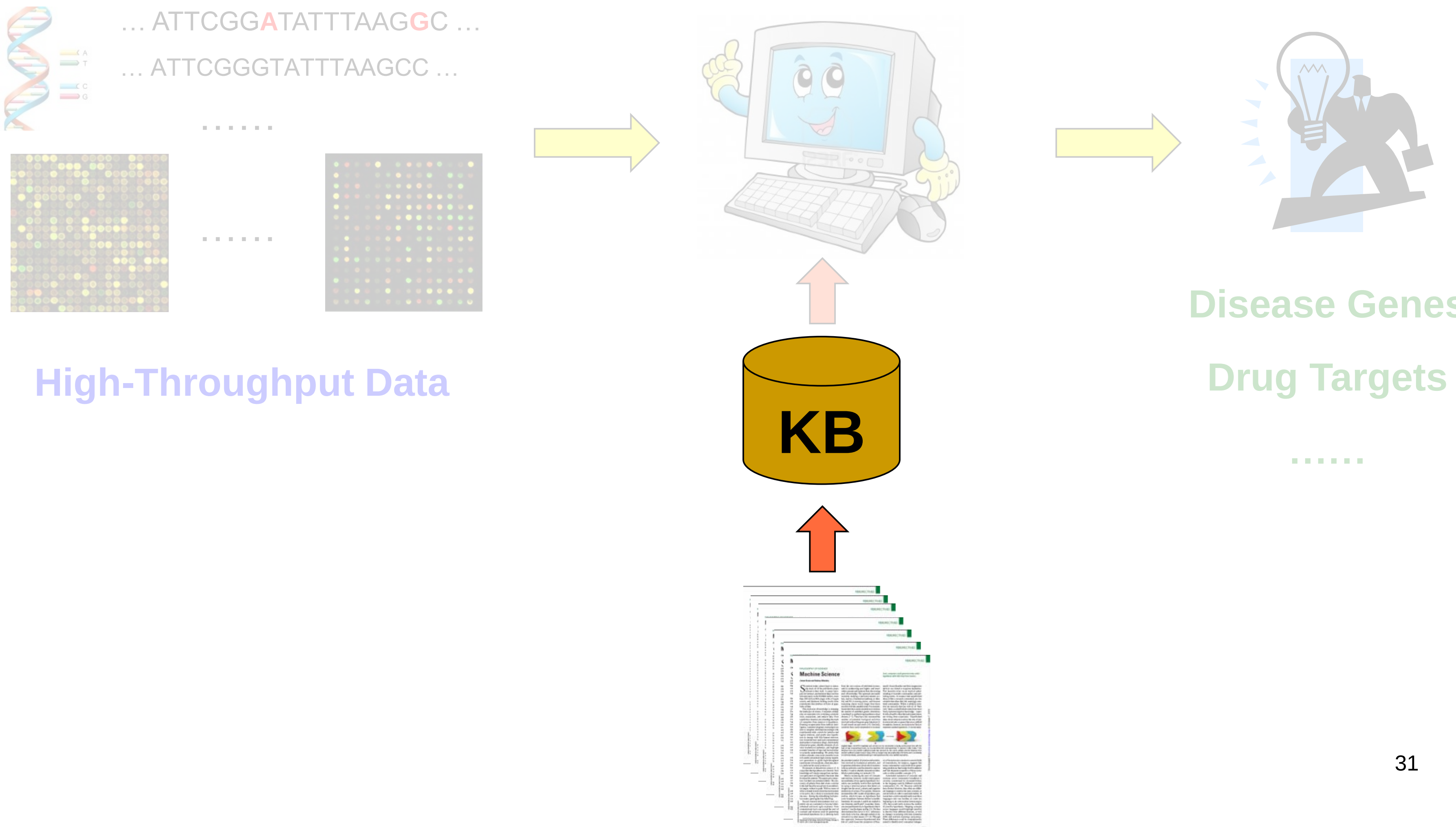




Approach:
Pathway → Graphical Model

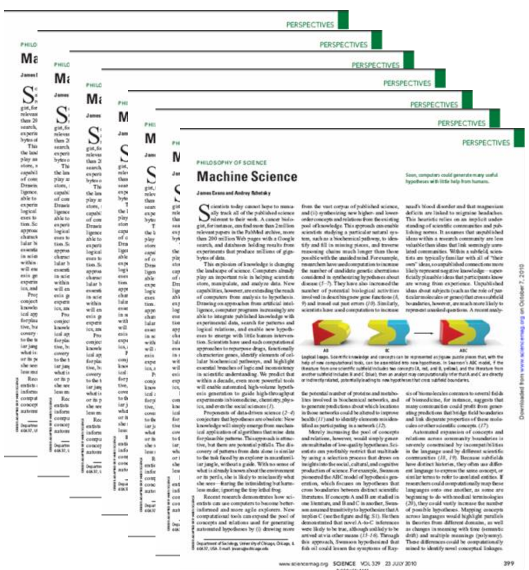


Extract Pathways from Pubmed



PubMed

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



Extract Pathways from Pubmed

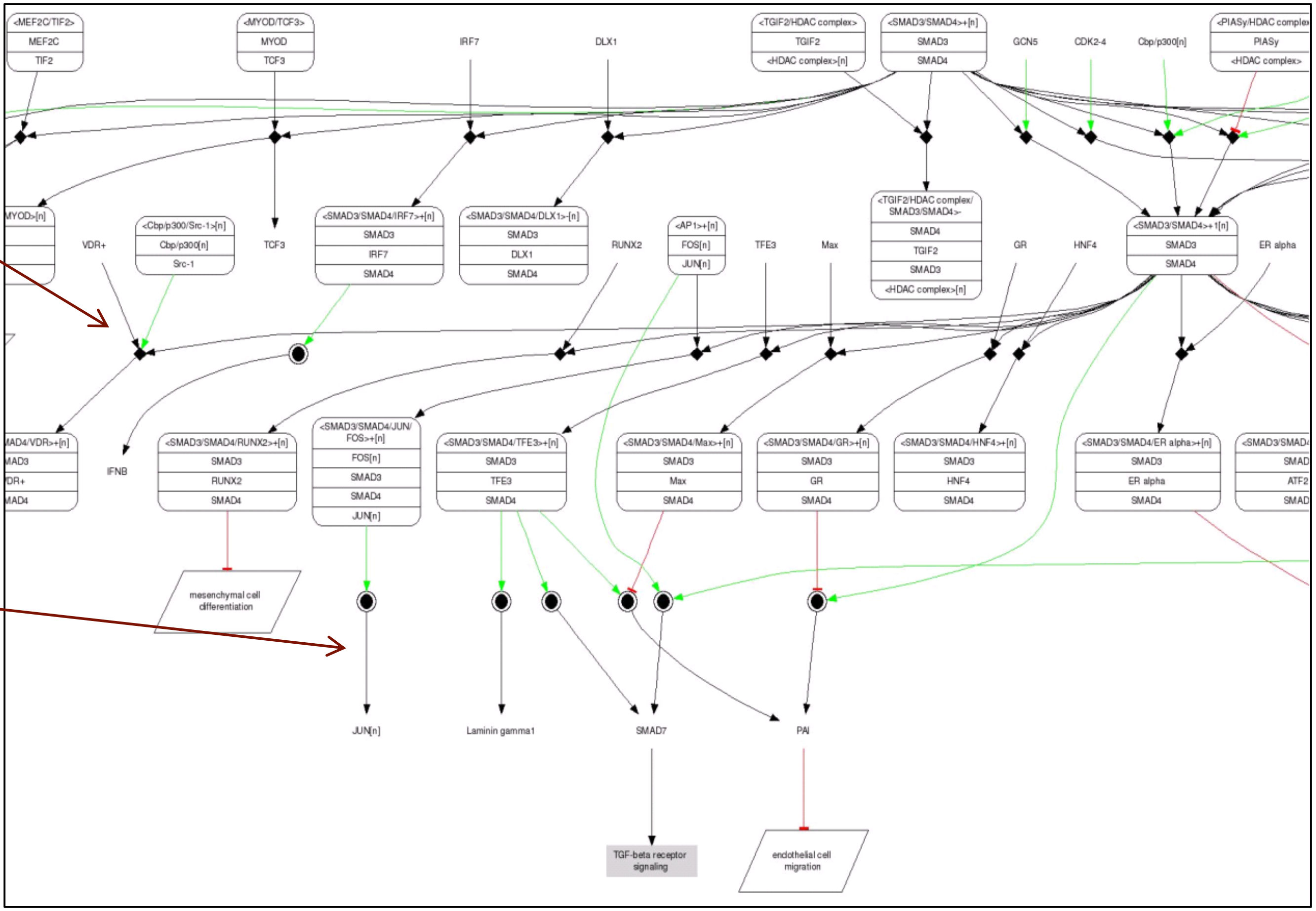
PMID: 123

…
VDR+ binds to SMAD3 to form
…

PMID: 456

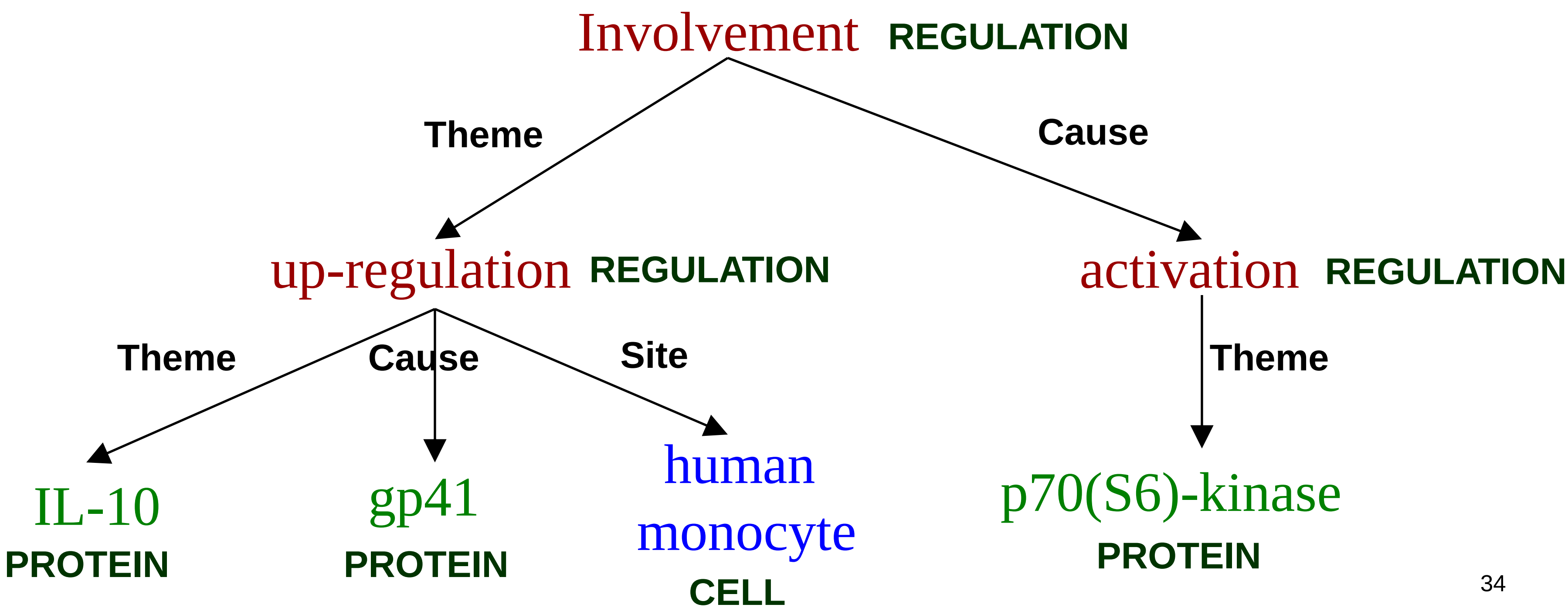
…
JUN expression is induced by SMAD3/4
…

⋮



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 ...



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

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.

Tied with state-of-the-art supervised learning

Shallow Semantics

Get flight from Toronto to San Diego stopping at DTW

INTENT: FLIGHT   FROM_CITY   TO_CITY

Information Extraction

Shallow Semantics

Get flight from Toronto to San Diego stopping at DTW

VERB ARG1

Semantic Role Labeling

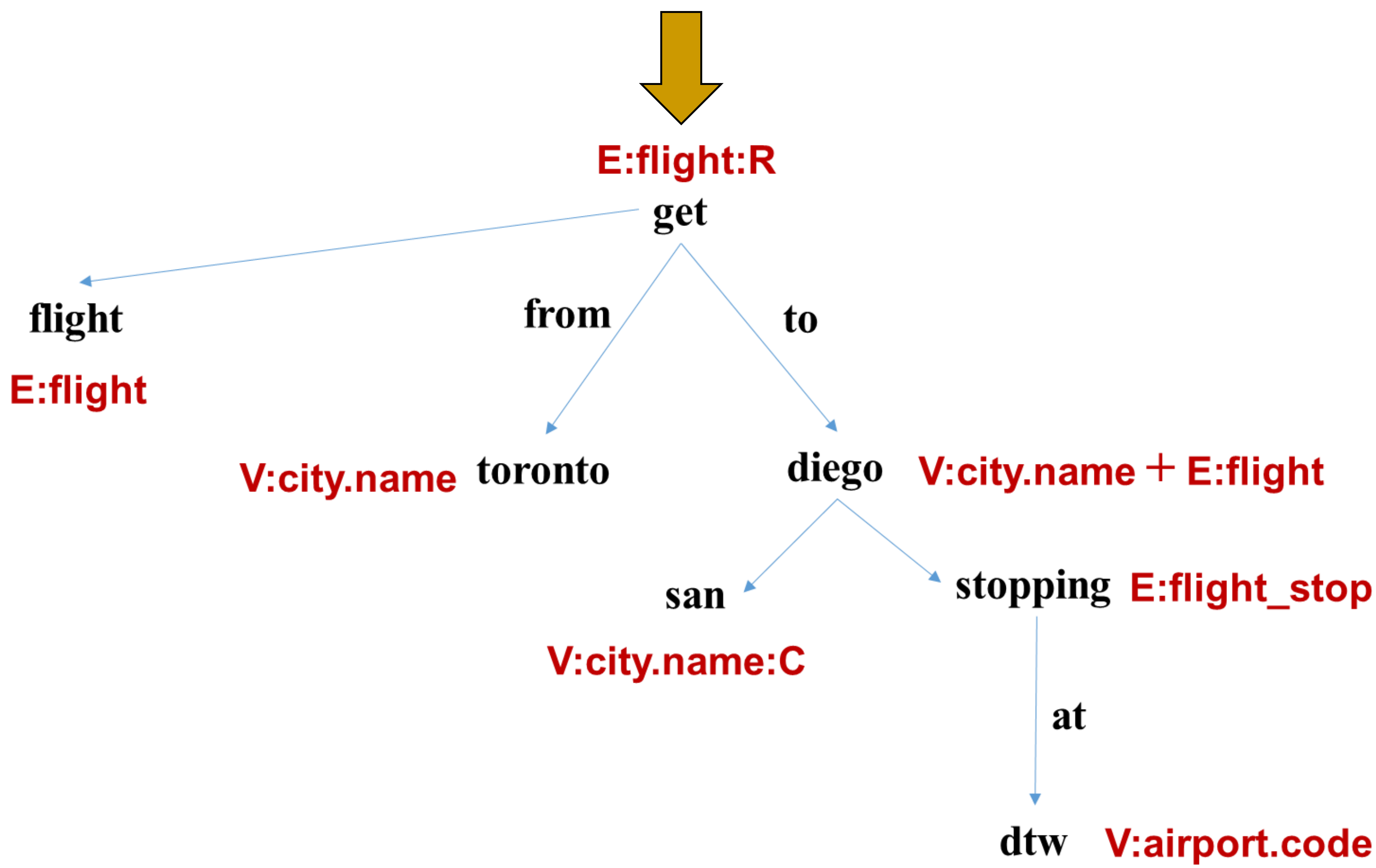
Semantic Parsing

Text $\Rightarrow$ Canonical meaning representation

- Ambiguity resolved
- Complete and detailed

Natural-Language Interface to Database

Get flight from Toronto to San Diego stopping at DTW



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

Supervised Learning

```
get first flight from oakland to salt lake city on thursday
(argmin $v1 (and (flight $v1) (from $v1 oakland:ci) (to $v1
salt_lake_city:ci) (day $v1 thursday:da)  ) (departure_time $v1))

get last flight from oakland to salt lake city on wednesday
(argmax $v1 (and (flight $v1) (from $v1 oakland:ci) (to $v1
salt_lake_city:ci) (day $v1 wednesday:da)  ) (departure_time $v1))

list last wednesday flight from oakland to salt lake city
(argmax $v1 (and (flight $v1) (from $v1 oakland:ci) (to $v1
salt_lake_city:ci) (day $v1 wednesday:da)  ) (departure_time $v1))

get flight from toronto to san diego stopping at dtw
(lambda $v0 e (and (flight $v0) (from $v0 toronto:ci) (to $v0
san_diego:ci) (stop $v0 dtw:ap) ))

get flights between st. petersburg and charlotte
(lambda $v0 e (and (flight $v0) (from $v0 st_petersburg:ci) (to
$v0 charlotte:ci) )


......
```

Supervised Learning

- Examples:
  - Zelle & Mooney [1993]
  - Zettlemoyer & Collins [2005, 2007, 2009]
  - Wong & Mooney [2007]
  - Lu et al. [2008]
  - Ge & Mooney [2009]
  - Kwiatkowski et al. [2011]
- Require annotated logical forms
- Costly and time-consuming

Grounded Learning

Get flight from Toronto to San Diego stopping at DTW

Flight ID
AS123
UA456
SW789

Annotate example question-answer pairs

Grounded Learning

- Examples:
  - Clarke et al. [2010]
  - Liang et al. [2011]
- Successful on JOBS, GeoQuery
- Still need to annotate answers
- Challenging in more complex domains

Unsupervised Semantic Parsing

- USP [Poon & Domingos 2009]
- Recursively cluster & compose synonymous meaning units
- Logical form = Self-induced clusters
- Not directly applicable to answer complex questions against an existing database
- Exact inference intractable

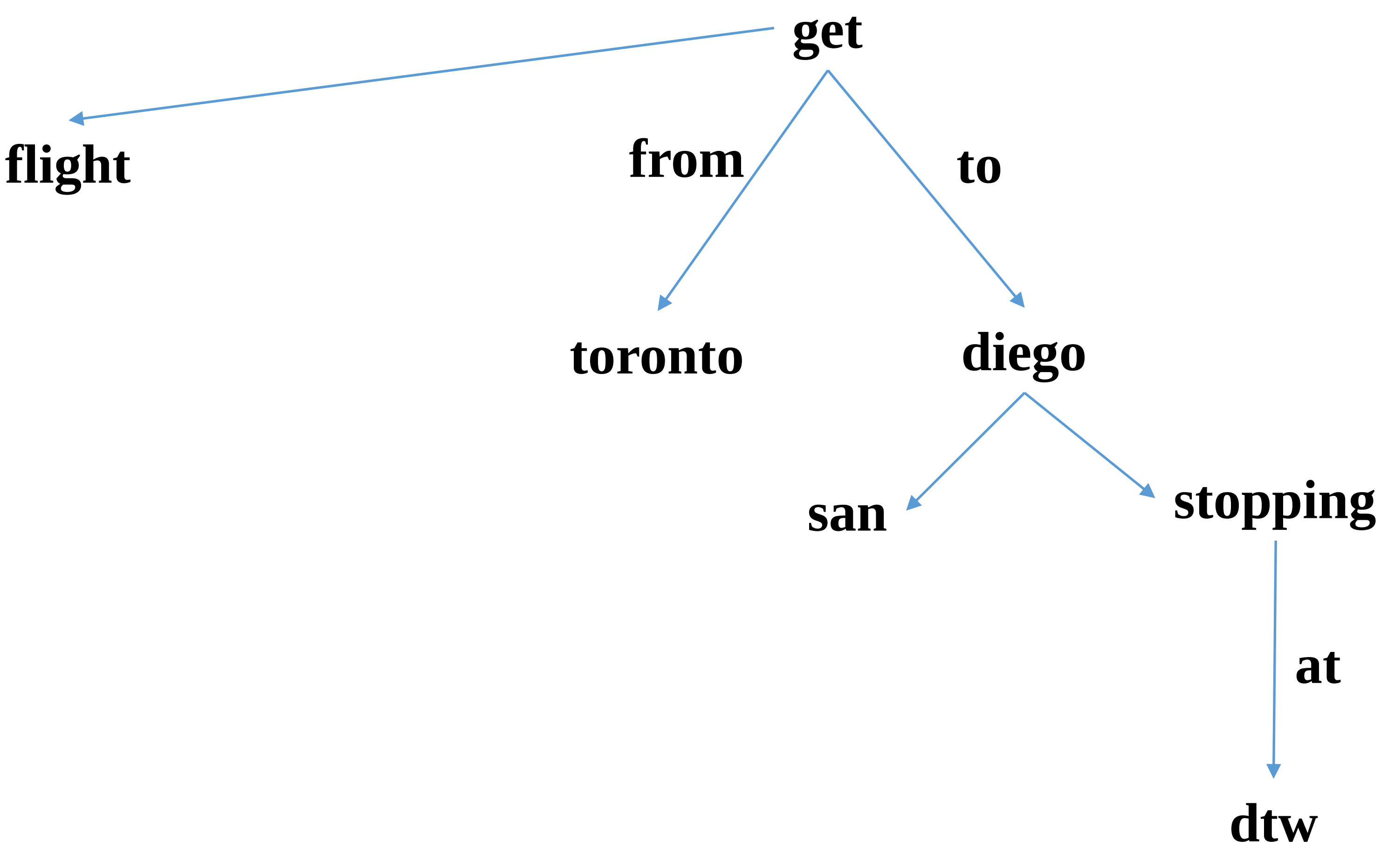
Other Related Work

- PRECISE [Popescu et al. 2003, 2004]
- Confidence-driven unsupervised semantic parsing [Goldwasser et al. 2011]
- Weak supervision [Artzi & Zettlemoyer 2011, 2013]
- Distant supervision
  - Mintz et al. [2009]
  - Riedel et al. [2010]
  - Hoffmann et al. [2011]
  - Krishnamurphy & Mitchell [2012]
  - Etc.

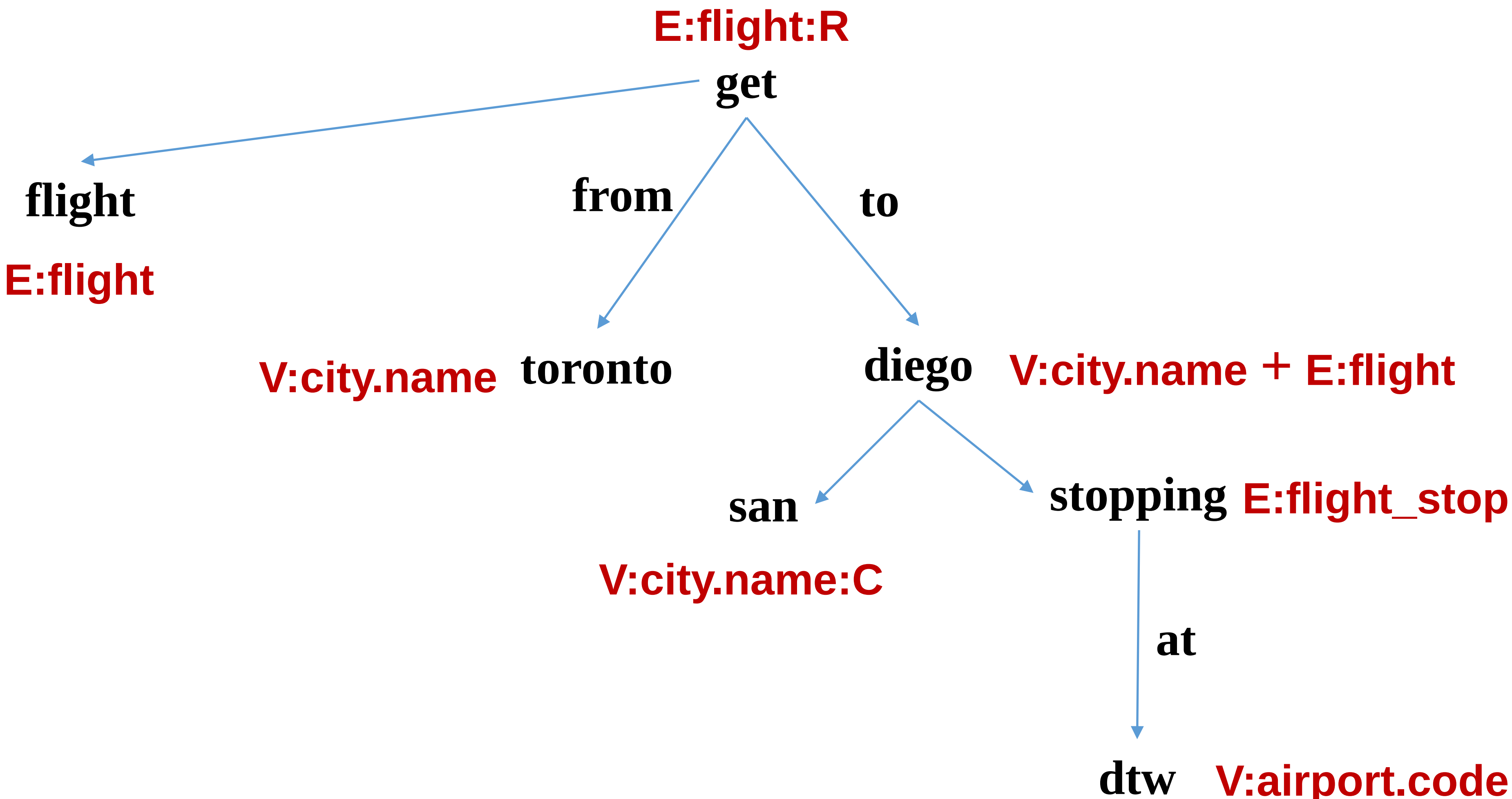
Grounded Unsupervised Semantic Parsing

- Many databases are available
- Database provides:
  - Schema: Concepts and relations
  - Contents: Element names and values
- **Idea:** Use databases as indirect supervision to bootstrap semantic learning

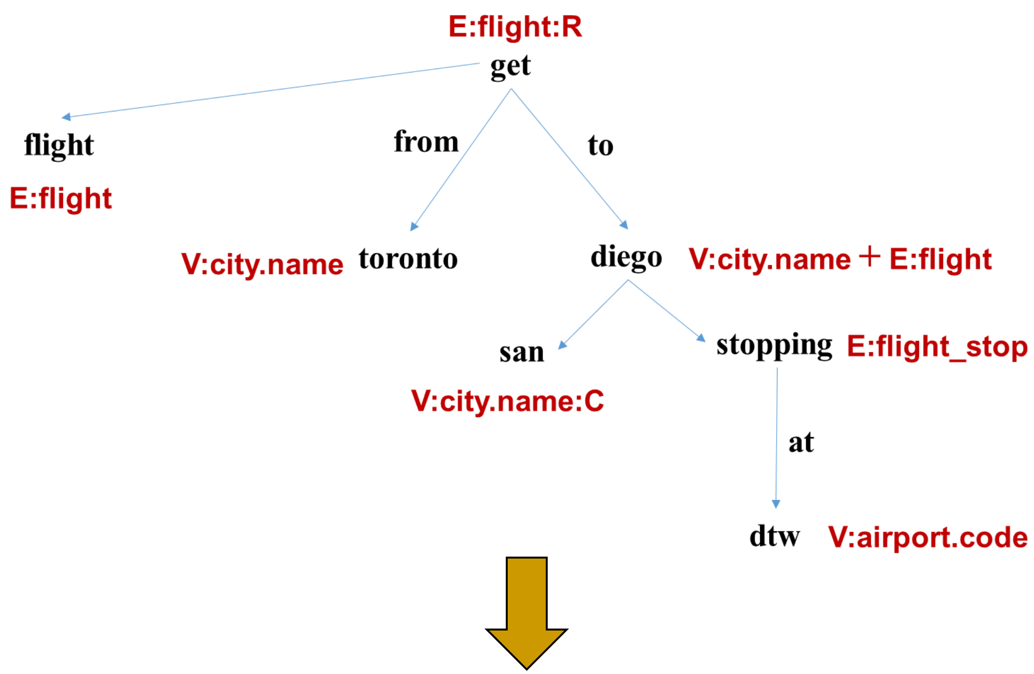
The GUSP System



The GUSP System



The GUSP System



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'

Problem Formulation

Dependency tree d Semantic parse z

Probability $P_\theta(d, z)$

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

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

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

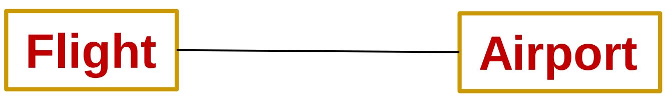
Prior: Favor Unix → System

GUSP: Key Ideas

- **Leverage target database**

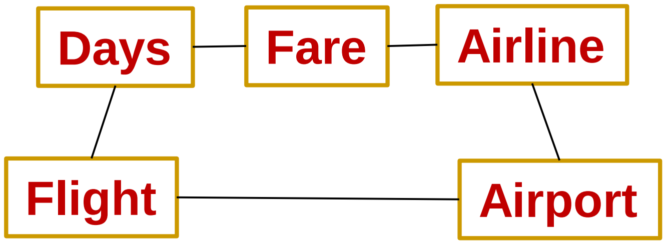
GUSP: Key Ideas

- **Leverage target database**



GUSP: Key Ideas

- **Leverage target database**



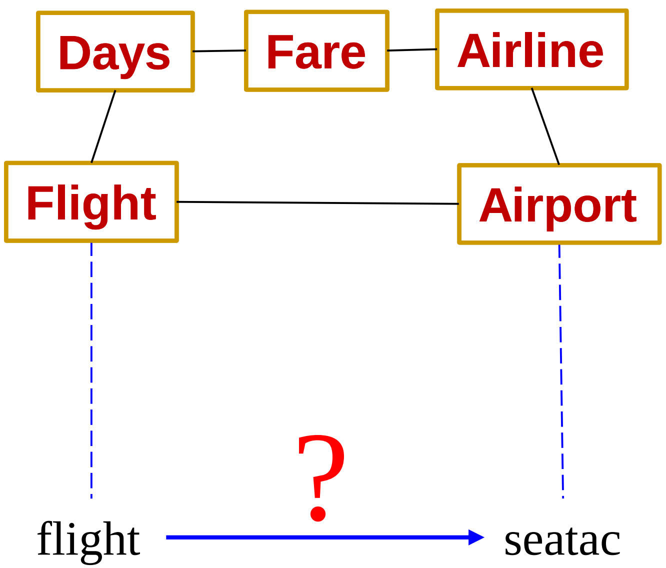
GUSP: Key Ideas

- **Leverage target database**

flight $\longrightarrow$ seatac

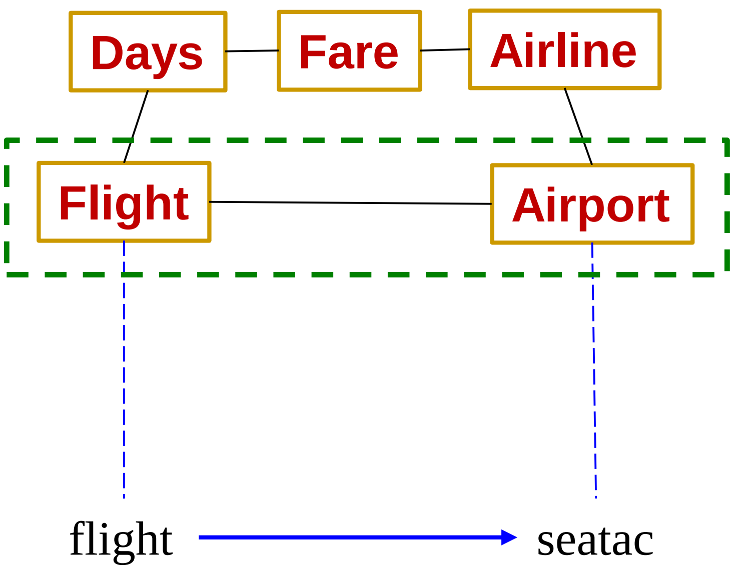
GUSP: Key Ideas

- **Leverage target database**



GUSP: Key Ideas

- **Leverage target database**



Leverage schema to guide learning

Prior: Favor shorter path

GUSP: Key Ideas

- Leverage target database
- **Start from syntactic parse**
  - Rich resources and available parsers
  - Reduce structured prediction to annotating latent semantic states
  - Need to handle syntax-semantics mismatch

Simple States

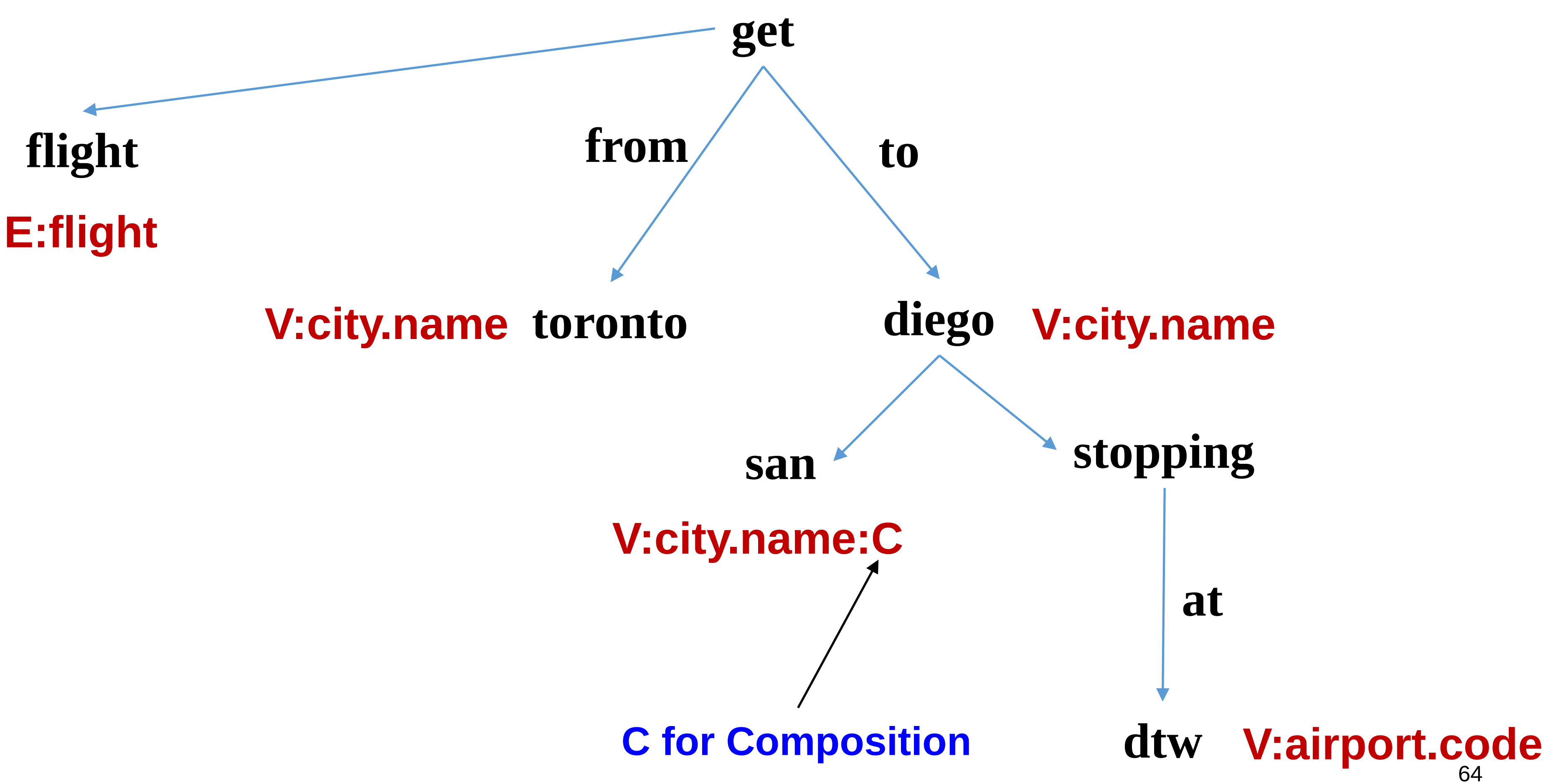
- Node states
- Edge states
- Domain-independent states

Node States

- Database entities, properties, values
- E.g.:

E:flight	flight
P:flight.departure_time	leaving
V:flight.departure_time	pm

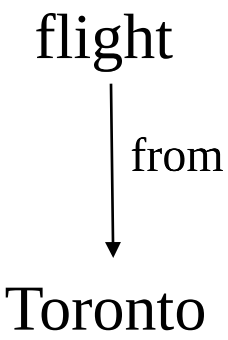
Simple Node States



Edge States

- Relational join paths
- E.g.

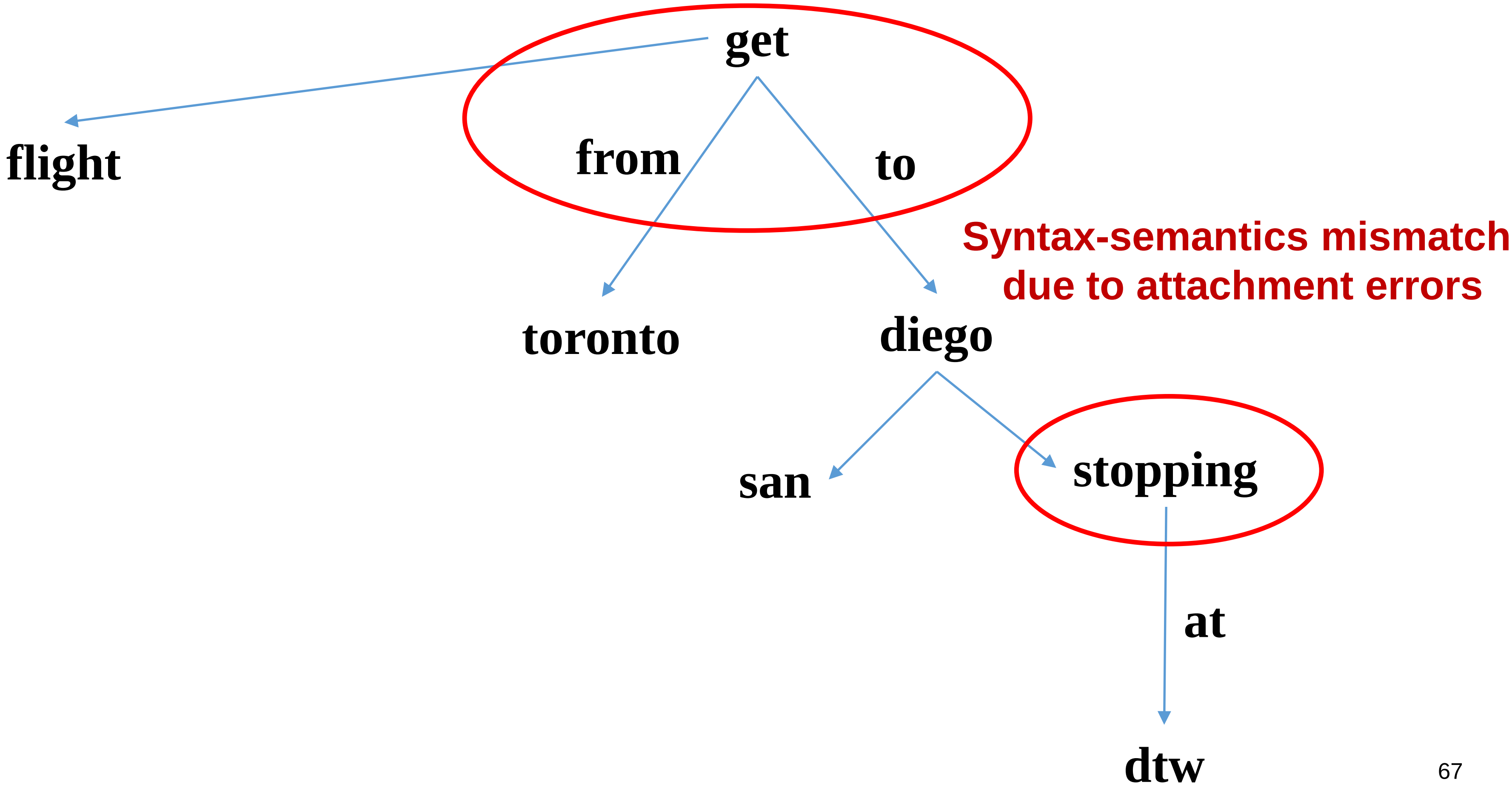
```
flight – flight.from_airport – airport_service.airport
 – airport_service.city – city.city_name
```



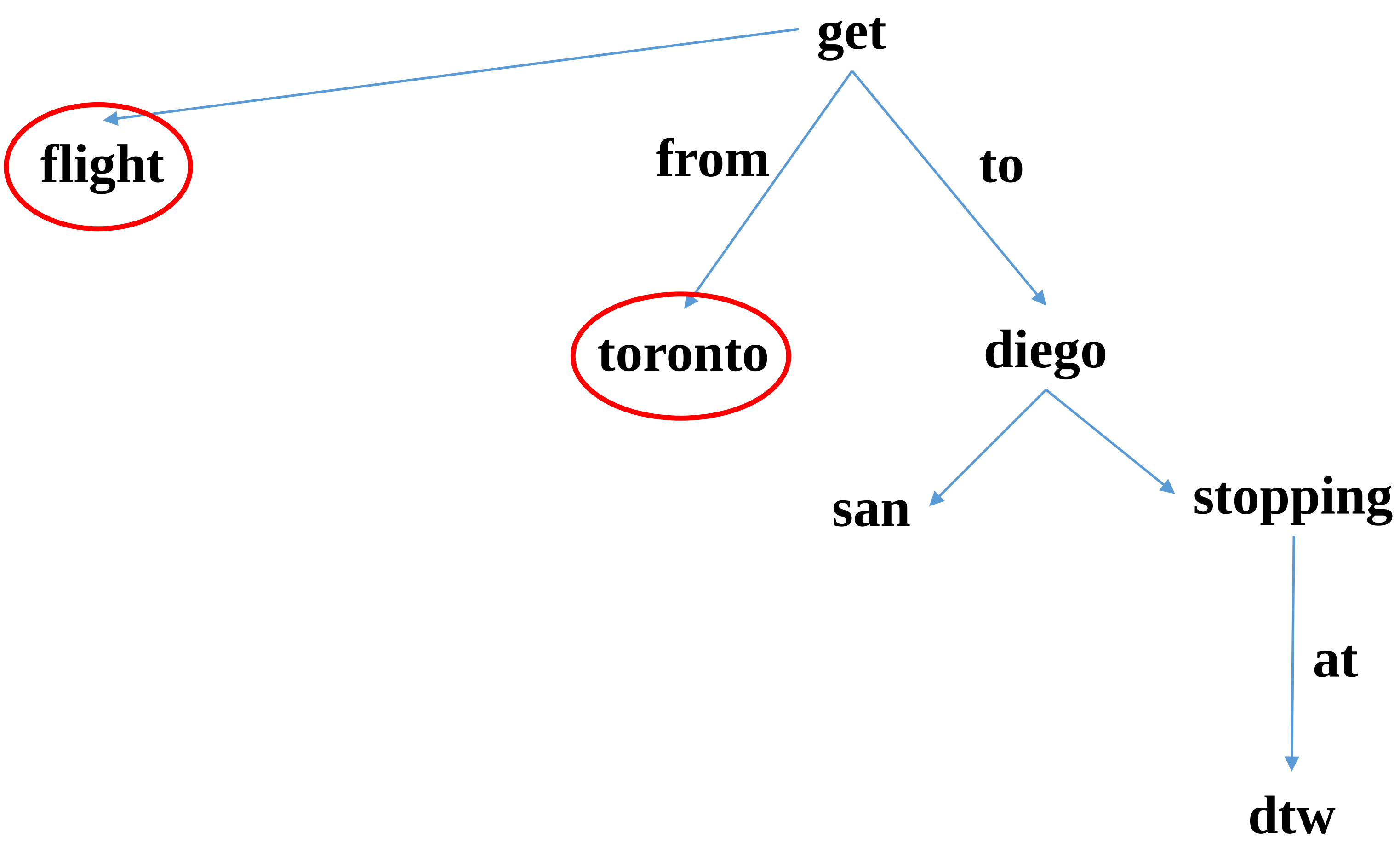
Domain-Independent States

- Logic: `AND,NOT,OR`
- Compare: `MORE,LESS,EQ`
- Superlative: `ARGMIN,ARGMAX`
- Aggregation: `SUM,COUNT`

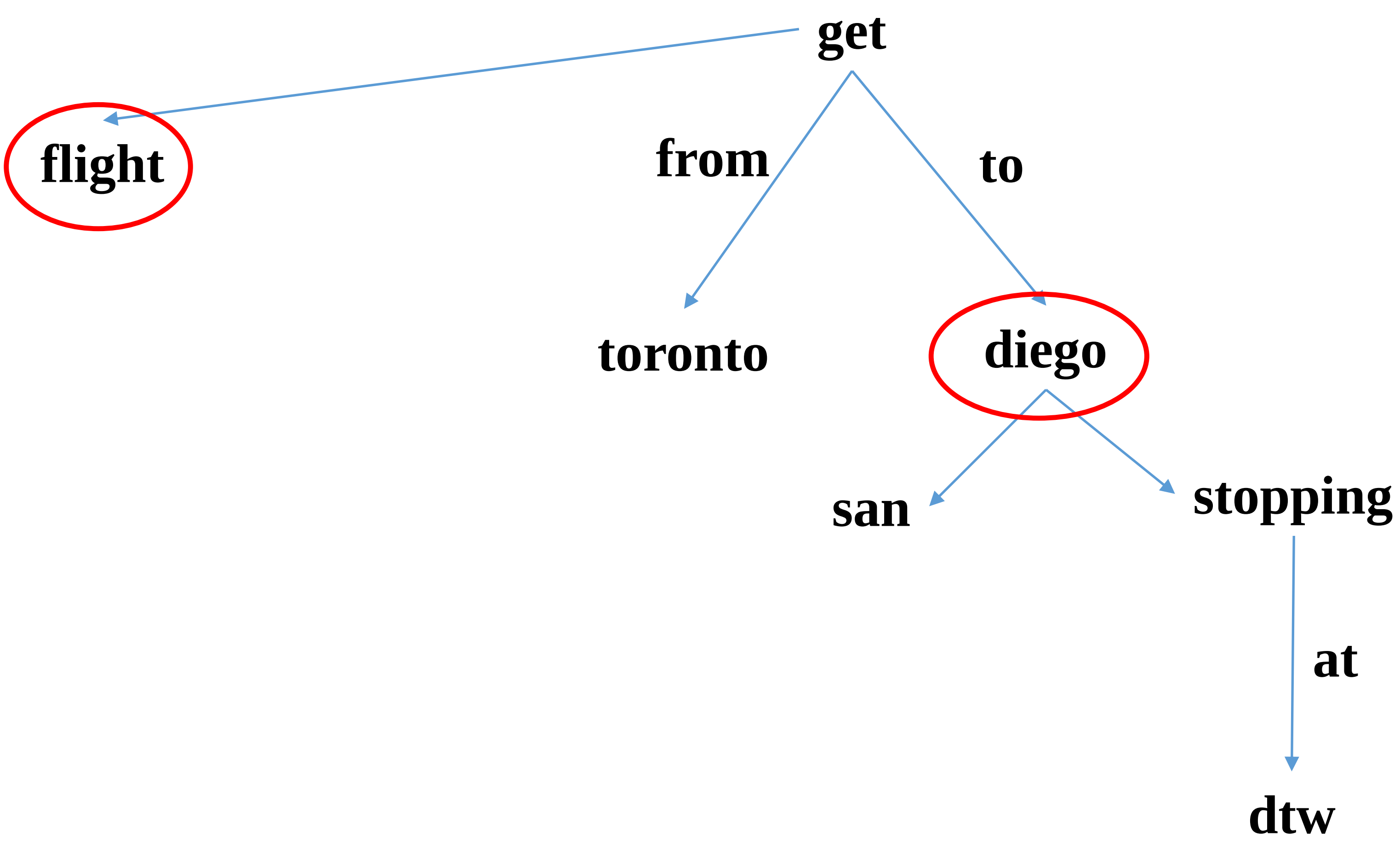
Why Complex States?



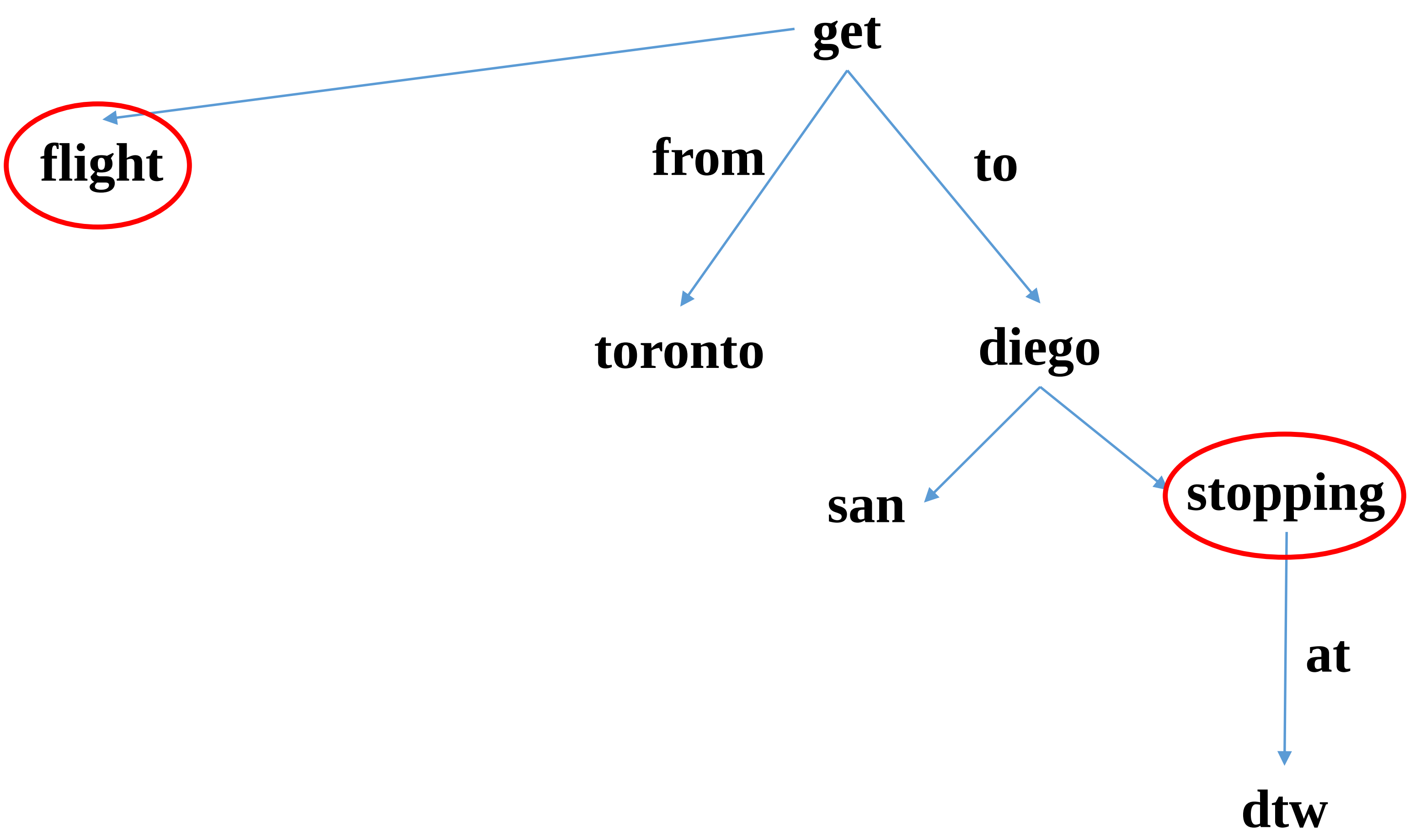
Why Complex States?



Why Complex States?



Why Complex States?



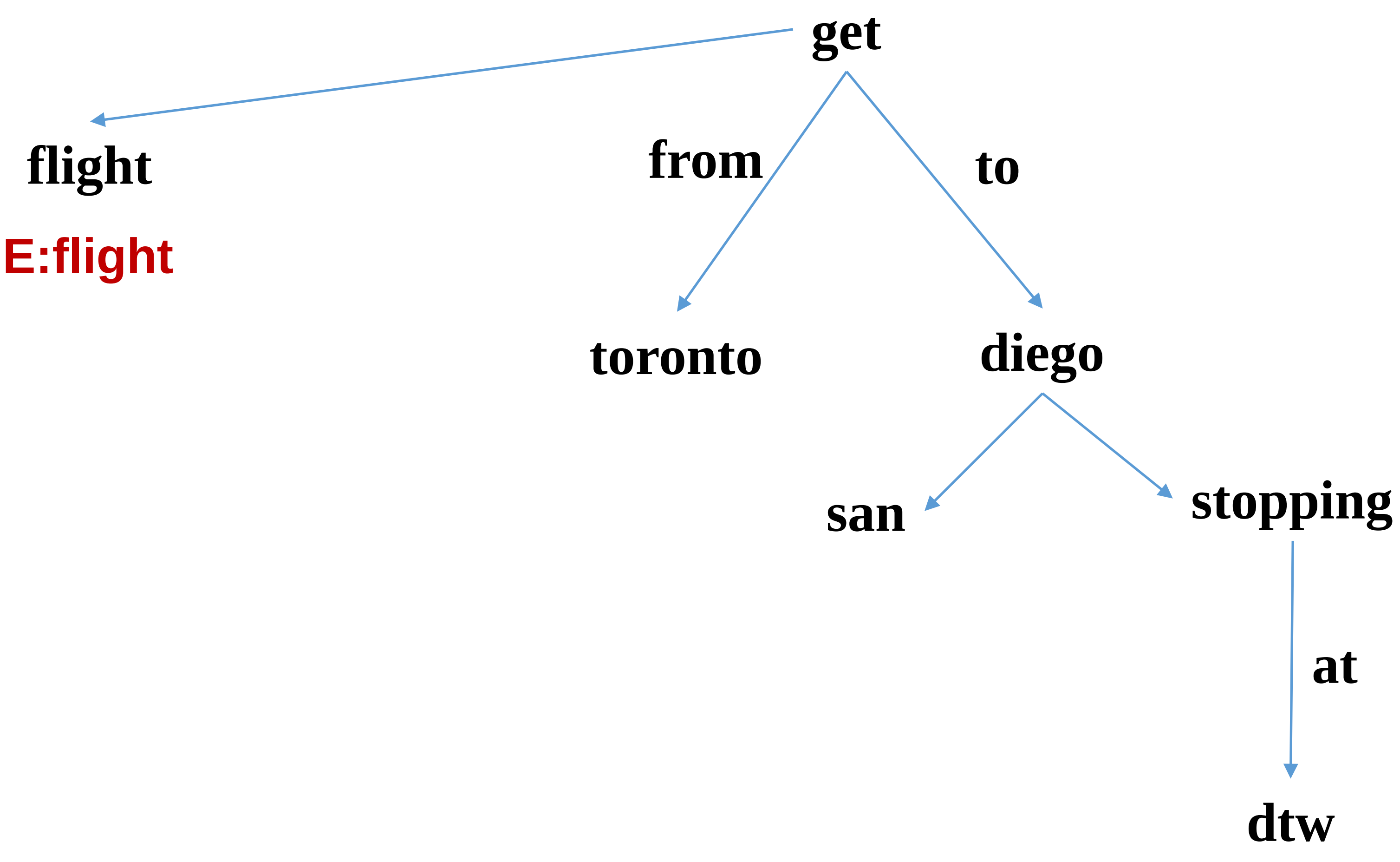
Complex States

- Raising
- Sinking
- Implicit

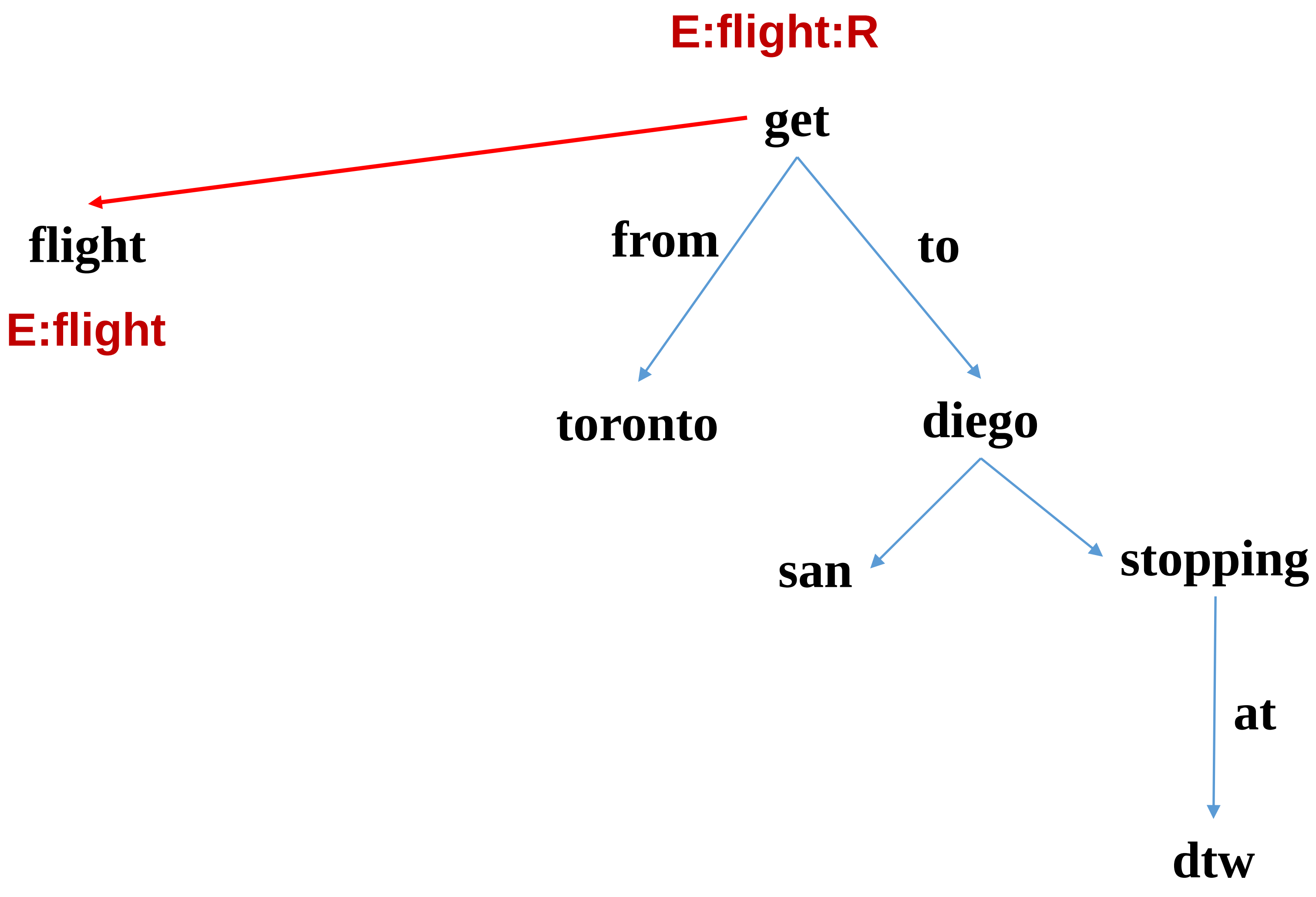
Raising

- For each simple node state (e.g., `E:flight`)
- Create a “raised” node state (e.g., `E:flight:R`)
- Create a “raising” edge state
  (e.g., `flight – R – flight`)

Raising


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

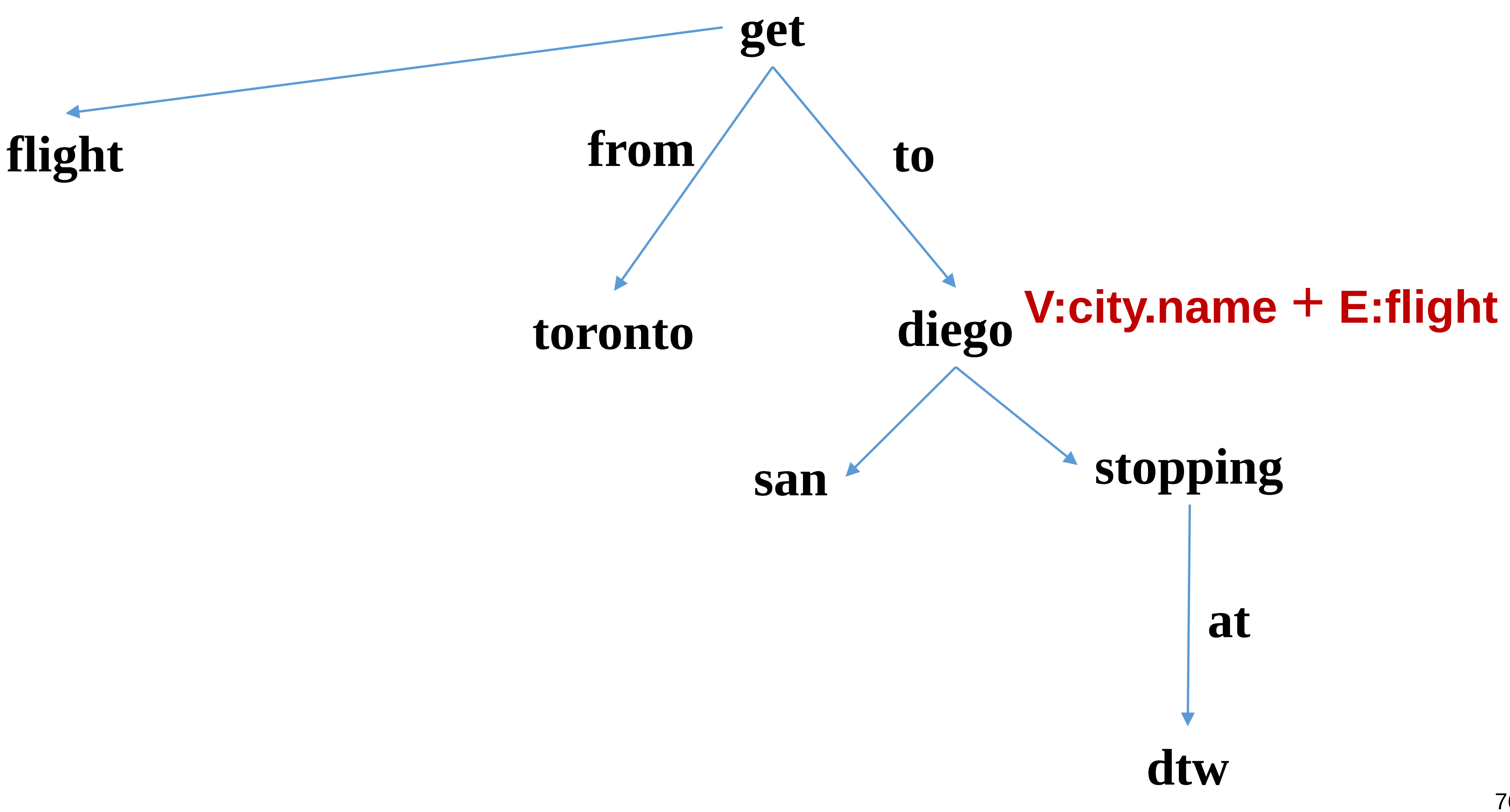
Raising


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

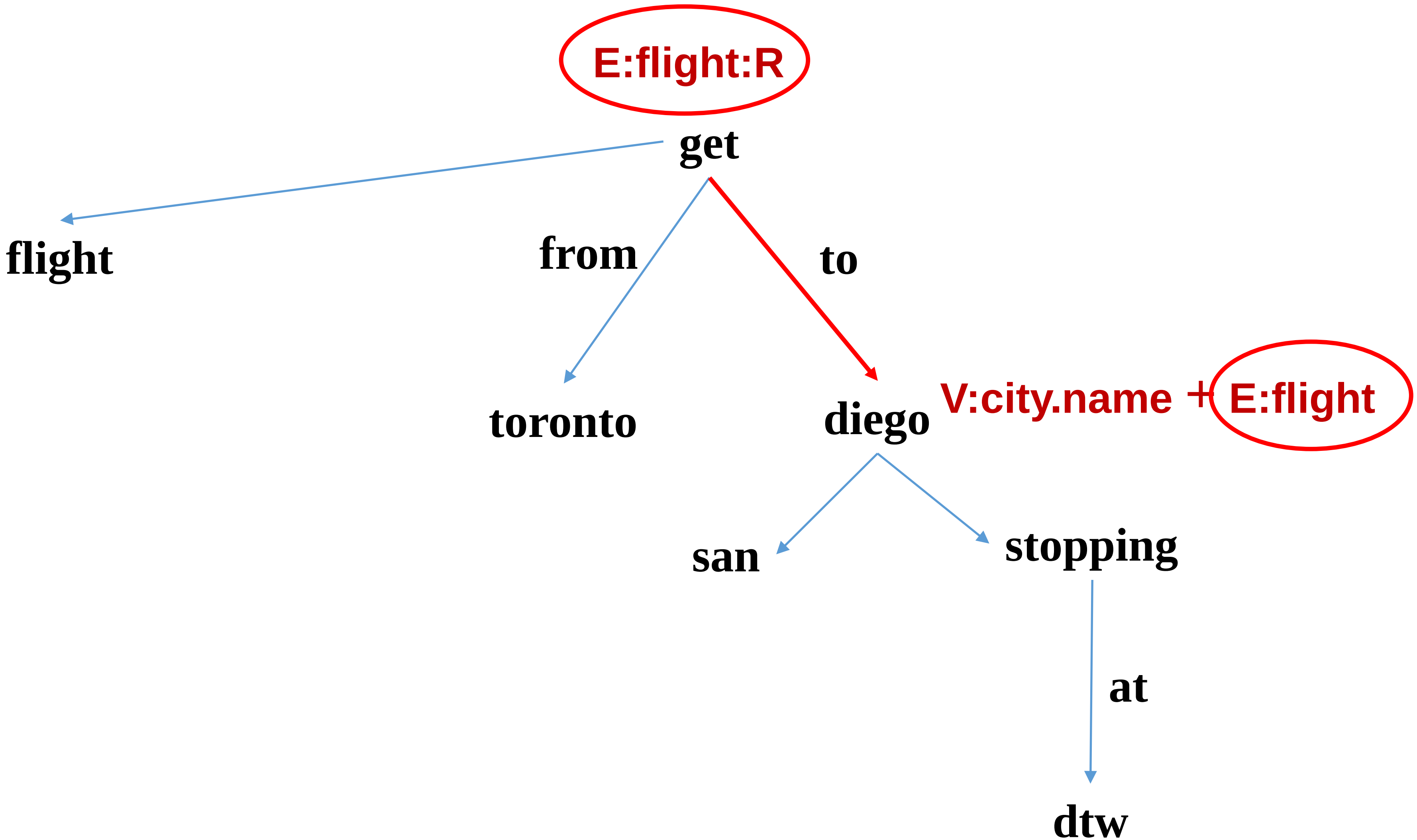
Sinking

- For each node state pairs `A`, `B`
- And for each connecting edge state `E`
- Create a “sinking” node state: `A+E+B`

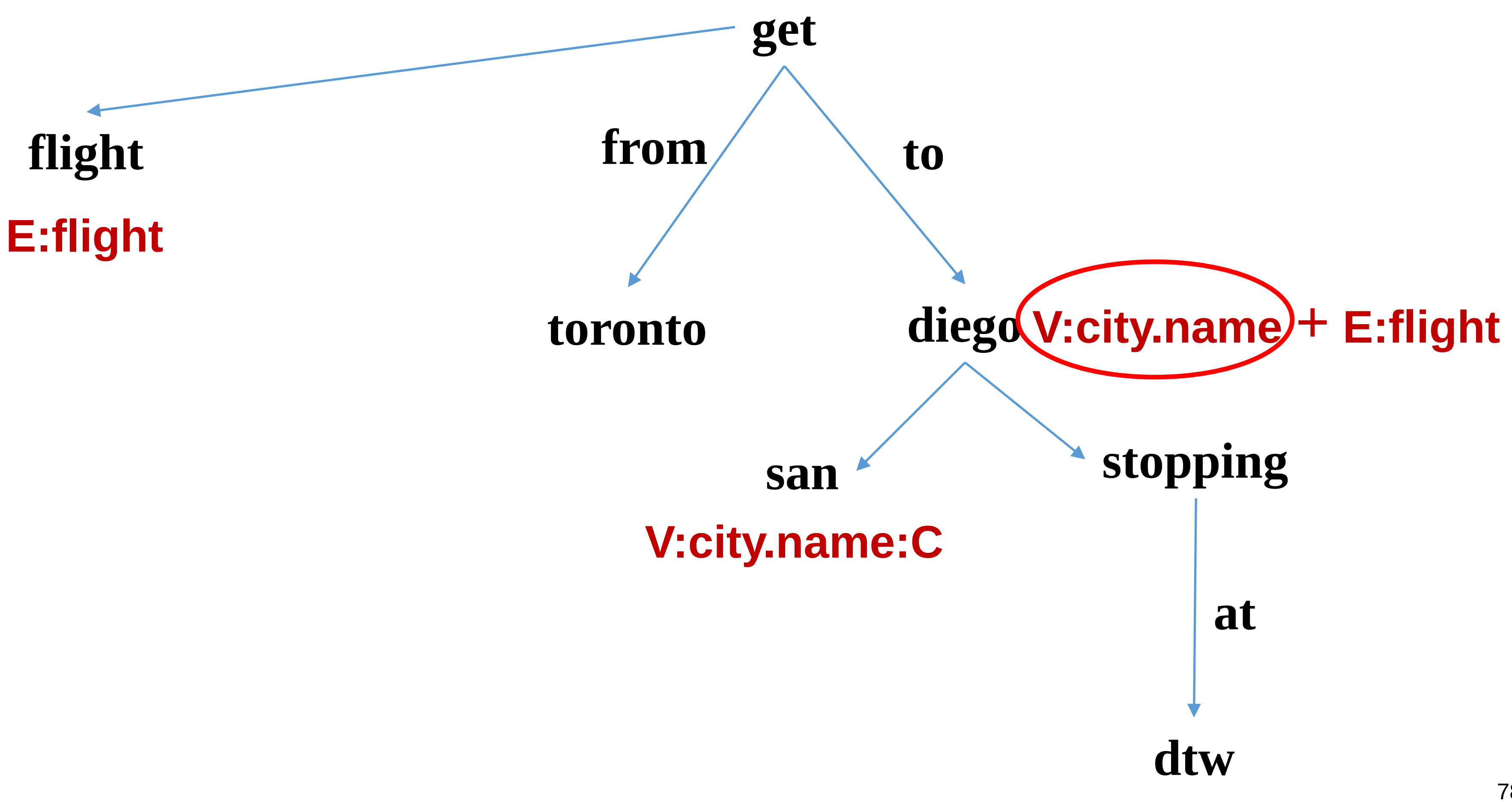
Sinking


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

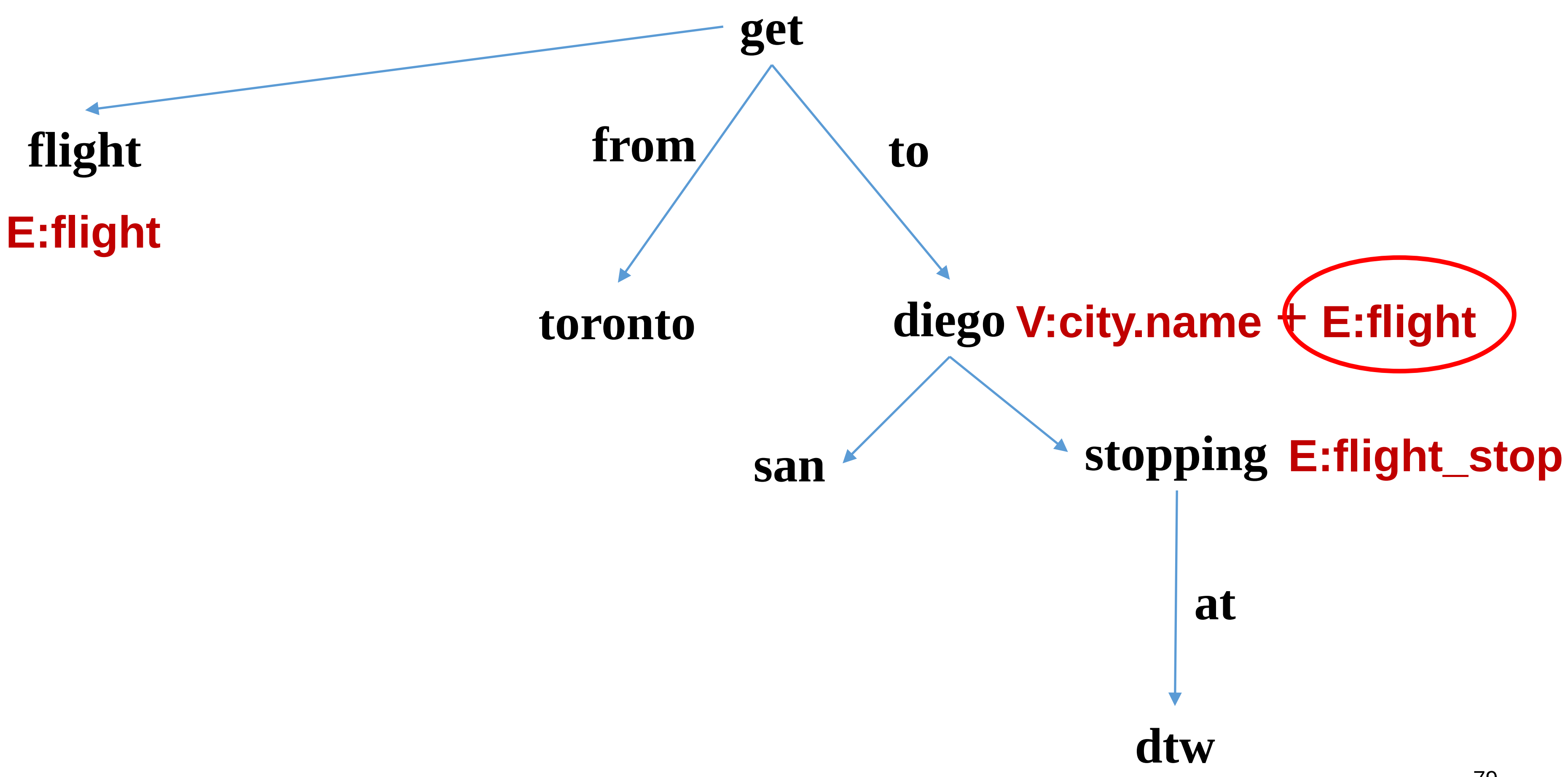
Sinking


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

Sinking


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

Sinking


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

Implicit

- Similar to sinking: Two simple node states
- But, “implicit” simple state not visible to parent

Implicit

Give me the fare from Seattle to Boston

Implicit

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

Implicit

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

fare

E:fare

fare

E:fare + E:flight

Lexical-Trigger Scores

- Entity name → Entity state
- Property name → Property state
- Property value → Value state

The GUSP Model

- Analogous to a tree HMM
- Inference: Viterbi, inside-outside

  Exact inference is linear-time
- Learning: EM
- Complexity prior

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

Future Work

- GUSP $\rightarrow$ Extract complex knowledge
  - Leverage distant supervision
  - Joint syntactic-semantic parsing
  - Continuous learning from interactions
- Pubmed-scaled extraction
  - Biological: Pathways, etc.
  - Medical: Drug-genome interactions, etc.
- Other domains: Financial, legal, etc.

Ongoing: Pubmed-Scaled Pathway Extraction

- Preliminary pass:
  - 500,000 instances
  - 7000 genes, 67,000 unique interactions
- Applications:
  - UCSC Genome Browser
  - Cancer Commons
  - Center for Cancer Innovation (U. Wash.)
  - Etc.

The Literome Project

Terms of Use | Trademarks | Privacy Statement Microsoft

The Literome Project

Welcome Hoifung Poon
Change to

SNPs (12491) Genes (13689) Diseases (3484) Drugs (2109)

Contact Us

Microsoft Research

BPNT1
BPTF
BPY2
BRAF
BRAP
BRCA1
BRCA2
BRCA3
BRCC3
BRD1
BRD2
BRD3
BRD4
BRD7
BRD8
BRD9
BRDT
BRE
BRF1

	via PDGFR-beta signaling pathway:STAT3				
Melanoma	PMID: 10023678 via PDGFR-beta signaling pathway:STAT3	Positive	Negative	Not Associated	Error!
	PMID: 10318823 via ErbB1 downstream signaling:ATF2	Positive	Negative	Not Associated	Error!
	PMID: 11325858 via ErbB1 downstream signaling:BAD	Positive	Negative	Not Associated	Error!
	PMID: 12068308	Positive	Negative	Not Associated	Error!
	PMID: 12150818	Positive	Negative	Not Associated	Error!

PMID: 12068308 (abstract provided by National Library of Medicine)

Mutations of the BRAF gene in human cancer .

Cancers arise owing to the accumulation of mutations in critical genes that alter normal programmes of cell proliferation , differentiation and **death** .

As the first stage of a systematic genome-wide screen for these genes , we have prioritized for analysis signalling pathways in which at least one gene is mutated in human cancer

The RAS RAF MEK ERK MAP kinase pathway mediates cellular responses to growth signals .

RAS is mutated to an oncogenic form in about 15 % of human cancer .

The three RAF genes code for cytoplasmic serine/threonine kinases that are regulated by binding RAS .

Here we repo

All mutations

Mutated BRAF proteins have elevated kinase activity and are transforming in NIH3T3 cells .

BRAF … mutations in 66% of malignant melanoma

http://research.microsoft.com/~hoifung/literome

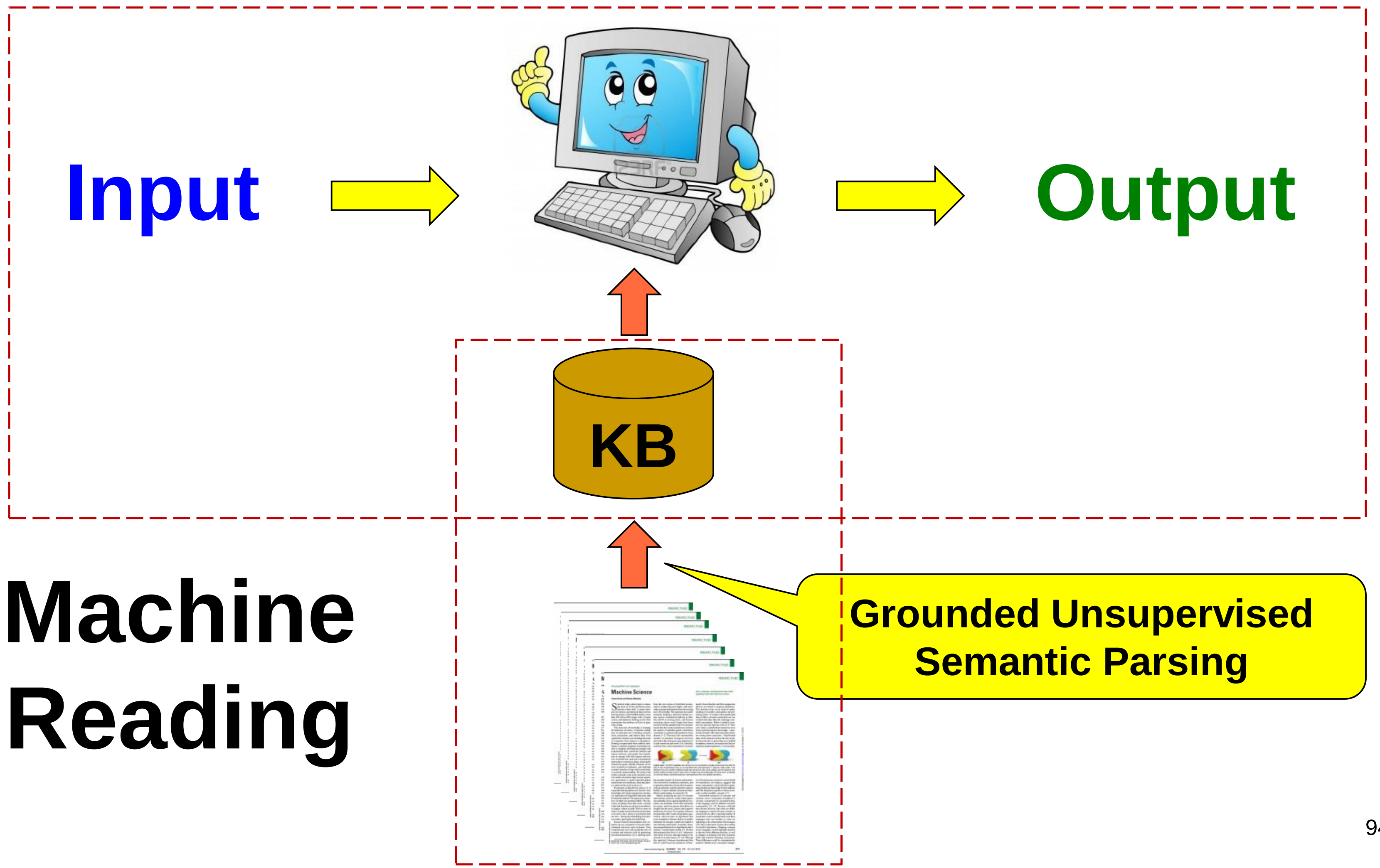
Summary

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

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

  Semantic parsing grounded in KBs

Knowledge-Rich Machine Learning



Acknowledgement

David Heckerman

Tony Gitter

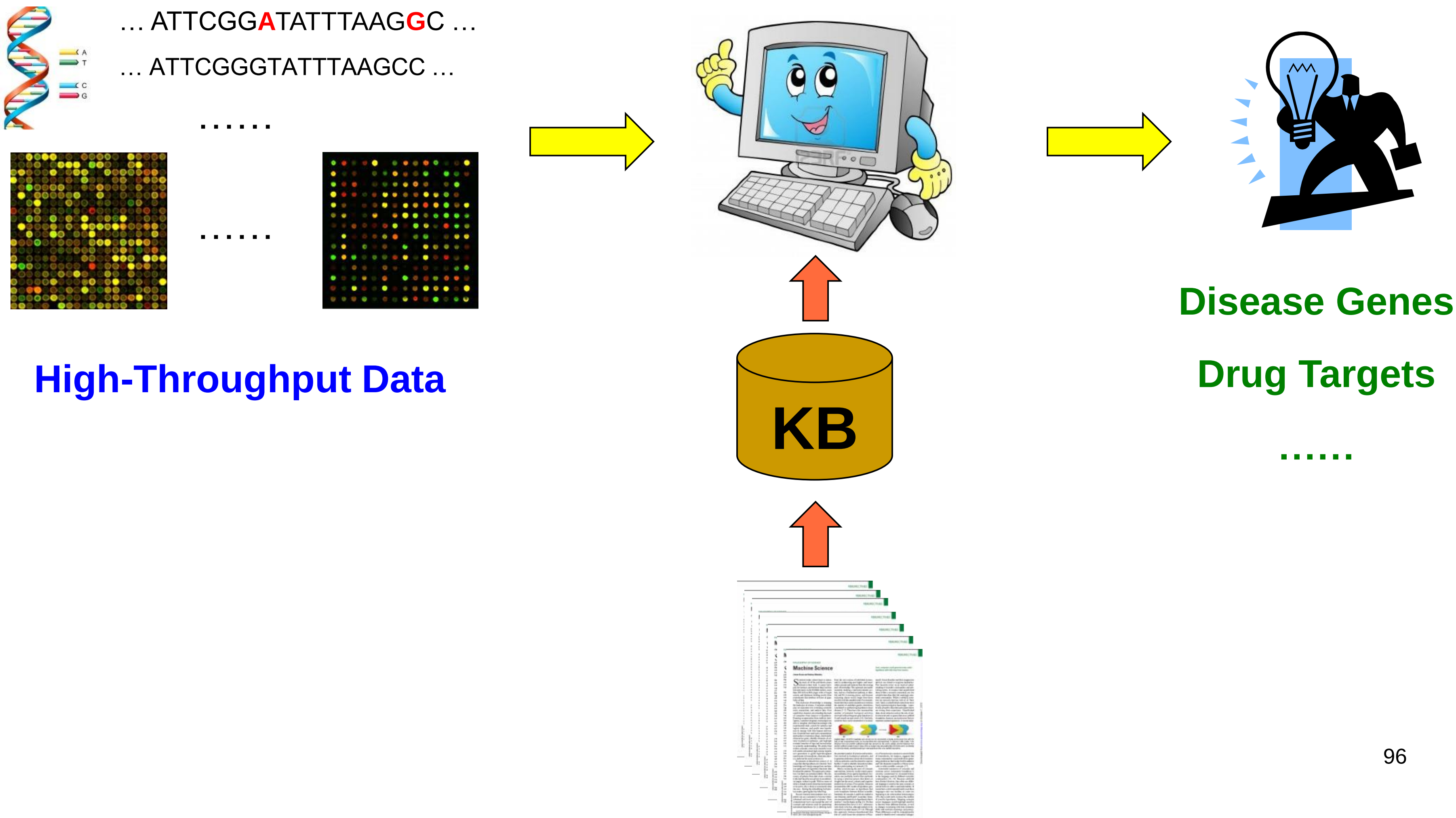
Chris Quirk

Lucy Vanderwende

Kristina Toutanova

Bob Davidson

Summary



… ATTCGGATATTTAAGGC …
… ATTCGGGTATTTAAGCC …
……
……
High-Throughput Data
KB
Disease Genes
Drug Targets
……