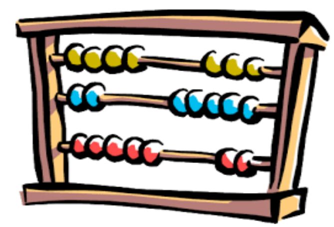
BIG DATA:
A COMPUTER SCIENCE PERSPECTIVE

Jesse Davis
ML Group, Department of Computer Science
http://dtai.cs.kuleuven.be
http://dtai.cs.kuleuven.be/sports

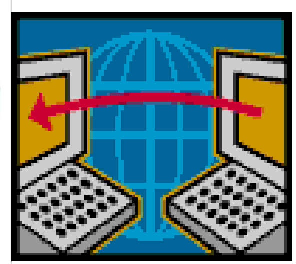
The Evolution of Information

Collect → Store → Retrieve → Analyze → Use

Before



Now

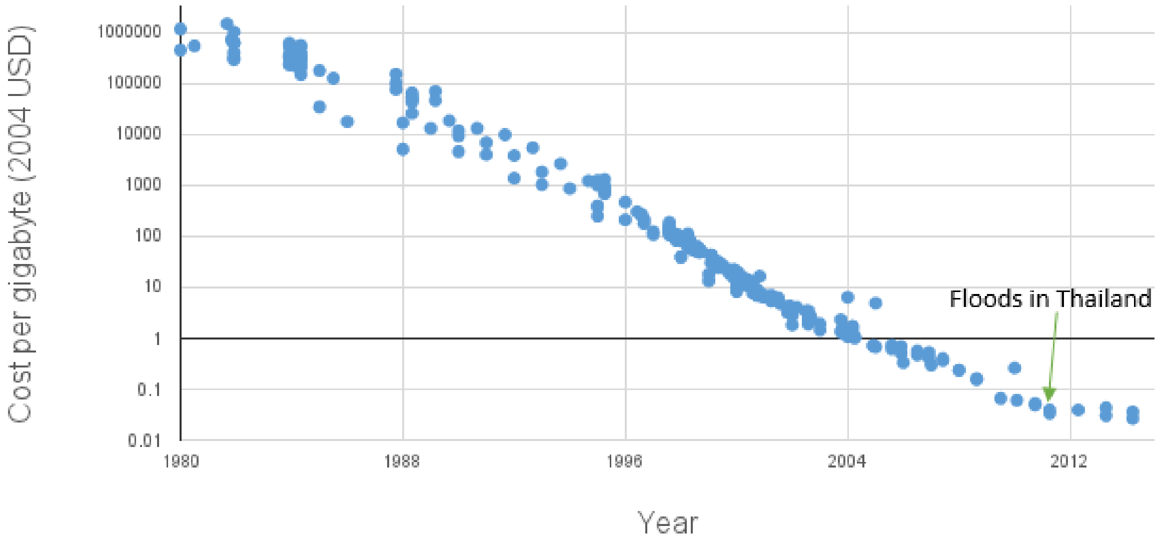


Two questions:

1. Why now?
2. What challenges does this create?

Technical Improvements in Computing Systems

- Storage is cheap



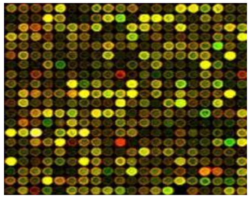
https://intelligence.org/2014/05/12/exponential-and-non-exponential/

- Computers are cheap and fast
- Lots of machines: Steve Ballmer in 2013: “We have something over a million servers in our datacenter infrastructure…Google is bigger…Amazon is a little bit smaller.”

Advances in Collection Techniques and Changes in Behavior

- Automatic collection

- High throughput collection



- Behavior changes

- Result: We generate huge amounts of data
  - **Twitter:** 500,000,000 tweets/day in 2013
  - **Internet Archive:** 15 petabytes of data in 2014
  - **Facebook:** >500 terabytes of data/day in 2012

Three Big Challenges

1. How can we store the data?

2. How can we process the data?

High-level, whirlwind tour

3. How can we find insights in the data?

More depth, with emphasis on ML group's work

Three Big Challenges

1. How can we store the data?

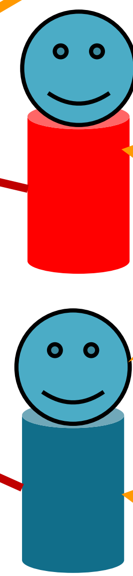
2. How can we process the data?

3. How can we find insights in the data?

Traditional Data Storage: Relational Databases

<u>CustID</u>	Name	Account #	Balance
1	Alex	1-100-101	25,230
1	Alex	1-200-101	1,320
2	Ben	2-200-102	978
3	Chuck	3-100-102	87,413
3	Chuck	3-200-103	3,201
4	Dave	4-200-104	4,243

Balance > 0



Simultaneously access acounts

What happened to my deposit?!?

<u>Goals of traditional databases</u>

- Concurrent access
- Real-time processing
- Constraints
- Recoverability

Databases designed based on
1. This data type
2. These desiderata
3. Resource costs/constraints

Today's World: New Data, New Tasks

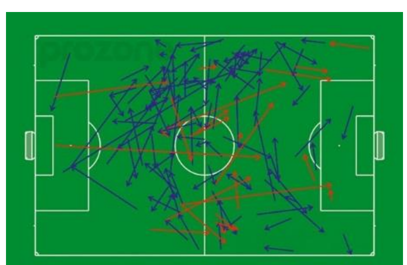
Data

Networks

Documents

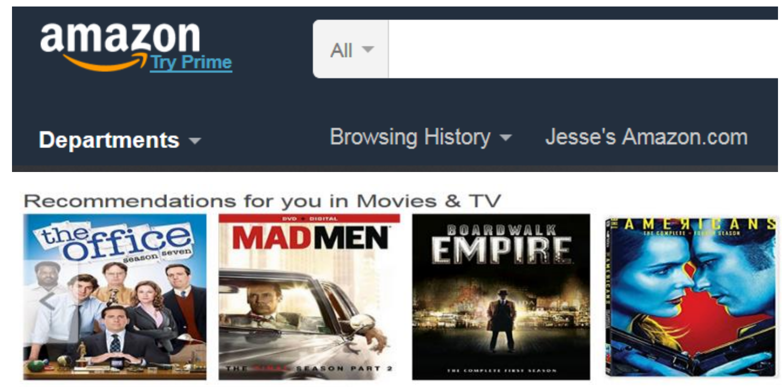
Images

Streams

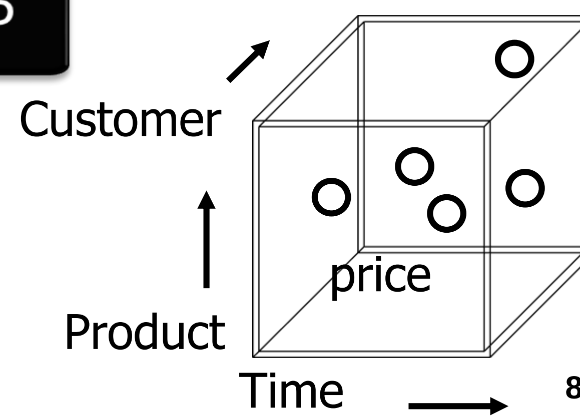


Tasks

Personalized Web Page



OLAP



Specialized Systems for Data Storage

- Key-value storages, e.g., Dynamo
- Column store RDBMS, e.g., Vertica
  [From Stonebraker, 2014 Turing Award winner,
  bought by HP for > 300M]
- Stream databases, e.g., STREAM
- Graph databases, e.g., Neo4j
- Document databases, e.g., MongoDB

Three Big Challenges

1. How can we store the data?

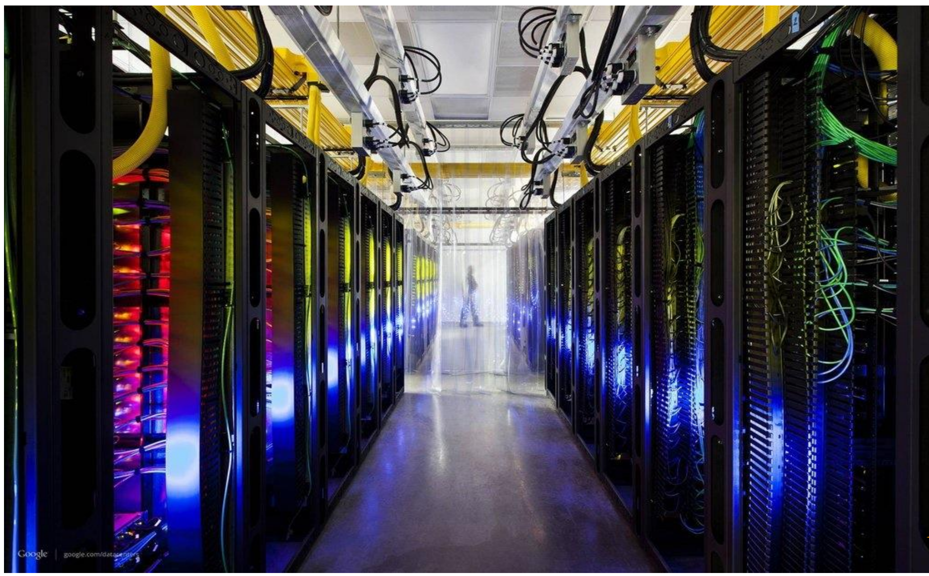
2. How can we process the data?

3. How can we find insights in the data?

Traditional Processing: Supercomputer

Custom architectures

New Processing: The Cloud



Azure Services Platform

amazon web services™

Lots and lots of commodity servers

Challenges:

1. How do we handle node failures?
2. How do we handle decentralized data?
3. How do we make parallel programming easy?

Solution: High-Level Distributed Programming Paradigms

- Three canonical systems
  - MapReduce from Google (less used there now)
  - Spark from UC Berkeley/Databricks
  - GraphLab from Turi (Apple bought for $200M)

- Prominent features include
  - Handle details of distributed programming
  - Coupled with distributed file systems
  - Gracefully cope with node failures

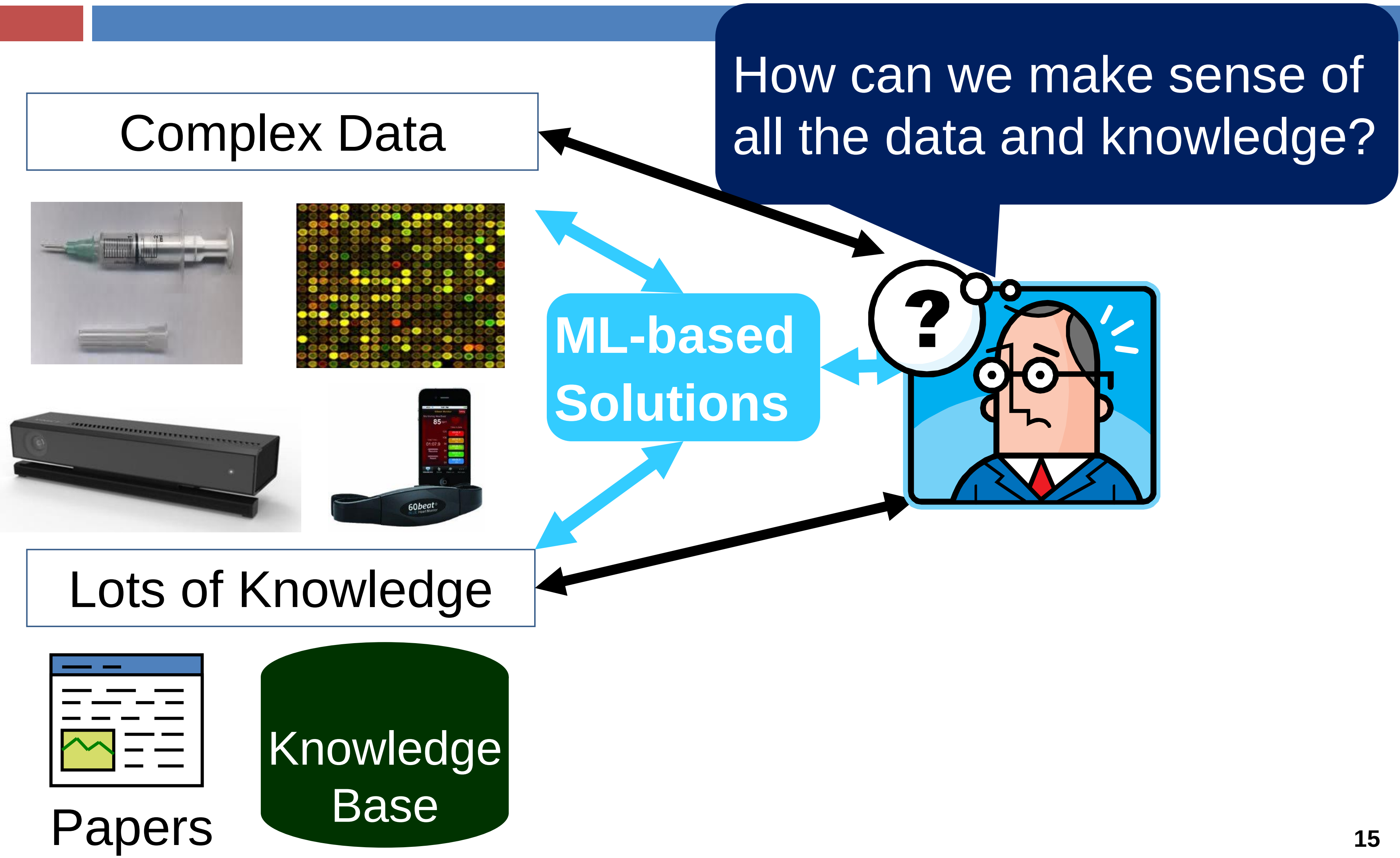
Three Big Challenges

1. How can we store the data?

2. How can we process the data?

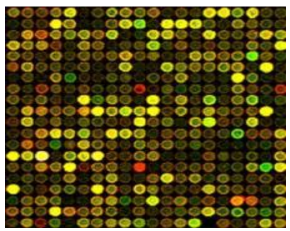
3. How can we find insights in the data?

Challenges Posed By Analysis



How Can Machine Learning Help

Complex Data



Lots of Knowledge

Papers

Knowledge Base

Learn predictive models

$\text{MassSize} > 10 \Rightarrow$ Cancer
Etc.

Reason about the data

Prob(Flu | Fever) = ?

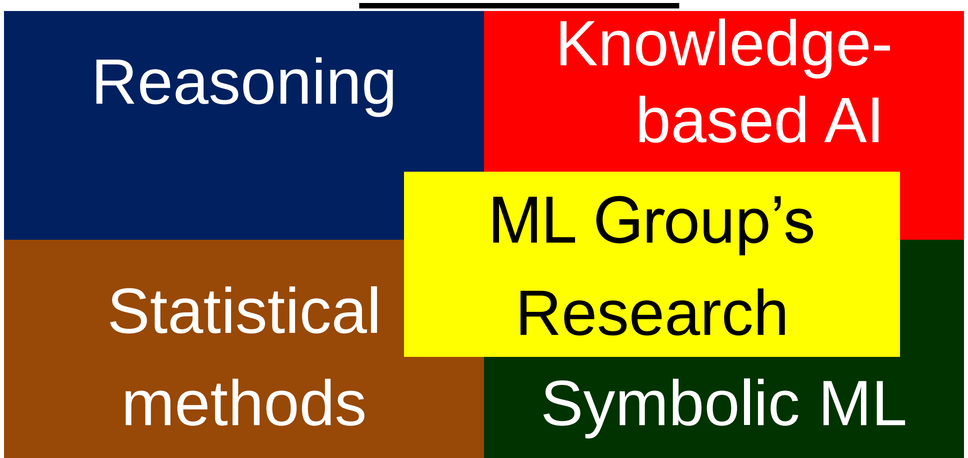
Discover patterns

Smoking $\wedge$ Cancer

ML Group

- Luc De Raedt (ERC advanced grant): AI, Probabilistic logics, automating data science
- Hendrik Blockeel: Prediction, clustering, scalability, anomalies
- Bettina Berendt: Privacy, ethical aspects, Web
- Jesse Davis: Machine learning, data mining, sports, health

<u>Research</u>



<u>Teaching</u>

- Master of AI: Big Data option
- Training courses
  - Data science in practice
  - Coping with big data

ML Group Research Goals: Desired Solution Characteristics

Is distance between players important?

Attacker

Location: (30,100)

Prefers right foot

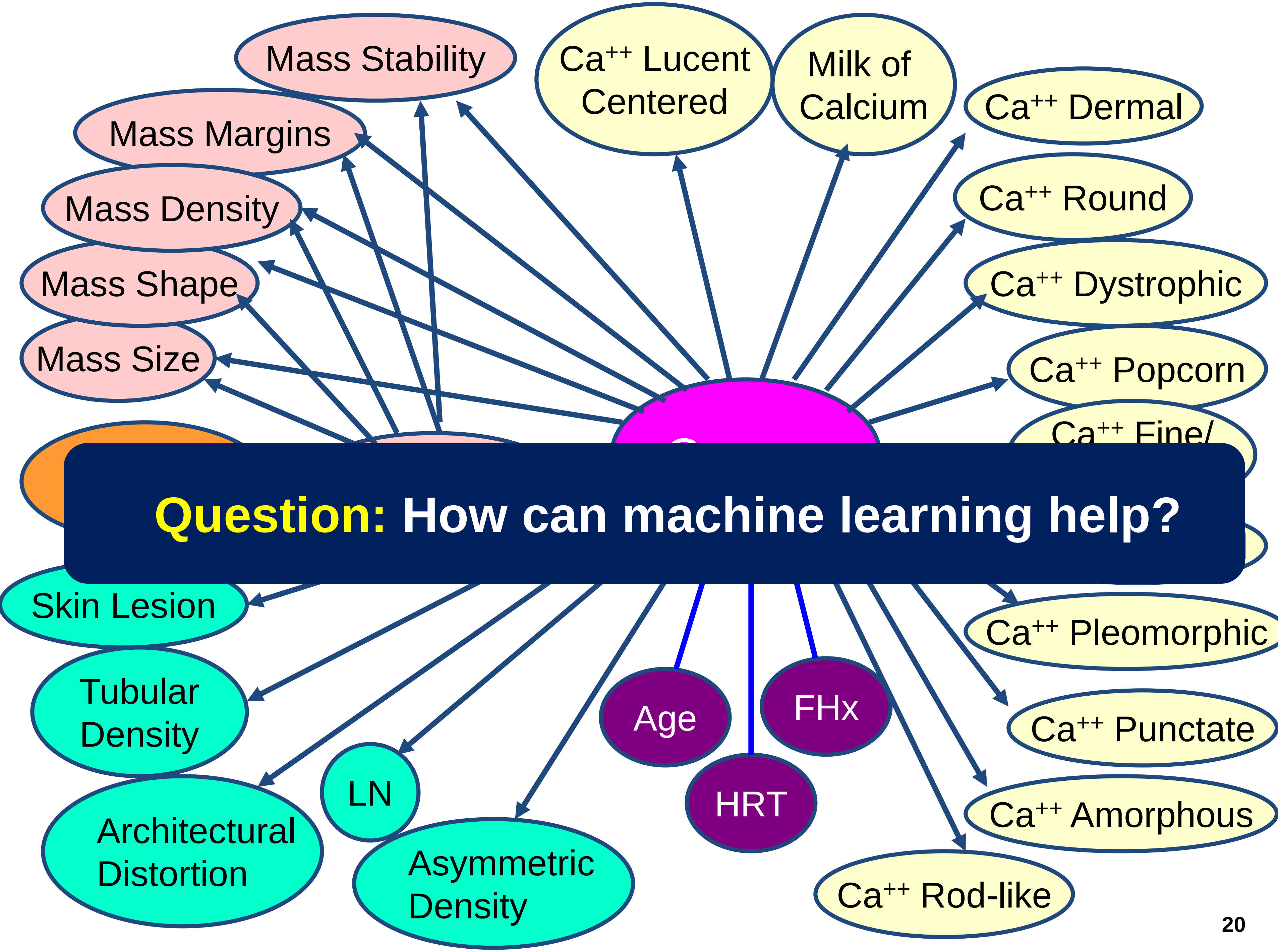
Tired?

$$\mathrm{dist}(p1,p2) < 2\mathrm{m} \wedge \mathrm{pr}(\mathrm{p1{=}tried}) > 0.8 \wedge \mathrm{prefRt}(\mathrm{p1}) \Rightarrow \mathrm{dribbleRt}(\mathrm{p1})$$

1. Represent discrete and continuous attributes
2. Model uncertainty
3. Capture important relationships
4. Incorporate domain knowledge
5. Produce interpretable output

Application: Mammography

- Provide decision support for radiologists
- Variability due to experience and training
- National mammography database schema
  - Pre-defined vocabulary to describe findings
  - Features of each abnormality
- How can we use this data?
  - Hand-crafted models
  - Machine learning

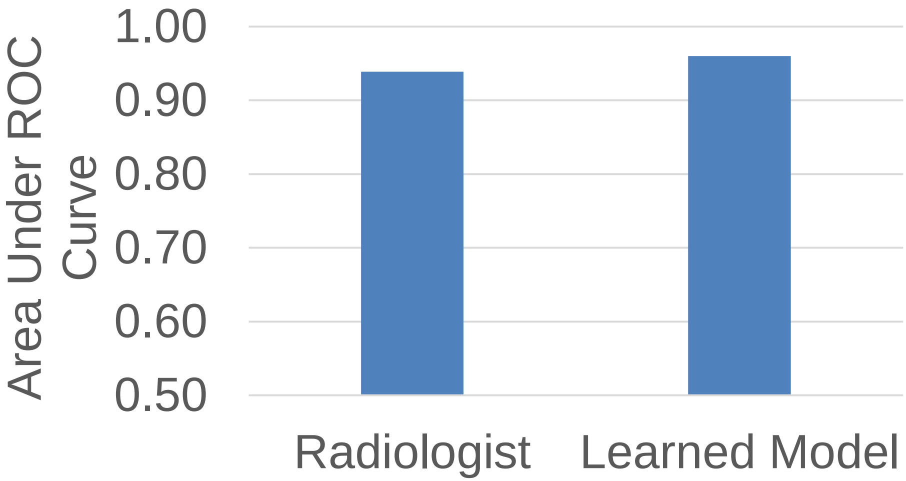


Question: How can machine learning help?

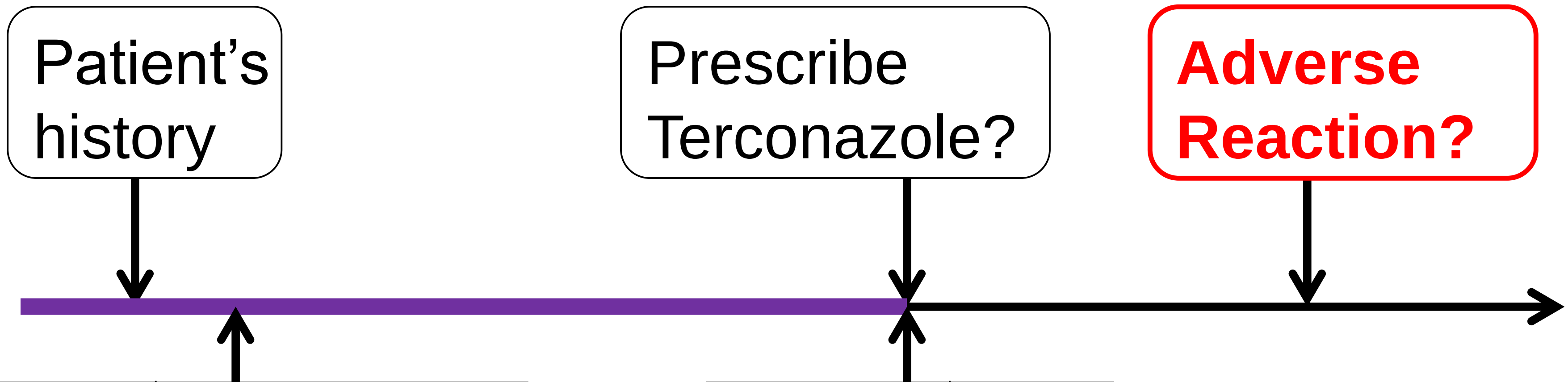
Methodology and Results

- Data: Matched to state-wide breast cancer registry to remove false negatives

- Compare: Radiologist vs. Bayesian network structure learning



Predicting Adverse Drug Events



PID	Date	Medication	Dose
P1	2/2/01	Warfarin	10mg

PID	Date	Weight
P1	2/2/03	120

Given: Patient's clinical history

Predict: At prescription time if the patient will have a known adverse reaction to drug

Challenge: Complex, Uncertain Data

Patient

PID	Birthday	Gender
P1	2/2/63	M
P2	4/7/55	M

Drug

PID	Date	Medication	Dose
P1	5/1/02	Warfarin	10mg
P1	2/2/03	Terconazole	10mg

Diseases

PID	Date	Diag.
P1	2/1/01	Flu
P1	5/2/03	Bleeding

Traditional Paradigms

Statistical Approach

+ Models uncertainty

− Ignores relations

Logical Approach

+ Models relations

− Ignores uncertainty

Statistical Relational Learning

- Combine graphical models (e.g., Bayes nets) with relational representations (e.g., first-order logic)
  - Problog (De Raedt et al., IJCAI'07)
  - Markov logic (Richardson & Domingos, MLJ'06)
  - Bayesian logic (Blog) (Milch et al., IJCAI'05)
  - Etc.
- Intuition: Attach probabilities to first-order rules to capture uncertainty
- Example: Smoking causes cancer

$$\mathrm{Smokes(person)} \Rightarrow \mathrm{Cancer(person)} \quad : 0.15$$

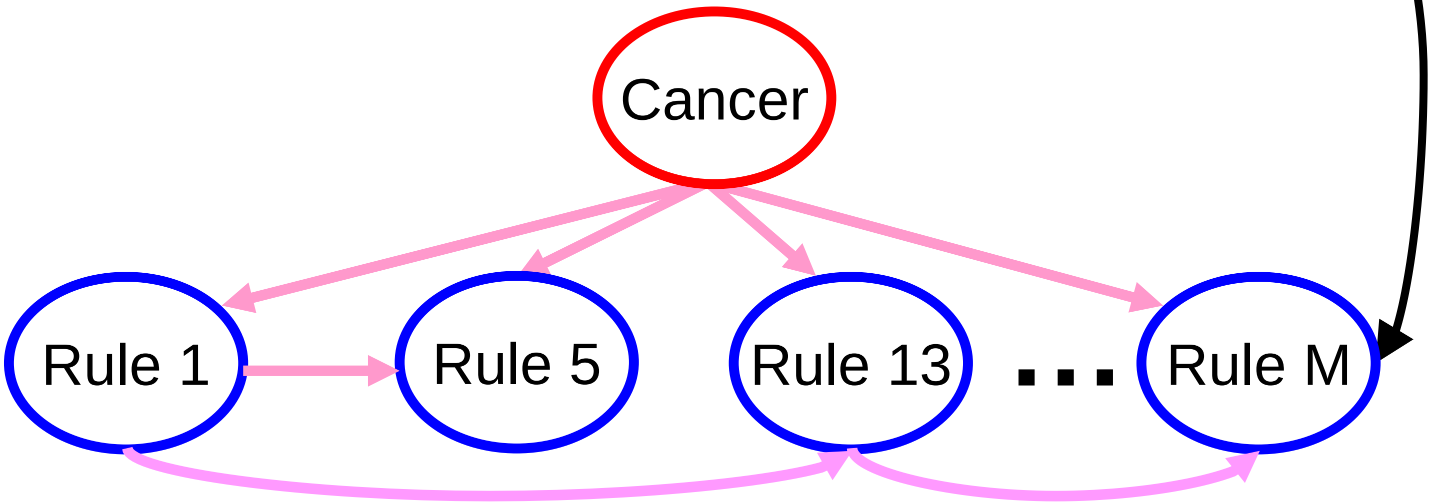
VISTA: A SRL Framework
[Davis et al., IJCAI'07, ICML'07, ICML'12]

Integrates feature induction and model construction

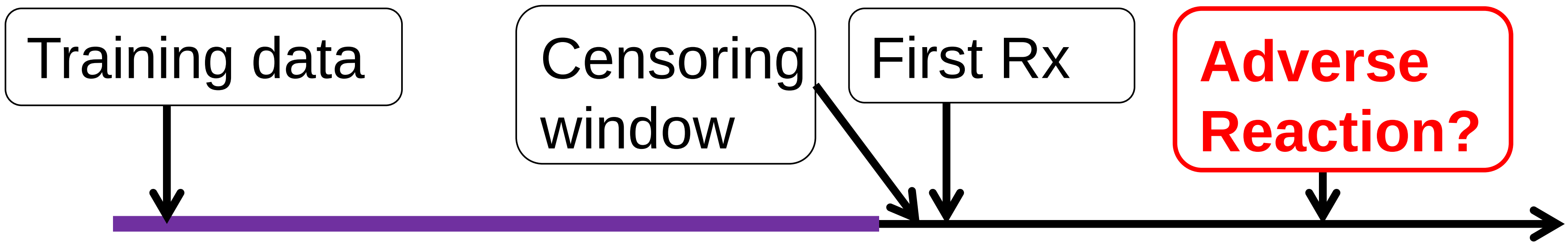
- If-then rules capture implicit, relational features

$$\boxed{\mathrm{Drug}(\mathrm{p}, \mathrm{Terconazole}) \wedge \mathrm{Wt}(\mathrm{p}, \mathrm{w}) \wedge \mathrm{w} < 120 \Rightarrow \mathrm{ADR}(\mathrm{p})}$$

- Rules become features in statistical model



Data Preparation

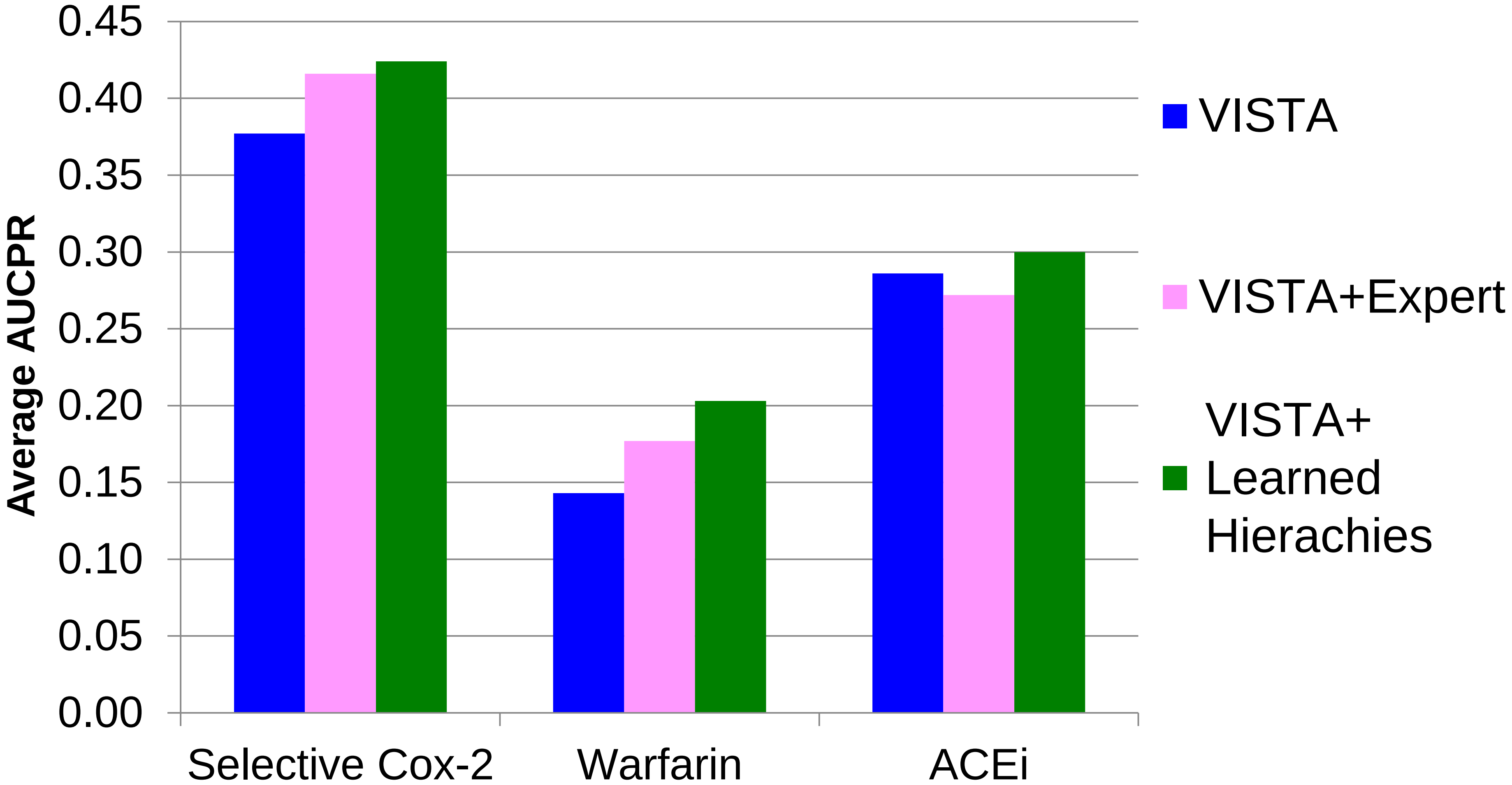


Positives: Adverse event after prescription

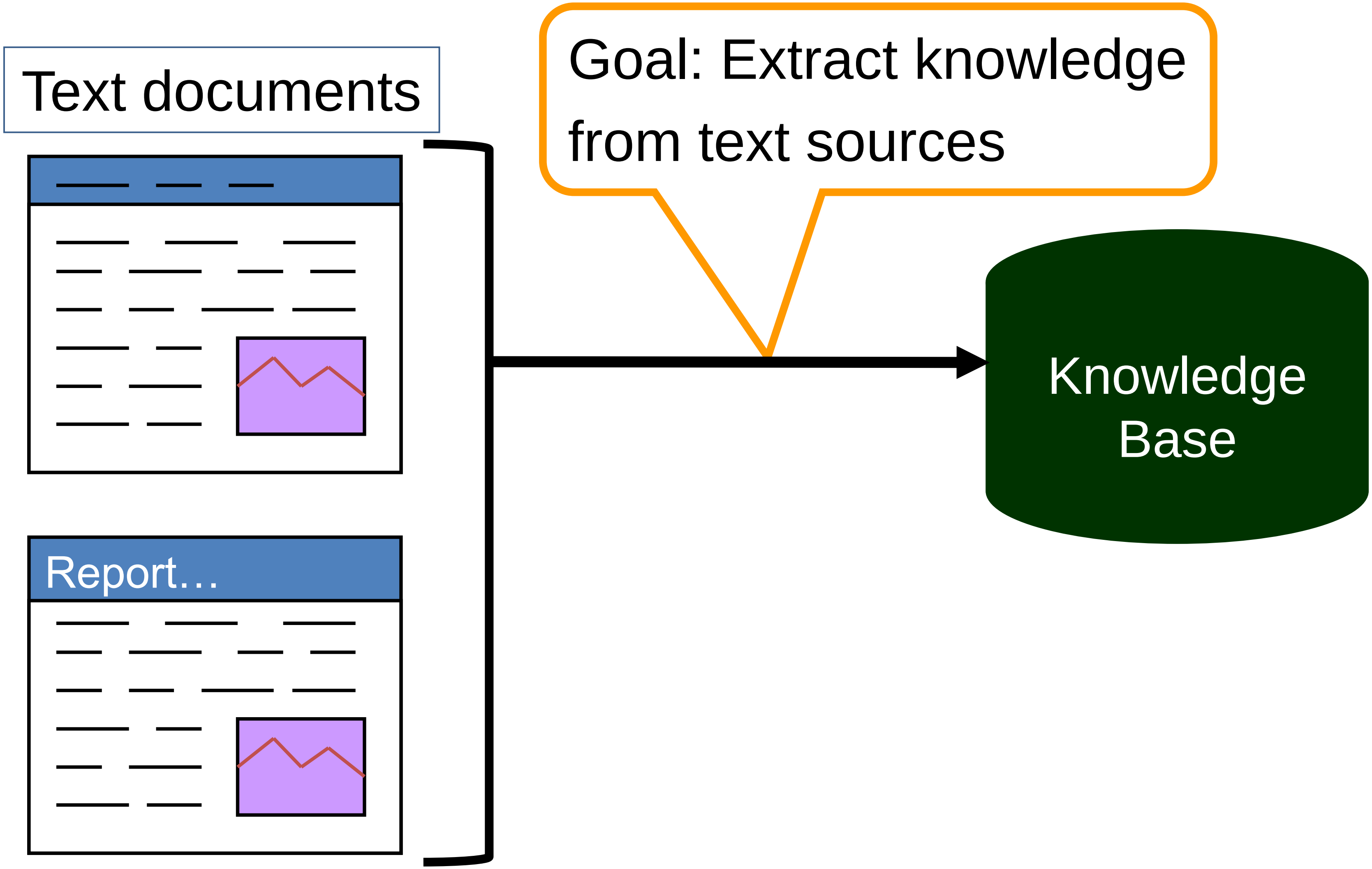
Negatives: Took medicine and no adverse event, matched on age and gender to positives

- Tasks considered:
  - Myocardial infarction on selective Cox-2 inhibitors
  - Internal bleeding with Warfarin
- Marshfield Clinic data: 1-10 million facts in DB
  - Diagnoses, Medications, Lab tests, Observations

Results


0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Average AUCPR
Selective Cox-2
Warfarin
ACEi
VISTA
VISTA+Expert
VISTA+ Learned Hierachies

Application: Automated Knowledge Base Construction



Example: Information Extraction

NELL: http://rtw.ml.cmu.edu/rtw/

Recently-Learned Facts twitter

Refresh

instance	iteration	date learned	confidence
kelly_andrews is a female	826	29-mar-2014	98.7
investment_next_year is an economic sector	829	10-apr-2014	95.3
shibenik is a geopolitical entity that is an organization	829	10-apr-2014	97.2
quality_web_design_work is a character trait	826	29-mar-2014	91.0
mercedes_benz_cls_by_carlsson is an automobile manufacturer	829	10-apr-2014	95.2
social_work is an academic program at the university rutgers_university	827	02-apr-2014	93.8
dante wrote the book the_divine_comedy	826	29-mar-2014	93.8
willie_aames was born in the city los_angeles	831	16-apr-2014	100.0
kitt_peak is a mountain in the state or province arizona	831	16-apr-2014	96.9
greenwich is a park in the city london	831	16-apr-2014	100.0

instances for many different relations

degree of certainty

Other systems: OpenIE (Oren Etzioni, AI2), DeepDive (Chris Re, Stanford), Google Knowledge Vault, etc.

Challenge: Fusing Information

Question: What prevents osteoporosis?



"gorillas prevent osteoporosis"

Web Images Videos News More Search tools

2 results (0.16 seconds)

Scholarly articles for cauliflower prevents osteoporosis

Onion and a mixture of vegetables, salads, and herbs ... - Mühlbauer - Cited by 103

... K in the prevention of fractures due to **osteoporosis** - Meunier - Cited by 32

Some vegetables (commonly consumed by humans) ... - Mühlbauer - Cited by 13

No results found for **"cauliflower prevents osteoporosis"**.

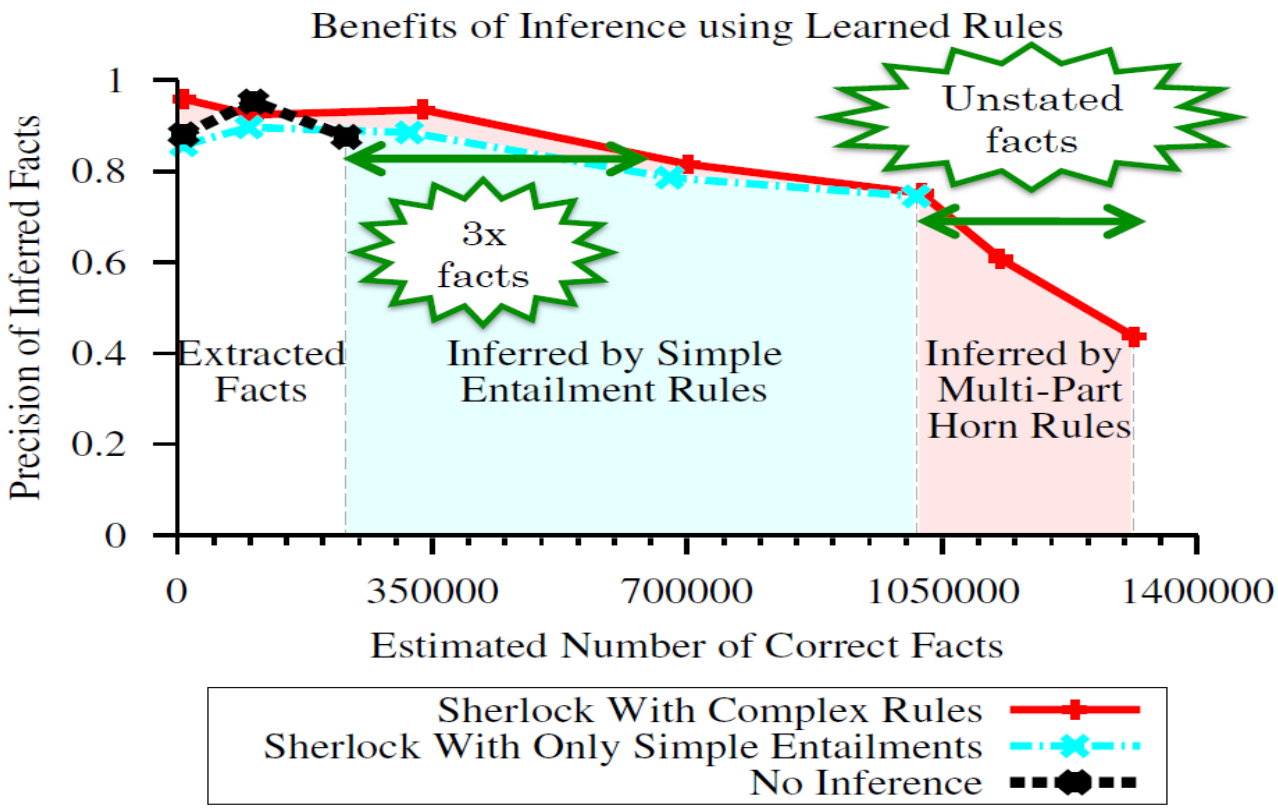
Solution: Rule Learning

[Schoenmakers et al., EMNLP'10, De Raedt et al., IJCAI'15, Zupanc & Davis, in preparation]

- Combine facts from multiple pages to **infer** the answer
  - Cauliflower contains calcium (7,700 pages)
  - Calcium prevents osteoporosis (43,500 pages)
  - ∴ Cauliflower prevents osteoporosis (with high probability)
- Key algorithmic challenges
  - KB only has "true facts": No negative examples!
  - Facts have associated confidences
  - Assign confidences to rules

* Results from google.com, November 2014

Results [Schoenmakers et al., EMNLP'10]



Took < 1 hour on 72 core cluster

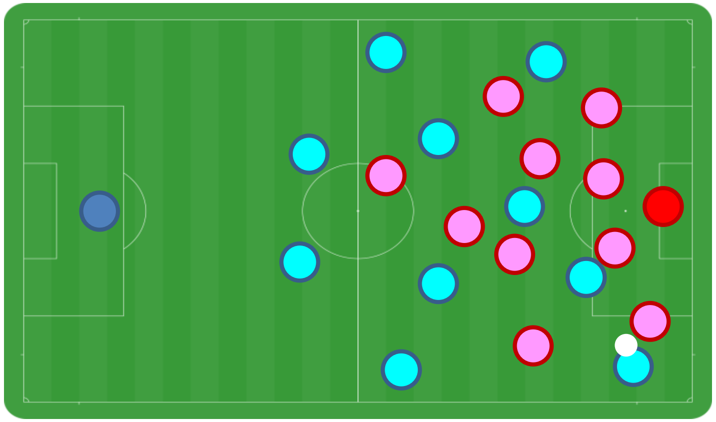
- Generate and evaluate >5M rules on 250k facts
- Make all inferences

Discover Offensive Strategies in Football Matches

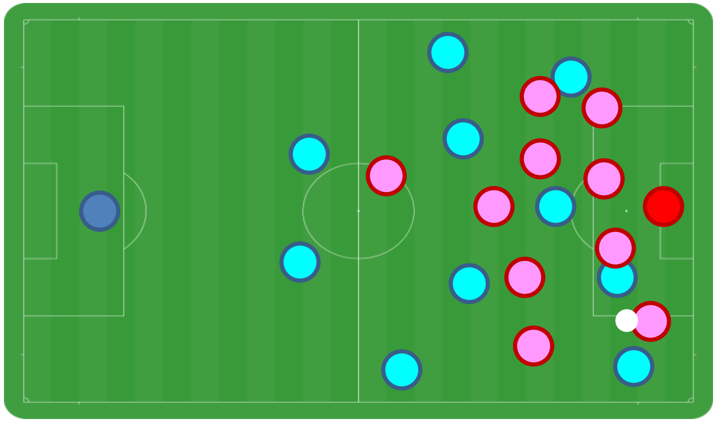
- **Given:** Streams with
  - Type (e.g., shot, pass, …) and location of all events
  - Locations of players and the ball (10 hz sample)

- **Find:** Typical offensive strategies
  - Film study is time consuming: Automation can help

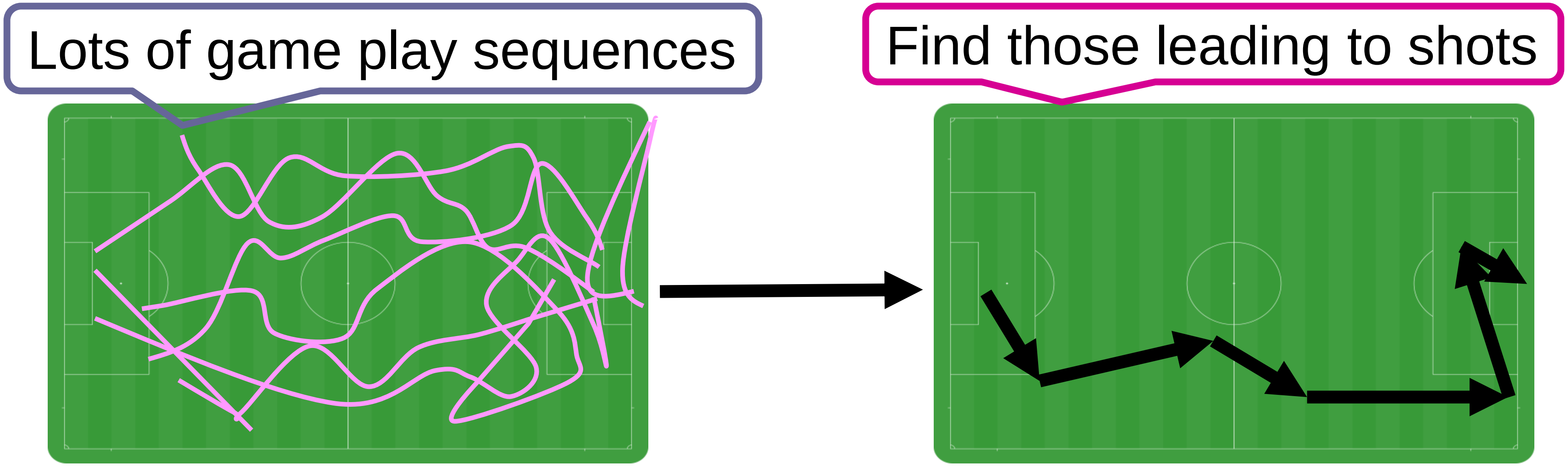
Time t



Time t+1



Task and Challenges

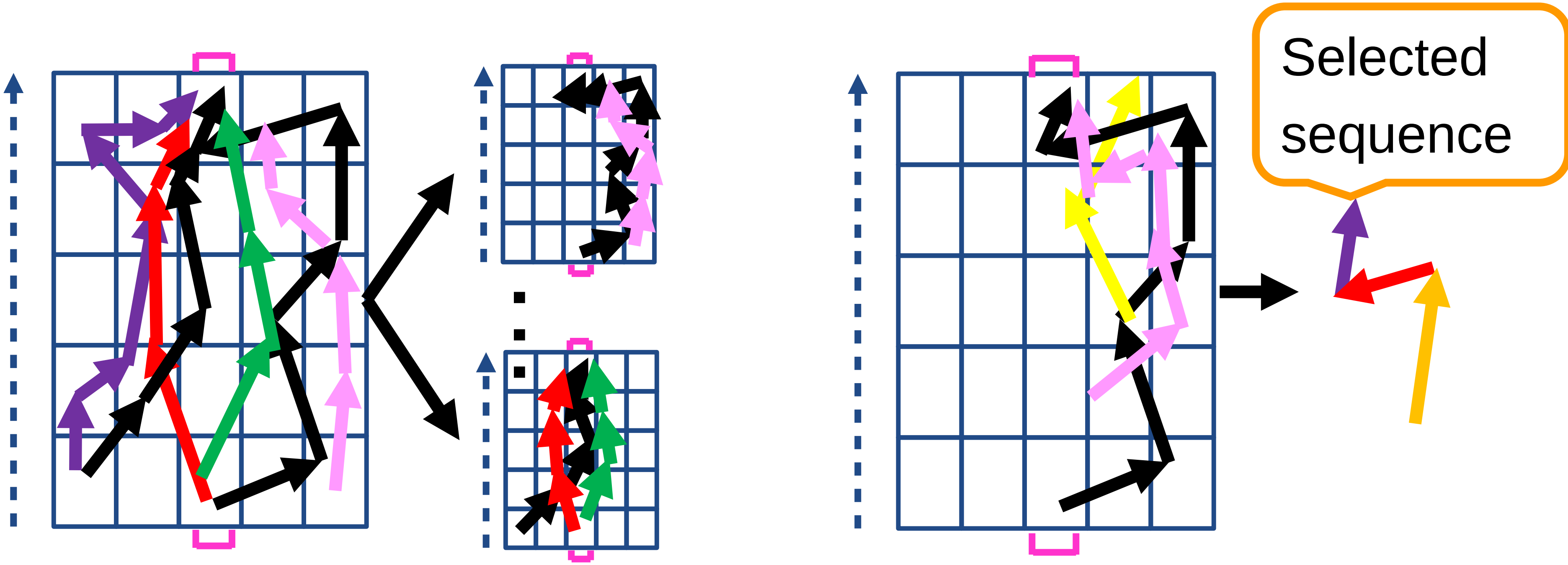


Challenges

- Data/match: ~1,250,000 locations, ~2000 events
- Model evolution of relationships among players over time and space
- No exact repetition of same sequence of events

Solution Sketch

[Van Haaren & Davis, LSSA'16]



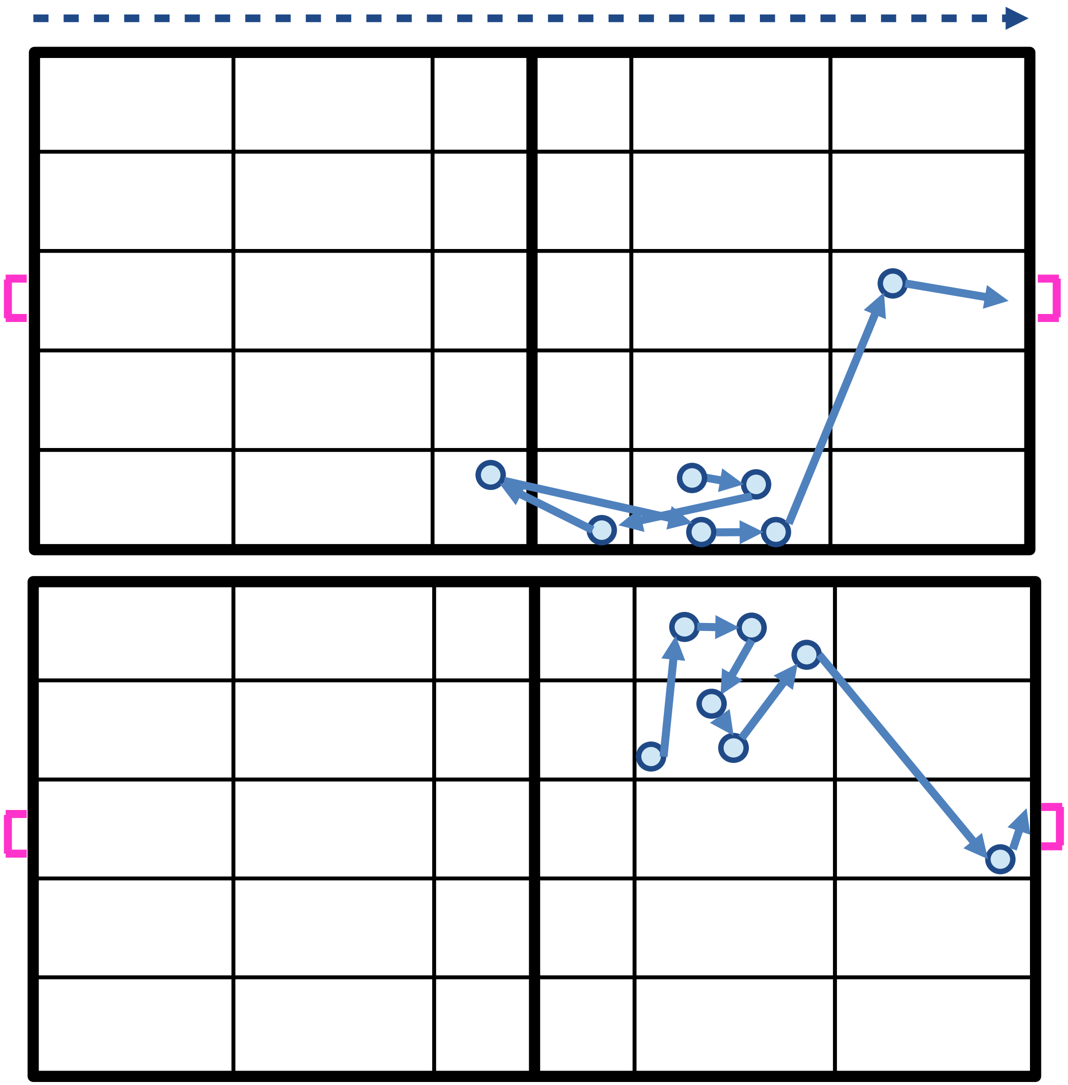
<u>Step 1: Cluster Data</u>

- Cluster ≈ strategy
- Generalize across locations
- Improved efficiency

<u>Step 2: Pattern Mining</u>

- Weigh actions
- Find commonly occurring subsequences

Two Representative Patterns



An attack down the right flank

An attack down the left flank

Conclusions

- Big data driving changes in many subfields of computer science
  - Systems, machine learning, data management, information retrieval, etc.
- ML group focus on novel analysis techniques
  - Reason about structured and uncertain data
  - Incorporate exert knowledge
  - Interpretable results
- Successes in health, sports, robotics, bioinformatics, etc.
- Always on the lookout for new collaborations and problems

Questions?

- Luc De Raedt
- Hendrik Blockeel
- Jan Van Haaren
- Wannes Meert
- Werner Helsen
- Benedicte Vanwanseele
- Jessa Bekker
- Tom Decroos
- Tim Op De Beeck
- Vincent Vercruyssen
- David Page
- Beth Burnside
- Vítor Santos Costa
- Ines Dutra
- Michael Caldwell
- Peggy Peissig
- Dan Weld
- Oren Etzioni
- Stef Schoenmakers

Conclusions

- Big data driving changes in many subfields of computer science
  - Systems, machine learning, data management, information retrieval, etc.
- ML group focus on novel analysis techniques
  - Reason about structured and uncertain data
  - Incorporate exert knowledge
  - Interpretable results
- Successes in health, sports, robotics, bioinformatics, etc.
- Always on the lookout for new collaborations and problems