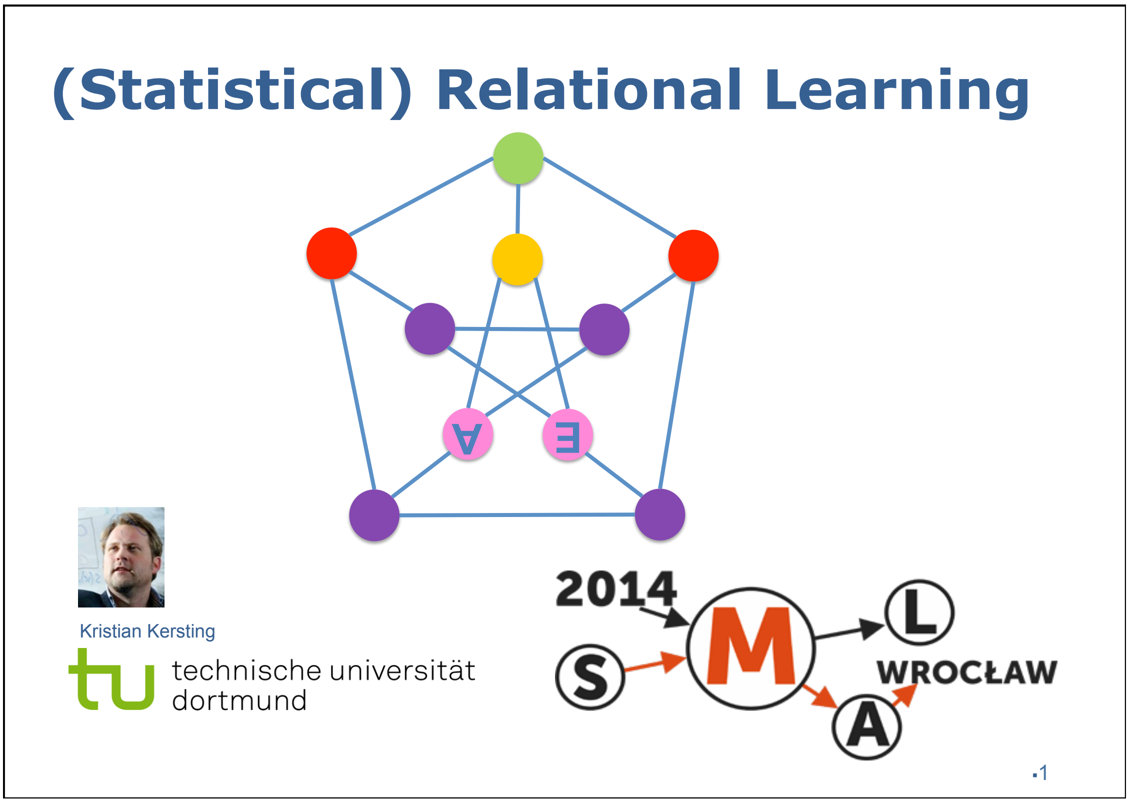


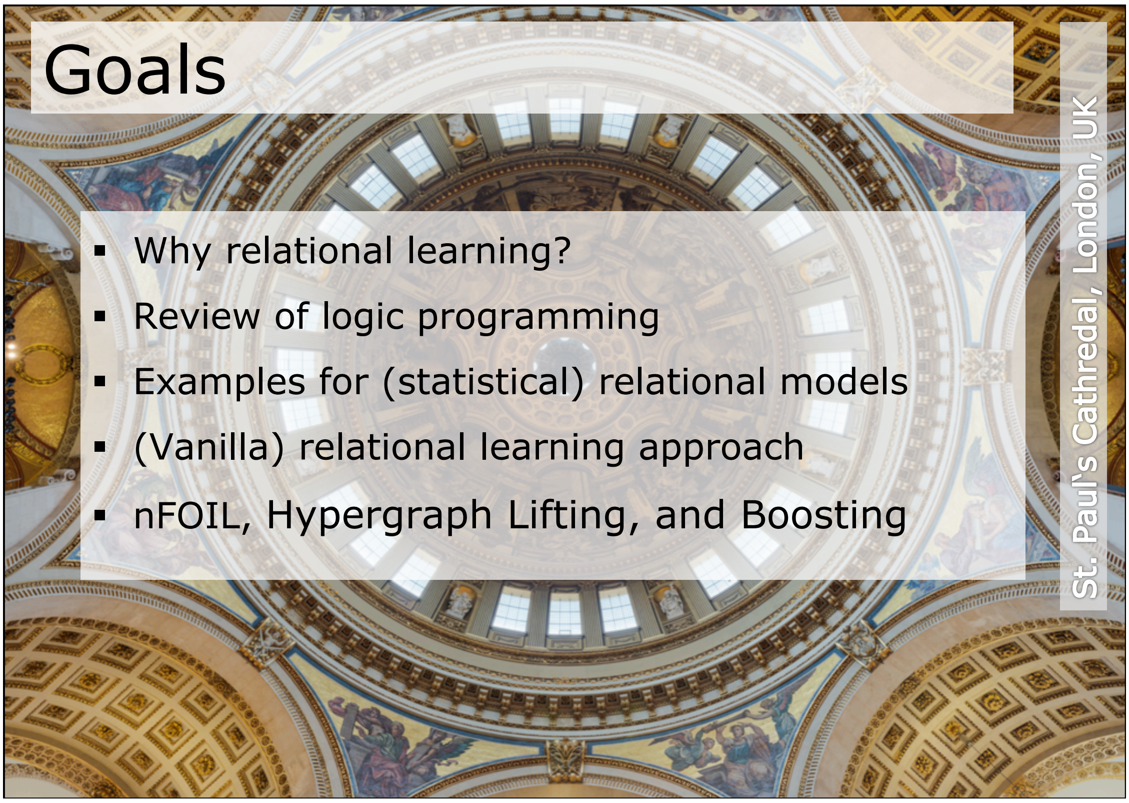
(Statistical) Relational Learning

Kristian Kersting

technische universität
dortmund

2014 S M L A WROCŁAW

▪1



Goals

- Why relational learning?
- Review of logic programming
- Examples for (statistical) relational models
- (Vanilla) relational learning approach
- nFOIL, Hypergraph Lifting, and Boosting

St. Paul's Cathredal, London, UK

Rorschach Test

Etzioni’s Rorschach Test for Computer S[illegible]ists

Moore’s Law?

Storage Capacity?

Number of Facebook Users?

Number of Scientific Publications?

Number of Web Pages?

Number of Actions?

Computing 2020: Science in an Exponential World

"The amount of scientific data is doubling every year"
[Szalay, Gray, Nature 440, 413-414 (23 March 2006)]

How to deal with millions of images ?

How to deal with millions of inter-related research papers ?

How to accumulate general knowledge automatically from the Web ?

How to deal with billions of shared users' perceptions stored at massive scale ?

How to realize the vision of social search?

Machine Learning in an Exponential World

ML = Structured Data + Model + Reasoning

Real world is structured in terms of objects and relations

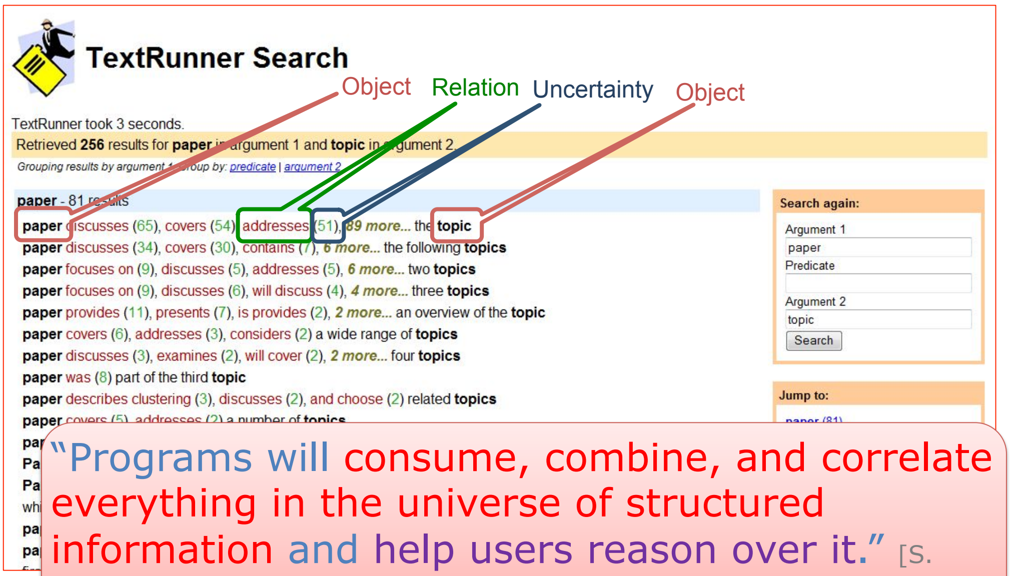
Relational knowledge can reveal additional correlations between variables of interest . Abstraction allows one to compactly model general knowledge and to move to complex inference

[Fergus et al. PAMI 30(11) 2008; Halevy et al., IEEE Intelligent Systems, 24 2009]

Most effort has gone into the modeling part

How much can the data itself help us to solve a problem?

[Etzioni et al. ACL08] http://www.cs.washington.edu/research/textrunner/



TextRunner Search

TextRunner took 3 seconds.

Retrieved **256** results for **paper** in argument 1 and **topic** in argument 2.

Grouping results by argument 1. Group by: predicate | argument 2

paper - 81 results

paper discusses (65), covers (54), addresses (51), *89 more...* the **topic**
paper discusses (34), covers (30), contains (7), *6 more...* the following **topics**
paper focuses on (9), discusses (5), addresses (5), *6 more...* two **topics**
paper focuses on (9), discusses (6), will discuss (4), *4 more...* three **topics**
paper provides (11), presents (7), is provides (2), *2 more...* an overview of the **topic**
paper covers (6), addresses (3), considers (2) a wide range of **topics**
paper discusses (3), examines (2), will cover (2), *2 more...* four **topics**
paper was (8) part of the third **topic**
paper describes clustering (3), discusses (2), and choose (2) related **topics**
paper covers (5), addresses (2) a number of **topics**

Search again:
Argument 1
paper
Predicate

Argument 2
topic
Search

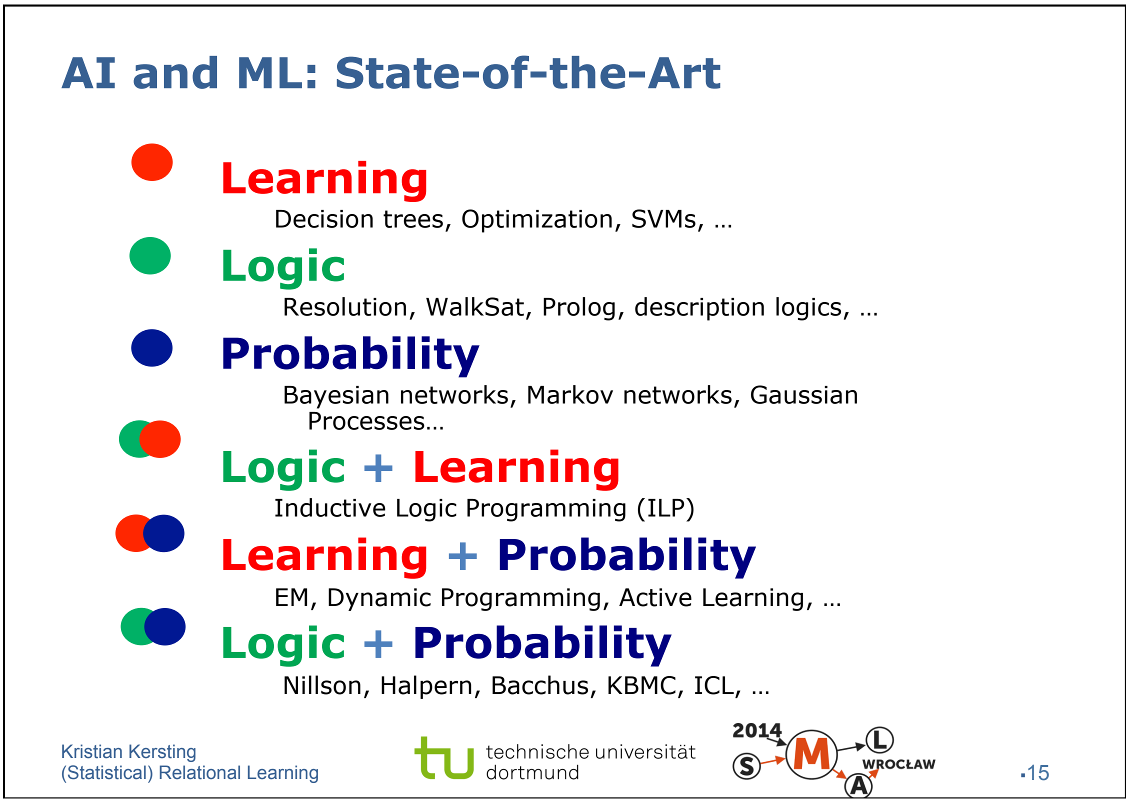
Jump to:
paper (81)

"Programs will consume, combine, and correlate everything in the universe of structured information and help users reason over it." [S. Parastatidis et al., Communications of the ACM Vol. 52(12):33-37]

So, the Real World is Complex and Uncertain

- Information overload
- Incomplete and contradictory information
- Many sources and modalities
- Variable number of objects and relations among them
- Rapid change

How can computer systems handle these ?



AI and ML: State-of-the-Art

● **Learning**

Decision trees, Optimization, SVMs, …

● **Logic**

Resolution, WalkSat, Prolog, description logics, …

● **Probability**

Bayesian networks, Markov networks, Gaussian Processes…

●● **Logic + Learning**

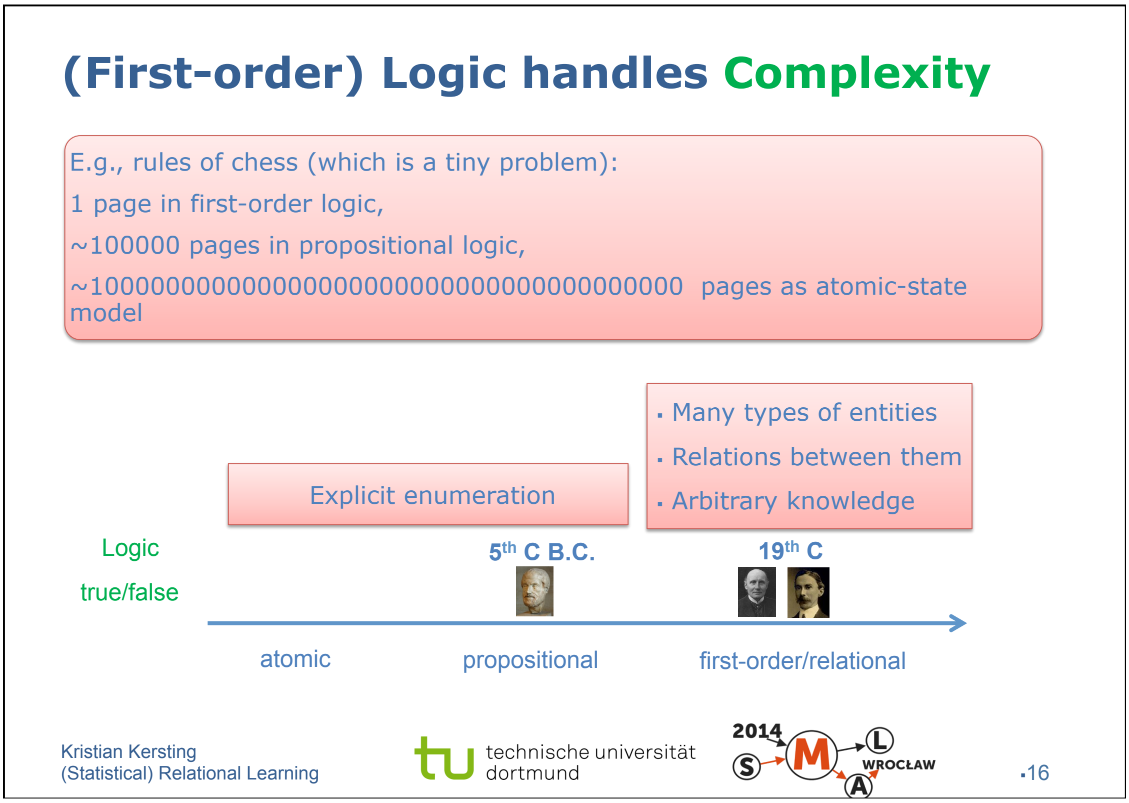
Inductive Logic Programming (ILP)

●● **Learning + Probability**

EM, Dynamic Programming, Active Learning, …

●● **Logic + Probability**

Nillson, Halpern, Bacchus, KBMC, ICL, …



(First-order) Logic handles Complexity

E.g., rules of chess (which is a tiny problem):

1 page in first-order logic,

~100000 pages in propositional logic,

~1000000000000000000000000000000000000000 pages as atomic-state model

Explicit enumeration

- Many types of entities
- Relations between them
- Arbitrary knowledge

Logic

true/false

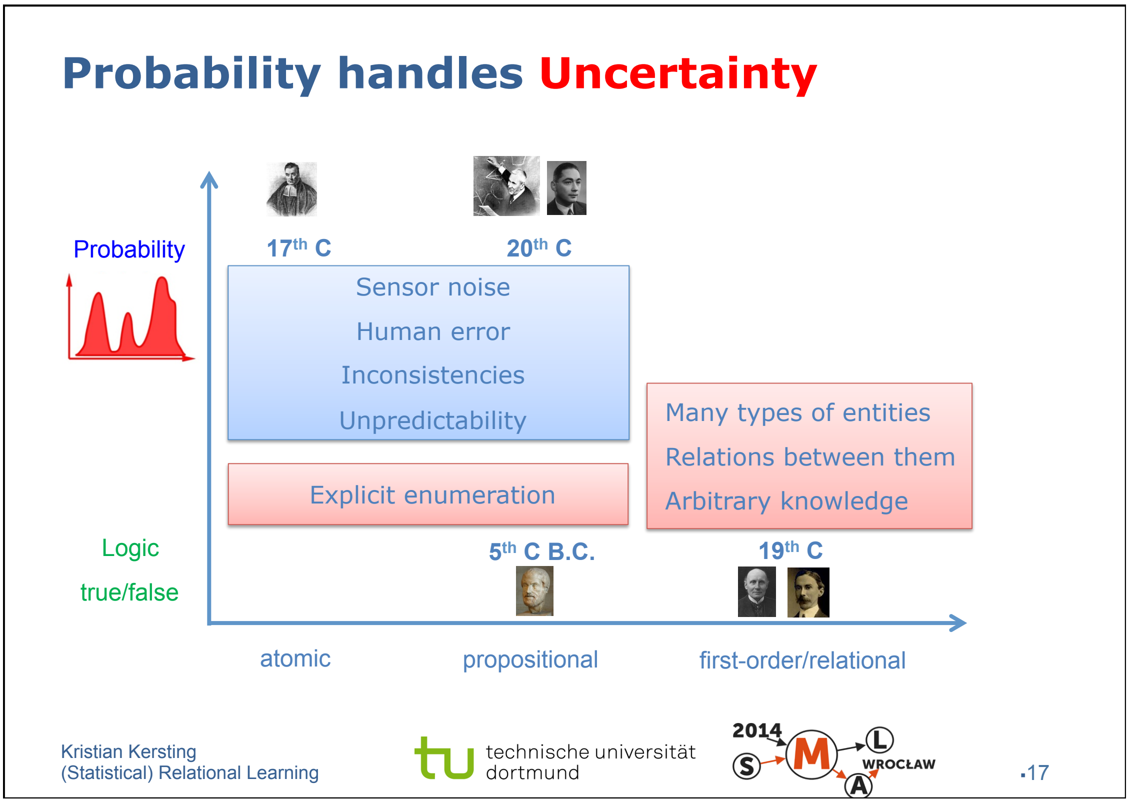
5th C B.C.

19th C

atomic

propositional

first-order/relational



Probability handles Uncertainty

Probability

Logic

true/false

17th C

20th C

Sensor noise

Human error

Inconsistencies

Unpredictability

Explicit enumeration

Many types of entities

Relations between them

Arbitrary knowledge

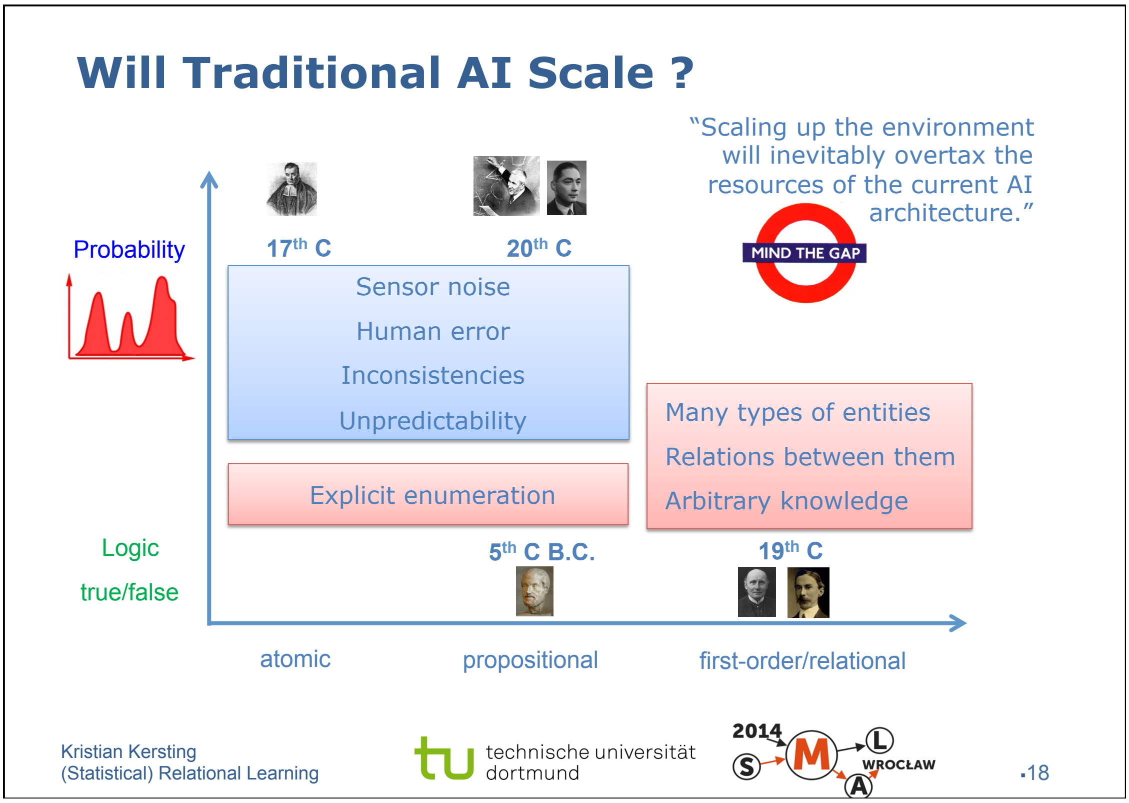
5th C B.C.

19th C

atomic

propositional

first-order/relational



Will Traditional AI Scale ?

"Scaling up the environment will inevitably overtax the resources of the current AI architecture."

Probability

Logic

true/false

17th C

20th C

Sensor noise

Human error

Inconsistencies

Unpredictability

Explicit enumeration

Many types of entities

Relations between them

Arbitrary knowledge

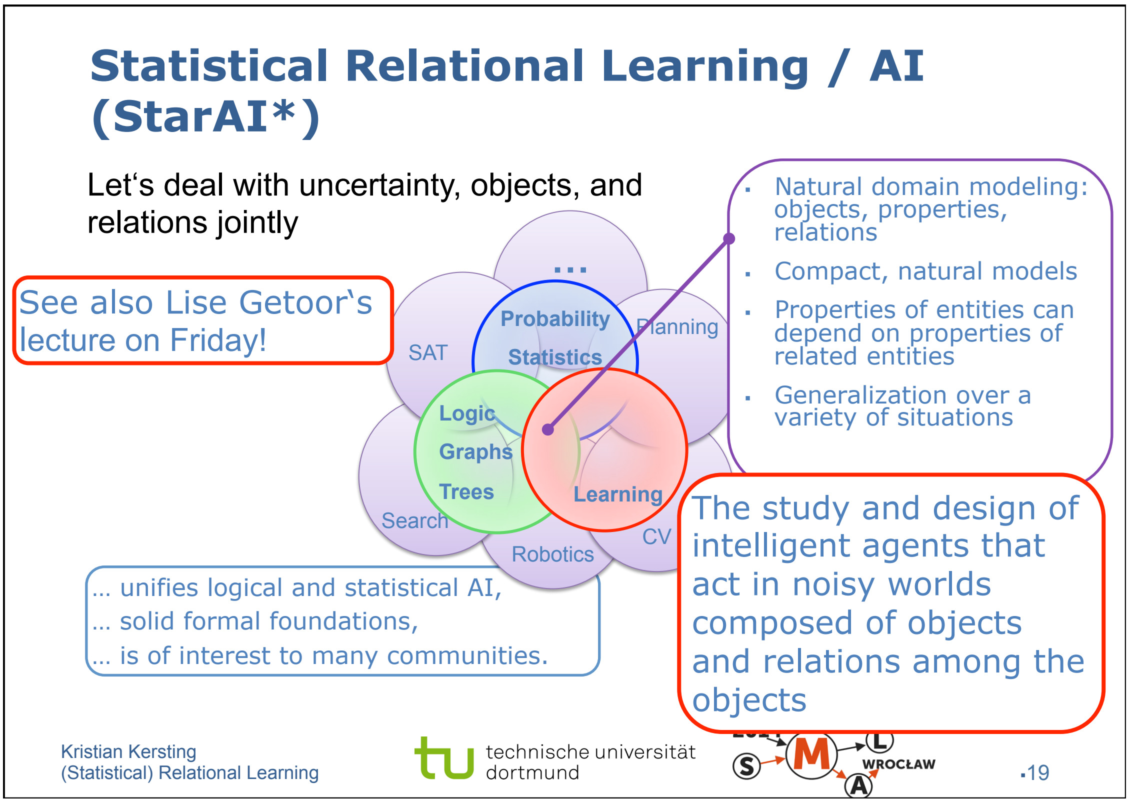
5th C B.C.

19th C

atomic

propositional

first-order/relational



Statistical Relational Learning / AI (StarAI*)

Let‘s deal with uncertainty, objects, and relations jointly

See also Lise Getoor‘s lecture on Friday!

- Natural domain modeling: objects, properties, relations
- Compact, natural models
- Properties of entities can depend on properties of related entities
- Generalization over a variety of situations

The study and design of intelligent agents that act in noisy worlds composed of objects and relations among the objects

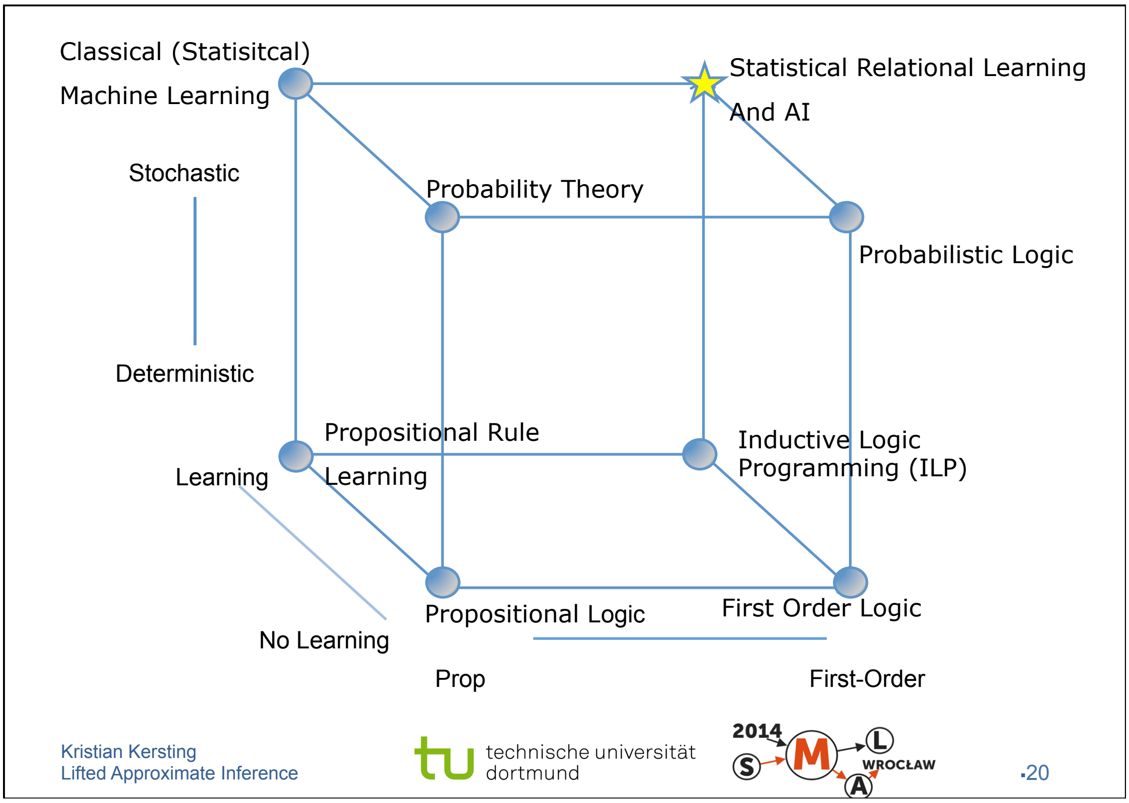
… unifies logical and statistical AI,
… solid formal foundations,
… is of interest to many communities.

Kristian Kersting
(Statistical) Relational Learning

technische universität dortmund

S M L A WROCŁAW

•19



Kristian Kersting
Lifted Approximate Inference

technische universität dortmund

2014 S M L A WROCŁAW

•20

Let’s consider a simple example: Reviewing Papers

- The grade of a paper at a conference depends on the paper’s quality and the difficulty of the conference.
  - **Good papers may get A’s at easy conferences**
  - **Good papers may get D’s at top conference**
  - **Weak papers may get B’s at good conferences**
  - **…**

Propositional Logic

- **Good papers get A’s at easy conferences**
  - good(p1)∧conference(c1,easy)⇒grade(p1,c1,a)
  - good(p2)∧conference(c1,easy)⇒grade(p2,c1,a)
  - good(p3)∧conference(c3,easy)⇒grade(p3,c3,a)

Number of statements explodes with the number of papers and conferences

No generalities, thus no (easy) generalization

First Order Logic

- The grade of a paper at a conference depends on the paper's quality and the difficulty of the conference.
  - **Good papers get A's at easy conferences**

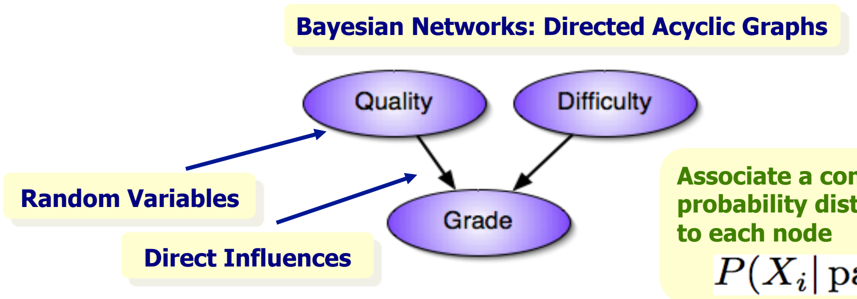
- ∀P,C [good(P)∧conference(C,easy)⇒grade(P,C,a)]

Many 'all universals' are (almost) false

Even good papers can get either A, B, C

True universals are rarely useful

Modeling the Uncertainty Explicitely

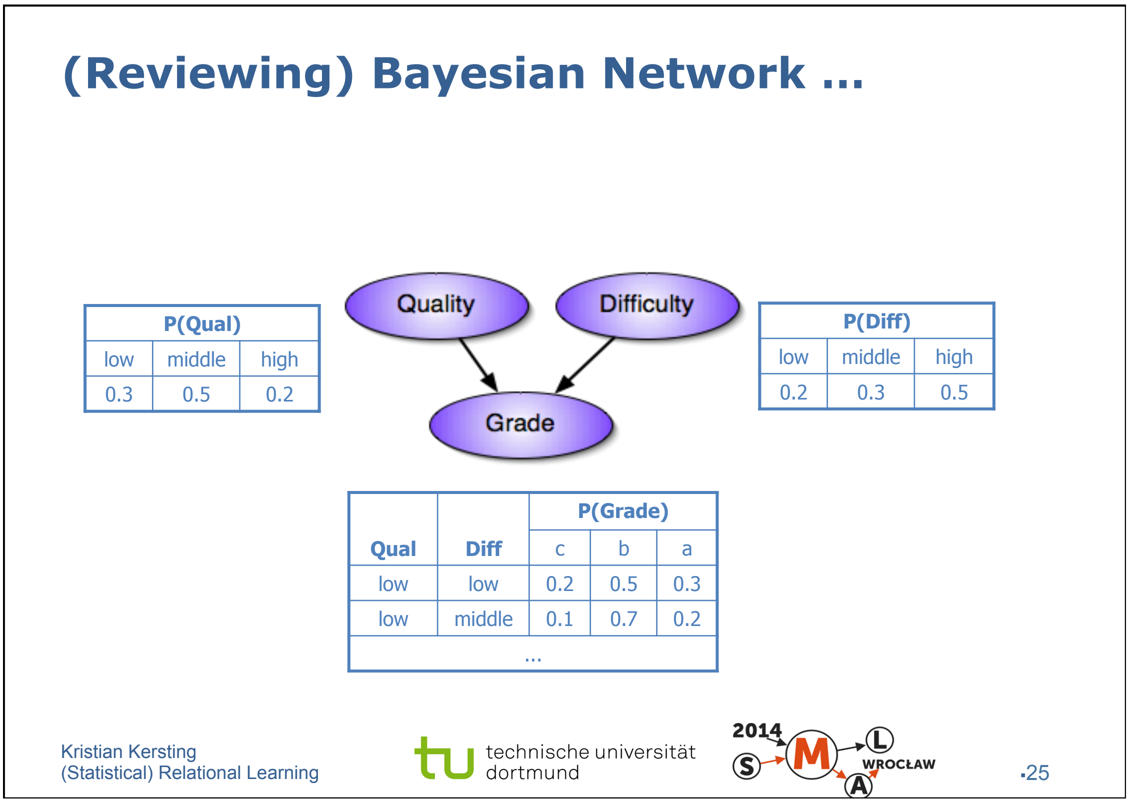


Associate a conditional probability distribution to each node

$$P(X_i|\,\mathrm{pa}(X_i))$$

Compact representation of the joint probability distribution

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P(X_i|\,\mathrm{pa}(X_i))$$



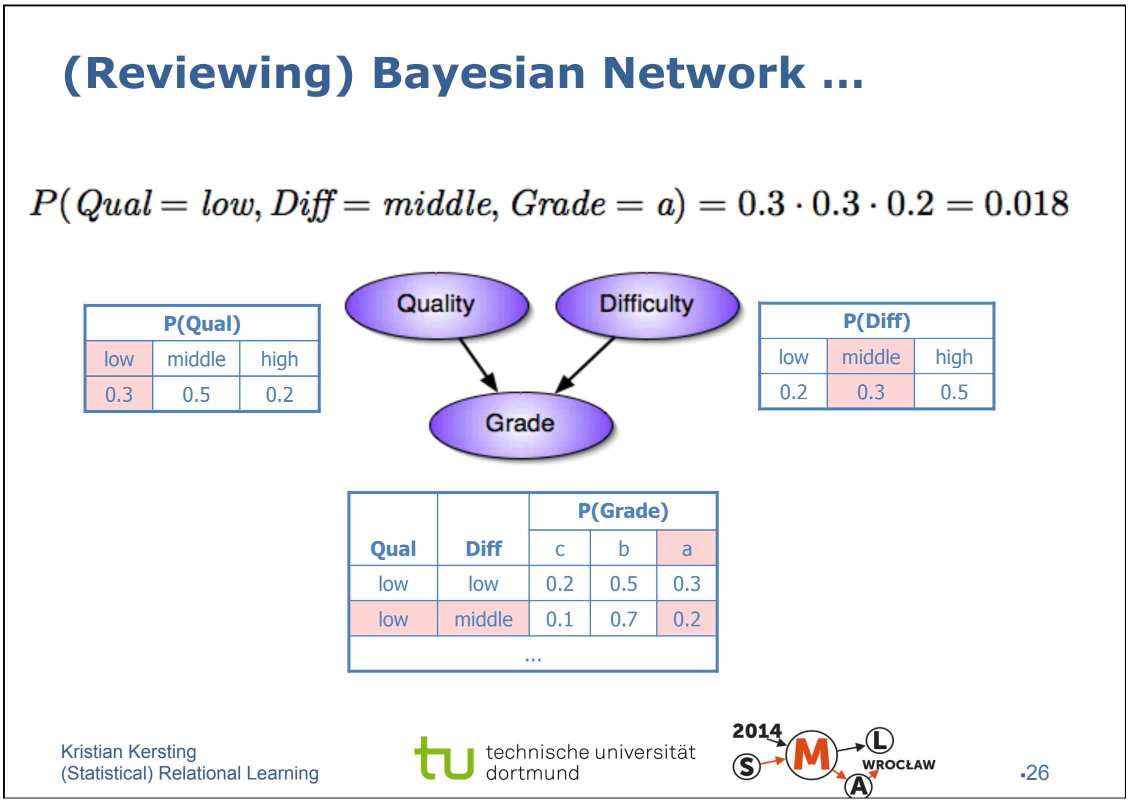
(Reviewing) Bayesian Network …

Quality
Difficulty
Grade

P(Qual)		
low	middle	high
0.3	0.5	0.2

P(Diff)		
low	middle	high
0.2	0.3	0.5

		P(Grade)		
Qual	Diff	c	b	a
low	low	0.2	0.5	0.3
low	middle	0.1	0.7	0.2
…				



(Reviewing) Bayesian Network …

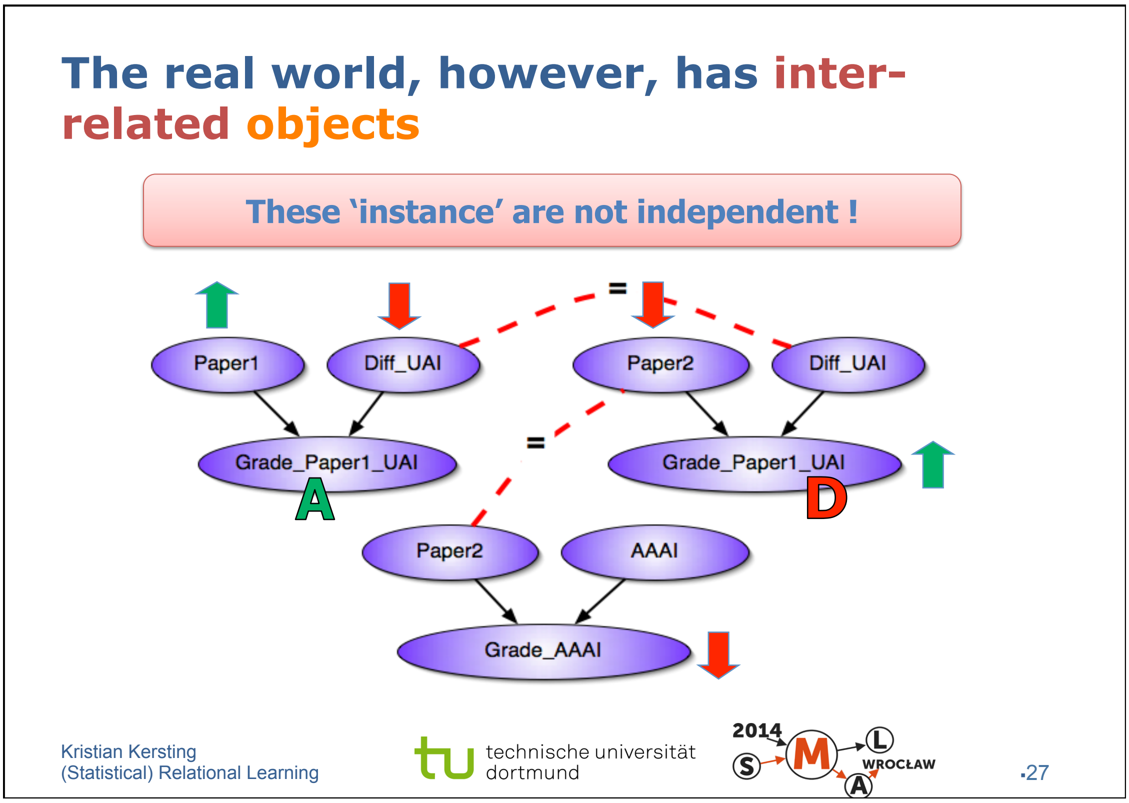
$$P(Qual = low, Diff = middle, Grade = a) = 0.3 \cdot 0.3 \cdot 0.2 = 0.018$$

Quality
Difficulty
Grade

P(Qual)		
low	middle	high
0.3	0.5	0.2

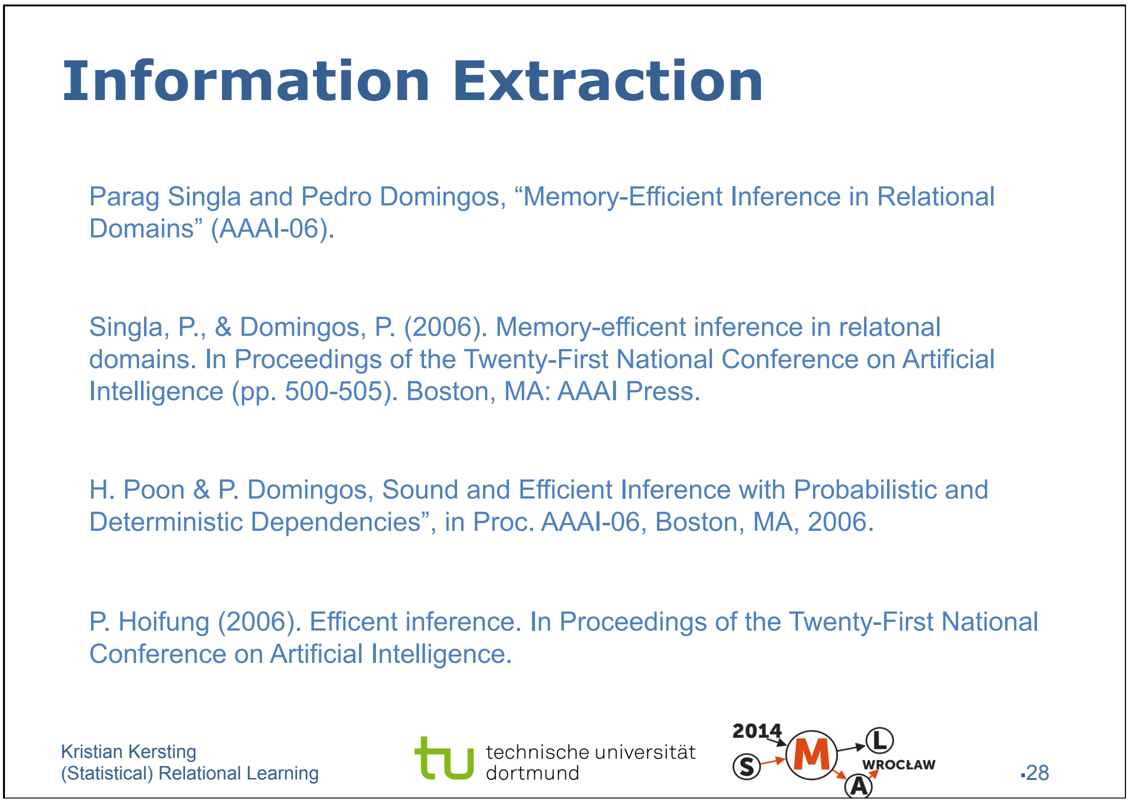
P(Diff)		
low	middle	high
0.2	0.3	0.5

		P(Grade)		
Qual	Diff	c	b	a
low	low	0.2	0.5	0.3
low	middle	0.1	0.7	0.2
…				



The real world, however, has inter-related objects

These 'instance' are not independent !



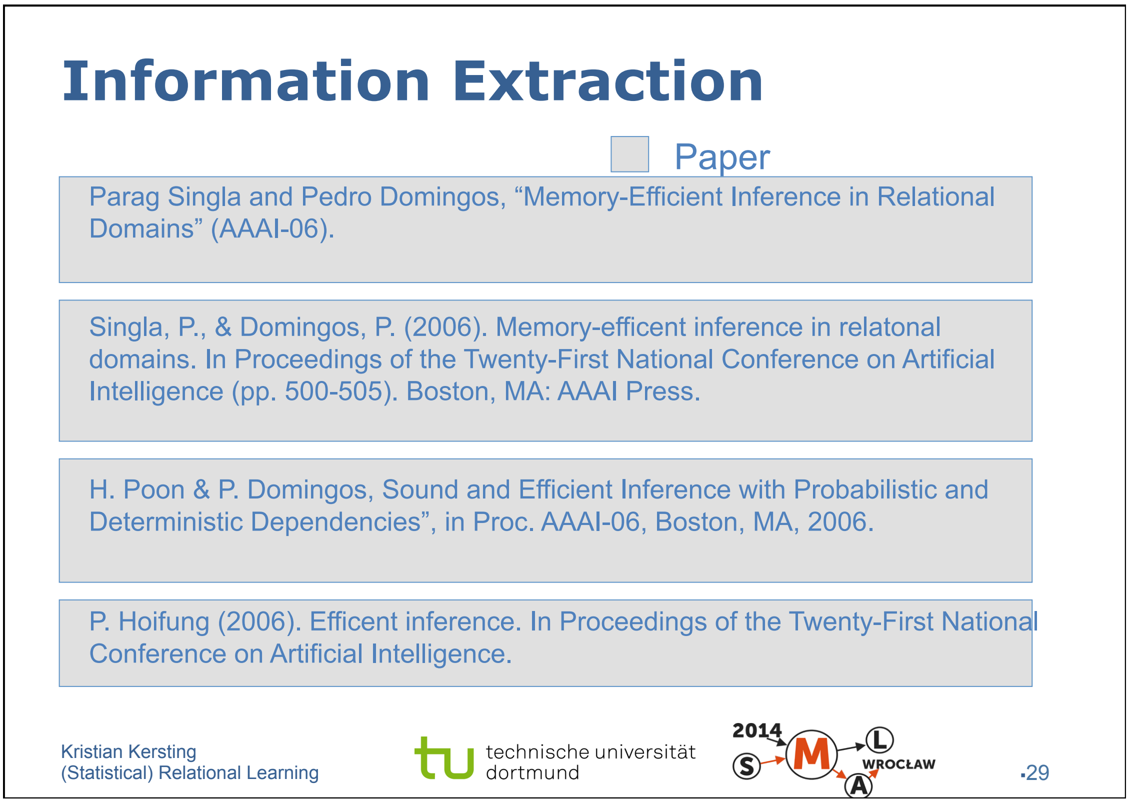
Information Extraction

Parag Singla and Pedro Domingos, "Memory-Efficient Inference in Relational Domains" (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies", in Proc. AAAI-06, Boston, MA, 2006.

P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.



Information Extraction

▨ Paper

Parag Singla and Pedro Domingos, “Memory-Efficient Inference in Relational Domains” (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies”, in Proc. AAAI-06, Boston, MA, 2006.

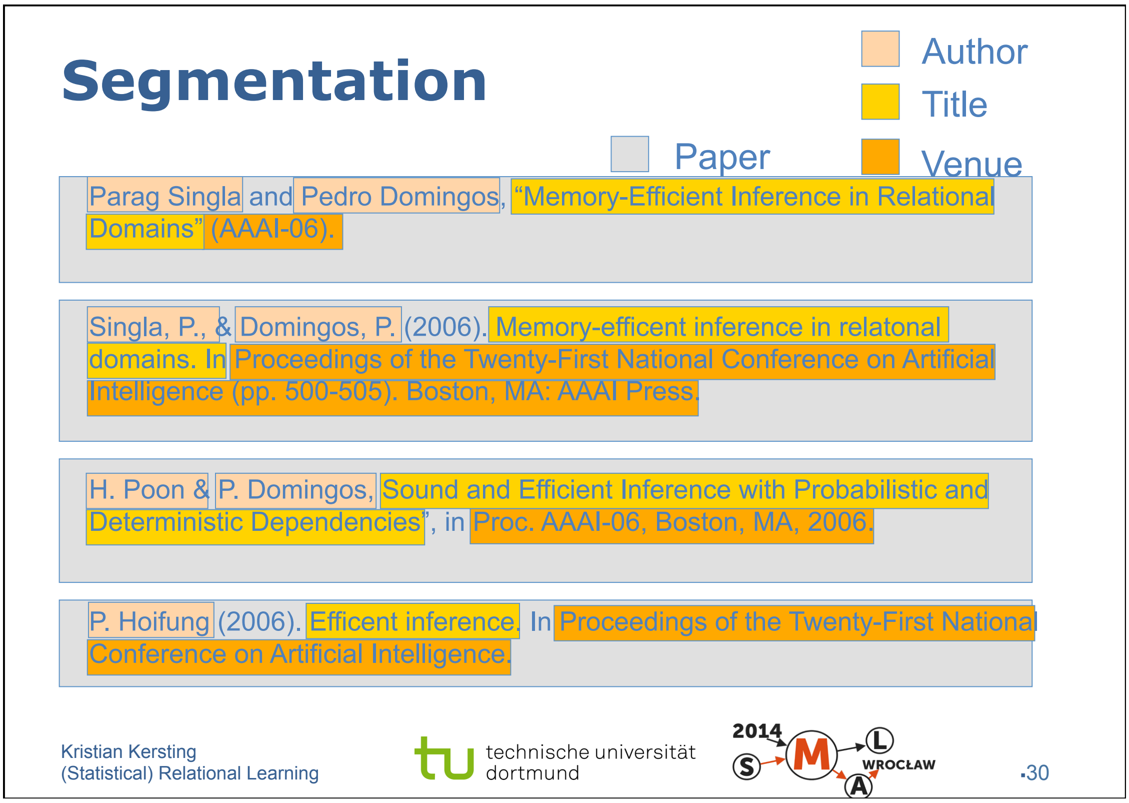
P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 M S L A WROCŁAW

▪29



Segmentation

🟧 Author
🟨 Title
▨ Paper
🟧 Venue

Parag Singla and Pedro Domingos, “Memory-Efficient Inference in Relational Domains” (AAAI-06).

Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies”, in Proc. AAAI-06, Boston, MA, 2006.

P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

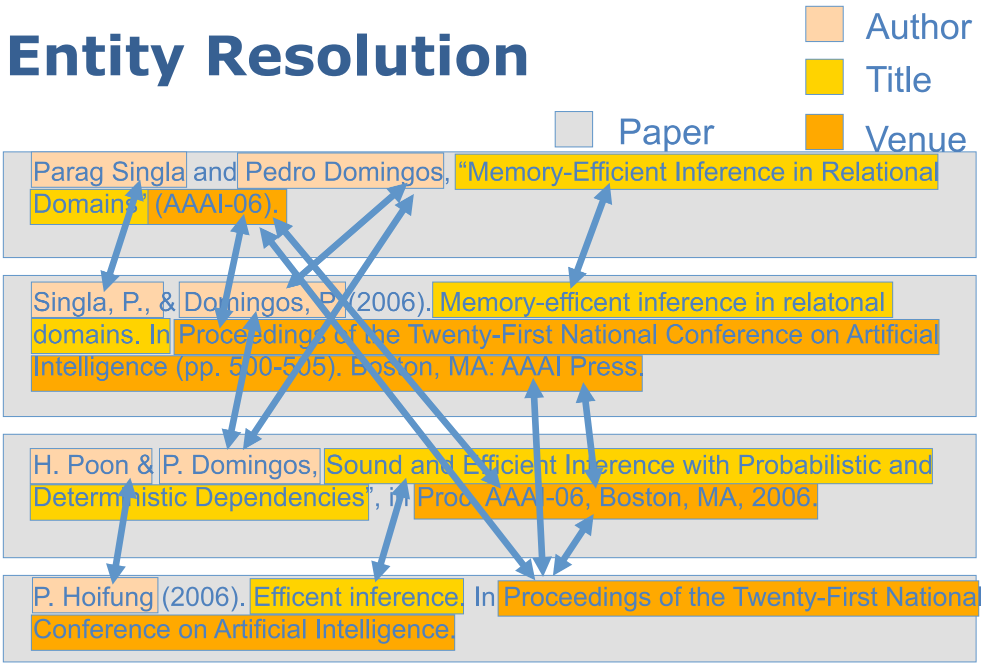
Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 M S L A WROCŁAW

▪30

Entity Resolution



Parag Singla and Pedro Domingos, “Memory-Efficient Inference in Relational Domains” (AAAI-06).

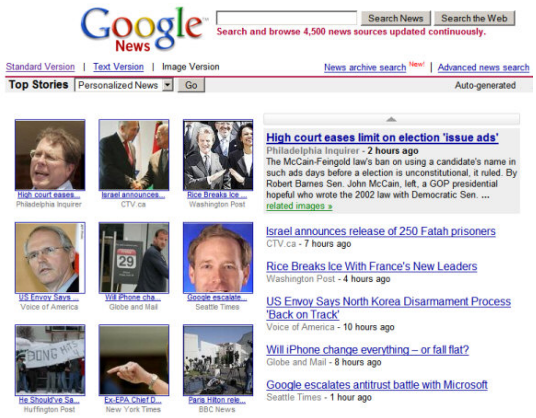
Singla, P., & Domingos, P. (2006). Memory-efficent inference in relatonal domains. In Proceedings of the Twenty-First National Conference on Artificial Intelligence (pp. 500-505). Boston, MA: AAAI Press.

H. Poon & P. Domingos, Sound and Efficient Inference with Probabilistic and Deterministic Dependencies”, in Proc. AAAI-06, Boston, MA, 2006.

P. Hoifung (2006). Efficent inference. In Proceedings of the Twenty-First National Conference on Artificial Intelligence.

Again, ‘instance’ are not independent !

Topic Models



High court eases limit on election 'issue ads'
Philadelphia Inquirer - 2 hours ago
The McCain-Feingold law's ban on using a candidate's name in such ads days before a election is unconstitutional, it ruled. By Robert Barnes Sen. John McCain, left, a GOP presidential hopeful who wrote the 2002 law with Democratic Sen. ...
related images »

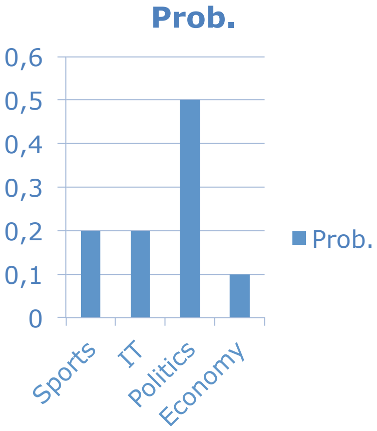
Israel announces release of 250 Fatah prisoners
CTV.ca - 7 hours ago

Rice Breaks Ice With France's New Leaders
Washington Post - 4 hours ago

US Envoy Says North Korea Disarmament Process 'Back on Track'
Voice of America - 10 hours ago

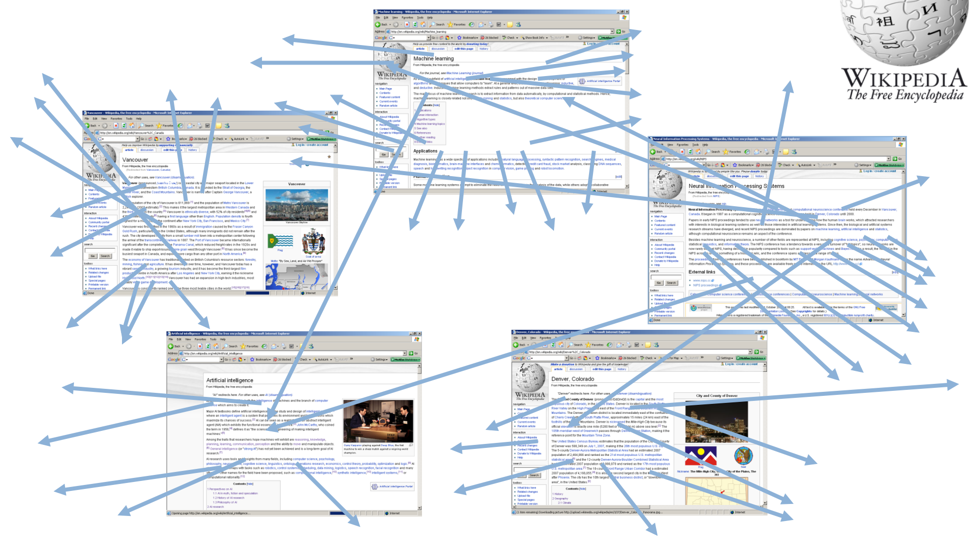
Will iPhone change everything – or fall flat?
Globe and Mail - 8 hours ago

Google escalates antitrust battle with Microsoft
Seattle Times - 1 hour ago



Wikipedia

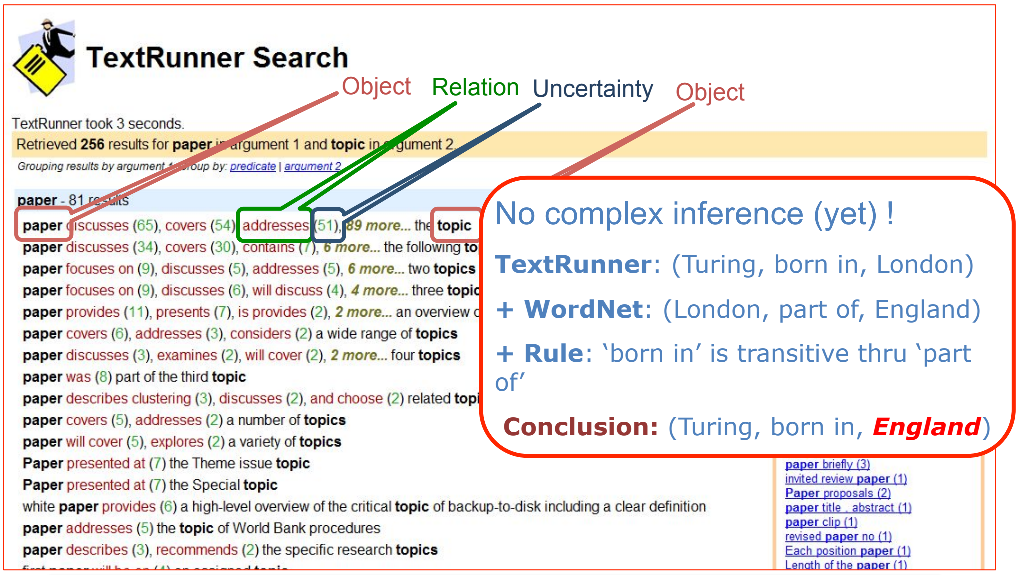
WIKIPEDIA
The Free Encyclopedia



Again, 'instance' are not independent !

[Etzioni et al. ACL08]

http://www.cs.washington.edu/research/textrunner/



TextRunner took 3 seconds.

Retrieved **256** results for **paper** in argument 1 and **topic** in argument 2.

Grouping results by argument 1. Group by: predicate | argument 2

paper - 81 results

paper discusses (65), covers (54), addresses (51), *89 more...* the **topic**
paper discusses (34), covers (30), contains (7), *6 more...* the following to
paper focuses on (9), discusses (5), addresses (5), *6 more...* two **topics**
paper focuses on (9), discusses (6), will discuss (4), *4 more...* three topic
paper provides (11), presents (7), is provides (2), *2 more...* an overview o
paper covers (6), addresses (3), considers (2) a wide range of **topics**
paper discusses (3), examines (2), will cover (2), *2 more...* four **topics**
paper was (8) part of the third **topic**
paper describes clustering (3), discusses (2), and choose (2) related topi
paper covers (5), addresses (2) a number of **topics**
paper will cover (5), explores (2) a variety of **topics**
Paper presented at (7) the Theme issue **topic**
Paper presented at (7) the Special **topic**
white **paper** provides (6) a high-level overview of the critical **topic** of backup-to-disk including a clear definition
paper addresses (5) the **topic** of World Bank procedures
paper describes (3), recommends (2) the specific research **topics**

paper briefly (3)
invited review **paper** (1)
Paper proposals (2)
paper title , abstract (1)
paper clip (1)
revised **paper** no (1)
Each position **paper** (1)
Length of the **paper** (1)

No complex inference (yet) !

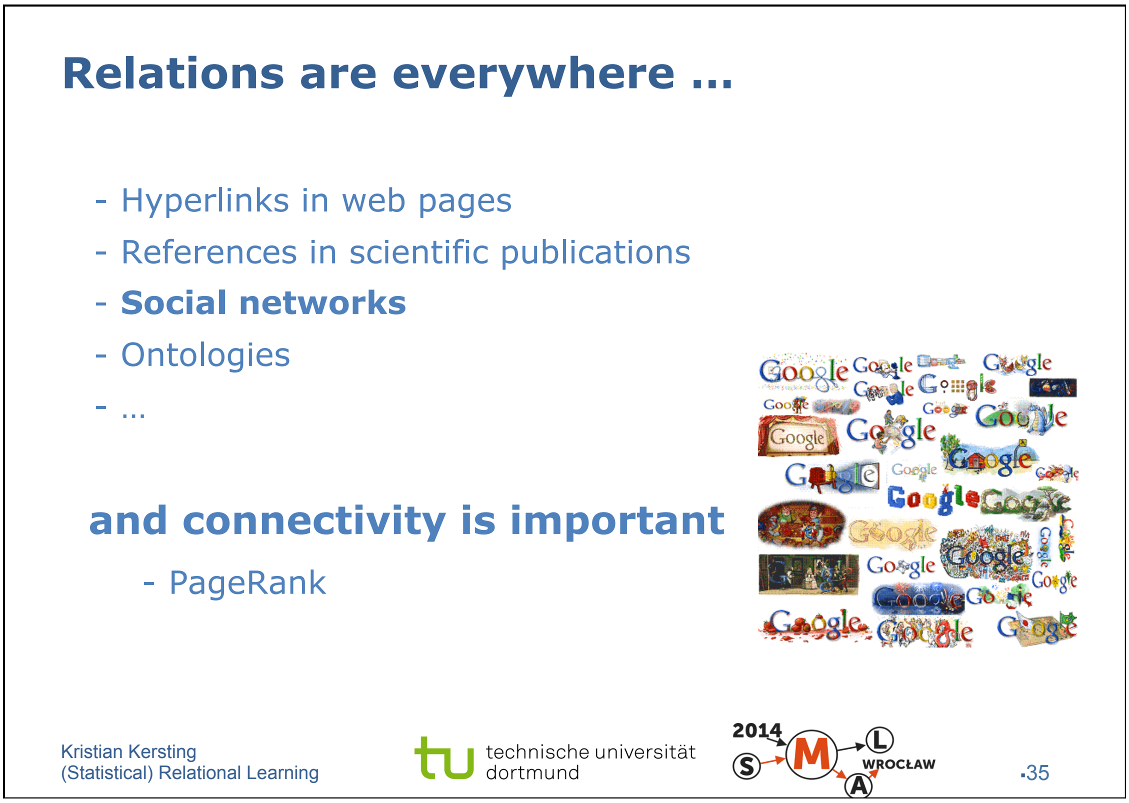
TextRunner: (Turing, born in, London)

+ **WordNet**: (London, part of, England)

+ **Rule**: 'born in' is transitive thru 'part of'

Conclusion: (Turing, born in, ***England***)

And again, 'instance' are not independent !

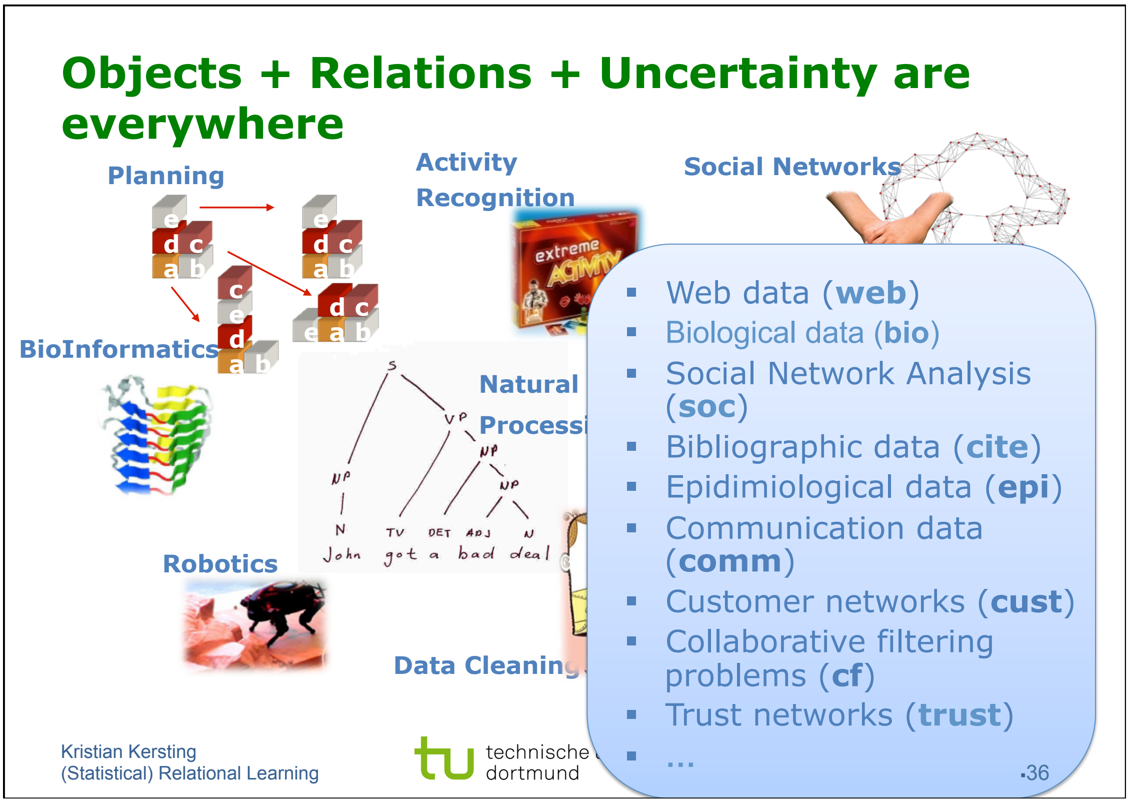


Relations are everywhere …

- Hyperlinks in web pages
- References in scientific publications
- **Social networks**
- Ontologies
- …

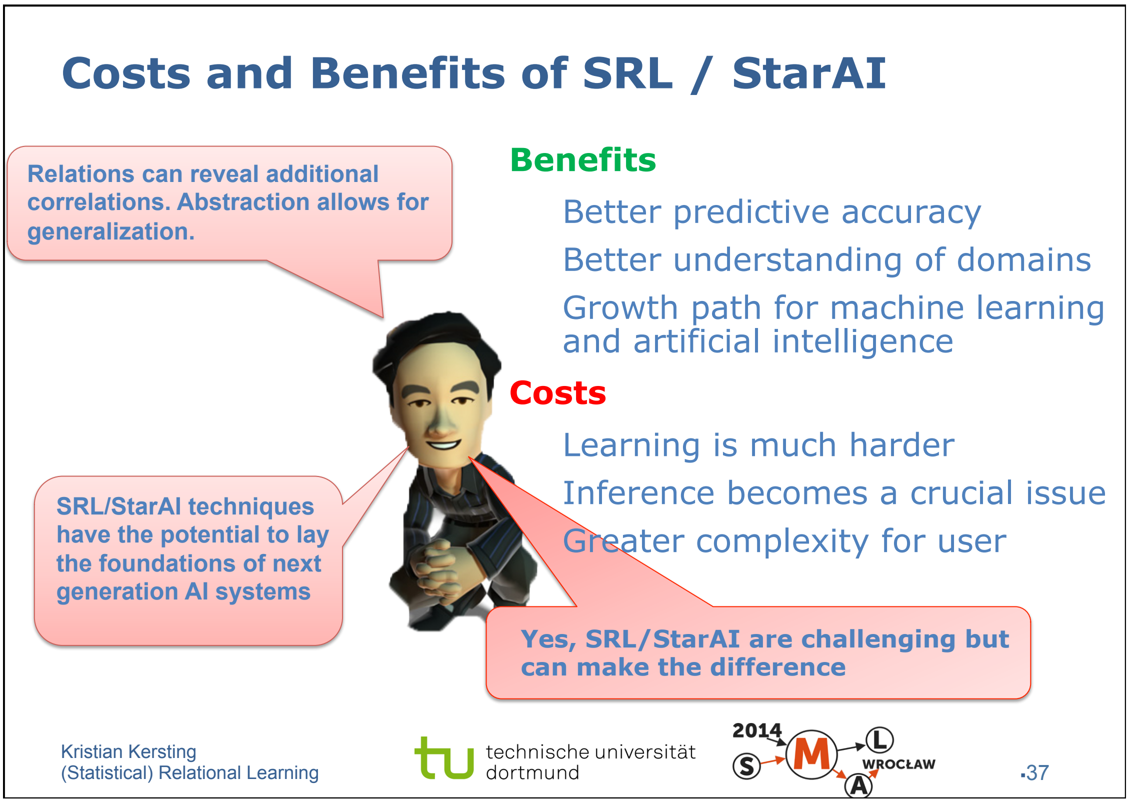
and connectivity is important

- PageRank



Objects + Relations + Uncertainty are everywhere

- Web data (**web**)
- Biological data (**bio**)
- Social Network Analysis (**soc**)
- Bibliographic data (**cite**)
- Epidimiological data (**epi**)
- Communication data (**comm**)
- Customer networks (**cust**)
- Collaborative filtering problems (**cf**)
- Trust networks (**trust**)
- …



Costs and Benefits of SRL / StarAI

Relations can reveal additional correlations. Abstraction allows for generalization.

SRL/StarAI techniques have the potential to lay the foundations of next generation AI systems

Benefits

Better predictive accuracy

Better understanding of domains

Growth path for machine learning and artificial intelligence

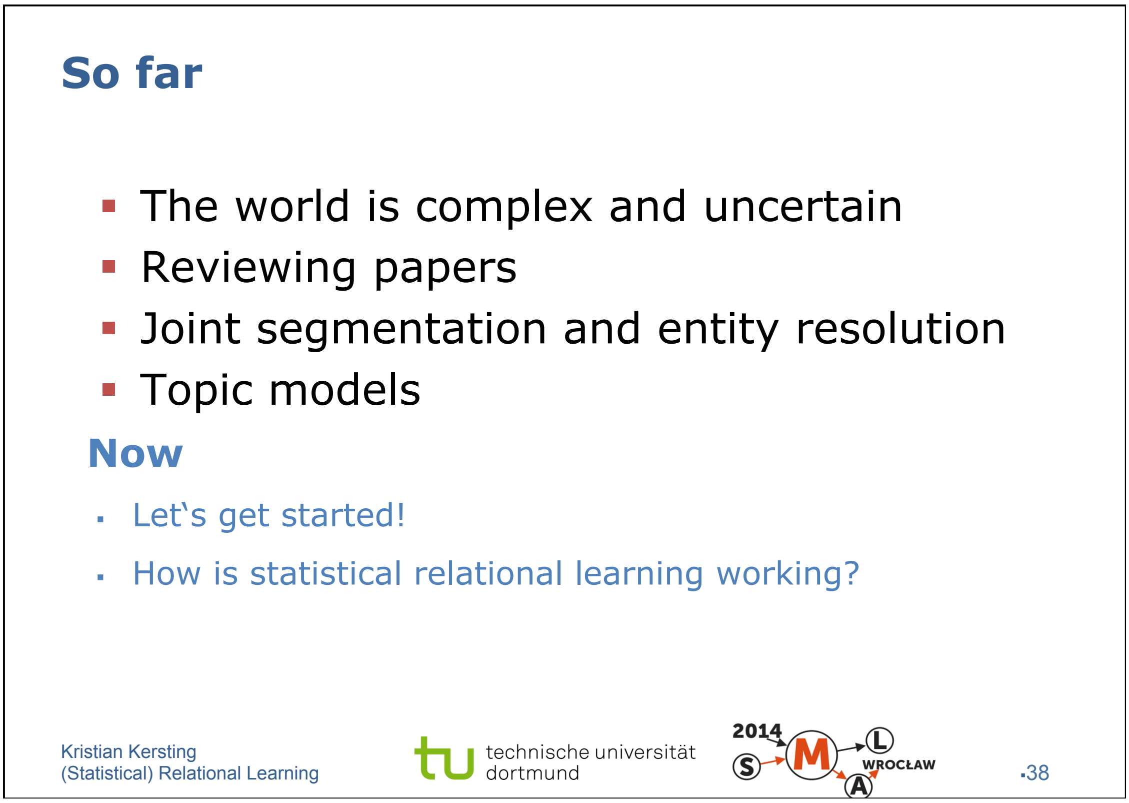
Costs

Learning is much harder

Inference becomes a crucial issue

Greater complexity for user

Yes, SRL/StarAI are challenging but can make the difference

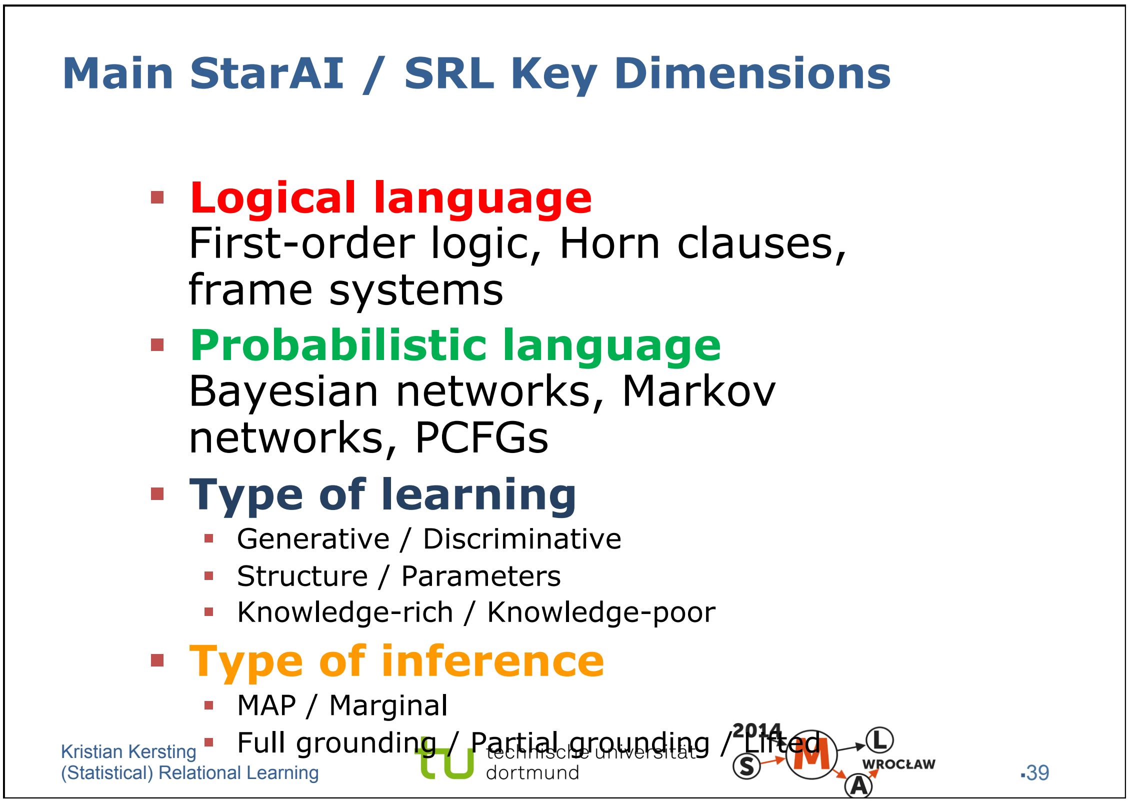


So far

- The world is complex and uncertain
- Reviewing papers
- Joint segmentation and entity resolution
- Topic models

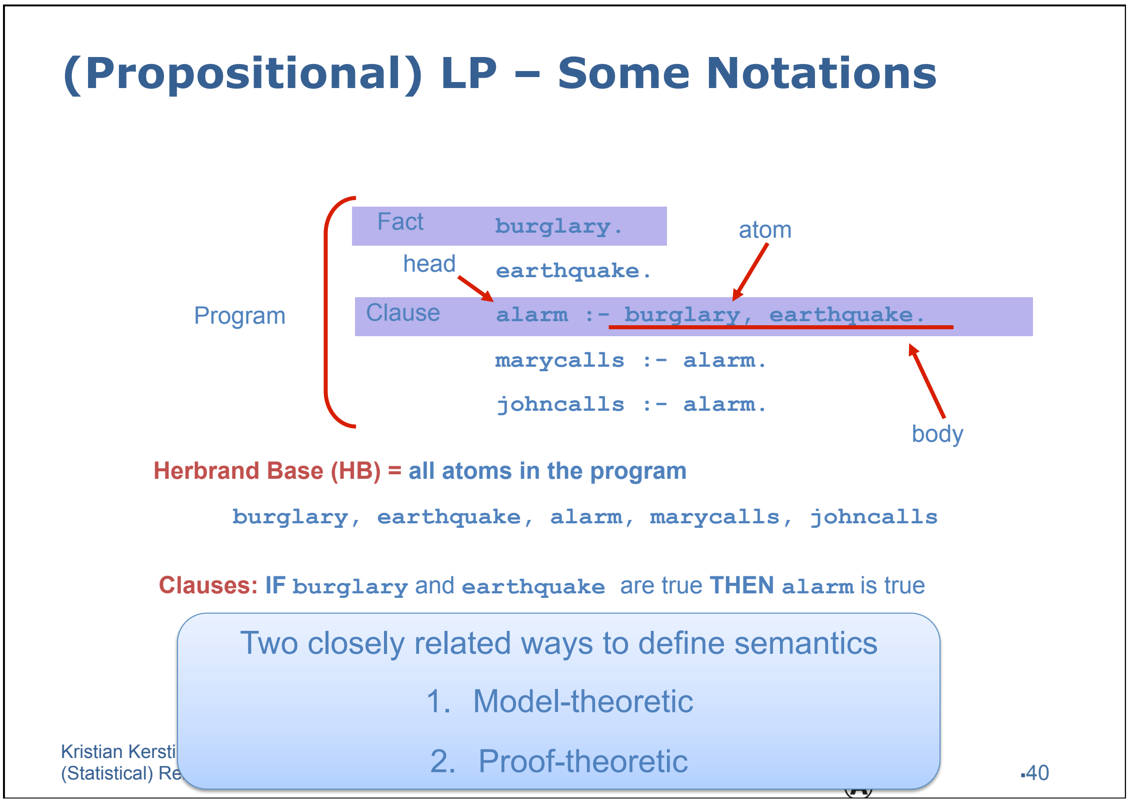
Now

- Let`s get started!
- How is statistical relational learning working?



Main StarAI / SRL Key Dimensions

- **Logical language**
  First-order logic, Horn clauses, frame systems
- **Probabilistic language**
  Bayesian networks, Markov networks, PCFGs
- **Type of learning**
  - Generative / Discriminative
  - Structure / Parameters
  - Knowledge-rich / Knowledge-poor
- **Type of inference**
  - MAP / Marginal
  - Full grounding / Partial grounding / Lifted



(Propositional) LP – Some Notations

Program
Fact
head
Clause
atom
body

```
burglary.
earthquake.
alarm :- burglary, earthquake.
marycalls :- alarm.
johncalls :- alarm.
```

Herbrand Base (HB) = all atoms in the program

`burglary, earthquake, alarm, marycalls, johncalls`

Clauses: **IF** `burglary` and `earthquake` are true **THEN** `alarm` is true

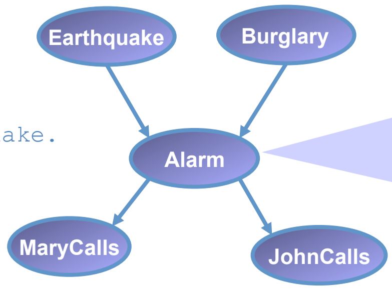
Two closely related ways to define semantics

1. Model-theoretic
2. Proof-theoretic

Model Theoretic: Restrictions on Possible Worlds

- Herbrand Interpretation
  - Truth assigments to all elements of HB
- An interpretation is a model of a clause C ⇔
  If the body of C holds then the head holds, too

```
burglary.
earthquake.
alarm :- burglary, earthquake.
marycalls :- alarm.
johncalls :- alarm.
```



E	B	P(A | B,E)	
e	b	0.9	0.1
e	$\overline{b}$	0.2	0.8
$\overline{e}$	b	0.9	0.1
$\overline{e}$	$\overline{b}$	0.01	0.99

Proof Theoretic: Restrictions on Possible Derivations

- A set of clauses can be used to prove that atoms are entailed by the set of clauses.

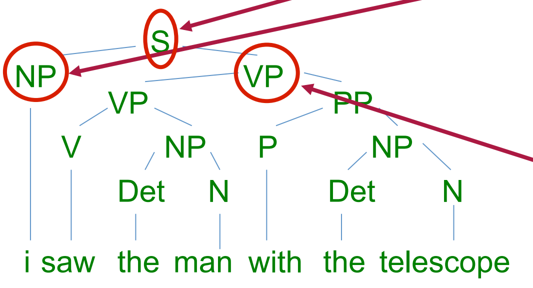
Goal

```
:- johncalls.
```

```
burglary.
earthquake.
earthquake.
alarm :- burglary, earthquake.
             :- burglary, earthquake.
marycalls :- alarm.
johncalls :- alarm.
                 :- alarm.
```

Stochastic Grammars

Upgrade HMMs (regular languages) to more complex languages such as context-free languages.



1.0 * 1/3 * 0.5 * 0.5 * 1.0 * ...
= 0.00231

Weighted Rewrite Rules

```
1.0 : S → NP, VP
1/3 : NP → i
1/3 : NP → Det, N
1/3 : NP → NP, PP
1.0 : Det → the
0.5 : N → man
0.5 : N → telescope
0.5 : VP → V, NP
0.5 : VP → VP, PP
1.0 : PP → P, NP
1.0 : V → saw
1.0 : P → with
```

Upgrading to First-Order Logic

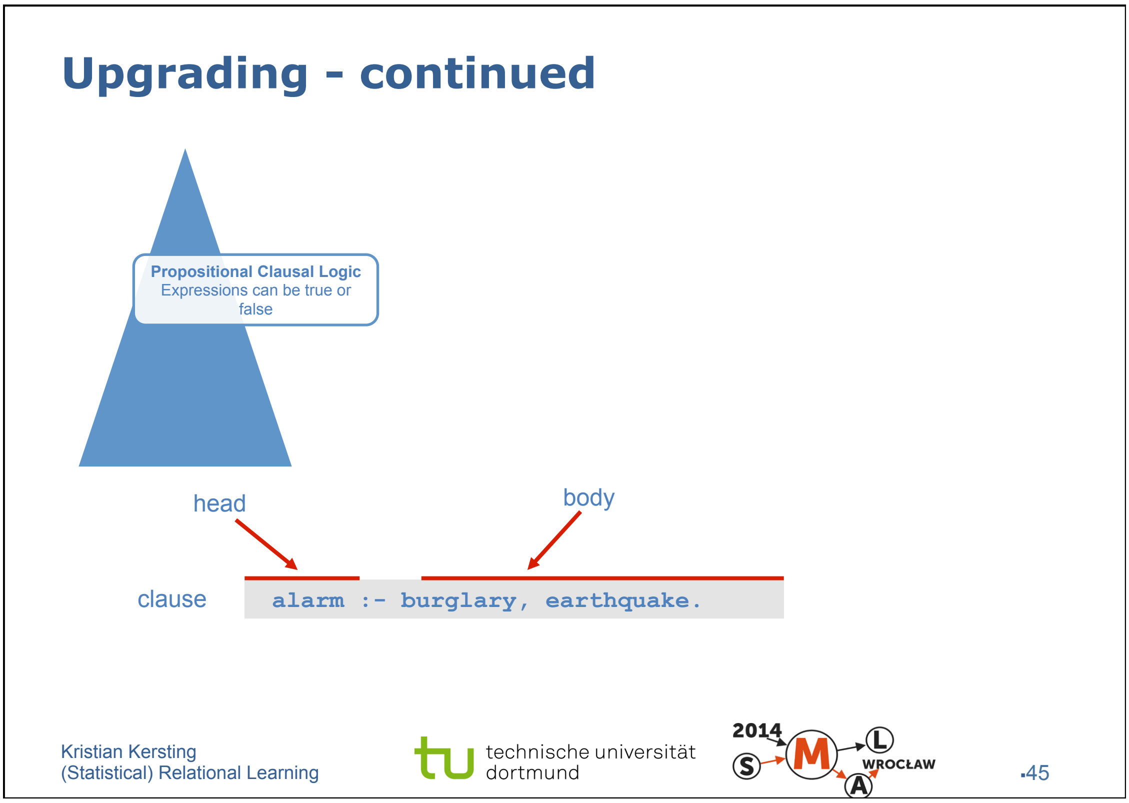
```
father(rex,fred).       mother(ann,fred).
father(brian,doro).     mother(utta, doro).
father(fred,henry).     mother(doro,henry).
pchrom(rex,a).  mchorm(rex,a).
pchrom(ann,a).  mchrom(ann,b).
...
```

The maternal information `mchrom/2` depends on the maternal and paternal `pchrom/2` information of the mother `mother/2`:

```
mchrom(fred,a). mchrom(fred,b),...
```

or better

```
mchrom(P,a) :- mother(M,P), pchrom(M,a), mchrom(M,a).
mchrom(P,a) :- mother(M,P), pchrom(M,a), mchrom(M,b).
mchrom(P,b) :- mother(M,P), pchrom(M,a), mchrom(M,b).
...
```



Upgrading - continued

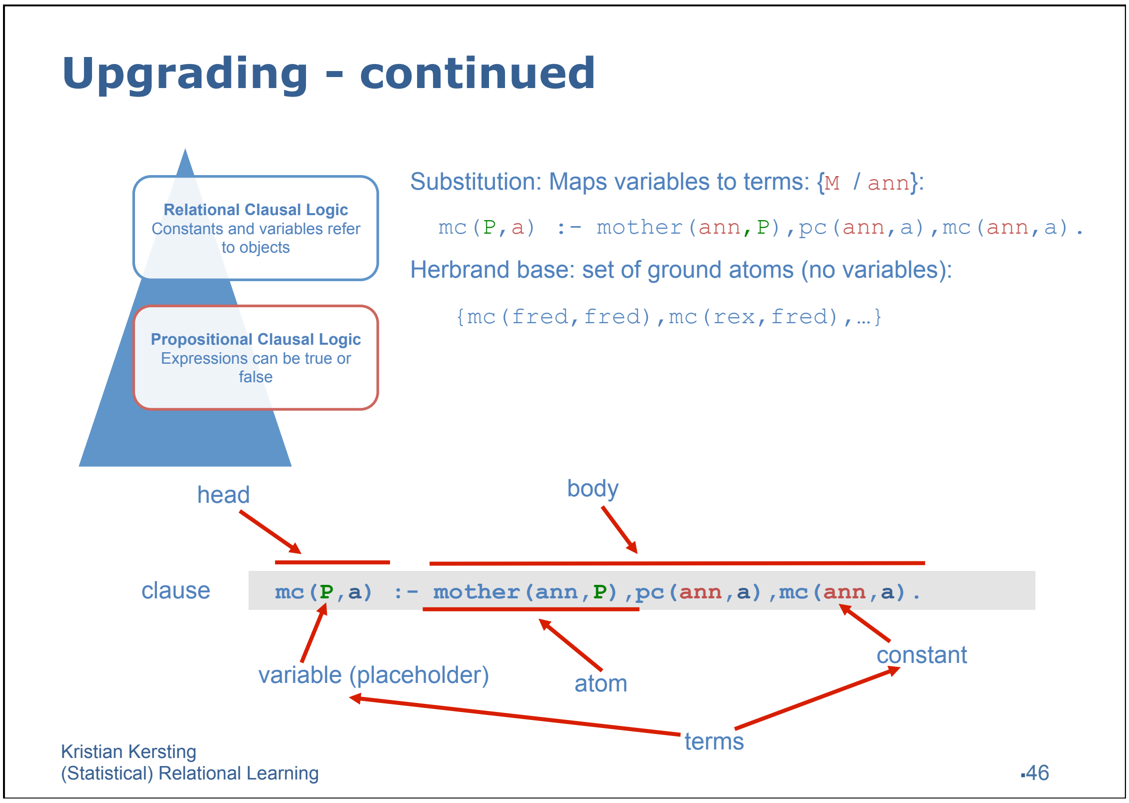
Propositional Clausal Logic
Expressions can be true or false

head

body

clause

```
alarm :- burglary, earthquake.
```



Upgrading - continued

Relational Clausal Logic
Constants and variables refer to objects

Propositional Clausal Logic
Expressions can be true or false

Substitution: Maps variables to terms: {M / ann}:

```
mc(P,a) :- mother(ann,P),pc(ann,a),mc(ann,a).
```

Herbrand base: set of ground atoms (no variables):

```
{mc(fred,fred),mc(rex,fred),…}
```

head

body

clause

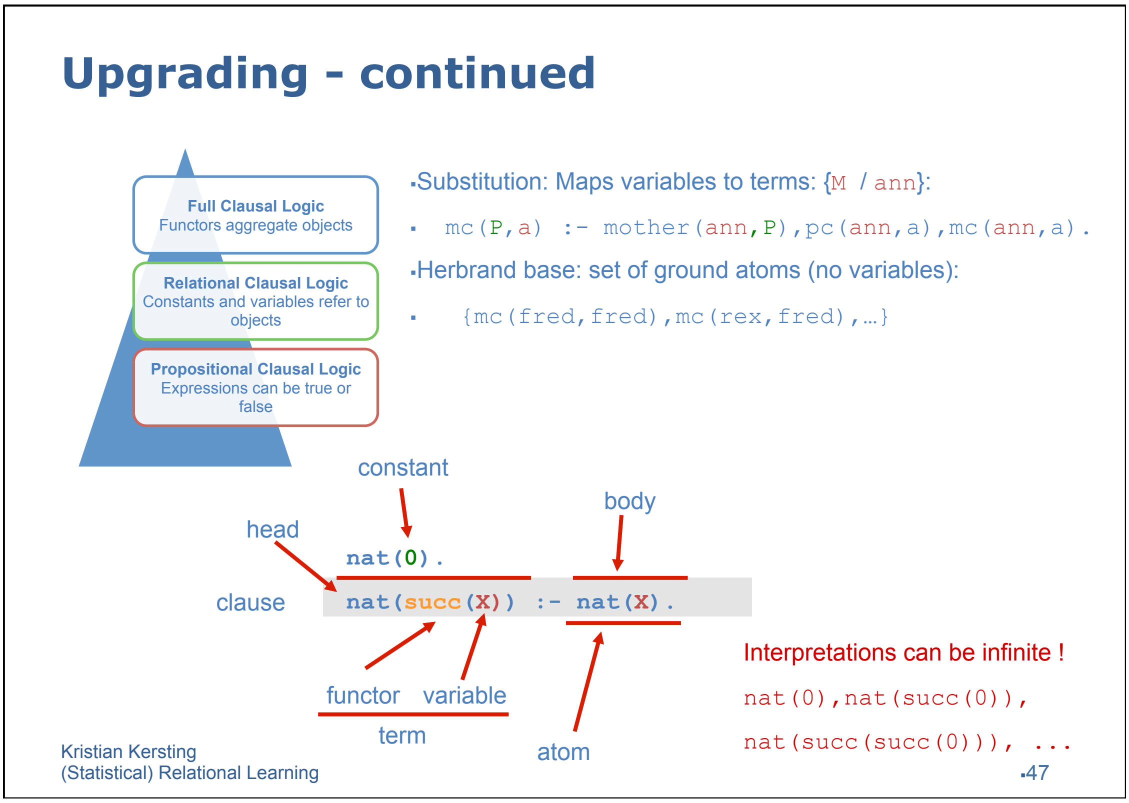
```
mc(P,a) :- mother(ann,P),pc(ann,a),mc(ann,a).
```

variable (placeholder)

atom

constant

terms



Upgrading - continued

Full Clausal Logic
Functors aggregate objects
Relational Clausal Logic
Constants and variables refer to objects
Propositional Clausal Logic
Expressions can be true or false

- Substitution: Maps variables to terms: {M / ann}:
  - `mc(P,a) :- mother(ann,P),pc(ann,a),mc(ann,a).`
- Herbrand base: set of ground atoms (no variables):
  - `{mc(fred,fred),mc(rex,fred),…}`

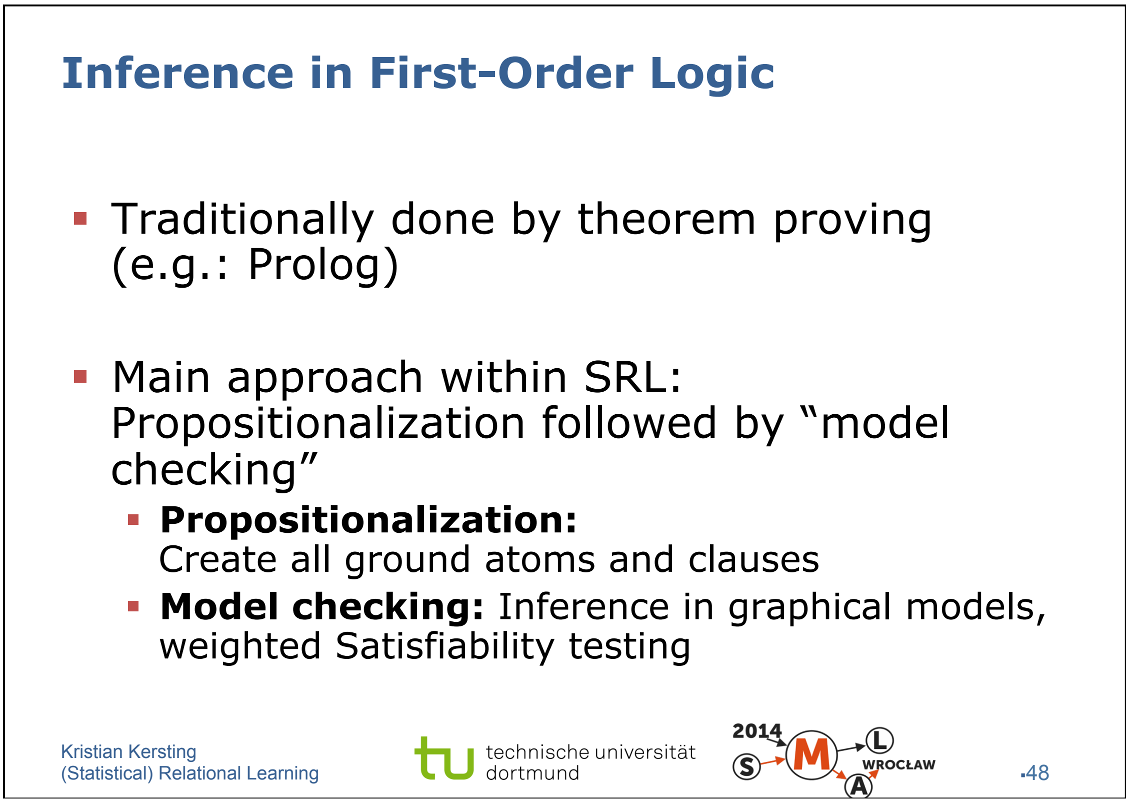
```
nat(0).
nat(succ(X)) :- nat(X).
```

constant
body
head
clause
functor
variable
term
atom

Interpretations can be infinite !

`nat(0),nat(succ(0)),`

`nat(succ(succ(0))), ...`



Inference in First-Order Logic

- Traditionally done by theorem proving (e.g.: Prolog)
- Main approach within SRL: Propositionalization followed by “model checking”
  - **Propositionalization:** Create all ground atoms and clauses
  - **Model checking:** Inference in graphical models, weighted Satisfiability testing

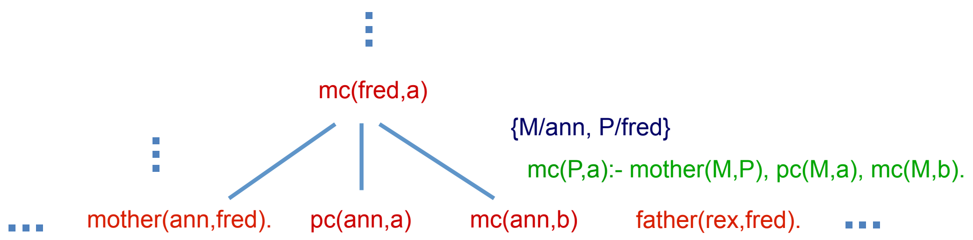
Forward Chaining

```
father(rex,fred).        mother(ann,fred).
father(brian,doro).      mother(utta, doro).
father(fred,henry).      mother(doro,henry).
pc(rex,a).  mc(rex,a).
pc(ann,a).  mc(ann,b).
...
```

mc(P,a) :- mother(M,P), pc(M,a), mc(M,a).

mc(P,a) :- mother(M,P), pc(M,a), mc(M,b).

Set of derivable ground atoms = least Herbrand model



mc(fred,a)

{M/ann, P/fred}

mc(P,a):- mother(M,P), pc(M,a), mc(M,b).

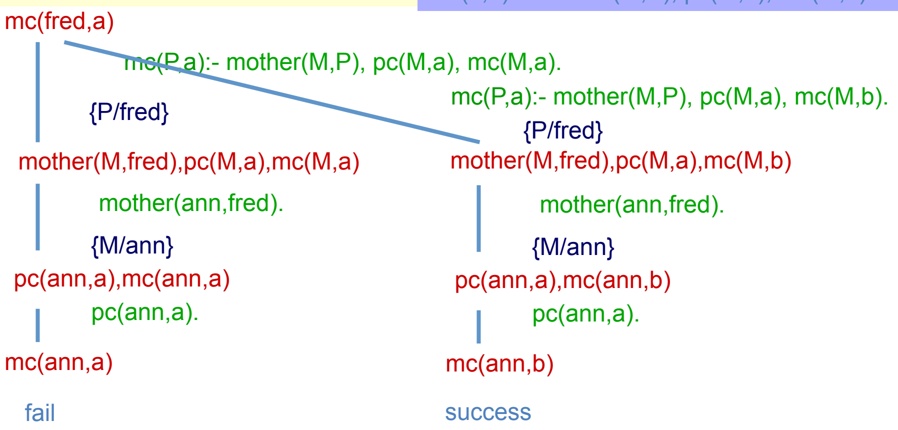
... mother(ann,fred). pc(ann,a) mc(ann,b) father(rex,fred). ...

Backward Chaining

```
father(rex,fred).        mother(ann,fred).
father(brian,doro).      mother(utta, doro).
father(fred,henry).      mother(doro,henry).
pc(rex,a).  mc(rex,a).
pc(ann,a).  mc(ann,b).
...
```

mc(P,a) :- mother(M,P), pc(M,a), mc(M,a).

mc(P,a) :- mother(M,P), pc(M,a), mc(M,b).



mc(fred,a)

mc(P,a):- mother(M,P), pc(M,a), mc(M,a).

{P/fred}

mother(M,fred),pc(M,a),mc(M,a)

mother(ann,fred).

{M/ann}

pc(ann,a),mc(ann,a)

pc(ann,a).

mc(ann,a)

fail

mc(P,a):- mother(M,P), pc(M,a), mc(M,b).

{P/fred}

mother(M,fred),pc(M,a),mc(M,b)

mother(ann,fred).

{M/ann}

pc(ann,a),mc(ann,b)

pc(ann,a).

mc(ann,b)

success

So far

- Motivation
- Brief review of logic

Now

- Let`s see some actual SRL frameworks

Alphabetic Soup of SRL

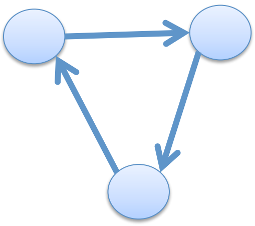
- Knowledge-based model construction [Wellman et al., 1992]
- PRISM [Sato & Kameya 1997]
- Stochastic logic programs [Muggleton, 1996]
- Probabilistic relational models [Friedman et al., 1999]
- Bayesian logic programs [Kersting & De Raedt, 2001]
- Bayesian logic [Milch et al., 2005]
- **Markov logic [Richardson & Domingos, 2006]**
- **Relational dependency networks [Neville & Jensen 2007]**
- ProbLog [De Raedt et al., 2007]

And many others!

[Neville & Jensen 2007]

Relational Dependency Networks

- **Logical language:** SQL queries
- **Probabilistic language:** Dependency networks
  - Conditional probability template for each predicate
  - Atoms depend on related atoms
  - >1 clause w/ head: aggregate functions
  - Cyclic dependencies



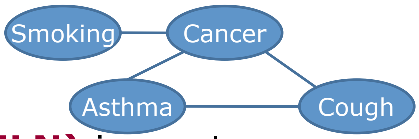
- **Learning:**
  - Parameters: EM based on Gibbs sampling
  - Structure: relational probability trees, boosting
- **Inference:** Gibbs sampling

[Richardson & Domingos, 2006]

Markov Logic

- **Logical language:** "First-order" logic
- **Probabilistic language:** Markov networks
  - **Syntax:** First-order formulas with weights
  - **Semantics:** Templates for Markov net features
- **Learning:**
  - **Parameters:** Generative or discriminative
  - **Structure:** ILP with arbitrary clauses and MAP score
- **Inference:**
  - **MAP:** Weighted satisfiability
  - **Marginal:** MCMC with moves proposed by SAT solver
  - Partial grounding + Lazy inference

Markov Logic



- A **Markov Logic Network (MLN)** is a set of pairs **(F, w)** where
  - **F** is a formula in first-order logic
  - **w** is a real number

$$P(X) = \frac{1}{Z} \exp\left(\sum_{i \in F} w_i n_i(x) \right)$$

# true groundings of *ith* clause

Normalization constant

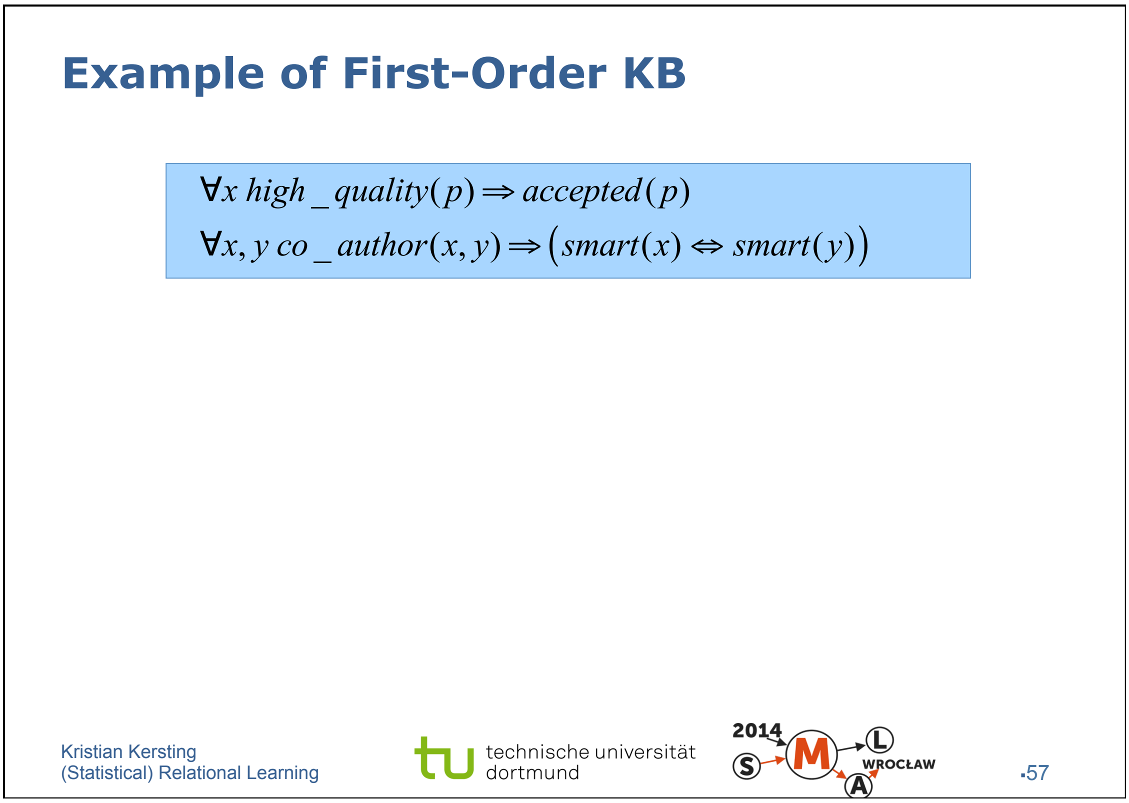
Iterate over all first-order MLN formulas

- Together with a finite set of constants, it defines a Markov network with
- Kind of undirected BLPs

Example of First-Order KB

High quality papers get accepted

Co-authors are either both smart or both not



Example of First-Order KB

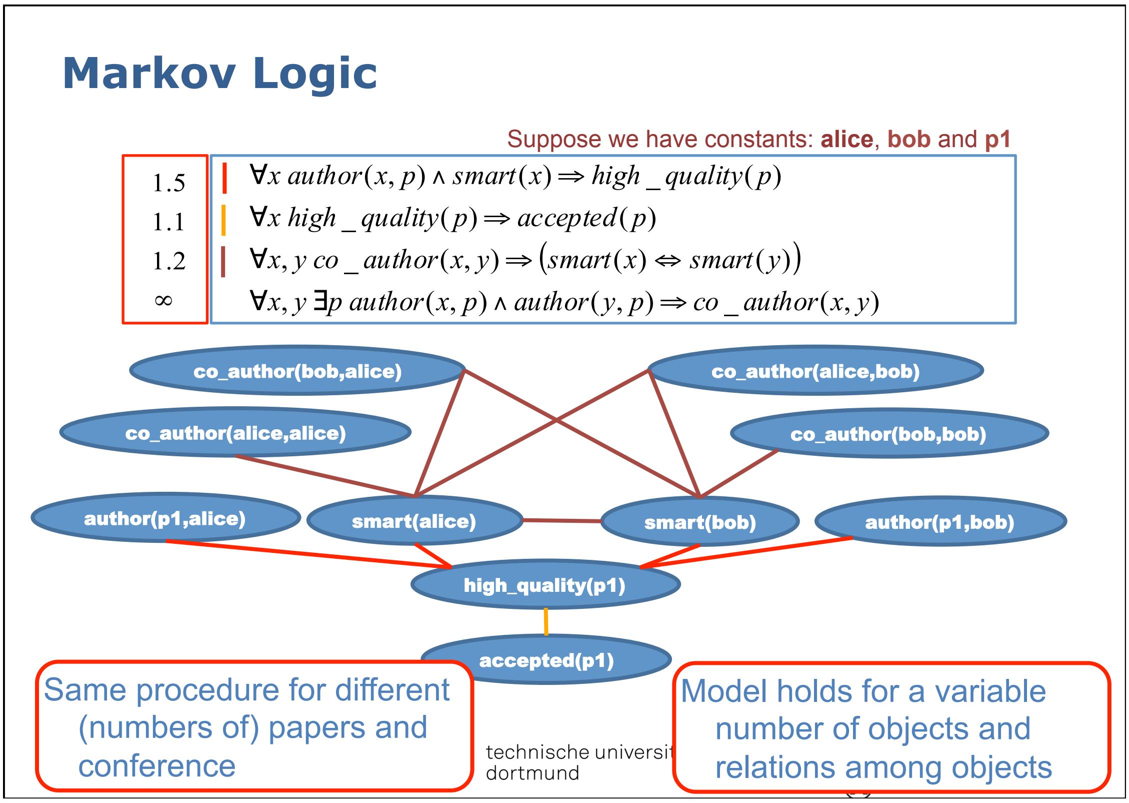
$$\forall x\ high\_quality(p) \Rightarrow accepted(p)$$
$$\forall x, y\ co\_author(x,y) \Rightarrow \big(smart(x) \Leftrightarrow smart(y)\big)$$

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014
S
M
L
A
WROCŁAW

▪57



Markov Logic

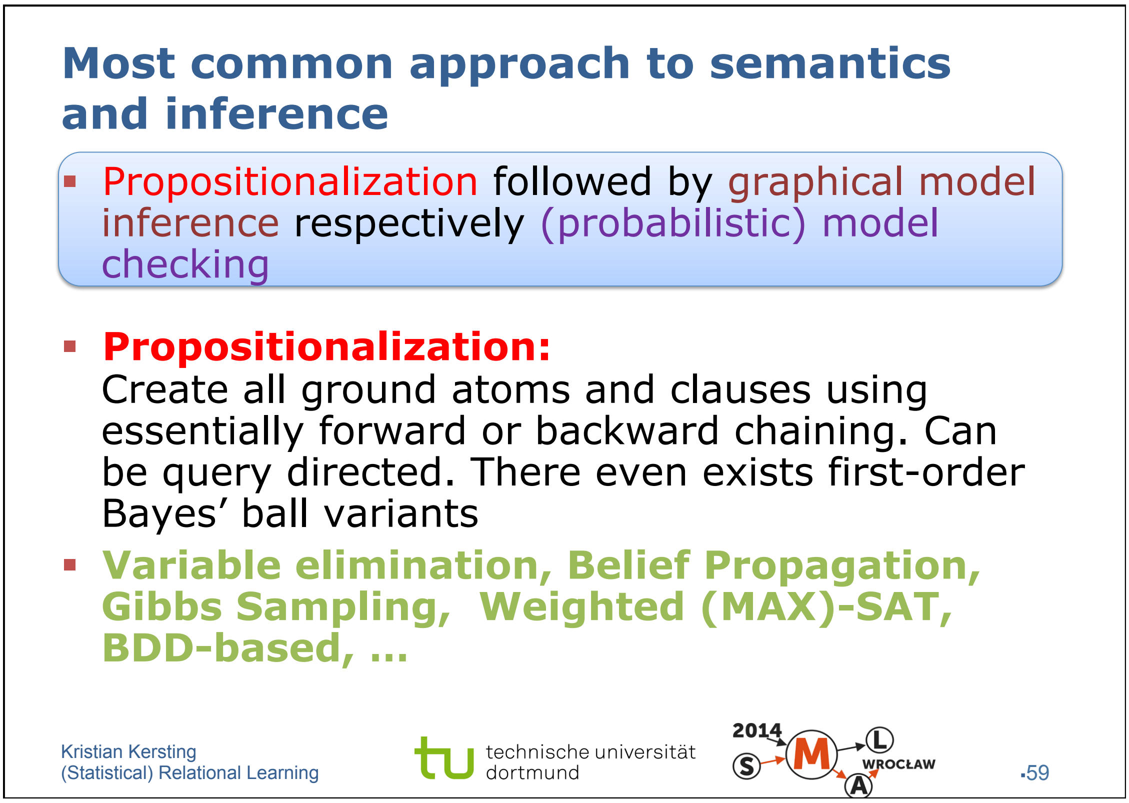
Suppose we have constants: **alice**, **bob** and **p1**

1.5	$\forall x\ author(x,p) \wedge smart(x) \Rightarrow high\_quality(p)$
1.1	$\forall x\ high\_quality(p) \Rightarrow accepted(p)$
1.2	$\forall x, y\ co\_author(x,y) \Rightarrow \big(smart(x) \Leftrightarrow smart(y)\big)$
∞	$\forall x, y\ \exists p\ author(x,p) \wedge author(y,p) \Rightarrow co\_author(x,y)$

Same procedure for different (numbers of) papers and conference

Model holds for a variable number of objects and relations among objects

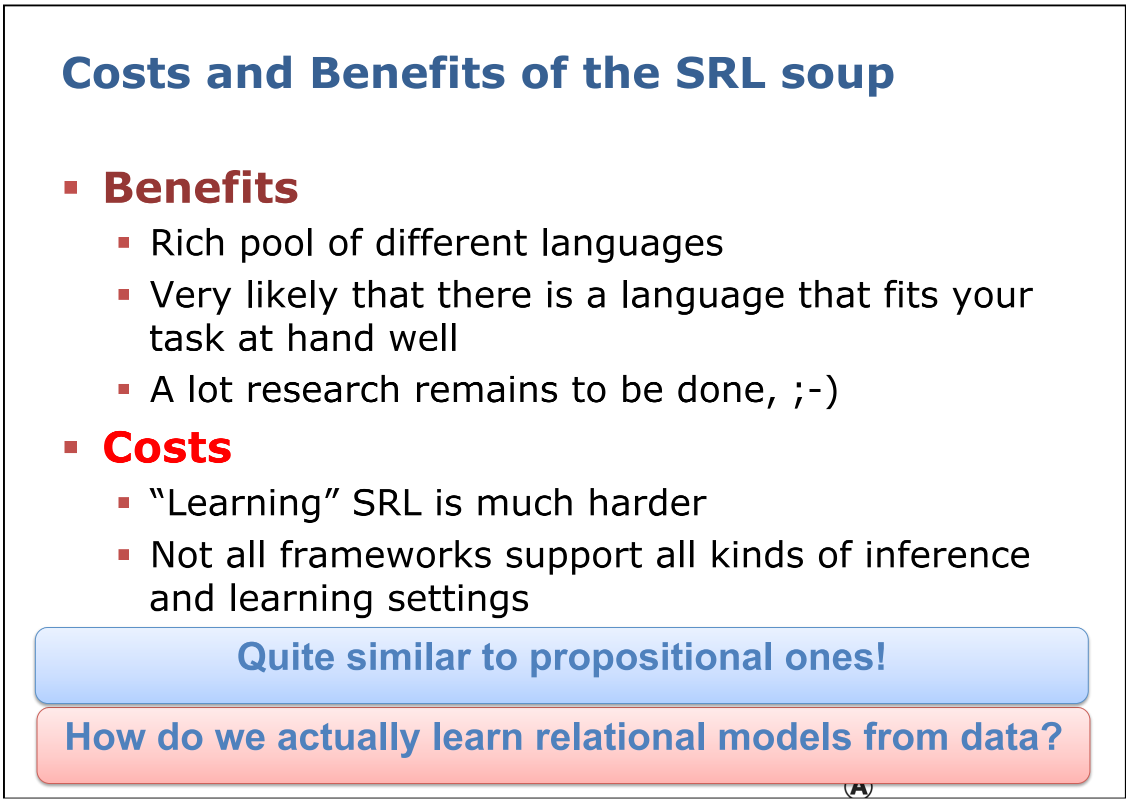
technische universit
dortmund



Most common approach to semantics and inference

- Propositionalization followed by graphical model inference respectively (probabilistic) model checking

- **Propositionalization:** Create all ground atoms and clauses using essentially forward or backward chaining. Can be query directed. There even exists first-order Bayes' ball variants
- **Variable elimination, Belief Propagation, Gibbs Sampling, Weighted (MAX)-SAT, BDD-based, …**

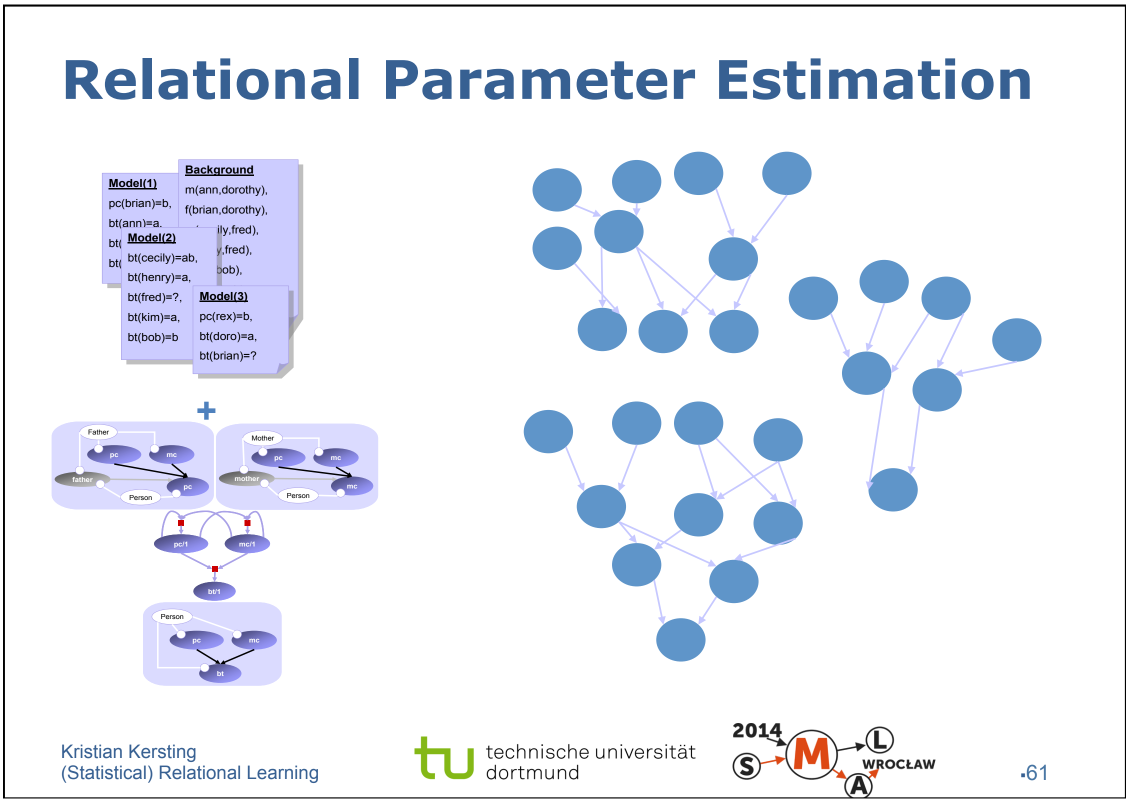


Costs and Benefits of the SRL soup

- **Benefits**
  - Rich pool of different languages
  - Very likely that there is a language that fits your task at hand well
  - A lot research remains to be done, ;-)
- **Costs**
  - "Learning" SRL is much harder
  - Not all frameworks support all kinds of inference and learning settings

Quite similar to propositional ones!

How do we actually learn relational models from data?

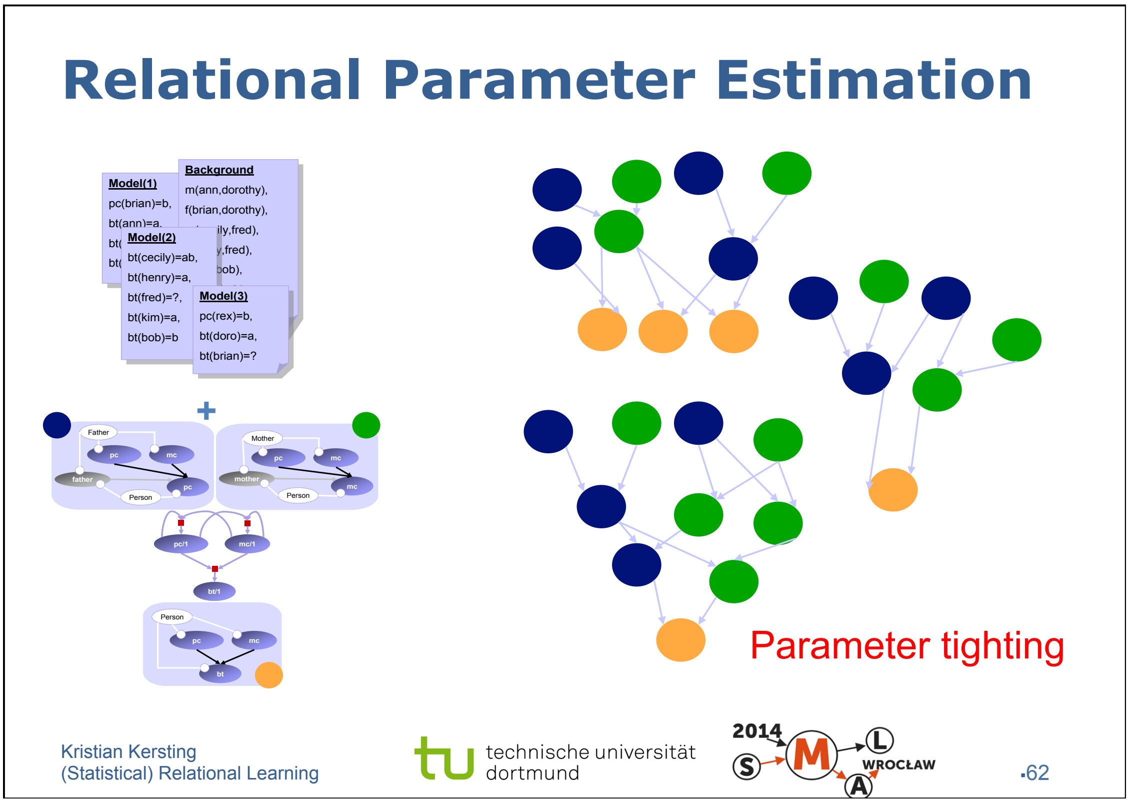


Relational Parameter Estimation

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 S M L A WROCŁAW



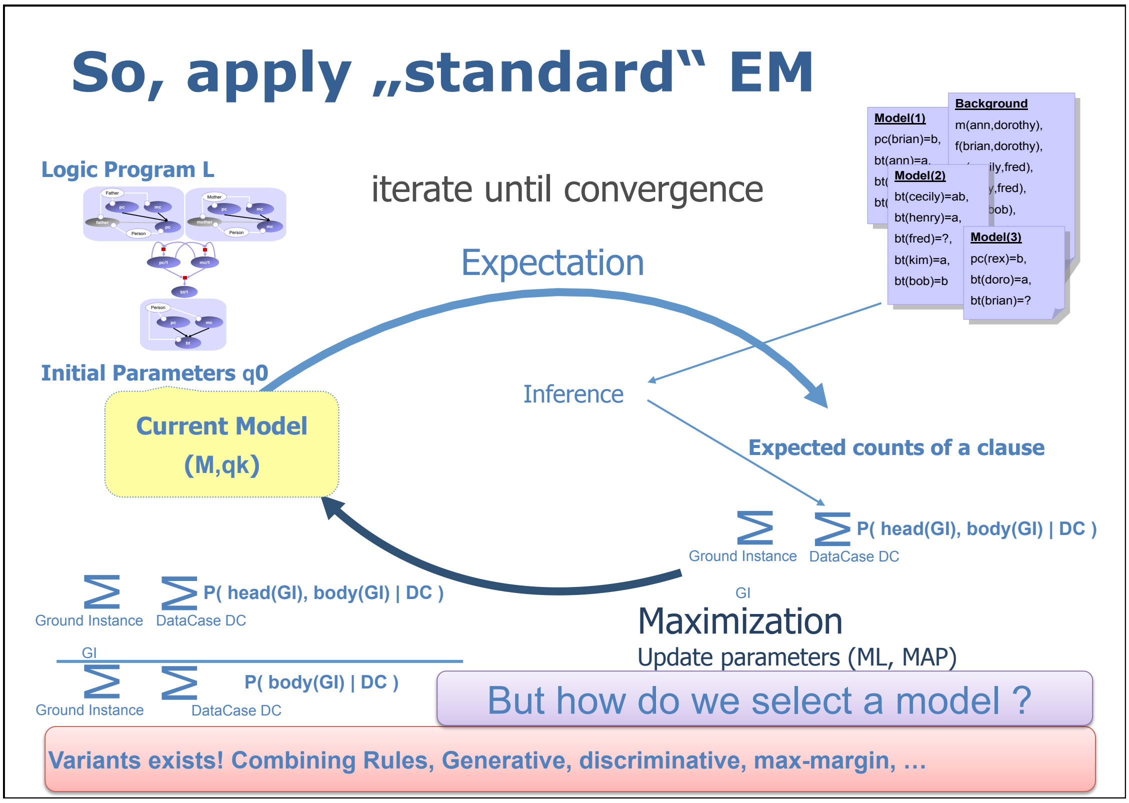
Relational Parameter Estimation

Parameter tighting

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 S M L A WROCŁAW



So, apply „standard" EM

Logic Program L

Initial Parameters q0

iterate until convergence

Expectation

Background
m(ann,dorothy),
f(brian,dorothy),
…ily,fred),
…y,fred),
…bob),

Model(1)
pc(brian)=b,
bt(ann)=a,
bt(…
bt(…

Model(2)
bt(cecily)=ab,
bt(henry)=a,
bt(fred)=?,
bt(kim)=a,
bt(bob)=b

Model(3)
pc(rex)=b,
bt(doro)=a,
bt(brian)=?

Current Model
(M,qk)

Inference

Expected counts of a clause

$$\sum_{\substack{\text{Ground Instance}\\ \text{GI}}} \sum_{\text{DataCase DC}} P(\text{head(GI), body(GI)} \mid DC)$$

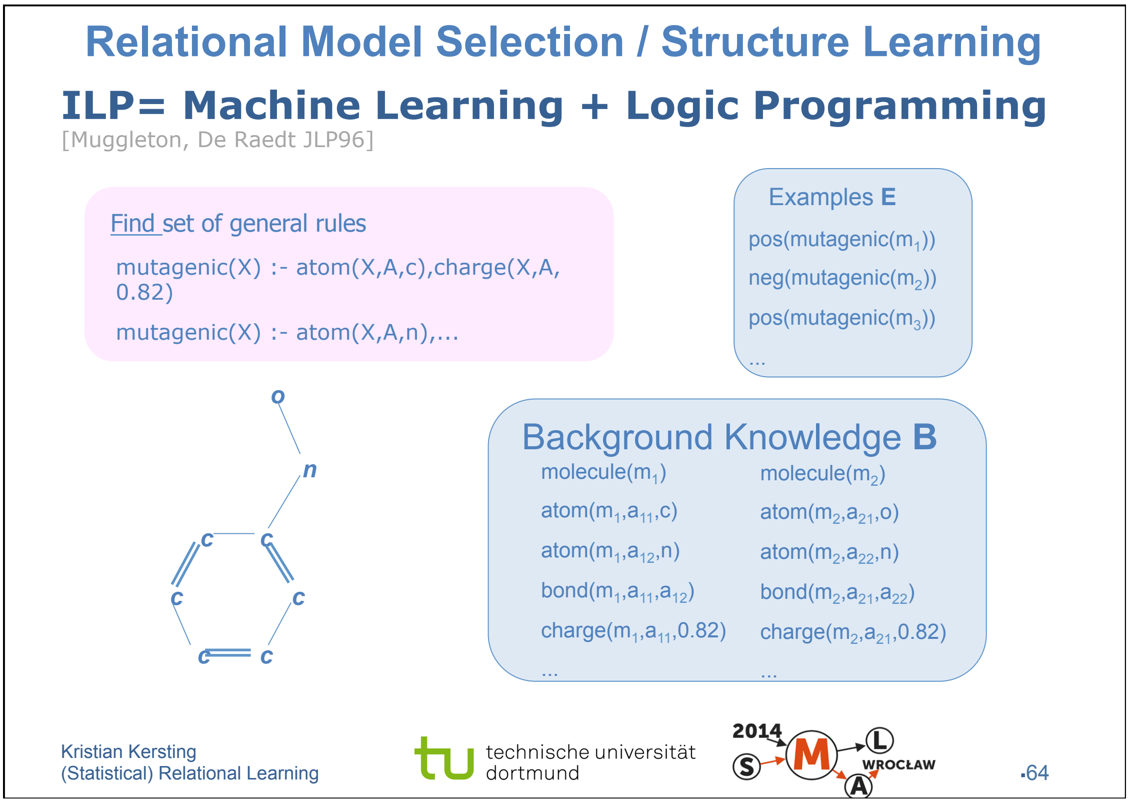
Maximization

Update parameters (ML, MAP)

$$\frac{\sum_{\substack{\text{Ground Instance}\\ \text{GI}}} \sum_{\text{DataCase DC}} P(\text{head(GI), body(GI)} \mid DC)}{\sum_{\text{Ground Instance}} \sum_{\text{DataCase DC}} P(\text{body(GI)} \mid DC)}$$

But how do we select a model ?

Variants exists! Combining Rules, Generative, discriminative, max-margin, …



Relational Model Selection / Structure Learning

ILP= Machine Learning + Logic Programming

[Muggleton, De Raedt JLP96]

Find set of general rules

mutagenic(X) :- atom(X,A,c),charge(X,A, 0.82)

mutagenic(X) :- atom(X,A,n),...

Examples **E**

pos(mutagenic(m_1))

neg(mutagenic(m_2))

pos(mutagenic(m_3))

...

Background Knowledge **B**

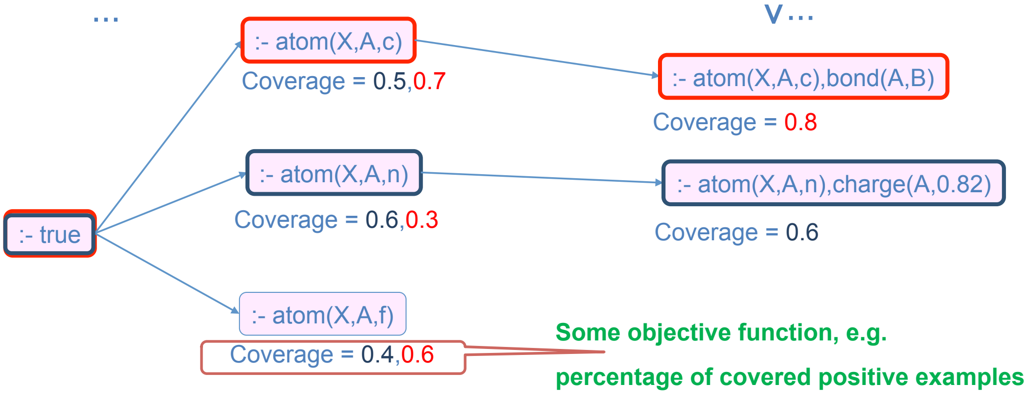
molecule(m_1)	molecule(m_2)
atom(m_1,a_{11},c)	atom(m_2,a_{21},o)
atom(m_1,a_{12},n)	atom(m_2,a_{22},n)
bond(m_1,a_{11},a_{12})	bond(m_2,a_{21},a_{22})
charge(m_1,a_{11},0.82)	charge(m_2,a_{21},0.82)
...	...

Example ILP Algorithm: FOIL [Quinlan MLJ 5:239-266, 1990]

mutagenic(X) :- atom(X,A,n),charge(A,0.82) 0

mutagenic(X) :- atom(X,A,c),bond(A,B) v 1 ≡ 1

… v…



Some objective function, e.g. percentage of covered positive examples

Vanilla SRL [De Raedt, Kersting ALT04]

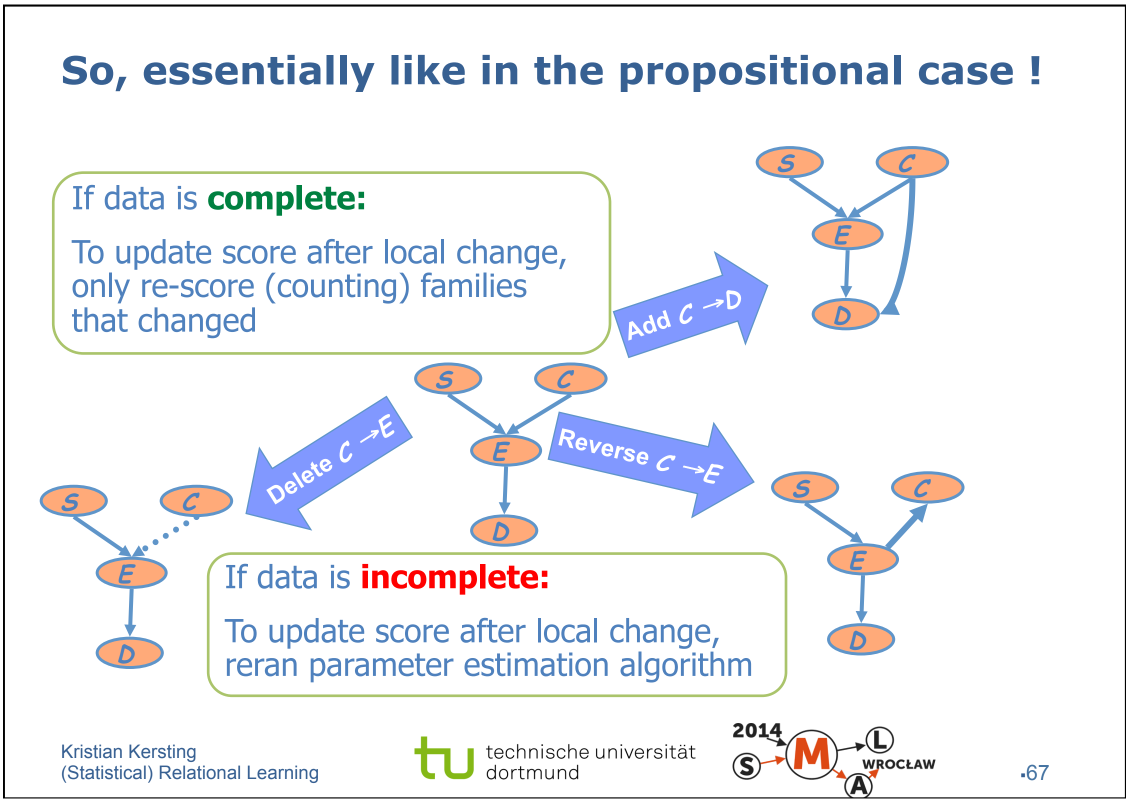
mutagenic(X) :- atom(X,A,n),charge(A,0.82)

mutagenic(X) :- atom(X,A,c),bond(A,B) =0.882

…

- Traverses the hypotheses space a la ILP
- Replaces ILP’s 0-1 covers relation by a “smooth”, probabilistic one [0,1]

$$\begin{aligned}\mathrm{cover}(e,H,B) &= P(e|H,B)\\ \mathrm{cover}(E,H,B) &= \prod\nolimits_{e\in E}\mathrm{cover}(e,H,B)\end{aligned}$$



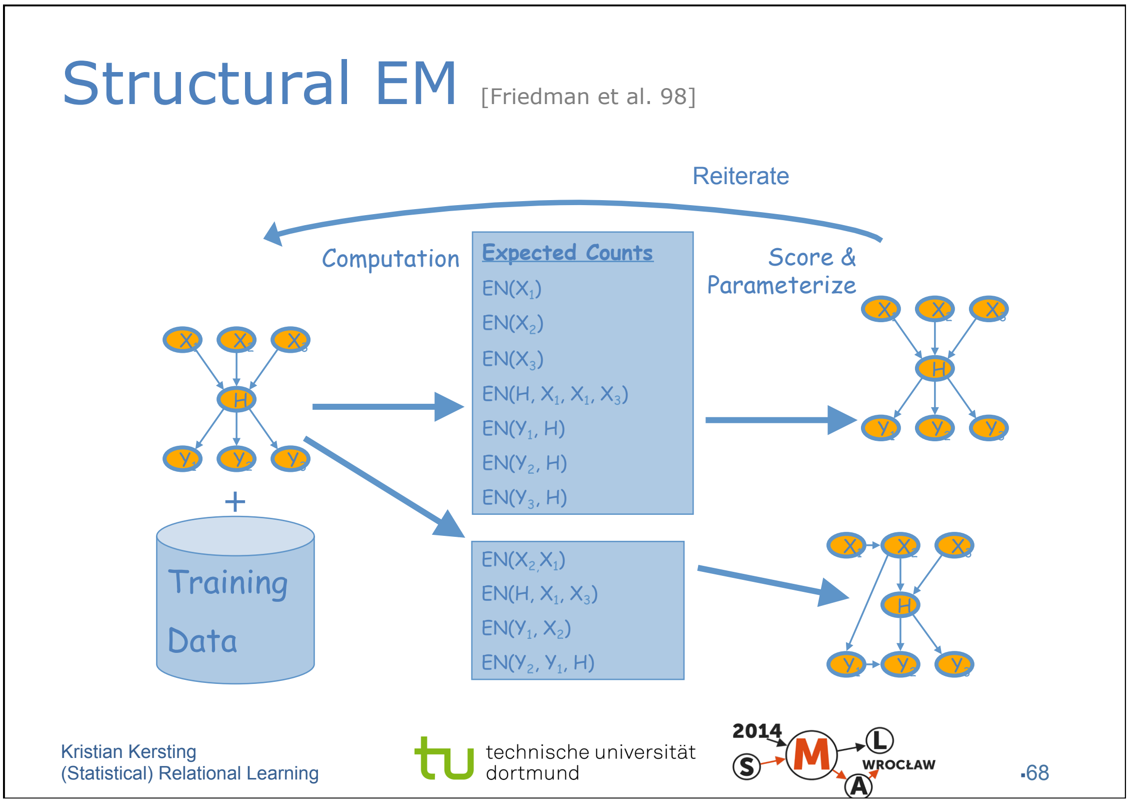
So, essentially like in the propositional case !

If data is **complete:**

To update score after local change, only re-score (counting) families that changed

If data is **incomplete:**

To update score after local change, reran parameter estimation algorithm



Structural EM [Friedman et al. 98]

Expected Counts

$EN(X_1)$

$EN(X_2)$

$EN(X_3)$

$EN(H, X_1, X_1, X_3)$

$EN(Y_1, H)$

$EN(Y_2, H)$

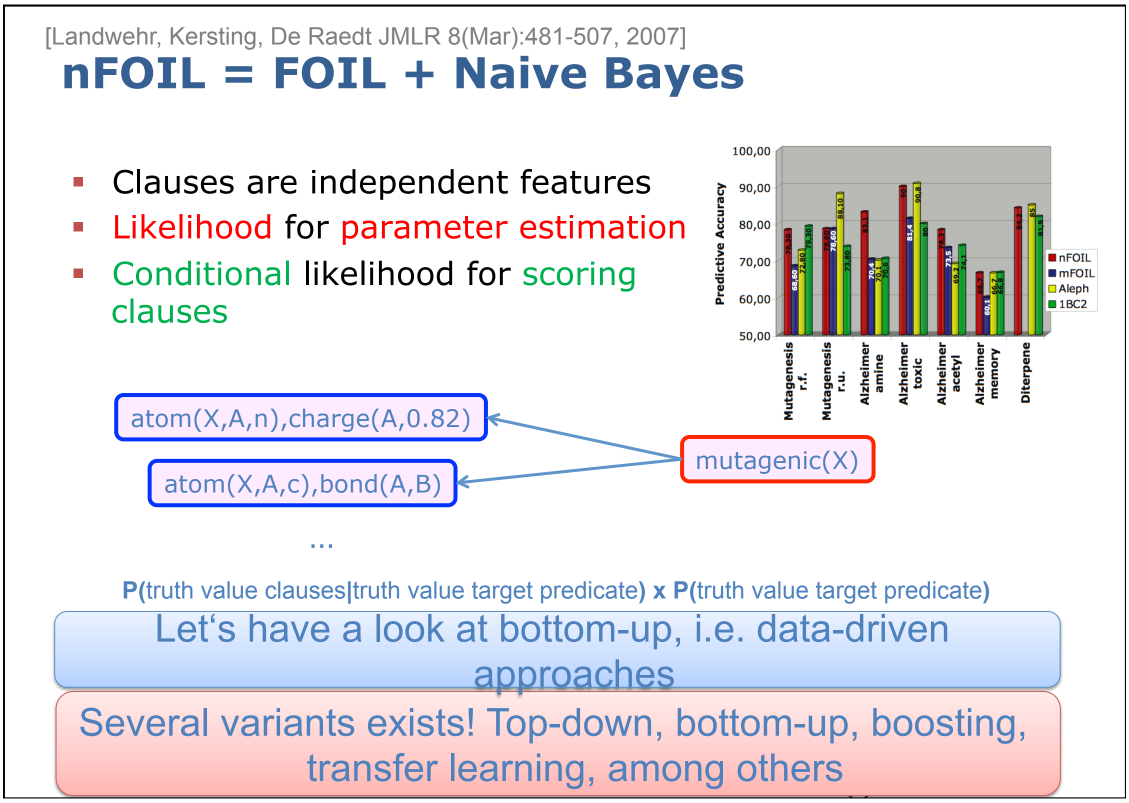
$EN(Y_3, H)$

$EN(X_2, X_1)$

$EN(H, X_1, X_3)$

$EN(Y_1, X_2)$

$EN(Y_2, Y_1, H)$



[Landwehr, Kersting, De Raedt JMLR 8(Mar):481-507, 2007]

nFOIL = FOIL + Naive Bayes

- Clauses are independent features
- Likelihood for parameter estimation
- Conditional likelihood for scoring clauses

Predictive Accuracy
100,00
90,00
80,00
70,00
60,00
50,00
Mutagenesis r.f.
Mutagenesis r.u.
Alzheimer amine
Alzheimer toxic
Alzheimer acetyl
Alzheimer memory
Diterpene
78,30
68,60
72,80
79,30
78,60
78,60
88,10
73,80
83,1
70,4
70,1
70,6
90
81,4
90,8
80
78,3
73,5
69,2
74,1
66,7
60,1
66,7
66,8
84,2
85
81,9
🟥 nFOIL
🟦 mFOIL
🟨 Aleph
🟩 1BC2

atom(X,A,n),charge(A,0.82)

atom(X,A,c),bond(A,B)

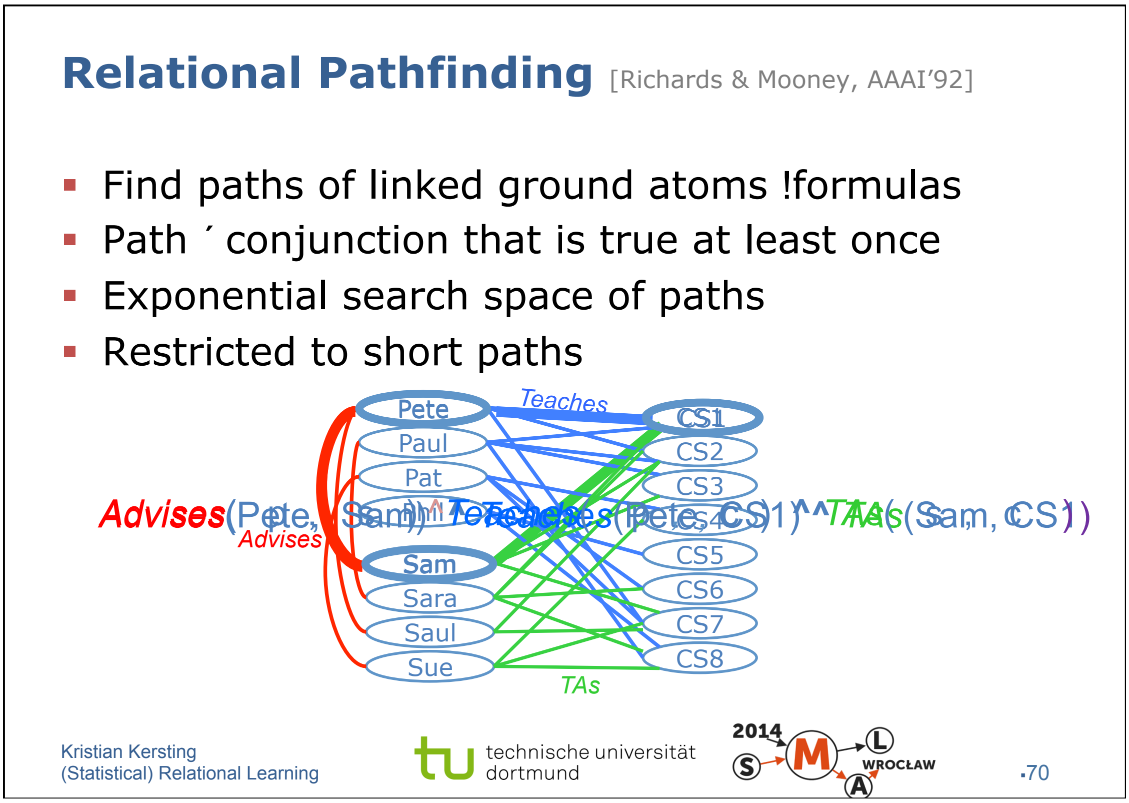
...

mutagenic(X)

P(truth value clauses|truth value target predicate**) x P(**truth value target predicate**)**

Let's have a look at bottom-up, i.e. data-driven approaches

Several variants exists! Top-down, bottom-up, boosting, transfer learning, among others



Relational Pathfinding [Richards & Mooney, AAAI'92]

- Find paths of linked ground atoms !formulas
- Path ´ conjunction that is true at least once
- Exponential search space of paths
- Restricted to short paths

Teaches
Pete
Paul
Pat
Sam
Sara
Saul
Sue
CS1
CS2
CS3
CS5
CS6
CS7
CS8
Advises
TAs

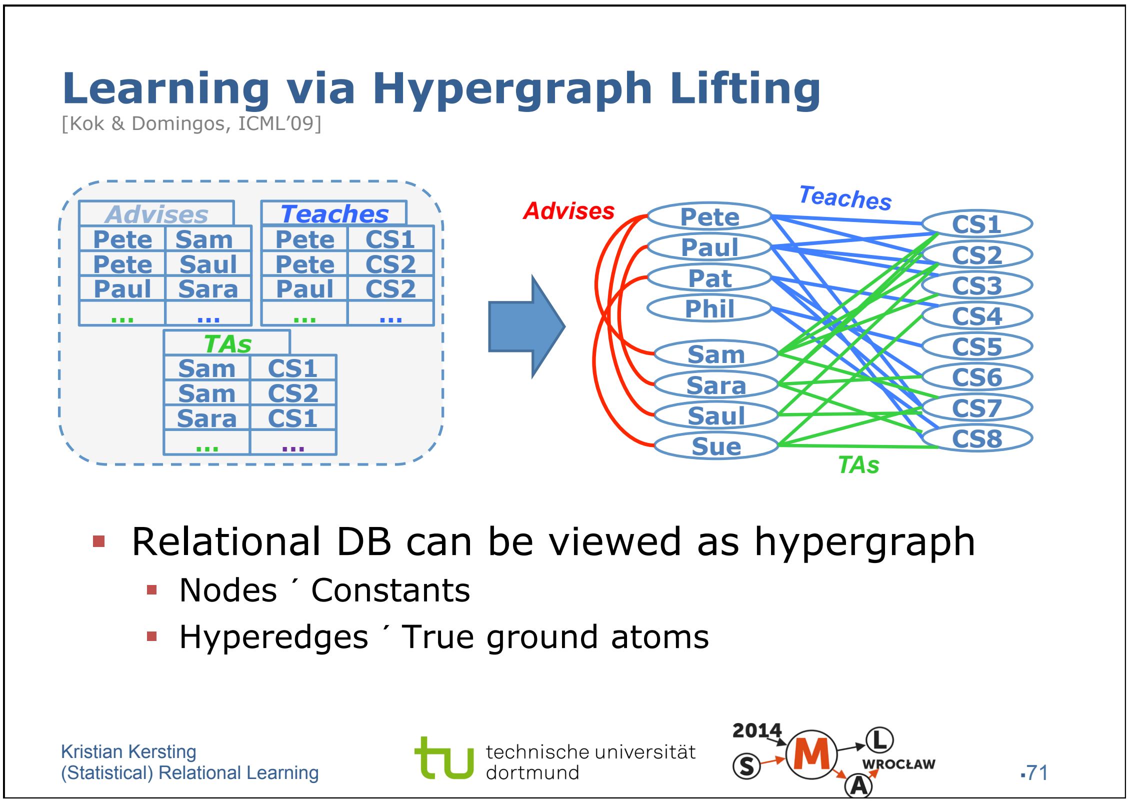
Advises(Pete, Sam) ^ Teaches(Pete, CS1) ^ TAs(Sam, CS1)

Kristian Kersting
(Statistical) Relational Learning

technische universität dortmund

2014 S M L A WROCŁAW

▪70



Learning via Hypergraph Lifting

[Kok & Domingos, ICML'09]

Advises	
Pete	Sam
Pete	Saul
Paul	Sara
…	…

Teaches	
Pete	CS1
Pete	CS2
Paul	CS2
…	…

TAs	
Sam	CS1
Sam	CS2
Sara	CS1
…	…

Advises
Teaches
Pete
Paul
Pat
Phil
Sam
Sara
Saul
Sue
CS1
CS2
CS3
CS4
CS5
CS6
CS7
CS8
TAs

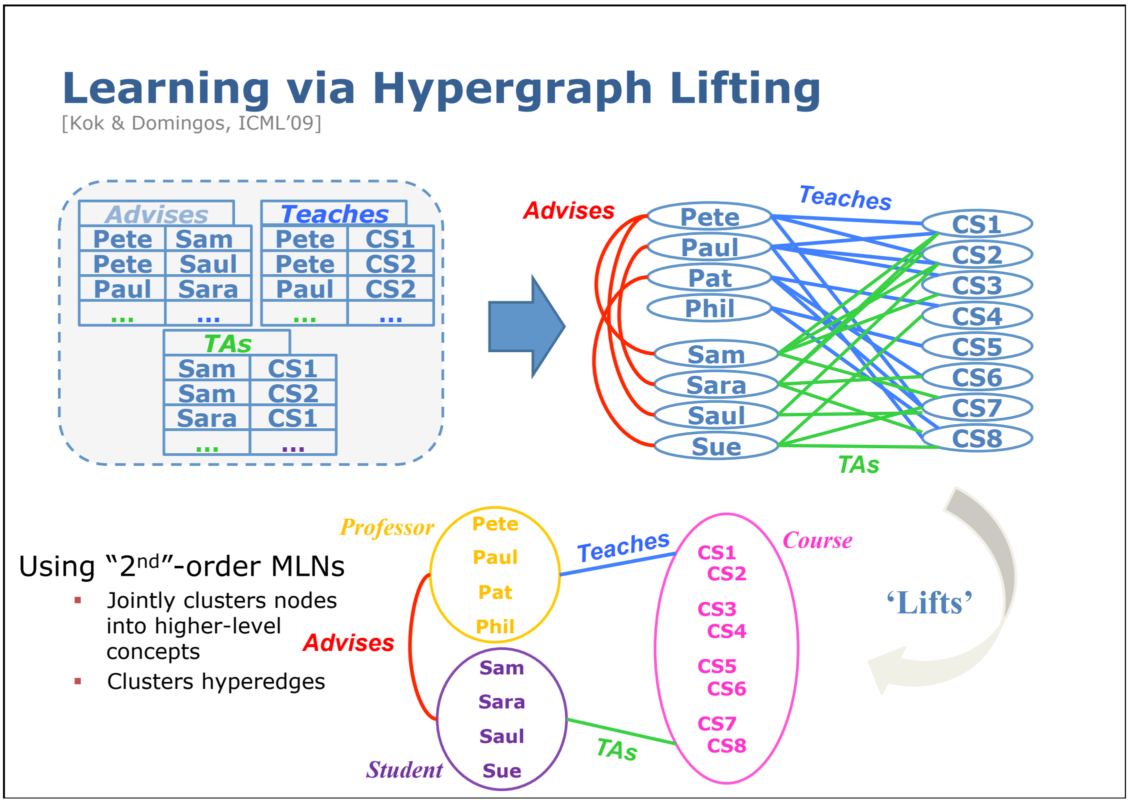
- Relational DB can be viewed as hypergraph
  - Nodes ´ Constants
  - Hyperedges ´ True ground atoms

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014
S
M
L
A
WROCŁAW

▪71



Learning via Hypergraph Lifting

[Kok & Domingos, ICML'09]

Advises	
Pete	Sam
Pete	Saul
Paul	Sara
…	…

Teaches	
Pete	CS1
Pete	CS2
Paul	CS2
…	…

TAs	
Sam	CS1
Sam	CS2
Sara	CS1
…	…

Advises
Teaches
Pete
Paul
Pat
Phil
Sam
Sara
Saul
Sue
CS1
CS2
CS3
CS4
CS5
CS6
CS7
CS8
TAs
Professor
Pete
Paul
Pat
Phil
Teaches
CS1
CS2
CS3
CS4
CS5
CS6
CS7
CS8
Course
'Lifts'
Advises
Sam
Sara
Saul
Sue
Student
TAs

Using "2nd"-order MLNs

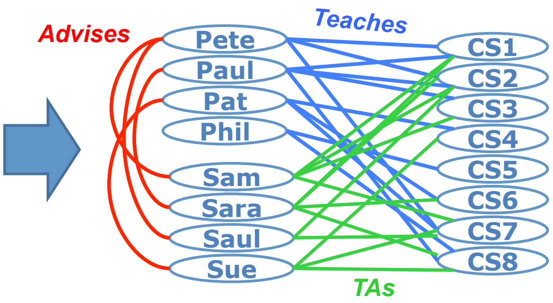
- Jointly clusters nodes into higher-level concepts
- Clusters hyperedges

Learning via Hypergraph Lifting[Kok & Domingos, ICML’09]

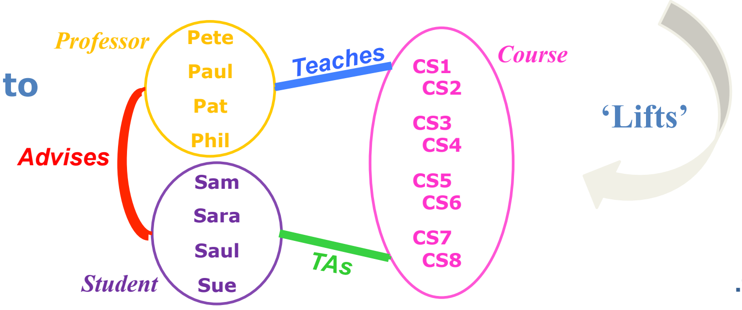
Advises	
Pete	Sam
Pete	Saul
Paul	Sara
…	…

Teaches	
Pete	CS1
Pete	CS2
Paul	CS2
…	…

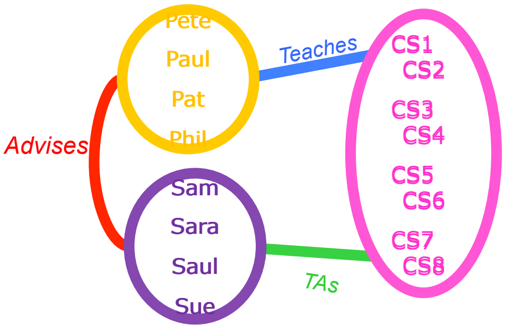
TAs	
Sam	CS1
Sam	CS2
Sara	CS1
…	…



Trace paths & convert paths to first-order clauses



FindPaths



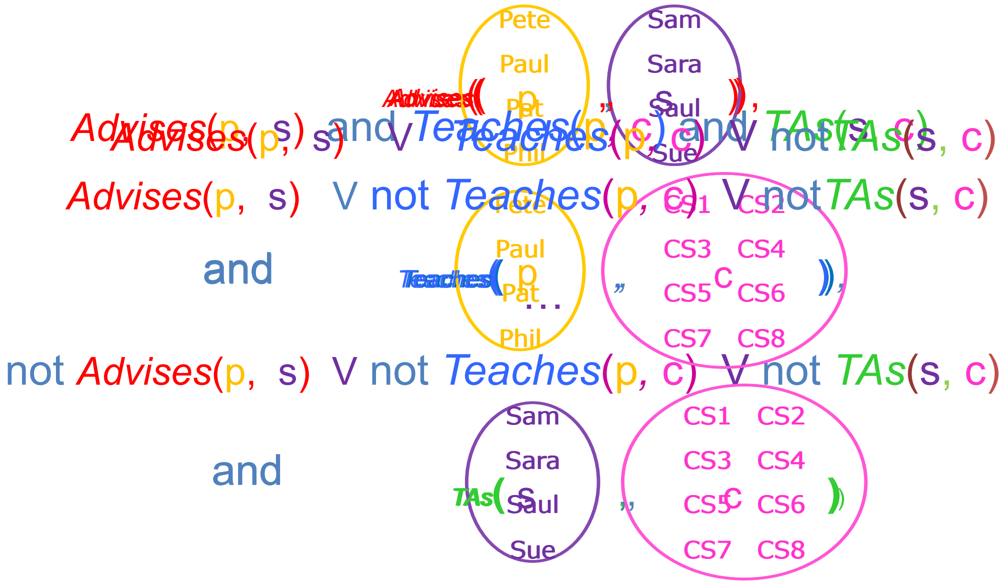
Paths Found

Advises(○, ○)

Advises(○, ○),
Teaches (○, ○)

Advises(○, ○),
Teaches (○, ○),
TAs(○, ○)

Clause Creation



Advises(p, s) and *Teaches*(p, c) and *TAs*(s, c)

Advises(p, s) ∨ *Teaches*(p, c) ∨ not *TAs*(s, c)

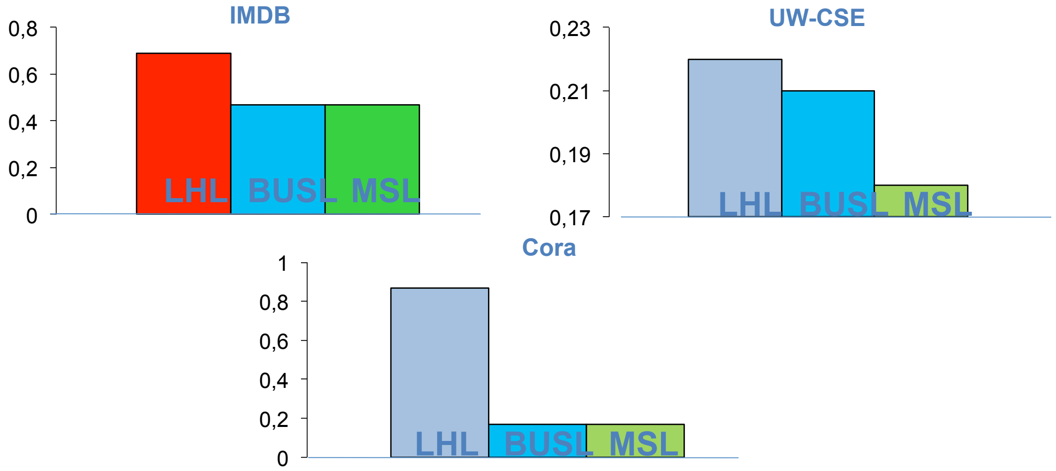
Advises(p, s) ∨ not *Teaches*(p, c) ∨ not *TAs*(s, c)

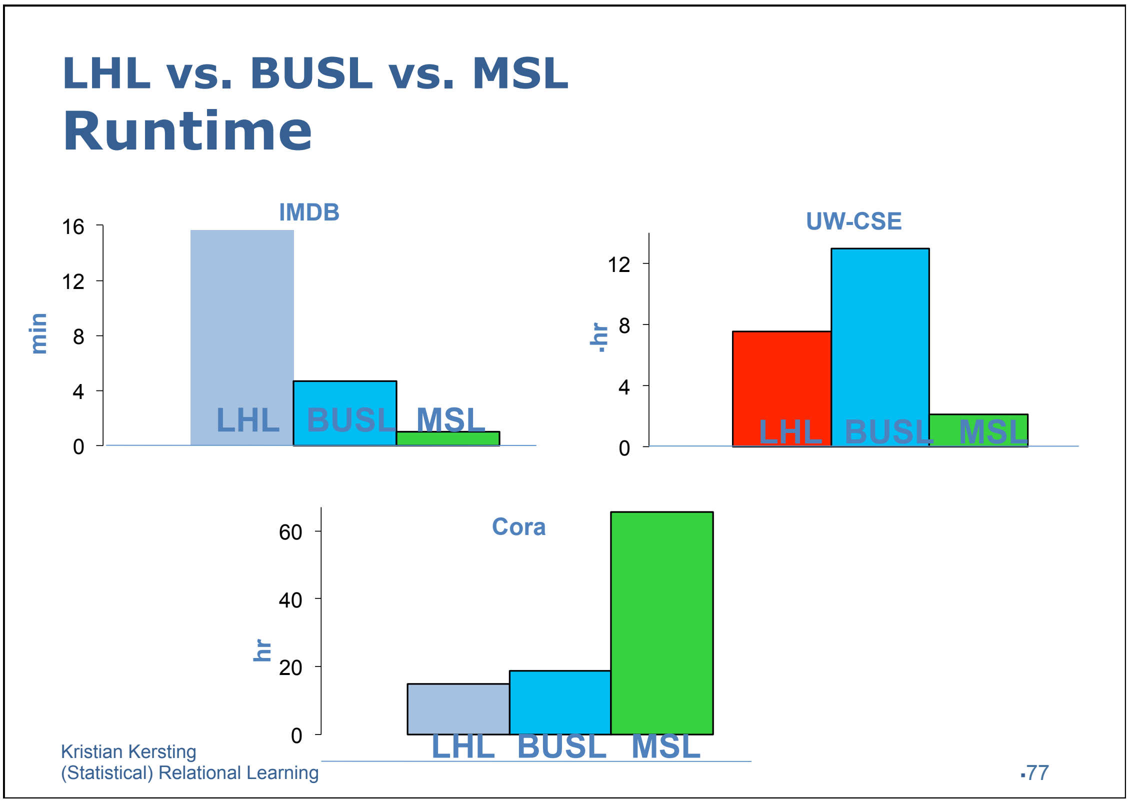
and

not *Advises*(p, s) ∨ not *Teaches*(p, c) ∨ not *TAs*(s, c)

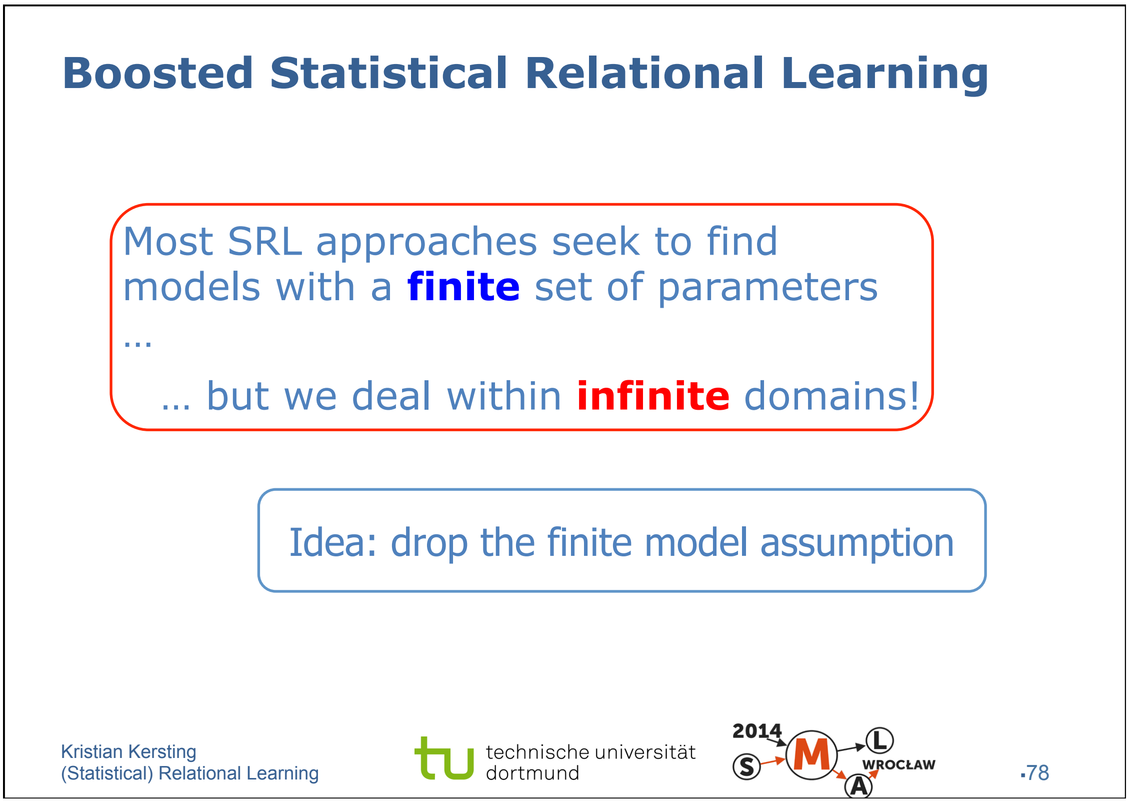
and

LHL vs. BUSL vs. MSL
Area under Prec-Recall Curve





LHL vs. BUSL vs. MSL
Runtime

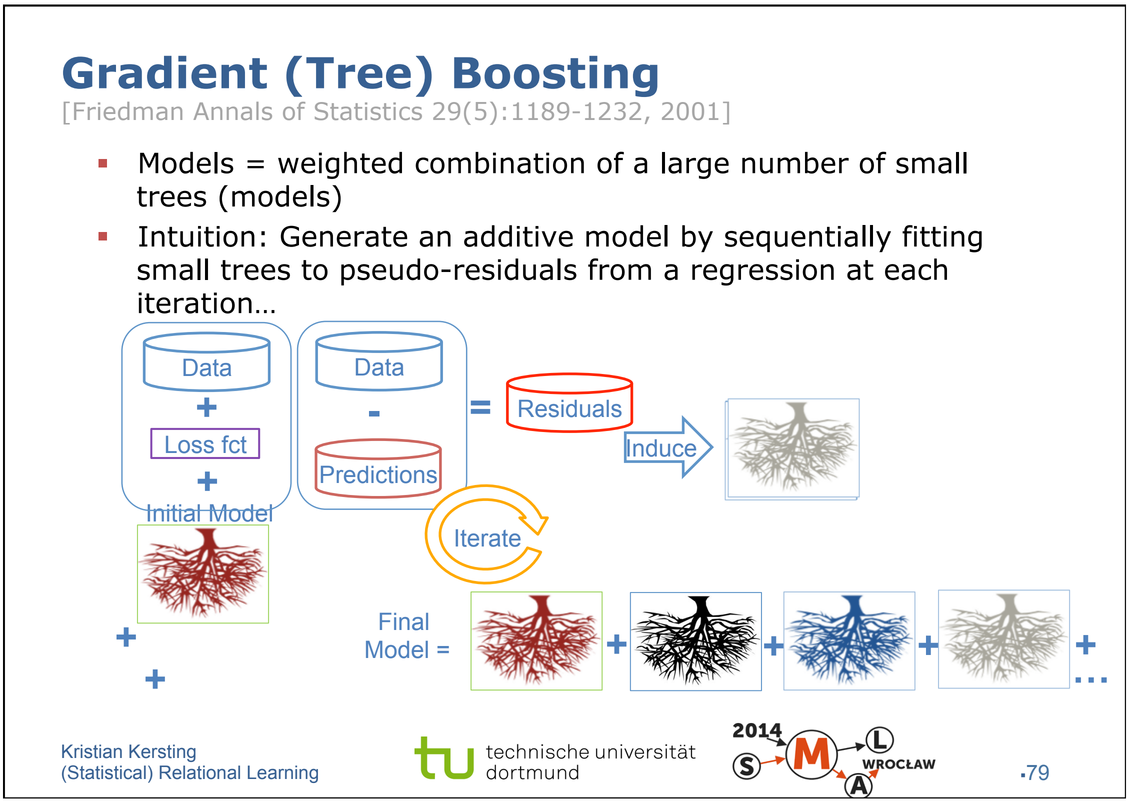


Boosted Statistical Relational Learning

Most SRL approaches seek to find models with a **finite** set of parameters …

… but we deal within **infinite** domains!

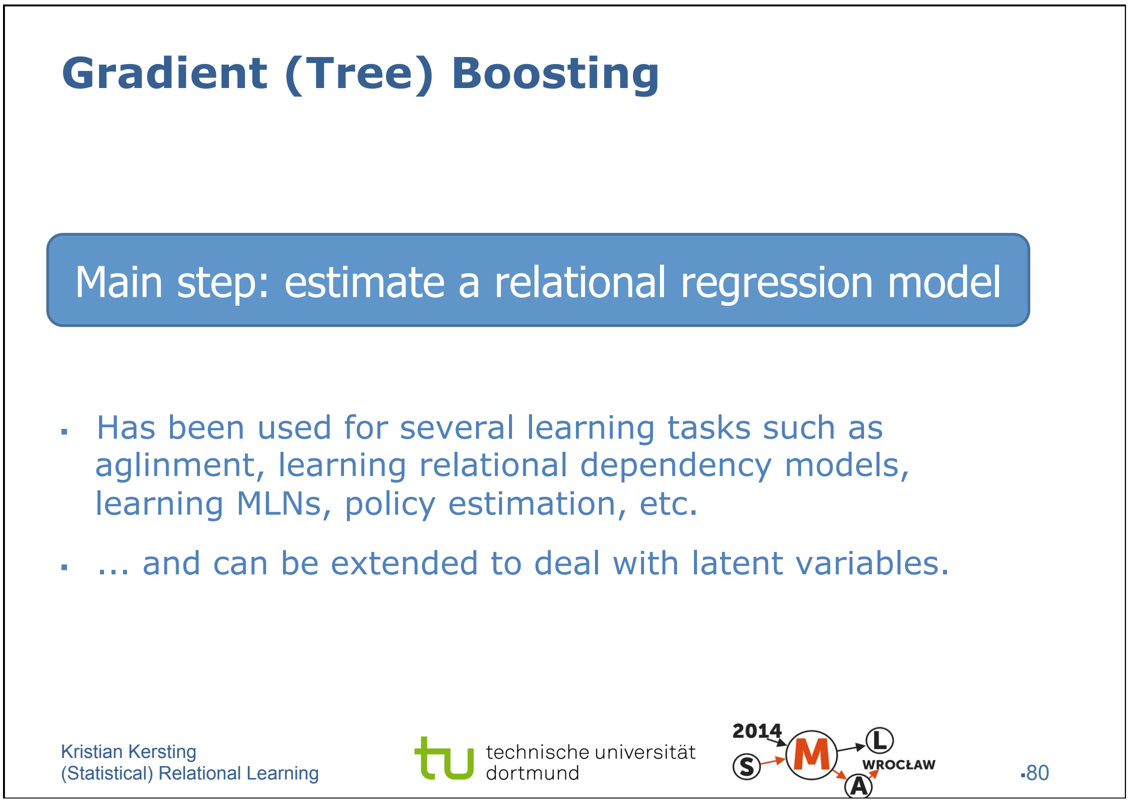
Idea: drop the finite model assumption



Gradient (Tree) Boosting

[Friedman Annals of Statistics 29(5):1189-1232, 2001]

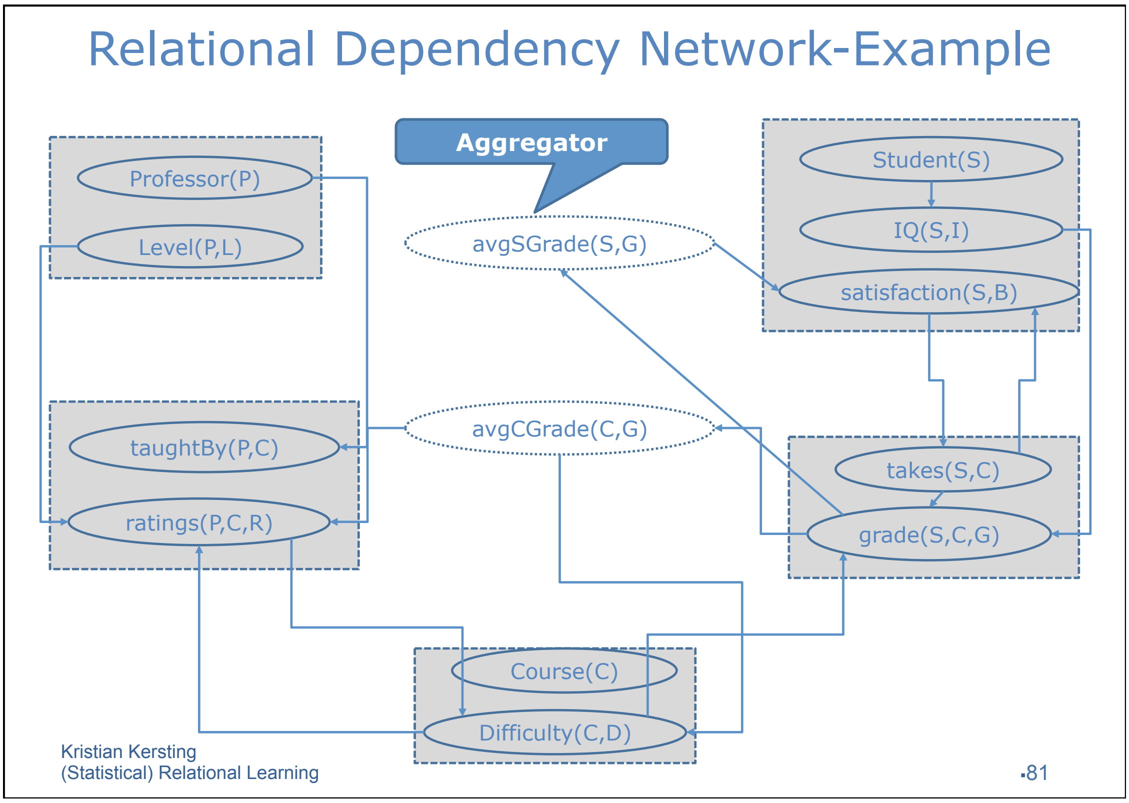
- Models = weighted combination of a large number of small trees (models)
- Intuition: Generate an additive model by sequentially fitting small trees to pseudo-residuals from a regression at each iteration…



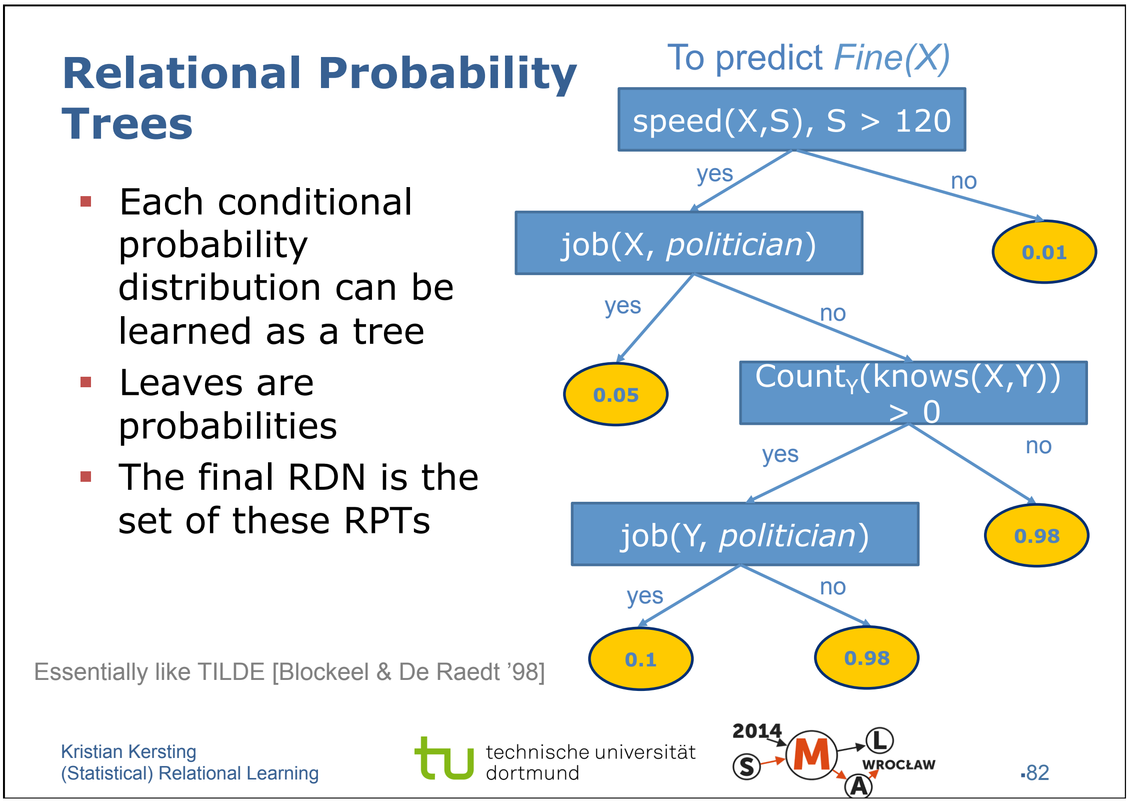
Gradient (Tree) Boosting

Main step: estimate a relational regression model

- Has been used for several learning tasks such as aglinment, learning relational dependency models, learning MLNs, policy estimation, etc.
- … and can be extended to deal with latent variables.



Relational Dependency Network-Example



Relational Probability Trees

- Each conditional probability distribution can be learned as a tree
- Leaves are probabilities
- The final RDN is the set of these RPTs

To predict *Fine(X)*

Essentially like TILDE [Blockeel & De Raedt ’98]

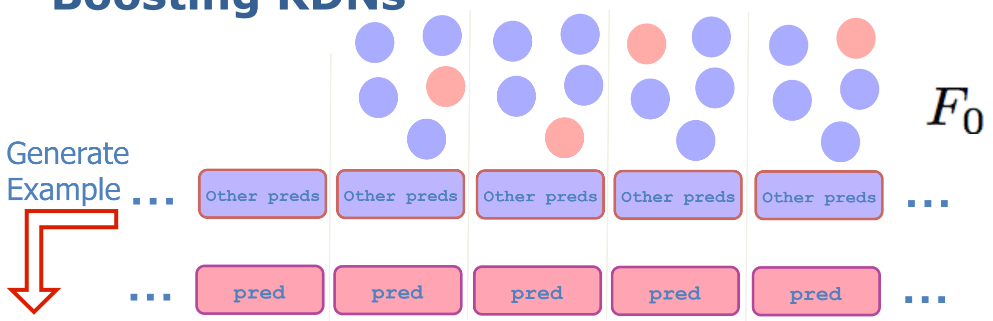
Gradient Tree Boosting

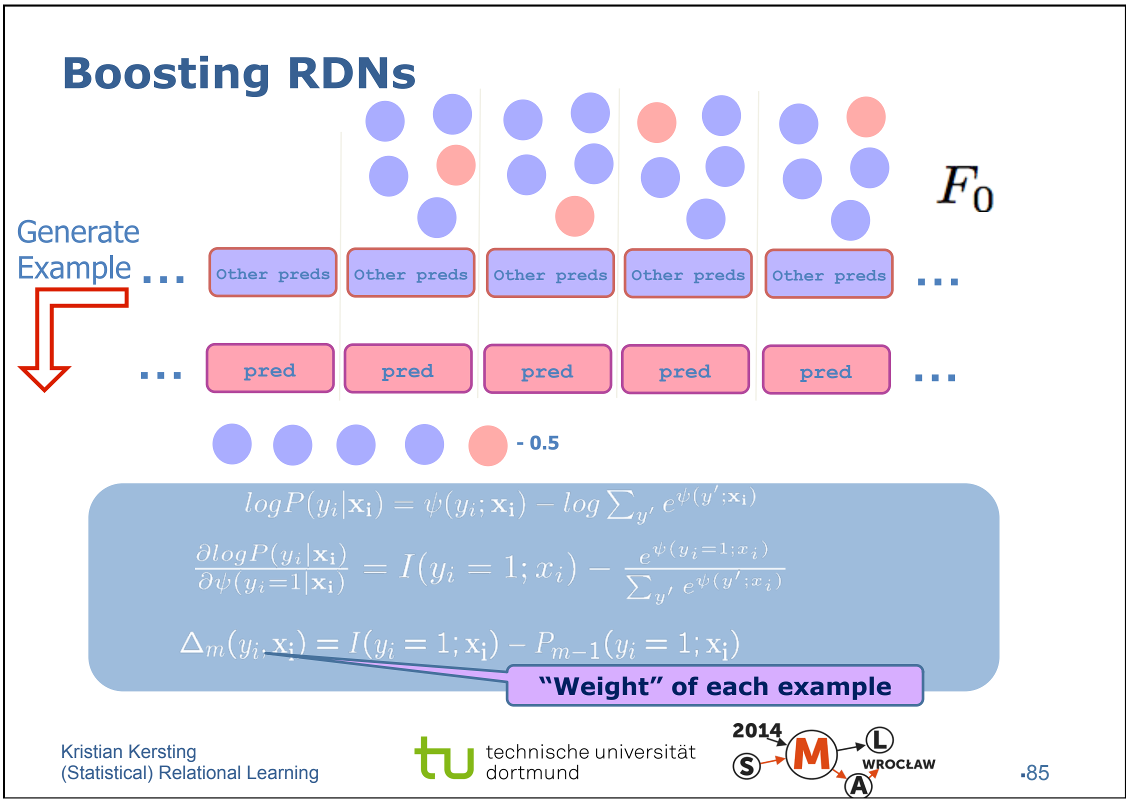
- Find ML parameters, i.e. maximize $\log P(Y|X)$ without fixing the model structure/features
- Functional Gradient

$$F_m = F_0 + \Delta_1 + \ldots + \Delta_m$$

$$\Delta_m = \eta_m \cdot E_{x,y}\left[\frac{\partial}{\partial F_{m-1}} \log P(y|x; F_{m-1})\right]$$

Boosting RDNs





Boosting RDNs

$$log P(y_i|\mathbf{x_i}) = \psi(y_i;\mathbf{x_i}) - log \sum_{y'} e^{\psi(y';\mathbf{x_i})}$$

$$\frac{\partial log P(y_i|\mathbf{x_i})}{\partial \psi(y_i=1|\mathbf{x_i})} = I(y_i=1;x_i) - \frac{e^{\psi(y_i=1;x_i)}}{\sum_{y'} e^{\psi(y';x_i)}}$$

$$\Delta_m(y_i;\mathbf{x_i}) = I(y_i=1;\mathbf{x_i}) - P_{m-1}(y_i=1;\mathbf{x_i})$$

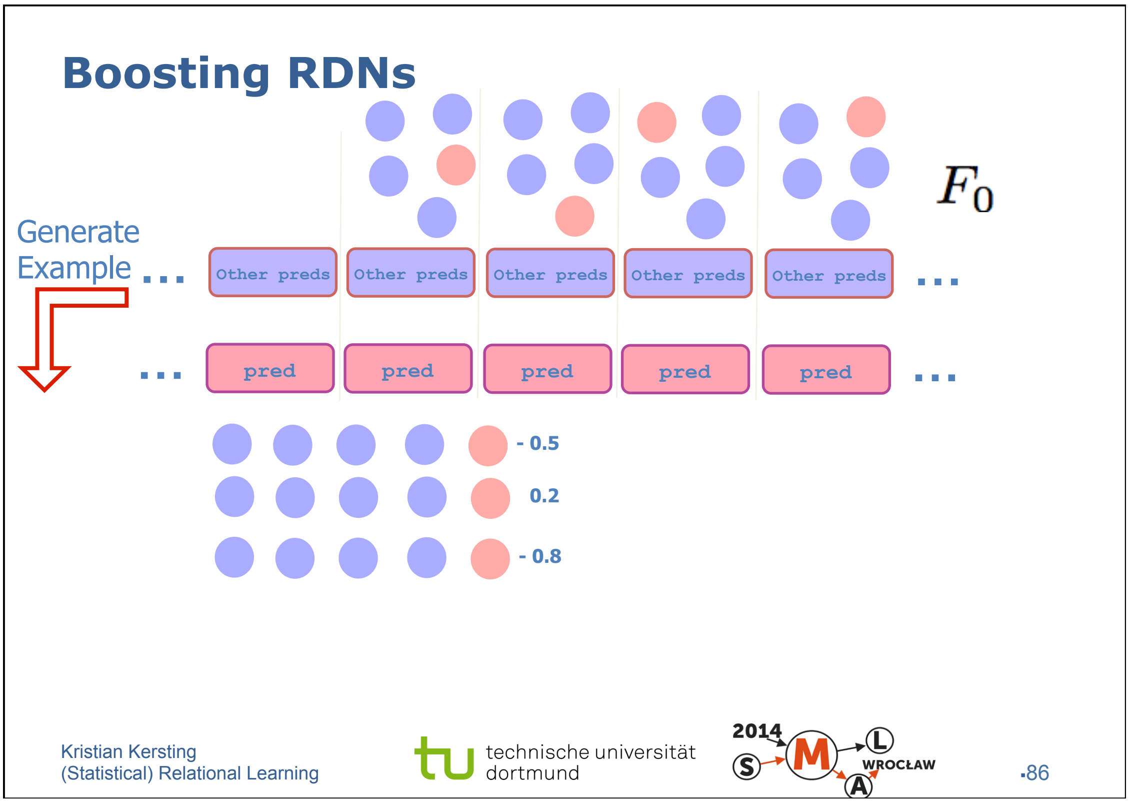
"Weight" of each example

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 S M L A WROCŁAW

▪85



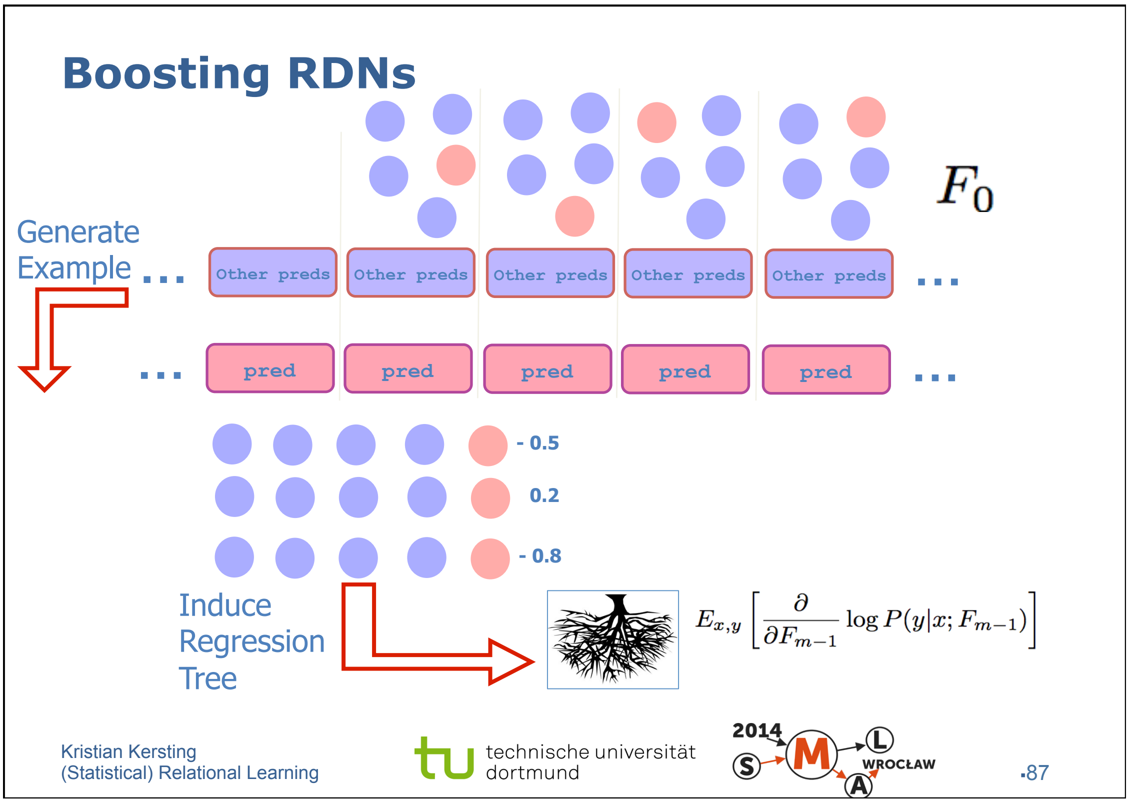
Boosting RDNs

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 S M L A WROCŁAW

▪86



Boosting RDNs

Generate Example

... Other preds Other preds Other preds Other preds Other preds ...

... pred pred pred pred pred ...

- 0.5

0.2

- 0.8

Induce Regression Tree

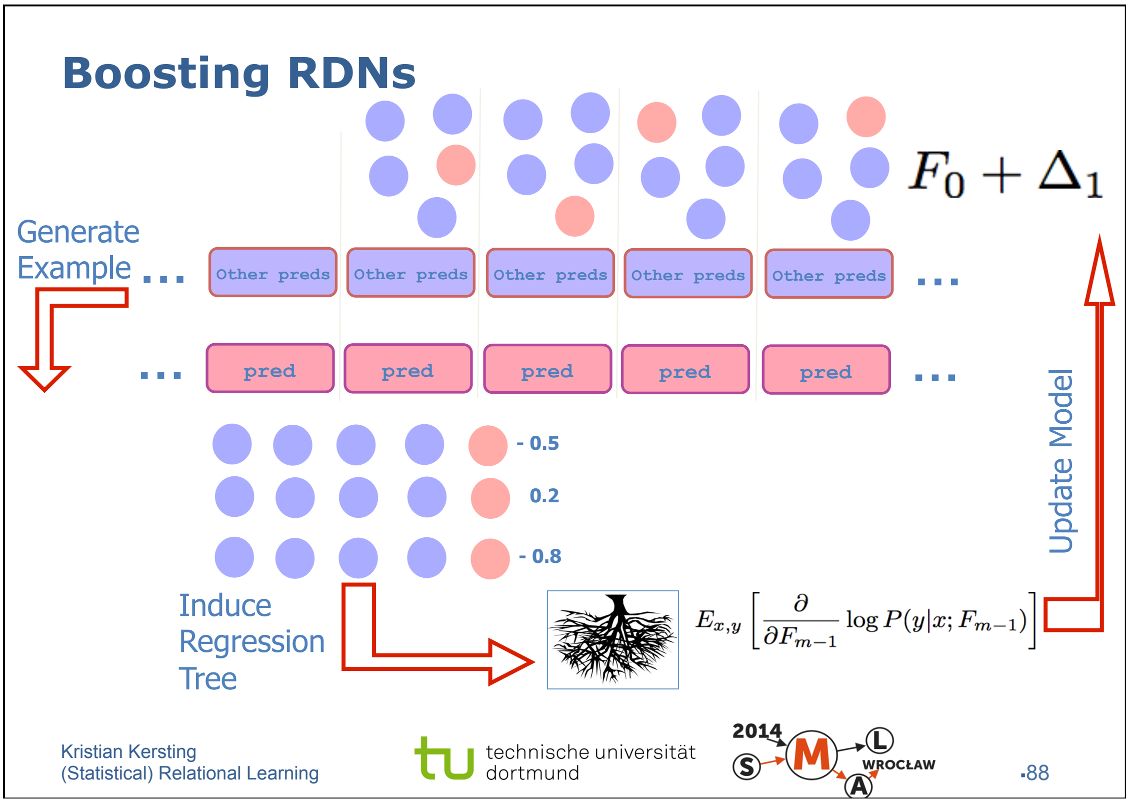
$$E_{x,y}\left[\frac{\partial}{\partial F_{m-1}}\log P(y|x;F_{m-1})\right]$$

F_0

Kristian Kersting
(Statistical) Relational Learning

technische universität
dortmund

2014 S M L A WROCŁAW



Boosting RDNs

Generate Example

... Other preds Other preds Other preds Other preds Other preds ...

... pred pred pred pred pred ...

- 0.5

0.2

- 0.8

Induce Regression Tree

$$E_{x,y}\left[\frac{\partial}{\partial F_{m-1}}\log P(y|x;F_{m-1})\right]$$

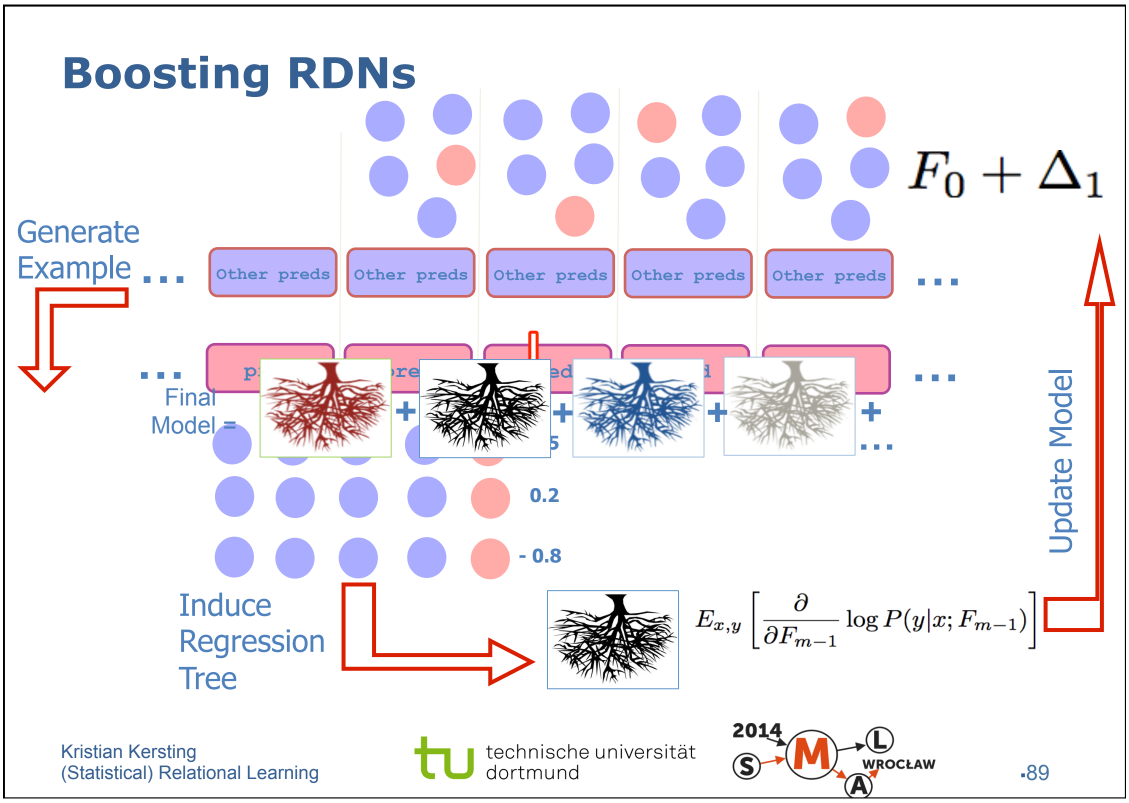
Update Model

$F_0 + \Delta_1$

Kristian Kersting
(Statistical) Relational Learning

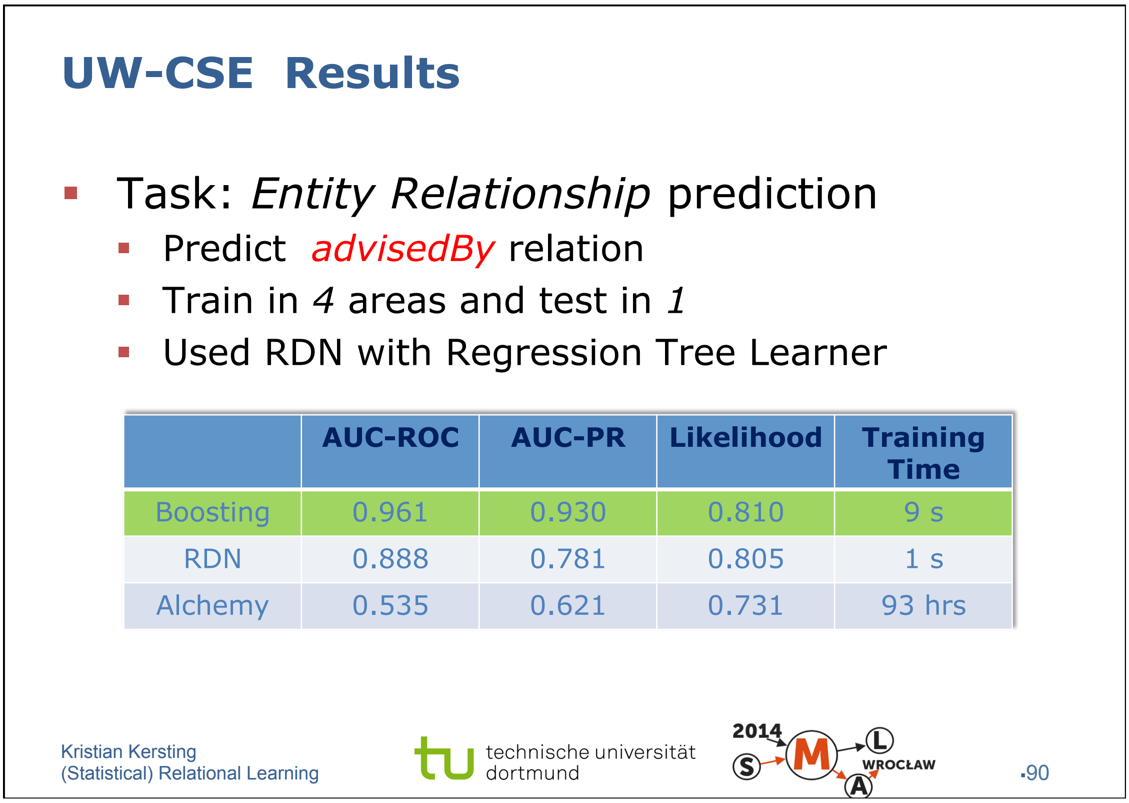
technische universität
dortmund

2014 S M L A WROCŁAW



Boosting RDNs

$$E_{x,y}\left[\frac{\partial}{\partial F_{m-1}} \log P(y|x;F_{m-1})\right]$$



UW-CSE Results

- Task: *Entity Relationship* prediction
  - Predict *advisedBy* relation
  - Train in *4* areas and test in *1*
  - Used RDN with Regression Tree Learner

	AUC-ROC	AUC-PR	Likelihood	Training Time
Boosting	0.961	0.930	0.810	9 s
RDN	0.888	0.781	0.805	1 s
Alchemy	0.535	0.621	0.731	93 hrs

OMOP Results

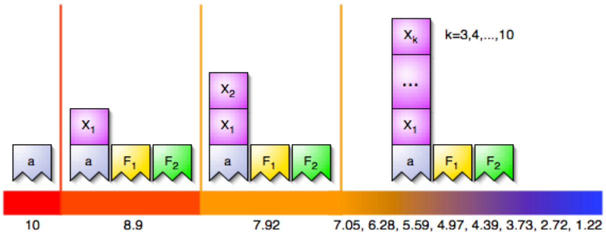
- Task: Predict *Adverse-drug events*
  - Input: Drugs and conditions (side-effects)
  - Goal: Predict if a patient is on a given drug (*onDrug(D,P)*)
  - Learning "in reverse"
  - Averaged over 5 train-test sets
  - Each set is a different drug

	AUC-ROC	AUC-PR	Accuracy	Training Time
Boosting	0.824	0.839	0.753	497.8 s
RDN	0.738	0.736	0.697	39.4 s
ILP + Noisy-Or	0.420	0.582	0.687	2400 s

Direct Policy Learning

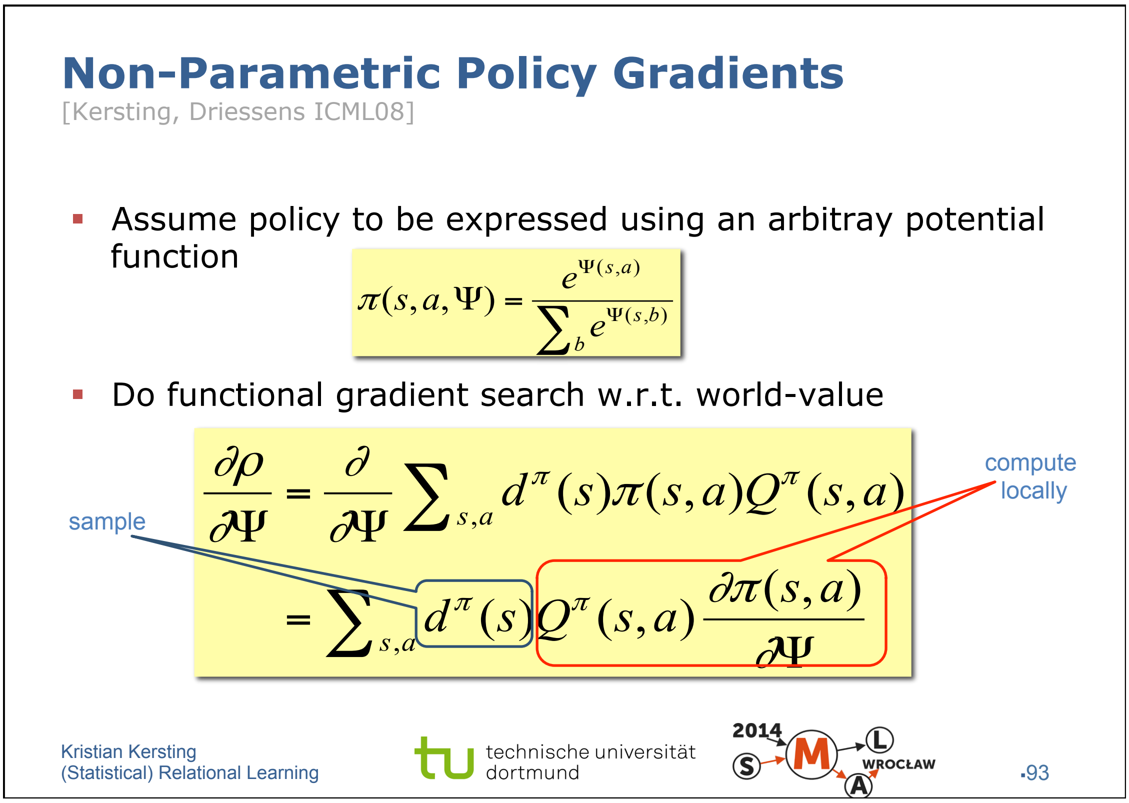
- Value functions can often be much more complex to represent than the corresponding policy

Goal: cl(a)



- When poli ıs than the correspon ı **policy space can be a good idea**

Policy: put each block on top of a on the floor



Non-Parametric Policy Gradients

[Kersting, Driessens ICML08]

- Assume policy to be expressed using an arbitray potential function

$$\pi(s,a,\Psi) = \frac{e^{\Psi(s,a)}}{\sum_b e^{\Psi(s,b)}}$$

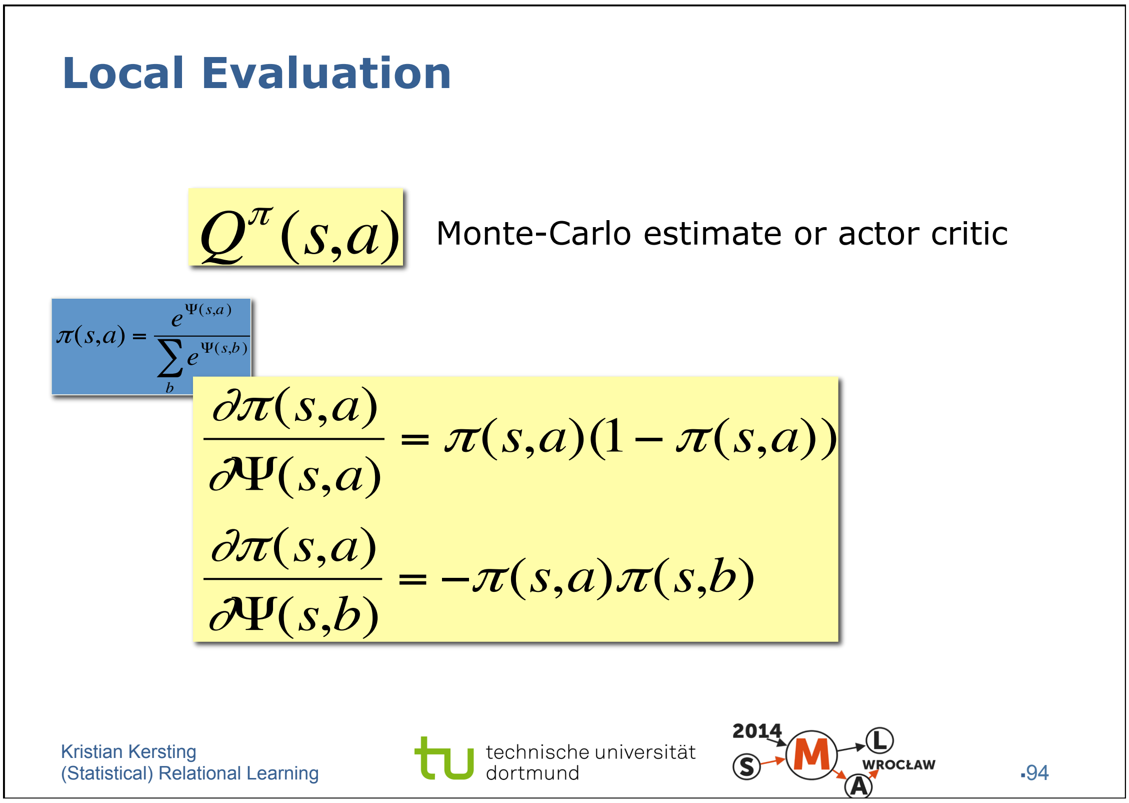
- Do functional gradient search w.r.t. world-value

$$\frac{\partial\rho}{\partial\Psi} = \frac{\partial}{\partial\Psi}\sum_{s,a} d^{\pi}(s)\pi(s,a)Q^{\pi}(s,a)$$

$$= \sum_{s,a} \boxed{d^{\pi}(s)}\boxed{Q^{\pi}(s,a)\frac{\partial\pi(s,a)}{\partial\Psi}}$$

sample

compute locally



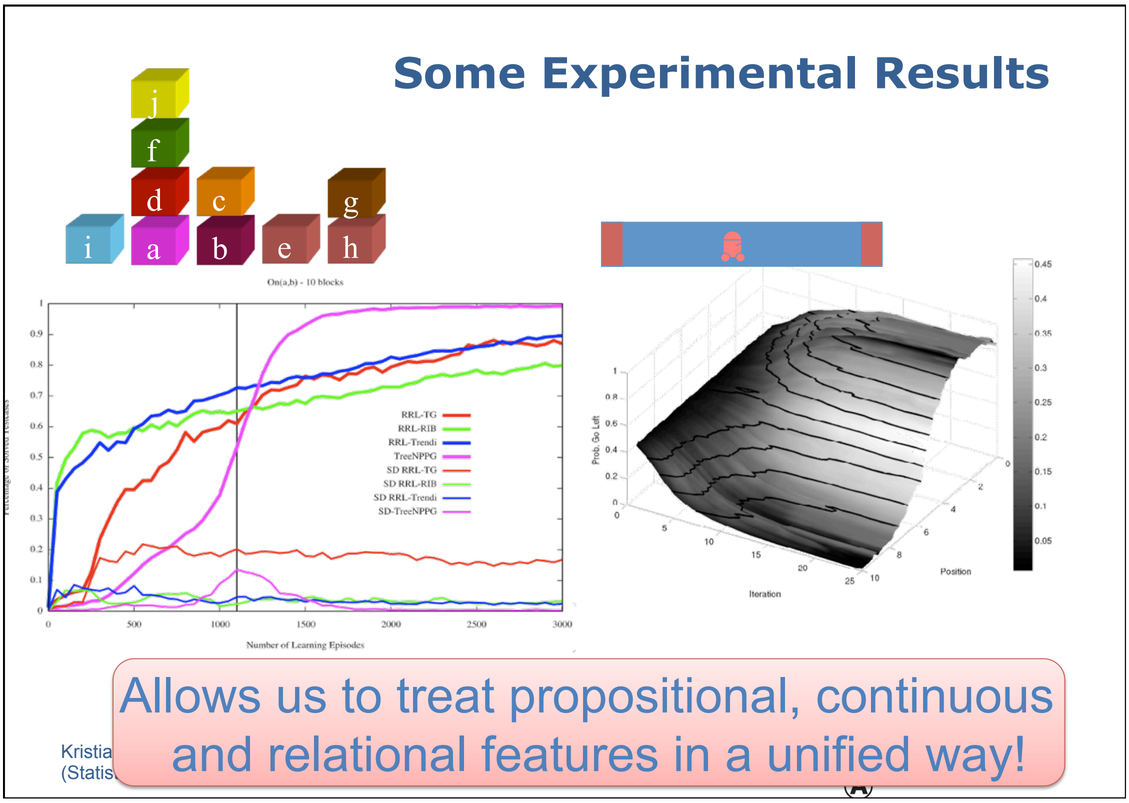
Local Evaluation

$Q^{\pi}(s,a)$ Monte-Carlo estimate or actor critic

$$\pi(s,a) = \frac{e^{\Psi(s,a)}}{\sum_b e^{\Psi(s,b)}}$$

$$\frac{\partial\pi(s,a)}{\partial\Psi(s,a)} = \pi(s,a)(1-\pi(s,a))$$

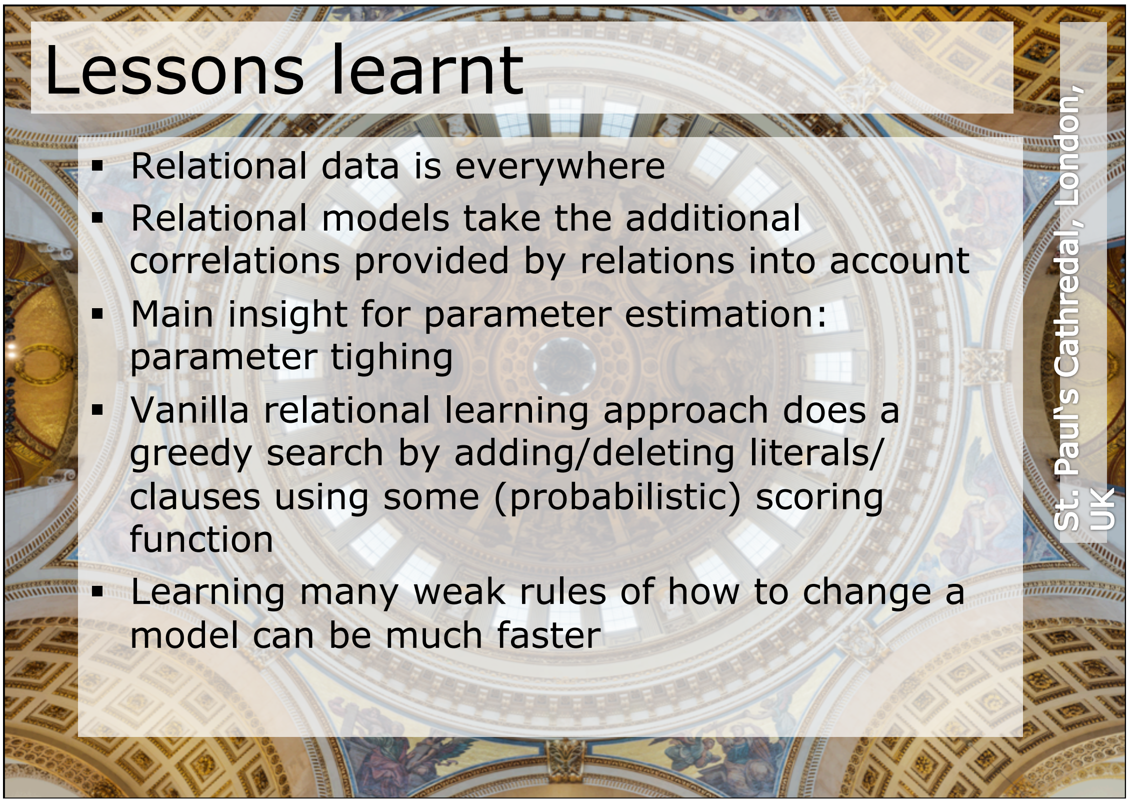
$$\frac{\partial\pi(s,a)}{\partial\Psi(s,b)} = -\pi(s,a)\pi(s,b)$$



Some Experimental Results

Allows us to treat propositional, continuous and relational features in a unified way!

Kristia
(Statis



Lessons learnt

- Relational data is everywhere
- Relational models take the additional correlations provided by relations into account
- Main insight for parameter estimation: parameter tighing
- Vanilla relational learning approach does a greedy search by adding/deleting literals/ clauses using some (probabilistic) scoring function
- Learning many weak rules of how to change a model can be much faster

St. Paul's Cathredal, London, UK