School of Computing Science

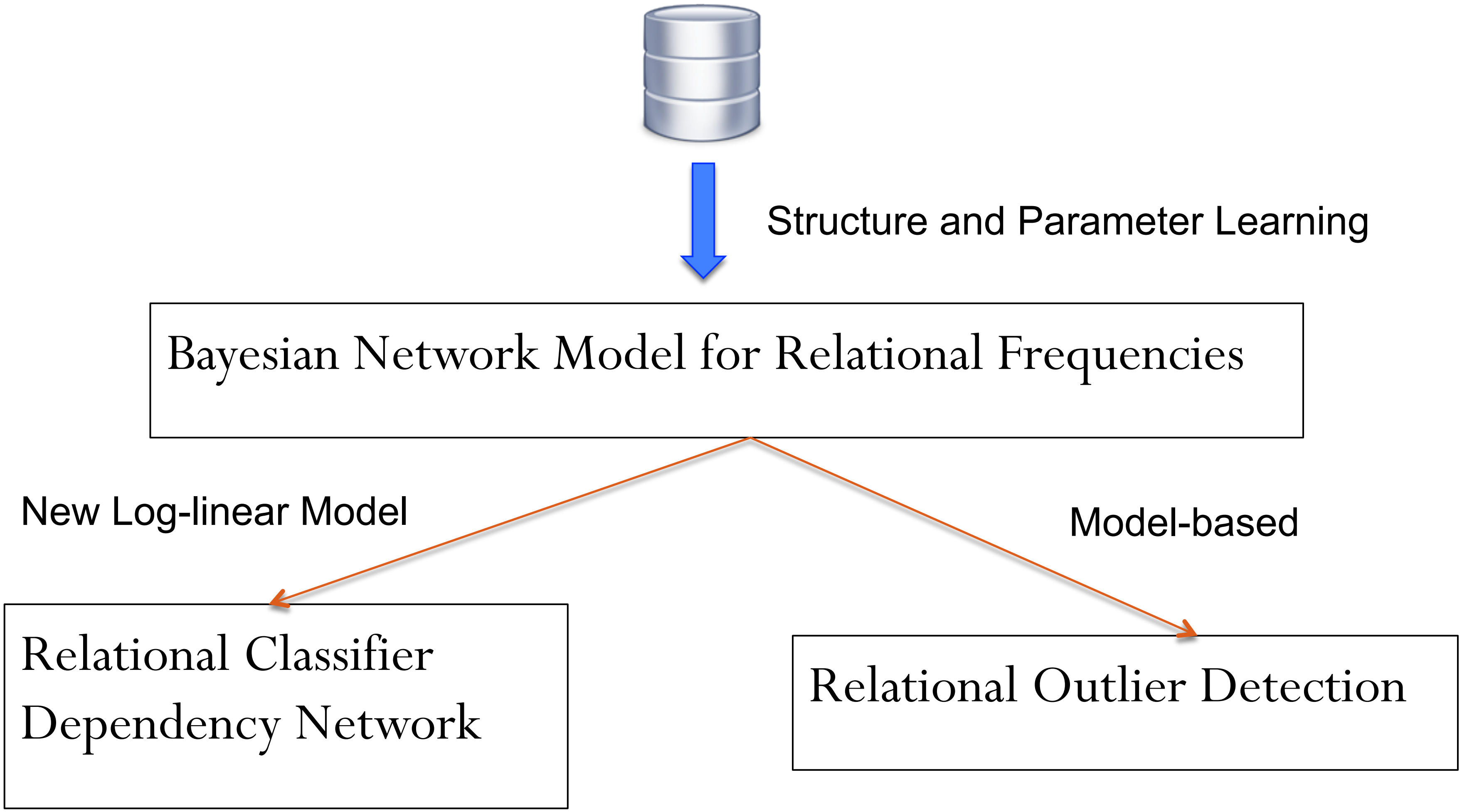
Simon Fraser University

Vancouver, Canada

From Relational Statistics to Degrees of Belief

Oliver Schulte	Hassan Khosravi	Arthur Kirkpatrick	Tianxiang Gao	Yuke Zhu

Overview


Structure and Parameter Learning
Bayesian Network Model for Relational Frequencies
New Log-linear Model
Model-based
Relational Classifier Dependency Network
Relational Outlier Detection

Relational Data and Logic

Lise Getoor

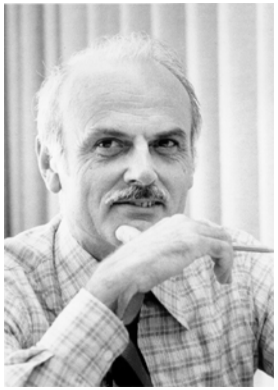
David Poole

Stuart Russsell

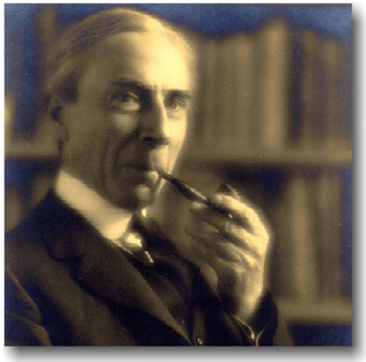
Database Management Systems

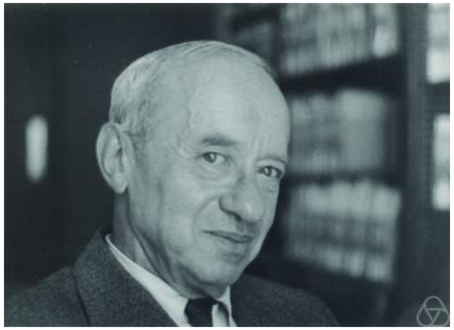
- Maintain data in linked tables.
- Structured Query Language (SQL) allows fast *data retrieval*.
  - E.g., find all movie ratings > 4 where the user is a woman.
- Multi-billion dollar industry, $Bn 15+ in 2006.
- IBM, Microsoft, Oracle, SAP, Peoplesoft.
- Much interest in analysis (big data, data mining, business intelligence, predictive analytics, OLAP…)

The Relational Data Model

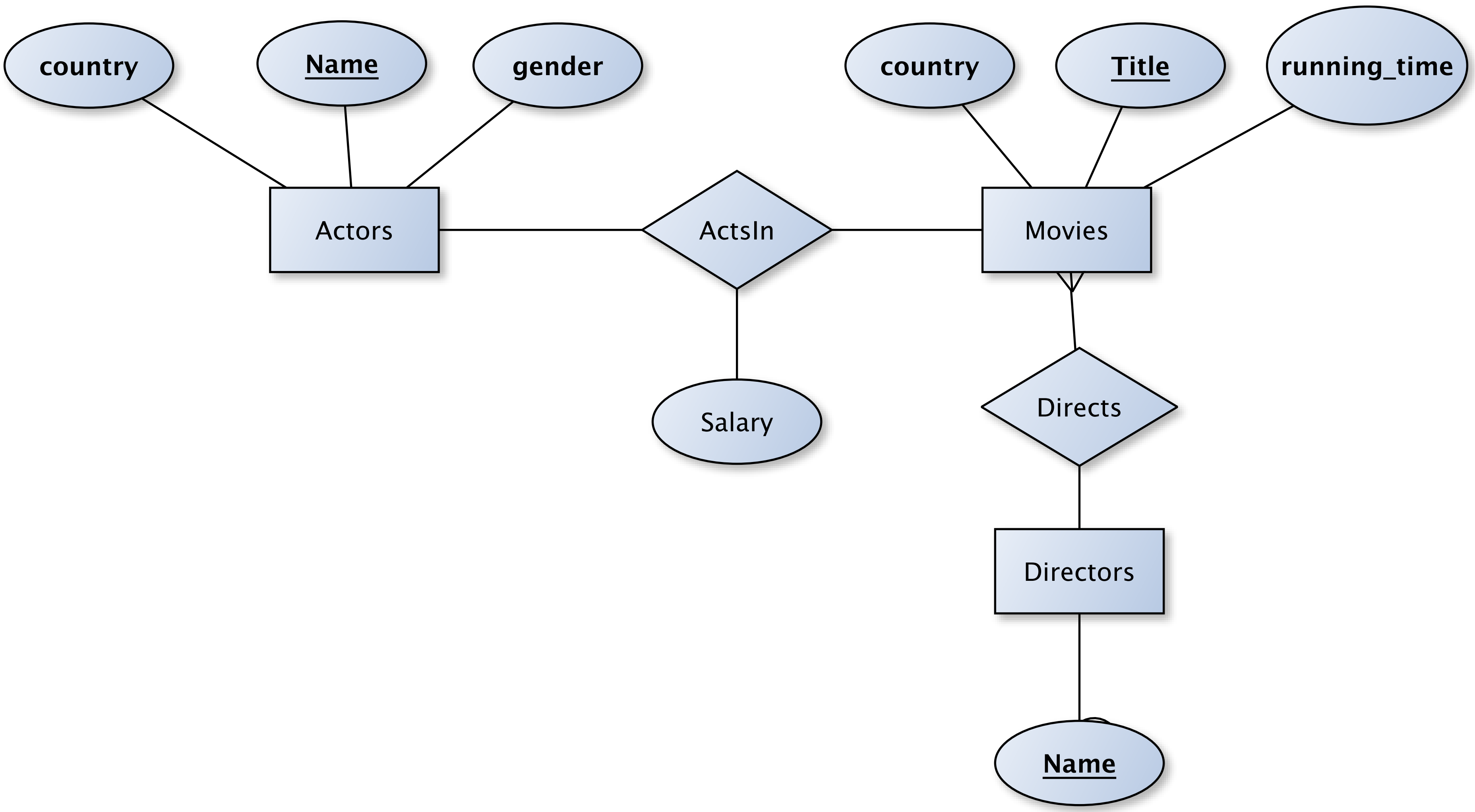


- Codd (IBM Research 1970)
- The fundamental question: *What kinds of information do users need to represent?*
- Answered by first-order predicate logic! (Russell, Tarski).
- The world consists of
  - Individuals/entities.
  - Relationships/links among them.

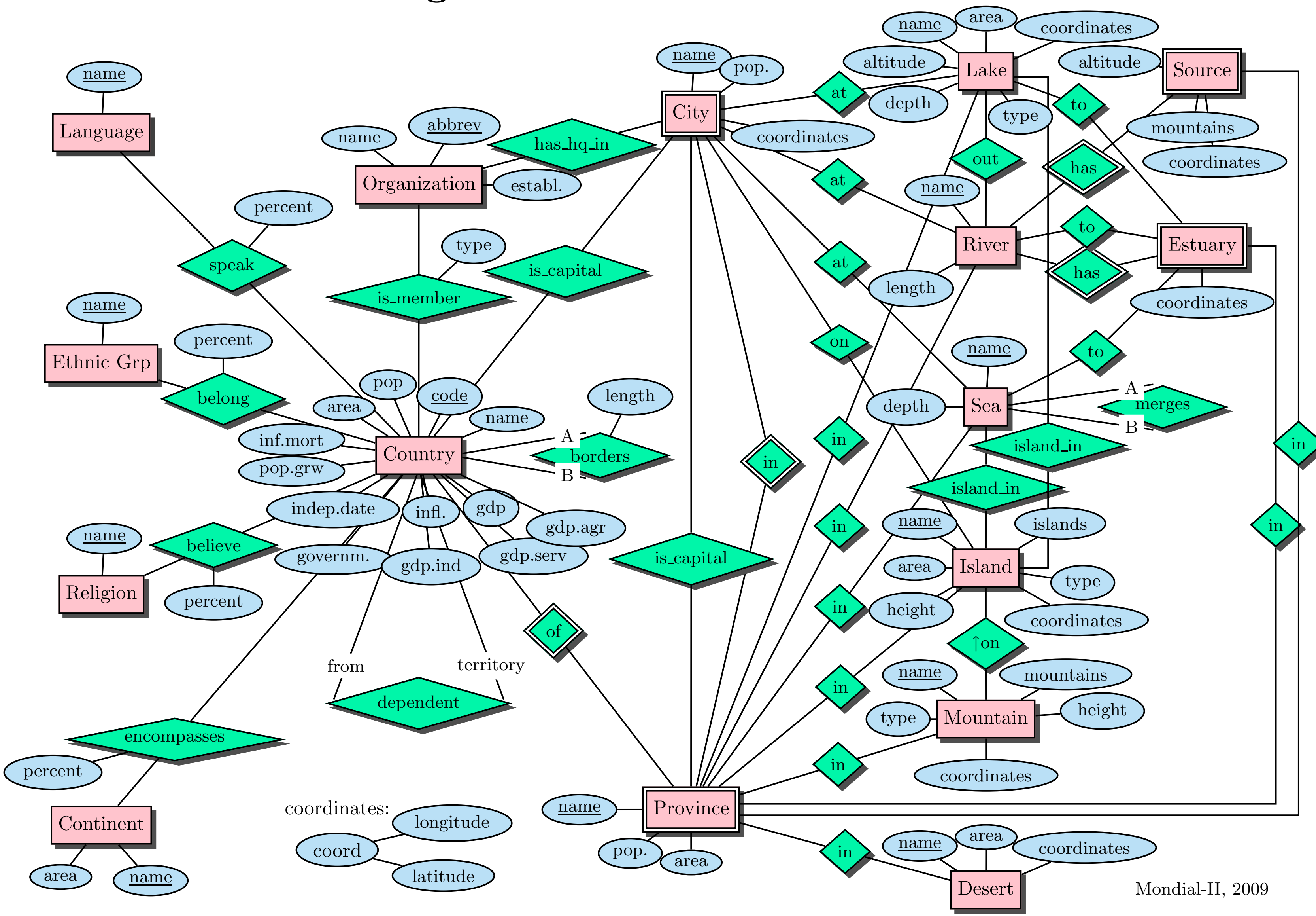




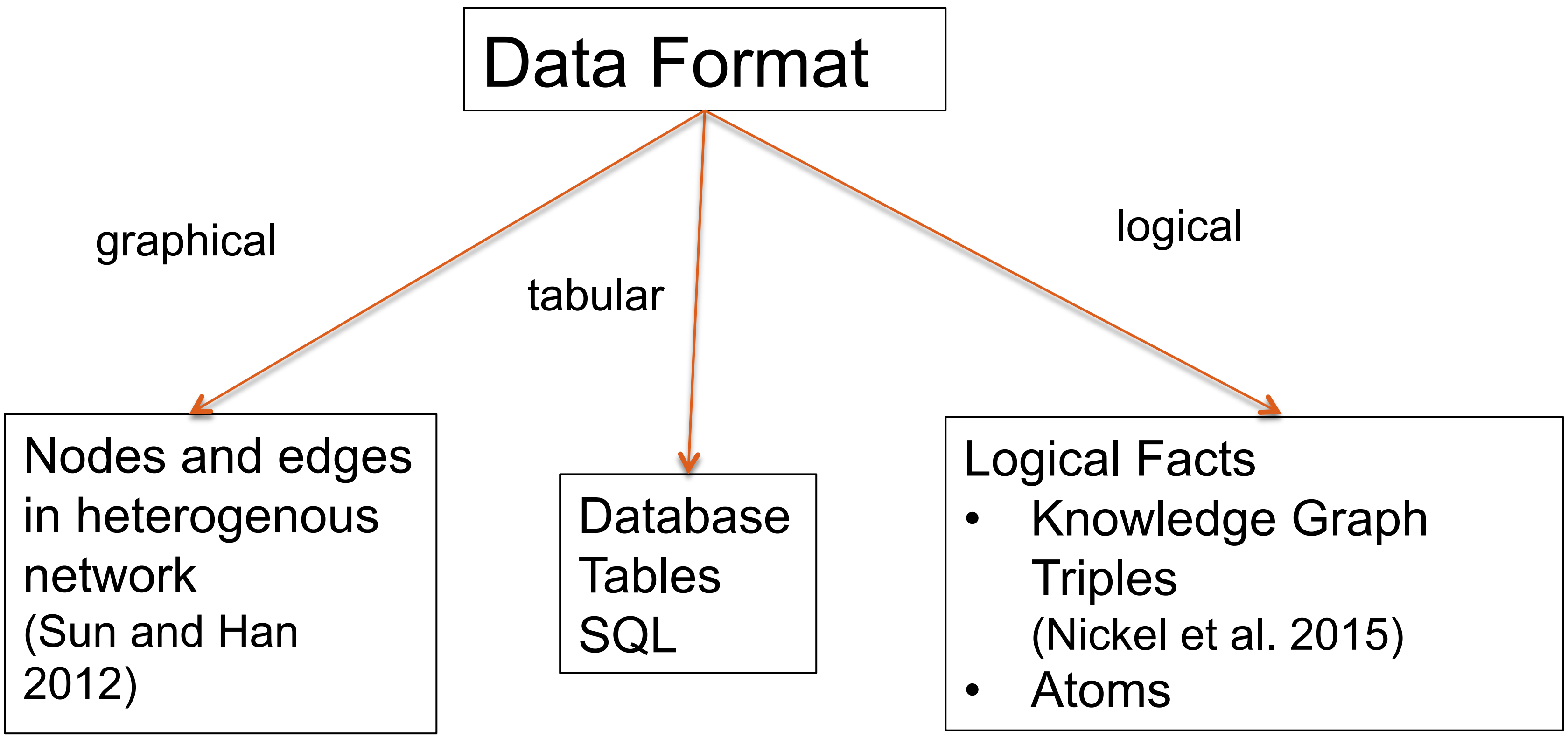
Entity-Relationship Diagram IMDb



ER-Diagram of the Mondial Database



Relational Data Formats



Sun, Y. & Han, J. (2012), Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers.

Nickel, M.; Murphy, K.; Tresp, V. & Gabrilovich, E. (2015), 'A Review of Relational Machine Learning for Knowledge Graphs', *ArXiv e-prints* .

Logical Representation

Edsger Dijkstra by Hamilton Richards

- Standard in database theory.
- Unify logic and probability.
- Equational logic (Dijkstra and Scholten 1990) is especially similar to random variable concepts in statistics.
  - Represent relational information using **functions** (functors) (Poole 2003).
  - Single table data : All functions take 1 argument (Nickel et al. 2015).
  - Relational data: Some functions take > 1 argument.

Poole, D. (2003), First-order probabilistic inference, *in 'IJCAI'*.

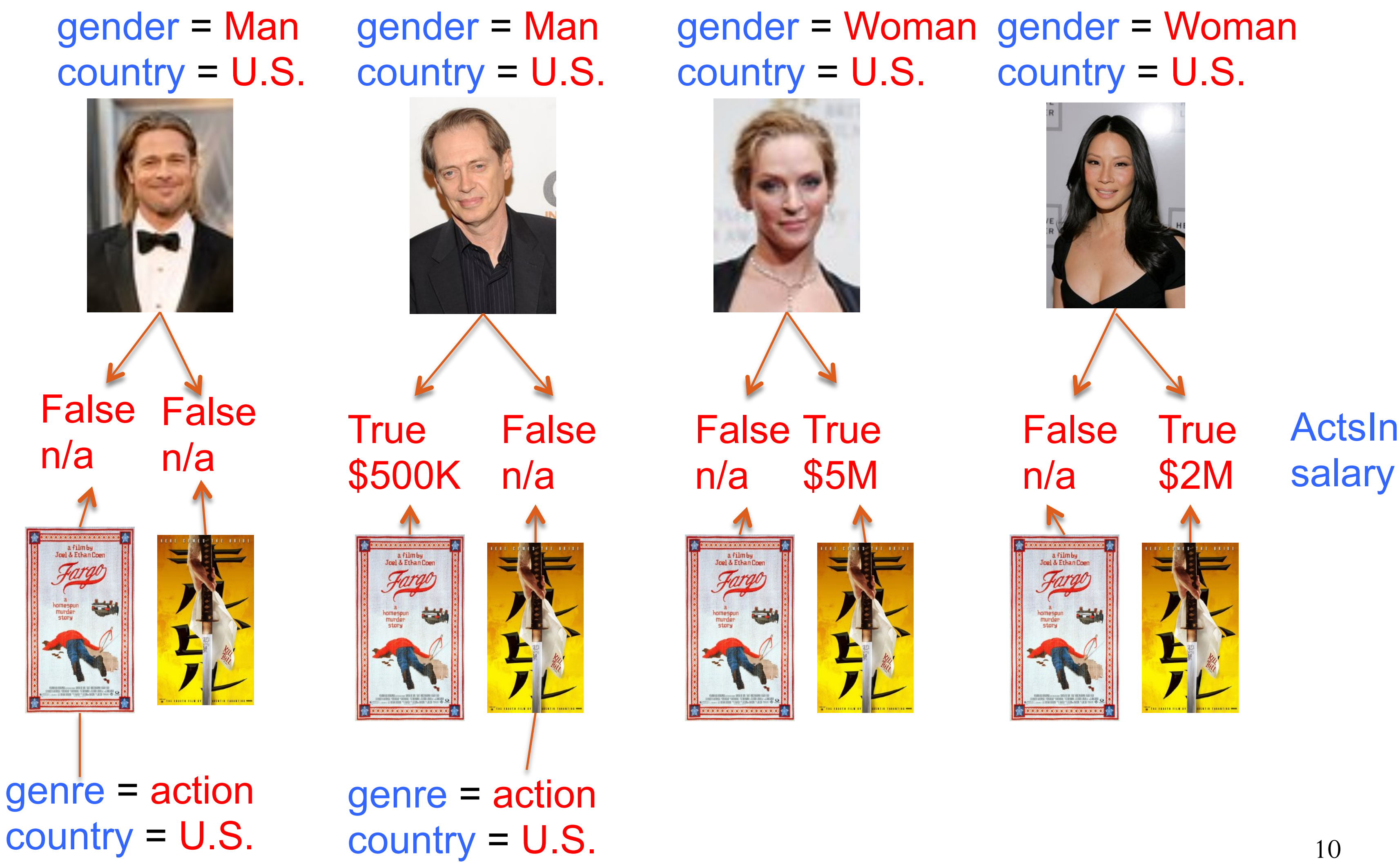
Getoor, L. & Grant, J. (2006), 'PRL: A probabilistic relational language', *Machine Learning* ***62(1-2), 7-31.***

Russell, S. & Norvig, P. (2010), *Artificial Intelligence: A Modern Approach, Prentice Hall.*

Ravkic, I.; Ramon, J. & Davis, J. (2015), 'Learning relational dependency networks in hybrid domains', *Machine Learning.*

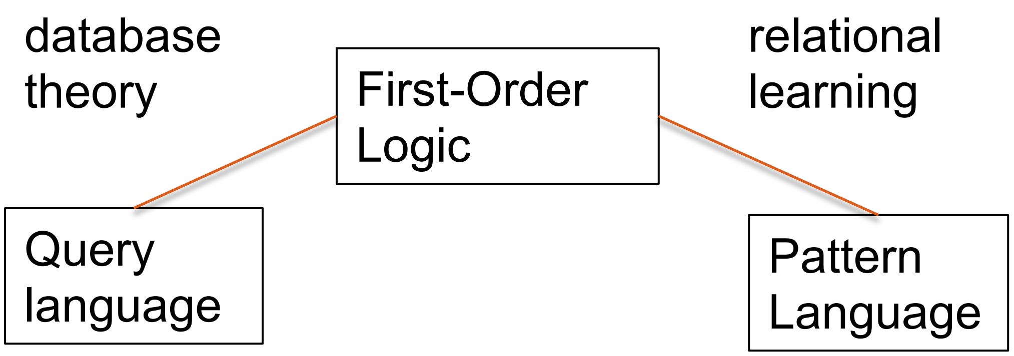
Dijskstra & Scholten (1990), *Predicate calculus and program semantics, Springer Verlag.*

Function Representation Example



First-Order Logic

An expressive formalism for specifying relational conditions.



Kimmig, A.; Mihalkova, L. & Getoor, L. (2014), 'Lifted graphical models: a survey', *Machine Learning, 1--45*.

First-Order Logic: Terms

- A constant refers to an individual.
  - e.g. “Fargo”
- A logical variable refers to a class of individuals
  - e.g. “Movie” refers to Movies.
- A ground term is of the form $f(a_1,..,a_n)$
  - e.g. “salary(UmaThurman, Fargo)”
- A first-order term is of the form $f(t_1,..,t_n)$ where at least one of the t_i is a first-order variable.
  - e.g. “salary(Actor, Movie)”.

Kimmig, A.; Mihalkova, L. & Getoor, L. (2014), 'Lifted graphical models: a survey', *Machine Learning, 1—45*.

Formulas (Equational Logic)

- A (conjunctive) formula is a conjunction $term_1 = value_1, \ldots, term_n = value_n$.
  - ActsIn(Actor, Movie) = T, gender(Actor) = W
- A *ground* formula contains only constants.
  - ActsIn(UmaThurman, KillBill) = T, gender(UmaThurman) = W

Two Kinds of Probability

Frequencies vs. Single Event Probabilities

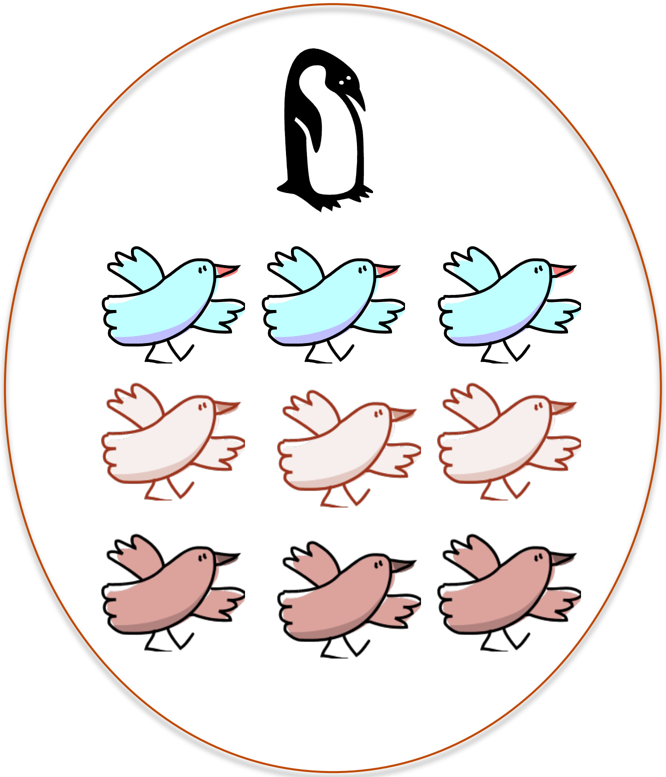
Joe Halpern

Fahim Bacchus

Frequencies/Proportions

- Classical statistics aims to estimate population frequencies or proportions.

Proportion
90% of birds fly.
0% of planes have crashed because of a turbulence.
5% of Brusselians speak only Dutch at home.
51.1% of U.S. voters voted for Barack Obama.



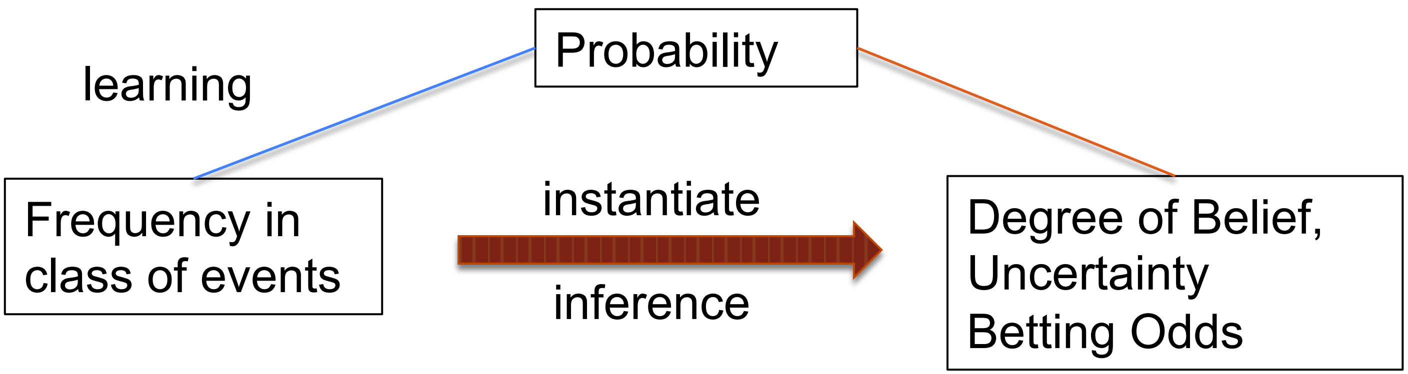
Probabilities for Single Events

- *Bayesian* statistics emphasizes assigning probabilities to single events.
  - Including the values of model parameters.

Proportion	Instance
90% of birds fly.	The probability that Tweety flies is 90%.
0% of planes have crashed because of a turbulence.	The probability that Flight 3202 to Brussels crashes because of a turbulence is 0%.
5% of Brusselians speak only Dutch.	Given that Marc lives in Brussels, the probability that he speaks only Dutch is 5%.
	The probability that the mean $\mu = 0$ for a Gaussian distribution is 0.01.

Heckerman, D. (1998), A tutorial on learning with Bayesian networks, *in 'NATO ASI on Learning in graphical models', pp. 301--354.*

Classical Statistics: Two Kinds of Probabilities



de Finetti 1937: La Prévision: ses lois logiques, ses sources subjectives, Annales de l'Institut Henri Poincaré

Two Kinds of Relational Probabilities

P(first-order formulas)

Relational Probability

P(ground formulas)

Relational Frequency
Statistics
type 1 probability
Class-level probability

instantiate (?)

inference (?)

Degree of Belief
type 2 probability
Instance-level probability

The Halpern instantiation principle:

$P(\varphi(X)) = p) \rightarrow P(\varphi(c)) = p$

where φ is a formula with free logical variable X, and c is a constant instantiating X.

Halpern, J.Y. (1990), 'An analysis of first-order logics of probability', *Artificial Intelligence 46(3), 311--350.*

Examples of the Instantiation Principle

First-Order	**Ground Instance**
90% of birds fly.	The probability that Tweety flies is 90%.
P(Flies(B)) = 90%	P(Flies(tweety)) = 90%
0% of planes have crashed because of a turbulence.	The probability that Flight 3202 to Brussels crashes because of a turbulence is 0%.
P(Turbulence_Crash(Plane)) = 0%.	P(Turbulence_Crash(3202)) = 0%.
x% of Brusselians speak Dutch.	Given that Marc lives in Brussels, the probability that he speaks Dutch is x%.
P(SpeaksOnly(Person,dutch) | FromBrussels(Person)) = 5%.	P(SpeaksOnly(marc,dutch) | FromBrussels(marc)) = 5%.

Previous SRL Work: Different Models for Different Probabilities

Many Model Types:
Parametrized Bayes Nets
Probabilistic Relational Models,
Markov Logic Networks,
Bayes Logic Programs,
Logical Bayesian Networks, …

Relational Probability

Statistical-Relational Models
(Lise Getoor, Taskar, Koller 2001)

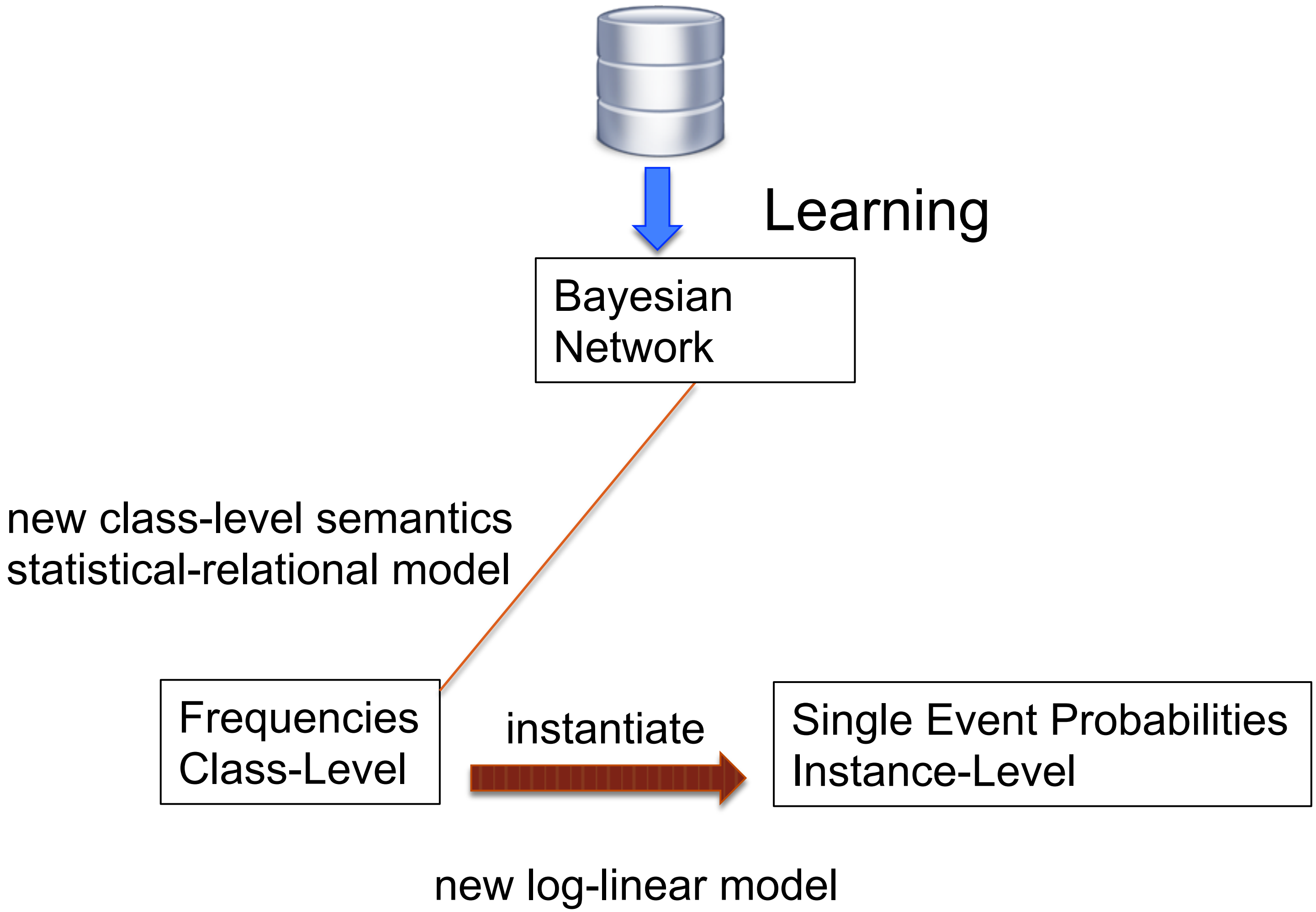
Model for Frequencies
Class-Level Probabilities

Model for Single Event Probabilities
Instance-Level Probabilities

Getoor, L. (2001), 'Learning Statistical Models From Relational Data', PhD thesis, Department of Computer Science, Stanford University.

Getoor, L.; Taskar, B. & Koller, D. (2001), 'Selectivity estimation using probabilistic models', *ACM SIGMOD Record 30(2), 461—472.*

Today: Unified Approach



Relational Frequencies

Joe Halpern

Fahim Bacchus

Applications of Relational Frequency Modelling

- First-order rule learning
  (e.g., "women users like movies with women actors").
- Strategic Planning
  (e.g., "increase SAT requirements to decrease student attrition").
- Query Optimization (Getoor, Taskar, Koller 2001).
  Class-level queries support selectivity estimation ➔ optimal evaluation order for SQL query .

Getoor, Lise, Taskar, Benjamin, and Koller, Daphne. Selectivity estimation using probabilistic models. ACM SIGMOD Record, 30(2):461–472, 2001.

Relational Frequencies

- Database probability of a first-order formula = number of satisfying instantiations/ number of possible instantiations.
- Examples:
  - $P_D(gender(Actor) = W) = 2/4.$
  - $P_D(gender(Actor) = W, ActsIn(Actor, Movie) = T) = 2/8.$

The Grounding Table

- P(gender(Actor) = W, ActsIn(Actor,Movie) = T, genre(Movie) = Action) = 2/8
- frequency = #of rows where the formula is true/# of all rows

Logical Variable

- Single data table that correctly represents relational frequencies.
- Schulte 2011, Riedel, Yao, McCallum (2013)

Actor	Movie	gender(Actor)	ActsIn(Actor,Movie)	genre(Movie)
Brad_Pitt	Fargo	M	F	Action
Brad_Pitt	Kill_Bill	M	F	Action
Lucy_Liu	Fargo	W	F	Action
Lucy_Liu	Kill_Bill	W	T	Action
Steve_Buscemi	Fargo	M	T	Action
Steve_Buscemi	Kill_Bill	M	F	Action
Uma_Thurman	Fargo	W	F	Action
Uma_Thurman	Kill_Bill	W	T	Action

Random Selection Semantics (Terms)

~~Logical~~ Random Variable

P(Movie = Fargo) = 1/2

Prob	Actor	Movie	gender(Actor)	ActsIn(Actor,Movie)	genre(Movie)
1/8	Brad_Pitt	Fargo	M	F	Action
1/8	Brad_Pitt	Kill_Bill	M	F	Action
1/8	Lucy_Liu	Fargo	W	F	Action
1/8	Lucy_Liu	Kill_Bill	W	T	Action
1/8	Steve_Buscemi	Fargo	M	T	Action
1/8	Steve_Buscemi	Kill_Bill	M	F	Action
1/8	Uma_Thurman	Fargo	W	F	Action
1/8	Uma_Thurman	Kill_Bill	W	T	Action

Halpern, J.Y. (1990), 'An analysis of first-order logics of probability', *Artificial Intelligence 46(3), 311--350.*

Random Selection Semantics

Population

Actors

Movies

Population variables

Actor
Random Selection from Actors.
P(*Actor* = *brad_pitt*) = 1/4.

Movie
Random Selection from Movies.
P(*Movie* = *fargo*) = ½.

First-Order Random Variables

gender(Actor)
Gender of selected actor.
P(*gender(Actor)* = *W*) = ½.

ActsIn(Actor, Movie) =
T if selected actor appears in selected movie, F otherwise
P(*ActsIn(Actor, Movie)* = T) = 3/8

genre(Movie)
Genre of selected movie.
P(*genre(Movie)*=*Action*) = 1.

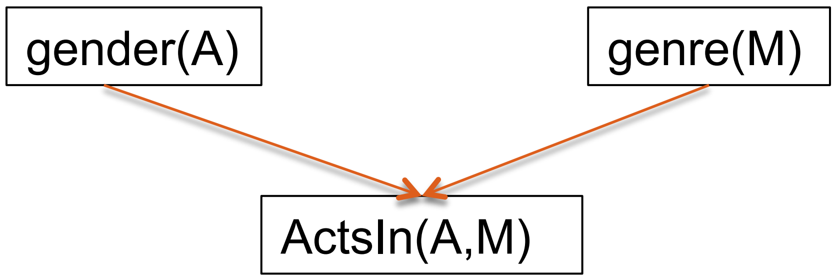
Bayesian Networks for Relational Statistics

Statistical-Relational Models (SRMs)
Random Selection Semantics for Bayesian Networks

Bayesian Network Model for Relational Frequencies

Bayesian networks for relational data

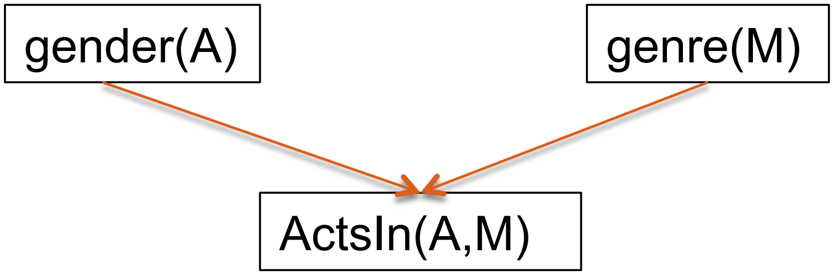
- A first-order Bayesian network is a Bayesian network whose nodes are first-order terms (Wang et al. 2008).
- aka parametrized Bayesian network (Poole 2003, Kimmig et al. 2014).



Wang, D. Z.; Michelakis, E.; Garofalakis, M. & Hellerstein, J. M. (2008), BayesStore: managing large, uncertain data repositories with probabilistic graphical models, in ,VLDB Endowment, , pp. 340--351.
Kimmig, A.; Mihalkova, L. & Getoor, L. (2014), 'Lifted graphical models: a survey', *Machine Learning, 1--45*.

Random Selection Semantics for First-Order Bayesian Networks

- P(gender(Actor) = W, ActsIn(Actor,Movie) = T, genre(Movie) = Action) = 2/8

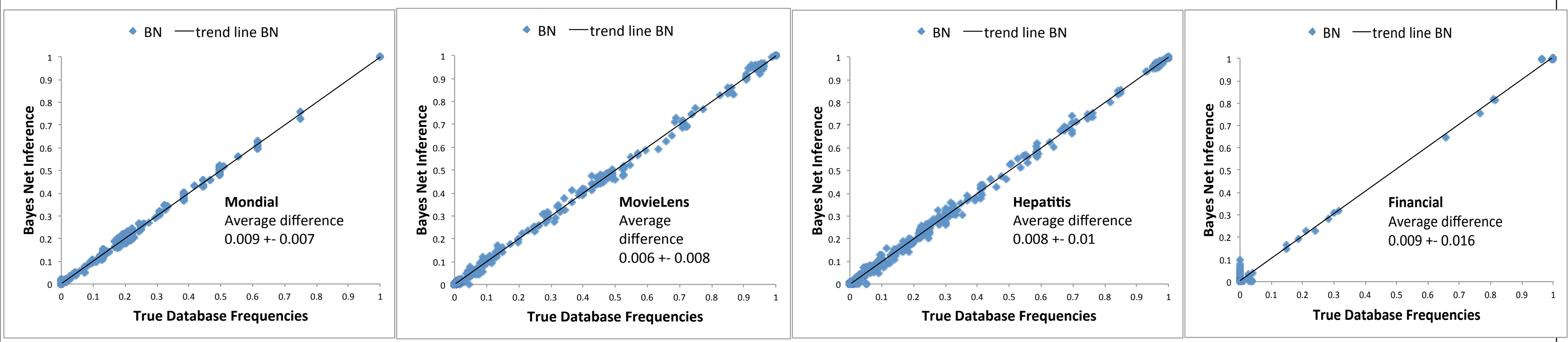


"if we randomly select an actor and a movie, the probability is 2/8 that the actor appears in the movie, the actor is a woman, and the movie is an action movie"

- Demo.

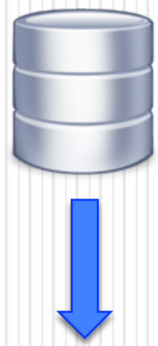
Bayesian Networks are Excellent Estimators of Relational Frequencies

- Queries Randomly Generated.
- Example: $P(gender(A) = W \mid ActsIn(A,M) = true, genre(M) = drama)$?
- Learn and test on entire database as in Getoor et al. 2001.



Schulte, O.; Khosravi, H.; Kirkpatrick, A.; Gao, T. & Zhu, Y. (2014), 'Modelling Relational Statistics With Bayes Nets', Machine Learning 94, 105-125.

How to upgrade propositional Bayesian network learners to first-order logic



Structure and Parameter Learning

Bayesian Network Model for Relational Frequencies

How to upgrade single-table Bayesian network learners to multi-relational data

Follow Luc's advice!

1. Search the space of *functors/predicates*, not literals. (Kersting and de Raedt 2007).
2. Organize search using the specialization/refinement *lattice* of models (Laer and de Raedt 2001).
3. Follow *the generalization principle* (Knobbe 2006): When we apply a relational model to a single i.i.d. data table, it should give the same result as the propositional model.

Kersting, K. & de Raedt, L. (2007), Bayesian Logic Programming: Theory and Tool'Introduction to Statistical Relational Learning', MIT Press, , pp. 291-318.
Laer, W. V. & de Raedt, L. (2001), How to upgrade propositional learners to first-order logic: A case study'Relational Data Mining', Springer Verlag, .
Knobbe, A. J. (2006), *Multi-relational data mining, Vol. 145, Ios Press.* \

General Structure Learning Schema

After Kimmig et al. 2014.

1. Initialize graph G := empty.
2. While not converged do
   a. Generate candidate graphs. ← lattice search
   b. For each candidate graph C, learn parameters θ_C that maximize score(C, θ, dataset). ← relational score
   c. G := argmax_C score(C, θ_C,dataset).
3. check convergence criterion.

Scoring Bayesian Networks for Relational Data

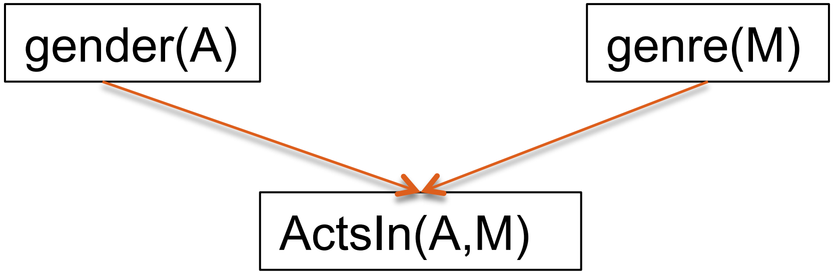
The Random Selection Pseudo Likelihood Function

1. Randomly select a grounding for **all** first-order variables in the first-order Bayesian network.
2. Compute the log-likelihood for the attributes of the selected grounding.
3. Pseudo log-likelihood = expected log-likelihood for a random grounding.

Generalizes i.i.d. log-likelihood.

Schulte, O. (2011), A tractable pseudo-likelihood function for Bayes Nets applied to relational data, *in 'SIAM SDM', pp. 462-473.*

Example

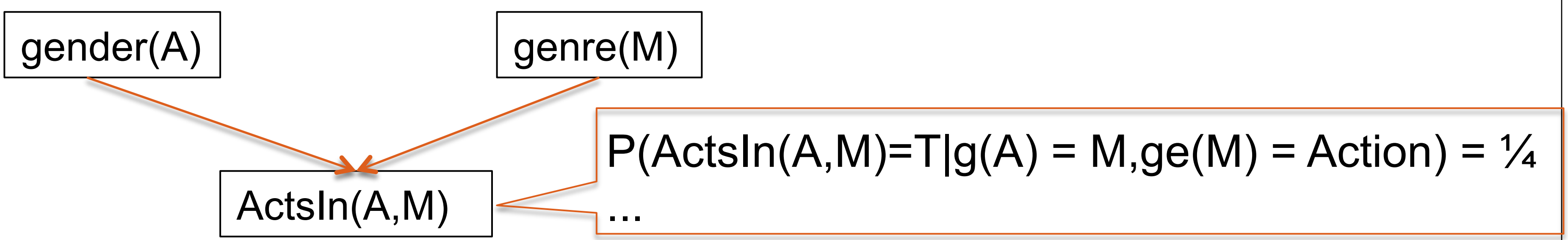


Prob	A	M	gender(A)	ActsIn(A,M)	genre(M)	P_B	$\ln(P_B)$
1/8	Brad_Pitt	Fargo	M	F	Action	3/8	-0.98
1/8	Brad_Pitt	Kill_Bill	M	F	Action	3/8	-0.98
1/8	Lucy_Liu	Fargo	W	F	Action	2/8	-1.39
1/8	Lucy_Liu	Kill_Bill	W	T	Action	2/8	-1.39
1/8	Steve_Buscemi	Fargo	M	T	Action	1/8	-2.08
1/8	Steve_Buscemi	Kill_Bill	M	F	Action	3/8	-0.98
1/8	Uma_Thurman	Fargo	W	F	Action	2/8	-0.98
1/8	Uma_Thurman	Kill_Bill	W	T	Action	2/8	-1.39
						0.27 geo	-1.32 arith

Schulte, O. (2011), A tractable pseudo-likelihood function for Bayes Nets applied to relational data, *in 'SIAM SDM', pp. 462-473.*

Observed Frequencies Maximize Pseudo-Likelihood

Proposition The random selection pseudo log-likelihood is maximized by setting the Bayesian network parameters to the observed conditional frequencies.



Equivalent Closed-Form Computation for Pseudo-Likelihood

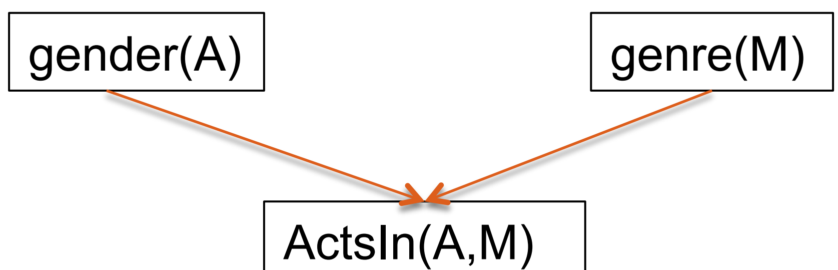
$$P(D;B) \propto \exp \sum_{\text{nodes } i} \sum_{\text{values } a} \sum_{\text{parent-state } j} p_D(a,j) \times \ln(P_B(a \mid j))$$

relational joint frequency of child node value and parent state

e.g. P_D(ActsIn,M,Action) = 1/8

Parameter of Bayes net

e.g. P(ActsIn|M, Action) = ¼



Parameter Learning

Maximum Likelihood Estimation

Computing Parameter Estimates

- Need to compute a **contingency table** with grounding counts.
- Well researched for all true relationships.

  SQL Count(*)

  Virtual Join

  Partition Function Reduction

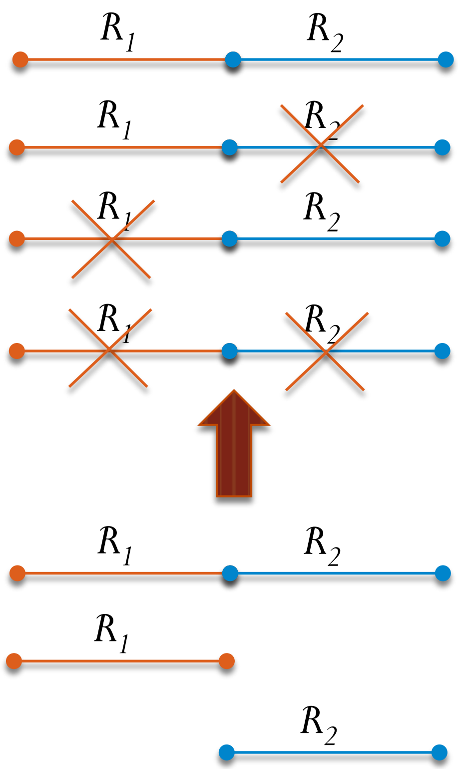
g(A)	Acts(A,M)	ge(M)	count
M	F	Action	3
M	F	Action	3
W	F	Action	2
W	T	Action	2
M	T	Action	1
M	F	Action	3
W	F	Action	2
W	T	Action	2

Yin, X.; Han, J.; Yang, J. & Yu, P. S. (2004), CrossMine: Efficient Classification Across Multiple Database Relations, *in 'ICDE'*.

Venugopal, D.; Sarkhel, S. & Gogate, V. (2015), Just Count the Satisfied Groundings: Scalable Local-Search and Sampling Based Inference in MLNs, *in AAAI, 2015, pp. 3606--3612*.

The Möbius Extension Theorem for negated relations

For two link types



nothing

Count(*)
Count(*)
Count(*)
Count(*)

Joint probabilities

Count(*)
Count(*)
Count(*)
Count(*)

Möbius Parameters

The Fast Möbius Transform

$ActsIn(A,M) = R_1$

$HasRated(U,M) = R_2$

Initial table with no false relationships

R_1	R_2	J.P.
T	T	0.2
*	T	0.3
T	*	0.4
*	*	1

J.P. = joint probability

R_1	R_2	J.P.
T	T	0.2
F	T	0.1
T	*	0.4
F	*	0.6

table with joint probabilities

R_1	R_2	J.P.
T	T	0.2
F	T	0.1
T	F	0.2
F	F	0.5

Kennes, R. & Smets, P. (1990), Computational aspects of the Möbius transformation, *in 'UAI', pp. 401-416.*

Using Presence and Absence of Relationships

- Fast Möbius Transform → almost free computationally!
- Allows us to find *correlations with relationships.*
  - *e.g.* users who search for an item on-line also watch a video about it.
- Relationship variables selected by standard data mining approaches (Qian et al 2014).
  - Interesting Association Rules.
  - Feature Selection Metrics.

Qian, Z.; Schulte, O. & Sun, Y. (2014), Computing Multi-Relational Sufficient Statistics for Large Databases, *in 'Computational Intelligence and Knowledge Management (CIKM)', pp. 1249--1258.*

Parameter Learning Time

- Fast Inverse Möbius transform (I**MT**) vs.
- Constructing **complement** tables using SQL.
- Times are in seconds.

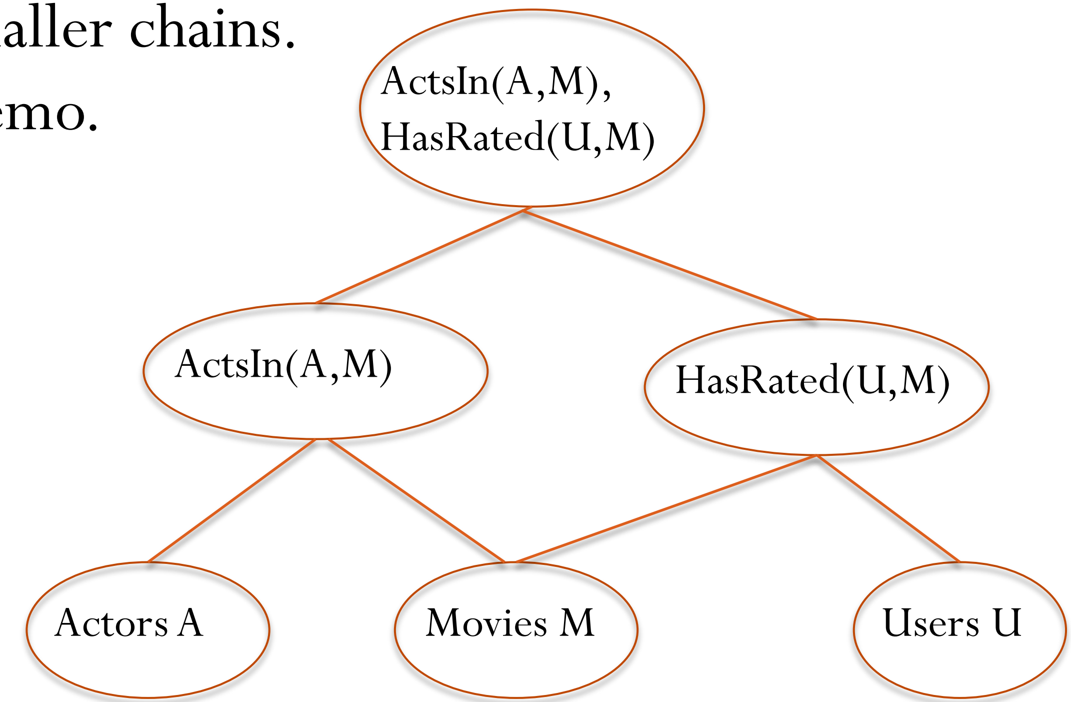
Database	Parameters	#tuples	Complement	IMT	Ratio
Mondial	1618	814	157	7	22
Hepatitis	1987	12,447	18,246	77	237
Financial	10926	17,912	228,114	14,821	15
MovieLens	326	82,623	2,070	50	41

Möbius transform is much faster, 15-200 times.

Structure Learning: Lattice Search

Learning a Bayesian Multi-Net

- Score = pseudo log-likelihood – parameter penalty
- Learn a Bayesian network for each relationship chain.
- Nodes and edges are propagated from shorter chains to smaller chains.
- Demo.



Khosravi, H.; Schulte, O.; Man, T.; Xu, X. & Bina, B. (2010), Structure Learning for Markov Logic Networks with Many Descriptive Attributes, *in 'AAAI', pp. 487-493.*

Comparison With Luc's advice

+ Search in space of functions/predicates.
+ Generalizes i.i.d. BN learning.
+ **Decompose specialization lattice into sublattices.**
  - Each sublattice corresponds to relational path.
  - Lattices at the same level can be analyzed separately ➔ distributed processing.
  - Results from lower levels are propagated to higher levels ➔dynamic programming style.

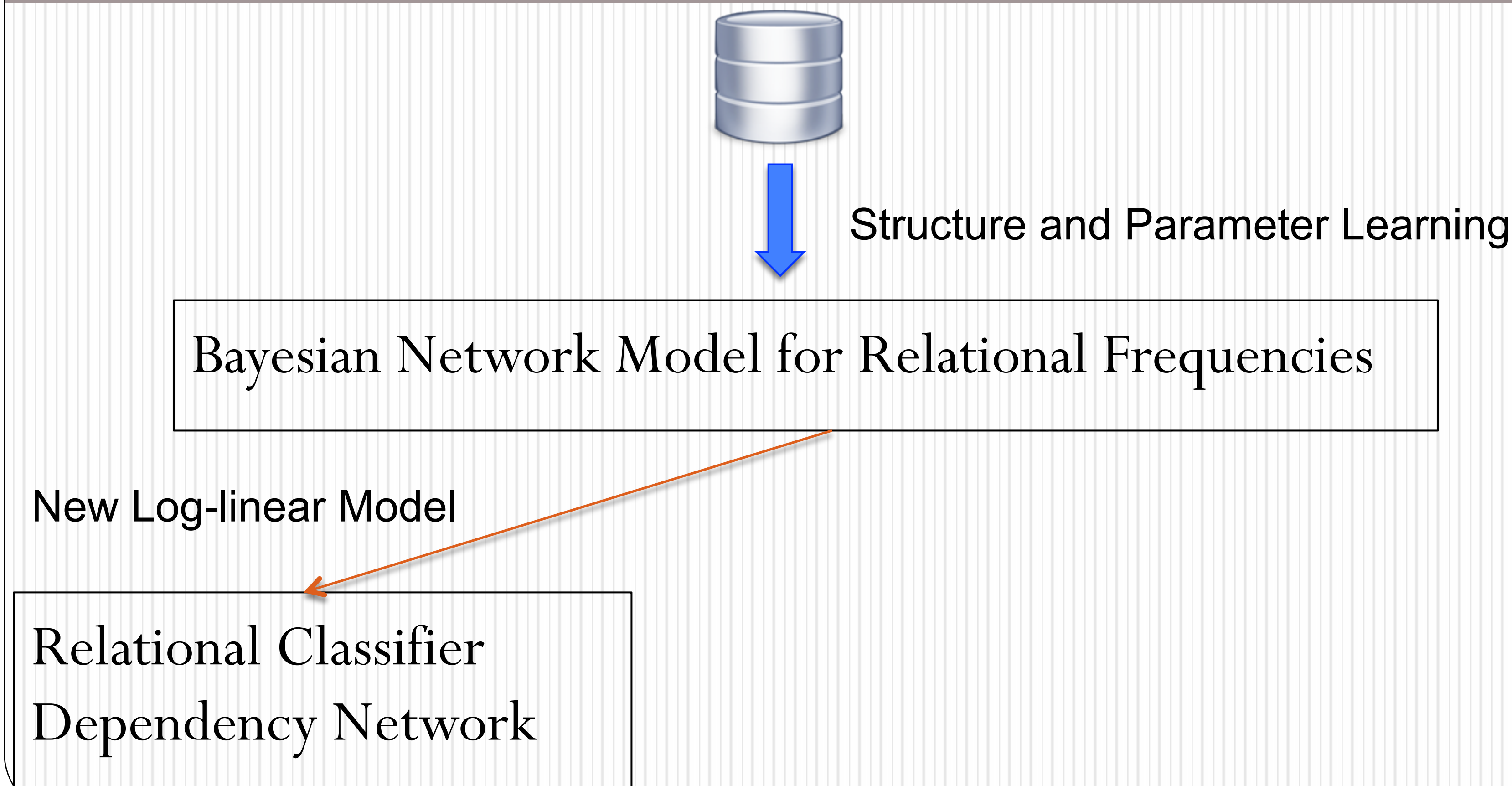
– First-order variables only ➔ half of theta-subsumption.

Fast Structure Learning

Dataset	# Predicates	# tuples	RDN_Boost	MLN_Boost	Lattice
UW	14	612	15±0.3	19±0.7	**1±0.0**
Mondial	18	870	27±0.9	42±1.0	102±6.9
Hepatitis	19	11,316	251±5.3	230±2.0	286±2.9
Mutagenesis	11	24,326	118±6.3	49±1.3	**1±0.0**
MovieLens(0.1M)	7	83,402	44±4.5 min	31±1.87 min	**1±0.0**
MovieLens(1M)	7	1,010,051	>24 hours	>24 hours	**10±0.1**
Imdb(1.5M)	17	1,538,400	>24 hours	>24 hours	**549**

- Standard deviations are shown for cross-validation.
- Units are *seconds/predicate or function*

From Relational Statistics to Degrees of Belief



Predicting Ground Facts

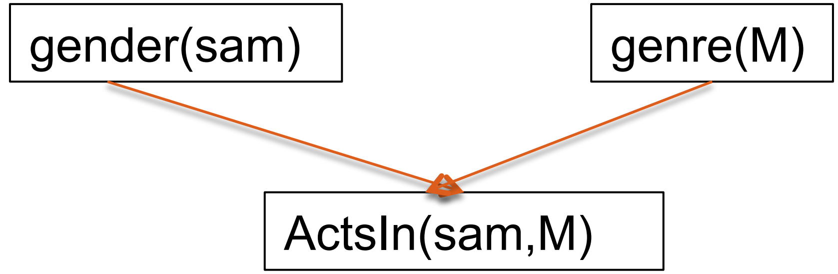
- Many relational models aim to predict specific facts, e.g.
  - Will KAA Gent win the Belgian football league 2015-2016?
  - Is Spectre likely to do well at the box office?
- The problem: relational data feature *multiple instantiations* of the same pattern.
  - E.g. 1,000 men give Spectre a high rating, 1,200 women give spectre a high rating.
- Halpern's project: from relational frequencies, derive a probability distribution over possible worlds (models, databases). (Halpern 1990, 1992, 2006).

Bacchus, F.; Grove, A. J.; Koller, D. & Halpern, J.Y. (1992), From Statistics to Beliefs, *in 'AAAI', pp. 602-608*.
Halpern, J.Y. (2006), From statistical knowledge bases to degrees of belief: an overview, *in 'PODS', ACM, , pp. 110—113*.

Bayesian Network Relational Classification

- Classification problem: Define $P(Y^*=y|X^*=x)$ for ground term Y^* given values for all other terms X^*.
- Strictly easier than defining joint probability $P(Y^*=y,X^*=x)$.
- Basic idea: score labels by comparing *pseudo-likelihood* $P(Y^*=0,X^*=x)$ to $P(Y^*=1,X^*=x)$.
- Restrict pseudo-likelihood to relevant groundings that involve the target term.
- Generalizes propositional Bayesian Network classification formula.

Example



-1.32 > -1.36: predict sam is a woman.

A	M	gender(A)	ActsIn(A,M)	genre(M)	P_B	$\ln(P_B)$
sam	Fargo	W	F	Action	2/8	-0.98
sam	Kill_Bill	W	T	Action	2/8	-1.39
						-1.32 arith

A	M	gender(A)	ActsIn(A,M)	genre(M)	P_B	$\ln(P_B)$
sam	Fargo	M	F	Action	3/8	-0.98
sam	Kill_Bill	M	T	Action	1/8	-1.39
						-1.36 arith

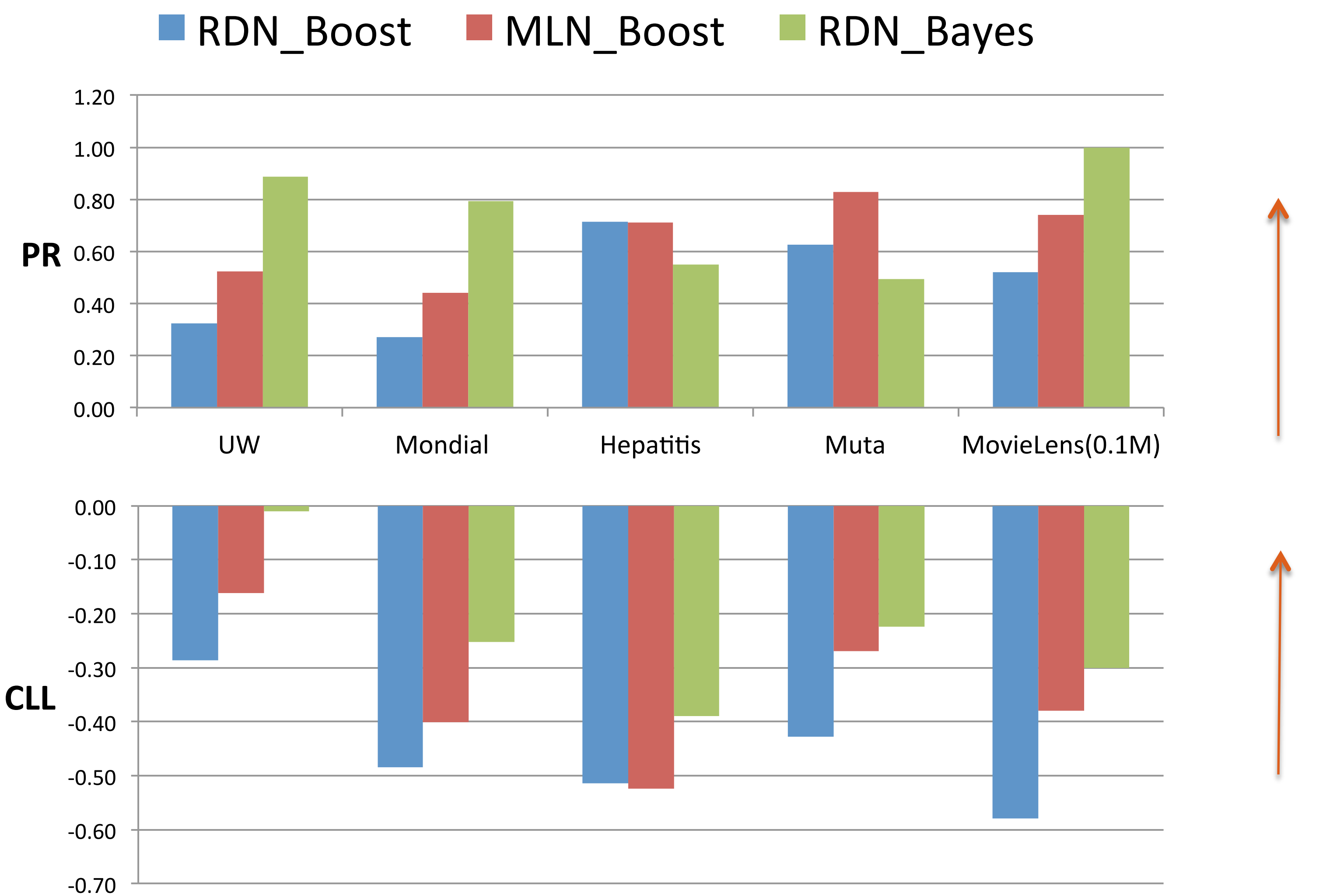
Schulte et al. (2014) 'Fast Learning of Relational Dependency Networks', in ILP 2014

Dependency Networks

- aka Markov blanket networks (Hoffmann and Tresp 1998).
- Defined by a local conditional distribution for each random variable Y*: P(Y*=y|X*=x).
- We just showed Bayesian network ➔ dependency network.
- Can compare with other dependency network learning.
- Recall that this is very fast (<12 min on 1M tuples).
- Finds complex dependencies
  - e.g. gender(User) correlates with gender(Actor) in movies they have rated.

Hofmann, R. & Tresp, V. (1998), Nonlinear Markov networks for continuous variables, *in 'Advances in Neural Information Processing Systems', pp. 521--527.*
Heckerman, D.; Chickering, D. M.; Meek, C.; Roundthwaite, R.; Kadie, C. & Kaelbling, P. (2000), 'Dependency Networks for Inference, Collaborative Filtering, and Data Visualization', *JMLR* ***1, 49—75.***

Accuracy Comparison



- Leave-one-out over all unary functors.
- PR = area under precision-recall curve.
  CLL: conditional log-likelihood

Model-Based Unsupervised Relational Outlier Detection

Model-Based Outlier Detection for I.I.D. data

ID	Attribute1	Attribute2	Attribute3
1	W	rich	10
2	….	…	…

Learning

Attribute1

Attribute2

Attribute3

ID	Att1	Att2	Att3
1000	M	rich	8

Likelihood of potential outlier
low likelihood → outlier

Cansado, A. & Soto, A. (2008), 'Unsupervised anomaly detection in large databases using Bayesian networks', *Applied Artifical Intelligence 22(4), 309--330.*

Model-Based Outlier Detection for Relational Data

Population Database
e.g. Premier League games

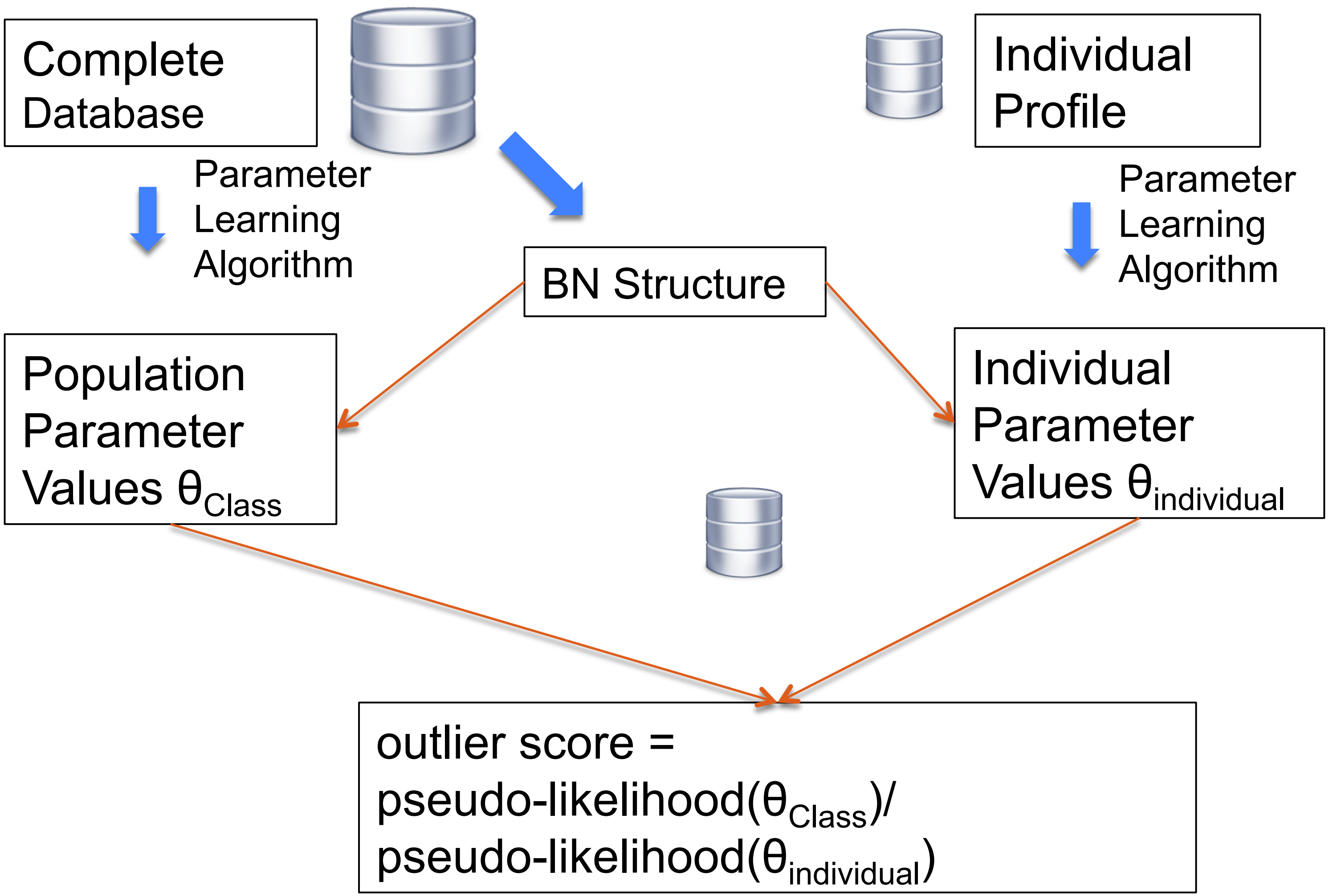
First-Order Bayesian network

Individual Profile
e.g. MU games

Pseudo likelihood of individual database
low likelihood → outlier

"Model-based Outlier Detection for Object-Relational Data". Riahi and Schulte (2015). IEEE SSCI.
Maervoet, J.; Vens, C.; Vanden Berghe, G.; Blockeel, H. & De Causmaecker, P. (2012), 'Outlier Detection in Relational Data: A Case Study in Geographical Information Systems', *Expert Systems With Applications 39(5), 4718—4728.*

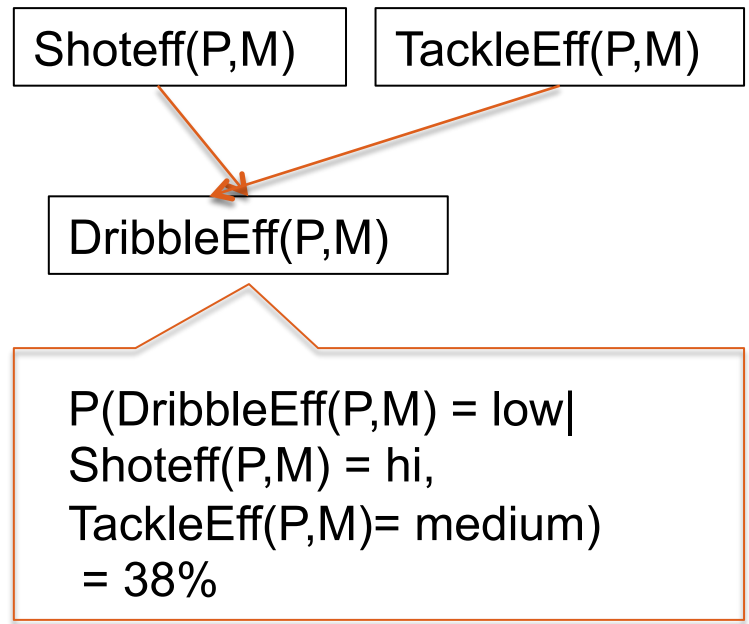
Likelihood Ratio Variant



Example Terms in Outlier Metrics

Striker Edin Dzeko.

- Pseudo log-likelihood:
  - Shoteff(dzeko,M) = hi, TackleEff(dzeko,M)= medium ➔ DribbleEff(dzeko,M) = low. Support=26% Confidence = 50%.
- Pseudo log-likelihood 26% x ln(38%).
- Pseudo Log-likelihood ratio 26% x (ln(38%)-ln(50%)).



Novak, P. K.; Lavrac, N. & Webb, G. I. (2009), 'Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining', *The Journal of Machine Learning Research* *10, 377--403.*

Interpretable (and Accurate)

Top Outliers from Selected Normal Classes:

- Strikers
- Midfielders
- Drama

Strikers (Normal) vs. Goalies (Outlier)							
PlayerName	Position	ELD Rank	ELD Max Node	ELD Node Score	FD Max feature Value	Object Probability	Class Probability
Edin Dzeko	Striker	1	DribbleEfficiency	83.84	DE=low	0.16	0.5
Paul Robinson	Goalie	2	SavesMade	49.4	SM=Medium	0.3	0.04
Michel Vorm	Goalie	3	SavesMade	85.9	SM=Medium	0.37	0.04
Midfielders (Normal) vs. Strikers (Outlier)							
PlayerName	Position	ELD Rank	ELD Max Node	ELD Node Score	FD Max feature Value	Object Probability	Class Probability
Robin Van Persie	Striker	1	ShotsOnTarget	153.18	ST=high	0.34	0.03
Wayne Rooney	Striker	2	ShotsOnTarget	113.14	ST=high	0.26	0.03
Scott Sinclair	Midfielder	6	DribbleEfficiency	71.9	DE=high	0.5	0.3
Drama (Normal) vs. Comedy (Outlier)							
MovieTitle	Genre	ELD Rank	ELD Max Node	ELD Node Score	FD Max feature Value	Object Probability	Class Probability
Brave Heart	Drama	1	ActorQuality	89995.4	a_quality=4	0.93	0.42
Austin Powers	Comedy	2	Cast_Position	61021.28	Cast_Num=3	0.78	0.49
Blue Brothers	Comedy	3	Cast_Position	24432.21	Cast_num=3	0.88	0.49

Riahi, S. and Schulte, O. (2015). 'Model-based Outlier Detection for Object-Relational Data' IEEE Symposium Series on Computing Intelligence. Forthcoming.

61

Summary, Review, Open Problems

Random Selection Semantics for First-Order Logic

- First-order variables and first-order terms are viewed as random variables.
- Associates relational frequency with each first-order formula.

Joe Halpern

Fahim Bacchus

Applying random selection to log-linear models

Graphical model ≈

Propositional log-linear model

$s = w_i \, x_i$

E(s) from random instantiation

relational log-linear model

$s = w_i \, f_i$

feature functions = frequency

- Combines multiple instantiations of the same feature.
- Defines relational pseudo log-likelihood score Bayes net.

Log-linear Models With Proportions

- Frequencies are on the same scale [0,1]: addresses "ill-conditioning" (Lowd and Domingos 2007).
- Surprisingly expressive: can "simulate" combining rules (Kazemi et al. 2014).
- Also effective for dependency networks with hybrid data types (Ravkic, Ramon, Davis 2015).
- Random selection semantics provides a theoretical foundation.

Lowd, D. & Domingos, P. (2007), Efficient Weight Learning for Markov Logic Networks, in 'PKDD', pp. 200—211.
Kazemi, S. M.; Buchman, D.; Kersting, K.; Natarajan, S. & Poole, D. (2014), Relational Logistic Regression, *in 'Principles of Knowledge Representation and Reasoning:, KR 2014.*
Ravkic, I.; Ramon, J. & Davis, J. (2015), 'Learning relational dependency networks in hybrid domains', *Machine Learning.*

Learning results

- Random selection *pseudo-likelihood score* for Bayesian networks.
- Closed-form *parameter estimation*.
  - Fast Möbius transform for computing parameters with *negated relationships.*
- Structure Learning: Decompose the *lattice of relationship chains.*
- Fast learning, competitive accuracy for:
  - modeling relational frequencies.
  - relational dependency networks.
  - relational outlier detection.

Open Problems

- Learning with constants (theta-subsumption).
- Generalize model scores like AIC, BIC with positive and negative relationships.
  - need to scale penalty terms as well as feature counts.

Thank you!

- Any questions?