Learning relational theories

Charles Kemp
Department of Brain and Cognitive Sciences
MIT

July 16, IPAM Summer School

Acknowledgments

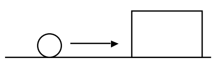
Noah Goodman
Tom Griffiths
Sourabh Niyogi
Josh Tenenbaum

Relational knowledge

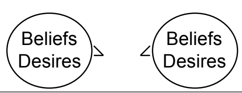
- Kinship

  $\forall x\, y\, z\; \text{Uncle}(x, z) \leftarrow \text{Brother}(x, y) \wedge \text{Parent}(y, z)$

- Folk physics



- Folk psychology



What is a theory?

"A theory is characterized by the phenomena in its domain, its **laws** and other explanatory mechanisms, and the **concepts** that articulate the laws and the representations of the phenomena."

(Carey 85)

Theories

CONCEPTS	**CONCEPTS**
force mass acceleration	life death food growth ⋮
RELATIONSHIPS	**RELATIONSHIPS**
F = ma	life *ends in* death food *sustains* life food *produces* growth ⋮

(Carey 85)

How to learn a T

- Search for T that maximizes

  $$P(T|\text{Data}) \propto P(\text{Data}|T)P(T)$$

- Prerequisites
  - Decide how Ts are represented.
  - Put a prior over a hypothesis space of Ts.
  - Decide how observable data are generated from an underlying T.

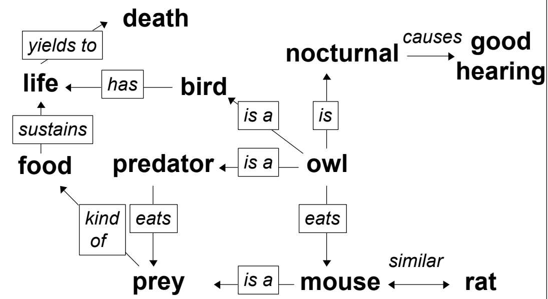
How to learn ~~a T~~ anything

- Search for T that maximizes

$$P(T|\text{Data}) \propto P(\text{Data}|T)P(T)$$

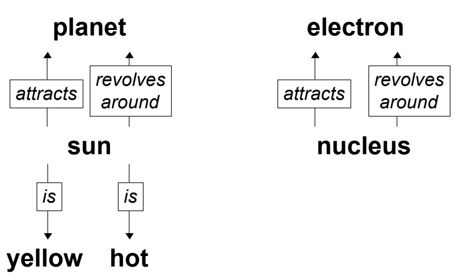
- Prerequisites
  - Decide how Ts are represented.
  - Put a prior over a hypothesis space of Ts.
  - Decide how observable data are generated from an underlying T.

Semantic knowledge



(Rumelhart, Norman & Lindsay 72; Anderson & Bower 73; …)

Analogy



(Genter, Holyoak, Goldstone, …)

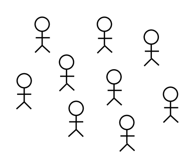
Outline

- Learning probabilistic theories
  - Computational models
  - Behavioral data (Tenenbaum and Niyogi)
- Learning deterministic theories
  - Computational models
  - Behavioral data

Learning probabilistic theories

- PRMs (FGKP models)
  - Getoor et al, 2001
- BLOG programs
  - Brian this morning
- Markov logic networks
  - Kok and Domingos, 2005
- Stochastic logic programs
  - Muggleton, 2000
- Stochastic blockmodels
  - Wang and Wong, 1987

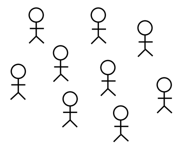
Learning a social theory



Learning a social theory

CONCEPTS

Professors
Graduate students
Undergraduates

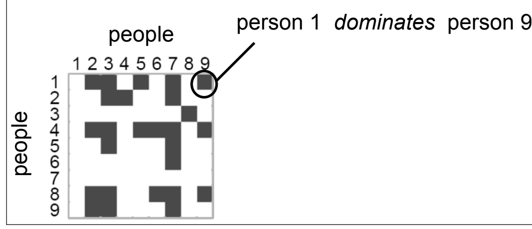


RELATIONSHIPS

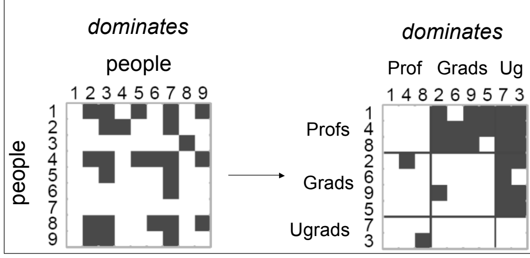
Professors *dominate* Graduate students
Undergraduates

Graduate students *dominate* Undergraduates

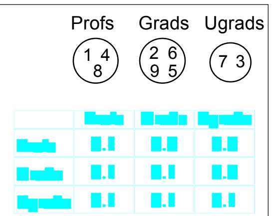
Learning a social theory

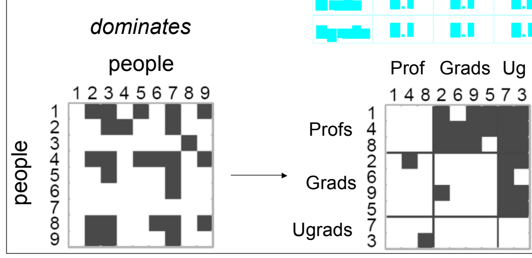


Learning a social theory

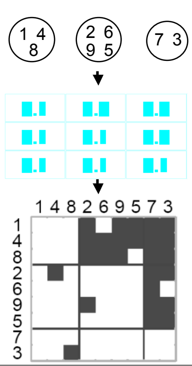


Learning a social theory





The Infinite Relational Model (IRM)

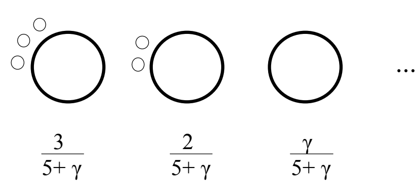


z $\quad z \mid \gamma \sim \mathrm{CRP}(\gamma)$

$\eta \quad \eta_{ab} \mid \alpha, \beta \sim \mathrm{Beta}(\alpha, \beta)$

R $\quad R_{ij} \mid z, \eta \sim \mathrm{Bernoulli}(\eta_{z_i z_j})$

P(z): The Chinese Restaurant Process

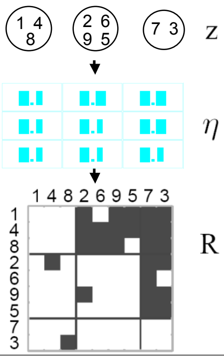


$$P(z_i = a \mid z_1, \ldots, z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & n_a > 0 \\ \frac{\gamma}{i-1+\gamma} & a \text{ is a new class} \end{cases}$$

where n_a is the number of entities in class *a*

(Aldous, 1985)

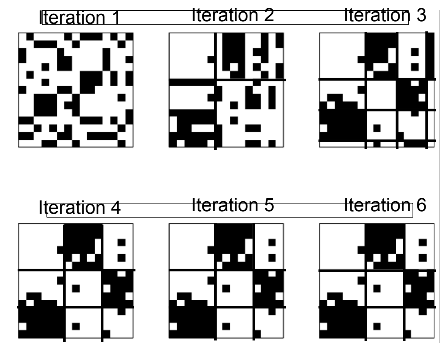
The Infinite Relational Model (IRM)



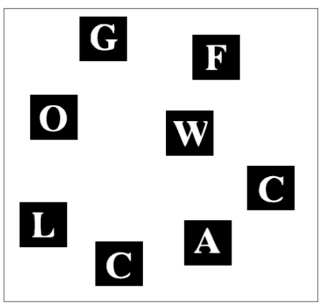
- Goal: find z and η that maximize

$$p(z,\eta|R) \propto P(R|z,\eta) \, p(\eta|z) \, P(z)$$

Searching for the best theory

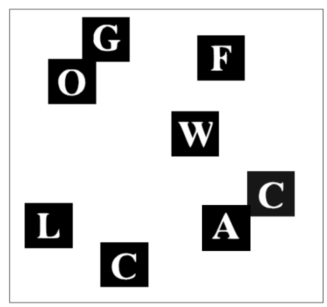


Theory-learning in the lab



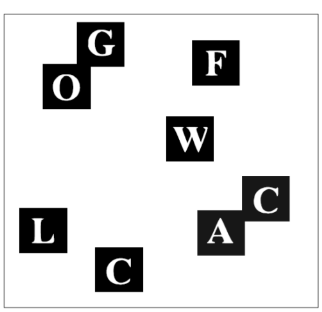
(Tenenbaum and Niyogi 03)

Theory-learning in the lab



(Tenenbaum and Niyogi 03)

Theory-learning in the lab

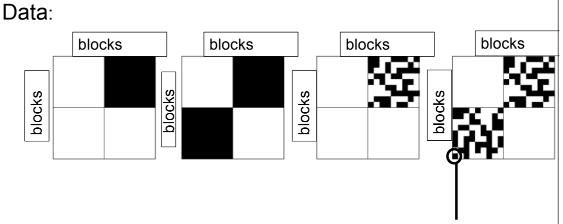


(Tenenbaum and Niyogi 03)

4 worlds

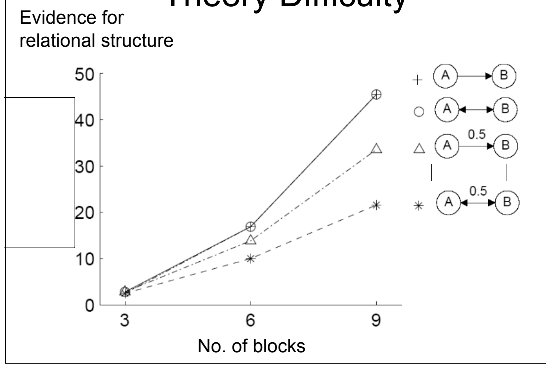
Theory:

Data:

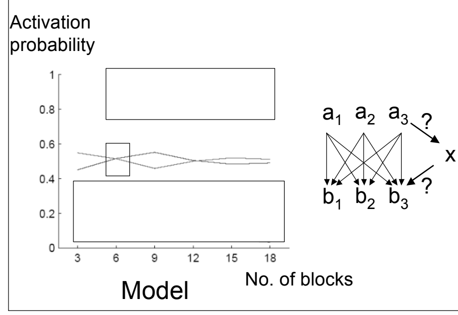


block 18 *activates* block 1

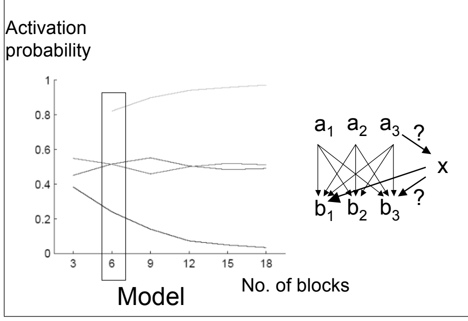
Theory Difficulty



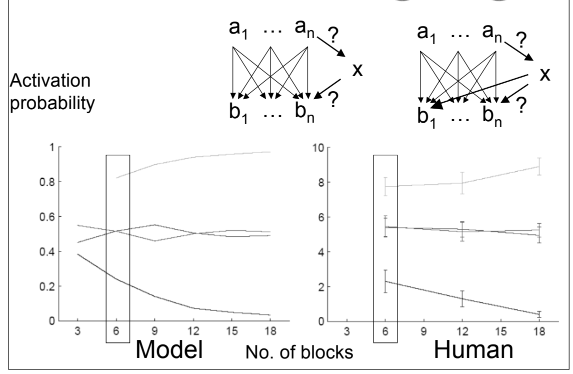
Learning curves for A → B



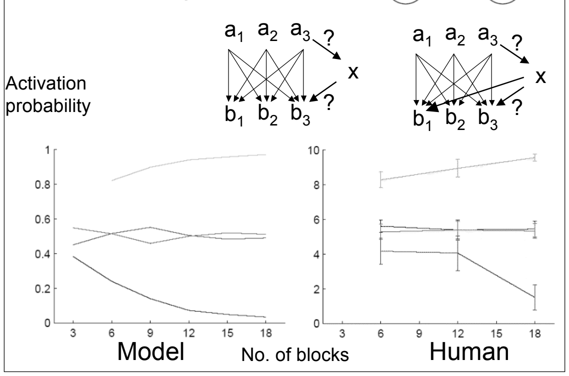
Learning curves for A → B



Learning curves for A → B



Learning curves for A ↔ B



More complex theories

Lead	Cancer	Coughing
Ammonia	Asthma	Weight loss
Chlorine	Pneumonia	Fatigue
Mercury	Flu	Fever

Cat
Dog
Human

More complex theories

CHEMICALS
Lead
Ammonia
Chlorine
Mercury

cause

DISEASES
Cancer
Asthma
Pneumonia
Flu

cause

SYMPTOMS
Coughing
Weight loss
Fatigue
Fever

affect

ANIMALS
Cat
Dog
Human

interact with

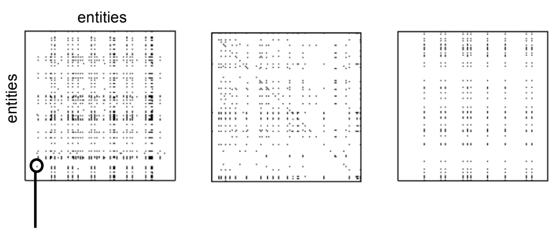
Discovering a medical theory

- 135 entities (bacterium, enzyme, cell function, mental process…)
- 49 relations (*causes*, *affects*, *disrupts*, ...)
- Data:
  - enzyme *affects* cell function
  - enzyme *disrupts* mental process …

(UMLS data, McCray,2003)

49 two dimensional relations

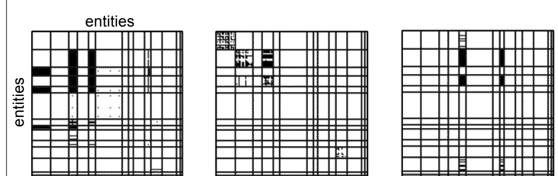
affects *interacts with* *causes*



enzyme *affects* cell function

49 two dimensional relations

affects *interacts with* *causes*



Concepts

1. Organisms
Alga
Amphibian
Animal
Archaeon
Bacterium
Bird

2. Chemicals
Amino acid
Carbohydrate
Chemical
Eicosanoid
Isotope
Steroid

3. Biological functions
Biological function
Cell function
Genetic function
Mental process
Molecular function
Physiological function

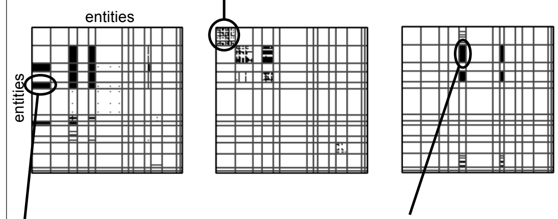
4. Bio-active substances
Antibiotic
Enzyme
Poisonous substance
Hormone
Pharmacologic substance
Vitamin

5. Diseases
Cell dysfunction
Disease
Mental dysfunction
Neoplastic process
Pathologic function
Experimental model of disease

Relationships between concepts

organisms *interact with* organisms

affects *interacts with* *causes*



diseases *affect* organisms

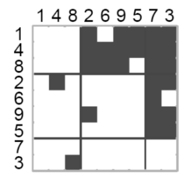
chemicals *cause* diseases

Outline

- Learning probabilistic theories
  – Computational models
  – Behavioral data (Tenenbaum and Niyogi)
- Learning deterministic theories
  – Computational models
  – Behavioral data

The Infinite Relational Model (IRM)

r(1,2). r(1,9). r(1,5) ...



a(1). a(6). a(4).
b(8). b(2). b(3). b(5).
c(9). c(7).

r(X,Y) ← a(X),a(Y). (0.0)
r(X,Y) ← a(X),b(Y). (0.9)
r(X,Y) ← a(X),c(Y). (1.0)
...

Deterministic theories

Sibling(victoria, arthur), Sibling(arthur,victoria),
Ancestor(chris,victoria), Ancestor(chris,colin),
Parent(chris,victoria), Parent(victoria,colin),
Uncle(arthur,colin), Brother(arthur,victoria) …

$\forall x\, y\, \text{Sibling}(x, y) \leftarrow \text{Sibling}(y, x)$

$\forall x\, y\, z\, \text{Ancestor}(x, z) \leftarrow \text{Ancestor}(x, y) \wedge \text{Ancestor}(y, z)$

$\forall x\, y\, \text{Ancestor}(x, y) \leftarrow \text{Parent}(x, y)$

$\forall x\, y\, z\, \text{Uncle}(x, z) \leftarrow \text{Brother}(x, y) \wedge \text{Parent}(y, z)$

Inductive Logic Programming

- A logic program is a set of *definite clauses*.

$\text{Sibling}(x, y) \leftarrow \text{Sibling}(y, x)$

$\text{Ancestor}(x, z) \leftarrow \text{Ancestor}(x, y), \text{Ancestor}(y, z)$

$\text{Ancestor}(x, y) \leftarrow \text{Parent}(x, y)$

$\text{Uncle}(x, z) \leftarrow \text{Brother}(x, y), \text{Parent}(y, z)$

- Definite clauses have one atom on the LHS, and all variables are universally quantified

(Muggleton, Quinlan, …)

A grammar-based prior

program	→	rule . program	(0.5)
program	→	Λ	(0.5)
rule	→	fact	(0.5)
rule	→	head :- body	(0.5)
head	→	atom	(1.0)
fact	→	atom	(1.0)
body	→	literal , body	(0.5)
body	→	literal	(0.5)
literal	→	atom	(0.5)
literal	→	not atom	(0.5)
atom	→	(see text)	

(Conklin and Witten, 1995)

First order theories

- Prior: $P(T)$
  – $P(T) = P(T|G)$ where G is a probabilistic grammar
- Likelihood: $P(\text{Data}|T)$
  – Assume the data are sampled at random from all facts that are true according to T
  (NB: only works for finite domains)

(Conklin and Witten, 1995)

How to learn a T

- Search for T that maximizes

$$P(T|\text{Data}) \propto P(\text{Data}|T)P(T)$$

- Prerequisites
  - Decide how Ts are represented.
  - Put a prior over a hypothesis space of Ts.
  - Decide how observable data are generated from an underlying T.

Searching the space of logic programs

- Top down: start with overly general rules and refine them
- Bottom-up: start with overly specific rules

$$P(x) \leftarrow A_1(x), A_2(x), A_3(x), A_4(x), B(x)$$
$$P(x) \leftarrow A_1(x), A_2(x), A_3(x), A_4(x), C(x)$$

$$\downarrow$$

$$A(x) \leftarrow A_1(x), A_2(x), A_3(x), A_4(x)$$
$$P(x) \leftarrow A(x), B(x)$$
$$P(x) \leftarrow A(x), C(x)$$

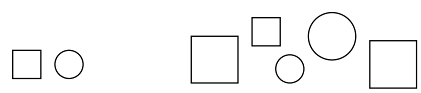
(Inter construction)

Theory-learning in the lab

- Which theories are hard or easy for people to learn?
- How do theories support inductive inferences?
- How are theories learned from positive examples?

Propositional theories

- Categorization and concept learning



Blickets Non-blickets

blicket(X) ← small(X), blue(X).

(Feldman 00, 06; Goodman et al 07; …)

Theory-learning in the lab

R(f,c) R(k,c) R(c,l) R(c,b)

R(f,k)

R(f,l) R(k,l) R(l,b)

R(l,h)

R(f,b) R(k,b) R(f,h) R(b,h)

R(k,h) R(c,h)

(cf Krueger 1979)

Theory-learning in the lab

f,c k,c c,l c,b

f,k

f,l k,l l,b

l,h

f,b k,b f,h b,h

k,h c,h

Theory-learning in the lab

1,2	1,3	1,4	1,5	1,6
	2,3	2,4	2,5	2,6
		3,4	3,5	3,6
			4,5	4,6
				5,6

Transitive: R(1,2). R(2,3). R(3,4). R(4,5). R(5,6).

R(X,Z) $\leftarrow$ R(X,Y), R(Y,Z).

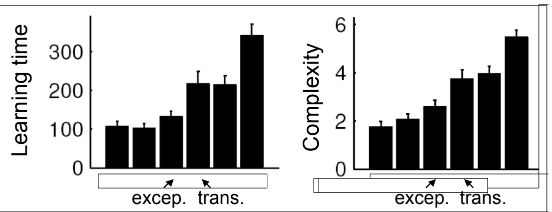
Theory-learning in the lab

1,1				
	2,6	3,6	4,6	5,6
1,7	2,7	3,7	4,7	5,7
1,8	2,8	3,8	4,8	5,8

Exception: T(6). T(7). T(8).

R(X,Y) $\leftarrow \overline{\text{T}}$(X), T(Y).

R(1,1). $\overline{\text{R}}$(1,6).



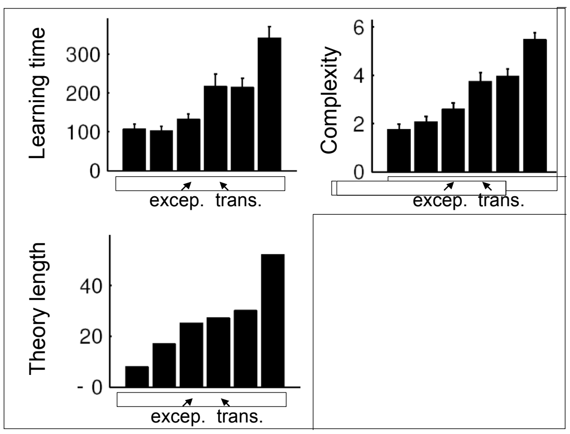
First order theories

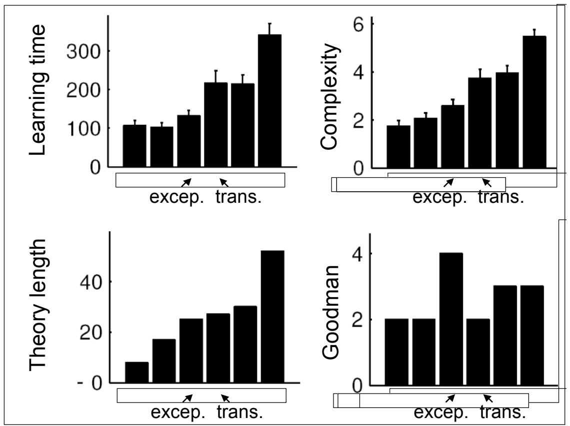
- Prior $P(T)$
  - $P(T) \propto 2^{-\text{length}(T)}$
- Likelihood $P(\text{Data}|T)$
  - Assume that the data include all facts that are true according to T

A length-based prior

1,1				
	2,6	3,6	4,6	5,6
1,7	2,7	3,7	4,7	5,7
1,8	2,8	3,8	4,8	5,8

Exception:	T(6). T(7). T(8).	9
	R(X,Y) $\leftarrow \overline{\text{T}}$(X), T(Y).	8
	R(1,1). $\overline{\text{R}}$(1,6).	8





Theories and induction

1,1

2,6 3,6 4,6 5,6 9,6

1,7 2,7 3,7 4,7 5,7 9,7

1,8 2,8 3,8 4,8 5,8 9,8

Exception: T(6). T(7). T(8).

R(X,Y) $\leftarrow$ $\overline{\text{T}}$(X), T(Y).

R(1,1). $\overline{\text{R}}$(1,6).

Theories and induction

1,1

2,6 3,6 4,6 5,6

1,7 2,7 3,7 4,7 5,7

1,8 2,8 3,8 4,8 5,8

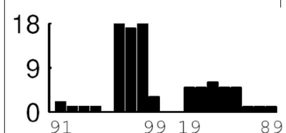
1,9 2,9 3,9 4,9 5,9

Exception: T(6). T(7). T(8). T(9).

R(X,Y) $\leftarrow$ $\overline{\text{T}}$(X), T(Y).

R(1,1). $\overline{\text{R}}$(1,6).

Theories and induction



People

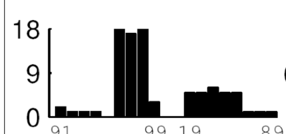
Predictions about new pairs

$$P(t_{\text{new}}|D) = \sum_T P(t_{\text{new}}|T)P(T|D)$$
$$= \sum_{T\,:\,t_{\text{new}} \in T} P(T|D)$$

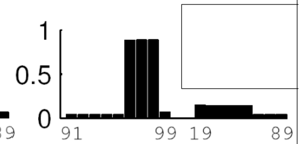
where D includes all training pairs

t_{new} is a pair including the new object

Theories and induction



People



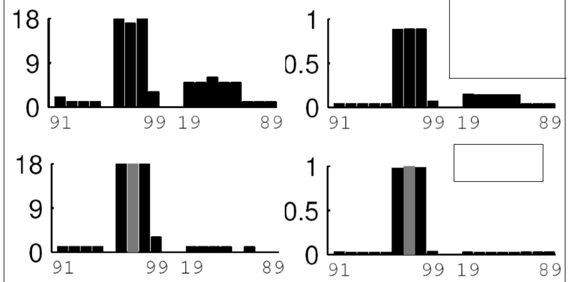
Model

Theories and induction

1,1
2,6 3,6 4,6 5,6 9,6
1,7 2,7 3,7 4,7 5,7 9,7
1,8 2,8 3,8 4,8 5,8 9,8

Exception: T(6). T(7). T(8).
R(X,Y) ← $\overline{\text{T}}$(X), T(Y).
R(1,1). $\overline{\text{R}}$(1,6).

Theories and induction



People Model

Learning from positive examples

1,1,1 2,2,2 2,2,1 3,3,1
3,3,3 4,4,4 4,4,1 5,5,1

Learning from positive examples

1,1,1 2,2,2 2,2,1 3,3,1
3,3,3 4,4,4 4,4,1 5,5,1

R(X,X,X). R(X,X,1).

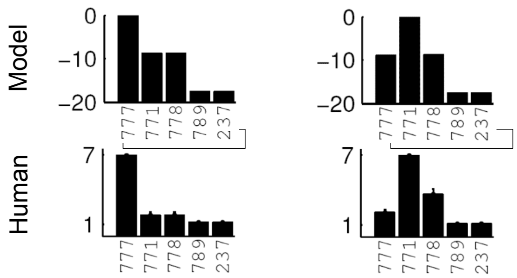
Learning from positive examples

1,1,1 2,2,2 2,2,1 3,3,1
3,3,3 4,4,4 4,4,1 5,5,1

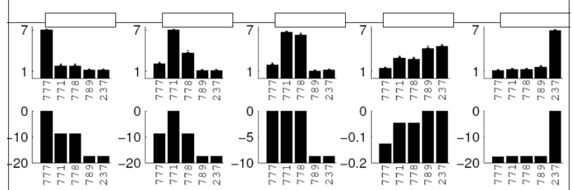
R(X,X,X). R(X,X,1).

First order theories

- Prior $P(T)$
  - $P(T) \propto 2^{-\text{length}(T)}$
- Likelihood $P(\text{Data}|T)$
  - Assume the data are sampled at random from all facts that are true according to T



1,1,1	2,2,2	2,2,1	3,3,1
3,3,3	4,4,4	4,4,1	5,5,1
R(X,X,X).		R(X,X,1).	



R(X,X,X)	R(X,X,1)	R(X,X,Y)	R(X,Y,Z)	R(2,3,X)
1,1,1	2,2,1	2,2,1	2,3,1	2,3,1
2,2,2	3,3,1	4,4,3	4,5,6	2,3,4
3,3,3	4,4,1	5,5,2	6,1,5	2,3,5
4,4,4	5,5,1	6,6,3	3,4,4	2,3,6

Conclusion

- Psychologists have argued that human knowledge is organized into systems of relations.
- Probabilistic inference can help to explain how these systems are acquired.

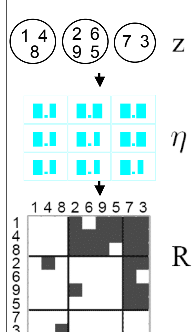
How to learn ~~a T~~ anything

- Search for T that maximizes

$$P(T|\text{Data}) \propto P(\text{Data}|T)P(T)$$

- Prerequisites
  - Decide how Ts are represented.
  - Put a prior over a hypothesis space of Ts.
  - Decide how observable data are generated from an underlying T.

Model Inference



Find *z* that maximizes

$$P(\mathbf{z}|R) \propto P(R|\mathbf{z})P(\mathbf{z})$$

$$P(R|\mathbf{z}) = \prod_{ab} \frac{B(m_{ab}+\alpha, \bar{m}_{ab}+\beta)}{B(\alpha,\beta)}$$

where m_{ab} is the number of 1-edges between classes *a* and *b*

$\bar{m}_{ab}$ is the number of 0-edges between classes *a* and *b*

$B(\cdot,\cdot)$ is the Beta function