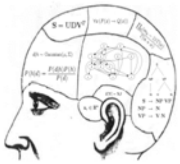


Computational models of cognitive development: the grammar analogy

Josh Tenenbaum
MIT

## Top 10 reasons to be Bayesian

1. Bayesian inference over hierarchies of structured representations provides integrated answers to these key questions of cognition:
   - What is the content and form of human knowledge, at multiple levels of abstraction?
   - How is abstract knowledge used to guide learning and inference of more specific representations?
   - How are abstract systems of knowledge themselves learned?

## The plan

- This morning…

*P*(grammar)
Linguistic grammar
↓ *P*(phrase structures | grammar)
Phrase structures
↓ *P*(utterances | phrase structures)
Utterances

*P*(grammar)
Scene grammar
↓ *P*(objects | grammar)
Object layout (scenes)
↓ *P*(features | objects)
Image features

- … this afternoon: higher-level cognition
  - Causal theories
  - Theories for object categories and properties

## (fragments of) Theories

- Physics
  - The natural state of matter is at rest. Every motion has a causal agent.
  - F = m a.
- Medicine
  - People have different amounts of four humors in the body, and illnesses are caused by them being out of balance.
  - Diseases result from pathogens transmitted through the environment, interacting with genetically controlled cellular mechanisms.
- Psychology
  - Rational agents choose efficient plans of action in order to satisfy their desires given their beliefs.
  - Action results from neuronal activity in the prefrontal cortex, which activates units in the motor cortex that control motor neurons projecting to the body's muscles.

## The structure of intuitive theories

"A theory consists of three interrelated components: a set of phenomena that are in its domain, the causal laws and other explanatory mechanisms in terms of which the phenomena are accounted for, and the concepts in terms of which the phenomena and explanatory apparatus are expressed."

Carey (1985), "Constraints on semantic development"

## The function of intuitive theories

- "Set the frame" of cognition by defining different domains of thought.
- Describe what things are found in that domain, as well as what behaviors and relations are possible
- Provide causal explanations appropriate to that domain.
- Guide inference and learning:
  - Highlights variables to attend to
  - Generates hypotheses/explanations to consider
  - Supports generalization from very limited data
  - Supports prediction and action planning.

## Theories in cognitive development

The big question: what develops from childhood to adulthood?

- One extreme: basically everything
  - Totally new ways of thinking, e.g. logical thinking.
- The other extreme: basically nothing
  - Just new facts (specific beliefs), e.g., trees can die.
- Intermediate view: something important
  - New theories, new ways of organizing beliefs.

## Hierarchical Bayesian models for learning abstract knowledge



| | | |
|---|---|---|
| Causal theory | Structural form | Linguistic grammar |
| ↓ | ↓ | ↓ |
| Causal network structures | Structure over objects | Phrase structures |
| ↓ | ↓ | ↓ |
| Observed events | Observed properties of objects | Observed utterances in the language |

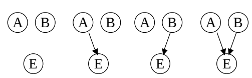
## Hierarchical Bayesian framework for causal induction

Causal theory

Objects can activate Machines
Activation requires contact
Machines are (near) deterministic

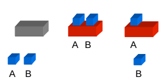
↓

Causal network structures




↓

Observed events



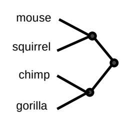

(Griffiths, Tenenbaum, Kemp et al.)

## Hierarchical Bayesian framework for property induction

Structural form — Tree with species at leaf nodes

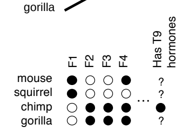
↓

Structure over objects




↓

Observed properties of objects




(Kemp & Tenenbaum)

## Questions

- How can the abstract knowledge be learned?
- How do we represent abstract knowledge?

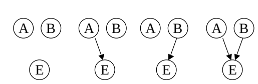
A tradeoff in representational expressiveness versus learnability...

Causal theory

**Types:** Block, Detector, Trial
**Predicates:**
Contact(block, detector, trial) non-random
Activates(block, detector) random
Active(detector, trial) random
**Dependency statements:**
Activates(b, d) ~ TabularCPD[[q (1-q)]];
Active(d, t) ~ NoisyORAggCPD[1 0]
({Contact(b, d, t) for block b: Activates(b, d)});

↓

Causal structure




Activates(A, detector)?
Activates(B, detector)?

Relational Skeleton (unknown)

↓

Observed events

Contact(A, detector, t1)
Contact(B, detector, t1)
Active(detector, t1)

Contact(A, detector, t2)
~Contact(B, detector, t2)
Active(detector, t2)

# Questions

- How can the abstract knowledge be learned?
- How do we represent abstract knowledge?

A tradeoff in representational expressiveness versus learnability...

… consider a representation for theories which is less expressive but more learnable.

# Learning causal networks

Causal structure

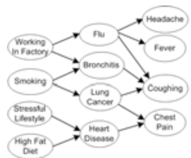
↓

Observed events


conditions

patients | has(patient,condition)

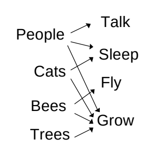
*Different* causal networks, *Same* domain theory

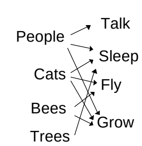


*Different* domain theories

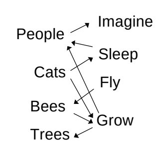
# The grammar analogy

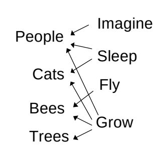
*Different* semantic networks, *Same* linguistic grammar




*Different* grammars




# The grammar analogy

<u>Natural Language Grammar</u>

Abstract classes: N, V, ....

Production rules: S → [N V], ....

("Nouns precede verbs")

N → {people, cats, bees, trees, ....}

V → {talks, sleep, fly, grow, ....}

Linguistic Grammar

↓

Syntactic structures

↓

Observed sentences in the language

<u>Causal Theory</u>

Abstract classes: D, S, ....

Causal laws: [D ⇢ S], ....

("Diseases cause symptoms")

D → {flu, bronchitis, lung cancer,....}

S → {fever, cough, chest pain, ....}

Causal Theory

↓

Causal network structures

↓

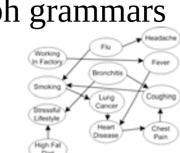
Observed data in the domain

# Theories as graph grammars

Classes = {B, D, S}

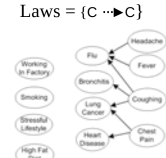
Laws = {B ⇢ D, D ⇢ S}

(⇢ : possible causal link)



Classes = {C}

Laws = {C ⇢ C}



Classes = {B, D, S}

Laws = {S ⇢ D}

# History of the grammar analogy

"The grammar of a language can be viewed as a theory of the structure of this language. Any scientific theory is based on a certain finite set of observations and, by establishing general laws stated in terms of certain hypothetical constructs, it attempts to account for these observations, to show how they are interrelated, and to predict an indefinite number of new phenomena…. Similarly, a grammar is based on a finite number of observed sentences… and it 'projects' this set to an infinite set of grammatical sentences by establishing general 'laws'… [framed in terms of] phonemes, words, phrases, and so on.…"

Chomsky (1956), "Three models for the description of language"

# Learning causal networks

Causal theory

↓

Causal structure

↓

Observed events

Classes = {B, D, S}

Laws = {B → D, D → S}

B: working in factory, smoking, stress, high fat diet, …

D: flu, bronchitis, lung cancer, heart disease, …

S: headache, fever, coughing, chest pain, …


| | conditions |
|---|---|
| patients | has(patient,condition) |

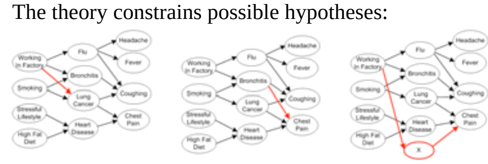
# Using a causal theory

Given current causal network beliefs . . .


. . . and some new observed data:

Correlation between "working in factory" and "chest pain".

The theory constrains possible hypotheses:



And rules out others:

- Allows strong inferences about causal structure from very limited data.
- Very different from conventional Bayes net learning.

# Learning causal theories

Causal theory

↓

Causal structure

↓

Observed events

Classes = {B, …

Laws = {B → …

…tory, … high fat

…, lung …sease,

…

S: headache, fever, coughing, chest pain, …

How do we define the theory as a probabilistic generative model?


| | conditions |
|---|---|
| patients | has(patient,condition) |

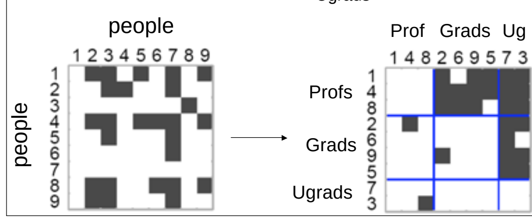
# Learning a relational theory

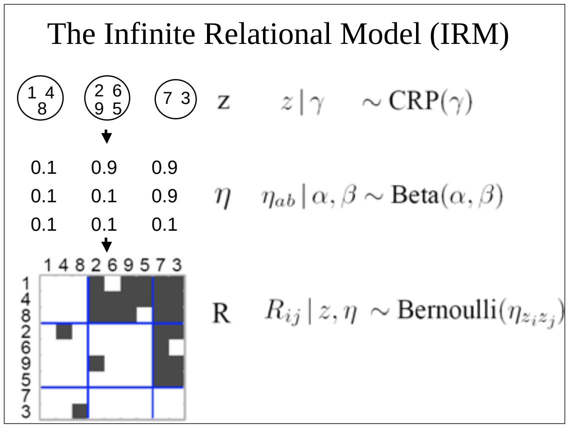
(c.f. Charles Kemp, Monday)

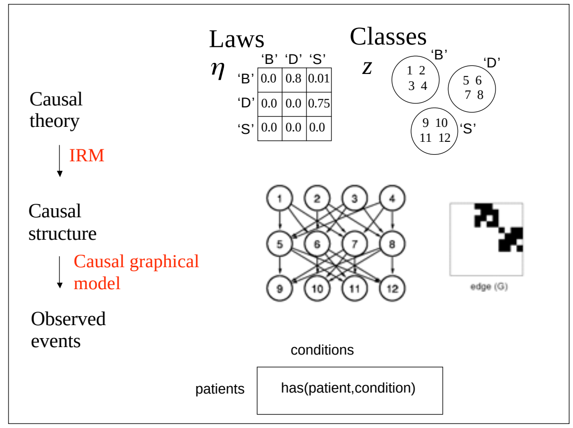
Profs: 1 4 8 — Grads: 2 6 9 5 — Ugrads: 7 3

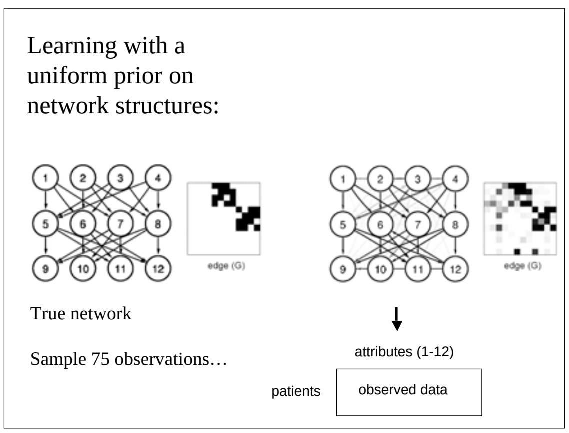
*dominates*

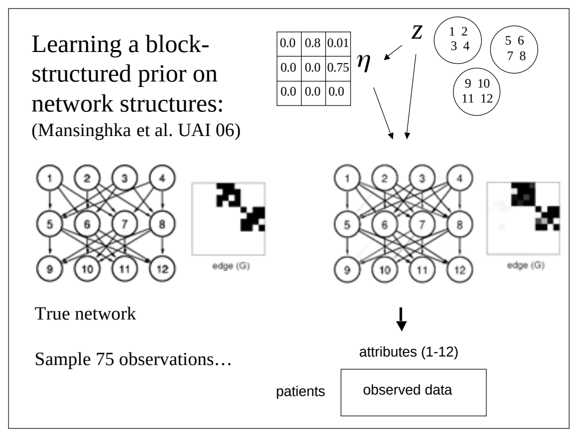
| | Profs | Grads | Ugrads |
|---|---|---|---|
| Profs | 0.1 | 0.9 | 0.9 |
| Grads | 0.1 | 0.1 | 0.9 |
| Ugrads | 0.1 | 0.1 | 0.1 |

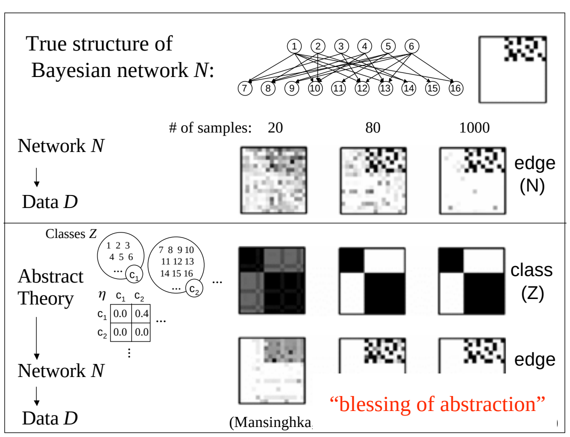


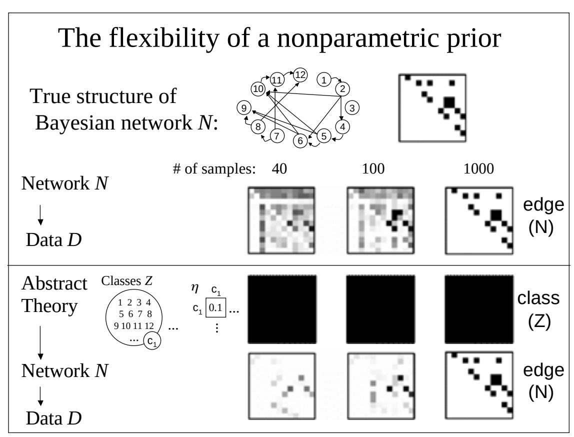

The Infinite Relational Model (IRM)
z
$z \mid \gamma \quad \sim \mathrm{CRP}(\gamma)$
0.1 0.9 0.9
0.1 0.1 0.9
0.1 0.1 0.1
$\eta$
$\eta_{ab} \mid \alpha, \beta \sim \mathrm{Beta}(\alpha, \beta)$
1 4 8 2 6 9 5 7 3
R
$R_{ij} \mid z, \eta \ \sim \mathrm{Bernoulli}(\eta_{z_i z_j})$


Laws
Classes
$\eta$
z
'B' 'D' 'S'
'B' 0.0 0.8 0.01
'D' 0.0 0.0 0.75
'S' 0.0 0.0 0.0
Causal theory
IRM
Causal structure
Causal graphical model
edge (G)
Observed events
conditions
patients
has(patient,condition)


Learning with a uniform prior on network structures:
edge (G)
True network
Sample 75 observations…
attributes (1-12)
patients
observed data


Learning a block-structured prior on network structures:
(Mansinghka et al. UAI 06)
0.0 0.8 0.01
0.0 0.0 0.75
0.0 0.0 0.0
$\eta$
Z
edge (G)
True network
Sample 75 observations…
attributes (1-12)
patients
observed data


True structure of Bayesian network N:
# of samples: 20 80 1000
Network N
Data D
edge (N)
Classes Z
Abstract Theory
class (Z)
Network N
Data D
edge
"blessing of abstraction"
(Mansinghka


The flexibility of a nonparametric prior
True structure of Bayesian network N:
# of samples: 40 100 1000
Network N
Data D
edge (N)
Abstract Theory
Classes Z
class (Z)
Network N
Data D
edge (N)

## Insights

- Abstract theories, which have traditionally either been neglected or presumed to be innate in cognitive development, could in fact be learned from data by rational inferential means, together with specific causal relations.
- Abstract theories can in some cases be learned *more quickly and more easily* than specific concrete causal relations (the "blessing of abstraction").
- Key to progress on learning abstract knowledge : finding sweet spots in the expressiveness – learnability tradeoff….

## Hierarchical Bayesian models for learning abstract knowledge

Causal theory
↓
Causal network structures
↓
Observed events

Structural form
↓
Structure over objects
↓
Observed properties of objects

Linguistic grammar
↓
Phrase structures
↓
Observed utterances in the language

## A framework for inductive learning

1. How does knowledge guide learning from sparsely observed data? Bayesian inference, with priors based on background knowledge.
2. What form does knowledge take, across different domains and tasks? Probabilities defined over structured representations: graphs, grammars, rules, logic, relational schemas, theories.
3. How is more abstract knowledge itself learned? Hierarchical Bayesian models, with inference at multiple levels of abstraction.
4. How can abstract knowledge constrain learning yet maintain flexibility, balancing assimilation and accommodation? Nonparametric models, growing in complexity as the data require.