Interoperability of probabilistic programs and data (probabilistic programs for representing scientific hypotheses)

David Poole

Department of Computer Science,
University of British Columbia
Leverhulme Trust visting professor at the University of Oxford

[Work with: Clinton Smyth, Rita Sharma, Jacek Kisynski, Lionel E. Jackon, Jr., Chia-Li Kuo, David Buchman]

April 2015

Research Agendas

- PubMed comprises over 24 million citations for biomedical literature. 10,000 added each week.
- IBM's Watson (and others) propose to read the literature to provide "evidence-based" advice for specific patients.
- Can we do better than data $\longrightarrow$ hypotheses $\longrightarrow$ research papers $\longrightarrow$ (mis)reading $\longrightarrow$ clinical practice?
- Wouldn't it be better to have the research published in machine readable form?

Example: Geology

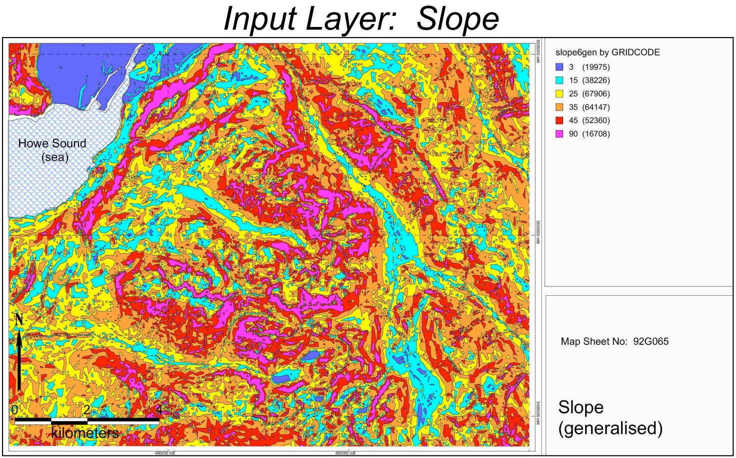
- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
- Geological "observations" are published by the geological surveys of counties and states/provinces and globally (onegeology.org)
- Geological hypotheses are published in research journals.
- We represented the hypotheses of hundreds of research papers, and matched them on thousands of descriptions of interesting places

[Research with Clinton Smyth, Georeference Online]

OneGeology.org

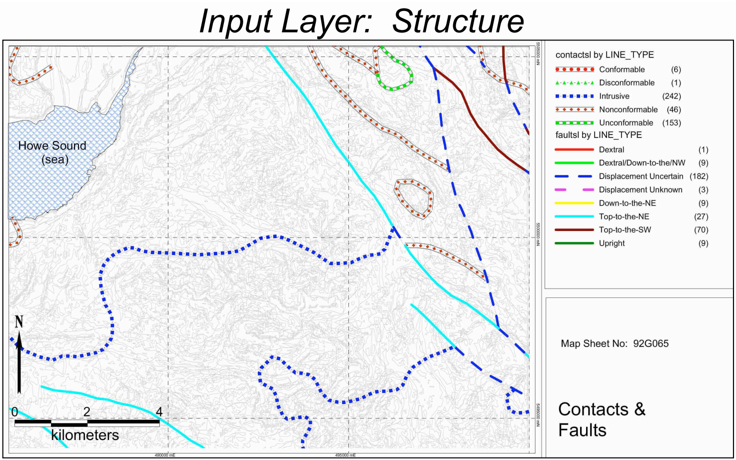
OneGeology.org

Example Data, Geology



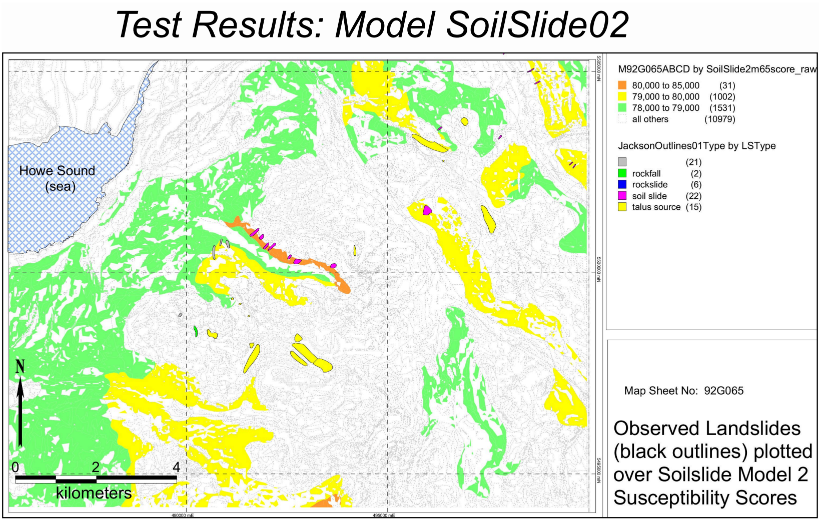
[Clinton Smyth, Georeference Online.]

Example Data, Geology



[Clinton Smyth, Georeference Online.]

Example Prediction from a Hypothesis

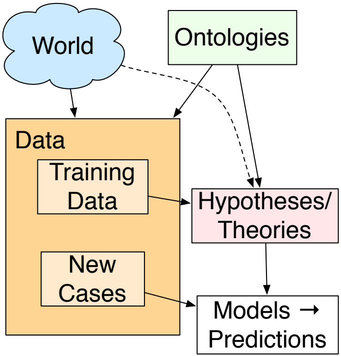


[Clinton Smyth, Georeference Online.]

Outline

1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Models: Ensembles of hypotheses
4. Property Domains and Undefined Random Variables

Semantic Science



- Ontologies represent the meaning of symbols.
- Observational data is published.
- Hypotheses make predictions on data.
- Data used to evaluate hypotheses.
- Hypotheses used for predictions on new cases.
- All evolve in time.

Science as the foundation of world-wide mind

Observations and hypotheses can be about anything:

- where and when landslides occur
- where to find gold
- what errors students make
- disease symptoms, prognosis and treatment
- what companies will be good to invest in
- what apartment Mary would like
- which celebrities are having affairs

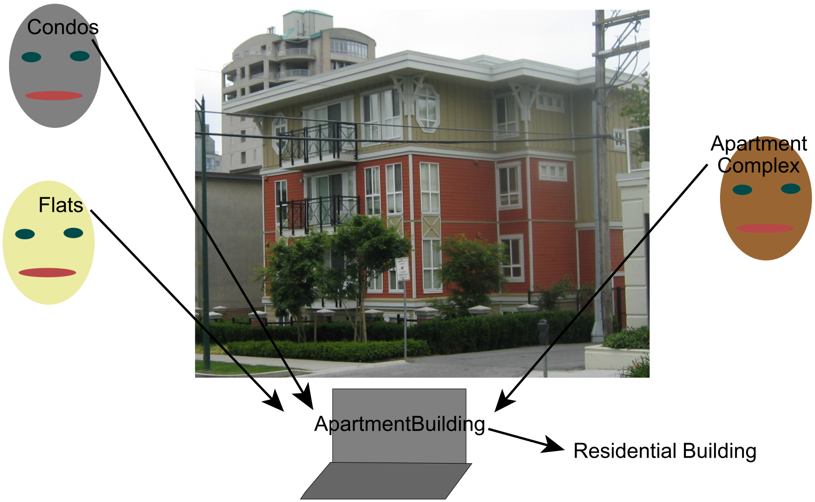
Outline

1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Models: Ensembles of hypotheses
4. Property Domains and Undefined Random Variables

Ontologies

- In philosophy, ontology the study of existence.
- In CS, an ontology is a (formal) specification of the meaning of the vocabulary used in an information system.
- Ontologies are needed so that information sources can inter-operate at a semantic level.

Ontologies


Condos
Flats
Apartment Complex
ApartmentBuilding
Residential Building

Main Components of an Ontology

- Individuals: the objects in the world
  (not usually specified as part of the ontology)
- Classes: sets of (potential) individuals
- Properties: between individuals and their values

$\langle \mathit{Individual}, \mathit{Property}, \mathit{Value} \rangle$ triples are universal representations of relations.

Aristotelian definitions

Aristotle [350 B.C.] suggested the definition if a class C in terms of:

- Genus: the super-class
- Differentia: the attributes that make members of the class C different from other members of the super-class

"If genera are different and co-ordinate, their differentiae are themselves different in kind. Take as an instance the genus 'animal' and the genus 'knowledge'. 'With feet', 'two-footed', 'winged', 'aquatic', are differentiae of 'animal'; the species of knowledge are not distinguished by the same differentiae. One species of knowledge does not differ from another in being 'two-footed'."

Aristotle, *Categories*, 350 B.C.

An Aristotelian definition

- An **apartment building** is a **residential building** with **multiple units** and **units are rented**.

$$\begin{aligned} ApartmentBuilding \equiv\ & ResidentialBuilding \& \\ & NumUnits = many \& \\ & Ownership = rental \end{aligned}$$

  NumUnits is a property with domain *ResidentialBuilding* and range $\{one, two, many\}$
  Ownership is a property with domain *Building* and range $\{owned, rental, coop\}$.

- All classes are defined in terms of properties.

Outline

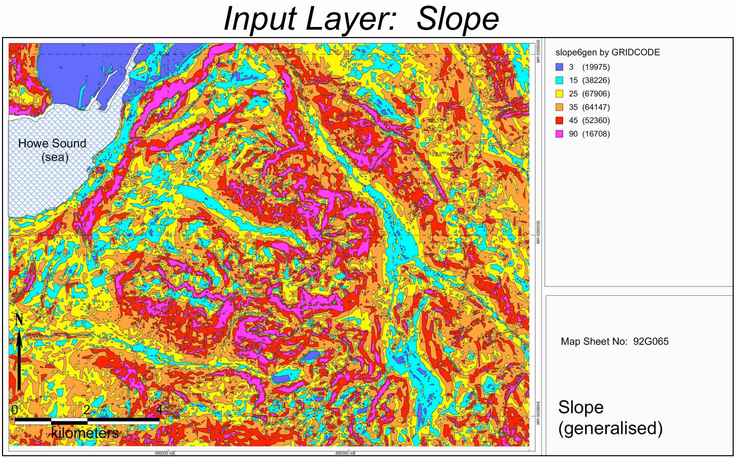
1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Models: Ensembles of hypotheses
4. Property Domains and Undefined Random Variables

Data

Real data is messy!

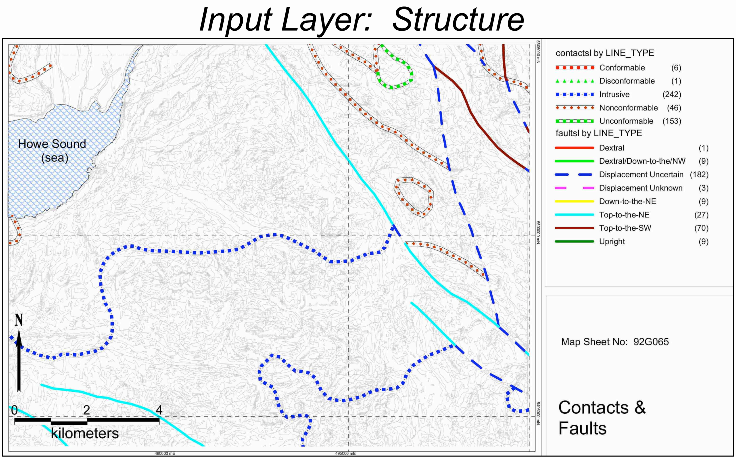
- Multiple levels of abstraction
- Multiple levels of detail
- Uses the vocabulary from many ontologies: rocks, minerals, top-level ontology,...
- Rich meta-data:
  - Who collected each datum? (identity and credentials)
  - Who transcribed the information?
  - What was the protocol used to collect the data? (Chosen at random or chosen because interesting?)
  - What were the controls — what was manipulated, when?
  - What sensors were used? What is their reliability and operating range?

Example Data, Geology



[Clinton Smyth, Georeference Online.]

Example Data, Geology



[Clinton Smyth, Georeference Online.]

Data is theory-laden

- Sapir-Whorf Hypothesis [Sapir 1929, Whorf 1940]: people's perception and thought are determined by what can be described in their language.
  (Controversial in linguistics!)
- A stronger version for information systems:

  What is stored and communicated by an information system is constrained by the representation and the ontology used by the information system.

- Ontologies must come logically prior to the data.
- Data can't make distinctions that can't be expressed in the ontology.
- Different ontologies result in different data.

Outline

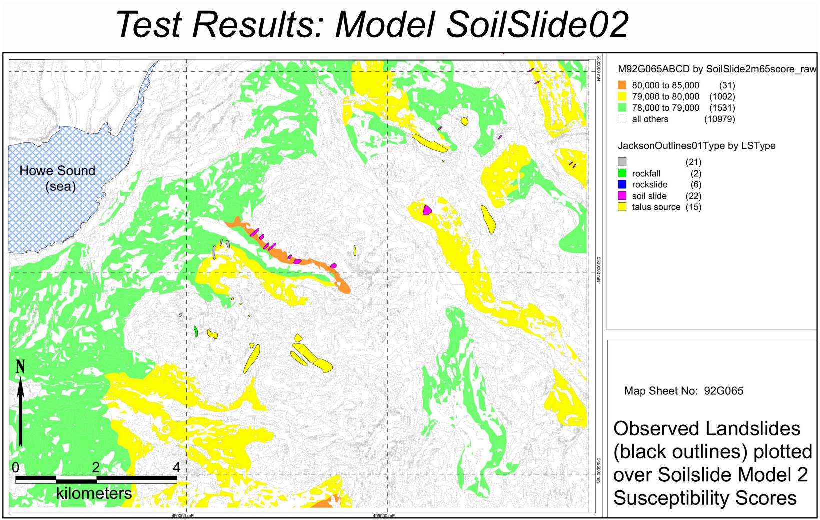
1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Models: Ensembles of hypotheses
4. Property Domains and Undefined Random Variables

Hypotheses make predictions on data

Hypotheses are programs that make predictions on data.
Theories are hypotheses that best fit the observational data.

- Hypotheses can make various predictions about data:
  - definitive predictions
  - point probabilities
  - probability ranges
  - ranges with confidence intervals
  - qualitative predictions
- Users can use whatever criteria they like to evaluate hypotheses (e.g., taking into account simplicity and elegance)
- Semantic science search engine: extract theories from published hypotheses.

Example Prediction from a Hypothesis



[Clinton Smyth, Georeference Online.]

Dynamics of Semantic Science

- New data and hypotheses are continually added.
- Anyone can design their own ontologies.
  — People vote with their feet what ontology they use.
  — Need for semantic interoperability leads to ontologies with mappings between them.
- Ontologies evolve with hypotheses:
  A hypothesis hypothesizes unobserved features or useful distinctions
  $\longrightarrow$ add these to an ontology
  $\longrightarrow$ other researchers can refer to them
  $\longrightarrow$ reinterpretation of data
- Ontologies can be judged by the predictions of the hypotheses that use them
  — role of a vocabulary is to describe useful distinctions.

Outline

Random Variables and Triples

- Reconcile:
  - random variables of probability theory
  - individuals, classes, properties of modern ontologies

Random Variables and Triples

- Reconcile:
  - random variables of probability theory
  - individuals, classes, properties of modern ontologies
- For **functional properties**:
  R is functional: $\langle x, R, y_1 \rangle$ and $\langle x, R, y_2 \rangle$ implies $y_1 = y_2$.
  random variable for each $\langle individual, property \rangle$ pair,
  range of the random variable is range of the property.
  E.g., if *Height* is functional, $\langle building17, Height \rangle$ is a random variable.

Random Variables and Triples

- Reconcile:
  - random variables of probability theory
  - individuals, classes, properties of modern ontologies
- For **functional properties**:
  R is functional: $\langle x, R, y_1 \rangle$ and $\langle x, R, y_2 \rangle$ implies $y_1 = y_2$.
  random variable for each $\langle \mathit{individual}, \mathit{property} \rangle$ pair,
  range of the random variable is range of the property.
  E.g., if *Height* is functional, $\langle \mathit{building}17, \mathit{Height} \rangle$ is a random variable.
- For **non-functional properties**:
  Boolean random variable for each $\langle \mathit{individual}, \mathit{property}, \mathit{value} \rangle$ triple.
  E.g., if *YearRestored* is non-functional $\langle \mathit{building}17, \mathit{YearRestored}, 1988 \rangle$ is a Boolean random var.

Ranges

	OWL	Probability
Datatype	Boolean, Real, Integer, String, DateTime...	Boolean, Real, Integer, String, DateTime...
ObjectProperty		$\begin{cases} \text{Discrete / Multinomial} \\ \text{Relational} \end{cases}$

E.g., consider the ranges:

- {very_tall, tall, medium, short}
- {10 High St, 22 Smith St, 57 Jericho Ave}

Probabilities and Aristotelian Definitions

Aristotelian definition

$$\begin{aligned} ApartmentBuilding \equiv\ & ResidentialBuilding\,\& \\ & NumUnits = many\,\& \\ & Ownership = rental \end{aligned}$$

leads to probability over class membership

$$\begin{aligned} & P(\langle A, type, ApartmentBuilding\rangle) \\ & = P(\langle A, type, ResidentialBuilding\rangle) \times \\ & \times P(\langle A, NumUnits\rangle = many \mid \langle A, type, ResidentialBuilding\rangle \\ & \times P(\langle A, Ownership, rental\rangle \mid \langle A, NumUnits\rangle = many, \\ & \qquad \langle A, type, ResidentialBuilding\rangle) \end{aligned}$$

No need to consider undefined propositions.

Outline

1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Models: Ensembles of hypotheses
4. Property Domains and Undefined Random Variables

Hypotheses, Models and Predictions

- Recall: a hypothesis is a program that makes predictions on data
- Hypotheses are often very narrow.
- We typically use many hypotheses to make a prediction.
- Hypotheses differ in
  - level of generality (high-level/low level)
    e.g., mammal vs poodle
  - level of detail (parts/subparts)
    e.g., mammal vs left eye

Applying hypotheses to new cases

- How can we compare hypotheses that differ in their generality?
- Hypothesis A makes predictions about all cancers. Hypothesis B makes predictions about lung cancers. Should the comparison between A and B take into account A's predictions on non-lung cancer?

Applying hypotheses to new cases

- How can we compare hypotheses that differ in their generality?
- Hypothesis A makes predictions about all cancers. Hypothesis B makes predictions about lung cancers. Should the comparison between A and B take into account A's predictions on non-lung cancer?
- What about C: *if lung cancer, use B's prediction, else use A's prediction*?

Applying hypotheses to new cases

- How can we compare hypotheses that differ in their generality?
- Hypothesis A makes predictions about all cancers. Hypothesis B makes predictions about lung cancers. Should the comparison between A and B take into account A's predictions on non-lung cancer?
- What about C: *if lung cancer, use B's prediction, else use A's prediction*?
- A model or hypothesis ensemble is a set of hypotheses applied to a particular case.
  - Judge hypotheses by how well they fit into models.
  - Models can be judged by simplicity.
  - Hypothesis designers don't need to game the system by manipulating the generality of hypotheses

Example Data

person visiting doctor:

Age	Sex	Coughs	HasLump
23	male	true	true
...	...	...	...

lump for person visiting doctor:

Location	LumpShape	Colour	CancerousLump
leg	oblong	red	false
...	...	...	...

person with cancer:

HasLungCancer	Treatment	Age	Outcome	Months
true	chemo	77	dies	7
...	...	...	...	...

Hypotheses

A hypothesis is of the form $\langle c, I, O, P \rangle$

- A context c in which specifies when it can be applied.
- A set of input features I about which it does not make predictions
- A set of output features O to predict (as a function of the input features).
- A program P to compute the output from the input.

Represents:

$$P(O \mid c, I)$$

or divide I into observation I_{obs} and intervention inputs I_{do}:

$$P(O \mid c, I_{obs}, do(I_{do}))$$

Example

Consider the following hypotheses:

- T_1 predicts the prognosis of people with lung cancer.
- T_2 predicts the prognosis of people with cancer.
- T_3 is the null hypothesis that predicts the prognosis of people in general.
- T_4 predicts whether people with cancer have lung cancer, as a function of coughing.
- T_5 predicts whether people have cancer.

What should be used to predict the prognosis of a patient with observed coughing?

Models

A model consists of multiple hypotheses, where each hypothesis can be used to predict a subset of its output features.

A model M needs to satisfy the following properties:

- M is coherent: it does not rely on the value of a feature in a context where the feature is not defined
- M is consistent: it does not make different predictions for any feature in any context.
- M is predictive: it makes a prediction in every context that is possible (probability > 0).
- M is minimal: no subset is also a model.

Model and Ensembles of Hypotheses

A hypothesis instance is a tuple of the form $\langle h, c, I, O \rangle$ such that:

- h is a hypothesis,
- c is a context in which the hypothesis will be used
- I is a set of inputs used by the hypothesis
- O is a set of outputs the hypothesis will be used to predict.

A model is a set of hypothesis instances that satisfy the previous conditions.

[Think of a model as a Bayesian belief network, but allowing for context-specific independence, avoiding undefined features, that interfaces with ontologies and used program to represent conditional probabilities.]

Example

- T_1 predicts the prognosis of people with lung cancer.
- T_2 predicts the prognosis of people with cancer.
- T_3 is the null hypothesis that predicts the prognosis of people in general.
- T_4 predicts (probabilistically) whether people with cancer have lung cancer, as a function of coughing.
- T_5 predicts (probabilistically) whether people have cancer.

A possible model for $P(Lives \mid person \wedge coughs)$:

- $\langle T_5, person, \{\}, \{HC\}\rangle$,
- $\langle T_3, person \wedge \neg hc, \{\}, \{Lives\}\rangle$,
- $\langle T_4, person \wedge hc, \{Coughs\}, \{HLC\}\rangle$,
- $\langle T_1, person \wedge hlc, \{\}, \{Lives\}\rangle$,
- $\langle T_2, person \wedge hc \wedge \neg hlc, \{\}, \{Lives\}\rangle$.

Programs and Meta-programs

Two sorts of probabilistic programs:

- Hypotheses are probabilistic programs that persist, are tuned to data.
- Models are probabilistic programs that are adapted to particular cases. Transient. Use hypotheses as subroutines.

Science versus application.
Always ask: "Why should we believe this prediction?"

Outline

Domains and Undefined Random Variables (Example)

Example (Ontology)

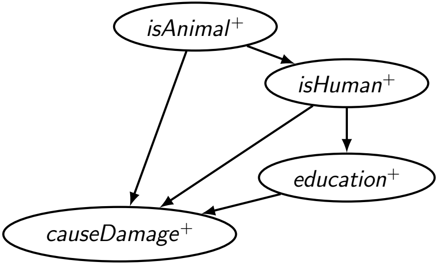
```
Classes:
  Thing
    Animal: Thing and isAnimal = true
      Human: Animal and isHuman = true

Properties:
  isAnimal:    domain: Thing   range: {true,false}
  isHuman:     domain: Animal  range: {true,false}
  education:   domain: Human   range: {low,high}
  causeDamage: domain: Thing   range: {true,false}
```

A property is only defined for individuals in its domain.

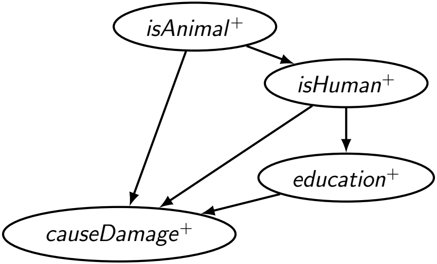
Extended Belief Networks (EBNs)

- Add "undefined" ($\perp$) to each range.
  - $range(isHuman^{+}) = \{true, false, \perp\}$.
  - $range(education^{+}) = \{low, high, \perp\}$.



- $education^{+}$ is like *education* but with an expanded range.
- Possible query: $P(education^{+} \mid causeDamage^{+} = true)$

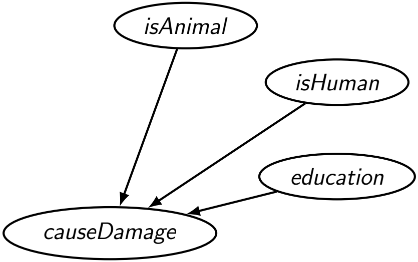
Extended Belief Networks (EBNs)



However...

- Expanding ranges is computationally expensive.
  - Exact inference has time complexity $\mathcal{O}(|range|^{treewidth})$.
- It may not be sensible to think about undefined values; no dataset would contain such values.
- Arcs $\langle isAnimal^+, isHuman^+ \rangle$ and $\langle isHuman^+, education^+ \rangle$ represent logical constraints

Ontologically-Based Belief Networks (OBBNs)



- OBBNs decouple the logical constraints (from the ontology) from the probabilistic dependencies.
- Don't model undefined ($\perp$) in ranges.
- The probabilistic network does not contain any ontological information.

Conditional Probabilities

isAnimal

true / \ *false*

isHuman (0.1, 0.9)

true / \ *false*

education (0.3, 0.7)

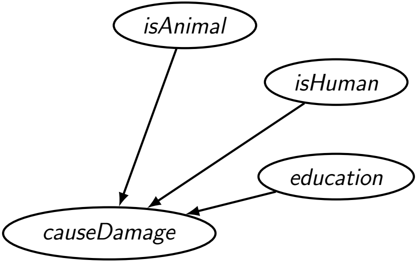
high / \ *low*

(0.9, 0.1) (0.5, 0.5)

$$P(causeDamage \mid isAnimal, isHuman, education)$$

- For each random variable, only specify (conditional) probabilities for well-defined contexts.

Ontologically-Based Belief Networks (OBBNs)



- The query $P(\mathit{education}^{+} \mid \mathit{causeDamage} = \mathit{true})$ has a non-zero probability of $\perp$
  — we can't ignore the undefined values.

Ontologically-Based Belief Networks (Inference)

The following give the same answer for $P(Q^+ \mid \mathcal{E} = e)$:

- Compute $P(Q^+ \mid \mathcal{E}^+ = e)$ using the extended belief network.
- From the OGBN:
    - Query the ontology for $domain(Q)$
    - Let $\alpha = P(domain(Q) \mid \mathcal{E} = e)$
    - If $\alpha \neq 0$ let $\beta = P(Q \mid \mathcal{E} = e \land domain(Q))$
    - Return

$$P(Q^+ = \bot \mid \mathcal{E} = e) = 1 - \alpha$$
$$P(Q \mid \mathcal{E} = e) = \alpha\beta$$

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?
- Surely we should have probabilistic ontologies?

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?
- Surely we should have probabilistic ontologies?
- How can we stop people from publishing fictional data?

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?
- Surely we should have probabilistic ontologies?
- How can we stop people from publishing fictional data?
- How can we test hypotheses if there is no "held-out" data? (Won't everyone cheat?)

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?
- Surely we should have probabilistic ontologies?
- How can we stop people from publishing fictional data?
- How can we test hypotheses if there is no "held-out" data? (Won't everyone cheat?)
- Why do you assume that probability is the right formalism?

Some Objections

- Currently hypotheses are buried in research articles; can't we just use IBM's Watson to read these?
- Surely we should have probabilistic ontologies?
- How can we stop people from publishing fictional data?
- How can we test hypotheses if there is no "held-out" data? (Won't everyone cheat?)
- Why do you assume that probability is the right formalism?
- How can you convince people to use maximally informed priors rather than maximally uninformed priors?

Conclusion

- Semantic science is a way to develop and deploy knowledge about how the world works.
  - Scientists (and others) develop hypotheses that refer to standardized ontologies and predict for new cases.
  - Multiple hypotheses—forming models—are needed to make predictions in particular cases.
  - For each prediction, we can ask what hypotheses it is based on.
  - For each hypothesis, we can ask about the evidence on which it can be evaluated.
- Ontologies, hypotheses and observations interact in complex ways.
- Many formalisms will be developed and discarded before we converge on useful representations.

To Do

- Representing, reasoning and learning complex (probabilistic) hypotheses. "probabilistic programming"
- Representations for observations that interacts with hypotheses.
- Build infrastructure to allow publishing and interaction of ontologies, data, hypotheses, models, evaluation criteria, meta-data.
- Build inverse semantic science web:
  - Given a hypothesis, find relevant data
  - Given data, find hypotheses that make predictions on the data
  - Given a new case, find relevant models with explanations