Ontologies, data and probabilistic hypotheses: Conditioning on all the knowledge in the world

David Poole

Department of Computer Science,
University of British Columbia

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

April 2016

For when I am presented with a false theorem, I do not need to examine or even to know the demonstration, since I shall discover its falsity *a posteriori* by means of an easy experiment, that is, by a calculation, costing no more than paper and ink, which will show the error no matter how small it is. . .

And if someone would doubt my results, I should say to him: "Let us calculate, Sir," and thus by taking to pen and ink, we should soon settle the question.

—Gottfried Wilhelm Leibniz [1677]

Questions considered

- Search engines give me the results of my query, but why should I believe these answers?

Questions considered

- Search engines give me the results of my query, but why should I believe these answers?
- The semantic web is supposed to make human knowledge accessible to computers, but how can we evaluate that knowledge?
- How could we go beyond the sum of human knowledge?

Questions considered

- Search engines give me the results of my query, but why should I believe these answers?
- The semantic web is supposed to make human knowledge accessible to computers, but how can we evaluate that knowledge?
- How could we go beyond the sum of human knowledge?
- There seems to be two branches of AI (Machine learning/uncertainty and the KR/logic/ontology); why do we have to choose one? Isn't there a coherent synthesis?

Questions considered

- Search engines give me the results of my query, but why should I believe these answers?
- The semantic web is supposed to make human knowledge accessible to computers, but how can we evaluate that knowledge?
- How could we go beyond the sum of human knowledge?
- There seems to be two branches of AI (Machine learning/uncertainty and the KR/logic/ontology); why do we have to choose one? Isn't there a coherent synthesis?
- What will AI and the web look like in 2030?

Example: medical diagnosis

Example: people give symptoms and want to know what is wrong with them.

Current Practice	An Alternative
— describe symptoms using keywords — results ranked by popularity (e.g., pagerank) and usually appeal to authority — text results	

Example: medical diagnosis

Example: people give symptoms and want to know what is wrong with them.

Current Practice	An Alternative
— describe symptoms using keywords	— use unambiguous terminology
— results ranked by popularity (e.g., pagerank) and usually appeal to authority	— predictions ranked by relevance and fit to data
— text results	— probabilistic predictions with references to sources

Believing information

2015	2030
• skeptics throw doubt on science and scientists say "trust us"	• data is available for all to view; all alternative hypotheses can be evaluated

Believing information

2015	2030
• skeptics throw doubt on science and scientists say "trust us"	• data is available for all to view; all alternative hypotheses can be evaluated
• evidence-based research is buried in research papers	• evidence-based results are available for everyday decisions

Believing information

2015	2030
• skeptics throw doubt on science and scientists say "trust us"	• data is available for all to view; all alternative hypotheses can be evaluated
• evidence-based research is buried in research papers	• evidence-based results are available for everyday decisions
• separation of uncertainty and KR issues	• uncertainty and ontologies are integral parts of world-wide mind

Believing information

2015	2030
• skeptics throw doubt on science and scientists say "trust us"	• data is available for all to view; all alternative hypotheses can be evaluated
• evidence-based research is buried in research papers	• evidence-based results are available for everyday decisions
• separation of uncertainty and KR issues	• uncertainty and ontologies are integral parts of world-wide mind
• relational representations starting to be used in ML	• rich representations with uncertainty ubiquitous

Believing information

2015	2030
• skeptics throw doubt on science and scientists say "trust us"	• data is available for all to view; all alternative hypotheses can be evaluated
• evidence-based research is buried in research papers	• evidence-based results are available for everyday decisions
• separation of uncertainty and KR issues	• uncertainty and ontologies are integral parts of world-wide mind
• relational representations starting to be used in ML	• rich representations with uncertainty ubiquitous
• data sets usable only by specialists	• data sets published, available, persistent and interoperable

Outline

Science is the foundation of belief

- A knowledge-based system should believe based on evidence. Not all beliefs are equally valid.

Science is the foundation of belief

- A knowledge-based system should believe based on evidence. Not all beliefs are equally valid.
- If a KR system makes a prediction, we should ask: what evidence is there?
  The system should be able to provide such evidence.

Science is the foundation of belief

- A knowledge-based system should believe based on evidence. Not all beliefs are equally valid.
- If a KR system makes a prediction, we should ask: what evidence is there?
  The system should be able to provide such evidence.
- The mechanism that has been developed for judging knowledge is called science. We trust scientific conclusions because they are based on evidence.

Science is the foundation of belief

- A knowledge-based system should believe based on evidence. Not all beliefs are equally valid.
- If a KR system makes a prediction, we should ask: what evidence is there?
  The system should be able to provide such evidence.
- The mechanism that has been developed for judging knowledge is called science. We trust scientific conclusions because they are based on evidence.
- The semantic web is an endeavor to make all of the world's knowledge accessible to computers.
  We use term semantic science, in an anaolgous way to the *semantic web*.
- Claim: semantic science will form the foundation of the world-wide mind.

Science as the foundation of world-wide mind

I mean *science* in the broadest sense:

- 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

Motivating Example

- 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

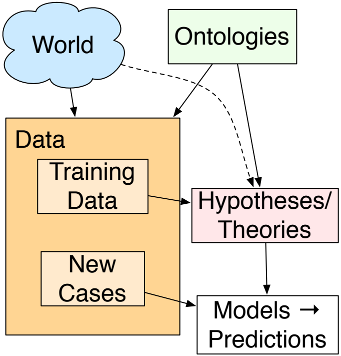
- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
- Geological "observations" are published by the geological surveys of counties and states/provinces and globally (onegeology.org)
- Geological hypotheses are published in research journals.
- We built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

[Work with Clinton Smyth, Georeference Online]

OneGeology.org

OneGeology.org

Semantic Science



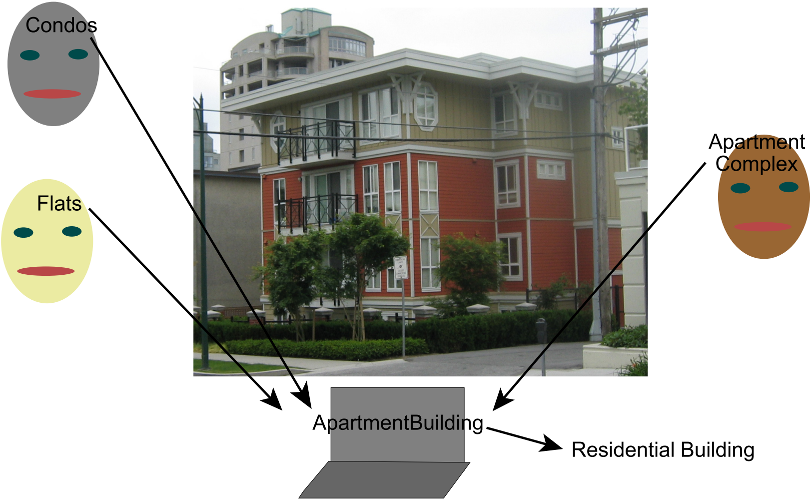
- Ontologies represent the meaning of symbols.
- Observational data is published.
- Hypotheses make predictions on data.
- Data used to evaluate hypotheses.
- Hypotheses used for predictions on new cases.
- All evolve in time.

Outline

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

$\langle Individual, Property, Value \rangle$ 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

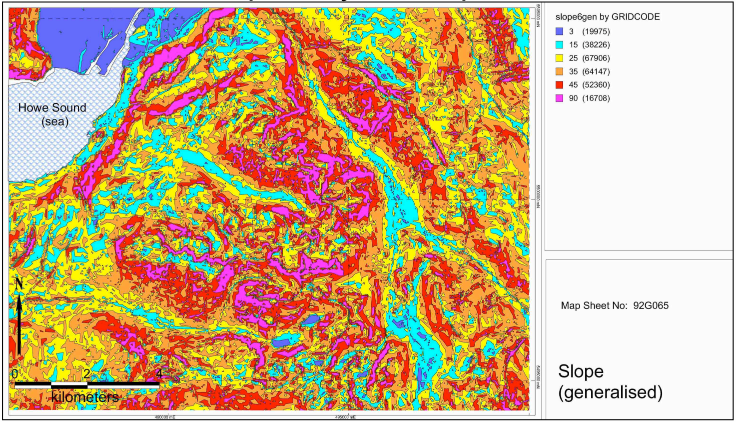
Data

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?
- Errors, forgeries, . . .

Example Data, Geology

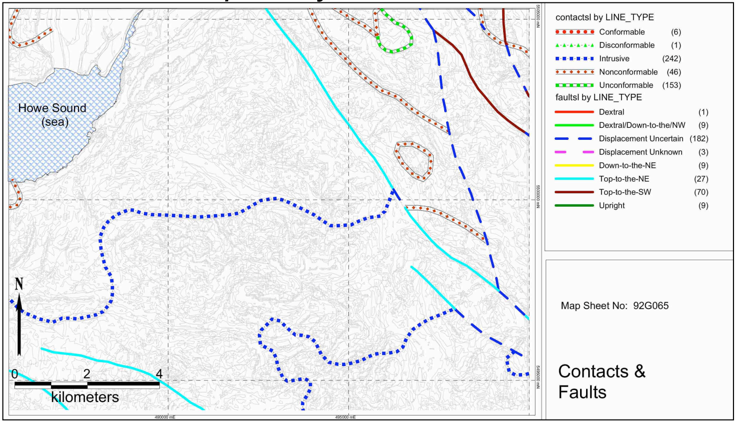
Input Layer: Slope



[Clinton Smyth, Georeference Online.]

Example Data, Geology

Input Layer: Structure



[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

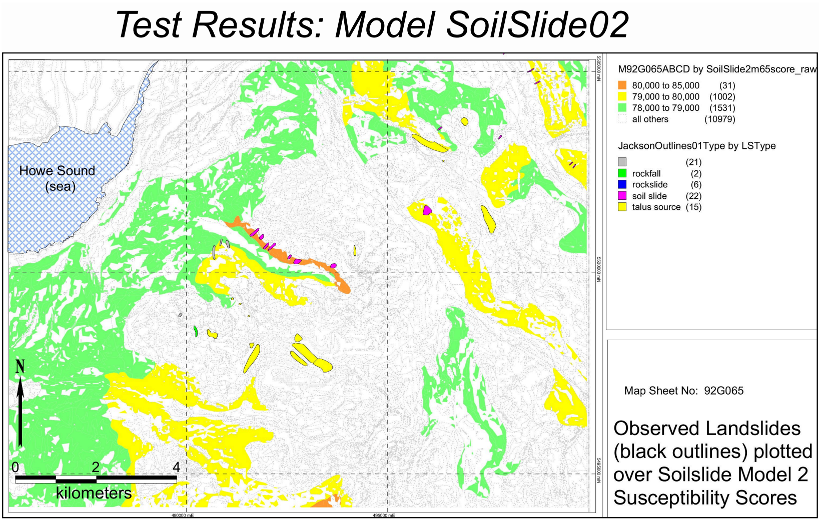
1. Semantic Science Overview
   - Ontologies
   - Data
   - Hypotheses
2. Probabilities with Ontologies
3. Property Domains and Undefined Random Variables
4. Models: Ensembles of hypotheses
5. Issues

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)

Example Prediction from a Hypothesis



[Clinton Smyth, Georeference Online.]

Levels of Semantic Science

0. Deterministic semantic science where all of the hypotheses make definitive predictions.
1. Feature-based semantic science, with non-deterministic predictions about feature values of data.
2. Relational semantic science, with predictions about the properties of (known) objects and relationships among objects.
3. First-order semantic science, with predictions about the existence of objects, identity, universally quantified statements and relations.

Outline

Random Variables and Triples

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

Random Variables and Triples

- Reconcile:
  - random variables (RVs) 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 RV is range of the property.
  E.g., if *Height* is functional, $\langle \mathit{building}17, \mathit{Height} \rangle$ is a RV.

Random Variables and Triples

- Reconcile:
  - random variables (RVs) 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 RV is range of the property.
  E.g., if *Height* is functional, $\langle \mathit{building}17, \mathit{Height} \rangle$ is a RV.
- For **non-functional properties**:
  Boolean RV 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 RV.

Ranges

	OWL	Probability
Datatype	Boolean, Real, Integer, String, DateTime...	Boolean, Real, Integer, String, DateTime...
ObjectProperty		{ Discrete / Multinomial Relational

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

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

leads to probability over class membership

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

(Conjunction here is not commutative — like $x \neq 0 \& y/x = z$)

Outline

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain.
- A property is almost always undefined:
  - *weight* is only defined for physical objects
  - *pitch* is only defined for sounds
  - *wavelength* is only defined for waves
  - *originality* is only defined for creative outputs
  - *hardness* (measured in Mohs scale) is only defined for minerals
  - *number_bedrooms* is only defined for buildings
- A dataset would not contain a triple with an undefined property

Domains and Undefined Random Variables (Example)

Example (Ontology)

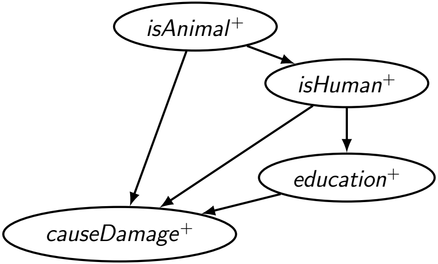
```
Classes:
  Thing
    Animal: Thing and isAnimal = true
      Human: Animal and isHuman = true

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}
```

education is not defined when $isHuman = false$.

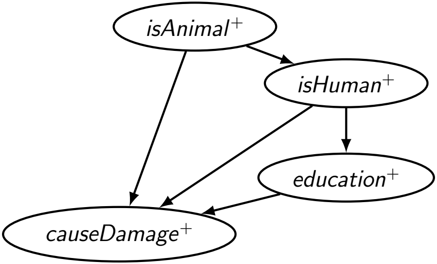
Extended Belief Networks (EBNs)

- 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: $P(education^+ \mid causeDamage^+ = true)$

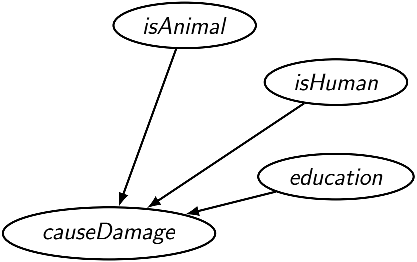
Extended Belief Networks (EBNs)



However...

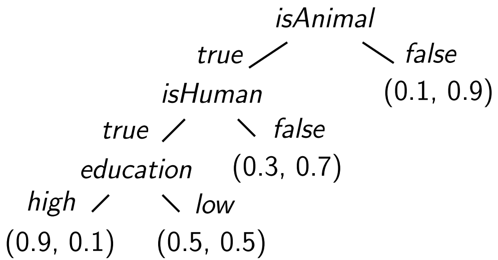
- Expanding ranges is computationally expensive.
  - Exact inference has time complexity $\mathcal{O}(|range|^{treewidth})$.
- It may not be sensible to think about undefined values; no dataset would contain such values.
- Arcs $\langle isAnimal^+, isHuman^+\rangle$ and $\langle isHuman^+, education^+\rangle$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

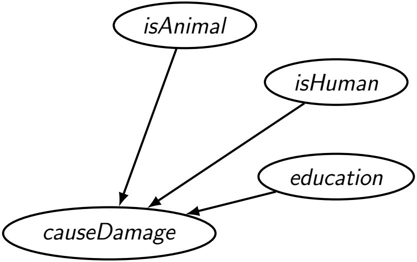
Conditional Probabilities



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

- 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 $\beta = P(Q \mid \mathcal{E} = e \wedge domain(Q))$
  - Return

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

Outline

Hypotheses, Models and Predictions

- Hypotheses are often very narrow.
- We need to 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 is a set of hypotheses applied to a particular case. "ensemble"
    - 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

Programs and Meta-programs

Two sorts of probabilistic programs:

- Hypotheses are probabilistic programs that persist, are tuned to data. Often very narrow.
- 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

Semantic Science Search Engine

Semantic Science Search Engine:

- Given a hypothesis, find data about which it makes predictions.
- Given a dataset, find hypotheses which make predictions on the dataset
- Given a new problem, find the best model (ensemble of hypotheses)

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 invents useful distinctions (latent features)
  $\longrightarrow$ 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.

Zero Probabilities

What do the following have in common?

- Ozone hole over Antarctica (1976-1985)
- Robot kidnap problem

Zero Probabilities

What do the following have in common?

- Ozone hole over Antarctica (1976-1985)
- Robot kidnap problem
  $\longrightarrow$ don't use zero probabilities for anything possible.

Zero Probabilities

What do the following have in common?

- Ozone hole over Antarctica (1976-1985)
- Robot kidnap problem
  $\longrightarrow$ don't use zero probabilities for anything possible.
- International Astronomical Union (IAU) in 2006 defined "planet" so Pluto is not a planet.
- Is there a dataset that says "Justin is an Mammal", "Justin is an animal" or "Justin is a holozoa"?
- What about "Justin is person but not an animal"?

Zero Probabilities

What do the following have in common?

- Ozone hole over Antarctica (1976-1985)
- Robot kidnap problem
  $\longrightarrow$ don't use zero probabilities for anything possible.
- International Astronomical Union (IAU) in 2006 defined "planet" so Pluto is not a planet.
- Is there a dataset that says "Justin is an Mammal", "Justin is an animal" or "Justin is a holozoa"?
- What about "Justin is person but not an animal"?
  $\longrightarrow$ all zero probabilities come from definitions.
  Ontologies give definitions — data that is inconsistent is rejected.
  Clarity principle. Clear definitions are useful!

More issues

- How can we stop people from publishing fictional data?

More issues

- How can we stop people from publishing fictional data? Standard hypotheses: data is just noise (null hypothesis), data is fake, . . .

More issues

- How can we stop people from publishing fictional data? Standard hypotheses: data is just noise (null hypothesis), data is fake, . . .
- If all data is published, how can we test hypotheses if there is no “held-out” data? (Won’t everyone cheat?)

More issues

- How can we stop people from publishing fictional data? Standard hypotheses: data is just noise (null hypothesis), data is fake, . . .
- If all data is published, how can we test hypotheses if there is no “held-out” data? (Won’t everyone cheat?)
- Meta-predictions: Will maximally informed priors or maximally uninformed priors win?

More issues

- How can we stop people from publishing fictional data?
  Standard hypotheses: data is just noise (null hypothesis), data is fake, . . .
- If all data is published, how can we test hypotheses if there is no "held-out" data? (Won't everyone cheat?)
- Meta-predictions: Will maximally informed priors or maximally uninformed priors win?
- How can we get there?
  Start in very narrow domains
  Few hypotheses, published data....

More issues

- How can we stop people from publishing fictional data?
  Standard hypotheses: data is just noise (null hypothesis), data is fake, . . .
- If all data is published, how can we test hypotheses if there is no “held-out” data? (Won’t everyone cheat?)
- Meta-predictions: Will maximally informed priors or maximally uninformed priors win?
- How can we get there?
  Start in very narrow domains
  Few hypotheses, published data....
- Users should be able to express data and hypotheses in their own terms. They shouldn’t have to be an expert in domain and statistics and (probabilistic) programming....
  They must see a value in representing data / hypotheses.

Conclusion

- Semantic science is a way to develop and deploy knowledge about how the world works.

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.

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.

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.

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”

To Do

- Representing, reasoning and learning complex (probabilistic) hypotheses. "probabilistic programming"
- Representations for observations that interacts with hypotheses.

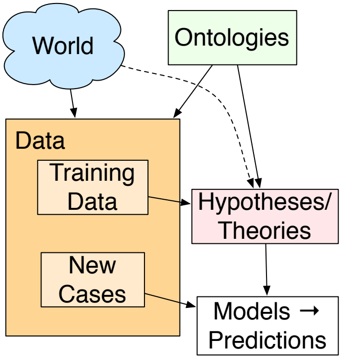
To Do

- 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.

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
- More complex models, e.g., for relational reinforcement learning where individuals are created and destroyed

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.