Semantic Science

David Poole

Department of Computer Science,
University of British Columbia

Work with: http://minervaintelligence.com, https://treatment.com/

April 3, 2019

There is a real world with real structure. The program of mind has been trained on vast interaction with this world and so contains code that reflects the structure of the world and knows how to exploit it. This code contains representations of real objects in the world and represents the interactions of real objects. . . .

You exploit the structure of the world to make decisions and take actions. Where you draw the line on categories, what constitutes a single object or a single class of objects for you, is determined by the program of your mind, which does the classification. This classification is not random but reflects a compact description of the world, and in particular a description useful for exploiting the structure of the world.

Eric Baum, *What is Thought?*, 2004, pages 169-170

Outline

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

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.
- Probability provides a calculus for how knowledge (observations) affects belief.

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.
- Probability provides a calculus for how knowledge (observations) affects belief. Bayes' rule:

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.
- Probability provides a calculus for how knowledge (observations) affects belief. Bayes' rule:

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

- What if e is a patient's symptoms and history, and h is the effect of a particular treatment on a particular patient?

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.
- Probability provides a calculus for how knowledge (observations) affects belief. Bayes' rule:

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

- What if e is a patient's symptoms and history, and h is the effect of a particular treatment on a particular patient?
- What if e is the electronic health records for all of the people in the province?

Informed decision making

- Acting in the world is gambling.
  Probability is the calculus of gambling.
- Probability provides a calculus for how knowledge (observations) affects belief. Bayes' rule:

$$P(h|e) = \frac{\overbrace{P(e|h)}^{\text{Likelihood}}\ \overbrace{P(h)}^{\text{Prior}}}{\underbrace{P(e)}_{\text{Normalizing constant}}}$$

- What if e is a patient's symptoms and history, and h is the effect of a particular treatment on a particular patient?
- What if e is the electronic health records for all of the people in the province?
- What if e is everything known about the geology of Earth?

Example: Decision making in Medicine

A patient walls into a GPs office....

Example: Decision making in Medicine

A patient walls into a GPs office....

Inputs	Outputs
	Top Diagnoses
	Suggested Tests
	Suggested Treatments
	. . . with justifications

Example: Decision making in Medicine

A patient walls into a GPs office....

Inputs	Outputs
Patient's complaint (reason for encounter)	Top Diagnoses
Receptionist's and Doctor's observations	Suggested Tests
Patient's History (EHR)	Suggested Treatments
Test results	. . . with justifications

Example: Decision making in Medicine

A patient walls into a GPs office....

Inputs	Outputs
Patient's complaint (reason for encounter)	Top Diagnoses
Receptionist's and Doctor's observations	Suggested Tests
Patient's History (EHR)	Suggested Treatments
Test results	. . . with justifications
Patient's preferences/utilities	

Example: Decision making in Medicine

A patient walls into a GPs office....

Inputs	Outputs
Patient's complaint (reason for encounter)	Top Diagnoses
Receptionist's and Doctor's observations	Suggested Tests
Patient's History (EHR)	Suggested Treatments
Test results	. . . with justifications
Patient's preferences/utilities	
Standardized vocabulary (ontologies)	
Best practices	
Latest Research Results	
Data from every other patient	

Example: Decision making in Medicine

A patient walls into a GPs office....

Inputs	Outputs
Patient's complaint (reason for encounter)	Top Diagnoses
Receptionist's and Doctor's observations	Suggested Tests
Patient's History (EHR)	Suggested Treatments
Test results	. . . with justifications
Patient's preferences/utilities	
Standardized vocabulary (ontologies)	
Best practices	
Latest Research Results	
Data from every other patient	

We want to make decisions conditioned on all of the information in the world

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
    - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, . . .

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
  - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, ...
  - provided because it is unusual (not sampled at random)

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
  - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, . . .
  - provided because it is unusual (not sampled at random)
  - at multiple levels of abstraction, in terms of more general or less general terms (e.g., "broken leg" vs "fractured leg")

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
  - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, ...
  - provided because it is unusual (not sampled at random)
  - at multiple levels of abstraction, in terms of more general or less general terms (e.g., "broken leg" vs "fractured leg")
  - at multiple level of detail, in terms of parts and subparts (e.g., "broken leg" vs "broken femur")

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
  - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, . . .
  - provided because it is unusual (not sampled at random)
  - at multiple levels of abstraction, in terms of more general or less general terms (e.g., "broken leg" vs "fractured leg")
  - at multiple level of detail, in terms of parts and subparts (e.g., "broken leg" vs "broken femur")
- Consider predicting the amount of a particular mineral at a particular location

Motivation

- Consider predicting the effect of a treatment on a particular patient in a GP's office. Information is:
  - heterogenous, provided from many sources at multiple points in time. E.g., from patient reports, nurse observation, doctor observation, lab tests, x-rays, . . .
  - provided because it is unusual (not sampled at random)
  - at multiple levels of abstraction, in terms of more general or less general terms (e.g., "broken leg" vs "fractured leg")
  - at multiple level of detail, in terms of parts and subparts (e.g., "broken leg" vs "broken femur")
- Consider predicting the amount of a particular mineral at a particular location
- Consider predicting whether a particular person will like a particular apartment

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations
- Relational, identity and existence uncertainty

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,...) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,...) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both
- There is lots of expert and textbook knowledge (that may be wrong)

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both
- There is lots of expert and textbook knowledge (that may be wrong)
- We want to use whatever evidence we can get, to learn from experience (but current EHRs are terrible).

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,...) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both
- There is lots of expert and textbook knowledge (that may be wrong)
- We want to use whatever evidence we can get, to learn from experience (but current EHRs are terrible).
- We need to justify recommendations

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both
- There is lots of expert and textbook knowledge (that may be wrong)
- We want to use whatever evidence we can get, to learn from experience (but current EHRs are terrible).
- We need to justify recommendations
- Always base decisions on best available evidence.

Challenges

- Problem is inherently relational: many types of objects (patients, body parts, tests, infections,. . .) and relations
- Relational, identity and existence uncertainty
- We need to interact with standardized vocabularies. E.g., SNOMED-CT has 350,000 medical concepts
- Sparse data: for almost every pair of symptoms, pair of diseases, or disease-treatment pair, *no one* in the world has both
- There is lots of expert and textbook knowledge (that may be wrong)
- We want to use whatever evidence we can get, to learn from experience (but current EHRs are terrible).
- We need to justify recommendations
- Always base decisions on best available evidence.
- Transportability: learn in Vancouver, apply in Beijing

Example: Medicine

- PubMed comprises over 29 million citations for biomedical literature. 10,000 added each week.

Example: Medicine

- PubMed comprises over 29 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.

Example: Medicine

- PubMed comprises over 29 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?

Example: Medicine

- PubMed comprises over 29 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

Example: Geology

- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
  Geology doesn't change at arbitrary political boundaries

Example: Geology

- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
  Geology doesn't change at arbitrary political boundaries
- Geological "observations" are published by the geological surveys of counties and states/provinces and globally (onegeology.org)

Example: Geology

- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
  Geology doesn't change at arbitrary political boundaries
- Geological "observations" are published by the geological surveys of counties and states/provinces and globally (onegeology.org)
- Geological hypotheses are published in research journals.

Example: Geology

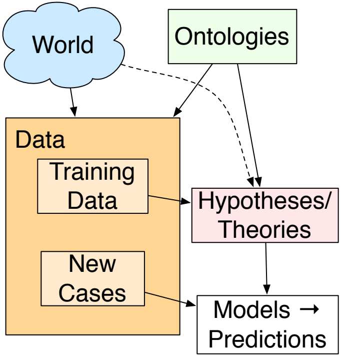
- Geologists know they need to make decisions under uncertainty
- Geologists know they need ontologies
  Geology doesn't change at arbitrary political boundaries
- Geological "observations" are published by the geological surveys of counties and states/provinces and globally (onegeology.org)
- Geological 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, Minerva Intelligence]

OneGeology.org

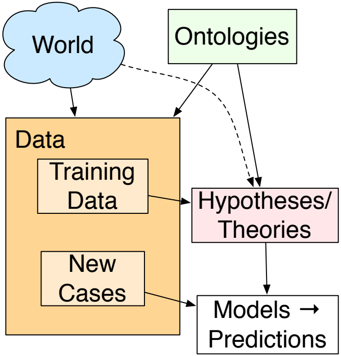
OneGeology.org

Semantic Science



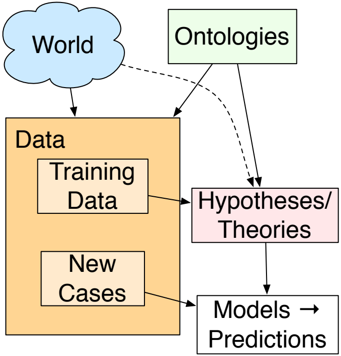
- Ontologies represent the meaning of symbols.

Semantic Science



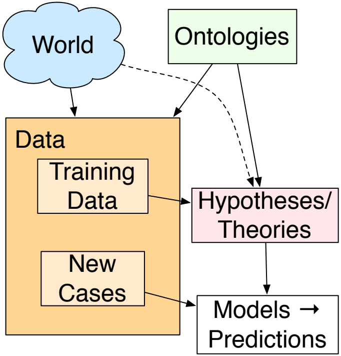
- Ontologies represent the meaning of symbols.
- Observational data describes world using symbols defined in ontology.

Semantic Science



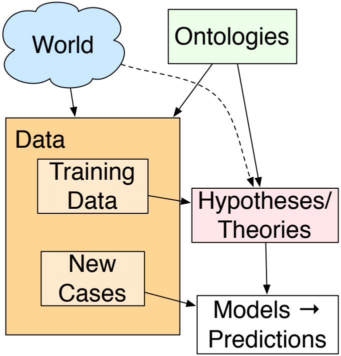
- Ontologies represent the meaning of symbols.
- Observational data describes world using symbols defined in ontology.
- Hypotheses make predictions on data.

Semantic Science



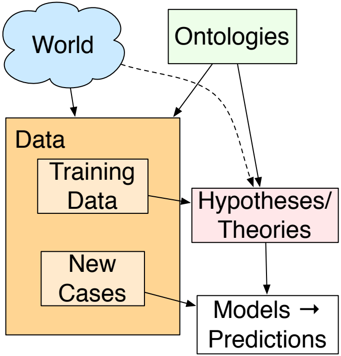
- Ontologies represent the meaning of symbols.
- Observational data describes world using symbols defined in ontology.
- Hypotheses make predictions on data.
- Data used to evaluate hypotheses.

Semantic Science



- Ontologies represent the meaning of symbols.
- Observational data describes world using symbols defined in ontology.
- Hypotheses make predictions on data.
- Data used to evaluate hypotheses.
- Hypotheses used for predictions on new cases.

Semantic Science



- Ontologies represent the meaning of symbols.
- Observational data describes world using symbols defined in ontology.
- 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.

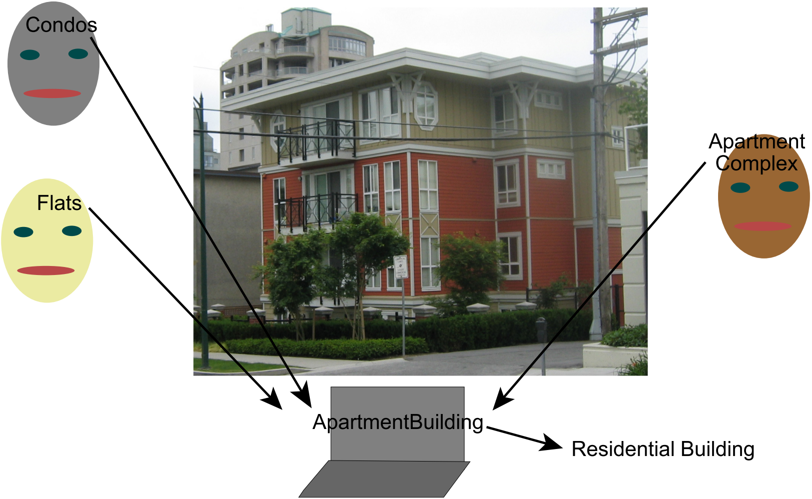
Ontologies

- 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.
- SNOMED-CT is a medical ontology with 349,548 concepts (January 31, 2019 release) in multiple languages

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.
- SNOMED-CT is a medical ontology with 349,548 concepts (January 31, 2019 release) in multiple languages
- Our geology ontology has 6022 minerals + 266 rocks in a "simplified" rock taxonomy + time + ...

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)

Main Components of an Ontology

- Individuals: the objects in the world
  (not usually specified as part of the ontology)
- Classes: sets of (potential) individuals.
  E.g., class of buildings is the set of things that would be apartment buildings (even those not yet built)

Main Components of an Ontology

- Individuals: the objects in the world
  (not usually specified as part of the ontology)
- Classes: sets of (potential) individuals.
  E.g., class of buildings is the set of things that would be apartment buildings (even those not yet built)
- Properties: between individuals and their values

Main Components of an Ontology

- Individuals: the objects in the world
  (not usually specified as part of the ontology)
- Classes: sets of (potential) individuals.
  E.g., class of buildings is the set of things that would be apartment buildings (even those not yet built)
- 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**.

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

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

- All classes are defined in terms of properties.

Outline

1. Motivation
   - Ontologies
   - Data
   - Hypotheses
2. Semantic Science
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

Data

Real data is messy!

- Multiple levels of abstraction
- Multiple levels of detail
- Uses the vocabulary from many ontologies: rocks, minerals, top-level ontology,. . .

Data

Real data is messy!

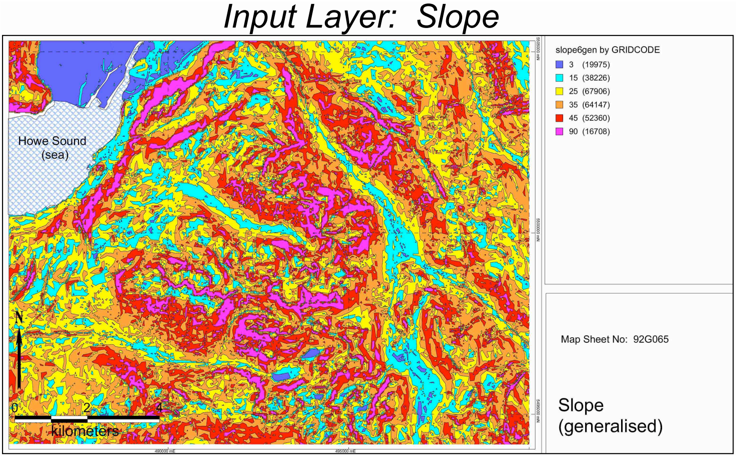
- Multiple levels of abstraction
- Multiple levels of detail
- Uses the vocabulary from many ontologies: rocks, minerals, top-level ontology,. . .
- Rich meta-data:
  - Who collected each datum? (identity and credentials)
  - Who transcribed the information?
  - What was the protocol used to collect 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?

Data

Real data is messy!

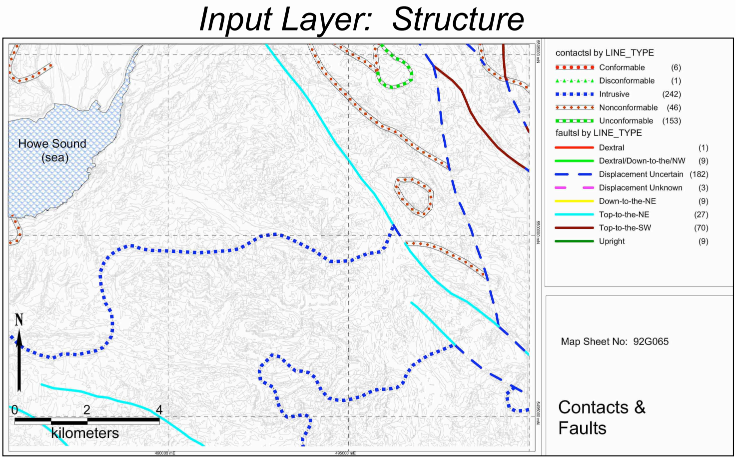
- Multiple levels of abstraction
- Multiple levels of detail
- Uses the vocabulary from many ontologies: rocks, minerals, top-level ontology,. . .
- Rich meta-data:
  - Who collected each datum? (identity and credentials)
  - Who transcribed the information?
  - What was the protocol used to collect 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



[Clinton Smyth, Minerva Intelligence]

Example Data, Geology



[Clinton Smyth, Minerva Intelligence]

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!)

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.

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 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.

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.

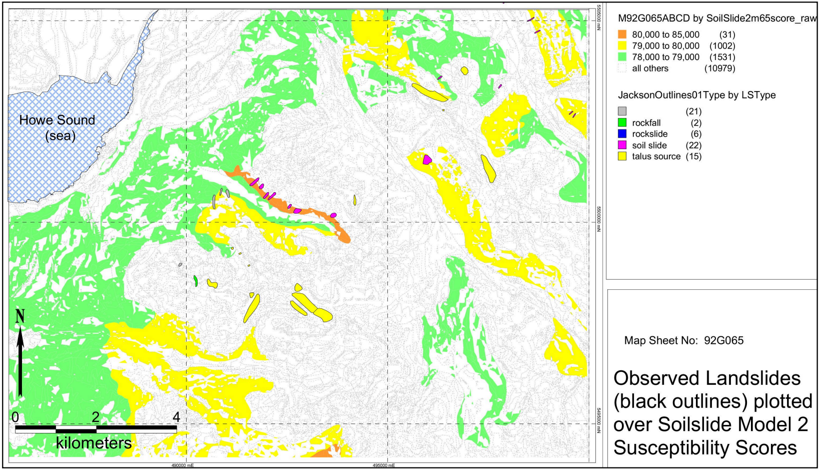
Outline

Hypotheses make predictions on data

- Hypotheses are programs that make predictions on data.
- To be useful for decision making, predictions should be probabilistic.
  $\longrightarrow$ probabilistic programs

Example Prediction from a Hypothesis

Test Results: Model SoilSlide02



[Clinton Smyth, Minerva Intelligence]

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
- Property R is functional means
  $\langle x, R, y_1 \rangle$ and $\langle x, R, y_2 \rangle$ implies $y_1 = y_2$.

Random Variables and Triples

- Reconcile:
  - random variables (RVs) of probability theory
  - individuals, classes, properties of modern ontologies
- Property R is functional means
  $\langle x, R, y_1 \rangle$ and $\langle x, R, y_2 \rangle$ implies $y_1 = y_2$.
- For **functional properties**:
  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
- Property R is functional means
  $\langle x, R, y_1 \rangle$ and $\langle x, R, y_2 \rangle$ implies $y_1 = y_2$.
- For **functional properties**:
  random variable for each $\langle individual, property \rangle$ pair,
  range of the RV is range of the property.
  E.g., if *Height* is functional, $\langle building17, Height \rangle$ is a RV.
- For **non-functional properties**:
  Boolean RV for each $\langle individual, property, value \rangle$ triple.
  E.g., if *YearRestored* is non-functional
  $\langle building17, YearRestored, 1988 \rangle$ is a Boolean RV.

Probabilities and Aristotelian Definitions

Aristotelian definition

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

leads to probability over class membership

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

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

Outline

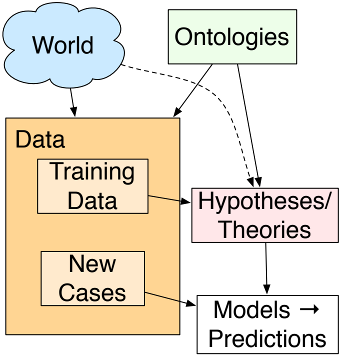
Semantic Science

- Governments are publishing data with rich ontologies.
  Journals are forcing authors to publish data.
  - European Union is mandating that all levels of government in EU publish all spatial (map) data using standardized vocabularies (INSPIRE `https://inspire.ec.europa.eu/`)

Semantic Science

- Governments are publishing data with rich ontologies.
  Journals are forcing authors to publish data.
    - European Union is mandating that all levels of government in EU publish all spatial (map) data using standardized vocabularies (INSPIRE `https://inspire.ec.europa.eu/`)
- Idea: also publish hypotheses that make (probabilistic) predictions.
  These must interact with standardized vocabularies

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.

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.

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.

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

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 a 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 a 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?)
- 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?)
- 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.

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

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

To make a prediction, multiple hypotheses need to be used together in a model.

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, and allowing a program to compute the 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 $P(Lives \mid person \wedge coughs)$:

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

Outline

1. Motivation
   - Ontologies
   - Data
   - Hypotheses

2. Semantic Science

3. Models: Ensembles of hypotheses

4. Property Domains and Undefined Random Variables

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, \mathit{domain}, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, \mathit{type}, C\rangle$

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- A property is almost always undefined:
  - *weight* is only defined for

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- A property is almost always undefined:
  - *weight* is only defined for physical objects
  - *pitch* is only defined for

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- A property is almost always undefined:
  - *weight* is only defined for physical objects
  - *pitch* is only defined for sounds
  - *wavelength* is only defined for

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- 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

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- 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

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- 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

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- 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

Properties, Domains and Undefined Random Variables

- Properties have domains.
- A property is only defined for individuals in its domain:
  If $\langle P, domain, C\rangle$ and $\langle i, P, j\rangle$ then $\langle i, type, C\rangle$
- 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}
```

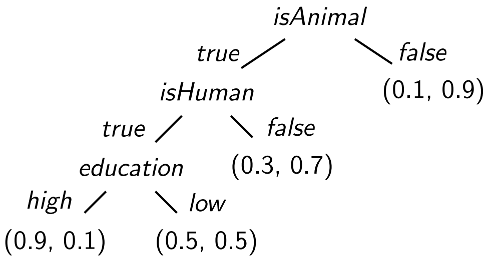
education is not defined when $isHuman = false$.

Well-defined Formulae

Well-defined conjunctions:

- $isAnimal = true \wedge isHuman = false$
  is well-defined.
- $isHuman = true \wedge isAnimal = false$
  is not well-defined.
- $isAnimal = true \wedge isHuman = true \wedge education = low$
  is well-defined.
- $isAnimal = true \wedge isHuman = false \wedge education = low$
  is not well-defined.

Conditional Probabilities

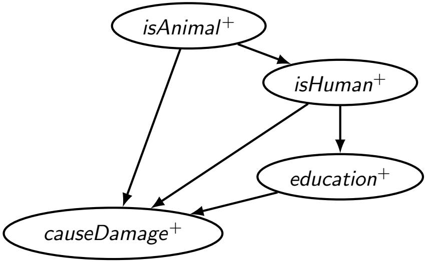


$$P(causeDamage \mid isAnimal, isHuman, education)$$

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

Extended Belief Networks (EBNs)

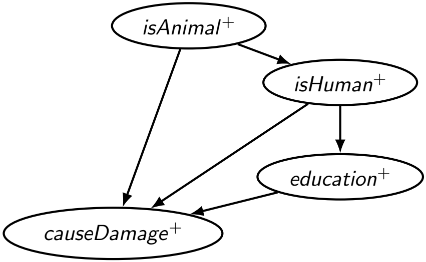
- Add "undefined" ($\perp$) to each range.
  - $range(isHuman^{+}) = \{true, false, \perp\}$.
  - $range(education^{+}) = \{low, high, \perp\}$.



- $education^{+}$ is like *education* but with an expanded range.

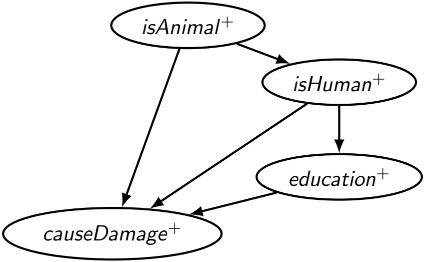
Extended Belief Networks (EBNs)

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

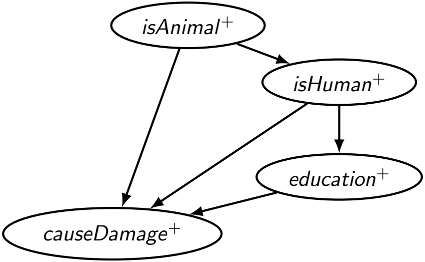
Extended Belief Networks (EBNs)



However...

- Expanding ranges is computationally expensive.
  - Exact inference has time complexity $\mathcal{O}(|range|^{treewidth})$.

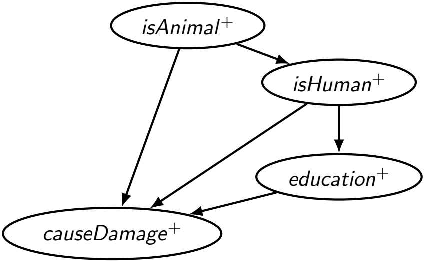
Extended Belief Networks (EBNs)



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.

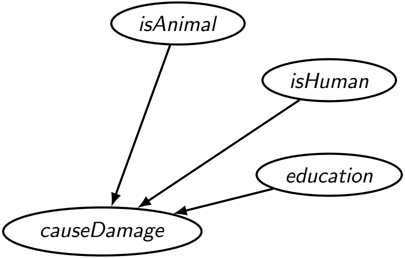
Extended Belief Networks (EBNs)



However...

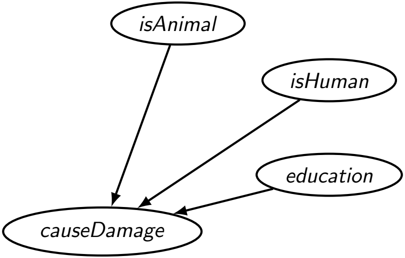
- Expanding ranges is computationally expensive.
  - Exact inference has time complexity $\mathcal{O}(|range|^{treewidth})$.
- It may not be sensible to think about undefined values; no dataset would contain such values.
- Arcs $\langle isAnimal^+, isHuman^+\rangle$ and $\langle isHuman^+, education^+\rangle$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

Ontologically-Based Belief Networks (OBBNs)



- The query $P(education^{+} \mid causeDamage = true)$ 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

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

Conclusion

- Rich history of probabilistic models of relational data
- Semantic science is a way to develop and deploy knowledge about how the world works.

Conclusion

- Rich history of probabilistic models of relational data
- 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

- Rich history of probabilistic models of relational data
- 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.
  - Justify predictions by hypotheses used
  - Justify hypotheses by relavant evidence

Conclusion

- Rich history of probabilistic models of relational data
- 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.
  - Justify predictions by hypotheses used
  - Justify hypotheses by relavant evidence
- Ontologies, hypotheses and observations interact in complex ways.

Conclusion

- Rich history of probabilistic models of relational data
- 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.
  - Justify predictions by hypotheses used
  - Justify hypotheses by relavant evidence
- 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.

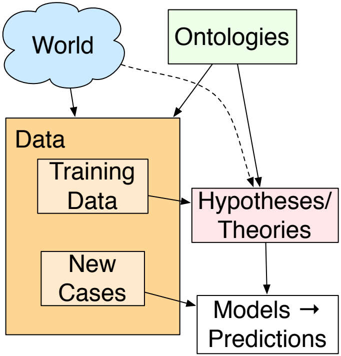
To Do

- 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

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.

What is now required is to give the greatest possible development to mathematical logic, to allow to the full the importance of relations, and then to found upon this secure basis a new philosophical logic, which may hope to borrow some of the exactitude and certainty of its mathematical foundation. If this can be successfully accomplished, there is every reason to hope that the near future will be as great an epoch in pure philosophy as the immediate past has been in the principles of mathematics. Great triumphs inspire great hopes; and pure thought may achieve, within our generation, such results as will place our time, in this respect, on a level with the greatest age of Greece.

– Bertrand Russell 1917