Semantic Science and Machine-Accessible Scientific Theories

David Poole* Clinton Smyth† Rita Sharma†

*Department of Computer Science, University of British Columbia
†Georeference Online

March 2008

Outline

1. Semantic Science Vision
   - Ontologies
   - Data
   - Theories
2. Theoretical Foundations
   - Probabilistic Prediction
   - Probabilities with Ontologies
   - Existence and Identity Uncertainty
3. Fielded Systems

Example: medical diagnosis

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

Current Practice (Google)	Vision
— describe symptoms using keywords — results ranked by popularity (pagerank) — text results	

Example: medical diagnosis

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

Current Practice (Google)	Vision
— describe symptoms using keywords	— use ontologies
— results ranked by popularity (pagerank)	
— text results	

Example: medical diagnosis

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

Current Practice (Google)	Vision
— describe symptoms using keywords	— use ontologies
— results ranked by popularity (pagerank)	— theories ranked by relevance and fit to data
— text results	

Example: medical diagnosis

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

Current Practice (Google)	Vision
— describe symptoms using keywords	— use ontologies
— results ranked by popularity (pagerank)	— theories ranked by relevance and fit to data
— text results	— probabilistic predictions with references to sources

Example: finding a location that contains gold

Given a model of where gold can be found and 25000 location descriptions:

Current Practice	Vision
— keyword database look-up	— describe model using ontology
— results (if any) unranked or ranked by popularity	— results ranked by fit to model
— text	— probabilistic prediction
— repeat for more and less general terms	

Example: finding minerals at a location

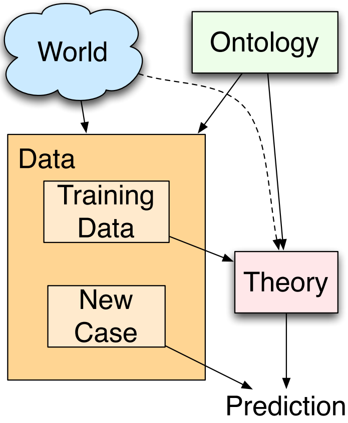
Given one location and 100 models of where minerals can be found:

Current Practice ????	Vision
— keyword database look-up	— describe location and models using ontology
— results (if any) unranked or ranked by popularity	— results ranked by relevance and fit to data
— text	— probabilistic prediction with references

Notational Minefield

- Theory / hypothesis / model / law (Science)
- Variable (probability and logic and programming languages)
- Model (science, probability and logic)
- Parameter (mathematics and statistics)
- Domain (science and logic and probability and mathematics)
- Object/class (object-oriented programming and ontologies)
- $=$ (probability and logic)
- First-order (logic and dynamical systems)

Our Semantic Science Vision



- Ontologies represent the meaning of symbols.
- Data that adheres to an ontology is published.
- Theories that make (probabilistic) predictions on data are published.
- Data can be used to evaluate theories.
- Theories make predictions on new cases.

AI Traditions

- Expert Systems of 70's and 80's (e.g., Prospector '74-83)
  - Probabilistic models and machine learning.
    Bayesian networks, Bayesian X...
  - Ontologies and Knowledge Representations.
    Description logic, X logic...

Science in Broadest Sense

We 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 house Mary would like

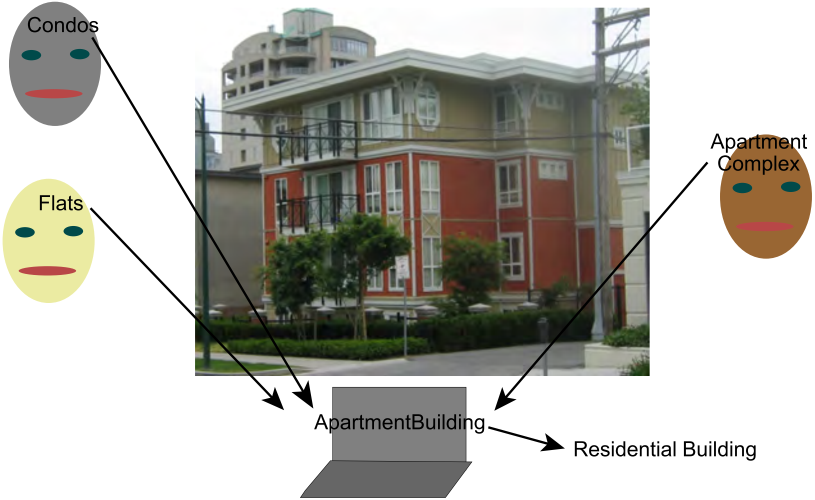
Outline

1. Semantic Science Vision
   - Ontologies
   - Data
   - Theories
2. Theoretical Foundations
   - Probabilistic Prediction
   - Probabilities with Ontologies
   - Existence and Identity Uncertainty
3. Fielded Systems

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

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**.

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

  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 can be defined in terms of properties.

Outline

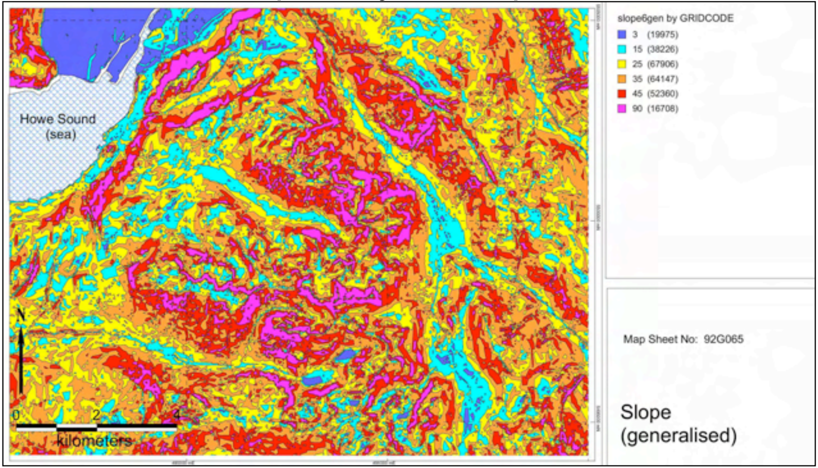
Data

Real data is messy!

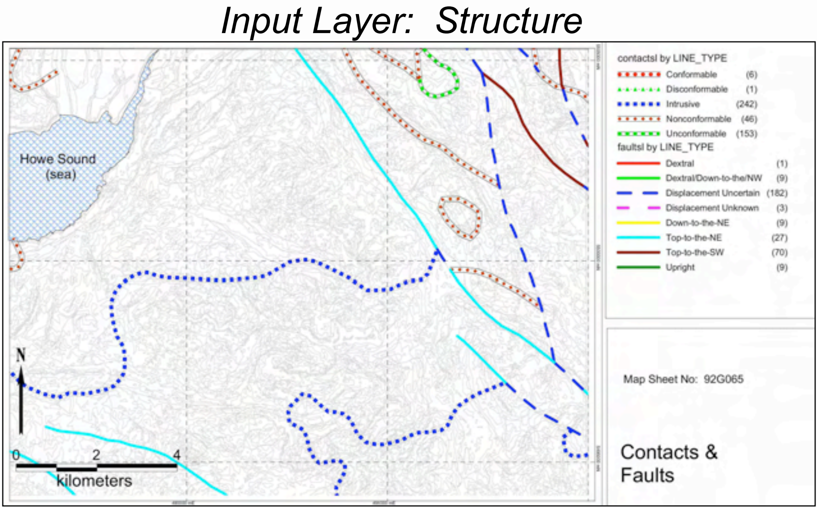
- Multiple levels of abstraction
- Multiple levels of detail
- Uses the vocabulary from many ontologies: rocks, minerals, top-level ontology,...
- Rich meta-data:
  - Who collected each datum? (identity and credentials)
  - Who transcribed the information?
  - What was the protocol used to collect the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

Example Data in Geology (I)

Input Layer: Slope



Example Data in Geology (II)



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 Vision
   - Ontologies
   - Data
   - Theories
2. Theoretical Foundations
   - Probabilistic Prediction
   - Probabilities with Ontologies
   - Existence and Identity Uncertainty
3. Fielded Systems

Theories make predictions on data

A theory is a procedure that makes a prediction on data.

- Theories can make whatever predictions they like about data:
  - definitive predictions
  - point probabilities
  - probability ranges
  - ranges with confidence intervals
  - qualitative predictions
- For each prediction type, we need ways to judge predictions on data
- Users can use whatever criteria they like to evaluate theories (e.g., taking into account simplicity and elegance)

Theory Ensembles

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

Theory Ensembles

- How can we compare theories that differ in their generality?
- Theory A makes predictions about all cancers.
  Theory 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 theory C: *if lung cancer, use B's prediction, else use A's prediction*?

Theory Ensembles

- How can we compare theories that differ in their generality?
- Theory A makes predictions about all cancers.
  Theory 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 theory C: *if lung cancer, use B's prediction, else use A's prediction*?
- Proposal: make theory ensembles the norm.
  - Judge theories by how well they fit into ensembles.
  - Ensembles can be judged by simplicity.
  - Theory designers don't need to game the system by manipulating the generality of theories

Dynamics of Semantic Science

- 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 theories:
  A theory 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 theories that use them
  — the role of the vocabulary is to describe useful distinctions.

Outline

1. Semantic Science Vision
   - Ontologies
   - Data
   - Theories
2. Theoretical Foundations
   - Probabilistic Prediction
   - Probabilities with Ontologies
   - Existence and Identity Uncertainty
3. Fielded Systems

Why Probabilistic Prediction?

- Probabilities are what you get from data.
  (Most suggested measures of prediction accuracy are optimized by probabilistic prediction!)
- There is a well defined procedure for combining background knowledge with data (conditioning).
- Probabilities are what is needed (with utilities) to make decisions.

Probabilistic Prediction

- The role of models in prediction:
  Given a description of a new case,

$$P(prediction|description) = \sum_{m \in Models} \left(\begin{array}{l} P(prediction|m\&description) \times \\ P(m|description) \end{array} \right)$$

  Models is a set of mutually exclusive and covering set of hypotheses.

Probabilistic Prediction

- The role of models in prediction:
  Given a description of a new case,

$$
\begin{aligned}
& P(prediction \mid description) \\
= & \sum_{m \in Models} \left(\begin{array}{l} P(prediction \mid m \& description) \times \\ P(m \mid description) \end{array} \right)
\end{aligned}
$$

  Models is a set of mutually exclusive and covering set of hypotheses.
- What features of the description are predictive?
- How do the features interact?
- What are the appropriate probabilities? (How can these be learned with limited data?)

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:
  random variable for each $\langle \mathit{individual}, \mathit{property} \rangle$ pair, where the domain of the random variable is the range of the property.
- For non-functional properties:
  Boolean random variable for each $\langle \mathit{individual}, \mathit{property}, \mathit{value} \rangle$ triple.

First-order probabilistic models

- Individuals are not known until run time.
- Therefore the random variables are not known until run time (and they change for each situation).
- We want to build the models before we know the random variables.

First-order probabilistic models

- Individuals are not known until run time.
- Therefore the random variables are not known until run time (and they change for each situation).
- We want to build the models before we know the random variables.

$\longrightarrow$ First-order probabilistic models

- Idea: if you are a Bayesian, you need to treat every individual that you have the same knowledge about the same (exchangability).
- Probabilities are specified for all individuals.

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 property values

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

Type uncertainty $\longrightarrow$ uncertainty over property values.

Outline

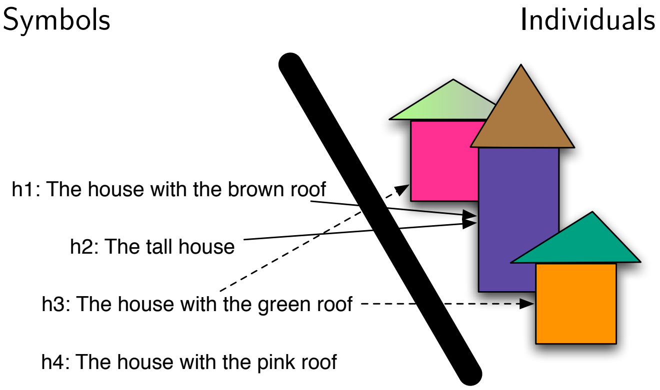
Existence and Identity Uncertainty

Theory about what house Mary would like:

Whether Mary likes an house depends on:

- Whether there is a bedroom for daughter Sam
- Whether Sam's room is green
- Whether there is a bedroom for Mary
- Whether Mary's room is large
- Whether they share

Existence and Identity



Clarity Principle

Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \wedge roof\_colour(h4, pink) \wedge \neg exists(h4)$

Clarity Principle

Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \wedge roof\_colour(h4, pink) \wedge \neg exists(h4)$ ✘

  Want: probability that there exists an object that matches some description. Name the the object that exists.

Clarity Principle

Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \wedge roof\_colour(h4, pink) \wedge \neg exists(h4)$ ✘

  Want: probability that there exists an object that matches some description. Name the the object that exists.
- What if more than one individual exists? Which one are we referring to?
  — In a house with three bedrooms, which is the second bedroom?

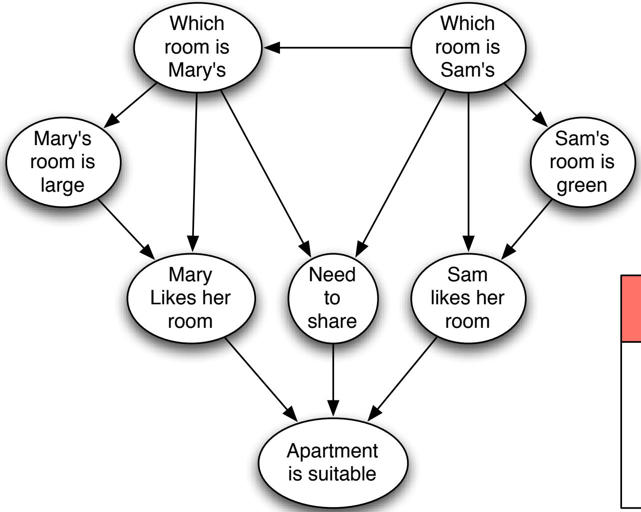
Clarity Principle

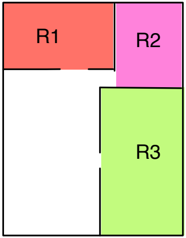
Clarity principle: probabilities must be over well-defined propositions.

- What if an individual doesn't exist?
  - $house(h4) \wedge roof\_colour(h4, pink) \wedge \neg exists(h4)$ ✘

  Want: probability that there exists an object that matches some description. Name the the object that exists.
- What if more than one individual exists? Which one are we referring to?
  — In a house with three bedrooms, which is the second bedroom?
- Note: Reified individuals are special:
  — Non-existence means the relation is false.
  — Well defined what doesn't exist when existence is false.
  — Same description implies the same individual.

Role assignments





Outline

Expert Models

What if the models are provided by the experts in the field?

- not covering — only provide positive models
- not exclusive — they are often refinements of each other
- described at various levels of abstraction and detail
- often the experts don't know the probabilities and there is little data to estimate them

Providing Probabilities

Experts are reluctant to give probabilities:

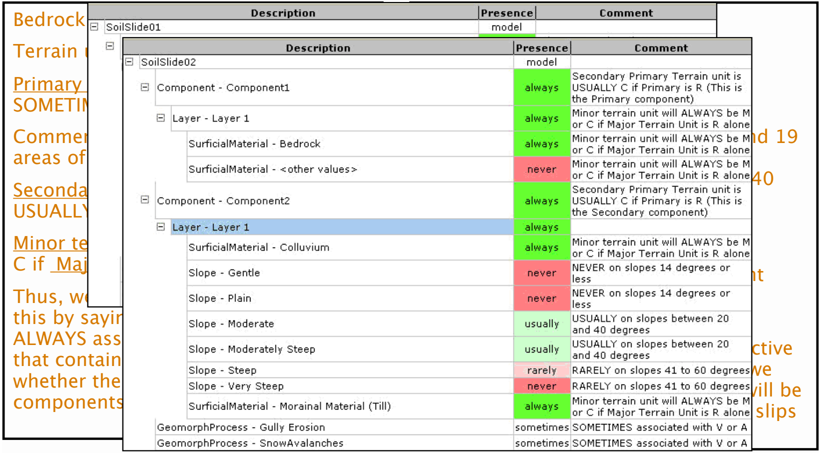
- No data from which to estimate them
- People who want to make decision use more information than provided in our theories
- Difficult to combine marginal probabilities with new information to make decisions
- It is *not* because decision theory is inappropriate. Decision makers use probabilities and utilities.

What we do

- Use qualitative probabilities: {*always*, *usually*, *sometimes*, *rarely*, *never*}.
- With thousands of instances and hundreds of models, find the most likely and the rationale.
- Independence assumptions.

Example Model

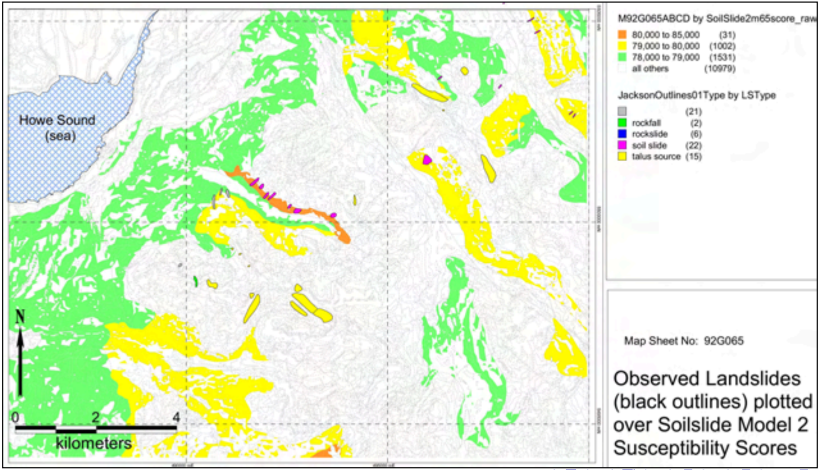
Prototype SoilSlide Model (Jackson, 2007)



Description	Presence	Comment
SoilSlide02	model	
Component - Component1	always	Secondary Primary Terrain unit is USUALLY C if Primary is R (This is the Primary component)
Layer - Layer 1	always	Minor terrain unit will ALWAYS be M or C if Major Terrain Unit is R alone
SurficialMaterial - Bedrock	always	Minor terrain unit will ALWAYS be M or C if Major Terrain Unit is R alone
SurficialMaterial - <other values>	never	Minor terrain unit will ALWAYS be M or C if Major Terrain Unit is R alone
Component - Component2	always	Secondary Primary Terrain unit is USUALLY C if Primary is R (This is the Secondary component)
Layer - Layer 1	always	
SurficialMaterial - Colluvium	always	Minor terrain unit will ALWAYS be M or C if Major Terrain Unit is R alone
Slope - Gentle	never	NEVER on slopes 14 degrees or less
Slope - Plain	never	NEVER on slopes 14 degrees or less
Slope - Moderate	usually	USUALLY on slopes between 20 and 40 degrees
Slope - Moderately Steep	usually	USUALLY on slopes between 20 and 40 degrees
Slope - Steep	rarely	RARELY on slopes 41 to 60 degrees
Slope - Very Steep	never	RARELY on slopes 41 to 60 degrees
SurficialMaterial - Morainal Material (Till)	always	Minor terrain unit will ALWAYS be M or C if Major Terrain Unit is R alone
GeomorphProcess - Gully Erosion	sometimes	SOMETIMES associated with V or A
GeomorphProcess - SnowAvalanches	sometimes	SOMETIMES associated with V or A

Example Prediction from a Model

Test Results: Model SoilSlide02



Conclusion

- Demand from funders, scientists and users.
- Complementary to Semantic web.
- Representing, reasoning and learning complex probabilistic theories is largely unexplored.
- Still lots of work to be done!

To Do

- Fundamental research on complex probabilistic models.
- Build infrastructure to allow publishing and interaction of ontologies, data, theories, theory ensembles, evaluation criteria, meta-data.
- Build inverse semantic science web:
  - Given a theory, find relevant data
  - Given data, find theory ensembles
  - Given a new case, find relevant theory ensembles with explanations