Introduction into Analysis of Methods and Tools for Hypothesis-Driven Scientific Experiment Support

L.A. Kalinichenko[1] **D.Y. Kovalev**[1] D.A. Kovaleva[2]
O.Y. Malkov[2]

[1] Institute of Informatics Problems
Russian Academy of Sciences

[2] Institute of Astronomy
Russian Academy of Sciences

ACM SIGMOD CHAPTER, Moscow
November 27, 2014

Motivation

- Science is increasingly dependent on data as the core source for discovery:
    - scientific instruments,
    - sensors,
    - simulations,
    - Web or social nets
- Big Data "movement": 3V's, new infrastructures, new algorithms

The basic objective of Data-Intensive Sciences (DIS) is to infer knowledge from the integrated data organized in networked infrastructures such as warehouses, grids, clouds. Open access to large volumes of data therefore becomes a key prerequisite for discoveries in the 21st century.

Motivation

> We have to do better at producing tools to support the whole research cycle – from data capture and data curation to data analysis and data visualization.
>
> — Jim Gray, 2007

> New tools are needed to bring humans into the data-analysis loop at all stages, recognizing that knowledge is often subjective and context-dependent and that some aspects of human intelligence will not be replaced anytime soon by machines.
>
> — Frontiers in Massive Data Analysis, 2013

Motivation

- Data deluge has affected the way scientific experiment is done
- X-informatics were the first to deal with it
- Providing valuable insight into how modern data intensive research is done (scheme, methods, algorithms, etc.)

Main Ideas of Talk

- Hypothesis remains the central research unit in DIS
- The most promising approaches and examples of problem solving in different DIS are collected
- The general scheme to do scientific research and provided possible methods for it is pinpointed
- Examples of complex DIS with various data and algorithm problems are highlighted

The Automation of Systems Biology

Revised Approach to Scientific Experiment

Example

This is a physically implemented laboratory automation system that exploits techniques from the field of artificial intelligence to execute cycles of scientific experiment

Adam formulated
and tested 20 hypotheses
concerning genes encoding
13 orphan enzymes . . . 12
hypotheses with no previous
evidence were confirmed.

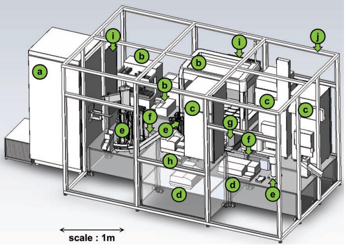
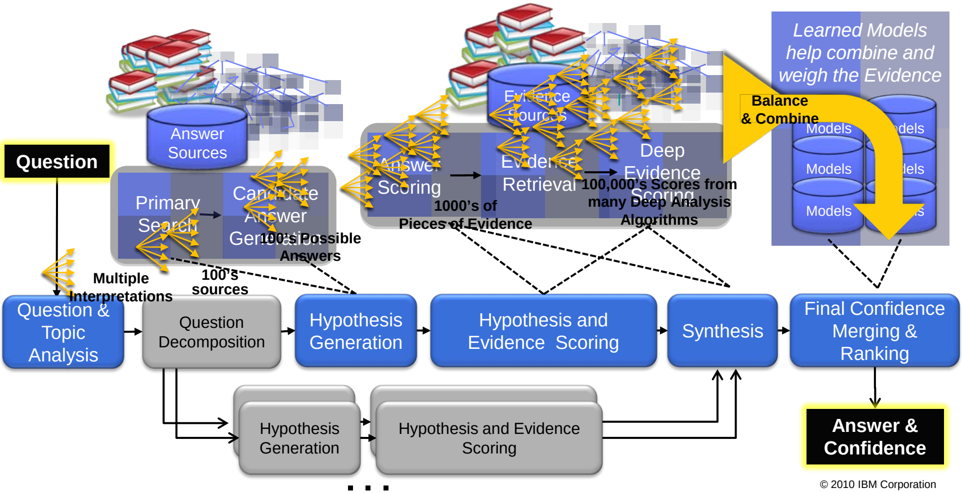


Figure: Lab equipment

DeepQA: The architecture underlying Inside Watson

*Generates many hypotheses, **collects a wide range of evidence** and balances the combined confidences of **over 100 different analytics that analyze the evidence form different dimensions***



Learned Models help combine and weigh the Evidence

Balance & Combine

Models

Models

Models

Answer Sources

Evidence Sources

Question

Primary Search

Candidate Answer Generation

100's Possible Answers

Answer Scoring

1000's of Pieces of Evidence

Evidence Retrieval

Deep Evidence Scoring

100,000's Scores from many Deep Analysis Algorithms

Multiple Interpretations

100's sources

Question & Topic Analysis

Question Decomposition

Hypothesis Generation

Hypothesis and Evidence Scoring

Synthesis

Final Confidence Merging & Ranking

Hypothesis Generation

Hypothesis and Evidence Scoring

. . .

Answer & Confidence

Outline

1. **Motivation**
2. **Hypotheses-Oriented Approach Revised**
   - Hypothesis Generation
   - Hypothesis Evaluation
   - Algorithmic Generation and Evaluation of Hypotheses
   - Bayesian Motivation for Discovery
3. **Facilities for Hypothesis-Driven Experiment Support**
   - Conceptualization of Scientific Experiment
   - Scientific Hypothesis Formalization
   - Hypotheses as Data in Probabilistic Databases
4. **Examples of Hypothesis-Driven Scientific Research**
   - Besançon Galaxy Model
   - Analysis of Connectome Based on Network Data
   - Climate in Australia
   - Financial Markets
5. **Summary**

Hypothesis-Oriented Approach to Scientific Experiment

First hypothesize, then experiment . . .

Definition

A scientific hypothesis is a proposed falsifiable explanation of a phenomenon which still has to be rigorously tested.

> “Without hypothesis there is no science”
>
> — M. Poincaré

Relationship between hypotheses, laws, theories

Definition

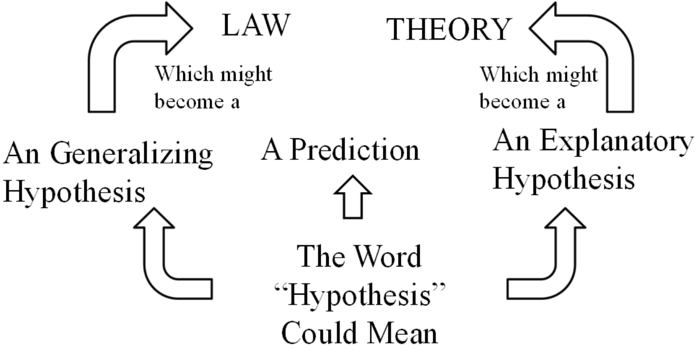
A scientific theory has undergone extensive testing and is generally accepted to be the accurate explanation behind an observation.

Definition

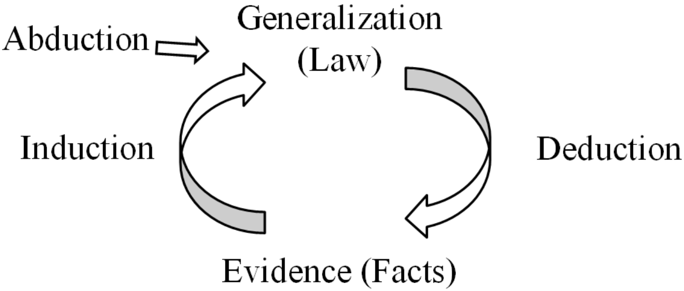
A scientific law is a proposition, which points out any such orderliness or regularity in nature, the prevalence of an invariable association between a particular set of conditions and particular phenomena.

- No essential difference between constructs used for expressing hypotheses, theories and laws.

Relationship between hypotheses, laws, theories


LAW
THEORY
Which might become a
Which might become a
An Generalizing Hypothesis
A Prediction
An Explanatory Hypothesis
The Word "Hypothesis" Could Mean

Enhanced Knowledge Production Diagram



Induction, Deduction, Abduction

Definition

Induction is a technique by which individual pieces of evidence are collected and examined until a law is discovered or a theory is invented

Definition

Deduction is the process of reasoning from one or more statements (premises) to reach a logically certain conclusions

Definition

Abduction is the process of validating a given hypothesis through reasoning by successive approximation

Different Representations of Scientific Hypothesis

From classical hypothesis to the DIS hypothesis

- Mathematical equation
  - $a(t) = -g$
  - $v(t) = -gt + v_0$
  - $s(t) = -(g/2)t^2 + v_0 t + s_0$
- Existentional formula
  - $\forall x \in X \, \forall y \in Y, \quad p(x) \to q(y)$

Database relation

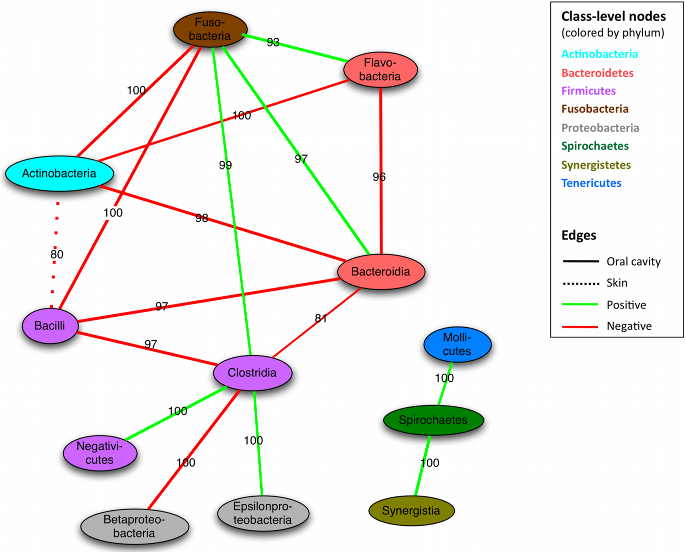
t	v	s
0	0	5000
1	-32	4984
2	-64	4936
3	-96	4856
4	-128	4744

Algorithm

```
for k = 0:n;
    t = k * dt;
    v = -g*t + v_0;
    s = -(g/2)*t^2 +
        v_0*t + s_0;
    t_plot(k) = t;
    v_plot(k) = v;
    s_plot(k) = s;
end
```

Different Representations of Scientific Hypothesis

Associative or causal relationship



Hypothetico-Deductive Method

- Scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data.
- A test that could and does run contrary to predictions of the hypothesis is taken as a falsification of the hypothesis.
- A test that could but does not run contrary to the hypothesis corroborates the theory.

Hypothetico-Deductive Method

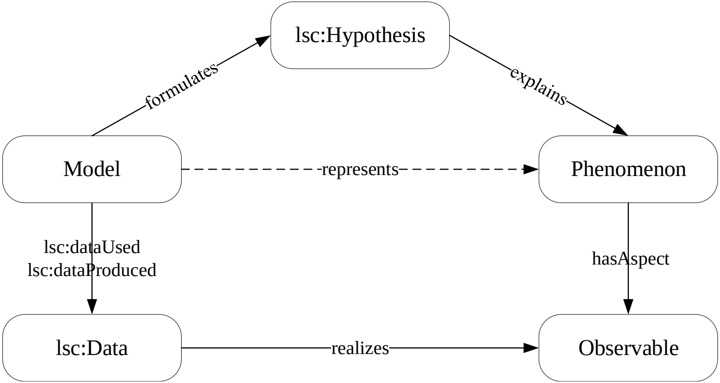
- Predictions are made from the *independent* variable to the *dependent* variable. It is the dependent variable that the researcher is interested in understanding.

Example

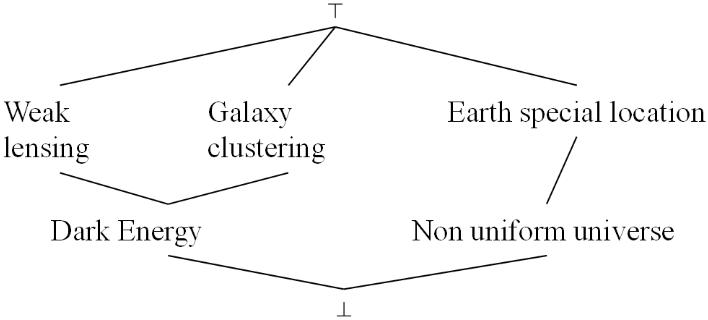
Effect of drug dosage on symptom severity. The independent variable is the dose and the dependent variable is the frequency/intensity of symptoms.

- Usually, statistical hypothesis testing is used:
  - Null and alternative hypothesis are stated;
  - The null hypothesis states that there is no relationship between the phenomena (variables) whose relation is under investigation, or at least not of the form given by the alternative hypothesis.
  - Rejecting the null hypothesis suggests that the alternative hypothesis may be true

Elements of Hypothesis-Driven Research



Research Lattice



⊤

Weak lensing

Galaxy clustering

Earth special location

Dark Energy

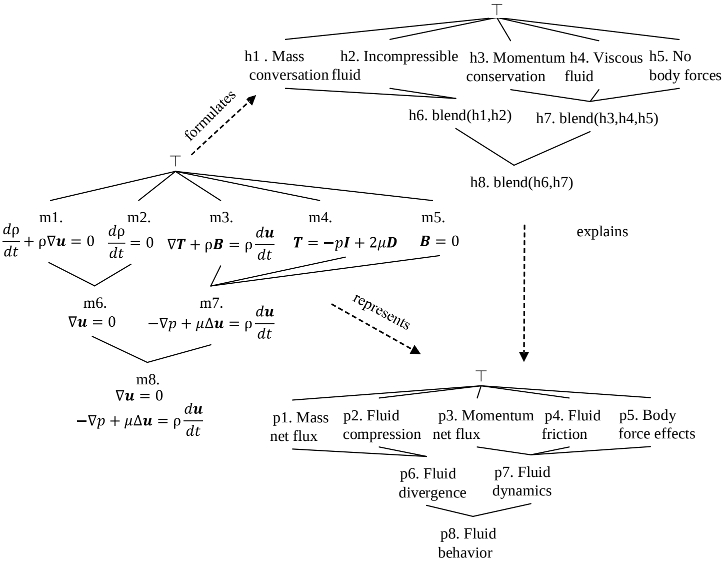
Non uniform universe

⊥

Definition

A hypothesis lattice is formed by considering a set of hypotheses equipped with wasDerivedFrom as a strict order < (from the bottom to the top). Hypotheses directly derived from exactly one hypothesis are *atomic*, while those directly derived from at least two hypotheses are *complex*.

Research Lattices I



$\top$

h1 . Mass conversation

h2. Incompressible fluid

h3. Momentum conservation

h4. Viscous fluid

h5. No body forces

h6. blend(h1,h2)

h7. blend(h3,h4,h5)

h8. blend(h6,h7)

formulates

$\top$

m1. $\frac{d\rho}{dt} + \rho\nabla\boldsymbol{u} = 0$

m2. $\frac{d\rho}{dt} = 0$

m3. $\nabla\boldsymbol{T} + \rho\boldsymbol{B} = \rho\frac{d\boldsymbol{u}}{dt}$

m4. $\boldsymbol{T} = -p\boldsymbol{I} + 2\mu\boldsymbol{D}$

m5. $\boldsymbol{B} = 0$

m6. $\nabla\boldsymbol{u} = 0$

m7. $-\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{d\boldsymbol{u}}{dt}$

m8. $\nabla\boldsymbol{u} = 0$ $-\nabla p + \mu\Delta\boldsymbol{u} = \rho\frac{d\boldsymbol{u}}{dt}$

explains

represents

$\top$

p1. Mass net flux

p2. Fluid compression

p3. Momentum net flux

p4. Fluid friction

p5. Body force effects

p6. Fluid divergence

p7. Fluid dynamics

p8. Fluid behavior

Research Lattices II

Definition

The lattices are isomorphic if one takes subsets of M (Model), H (Hypotheses) and P (Phenomenon) such that formulates, explains and represents are both one-to-one and onto mappings (i.e., bijections), seen as structure-preserving mappings (morphisms).

Hypothesis Generation

Few possible methods to produce hypotheses

- Discovery as abduction

P_1, P_2
$H_1, H_2 \rightarrow P_1$
$H_2, H_3 \rightarrow P_2$
$\Longrightarrow$ H_2 is a possible hypothesis

Abductive logic programming systems

ACLP, A-system, ABDUAL, ProLogICA

- Anomalies search
- Computational analogies

Approaches to hypothesis evaluation

- Logic-based approach
- Frequentist approach
    - relative frequencies of events
    - fixed unknown parameters
- Bayesian approach
    - degree of subject belief
    - inferences by producing probability distribution
- Parameter estimation

Logic-based hypothesis

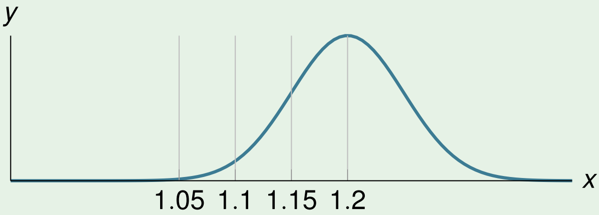
- Clean deduction exists in physics, very rare in biology
- The premises logically entail the conclusion, where the entailment means that the truth of the premises provides a guarantee of the truth of the conclusion
- Inductive logic:
  - Criterion of Adequacy (CoA): as evidence accumulates, the degree to which the collection of true evidence statements comes to support a hypothesis
  - Combinations with Bayesian approach

Statistical testing of hypothesis

Testing the effect of a drug

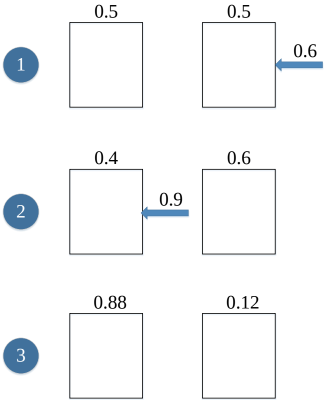
100 rats are injected a drug to test its effect on the response time. Not injected response time – 1.2 s. The mean of the 100 injected rats' response time is 1.05 s. with a sample standard deviation of 0.5 s. Does the drug have the effect on response time?

H_0 : Drug has no effect $\Longrightarrow \quad \mu = 1.2s.$
H_1 : Drug has an effect $\Longrightarrow \quad \mu \neq 1.2s.$
Assume H_0:
approximate $\overline{\sigma_x} = \overline{\sigma}/\sqrt{100} = 0.05$



How many σ away:
$z = (1.2 - 1.05)/0.05 = 3 \quad \Longrightarrow \quad p_{value} = 0.003$

Bayesian approach to hypothesis evaluation



Bayes formula is used to compute posterior probabilities:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Outline

1. Motivation

2. Hypotheses-Oriented Approach Revised
   - Hypothesis Generation
   - Hypothesis Evaluation
   - Algorithmic Generation and Evaluation of Hypotheses
   - Bayesian Motivation for Discovery

3. **Facilities for Hypothesis-Driven Experiment Support**
   - Conceptualization of Scientific Experiment
   - Scientific Hypothesis Formalization
   - Hypotheses as Data in Probabilistic Databases

4. Examples of Hypothesis-Driven Scientific Research
   - Besançon Galaxy Model
   - Analysis of Connectome Based on Network Data
   - Climate in Australia
   - Financial Markets

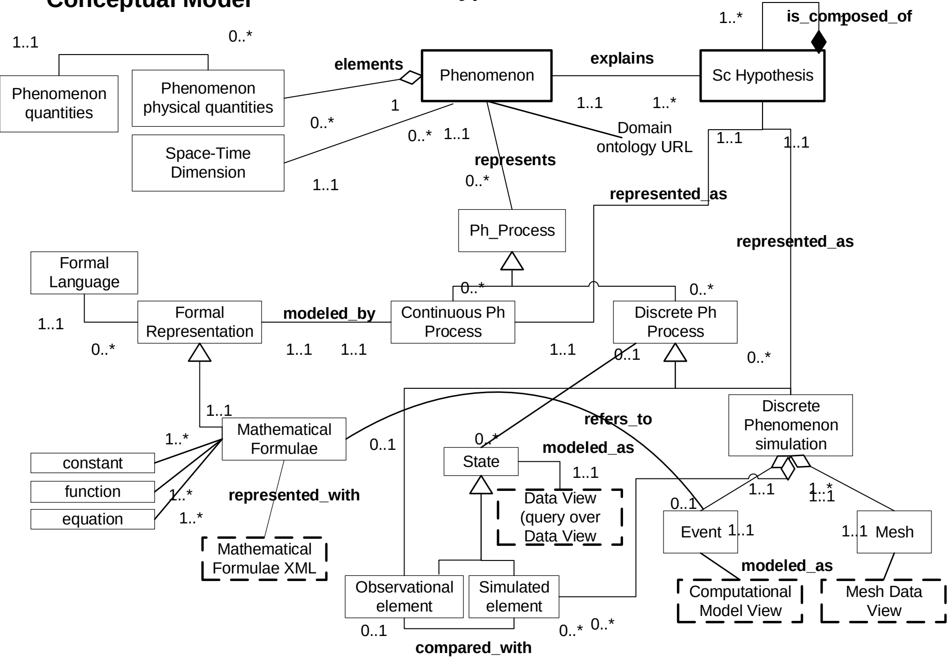
5. Summary

Conceptualization of Scientific Experiment

It becomes paramount to offer scientists mechanisms to manage the variety of knowledge produced during such investigations.

Sc Hypothesis Conceptual Model

Scientific Hypothesis Model



Scientific Hypothesis Formalization

Diversity of the components of scientific hypothesis model

- Borrowed from applications in NeuroScience and in human cardiovascular system in Computational Hemodynamics
- Provided
  - by a mathematical model;
  - a set of differential equations for continuous processes;
  - quantifying the variations of physical quantities in continuous space-time;
  - by the mathematical solver (HEMOLAB) for discrete processes.
- Mathematical equations in MathML (enabling models interchange and reuse);

Scientific Hypothesis Formalization

Possible problems and extensions

- Mostly monotonic and not suitable for incomplete knowledge representation
- Another framework implemented by extending BioSigNet-RR:
  1. support elaboration tolerant representation and non-monotonic reasoning;
  2. seamless integration of hypothesis formation with knowledge representation and reasoning;
  3. use of various resources of biological data as well as human expertise to intelligently generate hypotheses;
  4. support for ranking hypotheses and for designing experiments to verify hypotheses.
- Prototype of an intelligent research assistant of molecular biologists.

Hypotheses as Uncertain Data in Probabilistic Databases

γ–DB approach

- Set of MathML equations is used to build database relation schema and functional dependencies
- Database is filled with both simulated data and observed data
- It enables to maintain several hypotheses explaining some phenomena
- and provides evaluation mechanism based on Bayesian approach to rank them.

Outline

1. Motivation

2. Hypotheses-Oriented Approach Revised
   - Hypothesis Generation
   - Hypothesis Evaluation
   - Algorithmic Generation and Evaluation of Hypotheses
   - Bayesian Motivation for Discovery

3. Facilities for Hypothesis-Driven Experiment Support
   - Conceptualization of Scientific Experiment
   - Scientific Hypothesis Formalization
   - Hypotheses as Data in Probabilistic Databases

4. **Examples of Hypothesis-Driven Scientific Research**
   - Besançon Galaxy Model
   - Analysis of Connectome Based on Network Data
   - Climate in Australia
   - Financial Markets

5. Summary

Besançon Galaxy Model

Multiparameter Hypotheses Examples

Definition

Besançon Galaxy Model is a model of stellar population synthesis of the Galaxy, which allows to test hypotheses on the star formation history, star evolution, and chemical and dynamical evolution of the Galaxy

- BGM is based on a set of different interconnected hypotheses
- Some of the hypotheses are passed as free input parameters
- When a new data arrives (e.g. Tycho-2) a comparison is done between simulations from the model with data as a first test to verify and constrain the underlying hypotheses.

Besançon Galaxy Model

Example of underlying hypotheses

SFR

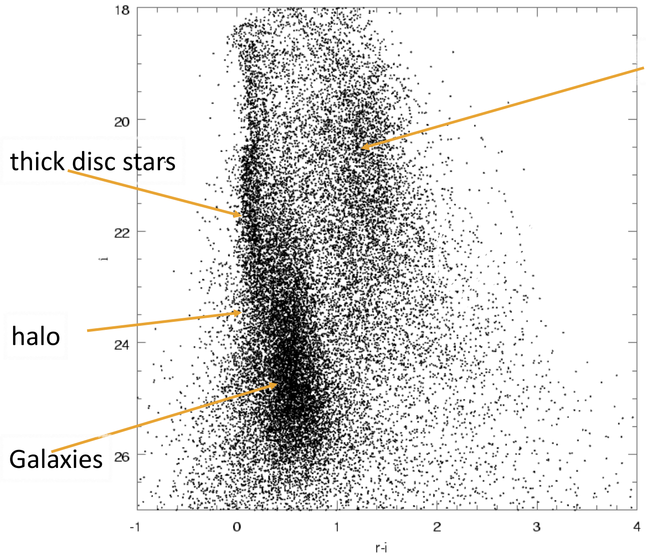
Star formation rate history: How many stars are formed at a given time

IMF

Initial Mass Function: How many stars of a given mass

▸ More on BGM hypotheses

Galaxy Model (Besançon): from data to Galactic formation and evolution



thin disc stars

thick disc stars

halo

Galaxies

<u>Sky survey data</u>:

Color-magnitude diagram of observed stars in a given direction

<u>Problem</u>: how to extract informations? Data on distances are not available.

GALAXY MODEL: Ingredients (basic complex multiparameter hypotheses)

- Star formation rate history : How many stars are formed at a given time
- Initial Mass Function (IMF) : How many stars of a given mass
- Stellar models (evolutionary tracks): How stars evolve with time
- Stellar atmosphere models : How the physical state of the star atmosphere are observed (getting magnitudes and colors)
- Stellar density distributions: How the star density changes across the sky
- Chemical evolution : How the chemical abundances of the stars change with their birth time
- Dynamical consistency: How to include dynamical constraints (based on gravity, conservation of energy, conservation of mass and of angular momentum)

Analysis of Connectome Based on Network Data

Definition

Connectome is a comprehensive map of neural connections in the human brain. Functional connectome denotes the collective set of functional connections in the human brain (its "wiring diagram").

- Uses fMRI, where associations are thought to represent functional connectivity, in the sense that the two regions of the brain participate together in the achievement of some higher-order function
- "Is there a difference between the networks of these two groups of subjects?"
- Projects:
  - 1000 Functional Connectomes Project (FCP)
  - Human Connectome Project (HCP) - build a network map of the human brain in healthy, living adults.

Connectomic challenges

- On the microscale , the raw data for a the size of the human brain would require a zettabyte (a trillion gigabytes) of memory, currently beyond the storage capacity of any computer
- Connectome data sets obtained at the macroscale with noninvasive imaging technology are manageable (has a petabyte scale).
- The network theory and respective data model should capture connectivity at any scale and thus naturally encompass multiscale architectures and nested connectivity patterns.
- Regarding the human mind, defining a classification scheme or ontology of mental processes has remained an unfulfilled challenge.

  Once a mapping between mental states and neural responses has been established, it should be possible to infer mental states from neural observations.

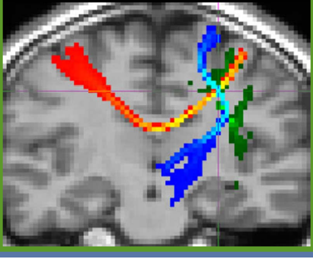
Multiscale Framework

- Virtual Brain and related modeling efforts explicitly build on a multiscale theoretical framework
- Multiscale approaches embrace models for dimension reduction where processes at smaller scales become part of compact descriptions of regularities at larger scales

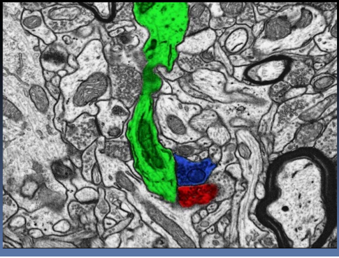
The Human Connectome Project:

- An NIH-funded effort to chart a comprehensive map of neuronal connections and its variability in healthy adults (on the macro-scale)

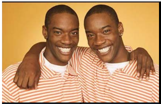
Macro-connectome
(whole-brain, long-distance)

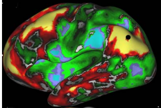


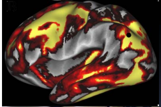
Micro-connectome
(synapses, neurons)

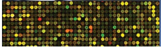


Project Goals I









- Study a large population:
  - 1,200 healthy adults
  - 300 twin pairs and their non-twin siblings
- Cutting-edge neuroimaging methods
  - 3T Skyra MRI, customized gradient (UMinn -> Wash U)
  - 7T MRI (UMinn, 200 subjects); perhaps also 10.5T
  - dMRI/tractography; R-fMRI; Task-fMRI
  - MEG/EEG (100 subjects)
- Extensive behavioral testing
- Blood samples for genotyping

Climate in Australia

Another view on hypothesis representation

- As data gets continuously aggregated
- Hypotheses should be represented as programs, that are executed repeatedly as new data arrives
- This method is tested by examining hypothesis about temperature trends in Australia during the 20th century

Hypothesis

Temperature series is not stationary and is integrated of order 1 (I(1))

▸ More on climate hypothesis

Climate in Australia

Data and tools

Data sources:

1. The National Oceanographic and Atmospheric Administration marine and weather information
2. Australian Bureau of Meteorology dataset

Statistical tests:

- Phillips-Perron test
- Kwiatkowski-Phillips-Schmidt-Shin (KPSS)

Tools:

- R *SPARQL*, *tseries* packages
- **agINFRA** for scientific workflows for natural sciences

Support for hypothesis

Authors received further evidence on different independent dataset that time series is integrated of order 1.

Efficient Market Hypothesis

Two approaches

Definition

Weak form of Efficient Market Hypothesis

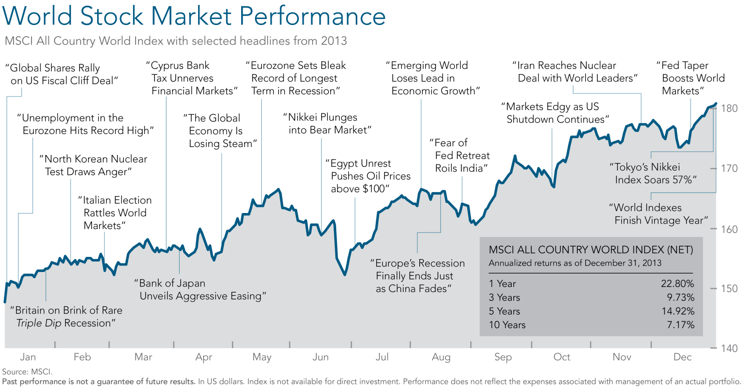
- The weak form of the weak form of efficient markets hypothesis (EMH) states that prices on traded assets (e.g., stocks, bonds, or property) already reflect all past publicly available information.
- One of possible formulations of the efficient market hypothesis used for weak form tests is that share prices follow a random walk.

Two approaches to evaluate this hypothesis are suggested:

1. by analyzing stock market movements for several countries' indices in selected period
2. by investigating how public sentiment(daily Twitter posts) predicts the stock market

Efficient Market Hypothesis

Classical approach



World Stock Market Performance

MSCI All Country World Index with selected headlines from 2013

MSCI ALL COUNTRY WORLD INDEX (NET)

Annualized returns as of December 31, 2013

1 Year	22.80%
3 Years	9.73%
5 Years	14.92%
10 Years	7.17%

Source: MSCI.

Past performance is not a guarantee of future results. In US dollars. Index is not available for direct investment. Performance does not reflect the expenses associated with management of an actual portfolio.

Efficient Market Hypothesis

Classical approach

1. Collected daily closing prices from the six European stock markets (France, Germany and UK, Greece, Portugal and Spain) during the period between 1993 and 2007
2. Stated null hypothesis (successive prices changes are independent (random walk)) and alternative hypothesis (they are dependent)
3. Applied
   - serial correlation test,
   - runs test
   - augmented Dickey-Fuller test
   - multiple variance ratio test

EMH is supported

The null hypothesis should not be rejected for all six markets. EMH is supported

Efficient Market Hypothesis

Sentiment approach

Committee
rate
percent
inflation
unemployment
market
purchases
labor
year
policy
economic
September.
participation
conditions
funds
growth
Conference
Bernanke
asset
longer
outlook
term
Chairman
PRELIMINARY
Press
program
time
improvement
projections
federal
progress
objective
guidance
monetary
today
Page
financial
see
concluded
course
month
generally
range
th
end
short
factors
appropriate
securities
meeting
recovery
population
price
important
likely
part
central
consistent
increase
still
CHAIRMAN
gradually
fallen
current
made
meaningful
developments
later
Federal
might
back
individual
employment
point
data
place
maximum
think
jobs
moderate
purchase
job
pace
stability
whether
months
percentage
due
continued
economy
question
forward
normal
projected
target
close
FOMC
risks
also
force
continue
subject
decline
substantial
gains
turn
two
example
rates
Reserve
BERNANKE
remains
first
moving
assessment
expect
participants
billion
past
toward
June
tendency
well
run
noted
including
addition
support
fiscal
know
tightening
high
take
context
remain
broader
ongoing
However
expectation
1/2
extent
expectations
long
per
earlier
statement
sustained
continuing
particular
trend

Efficient Market Hypothesis

Sentiment approach

1. Build public mood time series by sentiment analysis of tweets from February 28, 2008 to December 19, 2008
2. Null hypothesis states that the mood time series do not predict DJIA values
3. Granger causality analysis in which Dow Jones values and mood time series are correlated is used to test the null hypothesis

EMH is not supported

Results reject the null hypothesis and claim that public opinion is predictive of changes in DJIA closing values.

Outline

1. Motivation

2. Hypotheses-Oriented Approach Revised
   - Hypothesis Generation
   - Hypothesis Evaluation
   - Algorithmic Generation and Evaluation of Hypotheses
   - Bayesian Motivation for Discovery

3. Facilities for Hypothesis-Driven Experiment Support
   - Conceptualization of Scientific Experiment
   - Scientific Hypothesis Formalization
   - Hypotheses as Data in Probabilistic Databases

4. Examples of Hypothesis-Driven Scientific Research
   - Besançon Galaxy Model
   - Analysis of Connectome Based on Network Data
   - Climate in Australia
   - Financial Markets

5. **Summary**

Summary

- Role of hypotheses in DIR is emphasized
- Basic concepts defining the role of hypotheses in the formation of scientific knowledge and organization of the scientific experiments are presented
- Basic approaches for hypothesis formulation applying logical reasoning, various methods for hypothesis modeling and testing (including classical statistics, Bayesian hypothesis and parameter estimation methods, hypothetico-deductive approaches) are presented

Summary

- Facilities are aimed at the conceptualization of scientific experiment, hypothesis formulation and browsing in various domains (including biology, biomedical investigations, neuromedicine, astronomy), automatic organization of hypothesis-driven experiments
- Examples of scientific researches applying hypotheses include modeling of population and structure synthesis of the Galaxy, connectome-related hypothesis testing, studying of temperature trends in Australia, analysis of stock markets applying the EMH

Examples of BGM Hypotheses

◂ Return

Definitions for Climate Hypothesis

Definition

Non-stationarity means that the level of the time series is not stable in time and can show increasing and decreasing trends

Definition

I(1) means that by differentiating the stochastic process a stationary process (main statistical properties of the series remain unchanged) is obtained

◂ Return