Intelligent data analysis
Introduction

Peter Antal antal@mit.bme.hu

Overview

- The data-intensive age
- What is data analysis (i.e. inductive inference)?
- What is intelligence?
- Aspects of learning:
  - Neuronal : neurobiological adaptation
  - Individual: psychological development of a child, researcher
  - Social: scientific discovery
- A normative theory for inductive inference.
- Types of data
- Types of inference
- Current challenges
  - Fusion
  - Active learning
  - Transformational learning
  - Deep learning..

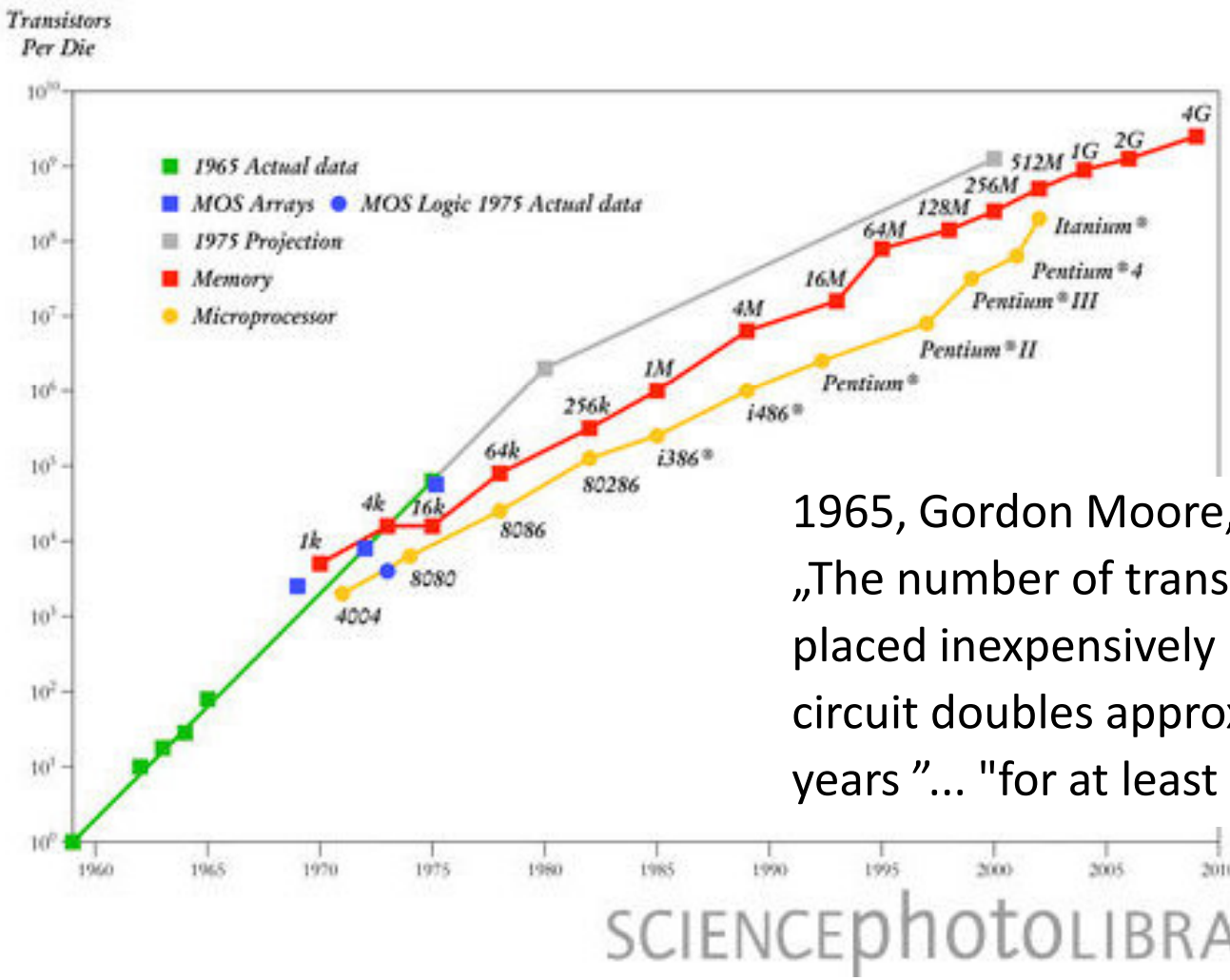
The data-intensive science

- Data analysis and knowledge fusion is more important than simulation of „simple” laws.
- 20th century: Physics vs. 21st century: Biology.
  - Tony Hey, Stewart Tansley, and Kristin Tolle: **The fourth paradigm (Data-Intensive Scientific Discovery**), http://research.microsoft.com/en-us/collaboration/fourthparadigm/, 2009
  - Gordon Bell, Tony Hey, Alex Szalay: **Beyond the Data Deluge**, Science, 323, pp 1297-1298, 2009

Data accumulation vs interpretation

- „New data--whole new types of data--are accumulating faster than researchers can make sense of them. The result is something like an optical illusion. Contradictory images [of Mars] seem to flicker in and out of focus in the mind's eye.” Hugh Keiffer
- In many fields:
  - Astronomy (Hubble)
  - Nuclear physics (LHC)
  - Biology (omics)
  - Chemistry (drug research)
  - Medicine (electronic patient records)
  - Ecosystems (climate change, pollution, weather forecast)
  - Social relations (Google ;-), security)
  - Economics (financial systems)
  - IT (server farms!!)

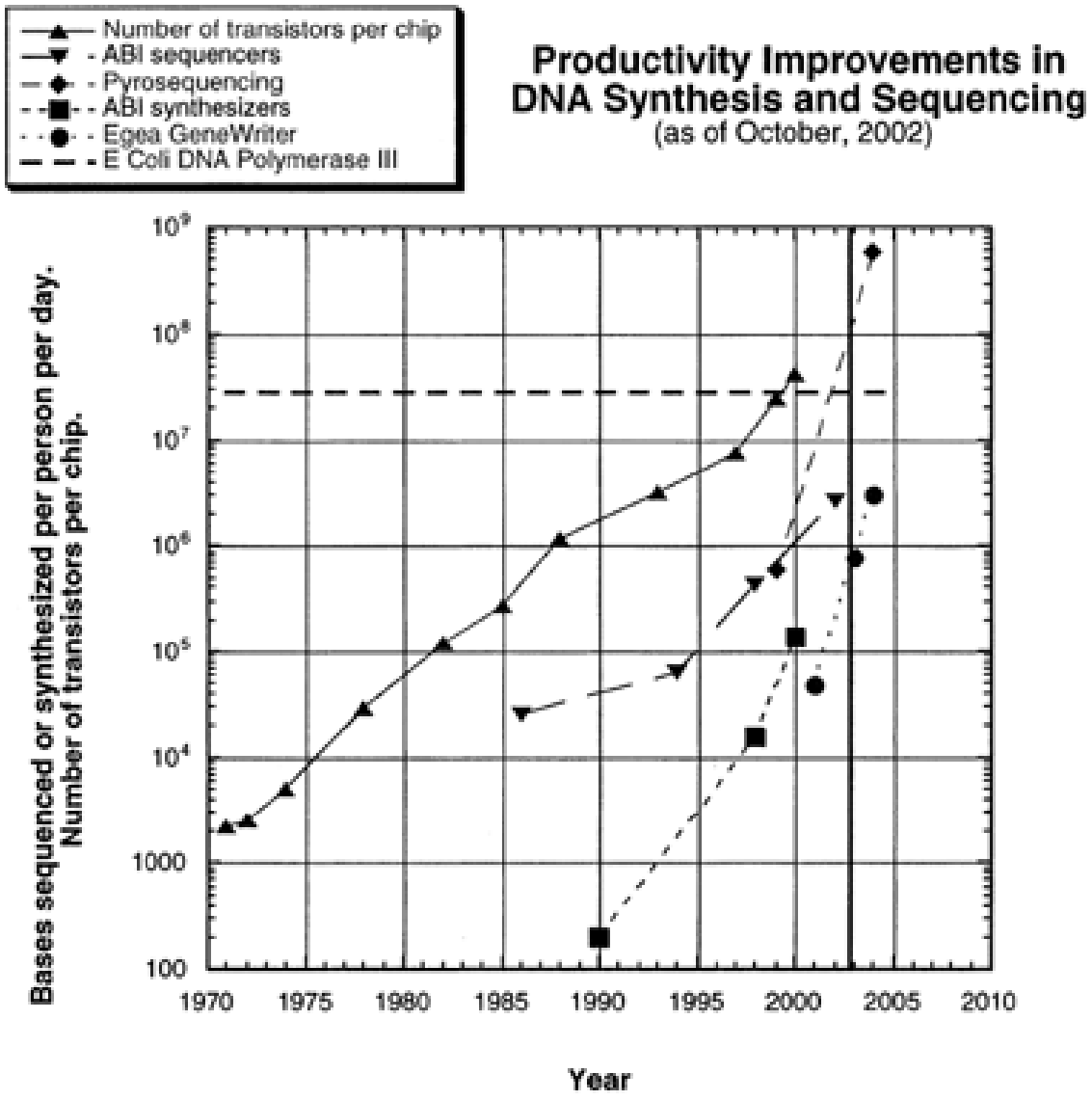
Moore's Law (in computation)



Integration and parallelization wont bring us further. End of Moor's law?

1965, Gordon Moore, founder of Intel: „The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years "... "for at least ten years"

Carlson's law I.

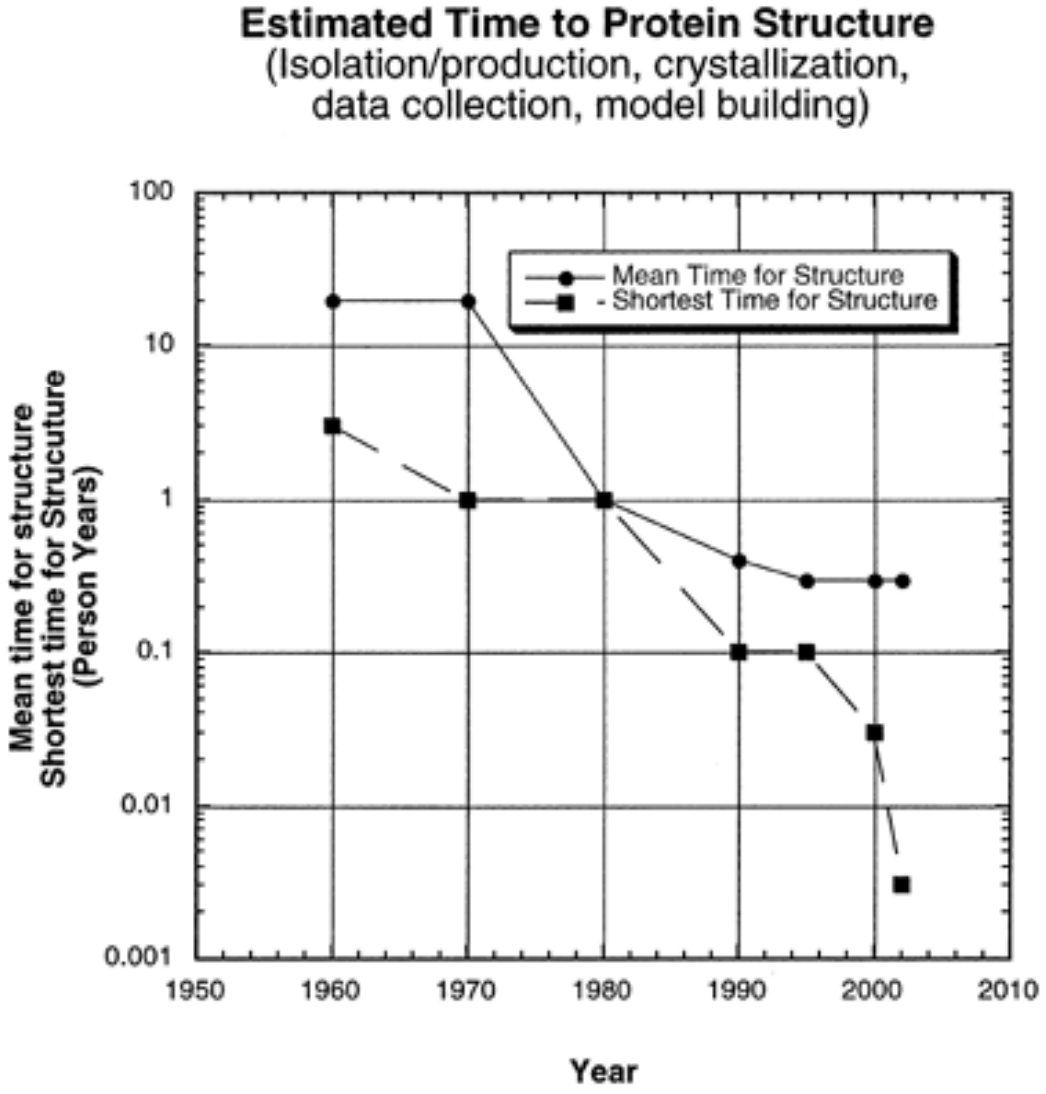


On this semi-log plot, DNA synthesis and sequencing productivity are both increasing at least as fast as Moore's Law (upwards triangles). Each of the remaining points is the amount of DNA that can be processed by one person running multiple machines for one eight hour day, defined by the time required for preprocessing and sample handling on each instrument.

Rob Carlson: The Pace and Proliferation of Biological Technologies,KurzweilAI.net March 4, 2004

http://www.kurzweilai.net/the-pace-and-proliferation-of-biological-technologies-2

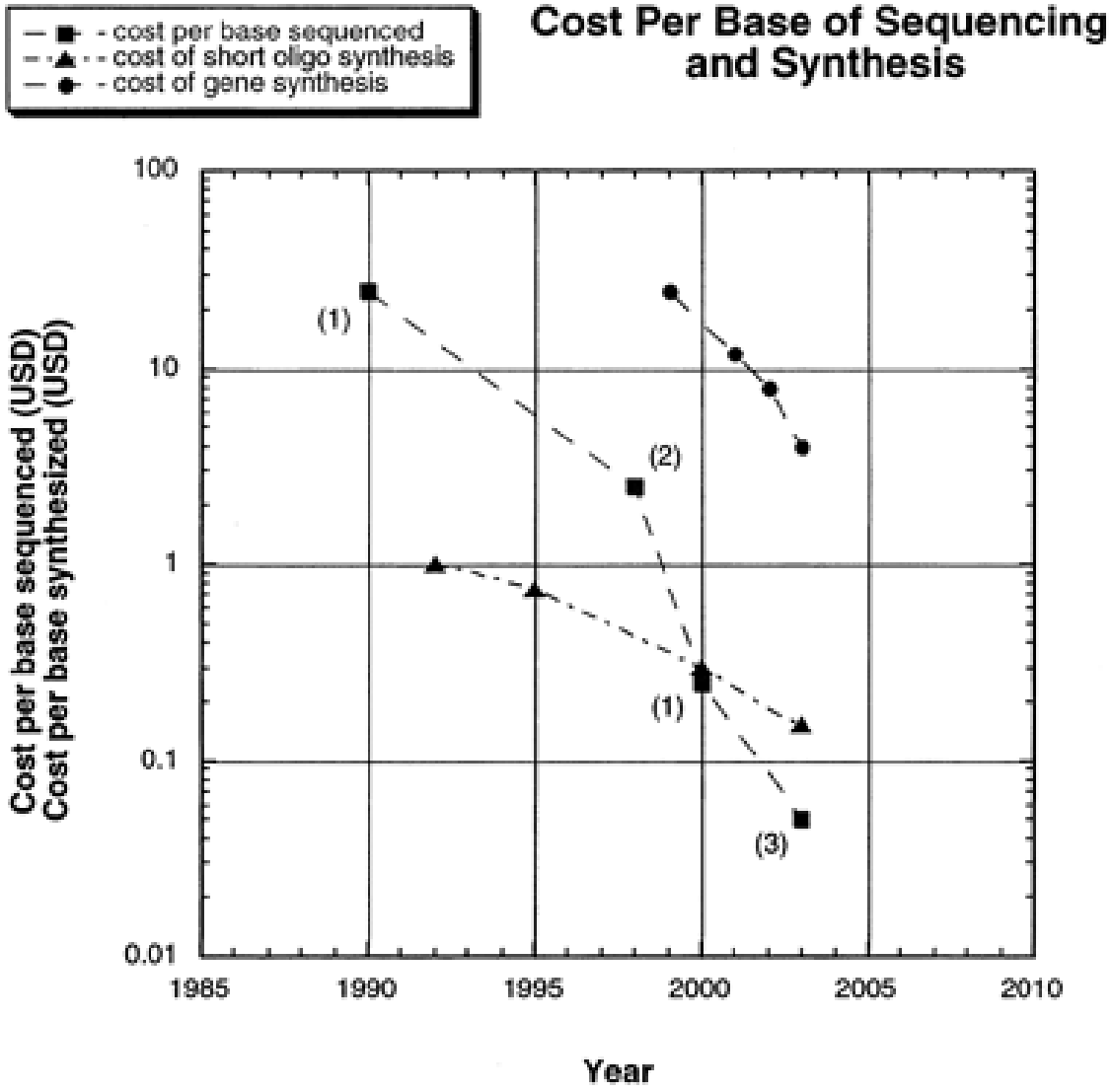
Carlson's law II.



The dramatic improvement in the time required to determine protein structures is evidence of a general trend towards increased productivity in biological technologies. Many of the technologies used in finding protein structures are used widely in biology for other purposes. Raw estimates of time to collect and crystallize recombinant proteins, to take x-ray data, and to build structural models were compiled by Richard Yu (The Molecular Sciences Institute, Berkeley, CA) based on his experience and a survey of five additional crystallographers.

Rob Carlson: The Pace and Proliferation of Biological Technologies,KurzweilAI.net March 4, 2004

Carlson's law III.



Rough estimates of the cost of synthesis and raw sequencing per base. Only very limited data are available. Estimates of synthesis costs are from John Mulligan, Blue Heron Biotechnology. Historical costs of sequencing are generally not available in the literature, have not been publicized by federally funded Genome Centers, and are, in general, surprisingly hard to come by:44(1) from Lander *et al.*;6 (2) from Dan Rokhsar, UC Berkeley; (3) approximate current commercial rate.

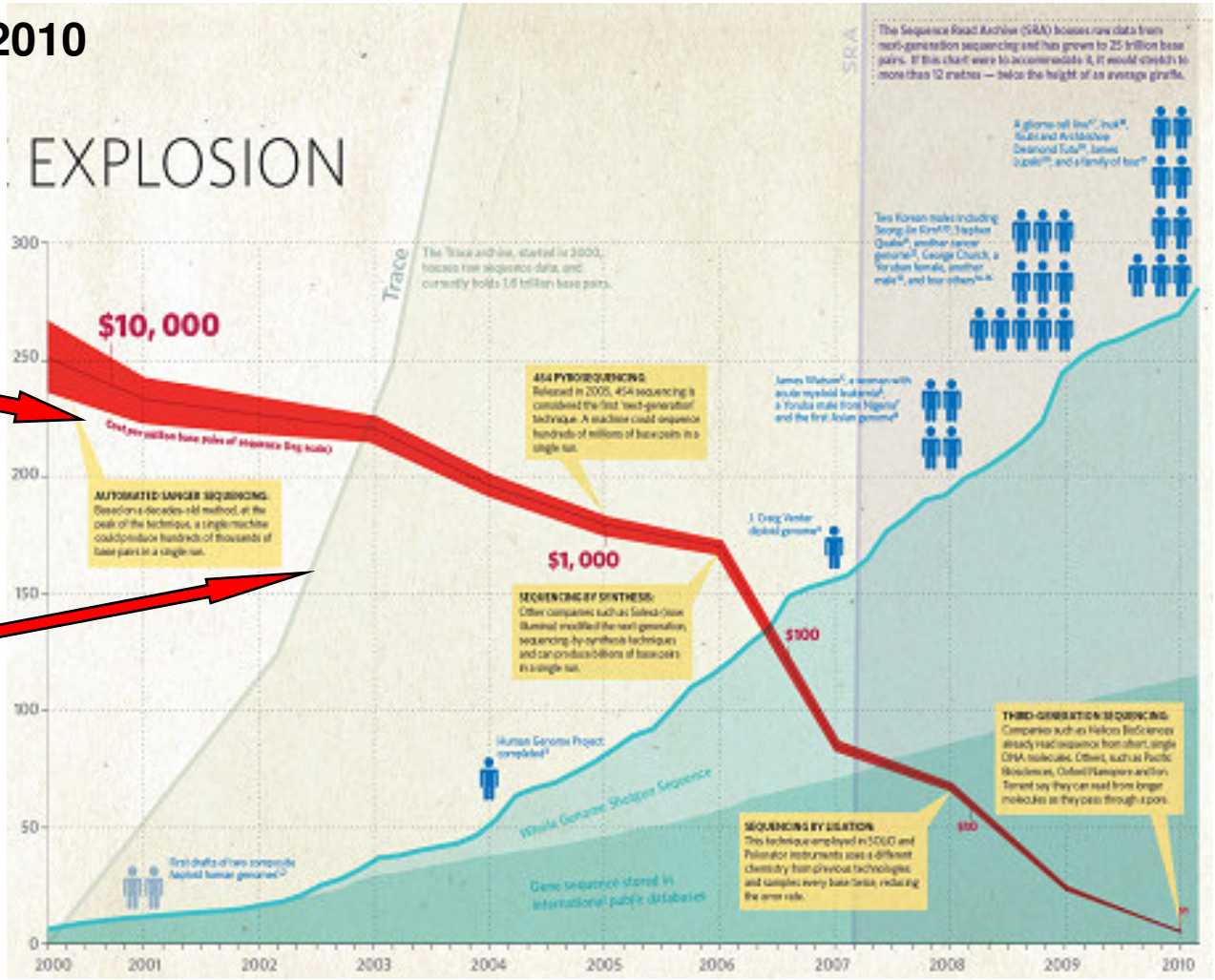
Rob Carlson: The Pace and Proliferation of Biological Technologies,KurzweilAI.net March 4, 2004

Moore's Law for Data Explosion

NATURE, Vol 464, April 2010

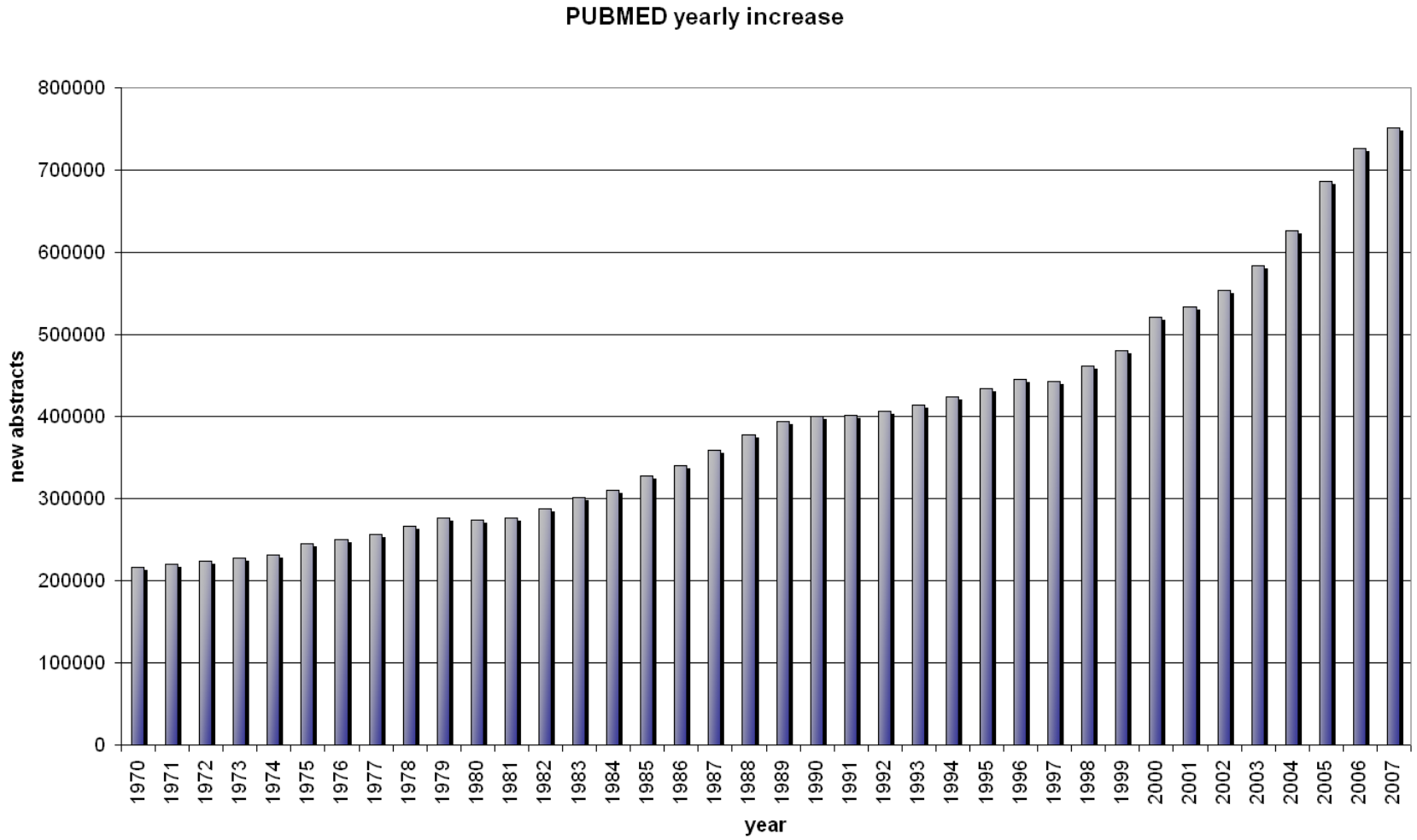
Sequencing costs per mill. base

Publicly available genetic data



- x10 every 2-3 years
- Data volumes and complexity that IT has never faced before…

Number of biomed publications



Rules in data-age

- Being close to data (e.g. clouds),
- fusion of heterogeneous information sources,
- incorporation of prior knowledge,
  (i.e. earlier data and earlier computation)
- averaging over complex, structured models,
- parallel, scalable methods,
- distributed, open, community-based methods
  are more important than the refined analysis of a single information source using a single method for a single researcher.

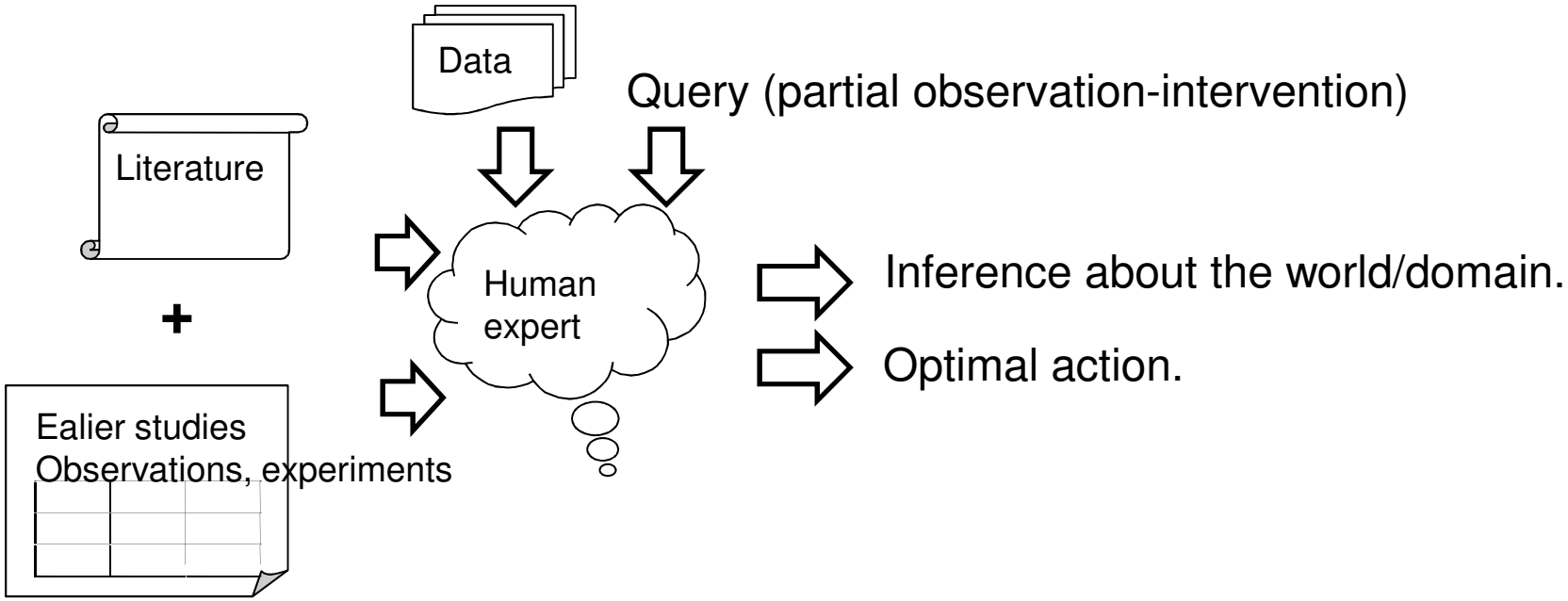
Challenges

Astronomy (Hubble)		
Nuclear physics (LHC)		
Biology (omics)		
Chemistry (drug research)		
Medicine (electronic patient records, @home!)		
Ecosystems (climate change, pollution, weather forecast)		
Social relations (Google ;-), security)		
Economics (financial systems, web, blogs,...)		

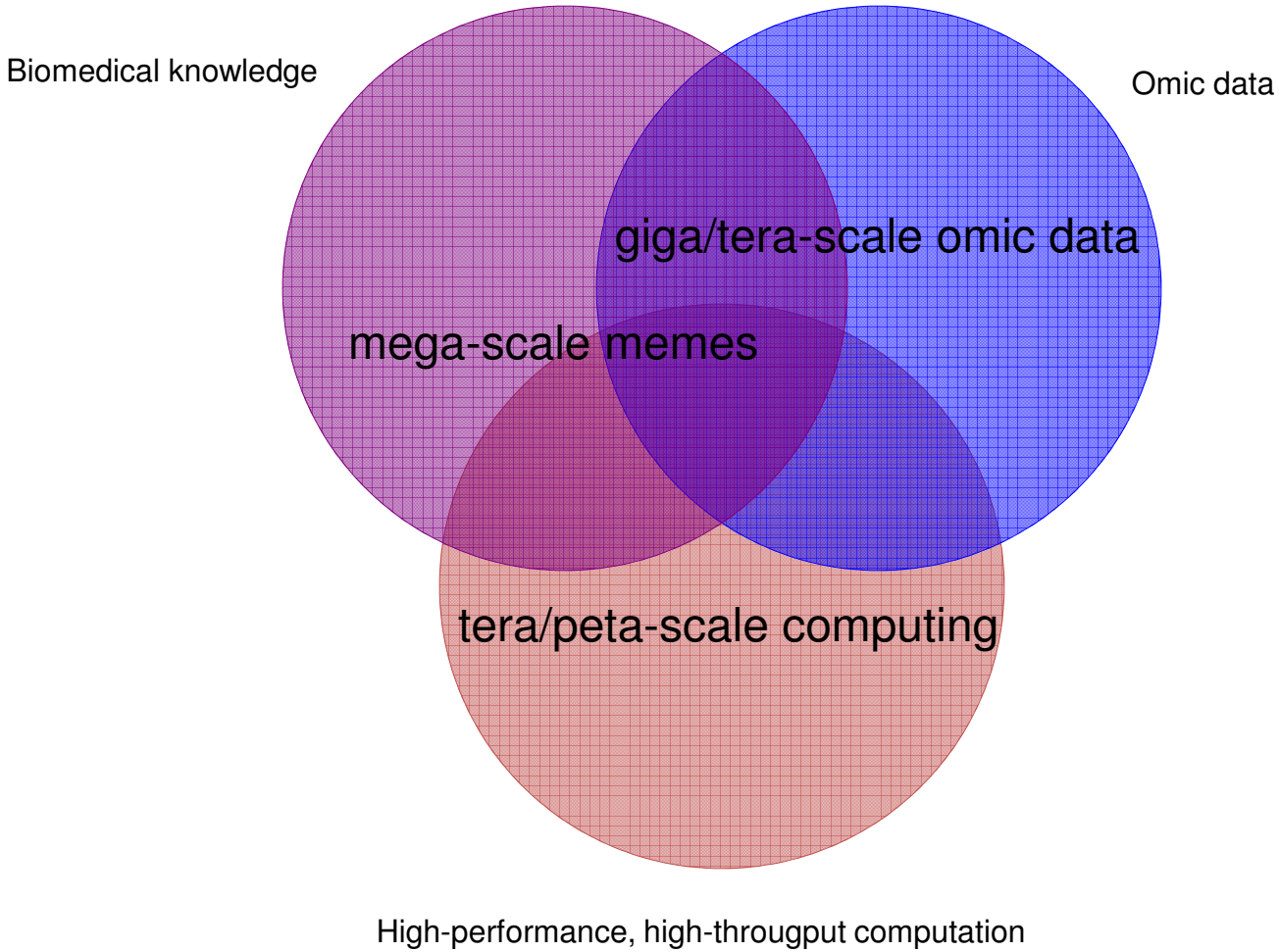
Challenges related to mankind

- Prediction, intervention at all level of humanity
  - Macromolecular
  - Cell
  - Organs
  - Individual
    - Body: health industry
    - Psychological:
      - scientific: external memories
      - commercial recommendation systems
  - Society
    - Economy

Data analysis = inductive inference



Vision: the fusion dream



A personal learning curve in induction

- Cognitive science: nature of expertise
  - Schemes,.., quantum computing :D
- Philosophy of science
  - Objective knowledge, .., paradigms
- Neurobiology: learning in neural networks
  - ART, case-based reasoning
- Cognitive psychology: development of a child
  - ACT-R, creativity,..
- Stastistical learnability
- Inference, causal inference
- Bioinformatics

Intelligent (inductive) inference (Intelligent data analysis)

Think like a human (data analyst?, child?)	Think rationally.
Act like a human.	Act rationally.

- Today the main bottleneck is fusion.
- Fusion of human experts cannot be imitated.
- ➔ rational bases is necessary for data and knowledge representation to support „limitless” fusion.

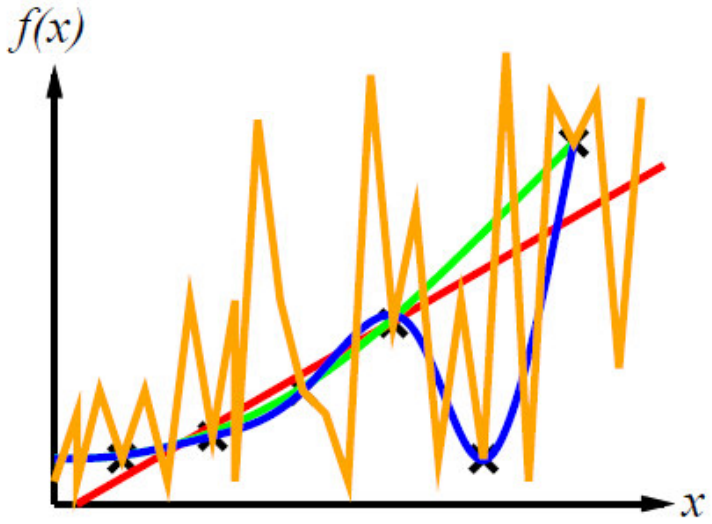
Rational bases for induction

- Ockham's razor: as a scientific principle.
- D.Hume: A the treatise of human nature: induction is logically impossible, both statistical and causal.
- Frequentist:
  - there is an unknown, fix world..
  - Fisher, Pearson: the hypothesis testing framework
  - K.R.Popper: falzification as a scientific paradigm
  - V.Vapnik: sharp(?) bounds for finite samples
- Subjectivist (Bayes,..):
  - Box: „all models are wrong, but some are useful"

Model complexity
- Ockham razor

Construct/adjust h to agree with f on training set
(h is consistent if it agrees with f on all examples)

E.g., curve fitting:



Ockham's razor: maximize a combination of consistency and simplicity

Foundation for induction: Probability theory?

„Probability theory= measure theory+independence”
(„a computer is a tensor”)

- Joint distribution
- Conditional probability
- Independence, conditional independence
- Bayes rule
- Marginalization/Expansion
- Chain rule
- Expectation, variance
- Independence map, decomposition....

Interpretation of probability

- Axioms in probability theory are the same (Kolmogorov)
- Sources of uncertainty
  - inherent uncertainty in the physical process;
  - inherent uncertainty at macroscopic level;
  - ignorance;
  - practical omissions;
- Interpretations of probabilities:
  - combinatoric;
  - physical propensities;
  - frequentist;
  - personal/subjectivist;
  - instrumentalist;
- The three „as if" theorems:
  - Uncertainty by probabilities
  - Preferences by utility function
  - Optimal action by maximum expected utility principle

$$\lim_{N\to\infty}\frac{N_A}{N}=\lim_{N\to\infty}\hat{p}_N(A)=p(A)\,?\,p(A\mid\xi)$$

Note: independence and convergence of frequencies are empirical observations (i.e., „laws of large numbers" consequences of independencies)

On the uniqueness of probability theory I.

- Bayesian framework for induction: we start with hypothesis space and wish to express relative preferences in terms of background information (the Cox-Jaynes axioms).
- **Axiom 0**: Transitivity of preferences.
- **Theorem** 1: Preferences can be represented by a real number π (A).
- **Axiom 1**: There exists a function f such that

$$\pi(\text{non A}) = f(\pi(A))$$

- **Axiom 2**: There exists a function F such that

$$\pi\,(A,B) = F(\pi(A),\, \pi(B|A))$$

- **Theorem2**: There is always a rescaling w such that $p(A)=w(\pi(A))$ is in [0,1], and satisfies the sum and product rules.

Probability theory II.

- **Sum Rule**:

$$P(\text{non } A) = 1 - P(A)$$

- **Product Rule**:

$$P(A \text{ and } B) = P(A)\, P(B|A)$$

- **BayesTheorem**:

$$P(B|A) = P(A|B)P(B)/P(A)$$

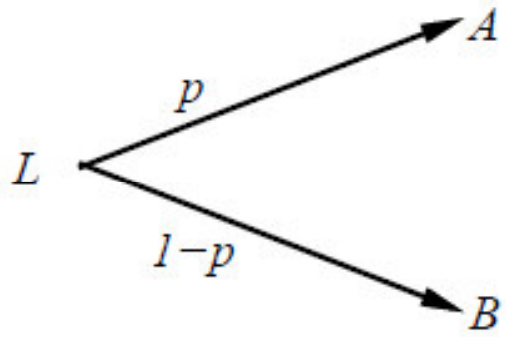
- **Induction Form**:

$$P(M|D) = P(D|M)P(M)/P(D)$$

Recall: utility theory!

From probability theory to decision theory: preferences

An agent chooses among prizes (A, B, etc.) and lotteries, i.e., situations with uncertain prizes



Lottery $L = [p, A;\ (1-p), B]$

Notation:

$A \succ B$	A preferred to B
$A \sim B$	indifference between A and B
$A \precsim B$	B not preferred to A

Rational preferences

Idea: preferences of a rational agent must obey constraints.
Rational preferences $\Rightarrow$
behavior describable as maximization of expected utility

Constraints:

Orderability

$$(A \succ B) \vee (B \succ A) \vee (A \sim B)$$

Transitivity

$$(A \succ B) \wedge (B \succ C) \Rightarrow (A \succ C)$$

Continuity

$$A \succ B \succ C \Rightarrow \exists p \ [p, A;\ 1-p, C] \sim B$$

Substitutability

$$A \sim B \Rightarrow [p, A;\ 1-p, C] \sim [p, B; 1-p, C]$$

Monotonicity

$$A \succ B \Rightarrow (p \geq q \Leftrightarrow [p, A;\ 1-p, B] \succsim [q, A;\ 1-q, B])$$

Utility, Maximum expected utility

Theorem (Ramsey, 1931; von Neumann and Morgenstern, 1944):
Given preferences satisfying the constraints
there exists a real-valued function U such that

$$U(A) \geq U(B) \quad \Leftrightarrow \quad A \succsim B$$
$$U([p_1, S_1; \ \ldots \ ; \ p_n, S_n]) = \sum_i p_i U(S_i)$$

MEU principle:

Choose the action that maximizes expected utility

Note: an agent can be entirely rational (consistent with MEU)
without ever representing or manipulating utilities and probabilities

E.g., a lookup table for perfect tictactoe

Milestones of the subjective approach

- [1713] Ars Conjectandi (The Art of Conjecture), Jacob Bernoulli
  - **Subjectivist interpretation** of probabilities
- [1718] The Doctrine of Chances, Abraham de Moivre
  - the first textbook on probability theory
  - **Forward predictions**
    - „..what is the probability of drawing a black ball?"
    - Predicted his own death
- [1764, posthumous] Essay Towards Solving a Problem in the Doctrine of Chances, Thomas Bayes
  - **Backward questions**: „given that one or more balls has been drawn, what can be said about the urn"
- [1812], Théorie analytique des probabilités, Pierre-Simon Laplace
  - General Bayes rule
- [1921]: **Correlation and causation**, S. Wright's diagrams
- -1950 **Frequentist statistics**
  - Ronald A. Fisher (J. Neyman and E. Pearson)
- [1937], "La prévision: ses lois logiques, ses sources subjectives", B. de Finetti
  - Exchangeability (instead of independency)
- [1939] "Theory of probability„, Harold Jeffreys
- 1950-: „**Bayesian" statistics** (as opposed to the „frequentist" school
  - I.J. Good, B.O. Koopman, Howard Raiffa, Robert Schlaifer and Alan Turing
- [1979] Conditional Independence in Statistical Theory, A.P. Dawid
  - Axiomatization of indepencies in **multivariate** distributions
- [1982] The decomposition of a multivariate distribution, S.Lauritzen
- [1988] Bayesian networks, J.Pearl
  - Representation of independencies

The hypothesis testing framework

- Terminology:
  - False/true x positive/negative
  - Null hypothesis: independence

reported	Ref.:0/N	Ref.1/P
0/N	TN	FN
1/P	FP	TP

  - Type I error/error of the first kind/α error/FP: $p(\neg H_0 | \underline{H_0})$
    - Specificity: $p(H_0 | \underline{H_0}) = 1-\alpha$
    - Significance: α
    - p-value: „probability of more extreme observations in repeated experiments”
  - Type II error/error of the second kind/β error/FN: $p(H_0 | \neg \underline{H_0})$:
    - Power or sensitivity: $p(\neg H_0 | \neg \underline{H_0}) = 1-\beta$

reported	Ref. $\underline{H_0}$	Ref.:$\neg \underline{H_0}$
H_0		Type II
$\neg H_0$	Type I („false rejection”)	

Bayes rule, Bayesianism

„all models are wrong, but some are useful”

$$p(X \mid Y) = \frac{p(Y \mid X)\, p(X)}{p(Y)}$$

A scientific research paradigm

$$p(Model \mid Data) \propto p(Data \mid Model)\, p(Model)$$

A practical method for inverting causal knowledge to diagnostic tool.

$$p(Cause \mid Effect) \propto p(Effect \mid Cause) \times p(Cause)$$

Frequentist vs Bayesian prediction

In the frequentist approach: Model identification (selection) is necessary

$$p(prediction \mid data) = p(prediction \mid BestModel(data))$$

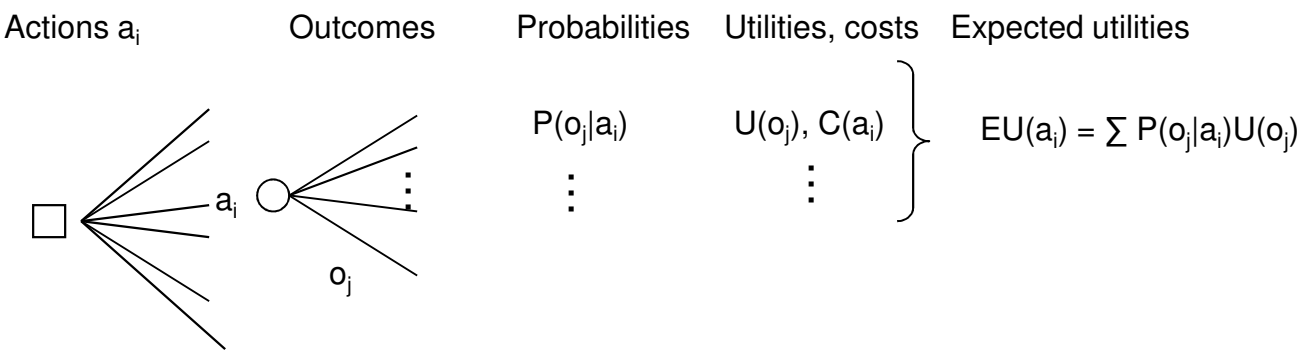
In the Bayesian approach models are weighted

$$p(prediction \mid data) = \sum_i p(pred. \mid Model_i)\, p(Model_i \mid data)$$

Note: in the Bayesian approach there is no need for model selection

Decision theory probability theory+utility theory

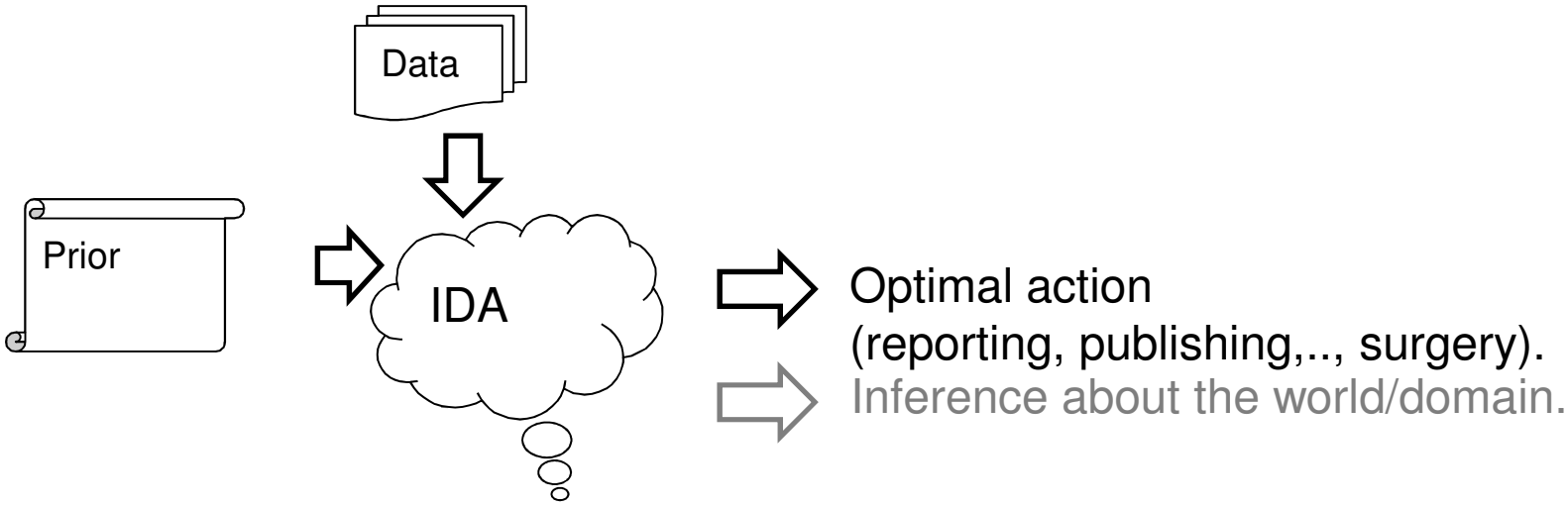
- Decision situation:
  - Actions a_i
  - Outcomes o_j
  - Probabilities of outcomes $p(o_j \mid a_i)$
  - Utilities/losses of outcomes $U(o_j \mid a_i)$
  - Maximum Expected Utility Principle (MEU) $EU(a_i) = \sum_j U(o_j \mid a_i)\, p(o_j \mid a_i)$
  - Best action is the one with maximum expected utility $a^* = \arg\max_i EU(a_i)$



Frequentist vs Bayesian statistics

Frequentist	Bayesian
-	Prior probabilities
Null hypothesis	-
Indirect: proving by refutation	Direct
Model selection	Model averaging
Likelihood ratio test	Bayes factor
p-value	-!
-!	Posterior probabilities
Confidence interval	Credible region
Significance level	Optimal decision based on Exp.Util.
Multiple testing problem	Remains, so → **complex model**
Model complexity dilemma	**Best achievable alternative**
hard to combine p/q-values	**Posteriors induce further distributions**

Data analysis=inductive inference



Types of Machine Learning (i.e. types of data-model-inference)

- unsupervised
- Semi-supervised (reinforcement, 1class)
- Supervised

Types of data

- Observational/Experimental
- Uncertainty? Noise?
- Completeness
- Discrete/Continuous
- Single table/Relational/ContextFreeGrammar
- Dimension?
- Sample size (with respect to dimension)

Models

- Abstraction level/granularity
  - Free text(?)
  - Semi-formal
  - Logical
  - Dependency
  - Causal
  - Parametric
- Conditional vs domain models
- Discrete vs continuous
- Deterministic vs stochastic
- Feedforward vs feedback

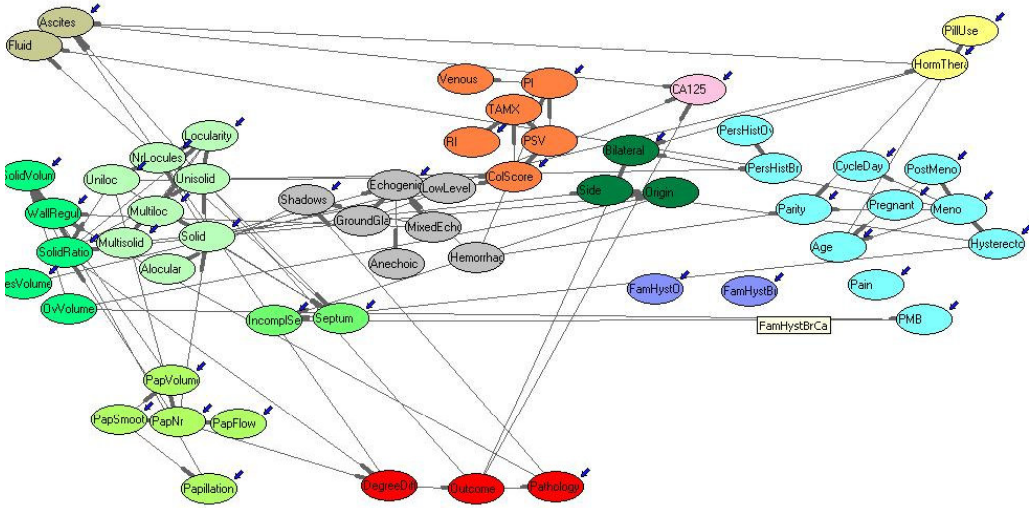
Types of inference

- (Passive, observational) inference
  - P(Query|Observations, Observational data)
- Interventionist inference
  - P(Query|Observations, Interventions)
- Counterfactual inference
  - P(Query| Observations, Counterfactual conditionals)

- Biomedical applications
  - Prevention
  - Screening
  - Diagnosis
  - Therapy selection
  - Therapy modification
  - Evaluation of therapic efficiance

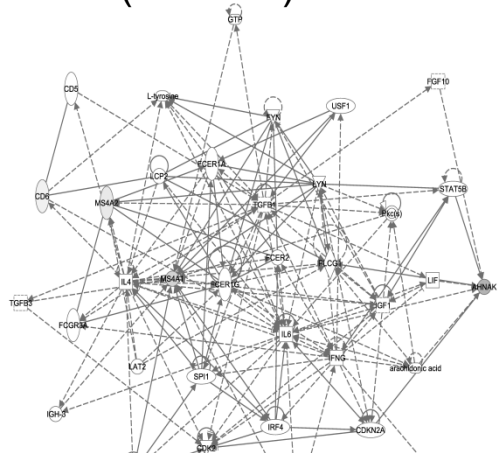
Association graphs, (in)dependence maps, causal networks, control systems

Clusters, modules

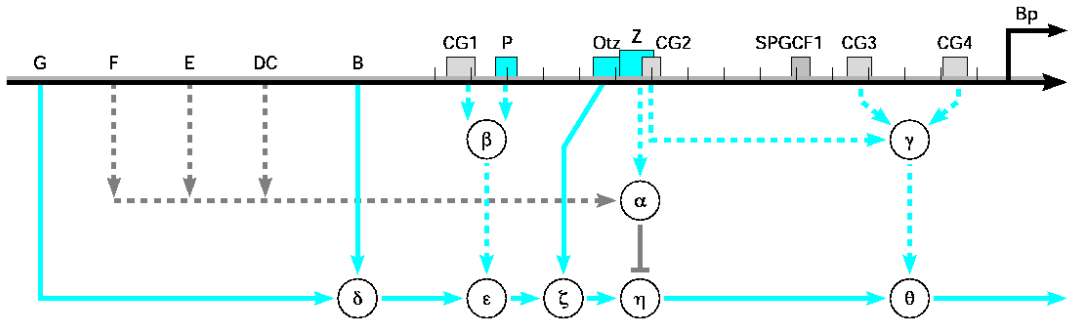


Conditional independencies

(Causal) Mechanisms



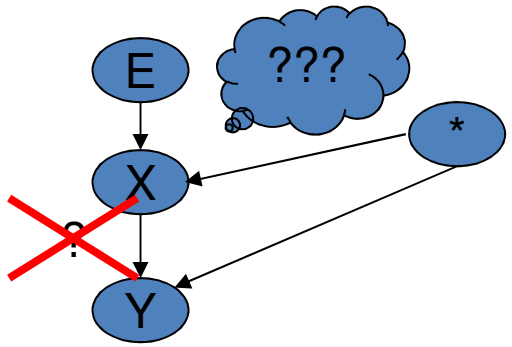
Parameters



Types of data and inference

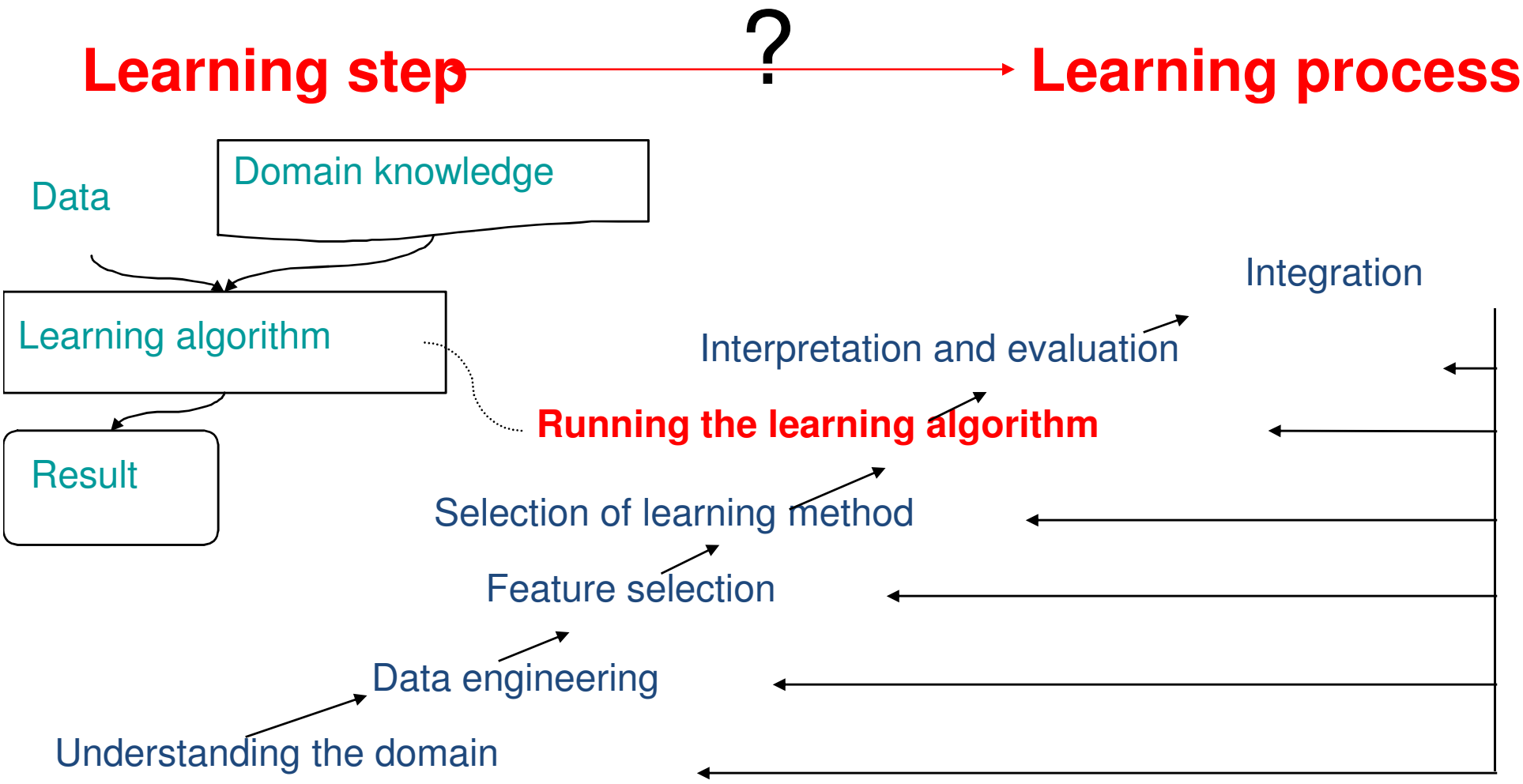
inference\Data	Observational	Interventional
Observational	OK	OK
Interventional	????	OK
Counterfactual	??????	??

- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded



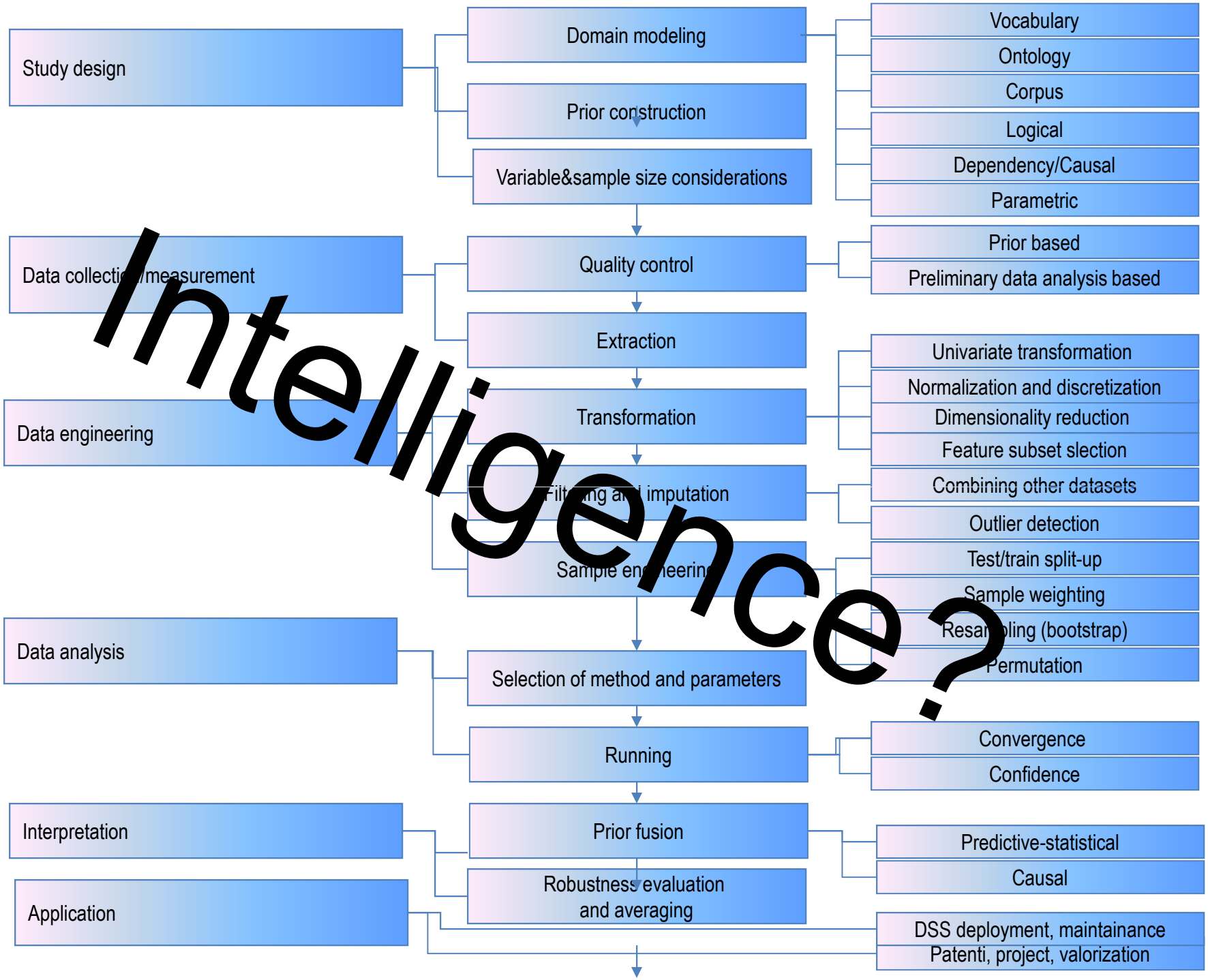
„Plato's two surprises:
1. Not all true theorems can be proved
2. Causal inference is possible from observations"

Learning step or learning **process?**

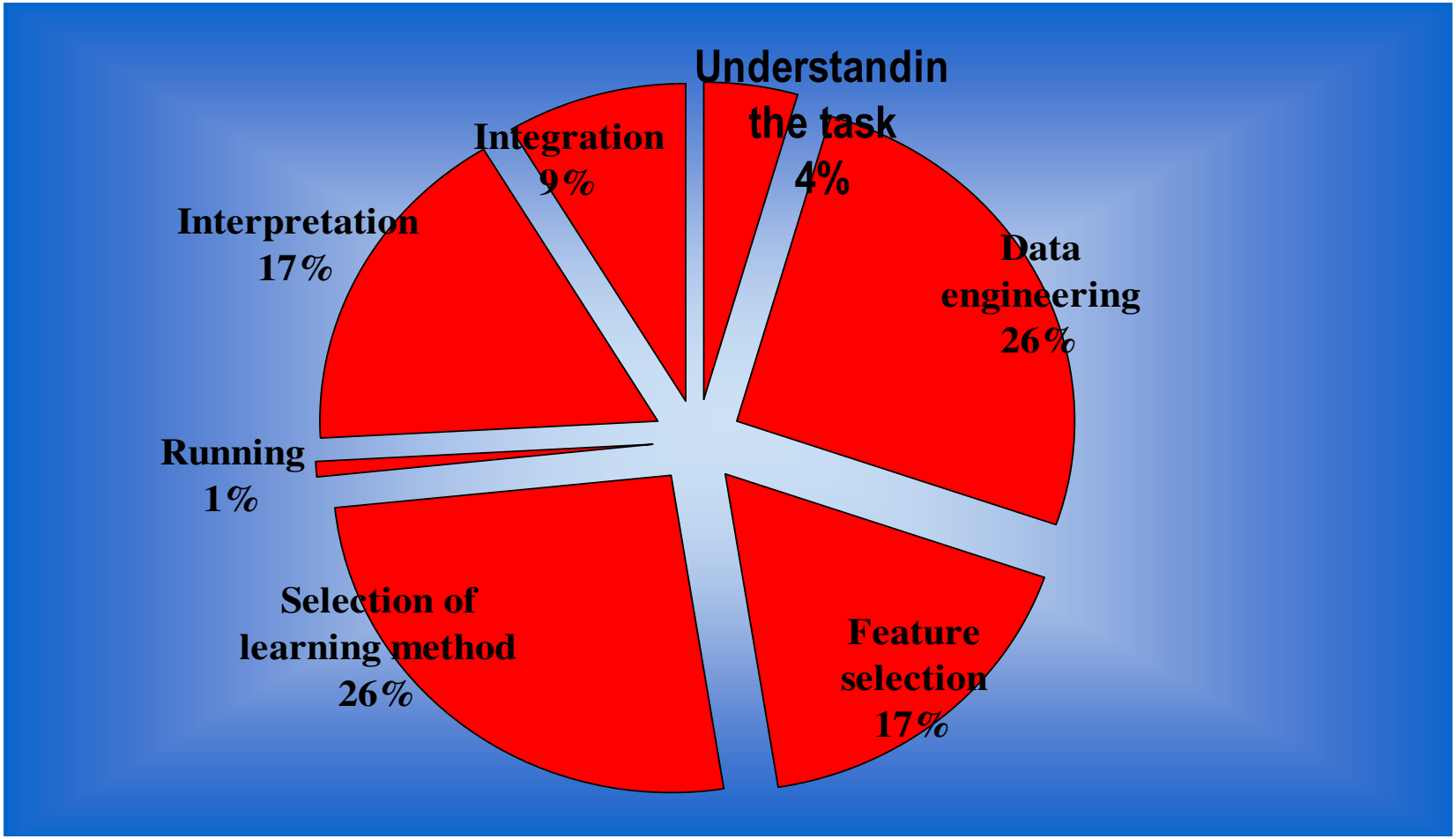


Data analysis in practice

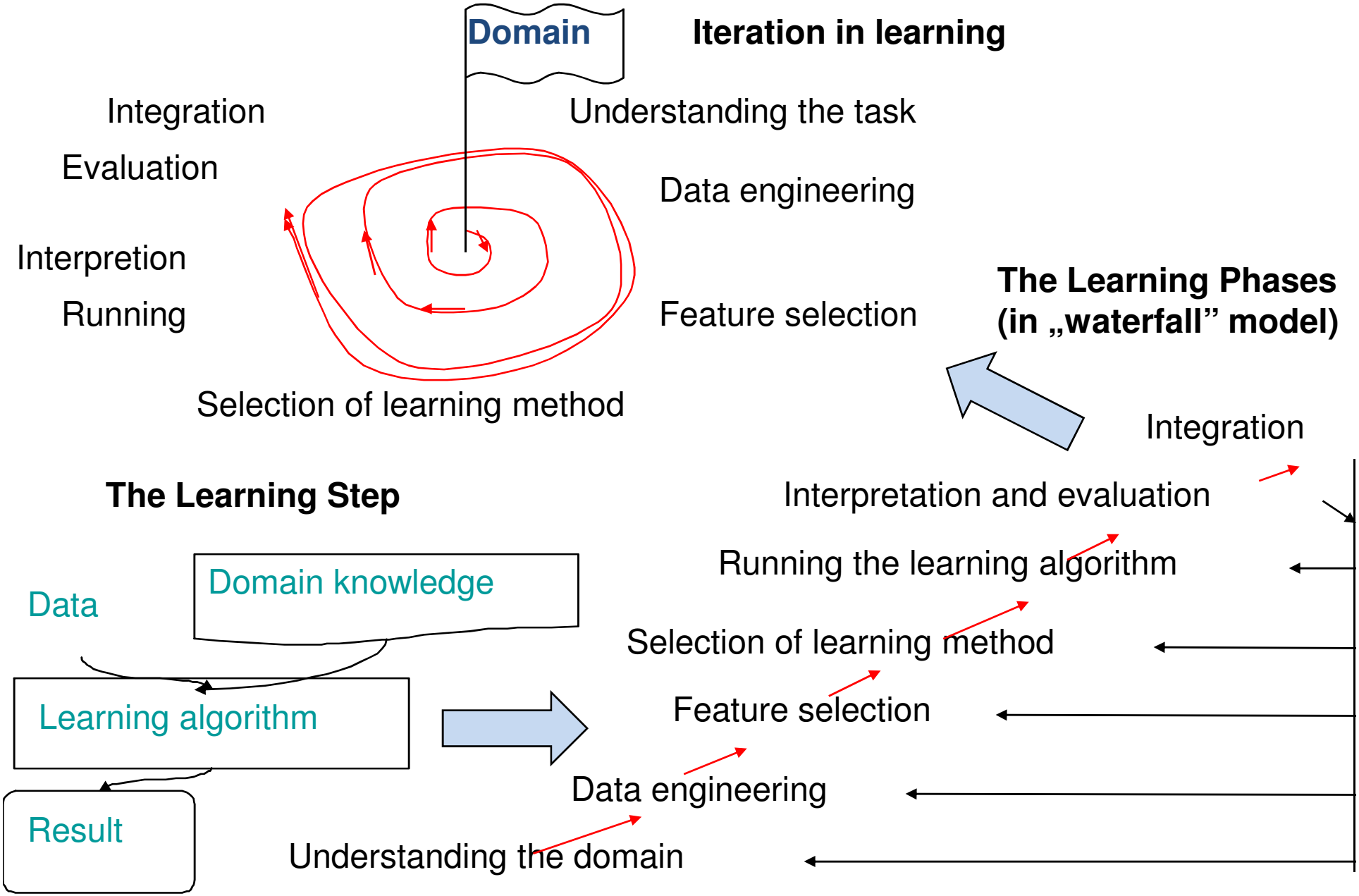
- Text mining
- Study design
- Data engineering
- Analysis
- Interpretation
- Application


Study design
Domain modeling
Prior construction
Variable&sample size considerations
Vocabulary
Ontology
Corpus
Logical
Dependency/Causal
Parametric
Data collection/measurement
Quality control
Prior based
Preliminary data analysis based
Extraction
Data engineering
Transformation
Univariate transformation
Normalization and discretization
Dimensionality reduction
Feature subset slection
Filtering and imputation
Combining other datasets
Outlier detection
Sample engineering
Test/train split-up
Sample weighting
Resampling (bootstrap)
Permutation
Data analysis
Selection of method and parameters
Running
Convergence
Confidence
Interpretation
Prior fusion
Predictive-statistical
Causal
Application
Robustness evaluation and averaging
DSS deployment, maintainance
Patenti, project, valorization
Intelligence?

Length of phases



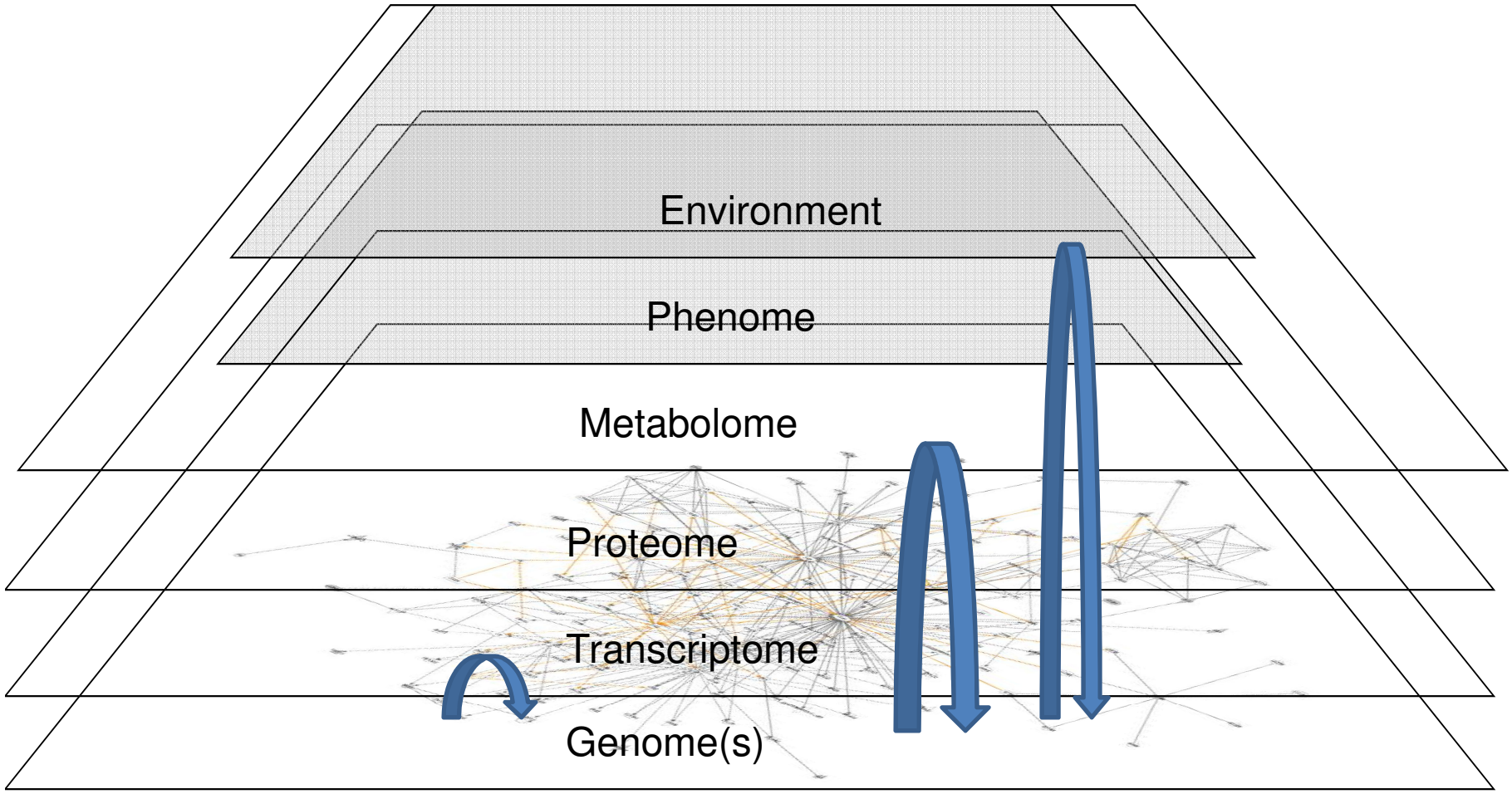
Forwards vs **iterative process**?



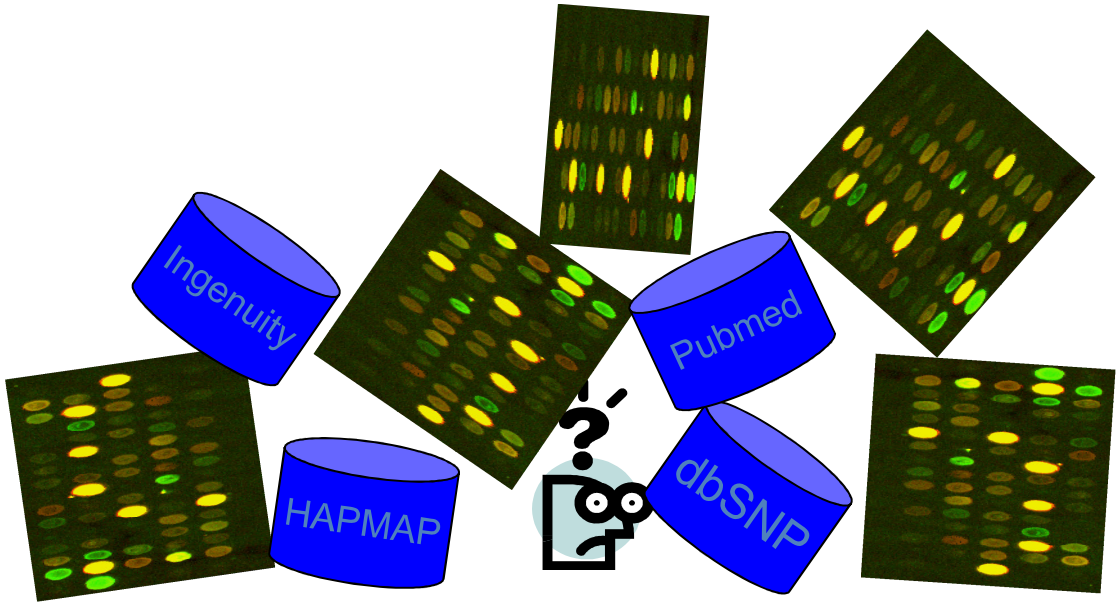
Bayesian positivism

- Positivism (19th century-)
  - experience-based knowledge
- Logical positivism (1920-)
  - L. Wittgenstein: all knowledge should be codifiable in a single standard language of science + logic for inference
- Bayesian(-www) positivism
  - Data are available in public repositories
  - Scientific papers are available public repositories
  - In a formal, single probabilistic representation the results of statistical data analyses are available in KBs
  - In a formal, single probabilistic representation models, hypotheses, conclusions linked to data are available in KBs
  - …

Multiple levels in biomedicine

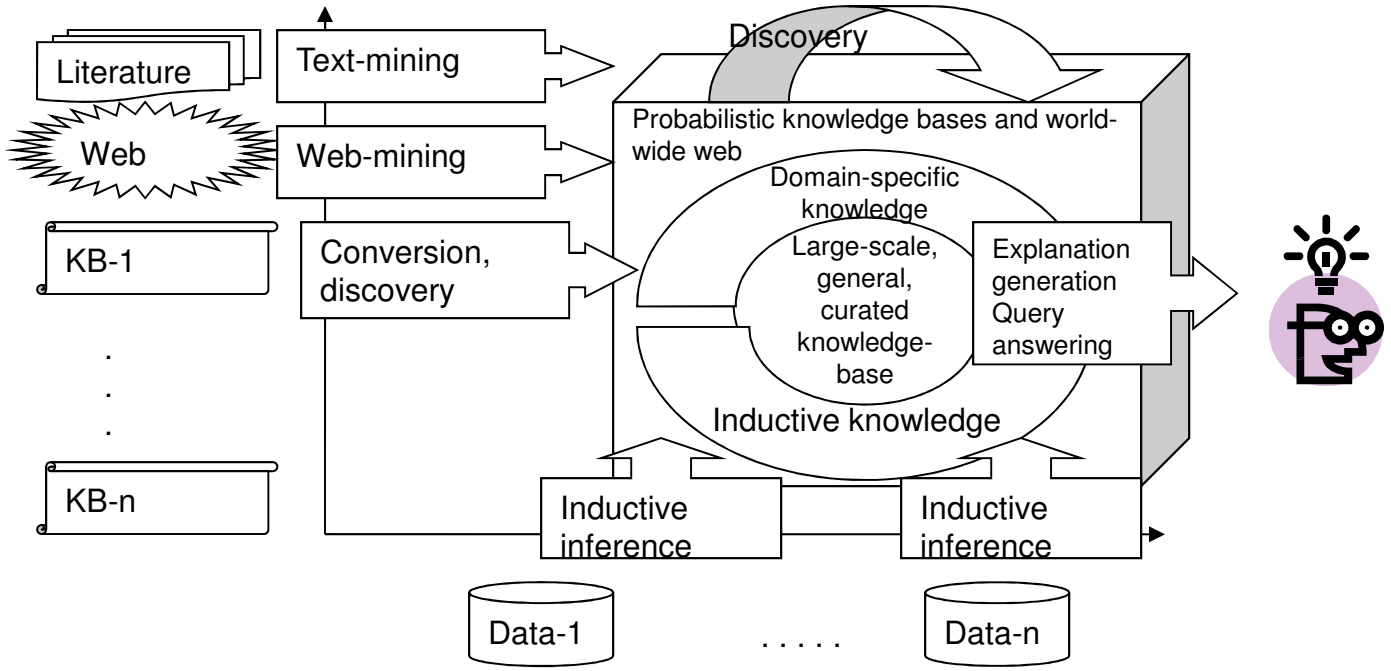


The interpretational bottleneck (~the fusion challenge)



- Free text repositories (with significances): e.g. SNPedia
- Manually curated logical knowledge bases: Ingenuity

Fusion using very large scale probabilistic logic KBs



- **multivariate Bayesian data analysis**,
- the use of more powerful **logic for representing knowledge**,
- **uncertainty management** in knowledge representation, induction, and inference,

syntactic semantics: $P(\phi \mid KB) = \sum_{\text{pKB}} P(\phi \text{ is provable} \mid pKB, lKB) p(pKB)$

“Ambient assisted” data analysis

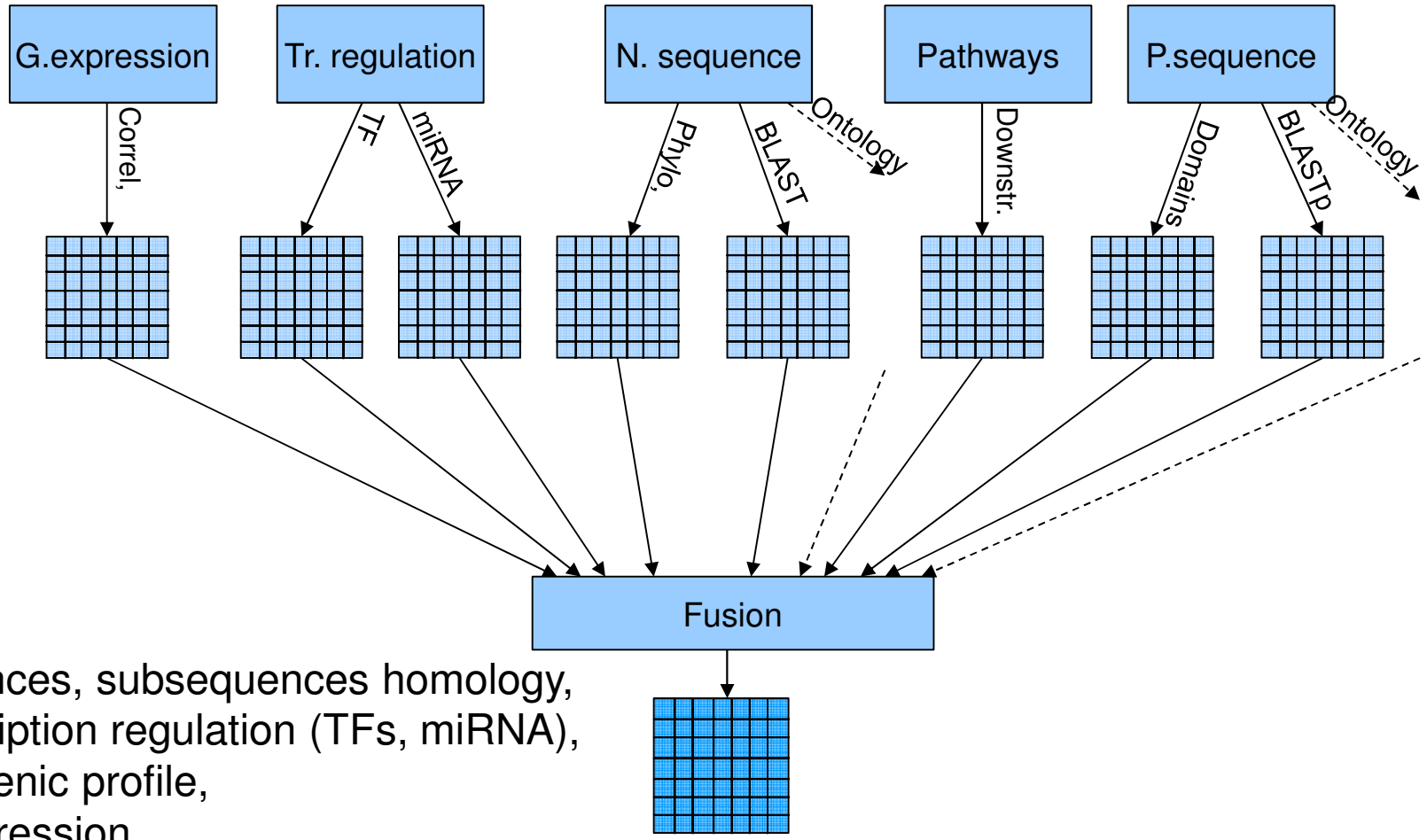
- J.Lamb: CMap
- The ultimate objective of biomedical research is to connect human diseases with the genes that underlie them and drugs that treat them. But this remains a daunting task, and even the most inspired researchers still have to resort to laborious screens of genetic or chemical libraries. What if at least some parts of this screening process could be systematized and centralized? **And hits found and hypotheses generated with something resembling an internet search engine?** These are the questions the Connectivity Map project set out to answer.

- http://www.altcancerweb.com/osteosarcoma/drug-design/connectivity-map-nat-rev-cancer-2007.pdf

Our objective, therefore, is to keep all of the computation in the background and provide a single analysis tool with no ‘knobs’ whatsoever; queries are executed from our website in real-time with a single click.

Gene Prioritization (GP)

GP Problem: Prioritizing relevance of genes to a given set/phenomena.



Similarity of

- sequences, subsequences homology,
- transcription regulation (TFs, miRNA),
- phylogenic profile,
- co-expression,
- homology, co-location, or common complex of products,
- taxonomic/semantic (Gene ontology),
- co-citation,
- role in pathway knowledge-bases.

Search engine in biomed? (CMAP: mRNA as universal language?)

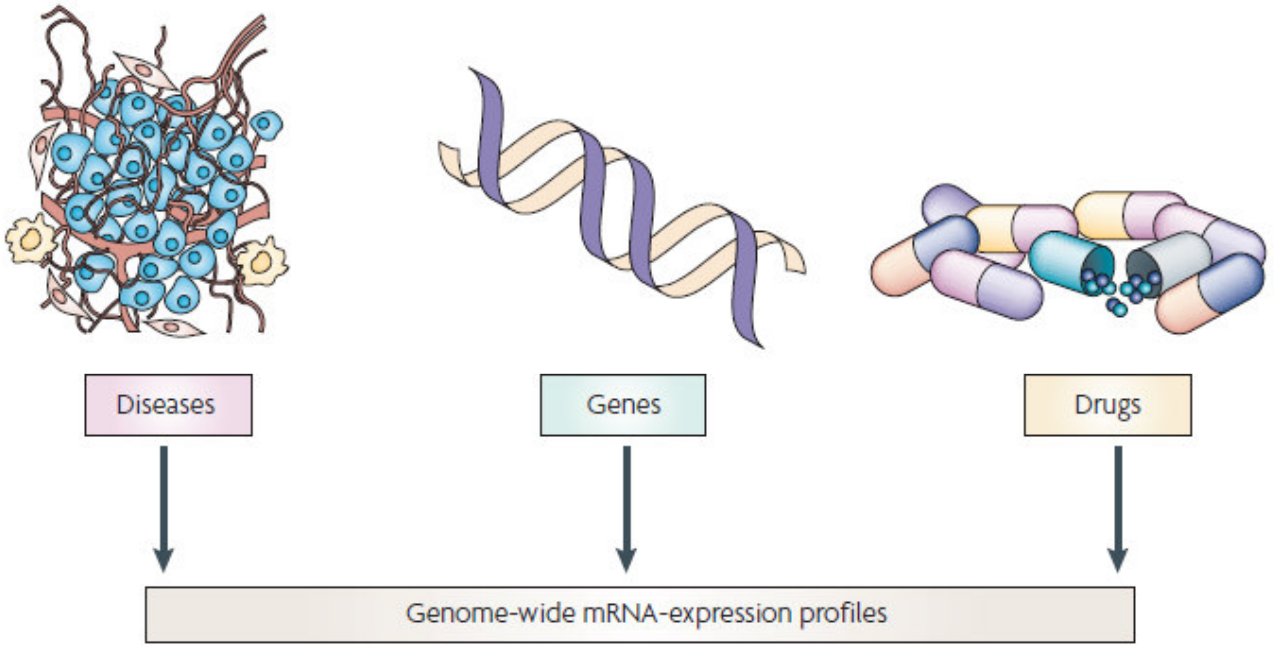
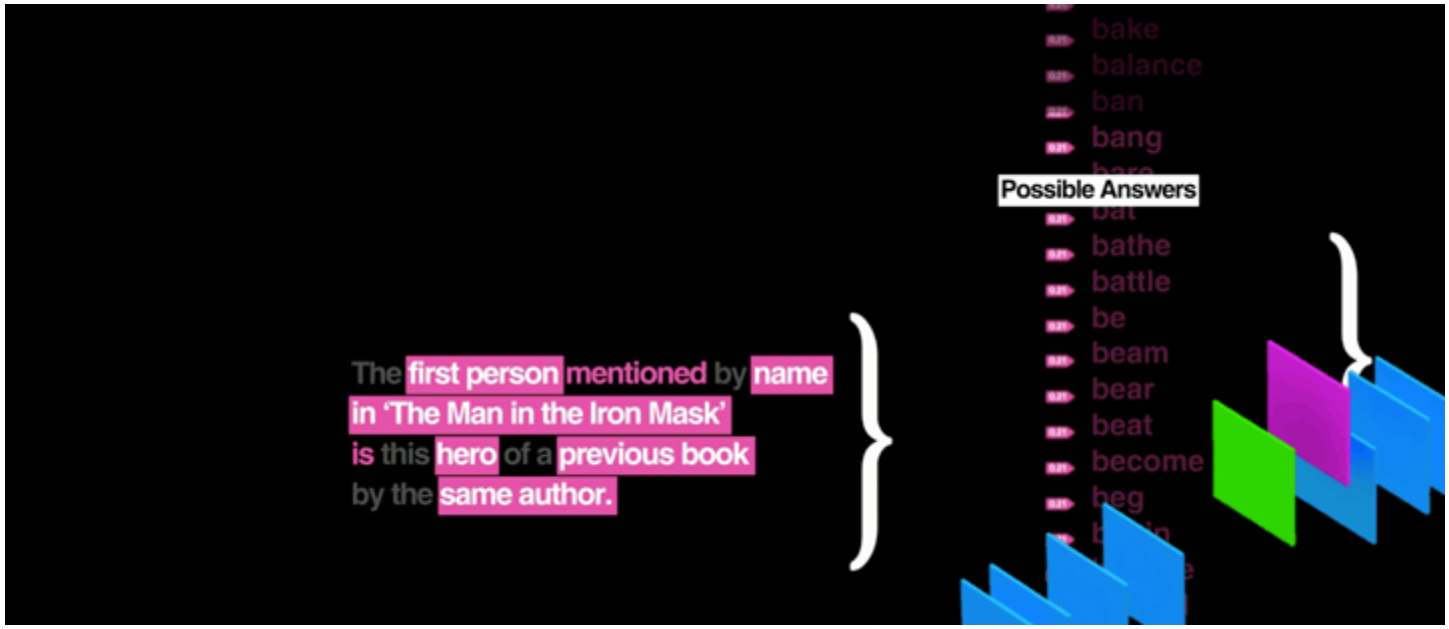


Figure 2 | **A universal functional bioassay.** As all of the transcripts are now known, and robust technologies for their simultaneous measurement are available, it is possible to capture objective high-dimensional depictions of all induced or organic biological conditions in a common global analytical space, and thereby readily appreciate similarities between them.

- Justin Lamb: The Connectivity Map: a new tool for biomedical research, Nature, 7,pp 54-60, 2007

Watson?
The Science Behind an Answer

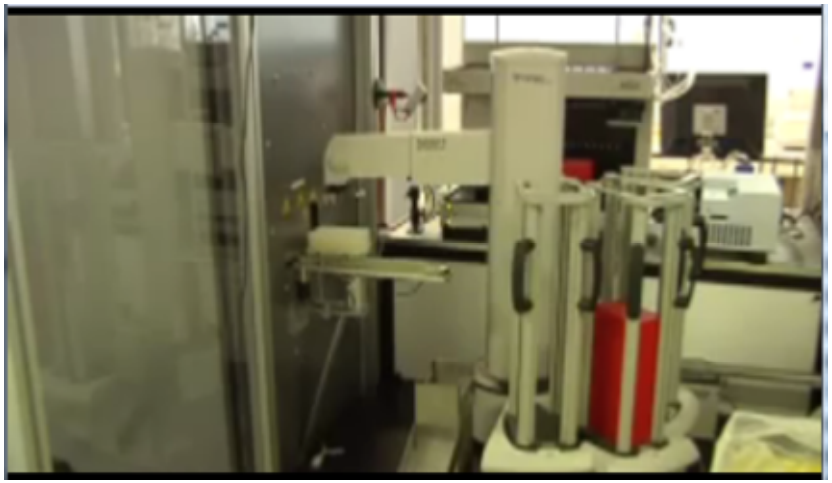


- http://www-03.ibm.com/innovation/us/watson/what-is-watson/science-behind-an-answer.html

IBM WATSON

Automated discovery systems

- Langley, P. (1978). Bacon: A general discovery system. Proceedings of the Second Biennial Conference of the Canadian Society for Computational Studies of Intelligence (pp. 173-180). Toronto, Ontario.
- …
- Chrisman, L., Langley, P., & Bay, S. (2003). Incorporating biological knowledge into evaluation of causal regulatory hypotheses. Proceedings of the Pacific Symposium on Biocomputing (pp. 128-139). Lihue, Hawaii.
- (Gene prioritization…)
- R.D.King et al.: The Automation of Science, Science, 2009



Lectures

1. Intro:
    1. Data-rich science
    2. bayesian dec theory as foundation
    3. +causal inference
2. Learnability: consistency, no free lunch, sample complexity, VC dimension
3. Bayesian inference: basics
4. Hidden Markov Models: inference, parameter learning (EM)
5. Graphical models: Markov networks + Bayesian networks
6. Posterior over parameters of BNs
7. Posterior over structural features of BNs
8. Resampling techniques: bootstrap, permutation tests
9. Monte Carlo methods
10. Causal inference
11. Assocation and relevance analysis, the featre subset selection problem
12. Inductive Logic Programming, Bayes Logic Programming
13. Incomplete data
14. Clustering
15. Fusion: prior and posterior fusion, rank/order statistics/kernel-fusion
16. Outlier detection: low prob, opt dec, 1class
17. Case study: genetic association analysis: haplotypes, imputation, confounders, compliance, effect strength
18. Case study: systems biology based relevance analysis: BayesEye