Intelligent data analysis
Introduction

Peter Antal antal@mit.bme.hu

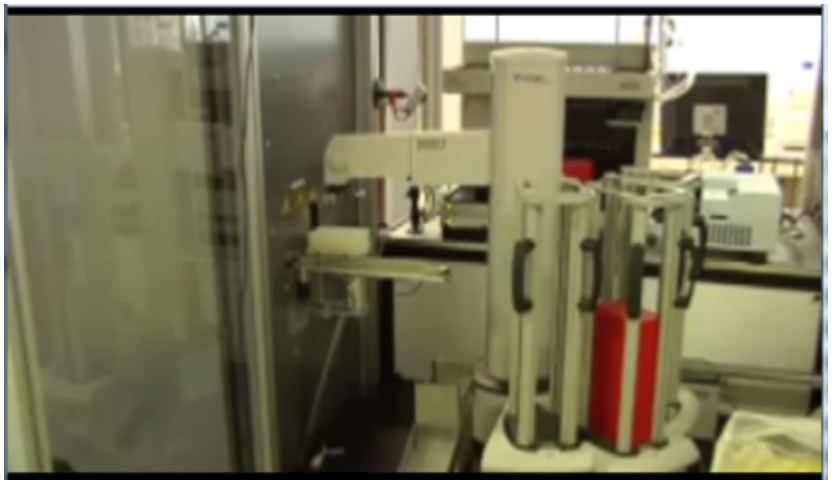
A historical overview of IDA

- Imitating human intelligence: automated statistician
- Efficient computational (time,space) complexity
- Incorporation of prior knowledge
  - Knowledge-based neural networks
  - Adaptive expert systems
    - Logic-based: inductive logic programming
    - Probabilistic: Bayesian networks
- Learning causal relations and models
- Adaptive study design, active learning
- Multitask learning, transfer learning, deep learning,..
- Data and knowledge fusion
- Interpretation of the results
- Integration of the results (gathering the gold dust)
- Big data: scalable, real time data analysis
- Google: automated statistician

Data analysis process
Bayesian decision theoretic foundation
Causal inference
Open access data, publication, model, computation

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**

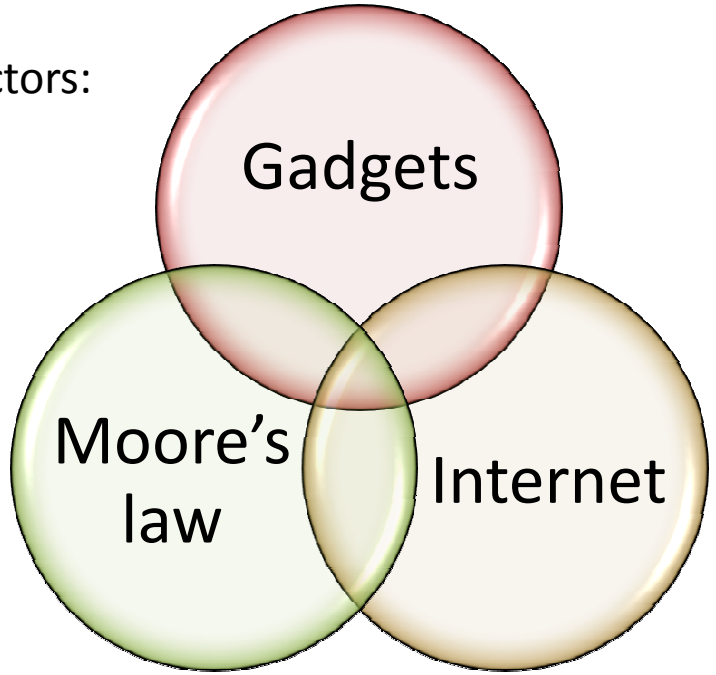


The „Common" big data

- Financial transaction data, mobile phone data, user (click) data, e-mail data, internet search data, social network data, sensor networks, ambient assisted living, intelligent home, wearable electronics,...

"The line between the virtual world of computing and our physical, organic world is blurring." E.Dumbill: Making sense of big data, Big Data, vol.1, no.1, 2013

Factors:



Definitions of „big data”

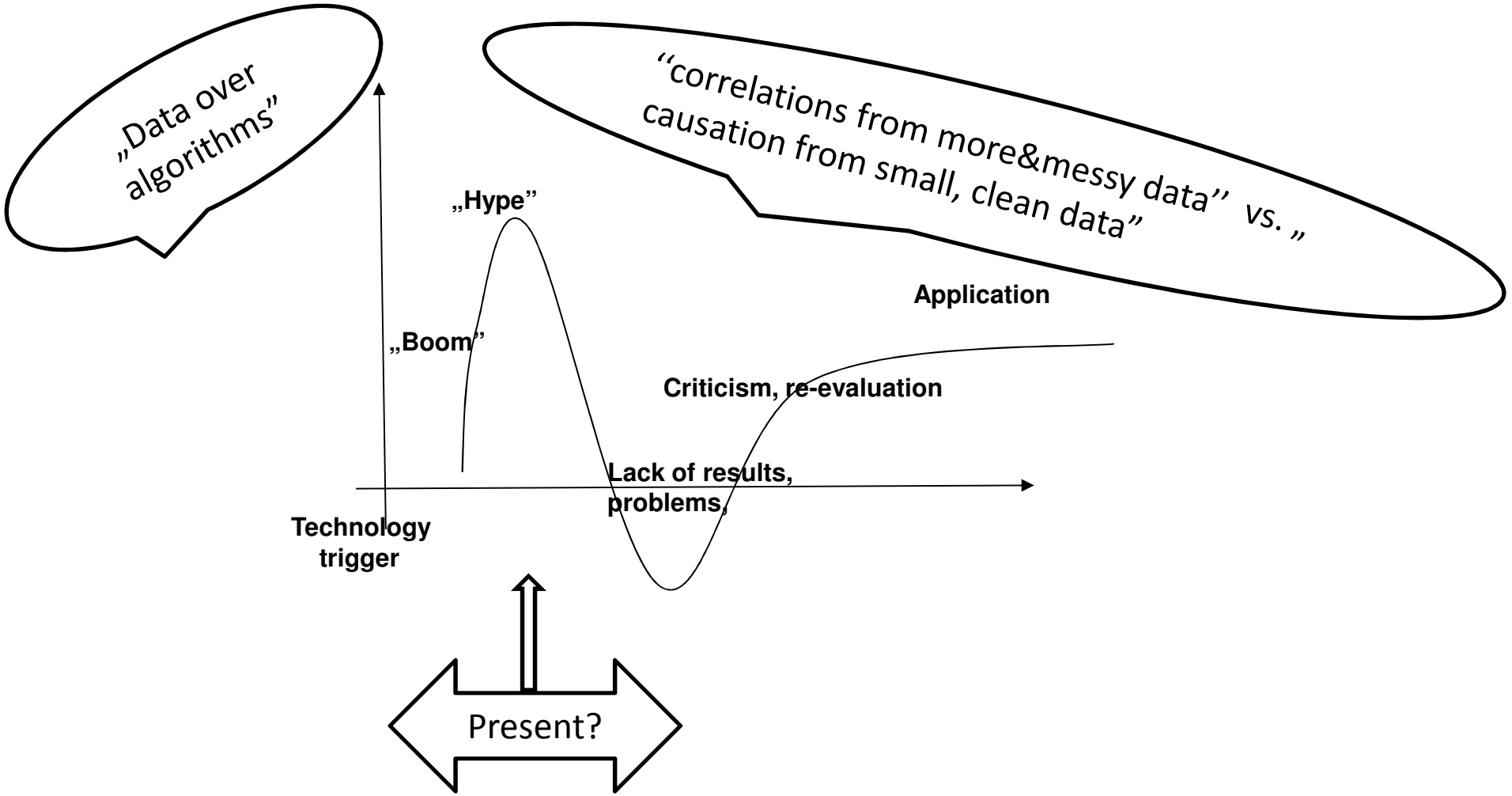
M. Cox and D. Ellsworth, “Managing **Big Data** for Scientific Visualization,” Proc. ACM Siggraph, ACM, 1997

The 3xV: **volume, variety, and velocity** (2001).

The 8xV: Vast, Volumes of Vigorously, Verified, Vexingly Variable Verbose yet Valuable Visualized high Velocity Data (2013)

Not „conventional” data: „Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn’t fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it (E.Dumbill: Making sense of big data, Big Data, vol.1, no.1, 2013)

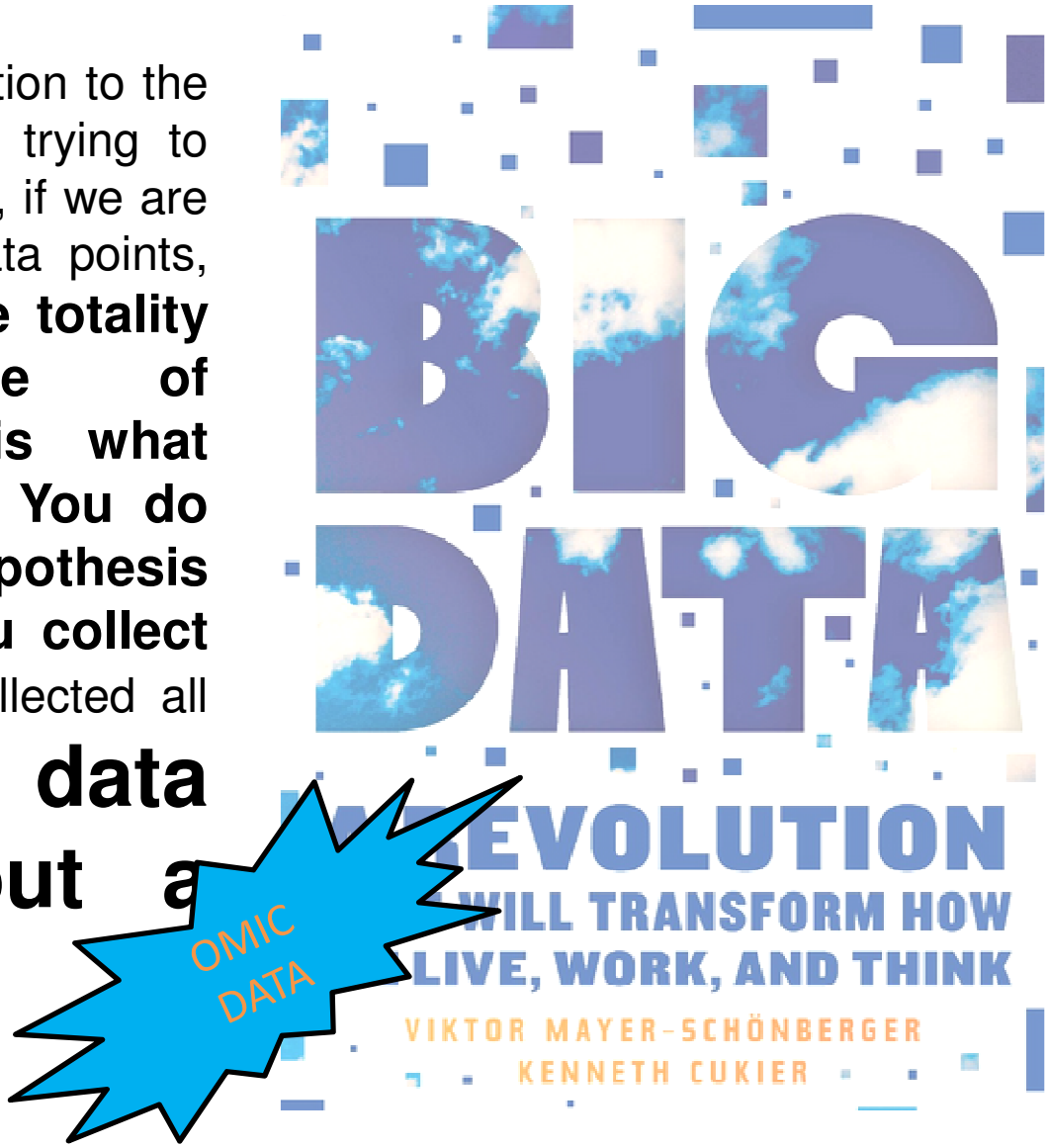
„Big data, big hype, big danger"??



Definitions of „big data”, cont’d

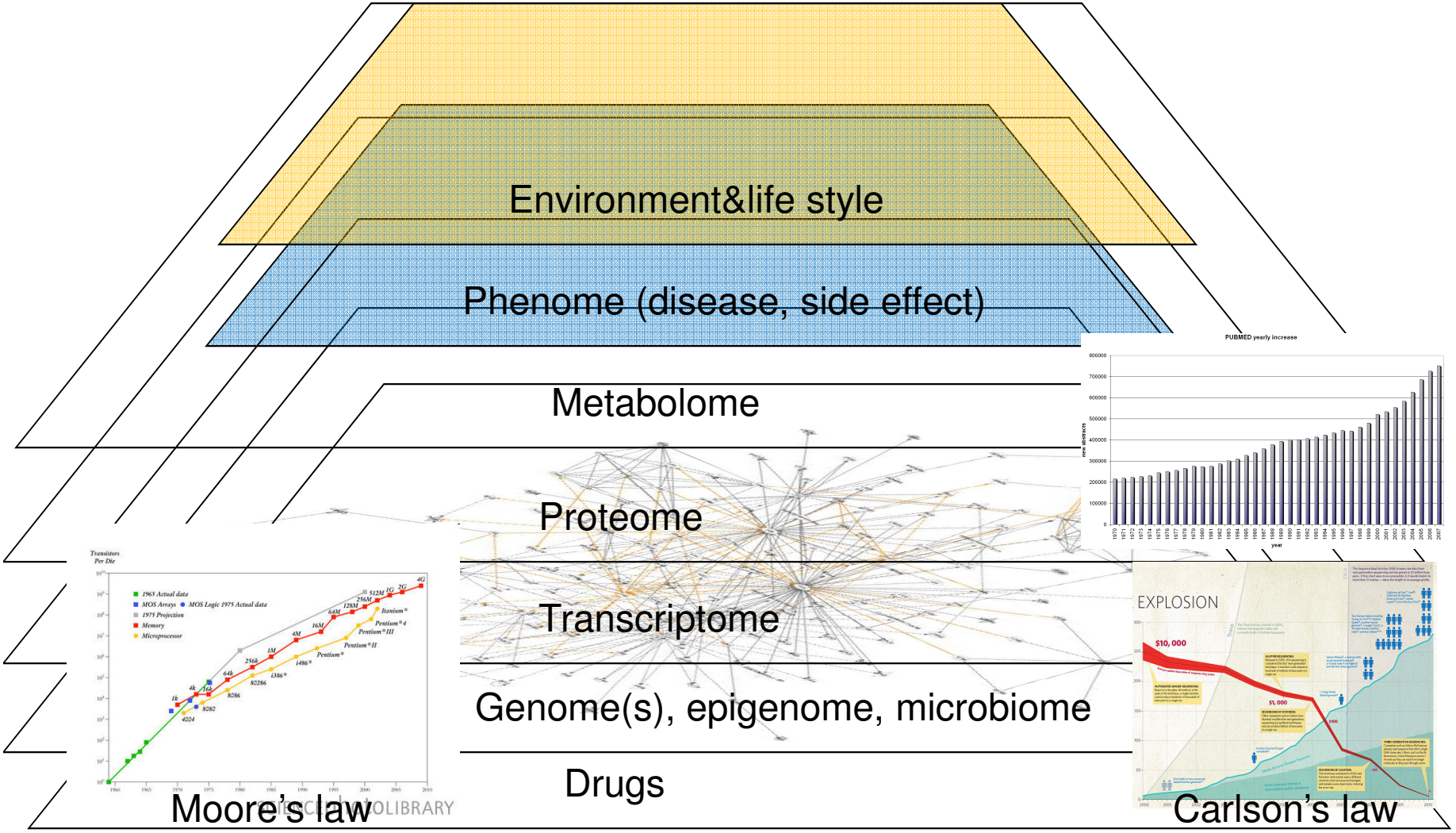
Definitions of „big data", cont'd

.. [data] is often big in relation to the phenomenon that we are trying to record and understand. So, if we are only looking at 64,000 data points, but **that represents the totality or the universe of observations. That is what qualifies as big data. You do not have to have a hypothesis in advance before you collect your data**. You have collected all there is—**all the data there is about a phenomenon.**

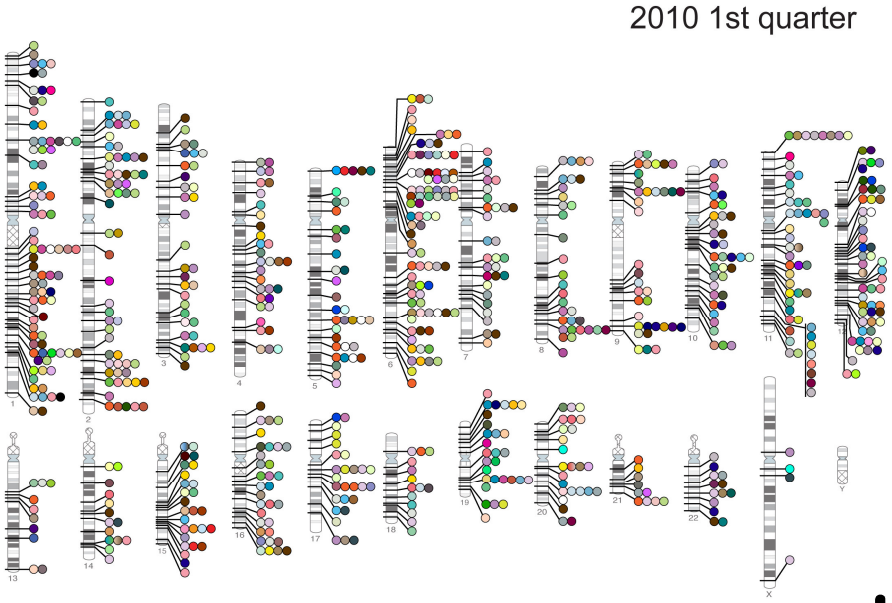


Biomedical omic data/big data

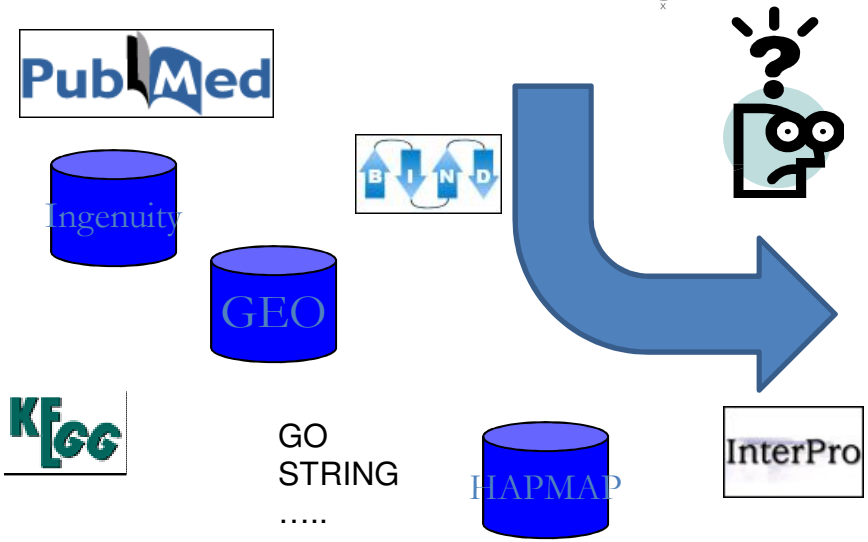
2010<: “Clinical phenotypic assay”/drugome: open clinical trials, adverse drug reaction DBs, adaptive licensing, Large/scale cohort studies (~100,000 samples)

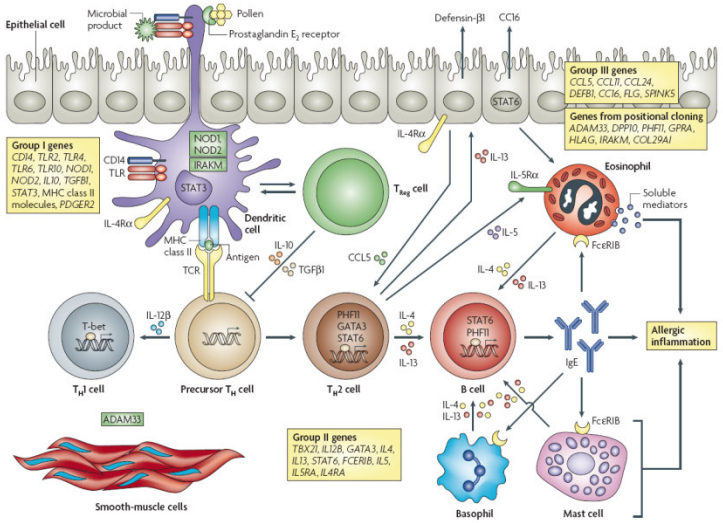


The hypothesis-free pitfall in omics



- Hypothesis-free measurement
- Hypothesis-free data analysis
- Interpretational/translational bottleneck

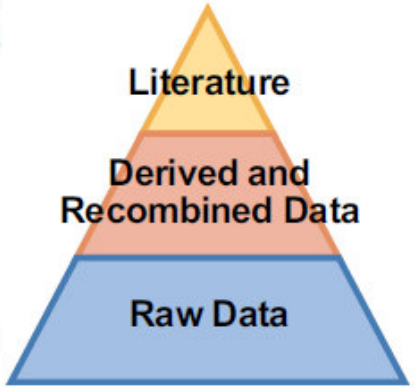


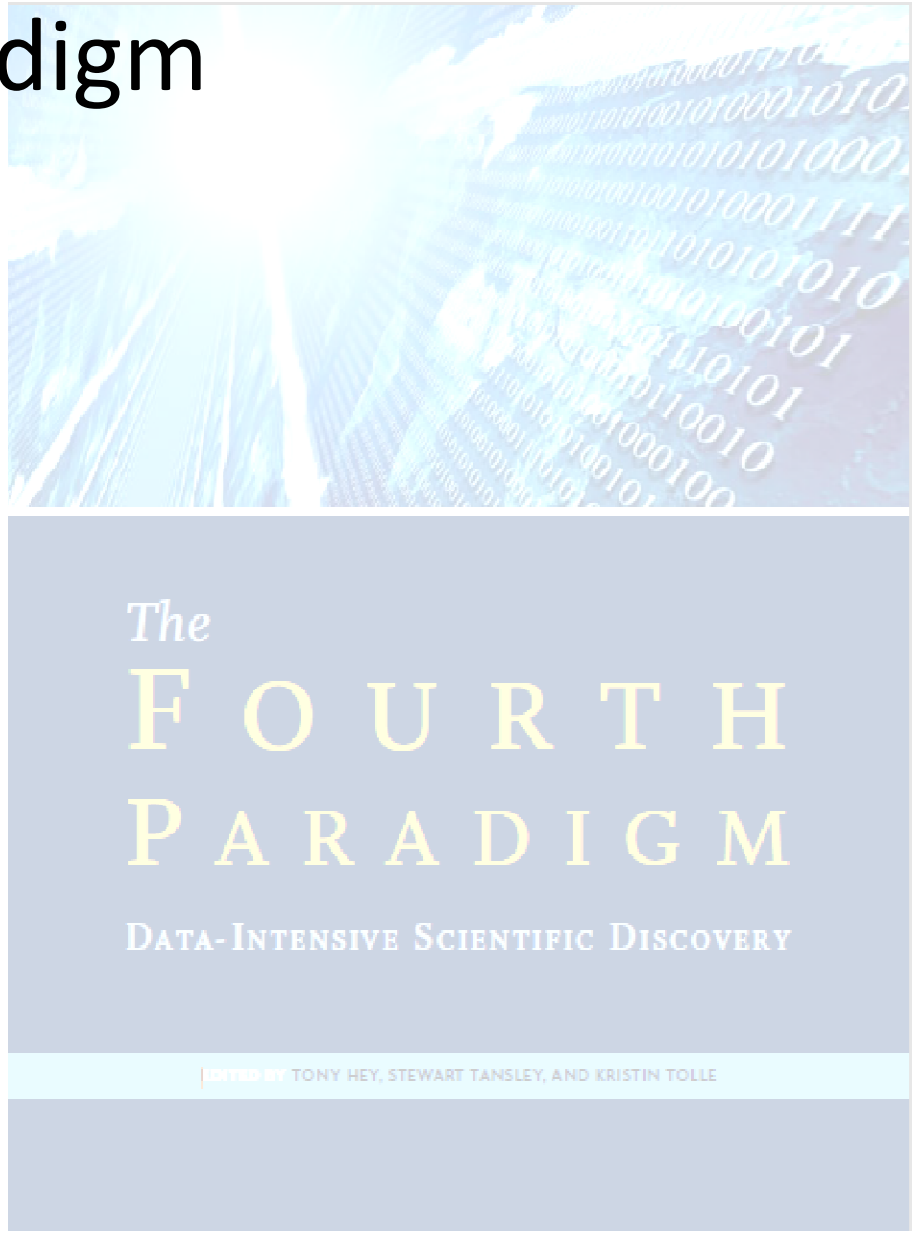


E-science, data-intensive science, the fourth paradigm

All Scientific Data Online

- Many disciplines overlap and use data from other sciences
- Internet can unify all literature and data
- Go from literature to computation to data back to literature
- Information at your fingertips for everyone-everywhere
- Increase Scientific Information Velocity
- Huge increase in Science Productivity





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!!)

„Data-driven" 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
- Data/www-driven 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
  - …

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"
  - Goal is an optimal action (and not model identification)

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;

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

- The three „as if” (representation) theorems:
  - Uncertainty by probabilities
  - Preferences by utility function
  - Optimal action by maximum expected utility principle

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

A chronology

- [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**
    - „given a specified number of white and black balls in an urn, what is the probability of drawing a black ball?"
    - 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 number of white and black balls in 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)
    - [Bayesianism is a] „fallacious rubbish"
    - His own approach was „Fiducial inference" ~ Bayesian statistics
    - He used informed priors in genetics

A chronology (cont'd)

- [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
- [1989] Exact general inference methods, S. Lauritzen

- … Markov Chain Monte Carlo methods – GPGPUs…

Bayes-omics

- **Thomas Bayes (c. 1702 – 1761)**
- Bayesian probability
- Bayes' rule
- Bayesian statistics
- Bayesian decision
- Bayesian model averaging

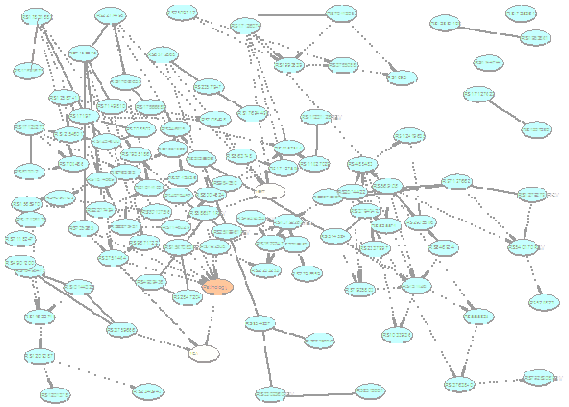
- Bayesian networks
- Bayes factor
- Bayes error
- Bayesian „communication"
- ...

$$p(\text{Model}|\text{Data}) \propto p(\text{Data}|\text{Model})\,p(\text{Model})$$

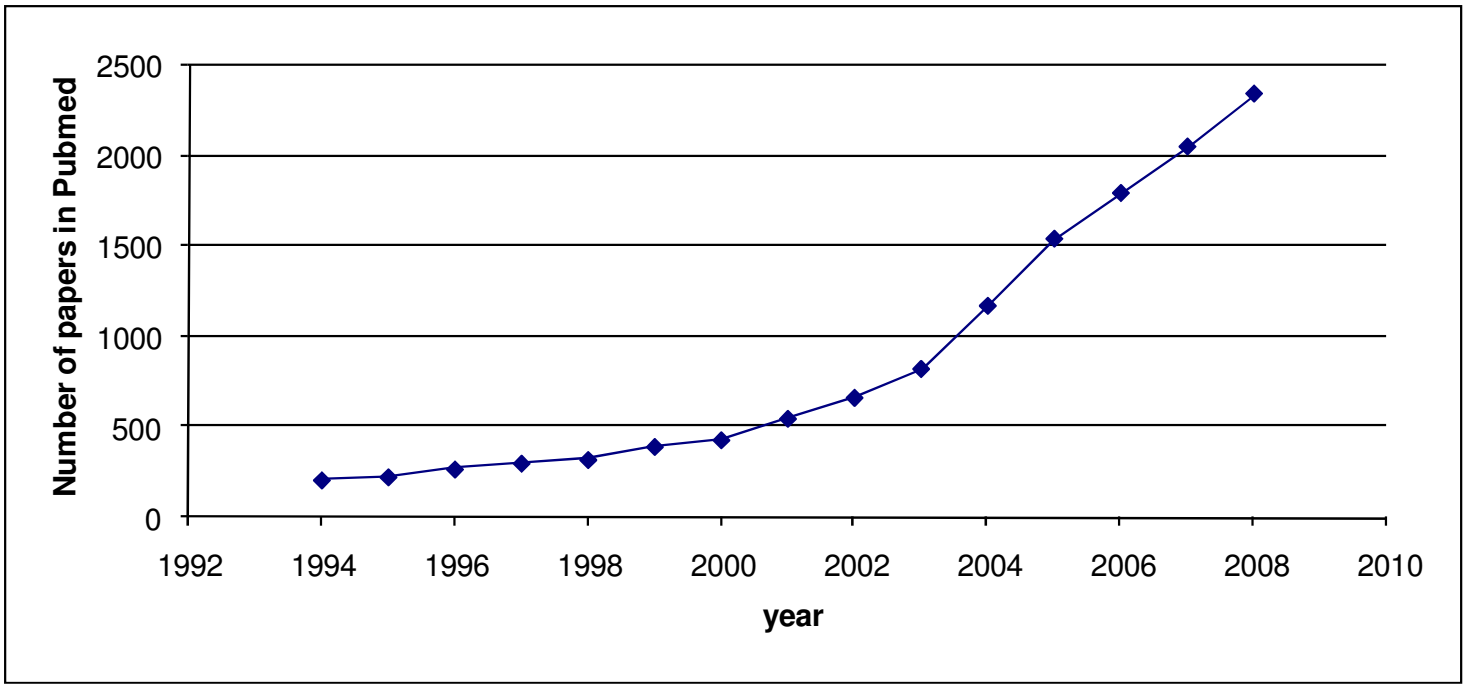
$$a^* = \arg\max_i \sum_j U(o_j)\,p(o_j \mid a_i)$$

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

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



Bayes in Pubmed



Number of papers in Pubmed for query „Bayesian”

What is the reason?

- Yet another statistical approach (in the hype phase)?
- A (statistical) paradigm shift?
- A new scientific paradigm?
- Something practical?

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!

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

Bayesian model averaging

View learning as Bayesian updating of a probability distribution over the hypothesis space

H is the hypothesis variable, values $h_1, h_2, \ldots$, prior $\mathbf{P}(H)$

jth observation d_j gives the outcome of random variable D_j
training data $\mathbf{d} = d_1, \ldots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

where $P(\mathbf{d}|h_i)$ is called the likelihood

Predictions use a likelihood-weighted average over the hypotheses:

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

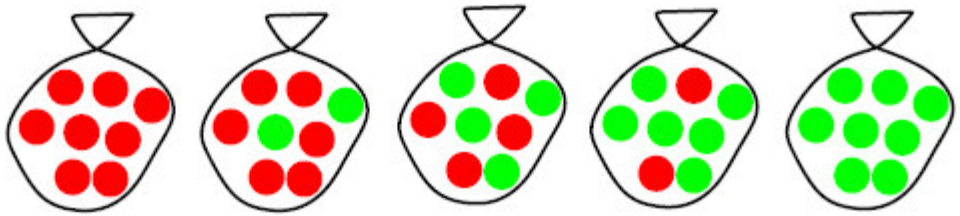
No need to pick one best-guess hypothesis!

Russel&Norvig: Artificial intelligence, ch.20

Bayesian Model Averaging example

Suppose there are five kinds of bags of candies:

10% are h_1: 100% cherry candies
20% are h_2: 75% cherry candies + 25% lime candies
40% are h_3: 50% cherry candies + 50% lime candies
20% are h_4: 25% cherry candies + 75% lime candies
10% are h_5: 100% lime candies

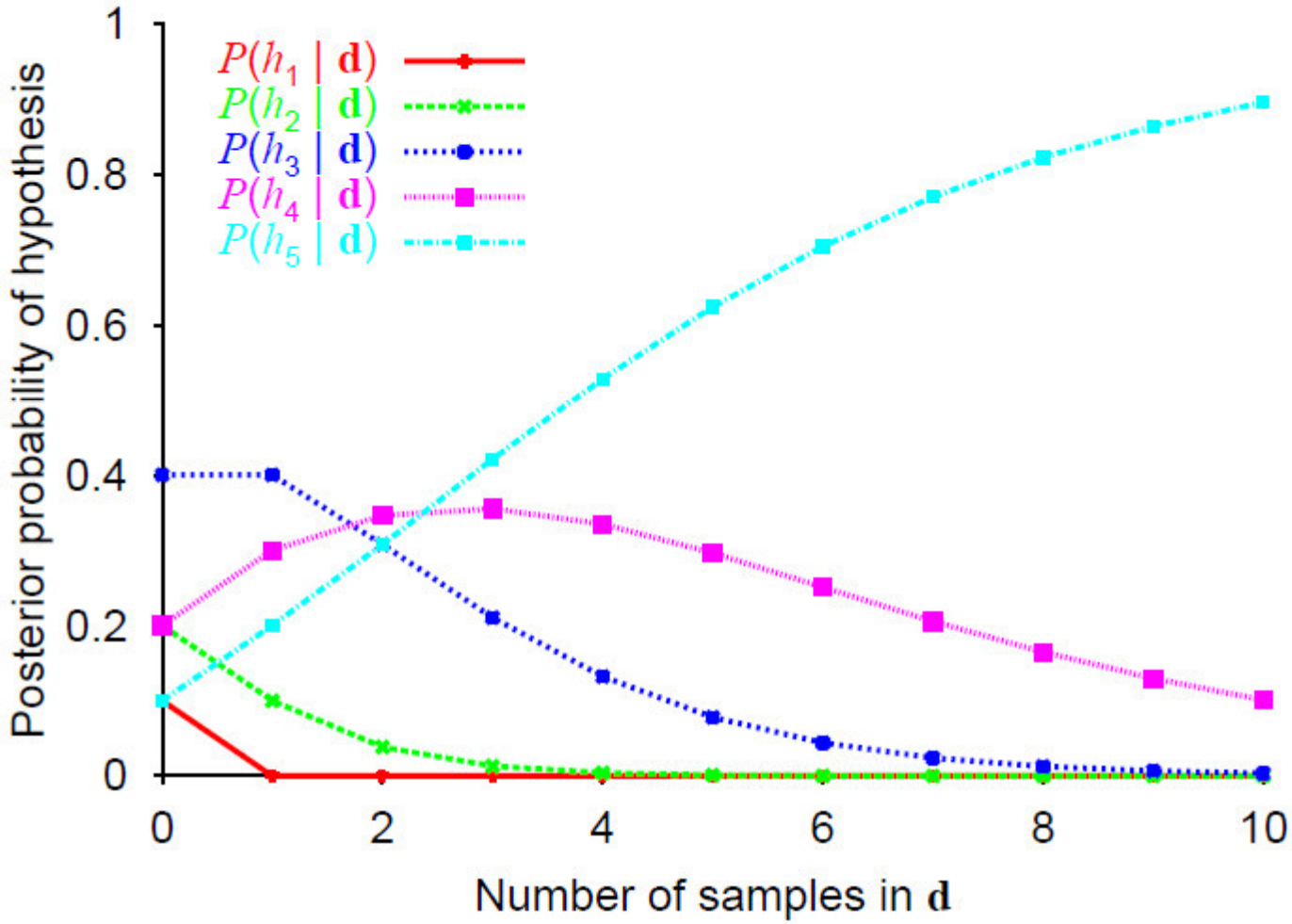


Then we observe candies drawn from some bag: 

What kind of bag is it? What flavour will the next candy be?

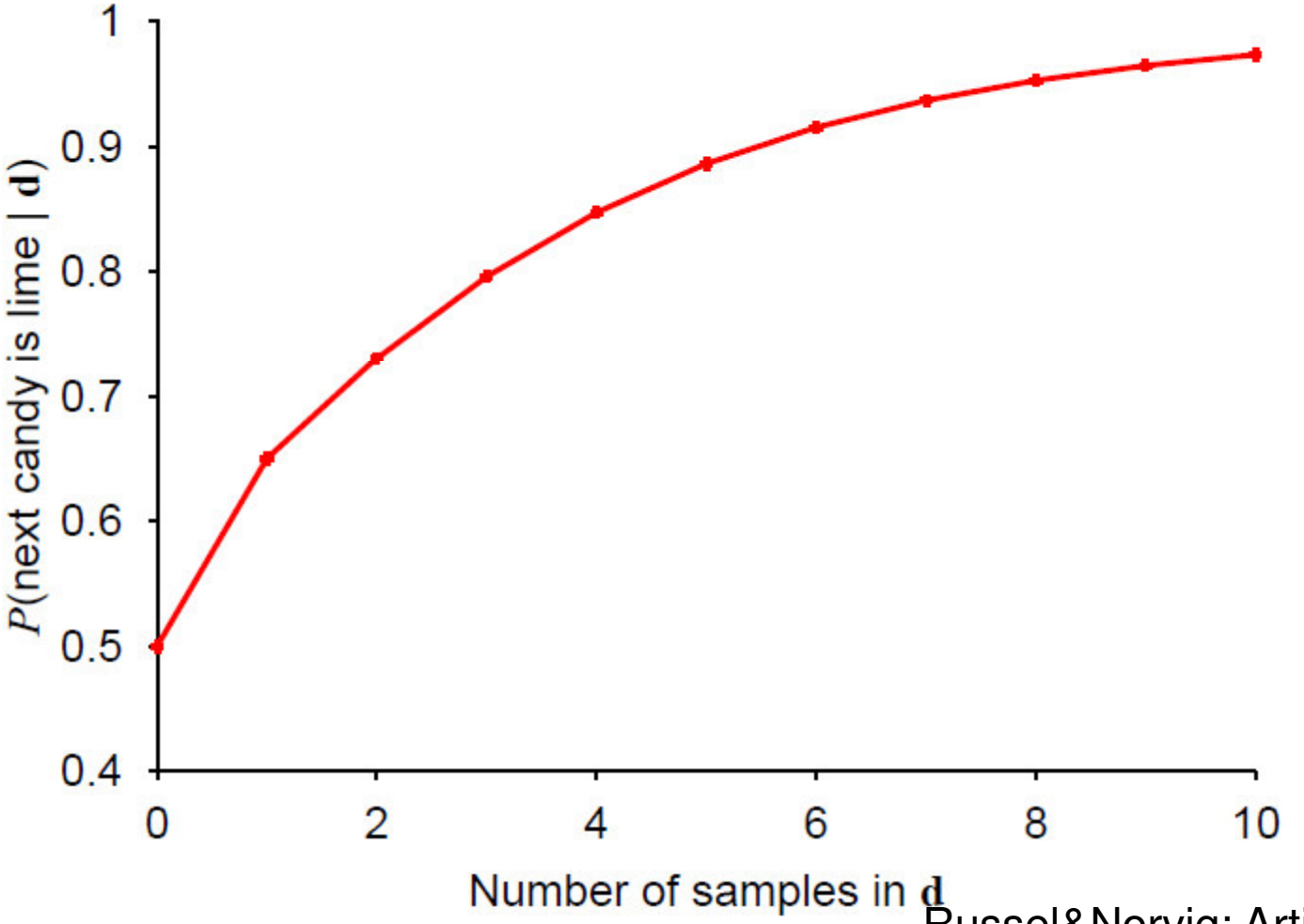
Russel&Norvig: Artificial intelligence

Learning rate for hypotheses



Russel&Norvig: Artificial intelligence

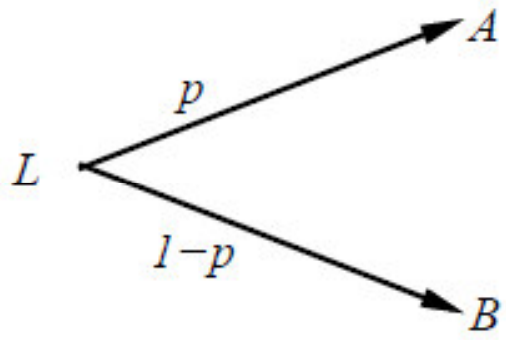
Learning rate for model predictions/properties



Russel&Norvig: Artificial intelligence

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

Utilities

Utilities map states to real numbers. Which numbers?

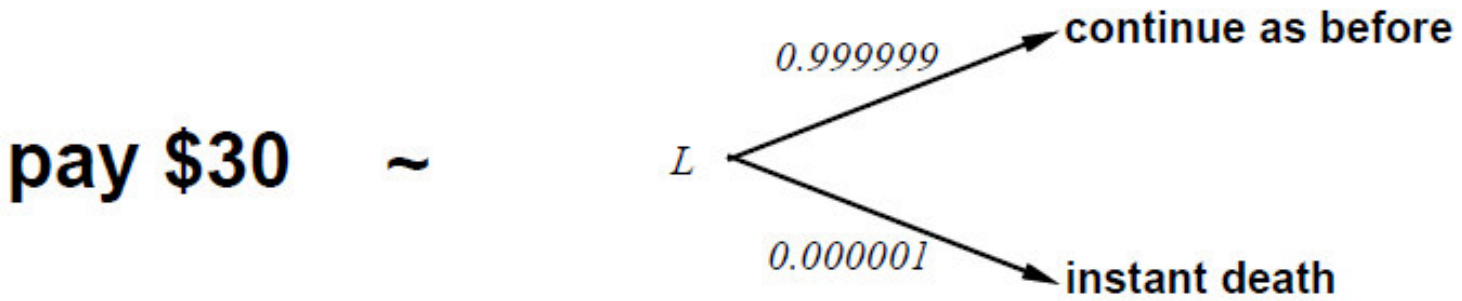
Standard approach to assessment of human utilities:

compare a given state A to a standard lottery L_p that has

"best possible prize" $u_\top$ with probability p

"worst possible catastrophe" $u_\bot$ with probability $(1-p)$

adjust lottery probability p until $A \sim L_p$



Russel&Norvig: Artificial intelligence, ch.16

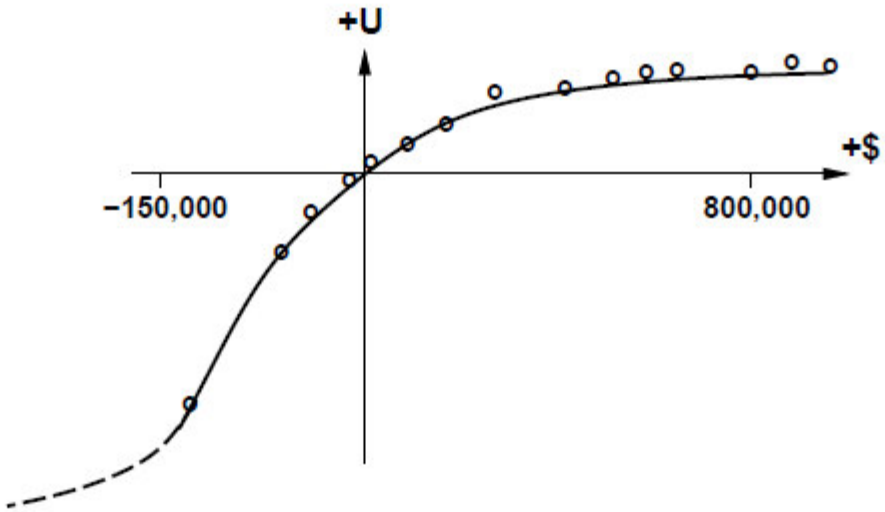
Utility of money

Money does **not** behave as a utility function

Given a lottery L with expected monetary value $EMV(L)$,
usually $U(L) < U(EMV(L))$, i.e., people are risk-averse

Utility curve: for what probability p am I indifferent between a prize x and a lottery $[p, \$M;\ (1-p), \$0]$ for large M?

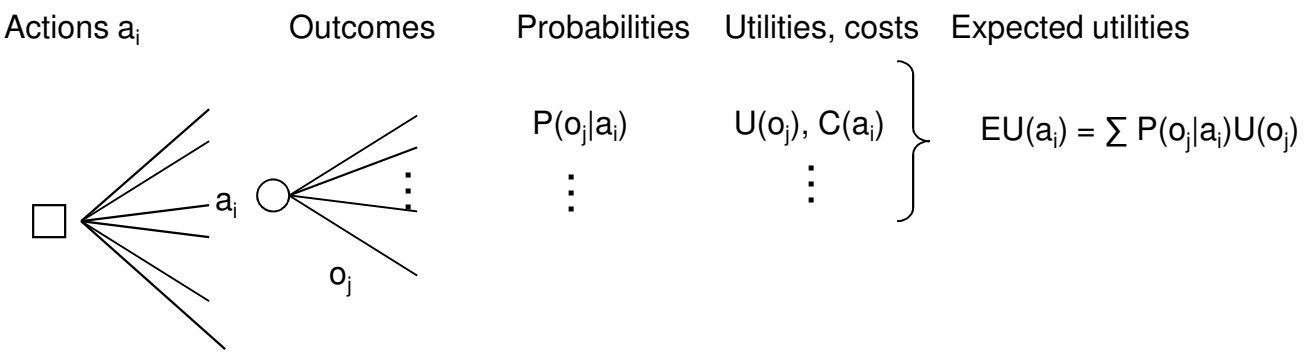
Typical empirical data, extrapolated with risk-prone behavior:



Russel&Norvig: Artificial intelligence, ch.16

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 $a^* = \arg\max_i EU(a_i)$



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”)	

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**

Universal theory of induction I.

- Universal machines: Turing, BSS, Kolmogorov
- R.Solomonoff: universal distribution?
  - how can one assign a probability to a hypothesis h *before* observing any data?
- Epicurus' (342? B.C. - 270 B.C.) principle of multiple explanations which states that one should *keep all hypotheses that are consistent with the data*.
- The principle of Occam's razor (1285 - 1349, sometimes spelt Ockham). Occam's razor states that when inferring causes *entities should not be multiplied beyond necessity*. This is widely understood to mean: Among all hypotheses consistent with the observations, choose the simplest. In terms of a prior distribution over hypotheses, this is the same as giving simpler hypotheses higher a priori probability, and more complex ones lower probability.

Universal theory of induction II.

- Universal distributions

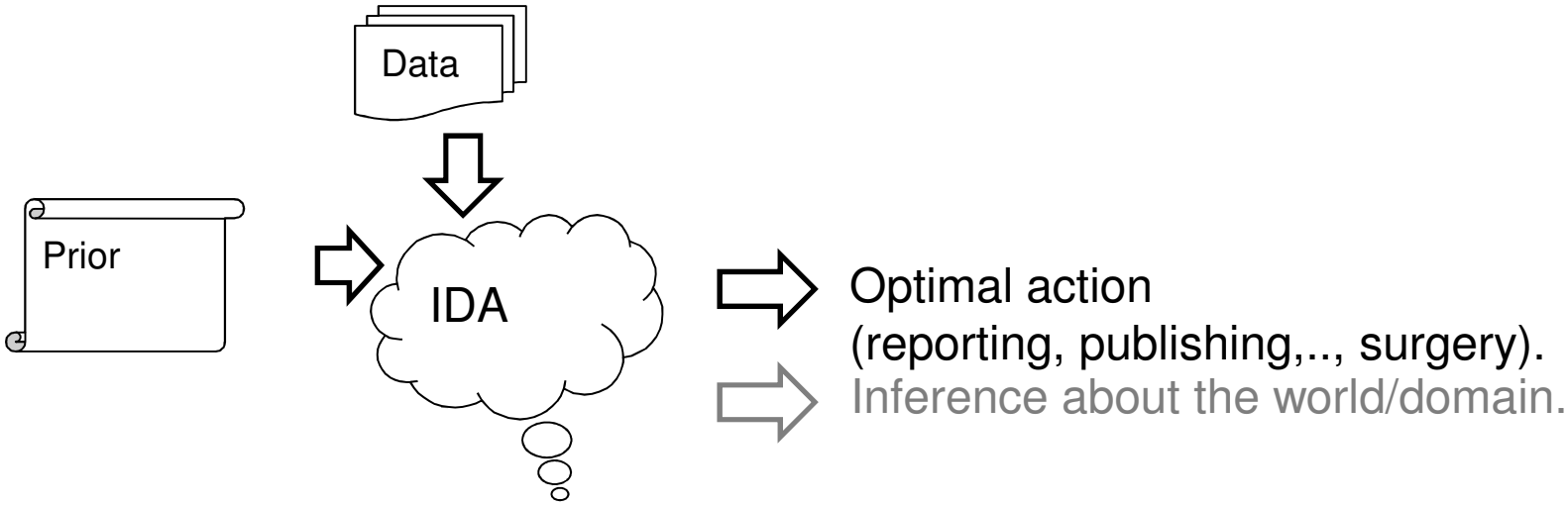
$$m(x) := \sum_{p\,:\,U(p)=x} 2^{-\ell(p)}, \qquad -\log m(x) = K(x) + O(1).$$

$$M(x) := \sum_{p\,:\,U(p)=x*} 2^{-\ell(p)}, \qquad -\log M(x) = K(x) - O(\log \ell(x))$$

If the infinite binary sequences are distributed according to a computable measure μ , then the predictive distribution
M(x n+1 |x 1 ...x n):=M(x 1 ...x n+1)/M(x 1 ...x n) converges rapidly to
μ(x n+1 |x 1 ...x n):=μ(x 1 ...x n+1)/μ(x 1 ...x n) with μ -probability 1.
Hence, M predicts almost as well as does the true distribution μ .

M. Li and P. M B. Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer, New York, 2nd edition 1997, and 3rd edition 2008

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)

Types of 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

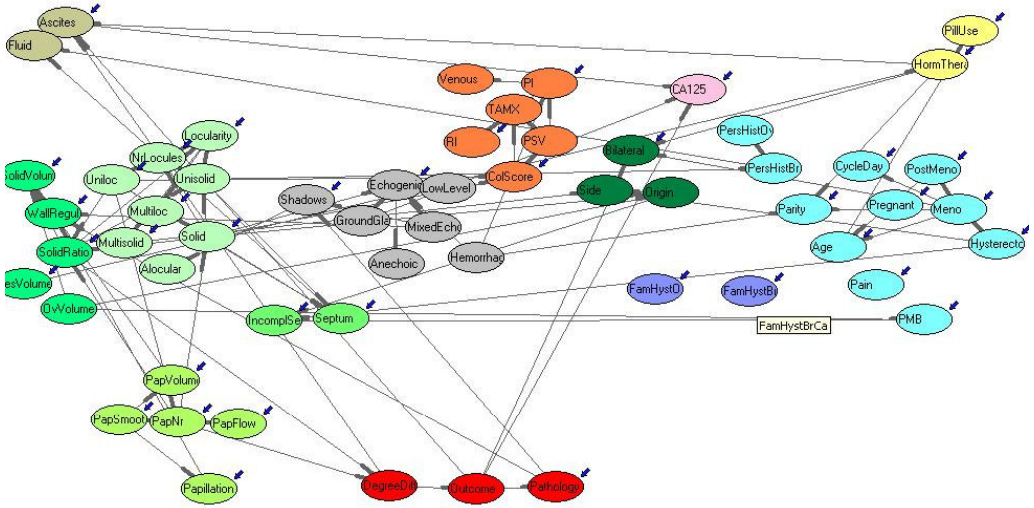
Types of inference

- (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

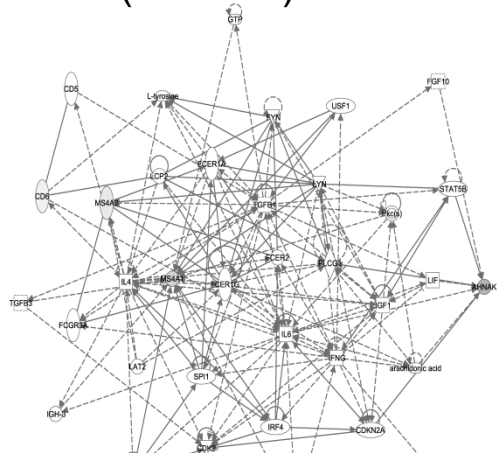
Association graphs, (in)dependence maps, causal networks, control systems

Clusters, modules

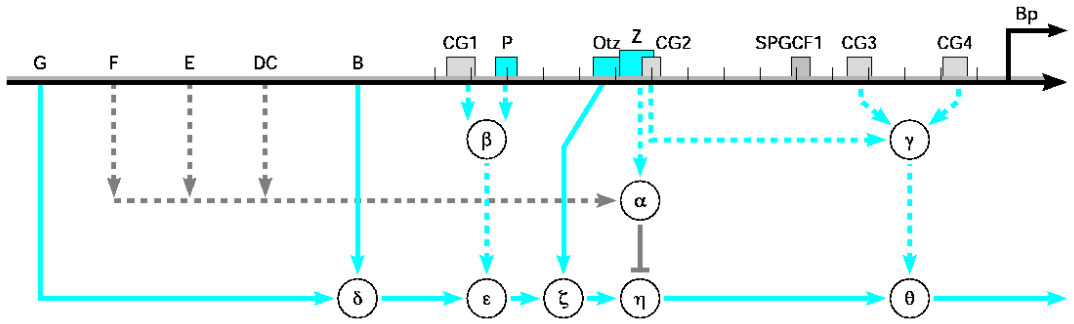


Conditional independencies

(Causal) Mechanisms



Parameters



Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  -
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

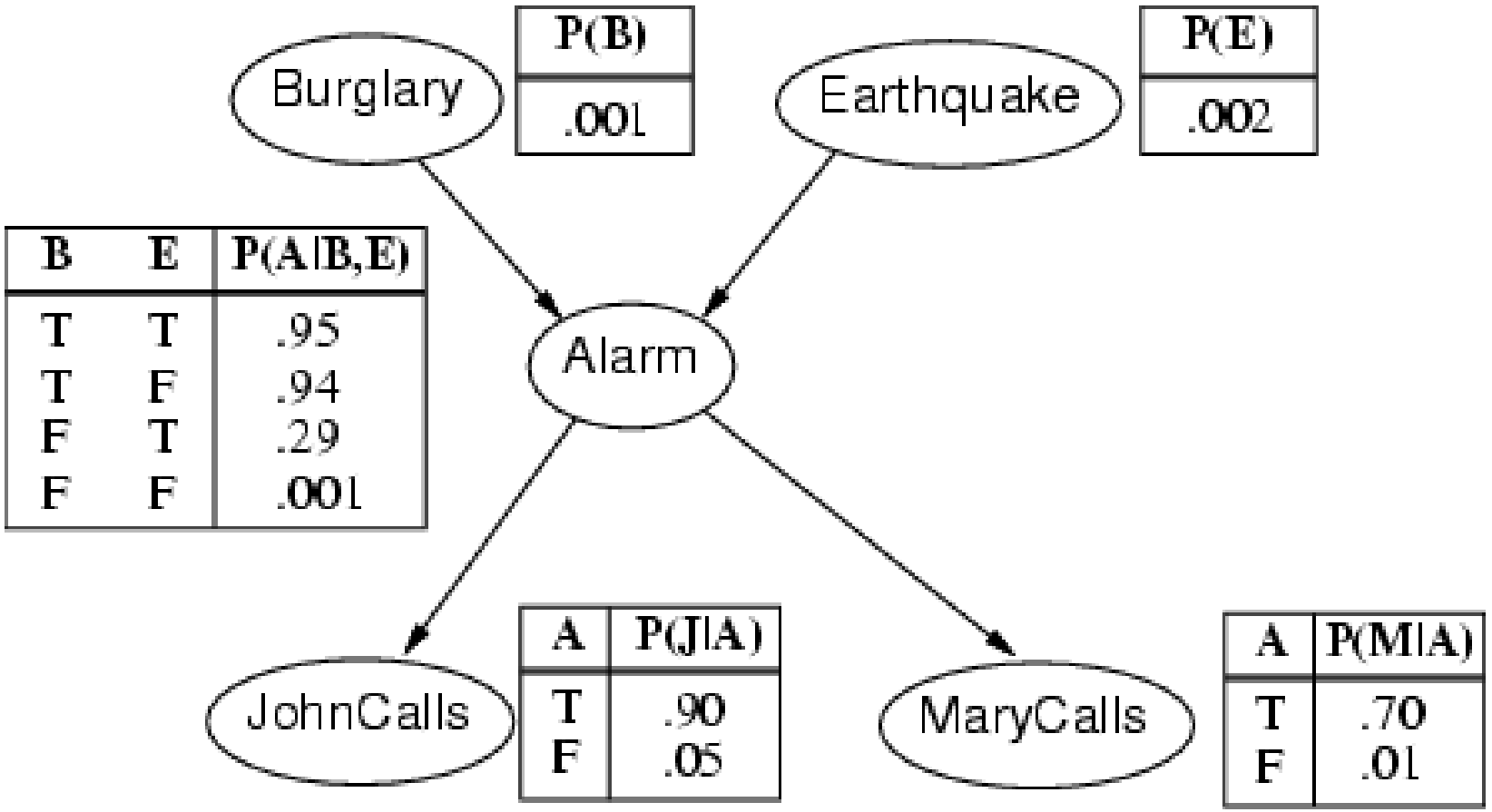
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

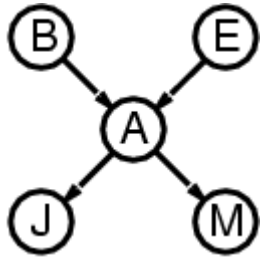
- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

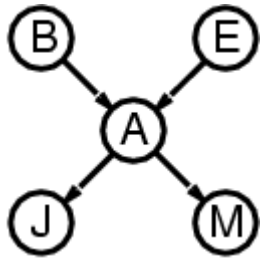


- Each row requires one number p for X_i = *true* (the number for X_i = *false* is just *1-p*)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. $2^5-1 = 31$)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:

$$\boldsymbol{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \boldsymbol{P}(X_i \mid Parents(X_i))$$



e.g., $\boldsymbol{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$= \boldsymbol{P}(j \mid a)\, \boldsymbol{P}(m \mid a)\, \boldsymbol{P}(a \mid \neg b, \neg e)\, \boldsymbol{P}(\neg b)\, \boldsymbol{P}(\neg e)$

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\boldsymbol{P}(X_i \mid Parents(X_i)) = \boldsymbol{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

$$\boldsymbol{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \boldsymbol{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad \text{//(chain rule)}$$

$$= \prod_{i=1} \boldsymbol{P}(X_i \mid Parents(X_i)) \quad \text{//(by construction)}$$

Inference by enumeration

Every question about a domain can be answered by the joint distribution.

Typically, we are interested in the posterior joint distribution of the query variables **Y** given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** - **Y** – **E**

Then the required summation of joint entries is done by summing out the hidden variables:

$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \alpha\Sigma_h\mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h})$

- The terms in the summation are joint entries because **Y**, **E** and **H** together exhaust the set of random variables
- Obvious problems:
  1. Worst-case time complexity $O(d^n)$ where d is the largest arity
  2. Space complexity $O(d^n)$ to store the joint distribution
  3. How to find the numbers for $O(d^n)$ entries?

Conditional independence

„Probability theory=measure theory+independence”

$I_P(X;Y|Z)$ or $(X \perp\!\!\!\perp Y|Z)_P$ denotes that X is independent of Y given Z:
$P(X;Y|z)=P(Y|z)\ P(X|z)$ for all z with $P(z)>0$.

(Almost) alternatively, $I_P(X;Y|Z)$ iff

$P(X|Z,Y)= P(X|Z)$ for all z,y with $P(z,y)>0$.

Other notations: $D_P(X;Y|Z)$ =def= $\neg I_P(X;Y|Z)$

Contextual independence: for not all z.

Naive Bayesian network (NBN)

Decomposition of the joint:

$P(Y,X_1,..,X_n) = P(Y)\prod_i P(X_i,|Y, X_1,..,X_{i-1})$ //by the chain rule

$= P(Y)\prod_i P(X_i,|Y)$ // by the N-BN assumption

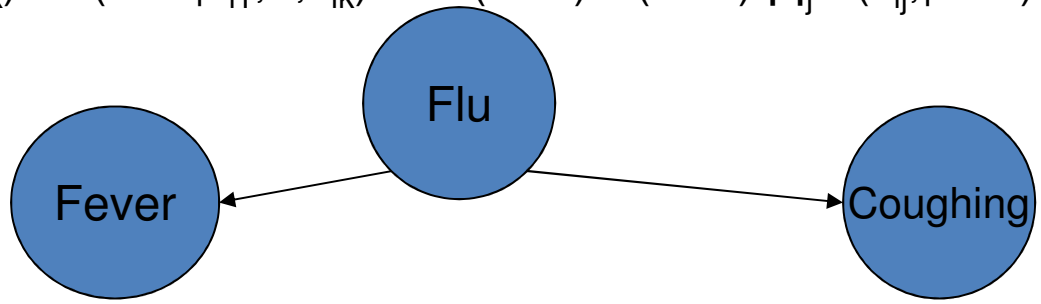
2n+1 parameteres!

Diagnostic inference:

$P(Y|x_{i1},..,x_{ik}) = P(Y)\prod_j P(x_{ij},|Y) / P(x_{i1},..,x_{ik})$

If Y is binary, then the odds

$P(Y=1|x_{i1},..,x_{ik}) / P(Y=0|x_{i1},..,x_{ik}) = P(Y=1)/P(Y=0) \prod_j P(x_{ij},|Y=1) / P(x_{ij},|Y=0)$



$$p(Flu = present \mid Fever = absent, Coughing = present)$$

$$\propto p(Flu = present)\, p(Fever = absent \mid Flu = present)\, p(Coughing = present \mid Flu = present)$$

Complexity of exact inference

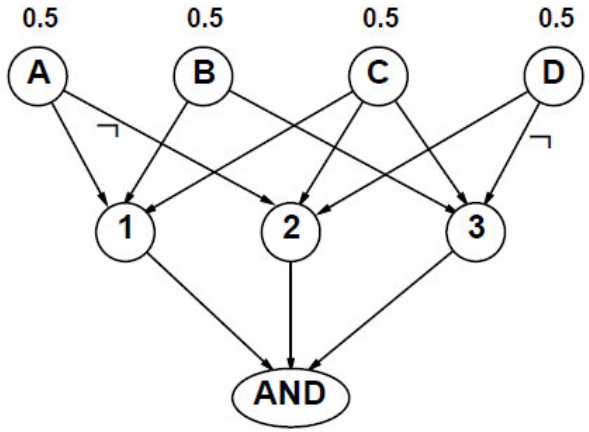
Singly connected networks (or polytrees):

- any two nodes are connected by at most one (undirected) path
- time and space cost of exact inference $O(d^k n)$

Multiply connected networks:

- can reduce 3SAT to exact inference: 0<p(AND)? $\Rightarrow$ NP-hard
- equivalent to **counting** 3SAT models $\Rightarrow$ #P-complete

1. A v B v C
2. C v D v ¬A
3. B v C v ¬D



Noisy-OR

Noisy-OR distributions model multiple noninteracting causes

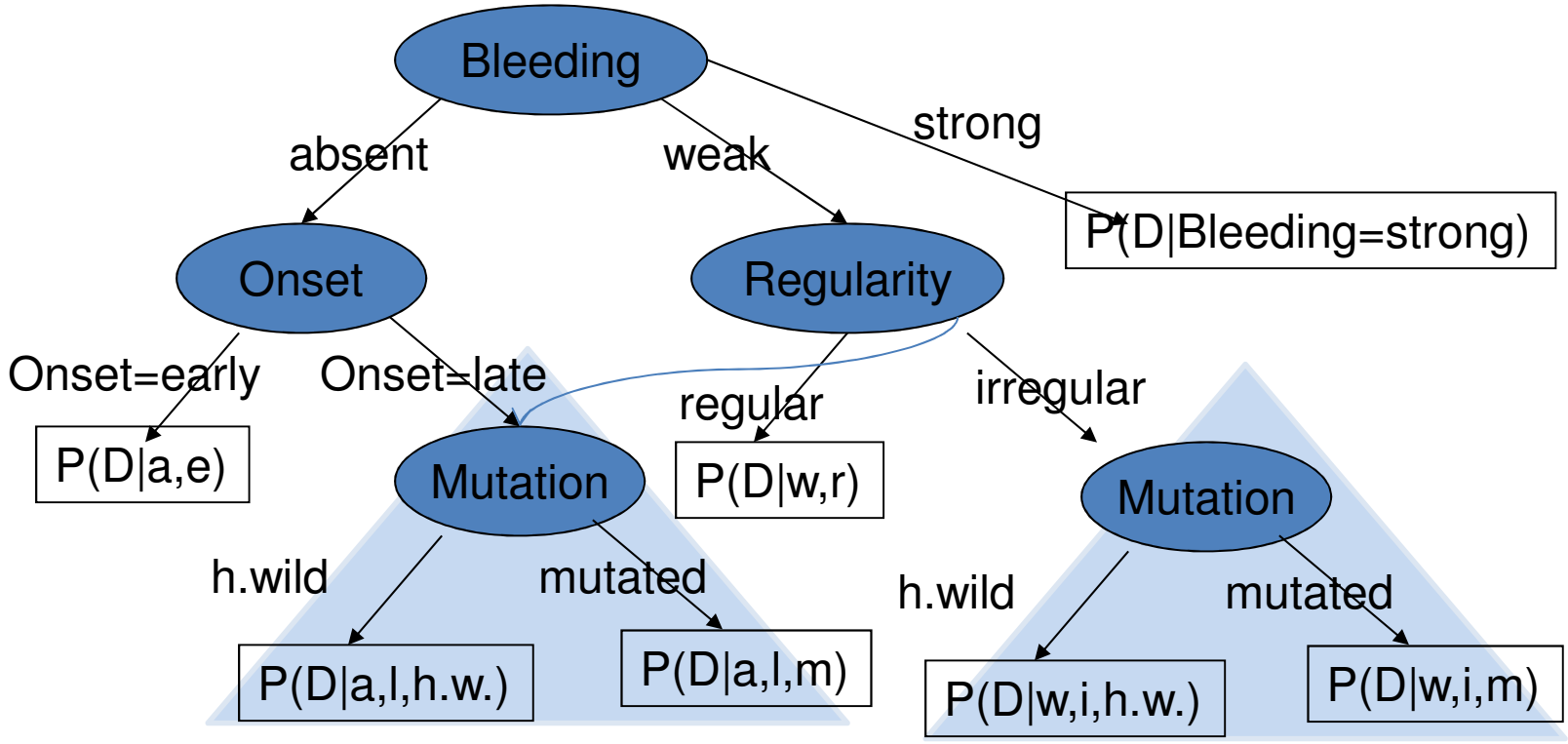
1) Parents $U_1 \ldots U_k$ include all causes (can add leak node)
2) Independent failure probability q_i for each cause alone

$$\Rightarrow P(X|U_1 \ldots U_j, \neg U_{j+1} \ldots \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

$Cold$	Flu	$Malaria$	$P(Fever)$	$P(\neg Fever)$
F	F	F	**0.0**	1.0
F	F	T	0.9	**0.1**
F	T	F	0.8	**0.2**
F	T	T	0.98	$0.02 = 0.2 \times 0.1$
T	F	F	0.4	**0.6**
T	F	T	0.94	$0.06 = 0.6 \times 0.1$
T	T	F	0.88	$0.12 = 0.6 \times 0.2$
T	T	T	0.988	$0.012 = 0.6 \times 0.2 \times 0.1$

Number of parameters **linear** in number of parents

Decision trees, decision graphs

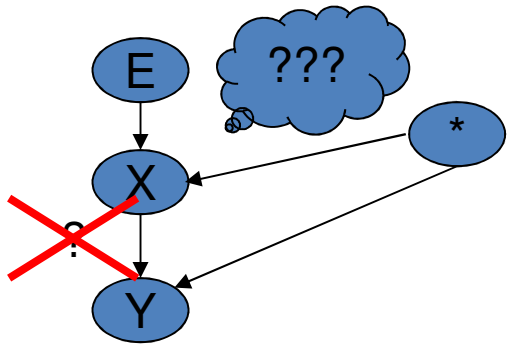


Decision tree: Each internal node represent a (univariate) test, the leafs contains the conditional probabilities given the values along the path.
Decision graph: If conditions are equivalent, then subtrees can be merged.
E.g. If (Bleeding=absent,Onset=late) ~ (Bleeding=weak,Regularity=irreg)

Types of data and inference, cont'd

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

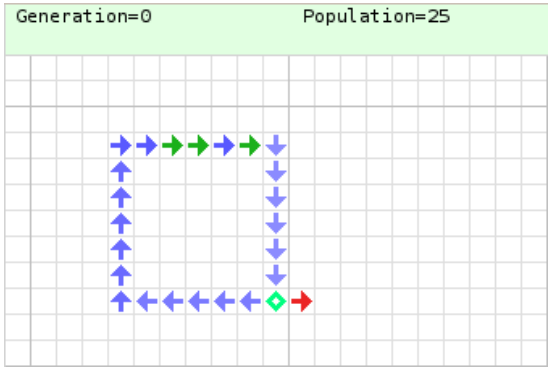
- Automated, tabula rasa causal inference from (passive) observation is possible, i.e. hidden, confounding variables can be excluded

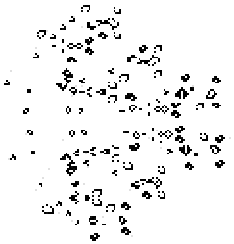


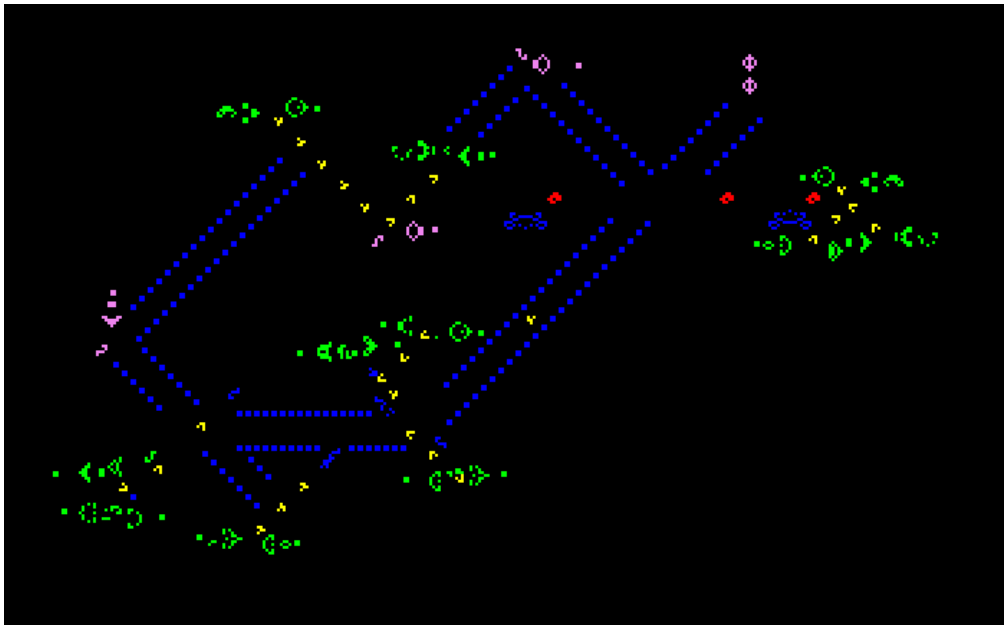
„Plato's two surprises:
1. Not all true theorems can be proved
2. Causal inference is possible from observations"

Causality, emergence, virtual/in silico data

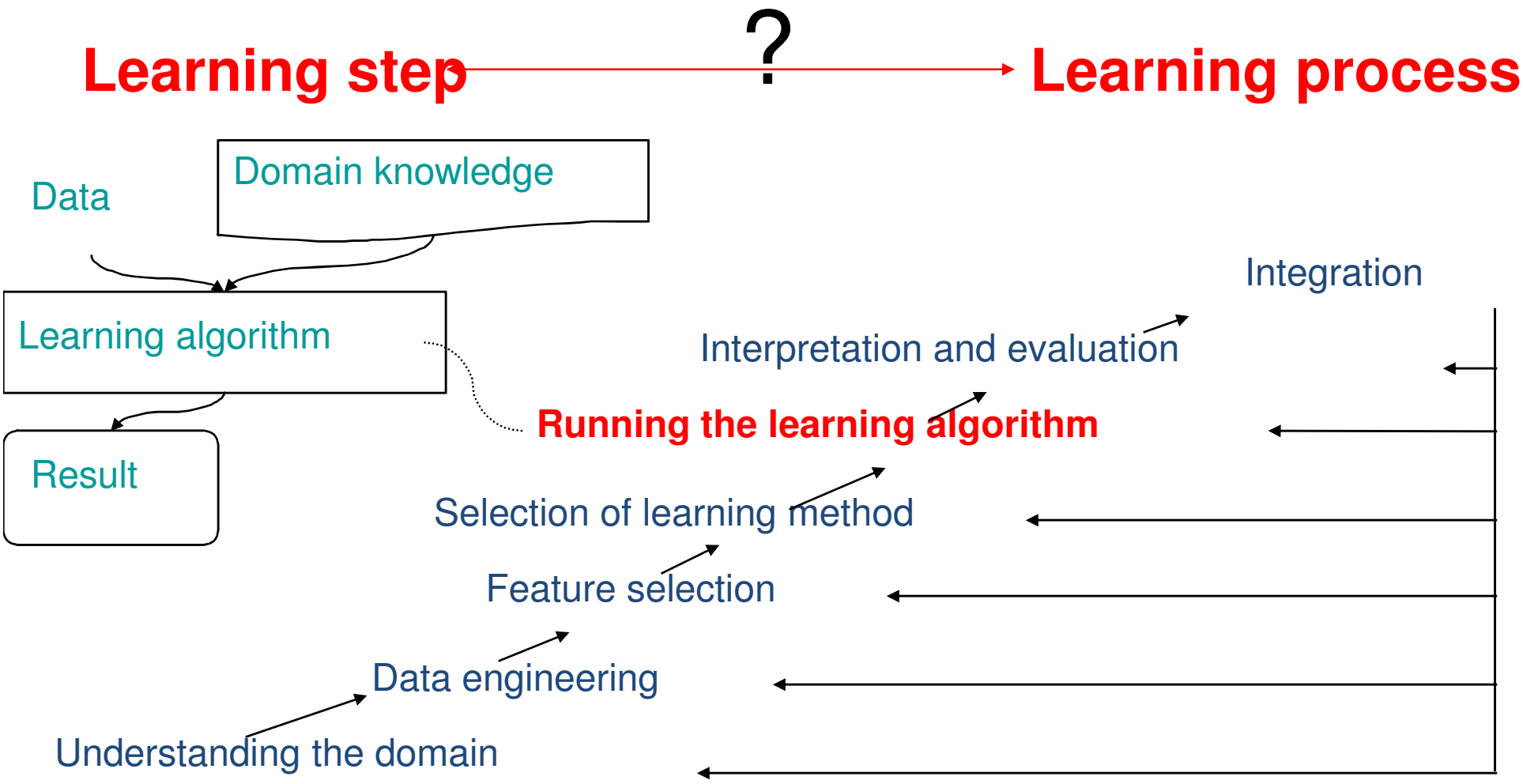
- Von Neumann, J. and A. W. Burks (1966): Theory of self-reproducing automata
- Whole cell/body simulations(!)





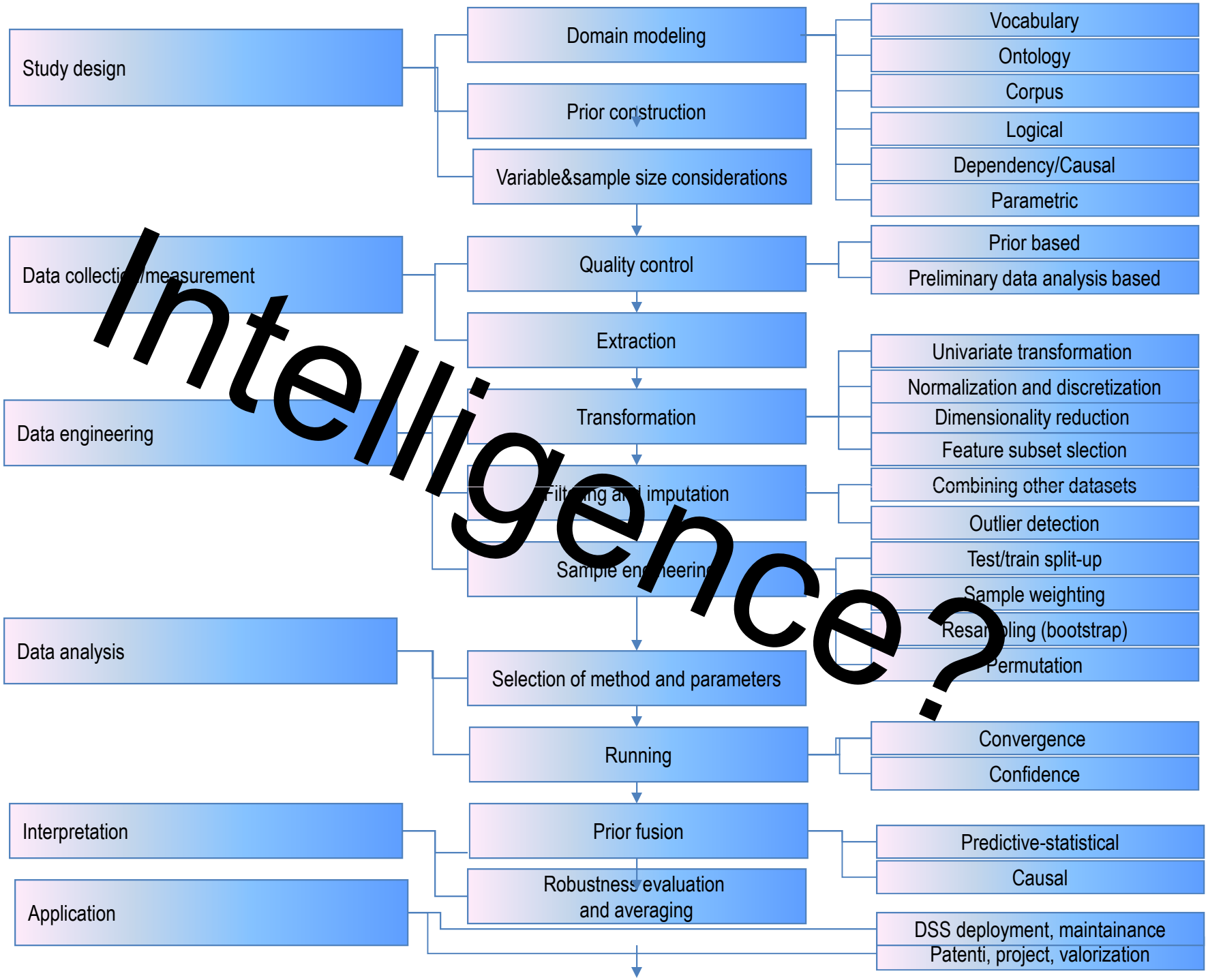


Learning step or learning **process?**

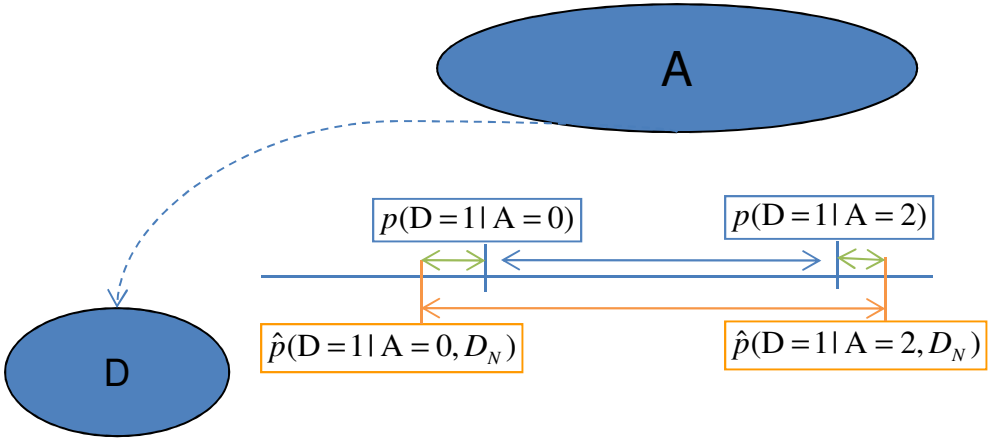


Data analysis in practice

- 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?

Fundamental questions in statistics



Real difference

Estimation errors

Estimated difference

Relative frequencies: $\hat{p}(\mathrm{D}=1|\mathrm{A}=0,D_N)=\hat{p}(\mathrm{D}=1,\mathrm{A}=0,D_N)/\hat{p}(\mathrm{A}=0,D_N)=N_{D=1,A=0}/N_{A=0}$

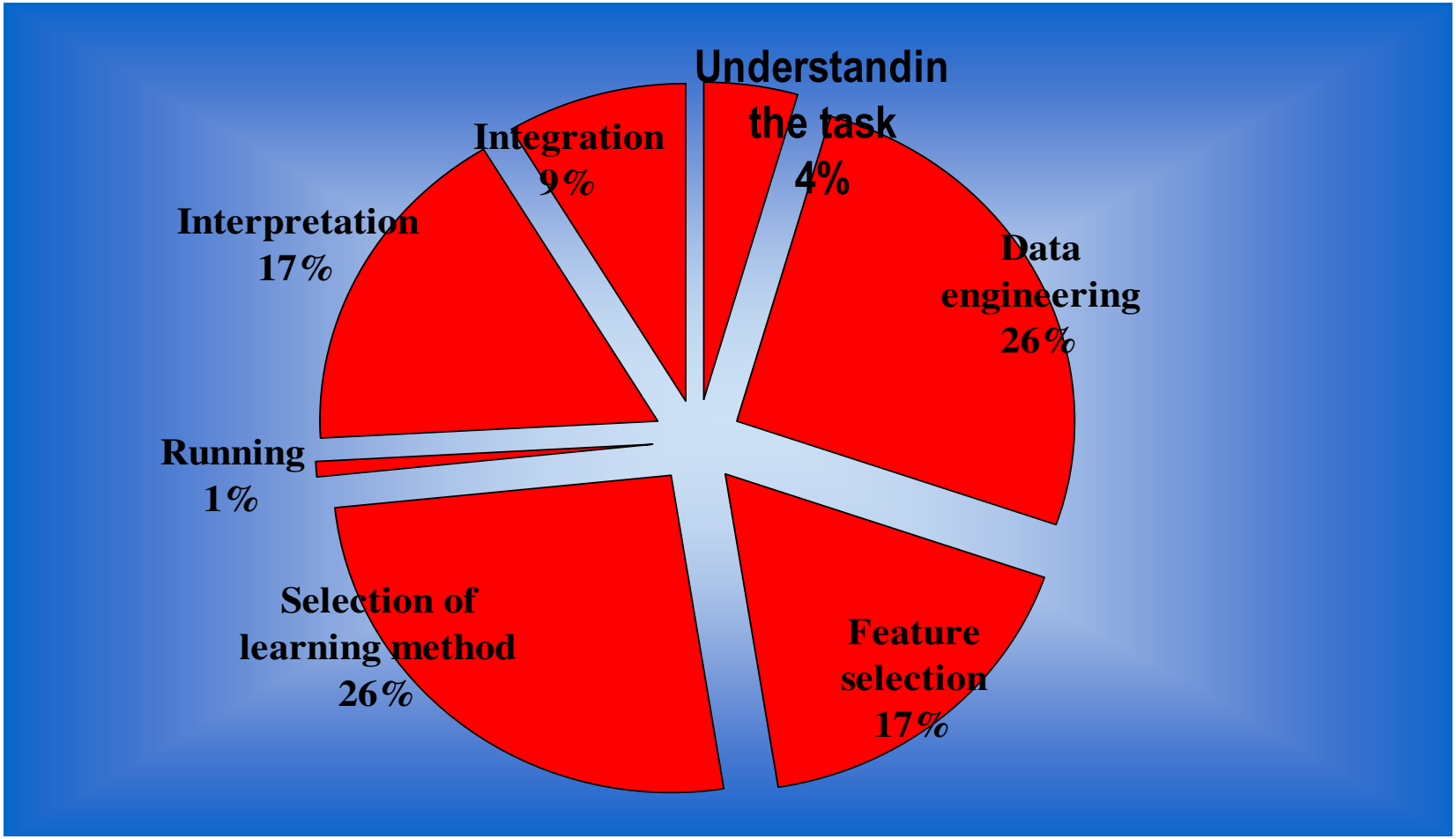
Law of large numbers: $p(\mathrm{D}=1|\mathrm{A}=0)\approx N_{D=1,A=0}/N_{A=0}$

Estimation error because of finite data D_N: $\hat{p}(\mathrm{D}=1|\mathrm{A}=0,D_N)$ - $p(\mathrm{D}=1|\mathrm{A}=0)$

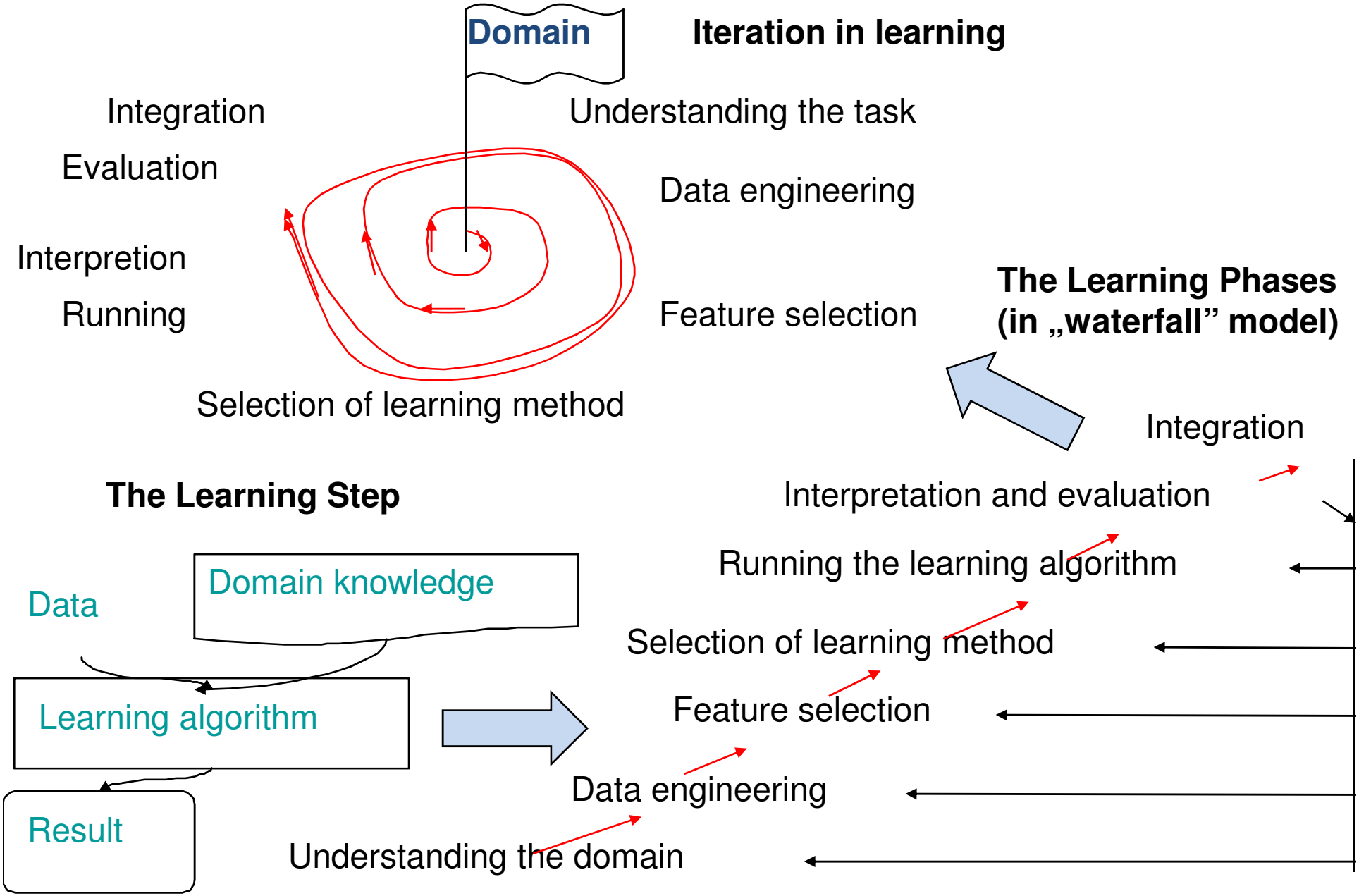
Central limit th. (asymptotic), Inequalities for finite(!) data (ε accuracy,δ confidence)

sample complexity: $N_{\varepsilon,\delta}$ $p(D_{N_{\varepsilon,\delta}}:\varepsilon<|\hat{p}(\mathrm{D}=1|\mathrm{A},D_{N\varepsilon,\delta})-p(\mathrm{D}=1|\mathrm{A})|)<\delta$

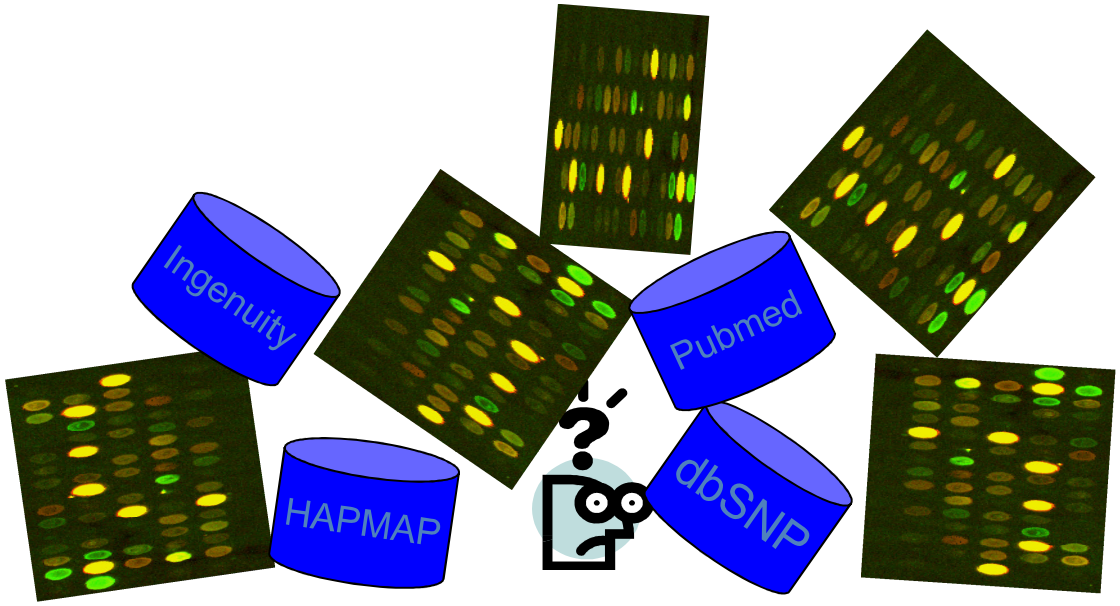
Length of phases



Forwards vs **iterative process**?



The interpretational bottleneck (~the fusion challenge)

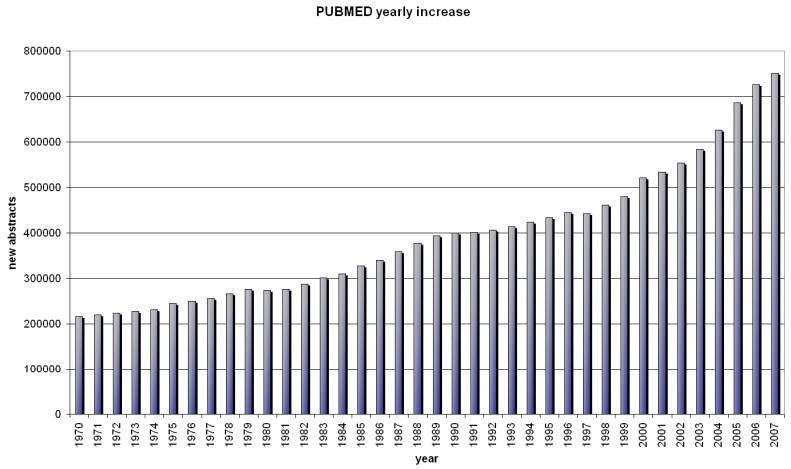


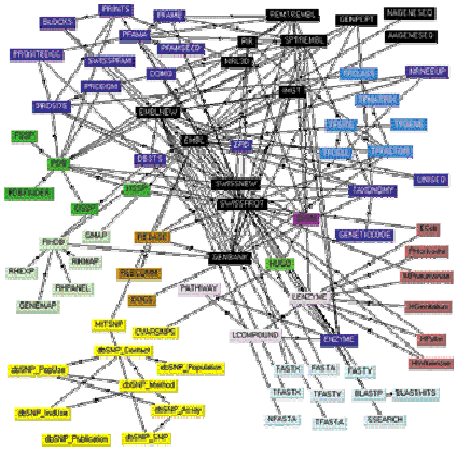
- Free text repositories (with significances): e.g. SNPedia
- Manually curated logical knowledge bases: Ingenuity

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
  - …

Semantic publishing: papers vs DBs/KBs





M. Gerstein, "E-publishing on the Web: Promises, pitfalls, and payoffs for bioinformatics," *Bioinformatics,* 1999

M. Gerstein: **Blurring the boundaries between scientific 'papers' and biological databases**, Nature, 2001

P. Bourne, "Will a biological database be different from a biological journal?," *Plos Computational Biology,* 2005

M. Gerstein et al: "Structured digital abstract makes text mining easy," *Nature,* 2007.

M. Seringhaus et al: "Publishing perishing? Towards tomorrow's information architecture," *Bmc Bioinformatics,* 2007.

M. Seringhaus: "Manually structured digital abstracts: A scaffold for automatic text mining," *Febs Letters,* 2008.

D. Shotton: "Semantic publishing: the coming revolution in scientific journal publishing," *Learned Publishing,* 2009

Databases, papers, knowledge bases

- An example system: Gene Expression Omnibus (integration of DB-KB-paper)
- Genetic association studies
  - Standardization of reporting,

H. Colhoun: "Problems of reporting genetic associations with complex outcomes," Lancet, 2003.
J. Attia et al: "How to Use an Article About Genetic Association A: Background Concepts," , B: Are the Results of the Study Valid?,"
C: What Are the Results and Will They Help Me in Caring for My Patients?," Jama, 2009.
J. Little et al: "STrengthening the REporting of Genetic Association studies (STREGA) " European Journal of Clinical Investigation, 2009.
J. Huang et al: "Minimum Information about a Genotyping Experiment (MIGEN)," Standards in Genomic Sciences, 2011
A. Janssens et al "Strengthening the reporting of Genetic Risk Prediction Studies: The GRIPS statement," Genetics in Medicine, 2011

Databases, papers, knowledge bases

- An example system: Gene Expression Omnibus (integration of DB-KB-paper)
- Genetic association studies
  - Standardization of reporting,
  - text-mining,

Leitner F, Valencia A (2008). A text-mining perspective on the requirements for electronically annotated abstracts. *Febs Letters* **582**(8): 1178-1181.
Lu Z, Hirschman L (2012). Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database* **Vol. 2012**(Article ID bas043).
Biograph, BioRat

Databases, papers, knowledge bases

- An example system: Gene Expression Omnibus (integration of DB-KB-paper)
- Genetic association studies
  - Standardization of reporting,
  - text-mining,
  - commercial omic KBs, inference (IBM's Watson)

HuGe, Ingenuity Pathway Analysis, Knome, Alamute, HuGe (W. Yu, et al: "A navigator for human genome epidemiology," *Nature Genetics,* 2008)

"The Science Behind an Answer"

Databases, papers, knowledge bases

- An example system: Gene Expression Omnibus (integration of DB-KB-paper)
- Genetic association studies
  - Standardization of reporting,
  - text-mining,
  - commercial KBs,
  - evidence-based medicine,

    K. Goddard *et al.*, "Building the evidence base for decision making in cancer genomic medicine using comparative effectiveness research," *Genetics in Medicine,* 2012

    M. Gwinn, *et al*, "Horizon scanning for new genomic tests," *Genetics in Medicine* 2011

Databases, papers, knowledge bases

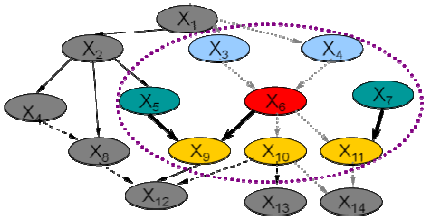
- An example system: Gene Expression Omnibus (integration of DB-KB-paper)
- Genetic association studies
  - Standardization of reporting,
  - text-mining,
  - commercial KBs, inference (IBM's Watson)
  - evidence-based medicine,
  - meta-analysis („gold dust")

G. Shi et al: "Mining Gold Dust Under the Genome Wide Significance Level: A Two-Stage Approach to Analysis of GWAS," Genetic Epidemiology, 2011
E. Evangelou et al: "Meta-analysis methods for genome-wide association studies and beyond," Nature Reviews Genetics, 2013

Systems-based, Bayesian biomarker discovery

„all models are wrong, but some are useful"

➔**e.g.** there are stable, interesting properties



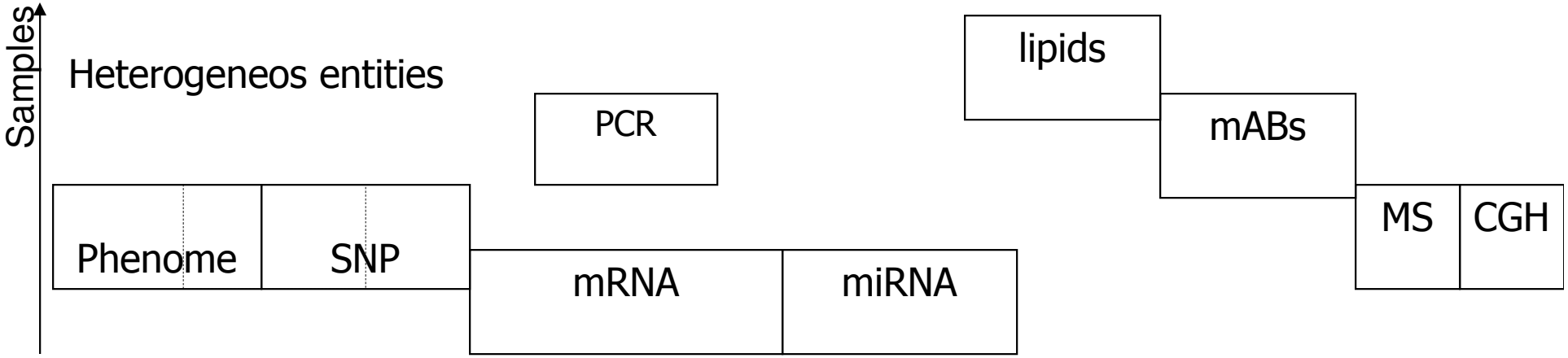
Bayes' rule:

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

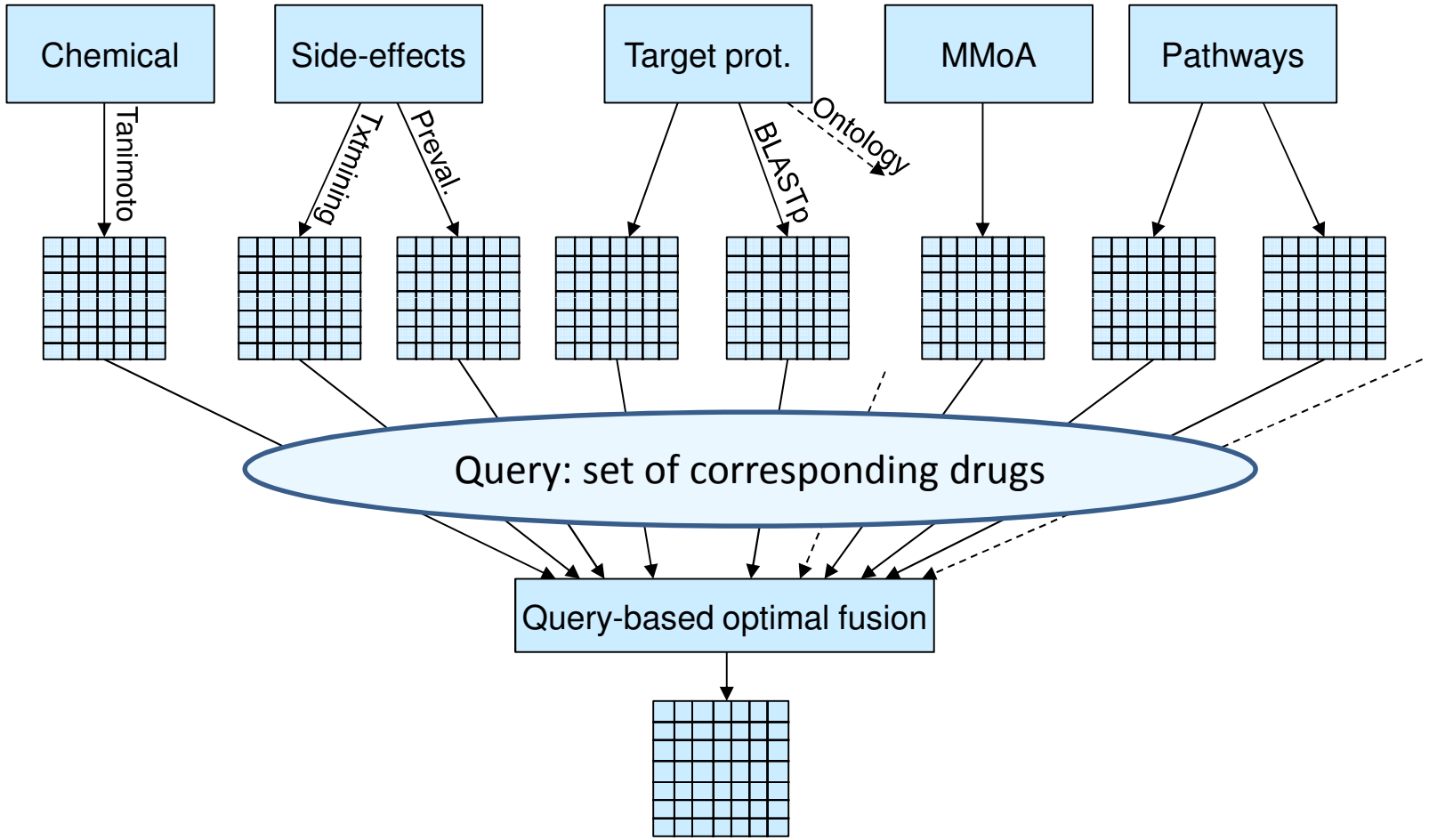
A priori distribution (prior), p(Model):

- *is a technical tool for inversion (to achieve a direct probabilistic statement),*
- *provides a principled solution to incorporate prior knowledge.*

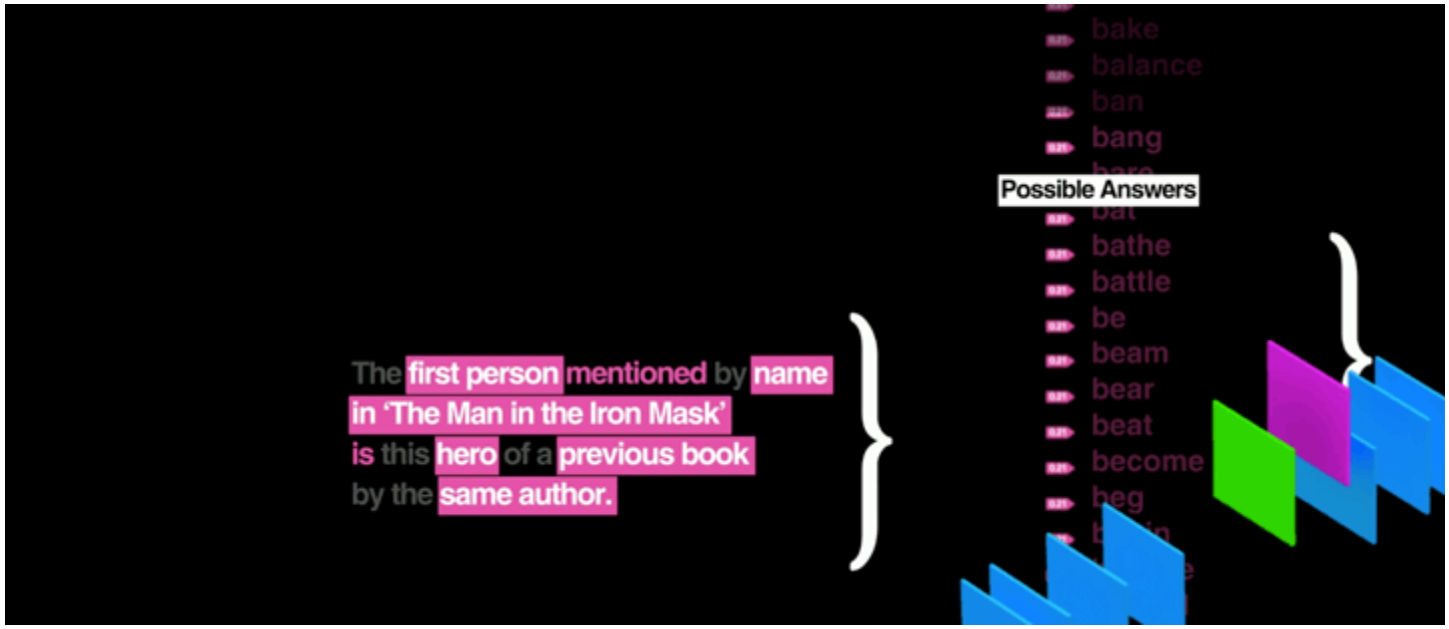
Open problem: derivation and formalization of informative (domain/data specific) priors.



Similarity-based fusion in drug repositioning



Watson?
Automated statistician??



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

IBM WATSON

Lectures

1. Introduction
    1. Data-rich/intensive science, big data
    2. Bayesian decision theory as foundation
    3. Causal inference
    4. Open access data, publication, model, computation
2. BSc recall: Bayesian inference
3. Nov 12 (kedd): TDK
4. BSc recall: Bayesian networks
5. Active learning, adaptive study design
6. Nov 22 (péntek): középiskolai
7. AA? Hidden Markov Models: inference, parameter learning (EM)
8. Learning graphical models: Markov networks + Bayesian networks
9. Monte Carlo methods: application for graphical models
10. Assocation and relevance analysis, the featre subset selection problem
11. ?Causal inference? HMM
12. Incomplete data