Intelligent data analysis
From coherent reasoning
to universal theory of induction

Peter Antal antal@mit.bme.hu

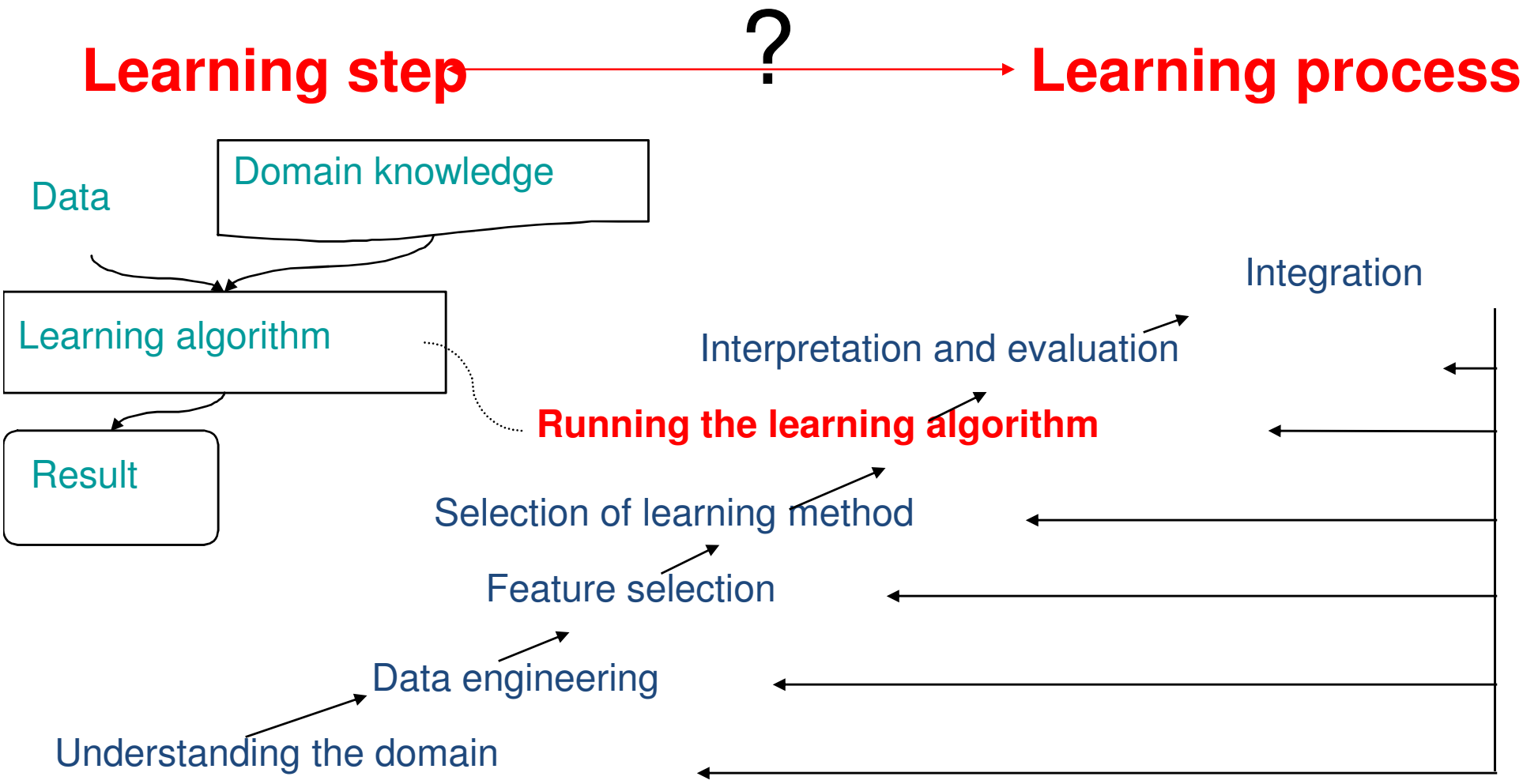
Overview

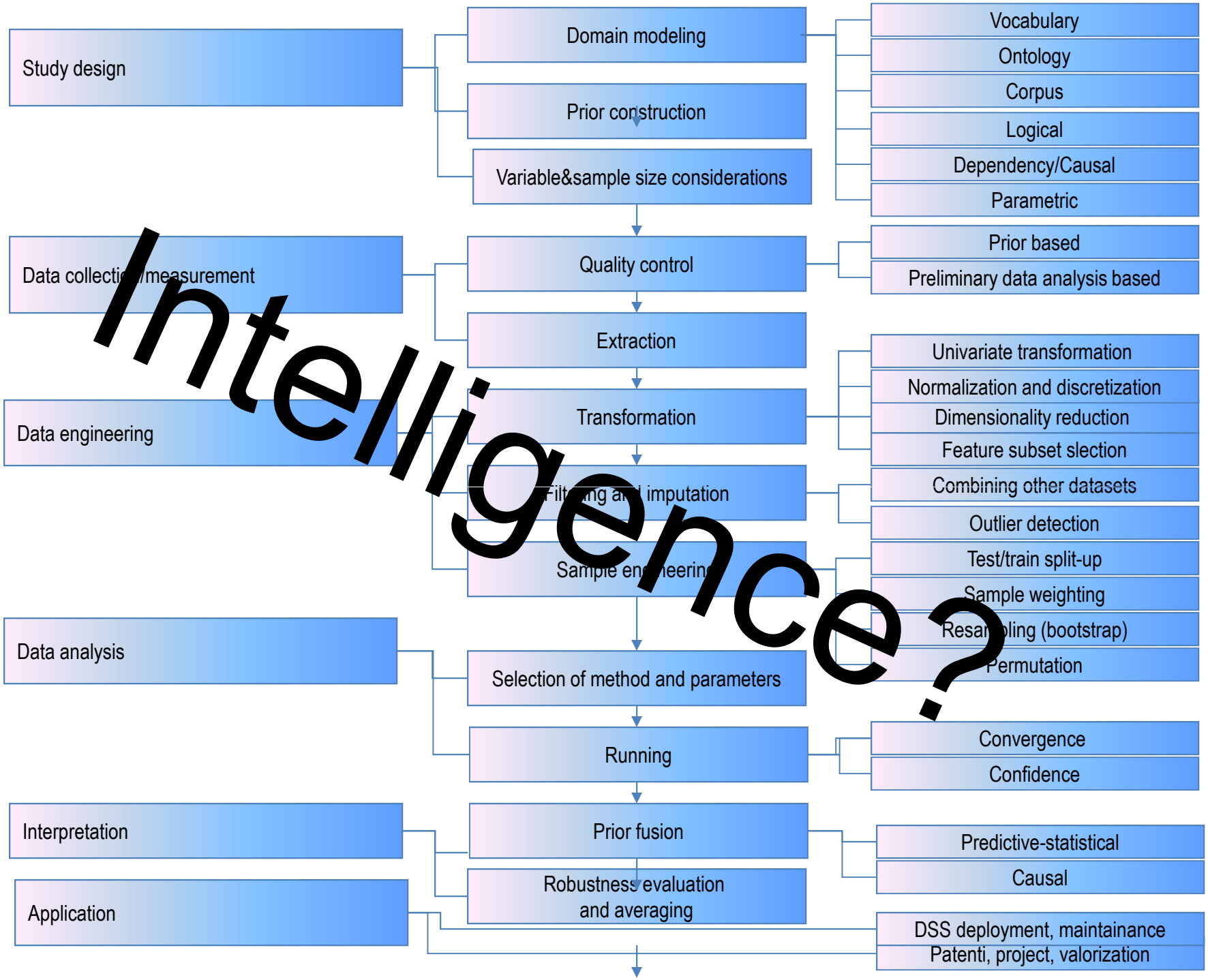
- Dimensions of intelligence in IDA
  - The real process…
  - The real questions, types of models…
  - Prior knowledge
- New/old ideas
  - Just visualization of data using priors af filters...
  - Just action: Ambient assisted data analysis
  - Persistent data analysis: the lifecycle of data (oneshot, meta)
- IDA: understanding or action
- Understanding your data: Cooper
- Optimal action using your data: ambient assisted data analysis
- Assume: passive observation+prediction context
- Improve induction➔why does it work? Does it work?
- Normativity of probability theory, utility theory.
- Universal theory of induction
- Prequential theory

Intelligent data analysis

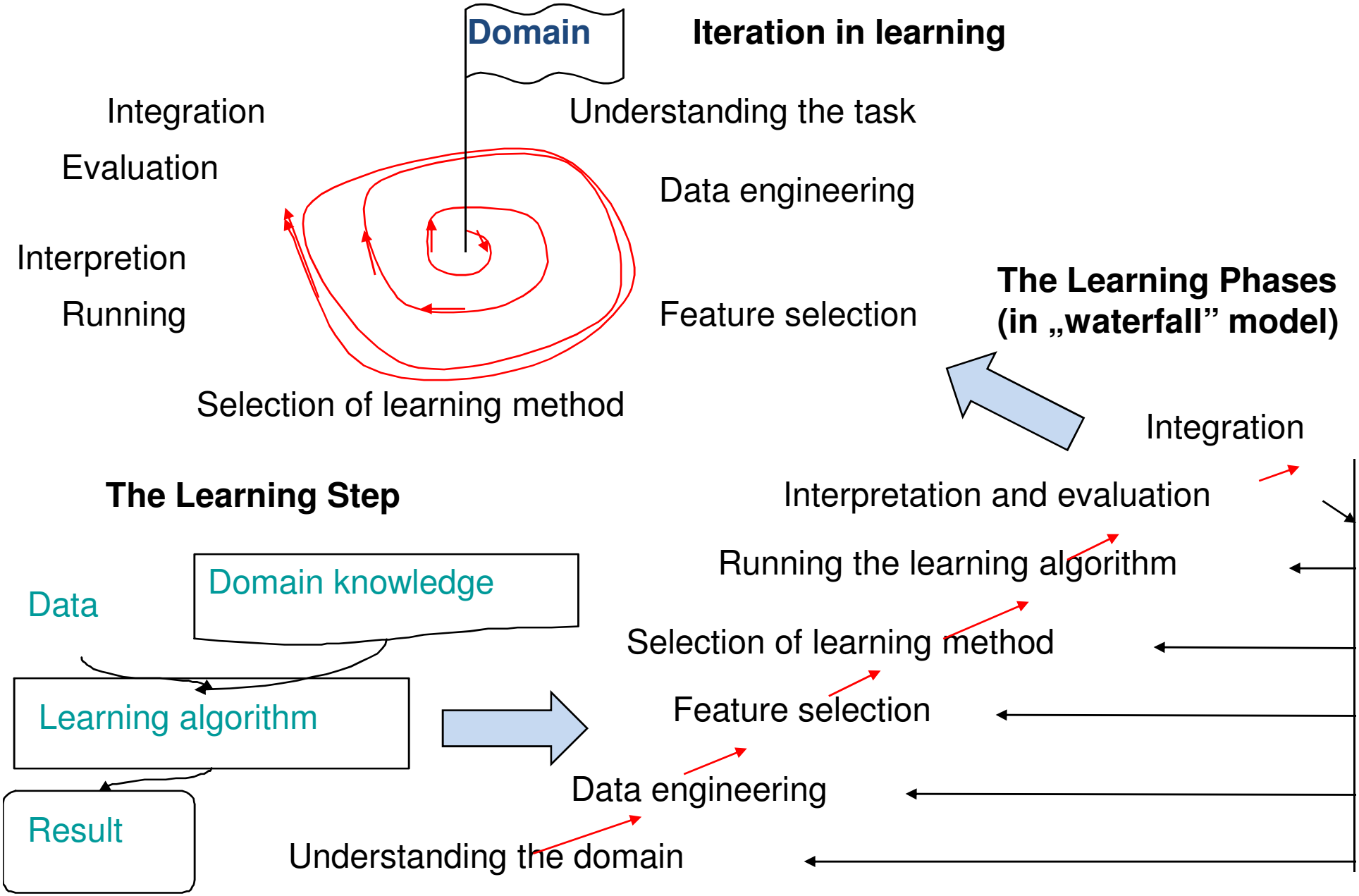
- Data analysis is a process (industrial process)
- Data analysis can be “deep”:
  - Predictive, causal, counterfactual
- Background knowledge
  - Wide: Common sense
  - Deep: domain knowledge
- Data analysis can be action oriented/ “ambient”
- Data analysis can be persistent.

Learning step or learning **process?**




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?

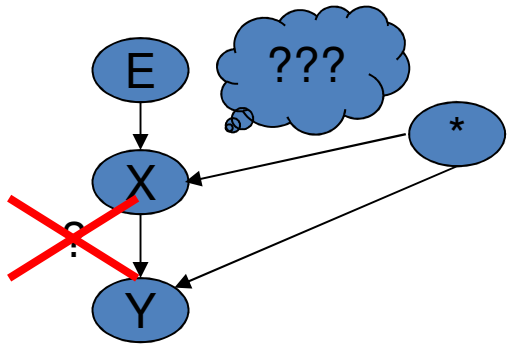
Forwards vs **iterative process**?



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"

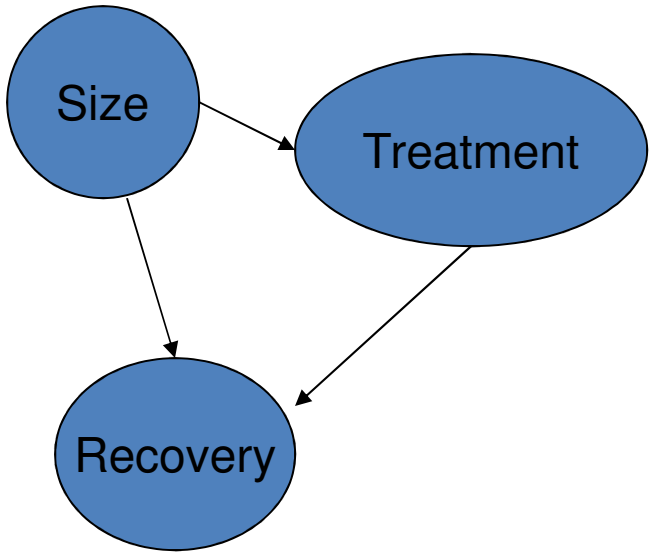
Intelligent (inductive) inference (Intelligent data analysis)

Think like a human (data analyst?, child?)	**Think rationally: understand your data**
Act like a human.	Act rationally.

Simpson's paradox

Kidney stone

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**



$P(R|S,T) \; != \sum_S P(R|S,T) \; P(S)$

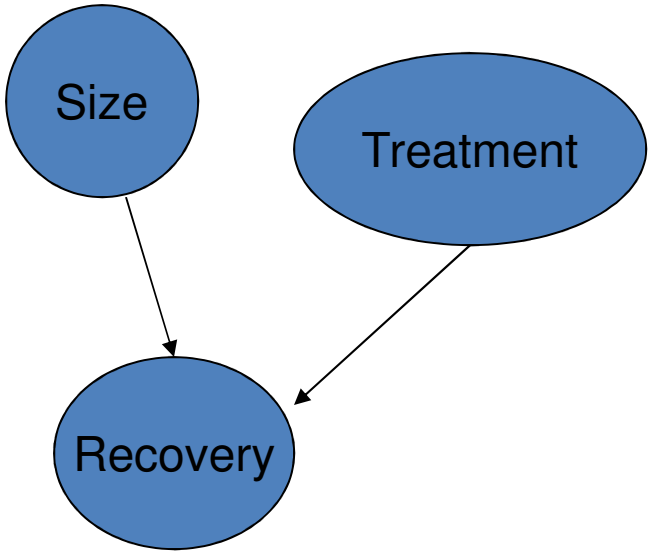
Berkeley sex bias case

	Applicants	% admitted
Men	8442	**44%**
Women	4321	35%

Department	Men		Women	
	Applicants	% admitted	Applicants	% admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

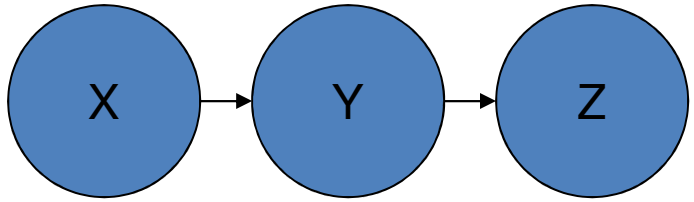
Berkson's paradox

- "The explaining away" phenomena
  - Size and treatment are conditionally dependent



Többváltozós függések intranzitivitása I.
„Az idő nyila”

- Intranzitív függés (pl. X és Z lehet független)



p(X),p(Z|X),p(Y|Z)

p(X|Z),p(Z|Y),p(Y)

p(X|Z),p(Z),p(Y|Z)

“tranzitív” M ≠ „intranzitív” M

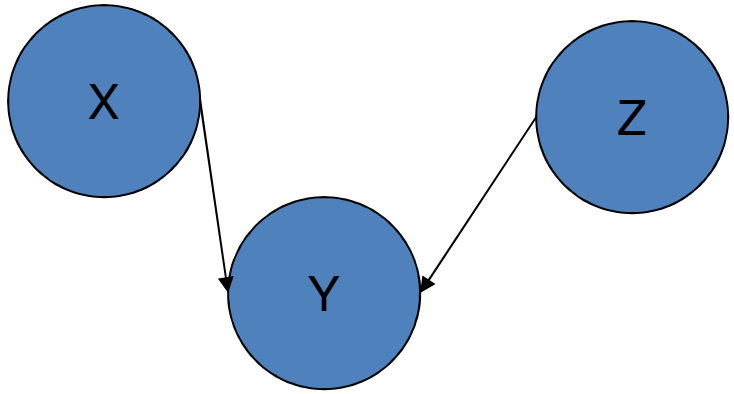
p(X),p(Z|X,Y),p(Y)

„Az idő nyíl”
‘v-struktúra’

Többváltozós függések intranzitivitása II. Interakciók

- Alsóbbrendű függetlenségből nem következik magasabbrendű

  (X,Z)-tól együtt függhet Y, mégha páronként független is



G.F.Cooper: Understanding your data I.

- Crisis: evidence-based medicine
- IV. CONCEPTS OF CAUSATION & CONFOUNDING
- 8. What does it mean that A causes B?
- a. Hume – counterfactuals
- b. Proof of causation seldom (ever?) obtainable in clinical research
- c. Definition of confounding
- V. CAUSATION, STATISTICS, AND JOINT DISTRIBUTIONS – THEIR RELATION
- 9. What is a joint distribution?
- 10. How do statistics relate to joint distributions?
- 11. How do statistics relate to causation?
- a. Statistical significance tells using nothing about causation
- b. Cartwright – “No causes in, no causes out”
- c. Data + Assumptions = Inference
- VI. CURRENT PRACTICE
- 12. Perform classical analysis
- a. Generate p-values, confidence intervals, etc.
- b. Interpret p-values as probabilities, confidence intervals as posterior intervals
- c. Making passing (non-quantitative) reference to potential problems with the methods in the limitations section

Understanding your data II.

- VII. IMPROVED PRACTICE
- 13. Fundamental shift in perspective
- a. Abandon hypothesis testing
- b. Emphasize estimation of effects
- c. Trying to learn, not trying to prove
- d. Be explicit about assumptions
- 14. Clarity regarding what is being studied
- a. Theoretical model
- i. Causal diagrams (Pearl)
- b. Relate study design to theoretical model
- 15. Understand your data
- a. Begin with graphical appreciation of data
- i. Check for reasonableness of univariate data (STATA *codebook command)*
- ii. Explore bivariate relationships
- iii. Explore multivariate relationships (Datadesk example)
- b. Explore your data
- i. Link effort to the theoretical model
- ii. Creative process – an interaction between the data and the analyst

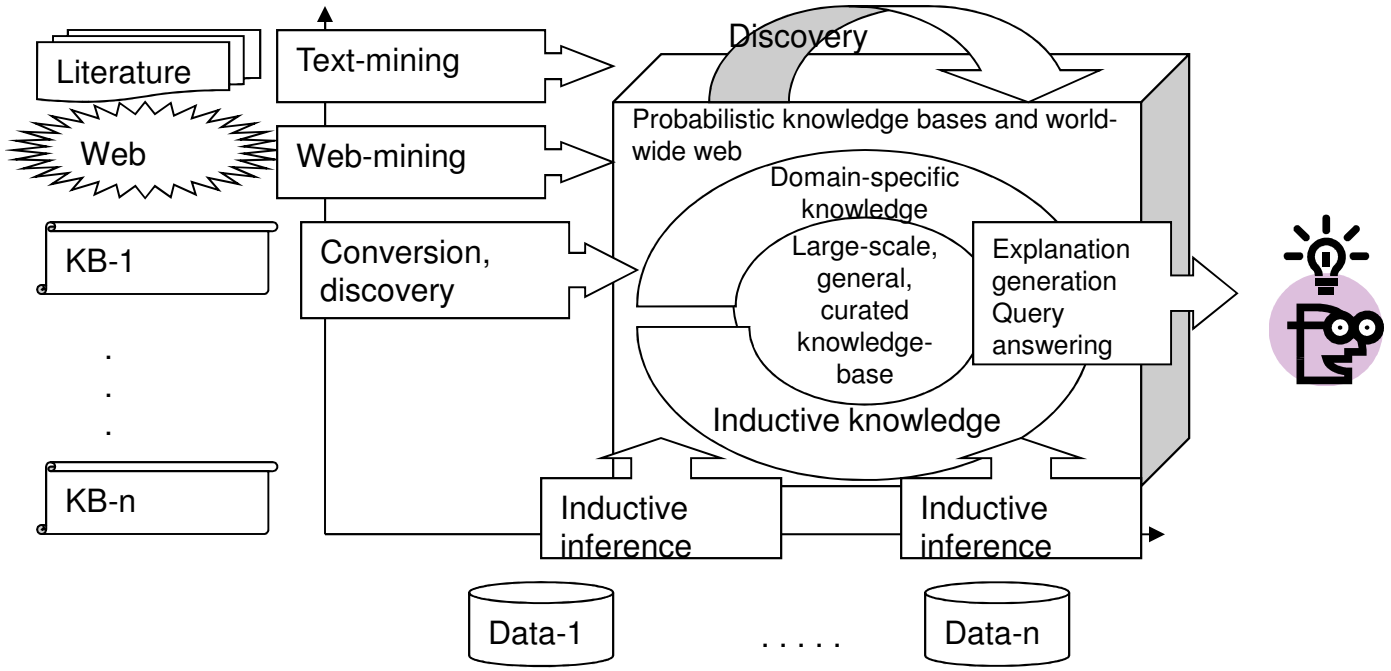
Understanding your data III.

- 16. Analyze your data
- a. Enumerate your assumptions
- i. a *priori knowledge*
- ii. selection bias
- iii. unmeasured confounding
- iv. measurement error
- v. (other factors per the situation)
- b. Perform crude analysis
- c. Perform analyses that include explicit adjustment for *a priori knowledge and assumptions*
- d. Perform additional sensitivity analyses
- 17. Presenting your investigation
- a. Begin with the theoretical model (Most disagreements can be found here)
- b. Explain your analytic strategy
- i. Link to theoretical model
- ii. Rationale for: 01. Lumping categories 02. Splitting categories (choice of cutpoints) 03. Choice of statistical model
- c. Show the data, not summaries of the data or statistics about the data
- iii. A graphical analysis may be all the analysis you need
- **01. Novel methods such as websites that allow for recalculation with user's prior are an attractive possibility**
- **i**ii. Incorporate sensitivity analyses into the Methods & Results

Understanding your data IV.

- VIII. BENEFITS OF IMPROVED PRACTICE
- 18. Subjective, but transparent and honest
- 19. Shifts the argument from one about conclusions to one about assumptions
- 20. Eliminates illusion of objectivity
- 21. Eliminates covert assumptions that are often false or unrealistic
- 22. Typically leads to greater uncertainty regarding results, thereby eliminating premature closure

The ultimate understanding…



- **multivariate Bayesian data analysis**,
- the use of more powerful **logic for representing knowledge**,
- **uncertainty management** in knowledge representation, induction, and inference,
  syntactic semantics: $P(\phi \mid KB) = \sum_{\text{pKB}} P(\phi \text{ is provable} \mid pKB, lKB) p(pKB)$

Meta-rationality

Nothing is more practical than a good theory (J.C.Maxwell)

The most incomprehensible thing about the world is that it is at all comprehensible.
Albert Einstein.

No theory of knowledge should attempt to explain why we are successful in our attempt to explain things.
K.R.Popper: Objective Knowledge, 1972

- Theory is where you know everything and nothing works. Practice is where things work, but noone knows why. Here, we combine theory and practice. Nothing works, and noone knows why. ;-)
-

Understanding or using your data

- Thinking/understanding or acting
- Prediction vs interpretation (predictive vs parametric)
- Optimal action using your data: ambient assisted data analysis
  - Intelligence without representation, Brooks...
  - Intelligent house…

“Ambient assisted” data analysis

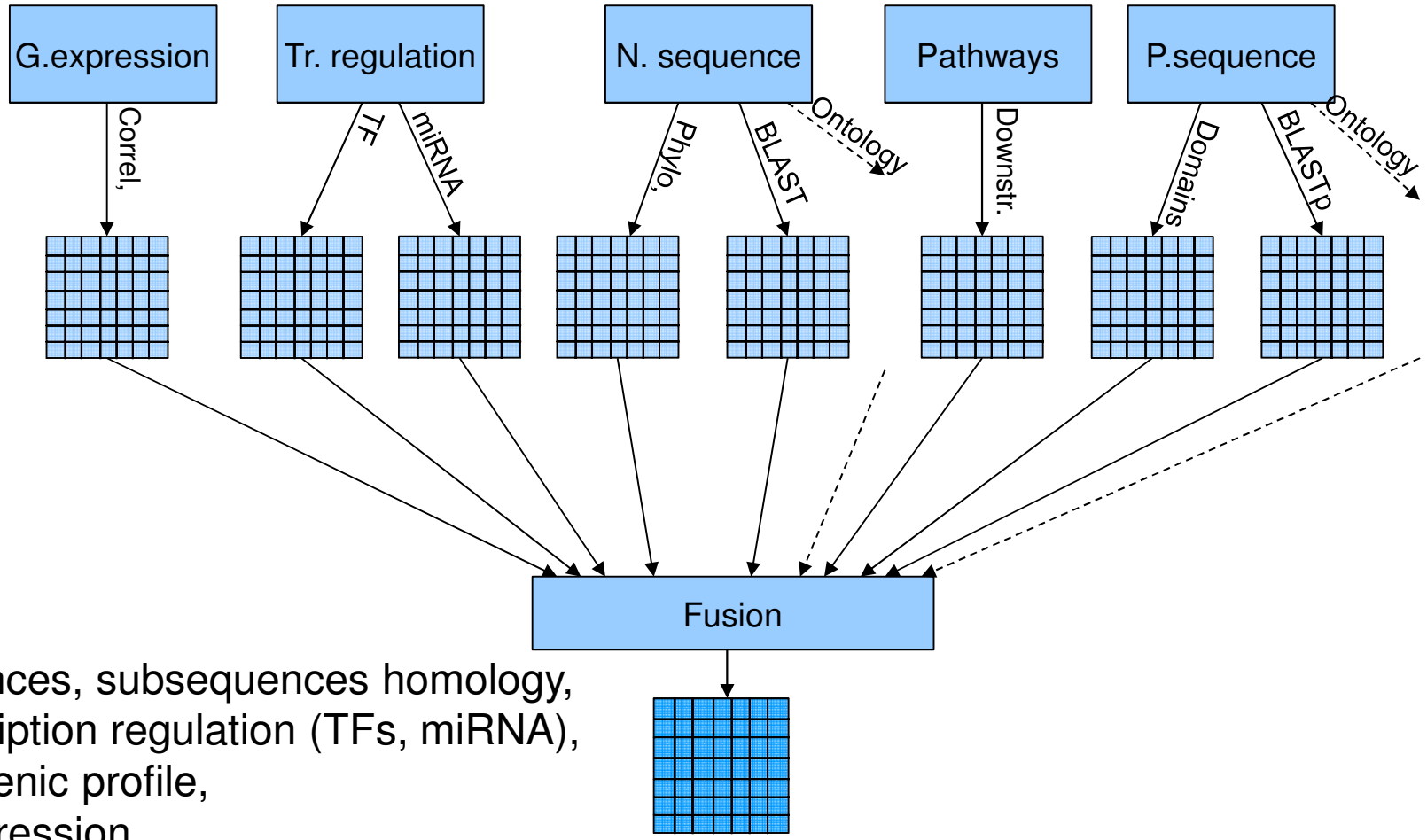
- J.Lamb: CMap
- The ultimate objective of biomedical research is to connect human diseases with the genes that underlie them and drugs that treat them. But this remains a daunting task, and even the most inspired researchers still have to resort to laborious screens of genetic or chemical libraries. 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.

Hopeless to understand??

- „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!!)

Foundation for induction: Probability theory, decision 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....

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

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

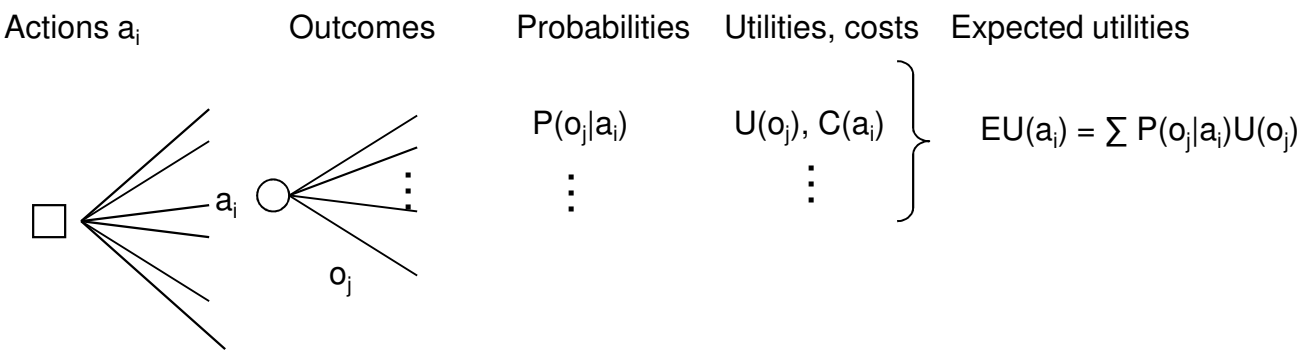
In the Bayesian approach models are weighted

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

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



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:

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

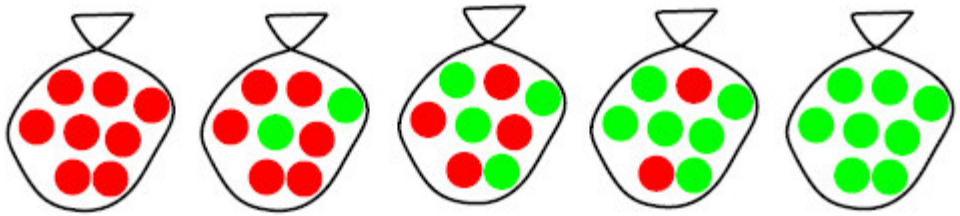
No need to pick one best-guess hypothesis!

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

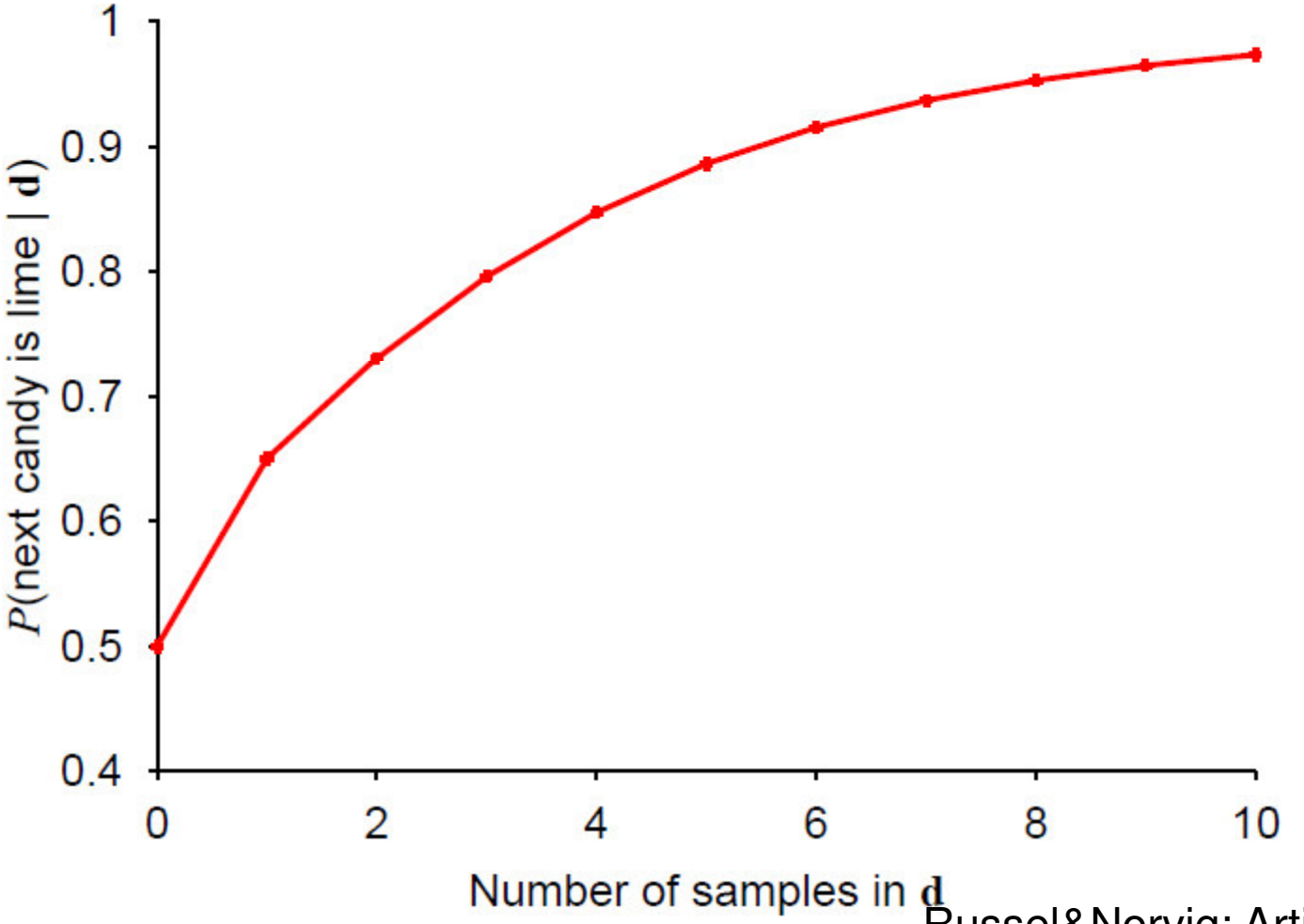


Then we observe candies drawn from some bag: 

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

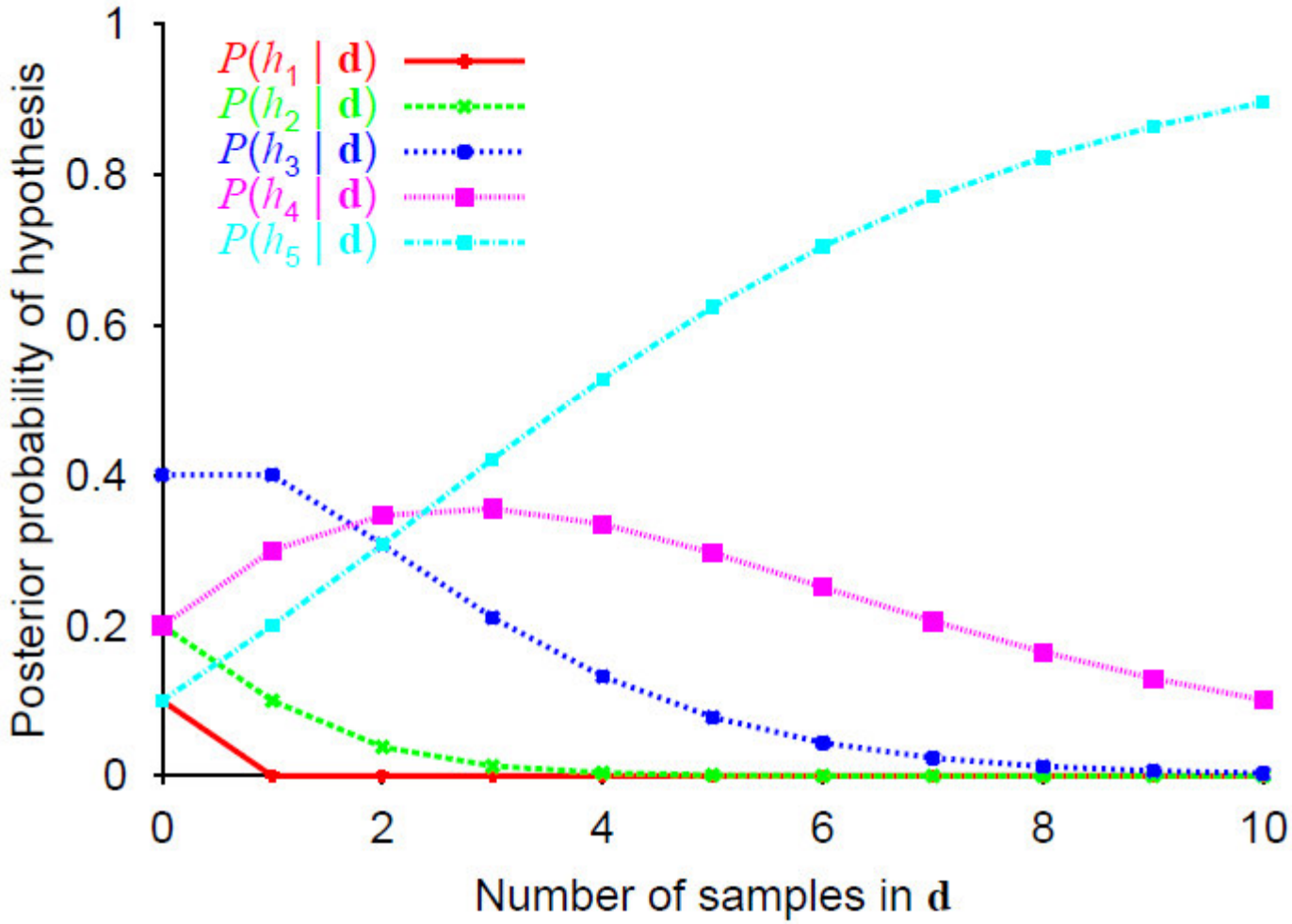
Russel&Norvig: Artificial intelligence

Learning rate for model predictions/properties



Russel&Norvig: Artificial intelligence

Learning rate for



Russel&Norvig: Artificial intelligence

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

The prequential approach

- The sequential update of the posterior

If we interpret the forecasts in a decision theoretic framework as reporting of the posteriors, then the score function is a loss function $S(\underline{q}, s_k)$ and $\underline{q_n}$ should correspond to the minimal loss forecast (see Eq. 2.9)

$$\arg\min_{\underline{q}} \sum_{k=1}^{r} S(\underline{q}, s_k) p_n(X_n = s_k | x_1, \ldots, x_{n-1}). \tag{2.28}$$

It can be shown that the requirements of honesty ("reporting true beliefs"), smoothness ("proportional penalty for errors") and decomposability (penalty depends on pairs of {forecasts-outcomes}) characterize a logarithmic score function, $S(\underline{q}, s_k) = A \log(q_k) + B_k$ where $A < 0$ and B_k are arbitrary constants [34].

$$\begin{aligned} S &= \sum_{i=1}^{n} S_i(p_i(X_i | x_1, \ldots, x_{i-1}), x_i) \\ &= -\log \prod_{i=1}^{n} p_i(x_i | x_1, \ldots, x_{i-1}) \\ &= -\log p(x_1, \ldots, x_n). \end{aligned}$$