Machine learning
&
data science

Peter Antal

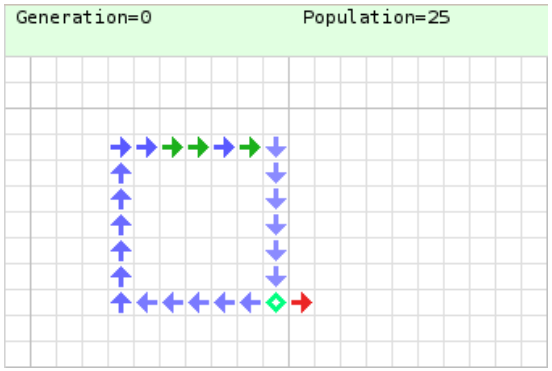
antal@mit.bme.hu

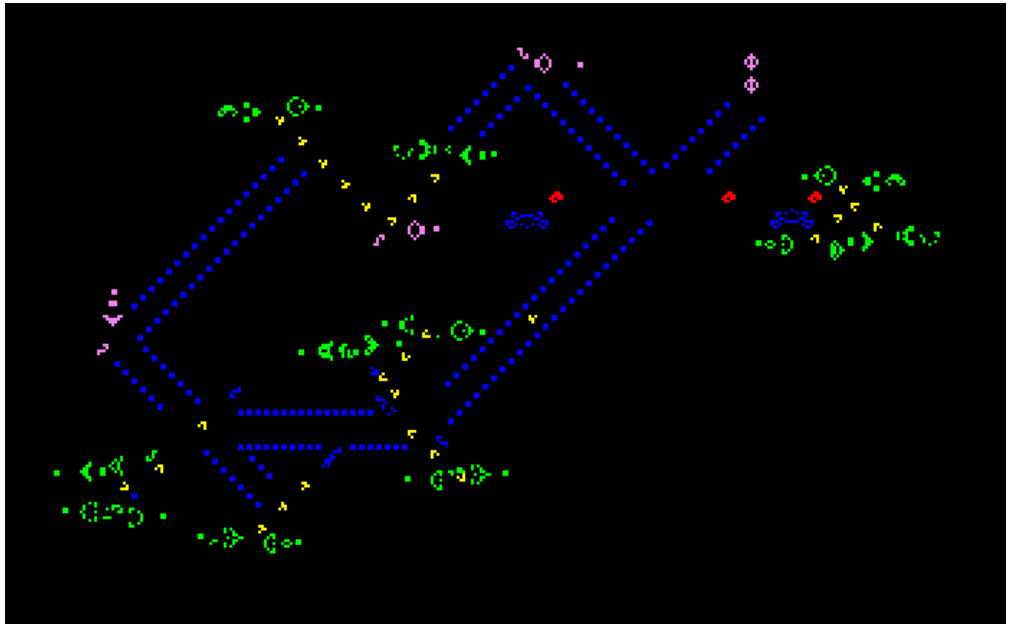
Overview

- Why can we learn?
- Learning as optimal decision and as probabilistic inference
- PAC learning
- The data flood and the big data
- Data science, e-science

Why can we learn? I.

- Regularities, compressibility,..
- The recursive universe
  - patterns, artificial life, emergence,..





Why can we learn? II.

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

Nobody knows ;-), predictability, understandability and computability are empirical observations.

➔Decision theory (DT) is at least a coherent framework.
➔➔ Bayesian model averaging (BMA) follows from DT.
➔➔➔ MAP/ML learning is at least a reasonable approximation of BMA.
➔➔➔➔ Regularized ML learning has a strong classical statistical background.

Optimal decision: decision theory probability theory+utility theory

- Decision situation:
  - Actions
  - Outcomes
  - Probabilities of outcomes
  - Utilities/losses of outcomes
  - Maximum Expected Utility Principle (MEU)
  - Best action is the one with maximum expected utility

$$a_i$$

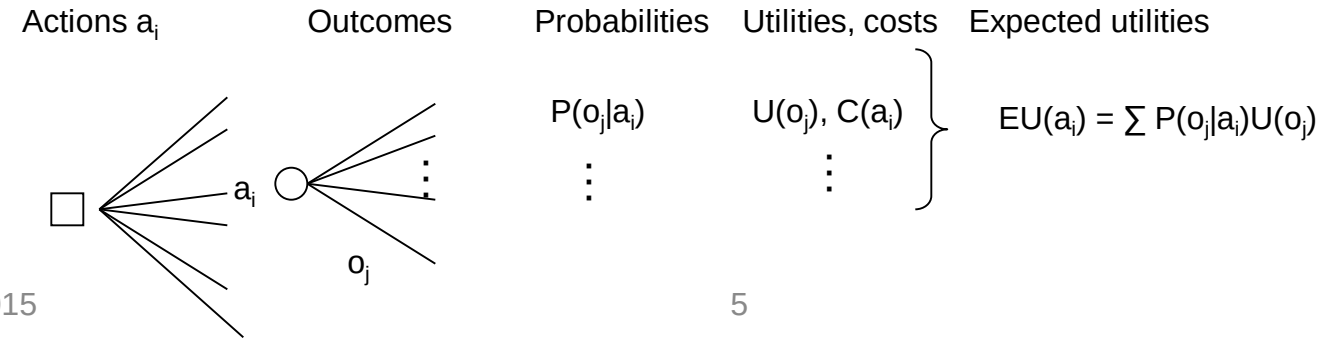
$$o_j$$

$$p(o_j \mid a_i)$$

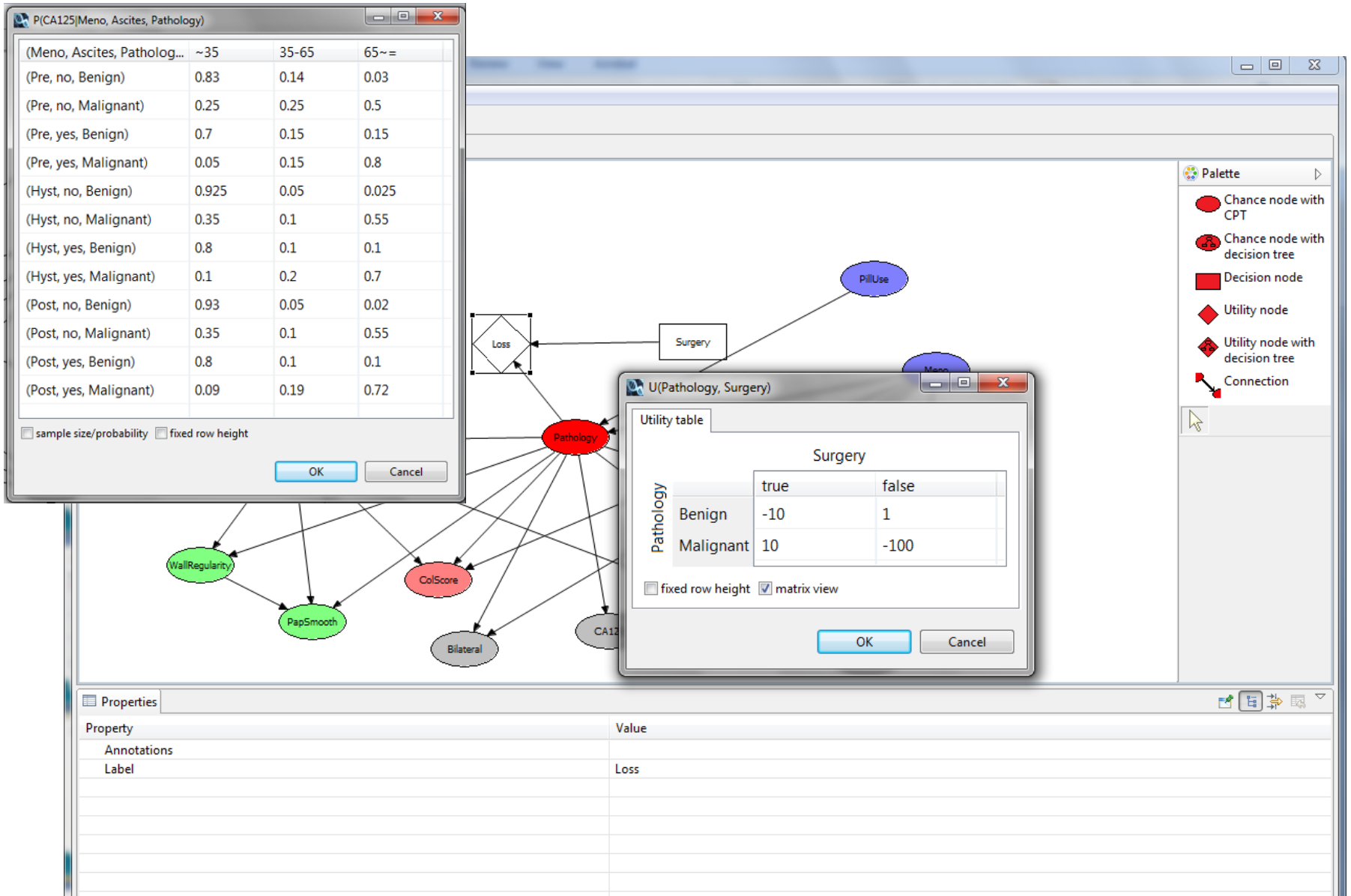
$$U(o_j \mid a_i)$$

$$EU(a_i) = \sum_j U(o_j \mid a_i)\, p(o_j \mid a_i)$$

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



Decision support systems, decision networks



Bayesian model averaging with data

Beside models, assume N multiple complete observations D_N.

The standard inference $p(Q = q|E = e, D_N)$ is defined as:

$$p(q|e, D_N) = \sum_{i=1,\ldots,M} p(q, M_i|e, D_N) = \sum_{i=1,\ldots,M} p(q|M_i, e, D_N)p(M_i|e, D_N)$$

Because $p(q|M_i, e, D_N) = p(q|M_i, e)$ and $p(M_i|e, D_N) \approx p(M_i|D_N)$:

$$p(q|e, D_N) \approx \sum_{i=1,\ldots,M} p(q|M_i, e)p(M_i|D_N)$$

where again $p(M_i|D_N)$ is a posterior after observations D_N:

$$p(M_i|D_N) = \frac{p(D_N|M_i)p(M_i)}{p(e)} \propto \underbrace{p(D_N|M_i)}_{likelihood} \underbrace{p(M_i)}_{prior}.$$

i.e., our rational foundation, probability theory, automatically includes and normatively defines learning from observations as standard Bayesian inference!

Frequentist vs Bayesian prediction

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

$$p(prediction|\, data) = p(prediction|\, BestModel(data))$$

In the Bayesian approach models are weighted

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

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

Principles for induction

- 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.

Laws of large numbers

$$\bar{f} = E_{\pi(X)}[f(X)]$$
$$\hat{f}_N = 1/N \sum_{t=1}^{N} f(X_t)$$

Markov's inequality: If X is any nonnegative integrable random variable and $a > 0$, then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

The estimate $\hat{f}_N$ is strongly consistent (by the "strong law of large number"), that is

$$P(lim_{N\to\infty} \hat{f}_N = \bar{f}) = 1 \tag{54}$$

The standardized of $\hat{f}_N$ has asymptotically Gaussian distribution (by the "central limit theorem"), that is

$$\frac{\hat{f}_N - \bar{f}}{\sigma_N} \to N(0,1)\ as\ N \to \infty\ \ where\ \sigma_N = Var(f(X))/\sqrt{N}. \tag{55}$$

If $f(X)$ is bounded, then non-asymptotic results about the speed of convergence are also available by the Hoeffding's inequality including the bound and by the Bernstein's inequality. Specifically, if $f(X)$ is within $[0,1]$, then

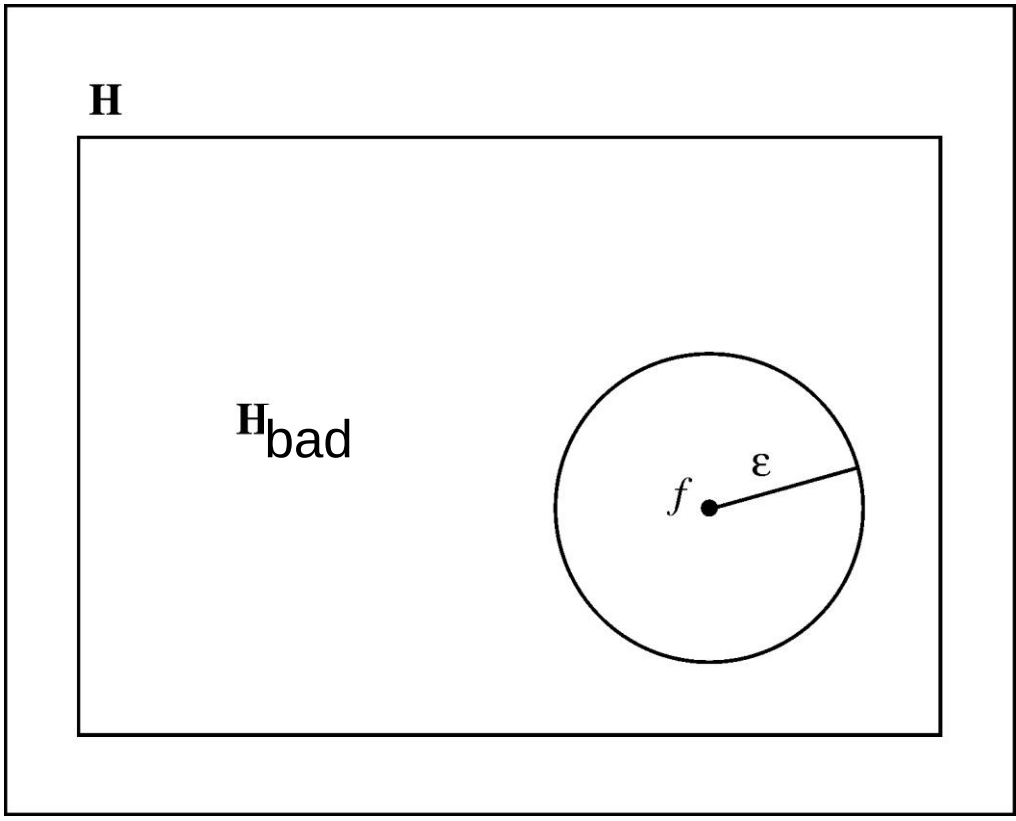
$$p(|\hat{f}_N - \bar{f}| \geq \epsilon) \leq 2\exp(-2\epsilon^2 N) \leq \delta \tag{56}$$
$$E[\ |\hat{f}_N - \bar{f}|\ ] \leq \sqrt{c_0/N}. \tag{57}$$

The Probably Approximately Correct PAC-learning

A single estimate is convergent, but can we estimate uniformly well the error/performance if many hypotheses?

Example from concept learning

X: i.i.d. samples.
m: sample size
H: hypotheses



Assume that the true hypothesis *f* is element of the hypothesis space **H**.

Define the error of a hypothesis h as its misclassification rate:

$$error(h) = p(h(x) \neq f(x))$$

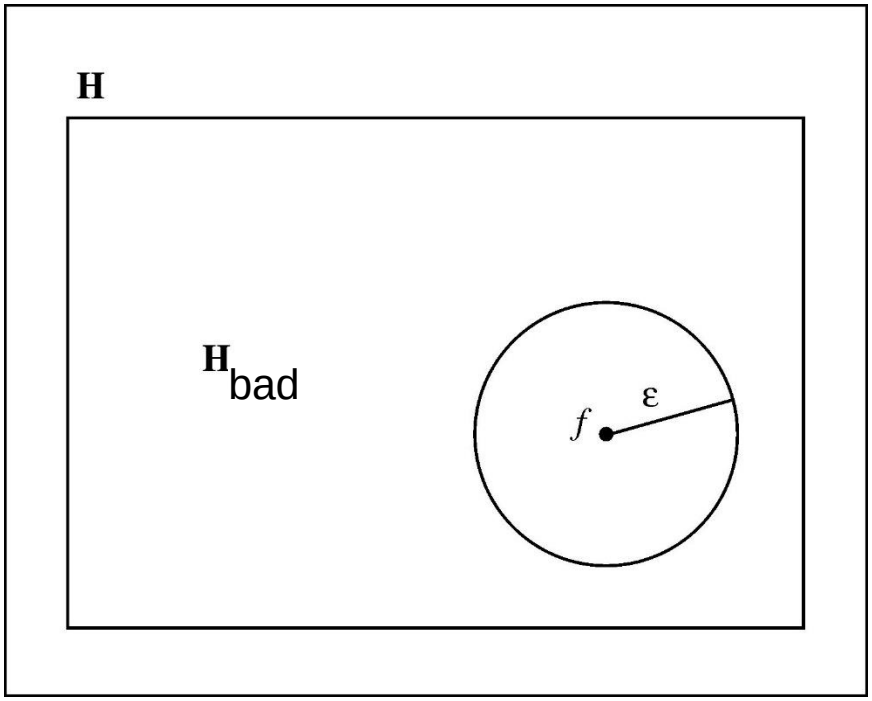
Hypothesis h is **approximately correct** if

$$error(h) < \varepsilon$$

(ε is the "accuracy")

For $h \in H_{bad}$ $\quad error(h) > \varepsilon$

H can be separated to $H_{<\varepsilon}$ and H_{bad} as $H_{\varepsilon<}$.



By definition for any h ∈ H_{bad}, the probability of error is larger than ε, thus the probability of no error is less than $\leq (1-\varepsilon)$

Thus for m samples for a $h_b \in H_{bad}$:

$$p\left(D_m{:}h_b(x) = f(x)\right) \leq (1-\varepsilon)^m$$

For any $h_b \in H_{bad}$, this can be bounded as

$$\begin{gathered} p\left(D_m{:}\forall h_b \in H, h_b\ (x) = f(x)\right) \leq \\ \leq |H_{bad}|(1-\varepsilon)^m \\ \leq |H|\ (1-\varepsilon)^m \end{gathered}$$

To have at least δ "probability" of approximate correctness:

$$\left|\mathbf{H}\right|(1-\varepsilon)^m \leq \delta$$

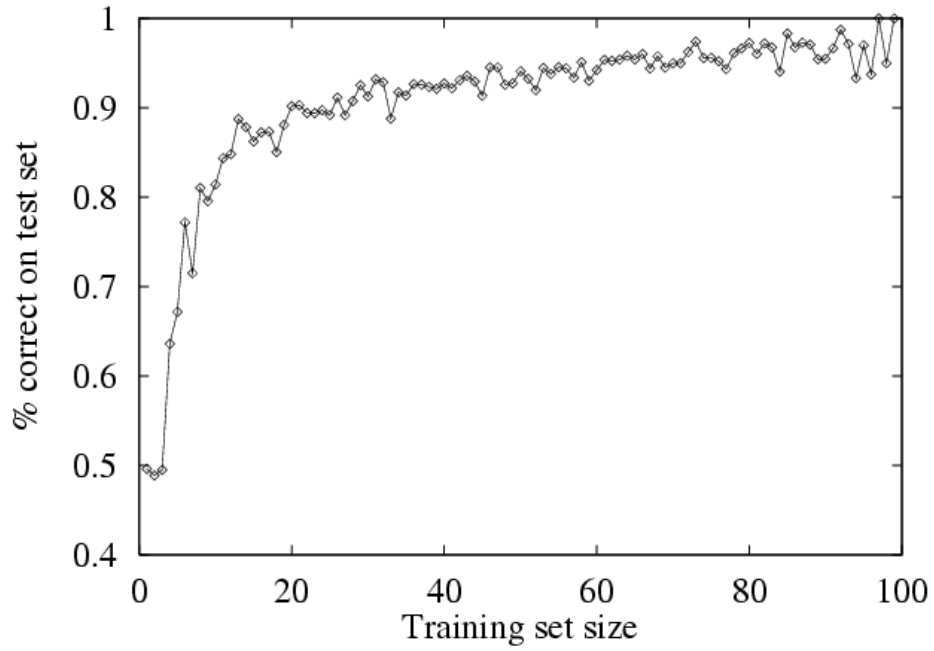
By expressing the sample size as function of *ε accuracy* and *δ confidence* we get a bound for ***sample complexity***

$$m \geq \frac{1}{\varepsilon}(\ln\frac{1}{\delta} + \ln\left|\mathbf{H}\right|)$$

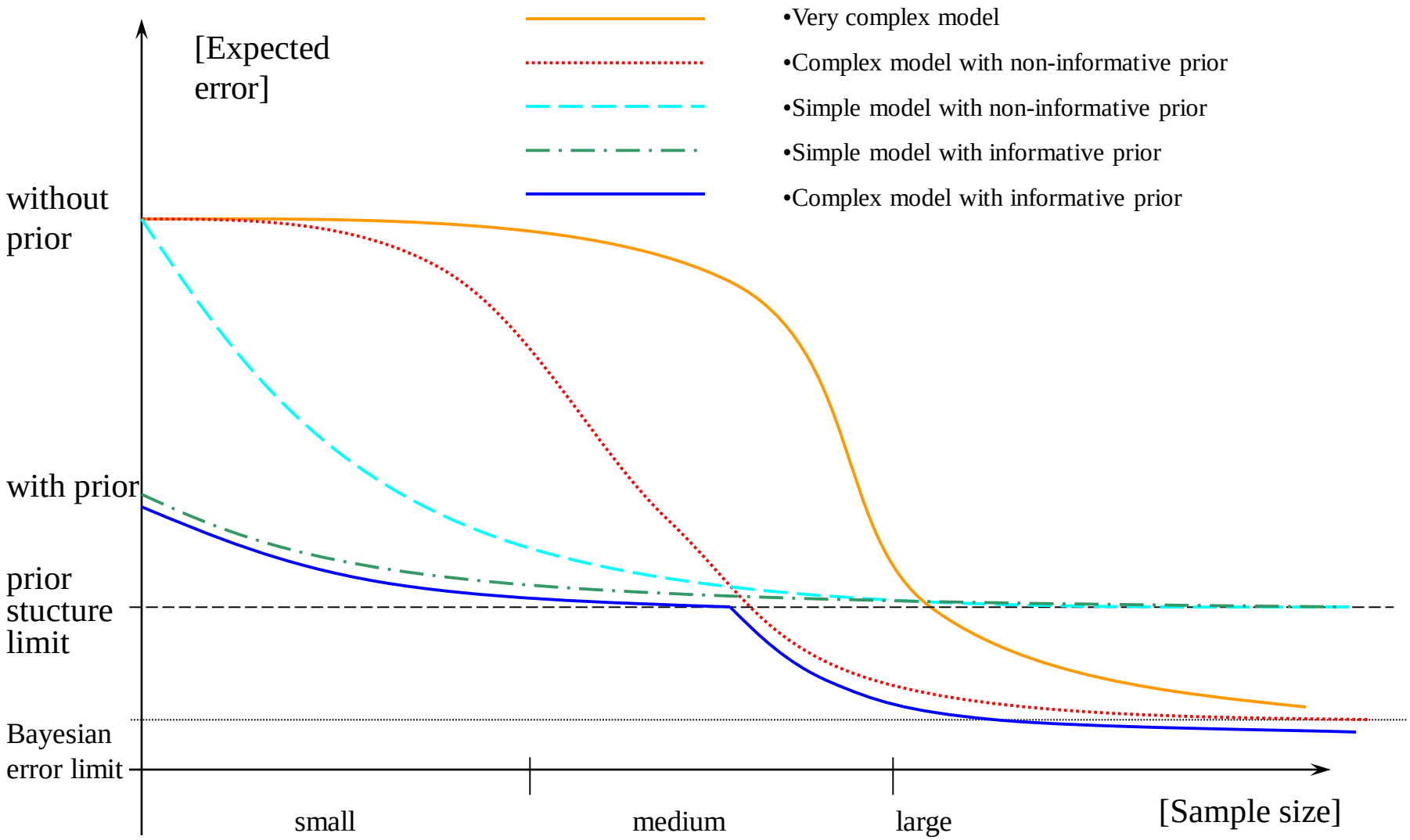
Estimation of future performance

- How do we know that $h \approx f$?
  1. Use theorems of computational/statistical learning theory
  2. Try h on a new test set of examples
     (use same distribution over example space as training set)

Learning curve = % correct on test set as a function of training set size



Learning characteristics of various methods

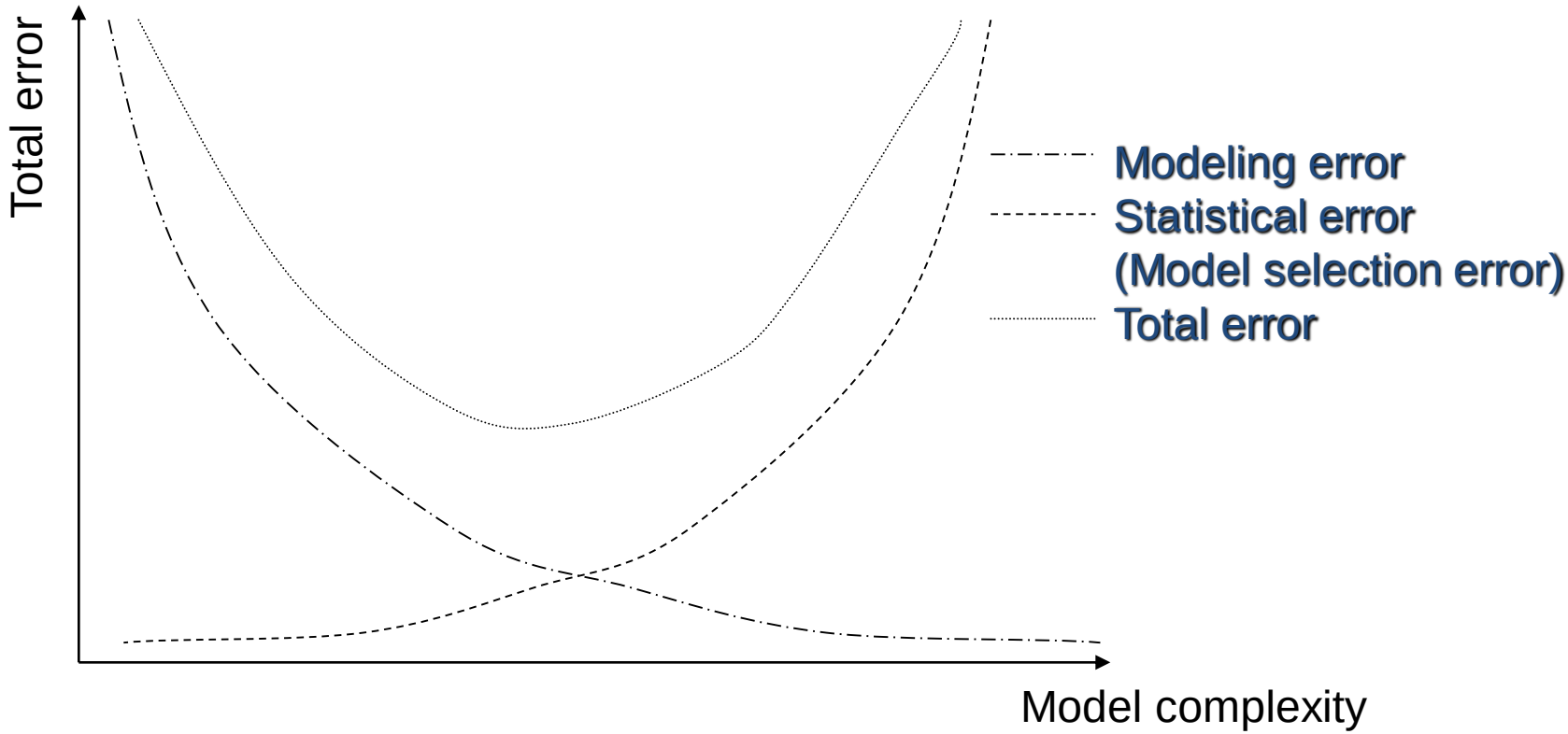


The bias-variance dilemma

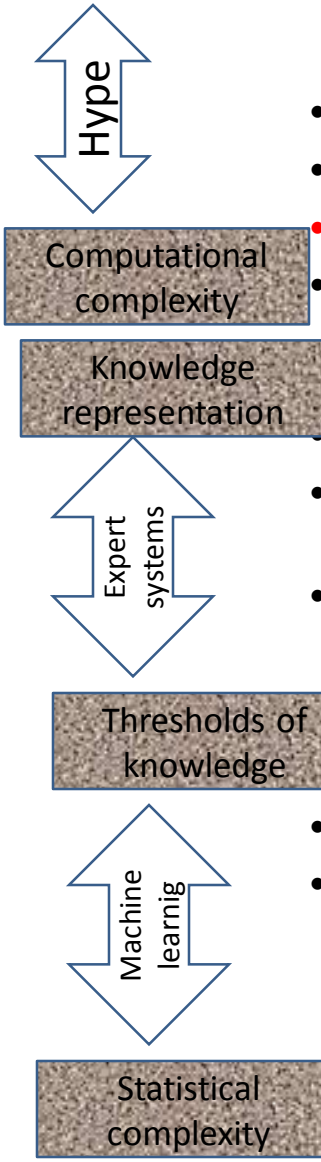
In practice, the target typically is not inside the hypothesis space: the total real error can be decomposed to “bias + variance”

- “bias”: expected error/modelling error
- “variance”: estimation/empirical selection error

For a given sample size the error is decomposed:



Can we bypass human (expertise) by data?

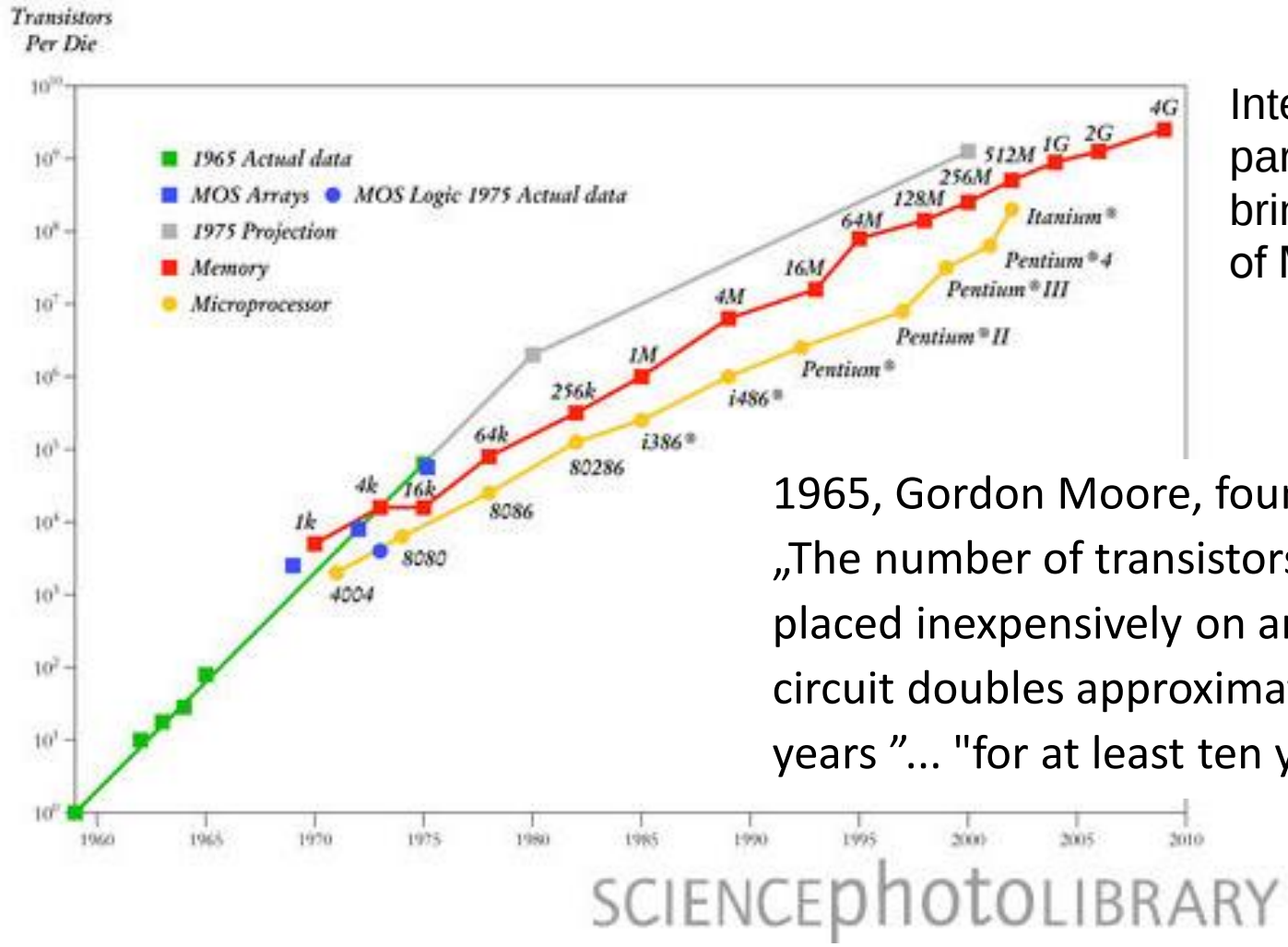


- 1943 McCulloch & Pitts: Boolean circuit model of brain
- 1950 Turing's "Computing Machinery and Intelligence"
- 1956 Dartmouth meeting: the term "Artificial Intelligence"
- 1950s Early AI programs, including Samuel's checkers program, Newell & Simon's Logic Theorist, Gelernter's Geometry Engine
- 1965 Robinson's complete algorithm for logical reasoning
- 1966—73 AI discovers computational complexity Neural network research almost disappears
- 1969—79 Early development of knowledge-based systems

- 1986-- Neural networks return to popularity
- 1988-- Probabilistic expert systems
- 1995-- Emergence of machine learning

Today: heterogeneous AI, data-intensive science, data and knowledge **fusion, automated science**

Moore's Law (in computation)



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

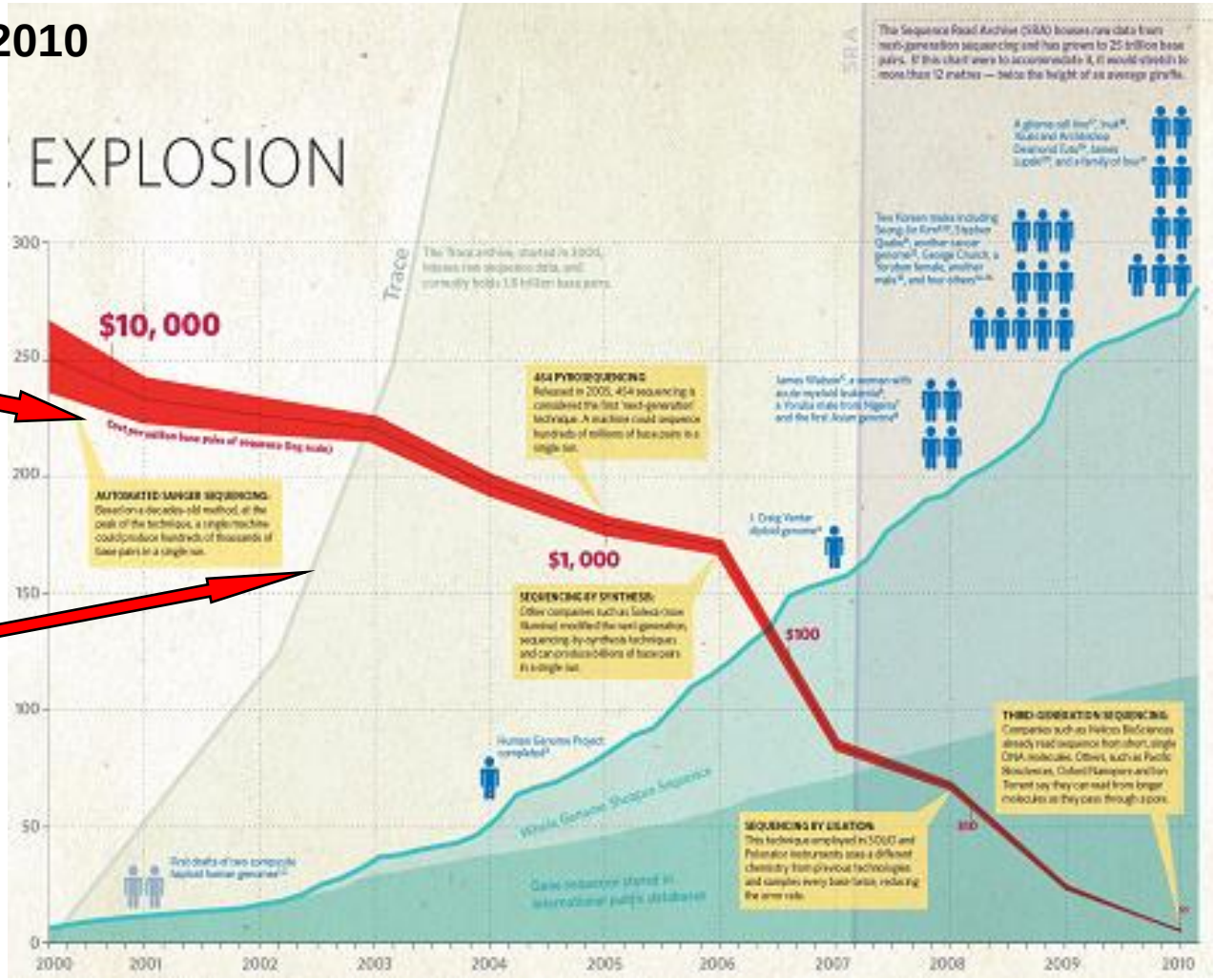
1965, Gordon Moore, founder of Intel: „The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years "... "for at least ten years"

Carlson's Law for Biological Data

NATURE, Vol 464, April 2010

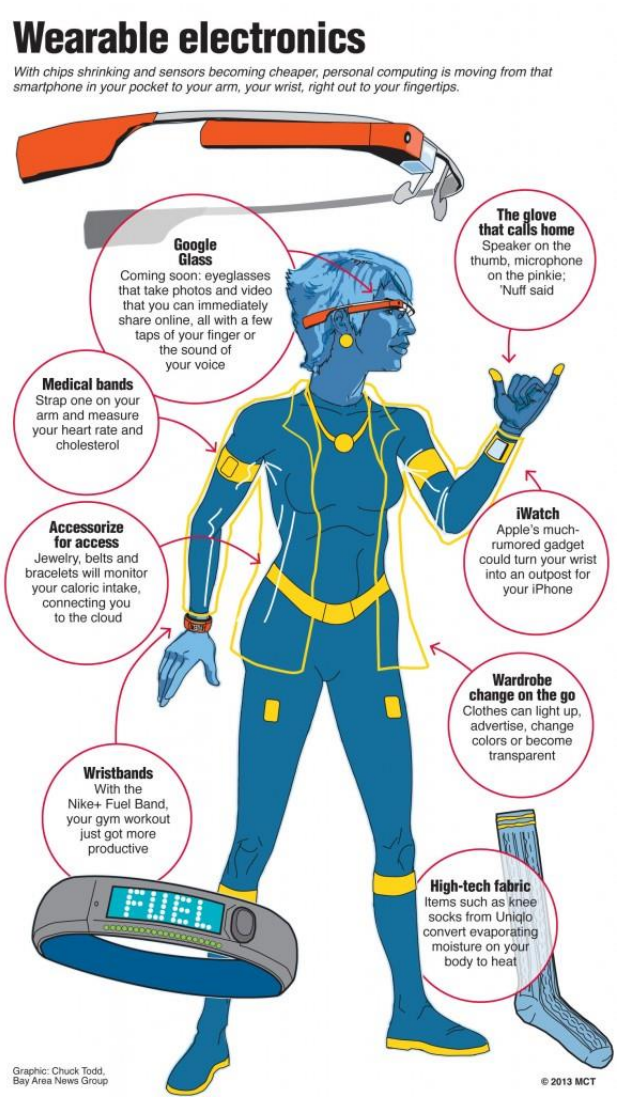
Sequencing costs per mill. base

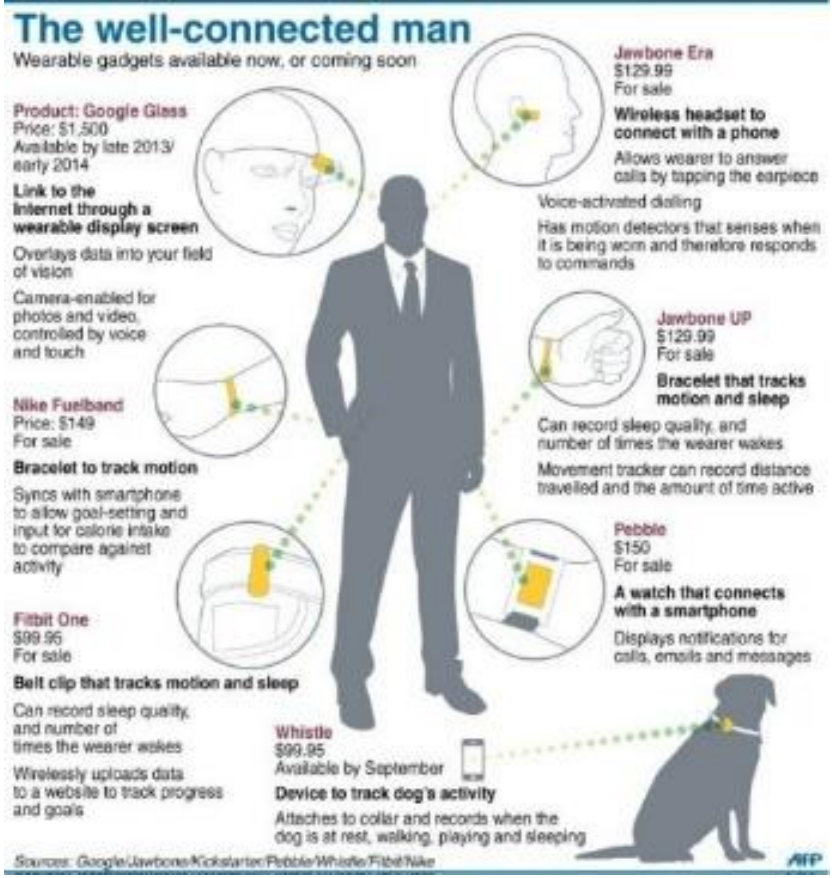
Publicly available genetic data

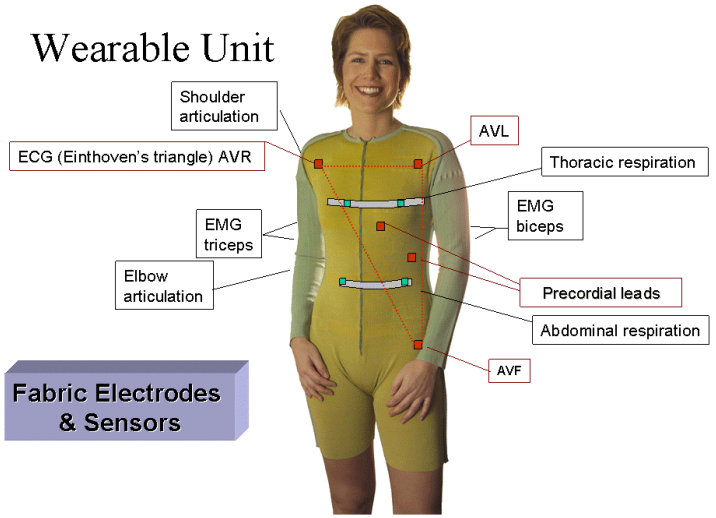


- x10 every 2-3 years
- Data volumes and complexity that IT has never faced before…

Quantified self





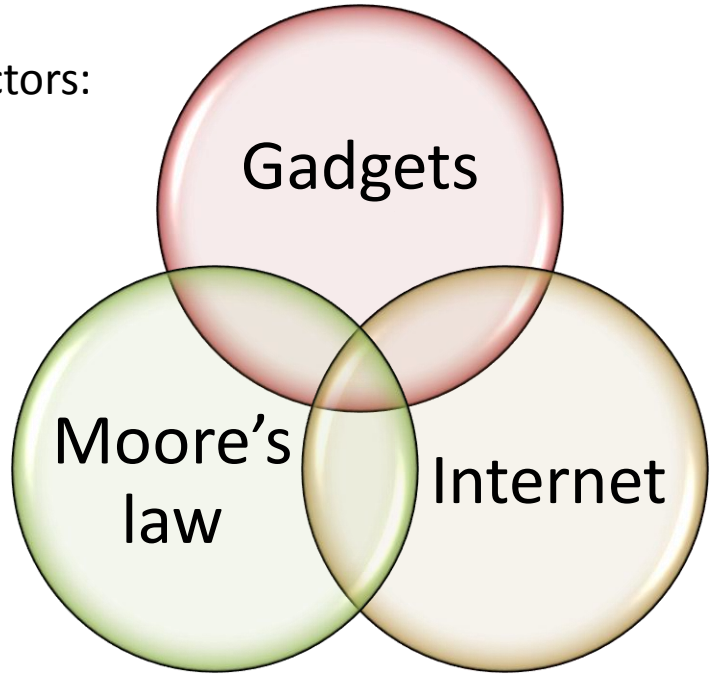


Precursors for the „big” data

- 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).

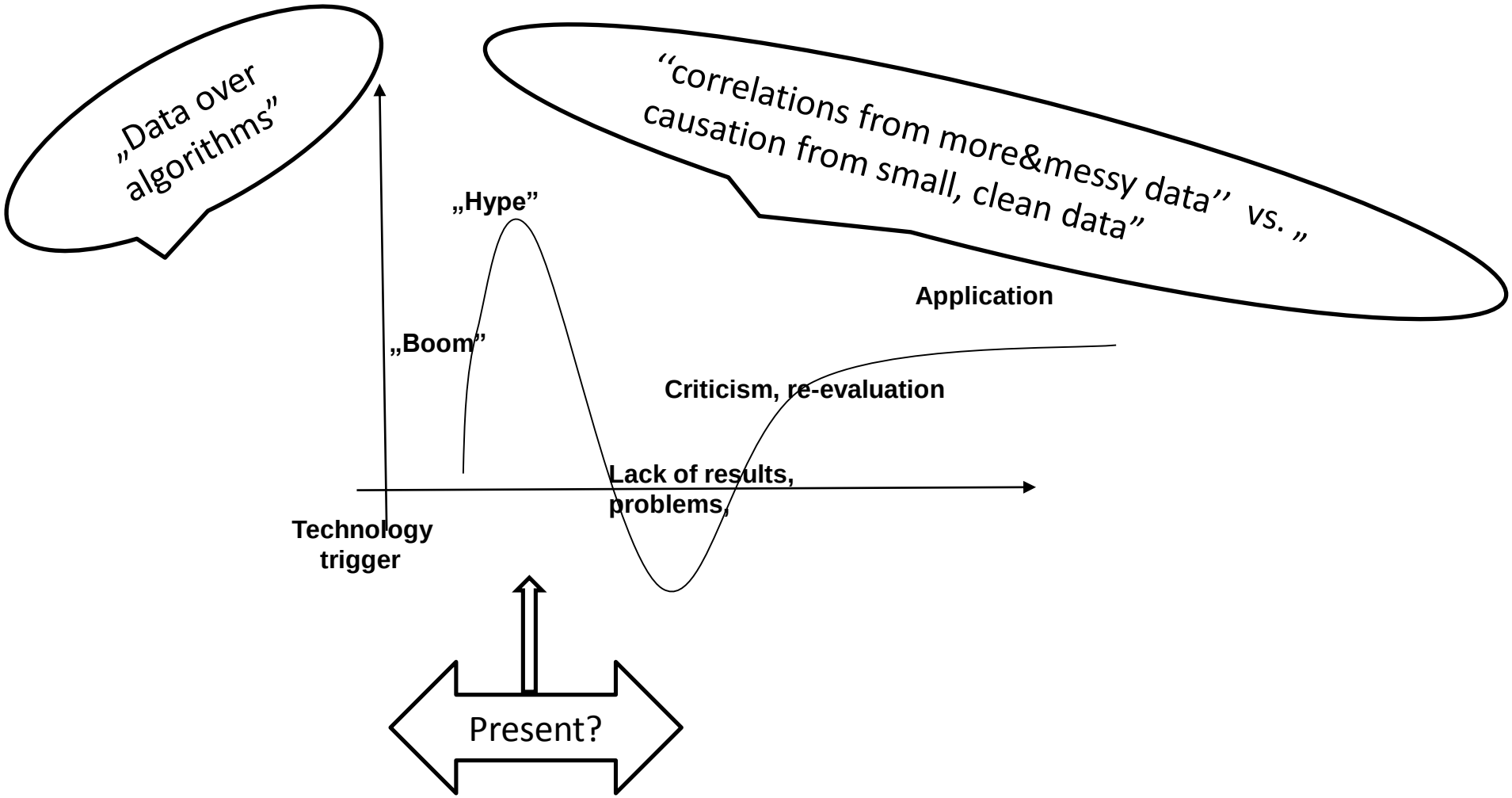
The 8xV: Vast, Volumes of Vigorously, Verified, Vexingly Variable Verbose yet Valuable Visualized high Velocity Data (2013)

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)

The „omic" definition of „big data"

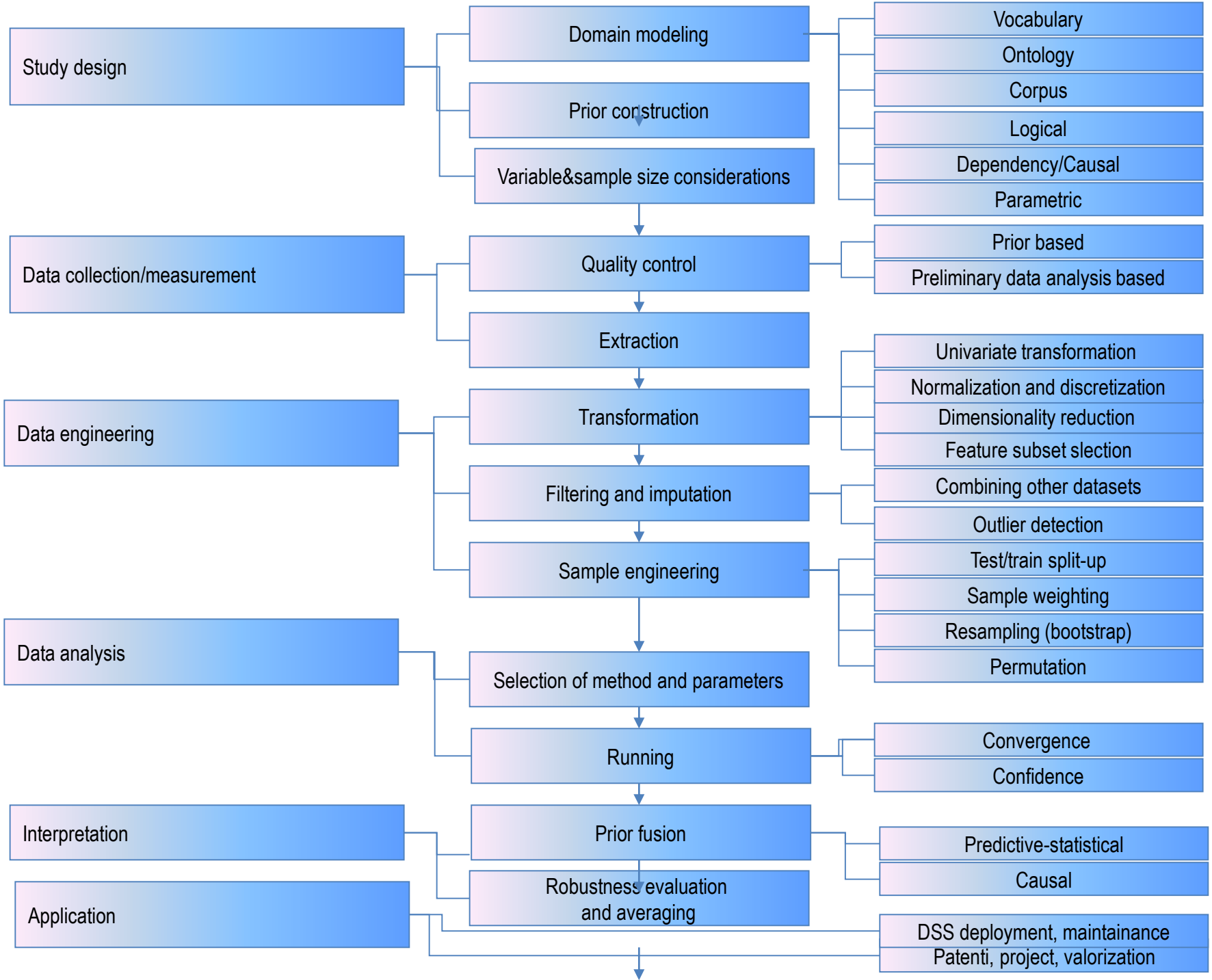
.. [data] is often big in relation to the phenomenon that we are trying to record and understand. So, if we are only looking at 64,000 data points, but **that represents the totality or the universe of observations. That is what qualifies as big data. You do not have to have a hypothesis in advance before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

„Big data, big hype”

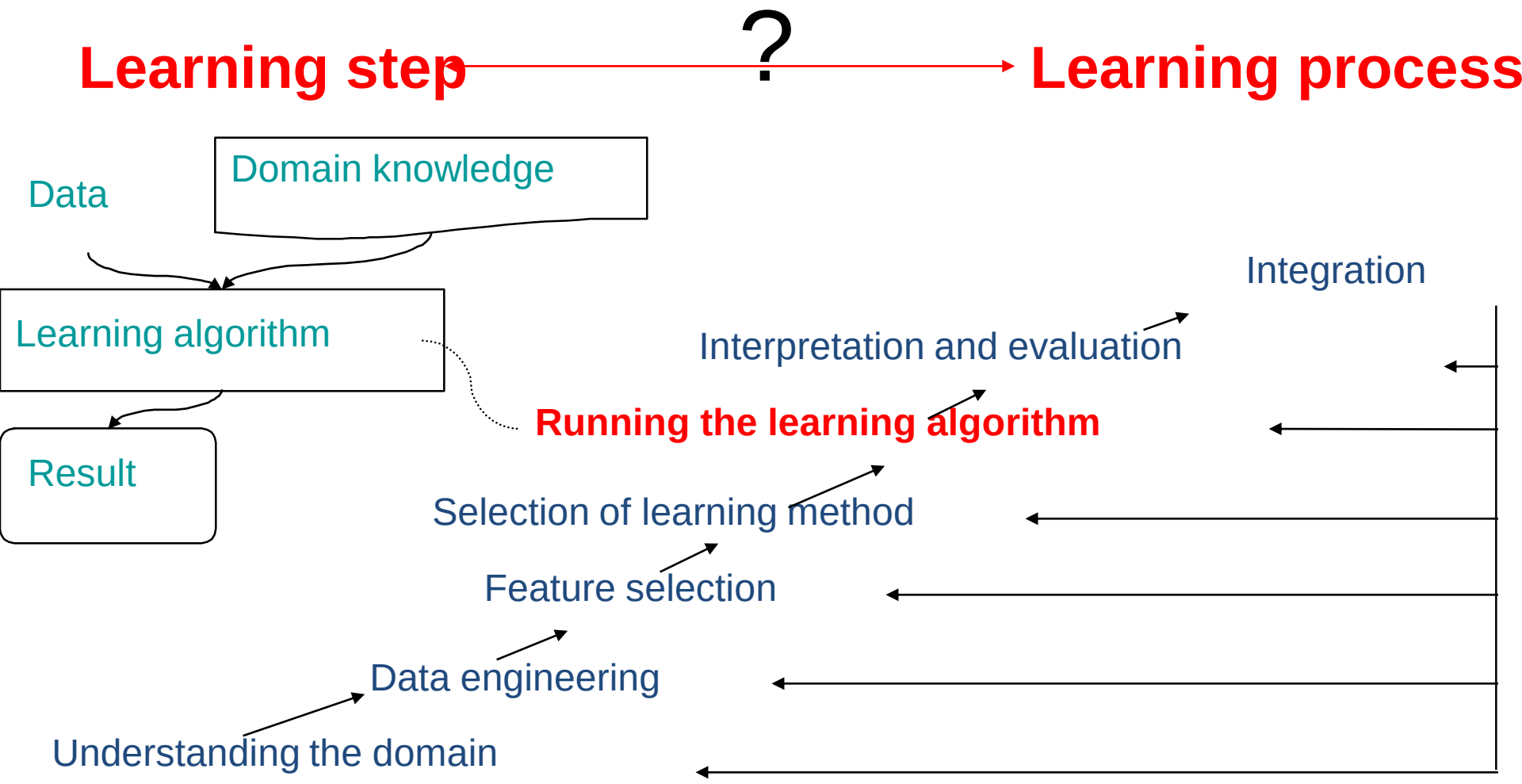


Intelligent data analysis

- 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.


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

Learning step or learning **process?**

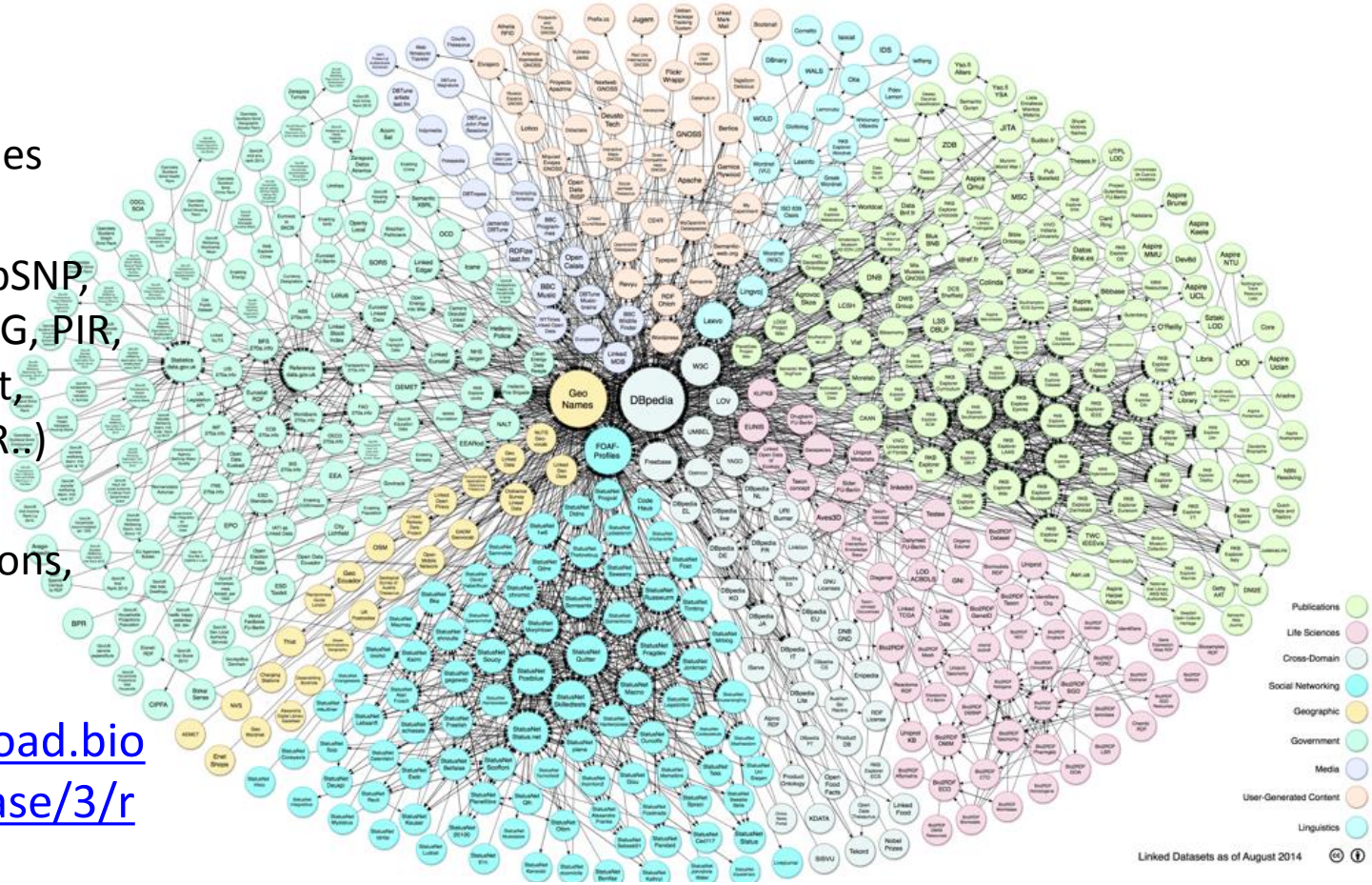


Open systems for machine learning

- R
- Julia
- IBM: SystemML
- Google: TensorFlow
- FB: Torch
- MS: Azure

Open linked data

- Bio2RDF
- ~11 billion triples
- 35 datasets: clinicaltrials, dbSNP, DrugBank, KEGG, PIR, GOA, OrphaNet, PubMed, SIDER..)
- local: chembl, pathwaycommons, reactome, wikipathways
- http://download.bio2rdf.org/release/3/release.html



Dataset	Date generated	# of triples	# of unique entities
		11900533153	1108204952

- **Discovery Platform** to cross barriers.
- The data sources you **already use**, **integrated** and **linked** together: *compounds*, *targets*, *pathways*, *diseases* and *tissues*.
- ChEBI, ChEMBL, ChemSpider, ConceptWiki, DisGeNET, DrugBank, Gene Ontology, neXtProt, UniProt and WikiPathways.
- **For questions** in drug discovery, **answers** from publications in peer reviewed scientific journals.

Top questions in the pharma industry I. (Open PHACTS)

Give me all oxidoreductase inhibitors active <100 nM in human and mouse

Given compound X, what is its predicted secondary pharmacology? What are the on- and off-target safety concerns for a compound? What is the evidence and how reliable is that evidence (journal impact factor, KOL) for findings associated with a compound?

Given a target, find me all actives against that target. Find/predict polypharmacology of actives. Determine ADMET profile of actives

For a given interaction profile – give me similar compounds

The current Factor Xa lead series is characterized by substructure X. Retrieve all bioactivity data in serine protease assays for molecules that contain substructure X

A project is considering protein kinase C alpha (PRKCA) as a target. What are all the compounds known to modulate the target directly? What are the compounds that could modulate the target directly? I.e. return all compounds active in assays where the resolution is at least at the level of the target family (i.e. PKC) from structured assay databases and the literature

Give me all active compounds on a given target with the relevant assay data

Identify all known protein–protein interaction inhibitors

For a given compound, give me the interaction profile with targets

For a given compound, summarize all 'similar compounds' and their activities

Retrieve all experimental and clinical data for a given list of compounds defined by their chemical structure (with options to match stereochemistry or not)

Top questions II. (OpenPHACTs)

For my given compound, which targets have been patented in the context of Alzheimer's disease?

Which ligands have been described for a particular target associated with transthyretin-related amyloidosis, what is their affinity for that target and how far are they advanced into preclinical/clinical phases, with links to publications/patents describing these interactions?

Target druggability: compounds directed against target X have been tested in which indications? Which new targets have appeared recently in the patent literature for a disease? Has the target been screened against in AZ before? What information on *in vitro* or *in vivo* screens has already been performed on a compound?

Which chemical series have been shown to be active against target X? Which new targets have been associated with disease Y? Which companies are working on target X or disease Y?

Which compounds are known to be activators of targets that relate to Parkinson's disease or Alzheimer's disease

For my specific target, which active compounds have been reported in the literature? What is also known about upstream and downstream targets?

Compounds that agonize targets in pathway X assayed in only functional assays with a potency $<1\ \mu M$

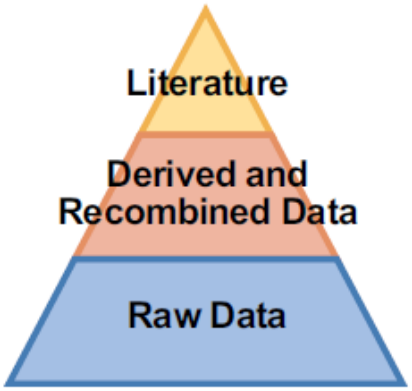
Give me the compound(s) that hit most specifically the multiple targets in a given pathway (disease)

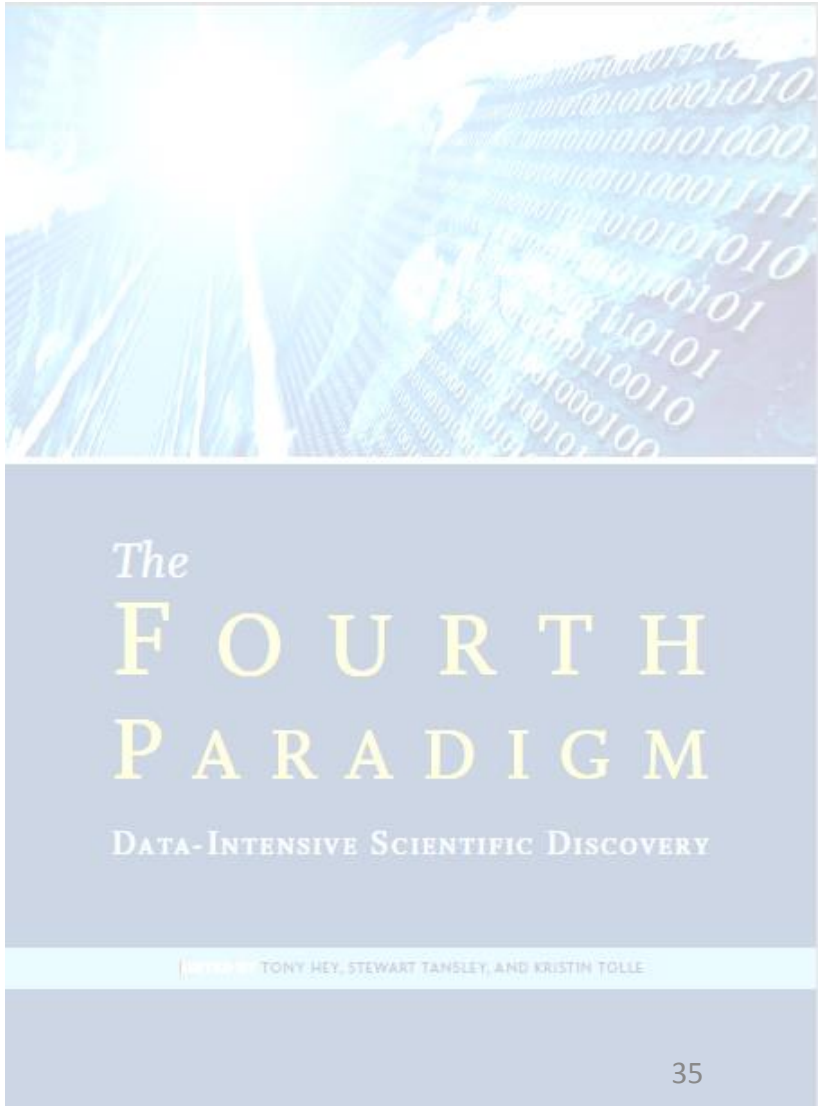
For a given disease/indication, give me all targets in the pathway and all active compounds hitting them

E-science, data-intensive science

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





„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
  - …

Overview

- Lectures
- Topics

Lectures

1. Sept 9th Introduction: what is AI
2. Decision theory and intelligent system design
3. 16th Sporting day (no lectures)
4. Problem solving with search: uninformed
5. 23rd Problem solving with search: informed, optimization
6. Problem solving with constraint based programming
7. 30th Search in game playing
8. Problem solving: exercises
9. 7th Logic: syntacs and semantics
10. Propositional logic
11. 14th Logic: first-order logic
12. Logic: exercises
13. 21st Earlier midterms: exercises
14. Day-off (Oct. 23)
15. Midterm exam
16. Discussion of midterm, Uncertainty
17. Probability theory
18. Nov. 6th : Probabilistic modeling, simple models: Markov chains, Naïve BNs, HMMs
19. Bayesian networks and inference (observational, interventional, counterfactual)
20. 13 Knowledge engineering of BNs: exercises, homework guide, applications
21. Inference in BNs: HMMs, poly-BNs, stochastic inference, Markov Chain Monte Carlo
22. 20. Decision theory, Inference with utilities and interventions, decision networks
23. Learning: induction/transduction/compression, PAC learning, sample complexity
24. Nov. 27.: Open university (no lectures)
25. Dec. 2.Repeated midterm exam
26. 4 Bayesian learning-approximations, Bayesian/MAP/ML decision tree learning
27. Overview – exercises for the final exam
28. Homework presentations – exercises for the final exam

midterm 28th of Oct.

repeated midterm 2nd of Dec.

Exams: 16th of December, 10am.
6th of January, 10am.
20th of January, 8.30am

Topics I.

1. Intro
    1. The four approaches to AI.
    2. The Turing test
    3. Acting rationally. The rational agent.
2. Agents
    1. Agent function, agent program, agent types/architectures.
    2. Environment properties: Observable, deterministic, static, single-agent.
    3. The reflex agent architecture, The utility-based agent architecture.
3. Problem-solving with search
    1. Problem types. The single-state problem formalization.
    2. The general tree search algorithm
    3. The four evaluation metric/properties for search strategies: completeness, space-complexity, time-complexity, optimality (branching factor, diameter of the state space).
    4. Uninformed search.
        a. Breadth-first (concept, pseudocode, properties), Depth-first (concept, pseudocode, properties), Iterative deepening (depth-limited depth-first) search (concept, pseudocode, properties), Comparison of properties.,
    5. Informed search
        a. Heuristic function
        b. Greedy search (concept, pseudocode, properties)
        c. A* (concept, pseudocode, properties) optimality with informal proof
4. Local search
    1. Applicability (when?)
    2. The hill-climbing algorithm (pseudocode)
    3. Problems with the hill-climbing algorithm
    4. Simulated annealing
5. Constraint satisfaction: heuristics
6. Game playing
    1. The game tree
    2. The MINIMAX algorithm
    3. Alpha-beta cuts
7. Logic
    1. The concept of general purpose inference and domain specific knowledge-base.
    2. Logic: syntax and semantics (conceptualization).
    3. The syntax of propositional logic.
    4. The concept of models wrt KBs and the model-based definition of semantic inference: entailment.
    5. (Syntactic) inference: elementary steps: modus ponens, resolution.
    6. Relation between entailment and (syntactic) inference: soundness, completeness.
    7. Definition of a Horn-clause
    8. The forward-chaining proof method
    9. The backward-chaining proof method
    10. Conversion of a KB to CNF form.
    11. The resolution-based proof method
    12. The first-order logic: Advantages, Quantifiers

Topics II.

1. Uncertainty
    1. The subjective interpretation of probability
    2. Decision theory: the binary decision problem (which action?)
    3. Probability theory
        a. Atomic events, composite events, joint distribution
        b. Conditional probability, the chain rule
        c. The Bayes rule,
            i. prior and posterior probabilities
            ii. relevance: causal and diagnostic direction
        d. Independence, conditional independence
    4. Inference by enumeration
    5. The naive Bayes model.
        a. The product form for the joint.
        b. Diagnostic inference
        c. The structure.
2. The Bayesian networks.
    1. Syntax.
    2. A complete example.
    3. Compactness (for binary random variables with max k parents).
    4. Global semantics (the product decomposition of the joint wrt the structure)
    5. Construction steps.
3. Inference in Bayesian networks.
    1. Tasks: simple query, composite query, relevance
    2. Inference by enumeration (pseudocode).
    3. Inference by stochastic simulation
        a. Sampling from an empty network (concept, pseudocode).
4. Temporal probability models
    1. Definition of a Markov process (homogeneous).
    2. Definition of a Hidden Markov model (homogeneous).
        a. Inference tasks: definitions of filtering, most likely explanation, smoothing.
            i. Filtering (concept, derivation, pseudocode)
    3. Connection between HMMs and Bayesian networks.
5. Decision theory
    1. Utility theory, preferences, the conditions for the existence of a utility function.
    2. The maximum expected utility principle.
    3. Decision network: elements and structure.
    4. Value of perfect information, formula

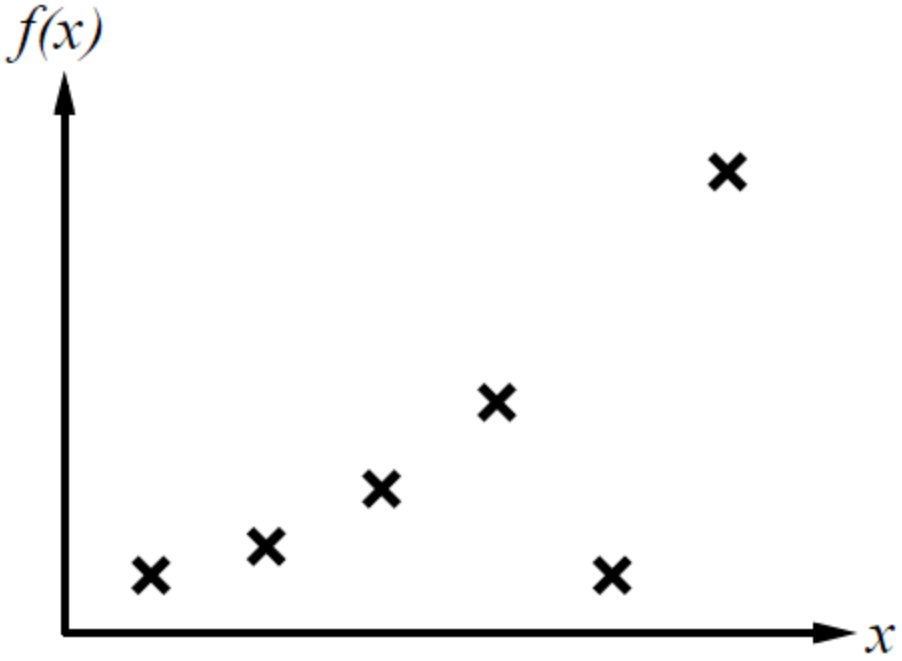
Topics III.

1. Learning
    1. The function approximation view of inductive learning.
    2. The Ockham principle
    3. Bayesian learning
        a. Bayes rule
        b. Posterior probability of a model/hypothesis
        c. Prediction using averaging, MAP and ML approximations.
2. decision theoretic foundation
    1. loss functions, error measures
    2. empirical vs expected loss: AUC
    3. asymptotic consistency
    4. rate of learning, speed of convergence
    5. The learning curve.
    6. The bias-variance dilemma
    7. Probably Approximately Correct (PAC) learning
        a. definition
        b. the misclassification rate as loss
        c. derivation of sample complexity of concept learning in i.i.d. context (independent identically distributed)
            i. within class
            ii. outside class
    8. concept learning methods
        a. version space
3. The decision tree representation.
    1. Expressivity
    2. Cardinality
    3. a learning method

Reminder: regression

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

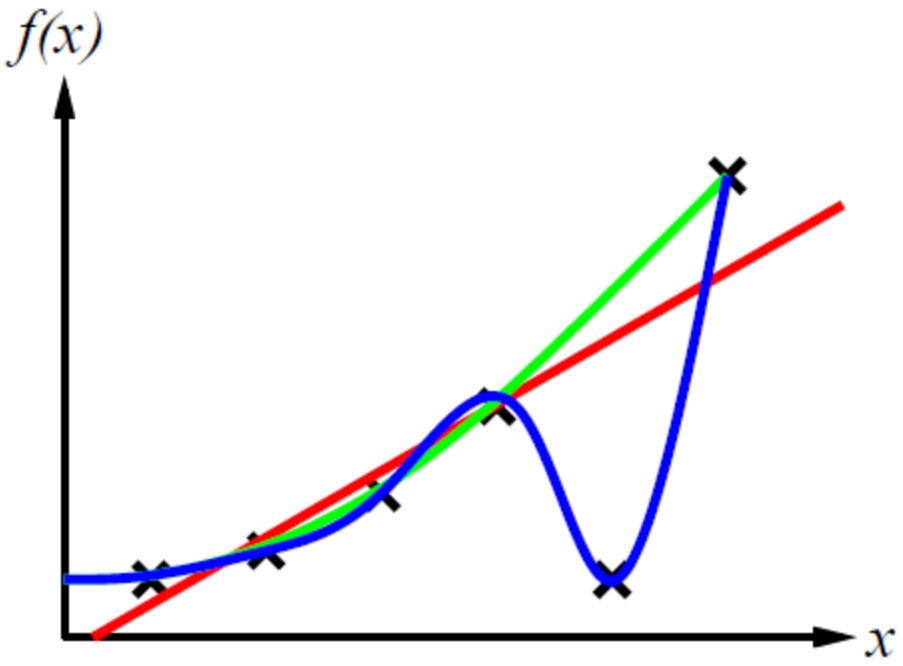
E.g., curve fitting:



Reminder: regression

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:



Reminder: concept learning/classification: learning decision trees (AIMA)

Problem: decide whether to wait for a table at a restaurant, based on the following attributes:

1. Alternate: is there an alternative restaurant nearby?
2. Bar: is there a comfortable bar area to wait in?
3. Fri/Sat: is today Friday or Saturday?
4. Hungry: are we hungry?
5. Patrons: number of people in the restaurant (None, Some, Full)
6. Price: price range ($, $$, $$$)
7. Raining: is it raining outside?
8. Reservation: have we made a reservation?
9. Type: kind of restaurant (French, Italian, Thai, Burger)
10. WaitEstimate: estimated waiting time (0-10, 10-30, 30-60, >60)

Attribute-based representations

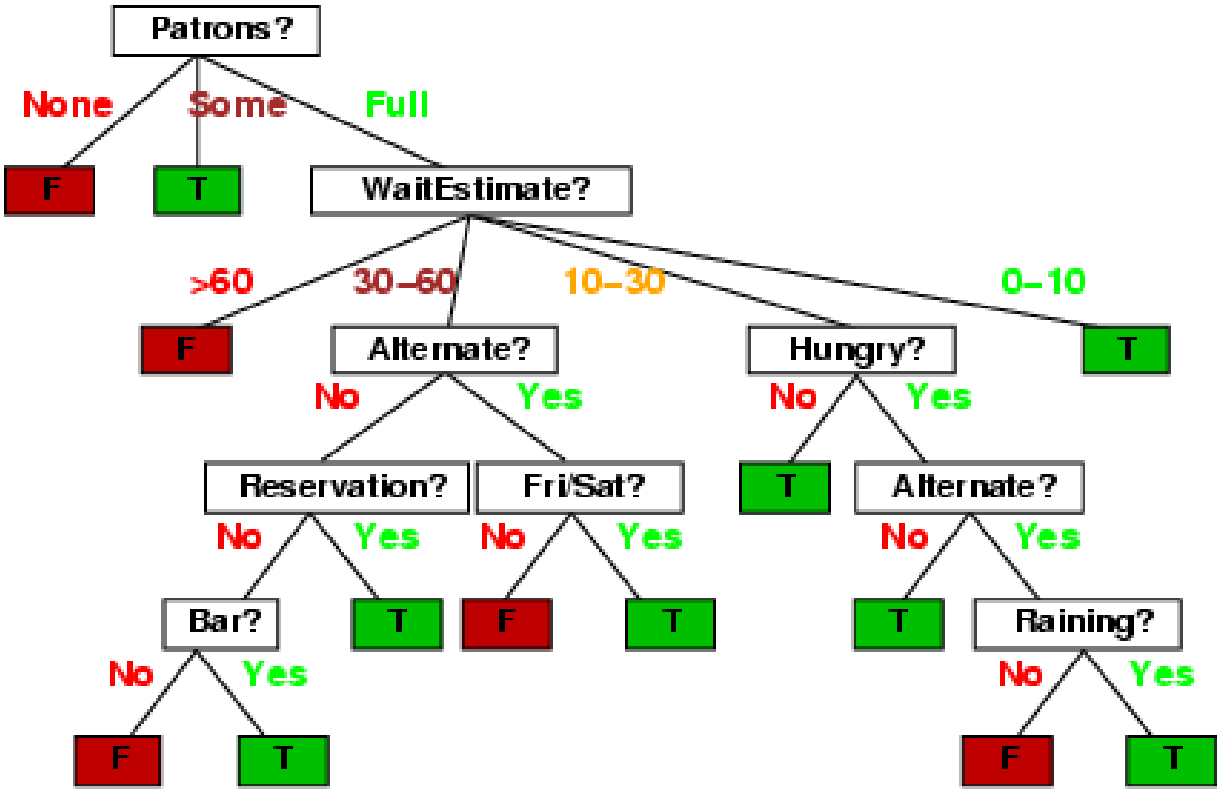
- Examples described by attribute values (Boolean, discrete, continuous)
- E.g., situations where I will/won't wait for a table:

Example	Attributes										Target
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*Wait*
X_1	T	F	F	T	Some	$$$	F	T	French	0–10	T
X_2	T	F	F	T	Full	$	F	F	Thai	30–60	F
X_3	F	T	F	F	Some	$	F	F	Burger	0–10	T
X_4	T	F	T	T	Full	$	F	F	Thai	10–30	T
X_5	T	F	T	F	Full	$$$	F	T	French	>60	F
X_6	F	T	F	T	Some	$$	T	T	Italian	0–10	T
X_7	F	T	F	F	None	$	T	F	Burger	0–10	F
X_8	F	F	F	T	Some	$$	T	T	Thai	0–10	T
X_9	F	T	T	F	Full	$	T	F	Burger	>60	F
X_{10}	T	T	T	T	Full	$$$	F	T	Italian	10–30	F
X_{11}	F	F	F	F	None	$	F	F	Thai	0–10	F
X_{12}	T	T	T	T	Full	$	F	F	Burger	30–60	T

- Classification of examples is positive (T) or negative (F)
-

Decision trees

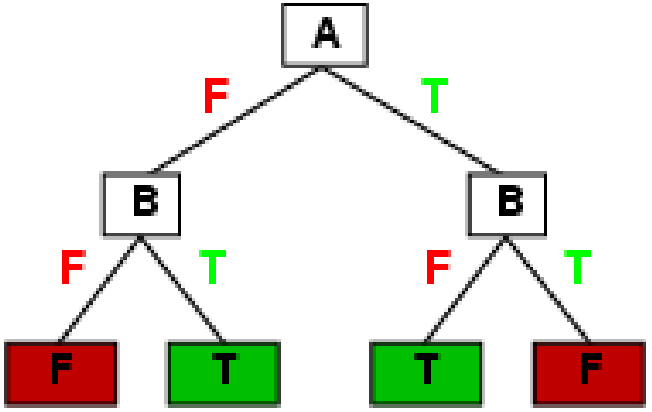
- One possible representation for hypotheses
- E.g., here is the "true" tree for deciding whether to wait:



Expressiveness

- Decision trees can express any function of the input attributes.
- E.g., for Boolean functions, truth table row → path to leaf:

A	B	A xor B
F	F	F
F	T	T
T	F	T
T	T	F



- Trivially, there is a consistent decision tree for any training set with one path to leaf for each example (unless *f* nondeterministic in *x*) but it probably won't generalize to new examples

- Prefer to find more compact decision trees

Hypothesis spaces

How many distinct decision trees with *n* Boolean attributes?

= number of Boolean functions

= number of distinct truth tables with 2^n rows = 2^{2^n}

- E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees

Hypothesis spaces

How many distinct decision trees with *n* Boolean attributes?

= number of Boolean functions

= number of distinct truth tables with 2^n rows = 2^{2^n}

- E.g., with 6 Boolean attributes, there are 18,446,744,073,709,551,616 trees

How many purely conjunctive hypotheses (e.g., $Hungry \wedge \neg Rain$)?

- Each attribute can be in (positive), in (negative), or out
  $\Rightarrow 3^n$ distinct conjunctive hypotheses
- More expressive hypothesis space
  - increases chance that target function can be expressed
  - increases number of hypotheses consistent with training set
    $\Rightarrow$ may get worse predictions

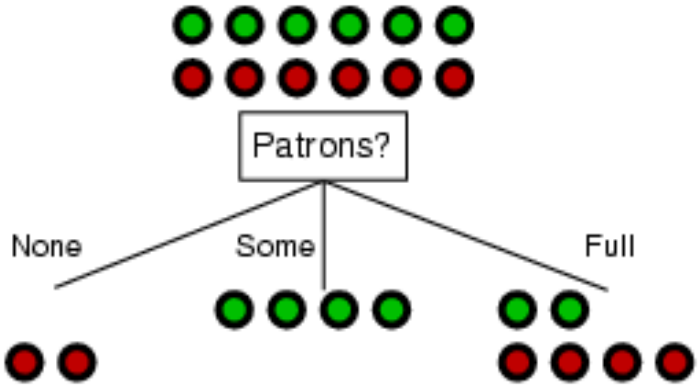
Decision tree learning

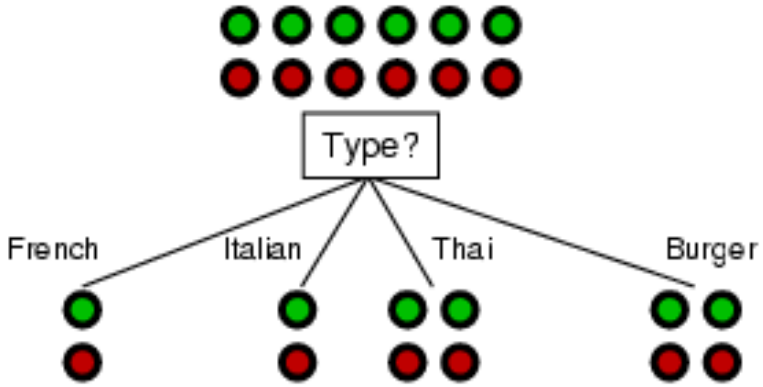
- Aim: find a small tree consistent with the training examples
- Idea: (recursively) choose "most significant" attribute as root of (sub)tree

```
function DTL(examples, attributes, default) returns a decision tree
    if examples is empty then return default
    else if all examples have the same classification then return the classification
    else if attributes is empty then return MODE(examples)
    else
        best ← CHOOSE-ATTRIBUTE(attributes, examples)
        tree ← a new decision tree with root test best
        for each value v_i of best do
            examples_i ← {elements of examples with best = v_i}
            subtree ← DTL(examples_i, attributes − best, MODE(examples))
            add a branch to tree with label v_i and subtree subtree
        return tree
```

Choosing an attribute

- Idea: a good attribute splits the examples into subsets that are (ideally) "all positive" or "all negative"





- *Patrons?* is a better choice

Using information theory

- To implement `Choose-Attribute` in the DTL algorithm
- Information Content (Entropy):

$$I(P(v_1), \ldots, P(v_n)) = \Sigma_{i=1} -P(v_i) \log_2 P(v_i)$$

- For a training set containing *p* positive examples and *n* negative examples:

$$I(\frac{p}{p+n}, \frac{n}{p+n}) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Information gain

- A chosen attribute *A* divides the training set *E* into subsets E_1, … , E_v according to their values for *A*, where *A* has v distinct values.

$$remainder(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p + n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

- Information Gain (IG) or reduction in entropy from the attribute test:

$$IG(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A)$$

- Choose the attribute with the largest IG

Information gain

For the training set, $p = n = 6$, $I(6/12, 6/12) = 1$ bit

Consider the attributes *Patrons* and *Type* (and others too):

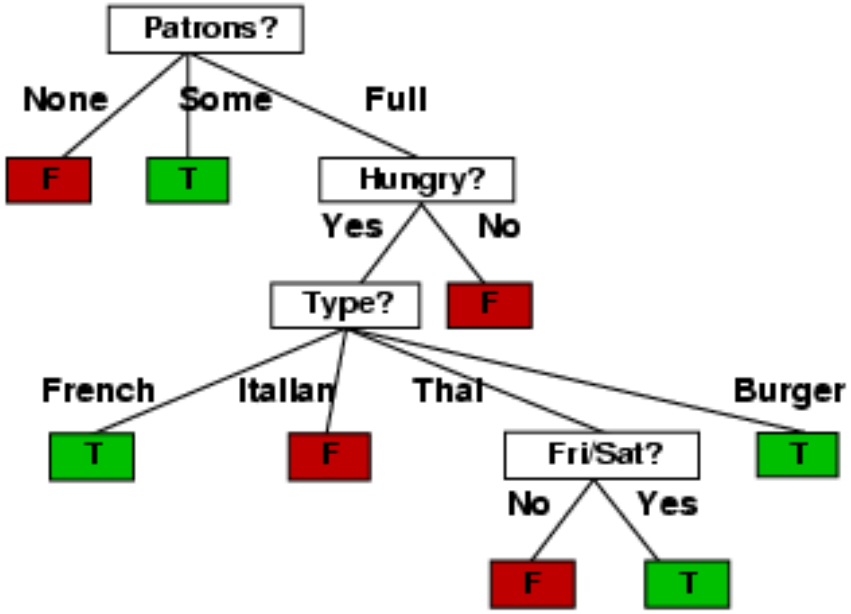
$$IG(Patrons) = 1 - [\frac{2}{12} I(0,1) + \frac{4}{12} I(1,0) + \frac{6}{12} I(\frac{2}{6}, \frac{4}{6})] = .0541 \text{ bits}$$

$$IG(Type) = 1 - [\frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{2}{12} I(\frac{1}{2}, \frac{1}{2}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4}) + \frac{4}{12} I(\frac{2}{4}, \frac{2}{4})] = 0 \text{ bits}$$

Patrons has the highest IG of all attributes and so is chosen by the DTL algorithm as the root

Example contd.

- Decision tree learned from the 12 examples:



- Substantially simpler than “true” tree---a more complex hypothesis isn’t justified by small amount of data