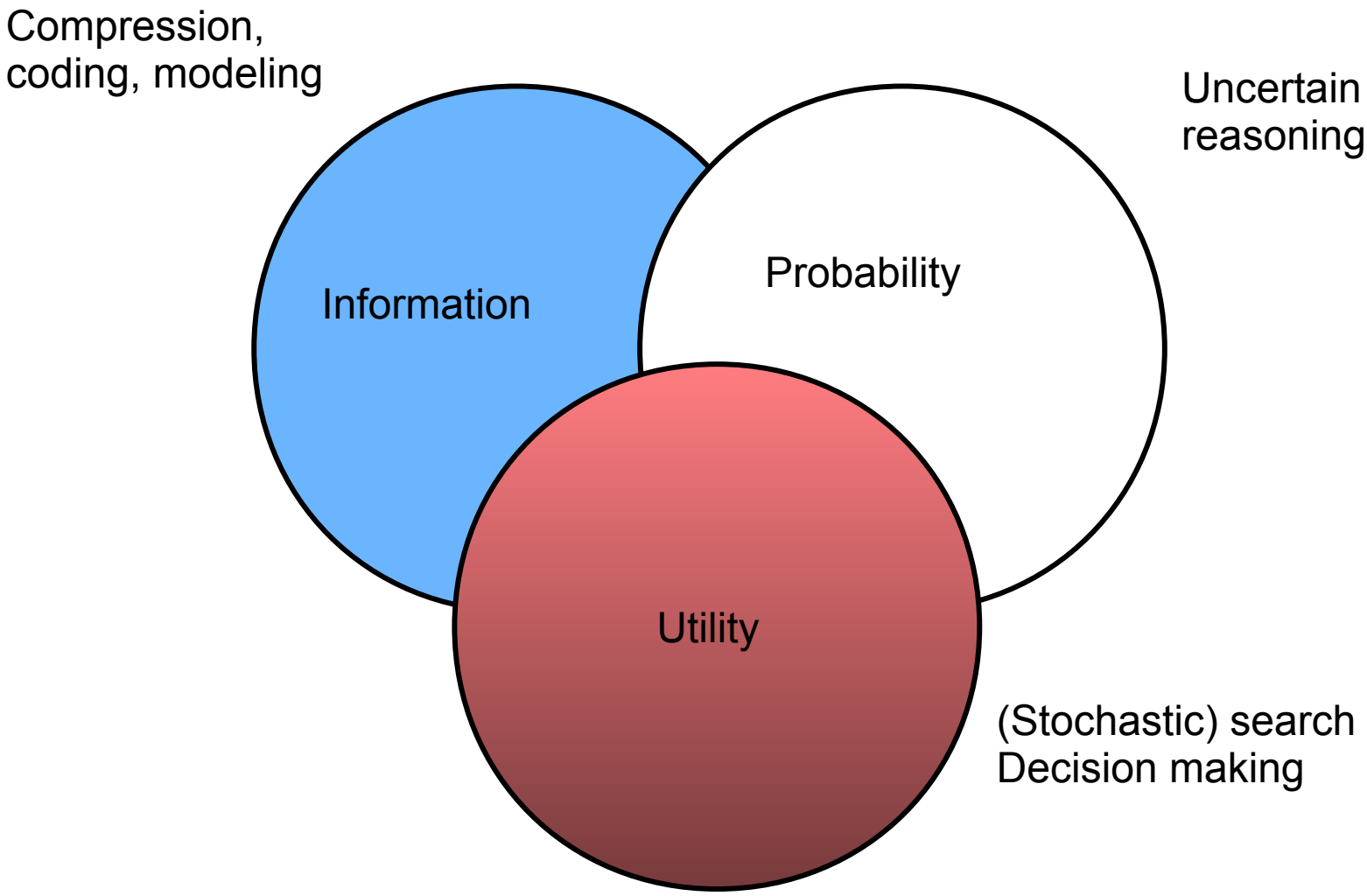
582636 Probabilistic Models

Spring 2011

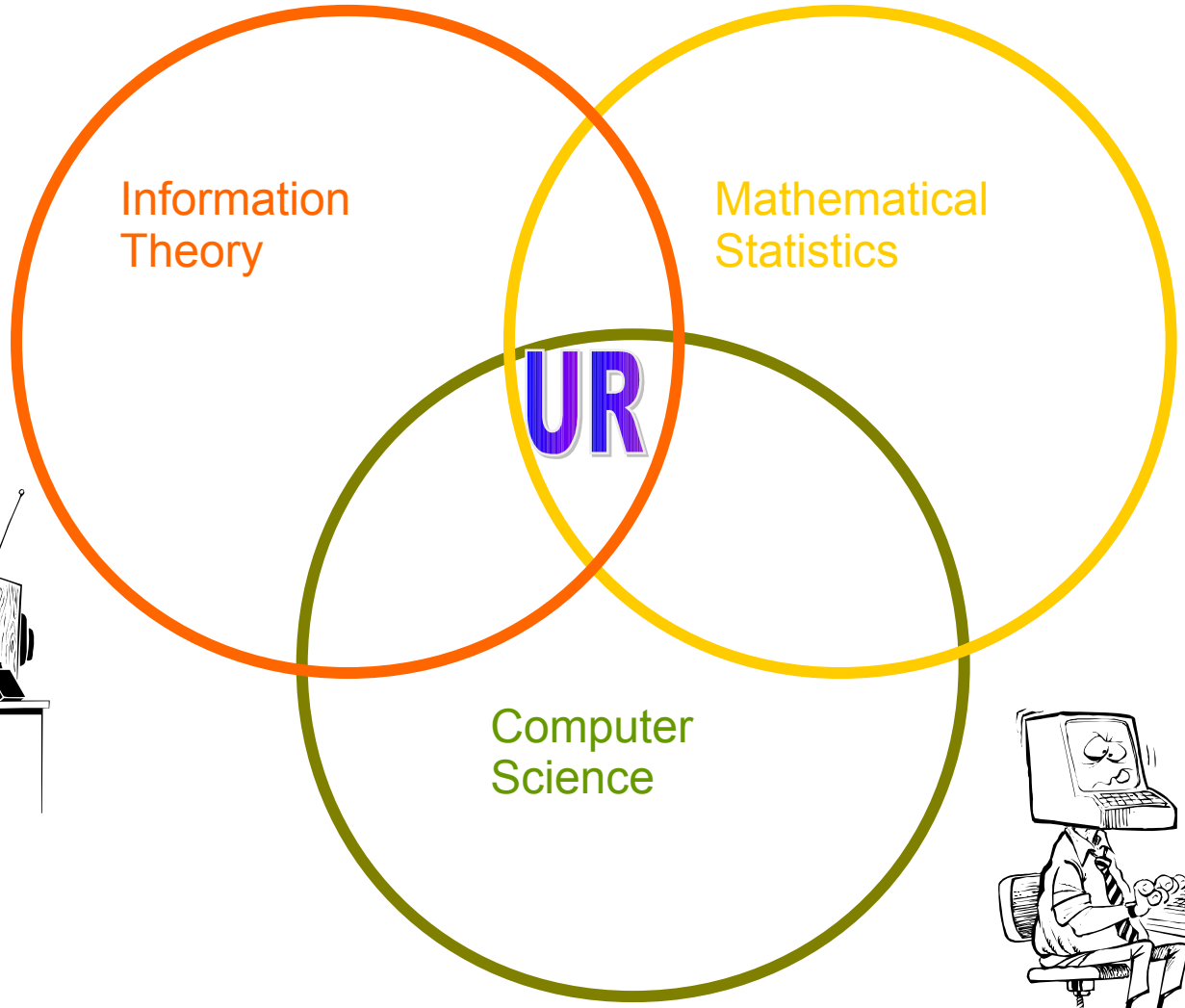
Petri Myllymäki
Department of Computer Science
University of Helsinki, Finland

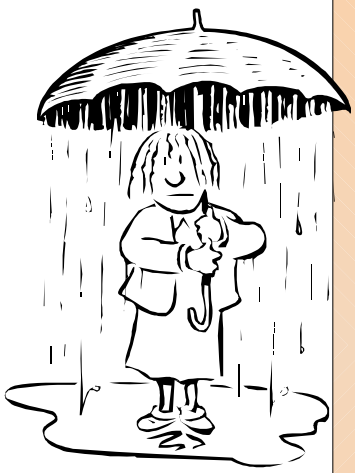
http://www.cs.helsinki.fi/courses/582636/2011/k/k/1

Three concepts



Our motivation: Uncertain reasoning





Yet another probability course?

- Computer science point of view
- Artificial Intelligence Point of View
  - Agent point of view
  - Knowledge representation (KR)
  - Reasoning
  - Rationality
  - Decision making
  - Grounding
- Machine learning point of view
  - Computational methods for data analysis
  - Large data bases

Reasoning under uncertainty

- The world is a very uncertain place
- Thirty years of Artificial Intelligence and Database research danced around this fact
- And then a few AI researchers decided to use some ideas from the eighteenth century
- *Uncertainty in Artificial Intelligence* conference series 1985-
- Probabilistic reasoning now mainstream AI

But first there was logic

- Historically too it was first - syllogisms
  - a model of rationality
- Certainty, correctness, modularity, monotonicity
- BUT limited applicability since

Agents almost never have access to the whole truth about their environment!

Acting by certain knowledge only?

- Is it enough to leave home 90 minutes before the flight departure?
- Anything can happen.
- How about X minutes before departure?
- Are you bound to stay home?

Qualification problem:
What are the things that have to be taken into account?

Knowledge representation in FOPL

- Let us try to use FOPL for dental diagnosis

$$\forall p\, Symptom(p, Toothache) \Rightarrow Disease(p, Cavity)$$

Wrong!

$$\forall p\, Symptom(p, Toothache) \Rightarrow$$
$$Disease(p, Cavity)$$
$$\vee\, Disease(p, GumDisease)$$
$$\vee\, Disease(p, Abscess) \ldots$$

Incomplete!

$$\forall p\, Disease(p, Cavity) \Rightarrow Symptom(p, Toothache)$$

Wrong again, need to add qualifications!

FOPL representation fails because

- Laziness
  - Its is too much work to list all the factors to ensure exceptionless rules
- Theoretical ignorance
  - We do not know all the factors that play role in the phenomenon
- Practical Ignorance
  - Even if we know all the factors in general, we do not know them for each particular case

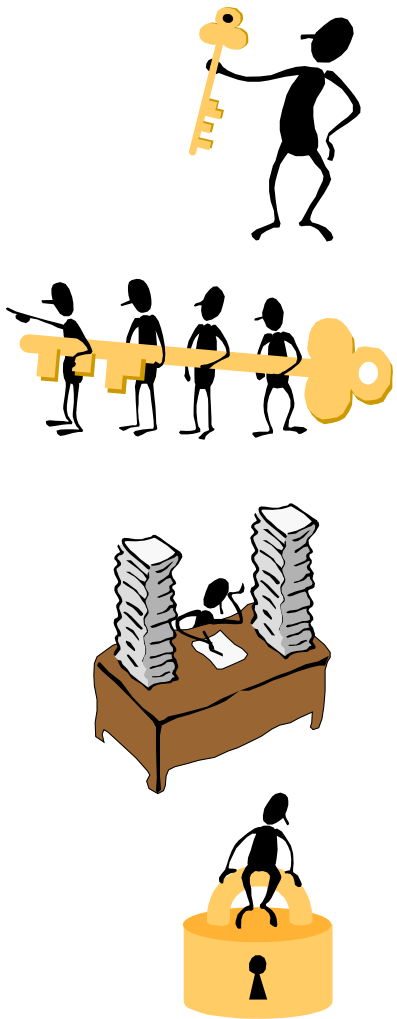
Probability to rescue

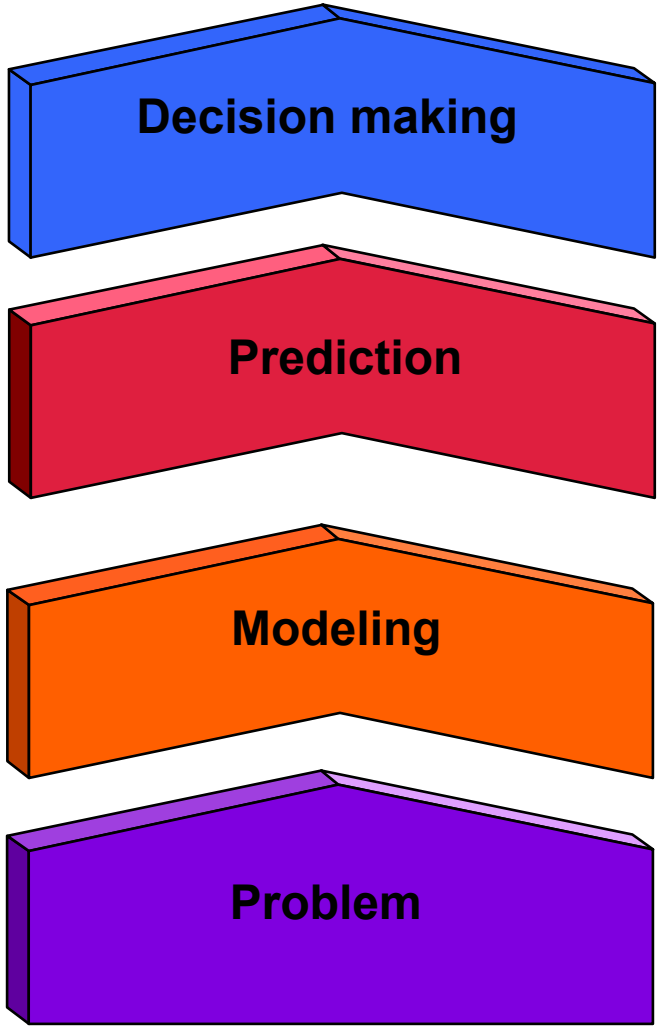
Probability provides a way to summarize the uncertainty that comes from our laziness and ignorance

On modeling

In building intelligent systems, in statistics and in the rest of the world ...

Modeling framework



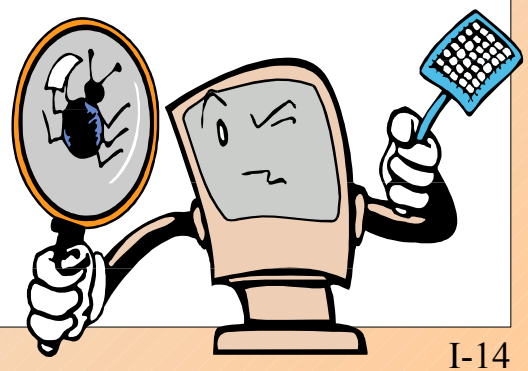


What does this mean?

- Problem: there is a need to model some part of the universe and make decisions based on the model
- Modeling: build the best model possible from a priori knowledge and data available
- Prediction: use the model to predict properties of interest
- Decision making: decide actions based on the predictions

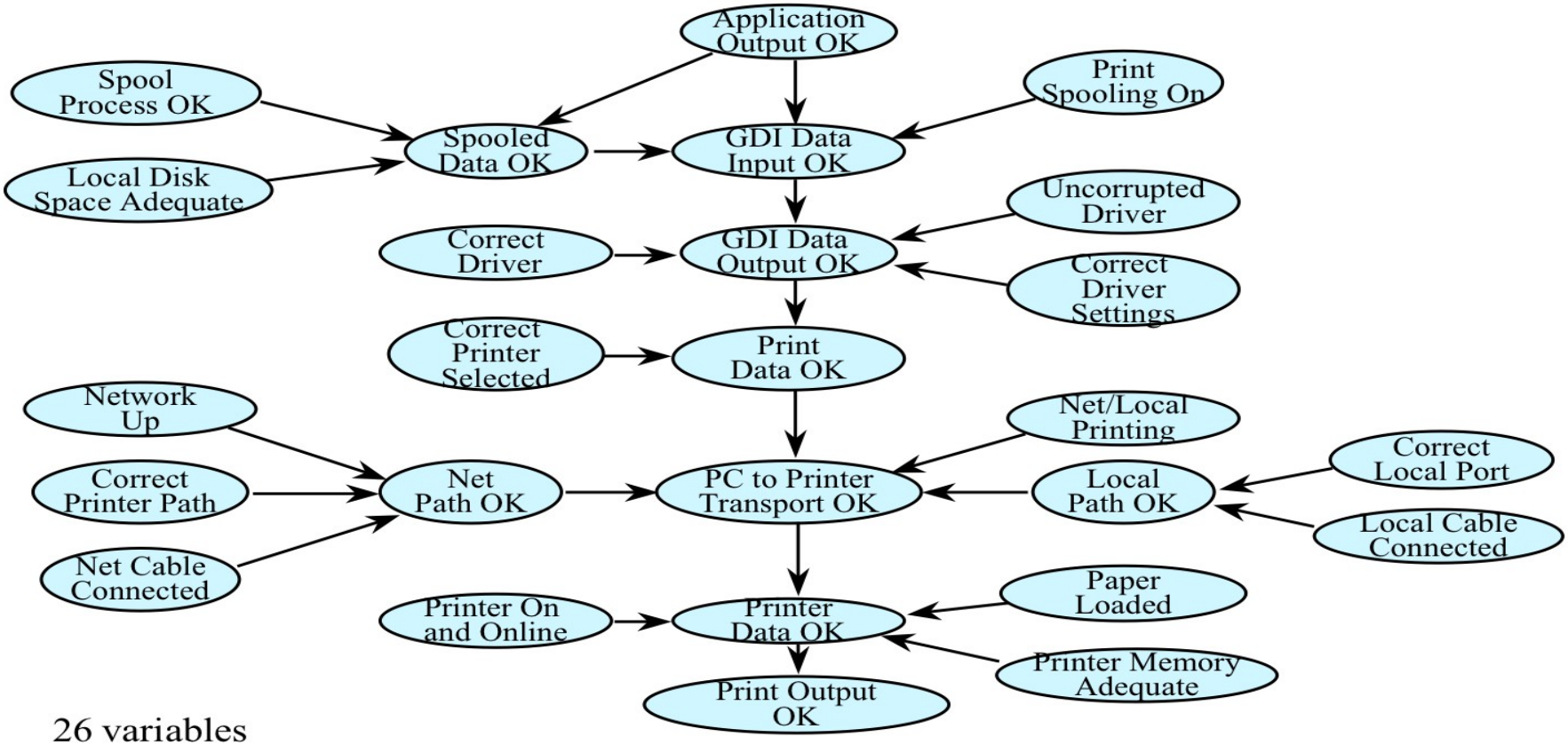
For example

- Problem: online troubleshooting of software/hardware
- Modeling: build a latent variable (Bayes) model of the problems user encounters based on knowledge about the software and symptom data
- Prediction: use the model to predict the underlying problem given symptoms
- Decision making: propose actions to remove the problem (or to find more symptoms)



Microsoft Technical support

Example: Printer Troubleshooter



26 variables

Instead of 2^{26} parameters we get

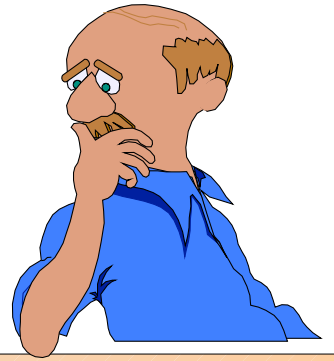
$99 = 17x1 + 1x2^1 + 2x2^2 + 3x2^3 + 3x2^4$

Bayesian email spam filters

- SpamBayes, OPFile, Outclass, bayespam, bogofilter, ifile, PASP, spamoracle, Spam Assassin Annoyance Filter, BSpam, Spam Bully, Death2Spam, InBoxer, …
- Software:
  - http://spambayes.sourceforge.net/related.html
- Background:
  - http://spambayes.sourceforge.net/background.html

Real questions are ...

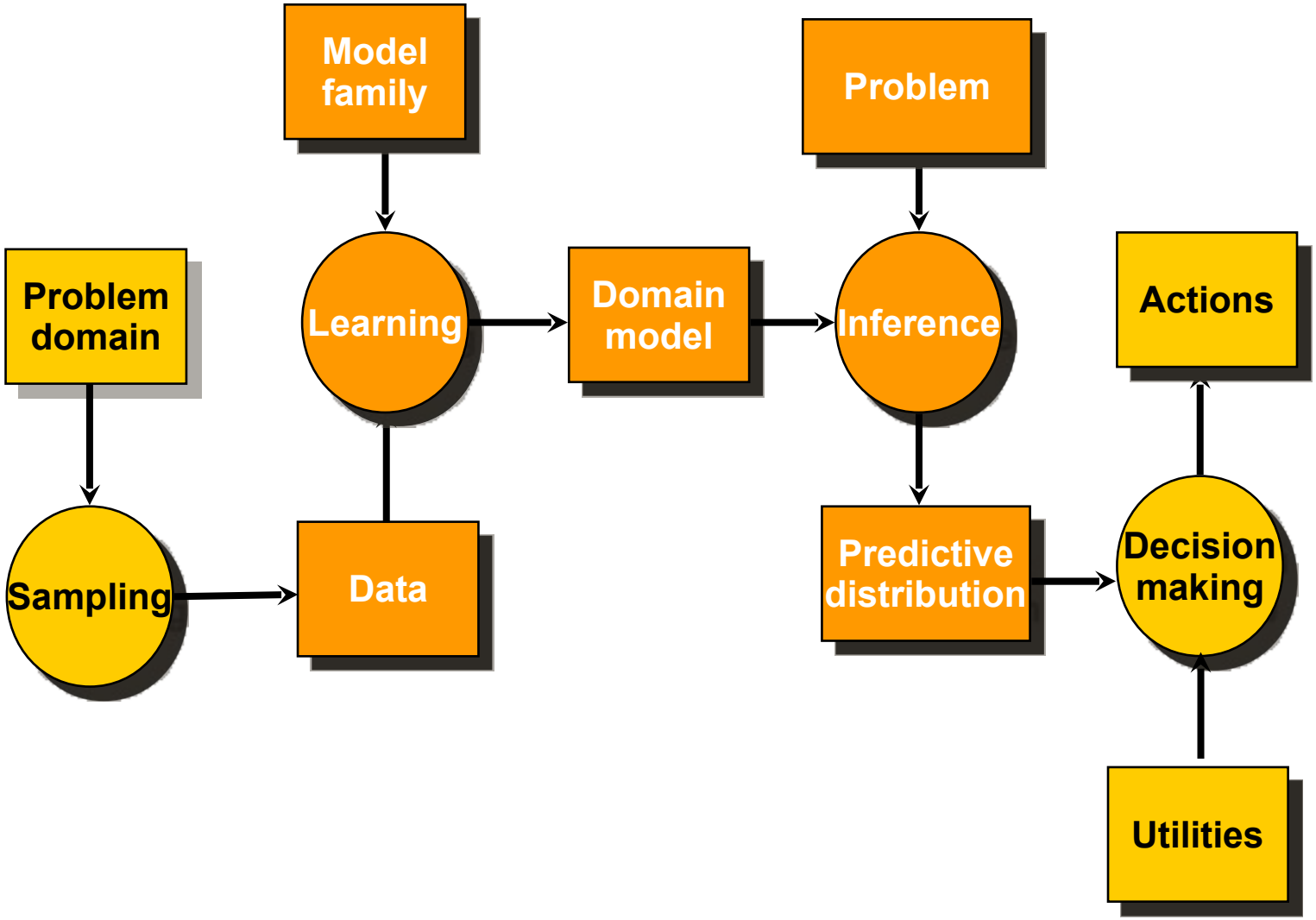
- Infinite number of models - what models do we consider?
  - Model is always chosen from a set of possible models!
- How do we compare models (i.e., measure that one model is better than another one) given some data?
- How do we find good models?

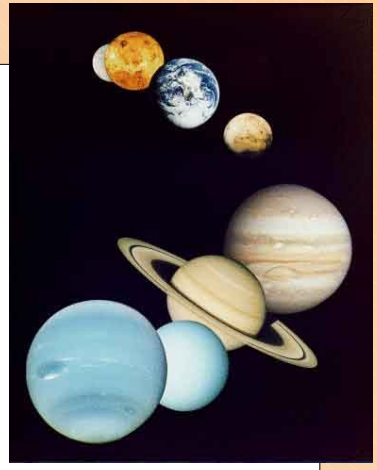


…and more

- How do we use the models to predict unobserved quantities of interest?
- What actions do we choose given the predictions?

General “rational agent” framework



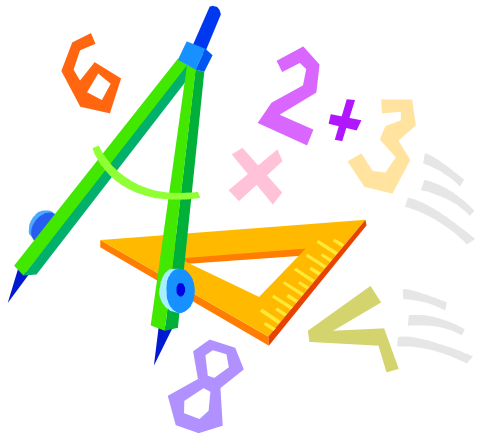


Choice of models

- Simple models vs. complex models
- Linear models vs. non-linear models
- Parametric models vs. non-parametric models
- Flat models vs. structural models
- What is complex is a totally nontrivial question
- One intuition: a complex model has more *effective* parameters

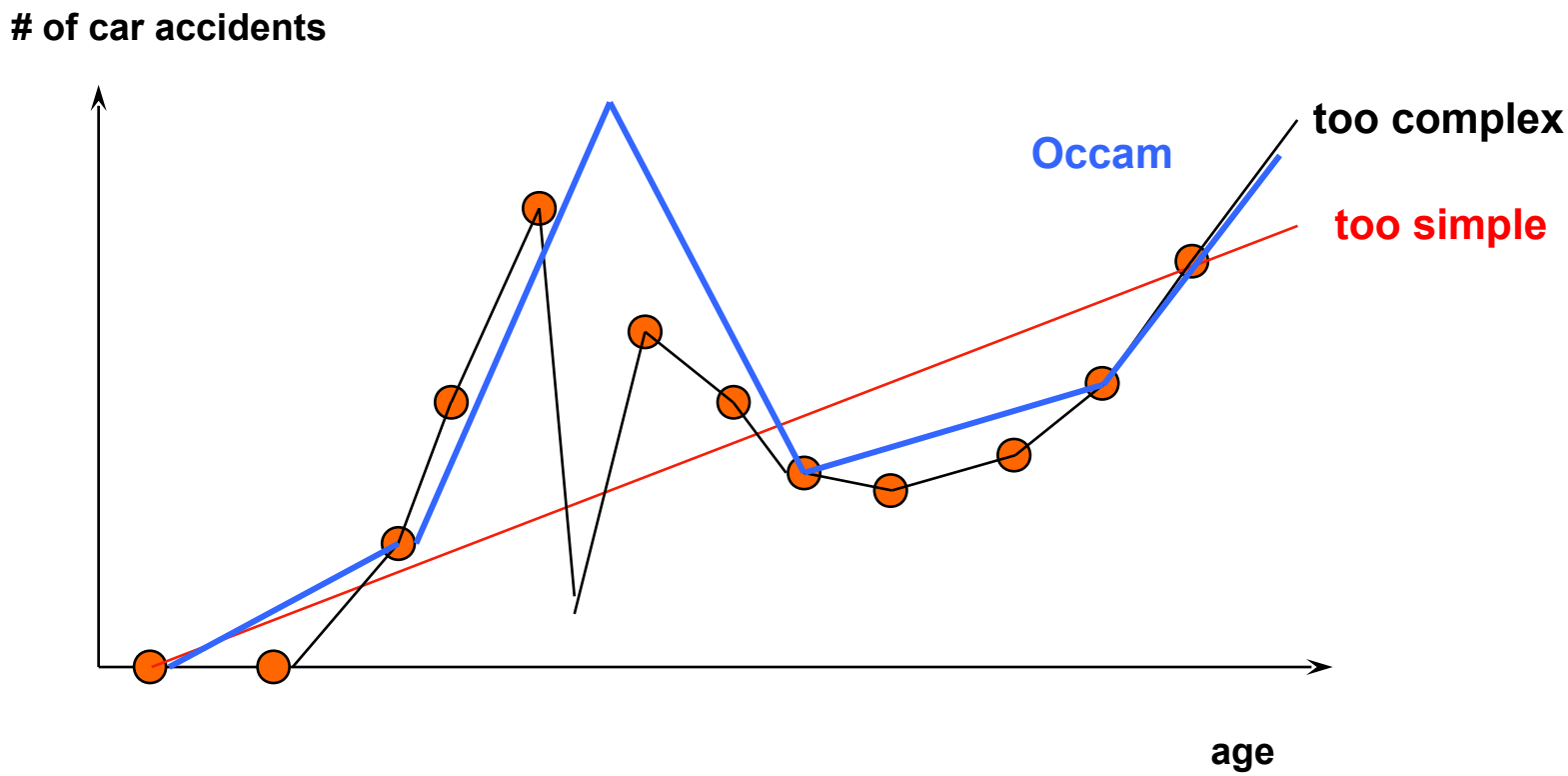
The Occam's razor principle

- The problem:
  - You are given the following sequence: -1, 3, 7, 11
  - Question: What are the two next numbers?
- Solution 1:
  - Answer: 15 and 19
  - Explanation: add 4 to the previous number
- Solution 2:
  - Answer: -19.9 and 1043.8
  - Explanation: if the previous number is x, the next one is $-x^3/11 + 9/11x^2 + 23/11$
- "Of two competing hypotheses both conforming to our observations, choose the simpler one."



Occam's Razor in Modeling

- there is a trade-off between the model complexity and fit to the data



Simpler models are better than complex models

- interpretation: they are easier to understand
- computation: predictions are typically easier to compute (not necessarily!)
- universality: they can be applied in more domains (more accurate predictions)
- “models should be only as complex as the data justifies”
- BUT: simpler models are NOT more probable a priori!
- Bayesian model selection: automatic Occam’s razor for model complexity regularization

Two types of modeling

- Descriptive models ("Statistical modeling")
  - describe objects (e.g., data) as they are
  - typically exploratory structures
- Predictive models ("Predictive inference")
  - models that are able to predict unknown objects (e.g., future data)
  - models of the underlying process

Some viewpoints

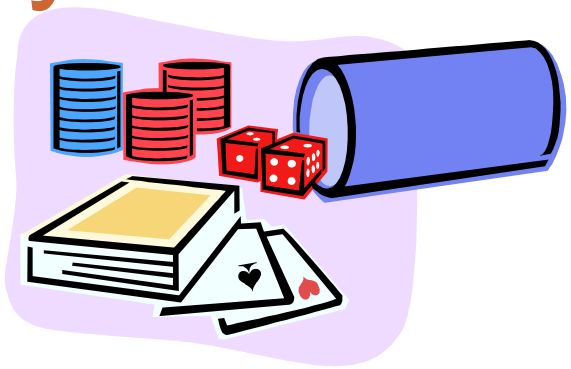
- "prediction is our business"
- why the best fit to data is not the best predictor
  - data can be erroneous - perfect fit is too "specialized" and models the errors also!
  - a sample can only "identify" up to a certain level of complexity
- intuitive goal: minimize model complexity + prediction error - it keeps you honest!

Alternatives

- Probabilistic inference
  - Statistical inference
  - Bayesian inference
- Fuzzy inference
- Dempster-Shafer inference
- Non-monotonic logic

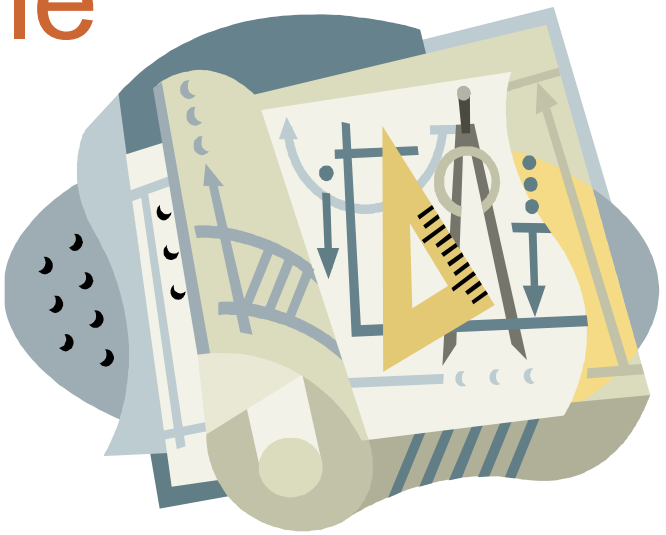
Bayesian inference: basic concepts

Some early history



- Bernoulli (1654-1705)
- Bayes (1701-1761)
- Laplace (1749-1827)
- Prediction problem ("forward probability"):
  - If the probability of an outcome in a single trial is *p*, what is the relative frequency of occurrence of this outcome in a series of trials?
- Learning problem ("inverse probability"):
  - Given a number of observations in a series of trials, what are the probabilities of the different possible outcomes?

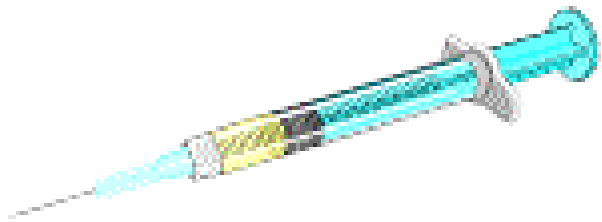
The Bayes rule



- Axioms of probability theory:
  - The sum rule:
    - $P(A \mid C) + P(\bar{A} \mid C) = 1$
  - The product rule:
    - $P(AB \mid C) = P(A \mid BC)\, P(B \mid C)$
- The Bayes rule:
  - $P(A \mid BC) = P(A \mid C)\, P(B \mid AC) / P(B \mid C)$
- A rule for updating our beliefs after obtaining new information
- H = hypothesis (model), I = background information, D = data (observations):
  - $P(H \mid D\, I) = P(H \mid I)\, P(D \mid H\, I) / P(D \mid I)$

Do I have a good test?

- A new home HIV test is assumed to have "95% sensitivity and 98% specificity"
- a population has HIV prevalence of 1/1000. If you use the test, what is the chance that someone testing positive actually has HIV?



Test continued ...

- P(HIV + | test HIV +) = ?
- We know that
  - P(test HIV + | HIV +) = .95
  - P(test HIV + | HIV -) = .02
- from Bayes we have learned that we can calculate the probability of having HIV given a positive test result by

$$\frac{P(\text{test HIV +}|\text{HIV +})P(\text{HIV +})}{P(\text{test HIV +}|\text{HIV +})P(\text{HIV +})+P(\text{test HIV +}|\text{HIV -})P(\text{HIV -})}$$

$$=\frac{0.95 \text{ x } 0.001}{0.95 \text{ x } 0.001 + 0.02 \text{ x } 0.99}=0.045$$

Thus finally

- thus over 95% of those testing positive will, in fact, not have HIV
- the right question is:

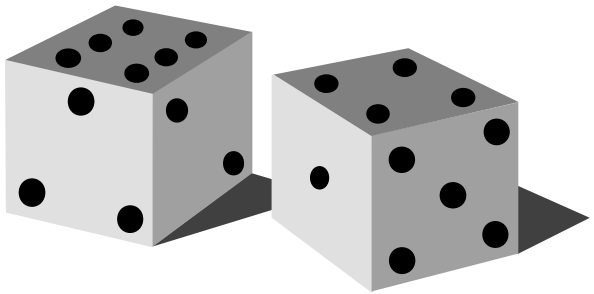
How should the test result change our belief that we are HIV positive?

Bayesian?

- Probabilities can be interpreted in various ways:
  - Frequentist interpretation (Fisher,Neyman, Cramer)
  - "Degree of belief" interpretation (Bernoulli, Bayes, Laplace, Jeffreys, Lindley, Jaynes)

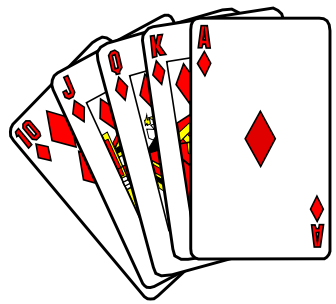
Frequentist says ...

- The long-run frequency of an event is the proportion of the time it occurs in a long sequence of trials - probability is this frequency
- probability can only be attached to “random variables” - not to individual events



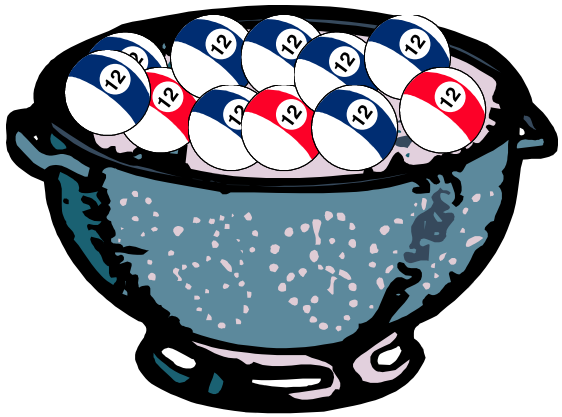
Bayesian says ...

- an event x = state of some part of the universe
- probability of x is the degree of belief that event x will occur
- probability will always depend on the state of knowledge
- p(x|y,C) means probability of event x given that event y is true and background knowledge C



Frequentist language for solving problems

- P(data | model)
- sampling distributions



Model

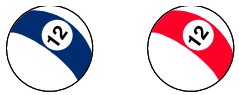
Data

?

Bayesian language for solving problems

- Bayesian: P(data | model) & P(model | data)

Prior knowledge



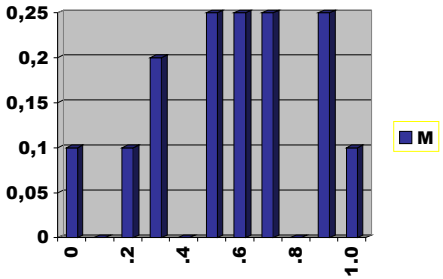
Data

?

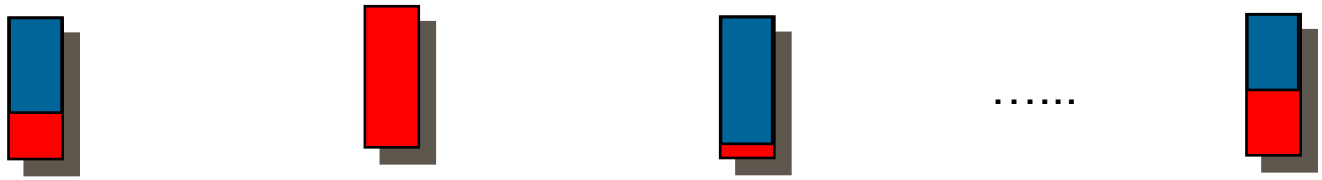
Isn't this what I already do? No.

Hypothesis testing

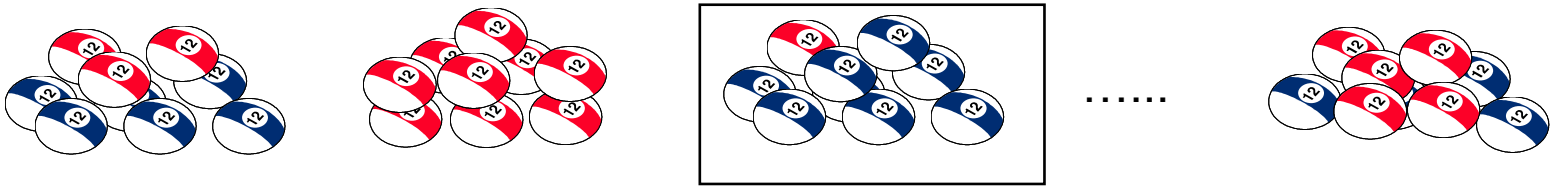
"Sampling distribution of the estimator"



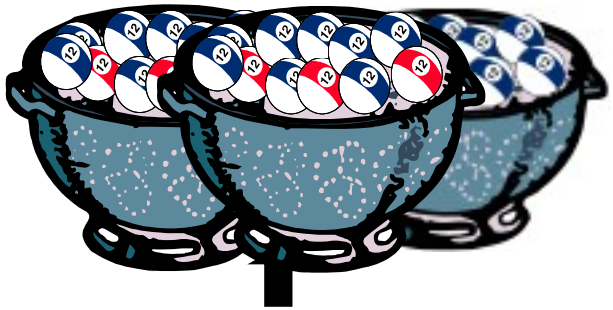
Estimator (function of data)

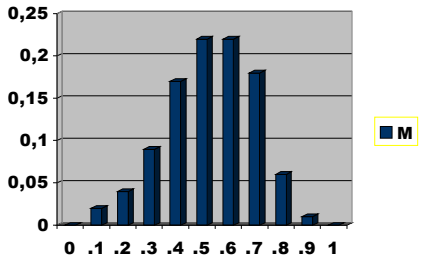


Data



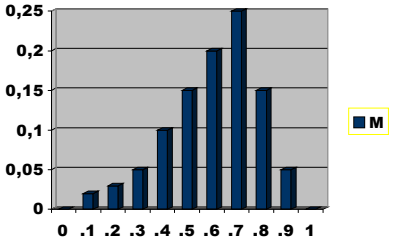
"The Bayesian way"



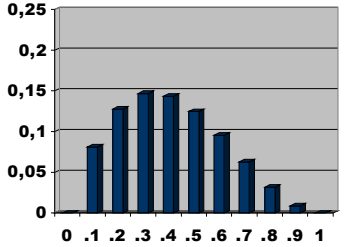


Posterior distribution of the models

Likelihood

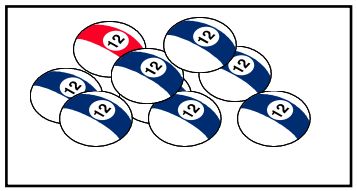


X



Prior distribution of the models

Data

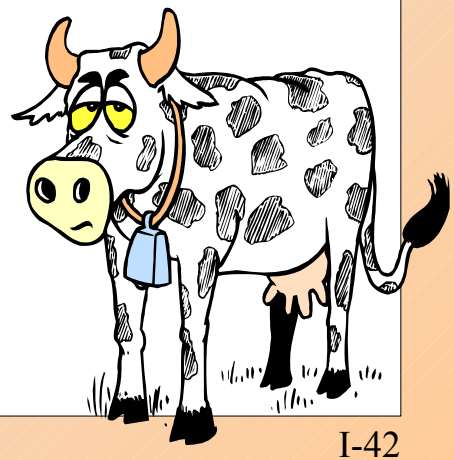


Reasons for using probability theory

- Cox/Jaynes argument: probability is an appealing choice as the language for plausible inference
- Berger argument: Decision theory offers a theoretical framework for optimal decision making, and decision theory needs probabilities
- Pragmatic argument: it is a very general framework and it works

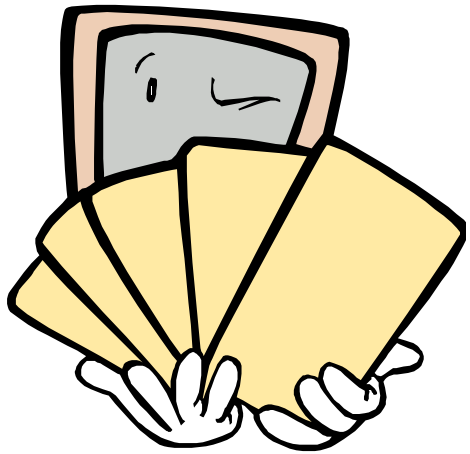
On plausible reasoning

- "*The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, non of which (fortunately) we have to reason on. Therefore the true logic for this world is the calculus of Probabilities, which takes account of the magnitude of the probability which is, or ought to be, in a reasonable man's mind*" (James Clerk Maxwell)
- Probabilistic reasoning is intuitively easy to understand, but on the other hand intuition may be a poor guide when facing probabilistic evidence
- "*Inside every non-Bayesian there is a Bayesian struggling to get out*" (Dennis V. Lindley)



Real questions

- Q1: Given plausibilities Plaus(A) and Plaus(B), what is Plaus(AB)?
- Q2: How is Plaus(~A) related to Plaus(A)?



Qualitative properties of p.r.

- D1. Degrees of plausibility are represented by real numbers
- D2. Direction of inference has a qualitative correspondence with common sense
  - For example: if Plaus(A | C´) > Plaus(A | C) and Plaus(B |C') = Plaus(B | C), then Plaus(AB | C') > Plaus(AB | C)
  - Ensures consistency in the limit (with perfect certainty) with deductive logic
- D3. If a conclusion can be inferred in more than one way, every possible way should lead to the same result
- D4. All relevant information is always taken into account
- D5. Equivalent states of knowledge must be represented by equivalent plausibility assignments

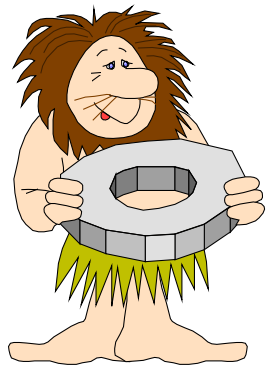
Cox/Jaynes/Cheeseman argument

- Every allowed extension of Aristotelian logic to plausibility theory is isomorphic to Bayesian probability theory
- Product rule (answers question Q1)
  - $P(AB \mid C) = P(A \mid BC)\, P(B \mid C)$
- Sum rule (answers question Q2)
  - $P(A \mid C) + P(\bar{A} \mid C) = 1$

Bayesian inference: How to update beliefs?

- Select the model space
- Use Bayes theorem to obtain the posterior probability of models (given data)

$$P(\text{Model}|\text{Data}) = \frac{P(\text{Data}|\text{Model})P(\text{Model})}{P(\text{Data})}$$



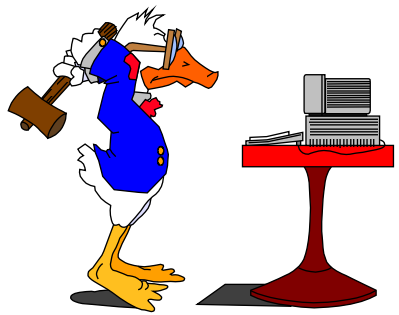
Posterior distribution is “the result” of the inference; what one needs from the posterior depends on what decisions are to be made

The Bayesian modeling viewpoint

- Explicitly include prediction (and intervention) in modeling

Models are a means (a language) to describe interesting properties of the phenomenon to be studied, but they are not intrinsic to the phenomenon itself.

"All models are false, but some are useful."



(Being predictive …)

- True prediction performance is a function of future data, not a model fit to current data

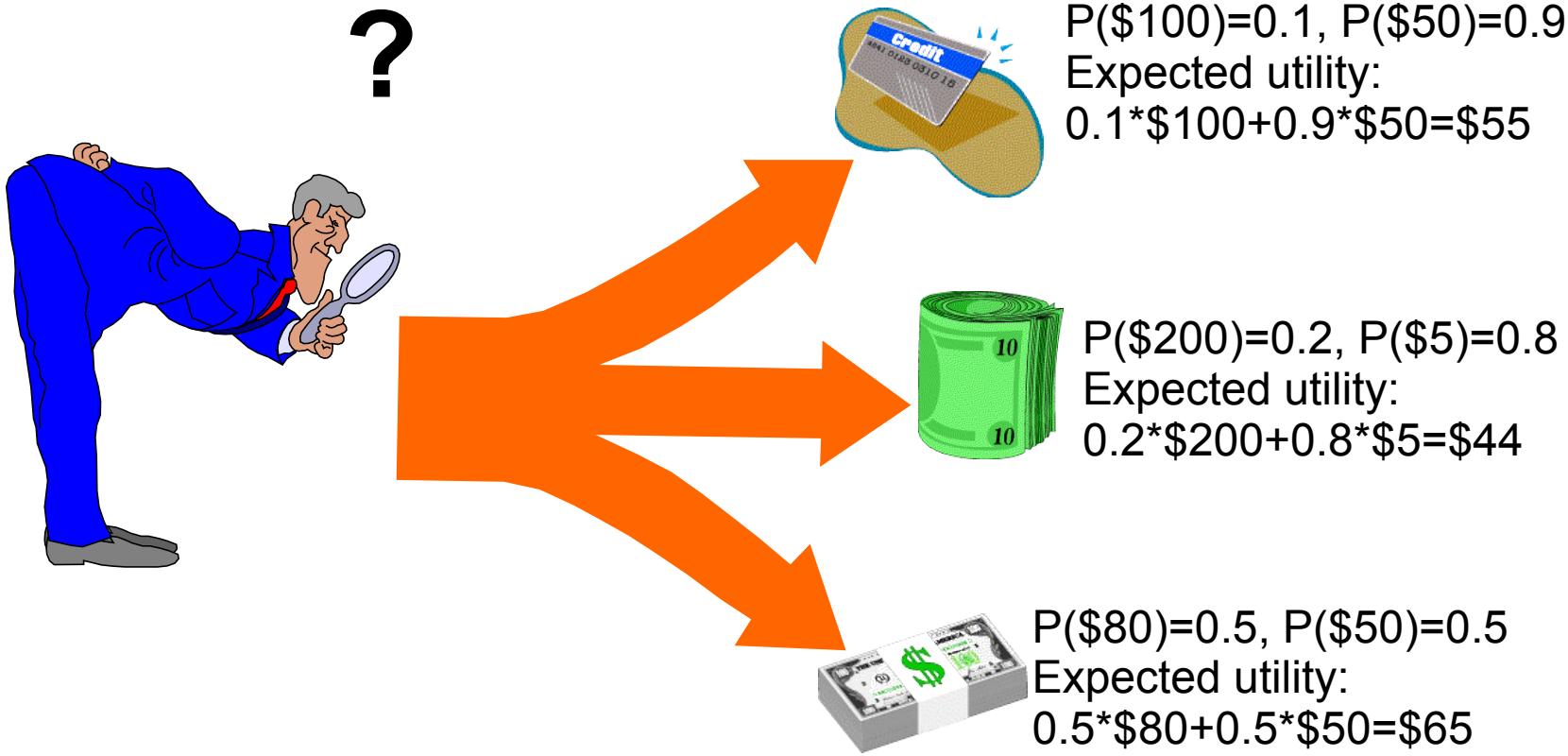
Good *predictive* models describe useful regularities of the data generating mechanism, while models that give a high probability to the observed *data* have possibly only learnt to memorize it.

Bayesian decision making for kids

- assign a benefit for every possible outcome (for every possible decision)
- assign a probability to every possible outcome given every possible decision
- what is the best decision?

Decision theory argument

- Decision theory offers a theoretical framework for optimal decision making

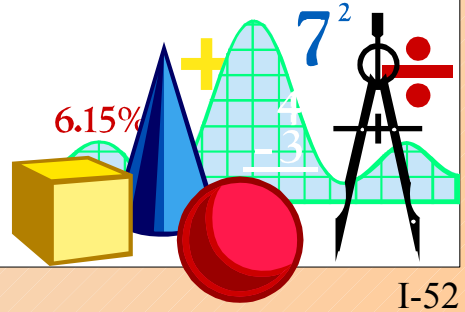


Optimal actions

- Optimal policy: choose the action with **maximal expected utility**
- The Dutch book argument: betting agencies **must be** Bayesians
- Where to get the utilities? (decision theory)

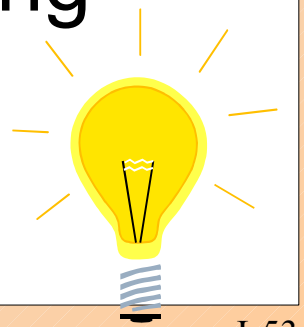
"Pragmatic" reasons for using probability theory

- The predictor and predicted variables (the inference task) do not have to be determined in advance
  - probabilistic models can be used for solving both classification (discriminative tasks), and configuration problems and prediction (regression problems)
  - predictions can also be used as a criteria for Data mining (explorative structures)



More pragmatic reasons for using probability theory

- consistent calculus
  - creating a consistent calculus for uncertain inference is not easy (the Cox theorem)
  - cf. fuzzy logic
- Probabilistic models can handle both discrete and continuous variables at the same time
- Various approaches for handling missing data (both in model building and in reasoning)



Nice theory, but...

- “isn’t probabilistic reasoning counter-intuitive, something totally different from human reasoning?”
- Cause for confusion: the old frequentist interpretation. But probabilities do NOT have to be thought of as frequencies, but as measures of belief
- The so called paradoxes are often misleading
  - A: P(€1.000.000)=1.0
  - B: P(€1.000.000)=0.25, P(€4.000.000)=0.25, P(€0)=0.5
- Even if that were true, maybe that would be a good thing!

Nice theory, but...

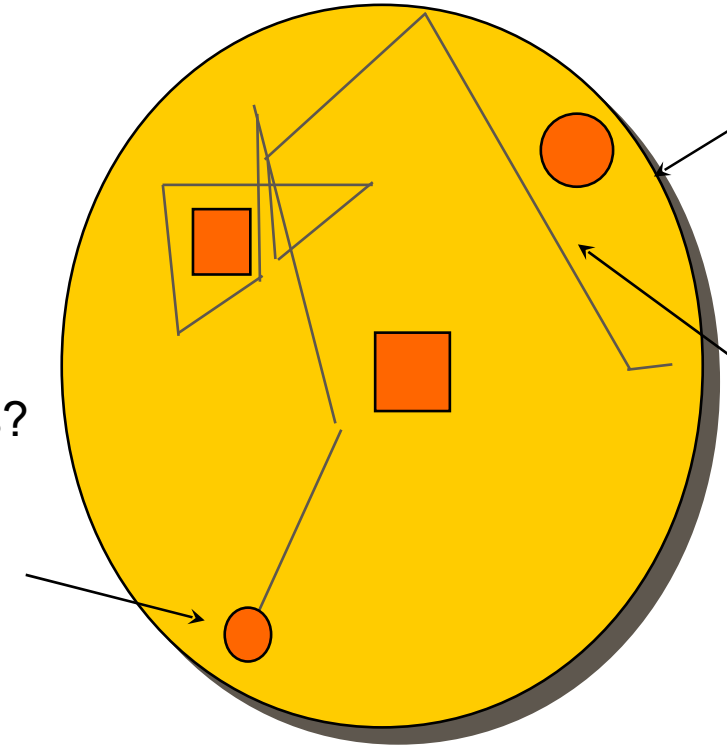
- "Where do all the numbers come from?"
  - Bayesian networks: small number of parameters
  - the numbers do not have to be accurate
  - probability theory offers a framework for constructing models from sample data, from domain knowledge, or from their combination

We can learn from Bayesians :-)

"I rest my case"

- Bayesian approaches never overfit (in principle)
- Bayesian approaches infer only from observed data (not possible data)
- Bayesian inference is always relative to a model family
- Does all this semi-philosophical debate really matter in practice?
  - **YES**!!
  - see e.g. "The great health hoax" by Robert Matthews, The Sunday Telegraph, September 13, 1998, or "Why Most Published Research Findings are False" by John Ioannidis, PLOS Medicine 2 (2005) 8.

Fundamental questions



What is the model space?

How do we compare models?

How do we search?

Bayesian answers

- Model family (space) is made explicit
- Comparison criteria is a probability
- No restrictions on the search algorithm

Classical statistics answers

- Model family is implicit (normal distributions)
- Comparison criteria is fit to data, deviation from "random" behavior, "model index"
- Simple deterministic "greedy" algorithms

Bayesian inference: basic operations

Probability of propositions

- Notation P(x) : read “probability of “x-pression”
- Expressions are statements about the contents of random variables
- Random variables are very much like variables in computer programming languages.
  - Boolean; statements, propositions
  - Enumerated, discrete; small set of possible values
  - Integers or natural numbers; idealized to infinity
  - Floating point (continuous); real numbers to ease calculations

Elementary “probositions”

- P(X=x)
  - probability that random variable X has value x
    - we like to use words starting with capital letters to denote random variables
- For example:
  - P(It_will_snow_tomorrow = true)
  - P(The_weekday_I'll_graduate = sunday)
  - P(Number_of_planets_around_Gliese_581 = 7)
  - P(The_average_height_of_adult Finns = 1702mm)

Semantics of P(X=x)=p

- So what does it mean?
  - P(The_weekday_I'll_graduate = sunday)=0.20
  - P(Number_of_planets_around_Gliese_581 = 7)=0.3
- Bayesian interpretation:
  - The proposition is either true or false, nothing in between, but we may be unsure about the truth. Probabilities measure that uncertainty.
  - The greater the p, the more we believe that X=x:
    - P(X=x) = 1 : Agent totally believes that X = x.
    - P(X=x) = 0 : Agent does not believe that X=x at all.

Compound “probositions”

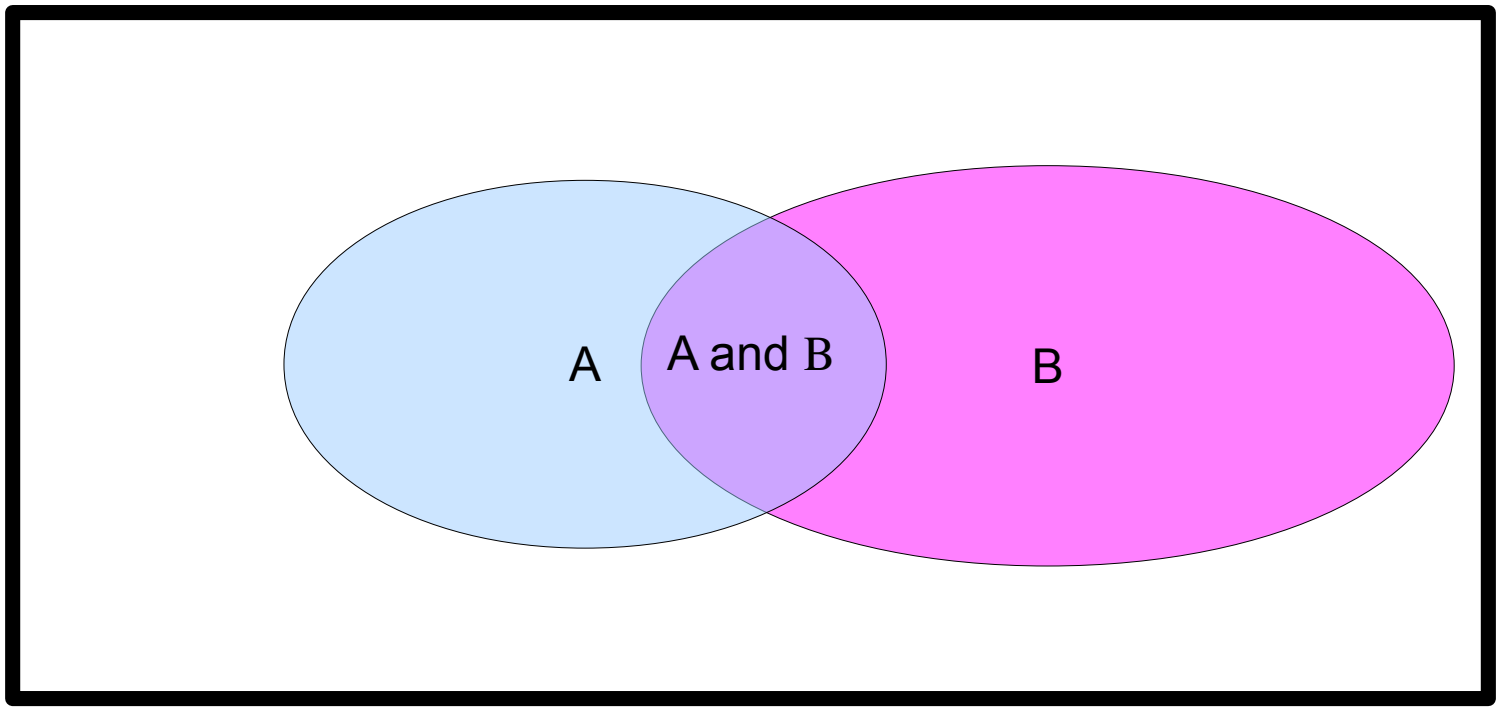
- Elementary propositions can be combined using logical operators ∧, ∨ and ¬.
  - like P(X=x ∧ ¬ Y=y) etc.
  - Possible shorthand: P(X ϵ S)
  - P(X≤x) for continuous variables
- Operator is the most common one, and often replaced by just comma like : P(A=a, B=b).
- Naturally other logical operators can also be defined as derivatives.

Axioms of probability

- Kolmogorov's axioms:
  1. $0 \leq P(x) \leq 1$
  2. $P(true) = 1, P(false)=0$
  3. $P(x \vee y) = P(x) + P(y) - P(x \wedge y)$
  - Some extra technical axioms needed to make theory rigorous
- Axioms can also be derived from common sense requirements (Cox/Jaynes argument)

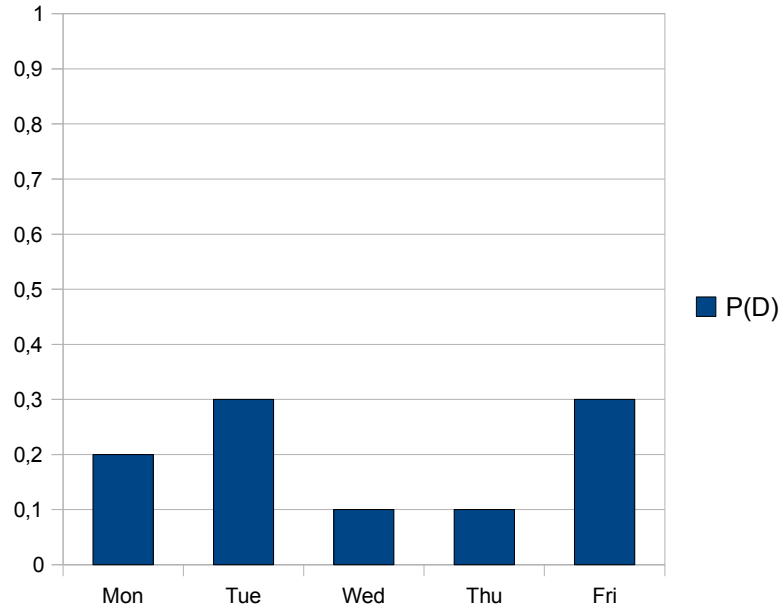
Axiom 3 again

- P(x or y) = P(x) + P(y) – P(x and y)
- It is there to avoid double counting:
- P(“day_is_sunday” or “day_is_in_July”) = 1/7 + 31/365 - 4/31.



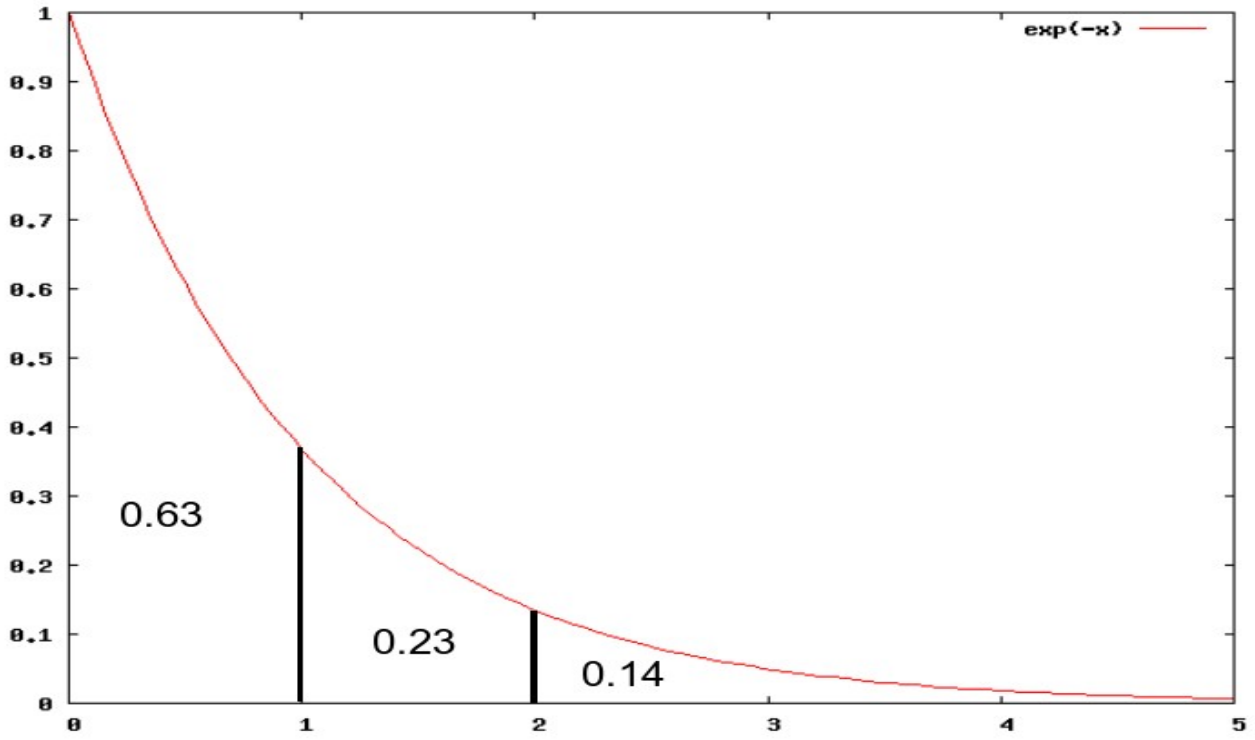
Discrete probability distribution

- Instead of stating that
  - $P(D=d_1)=p_1$,
  - $P(D=d_2)=p_2$,
  - ... and
  - $P(D=d_n)=p_n$
- we often compactly say
  - $P(D)=(p_1,p_2, ..., p_n)$.
- P(D) is called a probability distribution of D.
  - NB! $p_1 + p_2 + ... + p_n = 1$.



Continuous probability distribution

- In continuous case, the area under P(X=x) must equal one. For example P(X=x) = exp(-x):



Main toolbox of the Bayesians

- Definition of conditional probability
- Chain rule
- The Bayes rule
- Marginalization

- NB. These are all direct derivates of the axioms of probability theory

Conditional probability

- Let us define a notation for the probability of x given that we know (for sure) that y; and we know nothing else:

$$P(x|y) = \frac{P(x \wedge y)}{P(y)}$$

- Bayesians say that all probabilities are conditional since they are relative to the agent's knowledge K.

$$P(x|y, K) = \frac{P(x \wedge y|K)}{P(y|K)}$$

  - But Bayesians are lazy too, so they often drop K.
  - Notice that P(x,y) = P(x|y)P(y) is also very useful!

Chain rule

- From the definition of conditional probability, we get:

$$P(X_1, X_2) = P(X_2|X_1)P(X_1)$$

- And more generally:

$$P(X_1, \ldots, X_n) = \prod_i P(X_1)P(X_2|X_1)\ldots P(X_n|X_1, X_{2,}\ldots, X_{n-1})$$

Marginalization

- Let us assume we have a joint probability distribution for a set S of random variables.
- Let us further assume S1 and S2 partitions the set S.
- Now $$P(S_1 = s_1) = \sum_{s \in dom(S_2)} P(S_1 = s_1, S_2 = s)$$
$$= \sum_{s \in dom(S_2)} P(S_1 = s_1 | S_2 = s) P(S_2 = s),$$

  where s_1 and s are vectors of possible value combination of S1 and S2 respectively.

Joint probability distribution

- P(Toothache=x,Catch=y,Cavity=z) for all combinations of truth values (x,y,z):

Toothache	Catch	Cavity	probability
true	true	true	0,108
true	true	false	0,016
true	false	true	0,012
true	false	false	0,064
false	true	true	0,072
false	true	false	0,144
false	false	true	0,008
false	false	false	0,576
			1,000

- You may also think this as a P(Too_Cat_Cav=x), where x is a 3-dimensional vector of truth values.
- Generalizes naturally to any set of discrete variables, not only Booleans.

Joys of joint probability distribution

- Summing the condition matching numbers from the joint probability table you can calculate probability of any subset of events.
- P(Cavity=true or Toothache=true):

Toothache	Catch	Cavity	probability
true	true	true	0,108
true	true	false	0,016
true	false	true	0,012
true	false	false	0,064
false	true	true	0,072
false	true	false	0,144
false	false	true	0,008
false	false	false	0,576
			0,280

Marginal probabilities are probabilities too

- P(Cavity=x, Toothache=y)

Toothache	Catch	Cavity	probability
true	true	true	0,108
true	true	false	0,016
true	false	true	0,012
true	false	false	0,064
false	true	true	0,072
false	true	false	0,144
false	false	true	0,008
false	false	false	0,576
			1,000

- Probabilities of the lines with equal values for marginal variables are simply summed.

Conditioning

- Marginalization can be used to calculate conditional probability:

$$P(Cavity=t|Toothache=t)=\frac{P(Cavity=t\wedge Toothache=t)}{P(Toothache=t)}$$

Toothache	Catch	Cavity	probability
true	true	true	0,108
true	true	false	0,016
true	false	true	0,012
true	false	false	0,064
false	true	true	0,072
false	true	false	0,144
false	false	true	0,008
false	false	false	0,576
			1,000

$$\frac{0.108+0.012}{0.108+0.016+0.012+0.064}=0.6$$

Conditioning via marginalization

$$P(X|Y)$$

(definition)
$$= \frac{P(X,Y)}{P(Y)}$$

(marginalization)
$$= \frac{\sum_Z P(X,Z,Y)}{P(Y)}$$

(chain rule)
$$= \frac{\sum_Z P(X|Z,Y)P(Z|Y)P(Y)}{P(Y)}$$

$$= \sum_Z P(X|Z,Y)P(Z|Y).$$

Bayes' rule

Combining

$$P(x|y,K)=\frac{P(x\wedge y|K)}{P(y|K)}$$

$$P(x\wedge y|K)=P(y\wedge x|K)=P(y|x,K)P(x|K)$$

yields the famous Bayes formula

$$P(x|y,K)=\frac{P(x|K)P(y|x,K)}{P(y|K)}$$

or $$P(h|e)=\frac{P(h)P(e|h)}{P(e)}$$

Bayes formula as an update rule

- Prior belief P(h) is updated to posterior belief $P(h|e_1)$. This, in turn, gets updated to $P(h|e_1,e_2)$ using the very same formula with $P(h|e_1)$ as a prior. Finally, denoting $P(\cdot|e_1)$ with P_1 we get

$$P(h|e_1,e_2)=\frac{P(h,e_1,e_2)}{P(e_1,e_2)}$$

$$=\frac{P(h,e_1)P(e_2|h,e_1)}{P(e_1)P(e_2|e_1)}$$

$$=\frac{P(h|e_1)P(e_2|h,e_1)}{P(e_2|e_1)}=\frac{P_1(h)P_1(e_2|h)}{P_1(e_2)}$$

Bayes formula for diagnostics

- Bayes formula can be used to calculate the probabilities of possible causes for observed symptoms.

$$P(cause|symptoms)=\frac{P(cause)P(symptoms|cause)}{P(symptoms)}$$

- Causal probabilities P(symptoms|cause) are usually easier for experts to estimate than diagnostic probabilities P(cause|symptoms).

Bayes formula for model selection

- Bayes formula can be used to calculate the probabilities of hypotheses, given observations

$$P(H_1|D) = \frac{P(H_1)P(D|H_1)}{P(D)}$$

$$P(H_2|D) = \frac{P(H_2)P(D|H_2)}{P(D)}$$

...

General recipe for Bayesian inference

- X: something you don't know and need to know
- Y: the things you know
- Z: the things you don't know and don't need to know
- Compute:

$$P(X|Y) = \sum_Z P(X|Z,Y)\,P(Z|Y)$$

- That's it - we're done.