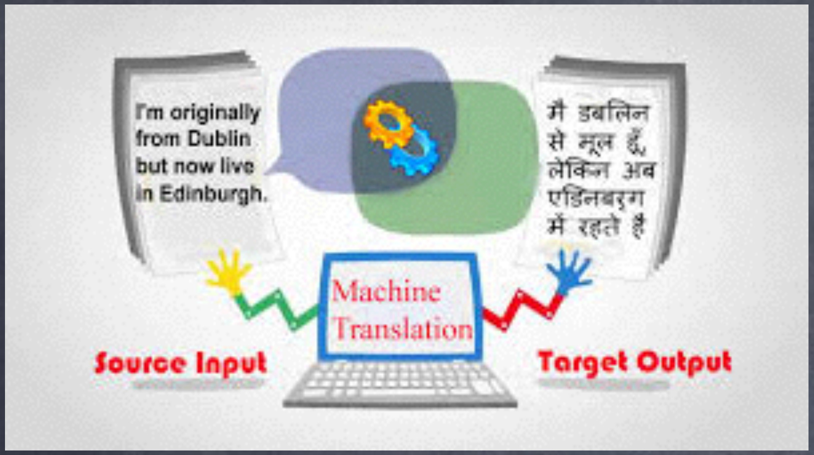
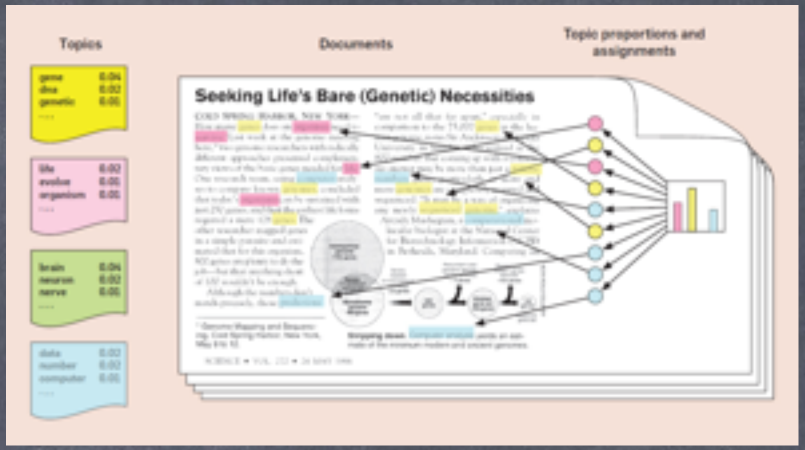
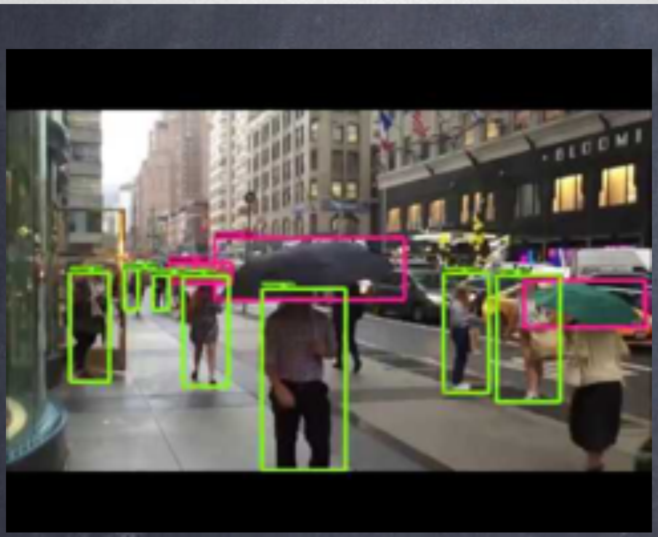
Probability in data science

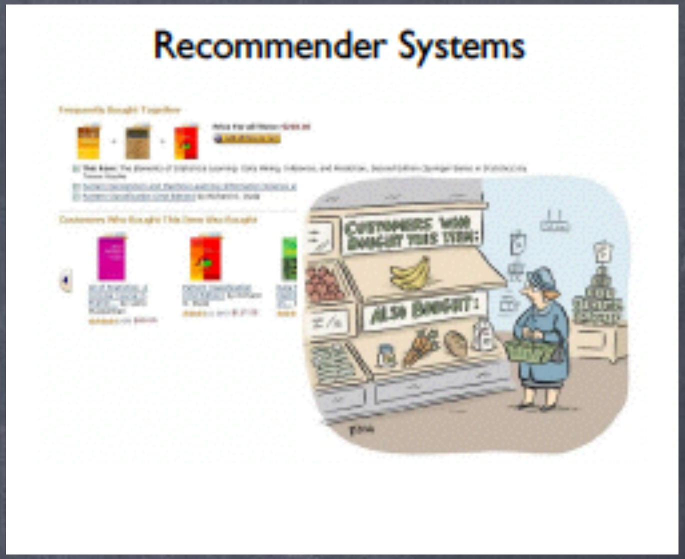
data Science


I'm originally from Dublin but now live in Edinburgh.
मै डबलिन से मूल हूँ, लेकिन अब एडिनबर्ग मे रहते है
Machine Translation
Source Input
Target Output





Internet of Things
Open Data
Smart Agriculture
Smart Retail
Smart Home
Smart Mobility
Education
SMART CITY
Smart Health
Smart Grid/ Smart Energy
Smart Government


Recommender Systems

Self-driving cars

Autonomous driving

- ALVINN – Drives 70mph on highways

Big data era

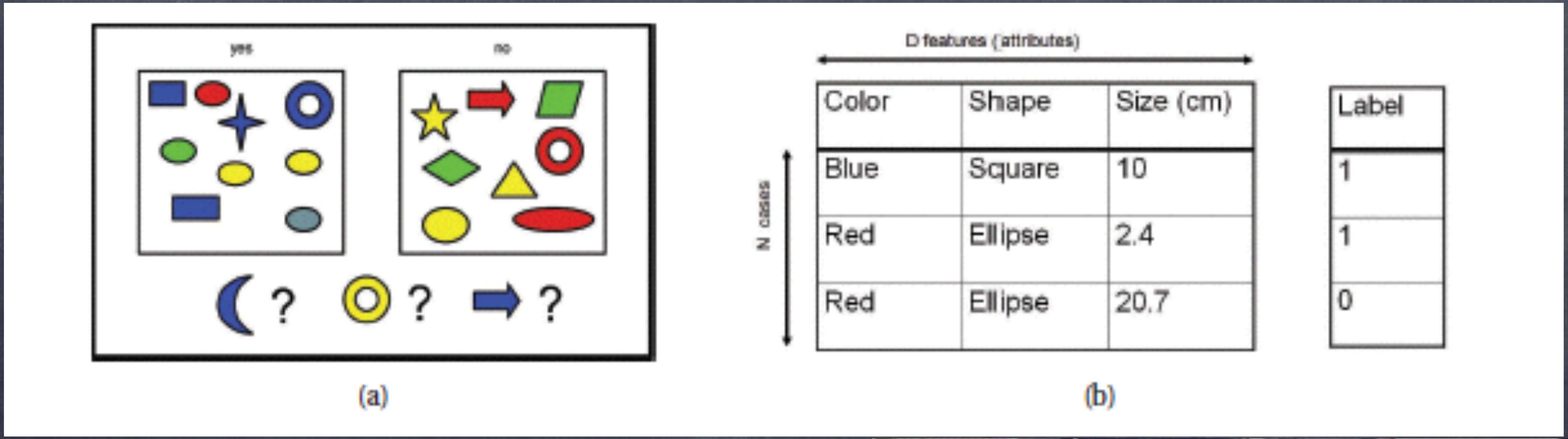
- There are about 1 trillion web pages
- one hour of video is uploaded to YouTube every second, amounting to 10 years of content every day
- Walmart handles more than 1M transactions per hour and has databases containing more than 2.5 petabytes of information

This deluge of data calls for automated methods of data analysis, which is what machine learning provides.

"set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty" - Kevin Murphy (Machine Learning : A Probabilistic Perspective)

Need for Probabilistic machine Learning

- Probability theory provides a mathematical framework to handle uncertainty



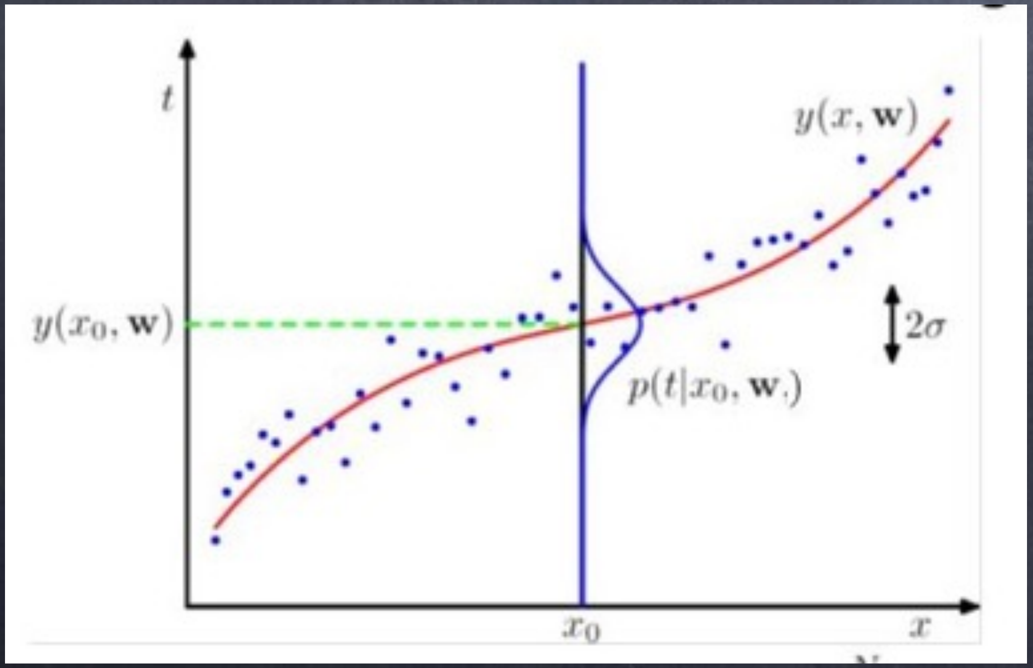
Color	Shape	Size (cm)	Label
Blue	Square	10	1
Red	Ellipse	2.4	1
Red	Ellipse	20.7	0

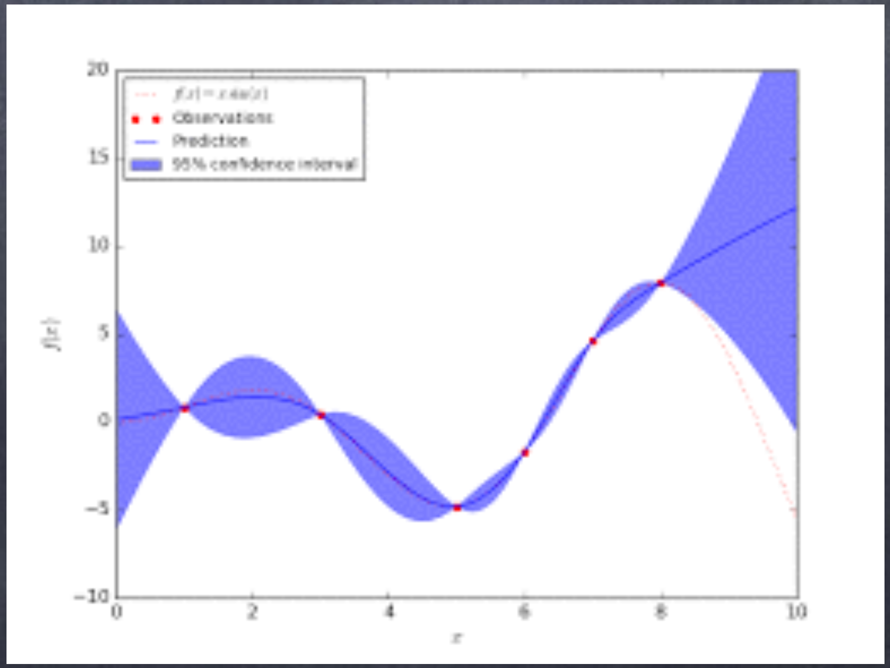
(a)

(b)

Probabilistic machine learning models - regression

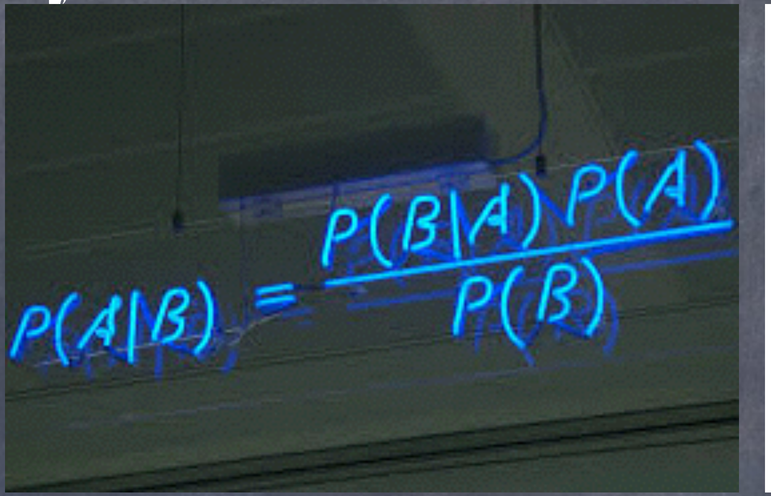
- Probabilistic linear regression
- Bayesian logistic regression
- Gaussian process regression

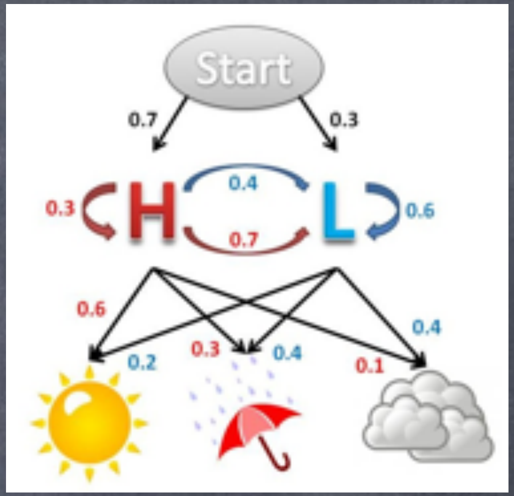


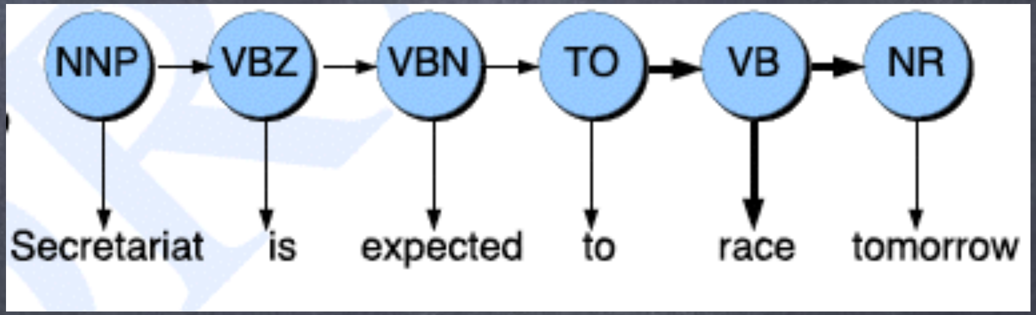


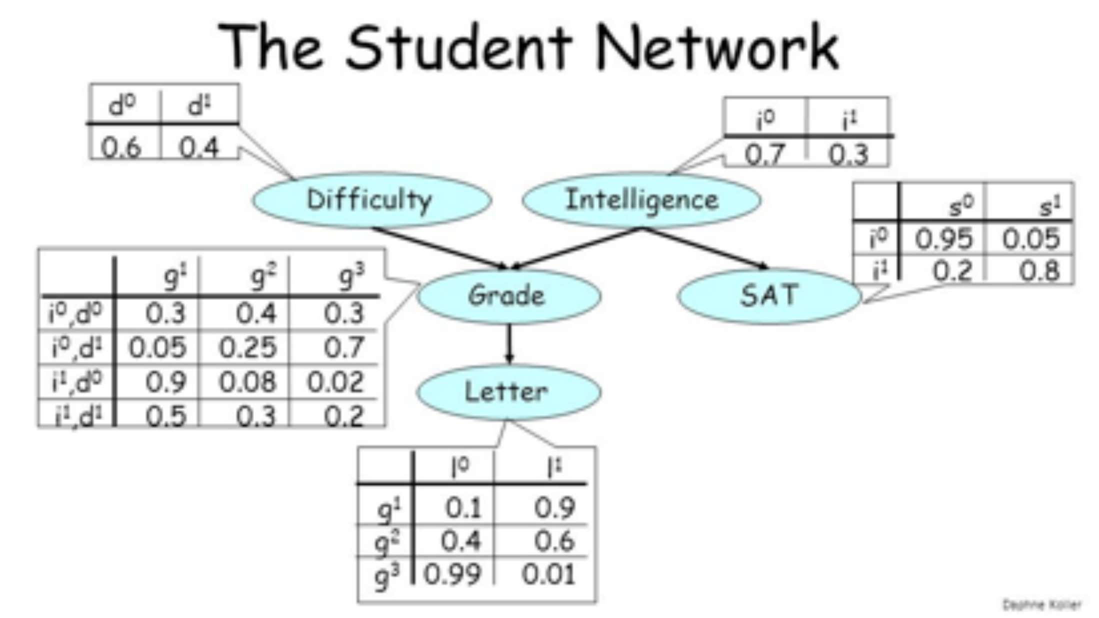
Probabilistic machine learning models - Classification

- Naive bayes classifier
- Logistic regression
- Sequence classification
    - Hidden markov models
    - Conditional Random Fields









The Student Network

d^0	d^1
0.6	0.4

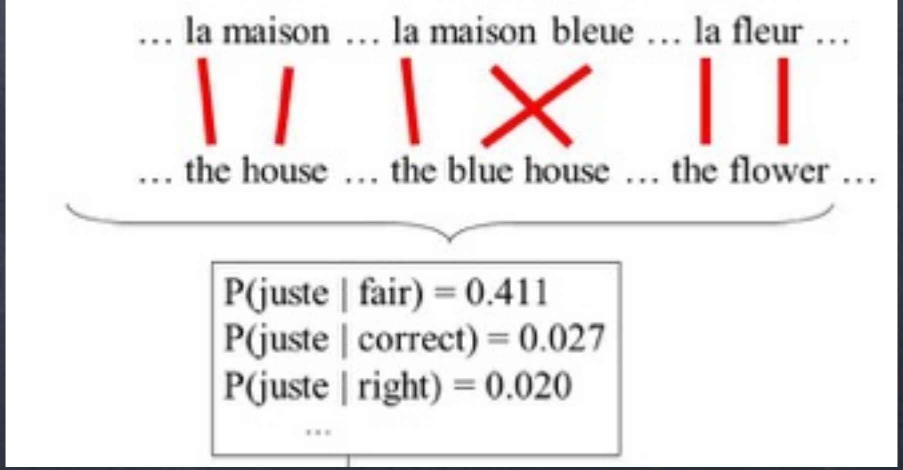
i^0	i^1
0.7	0.3

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

Daphne Koller



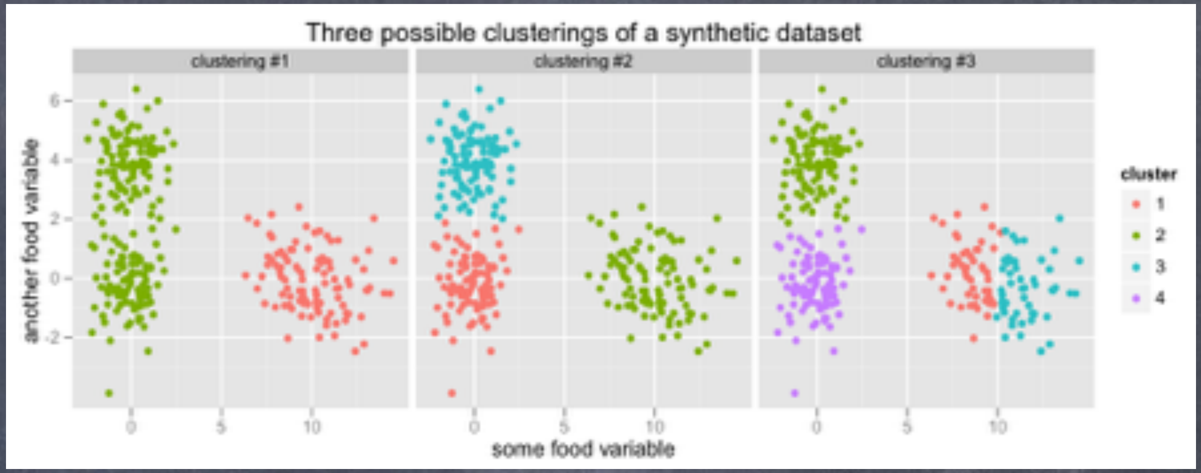
… la maison … la maison bleue … la fleur …

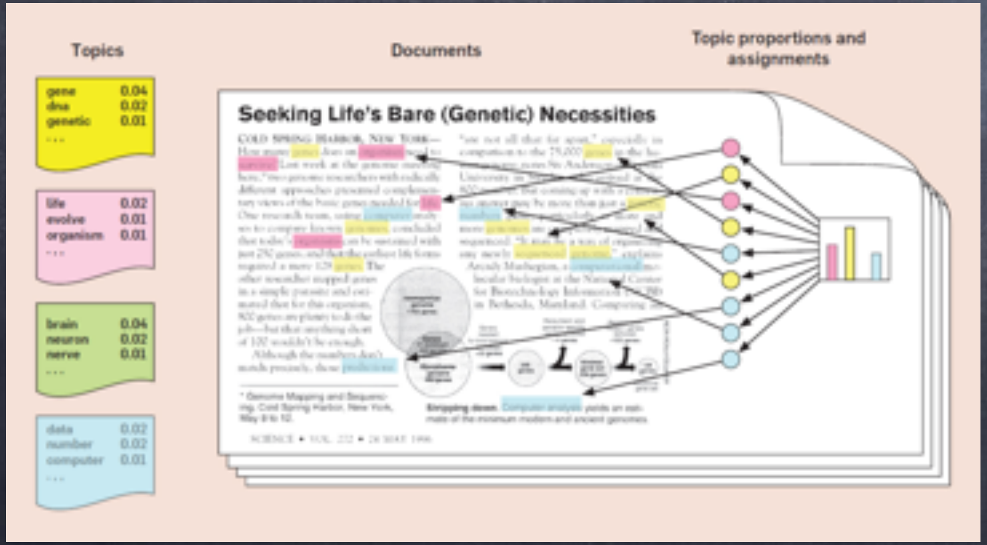
… the house … the blue house … the flower …

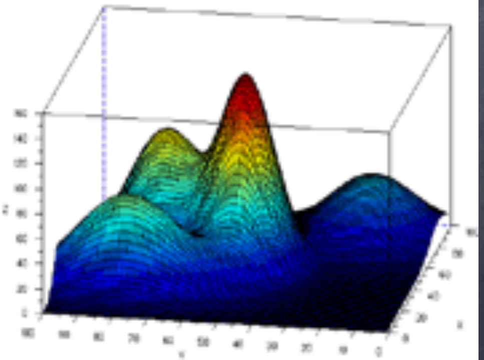
P(juste | fair) = 0.411
P(juste | correct) = 0.027
P(juste | right) = 0.020
…

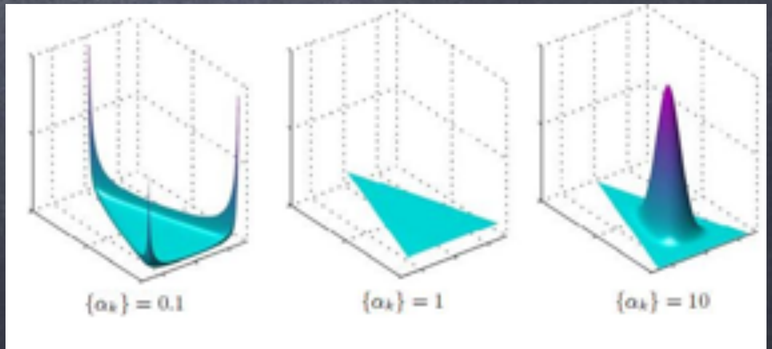
Probabilistic machine learning models - Clustering

- Gaussian Mixture Model
- Latent Dirichlet Allocation
- Dirichlet Process Mixture



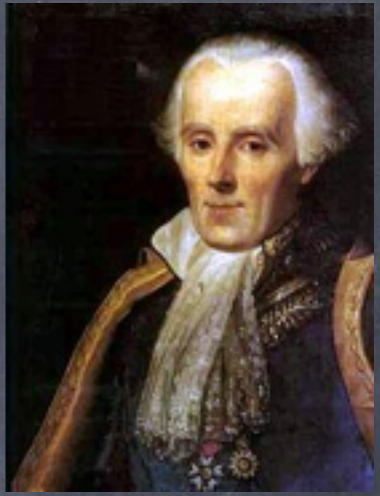

Topics
Documents
Topic proportions and assignments
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996





Probability

"Probability theory is nothing but common sense reduced to calculation" — Pierre Laplace, 1812



Whats Probability ?

"the probability that a coin will land heads is 0.5".
"probability that the polar ice cap will melt by 2020 CE"
"there is a 60 percent chance of oil in a certain region,"

- Frequentist interpretation
- Bayesian interpretation

compute the probability that a specific email message is spam.

SAMPLE SPACE AND EVENTS

- set of all possible outcomes of an experiment is known as the sample space (S)
- If the outcome of an experiment consists in the determination of the sex of a newborn child, then S = {g , b}
- Suppose we are interested in determining the amount of dosage that must be given to a patient until that patient reacts positively let S = $(0,\infty)$
- Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. E = {g }, then E is the event that the child is a girl

Sample Space and Events

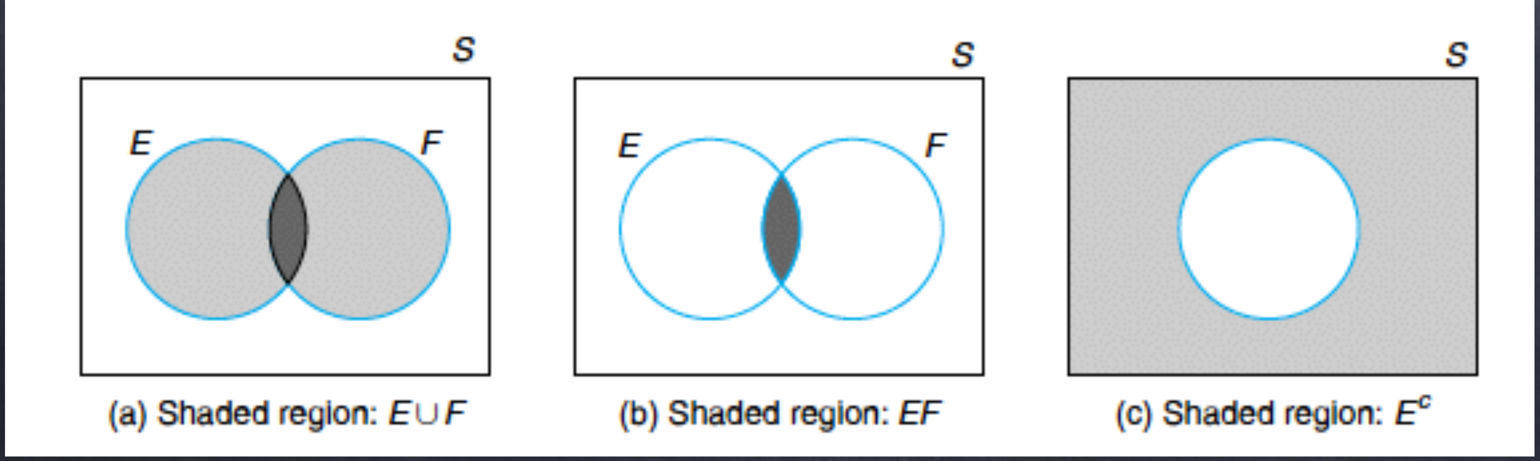
- union of the events E and F,consist of all outcomes that are either in E or in F or in both E and F.E = {g } and F = {b}, then $E \cup F$ = {g , b}.

- EF, called the intersection of E and F, to consist of all outcomes that are in both E and F. E = (0, 5), F =(2, 10), EF =(2, 5)

- EF = $\emptyset$, implying that E and F cannot both occur, then E and F are said to be mutually exclusive.

- Ec referred to as the complement of E, to consist of all outcomes in the sample space S that are not in E. if E = {b}Ec = {g }

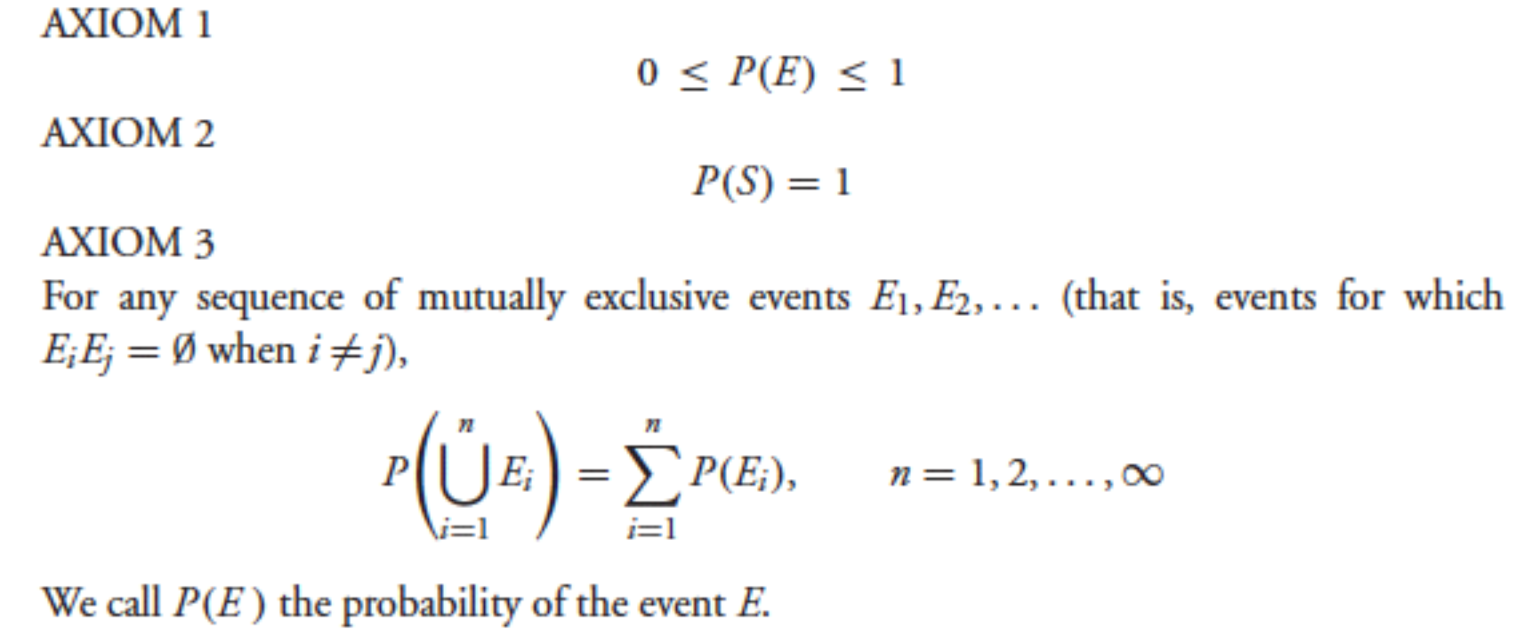
- E is contained in F and write $E \subset F$



(a) Shaded region: $E \cup F$ (b) Shaded region: EF (c) Shaded region: E^c

Axioms of Probability

- Probability : the proportion of time that the outcome is contained in E (frequentist)/strength of belief(bayesian)
- Kolmogorov Axioms/Cox theorem



AXIOM 1

$$0 \leq P(E) \leq 1$$

AXIOM 2

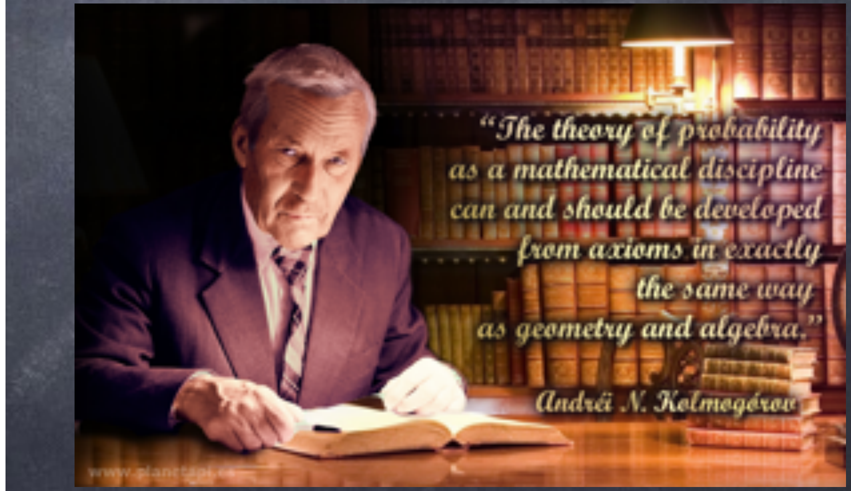
$$P(S) = 1$$

AXIOM 3

For any sequence of mutually exclusive events $E_1, E_2, \ldots$ (that is, events for which $E_i E_j = \emptyset$ when $i \neq j$),

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i), \qquad n = 1, 2, \ldots, \infty$$

We call $P(E)$ the probability of the event E.


"The theory of probability as a mathematical discipline can and should be developed from axioms in exactly the same way as geometry and algebra."
Andreï N. Kolmogórov

Probability

- $P(Ec\ ) = 1-P(E)$; $P(E \cup F\ ) = P(E)+P(F\ )-P(EF)$
- Each point in S={1,…,N} is equally likely

$$P(E) = \frac{\text{Number of points in } E}{N}$$

- A total of 28 percent of Indian males smoke cigarettes, 7 percent smoke cigars, and 5 percent smoke both cigars and cigarettes. What percentage of males smoke neither cigars nor cigarettes?

Probability Space

- A probability space is a mathematical triplet (Ω, F, P) consisting of

1. A sample space, Ω , which is the set of all possible outcomes.
2. A set of events F , where each event is a set containing zero or more outcomes.
3. The assignment of probabilities to the events; that is, a function P from events to probabilities.

If the experiment consists of just one flip of a fair coin, then the outcome is either heads or tails:Ω ={H,T} , in other words, F={{},{H},{T},{H,T}}. There is a fifty percent chance of tossing heads and fifty percent for tails, so the probability measure in this example isP({})=0,P({H})=0.5,P({T})=0.5,P({H,T})=1

Conditional Probability

- Calculating probabilities when some partial information concerning the result of the experiment is available

- Rolls a pair of dice ; S = {(i, j), i = 1, 2, 3, 4, 5, 6, j = 1, 2, 3, 4, 5, 6}, each possible outcomes is equally likely, first die lands on side 3 (F), sum of the two dice equals 8 (E)?

$$P(E|F) = \frac{P(EF)}{P(F)}$$

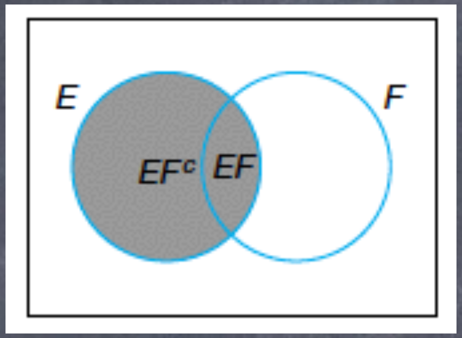
- Mr. John figures that there is a 30 percent chance that her company will set up a branch office in Mumbai. If it does, he is 60 percent certain that he will be made manager of this new operation. What is the probability that John will be a Mumbai branch office manager?

Conditional Probability

- $E = EF \cup EF^c$
- $P(E) = P(EF)+P(EF^c)$

  $= P(E|F)P(F)+P(E|F^c)P(F^c)$

  $= P(E|F)P(F)+P(E|F^c)[1-P(F)]$



- Helpful when it is difficult to compute the probability of an event directly
- An insurance company believes that people can be divided into two classes – those that are accident prone and those that are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability .4, whereas this probability decreases to .2 for a non-accident-prone person. If we assume that 30 percent of the population is accident prone, what is the probability that a new policy holder will have an accident within a year of purchasing a policy?

Independence

- Two events E and F are said to be independent iff P(EF) = P(E)P(F)
- If E and F are independent, then so are E and F c
- A card is selected at random from an ordinary deck of 52 playing cards. If A is the event that the selected card is an ace and H is the event that it is a heart, then A and H are independent

Bayes Theorem

- Suppose that a new policy holder has an accident within a year of purchasing his policy. What is the probability that he is accident prone?

- F1, F2, . . . , Fn are mutually exclusive events such that ,
  then $\bigcup_{i=1}^{n} F_i = S$ and $E = \bigcup_{i=1}^{n} EF_i$ $P(E) = \sum_{i=1}^{n} P(EF_i) = \sum_{i=1}^{n} P(E|F_i)P(F_i)$

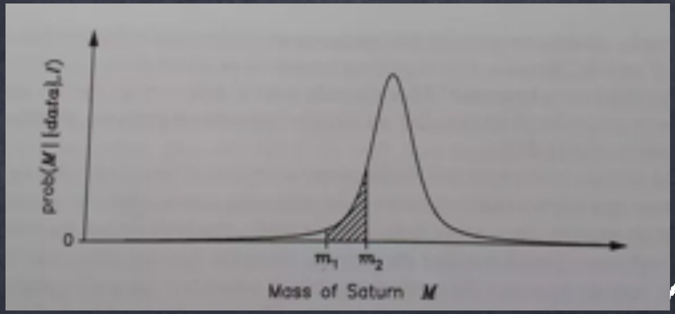
- Bayes formulae (Inverse Probability)

$$P(F_j|E) = \frac{P(EF_j)}{P(E)} = \frac{P(E|F_j)P(F_j)}{\sum_{i=1}^{n} P(E|F_i)P(F_i)}$$

Publication on Dec. 23, 1763 of *"An Essay towards solving a Problem in the Doctrine of Chances"* by the late Rev. Mr. Bayes, communicated by Mr. Price in the *Philosophical Transactions of the Royal Society of London.*

- http://www.stat.ucla.edu/history/essay.pdf

- Hypotheses held before the experiment P(Fj) to be modified by the evidence the experiment P(Fj|E)



- It provides a formulae to calculate probability P(Fj|E) from P(E|Fj)

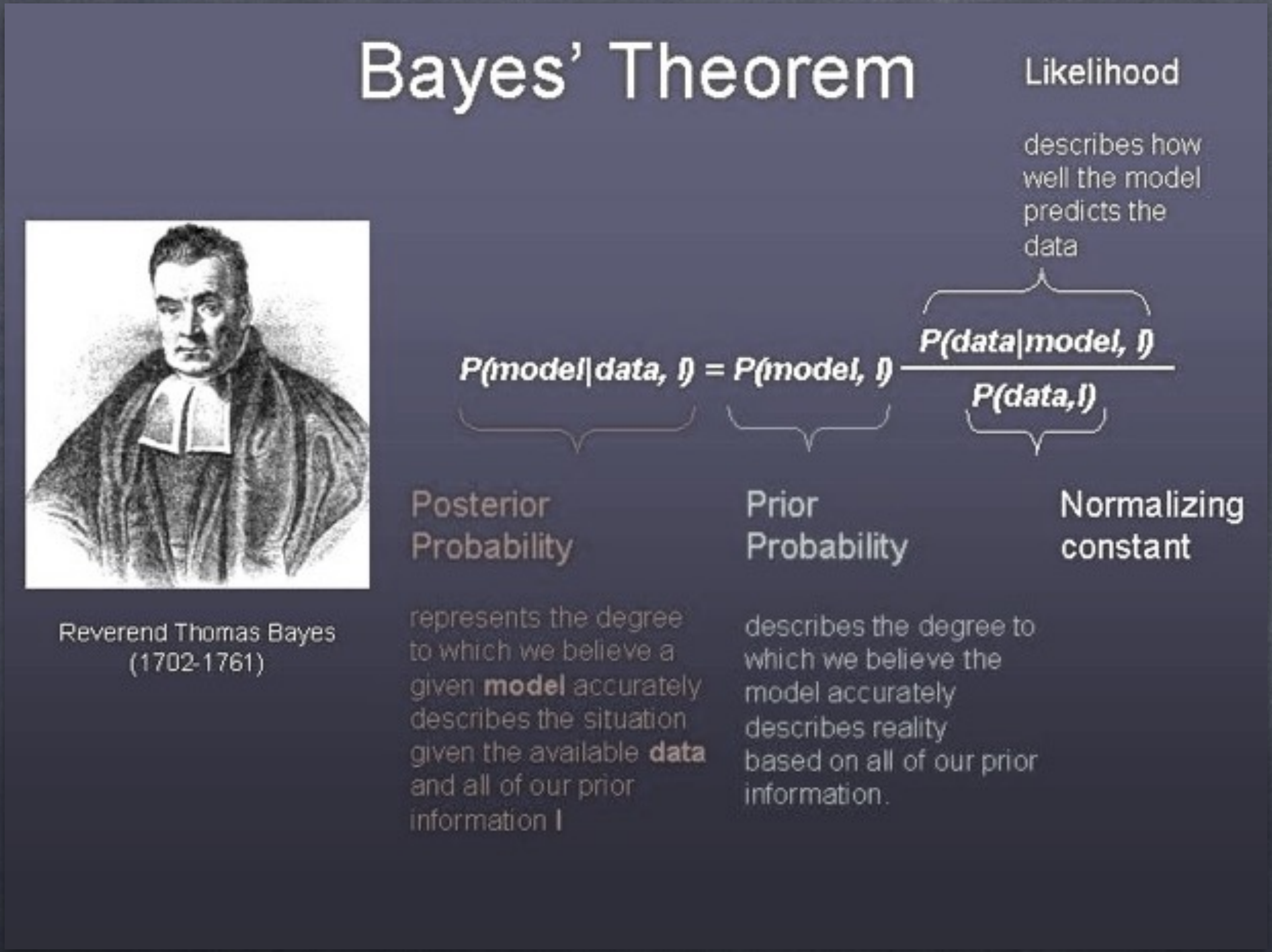
1812 "Théorie analytique des probabilités."

Bayes Theorem

- In a community of 100,000 people, 1,000 people will have cancer and 200 people will be 65 years old. Of the 1000 people with cancer, only 5 people will be 65 years old. Thus, of the 200 people who are 65 years old, how many is expected to have cancer?

- A laboratory blood test is 99 percent effective in detecting a certain disease when it is, in fact, present. However, the test also yields a "false positive" result for 1 percent of the healthy persons tested. (That is, if a healthy person is tested, then, with probability .01, the test result will imply he or she has the disease.) If .5 percent of the population actually has the disease, what is the probability a person has the disease given that his test result is positive?

Bayesian Modelling



Bayes' Theorem

Reverend Thomas Bayes
(1702-1761)

$$\underbrace{P(model|data, I)}_{\text{Posterior Probability}} = \underbrace{P(model, I)}_{\text{Prior Probability}} \frac{\overbrace{P(data|model, I)}^{\text{Likelihood}}}{\underbrace{P(data,I)}_{\text{Normalizing constant}}}$$

Likelihood

describes how well the model predicts the data

Posterior Probability

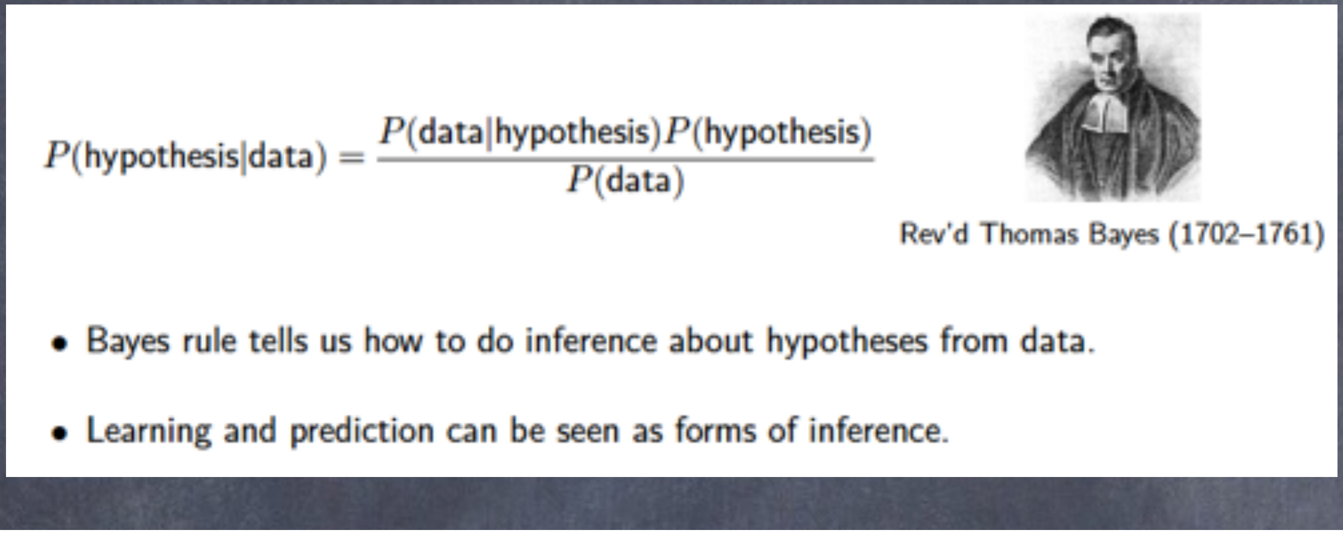
represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability

describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

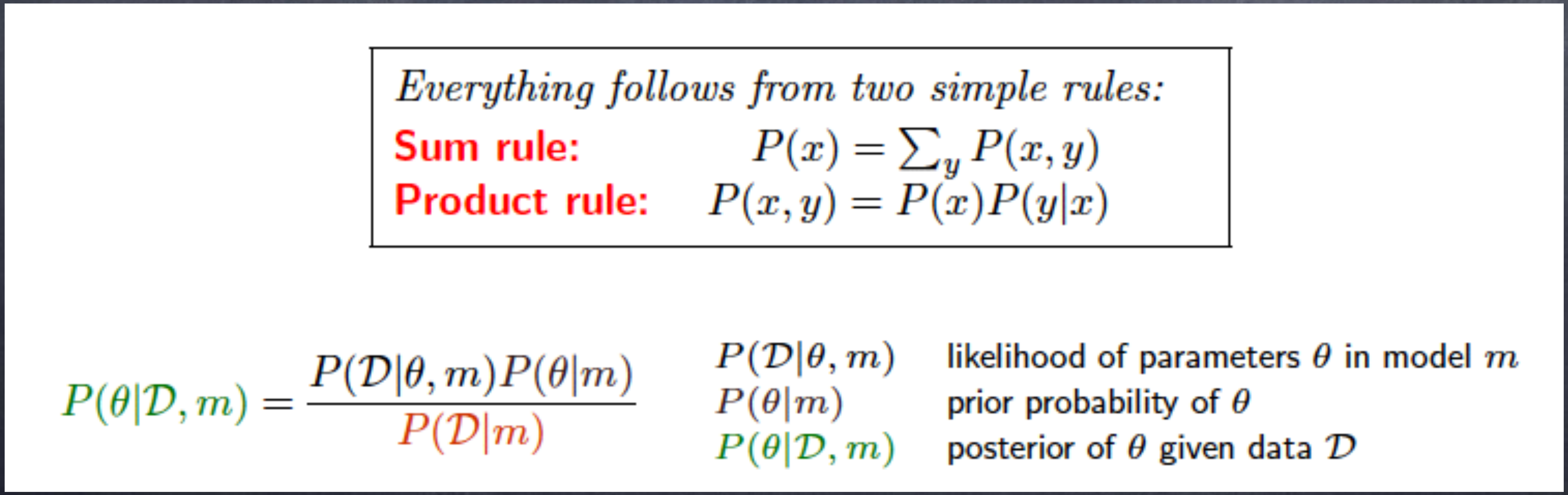
Bayesian Modelling



$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Rev'd Thomas Bayes (1702–1761)

- Bayes rule tells us how to do inference about hypotheses from data.
- Learning and prediction can be seen as forms of inference.



Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x,y)$

Product rule: $P(x,y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D},m) = \frac{P(\mathcal{D}|\theta,m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta,m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D},m)$	posterior of θ given data $\mathcal{D}$