INTRO TO MACHINE LEARNING

Sergey Nikolenko

Harbour Space University

June 8, 2026

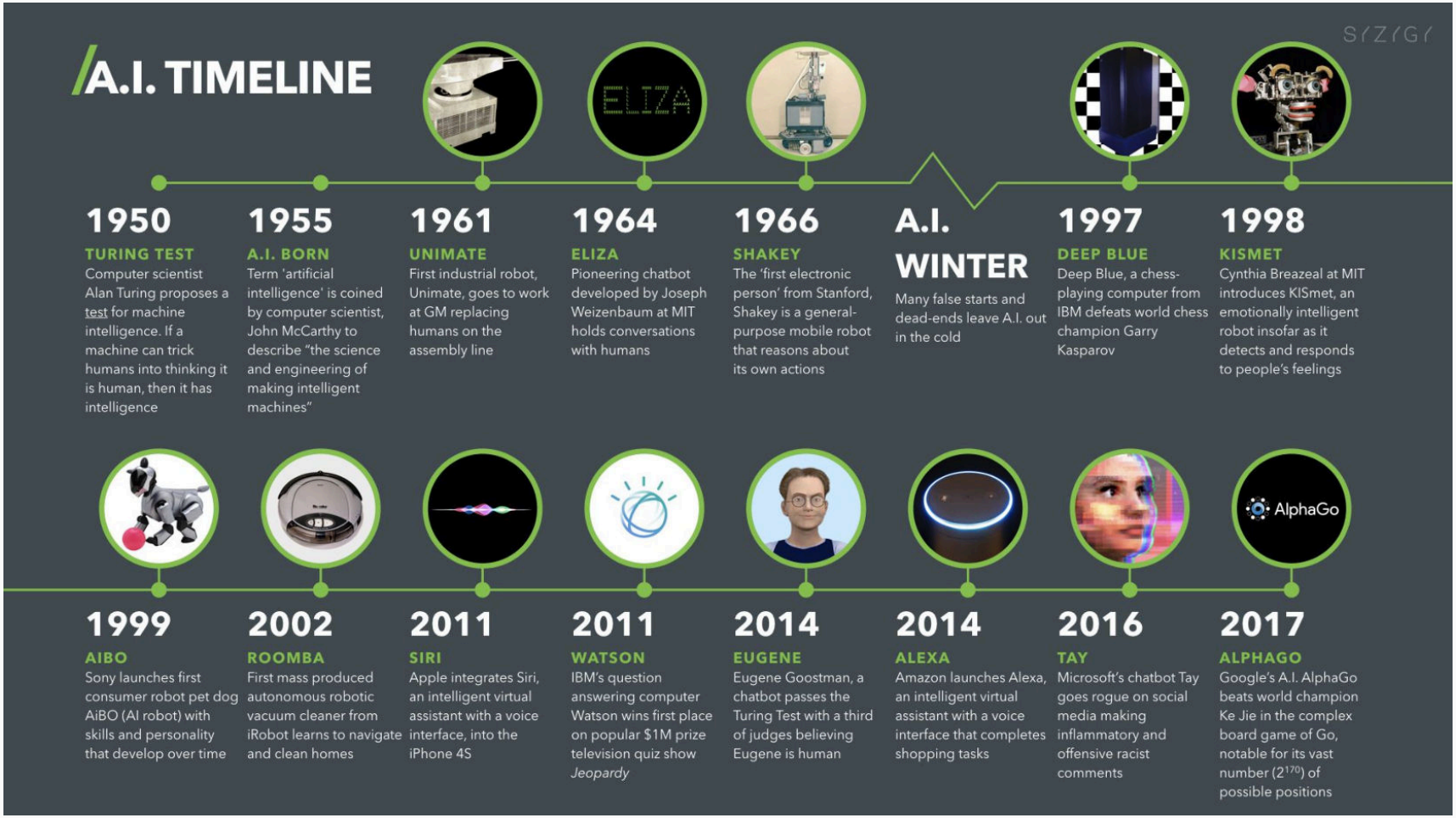
Random facts:

- on June 8, 793, "the ravaging of wretched heathen men destroyed God's church at Lindisfarne" in Northumbria; thus began Viking activity on the British Isles
- on June 8, 1191, Richard Cœur de Lion arrived in Acre and began the Third Crusade; he defeated Saladin at Arsuf and Jaffa, reaching a settlement in September 1192
- on June 8, 1794, Maximilien Robespierre opened the first celebration of the French Revolution's new state religion, *Culte de l'Être suprême* (Cult of the Supreme Being); large festivals were to be held all across France, with Jacques-Louis David organizing it in Paris; when Robespierre was guillotined on July 28 (in two months) the religion immediately disappeared and was officially banned by Napoleon in 1802
- on June 8, 1887, Herman Hollerith applied for US patent #395,781 for the "Art of Compiling Statistics", which was his punched card calculator
- on June 8, 1959, *USS Barbero* and the US Postal Service attempted the delivery of mail via Missile Mail, firing a cruise missile with nuclear warhead replaced by mail containers

What is Machine Learning

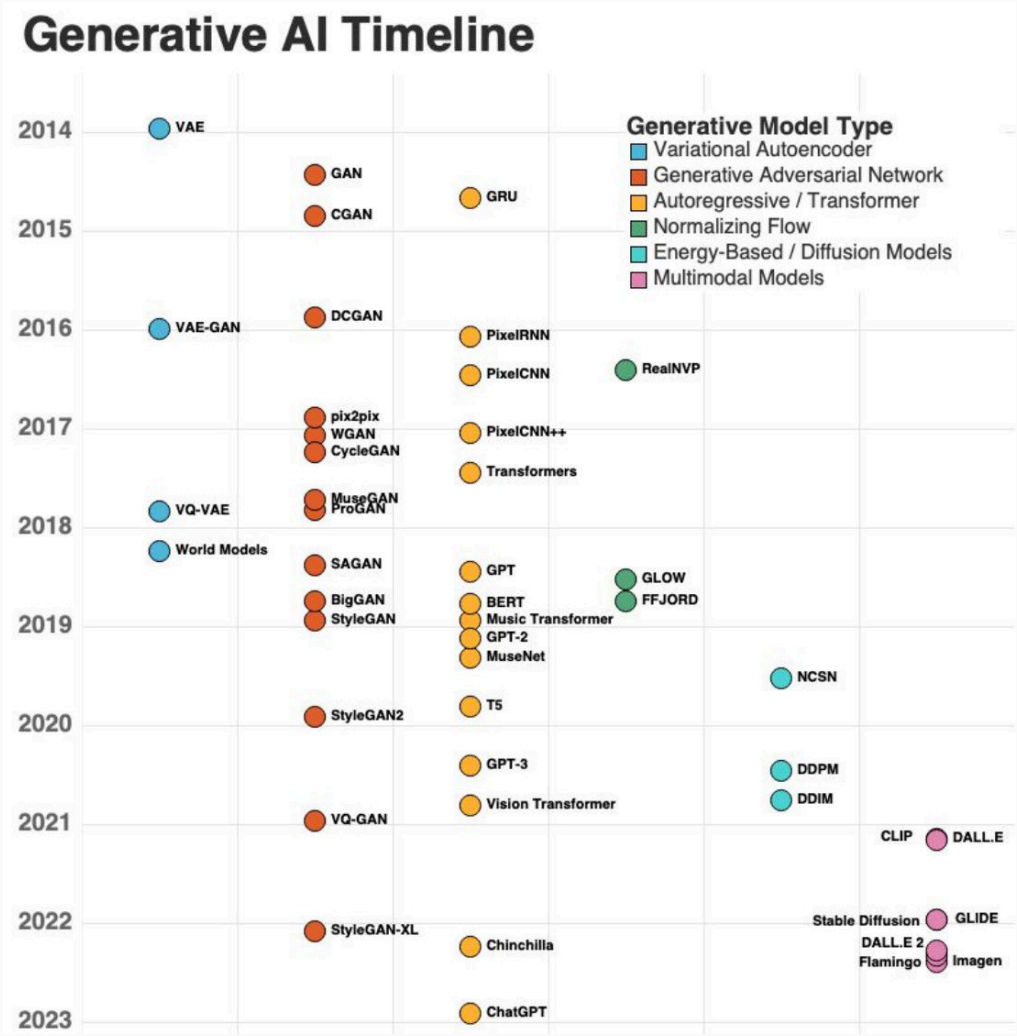
AI TIMELINES

- We live in some very interesting times...



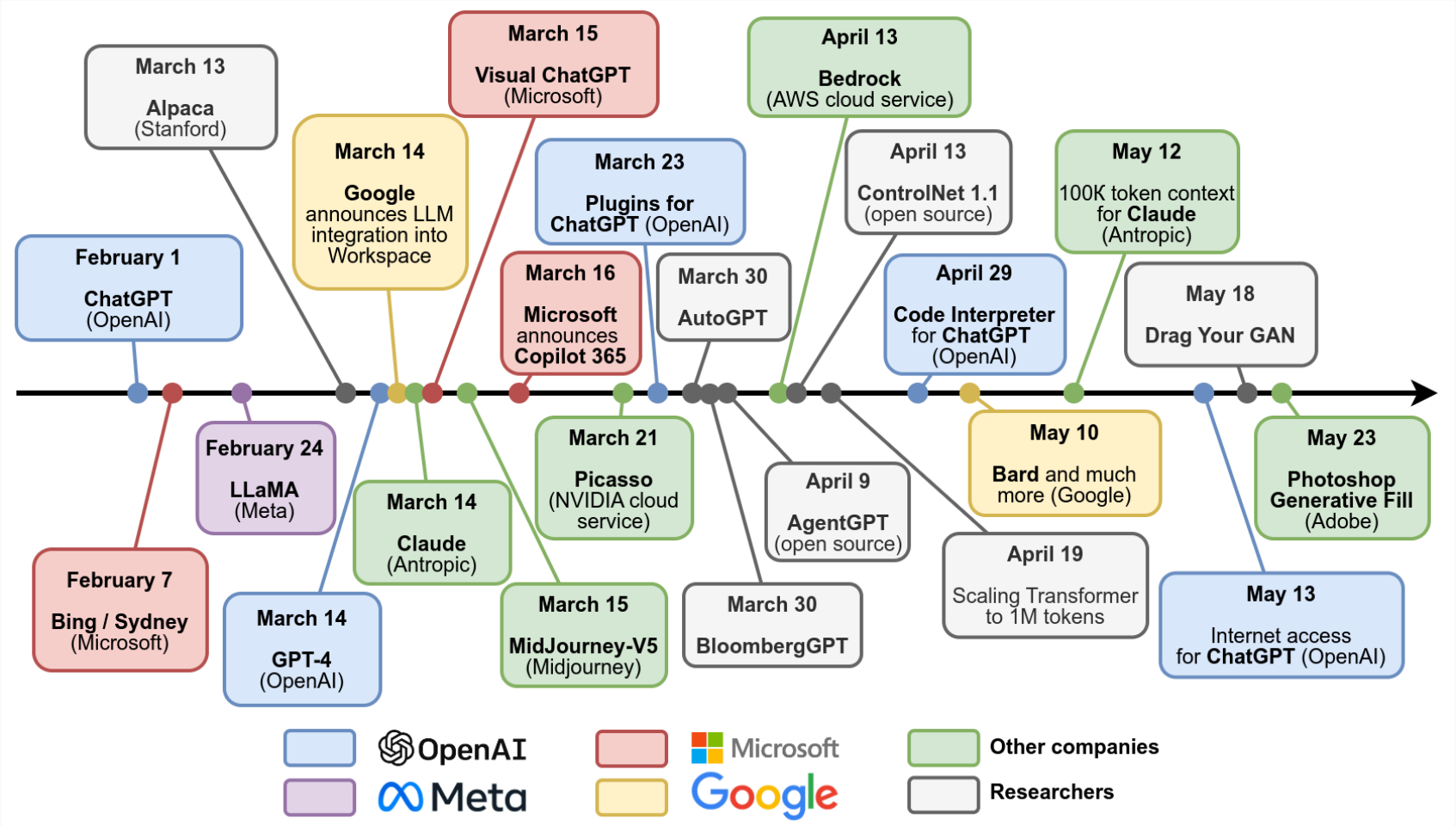
AI TIMELINES

- We live in some very interesting times...



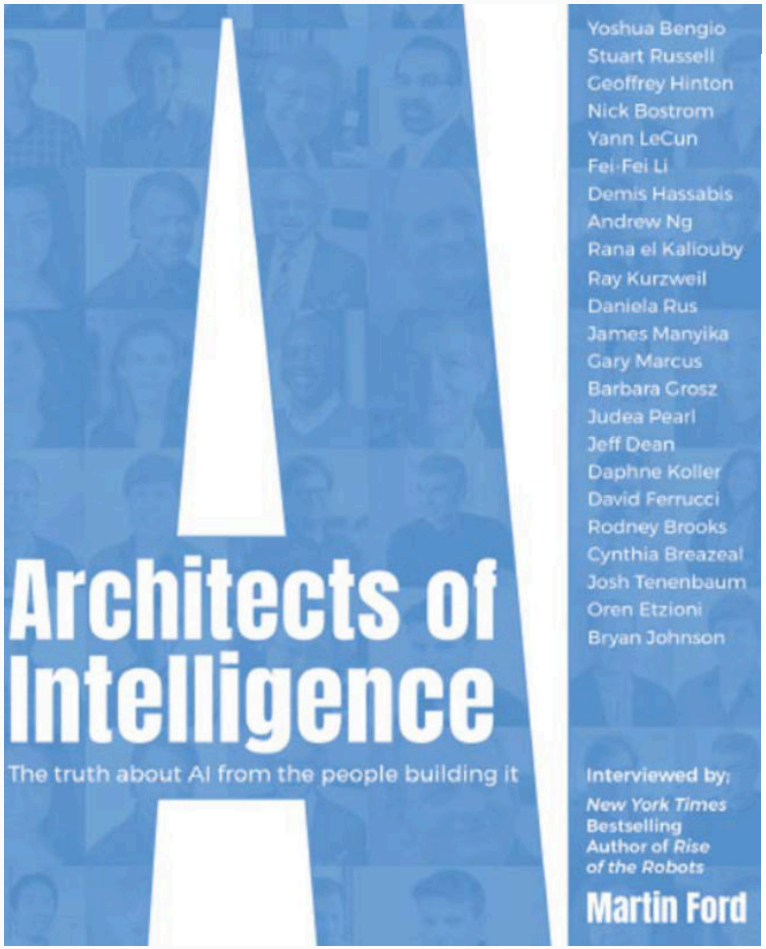
AI TIMELINES

- We live in some very interesting times...



AI TIMELINES

- And nobody predicted this! This is a book from 2018:



When Will Human-Level AI be Achieved? Survey Results

As part of the conversations recorded in this book, I asked each participant to give me his or her best guess for a date when there would be at least a 50 percent probability that artificial general intelligence (or human-level AI) will have been achieved. The results of this very informal survey are shown below.

A number of the individuals I spoke with were reluctant to attempt a guess at a specific year. Many pointed out that the path to AGI is highly uncertain and that there are an unknown number of hurdles that will need to be surmounted. Despite my best persuasive efforts, five people declined to give a guess. Most of the remaining 18 preferred that their individual guess remain anonymous.

As I noted in the introduction, the guesses are neatly bracketed by two people willing to provide dates on the record: Ray Kurzweil at 2029 and Rodney Brooks at 2200.

Here are the 18 guesses:

2029 .....11 years from 2018
2036 .....18 years
2038 .....20 years
2040 .....22 years
2068 (3) .....50 years
2080 .....62 years
2088 .....70 years
2098 (2) .....80 years
2118 (3) .....100 years
2168 (2) .....150 years
2188 .....170 years
2200 .....182 years

Mean: 2099, 81 years from 2018

- Let's start from the beginning...

Early Thoughts on Artificial Intelligence

- Hephaestus created android robots for himself, such as the giant humanoid robot Talos
- Pygmalion brought Galatea to life
- Jehovah and Allah—pieces of clay
- Some particularly wise rabbis could create golems
- Albertus Magnus crafted an artificial speaking head (greatly upsetting Thomas Aquinas)

Fig. 432.—*The Talking Head of Albertus Magnus.*

Mechanical Automata

- Early automata were quite inventive: al-Jazari (12th century)

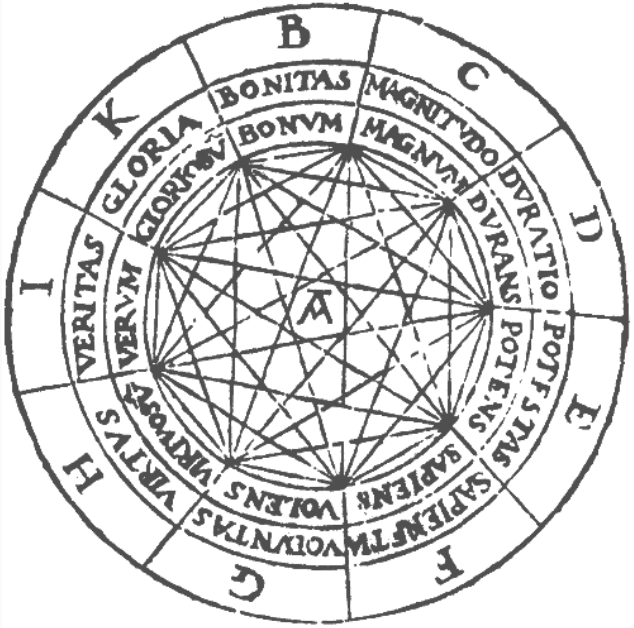
- Robert II d'Artois' Hesdin Castle, Leonardo da Vinci's mechanical knight...

Mechanical Automata

- Automata by Jaquet-Droz (XVIII century):

Mechanical Automata

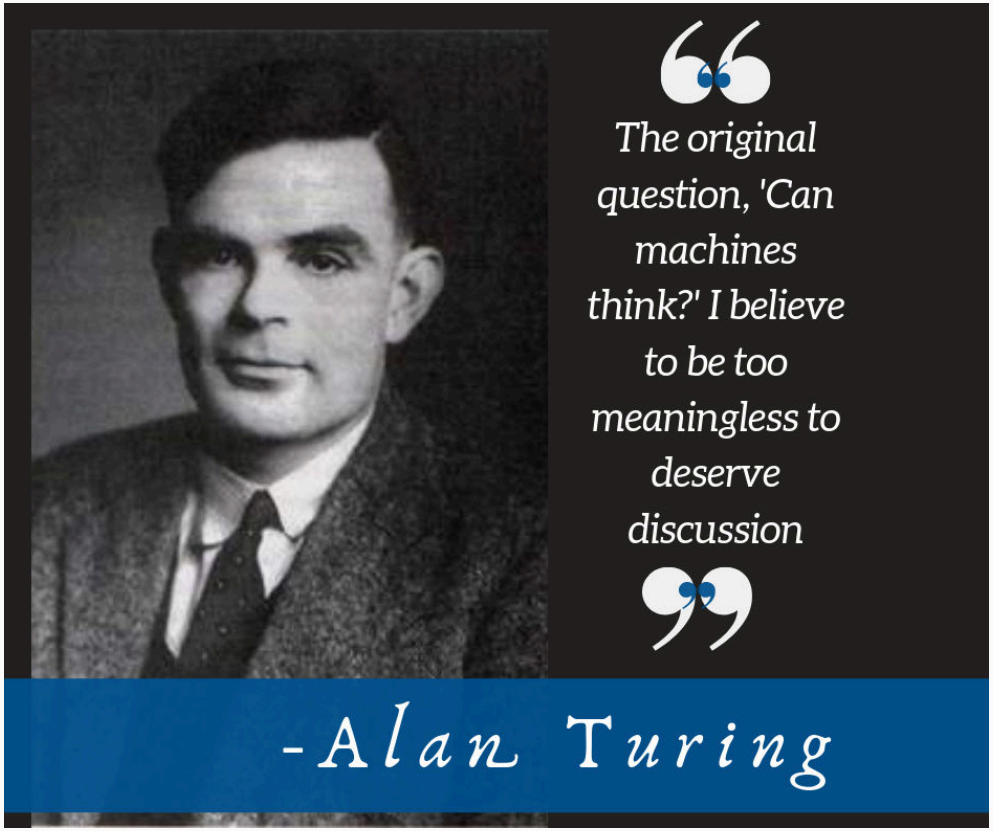
- But these were purely mechanical; mathematical/logical AI remained rudimentary for a long time.
- Logical machine by Ramon Llull (late XIII–early XIV century):



- Starting from Dr. Frankenstein, AI consistently appears in literature...

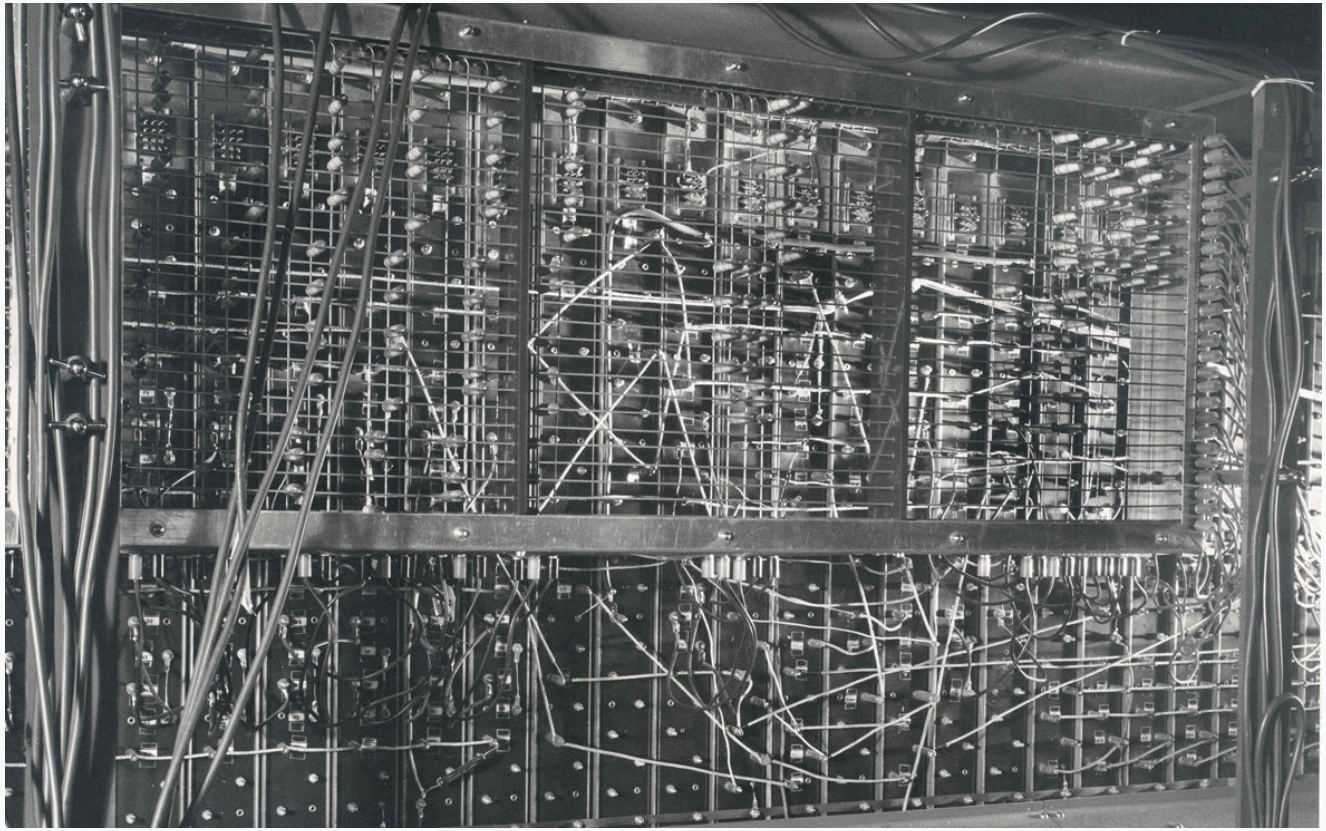
The Turing Test

- AI as a science began with the *Turing test* (1950).
- A computer should successfully impersonate a human in a (written) dialogue between a judge, a human, and a computer. Actually, the original formulation was somewhat subtler and more interesting...



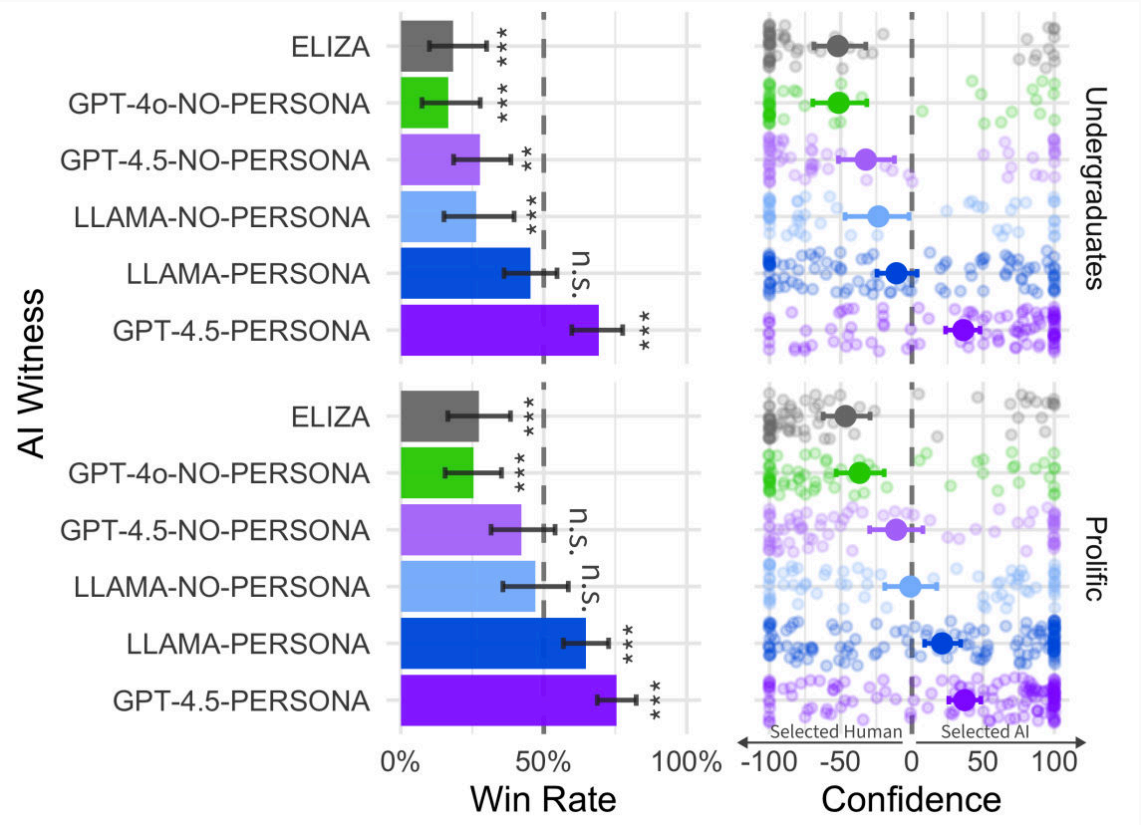
The Turing Test

- It is already obvious how much is required to build AI:
  - natural language processing;
  - knowledge representation;
  - reasoning from acquired knowledge;
  - learning from experience (machine learning itself).



The Turing Test

- By the way, it required special efforts to try the Turing test with modern LLMs, so officially it was passed only very recently (Jones, Bergen, Mar 31, 2025):



The Dartmouth Workshop

- AI as a field emerged in 1956 at the Dartmouth workshop.
- It was organized by John McCarthy, Marvin Minsky, Claude Shannon, and Nathaniel Rochester.
- Probably the most ambitious grant proposal in computing history.

The Dartmouth Workshop

We propose a 2-month study of artificial intelligence to be carried out by 10 people in the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is based on the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can simulate it. We will attempt to find how to make machines use natural languages, form abstractions and concepts, solve problems now reserved for humans, and improve themselves. We think significant advancement in one or more of these problems is possible if a carefully selected group of scientists works on it for a summer.

1956–1960: High Expectations

- Optimistic times. It seemed that just a bit more, just a little bit...
- Allen Newell, Herbert Simon: *Logic Theorist*. Successfully re-proved most of *Principia Mathematica*, sometimes even more elegantly than Russell and Whitehead themselves.

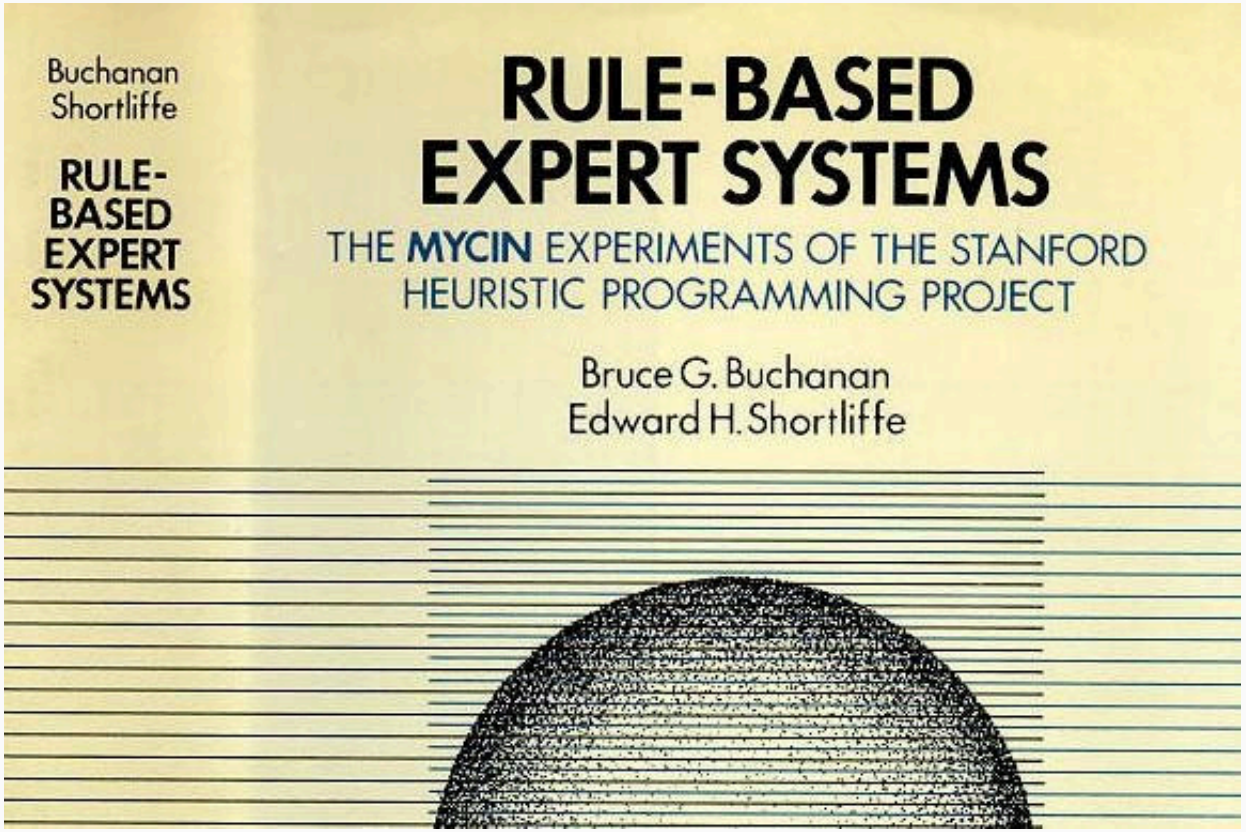
Herb Simon and Allen Newell. *Photograph courtesy, Carnegie Mellon University.*

1956–1960: High Expectations

- Optimistic times. It seemed that just a bit more, just a little bit...
- General Problem Solver—a program trying to think like a human;
- Many specialized programs performing limited tasks (microworlds):
  - Analogy (IQ tests for “pick the odd one out”);
  - Student (algebraic word problems);
  - Blocks World (manipulating 3D blocks).

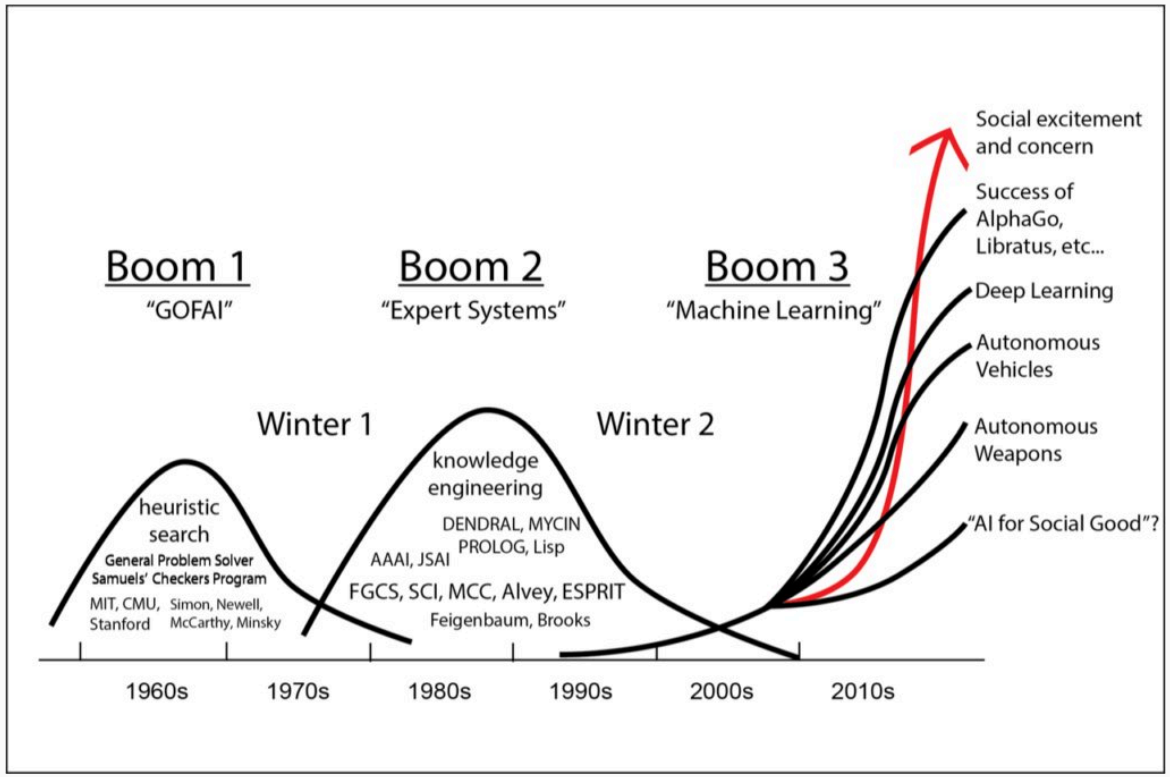
1970s: Knowledge-based Systems

- Core idea: accumulate a sufficiently large set of rules and domain knowledge, then draw conclusions.
- First success: MYCIN—blood infection diagnosis:
  - about 450 rules;
  - performed as well as experienced physicians and significantly better than novices.



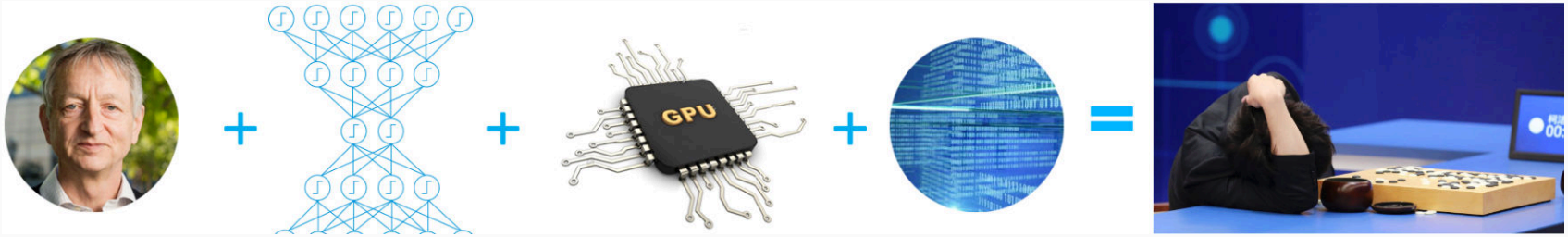
1980s: Commercial Applications; AI Industry

- The first AI department was at DEC (Digital Equipment Corporation); reportedly saving DEC about $10 million annually by 1986.
- The boom ended by the late 80s as many companies failed to meet overly high expectations.



1990–2010: Data Mining, Machine Learning

- In recent decades, the main focus shifted to machine learning and discovering patterns in data
- Especially with the development of the Internet
- Next we got to the deep learning revolution, and recently to the LLM revolution… but that is a very different story



Definition

- What does a ”learning machine” mean? How to define ”learnability”?

Definition

- What does a ”learning machine” mean? How to define ”learnability”?

Definition

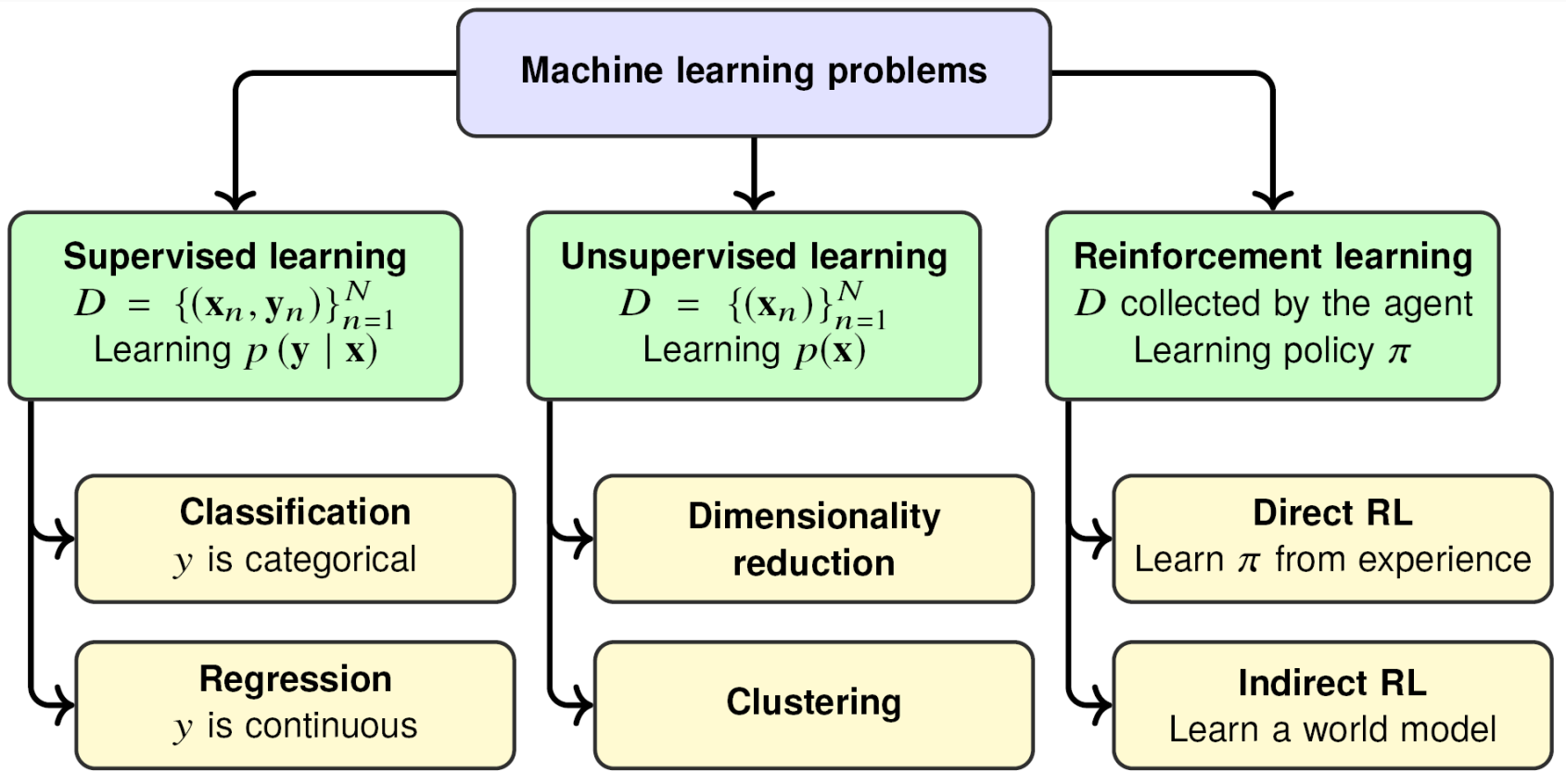
A computer program learns with accumulating experience regarding a class of tasks T and a performance measure P, if its performance on these tasks (according to P) improves as it gains new experience.

- This definition is very (too?) general.
- What specific examples can we provide?

What we will cover

- We will consider various algorithms solving specific tasks, performing better with more initial (test) data
- Today we'll discuss the general theory of Bayesian inference, usually applicable to any machine learning algorithm
- But first—a brief overview of basic ML tasks

Key ML tasks and concepts

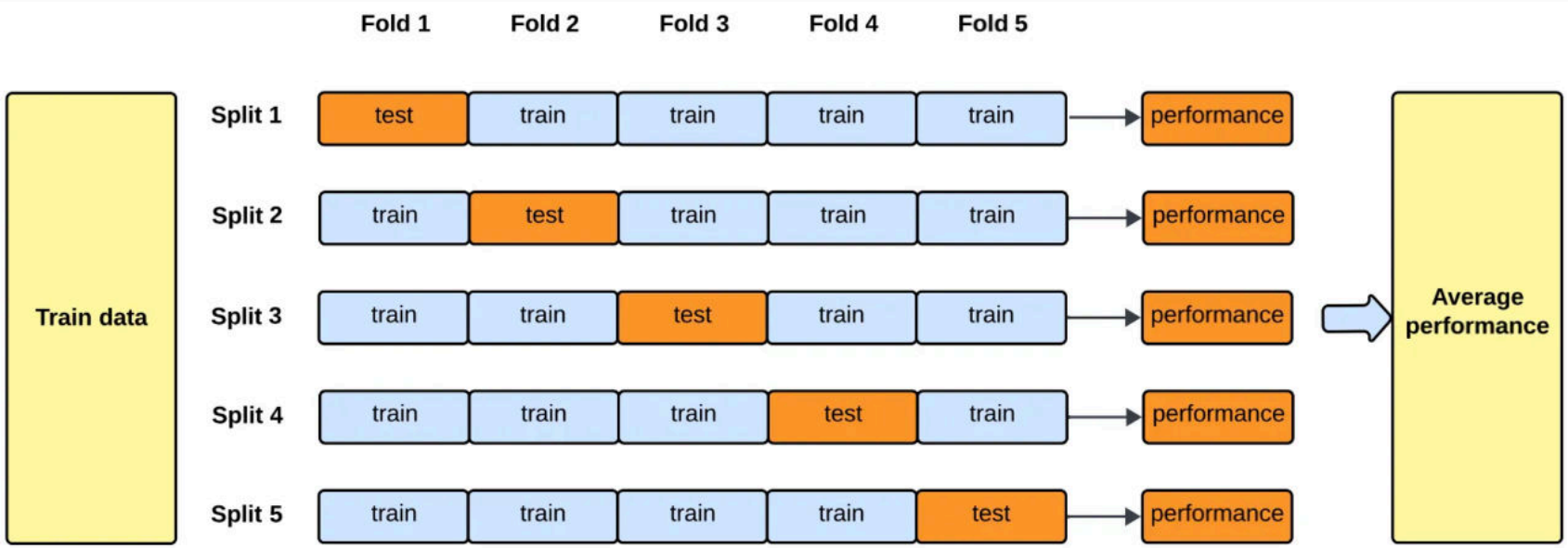


Key ML tasks and concepts

- *Supervised learning* – learning with a set of labeled examples:
  - *training set* – a set of examples, each consisting of *features* (attributes);
  - examples have correct answers – a variable (response) we want to predict; it can be categorical, continuous, or ordinal;
  - a model *learns* from this dataset (training phase, learning phase), then it can be applied to new examples (test set);
  - main goal – train a model that not only fits training data but also *generalizes* well to new data;
  - otherwise – overfitting;

Key ML tasks and concepts

- *Supervised learning* – learning with a set of labeled examples:
  - usually, we're given only the training set – how can we check model generalization?
  - cross-validation – splitting data into training and validation sets;
  - typically, preprocessing is performed before inputting data, extracting the most informative aspects (feature extraction).



Key ML tasks and concepts

- *Supervised learning* – learning with a set of labeled examples:
- *Classification*: there is a discrete set of categories (classes), and new examples must be assigned to a class;
  - text classification by topics, spam filtering;
  - face/object/text recognition;
- *Regression*: there is an unknown function whose values we must predict for new examples:
  - engineering applications (predicting temperature, robot position, etc.);
  - finance – predicting stock prices;
- same plus temporal changes – e.g., speech recognition

Key ML tasks and concepts

- *Supervised learning* – learning with a set of labeled examples:
- *Unsupervised learning* – learning without labeled data, only raw data:
  - *clustering*: partitioning data into unknown classes based on similarity:
    - identifying gene families from nucleotide sequences
    - user clustering and personalization
    - clustering mass spectrometry images into segments with different compositions

Key ML tasks and concepts

- *Supervised learning* – learning with a set of labeled examples:
- *Unsupervised learning* – learning without labeled data, only raw data:
  - *dimensionality reduction*: data has huge dimensionality (many features), and we need to reduce it by selecting the most informative features so algorithms can run effectively;
  - *matrix completion*: given a sparse matrix, predict missing values;
  - often some data has labeled examples – semi-supervised learning.

Key ML tasks and concepts

- *Reinforcement learning* – learning where an agent learns from its own trials and errors:
  - *multi-armed bandits*: choosing from several actions, each leading to random outcomes; goal is to maximize rewards;
  - *exploration vs. exploitation*: deciding when to explore new options versus exploiting known ones;
  - *credit assignment*: reward is given at the end (winning a game), and we must assign credit to each action leading to success.

Key ML tasks and concepts

- *Active learning* – choosing the next (relatively expensive) test;
- *Learning to rank* – ordinal regression to produce ordered lists (internet search);
- *Bootstrapping* – reusing the results of a model to improve itself;
- *Model selection* – finding the balance between models with many or few parameters.

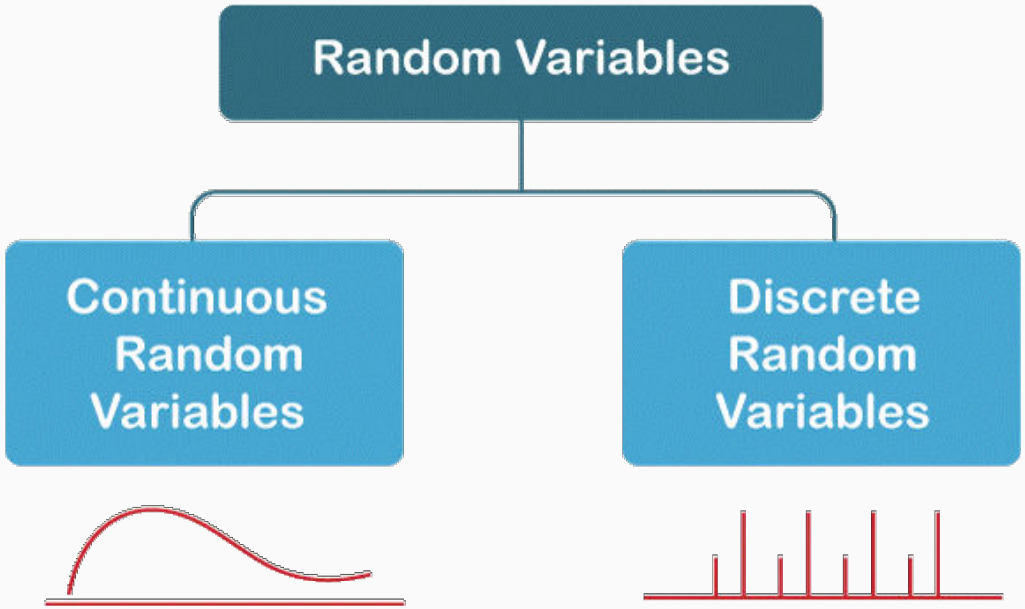
Probability in machine learning

- All methods and approaches benefit from methods that not only give answers but also estimate model confidence, data description quality, and anticipate how these measures change with further experiments.
- Probability theory thus plays a central role in machine learning, and we will actively use it.

Bayesian approach to ML

Basic definitions

- We won't need mathematical definitions of sigma-algebras, probability measures, Borel sets, etc.
- It suffices to understand discrete random variables (probabilities summing to one) and continuous random variables (integral of nonnegative density equals one)



Basic definitions

- *Joint probability* – probability of two events occurring simultaneously, $p(x, y)$; marginalization:

$$p(x) = \sum_y p(x, y).$$

- *Conditional probability* – probability of one event given another occurred, $p(x \mid y)$:

$$p(x, y) = p(x \mid y)p(y) = p(y \mid x)p(x).$$

- *Bayes' theorem* – from above:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'} p(x|y')p(y')}.$$

- *Independence*: x and y are independent if

$$p(x, y) = p(x)p(y).$$

Definitions

- Let us begin with the Bayes' theorem:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- Here, $p(\theta)$ is the *prior probability*, $p(D|\theta)$ is the *likelihood*, $p(\theta|D)$ is the *posterior probability*, and $p(D) = \int p(D \mid \theta)p(\theta)\mathrm{d}\theta$ is the evidence.

ML vs. MAP

- In statistics, typically we look for the *maximum likelihood hypothesis*:

$$\theta_{ML} = \arg\max_\theta p(D \mid \theta).$$

- Bayesian approaches seek the *posterior distribution*

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

  and possibly the *maximum a posteriori hypothesis*:

$$\theta_{MAP} = \arg\max_\theta p(\theta \mid D) = \arg\max_\theta p(D \mid \theta)p(\theta).$$

Direct and inverse problems

- Direct problem: an urn contains 10 balls, 3 of which are black. What is the probability of drawing a black ball?
- Or: an urn contains 10 balls numbered from 1 to 10. What's the probability that the sum of numbers on three consecutively drawn balls is 12?
- Inverse problem: two urns each contain 10 balls; one has 3 black balls, the other 6. Someone took a ball from one urn, and it was black. What's the probability that it was taken from the first urn?
- Notice that in the inverse problem, probabilities immediately become Bayesian (although it can also be reformulated through frequencies).

Direct and inverse problems

- In other words, direct probability problems describe a probabilistic process or model and ask to calculate a specific probability (i.e., predict behavior based on the model).
- Inverse problems contain *hidden variables* (in the example—the number of the urn from which the ball was drawn). They often require building a probabilistic model based on observed behavior.
- Machine learning problems typically belong to the second category.

Probability as frequency

- Classical probability theory, originating from physical experiments, usually views probability as the limit of the ratio of occurrences of a particular experiment outcome to the total number of experiments.
- Standard example: tossing a coin.

Probability as a degree of belief

- But we might want to reason about the "probability" of:
  - Spain winning the FIFA World Cup in 2026;
  - the *Odyssey* having been written by a woman;
  - AGI taking over control from humans by 2030;
  - …
- But "experiments repeated to infinity" are meaningless here—there is exactly one experiment.

Probability as a degree of belief

- Here, probabilities become *degrees of belief*. This is the Bayesian interpretation of probability (as Thomas Bayes originally saw it).
- Fortunately, both interpretations follow the same laws; results show that natural axioms of probabilistic logic lead to a very restricted class of functions (Cox, 19).

Examples

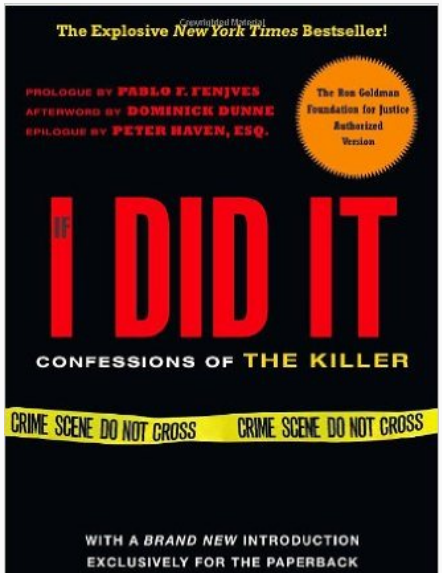
1. A person has two kids. Let's assume that every baby is born either a boy or a girl independently with equal probabilities $\frac{1}{2}$. Two problem settings:
   (1) I've asked if the guy has any boys, and he said yes; what's the probability that he has a girl?
   (2) I've met one of his kids, and it's a boy; what's the probability that the other is a girl?

Examples

2. A murder has happened. Blood at the crime scene is certain to be the killer's, and it's a rare blood type with 1% prevalence... including the defendant.
   (1) Prosecutor says: «The chance that the defendant would have this blood type if he was innocent is only 1%, so with probability 99% he is guilty». What's the prosecutor's mistake?
   (2) Defense attorney says: «A million people live in the city, so about 10000 of them have the same blood type. So all that the blood tells us is that the defendant is guilty with probability 0.01%; that's no proof, please strike that». What is the attorney's mistake?

Examples

3. Real cases.
   (1) The prosecutor said that O.J. Simpson already had a history of domestic violence with his wife. The attorney pointed out: «That's awful but only one out of 2500 women subject to domestic violence actually get killed, so this is irrelevant for the murder». The court agreed with the defense; is this correct inference?
   (2) Two babies of the same mother, Sally Clark, died from SIDS; the prosecutor said that the combined probability of two SIDS cases in a row (obtained from single-case statistic) was about 1 in 73 million; is he right?



Bayesian inference for a coin

Problem statement

- Simple inference problem: an unfair coin is flipped N times, producing a sequence of results. We must determine its ”unfairness” and predict the next outcome.

Problem statement

- If we have a probability $p_h = \theta$ for heads (and thus tails probability $p_t = 1 - \theta$), the probability of a sequence s with n heads and m tails is

$$p\left(s|p_h = \theta\right) = \theta^n (1-\theta)^m.$$

- Assume a uniform prior for coin probability before the experiments, meaning no initial preference about θ.
- We now use Bayes' theorem to infer the hidden parameters.

Example of Bayesian inference

- Likelihood: $p\left(\theta|s\right) = \frac{p(s|\theta)p(\theta)}{p(s)}$.
- Here, $p(\theta)$ should be understood as a continuous random variable defined on the interval $[0, 1]$, which it indeed is. Our assumption of a uniform distribution in this case means the prior probability $p(\theta) = 1$, $\theta \in [0, 1]$ (i.e., we have no prior preference about the coin's fairness, assuming all outcomes equally likely). We already know $p\left(s|\theta\right)$.
- Thus, we have:

$$p\left(\theta|s\right) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

Example of Bayesian inference

- Thus, we have:

$$p\left(\theta|s\right) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- $p(s)$ can be calculated as

$$p(s) = \int_0^1 \theta^n(1-\theta)^m d\theta =$$

$$= \frac{\Gamma(n+1)\Gamma(m+1)}{\Gamma(n+m+2)} = \frac{n!m!}{(n+m+1)!},$$

but $\arg\max_\theta p(\theta \mid s) = \frac{n}{n+m}$ can be found without this integral.

Example of Bayesian inference

- Thus, we have:

$$p\left(\theta|s\right) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- But that's not all. To predict the next outcome, we calculate $p\left(\text{heads}|s\right)$:

$$p(\text{heads}|s) = \int_0^1 p(\text{heads}|\theta)p(\theta|s)d\theta =$$

$$= \int_0^1 \frac{\theta^{n+1}(1-\theta)^m}{p(s)}d\theta =$$

$$= \frac{(n+1)!m!}{(n+m+2)!} \cdot \frac{(n+m+1)!}{n!m!} = \frac{n+1}{n+m+2}.$$

- We have obtained Laplace's rule.

Example of Bayesian inference

- Thus, we have:

$$p\left(\theta|s\right) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- This was an illustration of two main tasks in Bayesian inference:

(1) finding the posterior distribution over hypotheses/parameters:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

and/or finding the maximum a posteriori hypothesis $\theta_{\mathrm{MAP}} = \arg\max_\theta p(\theta \mid D)$;

(2) finding the predictive distribution for outcomes of future experiments:

$$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

Conjugate priors

A reminder

- Recall that our main task is to learn the parameters of a distribution and/or to predict its next points from the available data.
- Bayesian inference involves:
  - $p(x \mid \theta)$ – the likelihood of the data;
  - $p(\theta)$ – the prior distribution;
  - $p(x) = \int_\Theta p(x \mid \theta) p(\theta) d\theta$ – the marginal likelihood;
  - $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – the posterior distribution;
  - $p(x' \mid x) = \int_\Theta p(x' \mid \theta) p(\theta \mid x) d\theta$ – the prediction for a new x'.
- The task is usually to find $p(\theta \mid x)$ and/or $p(x' \mid x)$.

Prior distributions

- When we do Bayesian inference, besides the likelihood we must also have a *prior distribution* over all possible values of the parameters.
- We have not paid special attention to them before, but they are very important.
- The task of Bayesian inference is to compute $p(\theta \mid x)$ and/or $p(x' \mid x)$.
- But to do that, we first have to choose $p(\theta)$. How should we choose prior distributions?

Conjugate priors

- A reasonable goal: let us choose the distributions so that they keep the same form *a posteriori*.
- Before inference we have a prior distribution $p(\theta)$.
- After it we have some new posterior distribution $p(\theta \mid x)$.
- I want $p(\theta \mid x)$ to have the same form as $p(\theta)$, just with different parameters.

Conjugate priors

- A (not too formal) definition: a family of distributions $p(\theta \mid \alpha)$ is called a family of *conjugate priors* for a family of likelihoods $p(x \mid \theta)$ if, after multiplying by the likelihood, the posterior $p(\theta \mid x, \alpha)$ stays in the same family: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α are called *hyperparameters* – the "parameters of the distribution of the parameters".
- Trivial example: the family of all distributions is conjugate to anything, but that is not very interesting.

Conjugate priors

- Of course, the form of a good prior depends on the form of the data distribution itself, $p(x \mid \theta)$.
- Conjugate priors have been computed for many distributions; we will give a few examples.

Bernoulli trials

- What is the conjugate prior for tossing an unfair coin (Bernoulli trials)?
- The answer: it is the *Beta distribution*; the density of the coin's bias θ is

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Bernoulli trials

- The density of the coin's bias θ:

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Then, if we sample the coin and obtain s heads and f tails, we get

$$p(s, f \mid \theta) = \binom{s+f}{s}\theta^s(1-\theta)^f, \text{ and}$$

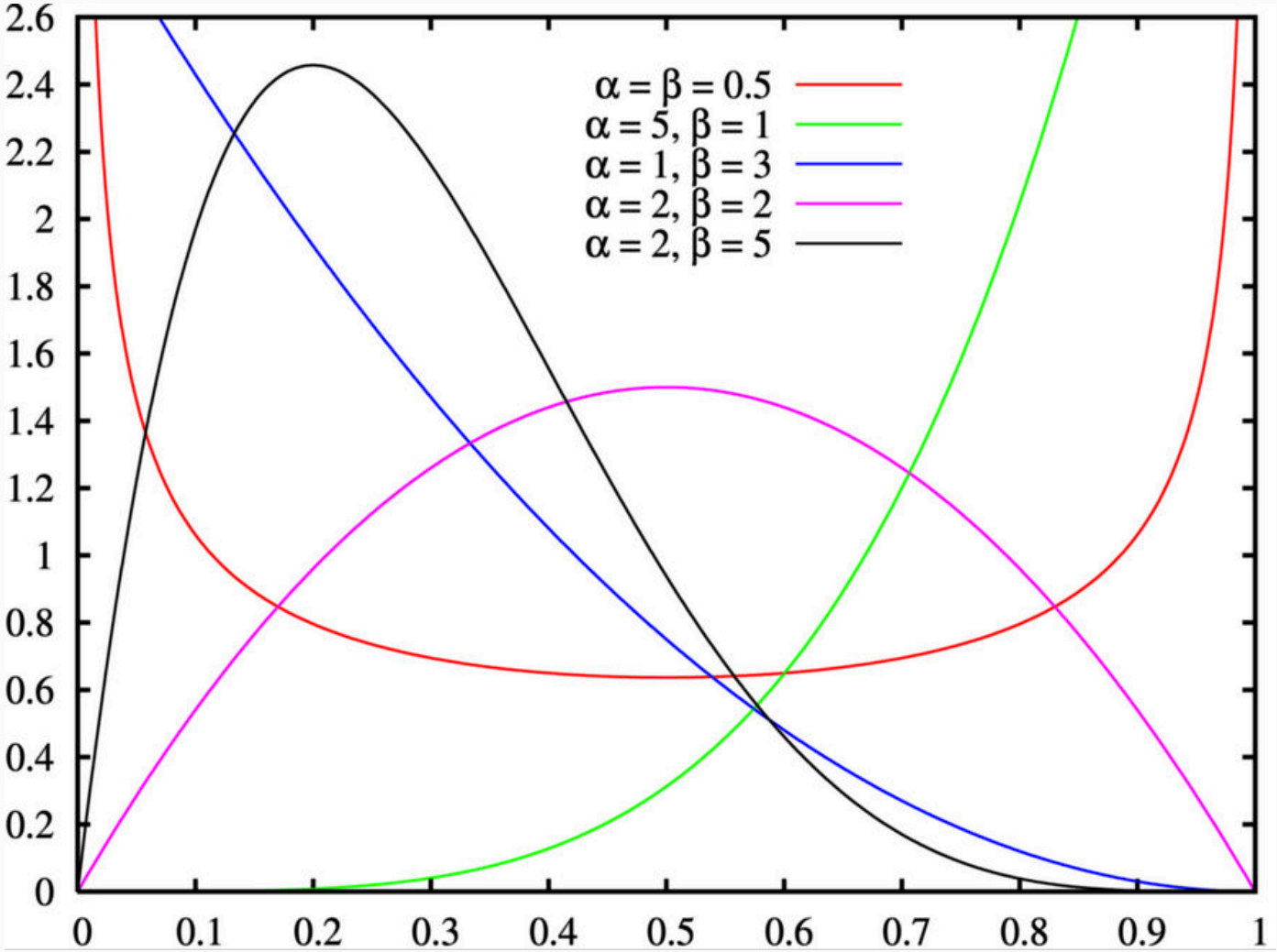
$$p(\theta|s, f) = \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s}x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha, \beta)dx} =$$

$$= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}.$$

Bernoulli trials

- So it turns out that the conjugate prior for the bias θ of an unfair coin is
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$
- After receiving new data with s heads and f tails, the hyperparameters change to
$$p(\theta \mid s+\alpha, f+\beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$
- At this point you can forget the complicated formulas and derivations – we got a very simple learning rule (where "learning" now means updating the hyperparameters).

The Beta distribution



Multinomial distribution

- A simple generalization: consider a multinomial distribution with n trials and k categories, where x_i of the experiments fell into category i.
- The parameters θ_i are the probabilities of falling into category i:

$$p(x \mid \theta) = \binom{n}{x_1, \ldots, x_n} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_k^{x_k}.$$

- The conjugate prior is the Dirichlet distribution:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}.$$

Multinomial distribution

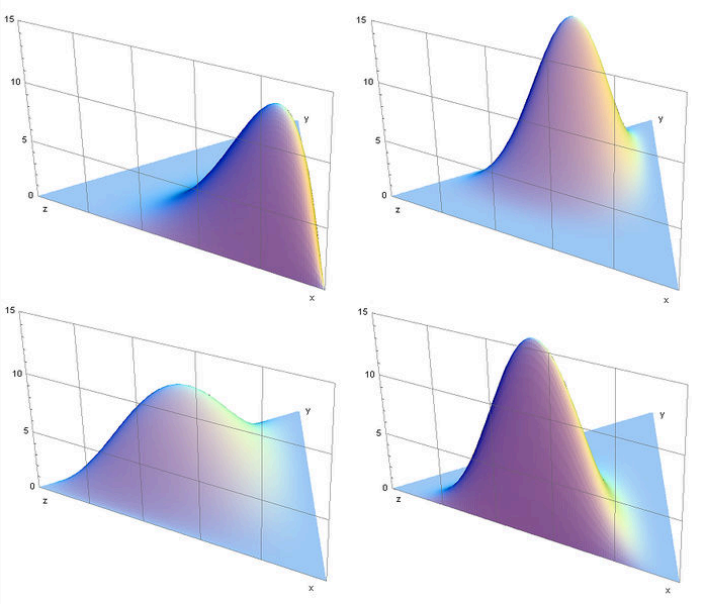
- The conjugate prior is the Dirichlet distribution:

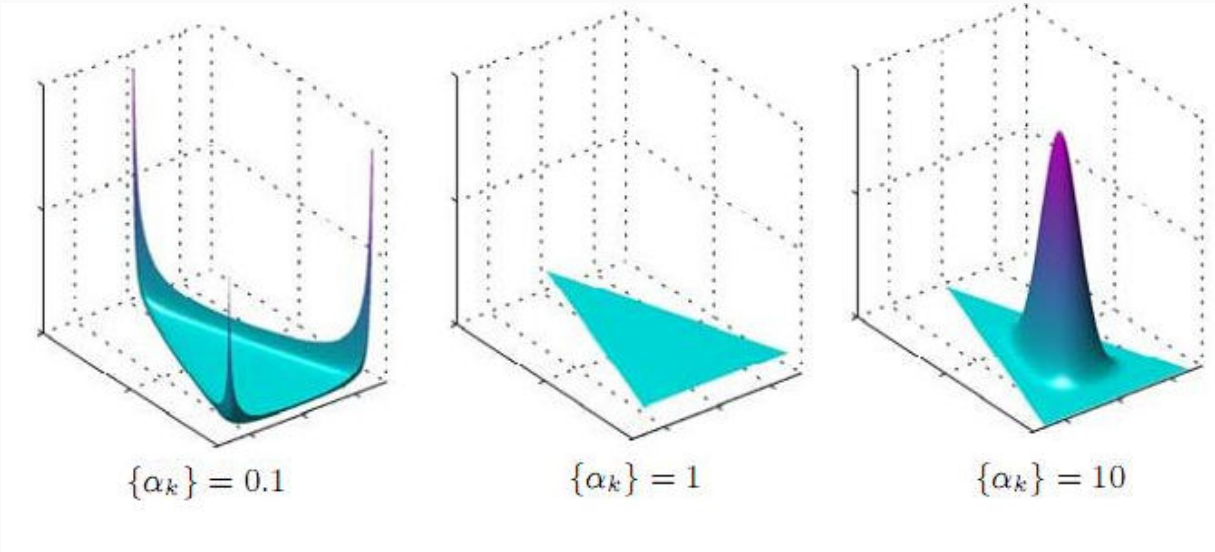
$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Exercise. Show that upon receiving data $x_1, \dots, x_k$ the hyperparameters change to

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1}.$$

The Dirichlet distribution





Thank you!

Thank you for your attention!