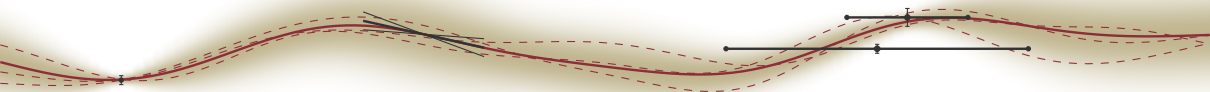
Probabilistic Machine Learning
Lecture 01
Introduction

Philipp Hennig

20 April 2020

Eberhard Karls
Universität Tübingen

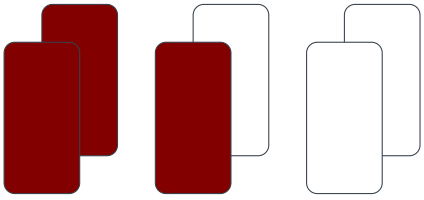
Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



Which Card?

an opening experiment

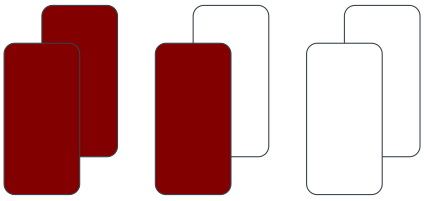
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Which Card?

an opening experiment

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



- $\frac{1}{2}$
- $2/3$
- something else
- I don't know yet

An *inference* problem requires statements about the value of an *unobserved* (latent) variable x based on observations y which are **related** to x, but may not be sufficient to fully determine x. This requires a notion of **uncertainty**.

Life is Uncertain

the ability to reason under uncertainty is the hallmark of intelligence

Life is Uncertain

the ability to reason under uncertainty is the hallmark of intelligence

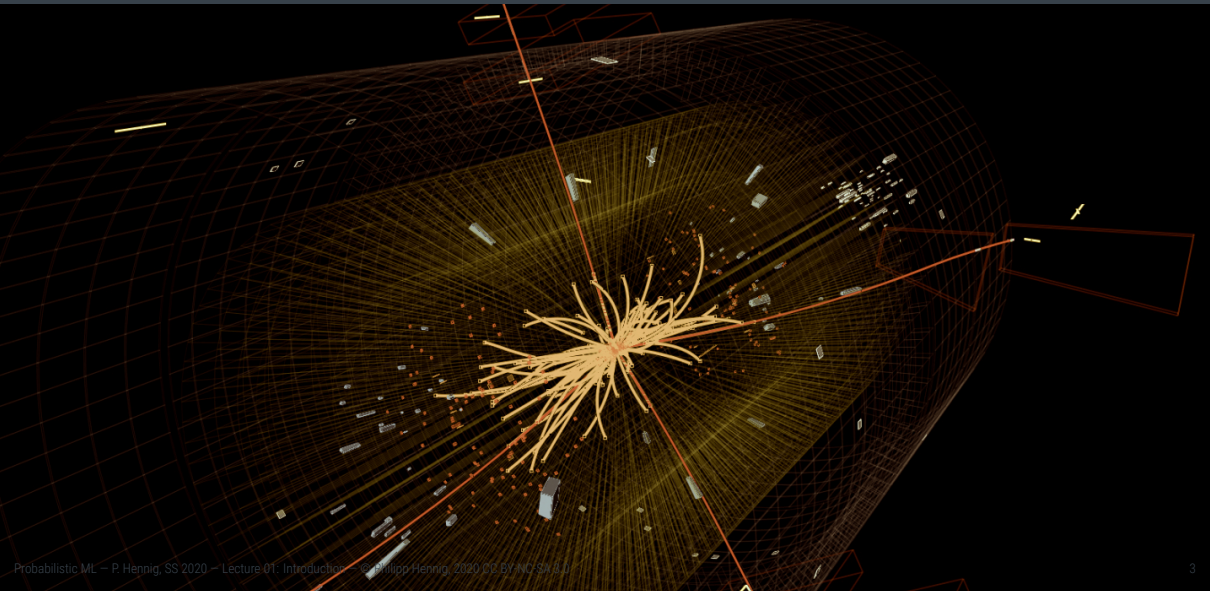
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

[ICJ den Haag, 2006]

Life is Uncertain

the ability to reason under uncertainty is the hallmark of intelligence

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



[© CERN / CMS Collaboration, 2011]

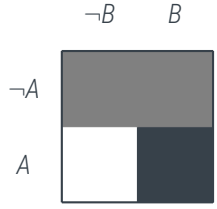
Life is Uncertain

the ability to reason under uncertainty is the hallmark of intelligence

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

[Luke Fildes, 1891 (Tate Modern)]

A = "It rains."
B = "The street is wet."



Deductive Reasoning: $A \Rightarrow B$ if A is true, then B is true

A is true	thus	B is true	modus ponens
B is false	thus	A is false	modus tollens

Plausible Reasoning: $P(B \mid A) > P(B)$ if A is true, the B becomes more plausible

A is true	thus	B becomes more plausible
B is false	thus	A becomes less plausible
B is true	thus	A becomes more plausible
A is false	thus	B becomes less plausible

The actual science of logic is conversant at present only with things either certain, impossible, or entirely doubtful, none 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.

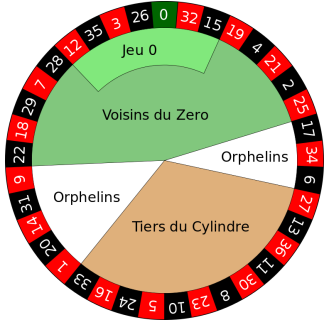
James Clerk Maxwell (1831–1879)

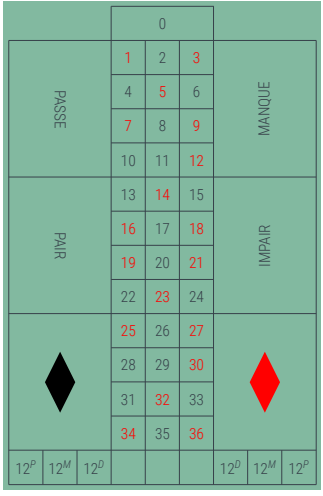
[image source: BBC / public domain]

Goals of this course:

- Establish a *formal framework* for **probable reasoning**
- Use it to build *powerful* inference mechanisms for **real-world problems**
- Develop the *technical tools* necessary to implement **inference** in practice

[image: wikipedia / Ann-Kathrin Schalkamp]





[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

§ 1. Axiome[2].

Es sei E eine Menge von Elementen $\xi, \eta, \zeta, \ldots$, welche man *elementare Ereignisse* nennt, und $\mathfrak{F}$ eine Menge von Teilmengen aus E; die Elemente der Menge $\mathfrak{F}$ werden weiter *zufällige Ereignisse* genannt.

I. $\mathfrak{F}$ *ist ein Mengenkörper*[3].

II. $\mathfrak{F}$ *enthält die Menge* E.

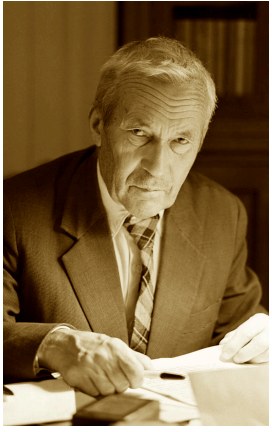
III. *Jeder Menge* A *aus* $\mathfrak{F}$ *ist eine nichtnegative reelle Zahl* $\mathsf{P}(A)$ *zugeordnet. Diese Zahl* $\mathsf{P}(A)$ *nennt man die Wahrscheinlichkeit des Ereignisses* A.

IV. $\mathsf{P}(E) = 1$.

V. *Wenn* A *und* B *disjunkt sind, so gilt*

$$\mathsf{P}(A + B) = \mathsf{P}(A) + \mathsf{P}(B).$$

Ein Mengensystem $\mathfrak{F}$ mit einer bestimmten Zuordnung der Zahlen $\mathsf{P}(A)$, welche den Axiomen I—V genügt, nennt man ein *Wahrscheinlichkeitsfeld*.



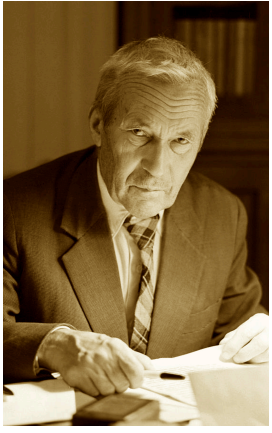
Andrey N. Kolmogorov
(1903–1987)

Kolmogorov's Axioms

Plausibility as a Measure

[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

[3] Vgl. HAUSDORFF: Mengenlehre 1927 S. 78. Ein Mengensystem heißt ein Körper, wenn Summe Durchschnitt und Differenz von zwei Mengen des Systems wieder dem System angehören. Jeder nicht leere Mengenkörper enthält die Nullmenge 0. Wir bezeichnen mit HAUSDORFF den Durchschnitt von A und B mit AB, die Vereinigungsmenge von A und B im Falle $AB = 0$ mit $A + B$, allgemein aber mit $A \dotplus B$, und die Differenz von A und B mit $A - B$. Das Komplement $E - A$ der Menge A wird durch A bezeichnet. Die elementaren Rechengesetze für Mengen und ihre Durchschnitte, Summen und Differenzen werden weiter als bekannt vorausgesetzt. Mengen aus $\mathfrak{F}$ werden weiter mit großen lateinischen Buchstaben bezeichnet.



Andrey N. Kolmogorov
(1903–1987)

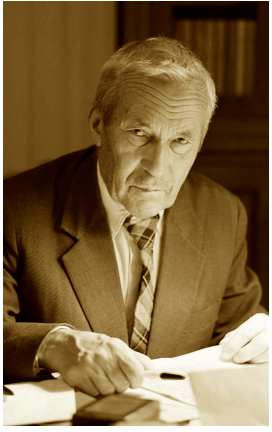
E is the roulette wheel
F is the table

Definition (σ-algebra, measurable sets & spaces)

Let E be a space of *elementary events*. Consider the power set 2^E, and let $\mathfrak{F} \subset 2^E$ be a set of subsets of E. Elements of $\mathfrak{F}$ are called *random events*. If $\mathfrak{F}$ satisfies the following properties, it is called a **σ-algebra**.

1. $E \in \mathfrak{F}$ II.
2. $(A, B \in \mathfrak{F}) \Rightarrow (A - B \in \mathfrak{F})$ F is closed under the difference I.
3. $(A_1, A_2, \cdots \in \mathfrak{F}) \Rightarrow (\bigcup_{i=1}^{\infty} A_i \in \mathfrak{F} \quad \wedge \quad \bigcap_{i=1}^{\infty} A_i \in \mathfrak{F})$ I.

(this implies $\varnothing \in \mathfrak{F}$. If E is countable, then 2^E is a σ-algebra). If $\mathfrak{F}$ is a σ-algebra, its elements are called **measurable sets**, and $(E, \mathfrak{F})$ is called a **measurable space** (or **Borel space**).



Andrey N. Kolmogorov
(1903–1987)

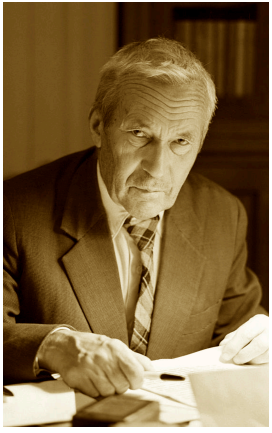
Definition (Measure & Probability Measure)

Let $(E, \mathfrak{F})$ be a **measurable space** (aka. Borel space). A nonnegative real function $P : \mathfrak{F} \to \mathbb{R}_{0,+}$ (III.) is called a **measure** if it satisfies the following properties:

1. $P(\varnothing) = 0$
2. For any countable sequence $\{A_i \in \mathfrak{F}\}_{i=1,\ldots,}$ of pairwise disjoint sets ($A_i \cap A_j = \varnothing$ if $i \neq j$), P satisfies **countable additivity** (aka. **σ-additivity**):

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \qquad \text{(V.)}$$

The measure P is called a **probability measure** if $P(E) = 1$. IV.
(For probability measures, 1. is unnecessary). Then, $(E, \mathfrak{F}, P)$ is called a **probability space**.



Andrey N. Kolmogorov
(1903–1987)

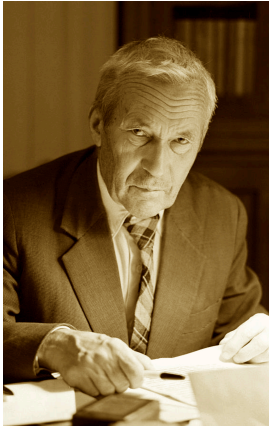
Theorem (Sum Rule)

From $A + \neg A = E$ we get

$$P(A) + P(\neg A) \overset{(V)}{=} P(E) \overset{(IV)}{=} 1, \quad \textit{thus} \quad P(A) = 1 - P(\neg A).$$

And from $A = A \cap (B + \neg B)$, using the notation $P(A, B) = P(A \cap B)$ for the ***joint probability*** *of A and B, we get the* ***Sum Rule***

$$P(A) \overset{(V)}{=} P(A, B) + P(A, \neg B).$$



Andrey N. Kolmogorov
(1903–1987)

Definition (Conditional Probability)

If $P(A) > 0$, the quotient

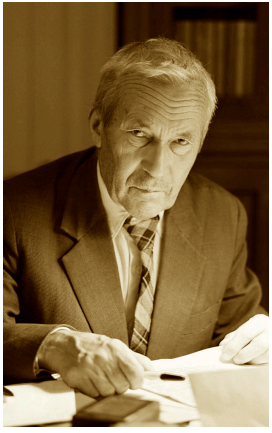
$$P(B \mid A) = \frac{P(A, B)}{P(A)}$$

is called the **conditional probability** of B given A. It immediately gives

$$P(A, B) = P(B \mid A)P(A) = P(A \mid B)P(B).$$

It is easy to show that $P(B \mid A) \geq 0$ (III), $P(E \mid A) = 1$ (IV), and for $B \cap C = \varnothing$, we have $P(B + C \mid A) = P(B \mid A) + P(C \mid A)$ (V). Thus, for a fixed A, $(E, \mathfrak{F}, P(\cdot \mid A))$ is a probability space.

Note that $P(A \mid A) = \frac{P(A \cap A)}{P(A)} = 1.$



Andrey N. Kolmogorov (1903–1987)

Theorem (Law of Total Probability)

Let $A_1 + A_2 + \cdots + A_n = E$ and $A_i \cap A_j = \varnothing$ if $i \neq j$. Then, for any $X \in \mathfrak{F}$,

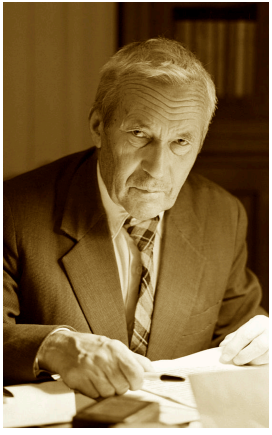
$$P(X) = \sum_{i=1}^{n} P(X \mid A_i) P(A_i).$$

Proof.

Because $X = E \cap X = \bigcup_{i=1}^{n} (A_i \cap X)$, we get from V that

$$P(X) = \sum_{i=1}^{n} P(A_i, X) \overset{def.}{=} \sum_{i=1}^{n} P(X \mid A_i) P(A_i).$$

□



Andrey N. Kolmogorov (1903–1987)

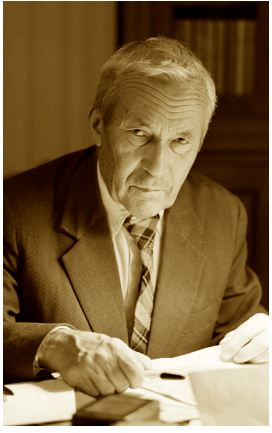
Theorem (Bayes' Theorem)

Let $A_1 + A_2 + \cdots + A_n = E$ and $A_i \cap A_j = \varnothing$ if $i \neq j$. Then, for any $X \in \mathfrak{F}$,

$$P(A_i \mid X) = \frac{P(A_i)P(X \mid A_i)}{\sum_{j=1}^{n} P(A_j)P(X \mid A_j)}.$$

Proof.

Apply the Sum Rule to the definition of the conditional probability. □



Andrey N. Kolmogorov
(1903–1987)

The Rules of Probability:

- the **Sum Rule:**

$$P(A) = P(A, B) + P(A, \neg B)$$

- the **Product Rule:**

$$P(A, B) = P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$$

- **Bayes' Theorem:**

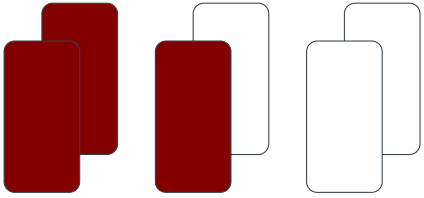
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B, A) + P(B, \neg A)}$$

likelihood is when we have a distribution that is a function of a variable (X), which is being conditioned -> P(D|X)

$$\underbrace{P(X \mid D)}_{\text{posterior for } X \text{ given } D} = \frac{\overbrace{P(X)}^{\text{prior for } X} \cdot \overbrace{P(D \mid X)}^{\text{likelihood for } X}}{\underbrace{P(D)}_{\text{evidence for the model}}} = \frac{P(X) \cdot P(D \mid X)}{\sum_{X \in \mathcal{X}} P(D \mid x) P(x)}$$

Bayes' Theorem tells us how to update the *belief* in a *hypothesis X* when observing *data D*.

- $P(D \mid X)$ is *the likelihood of X*, but the *(conditional) probability for D (given X)*
- the **model** is the entire thing – prior *and* likelihood
- despite the name, the prior is not necessarily what you know *before* seeing the data, but the marginal distribution $P(X) = \sum_{d \in \mathcal{D}} P(X, d)$ under *all* possible data.



- $\frac{1}{2}$
- $2/3$
- something else
- I don't know yet

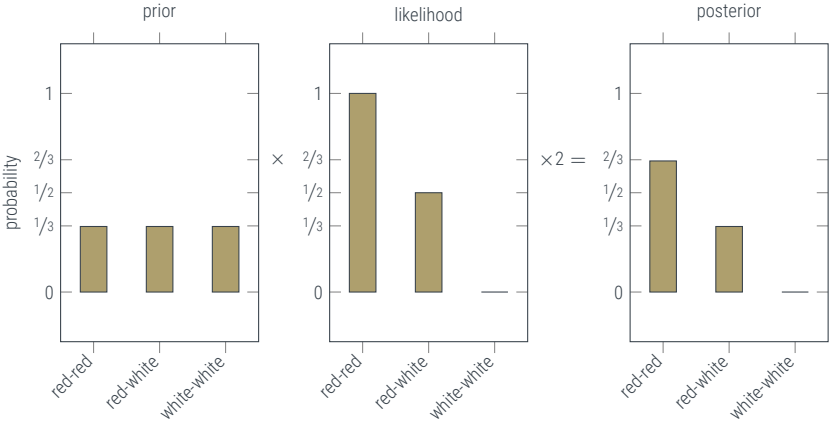
$$P(\text{card}|\text{color}) = \frac{P(\text{card}) \cdot P(\text{color} \mid \text{card})}{P(\text{color})} = \frac{P(\text{card}) \cdot P(\text{color} \mid \text{card})}{\sum_{i=1}^{3} P(\text{color} \mid \text{card} = i) P(\text{card} = i)}$$

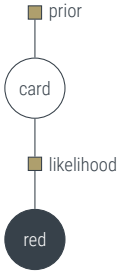
$$= \frac{1/3 \cdot P(\text{color} \mid \text{card})}{1/2} = \frac{2}{3} \cdot \{1 \quad 1/2 \quad 0\}$$

Note: It's all about the likelihood, not the prior!

Bayes' Theorem Appreciation Slides (1)

reasoning quantitatively under uncertainty





Plausible Reasoning, Revisited

have we succeeded in formalizing common sense?

A = "it will begin to rain by 6pm"
B = "the sky will become cloudy before 6pm"

Homework!

$A \Rightarrow B$
if A is true, the B is true

Assume: if A is true, then B is true ($A \Rightarrow B$)

A is true	thus	B is true	(modus ponens)
B is false	thus	A is false	(modus tollens)
B is true	thus	A becomes more plausible	
A is false	thus	B becomes less plausible	

if A is true, B becomes more plausible ($P(B \mid A) > P(B)$)

A is true	thus	B becomes more plausible
B is false	thus	A becomes less plausible
B is true	thus	A becomes more plausible
A is false	thus	B becomes less plausible

Bayes' Theorem Appreciation Slides (2)

Bayesian reasoning can help in unintuitive situations

T - test result
C - coronavirus immunity

$$P(T \mid C) = 0.938 \qquad P(\neg T \mid \neg C) = 0.96$$

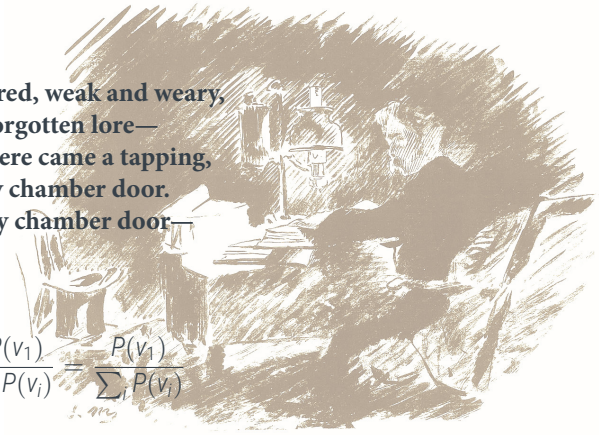
$$\begin{aligned} P(C \mid T) &= \frac{P(T \mid C)P(C)}{P(T \mid C)P(C) + P(T \mid \neg C)P(\neg C)} \\ &= \frac{P(T \mid C)P(C)}{P(T \mid C)P(C) + (1 - P(\neg T \mid \neg C))(1 - P(C))} \\ &= \frac{0.938 \cdot P(C)}{0.938P(C) + 0.04(1 - P(C))} \\ &= \frac{0.938 \cdot P(C)}{0.898P(C) + 0.04} \end{aligned}$$

- $P(C) = 0.01 \rightarrow P(C \mid T) = \boxed{0.19}$

Bayes' Theorem Appreciation Slides (3)

Bayesian reasoning matches human reasoning

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]



Once upon a midnight dreary, while I pondered, weak and weary,
Over many a quaint and curious volume of forgotten lore—
While I nodded, nearly napping, suddenly there came a tapping,
As of some one gently rapping, rapping at my chamber door.
"'Tis some visitor," I muttered, "tapping at my chamber door—
Only this and nothing more."

t - tapping the door
v1 - visitor 1

$$P(v_1 \mid t) = \frac{P(t \mid v_1) \cdot P(v_1)}{P(t)} = \frac{P(t \mid v_1) \cdot P(v_1)}{P(t \mid v_1) \sum_i P(v_i)} = \frac{P(v_1)}{\sum_i P(v_i)}$$

If the likelihood is constant, data does not help.

Bayes' Theorem Appreciation Slides (3)

rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

Ah, distinctly I remember it was in the bleak December;
And each separate dying ember wrought its ghost upon the floor.
Eagerly I wished the morrow;—vainly I had sought to borrow
From my books surcease of sorrow—sorrow for the lost Lenore—
For the rare and radiant maiden whom the angels name Lenore—
Nameless here for evermore.

$$P(\ell \mid t) = \frac{P(t \mid \ell) \cdot P(\ell)}{P(t)} = \frac{P(t \mid \ell) \cdot P(\ell)}{P(t \mid \ell)P(\ell) + \sum_{i \neq \ell} P(t \mid v_i)P(v_i)}$$
$$\lesssim 1$$

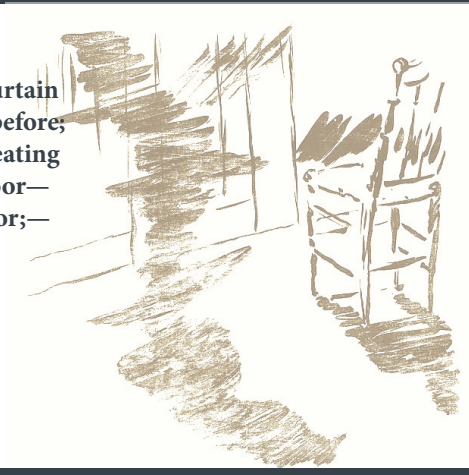
A very unlikely hypothesis can become dominant if it is the only one explaining the data well.

Bayes' Theorem Appreciation Slides (3)

rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

And the silken, sad, uncertain rustling of each purple curtain
Thrilled me—filled me with fantastic terrors never felt before;
So that now, to still the beating of my heart, I stood repeating
"'Tis some visitor entreating entrance at my chamber door—
Some late visitor entreating entrance at my chamber door;—
This it is and nothing more."



$$P(\ell \mid t) = \frac{P(t \mid \ell) \cdot P(\ell)}{P(t)} = \frac{P(t \mid \ell) \cdot 0}{P(t)}$$
$$= 0$$

No data can revive an a priori impossible hypothesis

Bayes' Theorem Appreciation Slides (3)

rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

Presently my soul grew stronger; hesitating then no longer,
"Sir," said I, "or Madam, truly your forgiveness I implore;
But the fact is I was napping, and so gently you came rapping,
And so faintly you came tapping, tapping at my chamber door,
That I scarce was sure I heard you"—here I opened wide the door;—
Darkness there and nothing more.

$$P(v) = 0$$

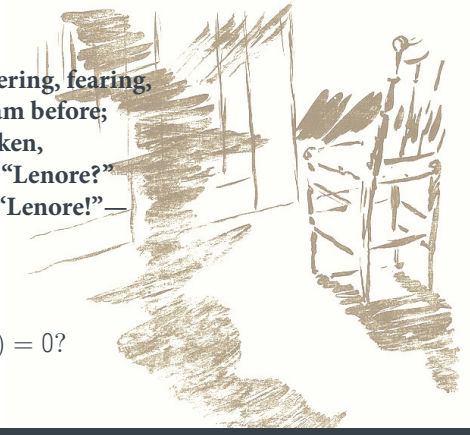
Additional evidence may force you to reconsider your prior.

Bayes' Theorem Appreciation Slides (3)

rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

Deep into that darkness peering, long I stood there wondering, fearing,
Doubting, dreaming dreams no mortal ever dared to dream before;
But the silence was unbroken, and the stillness gave no token,
And the only word there spoken was the whispered word, "Lenore?"
This I whispered, and an echo murmured back the word, "Lenore!"—
Merely this and nothing more.



$$P(v) = 0 \qquad P(\ell) = 0 \qquad P(t) = \sum_i P(t \mid v_i) P(v_i) = 0?$$

$P(E) = 1$. The hypothesis space has to contain *some* explanation for the data.

Bayes' Theorem Appreciation Slides (3)

rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

Back into the chamber turning, all my soul within me burning,
Soon again I heard a tapping somewhat louder than before.
"Surely," said I, "surely that is something at my window lattice;
Let me see, then, what thereat is, and this mystery explore—
Let my heart be still a moment and this mystery explore;—
'Tis the wind and nothing more!"

$$P(w \mid t) = \frac{P(t \mid w) \cdot P(w)}{P(t)} = \frac{P(t \mid w) \cdot P(w)}{P(t, w) + \sum_i P(t, v_i)}$$

The σ-algebra itself, rather than P, is often the most important prior assumption.

Bayes' Theorem Appreciation Slides (3)

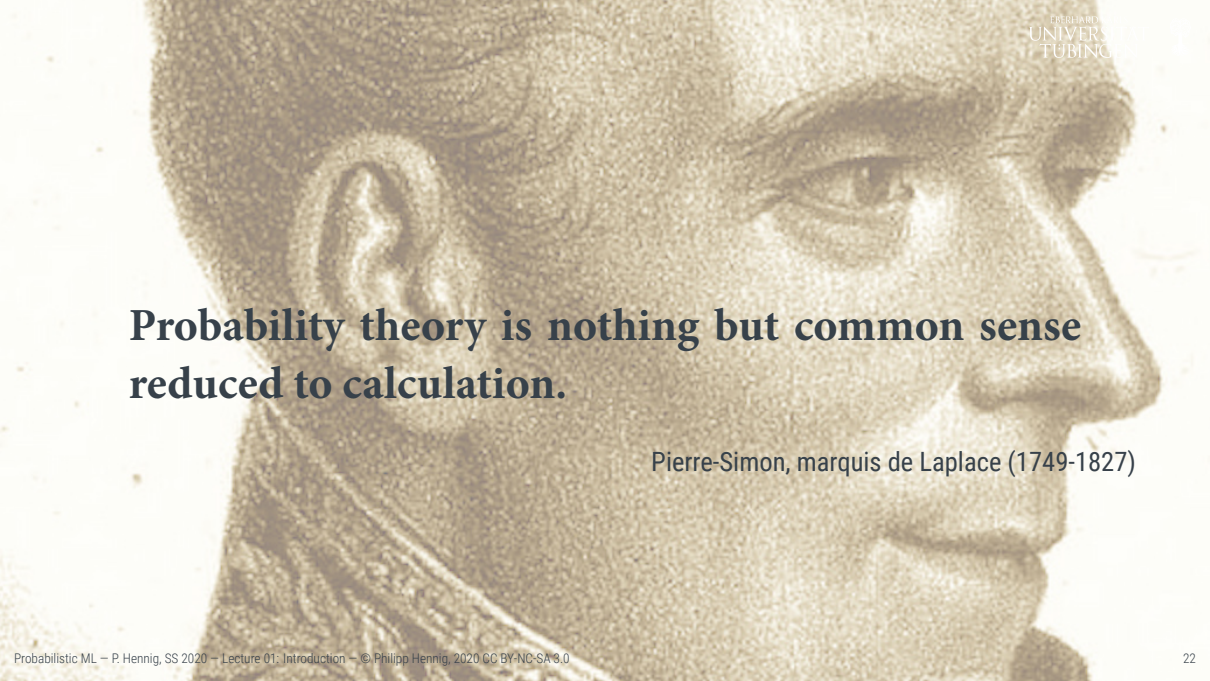
rationality and nothing more

[Edgar Allan Poe, 1845 / Édouard Manet, 1875]

Open here I flung the shutter, when, with many a flirt and flutter,
In there stepped a stately Raven of the saintly days of yore;
Not the least obeisance made he; not a minute stopped or stayed he;
But, with mien of lord or lady, perched above my chamber door—
Perched upon a bust of Pallas just above my chamber door—
Perched, and sat, and nothing more.

$$P(r \mid t) = \frac{P(t \mid r) \cdot P(r)}{P(t)} = \frac{P(t \mid r) \cdot P(r)}{P(t, r) + P(t, w) + \sum_i P(t, v_i)}$$

Probabilistic reasoning is a mechanism, it does not replace creativity.



Probability theory is nothing but common sense reduced to calculation.

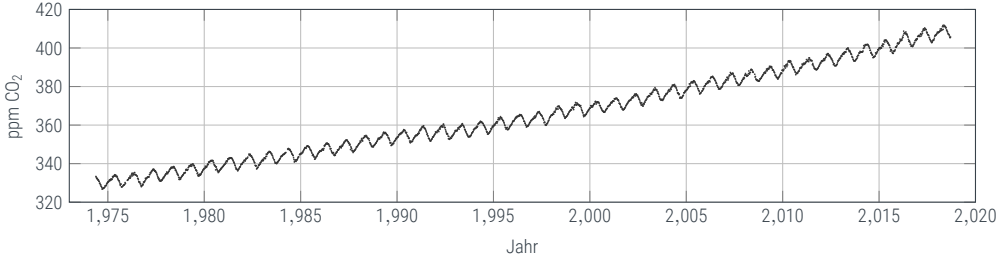
Pierre-Simon, marquis de Laplace (1749-1827)

Administrative Stuff

Let's meet!

- *Flipped Classroom* on **Tuesday, 21 April**, 10ct-12
- sign-in will be distributed via Ilias
- sign up on Ilias!
- please provide feedback on Ilias

`https://uni-tuebingen.de/de/134452`



- Goal: Reflect on the utility and power of deep learning (and what you know about it)
- Things to think about:
  - How do you include external structural knowledge?
  - Do you trust the prediction? Why, or why not?
  - Can your model do anything other than predict future data?

**To be able to reason in an uncertain world,
whether to build intelligent machines or find scientific insight,
you have to understand probabilities.**

- both philosophical (Cox) and formal (Kolmogorov) arguments lead to the same two rules:

  sum rule: $P(X) = P(X, Y) + P(X, \neg Y)$

  product rule: $P(X, Y) = P(X) \cdot P(Y \mid X)$

- their corollary, **Bayes' Theorem** provides the mechanism for **inference**:

$$\underbrace{P(X \mid D)}_{\text{posterior of } X \text{ given } D} = \frac{\overbrace{P(D \mid X)}^{\text{likelihood of } X \text{ under } D} \cdot \overbrace{P(X)}^{\text{prior of } X}}{\underbrace{P(D)}_{\text{evidence for the model}}}$$

- the fundamentality of probabilities has been debated at length. Probabilities are not the only inference system, but they are uniquely **general, expressive, and powerful**.
- **Machine learning and AI** can be approached in various ways. The probabilistic viewpoint is the closest we have to a **theory of everything** for ML.