https://attendance.lincoln.ac.uk

Access Code: **662942**

CMP9794M
Advanced Artificial Intelligence

Heriberto Cuayahuitl

UNIVERSITY OF LINCOLN

School of Computer Science

Delivery Team

Dr. Heriberto Cuayahuitl
hcuayahuitl@lincoln.ac.uk
office: INB3108

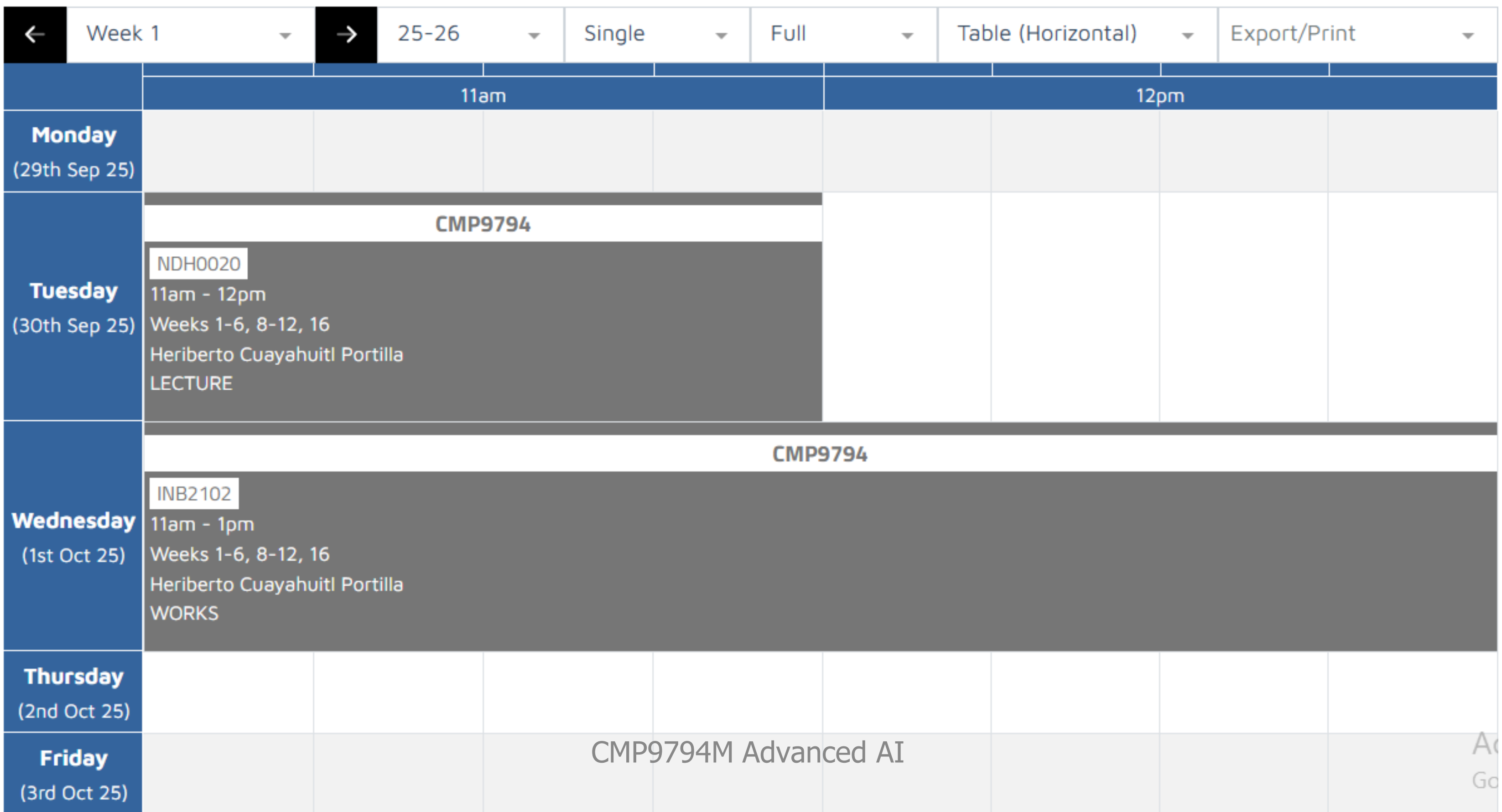
- British computer scientist (born in Mexico)
- PhD from the University of Edinburgh in 2009
- Areas of interest:
  - Conversational AI
  - Robot Learning

Recommendations

- Record your attendance in every session:
  - online https://attendance.lincoln.ac.uk
  - paper-based register (when online system is off)
- Check Blackboard and your email frequently for updates and notifications
- Mute your mobile phone during lectures
- If you want to ask/say something, raise your hand or just interrupt your lecturer (that’s okay)

Lectures and Workshops

- Lectures: Tueday 11am-12hrs in NDH0020
- Workshops: Wednesday 11am-1pm in INB2102



Main Topics in this Module

- **Quantifying uncertainty**
  - Introduction to probability theory
- **Probabilistic reasoning**
  - Bayesian Nets (BNs), Gaussian Processes (GPs)
- **Reasoning over time**
  - Dynamic BNs, Hidden Markov Models (HMMs)
- **Making complex decisions**
  - Markov Decision Processes (MDPs)
  - Reinforcement learning (without neural nets)

Agenda

Week	Commencing on	Topic	Delivered by
1	29/09/2024	Introduction to module	Heriberto Cuayahuitl
2	06/10/2024	Bayesian networks w/exact inference	Heriberto Cuayahuitl
3	13/10/2024	Structure learning	Heriberto Cuayahuitl
4	20/10/2024	Gaussian Bayesian networks	Heriberto Cuayahuitl
5	27/10/2024	Gaussian processes I—Exact Inference	Heriberto Cuayahuitl
6	03/11/2024	Gaussian processes II—Approx. Inference	Heriberto Cuayahuitl
7	10/11/2024	Reading week (no lectures/workshops)	
8	17/11/2024	Bayesian networks w/approx. inference	Heriberto Cuayahuitl
9	24/11/2024	Probabilistic reasoning over time	Heriberto Cuayahuitl
10	01/12/2024	Complex decision making	Heriberto Cuayahuitl
11	08/12/2024	Markov decision processes	Heriberto Cuayahuitl
12	15/12/2024	Reinforcement learning	Heriberto Cuayahuitl
13-15		Christmas break	
16	12/01/2025	Ethics and Agentic AI	Heriberto Cuayahuitl

Learning Objectives

LO1. Critically appraise a range of AI techniques for knowledge representation, reasoning and decision-making **under uncertainty**, identifying their strengths and weaknesses, and selecting appropriate methods to serve particular roles

LO2. Design and develop software algorithms for solving complex (non-trivial) AI problems in an application domain of interest.

Assessments*

- Assignment (50%):
  - Bayesian Networks (BNs)
  - Gaussian BNs
  - Gaussian Processes

- In-class test (50%):
  - Mock-test (last workshop)
  - Test (see Hand-in deadlines and your timetable)

*You should **READ** the Assessment docs in Blackboard

What is AI?

Many definitions. Example “the science and engineering for equipping machines/robots to acquire their own behaviour”.

Thinking humanly	**Thinking rationally**
Acting humanly	**Acting rationally**

See Russell & Norving, 2016. Chap. 1

Thinking Humanly: Cognitive Modelling

- We must have a way of determining how humans think:
  - Introspection (catching our own thoughts as they happen)
  - Psychological experiments (observing a person in action)
  - Brain imaging (observing brain in action)
- Cognitive science brings multiple fields together to try to construct precise and testable theories of the workings of the human mind.

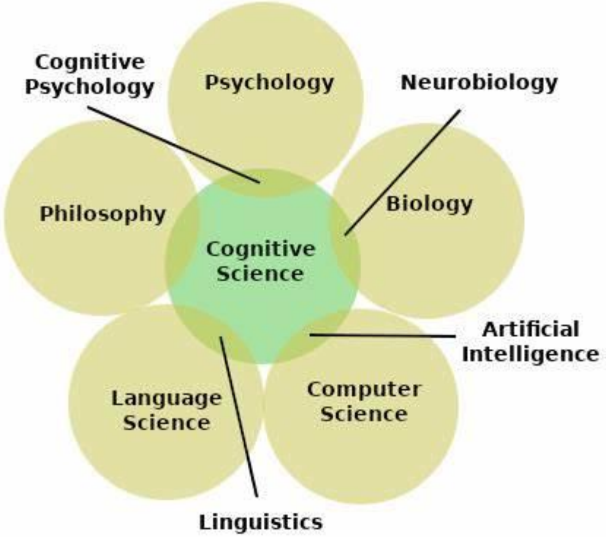
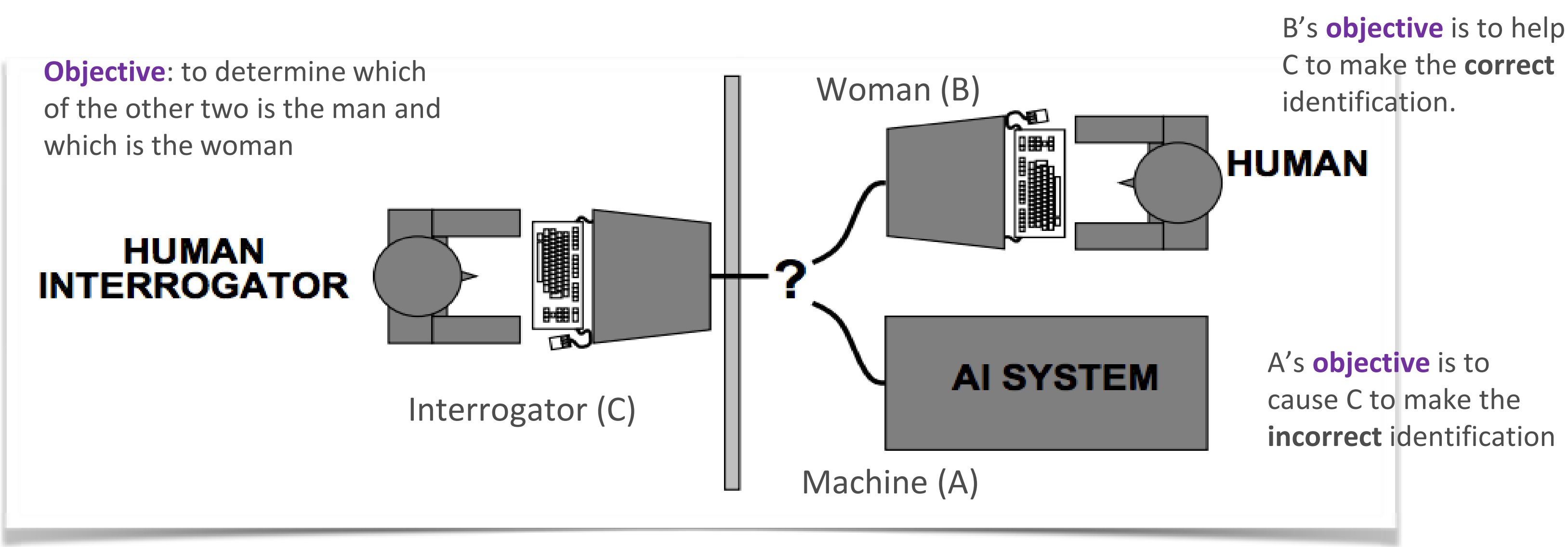


Image from: [Psychological Theory Definition: An Overview (mantracare.org)]

Acting Humanly: The Turing Test (1950)

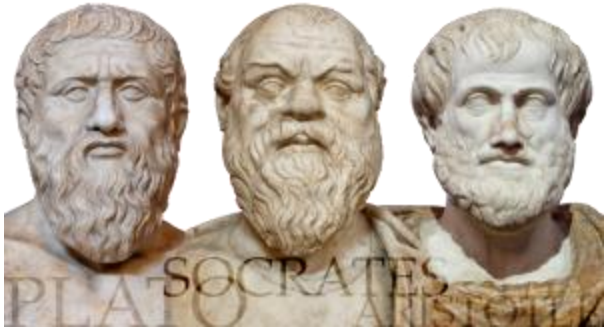
- Turing (1950) “Computing machinery and intelligence”
- Can machines behave intelligently?
- [Summary of Alan Turing’s paper (1950)]



https://www.csee.umbc.edu/courses/471/papers/turing.pdf
TURING TEST : Your Complete Guide 2020 (professional-ai.com)

Thinking Rationally: Laws of Thought

- Several Greek scholars developed various forms of logic: **notation** and **rules** for thoughts



- AI hopes to create intelligent systems using logic programming (e.g., Prolog).
- However, it is not easy to represent informal knowledge using logical notation, particularly when knowledge is **not 100% certain**.

Acting Rationally

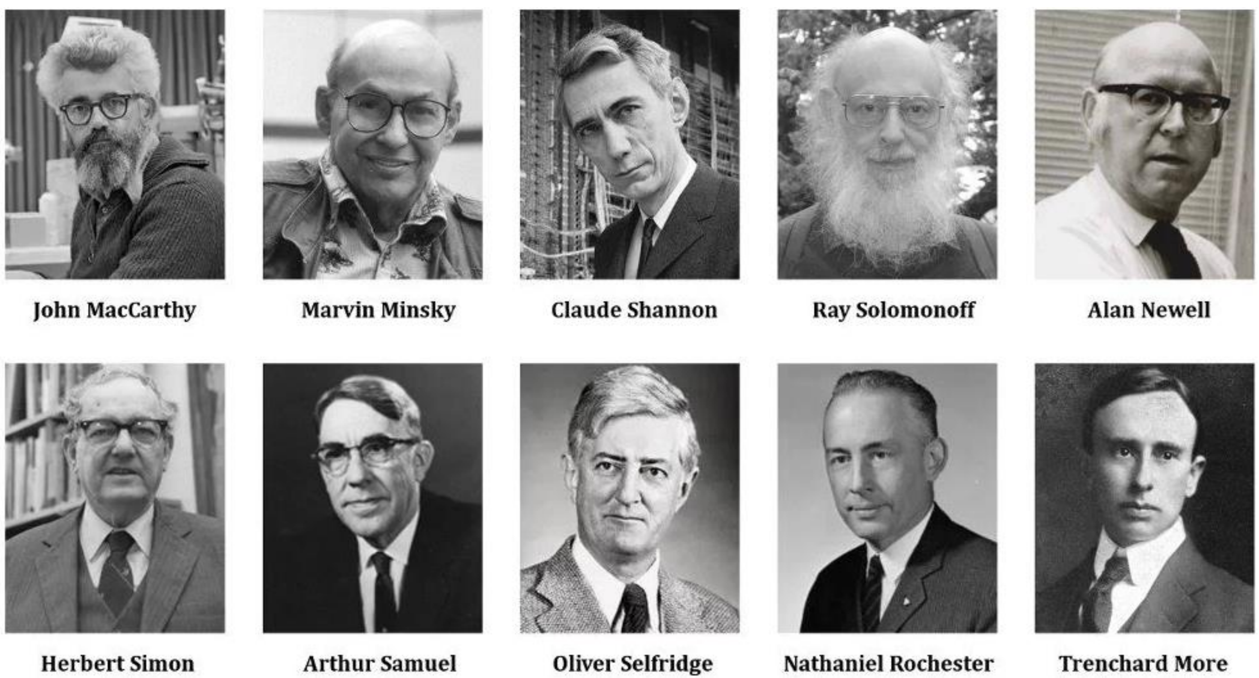
- **Rational behaviour**: doing the right thing

- **The right thing**: which is expected to maximise goal achievement – given the available information to the AI.

- **Rational Agent** is one that achieve the best outcome or, when there is **uncertainty**, the best expected outcome.

Birth of AI

- The Dartmouth Conference (1956) brought together researchers in a variety of topics:
  - complexity theory, language simulation, neuron nets, abstraction of content from sensory inputs, relationship of randomness to creative thinking, learning machines



Picture from https://www.scienceabc.com

Pillars of AI

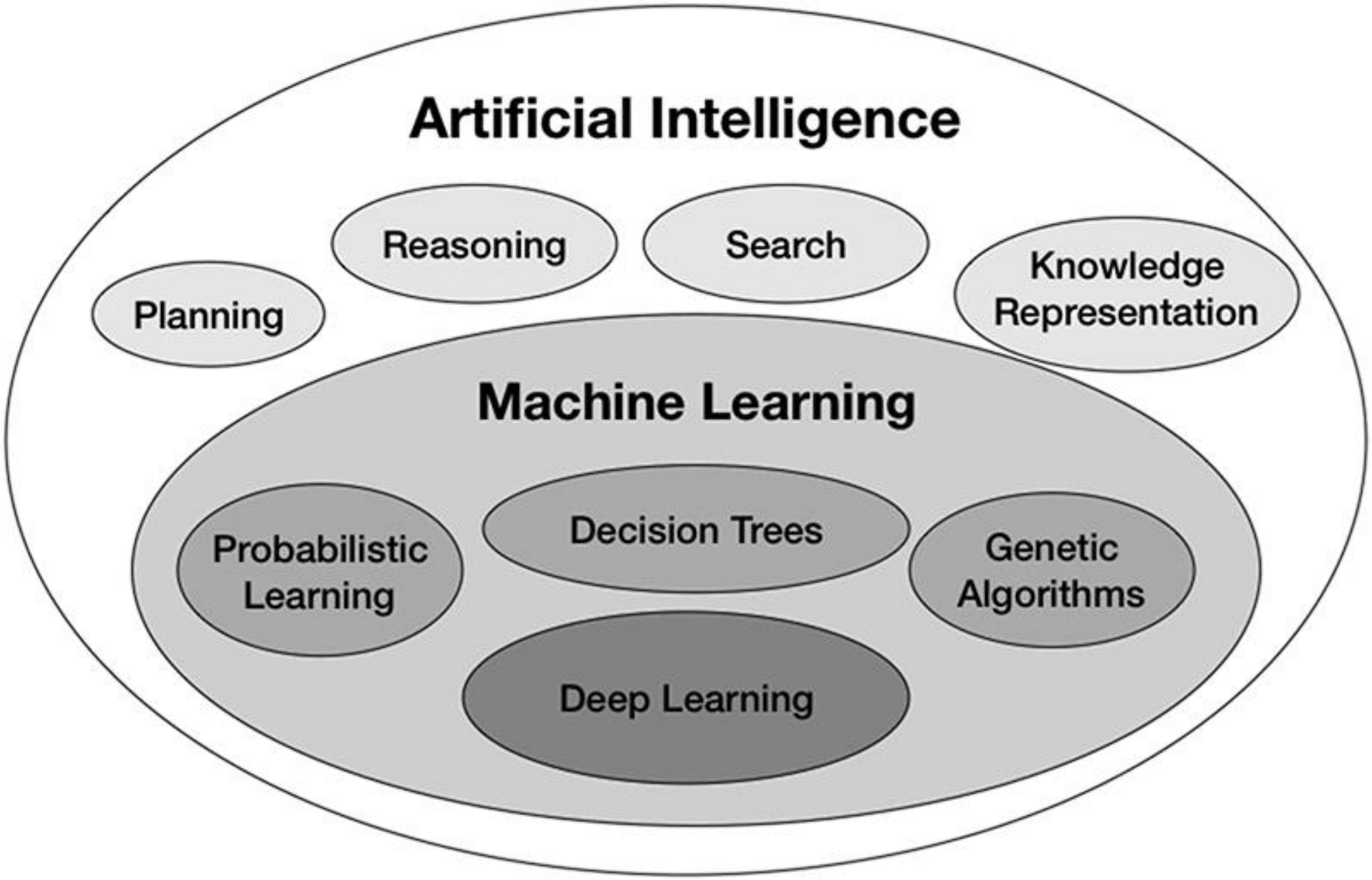


Image from
https://behavioralscientist.org

Major Areas in AI

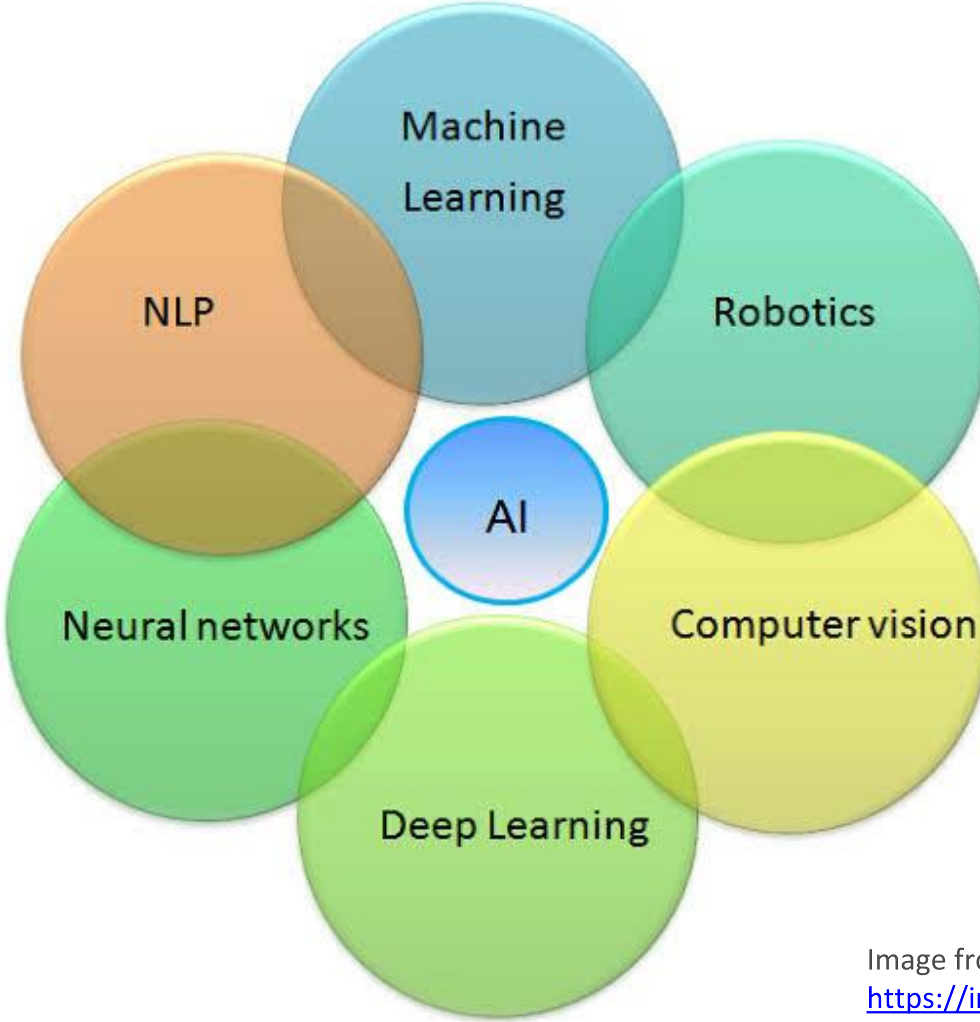


Image from https://infohubns.blogspot.com/2020/07/research-areas-of-ai.html

History of AI in a Nutshell – part 1

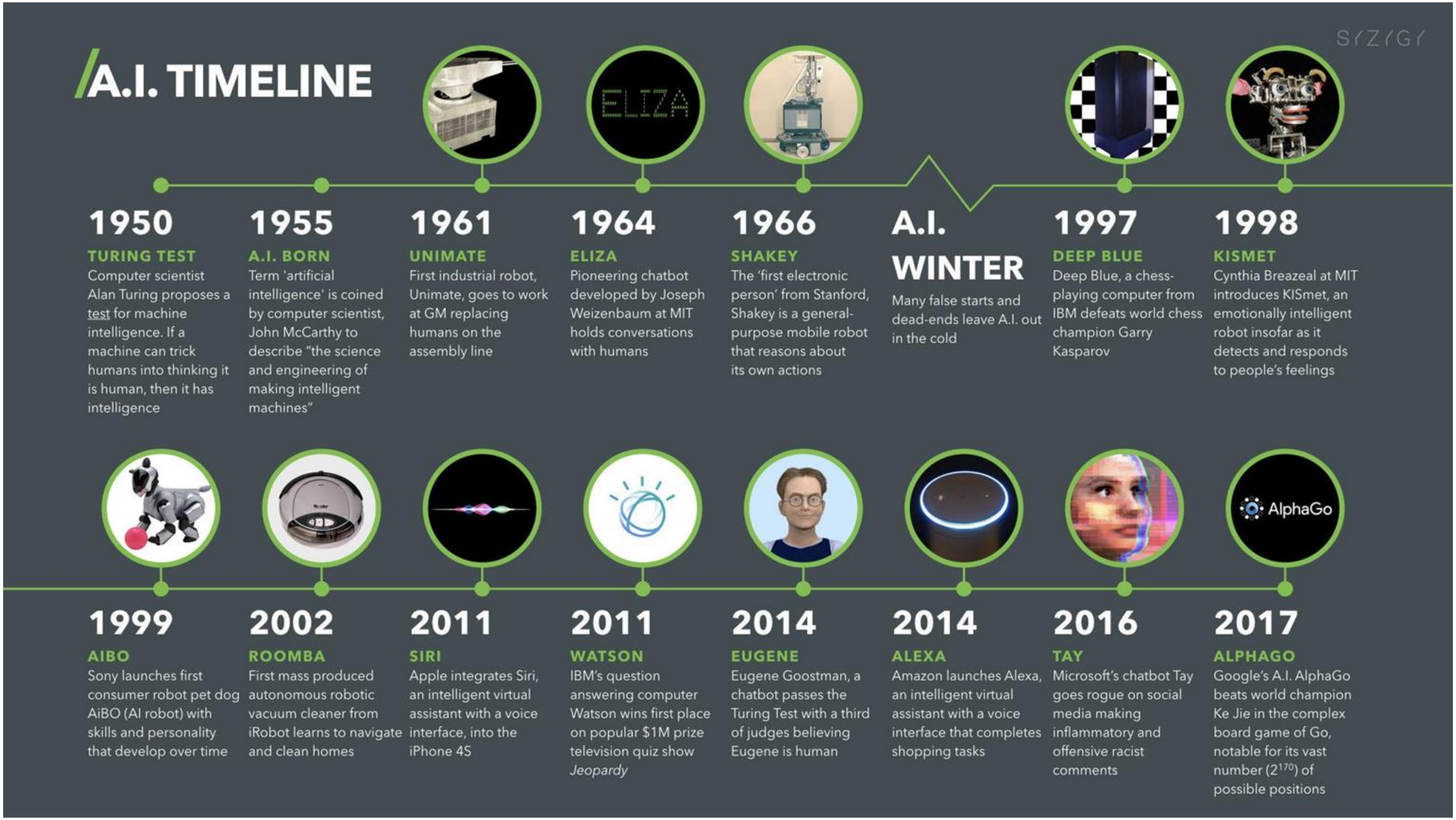


Image from https://digitalwellbeing.org/artificial-intelligence-timeline-infographic-from-eliza-to-tay-and-beyond/

History of AI in a Nutshell – part 2

Transformer networks, 2017, NIPS

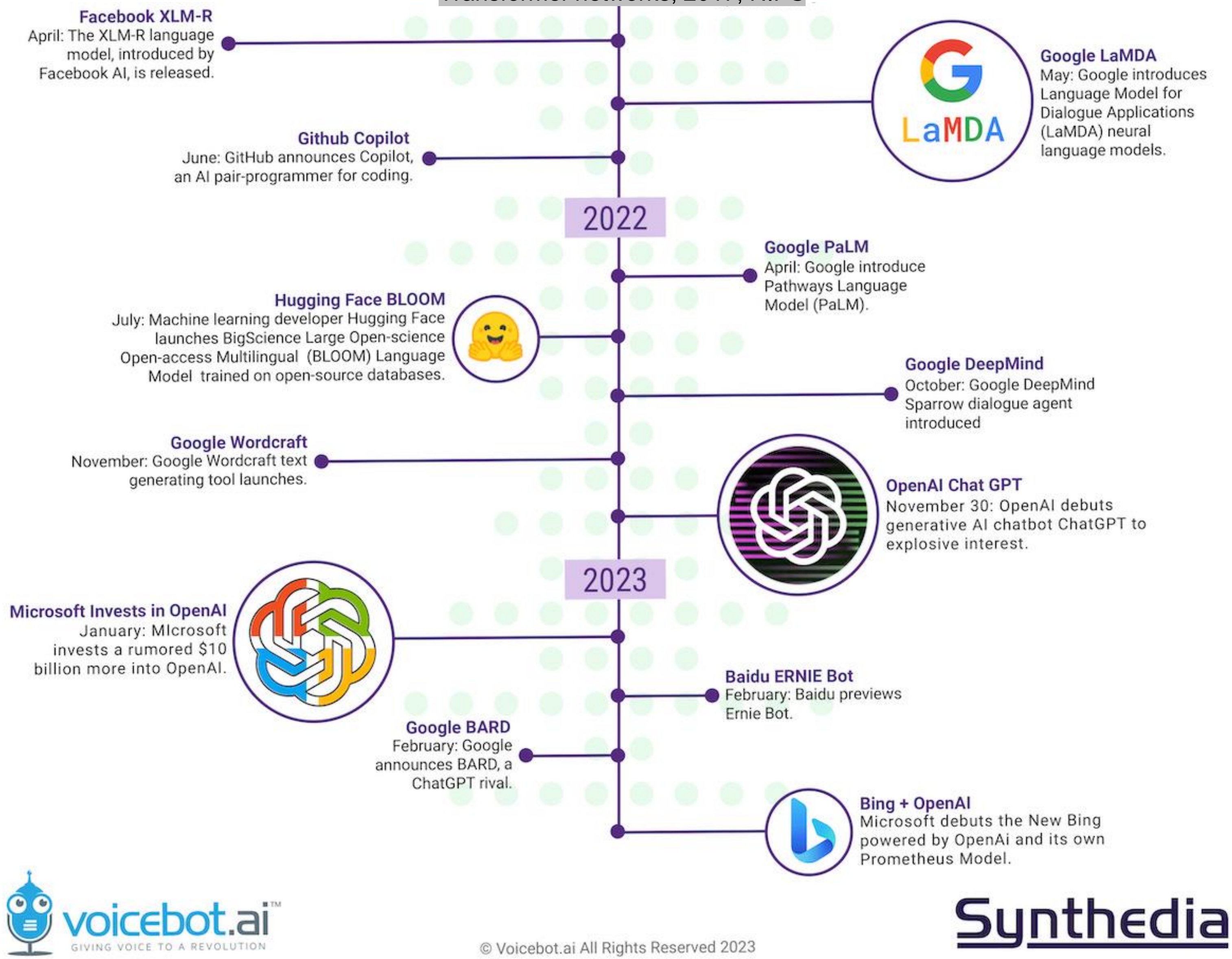
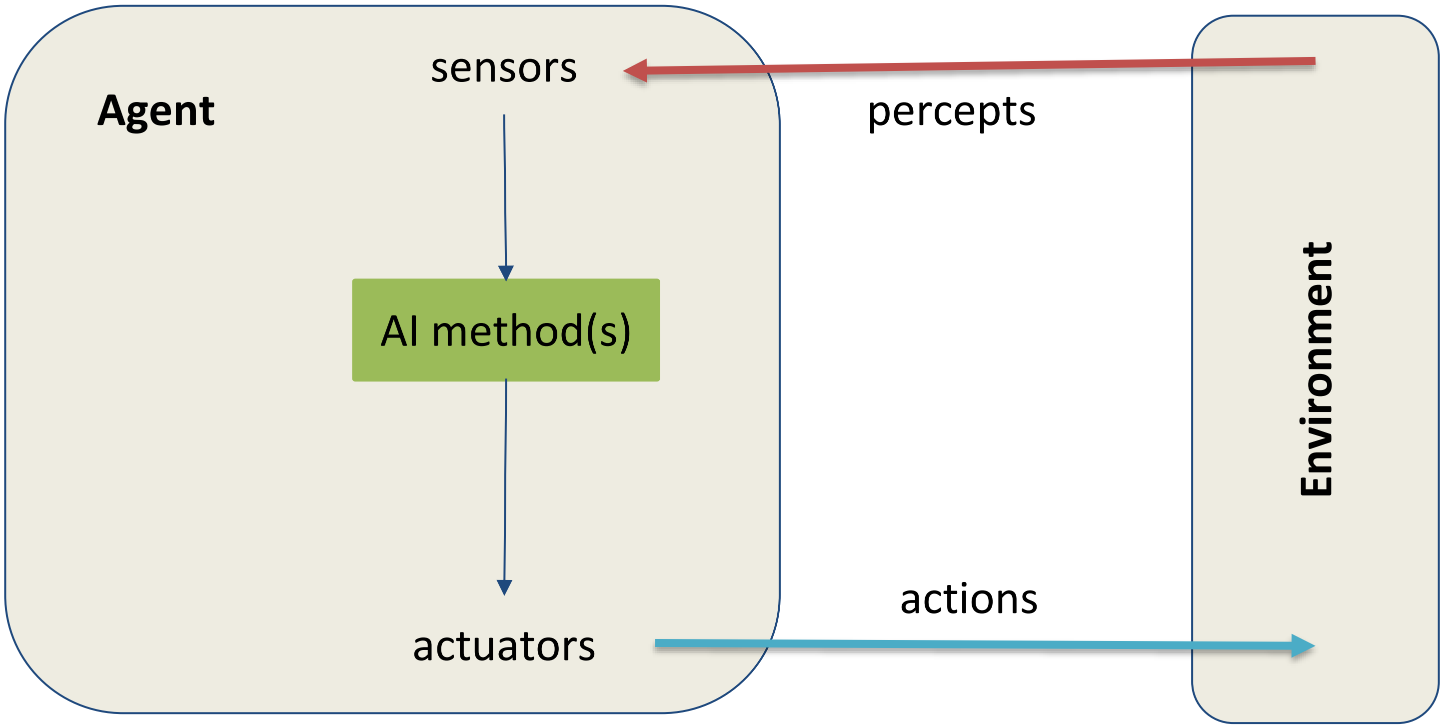


Image from https://synthedia.substack.com/p/the-history-of-large-language-models

AI Agents



AI Agent perceives its environment via its sensors, makes decisions using AI techniques, & executes decisions using its actuators.

Properties of Task Environments

- Fully-observable vs. partially observable
- Single-agent vs. multi-agent
- Deterministic vs. stochastic
- Episodic vs. sequential
- Static vs. dynamic
- Discrete vs. continuous
- Known vs. unknown

perception

behaviour

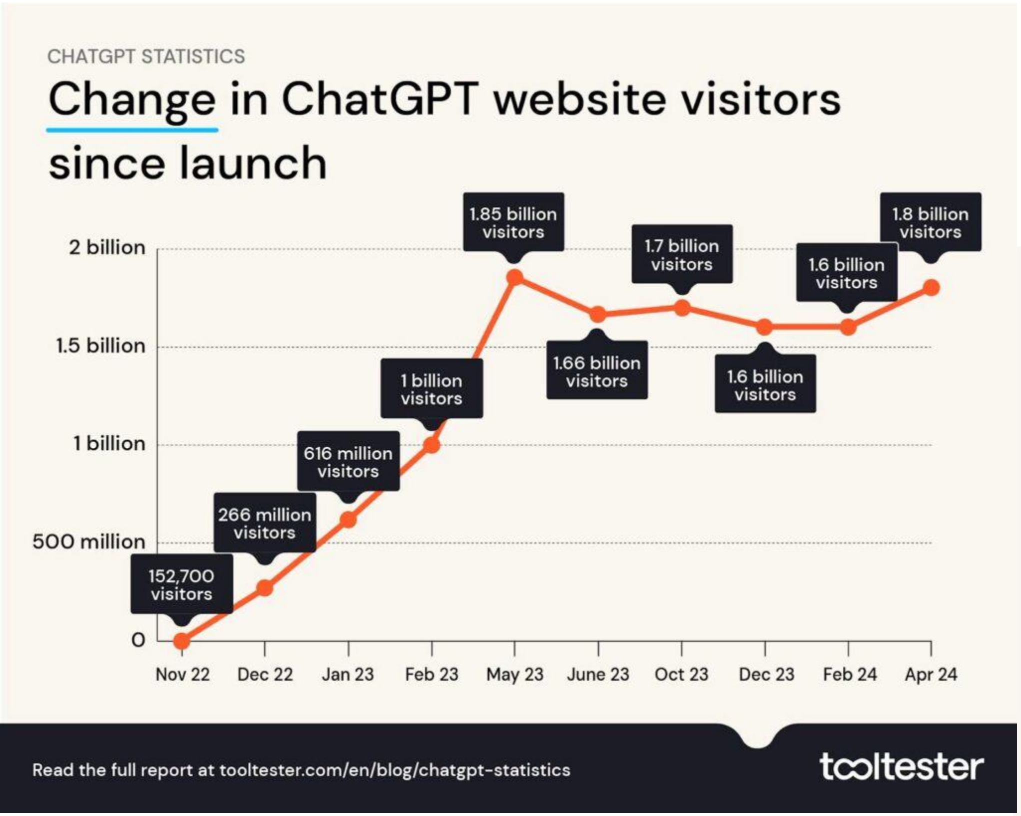
More details of those properties in Russell and Norvig Chapter 2.

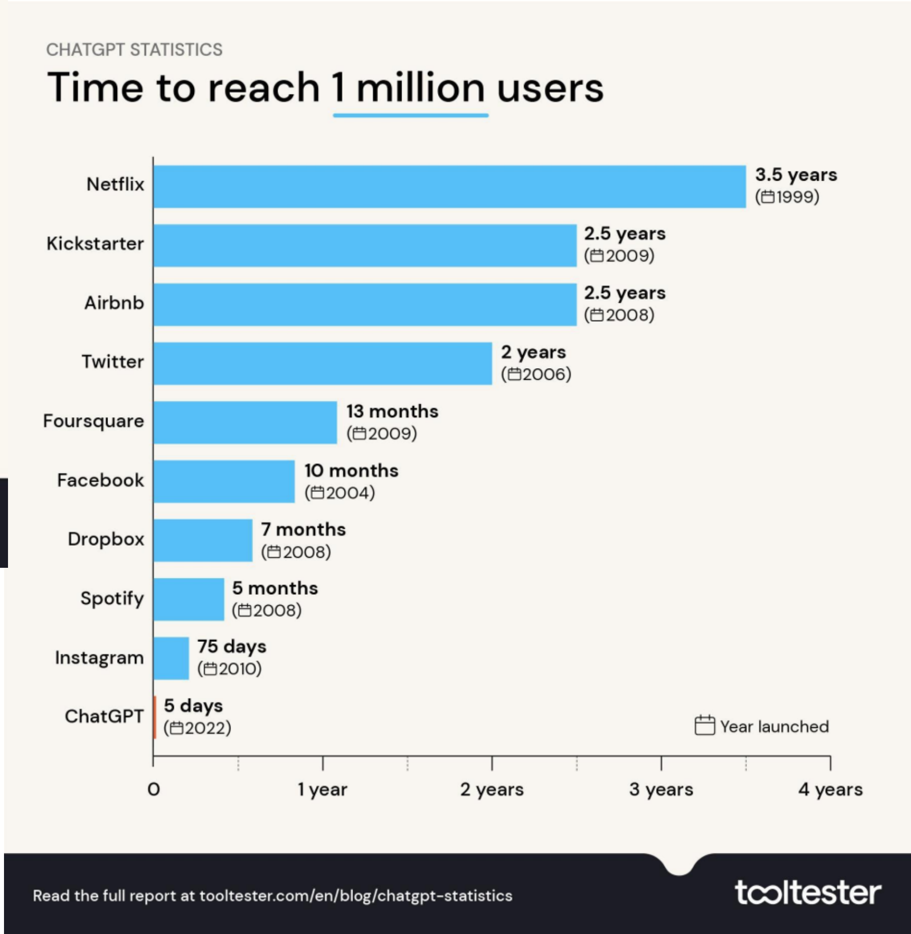
Hardest case (in red)

Example1: Autonomous cars driving on unfamiliar roads.

Example2: Robots cooking food in a changing/dynamic kitchen.

Some Recent Developments in AI





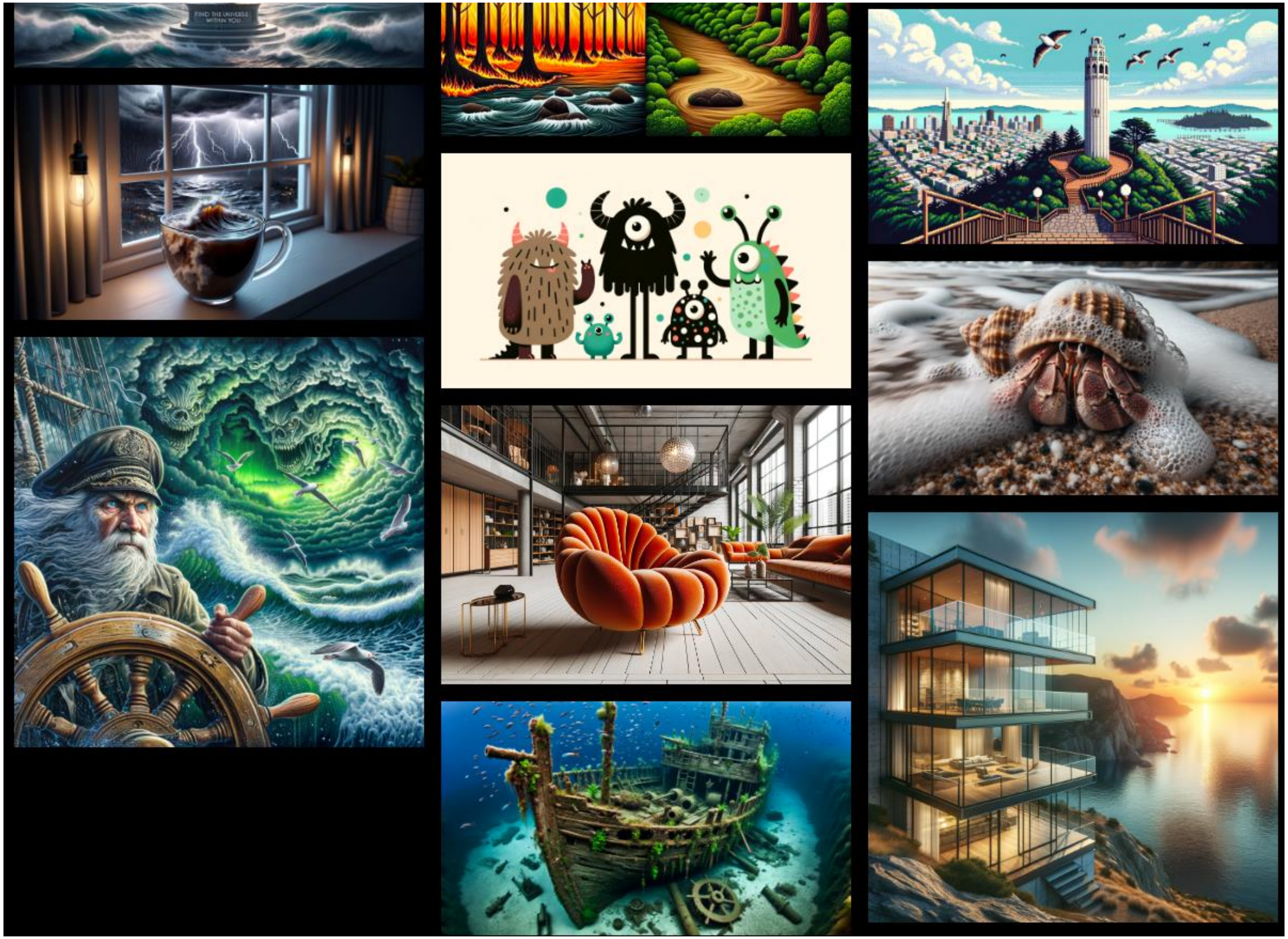
[ChatGPT received around 1.8 billion visitors in Aprill 2024, with ~100 million active users.]

[OpenAI's ChatGPT, a 175B parameter neural net, learns to chat on 570GB of text in 34 days, 2022]

Some Recent Developments in AI

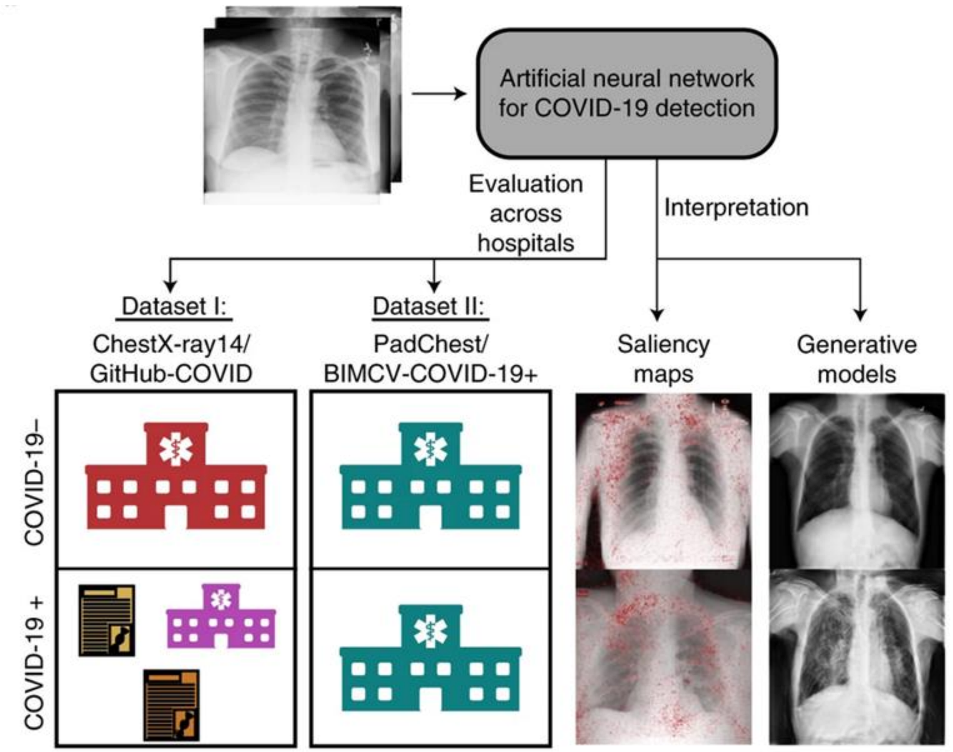
[Aachen+MPI+Tuebingen train neural nets to track multiple objects, 2019]

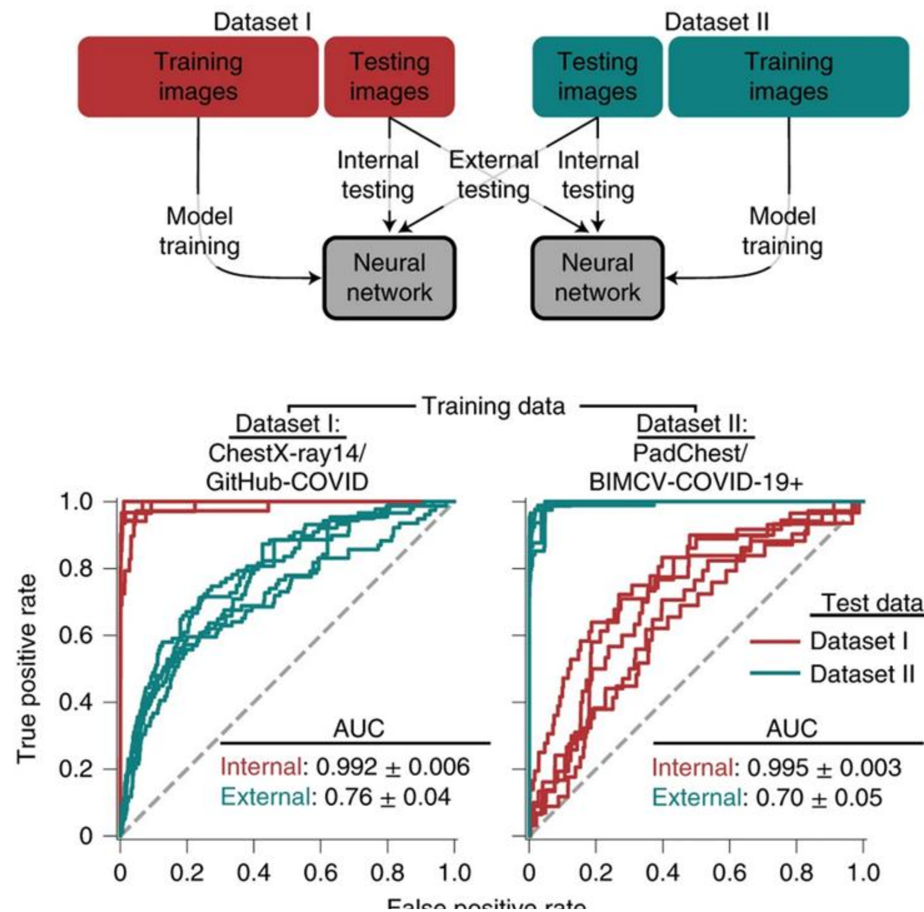
Some Recent Developments in AI



[OpenAI's DALL.E 3 is able to generate images that exactly adheres to text provided, 2023]

Some Recent Developments in AI



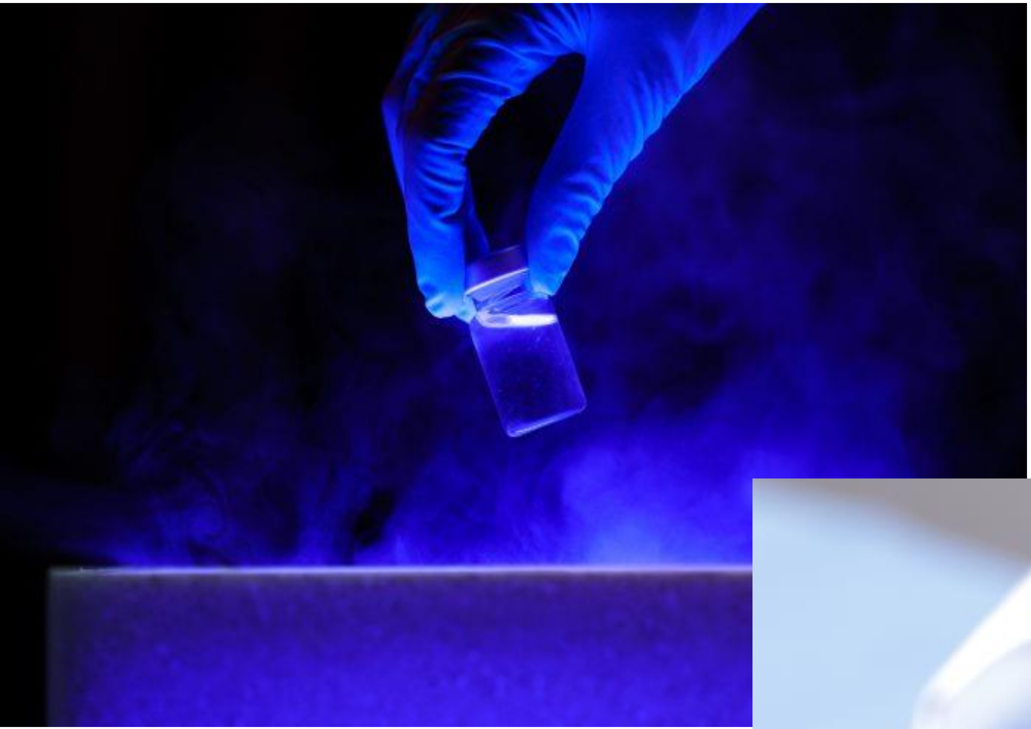


b

	Dataset I			Dataset II		
	Combined	Chest-X-ray14	GitHub-COVID	Combined	PadChest	BIMCV-COVID-19+
No. radiographs	112,528	112,120	408	97,866	96,270	1,596
No. patients	31,067	30,805	262	64,954	63,939	1,105

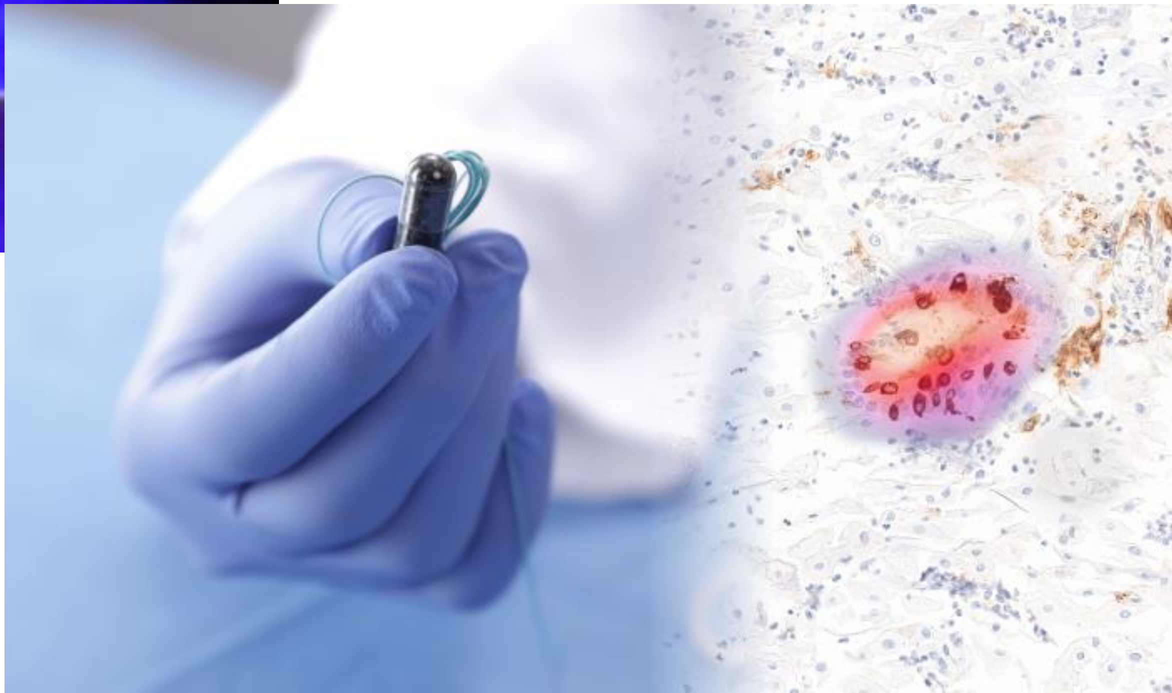
[Medical-imaging AI system learns to detect Covid-19 from chest radiographs, 2021]()

Some Recent Developments in AI

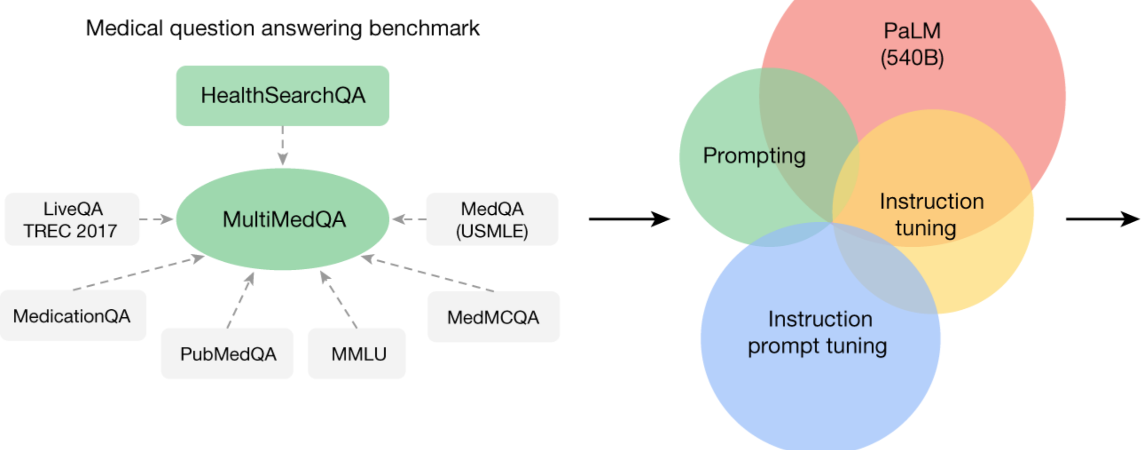


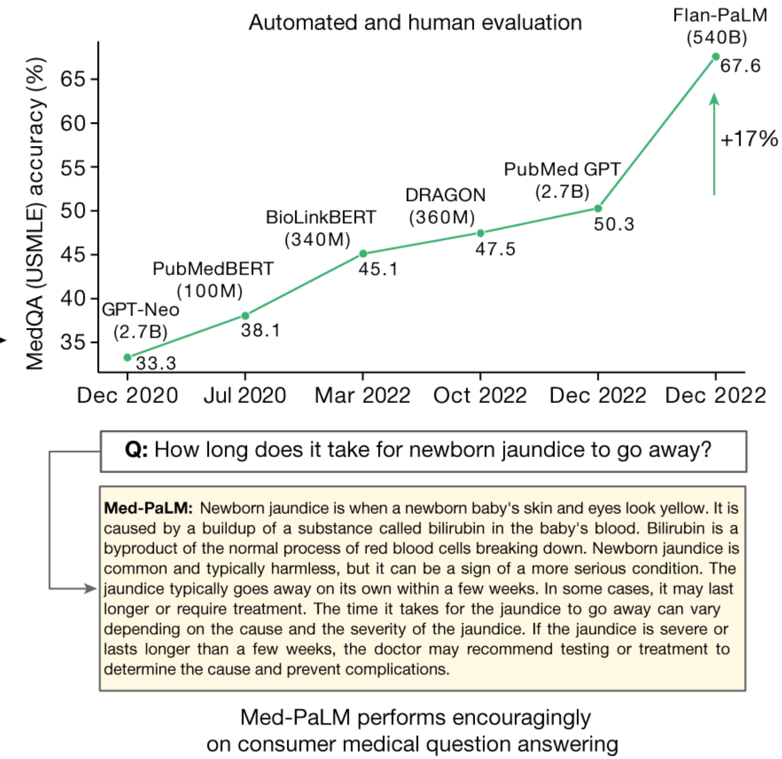
AI system can find COVID vaccine candidates within seconds, 2021

AI for early Cancer detection, 2021



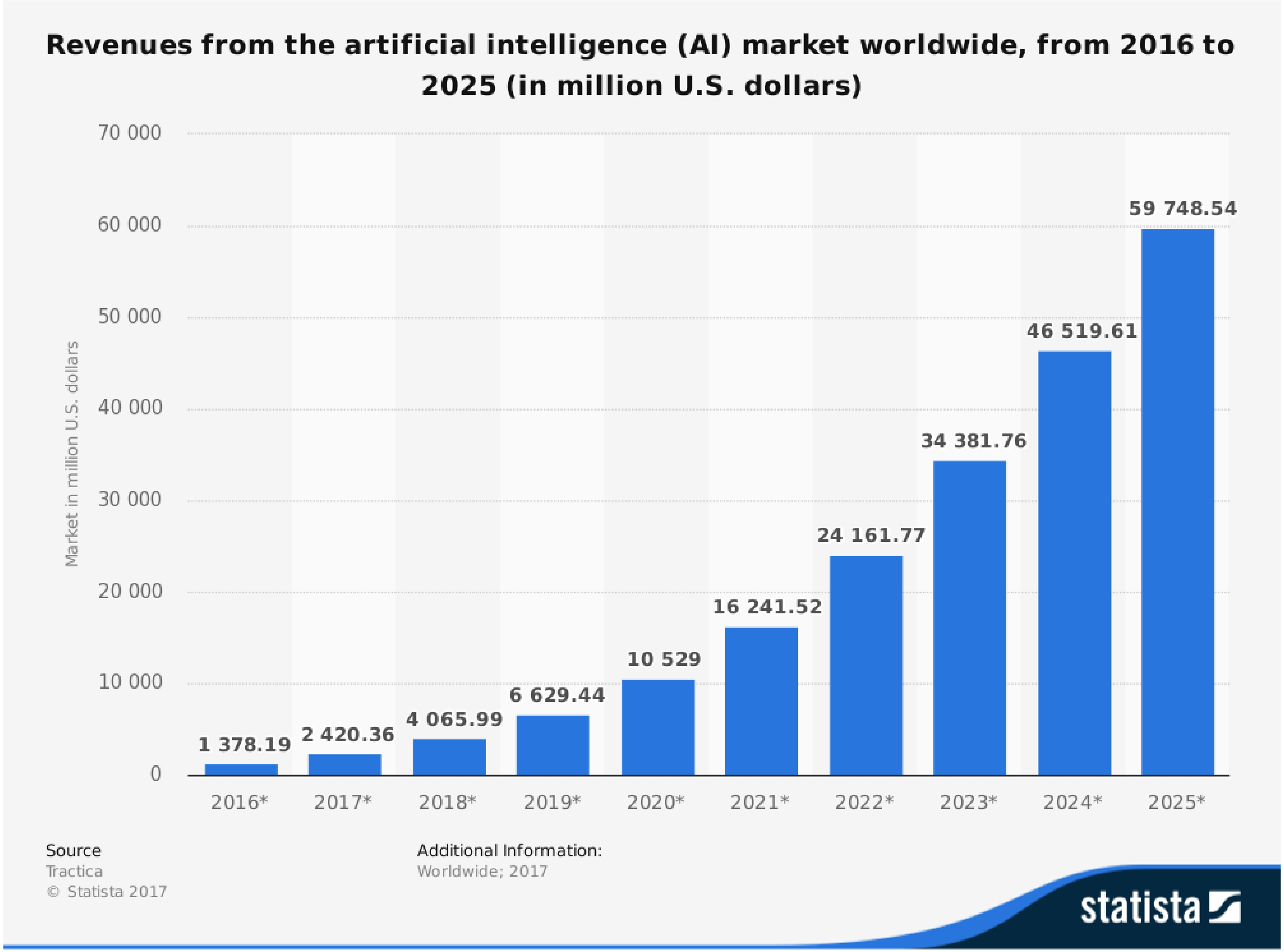
Some Recent Developments in AI





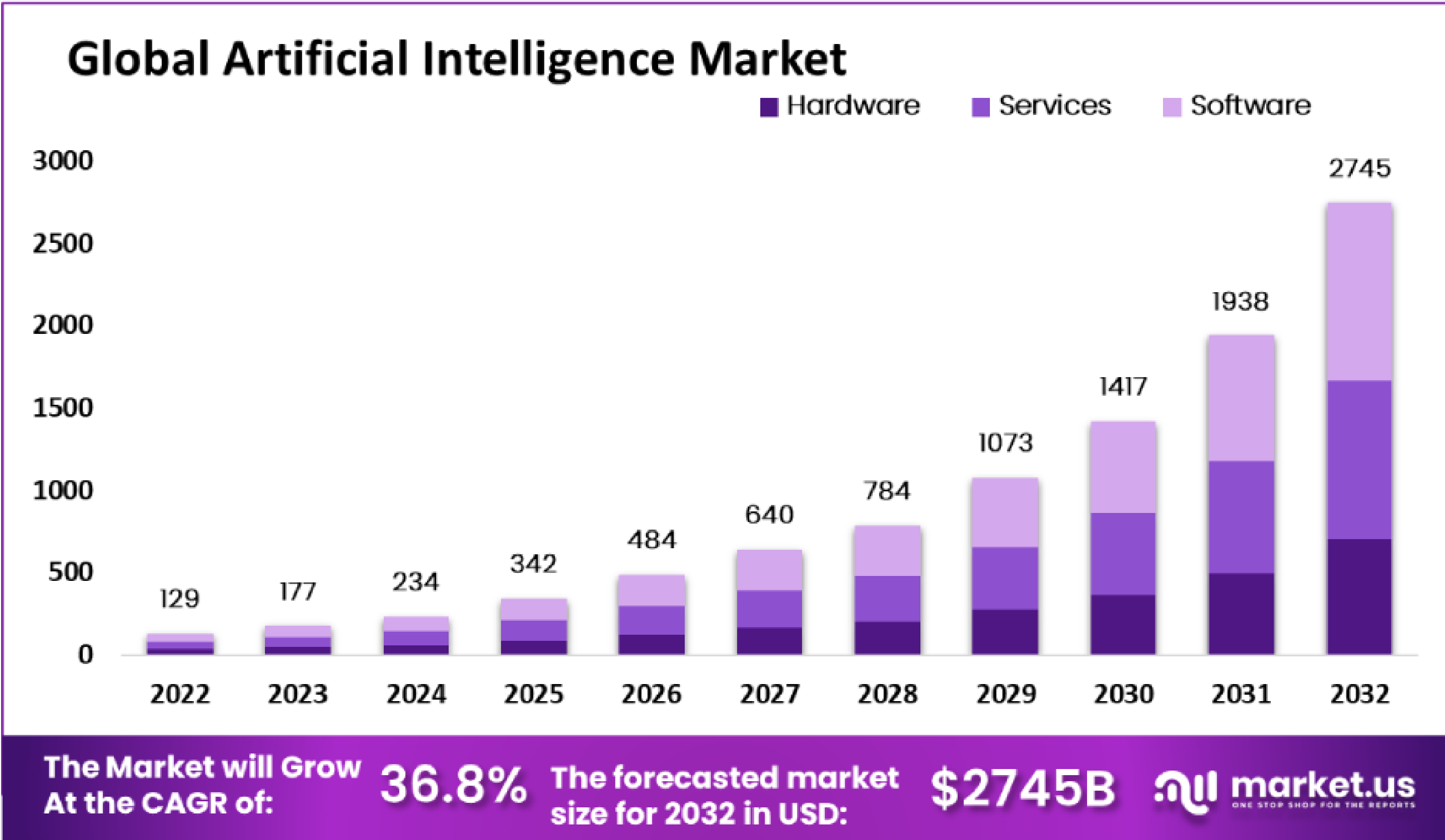
[Large language models encode clinical knowledge, 2023]

AI: Economic Impact



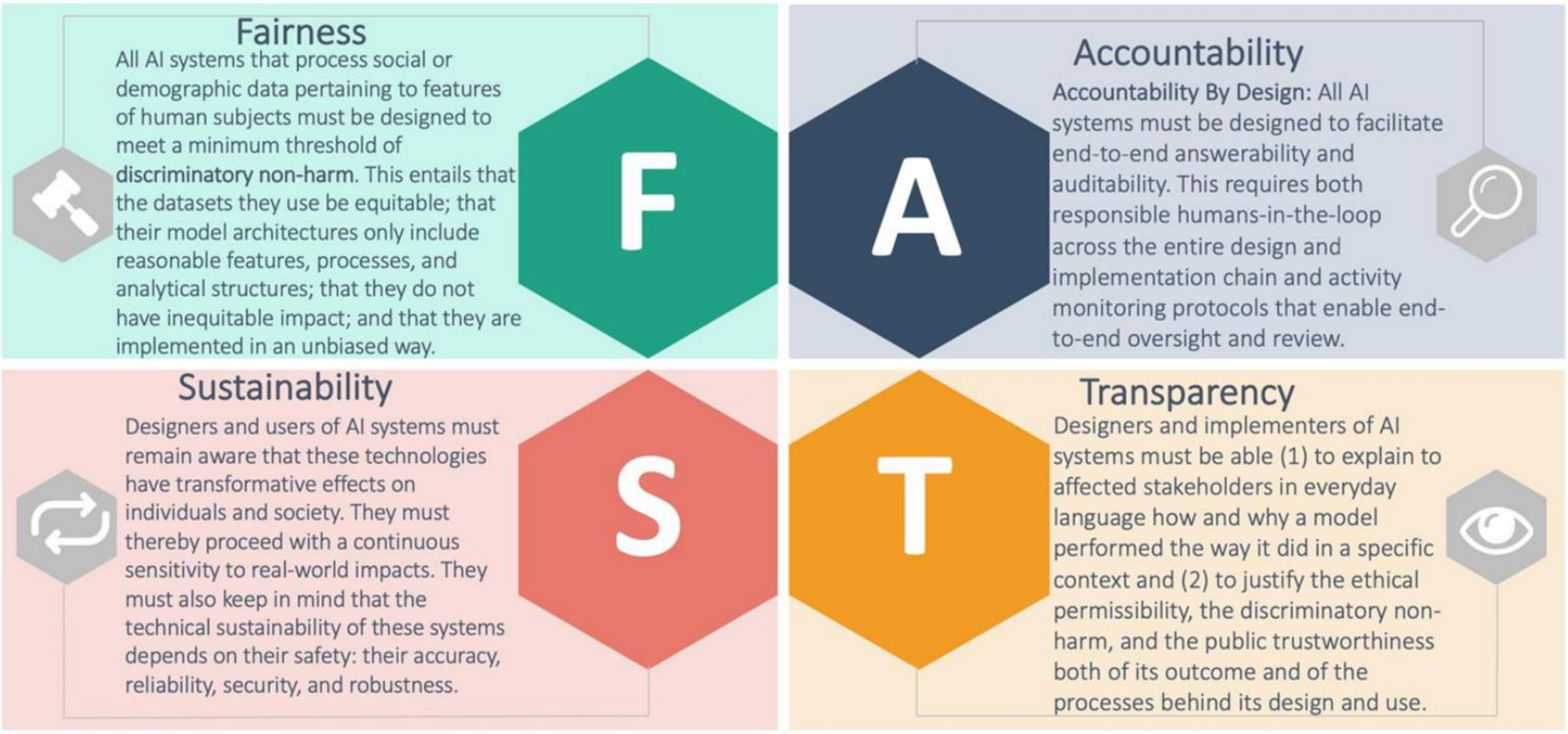
source

AI: Economic Impact



source

Ethics in AI



[Understanding Artificial Intelligence ethics and safety]

[Data ethics and AI guidance landscape]

(Discrete) Random Variables

- The events we are interested in have a set of possible values. Examples of random variables:

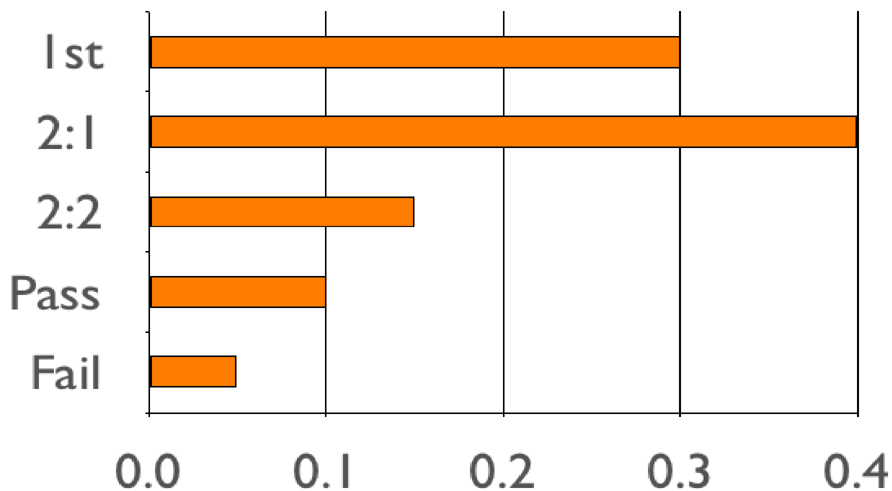
  Coin toss (C): {heads, tails}
  Roll a die (D): {1, 2, 3, 4, 5, 6}
  Weather (W): {snowy, sunny, rainy, foggy}
  Measles (M): {true, false}

- For each event, a random variable takes a value from the associated set. Then we have:

  P(C = tails) or P(tails)
  P(D = 1) or P(1)
  P(W = sunny) or P(sunny)
  P(M = true) or P(measles)

Discrete Probability Distribution

- A probability distribution is a listing of probabilities for *every possible* value of a single random variable



P(1st)=0.30

P(2:1)=0.40

P(2:2)=0.15

P(Pass)=0.10

P(Fail)=0.05

$$\sum_{e \in S} P(e) = 1$$

- Probability distributions can be estimated from data

Joint Probabilities

Joint probabilities represent the whole joint probability distribution

S1=1st	S2=1st	Probability
TRUE	TRUE	0.2
TRUE	FALSE	0.1
FALSE	TRUE	0.1
FALSE	FALSE	0.6

$$\sum_{e \in S} P(e) = 1$$

Joint Probability Distribution

Sometimes a joint probability distribution looks like the one below, which has the same information as the table on the previous slide.

	S2 =1st	¬(**S2** =1st)
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

Marginal Probabilities

	S2 =1st	**¬(S2** =1st**)**
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

Marginalisation => summing up the probabilities of the other variables—i.e. $P(X) = \sum_y P(X, Y = y)$

Example: What is the probability of S1 getting a 1st?

P(S1=1st) = P(S1=1st ∧ S2=1st) + P(S1=1st ∧ ¬(S2=1st))
= 0.2 + 0.1 = 0.3

Marginal Probabilities

	S2 =1st	**¬(S2** =1st**)**
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

P(S1=1st) = P(S1=1st ∧ S2=1st) + P(S1=1st ∧ ¬(S2=1st))
= 0.2 + 0.1 = 0.3

P(¬(S1=1st)) = P(¬(S1=1st) ∧ S2=1st) + P(¬(S1=1st) ∧ ¬(S2=1st))
= 0.1 + 0.6 = 0.7

Note that P(S1=1st)+ P(¬(S1=1st))=0.3+0.7=1

Marginal Probabilities

	S2 =1st	¬(**S2** =1st)
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

P(S2=1st) = P(S1=1st ∧ S2=1st) + P(¬(S1=1st) ∧ S2=1st)
= 0.2 + 0.1 = 0.3

P(¬(S2=1st)) = P(S1=1st ∧ ¬(S2=1st)) + P(¬(S1=1st) ∧ ¬(S2=1st))
= 0.1 + 0.6 = 0.7

Note that P(S2=1st)+ P(¬(S2=1st))=0.3+0.7=1

Conditional Probability

- Conditional probabilities answer the question: *Given that some event B happened, what is the probability of A happening too?*

- **P(A|B)**=?

- A conditional probability is defined as:
  $\underbrace{\mathbf{P(A \mid B) = P(A \wedge B) / P(B)}}_{\text{Product Rule}}$, where **P(B)** $\neq 0$

$$\boxed{\mathbf{P(A \wedge B) = P(A \mid B) * P(B)}}$$

Chain Rule for Probabilities

- **P(A ∧ B)= P(A | B) * P(B)**

- **P(A ∧ B ∧ C)= P(A | B ∧ C) * P(B | C) * P(C)**
  = P(C) * P(B | C) * P(A | B ∧ C)
  = P(A) * P(B | A) * P(C | B ∧ A)
  = P(C | B ∧ A) * P(B | A) * P(A)

- **P**$(A_1 \wedge A_2 \wedge \cdots \wedge A_n)$ **=** **P**$(A_n | A_{n-1} \wedge \cdots \wedge A_1) *$ **P**$(A_{n-1} | A_{n-2} \wedge \cdots \wedge A_1) * \cdots *$ **P**$(A_2 | A_1) *$ **P**(A_1)

Conditional Probability

	S2=1st	¬(S2=1st)
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

- P(S2=1st | S1=1st)?
- = P(S2=1st, S1=1st) / P(S1=1st)
- = 0.2 / 0.3 = 0.6666666

P(A | B) = P(A ∧ B) / P(B)

Substituting A with S2=1st

Substituting B with S1=1st

If the first student has a 1st, the second has a 66% chance of having a 1st too!

Conditional Probability

	S2=1st	¬(S2=1st)
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

- P(¬(S2=1st) | S1=1st)?
- = P(¬(S2=1st), S1=1st) / P(S1=1st)
- = 0.1 / 0.3 = 0.3333333

$P(A \mid B) = P(A \wedge B) / P(B)$

Substituting A with ¬(S2=1st)

Substituting B with S1=1st

If the first student (S1) has a 1st, the second has a 33% chance of not getting a 1st.

Conditional Probability

	S2=1st	**¬(S2=1st)**
S1=1st	$\boxed{0.2}$	$\boxed{0.1}$
¬(S1=1st)	0.1	0.6

Since P(S2=1st | S1=1st) = P(S2=1st, S1=1st) / P(S1=1st)

= 0.2 / 0.3 = 0.6666666

and P(¬(S2=1st) | S1=1st) = P(¬(S2=1st), S1=1st) / P(S1=1st)

= 0.1 / 0.3 = 0.3333333

Then P(S2=1st | S1=1st) + P(¬(S2=1st) | S1=1st) = 1

More on Joint Probabilities

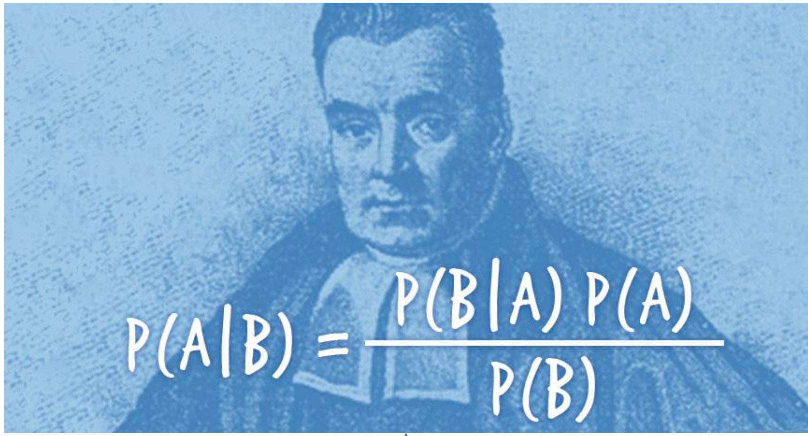
Given a joint probability table, we have all the information we need about the domain. We can calculate the probability of any logical formula.

P(S1=first ∨ S2=first) = 0.2 + 0.1 + 0.1 = 0.4

	S2 =1st	¬(**S2** =1st)
S1=1st	0.2	0.1
¬(S1=1st)	0.1	0.6

Bayes Rule

- **P(A ∧ B) = P(B ∧ A)**
- **P(A | B)*P(B) = P(B | A)*P(A)**
- **P(A | B) = (P(B | A)*P(A)) / P(B)**
- **P(B | A) = (P(A | B)*P(B)) / P(A)**



Thomas Bayes
(1701-1761)

Bayes Rule for Classification*

- Given inputs *X* and outputs *Y*, the Bayes rule can be written as

$$P(Y = y_k|X = x_i) = \frac{P(X = x_i|Y = y_k)P(Y = y_k)}{\sum_j P(X = x_i|Y = y_j)P(Y = y_j)}$$

where

y_k is the possible value for Y

x_i is the possible vector value for X

- Use training data to estimate $P(X|Y)$ and $P(Y)$.

* Accurately estimating $\boldsymbol{P(X \mid Y)}$ is expensive--
see Appendix 1.

Naïve Bayes Classifier

- Conditional independence: $P(X_1 \ldots X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$
- Thus

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)P(X_1 \ldots X_n|Y = y_k)}{\sum_j P(Y = y_j)P(X_1 \ldots X_n|Y = y_j)}$$

can be rewritten as

$$P(Y = y_k|X_1 \ldots X_n) = \frac{P(Y = y_k)\prod_i P(X_i|Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i|Y = y_j)}$$

Naïve Bayes Classifier

- Given a new instance/example $X^{new} =< X_1 \dots X_n >$, the most probable value of Y can be obtained with

$$Y = \arg\max_{y_k} \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- Since the denominator does not depend on y_k, the equation can be simplified as

$$Y = \underset{\mathrm{y_k}}{\operatorname{argmax}} P(Y = y_k) \prod_i P(X_i | Y = y_k)$$

Naïve Bayes for Discrete Inputs

- Estimate two sets of parameters:

$$\theta_{ijk} = P\left(X = x_{ij} \middle| Y = y_k\right)$$
$$\pi_k = P(Y = y_k)$$

- According to

$$\theta_{ijk} = P\left(X = x_{ij} \middle| Y = y_k\right) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = yk\} + lJ}$$
$$\pi_k = P(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

where J is the number of unique values in X_i, l (e.g., $l = 1$) avoids zero estimates, and $|D|$ is the num. of elements in the training set.

Example: Training Data

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Data from: Mitchell, T. "Machine Learning", McGraw Hill, 1997.

Example: Estimated Parameters

$$P(PlayTennis = yes) = \frac{9}{14} = 0.643$$
$$P(PlayTennis = no) = \frac{5}{14} = 0.357$$

$$P(Outlook = sunny|PlayTennis = yes) = 2/9 = 0.222$$
$$P(Outlook = sunny|PlayTennis = no) = 3/5 = 0.60$$
$$P(Outlook = overcast|PlayTennis = yes) = 4/9 = 0.444$$
$$P(Outlook = overcast|PlayTennis = no) = 0/5 = 0.0$$
$$P(Outlook = rain|PlayTennis = yes) = 3/9 = 0.333$$
$$P(Outlook = rain|PlayTennis = no) = 2/5 = 0.4$$

...

We want to avoid zero probabilities

Example: Estimated Parameters

$$P(Temperature = hot|PlayTennis = yes) = 2/9 = 0.222$$
$$P(Temperature = hot|PlayTennis = no) = 2/5 = 0.4$$
$$P(Temperature = mild|PlayTennis = yes) = 4/9 = 0.444$$
$$P(Temperature = mild|PlayTennis = no) = 2/5 = 0.4$$
$$P(Temperature = cool|PlayTennis = yes) = 3/9 = 0.333$$
$$P(Temperature = cool|PlayTennis = no) = 1/5 = 0.2$$

$$P(Humidity = high|PlayTennis = yes) = 3/9 = 0.333$$
$$P(Humidity = high|PlayTennis = no) = 4/5 = 0.8$$
$$P(Humidity = normal|PlayTennis = yes) = 6/9 = 0.666$$
$$P(Humidity = normal|PlayTennis = no) = 1/5 = 0.2$$

$$P(Wind = strong|PlayTennis = yes) = 3/9 = 0.333$$
$$P(Wind = strong|PlayTennis = no) = 3/5 = 0.6$$
$$P(Wind = weak|PlayTennis = yes) = 6/9 = 0.666$$
$$P(Wind = weak|PlayTennis = no) = 2/5 = 0.4$$

Example: Classifying a New Instance

$$P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes) =$$
$$= 0.643 * 0.222 * 0.333 * 0.333 * 0.333 = 0.0053$$

$$P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no) =$$
$$= 0.357 * 0.60 * 0.2 * 0.8 * 0.6 = 0.0206$$

$$P(PlayTennis = yes|evidence) = \frac{0.0053}{0.0053 + 0.0206} = 0.205$$

$$P(PlayTennis = no|evidence) = \frac{0.0206}{0.0053 + 0.0206} = 0.795$$

Outlook=sunny
Temperature=cool
Humidity=high
Wind=strong

The probabilities above can be calculated in log space, see Apendix 2.

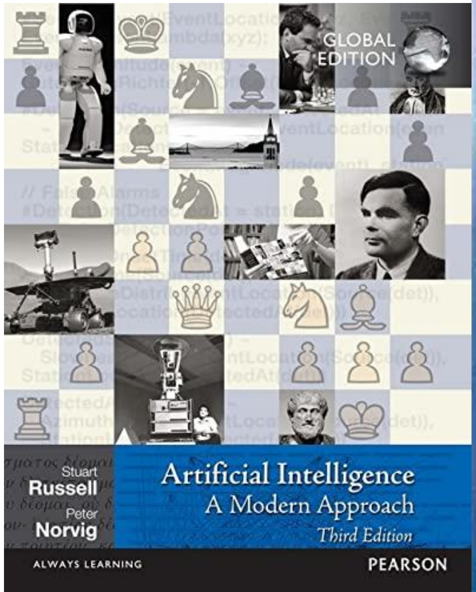
Today

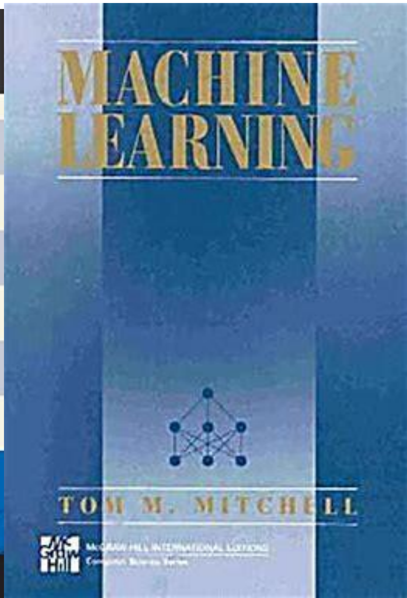
- Elaborated on the term `**Artificial Intelligence**'
- Some recent major developments in AI
- Introduction to probability theory
- Probabilistic reasoning with Naïve Bayes

Readings:

Russell & Norvig 2016. [Chapters 1, 2, 13]

Mitchell, T. 2017; 2nd Ed. [Chapter 3]





<u>This and Next Week</u>

Workshop (tomorrow):
Exercises on probability theory
Python programs for Bayesian classifiers

Lecture (next week):
Bayesian Networks with Exact Inference
Readings: Russell & Norvig 2016. Chapters 13-14.4

Questions now and during the workshop.

Appendix 1: Difficulty in Unbiased Bayesian Classifiers

- Accurately estimating $P(X|Y)$ requires a set of parameters such as $\theta_{ij} = P(X = x_i|Y = y_j)$, where

  index j refers to 2 possible values

  index i refers to 2^n possible values

- This requires 2^{n+1} parameters or probabilities.

- A vector X with 30 Boolean inputs requires 2.15B parameters

Appendix 2: Example with Log Probabilities

$$P(yes)P(sunny|yes)P(cool|yes)P(high|yes)P(strong|yes)$$
$$= (-0.442) + (-1.504) + (-1.098) + (-1.098) + (-1.098)$$
$$= -5{,}242$$

$$P(no)P(sunny|no)P(cool|no)P(high|no)P(strong|no)$$
$$= (-1.029) + (-0.51) + (-1.609) + (-0.223) + (-0.51)$$
$$= -3.883$$

$$P(PlayTennis = yes|evidence) = \frac{e^{-5{,}242}}{e^{-5{,}242} + e^{-3.883}} = 0.205$$

$$P(PlayTennis = no|evidence) = \frac{e^{-3.883}}{e^{-5{,}242} + e^{-3.883}} = 0.795$$