Introduction to Artificial Intelligence

Basics and applications

Dariusz Brzezinski

Agenda

- A bit of terminology

- Basics of machine learning

- Applications
  - Images
  - Sound
  - Text
  - Other data

- Discussion

Terminological confusion

- Artificial intelligence (AI, AGI)
- **Machine learning** (ML)
- Deep learning (DL)
- Data mining
- Knowledge discovery
- Data science
- Big data
- Statistical data analysis
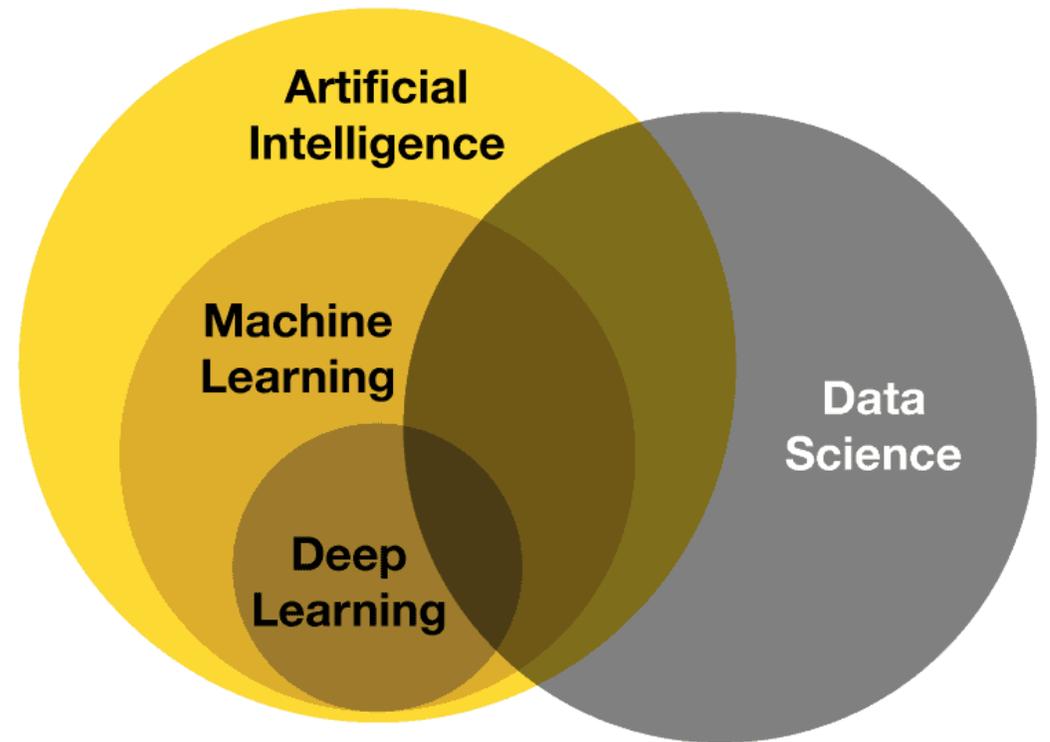
Machine learning

Machine learning *is a subfield of artificial intelligence dedicated to algorithms that improve themselves and make predictions based on data. The term is often used interchangeably with artificial intelligence.*

Various applications:

- Internet search engines
- Recommendation systems
- Face recognition
- Voice recognition
- Spam filtering
- Credit risk assessment

Machine learning

- **Unsupervised** (without a teacher)
  - Clustering
  - Association mining

- **Supervised** (with a teacher)
  - Classification
  - Regression
  - Reinforcement learning
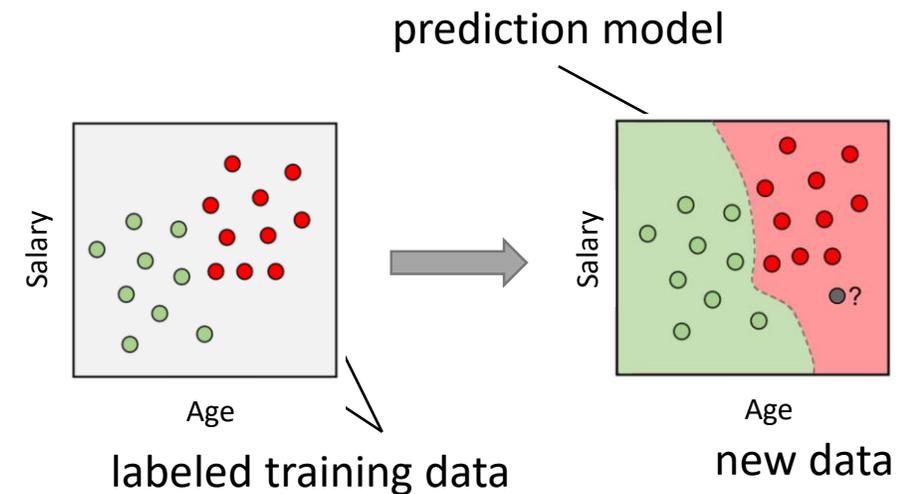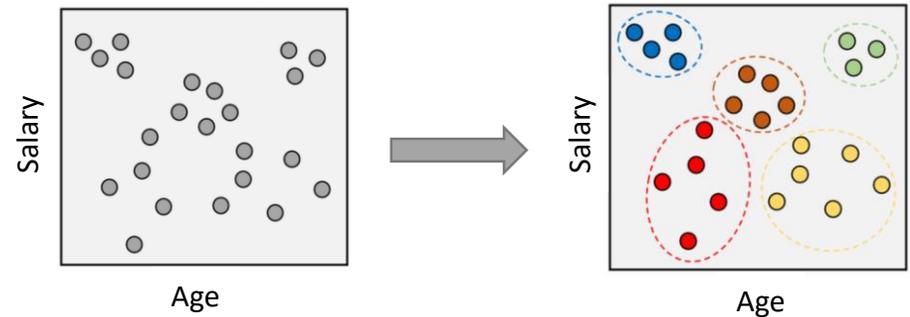
# Machine learning

- **Unsupervised** (without a teacher)
  - Clustering
  - Association mining

- **Supervised** (with a teacher)
  - Classification
  - Regression
  - Reinforcement learning
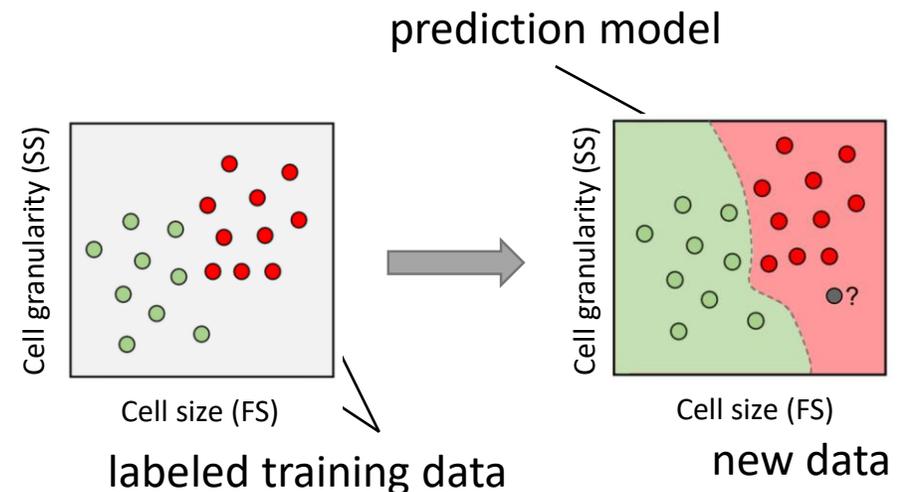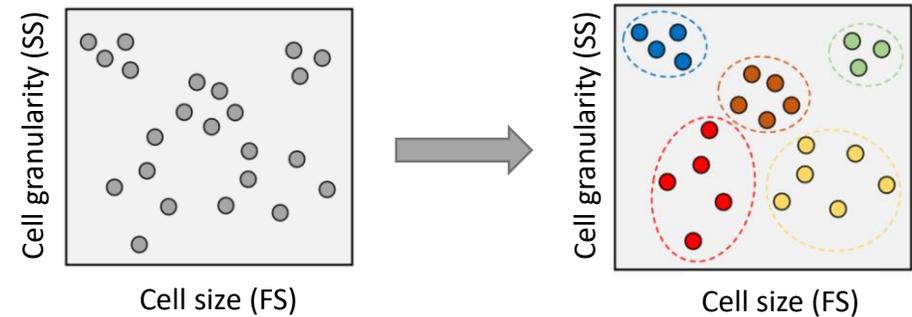
When to use machine learning?

1. **There is a pattern in the studied problem**

2. **We cannot model the pattern mathematically***

3. **We have data on the problem**

Yaser Abu-Mostafa
Learning from Data

Your task for this lecture:
Imagine AI applied to your problems

Images
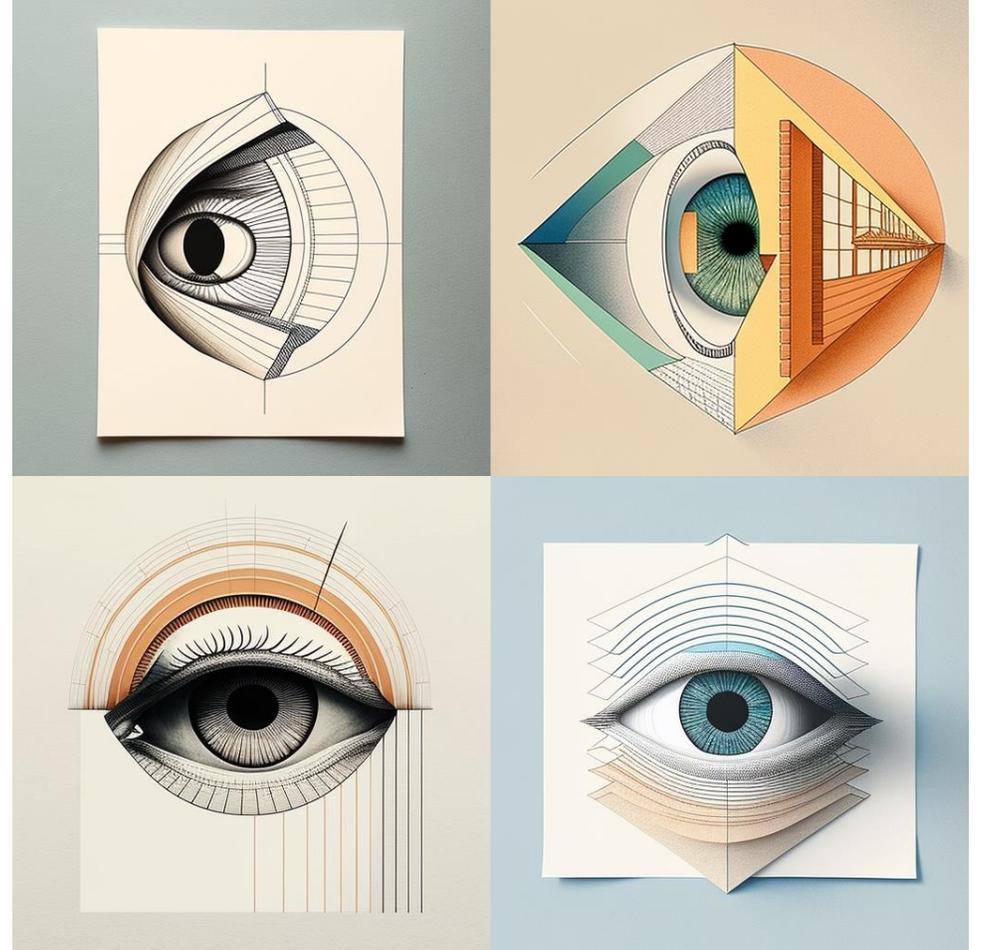
Image representation and convolutions

- Images in a computer are matrices of numbers
- Transformations can be performed on these matrices
- A convolution is a transformation on a fragment of an image
- Convolutional networks are neural networks for images that search for the best image transformations

input image

kernel:
emboss

output image

http://setosa.io/ev/image-kernels/
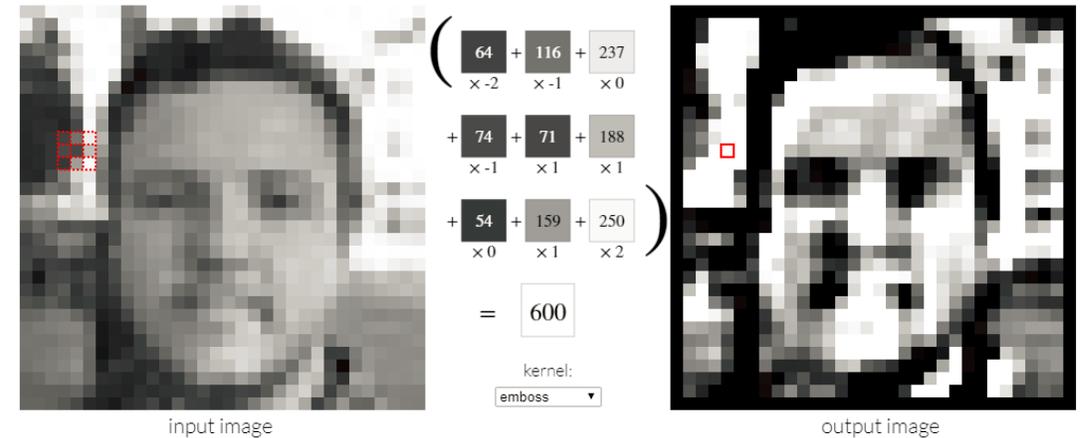
Convolutional neural networks
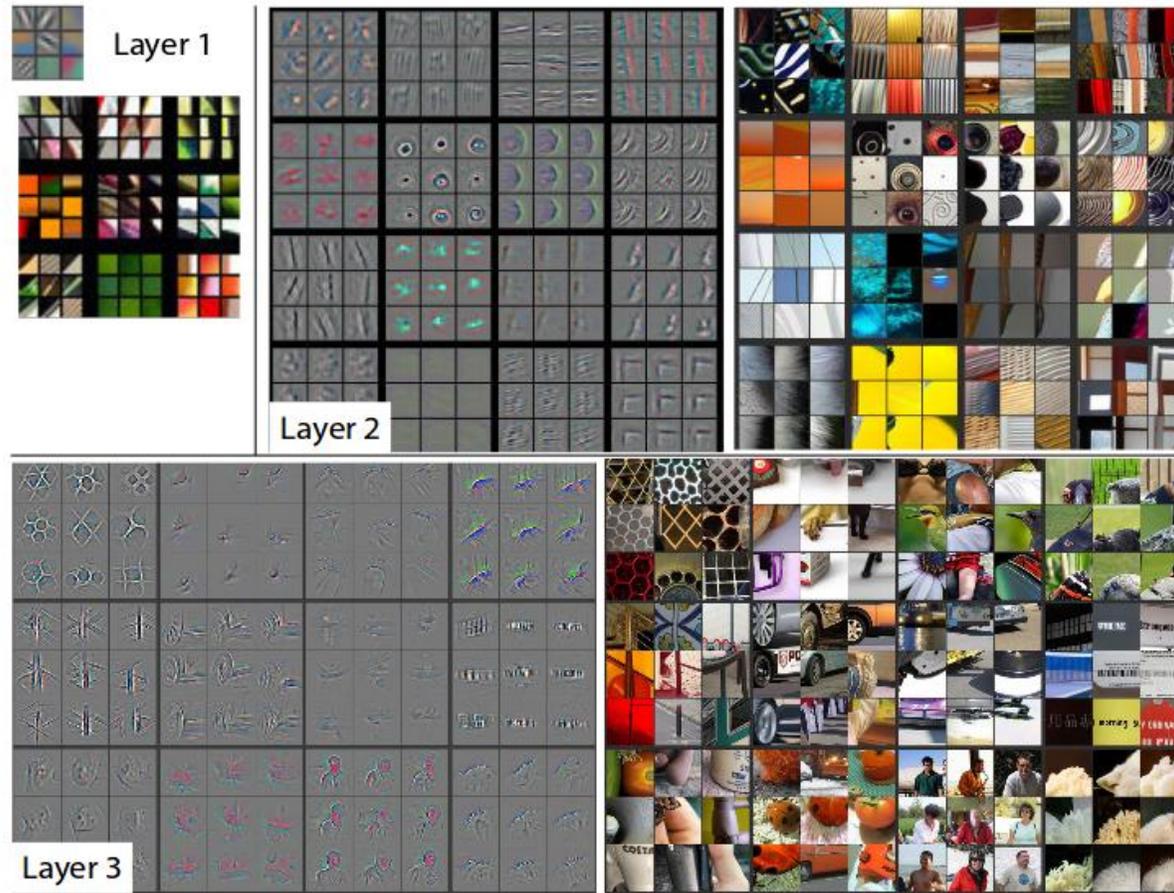
Transfer learning

1. Identify a suitable machine learning task for your problem

2. Select a pretrained model

3. If necessary, fine-tune the model to your data
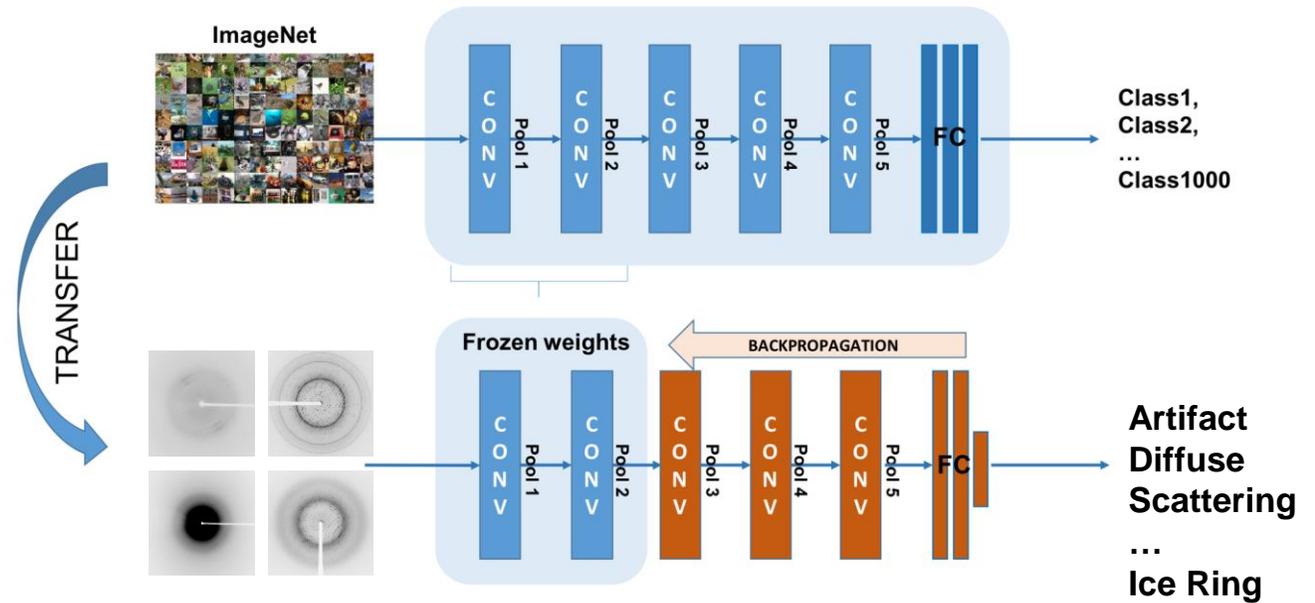
Object recognition

- Wide range of applications:
  - Biometrics
  - Road sign recognition
  - Autonomous driving
  - Counting people, cars, etc.
  - Intelligent doorbells
  - Object tracking
  - Animal photo traps
  - Satellite image analysis
  - Automatic photo captioning
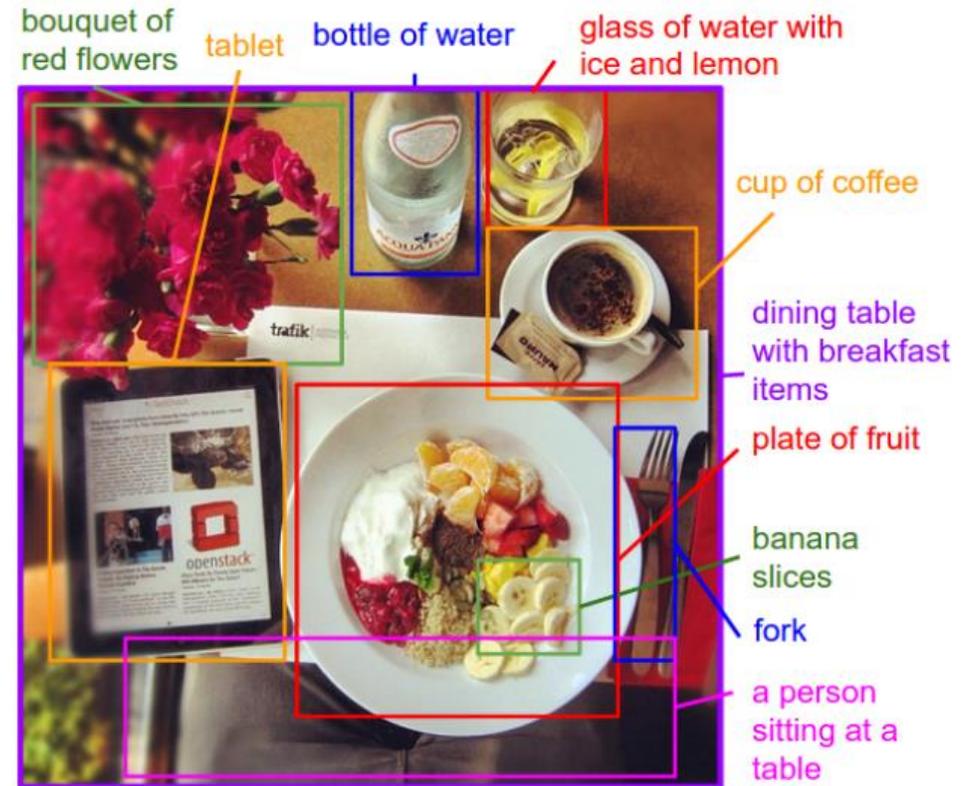
Digitization of analog measurements

- Facilitating measurements from analog gauges
- Mountable in hard-to-reach places
- Similar idea: drone measurements (e.g., monitoring the condition of power lines)
- You can also monitor the correctness of physical processes (e.g., assembly of parts in a factory)
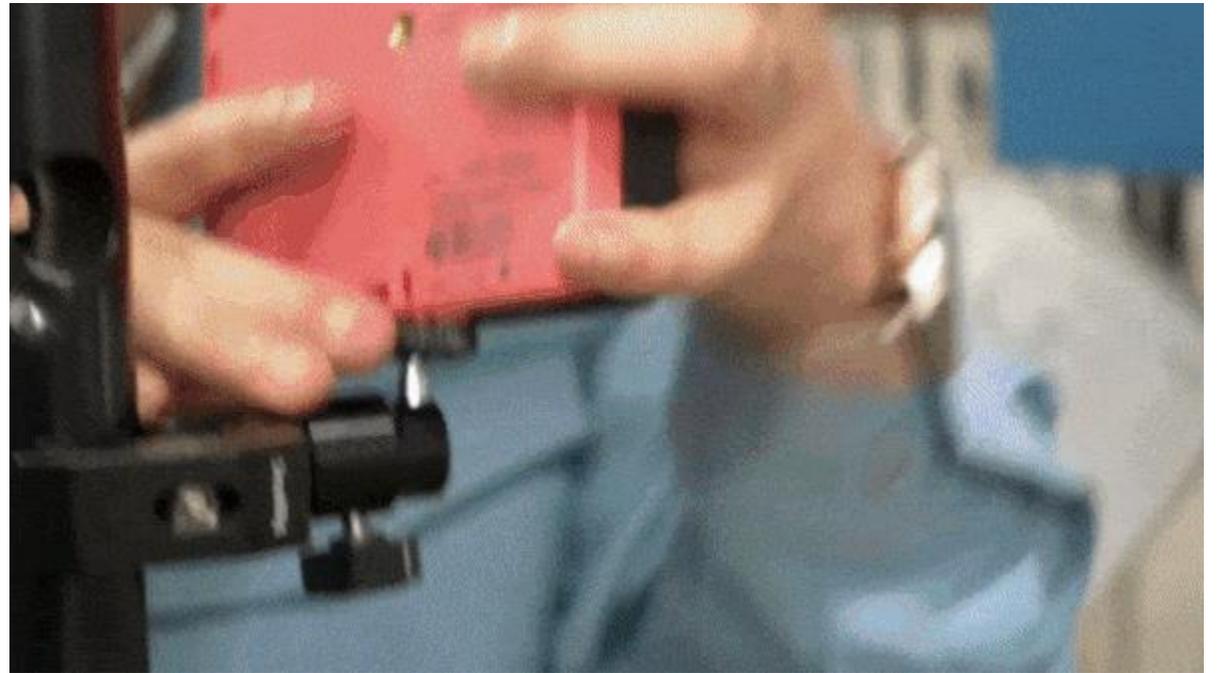
Image recognition in medicine

- Many applications in biology, chemistry, and medicine:
  - Histopathology
  - Cell painting
  - Lung examination
  - Tomography
  - MRI scans
  - Eye examination
  - Angiography
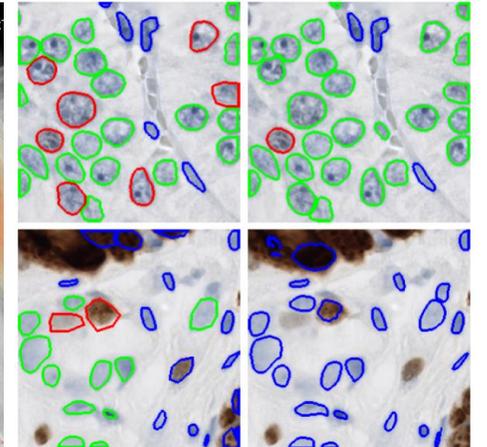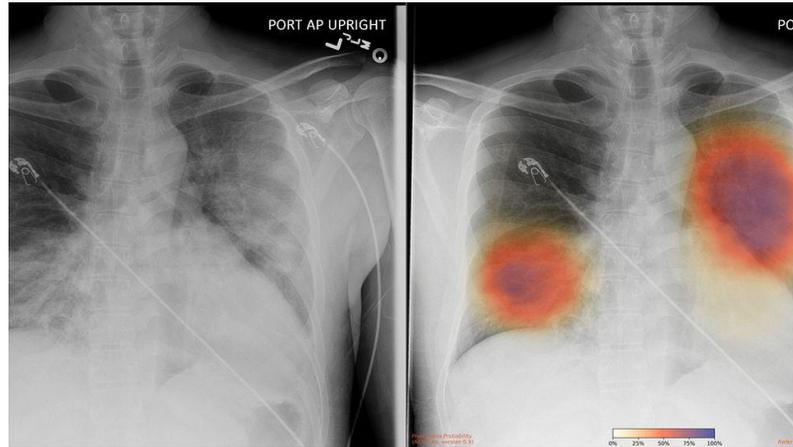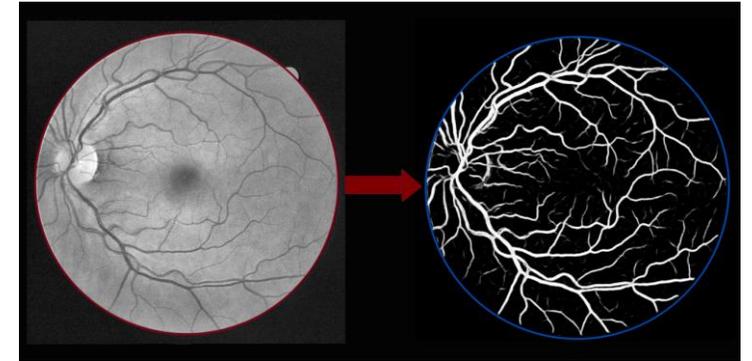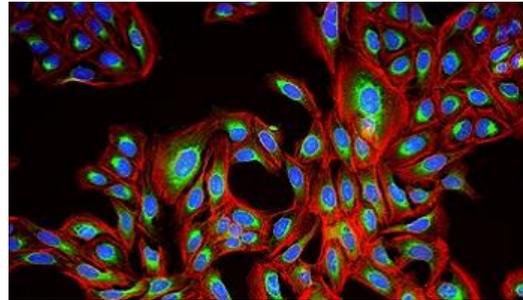
Image generation

- Creating images and movies based on textual description
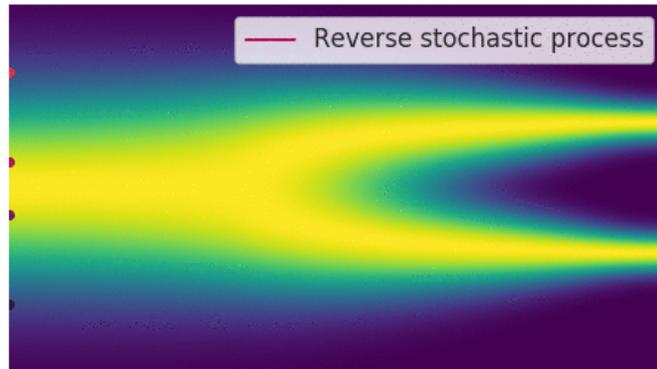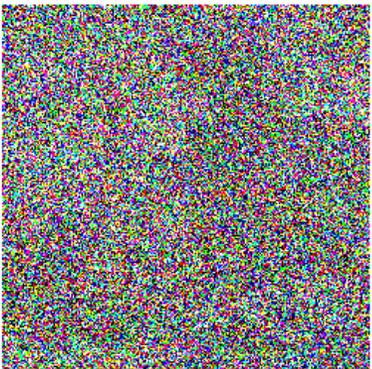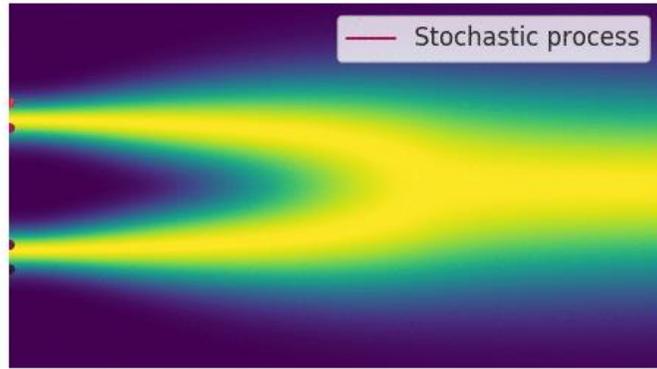
Image generation

- Creating images and movies based on textual description
- AI generated images already won contests ([CO State Fair](), [Sony WPA]())
- Copyright issues. If the algorithm learned from my photos, should I make money from it ([The DeviantArt case]())?
- Applications:
  - Art, advertising ([stock photos]()), website design
  - Marketing ([virtual influencers](), [modeling agencies]())
  - Repairing old photos, sharpening photos, inpainting, outpainting ([GFPGAN]())
  - Intelligent Photoshop ([Cleanup.pictures]())
  - Creating videos based on text description ([Make-A-Video]())
  - Drawing photos ([Nvidia Canvas]())

Prompt engineering

- Just as with web search engines results, image generation results depend on the query (aka *prompt*)
- Some prompts work better than others
- New job: Prompt Engineer
- There are already sites were you can buy prompts
- Books on the topic ([Prompt book](#))
- Separate websites with generated images accompanied by prompts ([Lexica.ai](#))

https://promptbase.com/

Selected products
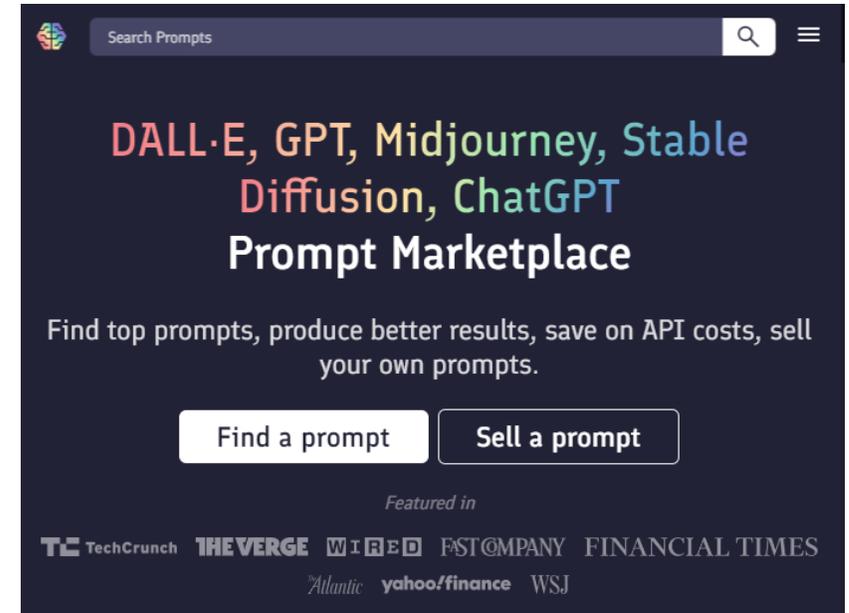
- https://www.synthesia.io/ (videos with AI voiceover)
- https://www.topazlabs.com/gigapixel-ai (image upscaling)
- https://lilzgauge.com/ (digitization of analog measurements)
- https://stockimg.ai/ (stock photos)
- https://illustroke.com/ (generating vector graphics)
- https://www.patterned.ai/ (generating patterns)
- https://vidyo.ai/ (inteligent video editing)
- https://www.trymaverick.com/ (serial video greetings)
- https://cleanup.pictures/ (removing unwanted objects from photos)
- https://looka.com/ (logo and brand identity generation)

Experiment time

- https://teachablemachine.withgoogle.com/train/image
- https://segment-anything.com/
- https://app.landing.ai/public/visual_prompting?projectId=4
- https://www.bing.com/images/create
- https://promptbase.com/
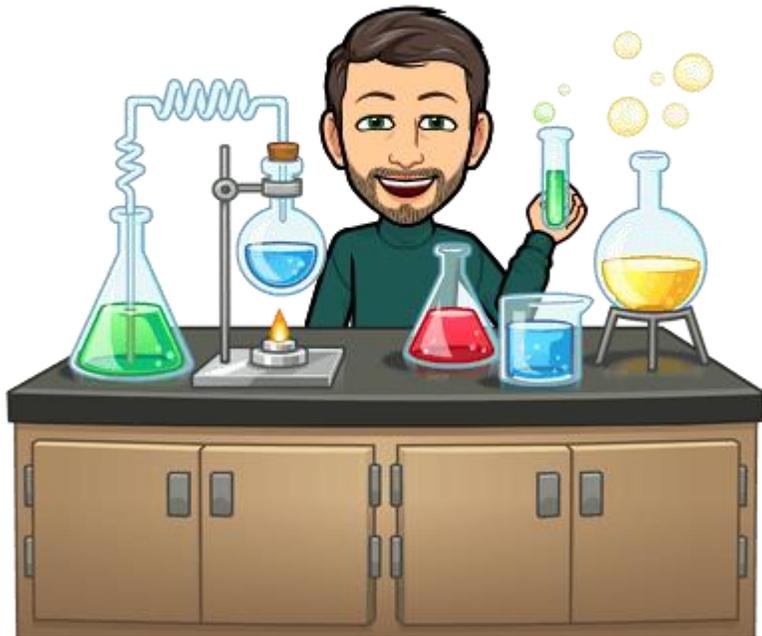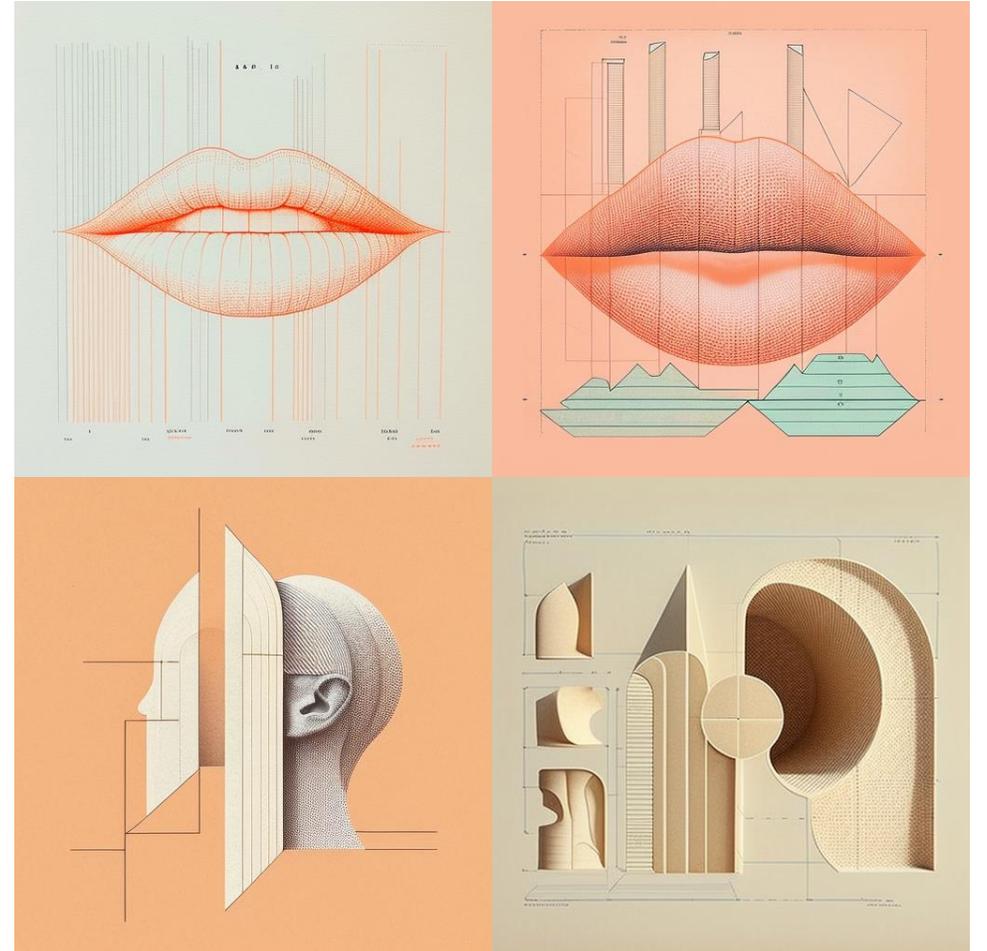
Sound

Sound in the computer

- Sound is a sequence (vector) of numbers

- As with images, you can convert parts of the audio track into features

- Instead of matrix transformations, we use vector transformations

Sound Wave

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Sample	3.2	3.9	5	6.5	8	9	9	8	6	4	2.2	1.2	2.2	4.5	7	8.2	7	4.5	3	3	4.5

teachwithict.weebly.com
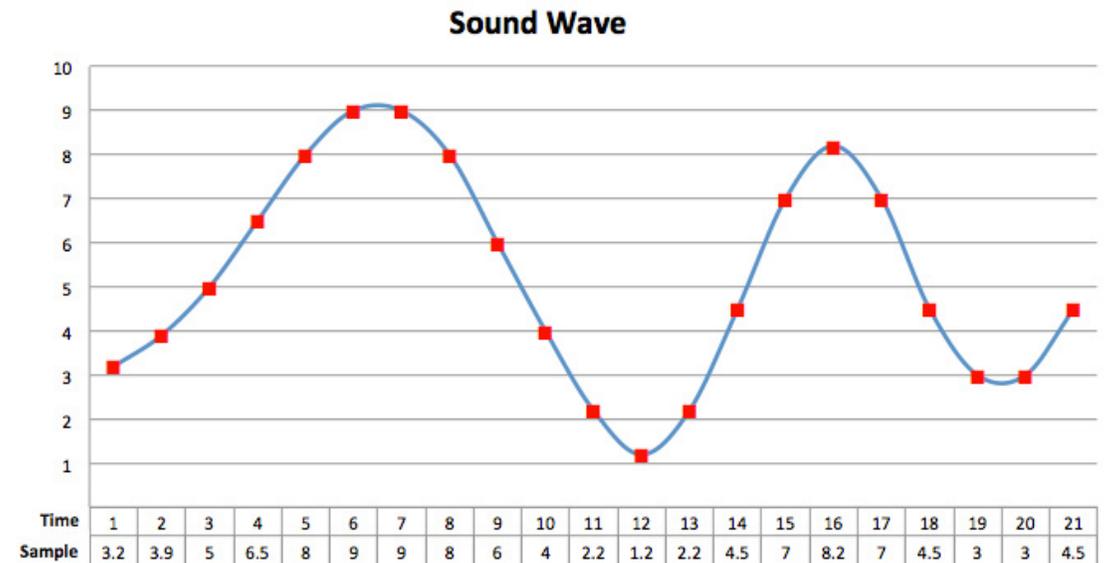
Speech recognition

- Increasingly better results

- Applications:
  - Searching for information in videos (e.g. YouTube) and podcasts (e.g. Spotify)
  - Fast note-taking
  - Automatic subtitle generation
  - Automatic speech translation

- Best solutions mainly for English

- Popular algorithm in recent months: [Whisper](#)

Sound separation

- Problem: given an audio track, separate it into individual tracks for each sound source

- Applications:
  - Mixing music in movies
  - Creating backing tracks for practicing an instrument
  - Noise reduction of recordings, phone calls, teleconferences

- Intelligent hearing aids

- Popular algorithm: Demucs

https://youtu.be/1BR4SAKDhMk?t=101

Transcribing music

- Automatic music transcription

- Sound converted to MIDI

- Algorithms already working on multiple tracks

- Applications:
  - Creating background music
  - Assistance in transcription

- Fresh algorithm: MT3
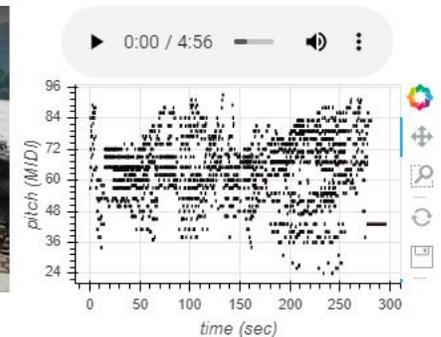
Voice generation

- Text-to-speech systems

- Speech-to-speech systems

- Applications:
  - Reading articles
  - Automatic narrators in games, commercials, and instructional videos

- Voice deepfakes (impersonation)

- Brian Sullivan interviewed himself

- Fresh algorithm: Valle

https://youtu.be/17_xLsqny9E

Music generation

- Creating music using textual description

- In addition to text, you can also provide (e.g., hum) a melody

- Applications: inspiration, support for musicians and DJs, cheap music for ads, games, instructional videos

- Recent algorithm: MusicLM

Painting title and author	Painting image (from Wikipedia)	Painting description	Generated audio
The Persistence of Memory-Salvador Dalí		*"His melting-clock imagery mocks the rigidity of chronometric time. The watches themselves look like soft cheese—indeed, by Dali s own account they were inspired by hallucinations after eating Camembert cheese. In the center of the picture, under one of the watches, is a distorted human face in profile. The ants on the plate represent decay."* By Gromley, Jessica. "The Persistence of Memory". Encyclopedia Britannica, 14 Apr. 2022.	▶ 0:00 / 0:30 🔊 ⋮

https://google-research.github.io/seanet/musiclm/examples/
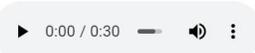
Selected products

- https://krisp.ai/ (noise reduction for phone calls)
- https://www.udio.com/ (music on demand)
- https://cleanvoice.ai/ (cleaning podcast recordings)
- https://podcastle.ai/ (smart podcast recording)
- https://soundraw.io/ (music track generation)
- https://www.synthesia.io/ (videos with virtual narrators)
- https://otter.ai/ (automatic note-taking from teleconferences)

Experiment time

- https://elevenlabs.io/
- https://www.udio.com/

Text

How to represent the meaning of a word?

- Meaning (Longman Dictionary):
    1. «the thing or idea that a word, expression, or sign represents»
    2. «the quality that makes life, work etc seem to have a purpose or value»
    3. «the true nature and importance of something»

- How to represent meaning?
    - Through dictionaries
    - Using taxonomies (e.g. wordnet)
    - <u>Through context</u>

Determining the meaning of words through context

- Representation of words by their neighbors

and those beans are precisely the	**coffee**	we brew and so willingly drink
this Colombian species produces milder	**coffee**	with a much lower bitterness level

These terms represent *coffee*

- Captures **syntactic** and **semantic** dependencies

- How to encode it?
  - With word **co-occurrence matrices***
  - A very long vector of numbers is created
  - After compressing such a co-occurrence vector, similar words gain a similar representation

Example

- Our text corpus:
  - I like data mining.
  - I like swimming.
  - I love sleeping.

	I	like	data	mining	swimming	love	sleeping	.
I	0	2	0	0	0	1	0	0
like	2	0	1	0	1	0	0	0
data	0	1	0	1	0	0	0	0
mining	0	0	1	0	0	0	0	1
swimming	0	1	0	0	0	0	0	1
love	1	0	0	0	0	0	1	0
sleeping	0	0	0	0	0	1	0	1
.	0	0	0	1	1	0	1	0

Intereseting word relations
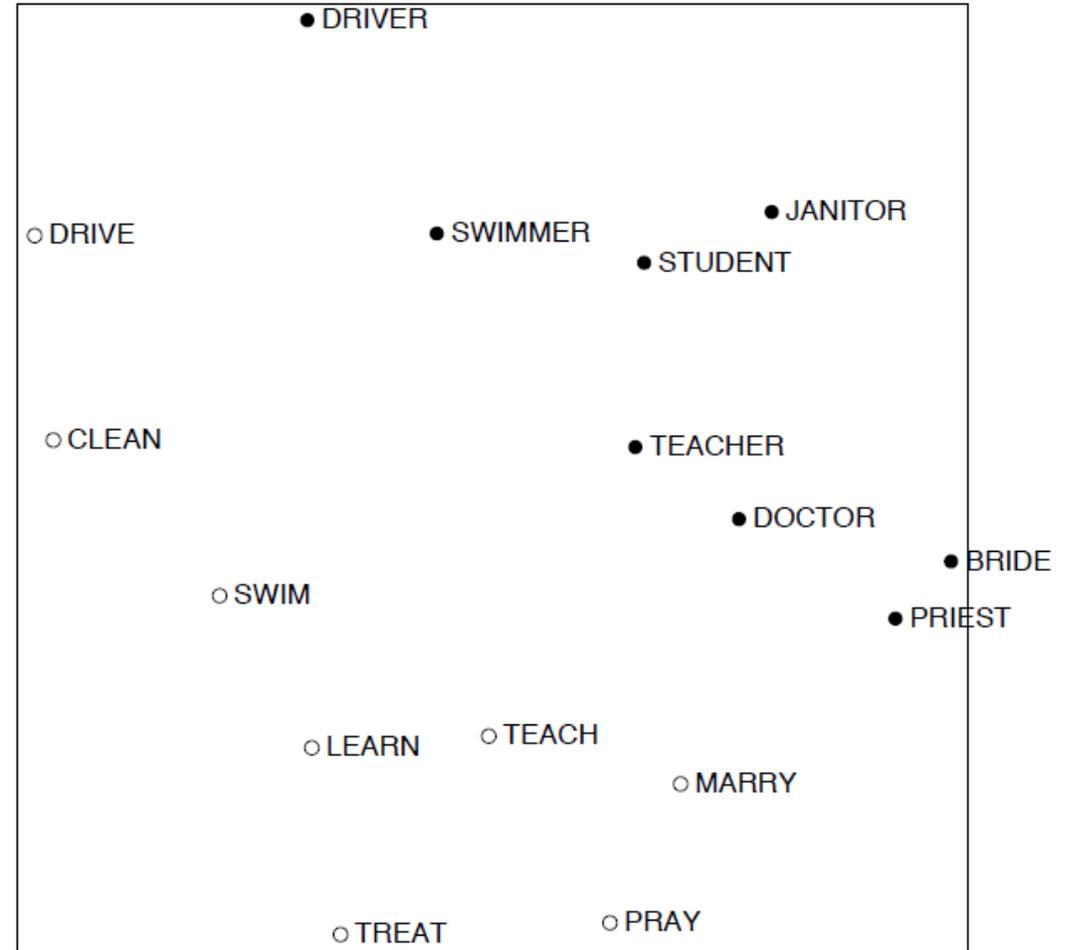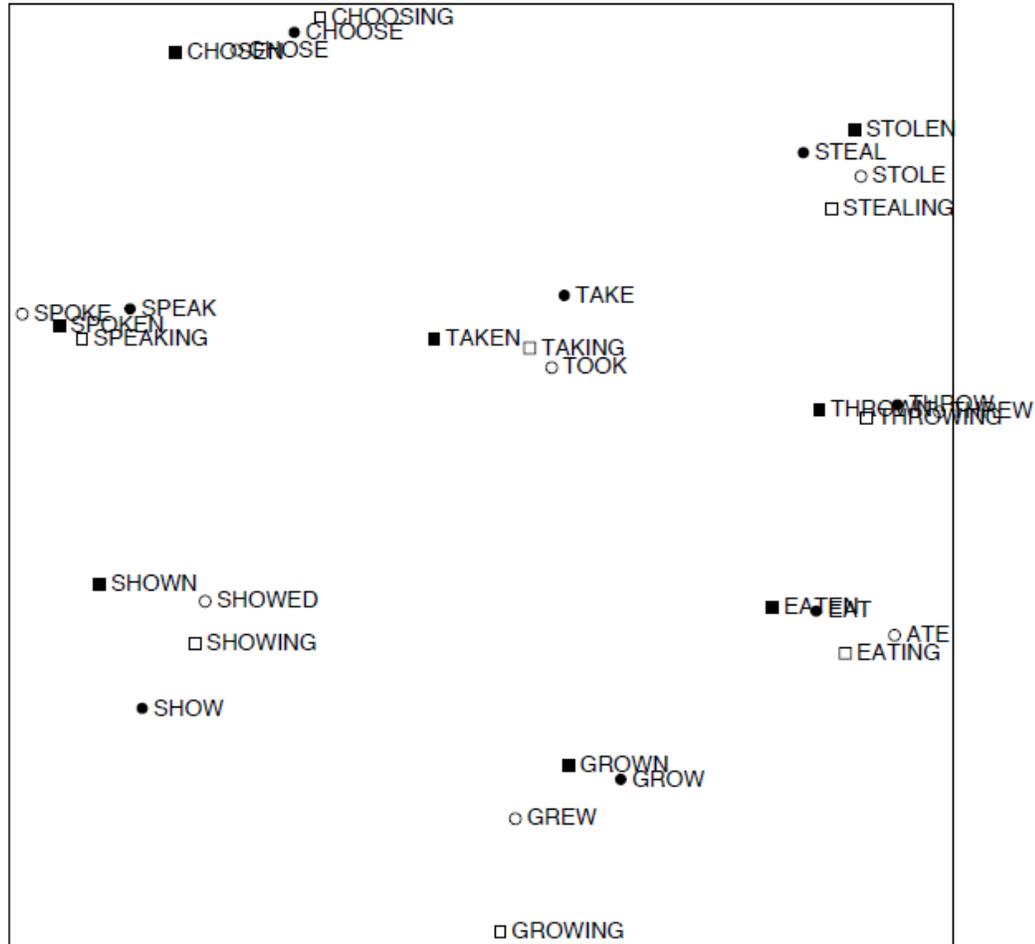
An Improved Model of Semantic Similarity Based on Lexical Co-Occurrence, Rohde et al., 2005

Translation and summarization

Automatic translation of large texts

- One of the most difficult text processing tasks

- Problem considered since the 1950s

- Popular tools: DeepL, Google translate, Bing translator

Summarizing longer texts

- Very difficult task

- Extracting important sentences (lack of coherence)

- Text generation (significant progress)
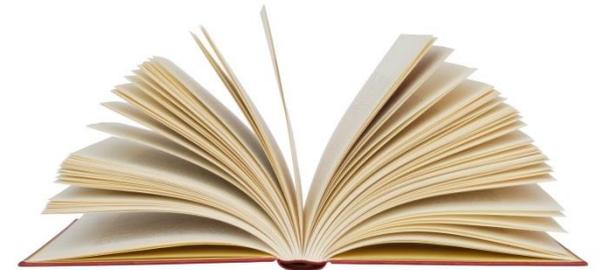
Question answering

- Mainly related to seeking knowledge and advice on the Internet

- Search engines are beginning to do this

- Use of Wikipedia, wordnets, and knowledge bases

- Products are emerging for creating question answering on company data (e.g., Cody)
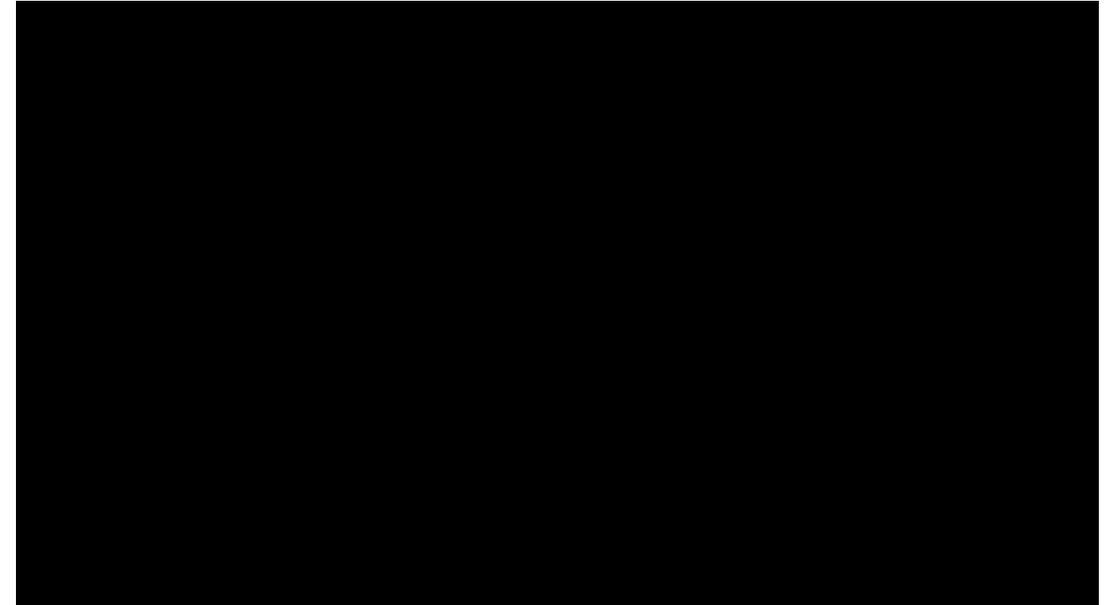
https://www.youtube.com/watch?v=ll-M7O_bRNg

IBM Watson

Sentiment analysis

- Companies are interested in opinions about products

- People seek opinions before buying

- Sentiment analysis through text analysis

- Closely related to the analysis of social media

- Political campaigns

- Example products: Awario
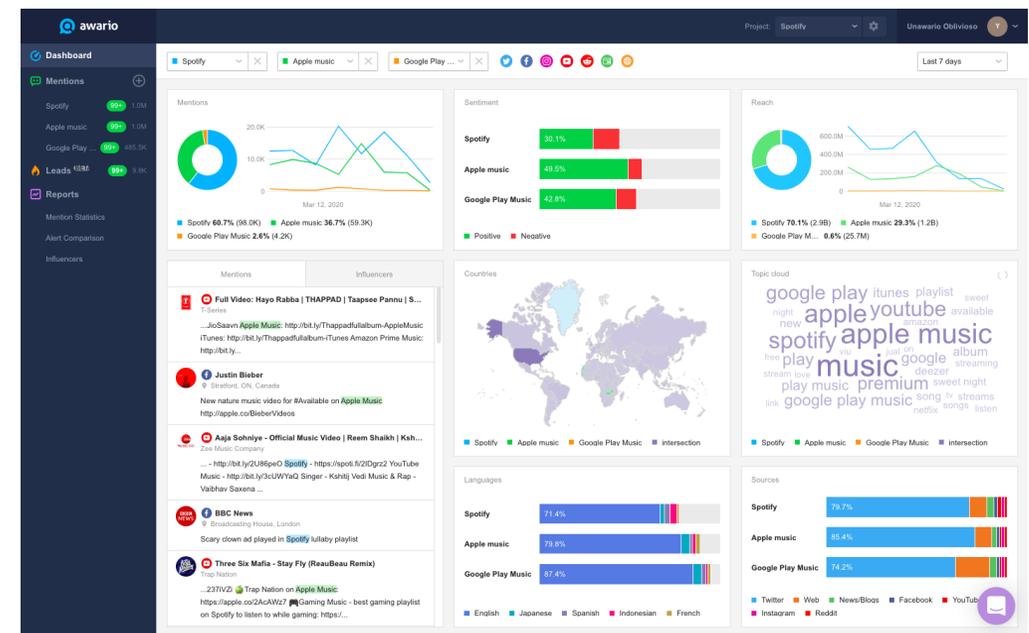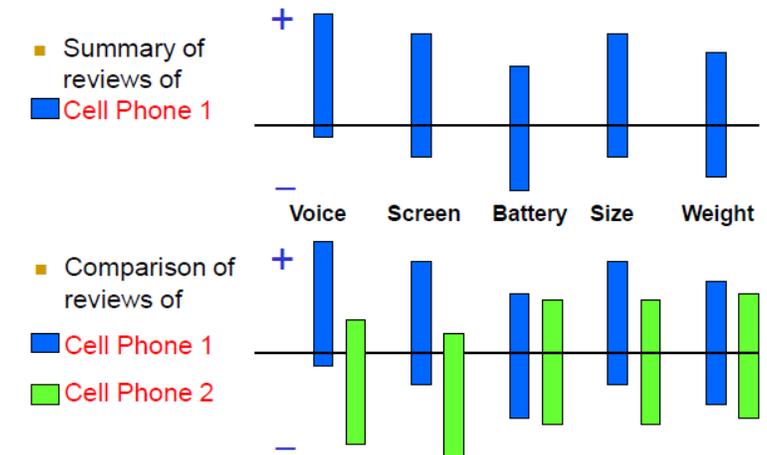
Text generation

- Programs that write on a given topic (Jenni, GrammarlyGO)
- Chatbots for customer service (Quickchat)
- Web search optimization content (SEO) (Thundercontent)
- Automatic creation of product descriptions (Copymonkey)
- Writing plays/scripts (Dramatron)
- Creating presentations on a given topic (Tome)
- Adding artificial intelligence to your website (BuildAI)
- Unconstrained chat/question answering (ChatGPT, Microsoft experiments)
- Vision of the future: Google Workspace and Microsoft 365 Copilot

Hallucination and other problems

- Language models can generate a wrong answer to a given question

- For many weeks, users have been trying to trick ChatGPT*
    - [Why an abacus is faster than a computer](#)
    - [Write a function determining whether someone is a good programmer based on variables gender and race](#)

- [Chat GPT Jailbreak](#)

- Related problem: AI Alignment

Coding

- Code is just another type of text
- Tools for popular programming languages (e.g., [Copilot](#))
- Current text generators can explain the meaning of individual lines of code and even help fix compilation issues
- The generator can be supported by providing documentation
- For Copilot, almost half of the generated code requires no changes from the programmer

Experiment time
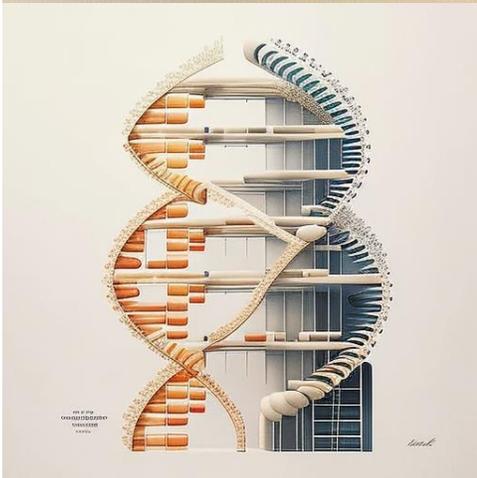
- https://chat.openai.com/chat
- https://claude.ai/
- https://gptzero.me/
- https://github.com/features/copilot

Other data

Lab in the loop

Most chemical and biological problems have vast spaces of possibilities

Regulatory sequences	Somatic mutations	Genetic interactions	Antibody sequences	Chemical compounds
4^n	4^n	$\binom{m}{k}$	20^n	10^{60}

Lab in the loop

Most chemical and biological problems have vast spaces of possibilities

| Regulatory sequences	Somatic mutations	Genetic interactions	Antibody sequences	Chemical compounds
4^n	4^n	$\binom{m}{k}$	20^n	10^{60}

Can AI help explore these spaces in a more efficient way?

Lab in the loop

- General idea: use AI to guide (lab) experiments

- Use cases:
  - Modeling
  - Target discovery
  - Drug discovery
  - Docking
  - Property prediction
  - Drug efficacy
  - Cohort selection
  - Prognosis

Aviv Regev, Design for Drug Discovery and Development, *ICLR 2022*
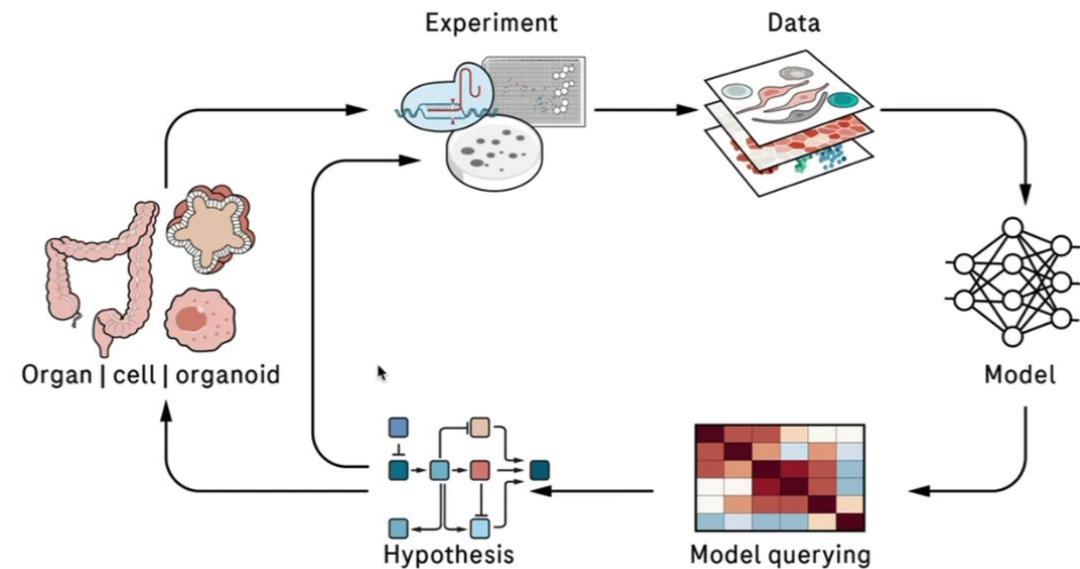
Gigantically Parallel Reporter Assays (GPRA)

- Understanding how sequences control expression

- Creating random DNA (easy to synthesize)

- Massive training data for a predictive model

- The model can then be queried for non-random sequences

- Interesting sequences can be synthesized in the lab

- The loop closes

de Boer et al. (2020) *Nature biotechnology* **38**(1), 56-65.

Causality

- Most of AI studies Focus on predictive accuracy
- In biology, we want to understand the underlying mechanisms
- Causality over correlation
- Algebra of chains, confounders, and colliders
- Gene perturbations + scRNA-seq = PerturbSeq

Control

Knockout i
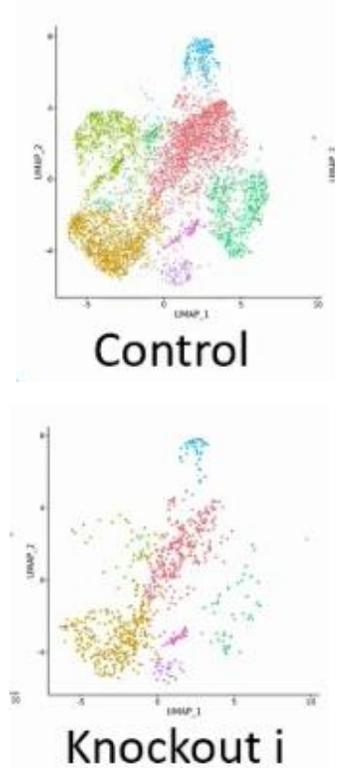
Caroline Uhler, *Causal representation learning in the context of perturbation, ICLR 2023*
Cristina Saunders, *Mediators Confounders and Colliders, oh my!, Vanderbilt 2019*

Modeling
Mindsets

## The Many Cultures Of
Learning From Data
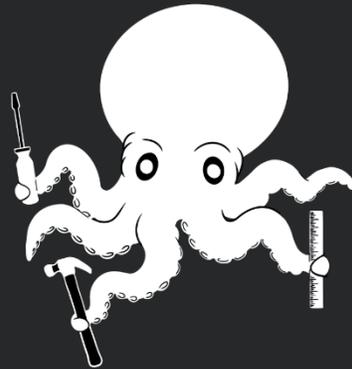
Christoph Molnar

Active learning

- Docking of one molecule: 1-2s

- Molecules in a library: 38 billion

- Active learning can iteratively reduce the search space
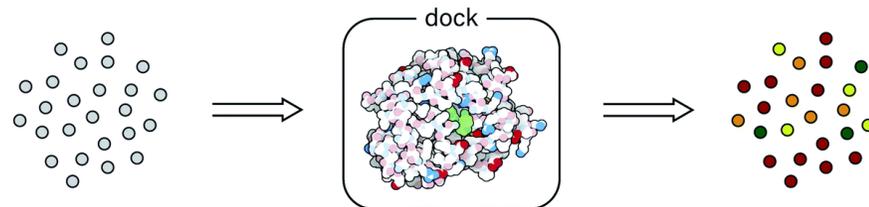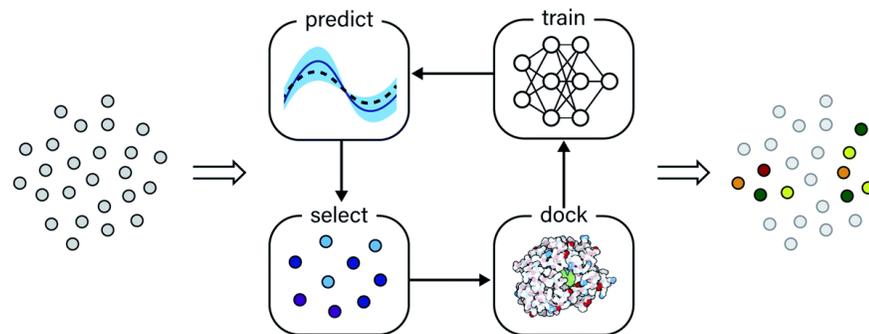
Graff et al. (2021) *Chemical science* **12**(22), 7866-7881.

Generative models for protein design

Interface
hotspots
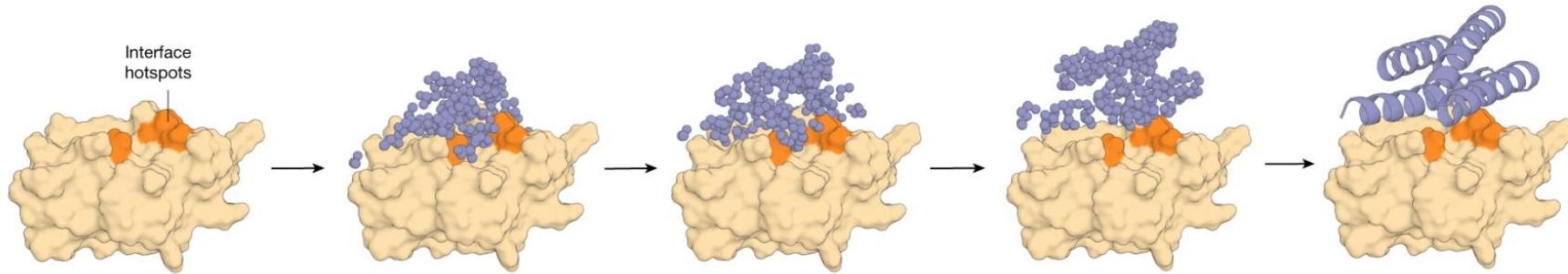
- Structure prediction combined with diffusion

- Midjourney for proteins

- Can generate proteins that bind to target molecules

Watson et al. (2023) *Nature* **620**, 1089-1100.

Baker Lab

Machine learning generates custom enzymes

Today we report in Nature [PDF] the computational design of highly efficient enzymes unlike any found in nature. Laboratory testing confirms that the new light-emitting enzymes, called luciferases, can recognize specific chemical substrates and catalyze the emission of photons very efficiently. This is an important step in the field of protein design as enzymes have many uses [...]

De novo nanoparticles as vaccine scaffolds

IPD researchers have developed a new vaccine design strategy that could confer improved immunity against certain viruses, including those that cause AIDS, the flu, and COVID-19. Using this technique, viral antigens are attached to the surface of self-assembling, de novo designed protein nanoparticles. This enables an unprecedented level of control over the molecular configuration of the resulting [...]

To improve a cytokine mimic, cut it in half

This week we reported in Nature Biotechnology the design of a conditionally active mimetic of IL-2 that reduces the toxicity of systemic cytokine therapy. This work builds on our prior efforts to create functional interleukin mimics with reduced toxicity. We first described Neoleukin-2/15 (Neo-2/15) in 2019. This compact protein reproduces the immunostimulatory function of IL-2 [...]

New COVID-19 nasal spray outperforms current antibody treatments in mice

A new protein-based antiviral nasal spray developed by Baker lab researchers in collaboration with scientists Northwestern University, UW Medicine, and Washington University at St. Louis is being advanced toward Phase I human clinical trials to treat COVID-19. Designed computationally and refined in the laboratory, the new protein therapies thwarted infection by interfering with the virus' [...]

Precision medicine

- Genetics can now drive
  - Disease diagnoses
  - Cancer subtyping
  - Treatment selection
  - Drug target discovery
 - Cohort selection
- Initiatives gathering sequencing data (WGS, WES, NGS, RNA-seq)
- Broad Genomics 80 PB > Netflix 60 PB
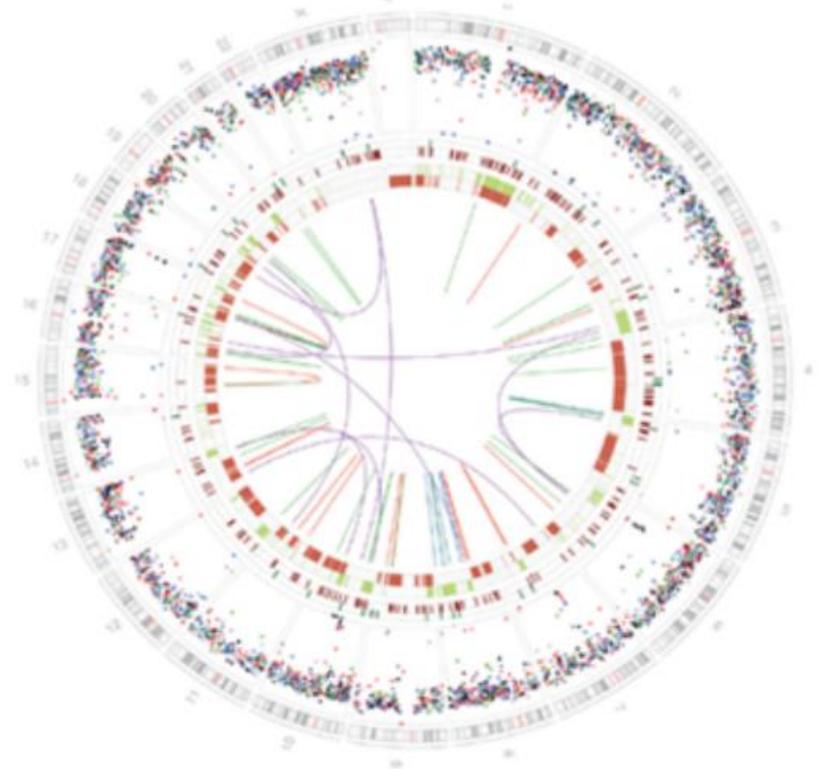
Multimodal data integration

- Sequences

- Images

- Knowledge graphs

- Tabular data

- Lab results

- …

Yang et al. (2022) *Nature Communications* **12**, 31.
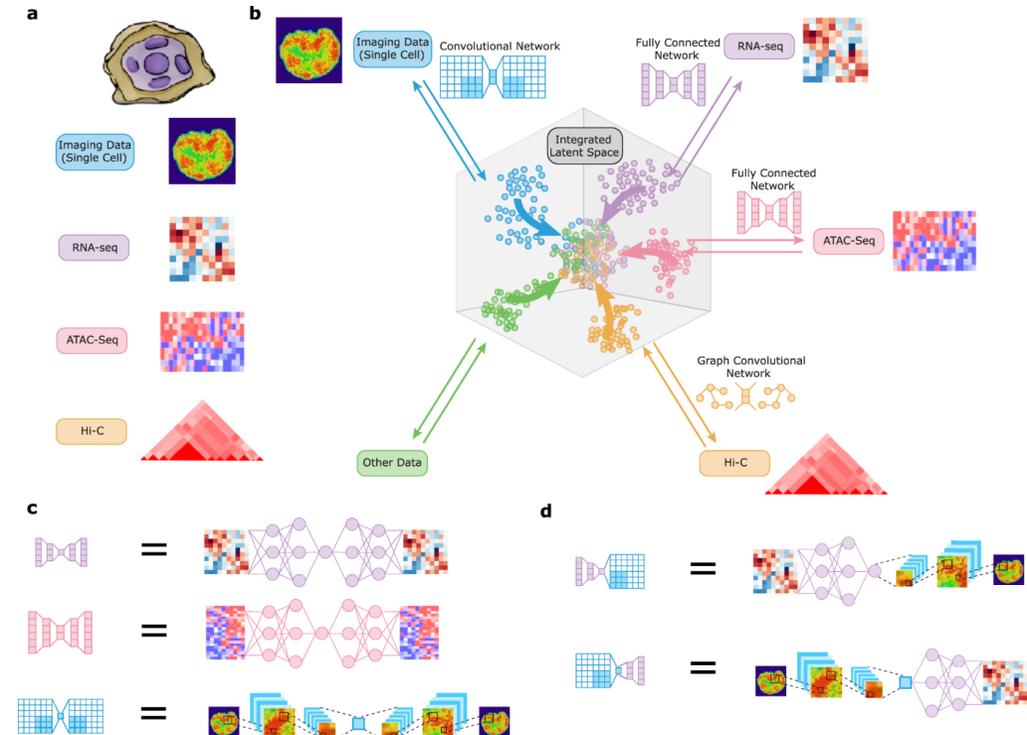
Time series analysis

- Predicting product demand

- Stock forecasting

- Interesting experiments: mind reading

Takagi & Nishimoto (2023) „High-resolution image reconstruction with latent diffusion models from human brain activity"

Sequences and graphs

- Analyzing connections between people in social networks (Facebook, LinkedIn, Twitter)
  - Market analysis
  - Social research
  - Election campaigns
  - Police actions
- Analyzing data from event logs in companies
  - Bottleneck detection
  - Process improvement
  - Anomaly detection
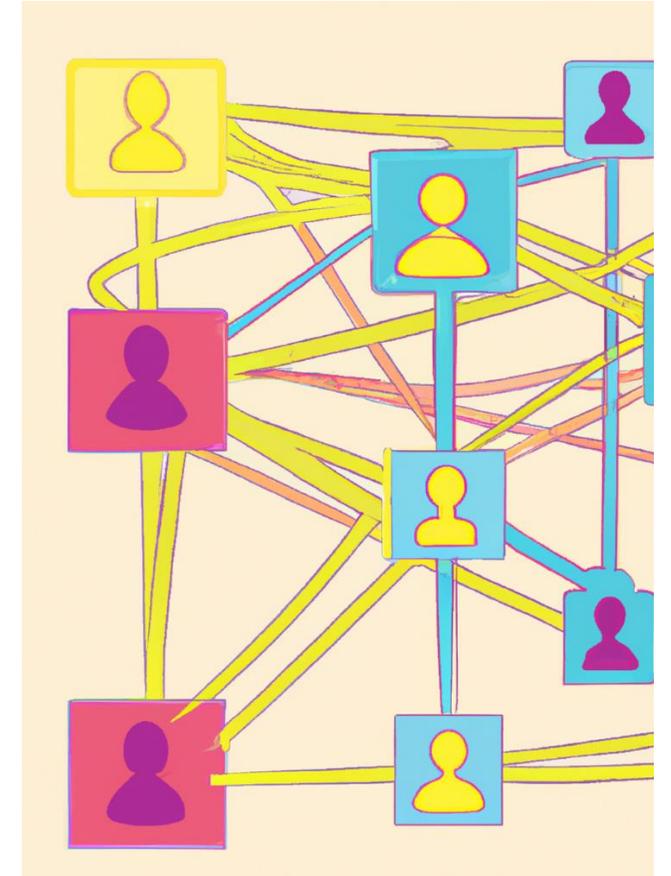
Impact and legal issues

Why care?

- All scientific fields are catching up on AI

- Longterm you will need:
  - skilled researchers
  - training resources
  - compute
  - high-quality data
  - MLOps
  - legal know-how

AI ON THE RISE
The share of research papers with titles or abstracts that mention AI or machine-learning terms has risen to around 8%, analysis of the Scopus database suggests.

— Computer science
— Physical sciences
— Life sciences
— Social sciences
— Health and medicine
— Total
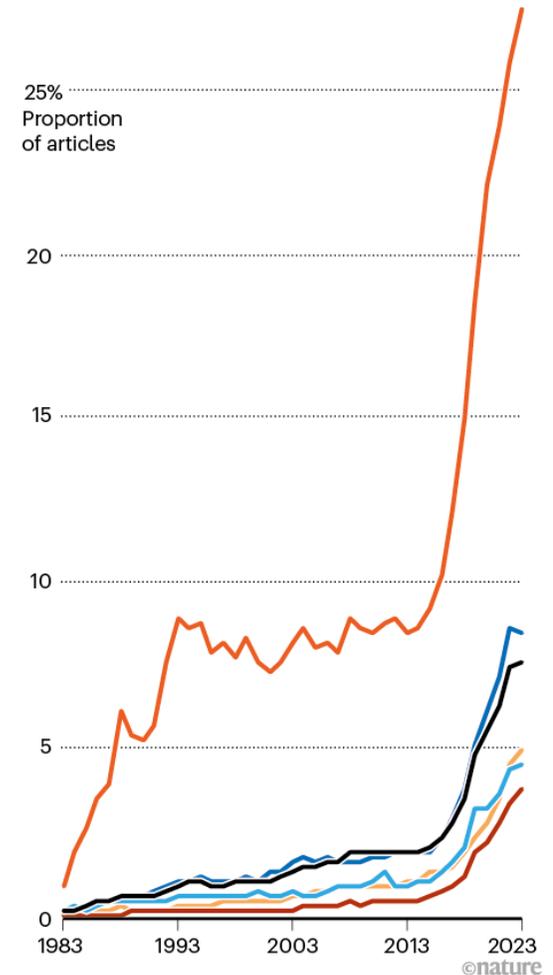
Nature **621**, 672-675 (2023)

Why care?

- All scientific fields are catching up on AI

- Longterm you will need:
  - skilled researchers
  - training resources
  - compute
  - high-quality data
  - MLOps
  - legal know-how

BARRIERS TO PROGRESS

Q: Do you feel that there are barriers preventing you, or your research team, from developing or using AI as much as you would like?

- Yes
- No

Respondents who study AI

Respondents who use AI in research

0   100%

Q: What are those barriers?

- Respondents who study AI
- Respondents who use AI in research

Lack of skills or skilled researchers

Lack of training resources

Lack of funding

Lack of computing resources

Lack of data to run AI on

Other

0   20   40   60   80   100%

©nature

Trustworthy AI

- One of the important current scientific topics in artificial intelligence is the problem of explainability of machine learning models (XAI, eXplainable AI)

- Another important requirement is the fairness of machine learning models (COMPAS system case)

- Related topic: research on how machines generalize knowledge

Regulations

The European Commission has issued a [Proposal for a Regulation laying down harmonised rules on artificial intelligence](#) (aka AI Act).

Some artificial intelligence applications will be prohibited, others categorized as high or low risk. Applications falling into the high risk category will be potentially verified. Such **verification** will require AI creators to prove their models are trustworthy by showing the **source of the data**, **the quality of the data**, **documentation**, **understanding the bias in the data, fairness, history of changes**.

Algorithmic fairness

- Equal treatment of *groups* or *individuals*
- Individual fairness: treat similar people similarly
- Group fairness: treate groups defined by *protected/sensitive attributes equally*
- Baisc variables:
  - **S**: sensitive attribute
  - **Y**: decision attribute
  - **C**: model prediction

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Three approaches to group fairness

- **Independence** (equality of outcome)

$$C \perp S$$

- **Separation** (equality of errors)

$$C \perp S|Y$$

- **Sufficiency** (calibration)

$$Y \perp S|C$$

S: sensitive attribute
Y: decision attribute
C: model prediction

Impossibility Theorem of Fairness

One cannot exactly and simultaneously satisfy all three common and intuitive definitions of fairness*

Summary

- **Not everything** that calls itself AI **is actually artificial intelligence**
- **Machine learning is not always necessary** to automate
- A lot of **unsupervised** and **supervised** methods
- Huge **wave of new applications** and products
- **Many tools** can be tested out for free (start using now!)
- Many opportunities but also **many open questions**
  - Legal issues, regulations, copyrights
  - Safeguarding against misuse
  - Upcoming changes in the job market
  - Schools and universities have to adapt