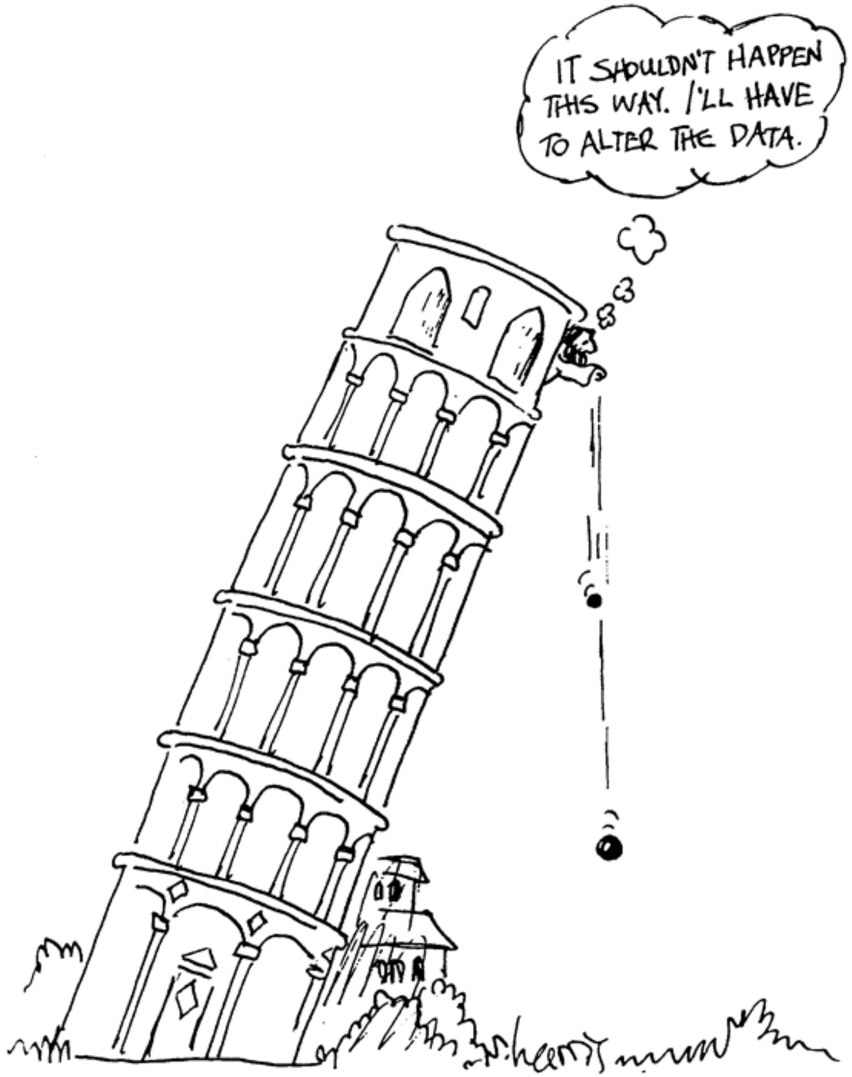
From Data to Learning

Announcements and Logistics

- Use website as reference: www.ml4science.com
- Assignment #0 posted: https://www.ml4science.com/content/hw/hw0/
- Use intro2ml.com as reference for math and python: https://intro2ml.com/schedule/
- Purpose of class: learn how to learn. **How?**
  - For scientific fields that are changing quickly, reading papers is the most reliable: https://www.scholar-inbox.com/. + YouTube, blogs, etc.
  - **Simple formula:** read, clone GitHub repository, reproduce results, come up with ideas, repeat
- Main deliverable: **presentation and paper.**
- **Learn from each other:** Discussions, questions, collaborations **on Slack**

Finding *constants* of nature that *generalize* across space and time



Galileo


676 EPITOMES ASTRONOMIÆ

LIBRI V.

PARS ALTERA.

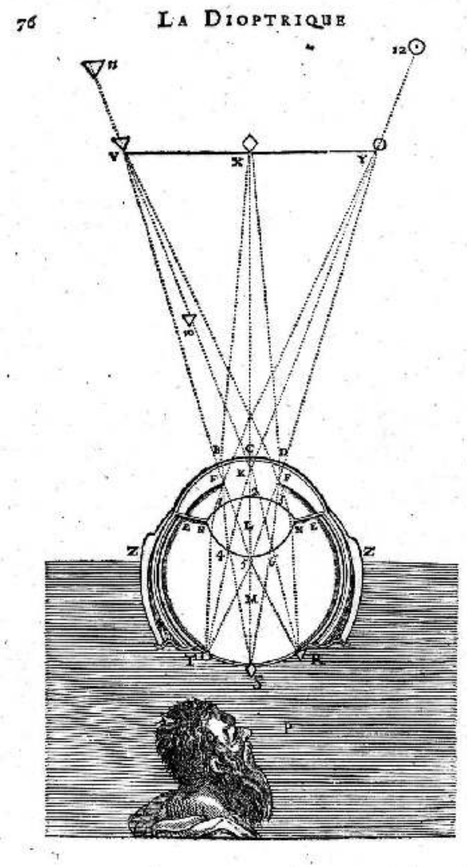
DE TERMINIS ASTRONOMICIS ET CALCVLO, EX ORBITA Excentrica orientibus.

Quomodo appellatur Orbita cuiusque Planetæ?

APpellatur veteri voce Eccentricus subaudi, Circulus. Etsi enim orbitæ sunt Ellipticæ, vt hic PERI. quæ habent duo quasi centra A L. quæ physicè Foci dicimus; & in eorum altero A. Sol ipse vt centrum mundi, inest: tamen etiam

punctum inter focos medium, vt B. à scriptoribus Conicis centrum figuræ, peculiari iure dicitur, & præter-

Kepler



Descartes

Compression



Measurements

Learning

World

Data

Modeling

Seen

Unseen

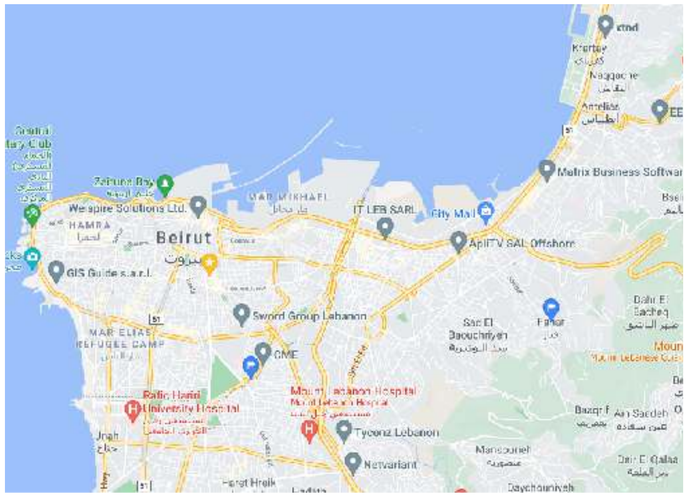
Inference


Projection

Reconstruction

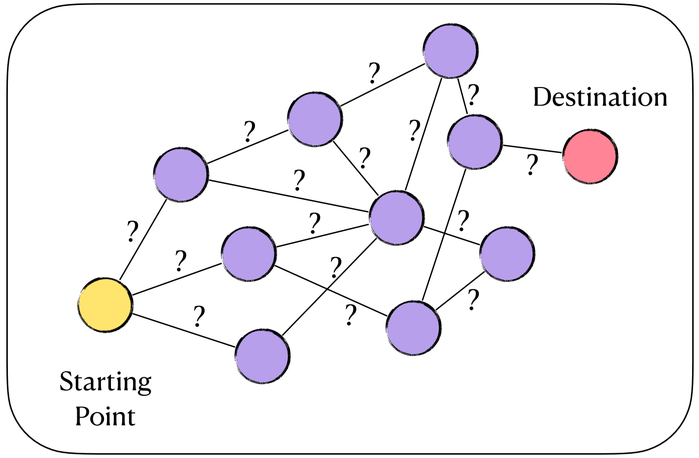
The 3 pillars of Artificial Intelligence

Real world

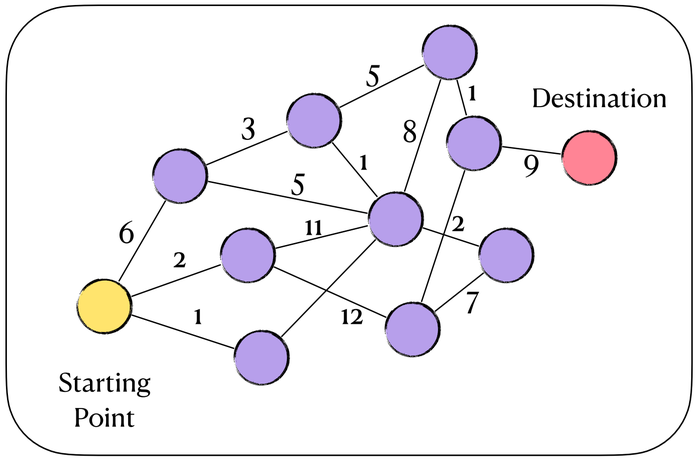


What's the shortest path from AUB to my home?

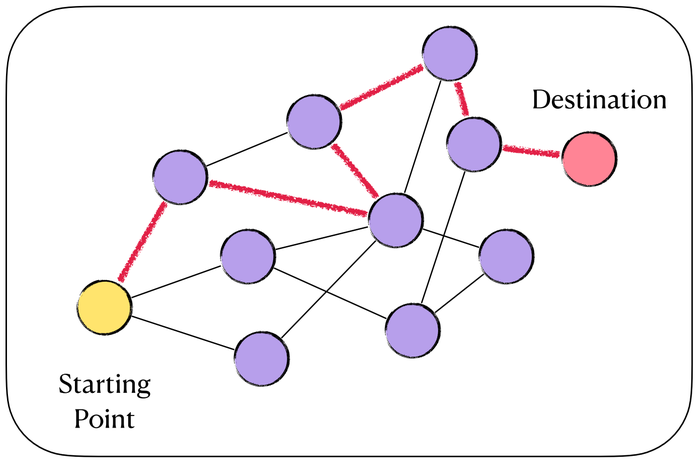
Learning



Modeling



Inference



The 3 pillars of Artificial Intelligence

Real world

Learning

$$\hat{p} = \arg\min_{p} \left\| \frac{\partial u}{\partial t} - \mathscr{L}(u(x,t);p) \right\|_2^2$$

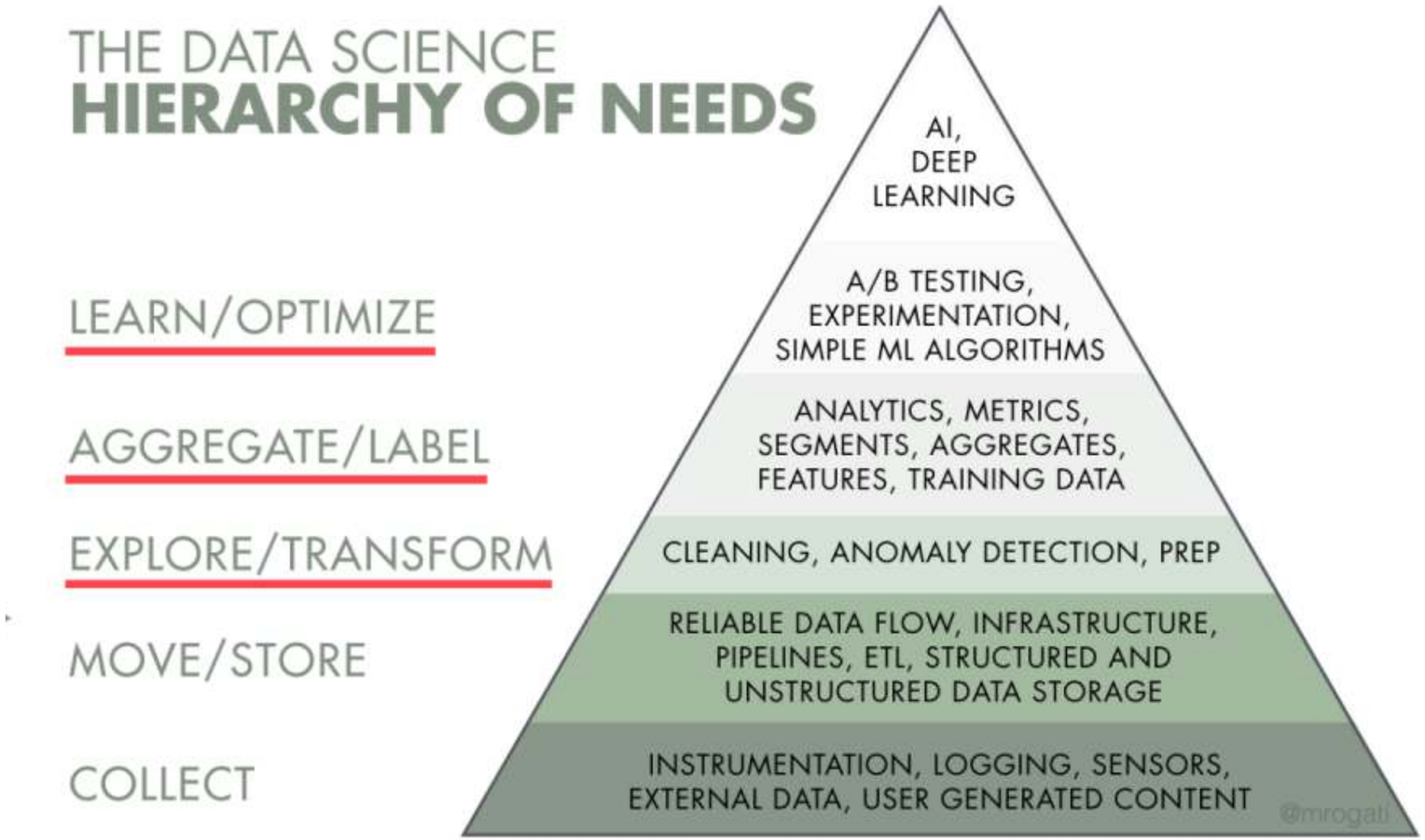
Inference

$$u(x_i, t_i) = \text{Integrator}\left(u(x_i, t_{i-1}), u(x_{i+1}, t_{i-1}), u(x_i, t_{i-1})\right)$$

Modeling

$$\frac{\partial u}{\partial t} = \mathscr{L}(u(x,t);p)$$

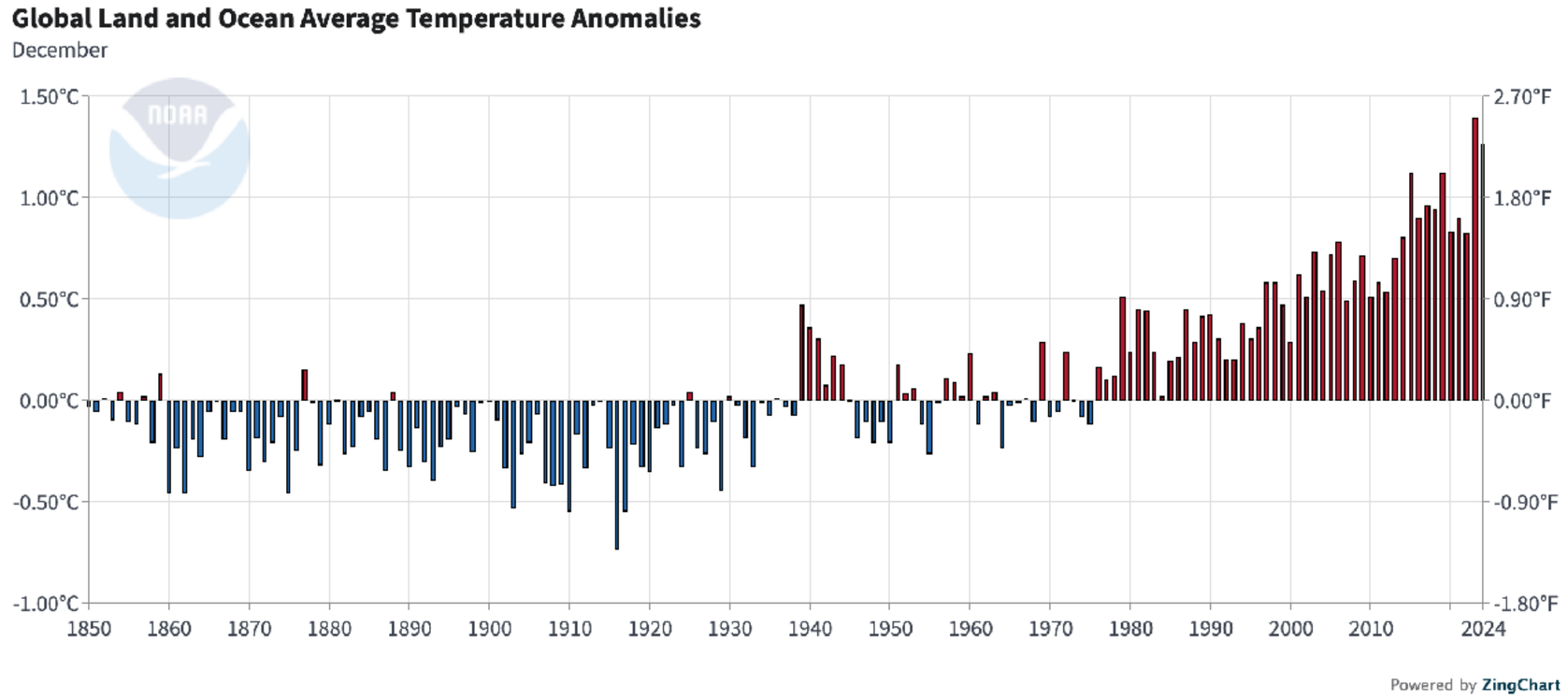
The data science hierarchy of needs



Data needs as illustrated by Monica Rogati

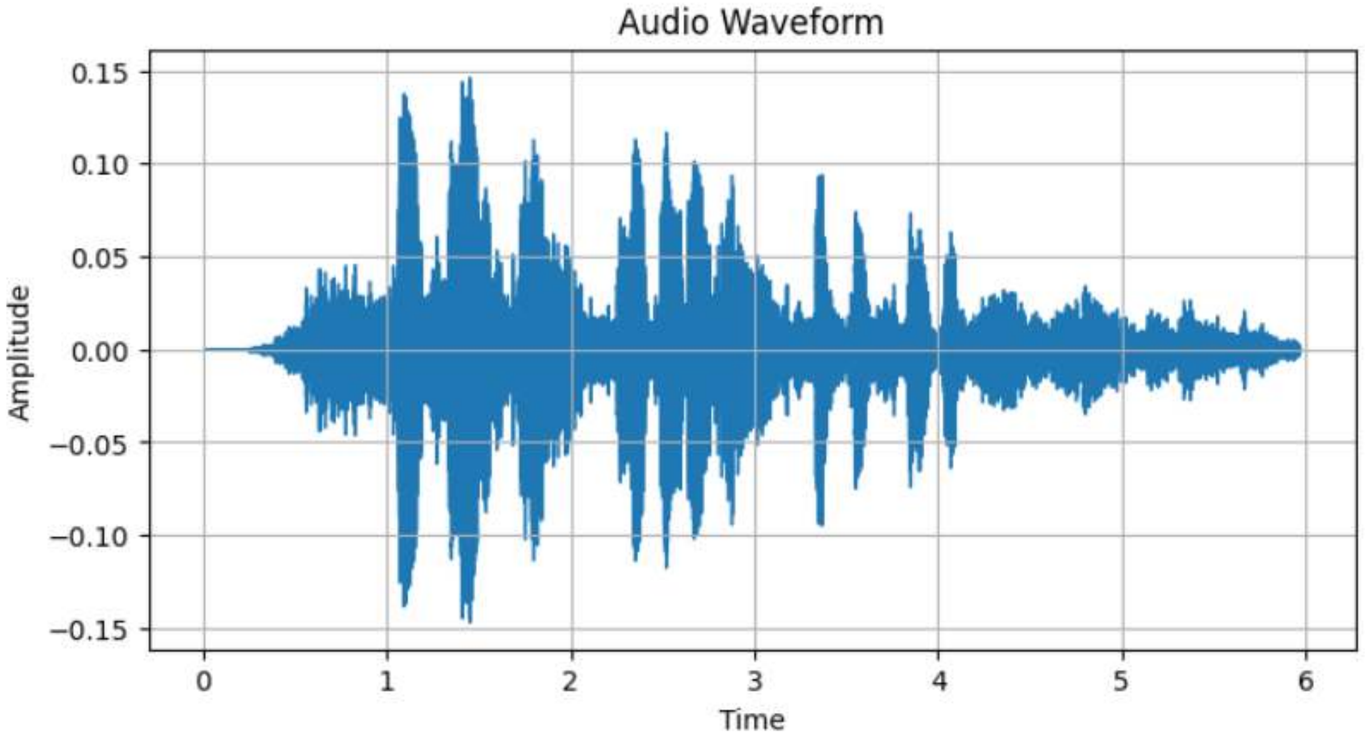
What **are*** Data?

Temperature



Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Temperature	0	1	-1	-2	-3	0	1	2	2	3	5	6	8	5	4	6

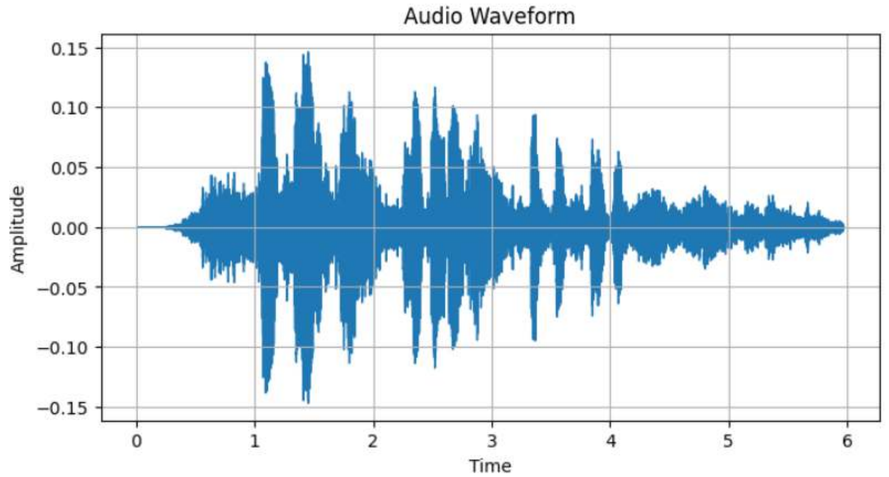
Audio



Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Amplitude	35	37	35	34	33	30	49	48	46	44	46	49	48	66	50	55

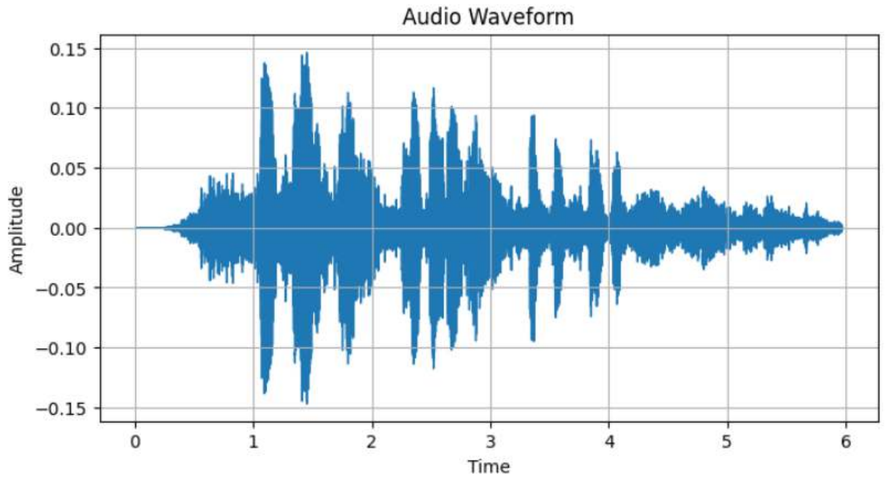
Audio

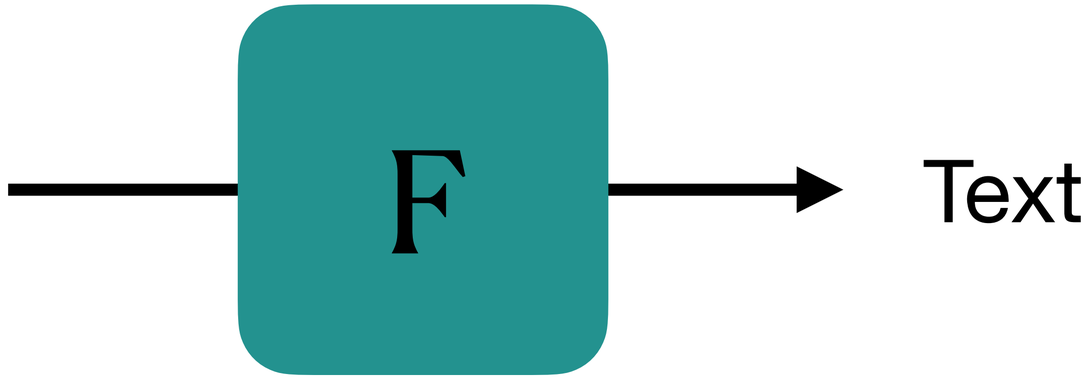
Text → F



Time	1.1	2.1	3.4	4.3	5	6	7	8	9	10	11	12	13	14	15	16
Amplitude	35	37	35	34	33	30	49	48	46	44	46	49	48	66	50	55

Audio





Time	1.1	2.1	3.4	4.3	5	6	7	8	9	10	11	12	13	14	15	16
Amplitude	35	37	35	34	33	30	49	48	46	44	46	49	48	66	50	55

Images

Google announces Waymo in 2010

Images

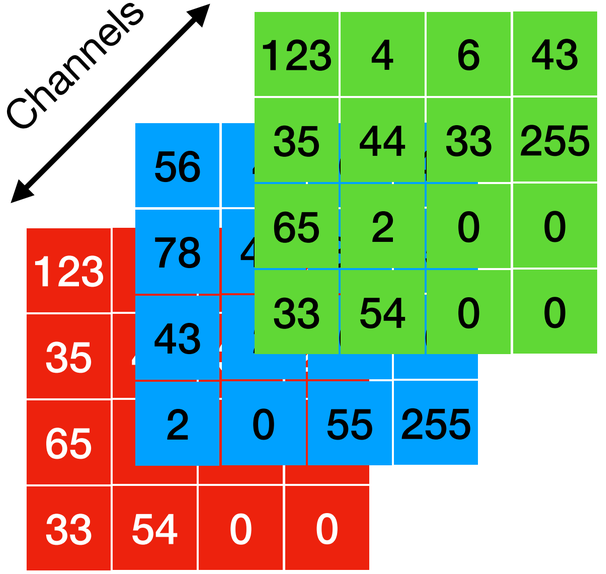
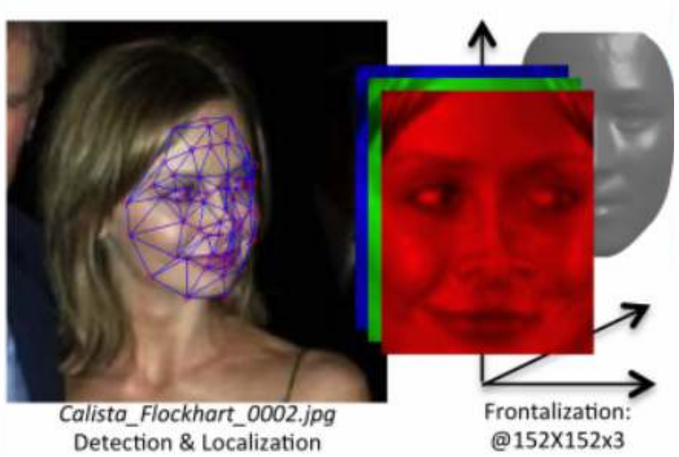
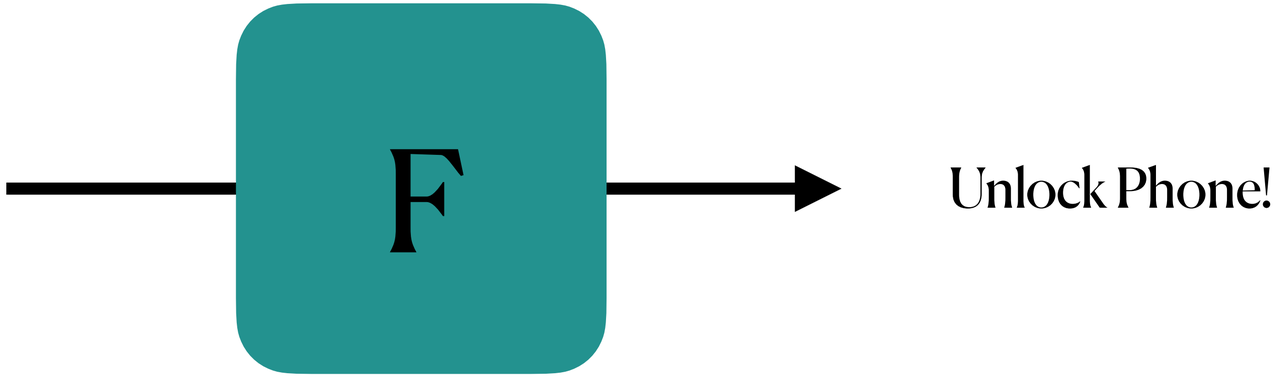


Image dimension: $h \times w \times 3$

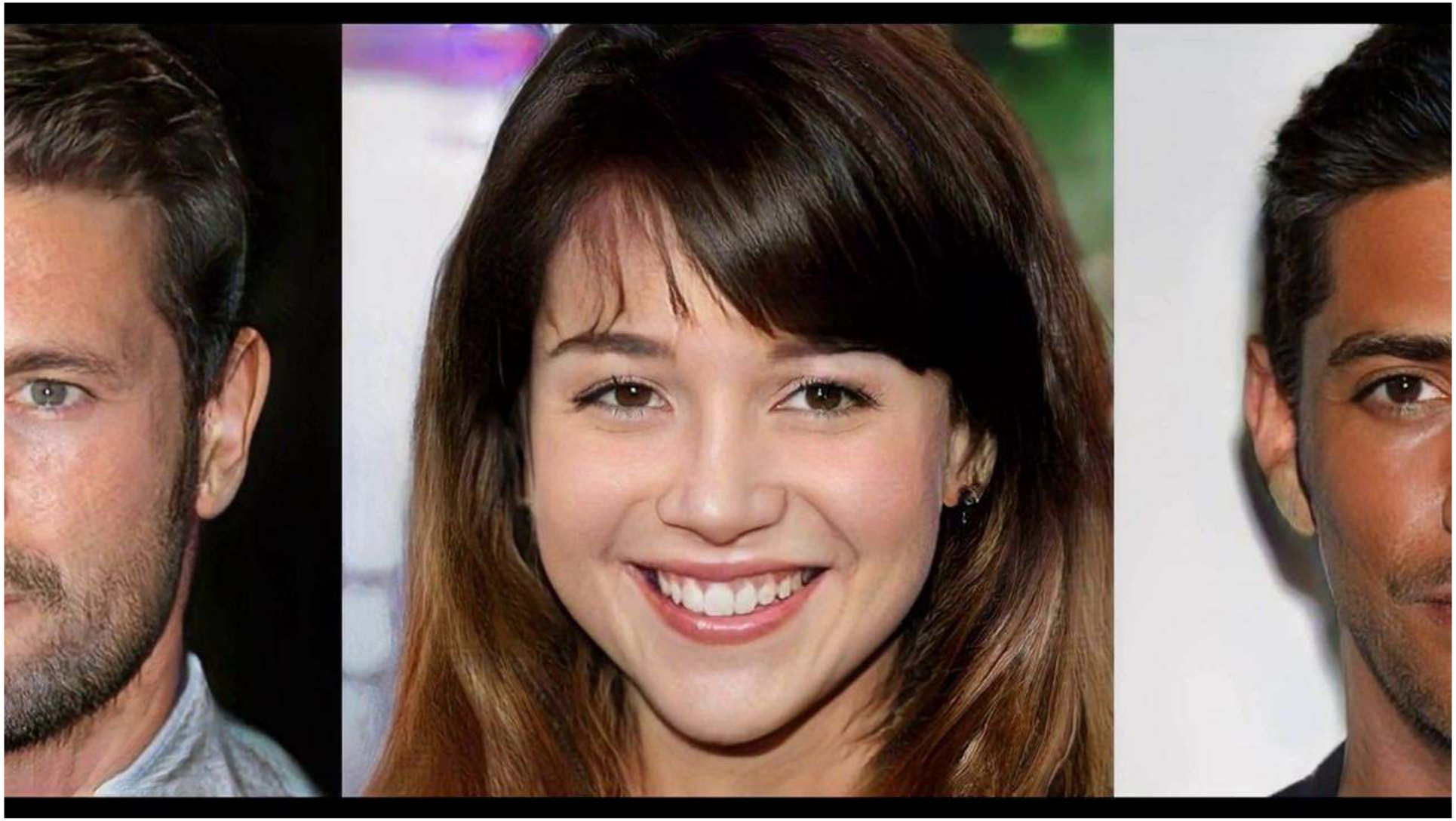
Image recognition





- Google Brain uses deep learning to recognize cats in images (2012)
- Facebook recognizes faces with 97% accuracy (2014)

Generative Adversarial Networks (2014)

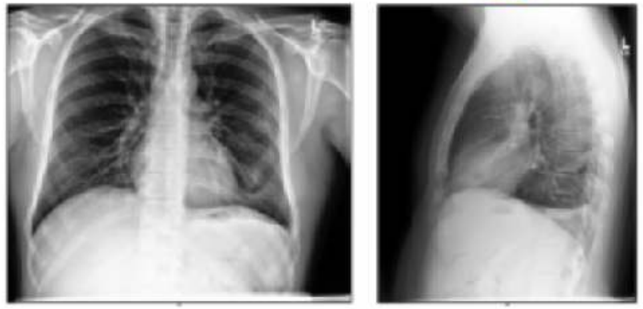


Progressive Growing of GANs for Improved Quality, Stability, and Variation

https://www.youtube.com/watch?v=G06dEcZ-QTg&t=36s&ab_channel=TeroKarrasFI

Image recognition

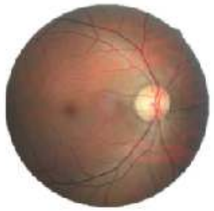
Chest radiology



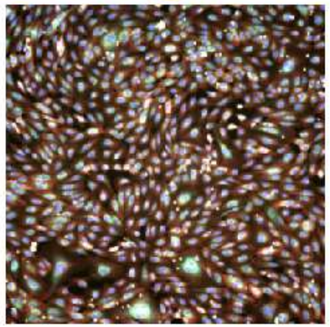
2019

Google AI outperforms radiologists at lung cancer detection

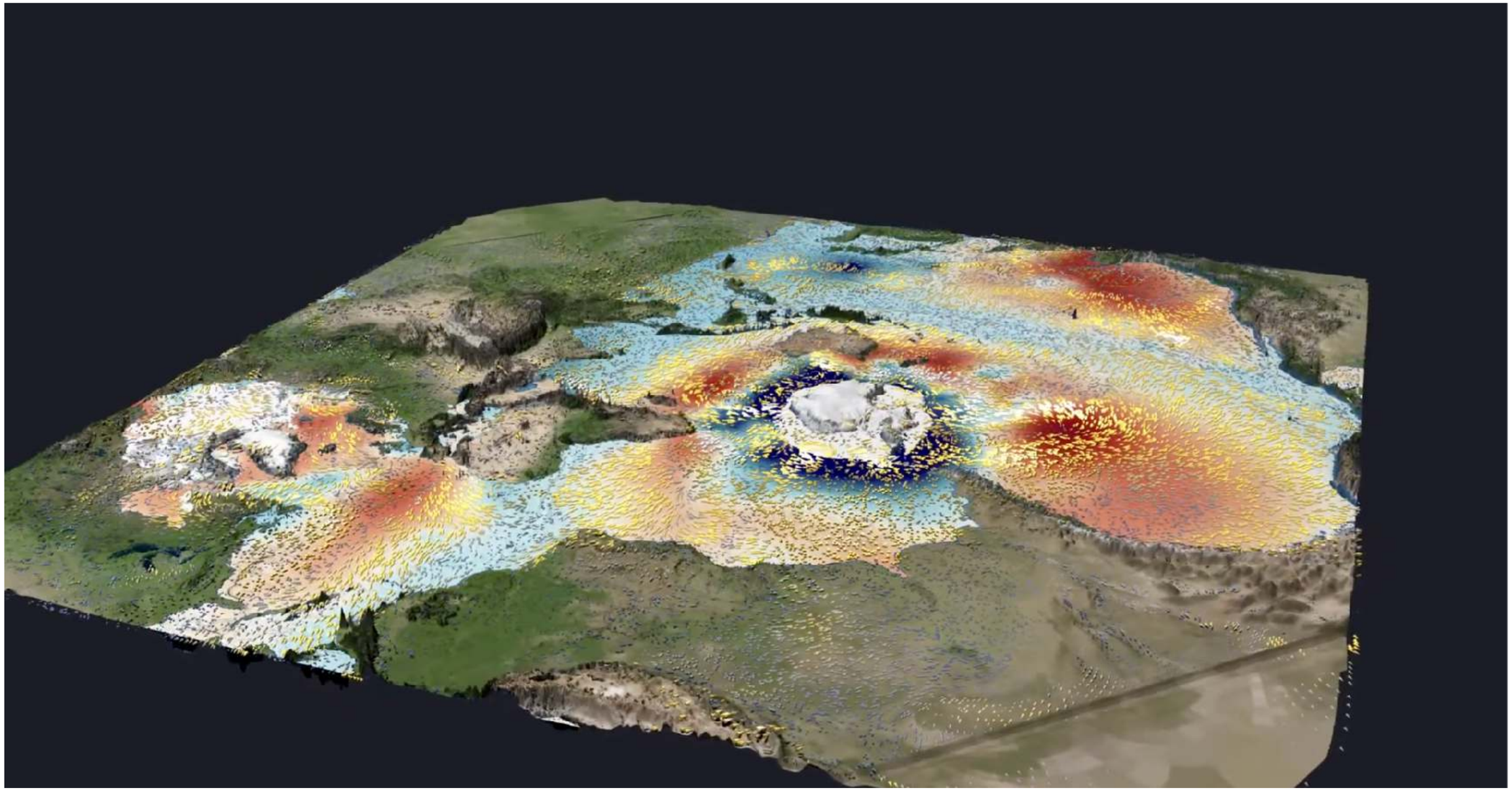
Diabetic retinopathy



Drug screening for COVID-19



Coupled Ocean Atmosphere Simulation



https://www.youtube.com/watch?v=4f0iVY2nd2M

Video

dimension: $T \times h \times w \times 3$

t_1

123	4	6	43
35	44	33	255
65	2	0	0
33	54	0	0

t_2

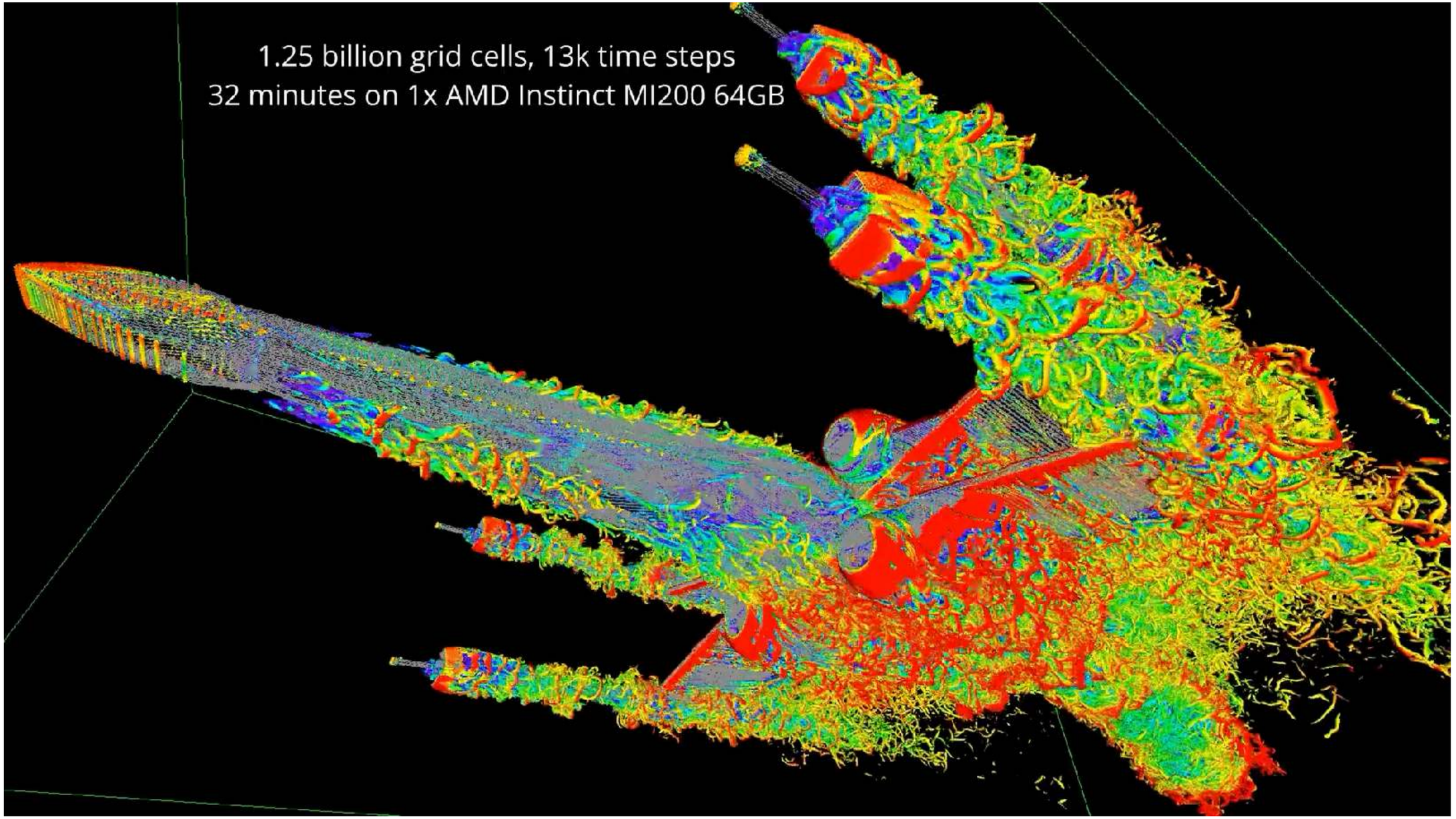
123	4	6	43
35	44	33	255
65	2	0	0
33	54	0	0

. . . . . .

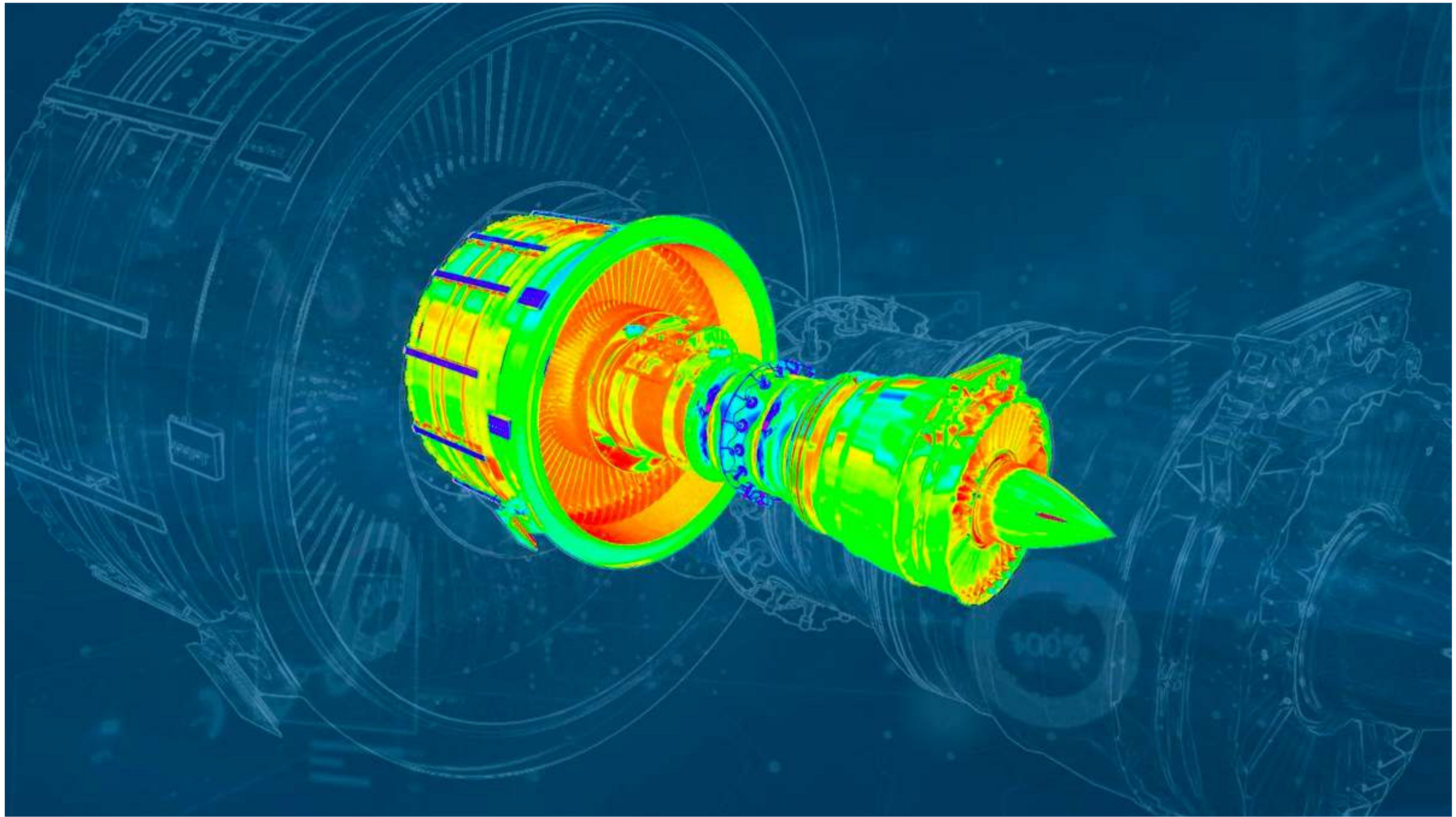
T

123	4	6	43
35	44	33	255
65	2	0	0
33	54	0	0

Fluid dynamics



Machine learning in materials research



Language

Deep learning surpasses traditional approaches in speech recognition in 2010

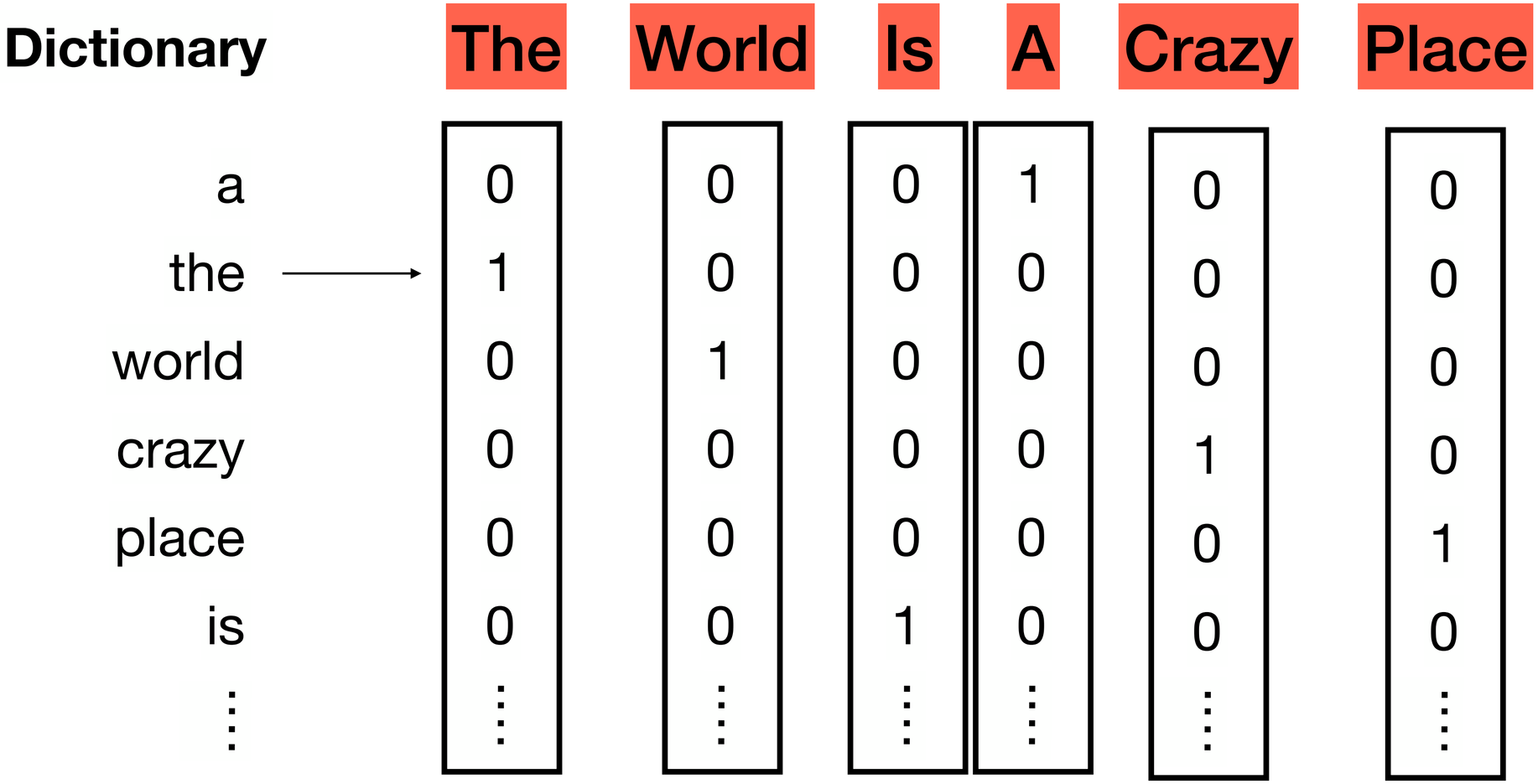
2016 2014 2011

IBM-Watson Defeats Humans in "Jeopardy!"

2011

Text

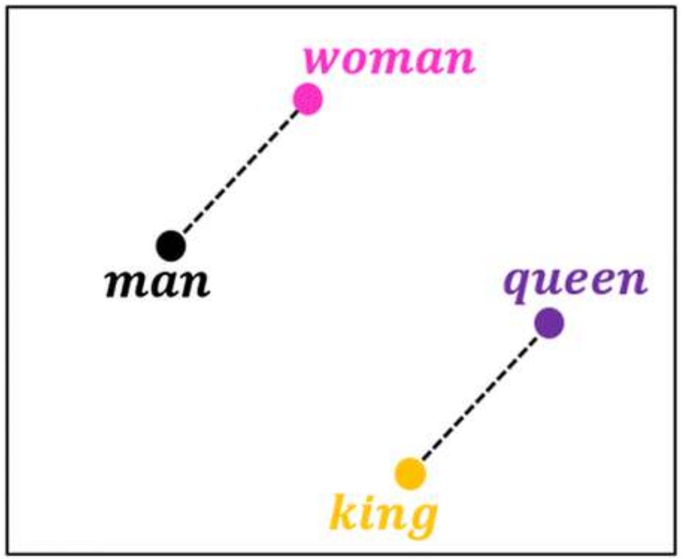
One-hot word representation



Text

Word Embedding Representation

man	→	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman	→	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king	→	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen	→	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9



Words

Embedding

Semantic Space

https://www.analyticsvidhya.com/blog/2021/07/feature-extraction-and-embeddings-in-nlp-a-beginners-guide-to-understand-natural-language-processing/

Text Generation

give me some project ideas on machine learning for music

Machine Learning Project Ideas for Music

Music Generation and Composition

- **Style Transfer:** Train a model to transfer the style of one musical piece to another. For instance, create a Bach piece in the style of a Beatles song.
- **Generative Adversarial Networks (GANs):** Generate new, original music pieces using GANs. You can train the model on specific genres or composers to create music in those styles.
- **Melody Generation:** Develop a model to generate catchy and melodic sequences, which can be used for jingles, background music, or as a starting point for composing full pieces.
- **Chord Progression:** Create a model to predict chord progressions based on given musical contexts, aiding in songwriting and composition.

Language Models that do Science

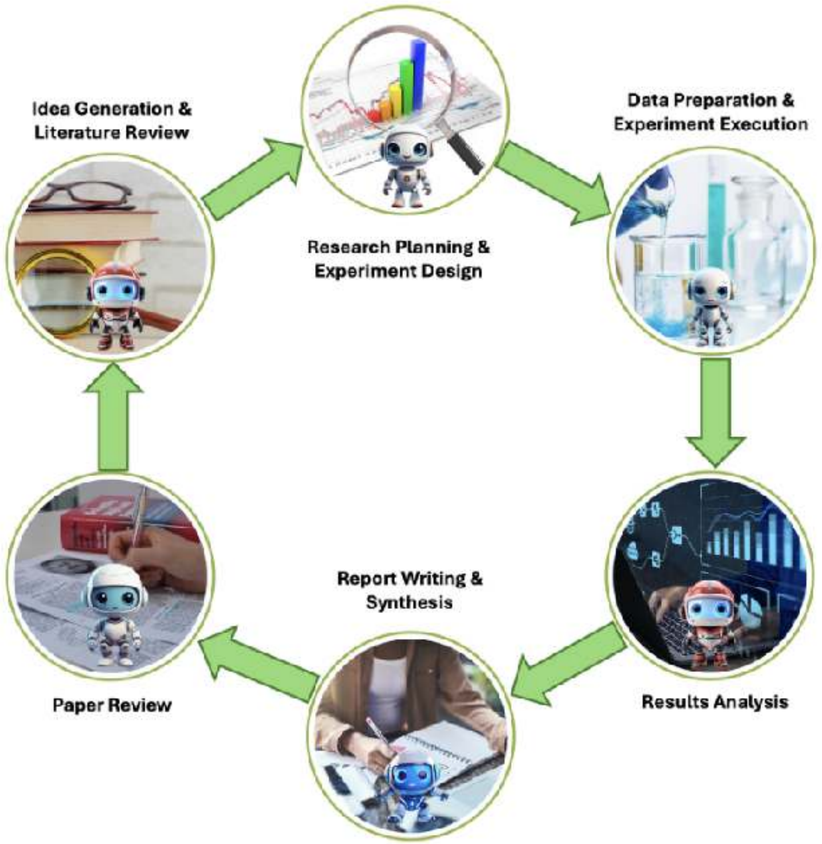


Figure 1: Agentic AI workflow for scientific discovery.

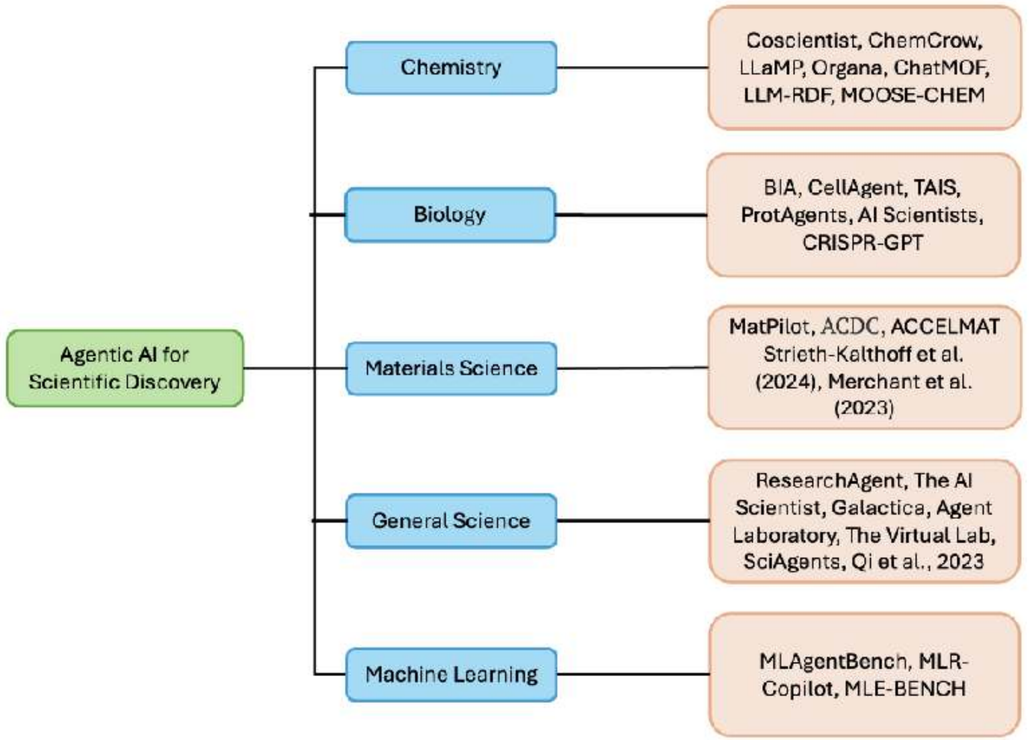


Figure 2: AI Agents frameworks for scientific discovery.

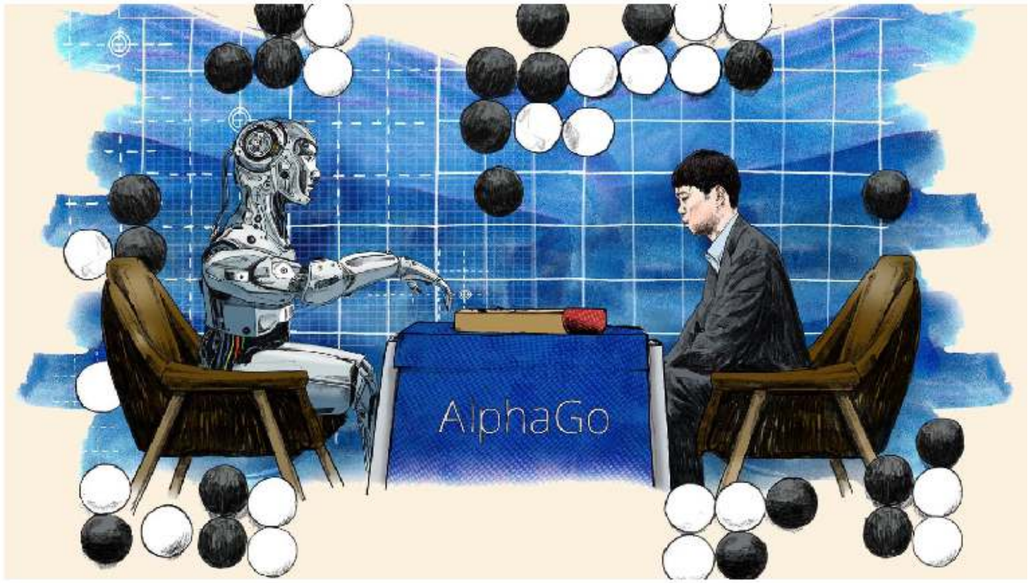
https://arxiv.org/abs/2503.08979

Games

IBM creates a checker playing program

1959

AlphaGo beats world's best Go player

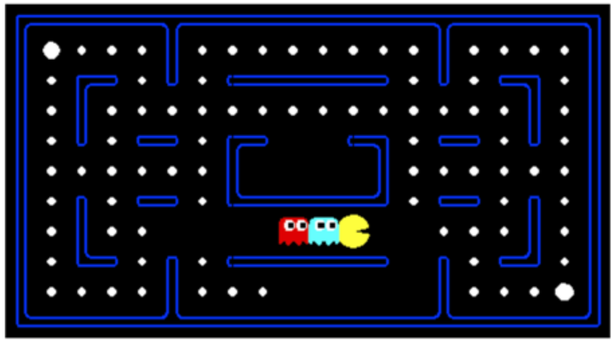


2016

IBM Deep Blue beats Kasparov

1997

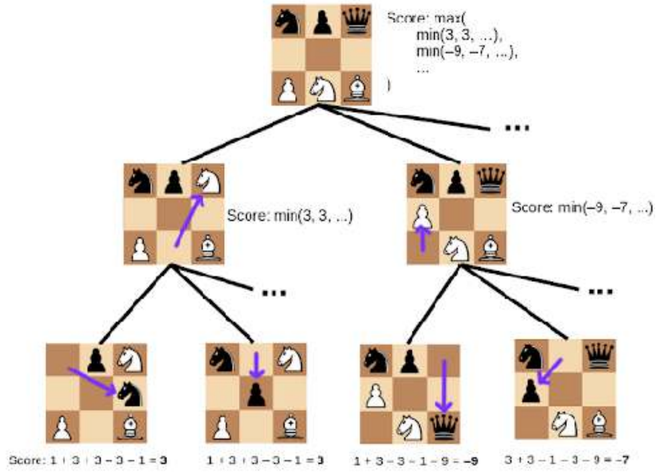
Solving puzzles



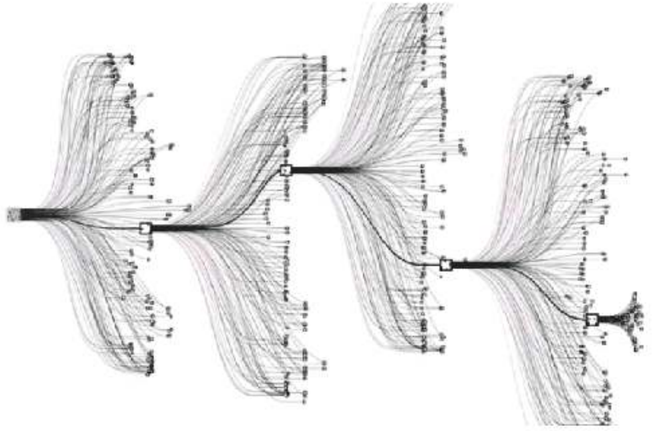
5	3			7				
6			1	9	5			
	9	8					6	
8				6				3
4			8		3			1
7				2				6
	6					2	8	
			4	1	9			5
				8			7	9

→

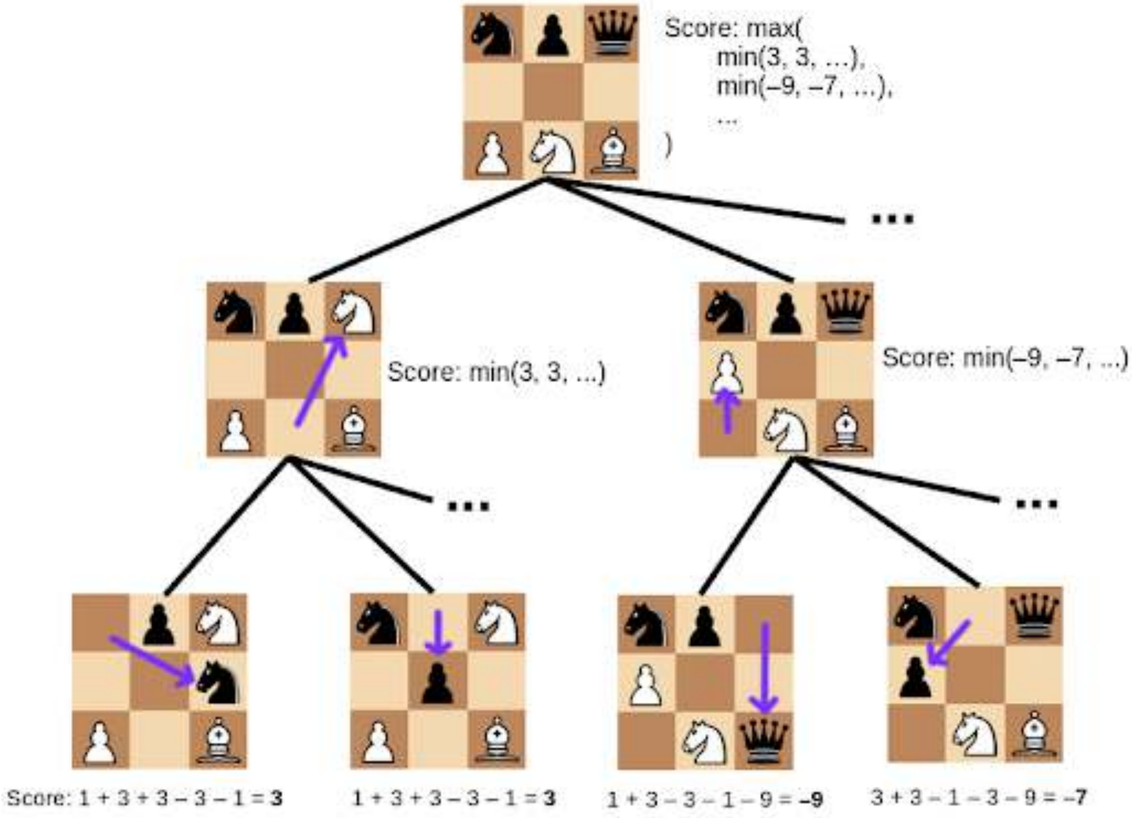
5	3	4	6	7	8	9	1	2
6	7	2	1	9	5	3	4	8
1	9	8	3	4	2	5	6	7
8	5	9	7	6	1	4	2	3
4	2	6	8	5	3	7	9	1
7	1	3	9	2	4	8	5	6
9	6	1	5	3	7	2	8	4
2	8	7	4	1	9	6	3	5
3	4	5	2	8	6	1	7	9



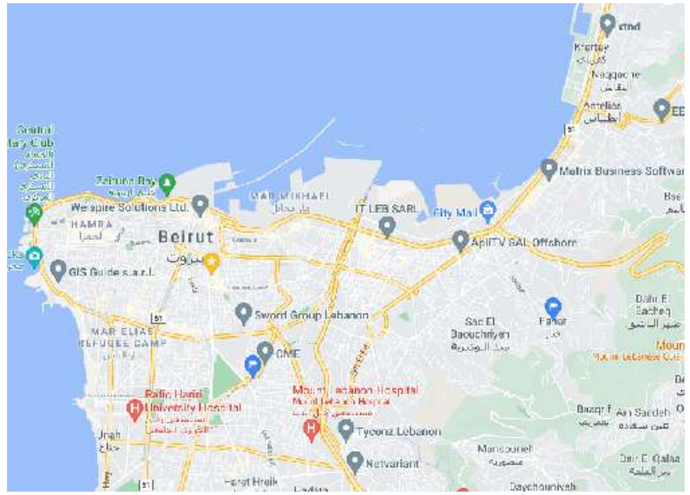
10^{120} possibilities

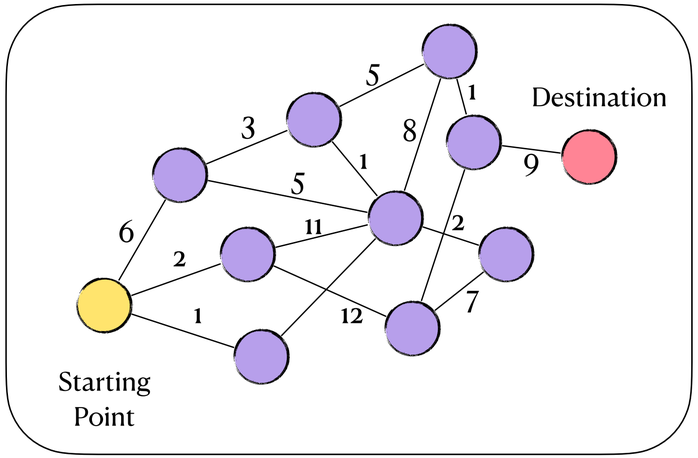


Graph Representation



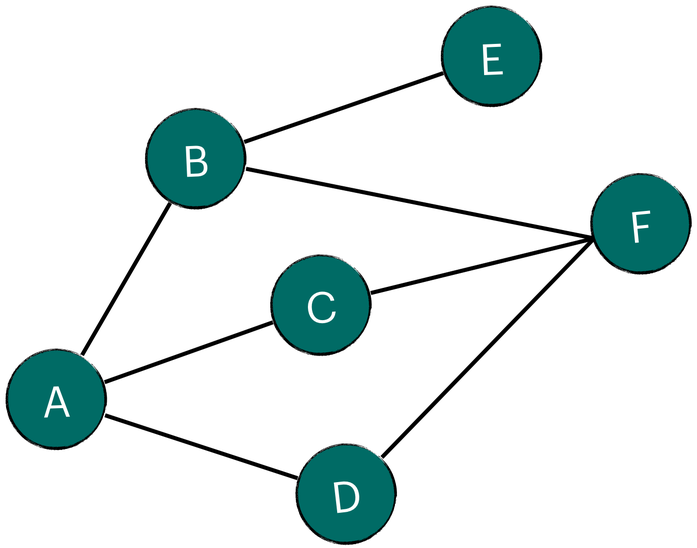
Real world





Graph Representation

Graph with Vertices and Edges



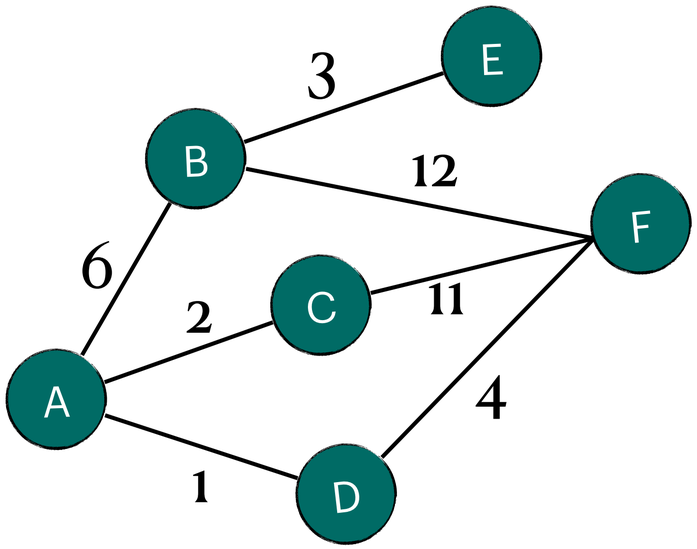
Adjacency Matrix

	A	B	C	D	E	F
A	0	1	1	1	0	0
B	1	0	0	0	1	1
C	1	0	0	0	0	1
D	1	0	0	0	0	1
E	0	1	0	0	0	0
F	0	1	1	1	0	0

https://www.cs.mtsu.edu/~xyang/3080/adjacencyMatrix.html

Graph Representation

Weighted Graph



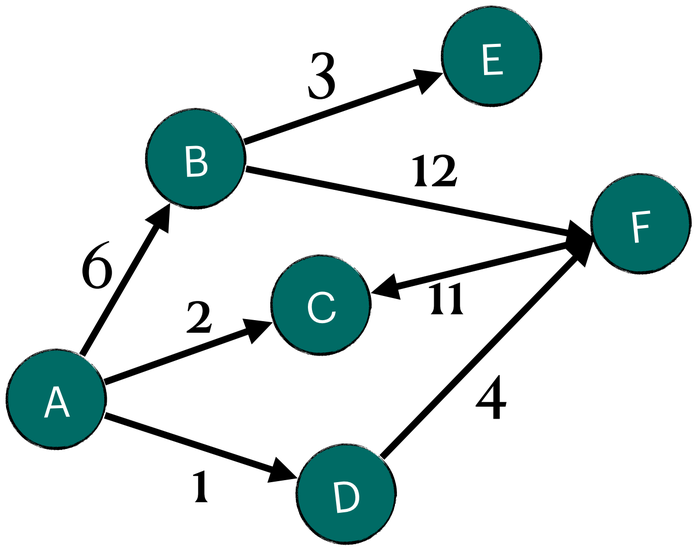
Adjacency Matrix

	A	B	C	D	E	F
A	0	6	2	1	0	0
B	6	0	0	0	3	12
C	2	0	0	0	0	11
D	1	0	0	0	0	4
E	0	3	0	0	0	0
F	0	12	11	4	0	0

https://www.cs.mtsu.edu/~xyang/3080/adjacencyMatrix.html

Graph Representation

Weighted Directed Graph

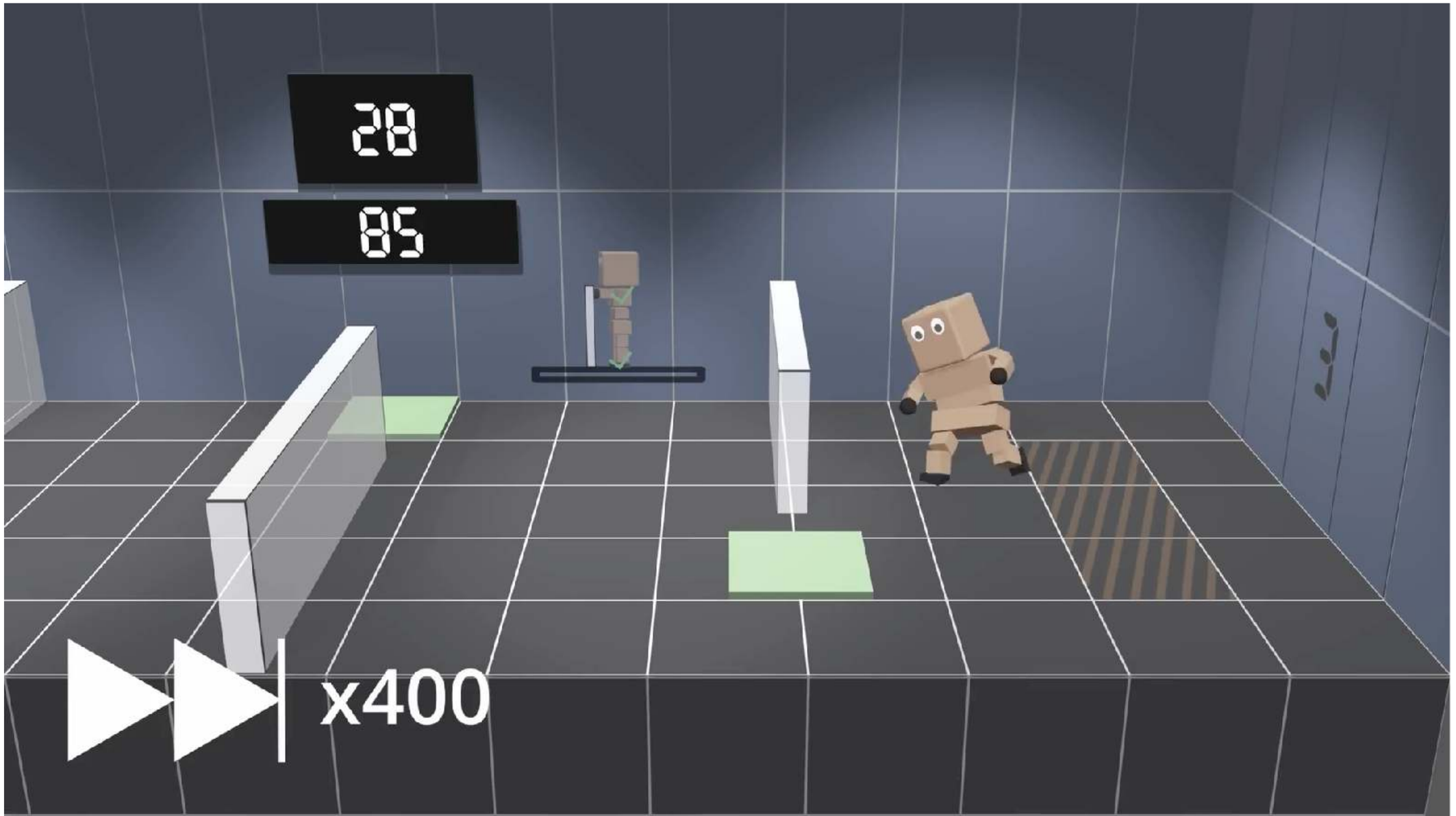


Adjacency Matrix

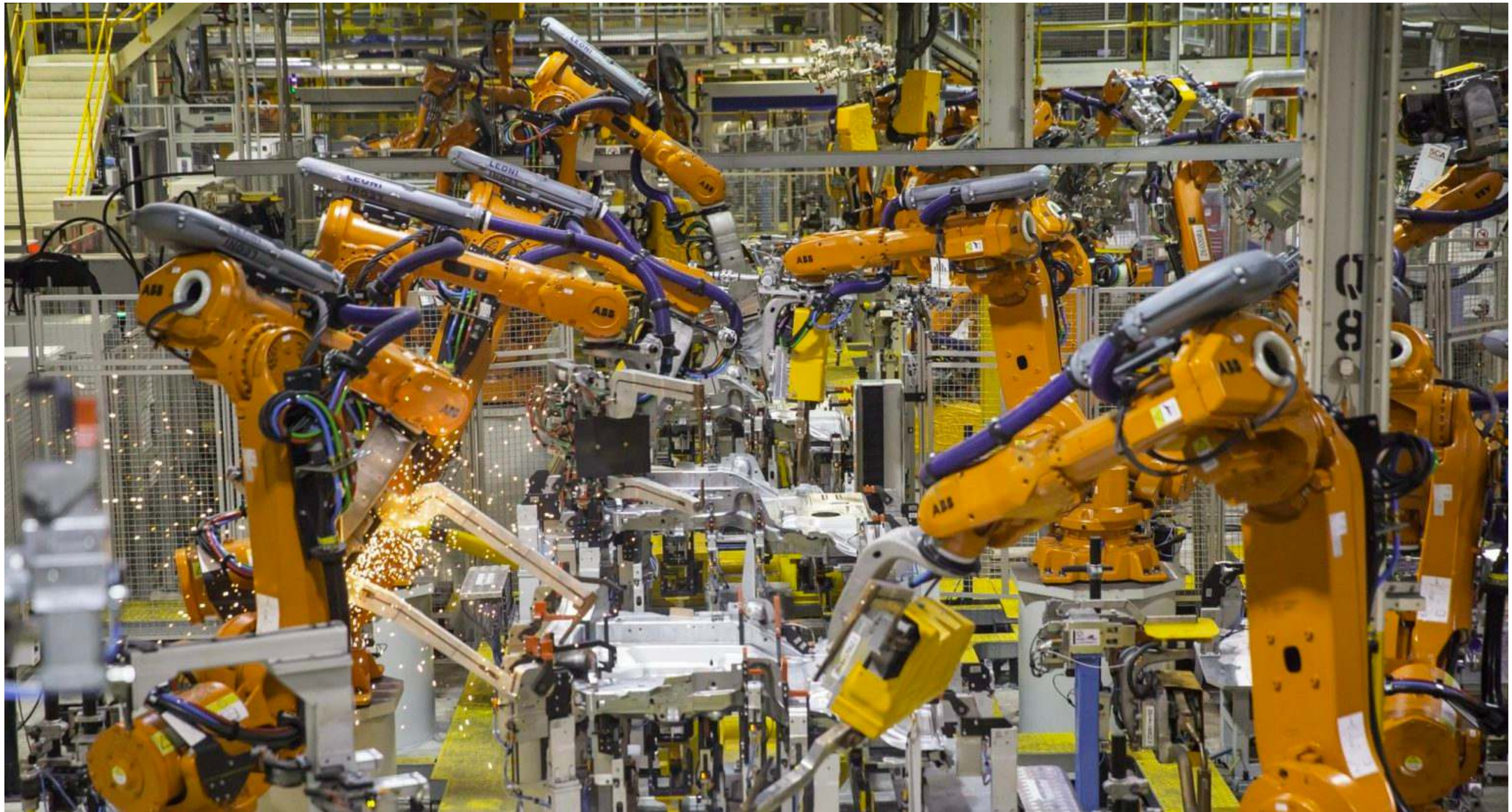
	A	B	C	D	E	F
A	0	6	2	1	0	0
B	0	0	0	0	3	12
C	0	0	0	0	0	0
D	0	0	0	0	0	4
E	0	0	0	0	0	0
F	0	0	11	0	0	0

https://www.cs.mtsu.edu/~xyang/3080/adjacencyMatrix.html

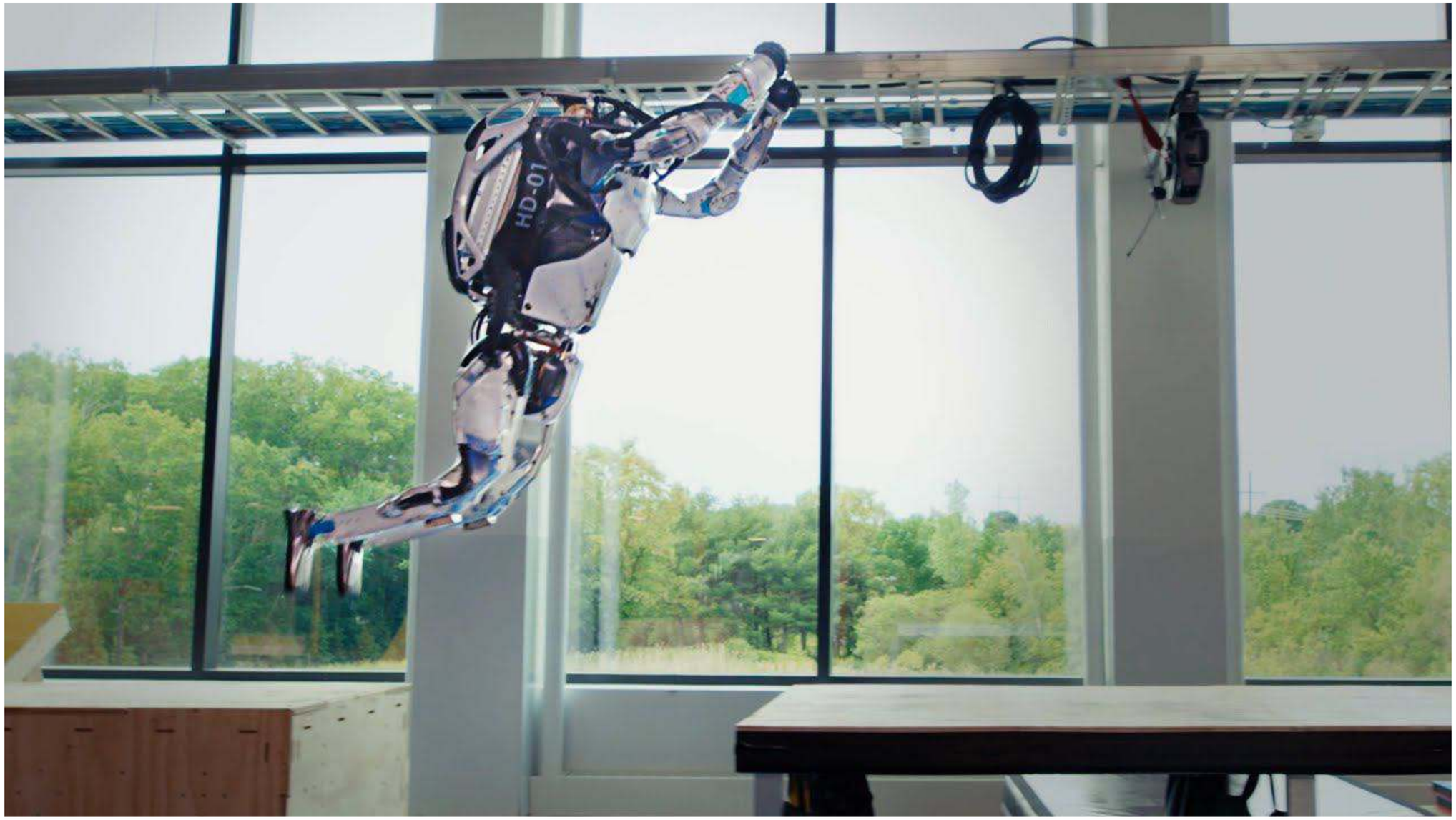
Robotics and games: Reinforcement Learning



Robotics and Manufacturing



Robotics



Boston Dynamics - https://youtube.com/watch?v=tF4DML7FIWk

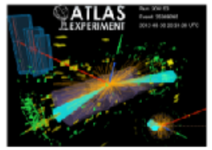
Classification

classifier

$$x \longrightarrow \boxed{f} \longrightarrow y \in \{+1, -1\}$$ label

Fraud detection: credit card transaction $\longrightarrow$ fraud or no fraud

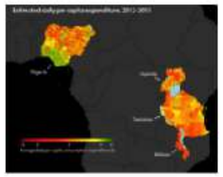
Toxic comments: online comment $\longrightarrow$ toxic or not toxic



Higgs boson: measurements of event $\longrightarrow$ decay event or background

Regression

$$x \longrightarrow \boxed{f} \longrightarrow y \in \mathbb{R} \text{ response}$$



Poverty mapping: satellite image $\longrightarrow$ asset wealth index

Housing: information about house $\longrightarrow$ price

Arrival times: destination, weather, time $\longrightarrow$ time of arrival

Structured Prediction

$x \longrightarrow \boxed{f} \longrightarrow y$ is a complex object

Machine translation: English sentence → Japanese sentence

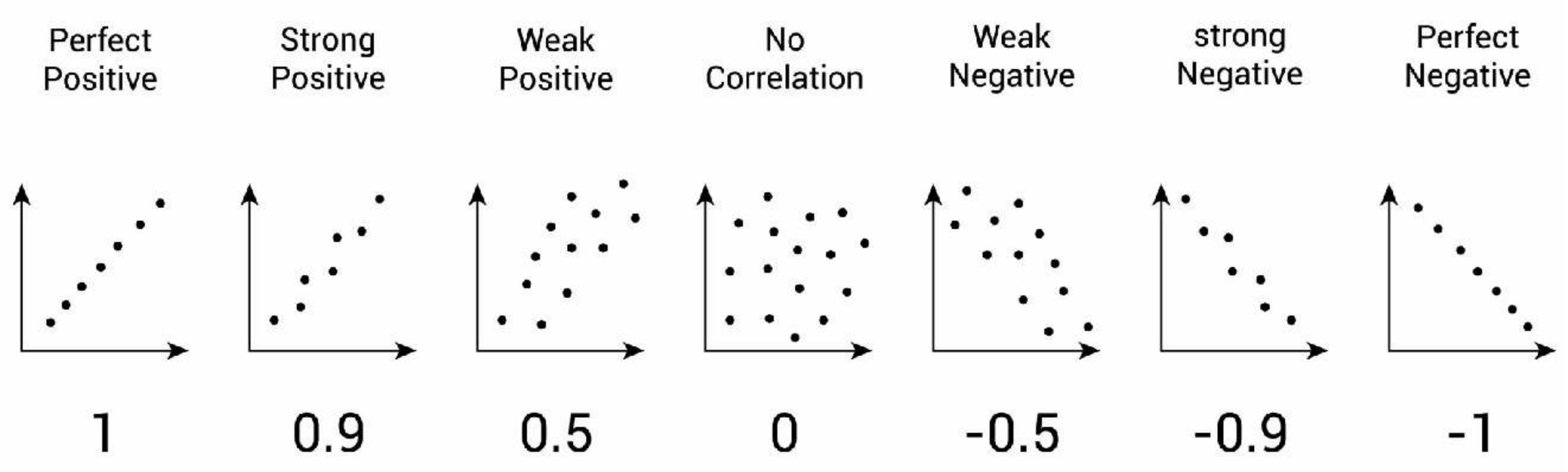
Dialogue: conversational history → next utterance

Image captioning: image → sentence describing image

Image segmentation: image → segmentation

Data Analysis

Correlation coefficient (2 variables)



$$r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata8.html

Correlation Matrix (multiple variables)

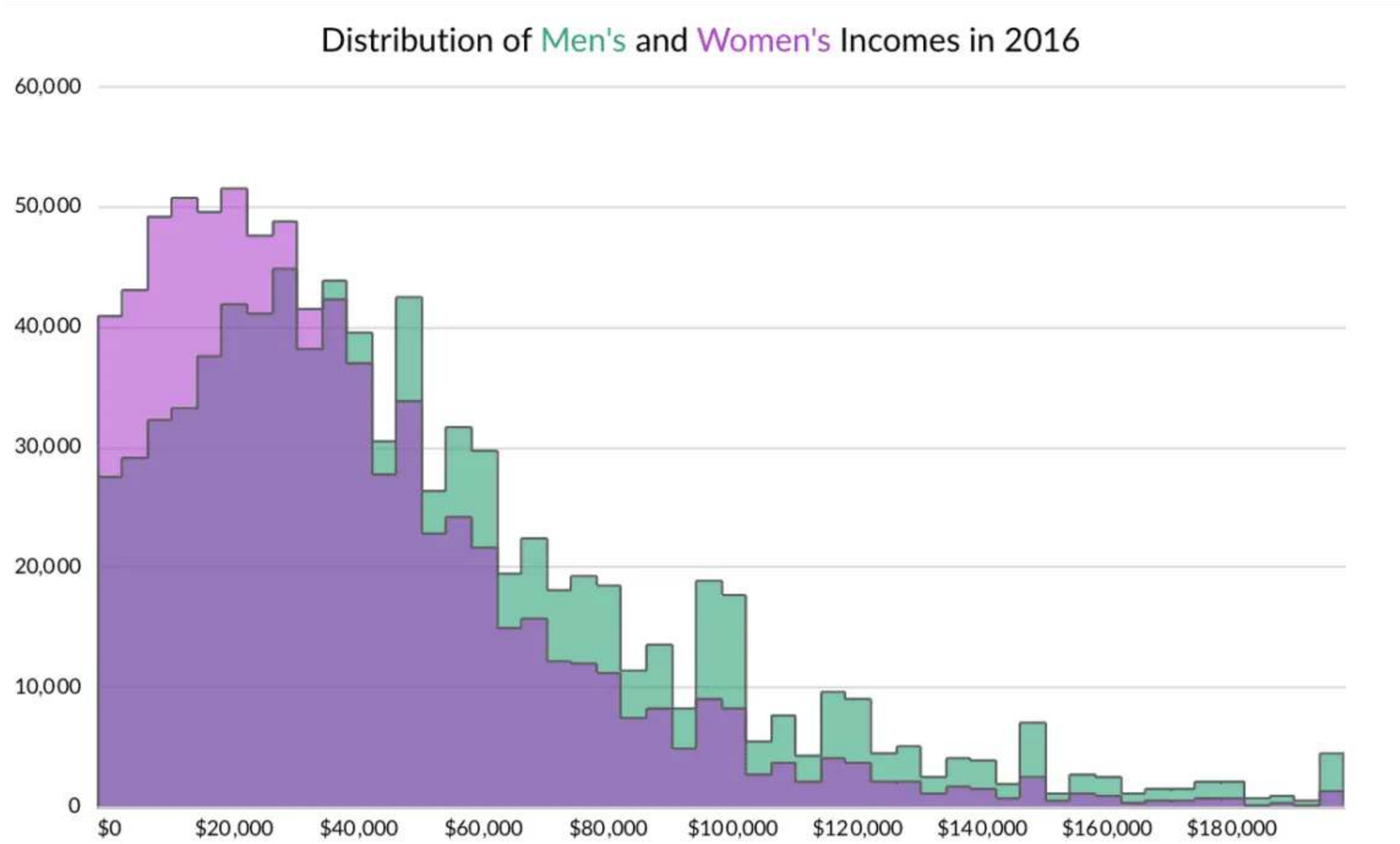
	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

$$r_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_{X_i}\sigma_{X_j}}$$

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

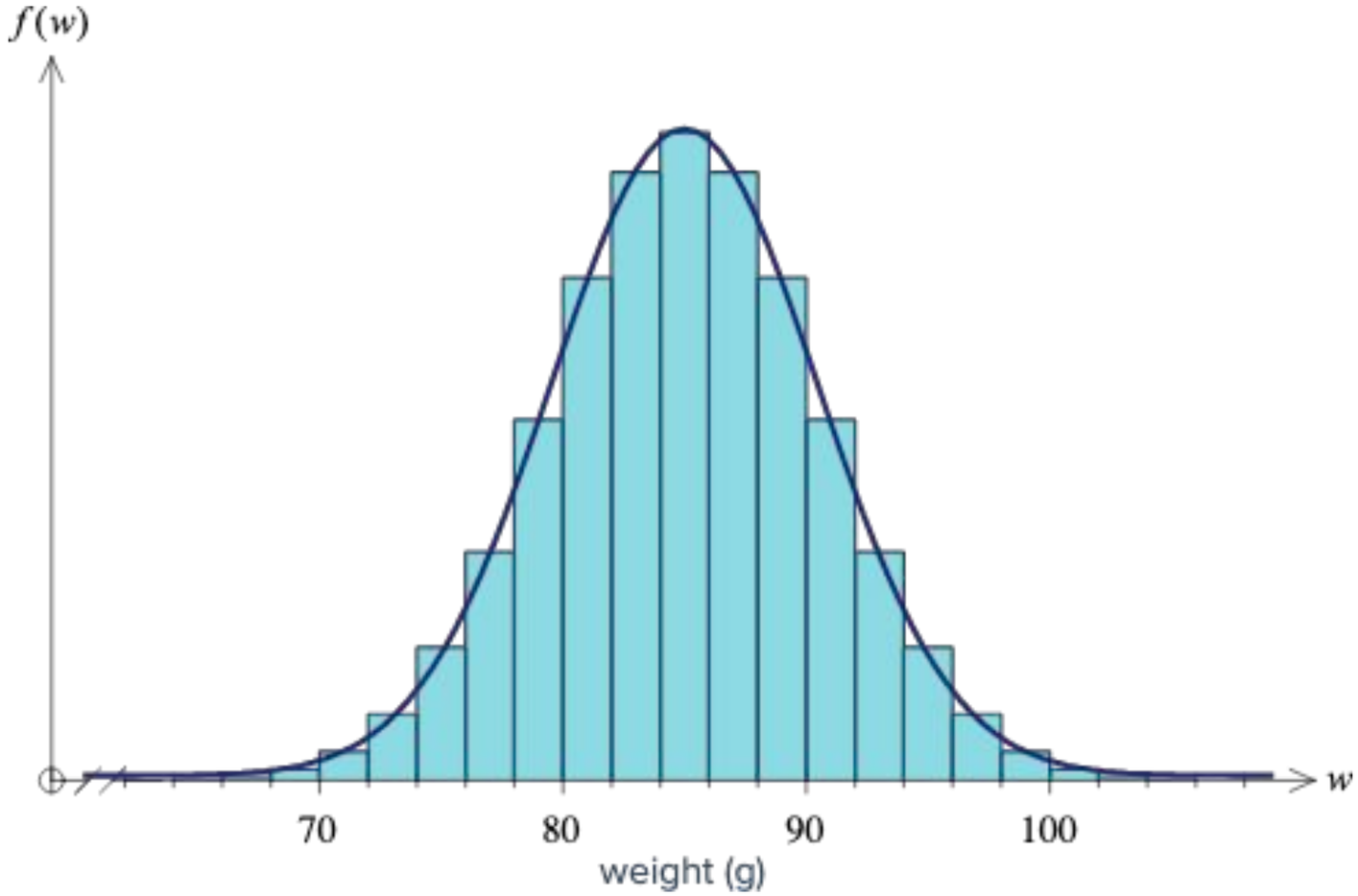
https://www.statology.org/how-to-read-a-correlation-matrix/

Histogram



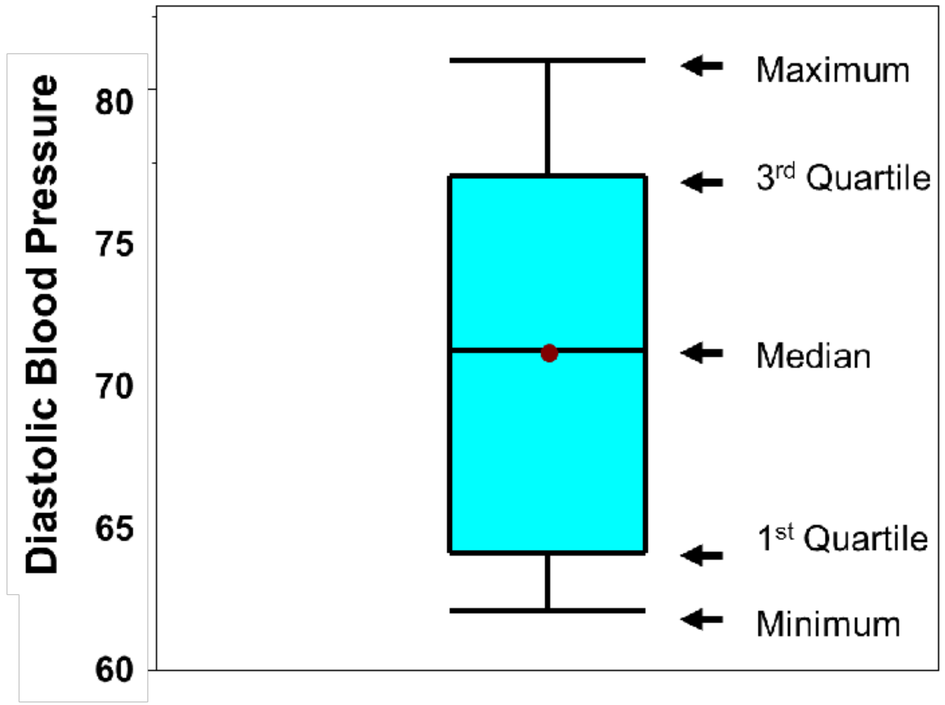
https://policyviz.com/2018/11/27/histogram-design-decisions/

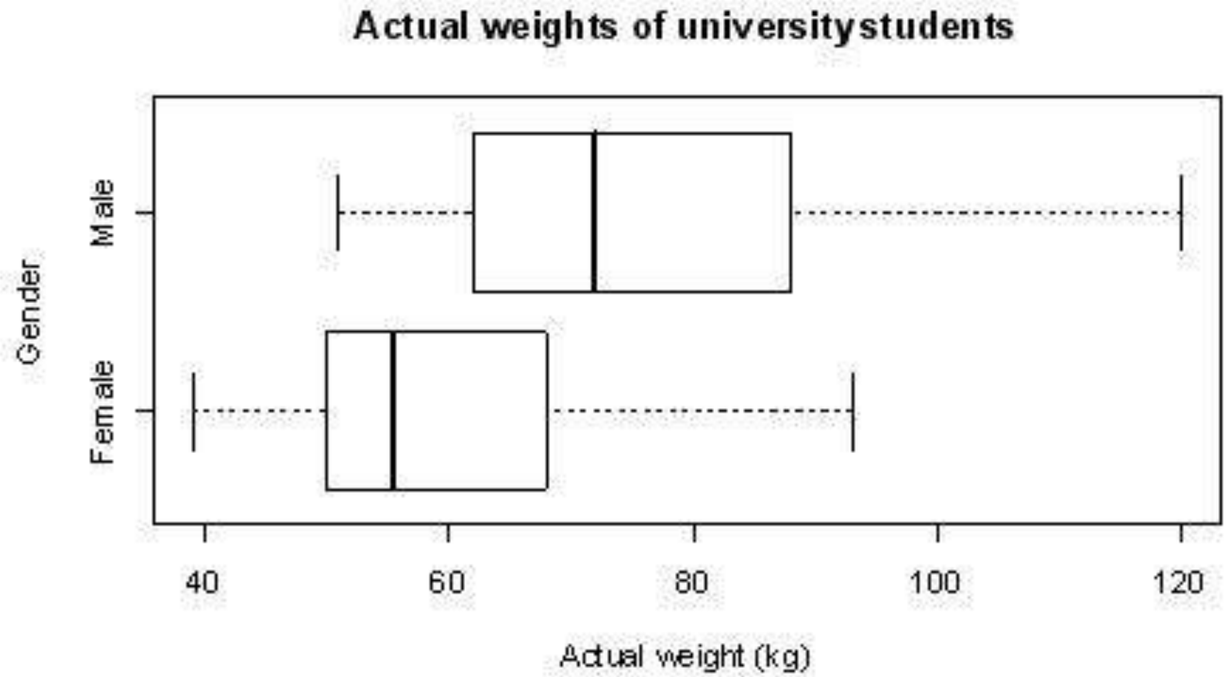
Probability Density Function



https://mathspace.co/textbooks/syllabuses/Syllabus-826/topics/Topic-18358/subtopics/Subtopic-279214/?activeTab=theory

Box and Whisker Plot



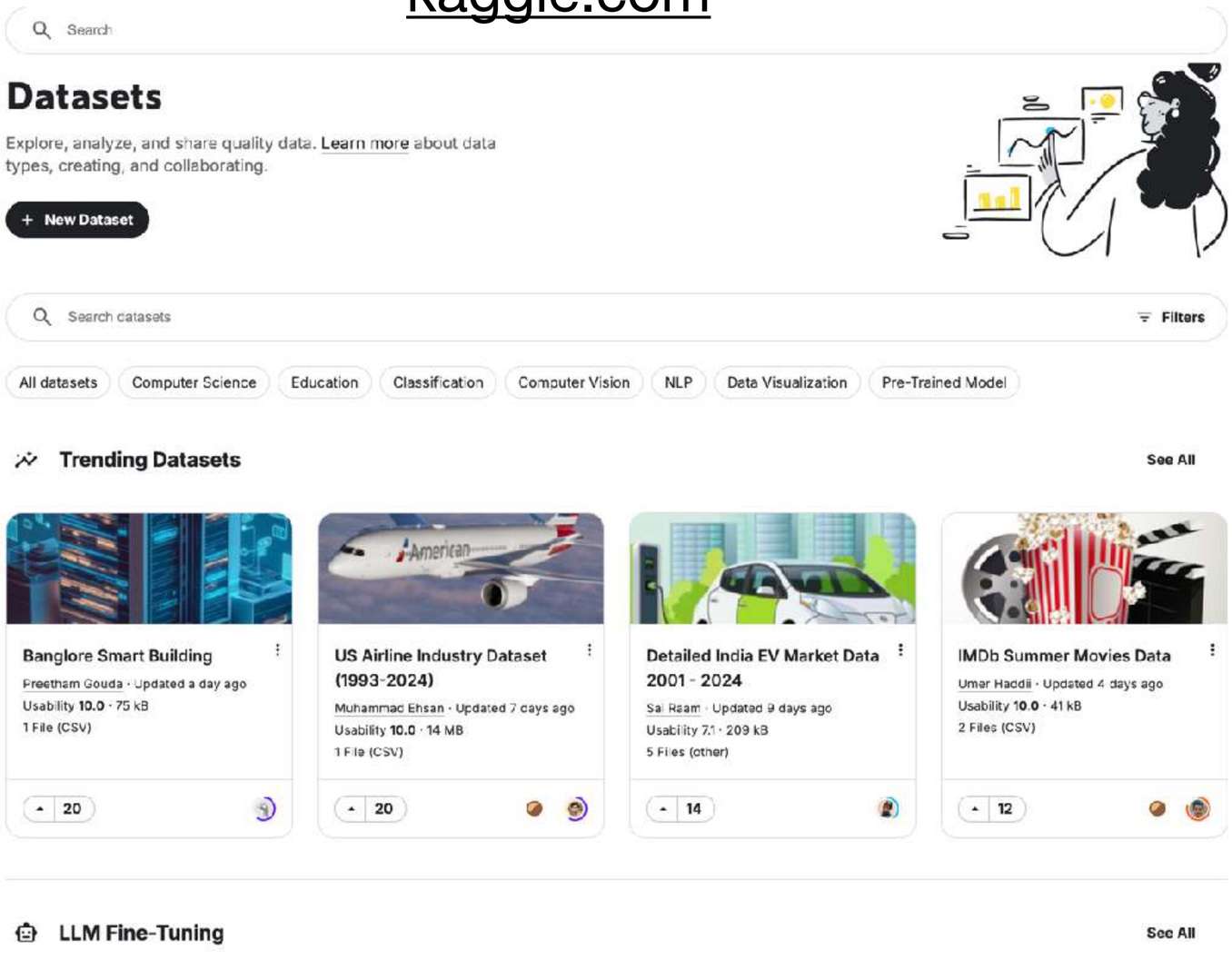


https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata8.html

Looking for data

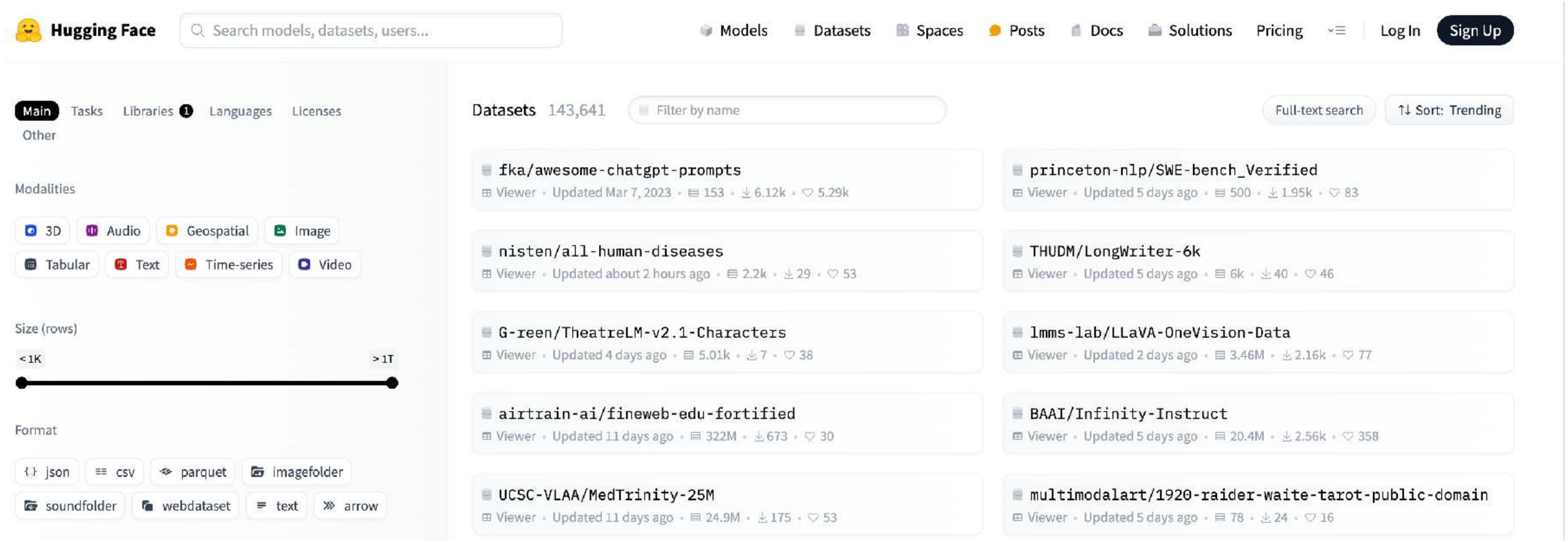
Where to look for data?

kaggle.com



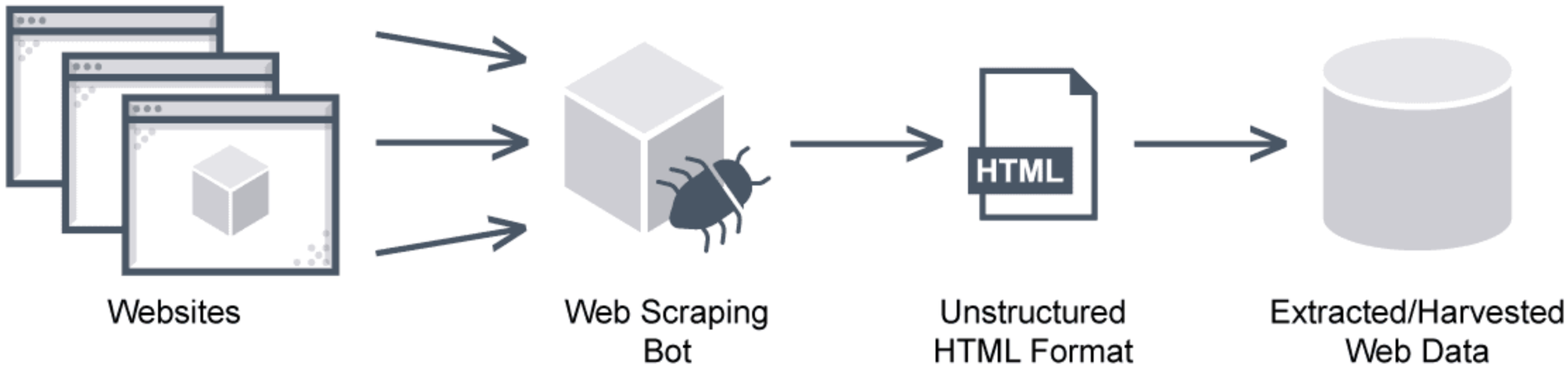
Where to look for data?

huggingface.co



Where to look for data?

Scrape the internet



https://avinetworks.com/glossary/web-scraping/

Where to look for data?

Ask GPT

I'm working on a machine learning project. Where do I look for data?

When you're starting a machine learning project, finding the right data is crucial. Here are some good places to look for data:

1. **Public Data Repositories:**
   - **Kaggle:** Offers a wide range of datasets across different domains, along with competitions that can help you refine your skills.
   - **UCI Machine Learning Repository:** A classic resource with numerous datasets often used in academic research.
   - **Google Dataset Search:** A search engine for datasets that pulls from various sources on the web.
   - **Data.gov:** For datasets related to U.S. government, health, and other sectors.
   - **AWS Public Datasets:** Datasets hosted on Amazon Web Services, spanning multiple fields.
2. **Specific Domain Databases:**
   - **ImageNet:** For image data, often used for training deep learning models.
   - **OpenStreetMap:** If you're working on geographic data or maps.
   - **National Centers for Environmental Information (NCEI):** If you're looking for climate and environmental data.
   - **Spotify API:** For music-related data.
3. **APIs:**
   - **Twitter API:** Great for text and sentiment analysis projects.
   - **Reddit API:** Another source for text data, particularly discussions and community interactions.
   - **OpenWeatherMap API:** For weather-related data.
4. **Web Scraping:**
   - If you have a specific niche in mind, web scraping could be your go-to. Tools like BeautifulSoup or Scrapy in Python can help you collect data from websites (just be sure to check the site's terms of service).
5. **Academic Journals and Research Papers:**
   - Sometimes datasets are shared as part of academic research. Sites like **arXiv** or **IEEE Xplore** might provide links to datasets used in published research.

How do you come up with project ideas?

google.com

gemini.google.com