

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**

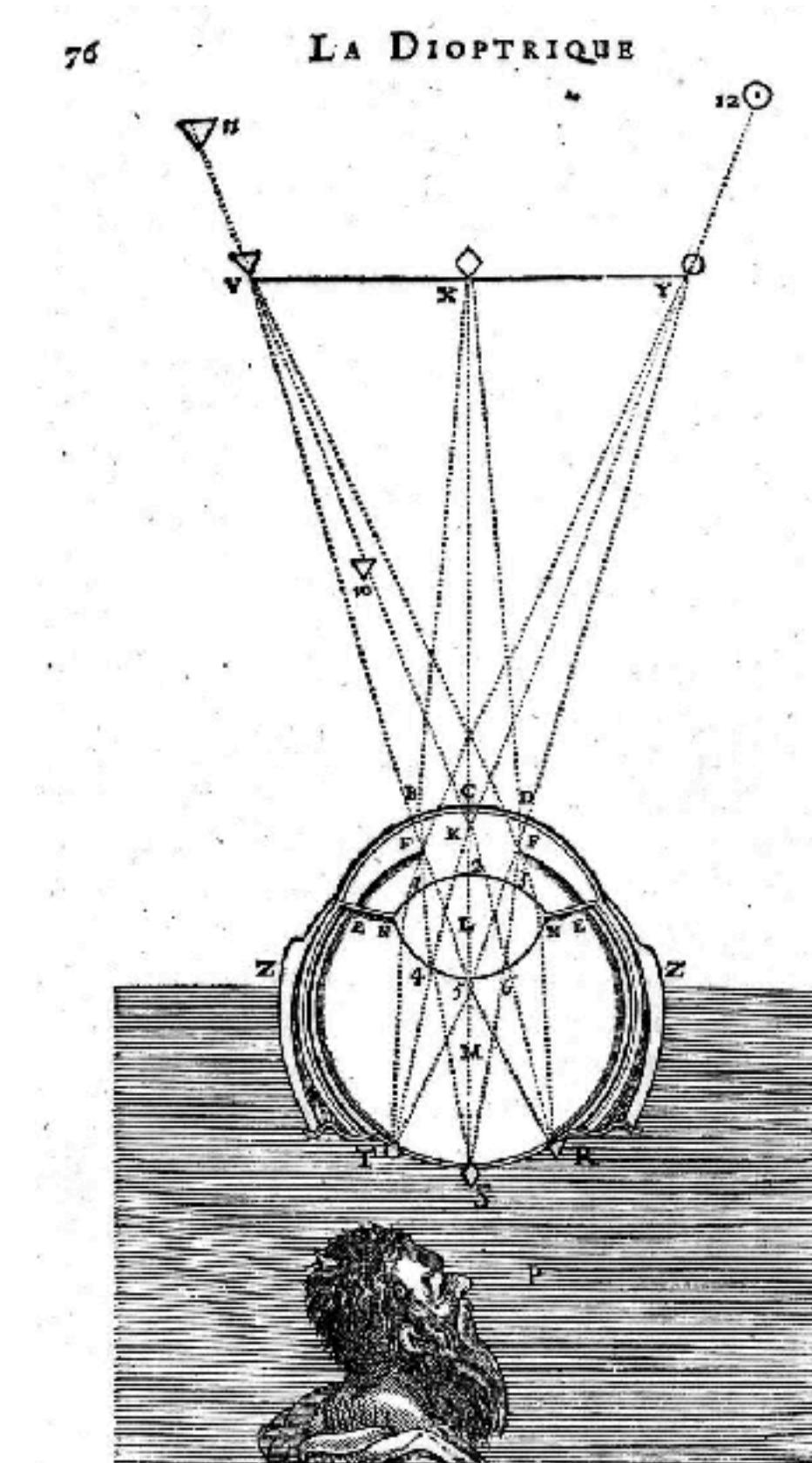
Finding constants of nature that generalize across space and time



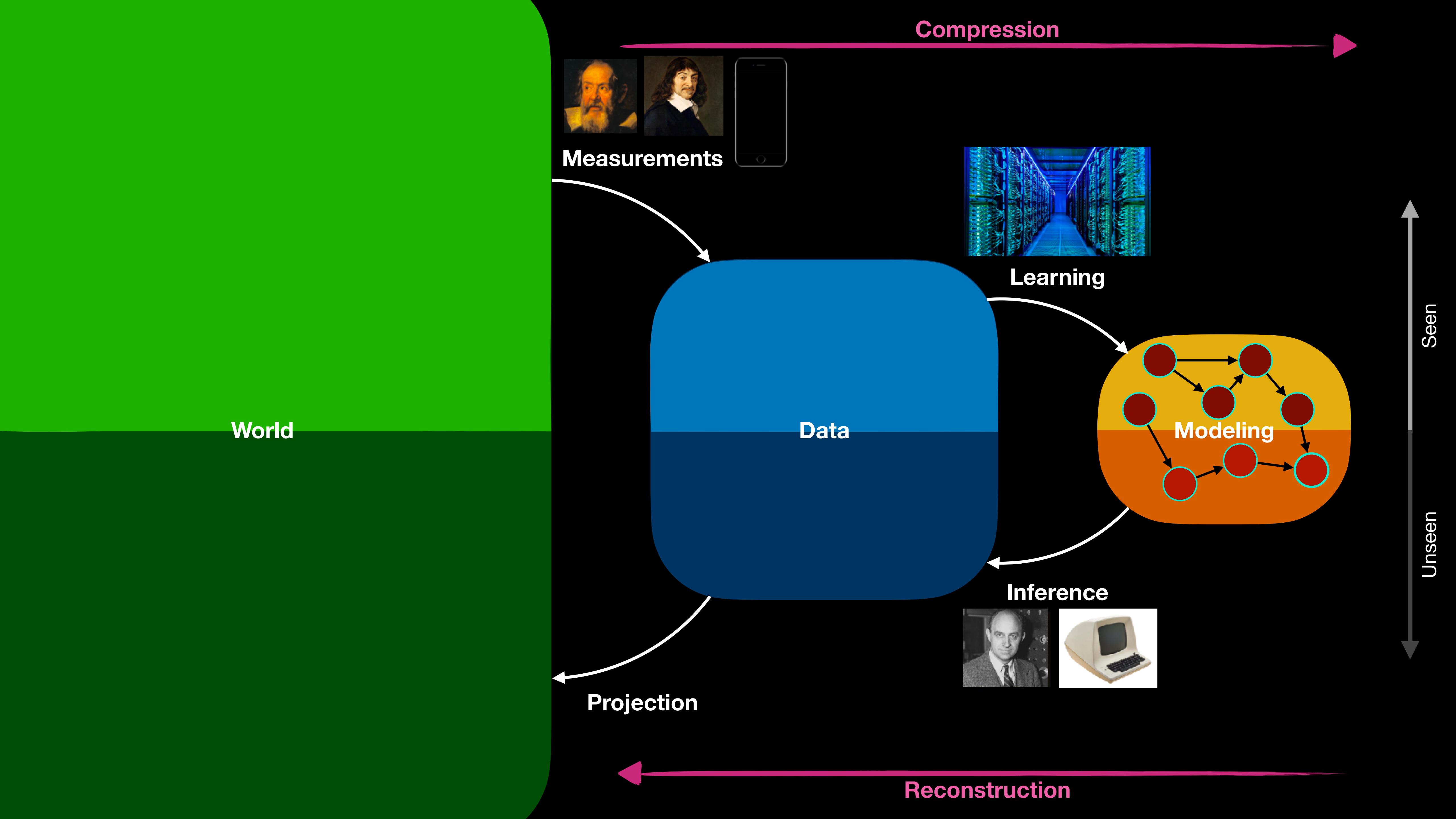
Galileo



Kepler

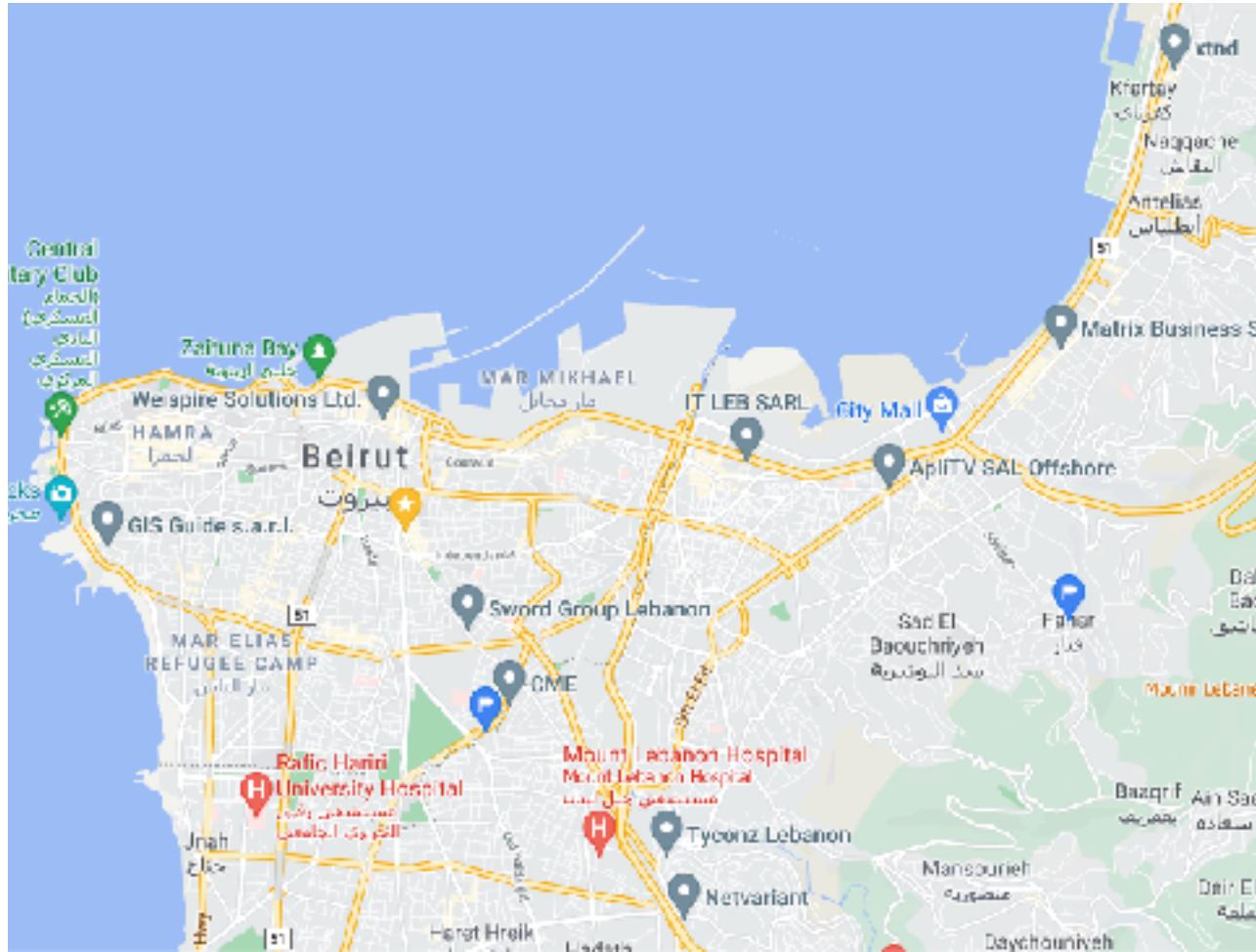


Descartes



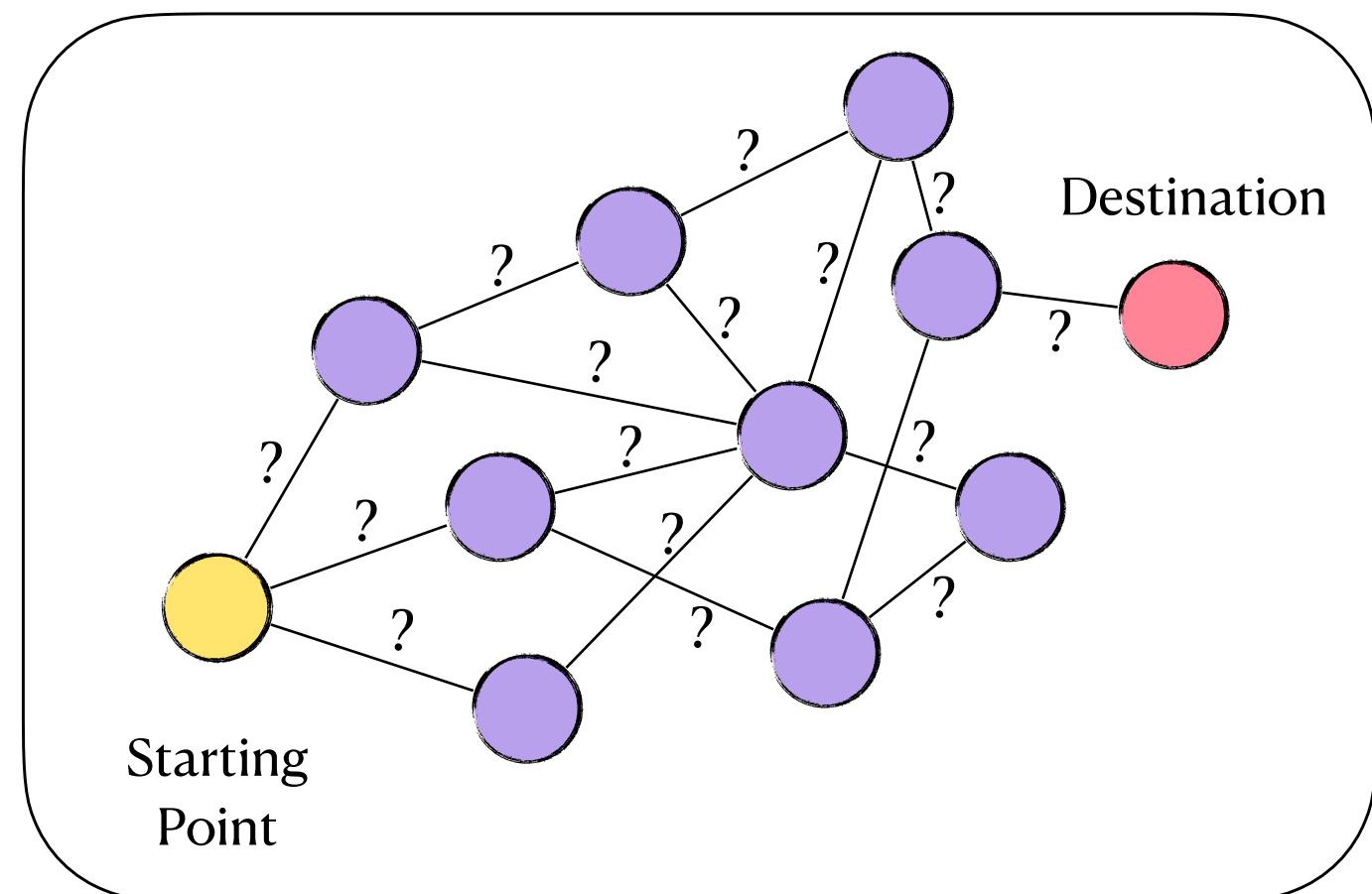
The 3 pillars of Artificial Intelligence

Real world

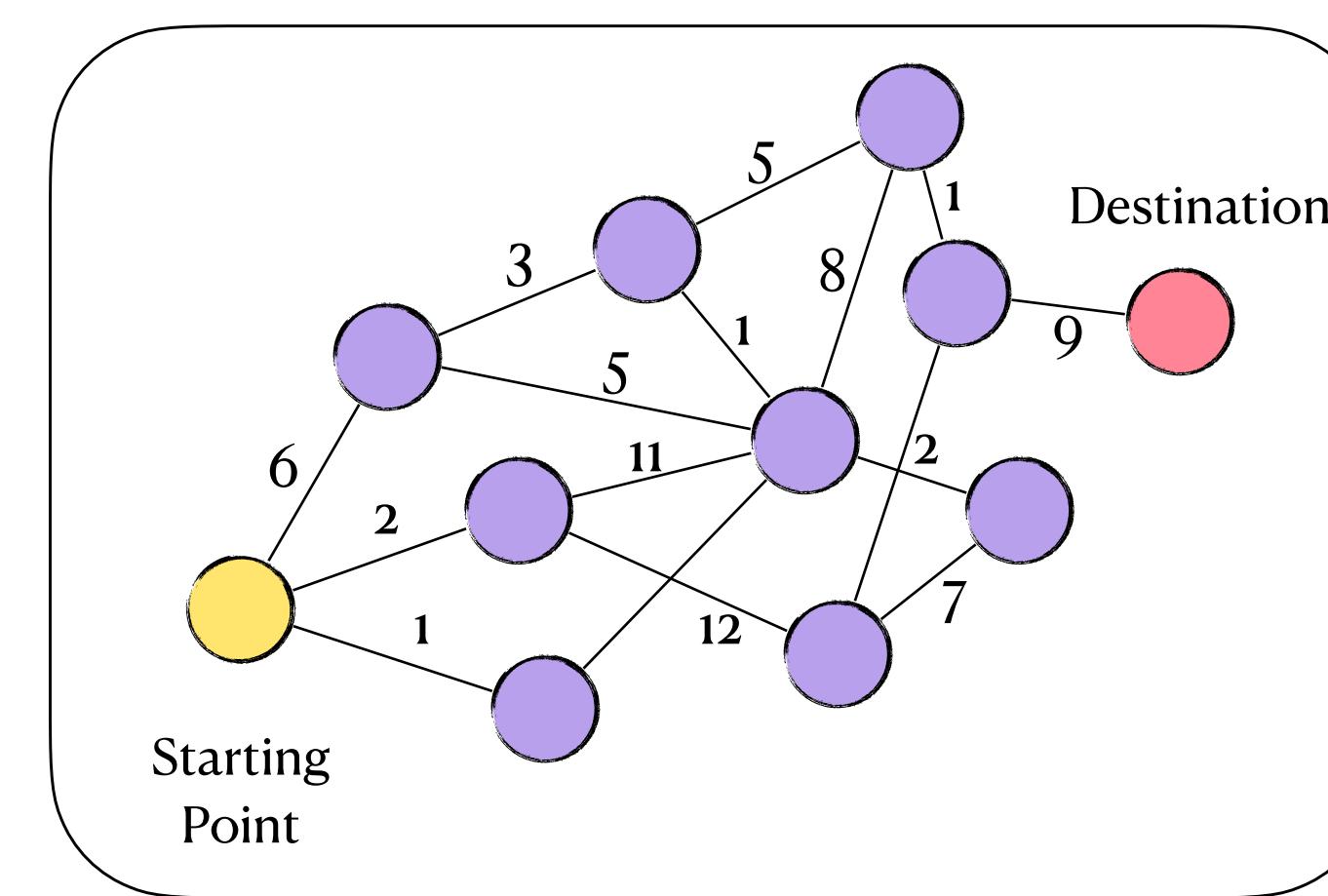


What's the shortest path
from AUB to my home?

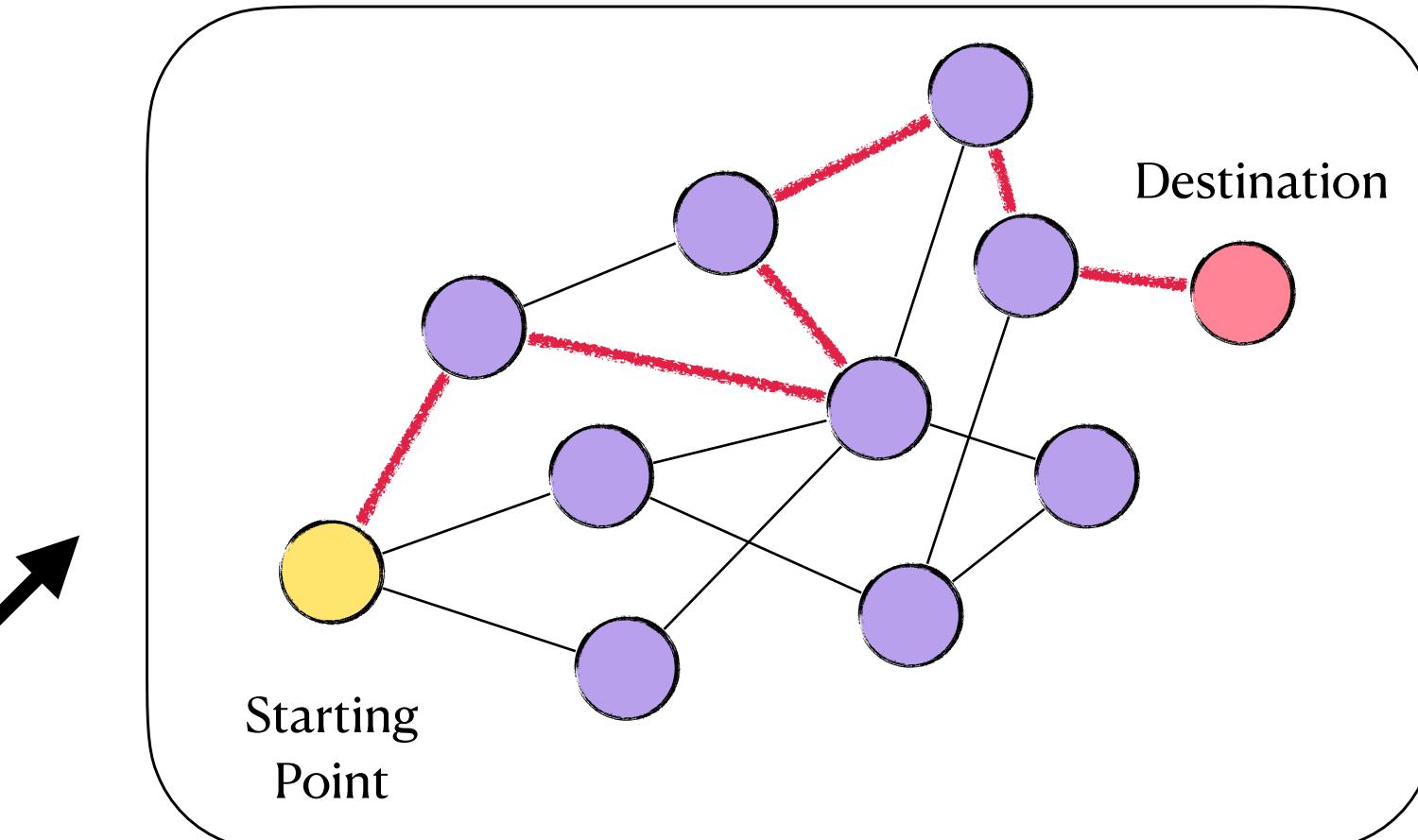
Learning



Modeling



Inference



The 3 pillars of Artificial Intelligence

Real world



Learning

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

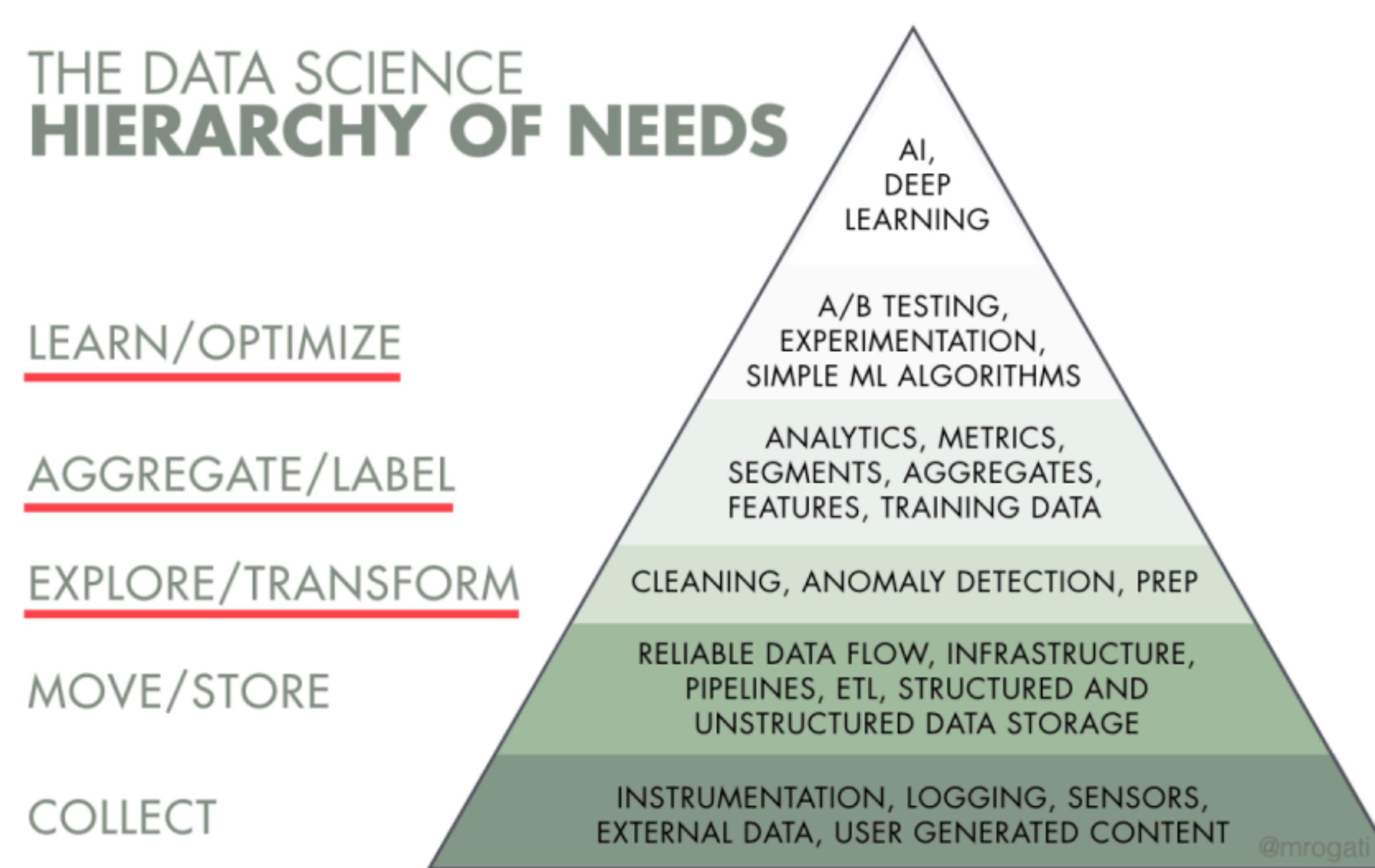
Inference

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

Modeling

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

The data science hierarchy of needs

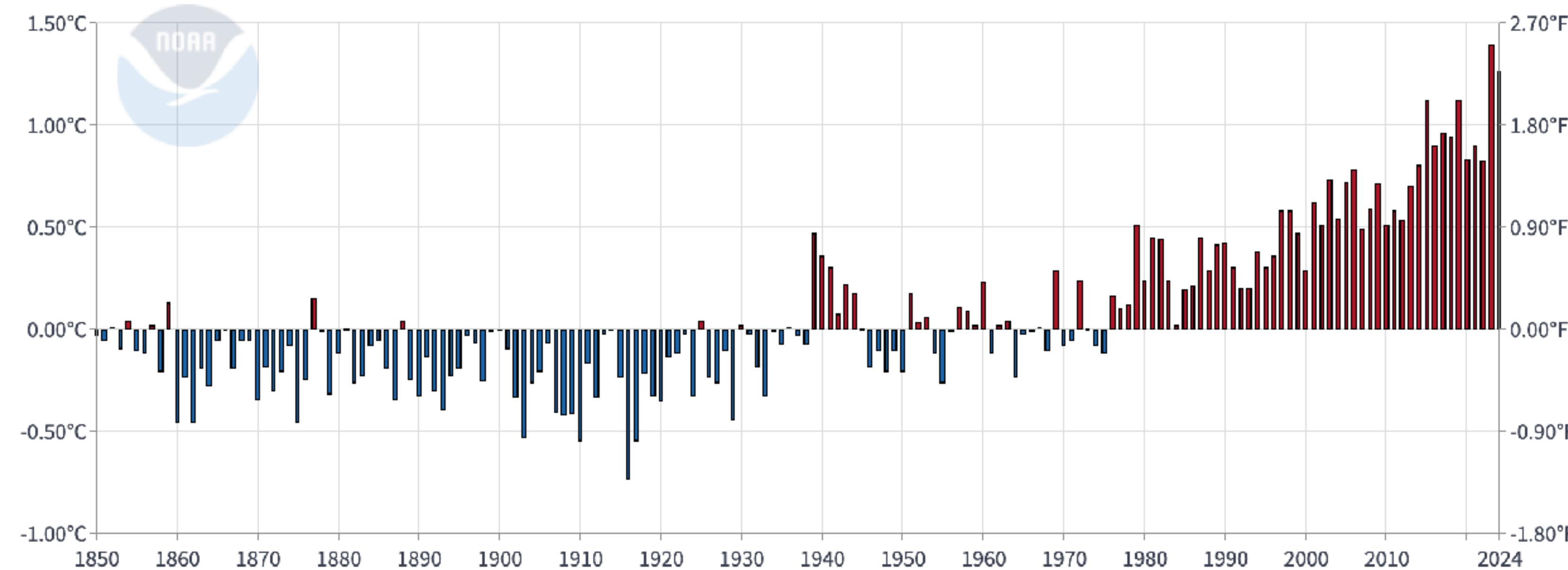


What are* Data?

Temperature

Global Land and Ocean Average Temperature Anomalies

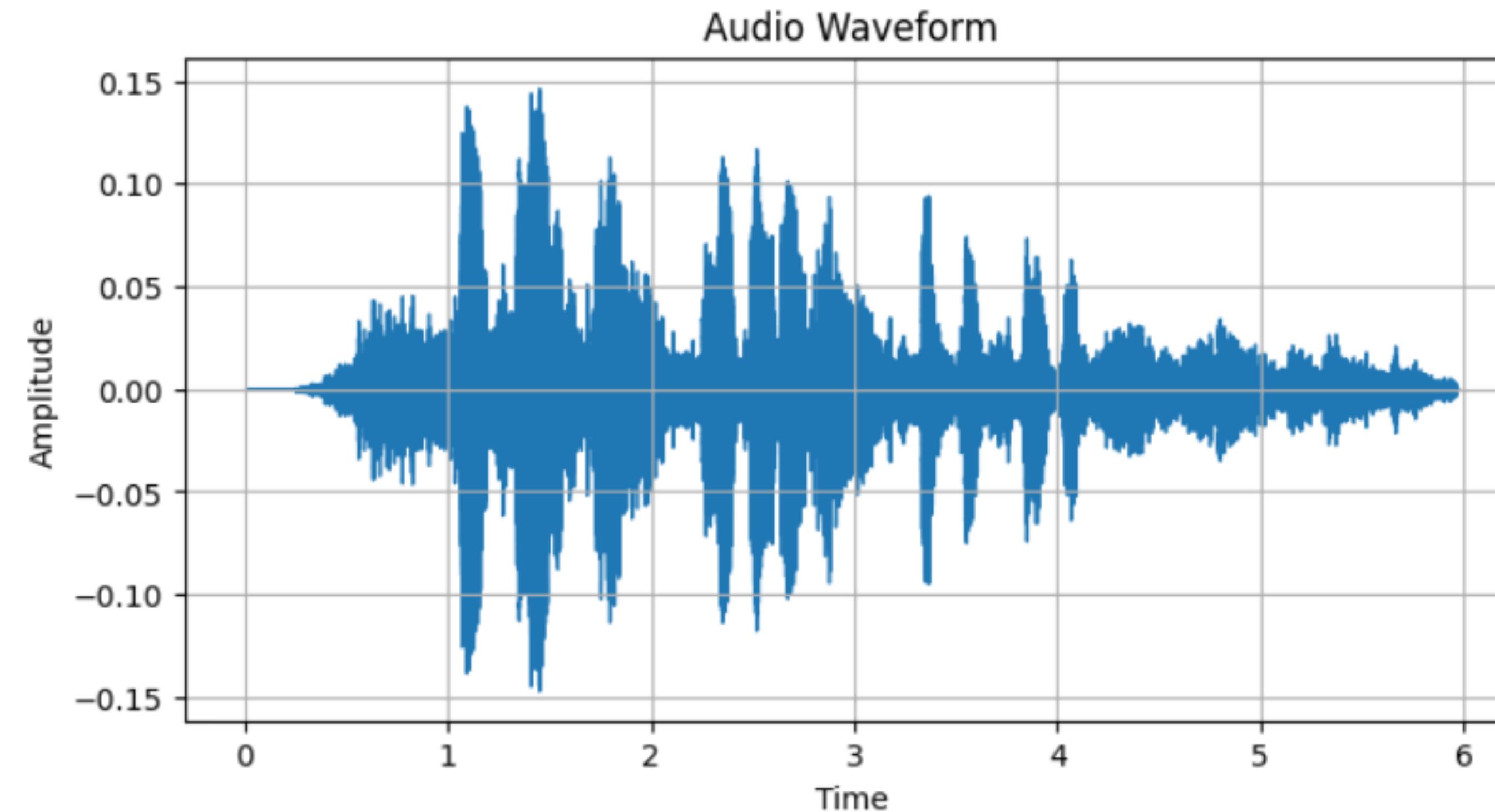
December



Powered by [ZingChart](#)

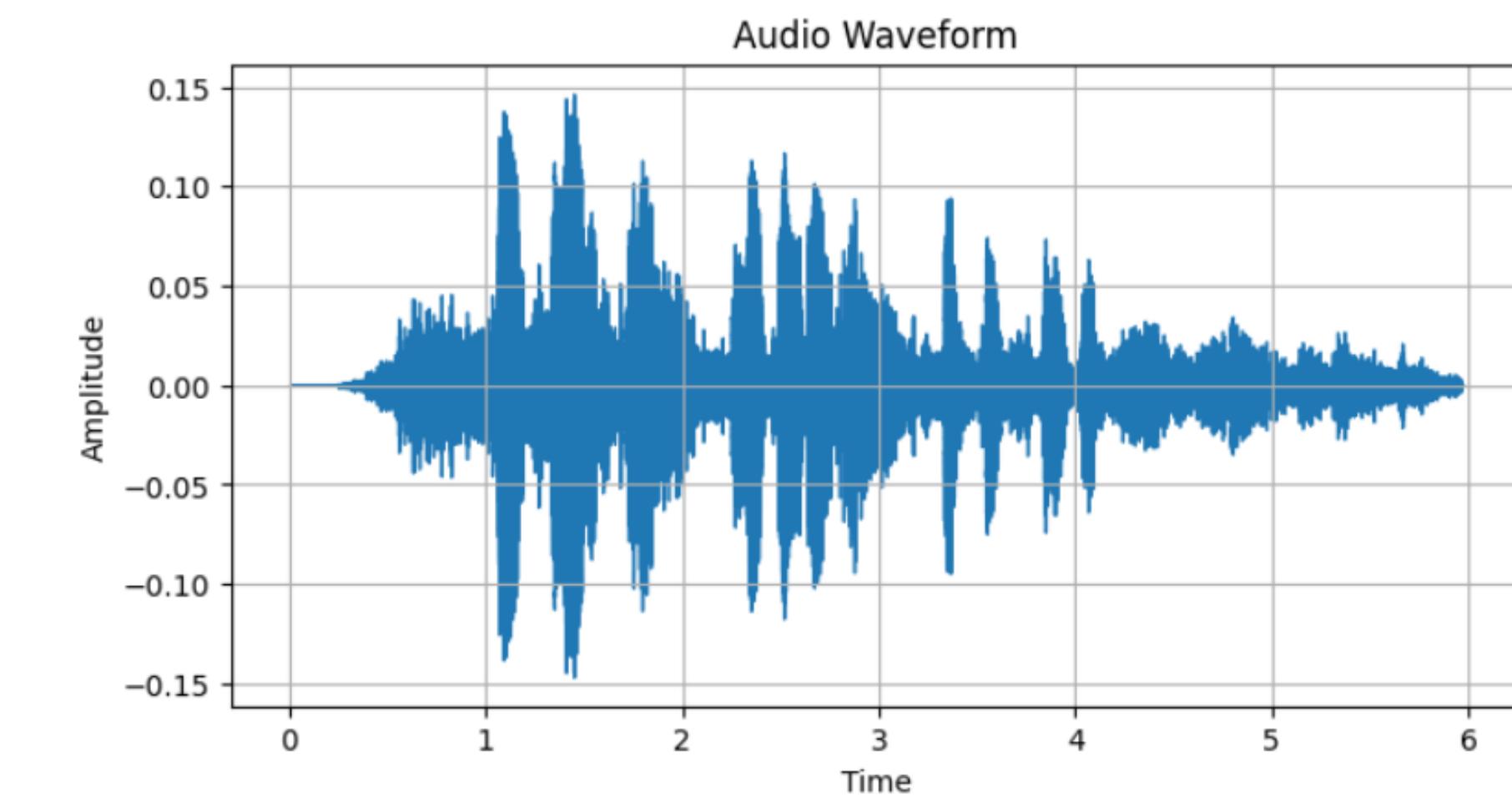
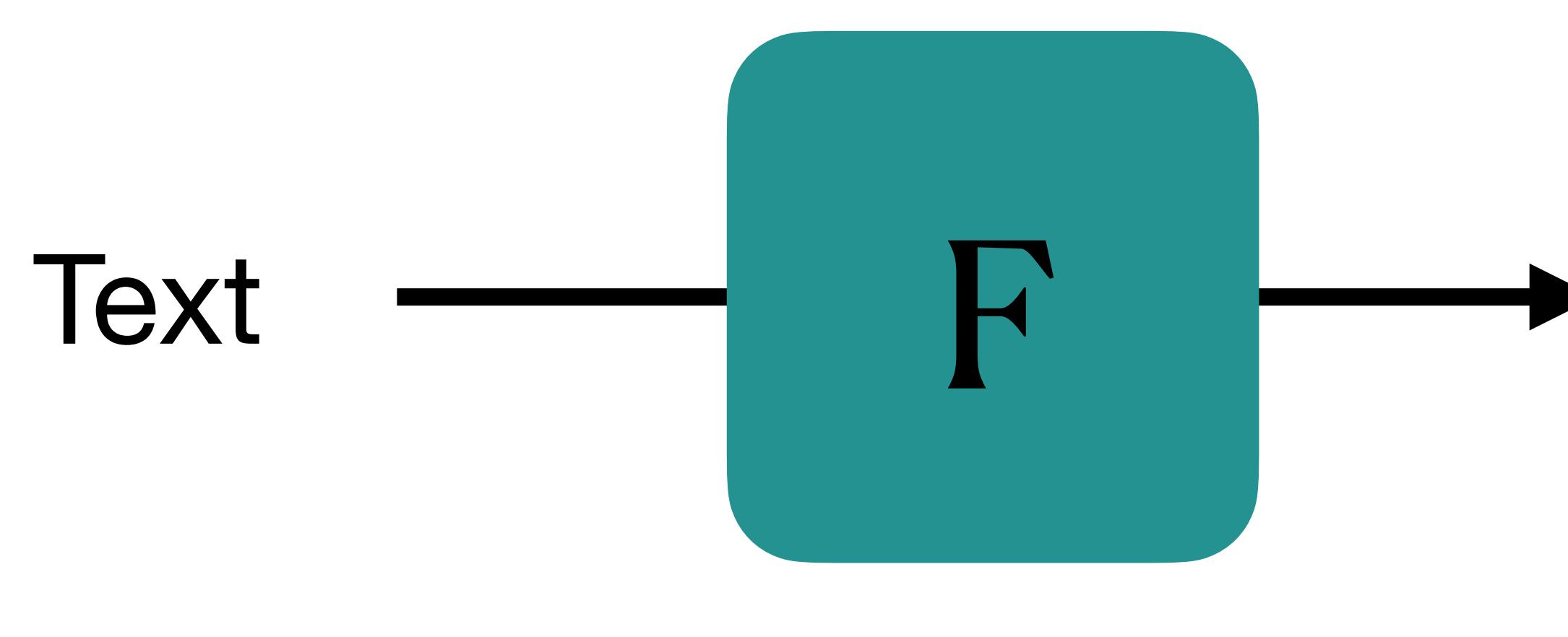
Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Temperature	0	1	-1	-2	-3	0	1	2	2	3	5	6	8	5	4	6

Audio



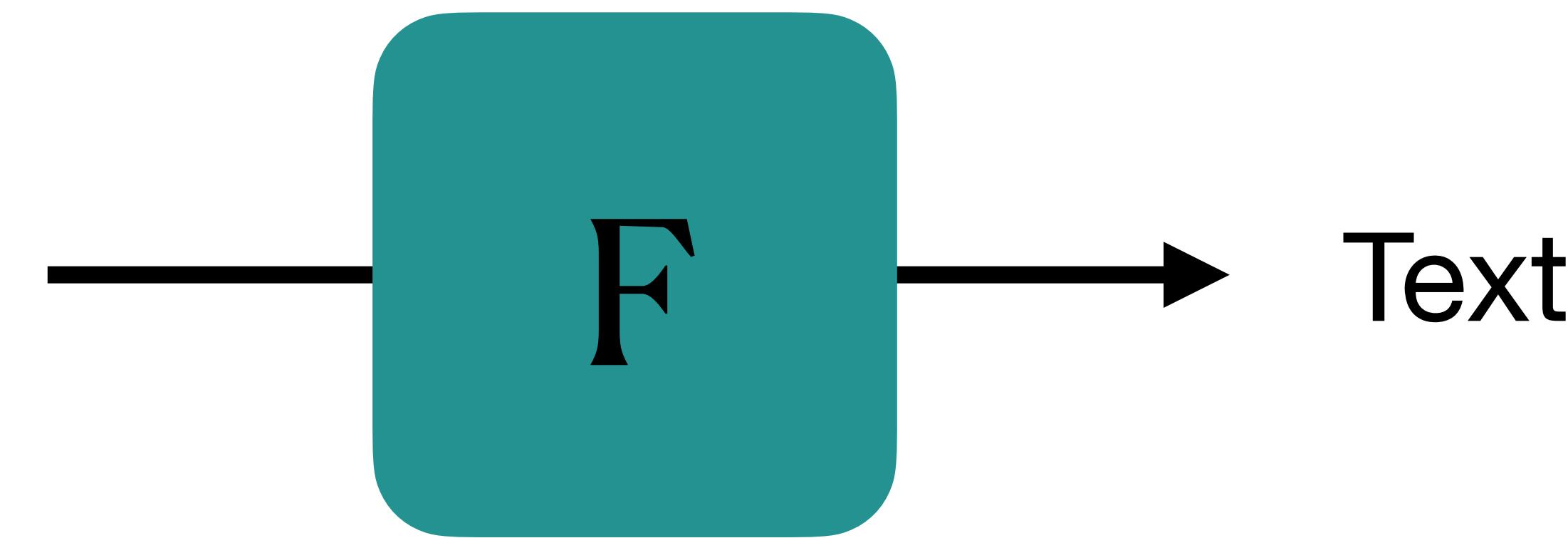
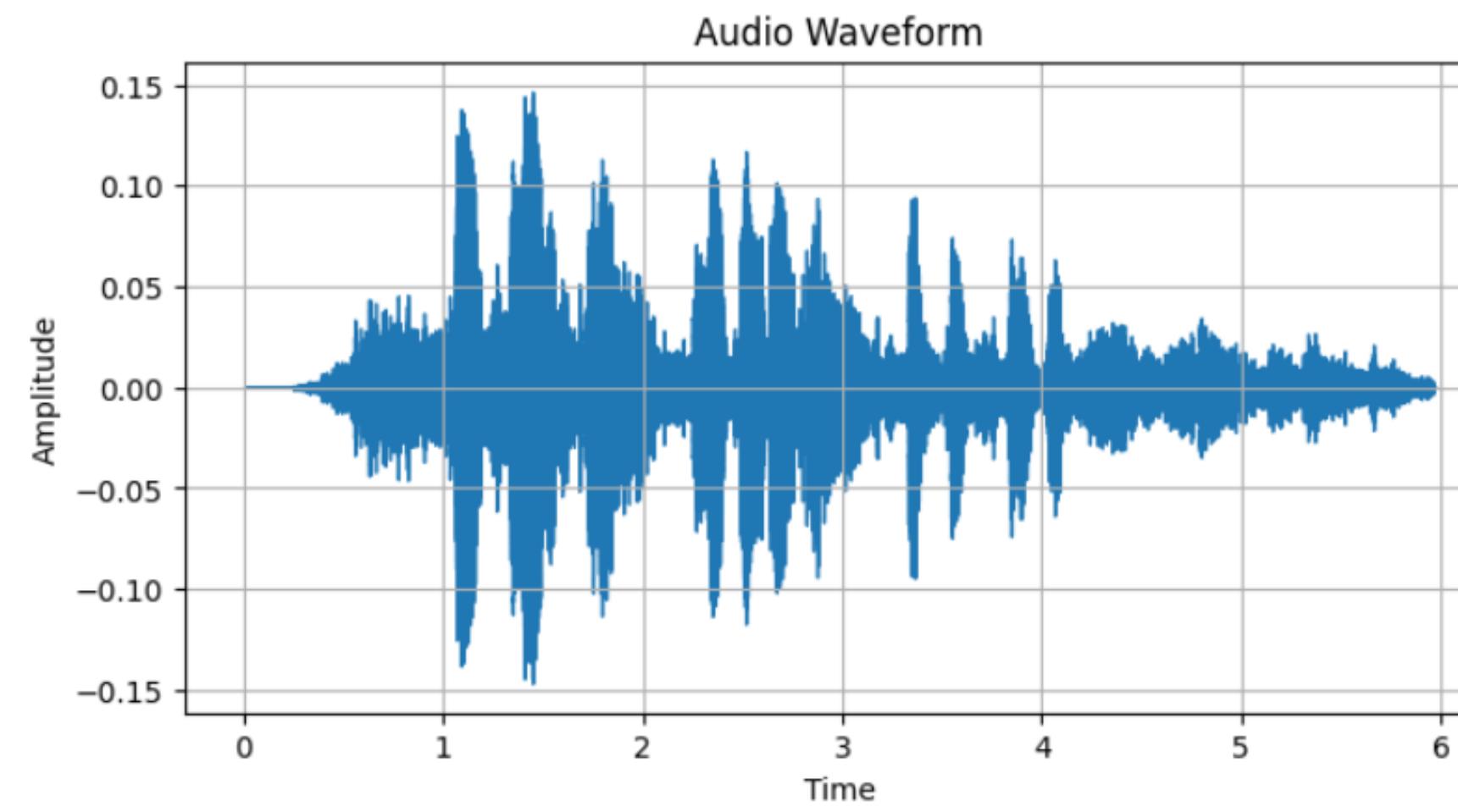
Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Amplitude	35	37	35	34	33	30	49	48	46	44	46	49	48	66	50	55

Audio



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

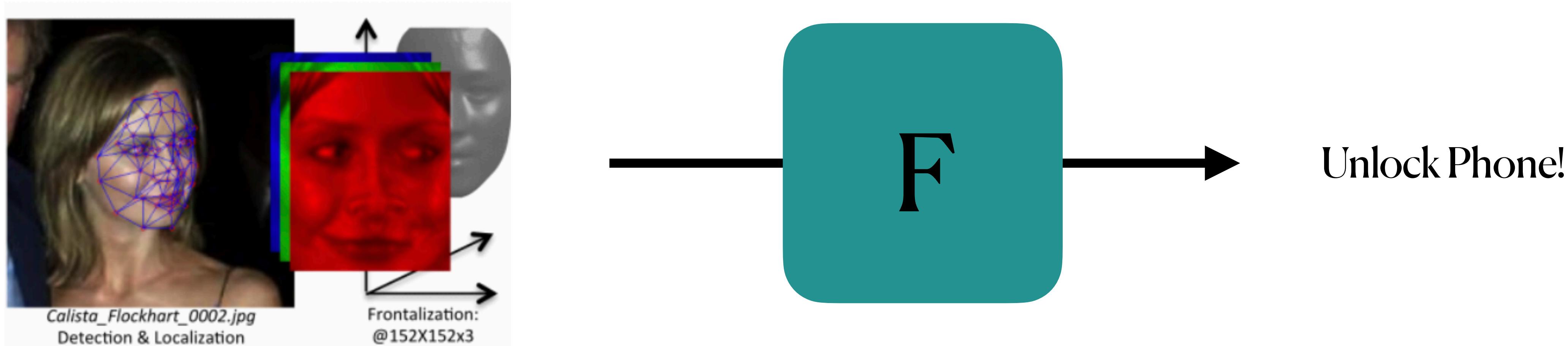


Channels

123	4	6	43
35	44	33	255
65	2	0	0
123	78	43	43
35	43	33	54
65	2	0	55
33	54	0	255
33	54	0	0

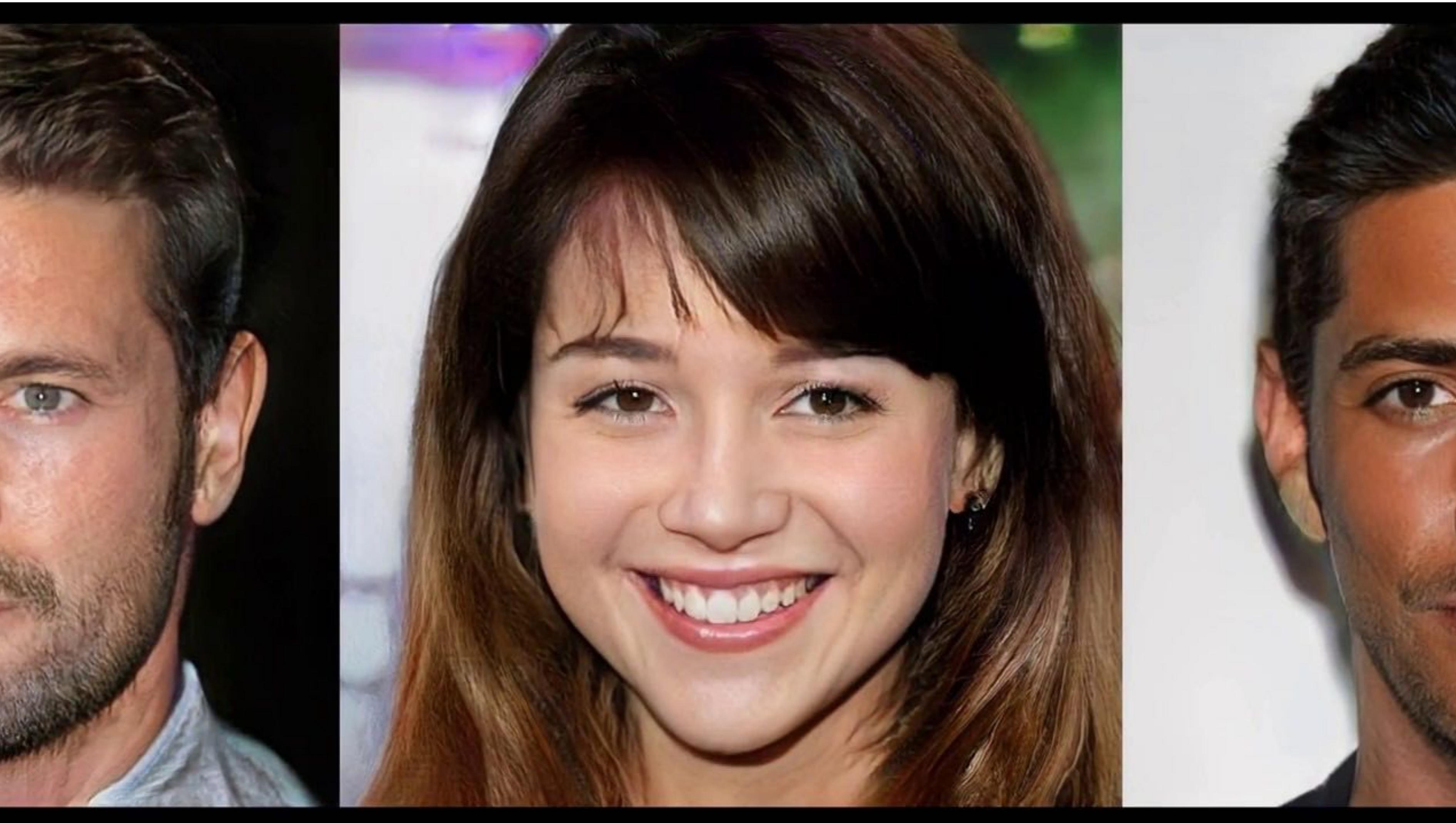
Image dimension: $h \times w \times 3$

Image recognition



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

Generative Adversarial Networks (2014)

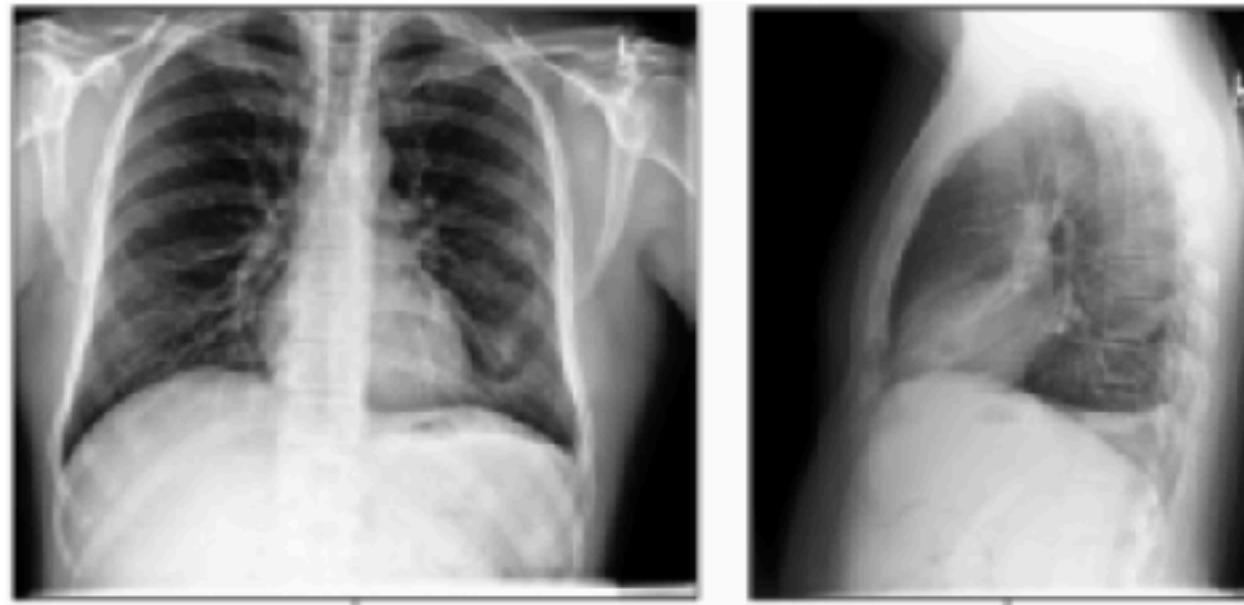


Progressive Growing of GANs for Improved Quality, Stability, and Variation

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

Image recognition

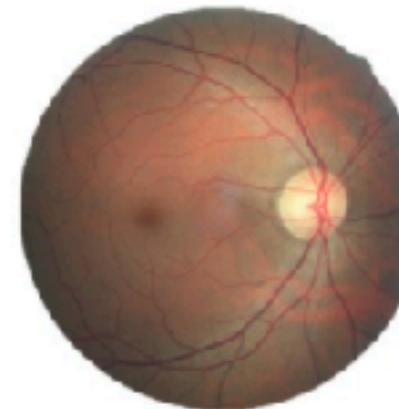
Chest radiology



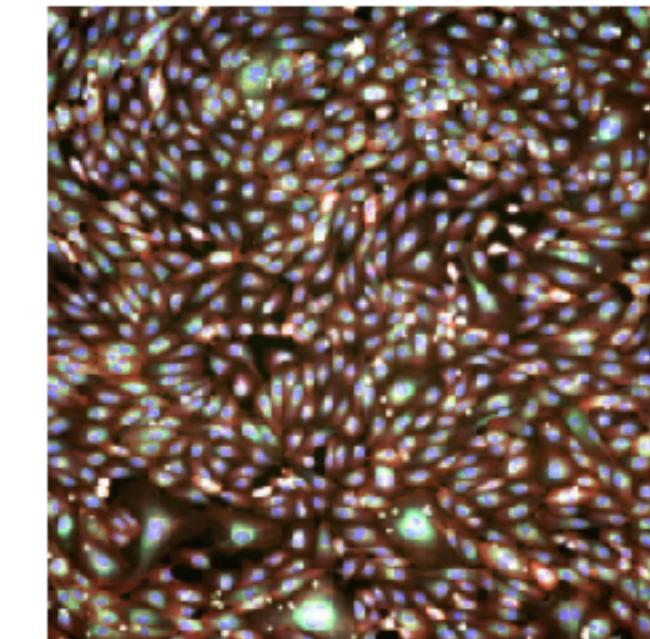
2019

**Google AI outperforms
radiologists at lung cancer
detection**

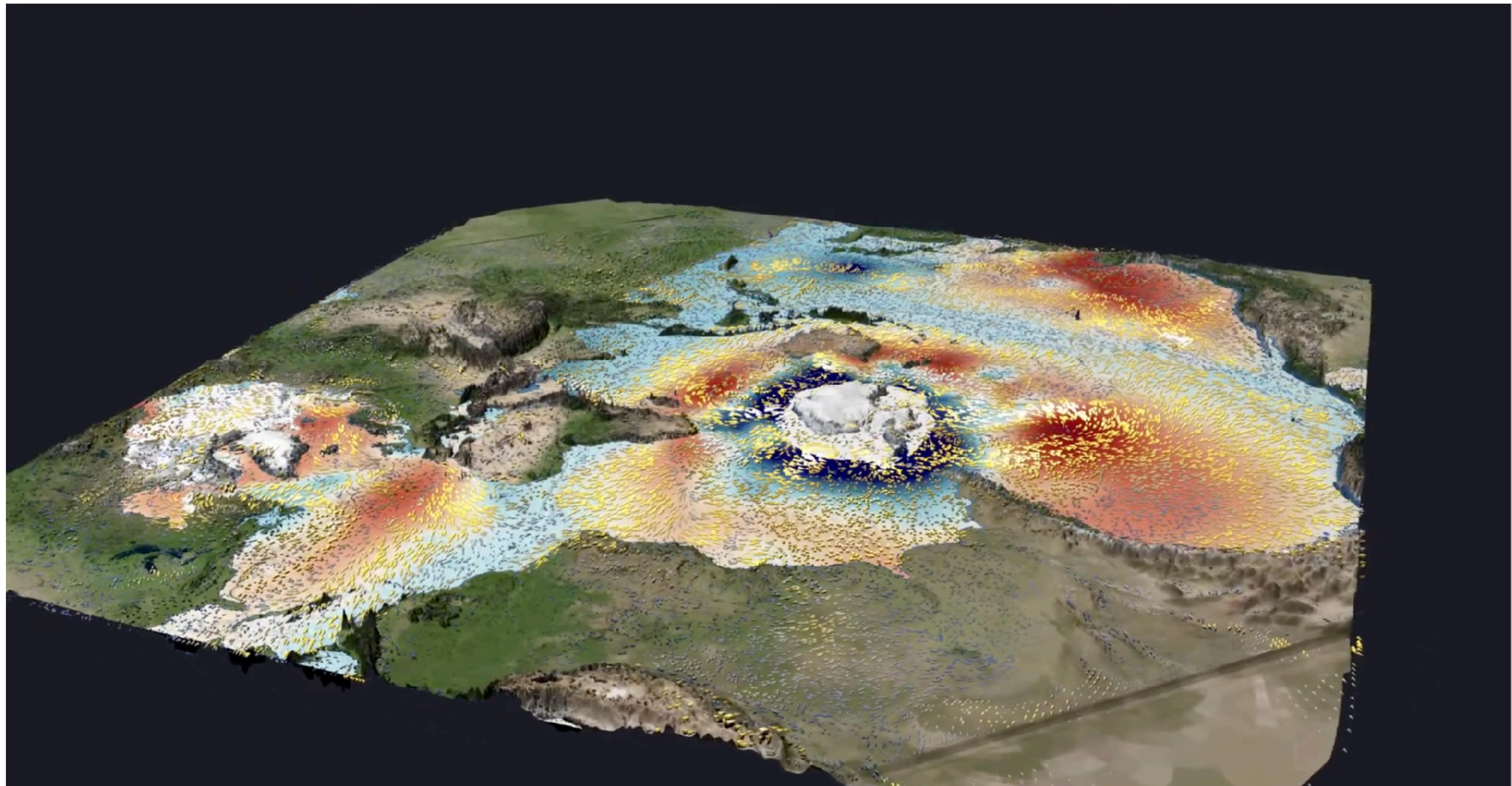
Diabetic retinopathy



Drug screening for COVID-19



Coupled Ocean Atmosphere Simulation



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

Video

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



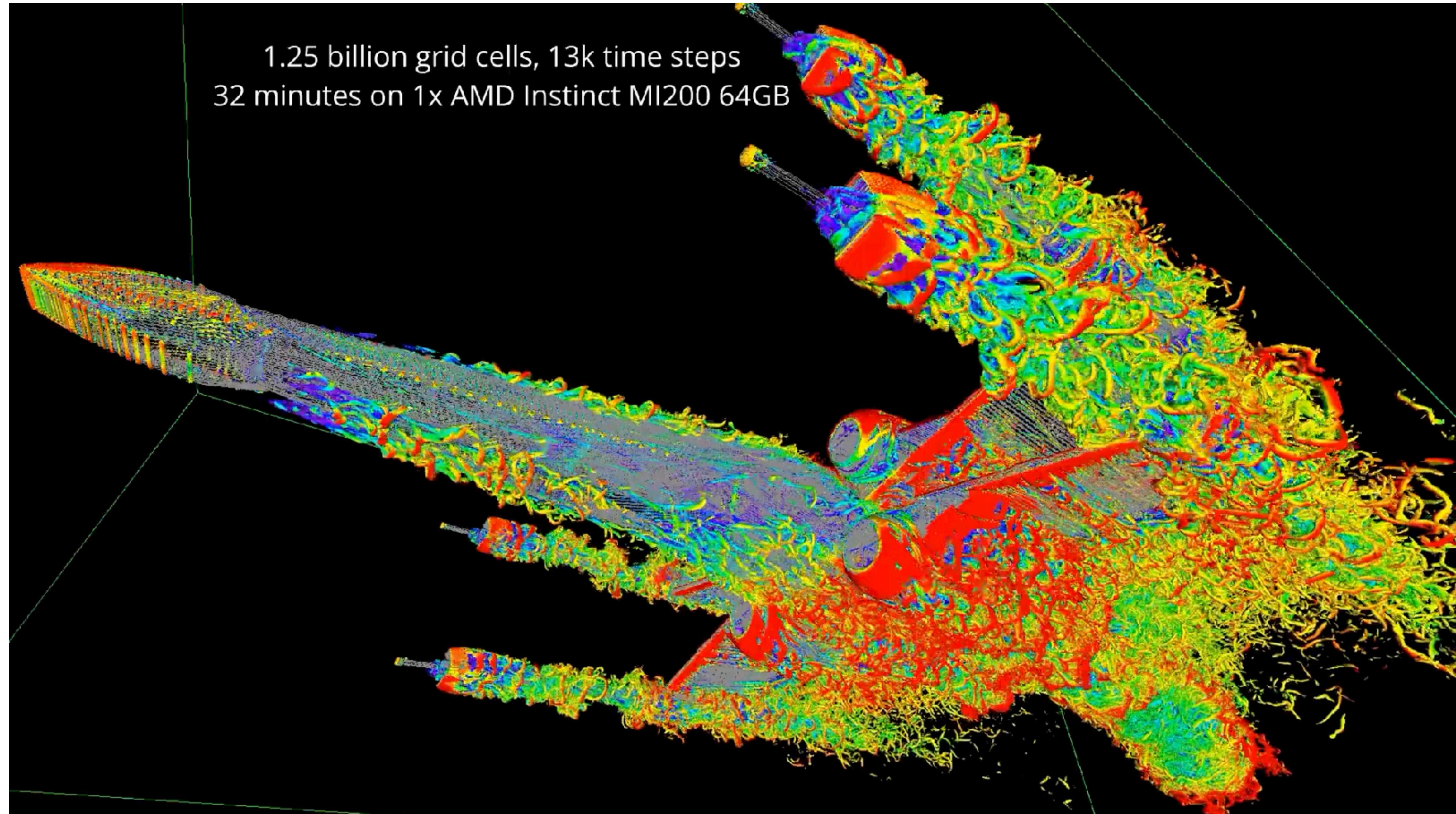
t_1	123	4	6	43
123	56	35	44	33 255
35	65	2	0	0
43	33	54	0	0
65	2	0	55	255
33	54	0	0	

t_2	123	4	6	43
123	56	35	44	33 255
35	65	2	0	0
43	33	54	0	0
65	2	0	55	255
33	54	0	0	

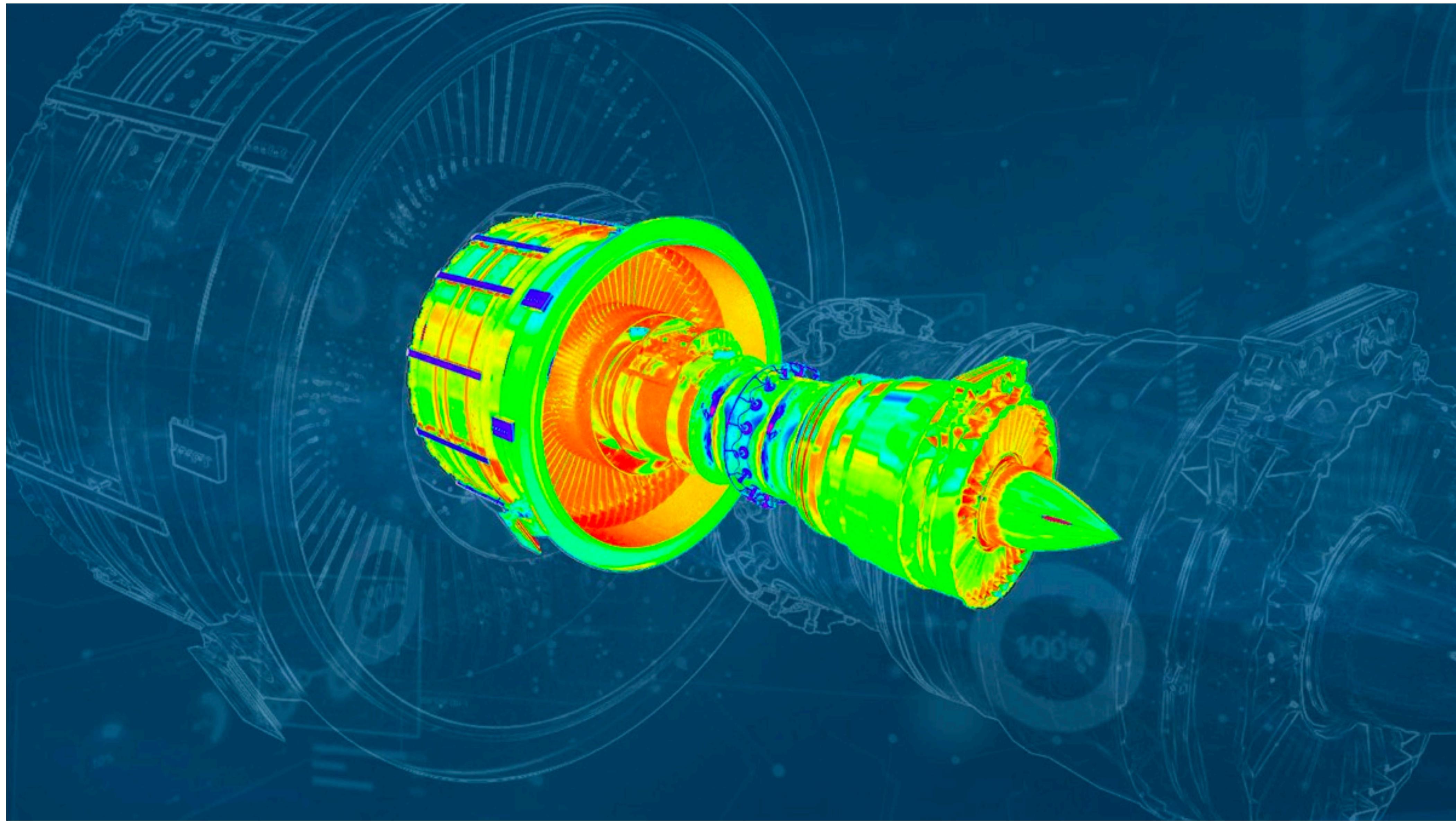
.....

T	123	4	6	43
123	56	35	44	33 255
35	65	2	0	0
43	33	54	0	0
65	2	0	55	255
33	54	0	0	

Fluid dynamics



Machine learning in materials research



Language

Deep learning surpasses traditional approaches
in speech recognition in 2010



Google
ASSISTANT

2016

amazon alexa

2014

Siri

2011

IBM-Watson Defeats Humans in "Jeopardy!"



2011

Text

One-hot word representation

Dictionary

The World Is A Crazy Place

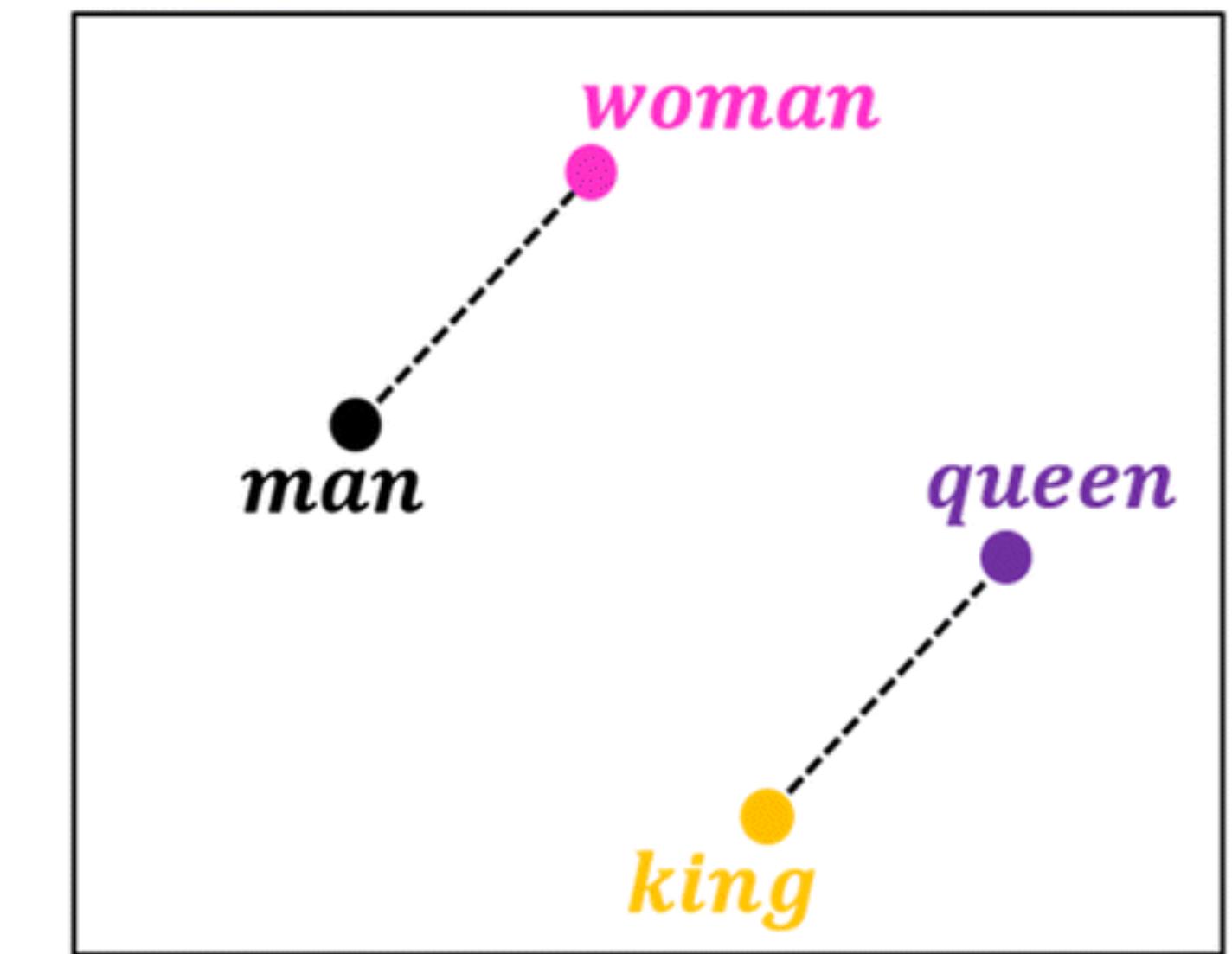
a
the →
world
crazy
place
is
:

0	0	0	1	0	0
1	0	0	0	0	0
0	1	0	0	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	1	0	0	0
:	:	:	:	:	:

Text

Word Embedding Representation

Words	Embedding
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]



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

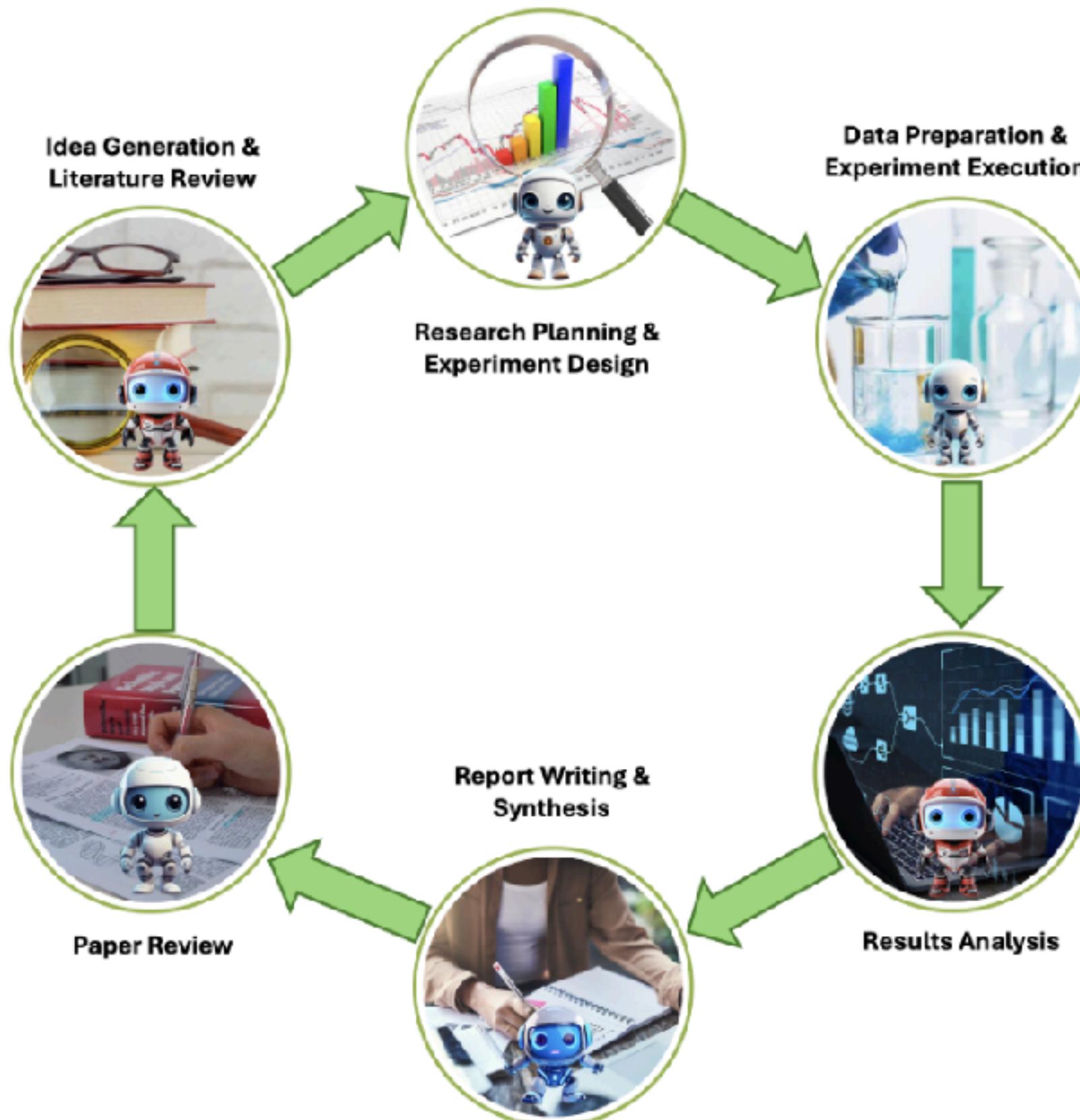


Figure 1: Agentic AI workflow for scientific discovery.

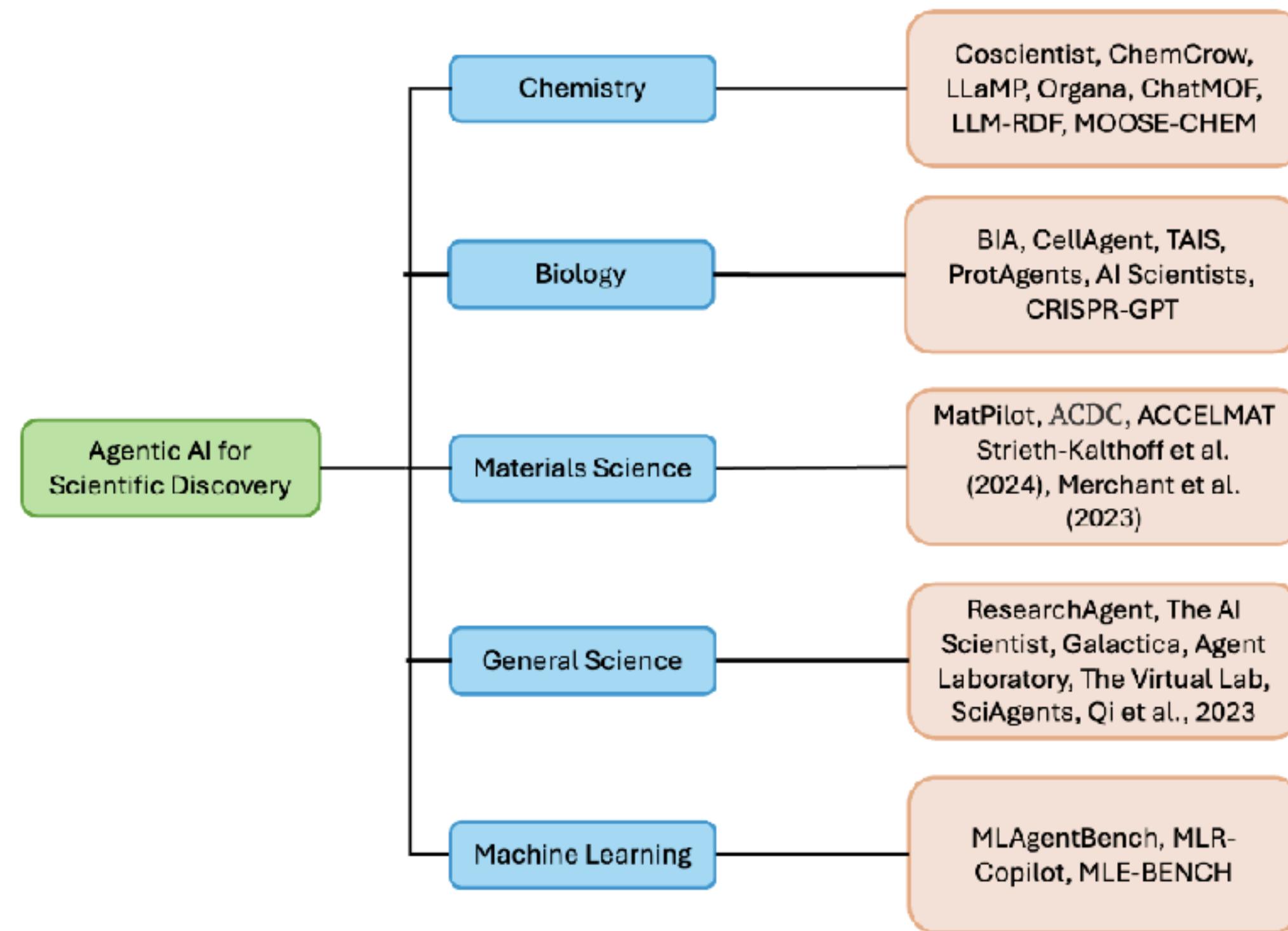


Figure 2: AI Agents frameworks for scientific discovery.

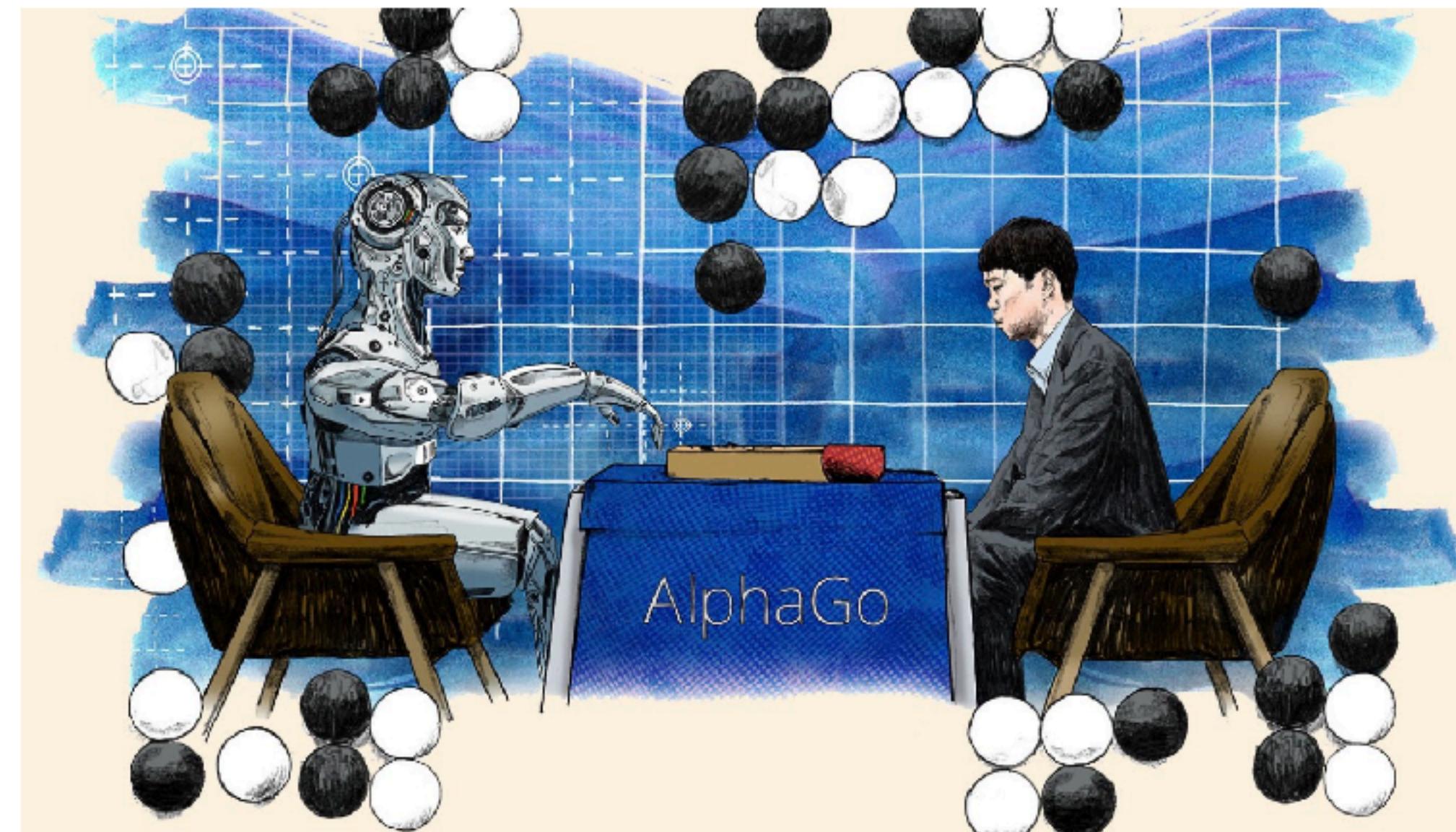
Games

IBM creates a checker playing program



1959

AlphaGo beats
world's best Go player



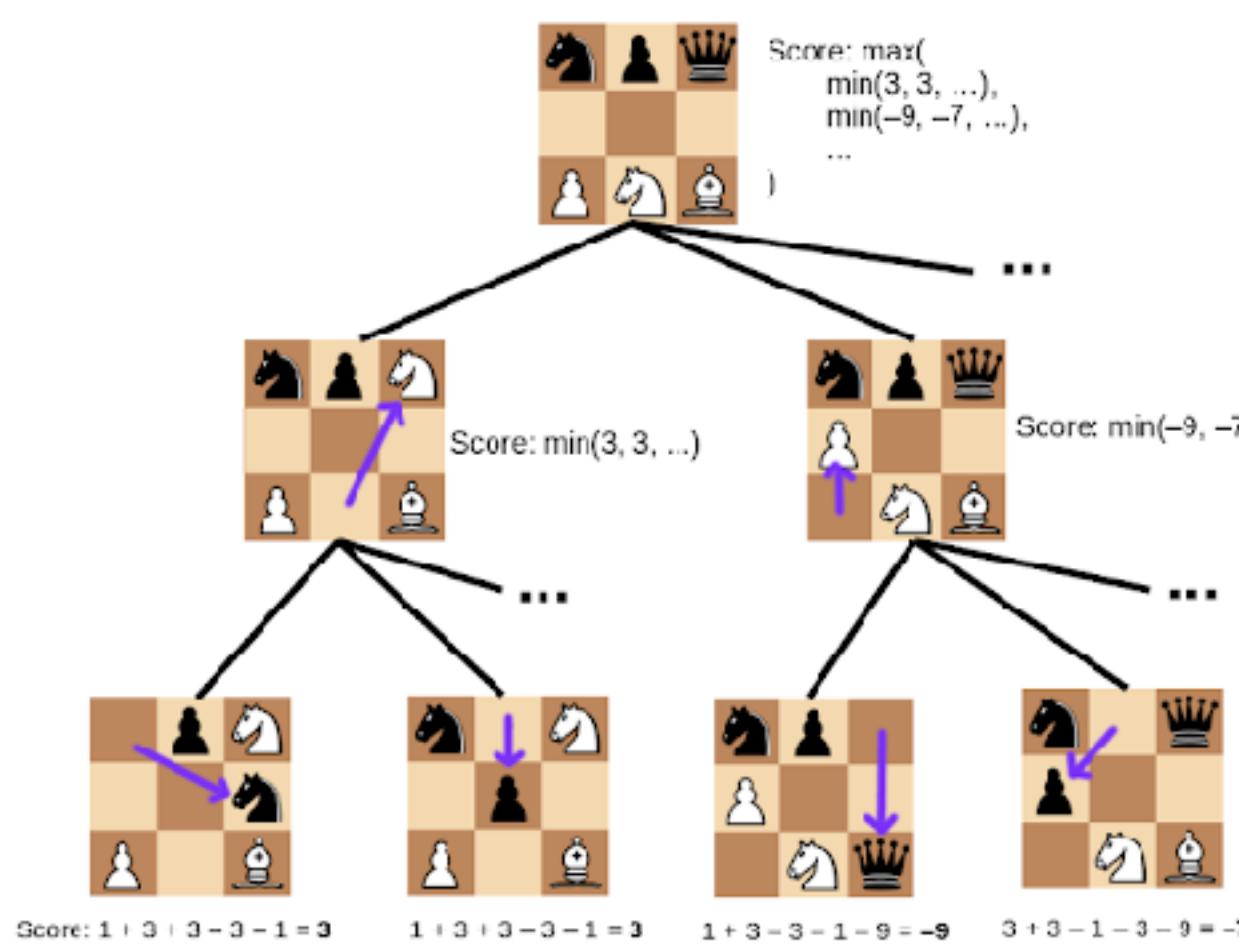
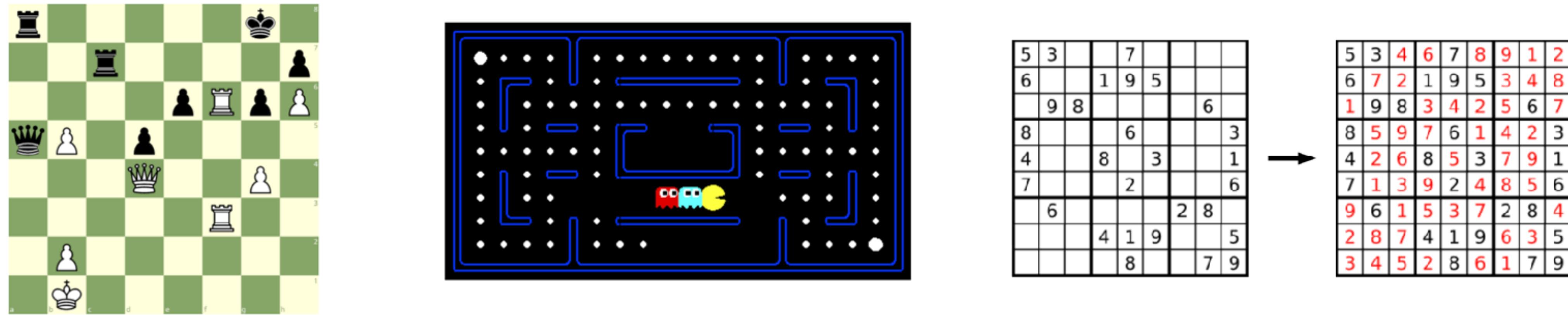
2016

IBM Deep Blue
beats Kasparov

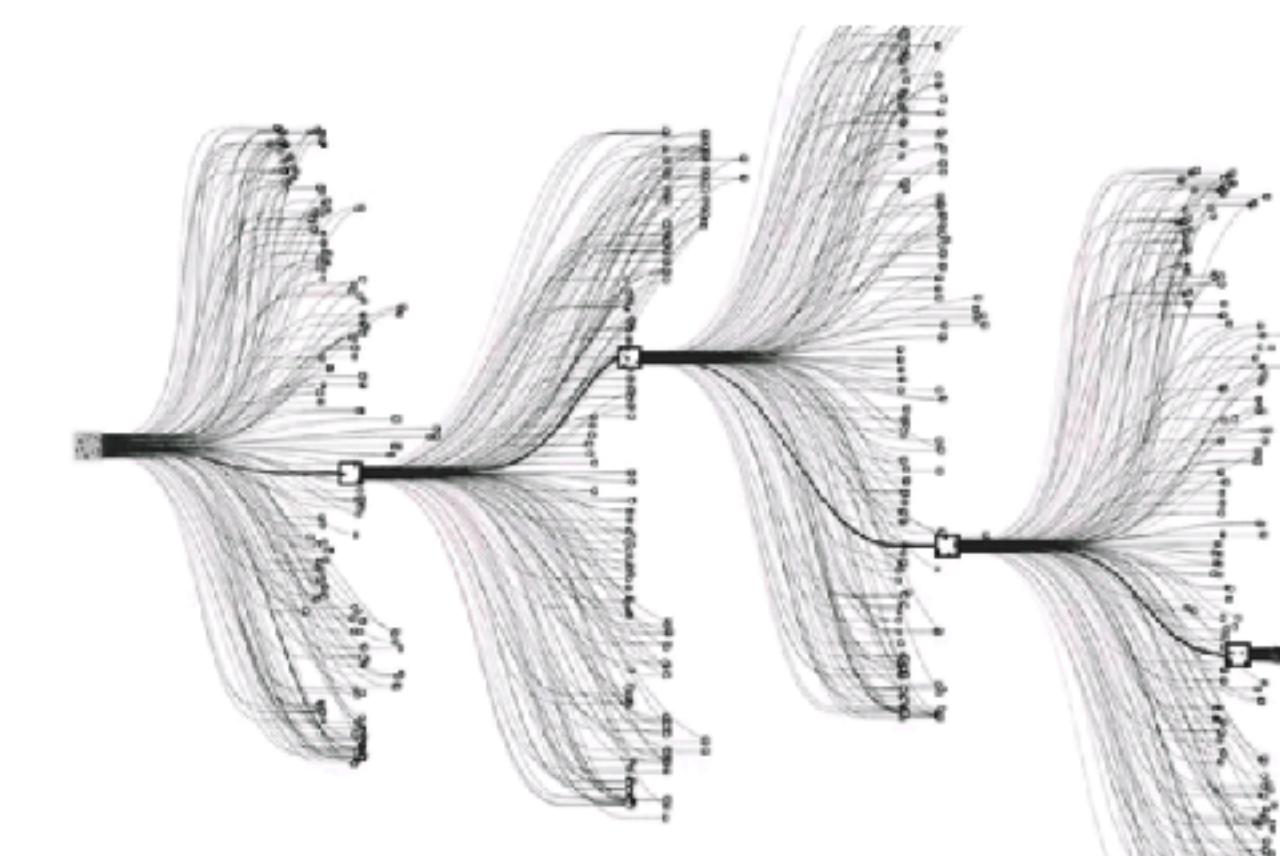


1997

Solving puzzles

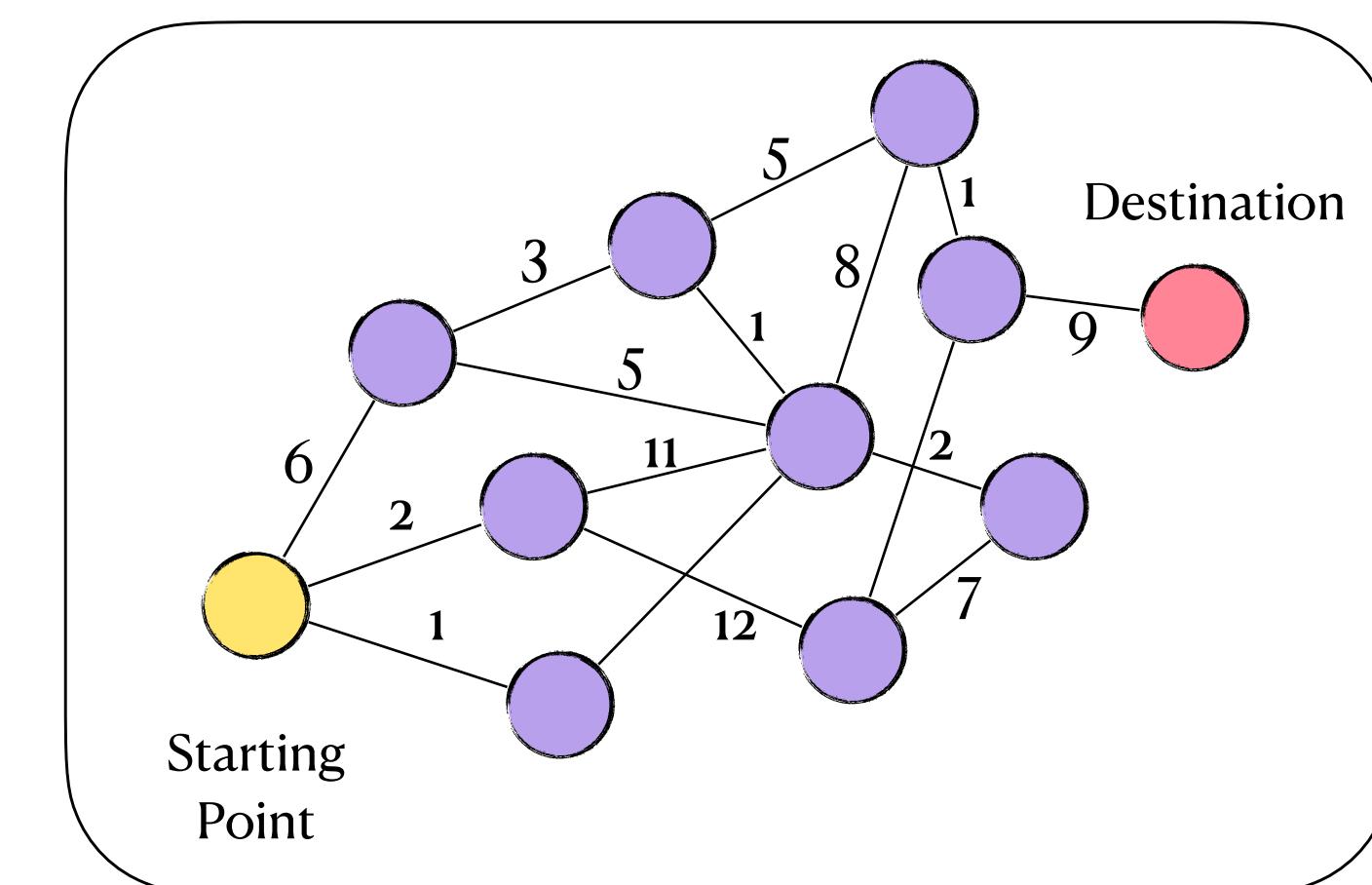
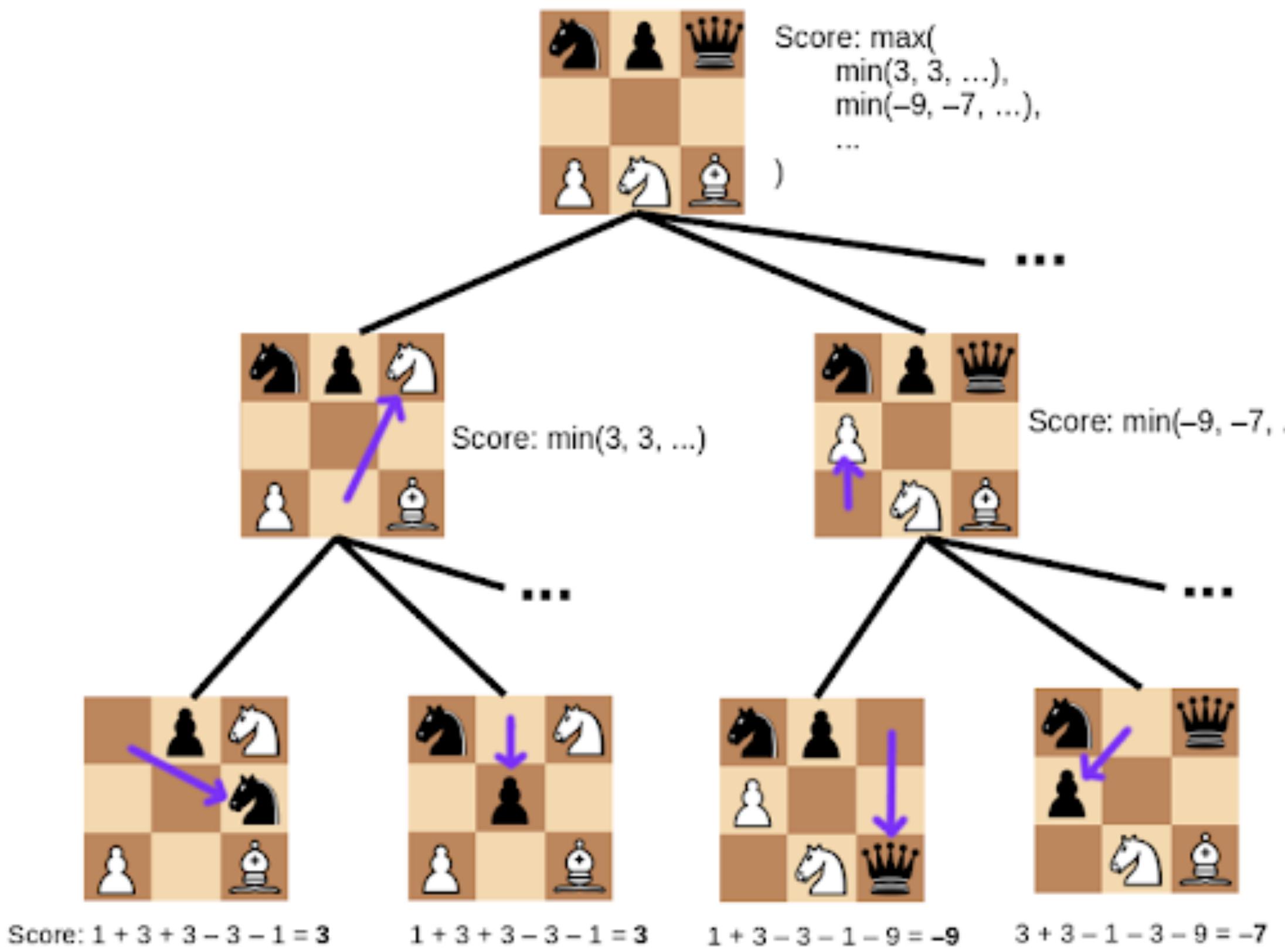
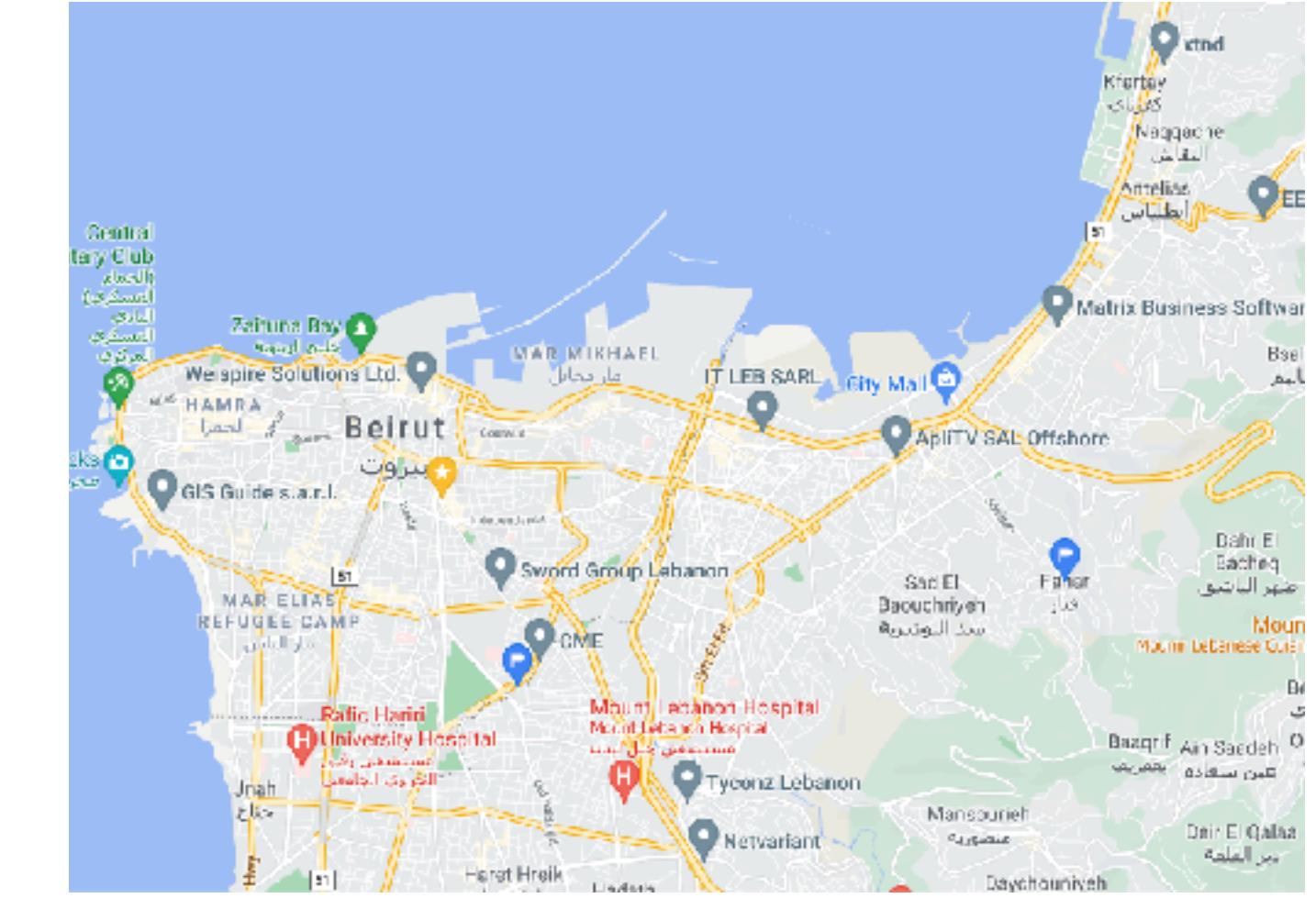


10^{120} possibilities



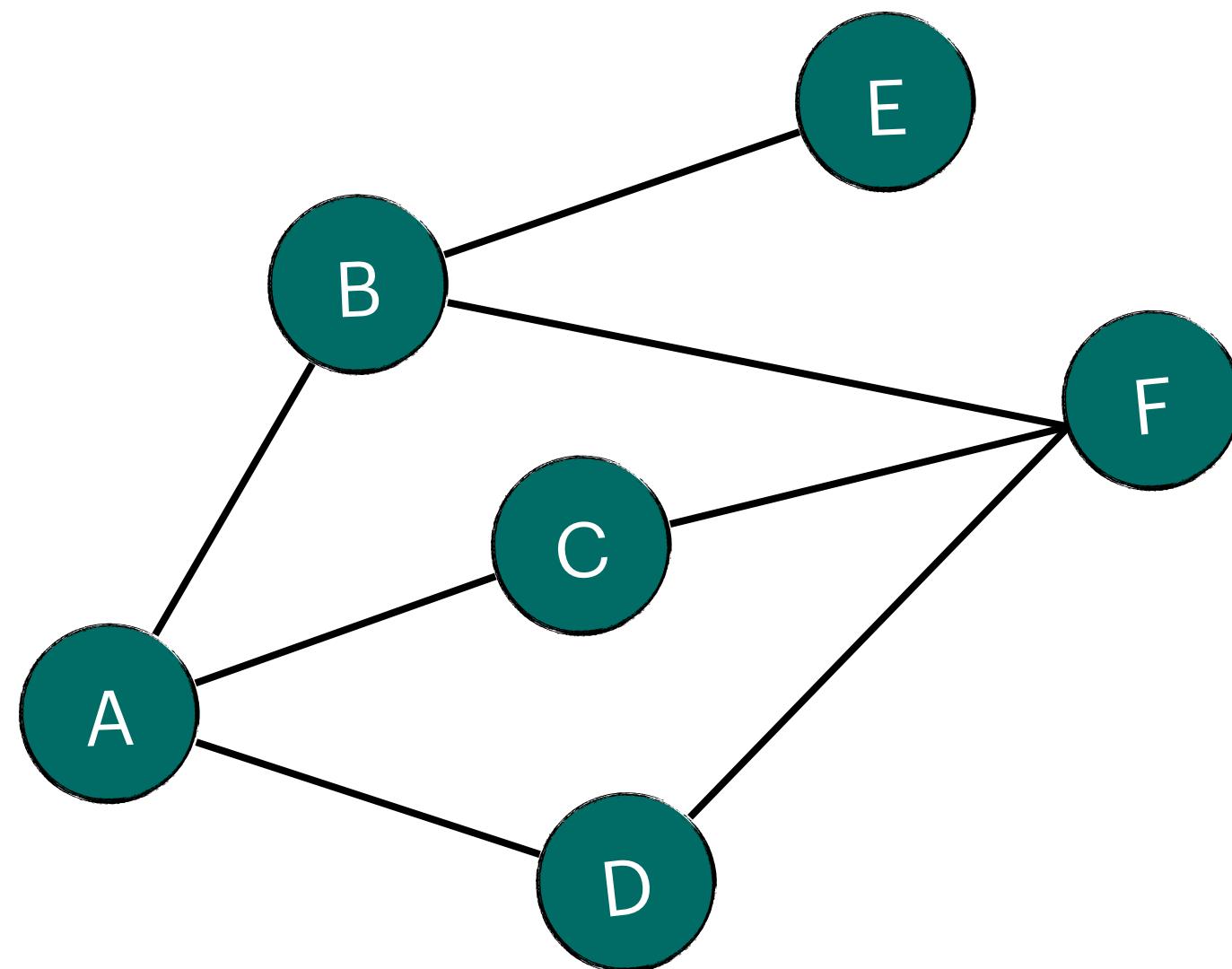
Graph Representation

Real world



Graph Representation

Graph with Vertices and Edges

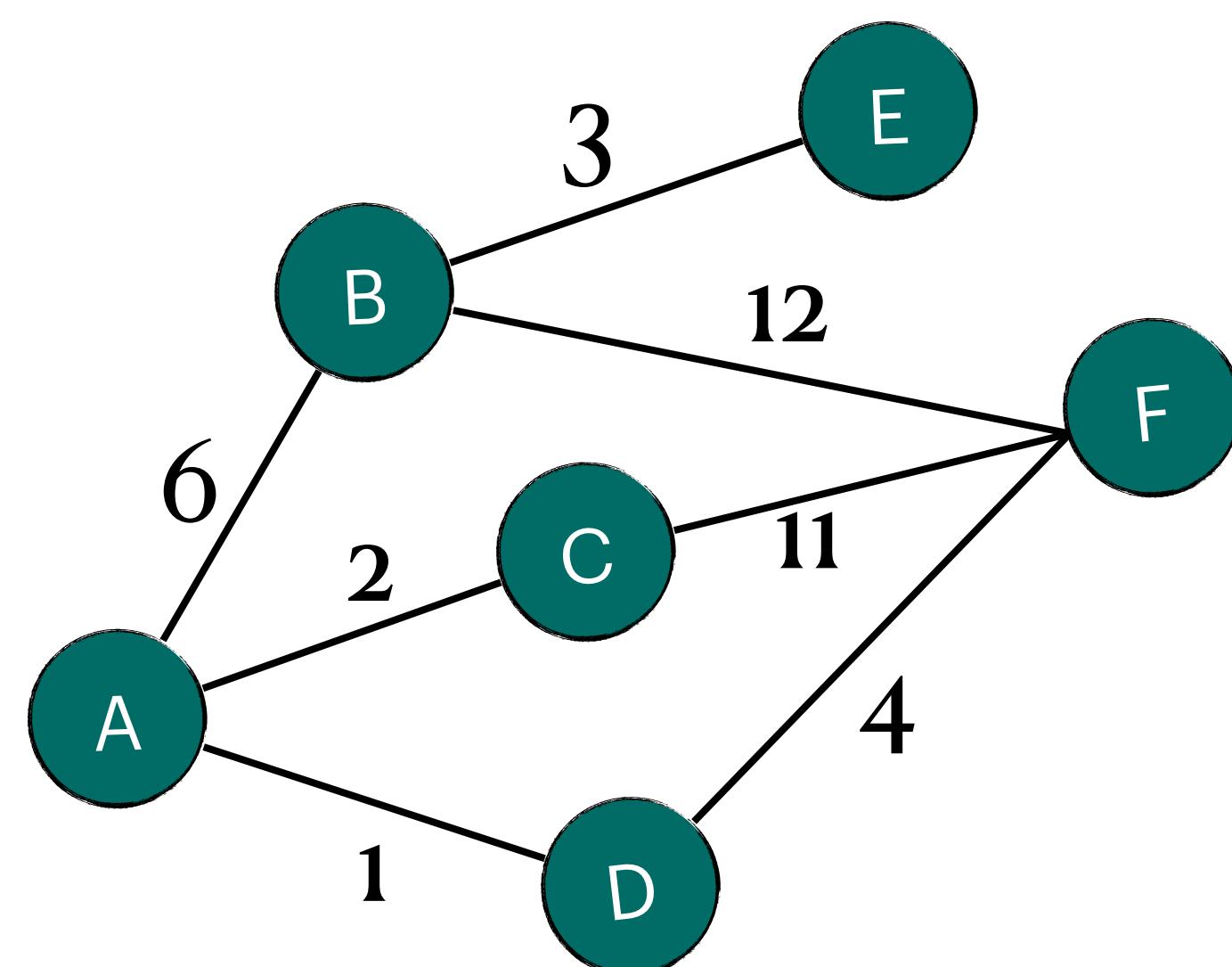


Adjacency Matrix

	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

Graph Representation

Weighted Graph

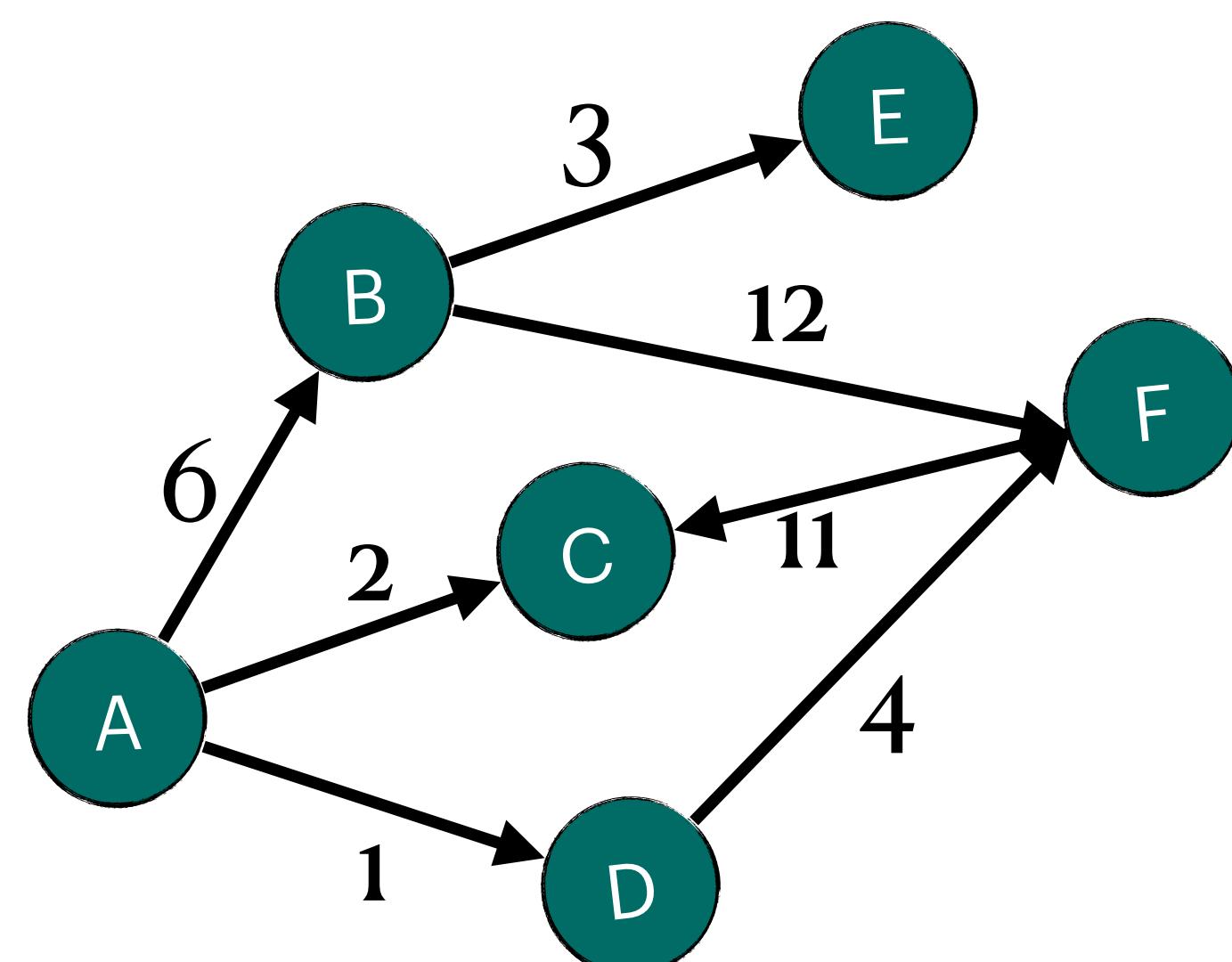


Adjacency Matrix

	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

Graph Representation

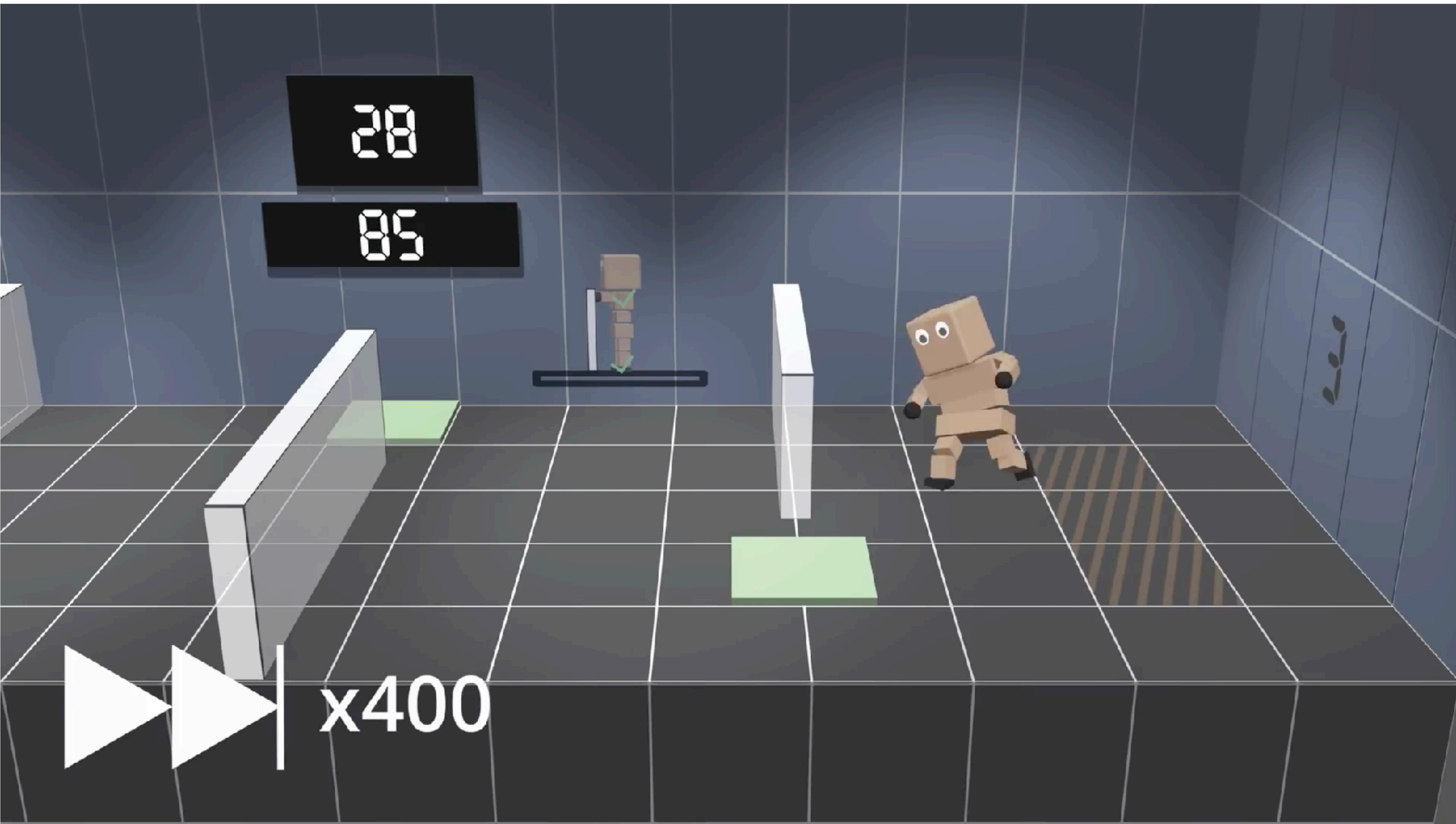
Weighted Directed Graph



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

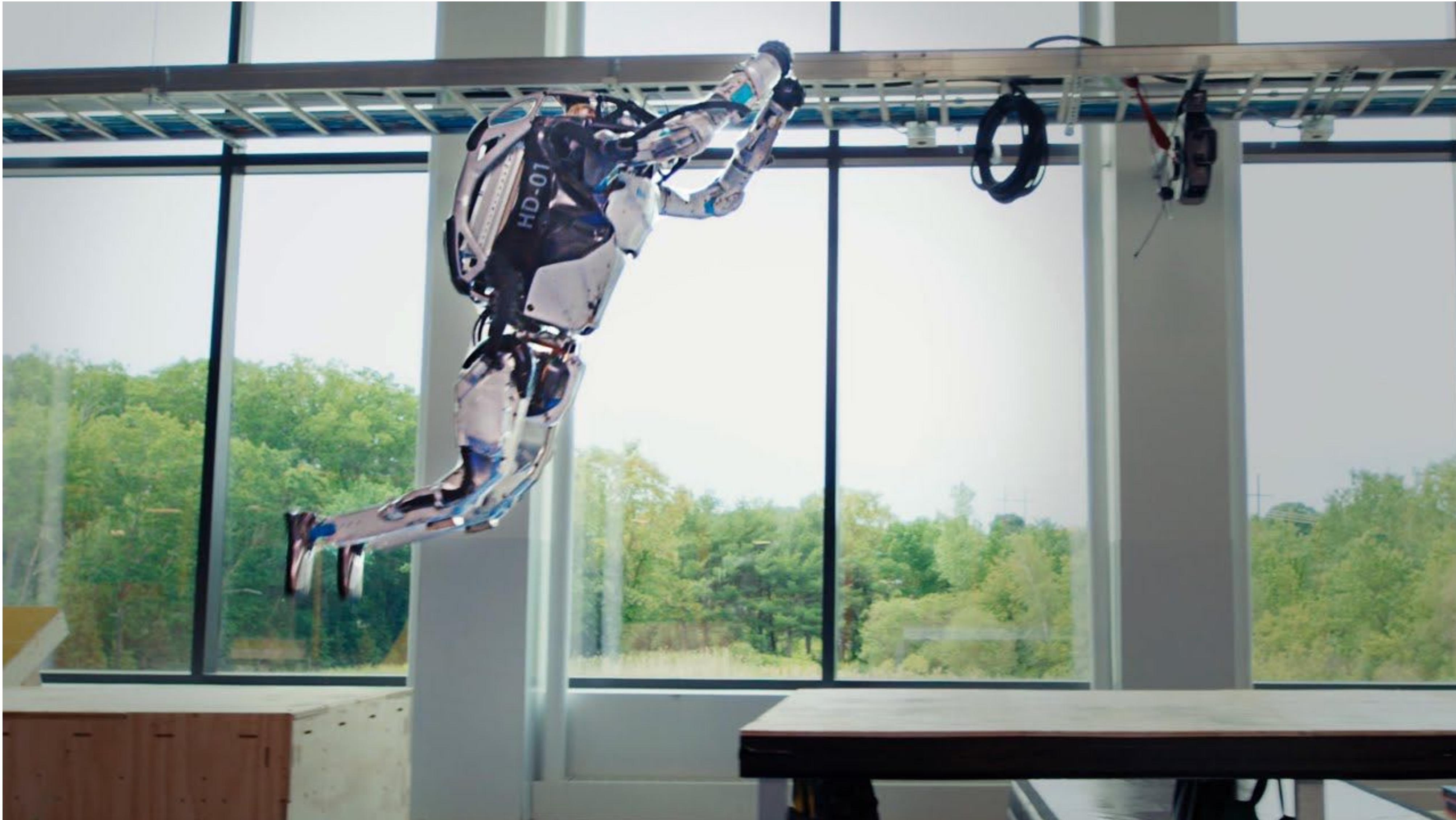
Robotics and games: Reinforcement Learning



Robotics and Manufacturing

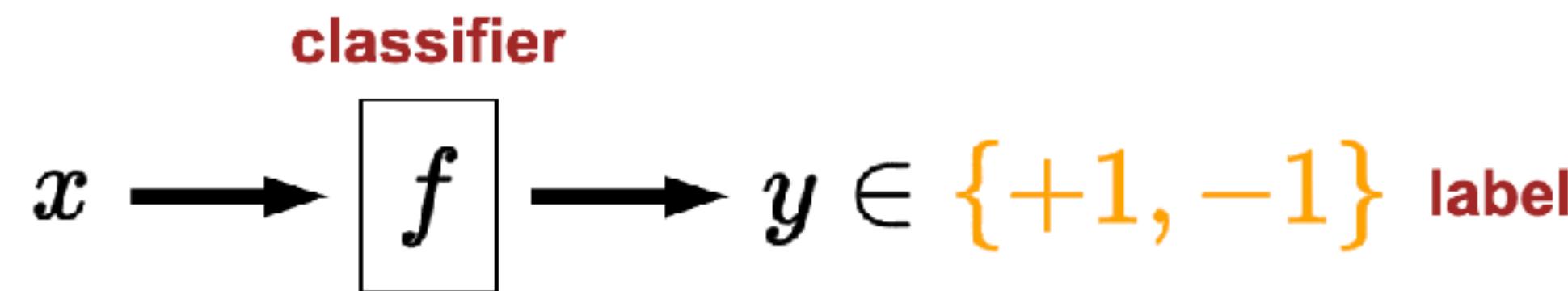


Robotics



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

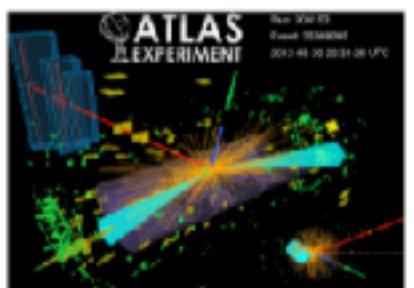
Classification



Fraud detection: credit card transaction → fraud or no fraud



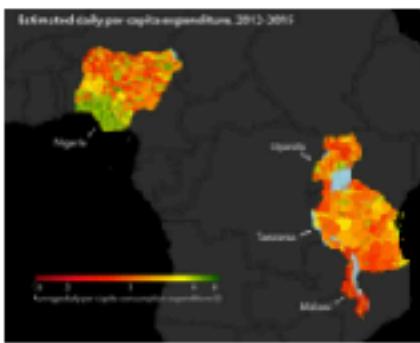
Toxic comments: online comment → toxic or not toxic



Higgs boson: measurements of event → decay event or background

Regression

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



Poverty mapping: satellite image → asset wealth index



Housing: information about house → price



Arrival times: destination, weather, time → time of arrival

Structured Prediction

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



Machine translation: English sentence → Japanese sentence



Dialogue: conversational history → next utterance



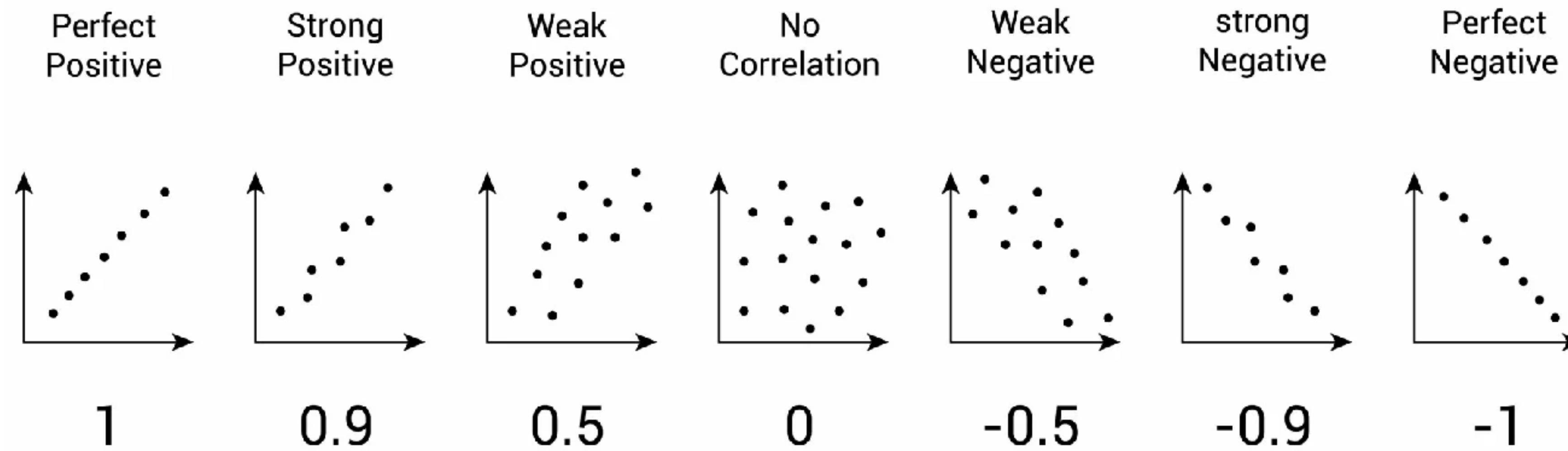
Image captioning: image → sentence describing image



Image segmentation: image → segmentation

Data Analysis

Correlation coefficient (2 variables)



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

Correlation Matrix (multiple variables)

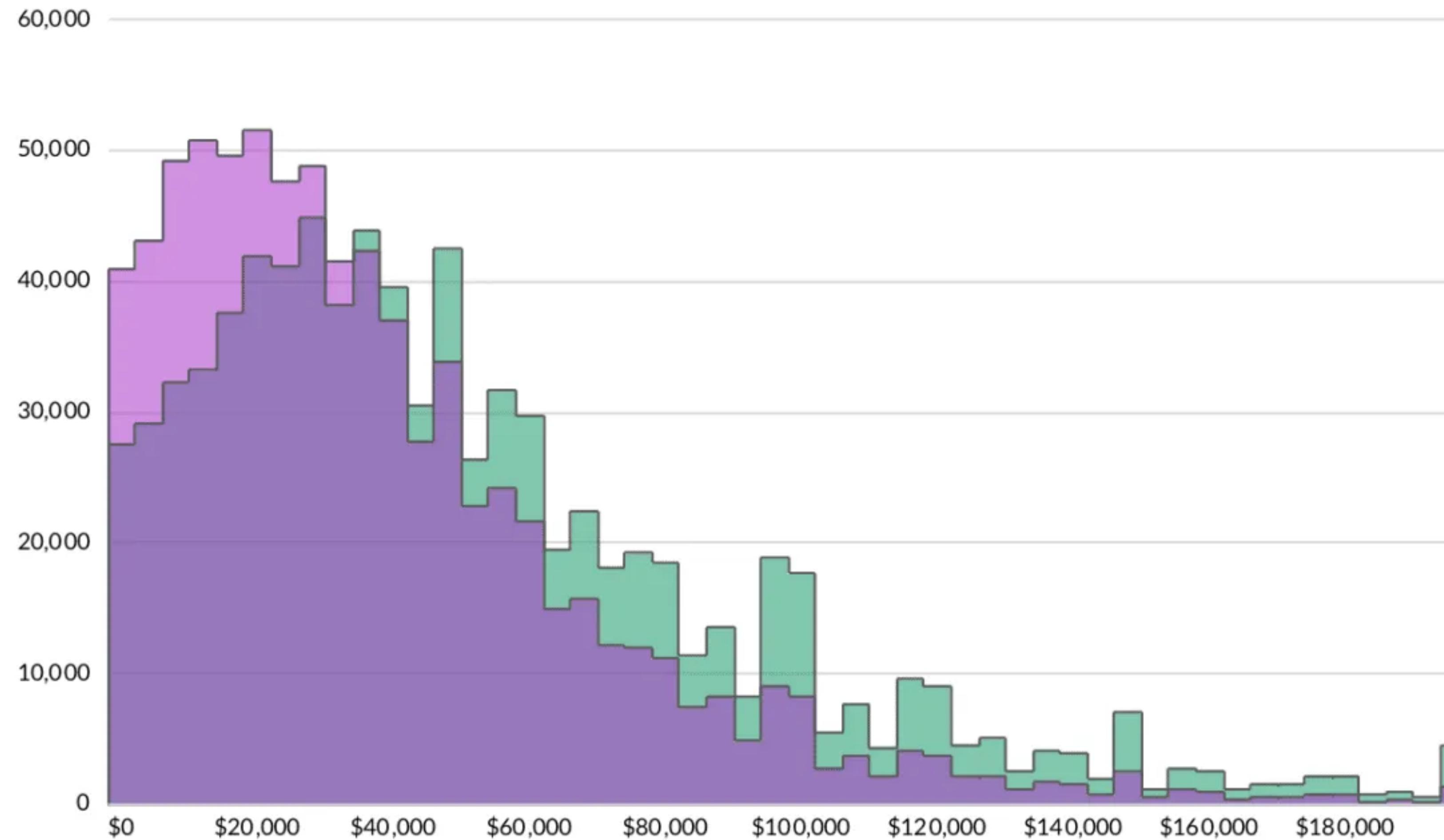
	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

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

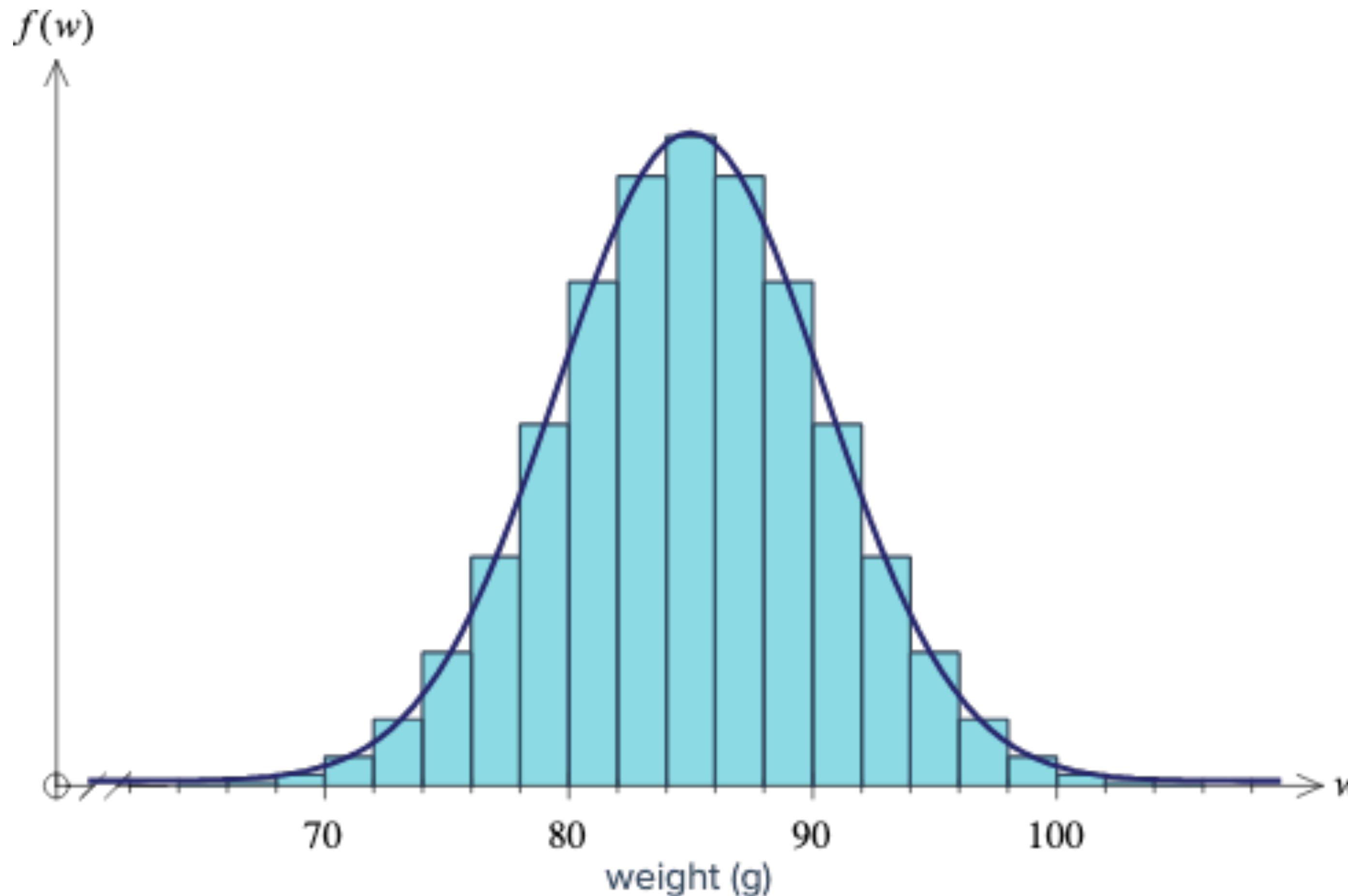
$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}$$

Histogram

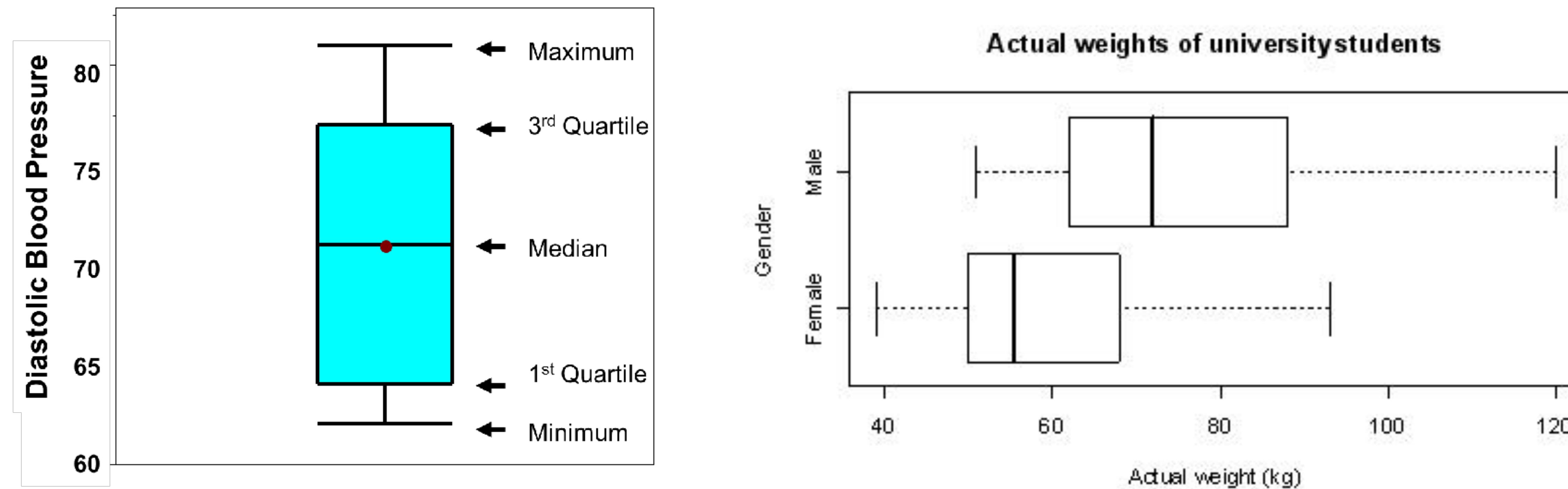
Distribution of Men's and Women's Incomes in 2016



Probability Density Function



Box and Whisker Plot



Looking for data

Where to look for data?

[kaggle.com](https://www.kaggle.com/datasets)

The screenshot shows the Kaggle datasets homepage. At the top, there's a search bar and a 'New Dataset' button. Below that is a section for 'Trending Datasets' featuring four cards: 'Banglore Smart Building', 'US Airline Industry Dataset (1993-2024)', 'Detailed India EV Market Data 2001 - 2024', and 'IMDb Summer Movies Data'. Each card includes a thumbnail, dataset name, author, last update, file count, and file type. Below the trending section are links for 'LLM Fine-Tuning' and 'See All'.

Datasets

Explore, analyze, and share quality data. [Learn more](#) about data types, creating, and collaborating.

+ New Dataset

Search datasets

Filters

All datasets Computer Science Education Classification Computer Vision NLP Data Visualization Pre-Trained Model

Trending Datasets

See All

Banglore Smart Building · Preetham Gouda · Updated a day ago · Usability 10.0 · 75 kB · 1 File (CSV)

US Airline Industry Dataset (1993-2024) · Muhammad Ehsan · Updated 7 days ago · Usability 10.0 · 14 MB · 1 File (CSV)

Detailed India EV Market Data 2001 - 2024 · Sai Raam · Updated 9 days ago · Usability 7.1 · 209 kB · 5 Files (other)

IMDb Summer Movies Data · Umer Haddii · Updated 4 days ago · Usability 10.0 · 41 kB · 2 Files (CSV)

LLM Fine-Tuning

See All

Where to look for data?

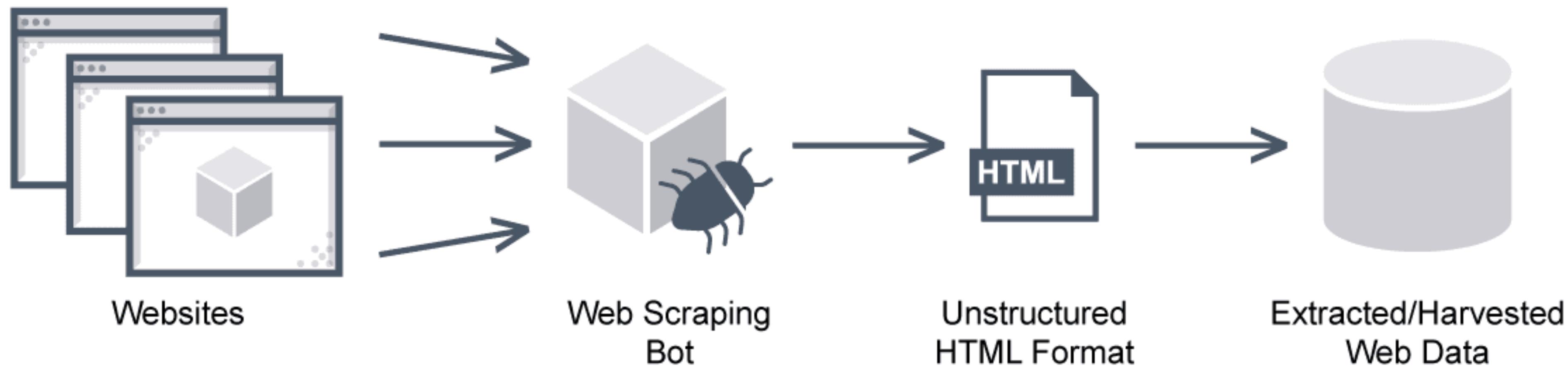
huggingface.co

The screenshot shows the Hugging Face website interface, specifically the 'Datasets' section. The top navigation bar includes links for Models, Datasets, Spaces, Posts, Docs, Solutions, Pricing, Log In, and Sign Up. On the left, there's a sidebar with filters for Main, Tasks, Libraries, Languages, Licenses, Other, Modalities, and Format. Below that are dropdowns for Size (rows) and Format. The main content area displays a grid of dataset cards, each showing the dataset name, a small icon, a 'Viewer' link, the last update date, size, and download count.

Dataset	Owner	Last Updated	Size	Downloads
fka/awesome-chatgpt-prompts	fka	Mar 7, 2023	153	6.12k
nisten/all-human-diseases	nisten	About 2 hours ago	2.2k	29
G-reen/TheatreLM-v2.1-Characters	G-reen	4 days ago	5.01k	7
airtrain-ai/fineweb-edu-fortified	airtrain-ai	11 days ago	322M	673
UCSC-VLAA/MedTrinity-25M	UCSC-VLAA	11 days ago	24.9M	175
princeton-nlp/SWE-bench_Verified	princeton-nlp	5 days ago	500	1.95k
THUDM/LongWriter-6k	THUDM	5 days ago	6k	40
lmms-lab/LLaVA-OneVision-Data	lmms-lab	2 days ago	3.46M	2.16k
BAAI/Infinity-Instruct	BAAI	5 days ago	20.4M	2.56k
multimodalart/1920-raider-waite-tarot-public-domain	multimodalart	5 days ago	78	24

Where to look for data?

Scrape the internet



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