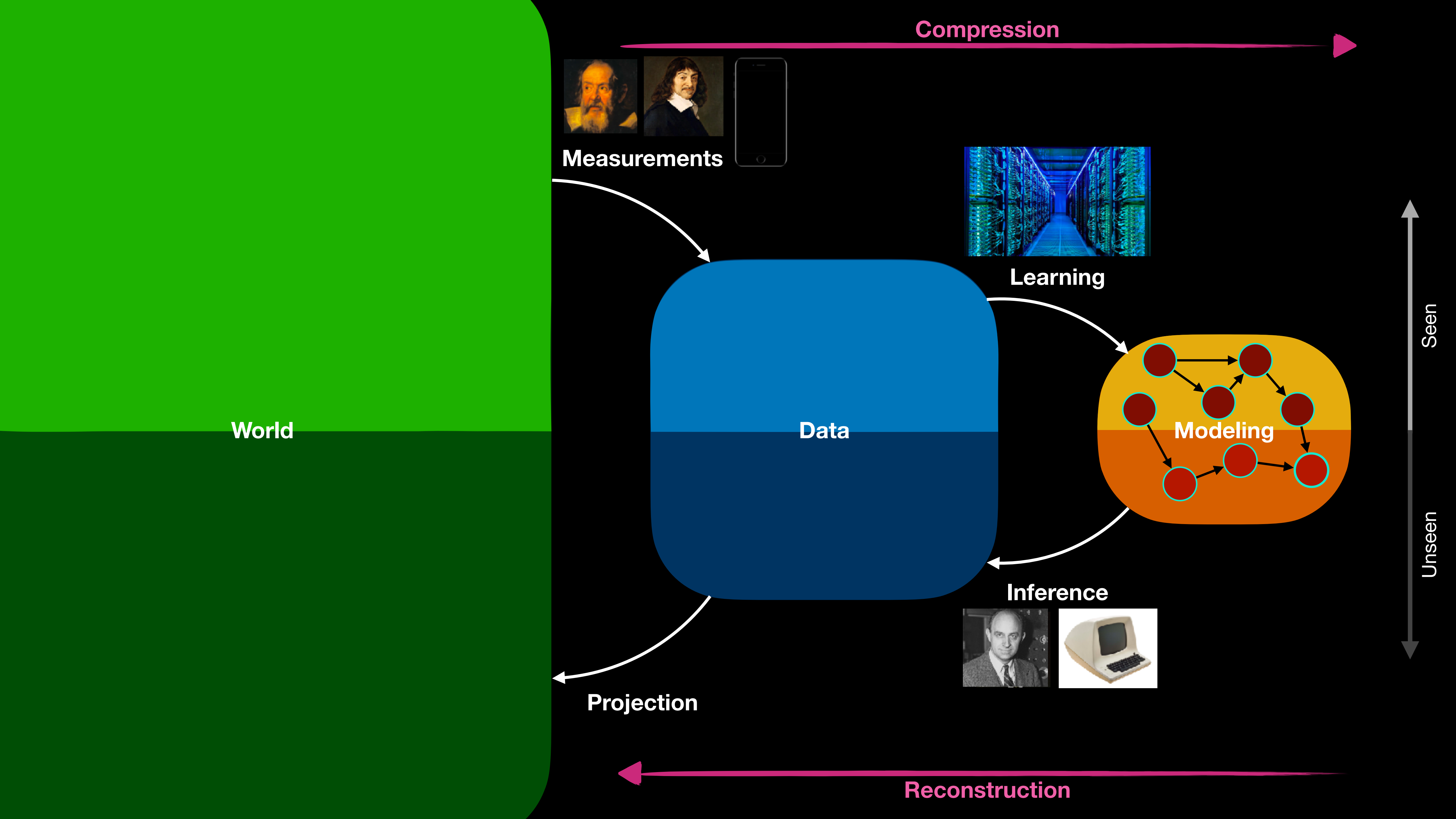


From Data to Learning



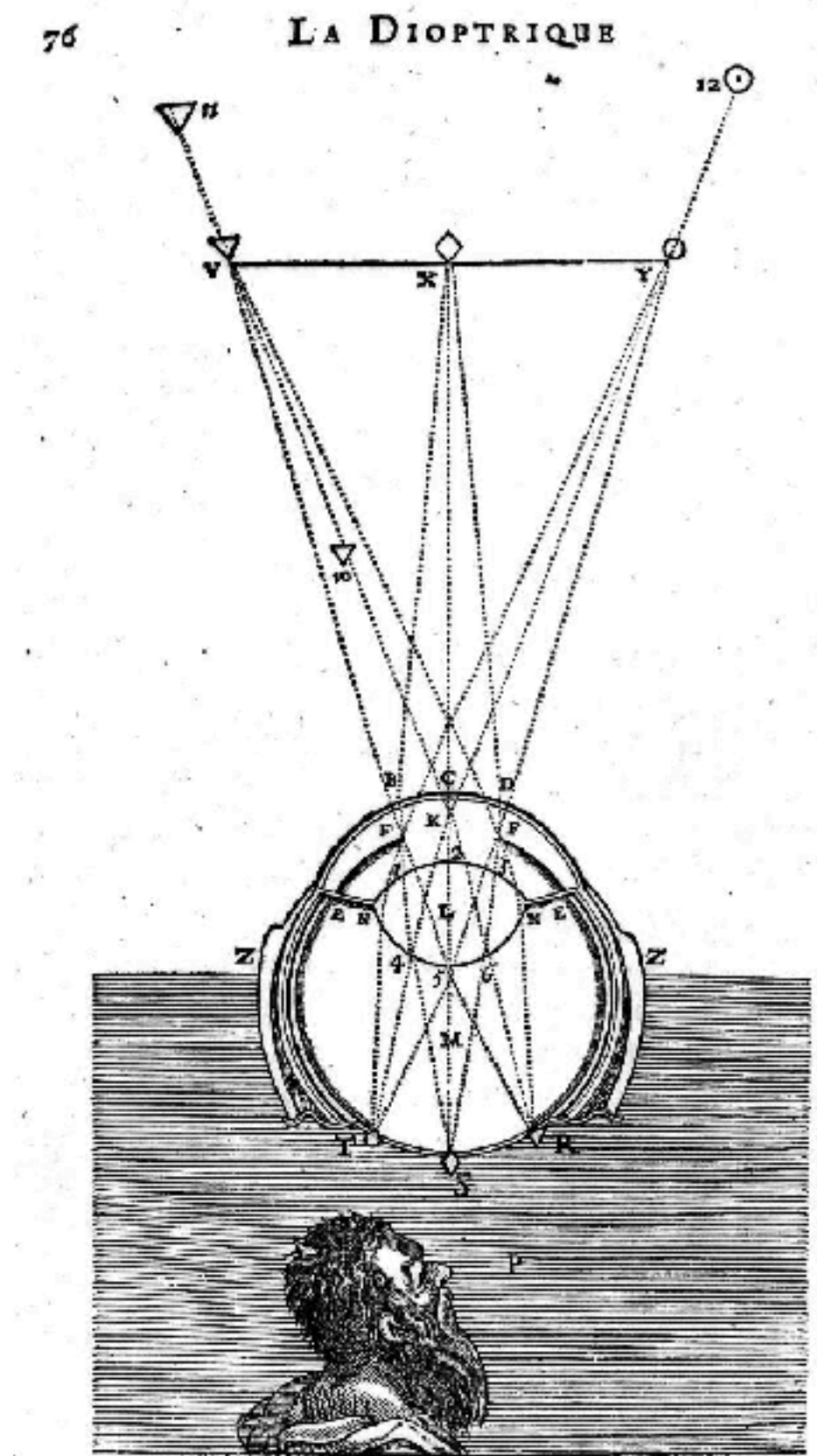
Finding *constants* of nature that *generalize* across space and time



Galileo



Kepler



Descartes

Empirical Laws are linear

Pascal's law (1653)



$$\Delta p = \rho g \Delta h$$

Hooke's law (1678)



$$F = -kx$$

Newton's law of viscosity (1701)



$$\tau = \mu \frac{du}{dy}$$

Ohm's law (1781)



$$I = V/R$$

Fourier's law (1822)



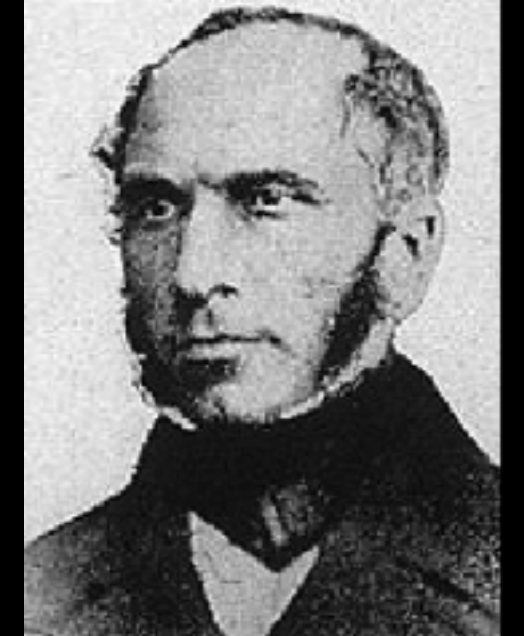
$$q = -k \frac{dT}{dx}$$

Fick's law (1855)



$$J = -D \frac{dC}{dx}$$

Darcy's law (1856)



$$Q = \frac{kA}{\mu L} \Delta p$$

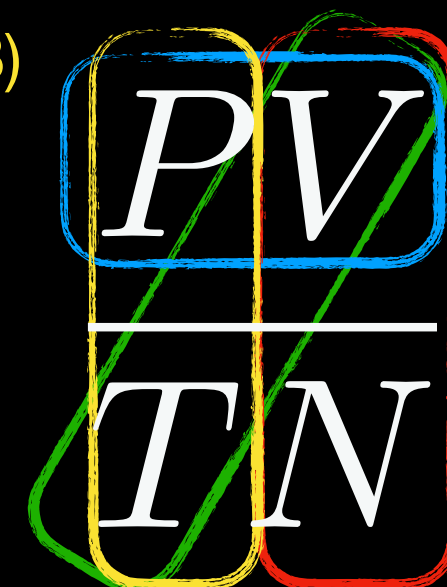
Ideal gas law (1834)

Amontons's law (1808)

Boyle's law (1662)

Charles's law (1787)

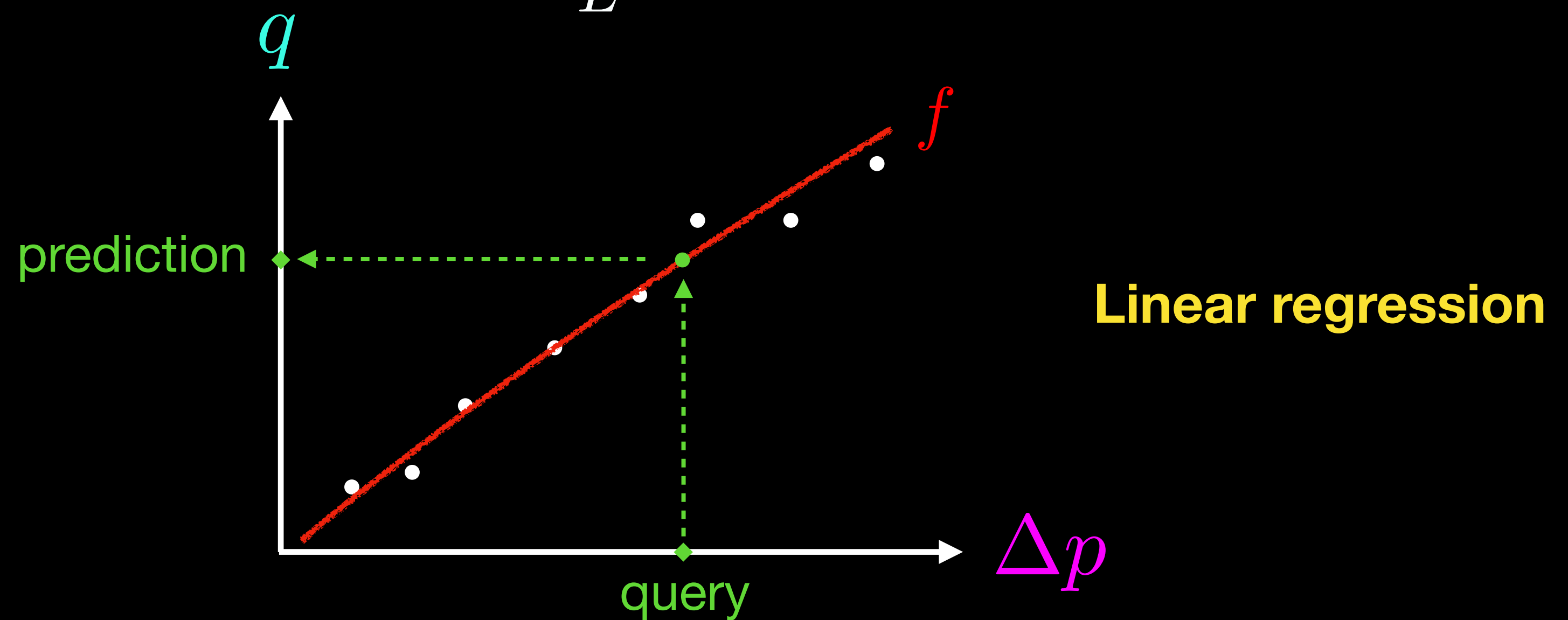
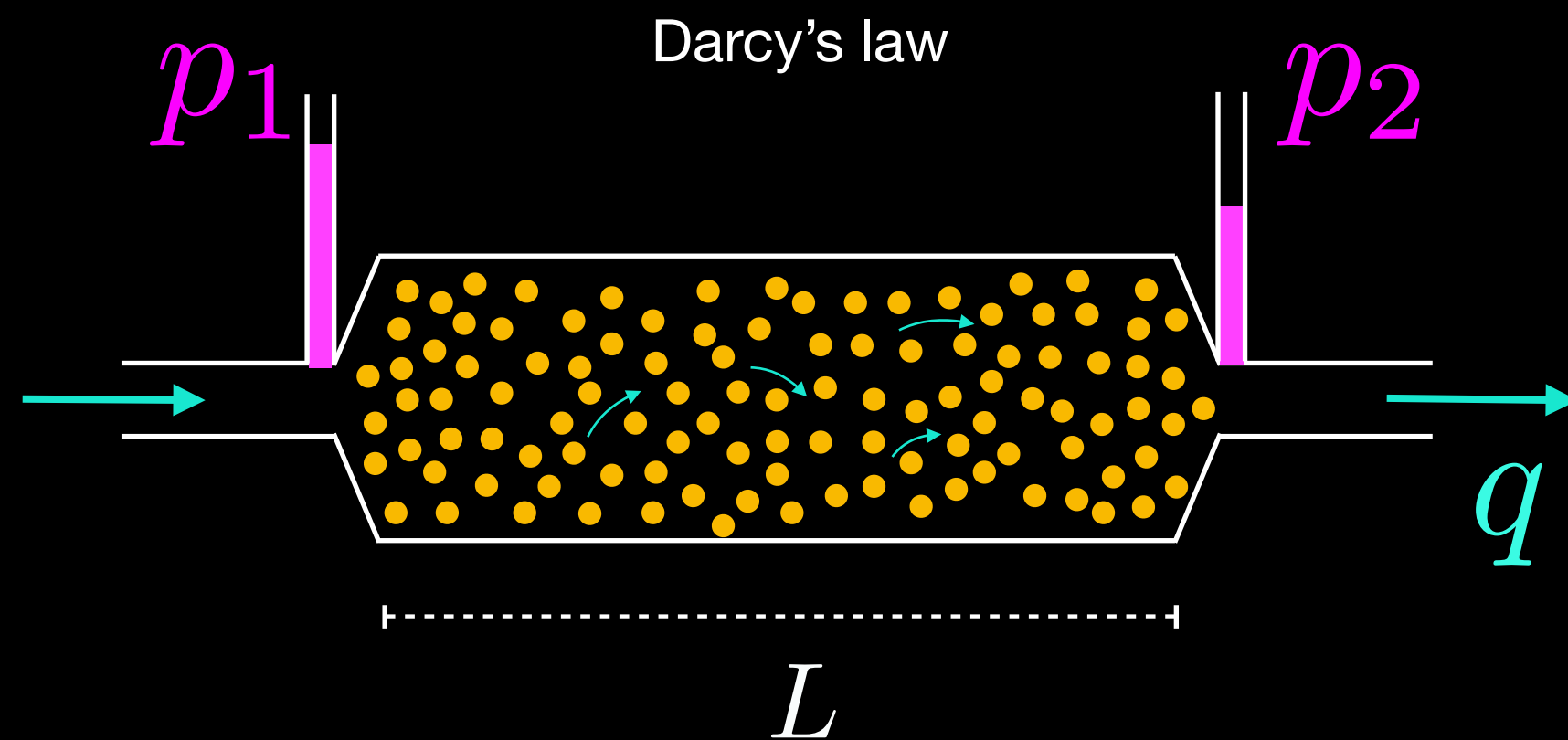
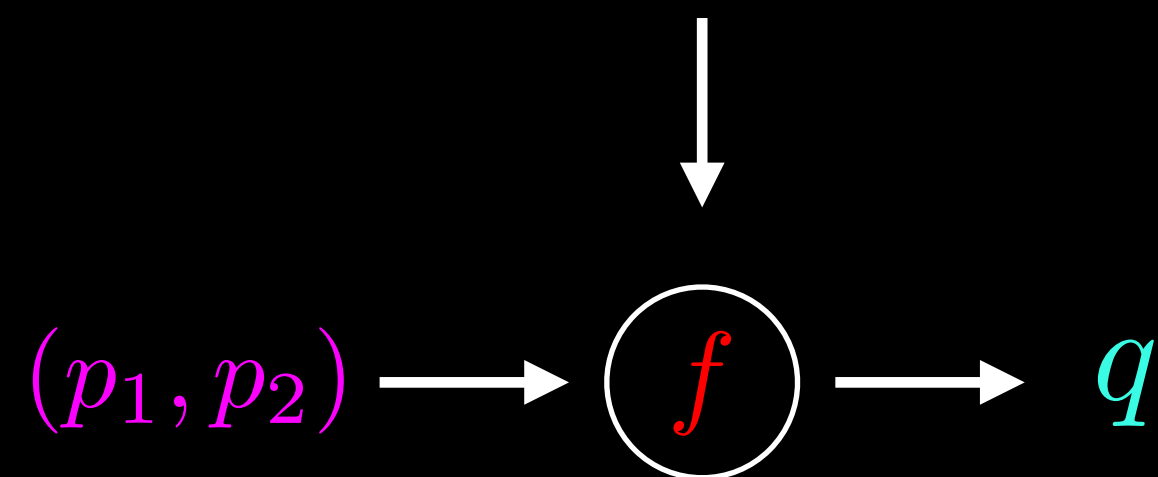
Avogadro's law (1811)



$$= k_B$$

From Experiment to Law

p_1	p_2	q
1.3	1.0	22
1.6	1.5	23
3.4	2.4	46
4.8	3.5	67
6.7	4.5	83
...		
2.3	1.4	?



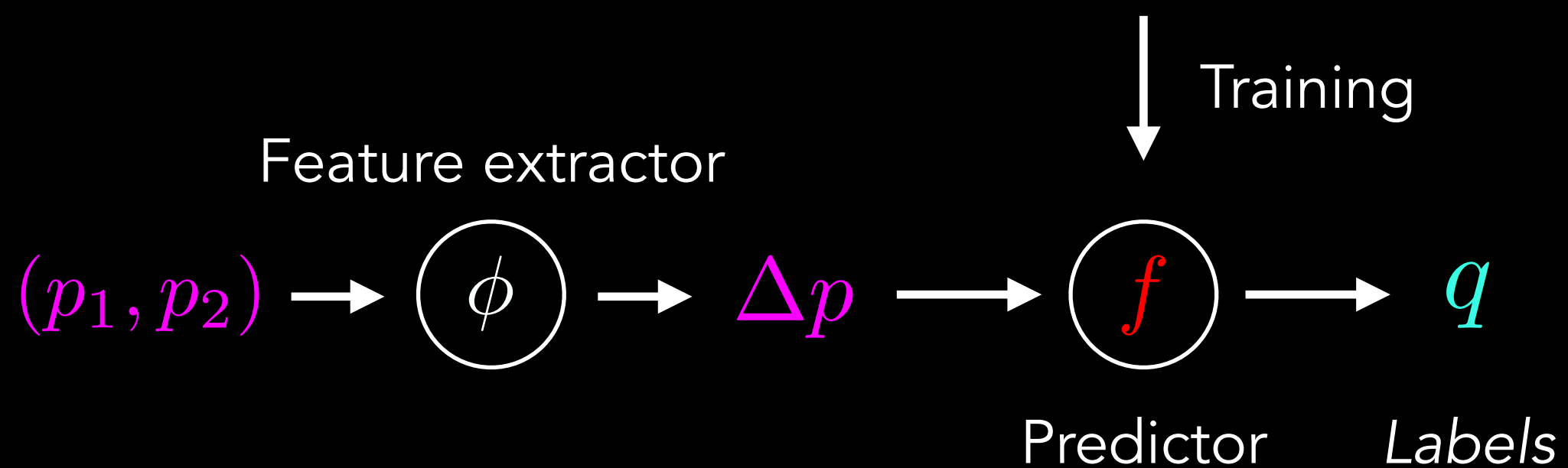
Machine or human learning

Training data: $\mathcal{D}_{\text{train}}$

Example →

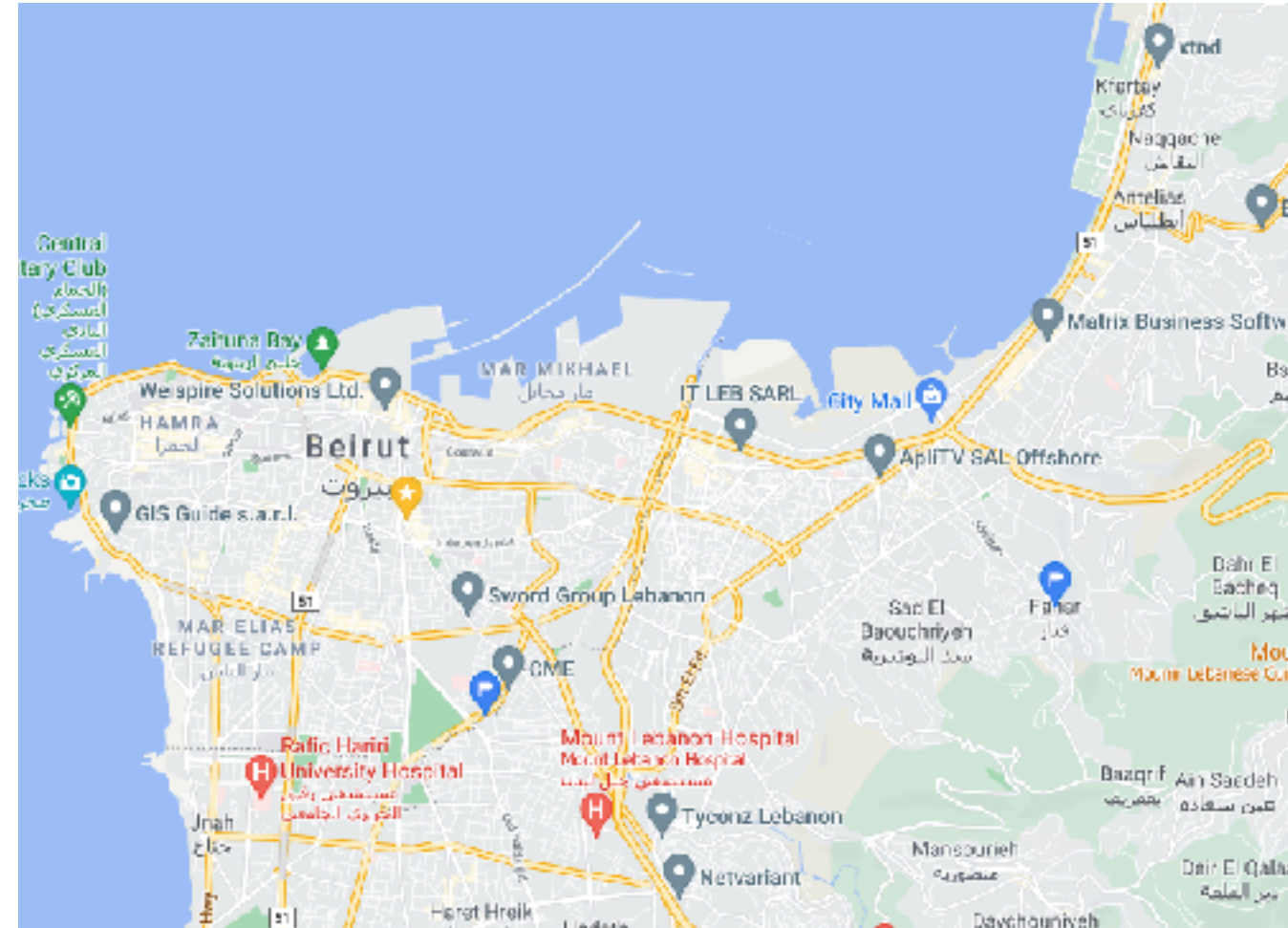
p_1	p_2	q
1.3	1.0	22
1.6	1.5	23
3.4	2.4	46
4.8	3.5	67
6.7	4.5	83
...		

- Which predictors are possible?
- How good is the predictor?
- How can we find the best predictor?



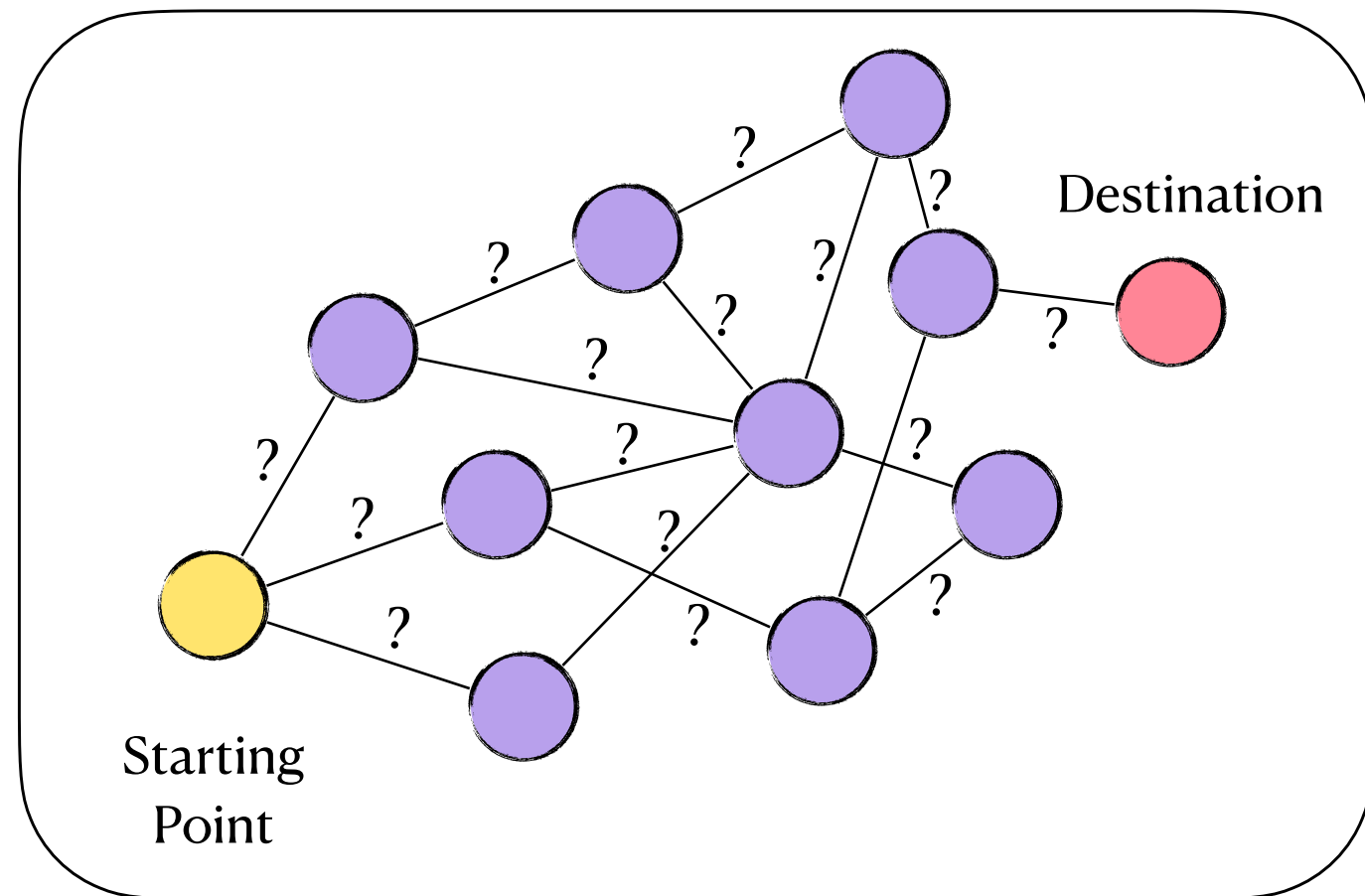
The 3 pillars of Artificial Intelligence

Real world

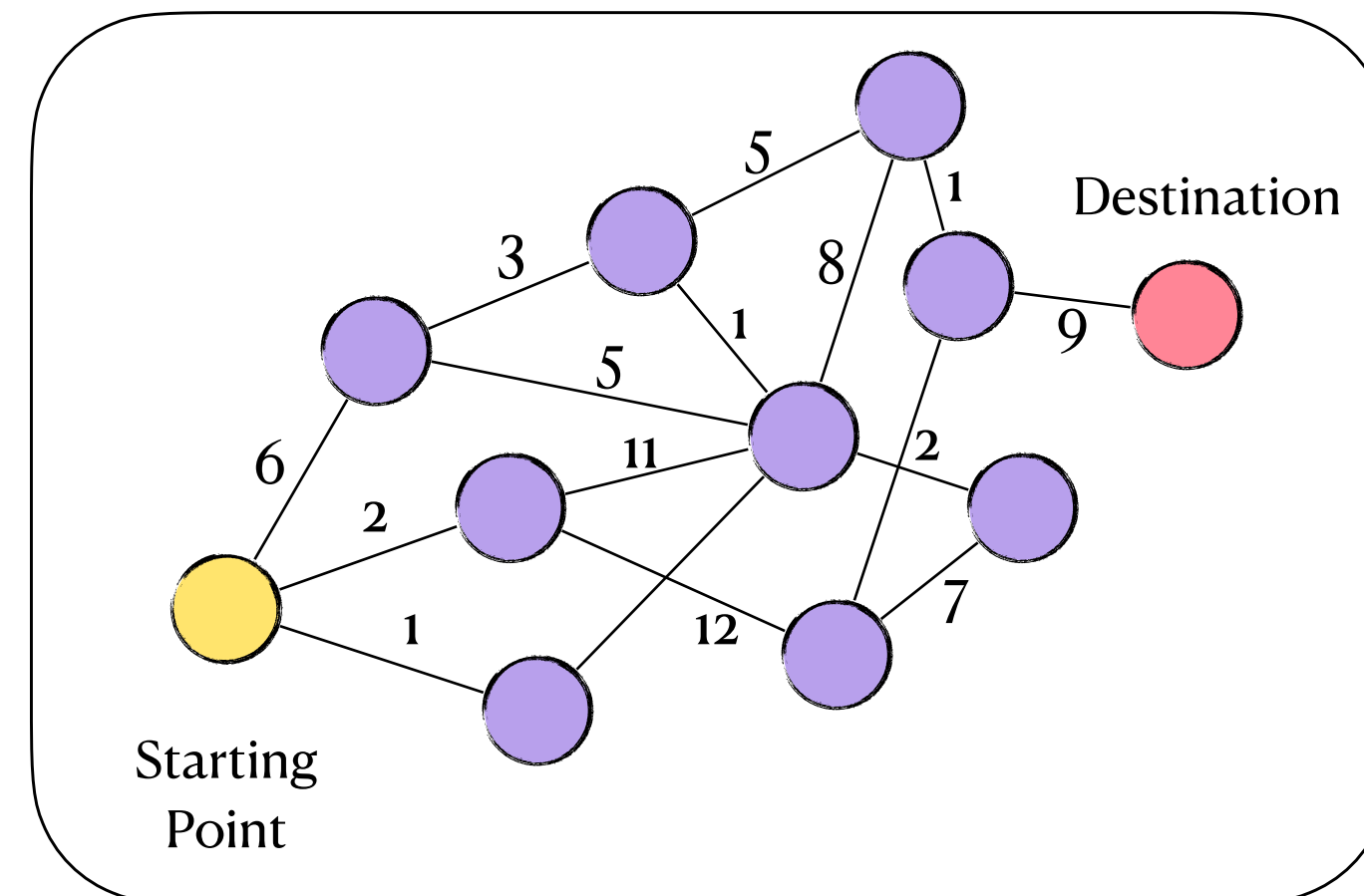


What's the shortest path from AUB to my home?

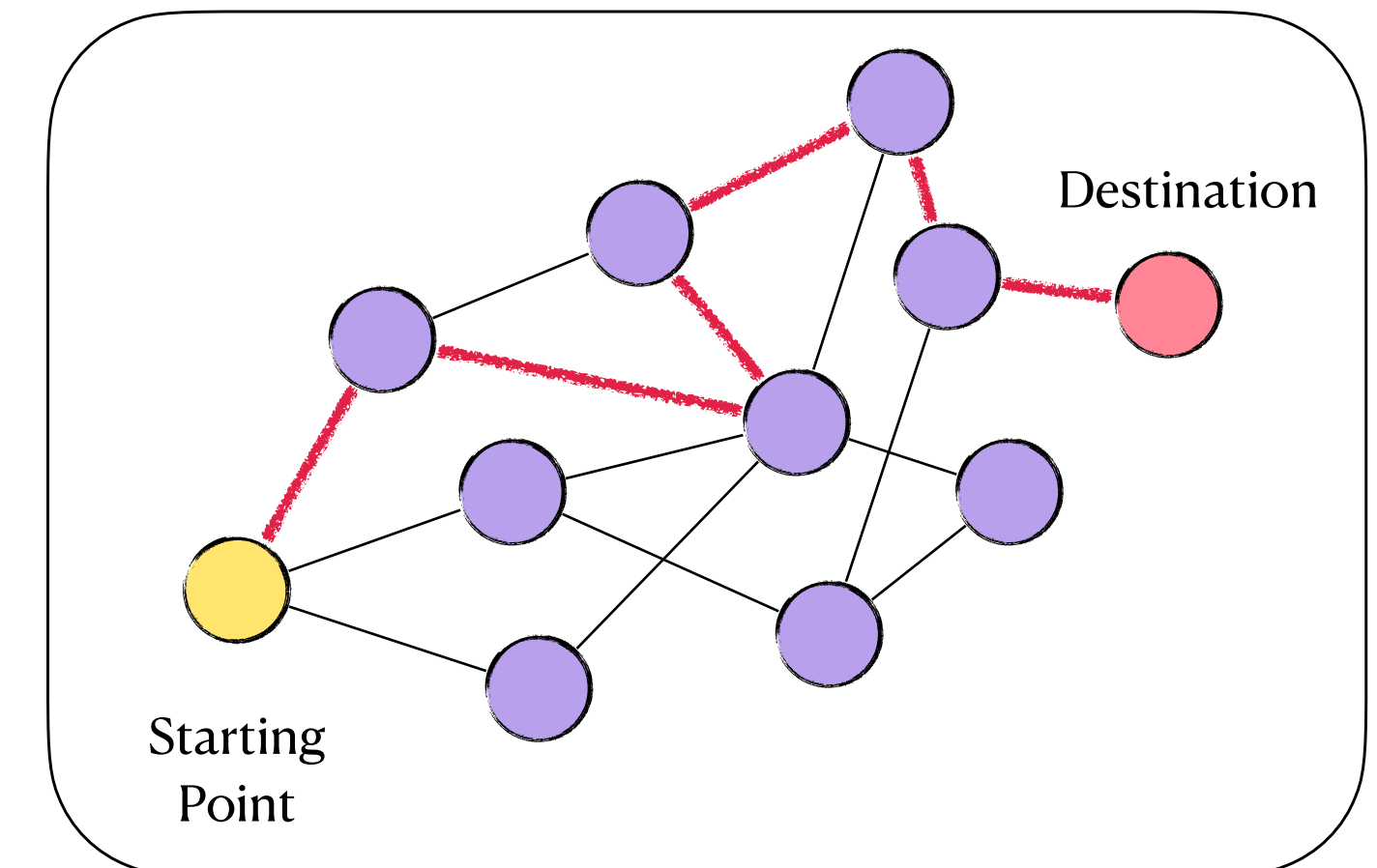
Learning



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

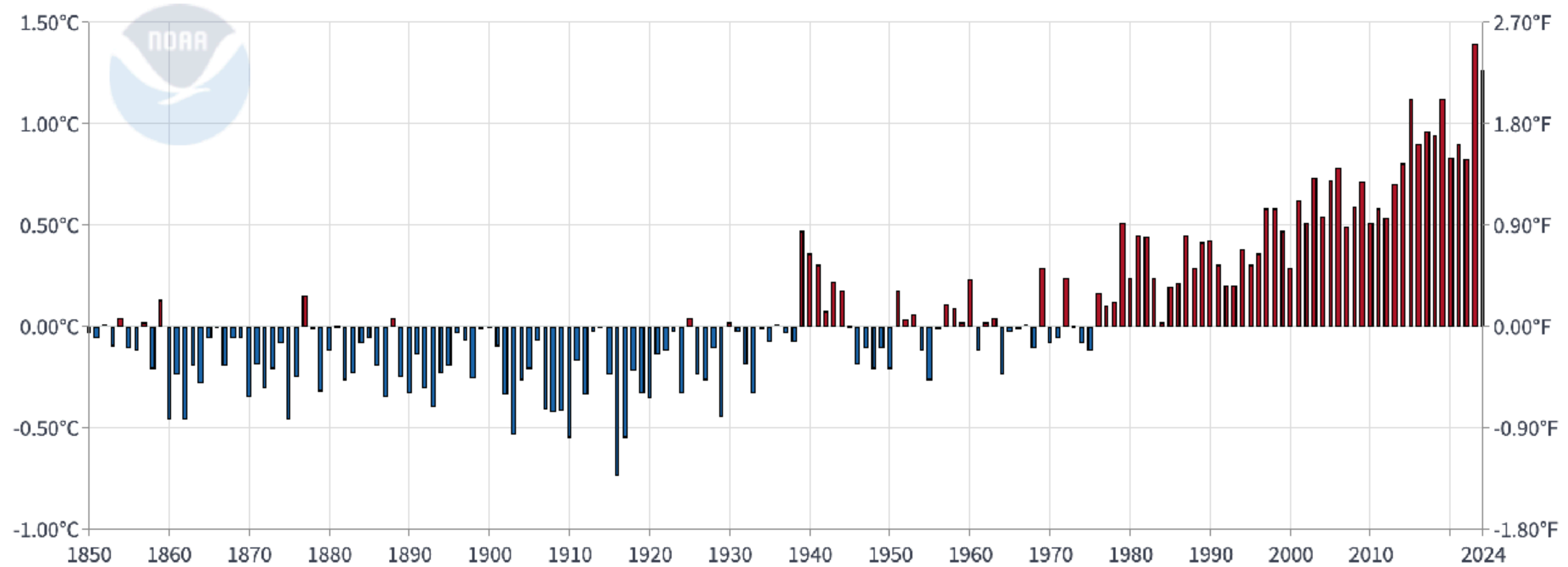
$$\frac{\partial u}{\partial t} = \mathcal{L}(u(x, t); p)$$

What are* Data?

Stocks

Global Land and Ocean Average Temperature Anomalies

December



Powered by ZingChart

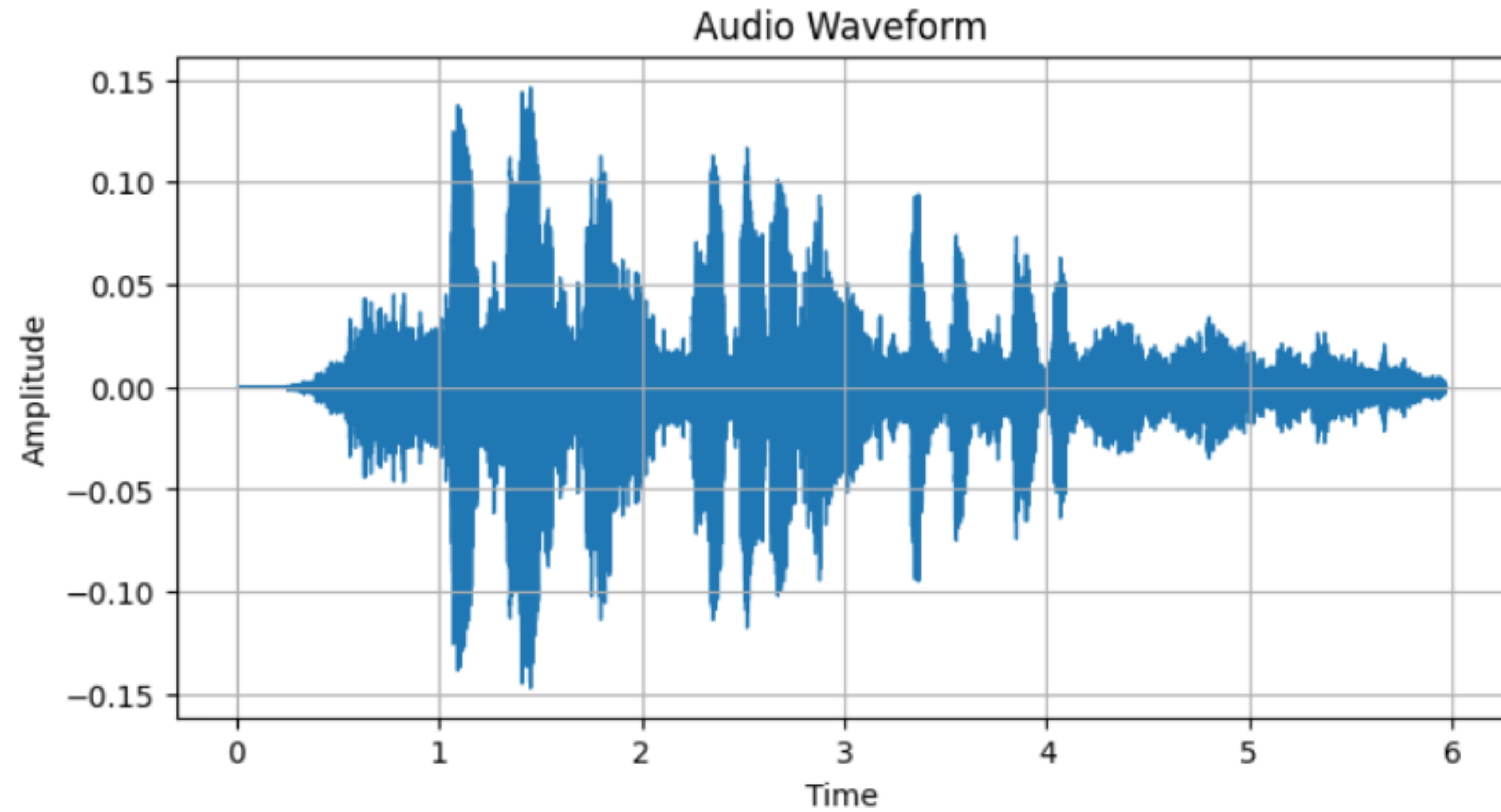
Time

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

Images

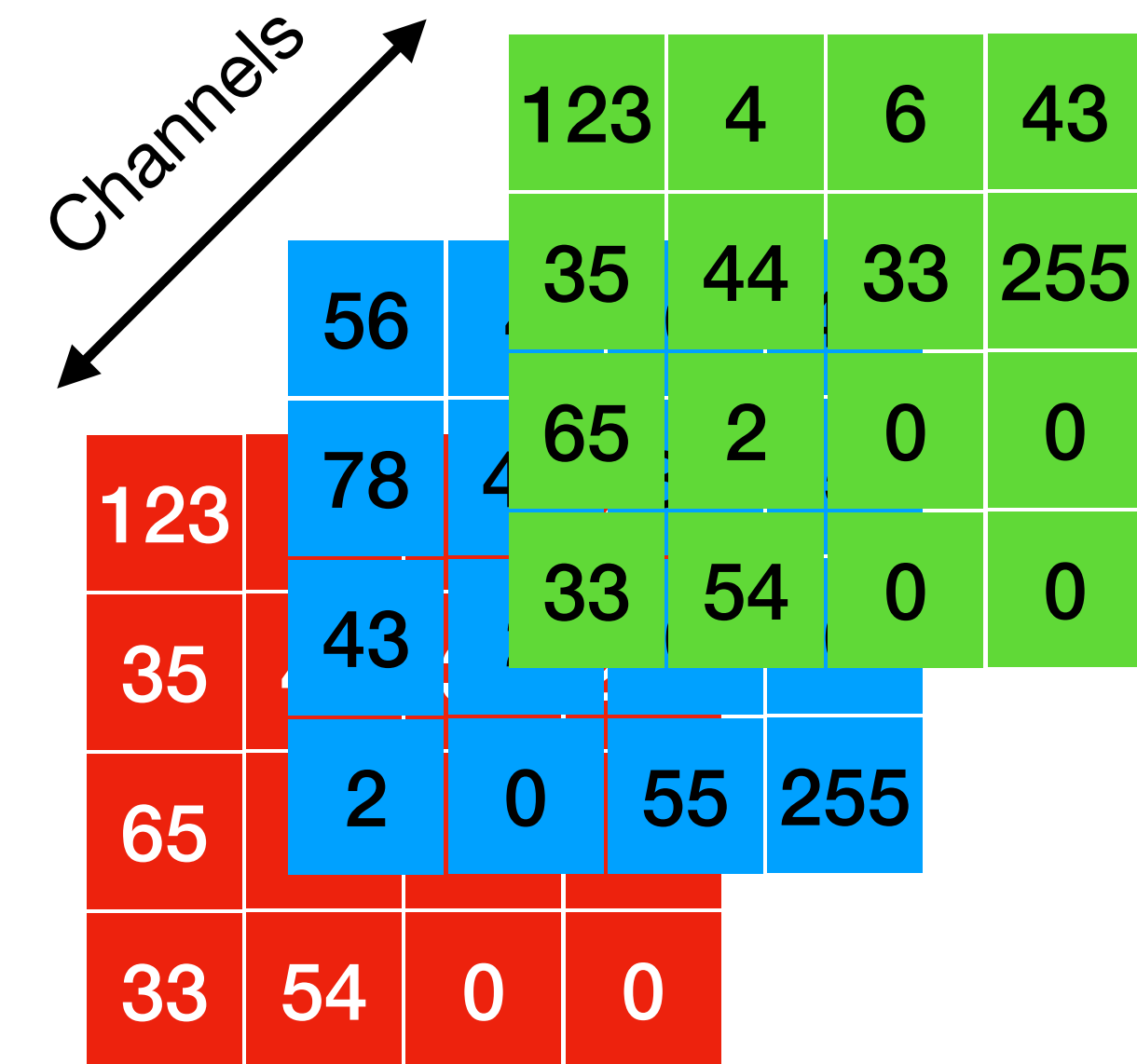
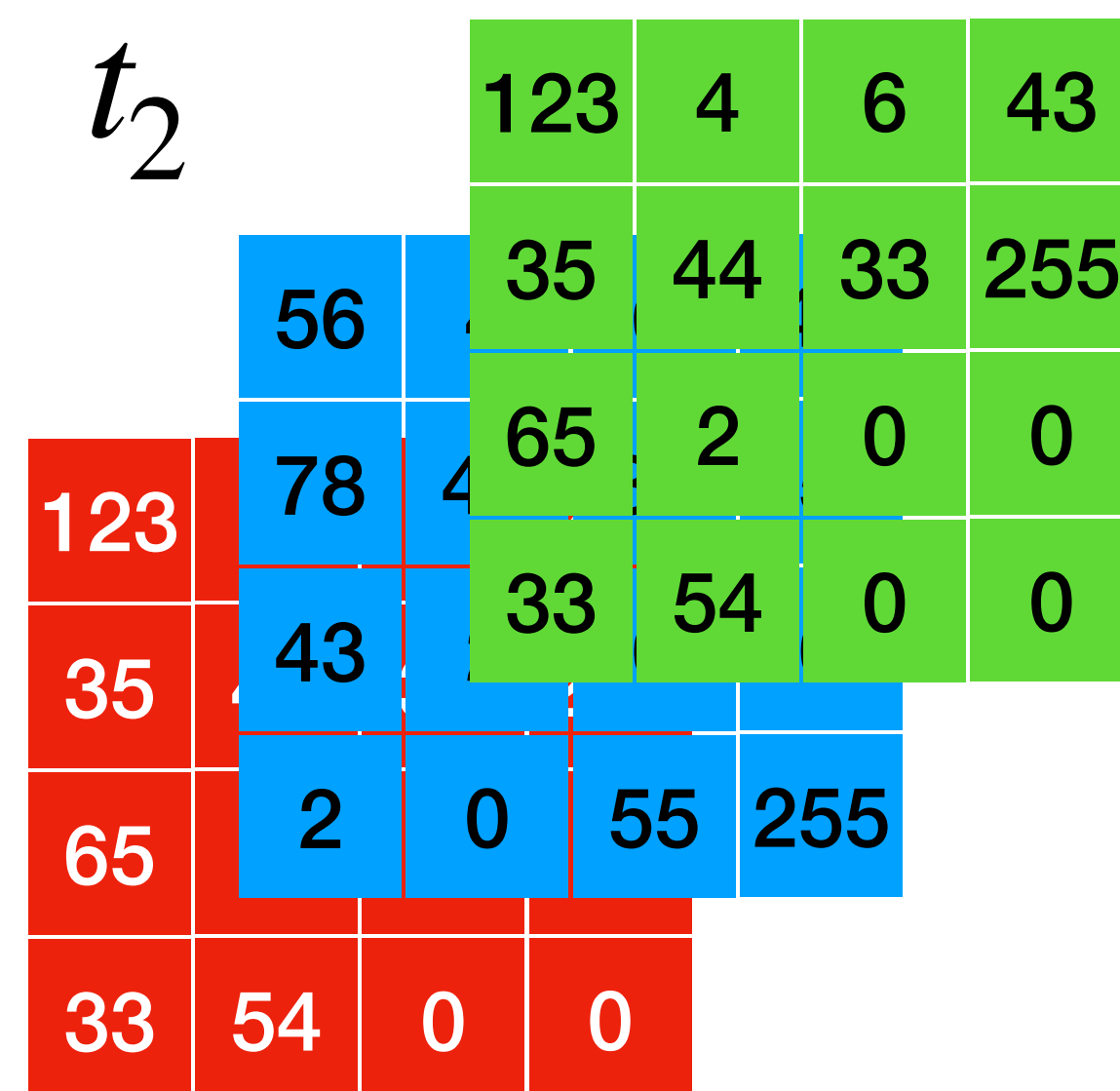
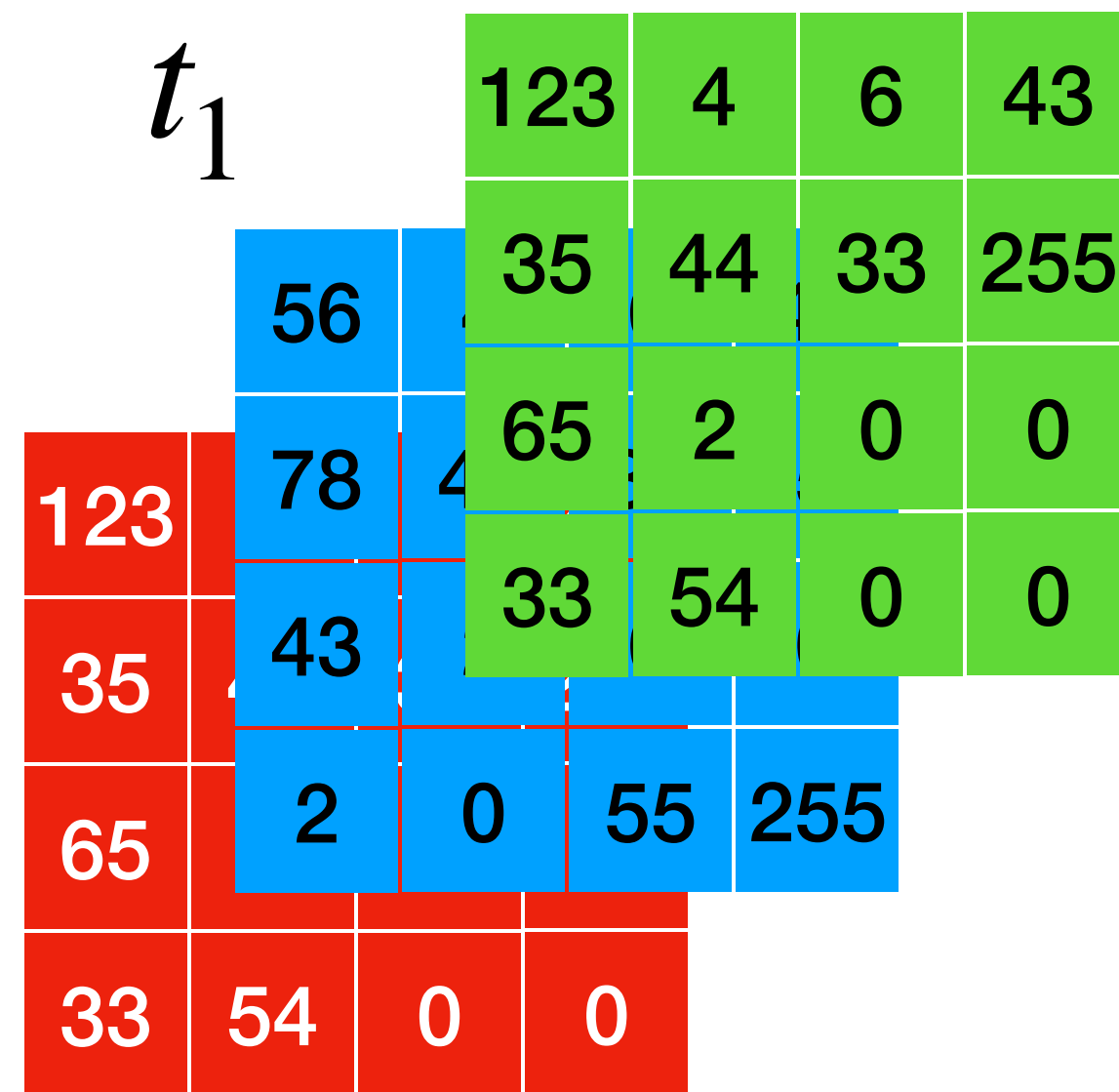


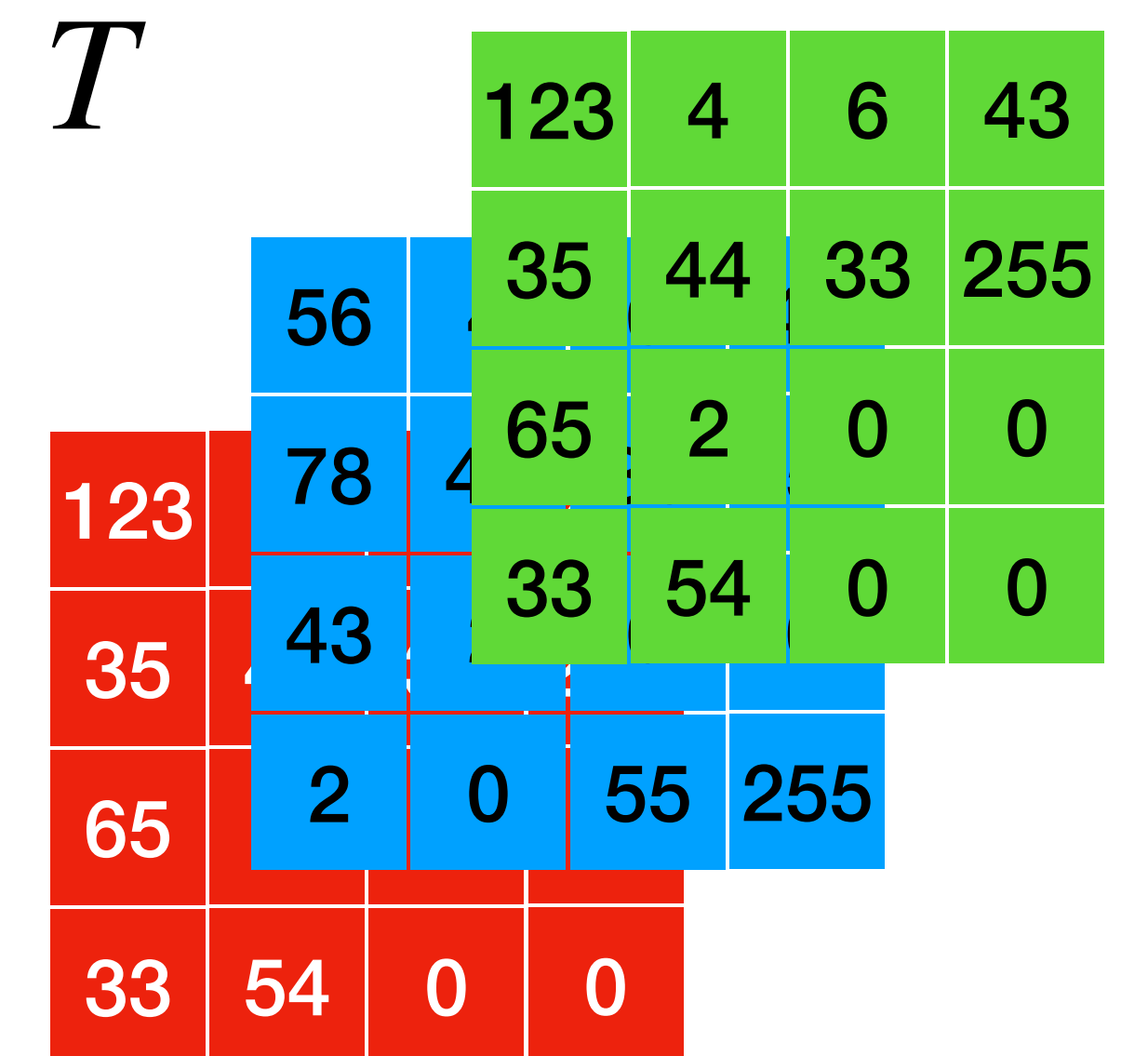
Image dimension: $h \times w \times 3$

Video

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

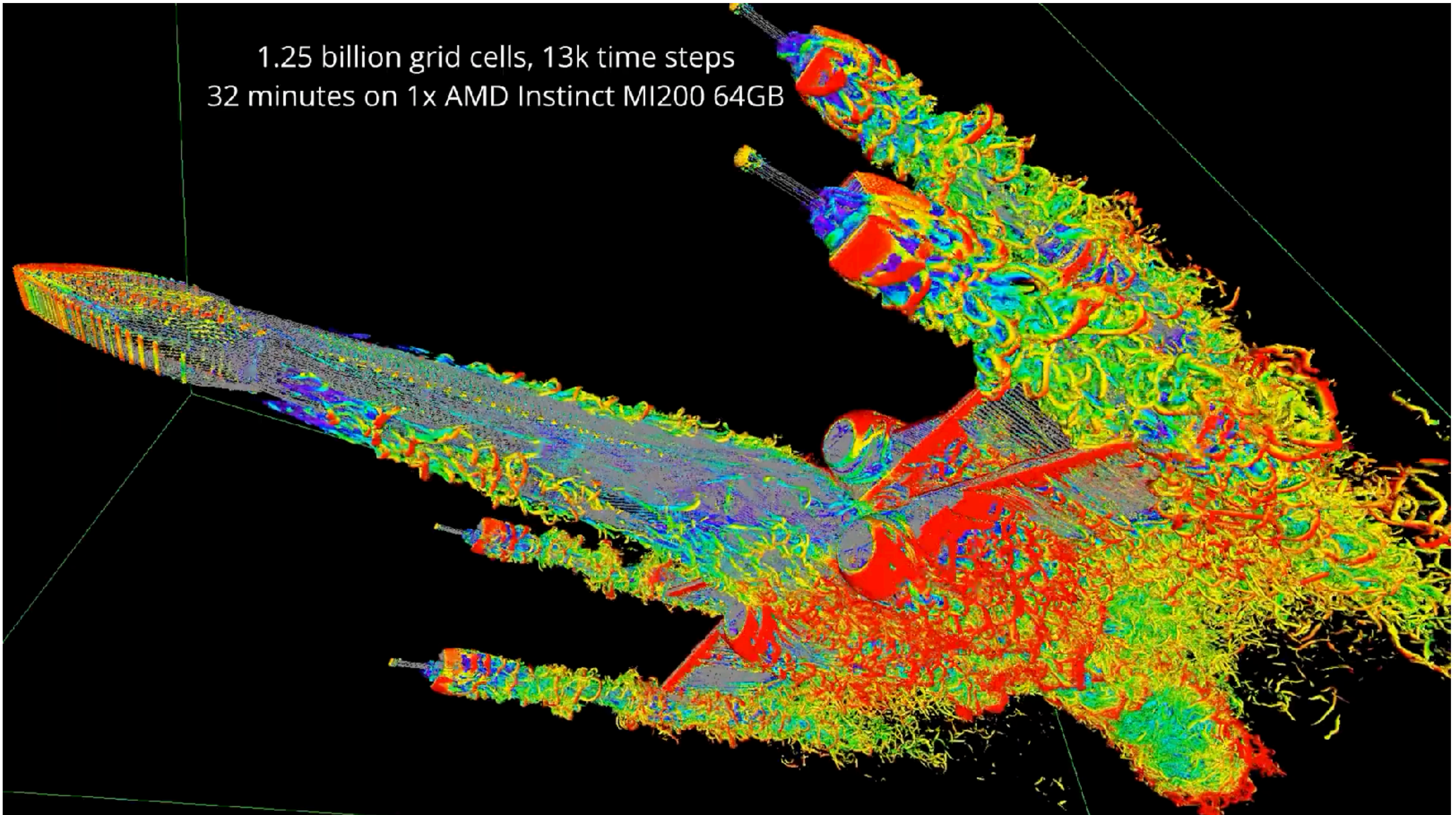


...



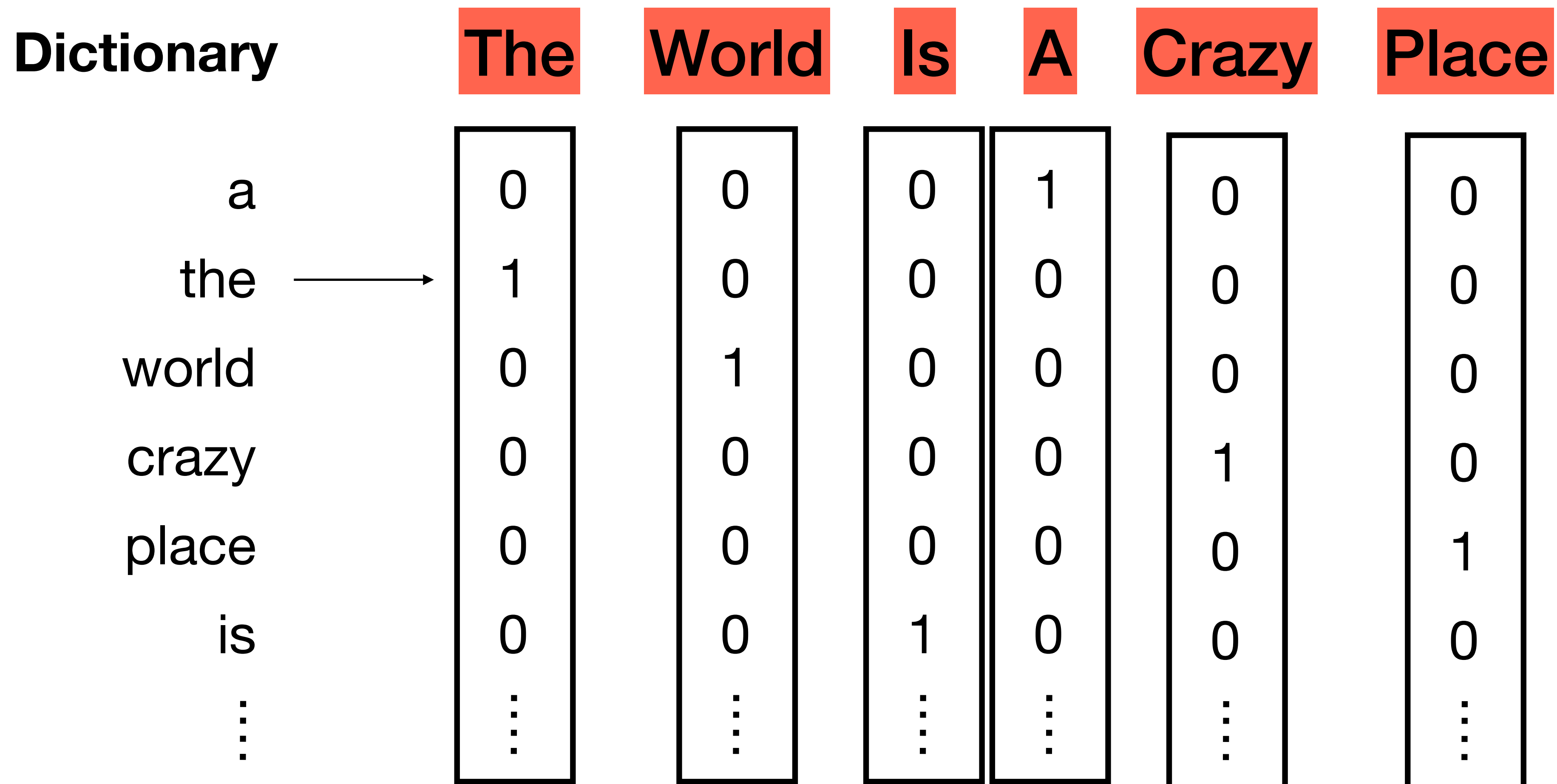
Fluid dynamics

1.25 billion grid cells, 13k time steps
32 minutes on 1x AMD Instinct MI200 64GB



Text

One-hot word representation



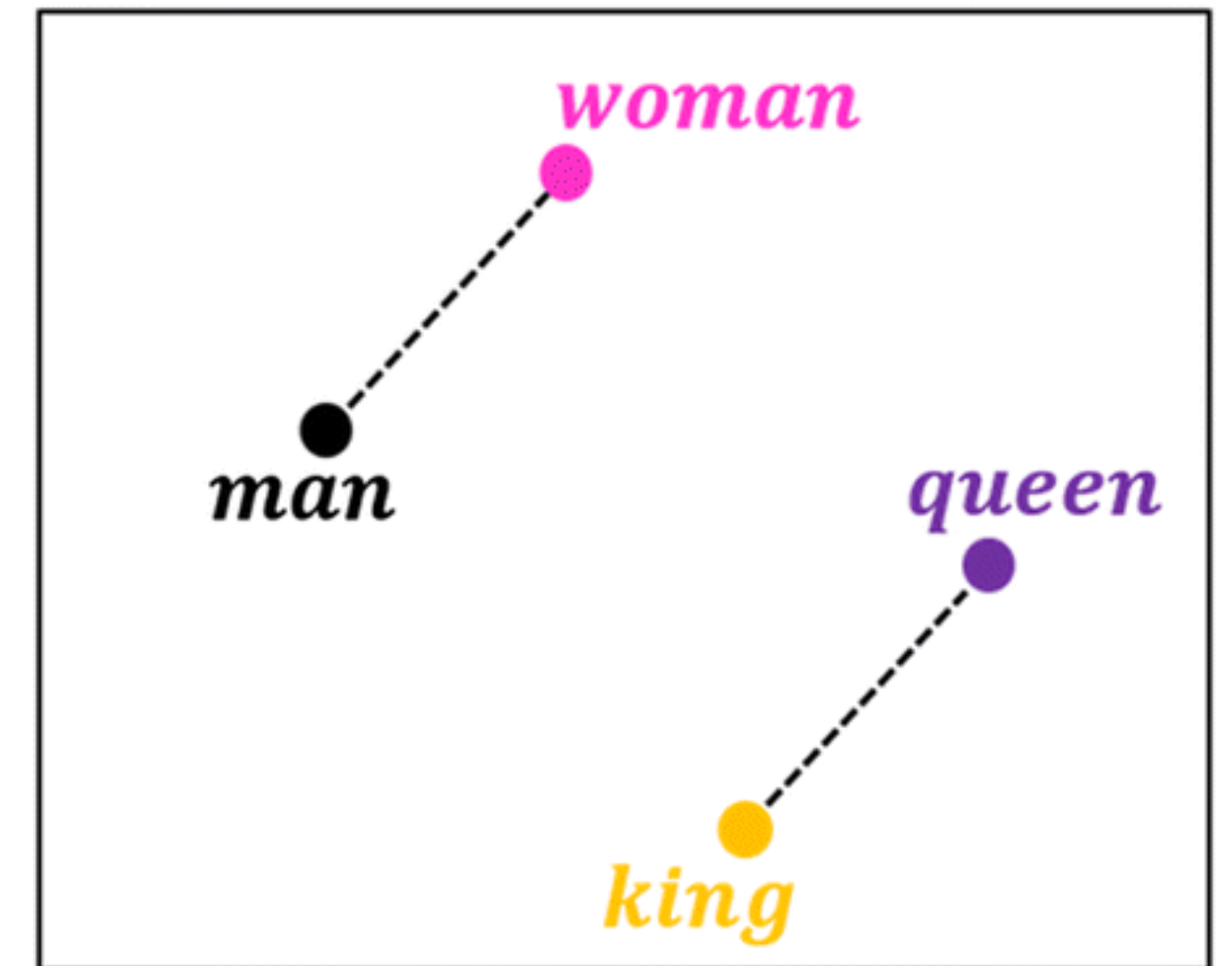
Text

Word Embedding Representation

<i>man</i>	→	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
<i>woman</i>	→	0.7	0.3	0.9	-0.7	0.1	-0.5	-0.4
<i>king</i>	→	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
<i>queen</i>	→	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

Words

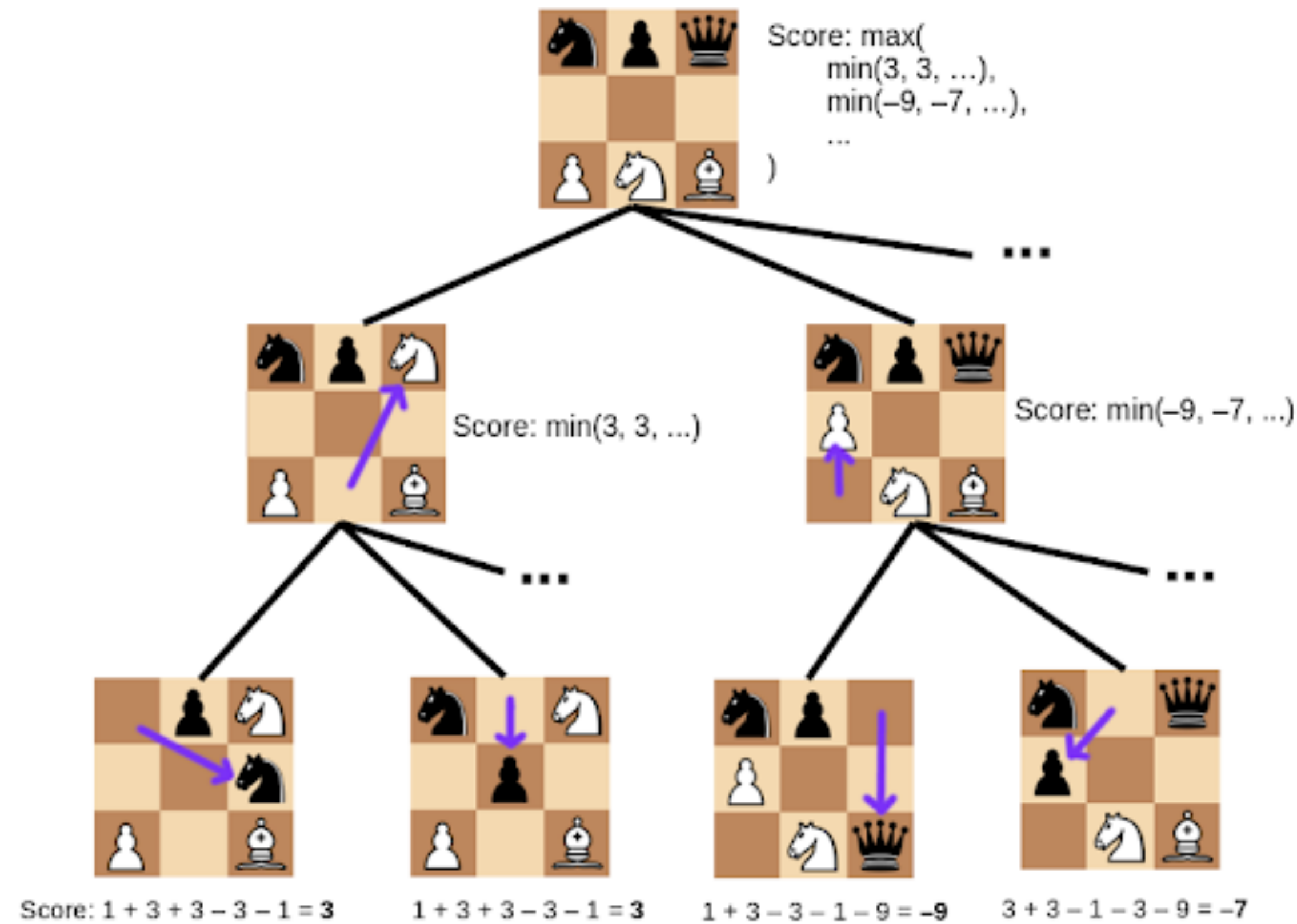
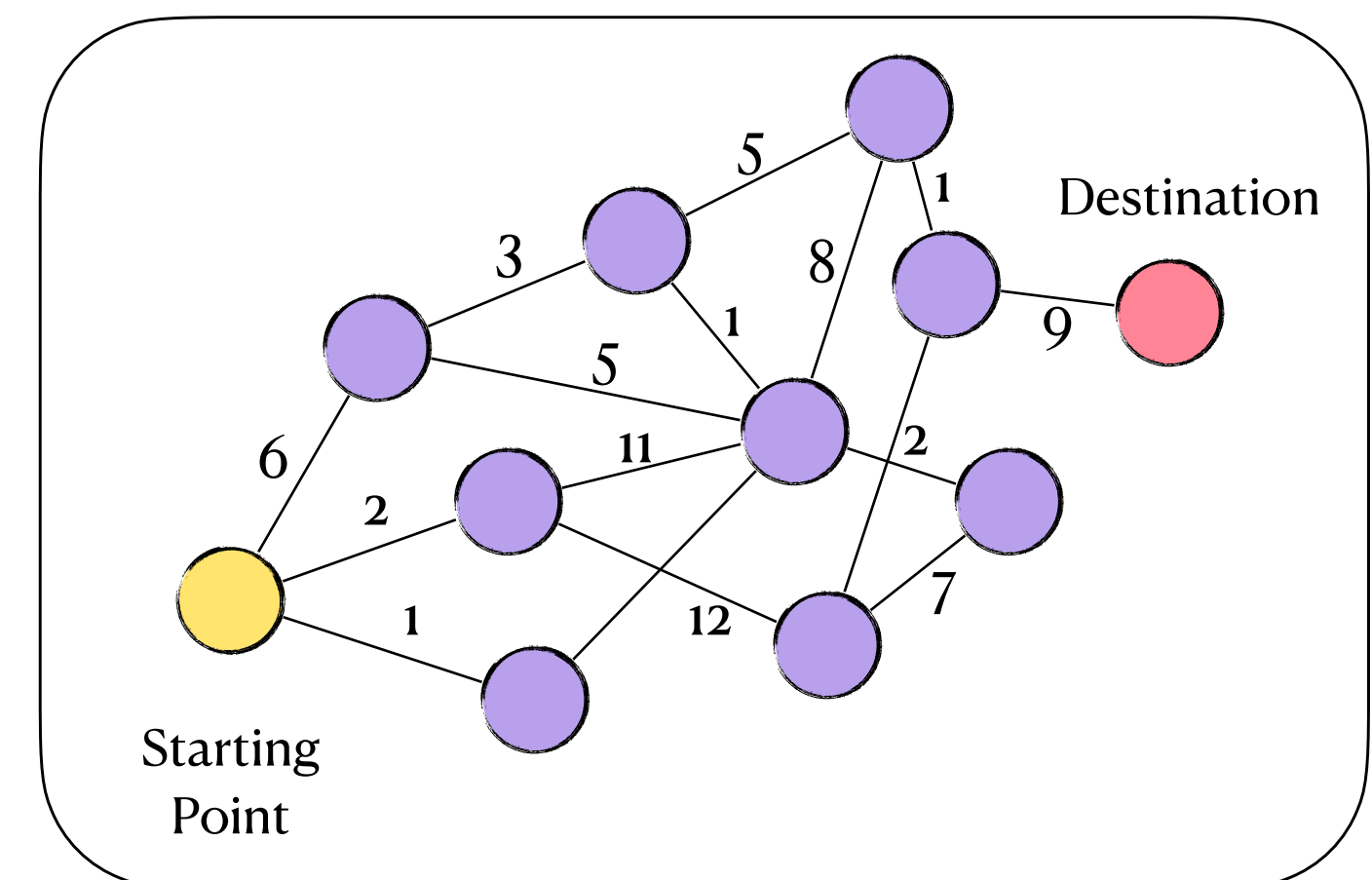
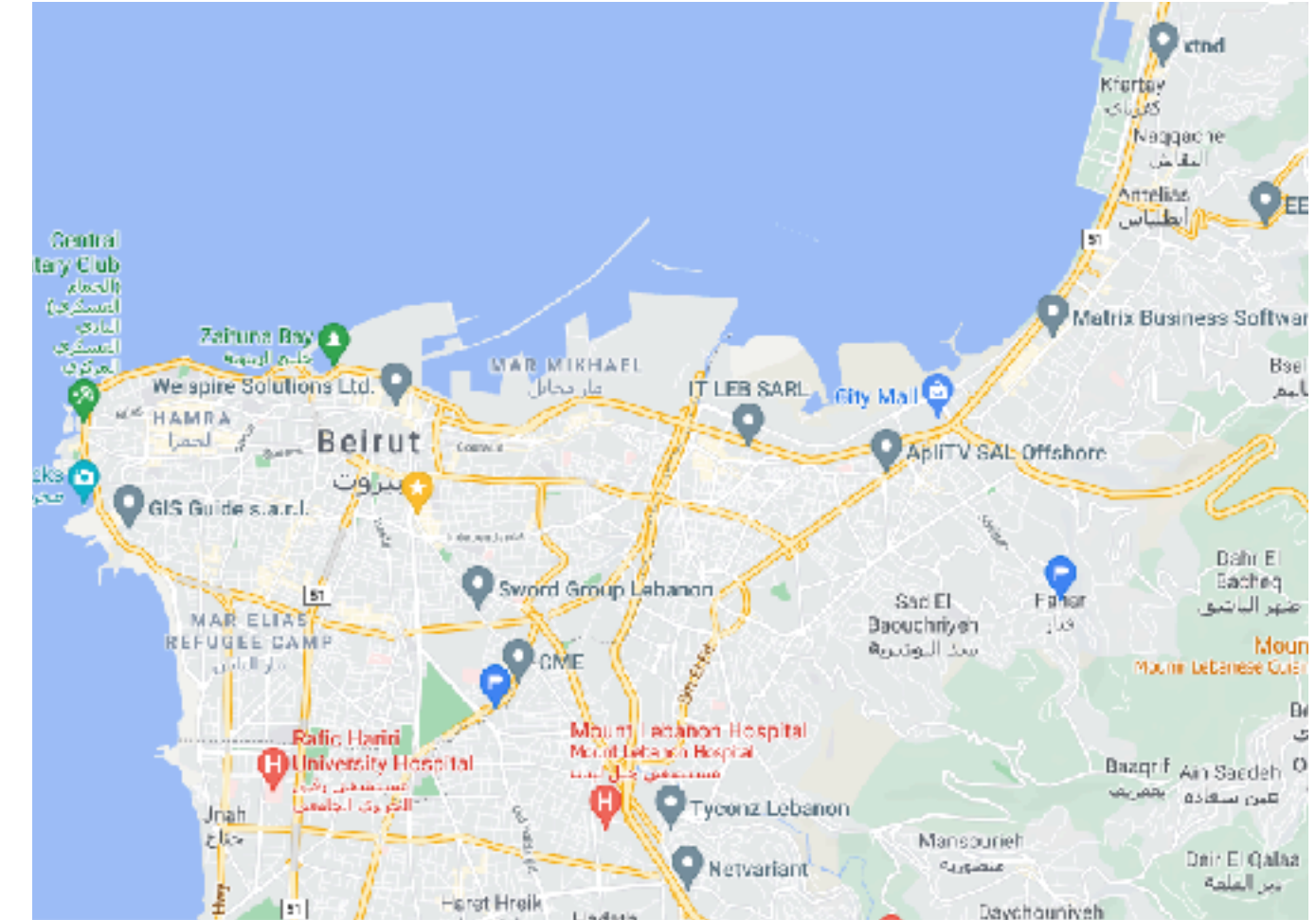
Embedding



Semantic Space

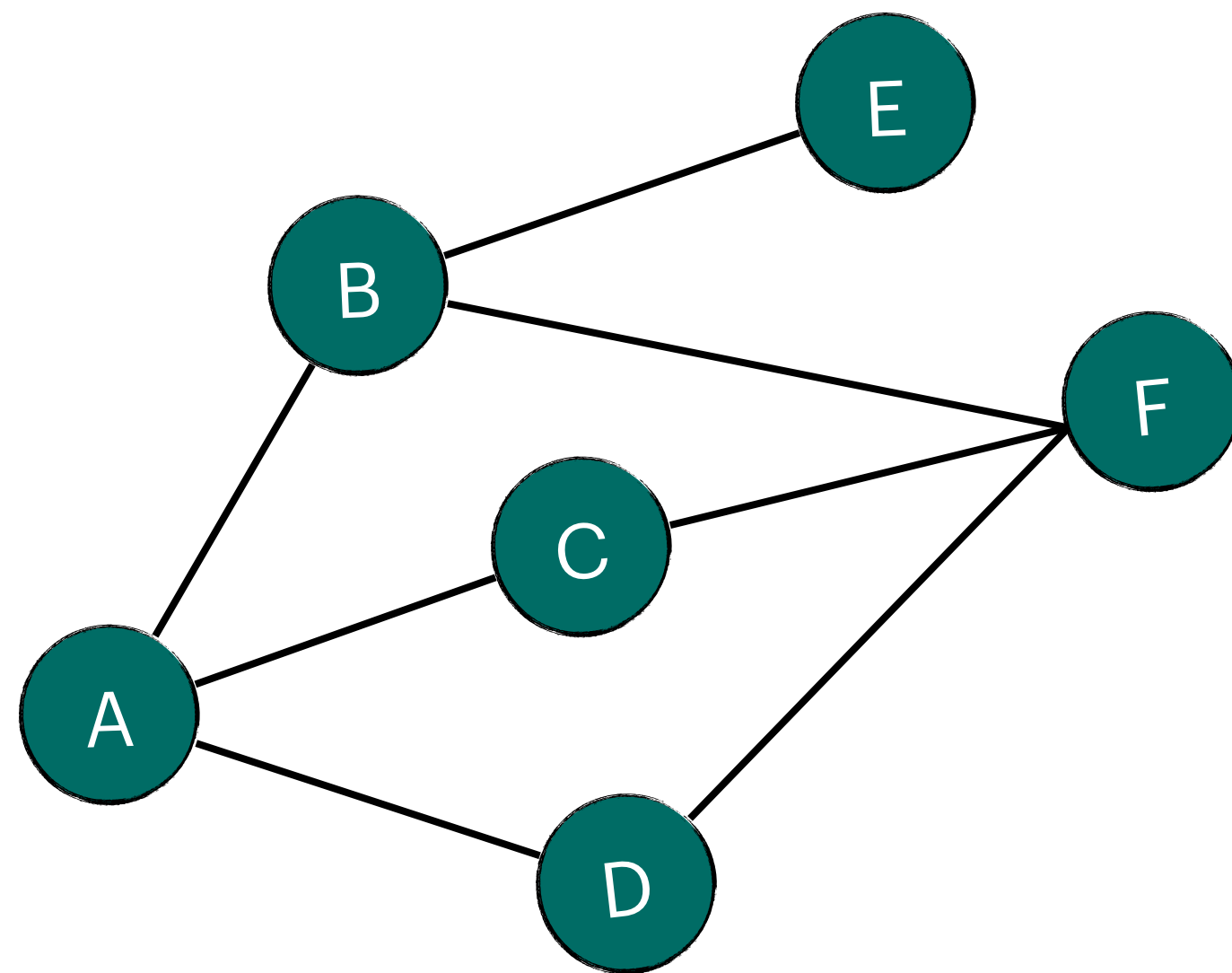
Graph Representation

Real world



Graph Representation

Graph with Vertices and Edges

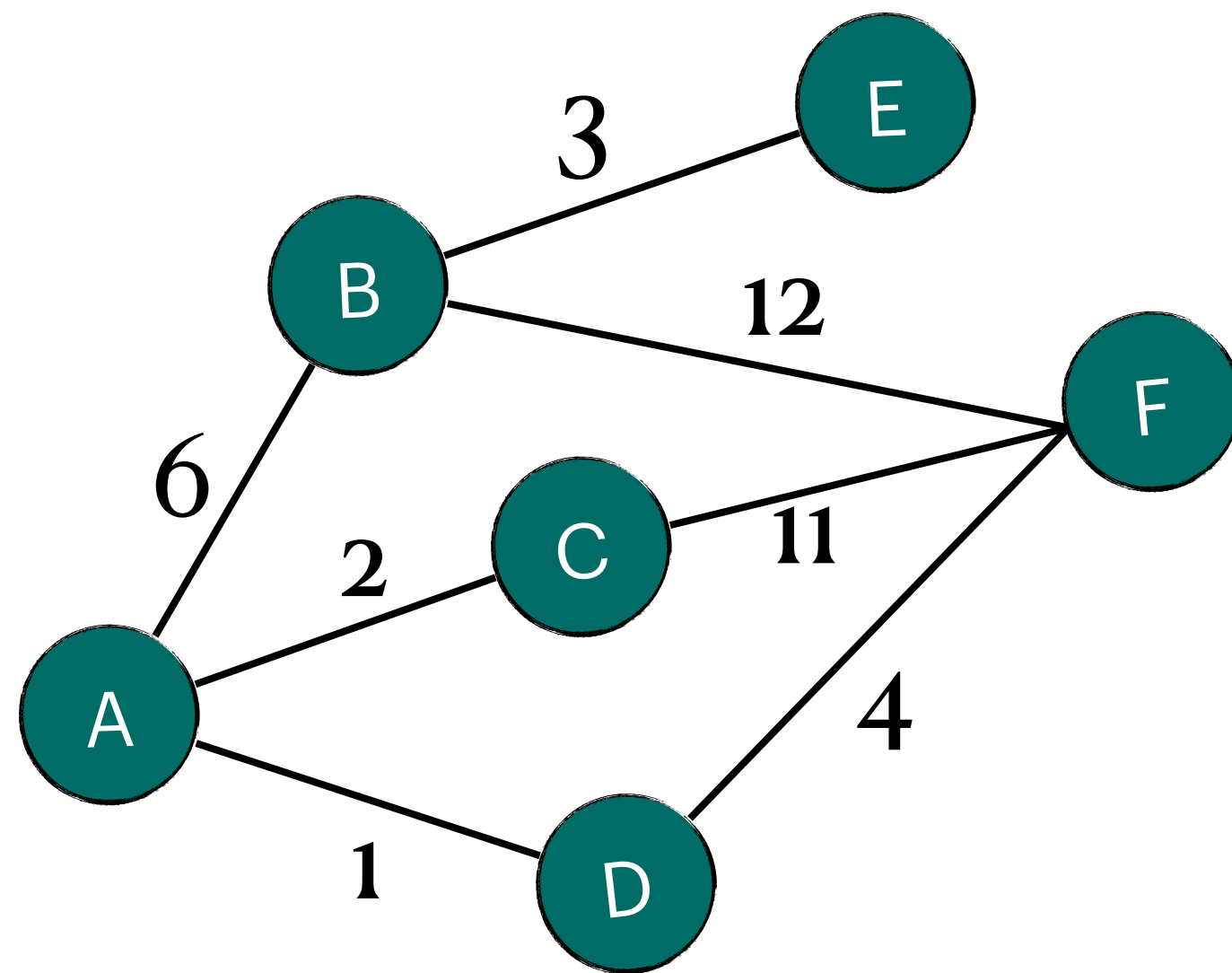


Adjacency Matrix

	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

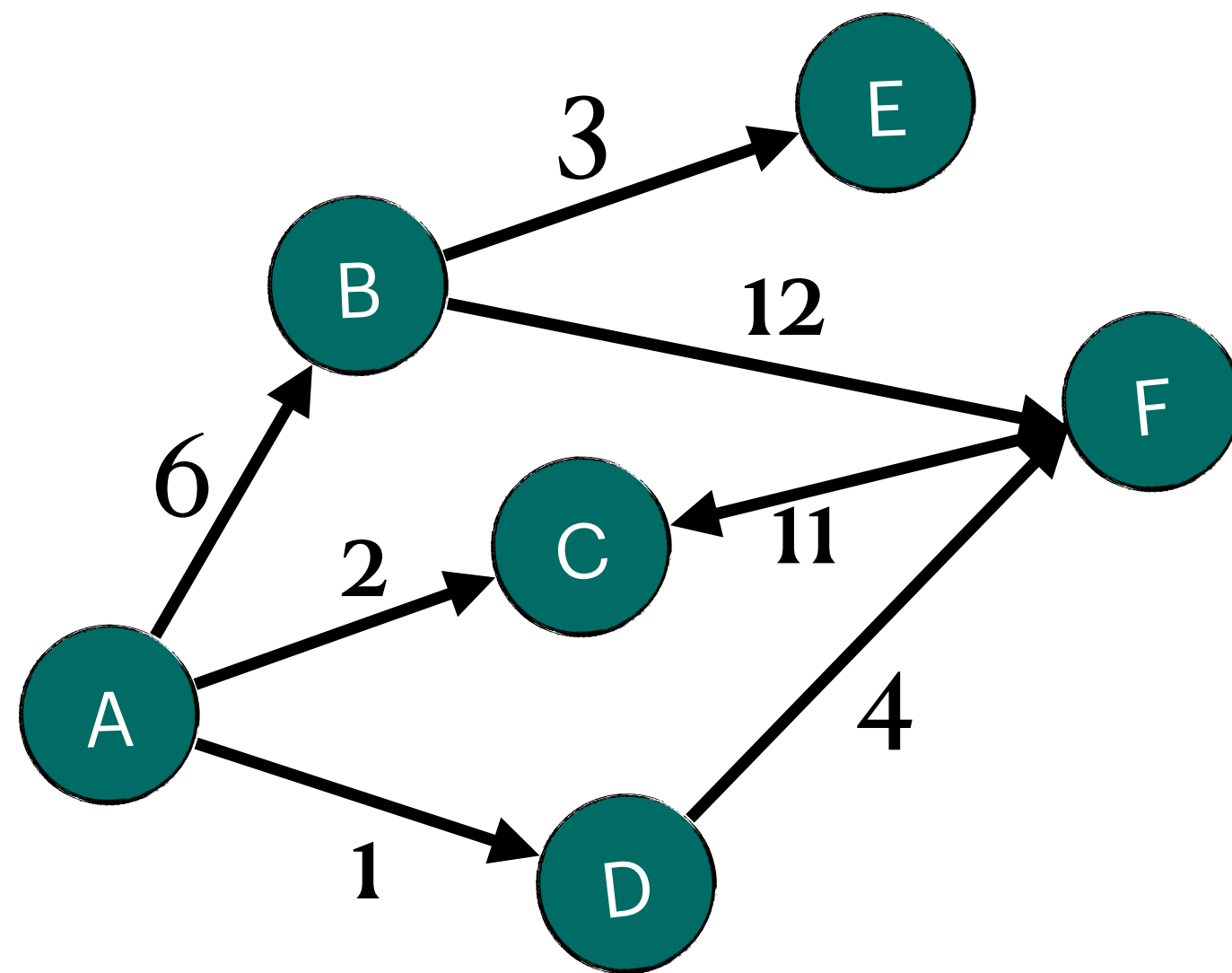


Adjacency Matrix

	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

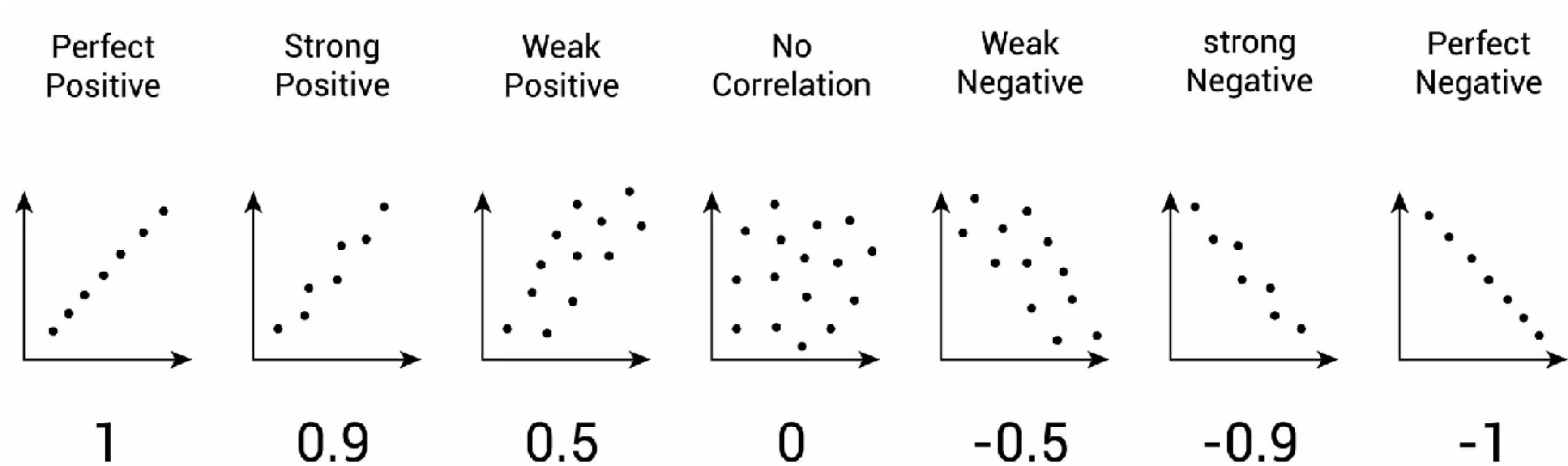


Adjacency Matrix

	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

Data Analysis

Correlation coefficient (2 variables)



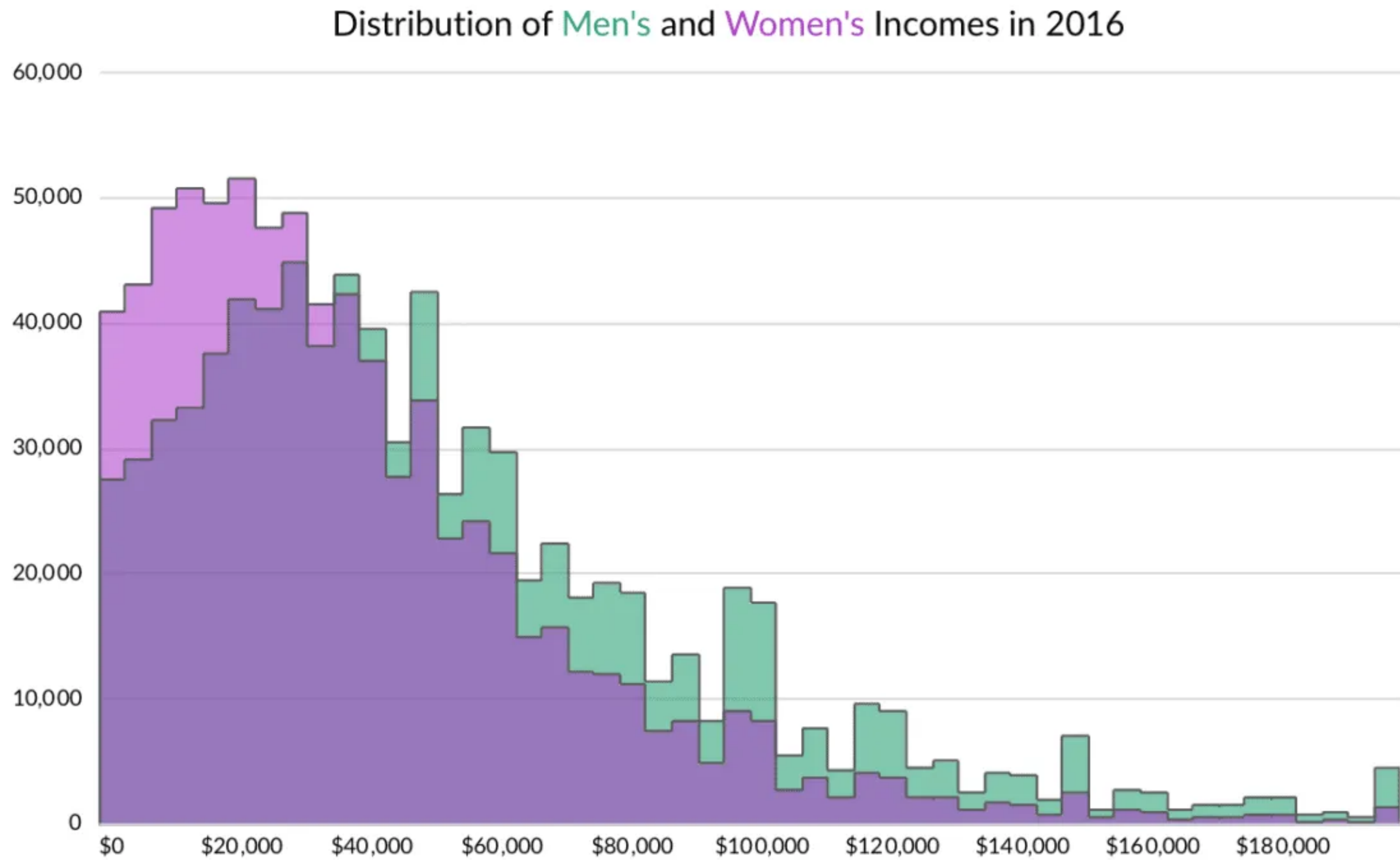
$$r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Correlation Matrix (multiple variables)

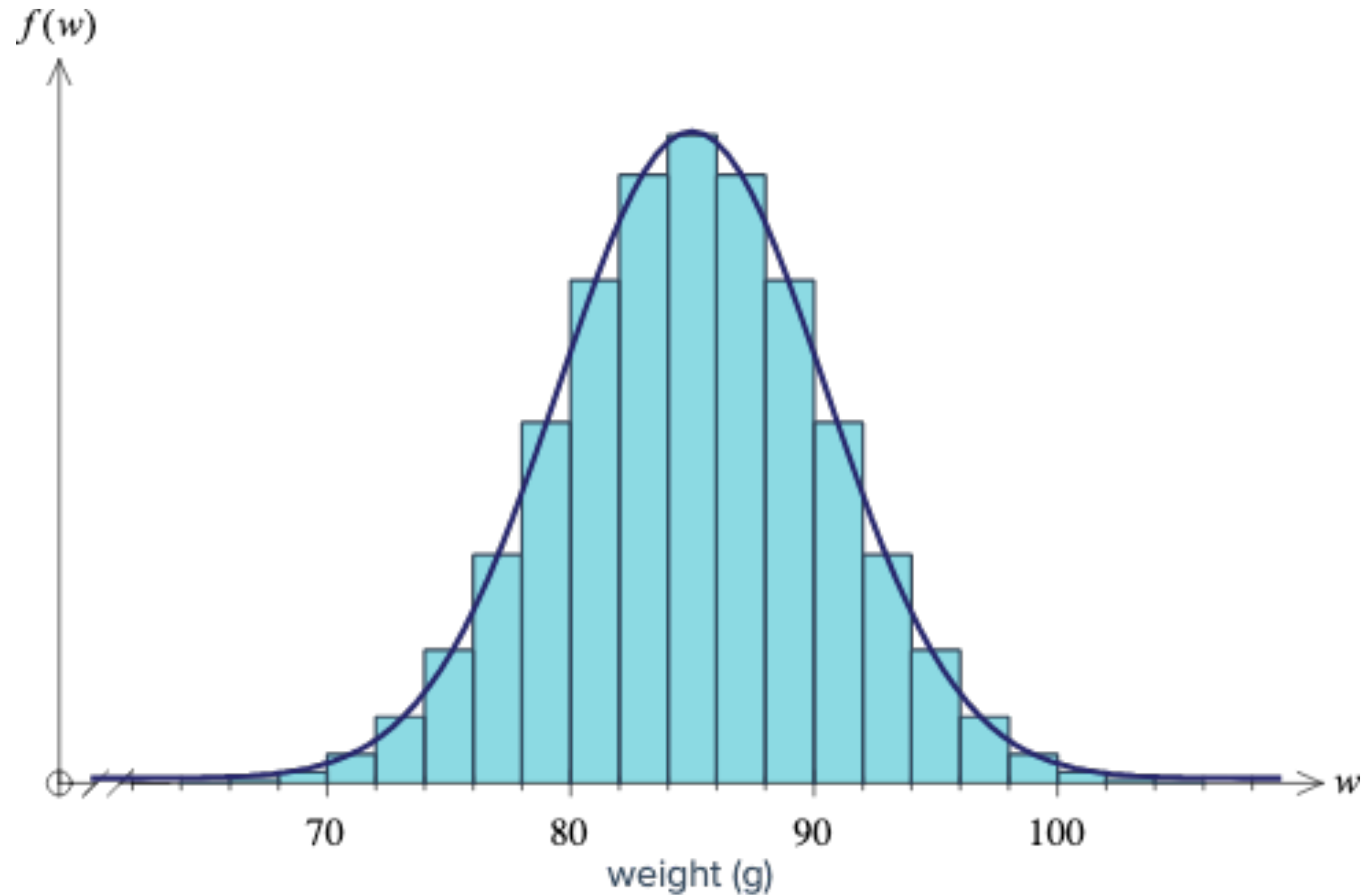
	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

$$r_{ij} = \frac{\text{Cov}(X_i, X_j)}{\sigma_{X_i} \sigma_{X_j}}$$
$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

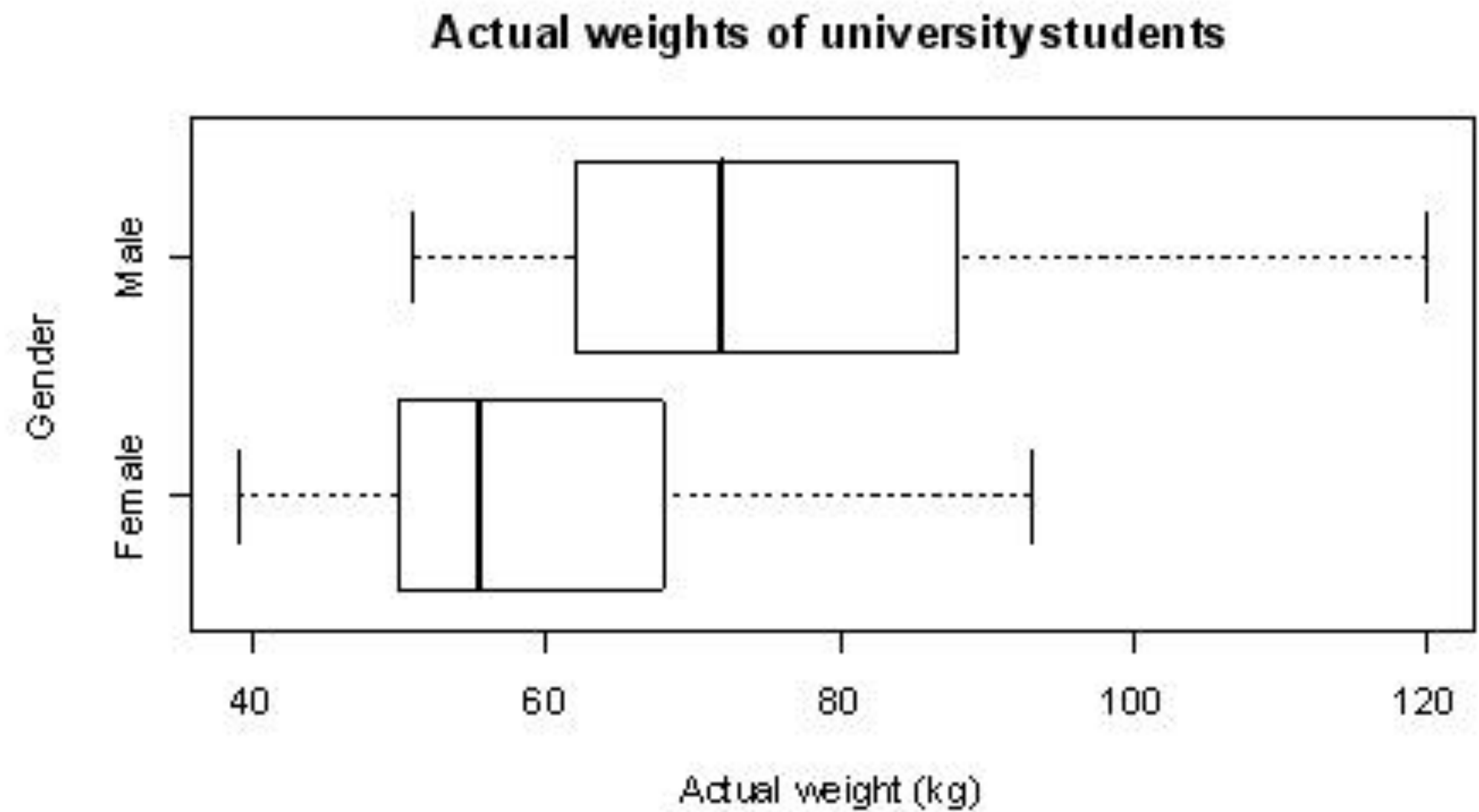
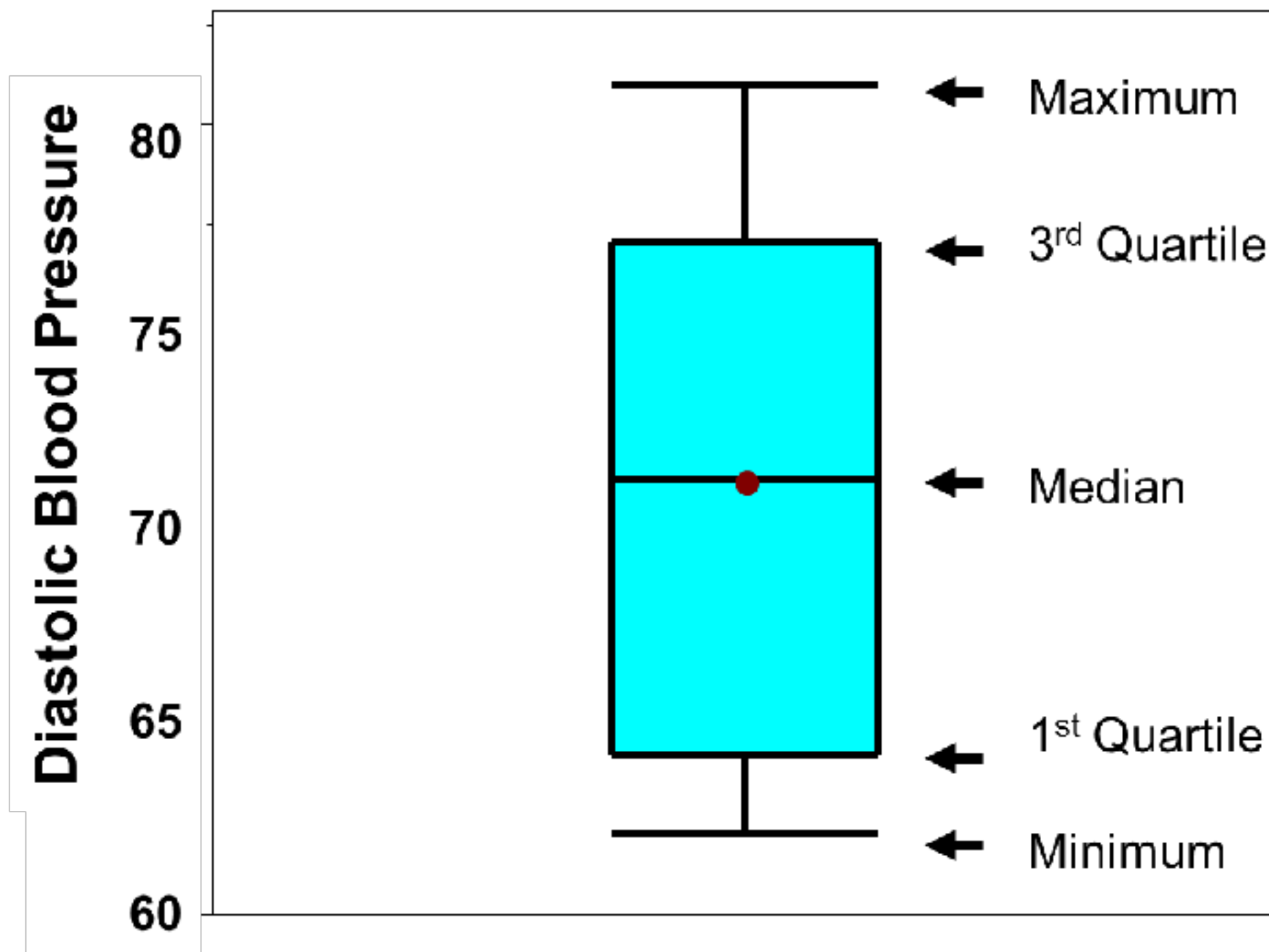
Histogram



Probability Density Function



Box and Whisker Plot



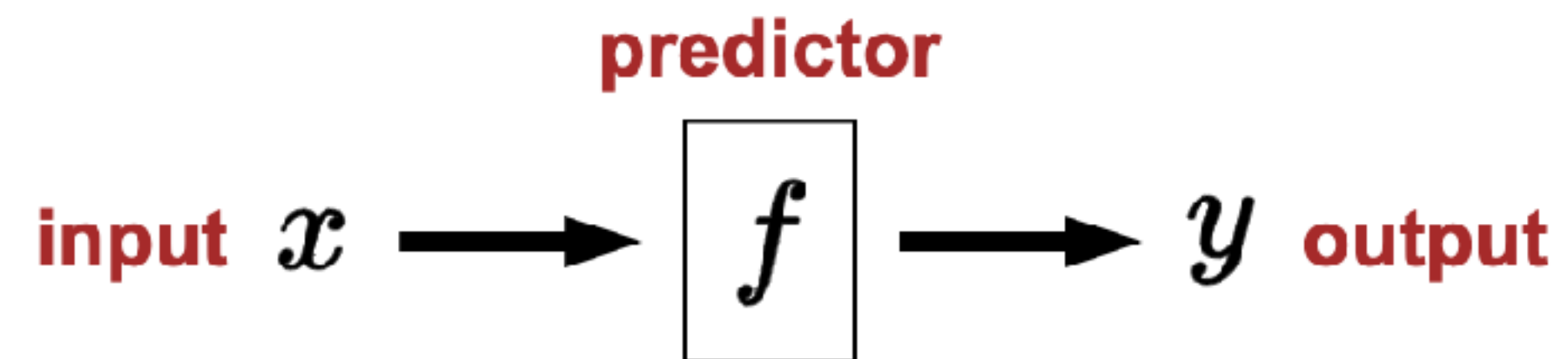
Discovering Functions From Data

Machine Learning

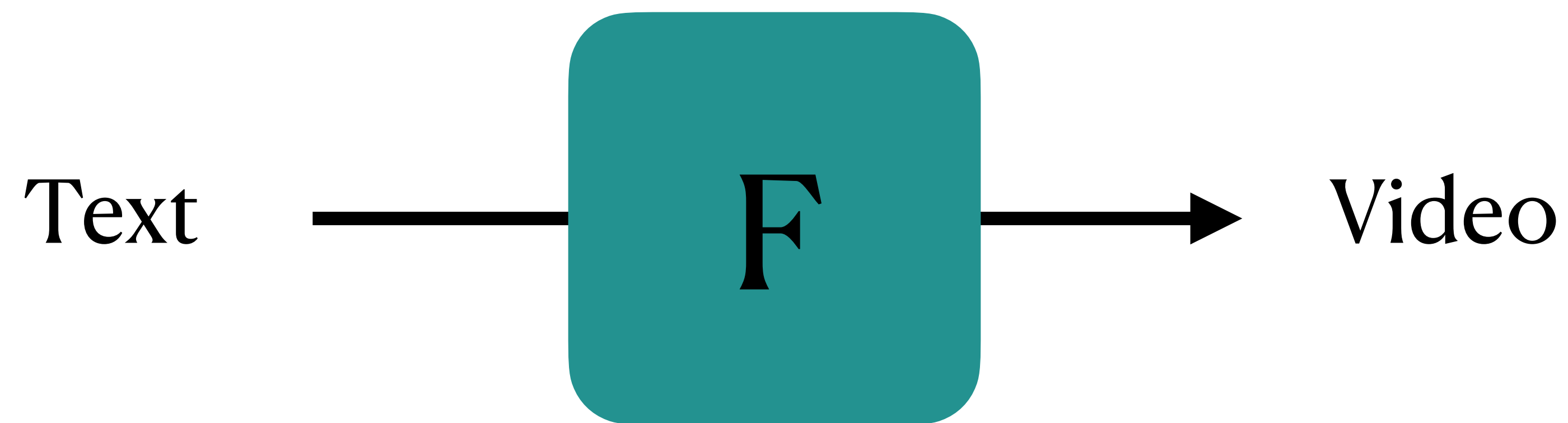
$$f(x) = 2^x \begin{pmatrix} x & 0, 1, 2, 3, 4, 5 \\ y & 1, 2, 4, 8, 16, ? \end{pmatrix}$$

$$\begin{matrix} x & y \\ \curvearrowright \end{matrix}$$

$$f(x) = 2x$$



Sora - Text-to-Video



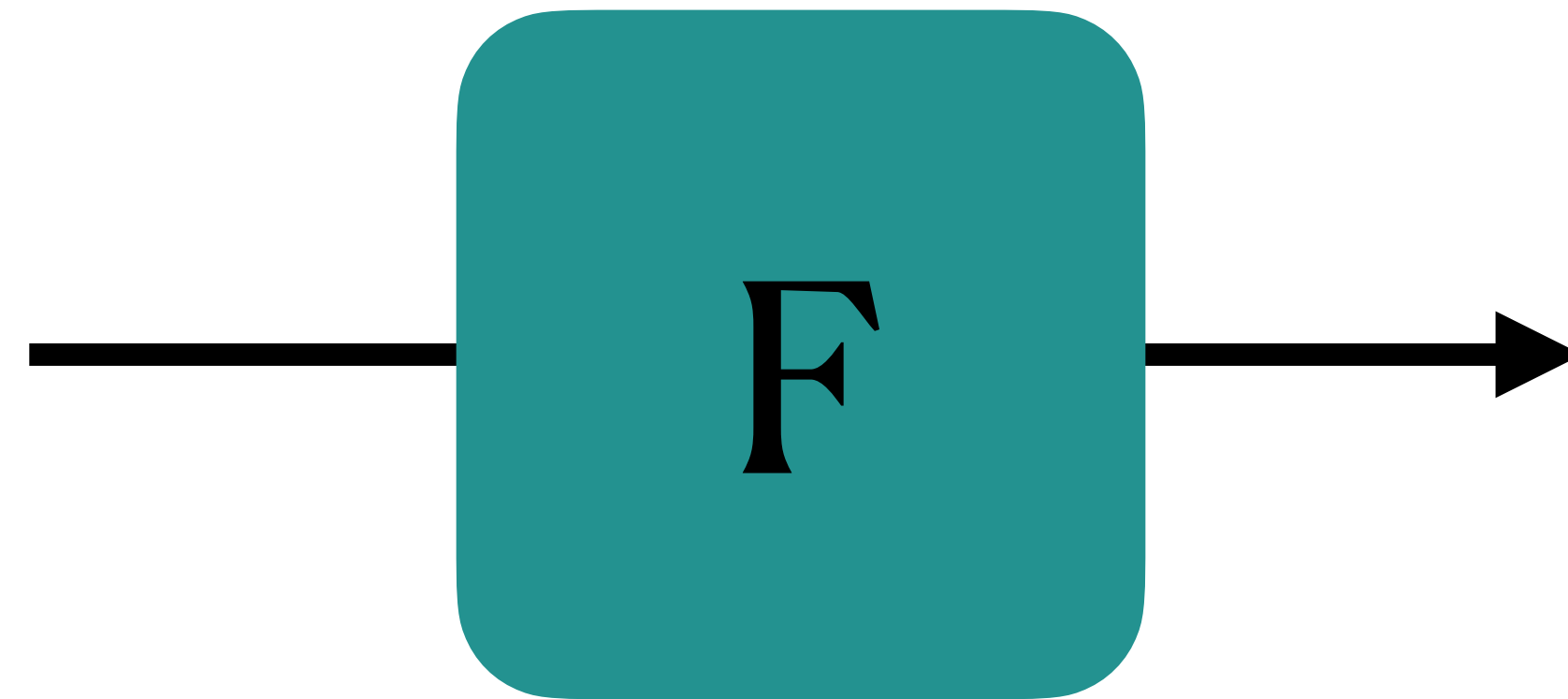
ChatGPT



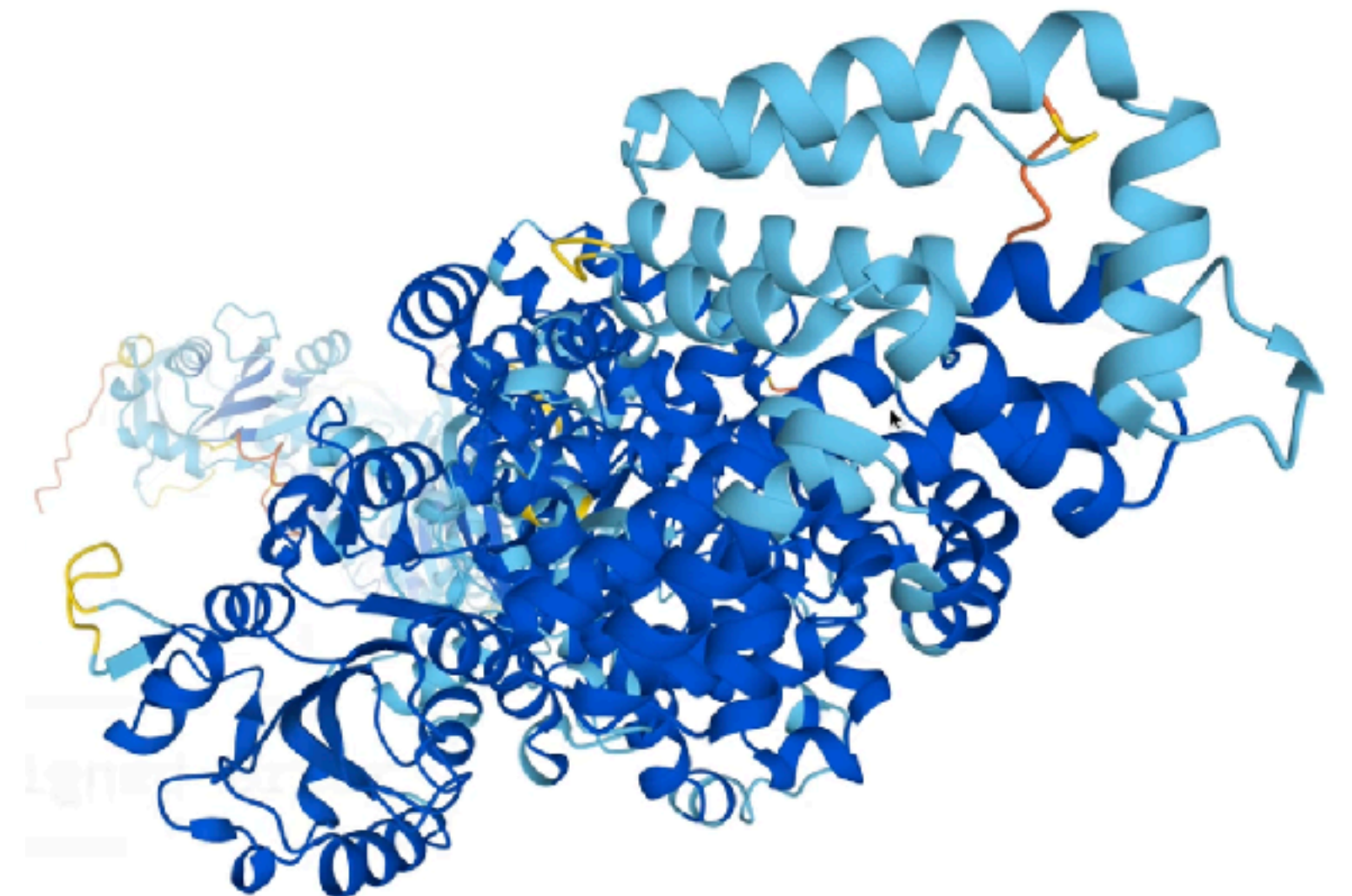
AlphaFold is an AI system developed by DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

AlphaFold

Amino Acid
Sequence

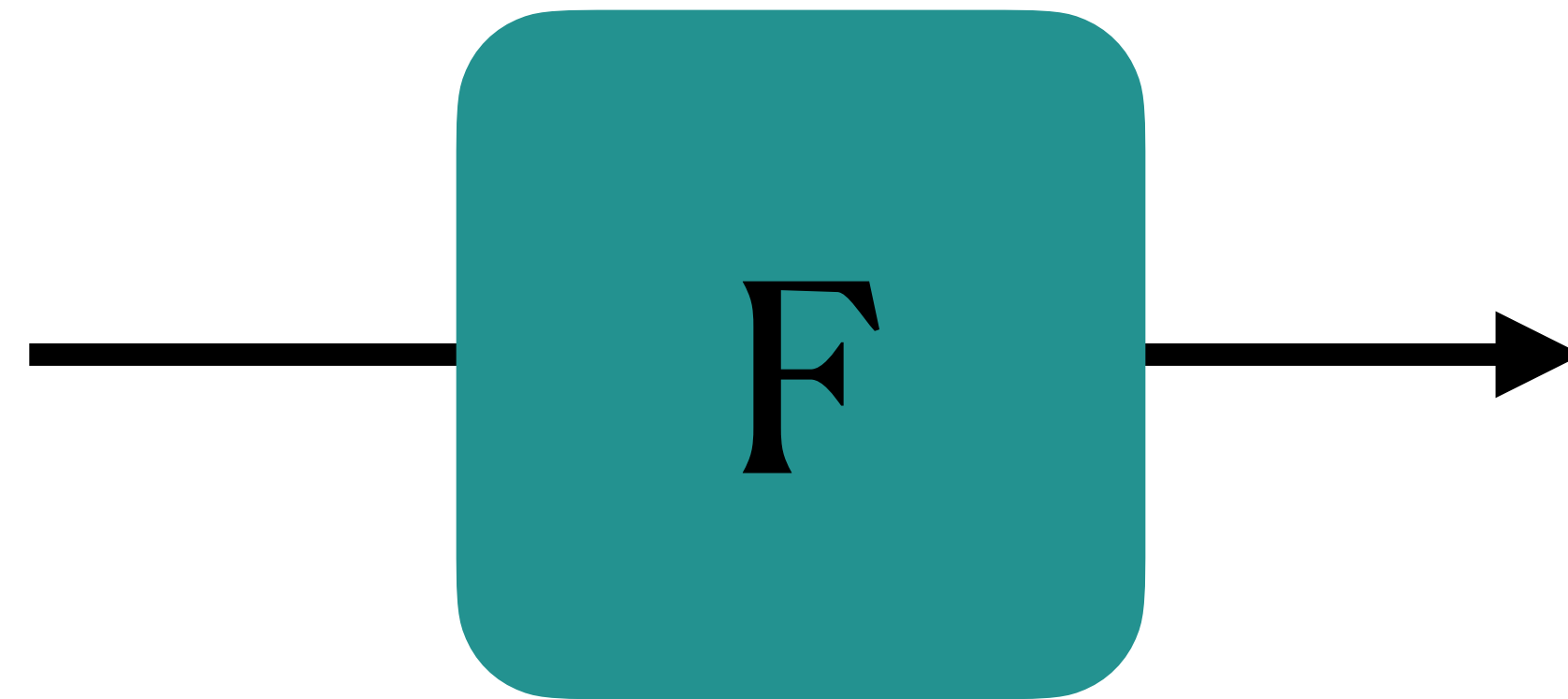
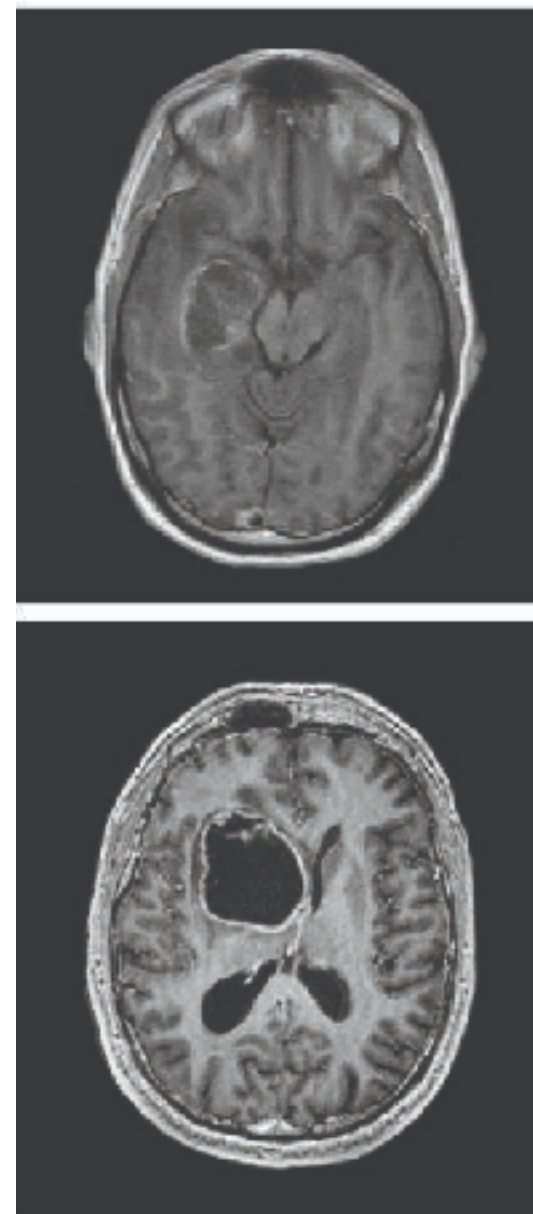


3D protein structure

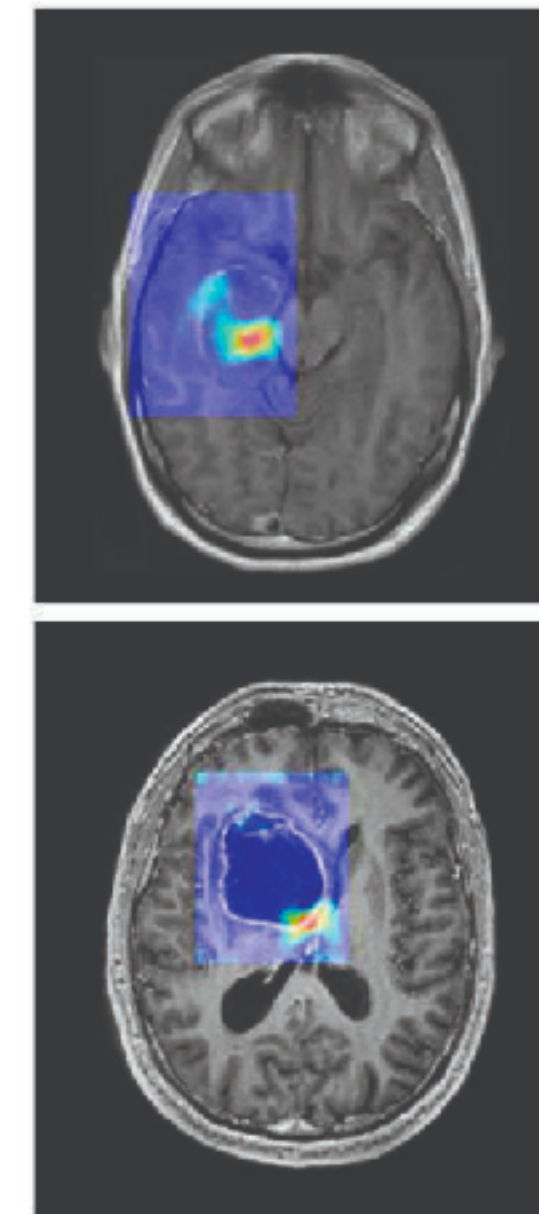


Healthcare

CT Scan



Cancer



Healthcare



Mindaugas Galvosas, MD @MGalvosas · Jan 18

FDA just cleared the first AI device detecting all major skin cancer by DermoSensor.

Its pivotal trial showed sensitivity of 96% and a 97% chance of accurately identifying a skin lesion as benign.

Notably, the approval process began back in 2016.



42

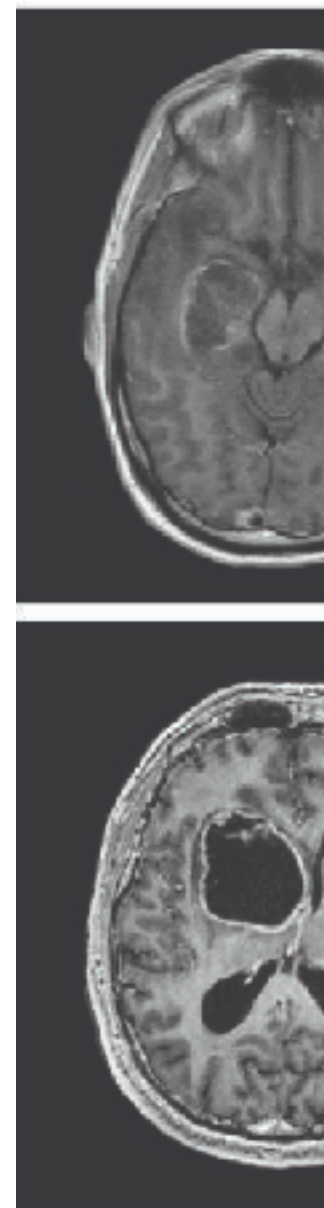
248

1K

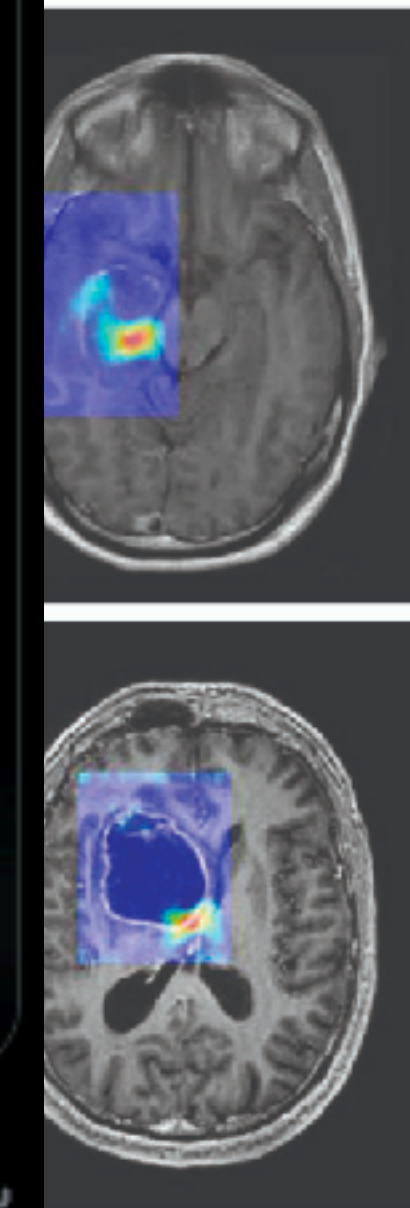
257K



CT S



ancer



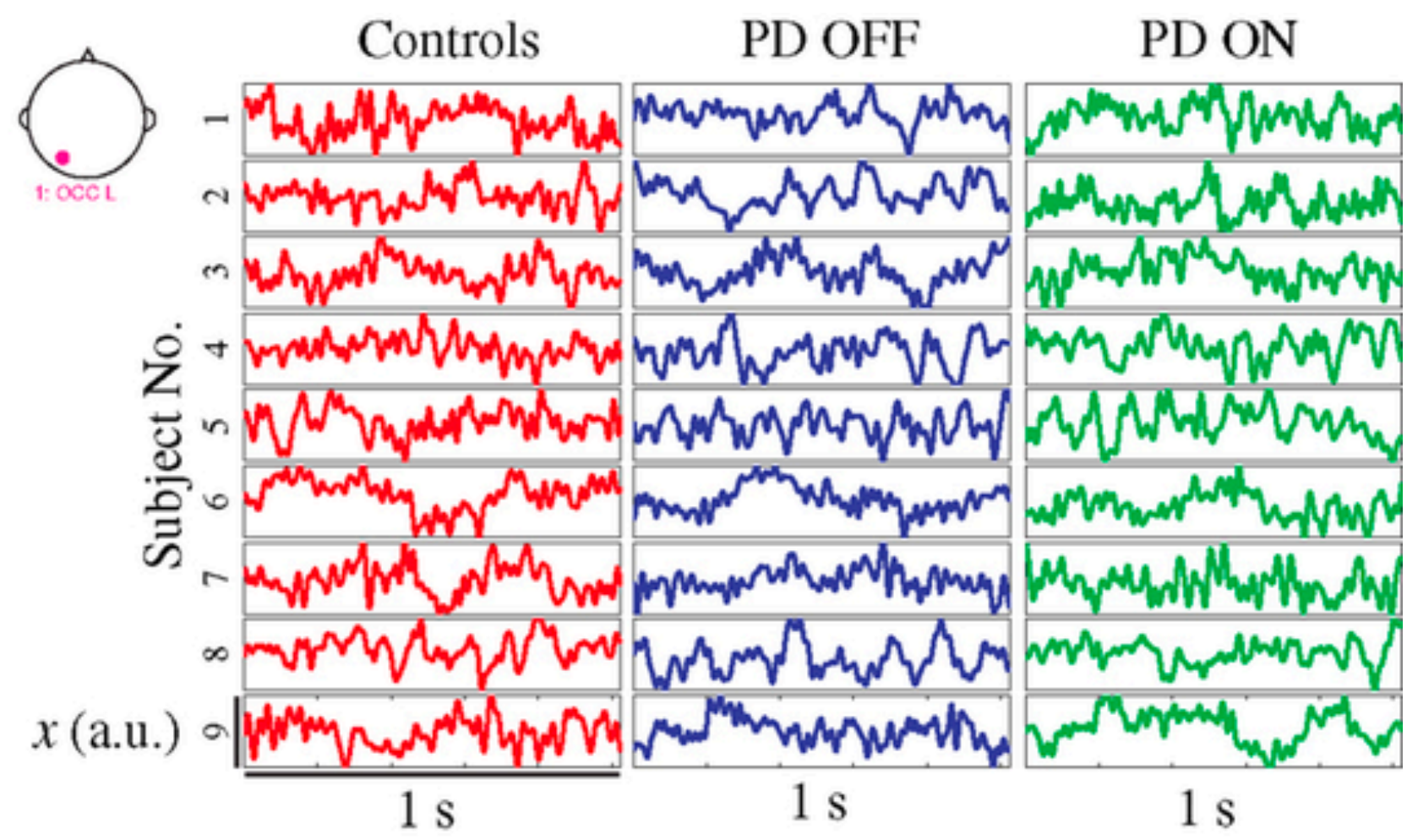
Neuroscience



EEG



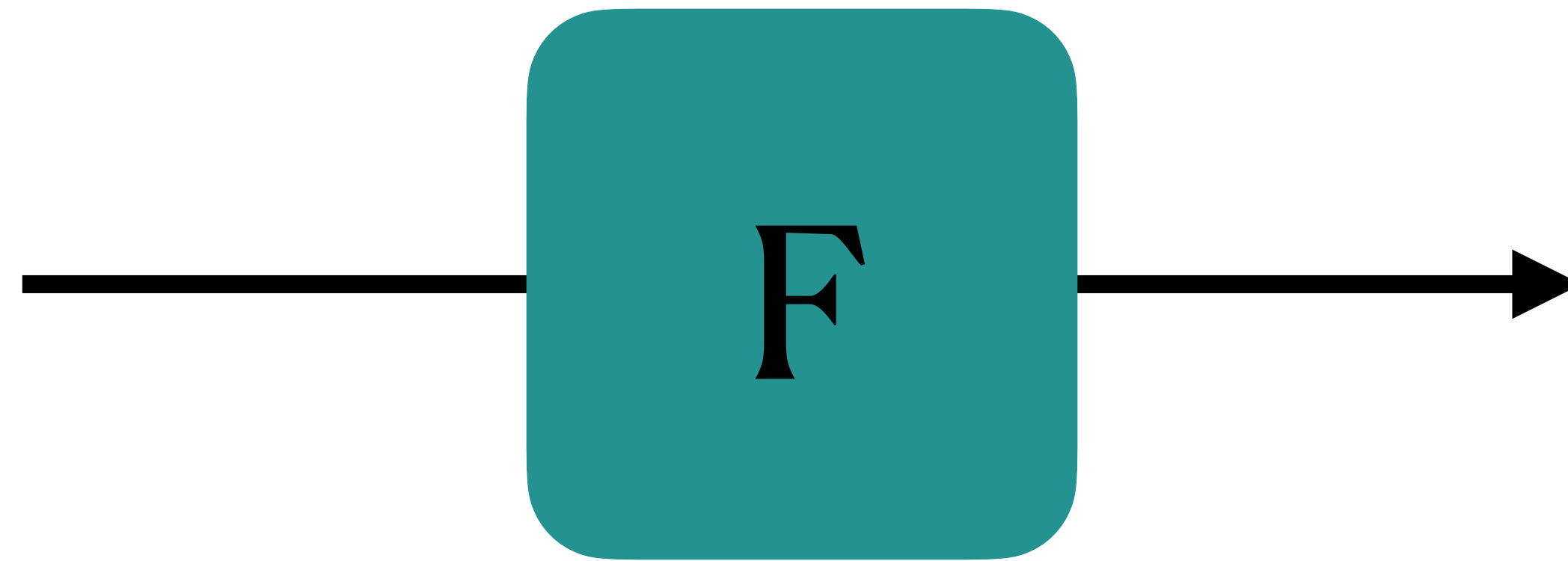
Behavior
Mechanism
Disease



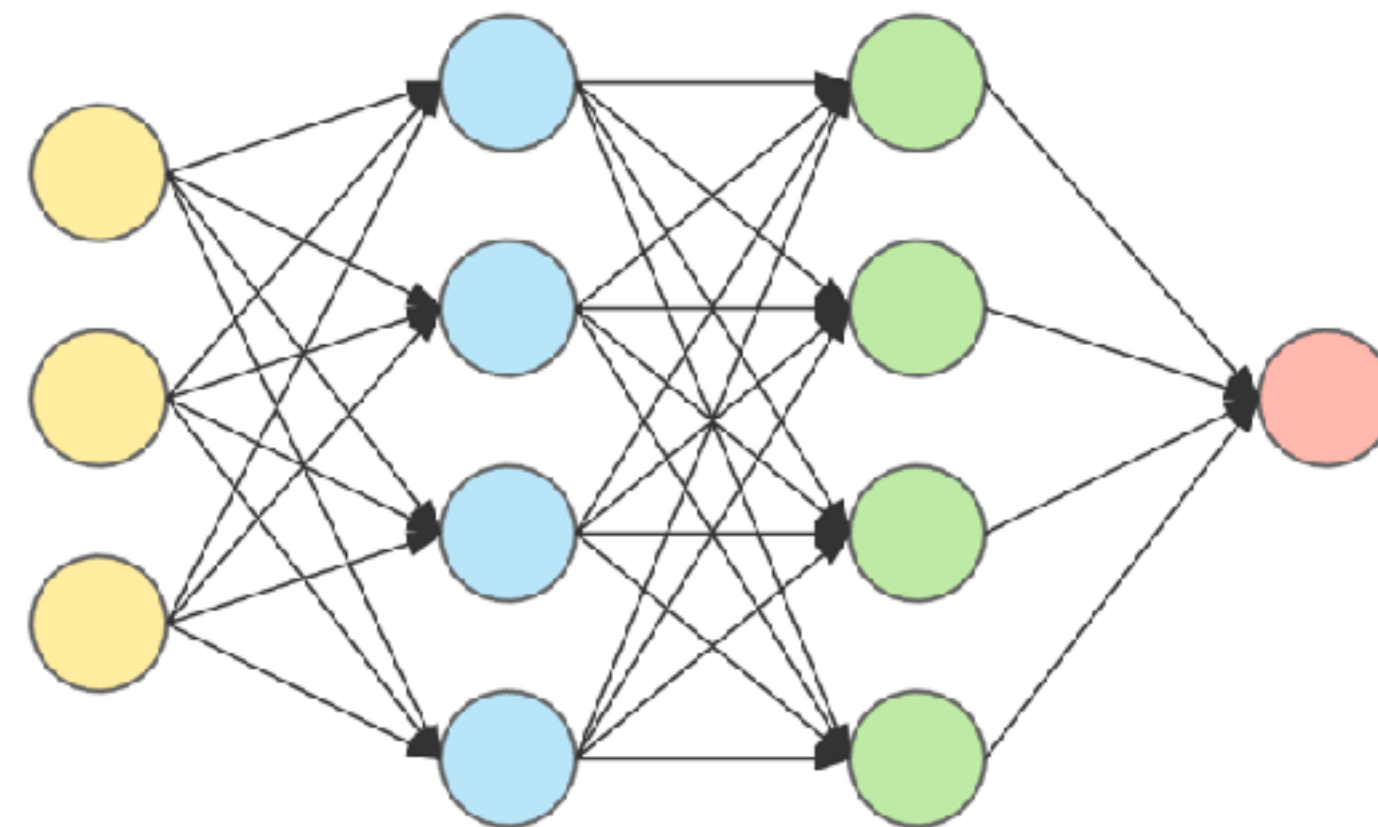
Lainscsek, Claudia, et al. "Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals." *Frontiers in neurology* 4 (2013): 200.

Neural Networks

Input



Output

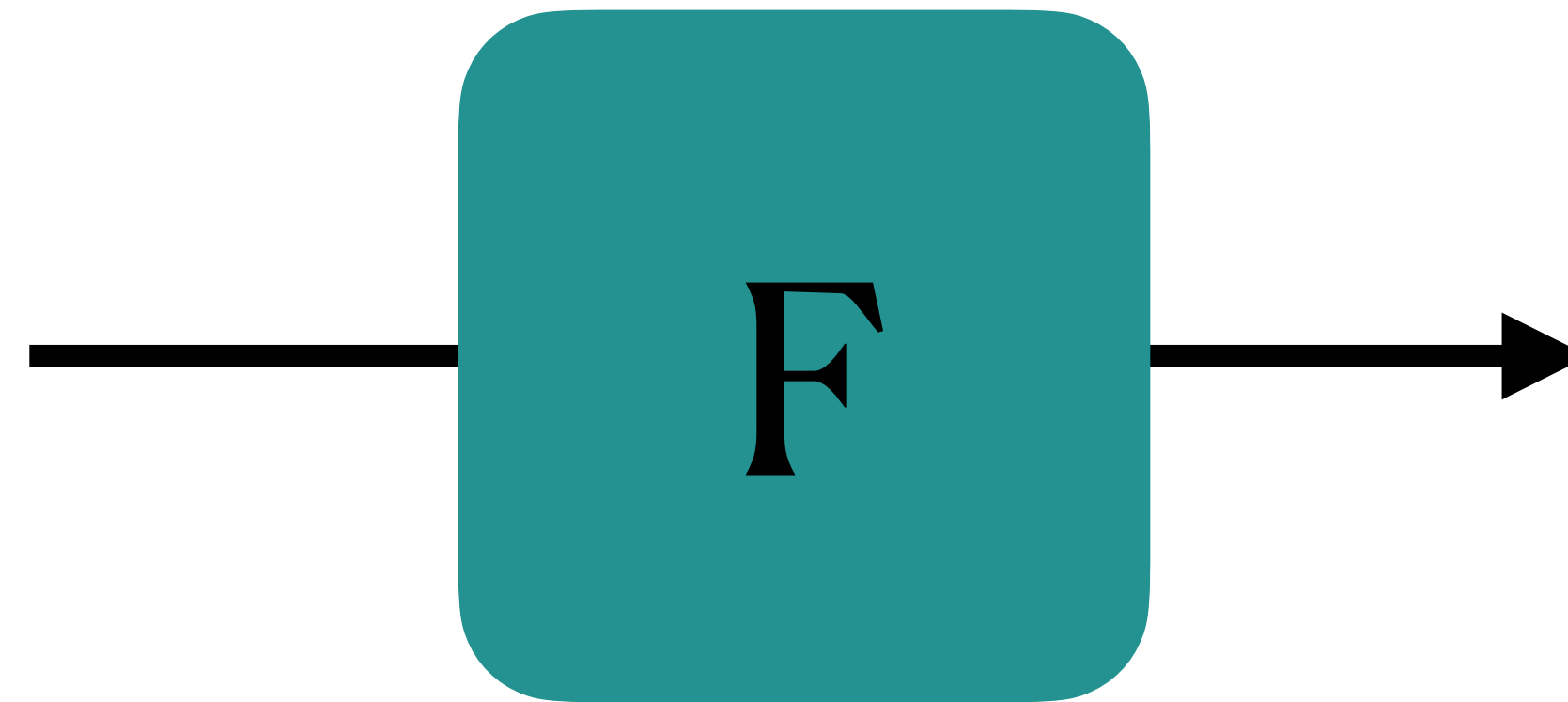
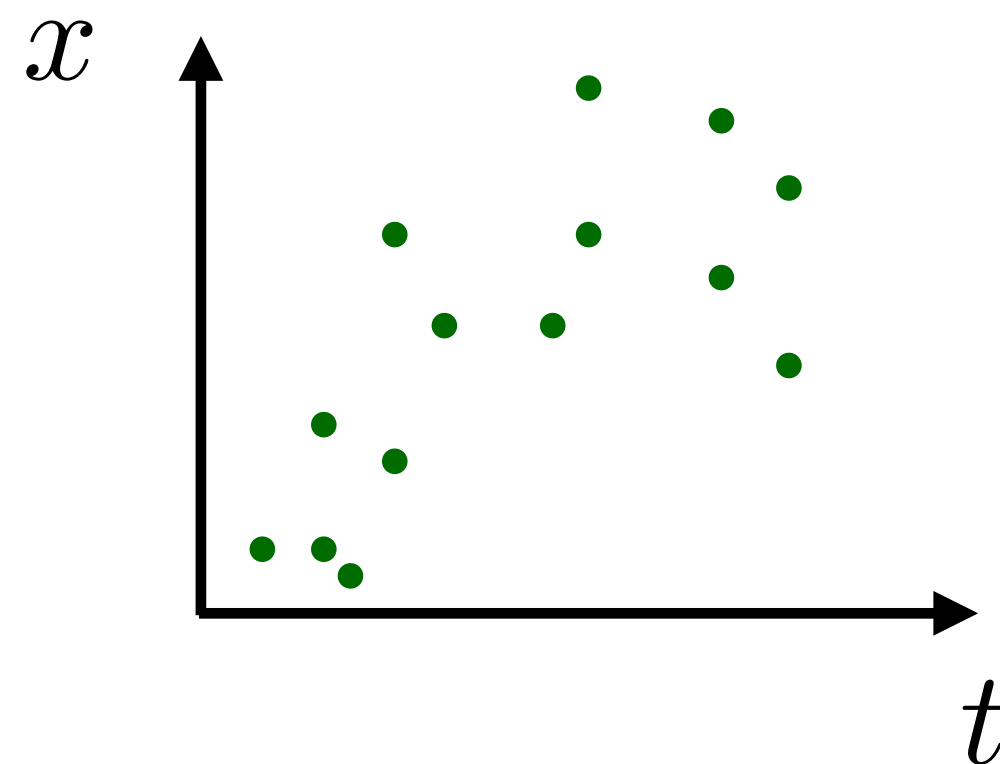


$$f \circ f$$

Applications in Scientific Computing

Symbolic Regression

Time series



Equation

$$\frac{dx}{dt} = f(x)$$

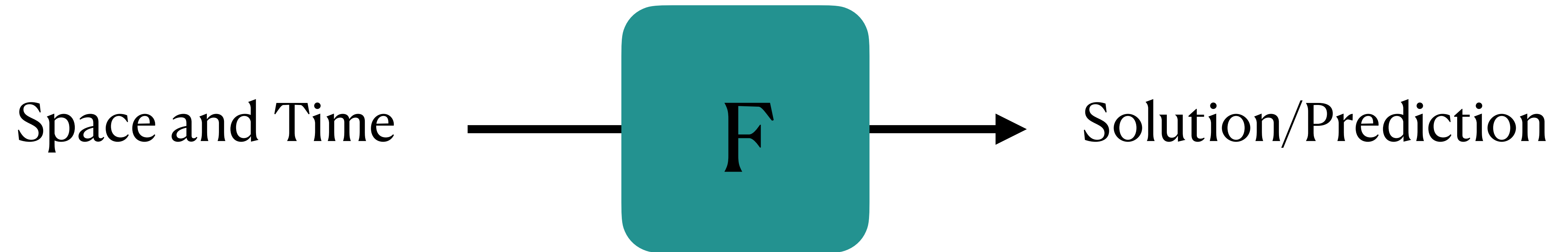
+ Physical Constraints

e.g. Conservation of Energy

???

Applications in Scientific Computing

Physics Informed Neural Networks



+ Physical Constraints

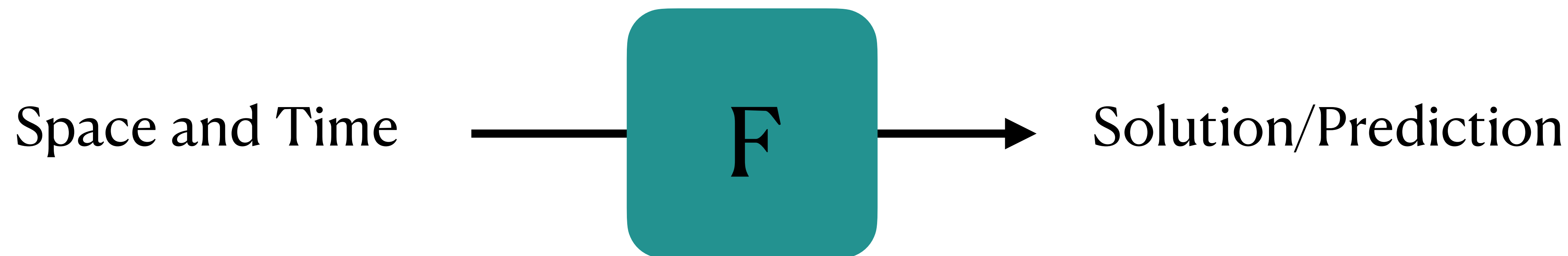
e.g. Conservation of Energy

???

A decoder-only foundation model for time-series forecasting

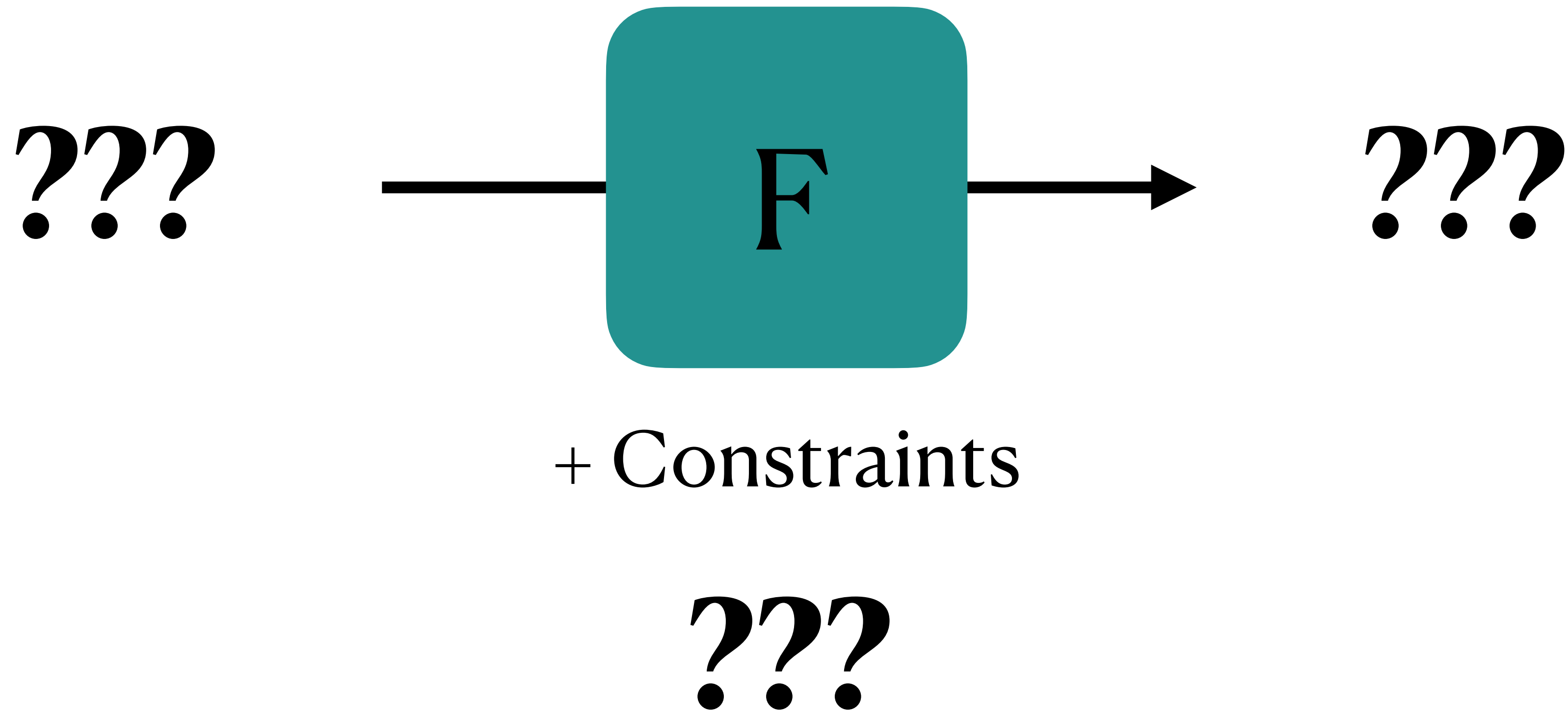
FRIDAY, FEBRUARY 02, 2024

Posted by Rajat Sen and Yichen Zhou, Google Research



Despite DL-based forecasters largely **outperforming** traditional methods and progress being made in **reducing training and inference costs**, they face challenges: most DL architectures require **long and involved training and validation cycles** before a customer can test the model on a new time-series. A foundation model for time-series forecasting, in contrast, can provide decent out-of-the-box forecasts on unseen time-series data with no additional training, enabling users to focus on refining forecasts for the actual downstream task like **retail demand planning**.

What is your input-output of interest?



Laws are linear

Pascal's law (1653)



$$\Delta p = \rho g \Delta h$$

Hooke's law (1678)



$$F = -kx$$

Newton's law of viscosity (1701)



$$\tau = \mu \frac{du}{dy}$$

Ohm's law (1781)



$$I = V/R$$

Fourier's law (1822)



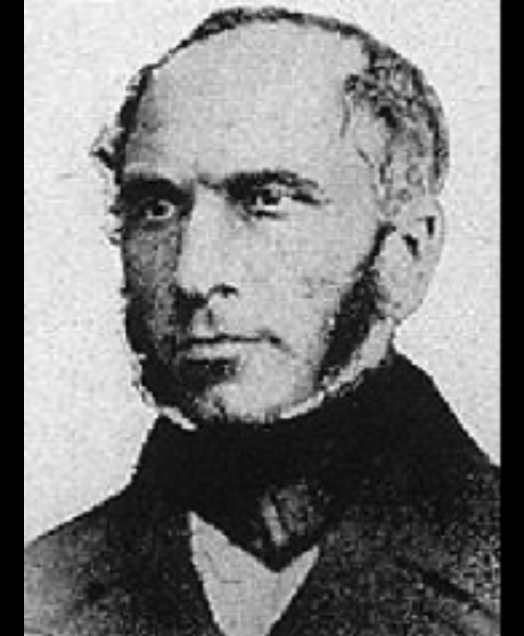
$$q = -k \frac{dT}{dx}$$

Fick's law (1855)



$$J = -D \frac{dC}{dx}$$

Darcy's law (1856)



$$Q = \frac{kA}{\mu L} \Delta p$$

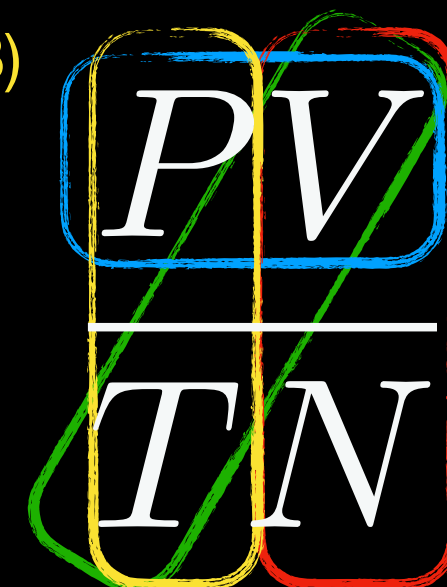
Ideal gas law (1834)

Amontons's law (1808)

Boyle's law (1662)

Charles's law (1787)

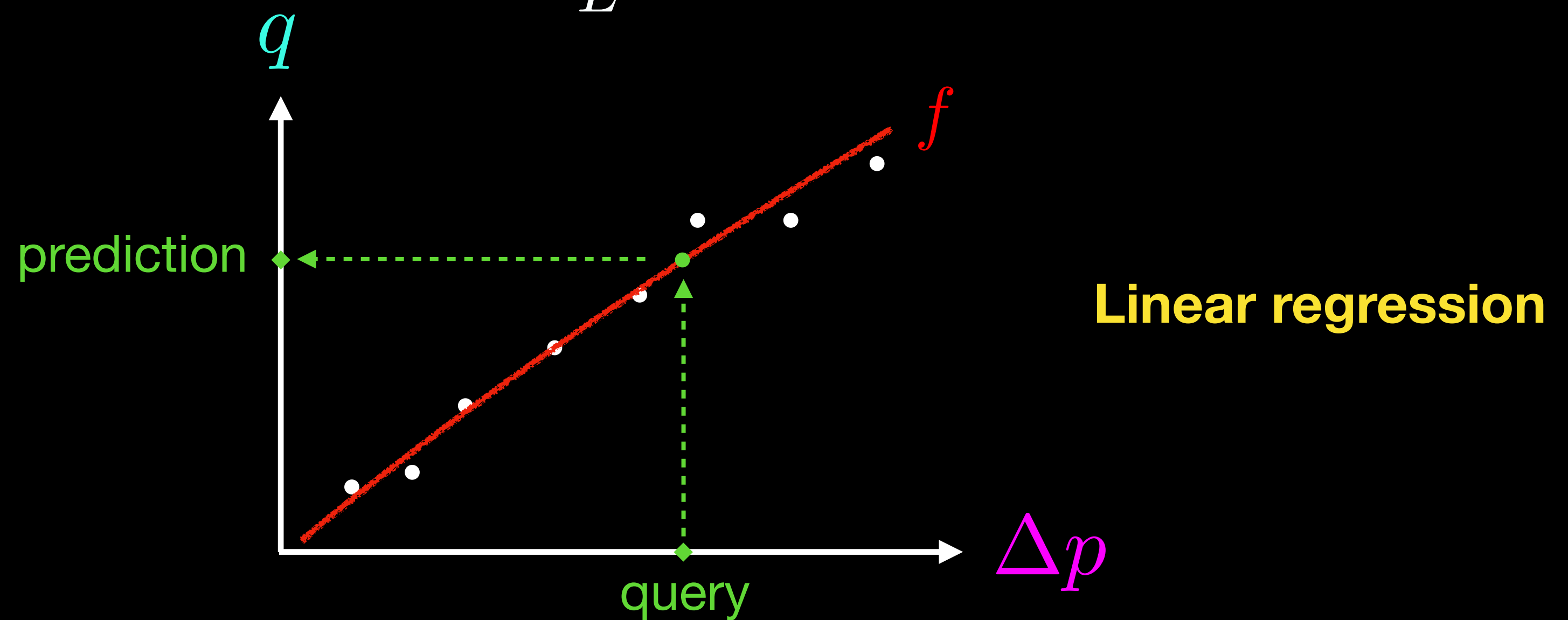
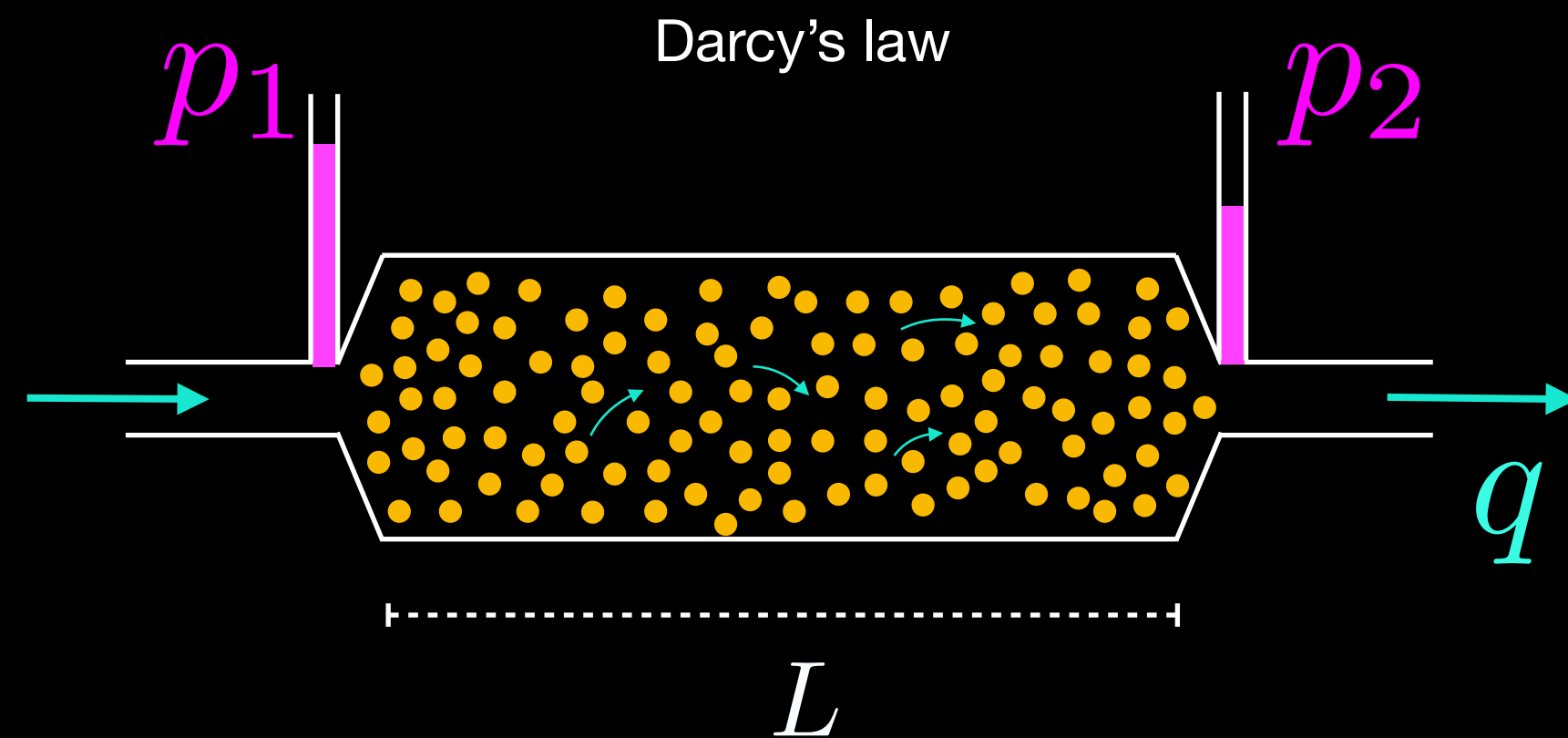
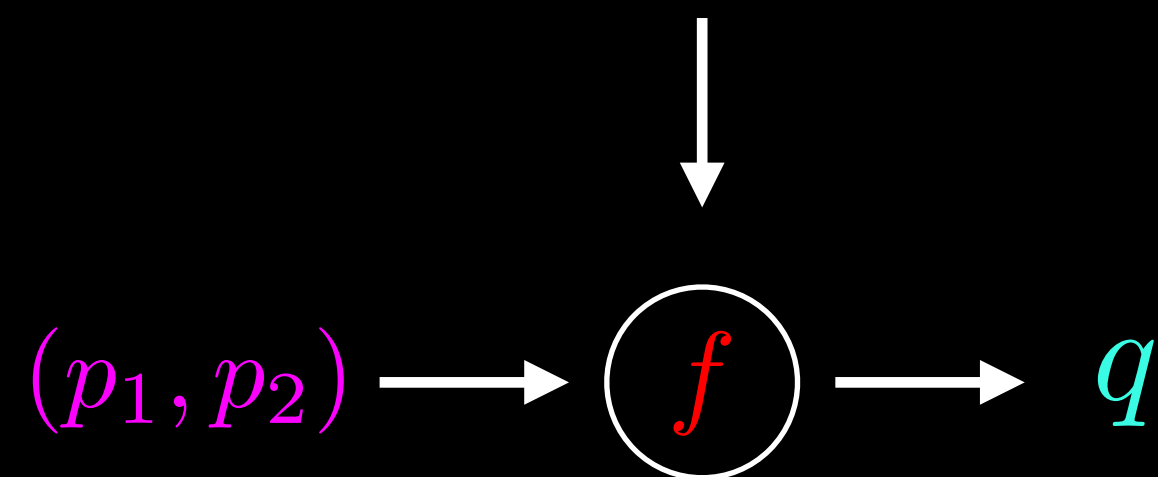
Avogadro's law (1811)



$$= k_B$$

From experiment to Law

p_1	p_2	q
1.3	1.0	22
1.6	1.5	23
3.4	2.4	46
4.8	3.5	67
6.7	4.5	83
...		
2.3	1.4	?



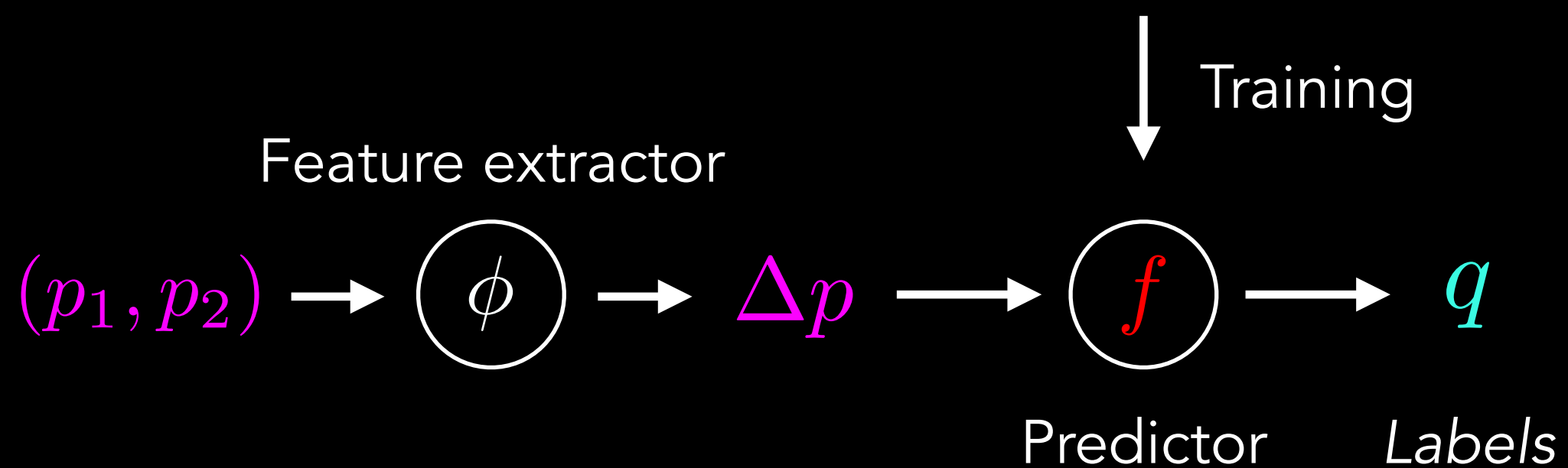
Machine or human learning

Training data: $\mathcal{D}_{\text{train}}$

Example →

p_1	p_2	q
1.3	1.0	22
1.6	1.5	23
3.4	2.4	46
4.8	3.5	67
6.7	4.5	83
...		

- Which predictors are possible?
- How good is the predictor?
- How can we find the best predictor?




Looking for data

The image shows a Google search interface with the query "time series temperature data". The search results are displayed in a dark theme. The first result is from the National Centers for Environmental Information (NCEI) (.gov), titled "Climate at a Glance | Global Time Series", with a snippet mentioning "Global and hemispheric temperature anomalies are with respect to the 1901-2000 average". The second result is from Kaggle, titled "Daily Climate time series data", with a snippet stating "This dataset provides data from 1st January 2013 to 24th April 2017 in the city of Delhi, India". Below the search results, there is a "Datasets" section with three entries: "Global Temperature Time Series" from DataHub, "Time Series Room Temperature Data" from Kaggle, and "Temperature - Historic Daily Time Series" from the Data Catalog of the U.S. Department of the Interior.

Google


time series temperature data

All Images Videos News Web Books Maps More ▾

 National Centers for Environmental Information (NCEI) (.gov)
[https://www.ncei.noaa.gov/access/monitoring/time-...](https://www.ncei.noaa.gov/access/monitoring/time-series)

Climate at a Glance | Global Time Series

14 hours ago — **Global and hemispheric temperature anomalies are with respect to the 1901-2000 average.** Coordinate temperature anomalies are with respect to the 1991-2020 ...

 Kaggle
[https://www.kaggle.com/datasets/sumanthvrao/dail...](https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series)

Daily Climate time series data

This dataset provides **data** from 1st January 2013 to 24th April 2017 in the city of Delhi, India. The 4 parameters here are meantemp, humidity, wind_speed, ...

Datasets :

<https://datahub.io/core/global-temp>
Global Temperature Time Series
Aug 29, 2017 — Data are included from the GISS Surface Temperature (GISTEMP) analysis and the global component of Climate at a Glance (GCAG). Two datasets are provided: 1) global monthly mean...
Licence: ODC Public Domain Dedication and Lic...

<https://www.kaggle.com/datasets/vitthalmadane/ts-temp-1>
Time Series Room Temperature Data
Nov 21, 2022 — Dataset is generated with help of an IOT Device data represents room air temperature values with respect time. In Time Series observations are function of time, each data corresponds to ...
Licence: Data files © Original Authors

<https://catalog.data.gov/dataset/temperature-historic-daily-time-series>
Temperature - Historic Daily Time Series
Nov 29, 2024 — Annual dataset covering the conterminous U.S., from 1981 to now. Contains spatially gridded annual average daily mean temperature at 4km grid cell resolution. Distribution of the point ...

Where to look for data?

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

Datasets

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


+ New Dataset

 Filters


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


Trending Datasets

See All





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

20 





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

20 




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

14 



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

12 

LLM Fine-Tuning

See All

Where to look for data?

huggingface.co



Search models, datasets, users...

Models

Datasets

Spaces

Posts

Docs

Solutions

Pricing

Log In

Sign Up

Main Tasks Libraries 1 Languages Licenses

Other

Modalities

3D Audio Geospatial Image
Tabular Text Time-series Video

Size (rows)

<1K >1T

Format

json csv parquet imagefolder
soundfolder webdataset text arrow

Datasets 143,641

Filter by name

Full-text search

Sort: Trending

fka/awesome-chatgpt-prompts

Viewer • Updated Mar 7, 2023 • 153 • 6.12k • 5.29k

princeton-nlp/SWE-bench_Verified

Viewer • Updated 5 days ago • 500 • 1.95k • 83

nisten/all-human-diseases

Viewer • Updated about 2 hours ago • 2.2k • 29 • 53

THUDM/LongWriter-6k

Viewer • Updated 5 days ago • 6k • 40 • 46

G-reen/TheatreLM-v2.1-Characters

Viewer • Updated 4 days ago • 5.01k • 7 • 38

Imms-lab/LLaVA-OneVision-Data

Viewer • Updated 2 days ago • 3.46M • 2.16k • 77

airtrain-ai/fineweb-edu-fortified

Viewer • Updated 11 days ago • 322M • 673 • 30

BAAI/Infinity-Instruct

Viewer • Updated 5 days ago • 20.4M • 2.56k • 358

UCSC-VLAA/MedTrinity-25M

Viewer • Updated 11 days ago • 24.9M • 175 • 53

multimodalart/1920-raider-waite-tarot-public-domain

Viewer • Updated 5 days ago • 78 • 24 • 16

Homework

Collect data from your phone and do something with it