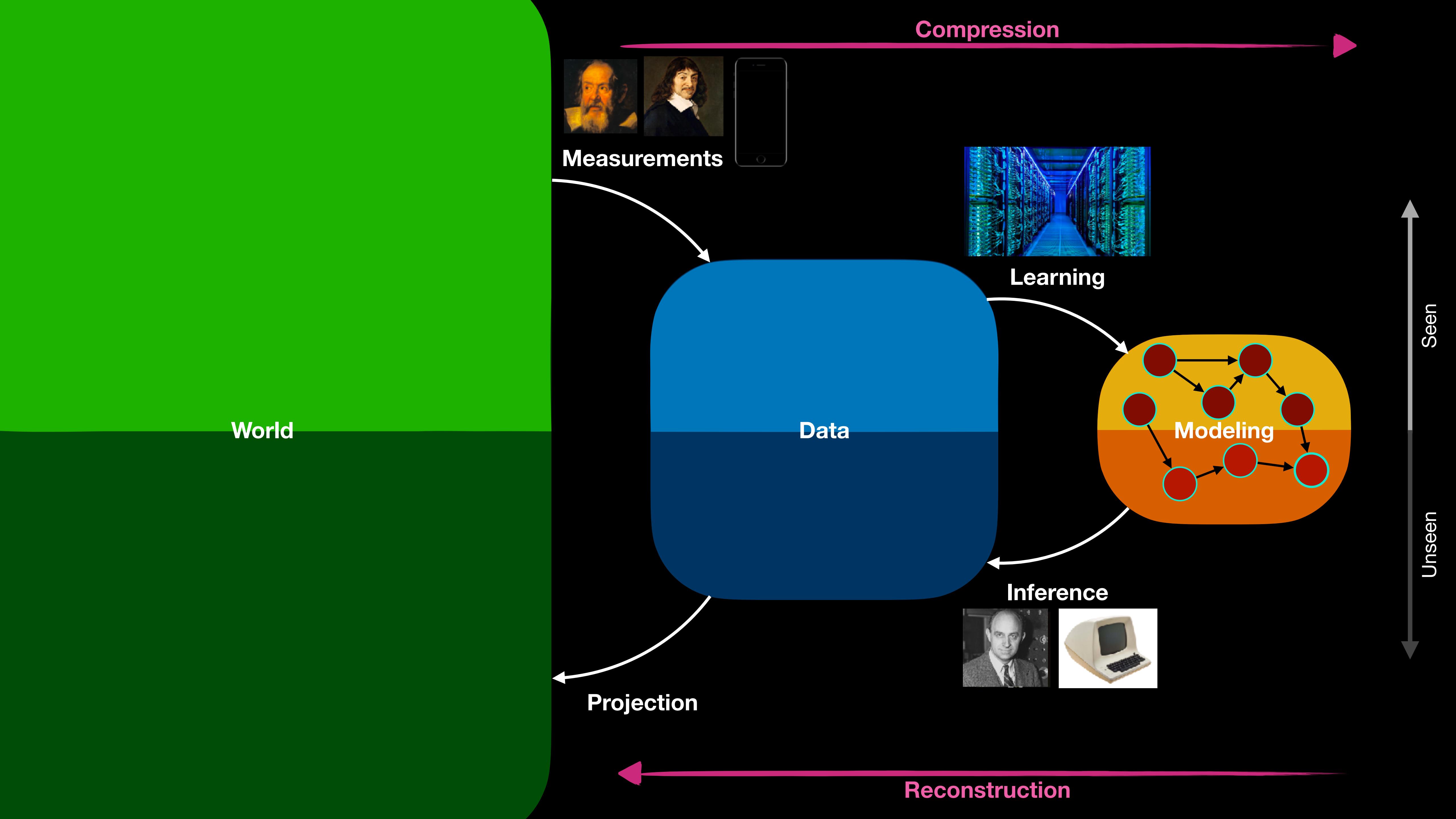


From Data to Learning



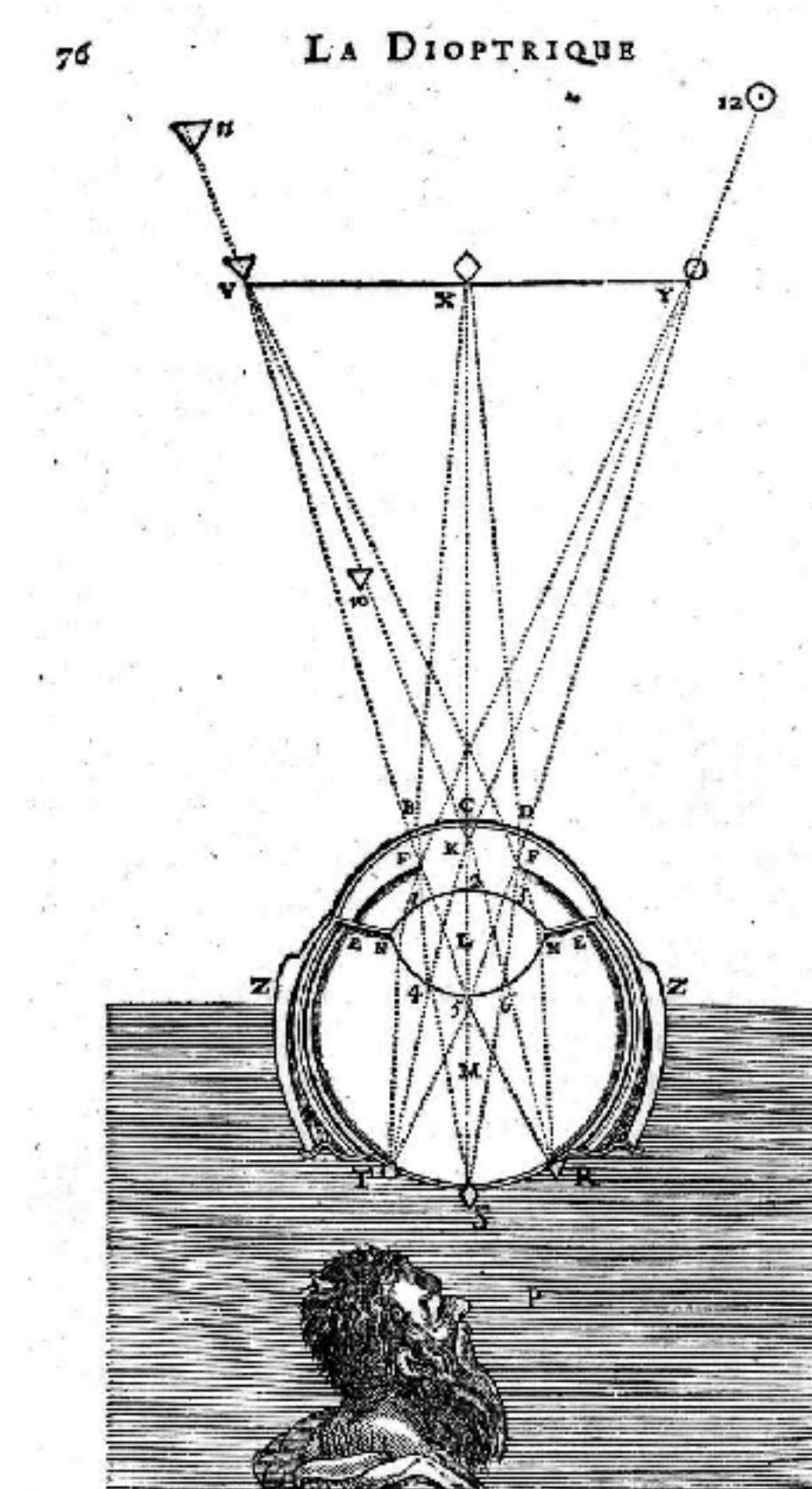
Finding constants of nature that generalize across space and time



Galileo



Kepler



Descartes

Empirical Laws are linear

Pascal's law (1653)



Hooke's law (1678)



Newton's law of viscosity (1701)



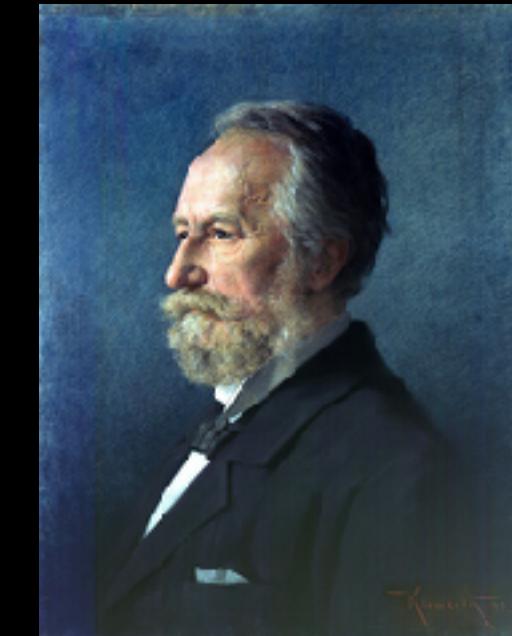
Ohm's law (1781)



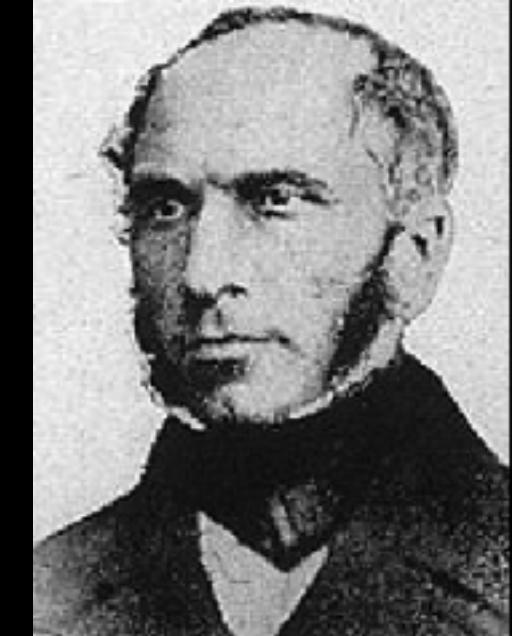
Fourier's law (1822)



Fick's law (1855)



Darcy's law (1856)



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

$$F = -kx$$

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

$$I = V/R$$

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

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

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

Ideal gas law (1834)

Amonton's law (1808)

Charles's law (1787)

Boyle's law (1662)

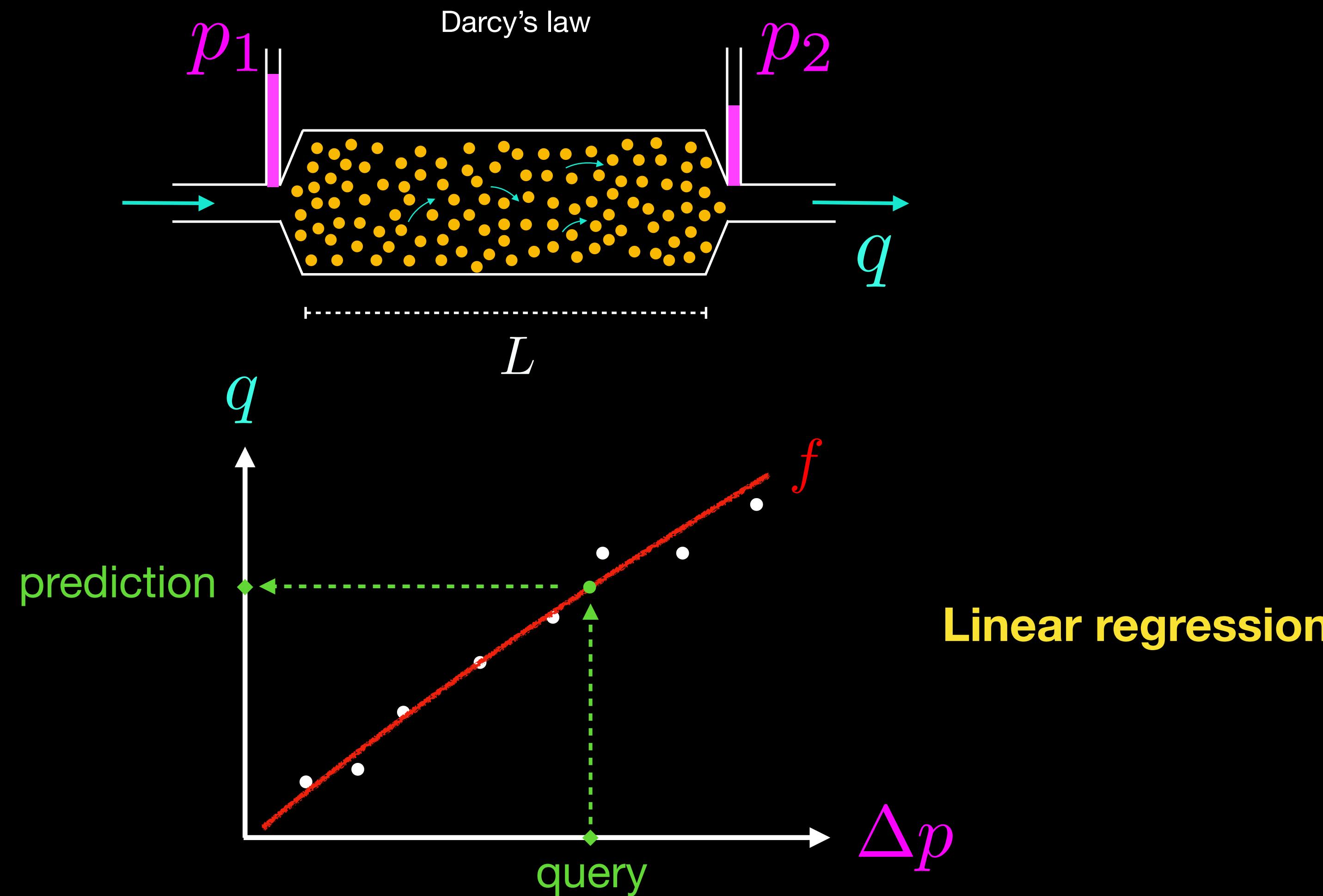
Avogadro's law (1811)

$$\frac{PV}{TN} = k_B$$

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	?

$$(p_1, p_2) \rightarrow f \rightarrow q$$



Machine or human learning

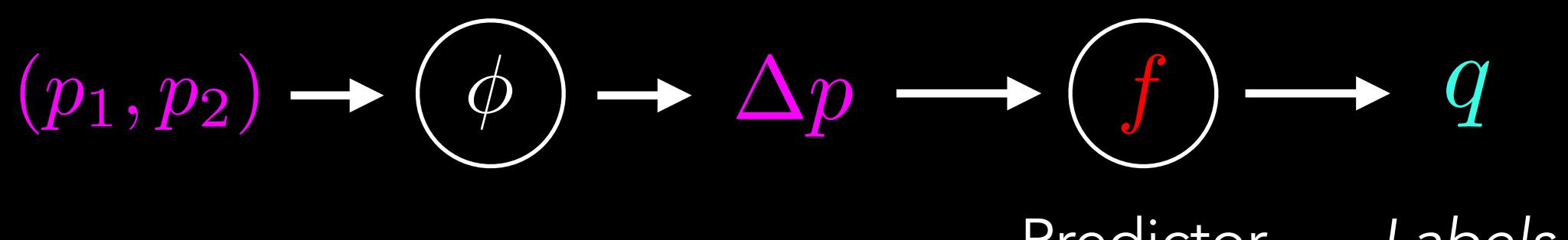
Training data: $\mathcal{D}_{\text{train}}$

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?

Feature extractor

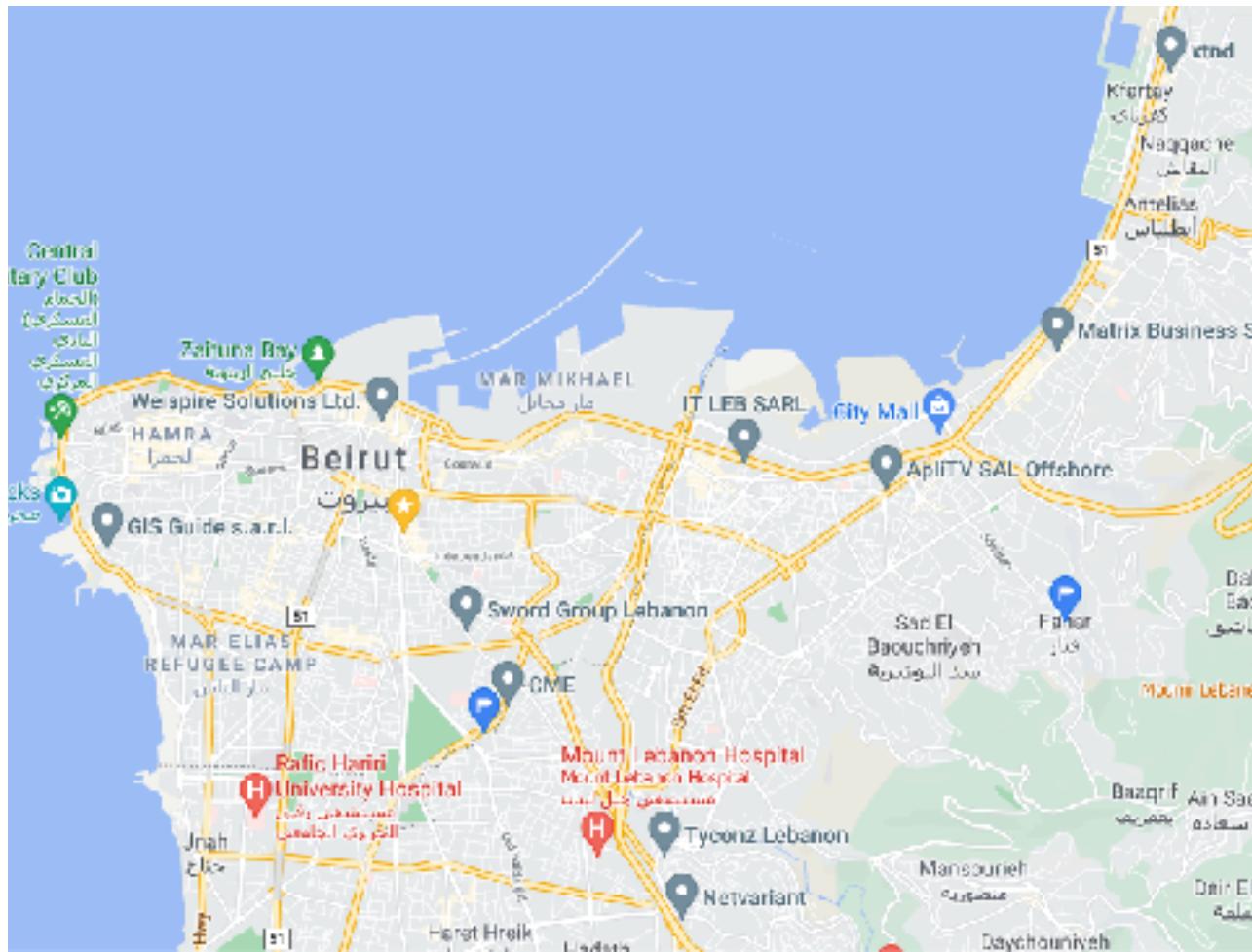


Predictor Labels

↓ Training

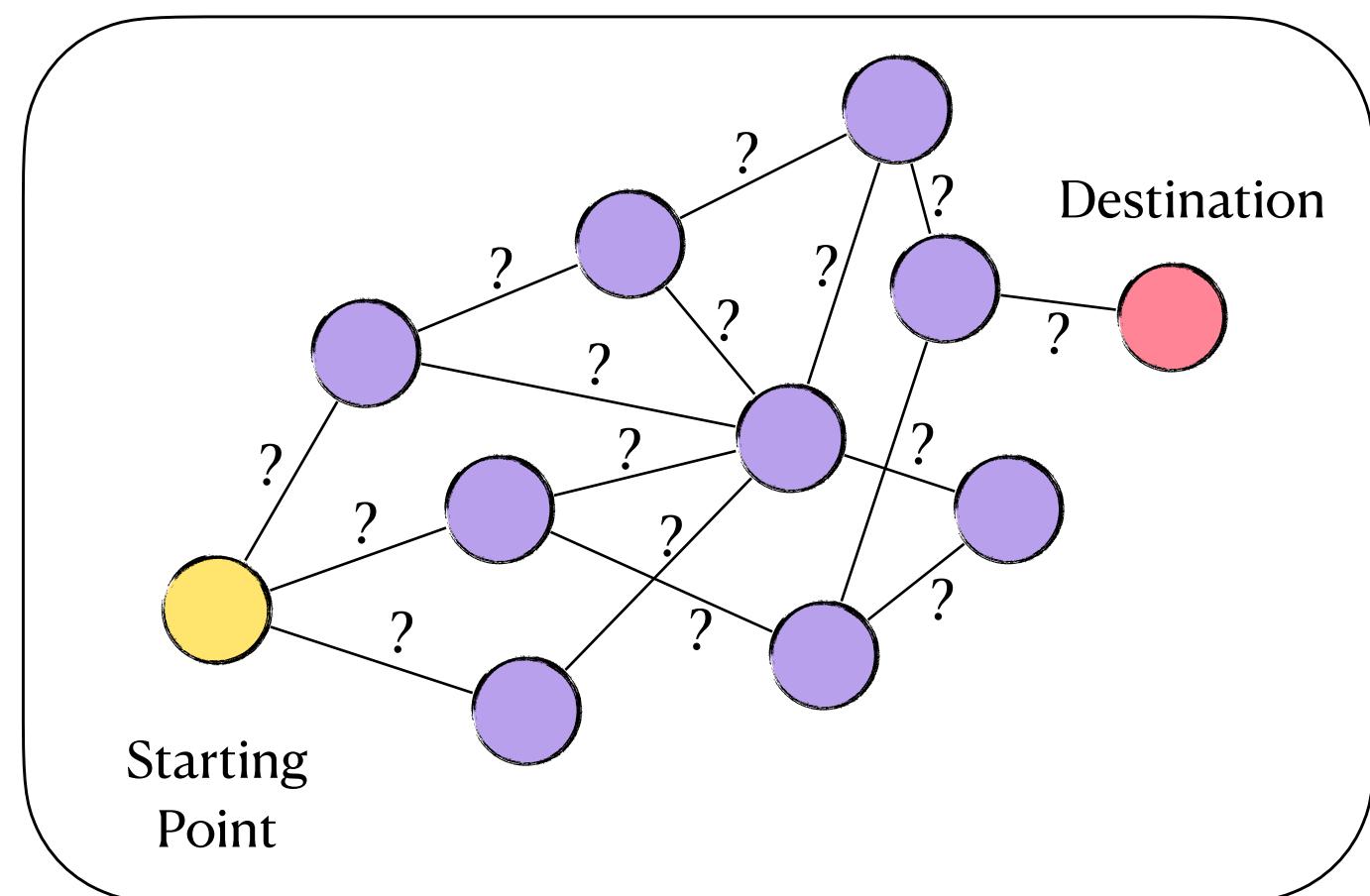
The 3 pillars of Artificial Intelligence

Real world

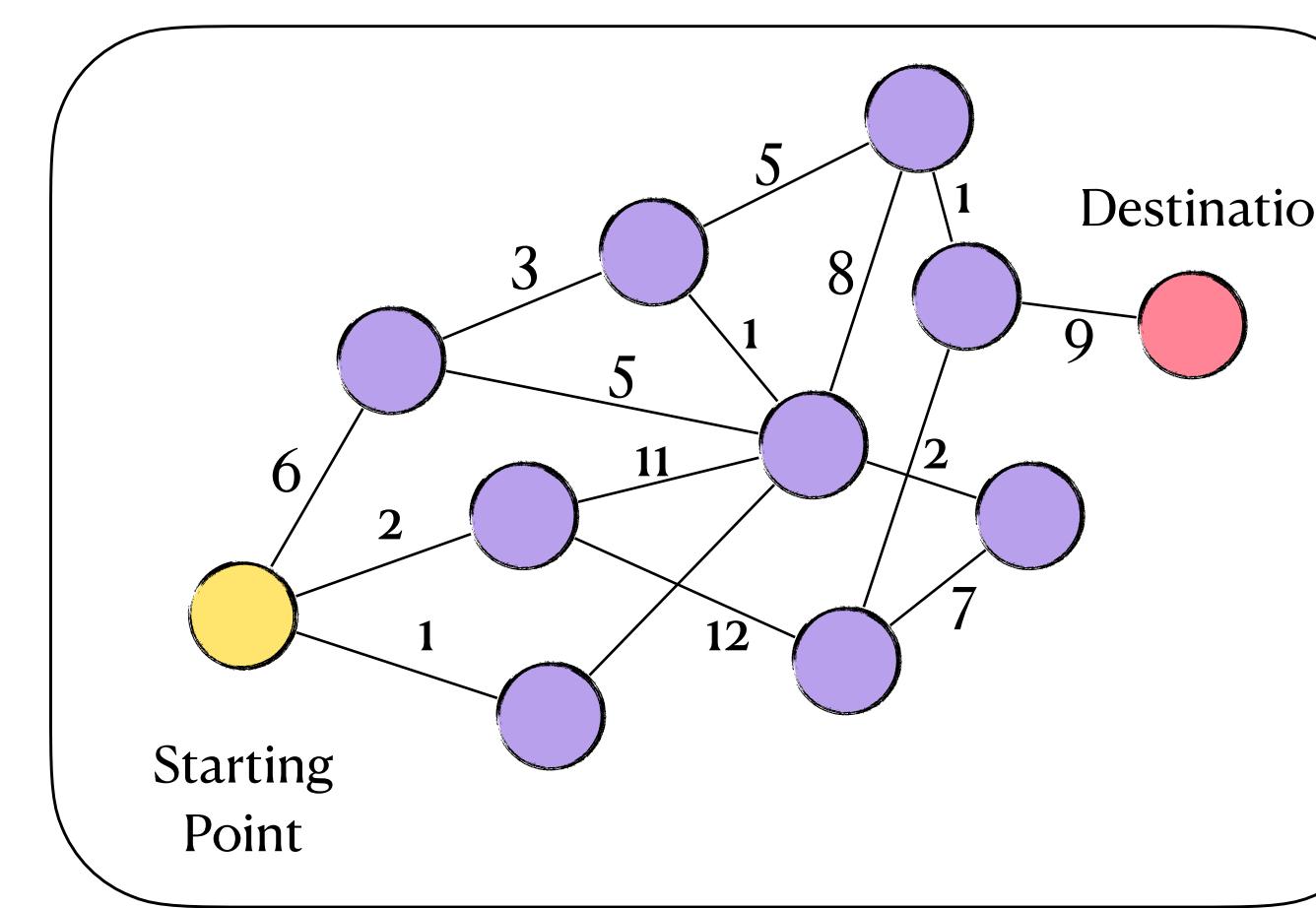


What's the shortest path
from AUB to my home?

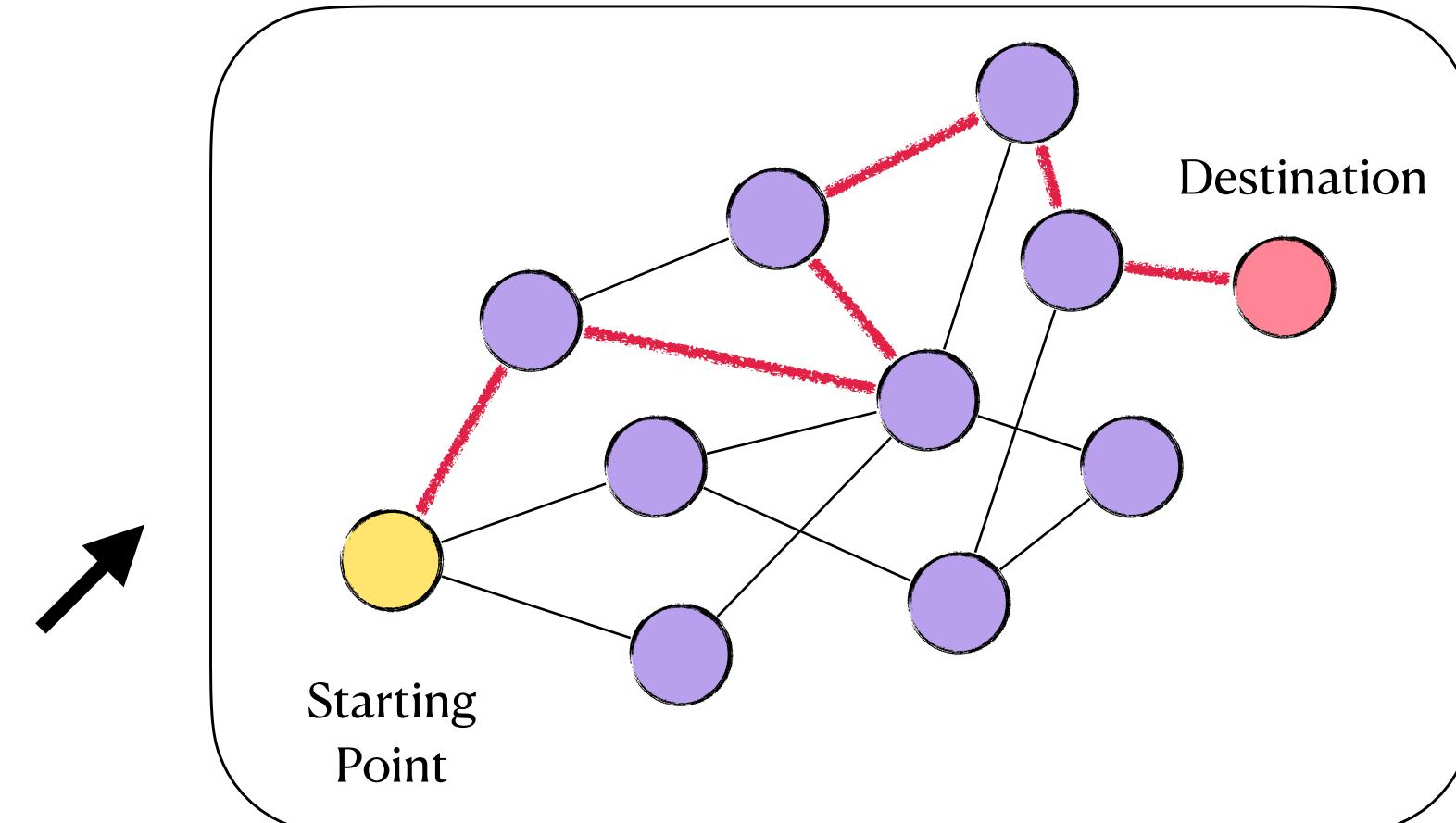
Learning



Modeling



Inference



The 3 pillars of Artificial Intelligence

Real world



Learning

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

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

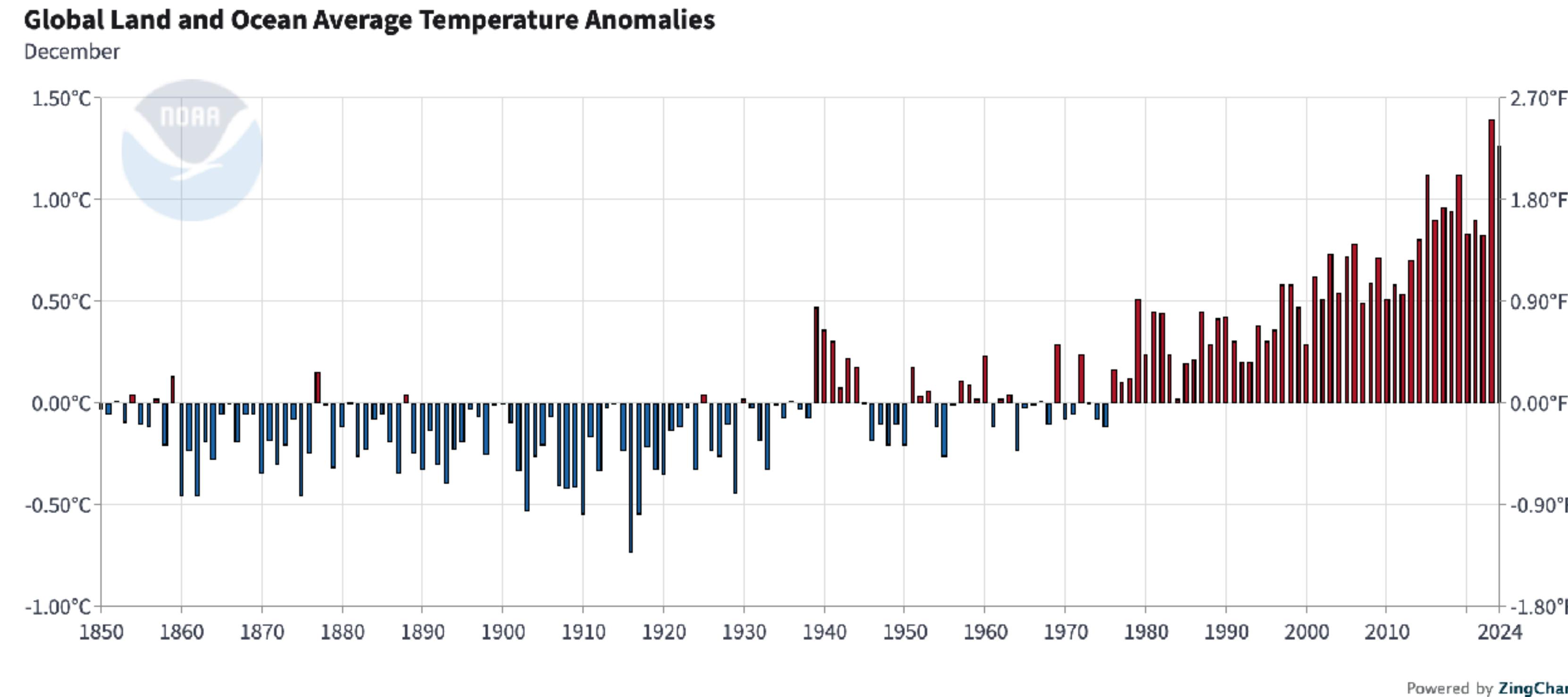
Modeling

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

Inference

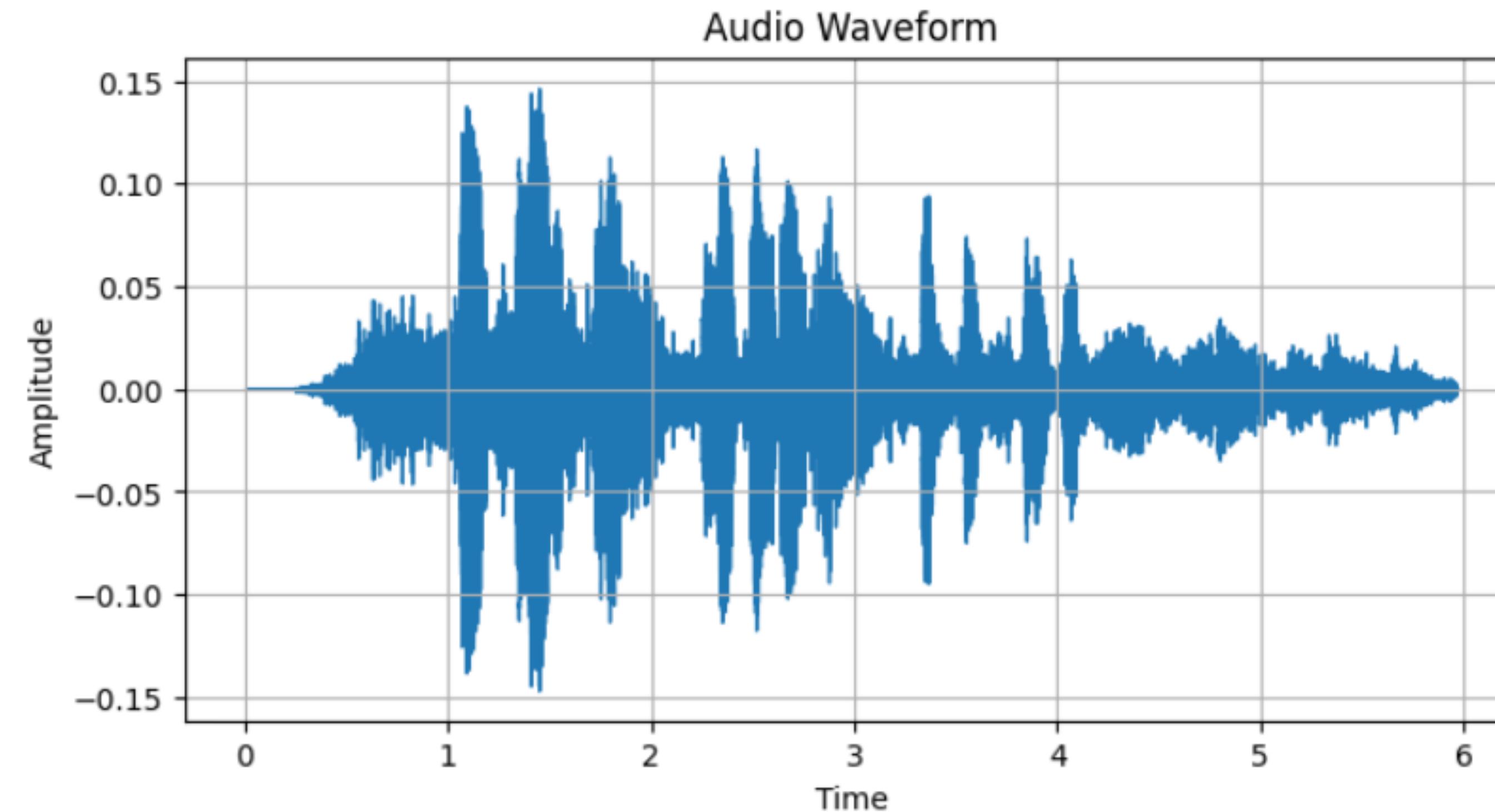
What are* Data?

Stocks



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



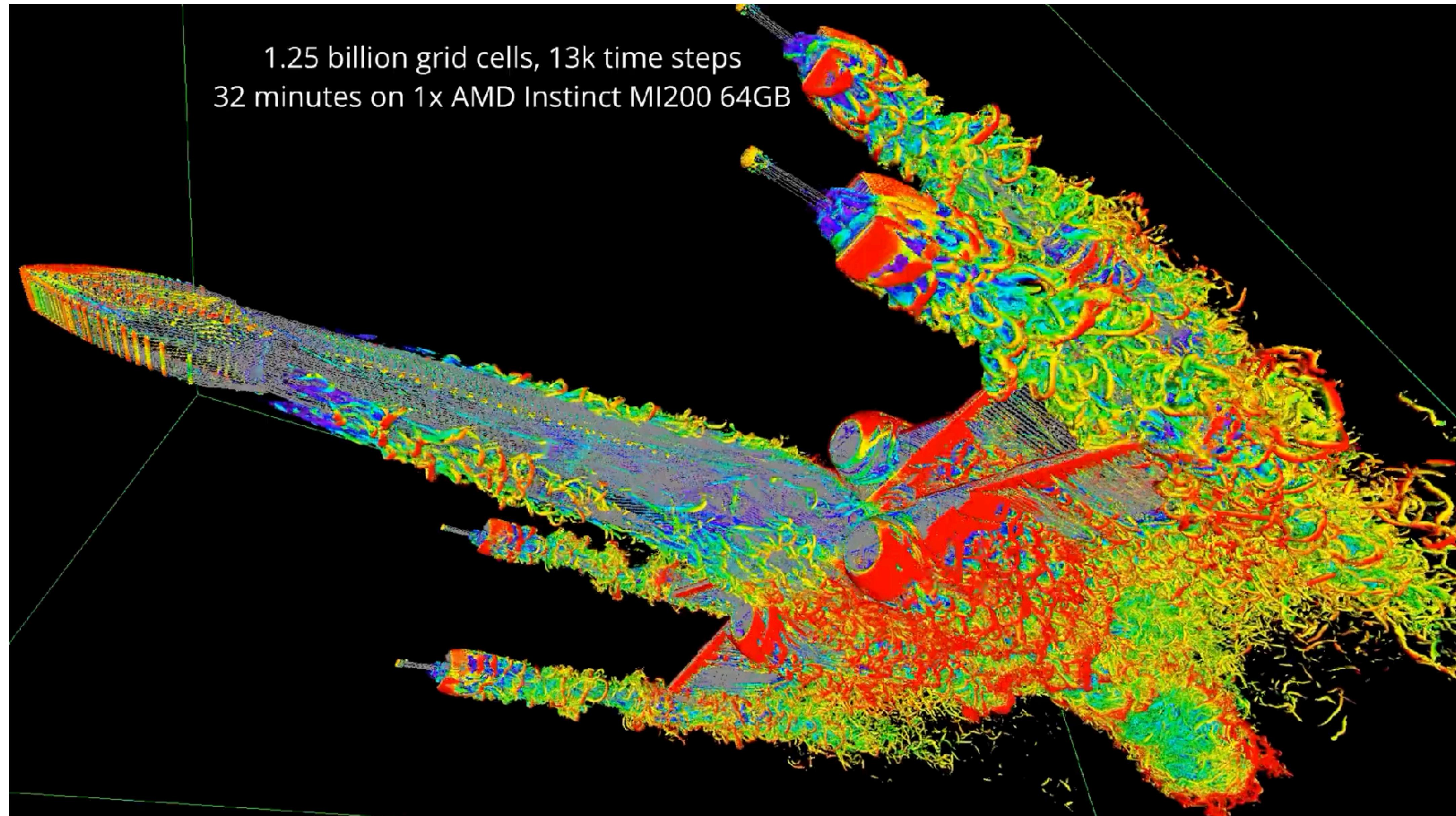
t_1	123	4	6	43
123	56	35	44	33 255
35	65	2	0	0
43	33	54	0	0
65	2	0	55	255
33	54	0	0	

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



Text

One-hot word representation

Dictionary

The World Is A Crazy Place

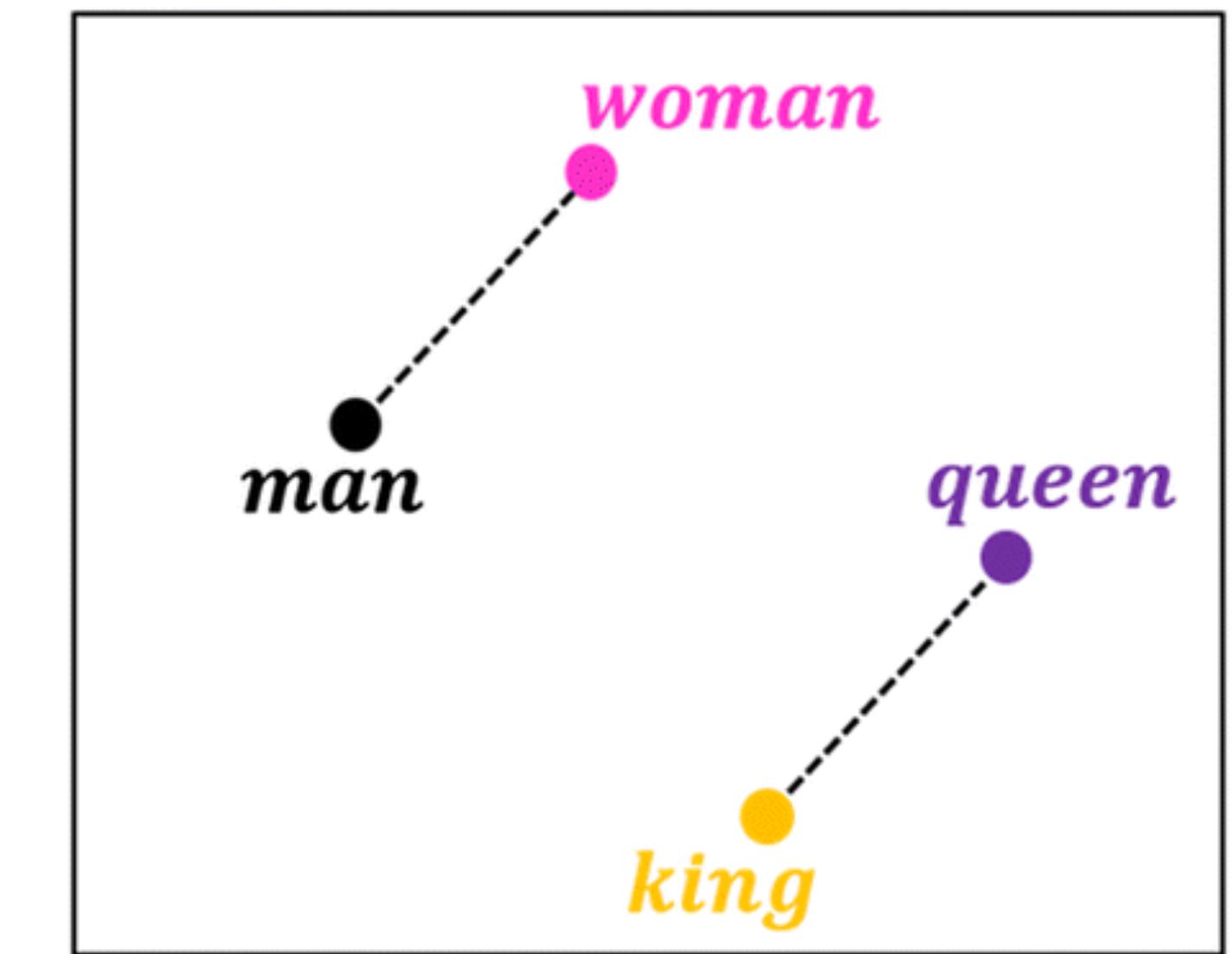
a
the →
world
crazy
place
is
:

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

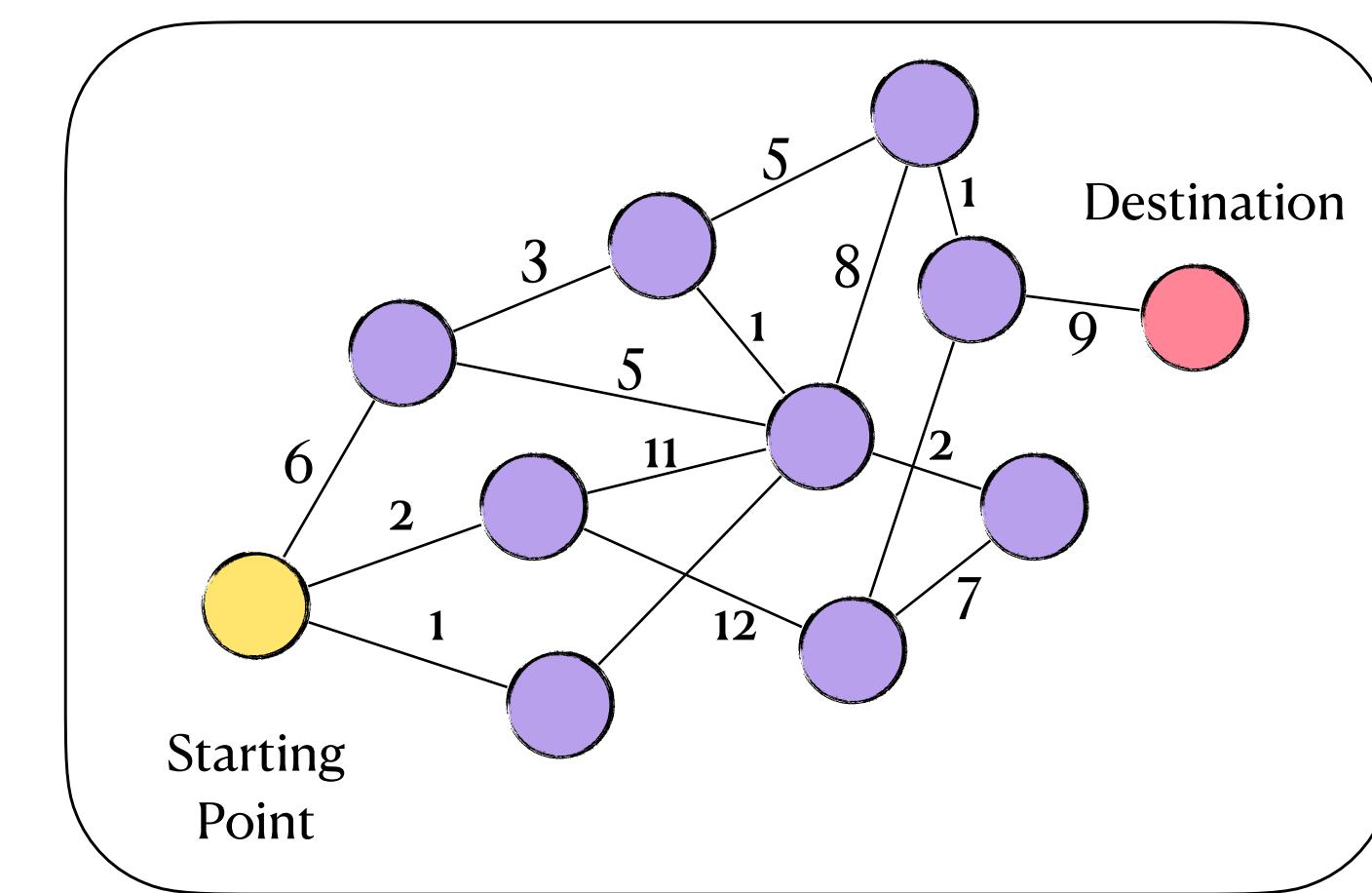
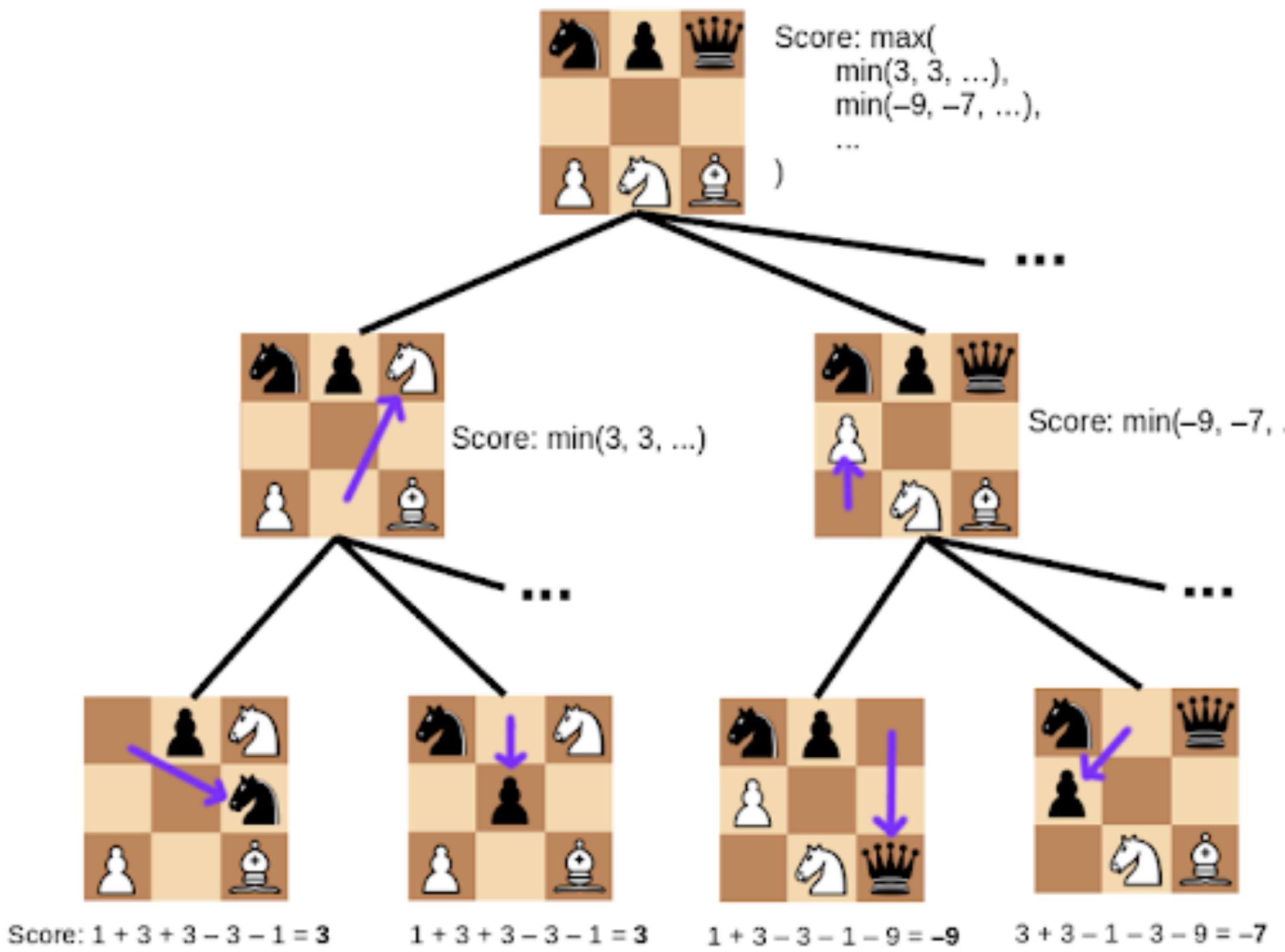
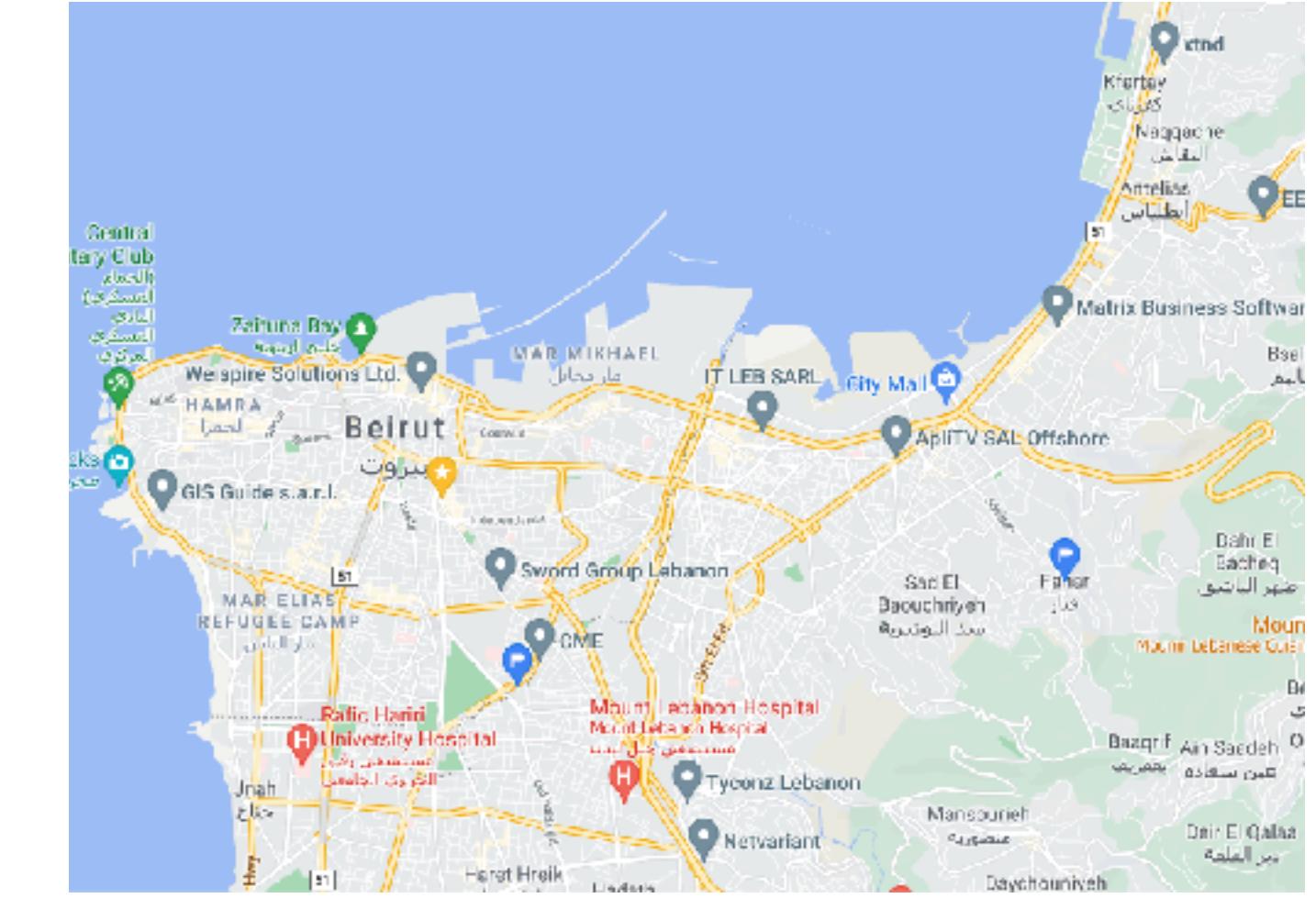
Word Embedding Representation

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]



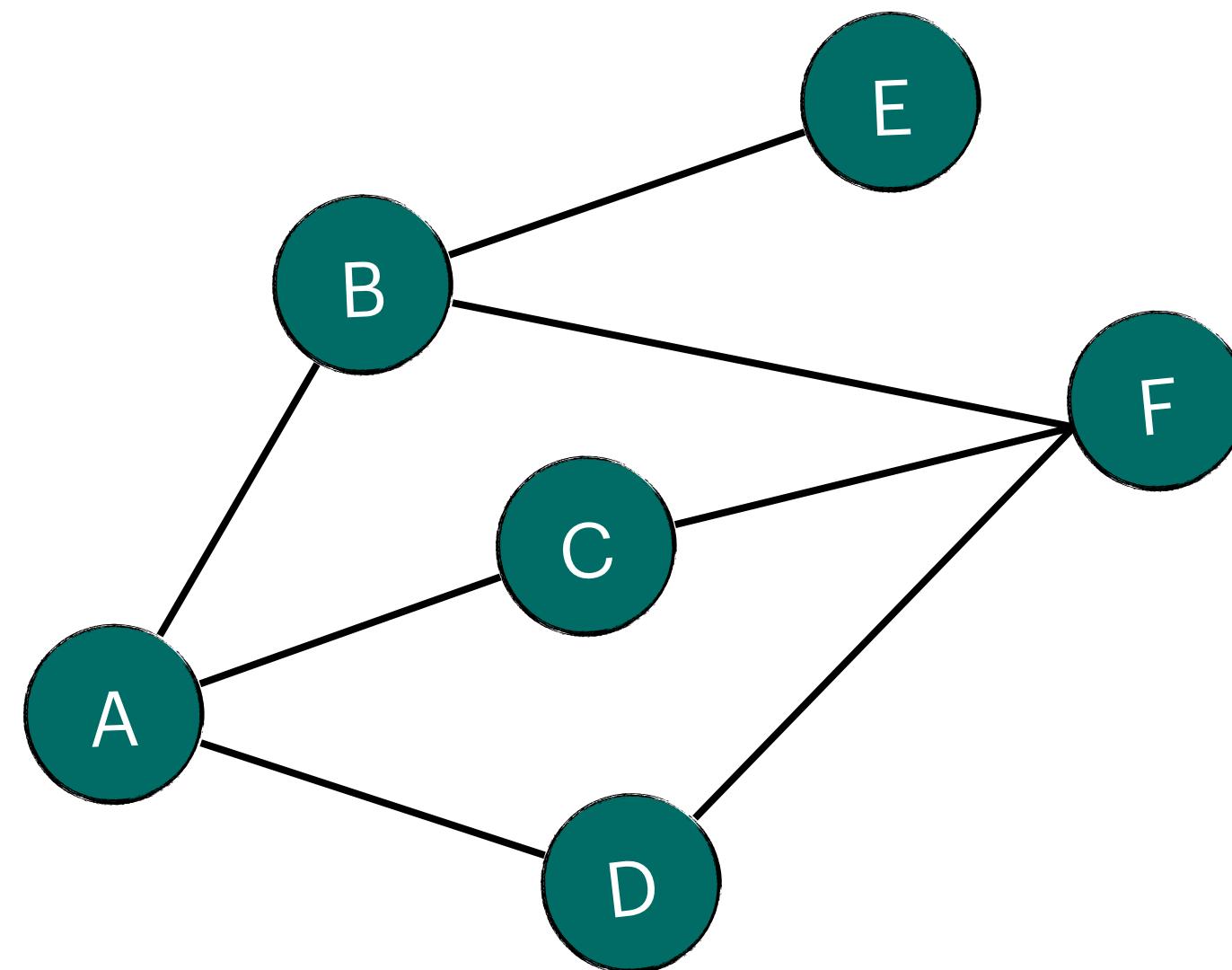
Graph Representation

Real world



Graph Representation

Graph with Vertices and Edges

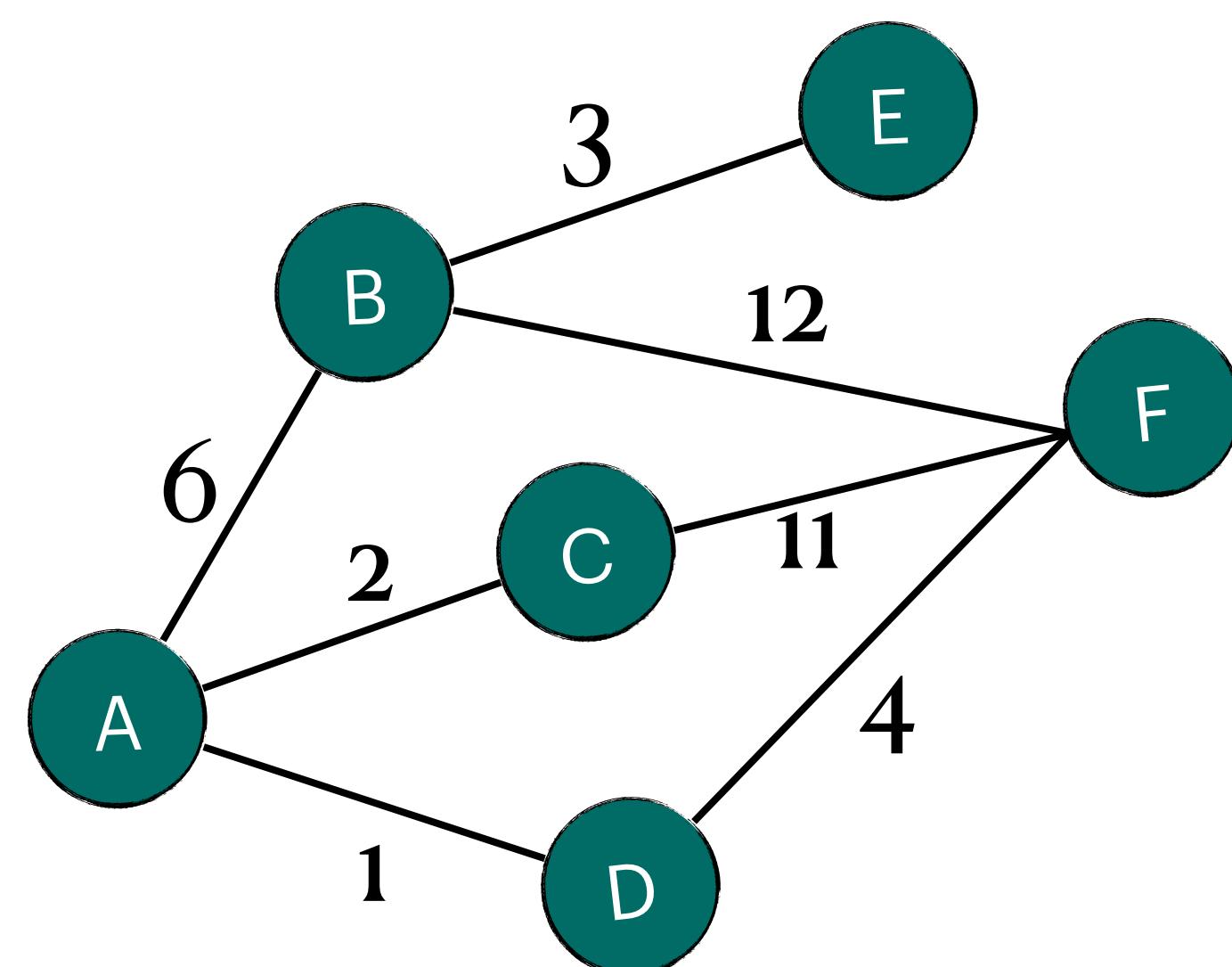


Adjacency Matrix

	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

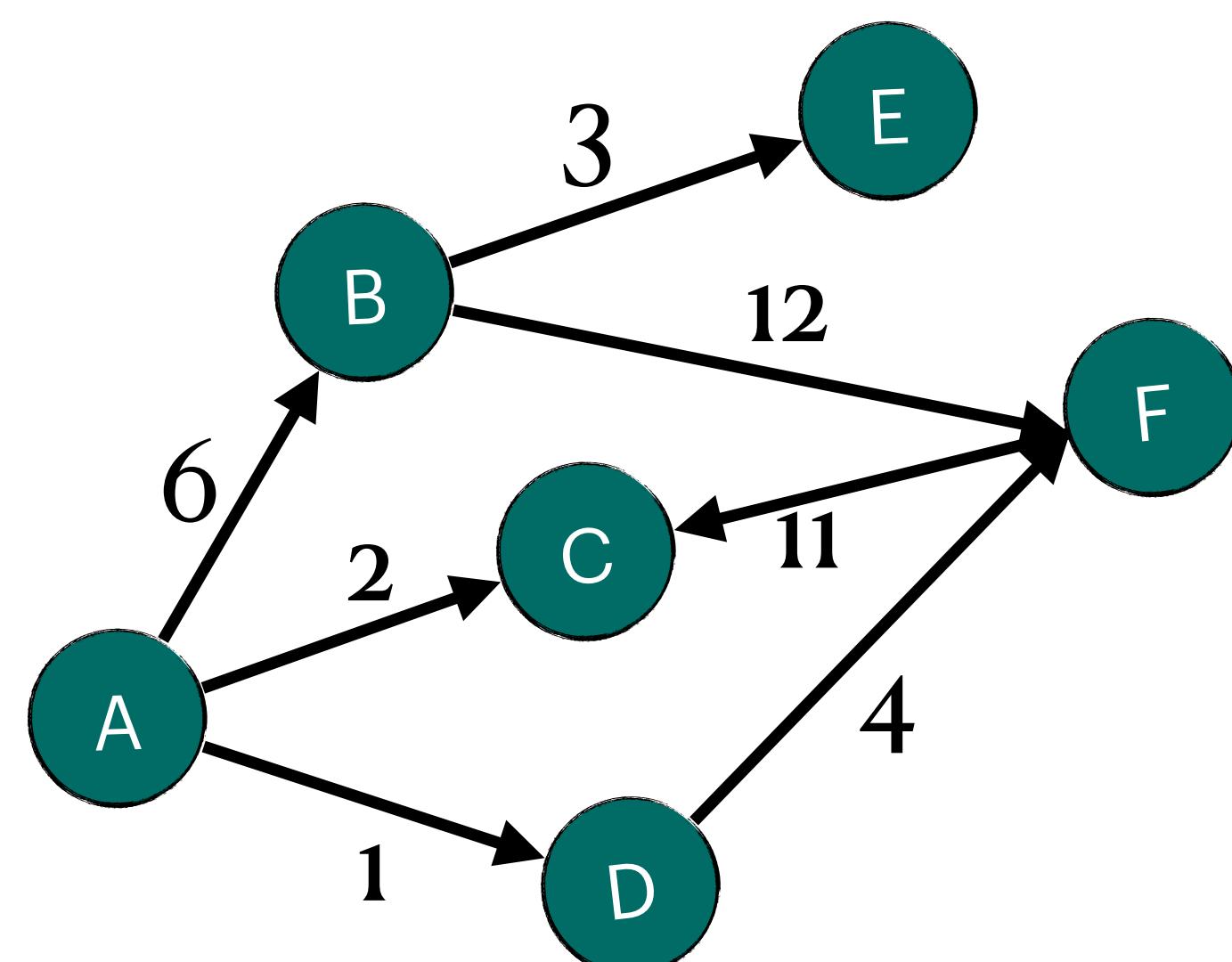


Adjacency Matrix

	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

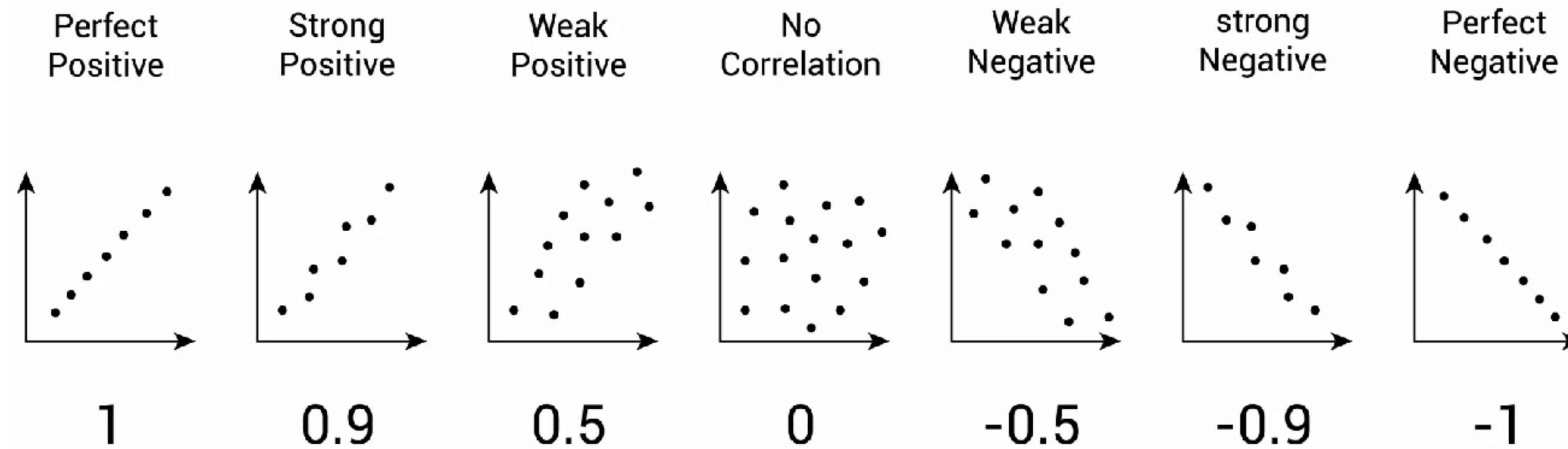


Adjacency Matrix

	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)

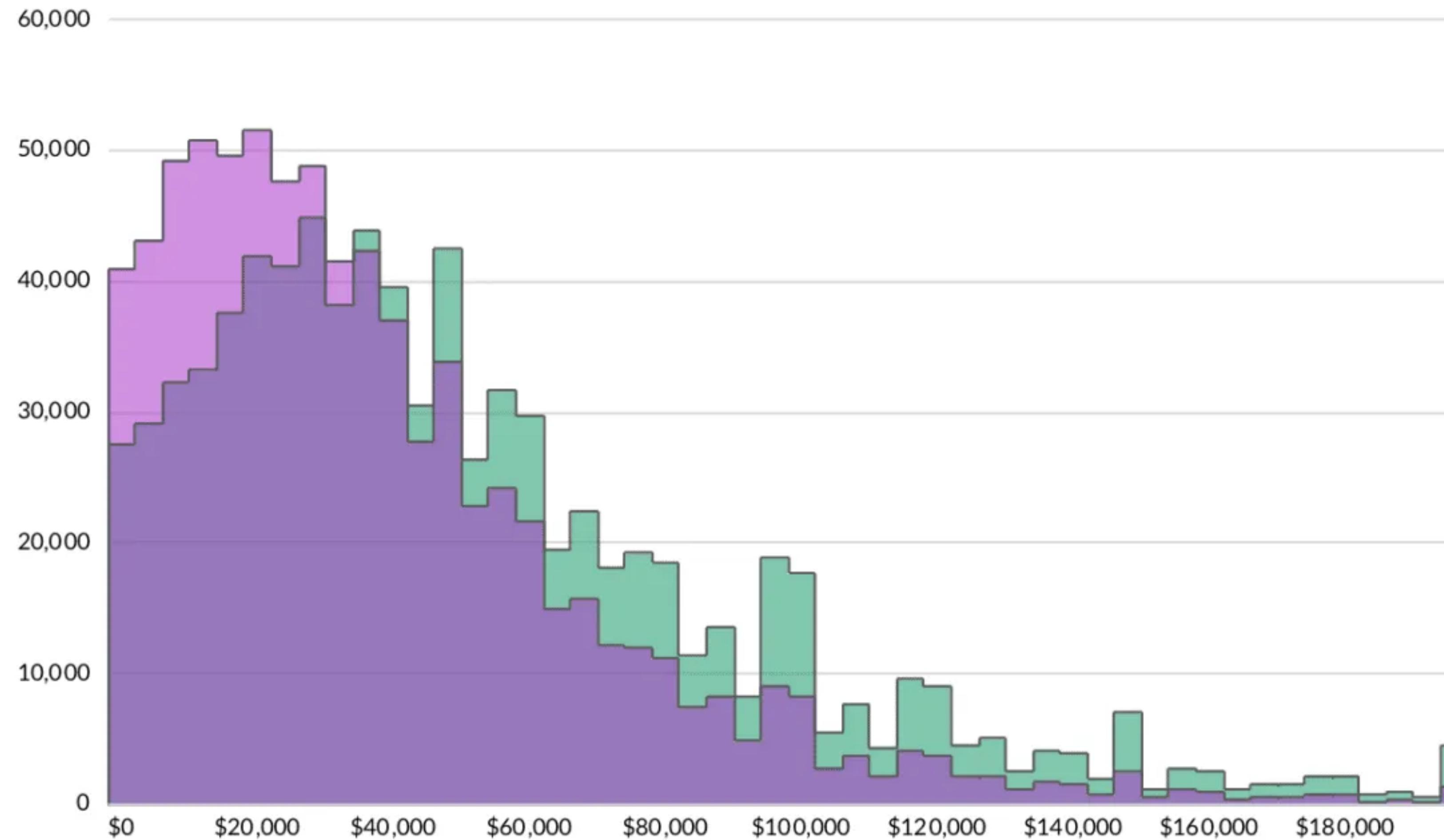
	Hours spent studying	Exam score	IQ score	Hours spent sleeping	School rating
Hours spent studying	1.00	0.82	0.48	-0.22	0.36
Exam score	0.82	1.00	0.33	-0.04	0.23
IQ score	0.08	0.33	1.00	0.06	0.02
Hours spent sleeping	-0.22	-0.04	0.06	1.00	0.12
School rating	0.36	0.23	0.02	0.12	1.00

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

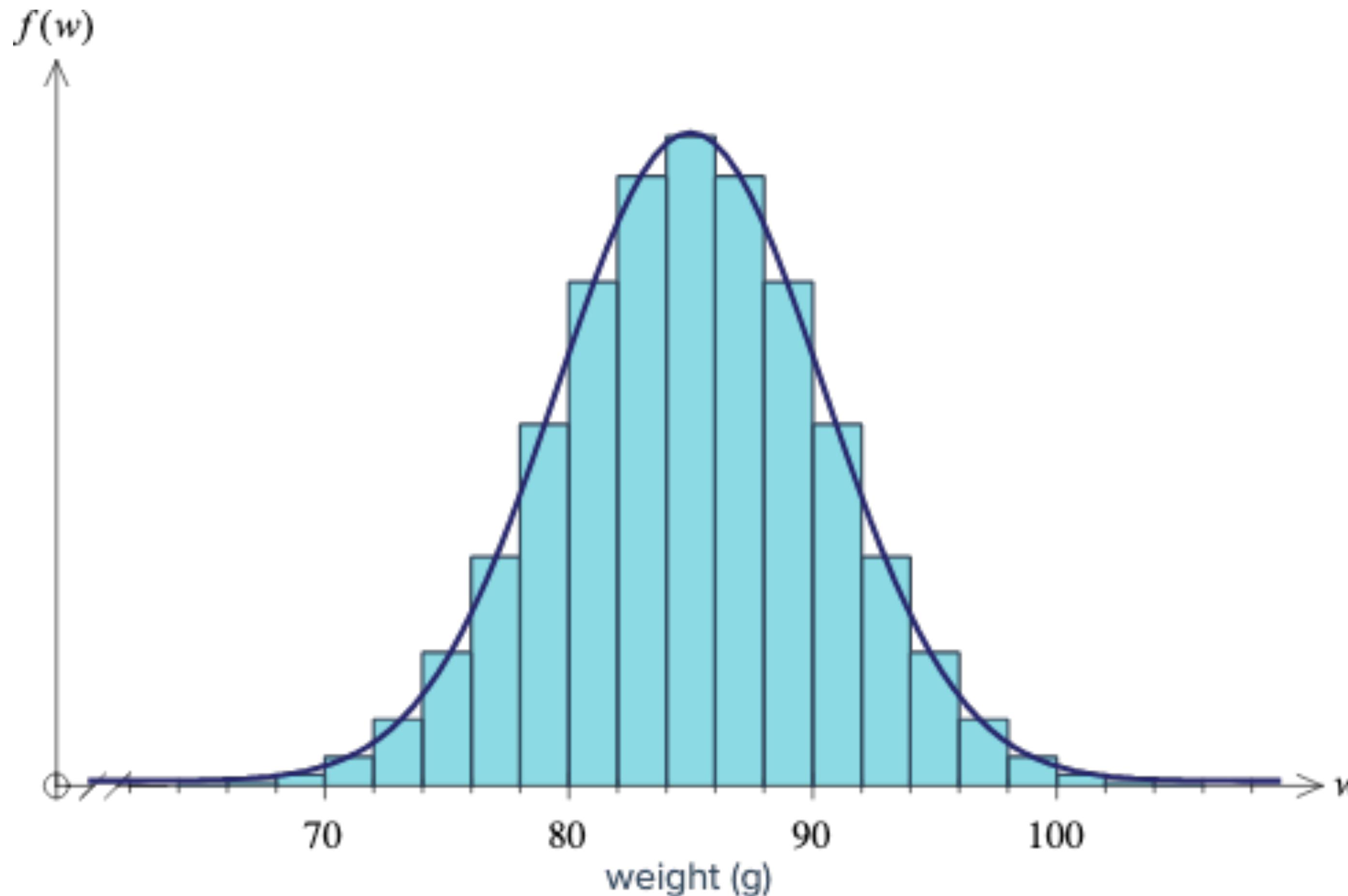
$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & r_{nn} \end{pmatrix}$$

Histogram

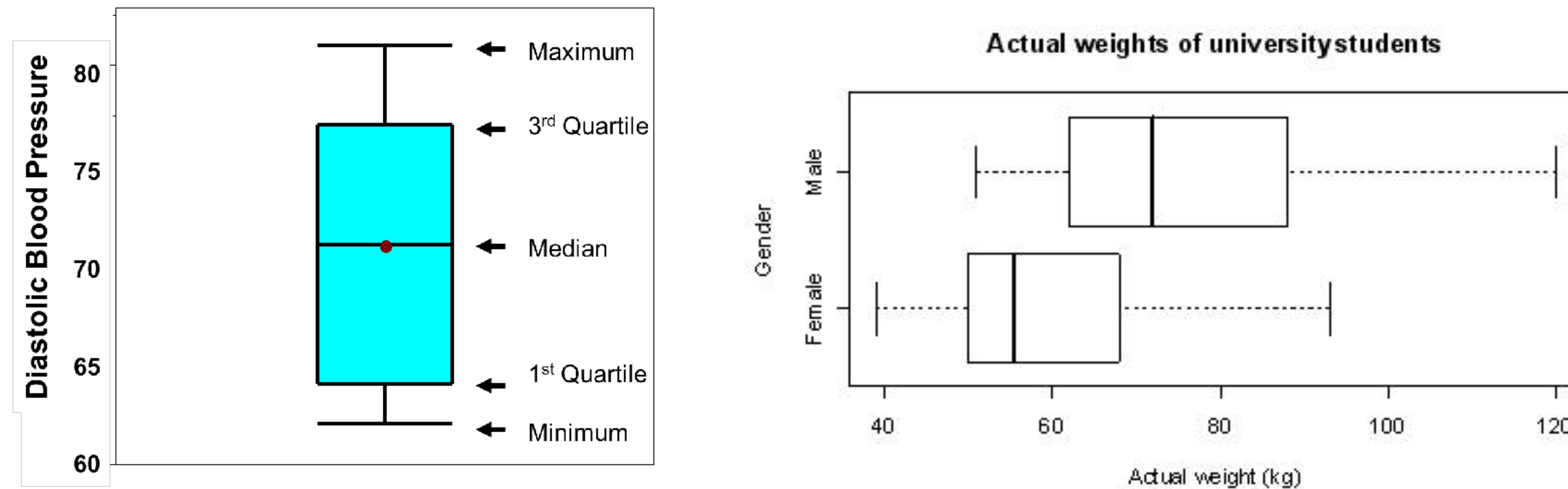
Distribution of Men's and Women's Incomes in 2016



Probability Density Function



Box and Whisker Plot



Discovering Functions From Data

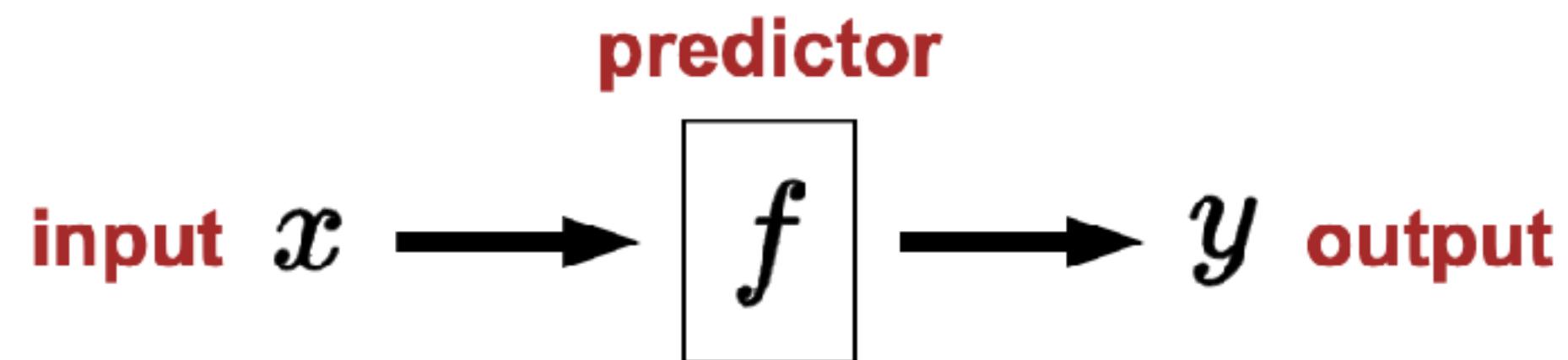
Machine Learning

$$f(x) = 2^x$$

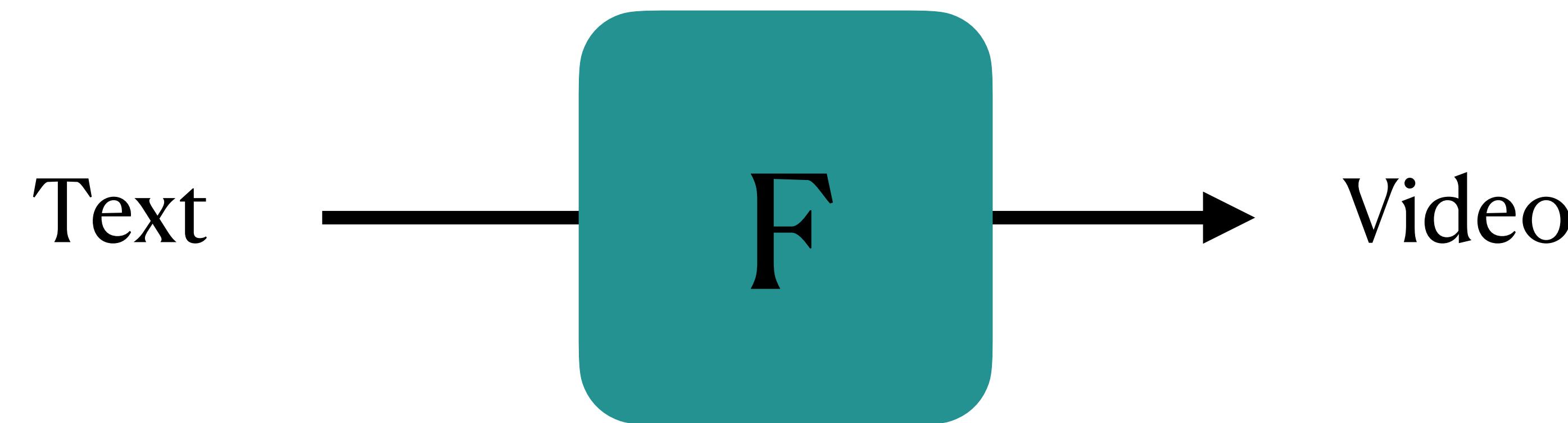
x	0, 1, 2, 3, 4, 5
y	1, 2, 4, 8, 16, ?

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

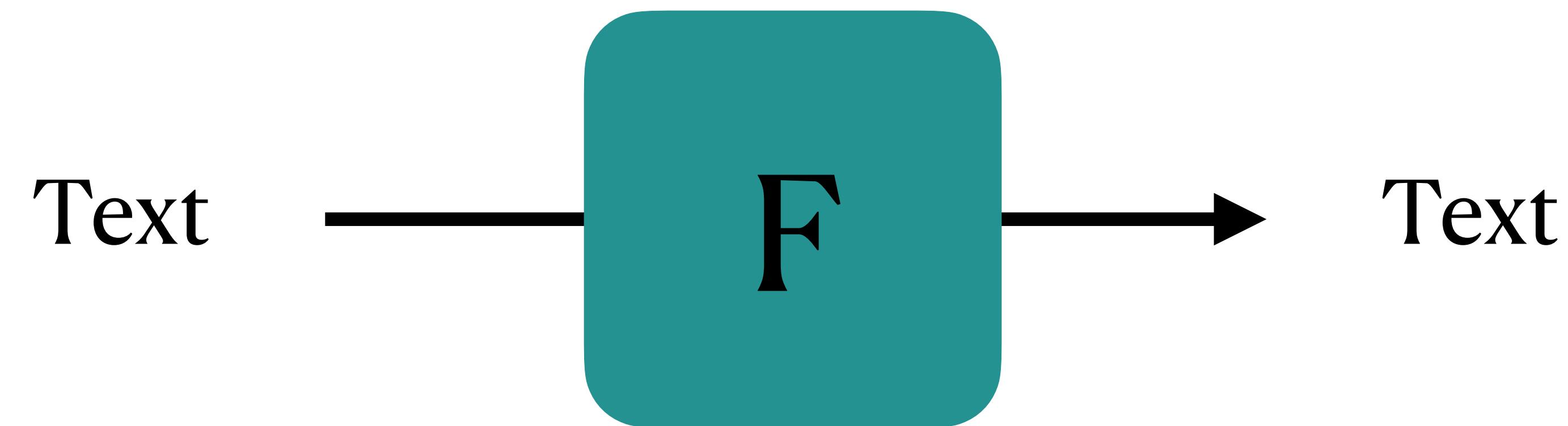
$$f(x) = 2x$$



Sora - Text-to-Video

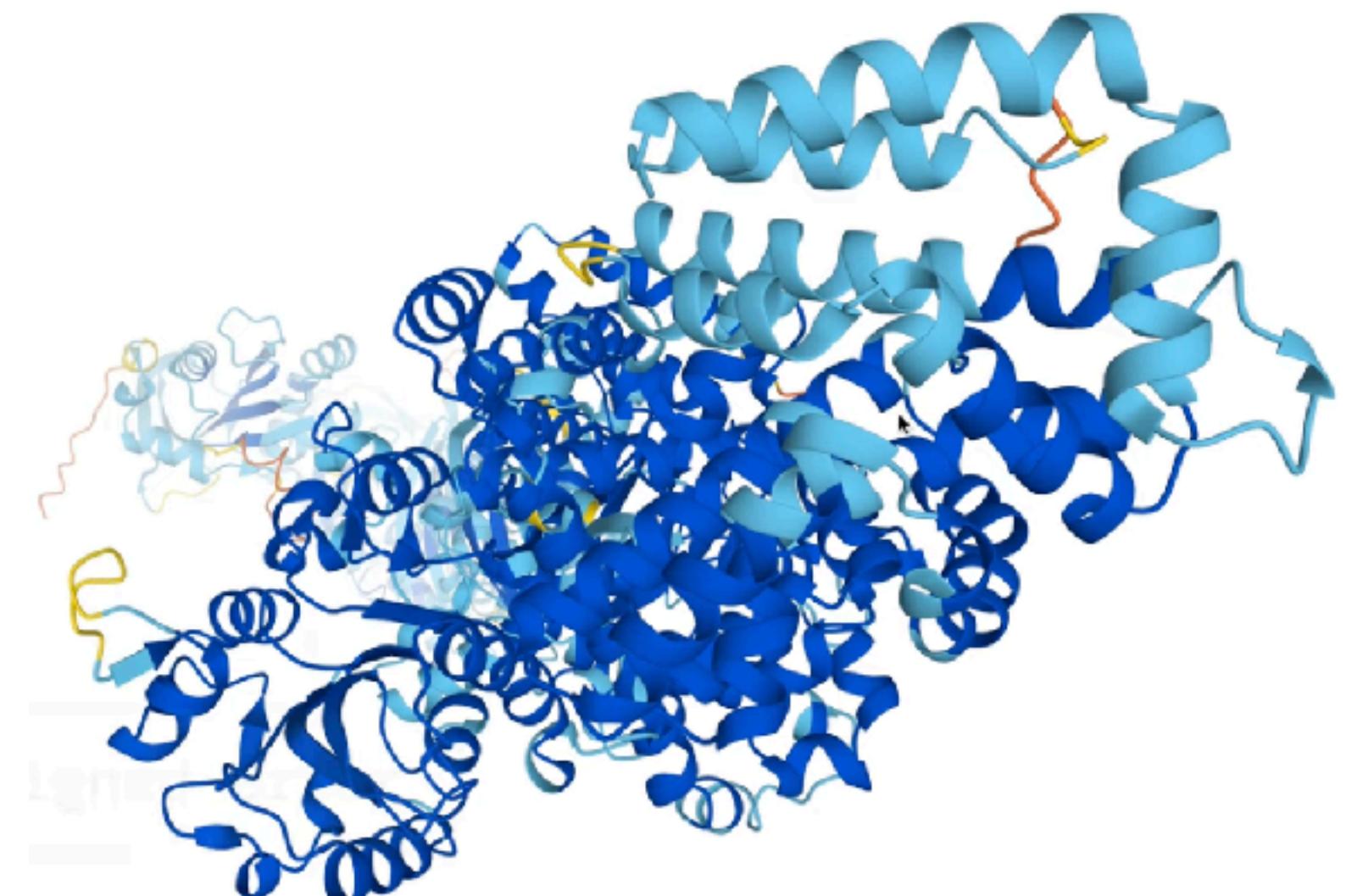
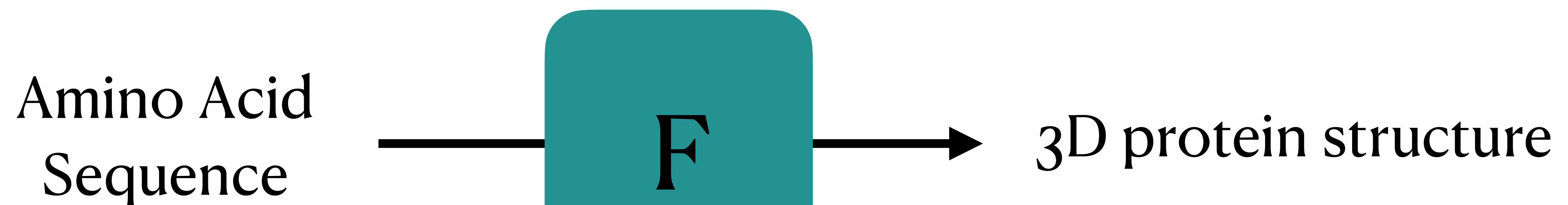


ChatGPT

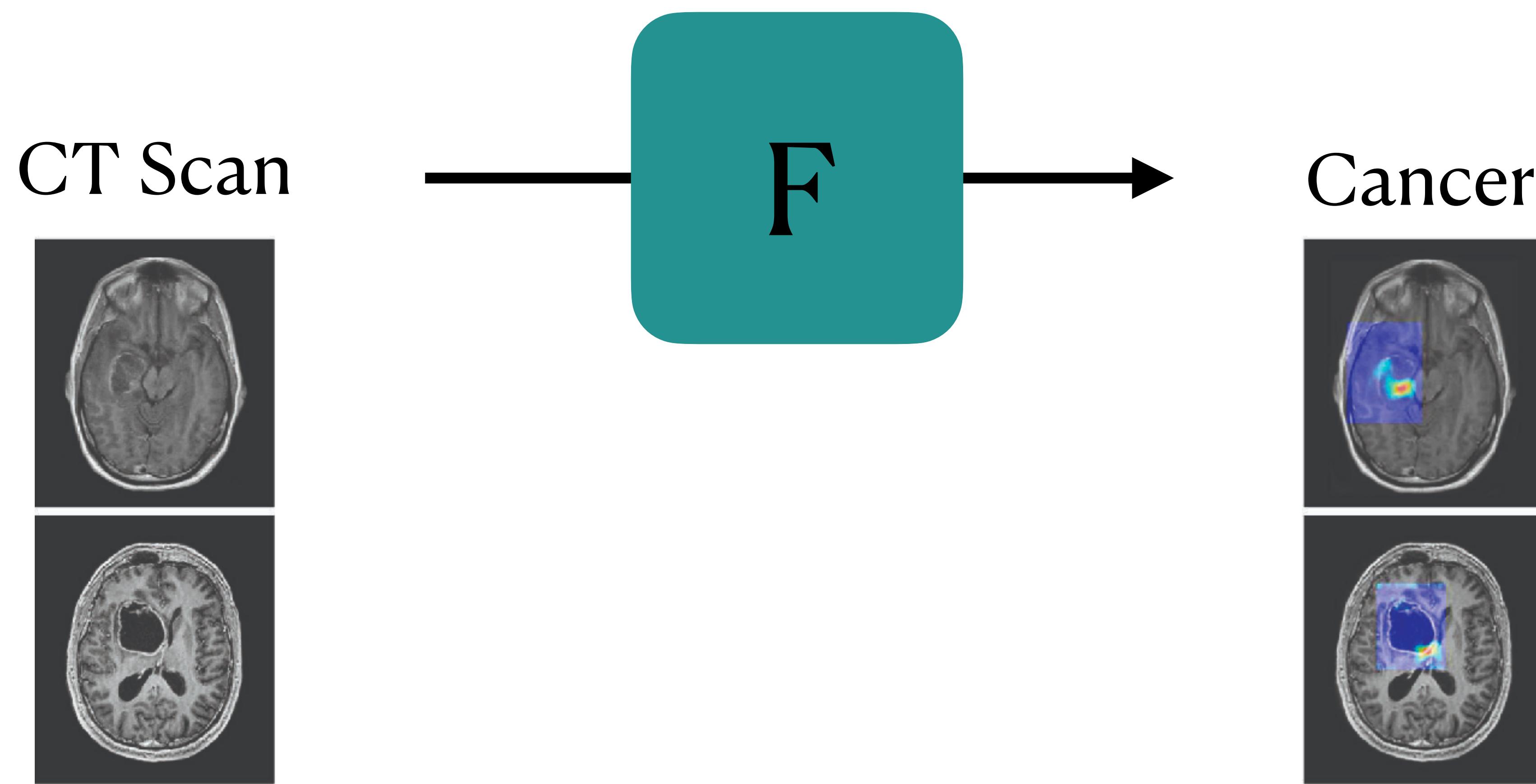


AlphaFold is an AI system developed by **DeepMind** that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

AlphaFold



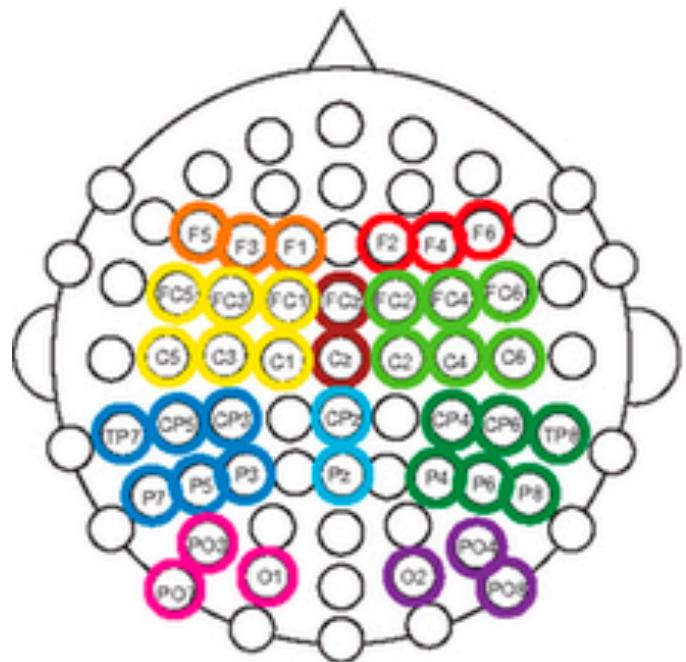
Healthcare



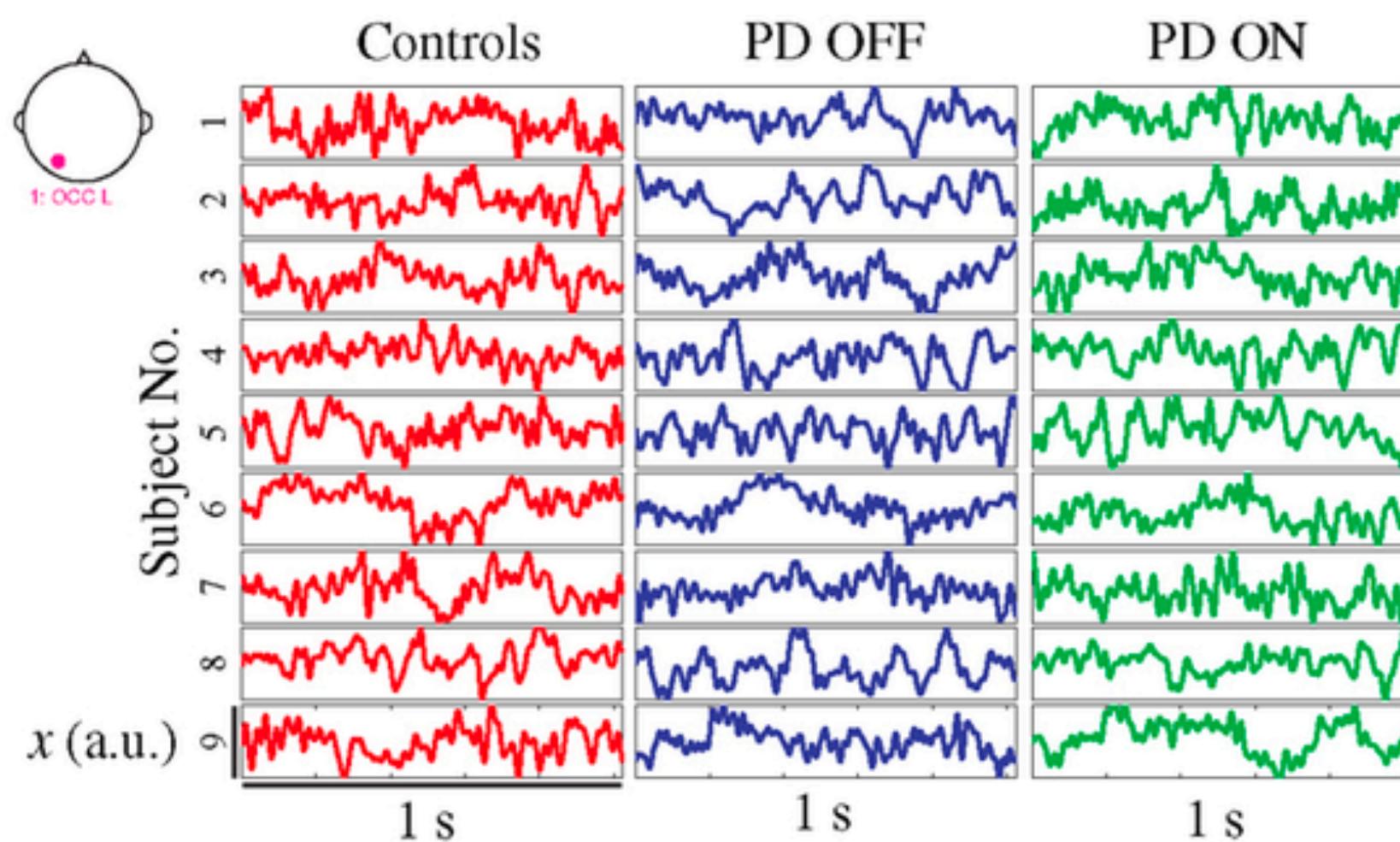
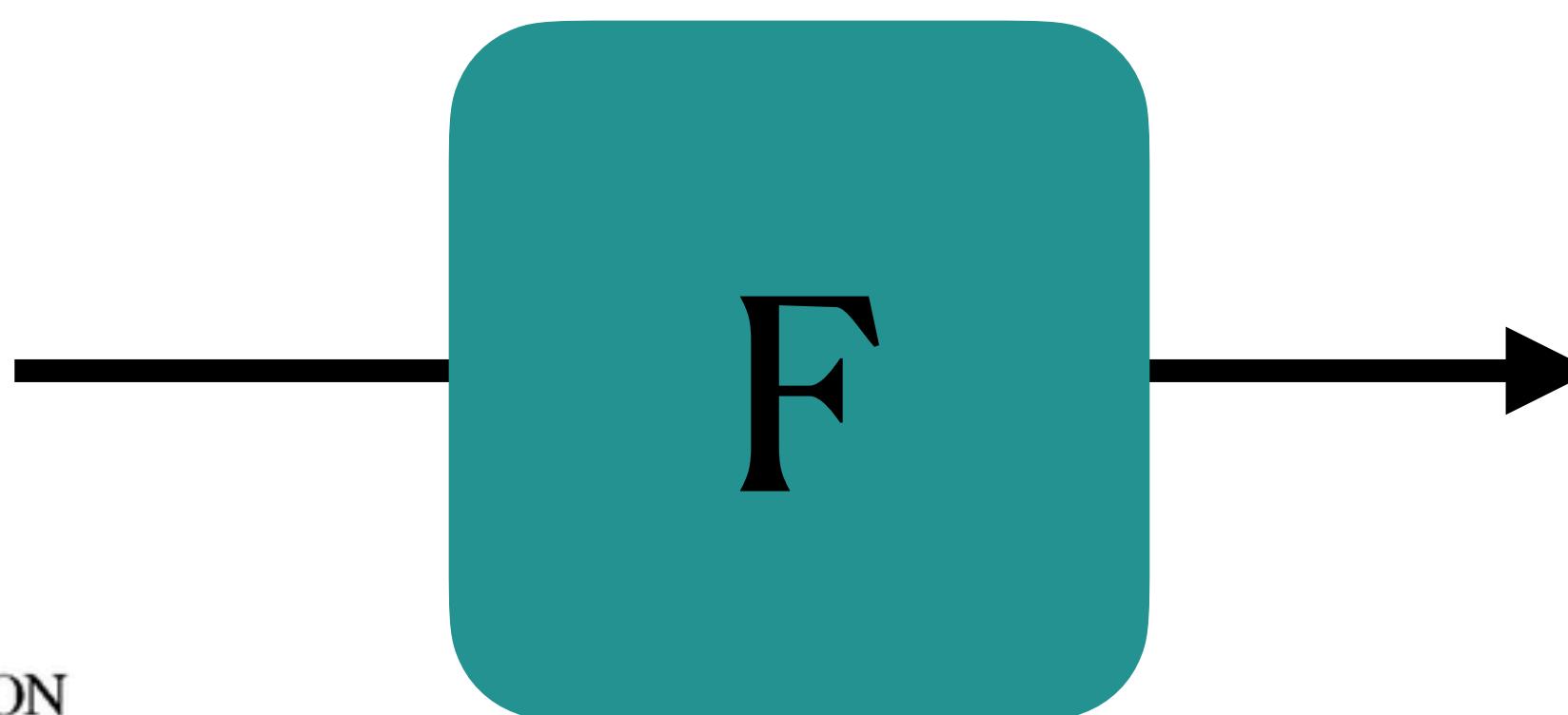
Healthcare



Neuroscience

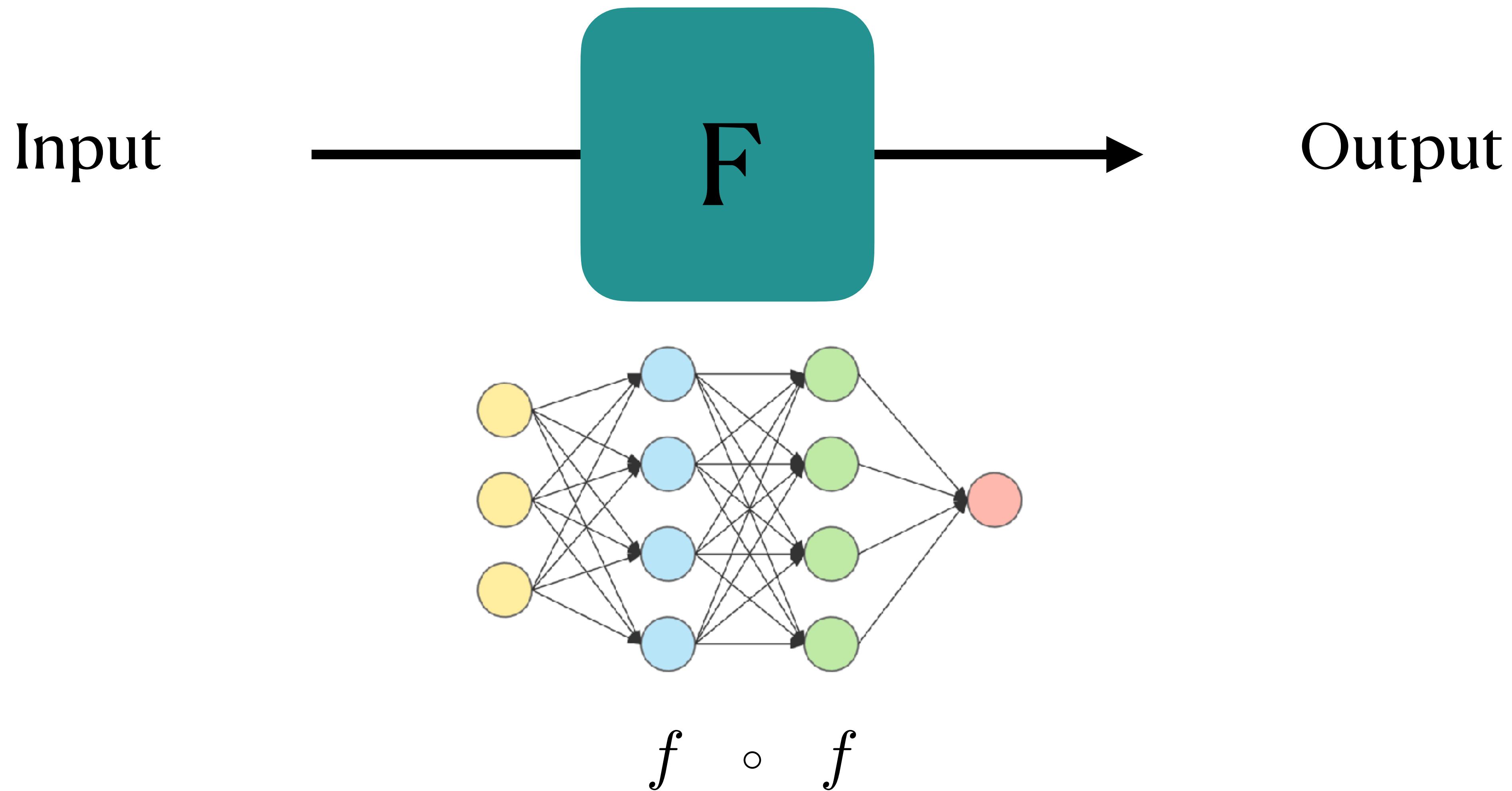


EEG



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.

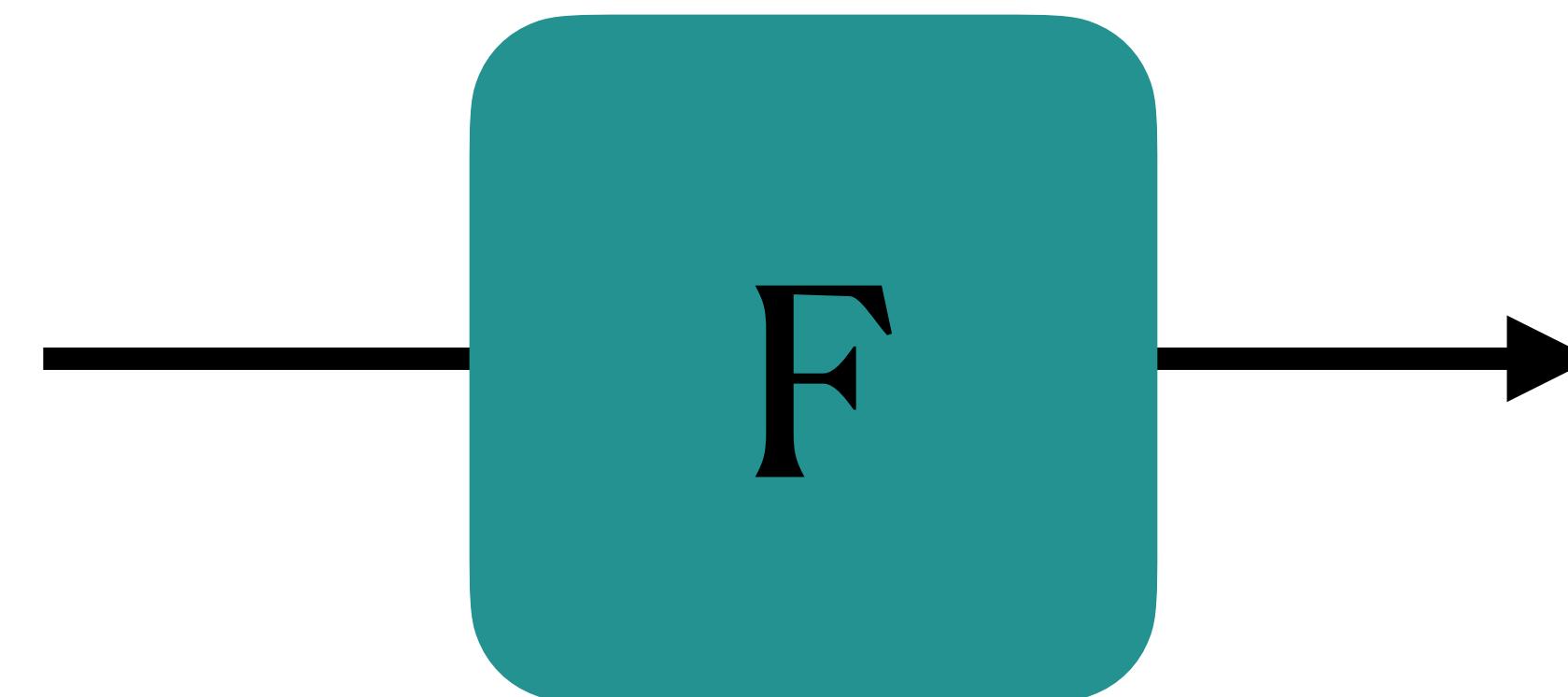
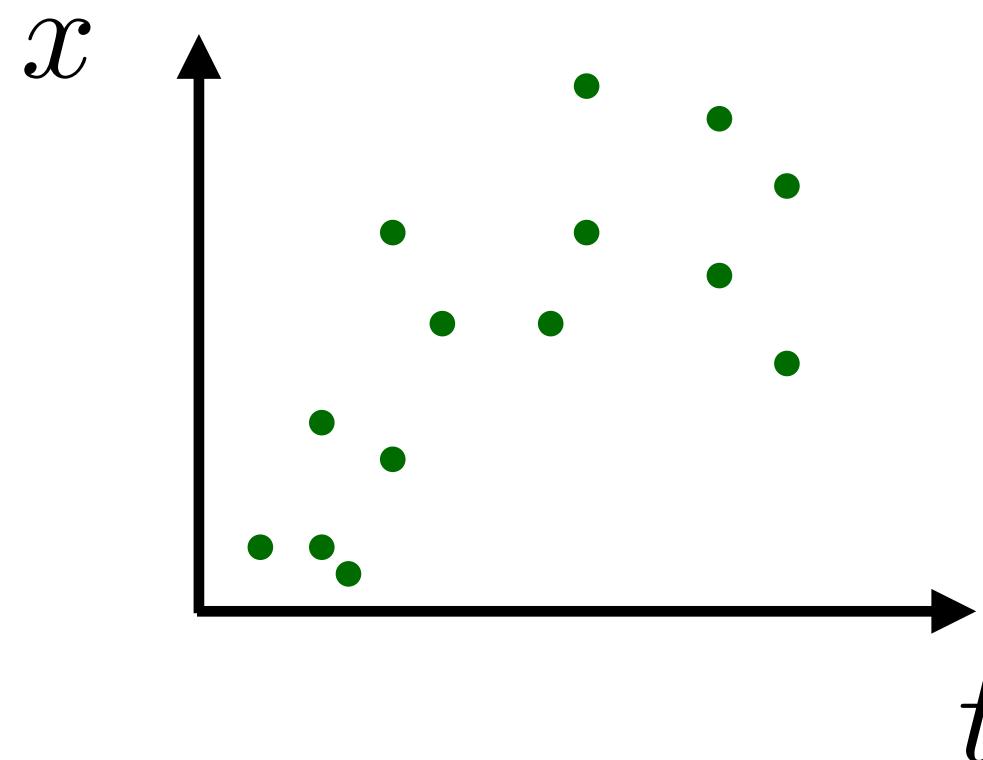
Neural Networks



Applications in Scientific Computing

Symbolic Regression

Time series



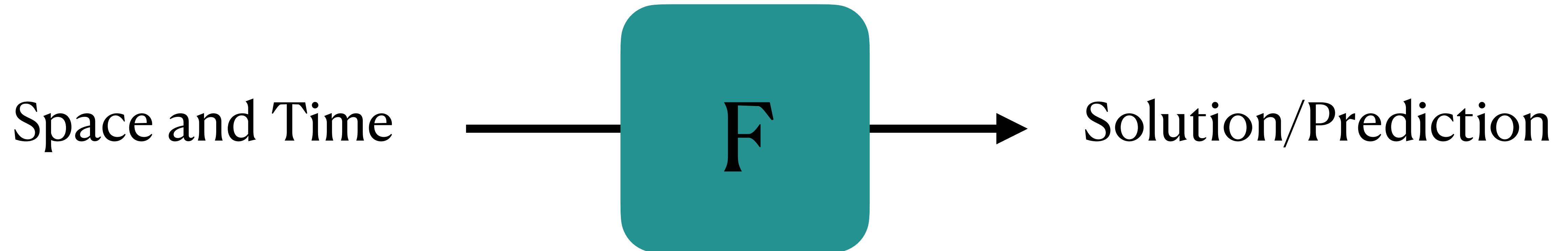
+ Physical Constraints
e.g. Conservation of Energy

???

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

Applications in Scientific Computing

Physics Informed Neural Networks

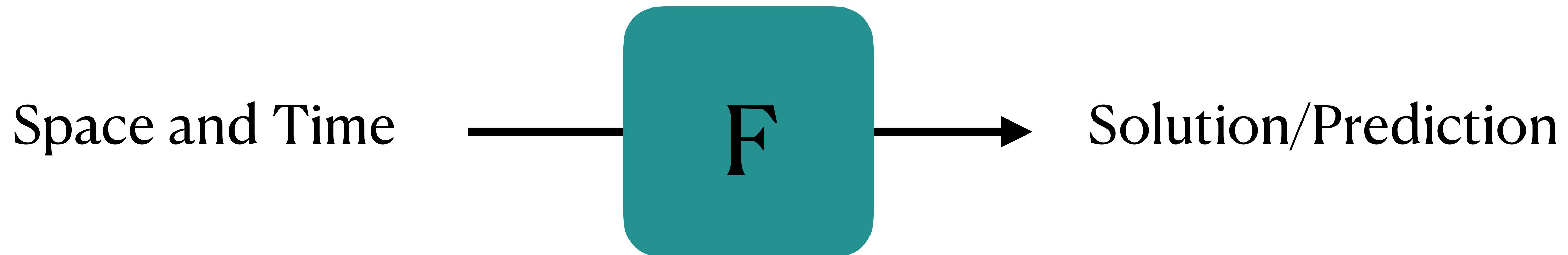


???

A decoder-only foundation model for time-series forecasting

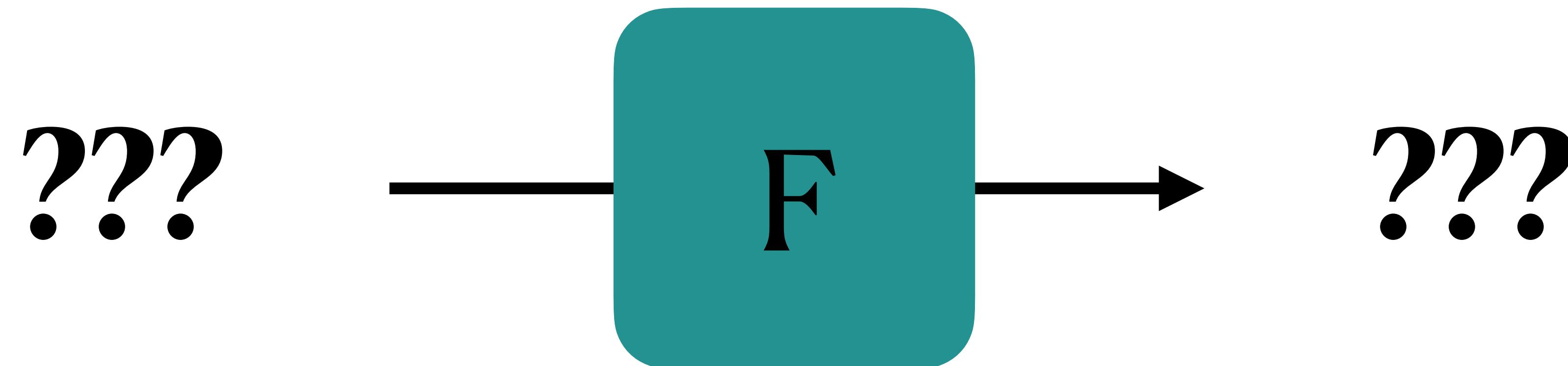
FRIDAY, FEBRUARY 02, 2024

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?



+ Constraints

???

Laws are linear

Pascal's law (1653)



Hooke's law (1678)



Newton's law of viscosity (1701)



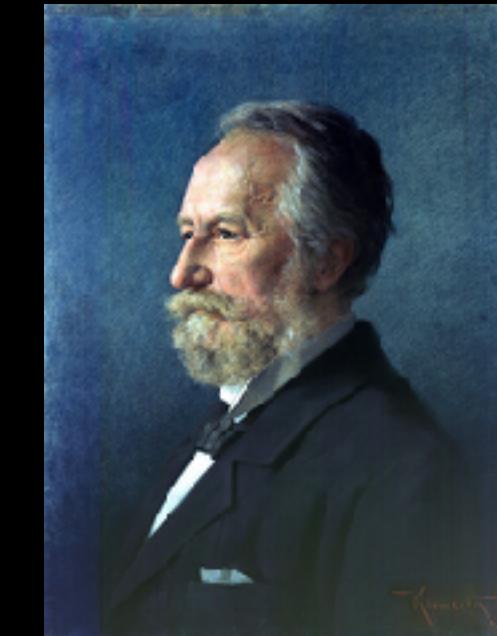
Ohm's law (1781)



Fourier's law (1822)



Fick's law (1855)



Darcy's law (1856)



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

$$F = -kx$$

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

$$I = V/R$$

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

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

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

Ideal gas law (1834)

Amonton's law (1808)

Charles's law (1787)

Boyle's law (1662)

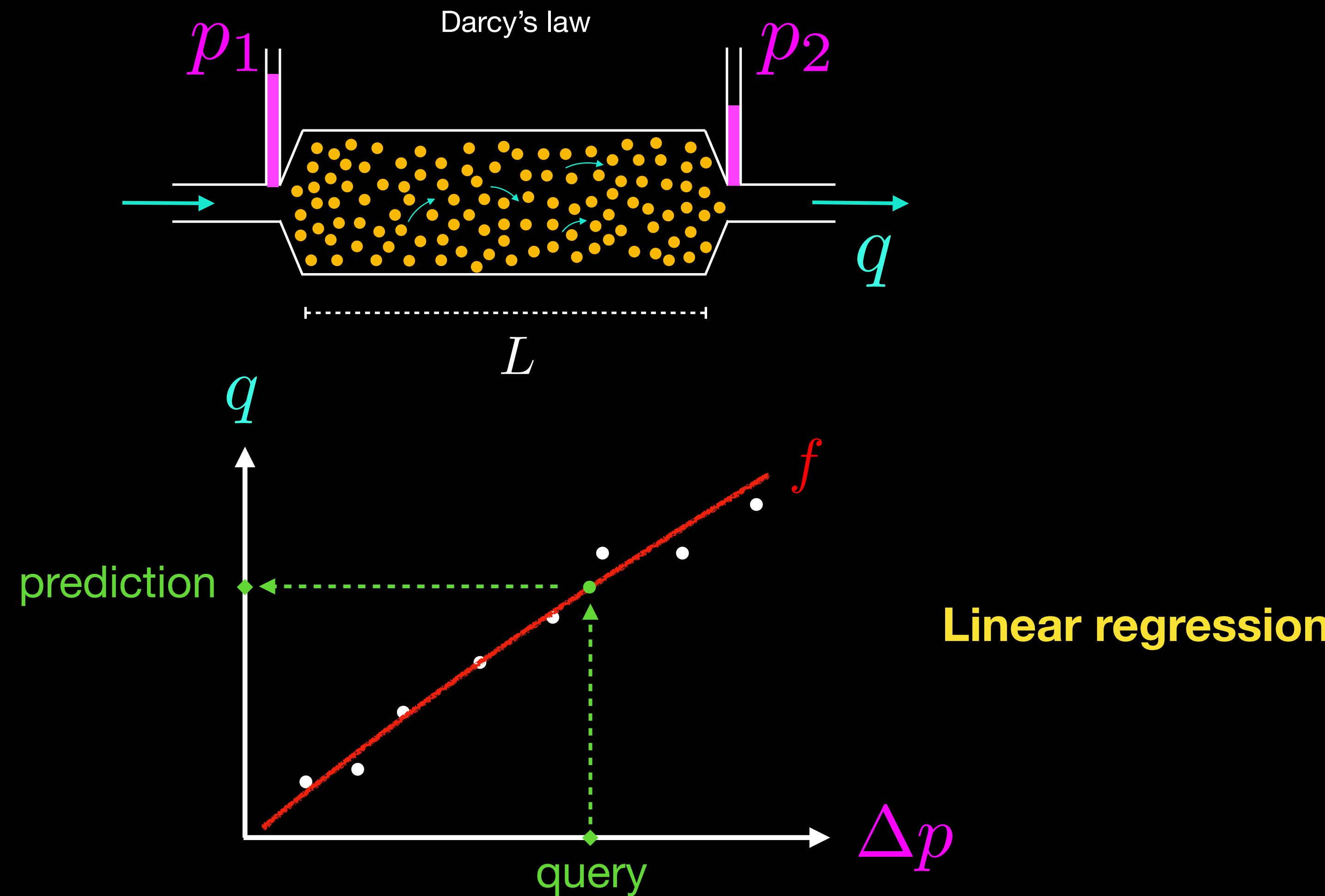
Avogadro's law (1811)

$$\frac{PV}{TN} = k_B$$

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	?

$$(p_1, p_2) \rightarrow f \rightarrow q$$



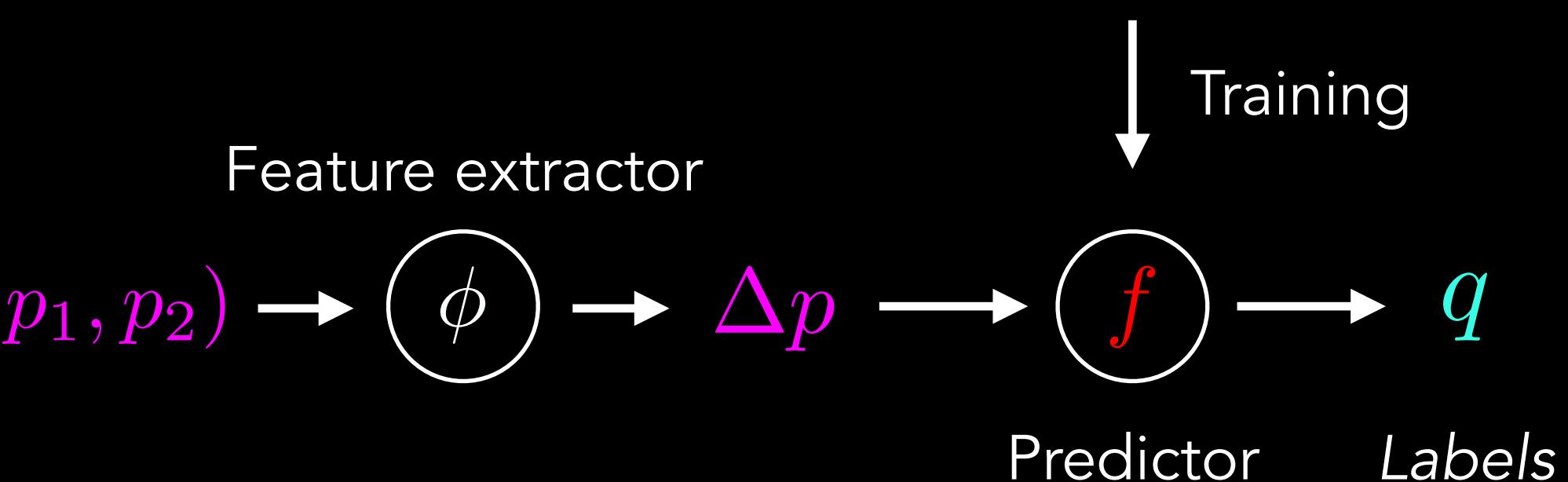
Machine or human learning

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

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-...
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 › 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 License Agreement

<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)

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

Homework

Collect data from your phone and do something with it