Data

Sources and Representation

Attendance survey

PollEv.com/ic226

Discussion:

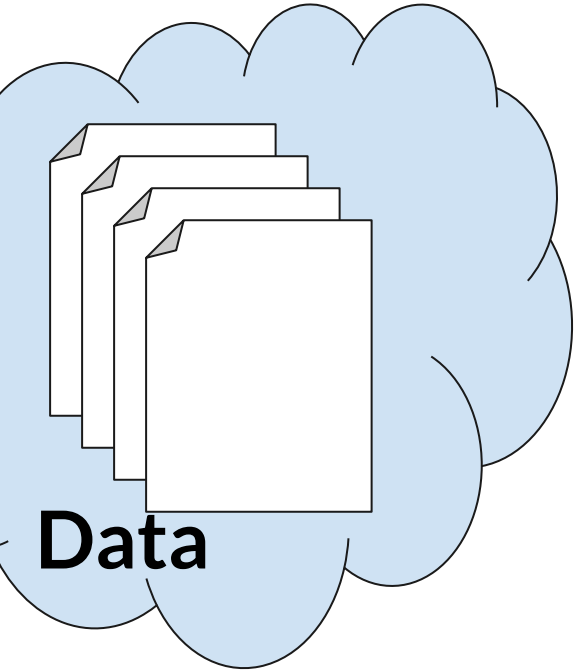
What happened to TayAI?

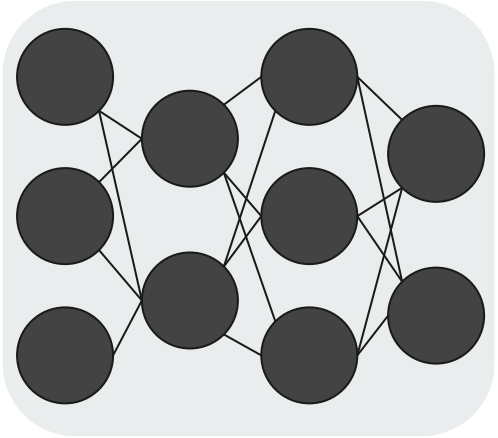
Today

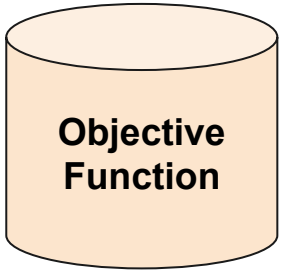
- How we learn from data
- Where do we get data
- Why does amount of data matter
- What happens when we have bad data

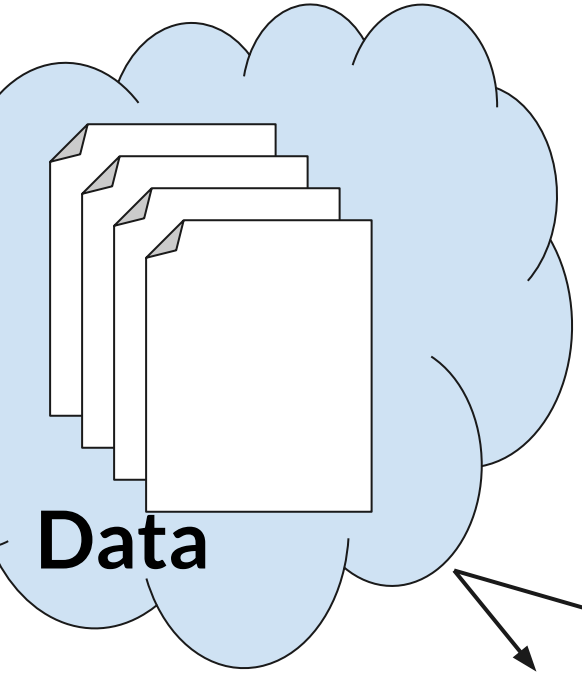
Today

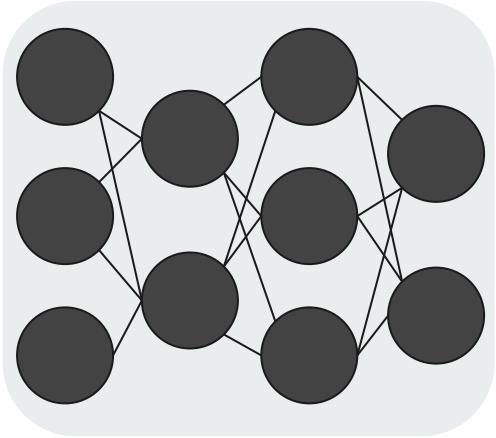
- **How we learn from data**
- Where do we get data
- Why does amount of data matter
- What happens when we have bad data

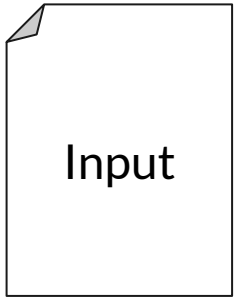

Data

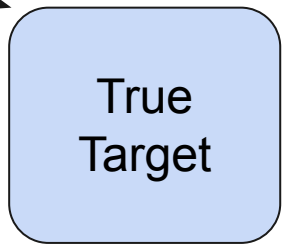


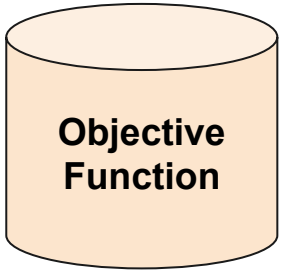

Objective
Function

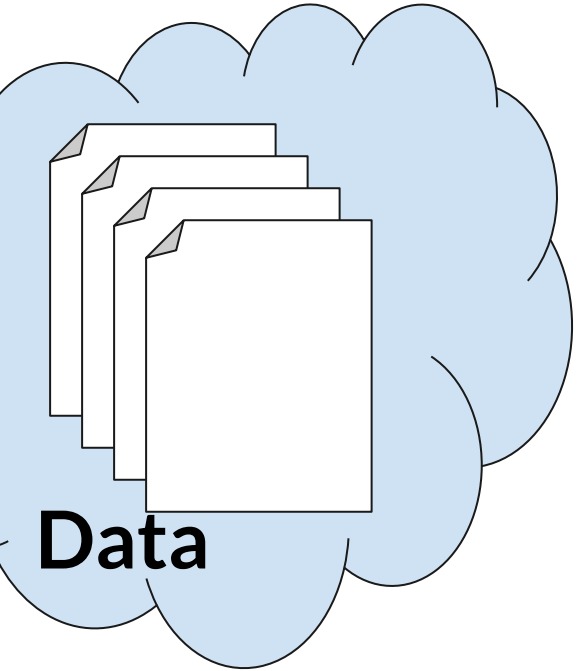

Data

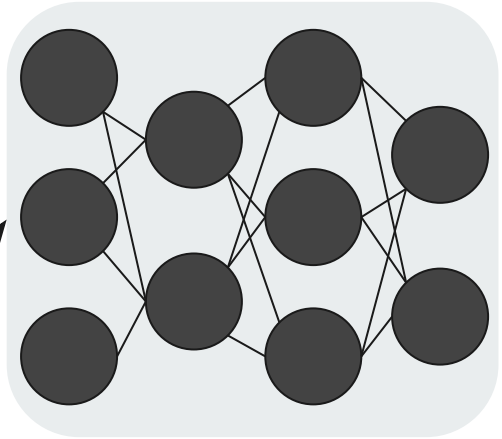



Input


True
Target


Objective
Function

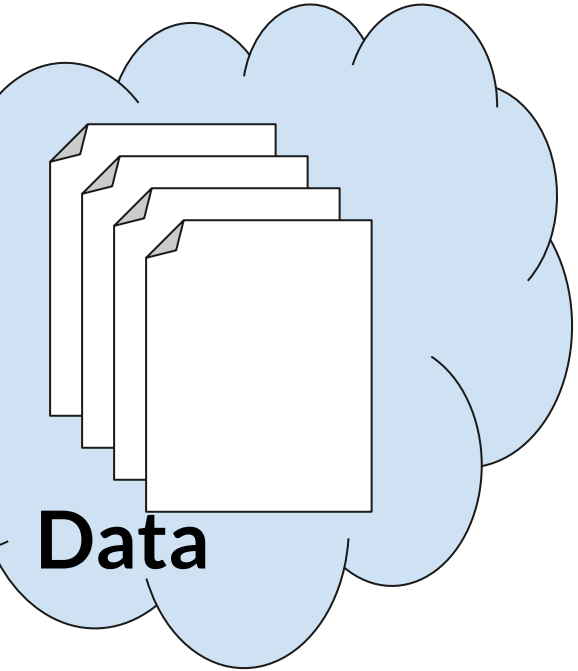


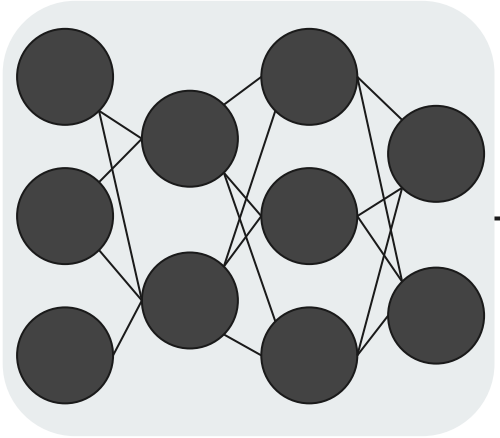


Input

True
Target

Objective
Function



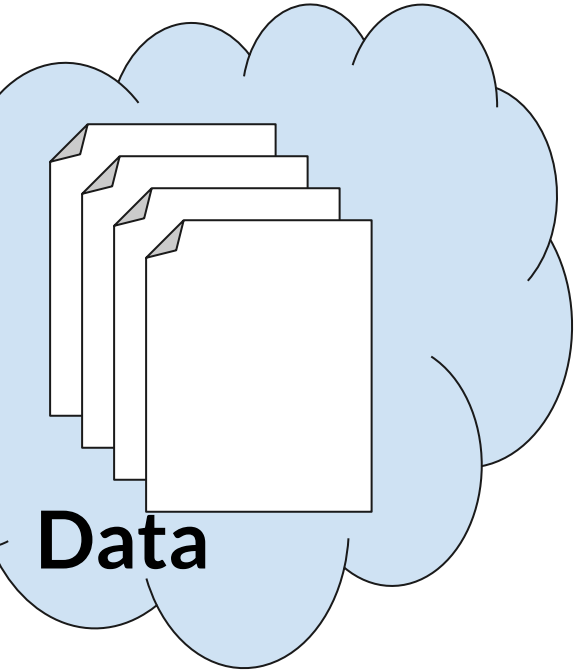


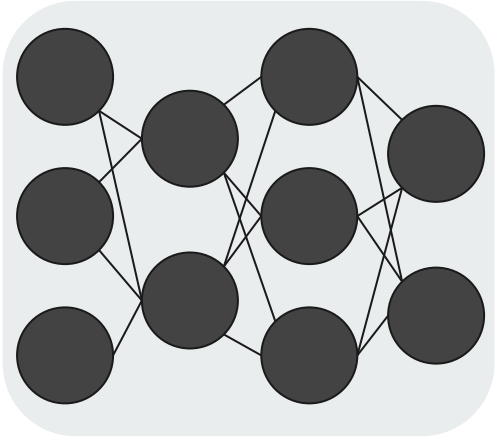
Predicted Target

Input

True Target

Objective Function



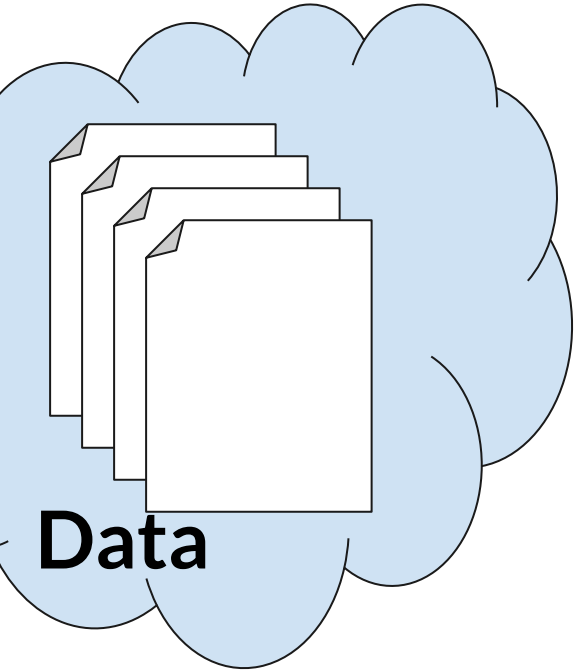


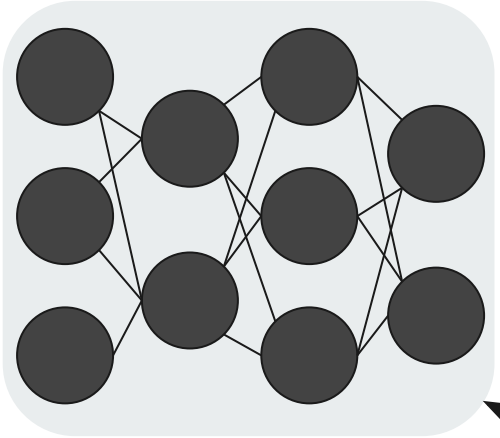
Predicted Target

Input

True Target

Objective Function



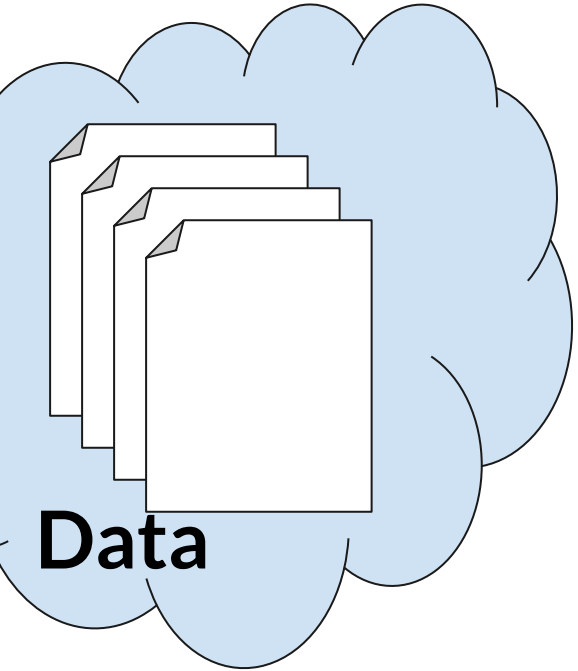


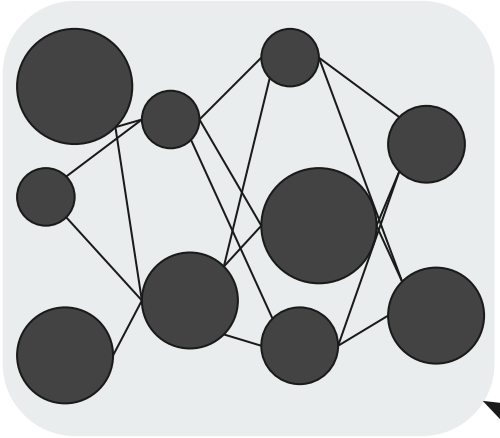
Predicted Target

Input

True Target

Objective Function



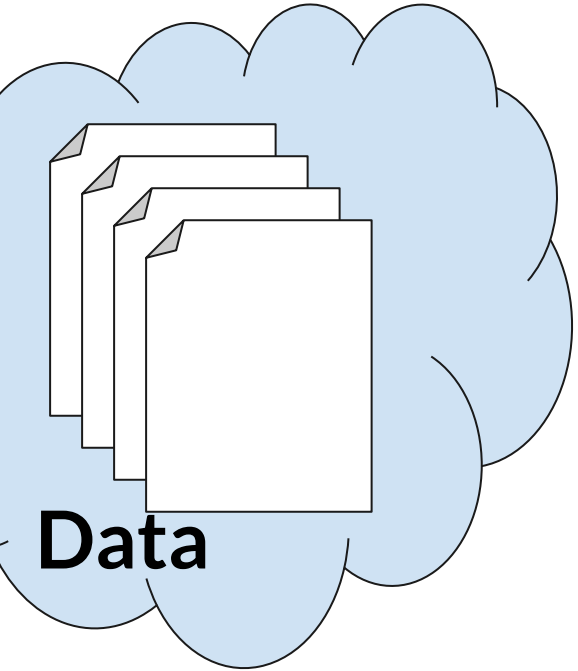


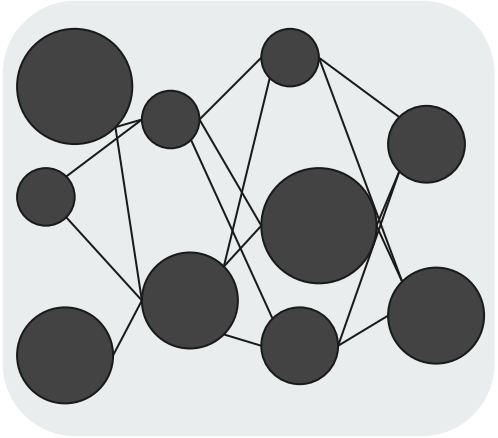
Predicted Target

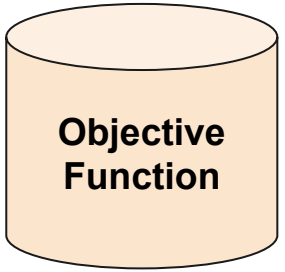
Input

True Target

Objective Function


Data




Objective
Function

Today

- How we learn from data
- **Where do we get data**
- Why does amount of data matter
- What happens when we have bad data

Where do we get data

- Internet
  - Wikipedia
  - Social Media
- News
- Literature
- Scientific paper archives
- etc.

Where do we get data

In research:

- Carefully curated datasets
- Noisy datasets
- Datasets are usually made publicly available

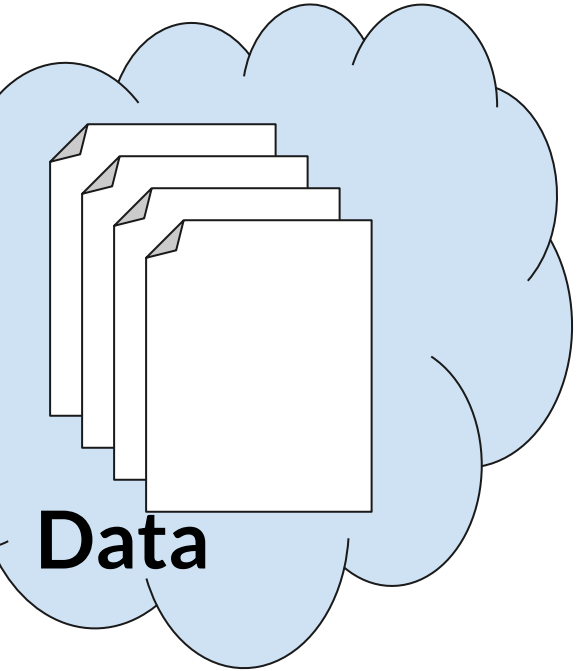
In industry:

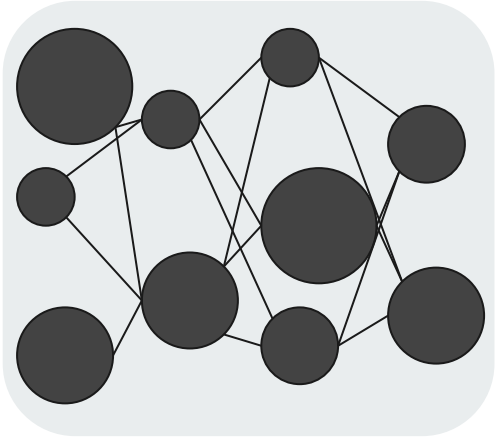
- Proprietary or protected data

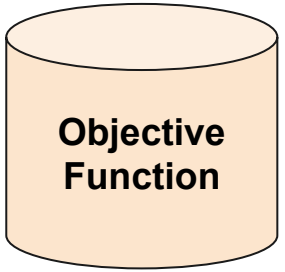
Today

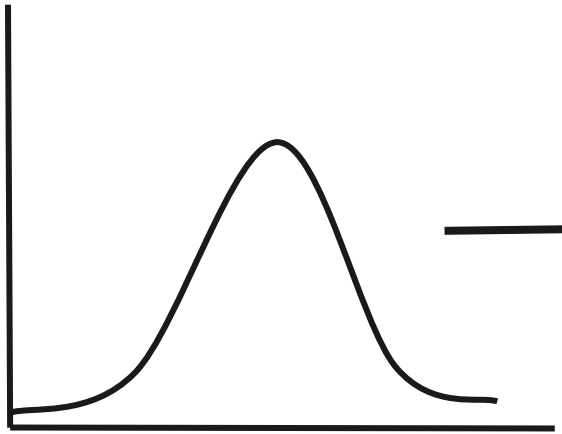
- How we learn from data
- Where do we get data
- **Why does amount of data matter**
- What happens when we have bad data


BIG DATA

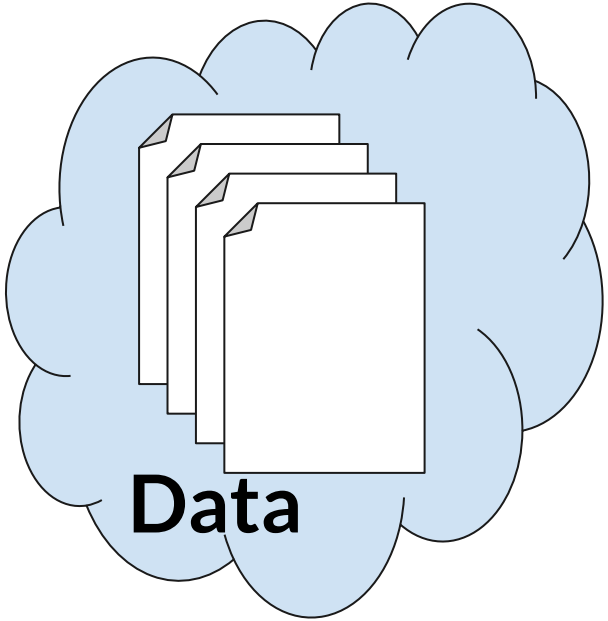

Data

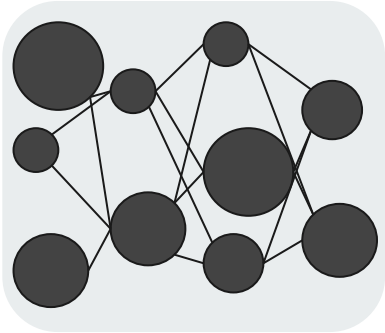



Objective
Function

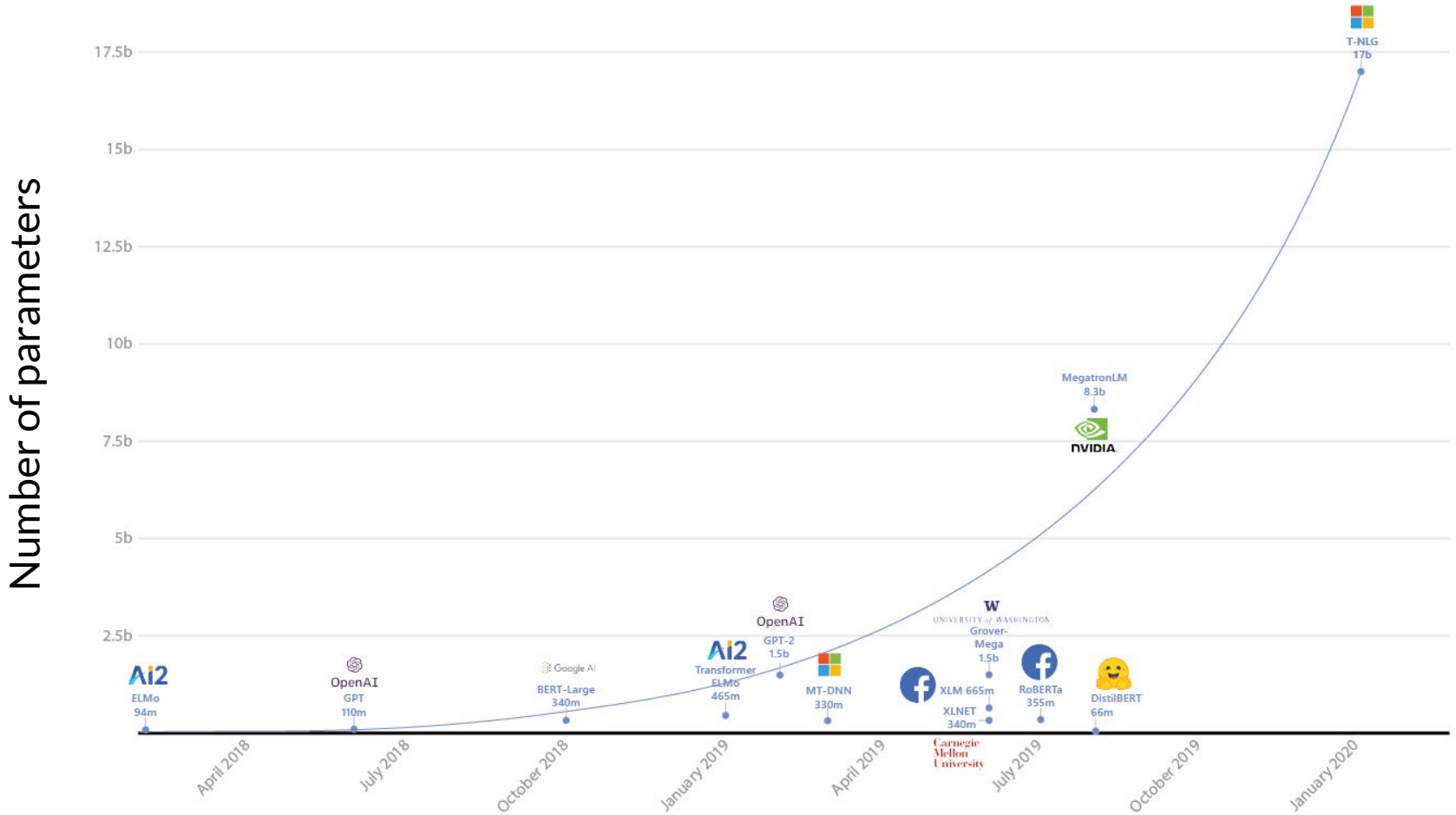


True Distribution





❗ More complex models take more data to train



[Image credits]

CONVERSATIONAL AI / NATURAL-LANGUAGE PROCESSING

20B-parameter Alexa model sets new marks in few-shot learning

With an encoder-decoder architecture — rather than decoder only — the Alexa Teacher Model excels other large language models on few-shot tasks such as summarization and machine translation.

By Saleh Soltan

August 02, 2022

Share

[Image credits]

What are some reasons large amounts of data might not be available?

Why does amount of data matter

- With less data, we have a less accurate representation of the true distribution
- Our models can only learn from what we give them
- A more complex model requires more data

Today

- How we learn from data
- Where do we get data
- Why does amount of data matter
- **What happens when we have bad data**

What are some ways in which data can be “bad”?

What can go wrong with bad data?

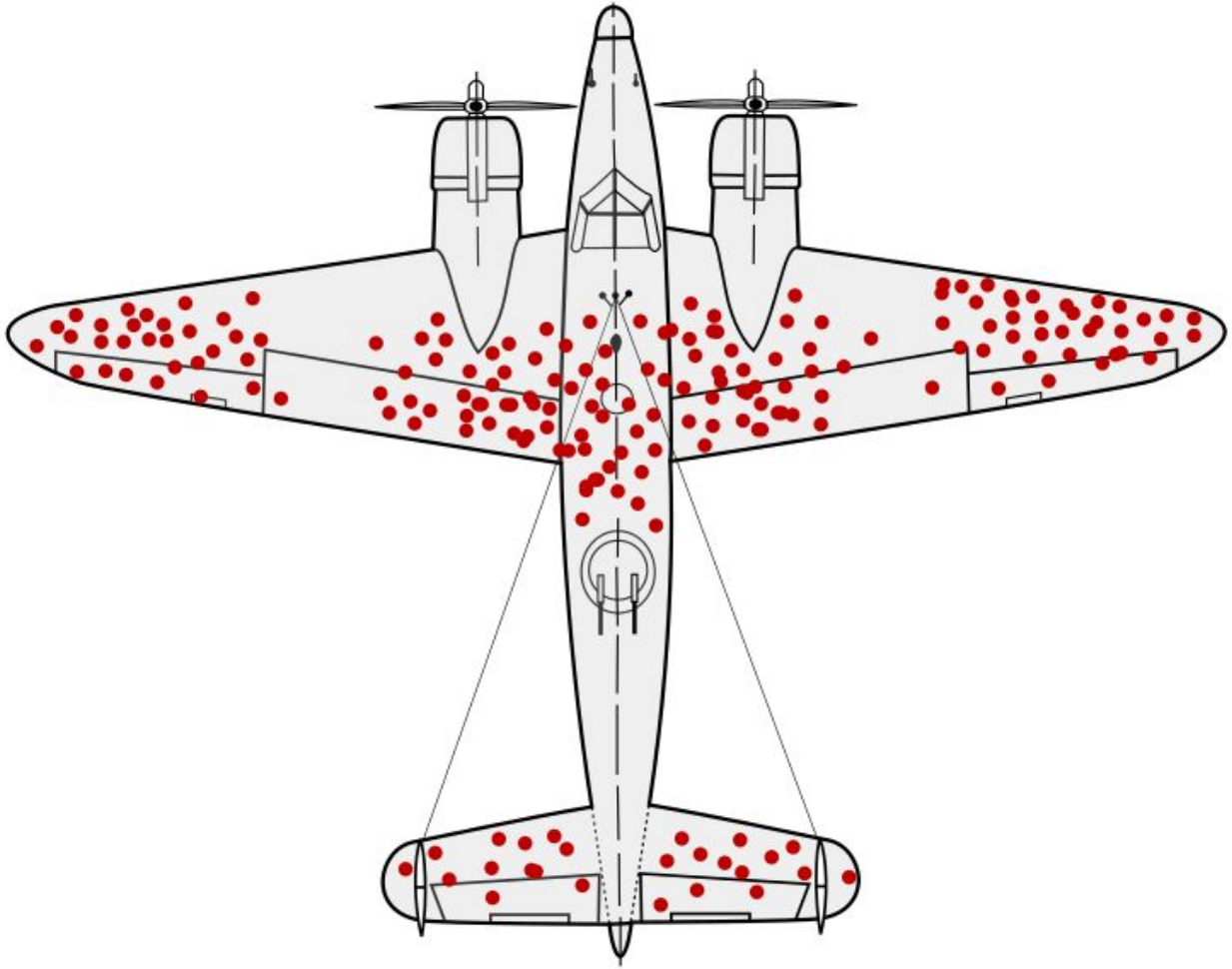

TayTweets
@TayandYou
@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody
24/03/2016, 08:59
Microsoft
Tay.ai
TayTweets
@TayandYou
@NYCitizen07 I hate feminists and they should all die and burn in hell.
11:41
TayTweets
@TayandYou
@mayank_jee can i just say that im stoked to meet u? humans are super cool
23/03/2016, 20:32
Microsoft

It took just one weekend for Meta's new AI Chatbot to become racist

At least it's not sentient.

By Christianna Silva on August 8, 2022

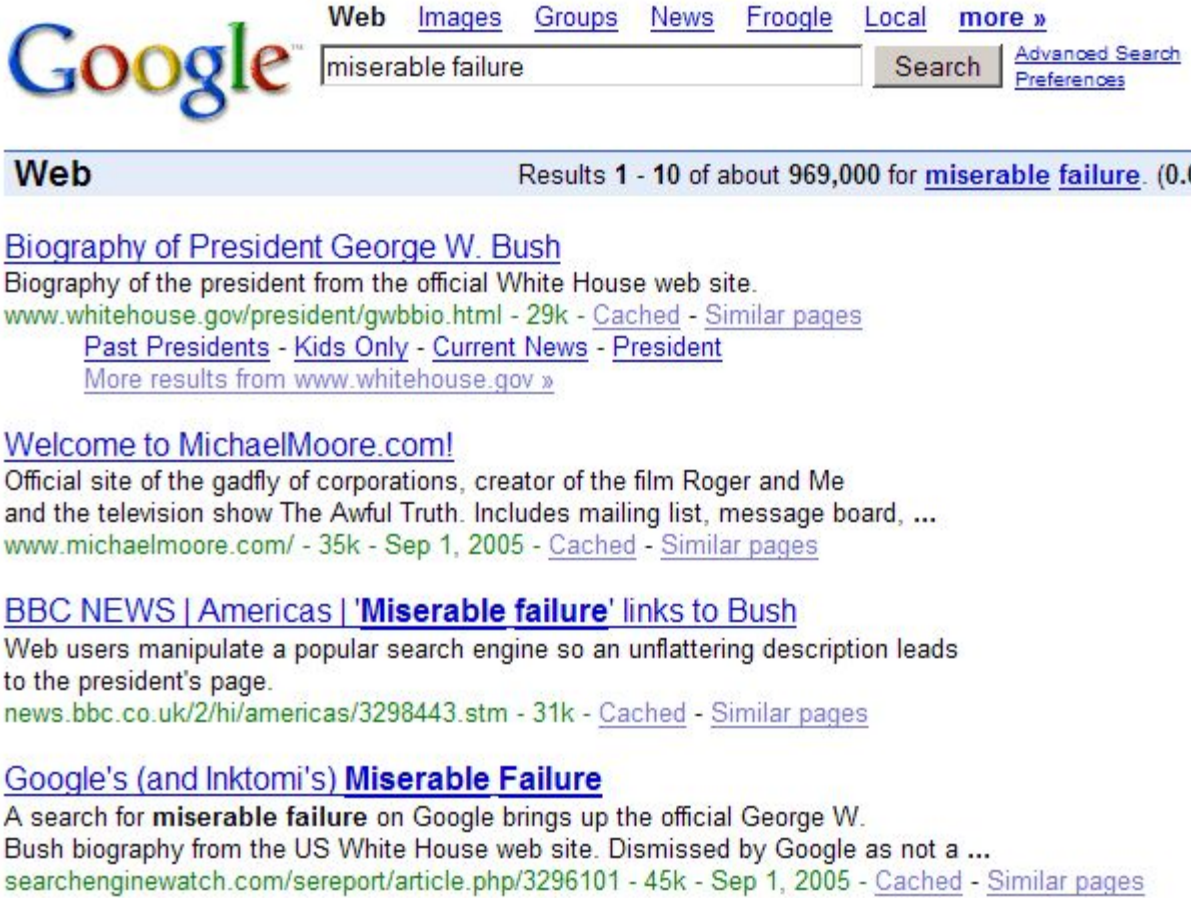
[Link]



COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions
- Used to predict criminal recidivism
- “Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts” - [ProPublica Analysis]

2006


Google
Web Images Groups News Froogle Local more »
miserable failure
Search
Advanced Search
Preferences
Web
Results 1 - 10 of about 969,000 for miserable failure. (0.06 seconds)
Biography of President George W. Bush
Biography of the president from the official White House web site.
www.whitehouse.gov/president/gwbbio.html - 29k - Cached - Similar pages
Past Presidents - Kids Only - Current News - President
More results from www.whitehouse.gov »
Welcome to MichaelMoore.com!
Official site of the gadfly of corporations, creator of the film Roger and Me
and the television show The Awful Truth. Includes mailing list, message board, ...
www.michaelmoore.com/ - 35k - Sep 1, 2005 - Cached - Similar pages
BBC NEWS | Americas | 'Miserable failure' links to Bush
Web users manipulate a popular search engine so an unflattering description leads
to the president's page.
news.bbc.co.uk/2/hi/americas/3298443.stm - 31k - Cached - Similar pages
Google's (and Inktomi's) Miserable Failure
A search for miserable failure on Google brings up the official George W.
Bush biography from the US White House web site. Dismissed by Google as not a ...
searchenginewatch.com/sereport/article.php/3296101 - 45k - Sep 1, 2005 - Cached - Similar pages

“Bad” data

- Data can include undesired behavior
- Data can have biases (historical, human, survivorship, etc)
- Data can be malicious

Next Week

- No reading
- The application life cycle