Basic data literacy for AI reporters

What we will cover in this module

- A brief and incomplete history of how data evolved
- Data in the realm of AI
- Three examples:
  - Structured data feeding decision-making algorithms
  - Longitudinal data fueling your social media timelines
  - Text data for large language models

Let’s start with an exercise

What kind of data is out there?

Let’s break down a tweet.

1. Read this.
2. Laugh a little.
3. Then tell me what kind of data this tweet may yield.

Person who posted it:
Picture link
Display name
Twitter handle

Text

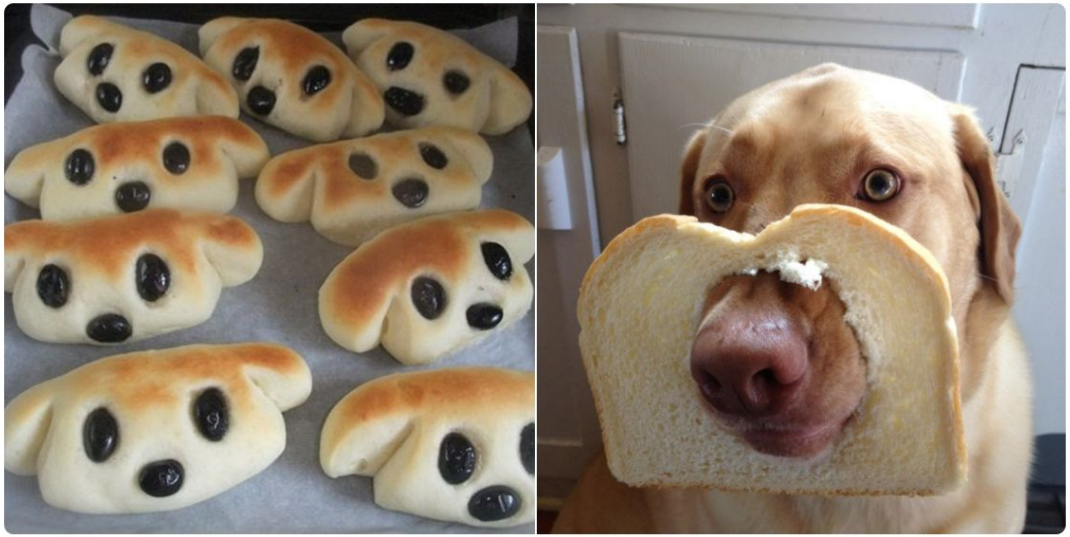
Media

Engagement:
Retweets
Likes
Replies
(further down)
Timestamp

Yael
@elle91

Follow

Purebread dogs vs. inbread dogs



RETWEETS 1,446 LIKES 2,758

12:25 PM - 22 Feb 2017

22 1.4K 2.8K

Same tweet, different format! The entire X.com site is fueled by data streams like this.

APIs or Application Programming Interface*

**fancy word for a stream of data in a format that robots/computers can understand* 🤖

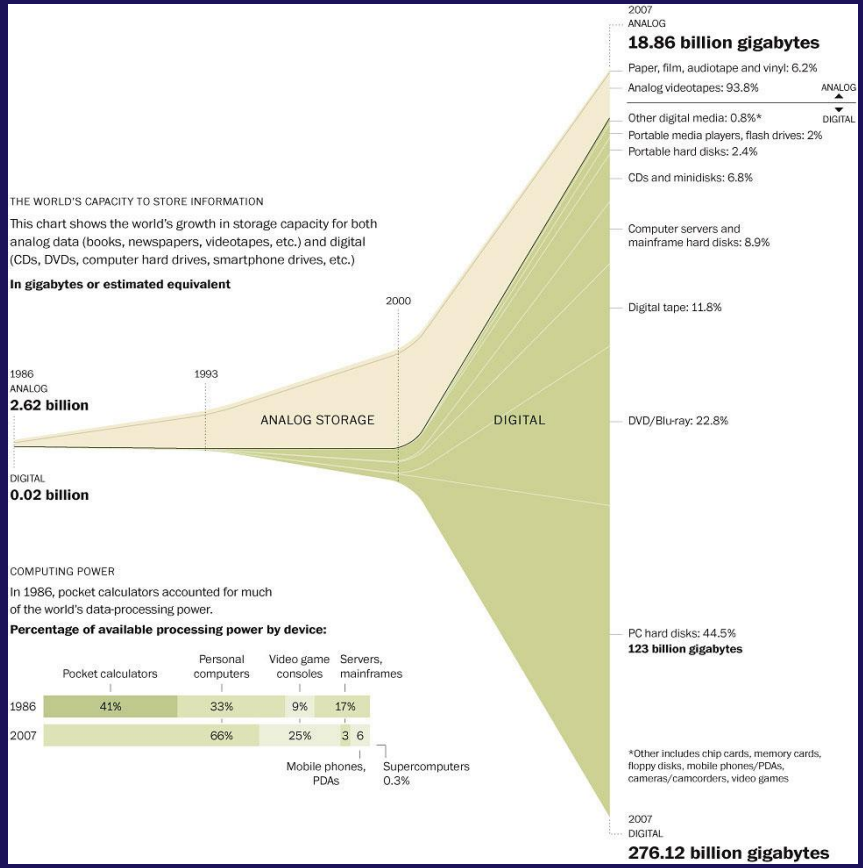

```
{
    "contributors": null,
    "truncated": false,
    "text": "Purebread dogs vs. inbread dogs https://t.co/e1nC0d7EkR",
    "is_quote_status": false,
    "in_reply_to_status_id": null,
    "id": 834454023260532737,
    "favorite_count": 2760,
    "source": "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for
    Android</a>",
    "retweeted": false,
    "coordinates": null,
    "entities": {
        "symbols": [],
        "user_mentions": [],
        "hashtags": [],
        "urls": [],
        "media": [ {
            "expanded_url": "https://twitter.com/elle91/status/834454023260532737/photo/1",
            "display_url": "pic.twitter.com/e1nC0d7EkR",
            "url": "https://t.co/e1nC0d7EkR",
            "media_url_https": "https://pbs.twimg.com/media/C5STfk1WQAEzRet.jpg",
            "id_str": "834454001643044865",
            "sizes": {
                "large": {
                    "h": 413, "resize": "fit", "w": 550
                }
                ,
                "small": {
                    "h": 413, "resize": "fit", "w": 550
                }
                ,
                "medium": {
                    "h": 413, "resize": "fit", "w": 550
```

Data is everywhere!

The world wide web is fueled by data

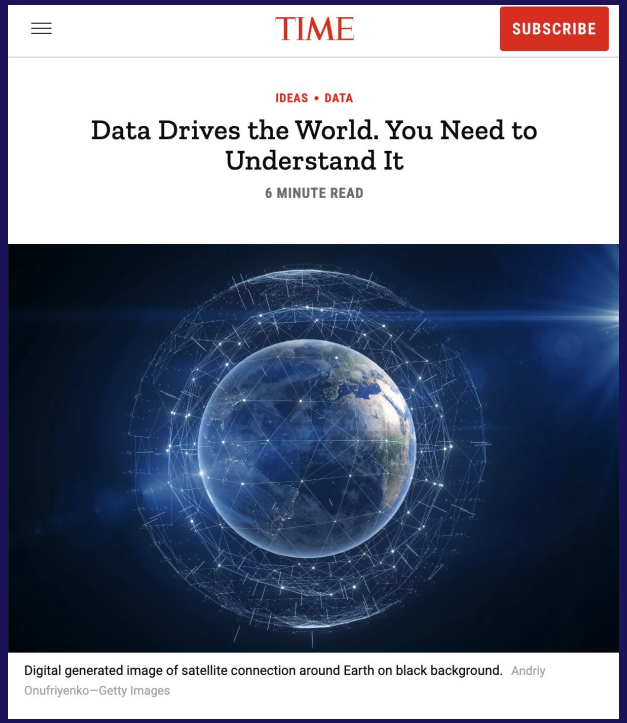
What happened to data?

There’s more and more of it



Source: Washington Post (2011)

There’s a lot more than just government data now



Source: IBM via Time

Source: CNET

Digital footprint example: WSJ

How Pizza Night Can Cost More in Data Than Dollars

Even a low-key evening at home can mean handing over a trove of personal information to high-tech companies

By Stephanie Stamm, Tripp Mickle and Jessica Kuronen

Published April 10, 2018 at 5:30 a.m. ET

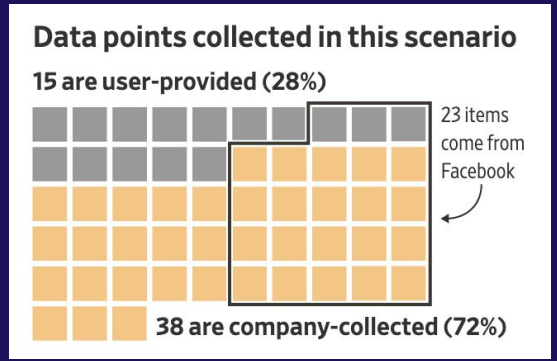
The smartphones, Facebook accounts and other technology products deeply embedded in modern life help people get more things done every day. They also gather more information about us than we often realize.

But Facebook's crisis over how it handles and protects user data has led some to ask: What data am I giving up?

Imagine "Sally" sets up a pizza-and-movie night with her friend "Kristen." The Wall Street Journal reviewed privacy statements to assess just how much data could be unknowingly shared on top of the price of that pepperoni pie.

The Cost in Data

Sally and Kristen potentially gave up at least 53 pieces of information together. The data detailed in the scenario reflect information the companies could collect according to their privacy statements, terms of service and related documents.



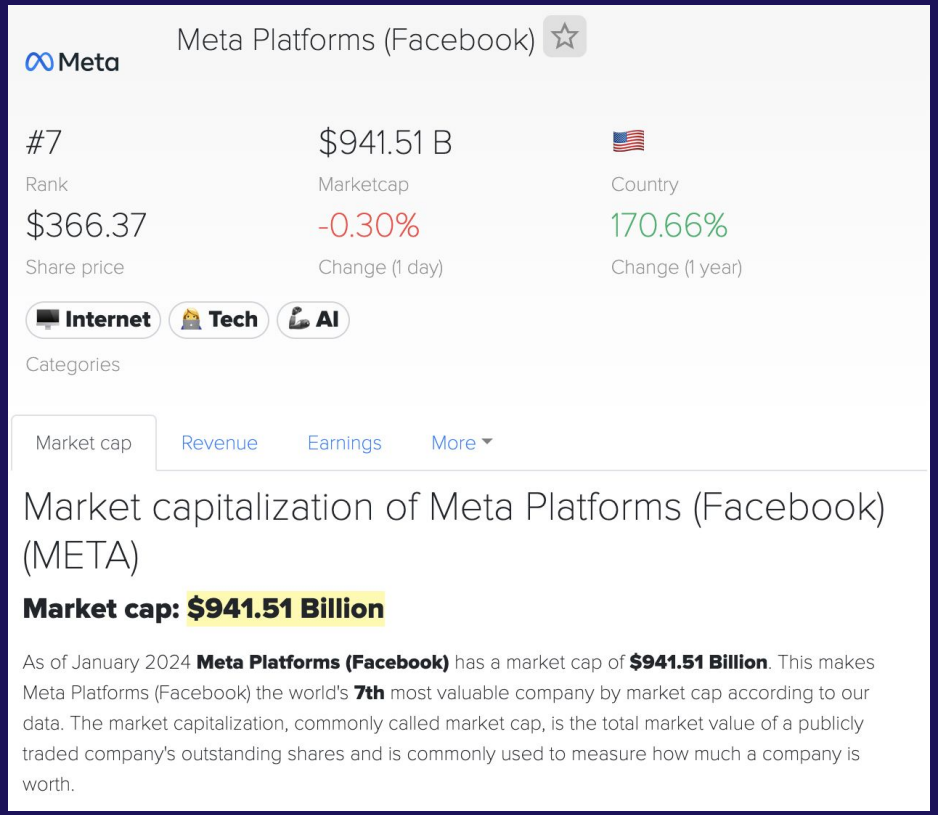
Source: WSJ

Digital footprint example: Meta data servers

Aerial view of some of Facebook’s data centers

Big Data

It’s making a lot of people a lot of money

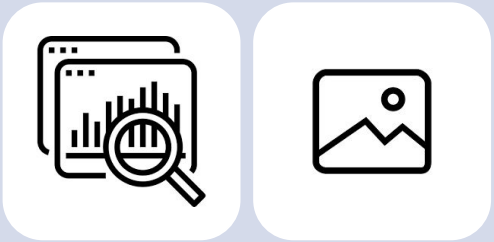


Data in the realm of AI

Lifecycle of AI

Where in the life cycle is training data?

Inputs or training data (text, imagery, statistics)



Model/Algorithm

AI

Outputs (‘predictions,’ evaluative scores, new images, text)

We predict with 95% certainty that this image pictures the pope

or

Lifecycle of AI

Where in the life cycle is training data?

Training data (text, imagery, statistics)

ML algo

model

new data or prompt

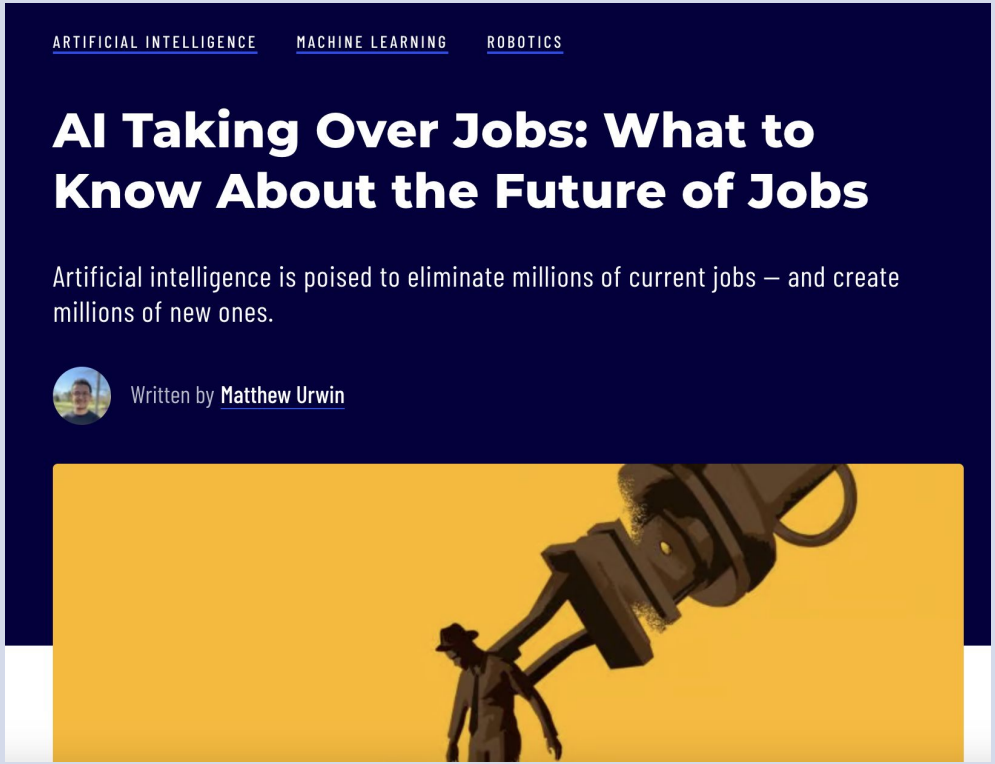
model

answer or new data ('predictions,' evaluative scores, new images, text)

We predict with 95% certainty that this image pictures the pope

or

AI: Data is changing our society



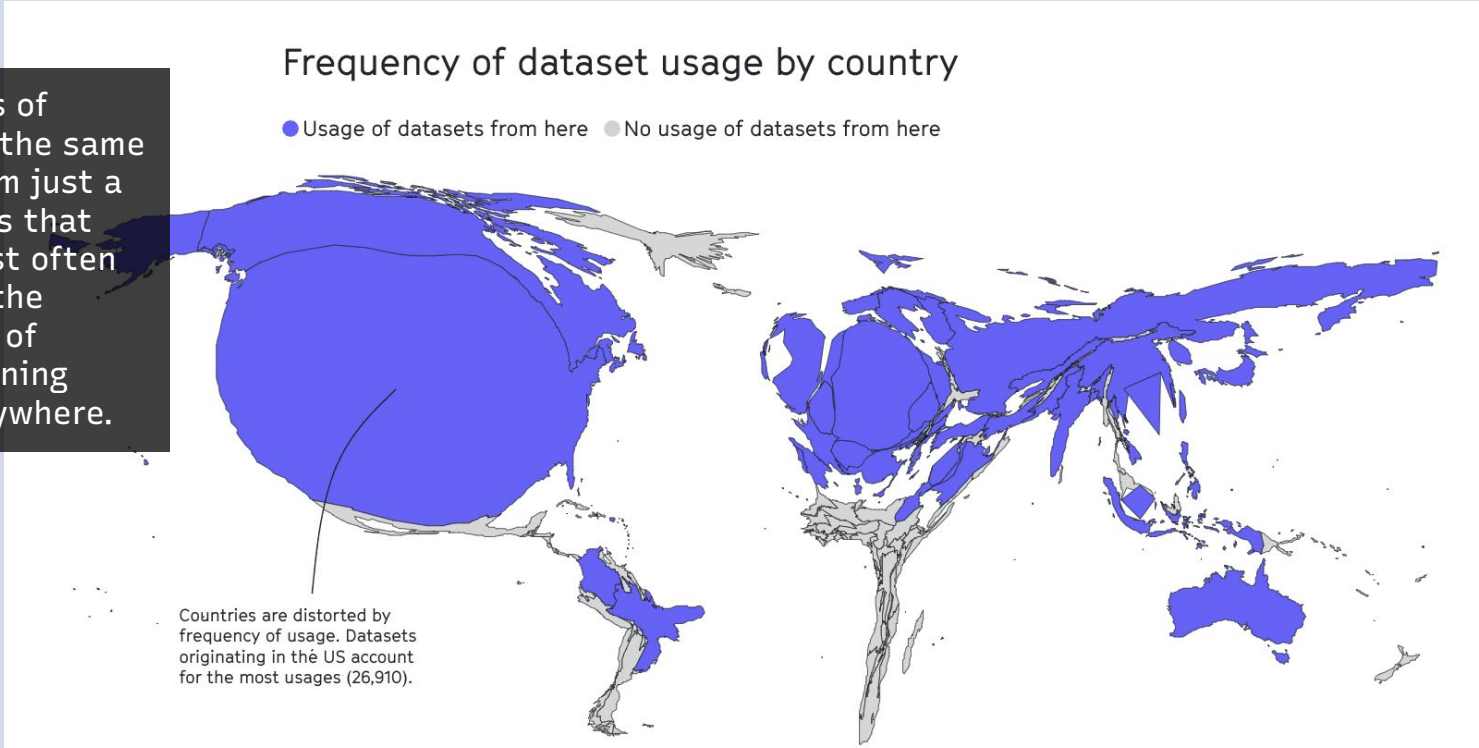
Source: https://builtin.com/artificial-intelligence/ai-replacing-jobs-creating-jobs

AI: Data is changing our society

Source: Lighthouse News

A lot of the data is reproducing bias

In thousands of papers, it is the same datasets from just a few countries that are used most often to evaluate the performance of machine learning models everywhere.



Source: Mozilla

Structured data feeding decision-making algorithms

Under Haryana's Old Age Samman Allowance scheme, people aged 60 years and above, whose income together with that of their spouse doesn't exceed 300,000 rupees ($3,600) per annum, are eligible for a monthly pension of 2,750 rupees ($33).

In June 2020, the state started using a newly built algorithmic system – the Family Identity Data Repository or the Parivar Pehchan Patra (PPP) database – to determine the eligibility of welfare claimants.

Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

Source: Pulitzer Center

The PPP is an eight-digit unique ID provided to each family in the state and has details of birth and death, marriage, employment, property, and income tax, among other data, of the family members. It maps every family's demographic and socioeconomic information by linking several government databases to check their eligibility for welfare schemes.

The state said that the PPP created "authentic, verified and reliable data of all families", and made it mandatory for citizens to access all welfare schemes.

But in practice, the PPP wrongly marked Chand as "dead", denying him his pension for several months. Worse, the authorities did not change his "dead" status even when he repeatedly met them in person.

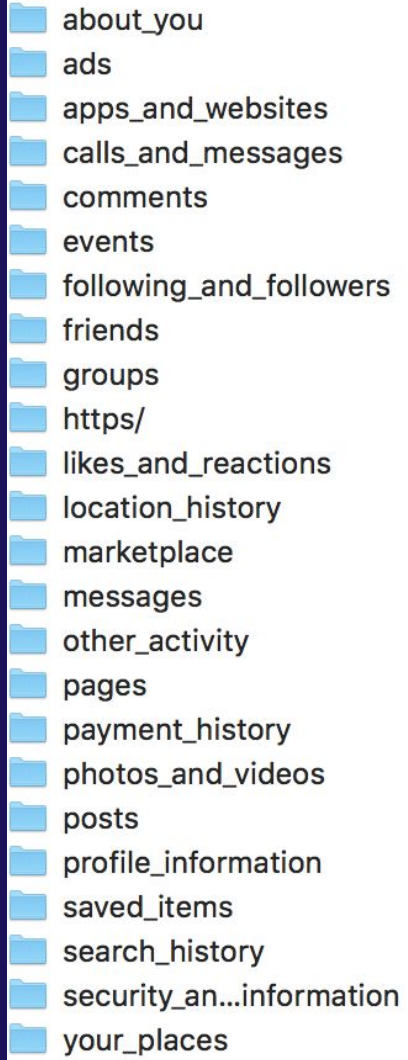
Dhuli Chand led a wedding procession to prove to government officials he was alive and should be paid his pension. Image courtesy of The Reporters' Collective.

Source: Pulitzer Center

Longitudinal data fueling your social media timelines

Data used by Facebook

Folders that are included in a downloadable Facebook data archive



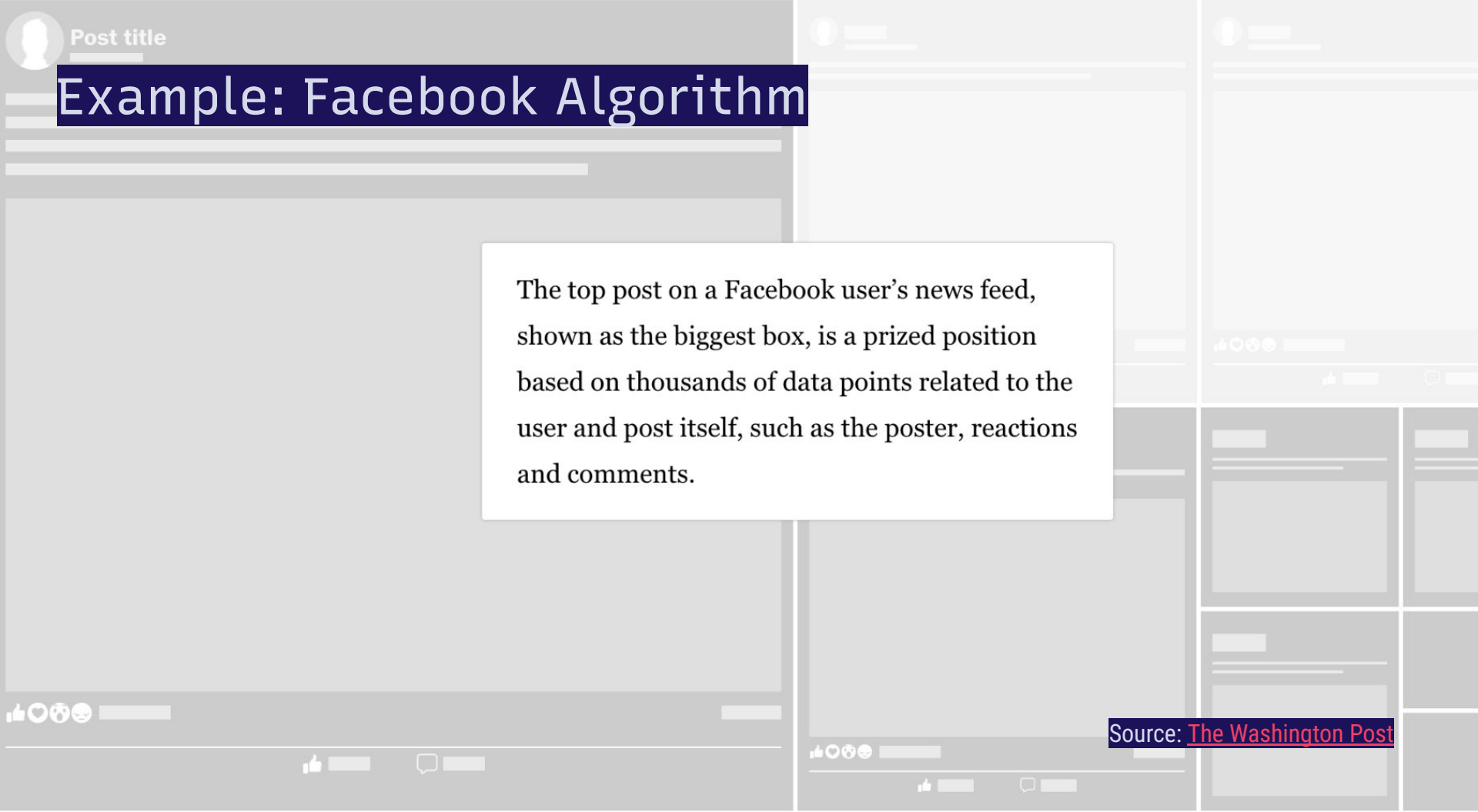
The Facebook algorithm is a ranking system that uses machine learning to arrange content in users' feeds. Data used:

- **Data you put into Facebook**
- **Data that you produce by using Facebook**
- **Data that is tracked of you by pages with Facebook buttons**

To get an idea of how deep this data goes users can download their data following Facebook's instructions:
https://www.facebook.com/help/1701730696756992?helpref=hc_global_nav

More on the subject here:
How Facebook watches you online (BuzzFeed News)
3 Simple Ways We Give Up A Ton Of Very Personal Information To Facebook And Random Apps (BuzzFeed News)

Example: Facebook Algorithm

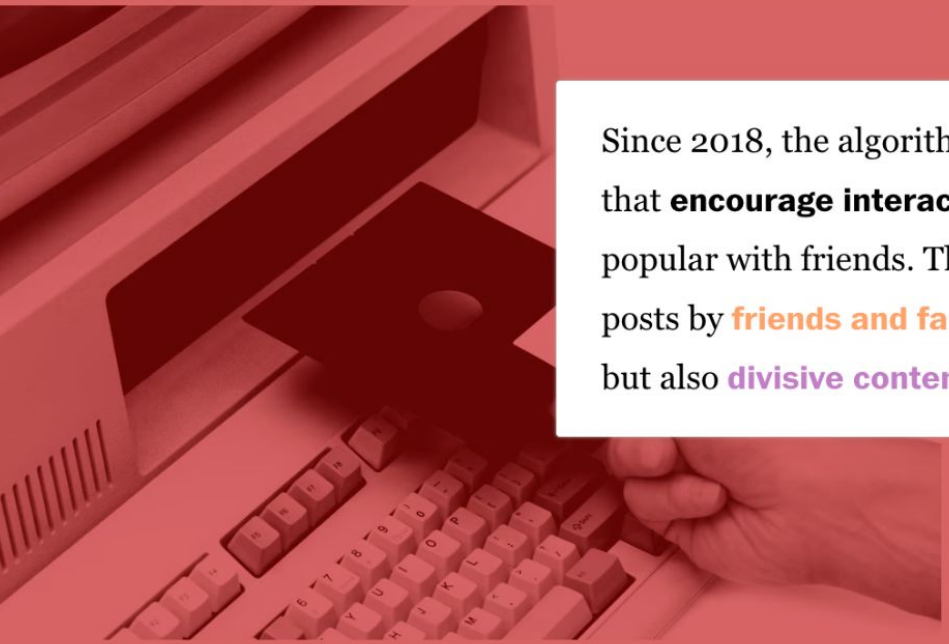


The top post on a Facebook user's news feed, shown as the biggest box, is a prized position based on thousands of data points related to the user and post itself, such as the poster, reactions and comments.

Source: The Washington Post

Example: Facebook Algorithm

Show your age



Zoe A. and 60.8K others

10.7K comments

Like Comment

Trip photos

Ray M. and 79 others

Like Comment

Since 2018, the algorithm has elevated posts that **encourage interaction**, such as ones popular with friends. This broadly prioritizes posts by **friends and family** and **viral memes**, but also **divisive content.**

Kai J. and 45.2K others

Like Comment

Engaged!

Rap name

Source: The Washington Post

Example: Facebook Algorithm

Player takes political stand

Lu Y. and 22.8K others

12.3K comments

Like Comment

Spicy opinion piece

Adi R. and 42 others

Like Co

J.R. and 6.7K others

Like Comment

Each user's feed reflects their expressed interests. For a subset of **extremely partisan** users, today's algorithm can turn their feeds into echo chambers of **divisive content** and **news**, of varying reputability, that support their outlook.

Source: The Washington Post

Text data for natural language processing

Example: Google

Predictive text in Google search is based on countless searches.

How many broken-hearted people googled Ryan Gosling's name to find out whether he was married?! (Plus a lot of people likely also googled a male celebrity name + wife)

Some concepts

1. **Corpus**
2. **Tokenizing**
3. **Stemming or lemmatizing**

Some concepts

1. **Corpus**
2. **Tokenizing**
3. **Stemming or lemmatizing**

Some concepts

1. **Corpus**
2. **Tokenizing**
3. **Stemming or lemmatizing**

Some concepts

1. **Corpus**
2. **Tokenizing**
3. **Stemming or lemmatizing**

The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

For instance:
am, are, is → be
car, cars, car's, cars' → car

The result of this mapping of text will be something like:

the boy's cars are different colors
→ the boy car be differ color

Some concepts

1. **Corpus**
2. **Tokenizing**
3. **Stemming or lemmatizing**

Stemming vs. **lemmatization**

Example: “saw”
(As in “I saw a bird, or was it a plane? I think it may have been Supergirl.”)

Stem:
saw → s

Lemma:
saw → see*

** if the token is a verb!*

Analyzing text

Once you've prepared your corpus into tokens (either as stems or lemmas) you can then analyze the text. This could include:

- **Creating lists of words that are used the most**
- **Predicting which words are most likely to follow one another (just like in the good search!)**
- **Detecting sentiment in a text**

AI Spotlight Series

The AI Accountability Network

Pulitzer Center