Bad Data In, Years of Useless Analysis - then Purged

Alisson Sol

June 4th, 2025

Sometimes, we capture data and beat it with years of analysis, hoping it will confess, as in the famous joke. But it doesn’t. Then, we resist the temptation to purge data, hoping some new tool or method in the future will bring us some magical insight. This session showcases the main issues that lead to bad data and then bad analysis.

Who is presenting?

- Alisson Sol has many years of experience in software development, having hired and managed several software teams that shipped many applications, services, and frameworks, focusing on image processing, computer vision, ERP, business intelligence, big data, machine learning, AI, cybersecurity, and distributed systems.
- He has a B.Sc. in Physics and an M.Sc. in Computer Science from the Federal University of Minas Gerais in Brazil and General Management training at the University of Cambridge-UK. When not coding, he likes to run half-marathons, play soccer, disassemble hardware, put it back to work, and reuse the spare parts elsewhere!
- Thanks to my current and previous employers for the experiences.
  All responsibility for the content is mine.

Setting expectations and hypotheses

Attention to these hypotheses

- We capture too much data of bad quality.
- We fear deleting data and keep torturing it.
- Eventually, there will be minimal cleanup instead of a methodical selection of the highest valuables to keep.

A tale of 3 data projects/teams/companies

DataVivid.com	DataVanity.com	DataVague.com
Only good data	Some good data, some bad!	Only bad data
Only correct analysis	Some correct analysis, some wrong!	Incorrect data analysis
Auto data cleansing	Reactive data purge emergencies	Never delete data

Reactive data purges

- Situation (trigger)
  - Storage crunch
  - Cyber incident
- Action
  - Delete data
  - “Clean servers”
- Result
  - Lost source code for major app
  - Lost wiki content in server cleanup

INTERESTING HISTORY

Toy Story 2 Was Accidentally Deleted but Saved by an Employee

By History and Mystery • August 14, 2024

You've probably heard of **Toy Story 2**, but did you know that the beloved animated film almost never made it to the big screen? In 1998, a simple command **nearly wiped out** months of hard work, deleting 90% of the movie's files. It's a nightmare scenario for any creative project, let alone a major studio production. But thanks to an **unexpected hero** working from home, the film was saved from oblivion. This near-disaster serves as a stark reminder of the importance of **robust backup systems** in the digital age. The story behind Toy Story 2's brush with deletion is as enthralling as the movie itself.

Pre-mortem: autoclean

- Practice: what if data was deleted?
- What are your “crown jewels”?

- Why keep other data streams?
  - Can it be moved to cold storage?
    - Circa 2025: ~$1 per TB per month + retrieval + transmission costs
      - Amazon S3 Glacier Deep Archive
      - Google Cloud Storage Archive
      - Azure Blob Storage Archive

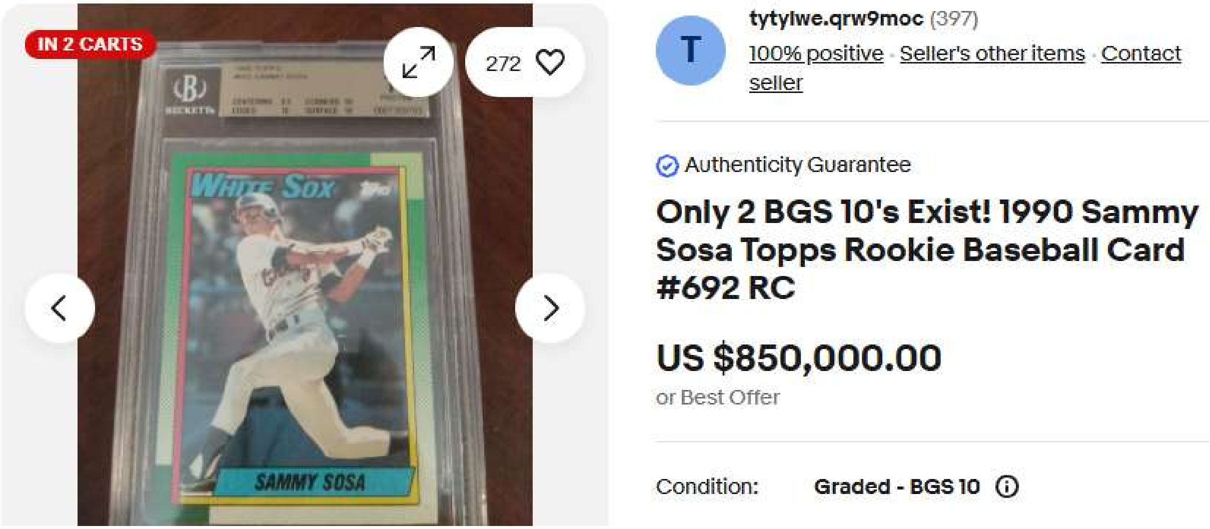
Who is doing autoclean?

- We capture too much data of bad quality.
- We fear deleting data and keep torturing it.
- ***Eventually, there will be minimal cleanup instead of a methodical selection of the highest valuables to keep.***

Why do we fear deleting data?

Why do we hoard?

- Hardship to "acquire"
- Fear of irreparable loss
- Endowment effect



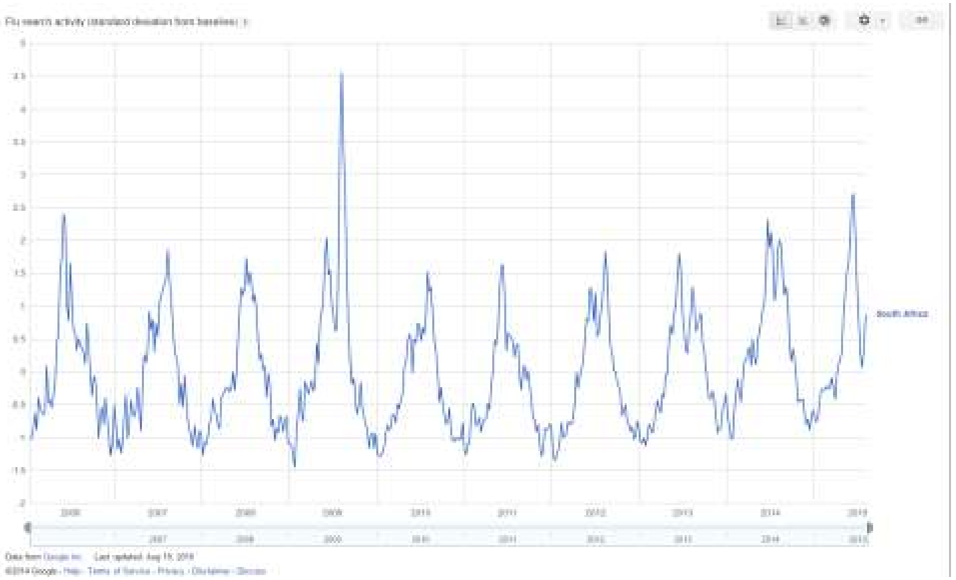
Has data torture produced confessions?

Google Flu Trends

Article Talk

From Wikipedia, the free encyclopedia

Google Flu Trends (**GFT**) was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]



Netflix Prize data

Dataset from Netflix's competition to improve their reccommendation algorithm

Data Card Code (156) Discussion (7) Suggestions (0)

About Dataset

Context

Netflix held the Netflix Prize open competition for the best algorithm to predict user ratings for films. The grand prize was $1,000,000 and was won by BellKor's Pragmatic Chaos team. This is the dataset that was used in that competition.

The BigChaos Solution to the Netflix Grand Prize

Andreas Töscher and Michael Jahrer

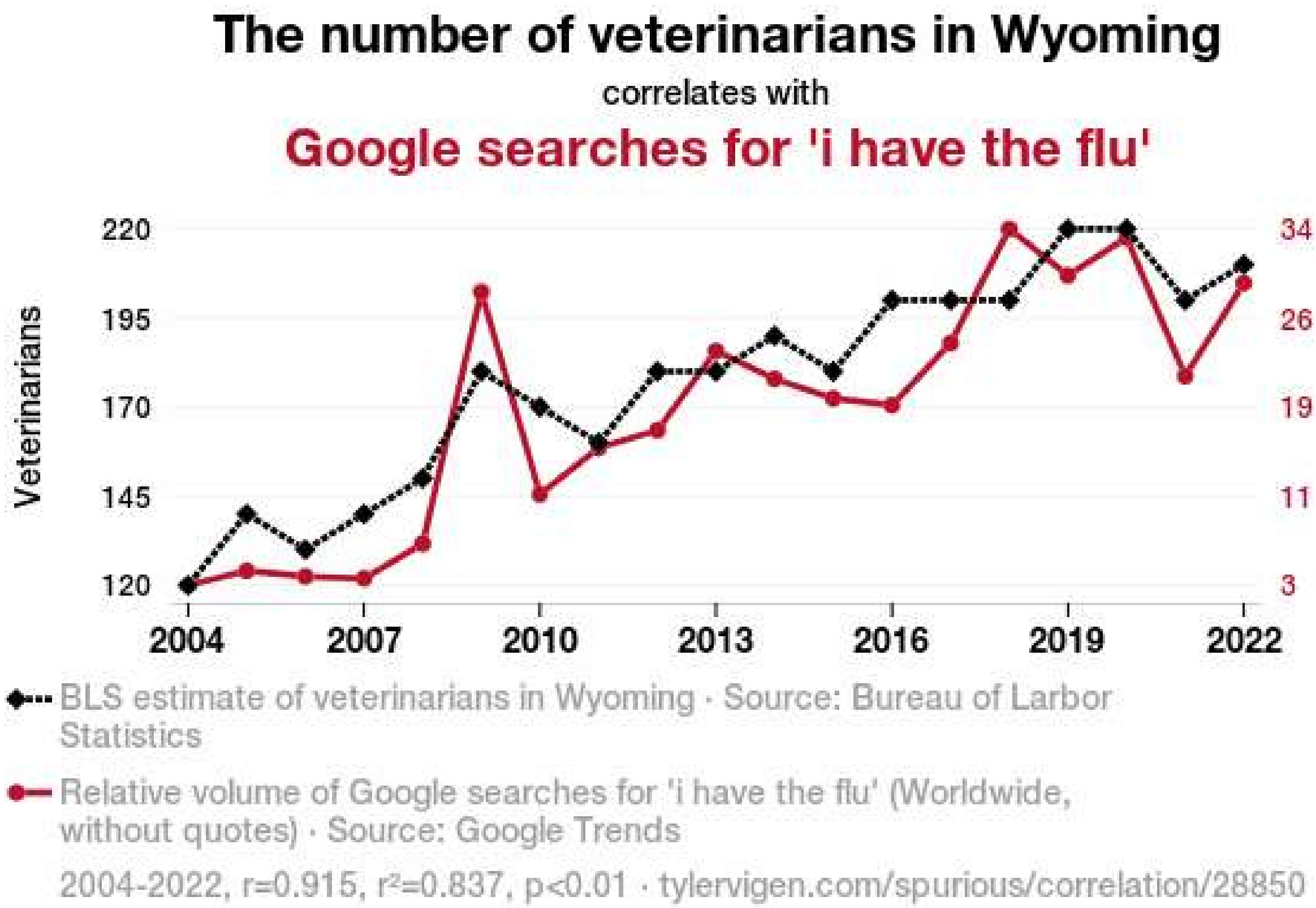
commendo research & consulting
Neuer Weg 23, A-8580 Köflach, Austria
{andreas.toescher,michael.jahrer}@commendo.at

Robert M. Bell*

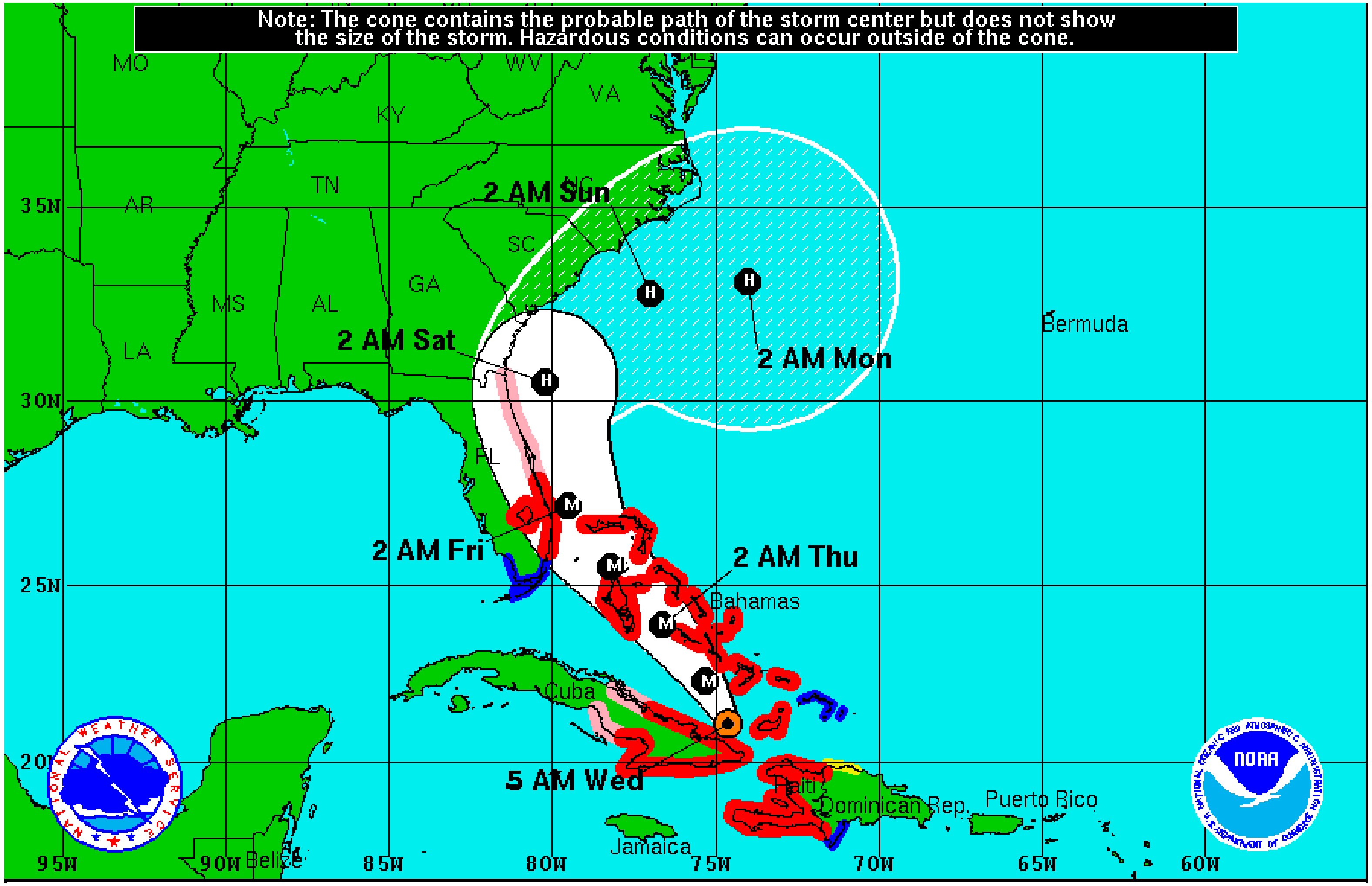
AT&T Labs - Research
Florham Park, NJ

September 5, 2009

The data may always confess “something”



Data analysis extrapolation: prediction error



Data is not information

- Simple test: data can be sorted, sharded, etc.
- Information value diminishes “out of order” or “out of context”
  - Consider: tomorrow (June 5th), the Bitcoin price will be ~$2,686.81

Bitcoin USD Price (BTC-USD) ☆ Follow + Add holdings

106,717.17 -871.12 (-0.81%)

As of 5:34:00 PM UTC. Market Open.

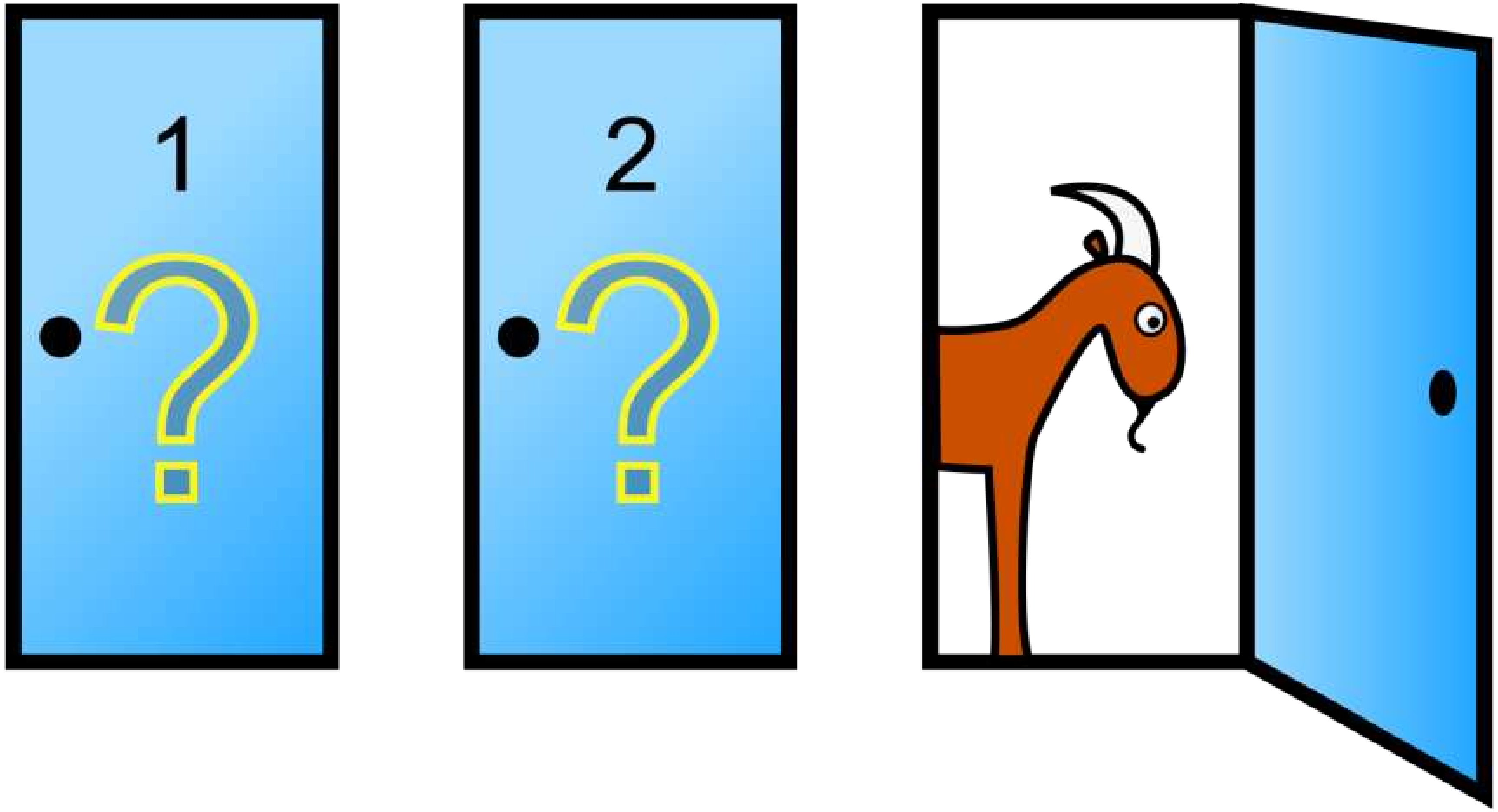
Data provided by CoinMarketCap

Jun 04, 2017 - Jun 06, 2017 | Historical Prices | Daily

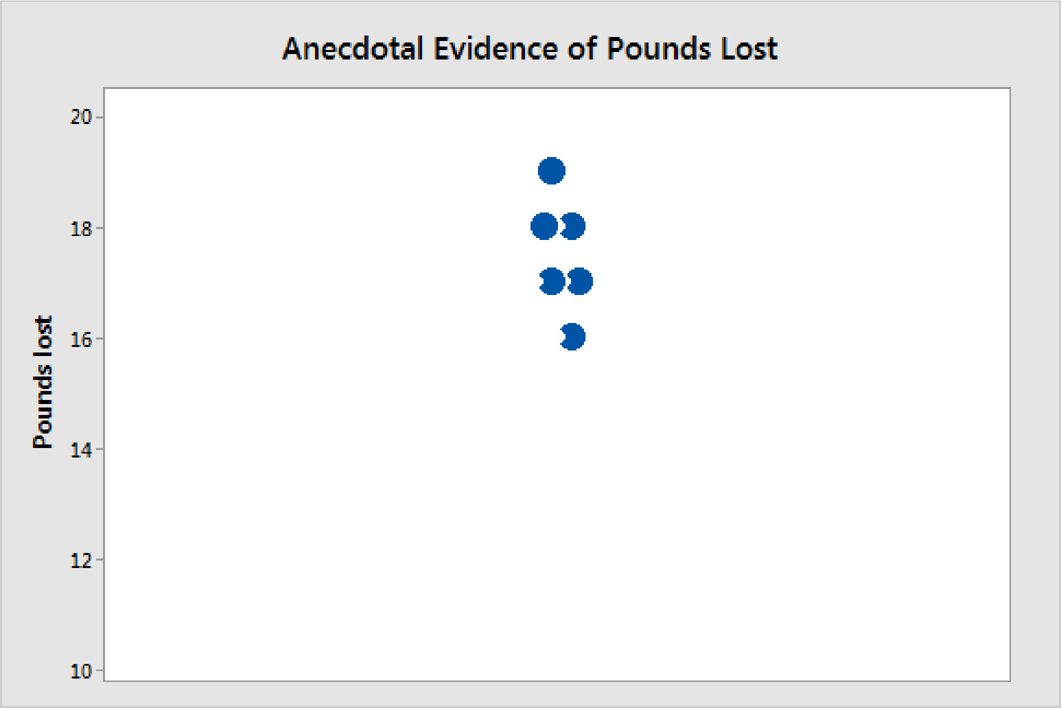
Currency in USD | Download

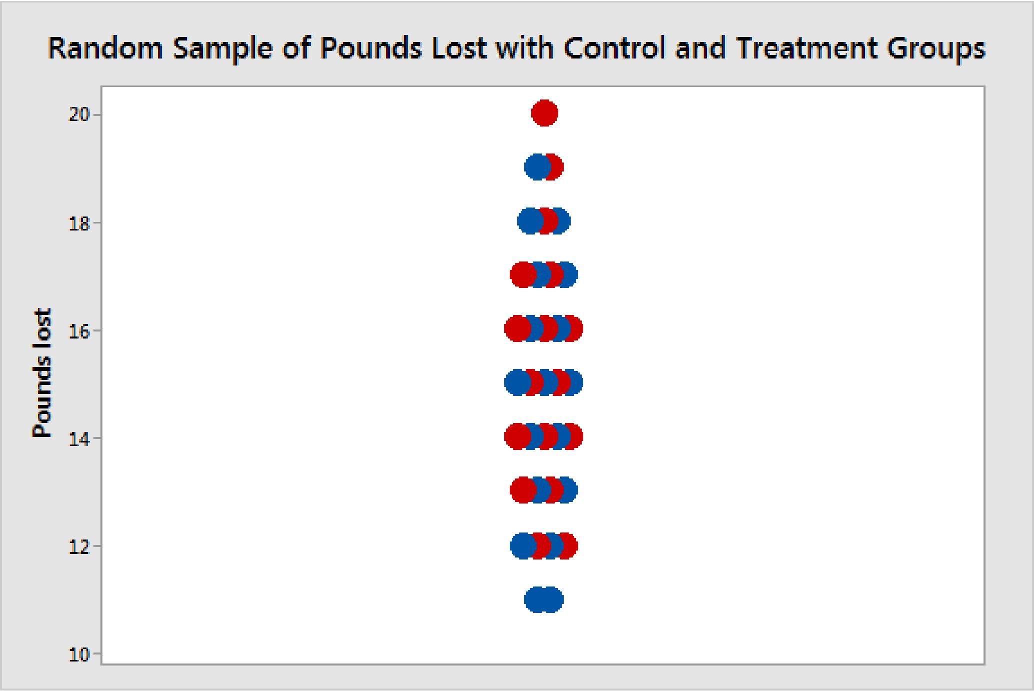
Date	Open	High	Low	Close	Adj Close	Volume
Jun 6, 2017	2,690.84	2,999.91	2,690.84	2,863.20	2,863.20	2,089,609,984
Jun 5, 2017	2,512.40	2,686.81	2,510.22	2,686.81	2,686.81	1,369,309,952
Jun 4, 2017	2,547.79	2,585.89	2,452.54	2,511.81	2,511.81	1,355,120,000

Statistics and anecdotes: Monty Hall problem



Anecdote induction





Beautiful stories

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › scientific correspondence › article

Scientific Correspondence | Published: 01 October 1993

Music and spatial task performance

Frances H. Rauscher, Gordon L. Shaw & Catherine N. Ky

Nature **365**, 611 (1993) | Cite this article

The influence of Mozart's music on brain activity in the process of learning.

Jausovec N [1], Jausovec K , Gerlic I

Author information ›

Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology, 06 Oct 2006, 117(12):2703-2714
https://doi.org/10.1016/j.clinph.2006.08.010 PMID: 17029951

Share this article

Abstract

Objective

The study investigated the influence Mozart's music has on brain activity in the process of learning. A second objective was to test priming explanation of the Mozart effect.

Behavior & Belief

The Mozart Effect Lives On

Stuart Vyse

September 21, 2023

Every now and again, it's worth looking back at old unsupported ideas that we thought were dead and buried because, like zombies, they sometimes climb out of their graves and stagger into the future. So, when I came across a recent mention of Mozart in a psychological study, I was not entirely surprised by what I dug up.

Mozart Effect Background

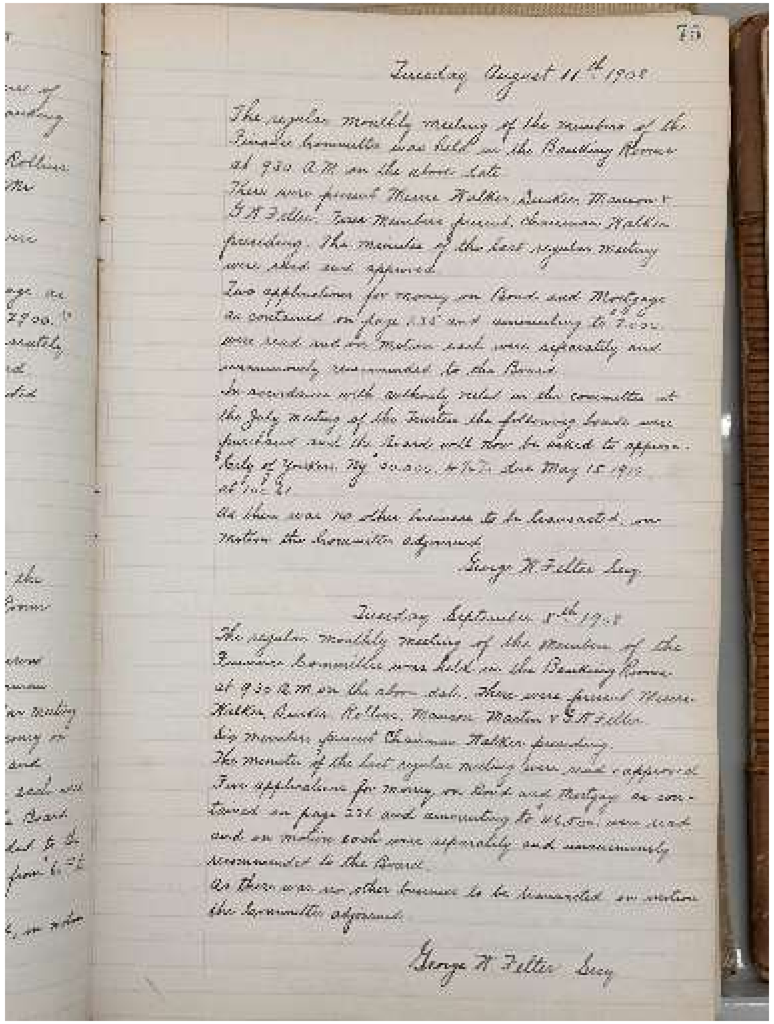
As you may recall, back in 1993, three University of California Irvine psychologists published a study in Nature, one of the world's most prestigious science journals, showing that college students who listened to ten minutes of Mozart's sonata in D for two pianos, K. 448 performed significantly better at a spatial reasoning test than when they heard a relaxation tape or silence (Rauscher et al 1993). Because spatial reasoning is a component of IQ, the authors calculated that the improved performance was equivalent to an eight- to nine-point improvement in spatial IQ. Before long, the media got wind of the Mozart study, and things got crazy.

Bad data analysis endowment effect!

Is your data analysis correct all the time?

- We capture too much data of bad quality.
- ***We fear deleting data and keep torturing it.***
- ***Eventually, there will be minimal cleanup instead of a methodical selection of the highest valuables to keep.***

Analog data with quality



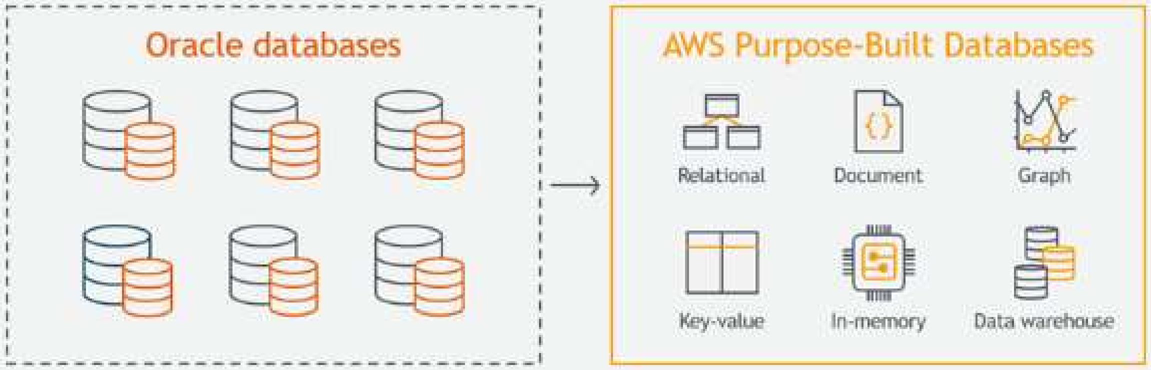
Digital quantity and quality

DataVivid.com	DataVanity.com	DataVague.com
Only good data	Some good data, some bad!	Only bad data
Only correct analysis	Some correct analysis, some wrong!	Incorrect data analysis
Auto data cleansing	Reactive data purge emergencies	Never delete data

Digital data “operations”

KTLO, BAU, RTE, Operations

Data migration
Degragmentation
Cache
Sharding
Geo-replication



AWS News Blog

Migration Complete – Amazon’s Consumer Business Just Turned off its Final Oracle Database

by Jeff Barr | on 15 OCT 2019 | in Database, Launch, Migration & Transfer Services, News | Permalink | Share

“We migrated 75 petabytes of internal data stored in nearly 7,500 Oracle databases to multiple AWS database services…”

Data pipelines

Data collection from “environment”

Vietnam war scenario

Q: *Have you taken any illegal drug in the last 12 months?*

Related scenarios

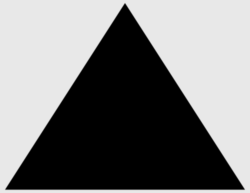
- *Would you vote for [candidate]?*
- *Would you buy [product]?*

Mitigation: statistical data sourcing

Have you taken any form of illegal drugs in the last 12 months?

400

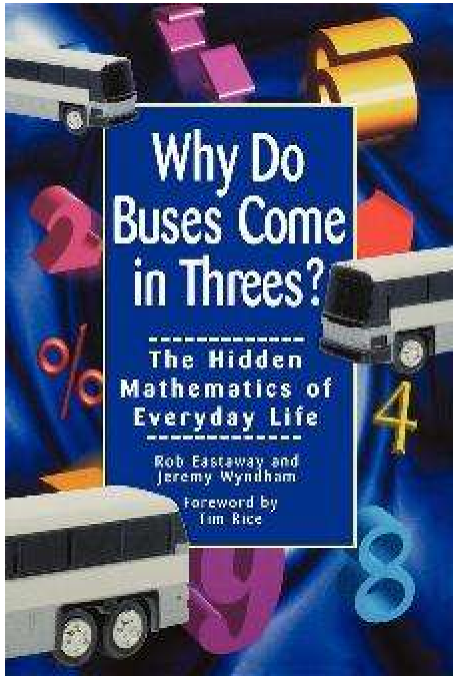
Is there a black triangle in this card?



400

Is there a black triangle in this card?

400



Data: 1,200 answers
Yes: 560
160/400 = 40%

Incongruent data is usually discarded…

Student	Before (Score)	After (Score)
Alice	70	85
Bob	65	80
Carol	80	82
David	90	88

Outliers!

- Incorrect data is wrong.
- Incongruent data doesn’t fit the surrounding information.

Imputed and synthetic data

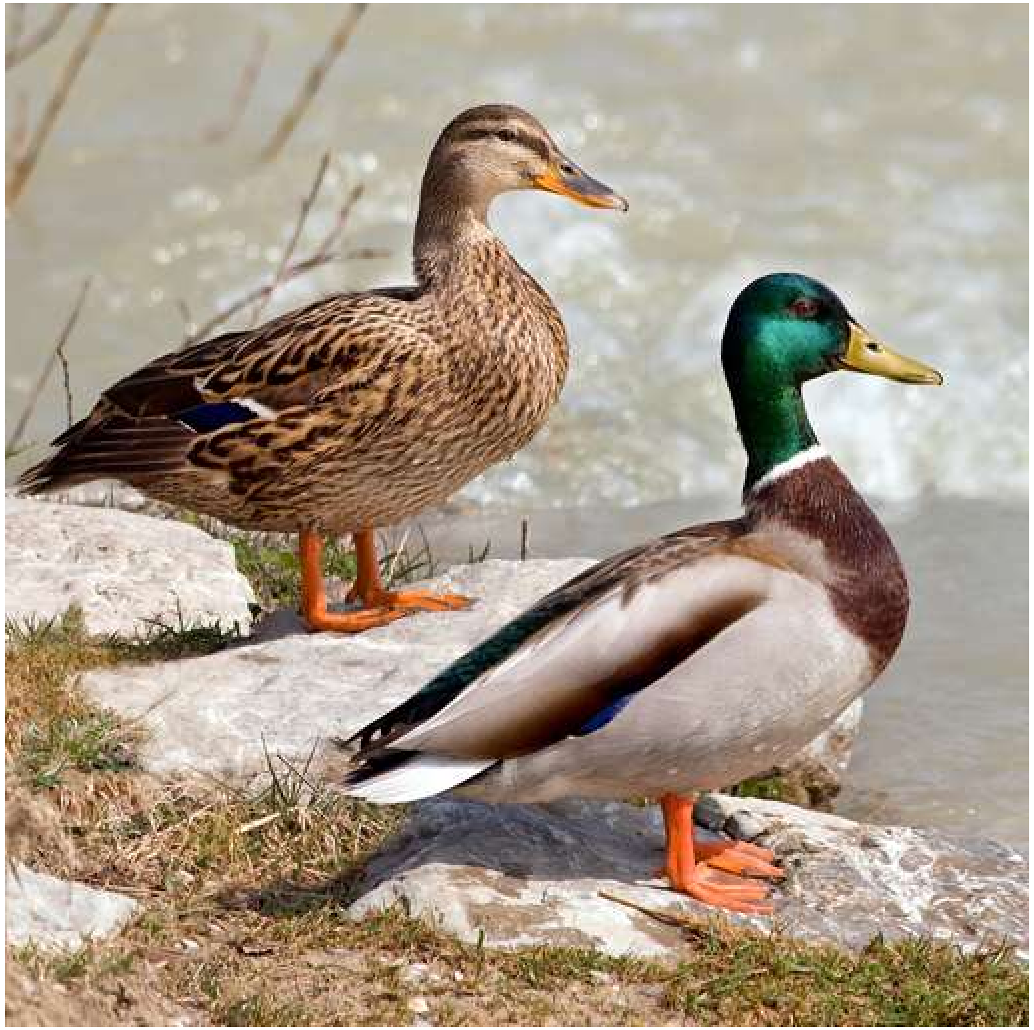
- Imputed: filling data gaps with estimated, plausible values
- Synthetic: artificially generated information

Historical Figure	Birth	Height (cm)	Cause of Death	Favorite Pastimes	Preferred Color
Cleopatra	-69	152	Suicide	Diplomacy, Studying	Gold
Henry VIII	1491	188	Heart Failure	Jousting, Hunting	Gold, Crimson
Abraham Lincoln	1809	193	Assassination	Reading, Storytelling	Black
Alan Turing	1912	175	Cyanide Poisoning	Running, Cryptography	Grey
Queen Elizabeth II	1926	163	Old Age	Horse Racing, Corgis	Blue

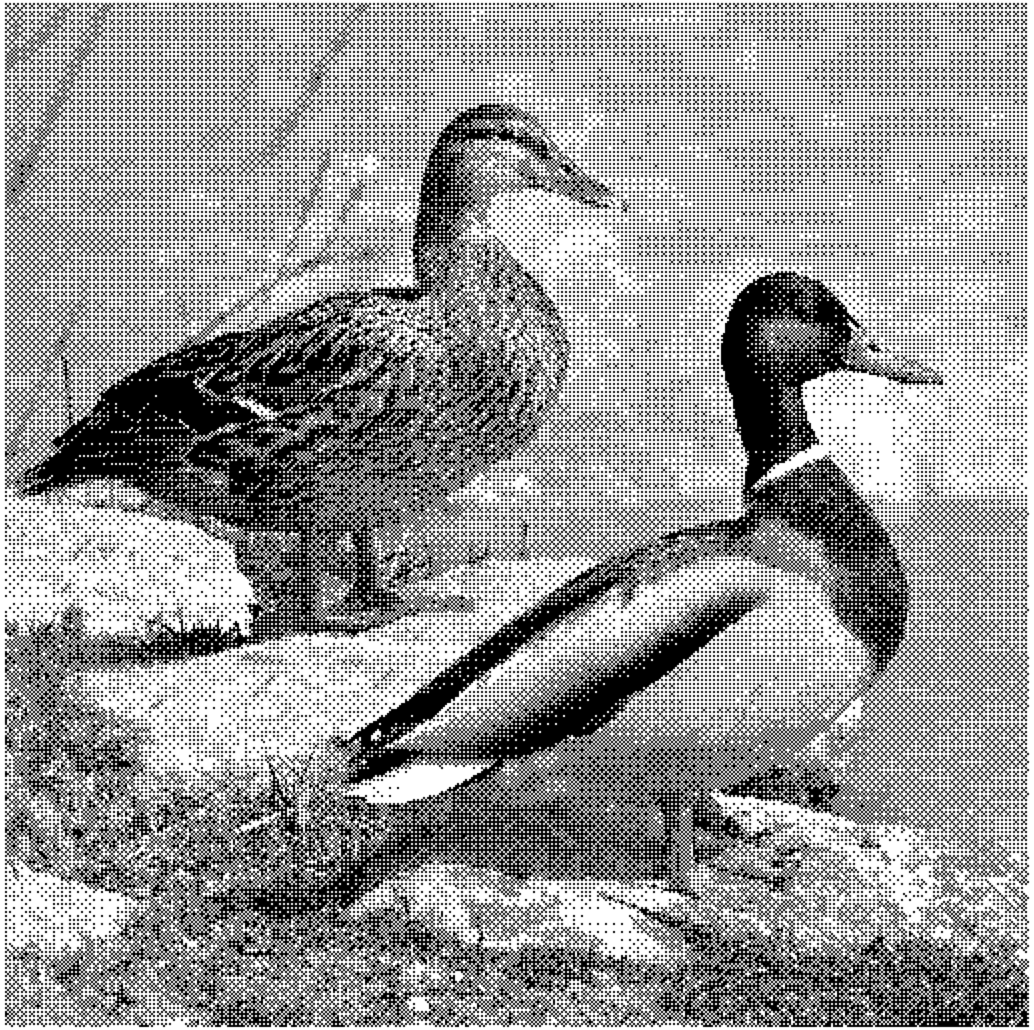
ETL: Transformations

- Quantization
- Encoding/embedding

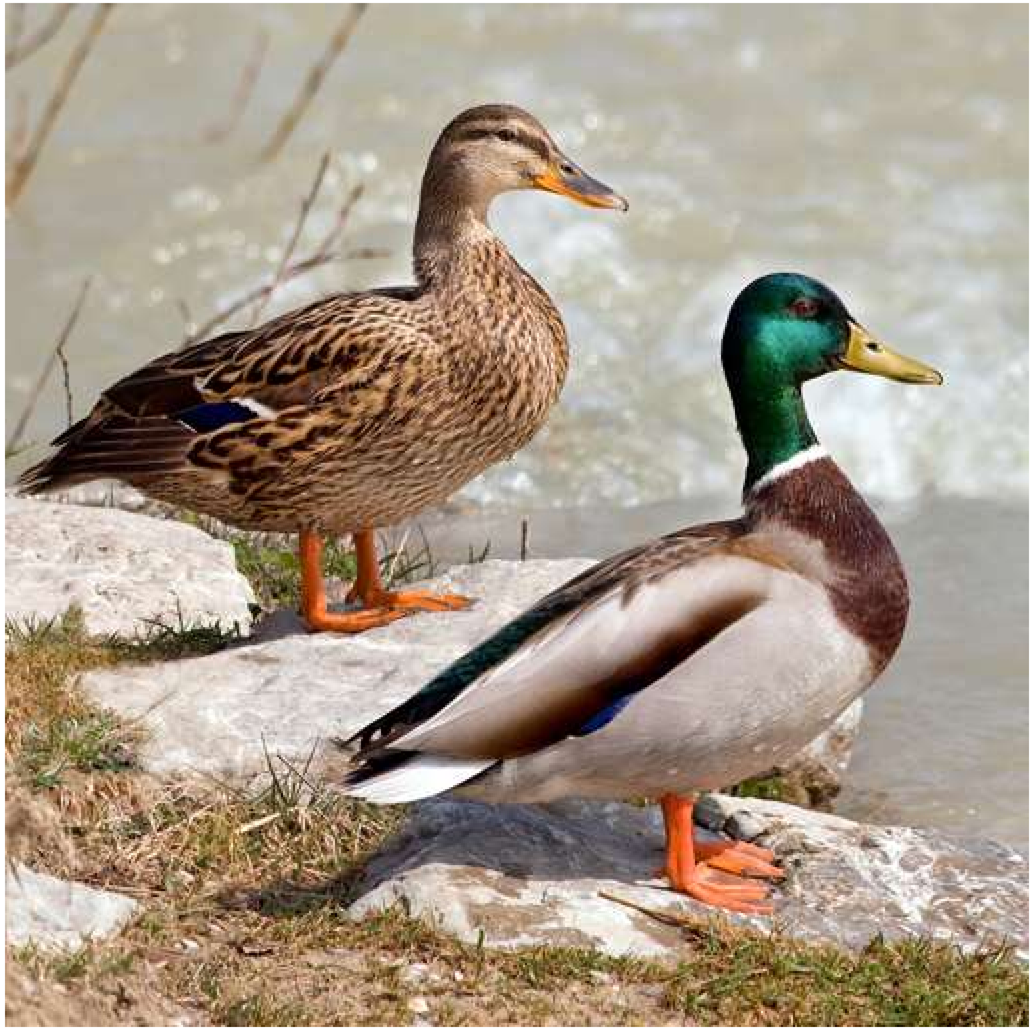
When analog signal become bits and bytes…

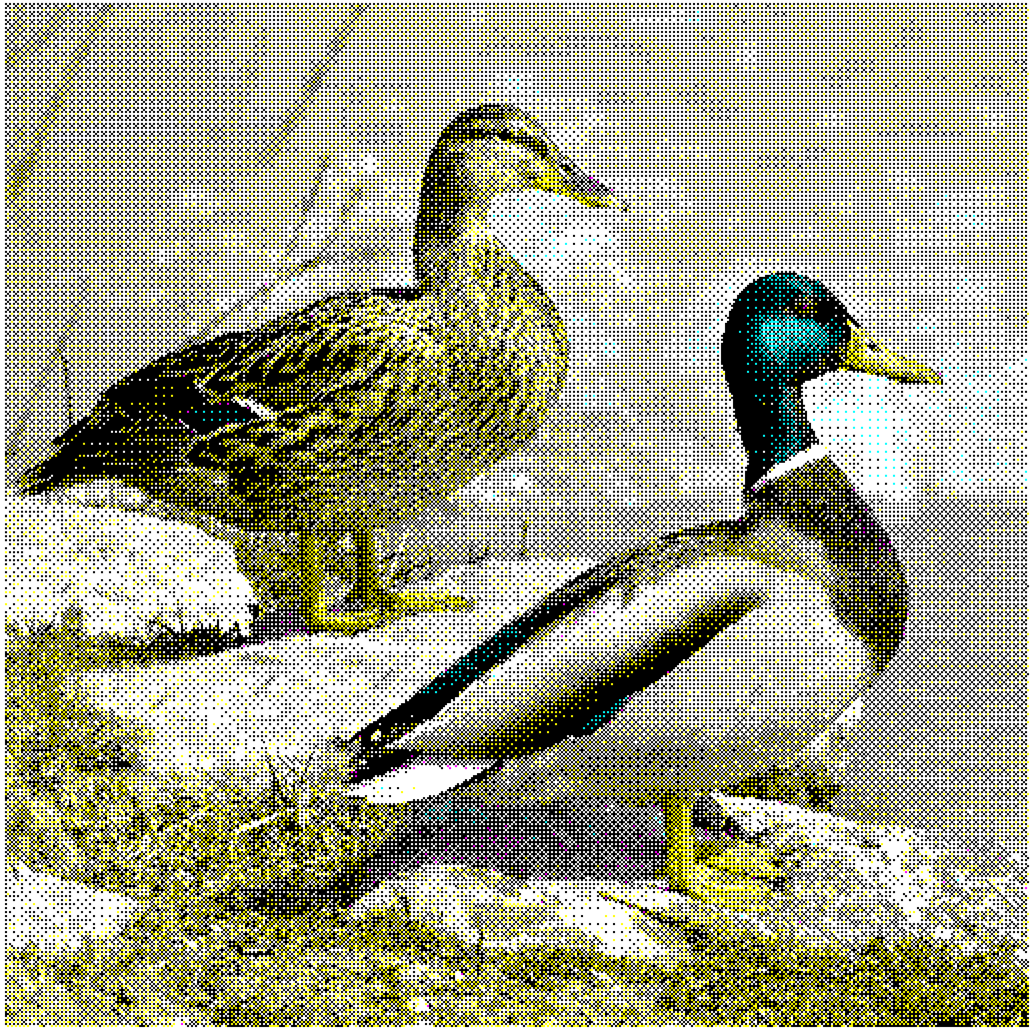


B&W + Bayer



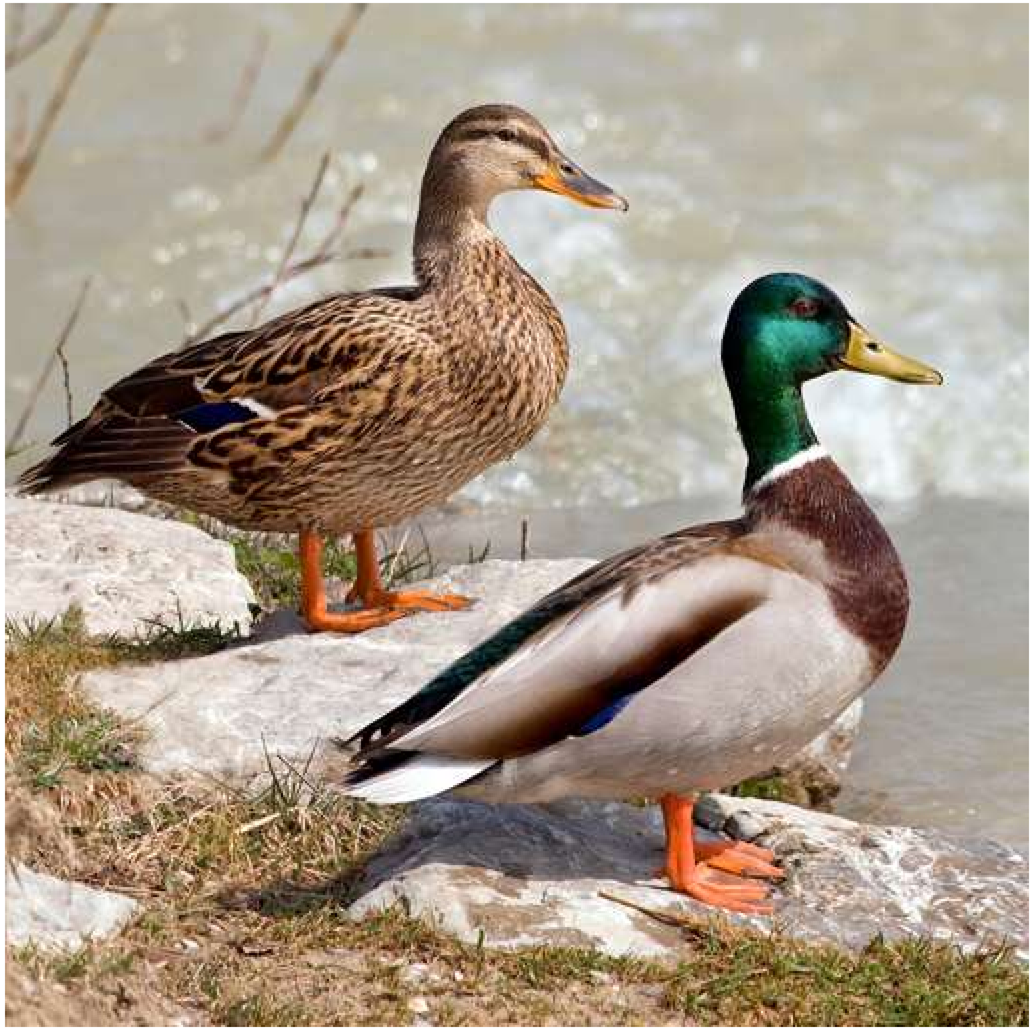
Dithering (CMYB + Bayer)

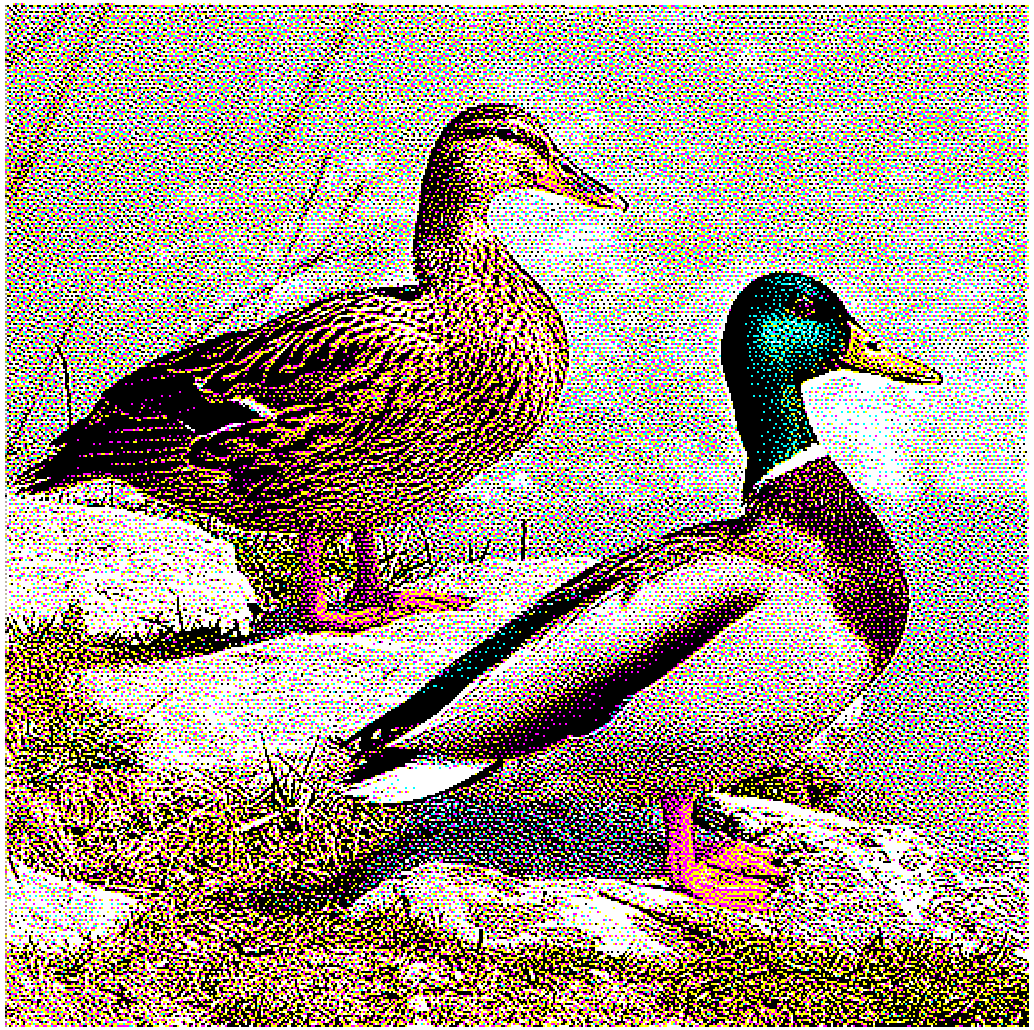




Dithering (CMYB, Err. Diffusion)

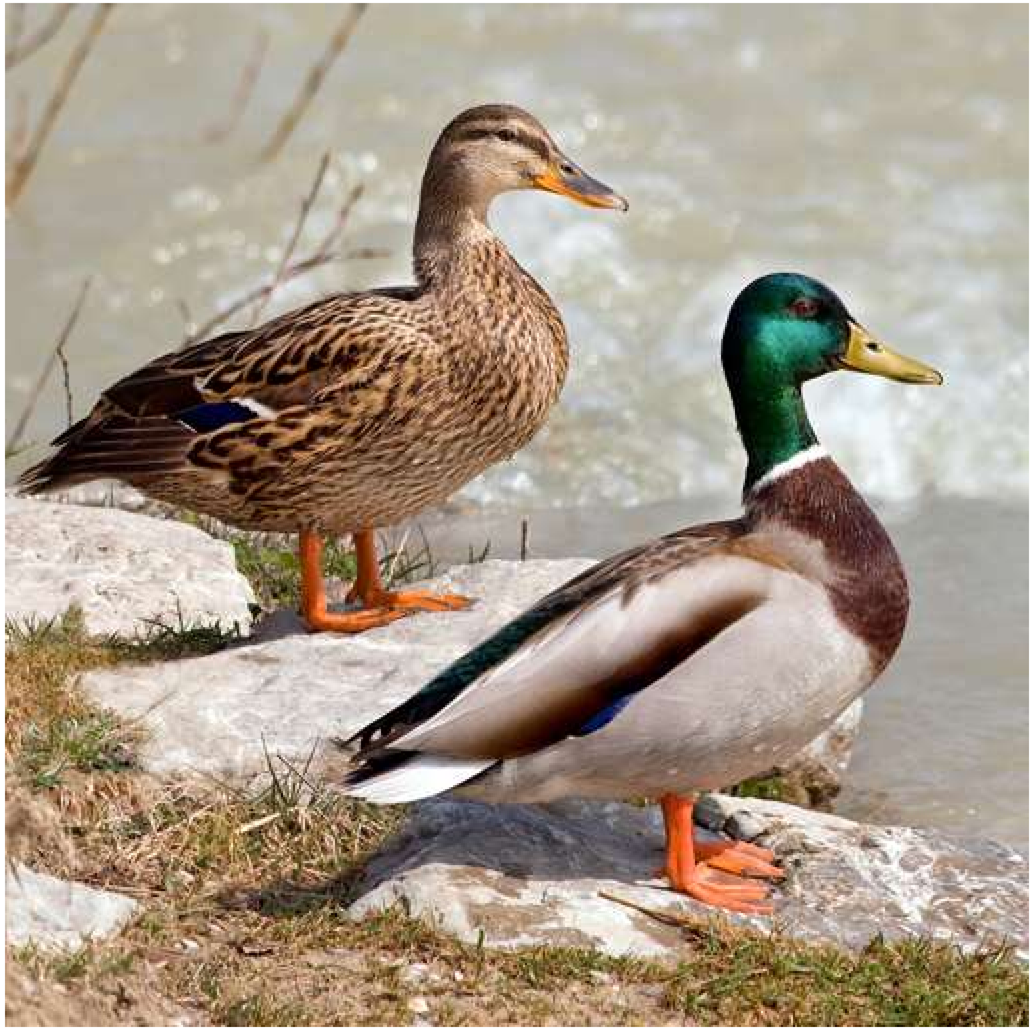
512px
Jarvis-Judice-Ninke
Serpentine Order

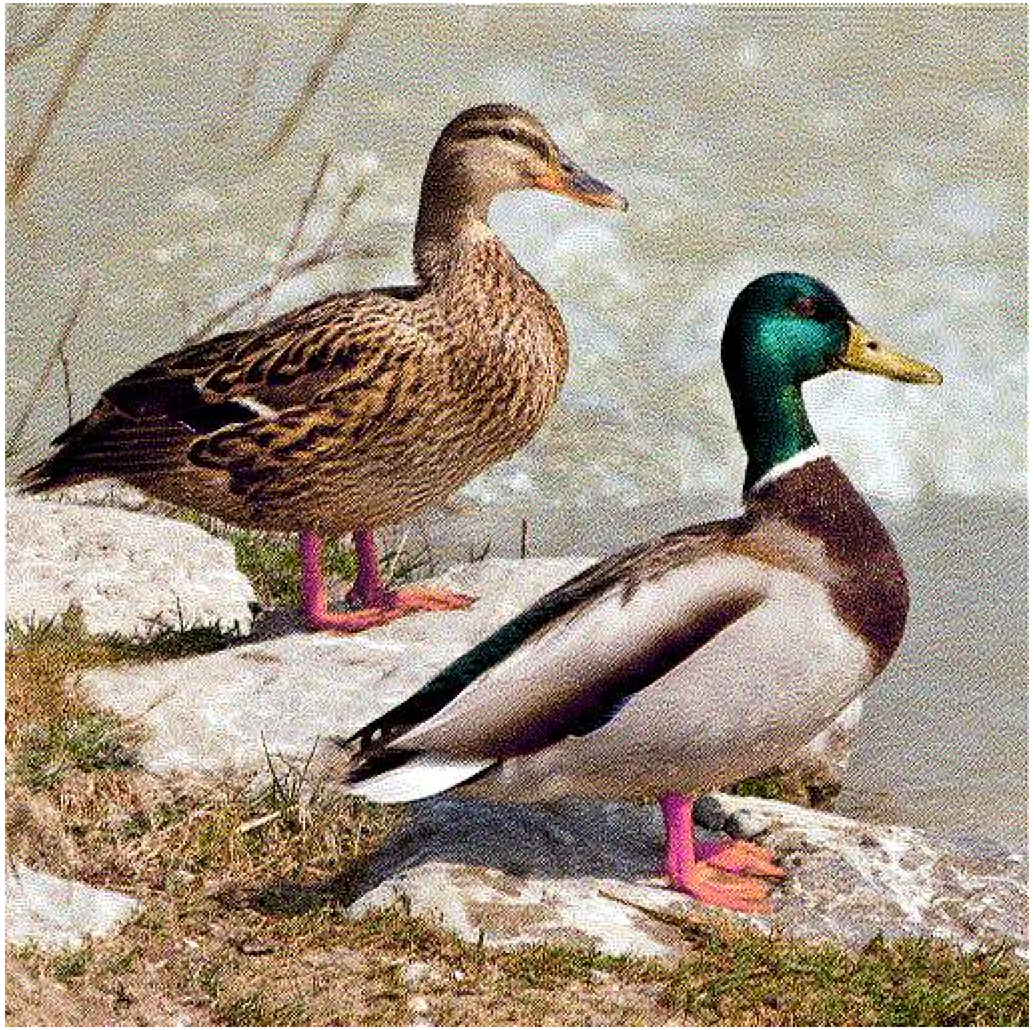




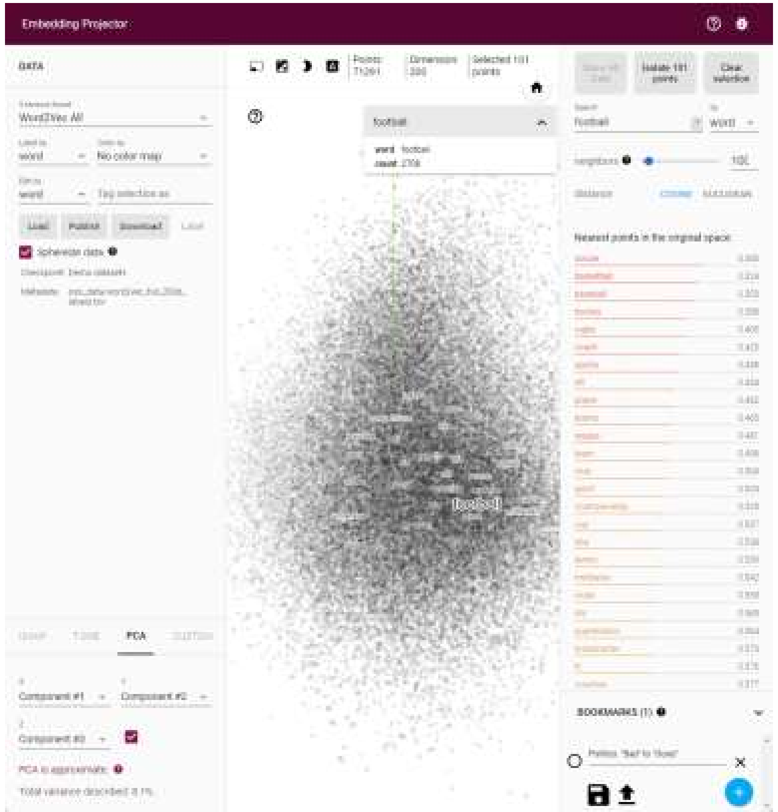
Dithering (CMYB, Err. Diffusion, +Res)

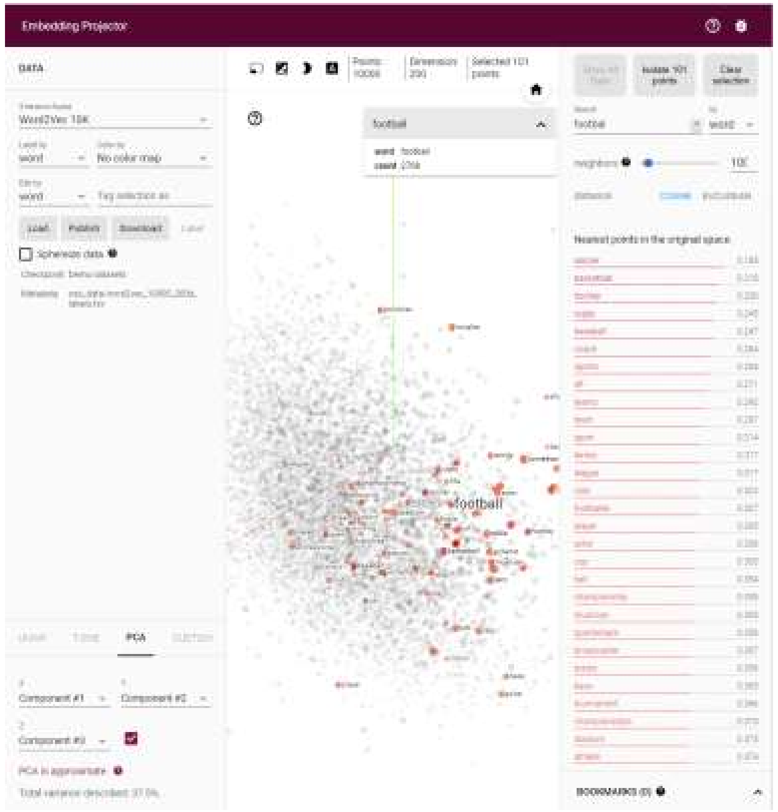
1,024px
Jarvis-Judice-Ninke
Serpentine Order

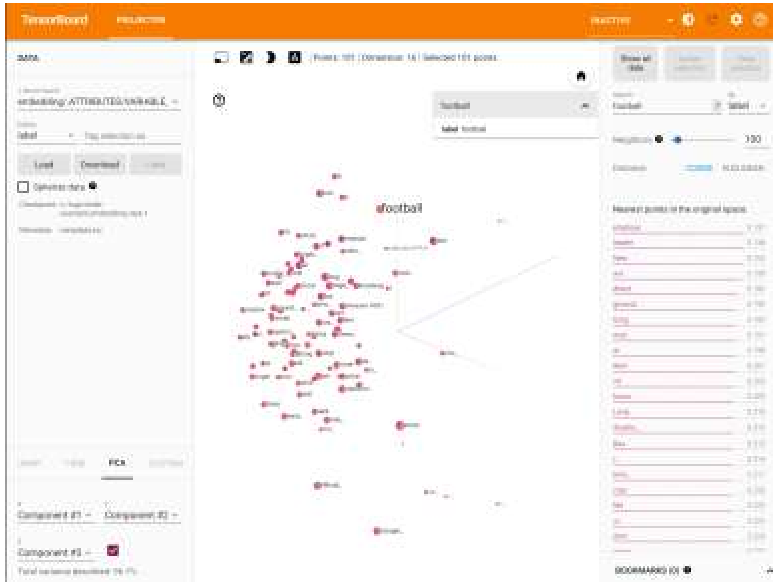




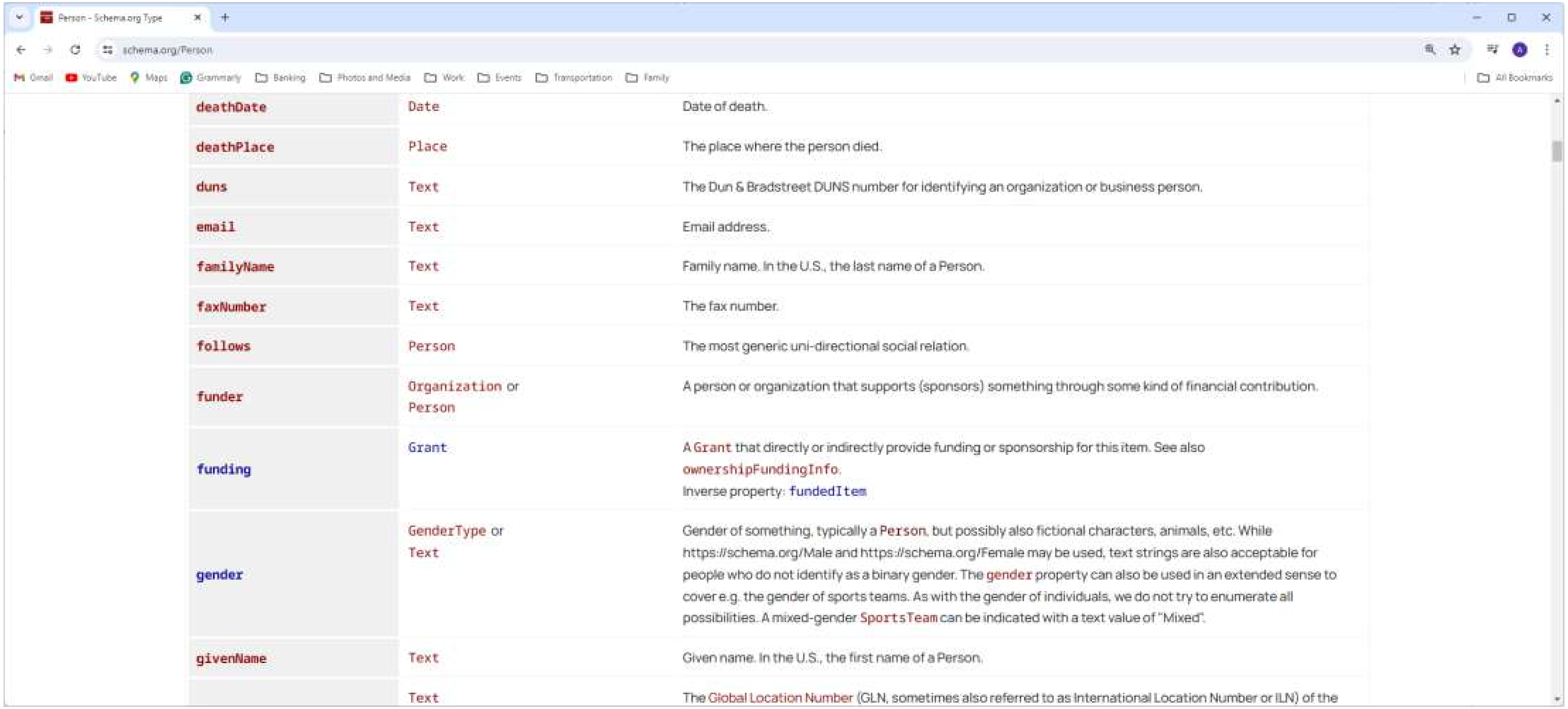
Words, Sentences, Vectors





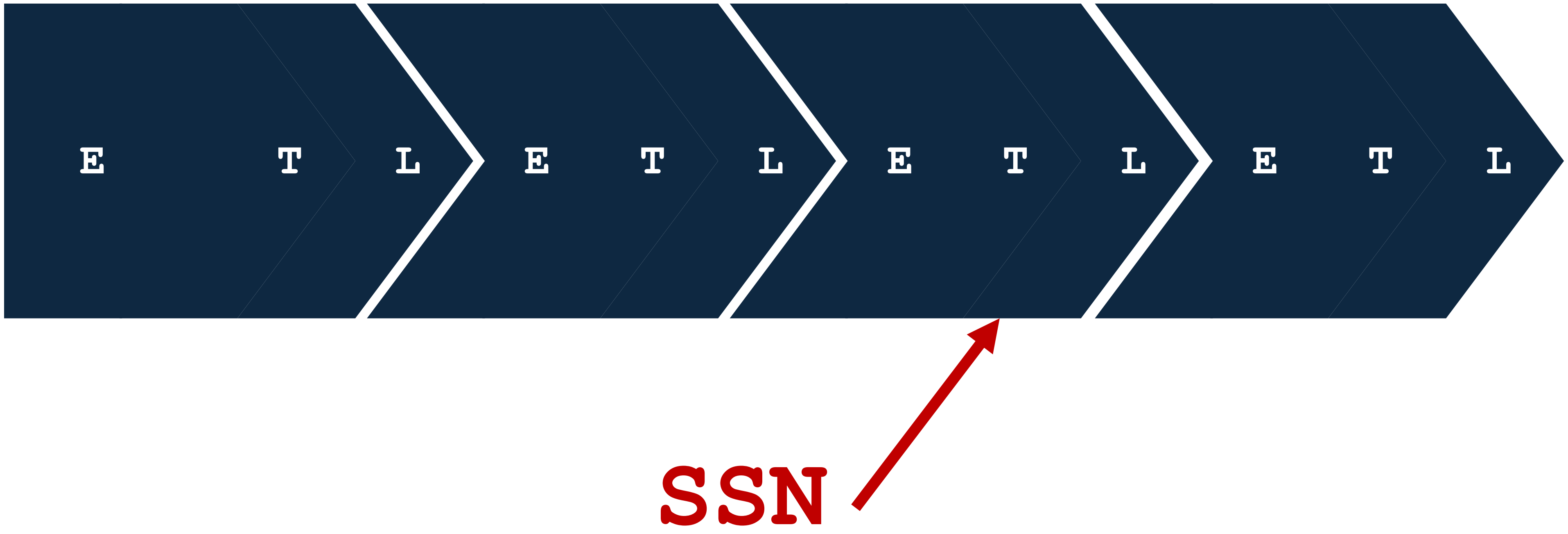


Loading into … destination data model



deathDate	Date	Date of death.
deathPlace	Place	The place where the person died.
duns	Text	The Dun & Bradstreet DUNS number for identifying an organization or business person.
email	Text	Email address.
familyName	Text	Family name. In the U.S., the last name of a Person.
faxNumber	Text	The fax number.
follows	Person	The most generic uni-directional social relation.
funder	Organization or Person	A person or organization that supports (sponsors) something through some kind of financial contribution.
funding	Grant	A Grant that directly or indirectly provide funding or sponsorship for this item. See also ownershipFundingInfo. Inverse property: fundedItem
gender	GenderType or Text	Gender of something, typically a Person, but possibly also fictional characters, animals, etc. While https://schema.org/Male and https://schema.org/Female may be used, text strings are also acceptable for people who do not identify as a binary gender. The gender property can also be used in an extended sense to cover e.g. the gender of sports teams. As with the gender of individuals, we do not try to enumerate all possibilities. A mixed-gender SportsTeam can be indicated with a text value of "Mixed".
givenName	Text	Given name. In the U.S., the first name of a Person.
	Text	The Global Location Number (GLN, sometimes also referred to as International Location Number or ILN) of the

Good data. **But it shouldn't be here!**



Are you sure all your data is “good”?

- ***We capture too much data of bad quality.***
- ***We fear deleting data and keep torturing it.***
- ***Eventually, there will be minimal cleanup instead of a methodical selection of the highest valuables to keep.***

Thank you!

- ***We capture too much data of bad quality.***
- ***We fear deleting data and keep torturing it.***
- ***Eventually, there is some minimal clean-up, instead of methodical selection of the highest valuables to keep.***