Bad Data Analysis

Alisson Sol

Presenter Bio

- 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!

Thank you for DECIDING to be here!

“Bad analysis happen but no miracle analysis can save you from bad data. This presentation will share stories of both bad data and bad analysis, and a few principles to avoid those.”

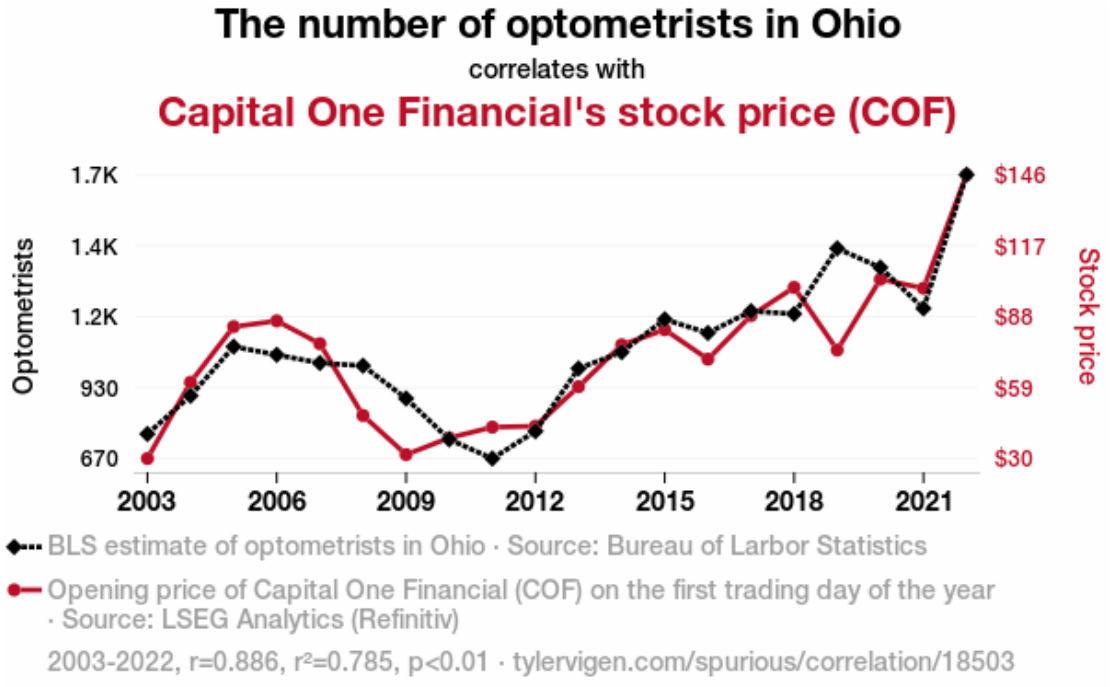
Agenda

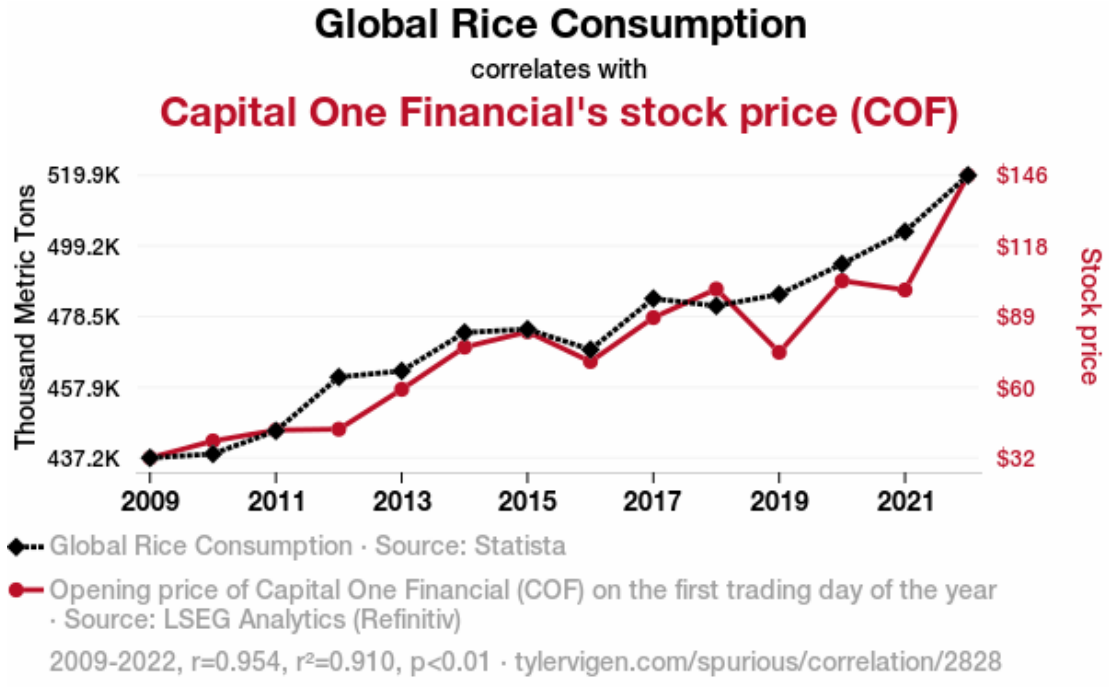
- Bad Data Analysis
- Bad Data, Analysis

Bad Data Analysis: Cherry Picking

Bad Data Analysis: Cherry Picking

- As compute power grew, cherry picking became easier!
- Spurious Correlations





IRL: Bad Data Analysis: Cherry Picking

THE LANCET

Submit Article

This journal Journals Publish Clinical Global health Multimedia Events About

Search for...

EARLY REPORT | VOLUME 351, ISSUE 9103, P637-641, FEBRUARY 28, 1998 Download Full Issue

PDF [942 KB] Figures

RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children

Dr AJ Wakefield, FRCS • SH Murch, MB • A Anthony, MB • J Linnell, PhD • DM Casson, MRCP • M Malik, MRCP • et al.

Show all authors

Published: February 28, 1998 • DOI: https://doi.org/10.1016/S0140-6736(97)11096-0

Summary
Introduction
Patients and methods
Results
Discussion
References
Article info
Figures
Tables
Linked Articles

Summary

Background

We investigated a consecutive series of children with chronic enterocolitis and regressive developmental disorder.

Methods

12 children (mean age 6 years [range 3–10], 11 boys) were referred to a paediatric gastroenterology unit with a history of normal development followed by loss of acquired skills, including language, together with diarrhoea and abdominal pain. Children underwent gastroenterological, neurological, and developmental assessment and review of developmental records. Ileocolonoscopy and biopsy sampling, magnetic-resonance imaging (MRI), electroencephalography (EEG), and lumbar puncture were done under sedation. Barium follow-through radiography was done where possible. Biochemical, haematological, and immunological profiles were examined.

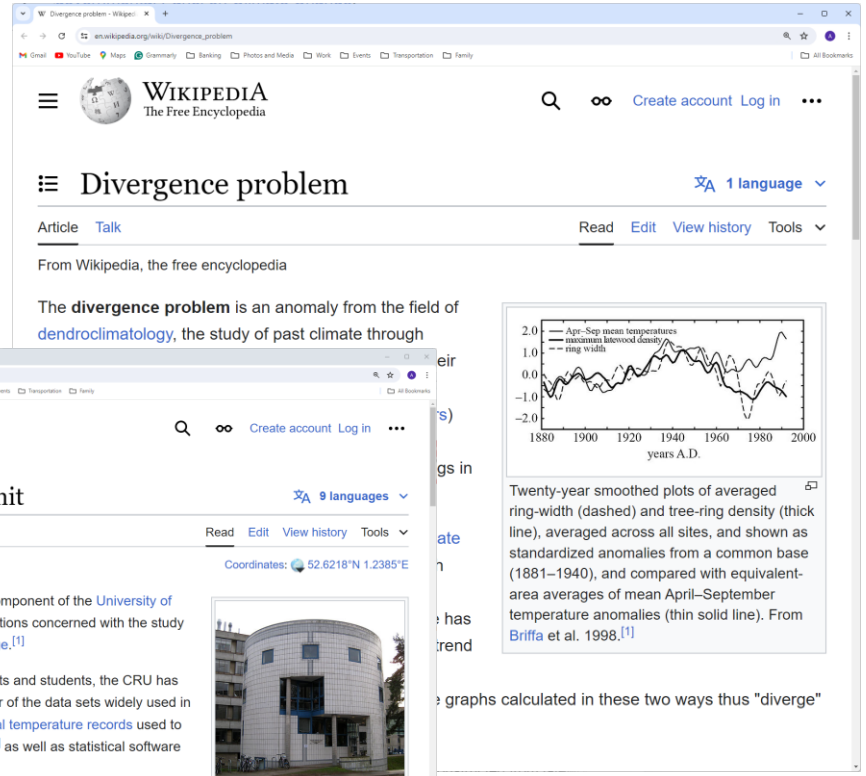
RETRACTED

Findings

Onset of behavioural symptoms was associated, by the parents, with measles, mumps, and rubella vaccination in eight of the 12 children, with measles infection in one child, and otitis media in another. All 12 children had intestinal abnormalities, ranging from lymphoid nodular hyperplasia to aphthoid ulceration. Histology showed patchy chronic inflammation in the colon in 11 children and reactive ileal lymphoid hyperplasia in seven, but no granulomas. Behavioural disorders included autism (nine), disintegrative psychosis (one), and possible postviral or vaccinal encephalitis (two). There were no focal neurological abnormalities and MRI and EEG tests were normal. Abnormal laboratory results were significantly raised urinary methylmalonic acid compared with agematched controls (p=0·003), low haemoglobin in four children, and a low serum IgA in four children.

Interpretation

We identified associated gastrointestinal disease and developmental regression in a group of previously normal children, which was generally associated in time with possible environmental triggers.



WIKIPEDIA
The Free Encyclopedia

Create account Log in

Divergence problem

1 language

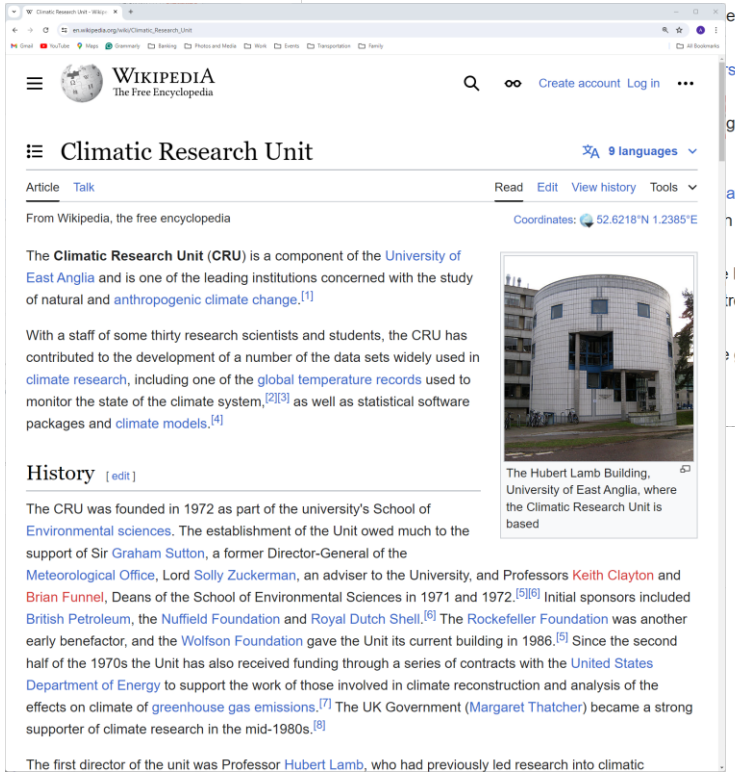
Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

The **divergence problem** is an anomaly from the field of dendroclimatology, the study of past climate through

Twenty-year smoothed plots of averaged ring-width (dashed) and tree-ring density (thick line), averaged across all sites, and shown as standardized anomalies from a common base (1881–1940), and compared with equivalent-area averages of mean April–September temperature anomalies (thin solid line). From Briffa et al. 1998.[1]

graphs calculated in these two ways thus "diverge"



WIKIPEDIA
The Free Encyclopedia

Create account Log in

Climatic Research Unit

9 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 52.6218°N 1.2385°E

The **Climatic Research Unit** (**CRU**) is a component of the University of East Anglia and is one of the leading institutions concerned with the study of natural and anthropogenic climate change.[1]

With a staff of some thirty research scientists and students, the CRU has contributed to the development of a number of the data sets widely used in climate research, including one of the global temperature records used to monitor the state of the climate system,[2][3] as well as statistical software packages and climate models.[4]

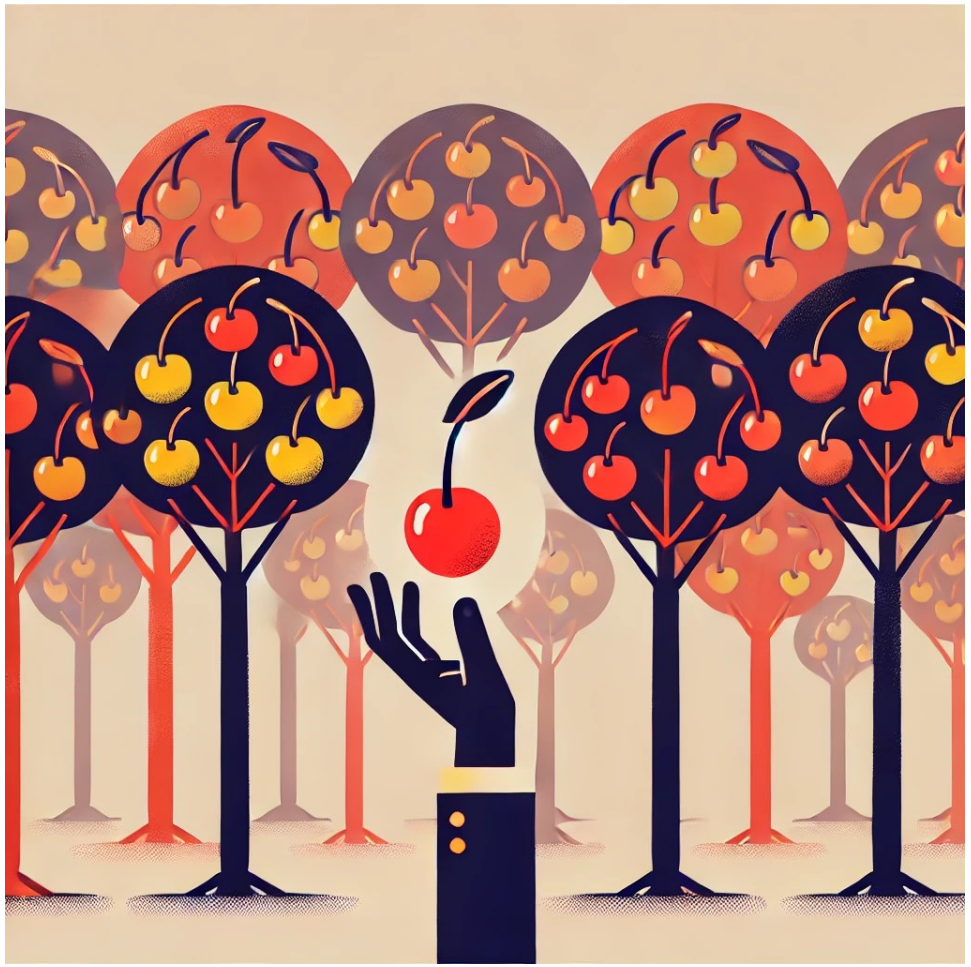
The Hubert Lamb Building, University of East Anglia, where the Climatic Research Unit is based

History [edit]

The CRU was founded in 1972 as part of the university's School of Environmental sciences. The establishment of the Unit owed much to the support of Sir Graham Sutton, a former Director-General of the Meteorological Office, Lord Solly Zuckerman, an adviser to the University, and Professors Keith Clayton and Brian Funnel, Deans of the School of Environmental Sciences in 1971 and 1972.[5][6] Initial sponsors included British Petroleum, the Nuffield Foundation and Royal Dutch Shell.[6] The Rockefeller Foundation was another early benefactor, and the Wolfson Foundation gave the Unit its current building in 1986.[5] Since the second half of the 1970s the Unit has also received funding through a series of contracts with the United States Department of Energy to support the work of those involved in climate reconstruction and analysis of the effects on climate of greenhouse gas emissions.[7] The UK Government (Margaret Thatcher) became a strong supporter of climate research in the mid-1980s.[8]

The first director of the unit was Professor Hubert Lamb, who had previously led research into climatic

Bad Data Analysis: Anecdote Induction

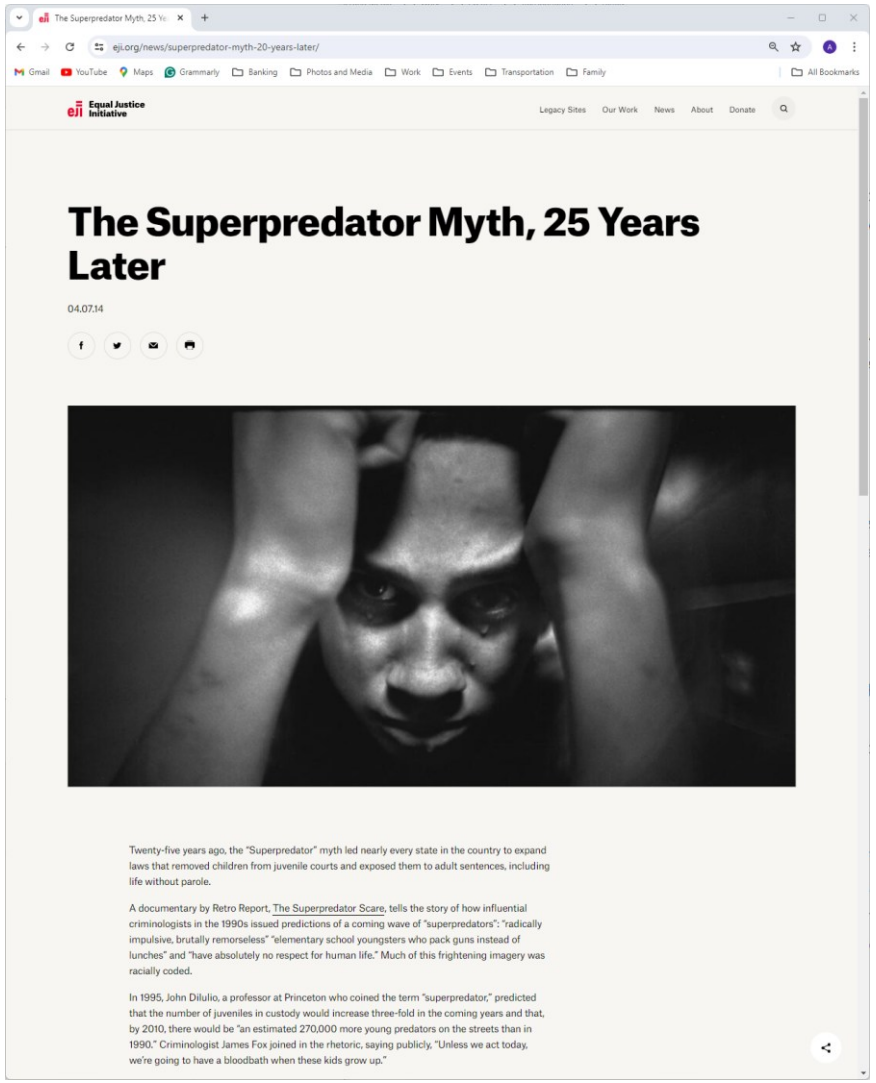


Bad Data Analysis: Anecdote Induction

- Pressing Carbon produces Diamonds
- Mathematical Induction

"Most Brazilians play soccer more than any other country."

IRL: Bad Data Analysis: Anecdote Induction



Equal Justice Initiative

Legacy Sites Our Work News About Donate

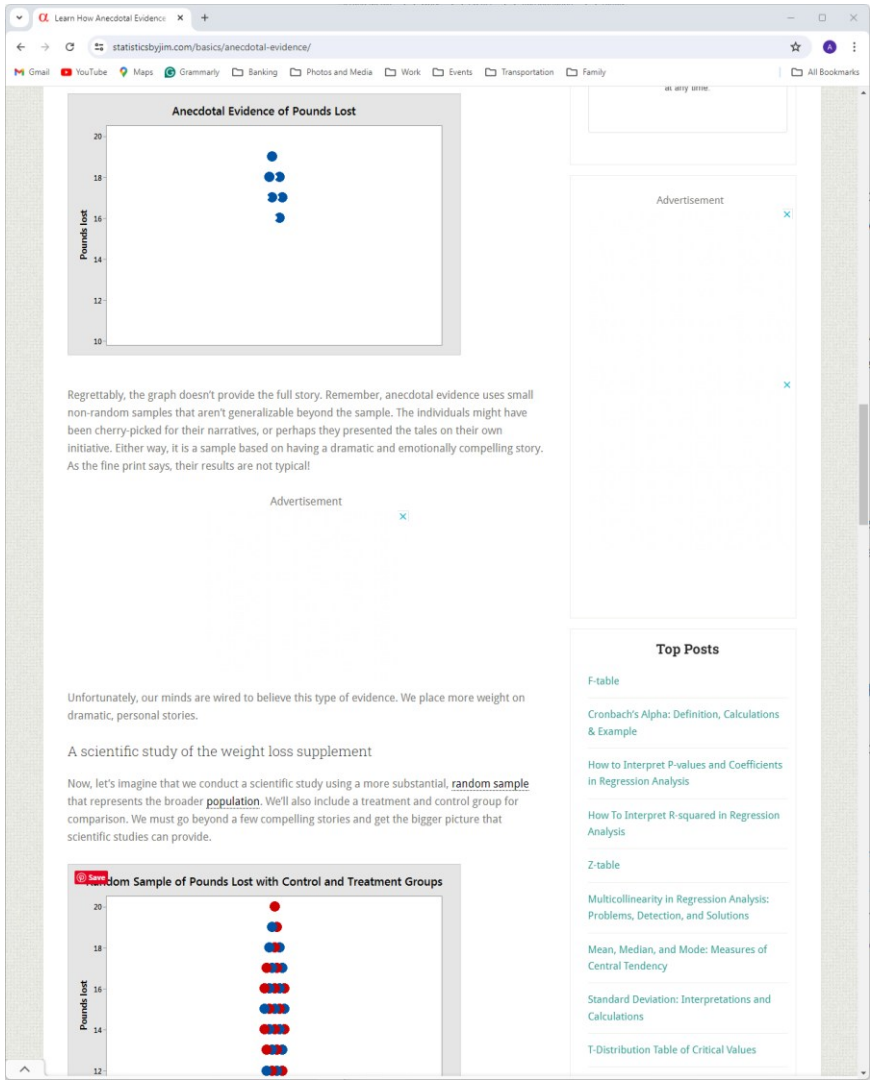
The Superpredator Myth, 25 Years Later

04.07.14

Twenty-five years ago, the "Superpredator" myth led nearly every state in the country to expand laws that removed children from juvenile courts and exposed them to adult sentences, including life without parole.

A documentary by Retro Report, The Superpredator Scare, tells the story of how influential criminologists in the 1990s issued predictions of a coming wave of "superpredators": "radically impulsive, brutally remorseless" "elementary school youngsters who pack guns instead of lunches" and "have absolutely no respect for human life." Much of this frightening imagery was racially coded.

In 1995, John DiIulio, a professor at Princeton who coined the term "superpredator," predicted that the number of juveniles in custody would increase three-fold in the coming years and that, by 2010, there would be "an estimated 270,000 more young predators on the streets than in 1990." Criminologist James Fox joined in the rhetoric, saying publicly, "Unless we act today, we're going to have a bloodbath when these kids grow up."



Anecdotal Evidence of Pounds Lost

Pounds lost

Regrettably, the graph doesn't provide the full story. Remember, anecdotal evidence uses small non-random samples that aren't generalizable beyond the sample. The individuals might have been cherry-picked for their narratives, or perhaps they presented the tales on their own initiative. Either way, it is a sample based on having a dramatic and emotionally compelling story. As the fine print says, their results are not typical!

Advertisement

Unfortunately, our minds are wired to believe this type of evidence. We place more weight on dramatic, personal stories.

A scientific study of the weight loss supplement

Now, let's imagine that we conduct a scientific study using a more substantial, random sample that represents the broader population. We'll also include a treatment and control group for comparison. We must go beyond a few compelling stories and get the bigger picture that scientific studies can provide.

dom Sample of Pounds Lost with Control and Treatment Groups

Pounds lost

Advertisement

Top Posts

F-table

Cronbach's Alpha: Definition, Calculations & Example

How to Interpret P-values and Coefficients in Regression Analysis

How To Interpret R-squared in Regression Analysis

Z-table

Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

Mean, Median, and Mode: Measures of Central Tendency

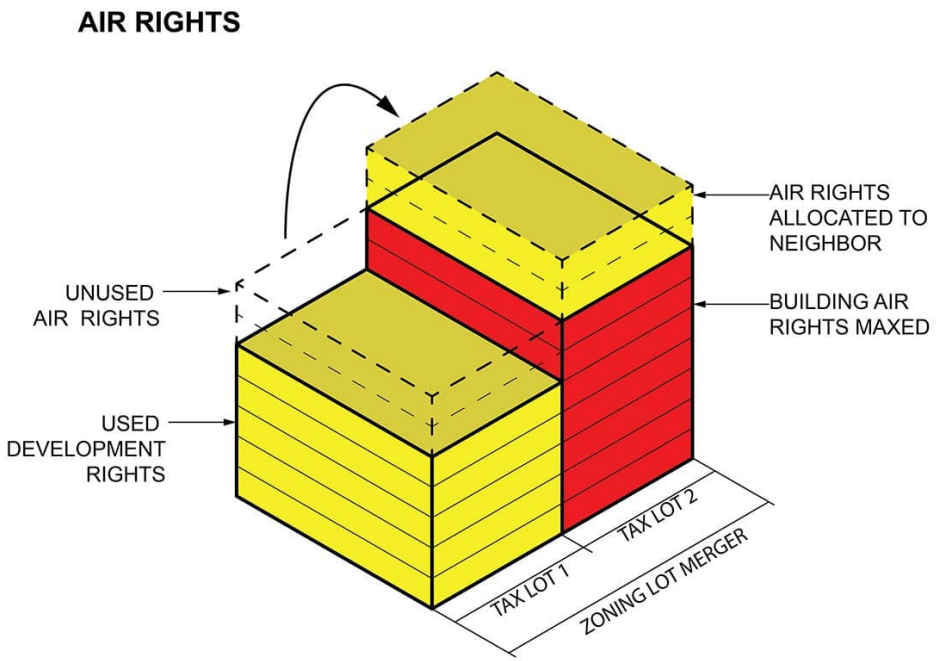
Standard Deviation: Interpretations and Calculations

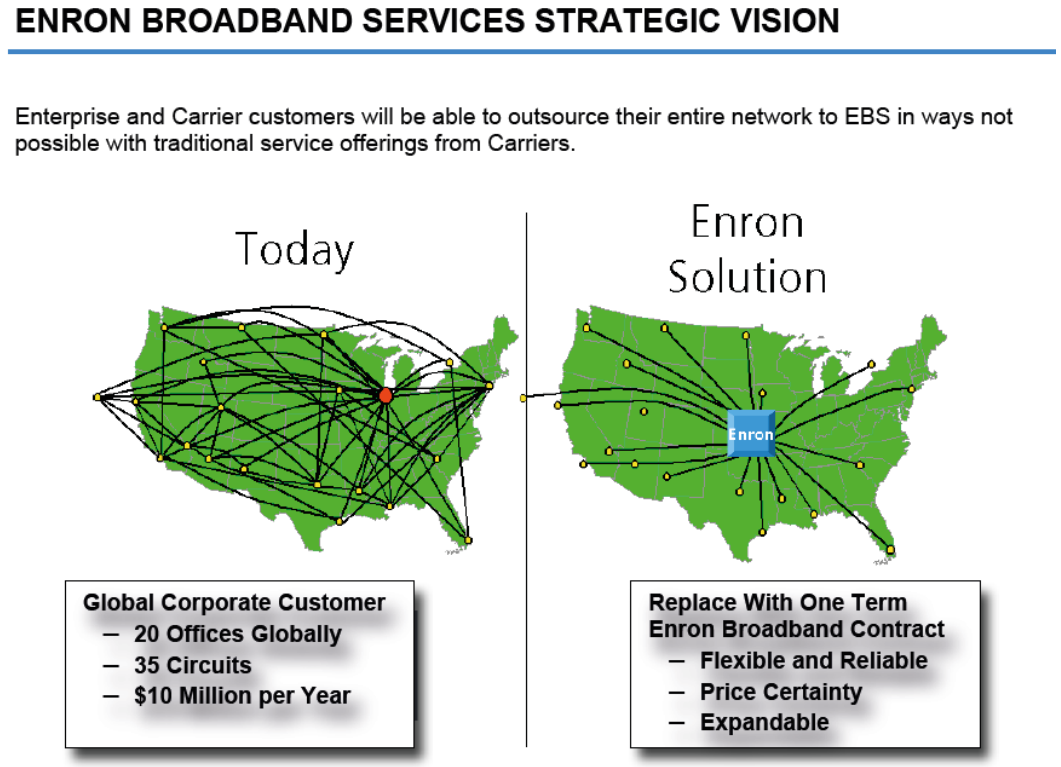
T-Distribution Table of Critical Values

Bad Data Analysis: Beautiful Stories

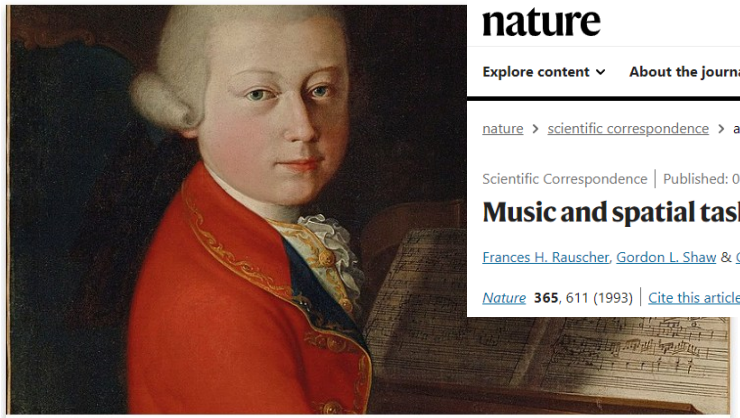
Bad Data Analysis: Beautiful Stories

- Storytelling as a disservice
- Dangerous analogies, with supporting data!





IRL: Bad Data Analysis: 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

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.

frontiers in Immunology

Front Immunol. 2020; 11: 574029. PMCID: PMC7655735
Published online 2020 Oct 28. doi: 10.3389/fimmu.2020.574029 PMID: 33193359

The Long History of Vitamin C: From Prevention of the Common Cold to Potential Aid in the Treatment of COVID-19

Giuseppe Cerullo,[1,†] Massimo Negro,[2,*†] Mauro Parimbelli,[2] Michela Pecoraro,[3] Simone Perna,[4] Giorgio Liguori,[1] Mariangela Rondanelli,[5,6] Hellas Cena,[6,7] and Giuseppe D'Antona[2,6]

Author information | Article notes | Copyright and License information PMC Disclaimer

See commentary "Commentary: The Long History of Vitamin C: From Prevention of the Common Cold to Potential Aid in the Treatment of COVID-19" in volume 12, 659001.

Journal Article Journal Article Journal Article

Abstract

Go to:

From Pauling's theories to the present, considerable understanding has been acquired of both the physiological role of vitamin C and of the impact of vitamin C supplementation on the health. Although it is well known that a balanced diet which satisfies the daily intake of vitamin C positively affects the immune system and reduces susceptibility to infections, available data do not support the theory that oral vitamin C supplements boost immunity. No current clinical recommendations support the possibility of significantly decreasing the risk of respiratory infections by using high-dose supplements of vitamin C in a well-nourished general population. Only in restricted subgroups (e.g., athletes or the military) and in subjects with a low plasma vitamin C concentration a supplementation may be justified. Furthermore, in categories at high risk of infection (i.e., the obese, diabetics, the elderly, etc.), a vitamin C supplementation can modulate inflammation, with potential positive effects on immune response to infections. The impact of an extra oral intake of vitamin C on the duration of a cold and the prevention or treatment of pneumonia is still questioned, while, based on critical illness studies, vitamin C infusion has recently been hypothesized as a treatment for COVID-19 hospitalized patients. In this review, we focused on the effects of vitamin C on immune function, summarizing the most relevant studies from the prevention and treatment of common respiratory diseases to the use of vitamin C in critical illness conditions, with the aim of clarifying its potential application during an acute SARS-CoV2 infection.

Keywords: vitamin C supplementation, viral infections, COVID-19, pneumonia, immune function, athletes, non-communicable diseases, frail elderly subjects

Bad Data Analysis: Mitigations

Hyphotheses & Placebos!

- **Cherry Picking:** *Representative and comprehensive dataset. Peer reviews and blind analysis.*
- **Anecdote Induction:** *Big data! Transparency.*
- **Beautiful Stories:** *Don't start with a story or the data will confess! Focus on key metrics.*

Roll over image to zoom in

BNT Dummy Fake Security Camera, with One Red LED Light at Night, for Home and Businesses Security Indoor/Outdoor (2 Pack, White)

Visit the BNT Store

4.3 ★★★★½ 7,454 ratings | Search this page

#1 Best Seller in Simulated Surveillance Cameras

1K+ bought in past month

-33% $9.99 ($5.00 / Count)

List Price: $14.99

prime One-Day

FREE Returns

Get a $200 Amazon Gift Card instantly upon approval for Prime Visa. No annual fee.

Color: white 2pack

$24.99 ($12.50 / Count) prime | $45.99 prime | $16.99 ($8.50 / Count) prime

$26.99 ($6.75 / Count) prime | $9.99 ($5.00 / Count) prime | $18.99 ($4.75 / Count) prime

Bundles with this item

BNT Dummy Fake Security Camera,Solar... -17% $32.99 List: $39.98

BNT Dummy Fake Security Camera, Sola... -11% $39.98 List: $44.98

See all bundles

Bad Data, Analysis

- Extract: Bad models, leading questions, data “shenanigans”
- Transform: Encoding and embeddings that help in misinformation
- Load: Poorly modeling data storage and visualization

Bad Data, Analysis: Extraction (Sourcing)

- **Incorrect**: *Volume*
- **Incomplete**: *Velocity*
- **Irrelevant**: *Variety*
- **Incongruent**: *Veracity*

IRL: Bad Data, ...: Intentionally Incorrect

How To Buy Reviews

Looking to buy online reviews? It's easy (and cheap!) to buy them for any review site you want to get more reviews on. But there's some risk involved too. Here we take a look at what's on the line as well as some alternative options.

You can buy a real coyote skull, bacon-flavored floss and a coffin on Amazon, so suffice it to say you can buy pretty much anything these days.

Business owners or managers might want to order bulk office supplies, a "World's Best Boss" mug, and maybe a few five-star reviews...

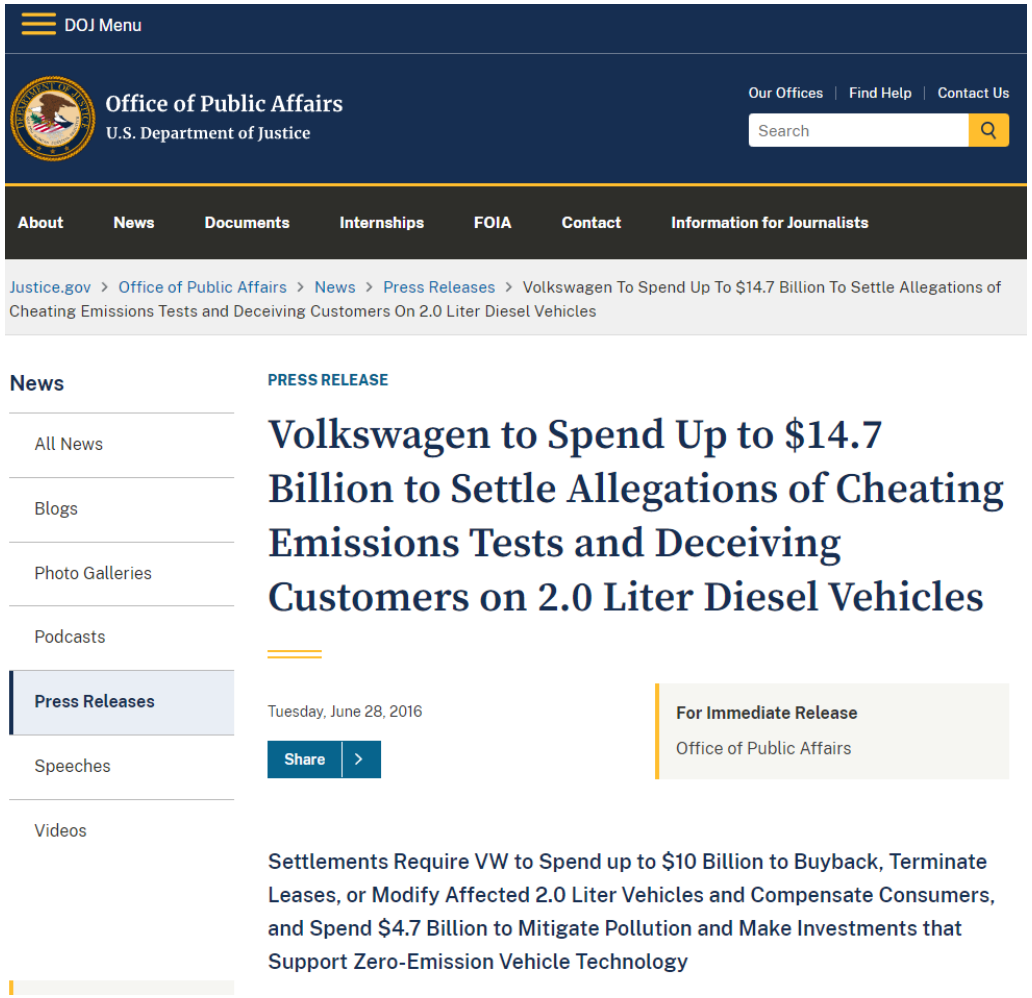
It's remarkably easy to pay for reviews, and often quite tempting too, especially since even just a few positive reviews can have a big impact on business. But there are some risks involved – sometimes *very big risks* (we'll get to that in just a minute).

Buying reviews is also totally unnecessary when getting positive reviews in an ethical way is just as easy (we'll get to *that* in a minute too).

Why Buy Reviews?

The temptation to buy reviews is real. **They're a powerful tool for generating new business.** Most consumers (almost all) consider online reviews when making buying decisions, and they take seriously what they read:

- **85 percent of people trust online reviews** as much as they trust recommendations from friends, according to a recent survey.
- **49 percent said they need a business to have a rating of at least four stars** before using it.



Bad Data, Analysis: Incorrect

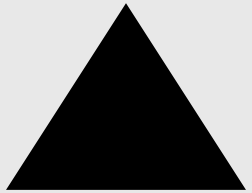
- Data acquisition is complex, time-consuming, and it may be hard to get truthful information

Statistical Data Sourcing

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

400

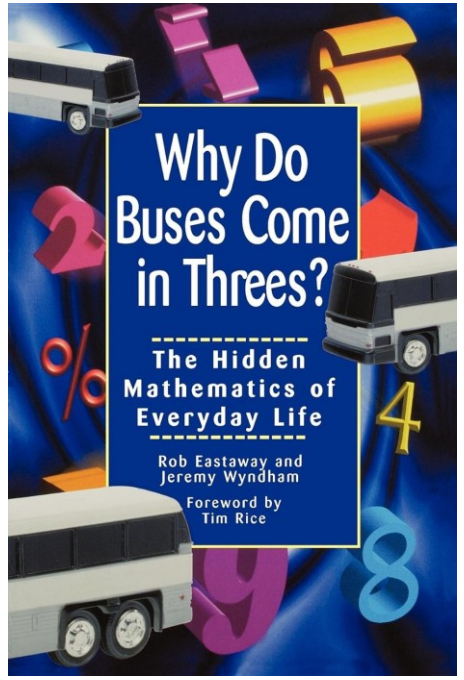
Is there a black triangle in this card?



400

Is there a black triangle in this card?

400



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

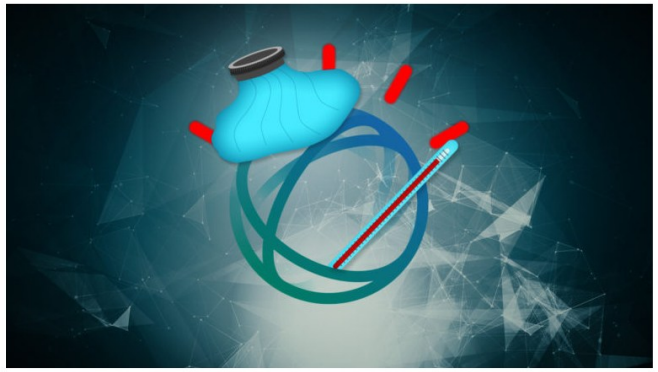
IRL: Bad Data, Analysis: Incorrect

EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By Casey Ross and Ike Swetlitz July 25, 2018 Reprints

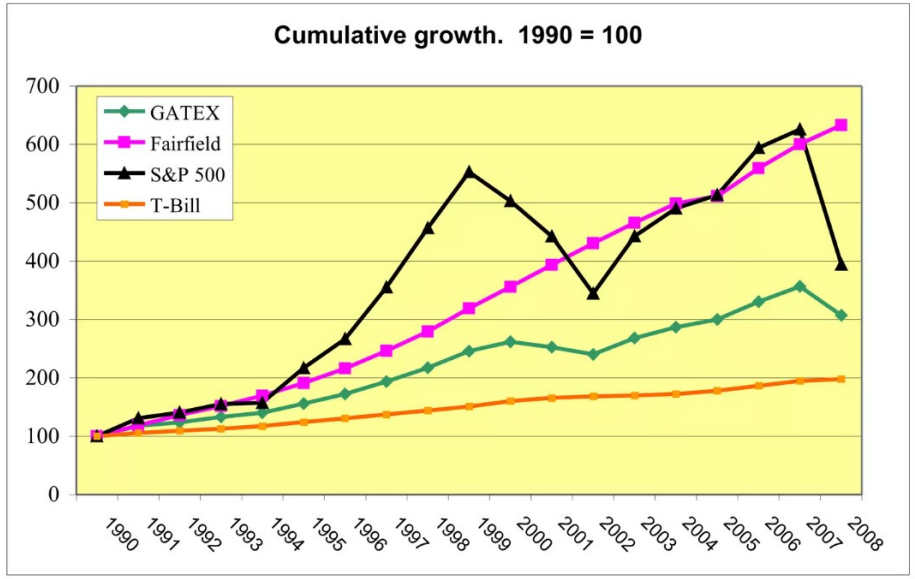


ALEX HOGAN/STAT

Internal IBM documents show that its Watson supercomputer often spit out erroneous cancer treatment advice and that company medical specialists and customers identified "multiple examples of unsafe and incorrect treatment recommendations" as IBM was promoting the product to hospitals and physicians around the world.

The documents — slide decks presented last summer by IBM Watson Health's deputy chief health officer — largely blame the problems on the training of Watson by IBM engineers and doctors at the renowned Memorial Sloan Kettering Cancer Center. The software was drilled with a small number of "synthetic" cancer cases, or hypothetical patients, rather than real patient data. Recommendations were based on the expertise of a few specialists for each cancer type, the documents say, instead of "guidelines or evidence."

And what Madoff's returns supposedly were

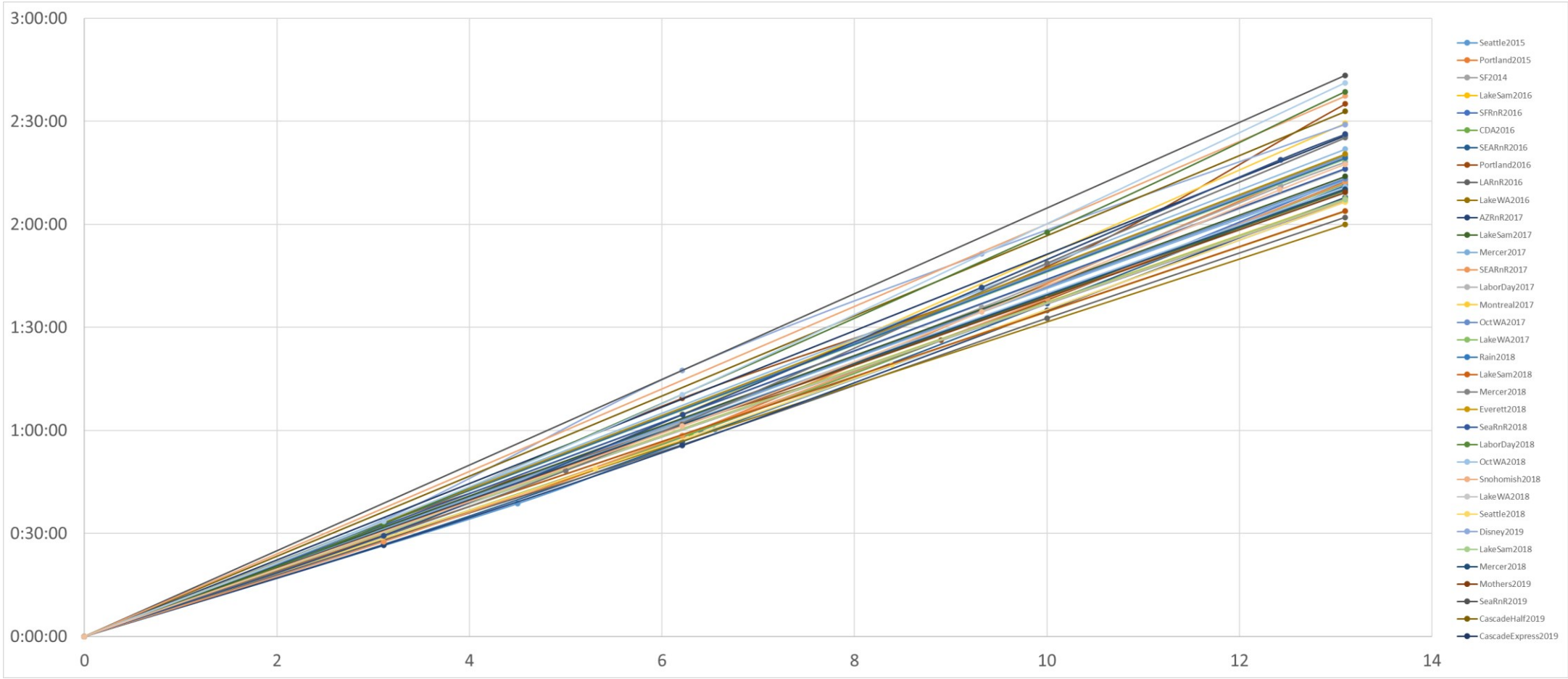


The pink line shows the cumulative growth for Fairfield Sentry, one of the largest feeder funds that was 100% invested with Madoff. So, it is a perfect window into Madoff's claimed record.

Markopolos knew immediately that one can't possibly avoid all monthly losses greater than -0.75% and at the same time handily beat out the long term returns of both GATEX and the S&P 500. Also, the near perfectly straight line with such a high positive slope was another give away for Markopolos. With higher slope (returns) comes wilder volatility (ups and downs). But, not for Madoff…

Bad Data, Analysis: Incomplete

- My half-marathon data…



IRL: Bad Data, Analysis: Incomplete

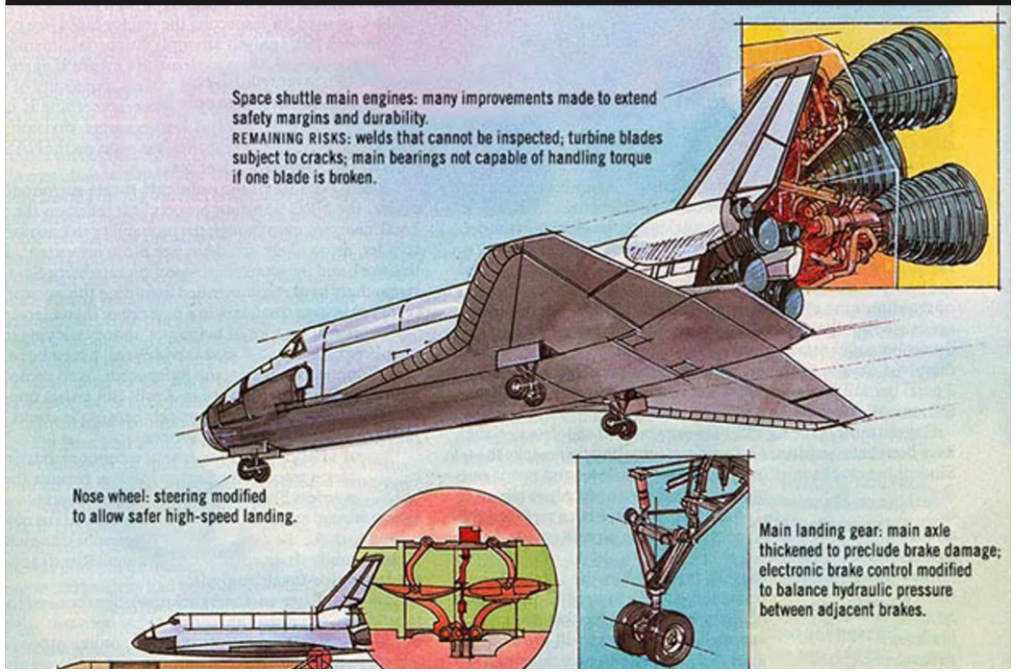
Gaussian assumptions!

FEATURE | HISTORY OF TECHNOLOGY

THE CHALLENGER DISASTER: A CASE OF SUBJECTIVE ENGINEERING

From the archives: NASA's resistance to probabilistic risk analysis contributed to the Challenger disaster

BY TRUDY E. BELL, KARL ESCH | 28 JAN 2016 | 14 MIN READ



Imperfect Information and the Housing Finance Crisis

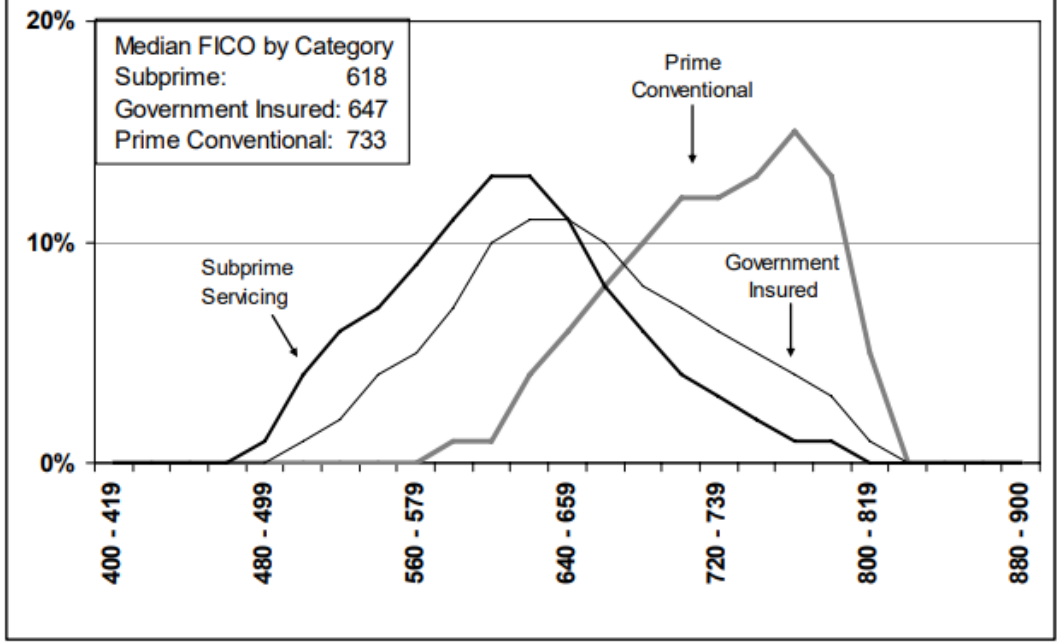
Edward Golding, Richard K. Green and Douglas A. McManus

February 2008

UCC08-6

Last revised: February 1, 2008

Figure 1: Distribution of Borrower FICO Scores in Outstanding Loans as of December 2006



IRL: Bad Data, Analysis: Irrelevant

The Enron Email Dataset

500,000+ emails from 150 employees of the Enron Corporation

Data Card | Code (257) | Discussion (6) | Suggestions (0)

About Dataset

The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.

This is the May 7, 2015 Version of dataset, as published at https://www.cs.cmu.edu/~./enron/

emails.csv (1.43 GB)

Detail | Compact | Column — 2 of 2 columns

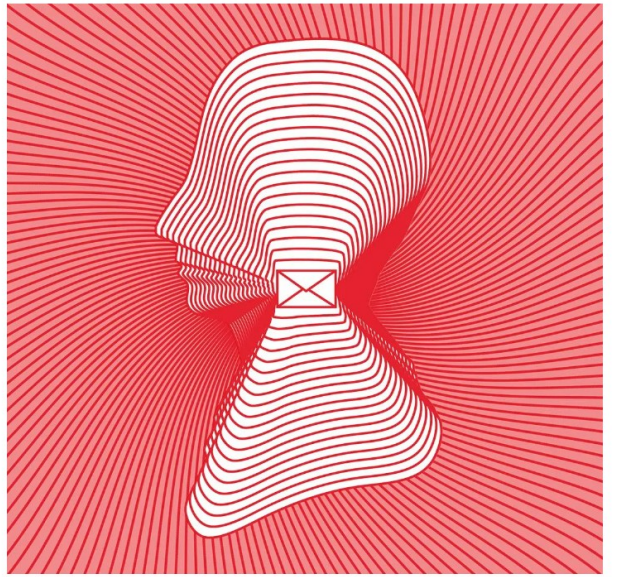
file	message
517401 unique values	517401 unique values
allen-p/_sent_mail/1.	Message-ID: <18782981.1075855378110.JavaMail.evans@thyme> Date: Mon, 14 May 2001 16:39:00 -0700 (PDT...

ANNALS OF TECHNOLOGY

WHAT THE ENRON E-MAILS SAY ABOUT US

Scholars have spent years analyzing the corporation's vast digital archive. What have they discovered?

By Nathan Heller
July 17, 2017



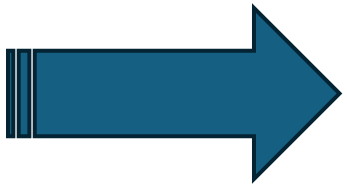
The Enron corpus provided a data dump of workplace communication styles. Illustration by Nicolas Ortega

A measure of industrial progress is the speed with which inventions grow insufferable. The elevator, once a marvel of efficiency, has become a social purgatory from which most of us cannot escape too quickly. The builders of the first commercial airplane couldn't have foreseen the crushed knees and the splattered salad dressings that their machine would visit on the world. "Hitherto it is questionable if all the mechanical inventions yet made have lightened the day's toil of any human being," John Stuart Mill wrote in the "Principles of Political Economy" (1848), and the precept holds for recent innovations, too. Think of e-mail. Or, rather, try *not* to think of e-mail, since, chances are, while you floss, steep tea, make love, or read these

Bad Data, Analysis: Incongruent… or not?

- “*Everyone needs a pair of shoes.*”
- “*All machine logs match actual events.*”
- “*Students in the course get higher test scores.*”

Outliers!



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

IRL: Bad Data, Analysis: Incongruent

Not so "twin" digital twin!

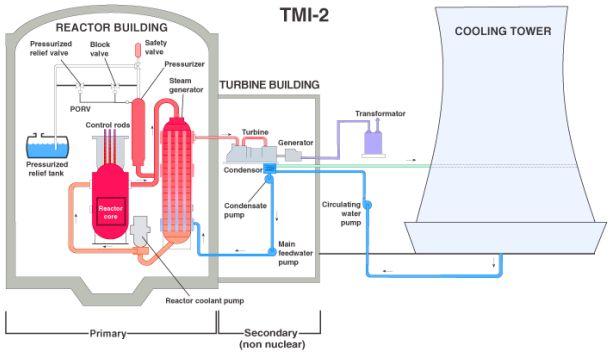
SAFETY AND SECURITY

Three Mile Island Accident

UPDATED TUESDAY, 11 OCTOBER 2022

- In 1979 at Three Mile Island nuclear power plant in USA a cooling malfunction caused part of the core to melt in the #2 reactor. The TMI-2 reactor was destroyed.
- Some radioactive gas was released a couple of days after the accident, but not enough to cause any dose above background levels to local residents.
- There were no injuries or adverse health effects from the Three Mile Island accident.

The Three Mile Island power station is near Harrisburg, Pennsylvania in the USA. It had two pressurized water reactors. TMI-1, a PWR of 880 MWe (819 MWe net) entered service in 1974, and remained one of the best-performing units in the USA until it was shut down in 2019. TMI-2 was of 959 MWe (880 MWe net) and almost brand new at the time of the accident.



The accident to unit 2 happened at 4 am on 28 March 1979 when the reactor was operating at 97% power. It involved a relatively minor malfunction in the secondary cooling circuit which caused the temperature in the primary coolant to rise. This in turn caused the reactor to shut down automatically. Shut down took about one second. At this point a relief valve failed to close, but instrumentation did not reveal the fact, and so much of the primary coolant drained away that the residual decay heat in the reactor core was not removed. The core suffered severe damage as a result.

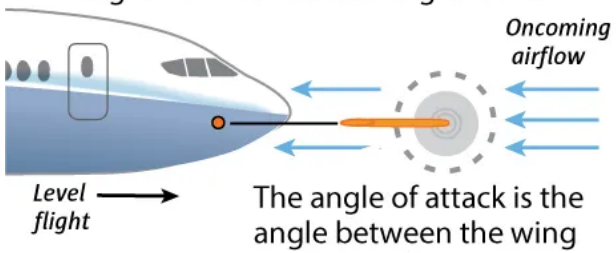
The operators were unable to diagnose or respond properly to the unplanned automatic shutdown of the reactor. Deficient control room instrumentation and inadequate emergency response training proved to be root causes of the accident.

The chain of events during the Three Mile Island accident

Within seconds of the shutdown, the pilot-operated relief valve (PORV) on the reactor cooling system opened, as it was supposed to. About 10 seconds later it should have closed. But it remained open, leaking vital reactor coolant water to the reactor coolant drain tank. The operators believed the relief valve had shut because instruments showed them that a "close" signal was sent to the valve. However, they did not have an instrument indicating the valve's actual position.

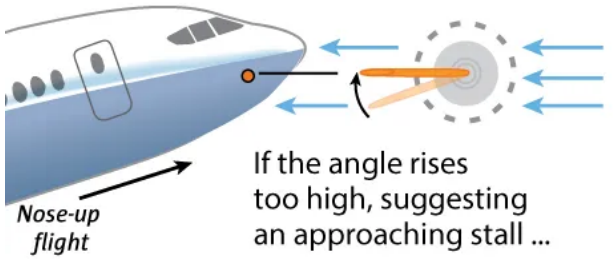
How the MCAS (Maneuvering Characteristics Augmentation System) works on the 737 MAX

1. The angle-of-attack sensor aligns itself with oncoming airflow.



The angle of attack is the angle between the wing and the airflow.

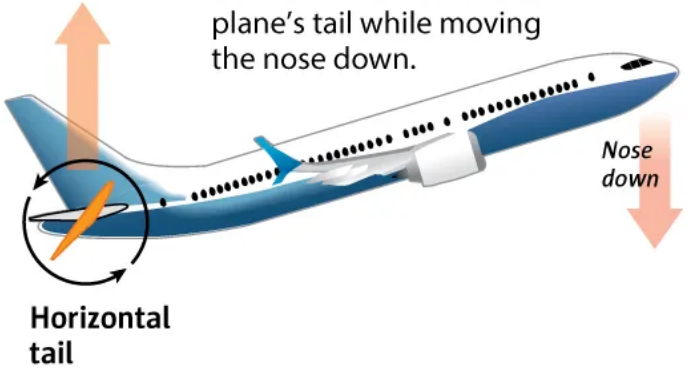
2. Data from the **sensor** is sent to the flight computer.



If the angle rises too high, suggesting an approaching stall ...

... the **MCAS** activates.

3. MCAS automatically swivels the **horizontal tail** to lift the plane's tail while moving the nose down.



Sources: Boeing, FAA, Indonesia National Transportation Safety Committee, Leeham.net, and The Air Current

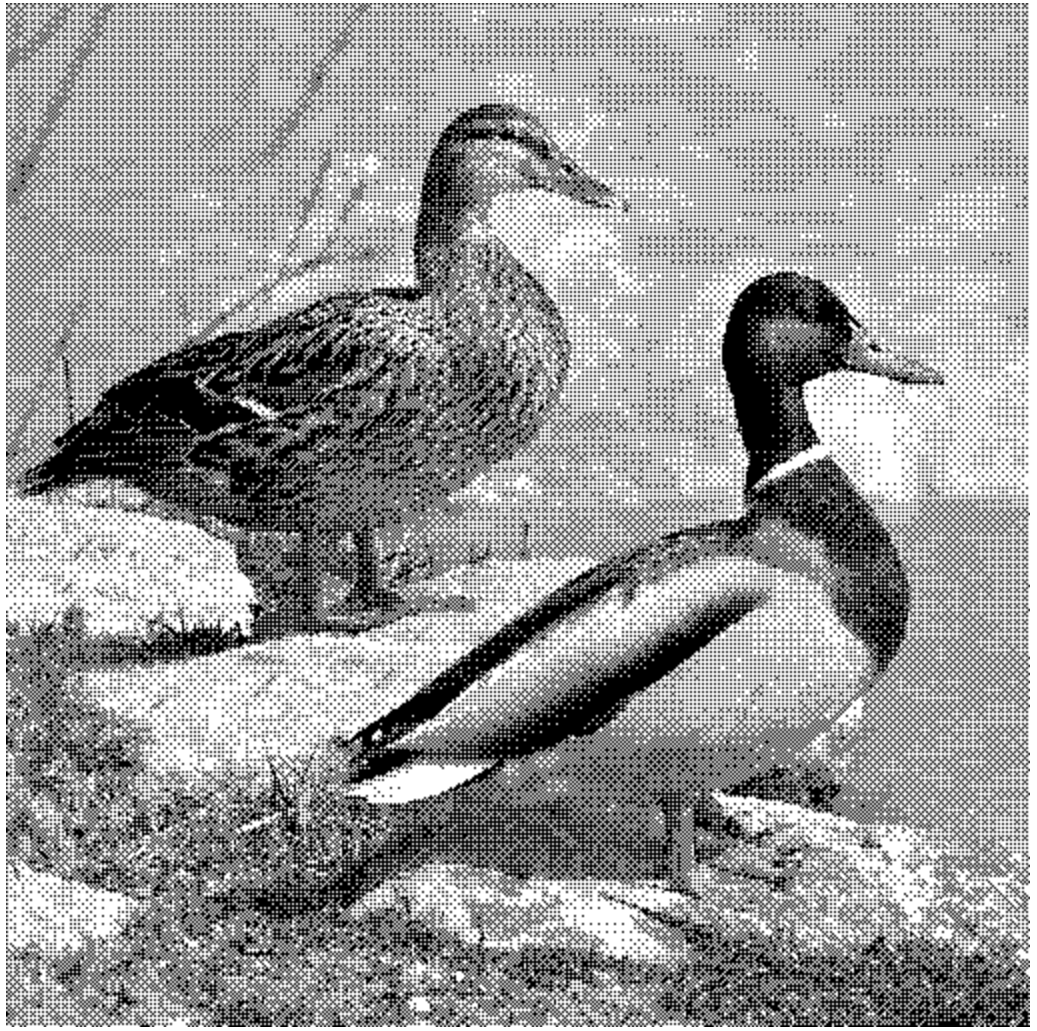
Reporting by DOMINIC GATES,
Graphic by MARK NOWLIN / THE SEATTLE TIMES

Bad Data, Analysis: Transform

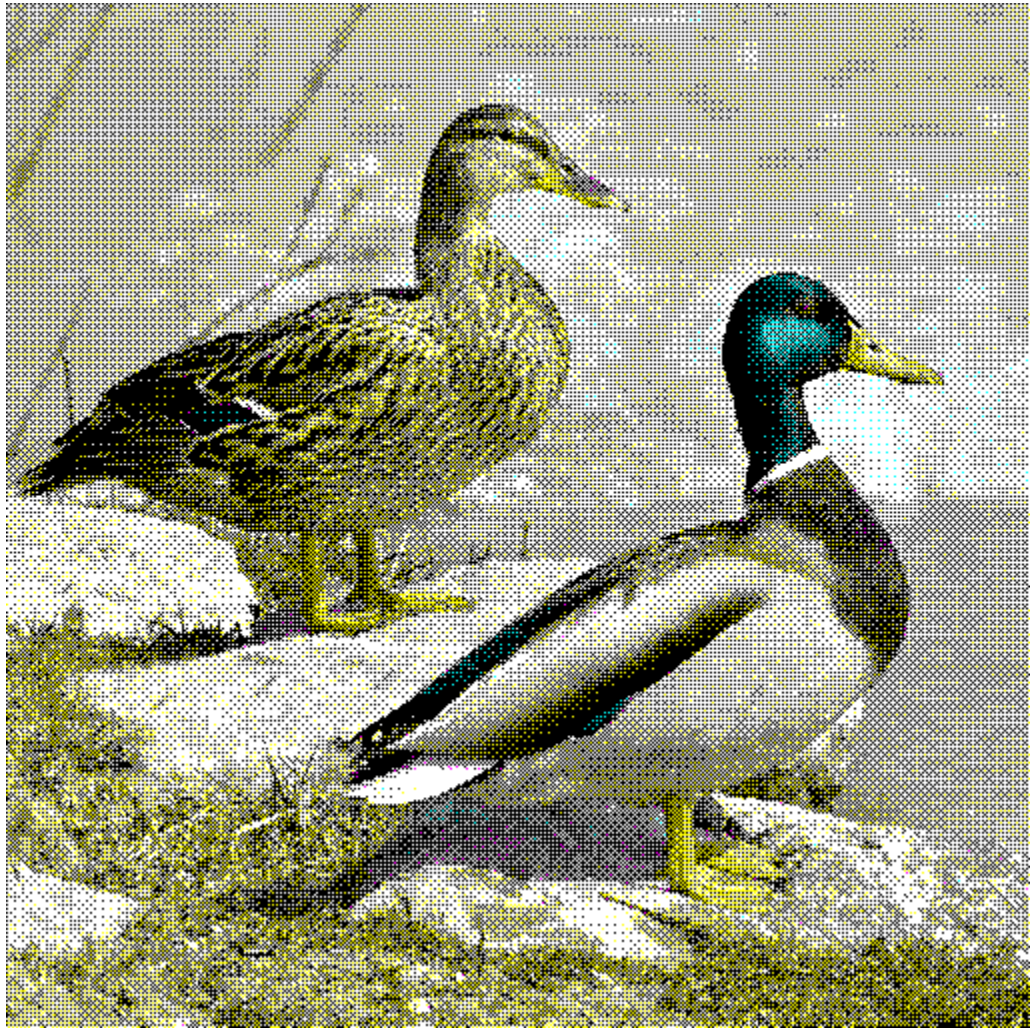
- Quantization
- Encoding/embedding

Bad Data, Analysis: “Original”

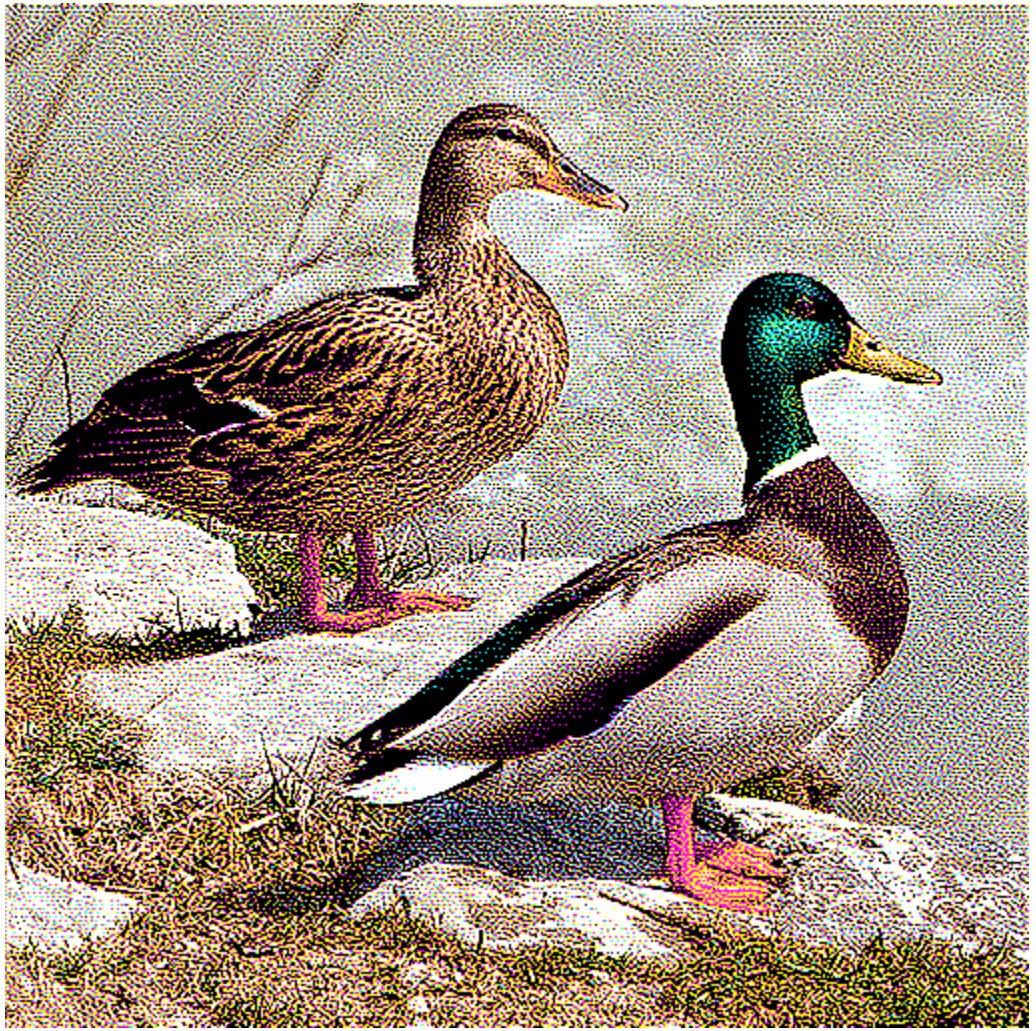
Bad Data, Analysis: Dithering (B&W + Bayer)



Bad Data, Analysis: Dithering (CMYB + Bayer)

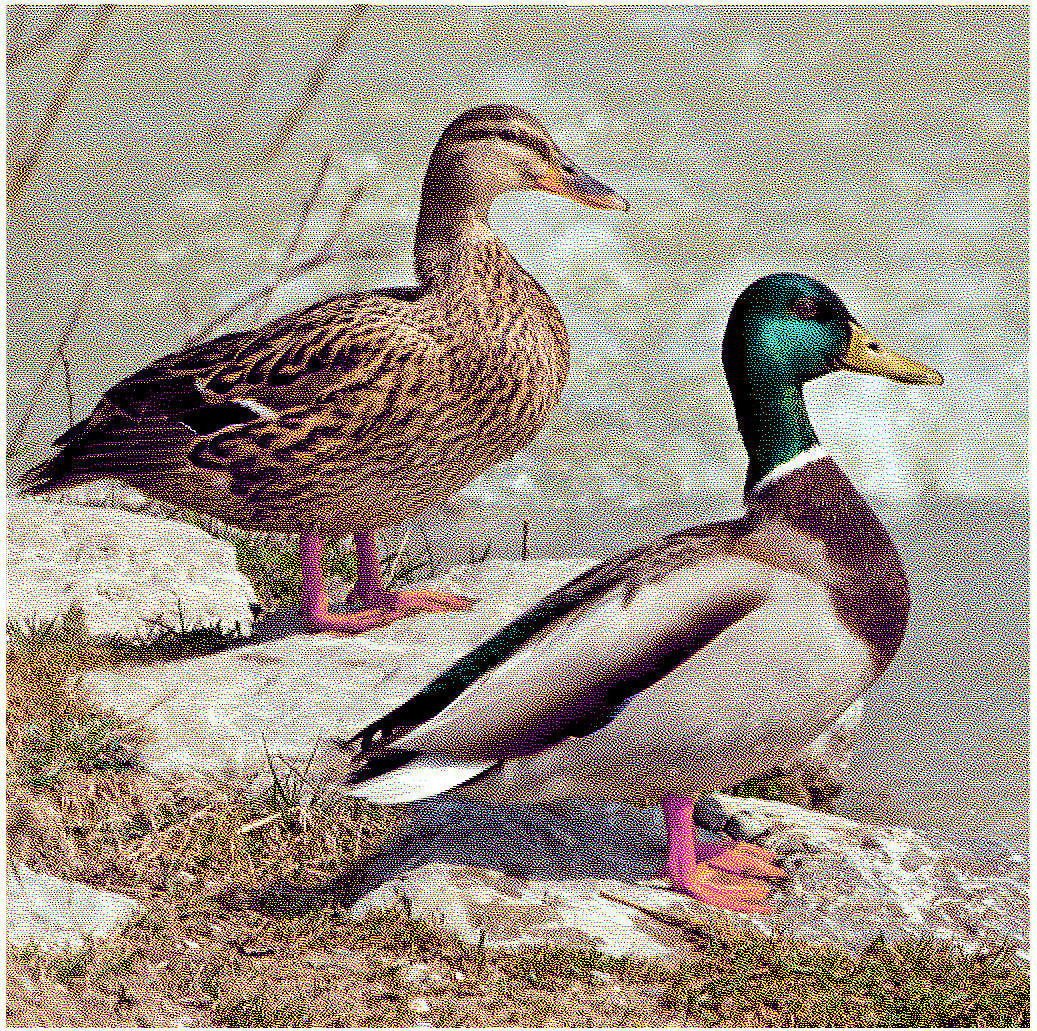


Bad Data, Analysis: Dithering (CMYB, Err. Diffusion)



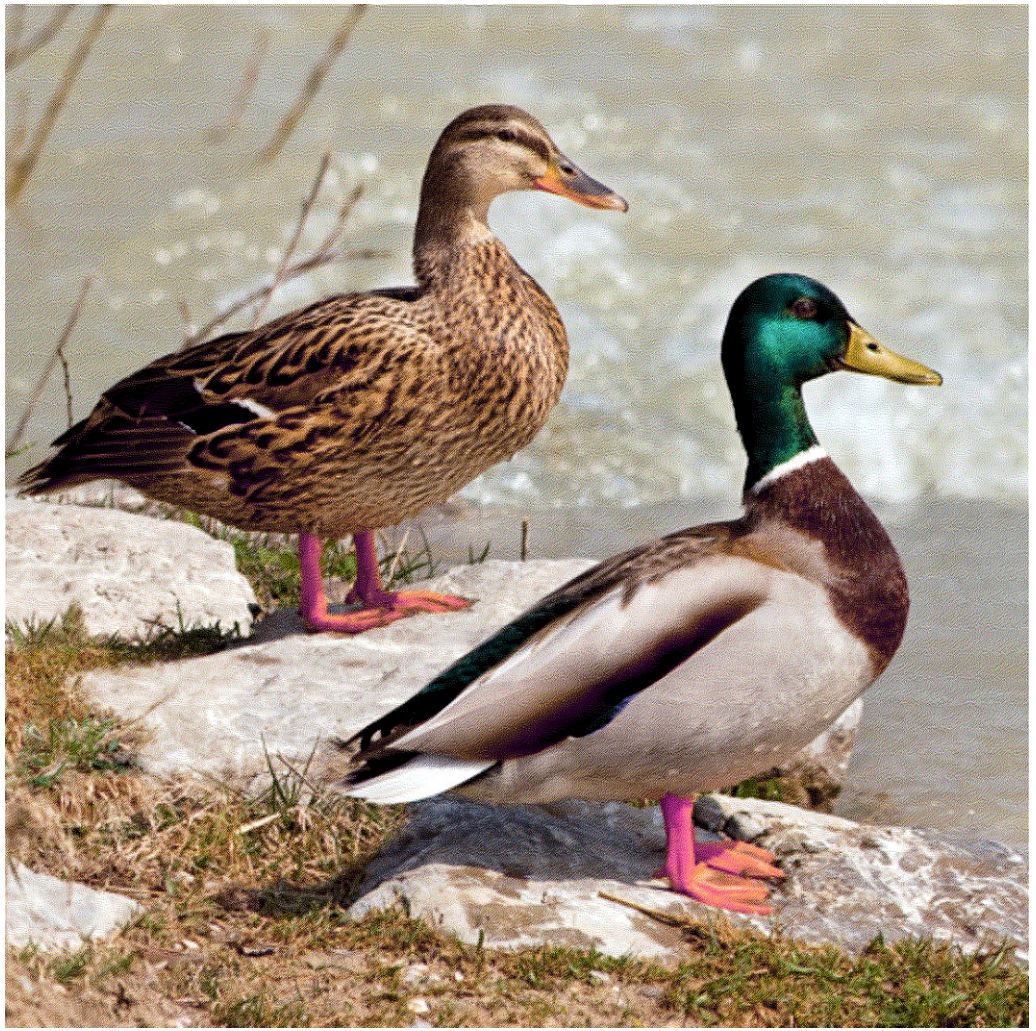
512px
Jarvis-Judice-Ninke
Serpentine Order

Bad Data, Analysis: Dithering (CMYB, Err. Diffusion)



1,024px
Jarvis-Judice-Ninke
Serpentine Order

Bad Data, Analysis: Dithering (CMYB, Err. Diffusion)

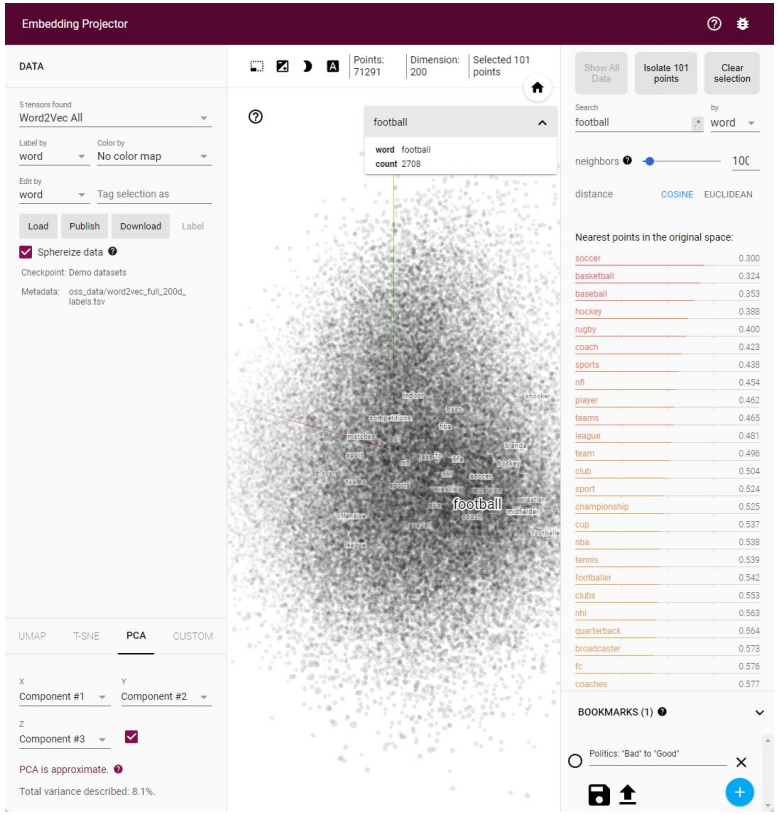


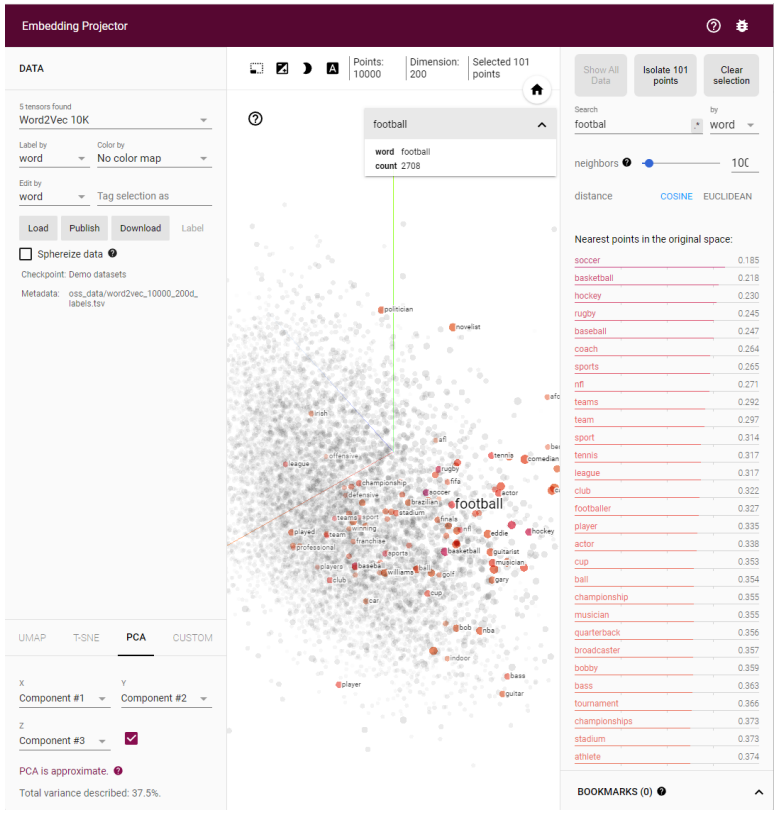
2,048px
Jarvis-Judice-Ninke
Serpentine Order

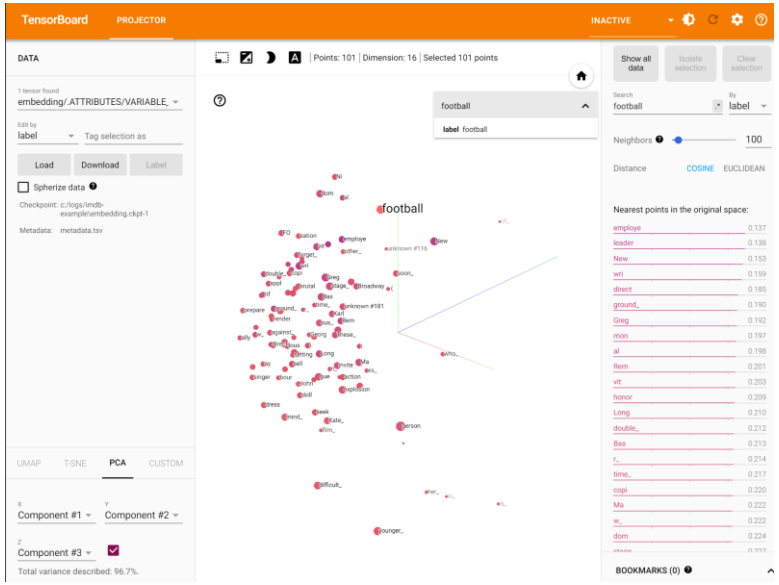
Bad Data, Analysis: Text Embedding

- Conflation and disambiguation
  - "*Tom Brady was a greatest football player*"
  - "*Pelé was the greatest football player*"

Bad Data, Analysis: Word, Sentences, Vectors







"Someone who spends enough time playing with tools named Python and Anaconda will eventually be bitten by versioning issues!"

"Premature optimization..."

IRL: Bad Data, Analysis: Transform

When NASA Lost a Spacecraft Due to a Metric Math Mistake

WRITTEN BY Ajay Harish | UPDATED ON December 8th, 2023 | APPROX READING TIME 11 Minutes

Blog › Aerospace & Defense › When NASA Lost a Spacecraft Due to a Metric Math Mistake

In September of 1999, after almost 10 months of travel to Mars, the Mars Climate Orbiter burned and broke into pieces. On a day when NASA engineers were expecting to celebrate, the ground reality turned out to be completely different, all because someone failed to use the right units, i.e., the metric units! The Scientific American Space Lab made a brief but interesting video on this very topic.

NASA'S LOST SPACECRAFT

The Metric System and NASA's Mars Climate Orbiter

The Mars Climate Orbiter, built at a cost of $125 million, was a 638-kilogram robotic space probe launched by NASA on December 11, 1998, to study the Martian climate, Martian atmosphere, and surface changes. In addition, its function was to act as the communications relay in the Mars Surveyor '98 program for the Mars Polar Lander. The navigation team at the Jet Propulsion Laboratory (JPL) used the metric system of millimeters and meters in its calculations, while Lockheed Martin Astronautics in Denver, Colorado, which designed and built the spacecraft, provided crucial acceleration data in the English system of inches, feet, and pounds. JPL engineers did not take into consideration that the units had been converted, i.e., the acceleration readings measured in English units of pound-seconds^2 for a metric measure of force called newton-seconds^2. In a sense, the spacecraft was lost in translation.

The $6.2B Excel error

This is how the error is described in the report (emphasis mine):

> Following that decision, further errors were discovered in the Basel II.5 model, including, most significantly, an operational error in the calculation of the relative changes in hazard rates and correlation estimates. Specifically, ***after subtracting the old rate from the new rate, the spreadsheet divided by their sum instead of their average***, as the modeler had intended.

Note: I don't have domain expertise in VaR models, synthetic credit derivatives, or trading in general. The following example is my over-simplification of the error based on what's written in the report.

The report talks about **hazard rates** (for what I assume relate to the default of corporate loans in this case) and how the changes in the hazard rates were improperly calculated. Here's a simple table from the **Google Sheet** showing fictitious dates, hazard rates, and the change in rates:

	A	B	C
1	**Day**	**Hazard Rate**	**Change in %**
2	8/2	8.0%	
3	8/3	10.0%	2.0%
4	8/4	13.0%	3.0%
5	8/5	18.0%	5.0%
6	8/6	13.0%	-5.0%
7	8/7	11.0%	-2.0%

Now here's what happens when you apply a SUM vs. an AVERAGE to the "Change in %" column:

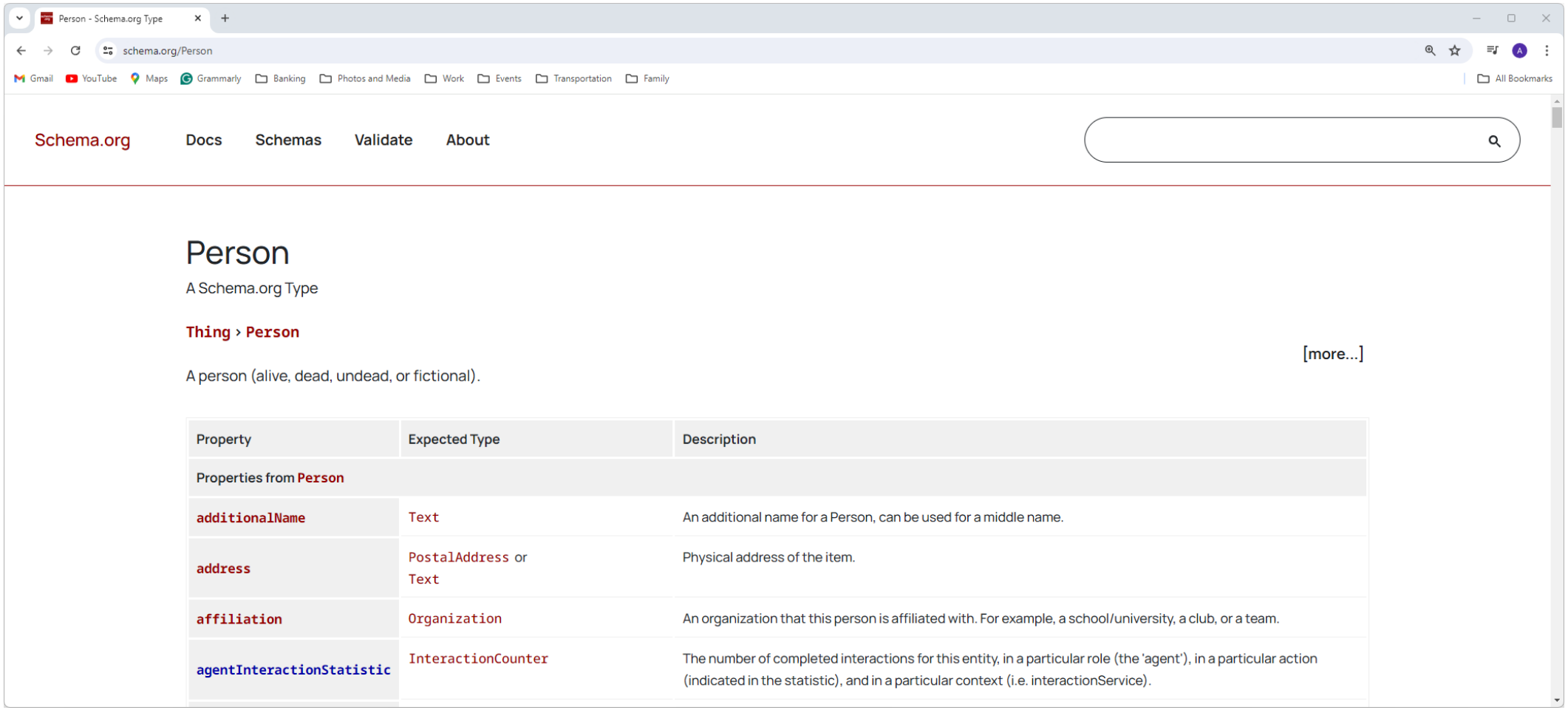
Day	Hazard Rate	Change in %
8/2	8.0%	
8/3	10.0%	2.0%
8/4	13.0%	3.0%
8/5	18.0%	5.0%
8/6	13.0%	-5.0%
8/7	11.0%	-2.0%
	Sum of changes	**3.0%**
	Average of changes	**0.6%**

This is hitting the border of my knowledge of growth rates and time periods, but the *sum of changes* will always be 5X the *average of changes* given there are 5 values we are summing/averaging.

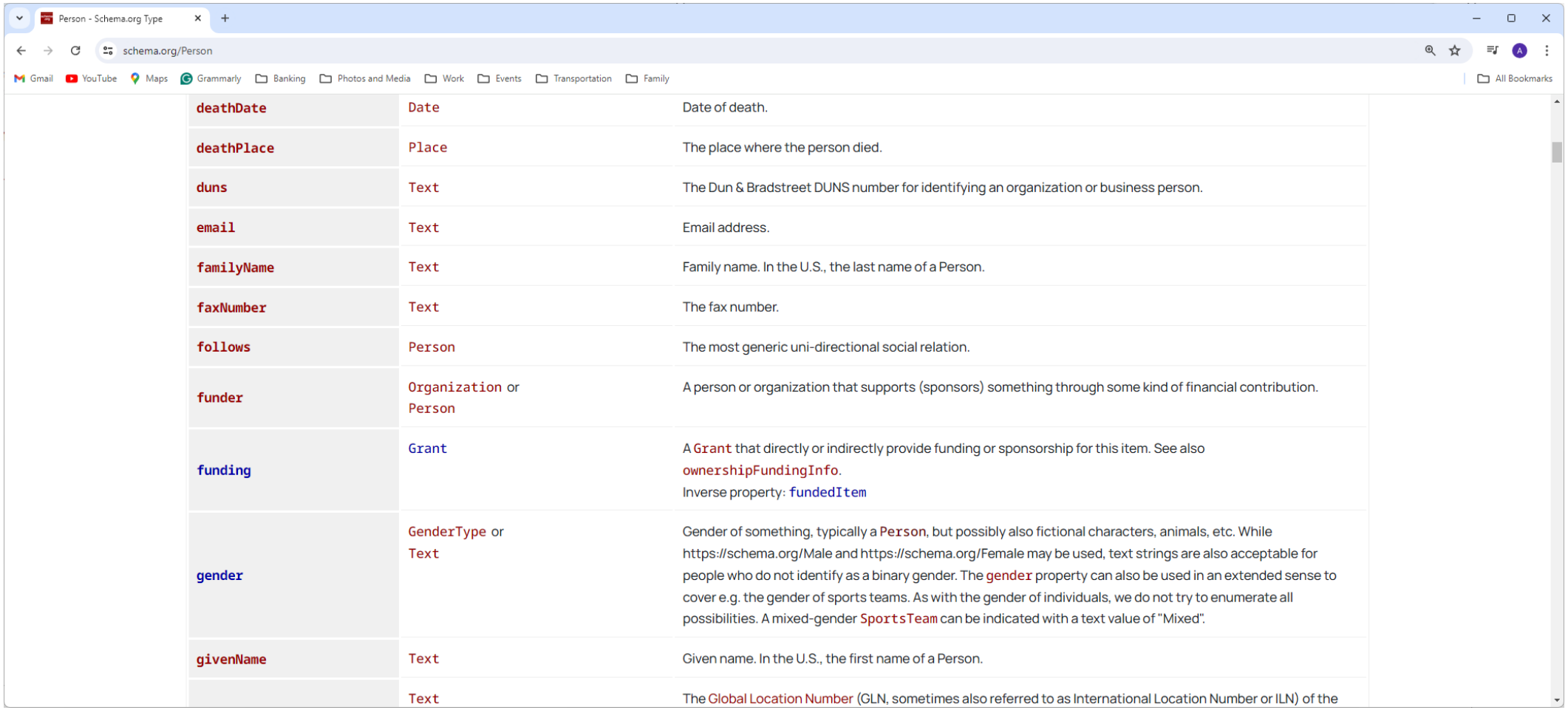
Bad Data, Analysis: Load

- Persistence
- Visualization

Bad Data, Analysis: Data Model



Bad Data, Analysis: Things Change…



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

IRL: Bad Data, Analysis: Persistence

The Trading Glitch, which cost Knight Capital $440 Million

Jee-Yu Yang · Follow
Published in CodeX · 3 min read · Jun 21, 2022

30

Software Testing Lessons learned from Knight Capital Fiasco 2012 — Part 1

knight capital @ nyse (foto by REUTER, ntv)

August 1st

08:01 a.m. EST — Knight Capital's personnel receive 97 emails that describe a disable of a specific trading algorithm from server NO.8. These emails referenced SMARS as an issue.

> ***SMARS*** *stands for Smart Market Access Routing System, and it was able to execute thousands of orders per second, comparing prices between dozens of different trading events within fractions of a second.*

These internal messages were neither designed for high-priority alerts nor the staff generally reviewed them in real-time. Therefore Knight Capital missed the opportunity to identify and fix this issue before the market opening.

FINTECH FUTURES NORTH AMERICAN EDITION

PAYTECH AWARDSUSA

News

Goldman Sachs trading error is "a warning to all"

Written by FinTech Futures 21st August 2013

Financial industry analysts have warned that investors should be careful about how they approach automated trading, following news that a trading error at Goldman Sachs cost the firm $100 million on Tuesday.

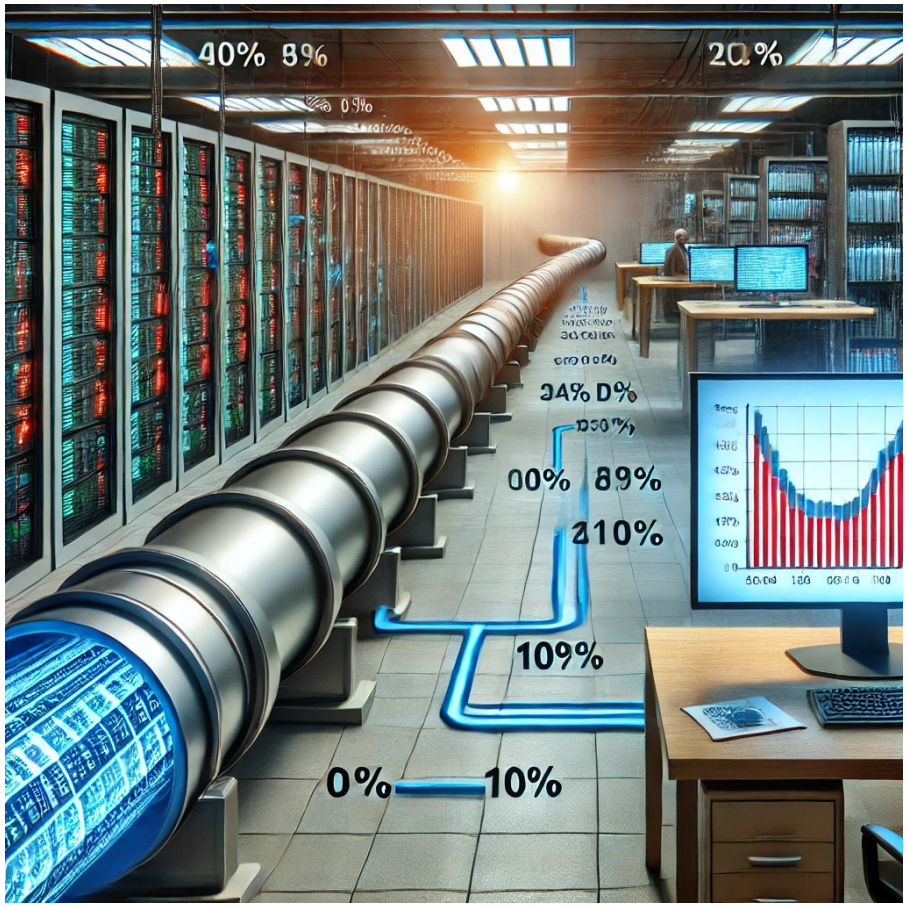
Goldman Sachs has lost an estimated $100 million due to a trading error

The glitch caused the firm to post a number of erroneous options trades that disrupted trading across US exchanges during the first 15 minutes of trading. The affected venues included CBOE, Nasdaq OMX and NYSE Euronext. Options on shares with listing symbols beginning with the letters H all the way through to L were affected.

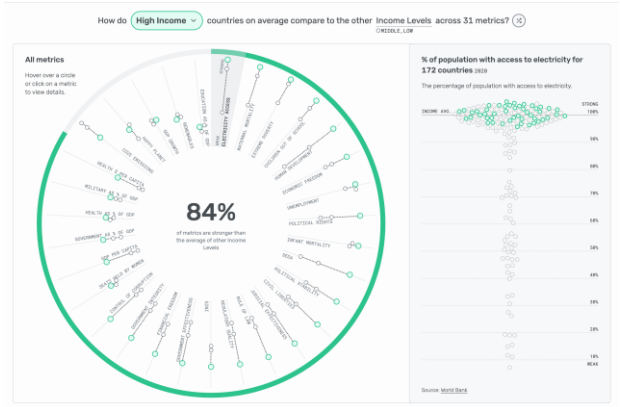
According to reports, the problem was caused by a computer error in which automated trading systems accidentally sent indications of interest as real orders to be filled at the exchanges. Goldman Sachs said in a statement that it faced no material loss or risk from the incident, but declined to comment further.

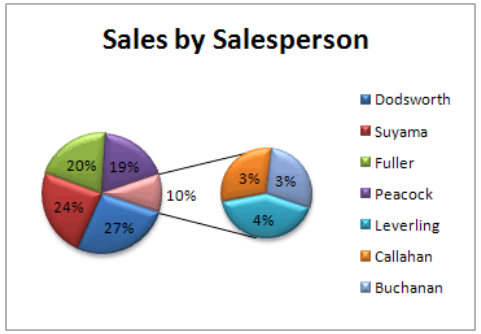
Bad Data, Analysis: Visualization

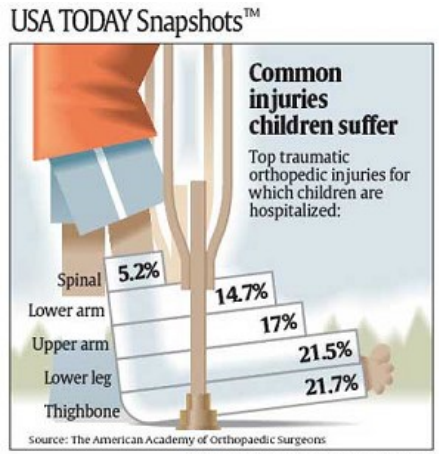
- It all worked. So far…



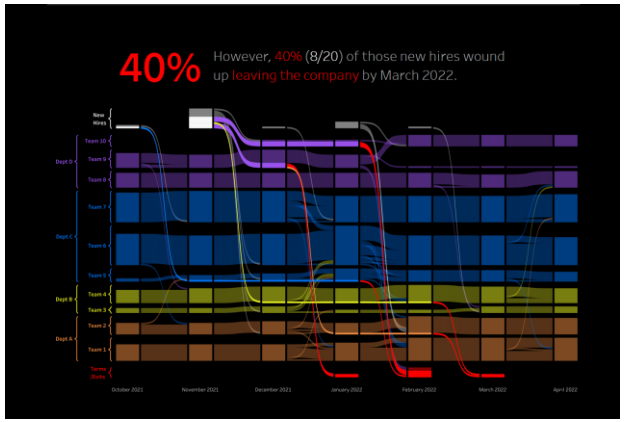
IRL: Bad Data, Analysis: Visualization

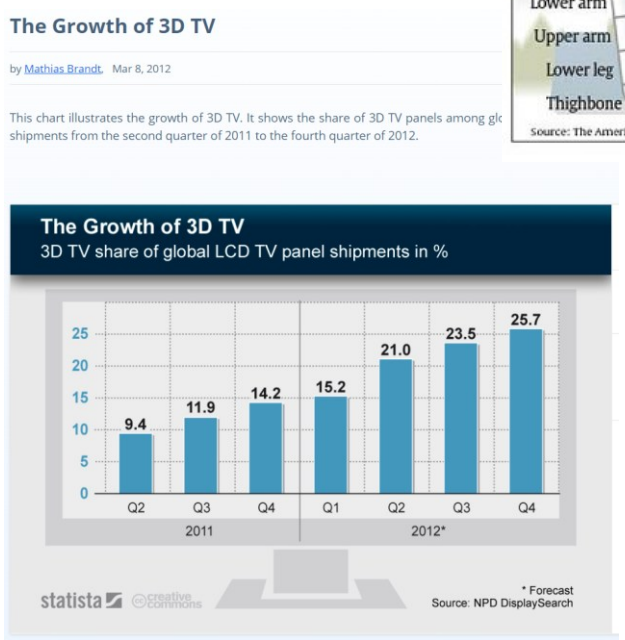






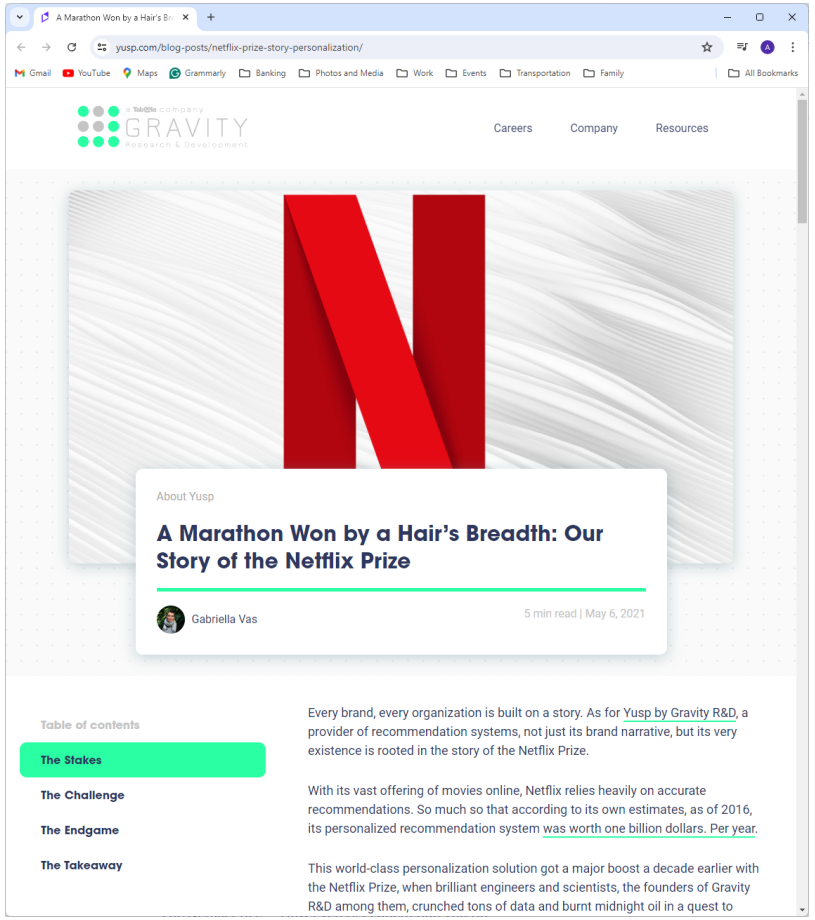
Hans Rosling's 200 Countries, 200 Years, 4 Minutes - The Joy of Stats - BBC Four





Bad Data, Analysis: Mitigations

- **Extract:** *Start, make mistakes, iterate your model. Always seek random samples.*
- **Transform:** *Keep "original data" and provenance (avoid premature optimization).*
- **Load:** "*Data dictionaries" (with History!)*



Thank you for DECIDING to be here!

“Bad analysis happen but no miracle analysis can save you from bad data. This presentation will share stories of both bad data and bad analysis, and a few principles to avoid those.”

A New Hope

Bad Data Analysis

- Cherry Picking
- Anecdote Induction
- Beautiful Stories

Bad Data, Analysis

- Extract (Source)
- Transform
- Load (Store, Visualize)