Data Science 110

Eldad Haber

Who are we

- Instructor:
    - Eldad
- TA’s:
    - Niloufar
    - Shahriar

We work in the interphase of data science, machine learning and applications.

Learning Goals

- General goals:
  - What is data
  - How do you present the data
  - How do you make sense of the data and have common sense
- Technical goals
  - A bit of python programming
  - A bit of linear algebra
  - A bit of calculus
  - Some science
  - Some common sense
- My goal
  - To make you passionate about data and models
  - To show you how such models can change the world
  - To show you how to make sense of data

Approximate Course plan

- Week 1 – Introduction
- Week 2 – A crash course in python
- Week 3 – Data and plotting it
- Week 4-13 – Every week we will analyze a different data set. Some of the data are:
    - World temperature data
    - Tied data
    - CO2 vs Temperature
    - Mineral occurrence
    - Vaccination and autism
- Week 14 – Summary

* If there is a data set that you would like to analyze please come forward

Some data to start

Grading

- Class participation 5%
- Homework (once every 3-4 weeks) 20%
- Projects (groups of 3-4) 20%
- Midterm 20%
- Final 15%

- This is an introductory course. You should be doing well

Some data to start

Grading

What was wrong with the last slide?

- Class participation 5%
- Homework (once every 3-4 weeks) 20%
- Projects (groups of 3-4) 20%
- Midterm 20%
- Final 15% Change to?

- This is an introductory course. You should be doing well

The data did not make sense

What is data?

Data is information

- Quantitative

- Qualitative

How to get it

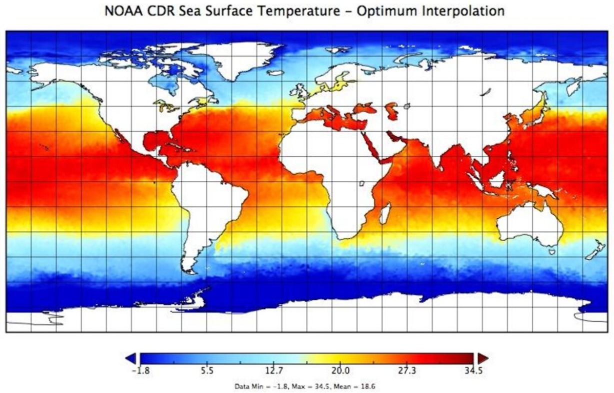
Data Sources

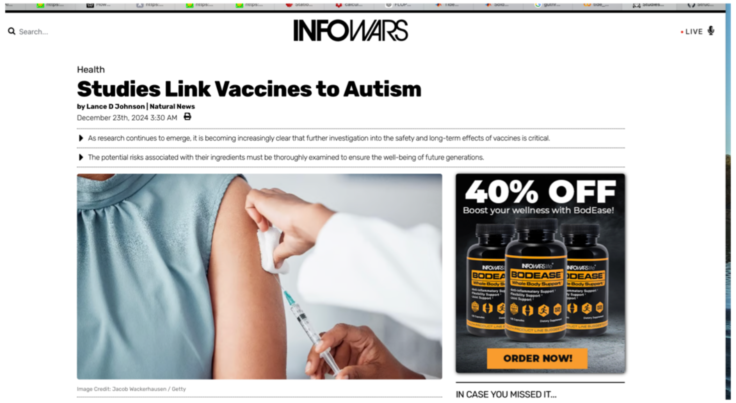
- Government, not for profit
- Media
- Private

Data Reliability

How do we know that the data is reliable?

- Clear vs fake


NOAA CDR Sea Surface Temperature – Optimum Interpolation
-1.8
5.5
12.7
20.0
27.3
34.5
Data Min = -1.8, Max = 34.5, Mean = 18.6



INFOWARS

Search...

• LIVE

Health

Studies Link Vaccines to Autism

by Lance D Johnson | Natural News

December 23th, 2024 3:30 AM

- As research continues to emerge, it is becoming increasingly clear that further investigation into the safety and long-term effects of vaccines is critical.
- The potential risks associated with their ingredients must be thoroughly examined to ensure the well-being of future generations.

Image Credit: Jacob Wackerhausen / Getty

40% OFF

Boost your wellness with BodEase!

ORDER NOW!

IN CASE YOU MISSED IT...

- Context
  - UBC tuition ~12K
  - Daycare ~24K
- Fake that looks like real
  - Opioids are great to control pain
  - Immigrants taking beds in Canadian hospitals

Data Reliability

One of the main goals of this course is to make you think about data, its sources and its reliability

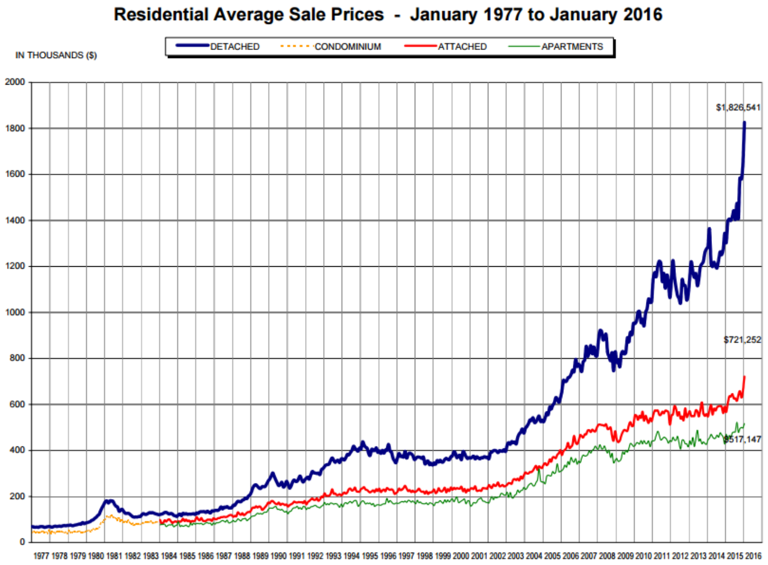
Use some technical tools to analyze the data

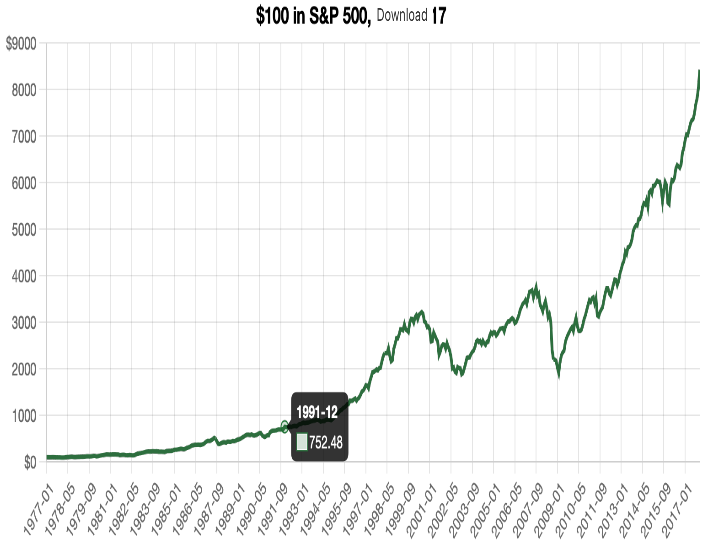
What do we do with data

- Share it
- Make decisions that impact our and other lives!

Make your decisions informed

Should I buy a house?





This chart shows the rate of gains and loss by month

Numerical Tabular Data

Most of the data we will be dealing with is tabular

- Temperature/CO2 vs time
- Tide data vs time
- Geospatial data
- Images and movies (pixels)
- …

- Advantage – Can use computers to analyze the data

Non numerical data

- Big effort to somehow make non-numerical data into numerical one and tabulate it

- Your mood
- Description of what is in the image (cat, dog …)
- Words in general (chatGPT)
- Unstructured data (stuff I tell you)

Data cleaning

- A lot of the data around us can be considered as noisy
- How do we know what is noise?
- One person noise is the other person's signal

Data cleaning

The cosmic microwave background (CMB) radiation was initially considered noise by its discoverers.

In 1964, Arno Penzias and Robert Wilson at Bell Telephone Laboratories were working with a sensitive microwave antenna when they detected an unexpected uniform signal that seemed to come from all directions

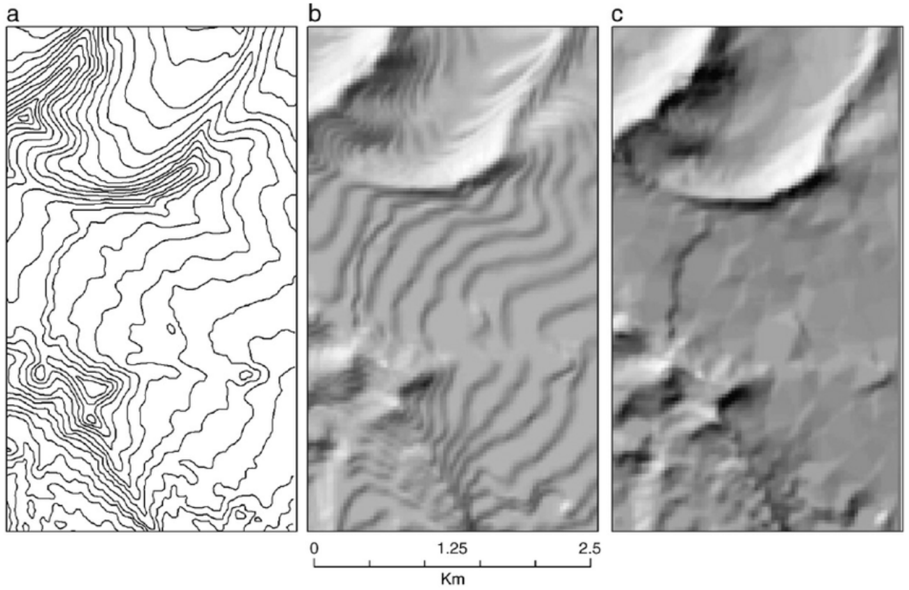
They initially thought this persistent microwave static was interference or noise, and spent months trying to identify and eliminate its source

They even checked for pigeon droppings inside the antenna as a potential cause of the mysterious signal

It wasn't until they consulted with physicist Robert Dicke that they realized they had inadvertently discovered the CMB radiation, which was predicted by the Big Bang theory

Tabulating the data

- Easier to work with tabular data
- Data interpolation
- Types of interpolation and interpolation effects



Working with data

- Until recently requires highly technical skills!

- Today, with a bit of effort and the help of LLMs can be done easily
- LLMs – Large Language Models
  - ChatGPT
  - Proplexity
  - Geminy
  - Claud
  - …

- We will learn how to use LLMs to code and play with data
- Bring your laptop and connect to the internet

Class activity

Divide into groups of 5

Find an example for data that is fake and look real

Explain how do you know that the data is fake

Rate the following data sources

- https://birdsarentreal.com
- Https://www.infowars.com/posts/a-young-child-died-during-moderna-covid-vaccine-clinical-trial-did-fda-know
- https://www.stonybrook.edu/commcms/geosciences/about/_LIG-Past-Conference-abstract-pdfs/2022-Abstracts/Tentomas.pdf