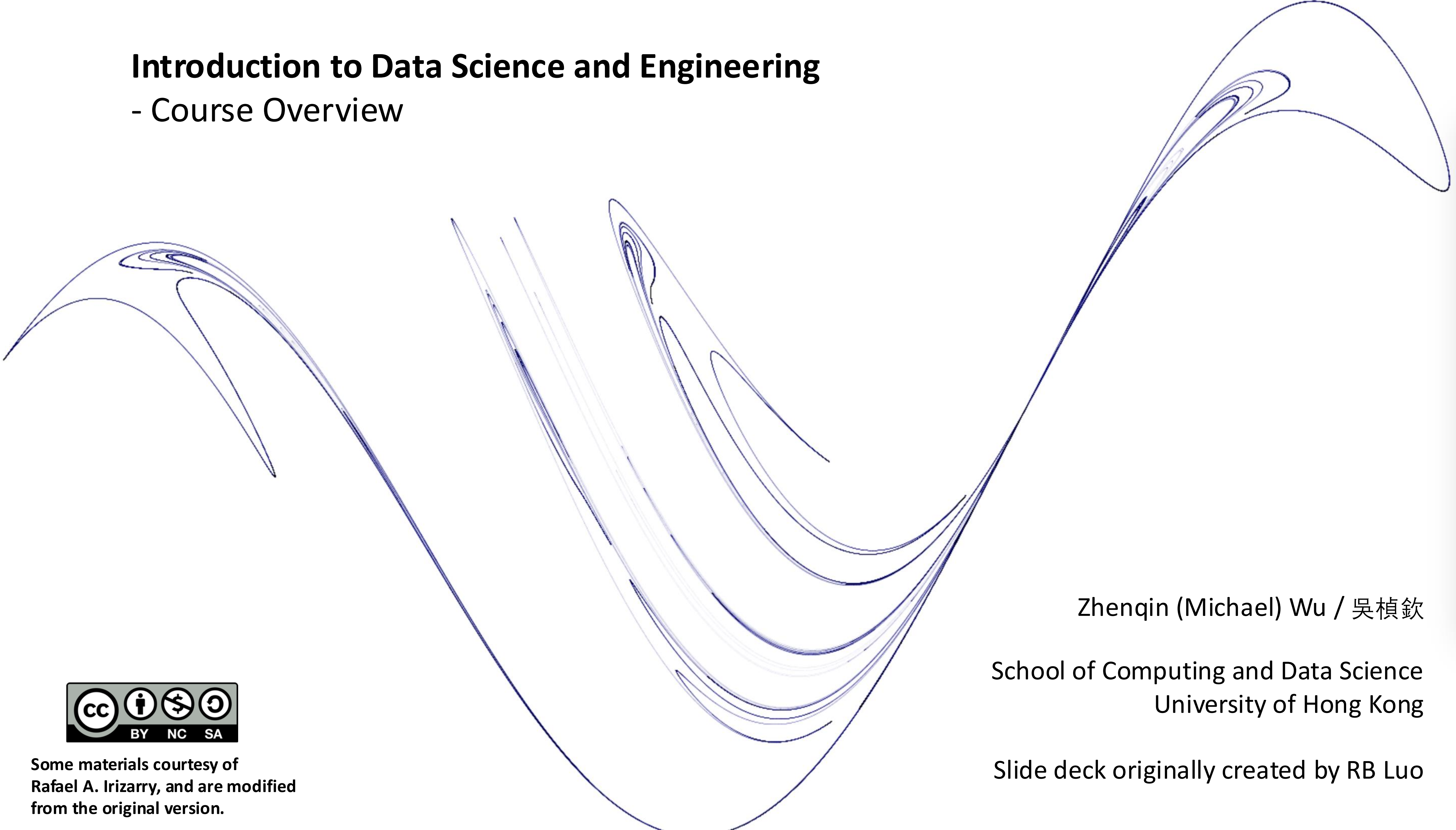
Introduction to Data Science and Engineering

- Course Overview



Zhenqin (Michael) Wu / 吴楨欽

School of Computing and Data Science
University of Hong Kong

Slide deck originally created by RB Luo

BY NC SA

**Some materials courtesy of
Rafael A. Irizarry, and are modified
from the original version.**

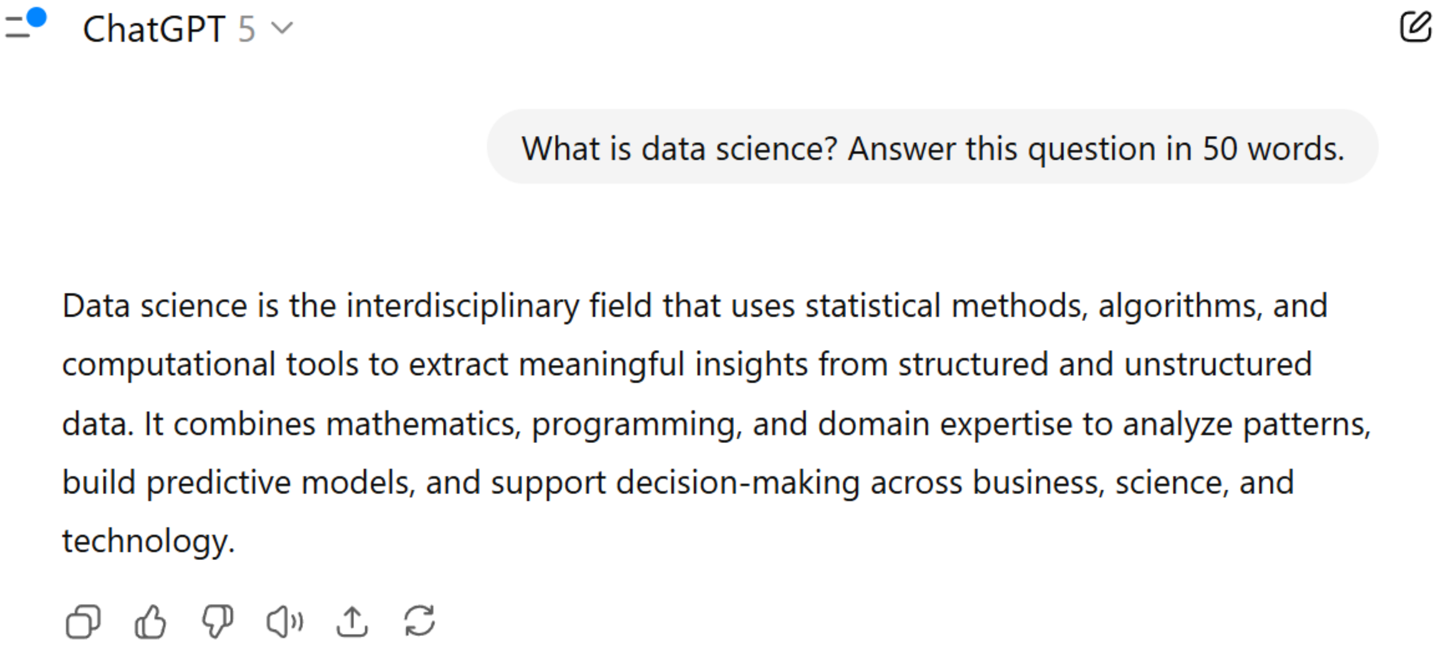
In this lecture

- Intro: four questions
- Course details

In this lecture

- **Intro: four questions**
- Course details

Q1: What’s data science?



What is data science? Answer this question in 50 words.

Data science is the interdisciplinary field that uses statistical methods, algorithms, and computational tools to extract meaningful insights from structured and unstructured data. It combines mathematics, programming, and domain expertise to analyze patterns, build predictive models, and support decision-making across business, science, and technology.

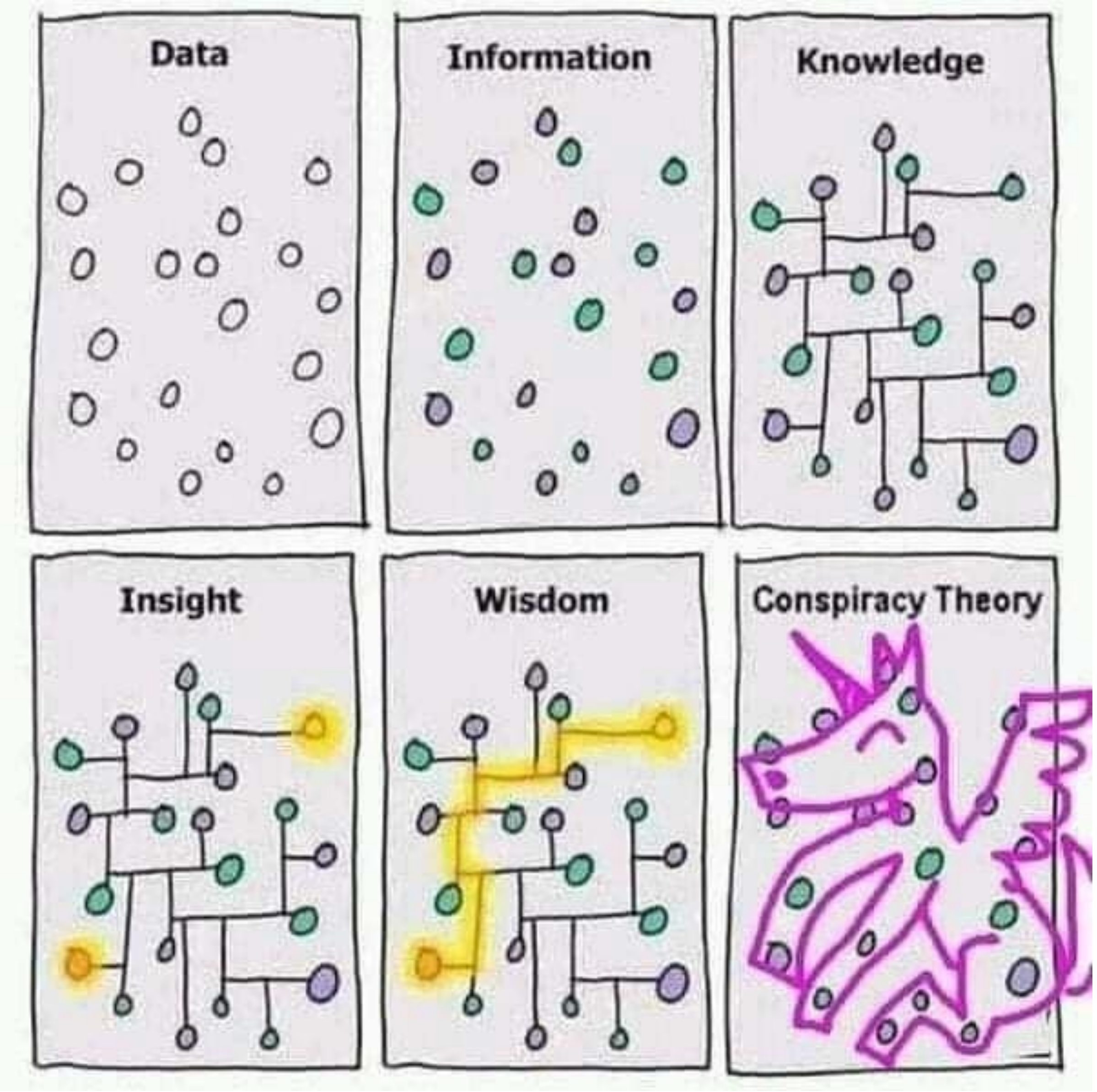
Key points here:

- Extract **insights from data**
- **Interdisciplinary**

Q1: What’s data science?

Key points here:

- Extract **insights from data**
  - What are insights?
  - Data is everywhere, how to extract insights?
  - How to avoid over-interpretations/conspiracies?
- **Interdisciplinary**



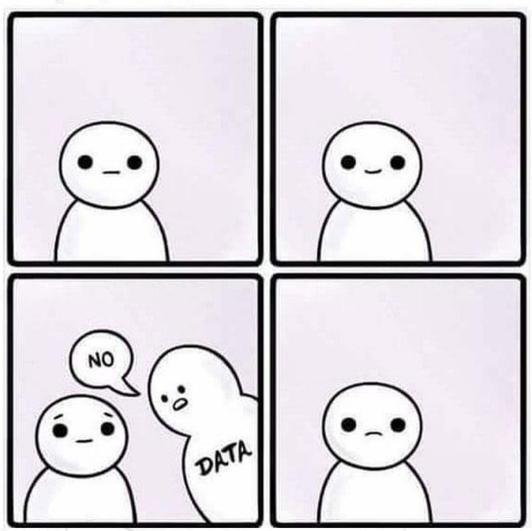
Q2: Is data science a real science?

Criticism: *"Big data analysis is mainly: garbage in, garbage out…*

Criticism: "*Data science can find anything in a large amount of data. By subsetting the data and building new features, they can prove anything. As Ronald H. Coase said: "If you torture the data long enough, it will confess.*“

Criticism: "*Data scientists can only build observational studies. They might spot a few correlations, but can't say anything about the underlying causes.*“

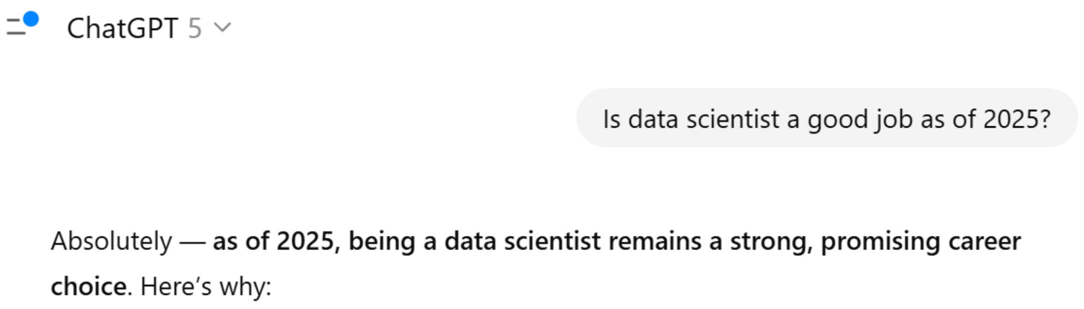
…


The (real) scientific method.

What could make data science **more scientific**?

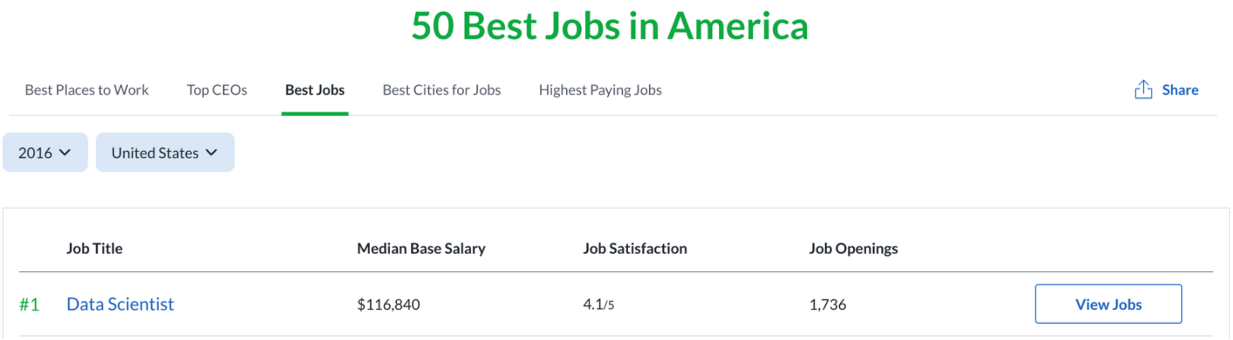
- Uphold **scientific values**: emphasize transparency and reproducibility
- Correct **methodology**: never presume your result, design your experiments/tests rigorously, acknowledge limitations

Q3: Is data scientist a good job?



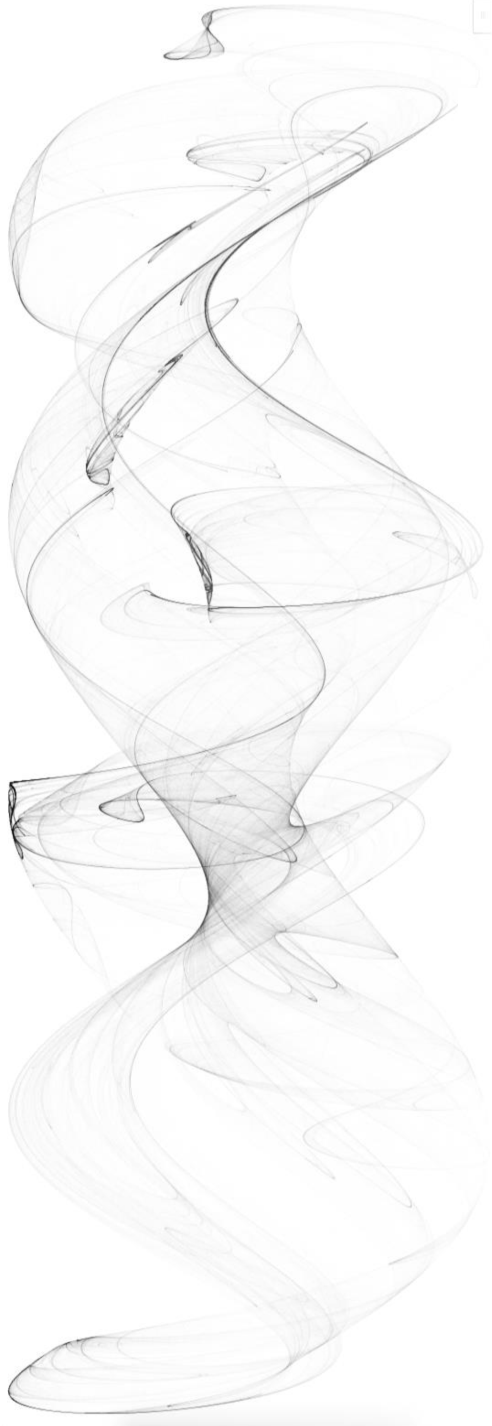
Is data scientist a good job as of 2025?

Absolutely — **as of 2025, being a data scientist remains a strong, promising career choice**. Here's why:

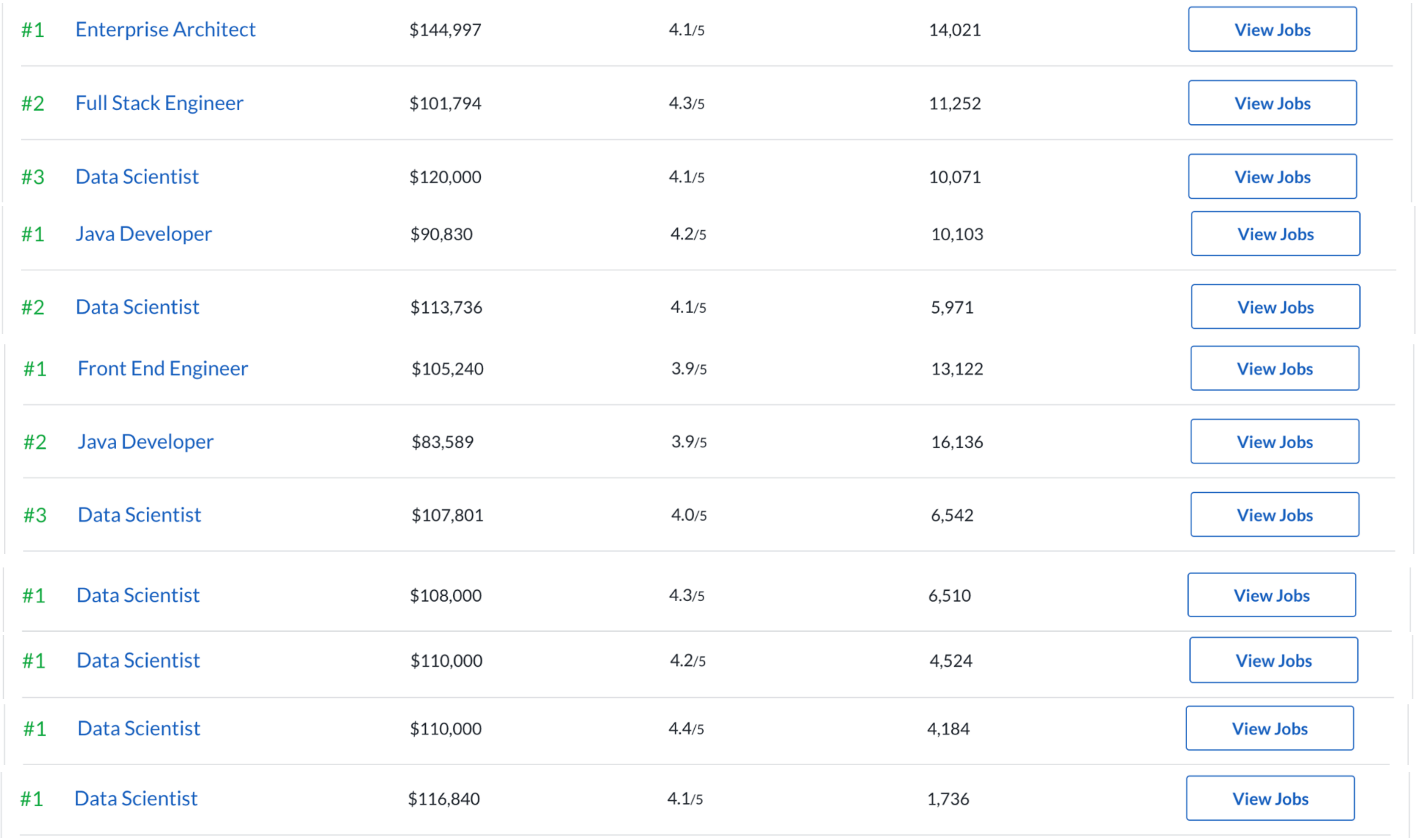


50 Best Jobs in America

	Job Title	Median Base Salary	Job Satisfaction	Job Openings	
#1	Data Scientist	$116,840	4.1/5	1,736	View Jobs



50 Best Jobs in America



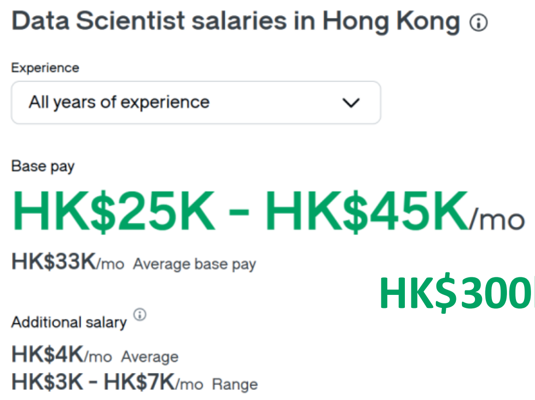
2022	#1	Enterprise Architect	$144,997	4.1/5	14,021	View Jobs
	#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
	#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs
2021	#1	Java Developer	$90,830	4.2/5	10,103	View Jobs
	#2	Data Scientist	$113,736	4.1/5	5,971	View Jobs
2020	#1	Front End Engineer	$105,240	3.9/5	13,122	View Jobs
	#2	Java Developer	$83,589	3.9/5	16,136	View Jobs
	#3	Data Scientist	$107,801	4.0/5	6,542	View Jobs
2019	#1	Data Scientist	$108,000	4.3/5	6,510	View Jobs
2018	#1	Data Scientist	$110,000	4.2/5	4,524	View Jobs
2017	#1	Data Scientist	$110,000	4.4/5	4,184	View Jobs
2016	#1	Data Scientist	$116,840	4.1/5	1,736	View Jobs

Q3: Is data scientist a good job?

Let’s get back to data:

In 2024, salary distribution of “Data Scientists” in the US:

- 10th Percentile: $ 63,650
- 25th percentile: $ 82,630
- Median: $112,590 (HK $631,629 after PPP adjustment)
- 75th percentile: $155,810
- 90th percentile: $194,410



Data Scientist salaries in Hong Kong

Experience

All years of experience

Base pay

HK$25K - HK$45K/mo

HK$33K/mo Average base pay

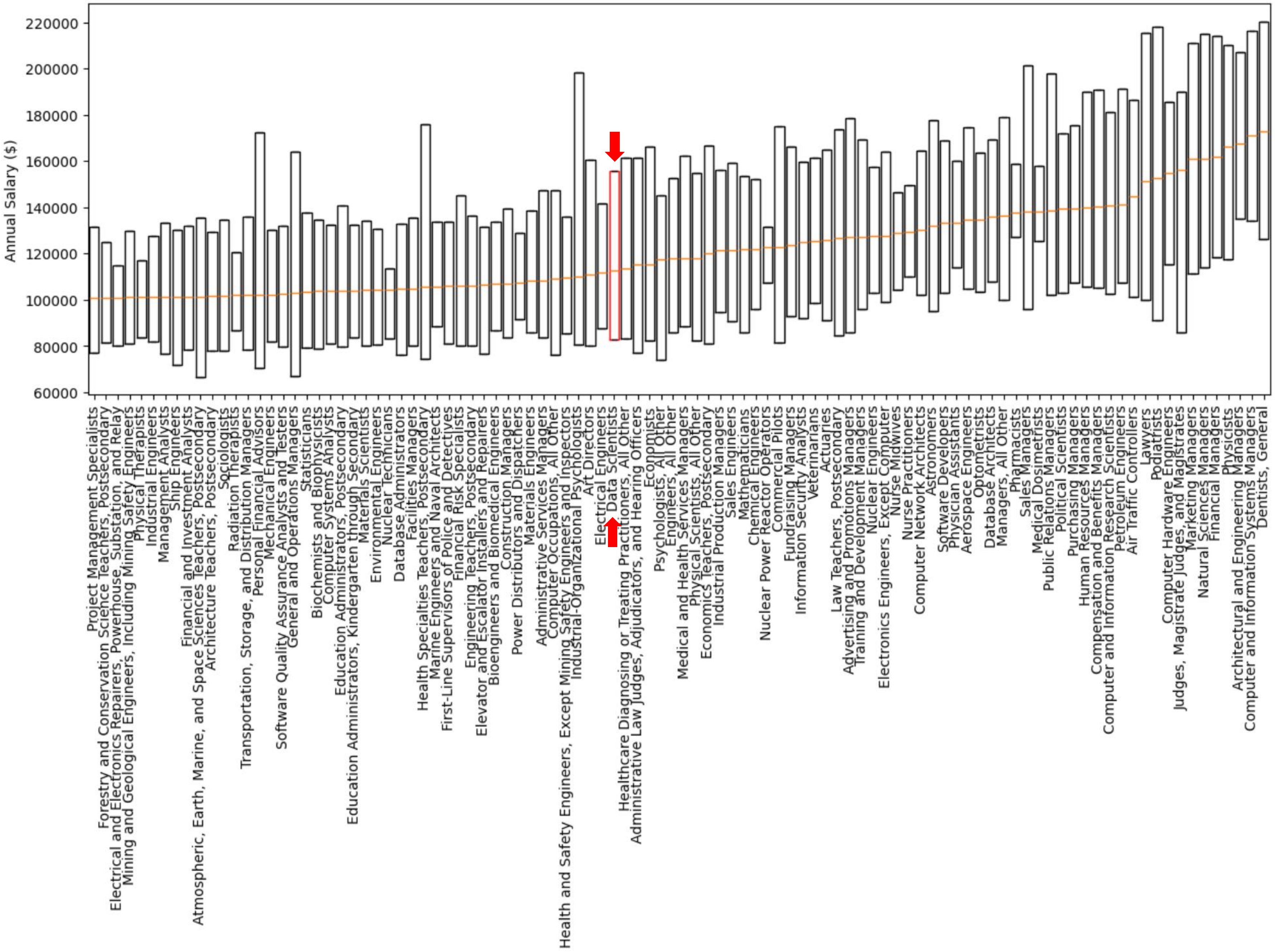
Additional salary

HK$4K/mo Average

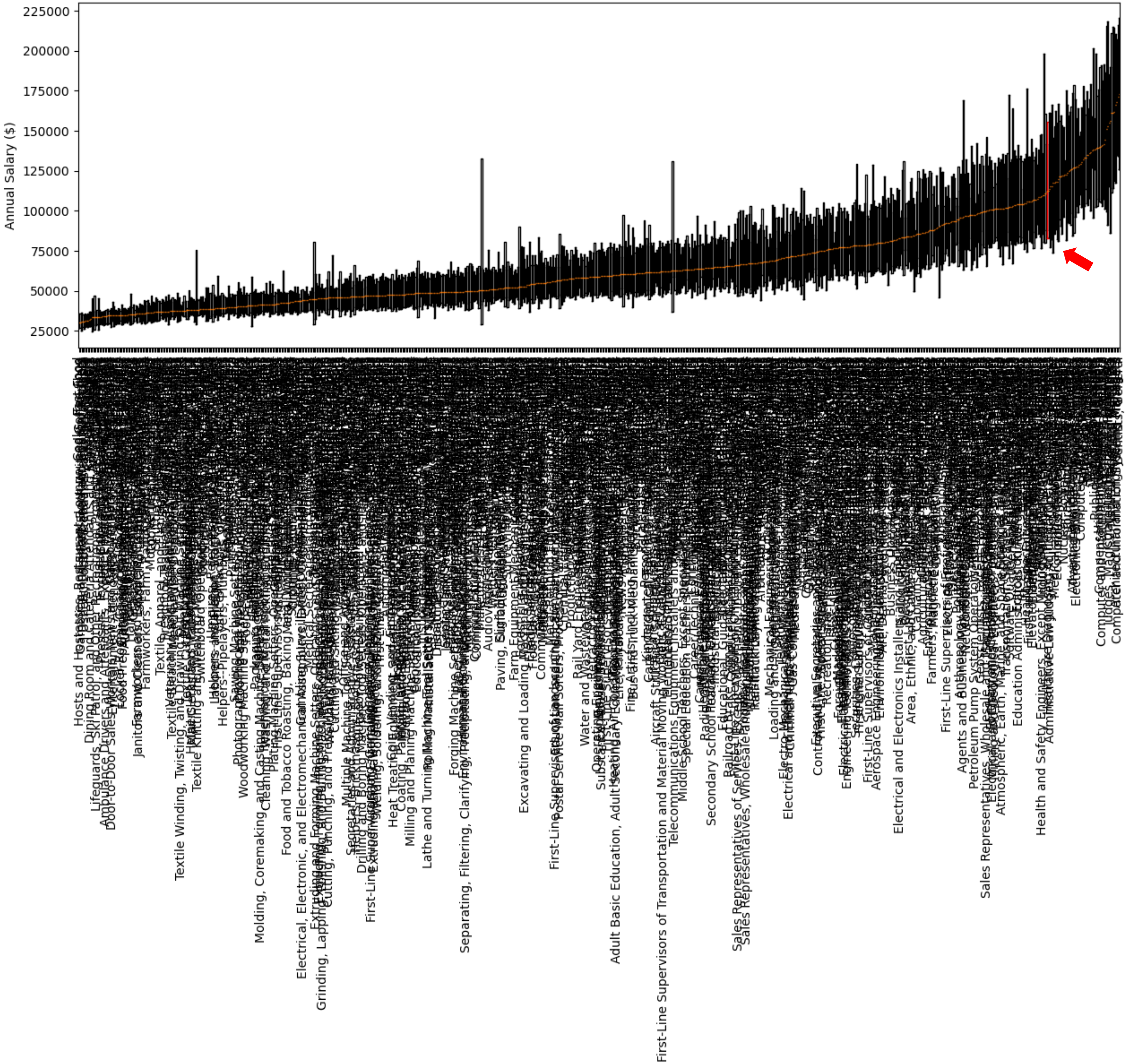
HK$3K - HK$7K/mo Range

HK$300k to 540k annual

Source: https://www.bls.gov/oes/tables.htm
https://www.glassdoor.com.hk/Salaries/data-scientist-salary-SRCH_KO0,14.htm



Source: https://www.bls.gov/oes/tables.htm

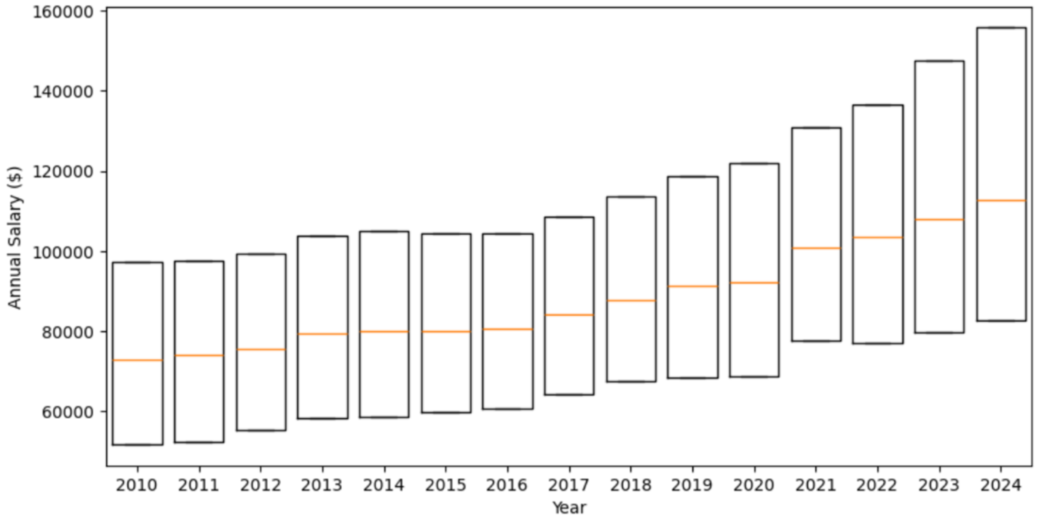

Annual Salary ($)
225000
200000
175000
150000
125000
100000
75000
50000
25000
Hosts and Hostesses, Restaurant
Lifeguards
Door-to-Door Sales
Janitors
Farmworkers, Farm
Textile Winding, Twisting, and Drawing Out
Molding, Coremaking, and Casting Machine
Food and Tobacco Roasting, Baking
Electrical, Electronic, and Electromechanical
Grinding, Lapping, Polishing
First-Line
Heat Treating
Milling and Planing Machine
Lathe and Turning Machine
Separating, Filtering, Clarifying
Excavating and Loading
Paving, Surfacing
First-Line Supervisors
Postal Service Mail Sorters
Water and
Adult Basic Education, Adult Secondary
First-Line Supervisors of Transportation and Material Moving
Telecommunications Equipment Installers
Secondary School
Sales Representatives of Services
Sales Representatives, Wholesale
Electrical
Electrical and Electronics Installers
Aerospace
Area, Ethnic
First-Line Supervisors
Agents and Business
Petroleum Pump System
Sales Representatives, Wholesale
Atmospheric, Earth
Education Administrators
Health and Safety Engineers, Except Mining
Computer

Source:
https://www.bls.gov/oes/tables.htm

Q3: Is data scientist a good job?

Does data scientist earn more/less over the year?

Note that OEWS does not have data scientists as a job category before 2021. We are using “Statistician” as surrogates.

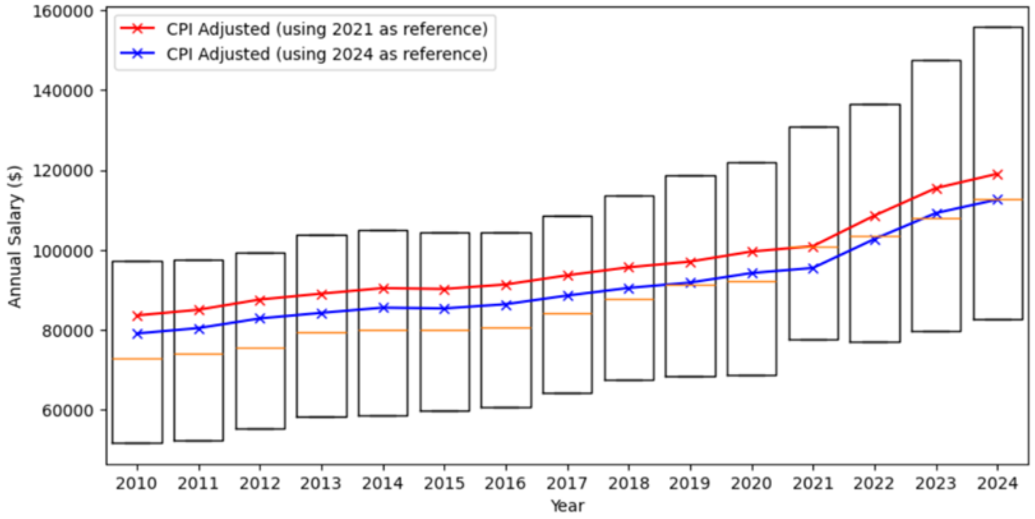


Source: https://www.bls.gov/oes/tables.htm, https://www.bls.gov/cpi/

Q3: Is data scientist a good job?

Does data scientist earn more/less over the year?

Note that OEWS does not have data scientists as a job category before 2021. We are using "Statistician" as surrogates.

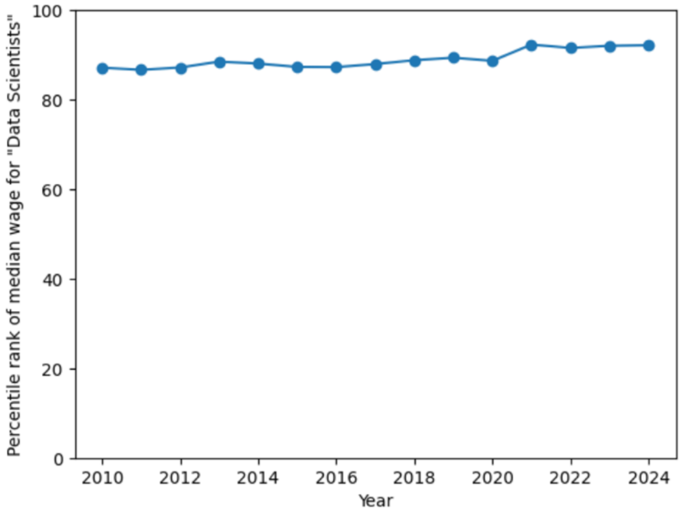


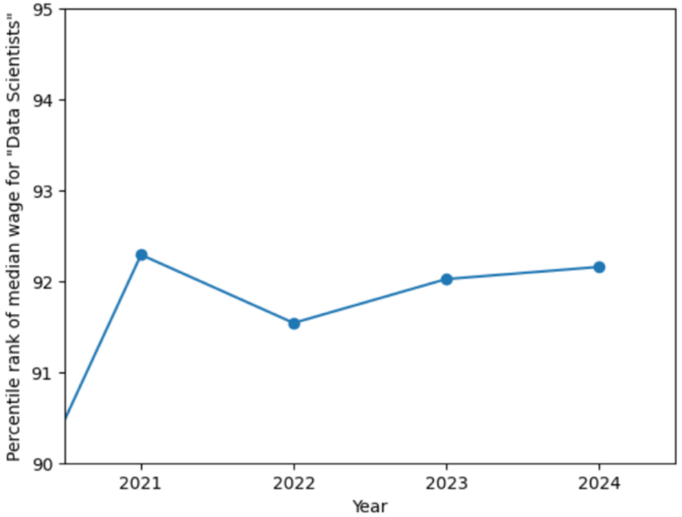
Source: https://www.bls.gov/oes/tables.htm, https://www.bls.gov/cpi/

Q3: Is data scientist a good job?

Does data scientist earn more/less over the year?

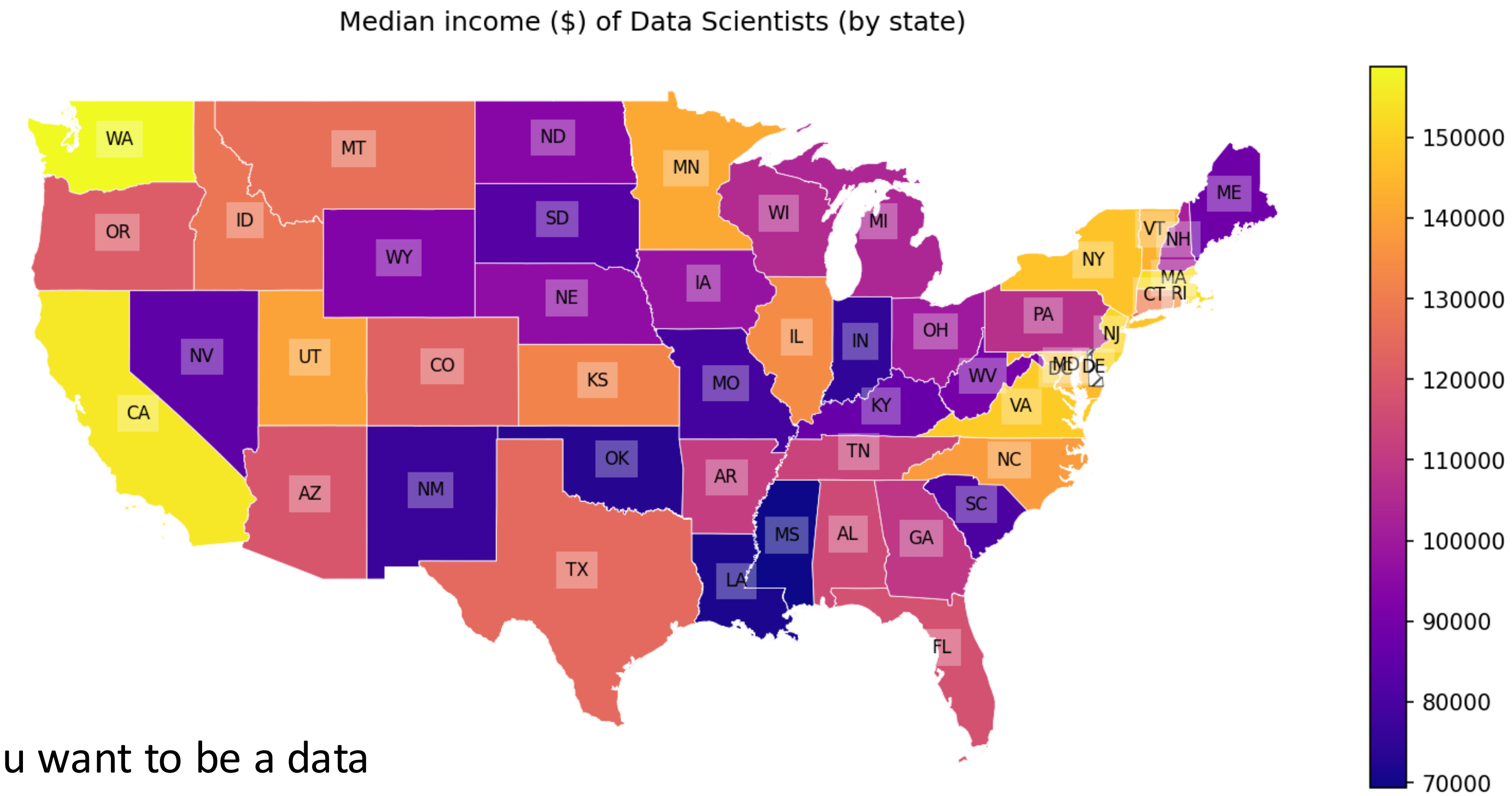
Note that OEWS does not have data scientists as a job category before 2021. We are using "Statistician" as surrogates.





Source: https://www.bls.gov/oes/tables.htm

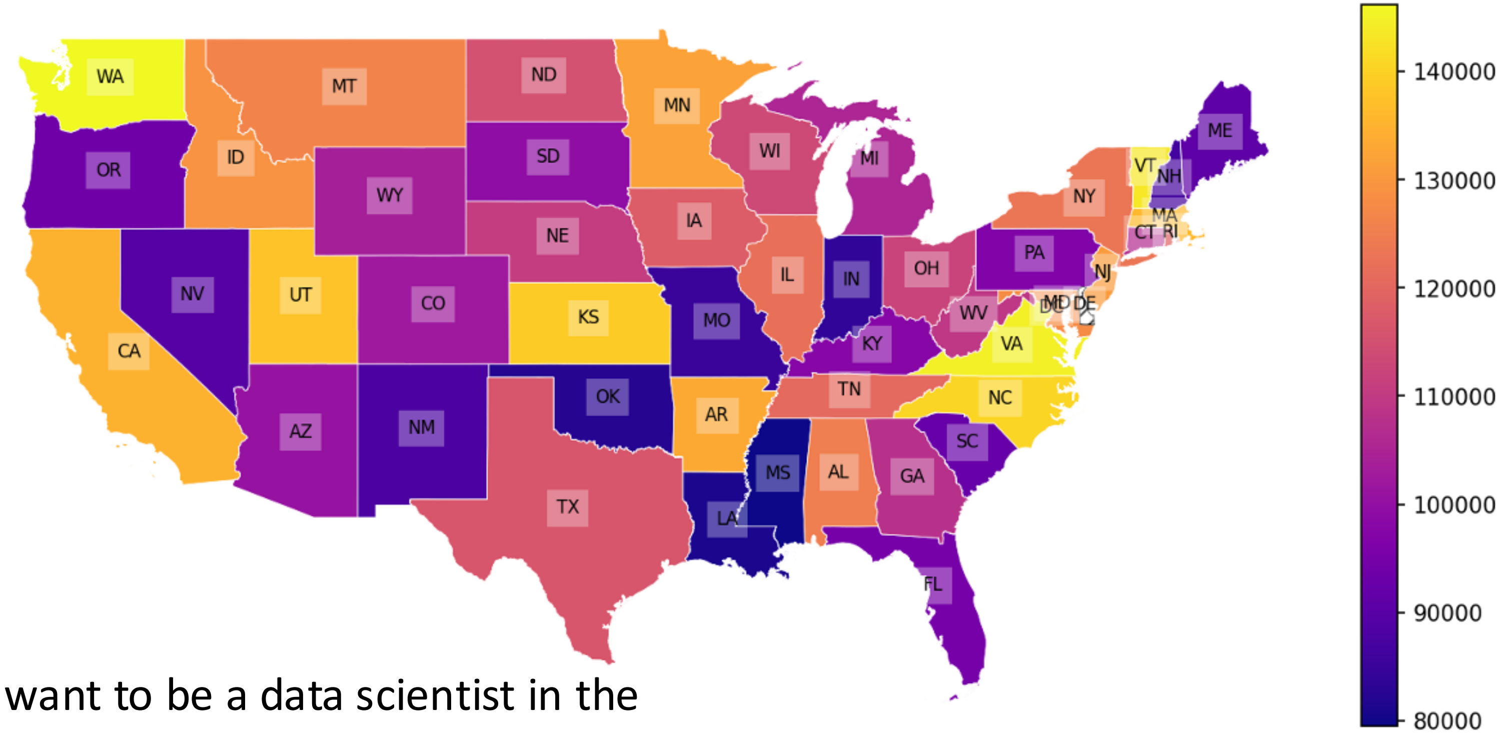
Q3: Is data scientist a good job?



If you want to be a data scientist in the U.S., where is the best place to get a job?

Source: https://www.bls.gov/oes/tables.htm

Q3: Is data scientist a good job?


Median income ($) of Data Scientists (by state), adjusted for regional price parities

If you want to be a data scientist in the U.S., where is the best place to get a job after considering living costs?

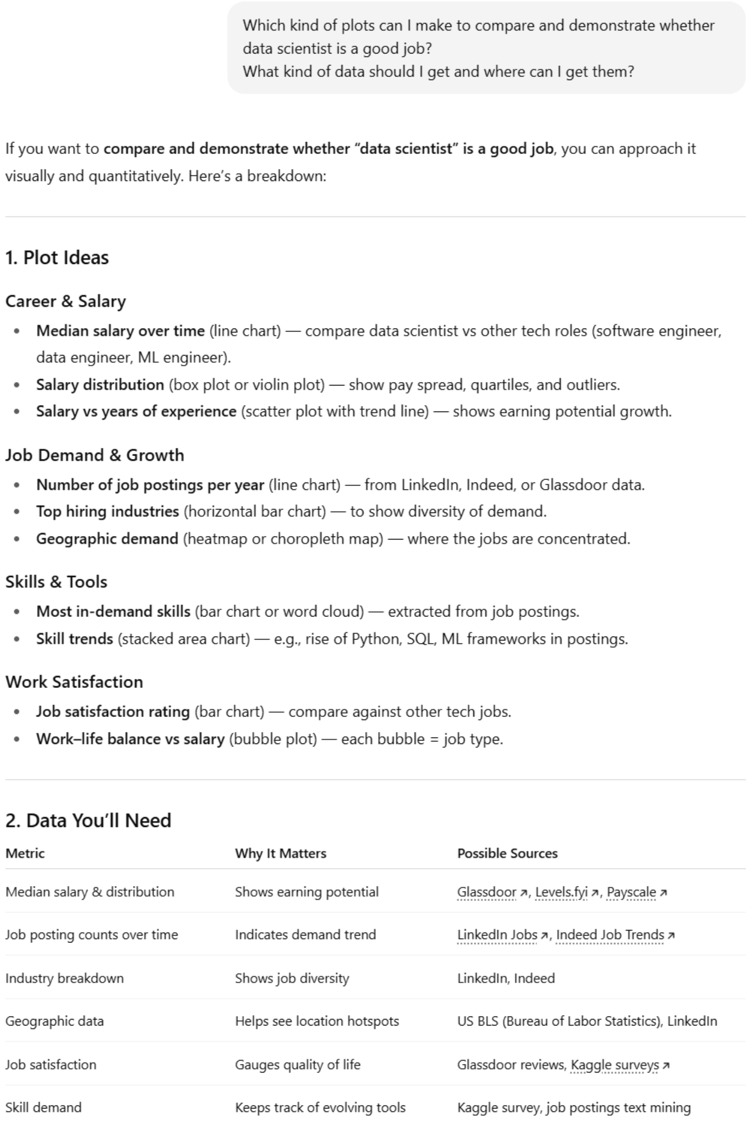
Source: https://www.bls.gov/oes/tables.htm, https://www.bea.gov/data/prices-inflation/regional-price-parities-state-and-metro-area

Q3: Is data scientist a good job?

So, is it a good job?

- More factors that we haven’t covered:
    - Training needed
    - Job demand
    - Work-life balance
    - Growth and satisfaction

- Data used for the previous plots are available in Moodle, make your own plots and derive your own **insights.**



Which kind of plots can I make to compare and demonstrate whether data scientist is a good job?
What kind of data should I get and where can I get them?

If you want to **compare and demonstrate whether “data scientist” is a good job**, you can approach it visually and quantitatively. Here’s a breakdown:

---

1. Plot Ideas

Career & Salary

- **Median salary over time** (line chart) — compare data scientist vs other tech roles (software engineer, data engineer, ML engineer).
- **Salary distribution** (box plot or violin plot) — show pay spread, quartiles, and outliers.
- **Salary vs years of experience** (scatter plot with trend line) — shows earning potential growth.

Job Demand & Growth

- **Number of job postings per year** (line chart) — from LinkedIn, Indeed, or Glassdoor data.
- **Top hiring industries** (horizontal bar chart) — to show diversity of demand.
- **Geographic demand** (heatmap or choropleth map) — where the jobs are concentrated.

Skills & Tools

- **Most in-demand skills** (bar chart or word cloud) — extracted from job postings.
- **Skill trends** (stacked area chart) — e.g., rise of Python, SQL, ML frameworks in postings.

Work Satisfaction

- **Job satisfaction rating** (bar chart) — compare against other tech jobs.
- **Work–life balance vs salary** (bubble plot) — each bubble = job type.

---

2. Data You’ll Need

Metric	Why It Matters	Possible Sources
Median salary & distribution	Shows earning potential	Glassdoor ↗, Levels.fyi ↗, Payscale ↗
Job posting counts over time	Indicates demand trend	LinkedIn Jobs ↗, Indeed Job Trends ↗
Industry breakdown	Shows job diversity	LinkedIn, Indeed
Geographic data	Helps see location hotspots	US BLS (Bureau of Labor Statistics), LinkedIn
Job satisfaction	Gauges quality of life	Glassdoor reviews, Kaggle surveys ↗
Skill demand	Keeps track of evolving tools	Kaggle survey, job postings text mining

Q4: How to be a good data scientist?

- Be good at extracting insights from data:
  - Read a lot, question a lot, think critically
  - Use the power of computing and plotting
  - Be scientifically rigorous

- Learn/know more than just data science

  ⇒Be interdisciplinary

  ⇒Incorporate your domain expertise

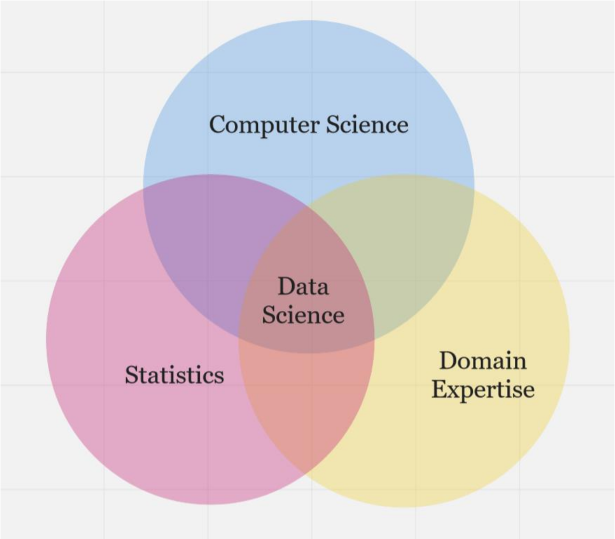
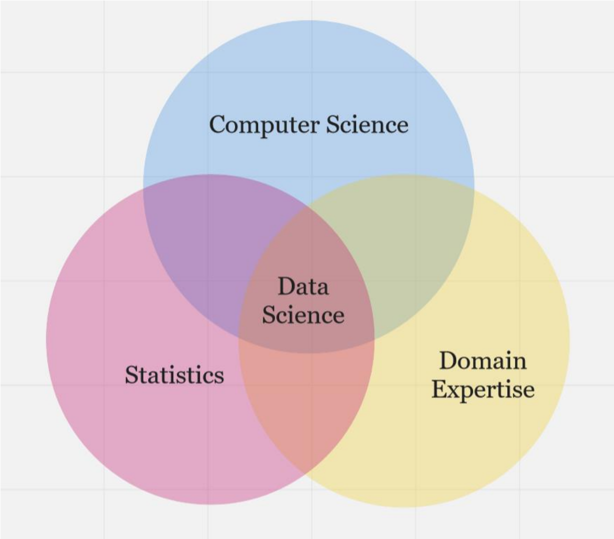


Diagram from: https://towardsdatascience.com/the-three-building-blocks-of-data-science-2923dc8c2d78/

Q4: How to be a good data scientist?

We learn/use data science to:

- Understand the big picture of the data
- Ask the right questions about the data
- Analyze pieces of data to answer the questions
- Communicate the resulting data insights

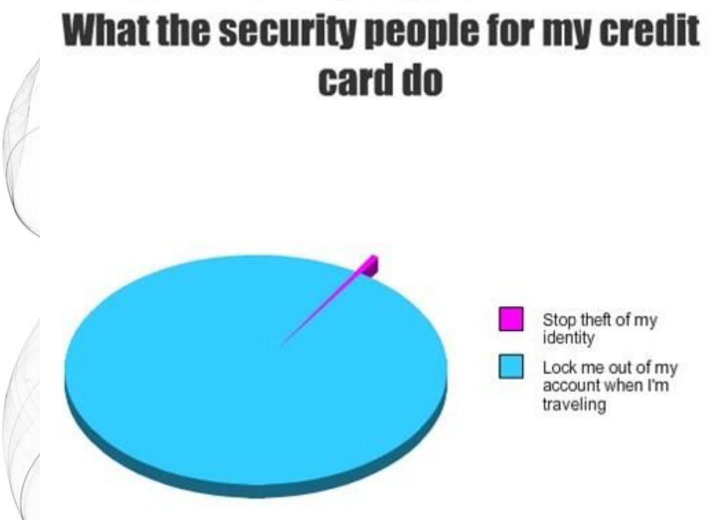


All of the above need us to have some understandings about the data, which is often called domain expertise.

"Data Science is not an end in itself, but a discipline that seeks to answer questions from other disciplines with data."

Diagram from: https://towardsdatascience.com/the-three-building-blocks-of-data-science-2923dc8c2d78/
Quote from: https://medium.com/codex/why-data-scientists-need-domain-expertise-111d33a6cf0a

Case study: Fraud Detection in Finance



DECIDE TO HAVE A SALAD FOR LUNCH, STOP AT THE HEALTH FOOD STORE
CREDIT CARD COMPANY CALLS TO ASK IF MY CARD HAS BEEN STOLEN
made on imgur

- Objective: develop a model that can automatically detect fraud based on user behavior
- Dataset: transactions made by credit cards in September 2013 by European cardholders.
- Questions before we proceed:
  - What defines a fraud? How frequent are frauds?
  - Which feature(s) is/are more likely to be important? Do we need to prioritize any feature(s)?
  - How to evaluate the (real-world) effectiveness of the model?
  - Should we avoid false positives or false negatives?
  - Is one month of data enough?

See also: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

My own experience: modeling a chemical property

- Objective: modeling atomization energy of molecules
- Dataset: QM7

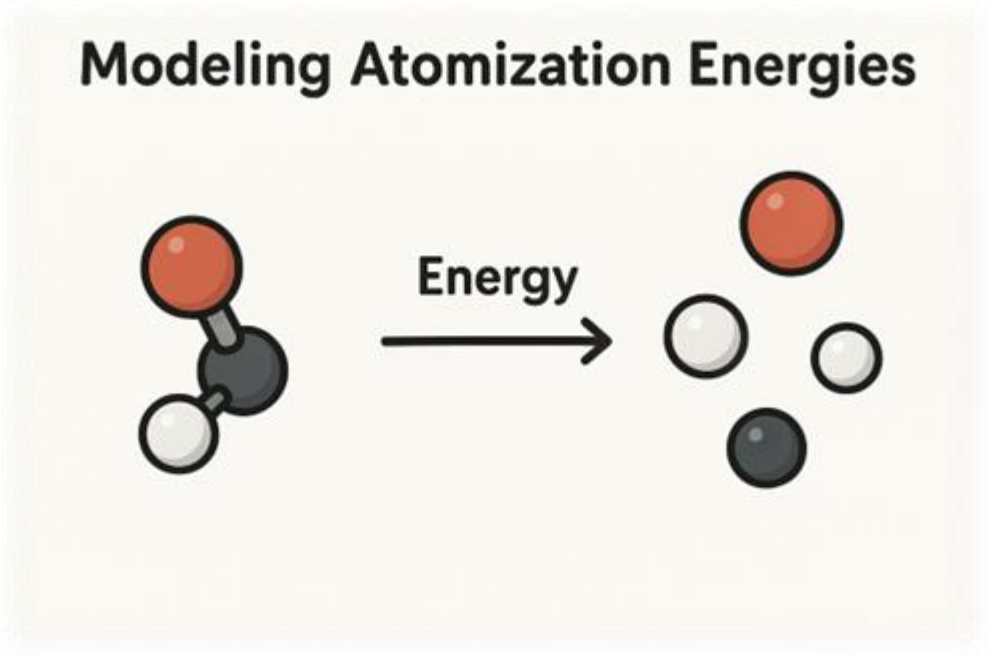
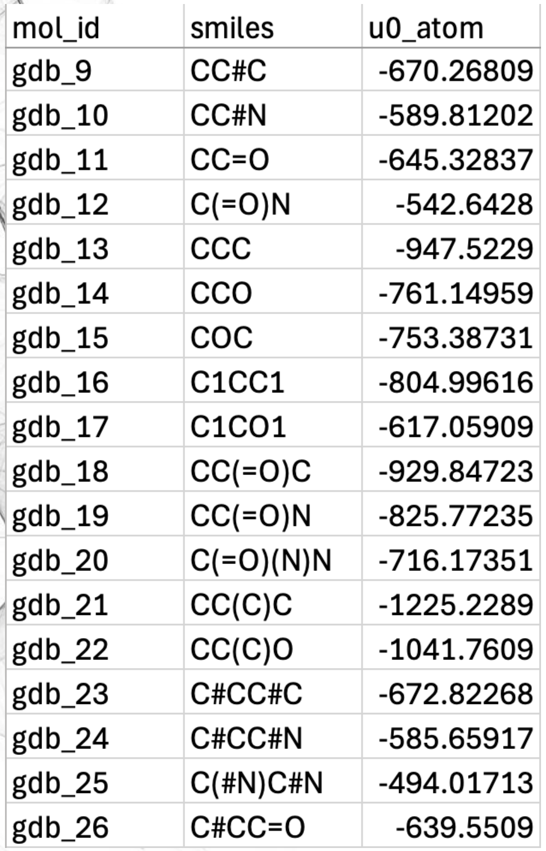


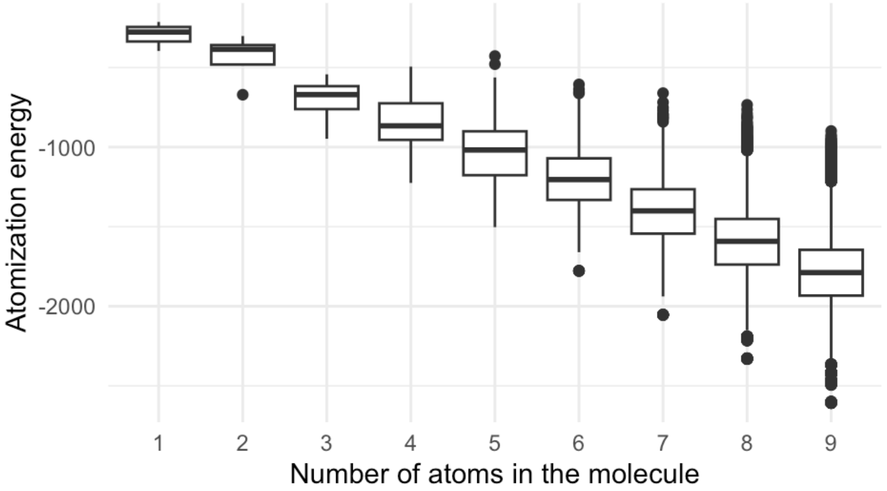
Figure created by GPT, See the full post at: https://medium.com/@pandelab/some-dangers-of-label-normalization-in-ml-6482d955882e

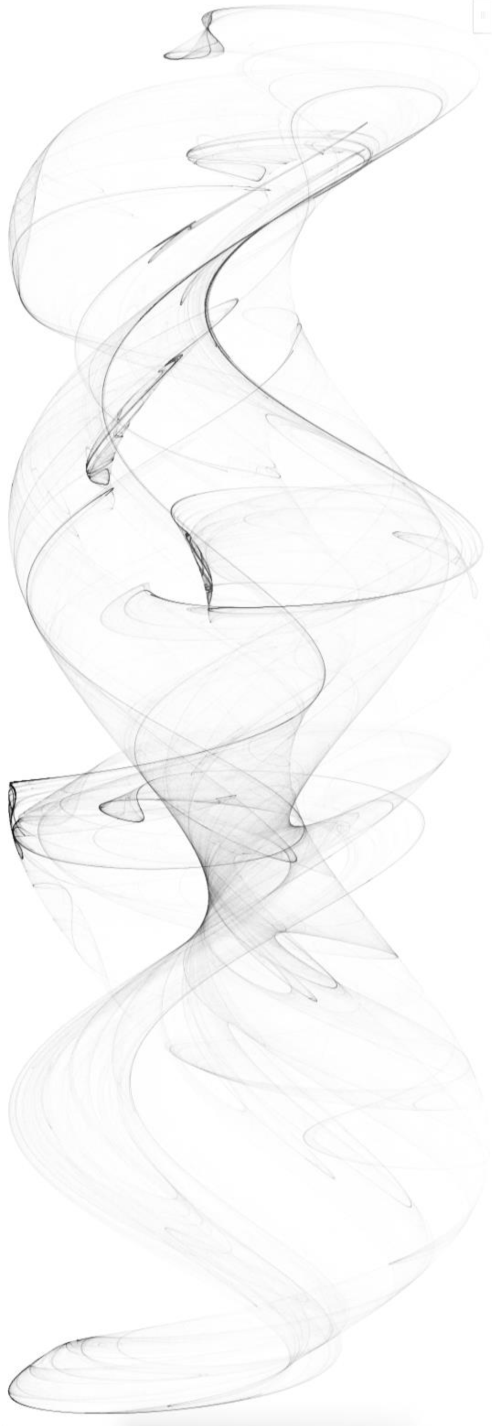
My own experience: modeling atomization energy



mol_id	smiles	u0_atom
gdb_9	CC#C	-670.26809
gdb_10	CC#N	-589.81202
gdb_11	CC=O	-645.32837
gdb_12	C(=O)N	-542.6428
gdb_13	CCC	-947.5229
gdb_14	CCO	-761.14959
gdb_15	COC	-753.38731
gdb_16	C1CC1	-804.99616
gdb_17	C1CO1	-617.05909
gdb_18	CC(=O)C	-929.84723
gdb_19	CC(=O)N	-825.77235
gdb_20	C(=O)(N)N	-716.17351
gdb_21	CC(C)C	-1225.2289
gdb_22	CC(C)O	-1041.7609
gdb_23	C#CC#C	-672.82268
gdb_24	C#CC#N	-585.65917
gdb_25	C(#N)C#N	-494.01713
gdb_26	C#CC=O	-639.5509

- Breakdown:
  - Input: molecule, defined as a set of atoms, connected through bonds
  - Task/label/output: energy required to break all the bonds and isolate all atoms



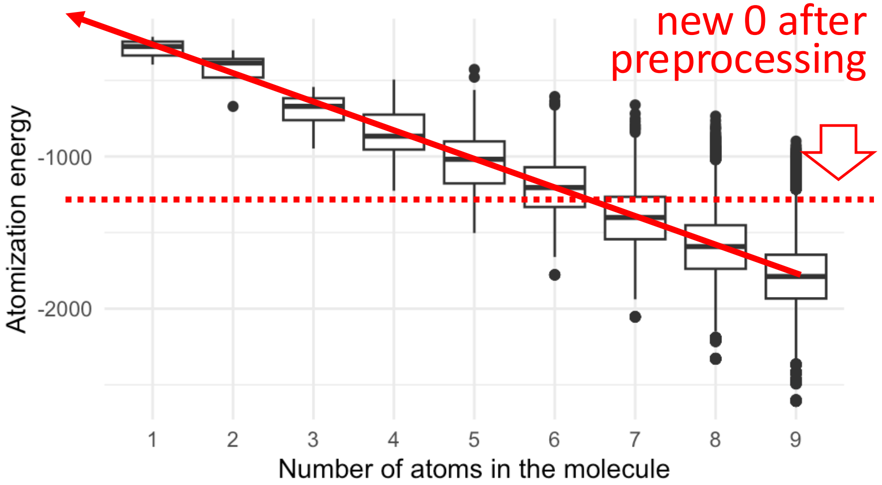


My own experience: modeling atomization energy

- Data preprocessing involved a z-scoring step:

$$y' = (y - \bar{y})/s_y$$

- However, atomization energies are intrinsically associated with molecule size
    - Smaller molecules => negative values
    - Larger molecules => positive values

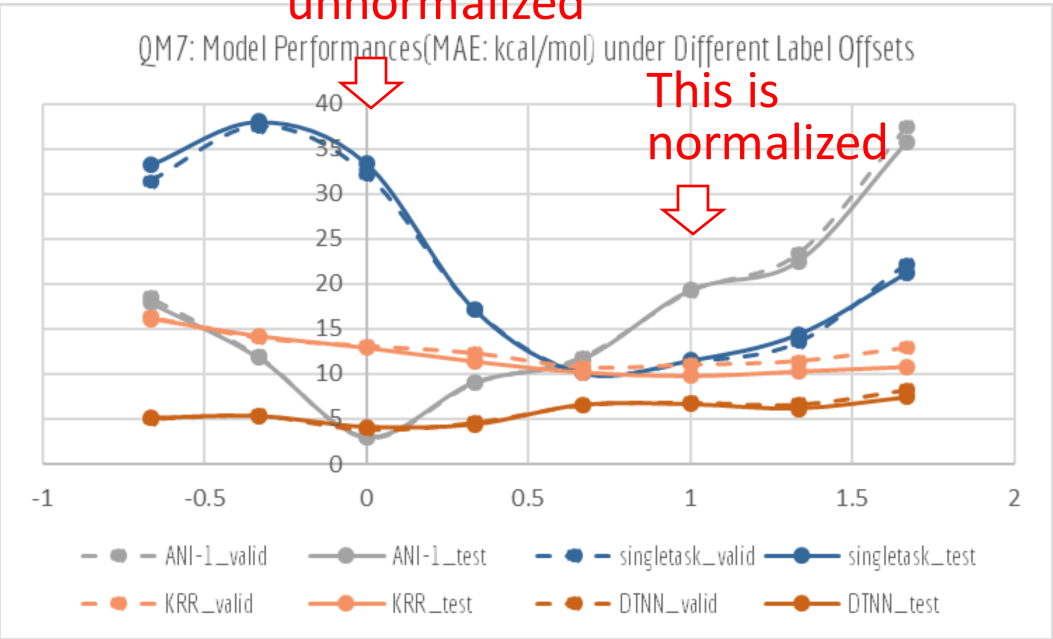


(0, 0)

This line becomes the new 0 after preprocessing

My own experience: modeling atomization energy

- Blue and light orange curves are methods built on full molecules
  - => They perform better on normalized output
- Grey and dark orange curves are methods built on individual atoms
  - => They perform better on unnormalized input



This is unnormalized

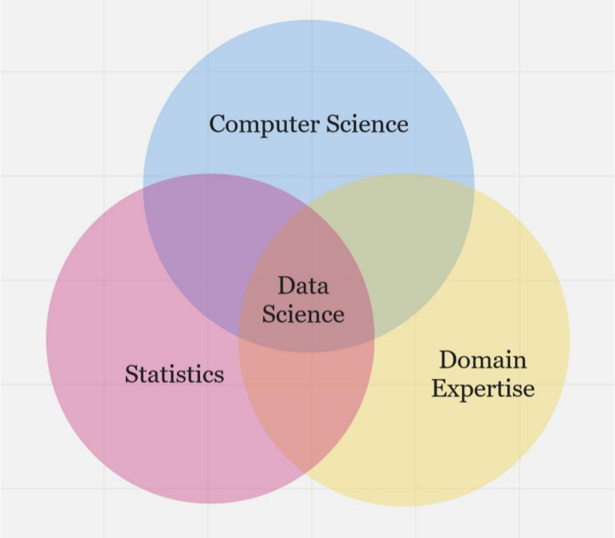
This is normalized

See the full post: https://medium.com/@pandelab/some-dangers-of-label-normalization-in-ml-6482d955882e

My own experience: modeling atomization energy

- What are the important factors in this data science project?
  - The task: how is it defined? What are known relevant factors?
  - The model: how does it treat the input?
  - The data processing: Will it change the intrinsic relations between inputs and outputs?

- All of the above need good understandings of the domain.



In this lecture

- Intro: four questions
- **Course details**

Welcome to the course

- No matter which level/type/domain of data scientist you want to be, this course will help you learn the very basics

- Aim of this course is to help you:
  - Understand the intuitions behind data science pipelines;
  - Be familiar with R programming and relevant packages;
  - Be able to collect and read data from different sources;
  - Know how to explore and visualize your data ;
  - Analyze your data rigorously using a variety of statistical and machine learning approaches .

Course detail and resources

- Course Materials
  - Moodle: **COMP2501 Introduction to data science [Section 1A, 2025]**
- Format
  - Lectures on Monday & Thursday: **CYPP3** (Chong Yuet Ming Physics Building); You are encouraged to bring your laptop.
  - Additional course materials will be shared through Moodle;
  - **No zoom streaming or recording will be provided.**
- Lecture Hours
  - **Monday** 9:00 am - 10:50 pm
  - **Thursday** 3:00 pm - 3:50 pm
- For Questions and Discussions
  - Google & GPT (more on that later)
  - Try Moodle forum “Q&A Forum”
  - Email to TA/Instructor
  - Consultation hour
    - Michael Wu (me): write to me (zqwu@cs.hku.hk) to set up an appointment;
    - TAs: consultation hours will be visible on Moodle later.

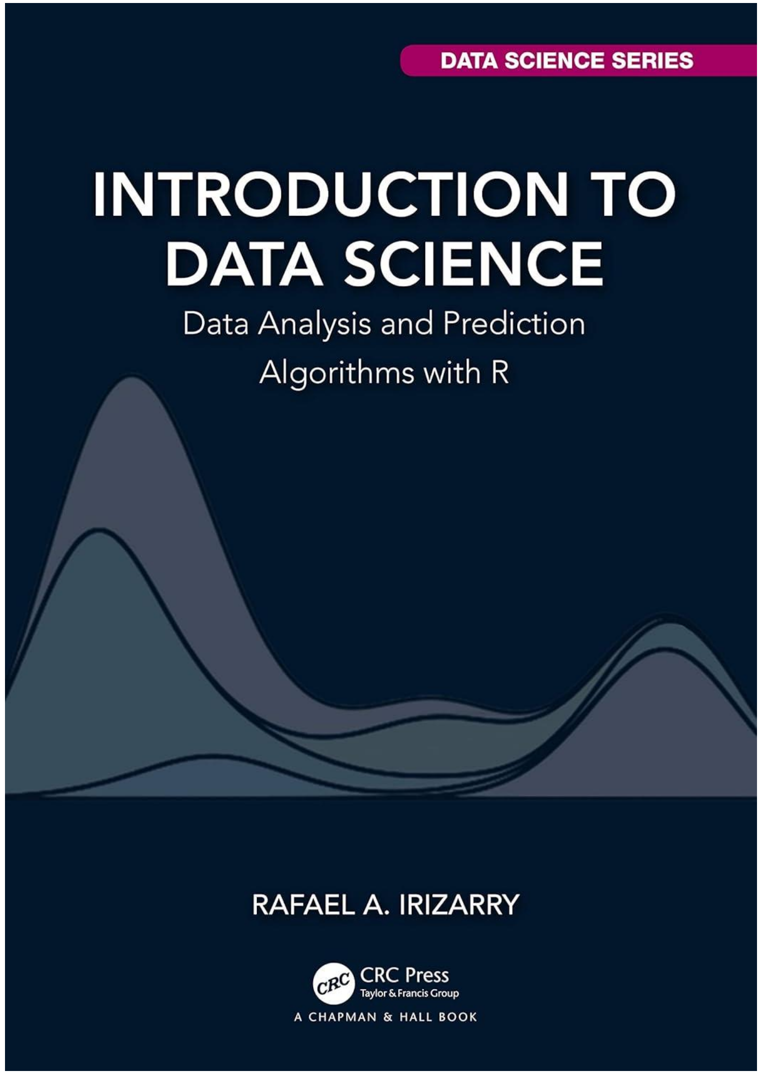
Teaching Staff

- Wu, Zhenqin (Michael); Instructor:
    - CB404, email: zqwu@cs.hku.hk
- CHENG, Man Yee (Jolly); Teaching Assistant 1
    - CB319, email: mycheng@cs.hku.hk
- LIU, Yufan; Teaching Assistant 2
    - HW307, email: yufan.liu@connect.hku.hk
- LI, Yiyao (Leo); Teaching Assistant 3
    - HW307, email: liyiyao@connect.hku.hk

Textbook

- Introduction to Data Science - Data Analysis and Prediction Algorithms with R
  - by Prof. Rafael A. Irizarry

- Freely available at http://rafalab.dfci.harvard.edu/dsbook/



Python vs R

- Two two major programming languages used in data science:
  - R: better at statistical analysis, tabular data manipulation (`tidyverse`), data visualization (`ggplot2`)
    - More specialized for “analysis”
  - Python: general-purpose programming, machine learning (`scikit-learn`) and deep learning (`pytorch`, `transformers`, etc.)
    - Better at “modeling”
- This course will be taught in R, but try to learn both

Syllabus

- 🟩 Lectures
- 🟩 Additional Topics
- 🟨 Interactive Learning Sessions
- 🟥 Exams

	Week	Day	Time	Hour	Content	Notes
Sept. 1st	1	Mon	9:00-9:50, 10:00-10:50 am	2	Introduction and R basics	
Sept. 4th	1	Thu	3:00-3:50pm	1	R basics & R markdown	
Sept. 8th	2	Mon	9:00-9:50, 10:00-10:50 am	2	Tidyverse	
Sept. 11th	2	Thu	3:00-3:50pm	1	Tidyverse, work with external datasets	Assignment 1
Sept. 15th	3	Mon	9:00-9:50, 10:00-10:50 am	2	Data visualization	
Sept. 18th	3	Thu	3:00-3:50pm	1	Data visualization in practice	
Sept. 22nd	4	Mon	9:00-9:50, 10:00-10:50 am	2	Data visualization principles	
Sept. 25th	4	Thu	3:00-3:50pm	1	**Lab session 1**	
Sept. 29th	5	Mon	9:00-9:50, 10:00-10:50 am	2	Data wrangling: reshaping, joining, web scraping	Assignment 2
Oct. 2nd	5	Thu	3:00-3:50pm	1	Data wrangling: regex & string processing	Assignment 1 deadline
Oct. 6th	6	Mon	9:00-9:50, 10:00-10:50 am	2	Text mining	
Oct. 9th	6	Thu	3:00-3:50pm	1	TBD: Recent topics in data science	
Oct. 13th	7	Mon		2	Reading Week	
Oct. 16th	7	Thu		1	Reading Week	
Oct. 20th	8	Mon	9:00-9:50, 10:00-10:50 am	2	**Mid-term**	
Oct. 23rd	8	Thu	3:00-3:50pm	1	**Project proposal write-up & Consultation**	Project proposal submission open
Oct. 27th	9	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: probability	
Oct. 30th	9	Thu	3:00-3:50pm	1	Statistics: probability & statistical methods	Assignment 2 deadline
Nov. 3rd	10	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: statistical methods	Assignment 3
Nov. 6th	10	Thu	3:00-3:50pm	1	Statistics: Bayesian statistics	
Nov. 10th	11	Mon	9:00-9:50, 10:00-10:50 am	2	Statistics: Inference & Regression	
Nov. 13th	11	Thu	3:00-3:50pm	1	Machine learning	
Nov. 17th	12	Mon	9:00-9:50, 10:00-10:50 am	2	Deep learning/artificial intelligence/LLM in practice	Project proposal submission deadline
Nov. 20th	12	Thu	3:00-3:50pm	1	**Lab session 2**	
Nov. 24th	13	Mon	9:00-9:50, 10:00-10:50 am	2	**Student presentations**	7 slots
Nov. 27th	13	Thu	3:00-3:50pm	1	**Student presentations**	3 slots; Assignment 3 deadline
TBD				2	**Final Exam**	Project report deadline

- Also available on Moodle;
- Any additional changes will be communicated through Moodle announcement.

Assessment

- 3 assignments: **30%**
  - Due at 11:59 pm on specified dates.
    - In the event that there are broad issues with the submission system, i.e., Moodle going down, we will change the deadline accordingly.
  - Format: R markdown, use only specified libraries;
  - **Plagiarism is strictly prohibited**!
- 1 project: **20%**
  - **ONE** student per project, propose and work on a real data-science problem;
  - Project proposal submission: Oct 23rd (9:00 am) to Nov 17th (11:59 pm);
  - Project report & presentation video submission: Nov 30th (11:59 pm);
  - 10 in-class presentation slots available with bonus points;
  - More details shared in the lecture on Oct 23rd.
- 1 mid-term exam: **20%**
  - In-class 2-hour Moodle quiz on Oct 20th, **NO make-up exam;**
  - Bring your own laptop or pad.
- 1 final exam: **30%**

Late policy

- Assignments and project report: 25% penalty for every late day
  - Deadline + 1 day: 75%
  - Deadline + 2 days: 50%
  - Deadline + 3 days: 25%
  - Deadline + 4 days: 0%
- Counted by the granularity of days (cutoff at 12:00 am). For example, if you turn in your assignment 30 minutes after the deadline (@12:29 am), you are late by 1 day.
- If you think you really really need an extension on a particular assignment due to unexpected circumstances, **contact the instructor as soon as possible and before the deadline.**
  - Acceptable: severe illness, major personal/family emergencies;
  - Unacceptable: competitions, interns, interviews, club-related events, etc.

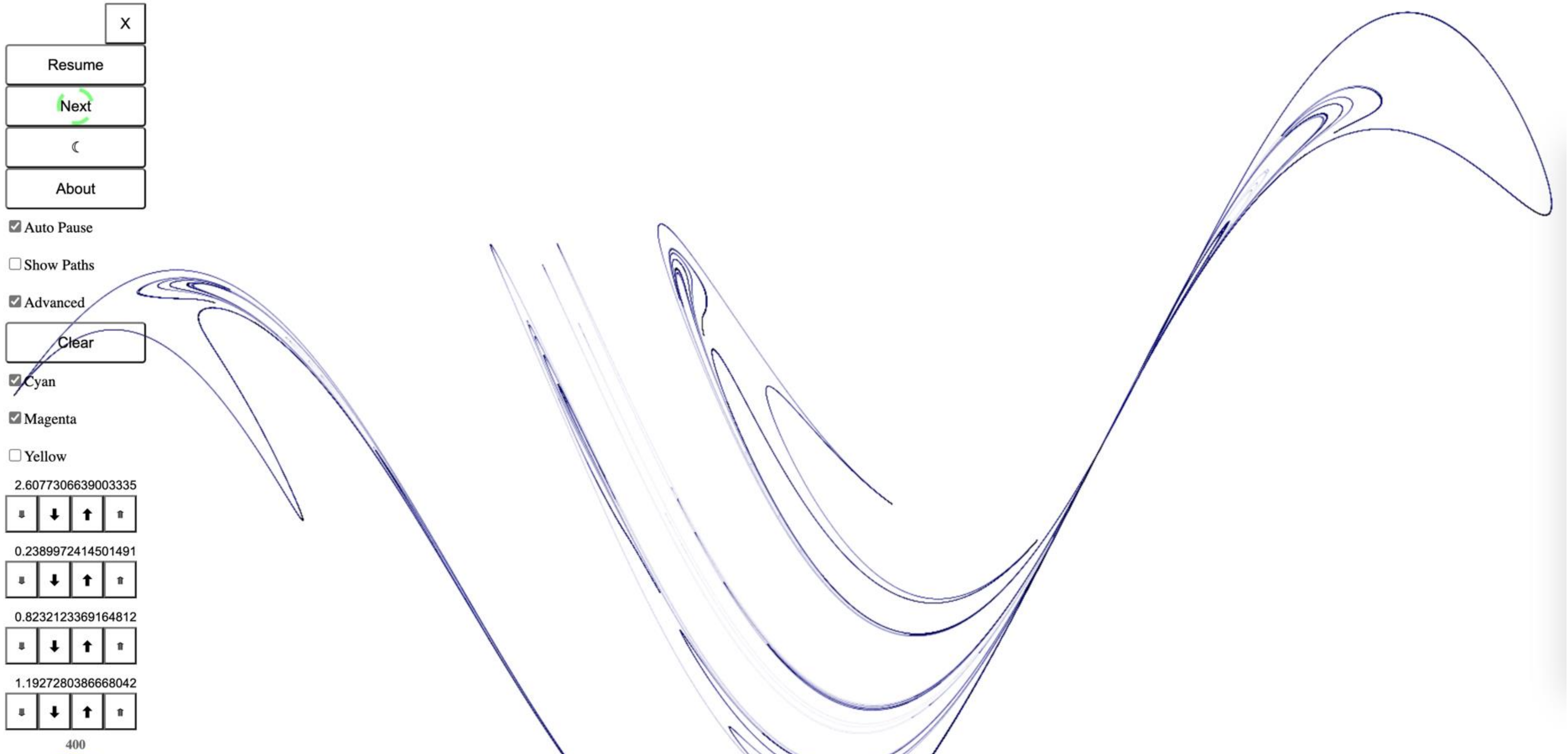
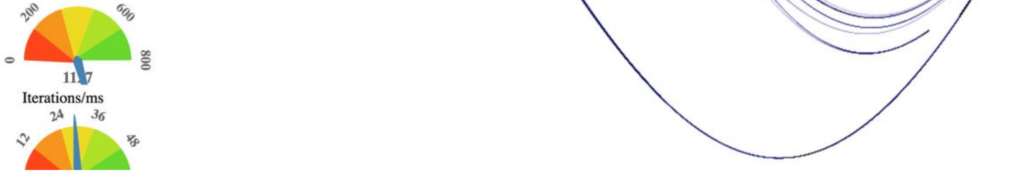
Regarding use of Generative AI

- University-wide guidelines
  - 1. Ethical Artificial Intelligence Framework
  - 2. Hong Kong Generative and Artificial Intelligence Technical and Application Guideline
- Course policy
  - Yes, you are encouraged to use LLMs: GPTs, DeepSeek, etc., to assist your study.
  - No, you are not encouraged (and there is no need) to use any image/video/voice generation tools.
- Rule of thumb: make sure you understand what LLMs are doing, especially when they are making mistakes.
  - When asking LLMs to generate code, make sure you understand every line of the output. Use APIs, documentations to help.
  - Be aware of hallucinations.
  - Don’t use LLMs for your assignments.

Regarding use of Generative AI

- Helpful prompt templates:
  - *"Provide a detailed explanation of [concept/topic] with examples and real-world applications."*
  - *"Compare and contrast [concept A] and [concept B], make a table to highlight their similarities and differences. "*
  - *"Write a [language] program that does [specific task]. Include detailed comments and documentations."*
  - *"What are some best practices for [specific aspect of coding, e.g., debugging, using APIs]?"*
- Which LLM to use?
  - ChatGPT through HKU: https://chatgpt.hku.hk/
  - OpenAI service through Microsoft Azure: https://its.hku.hk/software/azure/
  - POE (GPTs, Gemini, Claude, etc.): https://poe.com/
  - Grok 4 by xAI: https://grok.com/
  - DeepSeek: https://chat.deepseek.com/

About the title page picture

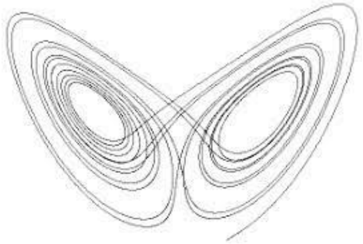



a 2.6077306639003335

b 0.2389972414501491

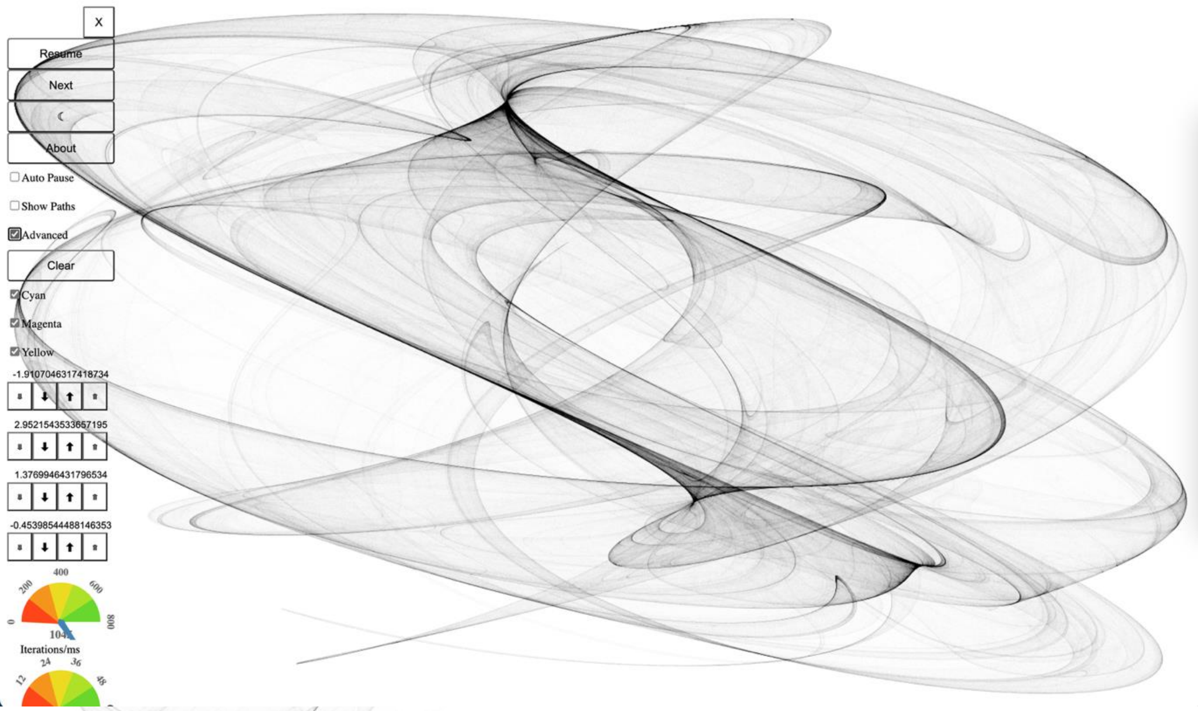
c 0.8232123369164812

d 1.1927280386668042



xn+1 = sin(b*yn) + c sin(b*xn)
yn+1 = sin(a*xn) + d cos(a*yn)

https://www.software-artist.com/chaotic-attractor/

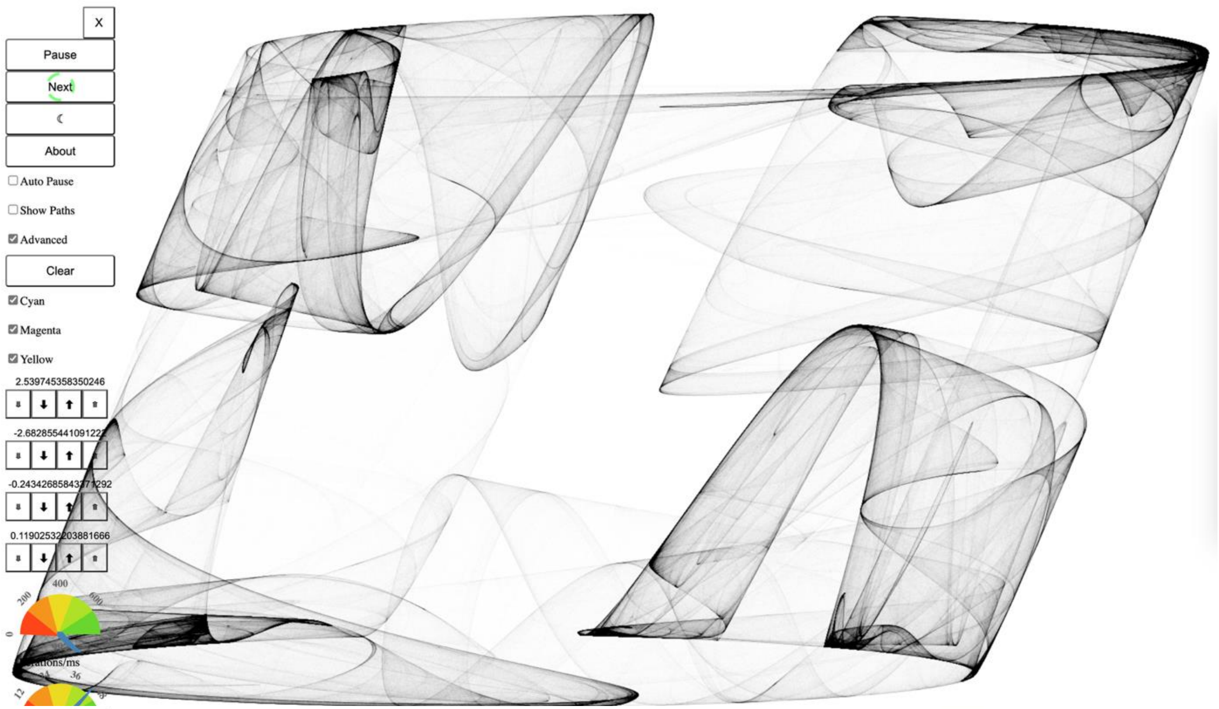


-1.9107046317418734

2.9521543533657195

1.3769946431796534

-0.4539854448814635З

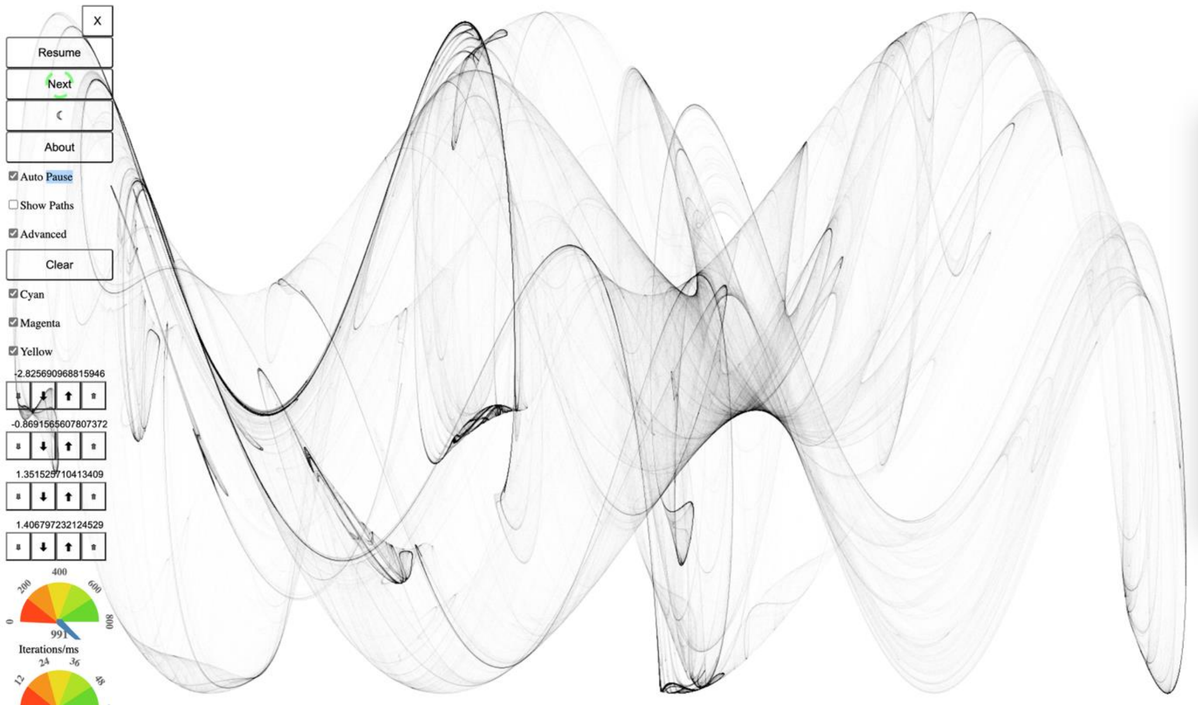


2.539745358350246

-2.68285544109122

-0.24342685843371292

0.11902532203881666

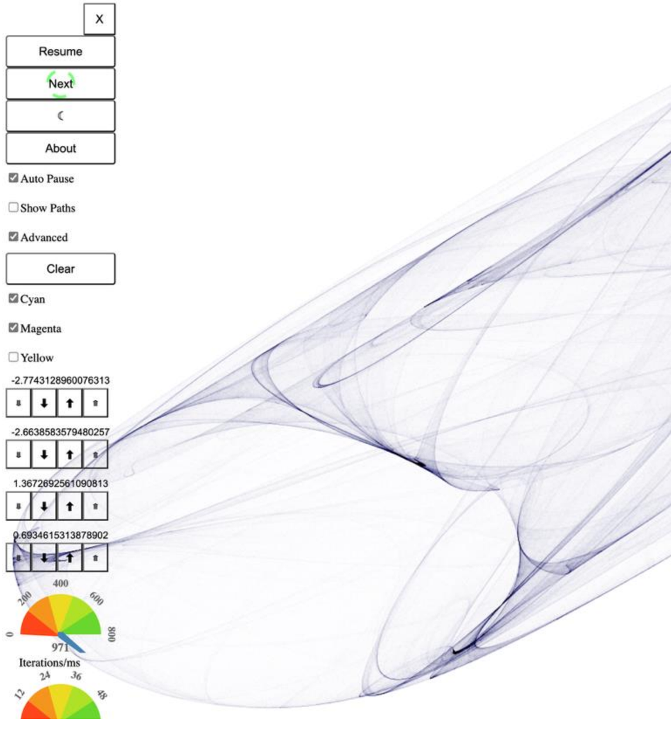


-2.825690968815946

-0.8691565607807372

1.351525710413409

1.406797232124529



-2.7743128960076313

-2.6638583579480257

1.3672692561090813

0.6934615313878902

$$x_{n+1} = \sin(b*y_n) + c \sin(b*x_n)$$

$$y_{n+1} = \sin(a*x_n) + d \cos(a*y_n)$$