Supervised Learning

Correlation

Errors & Artifacts

Variance

Gradient Descent

Sampling

Data Bias

Probability

Significance

Precision

Skew

Classification

Recall

F-Score

Charts & Plots

Unsupervised Learning

Machine Learning

Statistics

Prediction

Logistic Regression

Linear Regression

Clustering

Bias-Variance Tradeoffs

Data Science 1:
Introduction to Data Science

# Course Overview & Intro to Data Science

Winter 2025

## Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

Supervised Learning

Correlation

Errors & Artifacts

Variance

Gradient Descent

Sampling

Data Bias

Probability

Significance

Precision

Skew

Classification

Recall

F-Score

Charts & Plots

Unsupervised Learning

Machine Learning

Statistics

Prediction

Logistic Regression

Linear Regression

Clustering

Bias-Variance Tradeoffs

Data Science 1:
Introduction to Data Science

# Course Overview &

## Intro to Data Science

Winter 2025

## Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

# Meet Team Data



**Wolle**

Wolfram Wingerath
Data Science Prof

**Lectures**



**Jannik**

Jannik Schröder
PhD Student

**Exercises**



**Sophie**

Sophie Ferrlein,
Data Scientist at Finanzguru

**Awesome Guest Lecture**

# Meet Team Data: Wolle

**Research:**
- Stream Processing
- Real-Time Databases
- NoSQL & Cloud Systems
- …

**Wolfram Wingerath**
Data Science

+
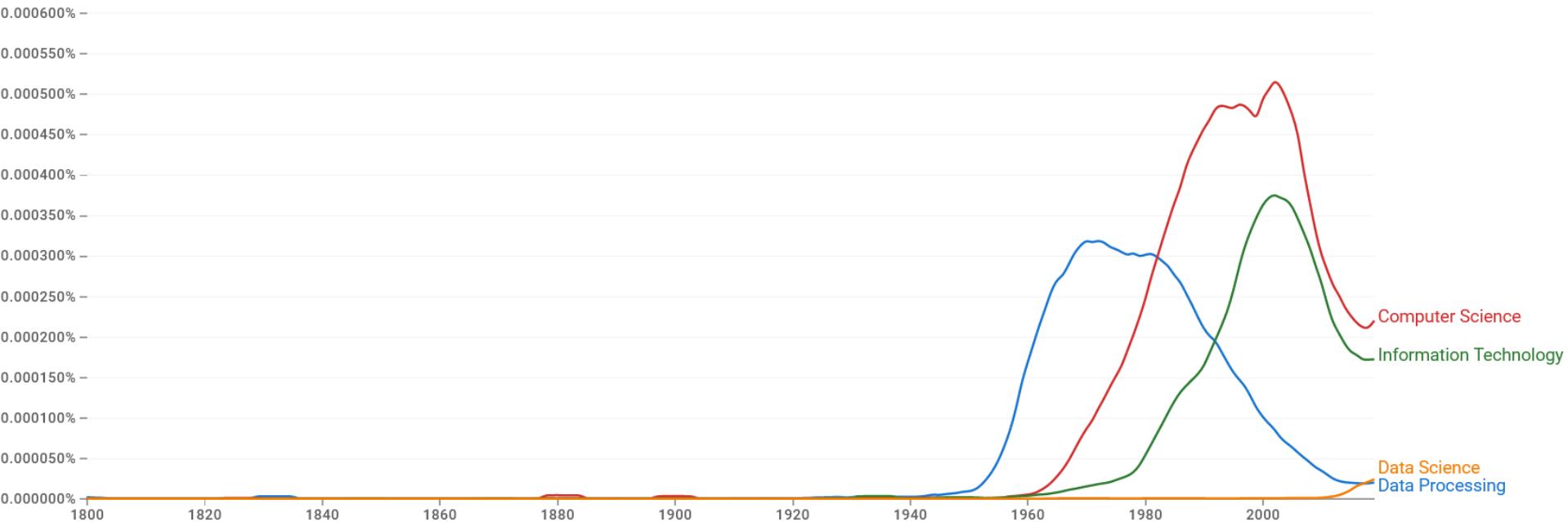
**Practice**:
Web Caching •
Big Data Analytics •
Anger Management •
… •

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Carl von Ossietzky
Universität
Oldenburg

Baqend

# Meet Team Data: Jannik
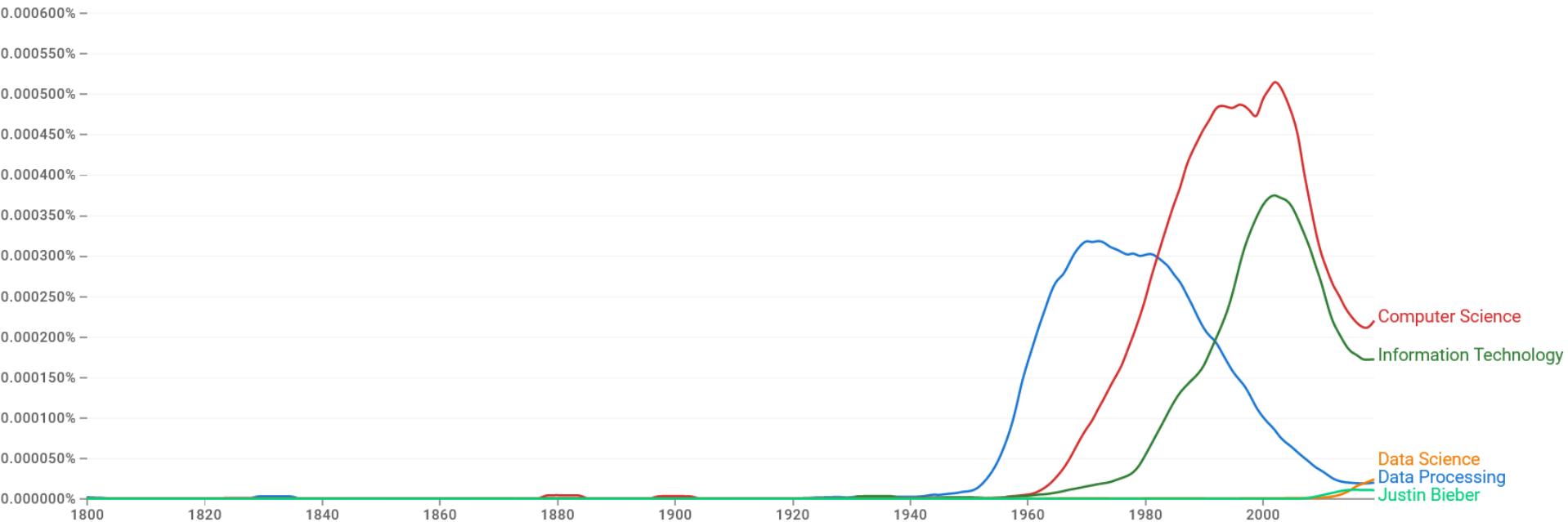
- Former digital forensic specialist, now research associate and PhD student in the DataScience Group.

- Research interests: Real-Time Databases and Anomaly Detection.

- PhD research: building a modular real-time query system.

# Data is not new to computing...



Google: Google Books Ngram Viewer (2023)

# Data is not new to computing...



Google: Google Books Ngram Viewer (2023)

# … So Why Data Science Now?

- New technology makes it possible to capture vast amounts of logging / sensor data.
- Computing advances make it possible to analyze data on ever increasing scales.
- Prominent role models (Google, Moneyball, hedge funds, Nate Silver, ...) have proven the power of modern data analytics.

# **Data Science is *not* Machine Learning**

- We will <u>not</u> be doing deep learning.
- But we will apply some machine learning.
- Our primary goal here is to learn to think about data: understanding, cleaning, visualization – not just building models.

# Context is Important

- Knowing how data is produced may be critical for interpreting it
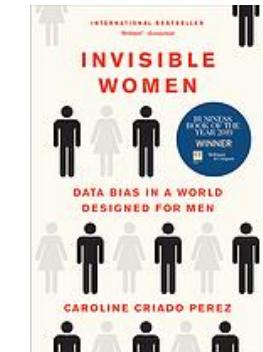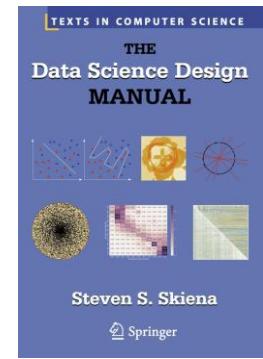
# My Experience with Data

- Core database topics at DB group in Hamburg
- Duplicate detection in probabilistic databases in bachelor's and master's thesis
- PhD on real-time databases (NoSQL + stream processing)
- Large-scale data analytics in web contexts
- Head of data engineering at Baqend (startup)

I have done <u>data engineering</u> for data science.

# The Data Science Design Manual

- This lecture is based on material from Steven Skiena, especially his book **"The Data Science Design Manual"** (Steven Skiena, 2017)
- Recommended texts include **"Invisible Women: Data Bias in a World Designed for Men"** (Caroline Criado-Perez, 2019)

# Semester Schedule

| | | | | | |
|---|---|---|---|---|---|
| CW 42 | 14. Oct | Lecture | 1 | Orga & Intro | 1-26 |
| CW 43 | 21. / 23. Oct | Lecture + Exercises | 2 | Probability, Statistics & Correlation | 27-56 |
| CW 44 | 28. Oct | Lecture | 3 | Data Munging, Cleaning & Bias | 57-94 / "Invisible Women" |
| CW 45 | 04. / 06. Nov | Lecture + Exercises | 4 | Scores & Rankings | 95-120 |
| CW 46 | 11. Nov | Lecture | 5 | Statistical Distributions & Significance | 121-154 |
| CW 47 | 18. / 20. Nov | Lecture + Exercises | 6 | Building & Evaluating Models | 201-236 |
| CW 48 | 25. Nov | Guest Lecture | 7 | Data Visualization | 155-200 |
| CW 49 | 02. / 04. Dec | Lecture + Exercises | 8 | Intro to Machine Learning | 351-390 |
| CW 50 | 09. Dec | Lecture | 9 | Linear Algebra | 237-266 |
| CW 51 | 16. / 18. Dec | Lecture + Exercises | 10 | Linear Regression & Gradient Descent | 267-288 |
| CW 02 | 06. Jan | Lecture | 11 | Logistic Regression & Classification | 289-302 |
| CW 03 | 13. / 15. Jan | Lecture + Exercises | 12 | Nearest Neighbor Methods & Clustering | 303-350 |
| CW 04 | 20. Jan | Lecture | 13 | Data Science in the Wild | 391-426 |
| CW 05 | 27. / 29. Jan | Lecture + Exercises | 14 | Q&A / Feedback | |
| CW 06 | 03. / 04. Feb | Oral Exams (Block 1) | | | |
| CW 13 | 24. / 25. Mar | Oral Exams (Block 2) | | | |

Preparation in our last session
("Oral Exam Briefing")

# Semester Schedule: Notes

When?

- Lecture <u>every Tuesday</u>
- Exercise session <u>every second Thursday</u>
- Always 14:15 to 15:45

**Note**: First Exercise Sheet out now and due next week!

Where?

- Default: Room A01 0-006
- Occasionally online (Big Blue Button)

# Participation

- We are trying to make this lecture as interactive as possible, so **please raise questions** as they come up.

- If there are technical issues (e.g. audio/video problems), please give us a hint.

- We will make recordings available after our lecture sessions.

# Hybrid Teaching

- We will be <u>trying</u> to enable online participation by streaming lectures and exercise sessions to Big Blue Button (BBB) where possible.

- However, we recommend attending the live sessions unless we have to move them online anyways.

# **Workload Estimation**

> 30 hours remain for working on the data science project!
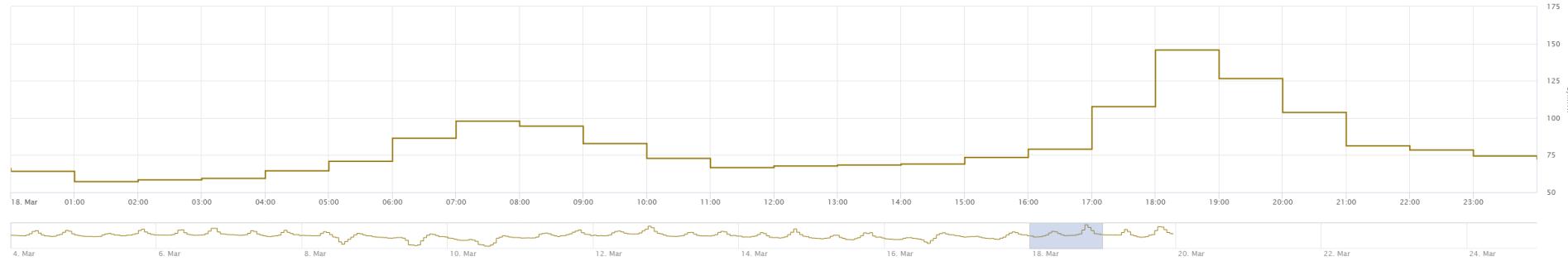
150 hours

- 6 credit points ≈ 150-180 work hours
  - 1 week of preparation before exam ≈ 40 hours
  - 14 lectures ≈ 14·90 min. ≈ 21 hours
  - 7 exercise sessions ≈ 7·90 min. ≈ 10.5 hours

- Remaining: ≥78.5 hours (>5.5 hours per week!)
  - If you spend 1.5 hours on reading / lecture notes, …
  - …, you still have 3 hours for exercises / tutorials …
  - … and 1 hour for catching up on Python or statistics

- Advice: Use spare time for data challenges!

# Data Science Project: Energy Prices

Analyze data and create a prediction model:

● Target: hourly day-ahead prices for January 27 (€ per MWh)

**Note**: That's 21:00 UTC!

● Deadline: January 25, 2026 at 22:00 CET

# Data Science Project: Report Structure

Project report resembles semester outline:

1. Gathering Domain Knowledge & Data Sources
2. Data Cleaning & Exploratory Data Analysis
3. Visualization & Storytelling
4. Predictive Modeling

Document your journey in a Jupyter notebook!

# **<u>Group</u> Exercises**

- Some students may have a substantial data science background already.
- Some students may not even have a background in Python.
- Please try to form exercise groups with complementary backgrounds.
- **<u>Please form groups of 3</u>**, so that grading submissions in time remains feasible 😉
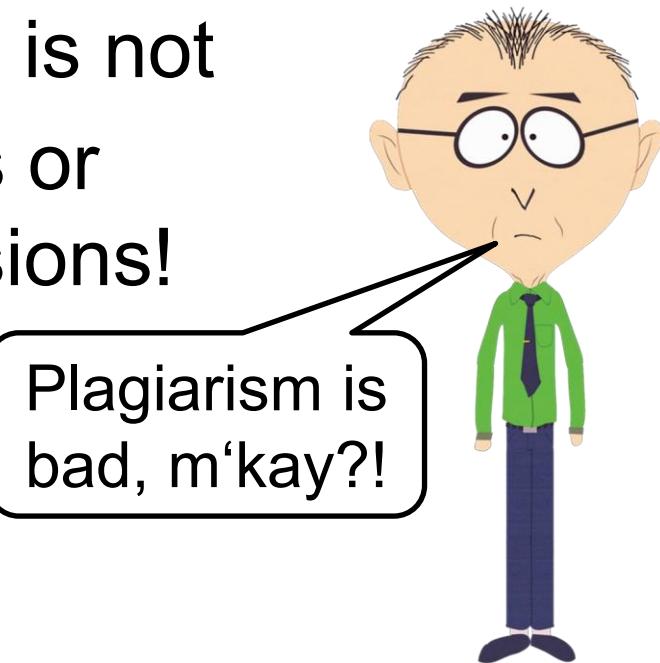
# **Exercise Submission**

1. Download from Stud.IP
   (PDF + Jupyter Notebook File)
2. Do exercises & compile <u>1</u> group submission
   (set up Jupyter environment in Exercise 1)
3. Upload your group submission to Stud.IP
   (there is a dedicated submission folder)

<u>Note</u>: The first session is Thursday next week!

# A Short Comment on Originality

- Working in groups is great
- Submitting the same solution is not
- Please always credit sources or collaboration in your submissions!
- Failure to do so can get you into <u>serious</u> trouble!

Plagiarism is bad, m'kay?!

# **Submission Feedback**

1. You will receive bonus points for every successful submission (no other feedback)


Lecture: 2.01.040 Data Science I
Overview   administration   Forum   Participants   Files   Schedule   Wiki   Blubber   Meetings   Cliqr   Examination dates   Bonus points   More …

2. If you have questions, ask in the exercise sessions or in the forum!

… But why bonus points at all?

# Academic Grading With Bumps!

- There will be an oral examination
- By doing well in practical sessions and project, you can earn up to 2 grading bumps (0.3 / 0.4):
  - 1 for *"doing well on the exercises"* in your group, i.e.:
    - achieving ≥50% points on 6 of 7 exercise sheets and
    - presenting at least 2 submitted solutions per group
  - 1 for achieving ≥50% points on the project report
- <u>Example</u>: With 1 grading bump, an oral exam graded with 2.0 leads to a final grade of 1.7!

# Oral Exam: Grading Scheme

- **1.0 (Excellent)**: The student demonstrates a deep understanding of the subject matter, can articulate complex concepts clearly, and can apply knowledge to new or hypothetical situations, showing evidence of critical thinking and original thought.
- **2.0 (Good)**: The student shows a solid understanding of the subject matter and can explain most concepts accurately. They can apply knowledge to practical situations and show some evidence of critical thinking.
- **3.0 (Satisfactory)**: The student understands the basic concepts of the subject matter, but may struggle with more complex ideas. They can apply knowledge to familiar situations, but may struggle with novel or hypothetical scenarios.
- **4.0 (Adequate)**: The student has a basic understanding of the subject matter and can articulate the main concepts, but struggles with applying knowledge to practical situations and shows limited evidence of critical thinking.
- **5.0 (Failed)**: The student demonstrates a lack of understanding of the subject matter, struggles to articulate even basic concepts, and cannot apply knowledge effectively. There is little to no evidence of critical thinking.

# Oral Exam: Basic Procedure

- **Intro & Outline**: welcome, procedure outline, brief pulse check
- **Examination (20 Minutes)**:
  - Questions to gauge how deep the student's understanding of the covered topics is, *starting with the student's topic of choice*
  - Discussion: we engage in an exchange on the students understanding, thoughts, opinions, and ideas in the context of a data science topic or application domain
  - Problem Solving: as a part of the discussion, the student may be asked to solve a specific problem or describe how they would approach it and why
- **Feedback**:
  - Internal discussion: we ask student to leave the room and discuss how the examination went
  - Immediate Feedback: we ask the student back in, explain how the exam went from our point of view (and whether they passed or failed)
  - <u>Note</u>: We do not give you the grade right away!
- **Results in Stud.IP**: we enter all grades (incl. bumps) when the exam block is over, typically by the end of the day or the day after your exam

# Instructor Style Disclaimer

- I try to make lectures fun through jokes and analogies, but always fear saying something that may offend someone in the class.
- I want everyone around me to feel comfortable.
- If anything I say bothers you, please come by and tell me so. I will apologize, and then do my best to understand the issue to avoid doing so again.

# Let's Have a Great Semester!
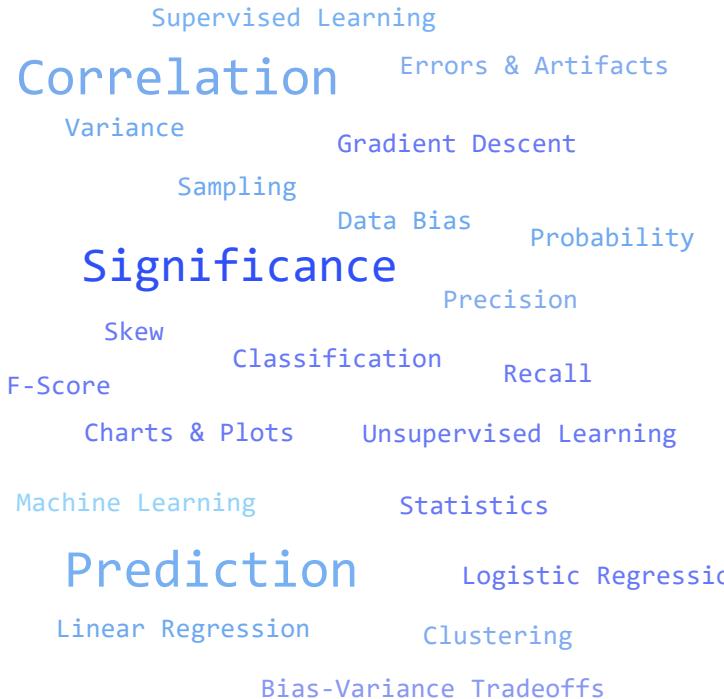
We are looking forward to teaching and learning with you!

Supervised Learning

Correlation        Errors & Artifacts

Variance

Gradient Descent

Sampling

Data Bias
Probability

Significance

Precision

Skew

Classification        Recall

F-Score

Charts & Plots        Unsupervised Learning

Machine Learning        Statistics

Prediction        Logistic Regression

Linear Regression        Clustering

Bias-Variance Tradeoffs

Data Science 1:
Introduction to Data Science

Course Overview &

# Intro to Data Science

Winter 2025

## Wolfram Wingerath, Jannik Schröder

Department for Computing Science
Data Science / Information Systems

# What is Data Science?

Like any emerging field, it isn't yet well defined, but incorporates elements of:
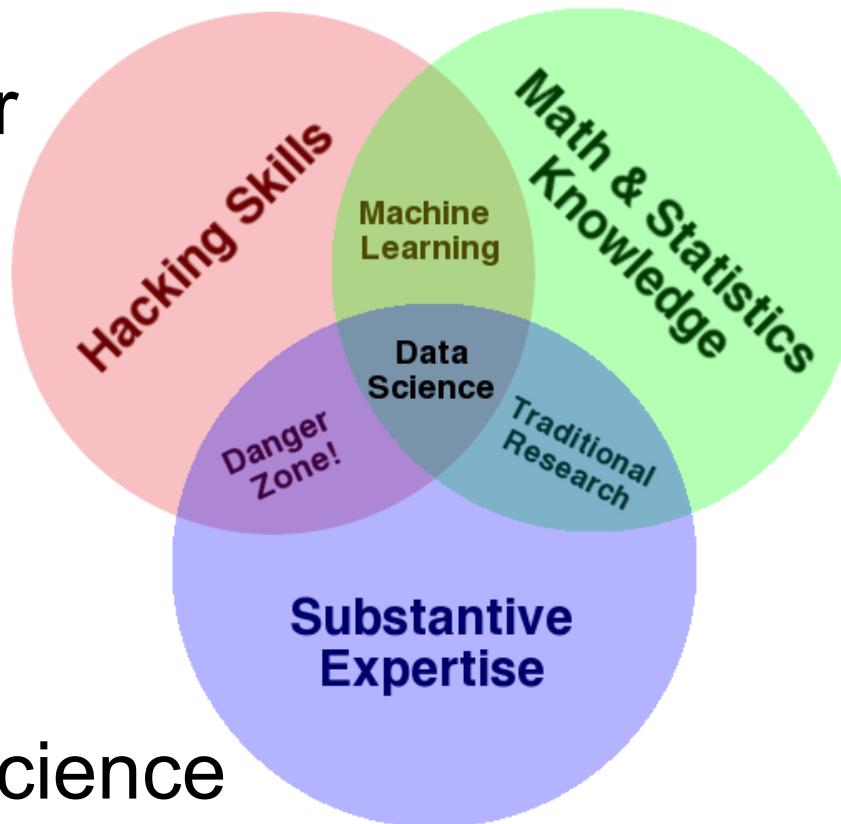
- Exploratory Data Analysis and Visualization
- Machine Learning and Statistics
- High-Performance Computing technologies for dealing with scale.

# Skill Sets for Data Science



Computer Science

Statistics

Domain Science

# **Appreciating Data**

Computer Scientists do not naturally appreciate data: it's just stuff to run through a program.

The usual way to test algorithm performance is to run the implementation on "random data".

But interesting data sets are a scarce resource, which requires hard work and imagination to obtain.

# Computer vs. Real Scientists (1)

- Scientists strive to understand the complicated and <u>messy natural world</u>, while computer scientists build their own clean and organized virtual worlds. Thus:
- Nothing is ever completely true or false in science, while everything is either true or false in Computer Science / Mathematics.

# Computer vs. Real Scientists (2)

- Scientists are data-driven, while computer scientists are algorithm-driven.
- Scientists obsess about discovering things, which computer scientists invent rather than discover.
- Scientists are comfortable with the idea that data has errors; computer scientists are not.

# Genius vs. Wisdom

Software developers are hired to produce code.

Data Scientists are hired to produce insights.

Genius shows in finding the right answer!!!

Wisdom shows in avoiding the wrong answers.

Data science (like most things) benefits more from wisdom than from genius.

# Developing Wisdom

- Wisdom comes from experience.
- Wisdom comes from general knowledge.
- Wisdom comes from listening to others.
- Wisdom comes from humility, observing how often you have been wrong and why/how.

I'm still seeking wisdom myself, but I try to get across how difficult it is to make good predictions.

# Developing Curiosity

- The good data scientist develops a curiosity about the domain/application they are working in.
- They talk shop with the people whose data they are working on.
- They read the newspaper every day, to get a broader perspective on the world.

# **Asking Good Questions**

Software developers are not encouraged to ask questions, but data scientists are:

- What exciting things might you be able to learn from a given data set?
- What things do you/your people really want to know?
- What data sets might get you there?

# Let's Practice Asking Questions!

Who, What, Where, When, and Why on the following datasets:

- Baseball-reference.com
- International Movie Database (IMBb)
- Google ngrams
- NYC taxi cab records
- Speedrun.com

# Baseball-Reference.com: biosketch

# Statistical Record of Play

Summary statistics of each years batting, pitching, and fielding record, with teams and awards.

# Baseball Questions

- How to best measure individual player's skill, value or performance?
- How fair do trades between teams work out?
- What is the trajectory of player's performances as they mature and age?
- To what extent does batting/pitching/fielding performance correlate with the position played?

# Demographic Questions

- Do left-handed people have shorter lifespans than right-handers?
- How often do people return to where they were born?
- Do player salaries reflect past, present, or future performance?
- Are heights and weights increasing in the population?

# IMDb: Movie Data

# IMDb: Actor Data



**James Stewart** (I) (1908–1997)

Top 5000

Actor | Soundtrack | Director

James Maitland Stewart was born on 20 May 1908 in Indiana, Pennsylvania, where his father owned a hardware store. He was educated at a local prep school, Mercersburg Academy, where he was a keen athlete (football and track), musician (singing and accordion playing), and sometime actor. In 1929 he won a place at Princeton, where he studied ... See full bio »

**Born:** James Maitland Stewart
May 20, 1908 in Indiana, Pennsylvania, USA

**Died:** July 2, 1997 (age 89) in Los Angeles, California, USA

230 photos | 42 videos | 1180 news articles »

**Won 1 Oscar.** Another 25 wins & 19 nominations. See more awards »

**Cast**                                                    Edit

Cast overview, first billed only:

| | James Stewart | ... | George Bailey |
| | Donna Reed | ... | Mary Hatch |
| | Lionel Barrymore | ... | Mr. Potter |
| | Thomas Mitchell | ... | Uncle Billy |
| | Henry Travers | ... | Clarence |
| | Beulah Bondi | ... | Mrs. Bailey |
| | Frank Faylen | ... | Ernie |
| | Ward Bond | ... | Bert |
| | Gloria Grahame | ... | Violet |
| | H.B. Warner | ... | Mr. Gower |

# Movie Questions
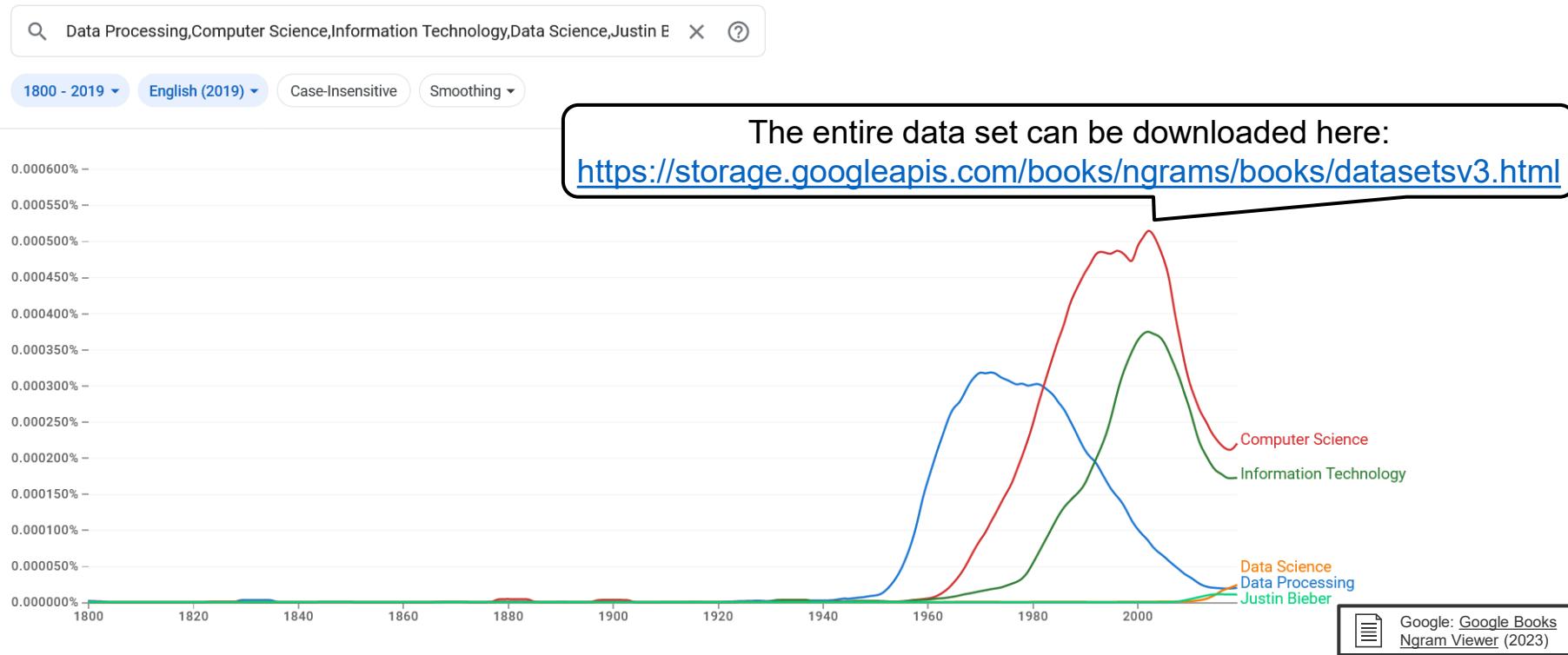
- Can we predict how well people will like a movie?  What about its gross?
- What does the social network of actors look like? (Six degrees of Kevin Bacon)
- What is the age distribution of actors and actresses in film?
- Do stars live longer or shorter lives than the bit players or public?

# Google Ngrams

- Presents an annual time series of the frequency of every "popular" word/phrase with 1 to 5 words occurs in scanned books.
- "popular" means appears >40 times in total.
- Google has scanned about 15% of all books ever published, making this resource quite comprehensive.

# Google Ngram Viewer

**Google** Books Ngram Viewer

🔍 Data Processing,Computer Science,Information Technology,Data Science,Justin B   ✕   ?

[1800 - 2019 ▾]   [English (2019) ▾]   Case-Insensitive   Smoothing ▾

The entire data set can be downloaded here:
https://storage.googleapis.com/books/ngrams/books/datasetsv3.html

Computer Science

Information Technology

Data Science
Data Processing
Justin Bieber

# Ngram Questions

- How has the amount of cursing changed over time?
- What is the lifespan of fame and technologies? Is it increasing/decreasing?
- How often do new words emerge?  Do they stay in common usage?
- What words are associated with other words, i.e. can you build a language model?

# NYC Taxi Cab Data

- Gives driver/owner, pickup/dropoff location, and fare data for every taxi trip taken.
- Data obtained from NYC via Freedom of Information Act Request (FOA)
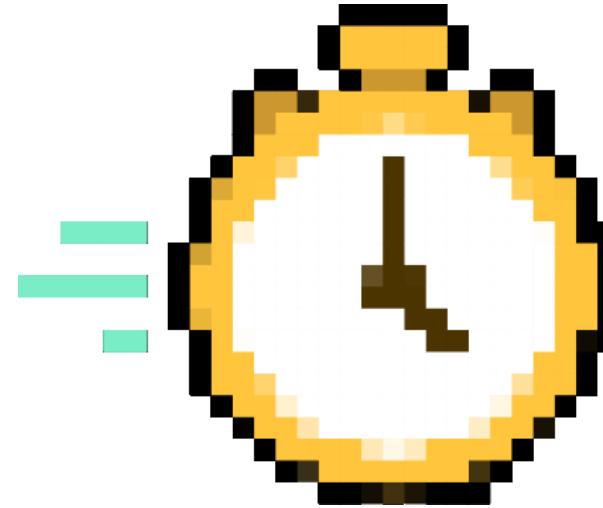
| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | | | | | | | | | | | | |
| 5 | Trip data, 2013 -> | | | | | | | | | | | |
| 6 | | | | | | | | | | | | |
| 7 | medallion | hack_license | vendor_id | rate_code | pickup_datetime | dropoff_datetim | passenger_c | trip_time_ | trip_distance | pickup_longitud | pickup_latitud | dropoff_longitude | dropoff_latitude |
| 8 | 89D227B655E5C82AEC | BA96DE419E7116! | CMT | 1 | 1/1/13 15:11 | 1/1/13 15:18 | 4 | 382 | 1 | -73.978165 | 40.757977 | -73.989838 | 40.751171 |
| 9 | 0BD7C8F5BA12B88E0B | 9FD8F69F0804BD! | CMT | 1 | 1/6/13 0:18 | 1/6/13 0:22 | 1 | 259 | 1.5 | -74.006683 | 40.731781 | -73.994499 | 40.75066 |
| 10 | 0BD7C8F5BA12B88E0B | 9FD8F69F0804BD! | CMT | 1 | 1/5/13 18:49 | 1/5/13 18:54 | 1 | 282 | 1.1 | -74.004707 | 40.73777 | -74.009834 | 40.726002 |
| 11 | ... | | | | | | | | | | | | |
| 12 | | | | | | | | | | | | |
| 13 | | | | | | | | | | | | |
| 14 | Fare data, 2013 -> | | | | | | | | | | | |
| 15 | | | | | | | | | | | | |
| 16 | medallion | hack_license | vendor_id | pickup_datetime | fare_amount | surcharge | mta_tax | tip_amou | tolls_amount | total_amount | | | |
| 17 | 89D227B655E5C82AEC | BA96DE419E7116! | CMT | 1/1/13 15:11 | 6.5 | 0 | 0.5 | 0 | 0 | 7 | | | |
| 18 | 0BD7C8F5BA12B88E0B | 9FD8F69F0804BD! | CMT | 1/6/13 0:18 | 6 | 0.5 | 0.5 | 0 | 0 | 7 | | | |
| 19 | 0BD7C8F5BA12B88E0B | 9FD8F69F0804BD! | CMT | 1/5/13 18:49 | 5.5 | 1 | 0.5 | 0 | 0 | 7 | | | |

# Taxicab Questions

- How much do drivers make each night?
- How far do they travel?
- How much slower is traffic during rush hour?
- Where are people traveling to/from at different times of the day?
- Do faster drivers get tipped better?
- Where should drivers go to pick up their next fare?

# Speedrun.com: Video Game Data

Leaderboards, statistics, forums, and other resources on nerds playing video games as fast as they can.

SPEEDRUN.COM

# Speed Run Questions
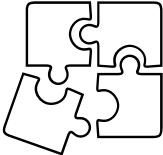
- What is the most popular game for speed running? Does this change often over time?
- What is the longest time that a record in a given game has remained unbeaten?
- Does player age correlate with success in speed running?
- Can we predict how fast people will be able to finish a game before its release?

# Properties of Data

- Structured vs. Unstructured
  - table vs. written text
- Quantitative vs. Categorical
  - measurements vs. labels
- Little vs. Big
  - Lecture evaluation vs. presidential election

# Classification & Regression

- Classification: assign a label
  - Who is going to win the election?
  - What genres does a movie belong to?
- Regression: forecast a quantitative value
  - How many people will vote for a given candidate?
  - How much box office the will a movie make on the first weekend?

# Wrapup: Intro to Data Science

- Data science is the study of extracting knowledge and insights from data.
- Data science includes skills from statistics, machine learning, and computer science.
- Collaboration and communication skills are essential for success in data science.
- Data science is constantly evolving.

# So What Now?

1. Please form groups of 3 (see Stud.IP)
2. Download & work on Exercise Sheet 1
   (feel free to solve the tasks on your own, but
   please agree on <u>1 submission per group</u>)
3. Submit group solution until before the deadline
4. Have a nice week!

See you next week! 😉