

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

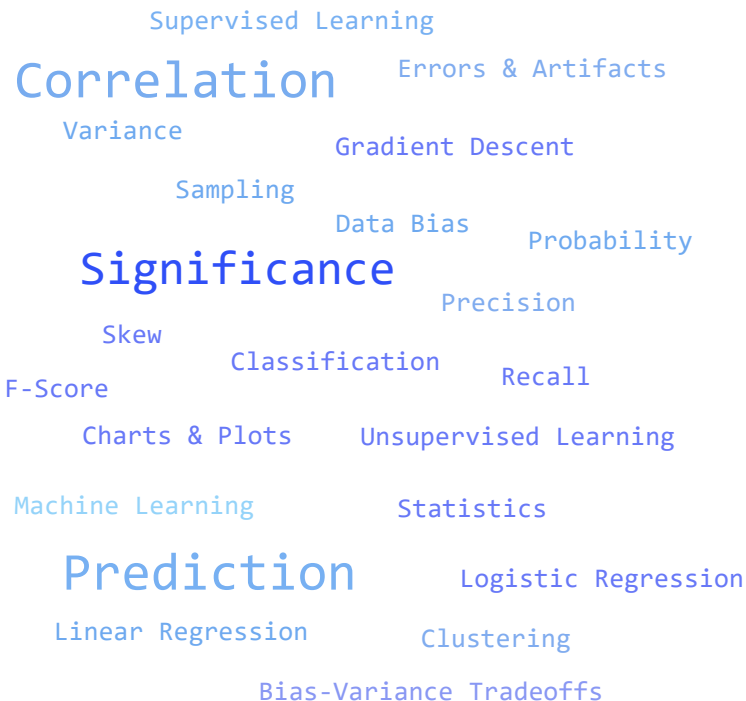
---

**Wolfram Wingerath, Jannik Schröder**

Department for Computing Science  
Data Science / Information Systems

Lecture slides based on content from "The Data Science Design Manual" (Steven Skiena, 2017) and associated course materials generously made available online by the author at <https://www3.cs.stonybrook.edu/~skiena/data-manual/>.

Special thanks to Professor Skiena for sharing these valuable teaching resources!



## 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



**Baqend**

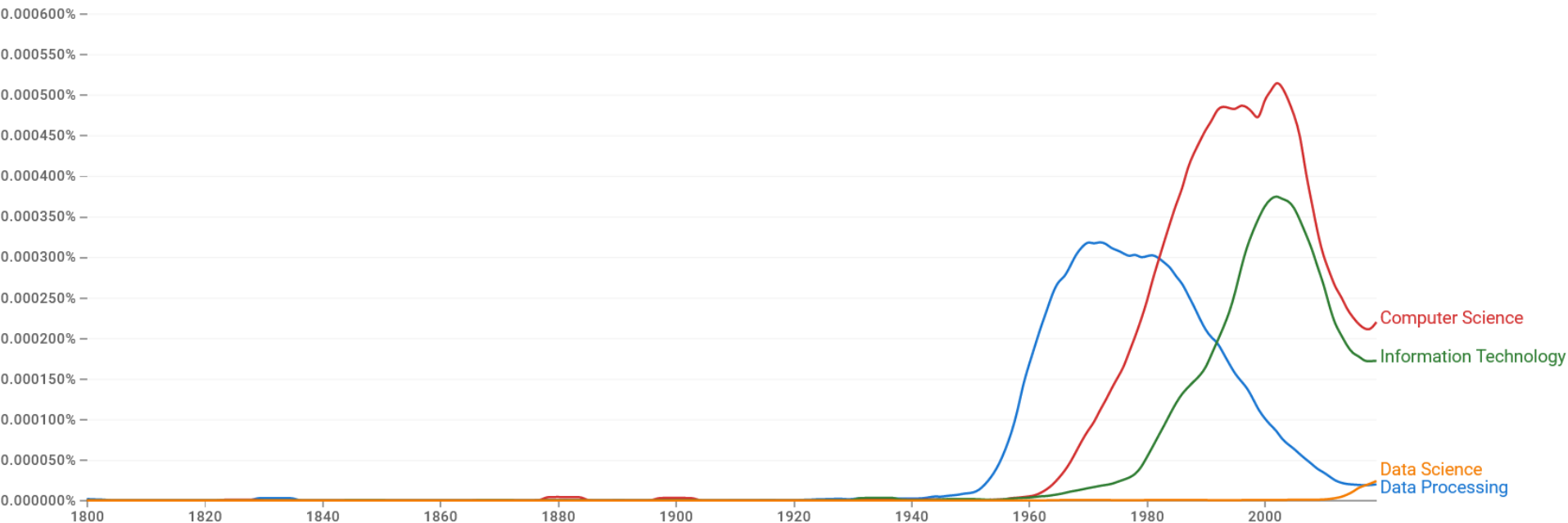


# Meet Team Data: Jannik

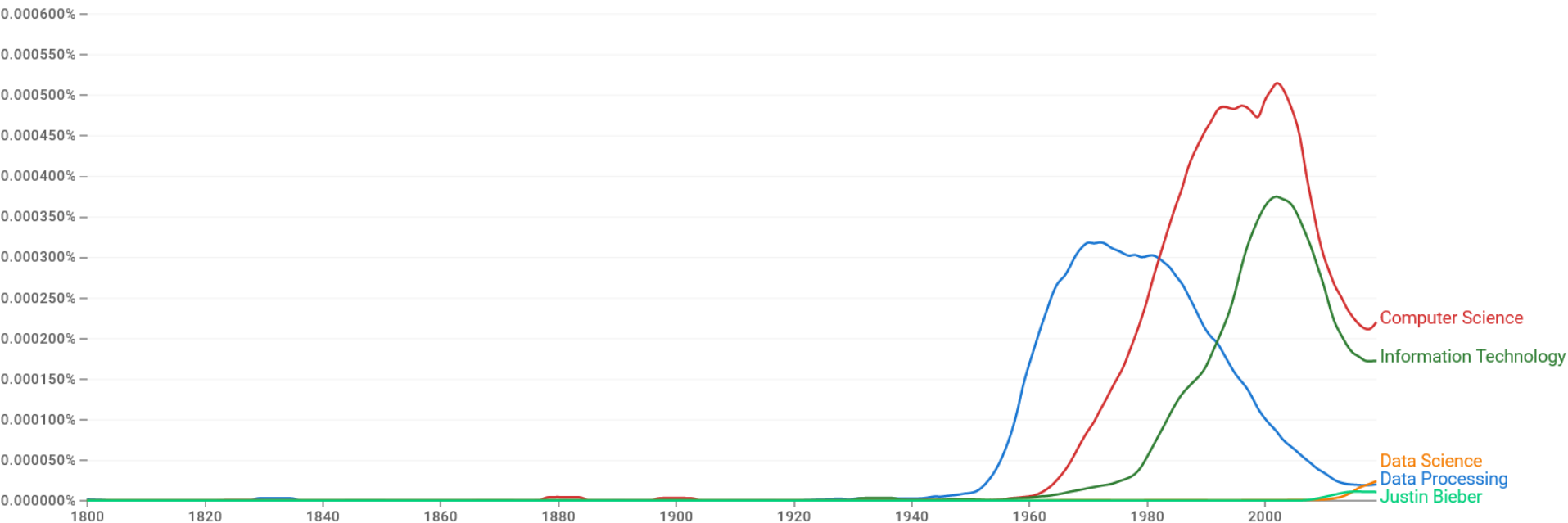
---

- 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...



# Data is not new to computing...



# ... 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 not 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

---

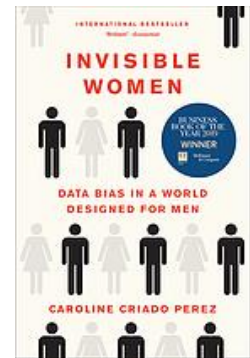
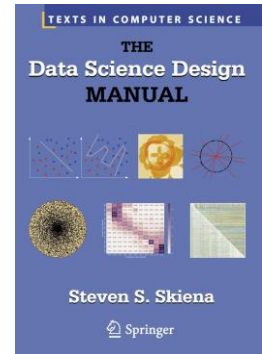
- Core database topics at DB group in Hamburg
- Duplicate detection in probabilistic databases in bachelor's and master's thesis
- PhD on real-time databases (NoSQL + stream processing)
- Large-scale data analytics in web contexts
- Head of data engineering at Baqend (startup)

I have done data engineering 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	<u>Guest Lecture</u>	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)	} Preparation in our last session („Oral Exam Briefing“)		
CW 13	24. / 25. Mar	Oral Exams (Block 2)			

# Semester Schedule: Notes

---

When?

- Lecture every Tuesday
- Exercise session every second Thursday
- 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 trying 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!

- 6 credit points  $\approx$  150-180 work hours
  - 1 week of preparation before exam  $\approx$  40 hours
  - 14 lectures  $\approx$  14·90 min.  $\approx$  21 hours
  - 7 exercise sessions  $\approx$  7·90 min.  $\approx$  10.5 hours
- Remaining:  $\geq 78.5$  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!

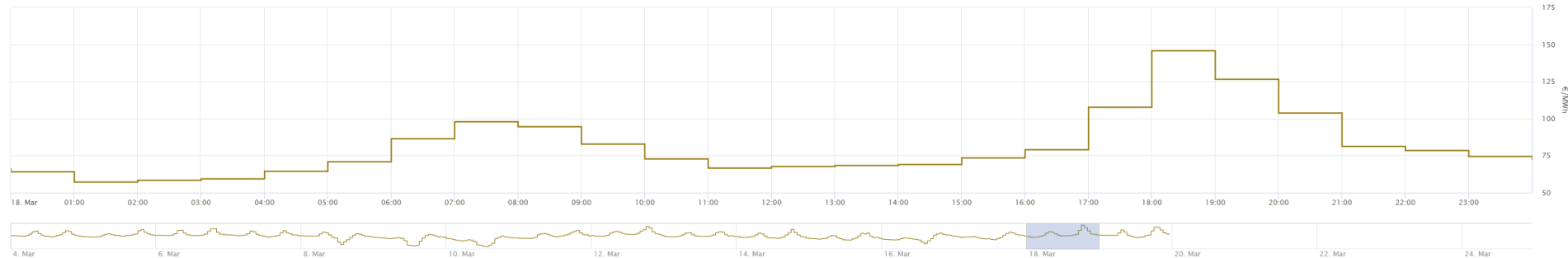
150 hours

# Data Science Project: Energy Prices

Analyze data and create a prediction model:

- Target: hourly day-ahead prices for January 27 (€ per MWh)
- Deadline: January 25, 2026 at 22:00 CET

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



# 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!

# Group 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.
- **Please form groups of 3**, so that grading submissions in time remains feasible 😊

# Exercise Submission

---

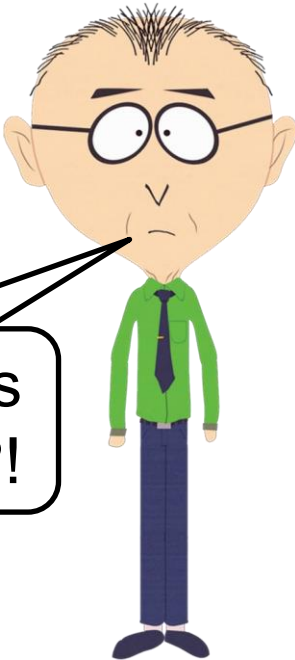
1. Download from Stud.IP  
(PDF + Jupyter Notebook File)
2. Do exercises & compile 1 group submission  
(set up Jupyter environment in Exercise 1)
3. Upload your group submission to Stud.IP  
(there is a dedicated submission folder)

Note: 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 serious trouble!

A cartoon character of a man with glasses, a green shirt, and a blue tie, looking slightly concerned. A speech bubble points to him from the text 'Plagiarism is bad, m'kay?!'.

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  $\geq 50\%$  points on 6 of 7 exercise sheets and
    - presenting at least 2 submitted solutions per group
  - 1 for achieving  $\geq 50\%$  points on the project report
- Example: 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)
  - Note: 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

---

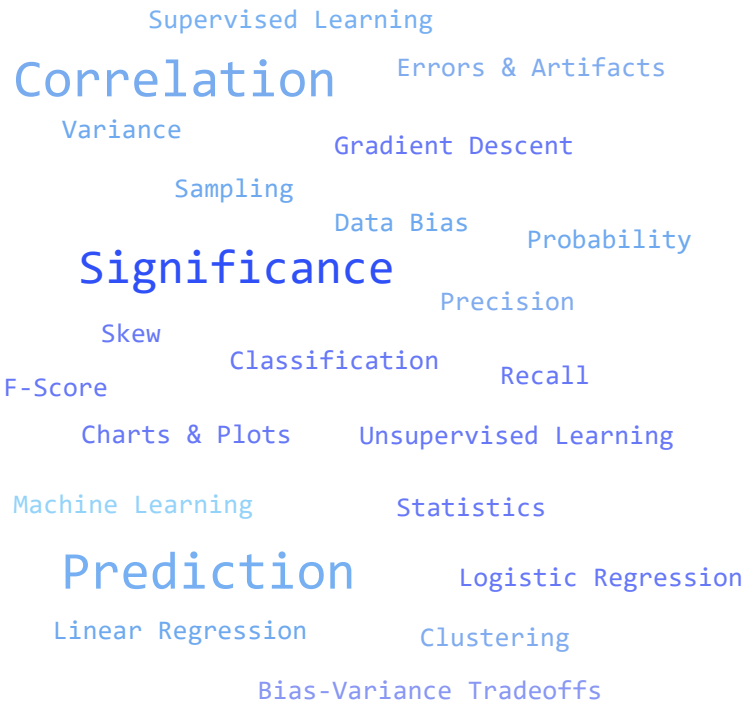
- I try to make lectures fun through jokes and analogies, but always fear saying something that may offend someone in the class.
- I want everyone around me to feel comfortable.
- If anything I say bothers you, please come by and tell me so. I will apologize, and then do my best to understand the issue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

# Let's Have a Great Semester!

---

We are looking forward to teaching and learning with you!





## 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

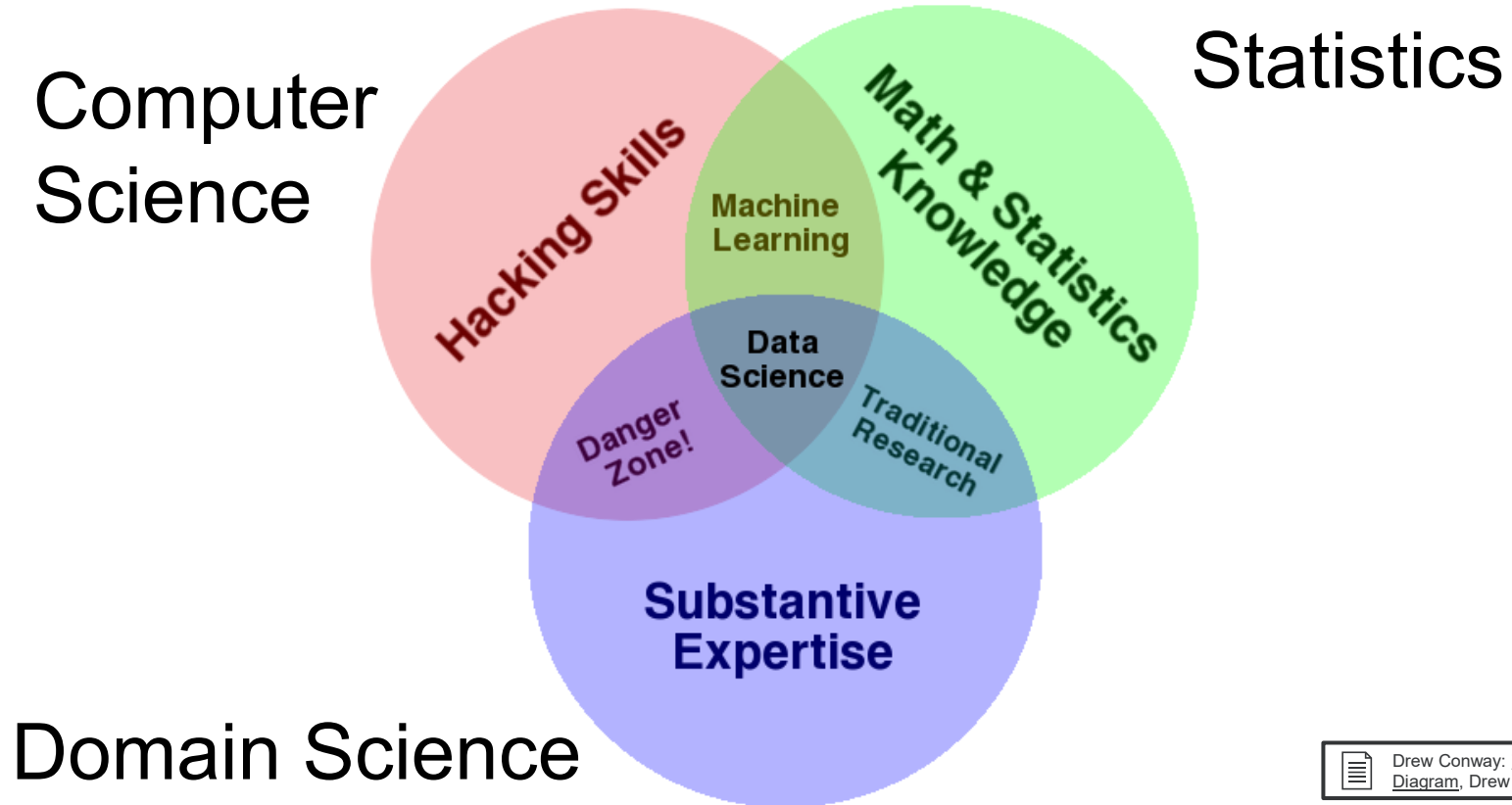
# What is Data Science?

---

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



# 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 messy natural world, 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](http://Baseball-reference.com)
- International Movie Database (IMDb)
- Google ngrams
- NYC taxi cab records
- [Speedrun.com](http://Speedrun.com)

# Baseball-Reference.com: biosketch



play index **players** teams seasons managers leaders awards postseason boxes japan nlb minors draft

Mobile Site You Are Here > Home > Encyclopedia of Players > R Listing > Babe Ruth Statistics and

News: s-r blog:KBO Stats back to 1999 - Baseball-Reference.com

Babe Ruth Player Page » Batting Pitching Fielding Minors News Archive (1456) Bullpen Oracle



## Babe Ruth

Like 1,213 people like this.

+25 Recommend this

George Herman Ruth (Babe, The Bambino or The Sultan Of Swat)

**Positions:** Outfielder and Pitcher

**Bats:** Left, **Throws:** Left

**Height:** 6' 2", **Weight:** 215 lb.

**Born:** February 6, 1895 in Baltimore, MD

**High School:** St. Mary's HS (Baltimore, MD) (All Transactions)

**Debut:** July 11, 1914 (Age 19.155)

**Rookie Status:** Exceeded rookie limits during 1915 season [\*]

**Teams** (by GP): Yankees/RedSox/Braves 1914-1935

**Final Game:** May 30, 1935 (Age 40.113)

**Inducted** into the Hall of Fame by BBWAA as Player in 1936 (215/226 ballots). Induction ceremony in View [Babe Ruth Page](#) at the Baseball Hall of Fame (plaque, photos, videos).

**Died:** August 16, 1948 in New York, NY (Aged 53.192)

**Buried:** Gate of Heaven Cemetery, Hawthorne, NY

**View Player Bio** from the [SABR BioProject](#)

[About biographical information](#)



S-R: M

## Transactions

**July 9, 1914:** Purchased with [Ernie Shore](#) and [Ben Egan](#) by the [Boston Red Sox](#) from Baltimore (International) for more than \$25000. more than \$25000

**December 26, 1919:** Purchased by the [New York Yankees](#) from the [Boston Red Sox](#) for \$100,000.

**February 26, 1935:** Released by the [New York Yankees](#).

**February 26, 1935:** Signed as a Free Agent with the [Boston Braves](#).

The transaction information used here was obtained free of charge from and is copyrighted by [RetroSheet](#). We attempt to update transactions throughout the season.

## Salaries

Convert to YYYY \$'s Salaries may not be complete (especially pre-1985) and may not include some earned bonuses

Year	Age	Team	Salary	ServTm(OnDay)	Sources	Notes/Other Sources
1914	19	Boston Red Sox	\$2,500		? Bill James Historical Abstract	Annualized rate; came up late in season
1915	20	Boston Red Sox	\$3,500		? Bill James Historical Abstract	
1916	21	Boston Red Sox	\$3,500		? Contract at HOF	
1917	22	Boston Red Sox	\$3,500		? Contract at HOF	BJHA: \$5,000; Baseball Timeline \$7,000
1918	23	Boston Red Sox	\$9,000		? Allan Wood, 1918, at 183	Includes \$1,000 midseason raise, \$1,000 WS' bonus
1919	24	New York Yankees	\$10,000*		? Michael Haupt research of HOF contracts	Contract at HOF: 10000.00,
1920	25	New York Yankees	\$20,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 20000.00,
1921	26	New York Yankees	\$20,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 30000.00, Plus \$5K for '20 and '21 exhibitions, \$50/HR (59)m
1922	27	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 52000.00,
1923	28	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 52000.00,
1924	29	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 52000.00,
1925	30	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 52000.00,
1926	31	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 52000.00,
1927	32	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	5/23/27 AL letter: 70000.00,
1928	33	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	5/23/27 AL letter: 70000.00,
1929	34	New York Yankees	\$52,000*		? Michael Haupt research of HOF contracts	5/23/27 AL letter: 70000.00,
1930	35	New York Yankees	\$70,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 80000.00,
1931	36	New York Yankees	\$70,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 80000.00,
1932	37	New York Yankees	\$70,000*		? Michael Haupt research of HOF contracts	M. Smelser, Life That Ruth Built, p. 441: 75000.00, Plus 25% of all exhibition-game profits
1933	38	New York Yankees	\$80,000*		? Michael Haupt research of HOF contracts	M. Smelser, Life That Ruth Built, p. 456: 52000.00, Plus 25% of revenue from in-season exhibitions
1934	39	New York Yankees	\$80,000*		? Michael Haupt research of HOF contracts	1/16/36 TSN, per government report: 36696.00, \$35,000 salary plus 25% of exhibition profits
1935	40	New York Yankees	\$75,000*		? Michael Haupt research of HOF contracts	Bill James Historical Abstract: 35000.00, Annualized rate; retired early in season
1936	41	New York Yankees	\$52,000		? Michael Haupt research of HOF contracts	
1937	42	New York Yankees	\$35,000		? Michael Haupt research of HOF contracts	
Career to date (may be incomplete)			\$1,020,000			



# Statistical Record of Play

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

Babe Ruth Player Page

BattingPitchingFieldingMinorsNews Archive (1456)BullpenOracle

Fan EloRater

[Fine Details](#) · Last updated Jun 3, 2014 9:17AM

All-Time Rank (among batters): #1. BABE RUTH... #2. Lou Gehrig... #3. Ted Williams... #4. Honus Wagner...

Vote

Standard BattingMore Stats

Glossary · Show Minors Stats · SHARE · Embed · CSV · PRE · [LINK](#) · ?

Minors	Game Logs	Splits	HR Log	Finders																												
Year	Age	Tm	Lg	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos	Awards			
1914	19	BOS	AL	5	10	10	1	2	1	0	0	0	0	0	0	4	.200	.200	.300	.500	49	3	0	0				/1				
1915	20	BOS	AL	42	103	92	16	29	10	1	4	20	0	0	9	23	.315	.376	.576	.952	188	53	0	2				1				
1916	21	BOS	AL	67	152	136	18	37	5	3	3	16	0		10	23	.272	.322	.419	.741	121	57	0	4				1				
1917	22	BOS	AL	52	142	123	14	40	6	3	2	14	0		12	18	.325	.385	.472	.857	162	58	0	7				1				
1918	23	BOS	AL	95	382	317	50	95	26	11	11	61	6		58	58	.300	.411	.555	.966	192	176	2	3				07138				
1919	24	BOS	AL	130	543	432	103	139	34	12	29	113	7		101	58	.322	.456	.657	1.114	217	284	6	3				*071/38				
1920	25	NYY	AL	142	616	458	158	172	36	9	54	135	14	14	150	80	.376	.532	.847	1.379	255	388	3	5				*0978/31				
1921	26	NYY	AL	152	693	540	177	204	44	16	59	168	17	13	145	81	.378	.512	.846	1.359	238	457	4	4				*078/31				
1922	27	NYY	AL	110	496	406	94	128	24	8	35	96	2	5	84	80	.315	.434	.672	1.106	182	273	1	4				*079/3				
1923	28	NYY	AL	152	697	522	151	205	45	13	41	130	17	21	170	93	.393	.545	.764	1.309	239	399	4	3				*097/83	MVP-1			
1924	29	NYY	AL	153	681	529	143	200	39	7	46	124	9	13	142	81	.378	.513	.739	1.252	220	391	4	6				*097/8				
1925	30	NYY	AL	98	426	359	61	104	12	2	25	67	2	4	59	68	.290	.393	.543	.936	137	195	2	6				097				
1926	31	NYY	AL	152	652	495	139	184	30	5	47	153	11	9	144	76	.372	.516	.737	1.253	225	365	3	10				*079/3				
1927	32	NYY	AL	151	691	540	158	192	29	8	60	165	7	6	137	89	.356	.486	.772	1.258	225	417	0	14				*097				
1928	33	NYY	AL	154	684	536	163	173	29	8	54	146	4	5	137	87	.323	.463	.709	1.172	206	380	3	8				*097				
1929	34	NYY	AL	135	587	499	121	172	26	6	46	154	5	3	72	60	.345	.430	.697	1.128	193	348	3	13				*097				
1930	35	NYY	AL	145	676	518	150	186	28	9	49	153	10	10	136	61	.359	.493	.732	1.225	211	379	1	21				*097/1				
1931	36	NYY	AL	145	663	534	149	199	31	3	46	162	5	4	128	51	.373	.495	.700	1.195	218	374	1	0				*097/3	MVP-5			
1932	37	NYY	AL	133	589	457	120	156	13	5	41	137	2	2	130	62	.341	.489	.661	1.150	201	302	2	0				*097/3	MVP-6			
1933	38	NYY	AL	137	576	459	97	138	21	3	34	104	4	5	114	90	.301	.442	.582	1.023	176	267	2	0				*097/31	AS			
1934	39	NYY	AL	125	471	365	78	105	17	4	22	84	1	3	104	63	.288	.448	.537	.985	160	196	2	0				*097	AS			
1935	40	BSN	NL	28	92	72	13	13	0	0	6	12	0		20	24	.181	.359	.431	.789	119	31	2	0	0				07/9			
22 Yrs				2503	10622	8399	2174	2873	506	136	714	2214	123	117	2062	1330	.342	.474	.690	1.164	206	5793	2	43	113							
162 Game Avg.				162	687	544	141	186	33	9	46	143	8	133	86	.342	.474	.690	1.164	206	375	3	7									
				G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos	Awards			
NYY (15 yrs)				2084	9198	7217	1959	2518	424	106	659	1978	110	117	1852	1122	.349	.484	.711	1.195	209	5131	35	94								
BOS (6 yrs)				391	1332	1110	202	342	82	30	49	224	13	0	190	184	.308	.413	.568	.981	190	631	8	19								
BSN (1 yr)				28	92	72	13	13	0	0	6	12	0		20	24	.181	.359	.431	.789	119	31	2	0	0							
AL (21 yrs)				2475	10530	8327	2161	2860	506	136	708	2202	123	117	2042	1306	.343	.475	.692	1.167	207	5762	43	113								
NL (1 yr)				28	92	72	13	13	0	0	6	12	0		20	24	.181	.359	.431	.789	119	31	2	0	0							

# 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



All

[Movies, TV & Showtimes](#)
[Celebs, Events & Photos](#)
[News & Community](#)
[Watchlist](#)



More at IMDbPro »

## It's a Wonderful Life (1946)

Approved 130 min - Drama | Family | Fantasy - 7 January 1947 (USA)

**Top 5000**

**Your rating:** ★★★★★★★★ -/10  
Ratings: **8.7**/10 from 202,743 users  
Reviews: 632 user | 187 critic

An angel helps a compassionate but despairingly frustrated businessman by showing what life would have been like if he never existed.

**Director:** Frank Capra  
**Writers:** Frances Goodrich (screenplay), Albert Hackett (screenplay), 4 more credits »  
**Stars:** James Stewart, Donna Reed, Lionel Barrymore | See full cast and crew »

[+ Watchlist](#) [Watch Trailer](#) [Share...](#)

## Details

**Country:** USA  
**Language:** English  
**Release Date:** 7 January 1947 (USA) [See more »](#)  
**Also Known As:** The Greatest Gift [See more »](#)  
**Filming Locations:** California, USA [See more »](#)

## Box Office

**Budget:** \$3,180,000 (estimated)  
**Opening Weekend:** £49,845 (UK) (19 December 2008)  
**Gross:** £682,222 (UK) (24 December 2010)  
[See more »](#)

## Company Credits

**Production Co:** Liberty Films (II) [See more »](#)  
[Show detailed company contact information on IMDbPro »](#)

## Technical Specs

**Runtime:** 130 min | 118 min (DVD edition)  
**Sound Mix:** Mono (RCA Sound System)  
**Color:** Color (colorized) | Black and White  
**Aspect Ratio:** 1.37 : 1  
[See full technical specs »](#)

# IMDb: Actor Data



**James Stewart** (I) (1908–1997)

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

Q Data Processing,Computer Science,Information Technology,Data Science,Justin E X ?

1800 - 2019

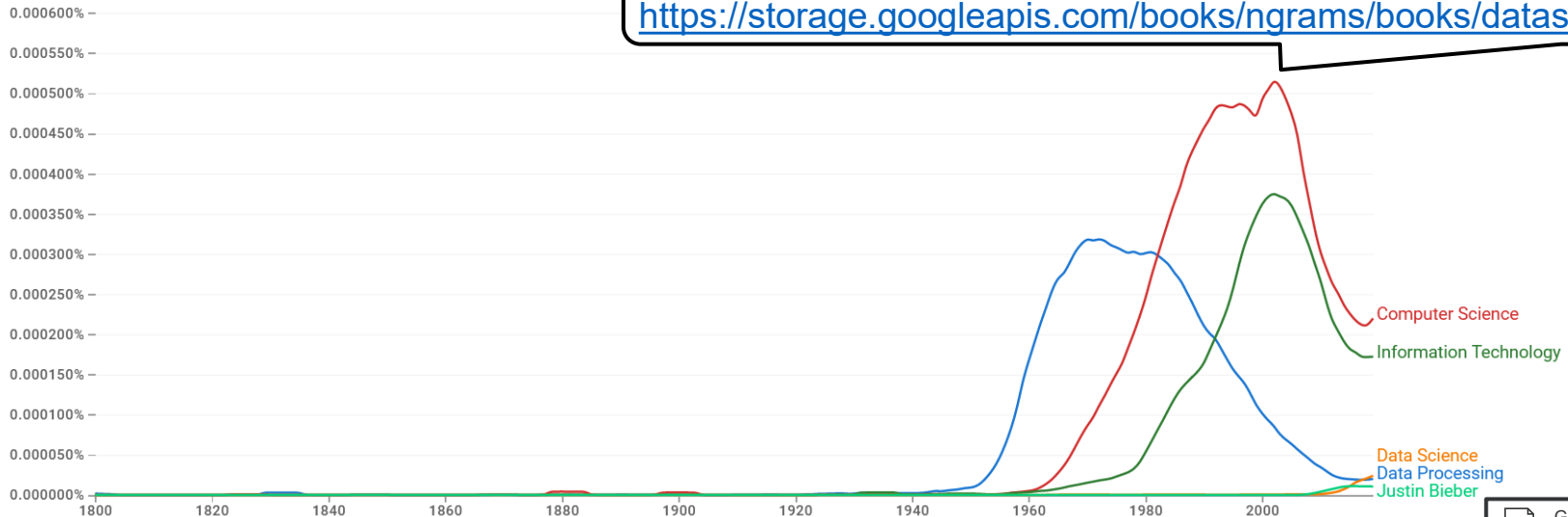
English (2019)

Case-Insensitive

Smoothing

The entire data set can be downloaded here:

<https://storage.googleapis.com/books/ngrams/books/datasetv3.html>





# 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_datetime	passenger_count	trip_time	trip_distance	pickup_longitude	pickup_latitude	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	9FD8F69F08048D	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	9FD8F69F08048D	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_amount	tolls_amount	total_amount			
17	89D227B655E5C82AEC	BA96DE419E7116	CMT	1/1/13 15:11	6.5	0	0.5	0	0	7			
18	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1/6/13 0:18	6	0.5	0.5	0	0	7			
19	0BD7C8F5BA12B88E0B	9FD8F69F08048D	CMT	1/5/13 18:49	5.5	1	0.5	0	0	7			

# Taxicab Questions

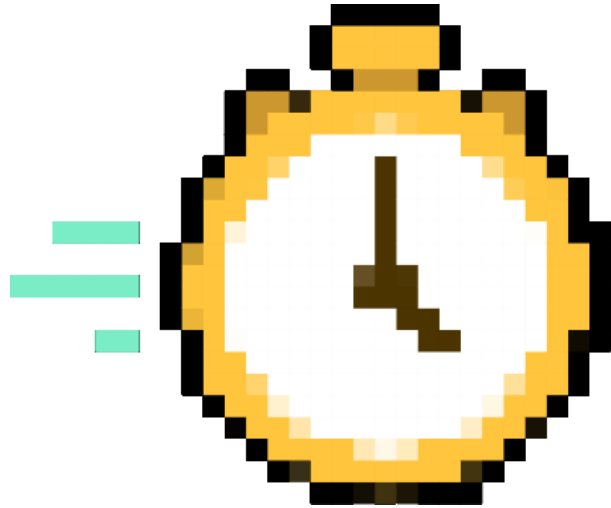
---

- 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

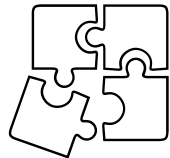
---

- 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 1 submission per group)
3. Submit group solution until before the deadline
4. Have a nice week!

See you next week! 😊