

Natural Language Processing, Data Science, and Large Language Models

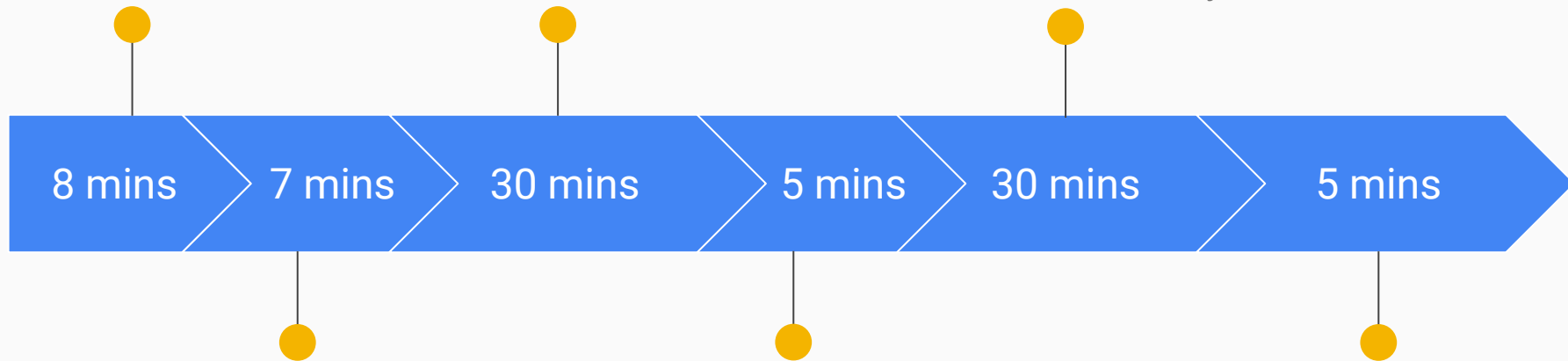
September 5, 2023



Syllabus changes &
Questions

History of Natural
Language Processing
(academic background)

The emergence of
Data Science and
NLP's role in industry



Project idea

Break

Conclusion

Problem set 1: Exploratory data analysis/dataset analysis (due Sept 21)

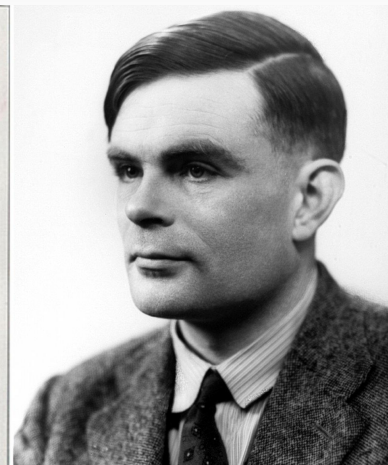
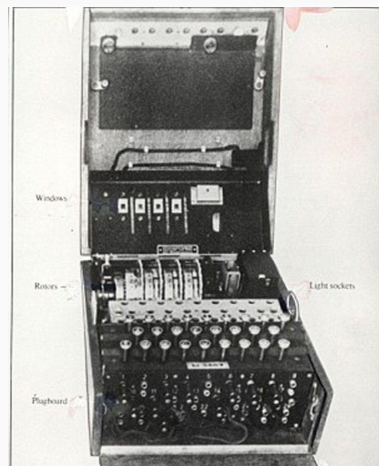
Lecture objective

Contextualize Natural Language Processing within the field of Data Science by understanding its importance and role in academia, industry, and culture at large

History of Natural Language Processing (academic background)

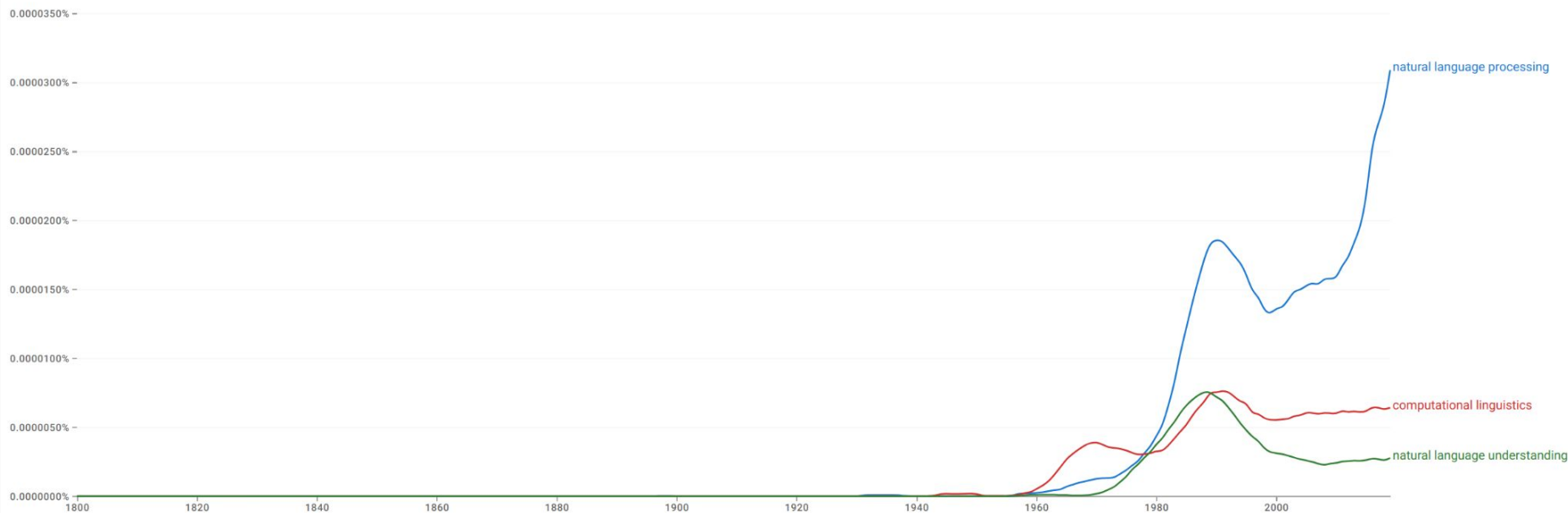
Natural Language Processing

The 'Turing Test'



Alan Turing
23 June 1912 – 7 June 1954

Natural Language Processing



https://books.google.com/ngrams/graph?content=natural+language+processing%2Ccomputational+linguistics%2Cnatural+language+understanding&year_start=1800&year_end=2019&corpus=en-2019&smoothing=3

Natural Language Processing



Please write me a sonnet on the subject of the Forth Bridge.



The Forth Bridge stands proud and strong, a feat
Of engineering, a symbol of man's might.
Its iron arms stretch out to join the street
Of water below, a bridge that sees no night.
A symbol of Scotland's grace and might,
It stands as a reminder of great skill
And those who toiled, through storm and wind and light,
To build a bridge to cross this tidal rill.
Its towers rise, a testament to time,
Its paint forever red, a blaze of hue.
It bridges not just water, but our climb
From past to future, with each step anew.
So let us hail the Forth, a bridge so grand,
A symbol of man's power, with a steadfast hand.

OpenAI: ChatGPT

<https://chat.openai.com/>



Write a funny joke about natural language processing and artificial intelligence possessing understanding



Why did the AI cross the road? To get to the other dataset!



give me another one



Why did the NLP model go to the therapist? Because it was having trouble parsing its feelings!

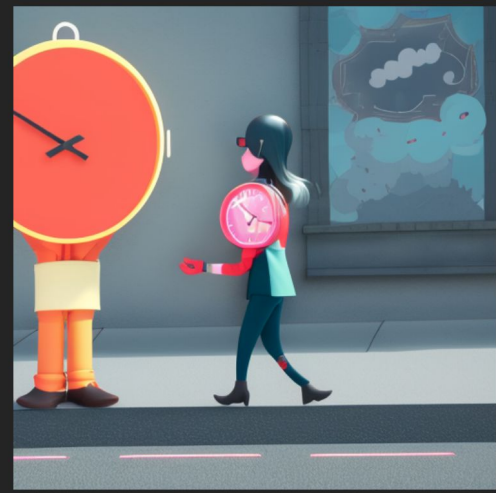
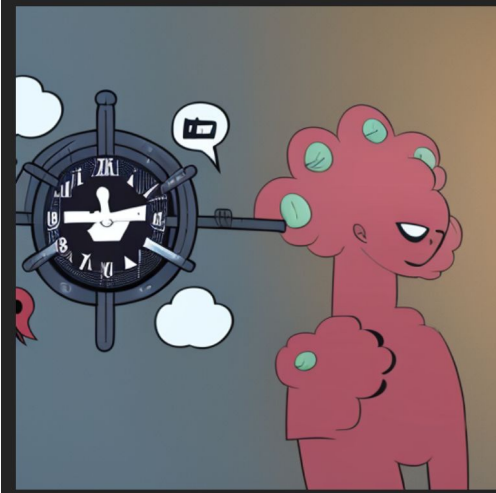
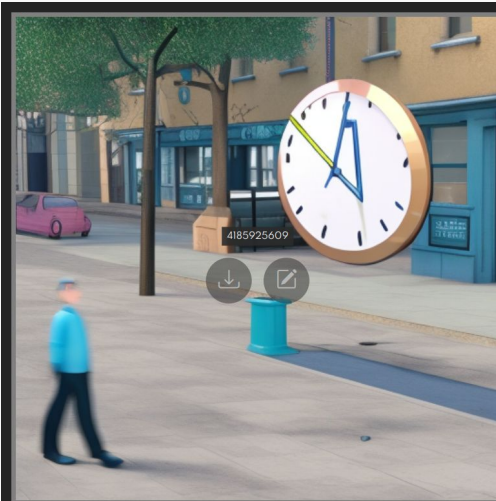


How did we get here?

Brief history of natural language processing

Stable diffusion

Prompt: 'a humanized clock walking down a street with a thought bubble over its head'



Evolution of computers and science

- Engineering breakthroughs (computer chips)
- Scientific developments and revolutions (Kuhn, *The Structure of Scientific Revolutions*)
- Sociological and cultural developments (the impact of tech on daily life)
- The need to process information and gain insight
- Statistical developments and evidence based reasoning
- Emergence of new jobs and academic disciplines
- Human curiosity and ingenuity
- Problems ...

Welcome to

EEEEEE	LL	IIII	ZZZZZZ	AAAAA
EE	LL	II	ZZ	AA AA
EEEEEE	LL	II	ZZZ	AAAAAAA
EE	LL	II	ZZ	AA AA
EEEEEE	LLLLLL	IIII	ZZZZZZ	AA AA

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.

This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?

YOU: Men are all alike.

ELIZA: What is the connection, do you suppose ?

YOU: They're always bugging us about something or other.

ELIZA: Can you think of a specific example ?

YOU: Well, my boyfriend made me come here.

ELIZA: Is it important to you that your boyfriend made you come here ?

YOU: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

YOU: It's true. I am unhappy.

ELIZA: Can you explain what made you unhappy ?

YOU:

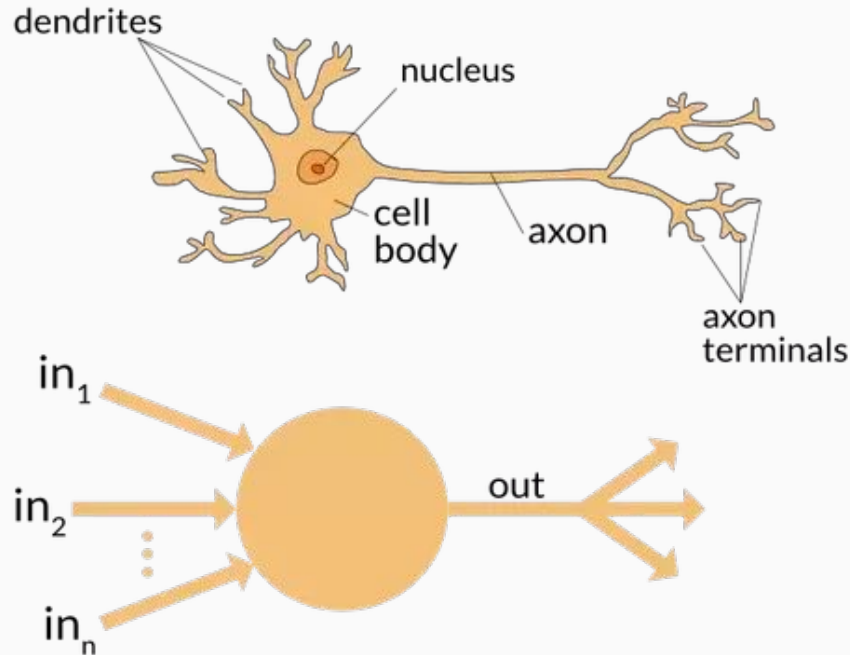
1940s - 1950s

- Automaton
- Probabilistic models
- Information theory
- Linguistics (Chomsky's generative syntax)

1940s - 1950s

- Automaton (McCulloch-Pitts neuron)
- Probabilistic models (Shannon, Tukey, and others)
- Information theory (Shannon and others)
- Linguistics (Chomsky's generative syntax)

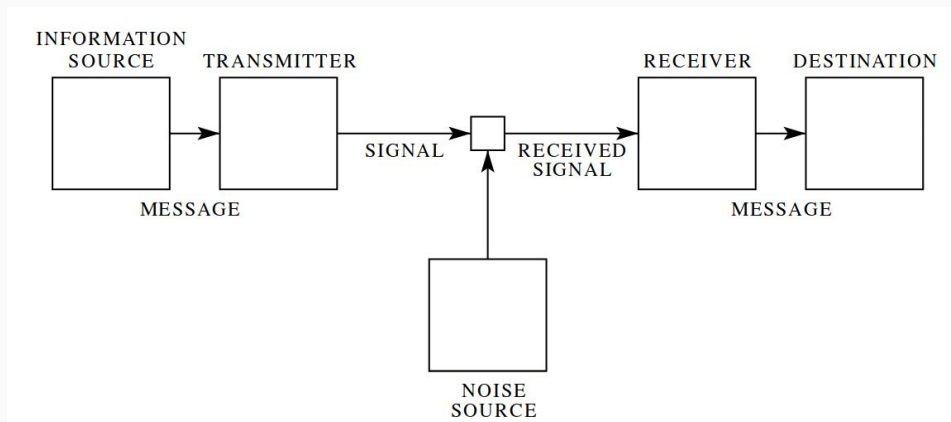
1940s - 1950s



- McCulloch-Pitts neuron - 1943
- Frank Rosenblatt (introduced the ability to train a neuron to recognize shapes)
- Marvin Minsky and Seymour Papert *Perceptrons* (MIT Press)

1940s - 1950s

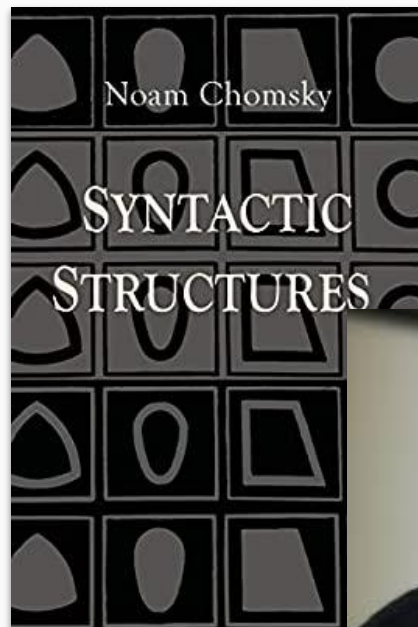
- Claude Shannon published “Mathematical Theory of Communication”
- Shannon borrowed the idea of entropy from physics to measure transmission of data and content of a language.



$$H(x,y) = -\sum_{i,j} p(i,j) \log p(i,j)$$

1940s - 1950s

- Noam Chomsky published *Syntactic Structures*
- Conceptualized grammar as a finite state machine
- Described syntax as a finite-state language and thus contributed to formalized theories of natural language



1960 - 1970 Two Camps

Two camps

Formal Language Theory (Symbolic)

- Frege, Russell, early Wittgenstein, etc.
- Symbolic

Artificial Intelligence (Stochastic)

- Marvin Minsky, Claude Shannon, Nathaniel Rochester
- probabilistic

1970 - 1993 Empiricism and Finite-State

- Stochastic paradigm - Hidden Markov Models
- Logic-based - Definite clause grammars; Lexical functional grammars; predicate logic and semantic representation
- Natural language understanding (NLU) - Winograd SHRDLU accepted natural language commands (first attempt to build an extensive English grammar)
- Thomas J. Watson of IBM - probabilistic models to predict parts of speech and word-sense ambiguities
- Data-driven methods with increasing compute power and labelled datasets

1970 - 1999 Empiricism and Finite-State

- Stochastic paradigm - Hidden Markov Models
- Logic-based - Definite clause grammars; Lexical functional grammars; predicate logic and semantic representation
- Natural language understanding (NLU) - Winograd SHRDLU accepted natural language commands (first attempt to build an extensive English grammar)
- Thomas J. Watson of IBM - probabilistic models to predict parts of speech and word-sense ambiguities
- Data-driven methods with increasing compute power and labelled datasets

2000 - 2023 Artificial Intelligence

- Linguistic Data Consortium (LDC) - large amounts of spoken and written data with linguistic annotations. Penn Tree Bank (1993), Prague Dependency Treebanks (1998), PropBank (2005), Penn Discourse Treebank (2004).
- Semantic analysis is cast as problem of supervised machine learning
- Support vector machines, Entropy techniques, Bayesian models, causation defined computational linguistics
- Data warehouses and increased computing power
- Vector semantics, unsupervised learning, and topic modelling
- Industry needs (medical, legal, financial knowledge domains)

The emergence of Data Science and NLP's role in industry

Where are we going?

Data science and NLP

Stable diffusion

Prompt: 'a futuristic city with people wearing various technology like watches or glasses and the sky is made of data like ones and zeros'



Data Science and NLP

Data Scientist: The Sexiest Job of the 21st Century

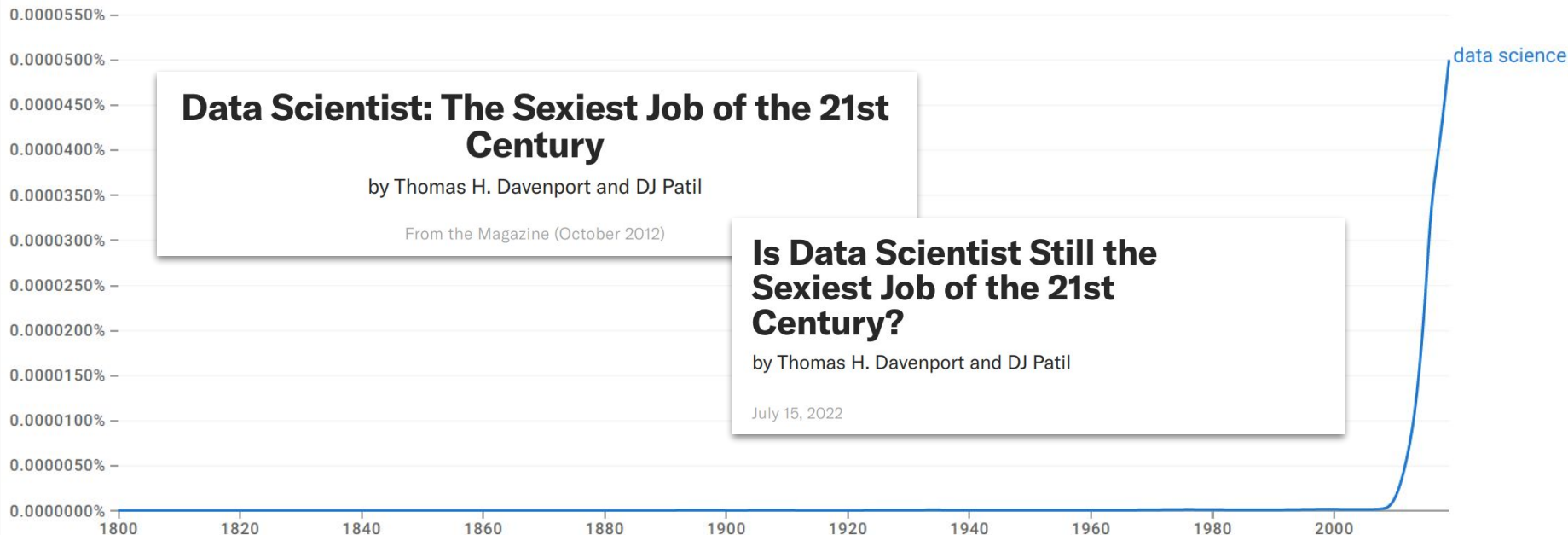
by Thomas H. Davenport and DJ Patil

From the Magazine (October 2012)

Is Data Scientist Still the Sexiest Job of the 21st Century?

by Thomas H. Davenport and DJ Patil

July 15, 2022



Jurafsky and Martin's second edition

“The goal of this new field is to get computers to perform useful tasks involving human language, tasks like enabling human-machine communication, improving human-human communication, or simply doing useful processing of text or speech.” (p. 1)



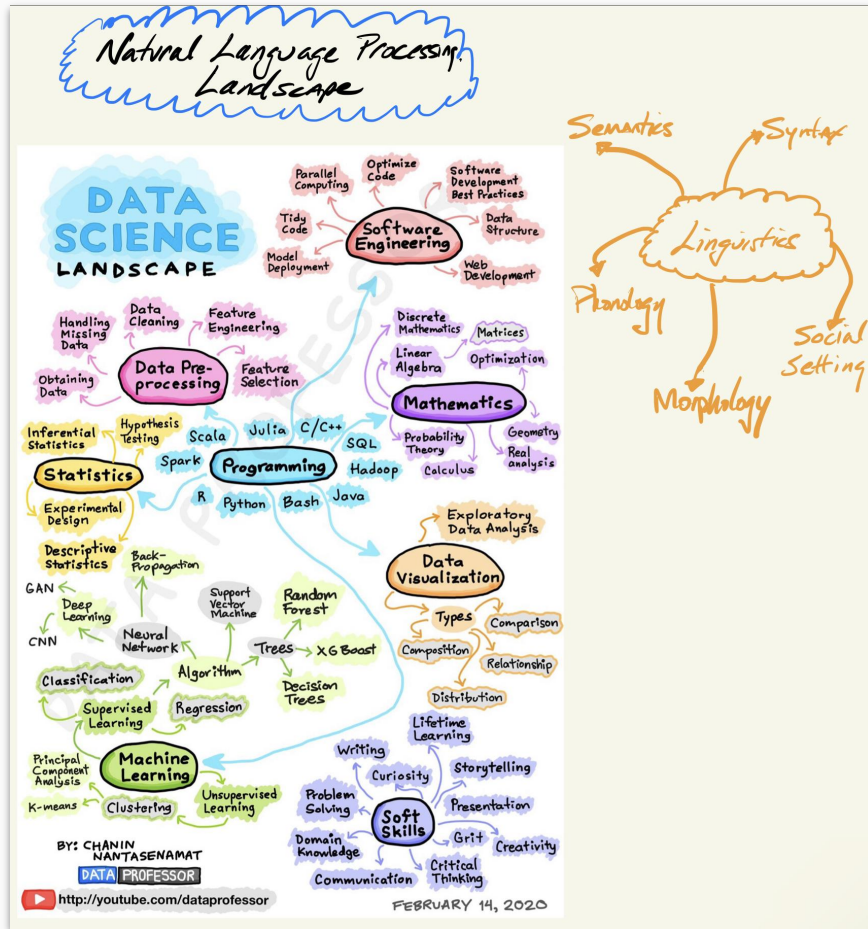


Data Professor

@DataProfessor

145K subscribers





Data Science & NLP

“Good data scientists have wide-ranging interests. They read the newspaper everyday to get a broader perspective on what is exciting. They understand that the world is an interesting place. Knowing a little of something about everything equips them to play in other people’s backyards. They are brave enough to get out of their comfort zones a bit, and driven to learn more once they get there.” Skiena, *The Data Science Design Manual*, 4.



