



# DATA 1301

# Introduction to Data Science

Student-Faculty connection day  
The Pillars of Science

Amir Shahmoradi

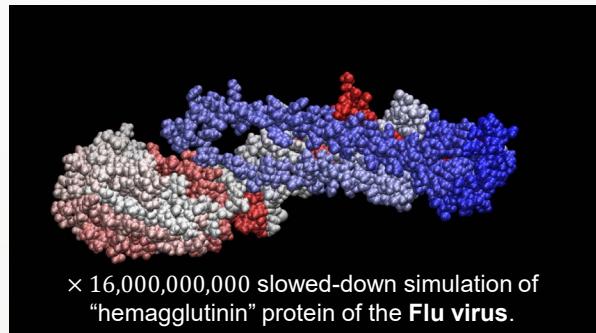
Department of Physics / College of Science  
Data Science Program / College of Science  
The University of Texas  
Arlington, Texas

Join us @ [cdslab.org](http://cdslab.org)

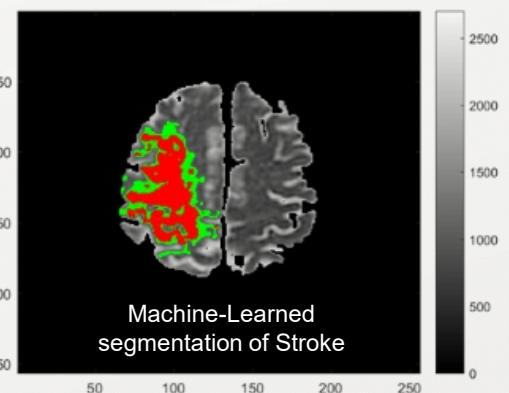
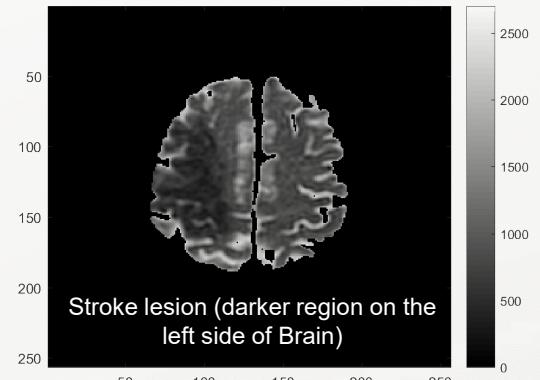
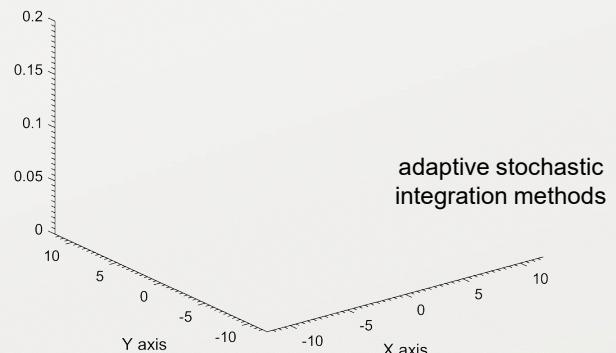
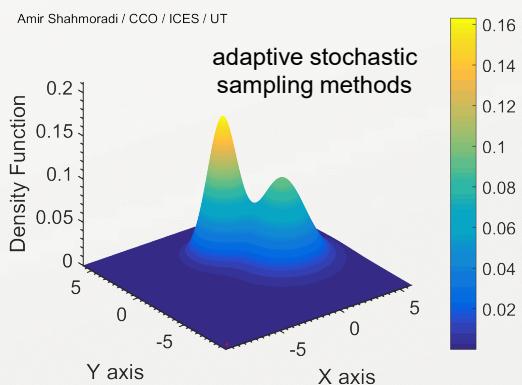


Physics of Gamma-Ray Bursts  
The most powerful explosions in the universe

Open-source software development:  
Machine Learning and Monte Carlo Methods



## Bioinformatics / Biophysics



## Biomedical Data Science

## Traffic Engineering



# Physicists' contributions to the foundations of Data Science

## Particle Physicists

Pioneers in high-throughput Big Data and the Internet technology



**WHERE THE  
WEB  
WAS BORN**

In the office of Tim Berners-Lee, all the fundamental technologies of the World Wide Web were developed.  
Started in 1990 from a proposal made by Tim Berners-Lee in 1989, the effort was first divided between an office in Building 11 of the Computing and Information Services (CIS) and an office in building 11 of the Electronics and Computing for Physics (ECP).  
In 1991 the teams came together in those offices, then belonging to ECP. It was composed of two CERN staff members, Tim Berners-Lee (CERN) and Robert Cailliau (then a member of Philips, Technical Research, a Computer and Semiconductors division).  
At the end of 1991 Tim Berners-Lee left CERN to direct the WWW Consortium at the Massachusetts Institute of Technology (MIT) in Cambridge, Massachusetts, USA. He was succeeded at CERN by Robert Cailliau.  
In 1995 Tim Berners-Lee and Robert Cailliau received the ACM Software System award for their work on the World Wide Web. In 2004 they received the first Millennium Technology Prize by the Finnish Technology Award Foundation.

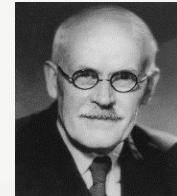
The CERN Library  
June 2004

## Physicists / Astronomers

The founding fathers and strong advocates of **Bayesian Inference** and **stochastic optimization / integration methods**



**Pierre Laplace**  
(1749 – 1827)  
[Astronomer / Mathematician](#)



**Harold Jeffreys**  
(1891 – 1989)  
[Astronomer / Geophysicist](#)



**Richard Cox**  
(1898 – 1991)  
[Physicist](#)



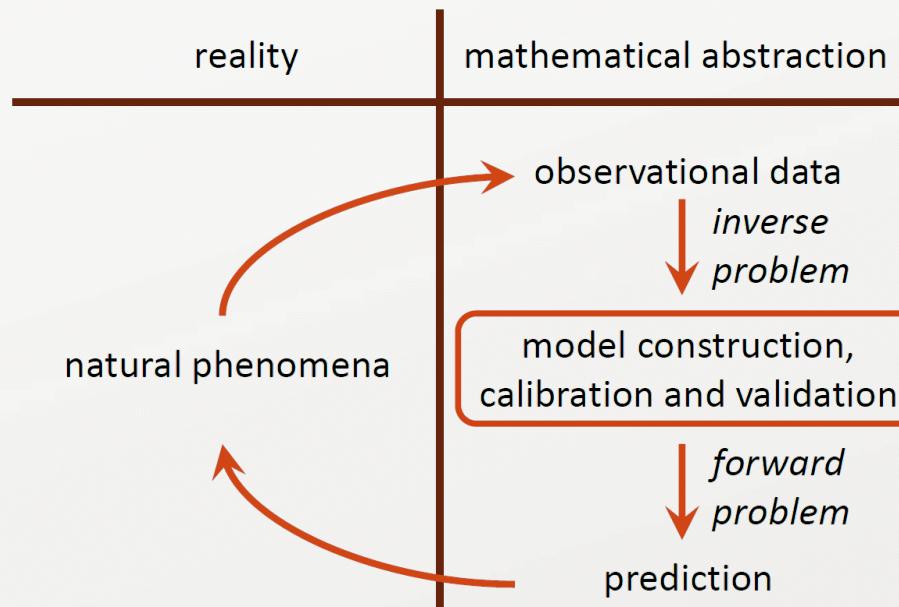
**Edwin Jaynes**  
(1922 – 1998)  
[Physicist](#)

Prior knowledge has a fundamental role in inference along with Data (via the Bayes' Rule)  
e.g., hyperparameter tuning

# The two classical pillars of science: Experiment and Theory

## How do we make a scientific inference?

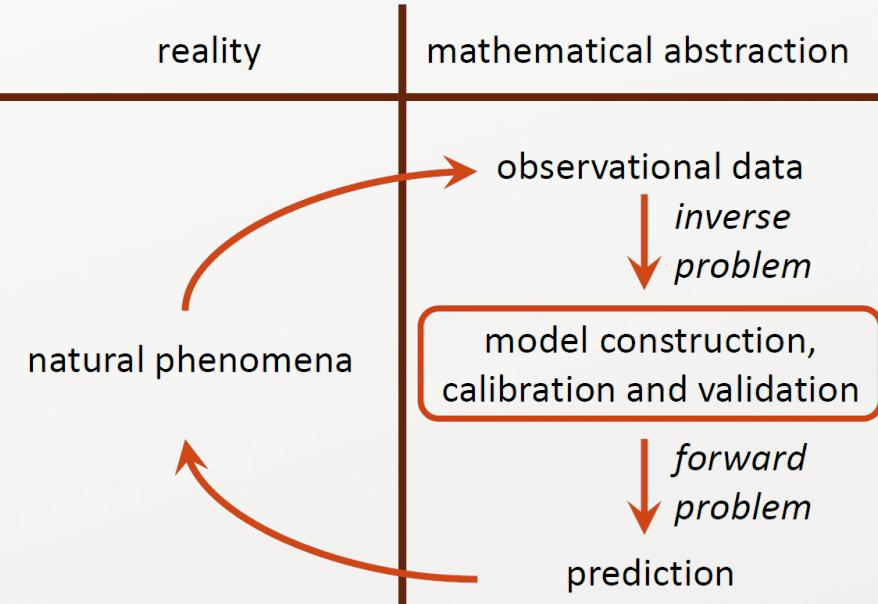
A very elementary depiction of the scientific method



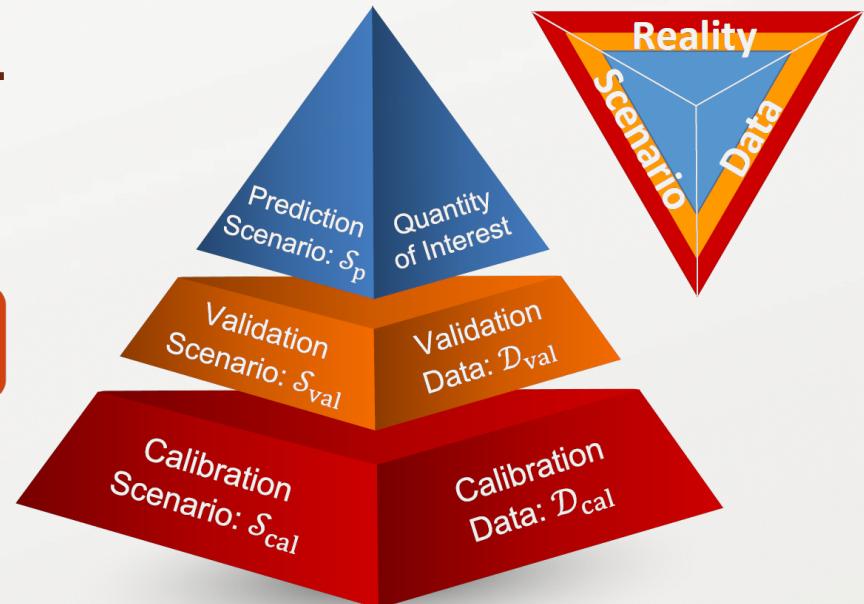
# The two classical pillars of science: Experiment and Theory

How do we make a scientific inference?

A very elementary depiction of the scientific method



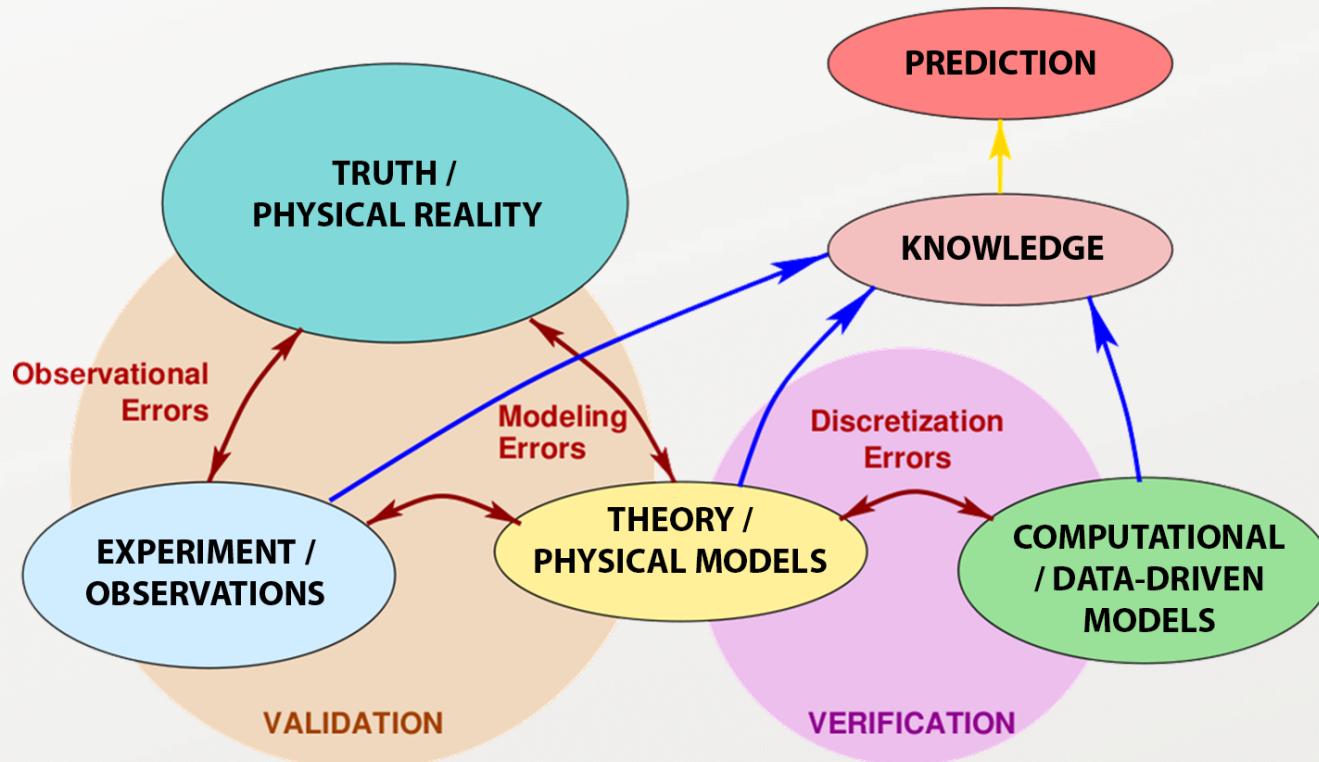
The prediction pyramid



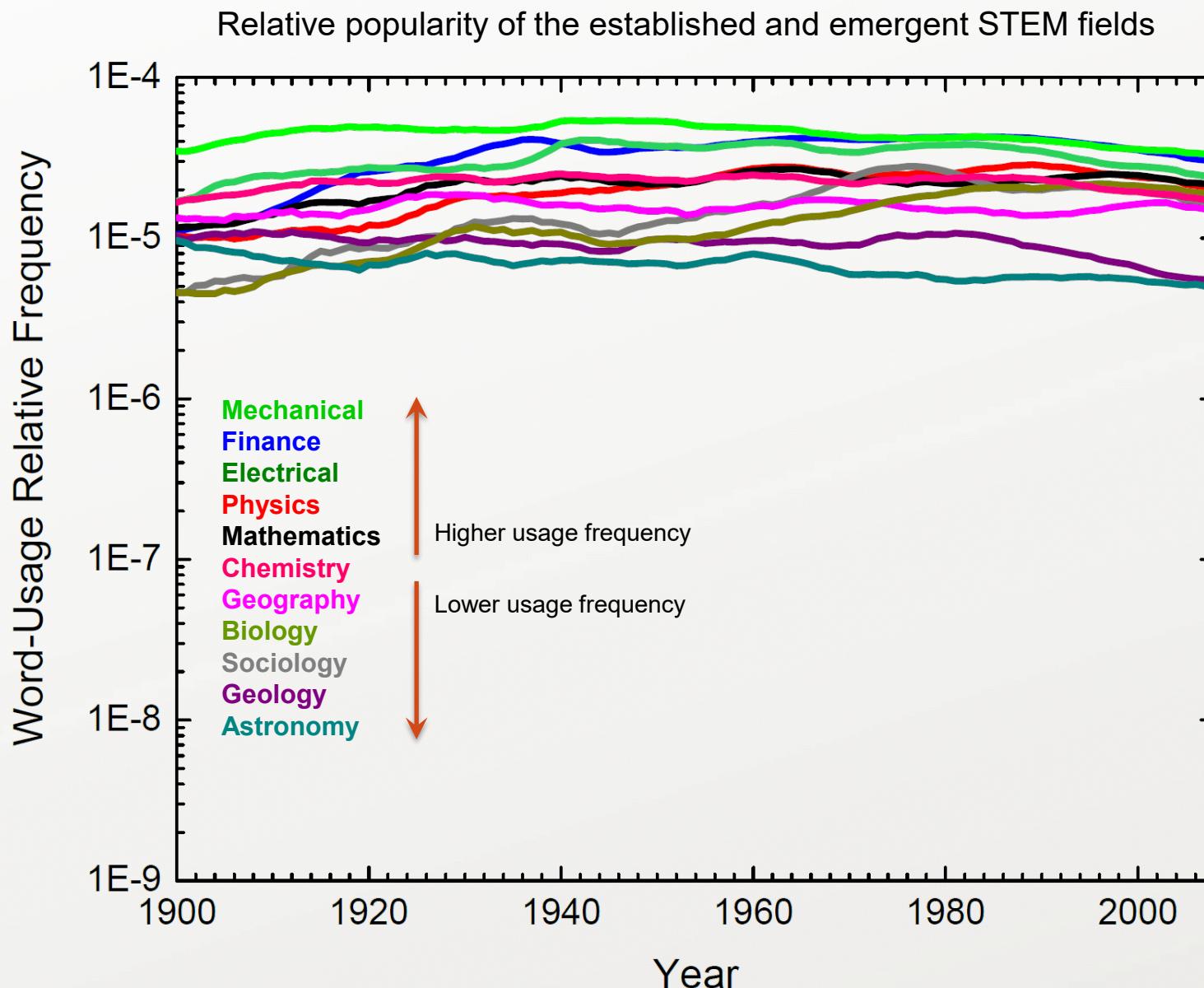
# The three pillars of science: Experiment, Theory, Computation+Data

## Three major roles of computational models in contemporary science

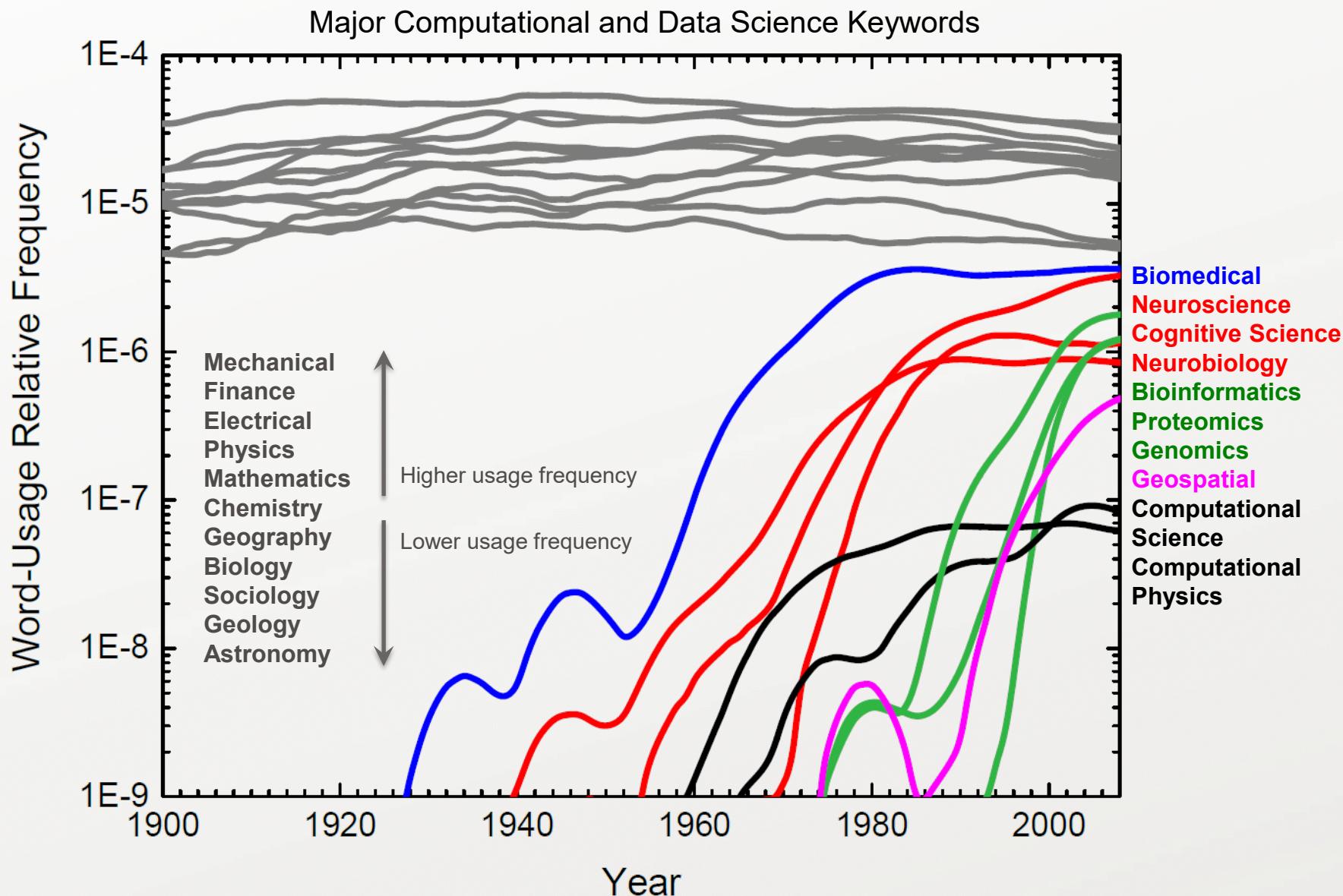
1. a workflow bridging data, hypothesis/theory, and prediction (**predictive computing**).
2. a substitute for experiment and observational data, where it is not available (**numerical simulation**).
3. a substitute for theory, where it is not available (**data-driven discovery via machine learning, deep learning, ...**).



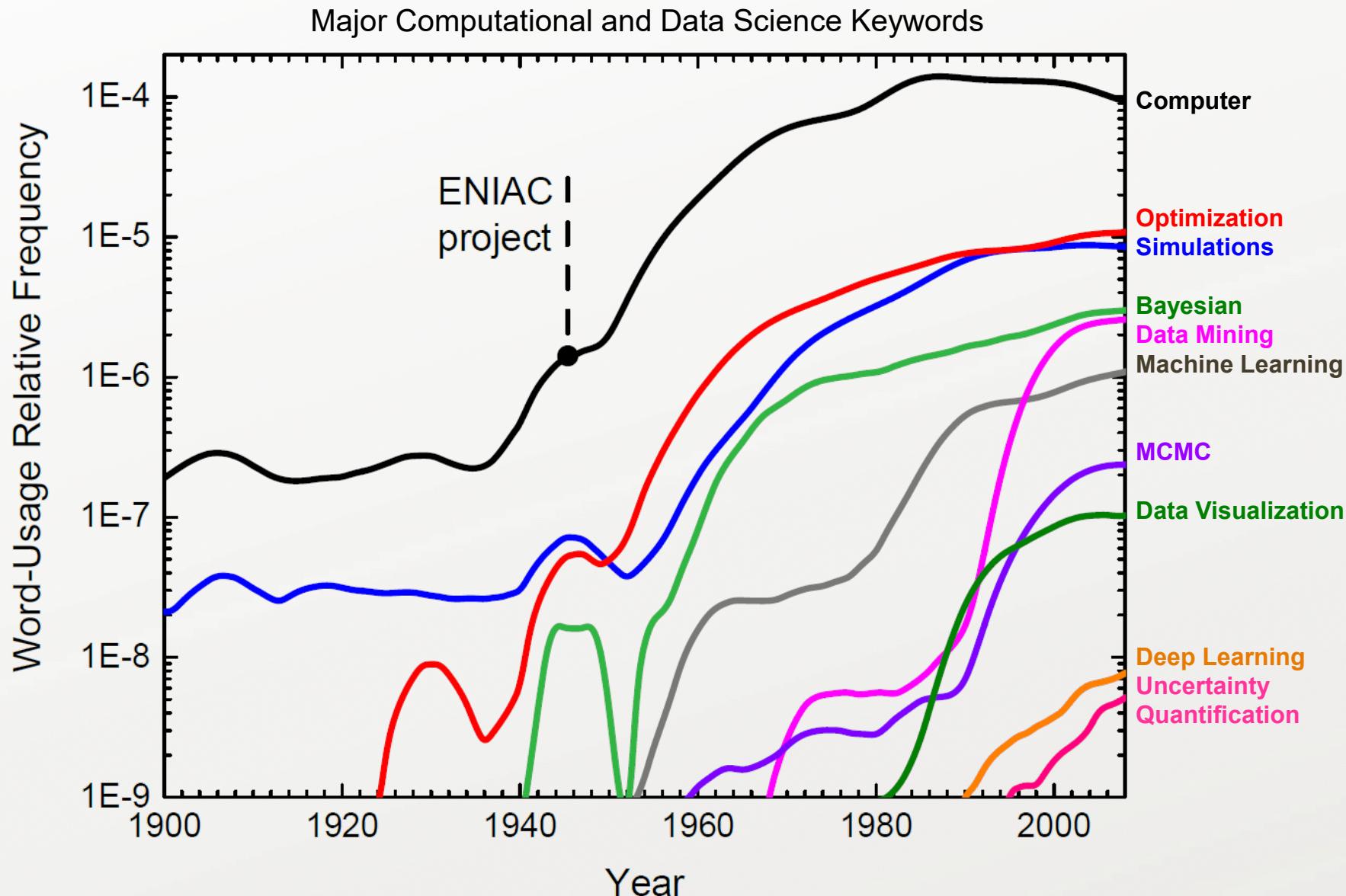
# The emerging third pillar of science: Computational and Data Sciences



# The emerging third pillar of science: Computational and Data Sciences



# The emerging third pillar of science: Computational and Data Sciences



## Data Science job market has grown by 700% over a half-decade

Skill	2013-2018 Growth
Machine Learning	809%
R	298%
Data Analysis	86%
Data Management	78%
SQL	45%

Role	2013-2018 Growth
Data Scientist	663%
Marketing Data Analyst	194%
Bioinformatician	75%
Financial Quantitative Analyst	57%

Data: Burning Glass Technologies 2019

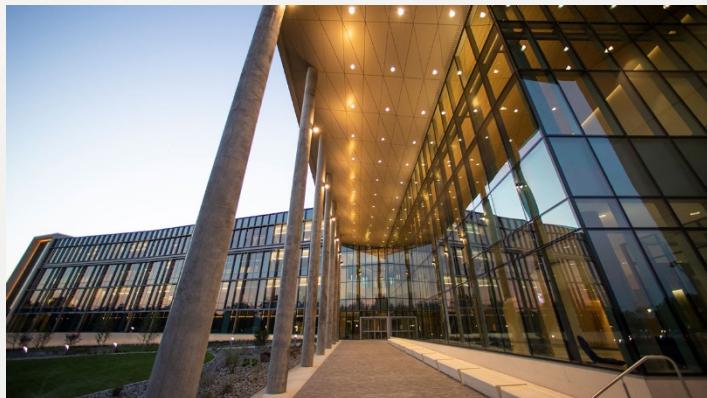
# The Data Science Program of The University of Texas Arlington

## UTA

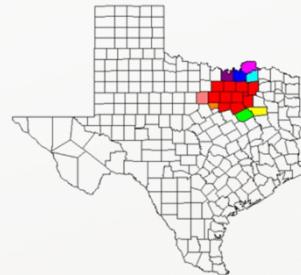
- Established 1895
- Located at DFW metroplex
  - 8,000,000 population
  - 3<sup>rd</sup>-largest concentration of Fortune 500
  - the largest economic engine of Texas
- >42,000 students
- Second-largest institution within the University of Texas System
- 5<sup>th</sup> most-diverse campus in the US (2018)
- The UTA Data Science program
  - established by the College of Science
  - Undergraduate Major and Minor
  - Final approval in April 2020

## UTA College of Science

- Biology
- Chemistry
- Environmental Science
- Mathematics
- Physics
- Psychology



Science & Engineering Innovation & Research Building



Dallas-Fort-Worth Metroplex



Dallas



Fort-Worth



Arlington



The UTA Data Science faculty

# The UTA Data Science program is unique in the nation for its focus on Natural Sciences

**Data Science Codes** of the Texas higher education coordinating board:

- **30.7001 Data Science, General.**

A program that focuses on the **analysis of large-scale data sources** from the interdisciplinary perspectives of **applied statistics, computer science, data storage, data representation**, data modeling, mathematics, and statistics. Includes instruction in **computer algorithms, computer programming, data management**, data mining, information policy, information retrieval, mathematical modeling, quantitative analysis, statistics, trend spotting, and visual analytics.

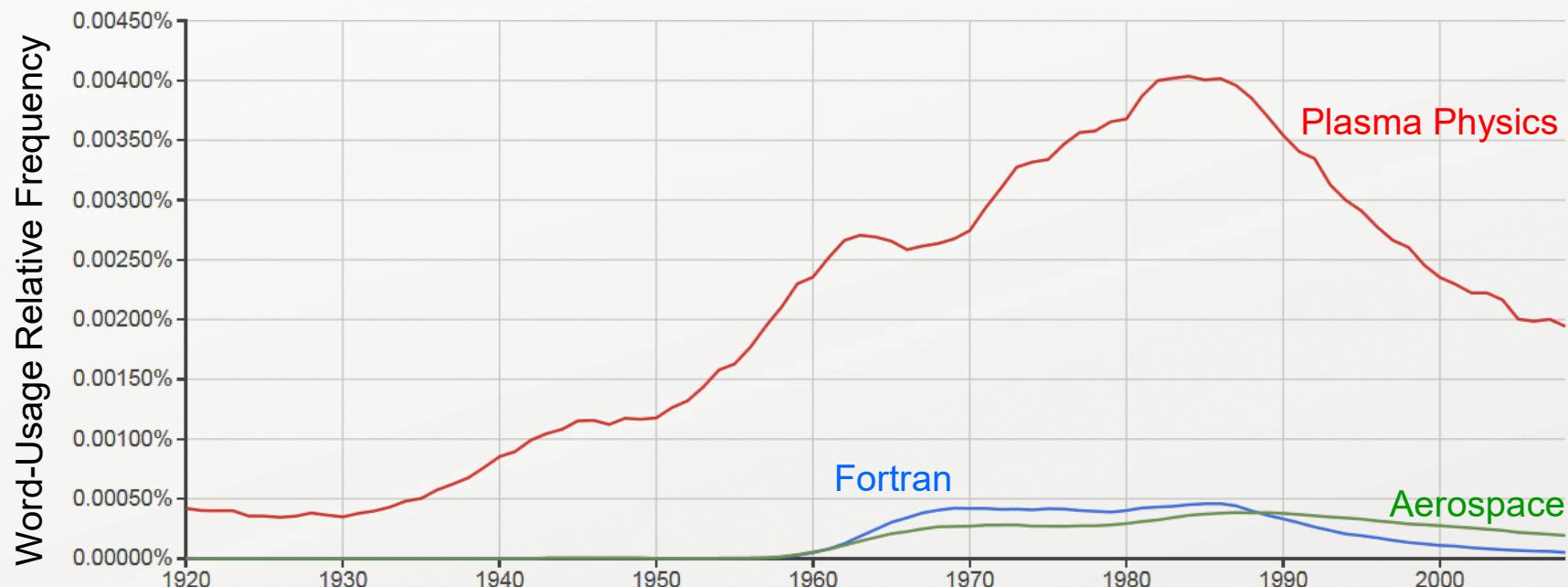
- **30.7101 Data Analytics, General.**

A program that prepares individuals to **apply data science** to generate insights from data and identify and predict trends. Includes **instruction in computer databases, computer programming**, inference, machine learning, optimization, probability and stochastic models, **statistics**, strategy, uncertainty quantification, and visual analytics.

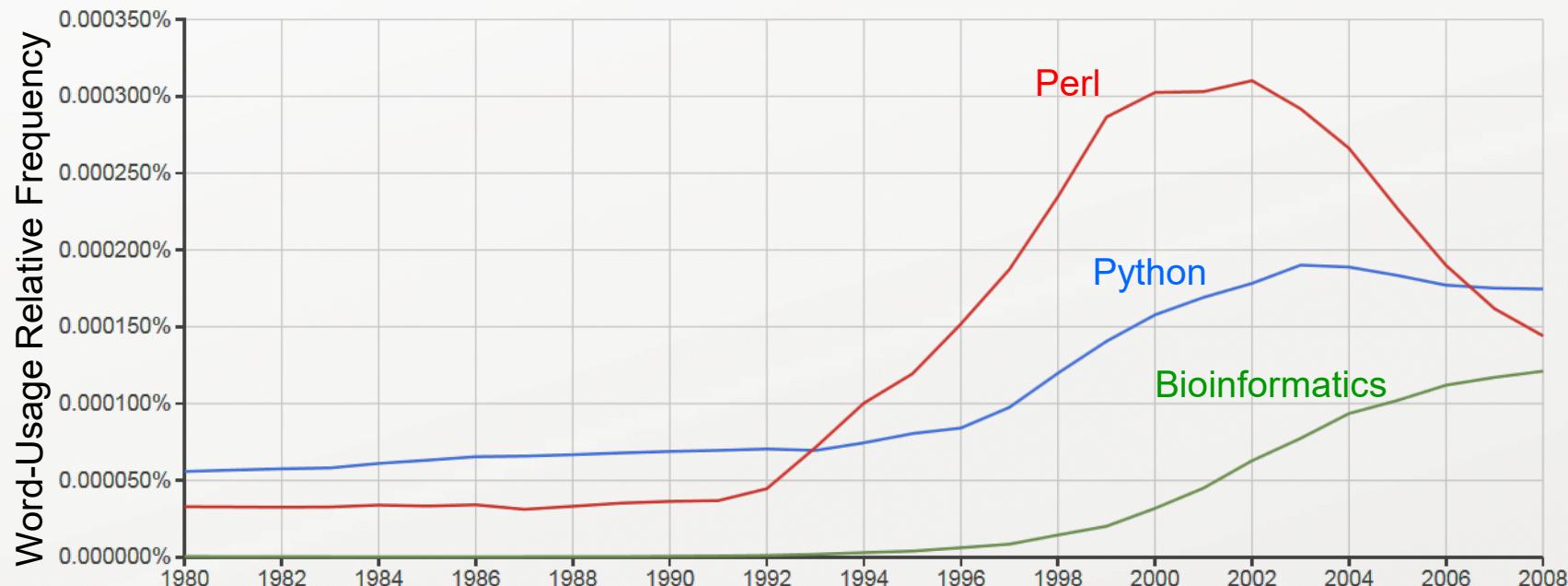
- **30.7099 Data Science, Other.**

Any instructional program in data science not listed above.

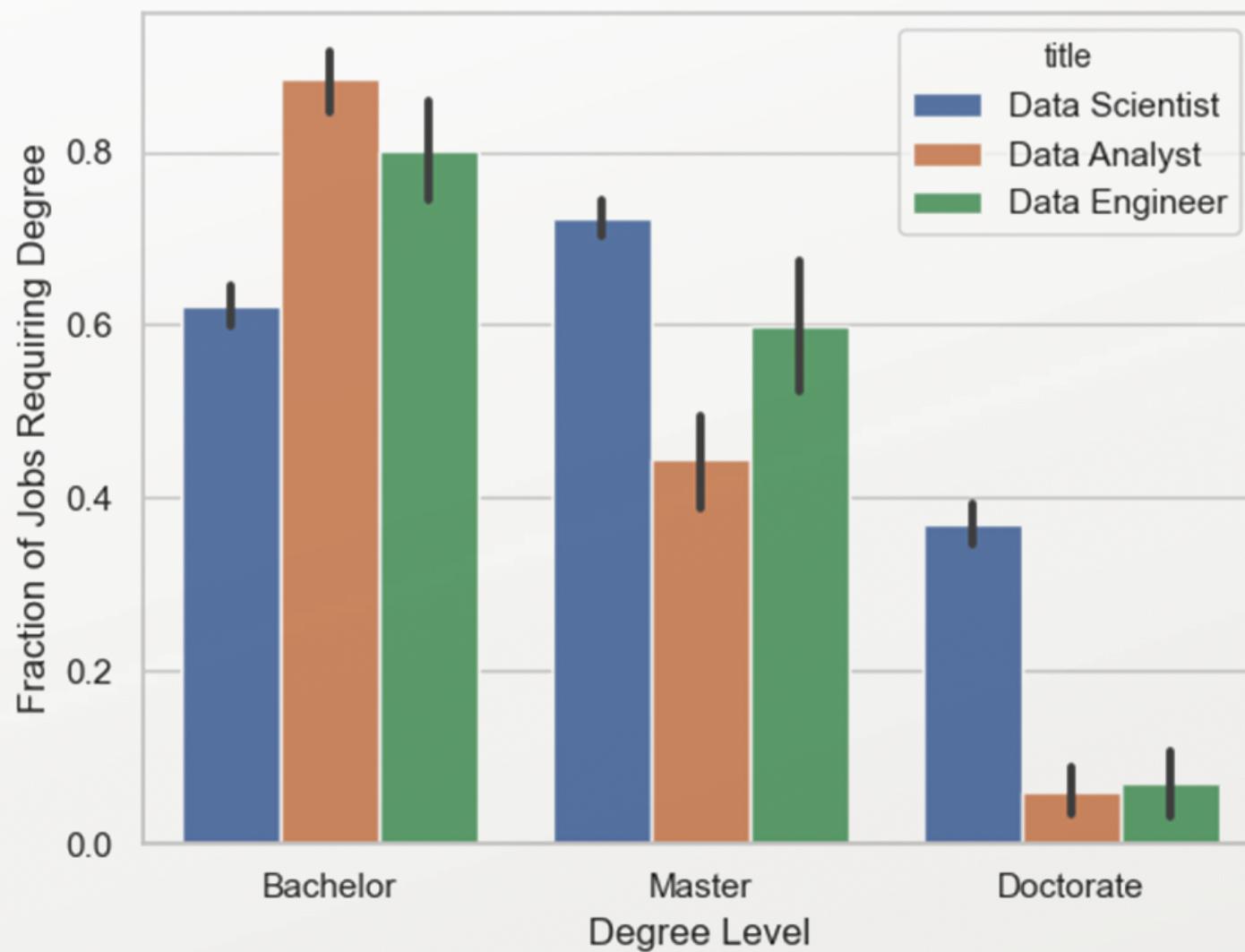
# The landscape of scientific tools is highly dynamic and domain-specific



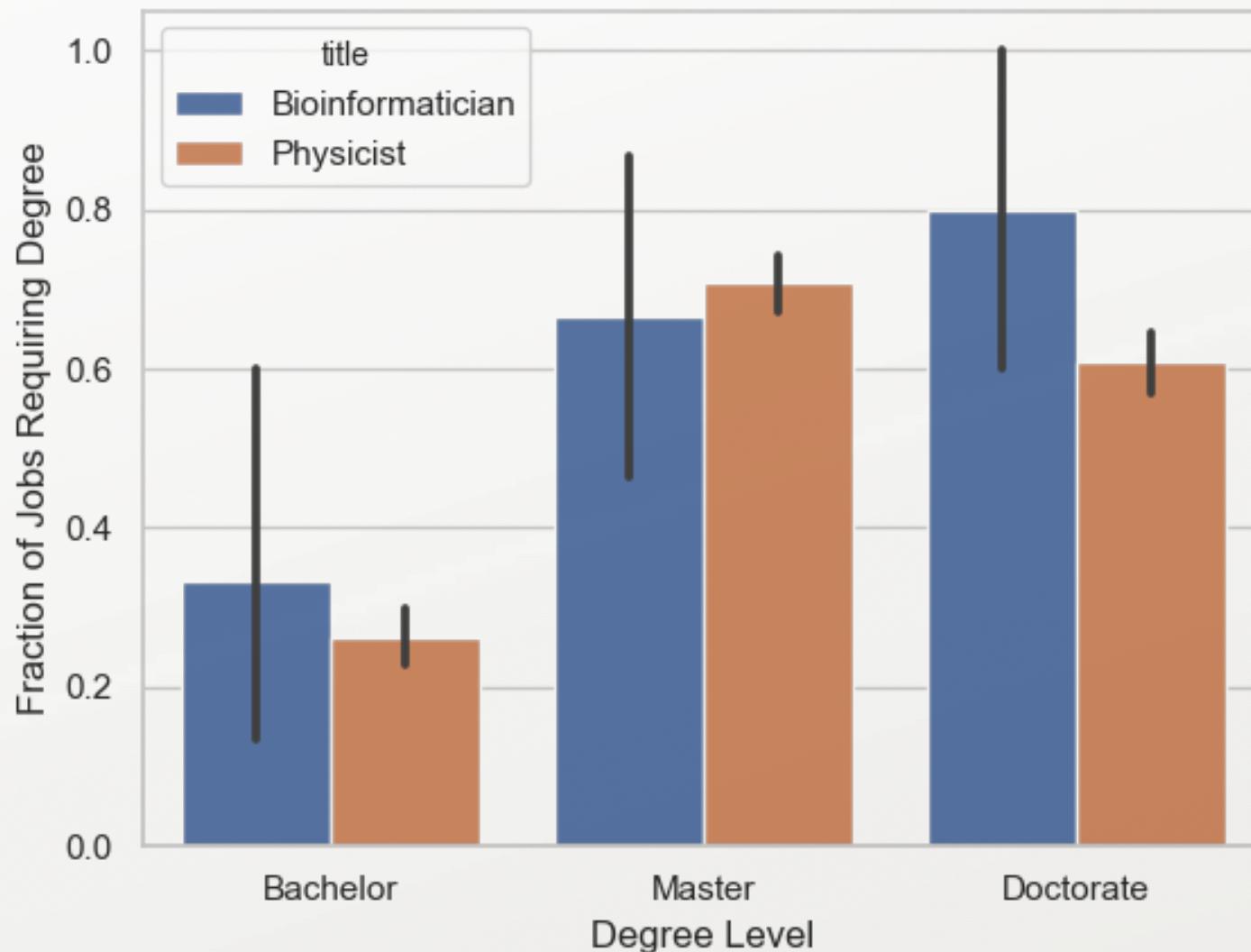
# The landscape of scientific tools is highly dynamic and domain-specific



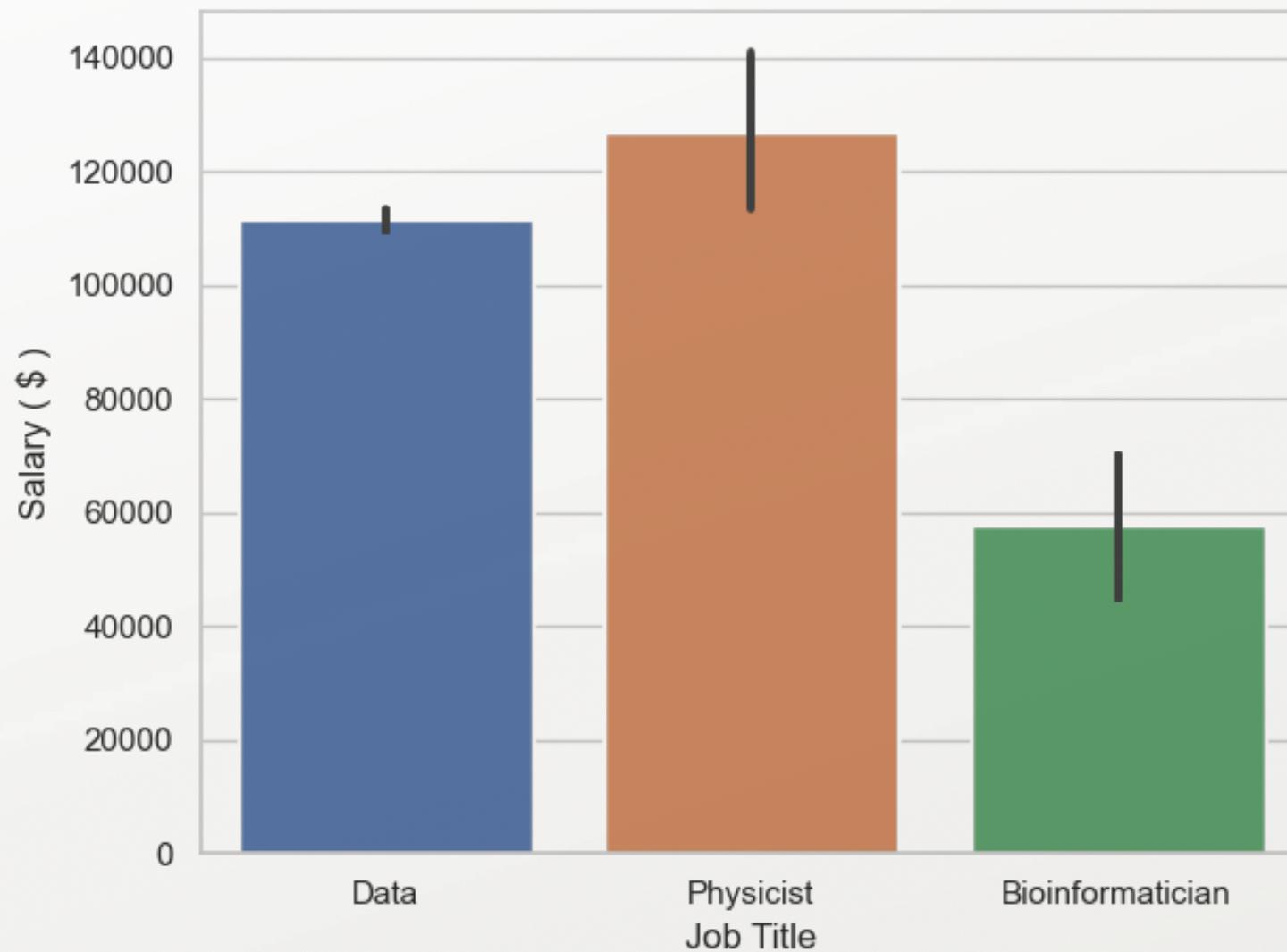
## Data-related jobs generally require less advanced degrees than STEM



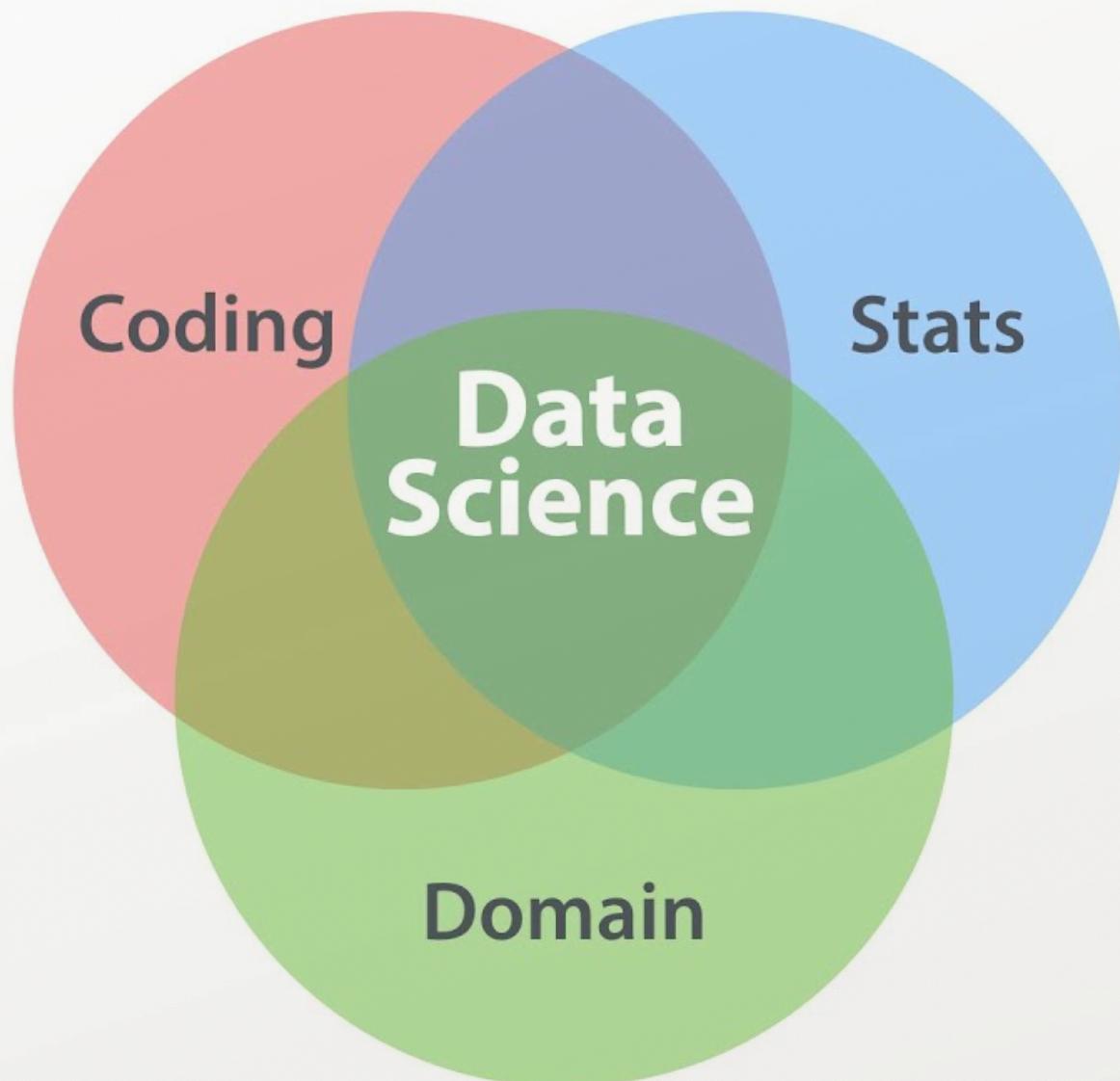
## Data-related jobs generally require less advanced degrees than STEM



## Data-related jobs generally pay much higher salaries than other STEM degrees



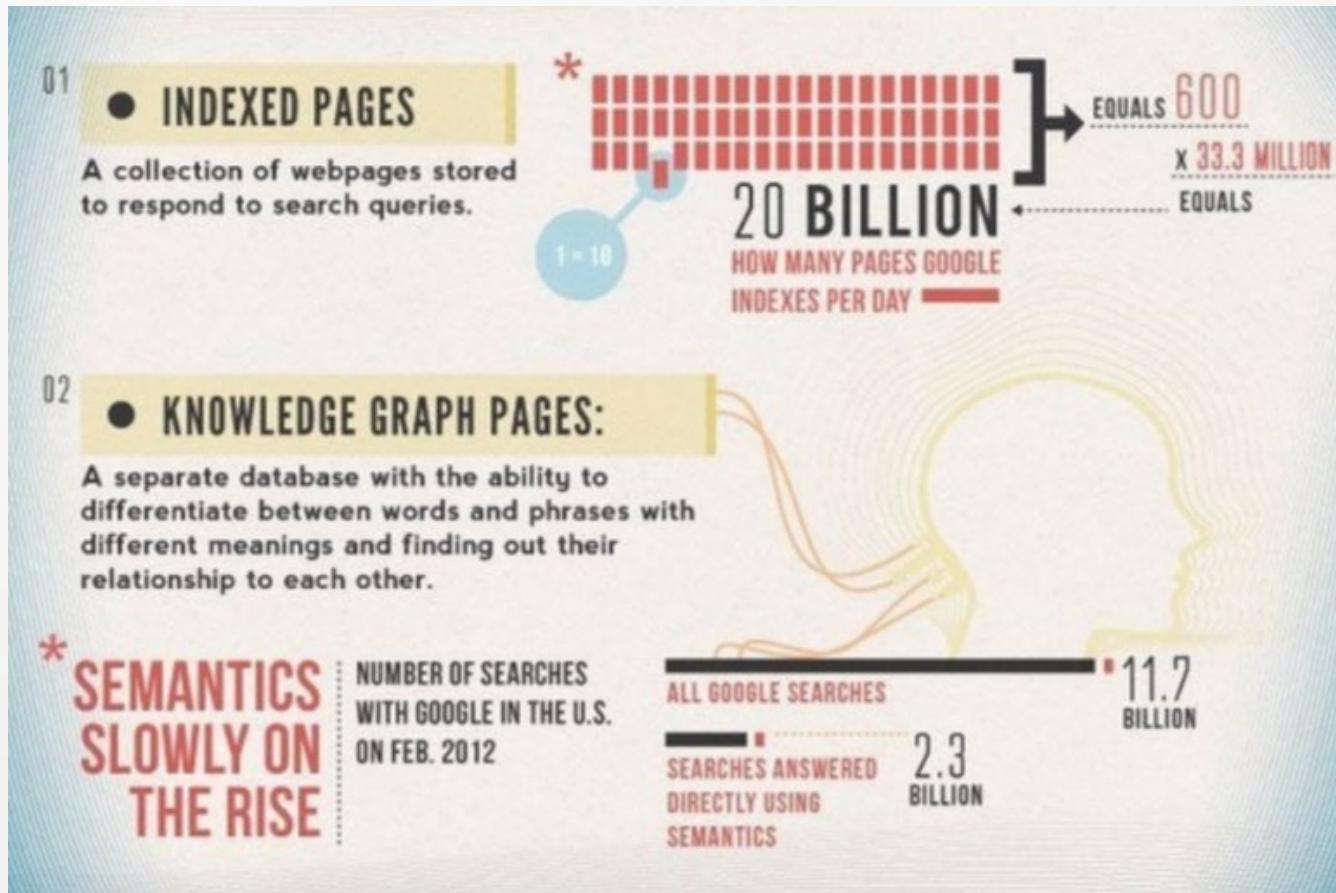
# Learning Data Science competencies via Word Embeddings





# BIG DATA: How does a search engine work

The search results come from indexed pages and knowledge graphs.



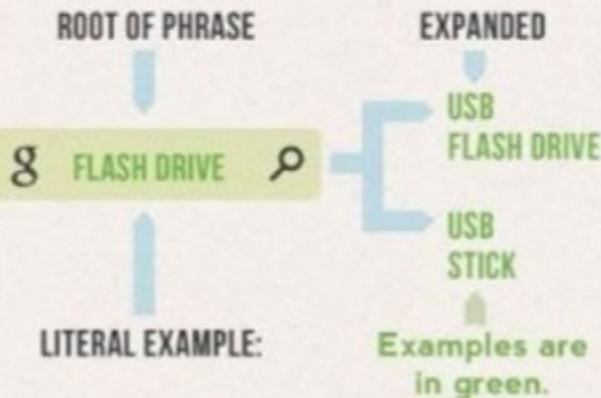
# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both

03

## ● LITERAL SEARCH:

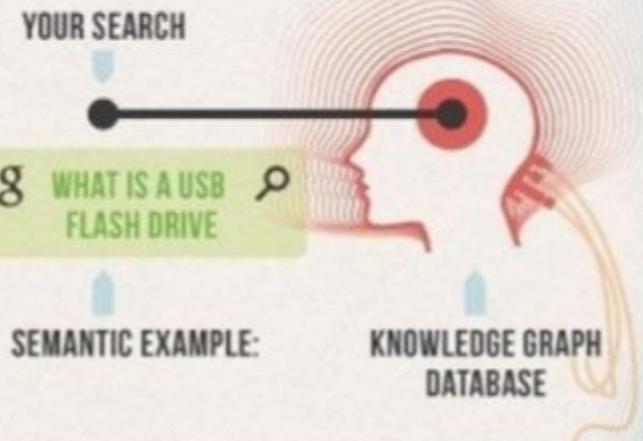
Search engines look for a match for some of or the entire phrase. The root of your search phrase is then found, examined and expanded upon to find a better result.



04

## ● SEMANTIC SEARCH:

These searches attempt to understand the context of a phrase by analyzing the terms and language in the Knowledge Graph database to directly answer a question with specific information.



# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both

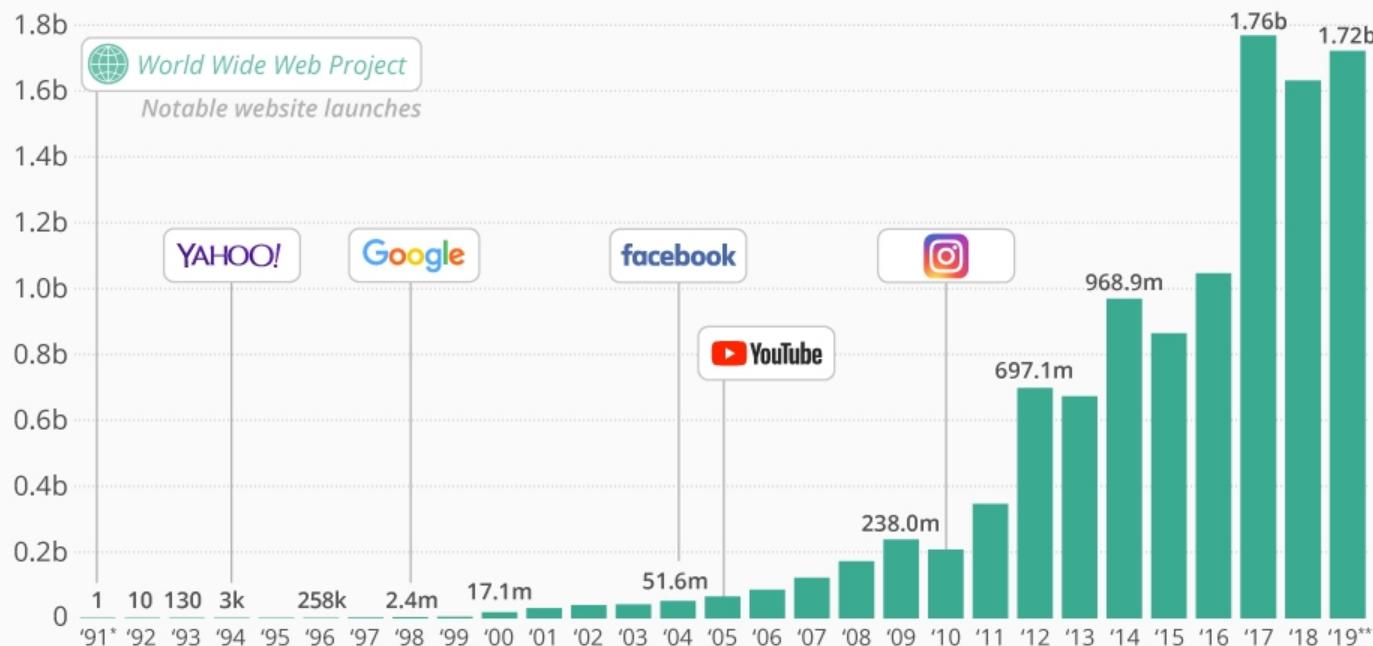


# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both

## How Many Websites Are There?

Number of websites online from 1991 to 2019



"Website" is defined as a unique hostname, i.e. a name which can be resolved, using a name server, into an IP Address.

\* As of August 1, 1991

\*\* As of October 28, 2019 at 10:00 CET

# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both

accept your query

parse your query

figure out the word order

look up the information in its database

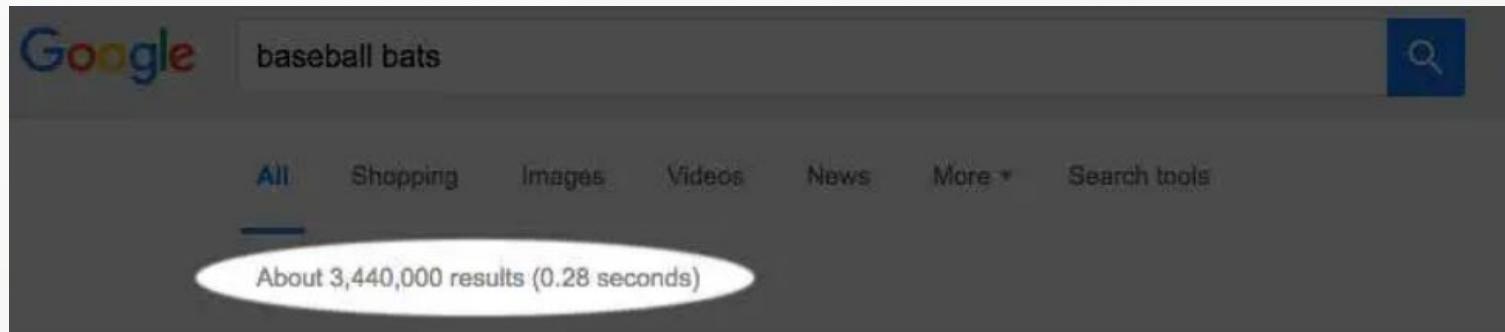
personalize your results by taking into account everything it knows about you

rank the results

send the results to your browser

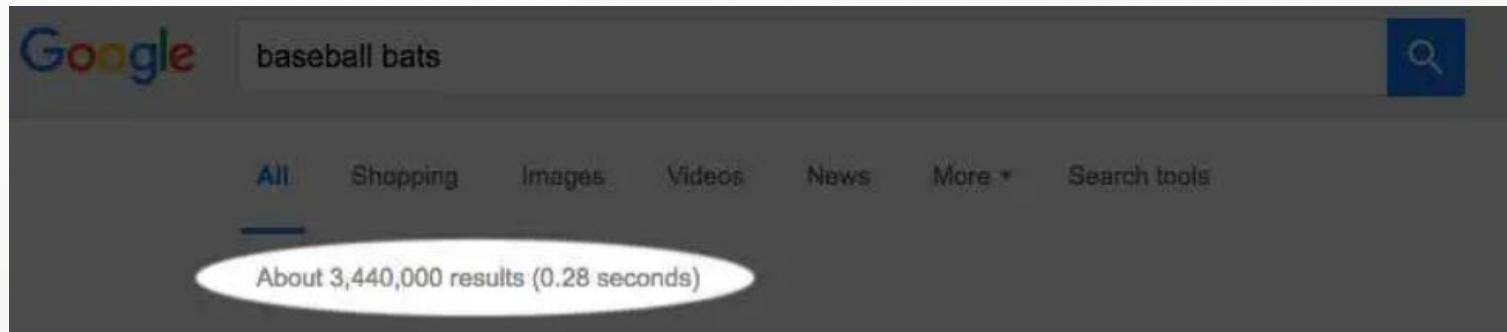
# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both



# How does a search engine work

When you type a phase into the search bar, Google analyzes the literal and semantics of the query and searches for both



Multiple Datacenters with a Worldwide Load Balancing Network

Hundreds of Computers in Each Datacenter Using Distributed Lookups

Custom File System and Custom Software

Caching, Index Stored in RAM, Pre-fetching Results

