

---

# Data Science

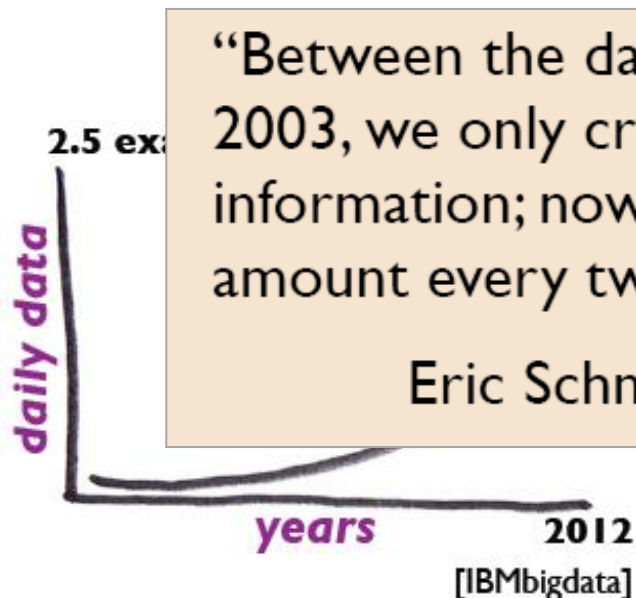
## Lecture 1: Overview / Data Munging

---

Acknowledgement: Lecture materials are prepared using The Data Science Design Manual by Steven S. Skiena, 2017.

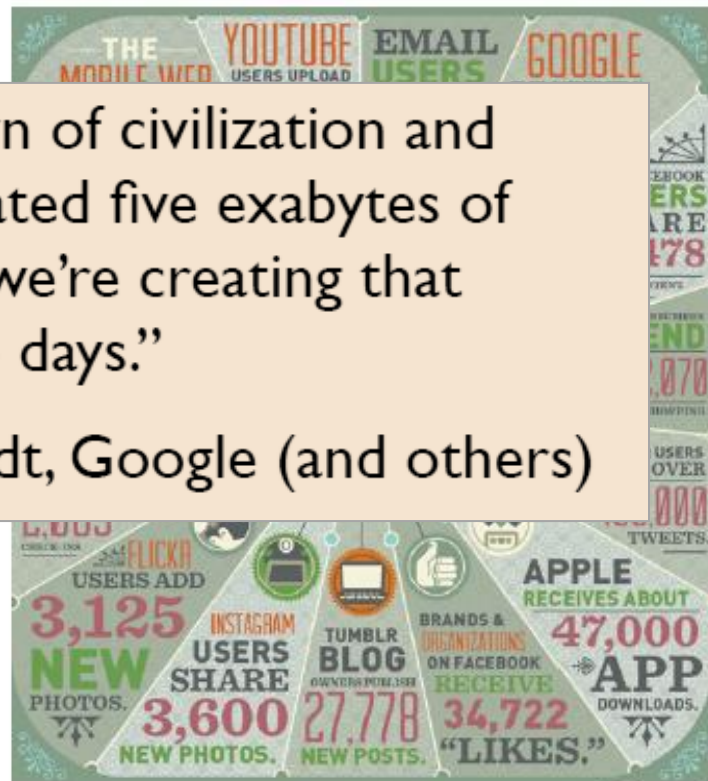
---

# Big Data

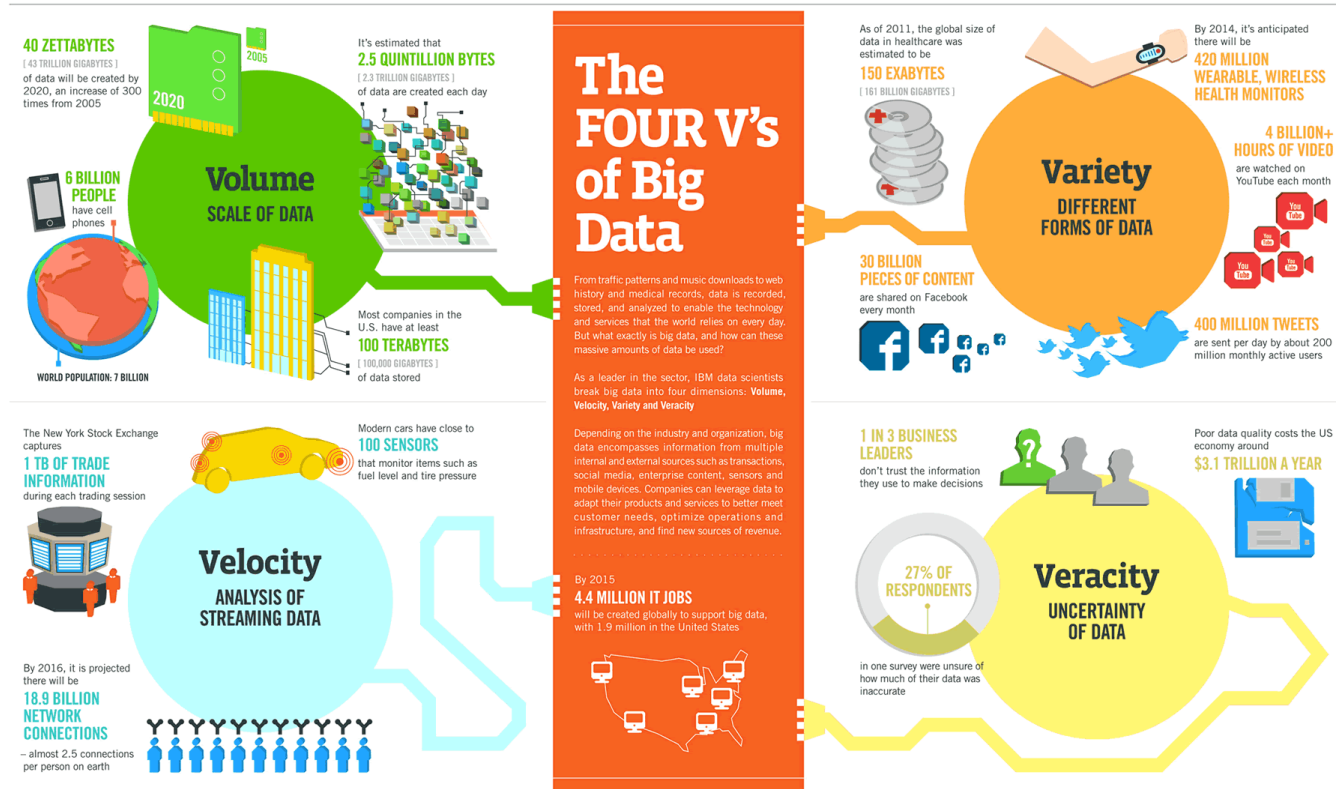


“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

Eric Schmidt, Google (and others)



# Big Data Explosion



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTCC, QAS

IBM

# Big Data Challenges

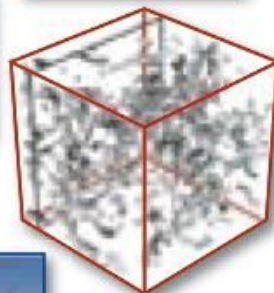
- Big Data
    - Large and complex data
    - E.g., social data, web data, financial transaction data, academic articles, genomic data etc.
  - Two challenges:
    - Efficient storage and access
    - **Data analytics** to mine valuable information
-

# Science Paradigms

- Thousand years ago:  
science was **empirical**  
*describing natural phenomena*
- Last few hundred years:  
**theoretical** branch  
*using models, generalizations*
- Last few decades:  
a **computational** branch  
*simulating complex phenomena*
- Today: **data exploration** (eScience)  
*unify theory, experiment, and simulation*
  - Data captured by instruments  
or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files  
using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

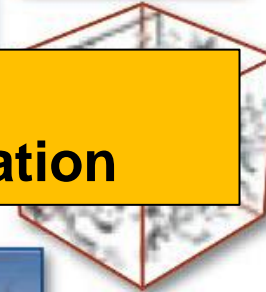


# Science Paradigms

- Thousand years ago:  
science was **empirical**  
*describing natural phenomena*
- Last few hundred years:  
**theoretical** branch  
*using models, generalizations*
- **Future: Data-driven Science**  
- **Data-driven Hypothesis Generation**  
*simulating complex phenomena*
- Today: **data exploration** (eScience)  
*unify theory, experiment, and simulation*
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = 4\pi G\rho - \frac{c^2}{a^2}$$





# Market demands for Big Data Scientists



(Harvard Business Review, 2012)

## Data Scientist: *The Sexiest Job of the 21st Century*

**Meet the people who  
can coax treasure out of  
messy, unstructured data.**  
by Thomas H. Davenport  
and D.J. Patil

**W**hen Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

# What is Data Science?

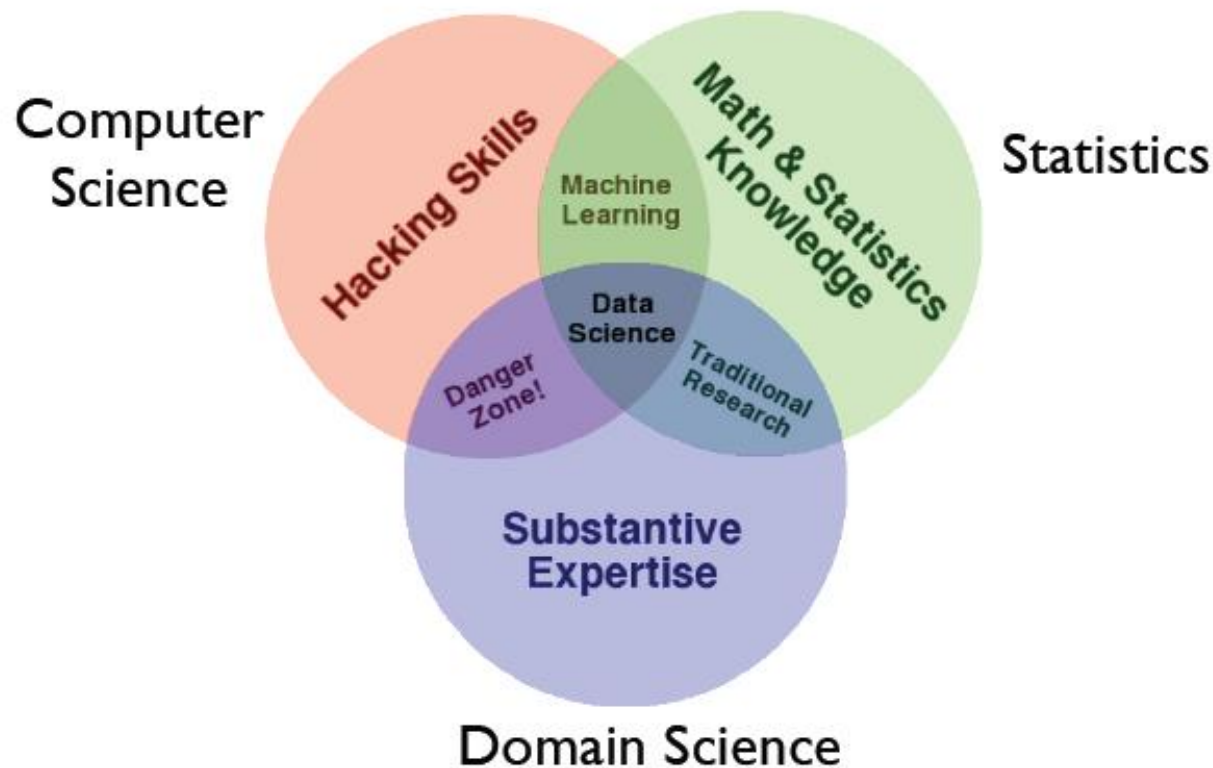
---

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.
-



# Data Science



# A Data Scientist Is...

“A data scientist is someone who knows more statistics than a computer scientist and more computer science than a statistician.”

- Josh Blumenstock

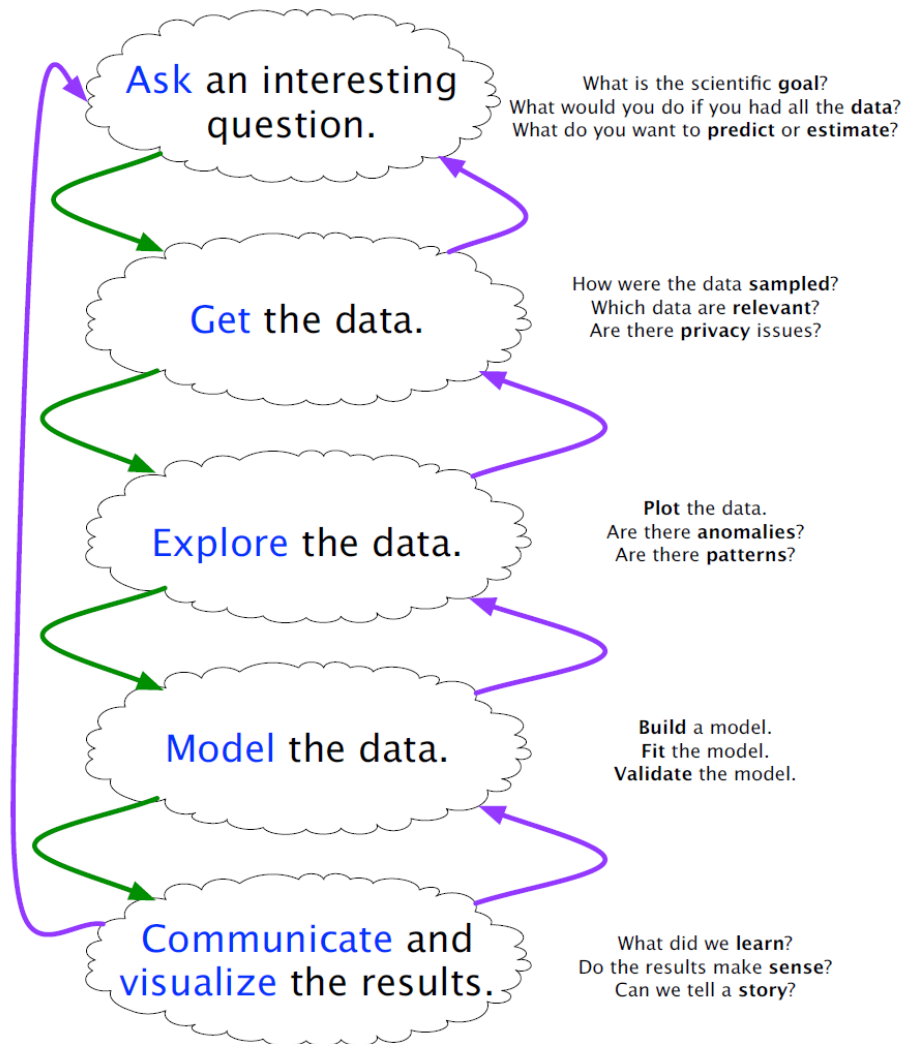
“Data Scientist = statistician + programmer + coach + storyteller + artist”

- Shlomo Aragmon

# Hal Varian Explains...

The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and **ubiquitous data.**” – Hal Varian

# Typical Data Science Pipeline



# What do data scientists do?

## Enterprise Data Analysis and Visualization: An Interview Study

Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer

**Abstract**—Organizations rely on data analysts to model customer engagement, streamline operations, improve production, inform business decisions, and combat fraud. Though numerous analysis and visualization tools have been built to improve the scale and efficiency at which analysts can work, there has been little research on how analysis takes place within the social and organizational context of companies. To better understand the enterprise analysts' ecosystem, we conducted semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance. Based on our interview data, we characterize the process of industrial data analysis and document how organizational features of an enterprise impact it. We describe recurring pain points, outstanding challenges, and barriers to adoption for visual analytic tools. Finally, we discuss design implications and opportunities for visual analysis research.

**Index Terms**—Data, analysis, visualization, enterprise.

---

### 1 INTRODUCTION

Organizations gather increasingly large and complex data sets each year. These organizations rely on data analysis to model customer engagement, streamline operations, improve production, inform sales and business decisions, and combat fraud. Within organizations, an increasing number of individuals—with varied titles such as “business analyst”, “data analyst” and “data scientist”—perform such analyses. These analysts constitute an important and rapidly growing user population for analysis and visualization tools.

Enterprise analysts perform their work within the context of a larger organization. Analysts often work as a part of an analysis team or business unit. Little research has observed how existing infrastructure, available data and tools, and administrative and social conventions within an organization impact the analysis process within the enterprise. Understanding how these issues shape analytic workflows can inform the design of future tools.

To better understand the day-to-day practices of enterprise analysts

and wrangling, often the most tedious and time-consuming aspects of an analysis, are underserved by existing visualization and analysis tools. We discuss recurring pain points within each task as well as difficulties in managing workflows across these tasks. Example pain points include integrating data from distributed data sources, visualizing data at scale and operationalizing workflows. These challenges are typically more acute within large organizations with a diverse and distributed set of data sources.

We conclude with a discussion of future trends and the implications of our interviews for future visualization and analysis tools. We argue that future visual analysis tools should leverage existing infrastructures for data processing to enable scale and limit data migration. One avenue for achieving better interoperability is through systems that specify analysis or data processing operations in a high-level language, enabling retargeting across tools or platforms. We also note that the current lack of reusable workflows could be improved via less

# What do data scientists do?

		Hacker																				Scripter					Application User									
		Analytics	Biology	Datamart	Finance	Finance	Healthcare	Healthcare	Healthcare	Insurance	Marketing	Marketing	News	Retail	Retail	Social Networking	Social Networking	Social Networking	Visualization	Web	Web	Analytics	Analytics	Analytics	Finance	Healthcare	Media	Retail	Finance	Insurance	Retail	Retail	Sports	Web	Security	
Process	Discovery	Locating Data	x	x	x	x	x	x	x	x					x	x	x	x	x																	
		Field Definitions	x	x	x	x	x	x	x		x	x		x		x	x	x		x																
	Wrangle	Data Integration	x	x	x	x	x	x	x					x	x	x	x				x															
		Parsing Semi-Structured	x	x	x	x					x	x										x														
		Advanced Aggregation and Filtering	x					x	x	x				x			x	x	x	x	x															
	Profile	Data Quality	x			x	x	x	x	x						x	x	x	x	x	x															
		Verifying Assumptions		x		x	x					x	x		x		x	x	x	x	x															
	Model	Feature Selection	x	x	x							x	x	x	x	x	x	x	x	x	x															
		Scale	x	x	x	x	x		x	x			x	x	x	x	x	x																		
		Advanced Analytics	x		x		x					x	x	x	x																					
Report	Communicating Assumptions						x	x					x	x	x	x	x																			
	Static Reports		x	x		x					x	x									x															
Workflow	Data Migration	x	x	x	x	x								x	x	x	x																			
	Operationalizing Workflows	x	x			x					x	x		x	x	x	x																			
Tools	Database	SQL	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x						x							
		Hadoop/Hive/Pig	x	x												x		x																		
		MongoDB	x																																	
		CustomDB	x																																	
	Scripting	Java	x	x		x								x	x	x		x	x	x																
		Perl																																		
		Python	x		x	x	x																													
		Clojure																																		
		Visual Basic		x																																
	Modeling	R	x		x																															
		Matlab				x																														
		SAS																																		
		Excel		x		x	x																													

Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

# Data Cleaning

“In our experience, the tasks of exploratory data mining and data cleaning constitute **80% of the effort** that determines 80% of the value of the ultimate data.”

T. Dasu and T. Johnson  
Authors of *Exploratory Data Mining  
and Data Cleaning*



## 3 Industries That Will Be Transformed By AI, Machine Learning And Big Data In The Next Decade



**Bernard Marr**, CONTRIBUTOR

*I write about big data, analytics and enterprise performance* [FULL BIO](#) ✓

Opinions expressed by Forbes Contributors are their own.

Historically, when new technologies become easier to use, they transform industries.

That's what's happening with artificial intelligence and big data; as the barriers to implementation disappear (cost, computing power, etc.), more and more industries will put the technologies into use, and more and more startups will appear with new ideas of how to disrupt the status quo with these technologies.

By my predictions, the AI revolution isn't coming, it's already here, and we'll see it first in a few key sectors.

### Healthcare

Most people agree that healthcare is broken, and many startups believe that the biggest answer is putting the power back in the hands of the patient.

We're all carrying the equivalent of Star Trek's tricorder around in our

### 3 Industries to be transformed by ML, BD

1. Healthcare
2. Finance
3. Insurance

Ad closed by Google

Report this ad

Ads by Google ⓘ

MAY 13, 2017 @ 09:21 PM 72,363 👁

Free Webcast: Investing In Bitcoin & Crypto Assets

# IBM Predicts Demand For Data Scientists Will Soar 28% By 2020



**Louis Columbus**, CONTRIBUTOR

[FULL BIO](#) ✓

Opinions expressed by Forbes Contributors are their own.

- Jobs requiring machine learning skills are paying an average of \$114,000. Advertised data scientist jobs pay an average of \$105,000 and advertised data engineering jobs pay an average of \$117,000.
- 59% of all Data Science and Analytics (DSA) job demand is in Finance and Insurance, Professional Services, and IT.
- Annual demand for the fast-growing new roles of data scientist, data developers, and data engineers will reach nearly 700,000 openings by 2020.





Ad closed by Google

Report this ad

Why this ad? ⓘ

# Is Data Science Still a Rising Career?

## 50 Best Jobs in America for 2022

Best Jobs	2022	United States	Share				
Job Title	Median Base Salary	Job Satisfaction	Job Openings				
#1 Enterprise Architect	\$144,997	4.1/5	14,021	<a href="#">View Jobs</a>			
#2 Full Stack Engineer	\$101,794	4.3/5	11,252	<a href="#">View Jobs</a>			
#3 Data Scientist	\$120,000	4.1/5	10,071	<a href="#">View Jobs</a>			
#4 Devops Engineer	\$120,095	4.2/5	8,548	<a href="#">View Jobs</a>			
#5 Strategy Manager	\$140,000	4.2/5	6,977	<a href="#">View Jobs</a>			
#6 Machine Learning Engineer	\$130,489	4.3/5	6,801	<a href="#">View Jobs</a>			
#7 Data Engineer	\$113,960	4.0/5	11,821	<a href="#">View Jobs</a>			
#8 Software Engineer	\$116,638	3.9/5	64,155	<a href="#">View Jobs</a>			
#9 Java Developer	\$107,099	4.1/5	10,201	<a href="#">View Jobs</a>			
#10 Product Manager	\$125,317	4.0/5	17,725	<a href="#">View Jobs</a>			

Source: [https://www.glassdoor.com/List/Best-Jobs-in-America-LST\\_KQ0,20.htm](https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm)

# Is Data Science Still a Rising Career?

The U.S. Bureau of Labor Statistics sees strong growth in the data science field and predicts the number of jobs will increase by about **28%** through 2026.

(Source: <https://towardsdatascience.com/is-data-science-still-a-rising-career-in-2021-722281f7074c>)

# Big Data Revolution!

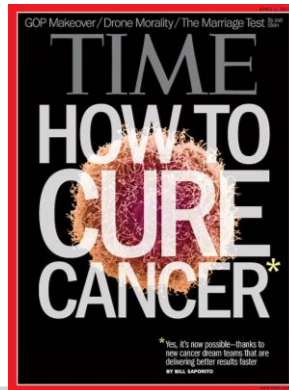
Google

amazon

Walmart\*

Renaissance

facebook



# Topics

- Lec 1: Introduction / Data Munging
  - Lec 2: Statistical Analysis / Visualizing Data
  - Lec 3: Modeling / Big Data
-

# Cultural Differences between Computer Science and Real Science (1)

---

## Scientists

- Data driven
- Try to understand messy natural world
- Focus on results (findings)
- Discover things
- Data is 1<sup>st</sup> class citizen

## Computer Scientists

- Algorithm driven
  - Build their own clean virtual world
  - Focus on methods
  - Invent things
  - Random data ok to prove correctness
-



# Cultural Differences between Computer Science and Real Science (2)

---

## Scientists

- $8/13 \approx 0.62$
- Care what it means
- Nothing is completely true
- Meaning

## Computer Scientists

- $8/13 = 0.61538461538$
  - Care what number is
  - Either correct or wrong
  - Accuracy
-

# Data Scientists

---

- Data scientists must learn to think like real scientists.
- Software developers are hired to produce code, while data scientists are hired to produce “ ”.

# 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:

- Internet Movie Database (IMDb)
  - NYC taxi cab records
  - Google Trends
-

---



**LIBERTY FILMS** -  
**Frank CAPRA** directs  
**IT'S A WONDERFUL LIFE**  
Starring **James STEWART**  
and **Donna REED**  
LIONEL BARRYMORE - THOMAS MITCHELL - HENRY TRUVERS  
BETTE DAVIS - NANA - BOB HOPE - GUY KIBBY - GUY KIBBY - GUY KIBBY  
Produced by FRANK CAPRA

**It's a Wonderful Life** (1946)

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

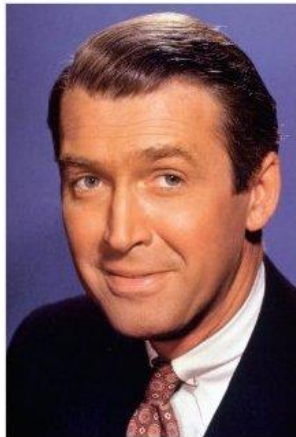
**Your rating:** ★★★★★★★★ -/10  
Ratings: **8.7**/10 from 202,743 users  
Reviews: 632 user | 187 critic

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...

# IMDb: Actor Data (<https://www.imdb.com/>)



## 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

---

- Predict movie ratings?
- What does the social network of actors look like? (Six degrees of Kevin Bacon, <https://oracleofbacon.org/> )



# 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

---

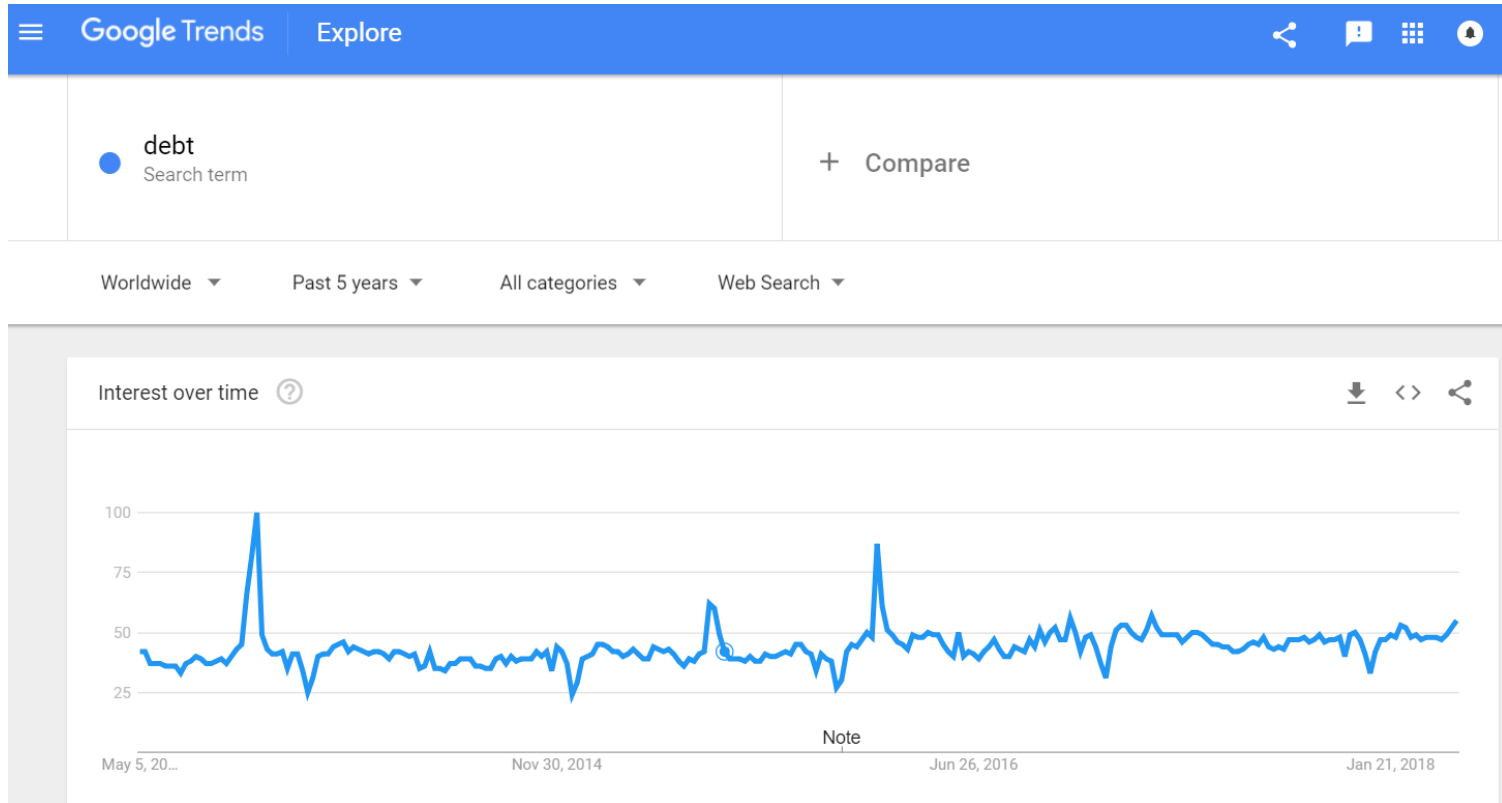
- How much do drivers make each night?
- How far do they travel?

# Google Trends

---

- Shows how often a particular search-term is entered relative to the total search-volume across various regions of the world, and in various languages.
  - Allows users to compare the relative search volume of searches between two or more terms.
-

# Google Trends (<https://trends.google.com/trends/>)



# An Example...

---

SCIENTIFIC  
REPORTS



SUBJECT AREAS:

STATISTICAL PHYSICS,  
THERMODYNAMICS AND  
NONLINEAR DYNAMICS

APPLIED PHYSICS

COMPUTATIONAL SCIENCE

INFORMATION THEORY AND  
COMPUTATION

## Quantifying Trading Behavior in Financial Markets Using *Google Trends*

Tobias Preis<sup>1\*</sup>, Helen Susannah Moat<sup>2,3\*</sup> & H. Eugene Stanley<sup>2\*</sup>

<sup>1</sup>Warwick Business School, University of Warwick, Scarman Road, Coventry, CV4 7AL, UK, <sup>2</sup>Department of Physics, Boston University, 590 Commonwealth Avenue, Boston, Massachusetts 02215, USA, <sup>3</sup>Department of Civil, Environmental and Geomatic Engineering, UCL, Gower Street, London, WC1E 6BT, UK.

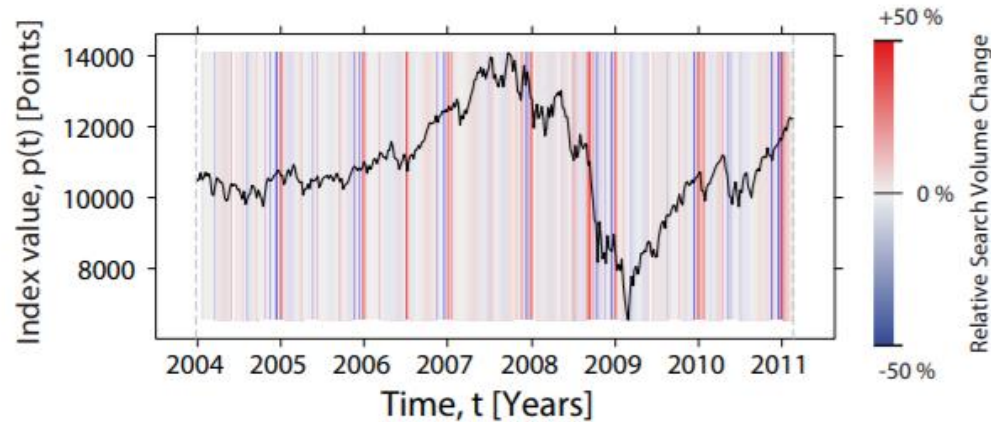
Crises in financial markets affect humans worldwide. Detailed market data on trading decisions reflect some of the complex human behavior that has led to these crises. We suggest that massive new data sources resulting from human interaction with the Internet may offer a new perspective on the behavior of market participants in periods of large market movements. By analyzing changes in *Google* query volumes for search terms related to finance, we find patterns that may be interpreted as “early warning signs” of stock market moves. Our results illustrate the potential that combining extensive behavioral data sets offers for a better understanding of collective human behavior.

Received  
25 February 2013

Accepted  
3 April 2013

# An Example...Cont'd

---



**Figure 1 | Search volume data and stock market moves.** Time series of closing prices  $p(t)$  of the *Dow Jones Industrial Average* (DJIA) on the first day of trading in each week  $t$  covering the period from 5 January 2004 until 22 February 2011. The color code corresponds to the relative search volume changes for the search term *debt*, with  $\Delta t = 3$  weeks. Search volume data are restricted to requests of users localized in the United States of America.

# An Example...Cont'd

---



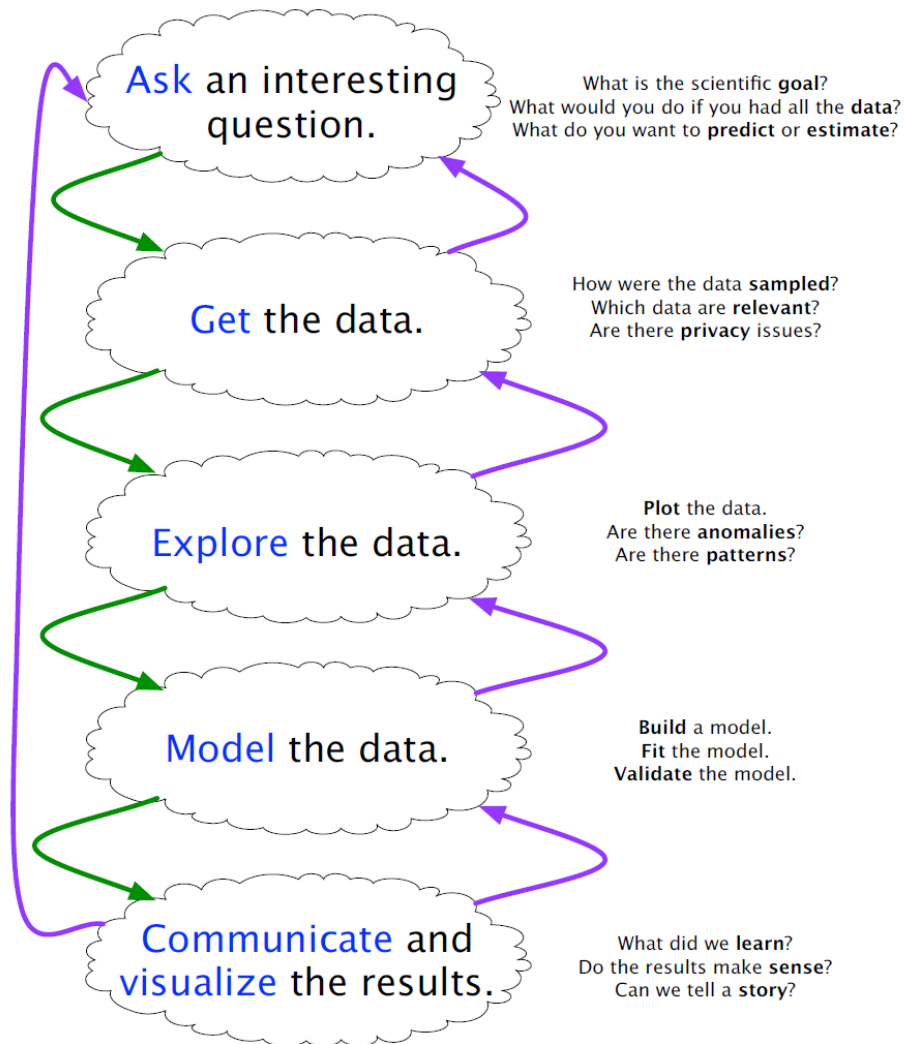
**Figure 2 | Cumulative performance of an investment strategy based on *Google Trends* data.** Profit and loss for an investment strategy based on the volume of the search term *debt*, the best performing keyword in our analysis, with  $\Delta t = 3$  weeks, plotted as a function of time (*blue line*). This is compared to the “buy and hold” strategy (*red line*) and the standard deviation of 10,000 simulations using a purely random investment strategy (*dashed lines*). The *Google Trends* strategy using the search volume of the term *debt* would have yielded a profit of 326%.



# Data Munging

---

# Typical Data Science Pipeline



# Data Munging

---

Good data scientists spend most of their time cleaning and formatting data.

The rest spend most of their time complaining there is no data available.

*Data munging* or *data wrangling* is the art of acquiring data and preparing (cleaning) it for analysis.

---

# Acquiring Data

---

- Mostly mean “find stuff on the internet!”
  - A lot of data stored in text files and on government websites
  - Files (CSV, XML, JSON)
  - Databases (SQL server)
  - API (Application Programming Interface)
  - Web Scraping (HTML)
-

# Common Data Formats

---

- CSV (Comma-Separated Values)

```
StudentID, Name, Dept, Birthdate, Gender, AdvisorID, GPA
3219875, William Lee, CS, 2000/1/1, M, 111212, 3.7
3219774, Emma Kim, CS, 2007/2/7,, 223123, 4.5
3219875, Noah Choi, CS, 9999/9/9, M, 111212, -0.5
:
```

StudentID	Name	Dept	Birthdate	Gender	AdvisorID	GPA
3219875	William Lee	CS	2000/1/1	M	111212	3.7
3219774	Emma Kim	CS	2007/2/7		223123	4.5
3219875	Noah Choi	CS	9999/9/9	M	111212	-0.5

- XML (eXtensible Markup Language)

```
<Document Element>
  <Student Table>
    <Student>
      <StudentID> 3219875 </StudentID>
      <Name> William Lee </Name>
      <Dept> CS </Dept>
      <Birthdate> 2000/1/1 </Birthdate>
      <Gender> M </Gender>
      <AdvisorID> 111212 </AdvisorID>
      <GPA> 3.7 </GPA>
    </Student>
    <Student>
      <StudentID> 3219774 </StudentID>
      <Name> Emma Kim </Name>
      <Dept> CS </Dept>
      :
```

## ● JSON (JavaScript Object Notation)

```
{
  "sedol": {
    "StudentID": 3219875,
    "Name": "William Lee",
    "Dept": "CS",
    "Birthdate": 2000/1/1,
    "Gender": "M",
    "AdvisorID": 111212,
    "GPA": 3.7
  },
  "alpha": {
    "StudentID": 3219774,
    "Name": "Emma Kim",
    "Dept": "CS",
    :
  }
}
```

# Loading data from files (CSV example)

---

```
import pandas  
student_data = pandas.read_csv("student_table.csv")
```

```
StudentID, Name, Dept, Birthdate, Gender, AdvisorID, GPA  
3219875, William Lee, CS, 2000/1/1, M, 111212, 3.7  
3219774, Emma Kim, CS, 2007/2/7,, 223123, 4.5  
3219875, Noah Choi, CS, 9999/9/9, M, 111212, -0.5  
:
```



Pandas  
dataframe

- Loading data from XML and JSON files is similar
-



# Getting data from Relational DB

---

```
import pandas
student_data = pandas.read_sql_query(sql_string, db_uri)
```

StudentID	Name	Dept	Birthdate	Gender	AdvisorID	GPA
3219875	William Lee	CS	2000/1/1	M	111212	3.7
3219774	Emma Kim	CS	2007/2/7		223123	4.5
3219875	Noah Choi	CS	9999/9/9	M	111212	-0.5

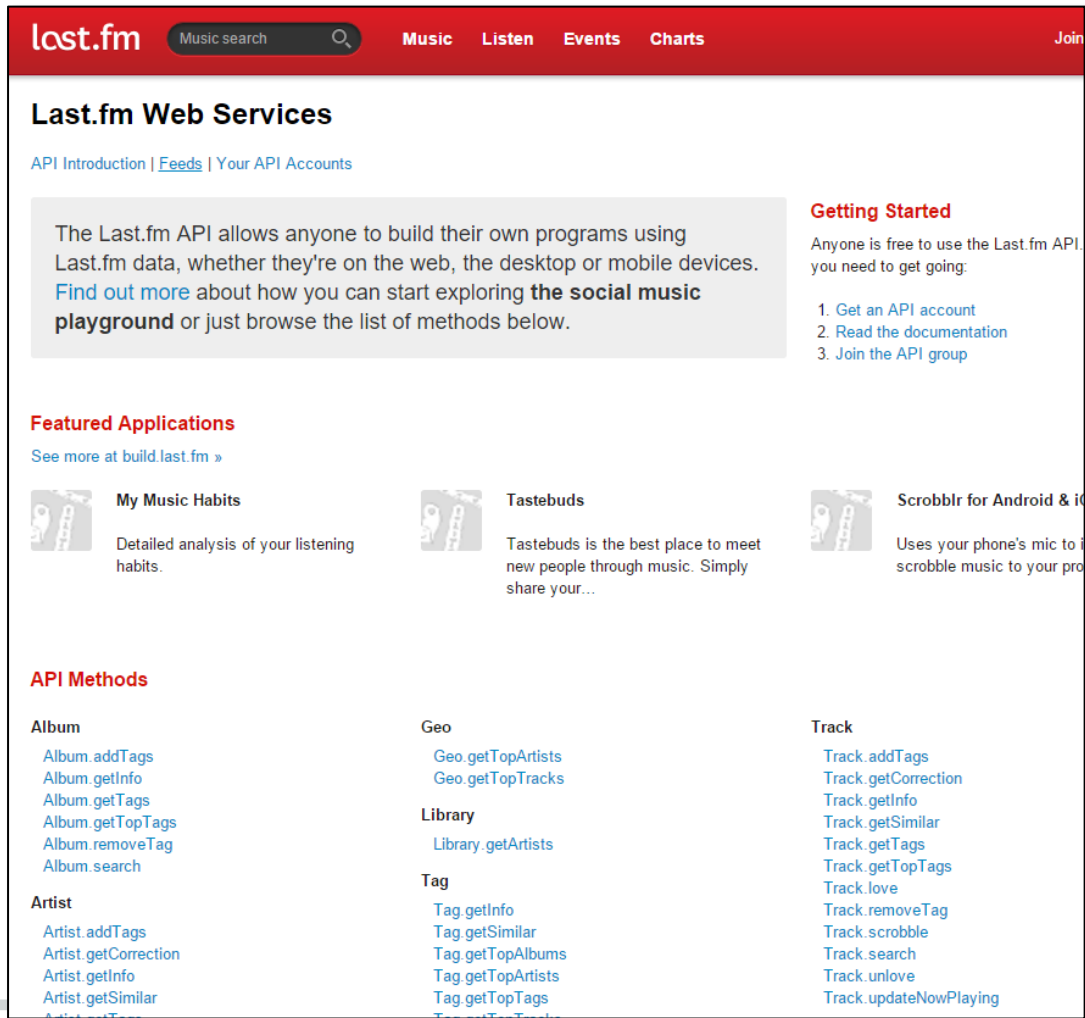
- `SELECT * FROM student_table`
  - `SELECT studentid, name FROM student_table WHERE gpa > 3.5`
  - `SELECT S.name, A.name`  
`FROM student_table S, advisor_table A`  
`WHERE S.advisorid = A.id`
-

StudentID	Name	Dept	Birthdate	Gender	AdvisorID	GPA
3219875	William Lee	CS	2000/1/1	M	111212	3.7
3219774	Emma Kim	CS	2007/2/7		223123	4.5
3219875	Noah Choi	CS	9999/9/9	M	111212	-0.5

- ```
SELECT S.advisorid, avg(S.gpa)
FROM student_table S
GROUP BY S.advisorid
```
  - ```
SELECT S.advisorid, sum(S.gpa)
FROM student_table S
WHERE S.birthdate >= 2000/1/1 AND S.birthdate < 2010/1/1
GROUP BY S.advisorid
```
-

# Getting data using API

- REST  
(Representational  
State Transfer) API

A screenshot of the Last.fm Web Services page. The page has a red header with the Last.fm logo, a search bar, and navigation links for Music, Listen, Events, Charts, and a Join button. The main content area is white. It starts with a section titled 'Last.fm Web Services' with links for API Introduction, Feeds, and Your API Accounts. Below this is a grey box with text explaining the API and a link to a social music playground. To the right is a 'Getting Started' section with a list of three steps: Get an API account, Read the documentation, and Join the API group. Further down is a 'Featured Applications' section with three items: 'My Music Habits', 'Tastebuds', and 'Scrobblr for Android & iOS'. At the bottom is an 'API Methods' section with three columns of methods: Album, Artist, Geo, Library, Tag, and Track.

lost.fm Music search Q Music Listen Events Charts Join

## Last.fm Web Services

[API Introduction](#) | [Feeds](#) | [Your API Accounts](#)

The Last.fm API allows anyone to build their own programs using Last.fm data, whether they're on the web, the desktop or mobile devices. [Find out more](#) about how you can start exploring **the social music playground** or just browse the list of methods below.


### Getting Started


Anyone is free to use the Last.fm API. you need to get going:


1. [Get an API account](#)
2. [Read the documentation](#)
3. [Join the API group](#)

### Featured Applications

[See more at build.last.fm »](#)

**My Music Habits**  
Detailed analysis of your listening habits.

**Tastebuds**  
Tastebuds is the best place to meet new people through music. Simply share your...

**Scrobblr for Android & iOS**  
Uses your phone's mic to i scrobble music to your pro

### API Methods

Album	Geo	Track
<a href="#">Album.addTags</a>	<a href="#">Geo.getTopArtists</a>	<a href="#">Track.addTags</a>
<a href="#">Album.getInfo</a>	<a href="#">Geo.getTopTracks</a>	<a href="#">Track.getCorrection</a>
<a href="#">Album.getTags</a>		<a href="#">Track.getInfo</a>
<a href="#">Album.getTopTags</a>	<b>Library</b>	<a href="#">Track.getSimilar</a>
<a href="#">Album.removeTag</a>	<a href="#">Library.getArtists</a>	<a href="#">Track.getTags</a>
<a href="#">Album.search</a>	<b>Tag</b>	<a href="#">Track.getTopTags</a>
	<a href="#">Tag.getInfo</a>	<a href="#">Track.love</a>
<b>Artist</b>	<a href="#">Tag.getSimilar</a>	<a href="#">Track.removeTag</a>
<a href="#">Artist.addTags</a>	<a href="#">Tag.getTopAlbums</a>	<a href="#">Track.scrobble</a>
<a href="#">Artist.getCorrection</a>	<a href="#">Tag.getTopArtists</a>	<a href="#">Track.search</a>
<a href="#">Artist.getInfo</a>	<a href="#">Tag.getTopTags</a>	<a href="#">Track.unlove</a>
<a href="#">Artist.getSimilar</a>		<a href="#">Track.updateNowPlaying</a>
<a href="#">Artist.getTopTags</a>		

# REST (Last.fm): album info

## /2.0/?method=album.getInfo&api\_key=KEY&artist=Cher&album=Believe

```
<album>
  <name>Believe</name>
  <artist>Cher</artist>
  <id>2026126</id>
  <mbid>61bf0388-b8a9-48f4-81d1-7eb02706dfb0</mbid>
  <url>http://www.last.fm/music/Cher/Believe</url>
  <releasedate>6 Apr 1999, 00:00</releasedate>
  <image size="small">...</image>
  <image size="medium">...</image>
  <image size="large">...</image>
  <listeners>47602</listeners>
  <playcount>212991</playcount>
  <toptags>
    <tag>
      <name>pop</name>
      <url>http://www.last.fm/tag/pop</url>
    </tag>
  </toptags>
</album>
```

**lost.fm**  [Music](#) [Listen](#) [Events](#) [Charts](#) [Join](#)

**Account**  
Your API accounts  
[Add API account](#)

**API Guides**  
[Introduction](#)  
[User Authentication](#)  
[Scrobbling](#)  
[Radio API](#)  
[Feeds](#)  
[Playlists API](#)  
[Tools](#)  
[REST requests](#)  
[XML-RPC requests](#)  
[Error codes](#)  
[Terms of Service](#)

**API Methods**  
**Album**  
[album.addTags](#)  
[album.getInfo](#)  
[album.getTags](#)  
[album.getTopTags](#)  
[album.removeTag](#)  
[album.search](#)  
**Artist**  
[artist.addTags](#)  
[artist.getCorrection](#)  
[artist.getInfo](#)  
[artist.getSimilar](#)  
[artist.getTags](#)  
[artist.getTopAlbums](#)  
[artist.getTopTags](#)  
[artist.getTopTracks](#)  
[artist.removeTag](#)  
[artist.search](#)

[Last.fm Web Services](#)  
**album.getInfo**  

Get the metadata and tracklist for an album on Last.fm using the album name or a musicbrainz id.

**Example URLs**

**JSON:** /2.0/?method=album.getInfo&api\_key=YOUR\_API\_KEY&artist=Cher&album=Believe&format=json  
**XML:** /2.0/?method=album.getInfo&api\_key=YOUR\_API\_KEY&artist=Cher&album=Believe

**Params**

**artist** (Required (unless mbid)) : The artist name  
**album** (Required (unless mbid)) : The album name  
**mbid** (Optional) : The musicbrainz id for the album  
**autocorrect[0|1]** (Optional) : Transform misspelled artist names into correct artist names, returning the correct instead. The corrected artist name will be returned in the response.  
**username** (Optional) : The username for the context of the request. If supplied, the user's playcount for this album is included in the response.  
**lang** (Optional) : The language to return the biography in, expressed as an ISO 639 alpha-2 code.  
**api\_key** (Required) : A Last.fm API key.

**Auth**

This service does not require authentication.

**Sample Response**

```
<album>
  <name>Believe</name>
  <artist>Cher</artist>
  <id>2026126</id>
  <mbid>61bf0388-b8a9-48f4-81d1-7eb02706dfb0</mbid>
  <url>http://www.last.fm/music/Cher/Believe</url>
  <releasedate>6 Apr 1999, 00:00</releasedate>
  <image size="small">...</image>
  <image size="medium">...</image>
  <image size="large">...</image>
  <listeners>47602</listeners>
</album>
```

# Getting data using API

---

```
import json
import requests
url =
'http://ws.audioscrobbler.com/2.0/?method=album
.getinfo&api_key=KEY&artist=Cher&album=Believe&
format=json'
data = requests.get(url).text
parsed_data = json.loads(data)
```

---

# Web Scrapping (HTML DOM)

---

Document Object Model: the hierarchical structure of HTML

```
<html>
  <head>
    <title> Data Science </title>
  </head>

  <body>
    <h1>Hello World!</h1>
    <p> Welcome to COSE 471 Data Science </p>
  </body>
</html>
```

---

# Sources of Data

---

- Proprietary data sources
  - Government data sets
  - Academic data sets
  - Web search /Scraping
  - Sensor data
  - Crowdsourcing
  - Sweat equity
-

# Proprietary Data Sources

---

Facebook, Google, Amazon, Blue Cross, etc. have exciting user/transaction/log data sets.

Most organizations have/should have internal data sets of interest to their business.

Companies sometimes release rate-limited APIs, including Twitter and Google.

---



# Proprietary Data Sources

---

However, getting outside access to proprietary corporate data is usually difficult for two reasons:

- Business issues: fear of helping competitors
- Privacy issues: fear of offending customers

Case Study: 2006 AOL search log release

- What business and privacy issues were there?
-

# Government Data Sources

---

- City, State, and Federal governments are increasingly committed to open data.
  - Data.gov has over 100,000 open data sets!
  - The Freedom of Information Act (FOI) enables you to ask if something is not open.
  - Preserving privacy is often the big issue in whether a data set can be released.
-

# Academic Data Sets

---

- Making data available is now a requirement for publication in many fields.
  - Expect to be able to find economic, medical, demographic, and meteorological data if you look hard enough.
  - Track down from relevant papers.
    - Find links to data, if none, ask authors.
-

# Web Search/Scraping

---

*Scraping* is the fine art of stripping text/data from a webpage.

Libraries exist in Python to help parse/scrape the web, but first search:

- Are APIs available from the source?
- Did someone previously write a scraper?

Terms of service limit what you can legally do.

---

# Available Data Sources

---

- Bulk Downloads: e.g. Wikipedia, IMDB, Million Song Database.
- API access: e.g. New York Times, Twitter, Facebook, Google.

Be aware of limits and terms of use.

---

# Crowdsourcing

---

Many amazing open data resources have been built up by teams of contributors:

- Wikipedia/Freebase
- IMDB

Crowdsourcing platforms like Amazon Turk and CrowdFlower enable you to pay for armies of people to help you gather data, like human annotation.

---

# Cleaning Data: Garbage In, Garbage Out

---

Many issues can arise in cleaning data for analysis:

- Distinguishing errors from artifacts.
  - Data compatibility / unification.
  - Imputation of missing values.
  - Estimating unobserved (zero) counts.
  - Outlier detection.
-

# Errors vs. Artifacts

---

- Data **errors** represent information that is fundamentally lost in acquisition.
- **Artifacts** are systematic problems arising from data processing.

The key to detecting artifacts is the **sniff test**, examining the product closely enough to get a whiff of something bad.

---



# Data Compatibility

---

Data needs to be carefully massaged to make “apple to apple” comparisons:

- Unit conversions
  - Number / character code representations
  - Name unification
  - Time/date unification
  - Financial unification
-

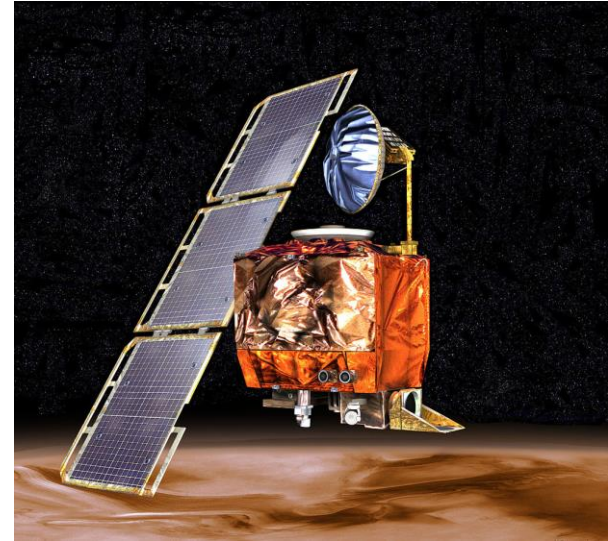
# Unit Conversions

---

NASA's Mars Climate Orbiter lost in 1999 due to a unit conversion issue between metric unit (kg, m) and US customary unit (lb, ft).

- Even sticking to the metric system has potential inconsistencies: cm, m, km?
- Bimodal distributions can indicate trouble
- Z-scores are dimensionless quantities.

Vigilance in data integration is essential.



# Normalization and Z-scores

---

It is critical to normalize different variables to make their range/distribution comparable.

Z-scores are computed:

$$Z_i = (X_i - \bar{X}) / \sigma$$

Z-scores of height measured in inches is the same as height measured in miles.

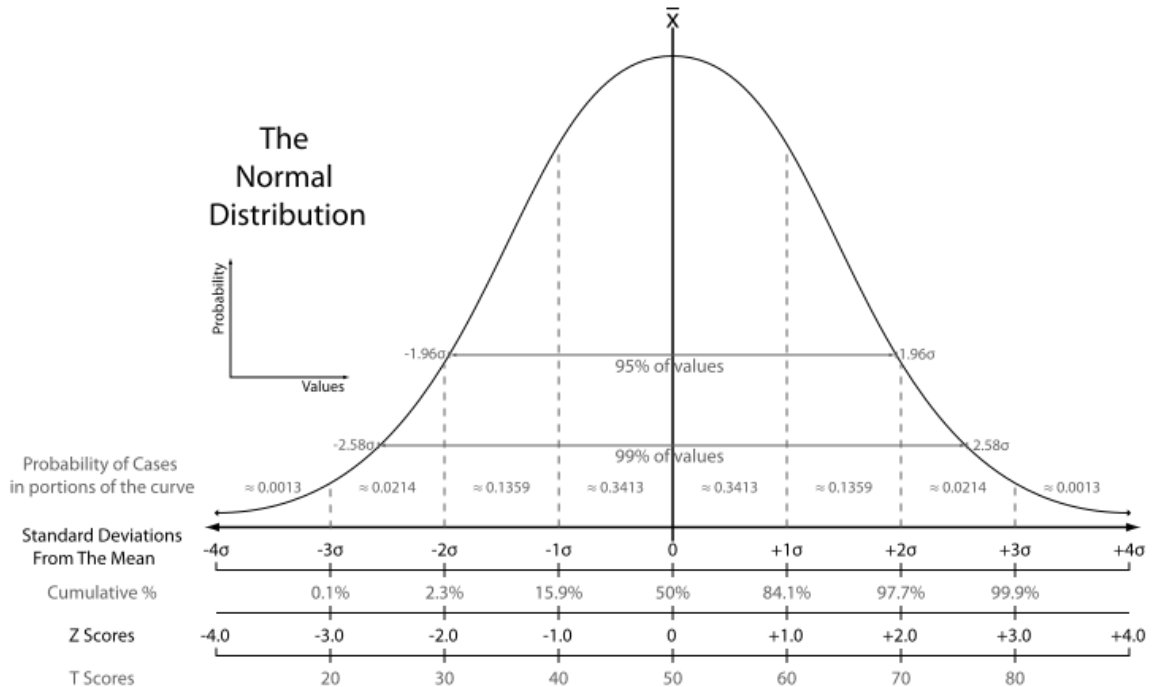
Z-scores have mean 0 and sigma=1.

---

# Z-score Examples

The sign identifies if it is above/below the mean.

Thus Z-scores of different variables are of comparable magnitude.



# Number / Character Representations

---

The Ariane 5 rocket exploded in 1996 due to a bad 64-bit float to 16-bit integer conversion.

- Avoid integer approximation of real numbers
- Measurements should generally be decimal numbers
- Counts should be integers.
- Fractional quantities should be decimal, not (q,r) like (pounds,ounce) or (feet,inches).

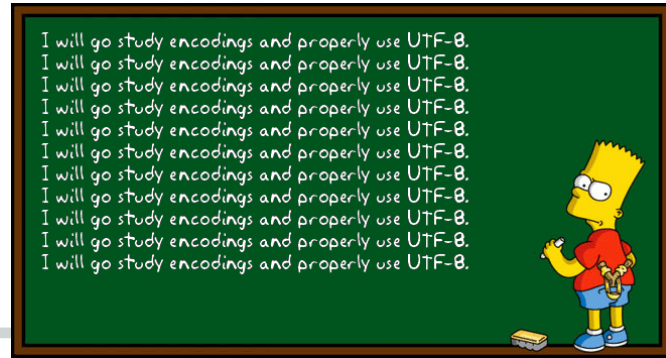


# Character Representations

A particularly nasty cleaning issue in textual data is unifying character code representations:

- ISO 8859-1 is a single byte code for ASCII
- UTF-8 is a multibyte encoding for all Unicode characters.

Unicode font, UTF8 format	Unicode font, XXX... format
搜索简体中文网页	????????
Recherche avancée	Recherche avancée
網路畫廊, 含中、港、澳參展作品	????????????????
โทรศัณยท์	????????????????
ウェブ全体から	??????
kehren Sie zur Suche zurück	kehren Sie zur Suche zurück
Сделайте Google стартовой	???????? Google ?????
إخترق بحث أقل وقت مطالعة أطول	?????? ??? ??? ??? ????? ????



# Name Unification

---

Same person appears on the web as:

*(Steve/Steven/S.) (S./Sol|\_) (Skiena/Skeina/Skienna)*

- Use simple transformations to unify names, like lower case, removing middle names or use initials instead, etc.

Tradeoff between false positives and negatives.

---

# Time / Date Unification

---

Aligning temporal events from different datasets/systems can be problematic.

- Use Coordinated Universal Time (UTC), a modern standard subsuming GMT.
- Financial time series are tricky because of weekends and holidays: how do you correlate stock prices and temperatures?

September 1752						
Su	M	Tu	W	Th	F	Sa
-	-	1	2	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30



# Financial Unification

---

- Currency conversion uses exchange rates.
  - Use returns / percentage change instead of absolute price changes.
  - Correct stock prices for splits and dividends.
  - The time value of money needs correction for inflation for fair long-term comparisons.
-

# Dealing with Missing Data

---

An important aspect of data cleaning is properly representing missing data:

- What is the year of death of a living person?
- What about a field left blank or filled with an obviously outlandish value?
- The frequency of events too rare to see?

Setting such values to zero is generally wrong

---

# Imputing Missing Values

---

With enough training data, one might drop all records with missing values, but we may want to use the model on records with missing fields.

Often it is better to estimate or **impute** missing values instead of leaving them blank.

---

# Imputation Methods

---

- *Mean value imputation* - leaves mean same.
  - *Heuristic-based imputation* – a good guess for your death year is birth year+80.
  - *Random value imputation* - repeatedly selecting random values permits statistical evaluation of the impact of imputation.
-

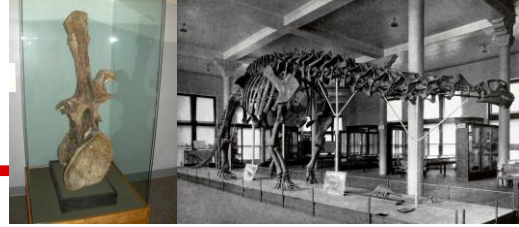
# Imputation Methods

---

- *Imputation by nearest neighbor* – identify closest record and use it to infer the missing values.
  - *Imputation by interpolation* - using linear regression to predict missing values works well if few fields are missing per record.
-

# Outlier Detection

---



The largest reported dinosaur vertebra is 50% larger than all others: presumably a data error.

- Look critically at the maximum and minimum values for all variables.
- Normally distributed data should not have large outliers,  $k$  sigma from the mean.

Fix why you have an outlier. Don't just delete.

---

# Detecting Outliers

---

- Visually, it is easy to detect outliers, but only in low dimensional spaces.
  - It can be thought of as an unsupervised learning problem, like clustering.
  - Points which are far from their cluster center are good candidates for outliers
-