Data Science

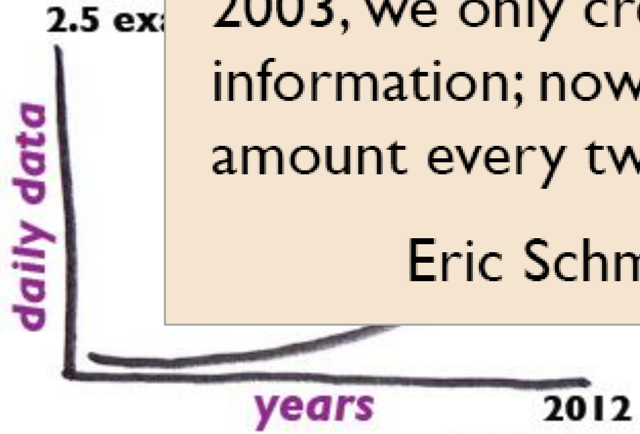
Lecture 1: Overview / Data Munging

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

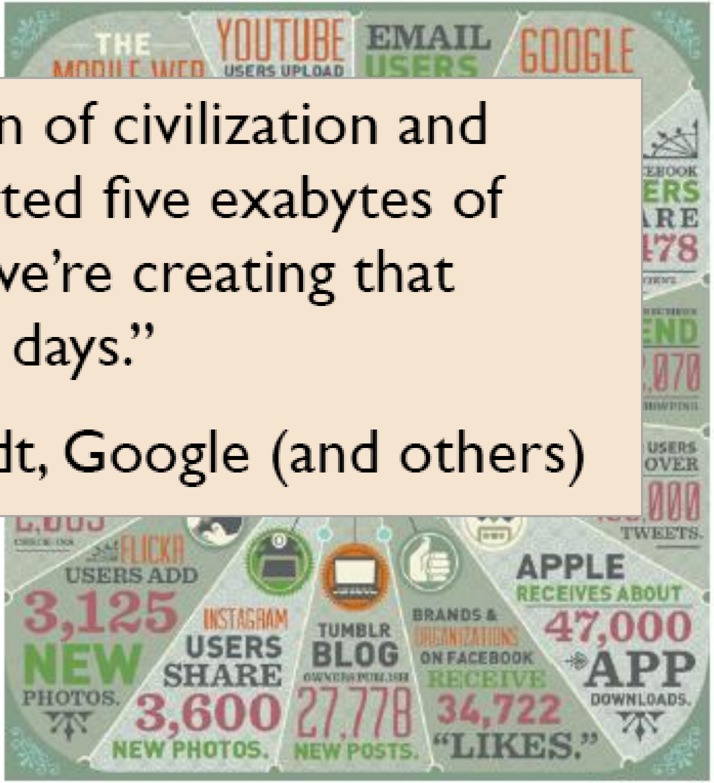
Big Data

"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)

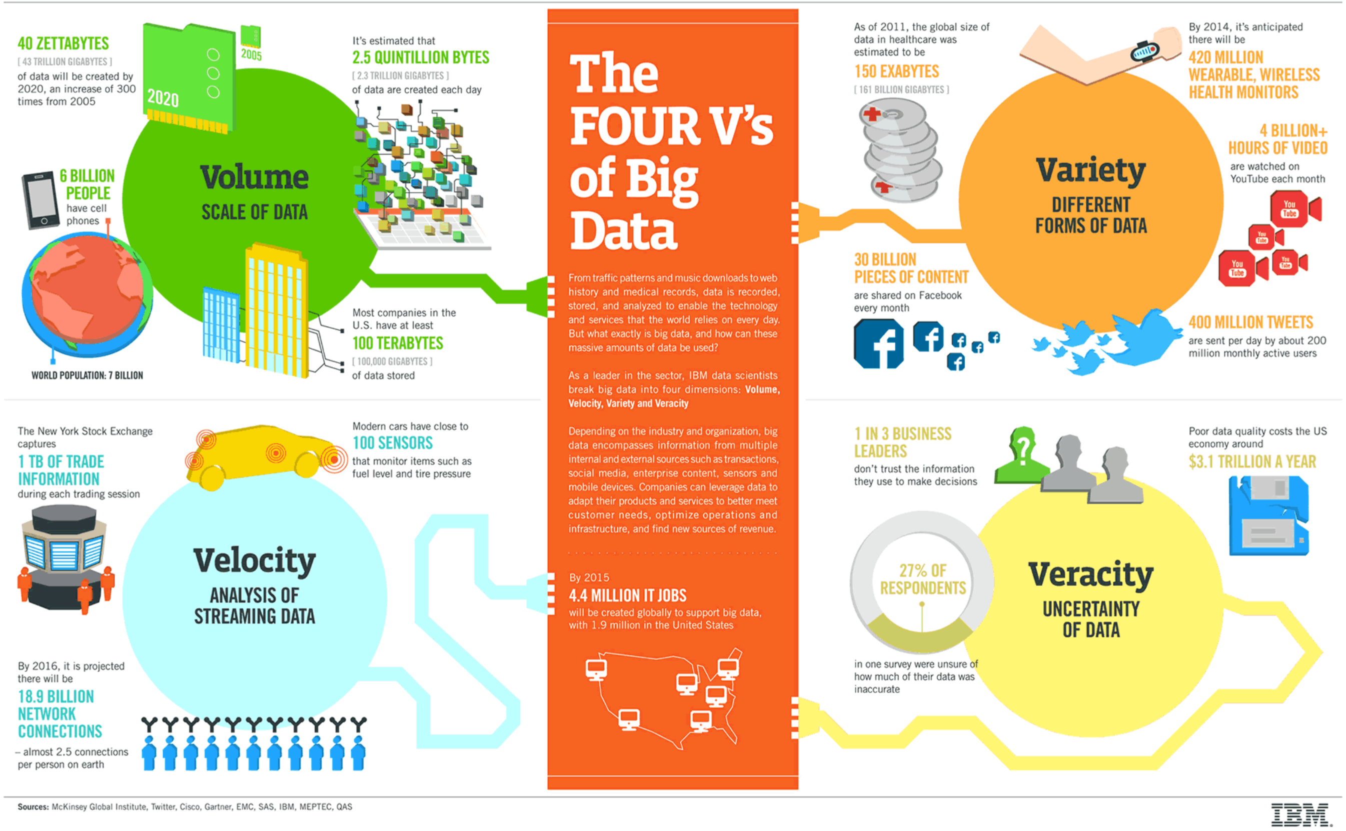


[IBMbigdata]



[Domo]

Big Data Explosion



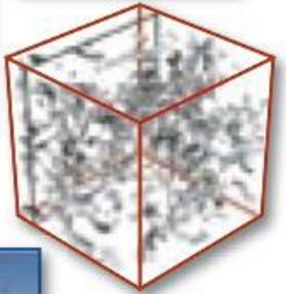
Big Data Challenges

- 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

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

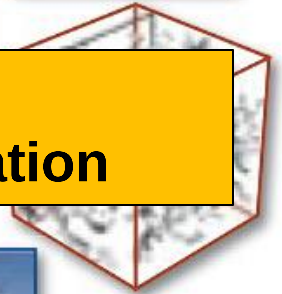


- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments
    or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files
    using data management and statistics

Jim Gray, Microsoft

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

Jim Gray, Microsoft

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

When 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."

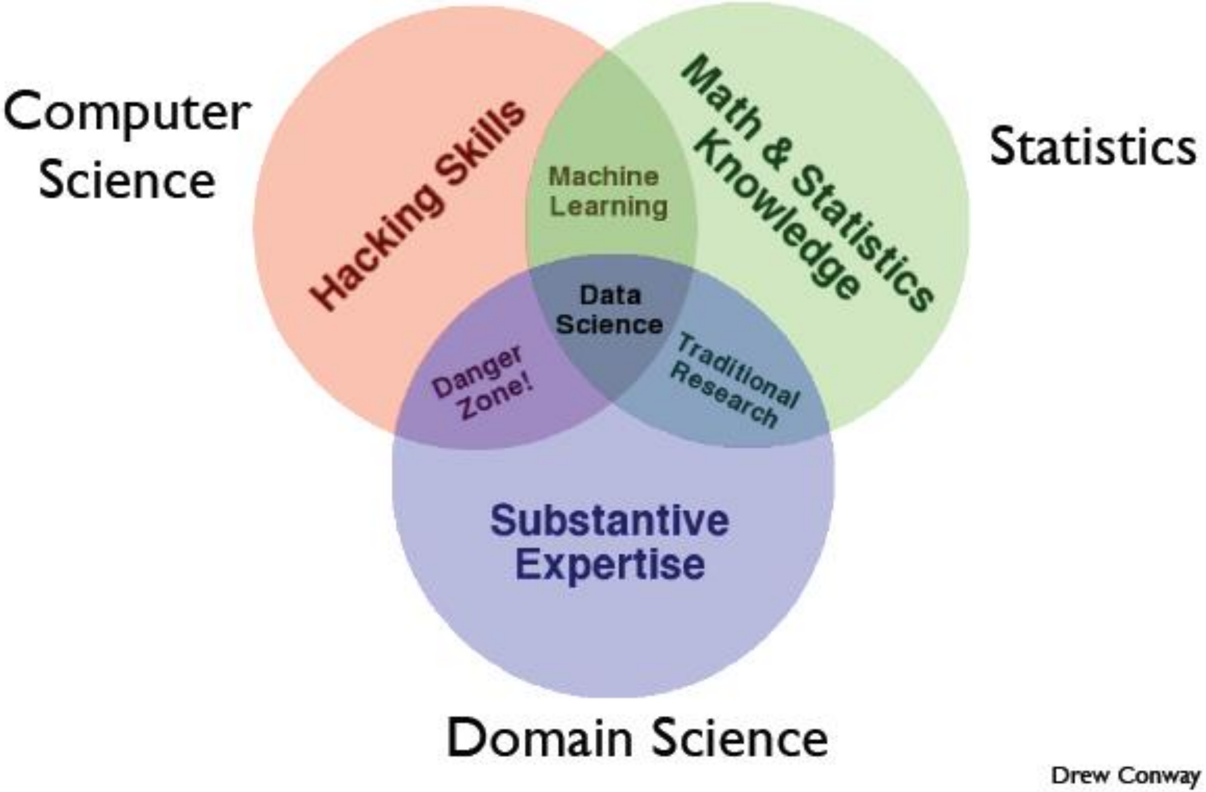
70 Harvard Business Review **October 2012**

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

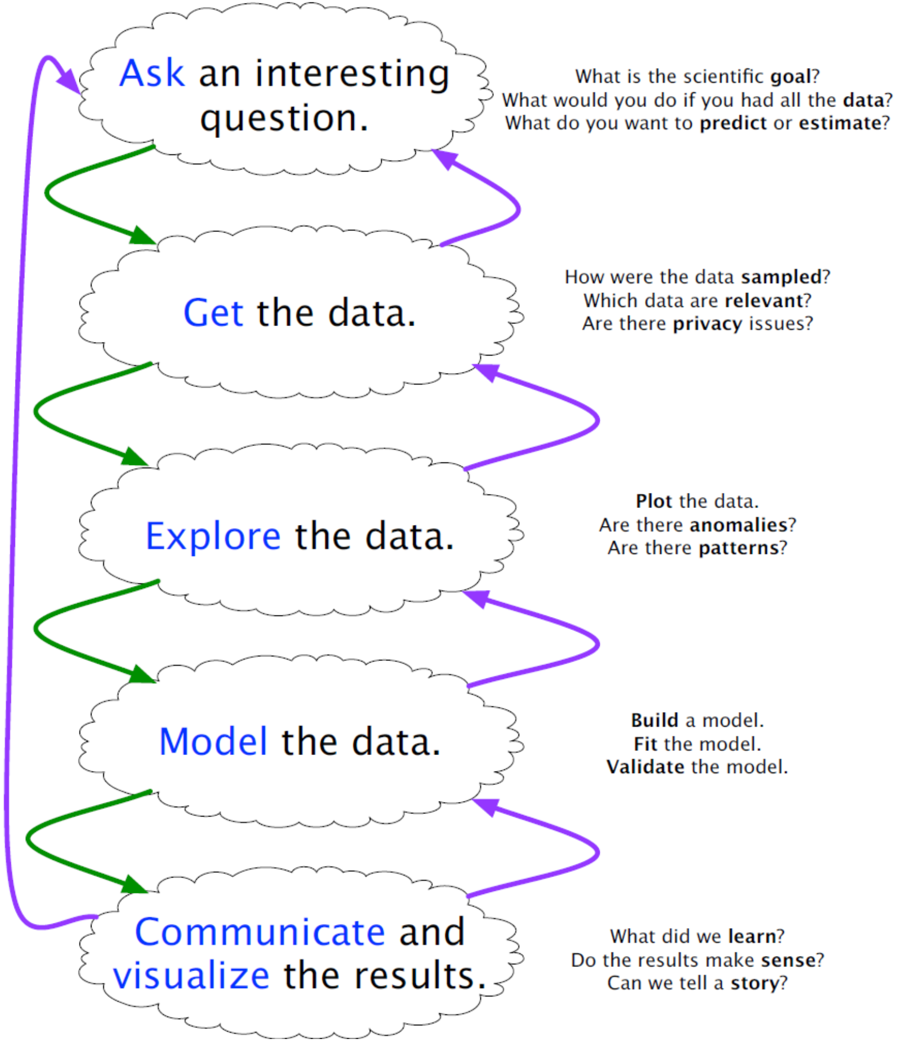
"Data Scientist = statistician + programmer + coach + storyteller + artist"

\- 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

ery 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?

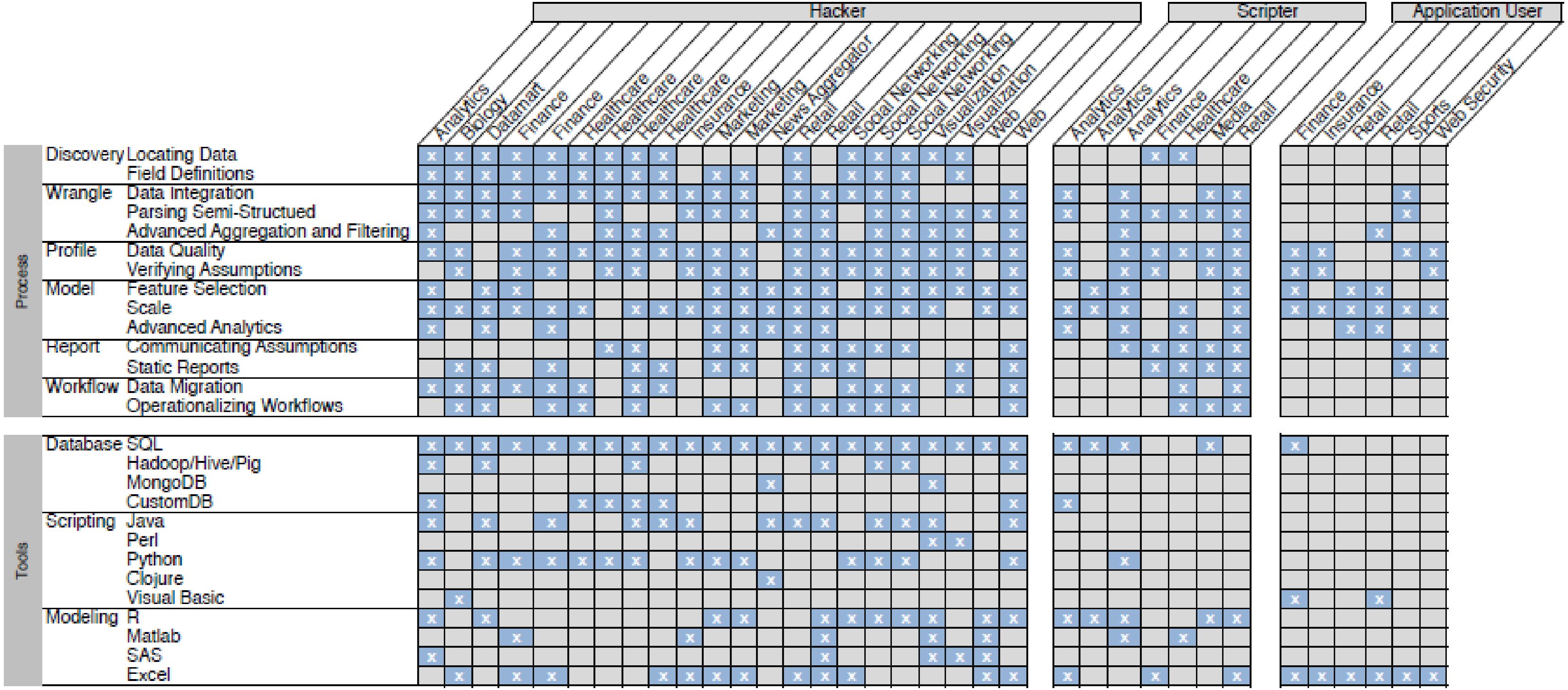


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* 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*

SEP 27, 2016 @ 01:30 AM 108,681 VIEWS

The Little Black Book of Bill

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.

Ad closed by Google
Report this ad
Why this ad? ⓘ

- 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.

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	View Jobs
#2	Full Stack Engineer	$101,794	4.3/5	11,252	View Jobs
#3	Data Scientist	$120,000	4.1/5	10,071	View Jobs
#4	Devops Engineer	$120,095	4.2/5	8,548	View Jobs
#5	Strategy Manager	$140,000	4.2/5	6,977	View Jobs
#6	Machine Learning Engineer	$130,489	4.3/5	6,801	View Jobs
#7	Data Engineer	$113,960	4.0/5	11,821	View Jobs
#8	Software Engineer	$116,638	3.9/5	64,155	View Jobs
#9	Java Developer	$107,099	4.1/5	10,201	View Jobs
#10	Product Manager	$125,317	4.0/5	17,725	View Jobs

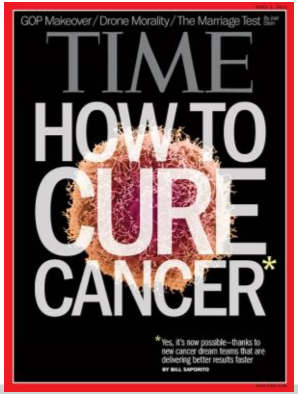
Source: https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

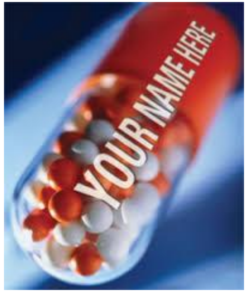
Is Data Science Still a Rising Career?

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!





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^{st} 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	Computer Scientists
• $8/13 \approx 0.62$	• $8/13 = 0.61538461538$
• Care what it means	• Care what number is
• Nothing is completely true	• Either correct or wrong
• Meaning	• 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

IMDb: Movie Data (https://www.imdb.com/)

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

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_datetim	passenger_c	trip_time_	trip_distance	pickup_longitud	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	9FD8F69F0804BD	CMT	1	1/6/13 0:18	1/6/13 0:22	1	259	1.5	-74.006683	40.731781	-73.994499	40.75066
10	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1	1/5/13 18:49	1/5/13 18:54	1	282	1.1	-74.004707	40.73777	-74.009834	40.726002
11	...												
12													
13													
14	Fare data, 2013 ->												
15													
16	medallion	hack_license	vendor_id	pickup_datetime	fare_amount	surcharge	mta_tax	tip_amou	tolls_amount	total_amount			
17	89D227B655E5C82AEC	BA96DE419E7116	CMT	1/1/13 15:11	6.5	0	0.5	0	0	7			
18	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1/6/13 0:18	6	0.5	0.5	0	0	7			
19	0BD7C8F5BA12B88E0B	9FD8F69F0804BD	CMT	1/5/13 18:49	5.5	1	0.5	0	0	7			

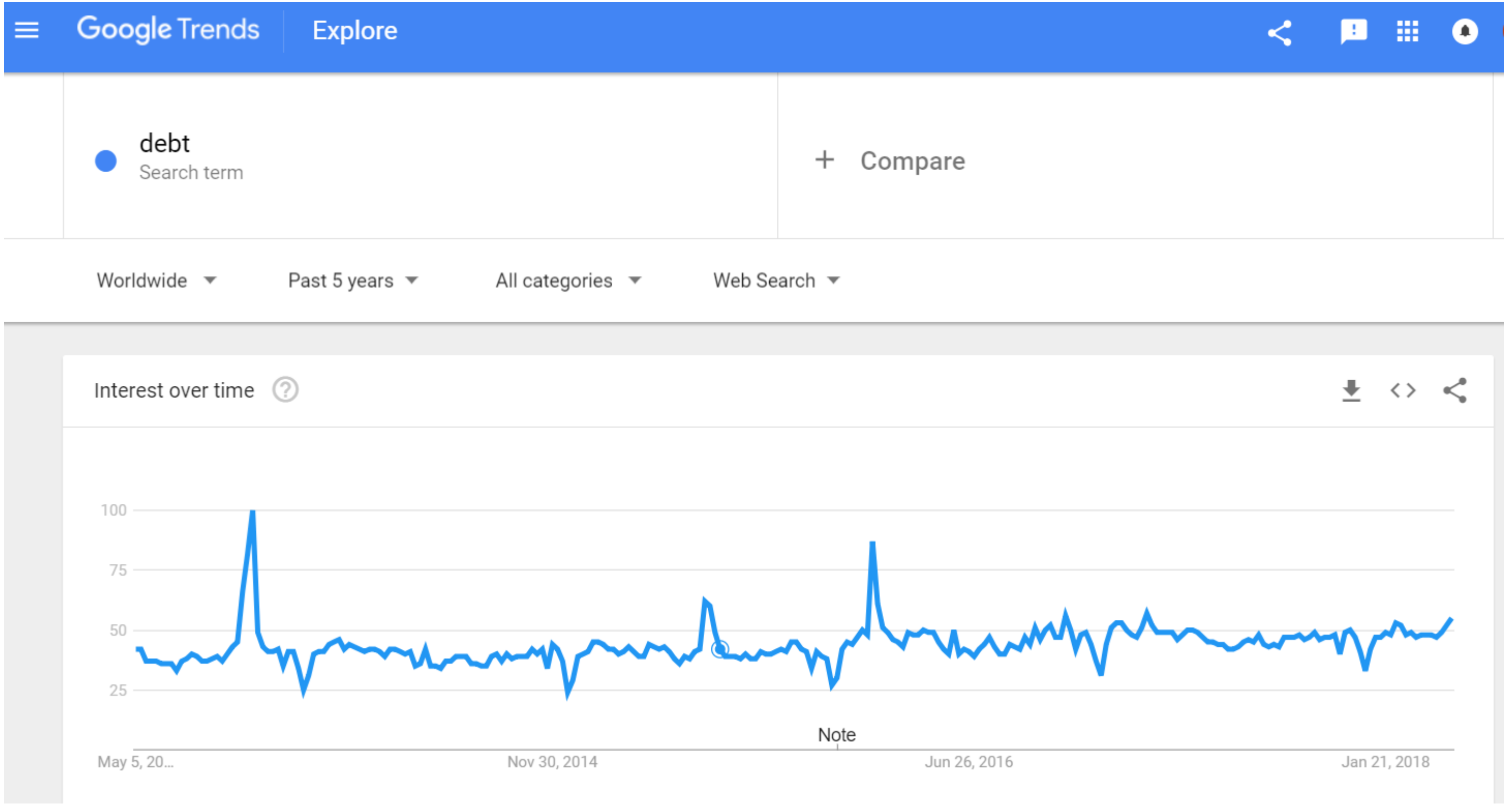
Taxicab Questions

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

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

OPEN

Quantifying Trading Behavior in Financial Markets Using *Google Trends*

Tobias Preis[1]*, Helen Susannah Moat[2,3]* & H. Eugene Stanley[2]*

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

SUBJECT AREAS:

STATISTICAL PHYSICS, THERMODYNAMICS AND NONLINEAR DYNAMICS

APPLIED PHYSICS

COMPUTATIONAL SCIENCE

INFORMATION THEORY AND COMPUTATION

Received
25 February 2013

Accepted
3 April 2013

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.

An Example…Cont'd

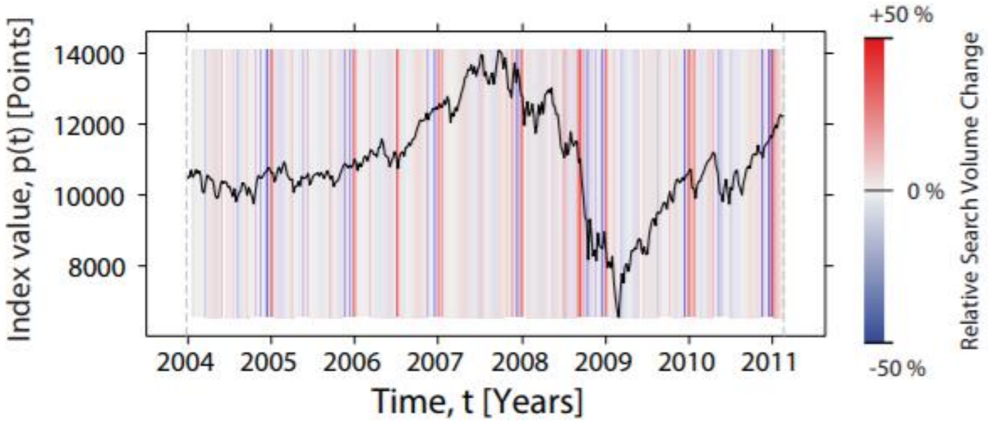


Figure 1 | Search volume data and stock market moves. Time series of closing prices *p(t)* of the *Dow Jones Industrial Average* (DJIA) on the first day of trading in each week *t* covering the period from 5 January 2004 until 22 February 2011. The color code corresponds to the relative search volume 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

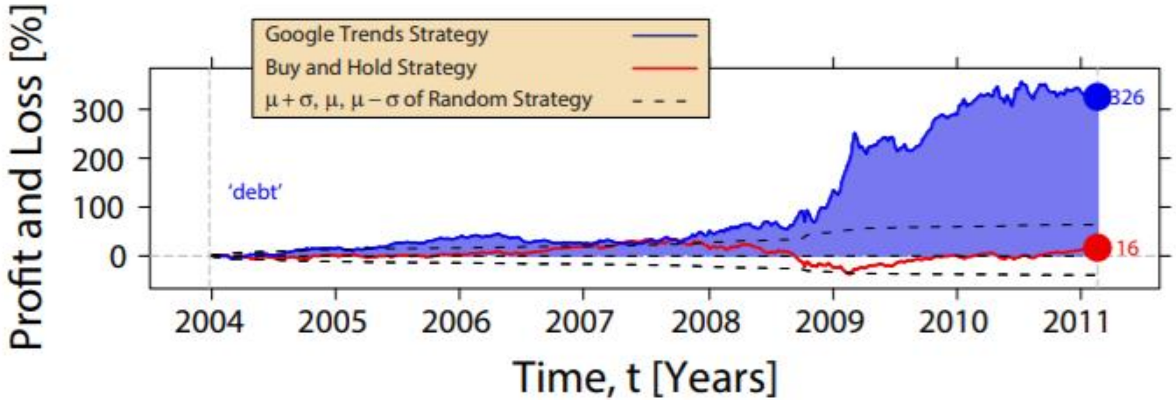
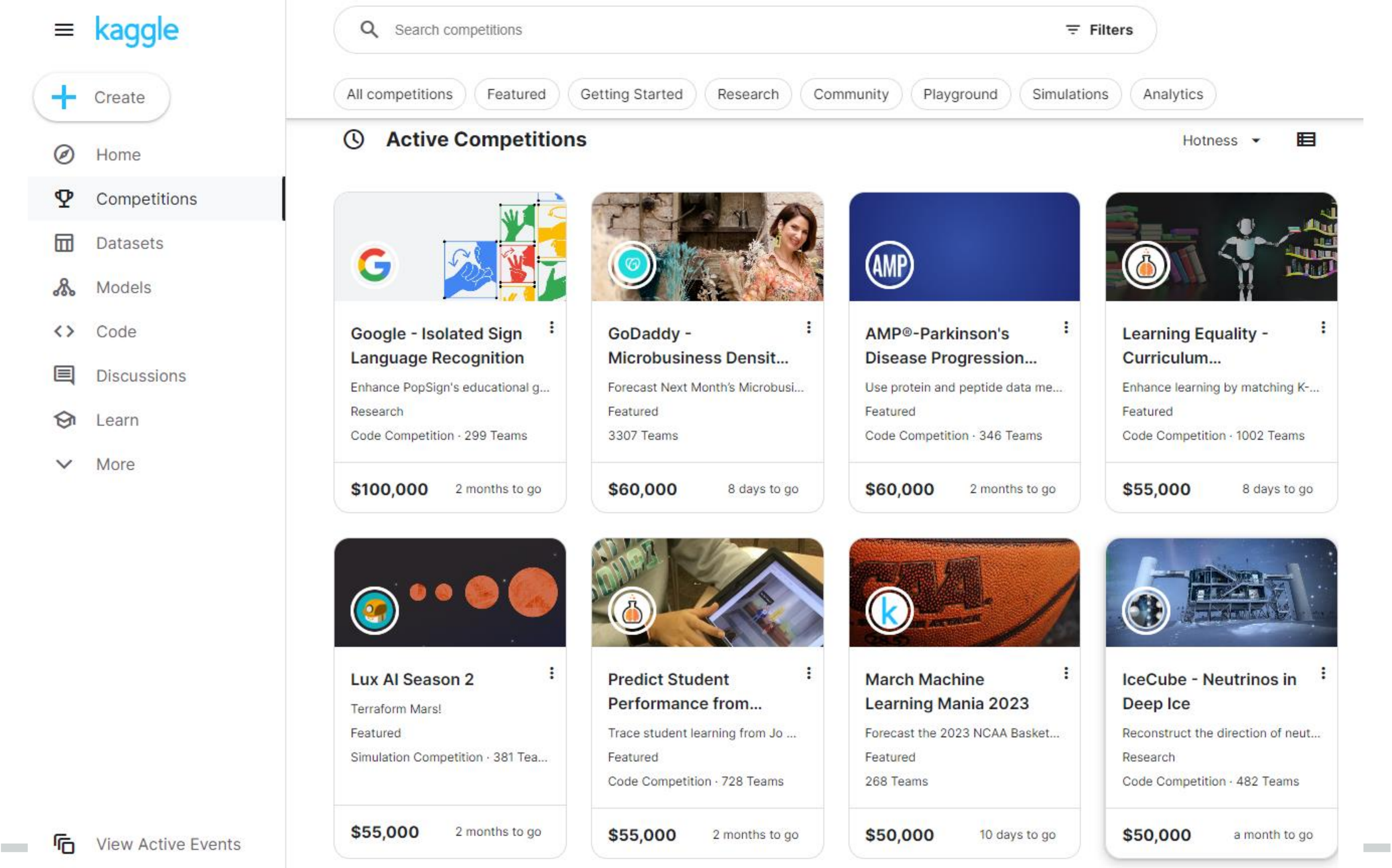


Figure 2 | Cumulative performance of an investment strategy based on *Google Trends* data. Profit and loss for an investment strategy based on the volume of the search term *debt*, the best performing keyword in our analysis, with $\Delta t = 3$ weeks, plotted as a function of time (*blue line*). This 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 Science Challenges

- Kaggle (https://www.kaggle.com/)
  - a platform for predictive modelling and analytics competitions.
- DreamChallenge (http://dreamchallenges.org/)
  - poses fundamental questions about systems biology and translational medicine.


kaggle
Search competitions
Filters
Create
Home
Competitions
Datasets
Models
Code
Discussions
Learn
More
View Active Events
All competitions
Featured
Getting Started
Research
Community
Playground
Simulations
Analytics
Active Competitions
Hotness
Google - Isolated Sign Language Recognition
Enhance PopSign's educational g...
Research
Code Competition · 299 Teams
$100,000
2 months to go
GoDaddy - Microbusiness Densit...
Forecast Next Month's Microbusi...
Featured
3307 Teams
$60,000
8 days to go
AMP®-Parkinson's Disease Progression...
Use protein and peptide data me...
Featured
Code Competition · 346 Teams
$60,000
2 months to go
Learning Equality - Curriculum...
Enhance learning by matching K-...
Featured
Code Competition · 1002 Teams
$55,000
8 days to go
Lux AI Season 2
Terraform Mars!
Featured
Simulation Competition · 381 Tea...
$55,000
2 months to go
Predict Student Performance from...
Trace student learning from Jo ...
Featured
Code Competition · 728 Teams
$55,000
2 months to go
March Machine Learning Mania 2023
Forecast the 2023 NCAA Basket...
Featured
268 Teams
$50,000
10 days to go
IceCube - Neutrinos in Deep Ice
Reconstruct the direction of neut...
Research
Code Competition · 482 Teams
$50,000
a month to go

Research Code Competition

Google - Isolated Sign Language Recognition

Enhance PopSign's educational games for learning ASL

Google · 299 teams · 2 months to go (2 months to go until merger deadline)

$100,000
Prize Money

Overview Data Code Discussion Leaderboard Rules

Join Competition

Overview

Description
Evaluation
Timeline
Prizes
Code Requirements
Acknowledgements
Data Card

Goal of the Competition

The goal of this competition is to classify isolated American Sign Language (ASL) signs. You will create a TensorFlow Lite model trained on labeled landmark data extracted using the MediaPipe Holistic Solution.

Your work may improve the ability of PopSign* to help relatives of deaf children learn basic signs and communicate better with their loved ones.^

**PopSign is an app developed by the Georgia Institute of Technology and the National Technical Institute for the Deaf at Rochester Institute of Technology. The app is available in beta on Android and iOS.*

^We cannot guarantee the competition will benefit the competitors or the disabled community directly.

Launch
11 days ago

Close
2 months

2 months
Rules Acceptance Deadline

299	319	1,707
Teams	Competitors	Entries

Points This competition awards ranking points

Tiers This competition counts towards tiers

Tags multiclass classification video classification custom metric

Citation Ashley Chow, eknight7, Glenn, Mark Sherwood, Phil Culliton, Sam Sepah, Sohier Dane, Thad Starner. (2023). Goog... 99 Cite

Google - Isolated Sign Language Recognition

Enhance PopSign's educational games for learning ASL

Google · 299 teams · 2 months to go (2 months to go until merger deadline)

Overview Data Code Discussion Leaderboard Rules

Join Competition ···

Leaderboard

Raw Data Refresh

Search leaderboard

Public Private

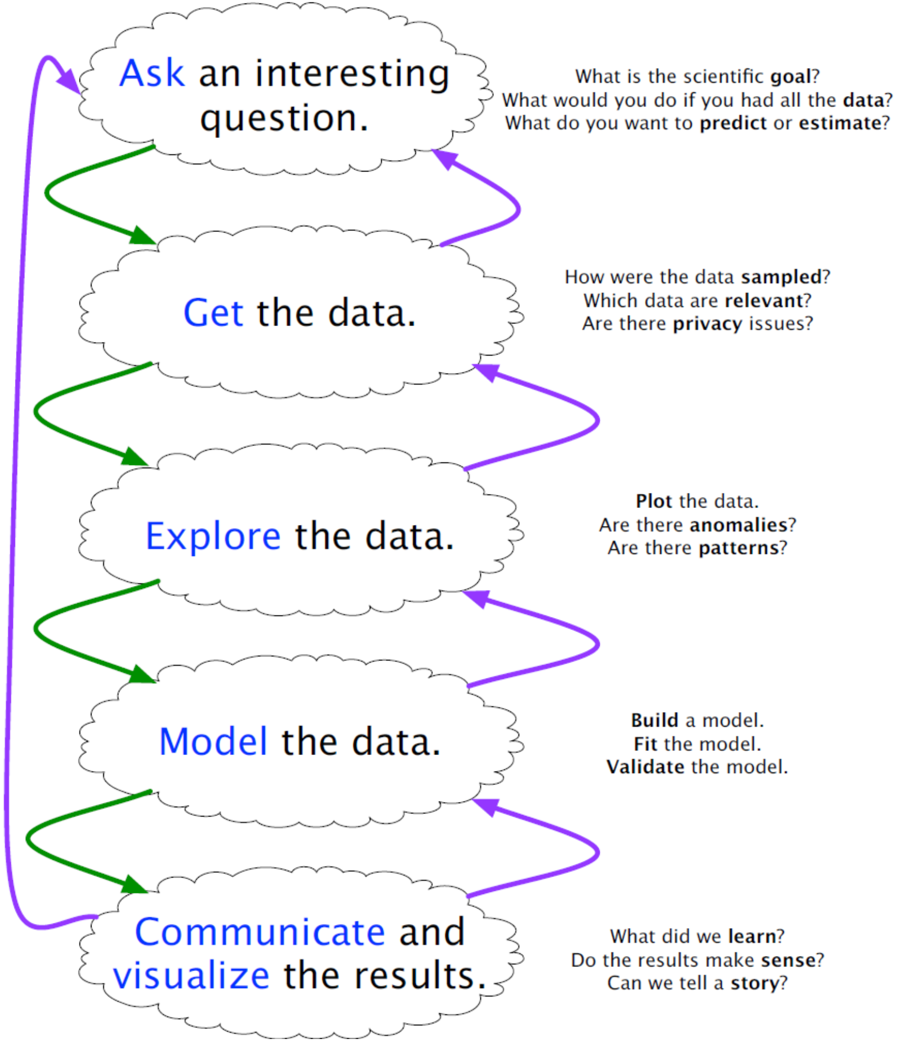
This leaderboard is calculated with approximately 50% of the test data. The final results will be based on the other 50%, so the final standings may be different.

Prize Contenders

#	Team	Members		Score	Entries	Last	Join
1	overtime submission			0.67	14	5h	
2	takuya			0.66	9	9h	
3	siwooyong			0.65	15	4h	
4	shadow_in_mirror			0.65	24	33m	
5	Rohith Ingilela			0.64	19	8h	

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.

Languages for Data Science

- *Python*: contains libraries and features (e.g regular expressions) for easier munging.

- *R*: programming language of statisticians.

- *Matlab*: fast and efficient matrix operations.
- *Java/C++*: language for Big Data systems.

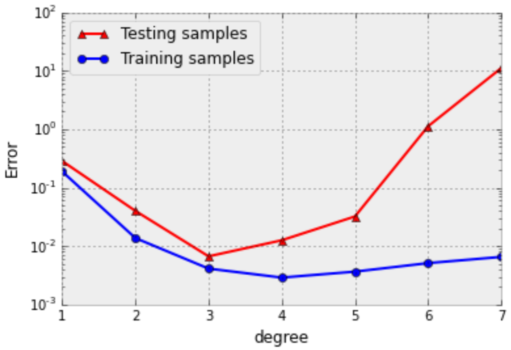
- *Excel*: bread and butter tool for exploration.

Notebook Environments

Result of data science project should be a computable notebook tying together the code, data, computational results, and written analysis.

- reproducible
- tweakable
- documented

In [40]:
```
degrees = range(1, 8)
errors = np.array([regressor3(d) for d in degrees])
plt.plot(degrees, errors[:, 0], marker='^', c='r', label='Testing samples
plt.plot(degrees, errors[:, 1], marker='o', c='b', label='Training sample
plt.yscale('log')
plt.xlabel("degree"); plt.ylabel("Error")
  = plt.legend(loc='best')
```



By sweeping the degree we discover two regions of model performance:

- **Underfitting** (degree < 3): Characterized by the fact that the testing error will get lower if we increase the model capacity.
- **Overfitting** (degree > 3): Characterized by the fact the testing will get higher if we increase the model capacity. Note, that the training error is getting lower or just staying the same!.

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, Lee Sedol, CS, 2000/1/1, M, 111212, 3.7
3219774, Alpha Go, CS, 2007/2/7,, 223123, 4.5
3219875, Hong Gildong, CS, 9999/9/9, M, 111212, -0.5
:
```

StudentID	Name	Dept	Birthdate	Gender	AdvisorID	GPA
3219875	Lee Sedol	CS	2000/1/1	M	111212	3.7
3219774	Alpha Go	CS	2007/2/7		223123	4.5
3219875	Hong Gildong	CS	9999/9/9	M	111212	-0.5

- XML (eXtensible Markup Language)

```
<Document Element>
        <Student Table>
            <Student>
                <StudentID> 3219875 </StudentID>
                <Name> Lee Sedol </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> Alpha Go </Name>
                <Dept> CS </Dept>
                        :
```

- JSON (JavaScript Object Notation)


```
{
    “sedol”: {
                    “StudentID”: 3219875,
                    “Name”: “Lee Sedol”,
                    “Dept”: “CS”,
                    “Birthdate”: 2000/1/1,
                    “Gender”: “M”,
                    “AdvisorID”: 111212,
                    “GPA”: 3.7
                    }
    “alpha”: {
                    “StudentID”: 3219774
                    “Name”: “Alpha Go”
                    “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, Lee Sedol, CS, 2000/1/1, M, 111212, 3.7
3219774, Alpha Go, CS, 2007/2/7,, 223123, 4.5
3219875, Hong Gildong, 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	Lee Sedol	CS	2000/1/1	M	111212	3.7
3219774	Alpha Go	CS	2007/2/7		223123	4.5
3219875	Hong Gildong	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

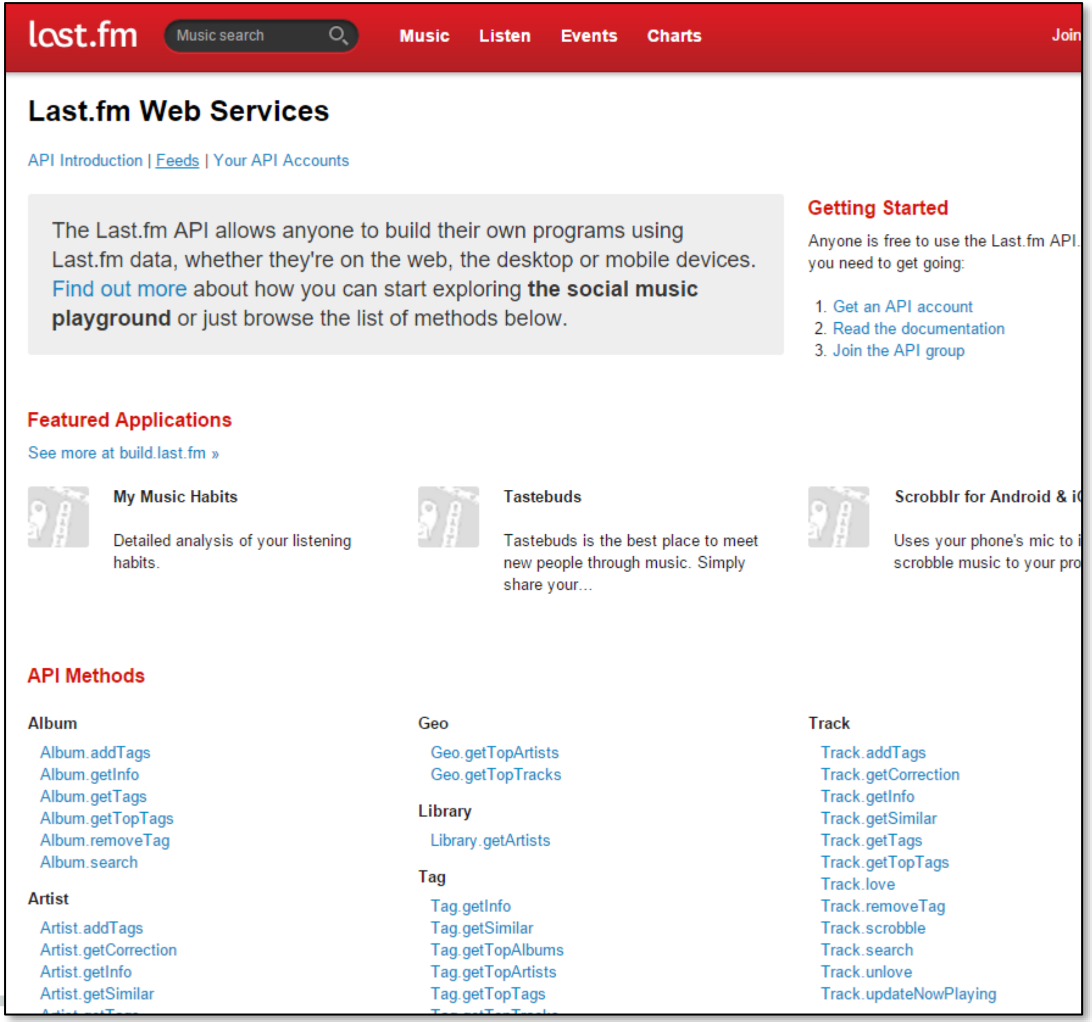
StudentID	Name	Dept	Birthdate	Gender	AdvisorID	GPA
3219875	Lee Sedol	CS	2000/1/1	M	111212	3.7
3219774	Alpha Go	CS	2007/2/7		223123	4.5
3219875	Hong Gildong	CS	9999/9/9	M	111212	-0.5

- 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

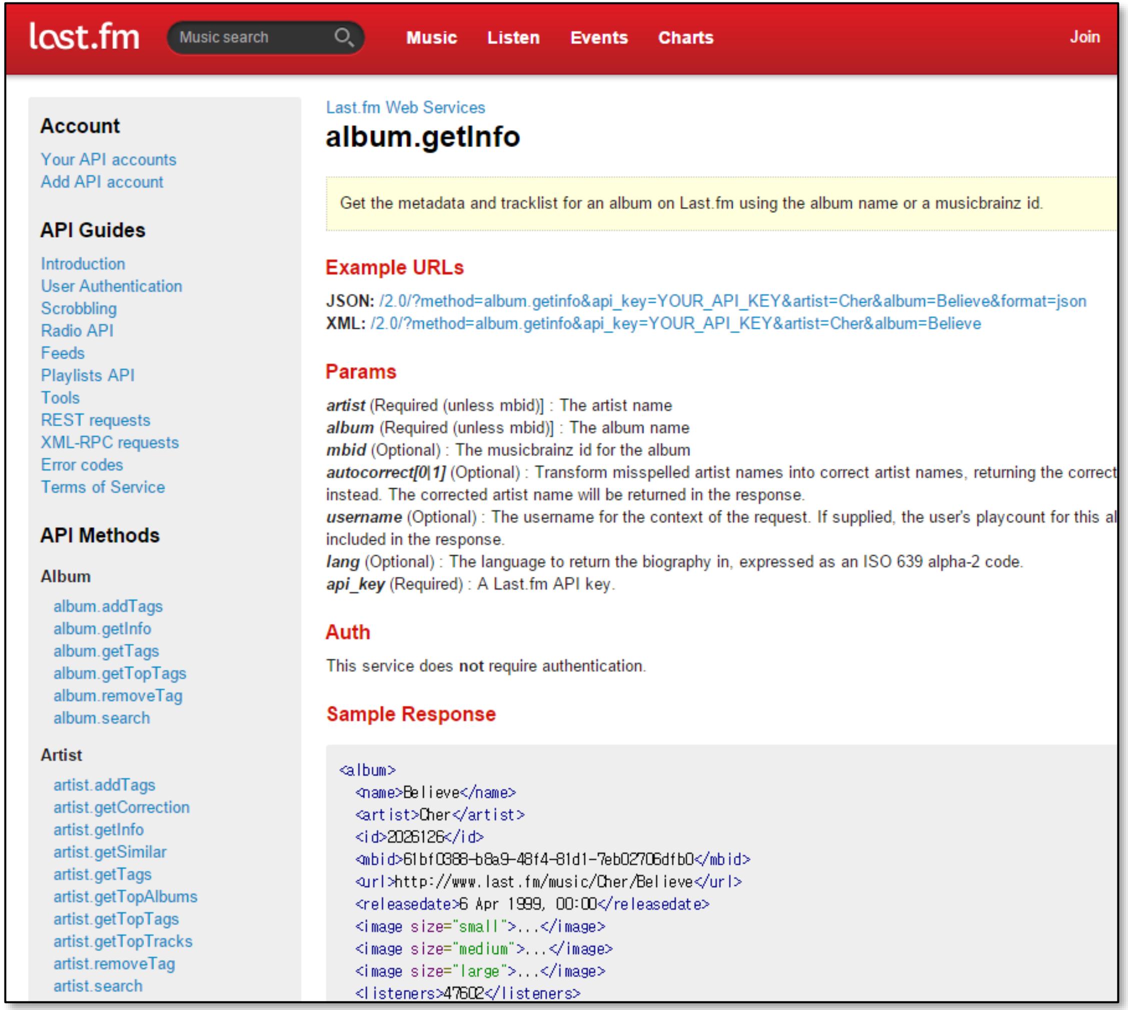
- REST (Representational State Transfer) API



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<
  <url>http://www.last.fm/music/Cher/Believe<
  <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>
```



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 Scraping (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:

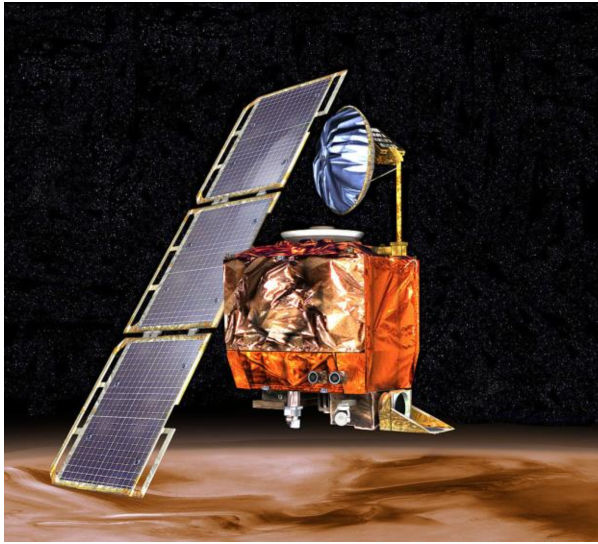
- Unit conversions
- Number / character code representations
- Name unification
- Time/date unification
- Financial unification

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:

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

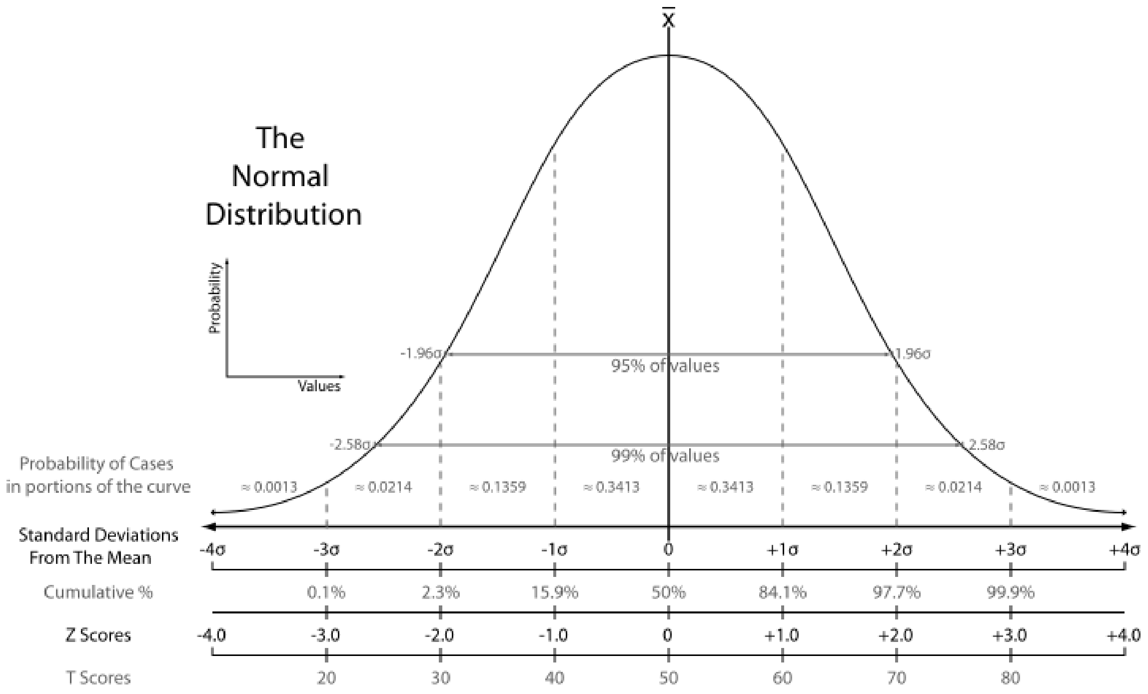
Z-scores of height measured in inches is the same as height measured in miles.

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

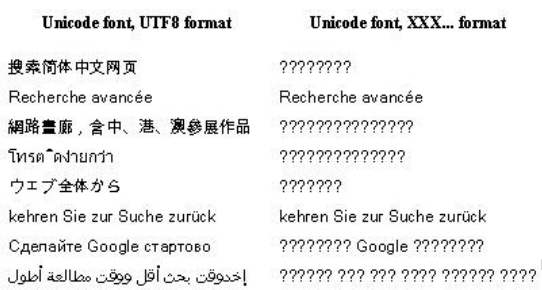
The Ariane 5 rocket exploded in 1996 due to a bad 64-bit float to 16-bit integer conversion.

- 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).

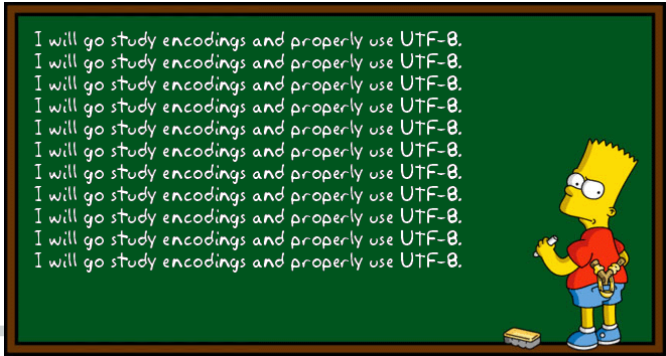
Character Representations

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	??????????????
โทรต̂ดง่ายกว่า	?????????????
ウェブ全体から	???????
kehren Sie zur Suche zurück	kehren Sie zur Suche zurück
Сделайте Google стартово	???????? Google ????????
إخدوقت بحث أقل ووقت 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

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

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?

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

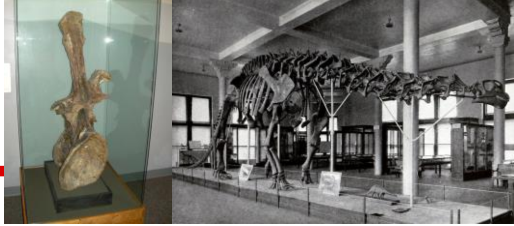
- *Mean value imputation* - leaves mean same.
- *Heuristic-based imputation* – a good guess for your death year is birth year+80.
- *Random value imputation* - repeatedly selecting random values permits statistical evaluation of the impact of imputation.

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