

Welcome to CSC 276

Data Science

CSC 276: Data Science

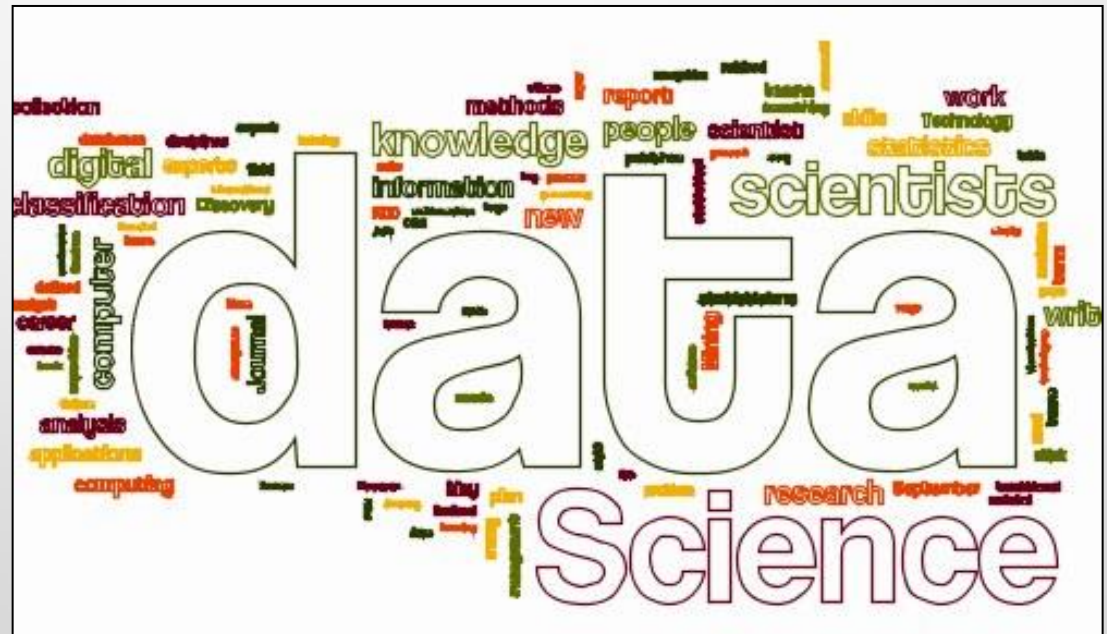
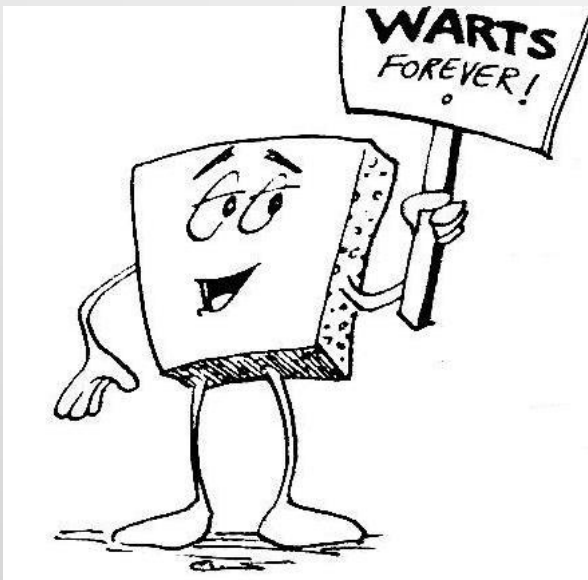
Lecture #2

Introduction

Dr.Fatema Nafa
Fall 2022

Welcome to CSC 276!

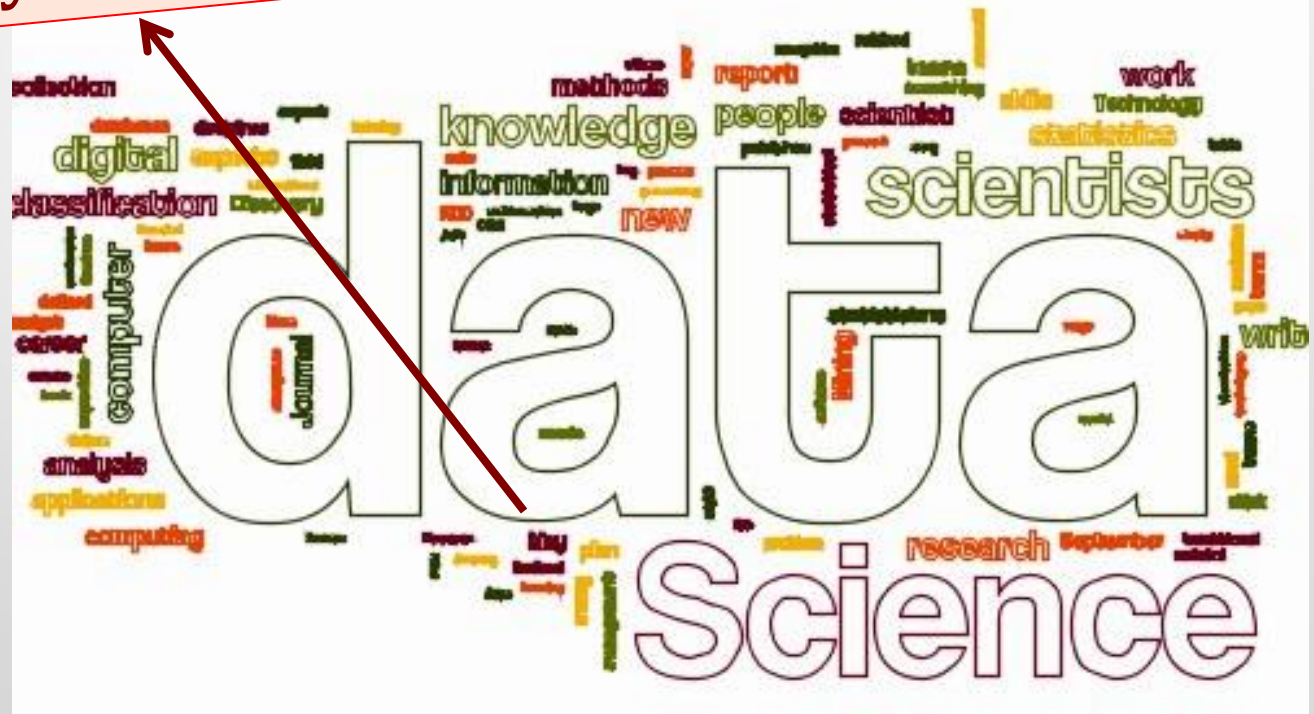
Data Science



Welcome to CSC 276!

This class is truly seminar-style: I'm here, as you are, in order to gain insights into this very new field....

Data Science



Lecture Outline

- **The Art of Data Science**
- Volume, Velocity, Variety
- The Logic of Data Science
- How to Be Agile
- Treating Data as Evidence
- Python
 - Fundamentals of Data Manipulation
 - Basic Data Processing with Pandas
 - Answering Questions with Messy Data



Data Science – A Definition

a data scientist is someone who asks unique, interesting questions of data based on formal or informal theory, to generate rigorous and useful insights.

It is likely to be an individual with multi-disciplinary training in computer science, business, economics, statistics, and armed with the **necessary quantity of domain knowledge relevant to the question at hand.**

The potential of the field is enormous for just a few well-trained data scientists armed with big data have the potential to transform organizations and societies.

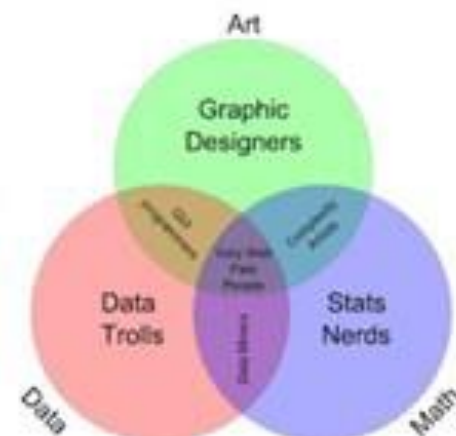
In the narrower domain of business life, the role of the data scientist is to generate applicable business intelligence.

Among all the new buzzwords in business – and there are many –

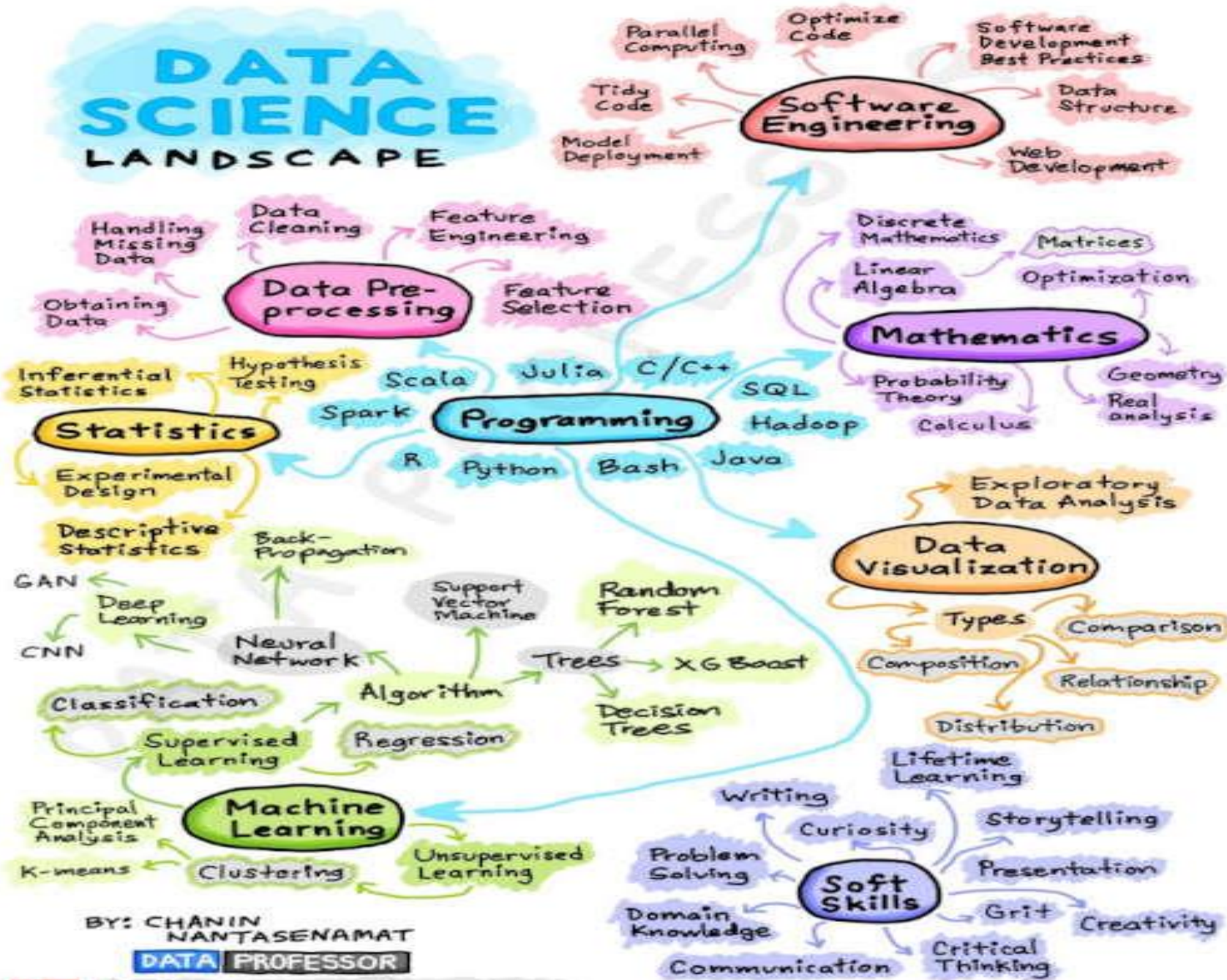
Data Science – A Definition

Being a data scientist is inherently interdisciplinary.

Good questions come from many disciplines, and the best answers are likely to come from people who are interested in multiple fields, or at least from teams that co-mingle varied skill sets.



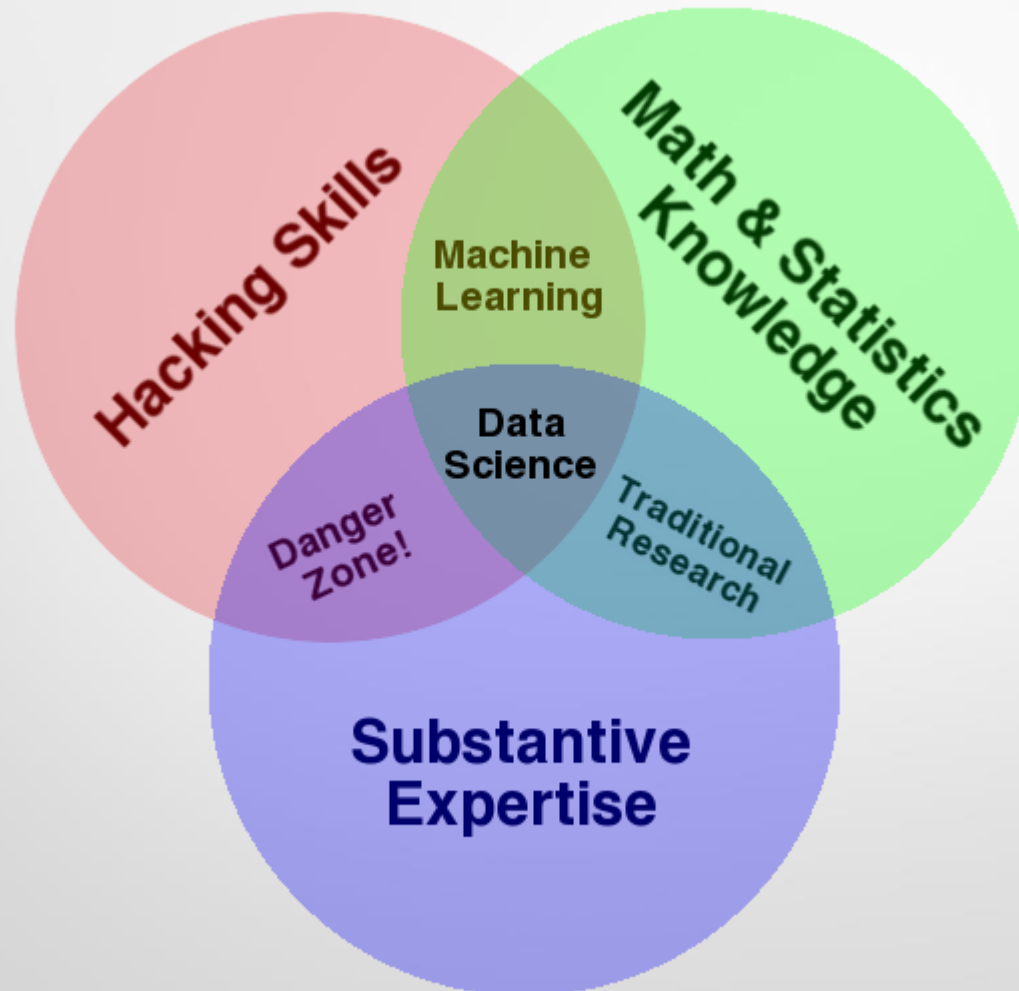
DATA SCIENCE LANDSCAPE



BY: CHANIN NANTASENAMAT

DATA PROFESSOR

Data Science – One Definition



Data Scientists are in high demand

Harvard Business Review

THE MAGAZINE | BLOGS | VIDEO | BOOKS | CASES | WEBINARS | COURSES

Guest: Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE
October 2012

ARTICLE PREVIEW To read the full article, [sign in](#) or [register](#). HBR subscribers click [here to register](#) for FREE access.

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Forbes | [Subscribe](#) | [Sign Up](#) | [Log In](#) | [Jobs](#) | [Videos](#) | [First look at Forbes](#) | [Forbes.com mobile app](#)

CEO's Edge: How The World's Highest-Paid Execs Stay on Top

Startup Success: How to Grow Your Tech Start-Up

Chart: The World's Most Valuable Tech Companies

EMC² The Hottest Jobs In IT: Training Tomorrow's Data Scientists

25 CNBC | [Enter Symbols](#) | [GO](#) | [Enter Keywords](#) | [GO](#)

HOME U.S. | NEWS | MARKETS | INVESTING | TECH | SMALL BIZ | VIDEO | SHOWS | PRIMETIME

NEW SHOW
SQUAWKalley | The Intersection of Wall St. & Tech

BIG DATA | A CNBC SPECIAL REPORT

Why your kids will want to be data scientists

John Phillips | @J_Phillips_IV
Tuesday, 3 Jun 2014 | 7:05 PM ET

TechRepublic / U.S. | [All Topics](#) | [Newsletters](#) | [Photos](#) | [Forums](#) | [Resource Library](#) | [Research](#)

CXO | Software | Startups | Cloud | Data Center | Mobile | Microsoft | Apple | Google | [Search TechRepublic](#)

SAP Is your business making the most out of today's technologies?

BIG DATA

Big data skills: Should data scientist be your next job?

Also, in academia

WHITE HOUSE TO UNIVERSITIES: WE NEED MORE DATA SCIENTISTS

NEW YORK UNIVERSITY, UNIVERSITY OF CALIFORNIA-BERKELEY, AND THE UNIVERSITY OF WASHINGTON ARE LAUNCHING A \$37.8 MILLION PROJECT TO BOOST THE NUMBERS OF AMERICAN DATA SCIENTISTS

BY NEAL UNGERLEIDER

It's official: America needs more data scientists. This week, a \$37.8 million project

Berkeley Research
UNIVERSITY OF CALIFORNIA

RESEARCH HIGHLIGHTS NEWS ACTIVITIES RESEARCH UNITS FACULTY EMPLOYEES RESEARCH POLICIES & ADMINISTRATION TECH TRENDS LONG-TERM RESEARCH

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

INSTITUTE FOR DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE

DATA SCIENCE



Data Science at UC Ber

SCIENTIFIC AMERICAN

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

Subscribe

How Big Data Can Transform Society for the Better

The digital traces we leave behind each day reveal more about us than we know. This could become a privacy nightmare—or it could be the foundation of a healthier, more prosperous world.

See Inside

See Inside

See Inside

See Inside

See Inside

See Inside

See Inside

See Inside

See Inside



RESEARCH CENTERS IN THE FIELD OF DATA SCIENCE

Center for Data Science (CDS)

The NYU Center for Data Science (CDS) is a focal point for New York University's university-wide initiative in data science. It was established to help advance NYU's goal of creating the country's leading data science training and research facilities, joining researchers and professionals with tools to harness the power of big data.

LEARN MORE

Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM)

The Center for the Promotion of Research Involving Innovative Statistical Methodology (PRISM) is a new center dedicated to improving the culture of research in quantitative social, behavioral, behavioral, health and policy science.

500k

The world's 500,000+ data centers are large enough to fit 10 5,000 football fields. (Source: Statista)

75%

75% of digital information is generated by individuals, which means you have probably left 80% of digital data at some point in the life. (Source: Statista)

UNIVERSITY of WASHINGTON

eScience Institute
Supporting Data-Driven Discovery In All Fields

WHO WE ARE

New Ph.D. Tracks in "Big Data"

Pays Well

Big Data, Big Paycheck

Median salary for analytics professionals and those specifically within data science, by level of experience.



Note: Data do not include managers Source: Burtch Works

The Wall Street Journal

Lecture Outline

- **The Art of Data Science**
- **Volume, Velocity, Variety**
- The Logic of Data Science
- How to Be Agile
- Treating Data as Evidence
- Fundamentals of Data Manipulation
- Basic Data Processing with Pandas
- Answering Questions with Messy Data

There are several "V"s of big data: three of these are volume, velocity, variety.⁸ Big data exceeds the storage capacity of conventional databases. This is its *volume* aspect. The scale of data generation is mind-boggling. Google's Eric Schmidt pointed out that until 2003, all of human kind had generated just 5 exabytes of data (an exabyte is 1000^6 bytes or a billion-billion bytes). Today we generate 5 exabytes of data every two days. The main reason for this is the explosion of "interaction" data, a new phenomenon in contrast to mere "transaction" data. Interaction data comes from recording activities in our day-to-day ever more digital lives, such as browser activity, geo-location data, RFID data, sensors, personal digital recorders such as the fitbit and phones, satellites, etc. We now live in the "internet of things" (or IoT), and it's producing a wild quantity of data, all of which we seem to have an endless need to analyze. In some quarters it is better to speak of 4 Vs of big data, as shown in Figure 1.1.

Goal of Data Science

Turn **data** into **data products**.

Analysis

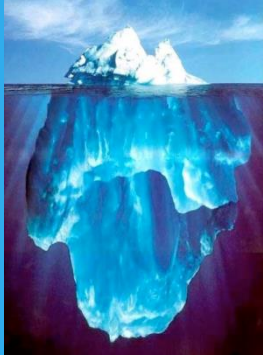
What kinds of data will you use?

- Almost anything is OK, except other predictions.
- **History:** individual or pair-wise?
- Team or players?
- Numerical or text?
- What kind of model will you build?
- What assumptions are safe to make?

Where does data come from?

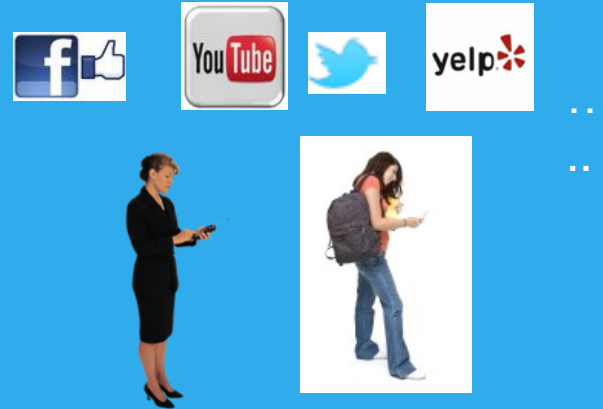
"Big Data" Sources

It's All Happening On-line



Every:
Click
Ad impression
Billing event
Fast Forward, pause,...
Server request
Transaction
Network message
Fault
...

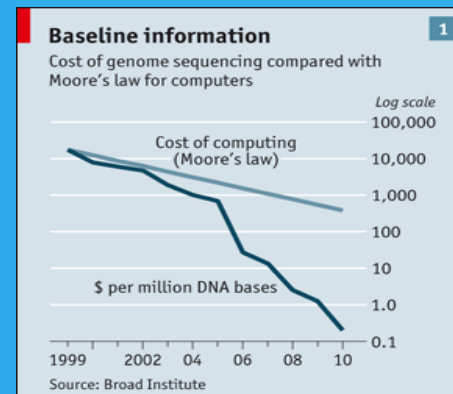
User Generated (Web & Mobile)



Internet of Things / M2M



Health/Scientific Computing

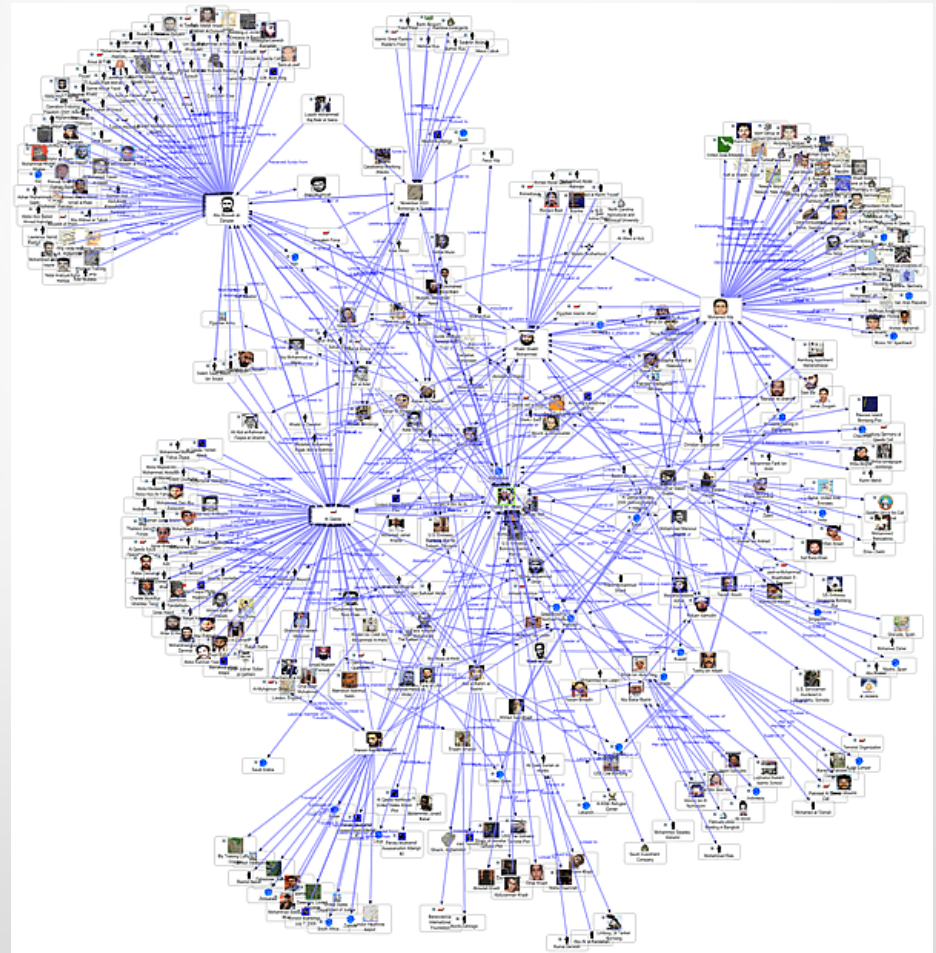


Graph Data

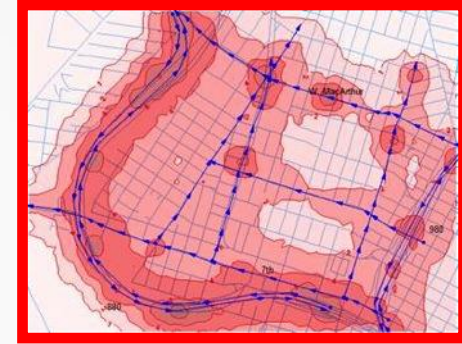
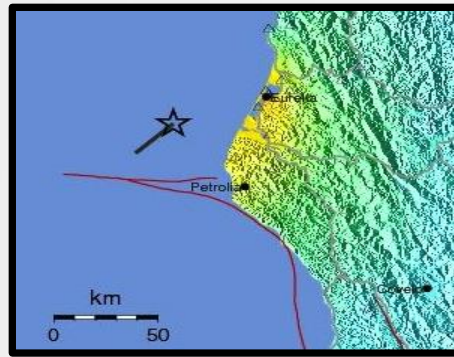
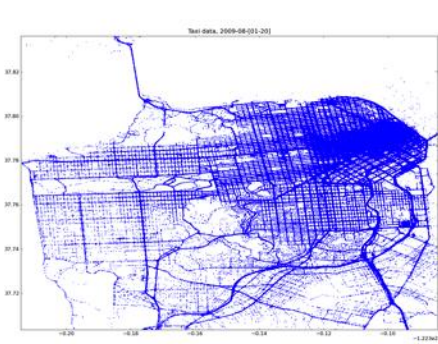
Lots of interesting data has a graph structure:

- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get quite large (e.g., Facebook* user graph)



What can you do with the data?



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



From Alex Bayen, UCB

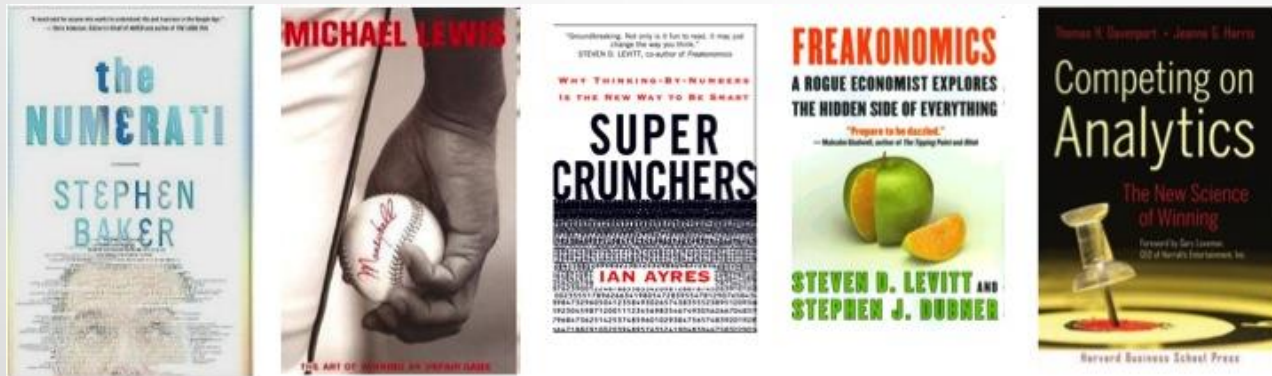


Contrast: Databases

	Databases	Data Science
Data Value	"Precious"	"Cheap"
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Realizations	SQL	NoSQL: Riak, Memcached, Apache River, MongoDB, CouchDB, Hbase, Cassandra,...

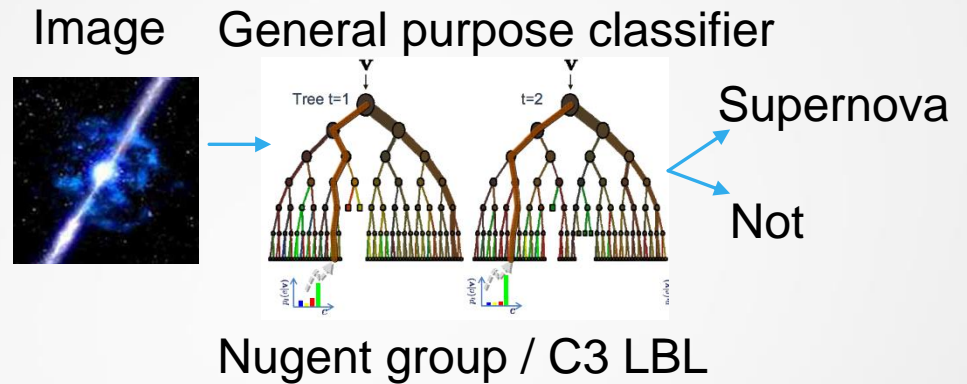
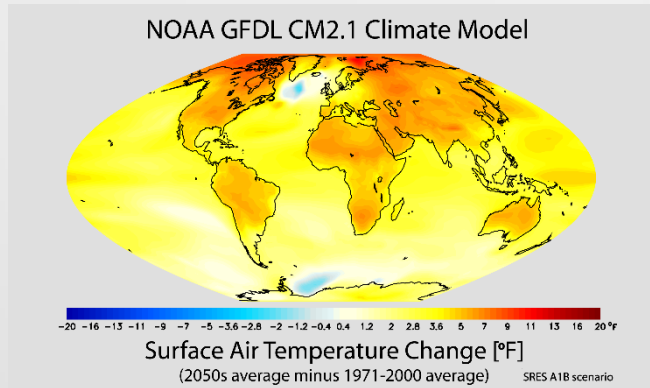
Contrast: Databases

Databases	Data Science
Querying the past	Querying the future



Business intelligence (BI) is the transformation of raw data into meaningful and useful information for [business analysis](#) purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia

Contrast: Scientific Computing



Scientific Modeling

Physics-based models

Problem-Structured

Mostly deterministic, precise

Run on Supercomputer or High-end Computing Cluster

Data-Driven Approach

General inference engine replaces model

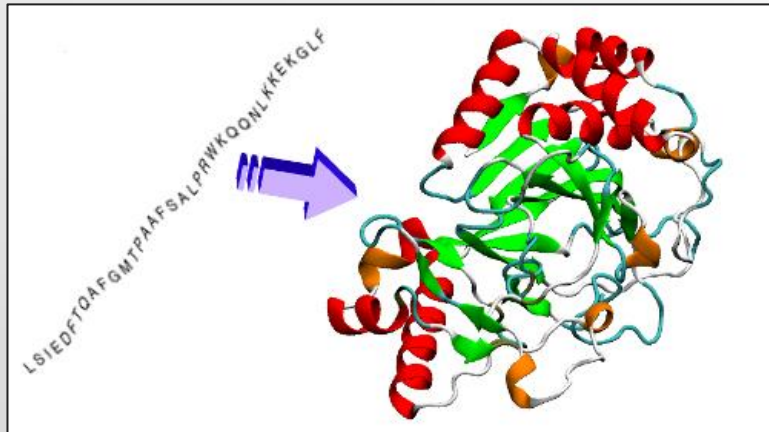
Structure not related to problem

Statistical models handle true randomness, and **unmodeled complexity**.

Run on cheaper computer Clusters (EC2)

Contrast: Computational Science

CASP: A Worldwide, Biannual Protein Folding Contest



Quark

Rich, Complex
Energy Models

Faithful, Physical
Simulation

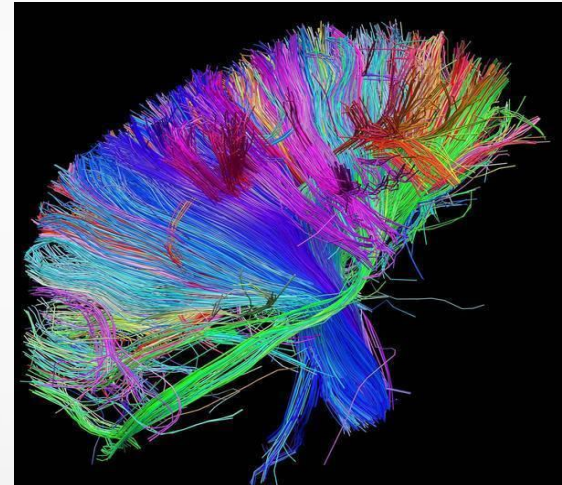
Raptor-X

Data-intensive,
general ML models

Feature-based inference

Conditional Neural Fields

Brain Mapping: Allen Institute, White House, Berkeley



Techniques (Massive ML)

Principal Component Analysis

Independent Component Analysis

Sparse Coding

Spatial (Image) Filtering

Contrast: Machine Learning

Machine Learning

Develop new (individual) models

Prove mathematical properties of models

Improve/validate on a few, relatively clean, small datasets

Publish a paper

Data Science

Explore many models, build and tune hybrids

Understand empirical properties of models

Develop/use tools that can handle massive datasets

Take action!

How to Be Agile

The nature of data science is experimental. You don't know the answer to the question asked of you—or even if an answer exists. You don't know how long it will take to produce a result or how much data you need. The easiest approach is to just come up with an idea and work on it until you have something. But for those of us with deadlines and expectations, that approach doesn't fly. Companies that issue you regular paychecks usually want insight into your progress.

This is where being agile matters. An agile data scientist works in small iterations, pivots based on results, and learns along the way. Being agile doesn't guarantee that an idea will succeed, but it does decrease the amount of time it takes to spot a dead end. Agile data science lets you deliver results on a regular basis and it keeps stakeholders engaged.

The key to agile data science is delivering data products in defined time boxes—say, two- to three-week sprints. Short delivery cycles force us to be creative and break our research into small chunks that can be tested using minimum viable experiments (Figure 6-1). We deliver something tangible after almost every sprint for our stakeholders to review and give us feedback. Our stakeholders get better visibility into our work, and we learn early on if we are on track.

Treating Data as Evidence

The logic of data science tells us what it means to treat data as evidence. But following the evidence does not necessarily lead to a smooth increase or decrease in confidence in a model. Models in real-world data science change, and sometimes these changes can be dramatic. New observations can change the models you should consider. New evidence can change confidence in a model. As we collected new employee satisfaction responses, factors like specific job titles became less important, while factors like advancement opportunities became crucial. We stuck with the methods described in this chapter, and as we collected more observations, our models became more stable and more reliable.

I believe that data science is the best technology we have for discovering business insights. At its best, data science is a competition of hypotheses about how a business really works. The logic of data science are the rules of the contest. For the practicing data scientist, simple rules like Ockham's Razor and Bayesian reasoning are all you need to make high-quality, real-world decisions.

Lecture Outline

- **The Art of Data Science**
- Volume, Velocity, Variety
- The Logic of Data Science
- How to Be Agile
- Treating Data as Evidence
- **Python**
 - Fundamentals of Data Manipulation
 - Basic Data Processing with Pandas
 - Answering Questions with Messy Data

1. Data sources

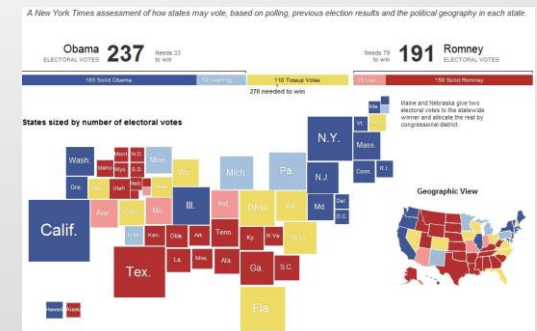
2. Collect data(**download**)

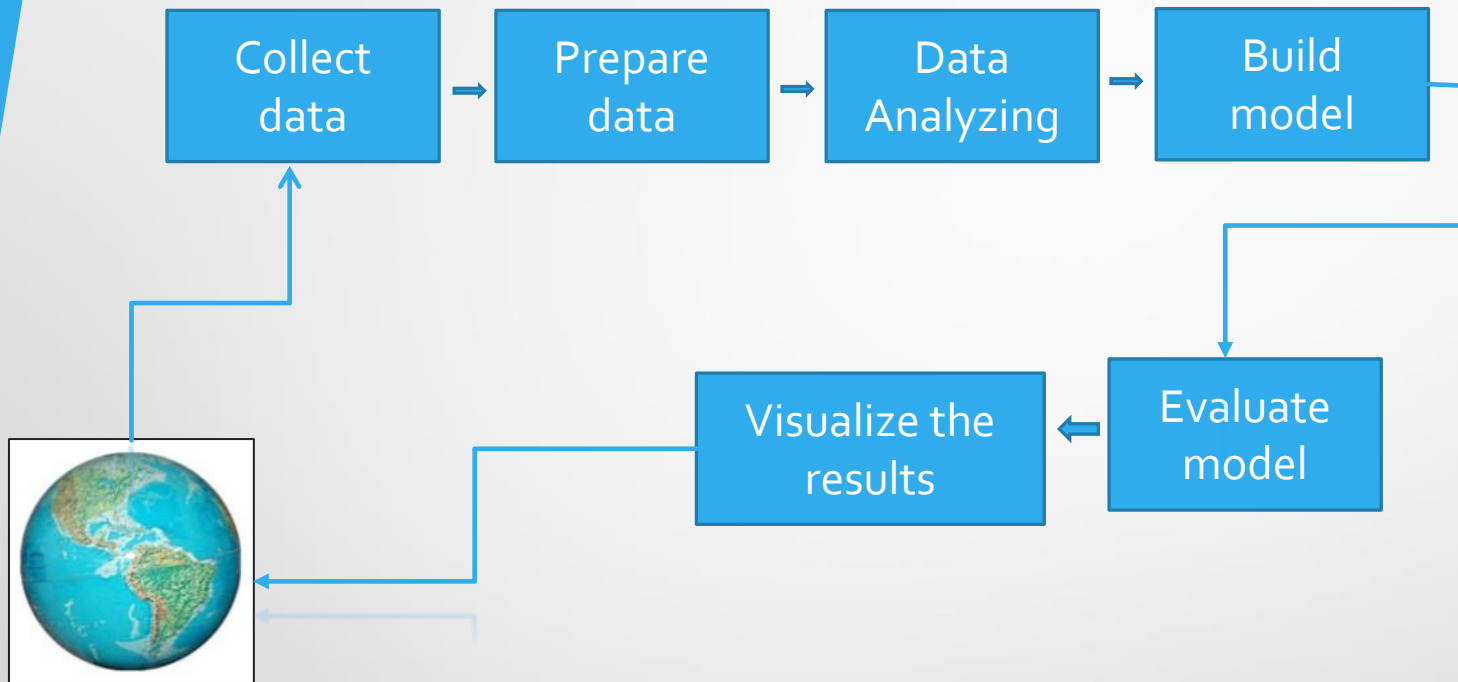
3. Prepare data (integrate, transform, clean, filter, aggregate)

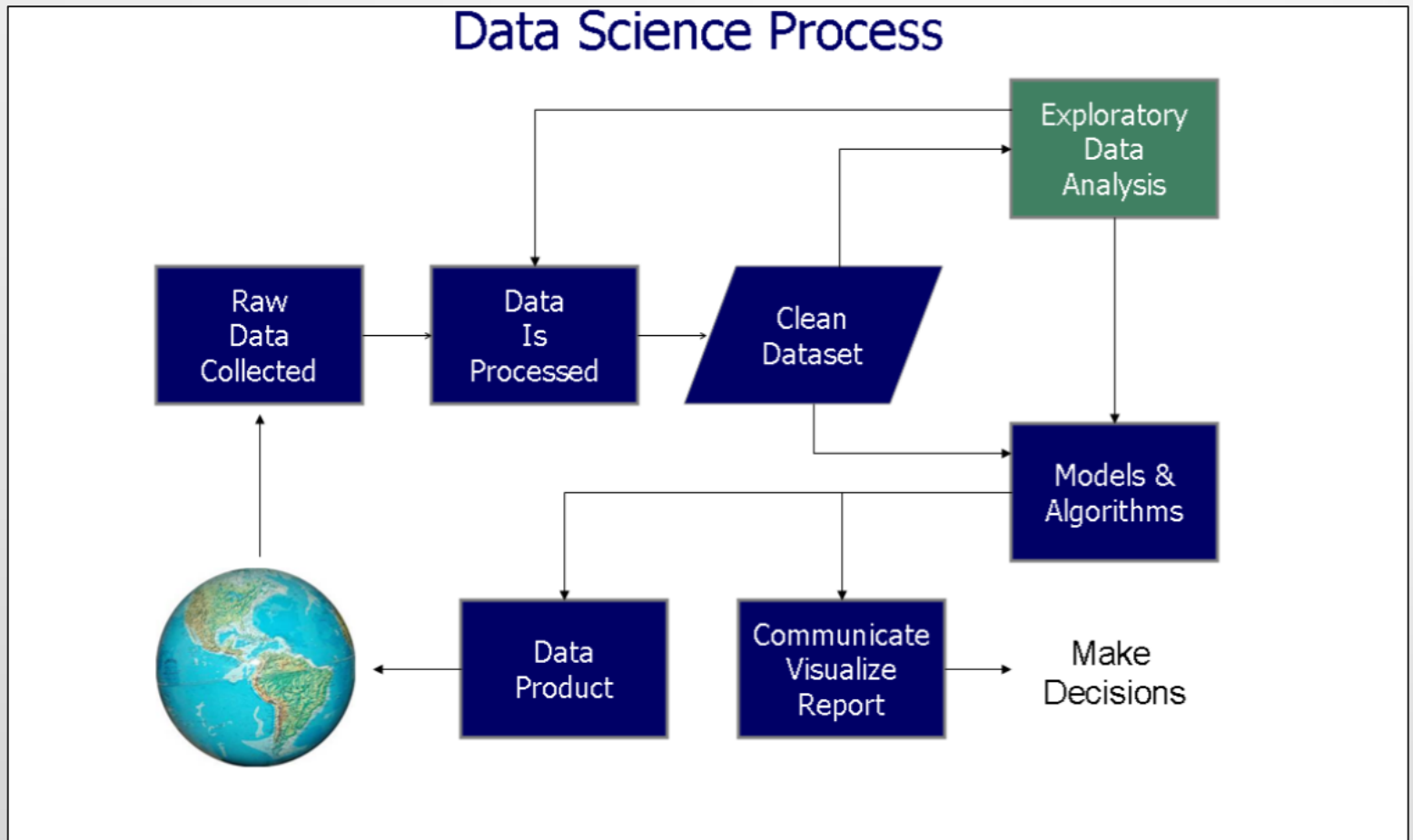
4. Build model

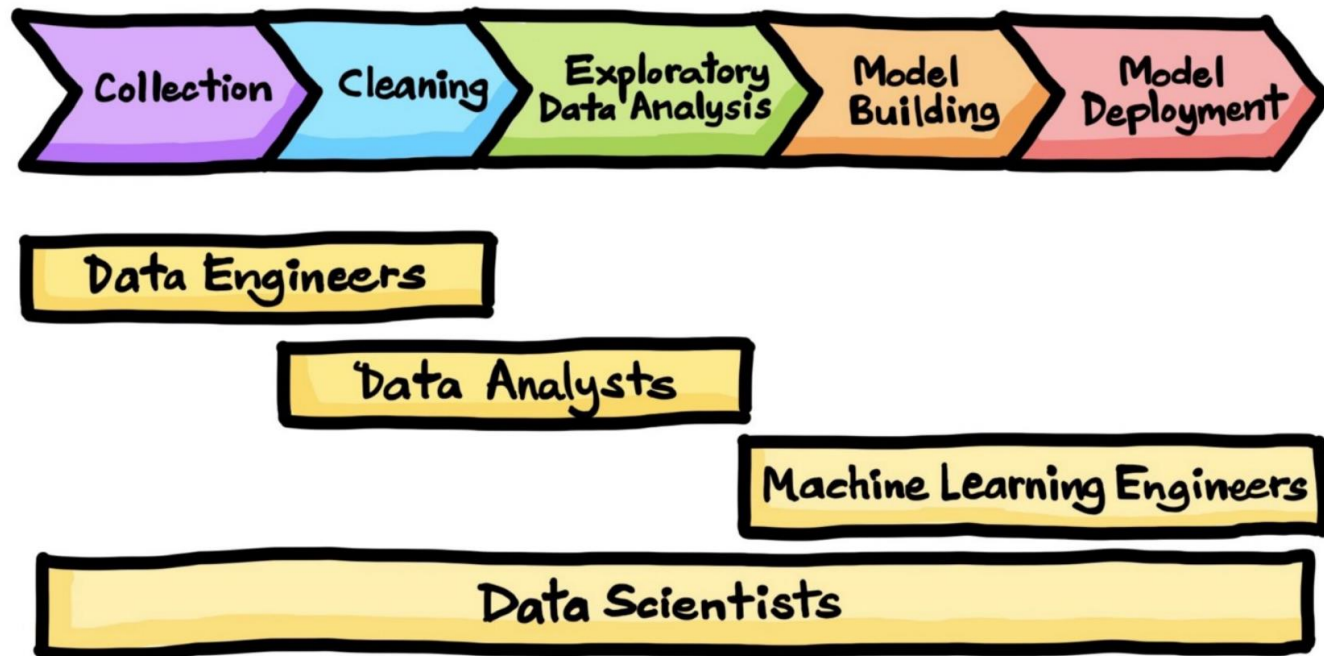
5. Evaluate model

6. Visualize the results









Asking Interesting Questions from Data

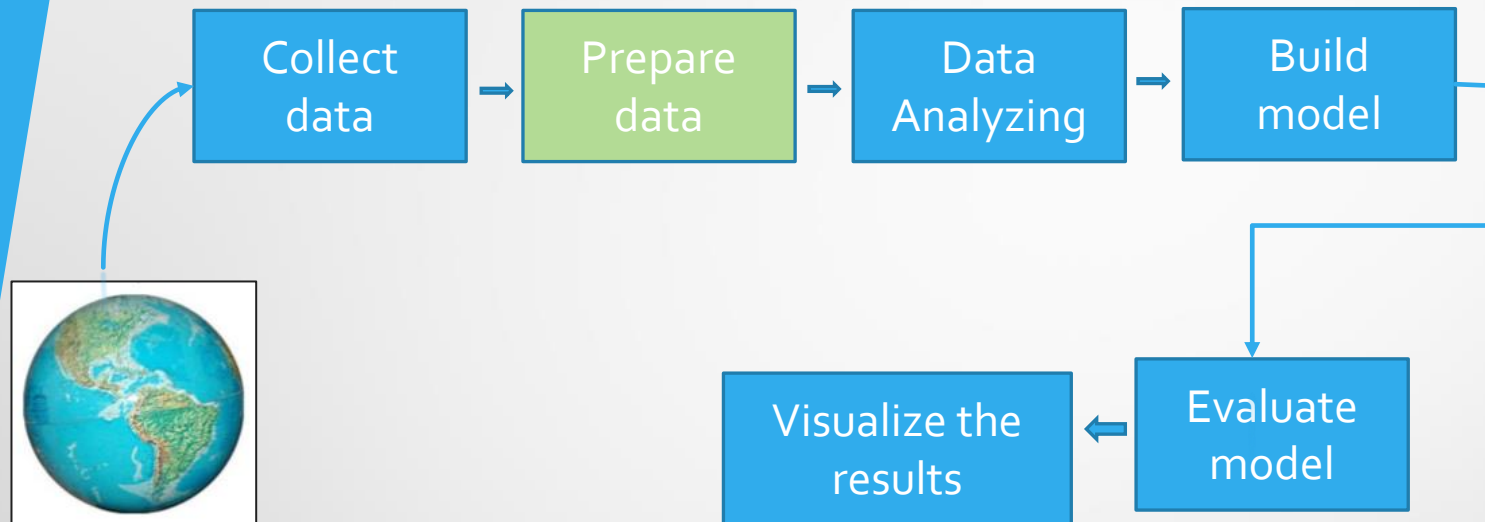
Good data scientists develop an inherent curiosity about the world around them, particularly in the associated domains and applications they are working on. They enjoy talking shop with the people whose data they work with. They ask them questions: What is the coolest thing you have learned about this field? Why did you get interested in it? What do you hope to learn by analyzing your data set? Data scientists always ask questions.

Good data scientists have wide-ranging interests. They read the newspaper every day to get a broader perspective on what is exciting. They understand that the world is an interesting place. Knowing a little something about everything equips them to play in other people's backyards. They are brave enough to get out of their comfort zones a bit, and driven to learn more once they get there.

Software developers are not really encouraged to ask questions, but data scientists are. We ask questions like:

- What things might you be able to learn from a given data set?
- What do you/your people really want to know about the world?
- What will it mean to you once you find out?

- After you understand what kind of information is available, try to come up with, say,
- 10 interesting questions you might explore/answer with access to the data set.



My DATA

Statistical information about my data

What's Hard about Data Science

- Overcoming assumptions
- Making ad-hoc explanations of data patterns
- Overgeneralizing
- Communication
- Not checking enough (validate models, data pipeline integrity, etc.)
- Using statistical tests correctly
- Prototype → Production transitions
- Data pipeline complexity (who do you ask?)

Readings

Read next week readings and complete it before next class.

- **Chapter Two**: Python Language Basics, IPython, and Jupyter Notebooks
- **Chapter Three**: Built-In Data Structures, Functions, and Files

python for data analysis 2nd edition pdf

