Data Science and Business Intelligence

BU.330.780

Session 1

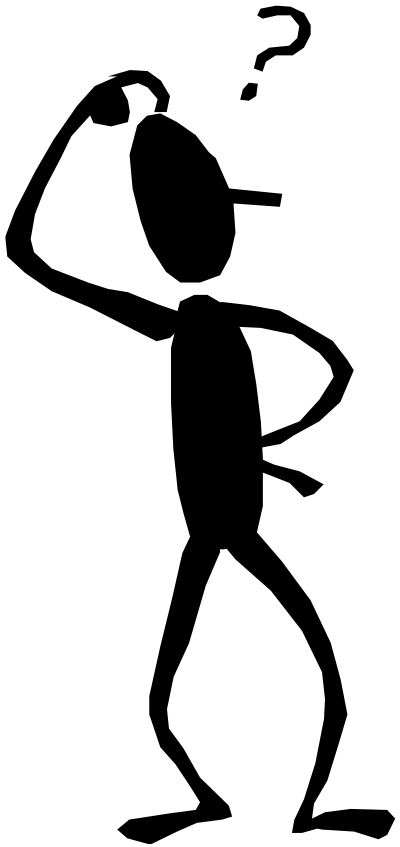
Instructor: Changmi Jung, Ph.D.

Today's Agenda

- Course Introduction
- Data Science Tasks
- Getting Started: Data prep and Tidyverse + R Markdown

JOHNS HOPKINS
CAREY BUSINESS SCHOOL

What is Data Science? How Are They Different?



Big Data

Business Intelligence

Data mining

Data Science

Machine Learning

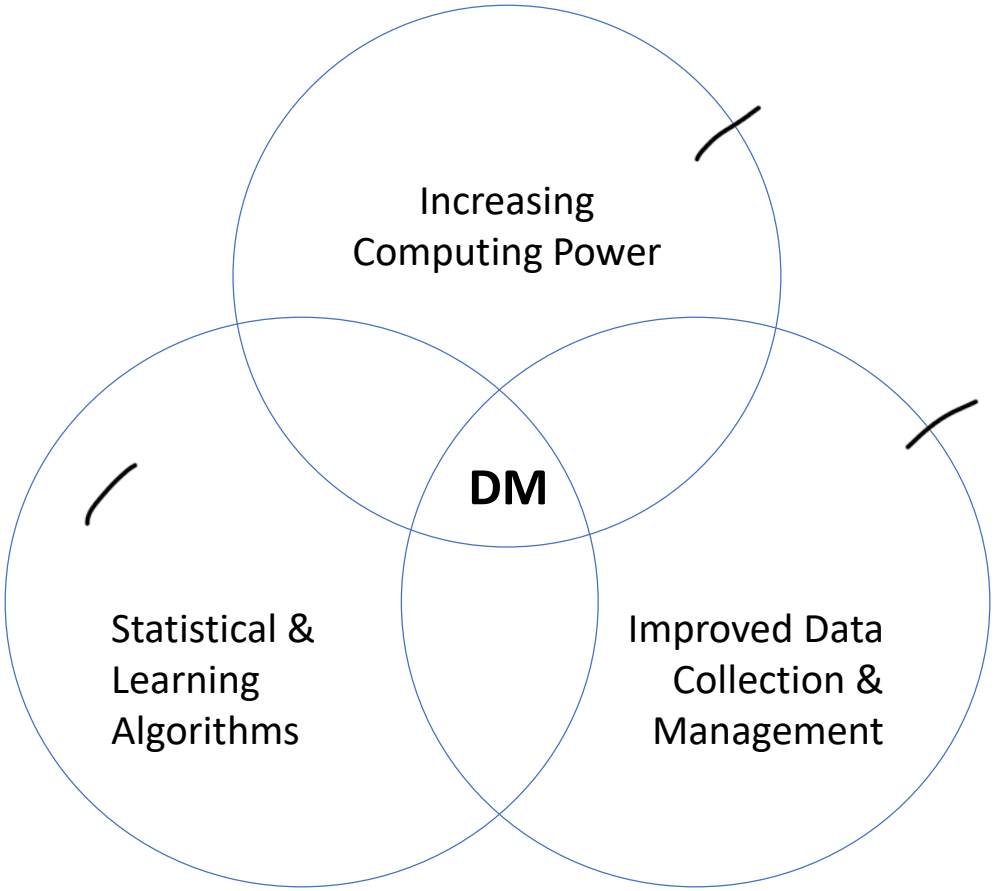
Data-driven
decision making

The Internet In Real Time

Source: https://www.betfy.co.uk/internet-realtime/

Data Opportunities

- 3V's of Big Data
  - Volume
  - Velocity
  - Variety
- Powerful computers
- Better algorithms



Data are widely available;

What is scarce is the ability to extract wisdom from them

– Hal Varian (2010), Chief Economist at Google

JOHNS HOPKINS

Source: http://www.economist.com/node/15557443

Automated Decision Making

- Credit scoring

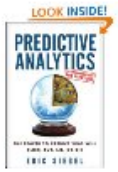
- Prevent customer churn

- Targeted marketing

Personalized Recommendation

Customers Who Bought This Item Also Bought

Data Smart: Using Data Science to Transform …
› John W. Foreman
(43)
Paperback
$26.77 Prime



Predictive Analytics: The Power to Predict …
› Eric Siegel
(183)
Hardcover
$15.82 Prime

Introduction to Data Mining
Pang-Ning Tan
(26)
Hardcover
$110.28 Prime

Doing Data Science: Straight Talk from the …
› Cathy O'Neil
(32)
Paperback
$26.67 Prime

Big Data: A Revolution That Will Transform …
› Viktor Mayer-Schönberge…
(274)
Paperback
$10.09 Prime

Big Data at Work: Dispelling the Myths, …
› Thomas H. Davenport
(64)
Hardcover
$18.98 Prime

PEOPLE YOU MAY KNOW

Jay Kreps
Principal Staff Engineer at LinkedIn
Connect

Igor Perisic
VP Engineering at LinkedIn
Connect

Sam Shah
Principal Engineer at LinkedIn
Connect

See more »

Best Technology Jobs

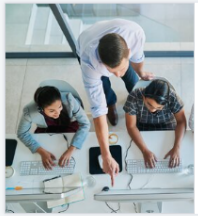
We use technology more than ever these days to connect with friends and family, stay up to date on the latest and greatest happenings in the world, and sometimes just to pass the time. With all the computers, tablets, smartphones and other high-tech devices society depends on, we need the skills of technology professionals. U.S. News' Best Technology Jobs of 2025 are high-paying and boast low unemployment rates. Check out what makes these gigs so great. For more information on how we rank, read the Best Jobs Methodology.

Rankings

Data Scientists rank #4 in Best Technology Jobs. Jobs are ranked according to their ability to offer an elusive mix of factors. Read more about how we rank the best jobs.

- #4 in **Best Technology Jobs**
- #6 in **Best STEM Jobs**
- #8 in **100 Best Jobs**

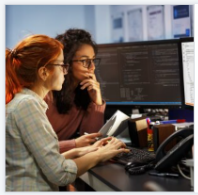
Scorecard	6.7
Wage Potential	6.8
Employment	4
Future Prospects	9.8
Comfort	4.4
Work Life Balance	5



IT Manage

#1 in **Best T**

Our increasingly computer-relate recommending software, secur new technologi



Software D

#2 in **Best T**

Software develc to succeed in th work better. Rea

Informatio

#3 in **Best T**

As concern abo analysts. It is th measures that protect a company's computer networks and systems. Read More »

Education Needed
Bachelor's

Data Scientist

#4 in **Best Technology Jobs**

Data scientists use technology to glean insights from large amounts of data they collect. Read More »

Projected Jobs
73,100

Median Salary
$108,020

Education Needed
Bachelor's

Actuary

#5 in **Best Technology Jobs**

Are you more of a risk calculator than a risk taker? Consider working as an actuary. These professionals are experts in uncertainty, using mathematics, statistics and financial theory to measure, manage and mitigate financial risk. Read More »

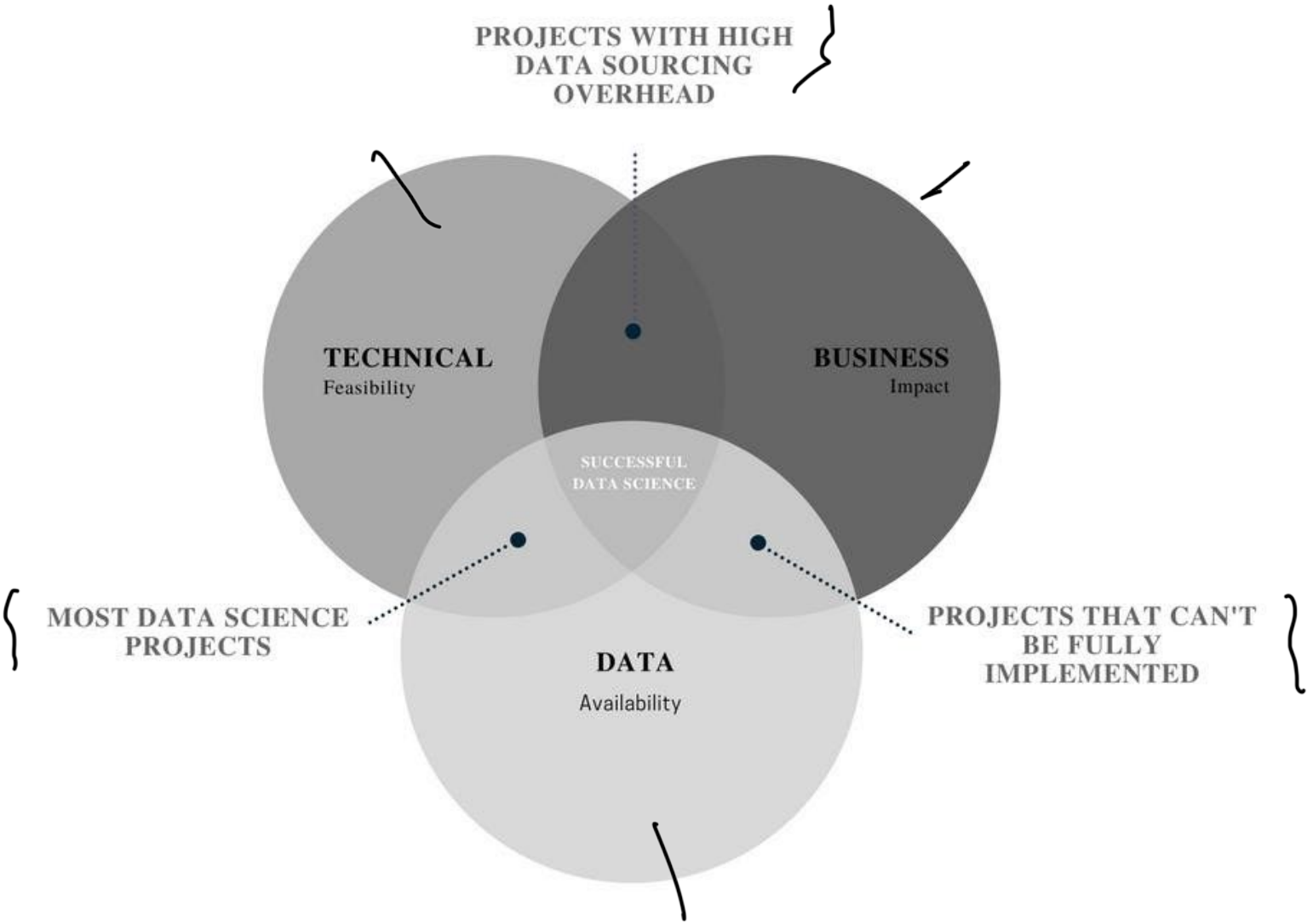
Projected Jobs
6,600

Median Salary
$120,000

Education Needed
Bachelor's

Source: US News

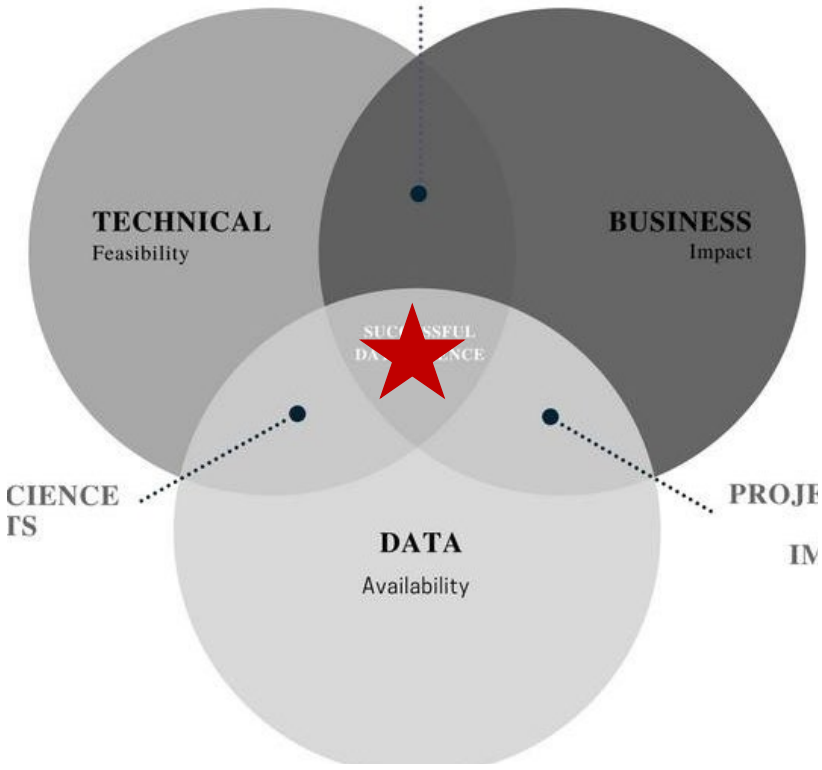
DESIGN THINKING MINDSET FOR DATA SCIENCE

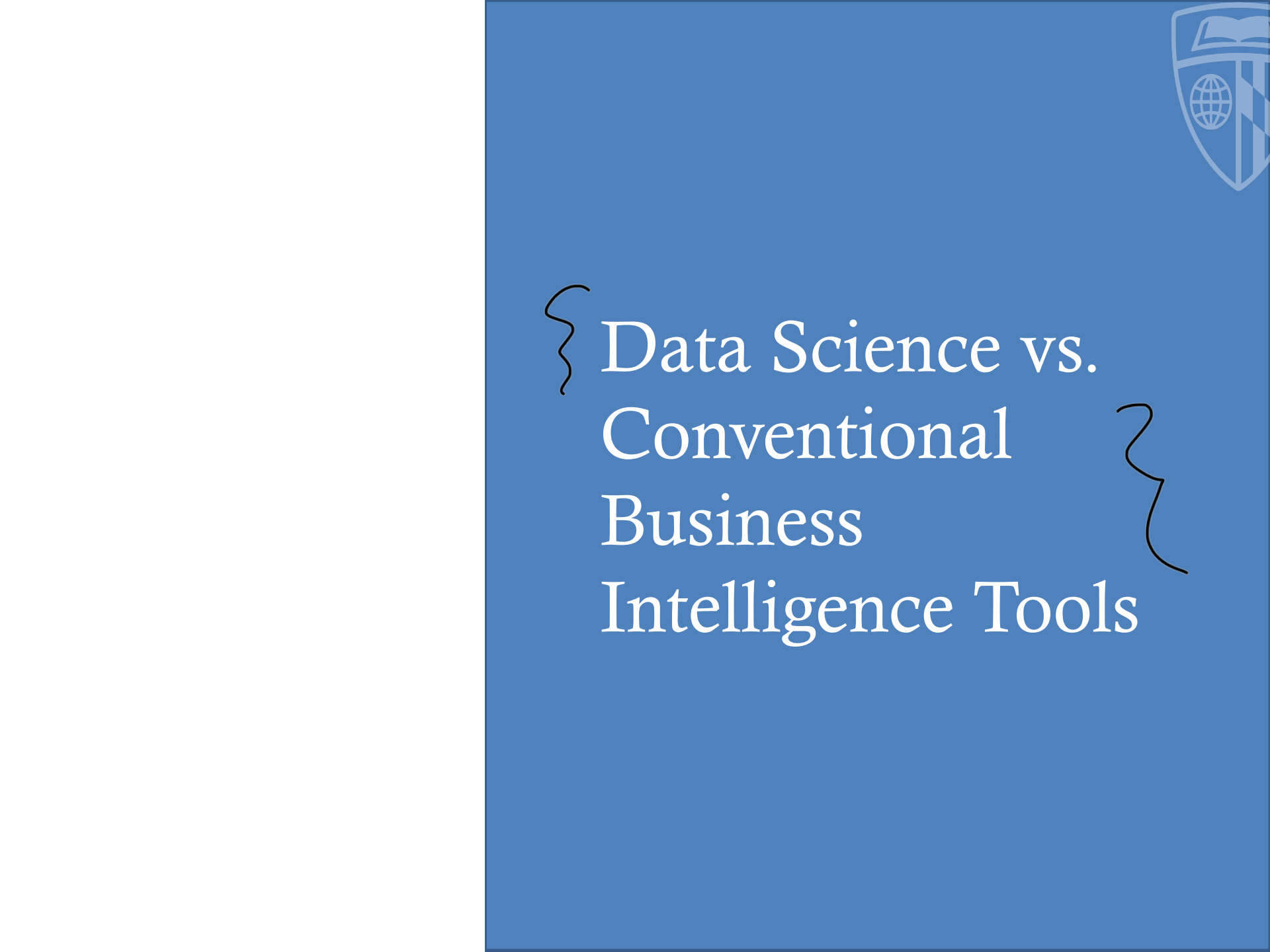


https://towardsdatascience.com/a-design-thinking-mindset-for-data-science-f94f1e27f90

After taking this course…

- Approach business problems data-analytically
- Interact competently on the topic of data science for business.
- Can translate from business to data science, and vice versa.
- Hands-on experience mining data from the start to the end.
- Ability to evaluate proposal and execution.




Data Science vs.
Conventional
Business
Intelligence Tools

Typical Job Descripti

Data Scientist
Booz Allen Hamilton · Fort Meade, MD

Apply | Save | •••

Job Number: R0216818

Data Scientist (AI Engineer) Co-Op
IBM · State College, PA (Hybrid)

Apply | Save

Data Scientist, Sp

amazon
Amazon.com, Inc.
Arlington, VA

Apply on Karkidi | Apply directly o

$ 120K–190K a year | Full-time

Job highlights
Identified by Google from the original job post

Qualifications

- Bachelor's Degree
- 3+ years of experience with data scrip R etc.) or statistical/mathematical so
- 2 years working as a Data Scientist
- Experience in as many of the followin variate testing & design, A/B testing & and regression analysis
- Good understanding of supervised an

Applicants for this job

542 Applicants
59 Applicants in the past day

Applicant seniority level

62% Entry level applicants

14% Senior level applicants

2% Director level applicants

1% CXO level applicants

Applicant education level

45% have a Master's Degree
29% have a Master of Science
12% have a Bachelor's Degree
14% have other degrees

g the secrets held by a data
T, machine learning, and
sive amounts of structured
entist at Booz Allen, you can
solve global challenges. Across
esearch to national intelligence

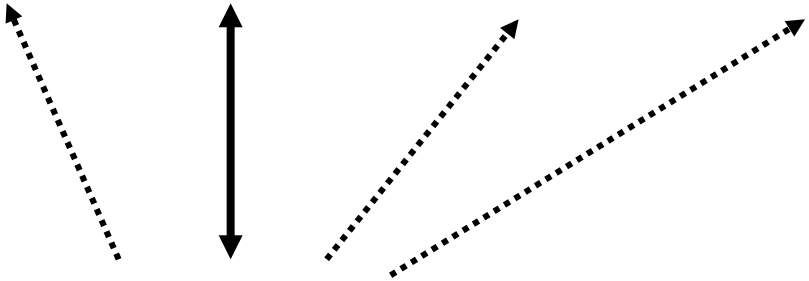
knowledge to create real-
and their questions and needs,
ces of their information
ght combination of tools and
e answers to help clients make
anding of the data, what it all

ch, or advanced mathematics

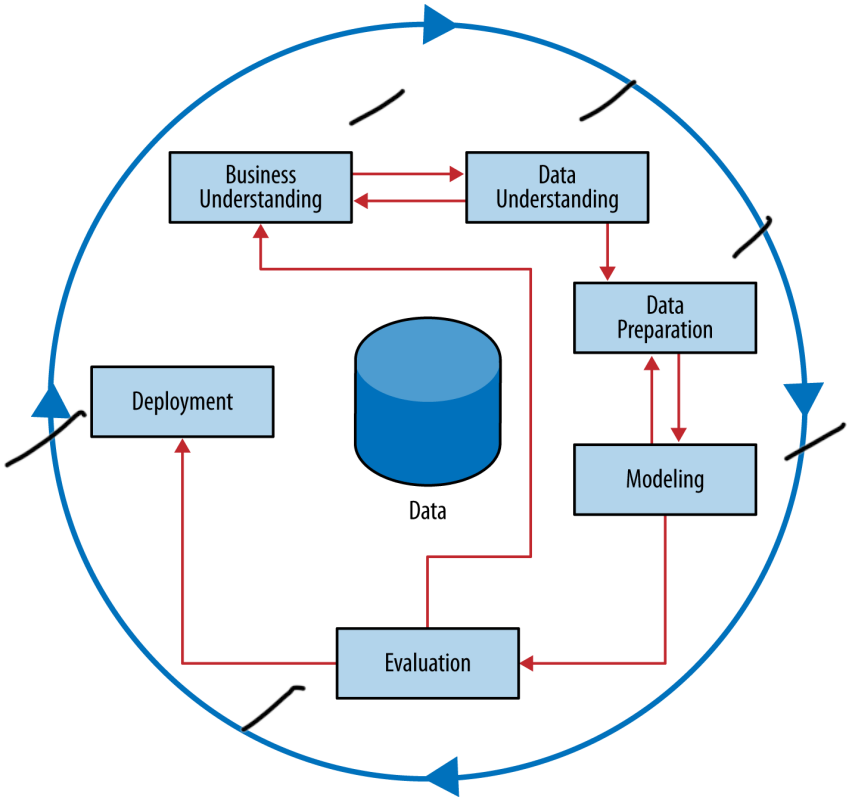
tive and quantitative analyses
paced environment

Data science is a process with well-understood stages

- science + craft + creativity + common sense



a *process*



Business Understanding

- **Understand the business problem** to be solved!
  - formulate a business problem
  - map the problem formulation to a data analysis task
  - understand the situation (available data, suitability of the task, …)
- Data science is an iterative process of discovery
- 80% of the knowledge of interest in a business context can be extracted from data using (*relatively*) conventional tools.

In a nutshell, these are the techniques and technologies that do not involve finding hidden patterns from data, which is the key part of data mining.

Business Understanding

- **Understand the business problem** to be solved!

1. Formulate a business problem

> We have a bookstore chain….. (2 pages of business background)…. our sales record is low during summer….. (another half page)…. So, we may try a new email promotion for subscribers…. (two more paragraphs explaining the promotion)…. Somehow, the promotion will increase the sales, but don't know what to do next….

> Was the promotion successful in the past?
> Which customer groups should receive the promotion?

2. Map the problem formulation to a data analysis task

> Was the promotion successful in the past? → ???
> Which customer groups should receive the promotion? → Classification

3. Understand situation (available data, suitability of the task, …)

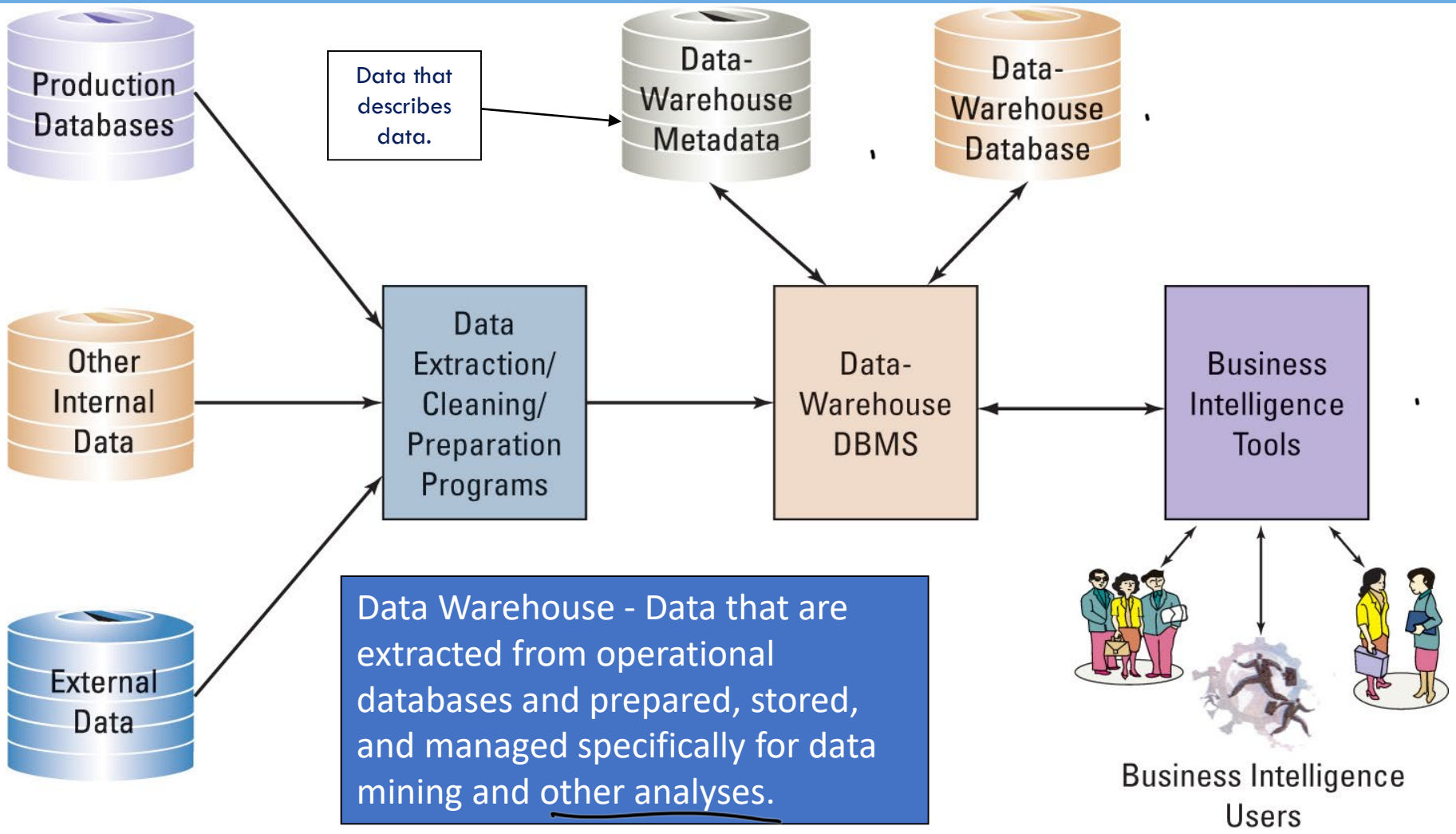
Business Understanding

- 80% of the knowledge of interest in a business context can be extracted from data using (*relatively*) conventional tools.

Was the promotion successful in the past?

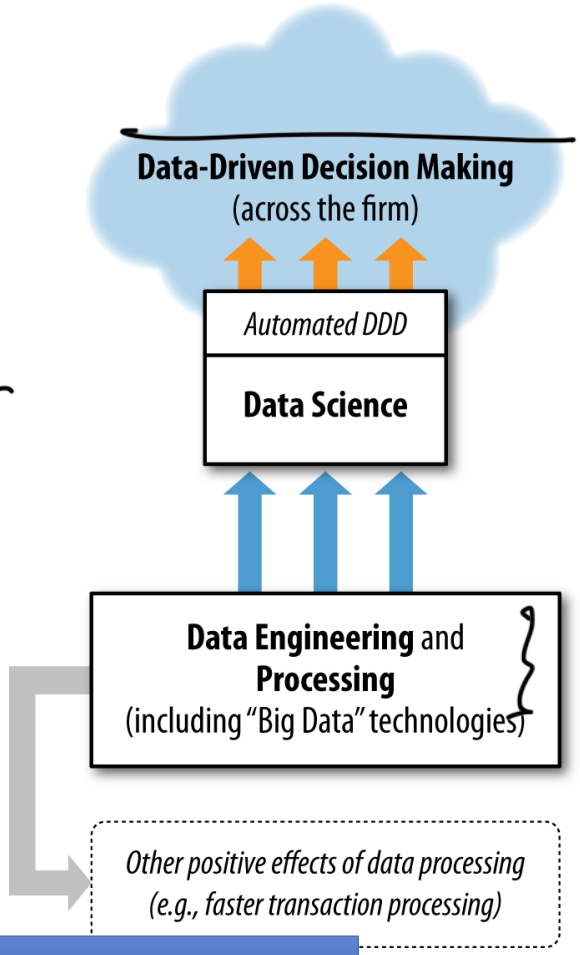
Conventional business intelligence tools such as SQL query, OLAP, statistical hypothesis testing can answer this question!

Data Mining versus… Data Warehouse



Data Mining versus… Data processing and "Big Data"

- Data engineering and data processing are critical to support data science
- But: data engineering is not data science



Data Mining versus… Database Querying

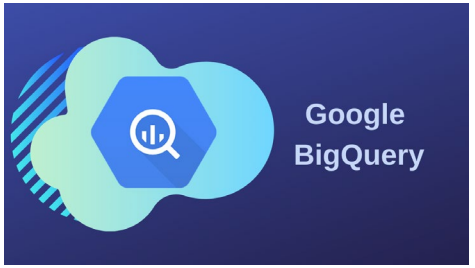
- A query is a specific request for a subset of data or for statistics about data, formulated in a technical language and posed to a database system
  - Structured Query Language (SQL)
  - Query-By-Example (QBE)

```
SELECT * FROM customers WHERE state='CA'
```

- No discovery of patterns or models
- Appropriate when an analyst already has an idea of what might be an interesting subpart of the data
- Extract the data you need for data mining (DM)

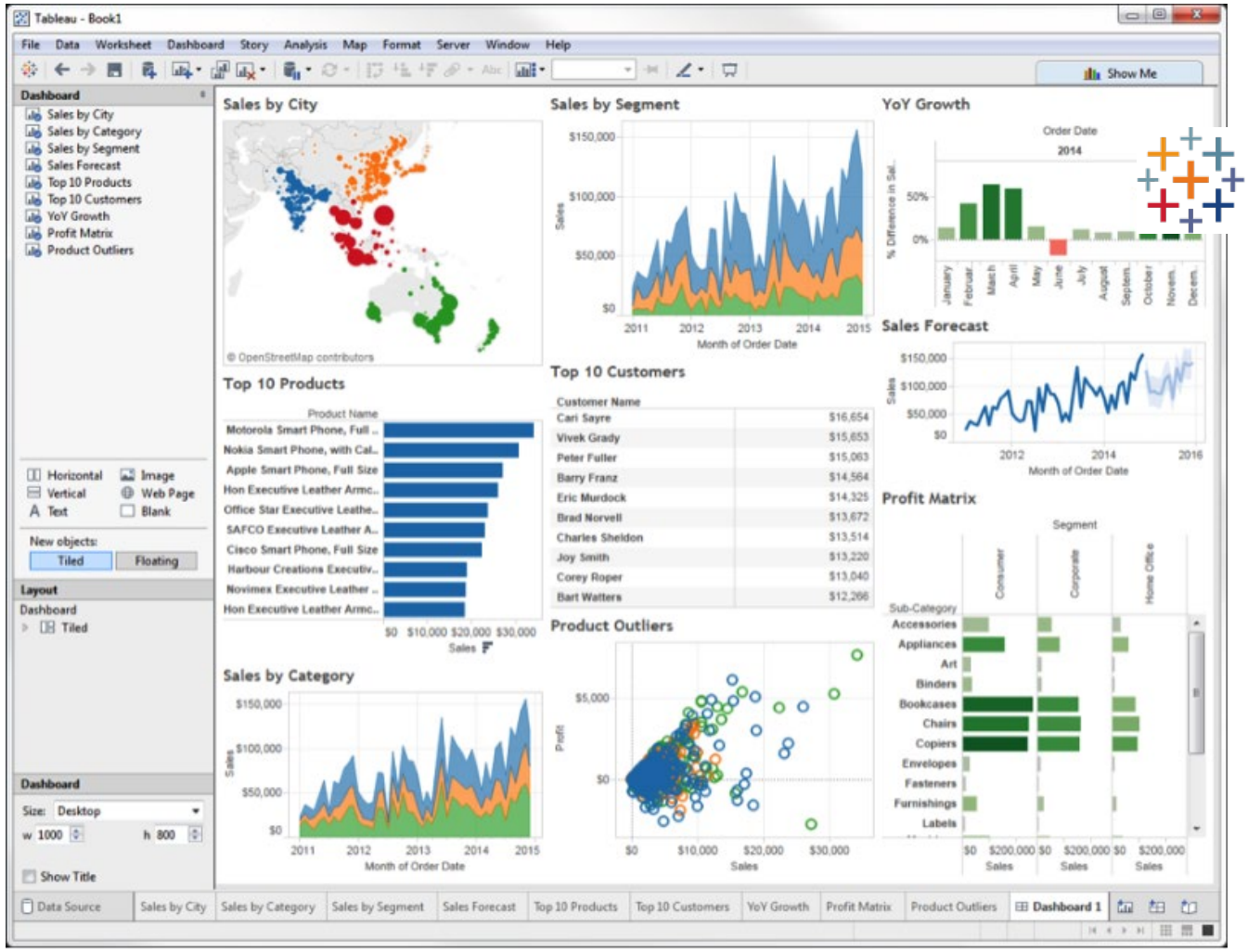
Data Mining versus… Online Analytical Processing (OLAP)

- **Online Analytical Processing (OLAP)** provides an easy-to-use GUI to explore large data collections, often in a Data Warehouse context
- Dimensions of analysis pre-programmed into the OLAP system -> evolved to columnar data lake
- No modeling or automatic pattern finding
- Tool for quickly answering **ad hoc** analytical questions
  - Online, interactive
  - Large, multi-dimensional data sets
  - Supports fundamental analytical needs



Data Mining versus… Digital Dashboards

BU. 520.650 Data Visualization



+ a b | e a u

Looker

Power BI

https://coronavirus.1point3acres.com/en/

Data Mining versus… Statistical Hypothesis Testing

- Understand different data distributions
- How to use data to test hypotheses
- Many of the DM techniques have their roots in statistics
- Quantification of uncertainty by confidence intervals

Data Mining (≈ Machine Learning)

- Objectives:
  - find **patterns** and **relationships**
  - **classify** and **predict**

- There is a large number of data mining algorithms available, but only a limited number of data mining tasks
  - Classification/ Regression/ Clustering
  - Co-occurrence grouping, Association Rule
  - Similarity matching / Profiling
  - Link prediction/ Data Reduction

- **Decompose** a data analytics problems into pieces if needed, such that you can solve a known task with a tool

- **Analysts' creativity** plays an important role

Q: Answering business questions with..

Let's consider a company that determines customer profitability based on the profit margins of products purchased over the past three months.

- Who are the most profitable customers?
  - Database querying
- Is there really a difference in profit margin between profitable customers and the average customer?
  - Statistical hypothesis testing
- But who really are these customers? Can I characterize them?
  - OLAP (manual search)
  - Data mining (automated pattern finding)
- Will a particular new customer be profitable? How much profit should I expect this customer to generate?
  - Data mining (predictive modeling)

Data Mining: Supervised vs. Unsupervised

- Is there a specific, quantifiable target (labeled) we are interested in or trying to predict?
  - If yes, supervised. Otherwise, unsupervised.

- Examples
  - "Do our customers naturally fall into different groups?"

    : no specific target → unsupervised

  - "Can we find groups of customers who have a particularly high likelihood of canceling their service soon after their contracts expire?"

    : specific target → supervised

  - "Can we predict show/no show of a specific patient to his/her appointment?"

    : specific target → supervised

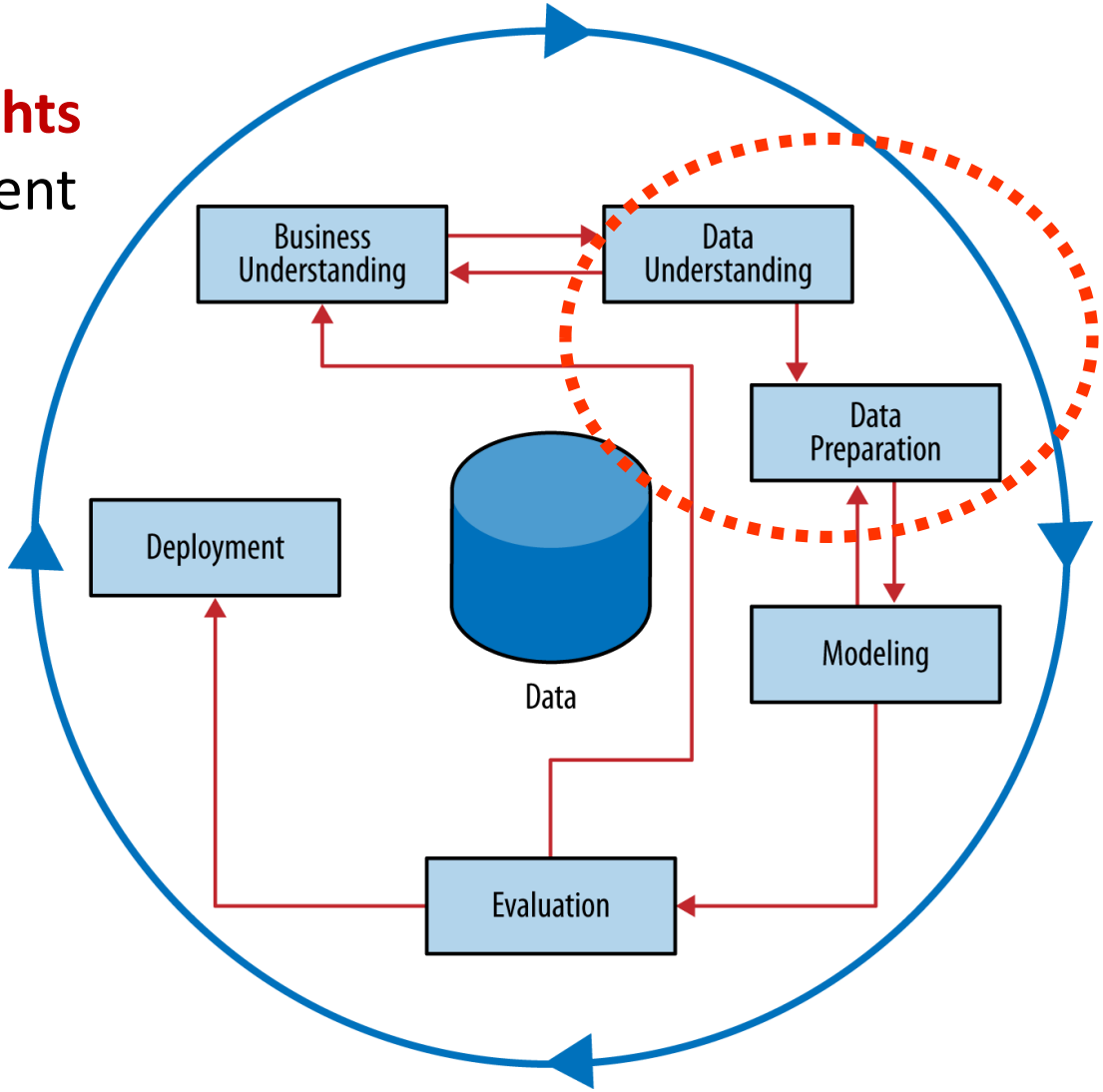
Supervised vs. Unsupervised

- Supervised: there is a specific target (or quantifiable target) that we want to predict. Because the outputs are labeled, we can measure the prediction accuracy. (Classification, regressions) have specific goals.
  - spam filtering, sentiment analysis
  - if a customer will respond to a promotional campaign
  - if a customer is going to churn
- Unsupervised: we can think about the opposite case – no specific target to predict, and data are not labeled. (Clustering, Association rules) mostly used to get insights
  - Is there any natural segmentation among our customers? (no specific goal or target)
  - Are there any differences between the groups of tissues? (no labeling, no idea what to expect)

Understanding and Preparing Data

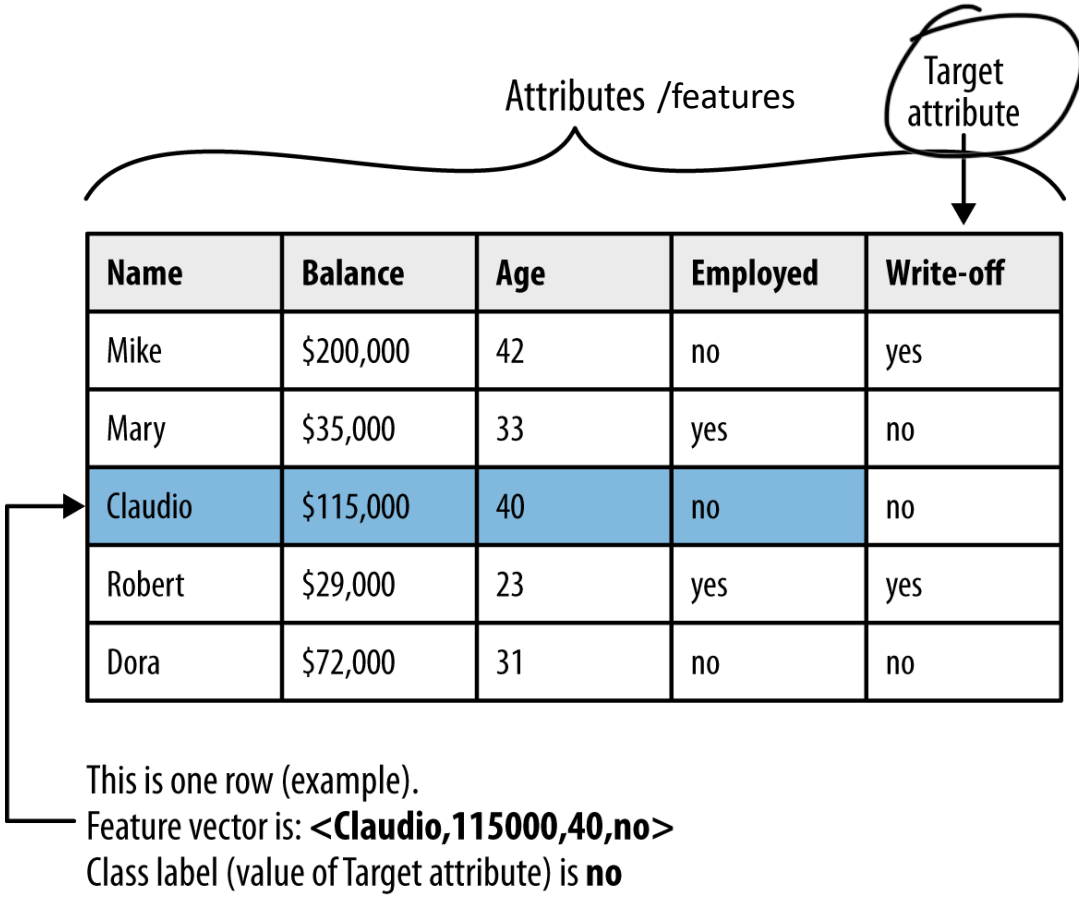
Data Understanding & Preparation

- Goals: Gain **general insights** about the data (independent of the project goal)



Terminology

- We often assume that the data set is provided in the form of a simple table
- The columns of the table → **attributes, features or variables**
- The rows of the table → **instances, examples, records, or feature vectors**
- A data set: A set of examples or data records



Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

Cross-sectional, Time-series, Panel

- Cross-sectional: Snapshot of several individuals ← Our Focus!
- Time-series: Over-time observations a single individual

- Panel = Cross-sectional + Time-series

Tidying Data with *tidyr*

- Principles of tidy data
  - Observations as rows
  - Variables as columns
  - One type of observational unit per table (person or pet, not both)

name	age	eye_color	height
Jake	34	Other	6'1"
Alice	55	Blue	5'9"
Tim	76	Brown	5'7"
Denise	19	Other	5'1"

Observation

Variable or Attribute

Hadley Wickham wrote a paper in the journal of statistical software called tidy data which summarizes these concepts in a clear and concise way.

Tidy enough? - Diagnosis

name	age	brown	blue	other	height
Jake	34	0	0	1	6'1"
Alice	55	0	1	0	5'9"
Tim	76	1	0	0	5'7"
Denise	19	0	0	1	5'1"

- Column headers are values, not variable names…
- Can use ‘Gather’ function (tidyr) to consolidate them

Data Scientists Spend 45-80% of Their Time On

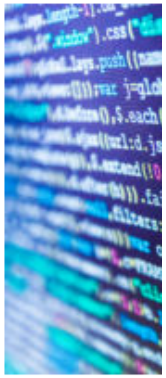
- Searching for data
- Correcting errors
- Verifying correctness

It is inevitable.....

July 6, 2020

Data Prep S

Alex Woodie



(BEST-

time, while model
the survey found.

Monica Rogati

Peter DaSilva for

Data scientists still spend too much of time cleaning data instead of creating actionable insights.

Frederic Jacquet

AI & Ethics | Digital Experience | Advanced technologies & Quantum Computing

June 5, 2022

Successful building data warehouses best practices shows no reason to tolerate an unstructured approach of the ETL process

Ralph Kimball says in his "***The Data Warehouse Toolkit***" that "*When asked about the best way to design and build the ETL system, many designers say, "Well, that depends." It depends on the source; it depends on limitations of the data; it depends on the scripting languages and ETL tools available; it depends on the staff's skills; and it depends on the BI tools. But the "it depends" response is dangerous because it becomes an excuse to take an unstructured approach to developing an ETL system, which in the worse-case scenario results in an undifferentiated spaghetti-mess of tables, modules, processes, scripts, triggers, alerts, and job schedules.* ***This "creative" design approach should not be tolerated.*** *With the wisdom of hindsight from thousands of successful data warehouses, a set of ETL best practices have emerged. There is no reason to tolerate an unstructured approach."*

> "Let's dump the sh#t in any Data Warehouse or Data Lakehouse, just like this, we'll figure out later how to make it actionable."

Find Dirty Stains in This Data!

Customer ID	Customer First Name	Customer Last Name	Address	City	State	Zip	Phone
1771	Larry	Shimk	143 S.	Denver	NY	178908	911
1771	Caroline	Shimk	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Shimk	Caroline	143 N. West St.	Buffalo	NY	14321	716-333-4567
1772	Heather	Schwiter	55 N. W. S. Miss	LaGrange	GA	14321	716-333-4567
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1772	Debbie	Fernandez	S. Main St.	Denver	CO	80252	333-8965
1773	Justin	Justin	34 Kerry Rd.	Littleton	CO	98987	716-67-9087
1774	Pam		66 S. Carlton	North Glen	CO	98765	343-456-6857
1775	D.	Fernandez	3514 S. Main	Denver	CO	80252	303-333-8965
1776	PepsiCo		15365 K St. NW	Washington	DC	20035	202-353-1535
1777	Sam	Esteban	4413 Madison Rd	Ann Arbor	MI	48109	734-140-2531 ext 354
1778	Caroline	Smith	143 N. West St.	Buffalo	NY	14321	716-333-4567

Why Do Data Get Dirty? (2/2)

- Think of Ms. Pamela Smith O'Brien
  - How many different names can she have?
- How about an address?
  - How many different addresses can be valid?
- Measurement can be inaccurate
- The question may be wrong or ambiguous
  - Phone number – home, work, or cell?
- The question can be answered inconsistently

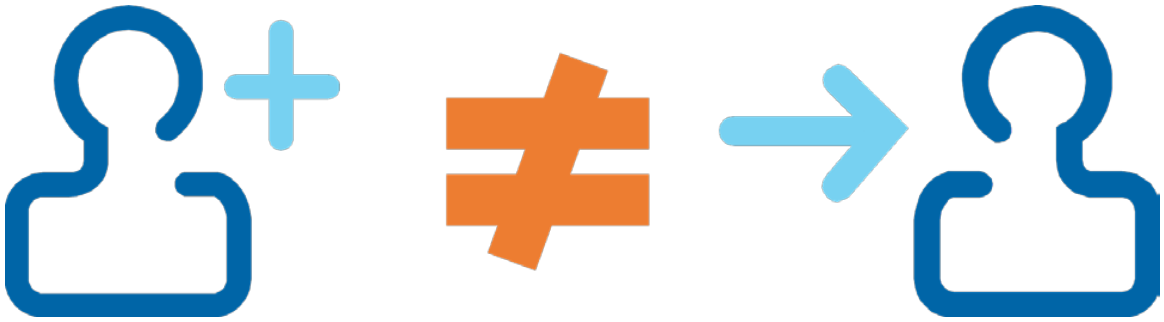
Pamela O'Brien
Pamela S. O'Brien
Pamela Smith O'Brien
O'Brien, Pamela S.

Why Does Data Get Dirty? (3/3)

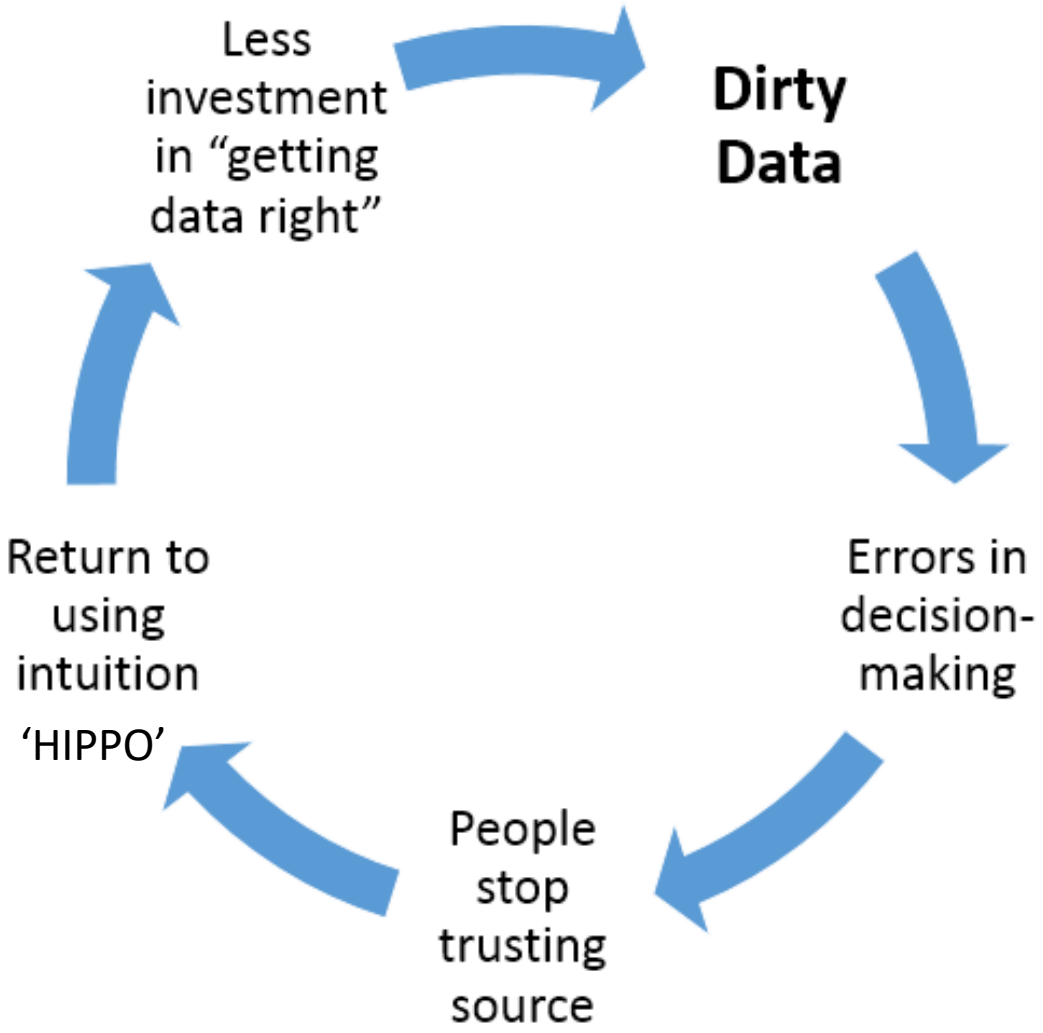
- ”The Agency Problem”

- The data creator is usually not the data consumer.
  - Data creator – sales, customer service
  - Data consumer –Marketing department, DS team

- When the creator doesn’t care much about how the data will be used, data are likely to get dirty.



Vicious Cycle from Dirty Data

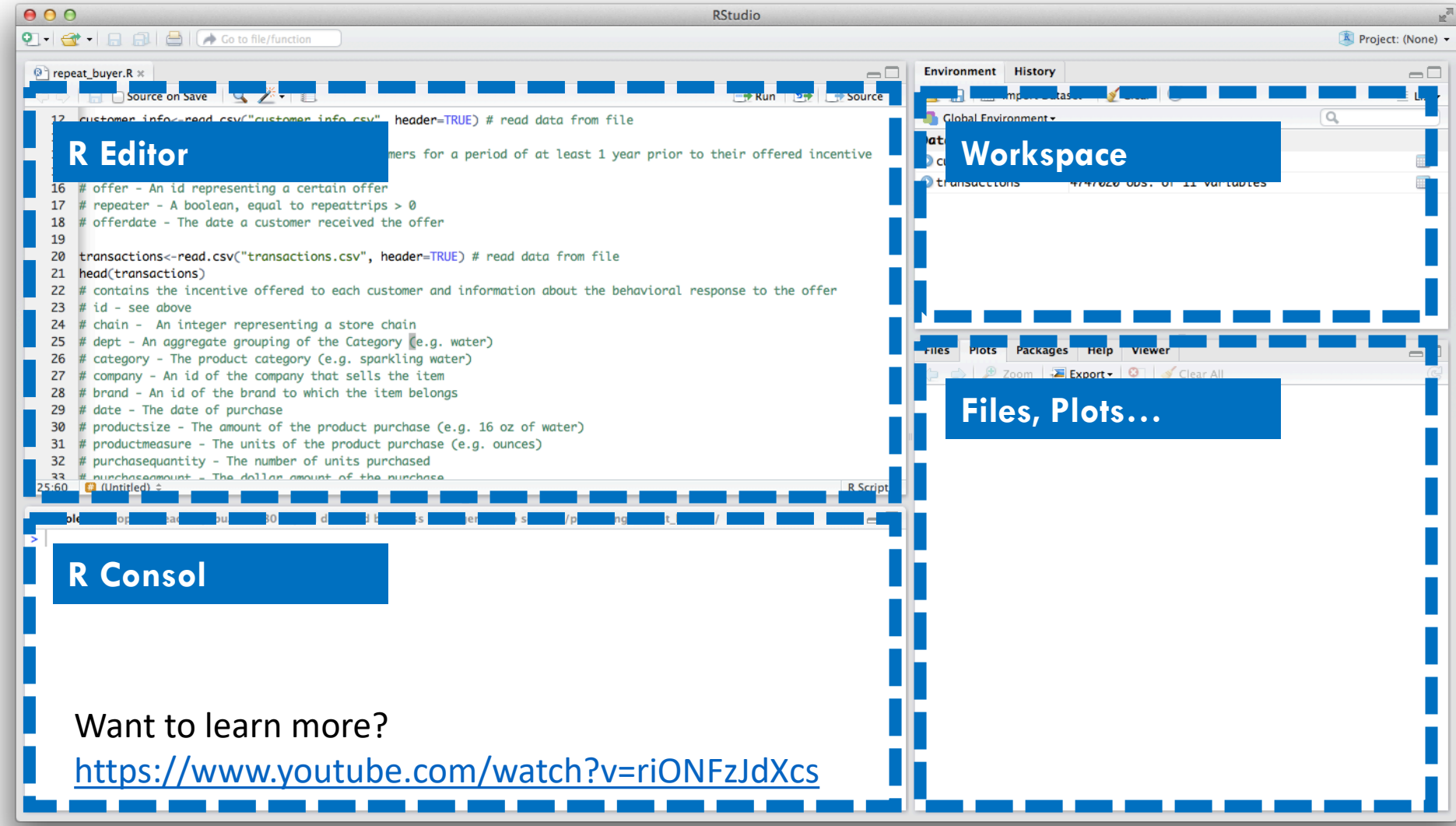


Today's R Session

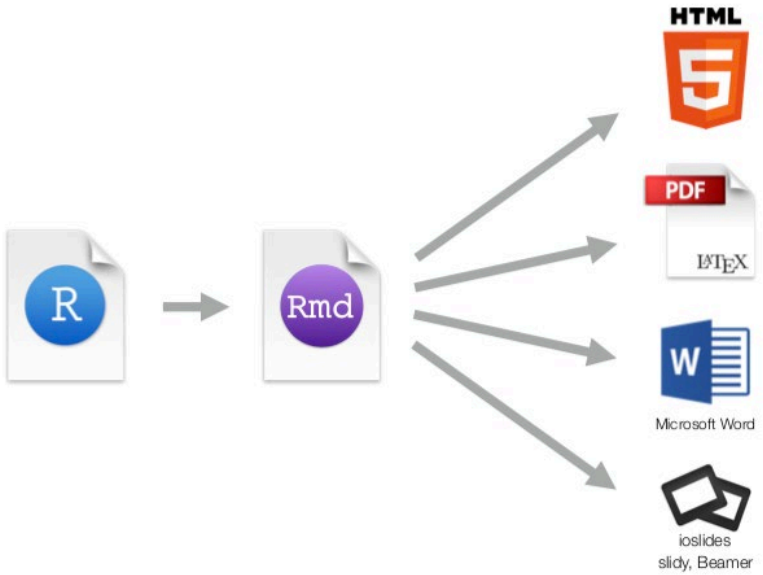
- Tools:
  - Reporting: R Markdown
  - Key Package: Tidyverse

- Tasks:
  - Explore raw data
  - Tidying data
  - Prepare data for analysis

R and R studio

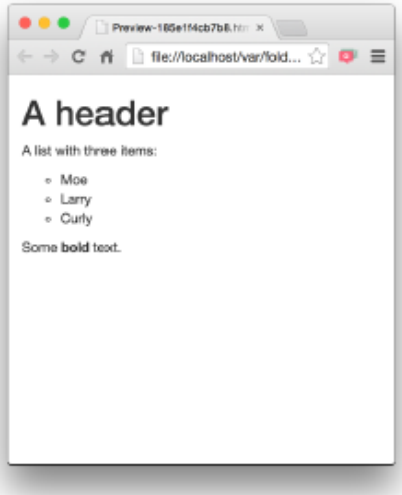


R Markdown



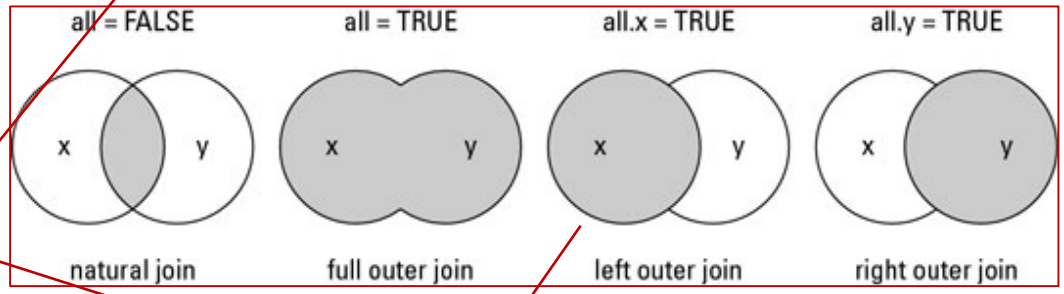

```
<!DOCTYPE html>
<html>
<head>
 <title>Page Title</title>
</head>

<body>

 <h1>A header</h1>

 <p>A list with three items:</p>
 <ul style="list-style-type:circle">
  <li>Moe</li>
  <li>Larry</li>
  <li>Curly</li>
 </ul>

 <p>Some <strong>bold</strong> text.</p>

</body>
```

html

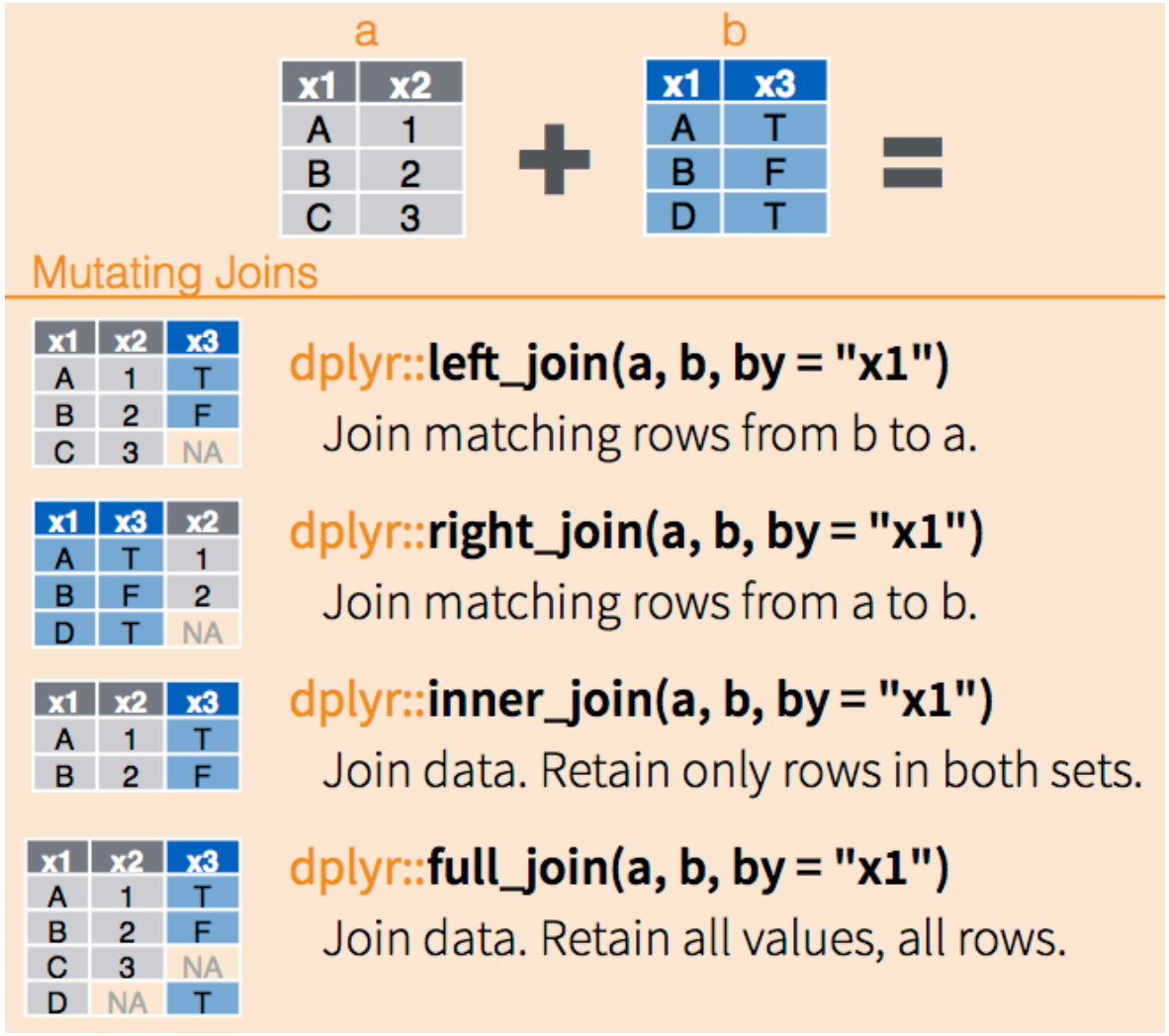


Merging dataset (Basic R)

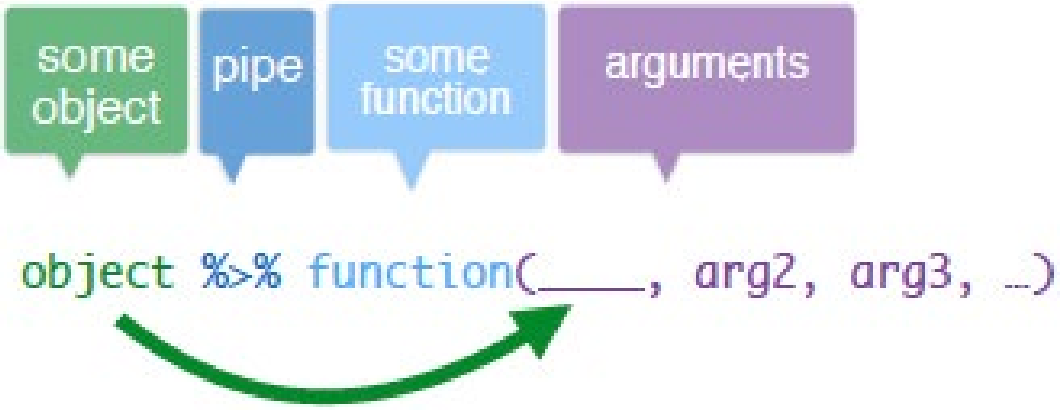
```
> head(order)
  customer_id order_id order_short_date order_total zip_code
1           1        1       2009-01-01      404.72    32435
2           1    33658       2013-02-15       71.10    32435
3           2        2       2009-01-01      288.62     1099
4           2     1653       2010-01-03      182.86     1099
5           2     4193       2010-09-12      108.43     1099
6           3        3       2009-01-01       27.15    66063
> head(zipcode)
  zip state
1 501    NY
2 544    NY
3 601    PR
4 602    PR
5 603    PR
6 604    PR
> order<-merge(order,zipcode,by.x="zip_code",by.y="zip",all.x=TRUE)
> head(order)
  zip_code customer_id order_id order_short_date order_total state
1    10013        3881     7382       2011-03-26       93.68    NY
2    10013        3881    11784       2011-09-30      360.39    NY
3    10013        3881    50215       2013-09-21       23.12    NY
4    10017         568      712       2009-08-06      237.23    NY
5    10017         568    24273       2012-09-04       30.08    NY
6    10017         568    21458       2012-07-05       34.35    NY
```



Merging dataset (in a tidy way)



Pipe Operation (%>%)

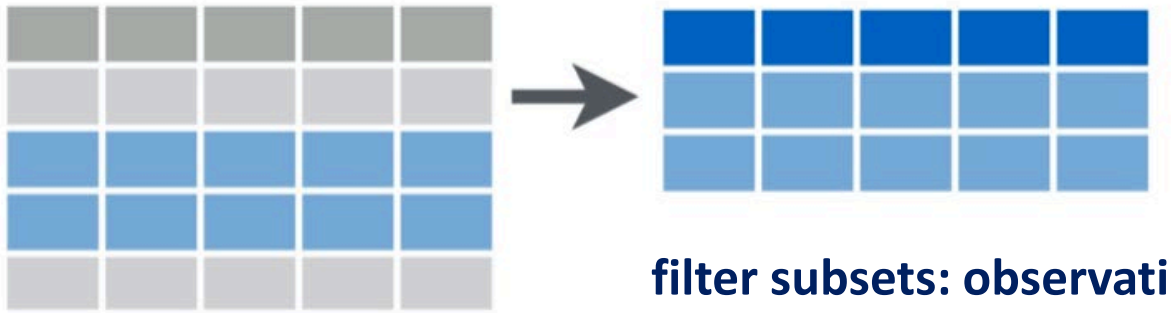


For example,

filter(data, state == “CA”) → data%>%filter(state == “CA”)

Filter(), Select() – Creating Subset

filter( )



filter subsets: observations

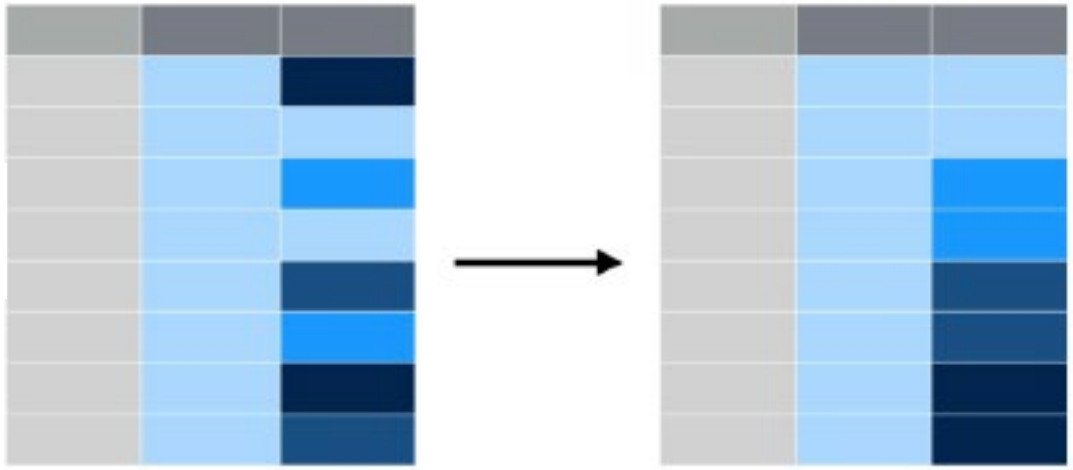
select ( )

filter subsets: variables

Arrange() and Mutate()

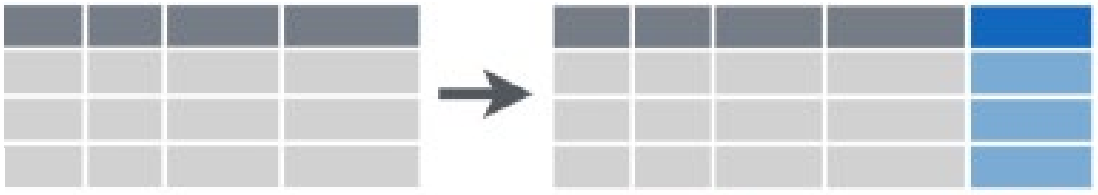
arrange( )

:sorts a table based on a variable

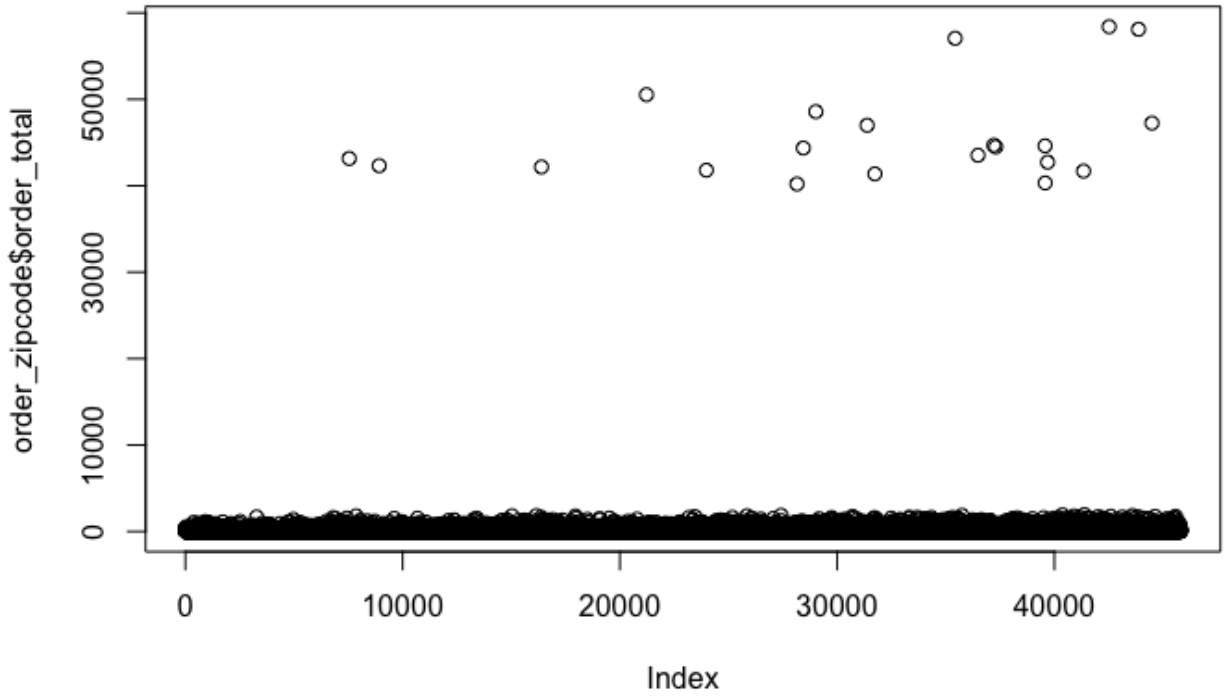


mutate( )

:changes or adds variables

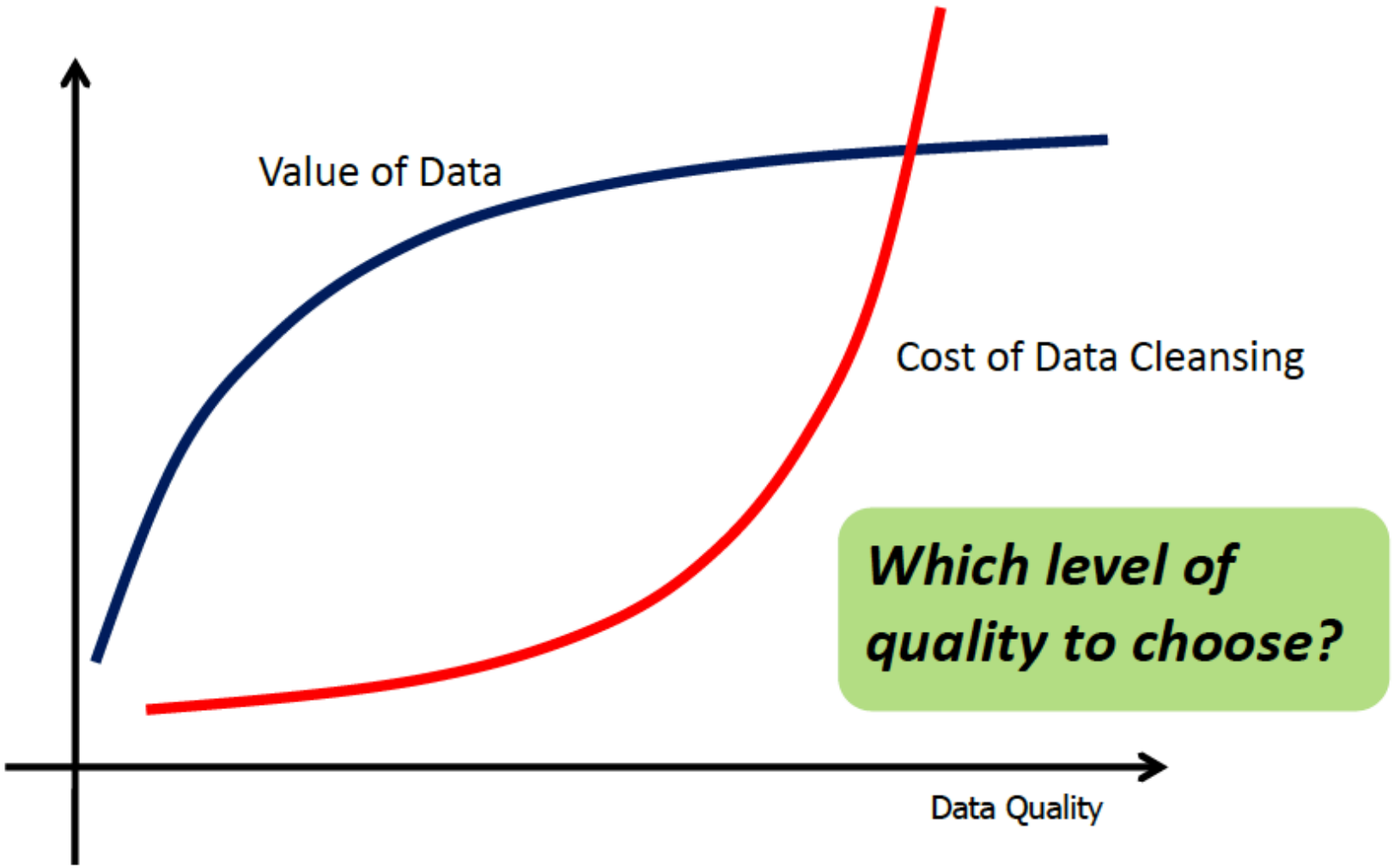


Be Careful in Cleansing Data

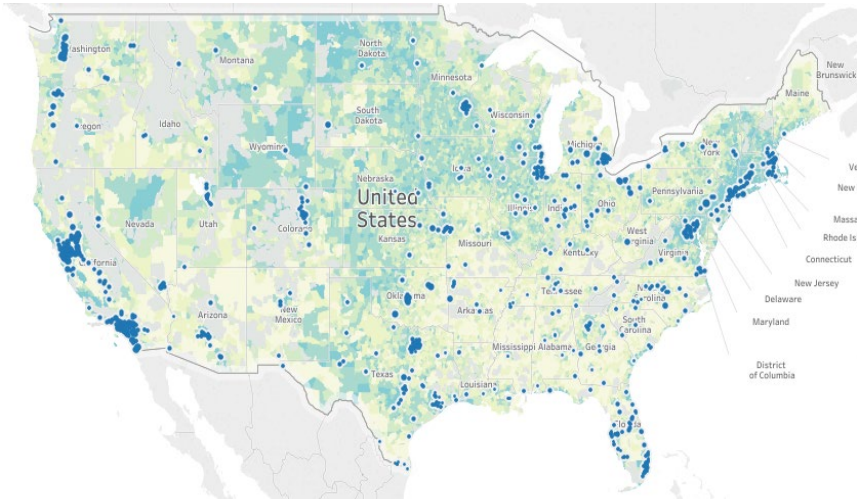


- Are they outliers?
- Do we need to “fix” it?
- It depends on cause.
- If yes, how can we tell if it is a real error or an unusual but correct data?

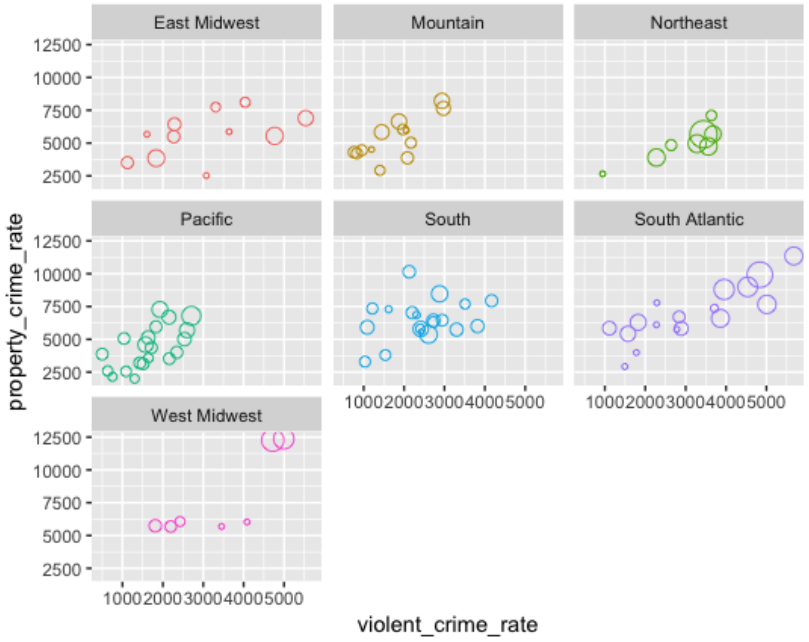
Trade-Off in Data Cleansing

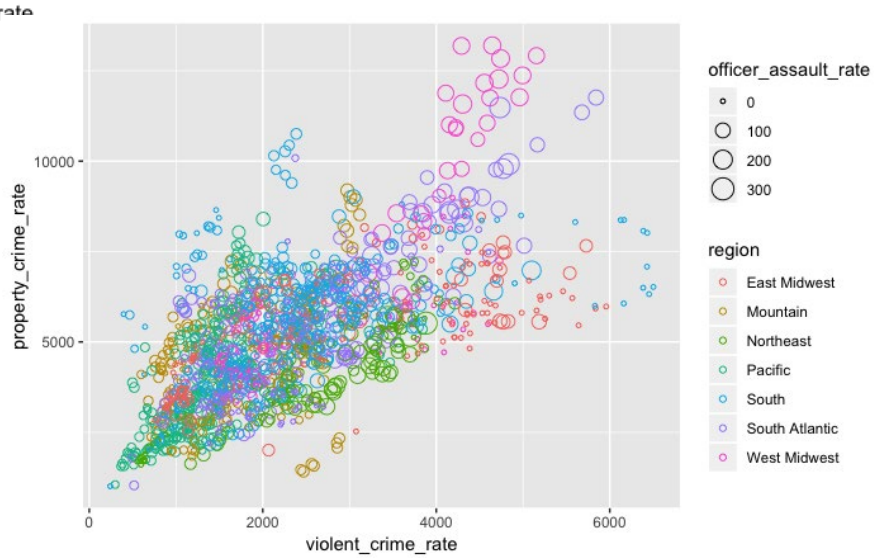


Preview of Next Week..



- More about Data Mining Tasks
- Data Visualization
  - R (ggplot2) & Tableau
  - Be sure to have Tableau ready on your computer!





JOHNS HOPKINS CAREY BUSINESS SCHOOL