

Introduction to Data Science

Lecture 1; October 5th, 2016

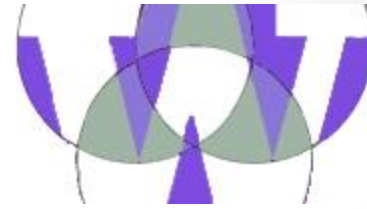
Ernst Henle

ErnstHe@UW.edu

Skype: ernst-henle

(1)

Agenda



- Announcements
 - Optional class on programming in R on October 8th from 9:00 AM to 12:00 noon. Use the following link (note the “b” at the end of the link):
<http://uweoconnect.extn.washington.edu/datasci250b>
 - Guest Lecture: The business side of data science by Marius Marcu
 - Guest Lecture: Data Science Trends in the Professional World by David Porter and Emily Nichols on Oct 19th 2016
- Introductions (Social aspect of this course)
- Break
- Class Structure
- What is Data Science?
- Quiz 01a Class Structure and In-class Assignment
- R Basics
- Quiz 01b (Basic R)
- Data Preparation
- Break
- Data Preparation in R
- Assignment (Complete all assignments items from all assignment slides. Must be submitted by Saturday 11:57 PM)

Introductions

Introductions (1)

Welcome to Data Science

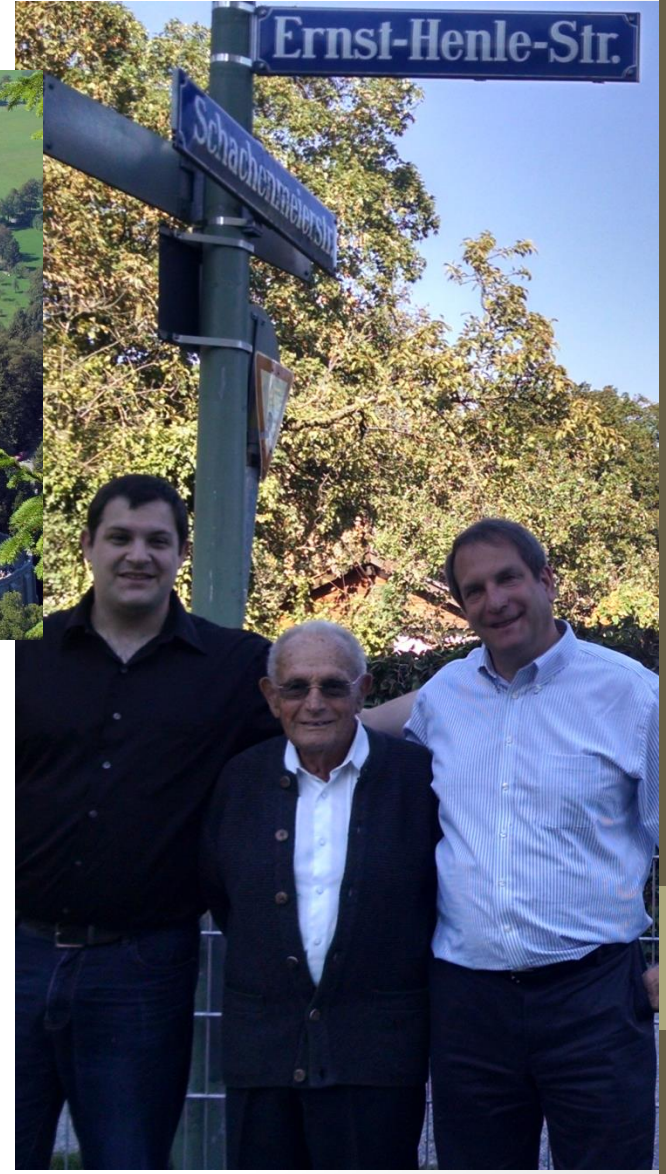
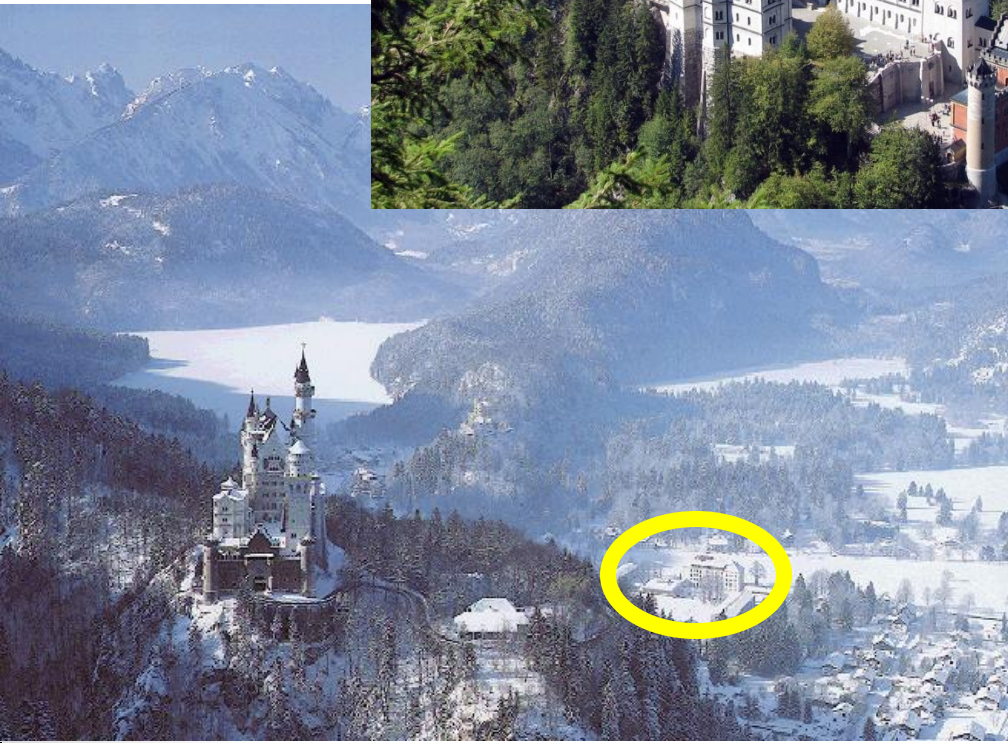
- A major component of this certificate program is the social component.
- Let's introduce ourselves.
 - Name
 - Professional background and interest in data science
 - Something personal



Introductions (2) My Background

- Ernst Henle
 - Email: ErnstHe@UW.edu
 - Skype: ernst-henle
- My interests
 - I use informatics, physics, math, and chemistry to solve problems in medicine
 - As an experimental scientist
 - My Ph.D. is in Biophysics. I studied: Ageing, DNA Damage, and Free Radicals
 - As a “data scientist”
 - Drug discovery using Genomics, Proteomics, and Metabolomics (these “-omics” are the equivalence of “data science”)
 - Adapt BI tools for medicine: IoT and Predictive Analytics
- Personal
 - I grew up in Bavaria and lived close to the castle that was the model for Disneyland's castle.
 - In Munich, there’s an Ernst Henle street, which was named after my grandfather.

Introductions (3) My Background



Introductions (4)

TA: Marius Marcu

- Marius Marcu is a strategic innovator with a great passion for high-tech. His career path includes product management and product marketing roles with big companies like Intel and Microsoft, but also nimbler, high-growth startups like Smartsheet.
- While at Microsoft, Marius fell in love with big data, cloud technologies and data science, which he thinks will make a lot of people rich in the next 10 years, including himself. He is an alumnus of this certificate program and he is very passionate about leveraging data power to drive growth in the enterprise technology business. Data Science professionals, like Marius, who come from a business background have unique opportunities. We are fortunate to have Marius as our TA
- <https://www.linkedin.com/in/mariusmarcu>
- mariusmarcu@global.t-bird.edu

Introductions (5)

Everybody

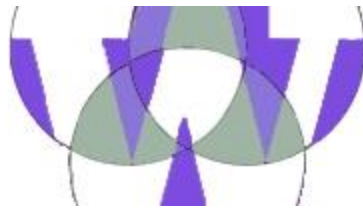
- Name
- Professional interests and interest in data science
- Something personal

Introductions (6)

Discussion Forum

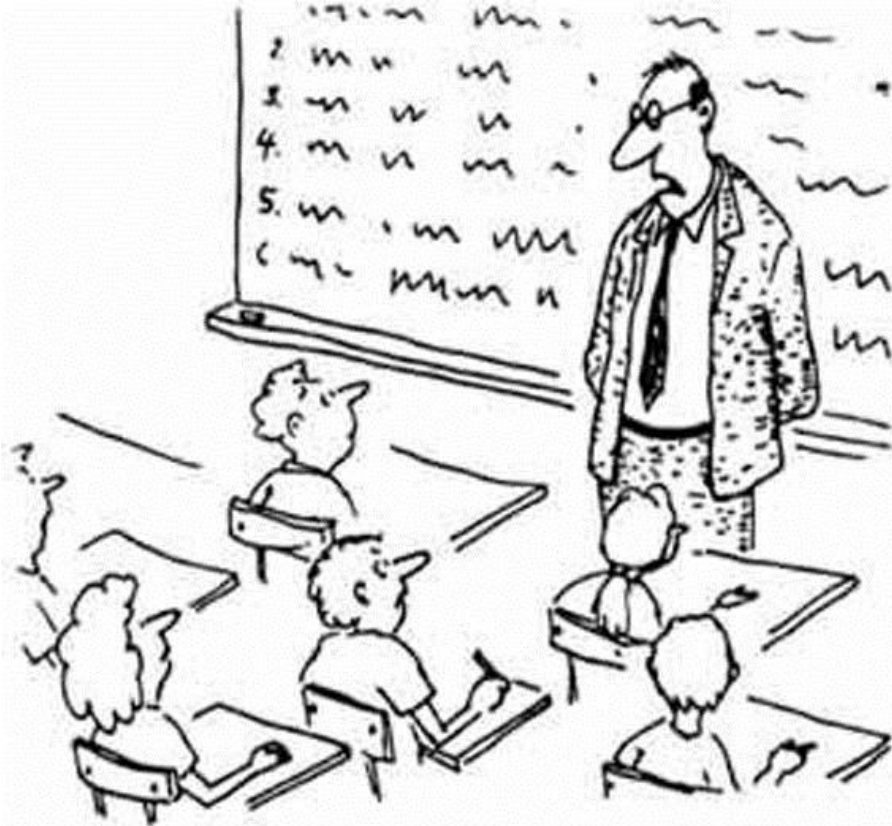
- A major component of this certificate program is the social component.
- Visit the LinkedIn group: **UW 2016 Data Science**. Comment on an ongoing discussion or create a new discussion

(<https://www.linkedin.com/groups/12008773/profile>)



Introductions

Break



"I expect you all to be independent, innovative, critical thinkers who will do exactly as I say!"

Class Structure

Class Structure (1)

- Quizzes
 - In-class quizzes are designed to re-inforce lecture material and provide clues for the assignments. You must get more than 50% of all quiz questions correct to pass the class. Generally, I will provide a quiz preview more than one day before class.
- Assignments
 - You must get at least 4/10 points on 7 of the 8 assignments. A homework with less than 4/10 points is considered “insufficient”.
 - A late assignment counts for maximally half. You must notify me before the deadline that you will hand in your assignment late.
 - Assignments are due by Saturday 11:57 PM for full credit.
 - Assignments should be done **collaboratively**. You will get more out of your assignment if you do it with a fellow student.
 - Assignments must be submitted individually by each student to the Canvas web site.

Class Structure (2)

- Class Prerequisites include:
 - UW NetID
 - Access to the Internet
 - Access to Canvas: <https://canvas.uw.edu>
 - Ability to use R/R Studio, Python, and other programs on your Computer
 - Ability to run VMWare or VirtualBox on your Computer
 - Ability to participate in the LinkedIn group called **UW 2016 Data Science**
 - Ability to submit homework to the Canvas Assignment submission sites
 - Ability to use Canvas
 - Get resources from class module where you can find Lesson Overviews, Assignment submission sites, and quizzes: <https://canvas.uw.edu/courses/1092103/modules>
 - For Example, go to the modules, expand Lesson 01 and click on Lesson 01 Overview. Download the items listed under Lecture Materials
 - For Example, go to the modules, expand Lesson 01 and click on Assignment 00 (In-class Submission).. See the submission site for the in-class assignment.
 - For Example, go to the modules, expand Lesson 01 and click on Quiz 01a (Class Structure) . Take the quiz.
 - Get Class Recordings here: <https://canvas.uw.edu/courses/1092103/pages/course-introduction>

Class Structure (3)

- Misc
 - Optional class on programming in R on Saturday October 8th 2016 from 9:00 AM to noon. Use this link and note the extra “b” at the end: <http://uweoconnect.extn.washington.edu/datasci250b>
 - Office hours by appointment. If requested, I can also establish office hours at a specific time each week. I use Skype (ernst-henle).

Class Structure (4)

Approximate Course Agenda

#	Date	Topics
01	Oct 5 th 2016	Introductions; What is Data Science? Class Structure; R; Data Preparation
R	Oct 8 th 2016	Optional on-line class on the basics of R: http://uweoconnect.extn.washington.edu/datasci250b
02	Oct 12 th 2016	Machine Learning; K-Means; Effect of Normalization on Linear vs Non-linear models; Clustering
03	Oct 19 th 2016	Supervised Learning; Classification; Data Science Trends in the Professional World
04	Oct 26 th 2016	Real World Predictive Analytics; Statistics; Sampling; ROC; Confusion Matrix; Accuracy Measures
05	Nov 2 nd 2016	RDBMS with Relation Algebra; Data Science – The business point of view
06	Nov 9 th 2016	Data Storage Concepts; Cloud Computing; CAP Theorem; No SQL; Scalability; Sparse Matrices and EAV
07	Nov 16 th 2016	Hadoop HDFS Sqoop Hive Impala Hue
	Nov 23 rd 2016	No Class! Thanksgiving
08	Nov 30 th 2016	Hadoop MapReduce exercise; Graph Data; SPARQL; Visualizations in Data Science: A developers perspective
09	Dec 7 th 2016	Page Rank; Web structure;
10	Dec 14 th 2016	Spark; Data Science: Coupling data selection, data preparation, and modeling.

Class Structure

What is Data Science?

What is Data Science? (1)

First measure, second estimate quantity, third calculate, fourth balance chances, fifth succeed.

By performing a large number of calculations you will enjoy victory but if you have only a small number of calculations you will suffer defeat.



Sun Tzu, *The Art of War*, 500 BC

What is Data Science? (2)

- My answer:
 - Data science is the generalization of the scientific method



What is Data Science? (3)

- My answer:
 - Data science is the generalization of the scientific method
- Why is data science a “new” discipline?
 1. Abundance of data outside of the traditional sciences
 2. Tools to investigate these data



What is Data Science? (4)

- My answer:
 - Data science is the generalization of the scientific method
- Why is data science a “new” discipline?
 1. Abundance of data outside of the traditional sciences
 2. Tools to investigate these data
- Decompose “Data Science” into “Data” and “Science”
 - What is Science?
 - What is Data?



What is Data Science? (5)

Science Past vs. Future

- Past
 - Science used to be about devising the best experiment to verify a specific hypothesis.
 - Data acquisition was coupled to a specific hypothesis

What is Data Science? (6)

Science Past vs. Future

- Past
 - Science used to be about devising the best experiment to verify a specific hypothesis.
 - Data acquisition was coupled to a specific hypothesis
- Future
 - The abundance of data has led to a new paradigm
 - Data is ubiquitous
 - We need methods to sift through the data and extract meaning.
 - Existing data supports many hypotheses

What is Data Science? (7)

Example Genomics

- “Omics” are the data-intense technologies in Bio-medicine
- First Human Genome
 - Time: 12 years (1990 – 2002) to sequence first genome
 - Cost: 3 Billion \$
 - Use in Science: Too expensive to devise an experiment that requires whole-genome sequencing

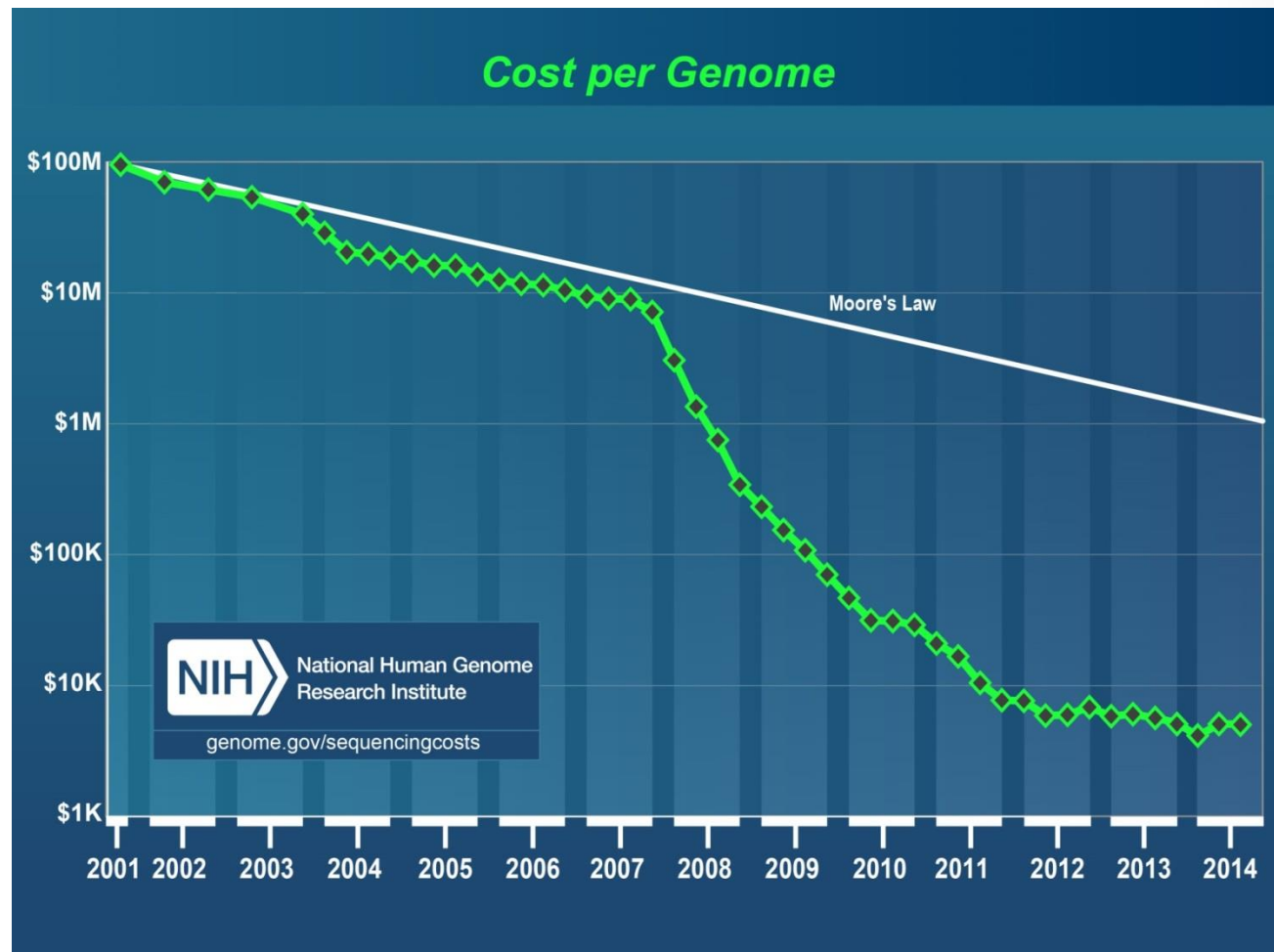
What is Data Science? (8)

Example Genomics

- “Omics” are the data-intense technologies in Bio-medicine
- First Human Genome
 - Time: 12 years (1990 – 2002) to sequence first genome
 - Cost: 3 Billion \$
 - Use in Science: Too expensive to devise an experiment that requires whole-genome sequencing
- Today
 - Time: 24 hours
 - Cost: \$1000.
 - Use in Science: Most investigators do not even need to pay for sequencing since enough sequences already exist
 - Expense and time are spent on sifting through the data.

What is Data Science? (9)

Example Genomics



What is Data Science? (10)

The Scientific Method

What is Data Science? (11)

The Scientific Method

1. A hypothesis is formulated that explains observations

What is Data Science? (12)

The Scientific Method

1. A hypothesis is formulated that explains observations
2. The hypothesis is tested
 - A verified hypothesis constitutes a theory


What is Data Science? (13)

The Scientific Method

1. A hypothesis is formulated that explains observations
2. The hypothesis is tested
 - A verified hypothesis constitutes a theory
3. New observations are considered in light of the theory

What is Data Science? (14)

The Scientific Method

- 
1. A hypothesis is formulated that explains observations
 2. The hypothesis is tested
 - A verified hypothesis constitutes a theory
 3. New observations are considered in light of the theory

What is Data Science? (15)

The Scientific Method in Data Science

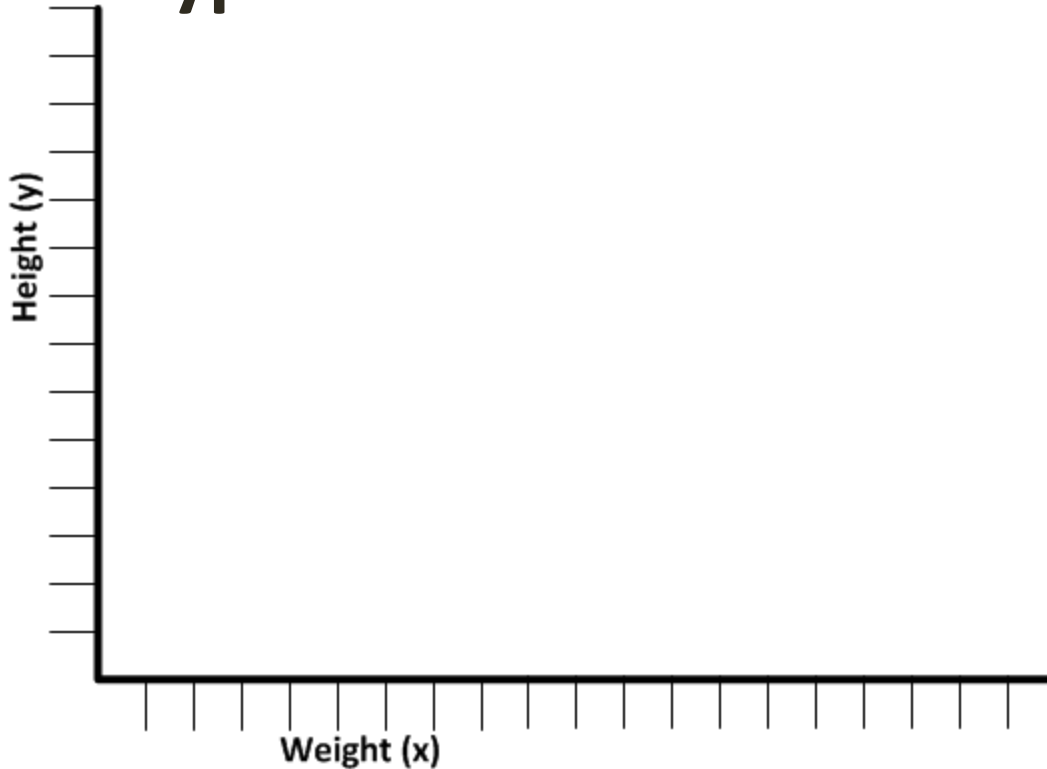
1. A data scientist receives a request from a hospital administrator:
 - Please, predict hospital readmissions!
2. The data scientist works with the hospital administrator to formulate a question:
 - "Can we predict hospital readmissions based on patient data?"
3. The data scientist re-works the question as a well-formed hypothesis. Formulating this hypothesis is at the heart of the project:
 - Inpatient hospital readmissions at Mercy General can be predicted with more than 85% certainty using a decision tree based on available patient data.
4. The hypothesis is tested
 - The testing requires data transformations
 - Transformations are verified by accuracy assessments
5. If the hypothesis is verified (withstands falsification) ask more questions like:
 - Why are patients being readmitted?
 - Can we prevent readmissions?
 - How can we increase our accuracy from 85% to 90%.

What is Data Science? (16)

Model as a Hypothesis

What is Data Science? (17)

Model as a Hypothesis

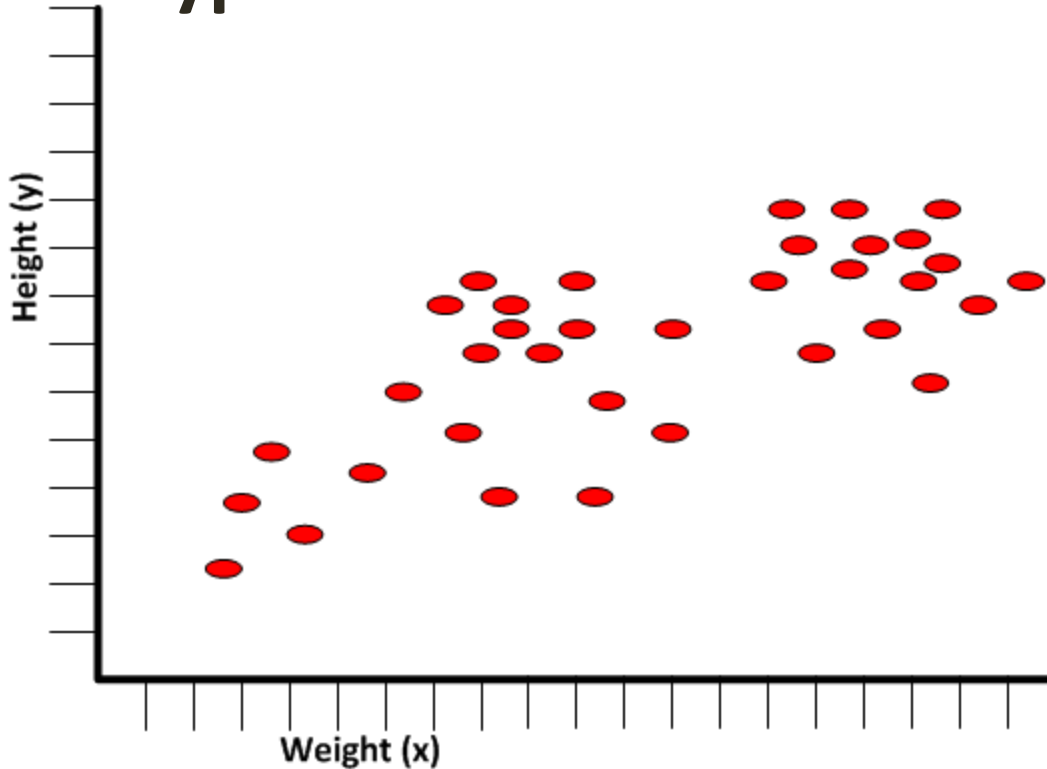


Schema, Space

This is a two dimensional space of
weight and height

What is Data Science? (18)

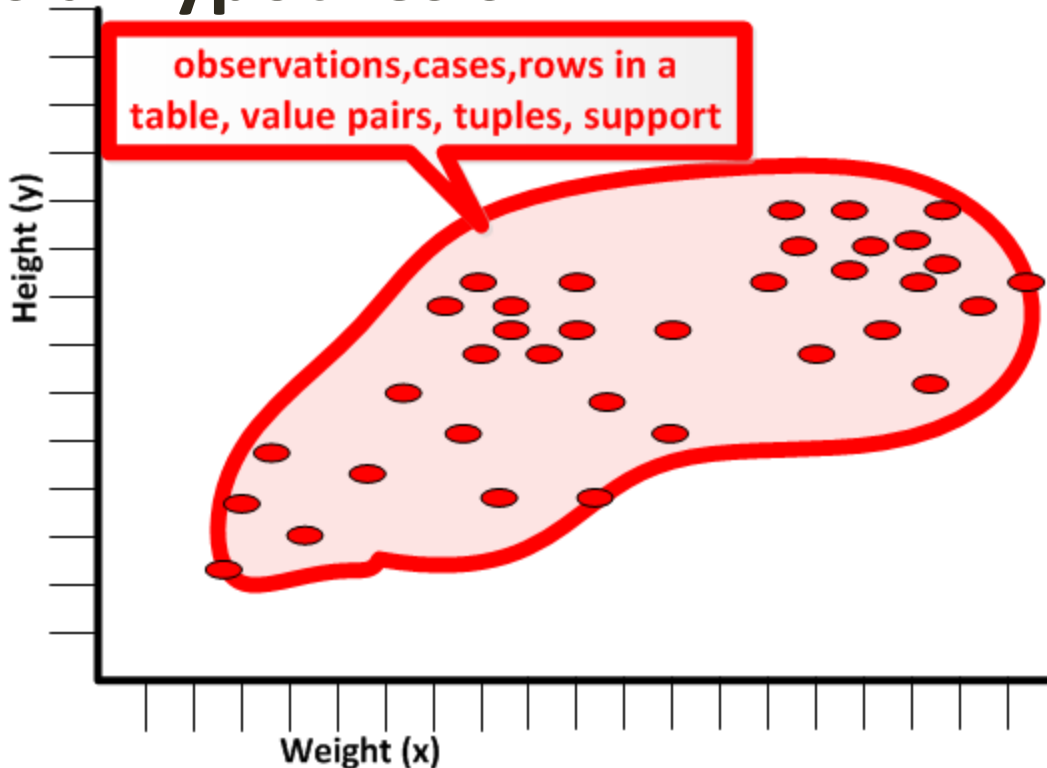
Model as a Hypothesis



Data: These data represent observed people.

What is Data Science? (19)

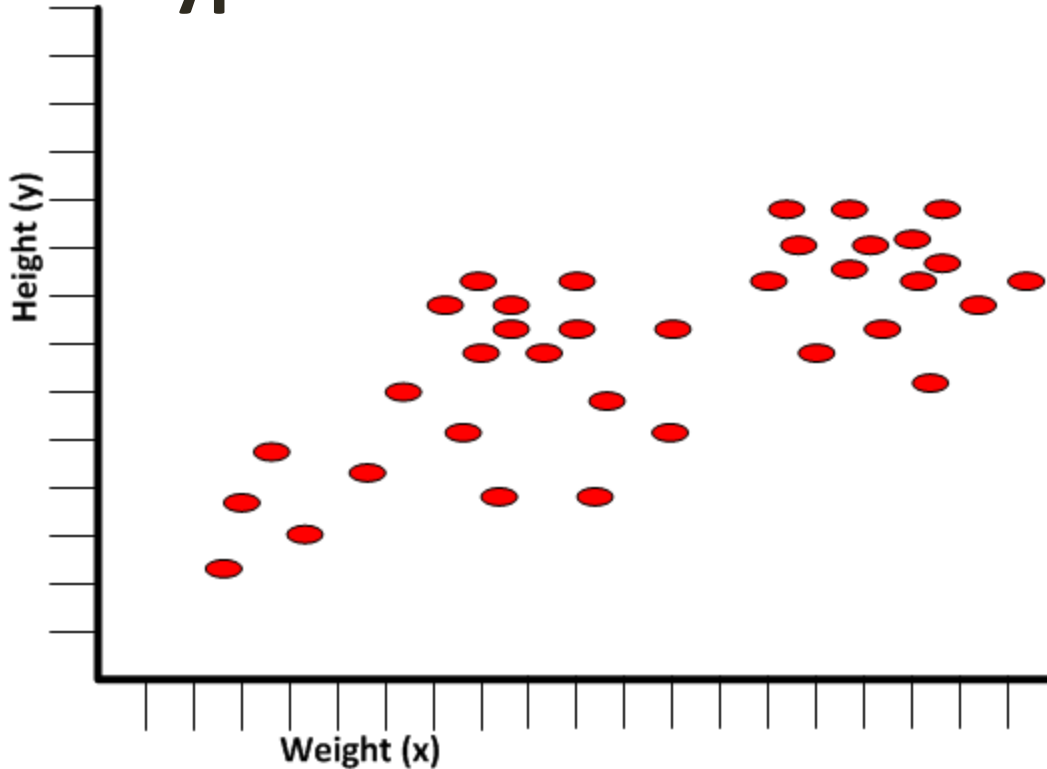
Model as a Hypothesis



Data: These data represent observed people. The data can be represented as a point in two-dimensional space

What is Data Science? (20)

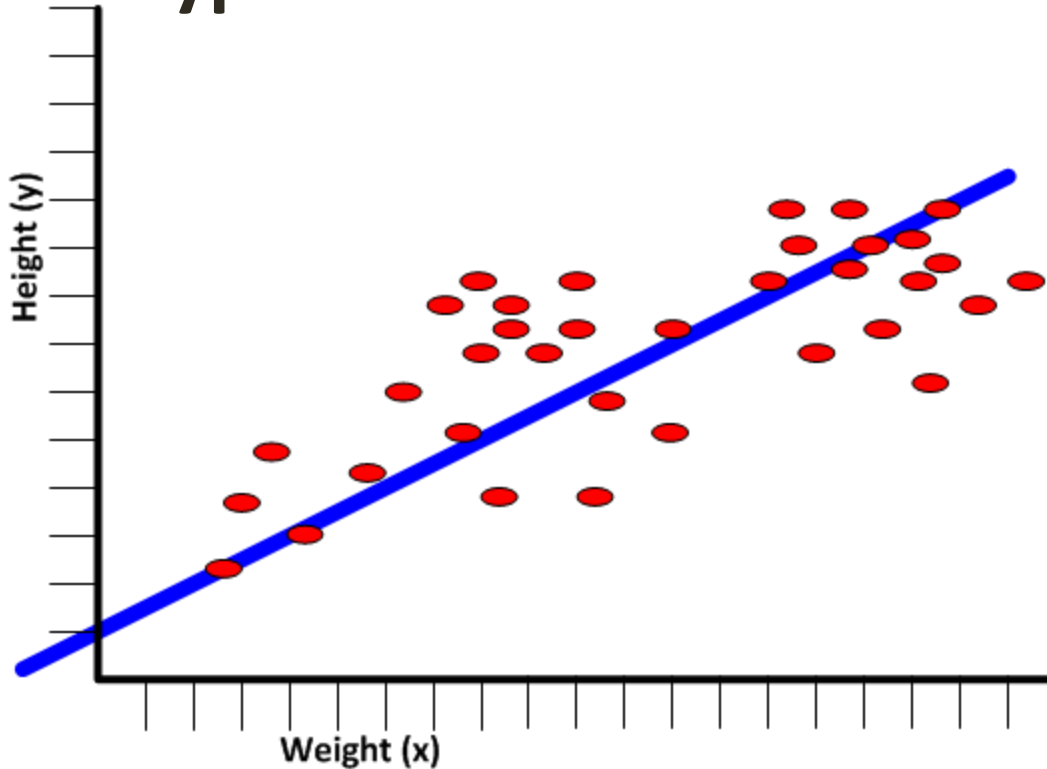
Model as a Hypothesis



Use Data to Create Hypothesis

What is Data Science? (21)

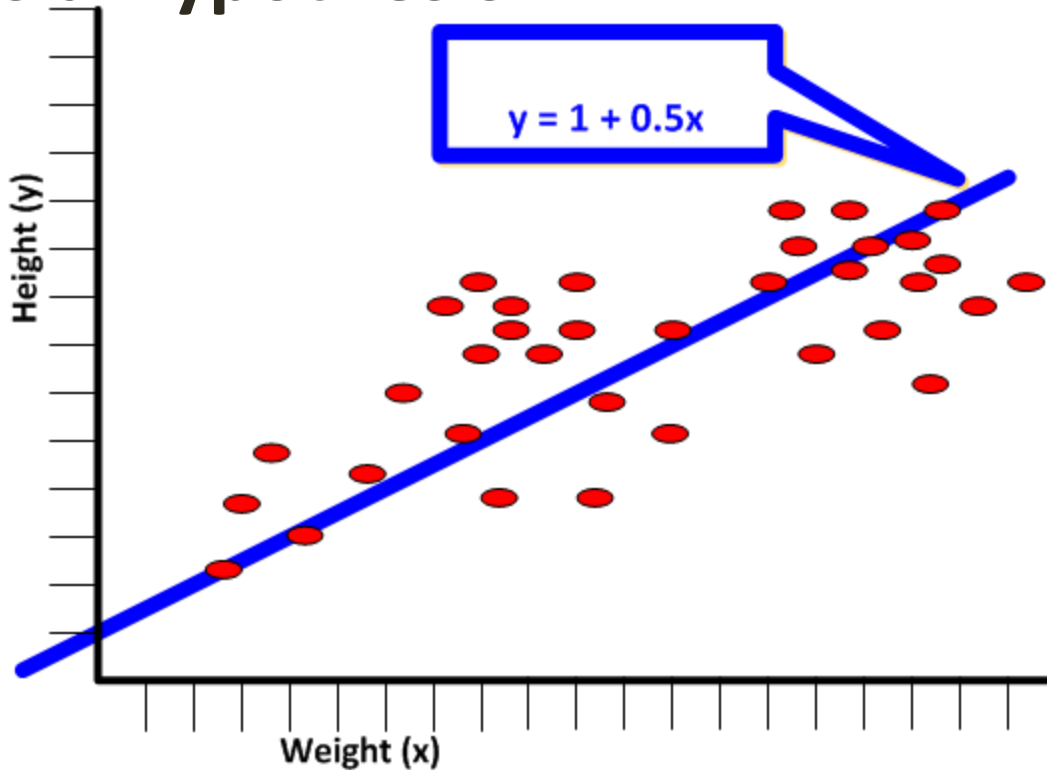
Model as a Hypothesis



The thick blue line represents the best fit through these data.

What is Data Science? (22)

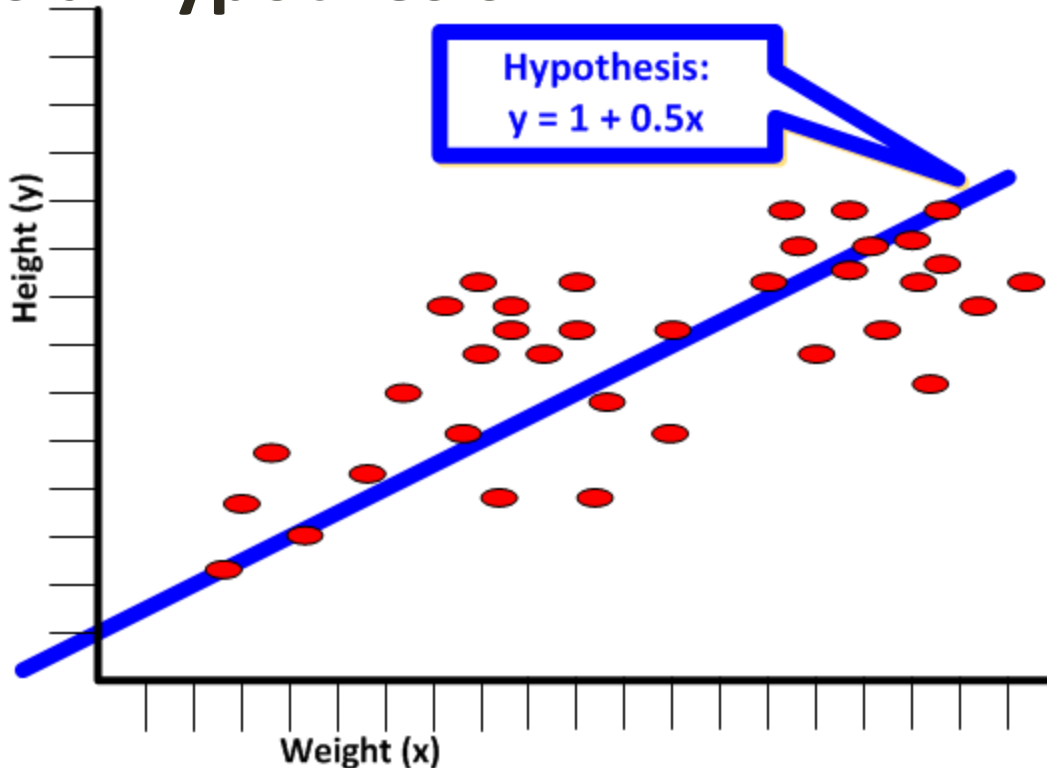
Model as a Hypothesis



The line can be represented algebraically.

What is Data Science? (23)

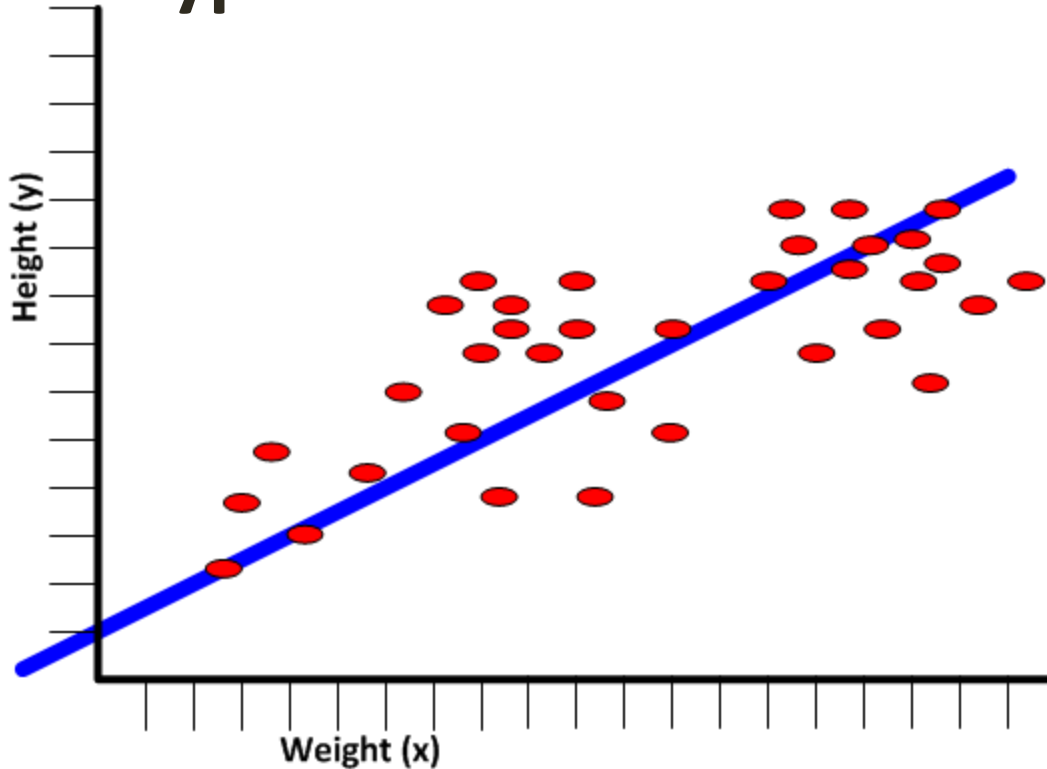
Model as a Hypothesis



The hypothesis can be represented algebraically. The hypothesis is the best fit line
 $y = 1 + 0.5x$

What is Data Science? (24)

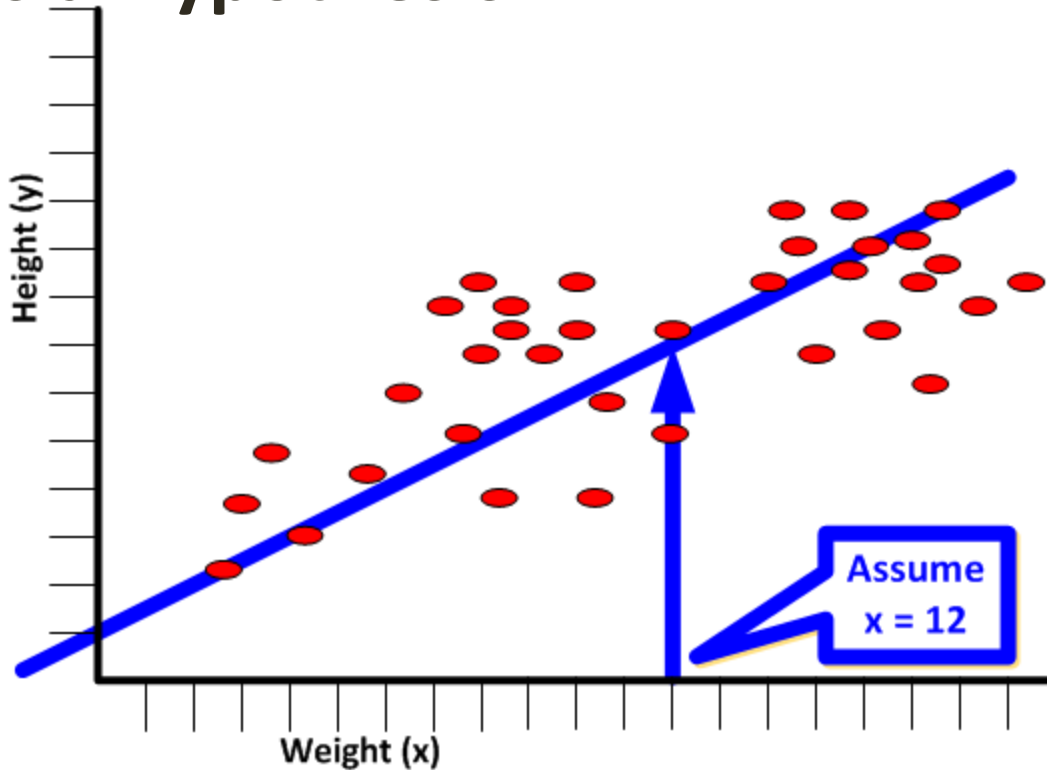
Model as a Hypothesis



The hypothesis can be used to predict

What is Data Science? (25)

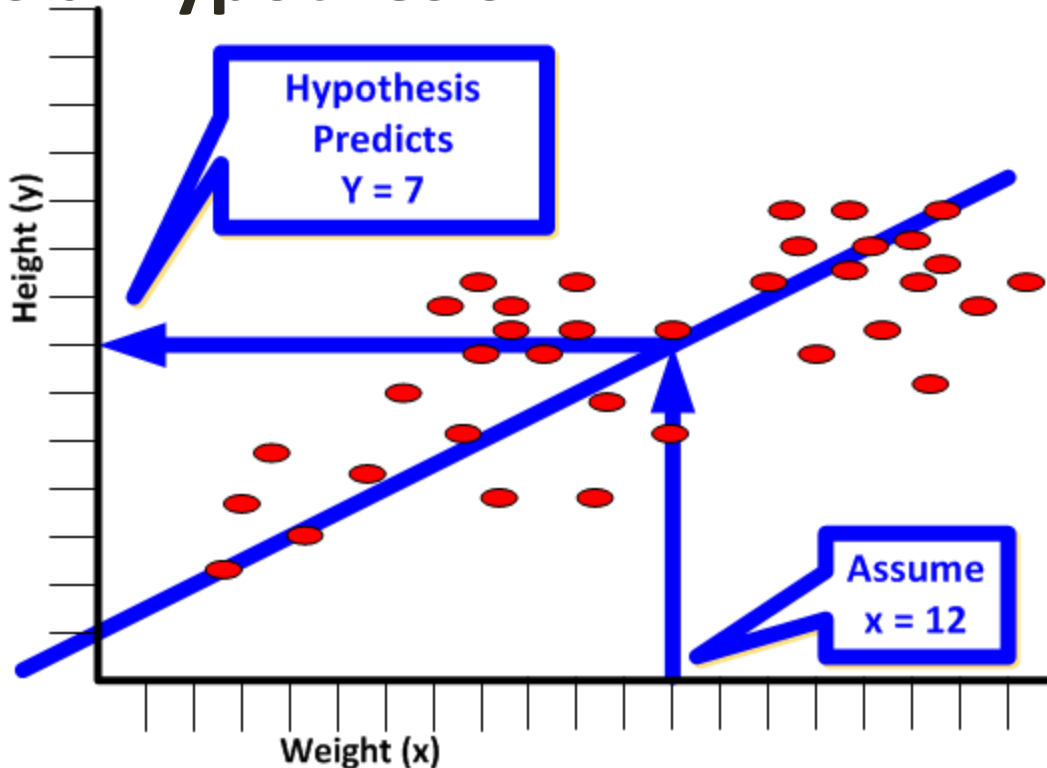
Model as a Hypothesis



The hypothesis can be used to predict. A prediction requires an input.

What is Data Science? (26)

Model as a Hypothesis



The prediction is an output.
Given an x value we can predict a
 y value. If $x = 12$, then $y = 7$.

What is Data Science? (27)

Some Terminology

- What is a model?
 - A model is a hypothesis based on data and a method (algorithm)
- What is a hypothesis? (<http://en.wikipedia.org/wiki/Hypothesis>)
 - Wikipedia: A proposed explanation for an observation that can be tested
 - A hypothesis is an explanation for the organization of a data set that allows a prediction
- Falsification (<http://en.wikipedia.org/wiki/Falsifiability>)
 - Falsification is the process that attempts to disprove a hypothesis
 - If a hypothesis is not falsifiable then it is not really a hypothesis
- What is a Theory?
 - A fact-based explanation for an observation (a well-tested hypothesis)
- What is a Law?
 - A prediction with no exceptions
 - A law does not attempt to explain the predictions as a theory would

What is Data Science? (28)

- Data are observations that are put into context
- Given that every observation has a context, an observation is a datum
 - Not Data:
 - 1 (one)
 - Chair
 - Diabetes
 - Data:
 - I see a chair
 - The patient has diabetes

What is Data Science? (29)

Unstructured Data

- Unstructured data does not exist.
 - The essence of data is that they are structured.
 - The context is what makes data.
- What is meant by unstructured data?
 - Answer: poorly structured data that are hard to analyze
 - Need to restructure the data through parsing, etc.
 - A list of tweets is often used as an example of unstructured data.
 - The tweets are organized into a list
 - Any single tweet comes from one source at one time
 - A tweet is a text with a length constraint.

What is Data Science? (30)

Data Structure leads to Data Types

- For example: a list of tweets
 - The data type is “a list of tweets”
 - A list inherits many characteristics of lists in general
 - A tweet inherits many characteristics from the data type text
- Typing makes Data
 - Typing is the context
- A list of tweets can be represented as a table
 - The column header provides context
 - The table structure states that all column values have comparable structures

What is Data Science? (31)

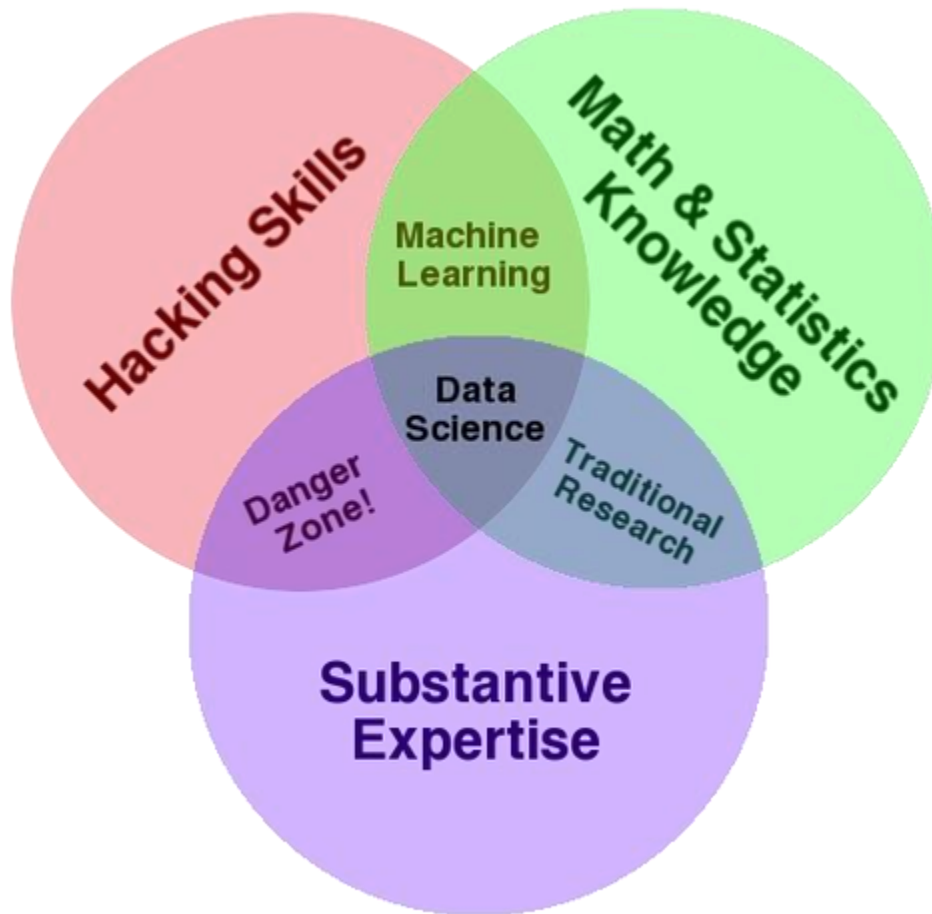
Data Structure leads to Data Types

- Physics
 - A data type is called a unit
 - Data are well-typed and can be universally converted
- Computer Science
 - Typing may demand structure: Strong Typing
 - [http://en.wikipedia.org/wiki/Strong typing](http://en.wikipedia.org/wiki/Strong_typing)
 - Structure and context determine typing: Weak typing
 - [http://en.wikipedia.org/wiki/Weak typing](http://en.wikipedia.org/wiki/Weak_typing)

What is Data Science? (32)

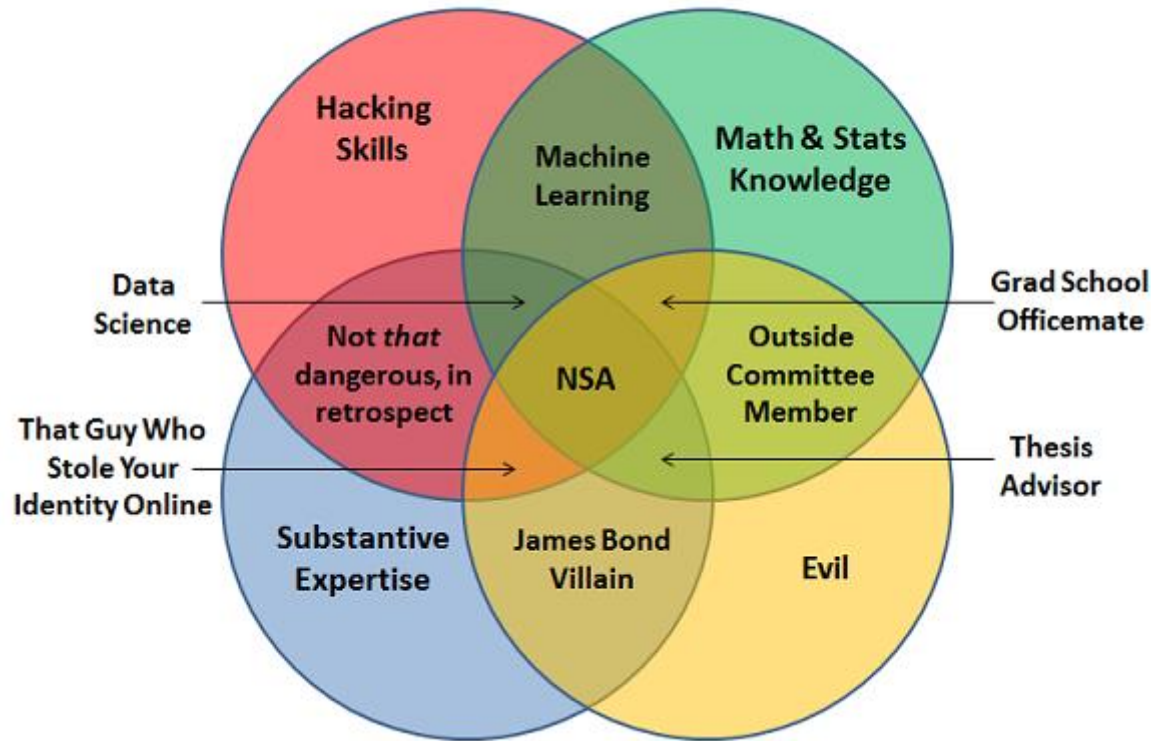
- What is Data Science?
- Data Science is made of two words: **Data** + **Science**
 - **Data** and their structures are well explained in computer science
 - The **Science** part of data science is explained by the scientific method.
 - The synthesis of these two disciplines allows
 - Data Visualization
 - Data Extraction
 - Data Processing / Transformation
 - Hypothesis Verification or Falsification

What is Data Science? (33)



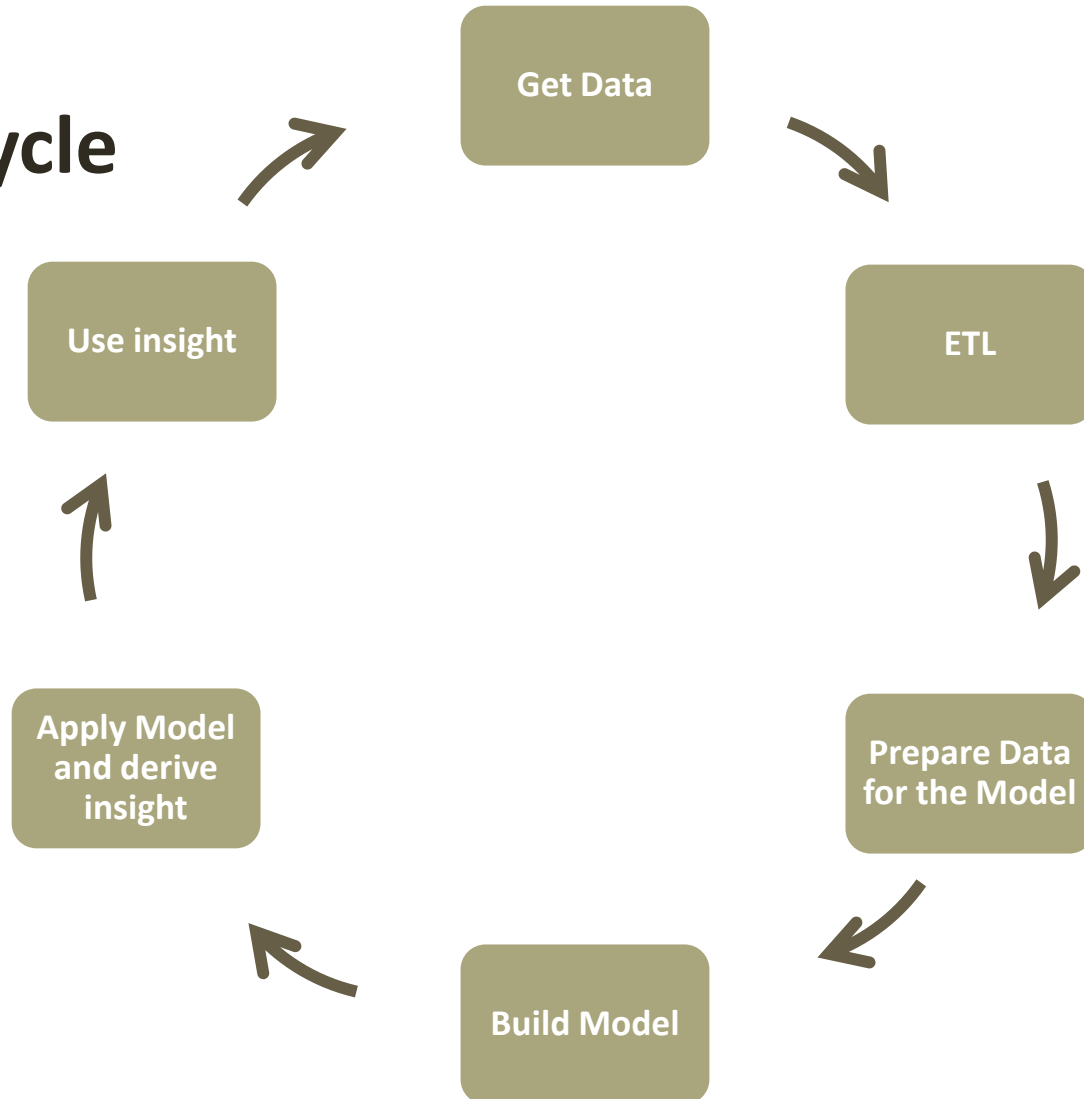
Drew Conway's Data Science Venn Diagram

What is Data Science? (34)



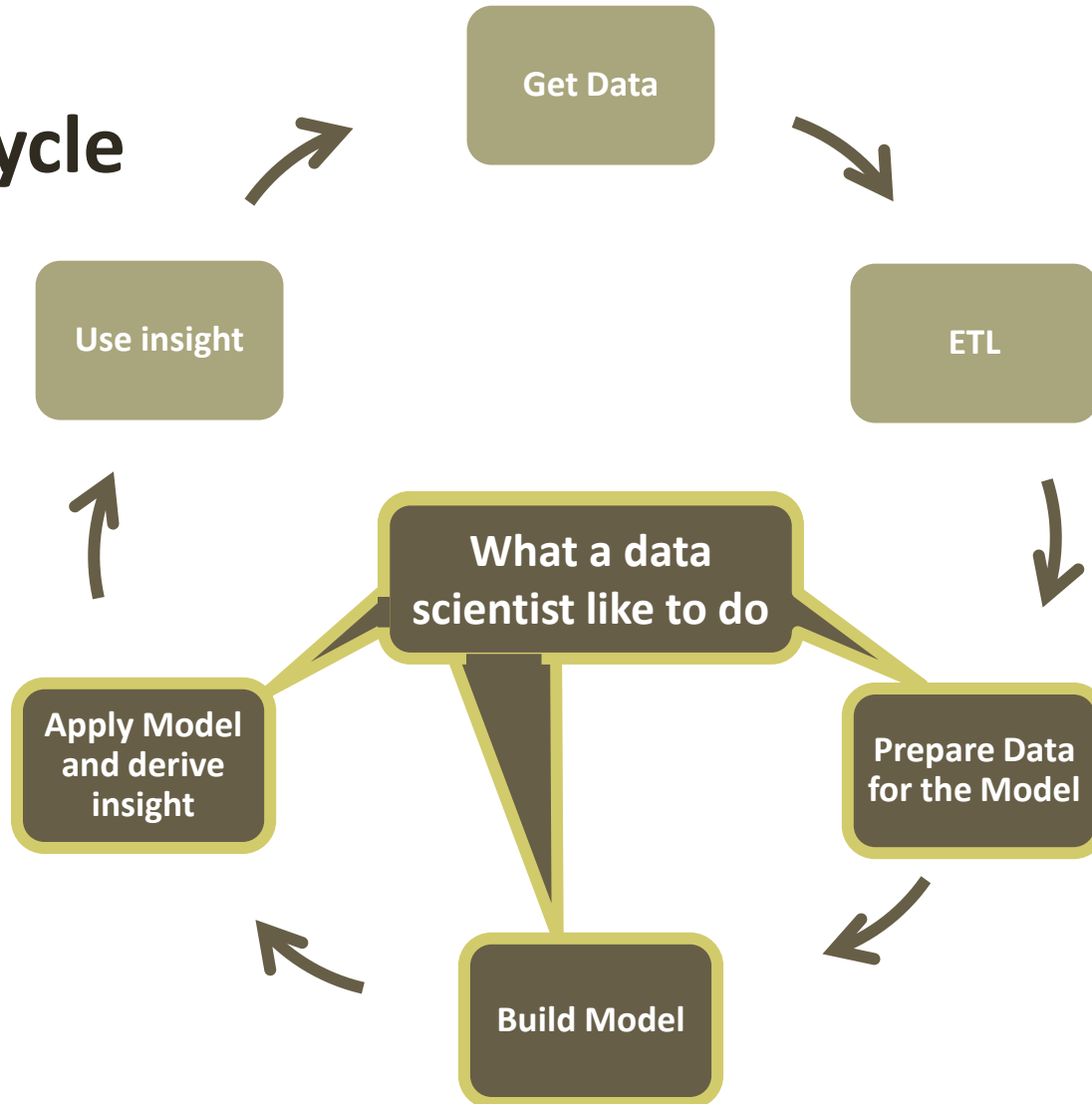
What is Data Science? (35)

BI Cycle



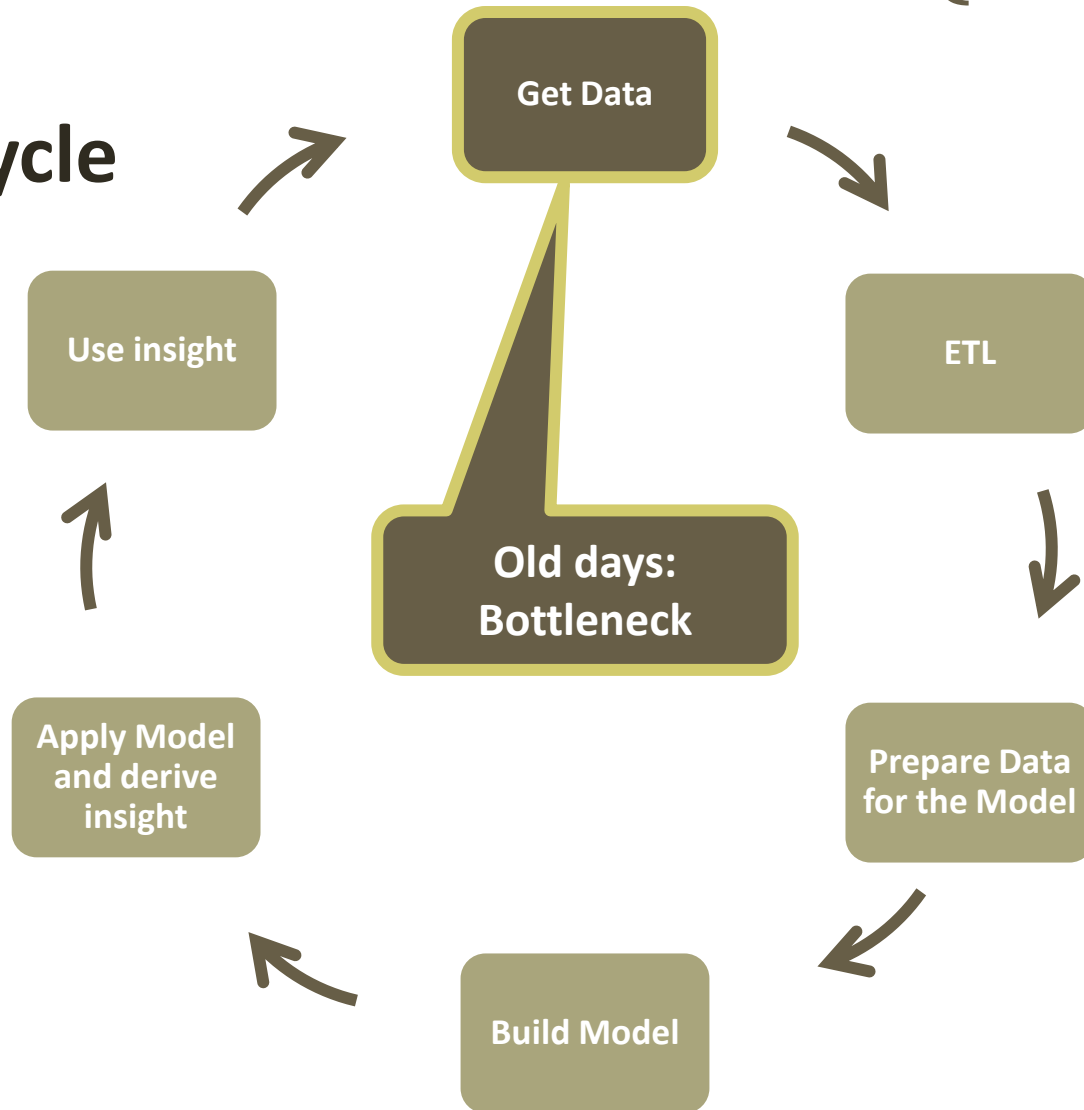
What is Data Science? (36)

BI Cycle



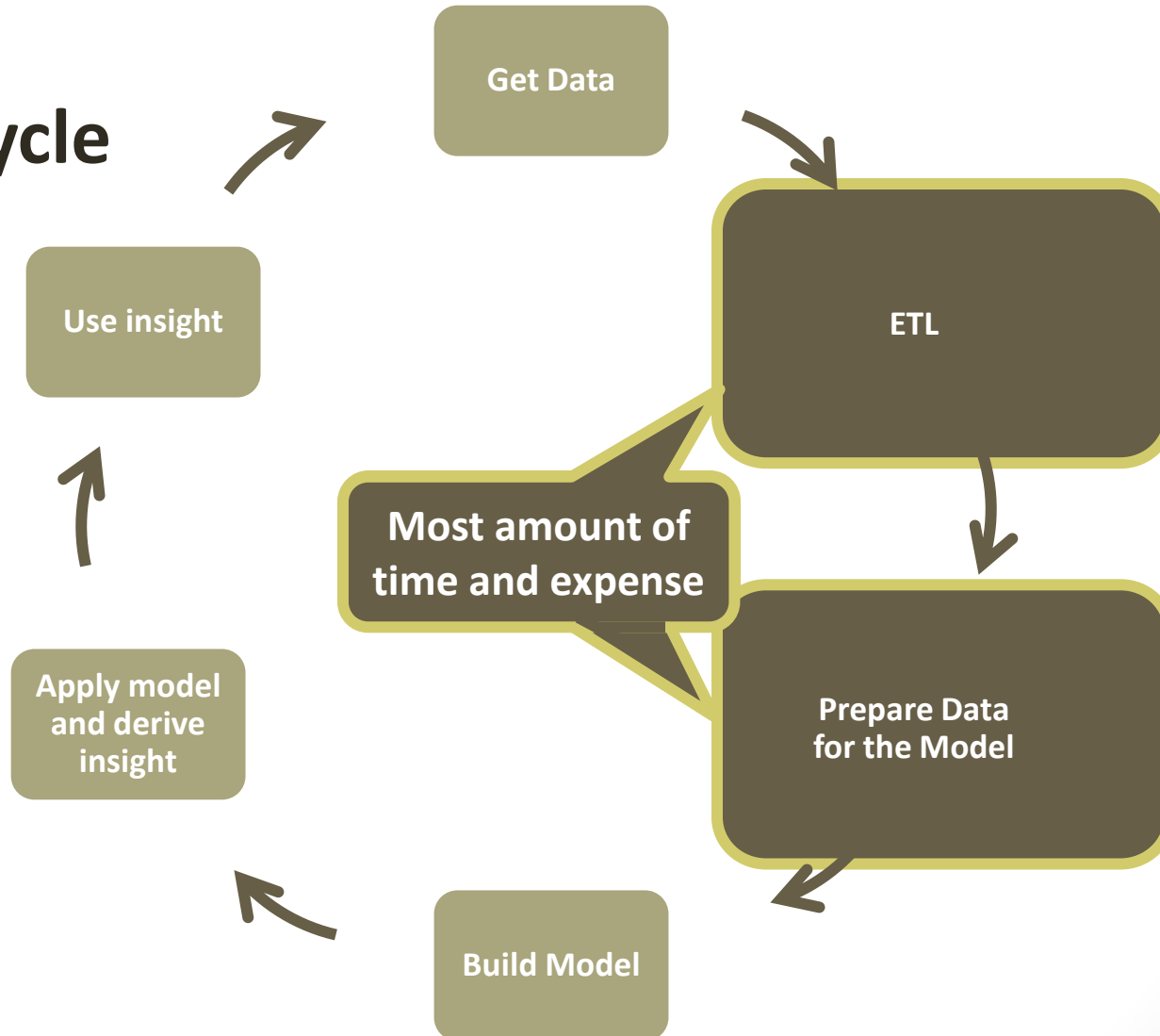
What is Data Science? (37)

BI Cycle



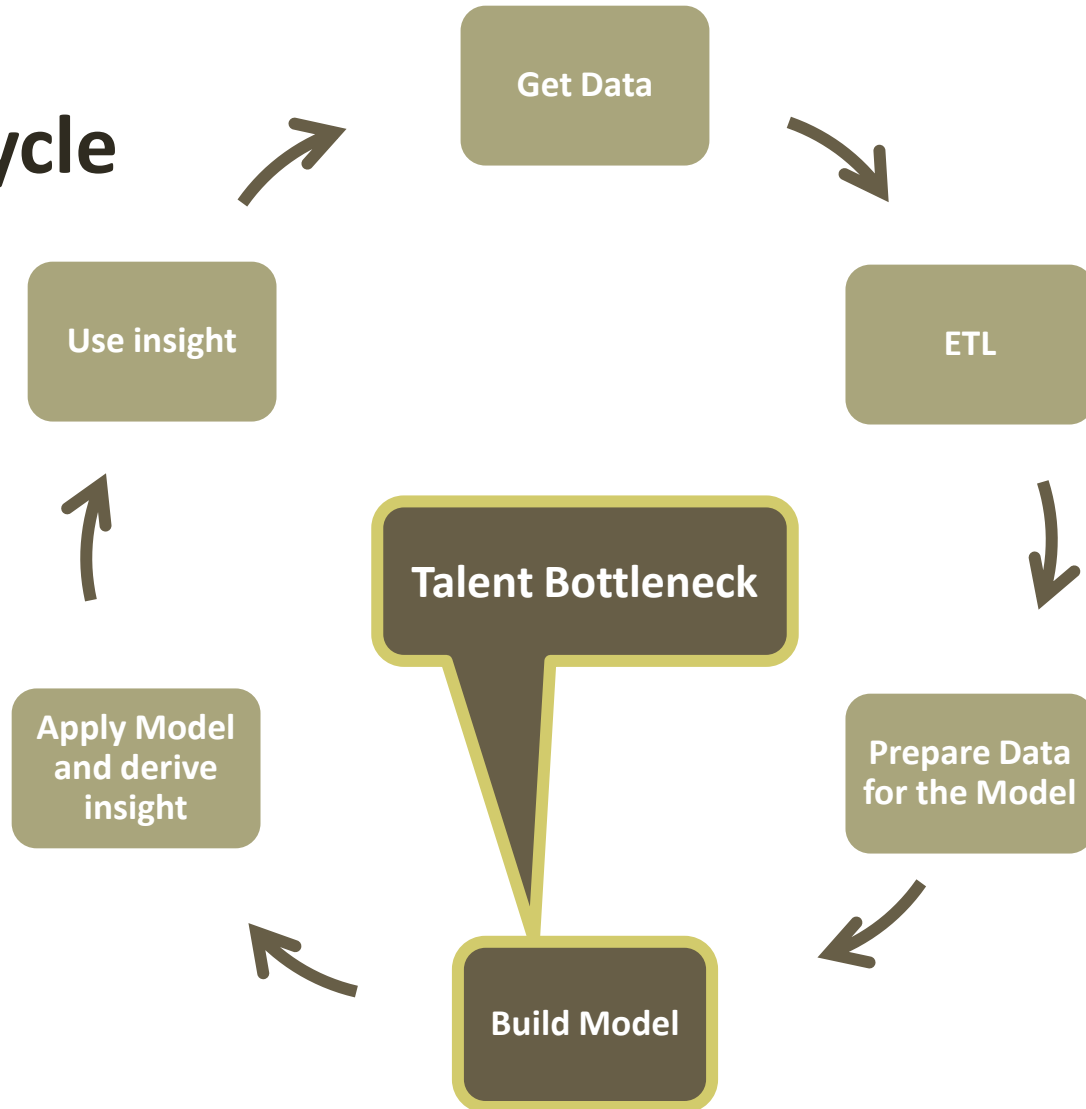
What is Data Science? (38)

BI Cycle



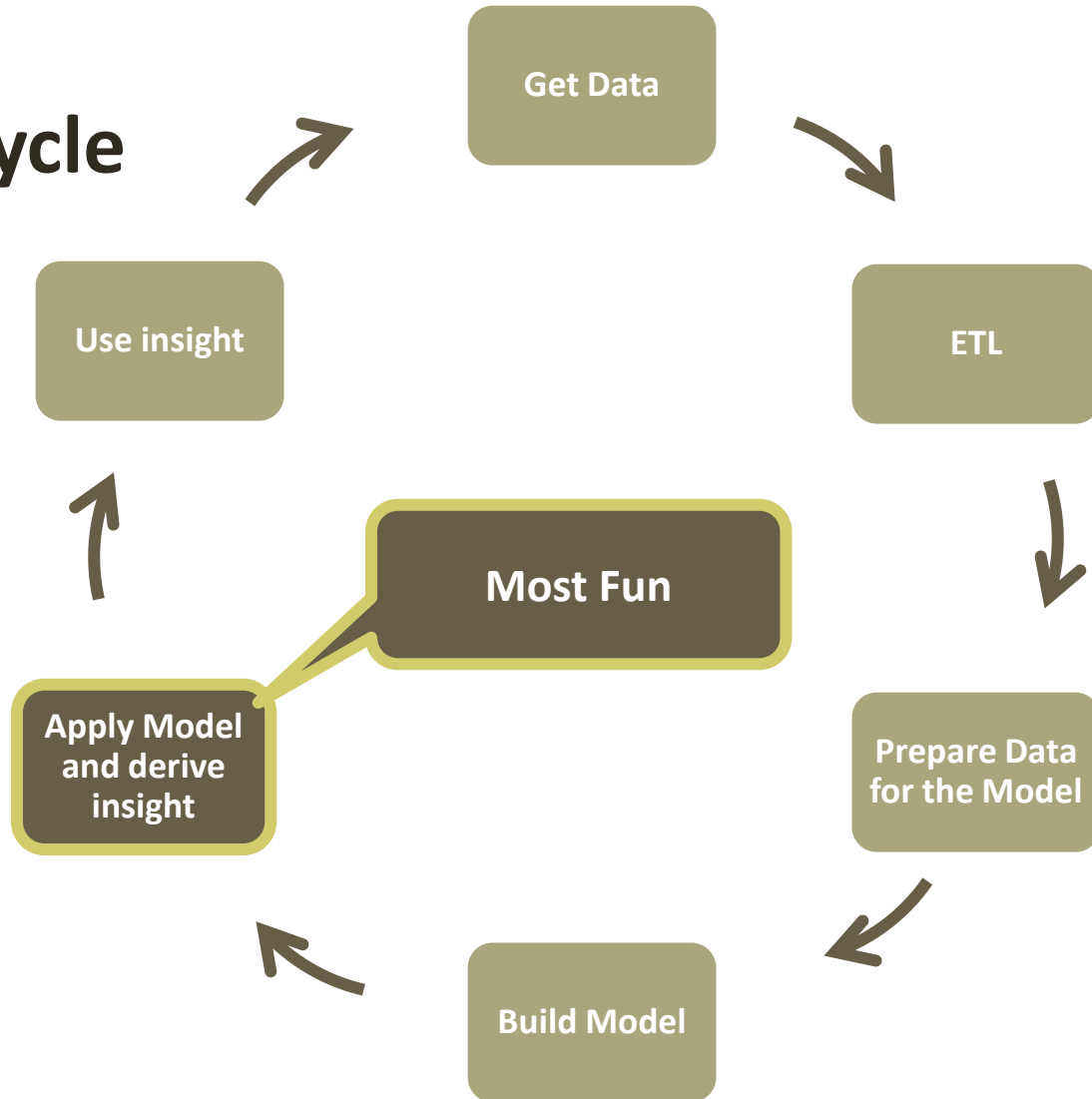
What is Data Science? (39)

BI Cycle



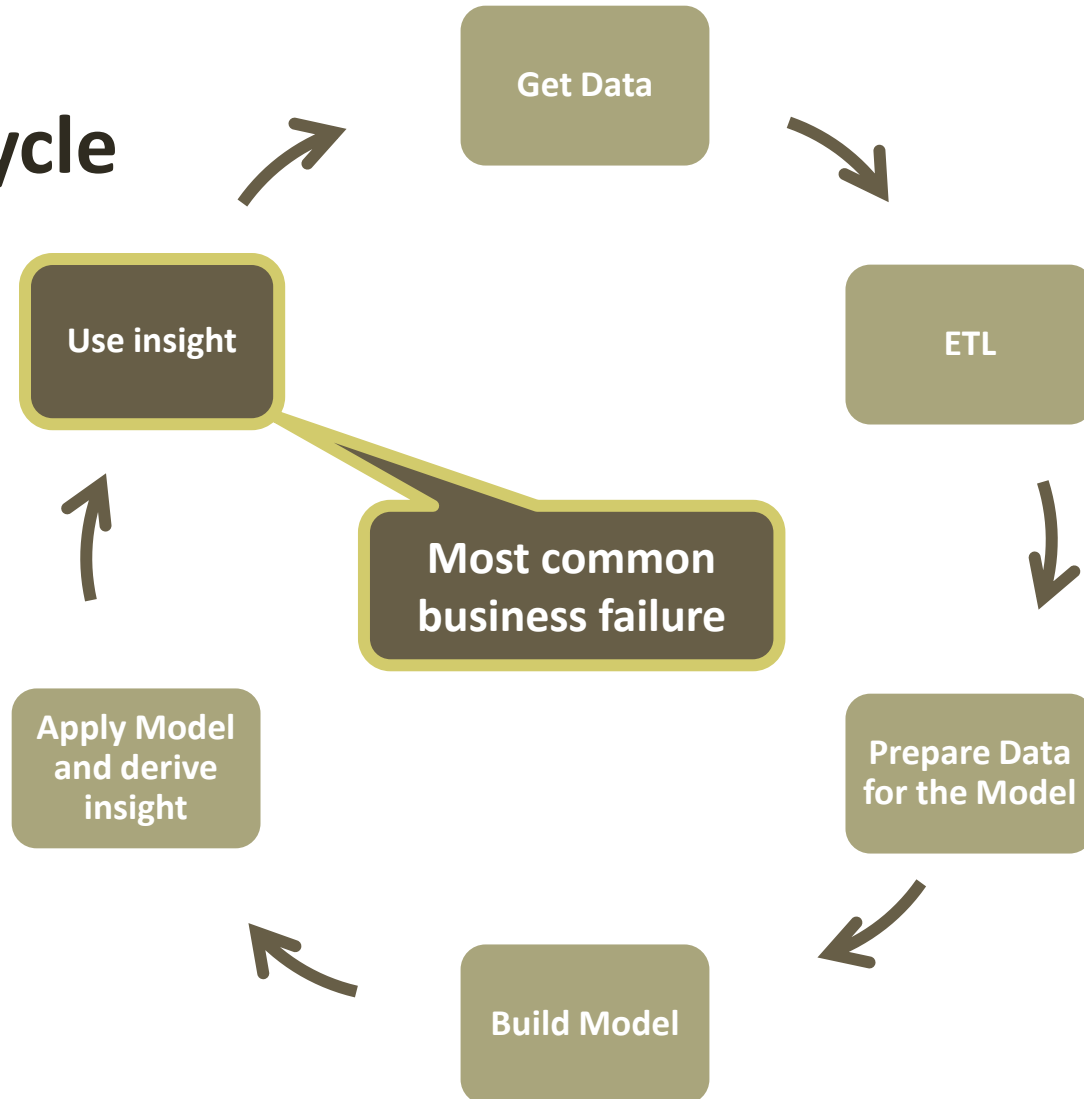
What is Data Science? (40)

BI Cycle



What is Data Science? (41)

BI Cycle



What is Data Science? (42)

Some data science tools and concepts in this course

- Some tools:
 - R (Maybe Python), XML
 - Database Engines: MySQL, Spark, Hadoop, RDF
 - Query languages and engines: SPARQL, SQL, HIVE, Impala
 - Data base accessories: Hue, sqldf, XML
 - Azure ML
- Some Concepts:
 - Data Types
 - Graph Analytics
 - Relational Algebra
 - Database structures
 - Predictive Modeling
 - NoSQL
 - CAP Theorem
 - MapReduce

What is Data Science? (43)

- Data Science
 - http://sqlblog.com/blogs/buck_woody/archive/2012/10/16/is-data-science-science.aspx
 - <http://radar.oreilly.com/2010/06/what-is-data-science.html>
 - <https://datajobs.com/what-is-data-science>
 - http://en.wikipedia.org/wiki/Data_science
- Data Mining (Mining of Massive Datasets)
 - <http://www.mmds.org/>
- Data Scientists in the Job Market
 - <http://www.kdnuggets.com/2015/03/salary-analytics-data-science-poll-well-compensated.html>
 - <http://www.hadoop360.com/blog/salaries-for-hadoop-professionals>
 - <http://www.analyticbridge.com/group/salary-trends-and-reports/forum/topics/salary-trends-for-data-science-professionals>
 - <http://www.analyticbridge.com/group/salary-trends-and-reports/forum/topics/the-10-highest-paying-jobs-for-math-geeks>

What is Data Science?

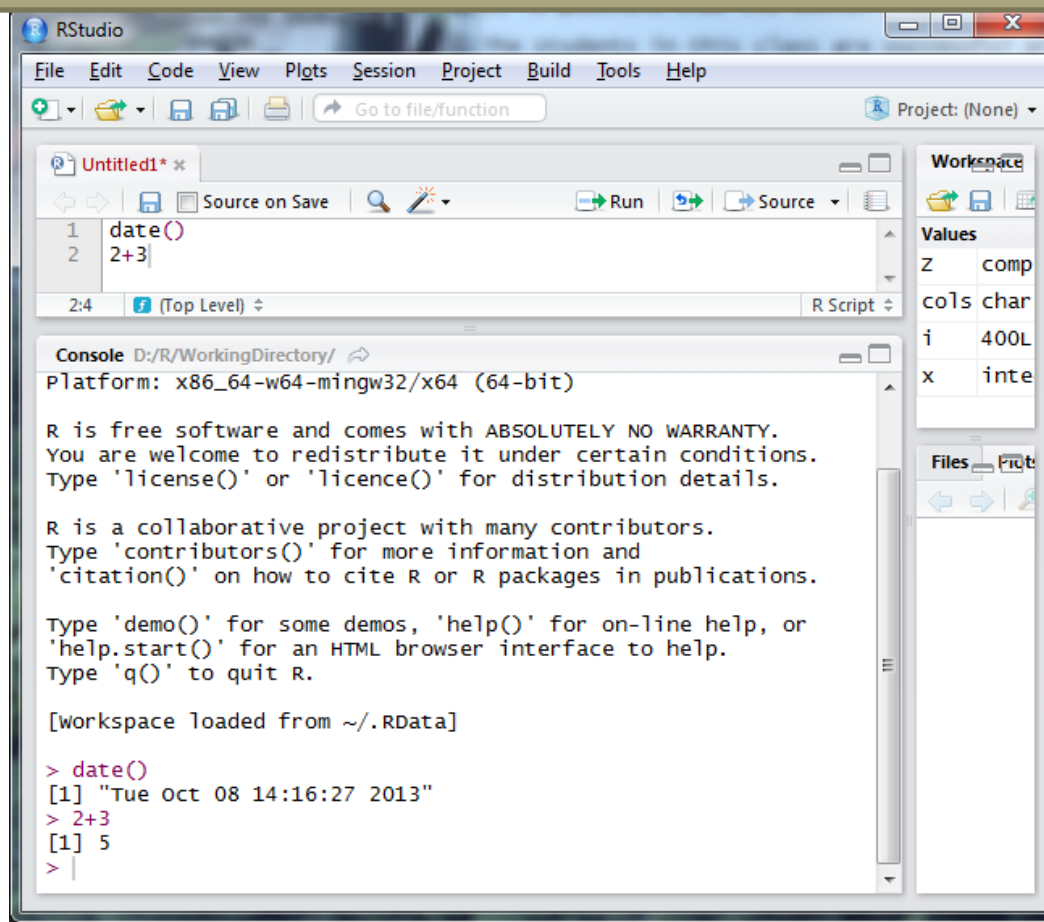
Quiz 01a (Class Structure)

- Go to the modules:
<https://canvas.uw.edu/courses/1092103/modules>
 - Expand Lesson 01
 - Click on Quiz 01a (Class Structure)
 - Take the quiz.
- If you cannot access canvas or the quiz, then please send me an email within the next 5 min. I will send the quiz to you later.
- Go to Canvas Module Lesson01. There is a submission site for "**Assignment00 (In-class Submission)**". Submit a text file named test.txt that contains your name.

R Basics

R Basics (1)

- <http://cran.r-project.org/bin/windows/base/>
- <http://www.rstudio.com/ide/download/>
- Get the latest version of R and the latest version of R studio. This screen shot is of an older version



R Basics (2)

- Go to Canvas
 - Go to the modules
 - Expand Lesson 01
 - Click on Lesson 01 Overview
 - Download DataScience01a.R and DataScience01b.R
- Open in R Studio: DataScience01a.R

R Basics (3)

- Some links for R
 - <http://cran.r-project.org/bin/windows/base/>
 - <http://cran.r-project.org/bin/macosx/>
 - <http://cran.r-project.org/bin/linux/>
 - <http://www.r-project.org/>
 - <http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf>
 - <http://www.rstudio.com/ide/download/desktop>
 - http://en.wikipedia.org/wiki/R_%28programming_language%29

R Basics

Quiz01b (Basic R)

- Find the quiz in the module for Lesson 01
- You should use R during the Quiz.

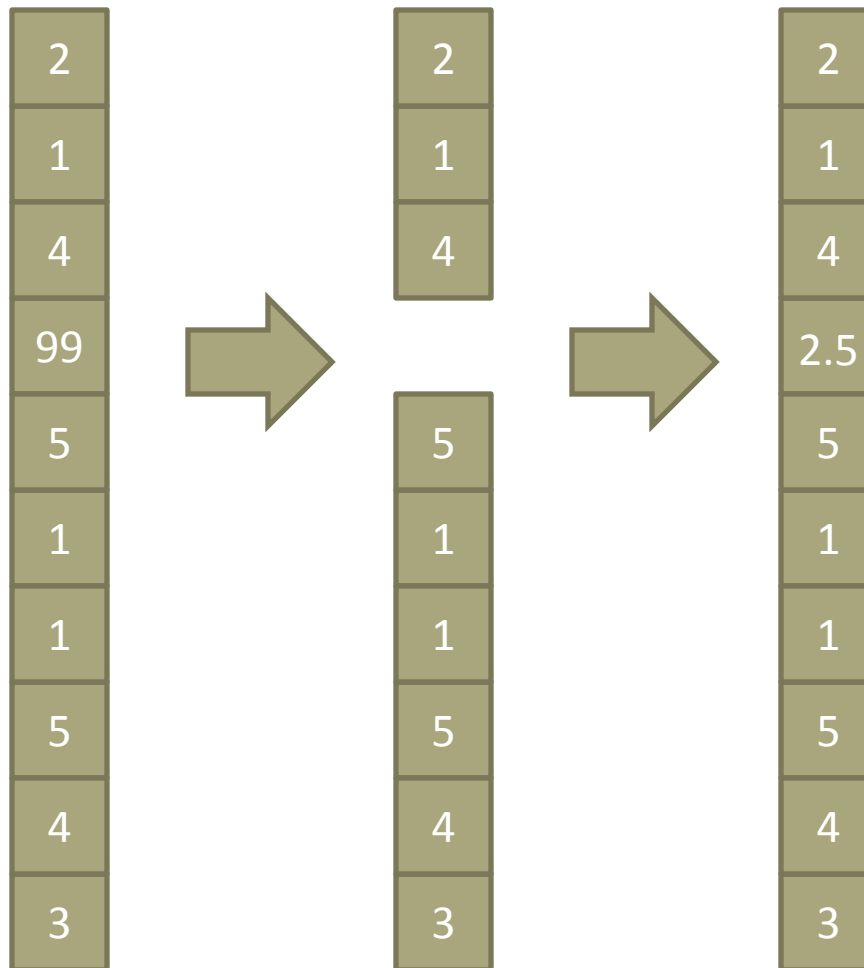
Data Preparation

Data Preparation (0)

- Schematize (Shape Data)
 - Create Tables
 - Relational Schemas
 - Star http://en.wikipedia.org/wiki/Star_schema
 - Snowflake http://en.wikipedia.org/wiki/Snowflake_schema
 - Flatten Relational Schemas
 - Specify Input vs. Target
 - Specify attributes that are neither Input nor Target
- Clean Data (Today's topic)
 - Later we will consider these techniques in DataScience01b.R

Data Preparation (1)

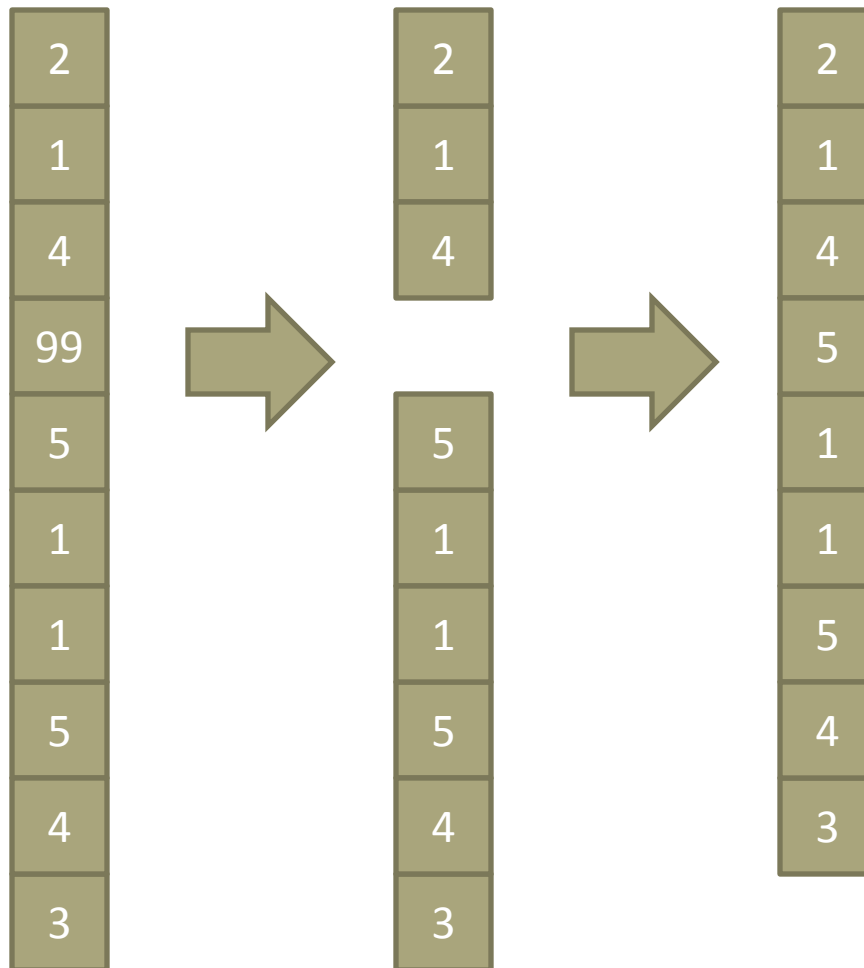
- Outlier Replacement



Data Preparation (2)

- Outlier Replacement
 - Numeric
 - Replace data beyond 3 standard deviations (1, 2, 2, 3, 3, 3, 4, 4, 5, 99)
 - `x[x > 10] <- median(x)`

Data Preparation (3)

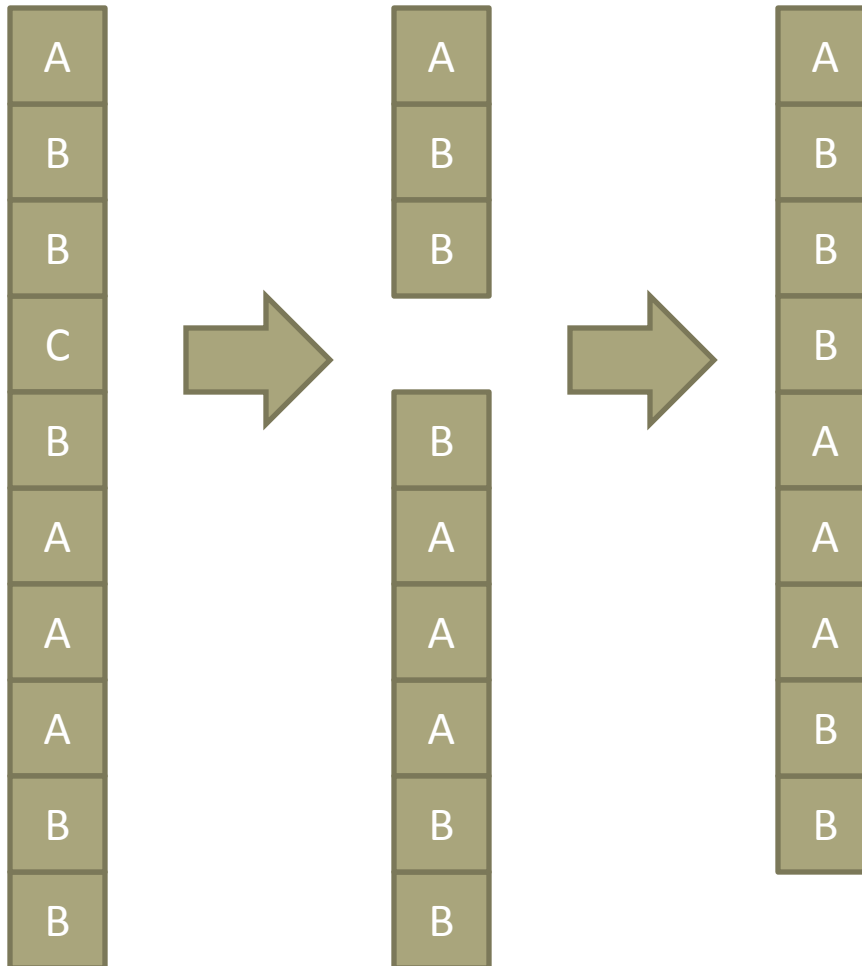


Data Preparation (4)

- Outlier Removal
 - Numeric
 - Remove data beyond 3 standard deviations (1, 2, 2, 3, 3, 3, 4, 4, 5, 99)
 - `x <- x[x < 10]`

Data Preparation (5)

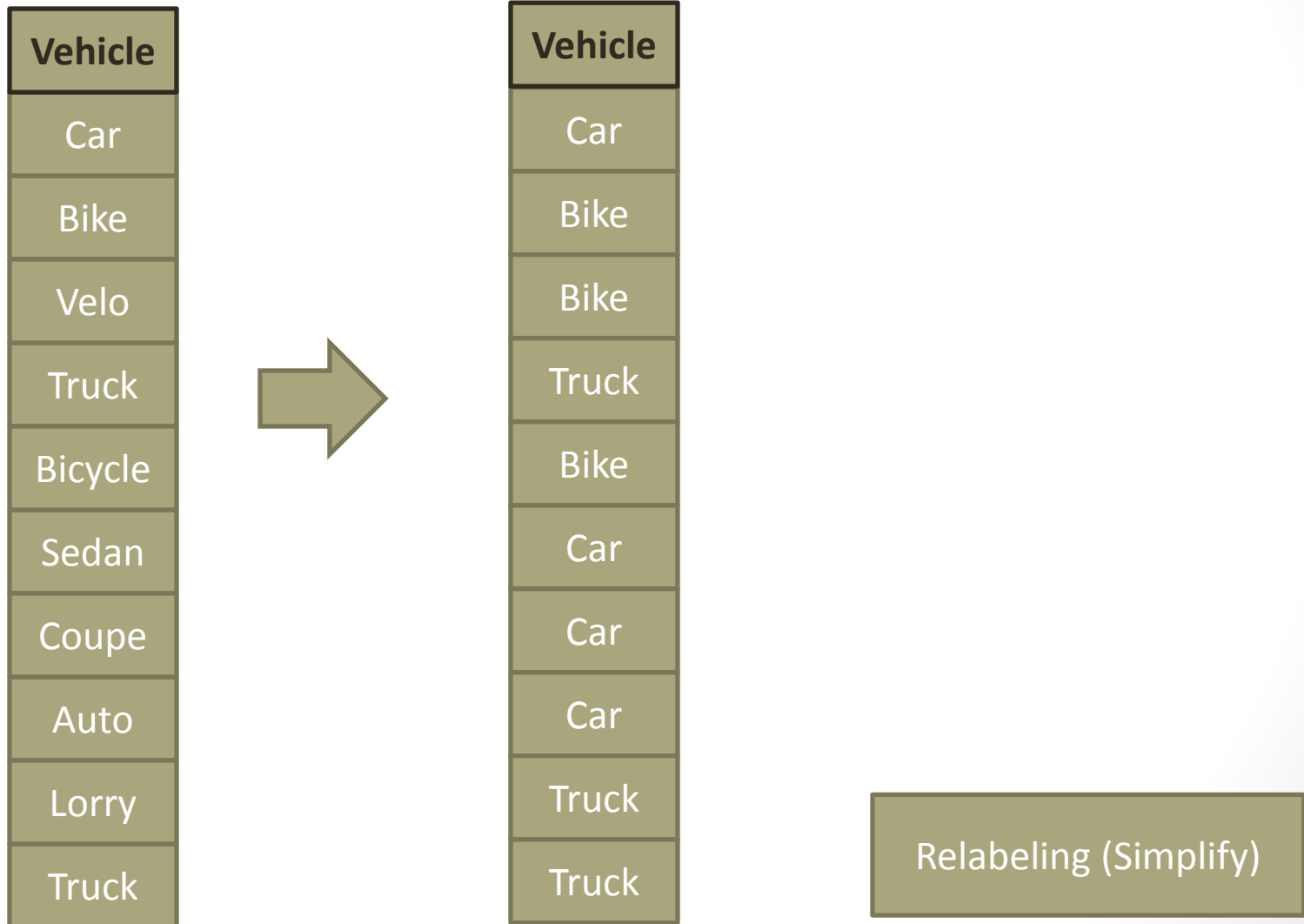
- Outlier Removal (Category)



Data Preparation (6)

- [illegible]

Data Preparation (7)



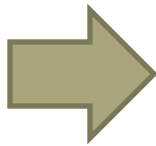
Data Preparation (8)

- Relabeling
 - Simplify (e.g. all 4 year degrees, like Bachelors, A.B. BSc, etc. as BS)
 - Example:
 - Vehicle: (Car, Automobile, Bike, Truck, Bicycle, Sedan, Coupe, Cycle, Truck, Velo, Automobile, Bike)
 - Car, Automobile, Sedan, Coupe -> Car
 - `x[x == "Automobile"] <- "Car"`
 - `x[x == "Sedan"] <- "Car"`
 - `x[x == "Coupe"] <- "Car"`
 - Bike, Bicycle, Cycle, Velo -> Bike
 - `x[x == "Bicycle"] <- "Bike"`
 - `x[x == "Cycle"] <- "Bike"`
 - `x[x == "Velo "] <- "Bike"`
 - Truck -> Truck
 - Vehicle: (Car, Car, Bike, Truck, Bike, Car, Car, Bike, Truck, Bike, Car, Bike)

Relabeling (Simplify)

Data Preparation (9)

Vehicle
1
2
2
3
2
1
1
1
1
3
3



Vehicle
Car
Bike
Bike
Truck
Bike
Car
Car
Car
Truck
Truck

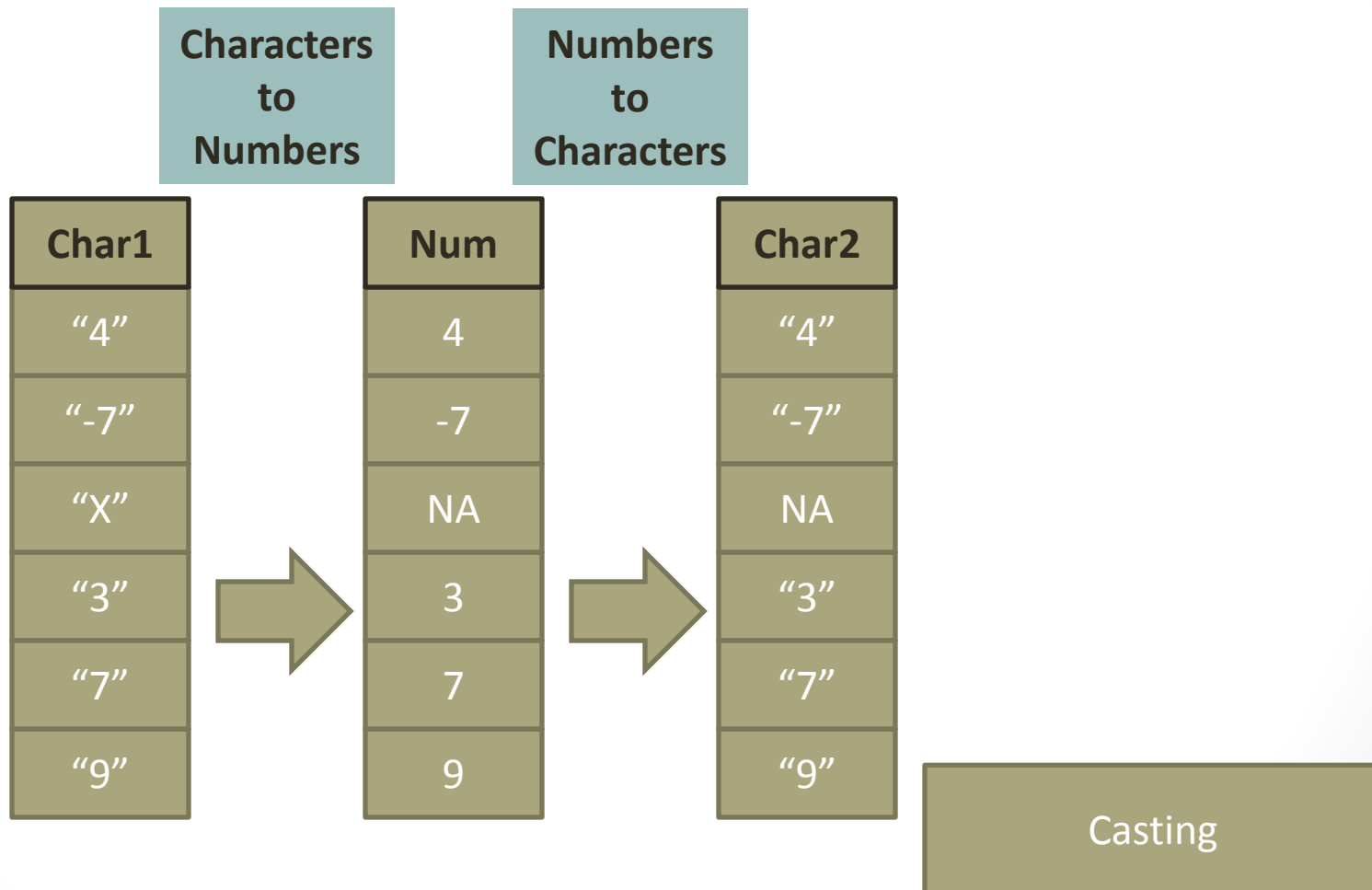
Code	Item
1	Car
2	Bike
3	Truck

Relabeling (Decode)

Data Preparation (10)

- Relabeling
 - De-code (numbers to categories)
 - Example1: Origin: (3, 1, 2, 1, 1, 2)
 - 1 is USA
 - 2 is Europe
 - 3 is Japan
 - Origin: (Japan, USA, Europe, USA, USA, Europe)
 - `x[x == 1] <- "USA"`
 - `x[x == 2] <- "Europe"`
 - `x[x == 3] <- "Japan"`
 - Example2: Origin: (3, 1, 2, 1, 1, 2)
 - (3, 1, 2, 1, 1, 2) -> ("3", "1", "2", "1", "1", "2")
 - `x <- as.character(3, 1, 2, 1, 1, 2)`

Data Preparation (11)



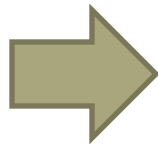
Data Preparation (12)

- Casting
 - Characters to Numbers: ("4", "-7", "X", "3") -> (4, -7, NA, 3)
 - `x <- as.numeric(x)`
 - Numbers to Characters : (4, -7, NA, 3) -> ("4", "-7", NA, "3")
 - `x <- as.character(x)`

Data Preparation (13)

Normalization

Orig		MM	Z
2		.3	-0.23
-1		0	-1.09
0		.1	-0.8
1		.2	-0.52
7		.8	1.2
9		1	1.78
7		.8	1.2
1		.2	-0.52
1		.2	-0.52
1		.2	-0.52



Data Preparation (14)

- Normalize
 - Normalize (Linear)
 - offset and multiplier $y = a + bx$ or $y = (x - c)/d$; Where: $a = -c/d$; $b = 1/d$
 - Min-Max where:
 - $c = \min$; $d = \max - \min$
 - $x \leftarrow (x - \min(x))/(\max(x) - \min(x))$
 - Z-score:
 - where $c = \text{mean}$; $d = \text{sigma}$
 - $x \leftarrow (x - \text{mean}(x))/\text{sd}(x)$
 - MAD (http://en.wikipedia.org/wiki/Median_absolute_deviation) where $c = \text{median}$; $d = \text{median of differences to median}$
 - Normalize (Non-Linear)
 - Log-normalization: $y = \text{Log}(x)$ or similar
 - Equalization

Data Preparation (15)

Discretization

Age		Range	Range		Cases	Cases
10		(0-33)	Low		(0-17)	Low
23		(0-33)	Low		(18-44)	Med
11		(0-33)	Low		(0-17)	Low
55		(34-66)	Med		(44-99)	High
60		(34-66)	Med		(44-99)	High
32		(0-33)	Low		(18-44)	Med
99		(67-99)	High		(44-99)	High
4		(0-33)	Low		(0-17)	Low
32		(0-33)	Low		(18-44)	Med
33		(0-33)	Low		(18-44)	Med
0		(0-33)	Low		(0-17)	Low

Data Preparation (16)

- Discretization
 - Age: (10, 23, 11, 55, 60, 32, 99, 4, 32, 33, 0) ->
 - Equal Range (0 – 33) (34 – 66) (67 – 99) -> (Low, Low, Low, Med, Med, Low, High, Low, Low, Low, Low)
 - $x[x > -\text{Inf} \ \& \ x < 34] <- \text{"Low"}$
 - $x[x \geq 34 \ \& \ x < 67] <- \text{"Med"}$
 - $x[x \geq 67 \ \& \ x < \text{Inf}] <- \text{"High"}$
- Equal Numbers (Equal Cases, Equal Area) (0 - 11) (23 - 33) (55 - 99) -> (Low, Med, Low, High, High, Med, High, Low, Med, Med, Low)

Data Preparation (17)

Dummy Variable

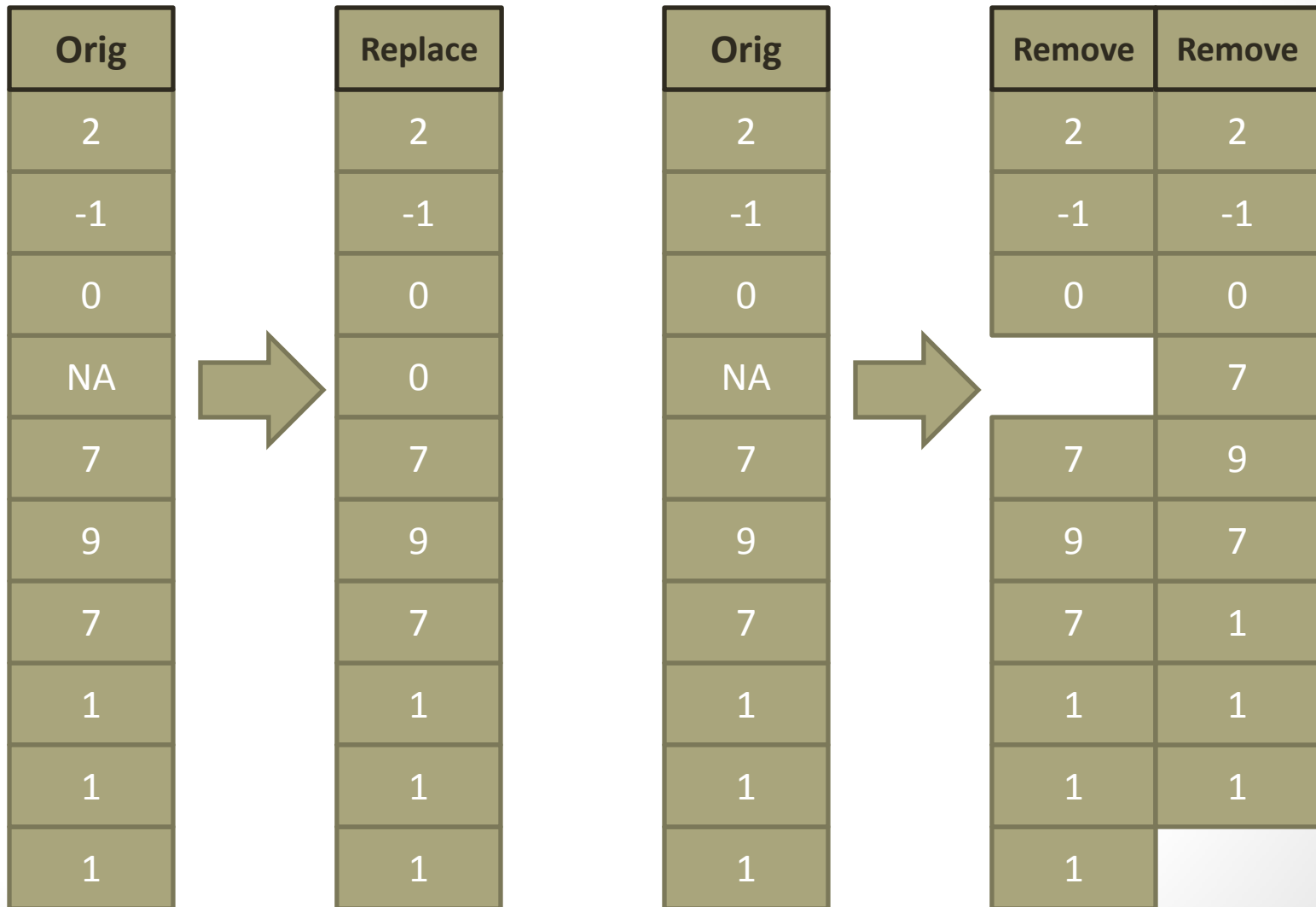
Vehicle		Car	Bike	Truck
Car		1	0	0
Bike		0	1	0
Bike		0	1	0
Truck		0	0	1
Bike		0	1	0
Car		1	0	0
Car		1	0	0
Car		1	0	0
Truck		0	0	1
Truck		0	0	1

Data Preparation (18)

- Categorical to Numerical (Binary)
- Binarize Categorical variables by creating multiple binary dummy variables
 - 1 column of Colors (Red, Green, Blue) -> three columns called isRed, isGreen, and isBlue
 - Color: `x <- Red, Green, Blue, Blue, Red, Red`
 - isRed:
 - Red, Green, Blue, Blue, Red, Red -> 1, 0, 0, 0, 1, 1
 - `r <- x == "Red"`
 - isGreen:
 - Red, Green, Blue, Blue, Red, Red -> 0, 1, 0, 0, 0, 0
 - `g <- x == "Green"`
 - isBlue:
 - Red, Green, Blue, Blue, Red, Red -> 0, 0, 1, 1, 0, 0
 - `b <- x == "Blue"`

Data Preparation (19)

Null Handling



Data Preparation (20)

- Null Handling
 - Value removal or Row Removal
 - $(4, -7, \text{NA}, 3) \rightarrow (4, -7, 3)$
 - `x <- x[!is.na(x)]`
 - Value substitution:
 - $(4, -7, \text{NA}, 3) \rightarrow (4, -7, 0, 3)$
 - `x[is.na(x)] <- 0`

Data Preparation

Break



Data preparation

Mise en place

Data Preparation in R

- Open in R Studio: DataScience01b.R

Assignment (0)

- All assignment items from all assignment slides are due by Saturday 11:57.

Assignment (1)

1. Download and Install R. Then download and install R studio. Calculate $2 + 3$ in R studio . Type your name into the console. Take a screenshot of R-studio (not just the console) and name the screenshot file: RStudio.jpg or RStudio.png or RStudio.pdf. The format should be jpg, png, or pdf.
2. Join the LinkedIn group for this course. Introduce yourself, start a discussion, or make a comment on an existing discussion. Write the topic of that discussion in a txt file called discussion.txt
3. Follow the patterns described in DataScience01a.R and use R to get the Indian Liver Patient Dataset from the UCI machine learning repository.
 - `url <- http://archive.ics.uci.edu/ml/machine-learning-databases/00225/Indian%20Liver%20Patient%20Dataset%20\(ILPD\).csv` # Copy this url carefully
 - `ILPD <- read.csv(url, header=FALSE, stringsAsFactors=FALSE)`

Assignment (2)

4. The following was not covered in class. Get the 11 column headers from this page:
[http://archive.ics.uci.edu/ml/datasets/ILPD+\(Indian+Liver+Patient+Dataset\)#](http://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)#)
 - Manually construct a vector of column headers using this pattern:
 - `a.headers <- c(<name1>, <name2>, ...)` # Each column has a name
 - Associate names with the dataframe using this pattern:
 - `a.names(<dataframe>) <- headers`
5. Use `head(ILPD)` to view the first 6 rows.
6. Follow the patterns described in DataScience01a.R. Write code to determine the mean, median, and standard deviation (sd) of each column and present their values in the console. Some calculations may fail. Where applicable, fix the failures by using “`na.rm = TRUE`”. You can see how to use “`na.rm = TRUE`” by typing `?median` into the console.
7. Follow the patterns described in DataScience01a.R Create Histograms (`hist`) for each column where possible.
8. Follow the patterns described in DataScience01a.R Use the `plot(ILPD)` function on this data frame to present a general overview of the data. You want to see a matrix of many plots. You may have some problems because the Gender column is not numeric. You can skip the Gender column, or you can turn the gender column into a numeric column.

Assignment (3)

9. Look at the plots from `plot(ILPD)` and answer:
 - How can you tell if a vector contains continuous numbers or binary data?
 - Which two vectors are most strongly correlated?
 - Give an example of two vectors that have little correlation
10. Follow the patterns described in `DataScience01b.R`. Write code to remove outliers from the following vector and present the result in the console: `c(1, -1, -1, 1, 1, 17, -3, 1, 1, 3)`
11. Follow the patterns described in `DataScience01b.R`. Write code to relabel the following vector. Use the shortest strings for each category in the relabeled version. Present the result in the console: `c('BS', 'MS', 'PhD', 'HS', 'Bachelors', 'Masters', 'High School', 'MS', 'BS', 'MS')`
12. Follow the patterns described in `DataScience01b.R`. Write code to normalize the following vector using a Min-Max normalization and present the result in the console. Do not remove outliers. `c(1, -1, -1, 1, 1, 17, -3, 1, 1, 3)`
13. Follow the patterns described in `DataScience01b.R`. Write code to normalize the following vector using a Z-score normalization and present the result in the console. Do not remove outliers. `c(1, -1, -1, 1, 1, 17, -3, 1, 1, 3)`
14. Follow the patterns described in `DataScience01b.R`. Write code to binarize: `c('Red', 'Green', 'Blue', 'Green', 'Blue', 'Blue', 'Red', 'Blue', 'Green', 'Blue')` and present the result in the console

Assignment (4)

15. Follow the patterns described in DataScience01b.R Write code to discretize the following vector into 3 bins of equal range and present the result in the console. Do not remove outliers. `c(81, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 12, 24, 24, 25)`
16. The following is a vector of ages of people in a kindergarten class. Included are some older siblings, teachers, and somebody's grandfather. Discretize this vector into 3 bins of equal or near equal number of people. No Code is necessary, just present the results as commented text in the R file. `c(81, 3, 3, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 12, 24, 24, 25)`
17. Submit to canvas the screenshot from item 1, the txt file from item 2, and an R script that contains the answers to items 3 through 16. Submit by Saturday 11:57 PM to the Homework Submission site on Canvas in the Module called "Lesson 01". The Assignment is called "Assignment 01". If you cannot submit the assignment on time, please notify me before the deadline at ErnstHe@UW.edu and put "Data Science UW 2016 Assignment 01 late" (without quotes) in the email subject line
18. Reading assignment
 - http://en.wikipedia.org/wiki/Cluster_analysis
 - http://en.wikipedia.org/wiki/K-means_clustering
 - http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/
 - <http://www.sqlserverdatamining.com/ArtOfClustering/default.aspx>
19. Look through Preview section of Lesson 01 Overview

Introduction to Data Science