Introduction to Data Science

Data Science Essentials

Meet your instructors!

Michael Holloway is a mathematician turned data scientist. Michael holds a PhD in Mathematics from the University of Tennessee, Knoxville and a Masters from Tennessee Technological University. Michael loves the challenge and excitement of keeping up with the ever-evolving landscape of data science. Prior to joining Nashville Software School in April 2019, Michael served as an Assistant Professor in the Department of Mathematics at Saint Augustine's University.

Mary van Valkenburg is a data enthusiast with experience in data wrangling and exploration, visualization, and machine learning. If you asked her what she loves more: data or learning, she would have a hard time answering. Mary has a BS in Psychology, an MS in Data Science, and 18 years of experience in software product development and consulting. She places high value on curiosity, empathy, and honesty.

NASHVILLE SOFTWARE SCHOOL

- Your name
- The place you call home
- Something people are usually surprised to discover about you

Classroom rules

- Ask lots of questions
- Help each other; learn from each other
- Coding tasks are a guide
  - You **don’t** have to get them **all** done
  - You **can** form your own ideas and do your own exploration beyond what has been suggested

NASHVILLE SOFTWARE SCHOOL

Class format

- Review of previous coding tasks
- Concepts/Code Lecture
- Coding tasks
- Interactive with instruction team!

NASHVILLE SOFTWARE SCHOOL

Goals for the class

- Get hands-on experience of what it might be like to work as a data scientist
- Get an idea of whether or not this might be a good fit for a career
- Make discoveries and have fun

NASHVILLE SOFTWARE SCHOOL

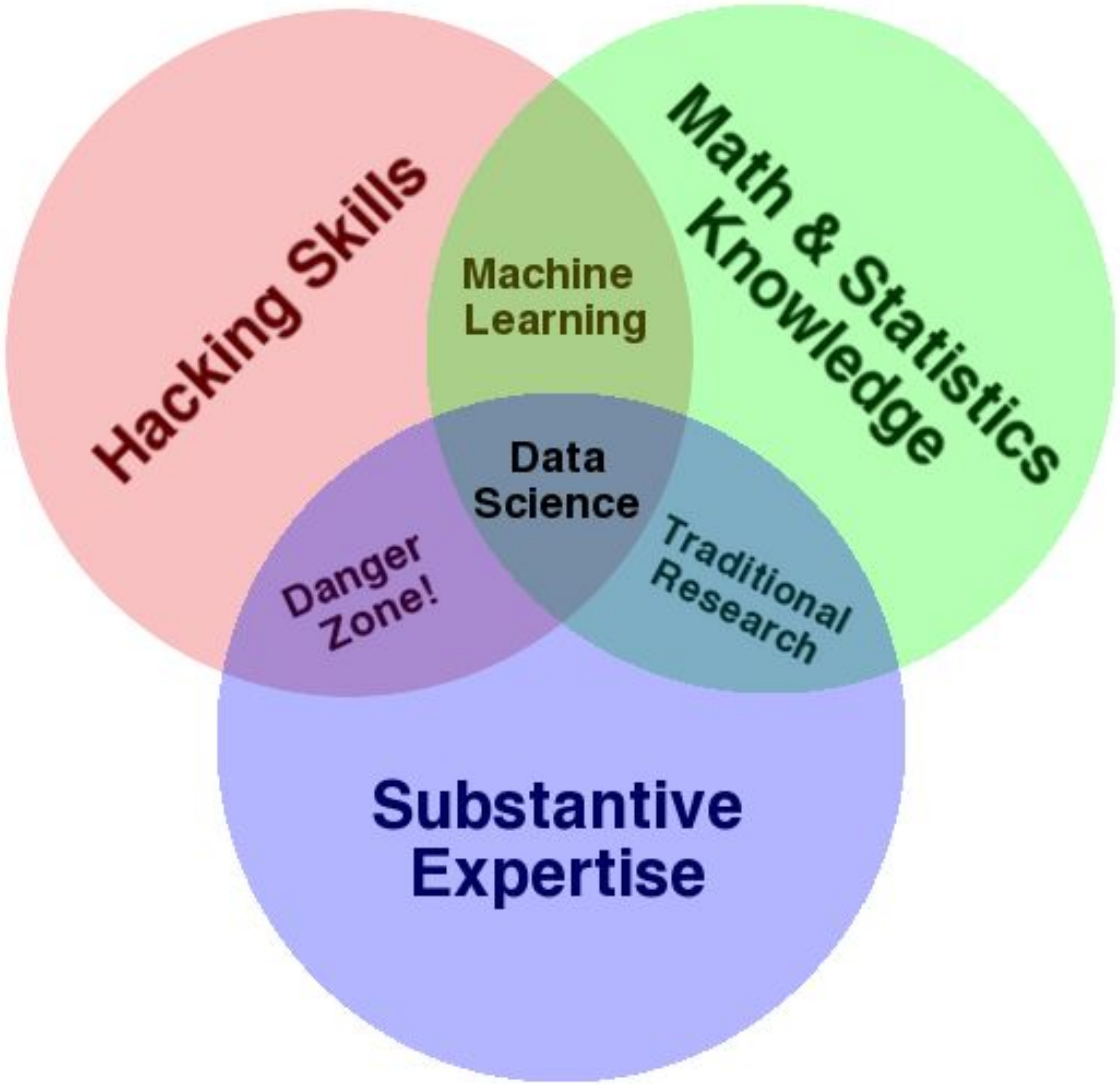
Goals for today

- Define Data Science and the Data Science Process
- DM Mary your github account name on Slack
- Learn a little pandas
- Understand the question
- Work in teams to complete coding tasks

NASHVILLE SOFTWARE SCHOOL

What is Data Science?

March 26, 2013



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

NASHVILLE SOFTWARE SCHOOL

January 9, 2018

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

VARIANCE EXPLAINED

David Robinson

Chief Data Scientist at DataCamp, works in R and Python.

Data science produces insights

Data science is distinguished from the other two fields because its goal is an especially human one: to gain insight and understanding. Jeff Leek has an excellent definition of the types of insights that data science can achieve, including descriptive ("the average client has a 70% chance of renewing") exploratory ("different salespeople have different rates of renewal") and causal ("a randomized experiment shows that customers assigned to Alice are more likely to renew than those assigned to Bob").

Again, not everything that produces insights qualifies as data science (the classic definition of data science is that it involves a combination of statistics, software engineering, and domain expertise). But we can use this definition to distinguish it from ML and AI. The main distinction is that in data science there's always a human in the loop: someone is understanding the insight, seeing the figure, or benefitting from the conclusion. It would make no sense to say "Our chess-playing algorithm uses data science to choose its next move," or "Google Maps uses data science to recommend driving directions".

This definition of data science thus emphasizes:

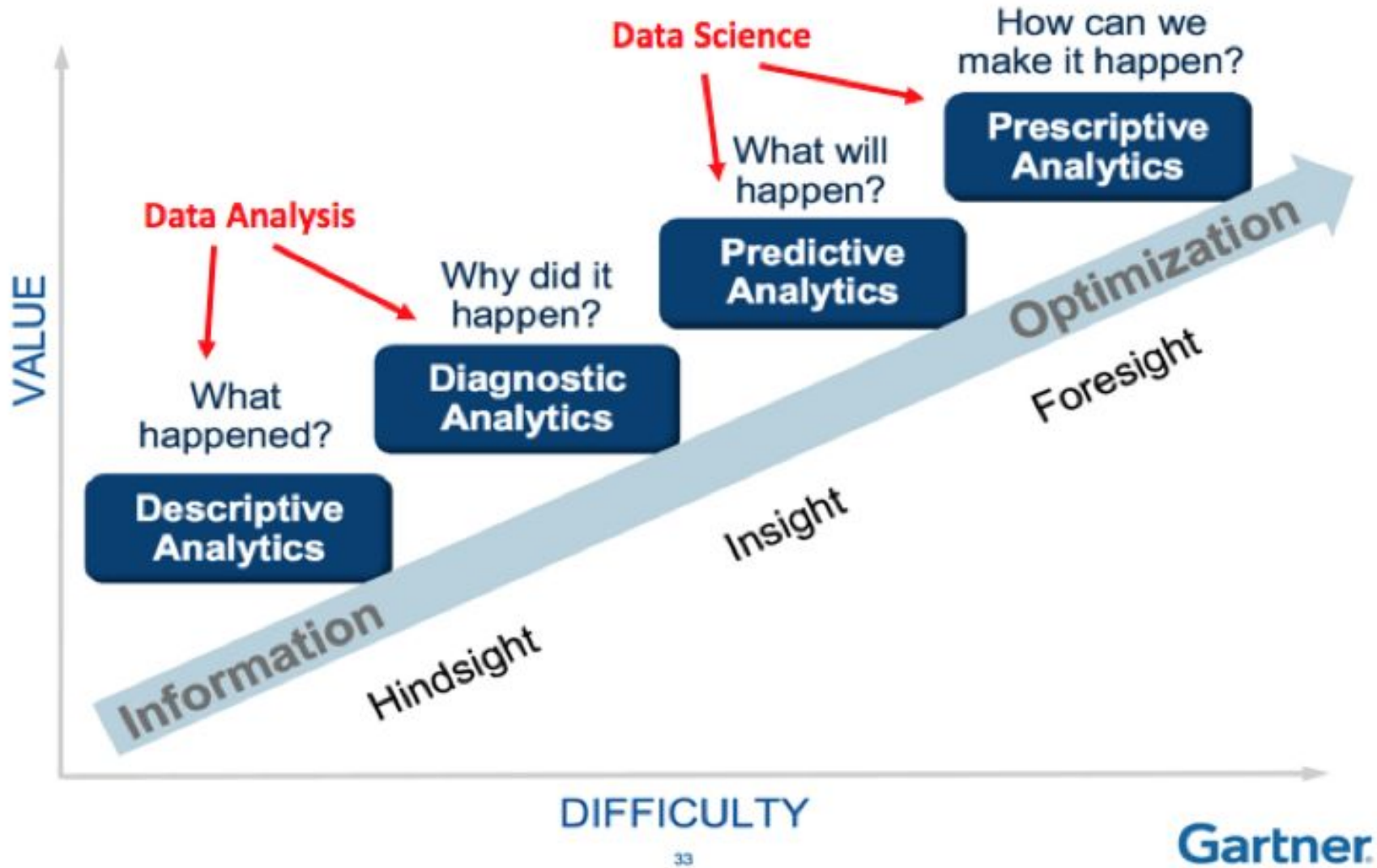
- Statistical inference
- Data visualization
- Experiment design
- Domain knowledge
- Communication

Data scientists might use simple tools: they could report percentages and make line graphs based on SQL queries. They could also use very complex methods: they might work with distributed data stores to analyze trillions of records, develop cutting-edge statistical techniques, and build interactive visualizations. Whatever they use, the goal is to gain a better understanding of their data.

http://varianceexplained.org/r/ds-ml-ai/

NASHVILLE SOFTWARE SCHOOL

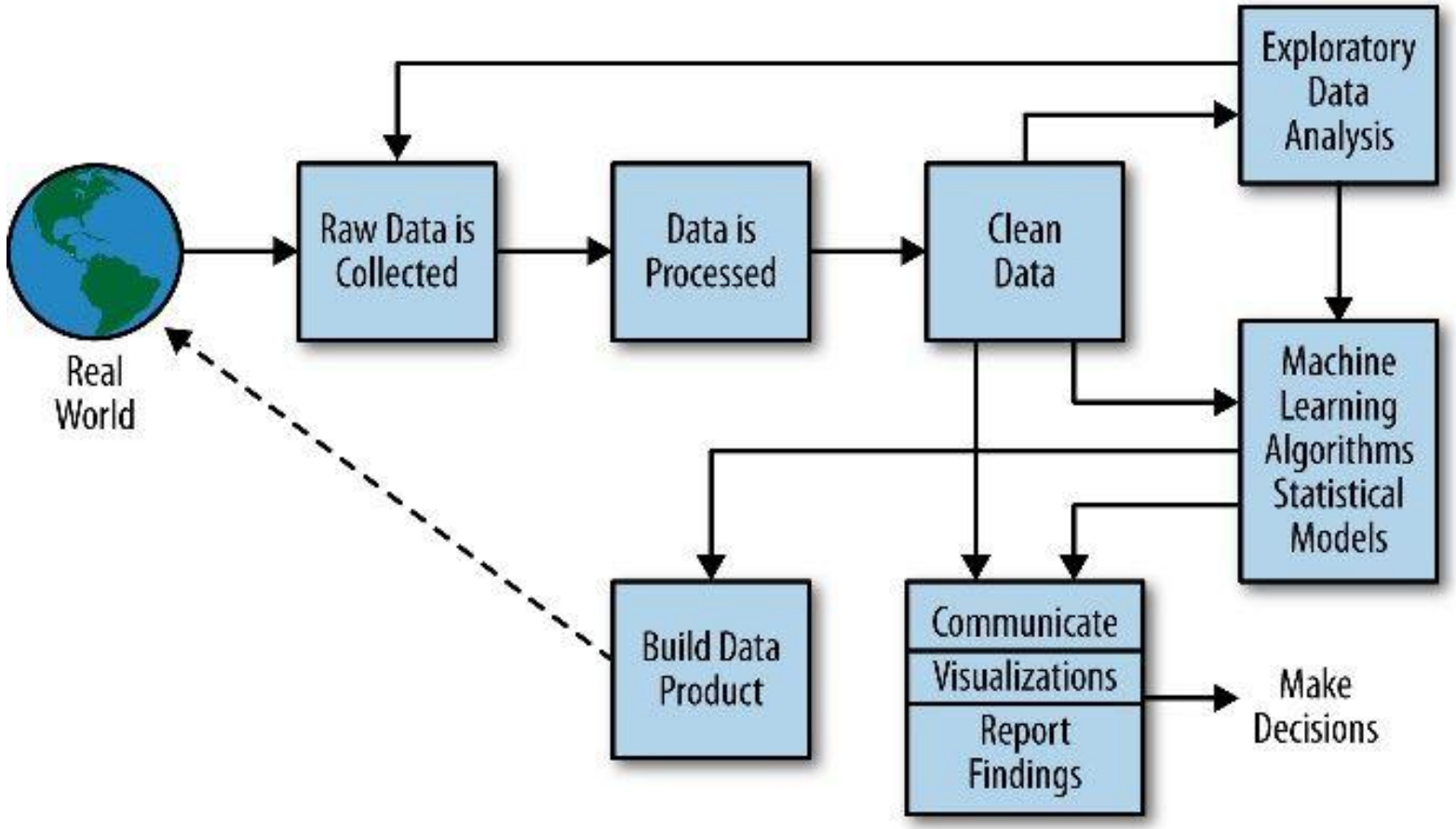
Gartner Analytic Ascendancy Model



Gartner

NASHVILLE SOFTWARE SCHOOL

The Data Science Process

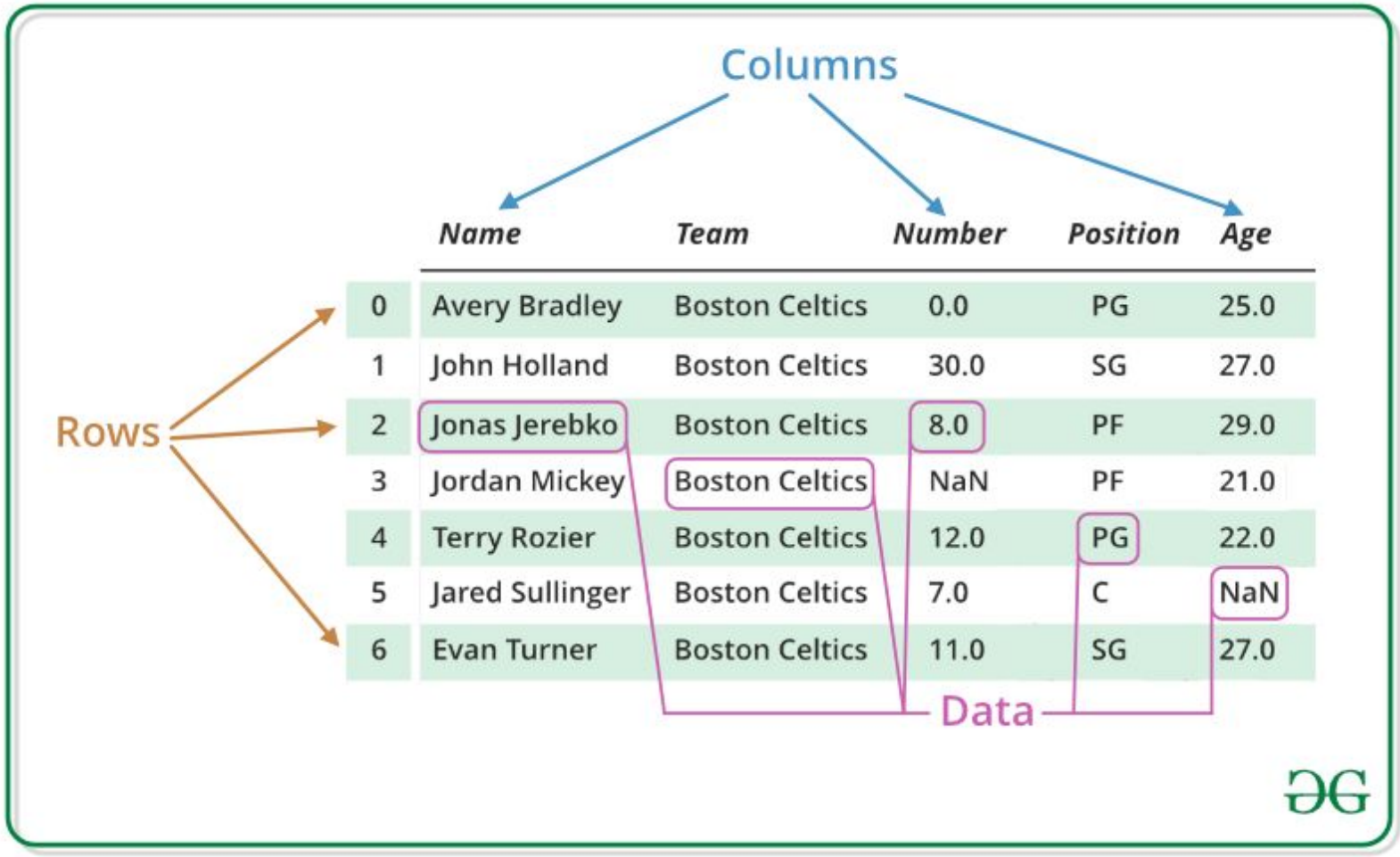


NASHVILLE SOFTWARE SCHOOL

We will work a lot with the pandas library which provides methods for working with ***DataFrames*** and ***Series***.

A DataFrame is a two-dimensional (tabular) data structure.

A Series is a one-dimensional data structure – could be a row or a column of data, but usually when we work with a series it is a column of data.



https://www.geeksforgeeks.org/python-pandas-dataframe/

NASHVILLE SOFTWARE SCHOOL

pandas – https://pandas.pydata.org/pandas-docs/stable/api.html

- **pd.read_csv()** – read a comma delimited file; good practice is to look at the raw file in a text editor (not excel); additional arguments may be needed to handle extra rows at the top and extra data (footnotes) at the bottom.
- **df.info()** – method to get information about the DataFrame
- **df.dtypes** – datatypes attribute for the Data Frame
- **df.head()** – looks at the top of the DataFrame; 5 rows by default
- **df.tail()** - looks at the bottom of the DataFrame; 5 rows by default
- **df.shape** – returns a tuple with number of rows and number of columns
- **df.columns** – column labels attribute
- **df.rename()** – rename values (can pass in a dictionary with existing columns as the key and new ones as the values)
- **df.loc[]** – pass in row name and column name to access data at that location
- **df.drop()** – drop the specified labels (either rows or columns) from the DataFrame
- **df[[ ]]** - creates a slice (subset) of the DataFrame including just the columns passed in

Get Data □ Process + Clean Data □ Exploratory Data Analysis

CSV Workflow

Open the file in text editor to determine

- is there a header? **pandas default is *header = 'infer'***
- are there notes (non-data) at the top? **pandas default is *skiprows = None***
- are there footnotes or other (non-data) at the bottom? **pandas default is *nrows = None***

Read the file into a pandas DataFrame ***pd.read_csv()***

Check the top and the bottom to ensure all data was read ***df.head()*** and ***df.tail()***

Look at the dimensions ***df.shape***

NASHVILLE SOFTWARE SCHOOL