Introduction to Analytics

LocalTek Healthcare Analytics

NASHVILLE SOFTWARE SCHOOL

Goals for the class

- Get hands-on experience accessing and manipulating data
- Get an introduction to analysis and visualization in Python
- Get an introduction to querying data in a database using SQL
- Make discoveries and have fun

NASHVILLE SOFTWARE SCHOOL

Goals for this lesson

- **Define Analytics**
- **Learn about commonly used tools in data analytics and data science**
- **Learn a little *pandas***

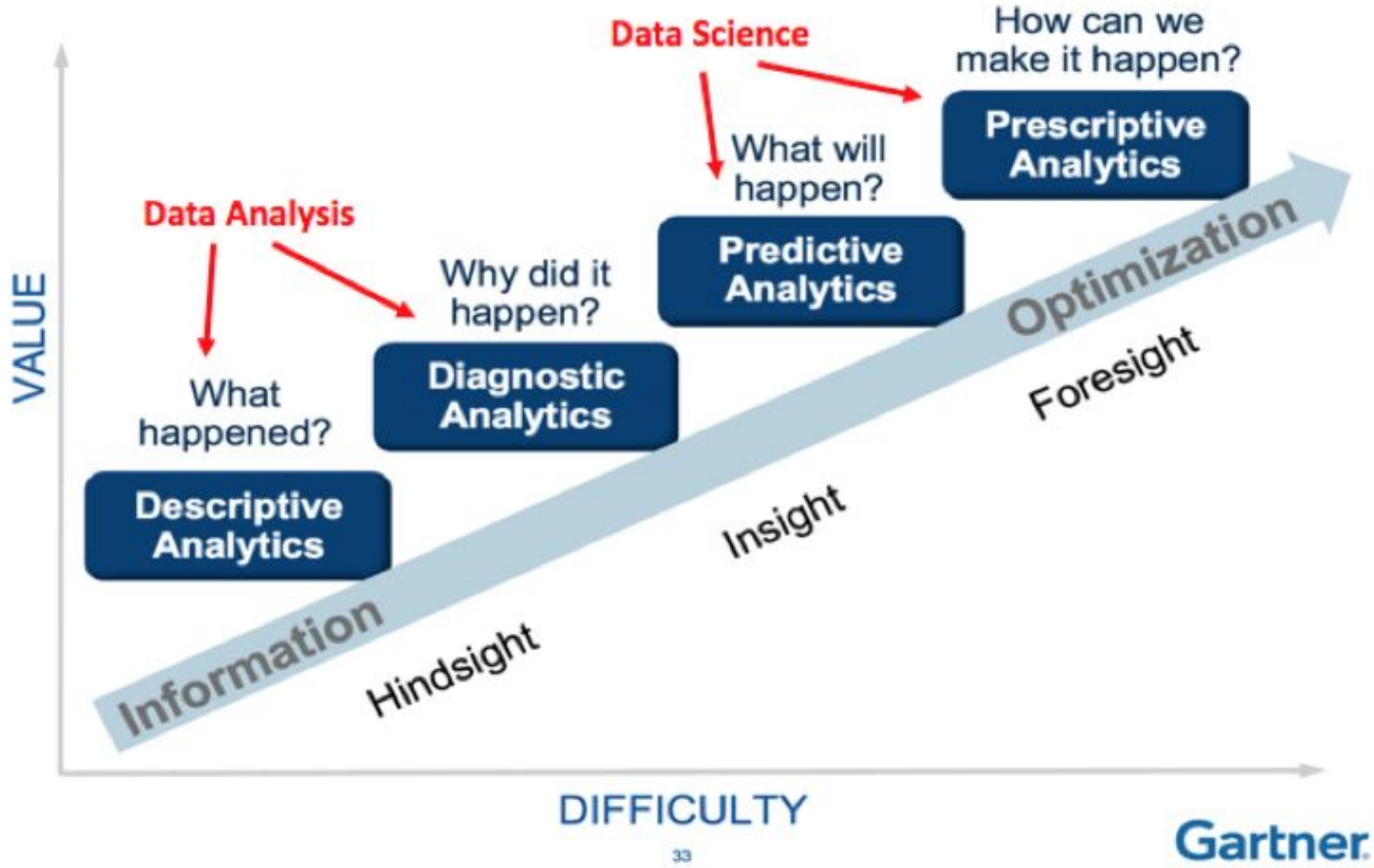
NASHVILLE SOFTWARE SCHOOL

What is Data Analytics?

The process of taking raw data, preparing and cleaning it, extracting insights, and communicating those findings to help drive better decision making.

NASHVILLE SOFTWARE SCHOOL

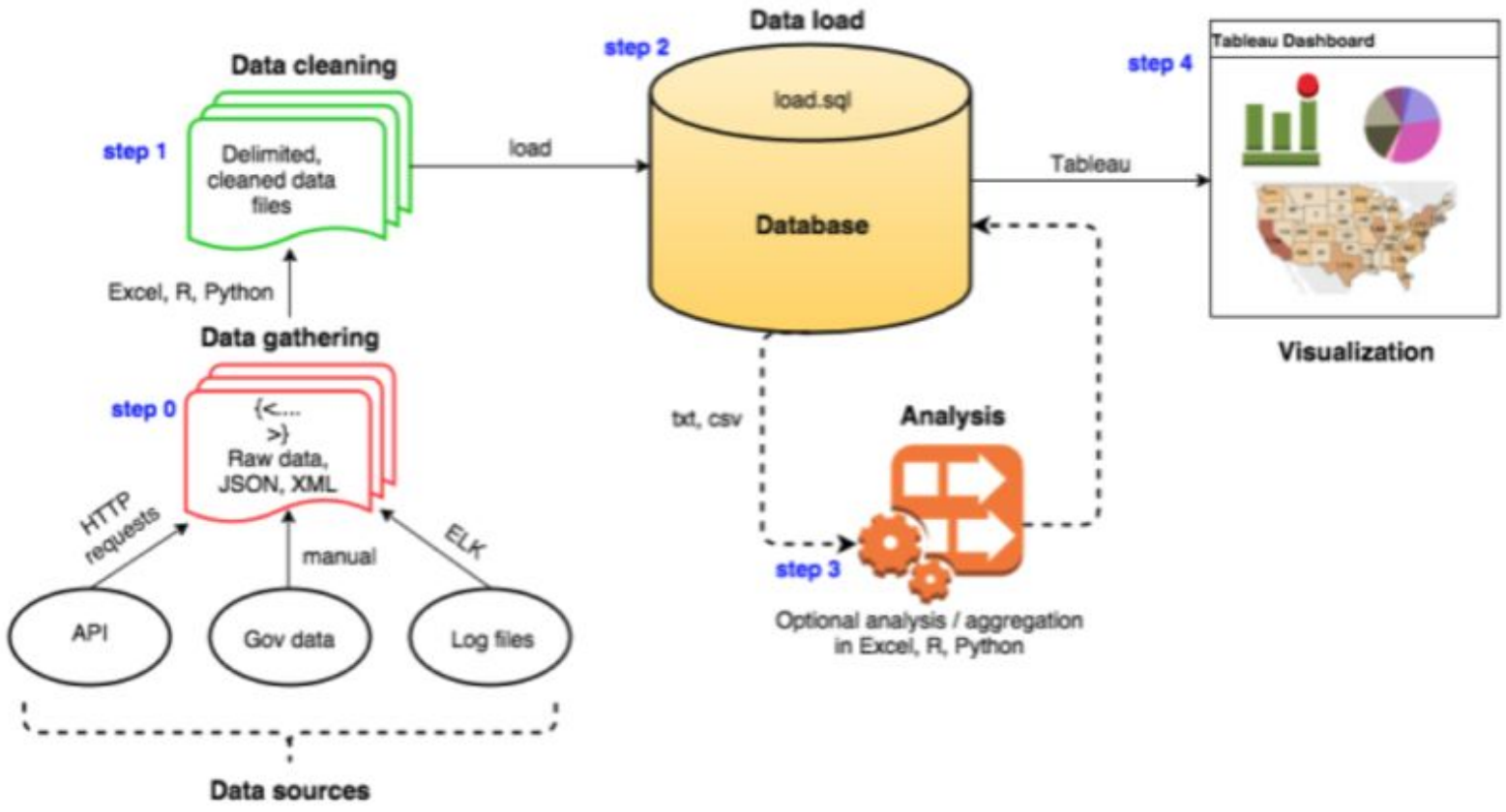
Gartner Analytic Ascendancy Model



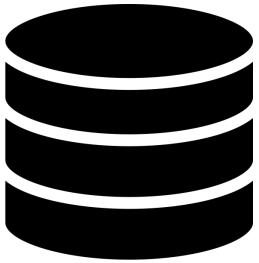
Gartner

NASHVILLE SOFTWARE SCHOOL

Analytics Workflow *- Ketaki Kekatpure*



Commonly Used Data Science and Analytics Tools



Microsoft® SQL Server®

python™

APACHE Spark™

tableau

Power BI

NASHVILLE SOFTWARE SCHOOL

Data Overview - COVID-19 Cases and Vaccinations

During this class, you will learn how to combine data from a number of different sources to explore where in the Metro Nashville area people have contracted the virus and where people are getting vaccinated.

Tools that you will use:

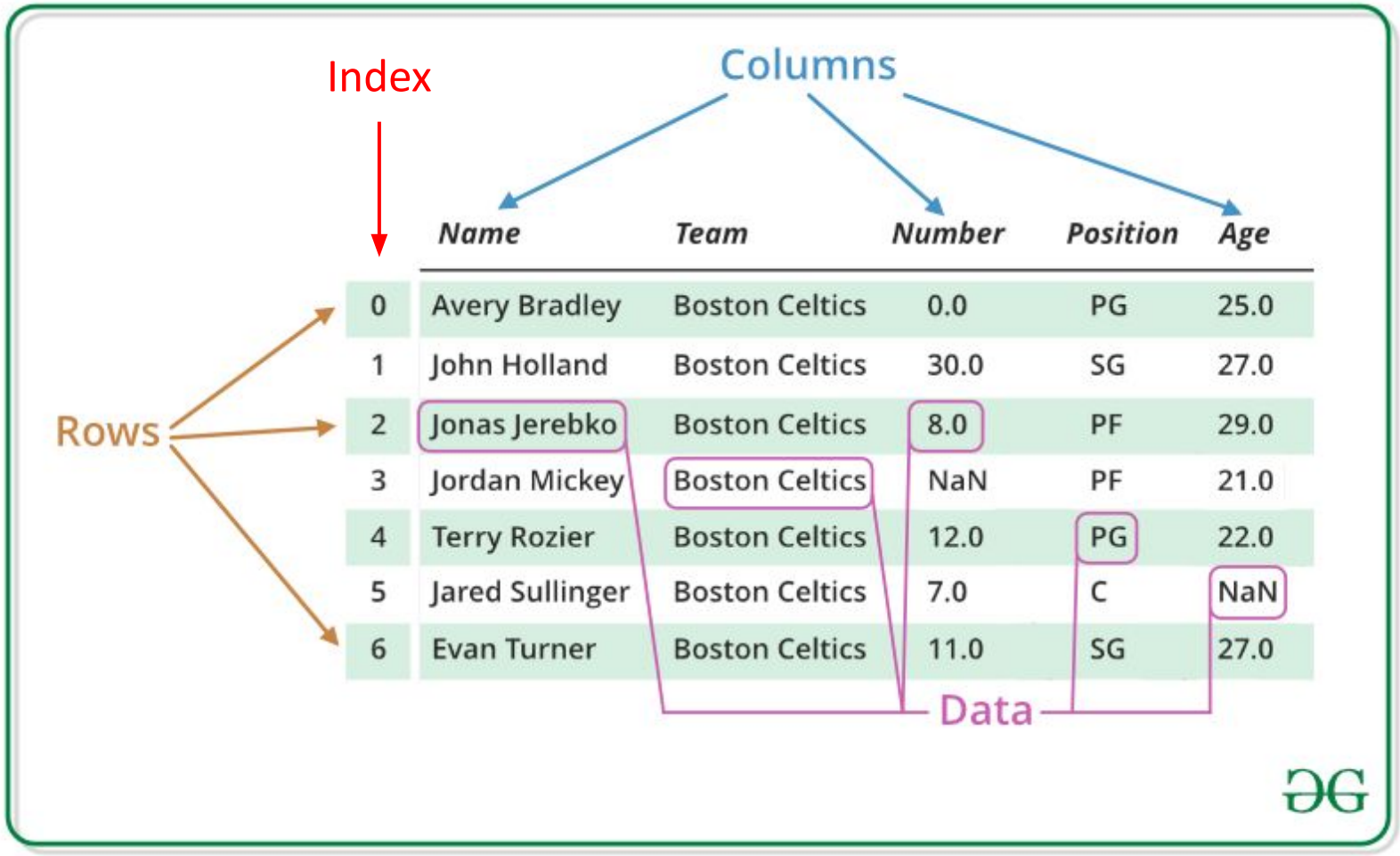
- Python
    - *pandas* for manipulating data read in from csv files, and databases
    - *matplotlib* for visualizations
- SQL
    - SQLite

NASHVILLE SOFTWARE SCHOOL

We will work a lot with Python's *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.



	Name	Team	Number	Position	Age
0	Avery Bradley	Boston Celtics	0.0	PG	25.0
1	John Holland	Boston Celtics	30.0	SG	27.0
2	Jonas Jerebko	Boston Celtics	8.0	PF	29.0
3	Jordan Mickey	Boston Celtics	NaN	PF	21.0
4	Terry Rozier	Boston Celtics	12.0	PG	22.0
5	Jared Sullinger	Boston Celtics	7.0	C	NaN
6	Evan Turner	Boston Celtics	11.0	SG	27.0

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

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

Importing Data

- **pd.read_csv()** – read a comma delimited file; good practice is to look at the raw file in a text editor (like Visual Studio Code, not Excel); additional arguments may be needed to handle extra rows at the top and extra data (footnotes) at the bottom.

Inspecting

- **df.info()** – method to get information about the DataFrame
- **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 the number of rows and number of columns

NASHVILLE SOFTWARE SCHOOL

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

Modifying

- **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.drop()** – drop the specified labels (either rows or columns) from the DataFrame

Summarizing

- **.unique()** – returns the unique values in a column
- **.nunique()** - returns the *number* of unique elements in a column
- **.value_counts()** - returns the unique elements in a column and the number of appearances of each

Slicing/Filtering

- **df.loc[]** – pass in row name and column name to access data at that location
- **df[[ ]]** - creates a slice (subset) of the DataFrame including just the columns passed in

NASHVILLE SOFTWARE SCHOOL

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