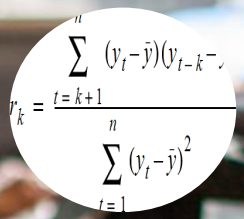
WHAT & WHY?

DATA SCIENCE INTRODUCTION



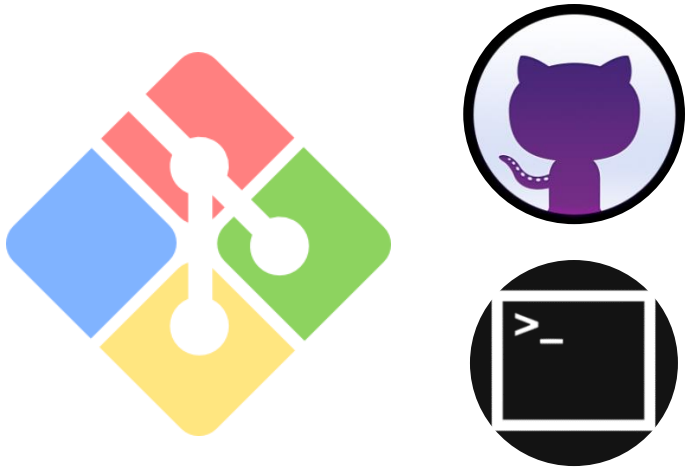
SETUP

1) GITBASH
    1) https://git-scm.com/downloads
2) ANACONDA 2.7
    1) https://www.continuum.io/downloads
3) SETUP A GITHUB PROFILE
    1) https://github.com

SETUP

ANACONDA

GITBASH



WHERE DOES DATA SCIENCE COME IN?

The world is full of situations where we want to:

1. Forecast likely outcomes given that we know now
2. Segment things into groups
3. Recommend something based on the likelihood it will be viewed or clicked on

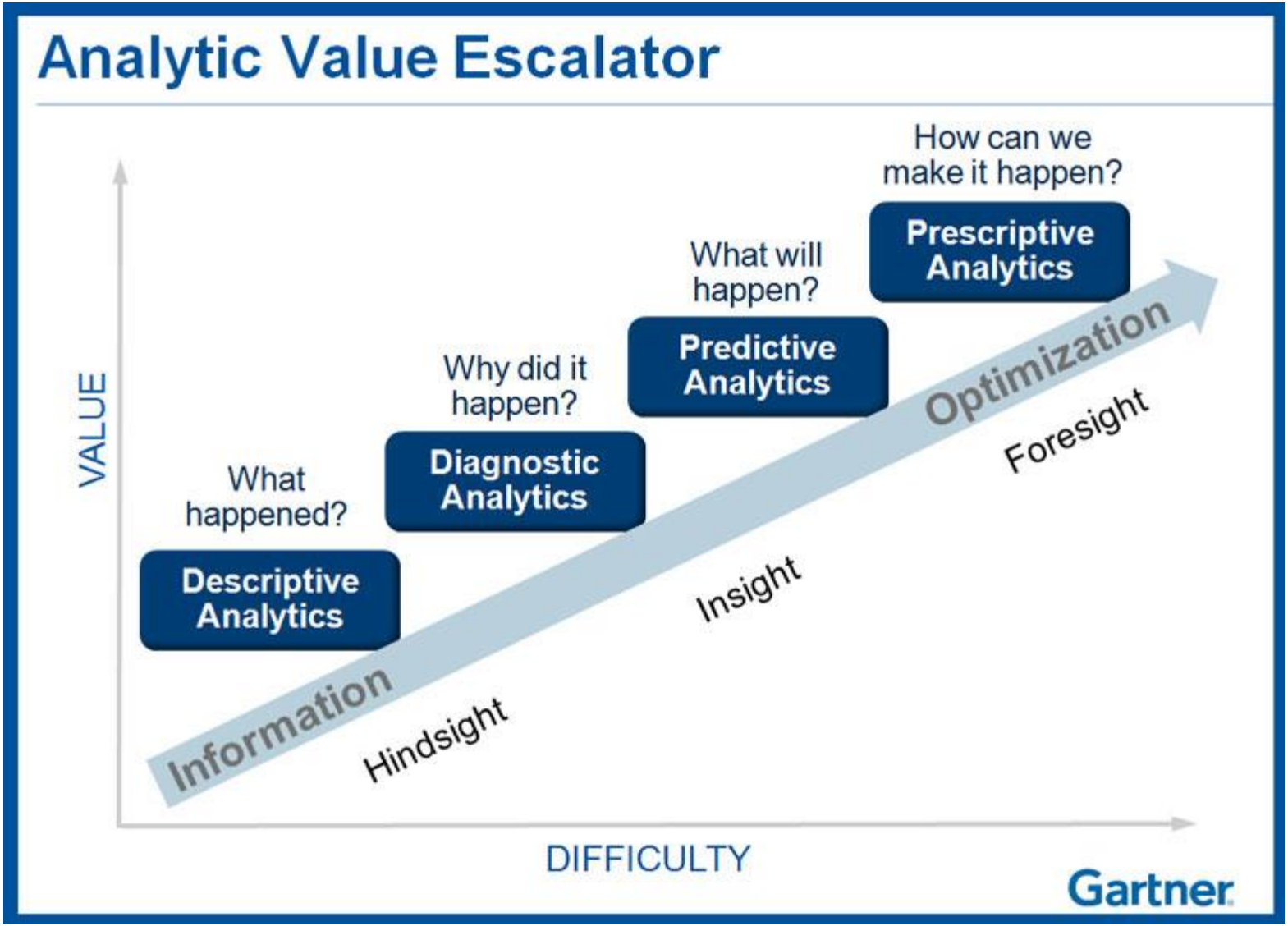
EXAMPLES OF DATA SCIENCE IN ACTION

- Facebook facial recognition in photos
- Netflix/Amazon/Spotify recommendations
- Siri/Echo/Cortana voice recognition assistants
- Building art with Neural Networks - https://github.com/jcjohnson/neural-style
- Faceswap - https://www.youtube.com/watch?v=UngUWA43q5o
- Stock Market - https://www.quantopian.com/ - building crowd source hedge funds
- Helping people
  https://www.drivendata.org/ - who is a good bet to give money to
  http://www.datakind.org/projects
  - Help find missing children - http://www.datakind.org/projects/finding-30000-missing-children
- Find correlations from sickness, grades, and attendance and try to find ways to improve them http://coolculture.org/webfm_send/62
- Additional examples https://www.kaggle.com/wiki/DataScienceUseCases

SOME OF THE TECHNIQUES APPLIED IN DATA SCIENCE

Forecast and prediction from *numeric* values	"What are our sales going to be next year given the trend in the sales of our product lines?"	Regression
Segmentation and cluster analysis	"What is a good grouping of our customers that I can use to think about how best to appeal to them?"	K-Means, DBSCAN
Spam filter	"Should this email message be classified as spam?"	Naïve Bayes
Matching web site users of similar interest	"What group is this new web page likely to appeal to"	Nearest neighbor, SVC, Ensemble Classifiers

BASIC DEFINITIONS



Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
907	418

12:55 PM - 3 May 2012

Javier Nogales
@fjnogales

Follow

Data Scientist (2/2): person who is worse at statistics than any statistician and worse at software engineering than any software engineer

RETWEET 1

FAVORITES 5

9:08 AM - 27 Jan 2014

WHAT IS A DATA SCIENTIST?

"Data Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type A (for Analysis):

- Primarily concerned with **making sense of data** or working with it in a fairly **static** way.
- Similar to a statistician, but knows all the **practical details of working with data** that aren't taught in statistics: data cleaning, dealing with large data sets, visualization, domain knowledge, etc.

Source: https://www.quora.com/What-is-data-science/answer/Michael-Hochster

WHAT IS A DATA SCIENTIST?

"**Data** Scientists are people with some mix of **coding and statistical skills** who work on **making data useful** in various ways."

Data Scientist Type B (for Building):

- Some statistical background, but **strong coder or software engineer**.
- Primarily concerned with **using data "in production"**: building models which interact with users (by giving recommendations, for example).

Our course is focused primarily on **Type A**.

BASIC DEFINITIONS

Data storage, Understanding schemas, tables, fields, relational and non relational databases is a foundation of data analytics

BASIC DEFINITIONS

Oracle express 11g edition/ Oracle SQL Developer

Postgres/Pgadmin4

BASIC DEFINITIONS

Business knowledge can include understanding: Knowing KPI's, Gather requirements, MetaData, Operational reports, Business acumen, communication and navigating politics and personalities of your business culture

Having a strong understanding of Lookup functions, string and numeric functions is necessary to understand the business and how the currently tackle problems.

$$r_k = \frac{\sum_{t=k+1}^{n} (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n} (y_t - \bar{y})^2}$$

Basic Statistics(Central Tendency) Understanding concepts is fine. Understanding long hand even better.

This is a must to understand the basic charts and graphs and be able to tell a story with them.

Structured Query Language: Unless someone is getting all of your data for you and cleaning it all for you, you will want to be proficient in SQL up to Advanced levels.

BASIC CONCEPTS

Python is a general purpose programing language. Allows you to give directions to a computer to tell it what to do.

R is a system for statistical computation and graphics.

SAS **SAS** (Statistical Analysis System) is a software suite developed by **SAS** Institute for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics.

SPSS Modeler IBM **SPSS Modeler** is a data mining and text analytics software application from IBM. It is used to build predictive models and conduct other analytic tasks

BASIC CONCEPTS

Jupyter Notebooks allows you to create and share documents that contain live code, equations, visualizations and explanatory text.

Plain text formatter that converts for use in html, used to create documentation within Jupyter

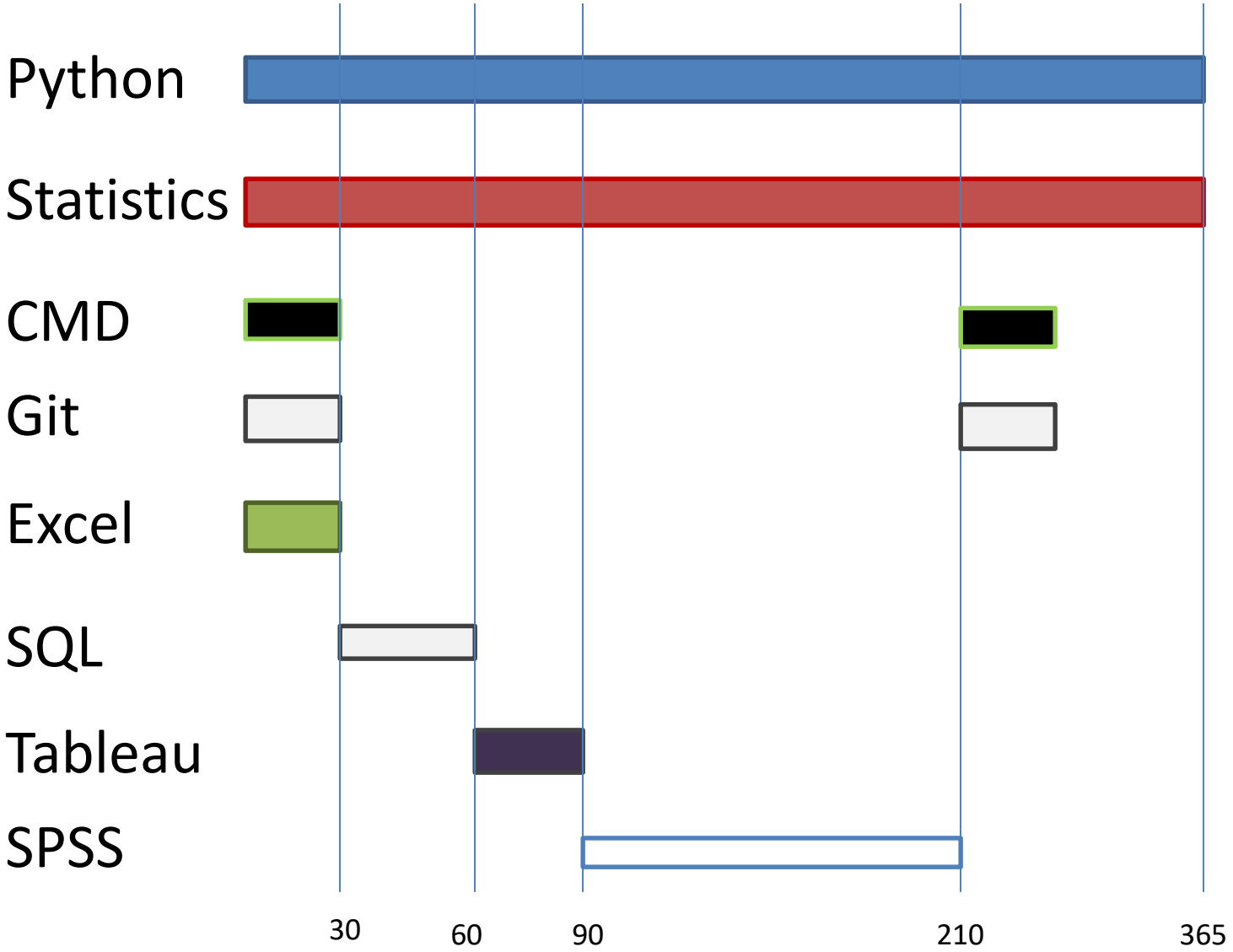
Purpose of git is to manage a project, or a set of files as they change over time. It allows for version control and collaboration.

Command line is a user interface to a computers operating system. It allows you to navigate, manipulate and analyze files, data and more.

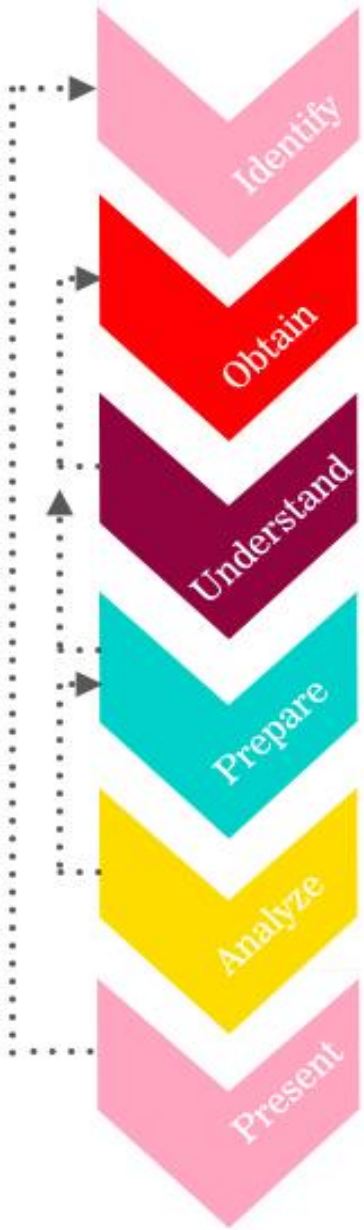
Python Vs R


adidas
BR
adidas
PETROBRAS
BOXE
ITALIA
MILANO
2009 AIBA
WORLD BOXING
CHAMPIONSHIPS

Training Path



ANALYTICS WORKFLOW



IDENTIFY THE PROBLEM

- ☐ Identify business/product objectives
- ☐ Identify and hypothesize goals and criteria for success
- ☐ Create a set of questions for identifying correct data set

OBTAIN THE DATA

- ☐ Identify the "right" data set(s)
- ☐ Import data and set up local or remote data structure
- ☐ Determine most appropriate tools to work with data

UNDERSTAND THE DATA

- ☐ Read any documentation provided with the data
- ☐ Perform exploratory data analysis
- ☐ Verify the quality of the data

PREPARE THE DATA

- ☐ Determine sampling methodology and sample data
- ☐ Format, clean and combine data
- ☐ Create necessary derived columns from the data (new data)

ANALYZE THE DATA

- ☐ Identify trends and outliers
- ☐ Apply descriptive and inferential statistics
- ☐ Visualize and transform data

PRESENT THE RESULTS

- ☐ Summarize findings with narrative, storytelling techniques
- ☐ Present limitations and assumptions of your analysis
- ☐ Identify follow up problems and questions for future analysis

Data Science Workflow

Define feature vector matrix

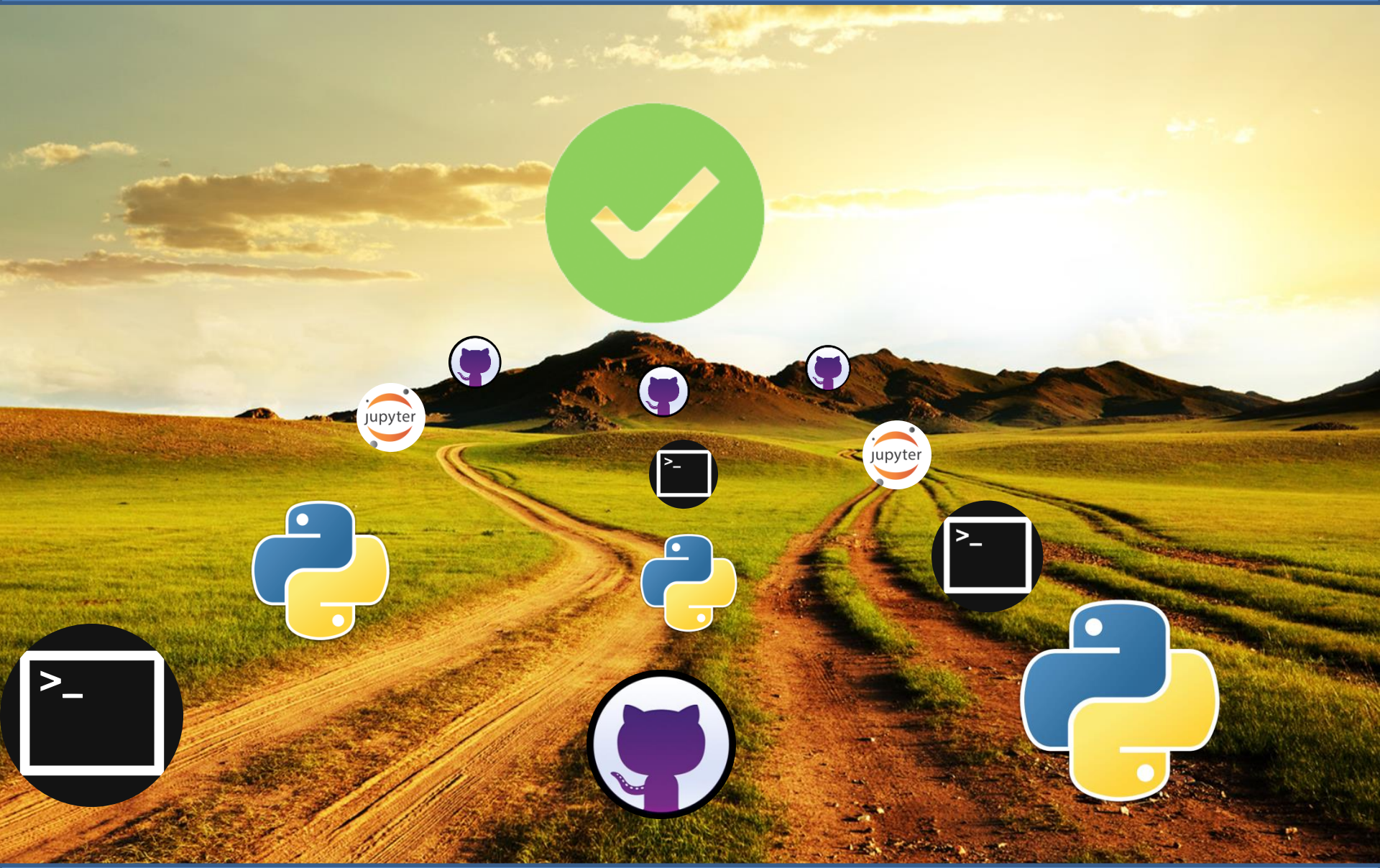
Choose an estimator

Insatiate the estimator

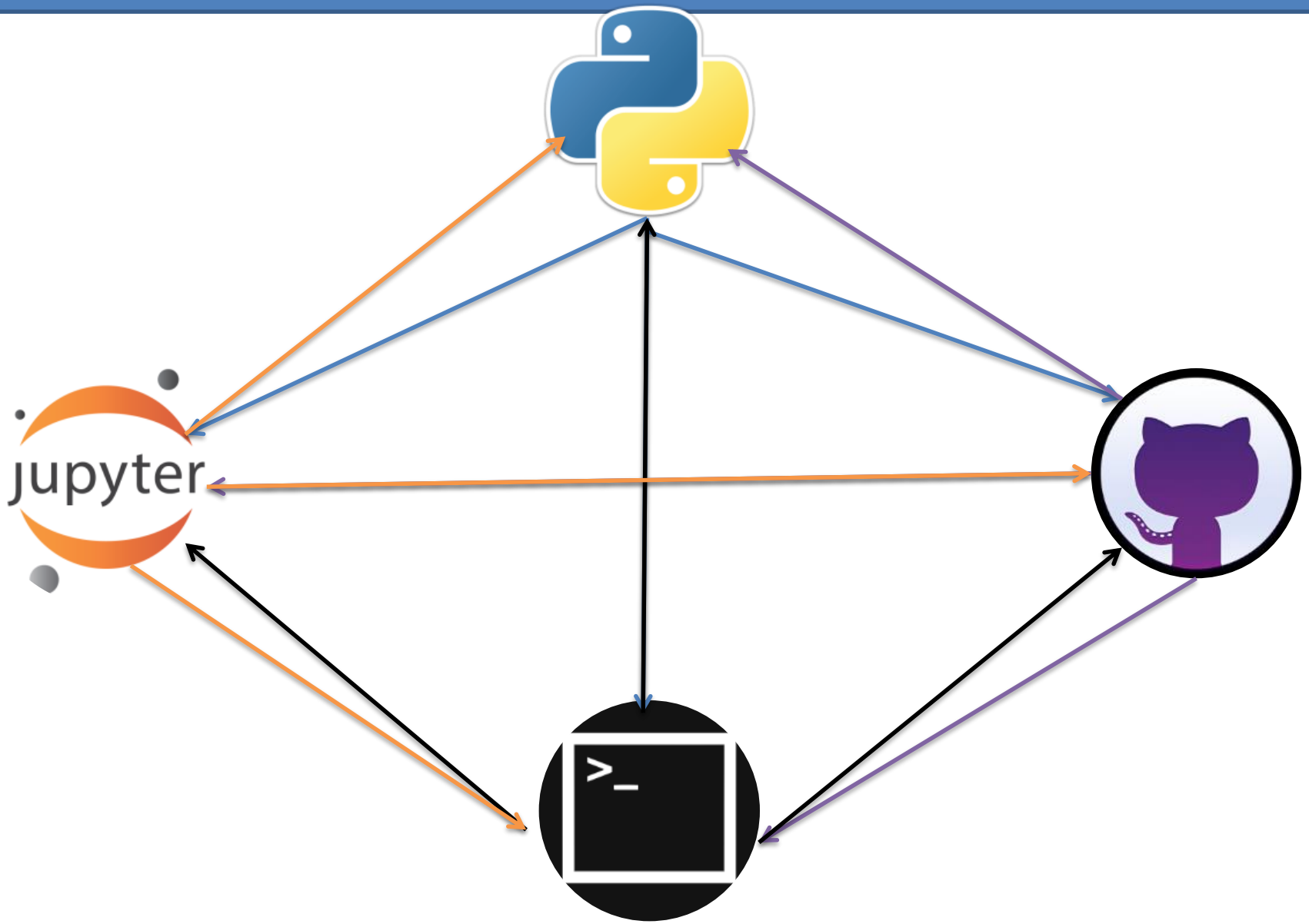
Make a prediction

Evaluate the model

Reality of Workflows



AGILE



Moving data around and exploring

1) Set up Sandbox using CMD
2) Pull data using git
3) Review data using CMD
4) Explore data using Python
5) Basic concepts of predictive models

SET UP SANDBOX USING CMDLINE

pwd Present working directory
ls- list files directories and subdirectories
cd Change directory
cd path/ Change directory and path name
mkdir- make a new directory
Git init – initialize new git repository

Exercise

Make a new directory on your desktop called Sandbox
In Sandbox make another directory called Dsintro

PULL DATA USING GIT

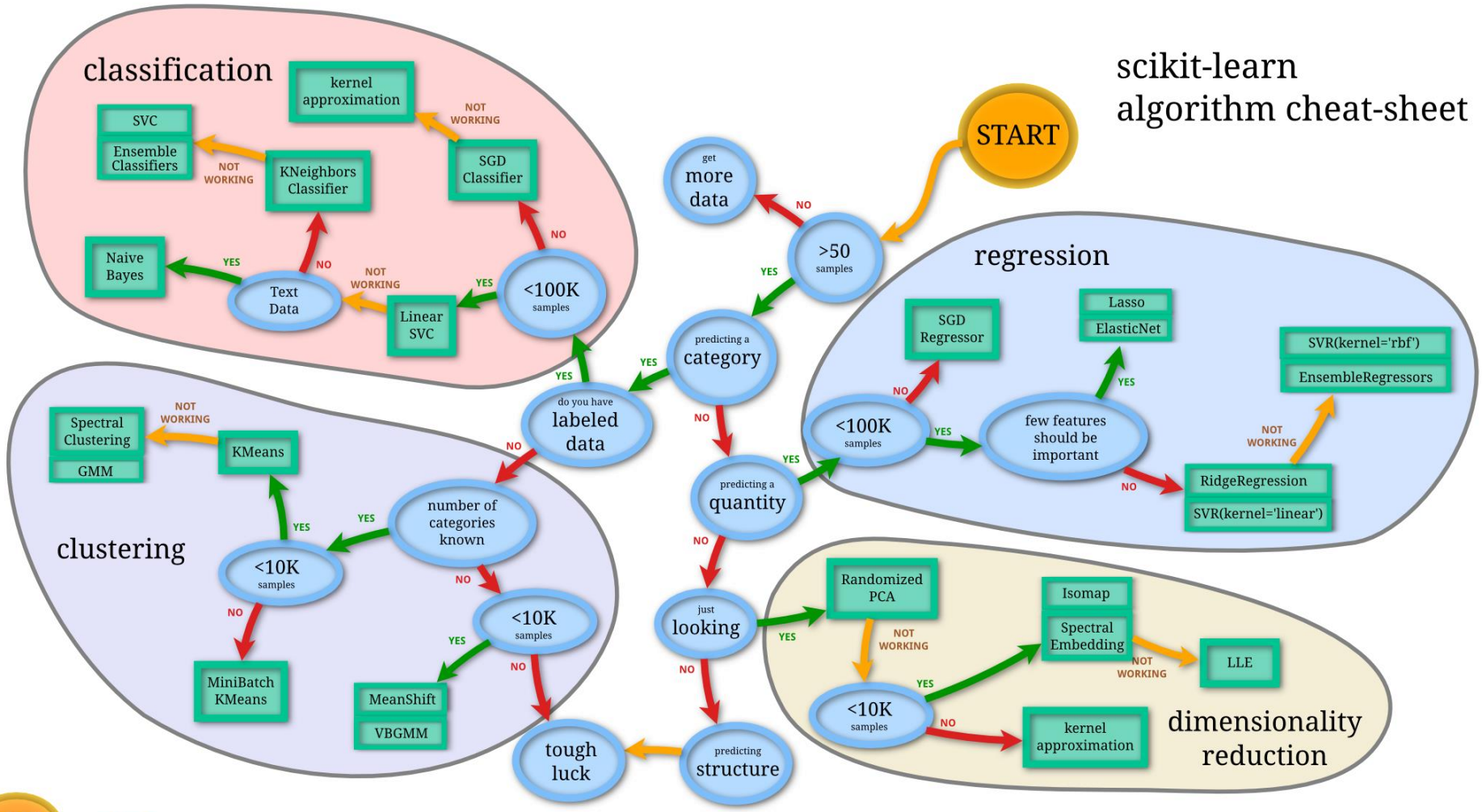
git pull https://github.com/Morrisdata/DataScience101.git

What just happened and why do you care?

head<filename>
 prints the head (the first 10 lines) of the file
head -n20 <filename>
prints the first 20 lines of the file
tail <filename>
 prints the tail (the last 10 lines) of the file

EXPLORE DATA USING PYTHON

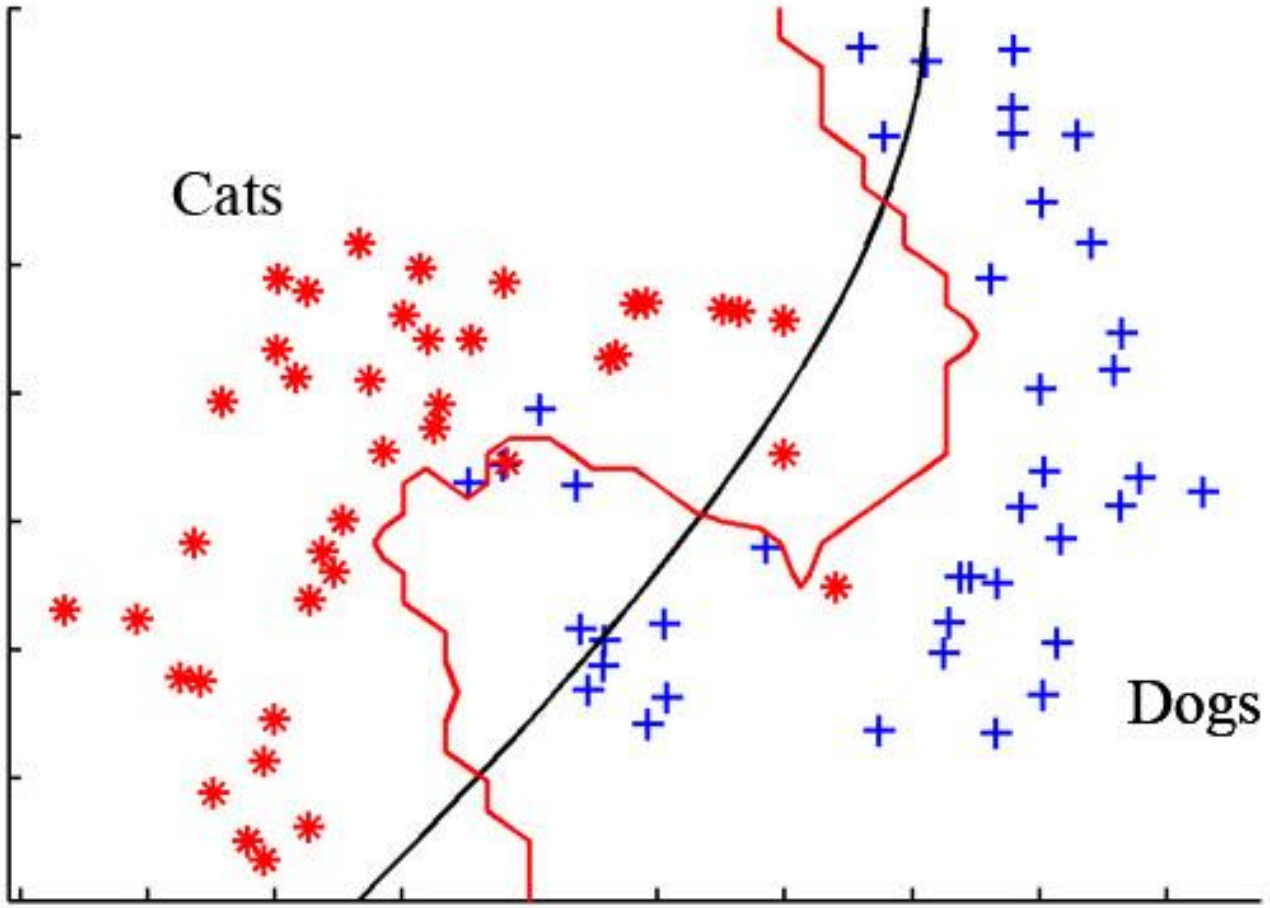
Training Path



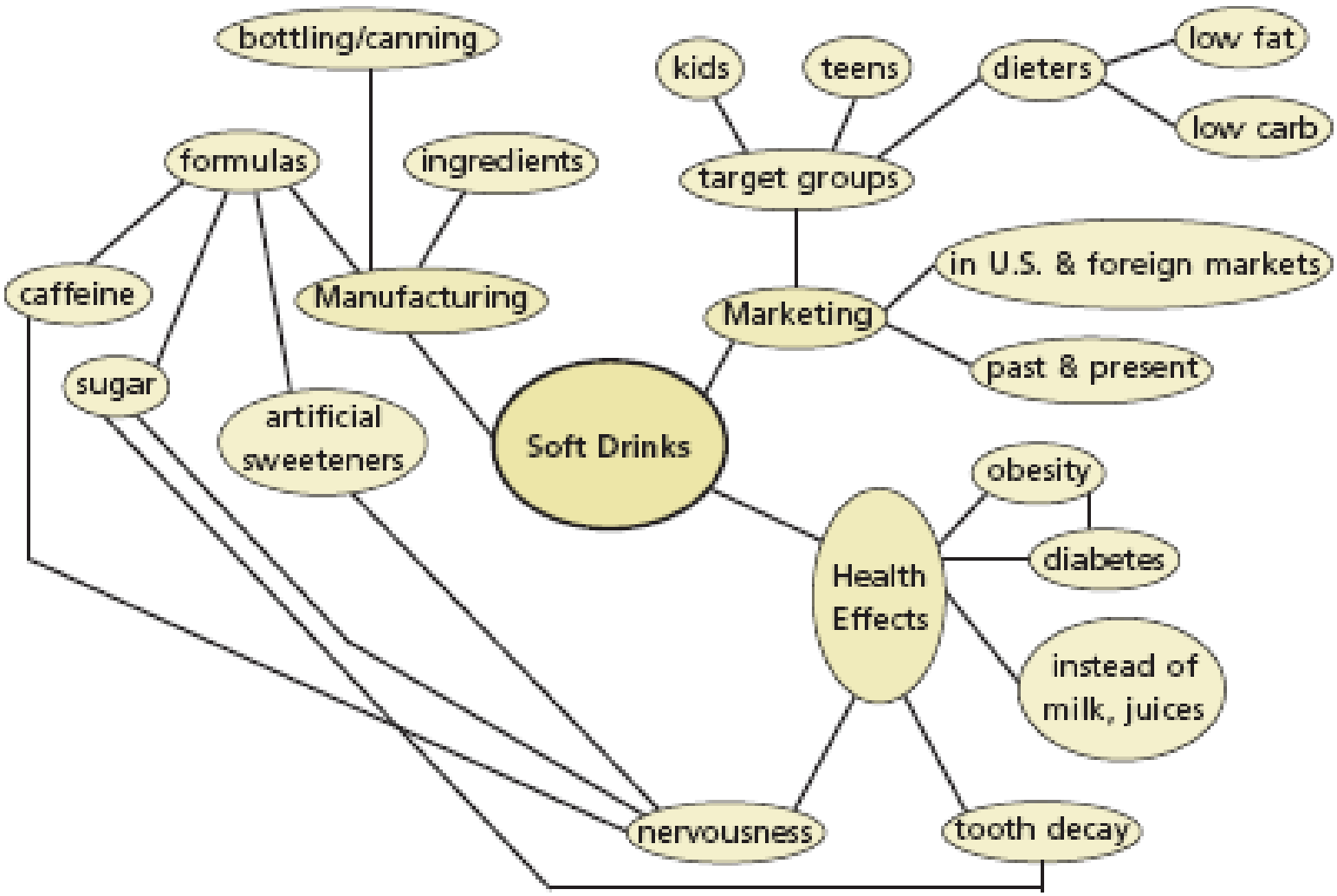
Back

scikit learn

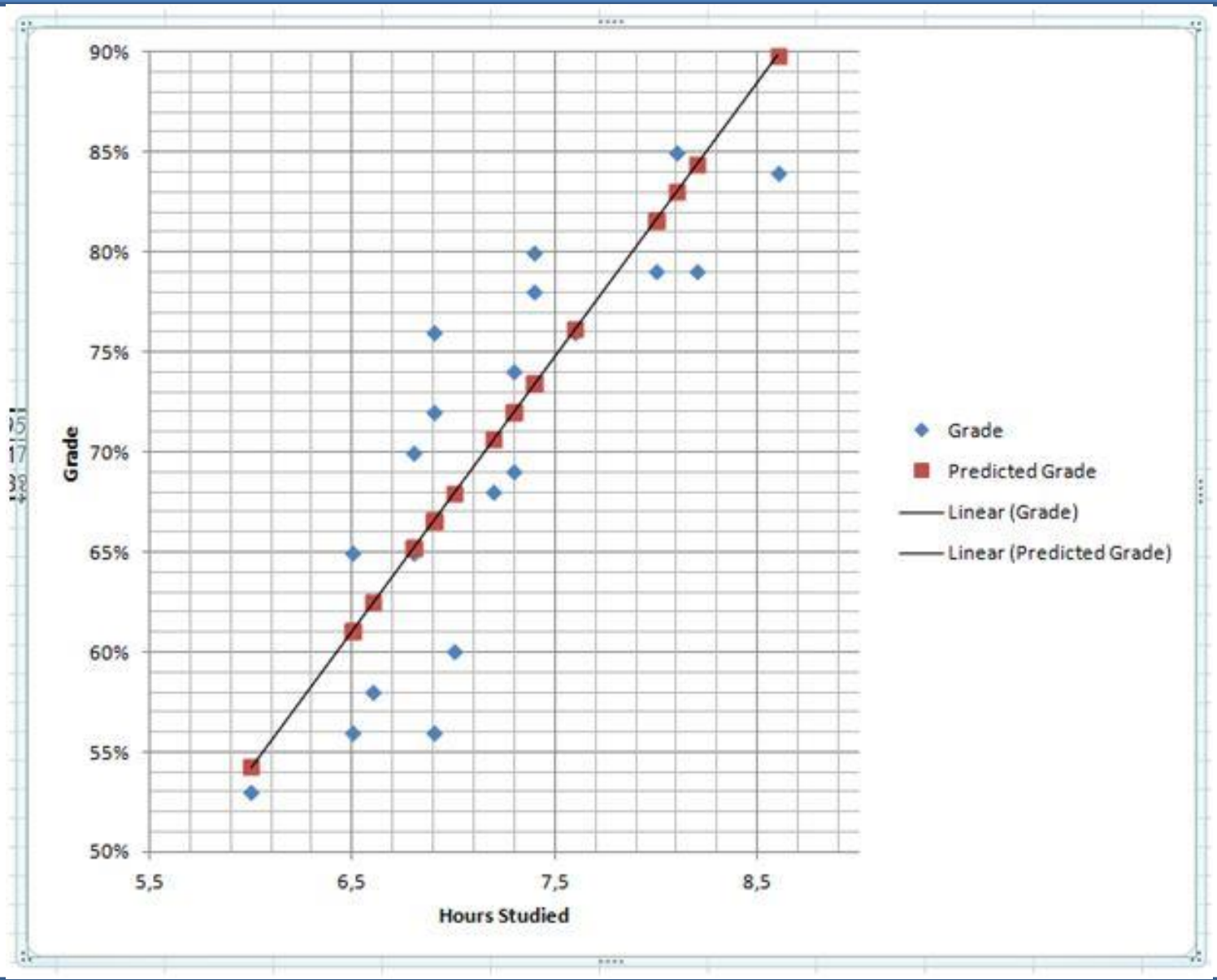
Classification



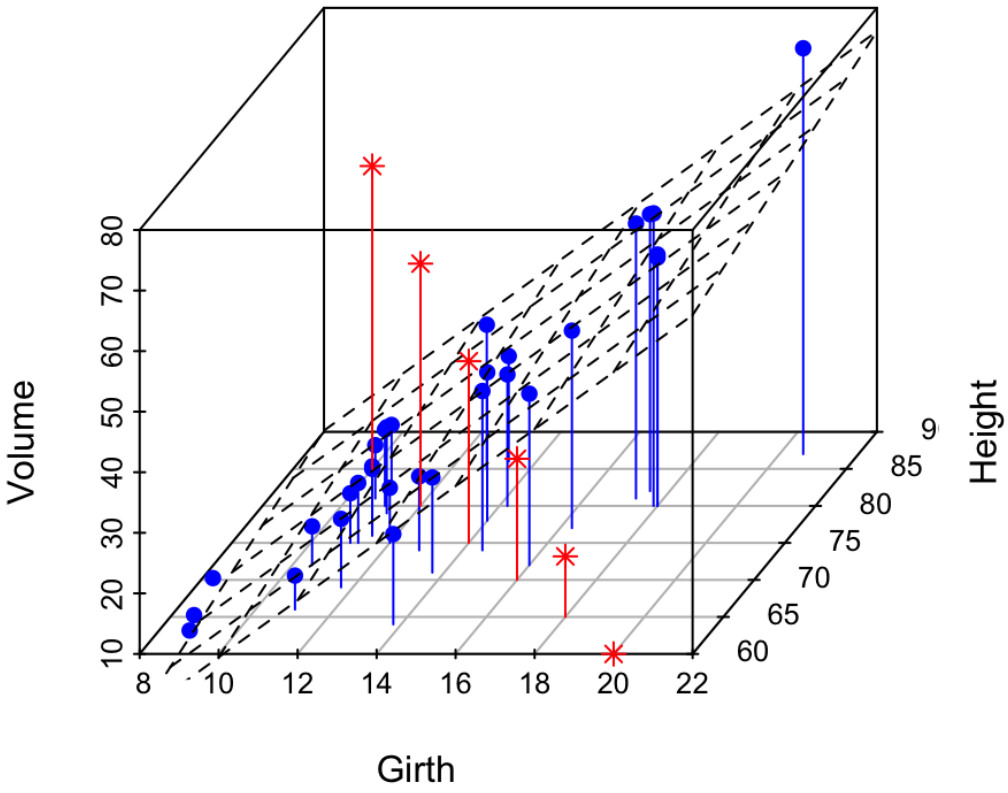
Clustering


bottling/canning
formulas
ingredients
caffeine
Manufacturing
sugar
artificial sweeteners
Soft Drinks
kids
teens
dieters
low fat
low carb
target groups
in U.S. & foreign markets
Marketing
past & present
obesity
diabetes
Health Effects
instead of milk, juices
nervousness
tooth decay

Regression


90%
85%
80%
75%
70%
65%
60%
55%
50%
Grade
5,5
6,5
7,5
8,5
Hours Studied
Grade
Predicted Grade
Linear (Grade)
Linear (Predicted Grade)

Dimensionality Reduction



WHAT DID WE JUST DO?

Set up Sandbox using command line
Pull data using git
Review data using cmd
Explore data using Python
Basic concepts of predictive models

ADDITIONAL RESOURCES

- How to Lie With Statistics - Darrell Huff
- What is a p-value anyway? 34 Stories to Help You Actually Understand Statistics - Andrew Vickers
- Teaching Statistics: A Bag of Tricks - Andrew Gelman and Deborah Nolan
- An introduction to Statistical Learning: with applications in R - James Gareth
- Python Machine Learning - Sebastian Raschka