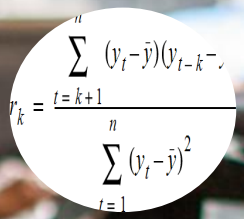
# WHAT & WHY?

BASIC SETUP





# 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 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 CONCEPTS

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

R

SAS

SPSS Modeler

# 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

Statistics
Python
Excel
SQL
Tableau
CMD line
Git
Jupyter
SPSS Modeler
Python
R

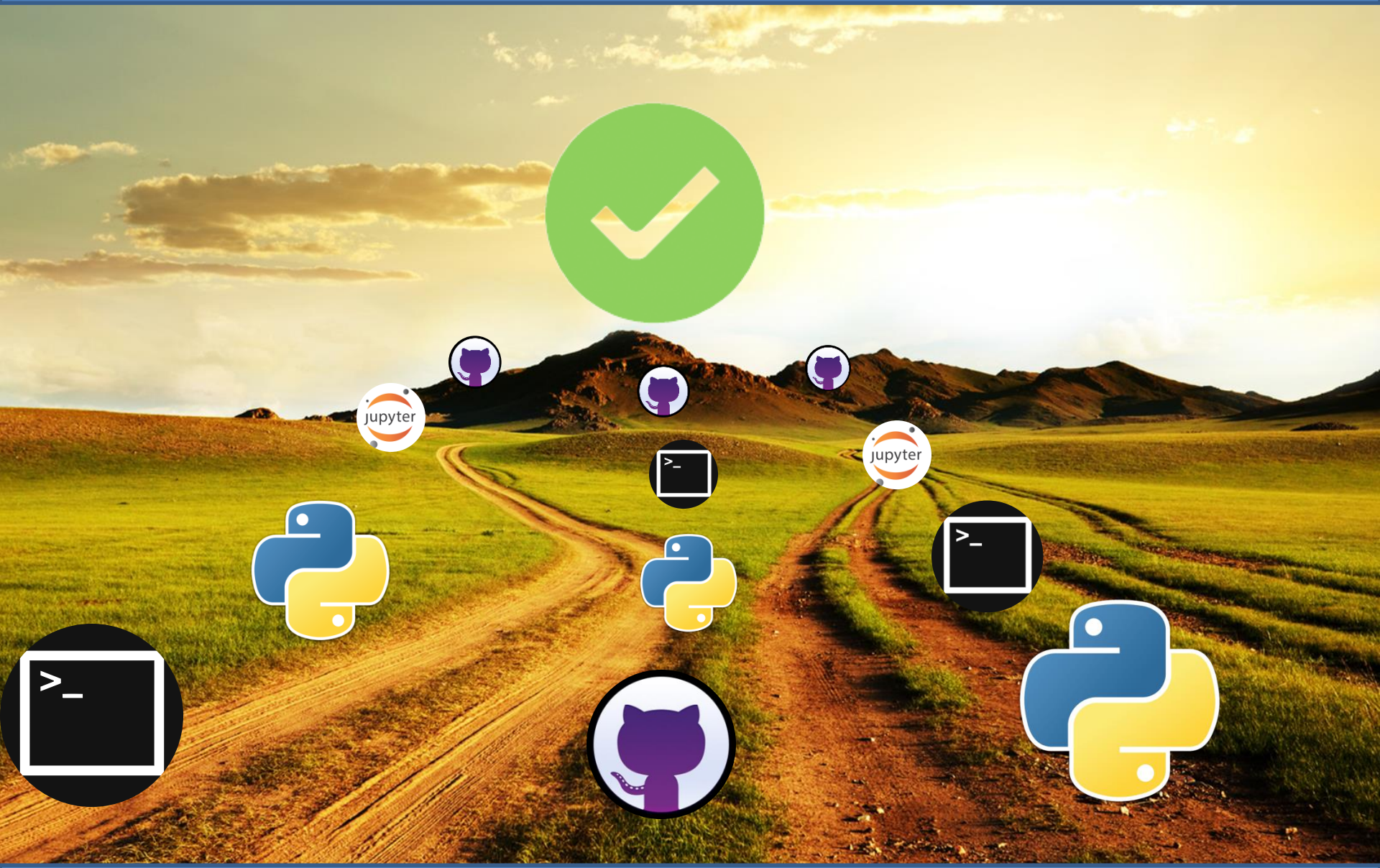
# Training Path

Data Analytics workflow
Identify a problem
Obtain Data
Understand Data
Prepare Data
Analyze  Data
Present Data

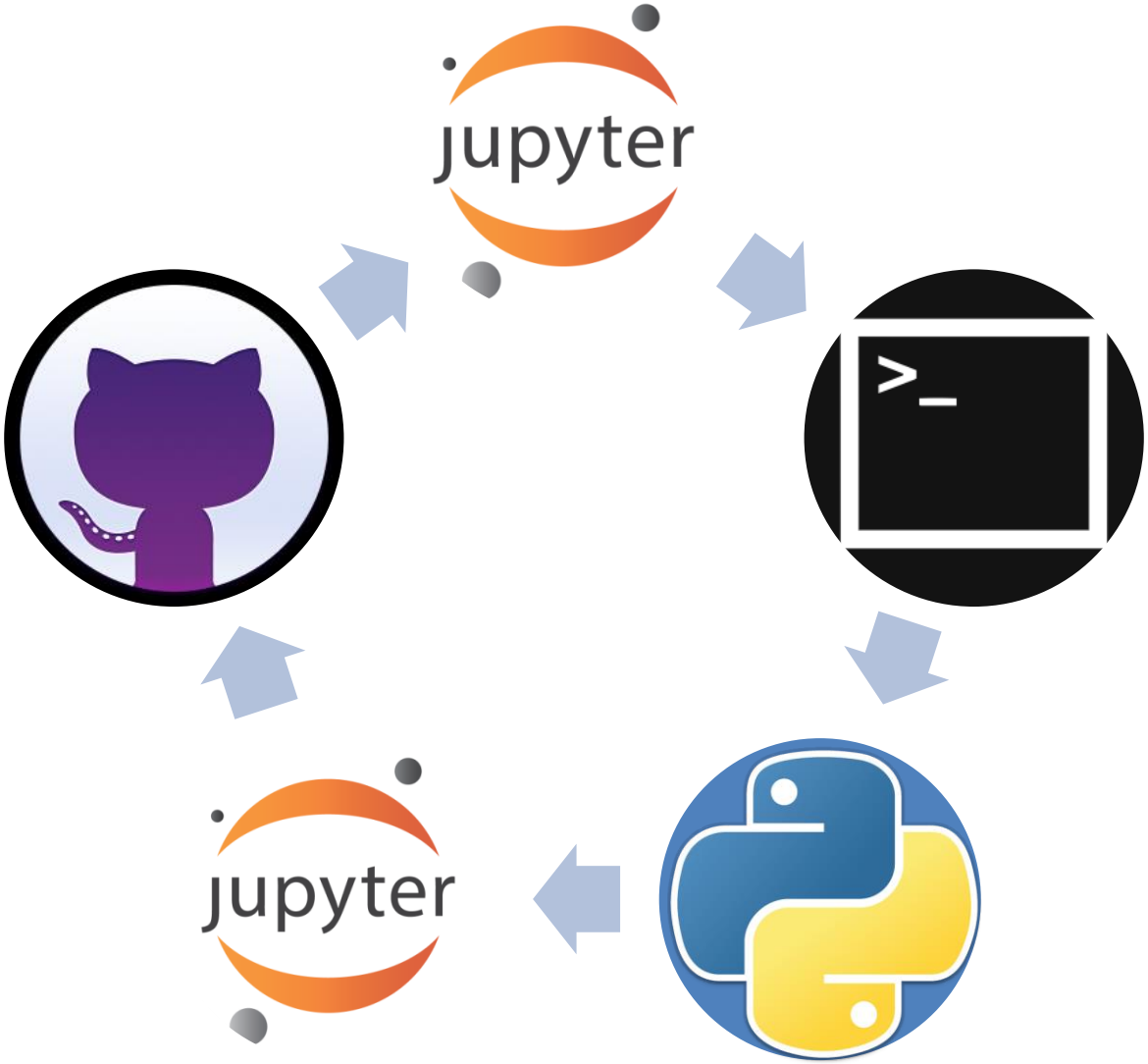
# Training Path

Data Science Workflow
Define feature vector matrix
Choose an estimator
Insatiate the estimator
Make a prediction
Evaluate the model

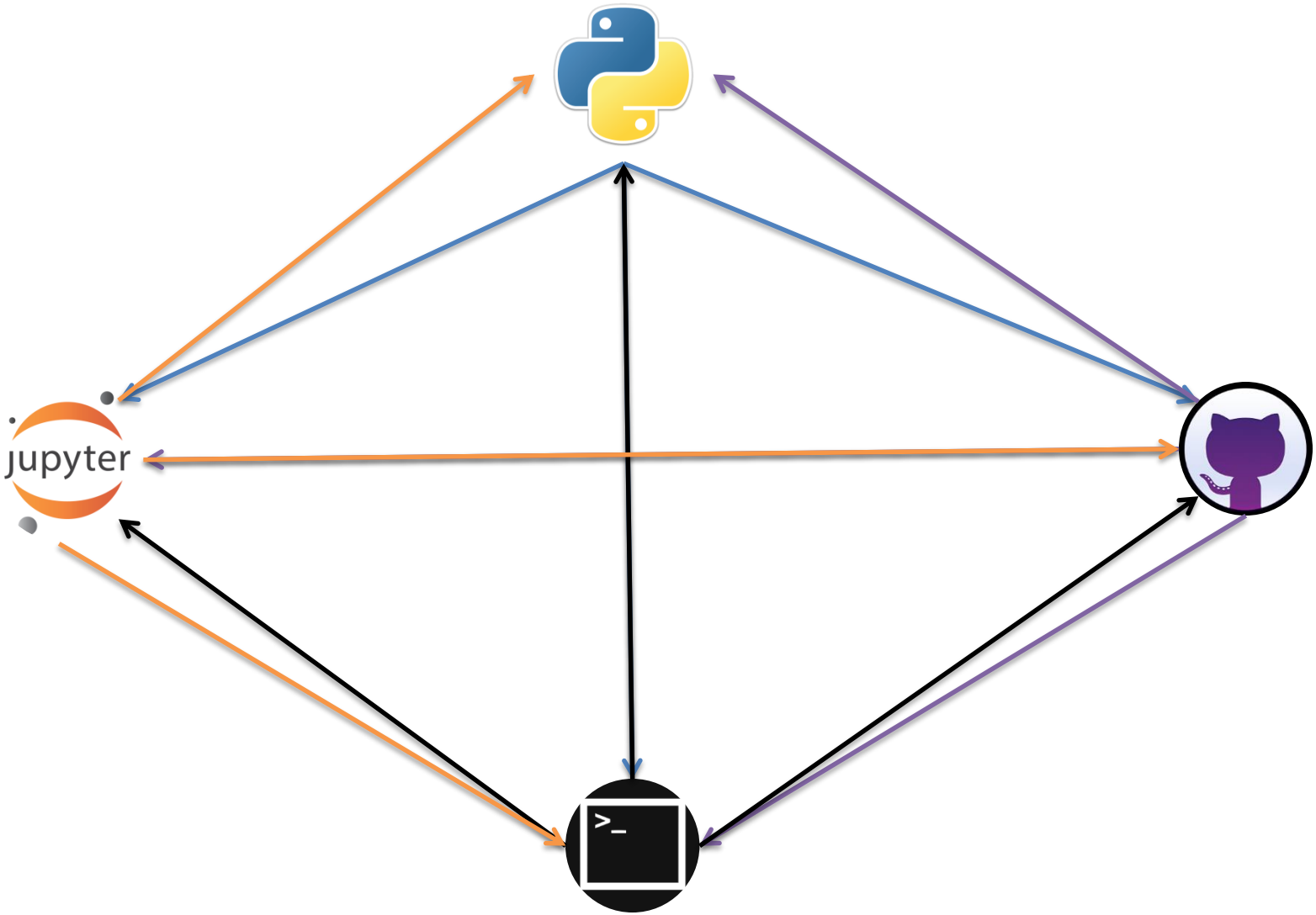
# Reality of Workflows



# WORKFLOWS ITERATE



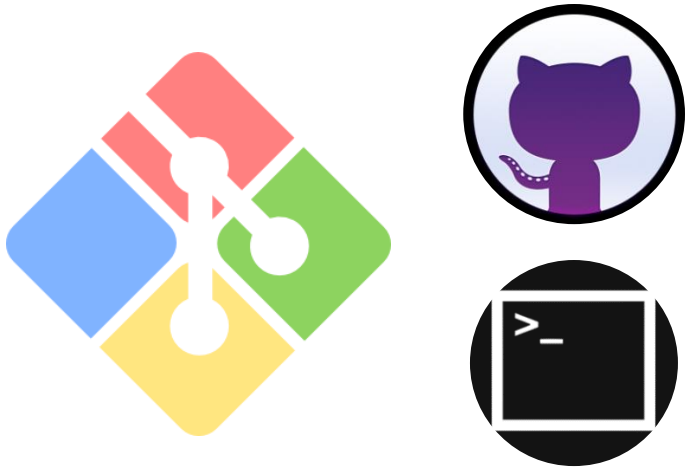
# AGILE



# SETUP


ANACONDA

GITBASH



# DEMO

Set up Sandbox using command line
Pull data using git
Analyze data using cmd
Create program in python
Run program in cmd line
Push data to git
Document using markdown in jupyter notebooks

# 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

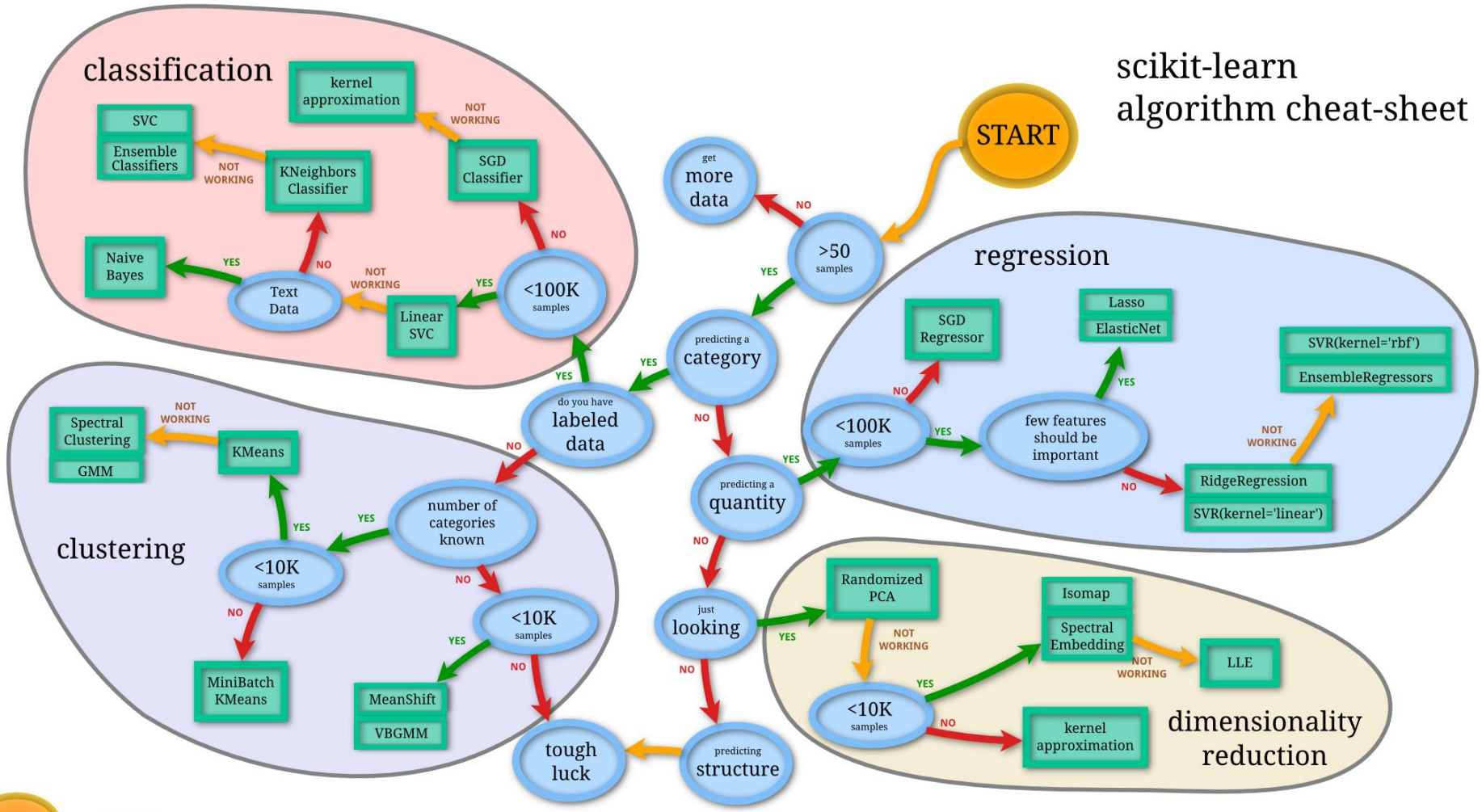
git fetch https://github.com/Morrisdata/Python_for_Data/

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

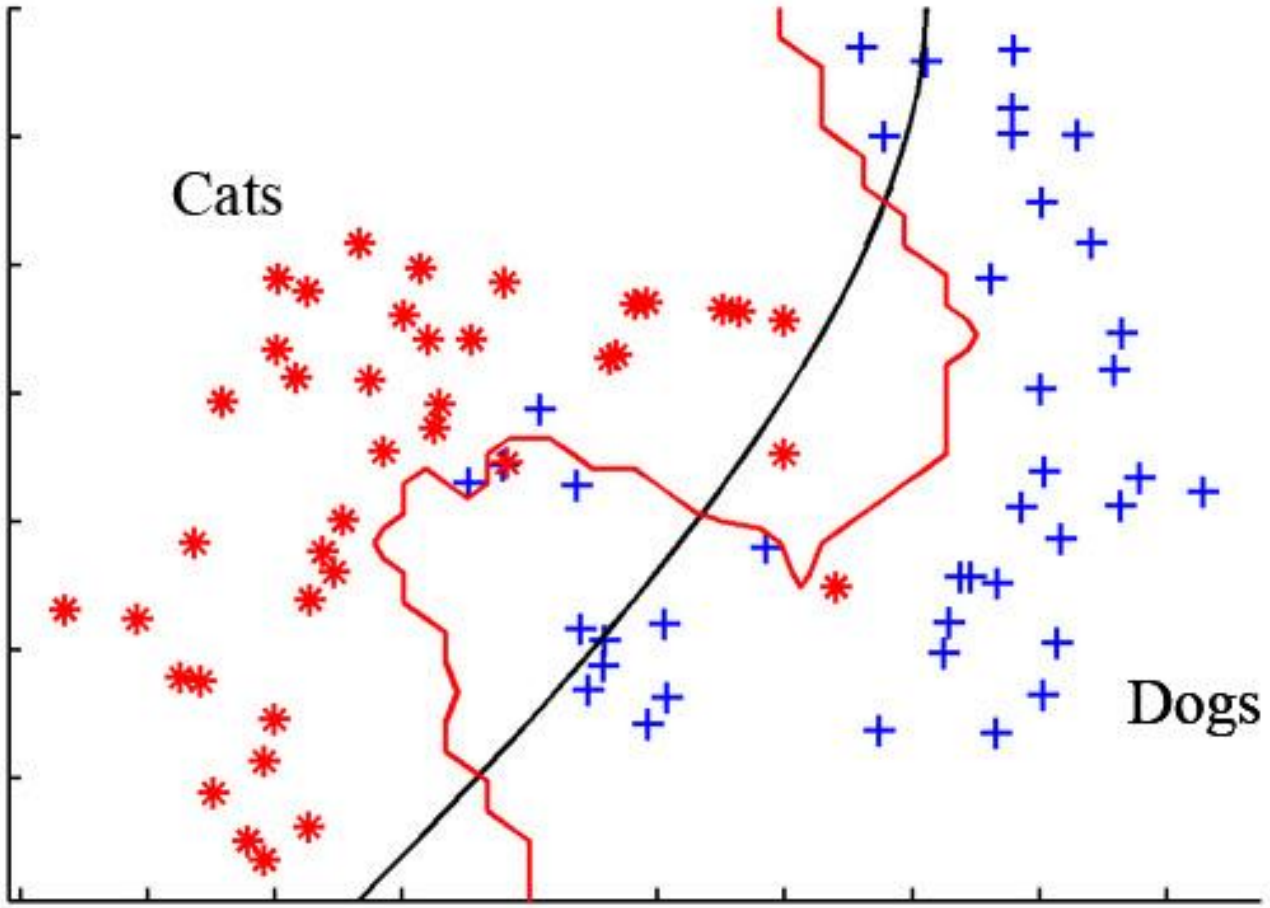
# Training Path



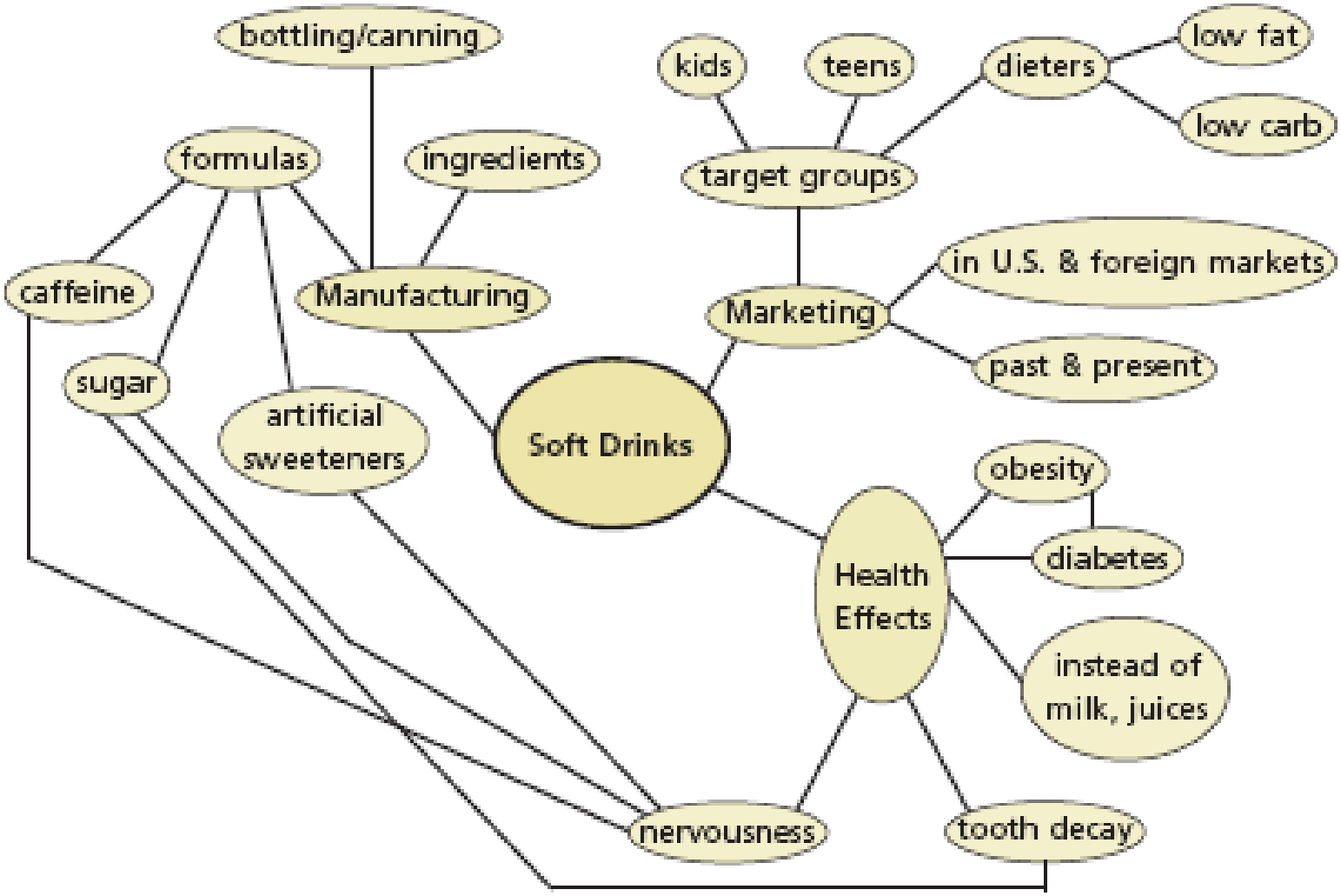

Back

scikit learn

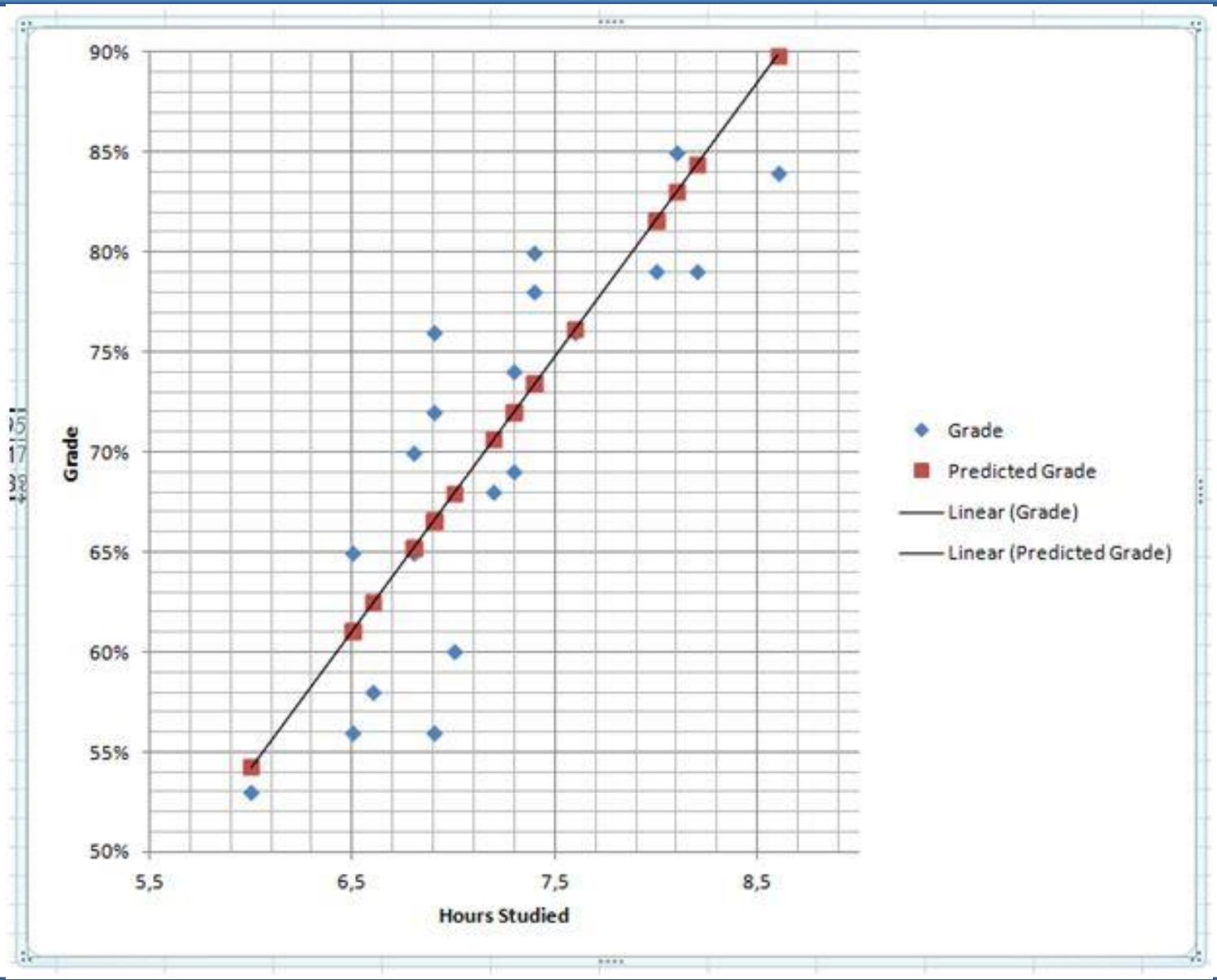
# Classification



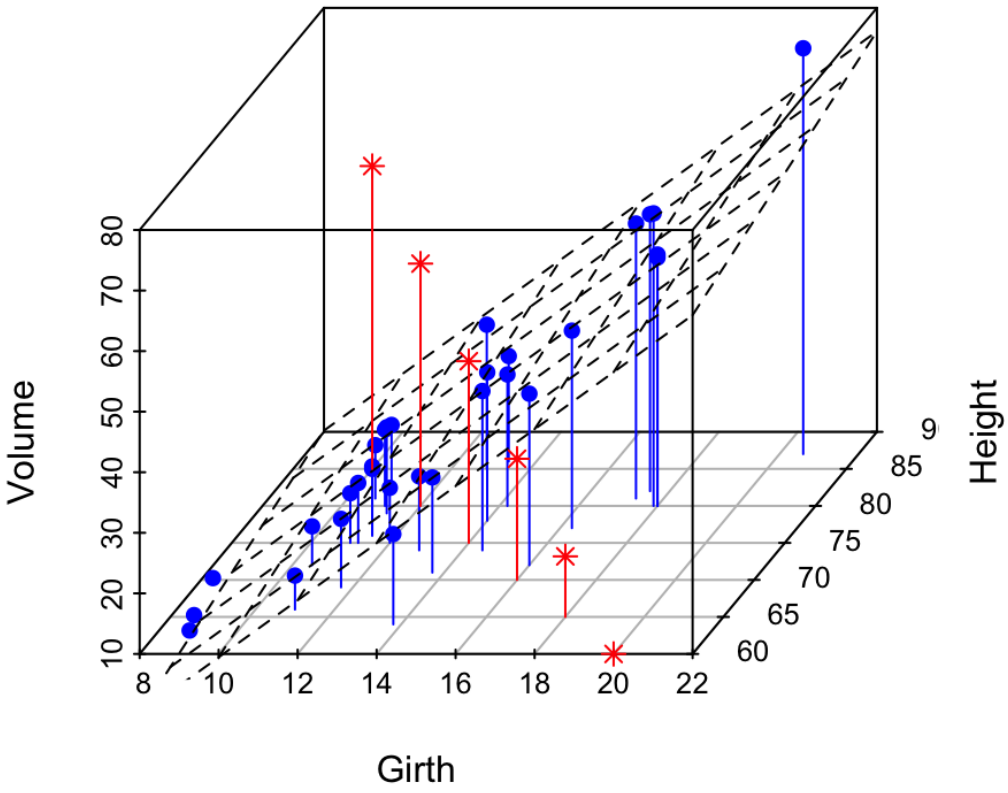
# Clustering


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