
GTK CYBER
GET ON THE FAST TRACK

# Applied Data Science for Cyber Security Professionals

www.gtkcyber.com

Join the Slack Channel at: http://bit.ly/30IvONs


# Course Agenda

**Day 1:  Introduction & Data Engineering**

- Intro: What is Data Science?
    - Overview of Machine Learning & Cyber Applications
- Regular Expressions
- Manipulating and Exploring Data
    - Exploratory Data Analysis in 1 Dimension

**Day 2:  Data Engineering & Data Exploration**

- Exploratory Data Analysis in 2 Dimensions
- Rapid Data Exploration with Apache Drill
- Data Visualization

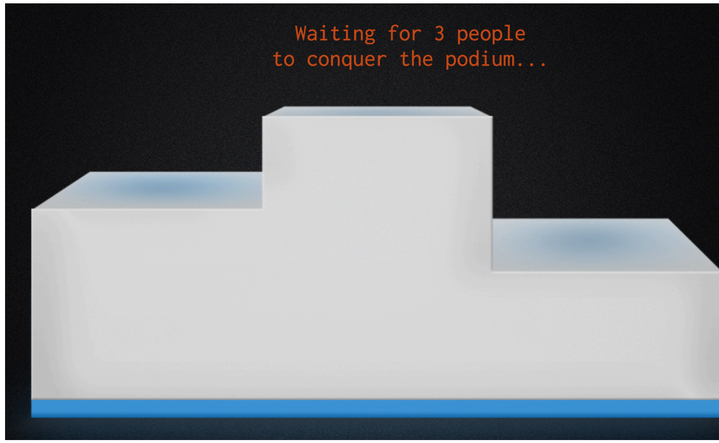
**Day 3:  Machine Learning**

- Feature Engineering
- Supervised Machine Learning
- Unsupervised Machine Learning
- Hacking Machine Learning Models

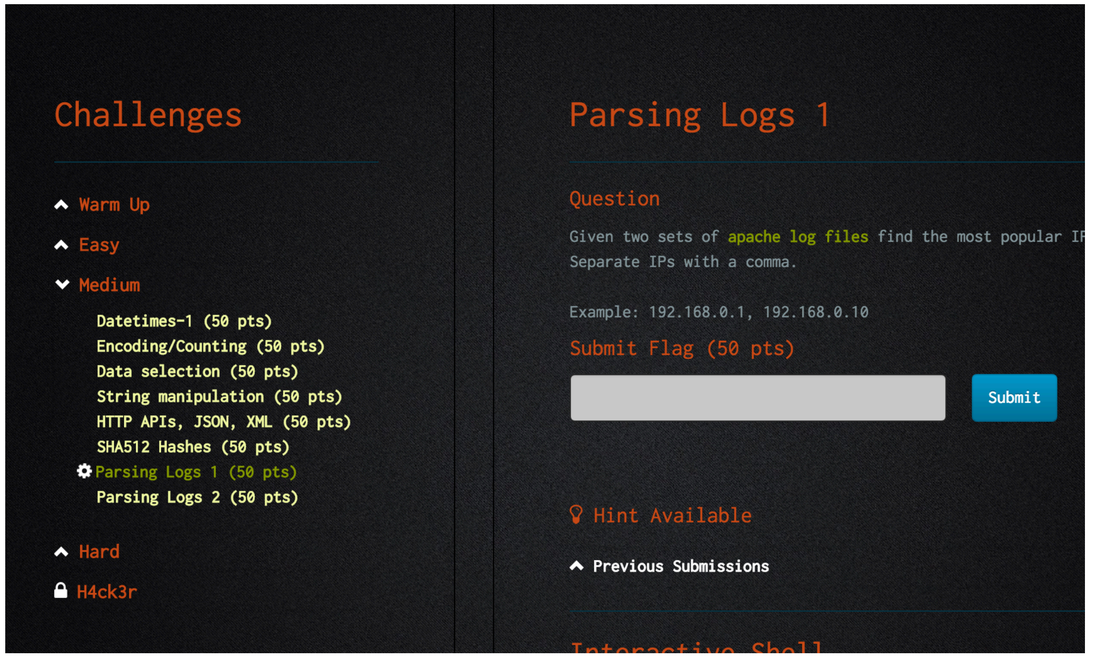
**Day 4: Advanced Topics**

- Introduction to Big Data Tools
- Anomaly Detection
    - PySpark, ELK & Kafka
- Hunting with Data Science

# Beer & Cyber CTF


- www.beerandcyber.com
- username: blackhat
  password: 2019
- Register and start!




# Expectations

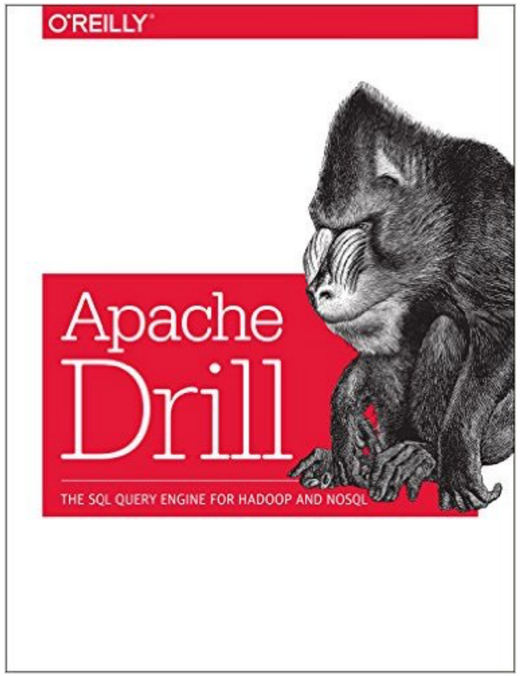
- Please participate and **ask questions.**
- Please follow along and **TRY OUT** the examples yourself during the class
- All the answers are in the slide decks or GitHub repository, but please try to complete the exercises **without looking at the answers**.
- Join the conversation in slack!
- Have fun!


# Introduction

# Our Lawyers Make Us Say This

# Charles Givre, CISSP

- Lead Data Scientist at Deutsche Bank
- PMC Chair for Apache Drill
- Senior Lead Data Scientist @ Booz Allen
- 5 Years @ CIA
- Undergraduate in Comp.Sci & Music




gtkcyber.com

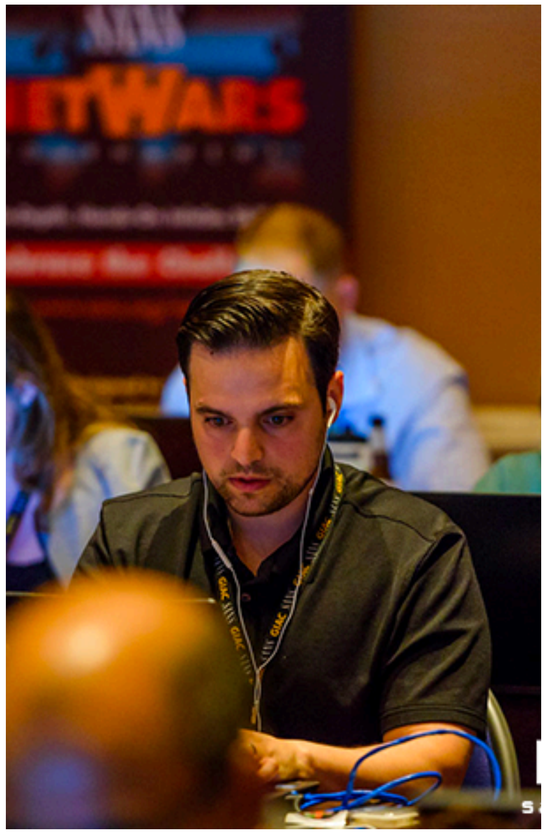
# Austin Taylor

Director of Cybersecurity R&D
@IronNetCyber

Cyber Warfare Operations Officer
@USAF

https://www.github.com/austin-taylor

- VulnWhisperer
- Flare
- Bluewall



**www.austintaylor.io**

**@HuntOperator**


# Who are you?

- Your name (or what you want me to call you)
- Where you work and/or your job role
- What you hope to get out of this class
- Your level of experience with coding


# What is Data Science?


Data Science is the automated extraction of **information** from **raw data**.

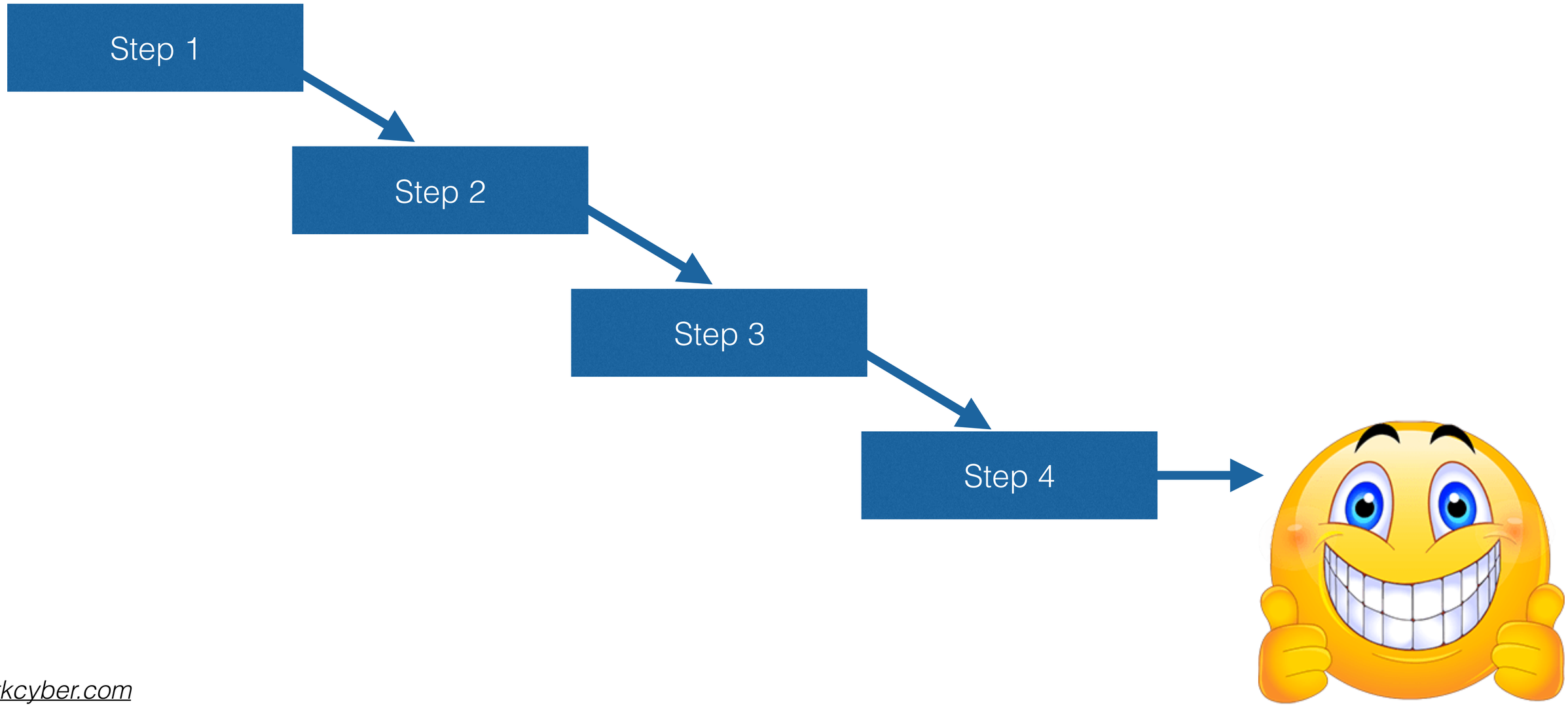

Data Science is the art of **turning data into actions**. This is accomplished through the creation of data products, which provide actionable information without exposing decision makers to the underlying data or analytics

Booz Allen Hamilton, Field Guide to Data Science, Pg. 17

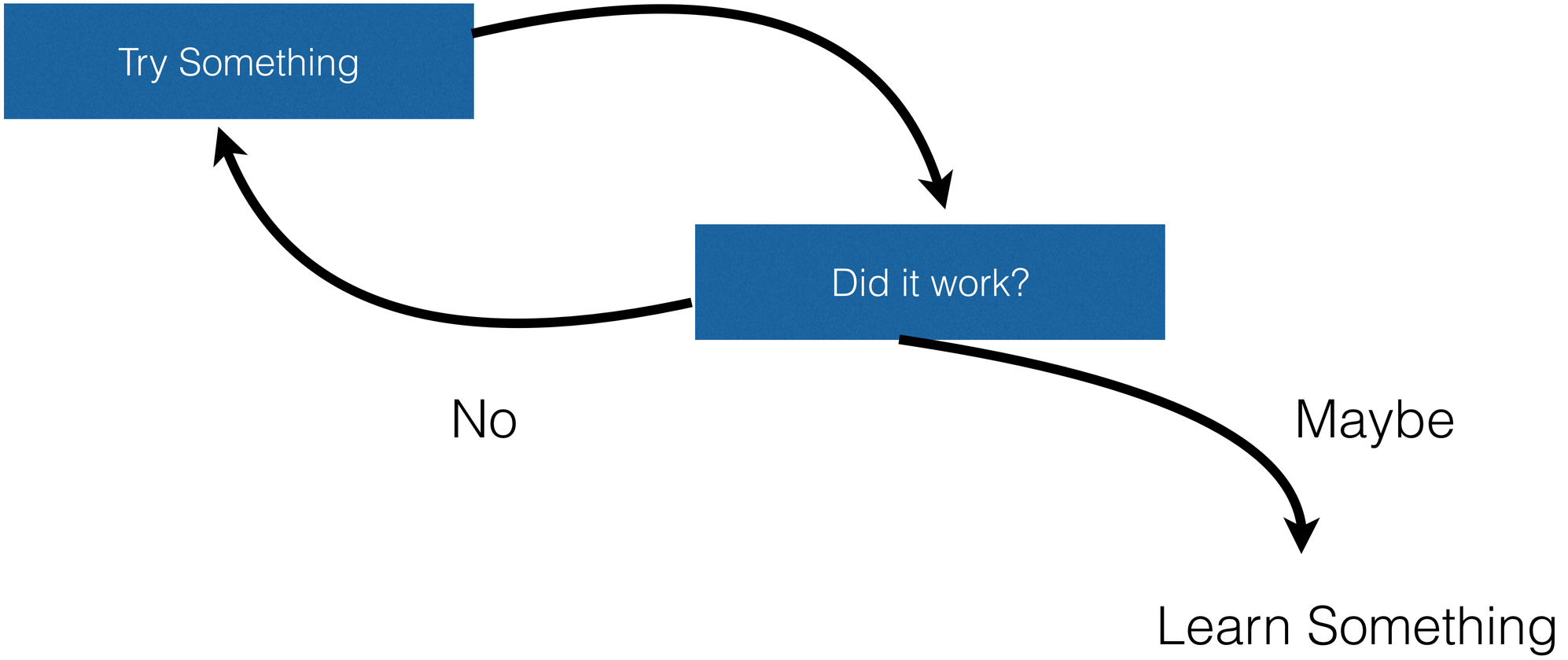

Analyst ⟷ Developer


# Analyst + Developer

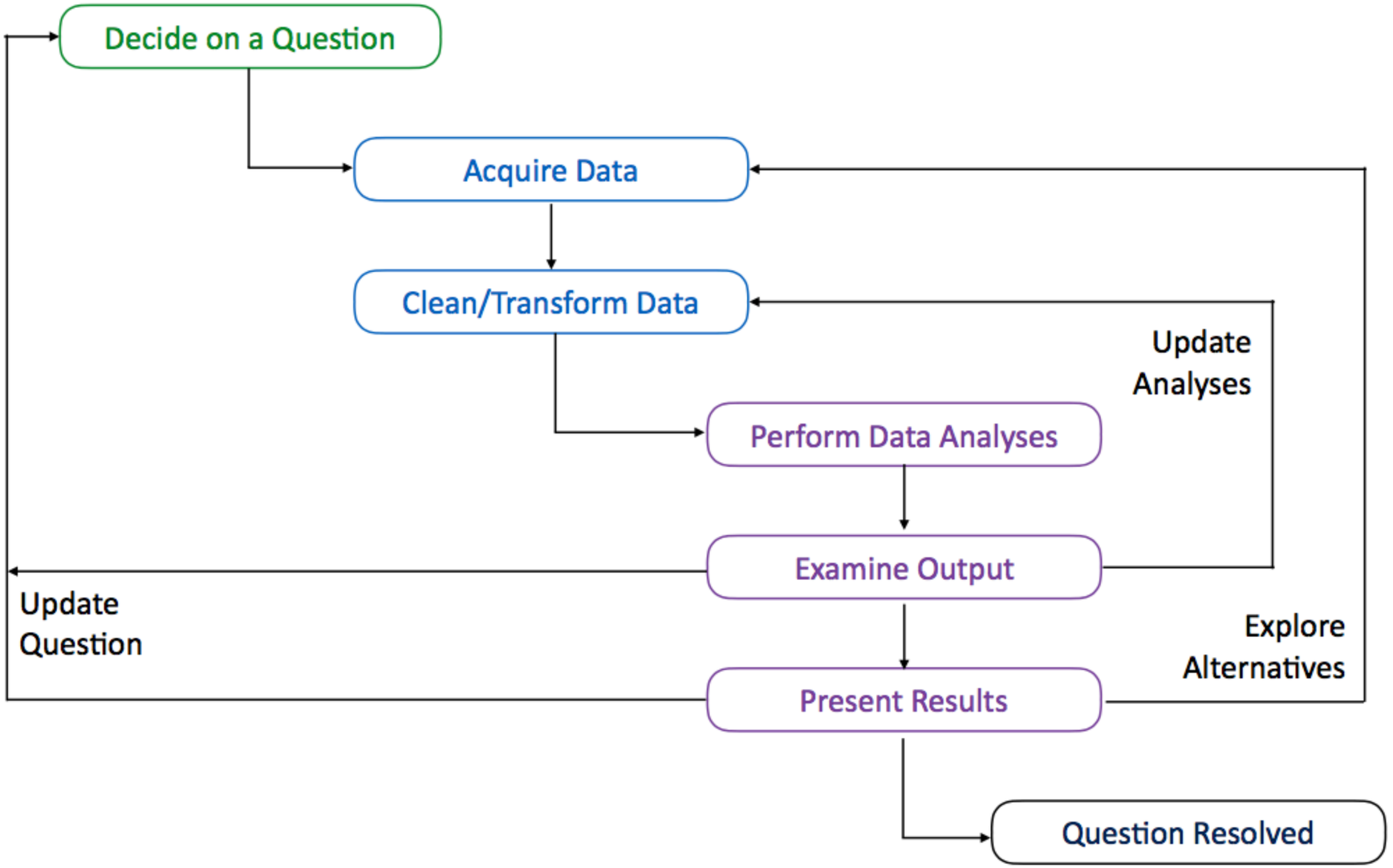
# What Data Science is Not



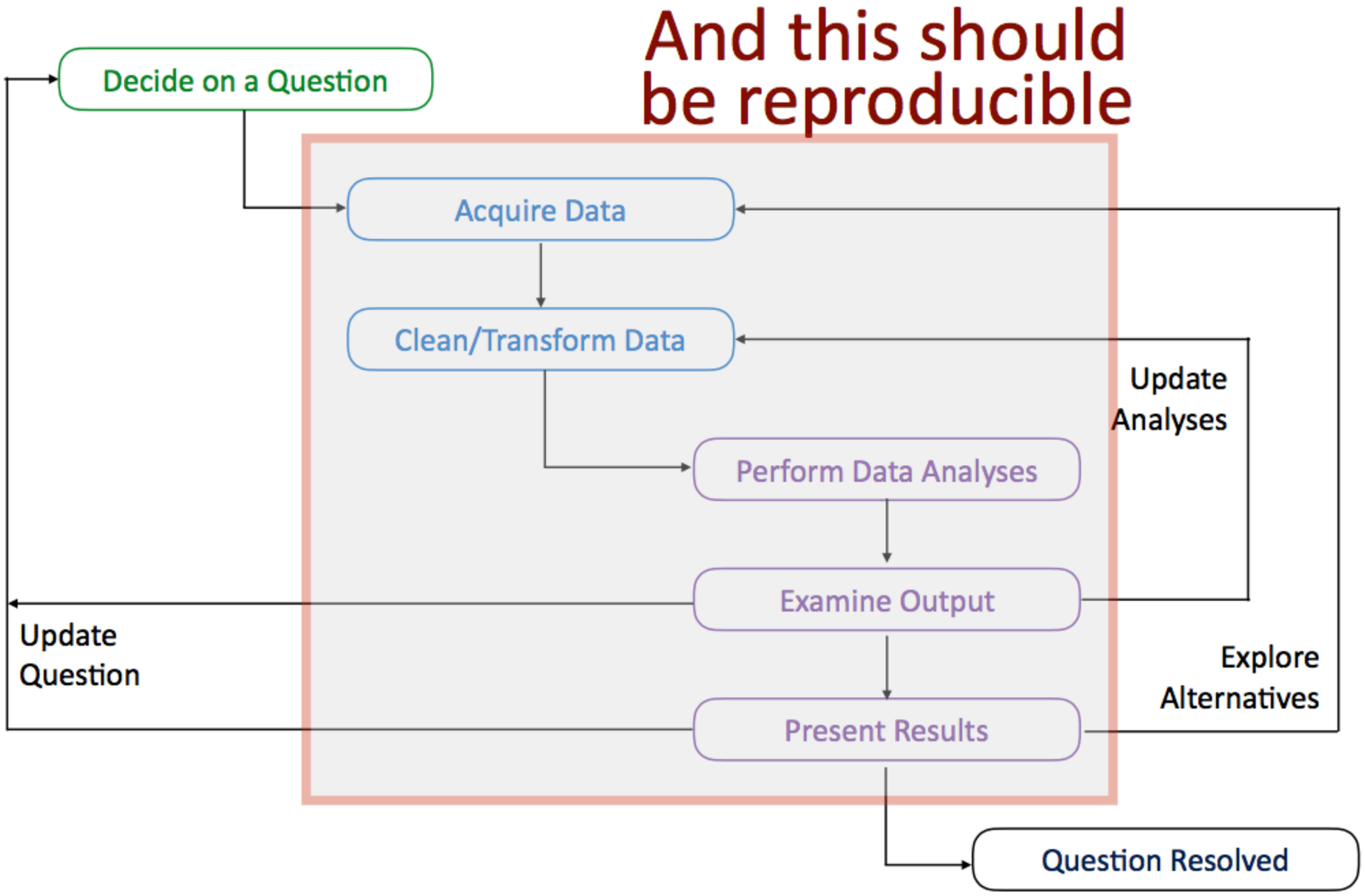
# What Data Science is



# Research Process



# Research Process




"The term "data scientist" will subside and may well sound dated five years from now. **The skills will become more commonplace and commoditized. When that happens, the real boom will begin**, because the technology will become widely adopted and thus more useful. … **Instead, we need self-service tools that empower smart and tenacious business people to perform Big Data analysis themselves.**

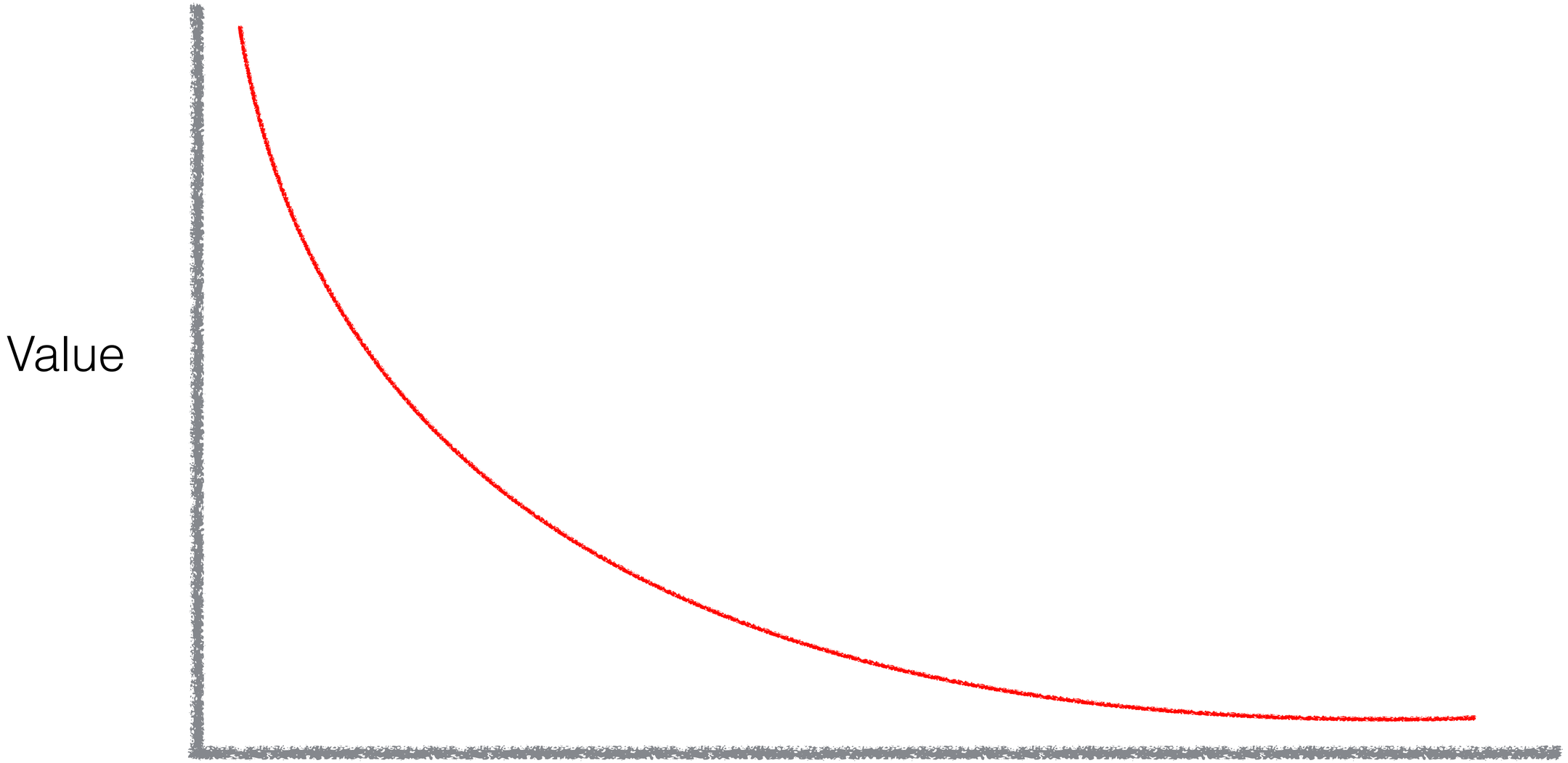
*–Andrew Brust, "Data scientists don't scale", http://www.zdnet.com/article/data-scientists-dont-scale/*


# Time to Insight

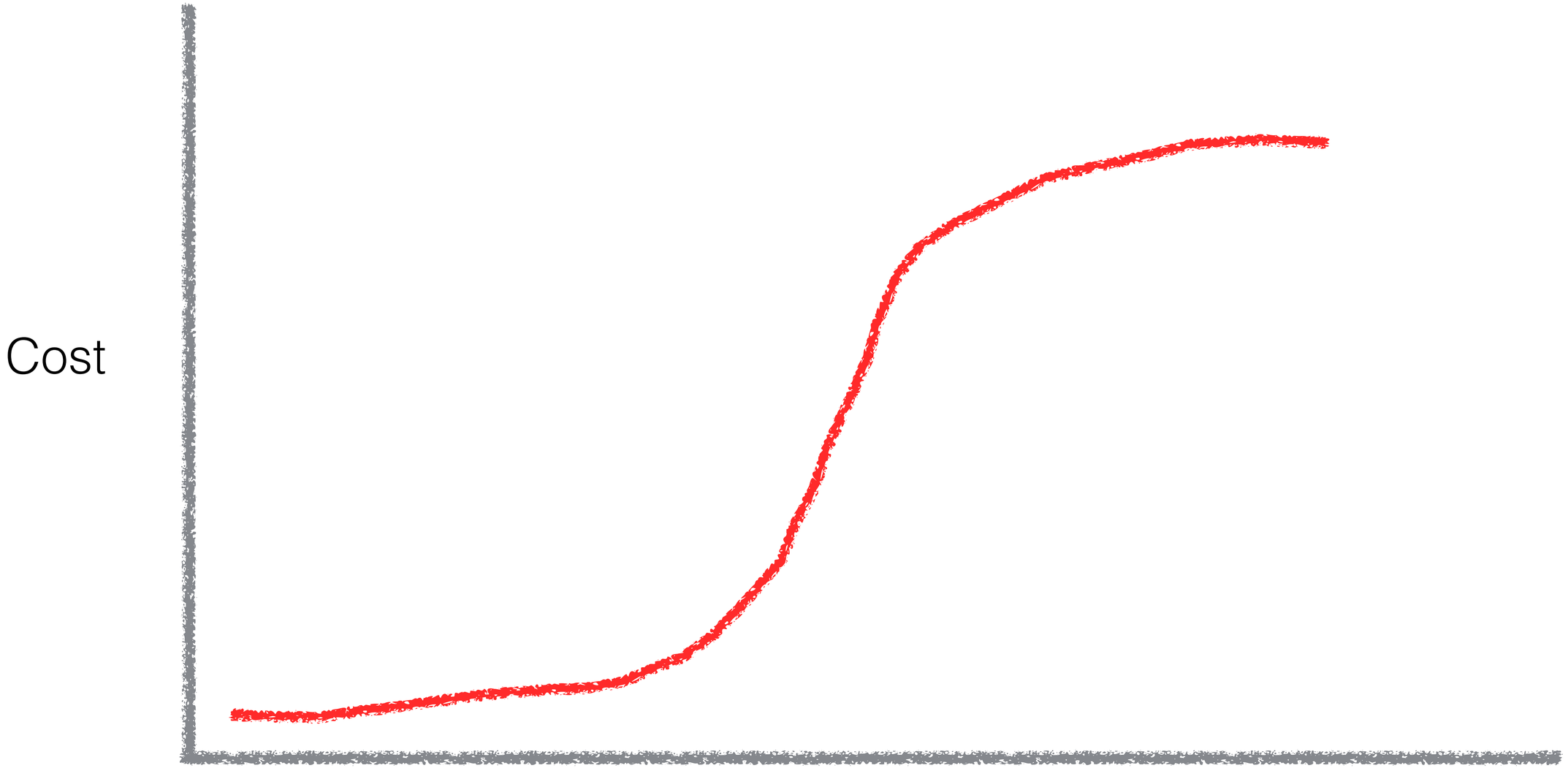

# Time to Insight

# Time = $$

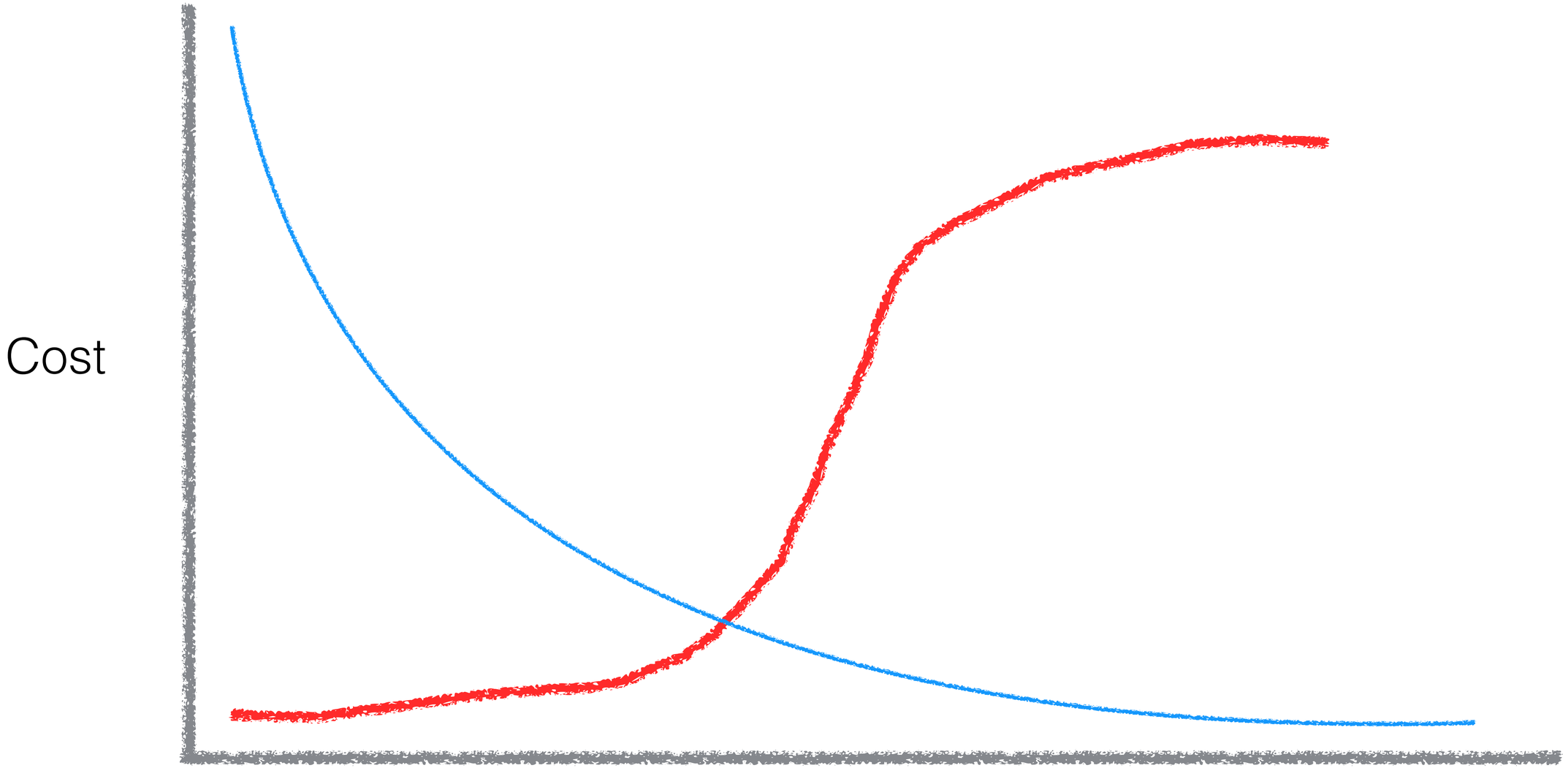
# Value of Insights over Time



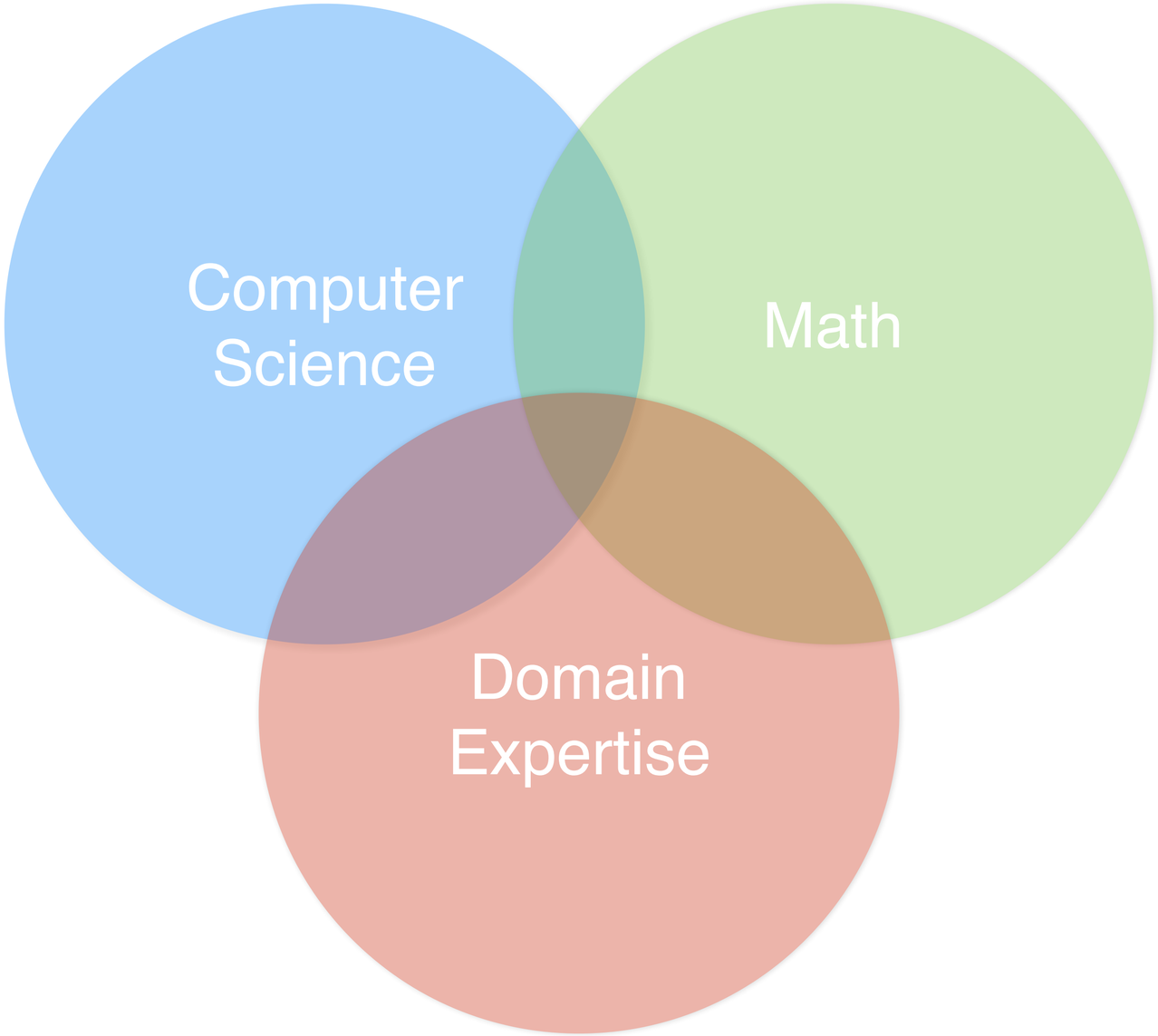
# Cost of Insights over Time



# Cost of Insights over Time

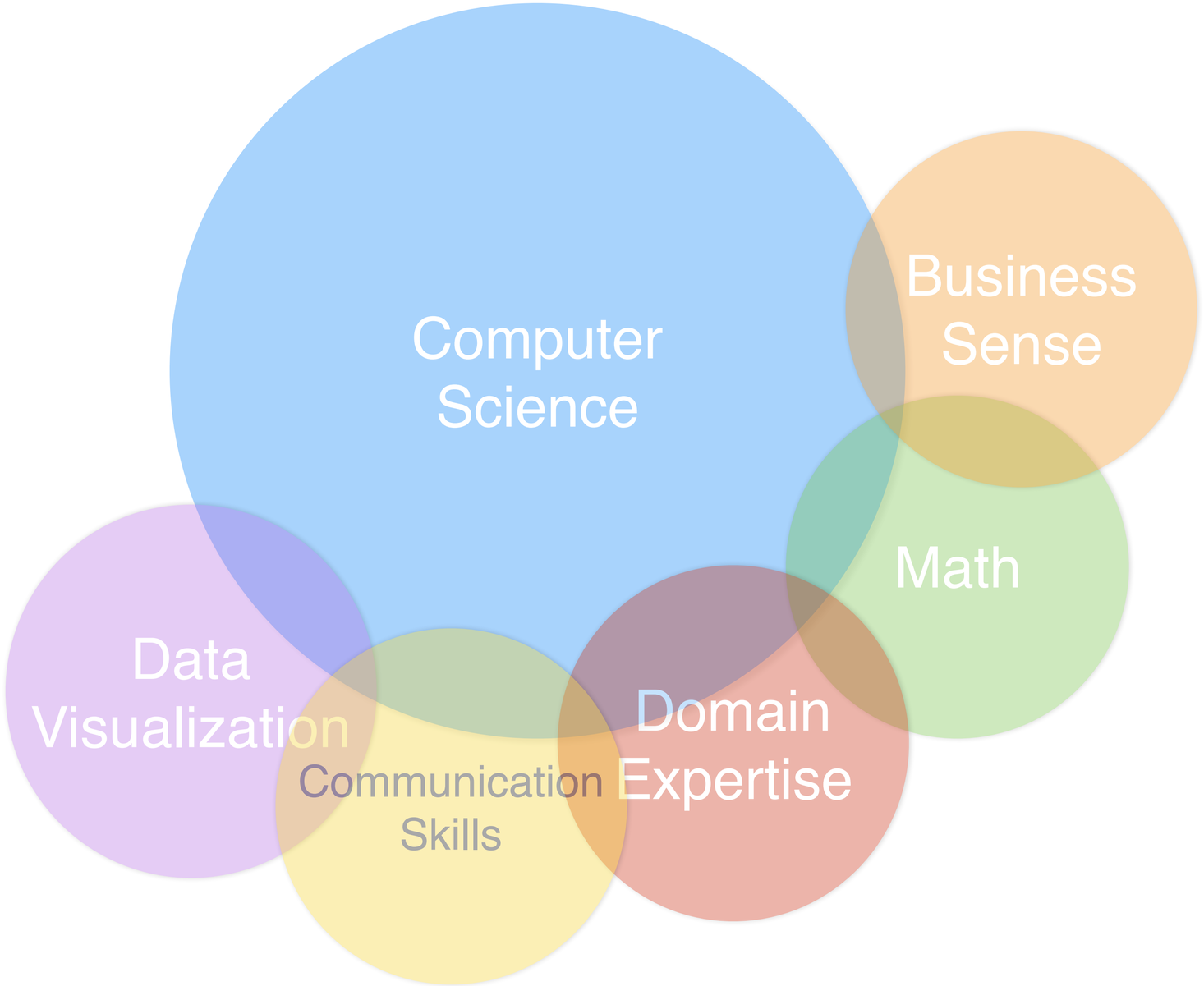



# What skills does a data scientist need?


Computer
Science
Math
Domain
Expertise

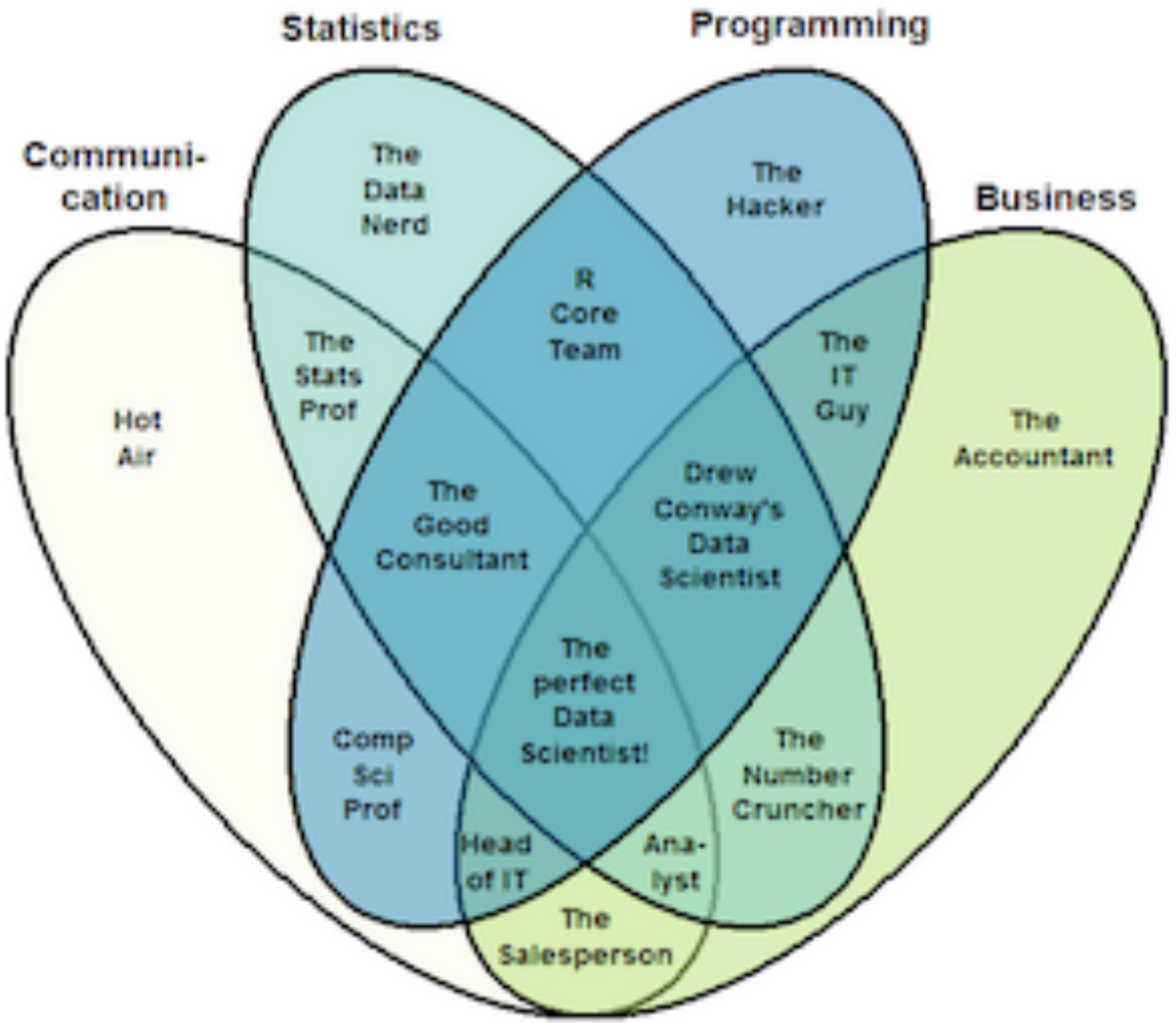
GTK CYBER

GET ON THE FAST TRACK



GTK CYBER
GET ON THE FAST TRACK

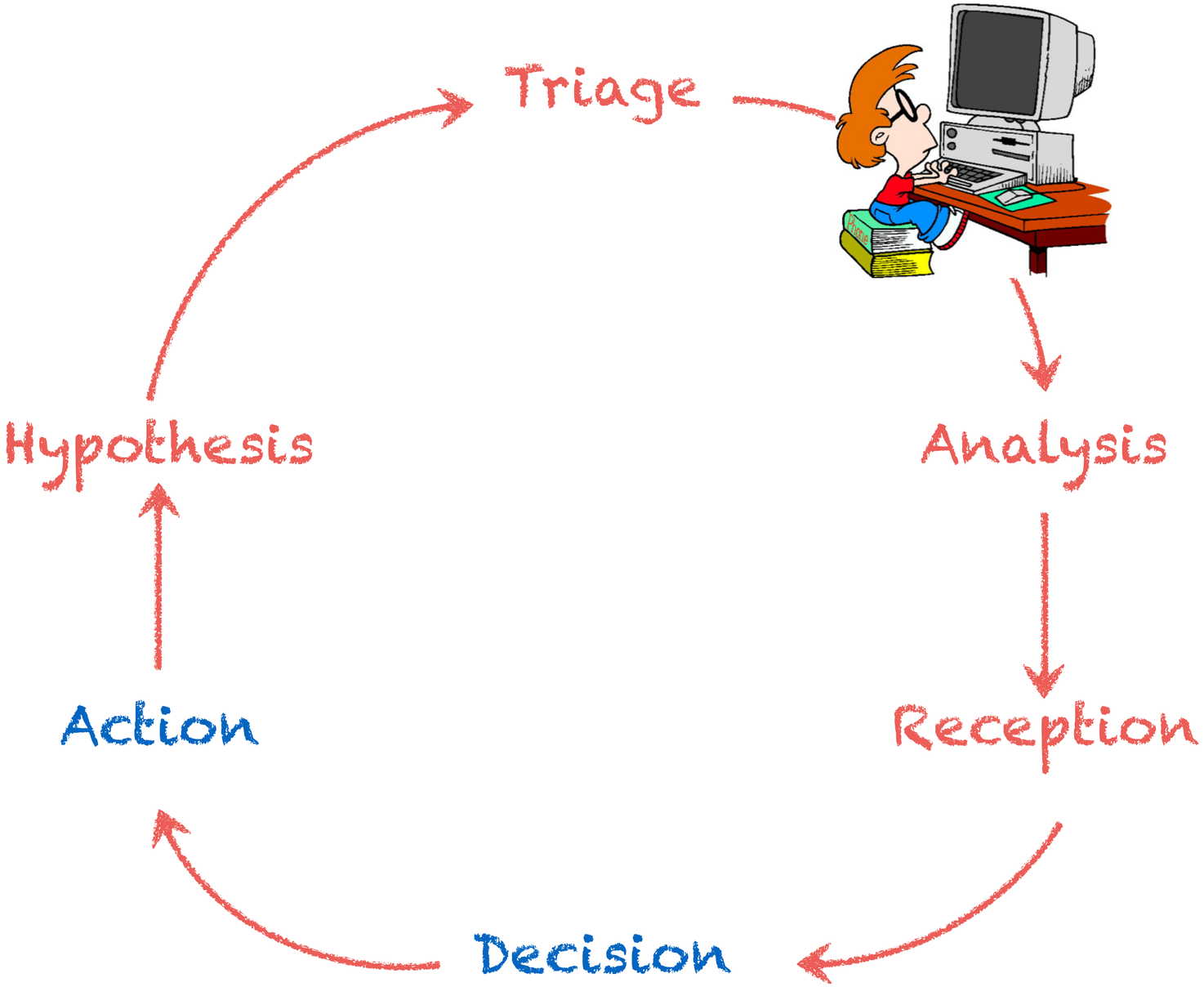
# The Data Scientist Venn Diagram




https://www.kdnuggets.com/2017/03/data-science-data-scientist-do.html


# Data Scientists spend 50-90% of their time being…


CAUTION
WET FLOOR


Triage
Analysis
Reception
Decision
Action
Hypothesis


# Thoughts for Data Science Success

# Data is a Strategic Asset... not a cost

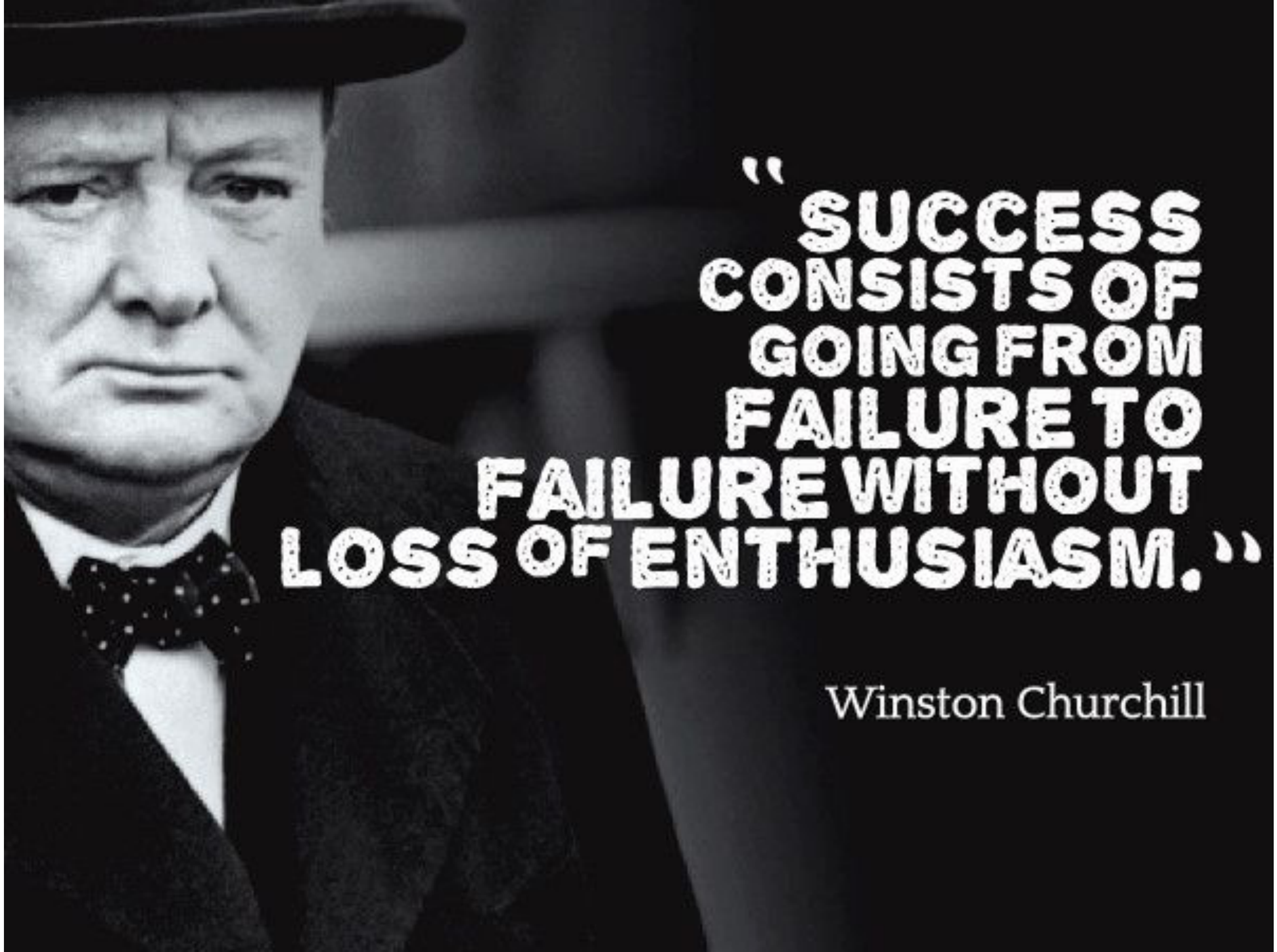
# Align Projects to Corporate Strategy

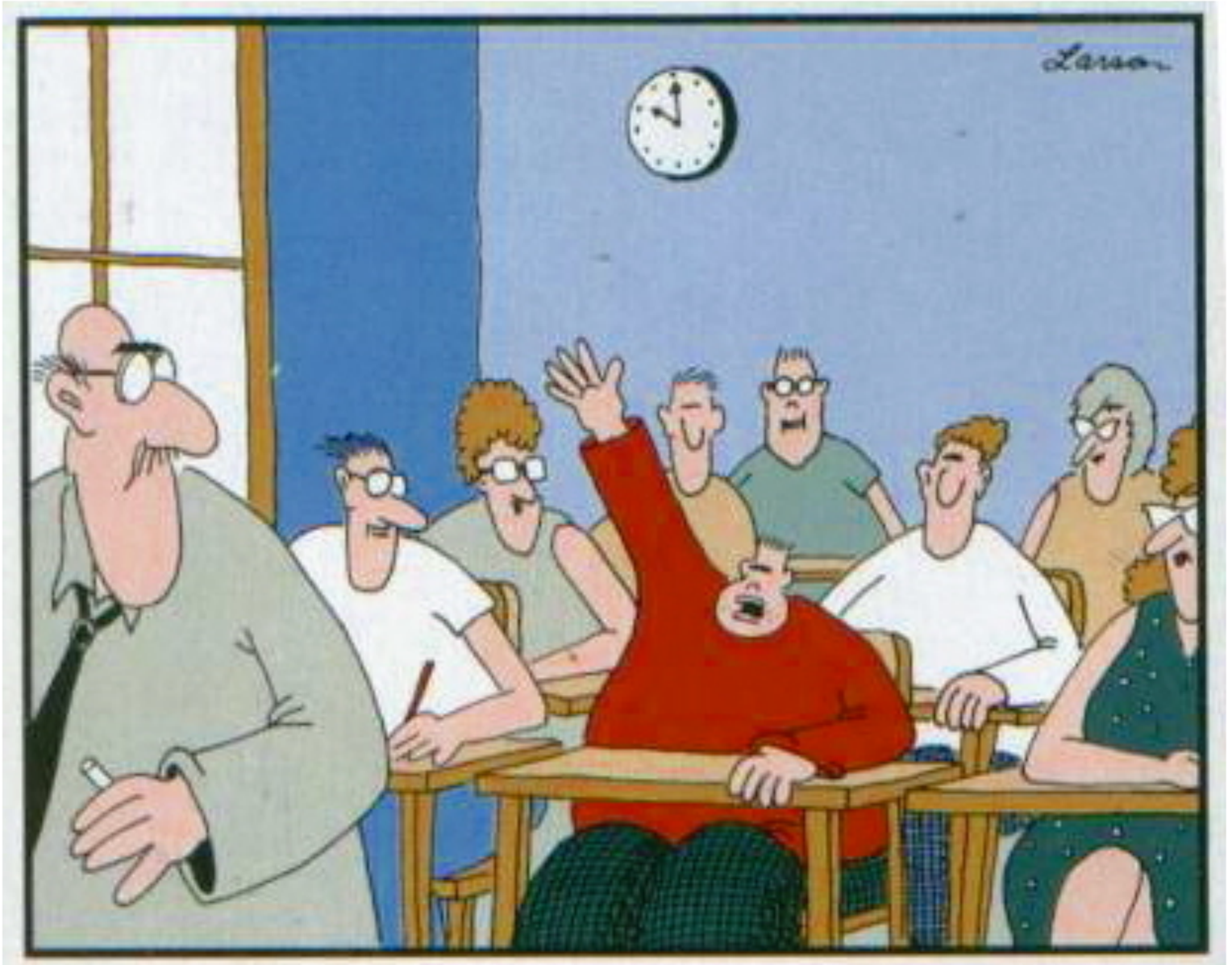
# Align Projects to Corporate Strategy

Your:
Time
Money
Job?

# Build the right team for Data Initiatives

# Prioritize building appropriate data platform


"SUCCESS CONSISTS OF GOING FROM FAILURE TO FAILURE WITHOUT LOSS OF ENTHUSIASM."
Winston Churchill




"Mr. Osborne, may I be excused?
My brain is full."


# Building a Data Science Team


# Building a Data Science Team

Programmer + Statistician + SME = DS Team?


# Building a Data Science Team

Programmer + Statistician + SME = DS Team?

Sort of…

# Building a Data Science Team

- Data Ambassador
- Data Scientist
- Data Storyteller
- Data Engineer

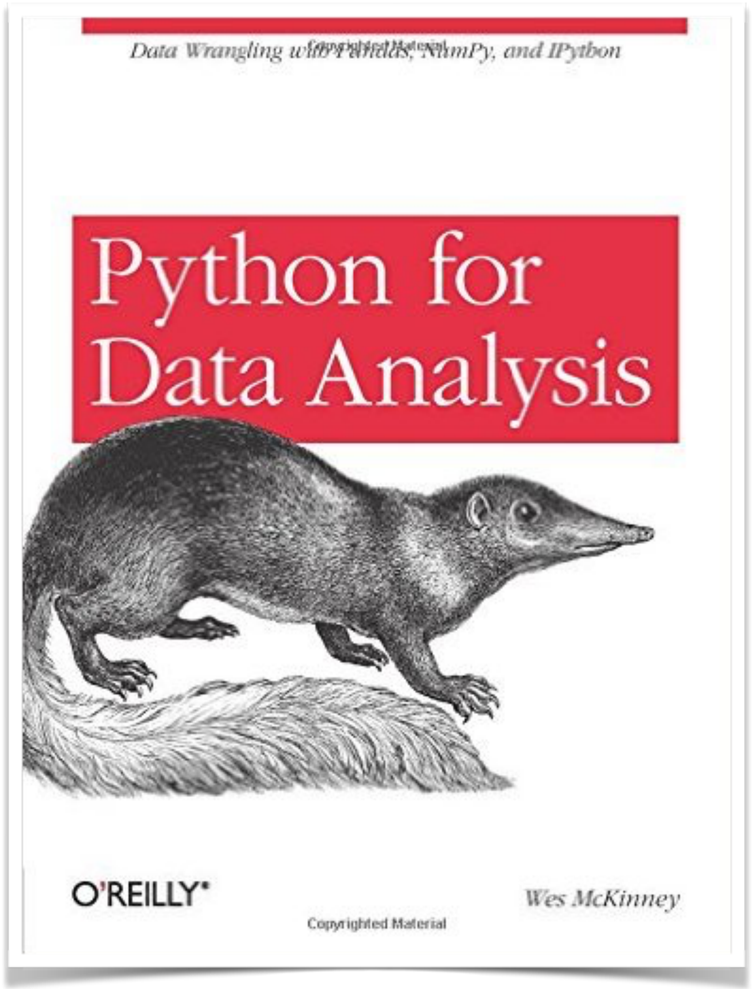

# By the end of the class you will be able to:

- Quickly and effectively prepare data for analysis
- Apply machine learning techniques to enhance security

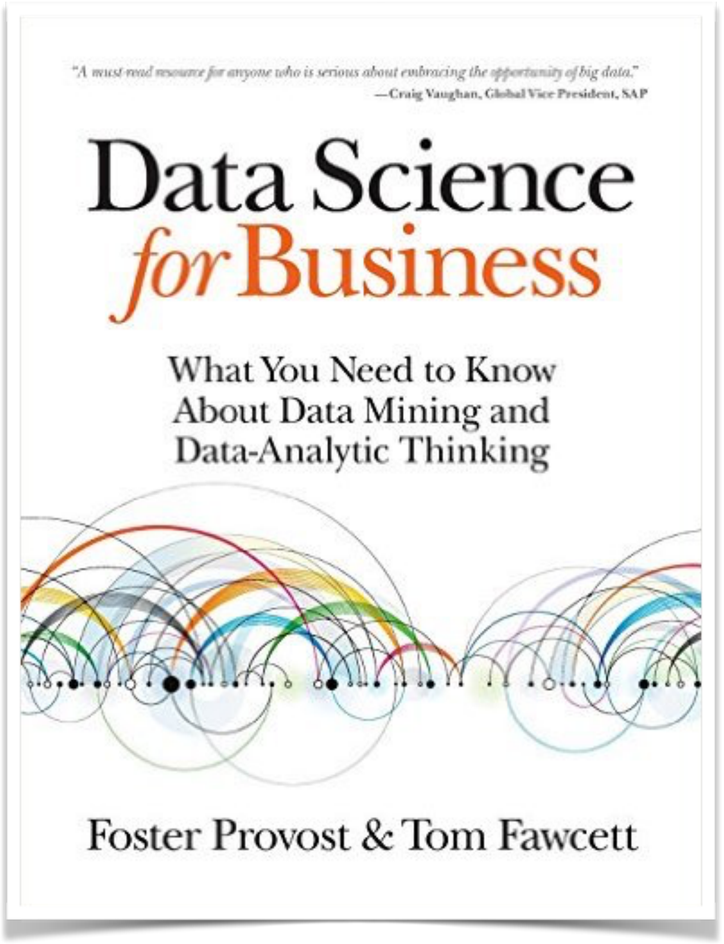

I THINK WE SHOULD BUILD AN SQL DATABASE.
UH-OH
DOES HE UNDERSTAND WHAT HE SAID OR IS IT SOMETHING HE SAW IN A TRADE MAGAZINE AD?
WHAT COLOR DO YOU WANT THAT DATABASE?
I THINK MAUVE HAS THE MOST RAM.
E-mail: SCOTTADAMS@AOL.COM
© 1996 United Feature Syndicate, Inc.(NYC)


Buzzword

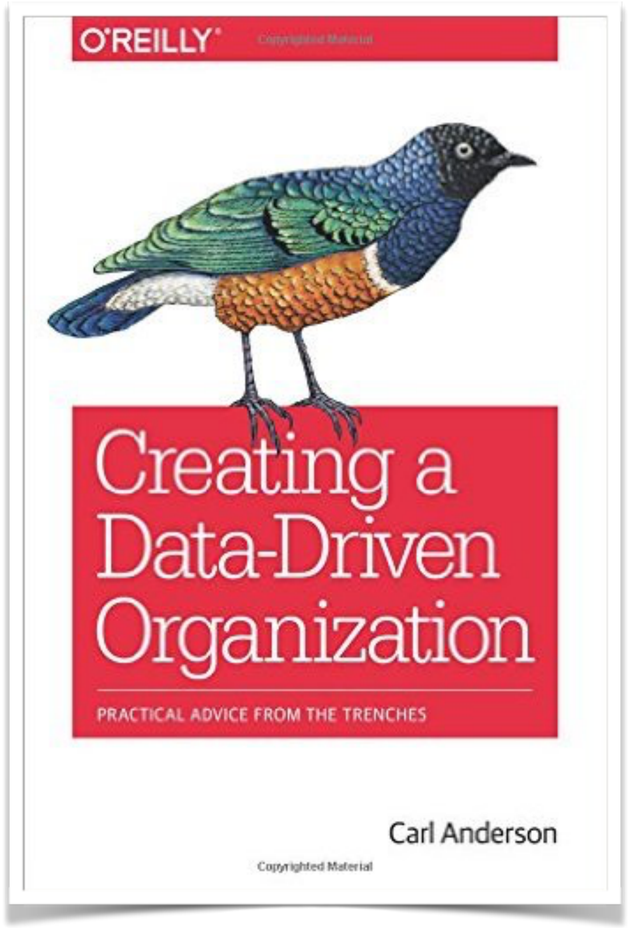
# Recommended Reading



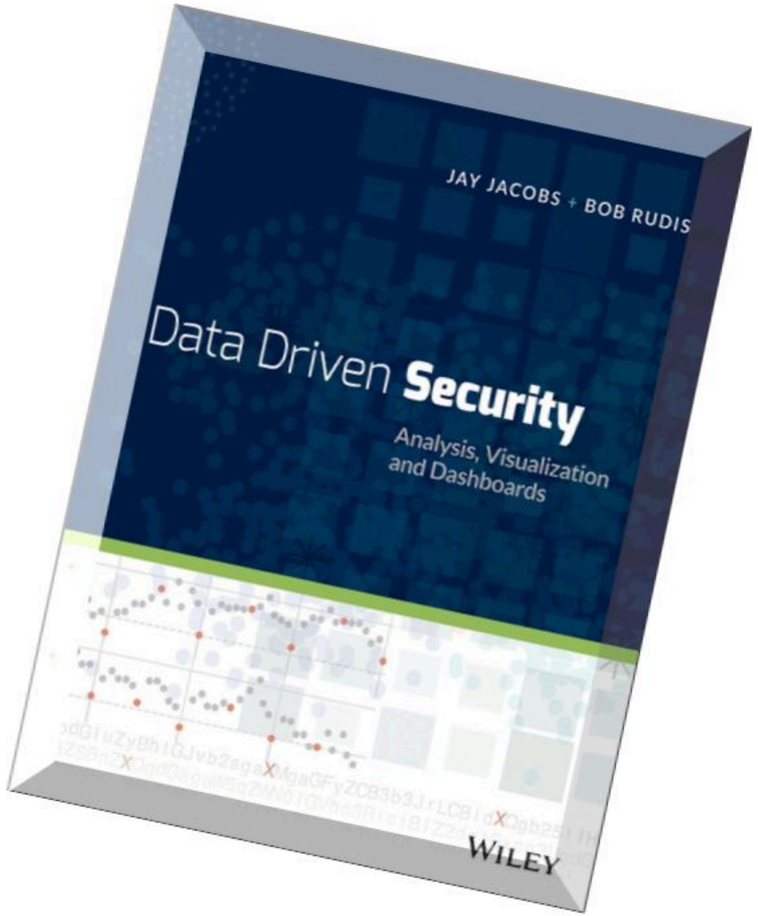
# Recommended Reading



# Recommended Reading

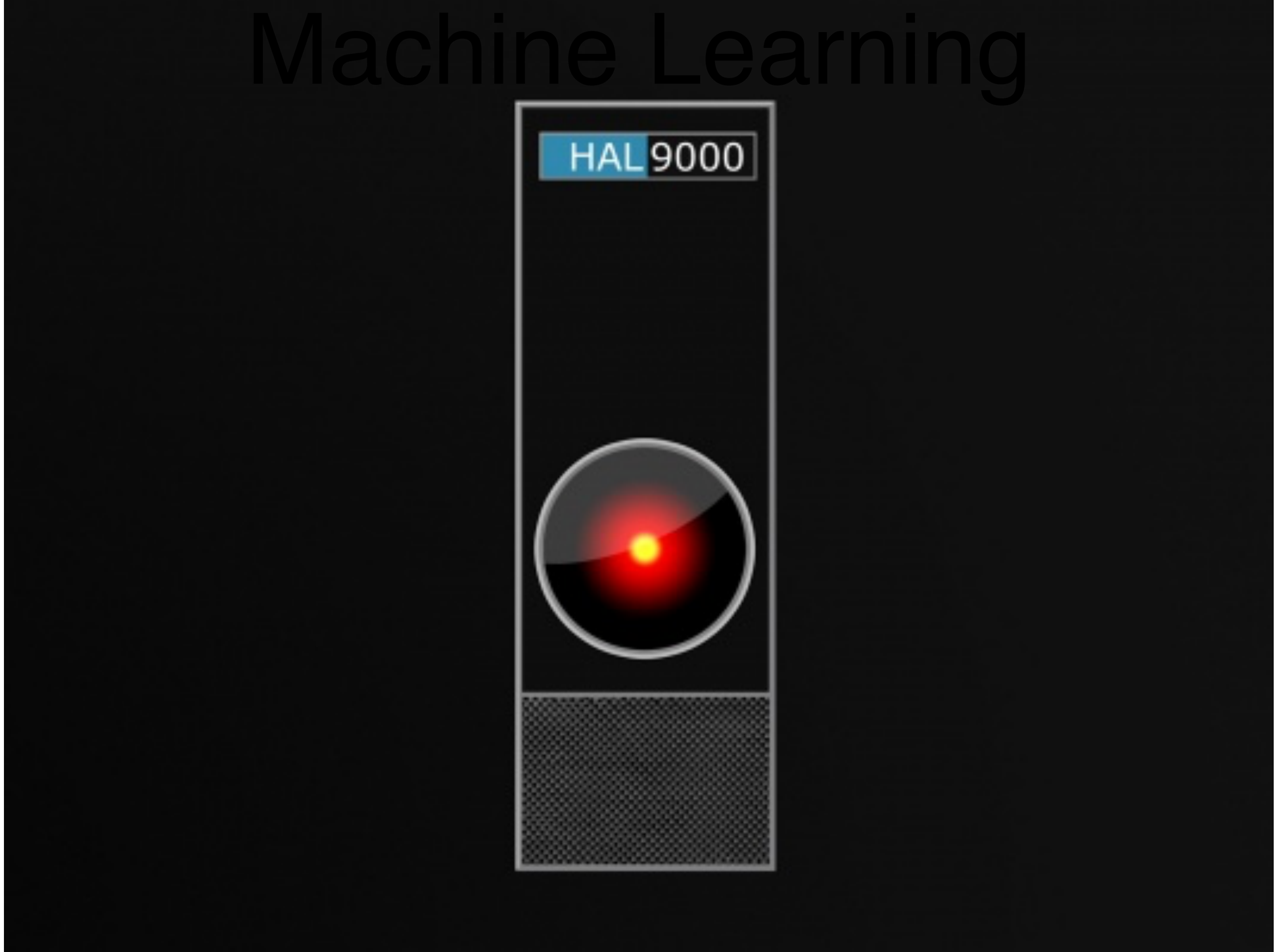


# Recommended Reading

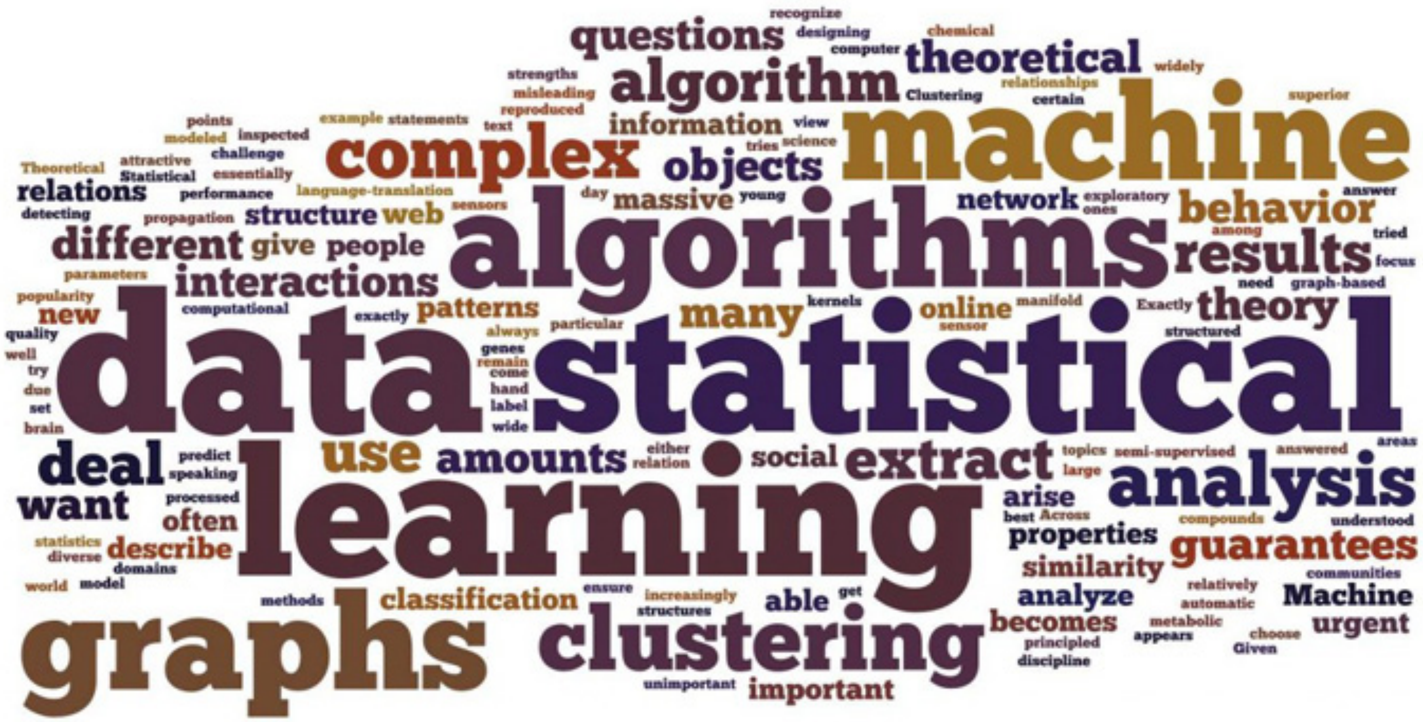



http://datadrivensecurity.info

# Machine Learning



# What is
# Machine Learning (ML)
# Artificial Intelligence (AI)

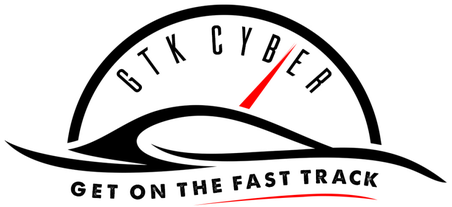



“I have no idea.  The latest buzzword? ”

–The guy in the back of the class


“Machine learning is the science of getting computers to **act without being explicitly programmed.** ”

– *https://www.coursera.org/course/ml*




"A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

*–Tom Mitchell, Carnegie Mellon University*


"Machine learning explores the construction and study of algorithms that can learn from and **make predictions on data.** Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, **rather than following strictly static program instructions.**"

*–https://en.wikipedia.org/wiki/Machine_learning*

GTK CYBER
GET ON THE FAST TRACK


SPAM®
CHOPPED PORK AND HAM
LICENCED BY:
Hormel Foods
Serving Suggestion


SPAM
TERIYAKI
Glorious SPAM!


SPAM
LITE
Glorious SPAM!


SPAM
Oven Roasted TURKEY
Hormel Foods


- Blacklists
- Simple keyword matching
- Naive Bayesian Classifiers
- Deep Learning

## Artificial Intelligence

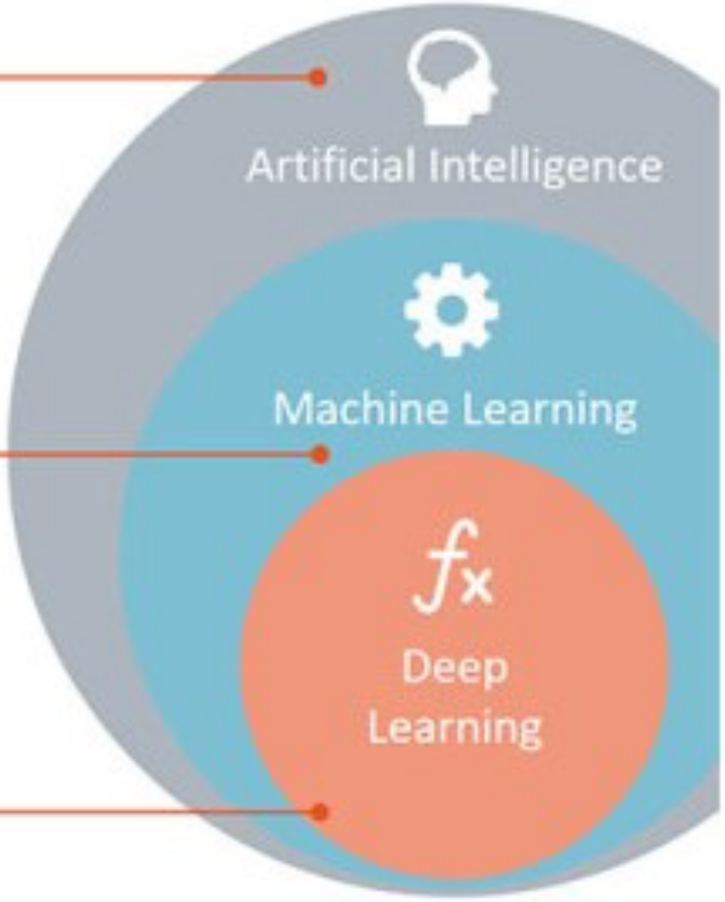
Any technique which enables computers to mimic human behavior.

## Machine Learning

Subset of AI techniques which use statistical methods to enable machines to improve with experiences.

## Deep Learning

Subset of ML which make the computation of multi-layer neural networks feasible.






https://www.kdnuggets.com/2017/07/rapidminer-ai-machine-learning-deep-learning.html

@katherinebailey Because marketing? Every time someone calls simple linear regression “AI” Gauss turns over in his grave.


# Machine Learning Problems

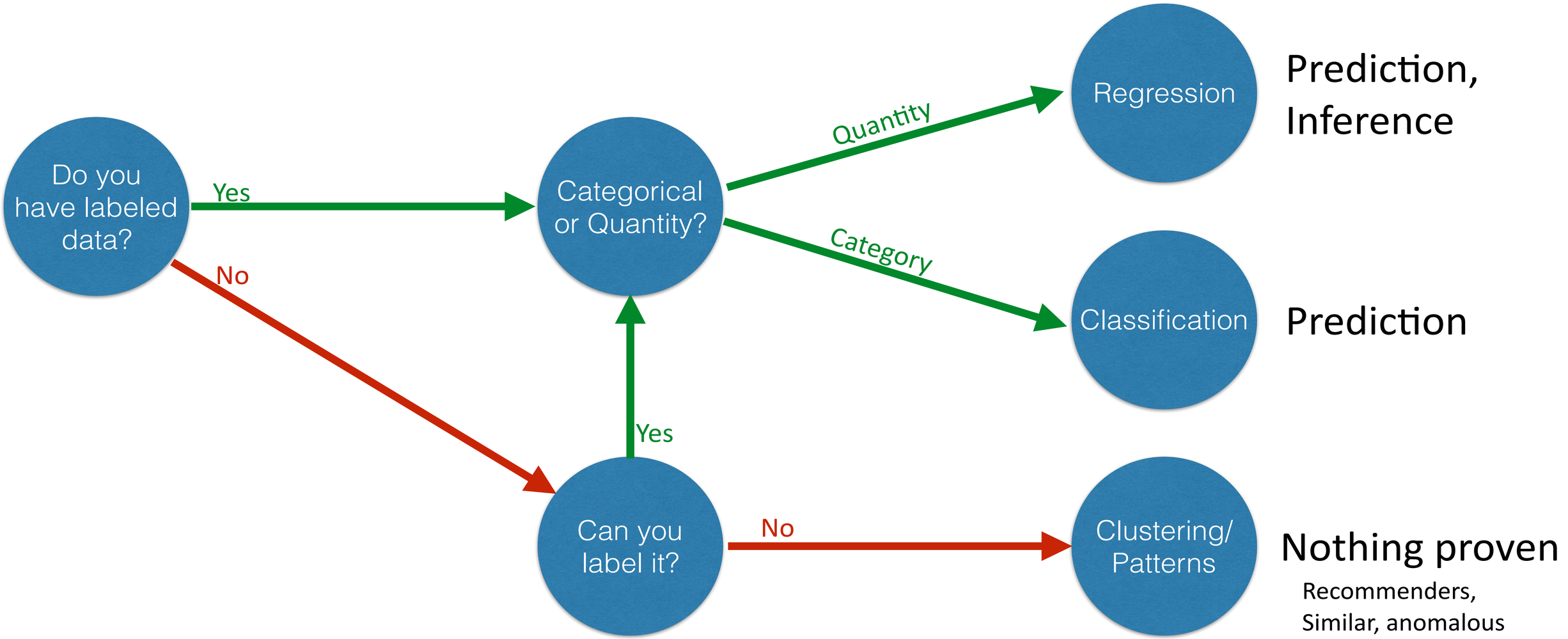
- **Supervised Learning:** Supervised Learning is a class of Machine Learning in which a model is "trained" using a set of pre-existing labeled data.

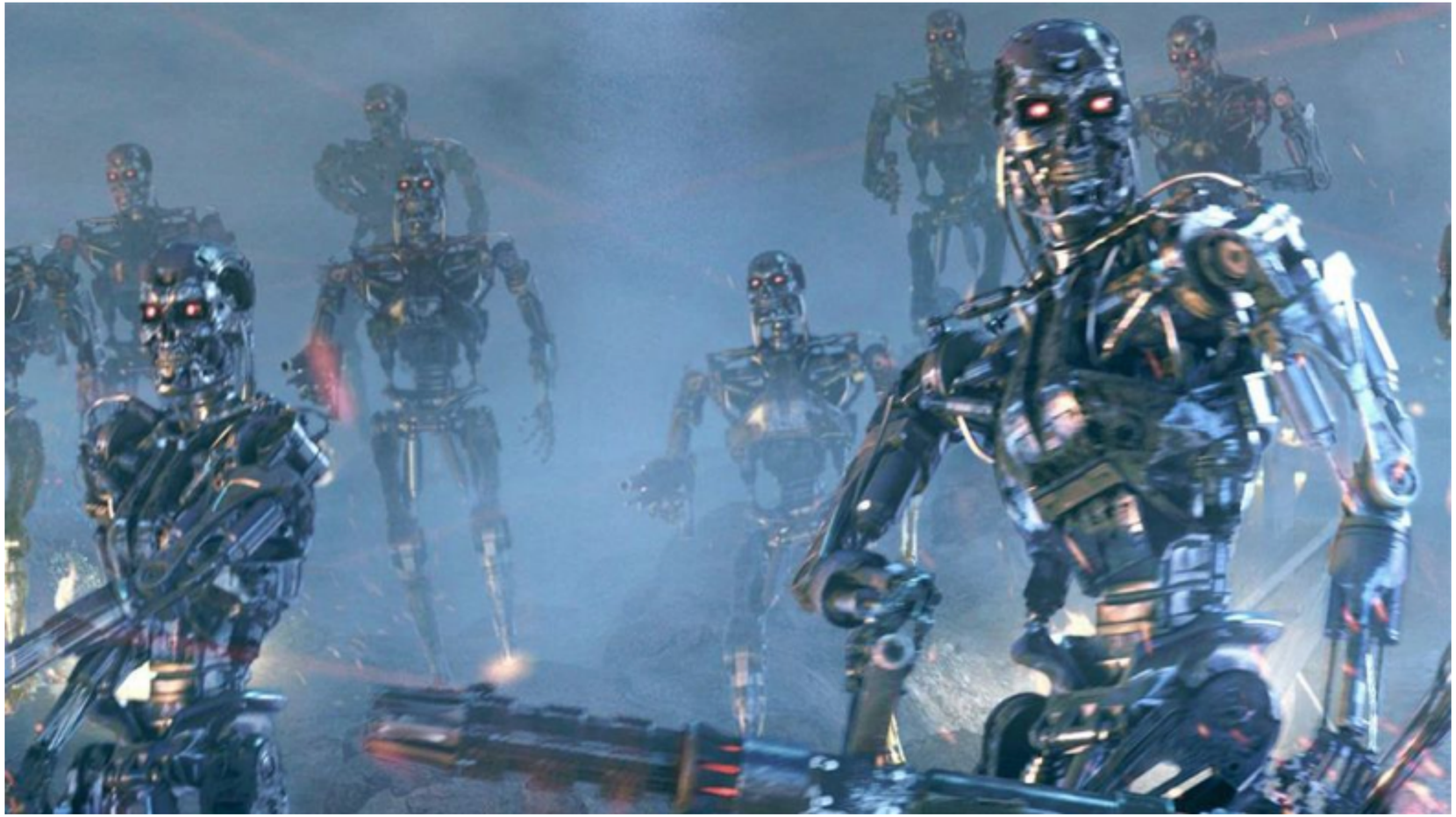
- **Unsupervised Learning:** A class of Machine Learning algorithms in which a model is built without the use of labeled data.


# Machine Learning Problem Types

- **Classification**: Assigning or predicting a observation's membership in discrete class
- **Regression:** Predicting a continuous value based on the observations' features
- **Clustering:** Identifying groupings within a dataset
- **Dimensionality Reduction:** Reducing the number of variables in a feature set

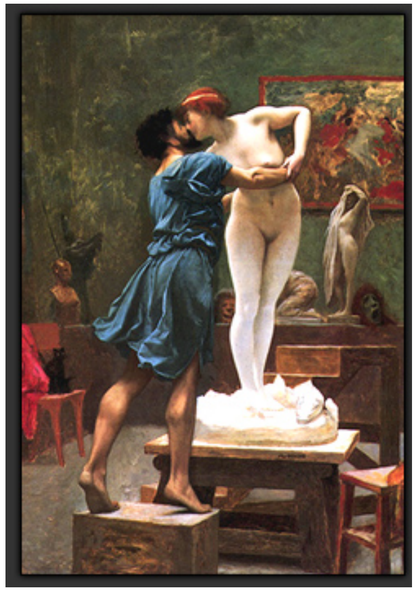
# What Problem am I solving?



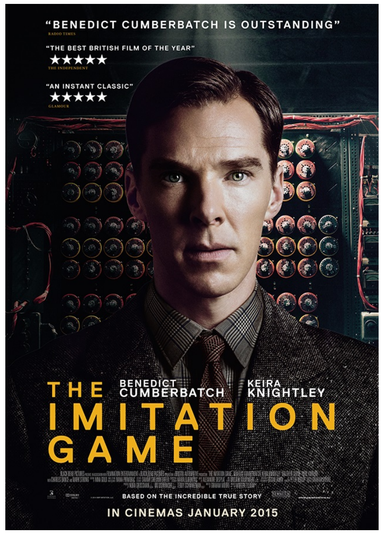


# What it is not

# History of AI and ML



- Mythical figure Pygmalion: Breath of life in a statue (ancient Greece)
- Turing Test 1950s by Alan Turing: Test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.
- Early AI: solve problems that are difficult for humans, but straightforward for computers (e.g. formal mathematical rules) IBM Deep Blue chess-playing system defeated world champion Garry Kasparov 1997.
- Real Challenge: Intuitive, easy problems for humans. They are hard to describe formally (e.g recognizing faces, words).



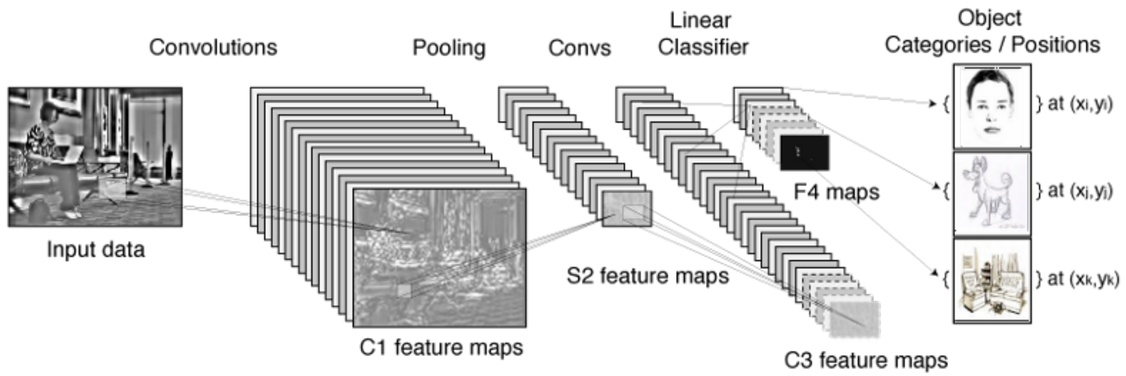




http://www.deeplearningbook.org/

# Path to Artificial Intelligence?

- Hard-coding knowledge failed (e.g. Cyc 1989).
- **Machine Learning: Gathering knowledge from experience** eliminates the need to formally specify all knowledge the computer needs.
- **AI** needs to acquire own knowledge by **extracting pattern from raw data.**
- Success depends heavily on **representation of data, aka features.**






http://www.deeplearningbook.org/


# Applications to Security

# Regression Example

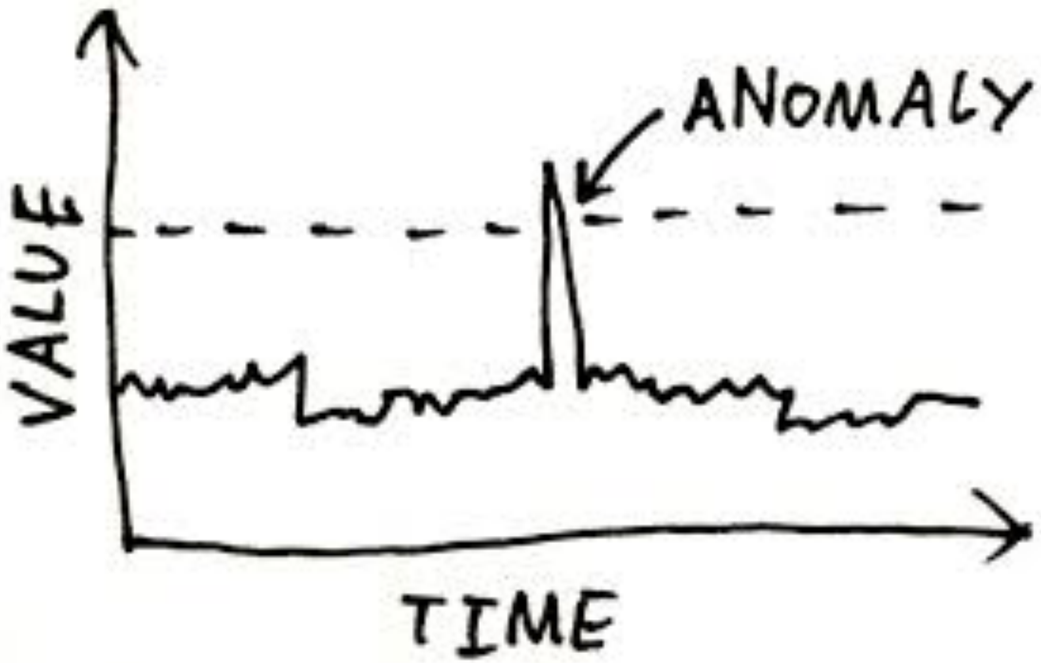
**Server Capacity Prediction:** Regression analysis can be used to predict a server's capacity (or CPU usage) based on the server's historical performance.

*https://www.researchgate.net/publication/256645877_LiRCUP_Linear_Regression_based_CPU_Usage_Prediction_Algorithm_for_Live_Migration_of_Virtual_Machines_in_Data_Centers*

*https://jgreenemi.com/predicting-capacity-with-linear-regression-ml/*

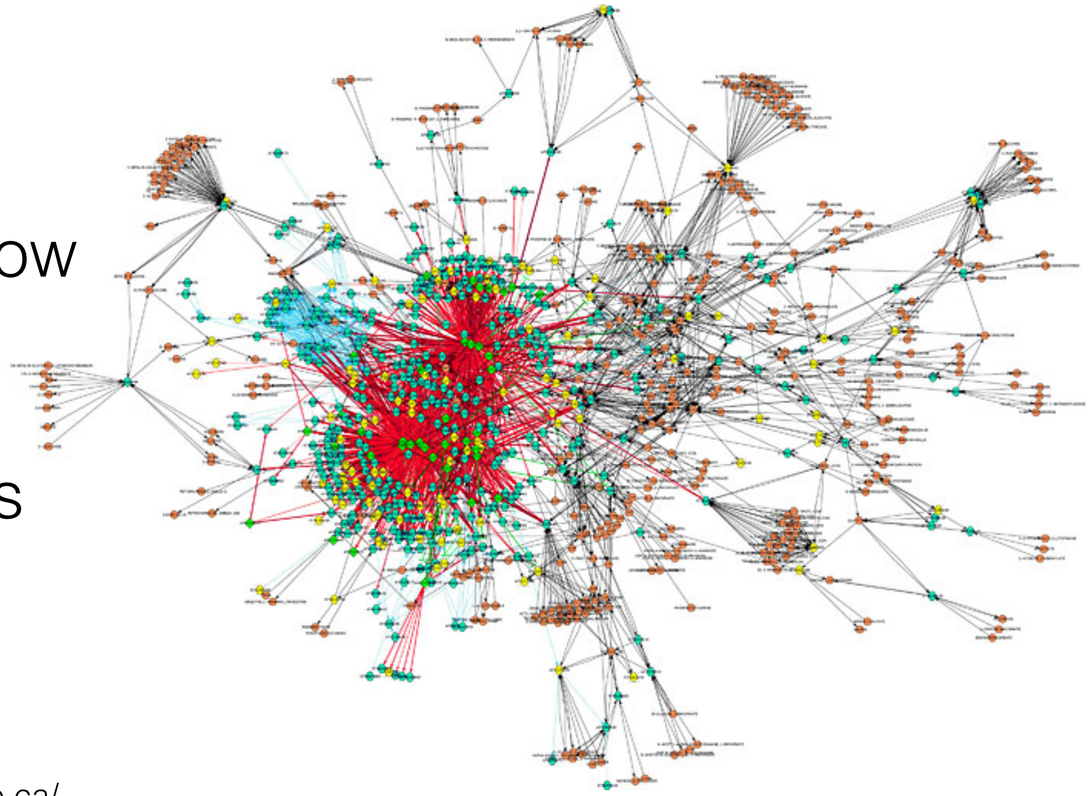
# Clustering Example

**Anomaly Detection:** Clustering techniques can be used to detect anomalous traffic or loads or anything really.



# Network-Based Intrusion Detection

- Derive Features from Network Traffic Captures “pcap” at packet level or NetFlow level (tools: tshark, tcpdump, bro…)
- Example Features based on header information: 2s-windowing of connections > duration, protocol, src and dat bytes, service.
- Get data sets: http://www.netresec.com/?page=PcapFiles, https://maccdc.org/, http://www.westpoint.edu/crc/SitePages/DataSets.aspx http://www.unb.ca/cic/research/datasets/,



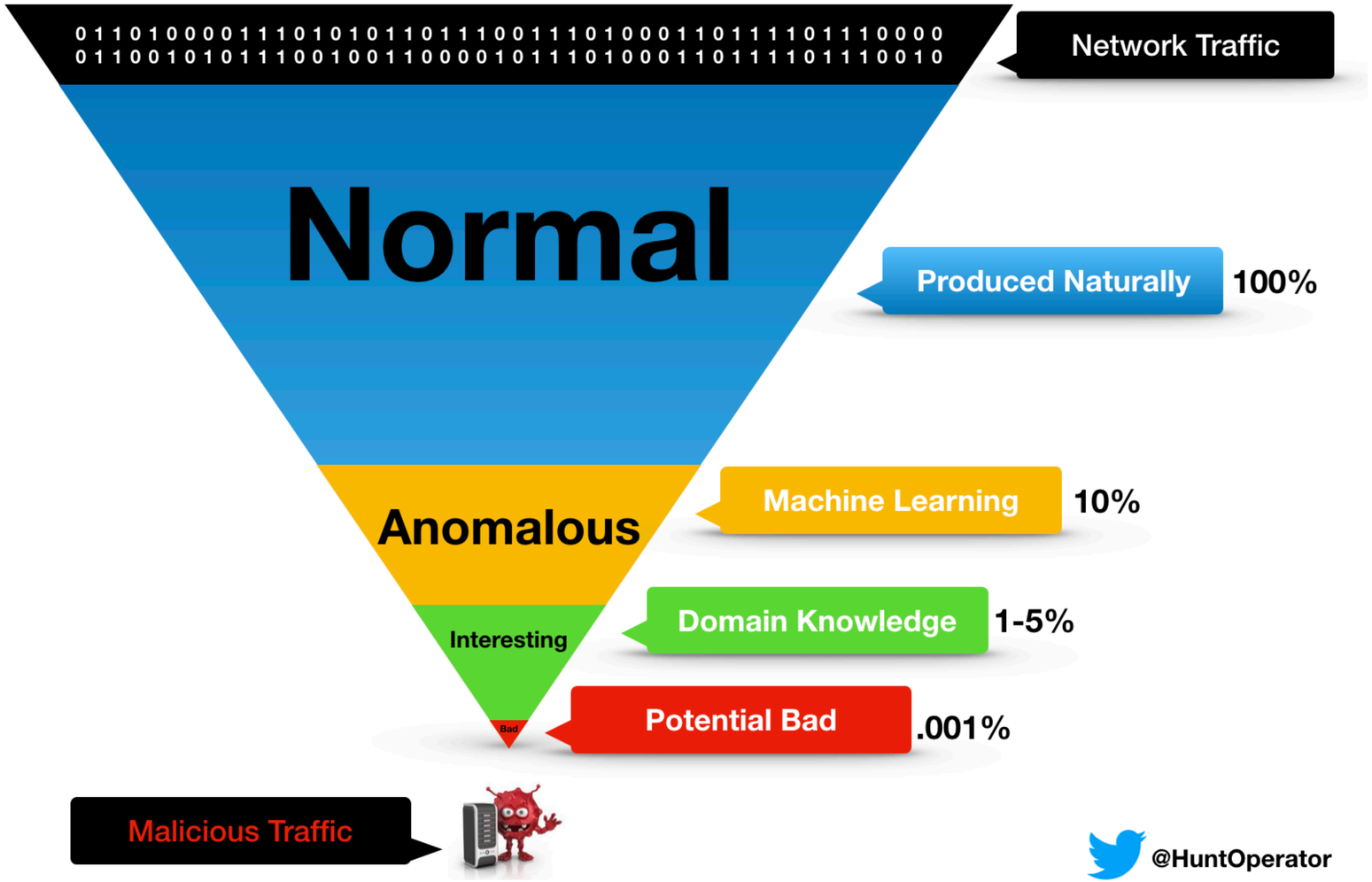
# Malware Detection/Classification

- Derive Features from Binary Content and metadata manifest (function calls, string obtained from IDA Disassembler)
- Example Features: opcode count (n-grams), segment count, asm pixel intensity, n-gramming of bytes, function name.
- Featureless Deep Learning with word2vec embedding
- Get open source malware samples: Vx Heaven, Virus Share, Maltrieve, Open Malware


# Security Applications of Machine Learning

- Domain Generation Algorithm (DGA) Detection  (Classification)
- Malicious URL Detection (Classification)
- Network Traffic: Beaconing Detection  (Classification/Clustering)
- Detection of new classes of malware  (Classification/Clustering)
- General Network Traffic Anomaly Detection  (Classification/Clustering)
- Log Analysis - Anomaly Detection  (Classification/Clustering)
- Phishing Detection (Classification)
- Identifying SQL Injection  (Classification)
- Identifying XSS cross-site scripting  (Classification)
- DOS/DDOS Detection  (Classification)
- Authentication  (Classification)

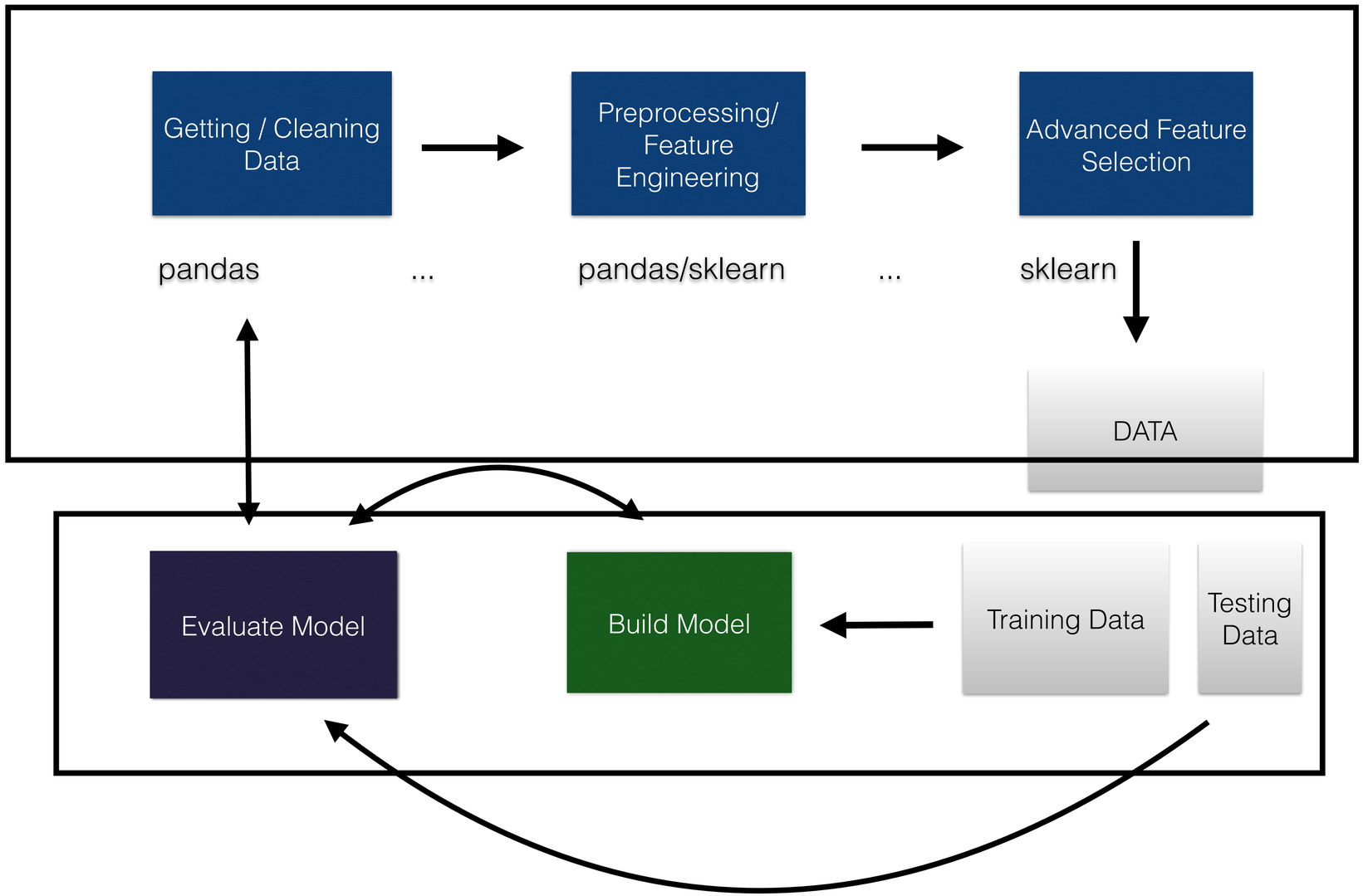
# Data Science Hunting Funnel



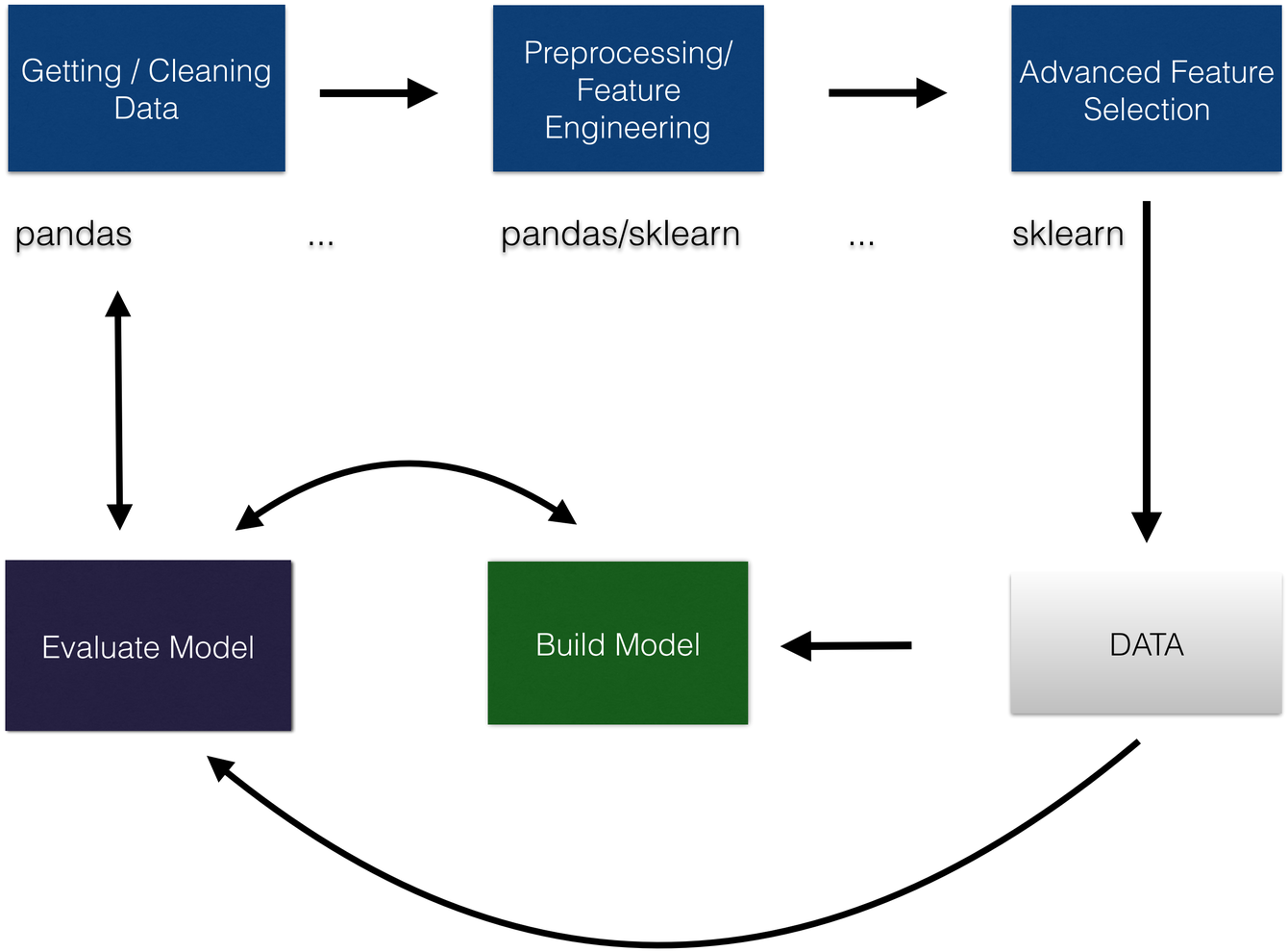

http://www.austintaylor.io/network/traffic/threat/data/science/hunting/funnel/machine/learning/domain/expertise/2017/07/11/data-science-hunting-funnel/


# The Machine Learning Process

# Supervised Machine Learning Process



# Unsupervised Machine Learning Process




# First, define your analytic question.


# What are you trying to do?


# How do you define success?
# What are you measuring?

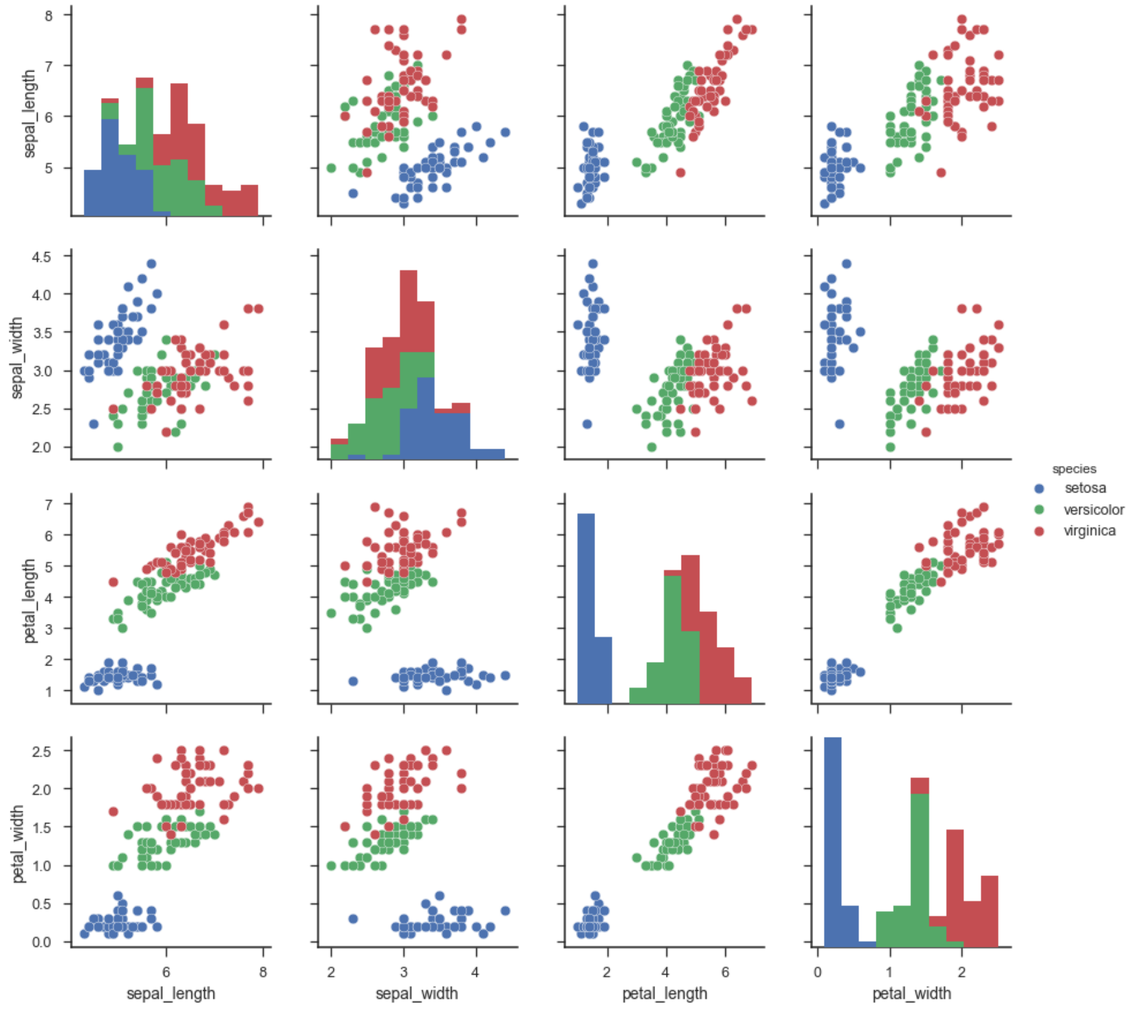

# Choose data sources

- What is available?
- Is it enough?
- Is the data reliable/clean/consistent?
- What other data could you use?


# Other Considerations

- Policies
- Legal contraints
- Biases in Data
- Latency
- Data size

# Gather and Explore Your Data




Is the data good enough?

What are the rules governing its use?

Do I have enough?

Do problems or biases exist in the data that could cause problems?

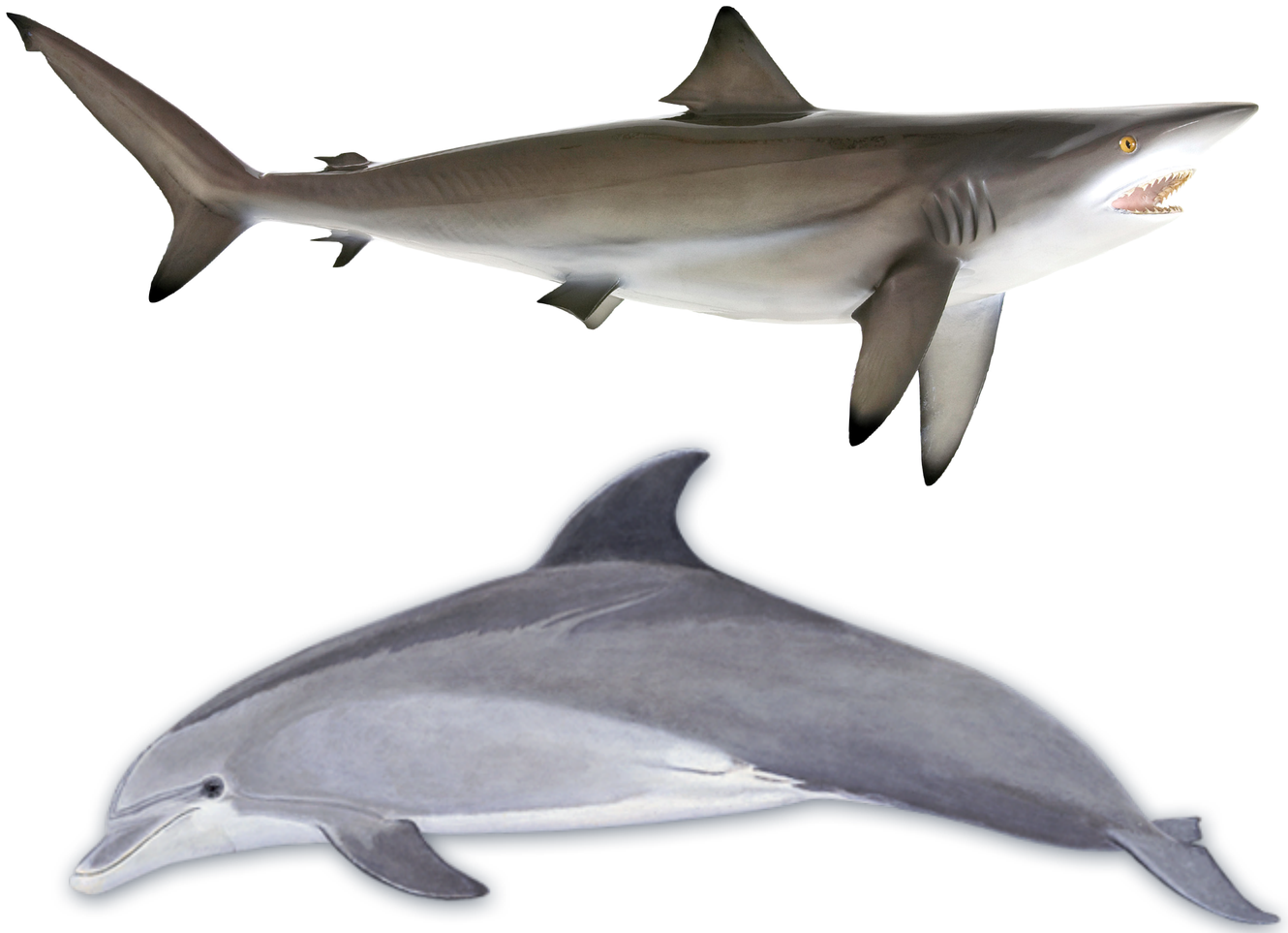

# Feature Engineering

- Define what you are trying to measure.  These will become the observations or rows of your final dataset
- Define how you will mathematically represent your data.  This will be come the features or columns of your final dataset.

# Feature Engineering

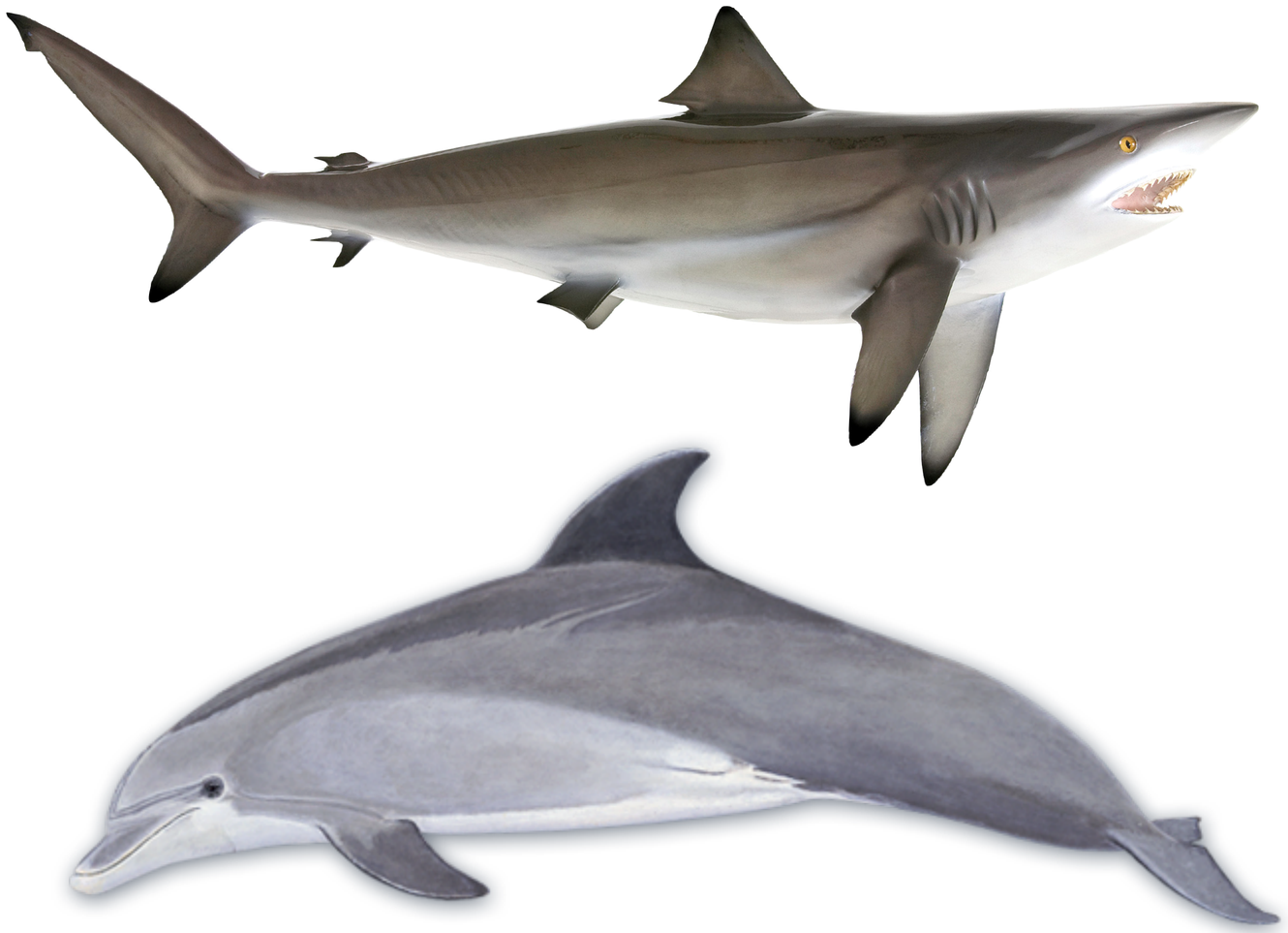
| Feature | Value |
| --- | --- |
| Color | Gray |
| Fins | 7 |
| Predator | TRUE |

# Feature Engineering



| Feature | Value |
|---|---|
| Color | Gray |
| Fins | 7 |
| Predator | TRUE |

# Feature Engineering



| Feature | Value |
|---|---|
| Color | Gray |
| Fins | 7 |
| Predator | TRUE |
| **Mammal** | **TRUE** |


# Build and Tune your Model

- Believe it or not, this is the easy part.
- Most of this is **done using libraries** like scikit-learn, mllib, tensorflow, caret or keras, and **many steps can be automated**.
- You can even do it in Splunk.

# Evaluate Performance

- Use various scoring methods, or write your own to determine model performance.
- Go back to step 1 and repeat!  (Do not pass go, do not collect $200)

# Discussion

- Consider that you are building a system to identify fraudulent credit card transactions.  What are some features that you would want to capture?

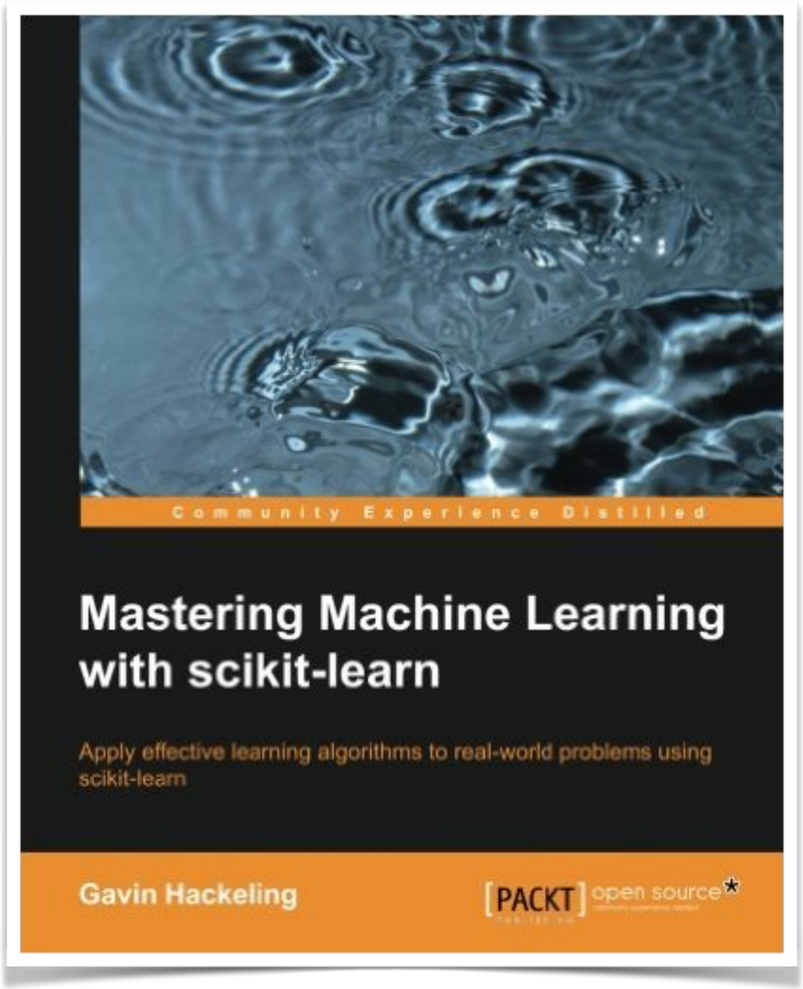
- What are the challenges you could foresee in this problem?


# The Python Data Science Ecosystem

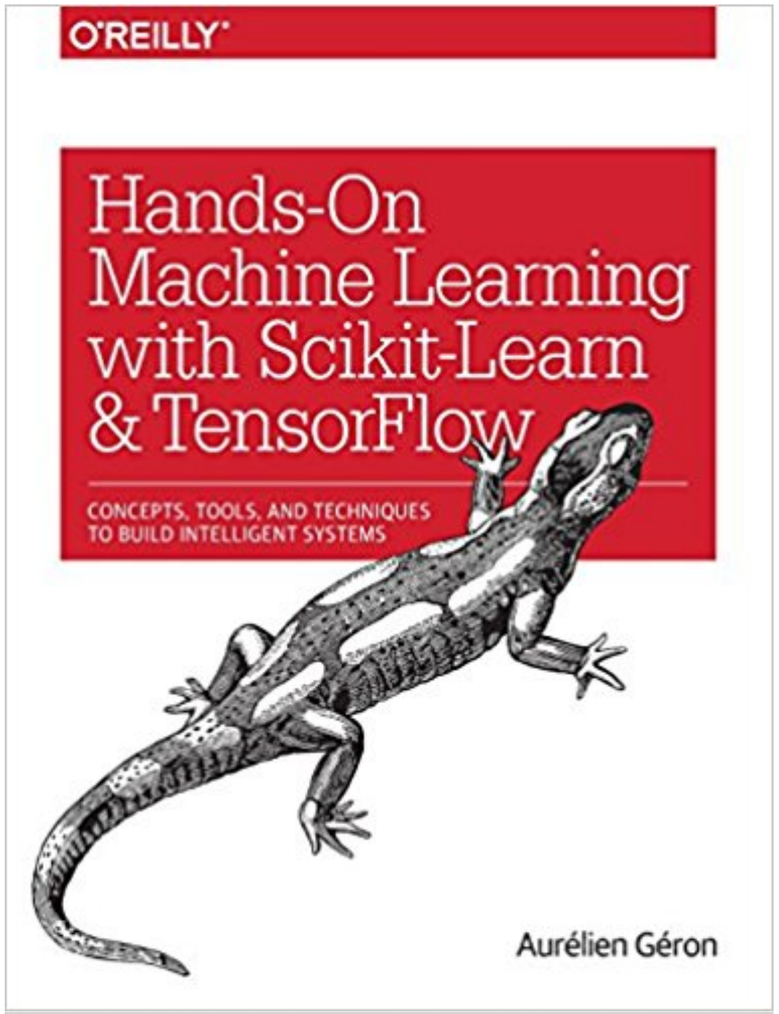

# Machine Learning Ecosystem

- **Data Gathering:** Pandas, Drill, BeautifulSoup, PyDBAPI, PyDAL, Boto3
- **Feature Extraction:** Pandas, NumPy, Featuretools
- **Machine Learning**
  - **"Regular" ML:** Scikit-learn (sklearn), h2o, mllib (PySpark)
  - **Deep Learning:** Tensorflow, Keras, Theano, Caffe, PyTorch
- **Visualization:** Matplotlib, Seaborn, Yellowbrick, LIME, ggplot, plot.ly,

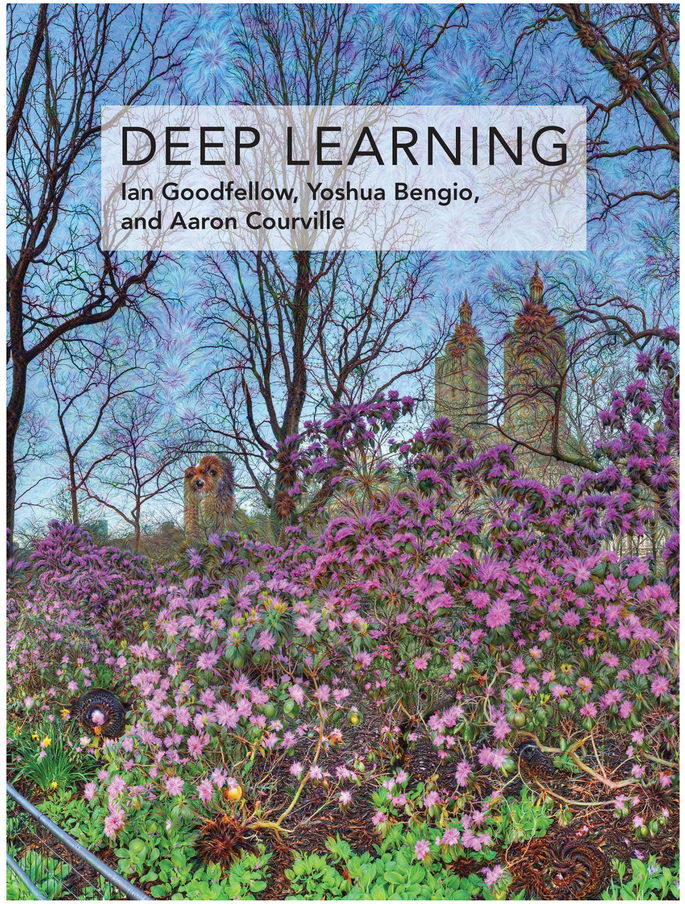
# Recommended Reading



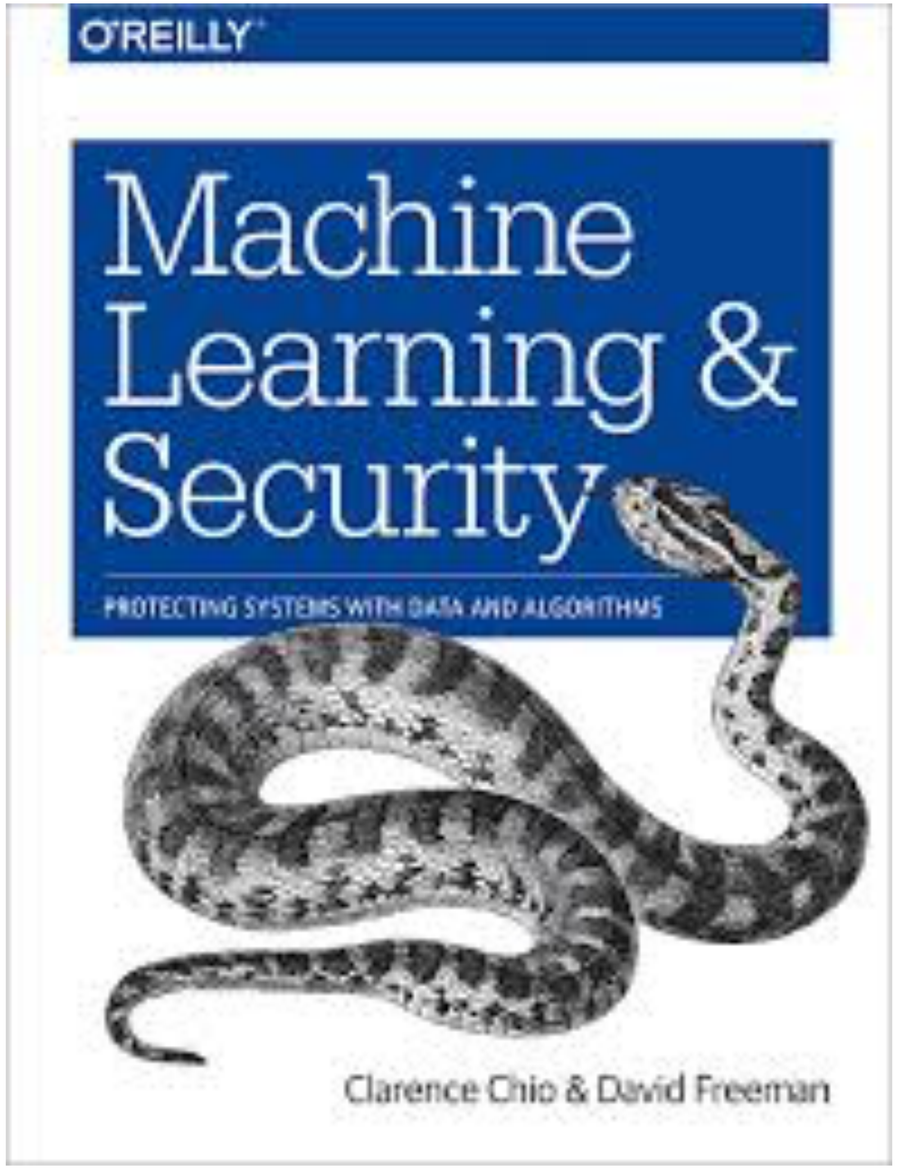
# Recommended Reading

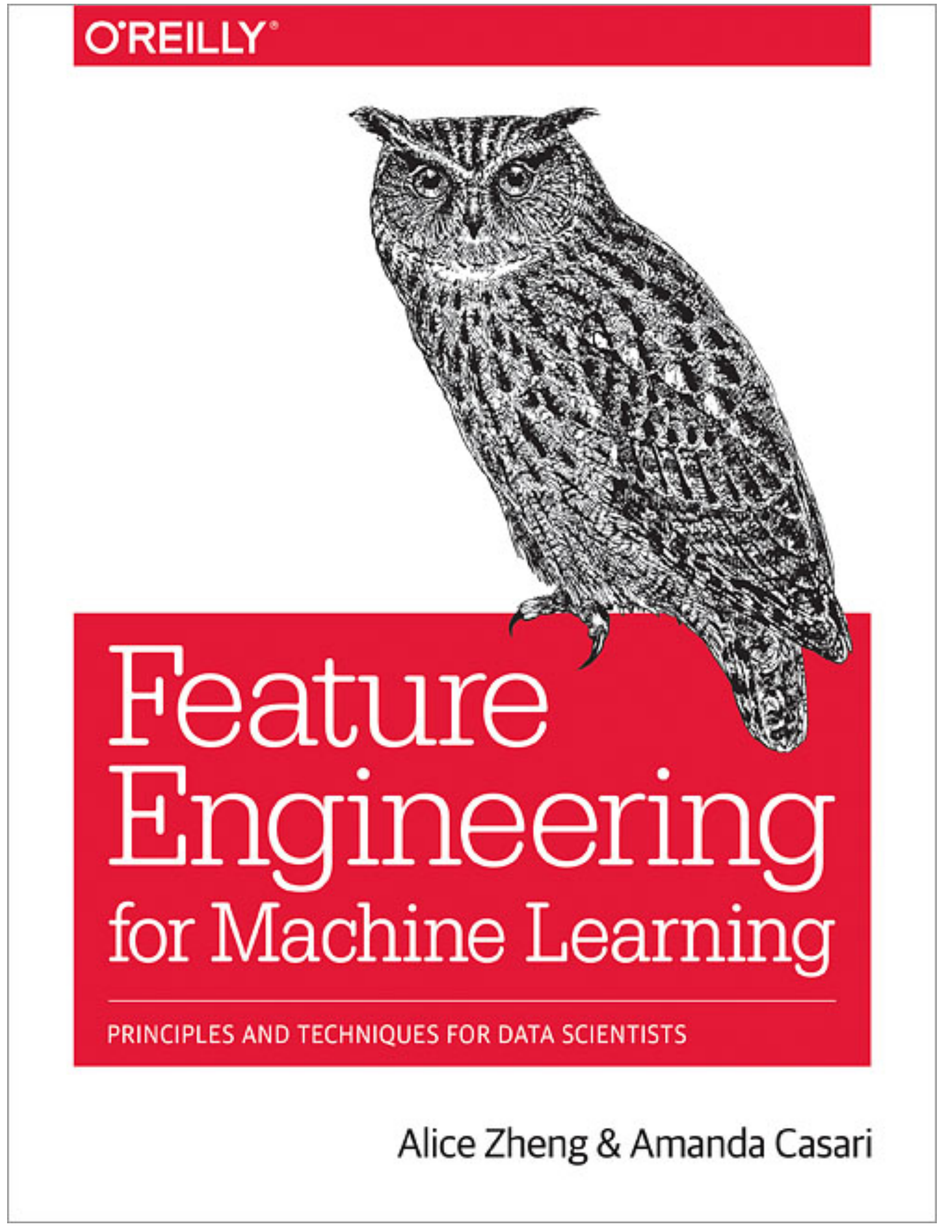


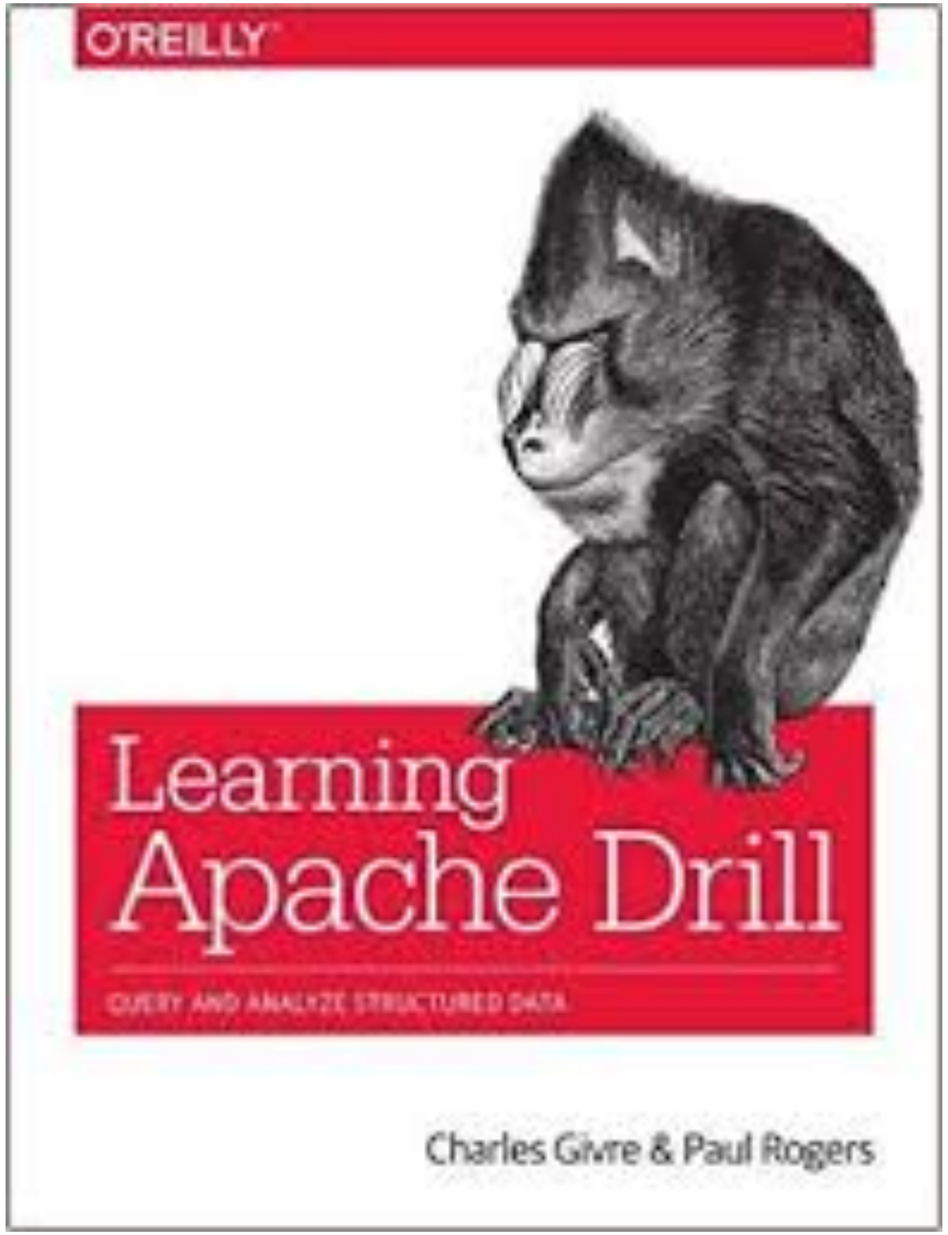
# Recommended Reading




http://www.deeplearningbook.org/


O'REILLY
Machine Learning & Security
PROTECTING SYSTEMS WITH DATA AND ALGORITHMS
Clarence Chio & David Freeman


O'REILLY
Feature Engineering for Machine Learning
PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS
Alice Zheng & Amanda Casari


O'REILLY
Learning
Apache Drill
QUERY AND ANALYZE STRUCTURED DATA
Charles Givre & Paul Rogers


# The Virtual Machine: Griffon


```
File  Edit  View  Search  Terminal  Help
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/hbase/hbase-1.1.3/lib/slf4j-log4j12-1
.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4
j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2016-05-16 13:04:54,887 WARN  [main] util.NativeCodeLoader: Unable to load nativ
e-hadoop library for your platform... using builtin-java classes where applicabl
e
2016-05-16 13:05:11,827 ERROR [main] zookeeper.RecoverableZooKeeper: ZooKeeper exists faile
d after 4 attempts
2016-05-16 13:05:11,828 WARN  [main] zookeeper.ZKUtil: hconnection-0x46a145ba0x0, quorum=lo
calhost:2181, baseZNode=/hbase Unable to set watcher on znode (/hbase/hbaseid)
org.apache.zookeeper.KeeperException$ConnectionLossException: KeeperErrorCode = ConnectionL
oss for /hbase/hbaseid
        at org.apache.zookeeper.KeeperException.create(KeeperException.java:99)
        at org.apache.zookeeper.KeeperException.create(KeeperException.java:51)
        at org.apache.zookeeper.ZooKeeper.exists(ZooKeeper.java:1045)
        at org.apache.hadoop.hbase.zookeeper.RecoverableZooKeeper.exists(RecoverableZooKeep
er.java:221)
        at org.apache.hadoop.hbase.zookeeper.ZKUtil.checkExists(ZKUtil.java:541)
        at org.apache.hadoop.hbase.zookeeper.ZKClusterId.readClusterIdZNode(ZKClusterId.jav
a:65)
        at org.apache.hadoop.hbase.client.ZooKeeperRegistry.getClusterId(ZooKeeperRegistry.
java:105)
```


# Do Data Science, Not Sysadmin


# Built on Ubuntu MATE


# Easy to Use


# Programming Languages

- Languages
- Libraries
- Editors and Notebooks
- Databases + Administrative tools
- Big Data Tools
- Machine Learning Libraries
- Data Visualization

# Scripting Languages







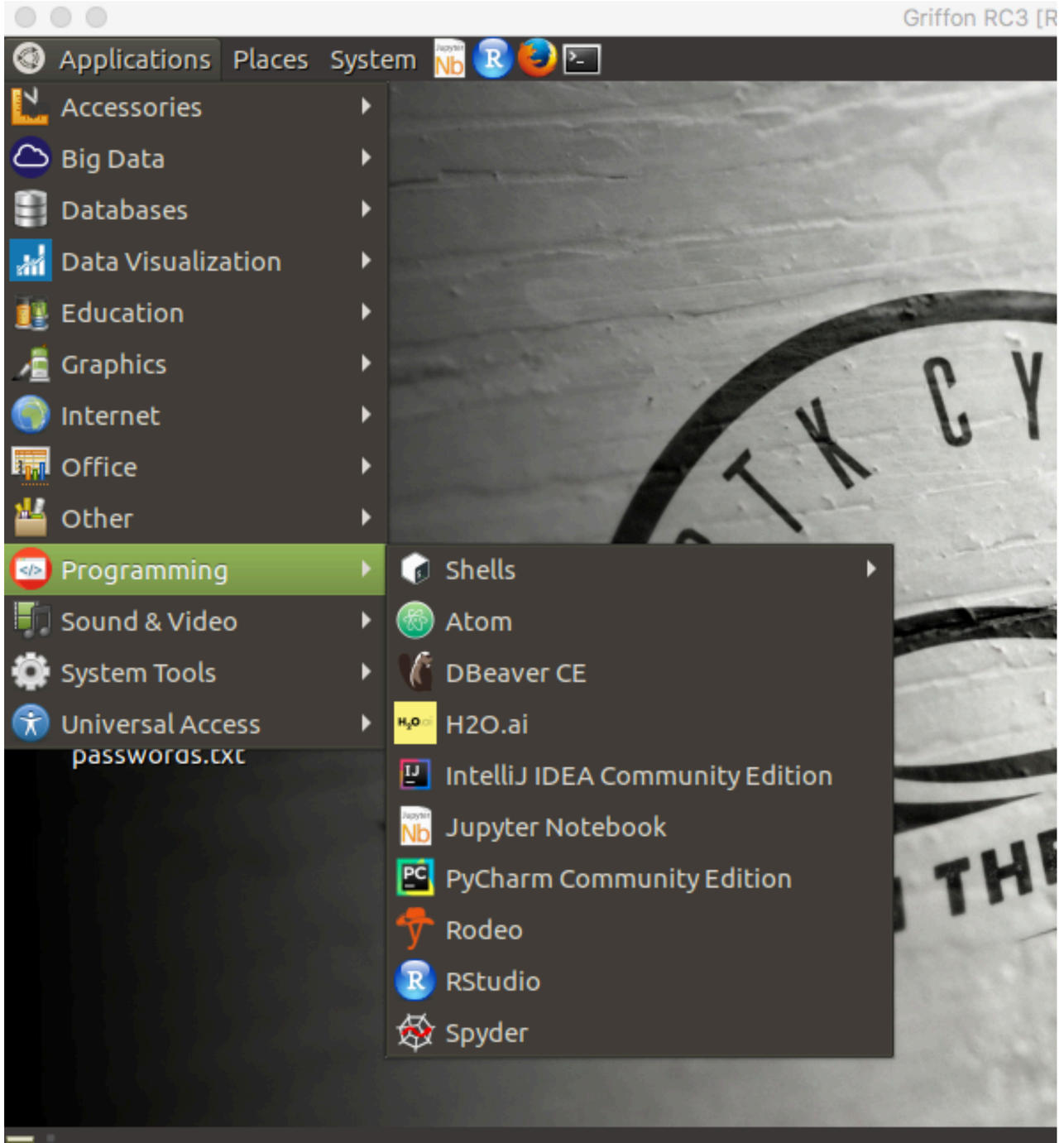
Griffon RC3 [Running]
Applications
Places
System
Accessories
Big Data
Databases
Data Visualization
Education
Graphics
Internet
Office
Other
Programming
Sound & Video
System Tools
Universal Access
passwords.txt
Shells
Atom
DBeaver CE
H2O.ai
IntelliJ IDEA Community Edition
Jupyter Notebook
PyCharm Community Edition
Rodeo
RStudio
Spyder
Julia
NodeJS
PHP Shell
Pig Shell
Python (v2.7)
Python (v3.5)
R
Ruby
Scala


# Editors & Notebooks

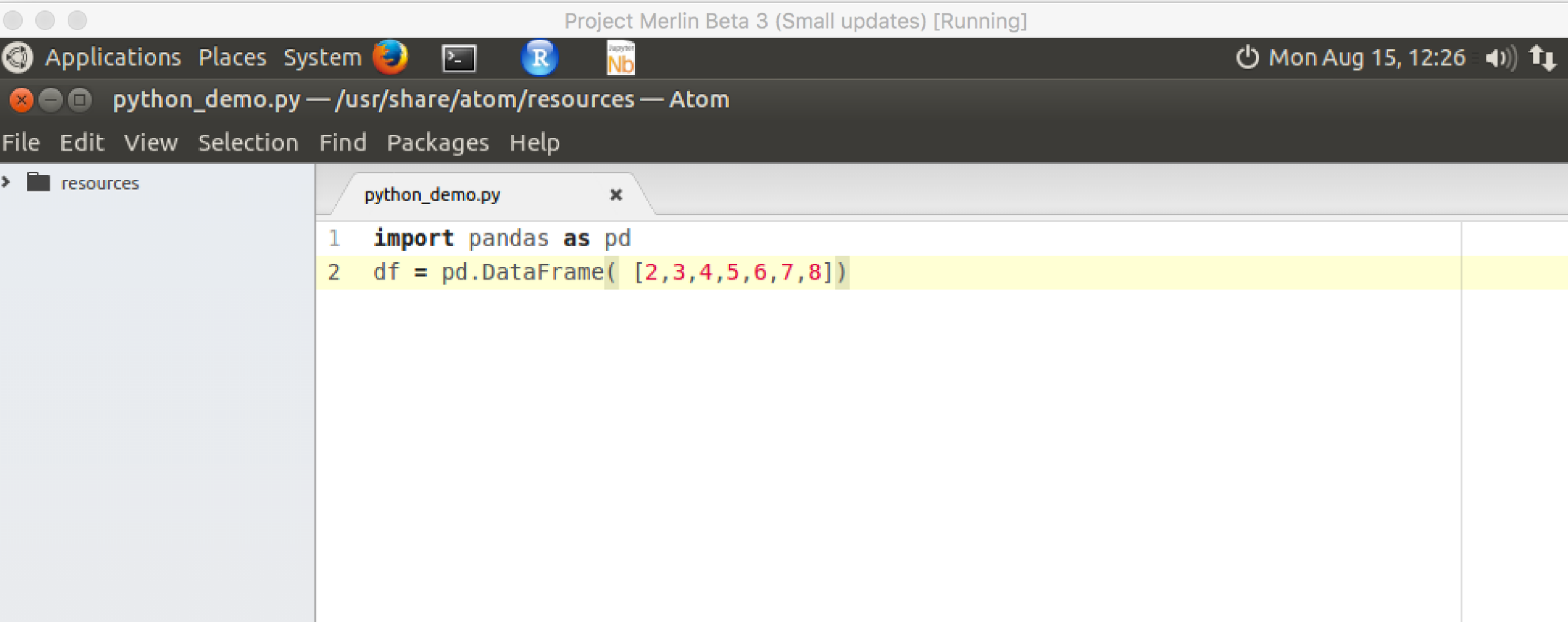




gtkcyber.com

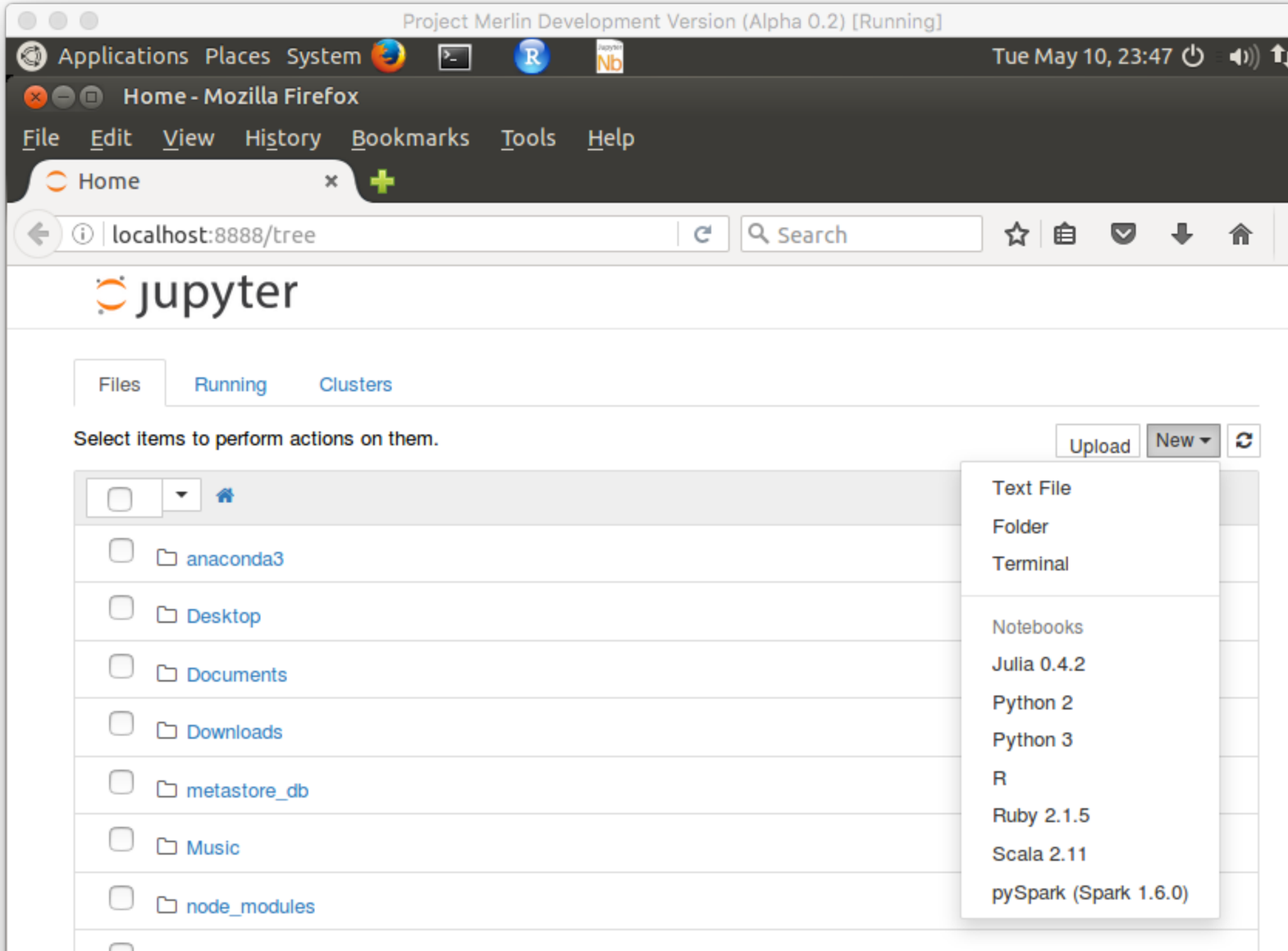


# Atom


Project Merlin Beta 3 (Small updates) [Running]
Applications Places System
Mon Aug 15, 12:26
python_demo.py — /usr/share/atom/resources — Atom
File Edit View Selection Find Packages Help
resources
python_demo.py
1 import pandas as pd
2 df = pd.DataFrame( [2,3,4,5,6,7,8])



gtkcyber.com





# Jupyter Notebook

- Python 2 & 3
- R
- Ruby
- Scala
- PySpark


GTK CYBER
GET ON THE FAST TRACK

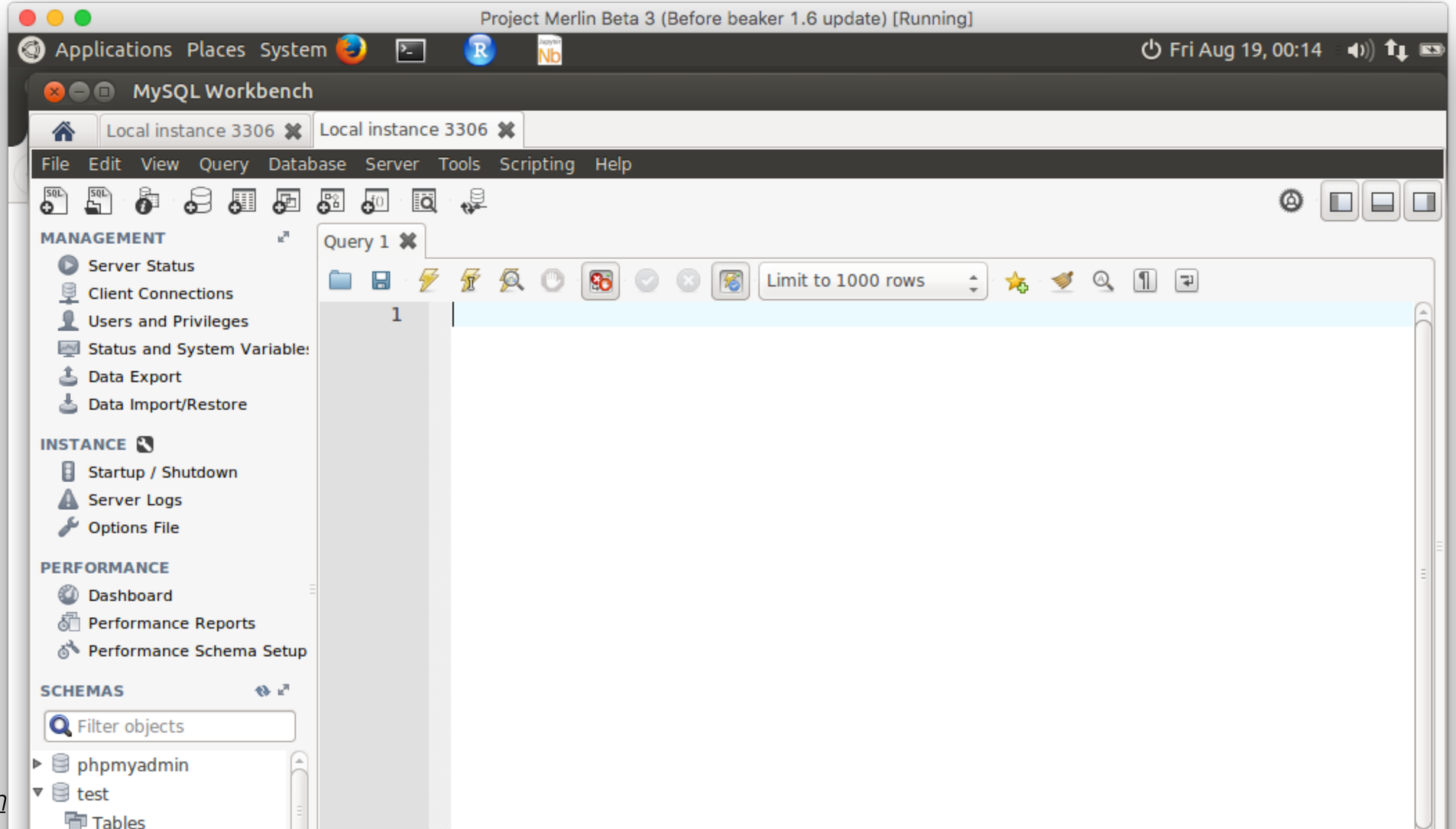

MySQL®


mongoDB®


PostgreSQL

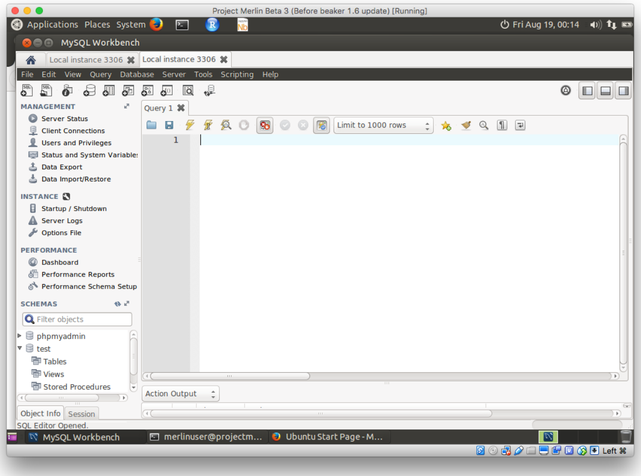

neo4j

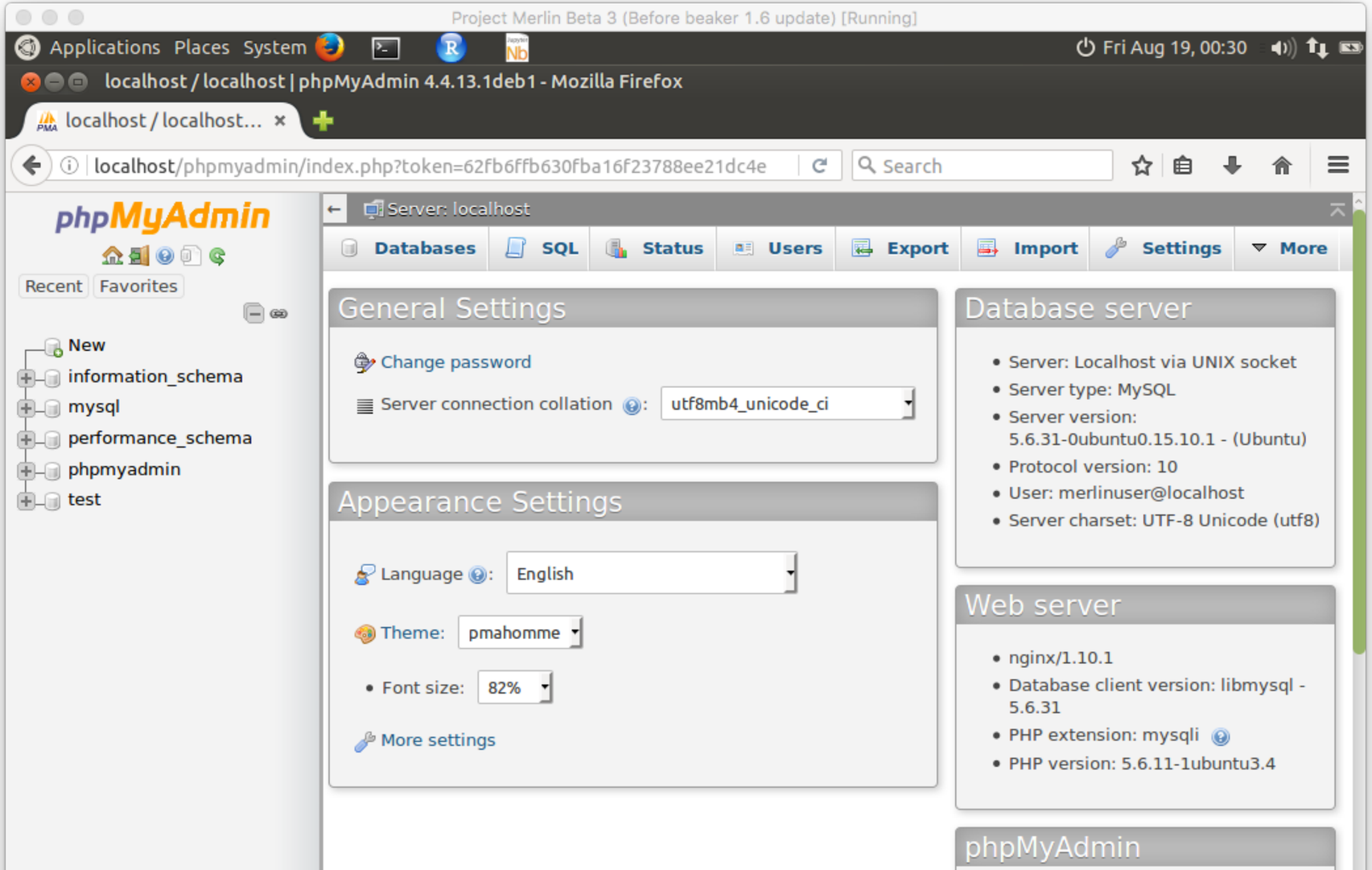

SQLite


Project Merlin Beta 3 (Before beaker 1.6 update) [Running]
Applications Places System
Fri Aug 19, 00:14
MySQL Workbench
Local instance 3306
Local instance 3306
File Edit View Query Database Server Tools Scripting Help
MANAGEMENT
Server Status
Client Connections
Users and Privileges
Status and System Variables
Data Export
Data Import/Restore
INSTANCE
Startup / Shutdown
Server Logs
Options File
PERFORMANCE
Dashboard
Performance Reports
Performance Schema Setup
SCHEMAS
Filter objects
phpmyadmin
test
Tables
Query 1
Limit to 1000 rows
1

GET ON THE FAST TRACK

MySQL®

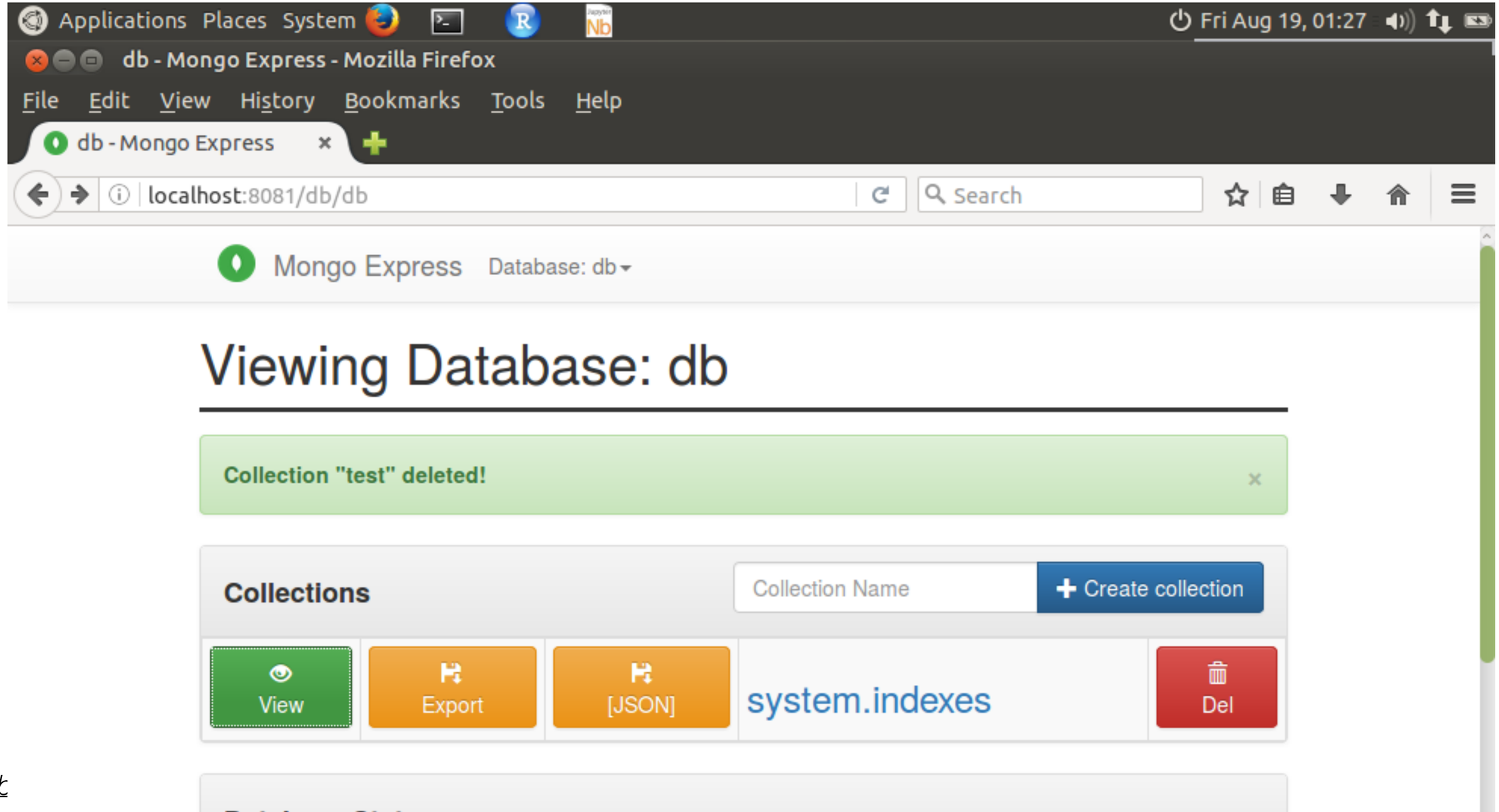


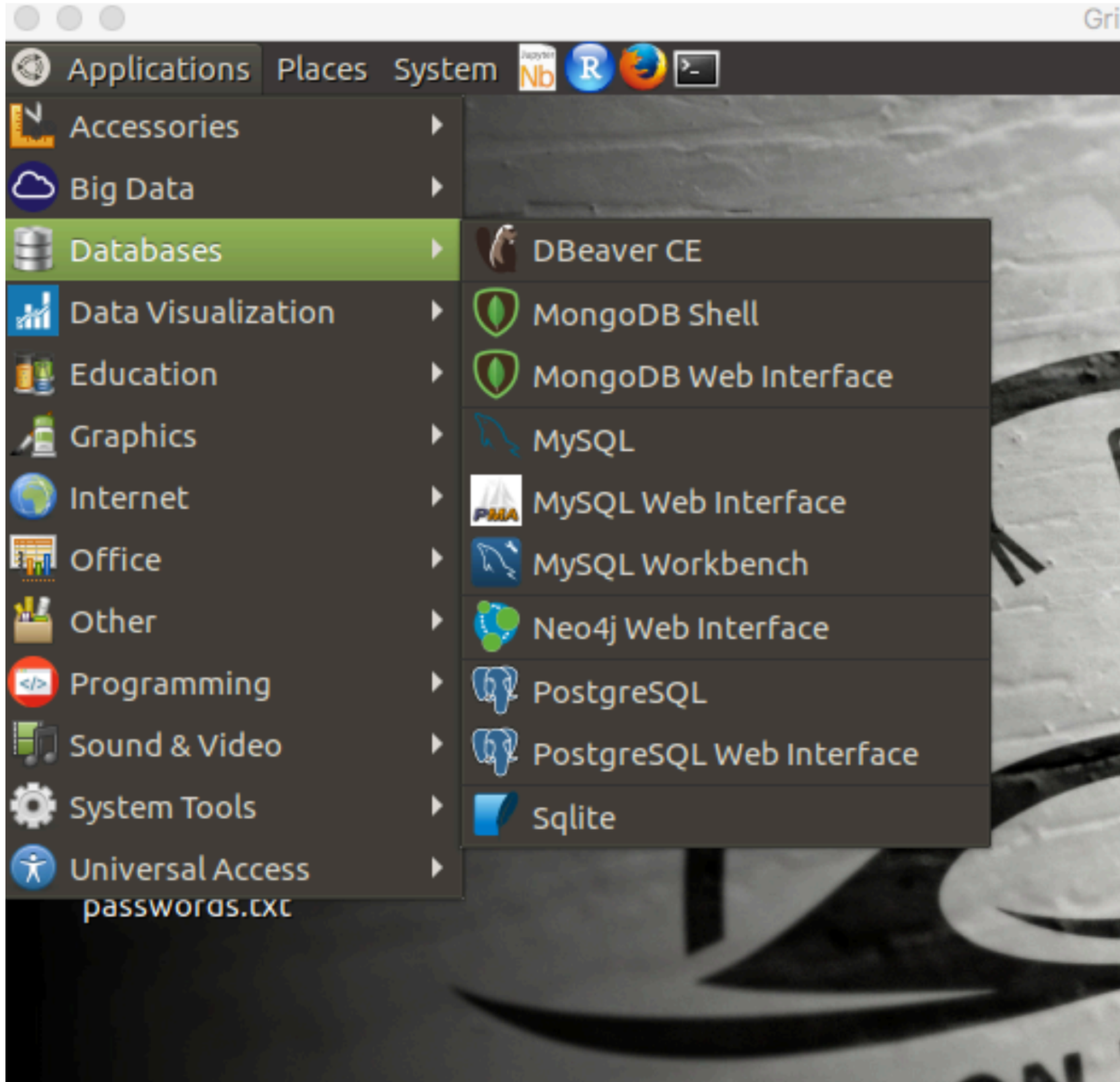




gtkcyber.com

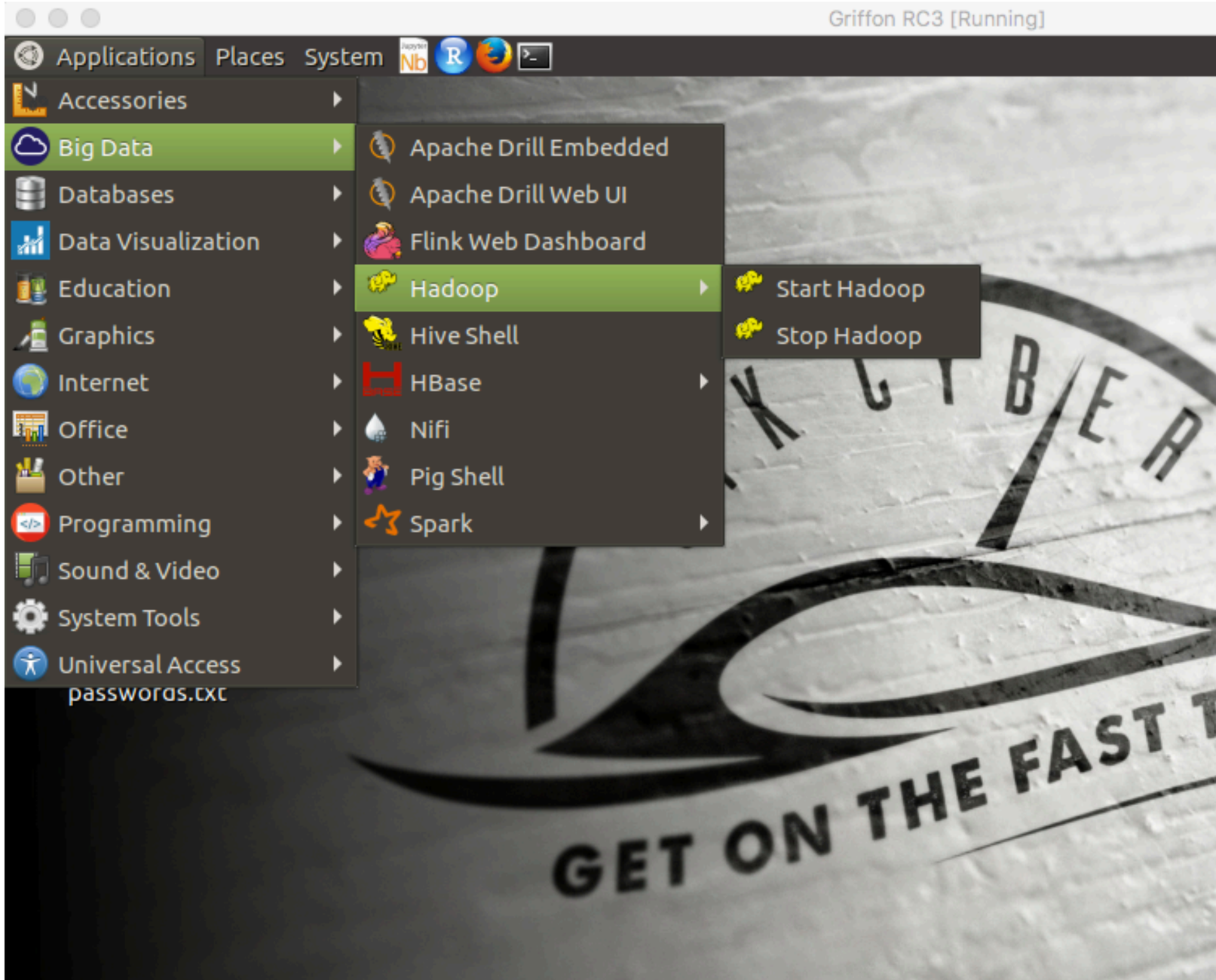

GTK CYBER
GET ON THE FAST TRACK

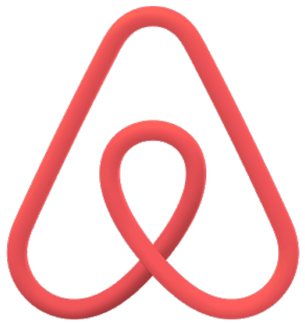

mongoDB.


Applications Places System
Fri Aug 19, 01:27
db - Mongo Express - Mozilla Firefox
File Edit View History Bookmarks Tools Help
db - Mongo Express
localhost:8081/db/db
Search
Mongo Express
Database: db
Viewing Database: db
Collection "test" deleted!
Collections
Collection Name
Create collection
View
Export
[JSON]
system.indexes
Del

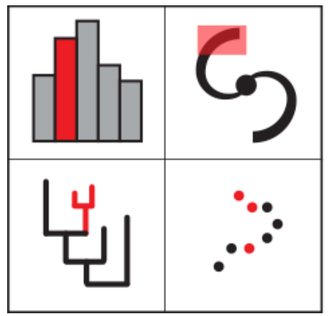

Applications
Places
System
Accessories
Big Data
Databases
Data Visualization
Education
Graphics
Internet
Office
Other
Programming
Sound & Video
System Tools
Universal Access
DBeaver CE
MongoDB Shell
MongoDB Web Interface
MySQL
MySQL Web Interface
MySQL Workbench
Neo4j Web Interface
PostgreSQL
PostgreSQL Web Interface
Sqlite
passwords.txt


merlin projectmerlin: /usr/local/hadoop/sbin
Edit View Terminal Help
merlinuser@project n:/usr/local/hadoop/sbin$ ./start-dfs.sh
16/05/11 23:23:39 WA il.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform.. ng builtin-java classes where applicable
Starting namenodes on [lo ost]
merlinuser@localhost's pas :
merlinuser@localhost's passw localhost: Permission denied, please try again.
merlinuser@localhost's password: lhost: Permission denied, please try again.
localhost: Permission denied (publi assword).
merlinuser@localhost's password:
merlinuser@localhost's password: localho ermission denied, please try again.


Griffon RC3 [Running]
Applications
Places
System
Accessories
Big Data
Databases
Data Visualization
Education
Graphics
Internet
Office
Other
Programming
Sound & Video
System Tools
Universal Access
Apache Drill Embedded
Apache Drill Web UI
Flink Web Dashboard
Hadoop
Hive Shell
HBase
Nifi
Pig Shell
Spark
Start Hadoop
Stop Hadoop
passwords.txt
GET ON THE FAST T





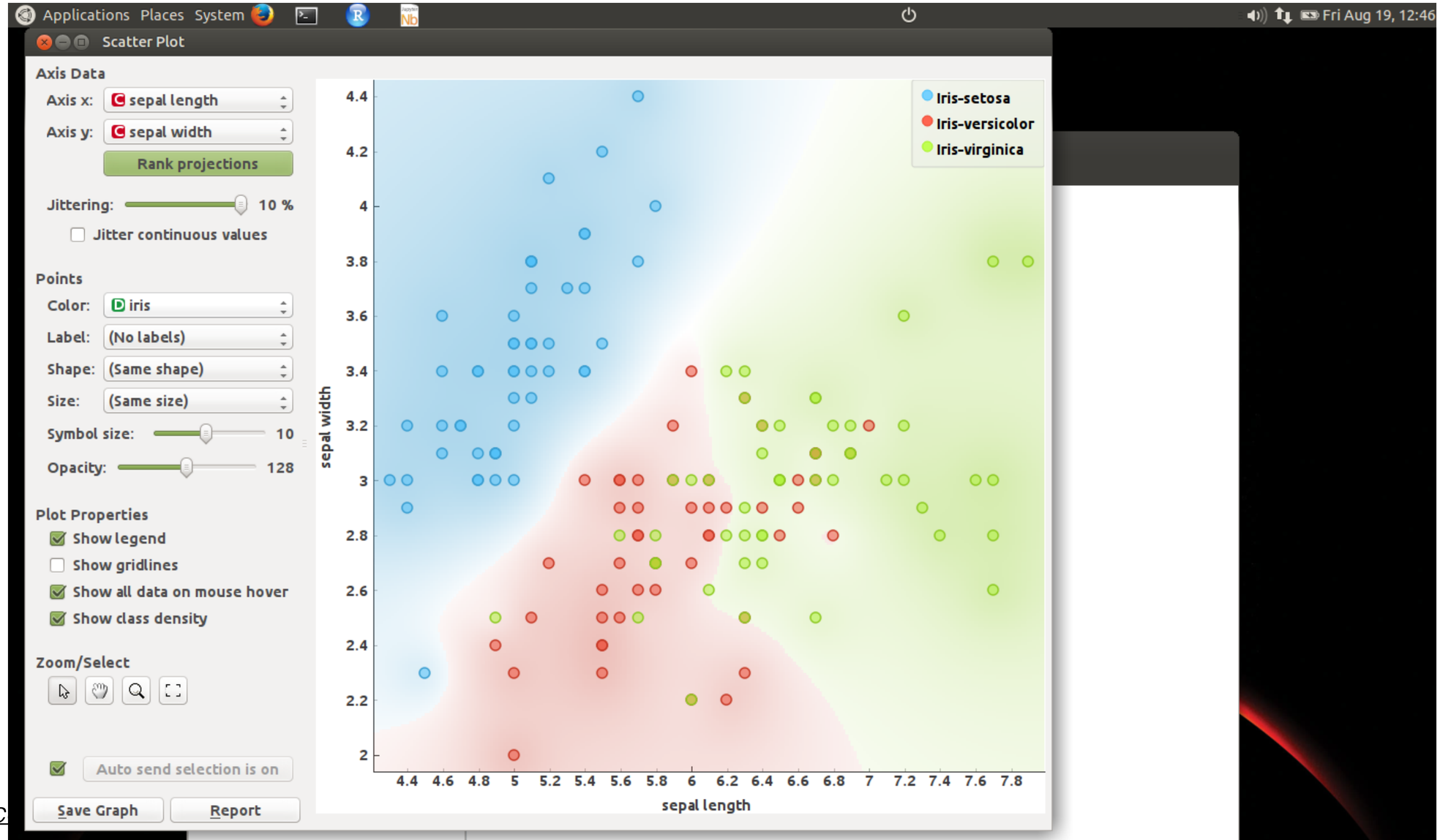
gtkcyber.com

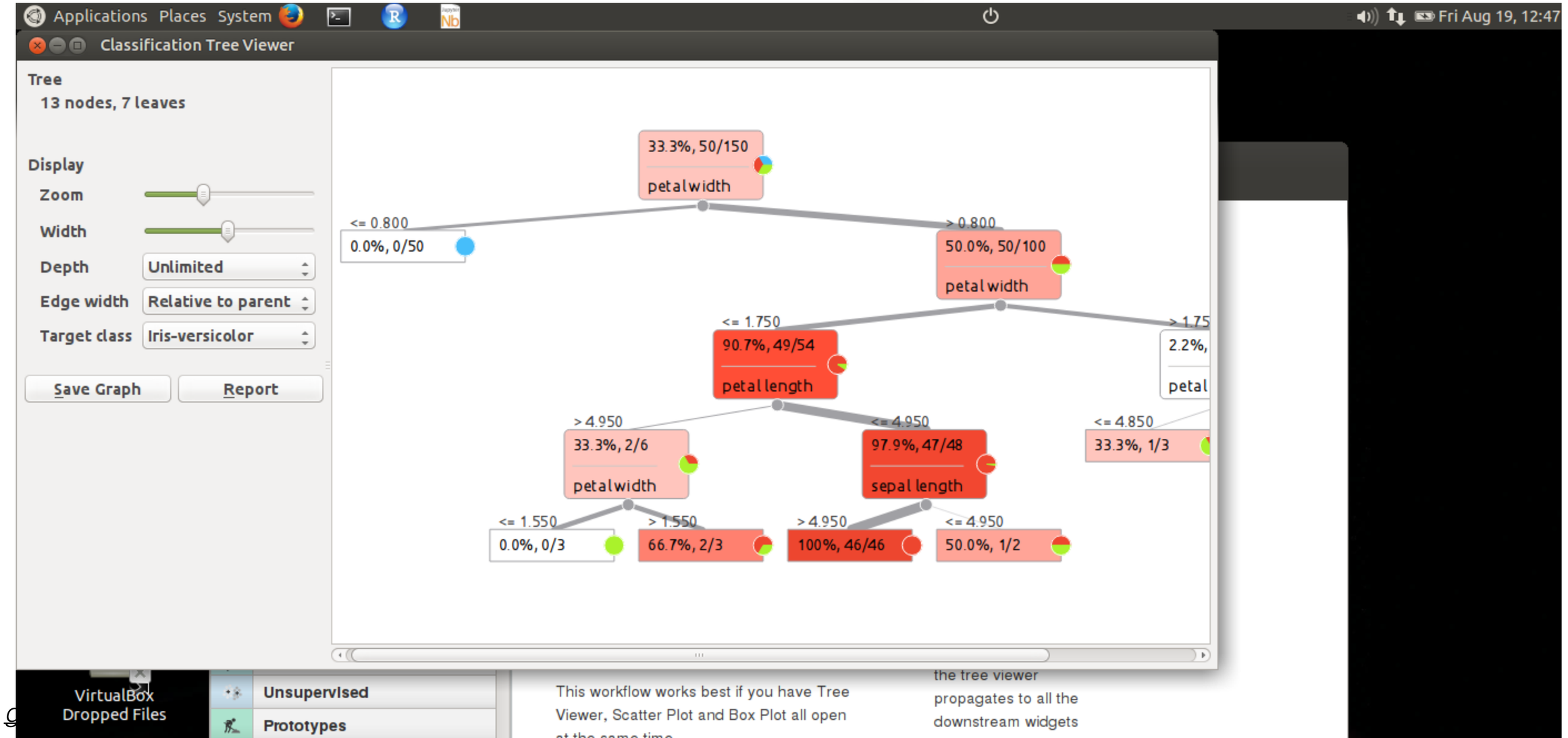

GTK CYBER
GET ON THE FAST TRACK


orange
DATA MINING
FRUITFUL&FUN

GTK CYBER
GET ON THE FAST TRACK

orange
DATA MINING
FRUITFUL&FUN

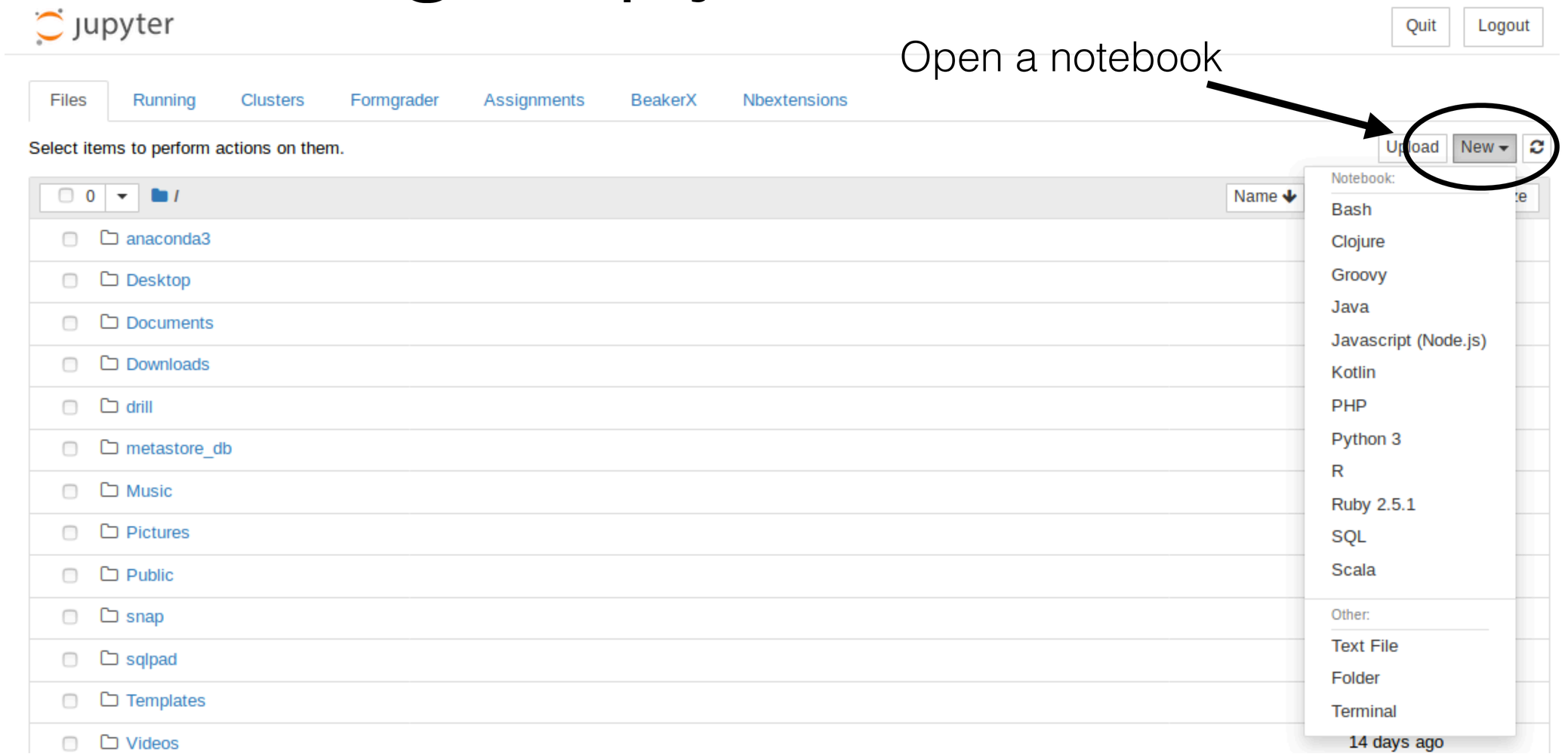



Applications Places System
Fri Aug 19, 12:47
Classification Tree Viewer
Tree
13 nodes, 7 leaves
Display
Zoom
Width
Depth
Unlimited
Edge width
Relative to parent
Target class
Iris-versicolor
Save Graph
Report
33.3%, 50/150
petal width
<= 0.800
0.0%, 0/50
> 0.800
50.0%, 50/100
petal width
<= 1.750
90.7%, 49/54
petal length
> 1.75
2.2%,
petal
> 4.950
33.3%, 2/6
petal width
<= 4.950
97.9%, 47/48
sepal length
<= 4.850
33.3%, 1/3
<= 1.550
0.0%, 0/3
> 1.550
66.7%, 2/3
> 4.950
100%, 46/46
<= 4.950
50.0%, 1/2
VirtualBox
Dropped Files
Unsupervised
Prototypes
This workflow works best if you have Tree
Viewer, Scatter Plot and Box Plot all open
the tree viewer
propagates to all the
downstream widgets

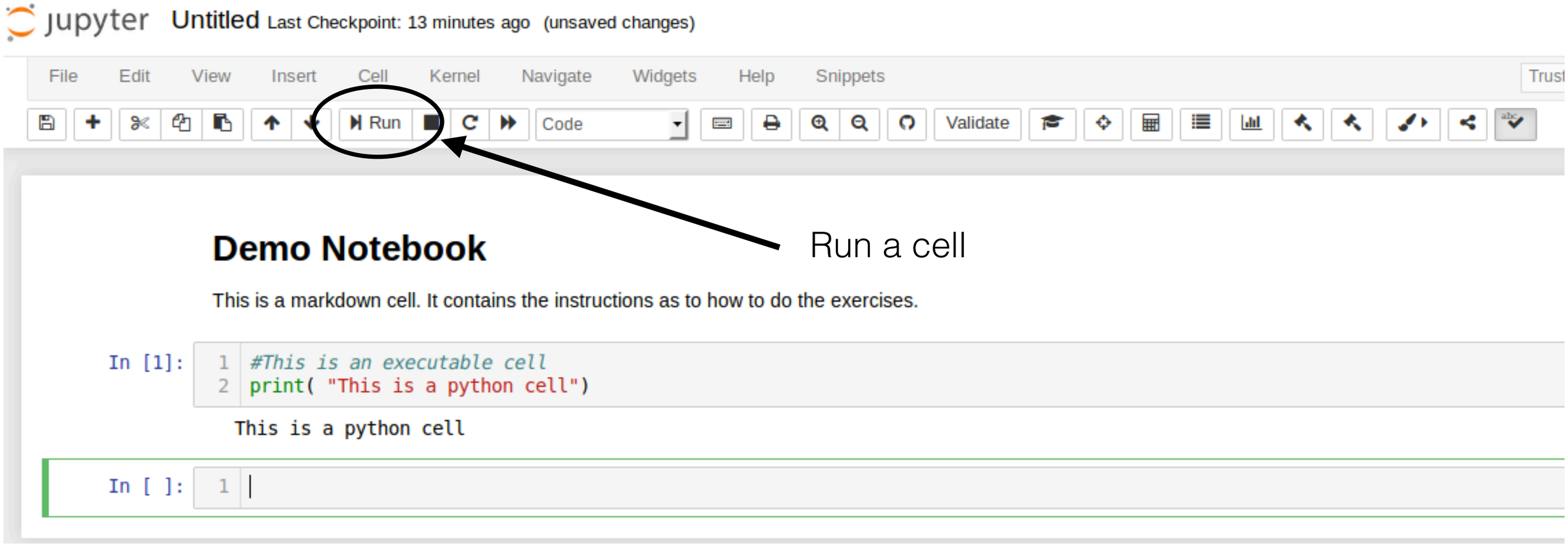

# Questions?


# Using Jupyter Notebook

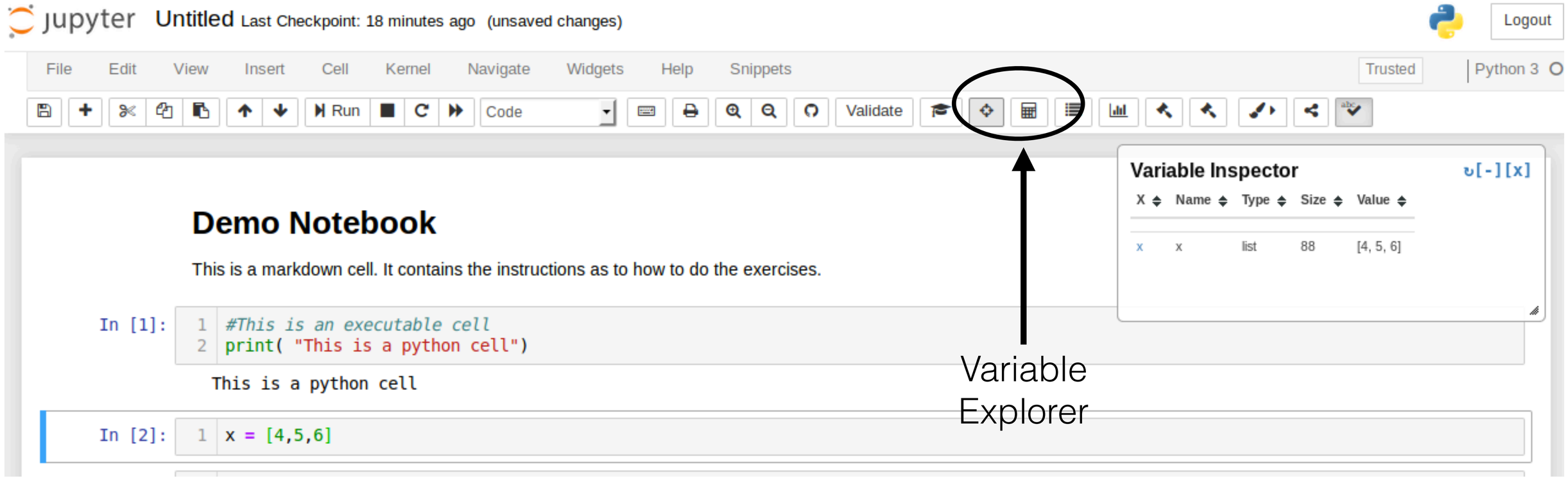
# Using Jupyter Notebook



# Using Jupyter Notebook



# Using Jupyter Notebook




# Questions?

# labs.gtkcyber.com

Username: student_0x
Password: LexGD92y#$P$

Slack https://bit.ly/2J2H1Di