Applied Supervised Learning for Cyber Security
Module 0: Introduction

Course Agenda

Day 1

- Intro: What is Data Science?
  - Overview of Machine Learning & Cyber Applications
- Manipulating and Exploring Data
  - Exploratory Data Analysis in 1 Dimension
  - Exploratory Data Analysis in 2 Dimensions
- Data Visualization

Day 2

- Machine Learning
  - Supervised & Unsupervised
- Anomaly Detection
- Attacking AI
- Hunting with Data Science
- The Future: LLMs & GPT

Expectations

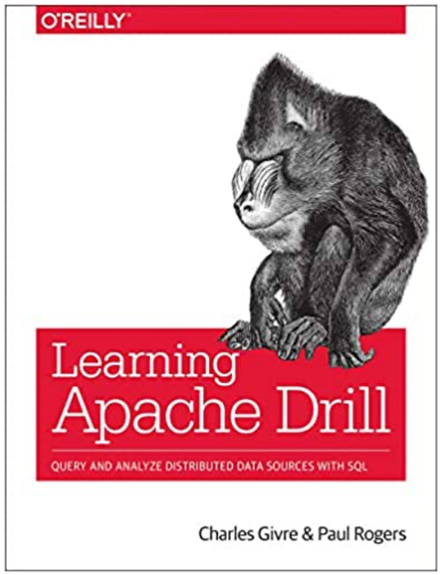
- Please participate and **ask questions.**
- Please follow along and **TRY OUT** the examples yourself during the class
- All the answers are in the slide decks or GitHub repository, but please try to complete the exercises **without looking at the answers**.
- Join the conversation in slack!
- Have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

Introduction

Our Lawyers Make Us Say This

Charles Givre, CISSP

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



JPMorgan Chase & Co.

Deutsche Bank

Booz | Allen | Hamilton

100 YEARS

ARIZONA

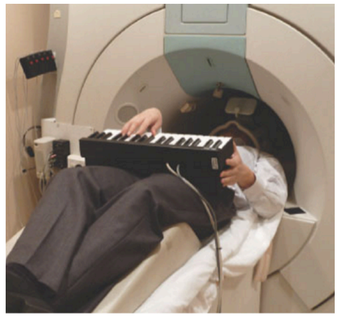
Brandeis

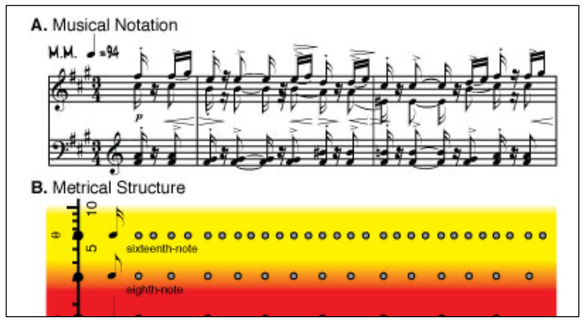
TRUTH · EVEN UNTO ITS INNERMOST PARTS · אמת

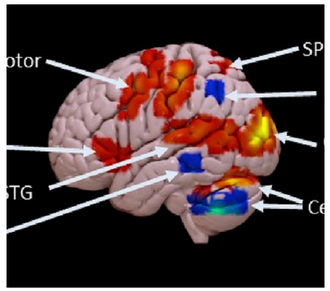
Summer Rankin, PhD

- Senior Lead Data Scientist @ Booz Allen Hamilton (5 years)
  - DoD, FDA, CDC, ONC, CMS
- Computational Neuroscientist in Auditory and Music Perception
- PhD in Complex Systems and Brain Sciences

JOHNS HOPKINS MEDICINE







Booz | Allen | Hamilton

100 YEARS

John Marion

- Data Engineer & Software Developer
- AWS Certified Cloud Developer
- Former BAH Data Scientist supporting DoD
- Former U.S. Marine Corps Intelligence Analyst
- Husband/Father
- Gardener
- Dungeon Master

Who are you?

- Your name (or what you want us to call you)
- Your job role
- What you hope to get out of this class
- Your level of experience with coding

Stop

What is Data Science?

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

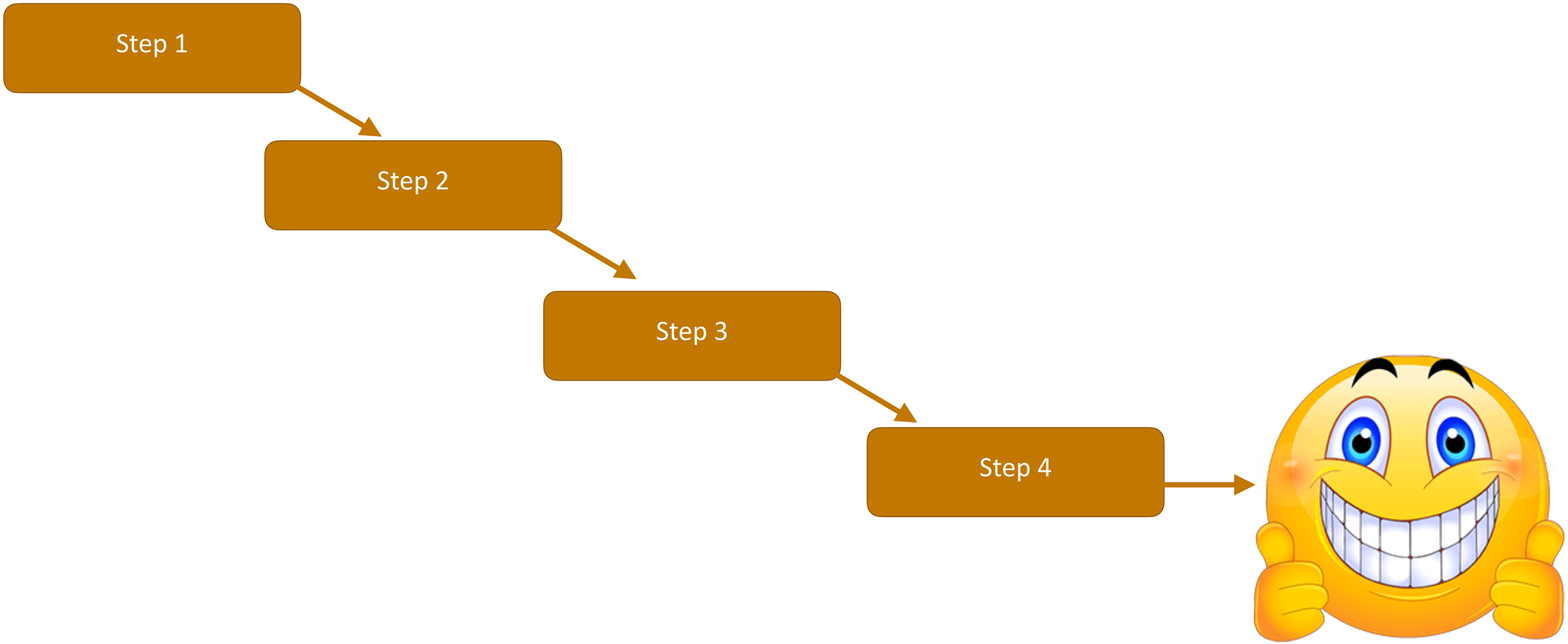
Data Science is the art of turning data into actions. This is accomplished through the creation of data products, which provide actionable information without exposing decision makers to the underlying data or analytics

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

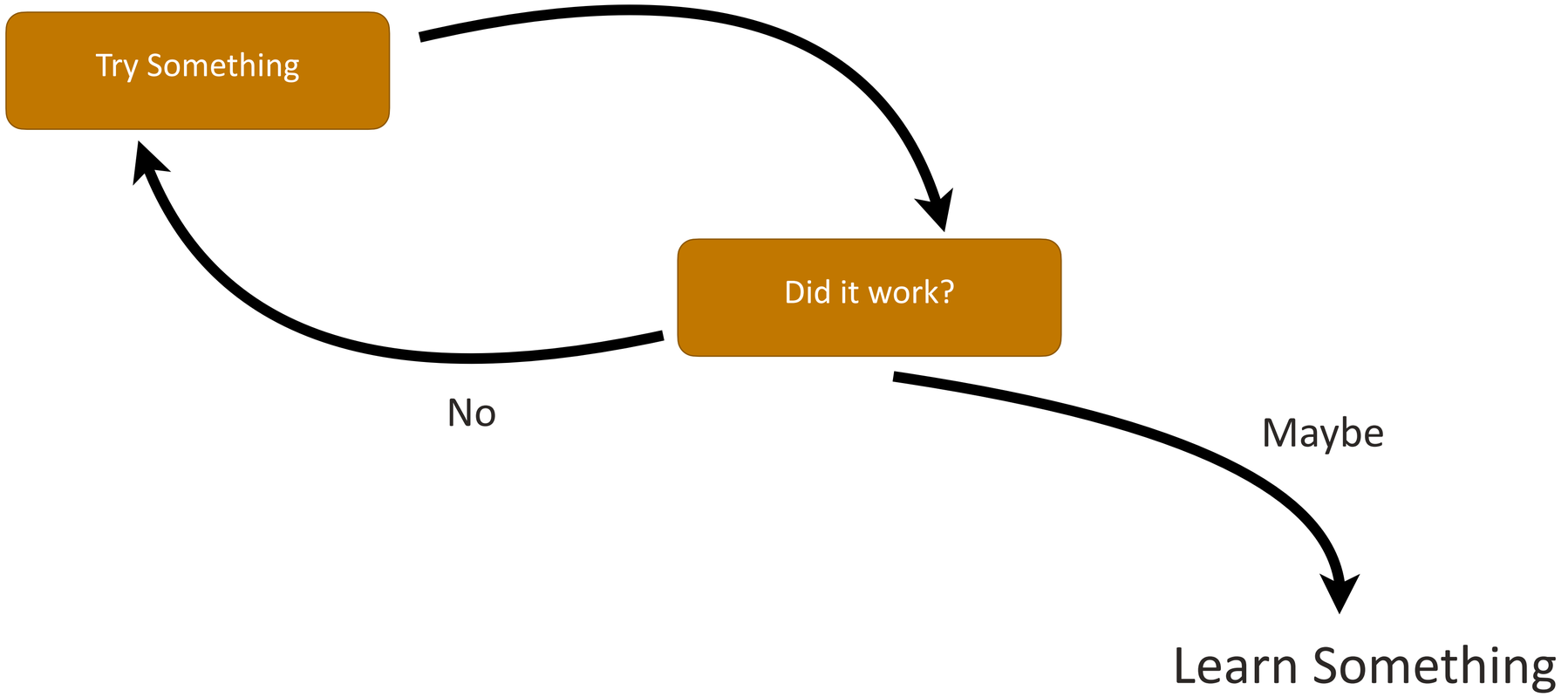
Analyst  Developer

Analyst + Developer

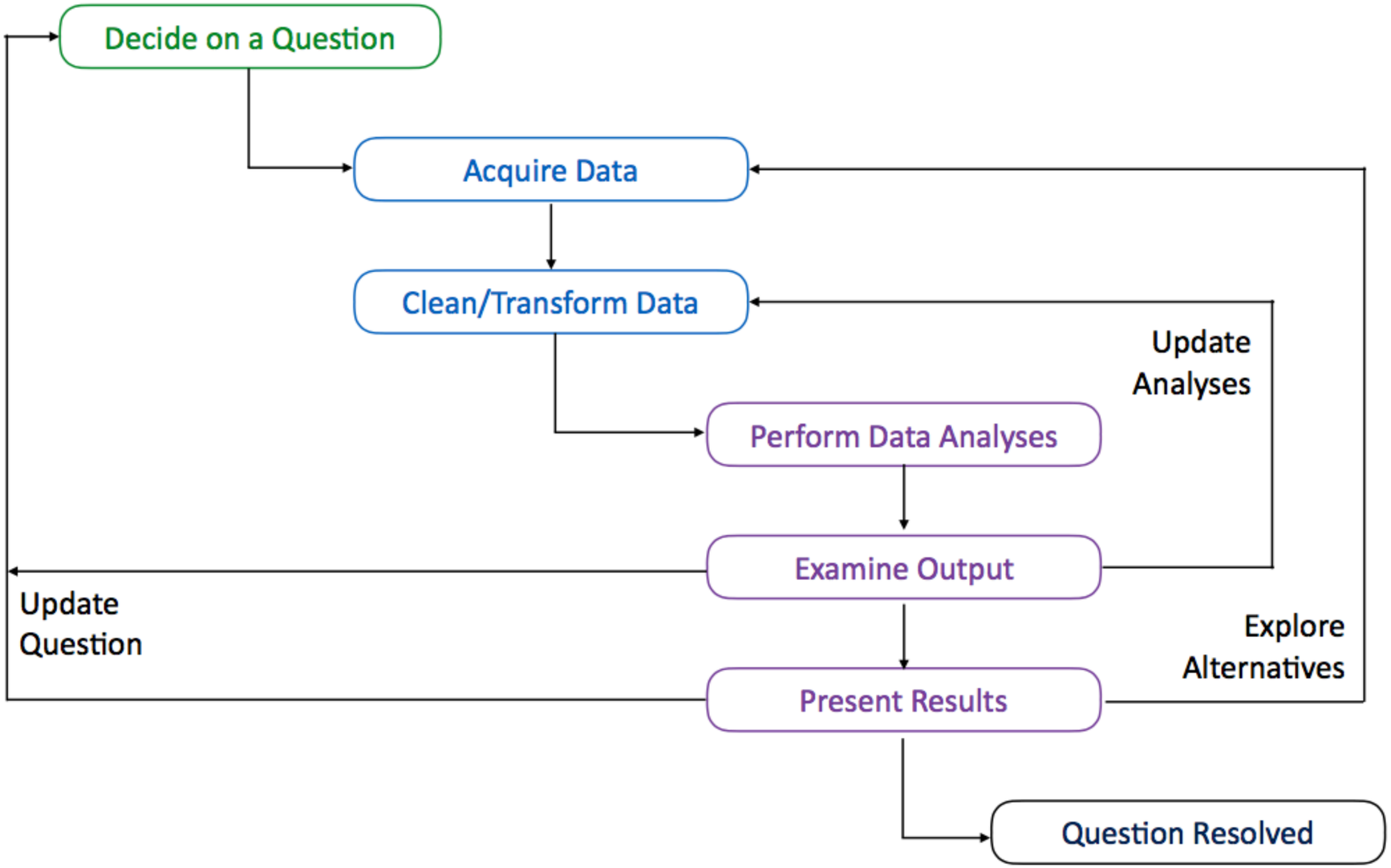
What Data Science is Not



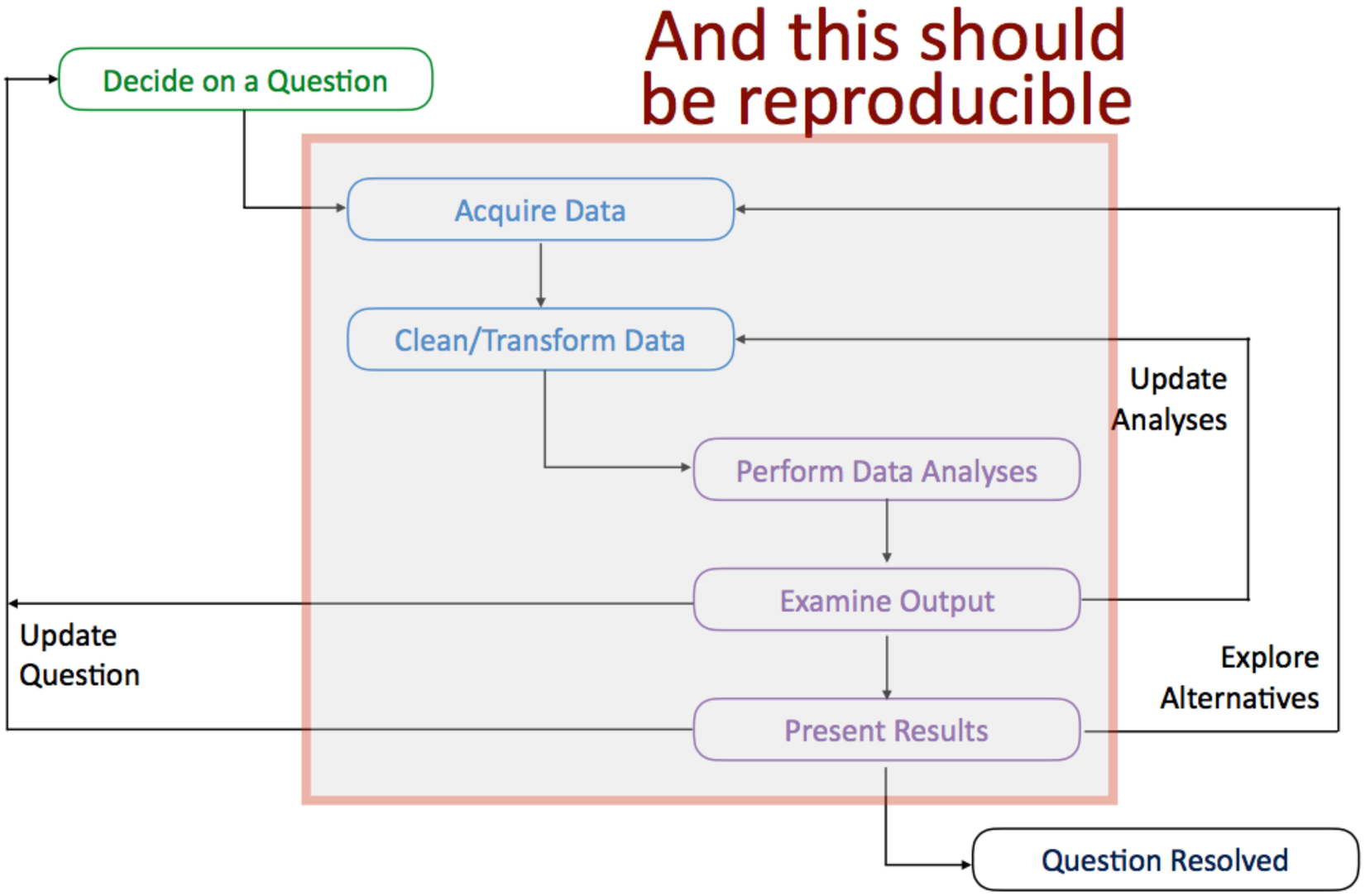
What Data Science Is



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.**

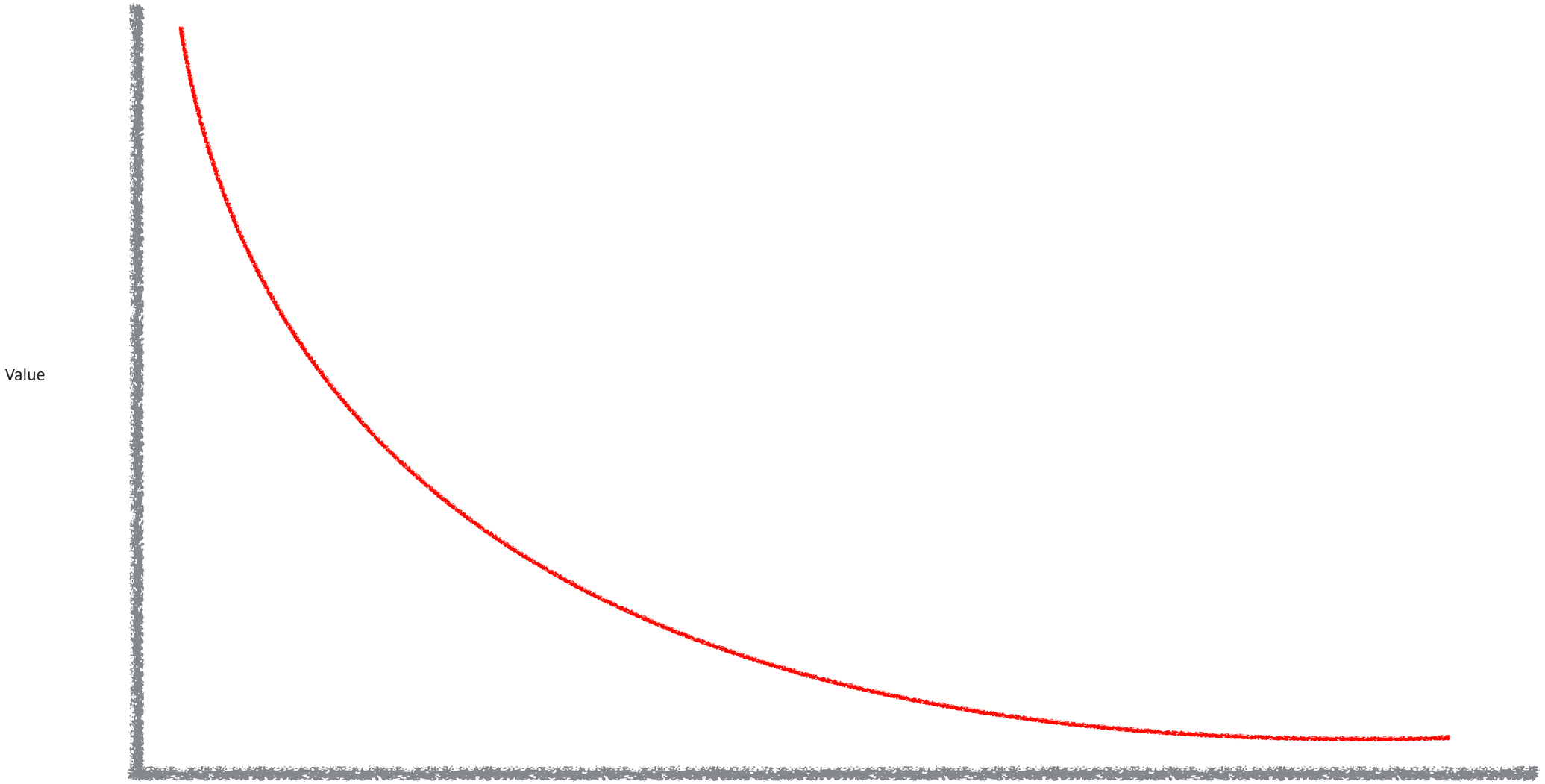
–Andrew Brust, “Data scientists don't scale”, http://www.zdnet.com/article/data-scientists-dont-scale/

Time to Insight

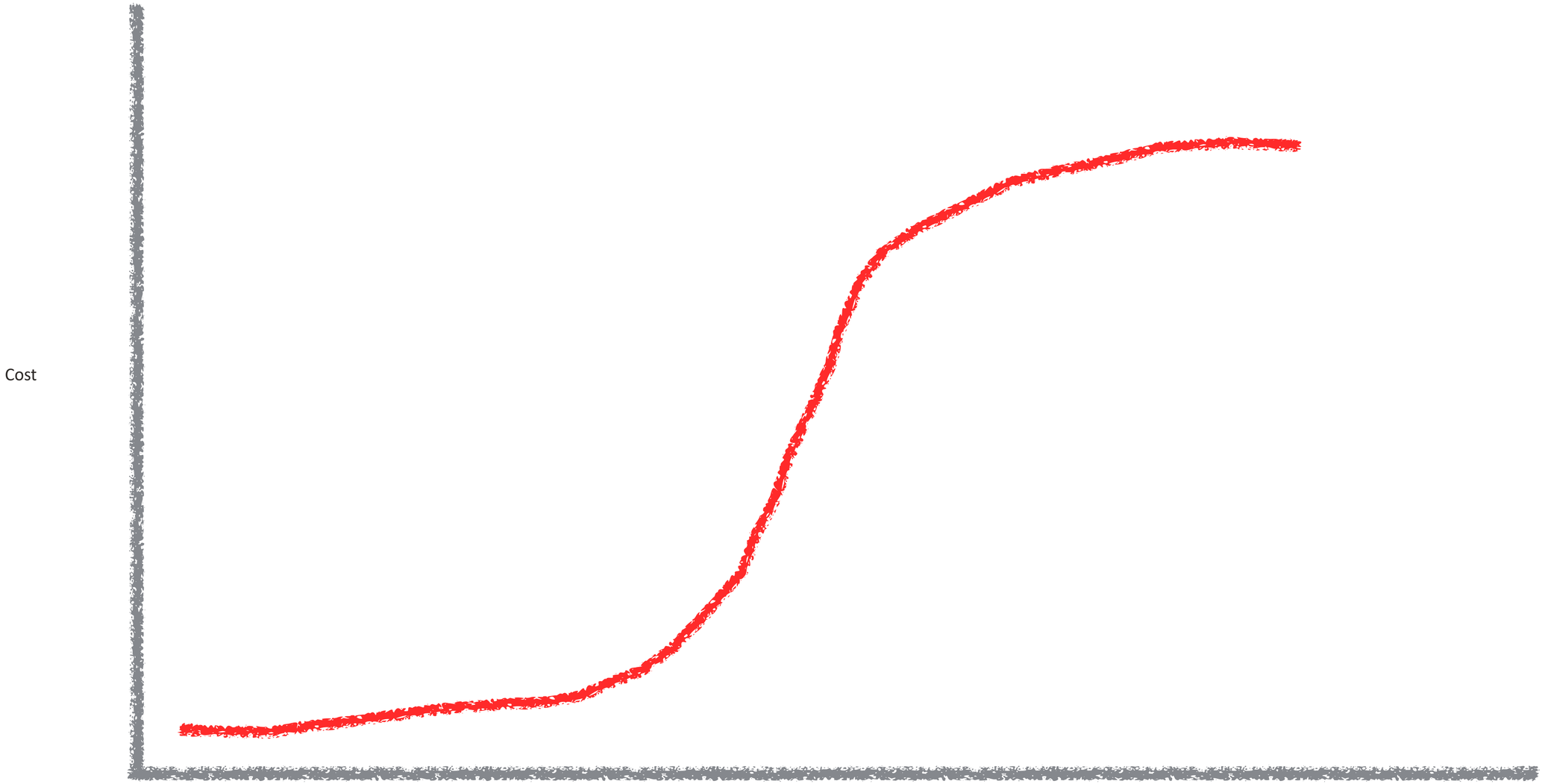
Time to Insight

Time = $$

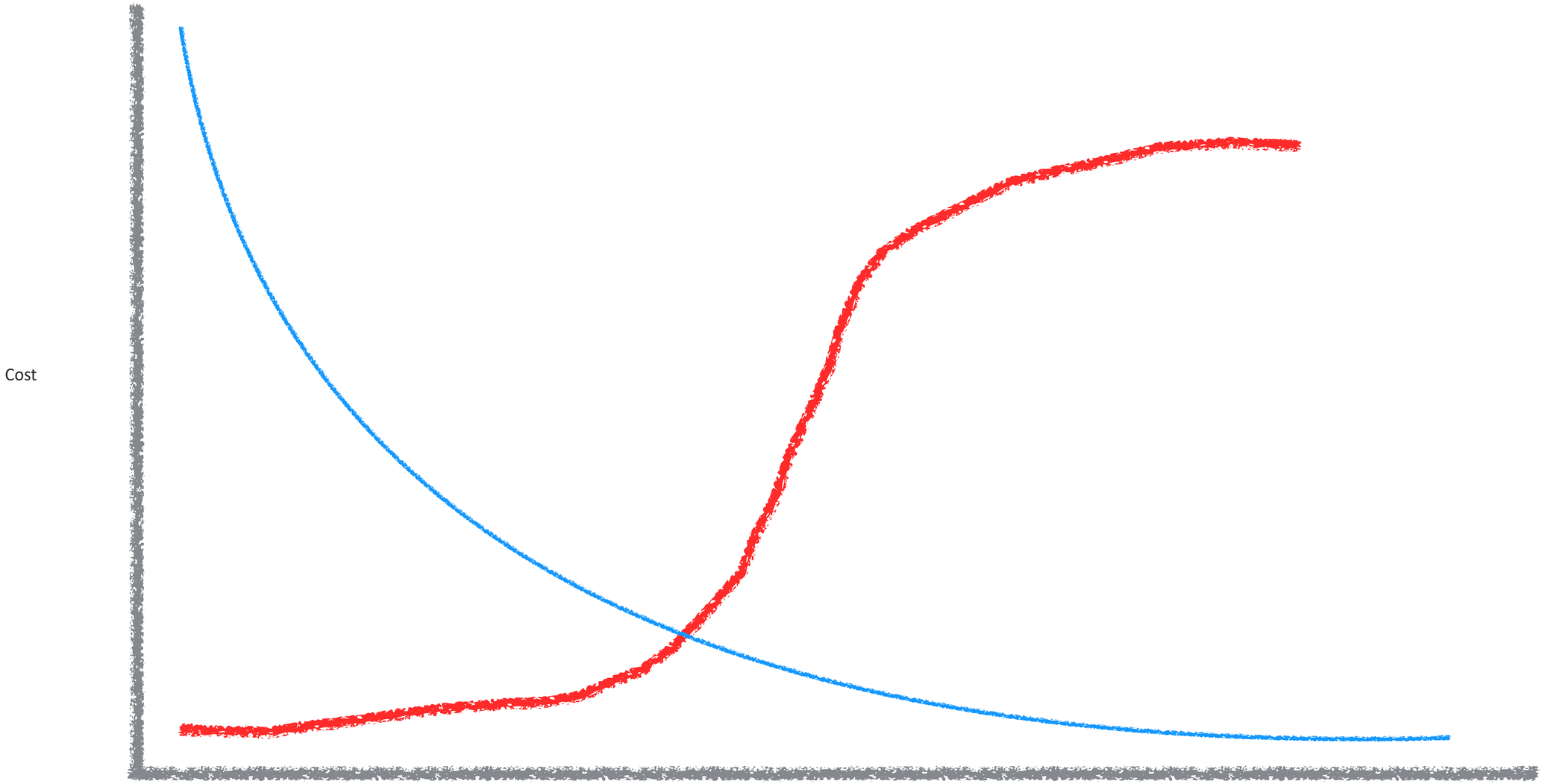
Value of Insights Over Time



Costs of Insights Over Time

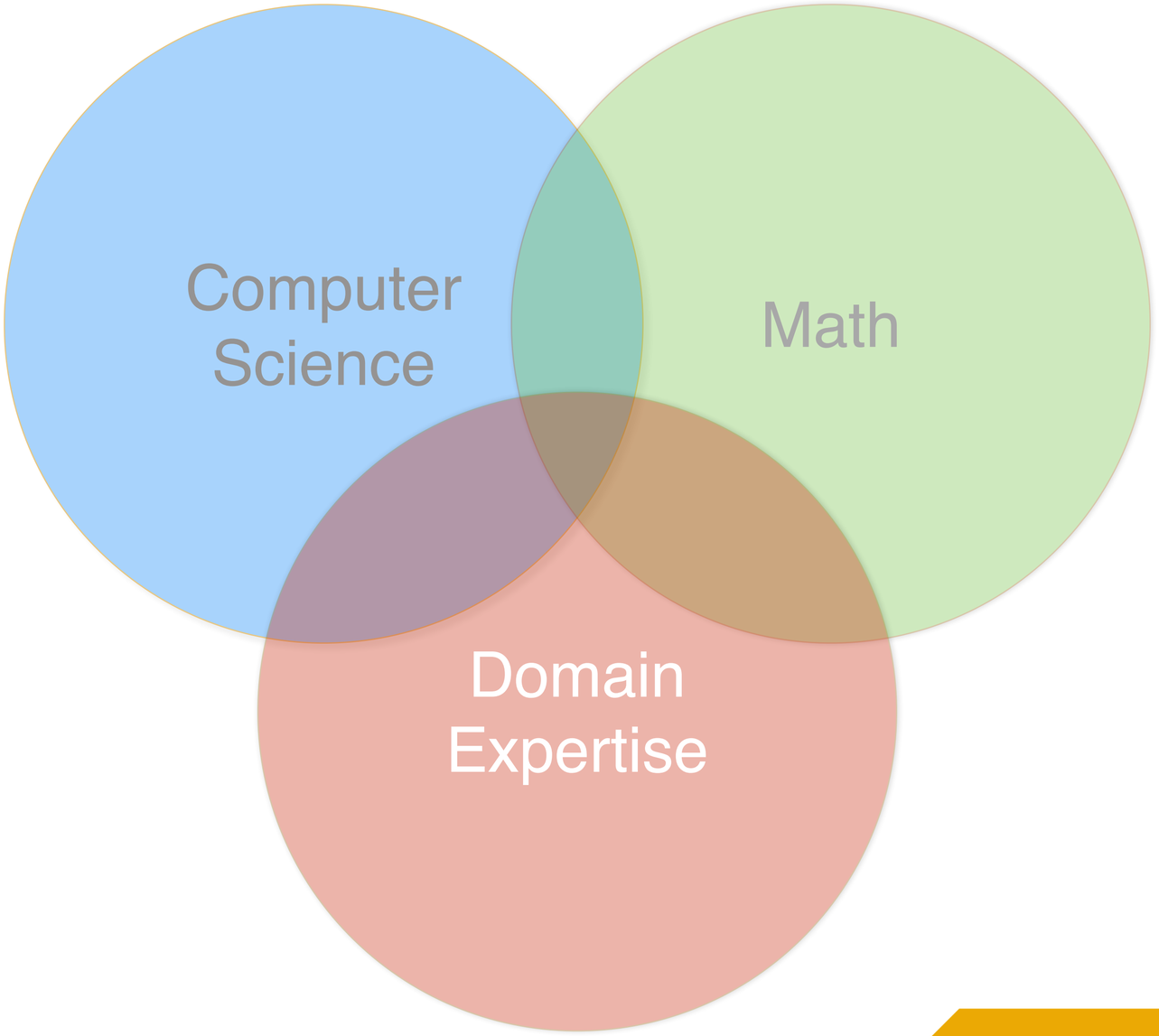


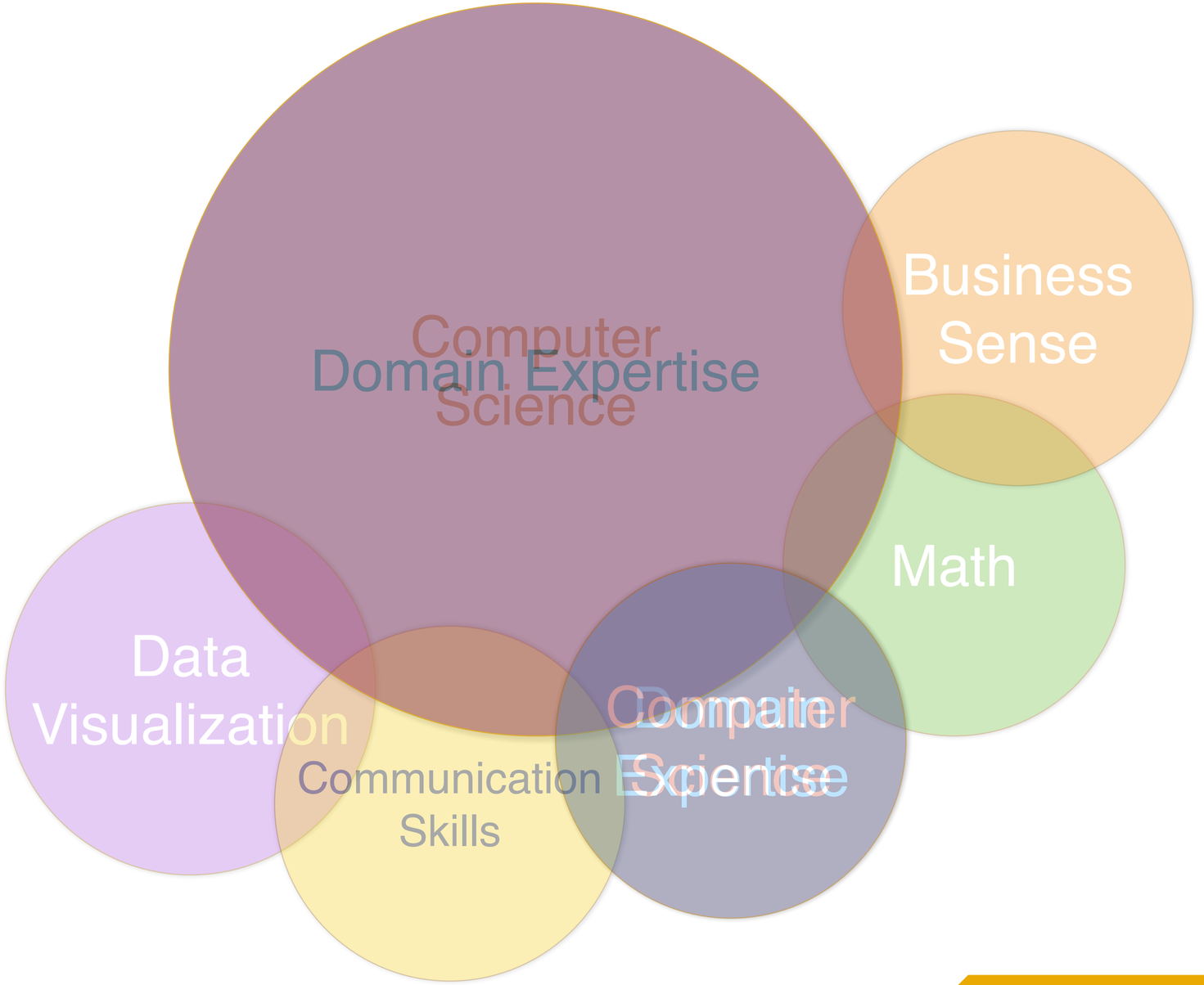
Cost of Insights Over Time

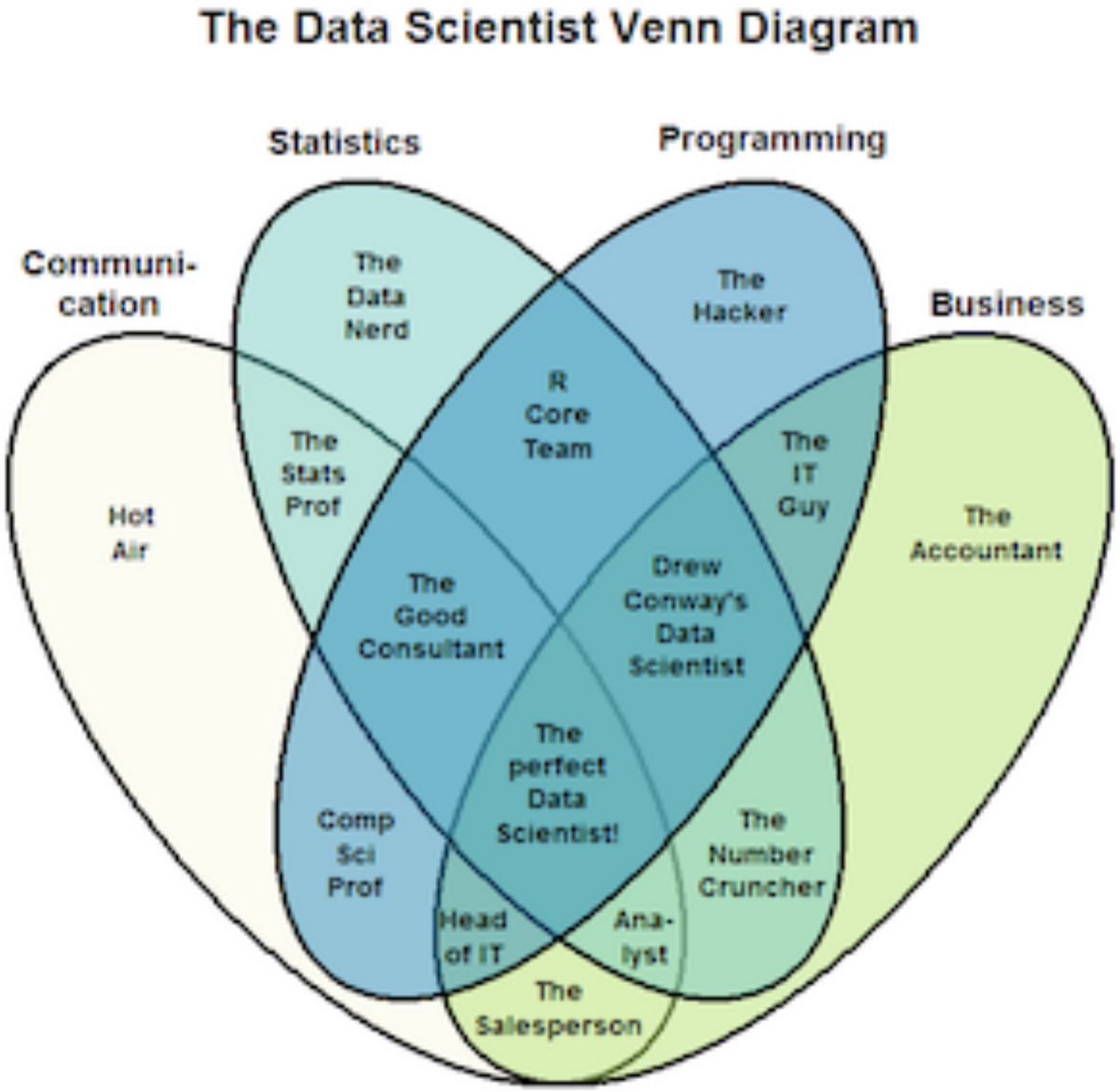


Stop

What Skills Does a Data Scientist Need?


Computer
Science
Math
Domain
Expertise

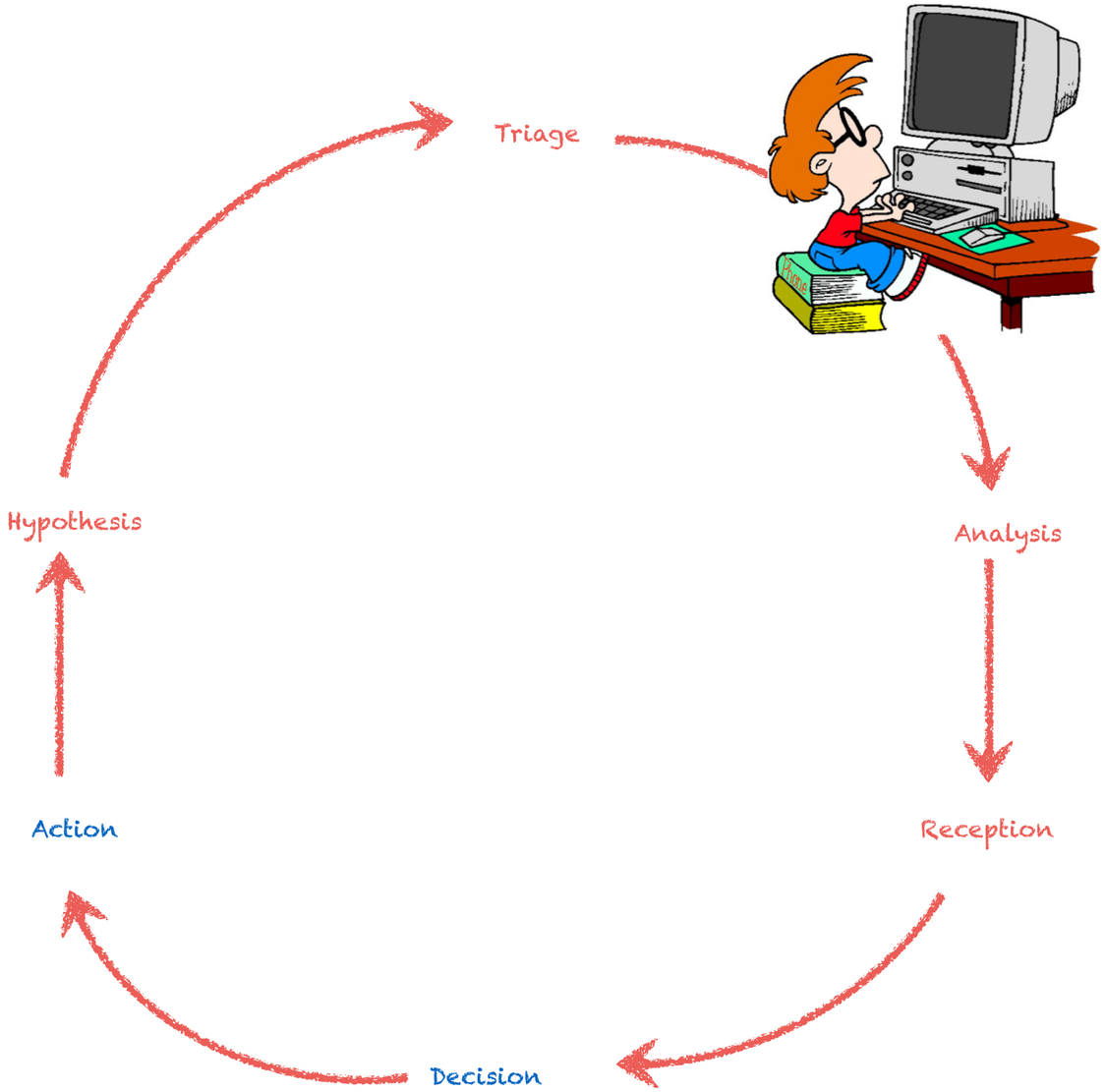

Computer
Domain Expertise
Science
Business
Sense
Math
Data
Visualization
Communication
Skills
Computer
Expertise



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

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

Data Janitors


Triage
Analysis
Reception
Decision
Action
Hypothesis

Thoughts for Data Science Success

Data is a Strategic Asset... not a cost

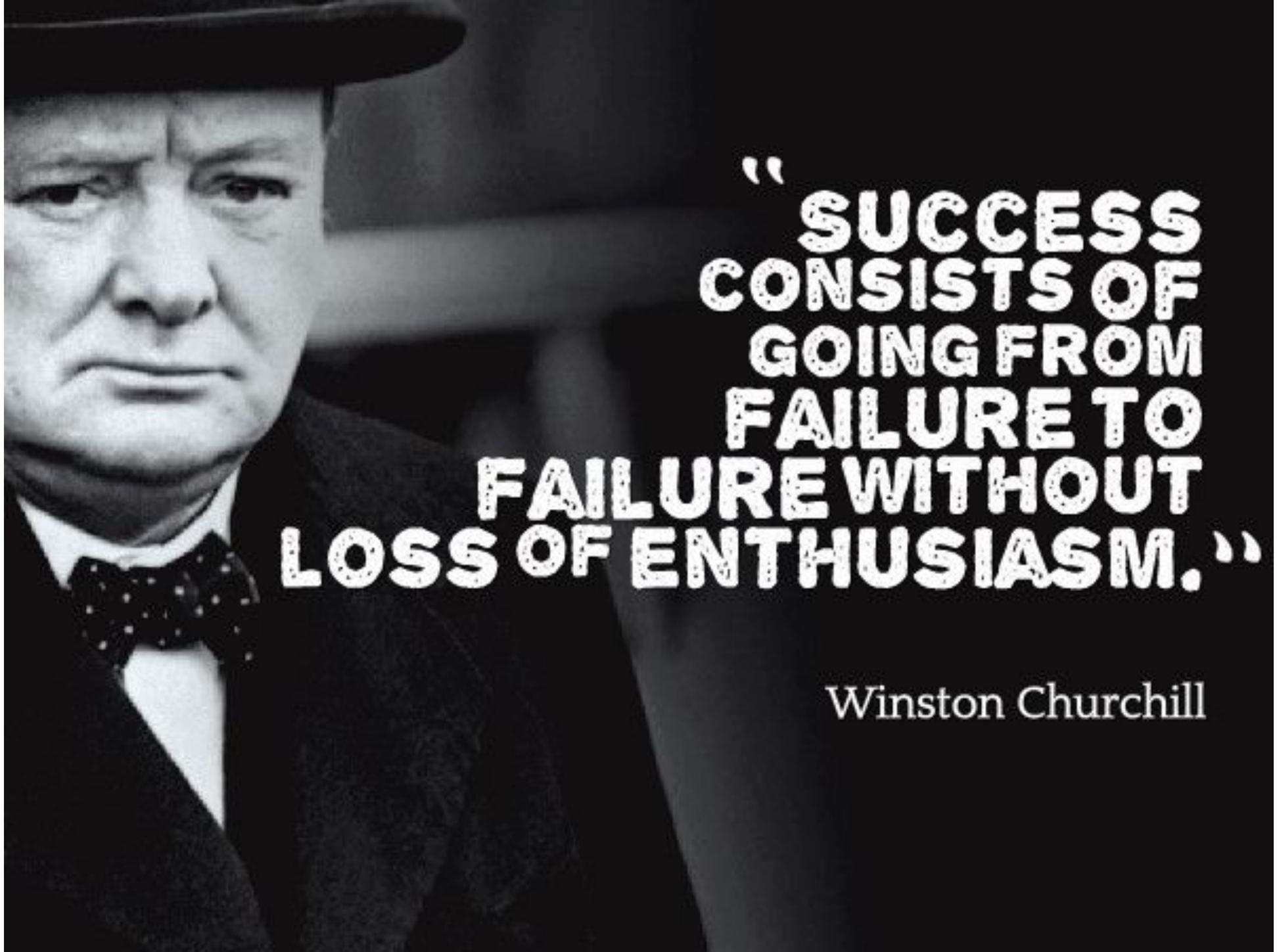
Align Projects to Corporate Strategy

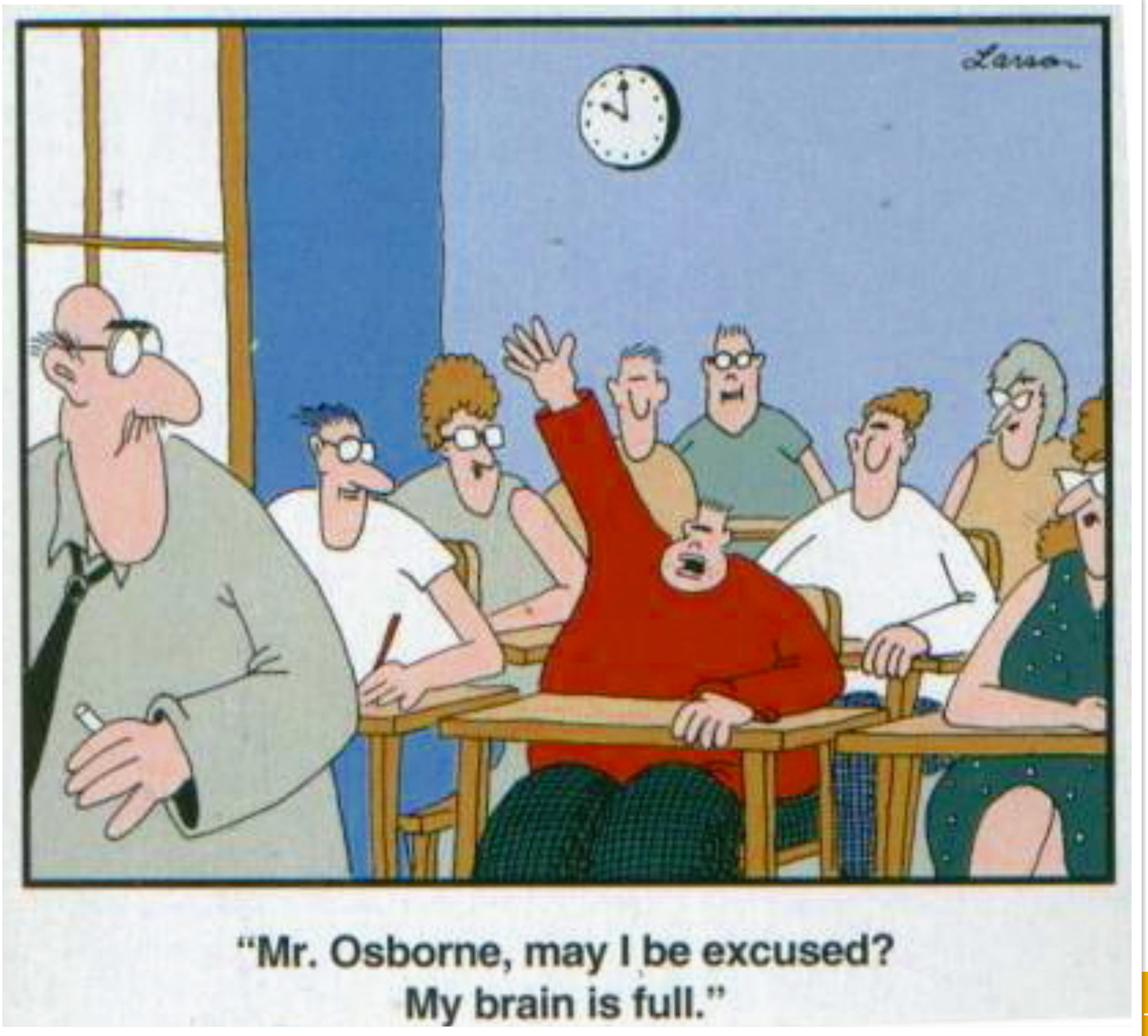
Align Projects to Corporate Strategy

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


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

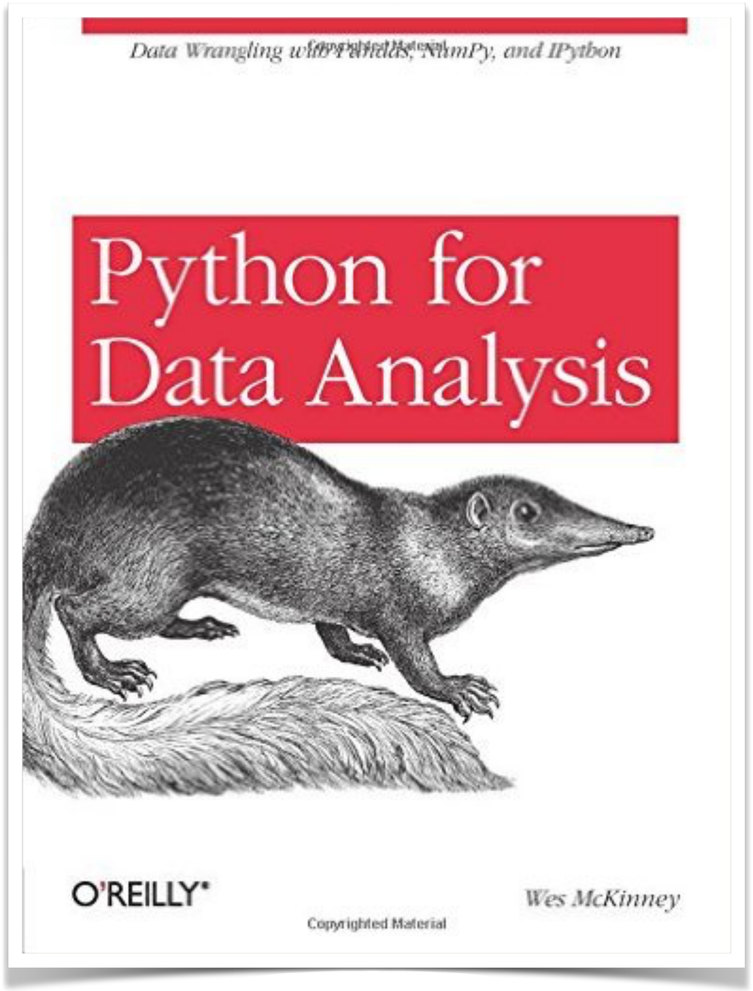
By The End of the Class, You Will Be Able To:

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

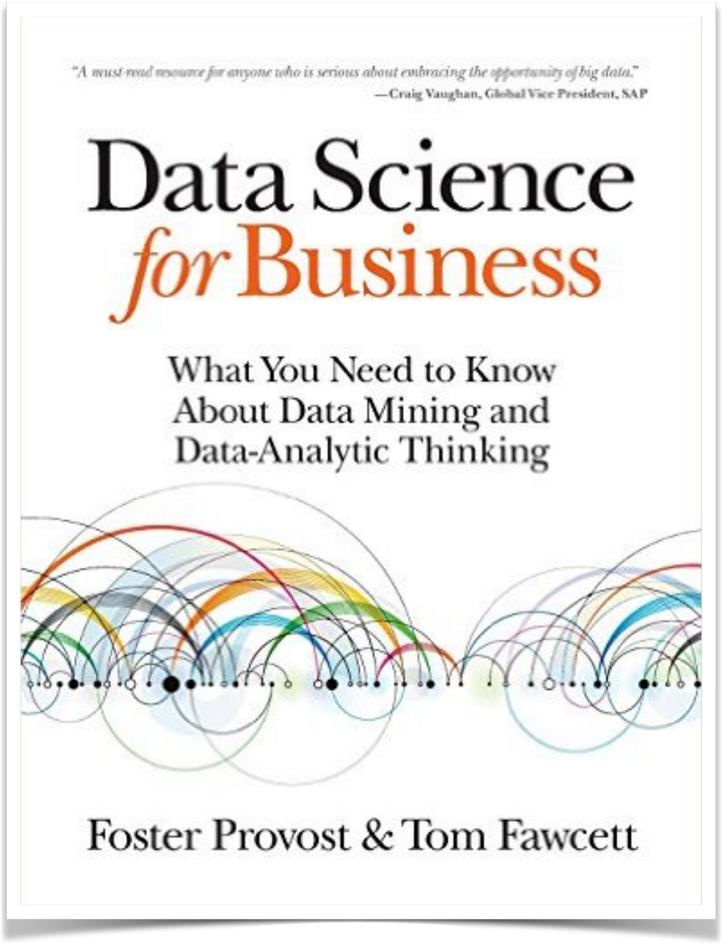

I THINK WE SHOULD BUILD AN SQL DATABASE.
UH-OH
E-mail: SCOTTADAMS@AOL.COM
DOES HE UNDERSTAND WHAT HE SAID OR IS IT SOMETHING HE SAW IN A TRADE MAGAZINE AD?
© 1996 United Feature Syndicate, Inc.(NYC)
WHAT COLOR DO YOU WANT THAT DATABASE?
I THINK MAUVE HAS THE MOST RAM.


Buzzword

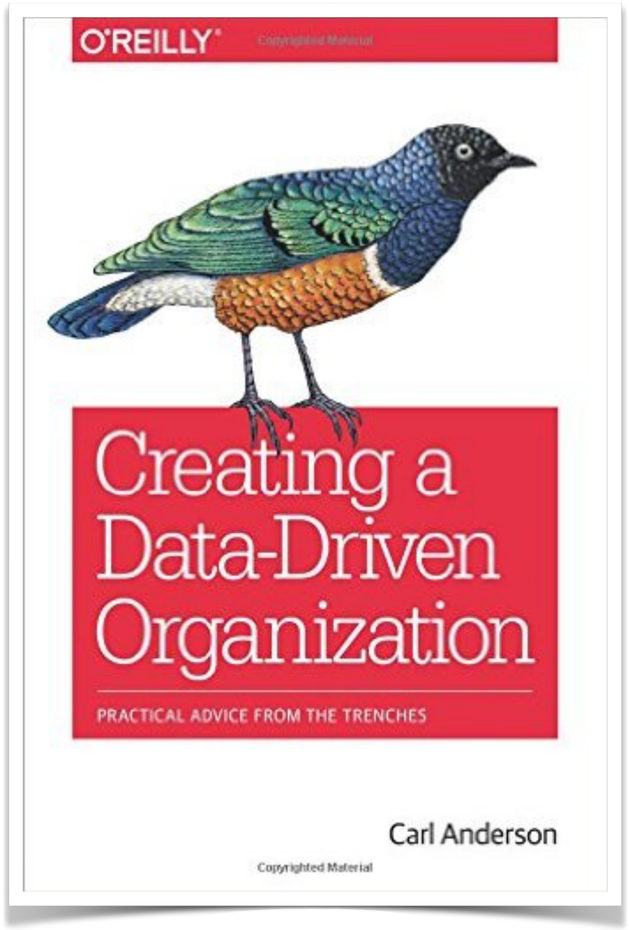
Recommended Reading



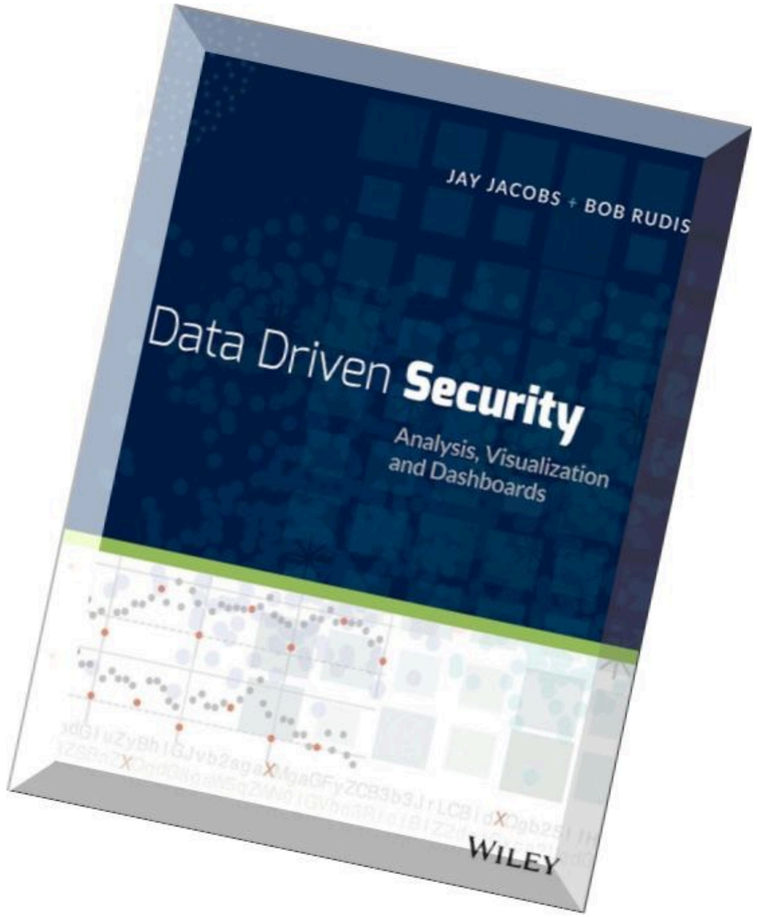
Recommended Reading



Recommended Reading



Recommended Reading



http://datadrivensecurity.info

Stop

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

“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


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


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

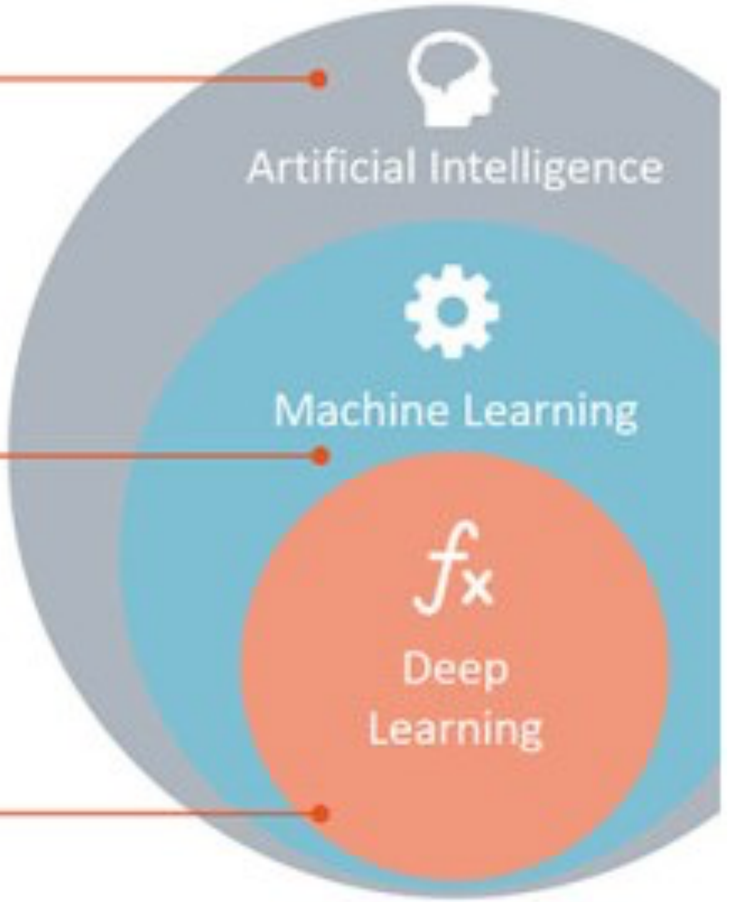
Any technique which enables computers to mimic human behavior.

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.

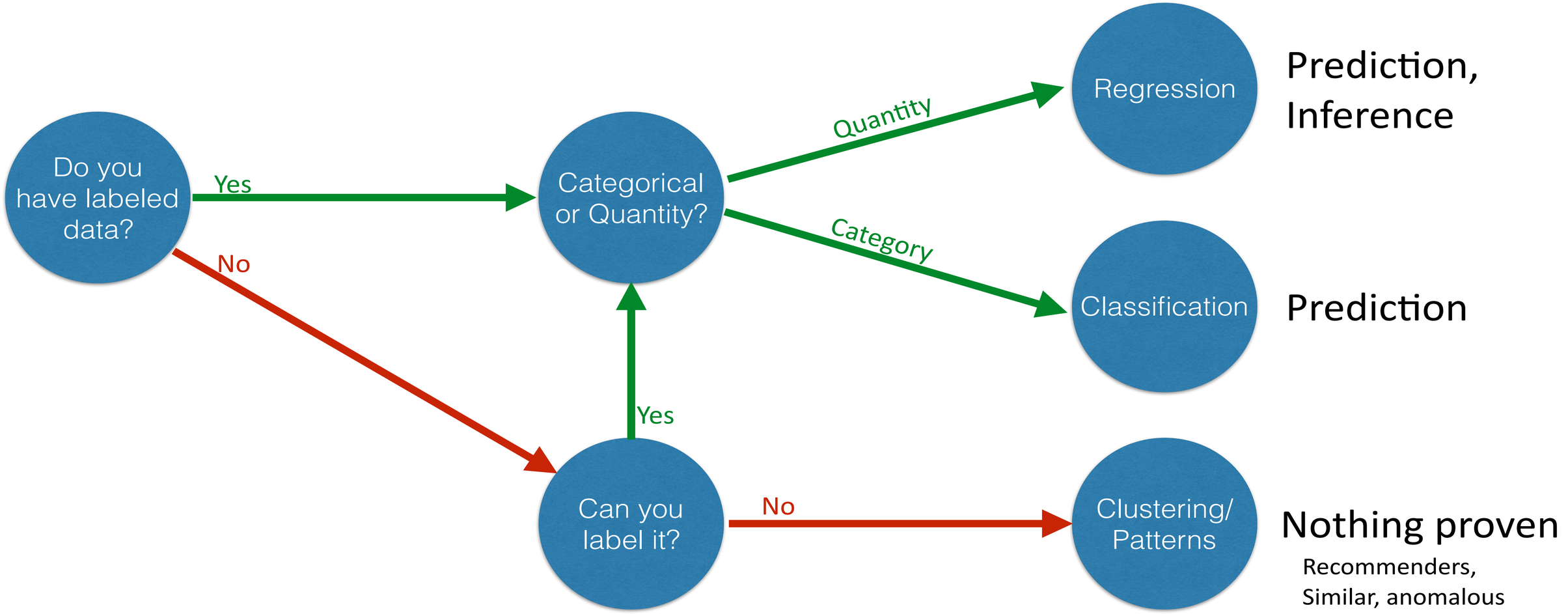
Machine Learning Problems

- **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.

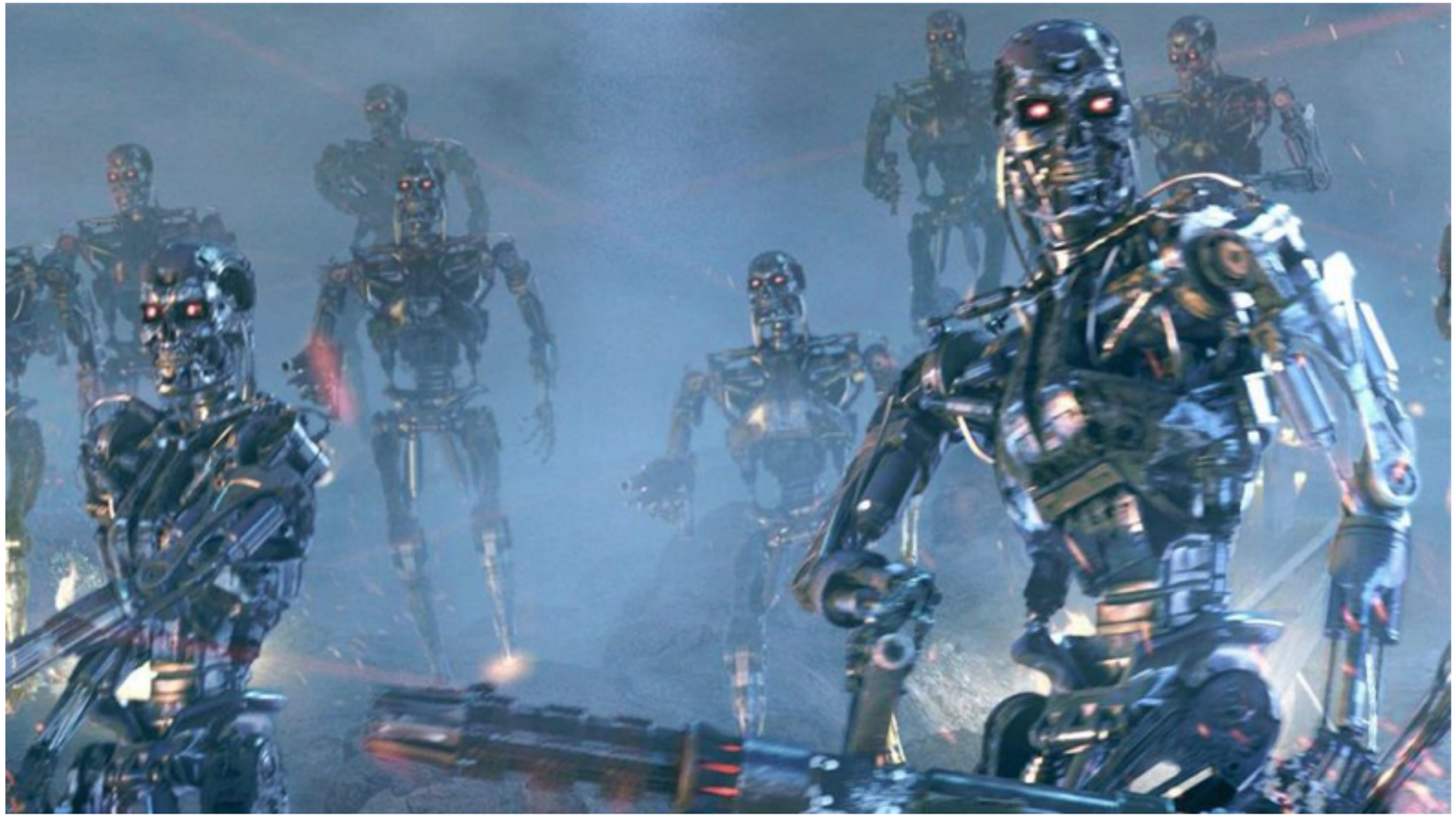
Machine Learning Problem Types

- **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



Stop

Applications to Security

Regression Example

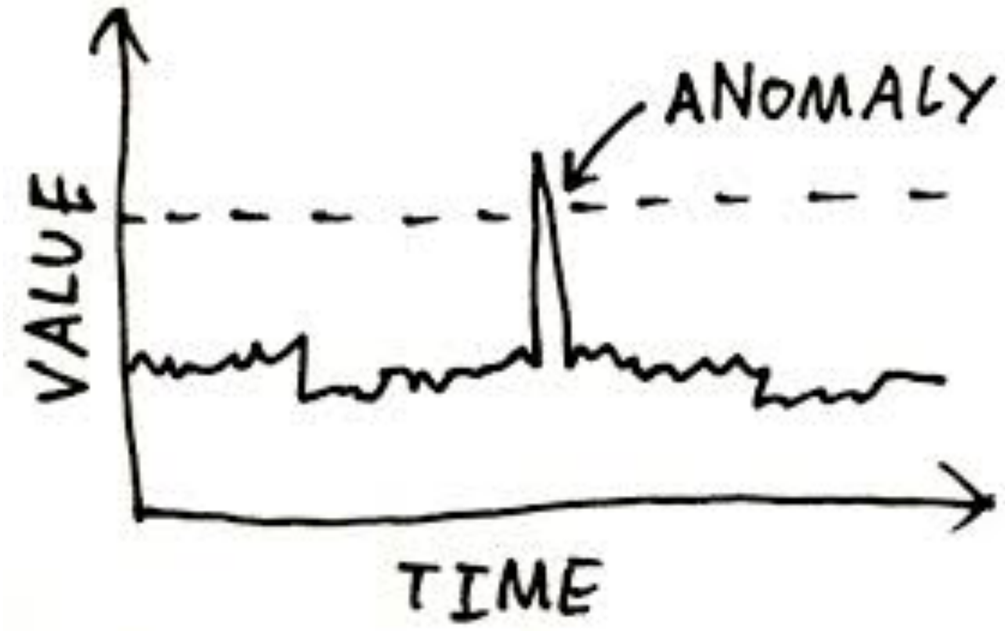
Server Capacity Prediction: Regression analysis can be used to predict a server's capacity (or CPU usage) based on the server's historical performance.

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/

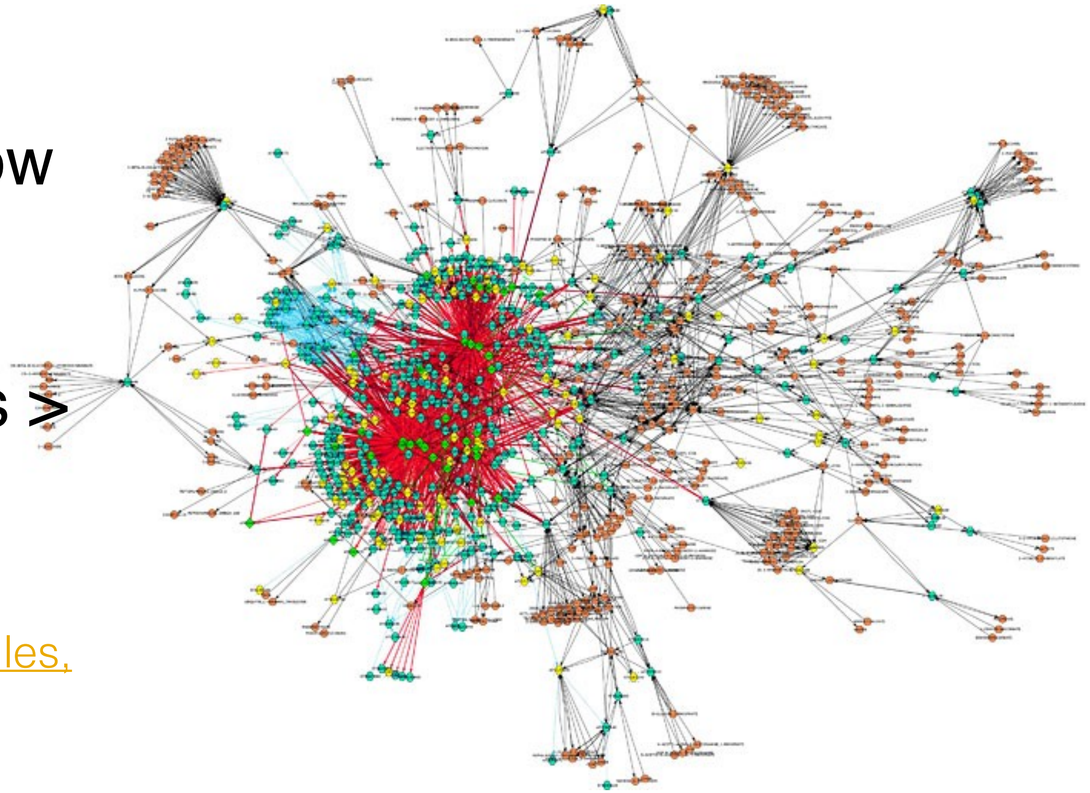
Clustering Example

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/,



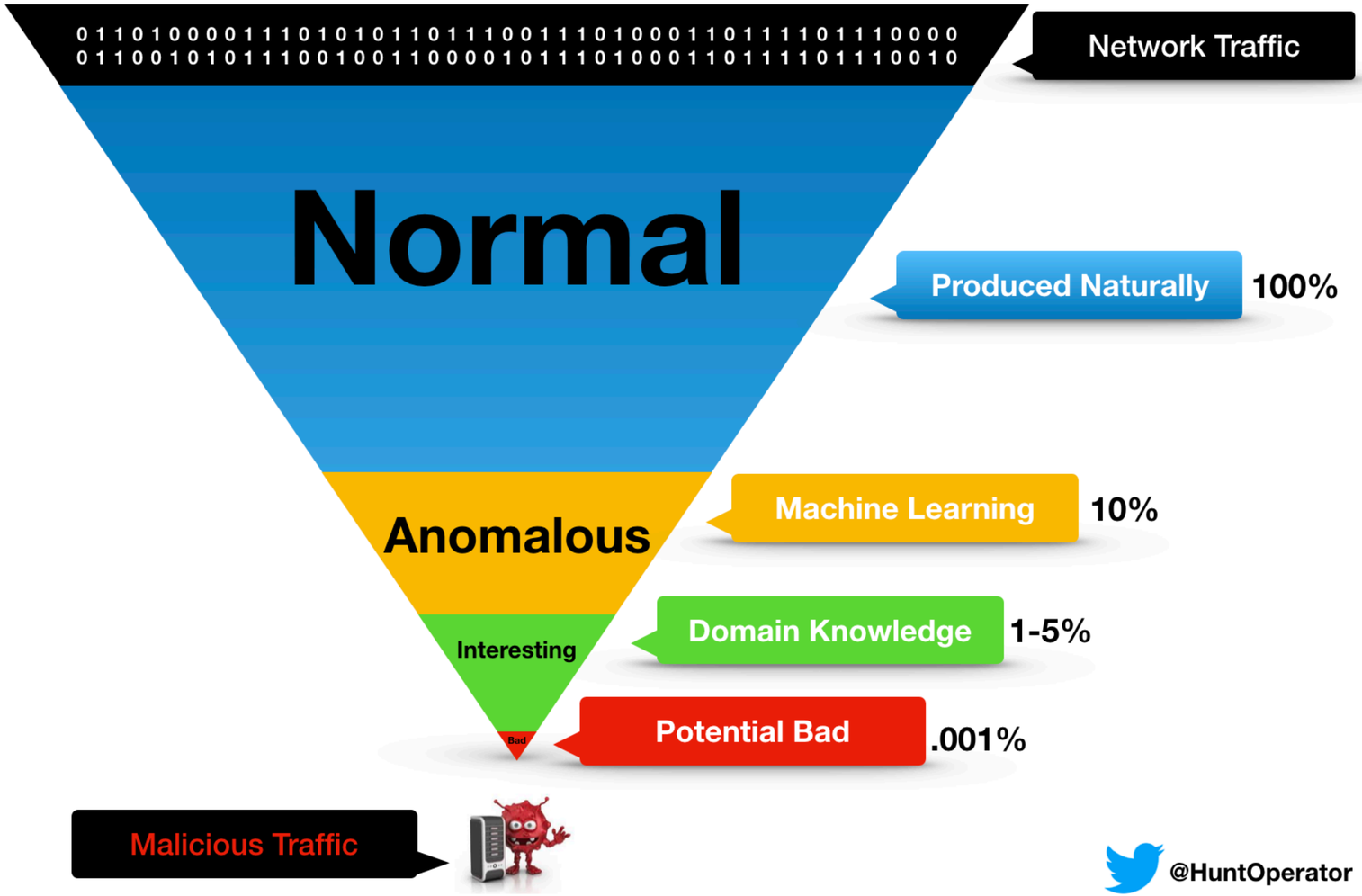
Malware Detection/Classification

- 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

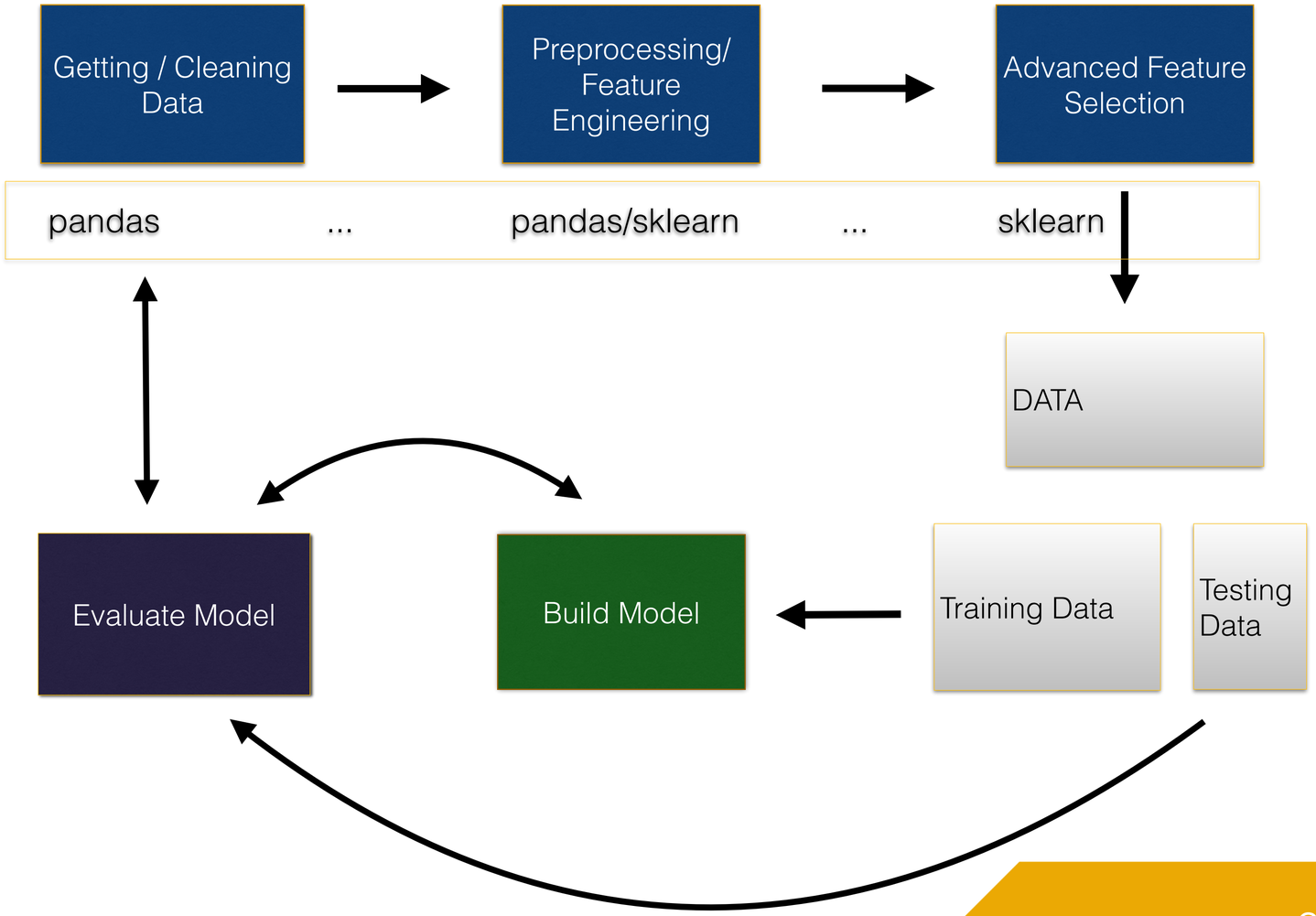


http://www.austintaylor.io/network/traffic/threat/data/science/hunting/funnel/machine/learning/domain/expertise/2017/07/11/data-science-hunting-funnel/

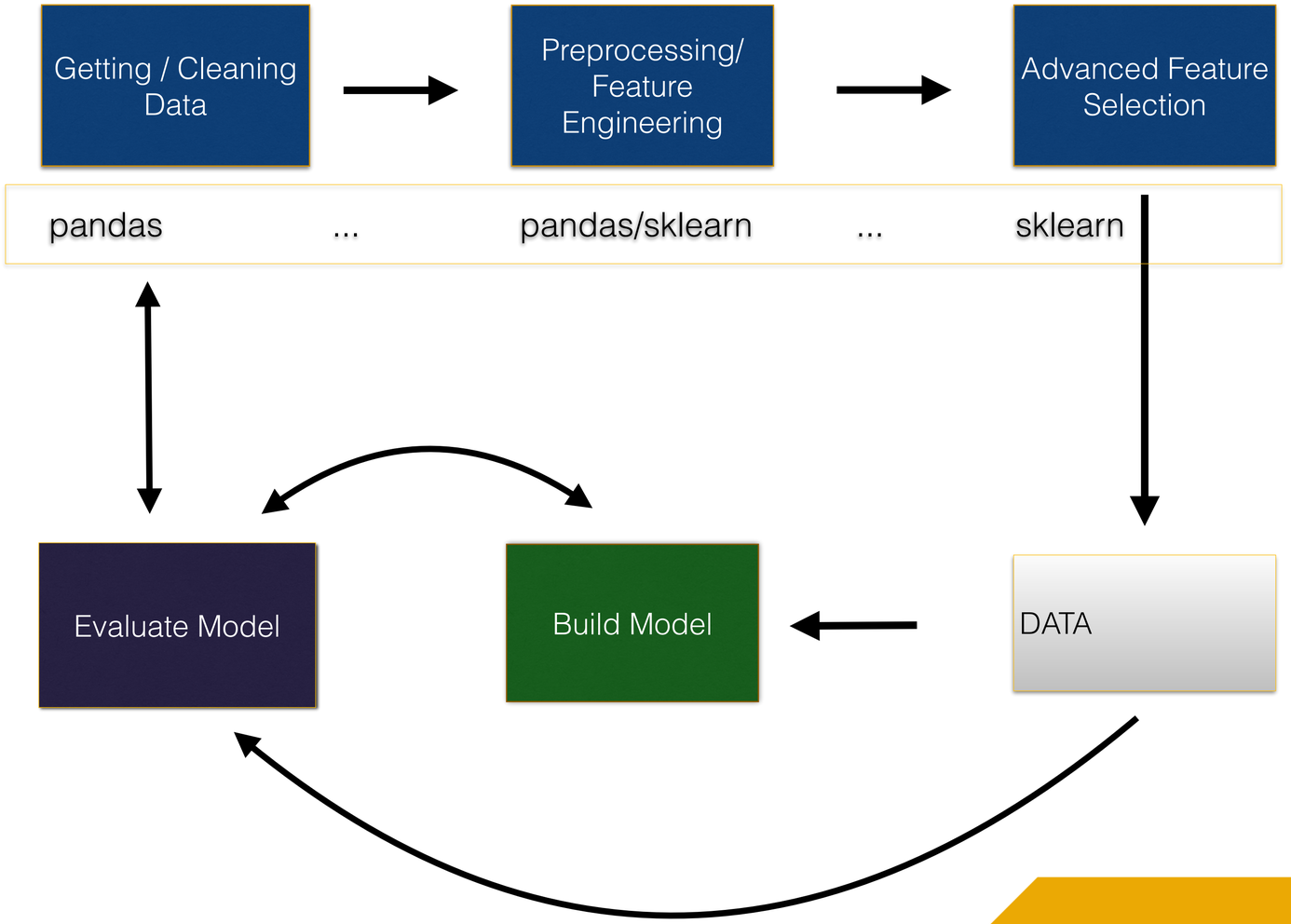
Stop

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?

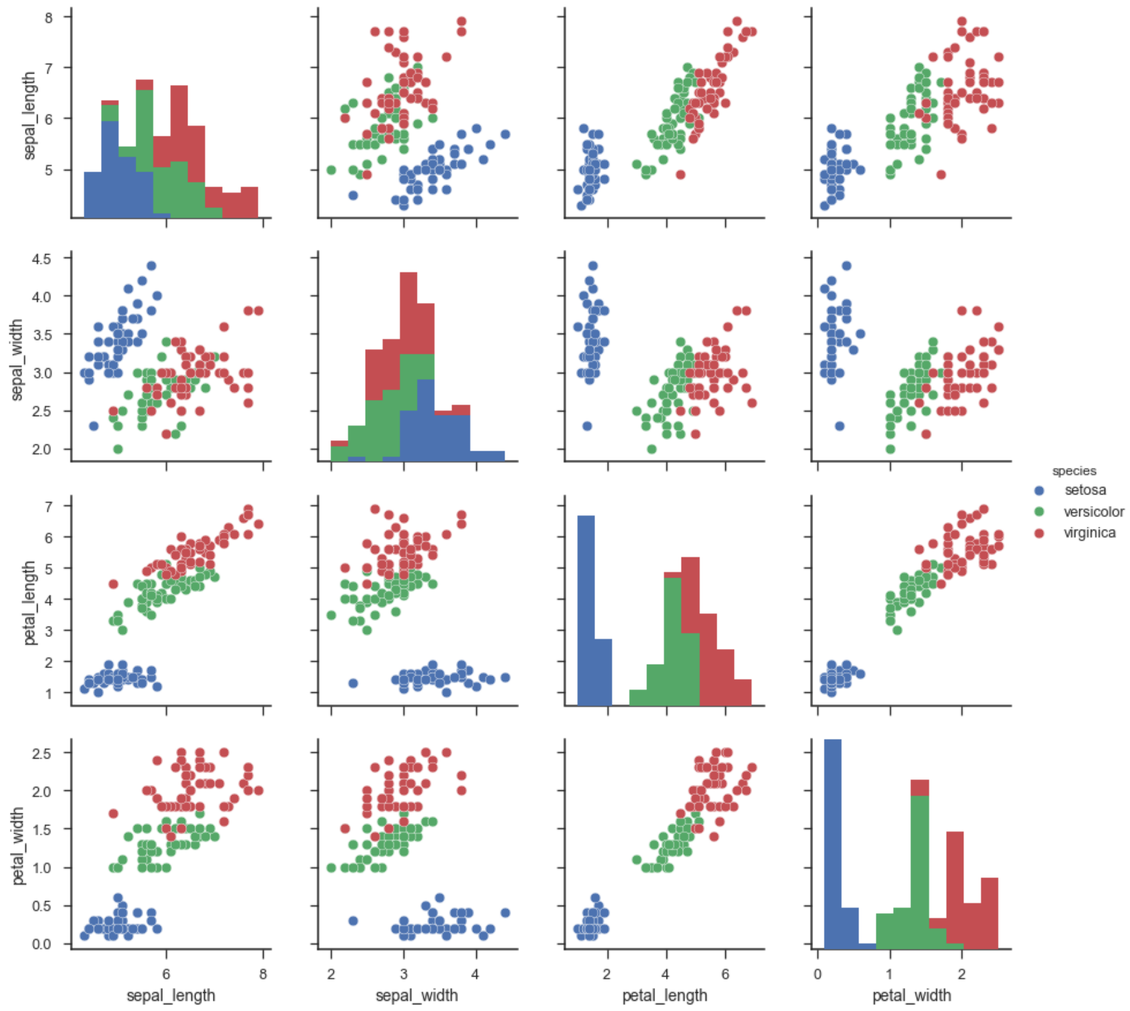
Choose data sources

- 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?

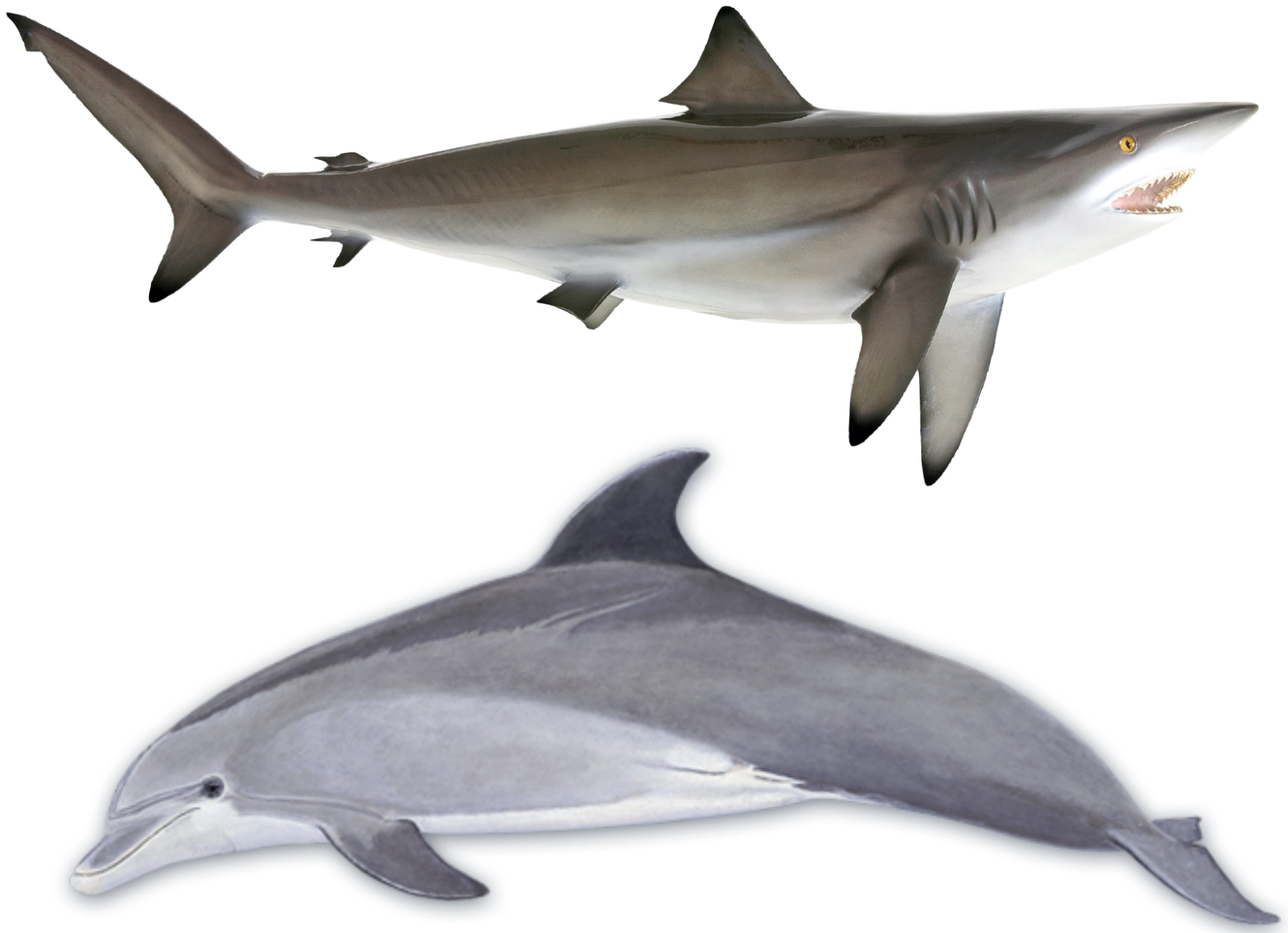
Feature Engineering

- 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

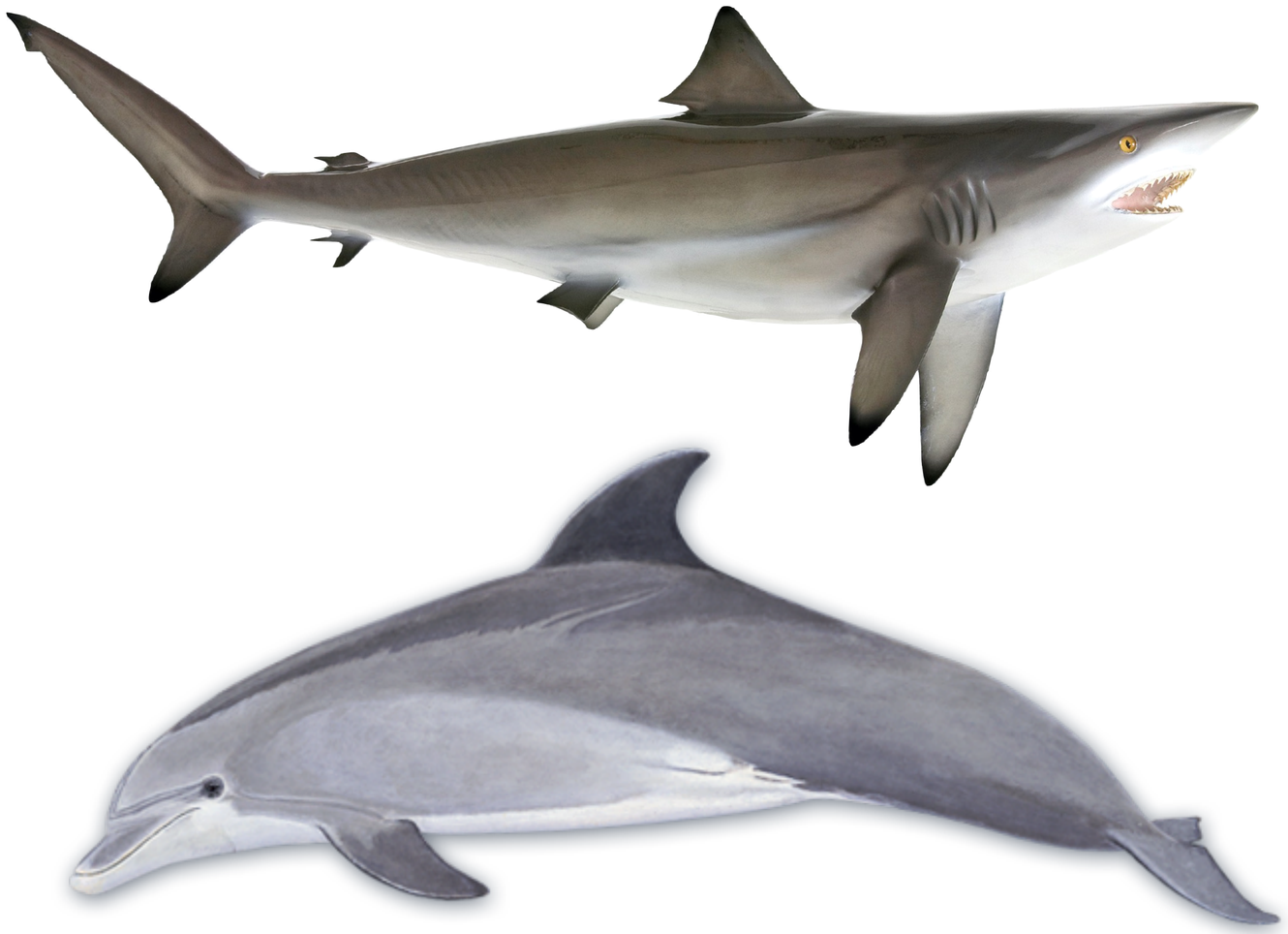
Feature	Value
Color	Gray
Fins	7
Predator	TRUE

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 or Elasticsearch.

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)

Group Discussion

Consider that you are building a system to identify fraudulent credit card transactions. In your groups, try to answer the following questions:

1. What are some features that you would want to capture?
2. What data sets will you need?
3. What legal and policy challenges might you face?
4. What other challenges you could foresee in this problem?
5. How will you define success?
6. How can you articulate the value of this model to stakeholders?

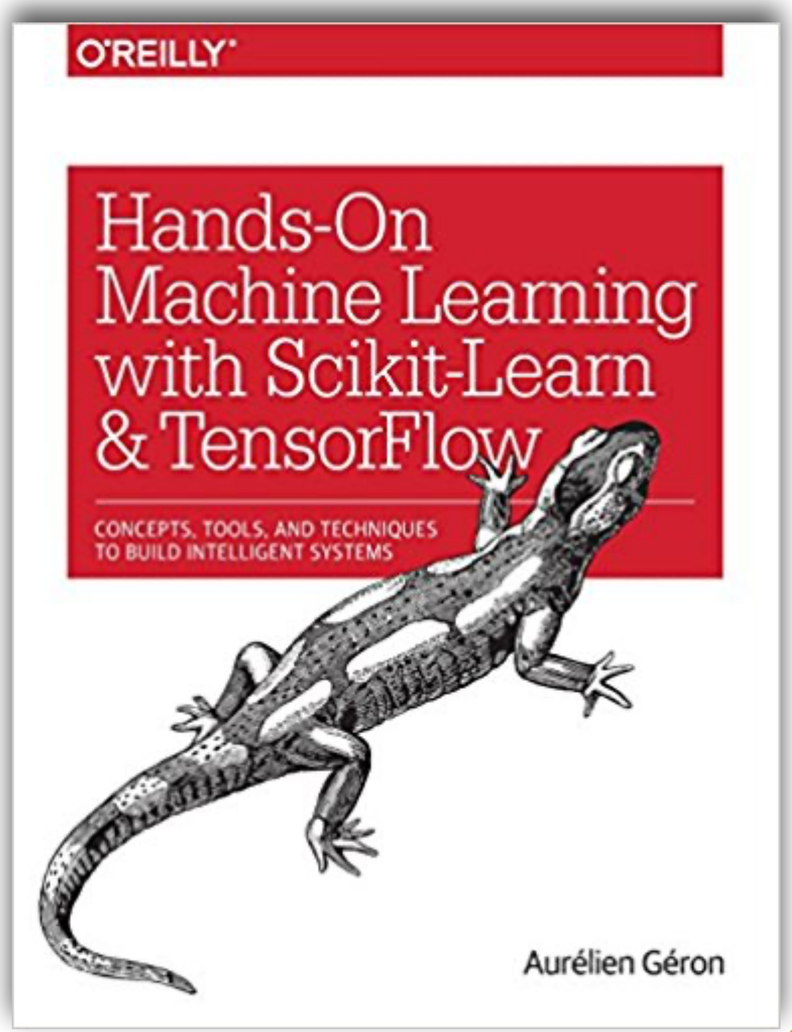
Stop

The Python Data Science Ecosystem

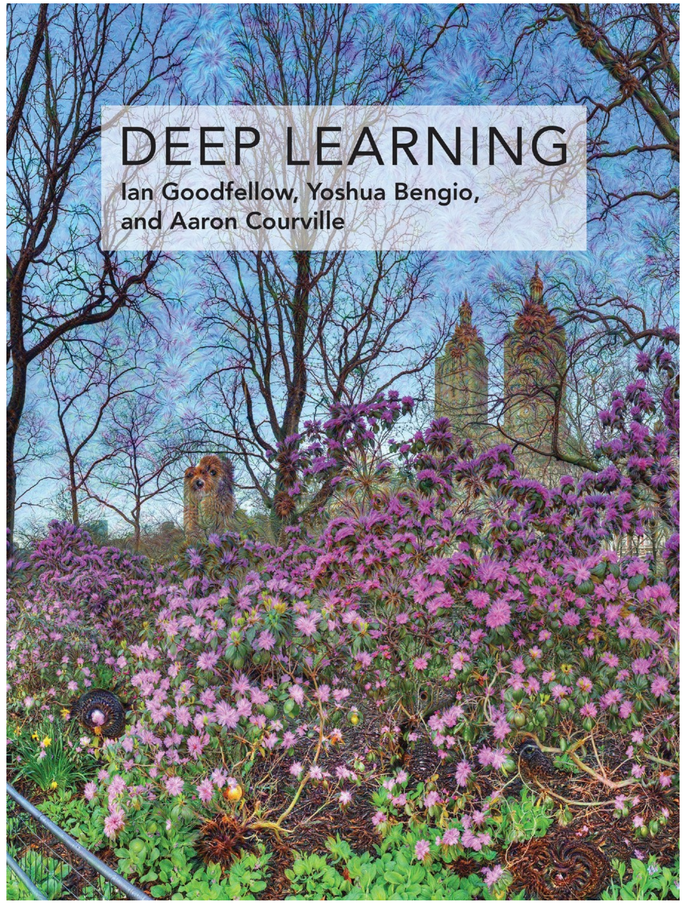
Machine Learning Ecosystem

- **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,

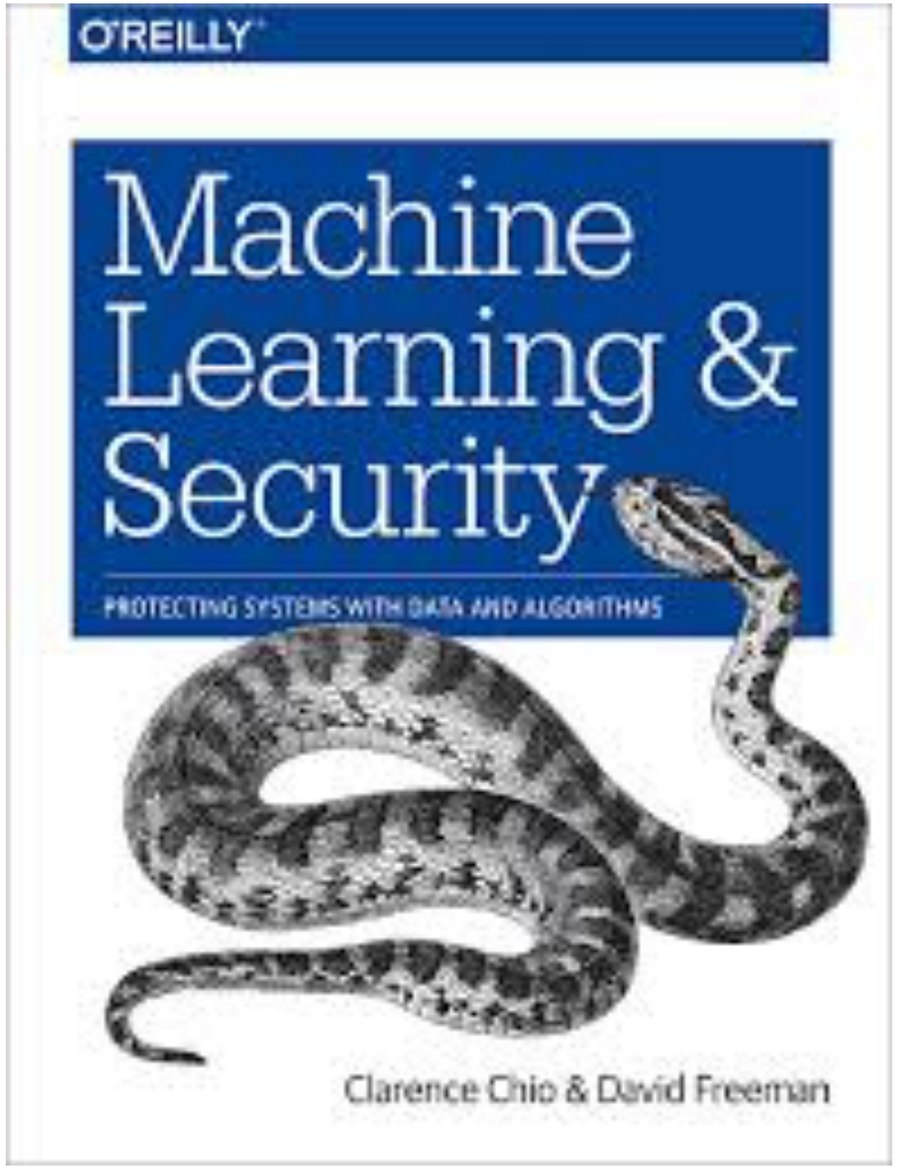
Recommended Reading

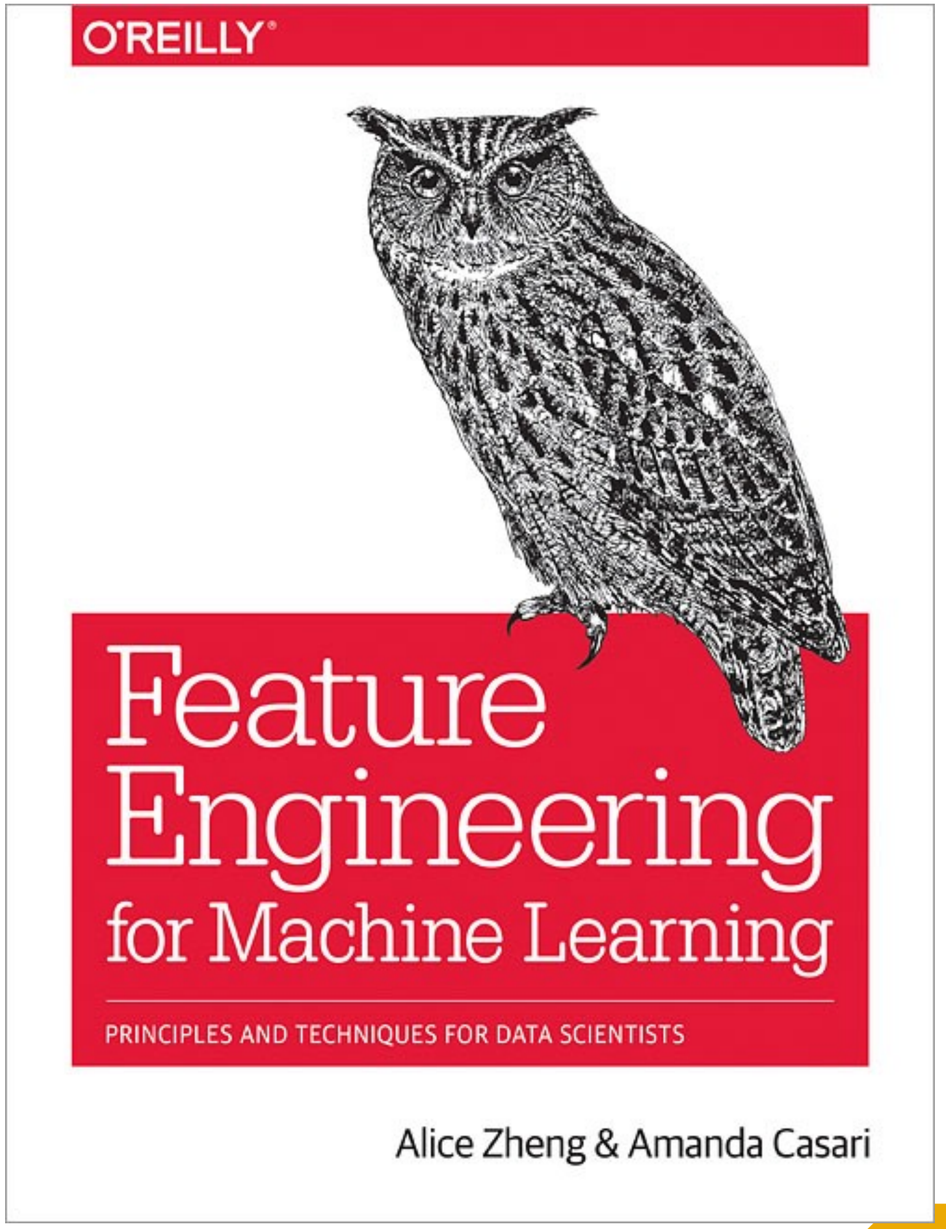


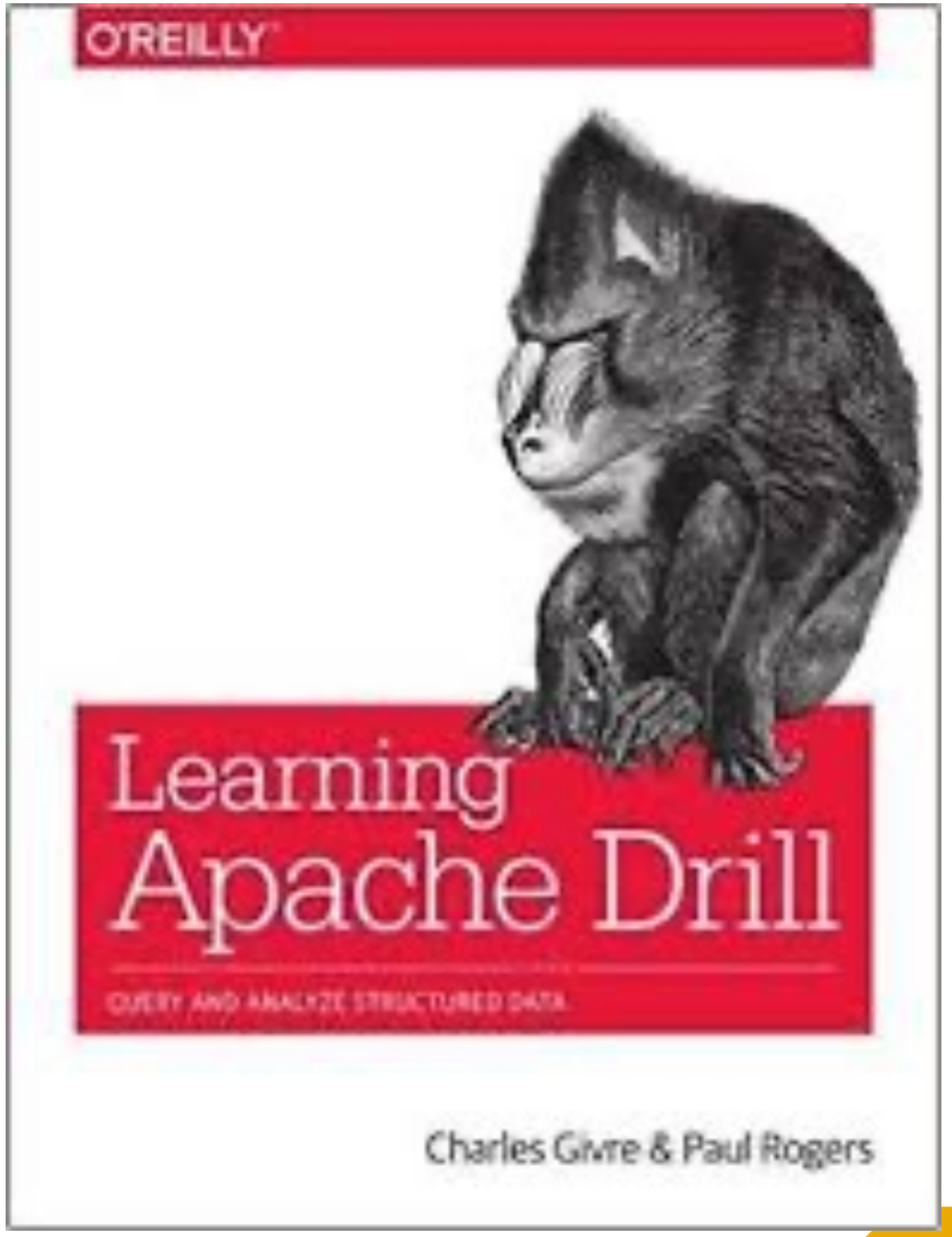
Recommended Reading



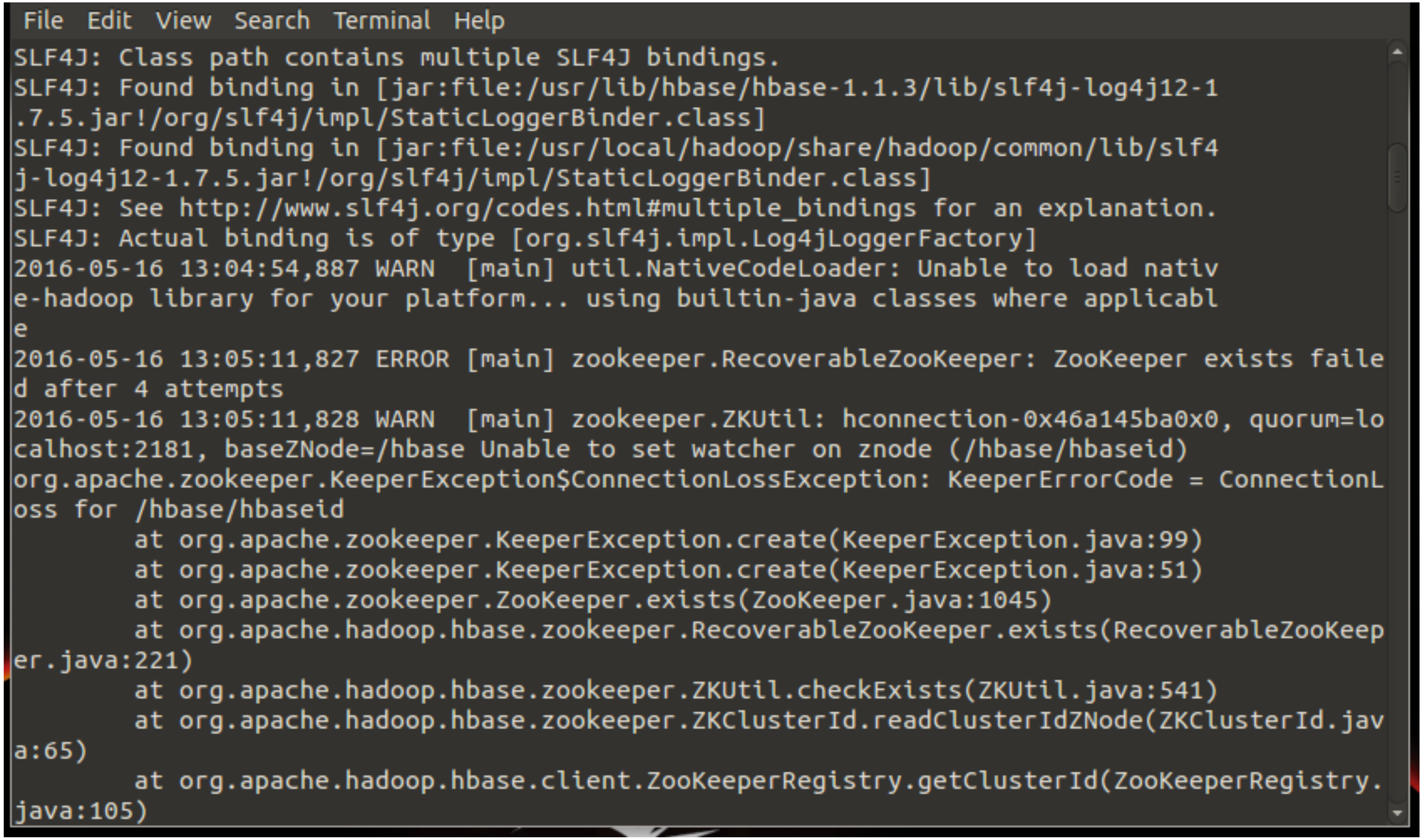
http://www.deeplearningbook.org/


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: Centaur


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