Introduction to Data Science

Joe Caserta
@joe_Caserta

Bill Walrond
@bill_walrond

caserta
CONCEPTS
@CasertaConcepts

About Caserta Concepts

- **Consulting** Data Innovation and Modern Data Engineering
  - Award-winning company
  - Internationally recognized work force
  - Strategy, Architecture, Implementation, Governance
- **Innovation** Partner
  - Strategic Consulting
  - Advanced Architecture
  - Build & Deploy
- **Leader** in Enterprise Data Solutions
  - Big Data Analytics
  - Data Warehousing
  - Business Intelligence
  - Data Science
  - Cloud Computing
  - Data Governance

Caserta Client Portfolio

Finance, Healthcare & Insurance

Digital Media/AdTech Education & Services

Retail/eCommerce & Manufacturing

Awards & Recognition

Gartner Market Guide to
Advanced Analytics
Service Providers

Top 10
Fastest Growing
Big Data Companies
2016

Top 20
Big Data
Companies
2015

Top 20
Most Powerful Big Data
Solution Provider
2014

Top 20
Healthcare
Solutions Providers
2015

Top 20
Data Analytics
Consulting Companies
2013

Our Partners

amazon web services	Hortonworks	cloudera	IBM	MAPR	Collibra
Lucidworks	tableau	basho	TIBCO Jaspersoft	talend	snowflake
DATASTAX	neo4j	Actian	syncsort	Datameer	Alation
pitney bowes	CISCO	DATAGUISE	REVOLUTION ANALYTICS	databricks	Microsoft Azure

Agenda

- Why we care about Big Data
- Challenges of working with Big Data
- Governing Big Data for Data Science
- Introducing the Data Pyramid
- Why Data Science is Cool?
- What does a Data Scientist do?
- Standards for Data Science
  - Business Objective
  - Data Discovery
  - Preparation
  - Models
  - Evaluation
  - Deployment
- Q & A

Big Data Analysis: Timeline of Society Media

Printing Press
1500s

Penny Post
1840s

Telegraph
1850s

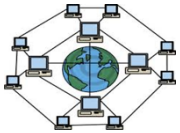
Rural Free Post
1850s

Telephone
1890s

Radio
1900s

TV
1950s

PCs
1970s



Internet
1980s

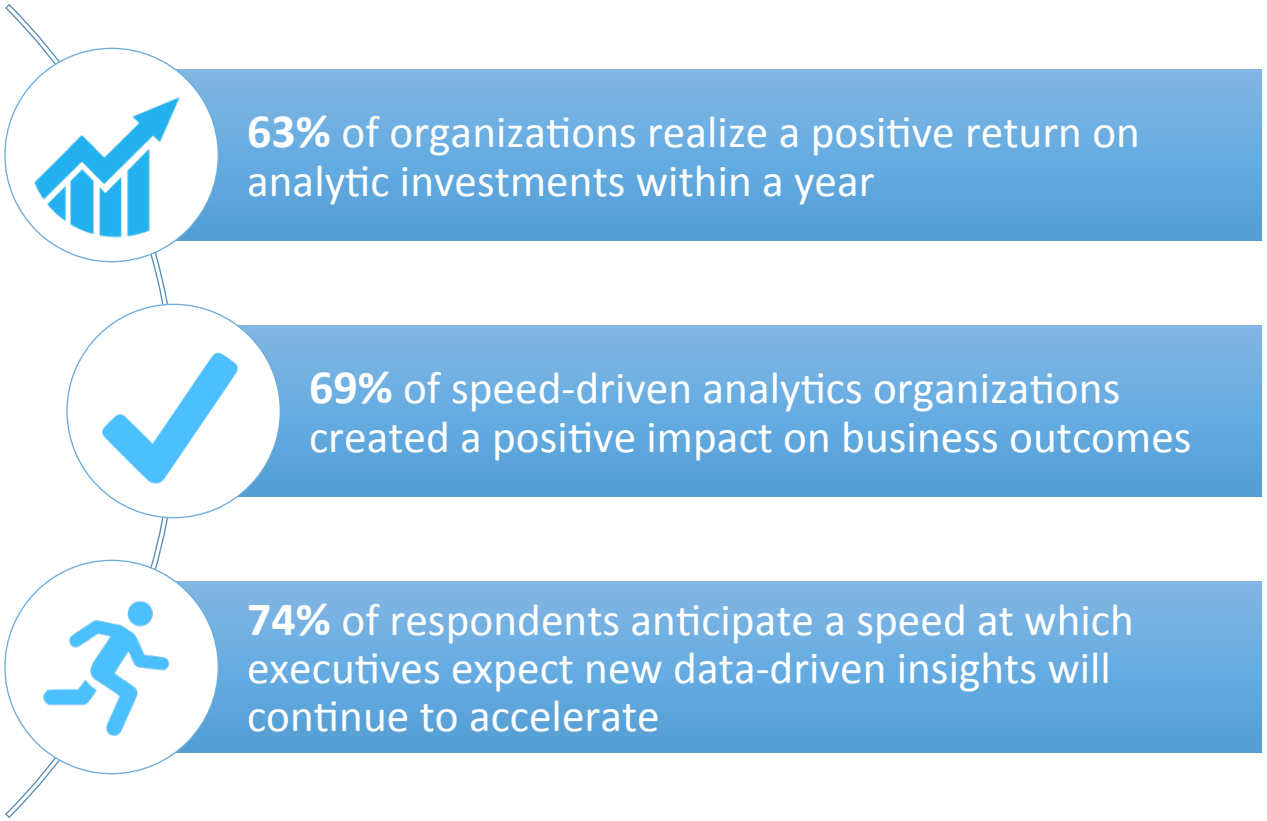
Web
1990s

Every 60 Seconds

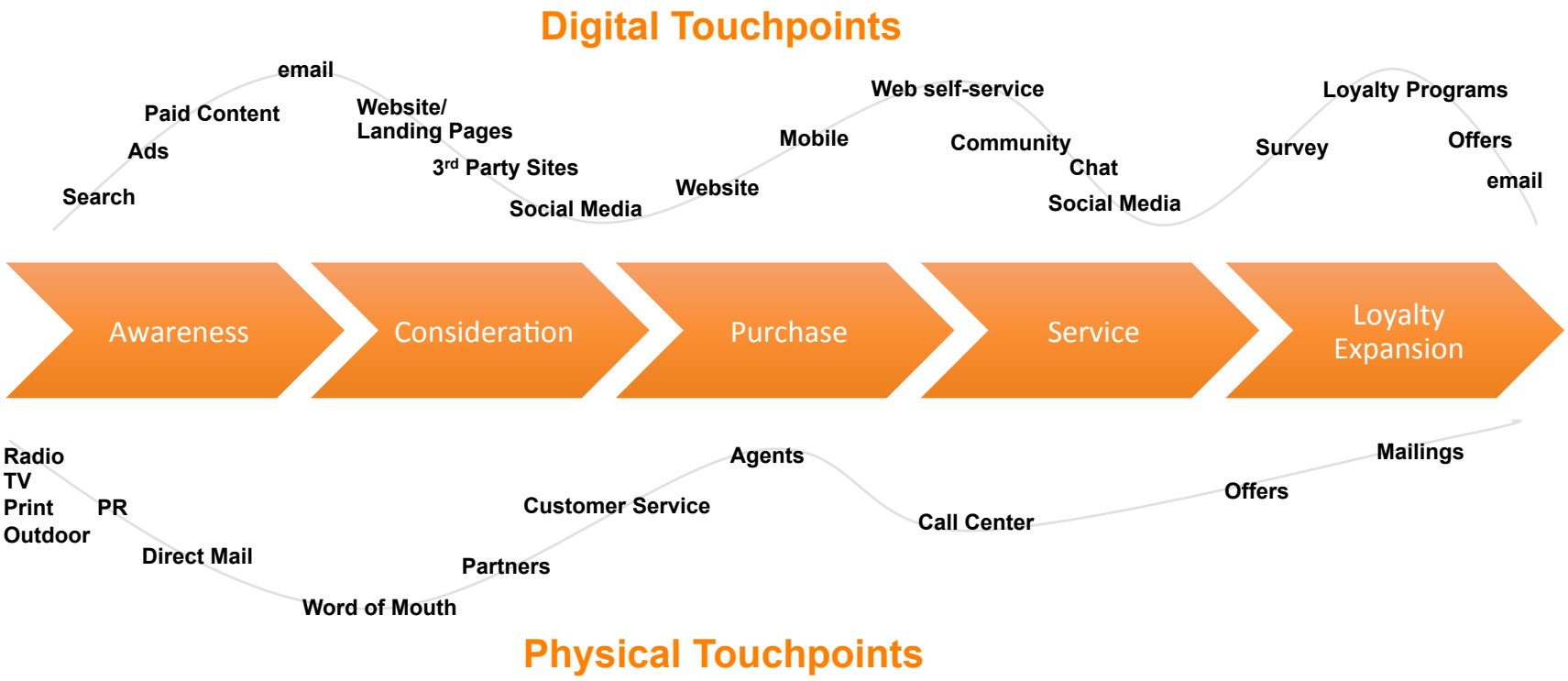
- **98,000+** Tweets
- **695,000** Status Updates
- **11 Million** instant messages
- **698,445** Google Searches
- **168 million+** emails sent
- **1,829 TB** of data created
- **217** new mobile web users

Social Media, Mobile, Big Data, Cloud
2000s

Data is your Differentiator



Understanding the Customer Journey



Understanding Touchpoint Methods

	Single Touch	Rules-Based	Statistically Driven
Type	Assign the credit to the first or last exposure	Assign the credit to each interaction based on business rules	Assign the credit to interactions based on data-driven model
	 100%	 33% 33% 33%	 27% 49% 24%
Comments	– Last touch only – Ignores bulk of customer journey – Undervalues other interactions and influencers	– Subjective – Assigns arbitrary values to each interaction – Lacks analytics rigor to determine weights	✓ Looks at full behavior patterns ✓ Consider all touch points ✓ Can apply different models for best results ✓ Use data to find correlations between touch points (winning combinations)

What is Data Science?

da·ta

/ˈdadə,ˈdādə/

noun

facts and statistics collected together for reference or analysis.

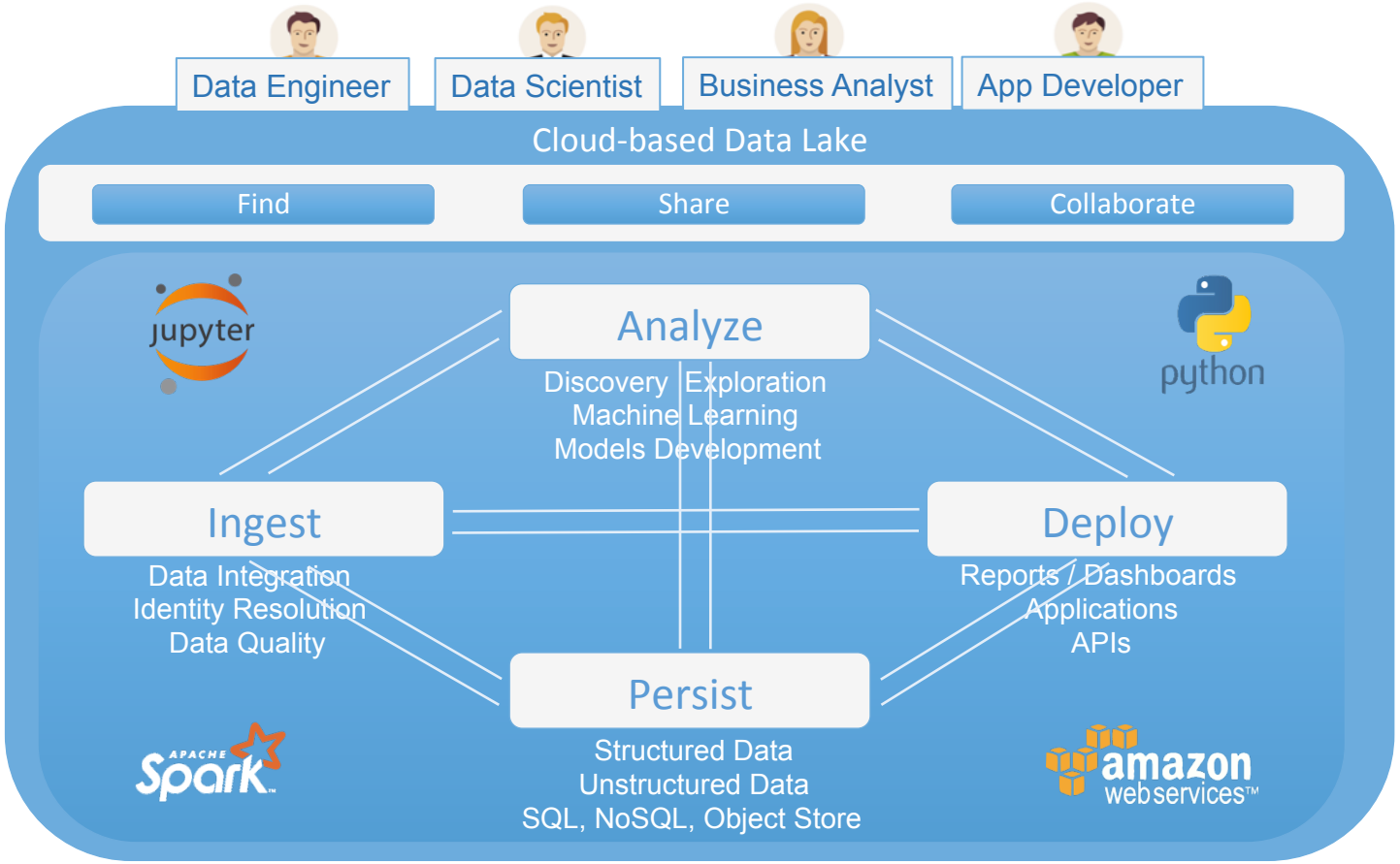
sci·ence

/ˈsīəns/

noun

the intellectual and practical activity encompassing the systematic study of the structure and behavior of the physical and natural world through observation and experiment.

Big Data Analysis: The Ecosystem of the future



Business Value

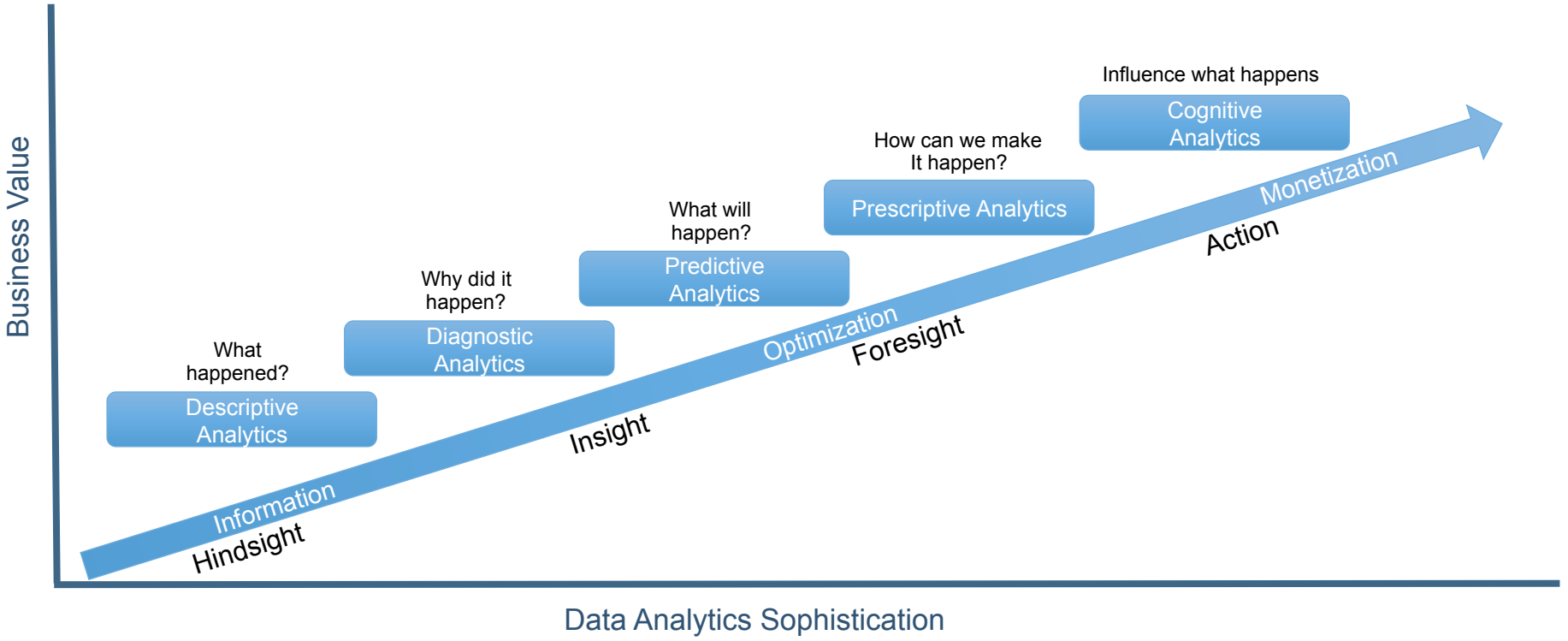
Provides innovative and industry leading technologies to rapidly be applied to the business without having to manage compatibility and data complexity.

Technical Value

Provides an open framework to reduce the number of integration points and testing environments to deliver business solutions.

Progression of Business Analytics to Data Science

Reports → Correlations → Predictions → Recommendations → Interactions



Progression of Data Science Maturity

- Timeline
- Tools
- Available libraries
- Best practices

What are the Realities of the Data Scientist

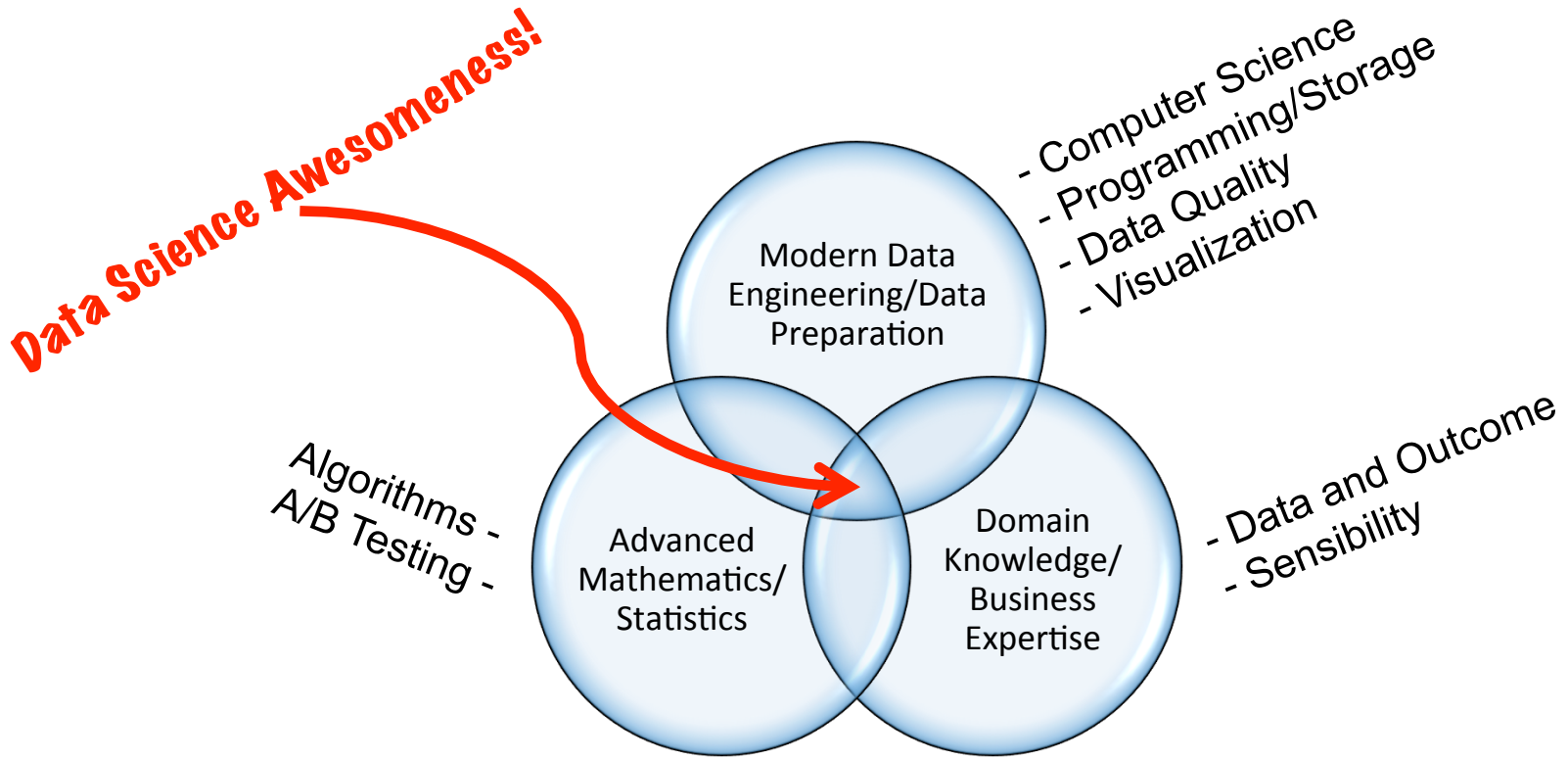
- Writes really cool and sophisticated algorithms that impacts the way the business runs.
- **NOT**

- Much of the time of a Data Scientist is spent:
- Searching for the data they need
- Making sense of the data
- Figuring why the data looks the way is does and assessing its validity
- Cleaning up all the garbage within the data so it represents true business
- Combining events with Reference data to give it context
- Correlating event data with other events
- **Finally, they implement algorithms to perform mining, clustering and predictive analytics**

Why Data Science now?

- Costs of compute and storage dramatically lower than just a few years ago
- Data generated by all aspects of society has dramatically increased
- Need to efficiently learn what there is to learn from our data

The Data Scientist Winning Trifecta

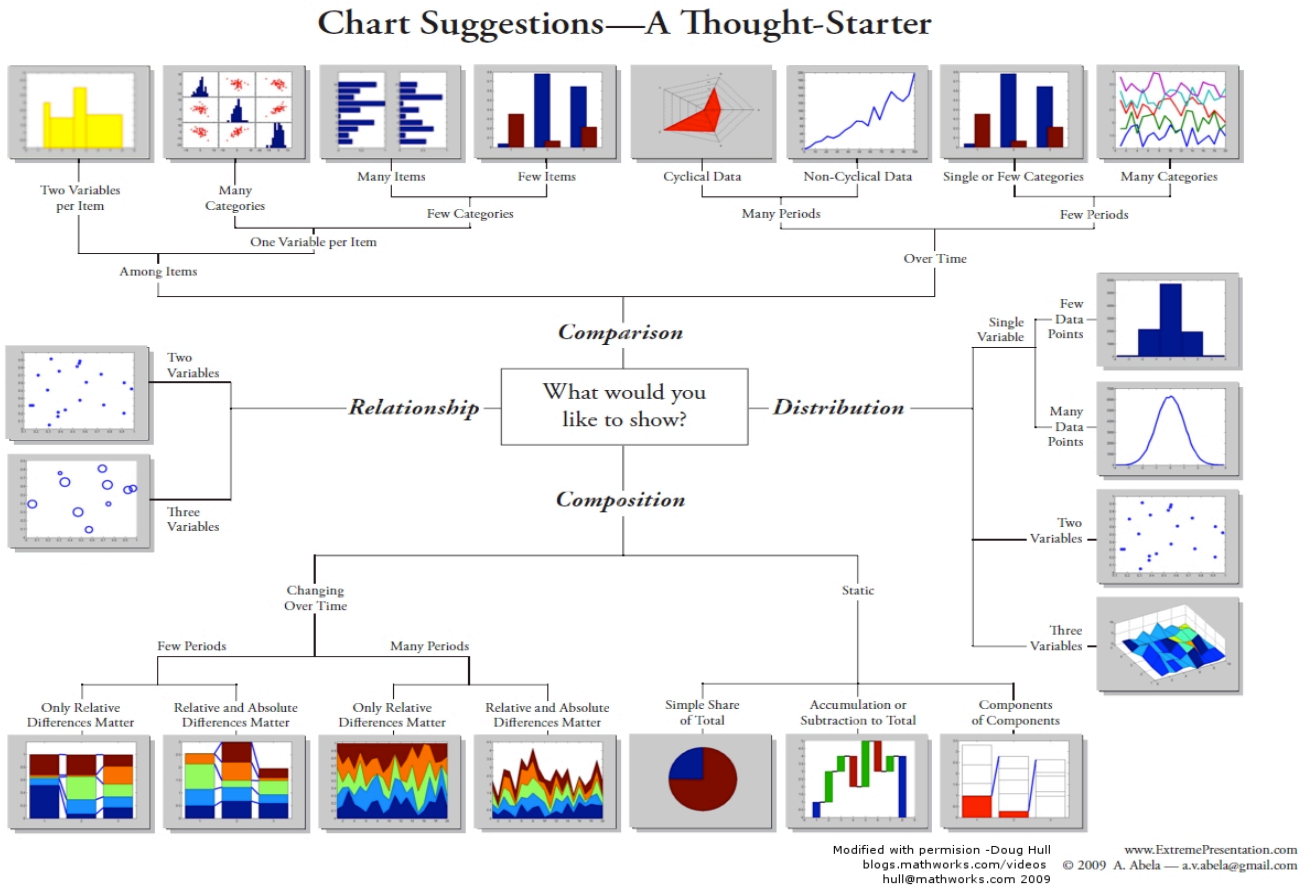


Modern Data Engineering

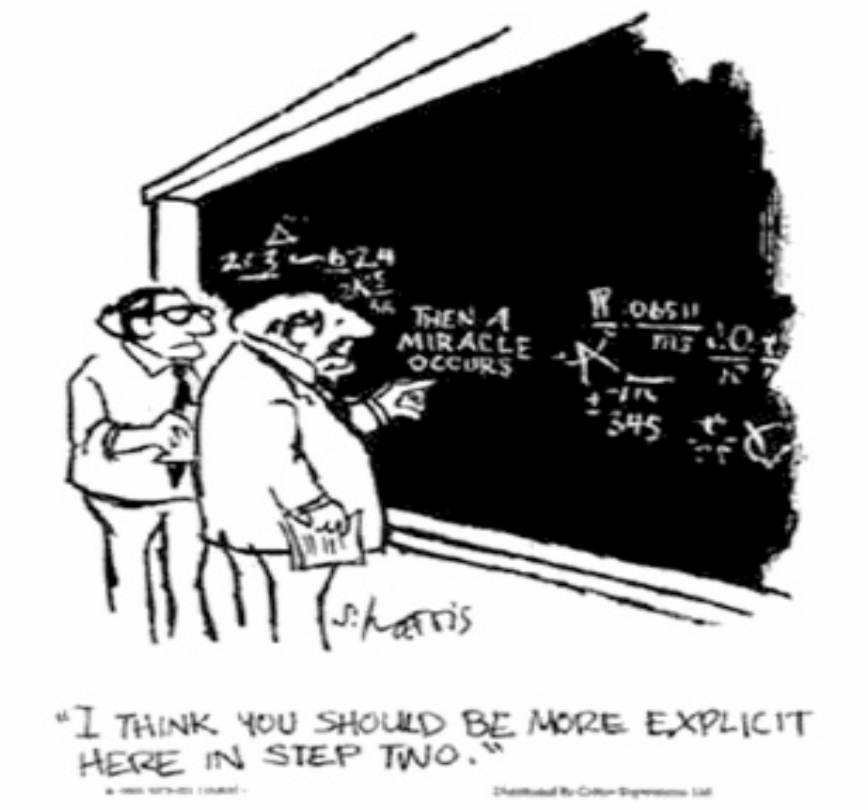
"It does look similar—but this one is powered by Hadoop"

"After careful consideration of all 437 charts, graphs, and metrics, I've decided to throw up my hands, hit the liquor store, and get snockered. Who's with me?!"

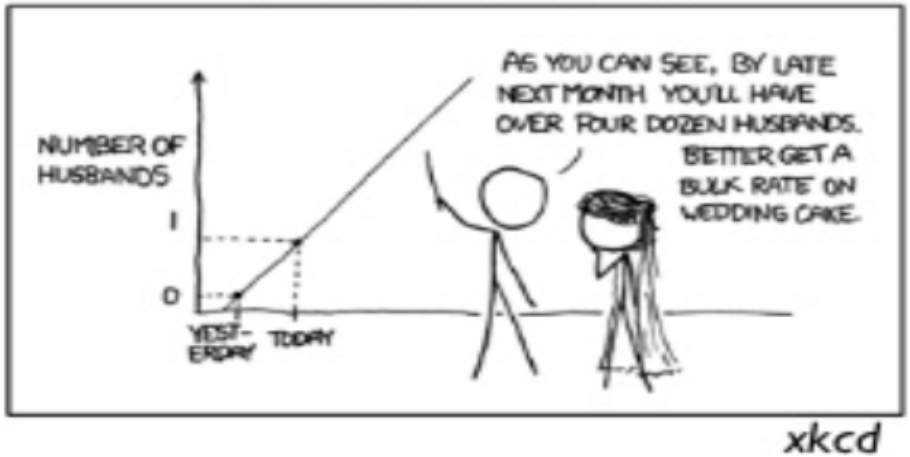
Which Visualization, When?



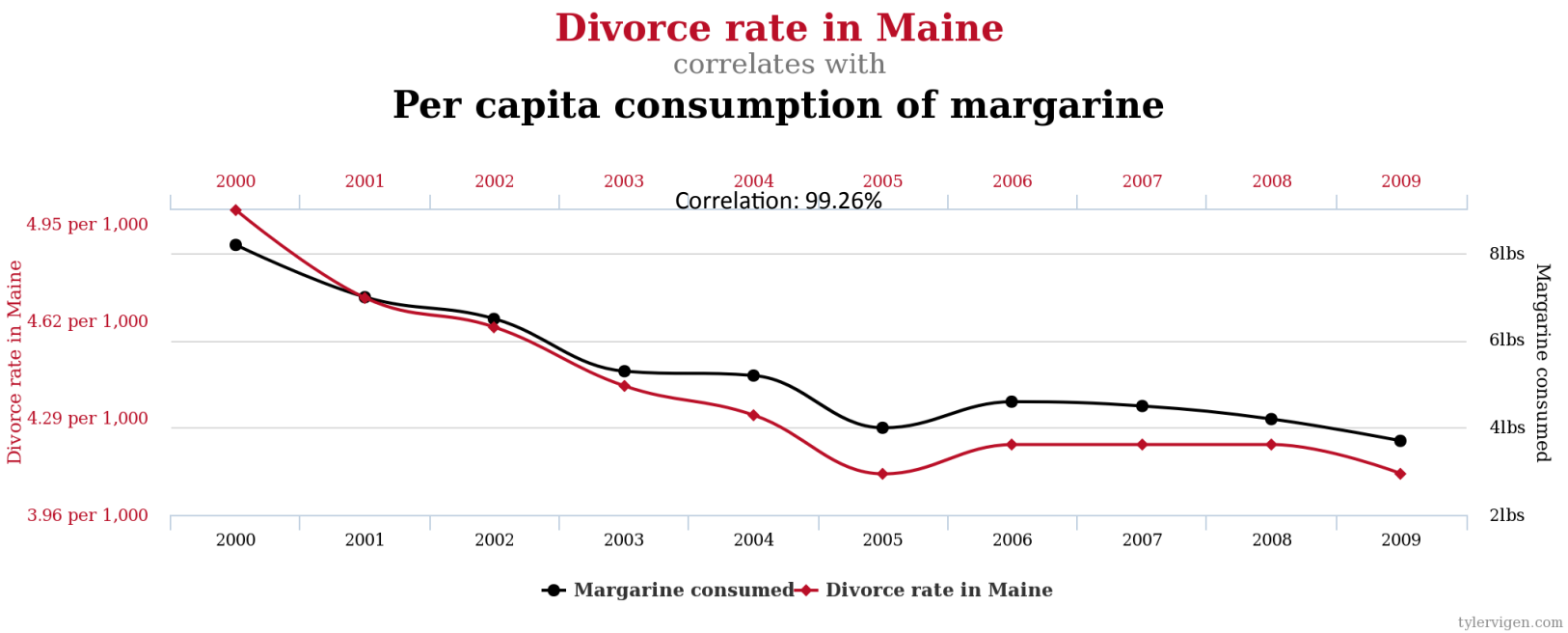
Advanced Mathematics / Statistics



Domain and Outcome Sensibility

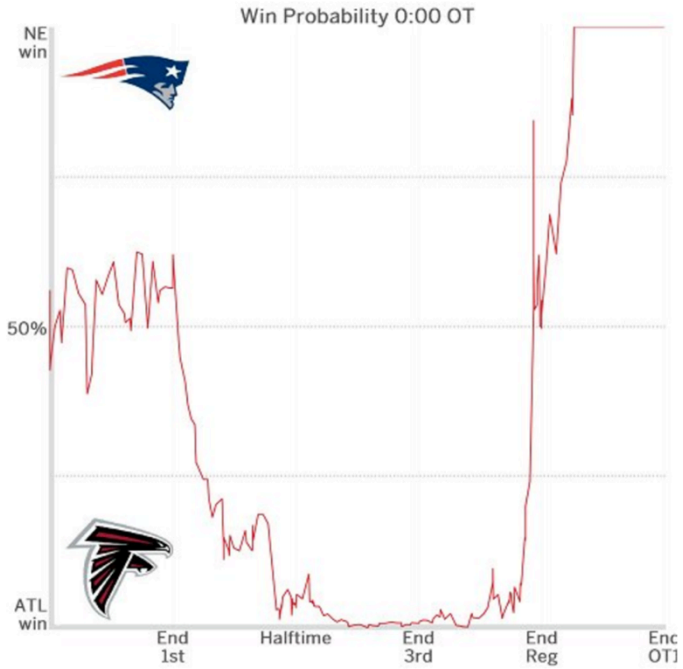


Is Data Trying To Trick You?



Are we Considering the Right Factors?

- 0.996: ESPN's estimate of the Falcons win probability with 9:44 remaining in the 4th quarter
- What is the problem with ESPN's estimate?
- How many times did the Patriots score at least 2 TDs in a quarter in 2016? 12%

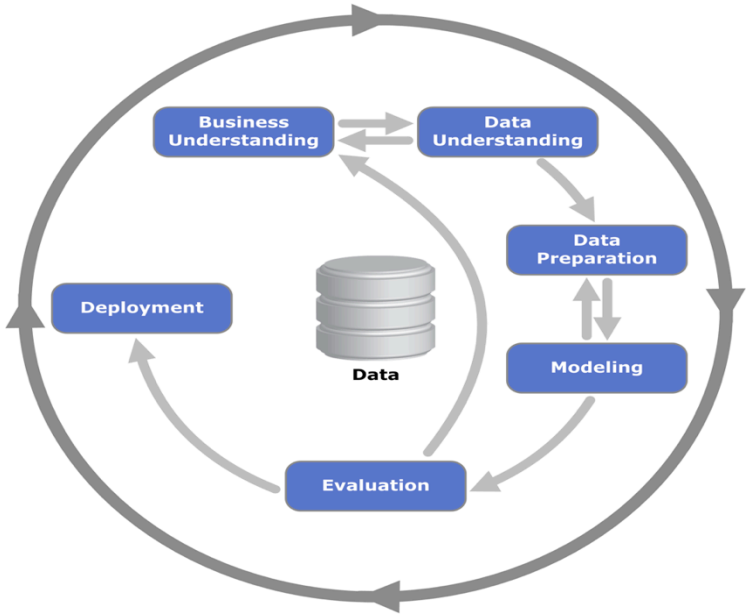


(Source: ESPN)

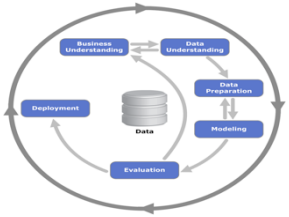
Are there Standards?

CRISP-DM: Cross Industry Standard Process for Data Mining

1. Business Understanding
    - Solve a single business problem
2. Data Understanding
    - Discovery
    - Data Munging
    - Cleansing Requirements
3. Data Preparation
    - ETL
4. Modeling
    - Evaluate various models
    - Iterative experimentation
5. Evaluation
    - Does the model achieve business objectives?
6. Deployment
    - PMML; application integration; data platform; **Excel**



1. Business Understanding

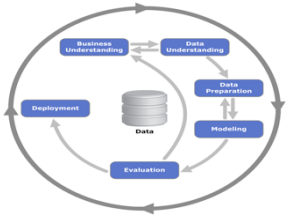


In this initial phase of the project we will need to **speak to humans.**

- It would be premature to jump in to the data, or begin selection of the appropriate model(s) or algorithm
- Understand the project objective
- Review the business requirements
- The output of this phase will be conversion of business requirements into a preliminary technical design (decision model) and plan.

Since this is an iterative process, this phase will be revisited throughout the entire process.

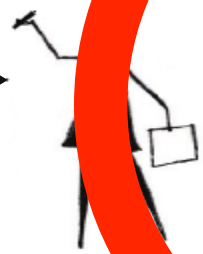
Gathering Requirements



Business Stakeholders

Business Stakeholders

Business Stakeholders

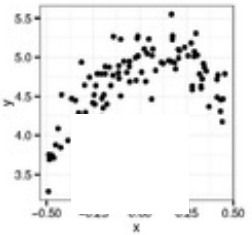


Business Analyst

Interview not

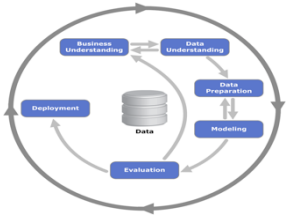
Requirement Document

Data Scientist



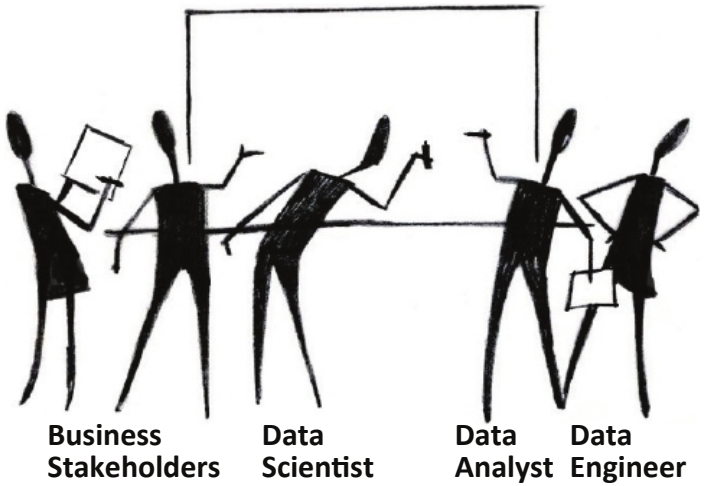
Models / Insights

Data Science Scrum Team



Efficient

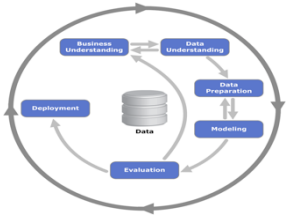
Inclusive



Interactive

Effective

2. Data Understanding

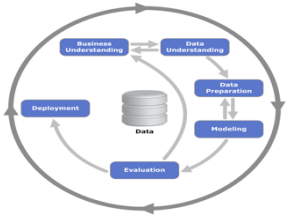


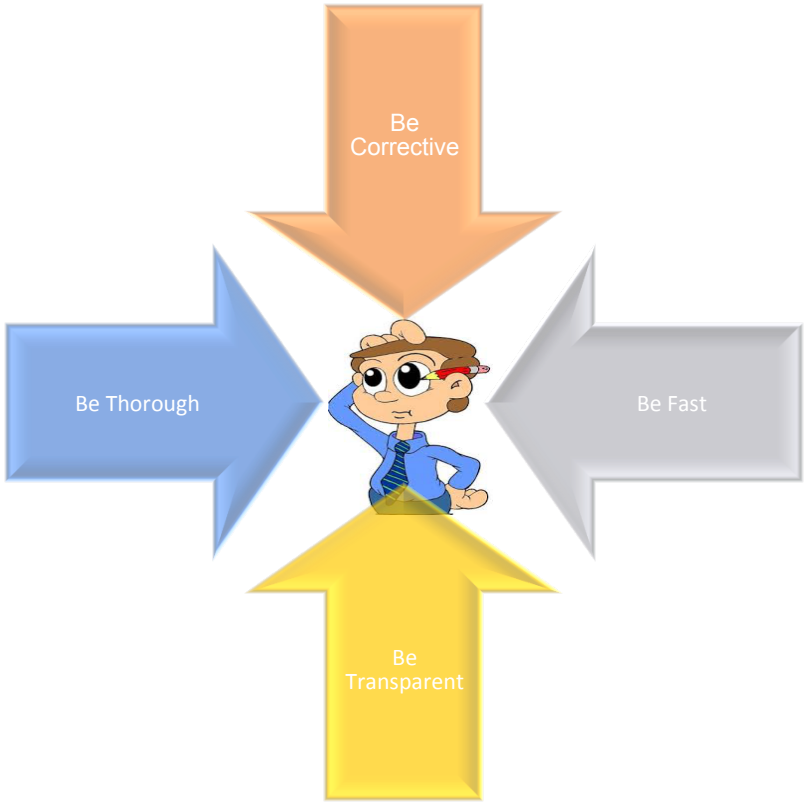
- **Data Discovery** → understand where the data you need comes from
- **Data Profiling** → interrogate the data at the entity level, understand key entities and fields that are relevant to the analysis.
- **Cleansing Requirements** → understand data quality, data density, skew, etc
- **Data Munging** → collocate, blend and analyze data for early insights! Valuable information can be achieved from simple group-by, aggregate queries, and even more with **SQL Jujitsu!**

Significant iteration between Business Understanding and Data Understanding phases.

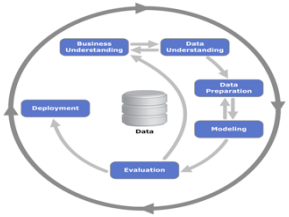
Sample Exploration tools for Hadoop: Trifacta, Paxata, Spark, Python, Waterline, Elasticsearch

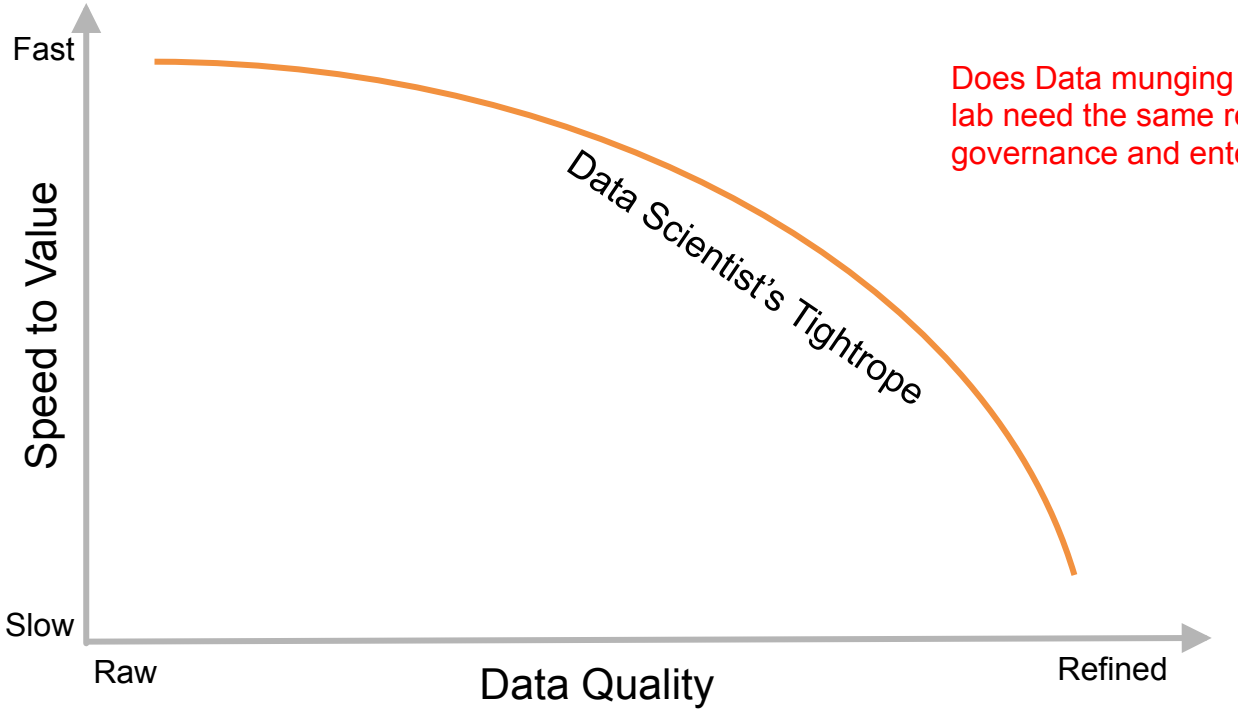
Data Science Data Quality Priorities





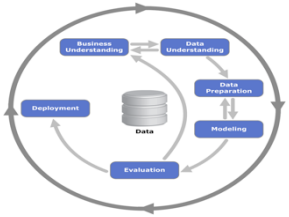
Data Science Data Quality Priorities





Does Data munging in a data science lab need the same restrictive governance and enterprise reporting?

3. Data Preparation

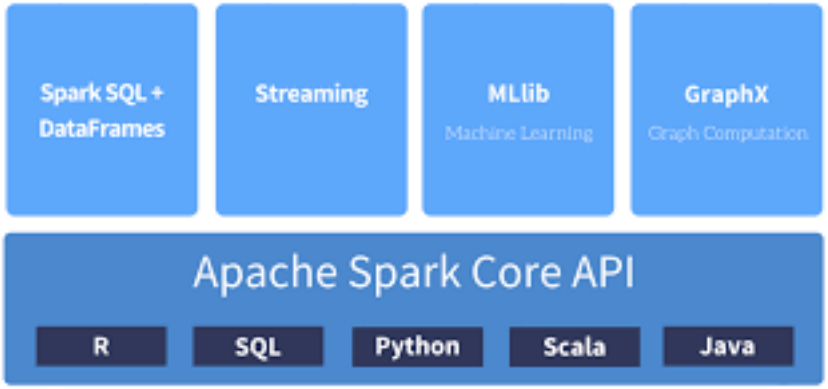


ETL (Extract Transform Load)

90+% of a Data Scientists time goes into Data Preparation!

- Locating and acquiring valuable data sources
- Select required entities/fields
- Address Data Quality issues: missing or incomplete values, whitespace, bad data-points
- Join/Enrich disparate datasets
- Derive behavioral features
- Transform/Aggregate data for intended use:
  - Sample
  - Aggregate
  - Pivot

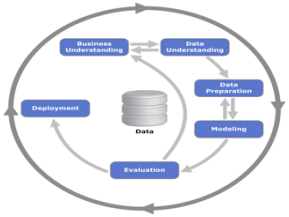
We ❤ Spark



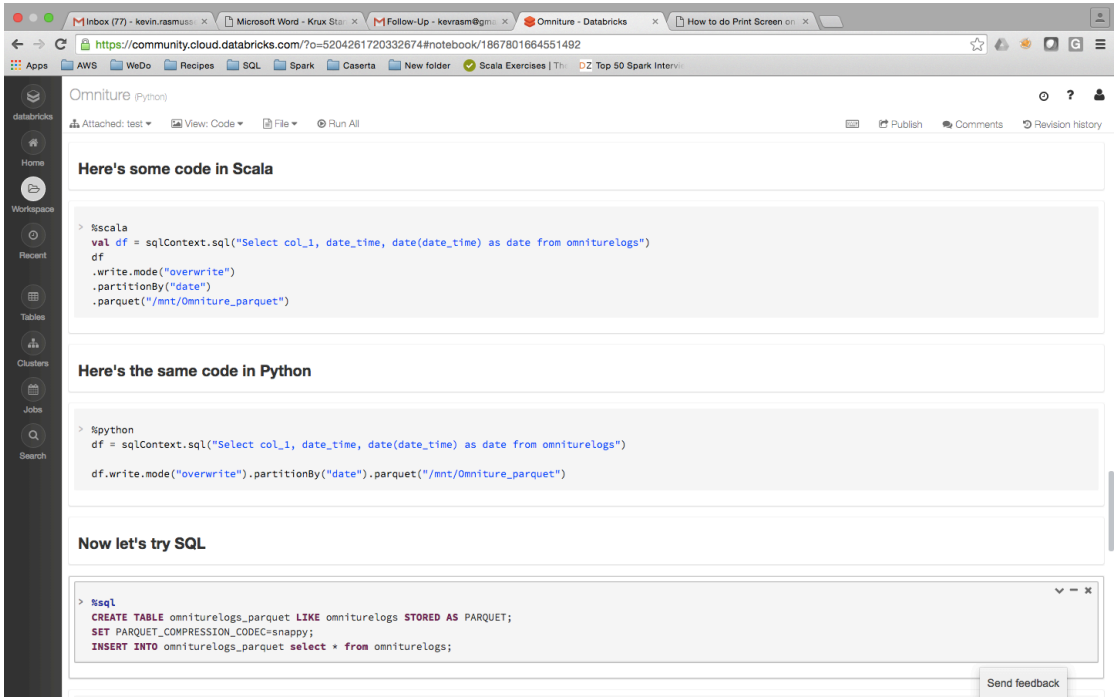
- Development local or distributed is identical
- Beautiful high level API's
- Full universe of Python modules
- Open source and Free
- Blazing fast!

Spark has become our *default* processing engine for a data engineering & science

Data Preparation

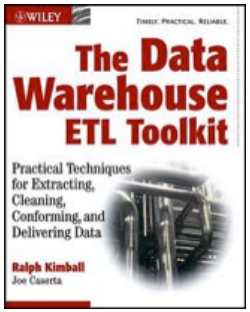


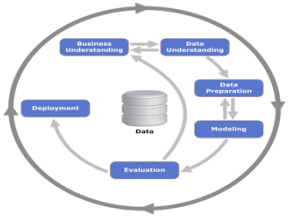
- We love Spark!
- ETL can be done in Scala, Python or SQL
- Cleansing, transformation, and standardization
- Address Parsing: usaddress, postal-address, etc
- Name Hashing: fuzzy, etc
- Genderization: sexmachine, etc
- And all the goodies of the standard Python library!
- Parallelize workload against a large number of machines in Hadoop cluster



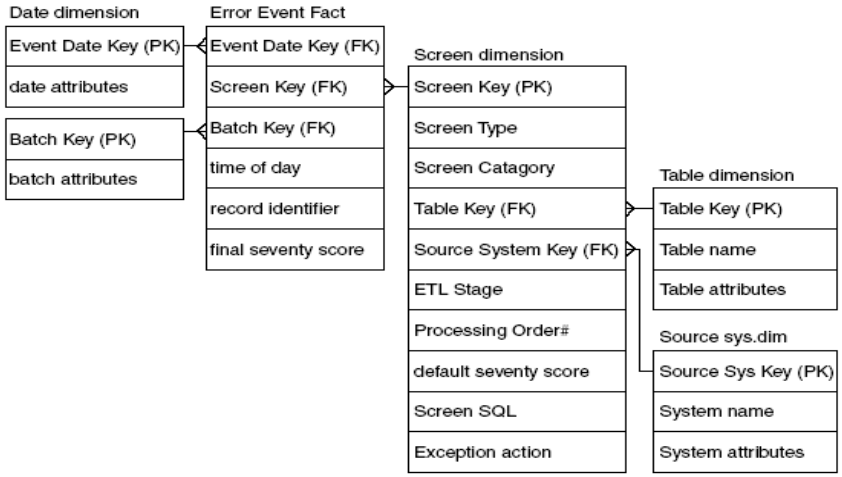
Data Quality and Monitoring

- BUILD a robust data quality subsystem:
- Metadata and error event facts
- Orchestration
- Based on Data Warehouse ETL Toolkit
- Each error instance of each data quality check is captured
- Implemented as sub-system after ingestion
- Each fact stores unique identifier of the defective source row





HAMBot: 'open source' project created in Caserta Innovation Lab (CIL)

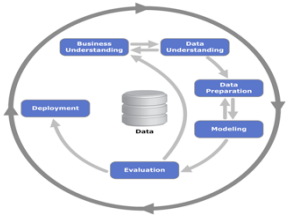


Data Preparation Demonstration!

Wifi:
Hilton Meeting Room Wifi
infotoday2017

Follow along: http://bit.ly/2r9ABcK
File: SanFranCrime.ipynb

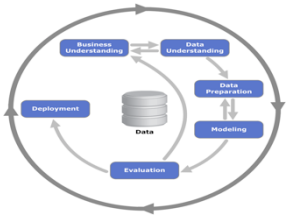
4. Modeling



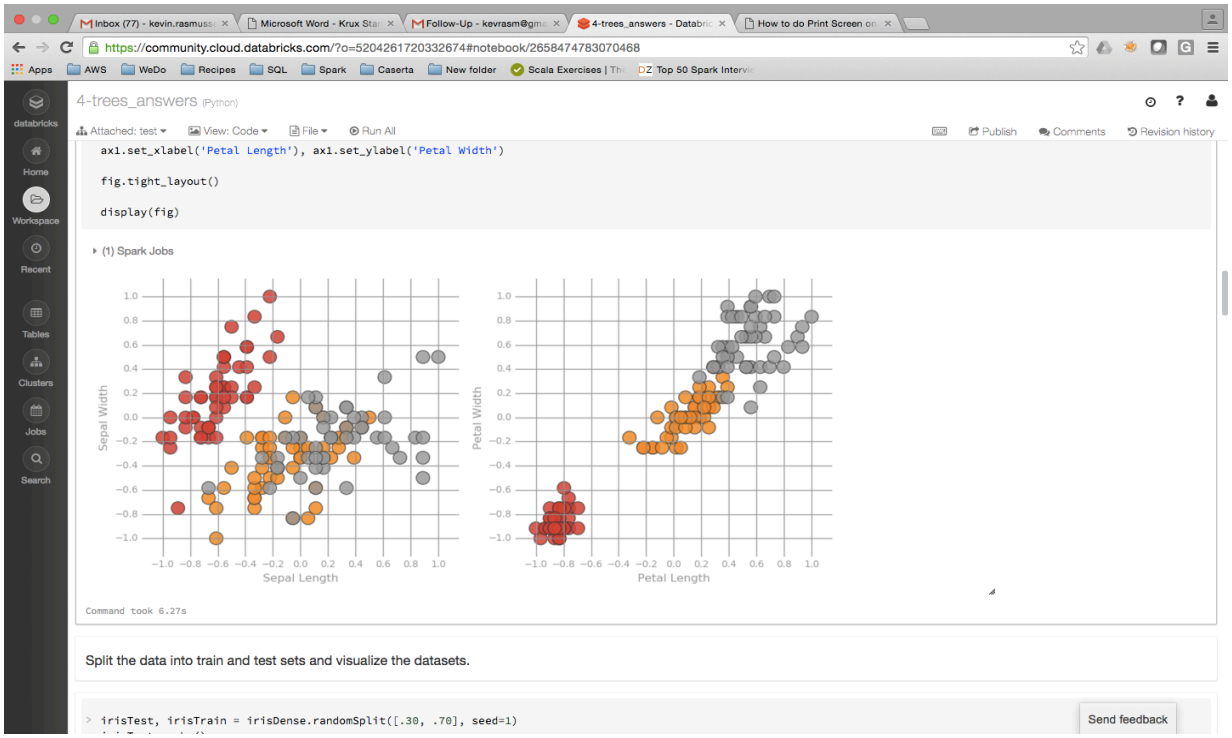
Do you love algebra & stats?

- Evaluate various models/algorithms
  - Classification
  - Clustering
  - Regression
  - Many others…..
- Tune parameters
- Iterative experimentation
- Different models may require different data preparation techniques (ie. Sparse Vector Format)
- Additionally we may discover the need for additional data points, or uncover additional data quality issues!

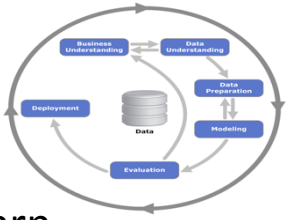
Modeling in Hadoop



- Spark works well
- SAS, SPSS, Etc. not native on Hadoop
- R and Python becoming new standard
- PMML can be used, but approach with caution



Machine Learning



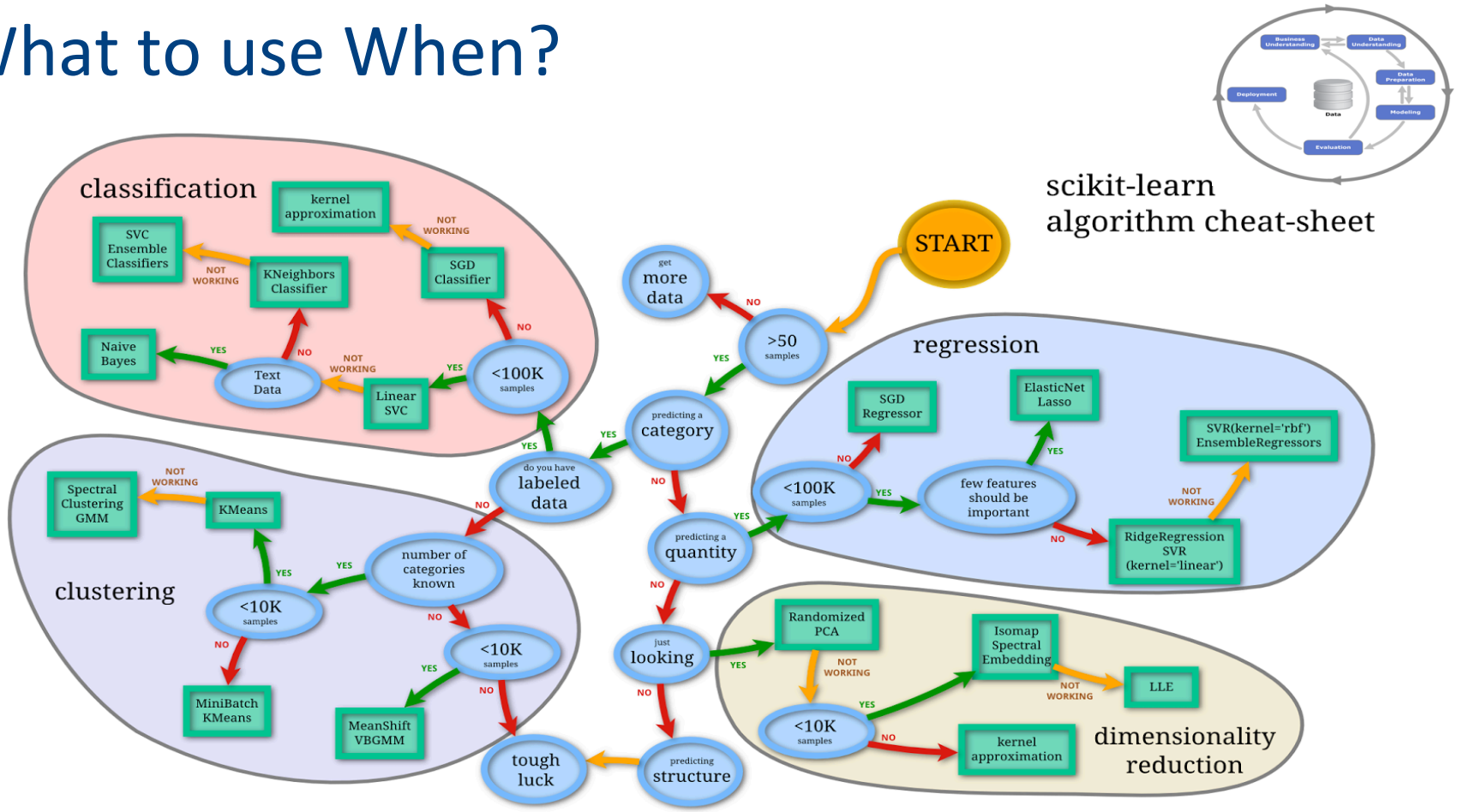
The goal of machine learning is to get software to make decisions and learn from data without being programed explicitly to do so

Machine Learning algorithms are broadly broken out into two groups:

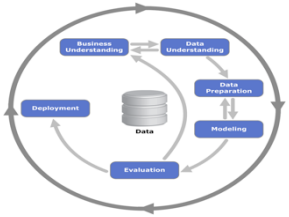
- **Supervised learning** → inferring functions based on labeled training data
- **Unsupervised learning** → finding hidden structure/patterns within data, no training data is supplied

We will review some popular, easy to understand machine learning algorithms

What to use When?



Supervised Learning



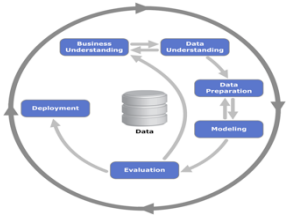
The training set is used to generate a **function**

Name	Weight	Color	Cat_or_Dog
Susie	9lbs	Orange	Cat
Fido	25lbs	Brown	Dog
Sparkles	6lbs	Black	Cat
Fido	9lbs	Black	Dog

..so we can predict if we have a cat or dog!

Name	Weight	Color	Cat_or_Dog
Misty	5lbs	Orange	?

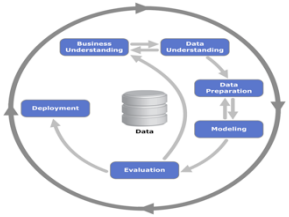
Category or Values?



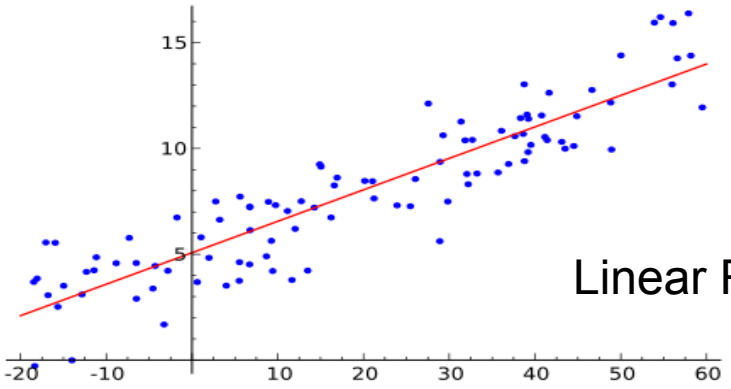
There are several classes of algorithms depending on whether the prediction is a category (like cat or dog) or a value, like the value of a home.

Classification algorithms are generally well fit for categorization, while algorithms like Regression and Decision Trees are well suited for predicting “continuous” values.

Regression

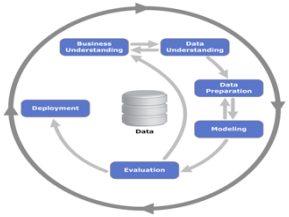


- Understanding the relationship between a given set of dependent variables and independent variables
- Typically regression is used to predict the output of a dependent variable based on variations in independent variables
- Very popular for prediction and forecasting

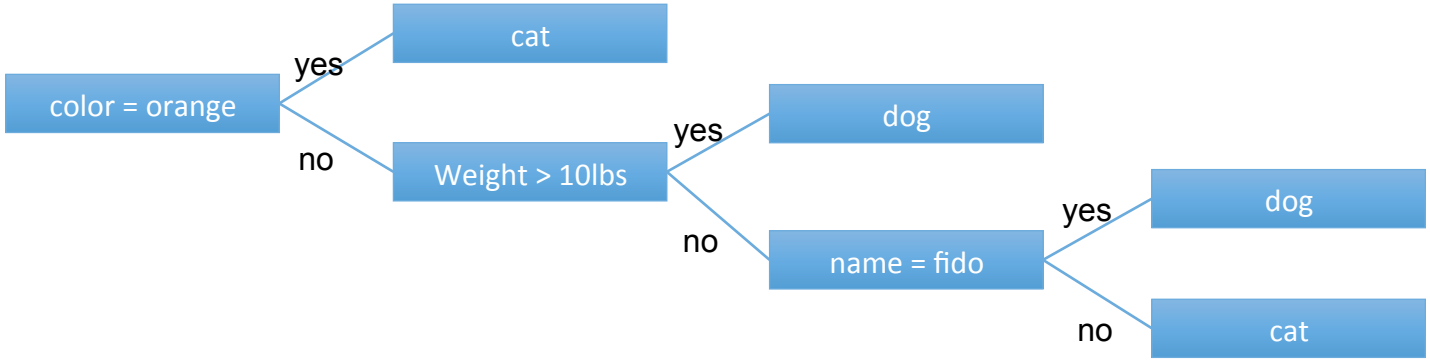


Linear Regression

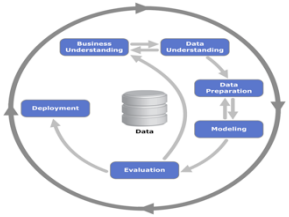
Decision Trees



- A method for predicting outcomes based on the features of data
- Model is represented a easy to understand tree structure of if-else statements

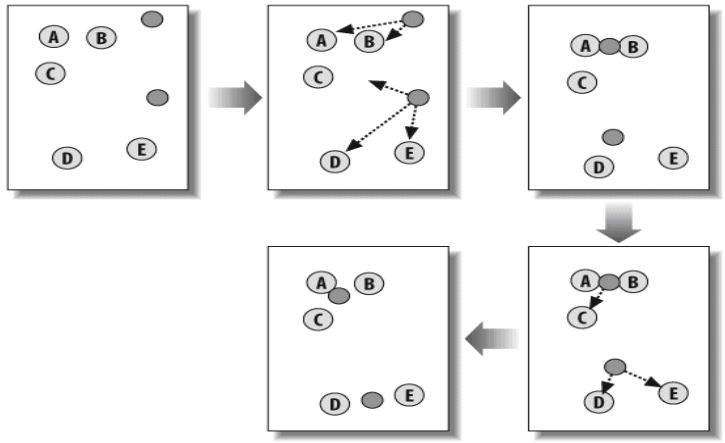


Unsupervised K-Means



Clustering of items into logical groups based on natural patterns in data

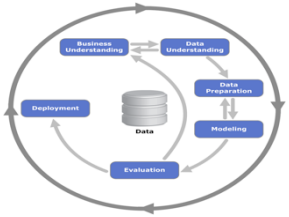
- Treats items as coordinates
- Places a number of random “centroids” and assigns the nearest items
- Moves the centroids around based on average location
- Process repeats until the assignments stop changing



Uses:

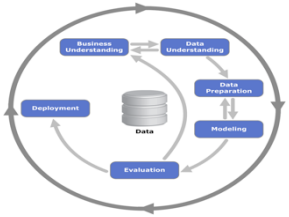
- Cluster Analysis
- Classification
- Content Filtering

Collaborative Filtering



- A hybrid of Supervised and Unsupervised Learning (Model Based vs. Memory Based)
- Leveraging **collaboration** between multiple agents to filter, project, or detect patterns
- Popular in recommender systems for projecting the “taste” for of specific individuals for items they have not yet expressed one.

Item-based



- A popular and simple memory-based collaborative filtering algorithm
- Projects preference based on item similarity (based on ratings):

```
for every item i that u has no preference for yet
    for every item j that u has a preference for
        compute a similarity s between i and j
        add u's preference for j, weighted by s, to a running average
return the top items, ranked by weighted average
```

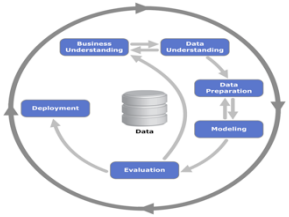
- First a matrix of Item to Item similarity is calculated based on user rating
- Then recommendations are created by producing a weighted sum of top items, based on the users previously rated items

Data Science Demonstration!

Wifi:
Hilton Meeting Room Wifi
infotoday2017

Follow along: http://bit.ly/2r9ABcK
File: SanFranCrime_model_DataSummit.ipynb

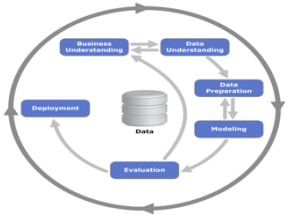
5. Evaluation



What problem are we trying to solve again?

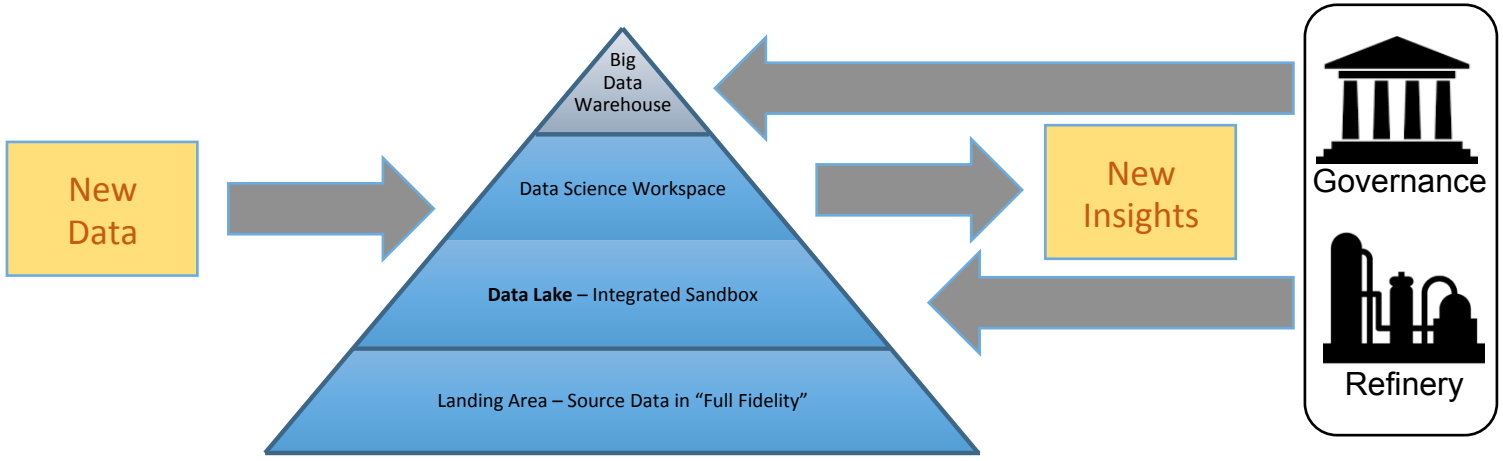
- Our final solution needs to be evaluated against original Business Understanding
- Did we meet our objectives?
- Did we address all issues?

6. Deployment



Engineering Time!

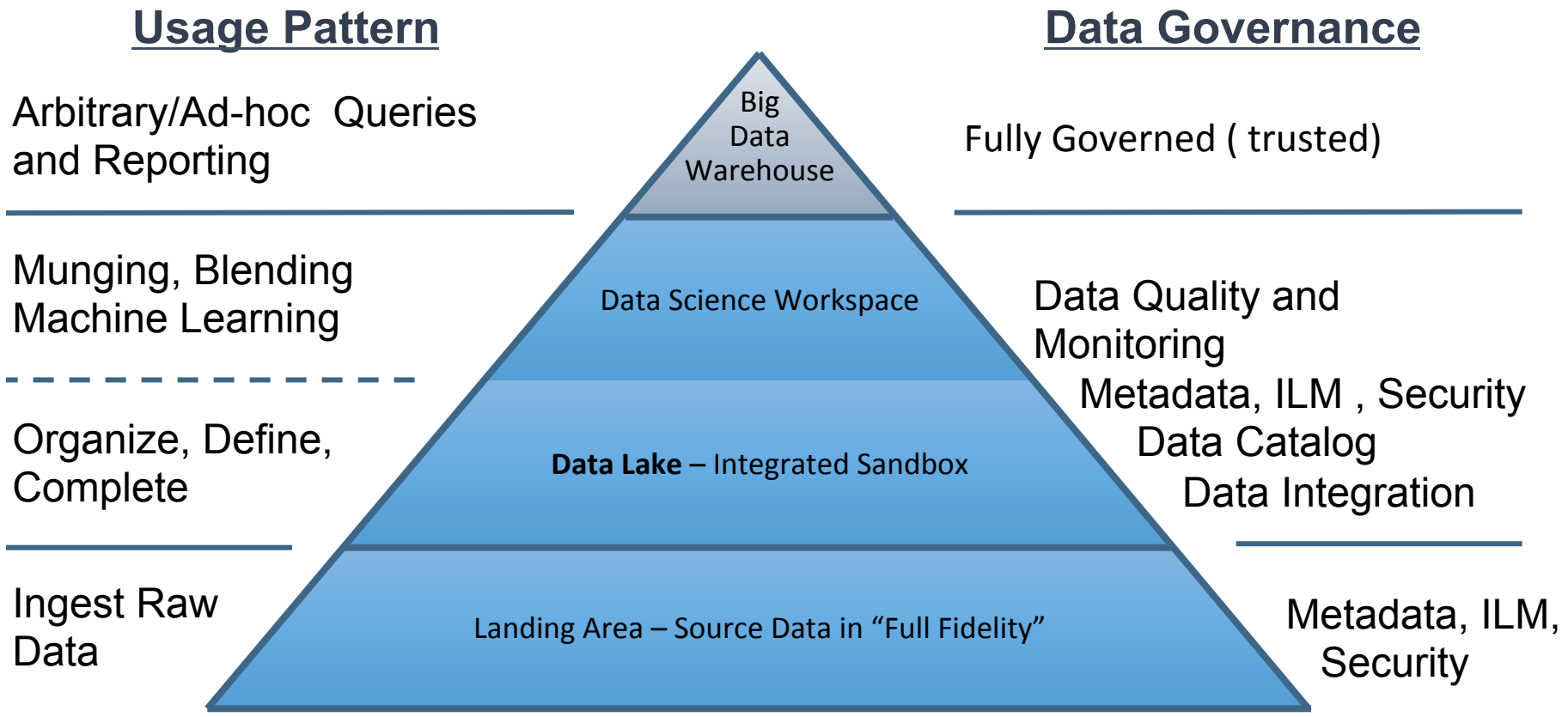
- It's time for the work products of data science to "graduate" from "new insights" to real applications.
- Processes must be hardened, repeatable, and generally perform well too!
- Data Governance applied
- PMML (Predictive Model Markup Langauge): XML based interchange format



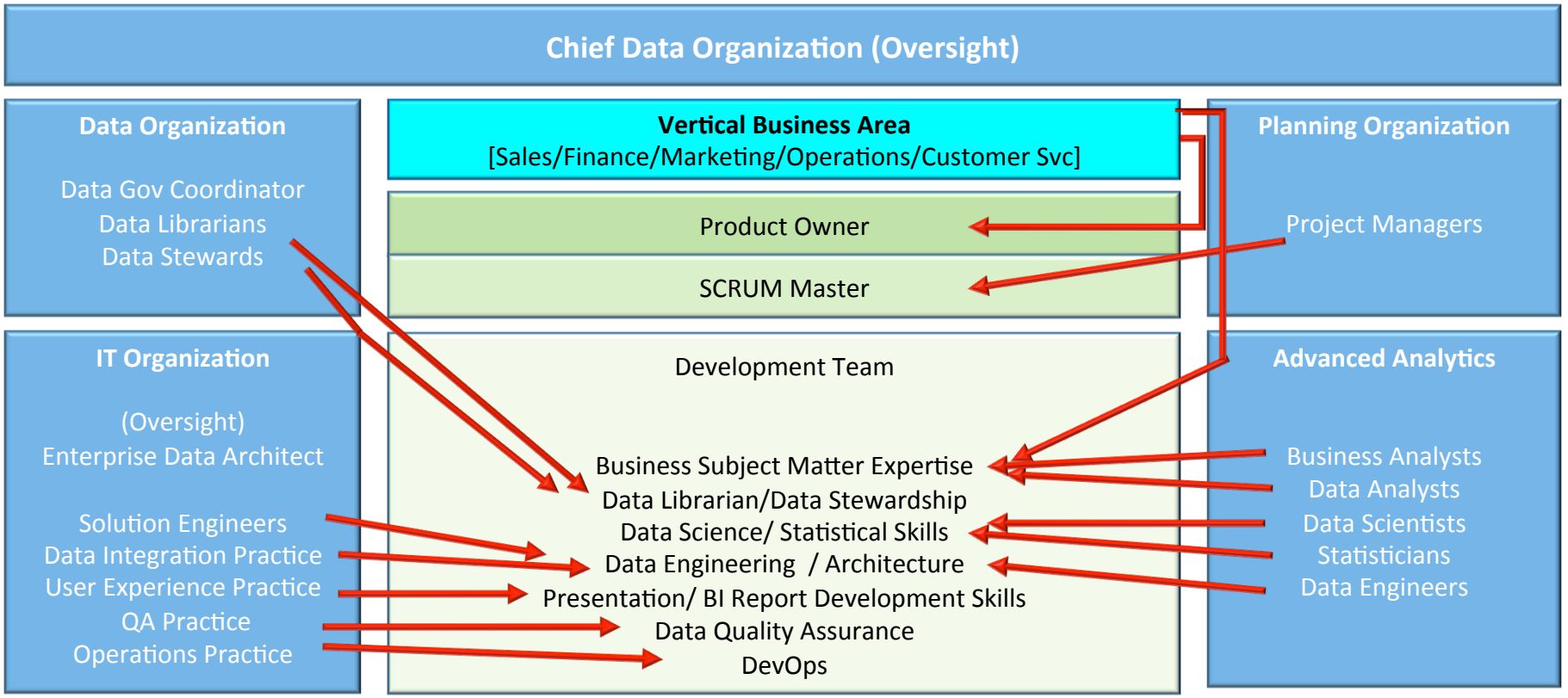
Components of Data Governance For Data Science

Organization	• Add Big Data to overall framework and assign responsibility • Add data scientists to the Stewardship program • Assign stewards to new data sets (twitter, call center logs, etc.)
Metadata	• Larger scale • New datatypes • Integrate with Hive Metastore, HCatalog, home grown tables
Privacy/Security	• Data detection and masking on unstructured data upon ingest
Data Quality and Monitoring	• Data Quality and Monitoring (probably home grown, drools?) • Quality checks not only SQL: machine learning, Pig and Map Reduce • Acting on large dataset quality checks may require distribution
Business Process Integration	• Near-zero latency, DevOps, Core component of business operations
Master Data Management	• Graph databases are more flexible than relational • Lower latency service required • Distributed data quality and matching algorithms
Information Lifecycle Management (ILM)	• Secure and mask multiple data types (not just tabular) • Deletes are more uncommon (unless there is regulatory requirement) • Take advantage of compression and archiving (like AWS Glacier)

Corporate Data Pyramid (CDP)



Analytics-Driven Organization



Technologies & Techniques

- The Cloud and Spark can provide a relatively low cost and extremely scalable platform for Data Science
- AWS S3 and Google GCS offers great scalability and speed to value without the overhead of structuring data
- Spark, with MLlib offers a great library of established Machine Learning algorithms, reducing development efforts
- Python and SQL are choices for Data Science
- Go Agile and follow Best Practices (CRISP-DM)
- Employ Data Pyramid concepts to ensure data has just enough governance

Some Thoughts – Enable the Future

- Data Science requires the convergence of data quality, advanced math, data engineering and visualization and business smarts
- Make sure your data can be trusted and people can be held accountable for impact caused by low data quality.
- Good data scientists are rare: It will take a village to achieve all the tasks required for effective data science
- **Get good!**
- **Be great!**
- **Blaze new trails!**

Data Science Training:

https://explore-data-science.thisismetis.com

- Big Data Warehousing Meetup
- New York City
- 4,300+ members
- Knowledge sharing

Thank You / Q&A

caserta
CONCEPTS

Joe Caserta
@joe_Caserta

Bill Walrond
@bill_walrond