Getting Value from Data

Joe Cline

DATA MODELER & ENGINEER

@mrjoedata www.linkedin.com/in/josephcline www.datanomicon.blog

In This Module

Data engineering ("Big Data")

- More than ETL development
- Data "munging" or "wrangling"

Business intelligence (BI)

- The BI developer role
- BI tools

Exploratory data analytics (EDA)

Statistical data analytics

- Data science

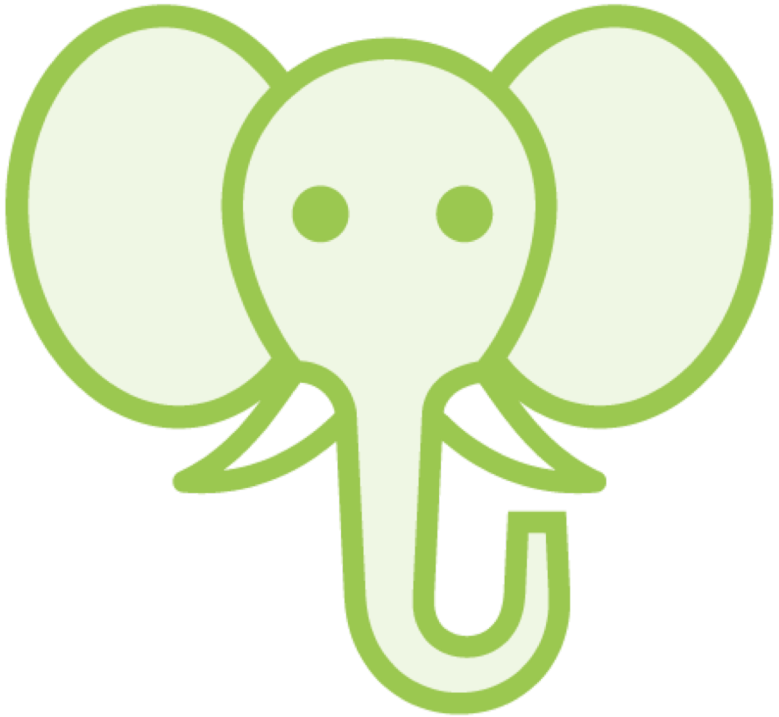
Predictive modeling

- Machine language (ML) algorithms
- Knowledge discovery and data mining (KDD)

Data visualization (Data viz)

Big Data

The 5 “V”s of Big Data



Volume – A lot of data > 1 TB

Variety – Structured, unstructured

Velocity – Always more coming

Variability – Data from disparate sources

Veracity – Potential data quality issues

Data Engineering

Data Engineering or Data Integration (ETL)

- How much data?
- What are you trying to do?
- How disparate are the sources?
- ETL tool access web services?
- Scrape websites?
- Connect to AWS or Google cloud?
- Connect to Hadoop?
- How permanent?

Productionalize:

To take a proof of concept and make it resilient and scalable so it can be put into a production environment

The Data Engineer

Data pipelines/connecting pipelines

Build proof-of-concepts (POC) fast

Fail fast, fail often

Productionalize

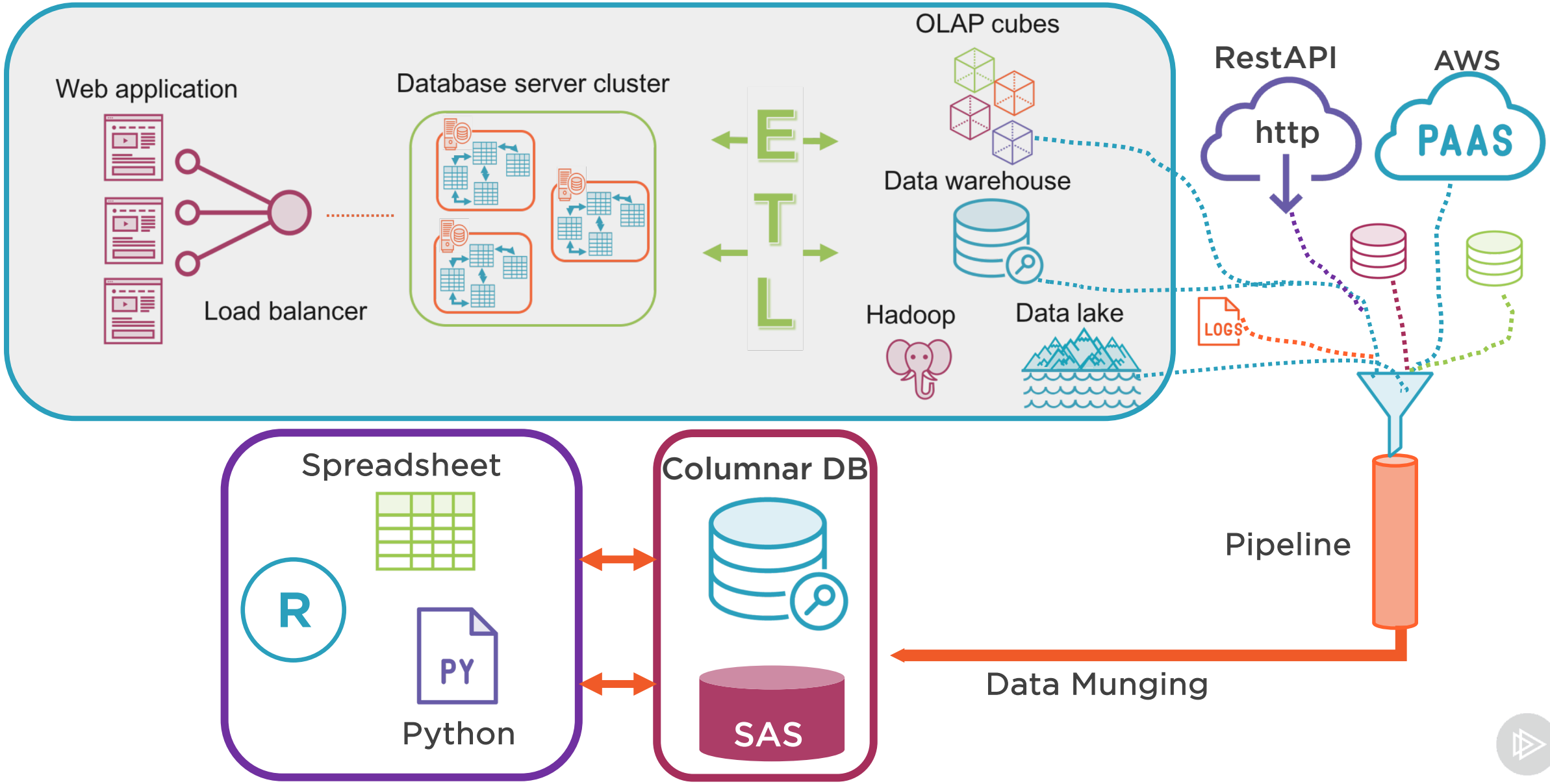
The Data Engineering Languages

- R
- Python
- SAS
- Java
- SQL
- Scala
- Go
- Julia

The Data Engineering Systems

- Vertica (or another columnar database)
- SAS environment
- R Studio Server
- Anaconda Enterprise Server (Python)
- Hadoop

Adding to Our Generic Architecture

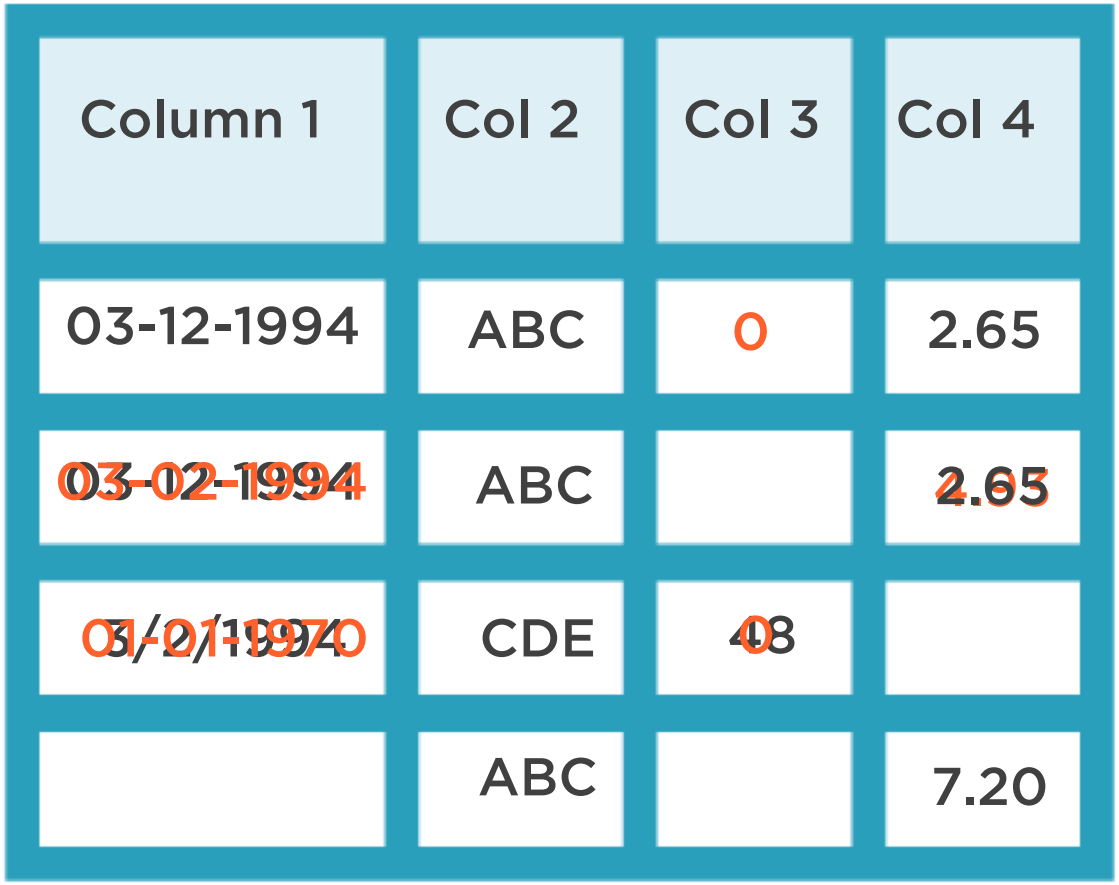


Tidy Data

Duplicates

De-dupe

Formats



Impute method

Zeros

Averages

Defaults

BI - Business Intelligence Reporting

Business Intelligence Developer

Uses BI tools like:

- MS SQL Server Reporting Services (SSRS)
- Business Objects (BO) Crystal Reports
- Tableau

Custom code for more complex reports

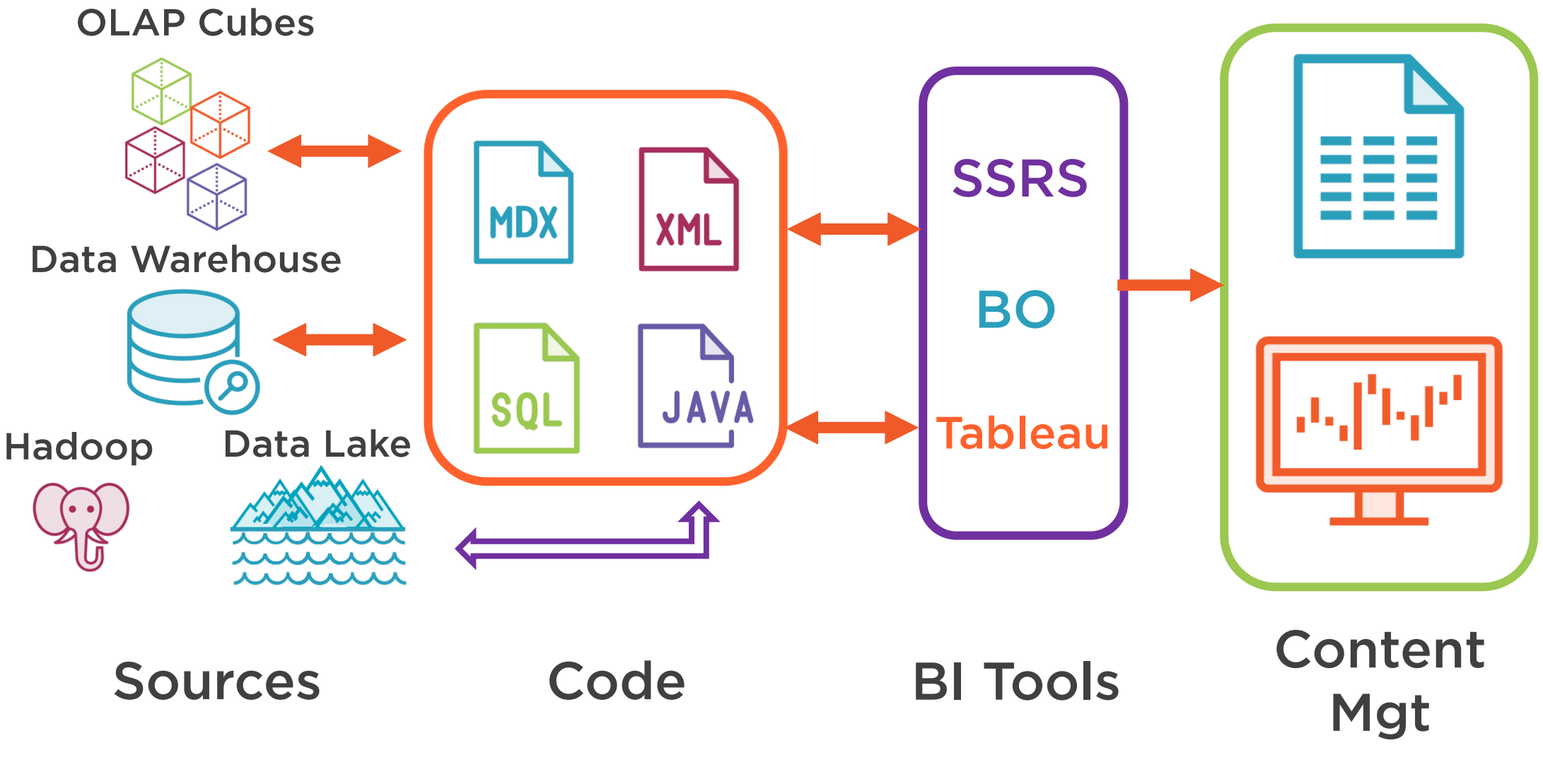
- MS-SSRS: C#
- Crystal Reports: XML
- Tableau SDK: C, C++, Java and Python

Designs Report Layout

Connects to data sources

- SQL script or stored procedure

The Typical BI Dataflow



Enterprise Report Content Management

SSRS Server

SharePoint

Business Objects Server

Tableau Server

Open Source Enterprise Content Management

Joombla

Plone

Magento

Wiki

Exploratory and Statistical Data Analytics

Exploratory Data Analysis – Who Does It?

Ecommerce SME

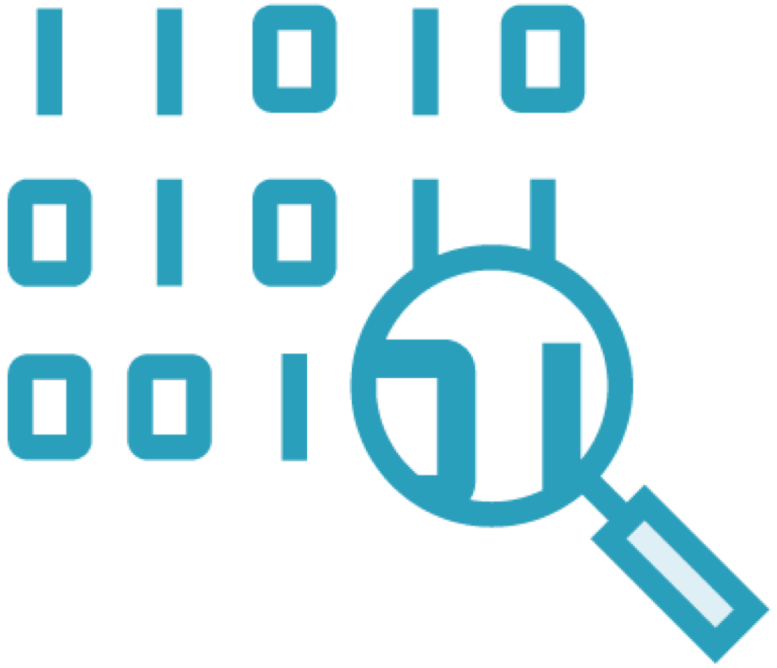
Sales SME

Call Center SME

Marketing SME

Subject Matter Experts

What is Exploratory Data Analysis?



Poking around, methodically

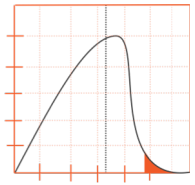
Patterns

"What if" analysis

Where data mining starts

A Data Science Tale

This is Maria



Maria does EDA and finds a curious pattern in the data

Maria tries a “what if” with a few values that further her curiosity

A Data Science Tale – Continued

Maria suspects as ad spending increases, sales numbers increase

But, Maria doesn't have the skill set to prove it.

Maria goes to Jamal, a data scientist at her company for help

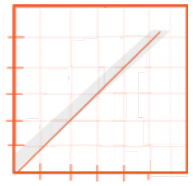
A Data Science Tale – Continued

Jamal needs to incorporate data from other data sets

Jamal asks Enrique, a data engineer to help him

Enrique develops a pipeline and tidies up the data for Jamal

A Data Science Tale – Continued



Jamal runs a linear regression analysis, the covariance is positive

Maria, Jamal and Enrique put together a data “viz” presentation

Maria gets a promotion, Jamal and Enrique get bonuses

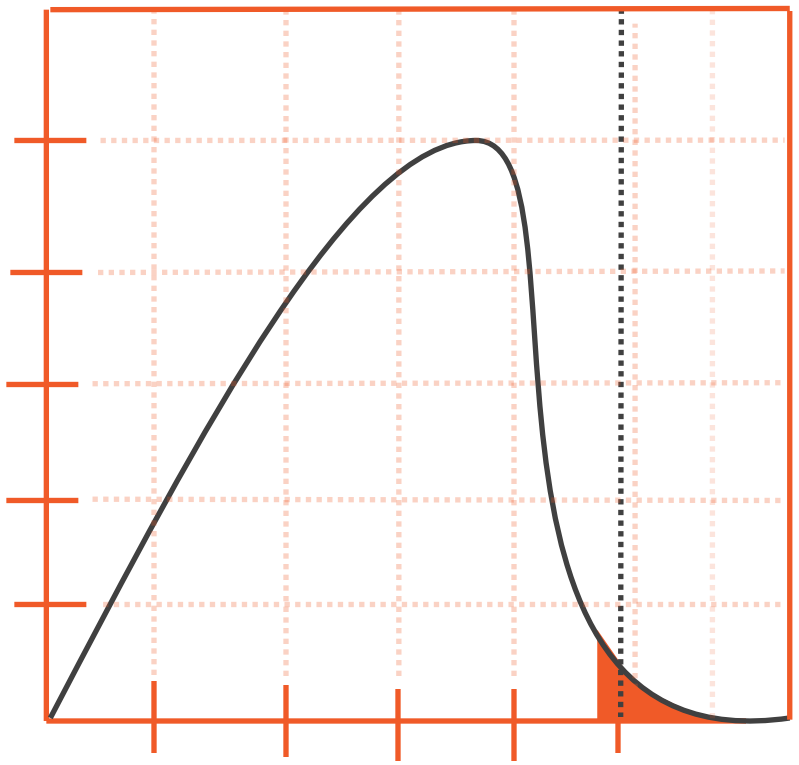
A Data Science Tale – Continued

Maria is happy

Jamal is happy

Enrique is happy

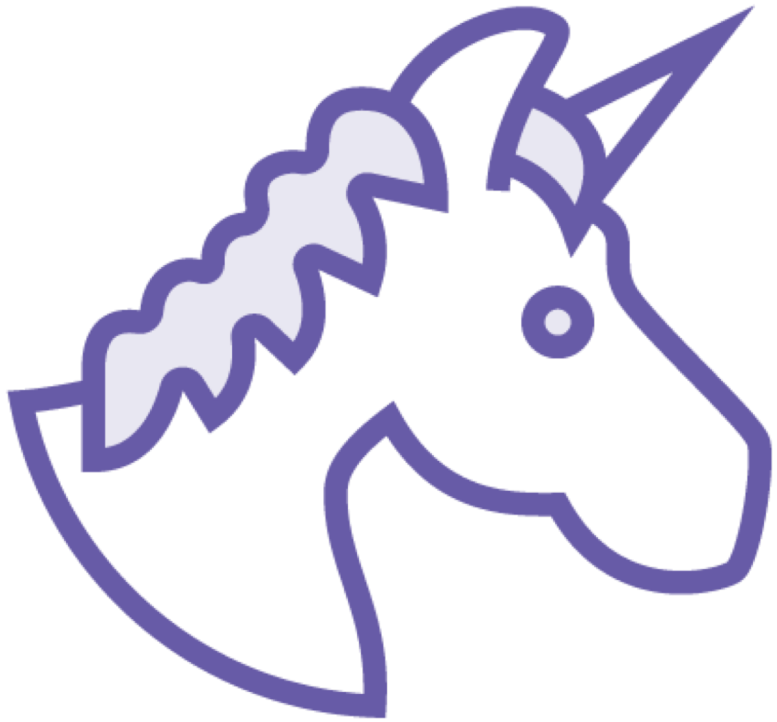
What Is Statistical Data Analysis?



The science part of data science

- Scientific method
- Sampling
- Hypotheses test
- Significance?

Who Is A Data Scientist?



Unicorns?

Statistician +

Data engineer +

Business analyst +

Graphic designer +

People person

MS/MA or PhD in a quant field

Predictive Modeling

What Is Predictive Modeling?

Forecasting from patterns in existing data

Choosing variables from a tidy dataset

Machine Learning Algorithms

- Supervised Learning
  - Train/Test datasets
  - Decision trees
  - Naïve Bayes Classification
  - Regression
  - Support Vector Machines (SVM)

What Is Predictive Modeling?

Machine Learning

- Unsupervised Learning
  - No training
  - Anomaly detection
  - Clustering
    - K-Means clustering
    - K-NN (a.k.a Nearest Neighbors)

What Is Predictive Modeling?

Machine Learning

- Semi - supervised learning
  - Supervised and unsupervised
- Reinforced
- Deep learning
  - Can be semi-, un-, supervised

Predictive Modeling Solutions

Paid solution packages

- SAS
- IBM's SPSS

Free tools

- Python libraries
- R libraries

Bringing It All Together

Exploratory data analysis

- Discover value

Statistical data analysis

- Understand the nature
- Provide evidence

Predictive modeling

- To make educated business decisions

Popular Machine Learning Libraries

TensorFlow (Python, C++, Java, and Go)

Scikit-Learn (Python)

Caret (R)

Data Visualization

What Is Data Visualization?

It's not your mama's pie chart

Telling a story

Infographics

- Canva
  - https://www.canva.com
- Venngage
  - https://venngage.com
- Pikochart
  - https://piktochart.com

Animated graphics

Google It:

Hans Rosling's 200 Countries,
200 Years, 4 Minutes

Data Visualization Libraries

- Matplotlib (Python)
- Bokeh (Python)
- Plotly (R)
- Ggplot2 (R)
- D3.js (custom Javascript)

Summary

- Data engineering (“Big Data”)
- Business intelligence (BI)
- Exploratory data analytics (EDA)
- Statistical data analytics
- Predictive modeling
- Data visualization (Data viz)

Next Up: Course Summary