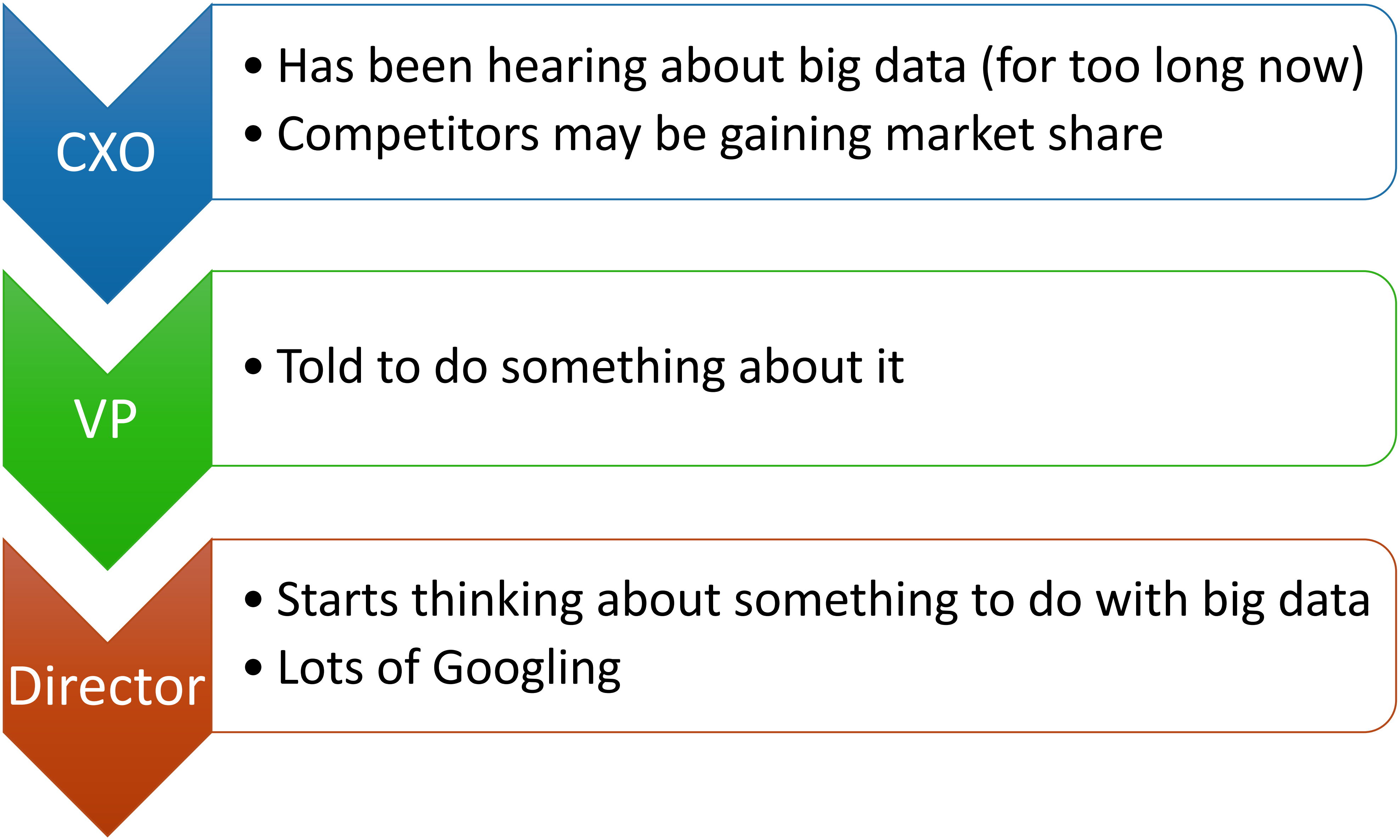
How to NOT miss the big picture in new big data initiatives

Jignesh M. Patel

University of Wisconsin

jignesh@cs.wisc.edu

A common scenario



CXO

- Has been hearing about big data (for too long now)
- Competitors may be gaining market share

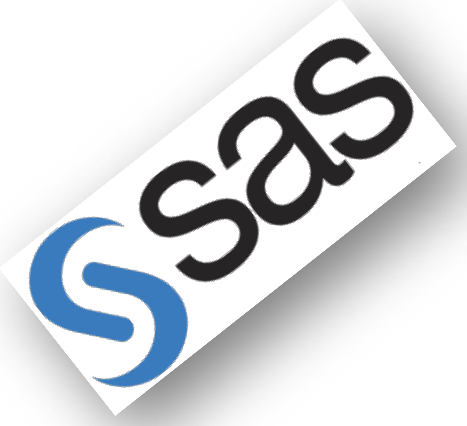
VP

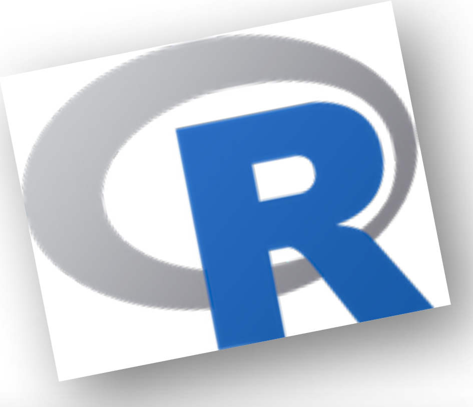
- Told to do something about it

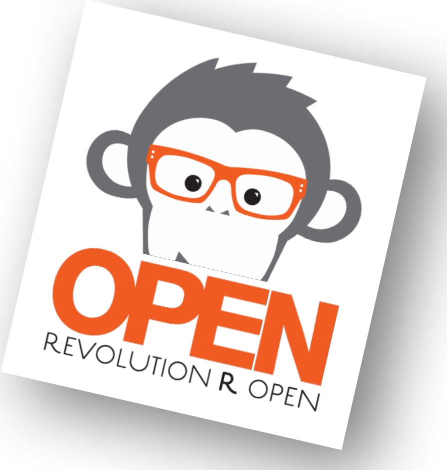
Director

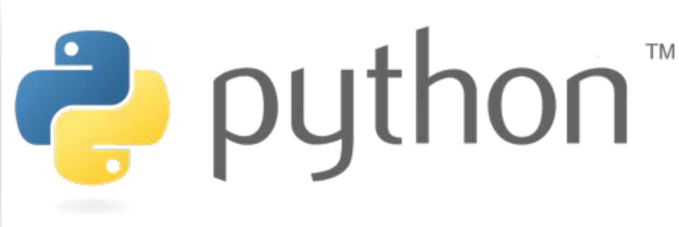
- Starts thinking about something to do with big data
- Lots of Googling

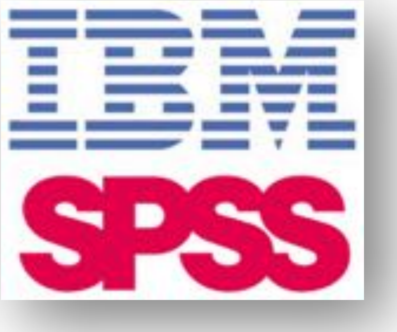
riak
cloudera
MySQL
IBM DB2
CouchDB relax
SAP
APACHE Spark
ORACLE
DATASTAX
NoSQL
SQL
Hortonworks
Microsoft SQL Server
Storm
Greenplum
HERON
APACHE HBASE
VERTICA An HP Company
NETEZZA an IBM Company
PostgreSQL
New SQL
HIVE
mongoDB
HAWQ

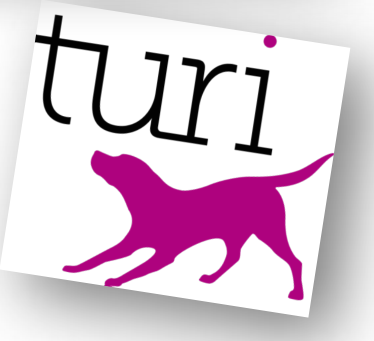


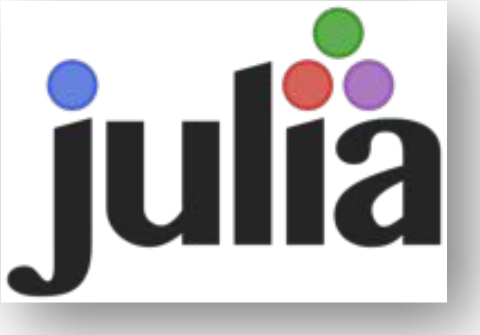






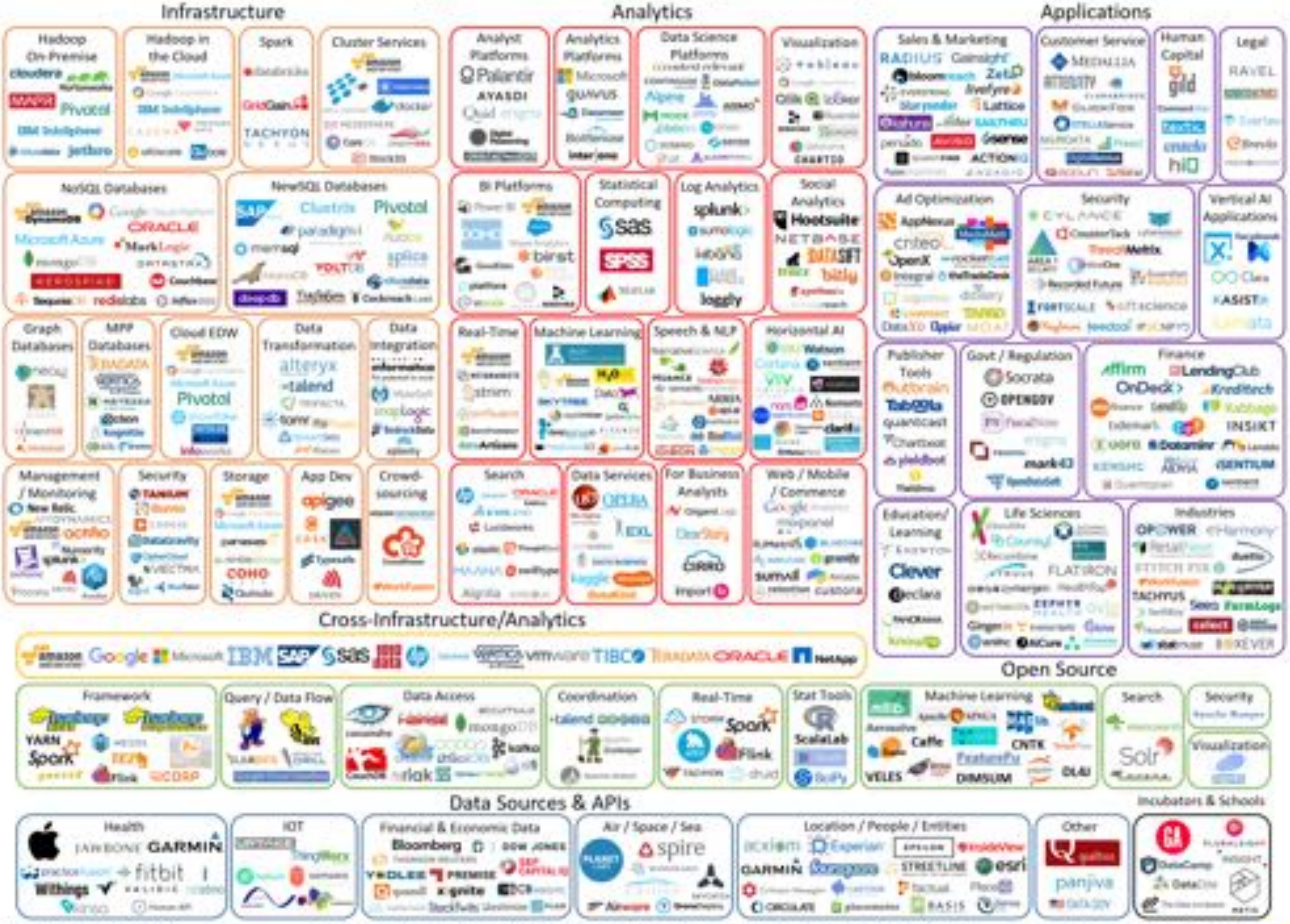






An even longer list at: https://en.wikipedia.org/wiki/List_of_statistical_packages

Big Data Landscape 2016 (Version 3.0)



Infrastructure

Hadoop On-Premise

Hadoop in the Cloud

Spark

Cluster Services

NoSQL Databases

NewSQL Databases

Graph Databases

MPP Databases

Cloud EDW

Data Transformation

Data Integration

Management / Monitoring

Security

Storage

App Dev

Crowd-sourcing

Analytics

Analyst Platforms

Analytics Platforms

Data Science Platforms

Visualization

BI Platforms

Statistical Computing

Log Analytics

Social Analytics

Real-Time

Machine Learning

Speech & NLP

Horizontal AI

Search

Data Services

For Business Analysts

Web / Mobile / Commerce

Applications

Sales & Marketing

Customer Service

Human Capital

Legal

Ad Optimization

Security

Vertical AI Applications

Publisher Tools

Govt / Regulation

Finance

Education/ Learning

Life Sciences

Industries

Cross-Infrastructure/Analytics

Open Source

Framework

Query / Data Flow

Data Access

Coordination

Real-Time

Stat Tools

Machine Learning

Search

Security

Visualization

Data Sources & APIs

Health

IOT

Financial & Economic Data

Air / Space / Sea

Location / People / Entities

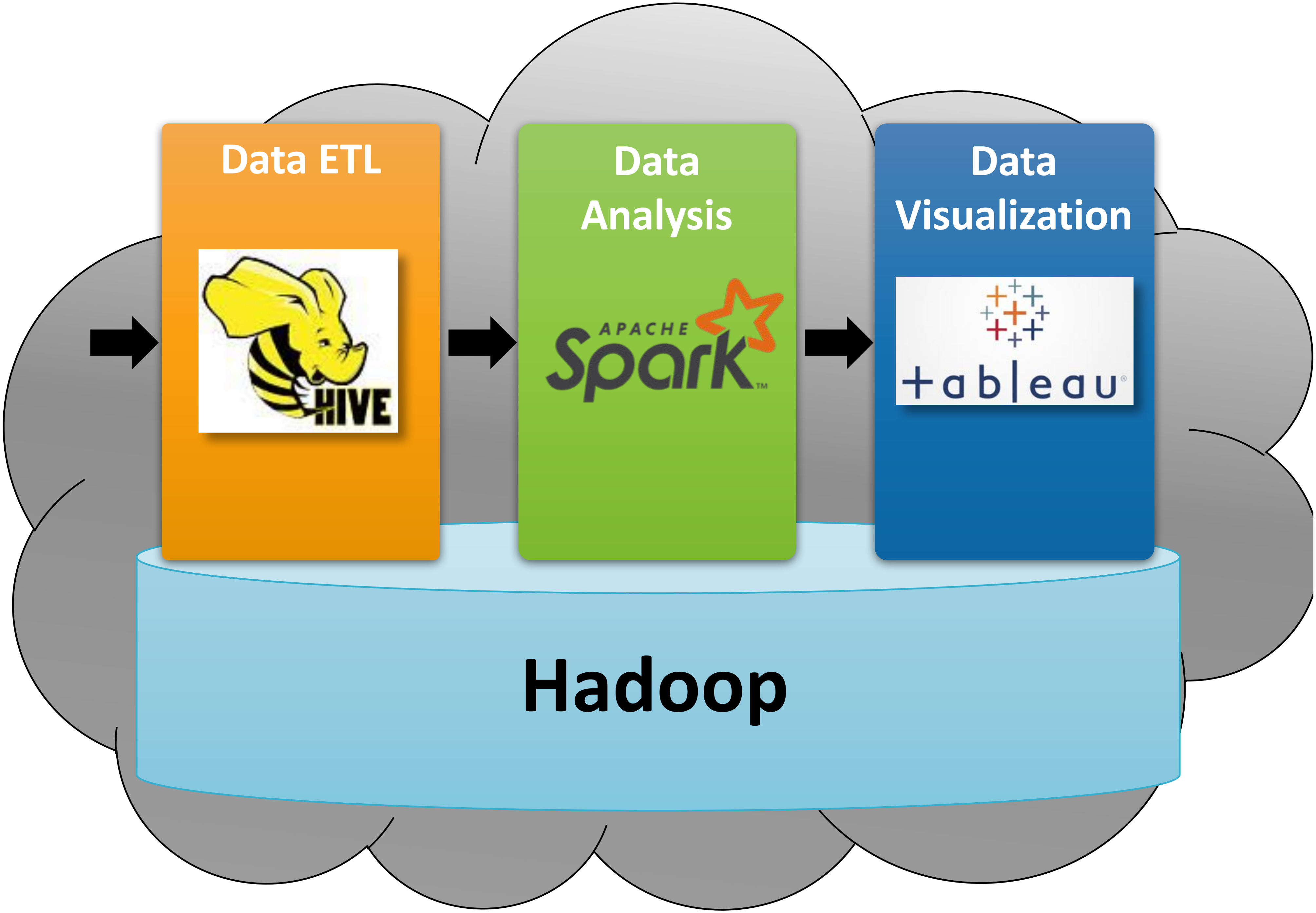
Other

Incubators & Schools

Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK



Data ETL

Data Analysis

Data Visualization

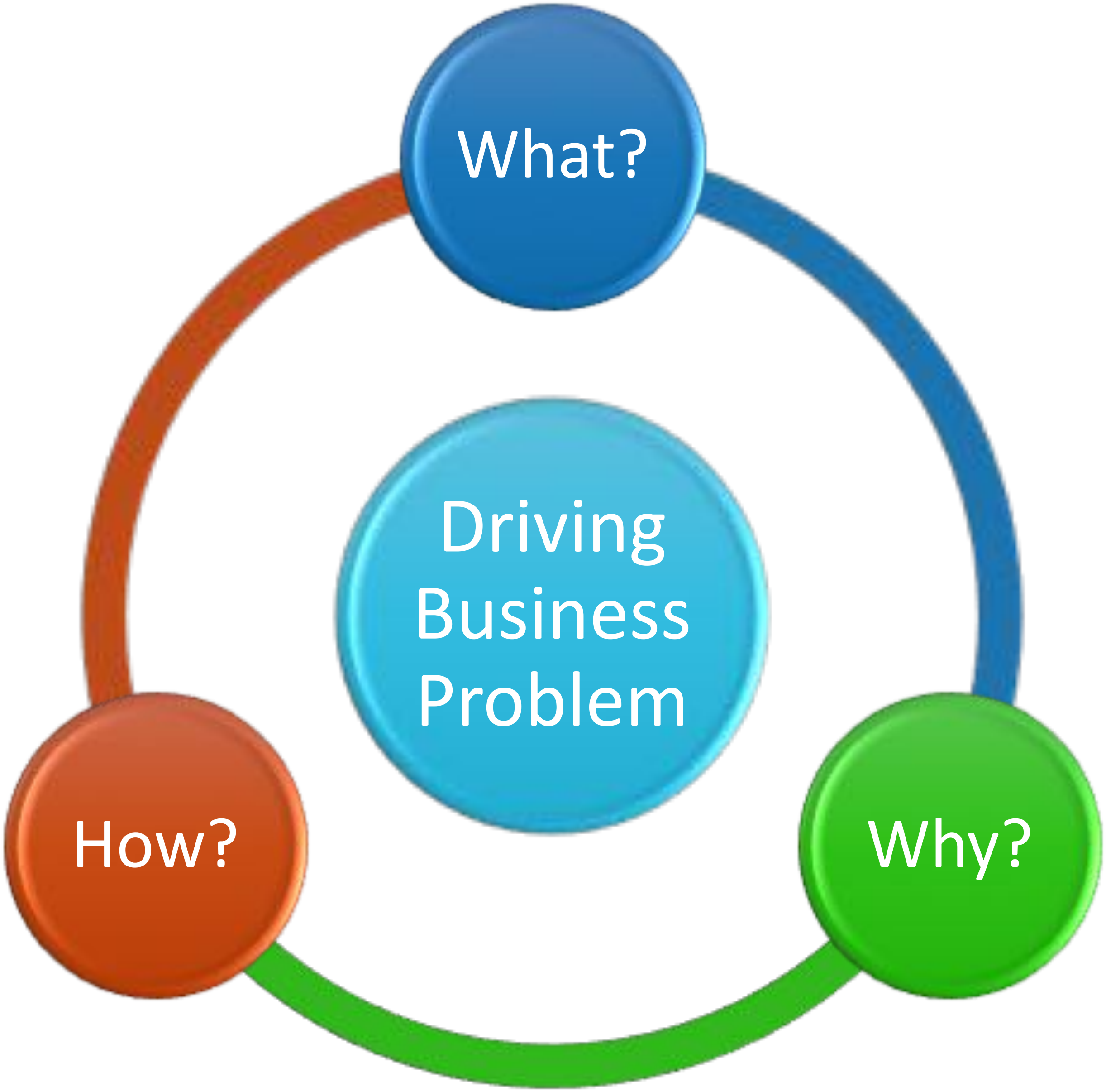
Hadoop

Outcome: Bundle of buzzword compliant technologies

Refine, retune, redesign …

Hire consultants …

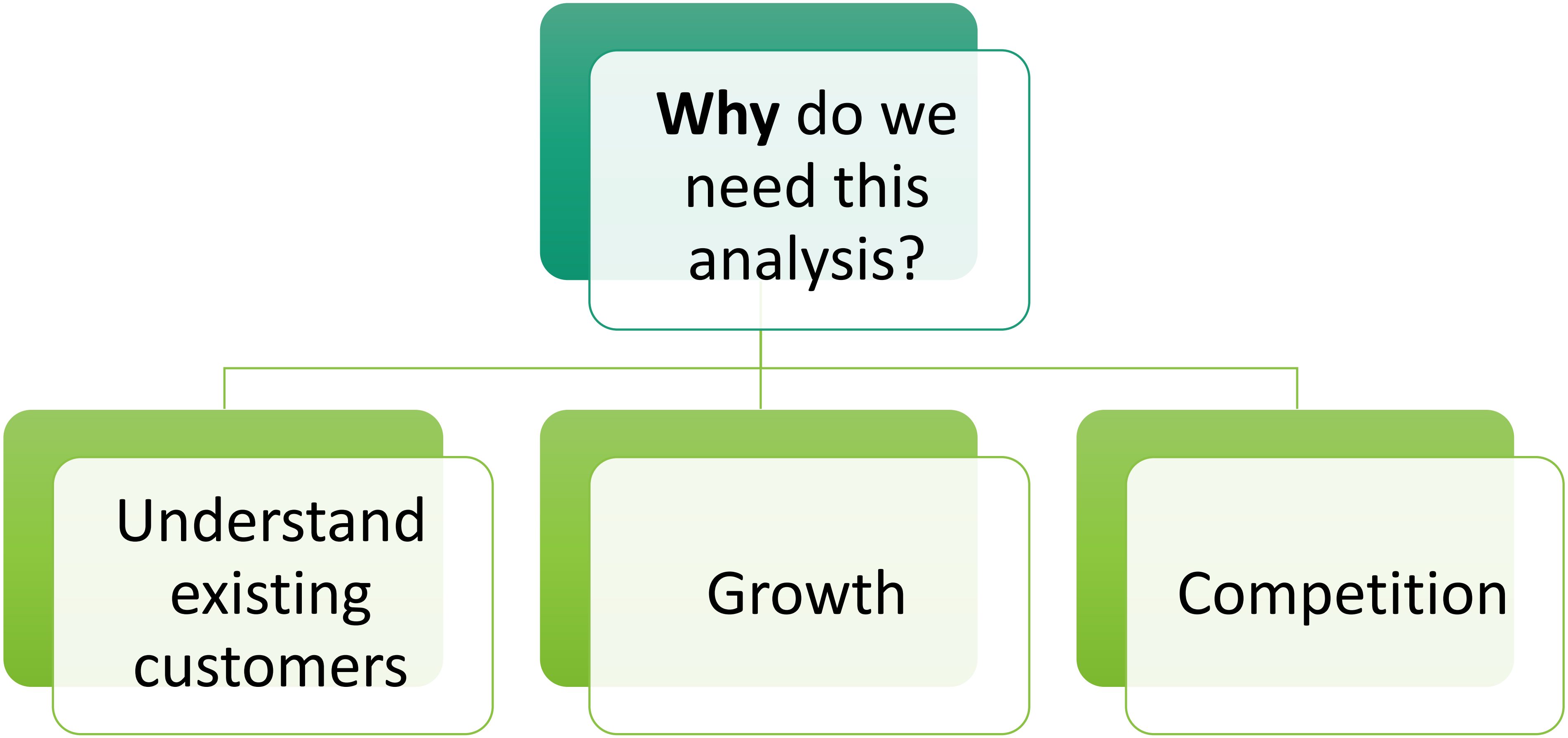
Repeat …


What?
Driving Business Problem
How?
Why?

What is the business problem?

ROI Analysis

It is an investment for a business!

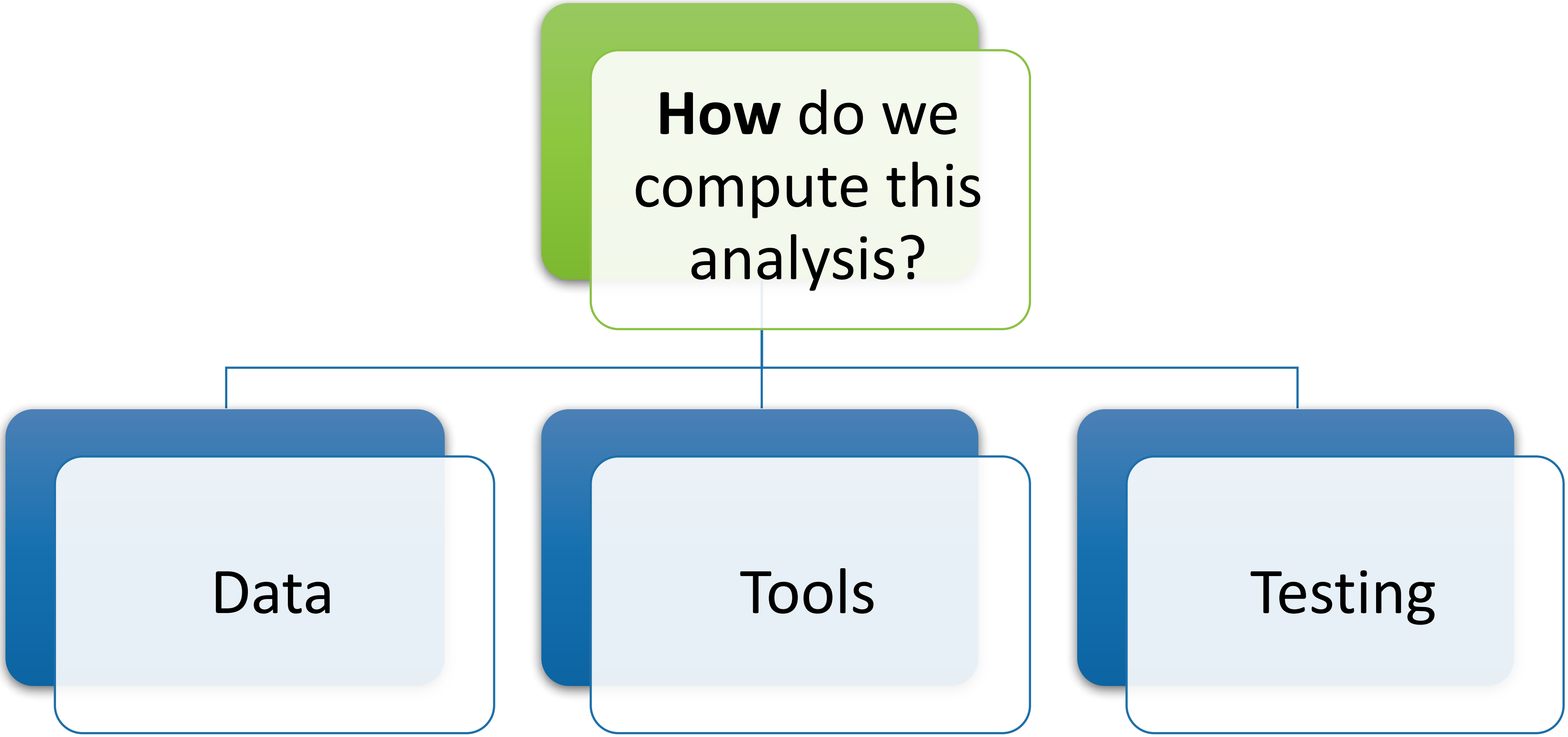


Why do we need this analysis?

Understand existing customers

Growth

Competition



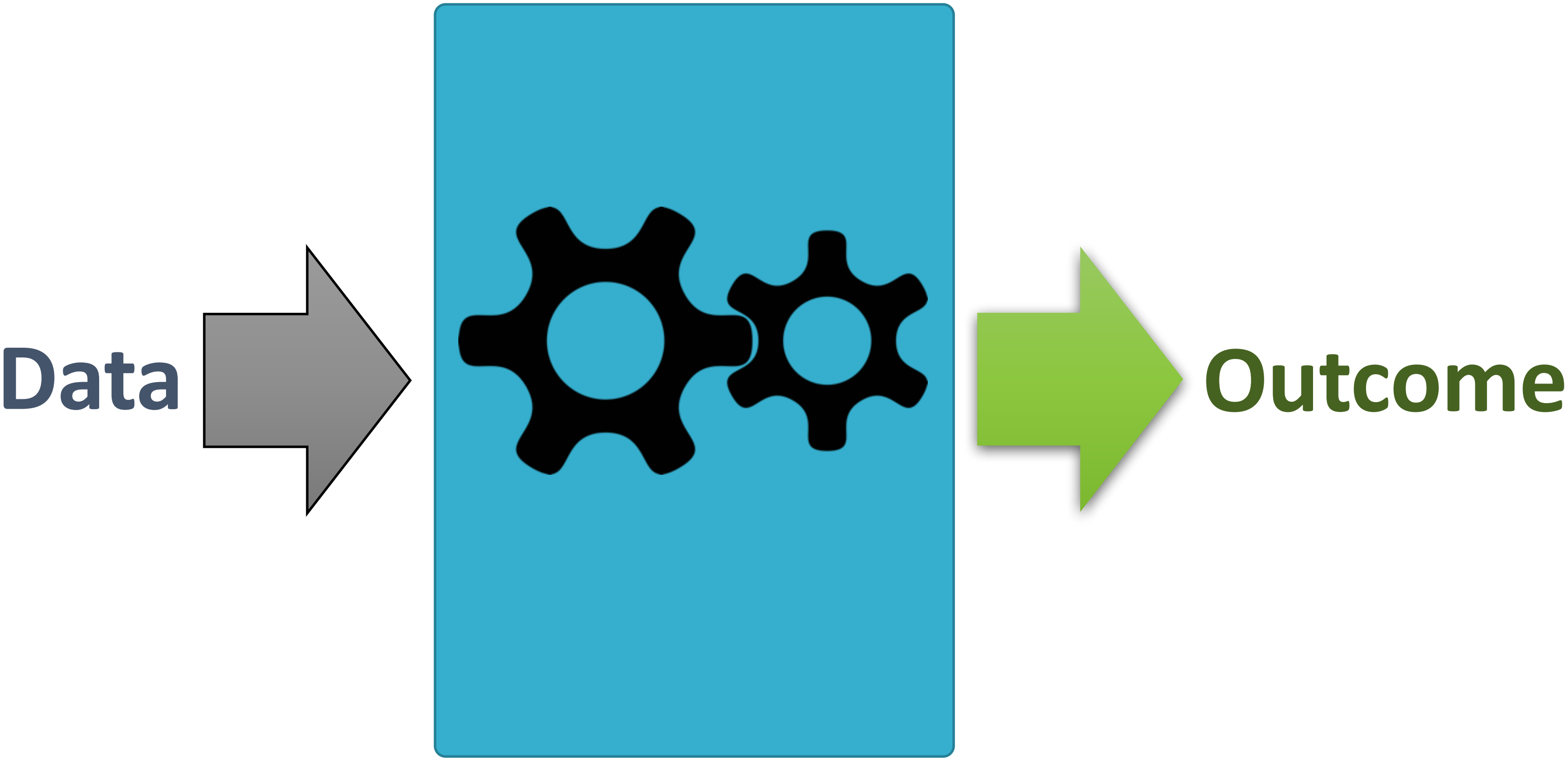
How do we compute this analysis?

Data

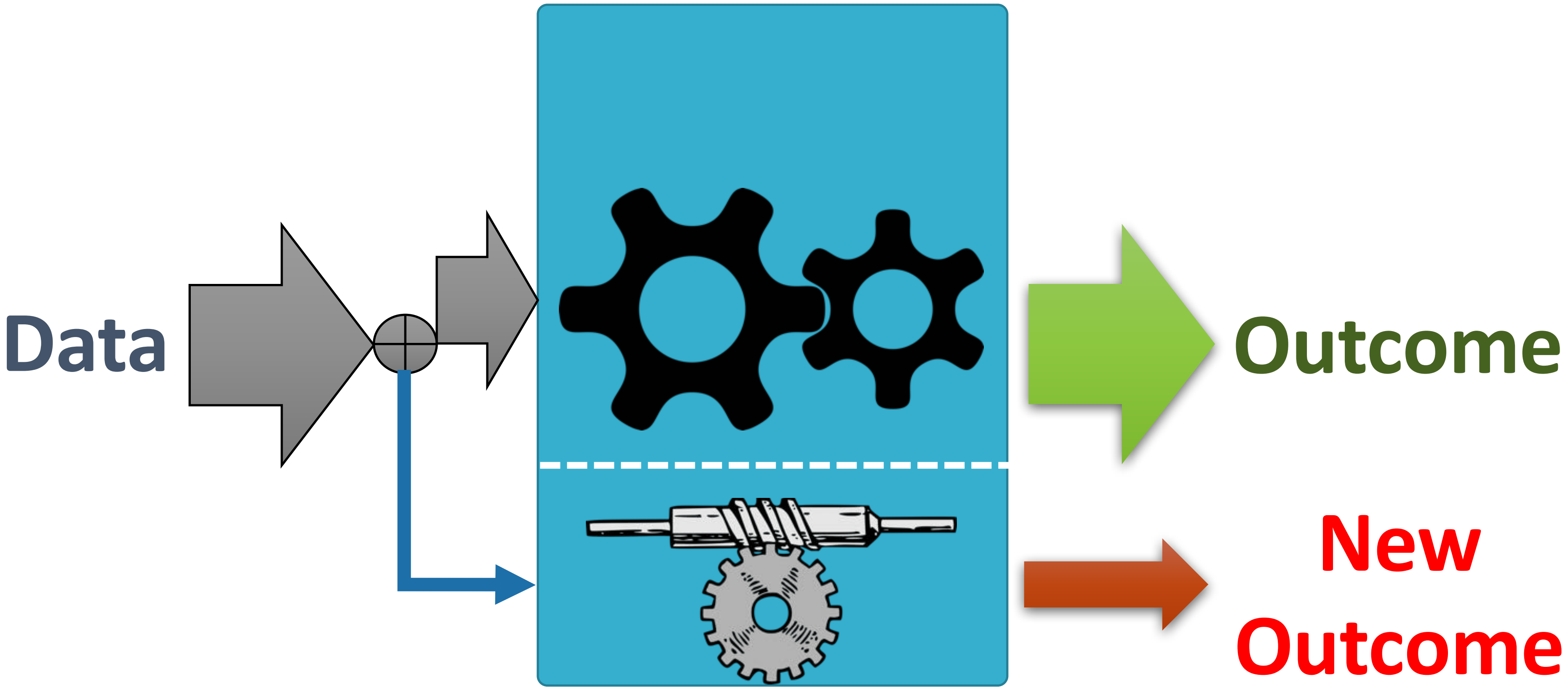
Tools

Testing

A/B testing infrastructure



A/B testing infrastructure

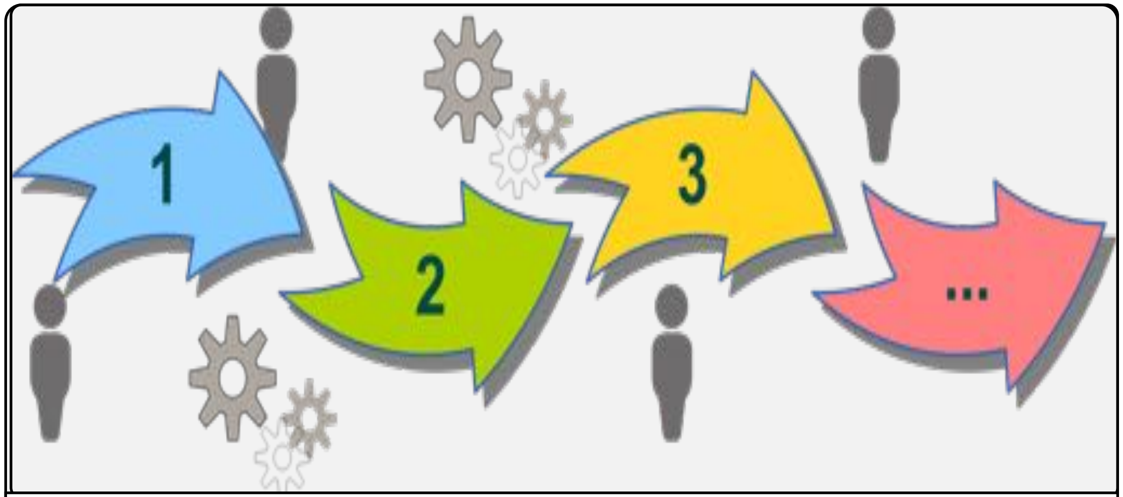


Data

Outcome

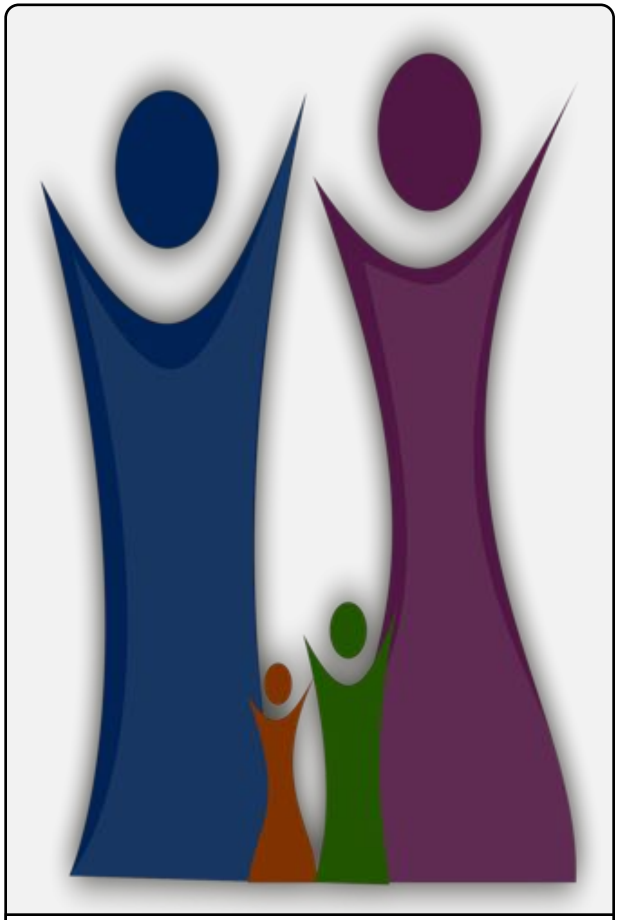
New Outcome

In addition …



Process

Agile, transparent, and **short** iterations. And ego-less.



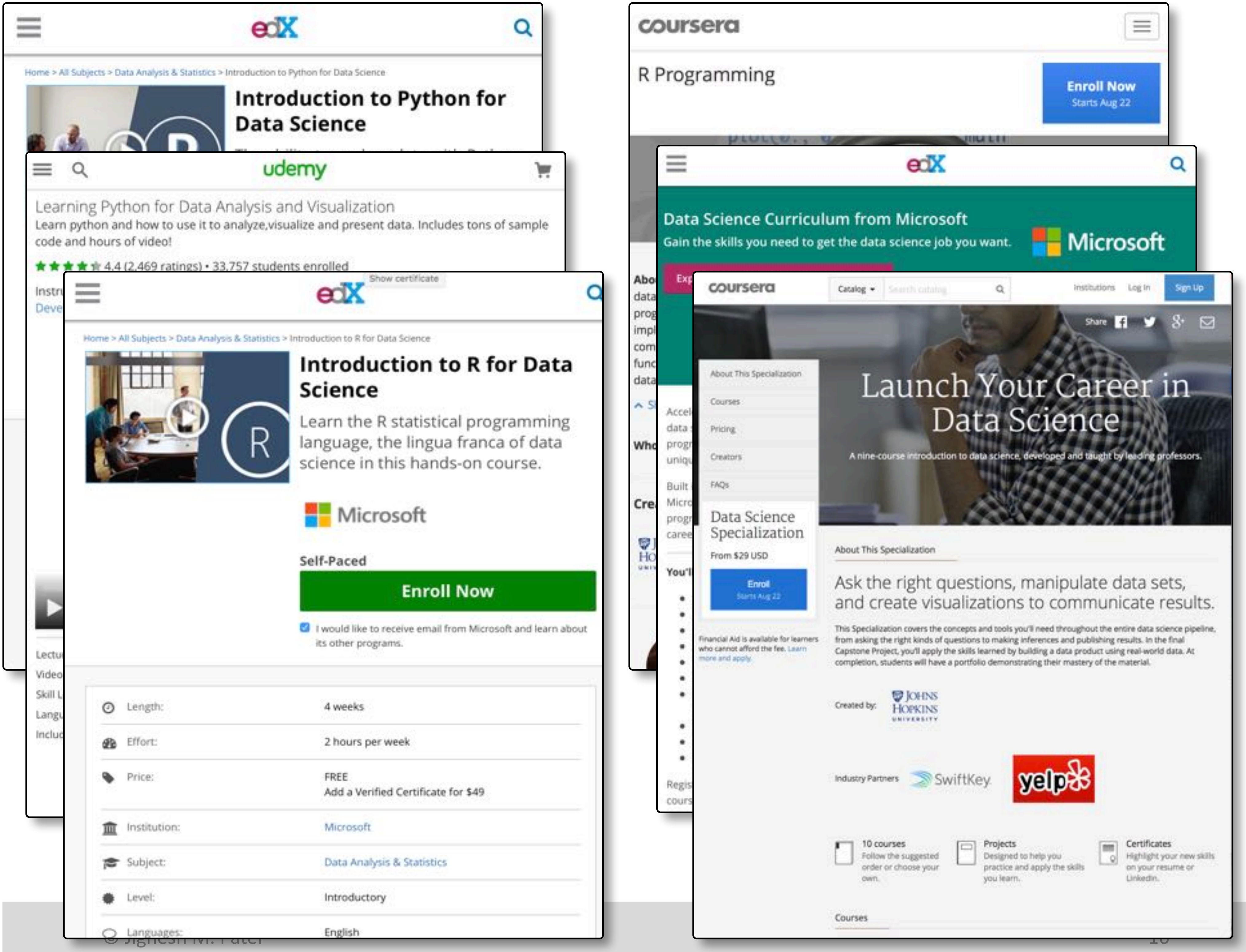
People

Invest in your existing people!

Democratization of data and tools

Open-source tools

Open access to education



edX

Home > All Subjects > Data Analysis & Statistics > Introduction to Python for Data Science

Introduction to Python for Data Science

udemy

Learning Python for Data Analysis and Visualization

Learn python and how to use it to analyze,visualize and present data. Includes tons of sample code and hours of video!

★★★★☆ 4.4 (2,469 ratings) • 33,757 students enrolled

coursera

R Programming

Enroll Now
Starts Aug 22

edX

Data Science Curriculum from Microsoft

Gain the skills you need to get the data science job you want.

Microsoft

edX

Show certificate

Home > All Subjects > Data Analysis & Statistics > Introduction to R for Data Science

Introduction to R for Data Science

Learn the R statistical programming language, the lingua franca of data science in this hands-on course.

Microsoft

Self-Paced

Enroll Now

☑ I would like to receive email from Microsoft and learn about its other programs.

Length:	4 weeks
Effort:	2 hours per week
Price:	FREE Add a Verified Certificate for $49
Institution:	Microsoft
Subject:	Data Analysis & Statistics
Level:	Introductory
Languages:	English

coursera

Catalog ▾ Search catalog

Institutions Log In Sign Up

Share

About This Specialization

Courses

Pricing

Creators

FAQs

Data Science Specialization

From $29 USD

Enroll
Starts Aug 22

Financial Aid is available for learners who cannot afford the fee. Learn more and apply.

Launch Your Career in Data Science

A nine-course introduction to data science, developed and taught by leading professors.

About This Specialization

Ask the right questions, manipulate data sets, and create visualizations to communicate results.

This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.

Created by: Johns Hopkins University

Industry Partners SwiftKey yelp

10 courses
Follow the suggested order or choose your own.

Projects
Designed to help you practice and apply the skills you learn.

Certificates
Highlight your new skills on your resume or LinkedIn.

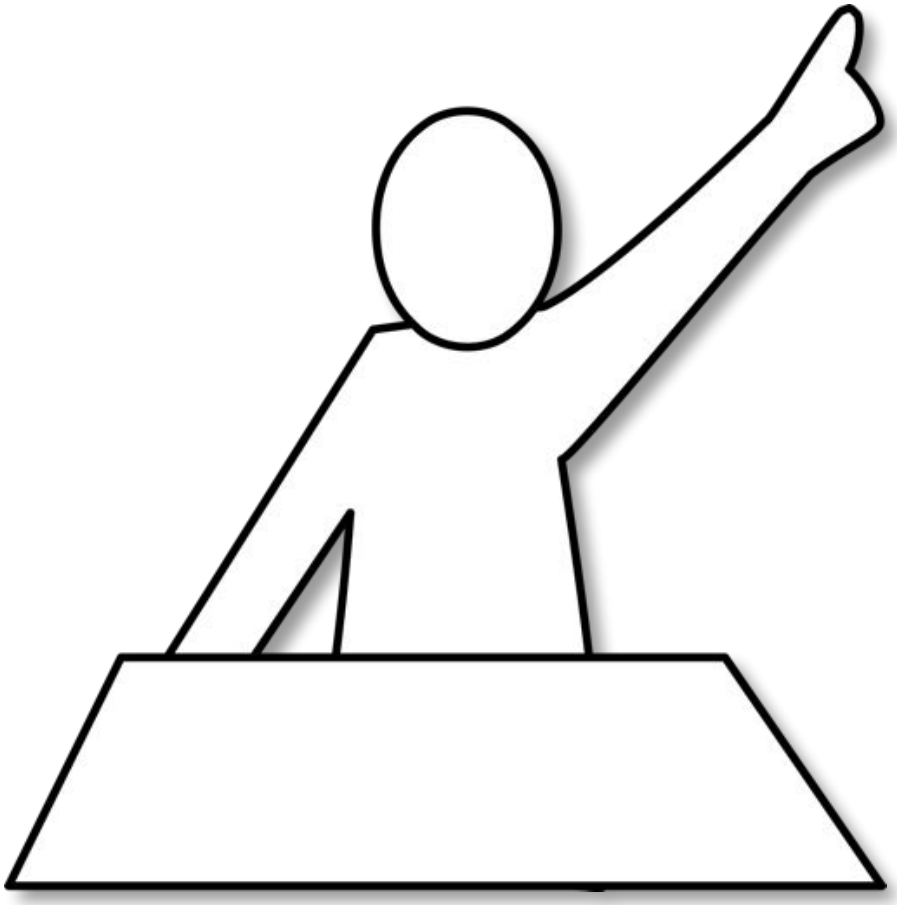
Courses

Summary

Goal-oriented approach, with clear ROI

Separate means from the end

(Re-)invest in your existing people



My email: jignesh@cs.wisc.edu