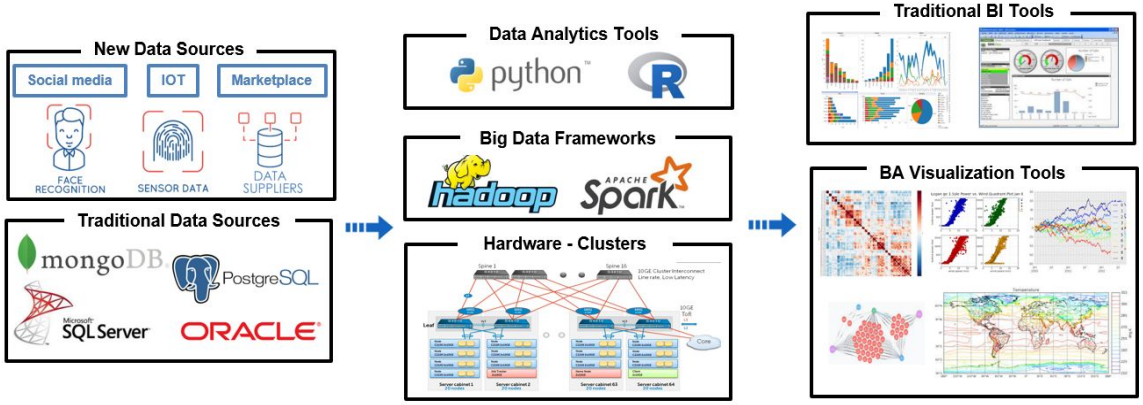


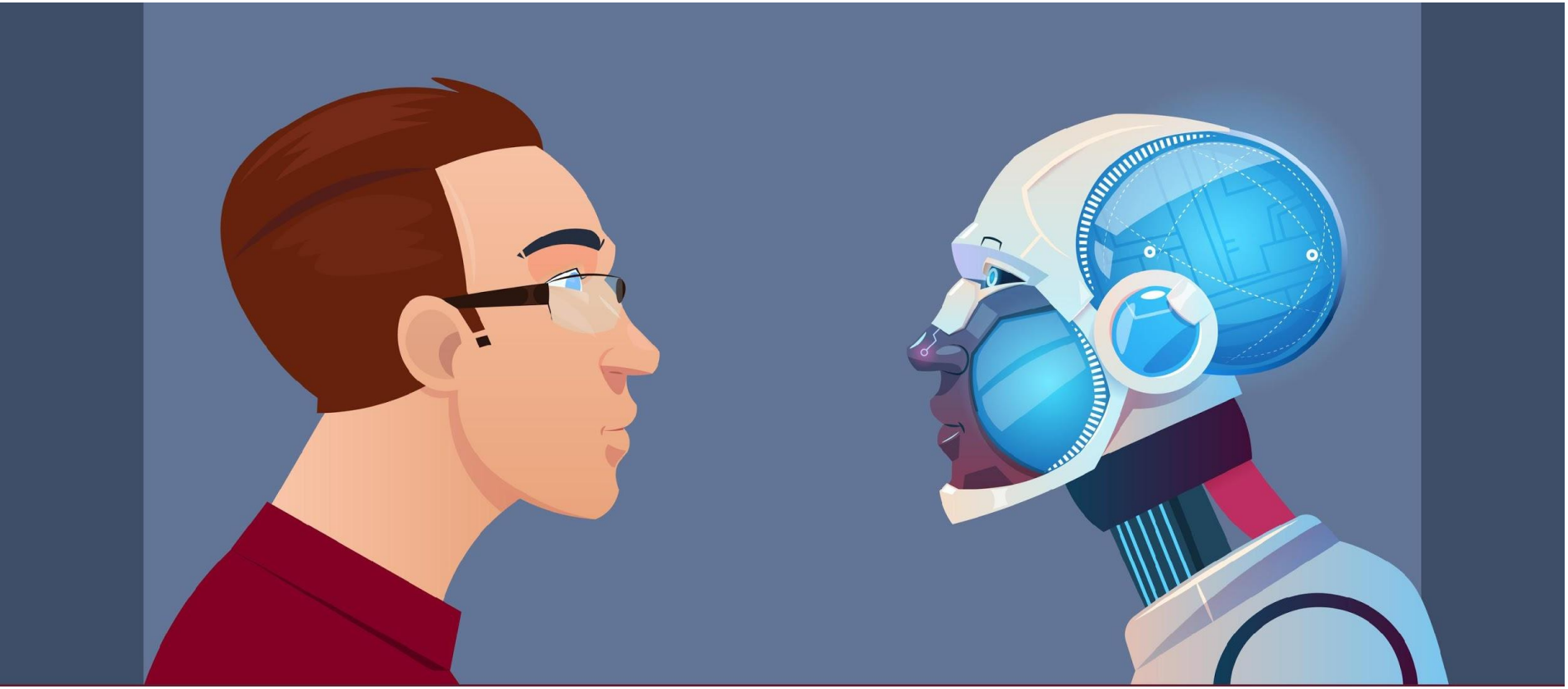
Data Analytics for Senior IT Managers (Day 1)

Laurence Liew
Dr Guo Lei
Jeanne Choo

NUS National University of Singapore | School of Continuing and Lifelong Education

What are the "data analytics" type questions you get asked and would like to know how to answer from this class?

OBJECTIVE OF THIS COURSE



OUTLINE (Day 1)

Day 1

1. Advancing Data Analytics at Organisations
   - Painting the big picture - big data and data analytics
   - Overview of Machine Learning
   - The Data Science Stack
   - The Data Science Team
   - Centre of Excellence for Analytics (a template)
   - Analytics Use Cases

2. Panel Discussion: A Deep Dive into 5 Categories of Analytics Use Cases with real-life examples
   - Understanding customers (e.g. behavioural segmentation for permit application)
   - Optimising delivery services (e.g. hospital bed optimisation and prediction of length of stay)
   - Measuring effectiveness (e.g. measuring how interventions impact school performance in the UK)
   - Fraud detection (e.g. detecting credit card fraud with Kaggle data, mapping Hillary Clinton's politicians social networks through her email contacts)
   - Policy review and planning (eg. Randomised controlled trial for policy testing)

3. Project Scoping Workshop Part 1
   - Analytics projects: What works and what doesn't
   - Introduction to take-home exercise: Plan your own project (discussion on Day 2)

4. Case Study (self review)
   - Government data analytics use cases (1 hour)
   - Data analytics training roadmap for WOG (0.5 hour)
   - Government infrastructure for WOG data analytics projects (1 hour)

OUTLINE (Day 2)

Day 2

1. Managing Data Analytics Projects
   - Open source and analytics
   - Data analytics infrastructure and tools
   - Reproducible Analytics Production Pipeline
   - The case for analytics in the cloud (public or on premise)
   - Introduction to artificial intelligence and deep learning

2. Project Scoping Workshop Part 2
   - Participants to plan their presentations
   - Participants present their ideas
   - Peer evaluation and discussion of project proposals

3. Case Study (self review)
   - Doing data analytics at agencies (1 hour)
     - Roles & resources GovTech plays in the WOG Data Science capability building
   - Moving up the maturity model (1 hour)
     - What agencies can do next

MACHINE LEARNING

A FAMOUS DATA SCIENTIST SAID THIS…

… tell us not to leave things to chance or luck. Everything must be examined, measured and calculated. You can know the outcome by calculation at the beginning.

- Sun Zi – The Art of War, 5th Century BC

NO MAGIC!

#MachineLearning itself does not magically solve your problems. Understanding your #data does. Use #CommonSense

~ Peter Czanik, balabit.com

TECH | 2/16/2012 @ 11:02AM | 2,492,601 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

335 comments, 174 called-out + Comment Now + Fo... Co...nts

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, w... need, and which coupons ar... most likely to make you h...ppy. Target, for example, has fig... how data-mine its way into your ...m, to figure out whether yo... a baby on the way long before you need to start buying diapers.

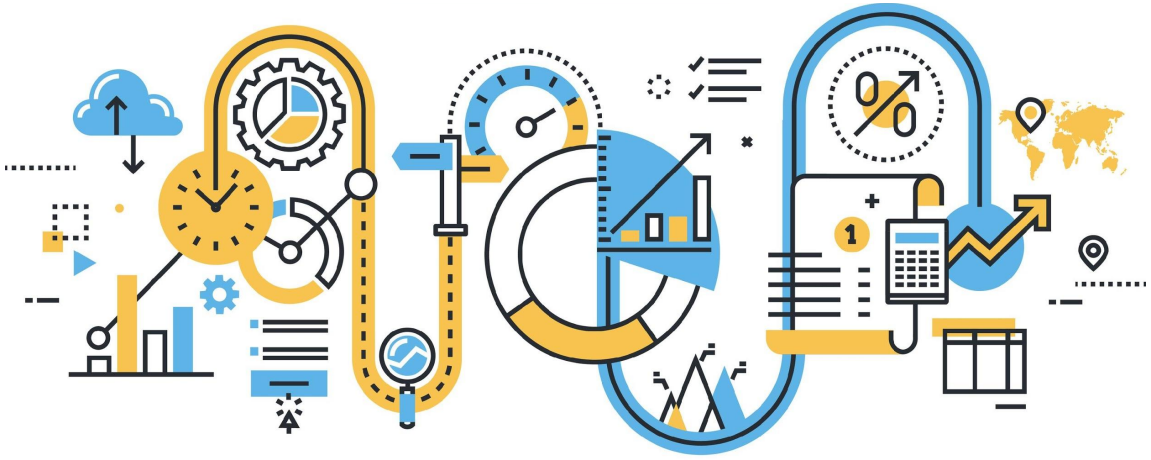
Target has got you in its aim

1k
Share
16.5k
Tweet
6.1k
Share
2.3k
reddit

$23Billion

Target knows before it shows.

Target's gangbusters revenue growth — $44 billion in 2002, when Pole was hired, to $67 billion in 2010 — is attributable to Pole's helping the retail giant corner the baby-on-board market, citing company president Gregg Steinhafel boasting to investors about the company's "heightened focus on items and categories that appeal to specific guest segments such as mom and baby."



@10,000m on Big Data, Analytics, Machine Learning/Artificial Intelligence

OVERVIEW

WHAT IS DATA SCIENCE

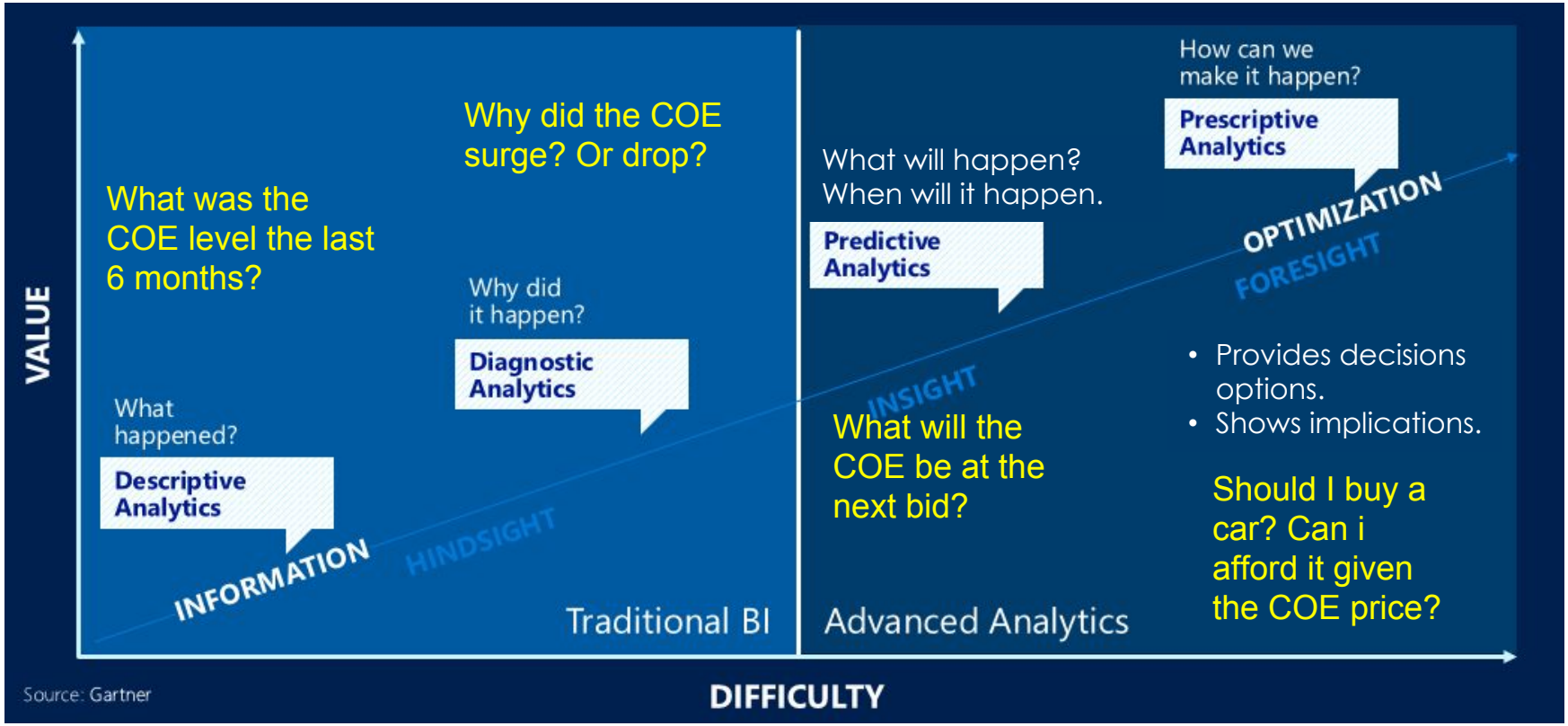
DATA ANALYTICS? MACHINE LEARNING

Data science (and analytics) is the transformation of data using mathematics and statistics into valuable insights, decisions, and products

- John Foreman, Chief Data Scientist, Mailchimp.com

Basically… allows computers to find hidden insights from the data without being told what to look for or program to do

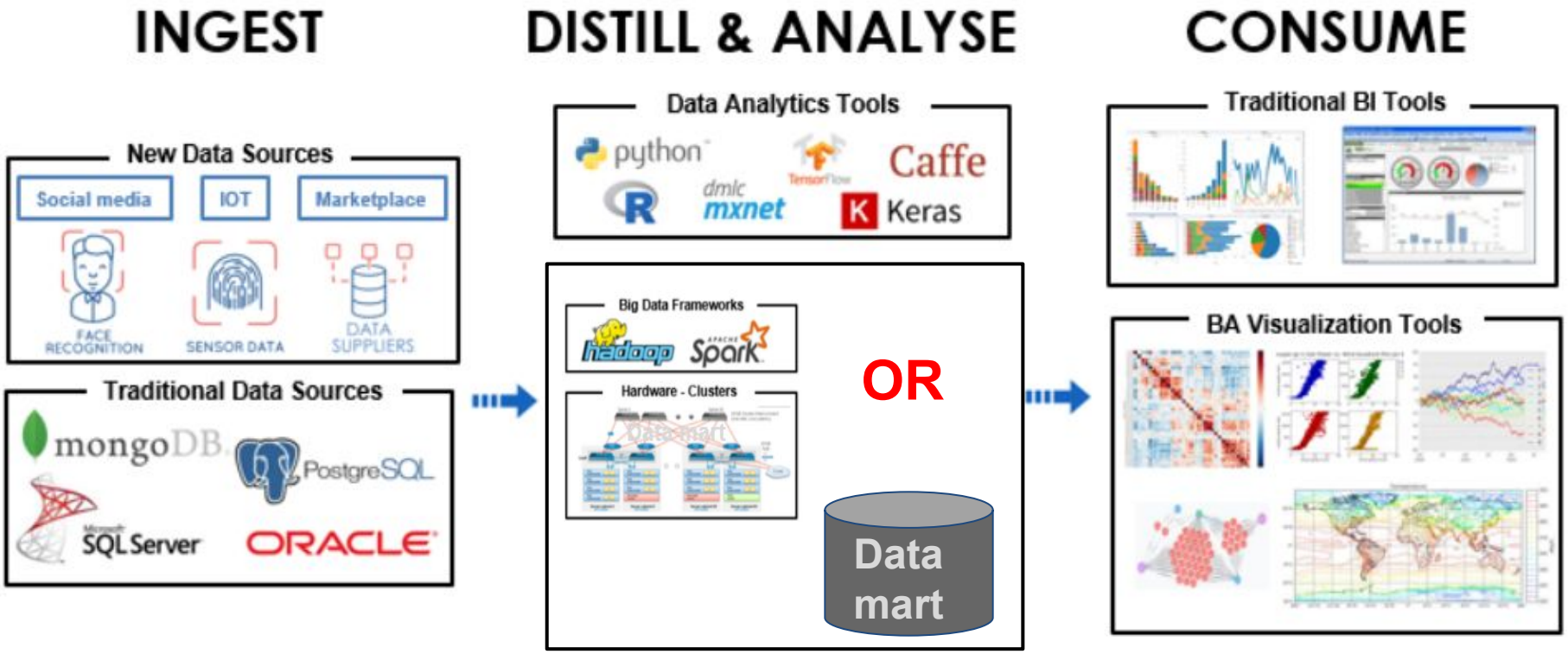
BEYOND BUSINESS INTELLIGENCE(BI)



Actionable Insights!

Source: Microsoft

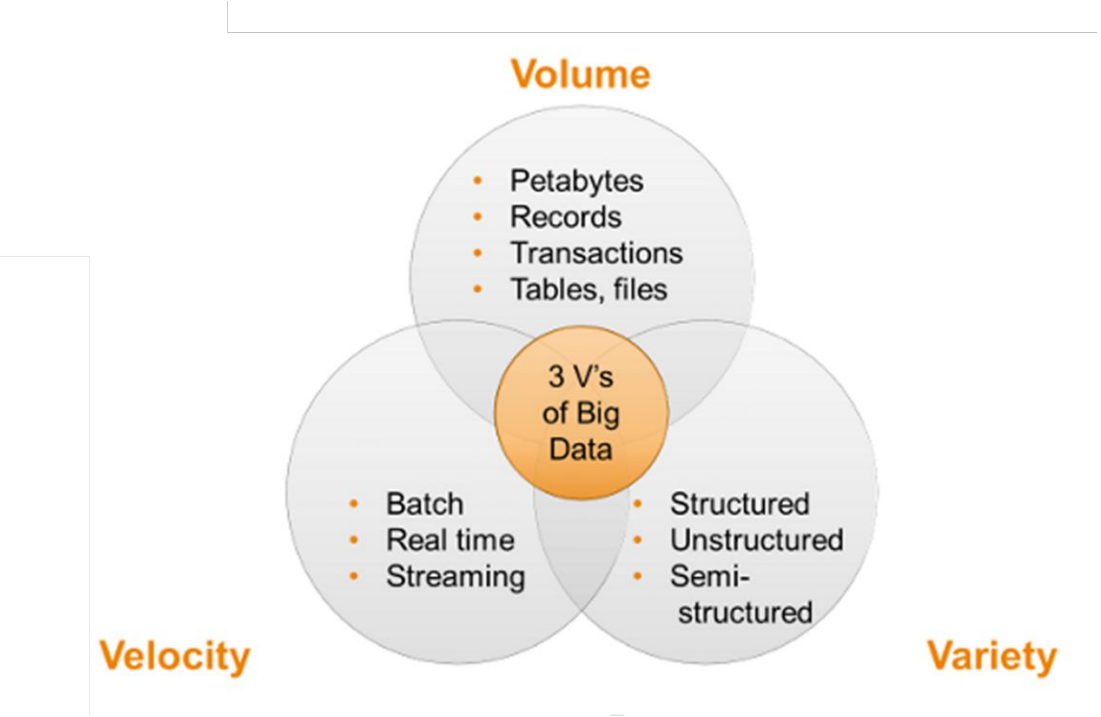
BIG DATA ANALYTICS OVERVIEW



We are going to dig deeper into the infrastructure required to support the analytics process.

Who thinks he/she have big data requirements?

WHAT IS BIG DATA?



When data itself becomes part of the problem!
Too large/fast for current enterprise systems.

BIG DATA INFRASTRUCTURE

Hadoop V1
Internet SCALE

Hadoop V2 aka SPARK
Enterprise SCALE (SINGAPORE)

R640, 2 CPU Xeon GOLD, 128GBRAM, 2 x M2 240 RAID 1 boot
4 x 1.92TB SSD 12Gbps RAID10 = 3.8GB (MDT)

R640, 2 CPU Xeon GOLD, 128GBRAM, 2 x M2 240 RAID 1 boot
4 x 1.92TB SSD 12Gbps RAID10 = 3.8GB (MDT)

R640, 2 CPU Xeon GOLD, 128GBRAM, 2 x M2 240 RAID 1 boot
4 x 1.92TB SSD 12Gbps RAID10 = 3.8GB (MDT)

R740xd, 2 CPU Xeon GOLD, 256GBRAM, 2 x M2 240 RAID 1 boot,
4 x 1.92TB SSD 12Gbps RAID10 = 3.8TB (MDT)
24 x 960GB SSD RAID6 (10+2) = 19.2TB (OST-FAST)

R740xd, 2 CPU Xeon GOLD, 256GBRAM, 2 x M2 240 RAID 1 boot,
4 x 1.92TB SSD 12Gbps RAID10 = 3.8TB (MDT)
24 x 960GB SSD RAID6 (10+2) = 19.2TB (OST-FAST)

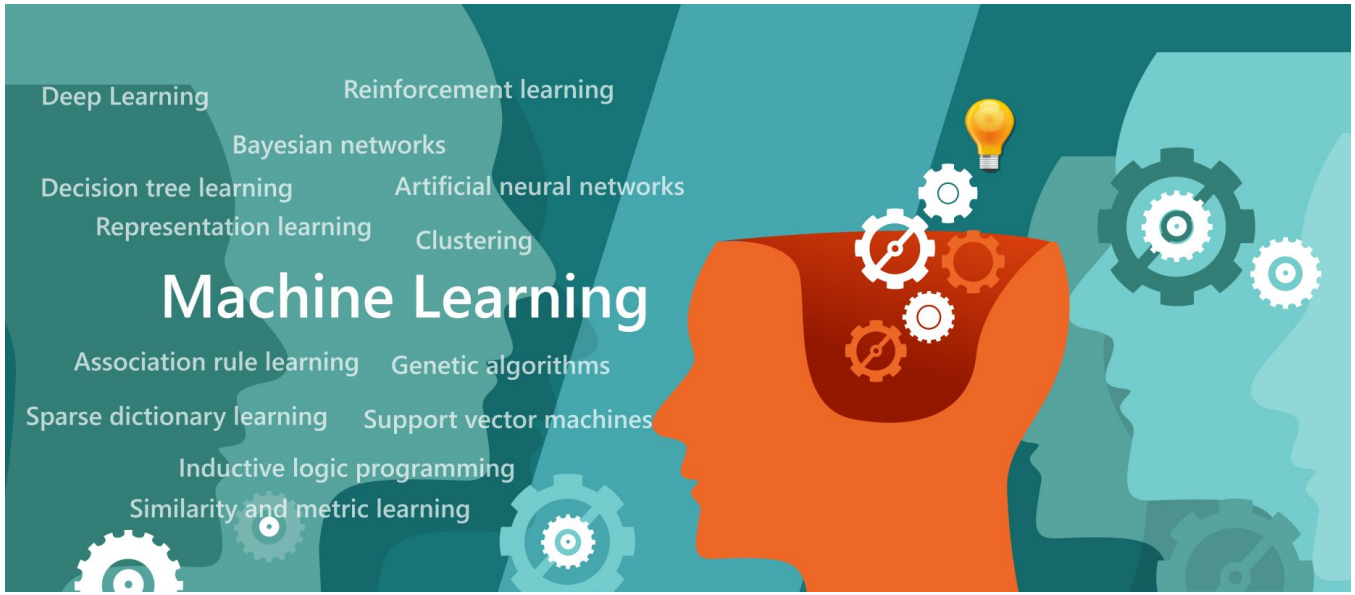
R740xd, 2 CPU Xeon GOLD, 256GBRAM, 2 x M2 240 RAID 1 boot,
4 x 1.92TB SSD 12Gbps RAID10 = 3.8TB (MDT)
24 x 960GB SSD RAID6 (10+2) = 19.2TB (OST-FAST)

19.2TB RAW per node

If Replication factor = 3
19.2TB usable in cluster

SUMMARY

- Business Intelligence (Reporting, Dashboards, Visualization) is not Data Science
- You do not need BIG Data to do Data Science



OVERVIEW

MACHINE LEARNING

DS | DA | ML ?

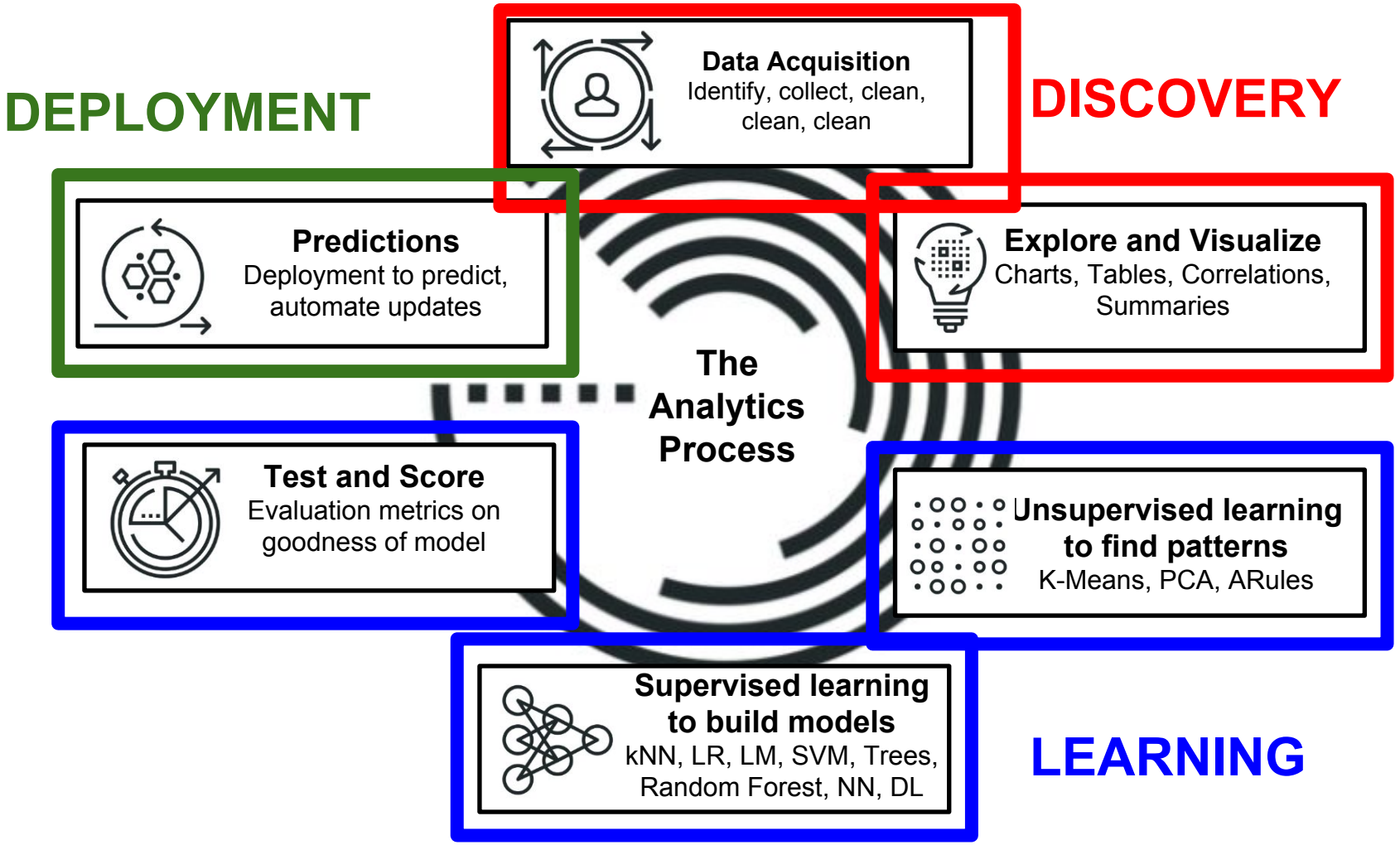
I will treat “data analytics” (DS) == “data science” (DS)

Machine learning (ML) systems automatically learn programs from data.
- *Pedro Domingos*

Machine learning is statistics minus any checking of models and assumptions
- *Brian D. Ripley*

What is the difference between these two fields?
The short answer is: None. They are both concerned with the same question: how do we learn from data?
- *Larry Wasserman*

THE ANALYTICS/ML PROCESS



MACHINE LEARNING

DATA DEFINITIONS: LABELED AND UNLABELED DATA

We collected the data from a known group of iris flower types.

The iris column is the target or labels.

Hence labeled data.

Classification or Prediction techniques is used in these cases.

Target (labels) **Variables or Features**

	iris	sepal length	sepal width	petal length	petal width
43	Iris-setosa	4.400	3.200	1.300	0.200
44	Iris-setosa	5.000	3.500	1.600	0.600
45	Iris-setosa	5.100	3.800	1.900	0.400
46	Iris-setosa	4.800	3.000	1.400	0.300
47	Iris-setosa	5.100	3.800	1.600	0.200
48	Iris-setosa	4.600	3.200	1.400	0.200
49	Iris-setosa	5.300	3.700	1.500	0.200
50	Iris-setosa	5.000	3.300	1.400	0.200
51	Iris-versicolor	7.000	3.200	4.700	1.400
52	Iris-versicolor	6.400	3.200	4.500	1.500
53	Iris-versicolor	6.900	3.100	4.900	1.500
54	Iris-versicolor	5.500	2.300	4.000	1.300
55	Iris-versicolor	6.500	2.800	4.600	1.500

No Target (labels) **Variables or Features**

We do not know what type of iris flowers they are.

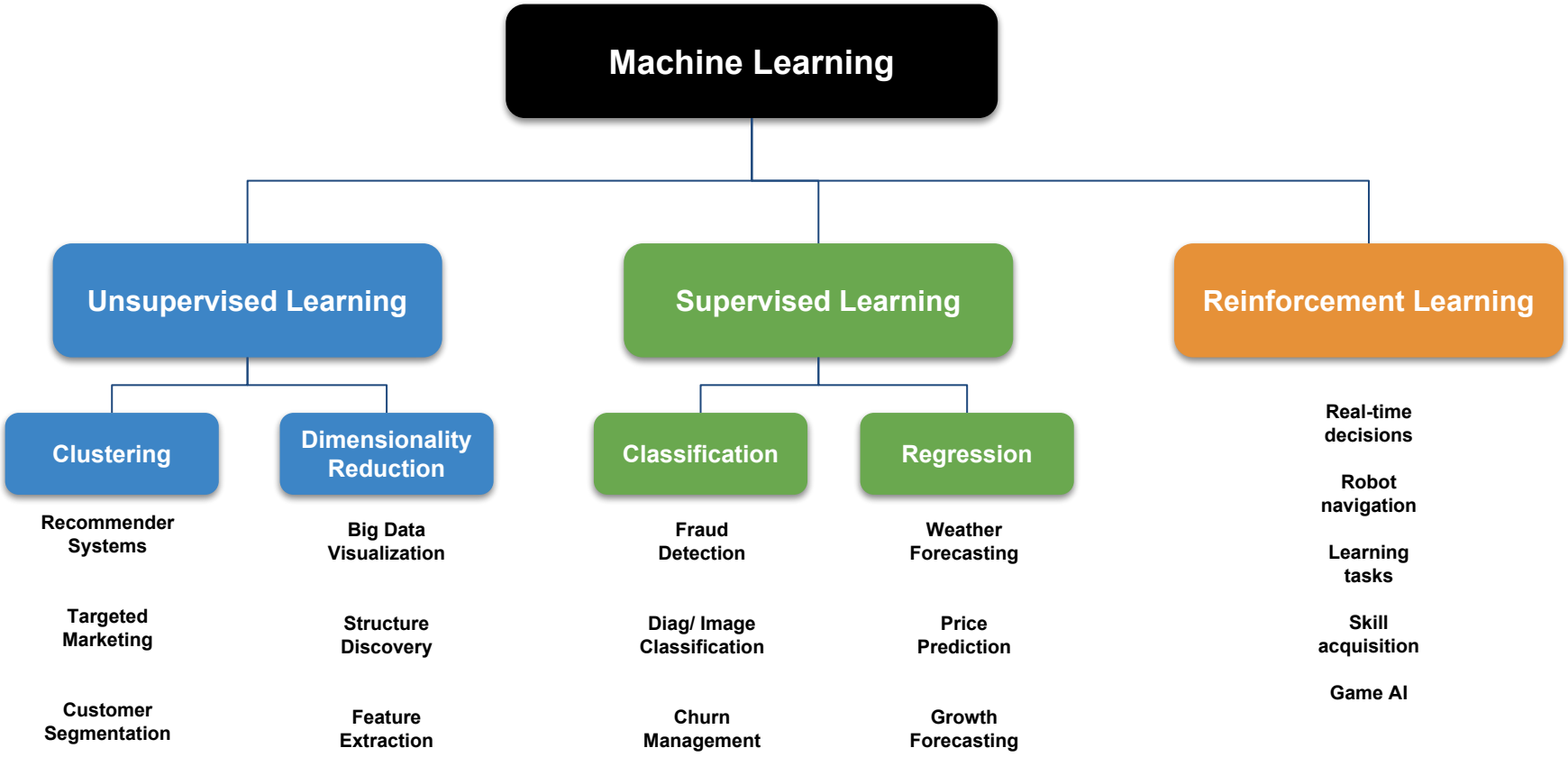
There is no labels.

Clustering techniques are typically used here.

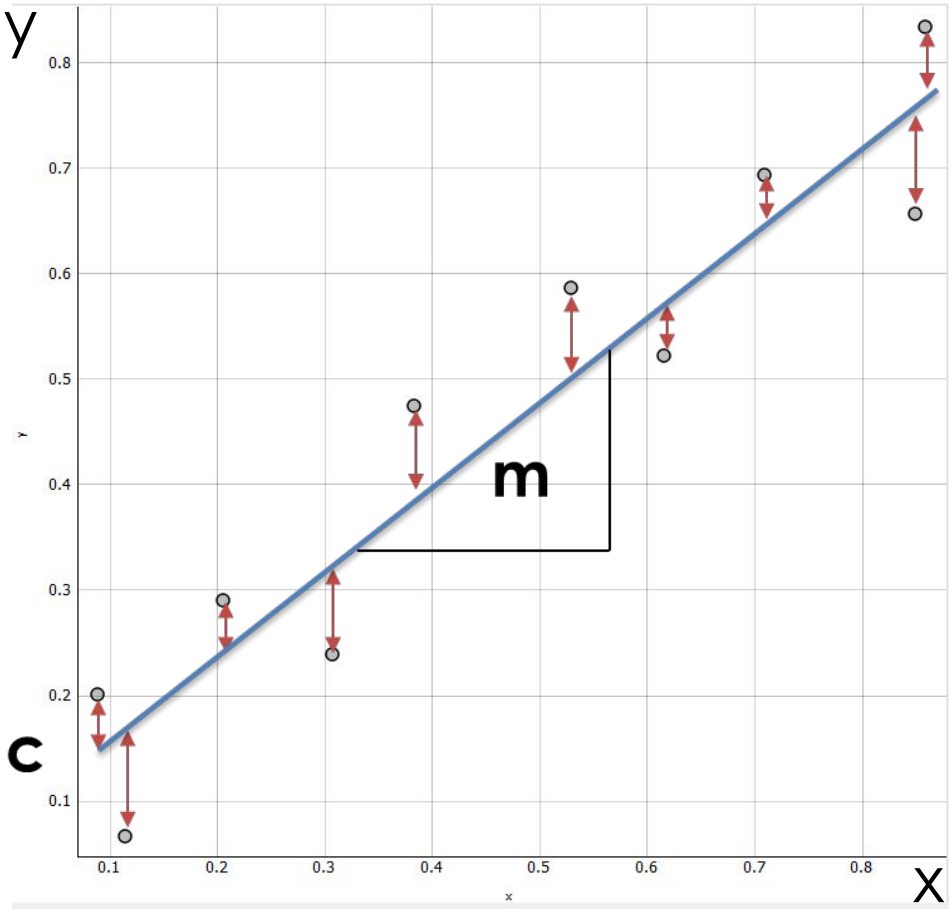
?

	sepal length	sepal width	petal length	petal width
45	5.100	3.800	1.900	0.400
46	4.800	3.000	1.400	0.300
47	5.100	3.800	1.600	0.200
48	4.600	3.200	1.400	0.200
49	5.300	3.700	1.500	0.200
50	5.000	3.300	1.400	0.200
51	7.000	3.200	4.700	1.400
52	6.400	3.200	4.500	1.500
53	6.900	3.100	4.900	1.500

MACHINE LEARNING OVERVIEW



WHAT IS MACHINE LEARNING?

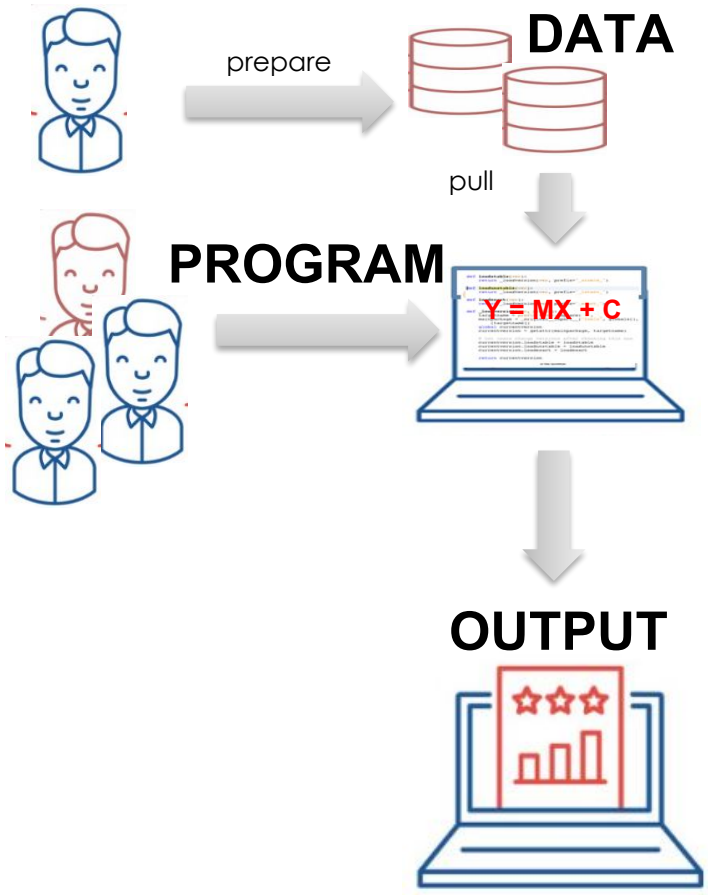


- Best-fit linear regression line in school (and Excel)
- **Find the minimum sum of … technically**
  - **Sum of squares of the error as small as possible**
  - **LEAST SQUARES method**

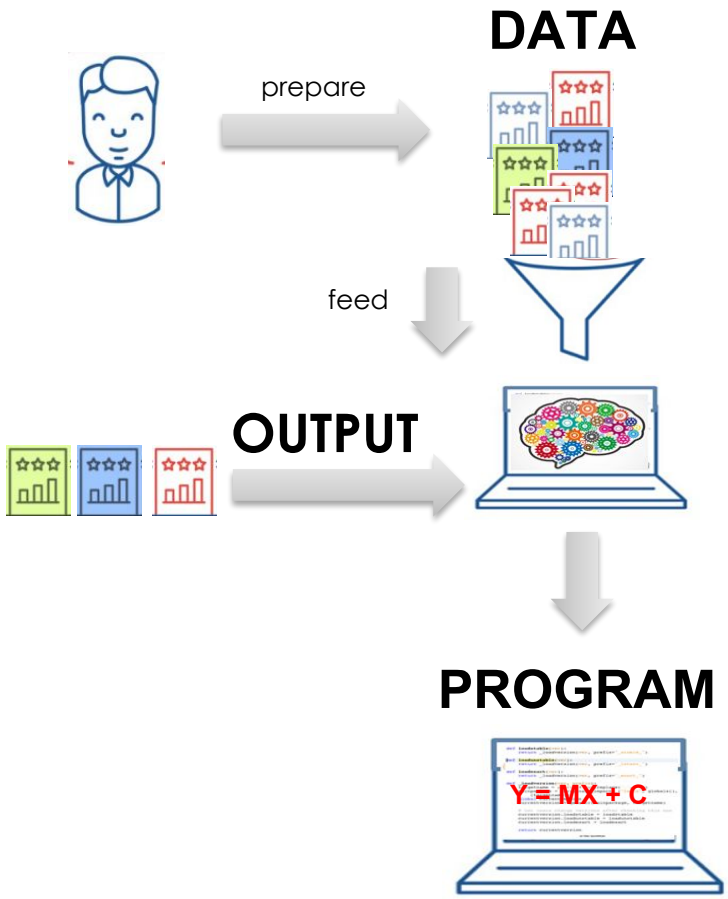
$$y = mx + c$$

PROGRAMMING VS MACHINE LEARNING

Traditional Programming



Machine Learning



LINEAR REGRESSION

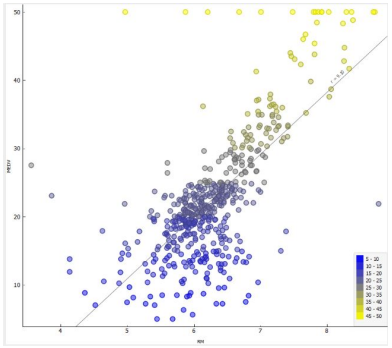
(a more detailed example)

	MEDV	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT
1	24.000	0.006	18.000	2.310	0.000	0.538	6.575	65.200	4.090	1.000	296.000	15.300	396.900	4.980
2	21.600	0.027	0.000	7.070	0.000	0.469	6.421	78.900	4.967	2.000	242.000	17.800	396.900	9.140
3	34.700	0.027	0.000	7.070	0.000	0.469	7.185	61.100	4.967	2.000	242.000	17.800	392.830	4.030
4	33.400	0.032	0.000	2.180	0.000	0.458	6.998	45.800	6.062	3.000	222.000	18.700	394.630	2.940
5	36.200	0.069	0.000	2.180	0.000	0.458	7.147	54.200	6.062	3.000	222.000	18.700	396.900	5.330
6	28.700	0.030	0.000	2.180	0.000	0.458	6.430	58.700	6.062	3.000	222.000	18.700	394.120	5.210
7	22.900	0.088	12.500	7.870	0.000	0.524	6.012	66.600	5.561	5.000	311.000	15.200	395.600	12.430
8	27.100	0.145	12.500	7.870	0.000	0.524	6.172	96.100	5.950	5.000	311.000	15.200	396.900	19.150

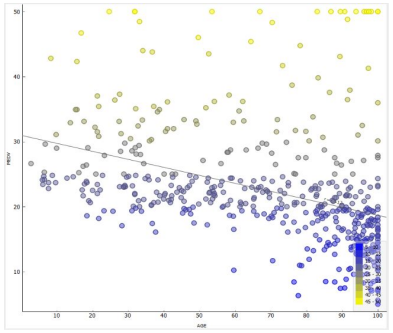
Target
Y

Variables, Features
X1, X2,... Xn

MEDV vs RM



MEDV vs AGE



	name	coef
1	intercept	36.4594884
2	CRIM	-0.1080114
3	ZN	0.0464205
4	INDUS	0.0205586
5	CHAS	2.6867338
6	NOX	-17.7666112
7	RM	3.8098652
8	AGE	0.0006922
9	DIS	-1.4755668
10	RAD	0.3060495
11	TAX	-0.0123346
12	PTRATIO	-0.9527472
13	B	0.0093117
14	LSTAT	-0.5247584

Y = MX + C

Y = M1X1 + M2X2 +... + C

MEDV = -0.108CRIM + 0.046ZN + 0.02INDUS + 2.68CHAS + ….3.8RM + 36.45

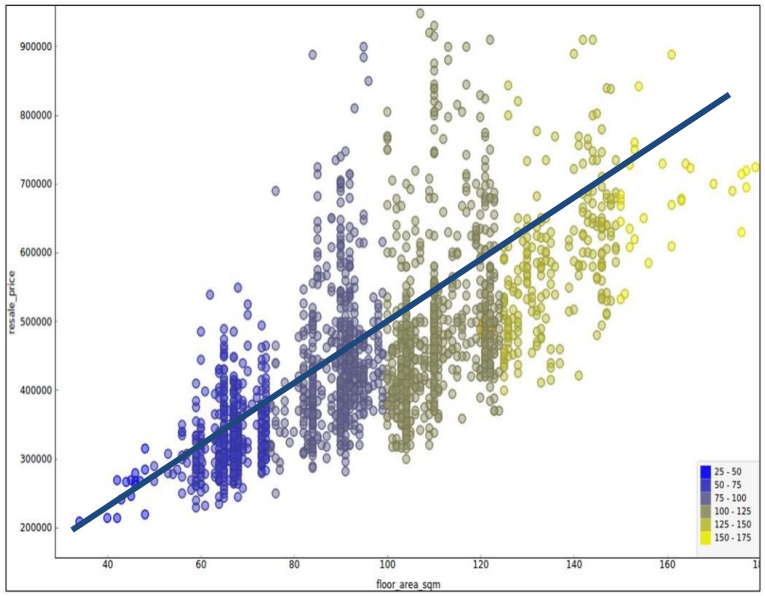
FIVE QUESTIONS THAT DATA SCIENCE ANSWERS

- Is this X or Y?
- Is this abnormal?
- How much does it cost in the future? Or how many will there be?
- How are they organized?
- What should I do next?

Each of these questions can be answered by an algorithm or recipe by using your data.

TYPES OF LEARNING (1/4)

- **Supervised – (Data with Labels)**
  - Make predictions based on the patterns in my data
  - **Regression Algorithms**
  - These are used to predict **one or more continuous** variables, such as height or weight, based on other attributes in the dataset.

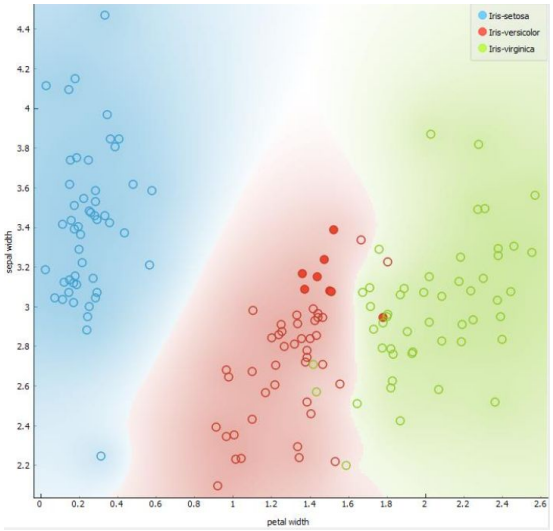


HOW MUCH WILL IT BE IN THE FUTURE?

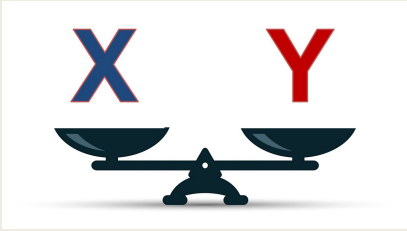
How much will my HDB flat be worth in the future?

TYPES OF LEARNING (2/4)

- **Supervised – (Data with Labels)**
  - Make predictions based on the patterns in my data
  - **Classification Algorithms**
  - These are used to **classify data into different categories**, such as whether an object is a fruit or vegetable, based on the other attributes in the dataset (calories, sugar etc)



Is this X or Y?



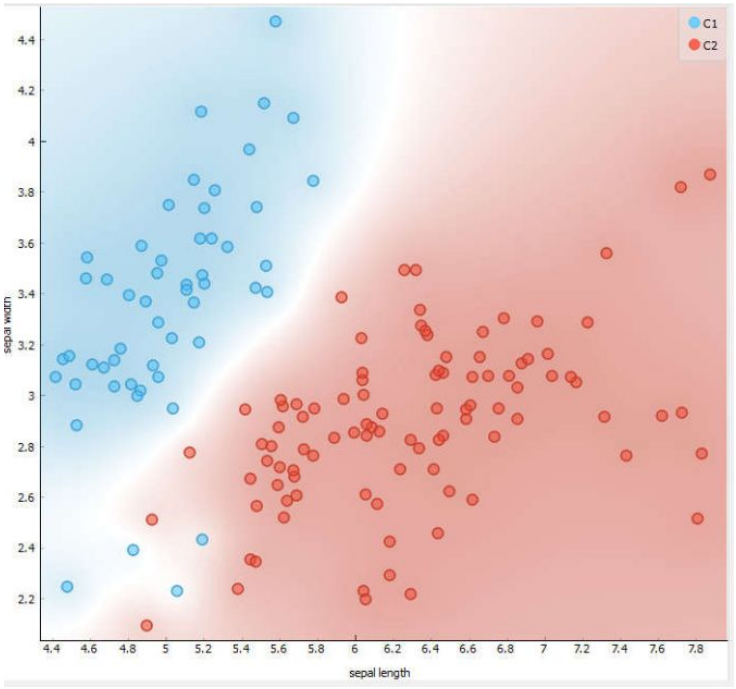
Is this a CAT or a DOG?

Is this abnormal?

Is this a fraudulent transaction?

TYPES OF LEARNING (3/4)

- **Unsupervised – (Data with no labels)**
  - Show me the patterns in my data
  - **Clustering** Algorithms which finds **natural groupings** and patterns in datasets.



How are they organized?

TECH | 2/16/2012 @ 11:02AM | 2,492,601 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

335 comments, 174 called-out + Comment Now + Follow Comments

Every time you go shopping, you share intimate details about your consumption patterns with retailers. And many of those retailers are studying those details to figure out what you like, what you need, and which coupons are most likely to make you happy. Target, for example, has figured out how to data-mine its way into your womb, to figure out whether you have a baby on the way long before you need to start buying diapers.

TARGET

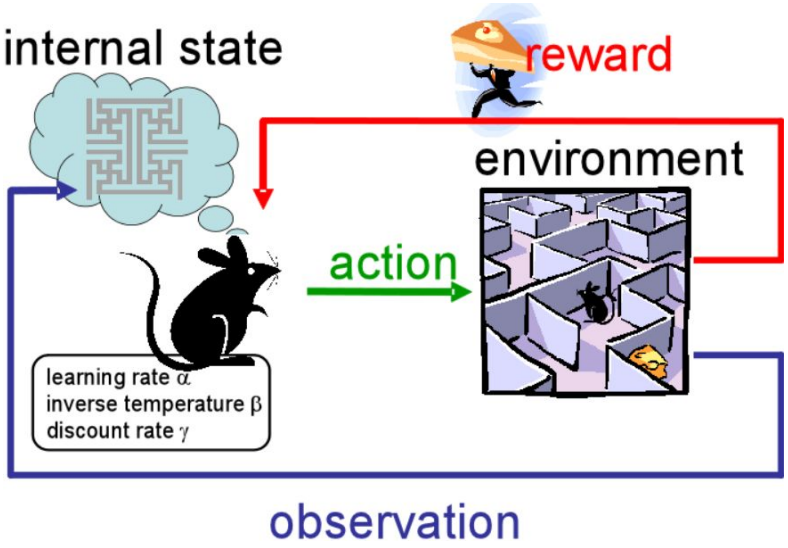
Target has got you in its aim

51.1k Share 16.5k Tweet 6.1k Share 2.3k reddit

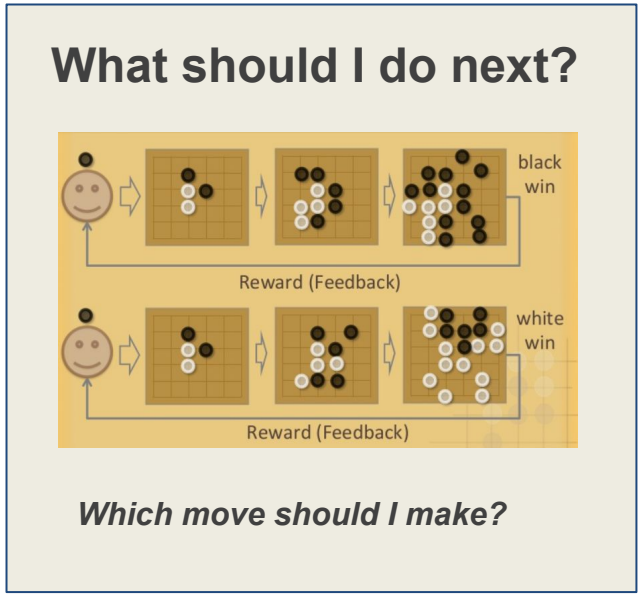
Group my customers in 5 groups!

TYPES OF LEARNING (4/4)

- **Reinforcement Learning**
  - Agent + Environment
  - Reward actions based on policies.

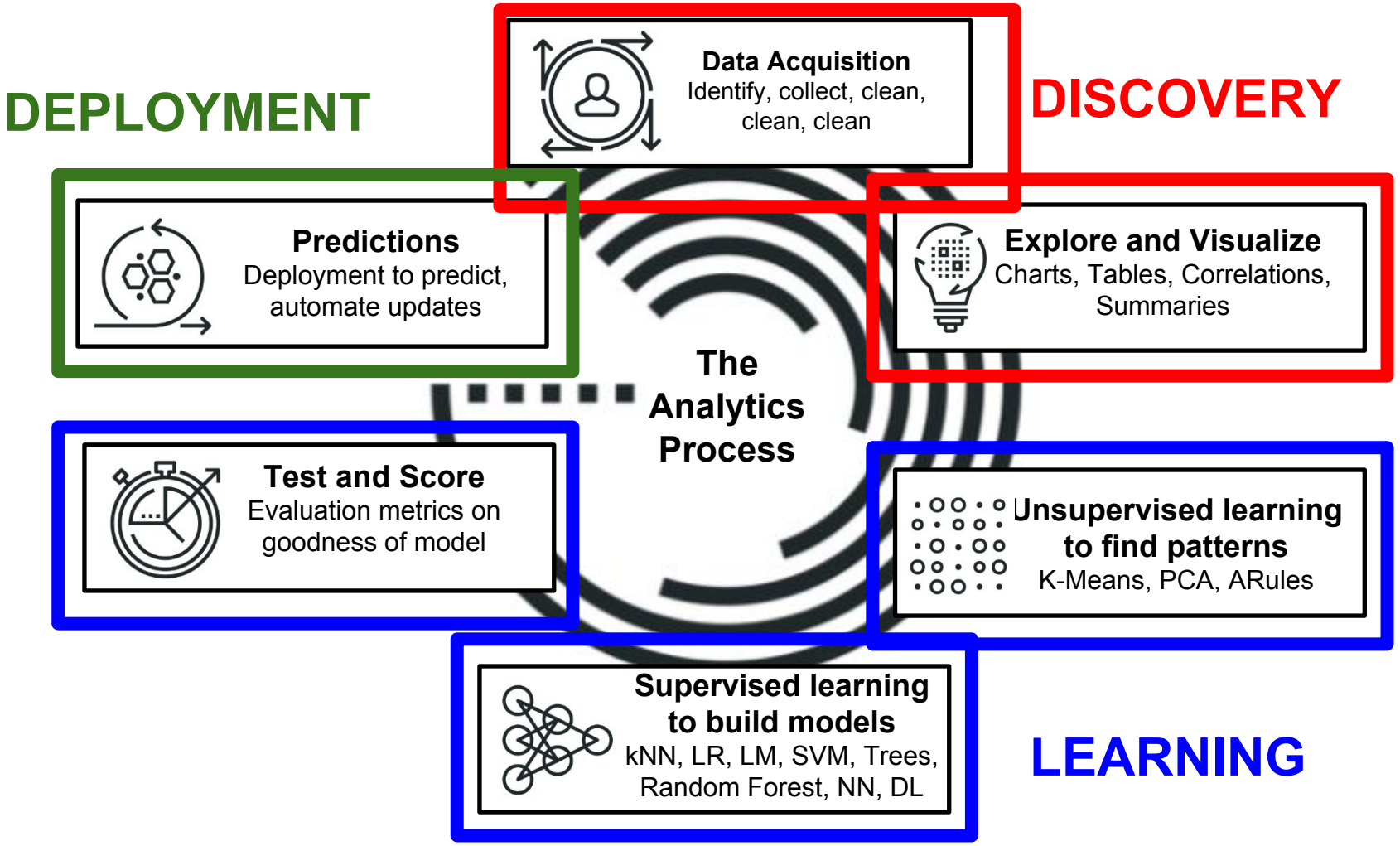


Source: UT Computer Science



Source:
https://www.slideshare.net/ckmarkohchang/alphago-in-dept

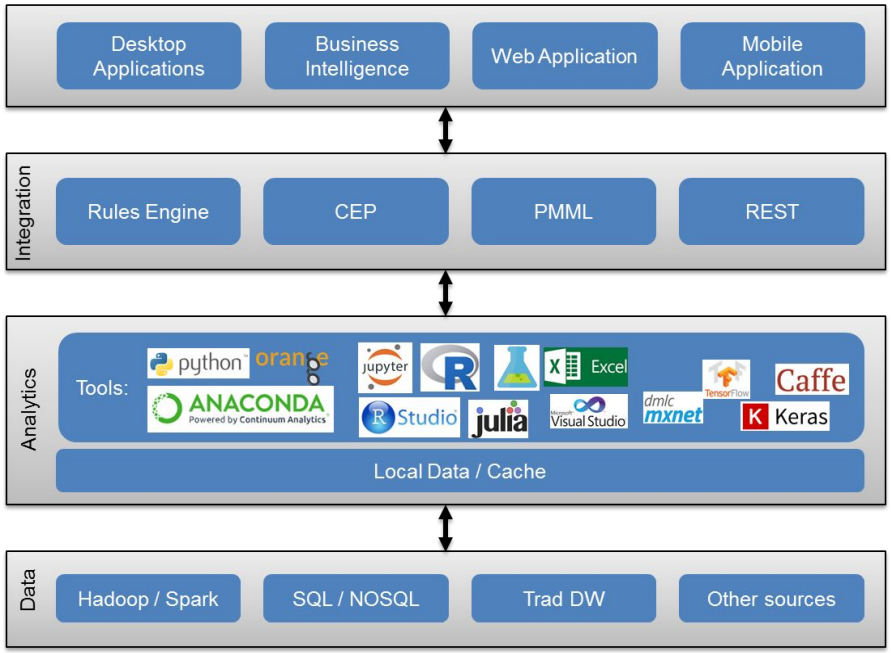
SUMMARY



Overview

THE DATA SCIENCE STACK

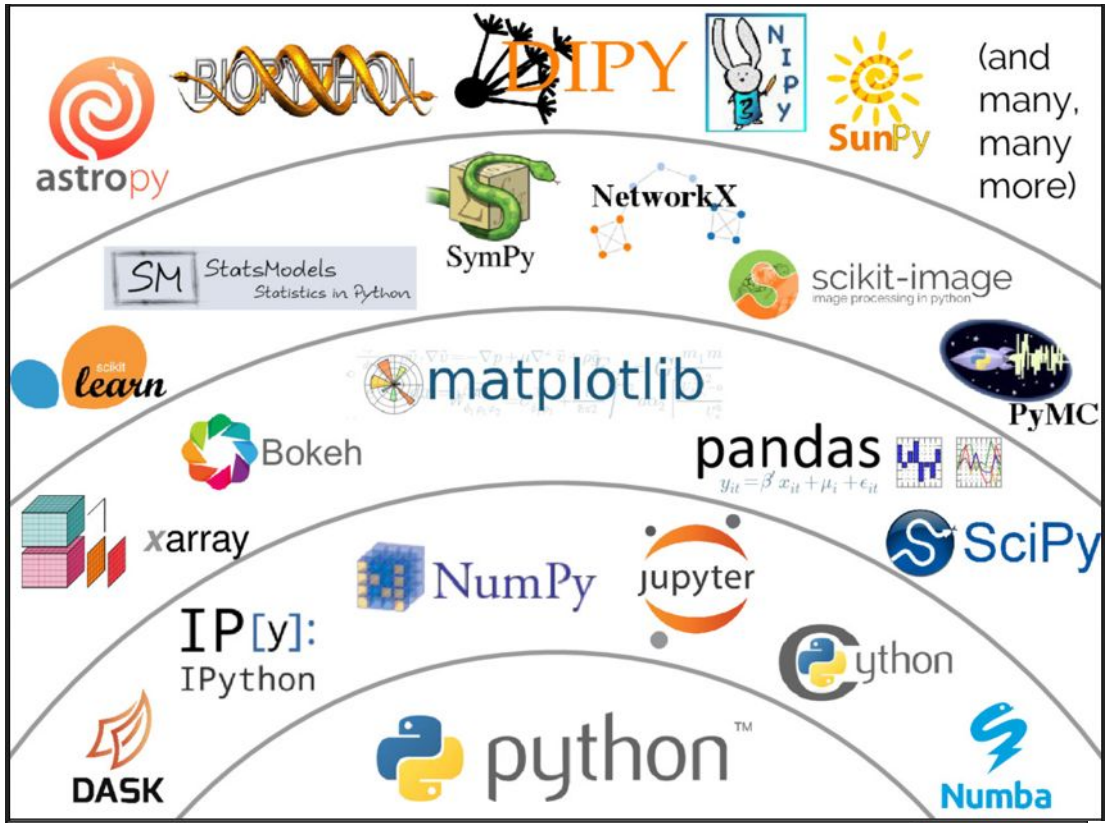
The Data Science Stack



What is R?

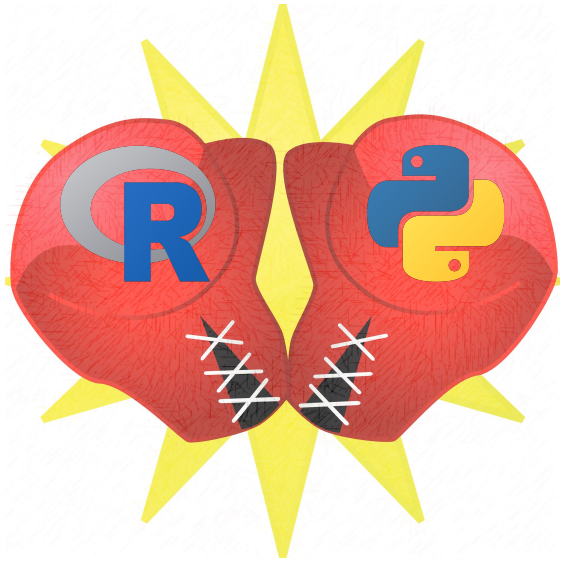
http://www.youtube.com/watch?v=TR2bHSJ_eck

PYTHON'S SCIENTIFIC AND ANALYTICS STACK

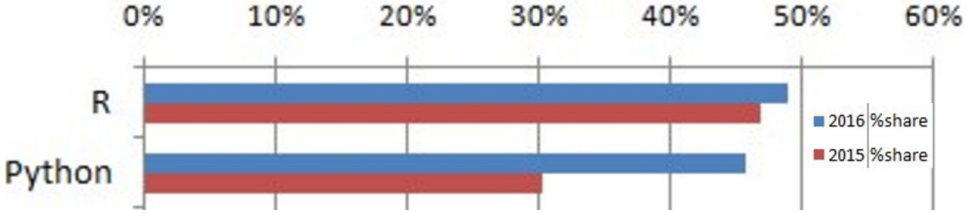


Source: Jake VanderPlas
The unexpected effectiveness of python in science

R VS PYTHON



KDnuggets Analytics/Data Science 2016 Software Poll, top 10 tools



2016 vs 2015

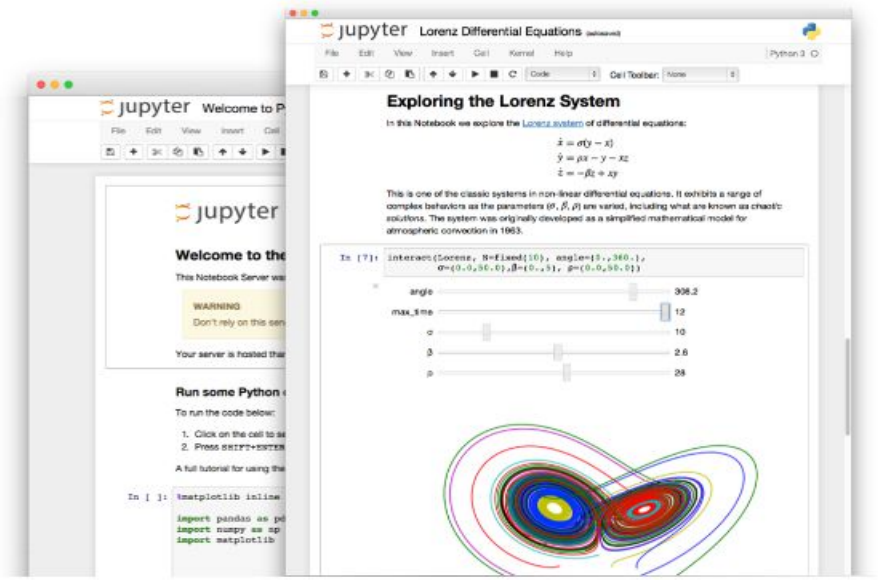
Tool	2016 % share	% change	% alone
R	49%	+4.5%	1.4%
Python	45.8%	+51%	0.1%
SQL	35.5%	+15%	0%
Excel	33.6%	+47%	0.2%
RapidMiner	32.6%	+3.5%	11.7%
Hadoop	22.1%	+20%	0%
Spark	21.6%	+91%	0.2%
Tableau	18.5%	+49%	0.2%
KNIME	18.0%	-10%	4.4%
scikit-learn	17.2%	+107%	0%

Source: KDnuggets™

JUPYTER
NOTEBOOKS

NUS National University of Singapore

jupyter Install About Resources Documentation NBViewer Widgets Blog Donate



Multi-user / Enterprise

jupyterhub

A multi-user version of the notebook designed for companies, classrooms and research labs

</>

Language of choice

The Notebook has support for over 40 programming languages, including those popular in Data Science such as Python, R, Julia and Scala.

Share notebooks

Notebooks can be shared with others using email, Dropbox, GitHub and the Jupyter Notebook Viewer.

Interactive widgets

Code can produce rich output such as images, videos, LaTeX, and JavaScript. Interactive widgets can be used to manipulate and visualize data in realtime.

Spark

Big data integration

Leverage big data tools, such as Apache Spark, from Python, R and Scala. Explore that same data with pandas, scikit-learn, ggplot2, dplyr, etc.

END TO END BIG DATA, ANALYTICS AND AI STACK

NUS National University of Singapore

- Easy of use
- Multiple data sources
- Distributed Algorithms
- Industry Templates Library
- Intuitive graphical tool
- Python/Java/Scala Extension
- GPU Cluster Accelerated
- Enterprise Security

Midas
Interactive Development Environment

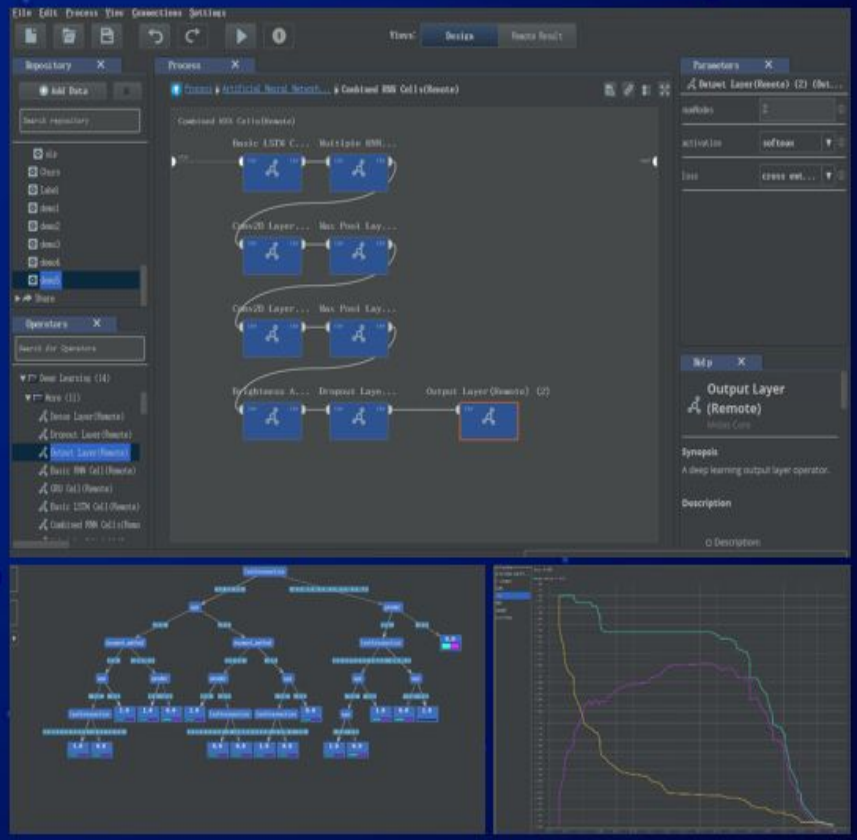
Feature Engineering | Industrial Templates | Intelligent Modeling

Deep Learning Algorithms Library

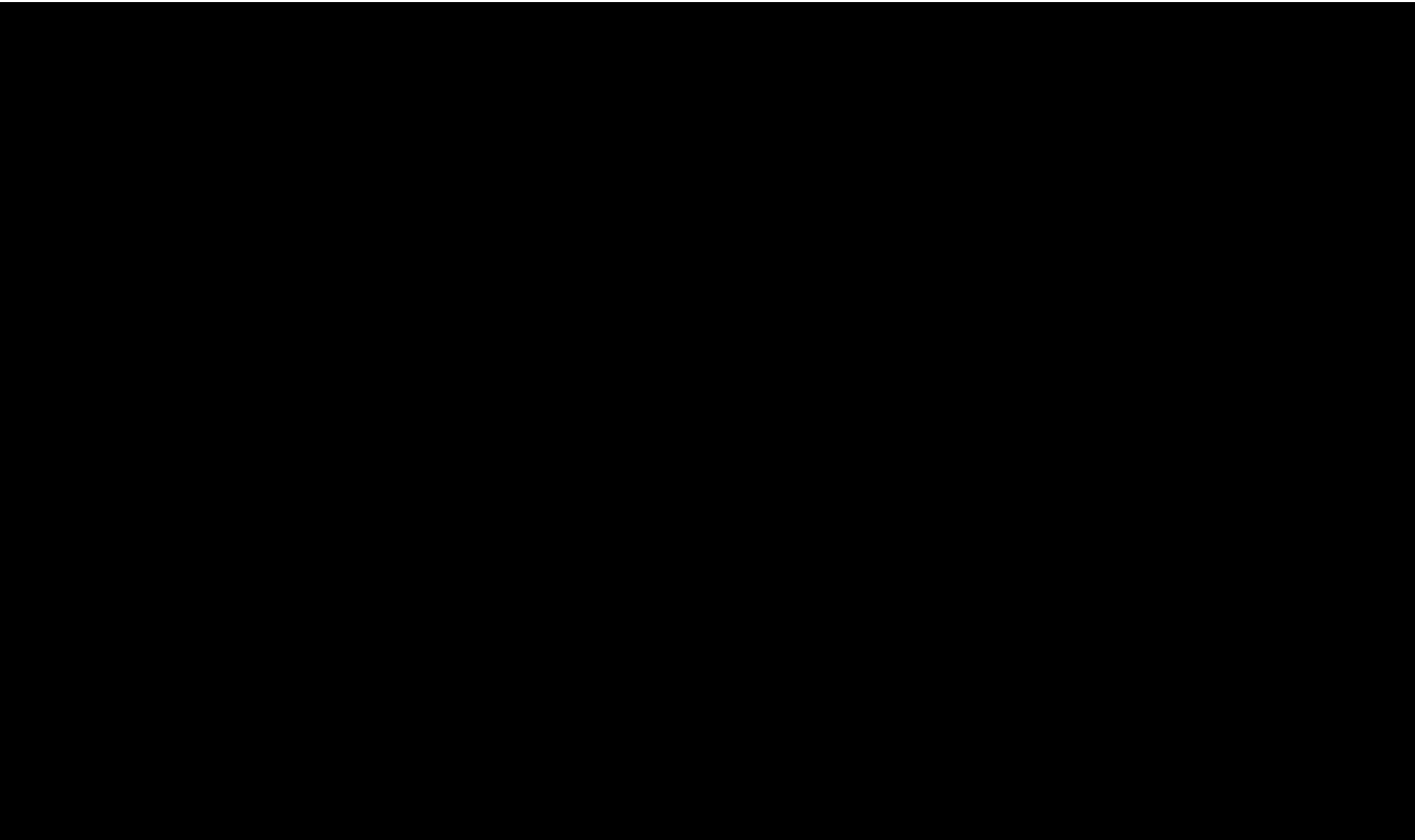
Transwarp Hubble Bridge

Spark | TensorFlow | mxnet | Graphene

Transwarp Sophon



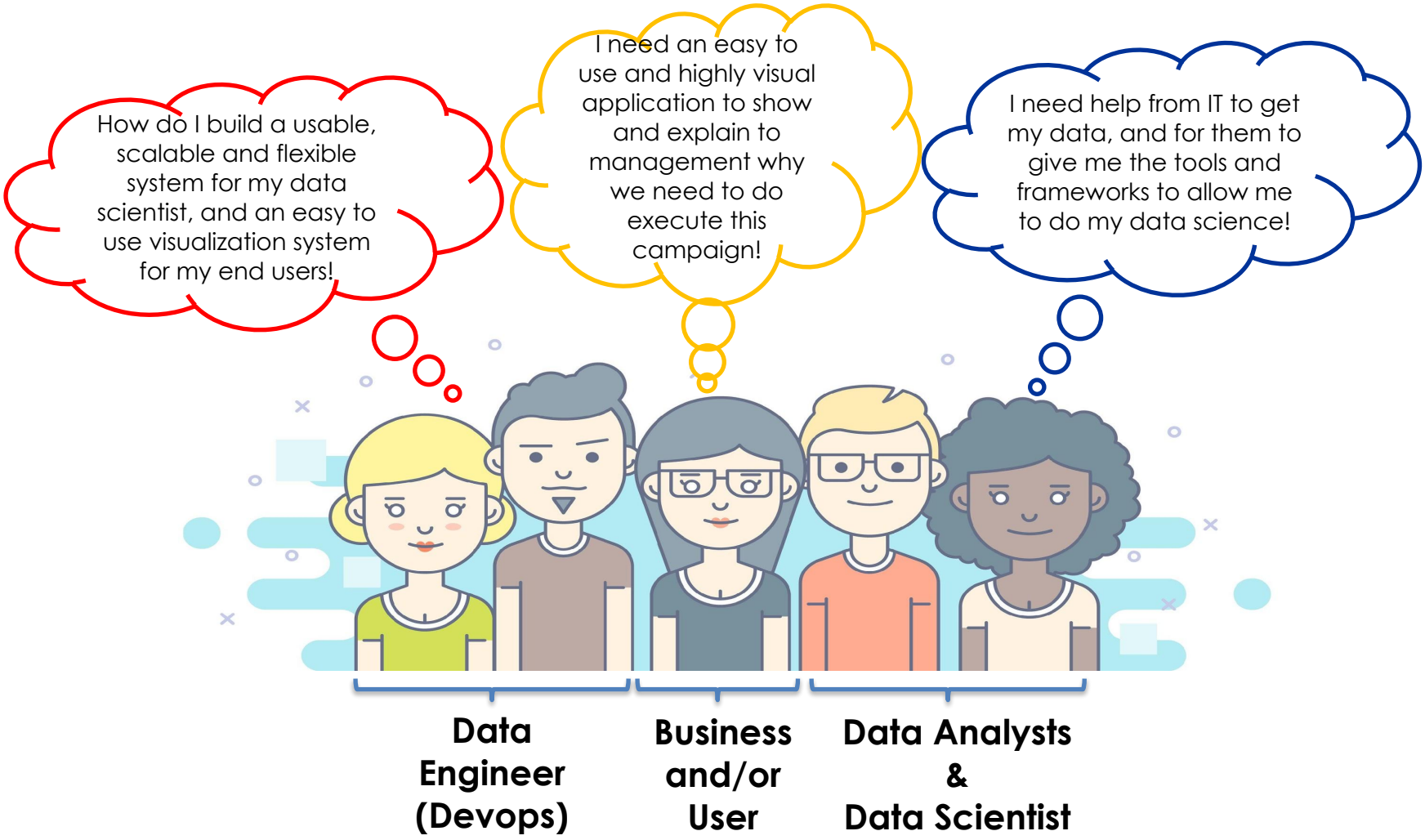
Demo



Build or outsource? How do you go about building your team?

THE DATA SCIENCE TEAM

THE DATA SCIENCE TEAM



PERSONAS

Debra (Data Scientist)
Data Prep & Modeling

Both:
Operationalize
Put model in production and dashboard.

Data Scientist Workstation

R IDE

Microsoft R Open

R Server

Train Model
Predict LOS

SQL Server 2016

Advanced Analytics Extensions

Microsoft R Open

R Server

Visualize Results

PowerBI Dashboard

Caroline (LOB):
Visualize, makes decision and recommendations

Danny (Devops):
DB server setup

THE DATA SCIENTIST

MYTHICAL CREATURE OR REAL?

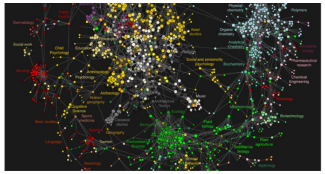
Data Scientists: The New Sex Symbol

April 15, 2013 by john Leave a Comment

Last October an article in the *Harvard Business Review* appeared with the headline, "Data Scientist: The Sexiest Job of the 21st Century."

Last week Claire Cain Miller, a technology reporter for The New York Times, used the head to introduce her own story on the fast growing field of data science and the efforts on the part of the academic community to create a new crop of data scientists – whom she calls "…the magicians of the Big Data era."

https://www.linkedin.com/pulse/youre-data-scientist-chuck-russell



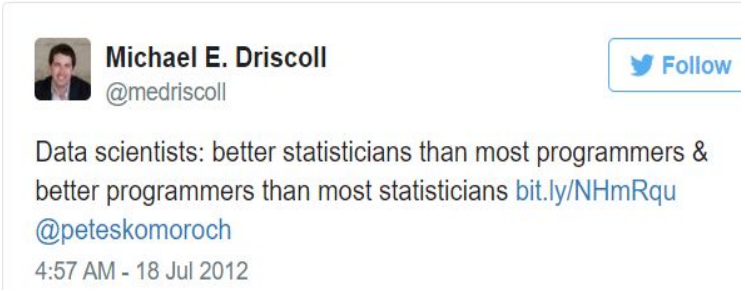
You're not a Data Scientist

Published on April 28, 2015

Chuck Russell Follow
Founder - Collective Intelligence Inc.

300 54 97

The 8-week course you took on Coursera or the Data Science boot camp you attended no more makes you a data scientist than my recent golf lessons make me a golf pro. I believe in lifelong learning and I'm all for self-improvement but this is self-delusion.

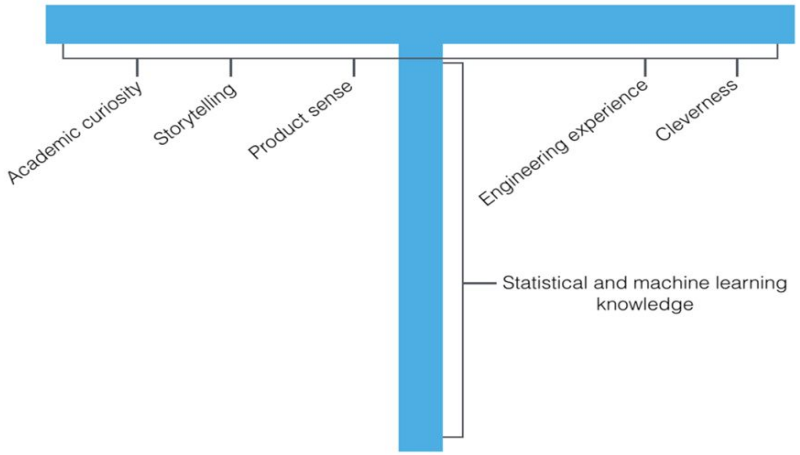


Michael E. Driscoll
@medriscoll

Follow

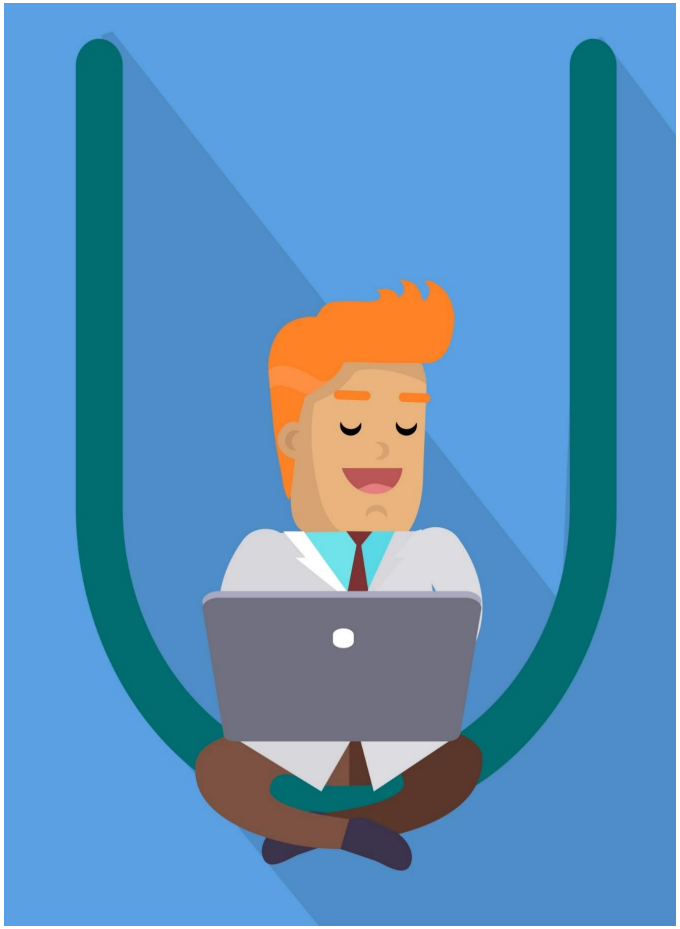
Data scientists: better statisticians than most programmers & better programmers than most statisticians bit.ly/NHmRqu @peteskomoroch

4:57 AM - 18 Jul 2012



https://www.import.io/post/data-scientists-vs-data-analysts-why-the-distinction-matters/

THE MODERN DATA SCIENTIST



Math & Stats
Unsupervised Statistical/Machine learning: clustering, PCAs etc
Supervised Statistical/Machine learning: regressions, SVM, trees, random forest etc
Bayesian learning
Deep learning: CNN, RNN/LSTM, GANs, RL
Search and Optimization: SGD, ADAM etc
Design of experiments

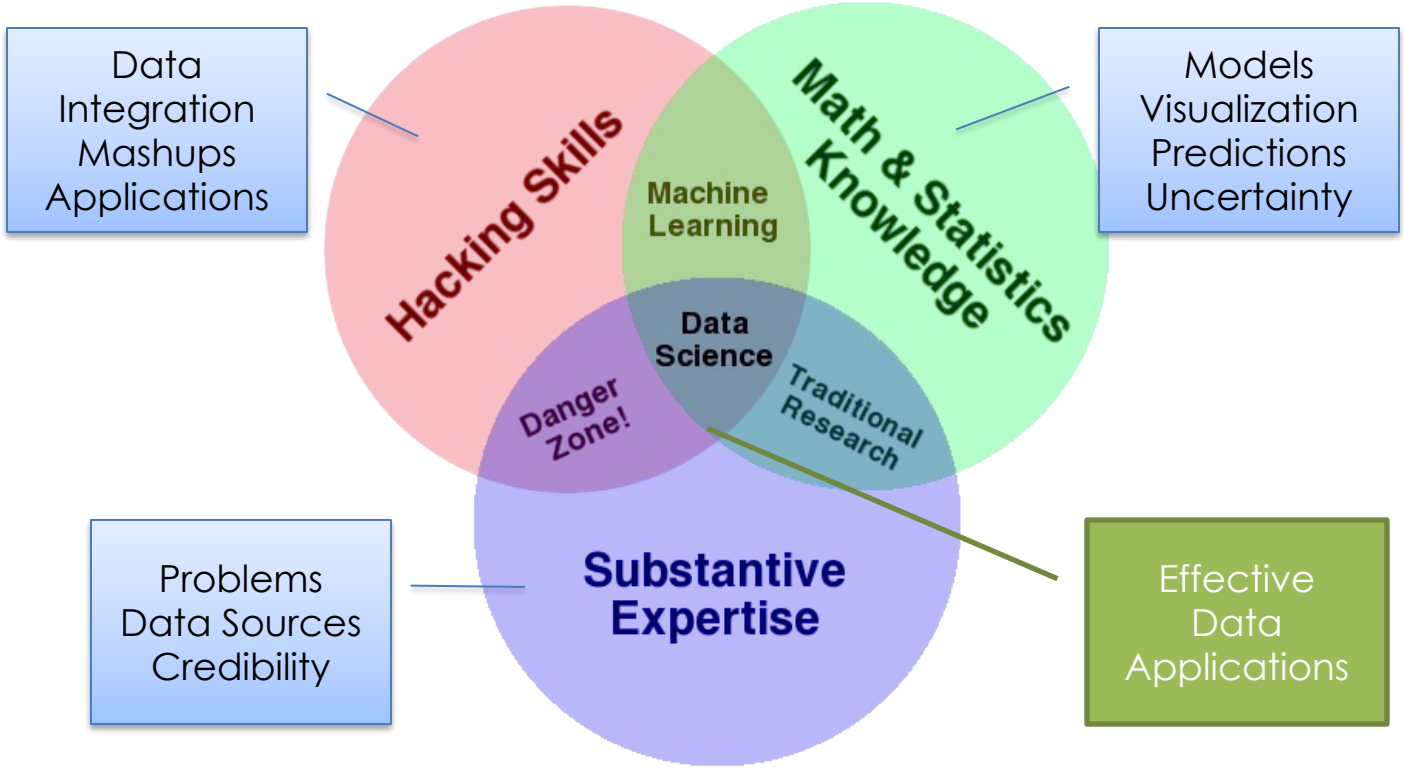
Programming & Databases
Computer science fundamentals
Scripting: Python, Bash
Statistical: R, Python/Scikit-learn
Distributed and parallel computing frameworks such as
Hadoop/Pig/Hive, Spark
Databases: SQL, NoSQL, Parallel databases
Cloud: AWS, Azure, Google Compute

Communication & Story-telling
Be able to interact-well with working-level to distill domain knowledge
Be able to engage with management to sell your story and analysis
Ability to translate data-driven insights into recommendations and actions
Good sense of visual design
R: ggplot, lattice
Python: MatplotLib, Bokeh, Seaborne
Others: D3.js, Tableau

Domain knowledge & soft-skills
Passionate about data
Hacker mindset
Curious
Problem solver
Think Strategic, work local
Collaborative

THE DATA SCIENTIST

REALITY – THE DATA SCIENCE TEAM



Drew Conway
http://www.dataists.com/2010/09/the-data-science-venn-diagram/

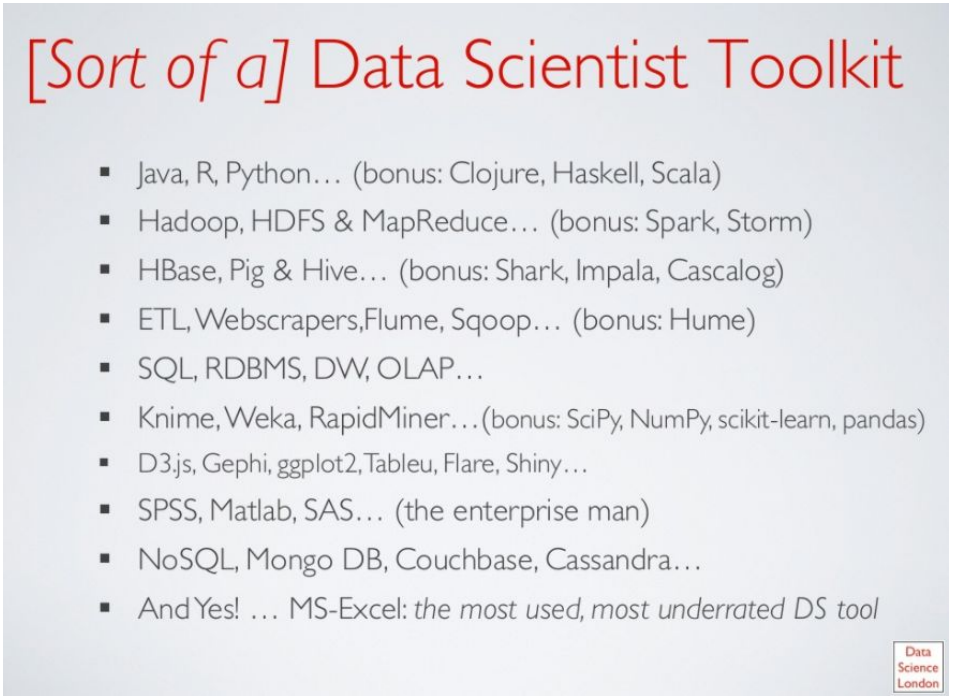
ROLE: DATA ANALYST

- Data Analyst
  - Essentially a junior data scientist
  - Don’t have the mathematical or research background to invent new algorithms
  - But they have a strong understanding of how to use existing tools to solve problems
  - Need to have a baseline understanding of five core competencies: programming, statistics, machine learning, data munging, and data visualization.

Source: http://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html

ROLE: DATA SCIENTIST

- Data Scientist
  - Typically have advanced degrees in computer science, physics, statistics, or applied mathematics,
  - Have knowledge to invent new algorithms to solve data problems
  - Essentially leverage data to solve business problems.
  - They interpret, extrapolate from, and prescribe from data to deliver actionable recommendations.



A data analyst summarizes the past; a data scientist strategizes for the future.

Source: http://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html

Reference only

ROLE: DATA ENGINEER

- Data Engineer
  - Builds a robust, fault-tolerant data pipeline that cleans, transforms, and aggregates unorganized and messy data into data sources for a data analyst or data scientist to easily retrieve the needed data for their evaluations and experiments.
  - Skills: Hadoop (MapReduce, Hive, Pig), Spark, SQL(Oracle, MySQL, Postgresql), NoSQL (MongoDB, Cassandra, Neo4j), data warehousing arhcitectures

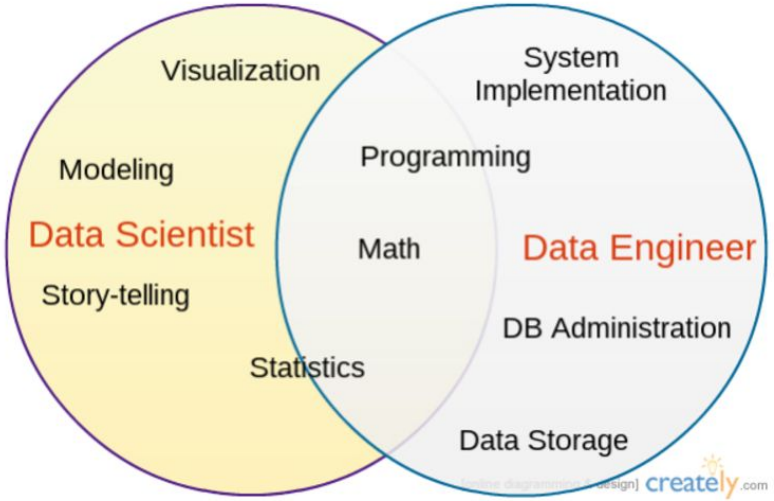


Image via Data Science 101

Source: http://blog.udacity.com/2014/12/data-analyst-vs-data-scientist-vs-data-engineer.html

DATA SCIENTIST – HOW TO HIRE

- From the R or Python user group ! There is nearly always one in the country!
- From the university / research labs
  - Masters or PhDs in statistics
  - But will require training in big data infrastructure
  - And throw them into the deep end of the pool in business domain understanding (we should expect these people able to learn FAST!)
  - Caution: Maybe too academic focused, probably have not handled big data before
  - Hire the smartest you can find
- Think like a soccer manager
  - Stalk the competitors, playing field (Kaggle)
  - Hire the team, with each member bring in domain expertise in different fields
  - But you should expect them to easily cross train/learn from different domains

NOT NECESSARILY COMPUTER SCIENCE!

- People often assume that data scientists need a background in computer science. In my experience, that hasn't been the case: my best data scientists have come from very different backgrounds.
- The inventor of LinkedIn's People You May Know was an experimental physicist.
- An oceanographer made major impacts on the way we identify fraud.
- Perhaps most surprising was the neurosurgeon who turned out to be a wizard at identifying rich underlying trends in the data.

-- DJ Patil, former US Chief Data Scientist, Linked and coined the term data scientist

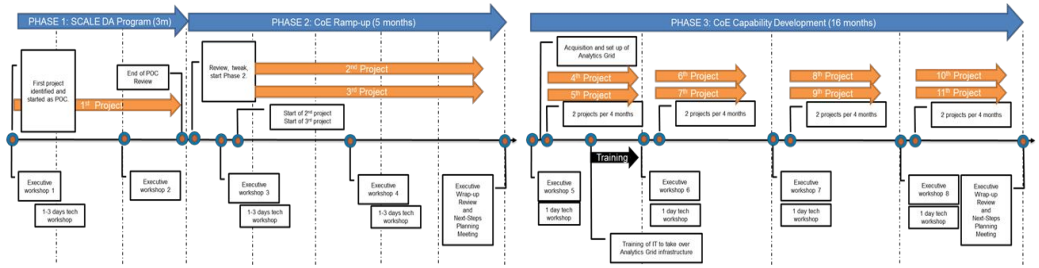
FORMING THE TEAM

YOUR TEAM NEEDS TOOLS!

- Invest in tools to allow the team to experiment faster
- Open source and/or commercial
- Automate, automate, automate!
- Scale of experiments:
  - 1 day to run -> I can only ask questions which I think I know the answer (the query is to support my hypotheses)
  - 1 min to run -> I can ask any questions!

SUMMARY

- Always build your own team **(IF YOU CAN and MAKE SENSE TO DO SO)**
- Risk of depending on external consultants: your organization's data/knowledge stays/becomes external and not retained within
- One strategy is to develop a wide base of "data literate" team members, with a few "power users" who can themselves connect to other experts in other departments



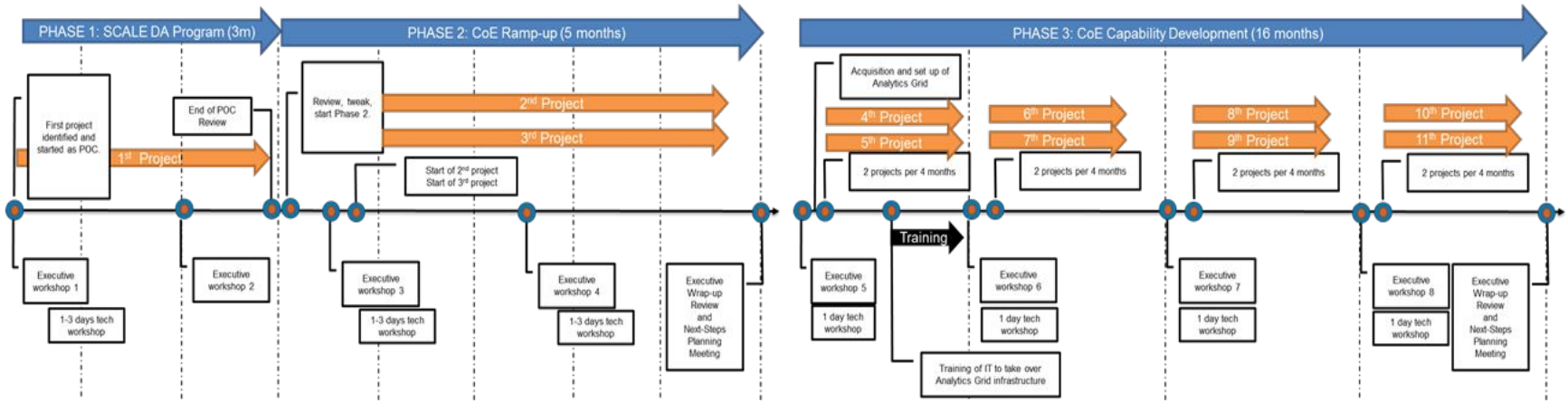
Build or outsource? How do you go about building your team?

CENTRE OF EXCELLENCE FOR ANALYTICS

COE FOR ANALYTICS

- **3 Phase (24-months) ramp-up**
  - **Phase 1**: 3 months, small initial POC
  - **Phase 2**: 5 months, 2 bigger projects, setup of data science/analytics/AI Stack
  - **Phase 3**: 16 months, 6 projects, full DA Stack up and running
- Ideas and best practices from experience in building Data Science and DevOps team
  - Revolution APAC office
  - Customers

Revolution Analytics 2012-2016



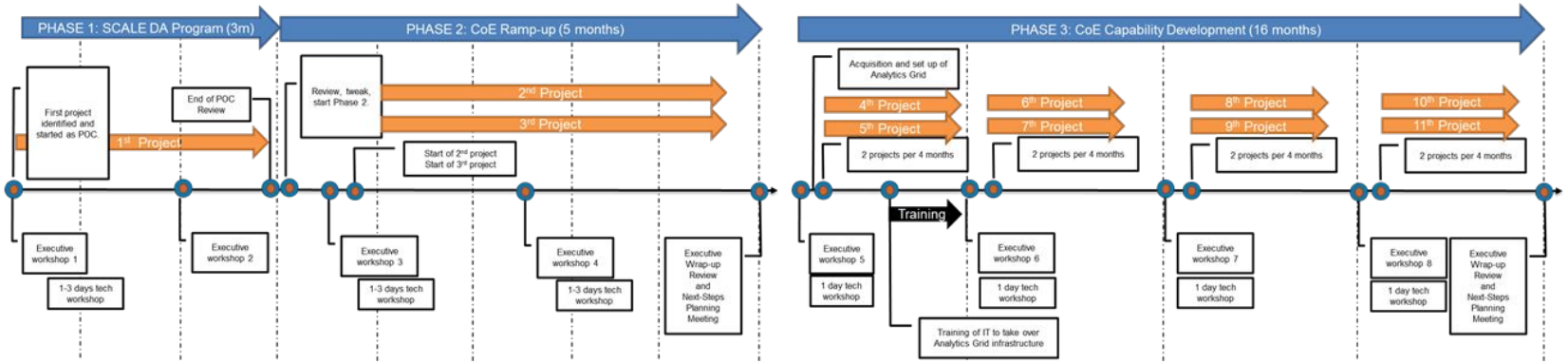
BEYOND THE DATA SCIENCE TEAM

- **Structure**
  - Data science teams needs to work closely with business, learn their domain, speak their language
    - Else disjointed, far away, not trusted by the business users
  - However, if too deeply involved, biases sets in
    - Predictions may “make the boss look good”
- **Data-driven culture**
  - Too much of a good thing
    - Trust the data and model too much and upset the sales (“gut feel”) folks
  - Too much gut-instinct
    - When data shows otherwise – ignore or push back
- **Avoid complexity (KISS)**
  - Too complicated (long development time, expensive, difficult to change) models not necessary better
  - Models must be deployable! Eg Netflix $1M prize 1st runner-up was deployed and not the 1st prize winner
  - Too simple also an issue if the model cannot capture the nuances of the underlying data.

BEYOND THE DATA SCIENCE TEAM

- Core Analytics Team
- Have a Chief Data / Analytics Officer at the C-Suite
- Start small
  - Short 3- 6 months MVP
  - Use this as the champion to gain mindshare and support from stakeholders

24-MONTHS COE PLAN

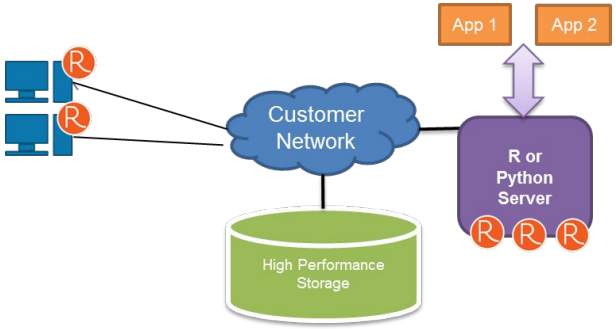


PHASE 1

PEOPLE AND SETUP

- PARTNER:
  - 1 x Lead Data Scientist (20%)
  - 1 x Senior data scientist (100%)
  - 1 x Big Data developer (50%)

- Customer
  - 1 x COA project
  - 2 data scientist to be trained in R or Python
  - 1 x IT to be trained on R or Python enterprise integration

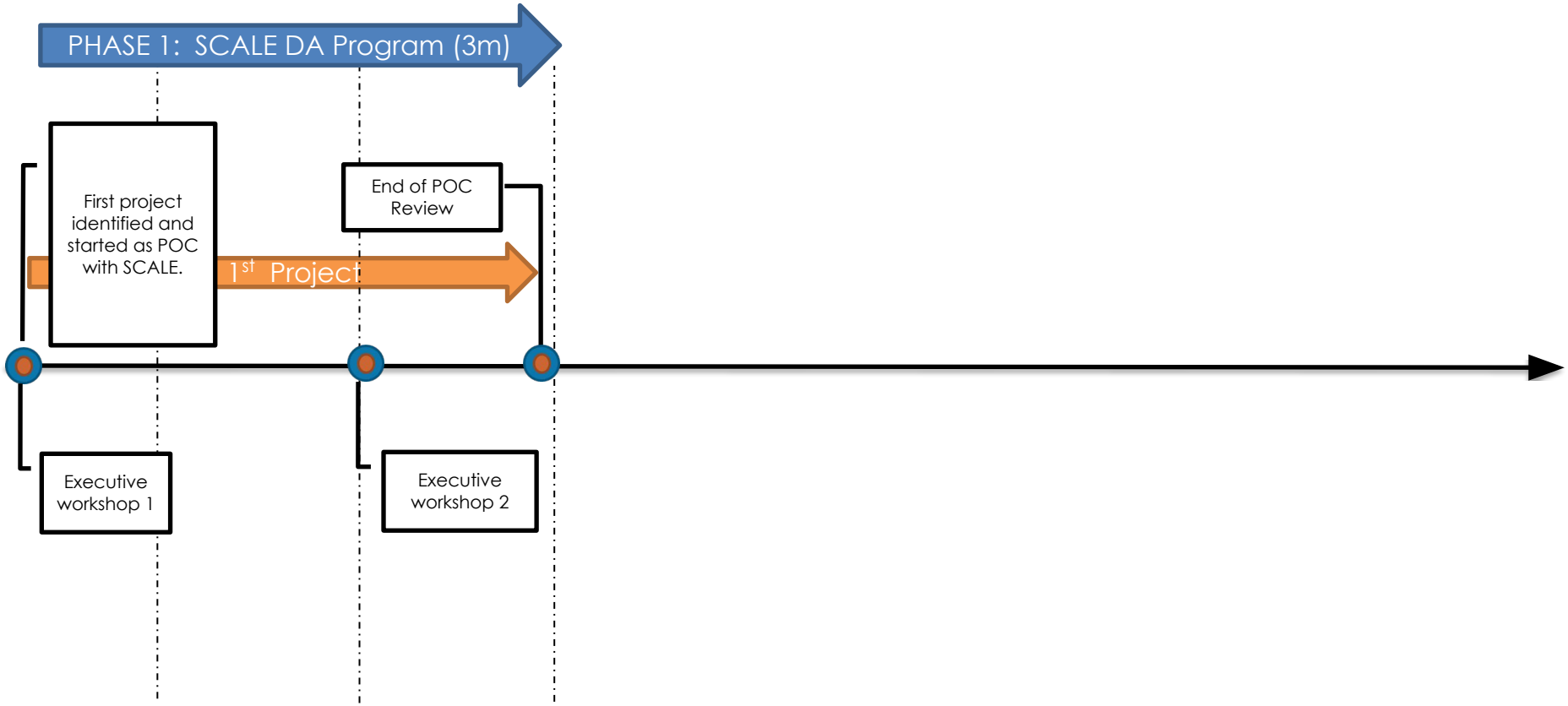


Usage model:

1. User build models on R/Python workstations
2. User run bigger models on R/Python/Jupyter Server
   - Each user to limit himself/herself to 4-cores max per run so as not to starve others of CPU processing
   - So max 8-concurrent jobs of “4-cores 64GB ram”
3. IT will build end-user facing applications using the models built
4. WEB or MOBILE applications as frontend, executing R/Python models or making use of analytics output from R/Python models built.

PHASE 1

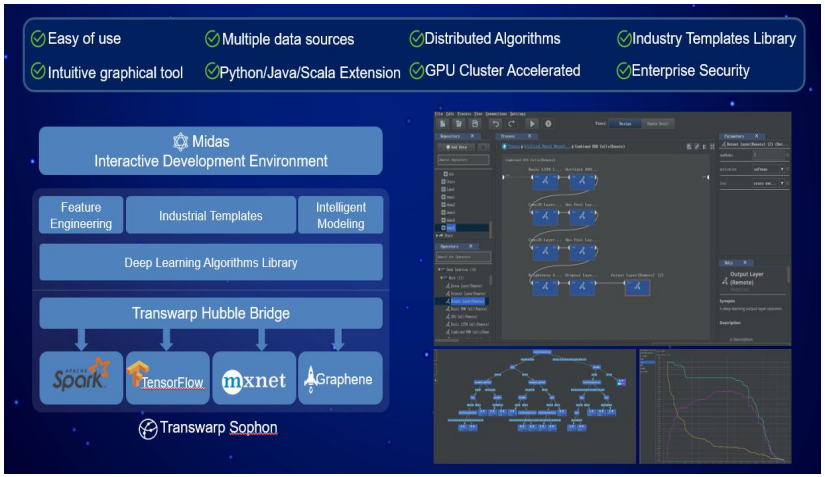
3 MONTHS – KEEP IT SIMPLE!



PHASE 2

PHASE 2: 5 MONTHS COE RAMP-UP

- PARTNER:
  - 1 x Big Data Analytics Architect (20%)
  - 1 x Lead Data Scientist (20%)
  - 1 x Senior data scientist (100%)
  - 1 x Big Data developer (50%)



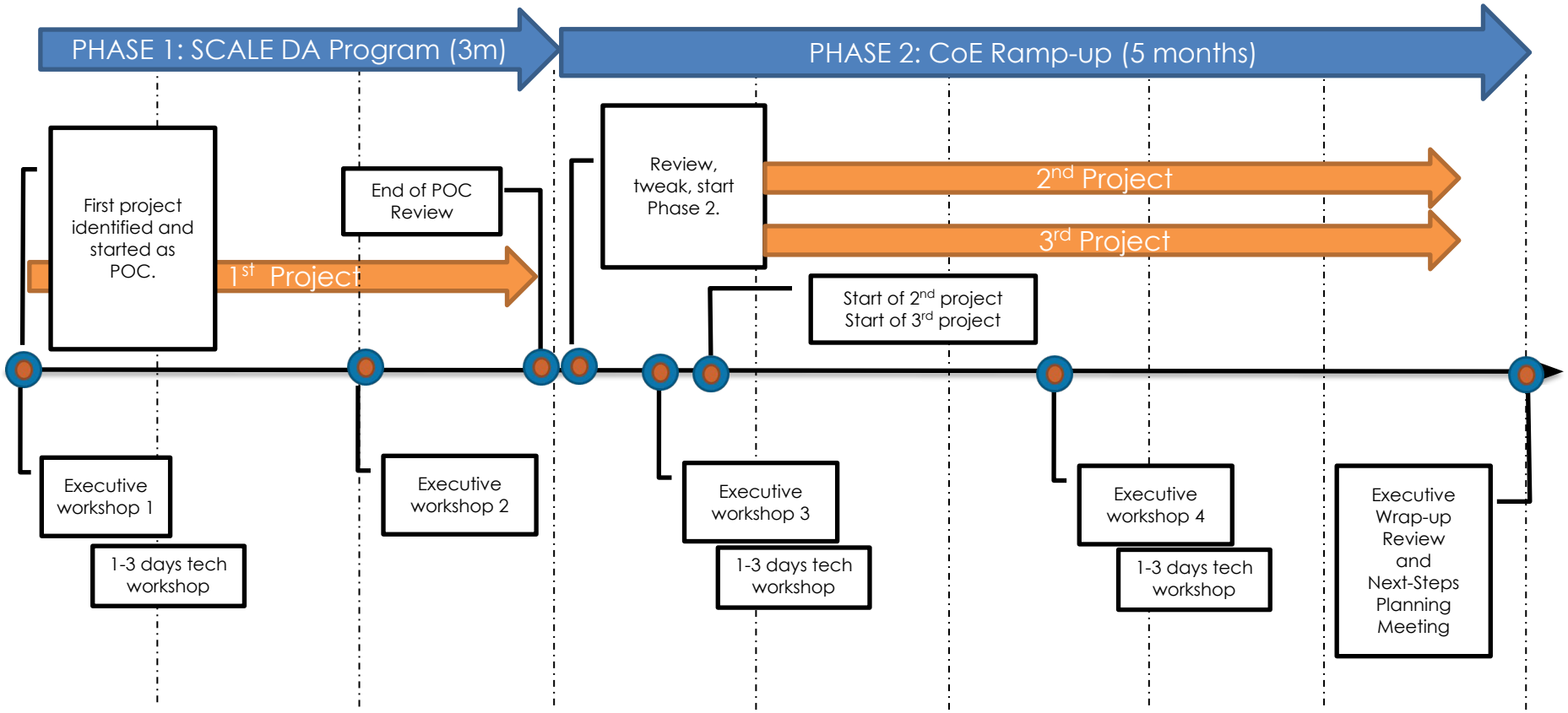
- Customer
  - 2 x project
  - Additional 2-4 data scientist to be trained in R/Python (total 6 data scientist)
  - Additional 2 x IT to be trained on R/Python enterprise integration and Analytics Stack (total 2 IT staff)

- **Build bigger/more complex models**
- **Build up your Analytics Stack**

Reference only

PHASE 2

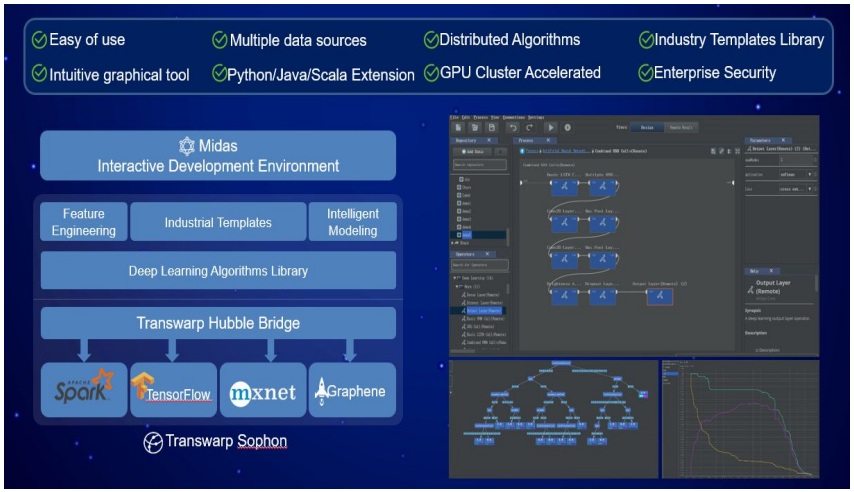
BIGGER PROJECTS + ANALYTICS STACK



PHASE 3:

16 MONTHS COE CAPABILITY DEVELOPMENT

- **PARTNER:**
  - 1 x Big Data Analytics Architect (20%)
  - 1 x Lead Data Scientist (20%)
  - 1 x Senior data scientist (100%)
  - 1 x Big Data developer (50%)



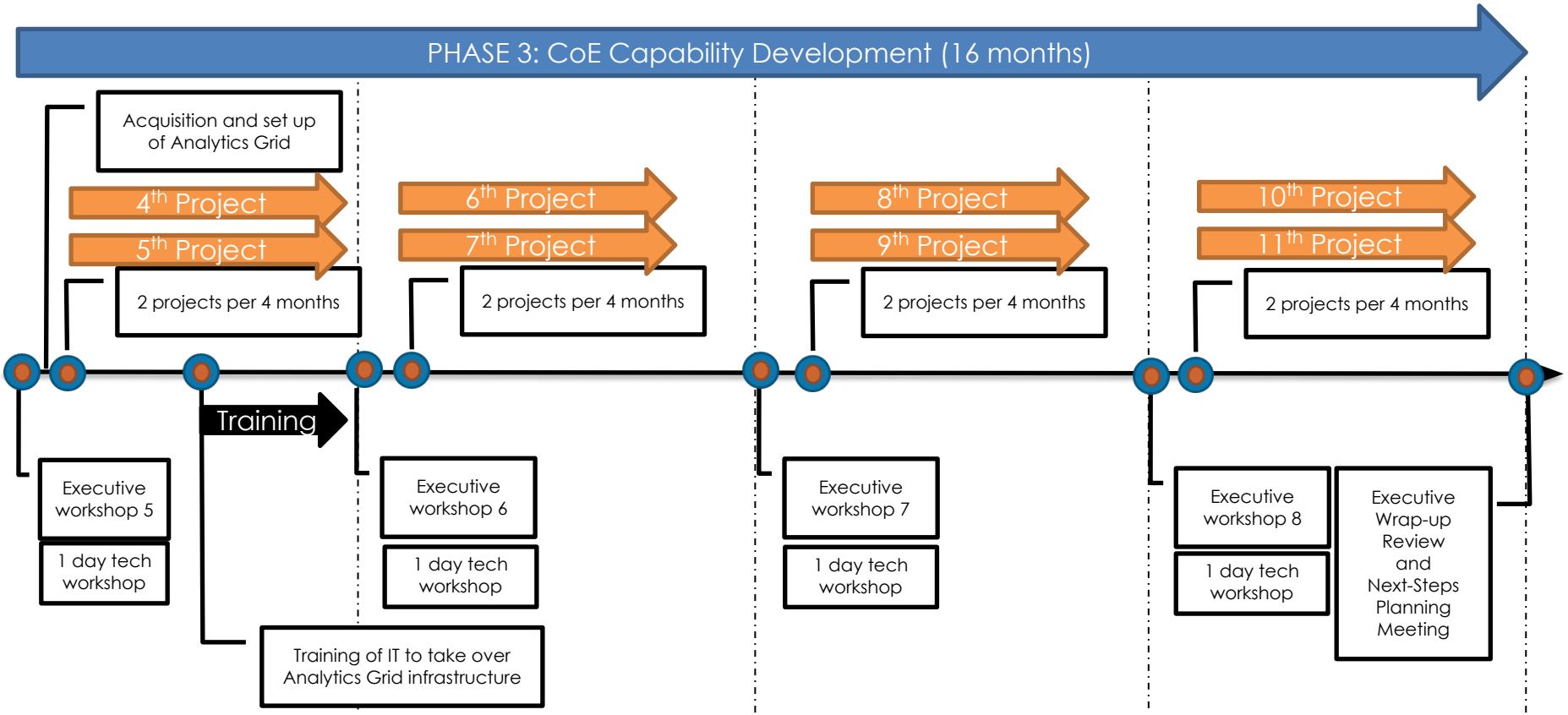
- **Customer**
  - 6 - 8 x projects
  - Additional data scientists to be trained in R
  - Additional IT to be trained on R enterprise integration

- **Build FASTER, bigger/more complex models**
- **Full Analytics Stack up and running**

Reference only

PHASE 3

FULL ANALYTICS STACK ACHIEVED AND TEAM IN PLACE



PHASE 3

ORGANIZATION-WIDE INNOVATION WITH ANALYTICS

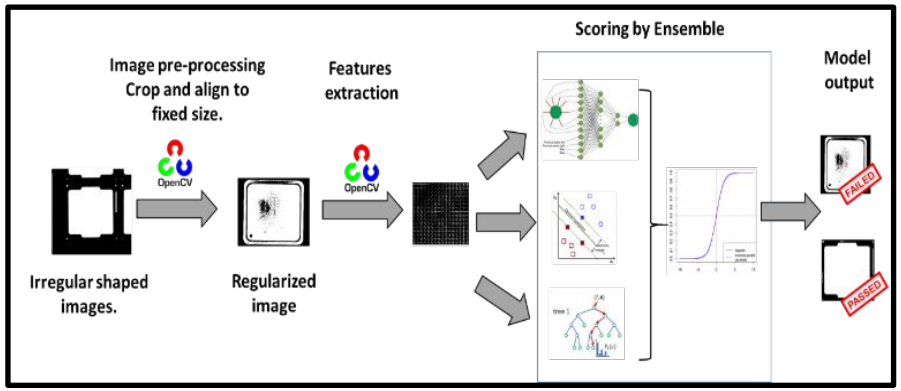
- Customer Organization
  - Core Data Science team formed
  - Core Devops team formed
  - A scalable analytics platform built to undertake the various analytics workload
  - Target a sustainable 4 - 6 projects per year
  - Agile/Scrum methodology recommended to ensure success

- SCALE
  - Step down and provide on-going advisory if required

- End of 24 months
  - Review next steps

SUMMARY

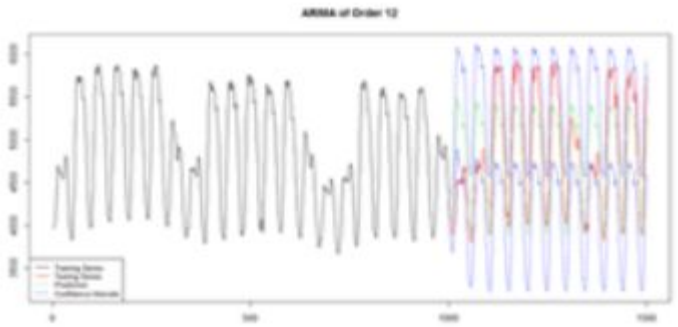
- Its only a template!
- Every organization will be different



Well-known to personal experiences

ANALYTICS USE CASES

ARIMA FOR TIME SERIES PREDICTION



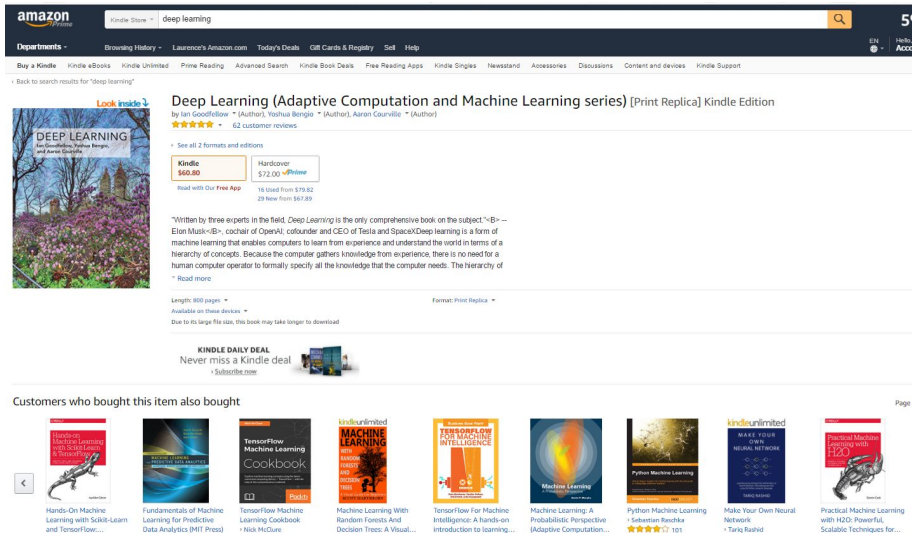
Energy Pricing is usually a function of supply and demand. However, making predictions on pricing at per-hour level can be difficult. Objective here was to create a machine learning model that could predict pricing at real-time intervals.

- First plot – ARIMA based model from Forecast Package in R
- Blue lines are the prediction intervals. Red is actual, and green is predicted
- Model captures seasonality nature of energy demand, low at night, high in day
- Prediction is accurate in the low demand states, ie between 10PM – 6AM, a lot of variation in the day
- Also, model does not take into account weekday/weekend effect; which is evident in the plot; where there are 5 big "humps" followed by 2-3 small "humps". That chart also shows a public holiday next to the weekend.

COLLABORATIVE FILTERING FOR RECOMMENDATION ENGINES

NUS National University of Singapore

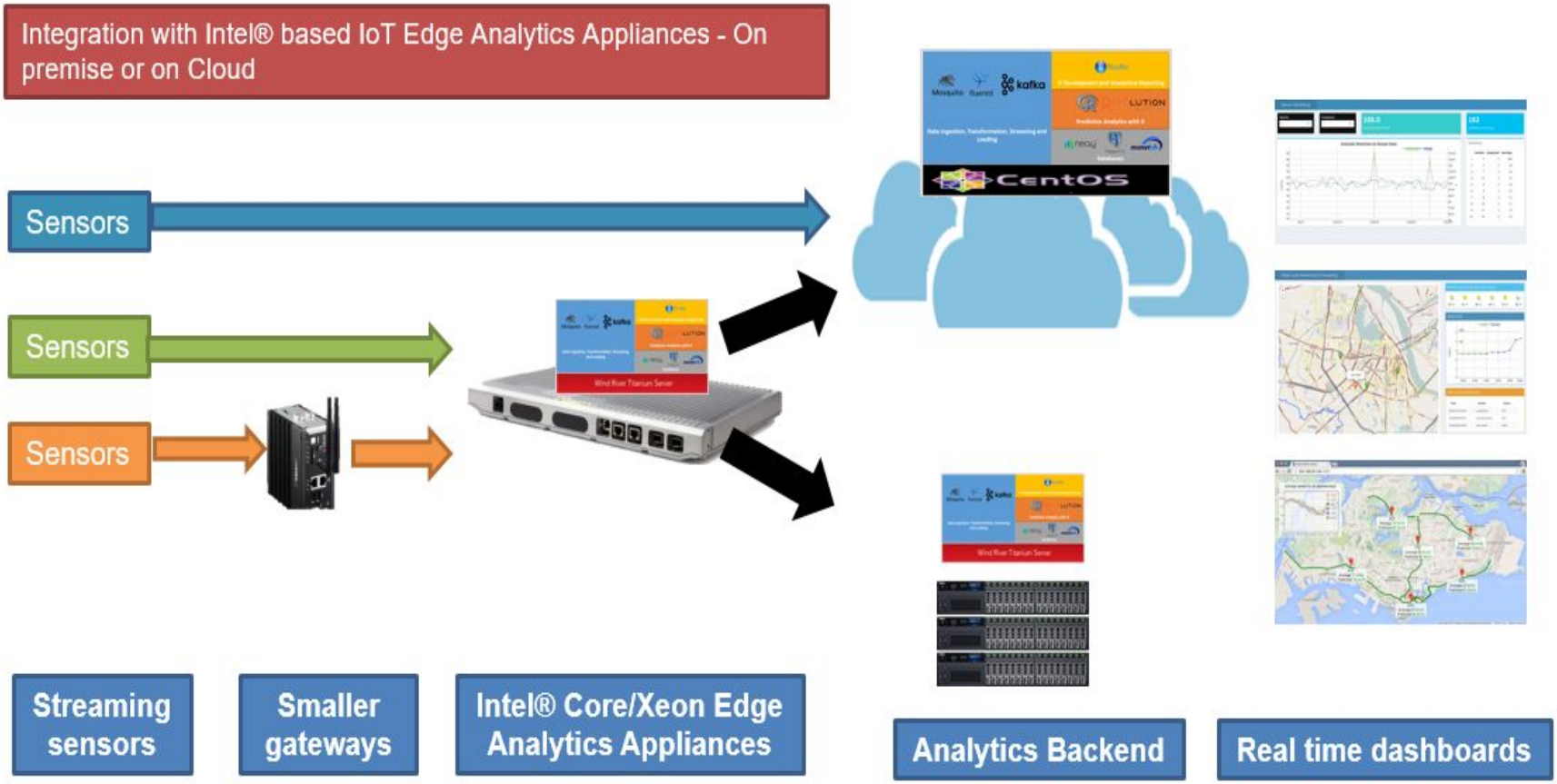
Indonesian telco with e-commerce site



https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html

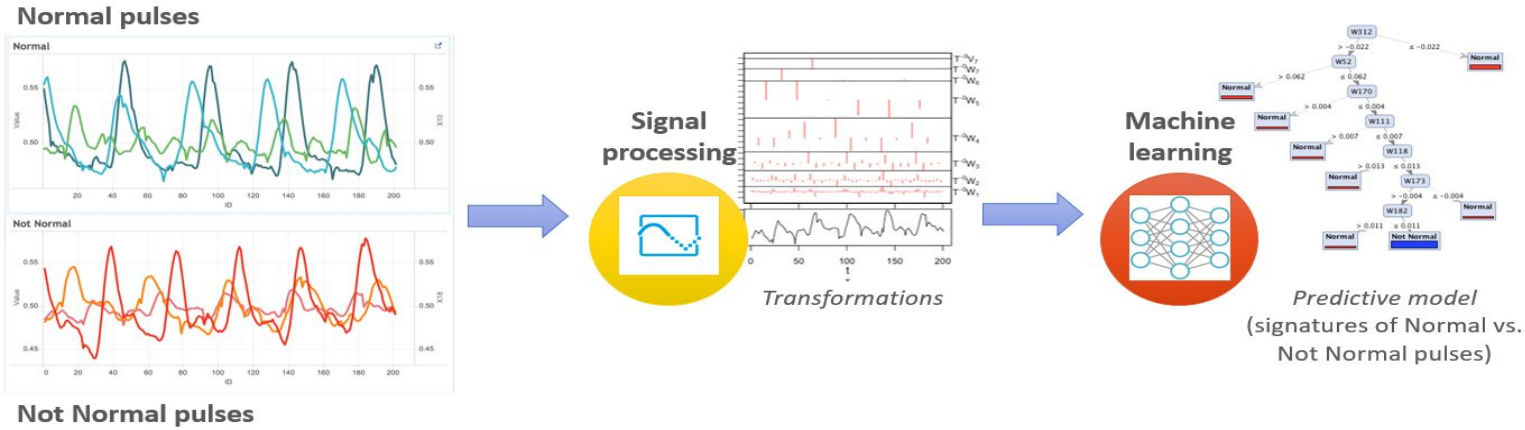
https://cran.r-project.org/web/packages/recommenderlab/index.html

EDGE ANALYTICS



APPLYING DA/AI TO TCM

Classification of normal and abnormal TCM pusles



Left0 Pulses	True Not Normal	True Normal	Precision
Predicted Not Normal	14	7	66.67%
Predicted Normal	6	15	71.43%
Recall	70.00%	68.18%	

Left1 Pulses	True Not Normal	True Normal	Precision
Predicted Not Normal	16	6	72.73%
Predicted Normal	4	16	80.00%
Recall	80.00%	72.73%	

Right1 Pulses	True Not Normal	True Normal	Precision
Predicted Not Normal	17	7	70.83%
Predicted Normal	5	13	72.22%
Recall	77.27%	65.00%	

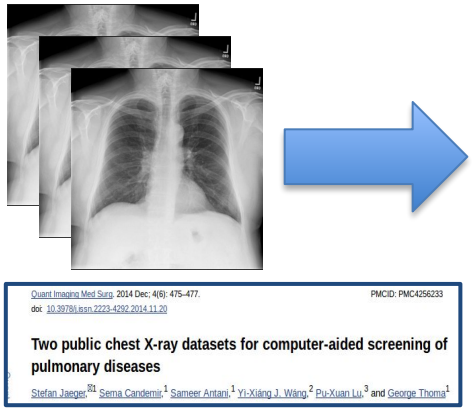
DECISION TREE/RANDOM FOREST FOR ROOT-CAUSE ANALYSIS

1. Component (hard disk manufacturing) involves multiple steps and the effect could be combinatorial when determining yield outcomes.
2. Objective was quickly identify important effects to the yield process
3. Visualization of Decision Trees
4. First branch of the tree indicates the most important variable
5. End result is a machine learning way to further narrow down and identify cause of low yield

DEEP LEARNING FOR IMAGE CLASSIFICATION

To classify normal vs abnormal chest-x-rays for a hospital

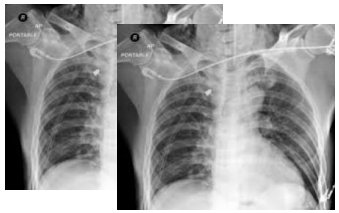
Validating the idea



Quant Imaging Med Surg. 2014 Dec; 4(6): 475–477. PMCID: PMC4256233
doi: 10.3978/j.issn.2223-4292.2014.11.20

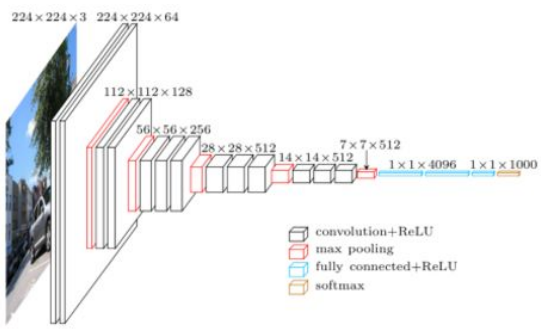
Two public chest X-ray datasets for computer-aided screening of pulmonary diseases

Stefan Jaeger,[&] Sema Candemir,[1] Sameer Antani,[1] Yi-Xiang J. Wáng,[2] Pu-Xuan Lu,[3] and George Thoma[1]



How does it perform on real customer data?

Pre-trained network vgg16



Performance on public chest-x-rays

Initial Results :
0.96 Training Accuracy
0.81 Validation Accuracy

Further finetuning is possible:
Image Augmentation to deal with small data size
Finetune final pre-trained layer

Performance on sample customer chest-x-rays

Initial Results :
0.96 Training Accuracy
0.81 Validation Accuracy

Further finetuning is possible:
Image Augmentation to deal with small data size
Finetune final pre-trained layer

MANUFACTURING ANALYTICS

IOT + ANALYTICS

Optimizing Manufacturing with IOT and Analytics

White Paper
Big Data Analytics Pilot, Internet of Things
Manufacturing

Optimizing Manufacturing with the Internet of Things

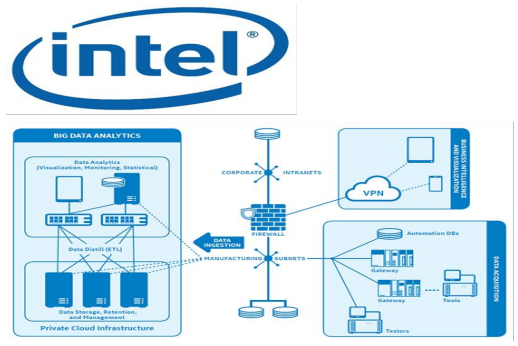
Intel manufacturing advances operational efficiencies and boosts the bottom line with an IoT and big data analytics pilot

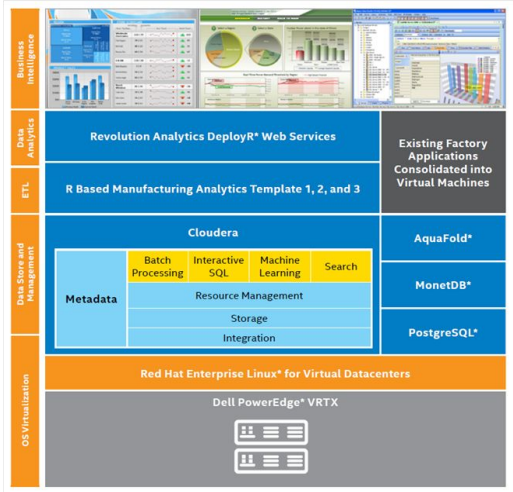
Intel proves IoT success by using big data analytics to bring cost savings, predictive maintenance, and higher product yields to its own manufacturing processes.

Introduction

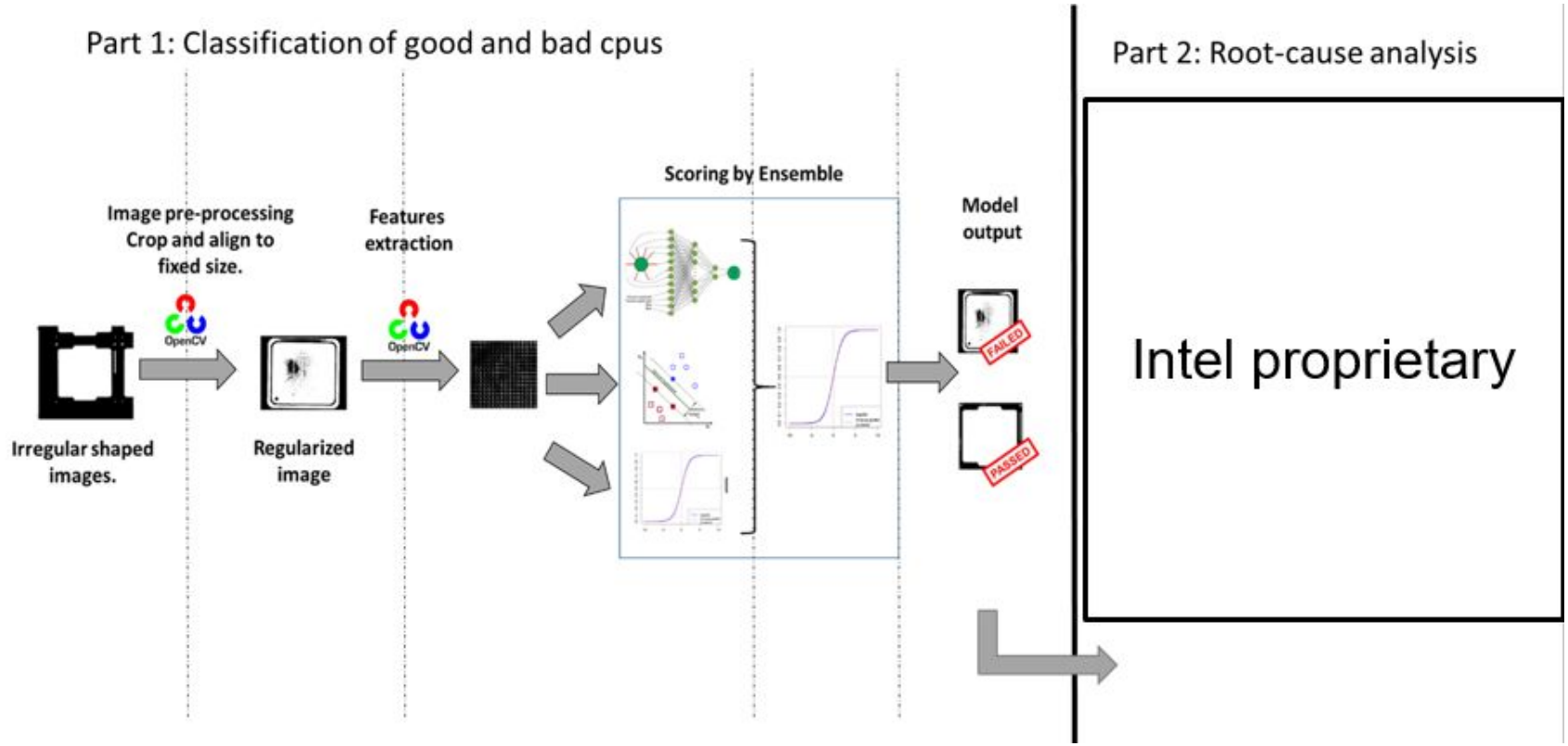
The volume, variety, and velocity of data[1] being produced in manufacturing is growing exponentially, creating opportunities for those diligently analyzing this data to gain a competitive edge, respond to changing market dynamics, and increase manufacturing margins, productivity, and efficiency. Equipment across the factory floor is generating thousands of different data types, such as unit production data at multiple levels, equipment operation data, process data, and human operator data, which can be stored to be made available for analysis.

Although large manufacturers have been using statistical process control and statistical data analysis designs for manufacturability, thereby improving supply chain management and introducing the use of customized



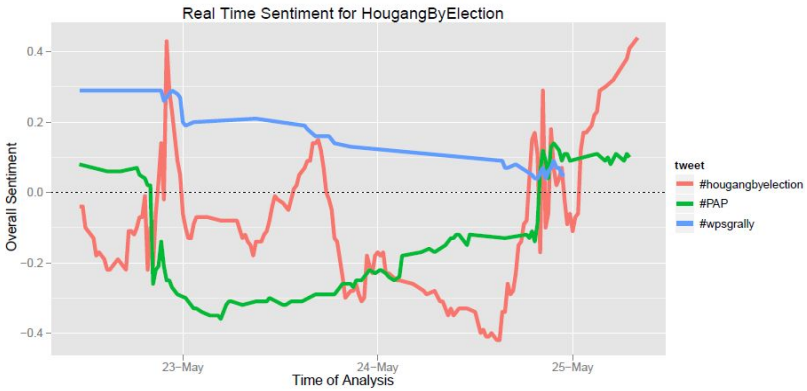


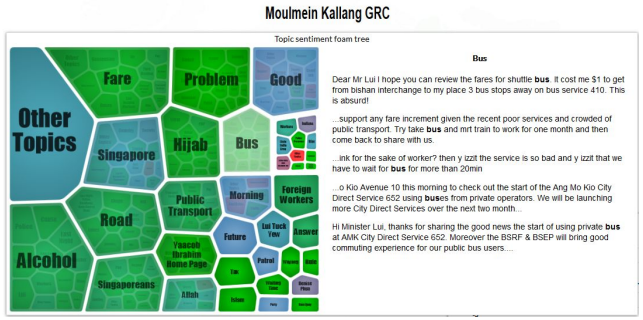
ENSEMBLE OF NN, SVM, LOGIT TO CLASSIFY CPU
INTEL USE CASE 3

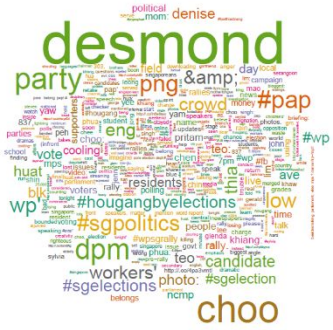


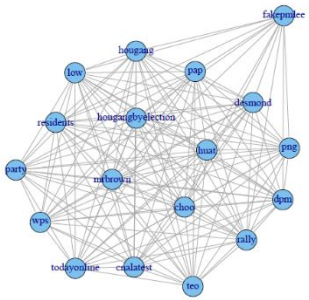
SOCIAL NETWORK ANALYSIS FOR KEY INFLUENCER DETECTION

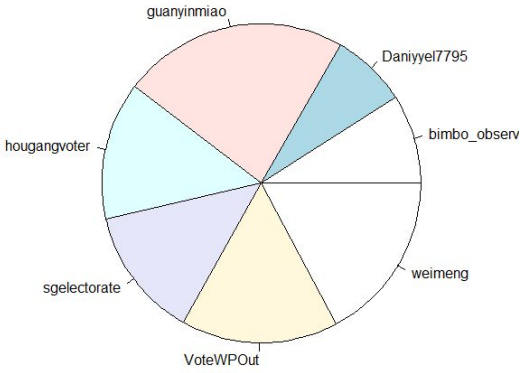
Sentiment Analysis of Hougang By Election (2012)

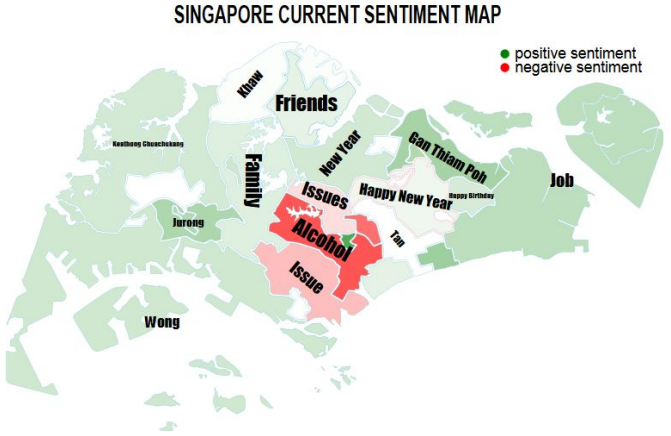








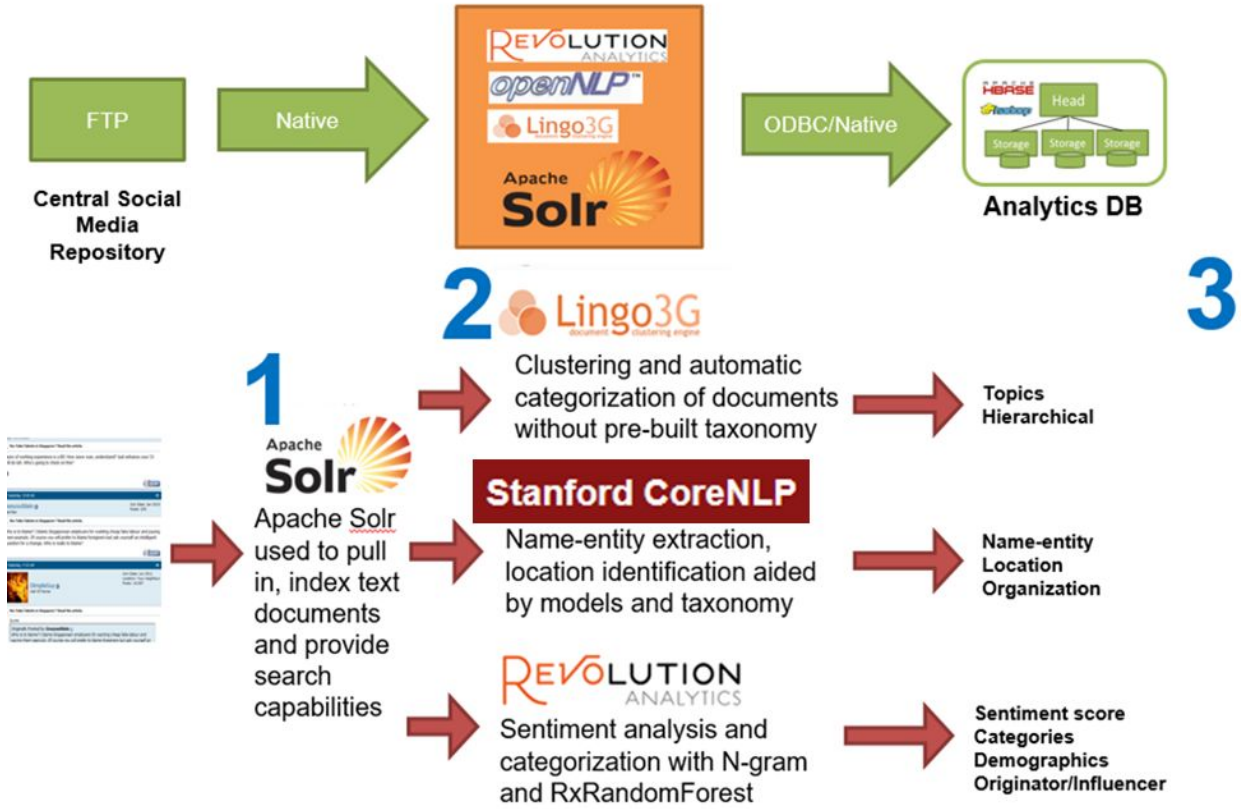




NATURAL LANGUAGE PROCESSING (NLP)

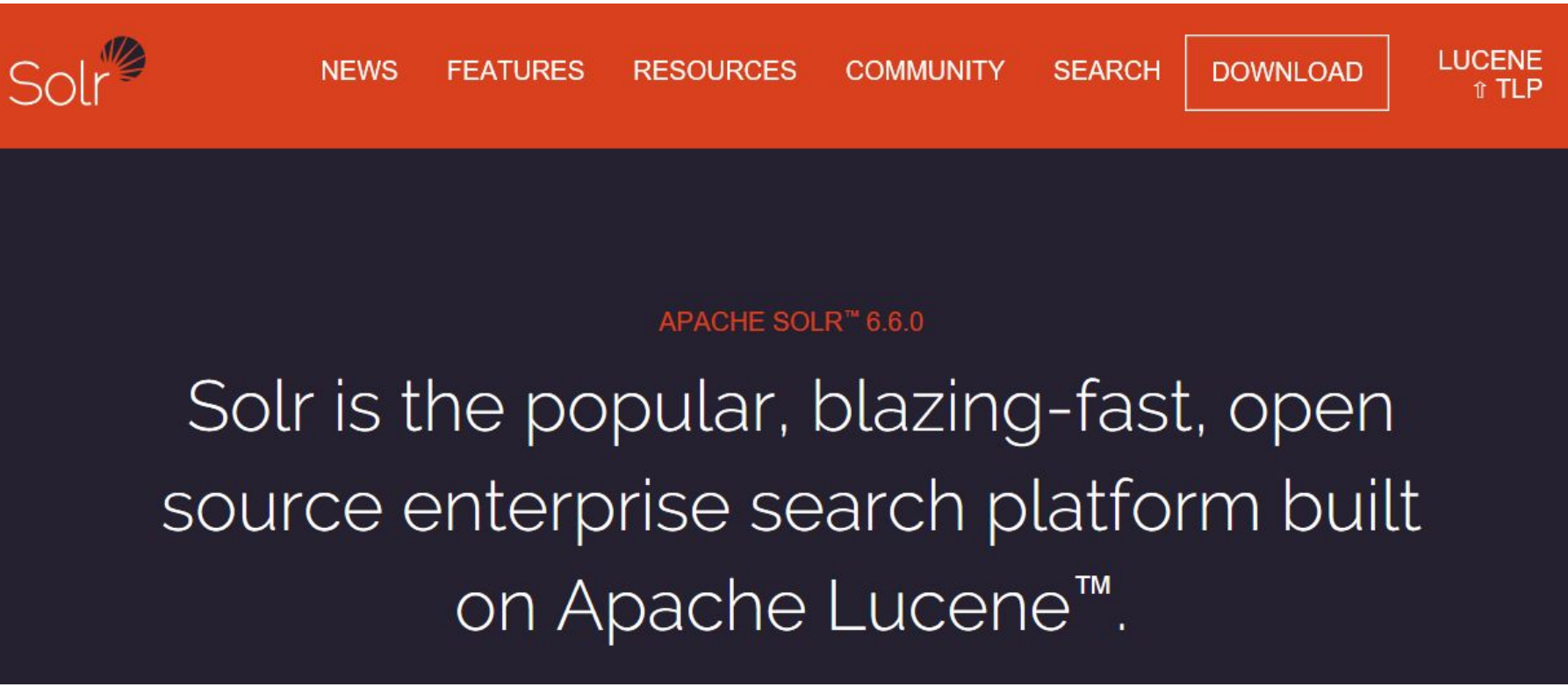
TEXT ANALYTICS AND SENTIMENT ANALYSIS

The "Brains"



TEXT PROCESSING

APACHE LUCENE & SOLR



Solr is highly reliable, scalable and fault tolerant, providing distributed indexing, replication and load-balanced querying, automated failover and recovery, centralized configuration and more. Solr powers the search and navigation features of many of the world's largest internet sites.

TEXT CLUSTERING & VISUALIZATION

SOFTWARE COMPONENTS FOR INNOVATIVE APPS, SERVICES AND ORGANIZATIONS

REAL-TIME TEXT CLUSTERING

Lingo3G

Organizes text documents into clearly labelled thematic groups. Accurately, on-the-fly, without knowledge bases.

Learn more

LARGE-SCALE TEXT CLUSTERING

Lingo4G

Identifies clearly labelled themes among millions of documents and gigabytes of text. Automatically, in near-real-time.

Learn more

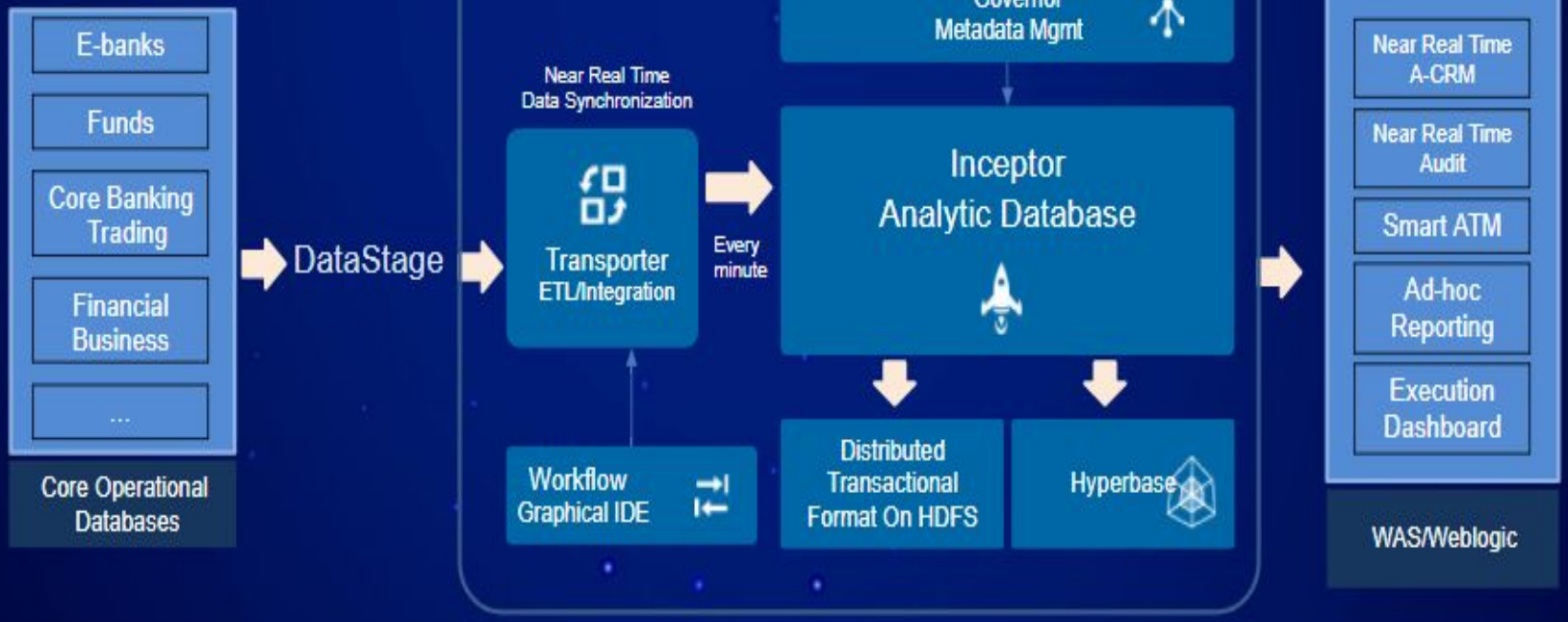
SEARCH RESULTS CLUSTERING

Carrot²

Popular open source search results clustering APIs and tools. Commercially friendly BSD license.

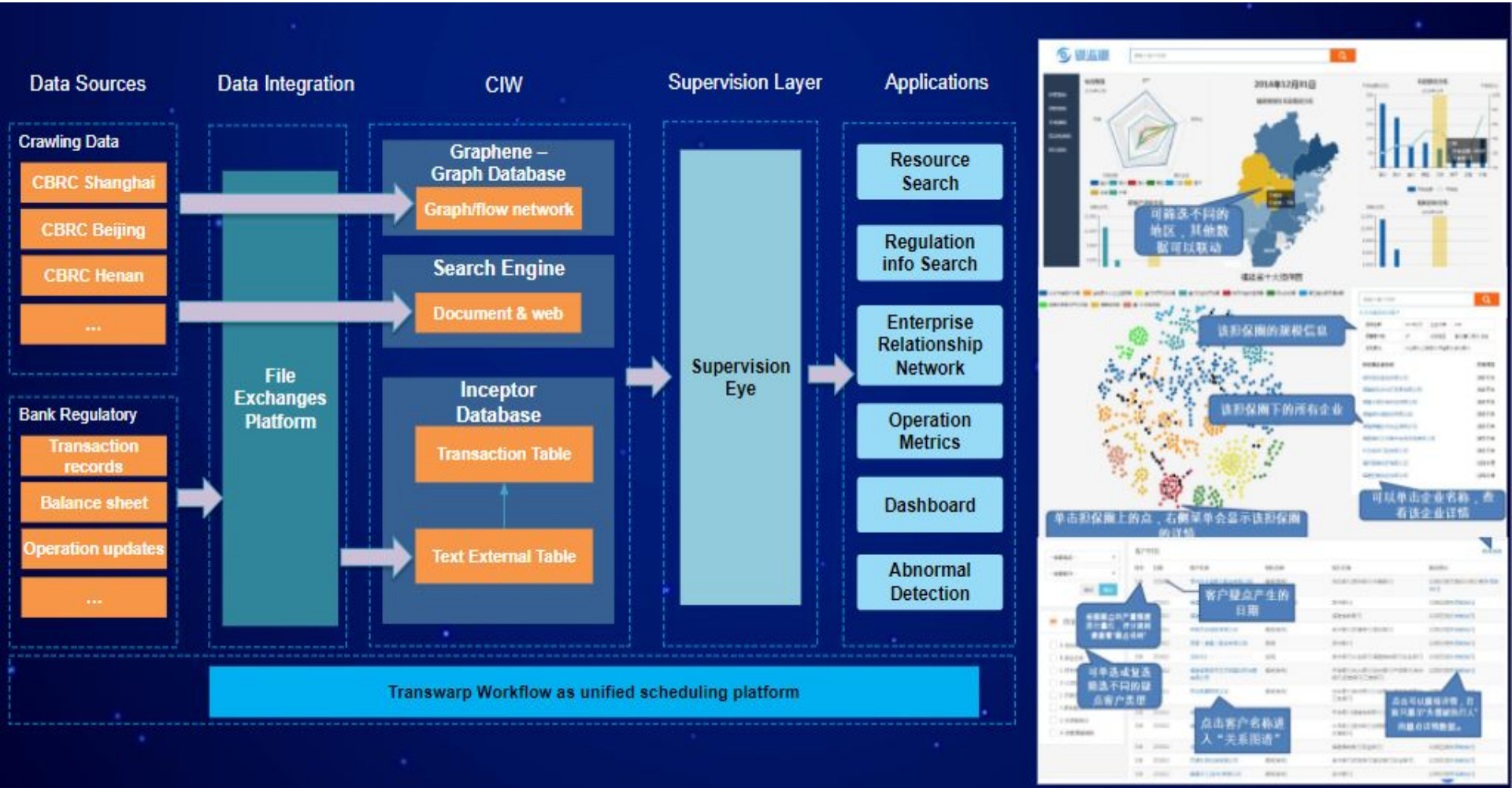
Learn more

NEAR REAL TIME DATA WAREHOUSE



ZJRC (Zhejiang Rural Commercial Bank) is the top 3 rural banks in China. It builds near real-time data warehouse and synchronize data from transaction system every minutes, and builds new near real time applications, including A-CRM, Audit System, Reporting System, etc.

LARGE SCALE FRAUD/MONITORING



SUMMARY

- Analytics is used by both large and smaller companies.
- Companies compete based on data today
- Wide range of use cases where analytics can bring business benefits.

PROJECT MANAGEMENT

DAY 1 WRAP-UP

Project Scoping Exercise

NUS National University of Singapore | School of Continuing and Lifelong Education

Phases of a Data Science Project

SCOPING

- Formulate problem statements and hypothesis
- Obtain stakeholder buy-in
- Sign data sharing agreement

DATA PREPARATION

- Collect raw data
- Manage privacy concerns through aggregation, anonymisation, hashing of fields
- Prepare the data for analysis

ITERATIVE ANALYSIS

- Analyse data to gain insights, discover patterns and validate hypothesis
- Apply relevant machine learning techniques

ACTIONABLE INSIGHTS

- Develop clear narrative to communicate insights
- Devise action plan and see through execution
- Recommend metrics to measure success

A bad example on “Improving NSmen Health”

What is the problem you are trying to solve and what does an improvement look like?

I want to make our NSmen healthier

What data do you collect that are relevant to the problem?

We can base it on their IPPT results. We don’t have the data now but we can approach MINDEF… probably…

What is the measurable impact of the project and are the key stakeholders bought in to recognise its value?

We want to make people healthier. I will need to check in with the bosses but should be fine…

Better

What is the problem you are trying to solve and what does an improvement look like?

How do I improve the passing rate of NSmen taking the physical fitness test?

What data do you collect that are relevant to the problem?

Physical fitness test results, demographics, training schedules for all NSmen across 2000 to 2015 stored in a relational database

What is the measurable impact of the project and are the key stakeholders bought in to recognise its value?

Measure and assess passing rate of NSmen after intervention. Agency senior management have approved this project with a monthly progress update meeting schedule

Project Activities Mapping

Activity	Phase 1	Phase 2	Phase 3
Project Planning	X		
Business Understanding and Analytic Goals	X	X	
Data Preparation	X	X	
Analysis & Modelling	X	X	X
Results Validation	X	X	X
Communicating Results	X	X	X

1. Business Objectives Understanding

Activities

- Assessment of current situation and Determine data analytic goals
- Initial assessment of data sources and Identify appropriate analysis approaches
- Evaluate alternative analysis approaches and Exploratory analysis and modelling
- Produce project plan

Deliverables, consisting of:

- Statement of business objectives and mapping to analysis objectives
- Assessment, selection and justification of appropriate analysis approaches
- Discussion of results of exploratory analysis and modelling
- Project plan

2. Data Preparation and Initial Modelling

Activities

- Application of chosen analysis techniques
- Initial analysis and modelling assessment
- Consideration of possible change to analysis approach
- Construction of validation tests
- Project plan update

Deliverables, consisting of :

- Data description
- Evaluation of appropriate analysis approaches
- Initial modelling results

3. Modelling, Validation and Conclusions

Activities

- Further analysis and modelling
- Testing, validation and assessment of results
- Create business conclusions from analysis conclusions
- Document and present results and recommendations

Deliverables, consisting of:

- Full modelling assessment & results
- Validation and testing results
- Final project conclusions

Some Assessment Criteria

Assessment

- Is the scope and content of the project sufficiently technically demanding?
- Have you adopted an appropriate analytical strategy?
- Have you found adequate solutions to business problems?
- Have you evaluated suitable alternatives to the approach chosen?
- Have you validated the results produced?
- Have you captured and documented the results produced?

…

Use these scoping questions, think about a problem you face in your work which you can possibly embark on a Data Analytics project to solve

Clear and Focused Problem Statement

1. How do you know the problem exists?
2. What is wrong with the current situation and why can't it continue the way it is?
3. What would an improvement look like?

Available Data

1. What data do you collect that are relevant to the problem?
2. Provide a sample set of the data

Value and Impact

1. How do you measure the improvements after the solution is implemented?
2. Who are the key stakeholders of the data acquisition and how onboard are they?
3. Who are the key stakeholders of the data implementation and how onboard are they?

Start a project...

How to use data analytics to solve…

Tasks:

- Scope and define the problem (e.g. value and impact)
- Choose analytical methods (e.g. descriptive, diagnostic or predictive)
- Select the sample and data sources
- Plan the project (e.g. time and resources)
- Communicate: 5 minutes elevator pitch to the class

Time: 30 minutes

THANK YOU