Data Science, the Key for Competing in Data-driven World

Setia Pramana

Politeknik Statistika STIS

Outline

- Data Driven World
- Big Data Sources
- Data Science
- What we should learn?
- Summary

The World is Changing

Big Data

90%
of the data created in the last two years alone.

Mobile

1 billion (plus)
(plus) smart devices shipped in 2013 alone.

Social

81%
of customers depend on social sites for purchasing advice.

Cloud

62%
of total workloads will be in the cloud by 2016.

Internet of Things

50 billion
devices connected to the internet by 2020.

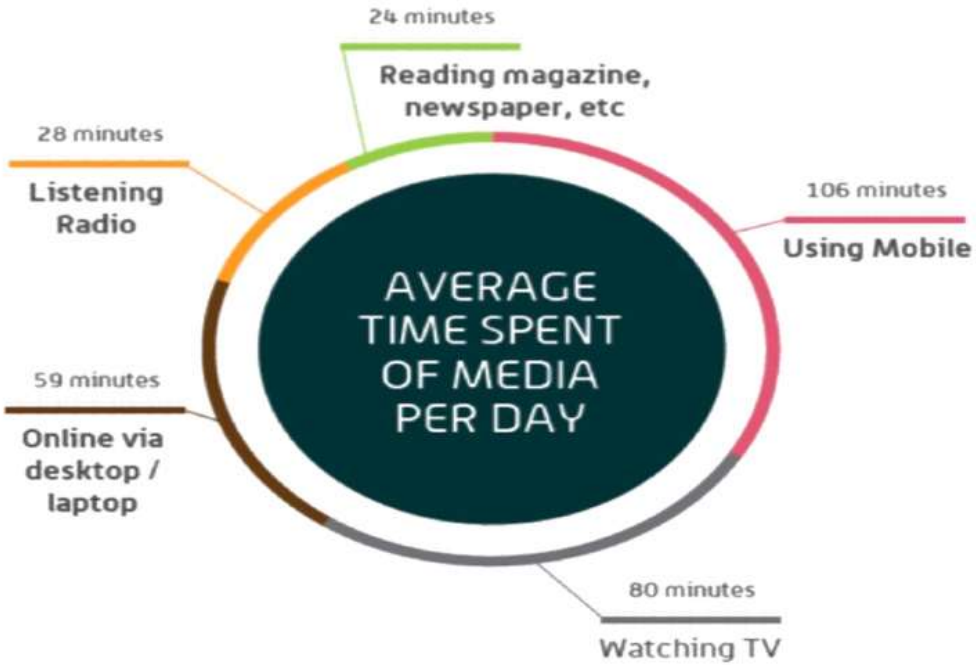
API Economy

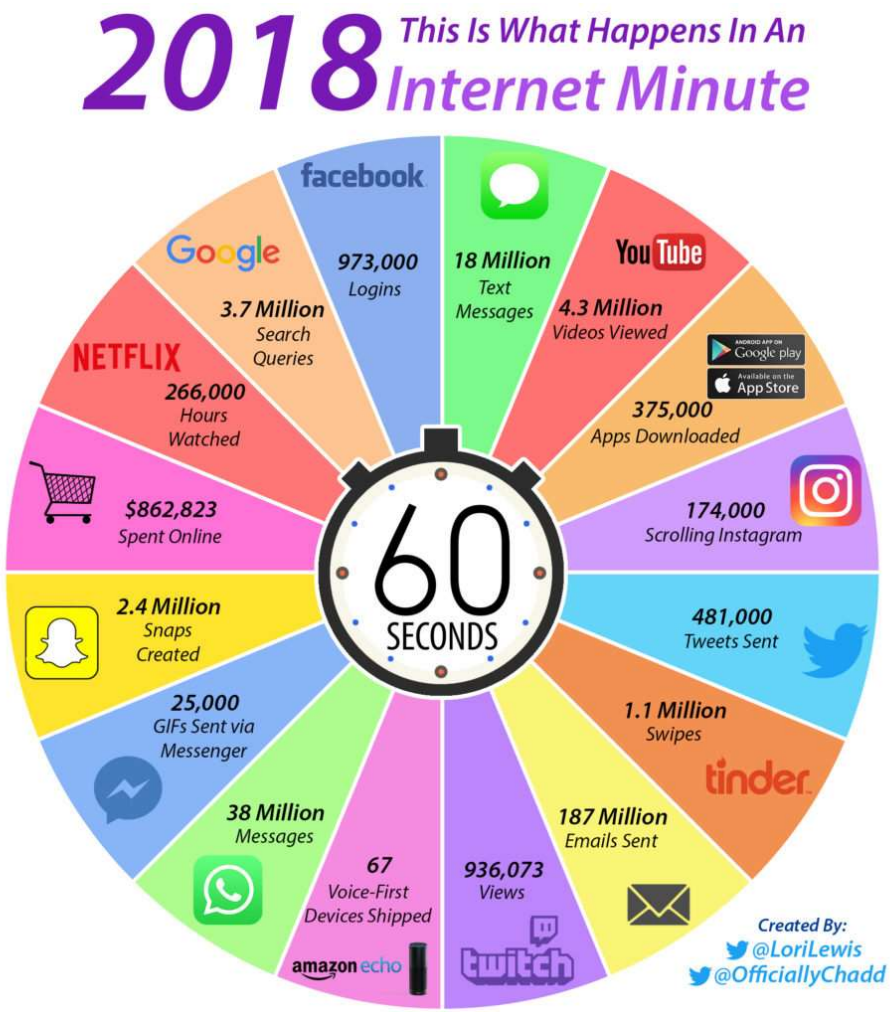
Global m-commerce sales were

85 billion
in 2013 and forecast to rise to $120 billion by 2015 and an estimated $1 trillion by 2017

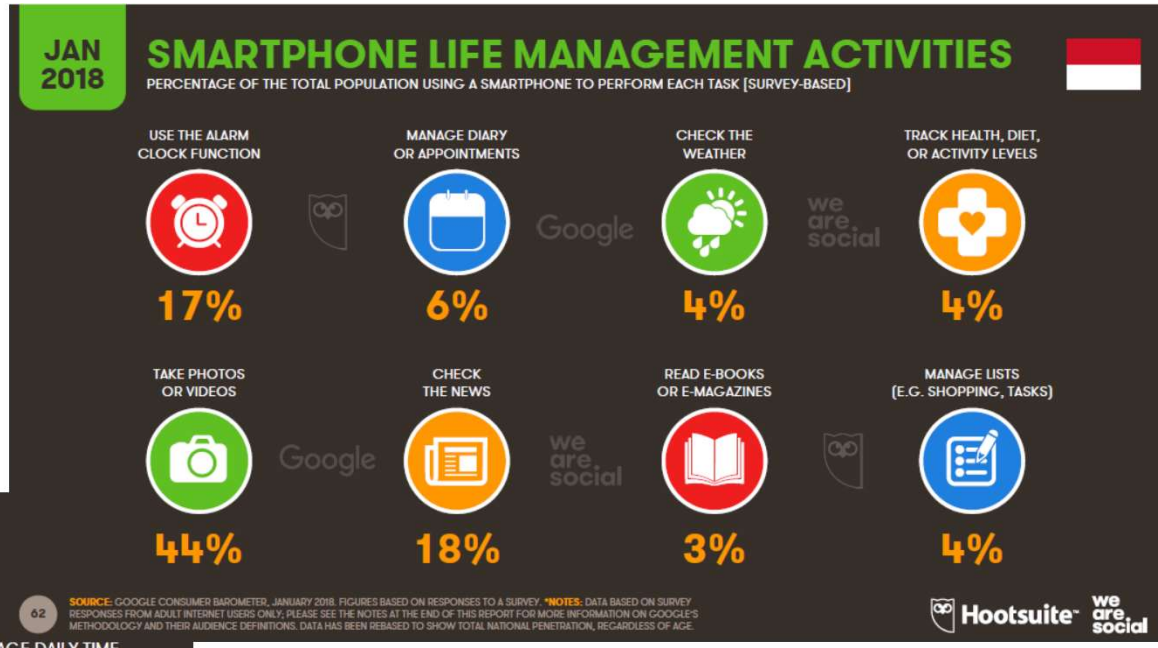
We can do everything Online!

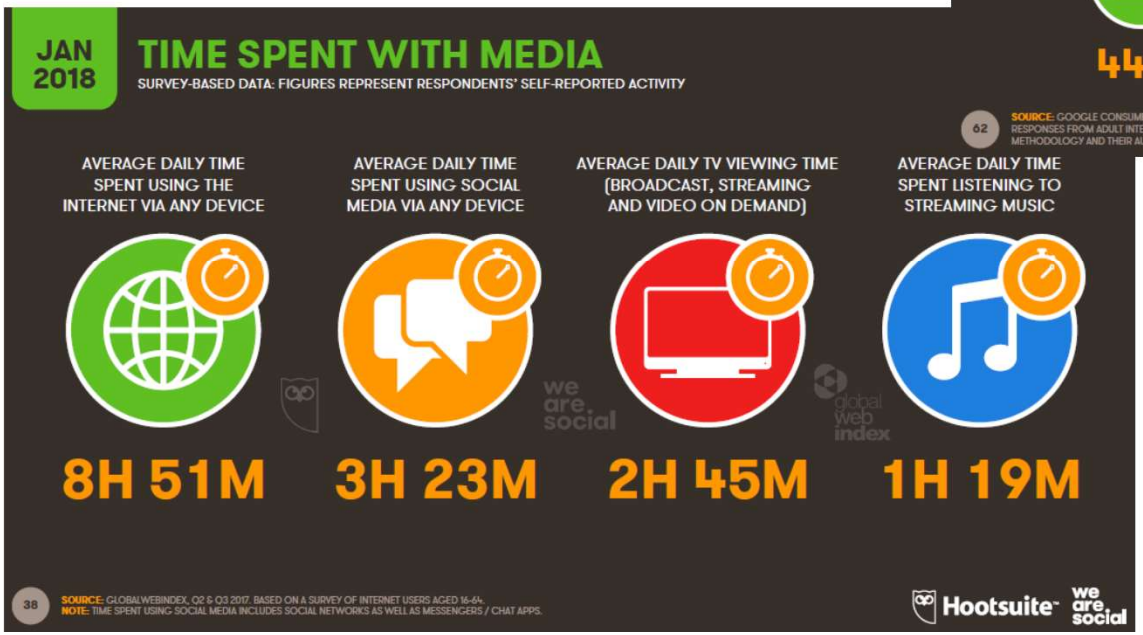
What we "produce"?



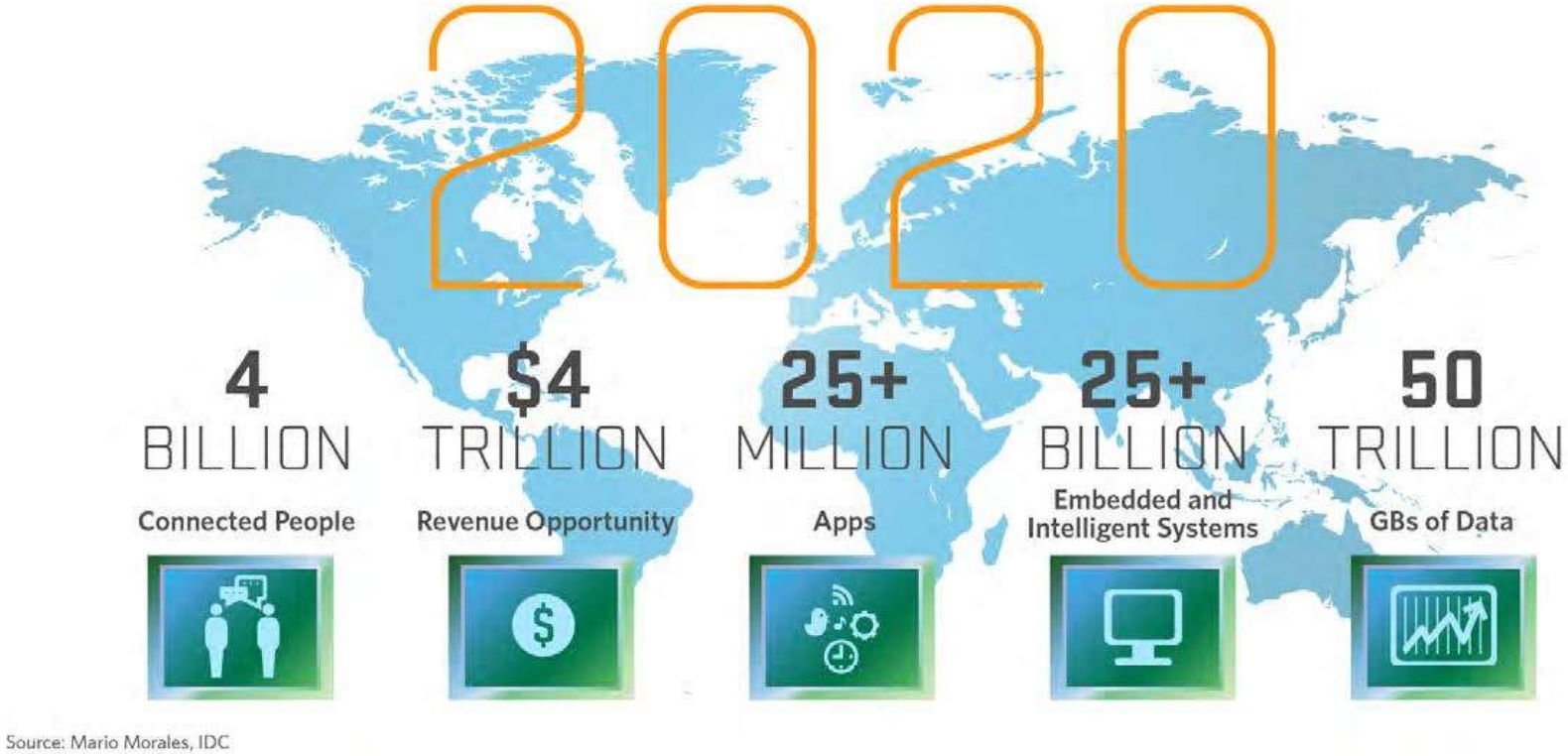


How about Indonesia?





Data Explosion

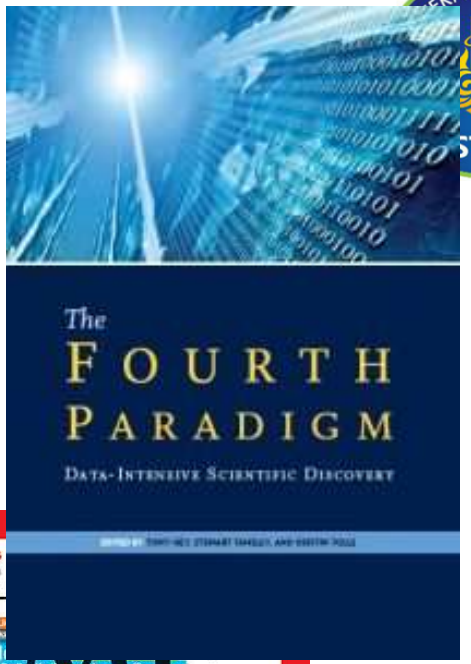


Source: Mario Morales, IDC


4 September 2008 | www.nature.com/nature | £10
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
nature
THE BITER BIT
Viral infections for viruses
TROPICAL CYCLONES
The strong get stronger
BLACK HOLE PHYSICS
A new window on the Galactic Centre
BIG DATA
SCIENCE IN TH
PETABYTE ER

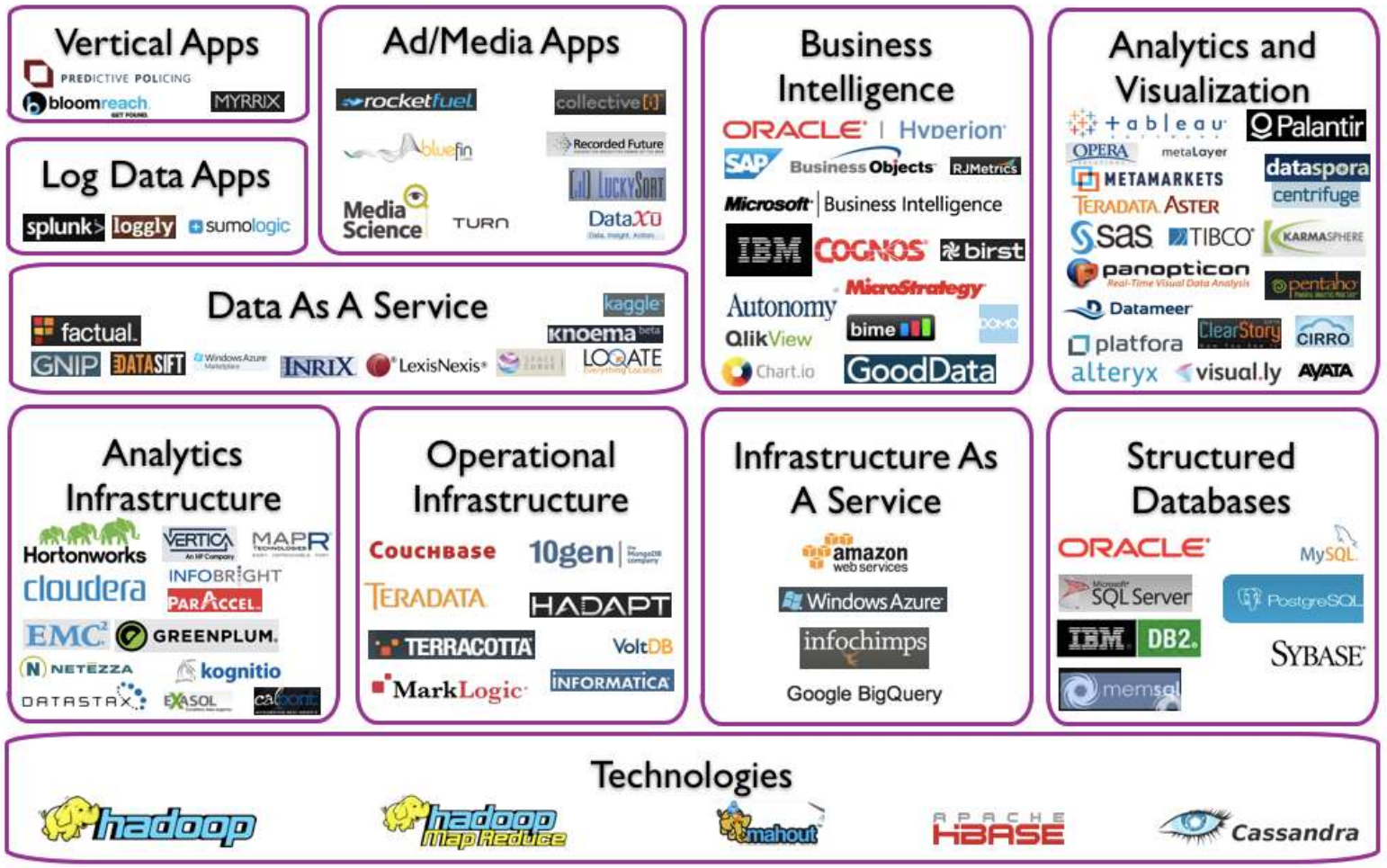

POPULAR SCIENCE
THE CONTROL CENTERS
Using Data to Feed the World, Solve Cold Cases, Battle Malware, Predict Our Fate
OFFICER ALGORITHM
Can a Crime Be Prevented Before It Begins?
NEW WAYS OF SEEING
A Gallery of Extraordinary Infographics
PLUS
Juan Enriquez
James Gleick
AND
Lawrence Weschler
Questions the Cloud
SPECIAL ISSUE
DATA IS POWER
HOW INFORMATION IS DRIVING THE FUTURE


MIT Technology Review PAKISTAN
Published by ITU
Special Feature
How President Obama's Campaign Used Big Data to Rally Individual Voters
p.64
BIG ISSUE
SOCIAL MEDIA AND
BIG DATA
IN POLITIC
Growing Role of Social Media and Technology in Pakistani Politics


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY

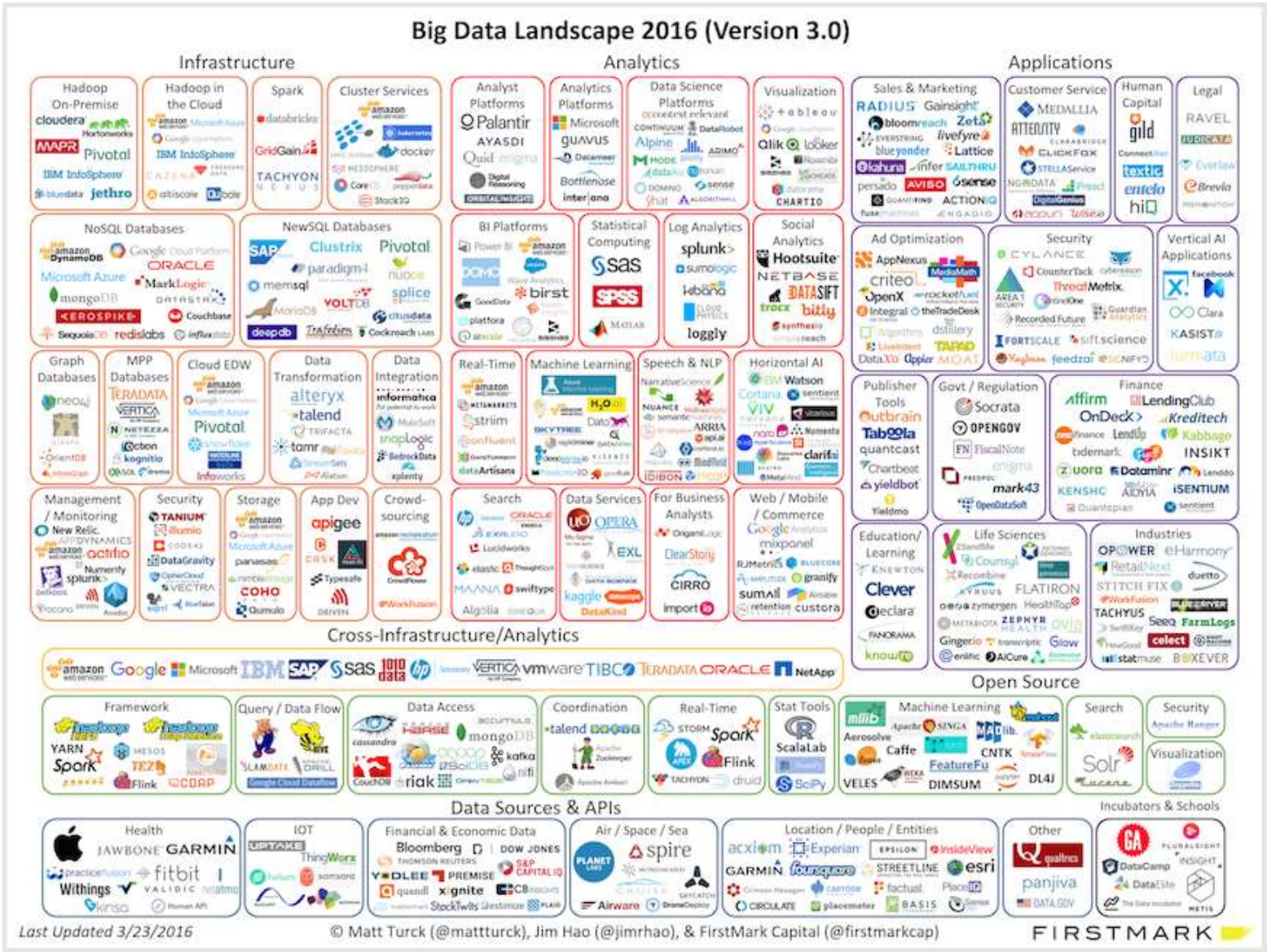

A tale of two Libyas
Plus: Why the U.S. can't sit on the sidelines BY FAREED ZAKARIA
TIME
YOUR DATA
FOR SALE
Everything about you is being tracked—get over it
BY JOEL STEIN
Household income: $100,000+
Owns a laptop
Age: 38-39
Likes: online news
Likes: Asian cuisi
Likes: green living
Favorite celebrities:
ZIP code: 1070
Wi-fi warrior
Likes: business & financ
Sister is a la
Likes: parenting
Spent $180 on intimate app. & underarments on Oct. 10, 2010
Likes: hiking
Dislikes: autos & vehicles
Married
Likes: movies
Magazine subscriber
Smart-phone user
Has had LASIK surgery
Likes: restaurants

Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib


Big Data Landscape 2016 (Version 3.0)
Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt / Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
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

Analytics Approaches

- Descriptive: What happened or what is happening now?
- Diagnostic: Why did it happen or Why is it happening now?
- Predictive: What will happen next? What will happen under various conditions?
- Prescriptive: What are the options to create the most optimal/high value result/outcome?

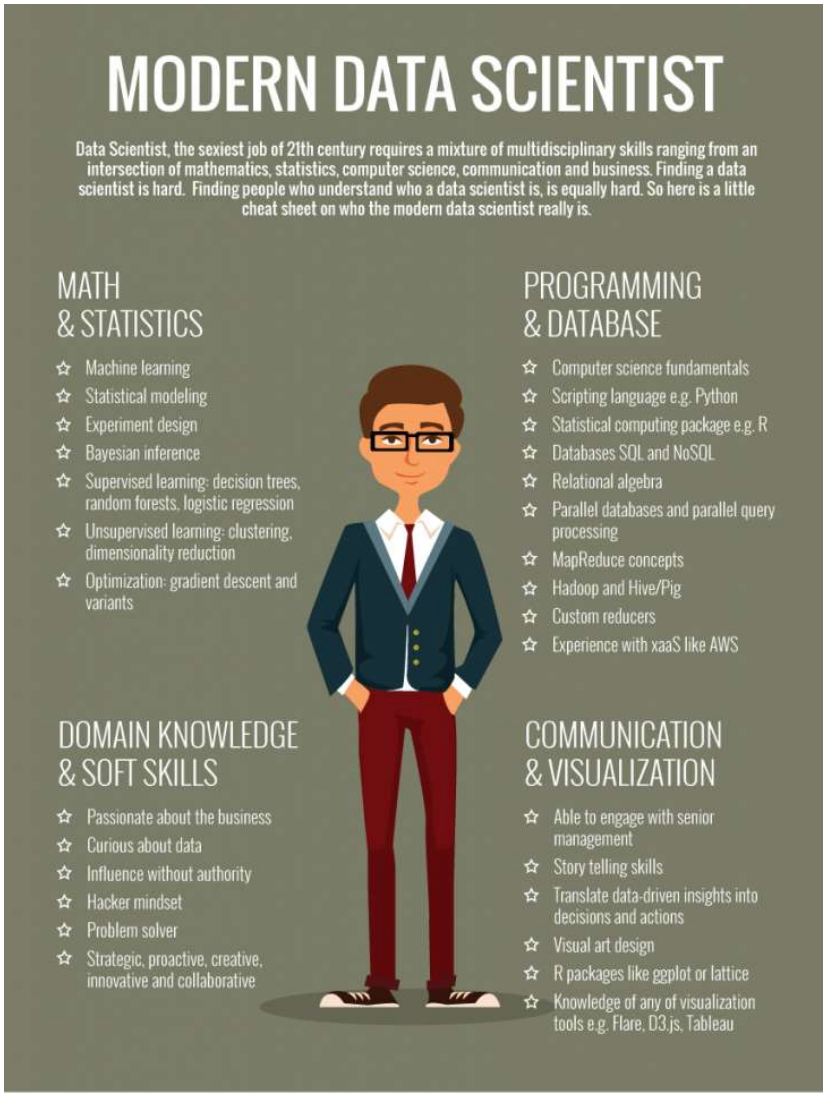
Expertise Needed

“The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.”

Hal Varian, chief economist at Google and UC Berkeley professor of information sciences, business, and economics

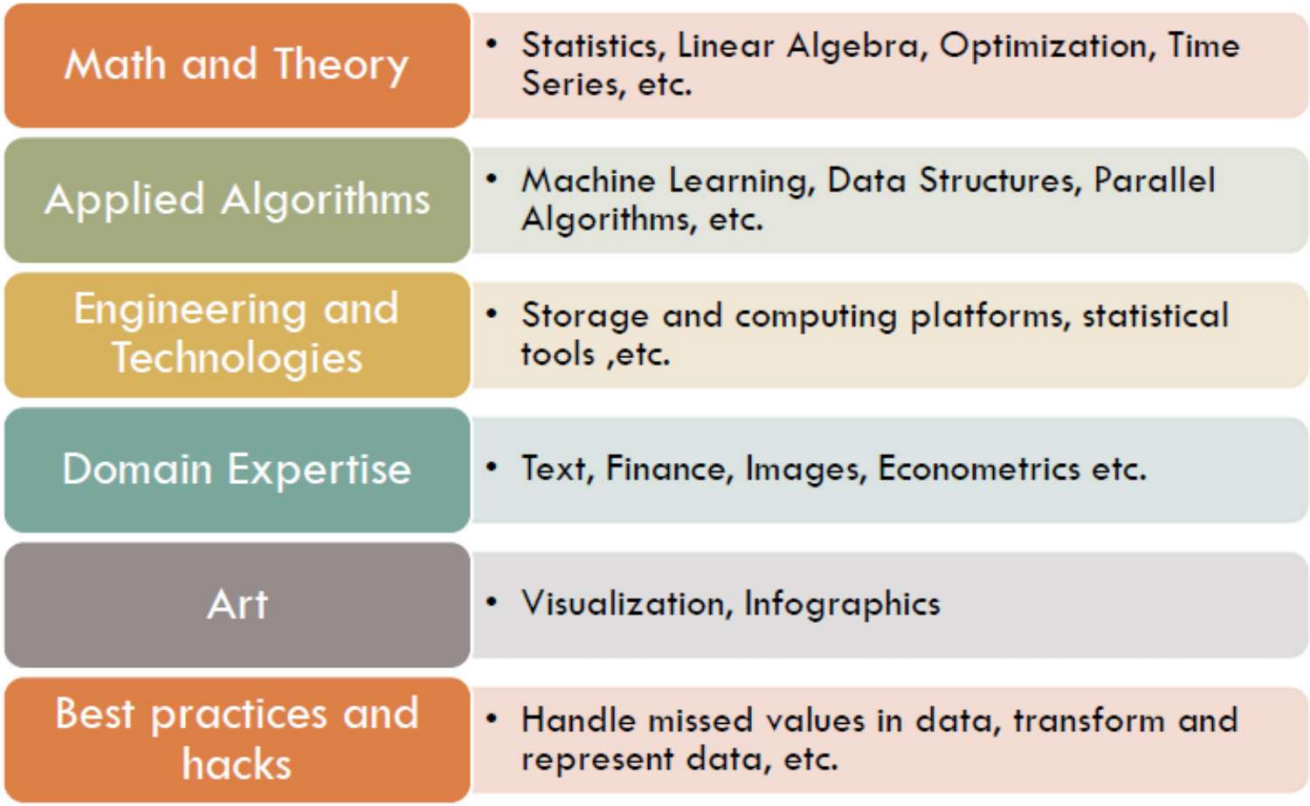

ANATOMY OF A
DATA SCIENTIST
SALARY
Average salary of data scientists is
$120,000/year
BENEFITS
• Harvard Business Review called data science the "Sexiest Job of the 21st Century"
• One of the fastest growing careers in the United States
• 94% of data science graduates have found jobs since 2011
RESPONSIBILITIES
• Conduct research
• Extract, clean, and analyze data from varied sources
• Solve problems
• Build automation tools
• Communicate findings to management
EDUCATION
• 88% of all data scientists have at least a Master's degree
• 46% of data scientists have a PhD
SKILLS
• Programming languages (R, Python, SQL, Hive, etc.)
• Statistics
• Multivariable calculus and linear algebra
• Machine learning
• Software engineering
• Wrangle, visualize, and communicate data to management
CAREER POSSIBILITIES
• The majority of data scientists work in the technology industry.
• Other options include marketing, consulting, healthcare and pharmaceuticals, finance, government, gaming, and many more.
RESOURCES:
https://insidebigdata.com/2017/08/05/benefits-data-scientist-career/
https://www.glassdoor.com/Salaries/us-data-scientist-salary-SRCH_IL.0,2_IN1_KO3,17.htm
https://blog.udacity.com/2014/11/data-science-job-skills.html
THE COMPUTER MERCHANT, LTD.

orate

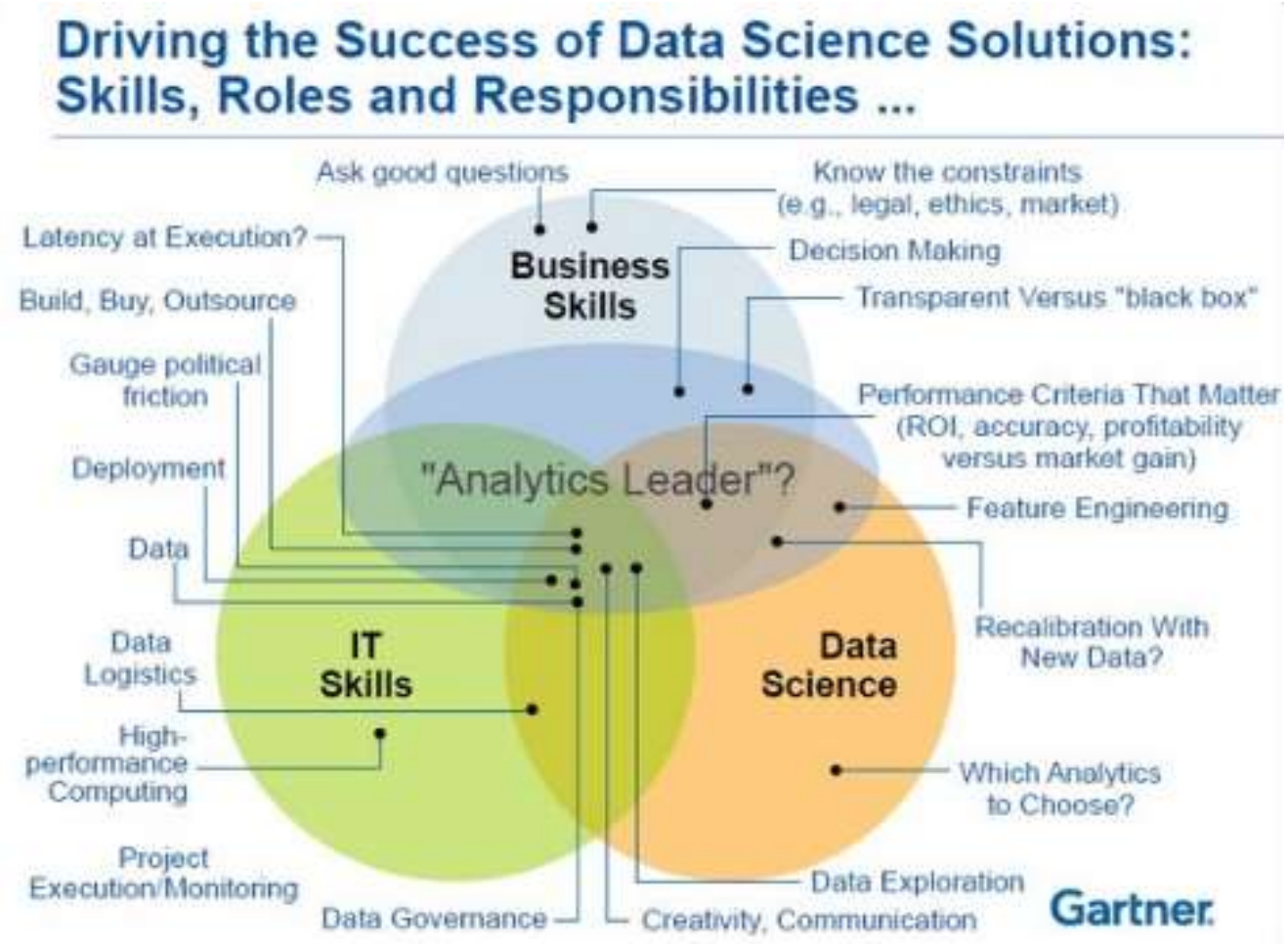

MODERN DATA SCIENTIST
Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.
MATH & STATISTICS
☆ Machine learning
☆ Statistical modeling
☆ Experiment design
☆ Bayesian inference
☆ Supervised learning: decision trees, random forests, logistic regression
☆ Unsupervised learning: clustering, dimensionality reduction
☆ Optimization: gradient descent and variants
PROGRAMMING & DATABASE
☆ Computer science fundamentals
☆ Scripting language e.g. Python
☆ Statistical computing package e.g. R
☆ Databases SQL and NoSQL
☆ Relational algebra
☆ Parallel databases and parallel query processing
☆ MapReduce concepts
☆ Hadoop and Hive/Pig
☆ Custom reducers
☆ Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS
☆ Passionate about the business
☆ Curious about data
☆ Influence without authority
☆ Hacker mindset
☆ Problem solver
☆ Strategic, proactive, creative, innovative and collaborative
COMMUNICATION & VISUALIZATION
☆ Able to engage with senior management
☆ Story telling skills
☆ Translate data-driven insights into decisions and actions
☆ Visual art design
☆ R packages like ggplot or lattice
☆ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.
Marketing DISTILLERY

Data Science

- A Mashed Up Discipline



Data Science



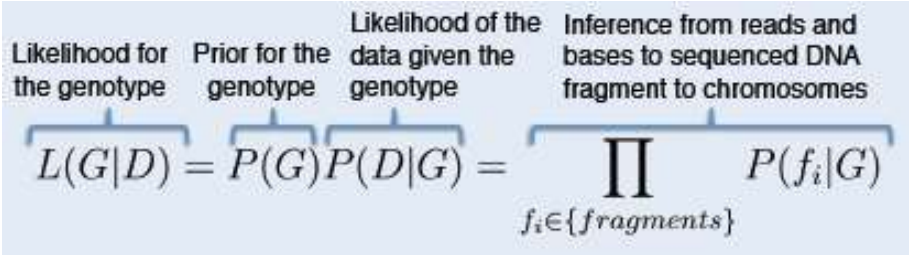
Huge Demands

28%	4,524	$120,931	#1
Demand Increase by 2020	Number of Job Openings	Average Base Salary	Best Job in America 2016, 2017, 2018

Sources: **Glassdoor** and **Forbes**

Data Scientist Roles

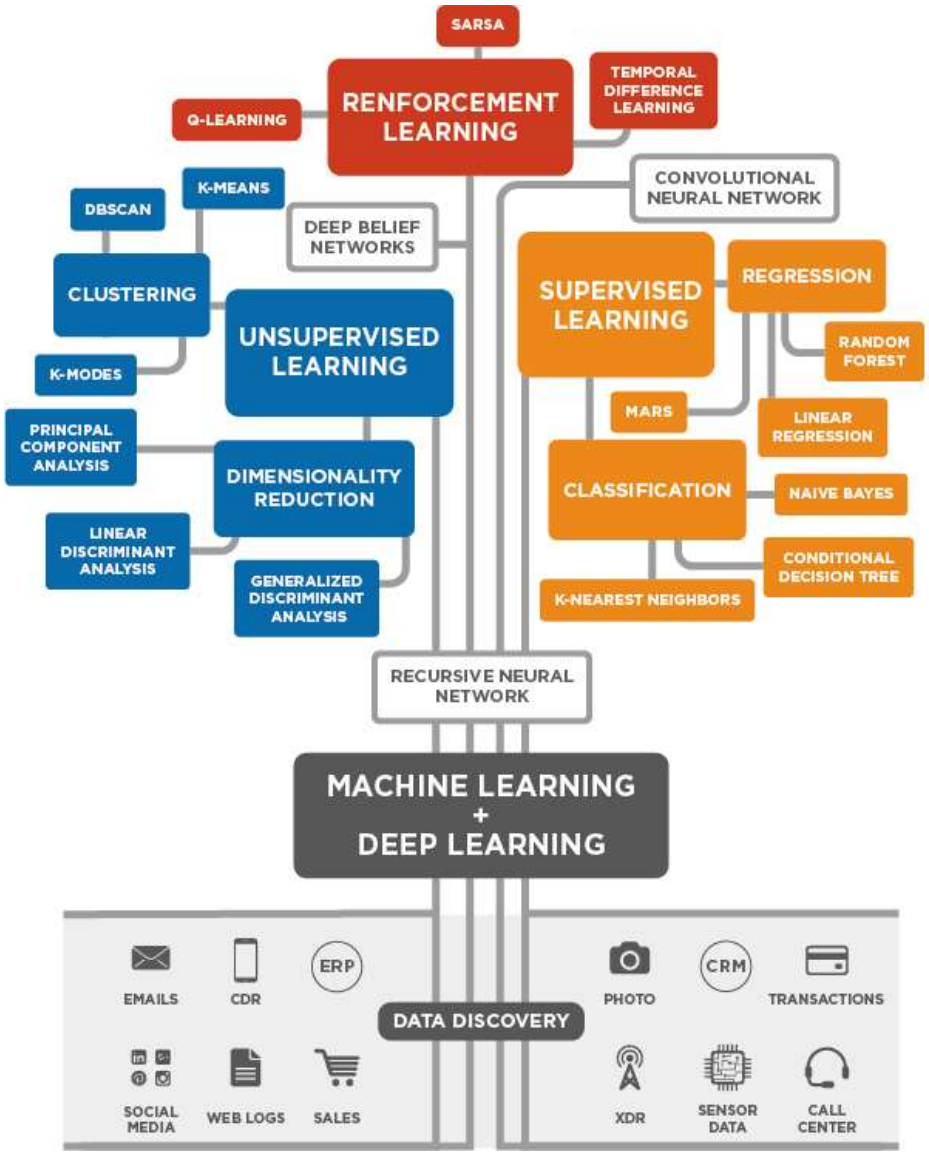
- **Math and Statistics**
- Programming and Database
- Domain Knowledge
- Communication and Visualization



$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2), \quad i = 0, \cdots, K, \quad j = 1, 2, \cdots, n_i,$$

$$\begin{aligned} Y_{ij} &\sim \mathrm{N}(\mu_i, \sigma_1^{-2}) && \text{likelihood,} \\ \mu_i &\sim \mathrm{N}(\eta_{\mu_i}, \sigma_{\mu_i}^{-2}) I(\mu_{i-1}, \mu_{i+1}) && \text{prior,} \end{aligned}$$

Numerous (New) Algorithms



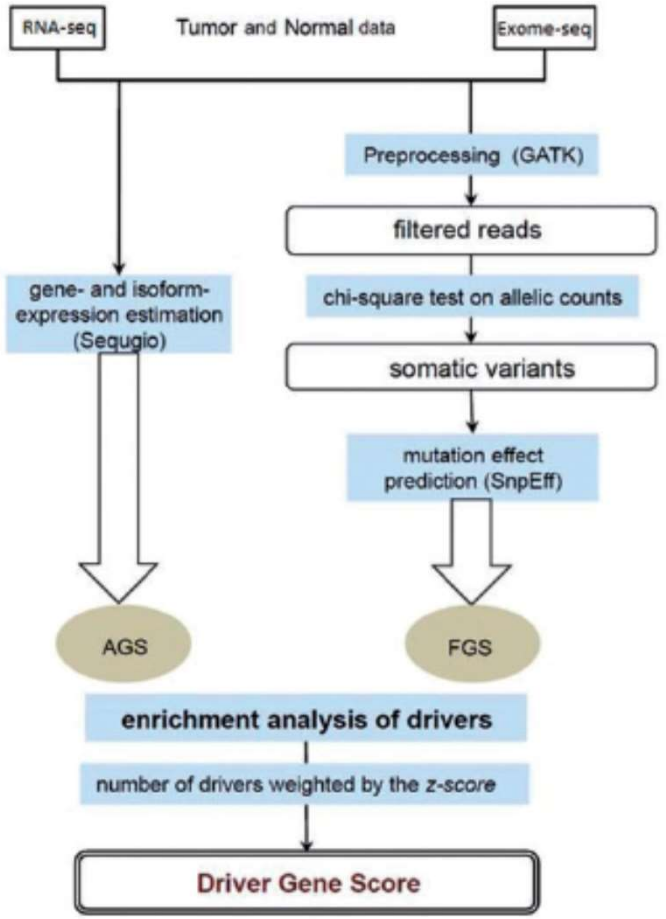
Numerous (New) Algorithms

Classification algorithms considered in the benchmarking study.

	BM selection	Classification algorithm	Acronym
Individual classifier	n.a.	Bayesian Network	B-Net
		CART	CART
		Extreme learning machine	ELM
		Kernalized ELM	ELM-K
		k-nearest neighbor	kNN
		J4.8	J4.8
		Linear discriminant analysis[3]	LDA
		Linear support vector machine	SVM-L
		Logistic regression[2]	LR
		Multilayer perceptron artificial neural network	ANN
		Naive Bayes	NB
		Quadratic discriminant analysis[4]	QDA
		Radial basis function neural network	RbfNN
		Regularized logistic regression	LR-R
		SVM with radial basis kernel function	SVM- Rbf
		Voted perceptron	VP
Classification models from individual classifiers			**16**
Homogenous ensembles	n.a.	Alternating decision tree	ADT
		Bagged decision trees	Bag
		Bagged MLP	BagNN
		Boosted decision trees	Boost
		Logistic model tree	LMT
		Random forest	RF
		Rotation forest	RotFor
		Stochastic gradient boosting	SGB
Heterogeneous ensembles	n.a.	Simple average ensemble	AvgS
		Weighted average ensemble	AvgW
		Stacking	Stack
	Static direct	Complementary measure	CompM
		Ensemble pruning via reinforcement learning	EPVRL
		GASEN	GASEN
		Hill-climbing ensemble selection	HCES
		HCES with bootstrap sampling	HCES-B
		Matching pursuit optimization ensemble	MPOE
		Top-*T* ensemble	Top-*T*
	Static indirect	Clustering using compound error	CuCE
		k-Means clustering	k-Mean
		Kappa pruning	KaPru
		Margin distance minimization	MDM
		Uncertainty weighted accuracy	UWA
	Dynamic	Probabilistic model for classifier competence	PMCC
		k-nearest oracle	kNORA
Classification models from heterogeneous ensembles			**17**

Data Scientist Roles

- Math And Statistics
- **Programming and Database**
- Domain Knowledge
- Communication and Visualization



Vu et. al, 2016

Technical Skills

R	Apache Spark	Apache Pig
Python	NoSQL databases	Tableau
Apache Hadoop	Cloud computing	iPython notebooks
MapReduce	D3	GitHub

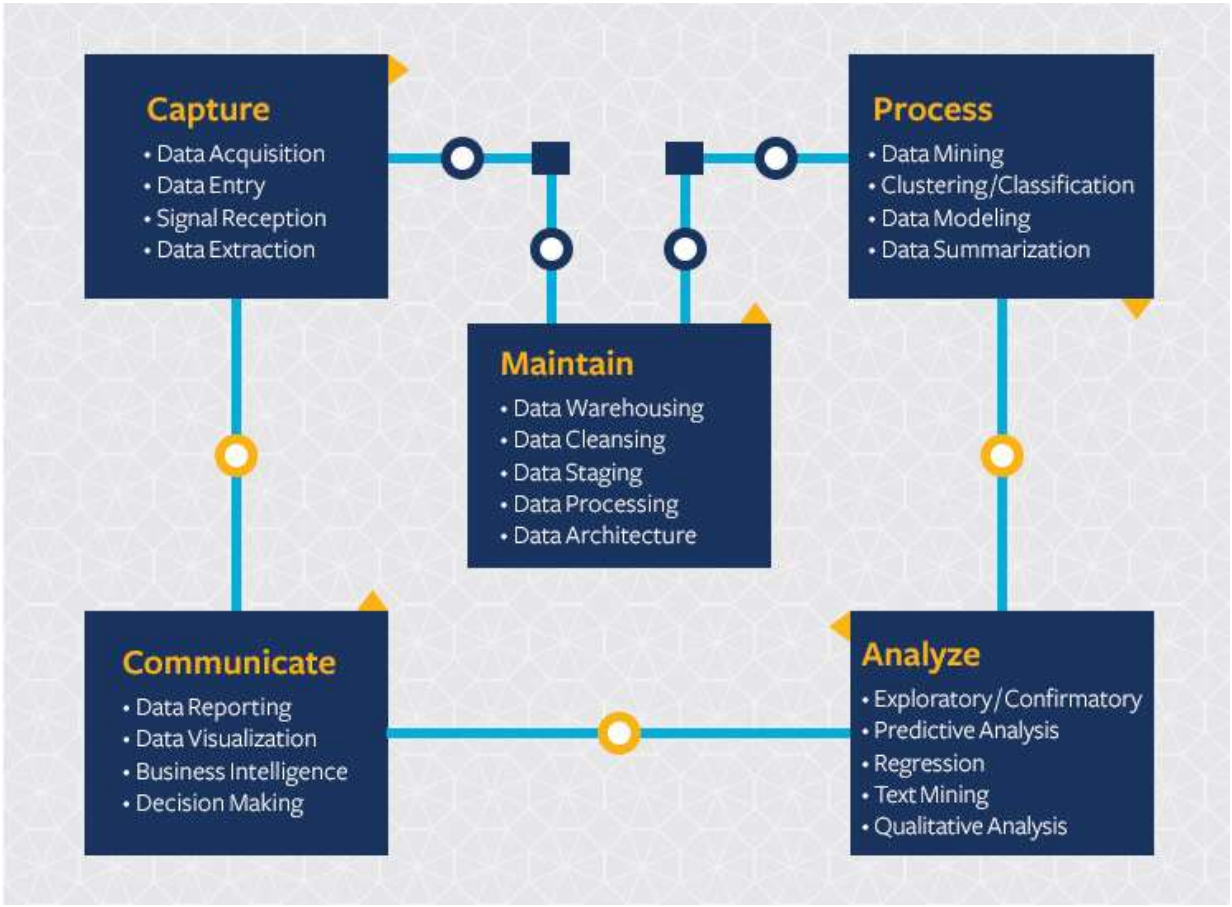
Data Scientist Roles

- Math And Statistics
- Programming and Database
- **Domain Knowledge**
- Communication and Visualization

Data Scientist Roles

- Math And Statistics
- Programming and Database
- Domain Knowledge
- **Communication and Visualization**

Data Science Life Cycle



datascience.berkeley.edu

Which one?

- **Data Scientist**: examine which questions need answering and where to find the related data. Have analytical skills as well as the ability to mine, clean, and present data, then synthesize and communicate to key stakeholders.
  - **Skills needed:** Programming skills (SAS, R, Python), statistical and mathematical skills, storytelling and data visualization, Hadoop, SQL, machine learning
- **Data Analyst**: are responsible for translating technical analysis to qualitative action items and effectively communicating their findings to diverse stakeholders.
  - **Skills needed:** Programming skills (SAS, R, Python), statistical and mathematical skills, data wrangling, data visualization
- **Data engineers** manage exponential amounts of rapidly changing data and focus on the development, deployment, management, and optimization of data pipelines and infrastructure to transform and transfer data to data scientists for querying.
  - **Skills needed:** Programming languages (Java, Scala), NoSQL databases (MongoDB, Cassandra DB), frameworks (Apache Hadoop)

Data Science, What should we learn?

- 70% Statistical Machine Learning (7 weeks)
  - Focus on practical aspects
  - Classes
    - Necessary theoretical background
    - Basic R programming lab
- 20% Big Data Algorithms (2 weeks)
  - Focus on algorithms not on big data technologies
- 10% Data Visualization (1 weeks)
  - Grammar of graphics in R

Data Science, What should we learn?

- Appropriate skills in statistical computing and computational topics relevant to statistics
- Introduce multivariate data early
- Teach about, and with, interactive graphics
- Work with multiple data sources (e.g, Open data etc,.) and types (text, images)
- Teach with, and about, familiar technologies: smartphones, fitness trainers and tablets offer opportunities to provoke statistical thinking
- Use Internet resources to revitalize teaching

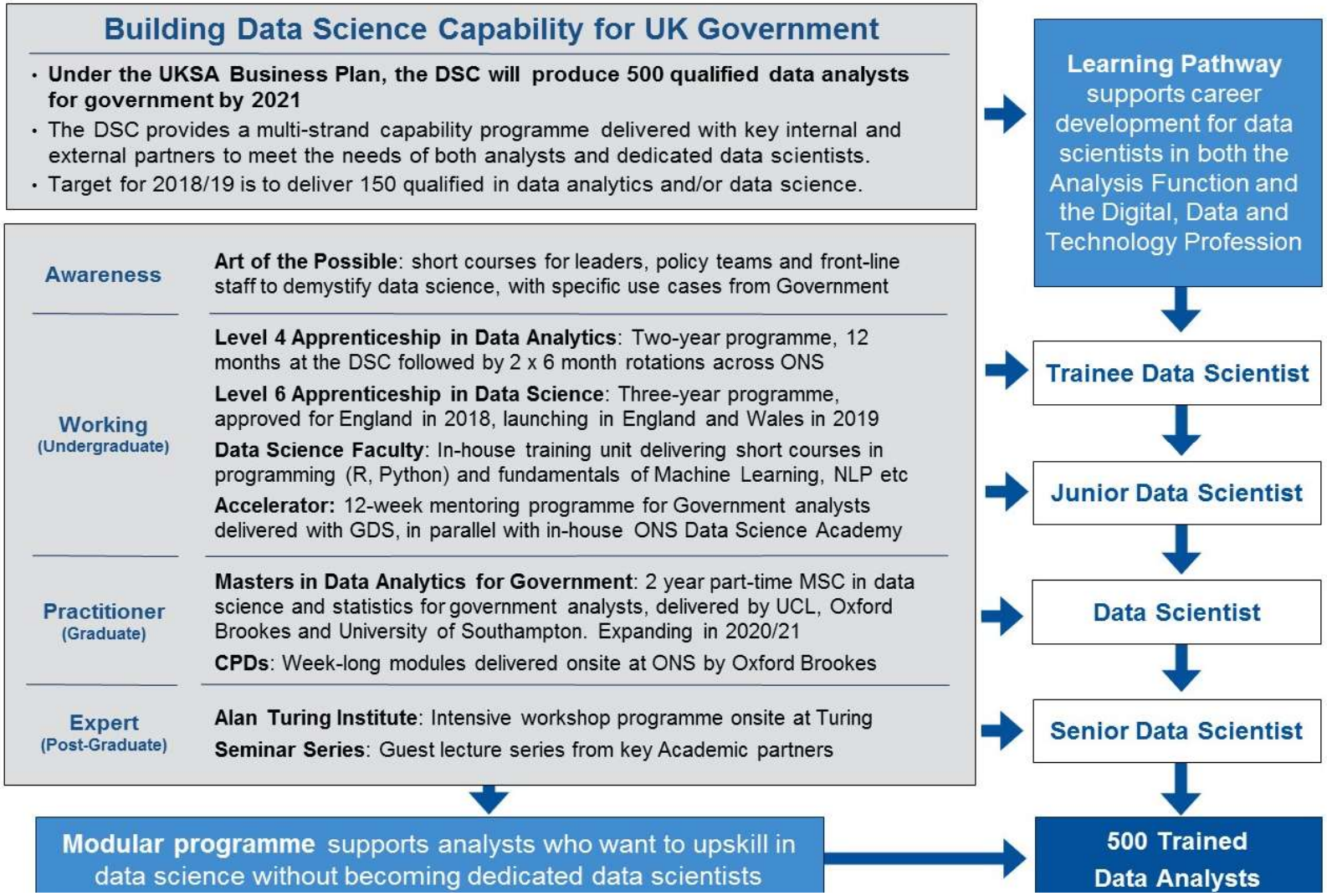
Office for National Statistics

Data Science Campus

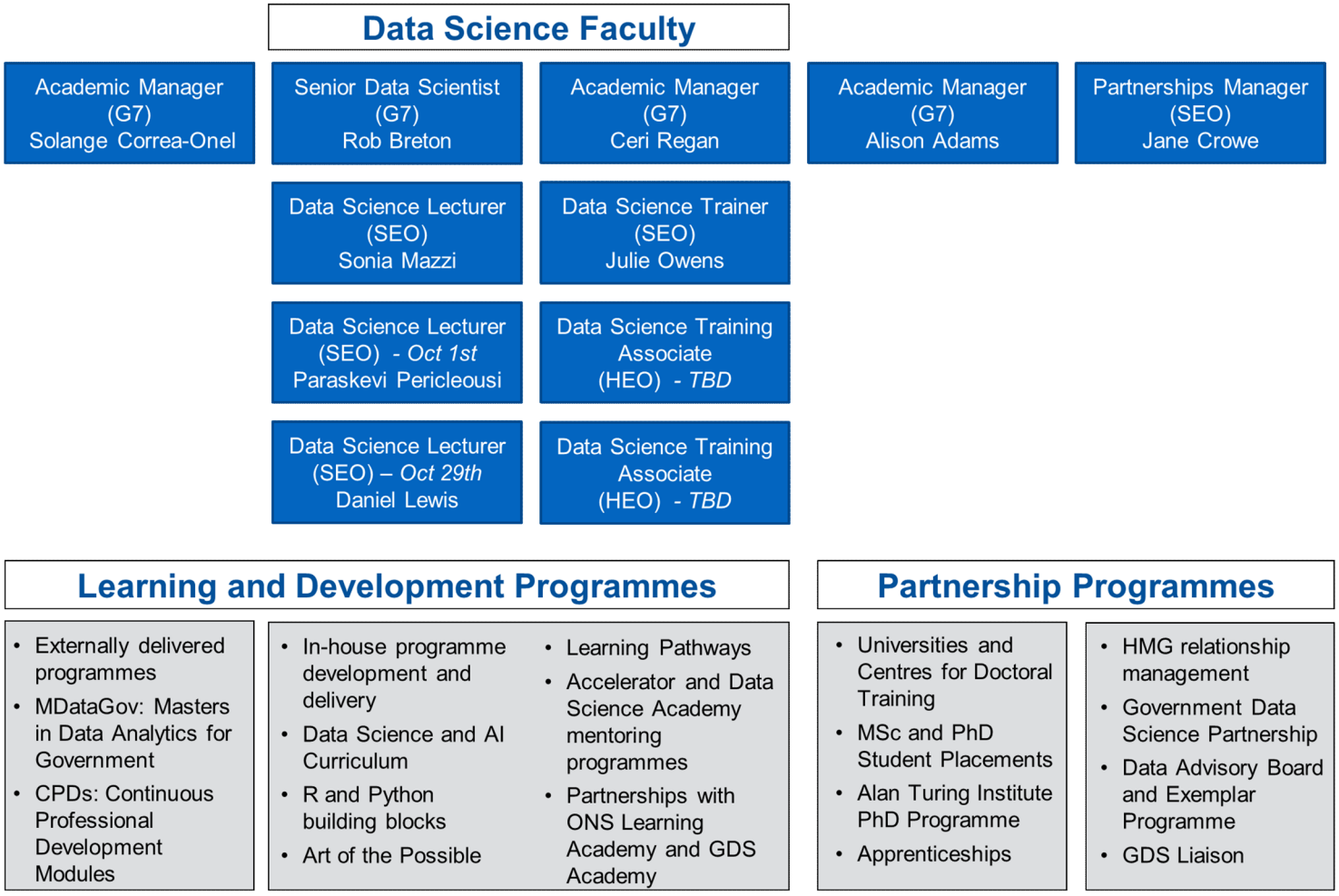
"Data Science for public good"

Data Science	Applying the tools, methods and practices of the digital and data age to create new understanding and improve decision-making
Purpose	We apply data science, and build skills, for public good across the UK and internationally
Mission	We work at the frontier of data science and AI – building skills and applying tools, methods and practices – to create new understanding which improves decision-making for public good

Learning Pathway



Knowledge Exchange



Take Home Message

- Data Size has grown Exponentially and introduce a new type of data
- More Sophisticated Algorithms have been developed
- Computational Power and Storage have been improved
- Challenges in:
  - Extracting Value from Data
  - Get the right talent for Data Scientist and Business translator
- Universities (MSc Statistics and others) have to revolutionize their courses
- Needs Statistical Literacy in all Aspects

Thank you!

setia.pramana@stis.ac.id