WQD7001

Big Data and Data-Drivenness

Prepared by Dr. Salimah M

Learning Objectives:

1. To describe how the **world has changed**.
2. To illustrate the concept of "***datafication***"
3. To explain the phenomena of **big data**.
4. To determine the **value** of big data.
5. To summarize the **strategy of data-driveness**.

THE WORLD HAS CHANGED

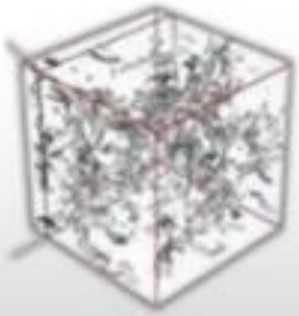
AND CONSTANTLY CHANGING


2005
Luca Bruno /AP


2013
NBC NEWS
Michael Sohn / AP

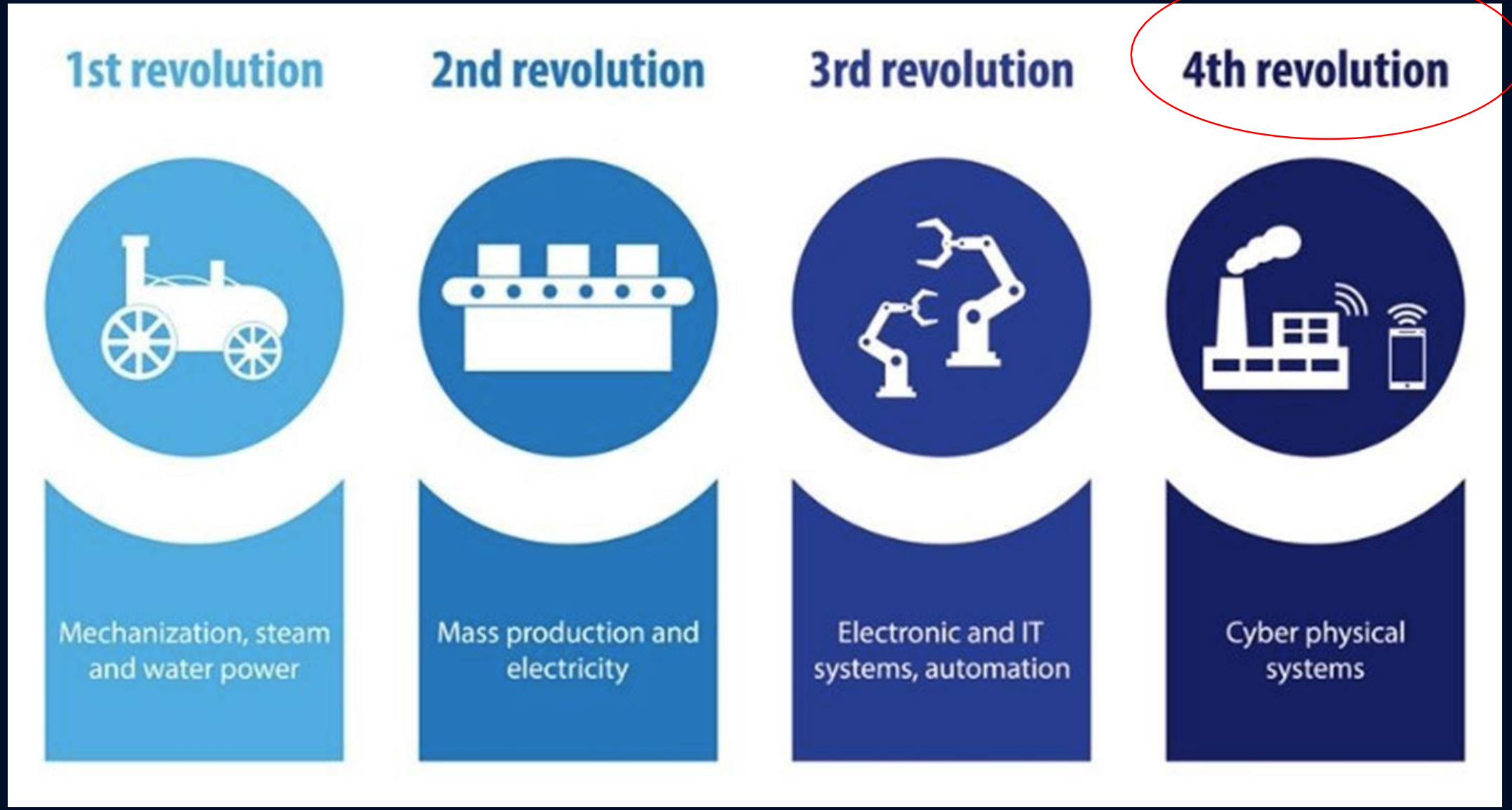
Data Science Landscape

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton's laws, Maxwell's equations...*	*Simulation of complex phenomena*	*Unify theory, experiment and simulation with large multidisciplinary Data*
			Using data exploration

Source: https://www.slideshare.net/TechnetFrance/rec201-mstechdaysfinal130213033305phpapp02-19779391


1st revolution
2nd revolution
3rd revolution
4th revolution
Mechanization, steam and water power
Mass production and electricity
Electronic and IT systems, automation
Cyber physical systems

The 4th Industrial Revolution - *„Industry 4.0“*

Drivers
Quality of life
Engineering Sciences

1st

steam engine

GB

1782

Power generation
Mechanical automation

Mobility

2nd

conveyor belt

US

1913

Industrialization

Mass production
& assembly line

µelectronics

3rd

Computer, NC,PLC

US/EU

1954

Electronic Automation

ICT

4th

Cyber Physical Systems

EU

2015

Smart Automation

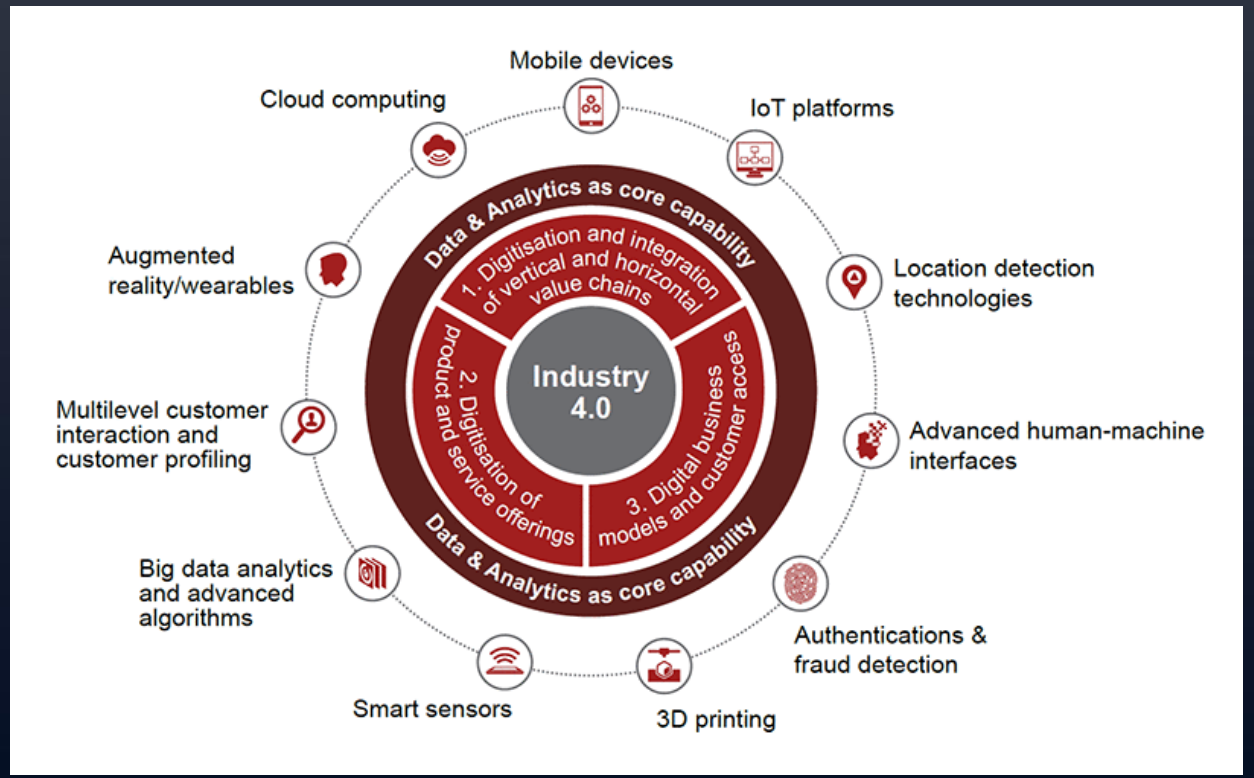
Cyber-physical system

IR 4.0 – Making Sense of the Buzzword

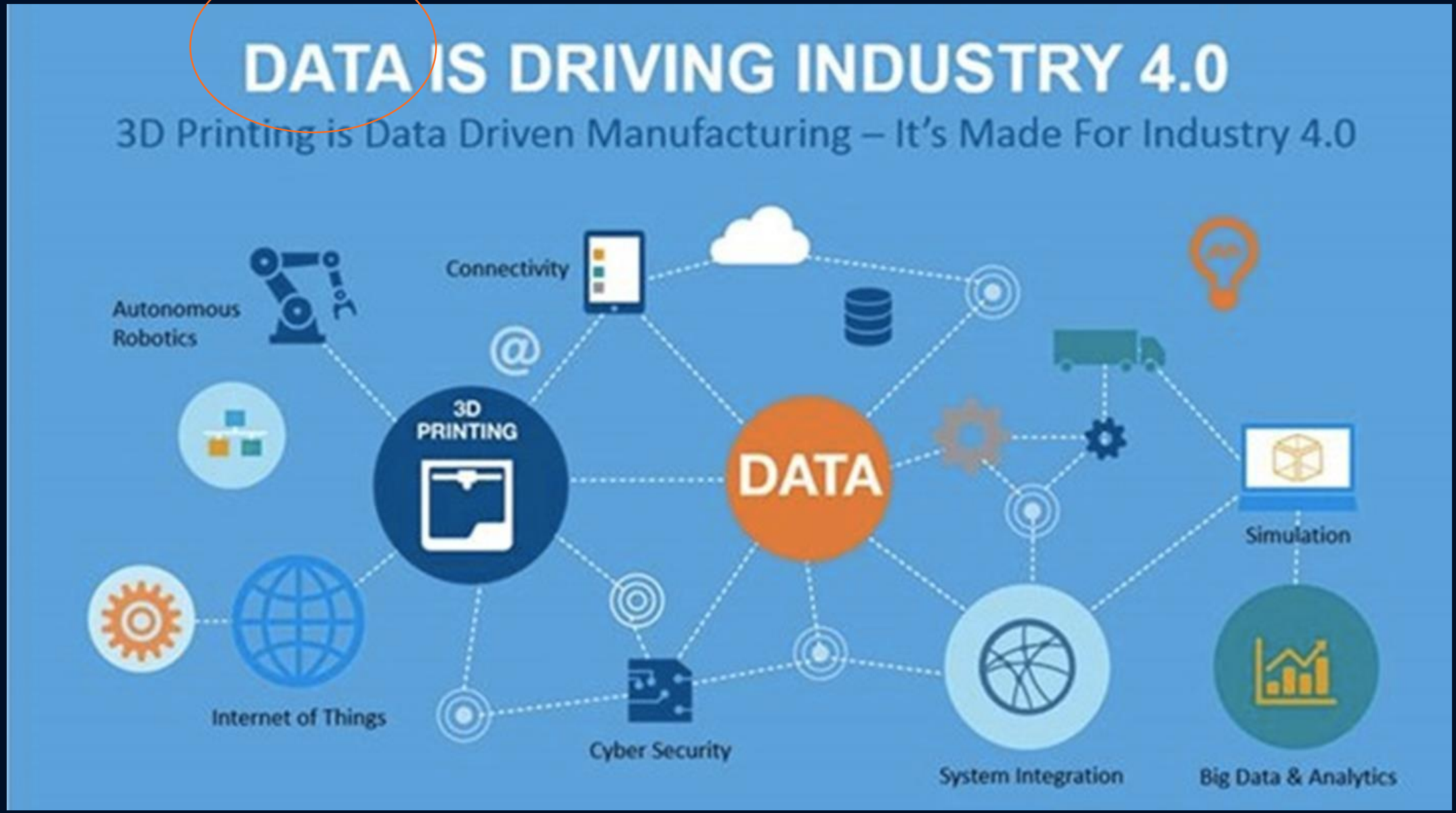
The term "**Industry 4.0**", originates from a project in the high-tech strategy of the
German government, which promotes the computerization of manufacturing.

3 differences between IR 4.0 and the previous ones:

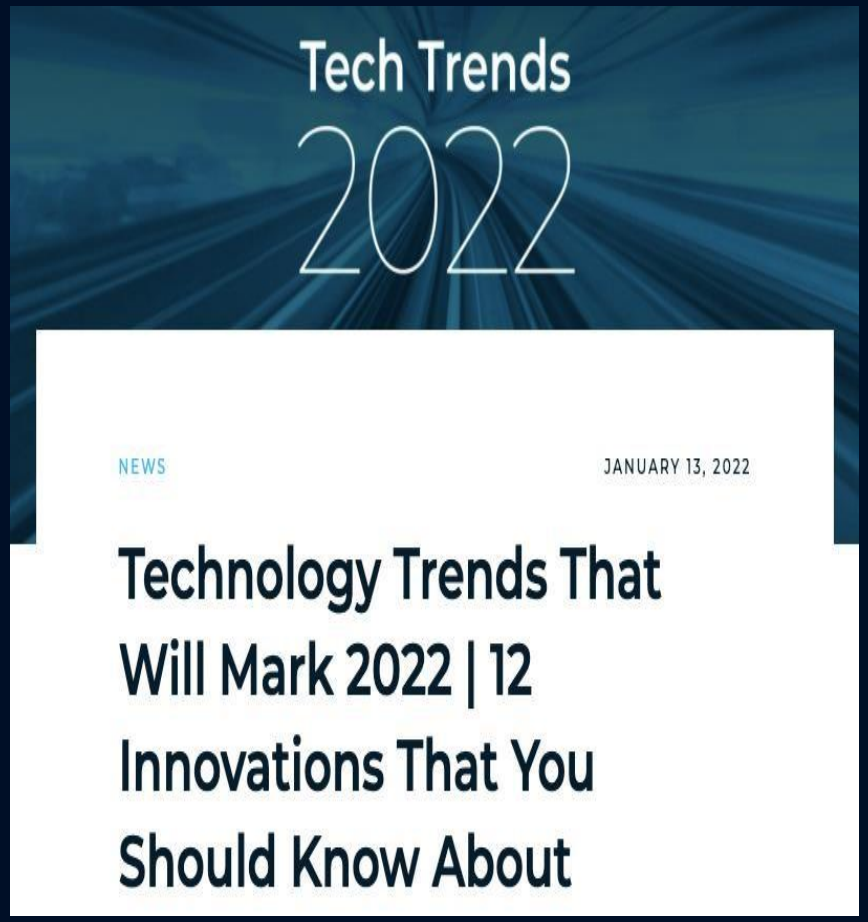
1. **Speed** – this revolution is upon us rapidly
2. **Pervasiveness** – it's about mobile networks, sensors, nanotechnology, brain research, computing, networks, etc. being accessible and affordable.
3. **Entirety** – it is creating the global shared economy.



https://www.thedualarity.com/industrial-revolution-4-0-2/


DATA IS DRIVING INDUSTRY 4.0
3D Printing is Data Driven Manufacturing – It's Made For Industry 4.0
Autonomous Robotics
Connectivity
3D PRINTING
DATA
Simulation
Internet of Things
Cyber Security
System Integration
Big Data & Analytics

https://www.forbes.com/sites/bernardmarr/2022/11/21/the-top-10-tech-trends-in-2023-everyone-must-be-ready-for/?sh=1554b0557df0



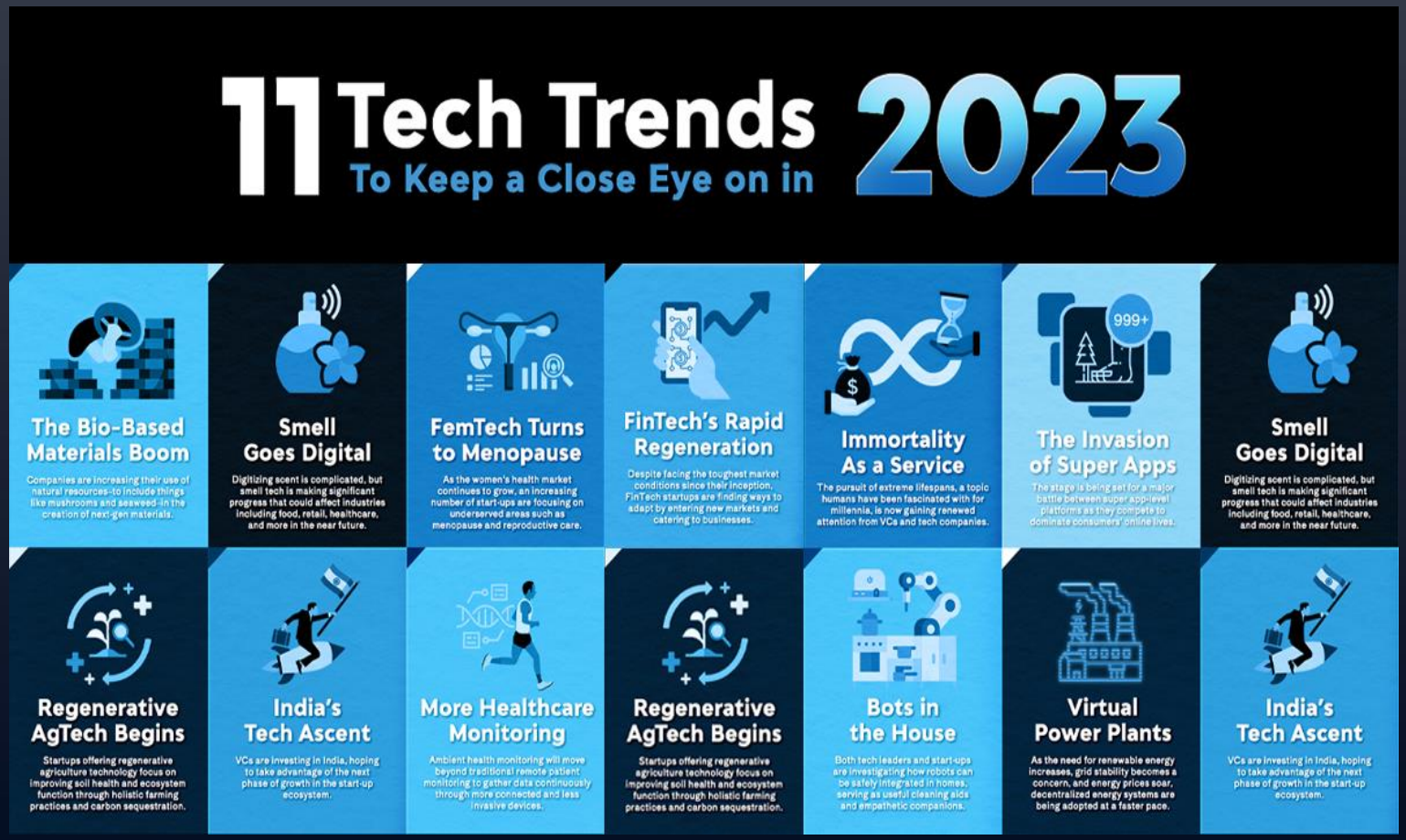
Below, we take a look at the **technology trends that will mark 2022** and that you should keep an eye on.

1 Automation and Hyper-Automation
2 Metaverse
3 New Convolutional Neural Network Architectures: ConvNext
4 Predictive Analytics
5 Edge Computing and Native Clouds
6 Distributed Companies
7 Low-Code Services
8 Decision Intelligence
9 Total Experience (TX)
10 Computer Vision and Pattern Recognition
11 Data Fabric
12 Sustainable Technology Trends
13 NFT (Non-Fungible Tokens) with AI

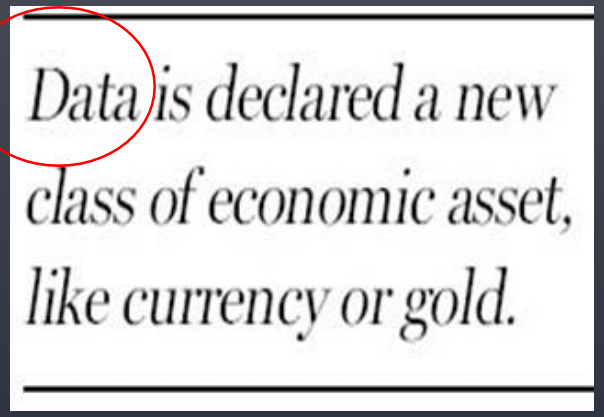
https://www.plainconcepts.com/technology-trends-2022/

2023 :Top Technology & Trends

https://www.forbes.com/sites/bernardmarr/2022/11/21/the-top-10-tech-trends-in-2023-everyone-must-be-ready-for/?sh=1554b0557df0



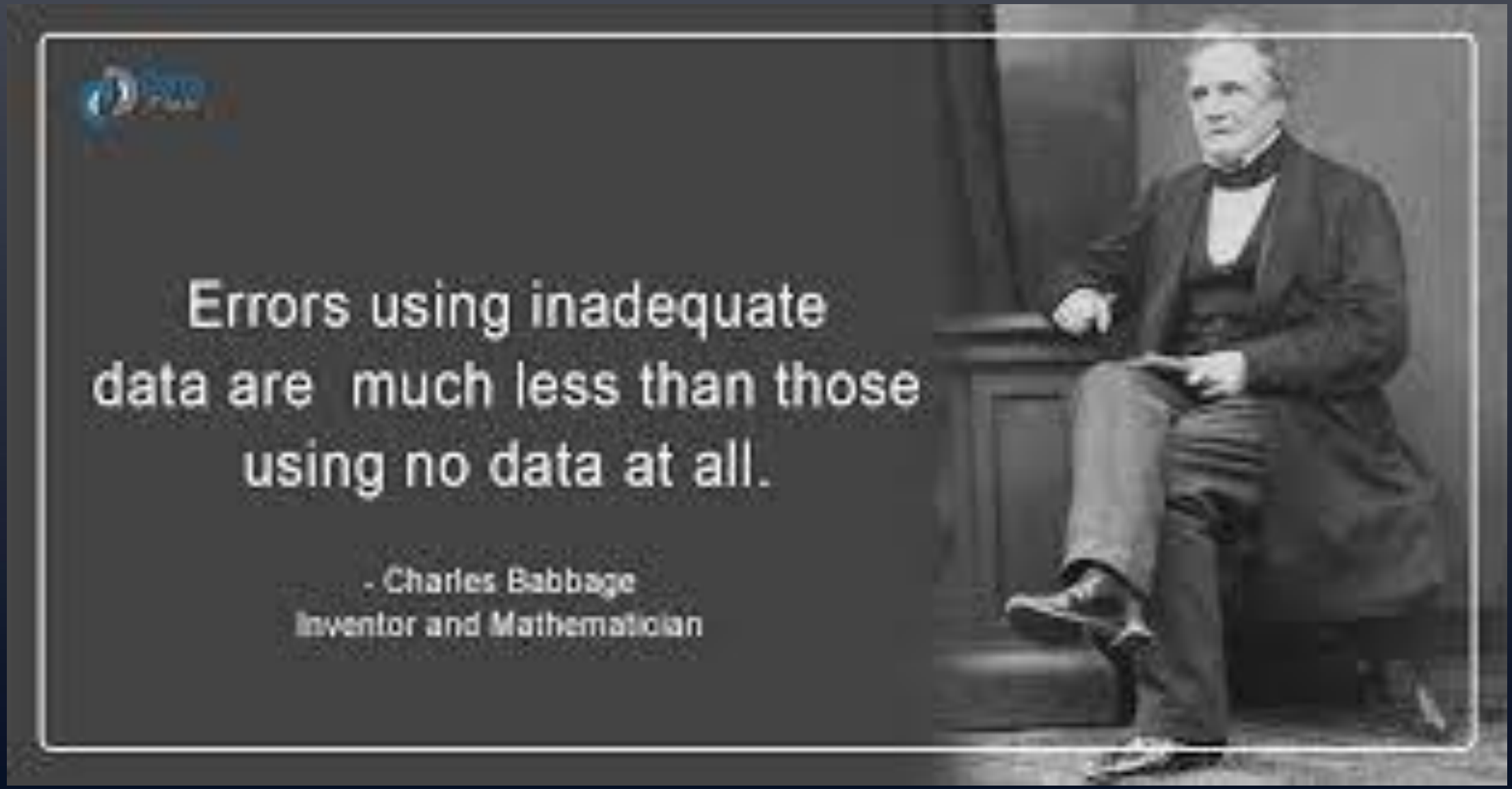

SPECIAL ISSUE
THE PERILS OF LONG COVID
A BOOM IN DIAGNOSTICS
SOCIAL MOVEMENTS AND BACKLASH
SCIENTIFIC AMERICAN
SCIENTIFICAMERICAN.COM
MARCH 2022
HOW COVID CHANGED THE WORLD
Lessons from two years of emergency science, upheaval and loss
INSIDE
• A virus showed the dangers of rugged individualism
• Global health institutions lost trust
• Messenger RNA vaccines opened the door to new therapies
• Conspiracy theories made everything harder
• And more



Data can be sold to monetization or organizations can use it to provide value-added services to their customers.
IT and business leaders regularly state information is their most valuable asset, they fail to value or manage it like one.
Therefore businesses should measure the value of their data to know how much it would value to your organization.

Infonomics - emerging discipline of managing and accounting for information with the same or similar rigor and formality as other traditional assets and liabilities (such as financial, physical and intangible assets and human capital).
Infonomics posits that information itself meets all the criteria of formal company assets, and, although not yet recognized by generally accepted accounting practices (GAAP), it is increasingly incumbent on organizations to behave as if information were a real asset.

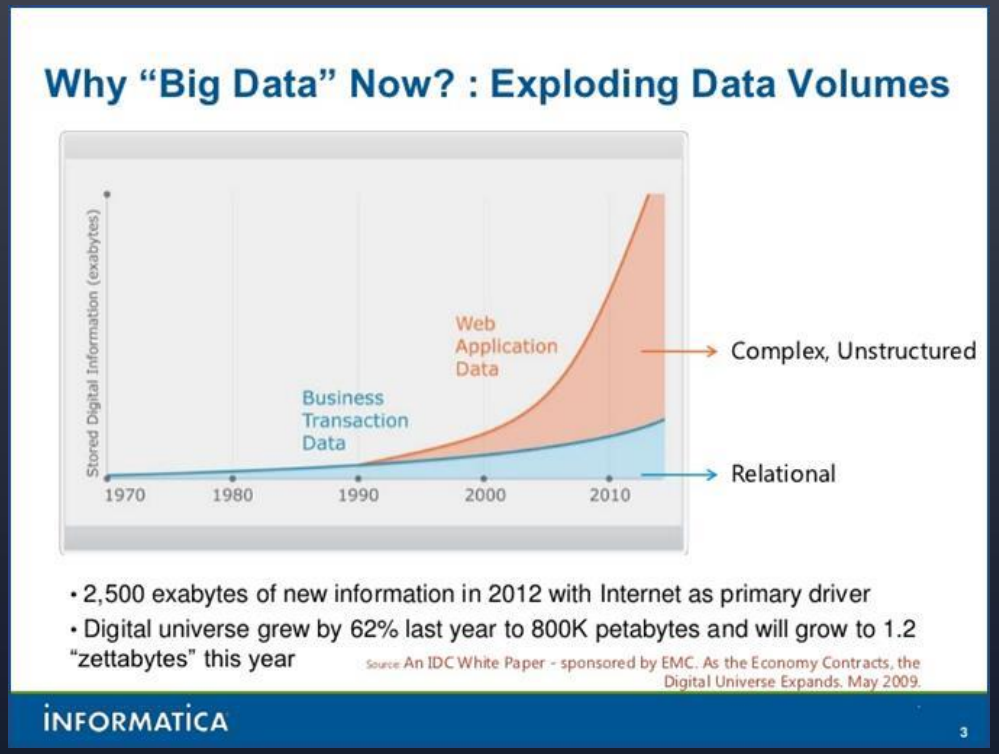
https://research.aimultiple.com/infonomics/


Errors using inadequate
data are much less than those
using no data at all.
- Charles Babbage
Inventor and Mathematician

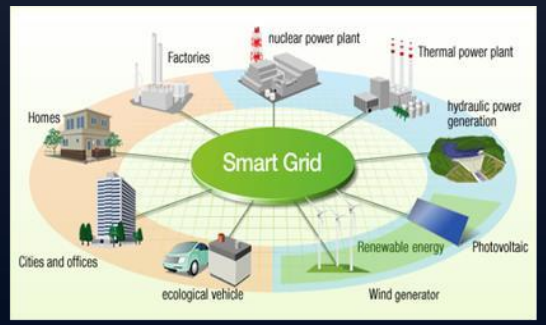
- The Trigger - Data Explosion
  - Defining Big Data
    - Big Data as Research & Scientific Topic
      - Opportunities of Big Data
        - Challenges of Big Data

Data Explosion

Data is everywhere! Take a look at the data that is being produced throughout the world today.



There's a lot of data, but little information!

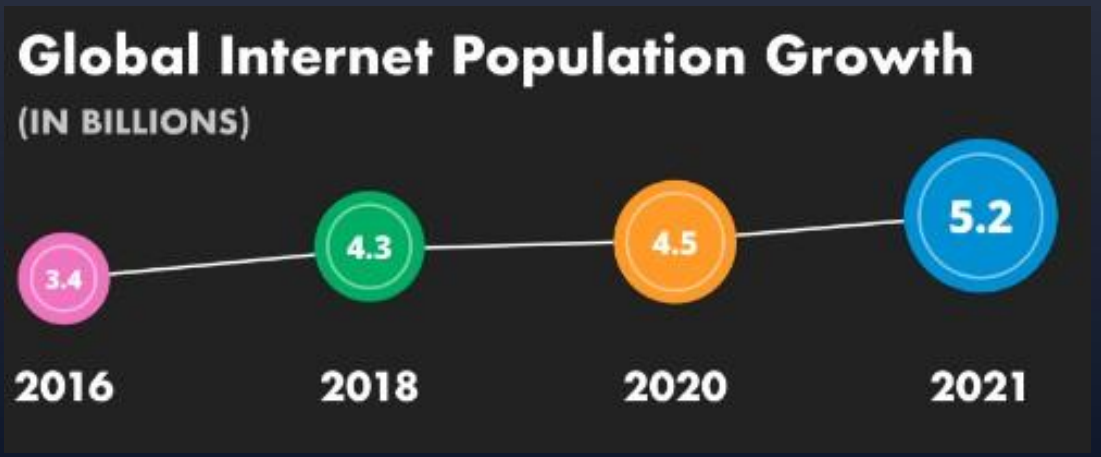


Internet of things (IoT)

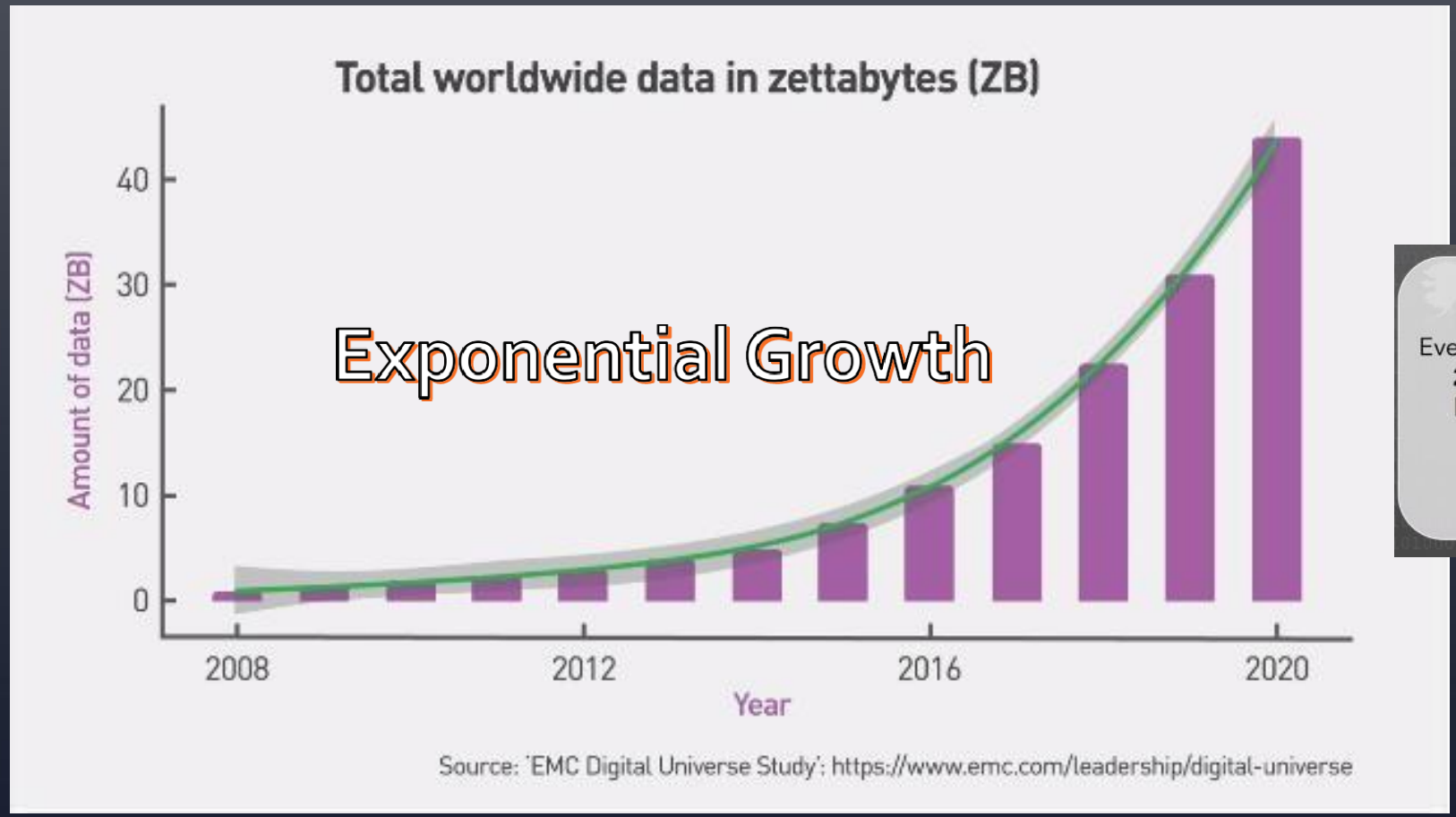
Data Never Sleeps 9.0

How much data is generated *every minute*?

The 2020 pandemic upended everything, from how we engage with each other to how we engage with brands and the digital world. At the same time, it transformed how we eat, how we work and how we entertain ourselves. Data never sleeps and it shows no signs of slowing down. In our 9th edition of the "Data Never Sleeps" infographic, we bring you a glimpse of how much data is created every digital minute in our increasingly data-driven world.



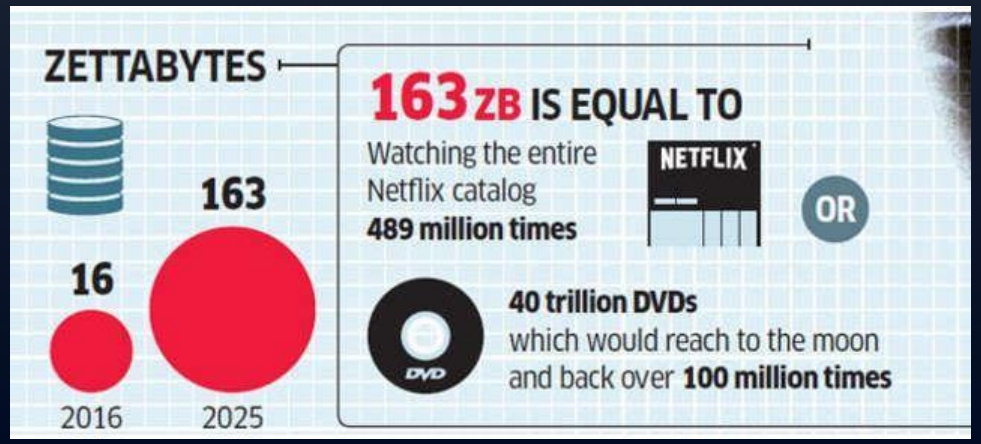
https://www.domo.com/learn/infographic/data-never-sleeps-9

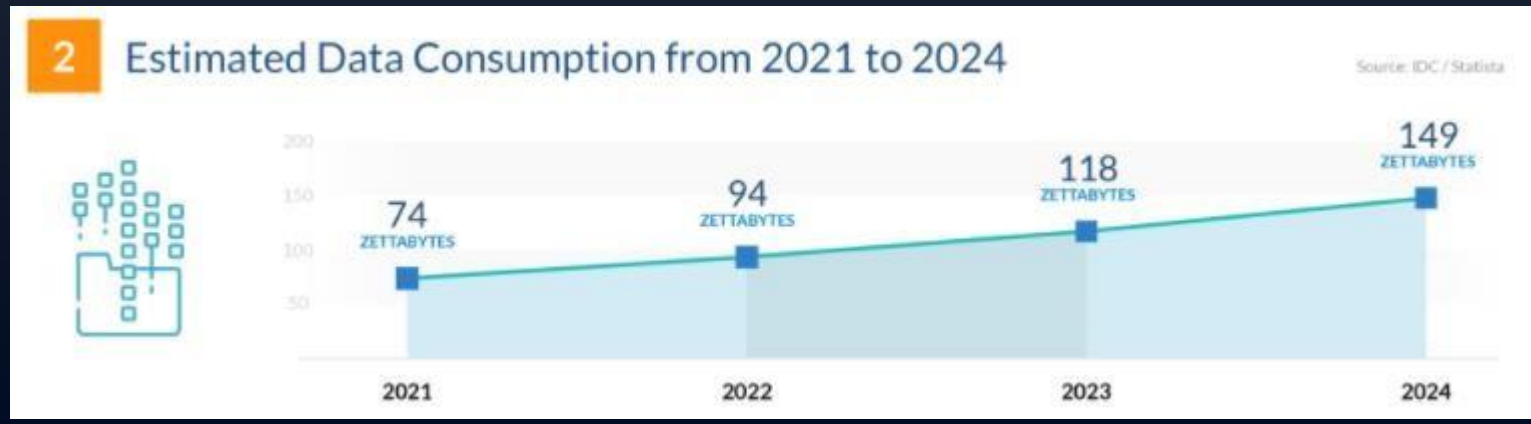


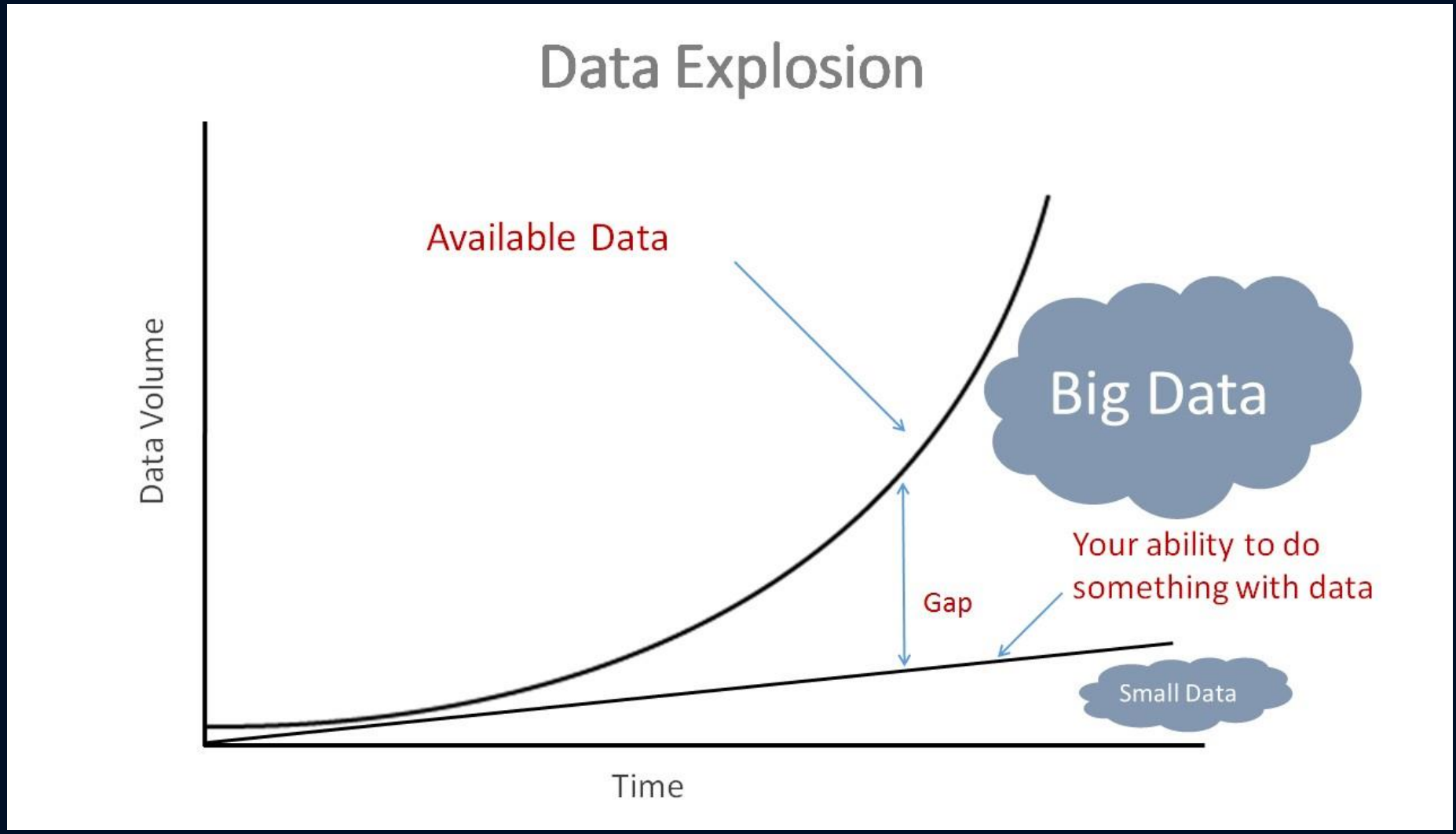
Every day, we create **2.5 quintillion bytes** of data.
- IBM

90% of the worlds data **was created in the past two years.**
- IBM

Of that data, **only 0.05% has been analyzed.**
- Google



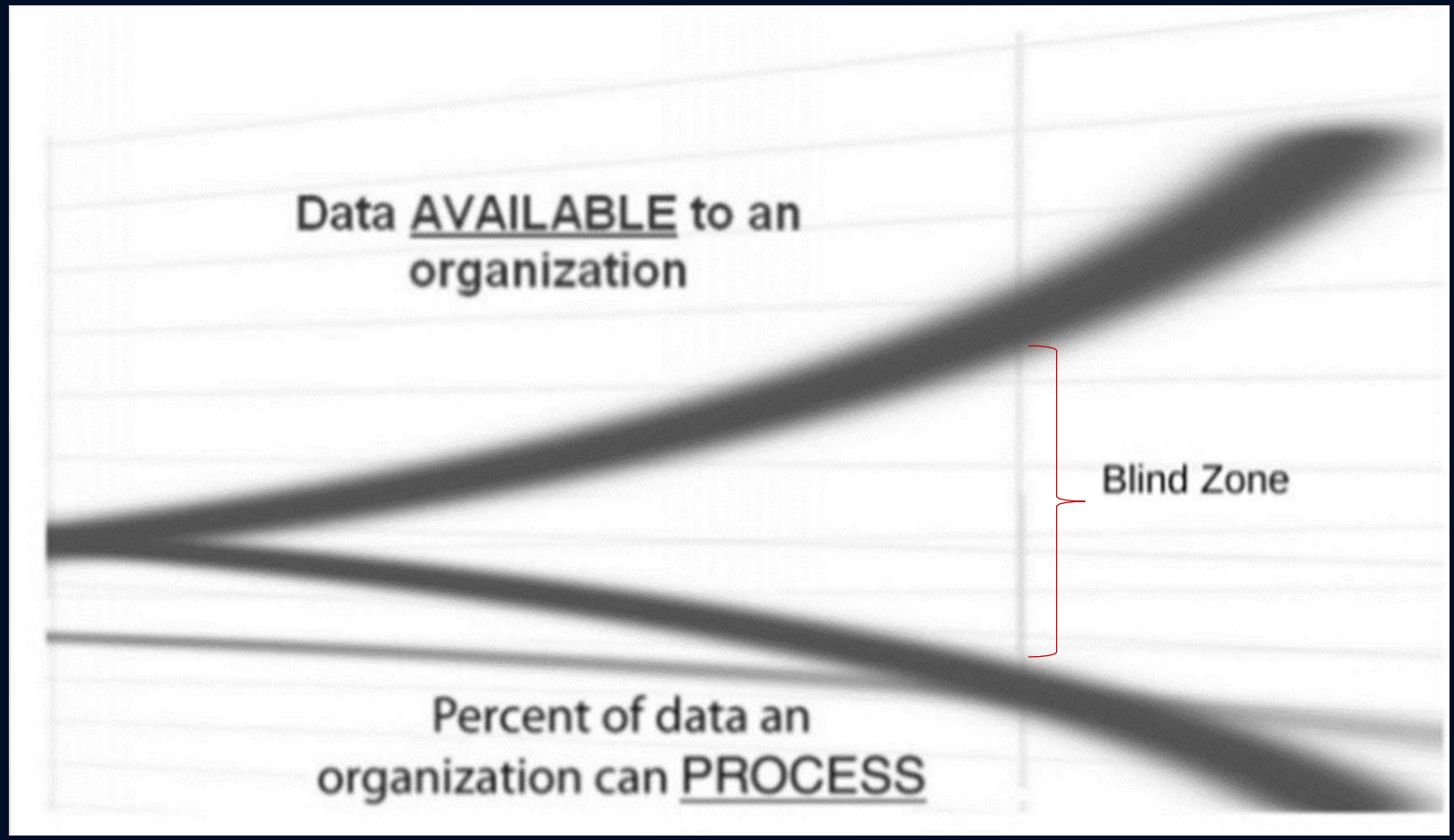




Two key consequences result:

Knowledge Gap: The difference between collecting data and understanding data

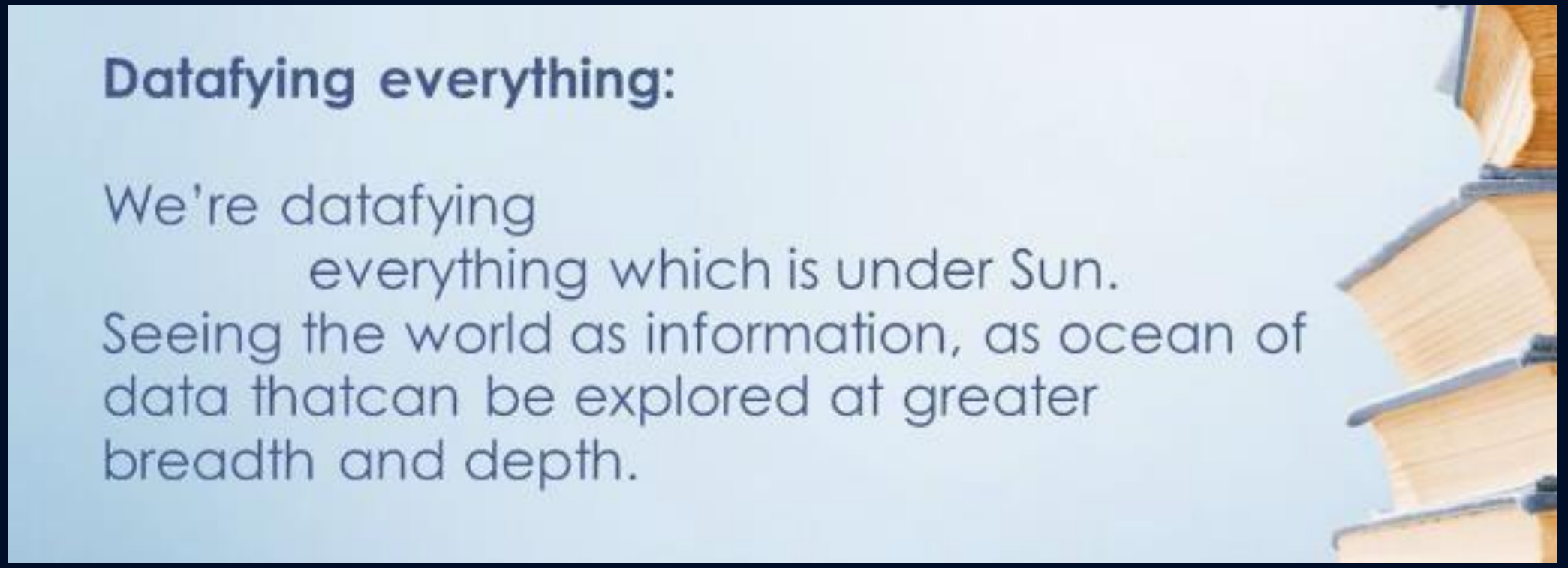
Execution Gap: The difference between understanding data and acting on it



As the amount of data available increased, the % of data organizations can process is decreased. Thus creates a "**blind zone**" – an uncertainty concerning the value of all captured and yet-unexplored data.

Datafication Explained

https://www.youtube.com/watch?v=c4J7-Oiqiy0



Real Challenge: How to extract data from the unlikeliest of places?

Unlikeliest Place

- Case 1: **pathfinder of seas**, Matthew Maury, Us Navy officer, 1855, one of pioneers of Datafication,
- Retrieved information from old logbooks and try a new navigational charts by extracting and tabulating them.
- Case 2: **anti-theft systems in car based on the way somebody is seated**, Professor Koshimizu, took sth. that had never treated as data and transformed it to numerically quantified format and create unique value.

Datafication of Anti-Theft Device

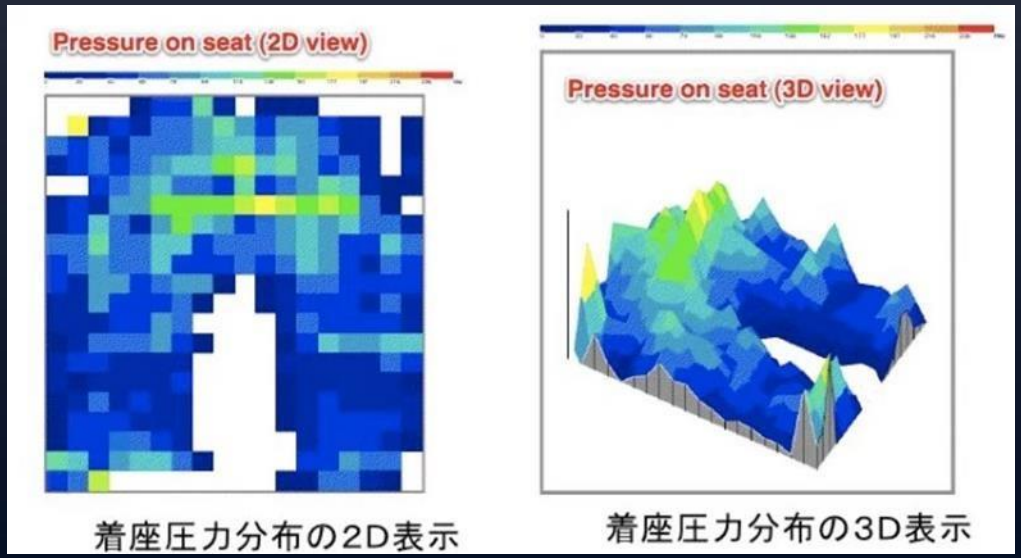
- A Japanese team convert backsides (posture – the way people sit) into data by measuring the pressure at 360 different points from sensors in a car seat and indexing each point on a scale from zero to 256.
- The result is a **digital code** that is unique for each individual.
- This system can distinguish drivers with 98% accuracy, and later this technology is being developed as an **anti-theft device** in cars by recognizing if the driver is the car owner.

- Shigeomi Koshimizu, from the Advanced Institute of Industrial Technology in Tokyo, Japan

To Prevent Theft: Car Seat Identifies Drivers Sitting Down

The trick is that the system measures the pressure people apply on the seat through a set of 360 sensors.

Each sensor is measuring pressure by its own and sends the information to a laptop, which aggregates the information to show key data like the highest value of pressure, area of contact on the seat, and other factors.



The 7 Ways You are totally Unique

1. Backside
2. Body odor
3. Way you walk
4. Ears
5. Skull
6. Fingernails
7. Pores on your nose

Unlikeliest place to explore!

http://www.bbc.com/future/story/20170109-the-seven-ways-you-are-totally-unique

Digitization and Datafication

- The process of converting analog information into zeros (0) and ones (1) so computers can manipulate it --- **digitization**.
- Digitizing makes datafying vastly more efficient and enables mathematical analysis of data to uncover its hidden value.
- Digitization turbocharges datafication.

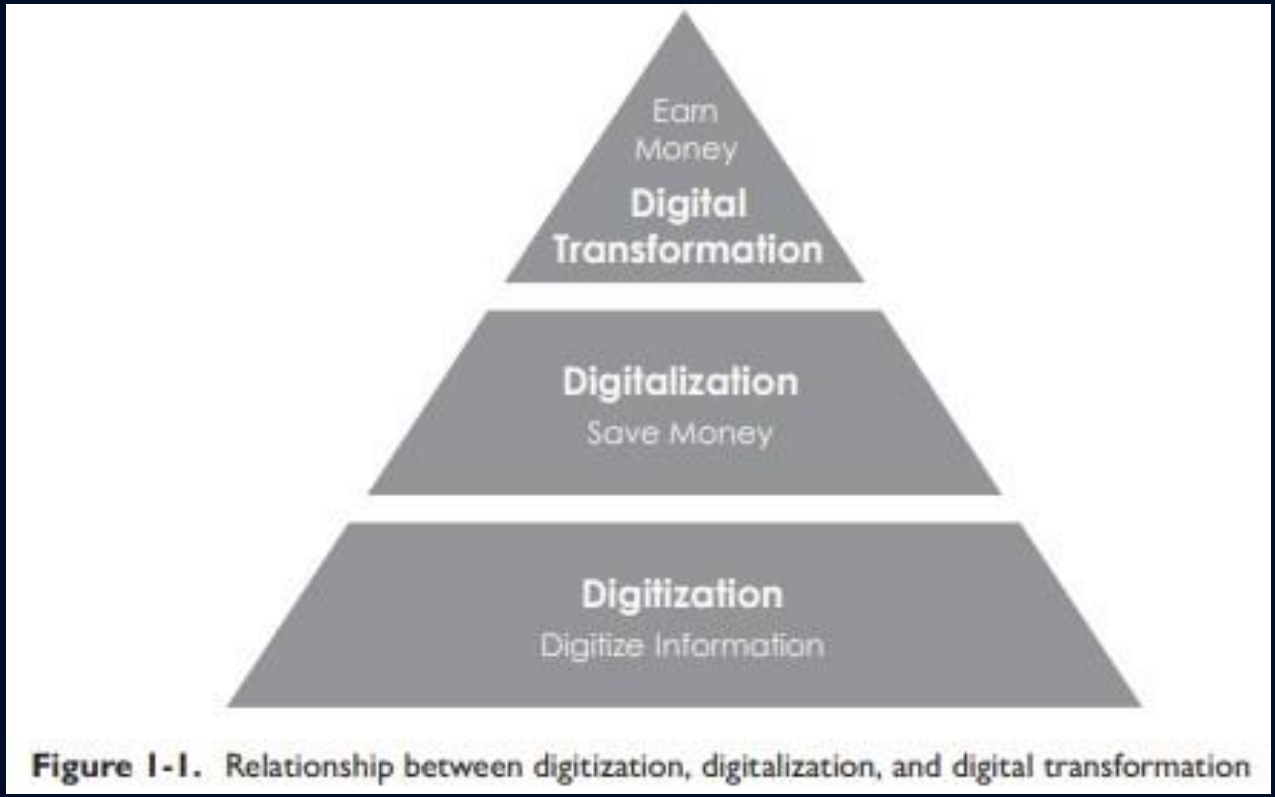
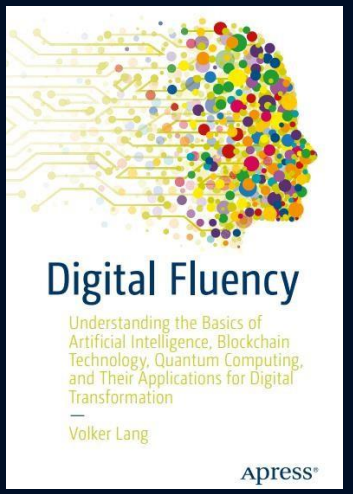


Figure 1-1. Relationship between digitization, digitalization, and digital transformation

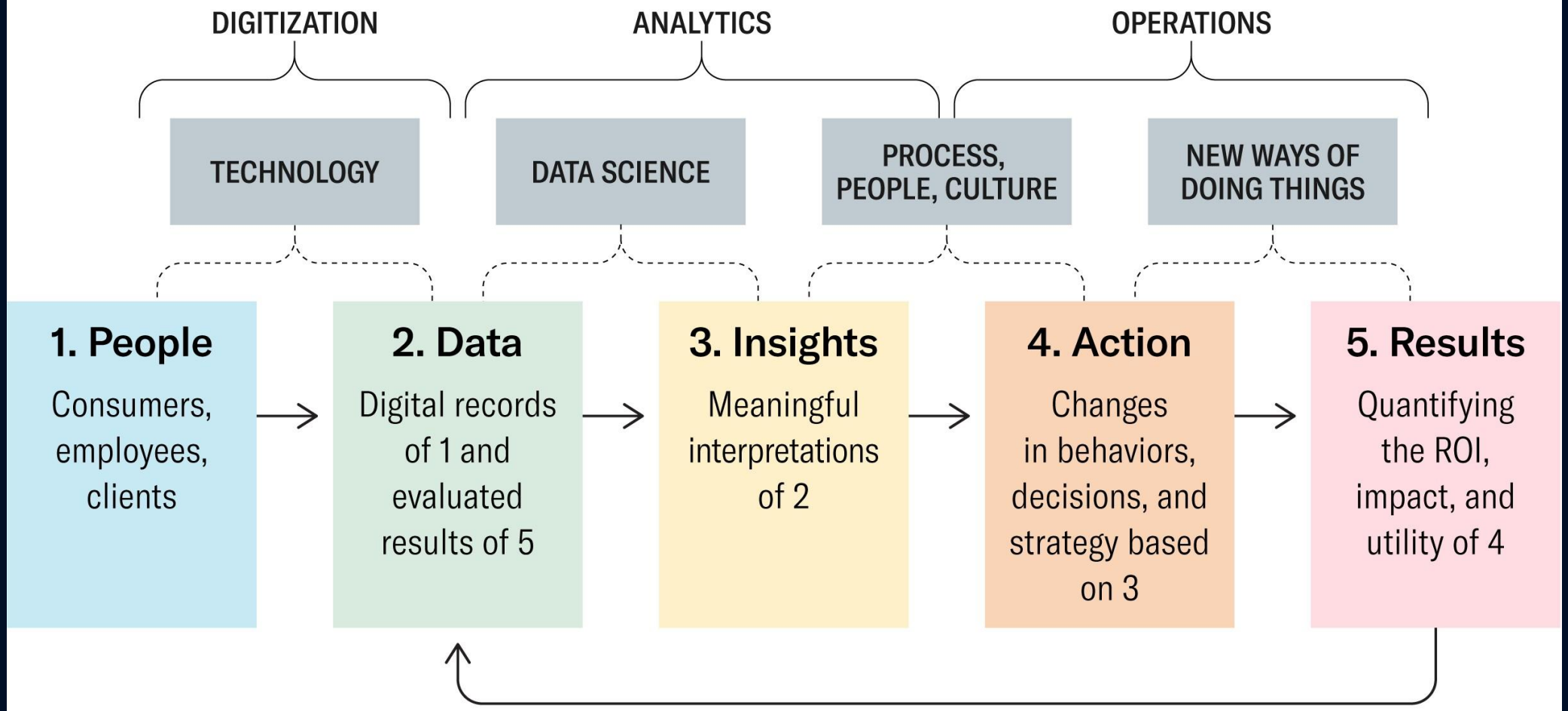


Digitalization is the use of digital technologies to change a business model and provide new revenue and value producing opportunities; it is the process of moving to a digital business.

Digital transformation can refer to anything from IT modernization, to digital optimization, to the invention of new digital business models.

The 5 Essential Components of a Digital Transformation

Mapping the journey to becoming a data-centric organization.



https://hbr.org/2021/11/the-essential-components-of-digital-transformation

Words become Data

- **Book domain** is where both digitization and datafication have happened.
- **Google Book Project** – every page, every book, everyone, every where freely.
- Started in 2004, after 8 years, in 2012 had scanned 20 millions titles which is more than 15% of the world's whole written book from mid-15th century.
- **First step** – Digitization (solid high-resolution digital images)
- **Main step (challenges)** – Datafication (datafied text makes it indexable, searchable, analyzable)

Locations as Data

- Quickly, cheaply and without any specialized instrument we can measure every square inch of **area on Earth**.
- **GPS**, **sensors** and **wireless modules** are used in Google, Microsoft and Apple devices to **trace people** and **object**.
- **Reveal trends** - to predict the problem and correct it before the user realize that there is something wrong.
- **Reality Mining** – processing huge amounts of data from mobile phones to make interpretations and predictions about human behaviors.

Interactions become Data

– Facebook has “datafied” our friend network.
– Google has "datafied" our search and information retrieval.
– LinkedIn has "datafied" our long past professional experiences & connections.
– Twitter is "datafying" emotions and thoughts.
– Waze is "datafying" our driving.

- **Datafication** is a resource and a tool. It is meant **to inform**, rather than explain; it points toward **understanding**.

According to Cukier-Mayer-Schoenberger; the datafication revolution consists of three things:
1. *Collecting and using a lot of data rather than small samples.*
2. *Accepting messiness in your data.*
3. *Giving up on knowing the causes.*

https://www.youtube.com/watch?v=FUj9Ug5kGHM

Other industries where *datafication* process is actively used:

- Insurance: Data used to update risk profile development and business models.
- Banking: Data used to establish trustworthiness and likelihood of a person paying back a loan.
- Human resources: Data used to identify e.g. employees risk-taking profiles.
- Hiring and recruitment: Data used to replace personality tests.
- Social science research: Datafication replaces sampling techniques and restructures the manner in which social science research is performed.

https://www.datasciencecentral.com/profiles/blogs/the-concept-of-datafication-definition-amp-examples

What is Big Data?

Many businesses are only just waking up to the realisation that **data is a valuable asset** that they need to protect and exploit.

Big data is a phenomenon resulting from the rapid acceleration in the expanding volume of high velocity, complex, and diverse types of data.

"Big Data is a term that describes **large volumes of high velocity, complex and variable data** that require **advanced techniques and technologies** to enable the capture, storage, distribution, management, and analysis of the information."

No single standard definition !

Challenge: How to derive value through insight from an explosion of data at greater speed, scale and efficiency?

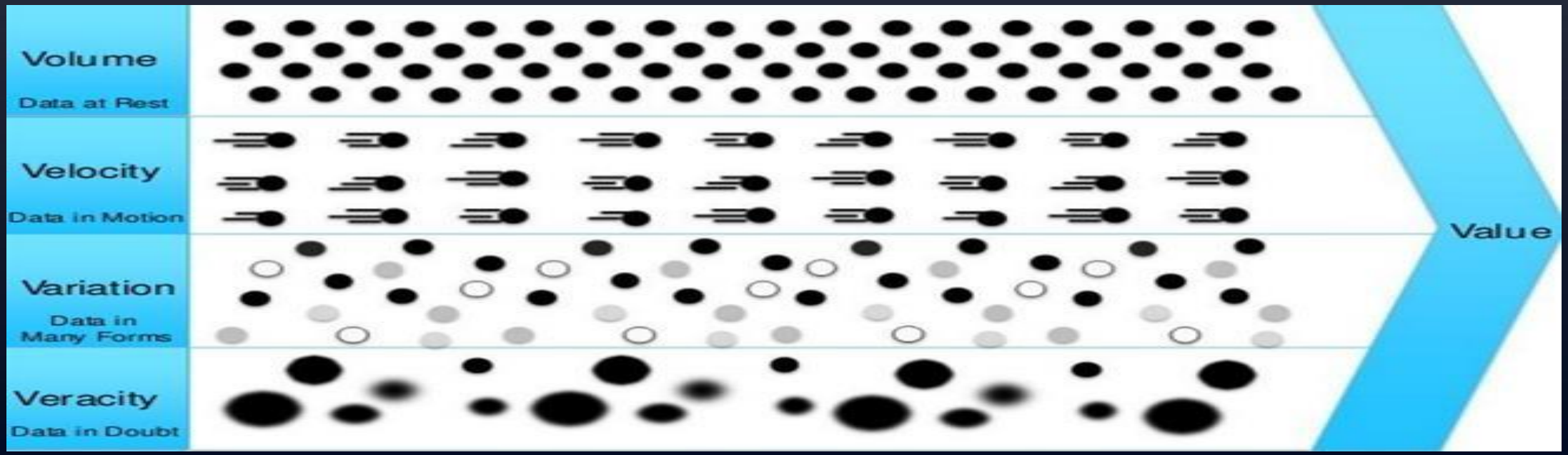
http://datascience.berkeley.edu/what-is-big-data/

Big data is an term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity, and variety.

This term is also typically applied to technologies and strategies to work with this type (size) of data.

The Dimensions of Big Data – 5Vs

Volume: Large volumes of data (i.e Terabytes, Records/Archive, Transactions, Tables, Files).
Velocity: Quickly moving data (Batch, Real / near time, Processes, Streams)
Variety: Structured, unstructured, images, etc.
Veracity: Trust and integrity is a challenge and a must and is important for big data just as for traditional relational DBs.
Value: generate transformative amazing value from big data.

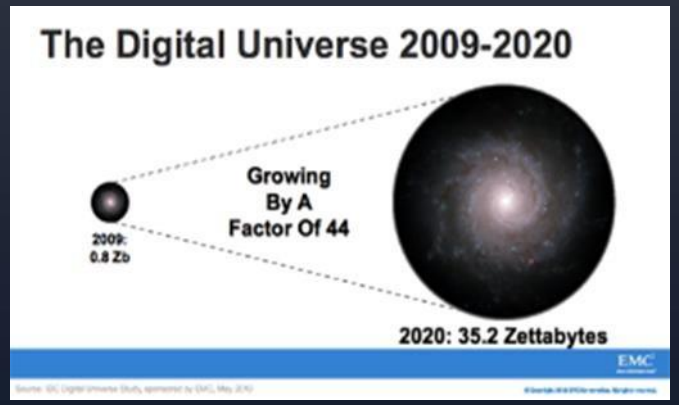


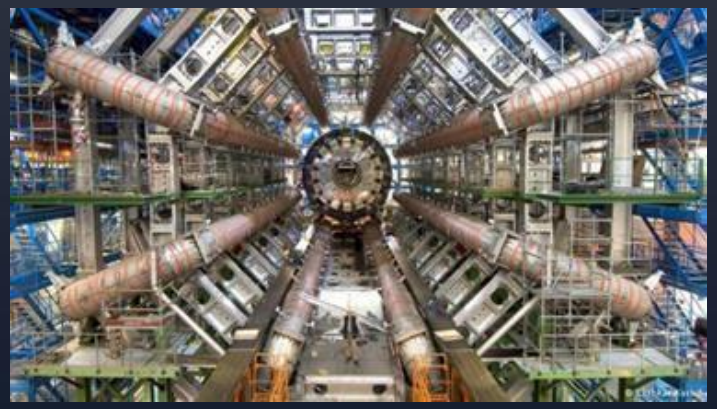
Volume - The Big Data Numbers

Data Volume

- **44x increase from 2009 to 2020**
- **From 0.8 zettabytes to 35zb**

Data volume is increasing exponentially.



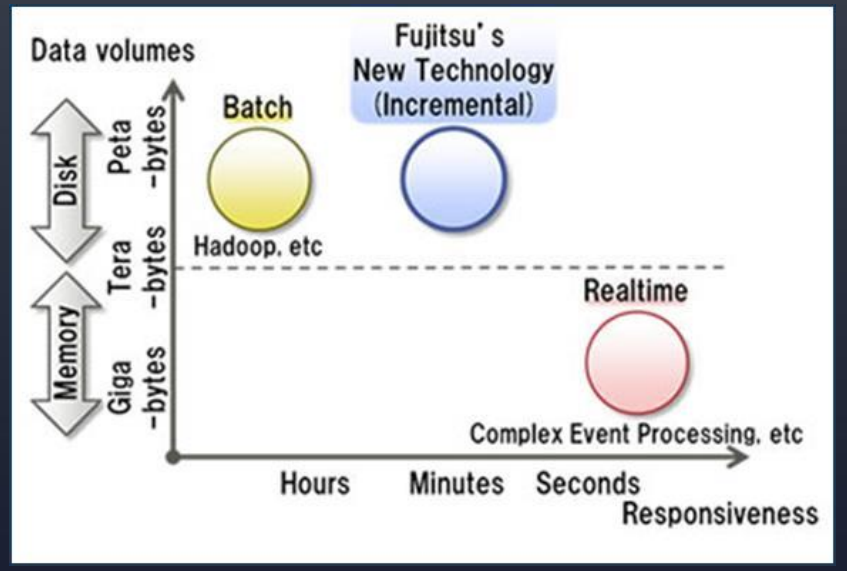


Units of Volume[1]	
1000 Kilobytes	1 Megabyte
1000 Megabytes	1 Gigabyte
1000 Gigabytes	1 Terabyte
1000 Terabytes	1 Petabyte [where most small-to-medium corporations are]
1000 Petabytes	1 Exabyte [where most large corporations are]
1000 Exabyte	1 Zettabyte [where leaders like Facebook and Google are]
1000 Zettabytes	1 Yottabyte
1000 Yottabytes	1 Brontobyte
1000 Brontobytes	1 Geopbyte

(a) CERN,Large Hadron Collider: Generated 15PB of data.
(b) YouTube: 72 hours of Video per hour.
(c) Human Genomics: 7000 PB
(d) Large Synoptic Survey Telescope: 30 TB of Images per day.
(e) Annual Email Traffic (No SPAM): 300 PB +

Velocity (Speed)

Data that is being generated fast, need to be processed fast.
Online Data Analytics
Late decisions → missing opportunities.



Data on time - Accelerating time- to-value from data creation to data usage.

Examples

E-Promotions: Based on your current location, your purchase history, what you like → send promotions right now for store next to you.

Healthcare monitoring: sensors monitoring your activities and body → any abnormal measurements require immediate reaction.

Variety of Data

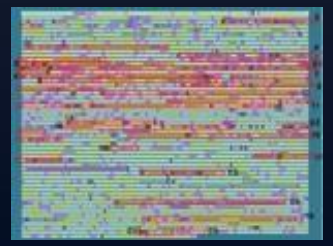
Structured data – neatly organized in databases.

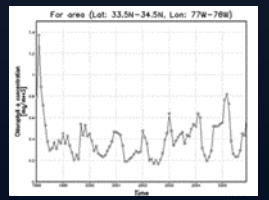
- Could be translated into strings of 1 and 0 capable of being recorded, stored, searched, and analysed.

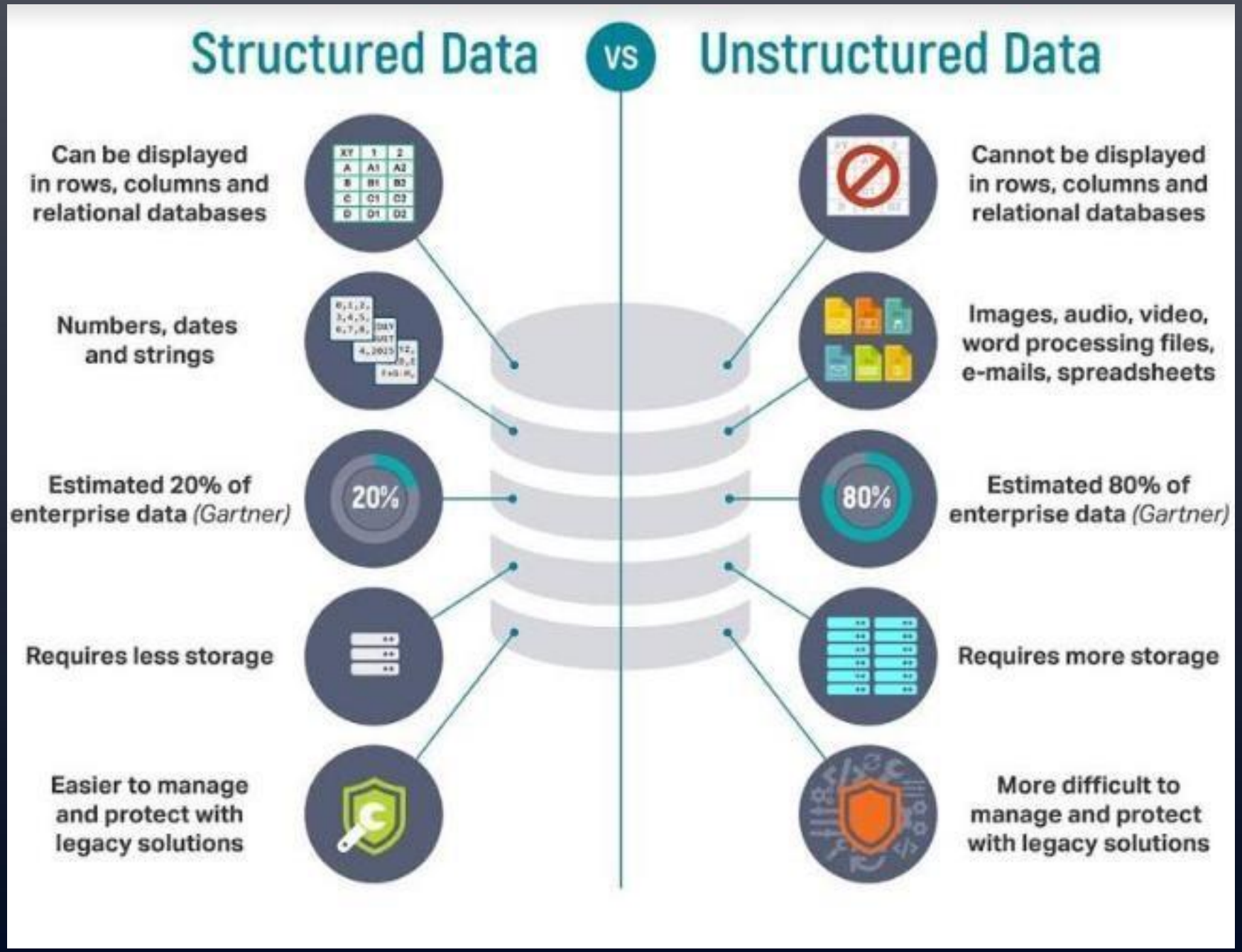
Unstructured data - generated by all our digital interactions, from email to online shopping, text messages to tweets, Facebook updates to YouTube videos.

- The number of gadgets recording and transmitting data, from smart-phones to intelligent fridges, industrial sensors to CCTV cameras, has proliferated globally, leading to an explosion in the volume of data.
- These data sets are now so large and complex that we need new tools and approaches to make the most of them.
- **Data categories** - social media; server logs; Web clickstream; machine/sensor; and geolocation

To extract knowledge➔ all these types of data need to linked together





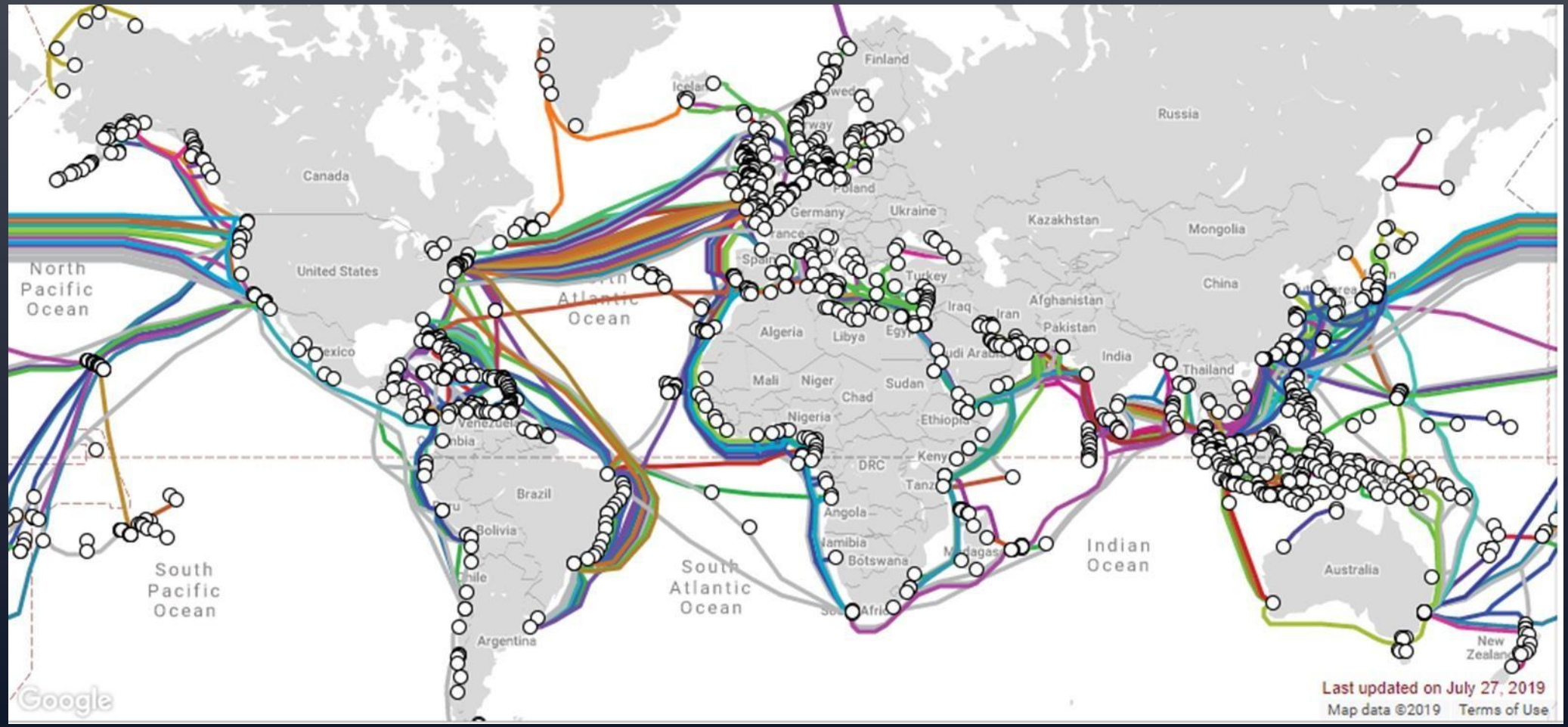

Structured Data
vs
Unstructured Data
Can be displayed in rows, columns and relational databases
Numbers, dates and strings
Estimated 20% of enterprise data (Gartner)
20%
Requires less storage
Easier to manage and protect with legacy solutions
Cannot be displayed in rows, columns and relational databases
Images, audio, video, word processing files, e-mails, spreadsheets
Estimated 80% of enterprise data (Gartner)
80%
Requires more storage
More difficult to manage and protect with legacy solutions

Where Is It All Stored?

- The first computers came with memories measured in kilobytes, but the latest smart-phones can now store 32GB and many laptops now have one terabyte (1,000GB) hard drives as standard. Storage is not really an issue anymore.
- Businesses can either keep all their **data on-site**, in their own **remote data centres**, or farm it out to "**cloud-based**" data storage providers.
- A number of **open source platforms** have grown up specifically to handle these vast amounts of data quickly and efficiently, including Hadoop, MongoDB, Cassandra, and NoSQL.

NoSQL (Not Only SQL) is a distributed database infrastructure that can handle the heavy demands of big data, **Hadoop** is a file system that allows for massively parallel computing.

Submarine Cable Map



https://www.submarinecablemap.com/

Cloud data is stored off-site, and largely travels through submarine cables that are laid at the bottom of the ocean. So the cloud is not in the air as we might think, but underwater.

Digital Transformation is Driving Hybrid Transaction/Analytical Processing (HTAP)

"IMC-enabled HTAP can have a transformational impact on the business." — Gartner 2/17

Traditional Architecture

HTAP Architecture

HTAP enables real-time analytics and situation awareness on live transaction data as opposed to after-the-fact analysis on stale data (in traditional architectures).

Veracity of Data

Data veracity - how accurate or truthful a data set may be.

- When it comes to the accuracy of big data, it's not just the **quality** of the data itself but how **trustworthy** the data source, type, and processing (processing method of the actual data makes sense based on business needs and the output is pertinent to objectives) of it is.
- Removing things like **bias**, **abnormalities** or **inconsistencies**, **duplication**, and **volatility** (rate of change and lifetime of the data) are just a few aspects that factor into improving the accuracy of big data.

Veracity helps to filter through what is important and what is not

Source: shorturl.at/crtBU

Value of Big Data

Data's value shifts from its primary use to its potential future use. Unlike material things, **data's value doesn't diminish** when it is used. "Non-rivalrous" good – one person's use of it does not impede another's.

Amazon can use data from past transactions when making recommendations – and use it repeatedly, not only for the customer who generated the data but for many others as well.

Transformational value

- ✓ Customer delight (beyond customer satisfaction)
- ✓ Competitive advantage
- ✓ World-class risk management
- ✓ Disruptive new business model, e.g. cloud computing
- ✓ Digital disruptive companies

Digital Disruption

Digital disruption is the change that occurs when new digital technologies and business models affect the value proposition of existing goods and services.
a.k.a **Disruptive innovation**

skype ...the world's largest phone company... OWNS NO TELCO INFRASTRUCTURE!

/ Google ...the largest mobile software vendors... WHO DON'T WRITE MOST APPS!

Disruptive Technology

- One that **displaces an established technology** and impacts the industry associated with that technology.
- Or a **groundbreaking product** that creates a completely new industry.
- **Examples**: Personal computer, email, digital cameras, smart phone, social networking, cloud technology, IOT, 3D printing, etc.

Innovation Types

- Sustaining:
  An innovation that does not affect existing markets.
- Evolutionary:
  An innovation that improves a product in an existing market in ways that customers are expecting.
- Revolutionary:
  An innovation that is unexpected, but nevertheless does not affect existing markets.
- Disruptive:
  An innovation that creates a new market by applying a different set of values, which ultimately and unexpectedly overtakes an existing market.

The “Option Value” of Data

The value of data is more like an iceberg; invisible value far exceeds the immediately realizable value.

This means the system takes information generated for one purpose (primary) and reuses it (secondary) for another.

Primary use:

- The car’s battery indicator tells when to fill gas.
- Power grid usage data is for manage the stability of grid.

Secondary use:

- Determine when and where to recharge, and where to build electric vehicle service stations.

Example: https://www.nytimes.com/interactive/2020/03/15/business/economy/coronavirus-worker-risk.html

The “Option Value” of Data (continue)

The crux of data’s worth is its seemingly unlimited potential for reuse: its option value.

Most of data’s value lies in its use, not its mere possession. There are 3 potent ways to unleash data’s option value:

1. basic reuse
2. merging datasets
3. finding “twofers”

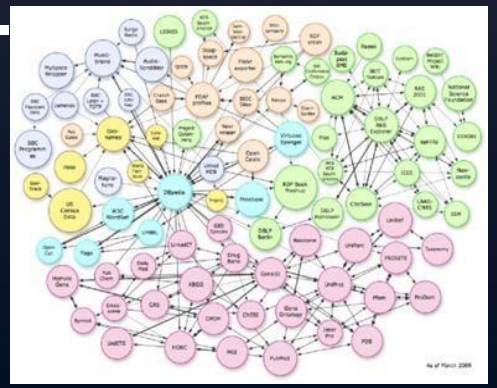
The Reuse of Data

- There is **initial** (**primary**) use of data and subsequent (**secondary**) use of data, i.e. the **reuse of data**.
- The value in data's reuse is good news for organizations that collect or control large datasets but currently make little use of them.
- A **classic example** of data's innovative reuse, is search terms. Old queries can be extraordinarily valuable.
- SWIFT, the global interbank system, offer GDP forecasts based on fund transfer data passing over its network.
- **Captcha**: The data had a primary use—to prove the user was human—but it also had a secondary purpose: to decipher unclear words in digitized texts.

Recombinant of Data

- Combining multiple (and disparate) datasets.
- We can do innovative things by **commingling data** in new ways.
- With big data, the sum is more valuable than its parts, and when we combine the sums of multiple datasets together, the sum too is worth more than its individual ingredients.
- The Danish Cancer Society meshed cell-phone transaction data with socioeconomic and cancer registry files to test the influence of cell- phone use on cancer prevalence in Denmark.

Today Internet users are familiar with more basic "**mashups**," which combine two or data sources in a novel way



Extensible Data

- One way to enable the reuse of data is to design **extensibility** into it from the outset so that it is suitable for **multiple uses**.
- Example: some retailers are positioning store **surveillance cameras** so that they not only spot shoplifters but can also track the flow of customers through the store and where they stop to look.
- That increases the **data's option value**.
- The point is to look for "**twofers**"—where a single dataset can be used in multiple instances if it can be collected in a certain way. Thus the data can do **double duty**.

The Value of Data Exhaust

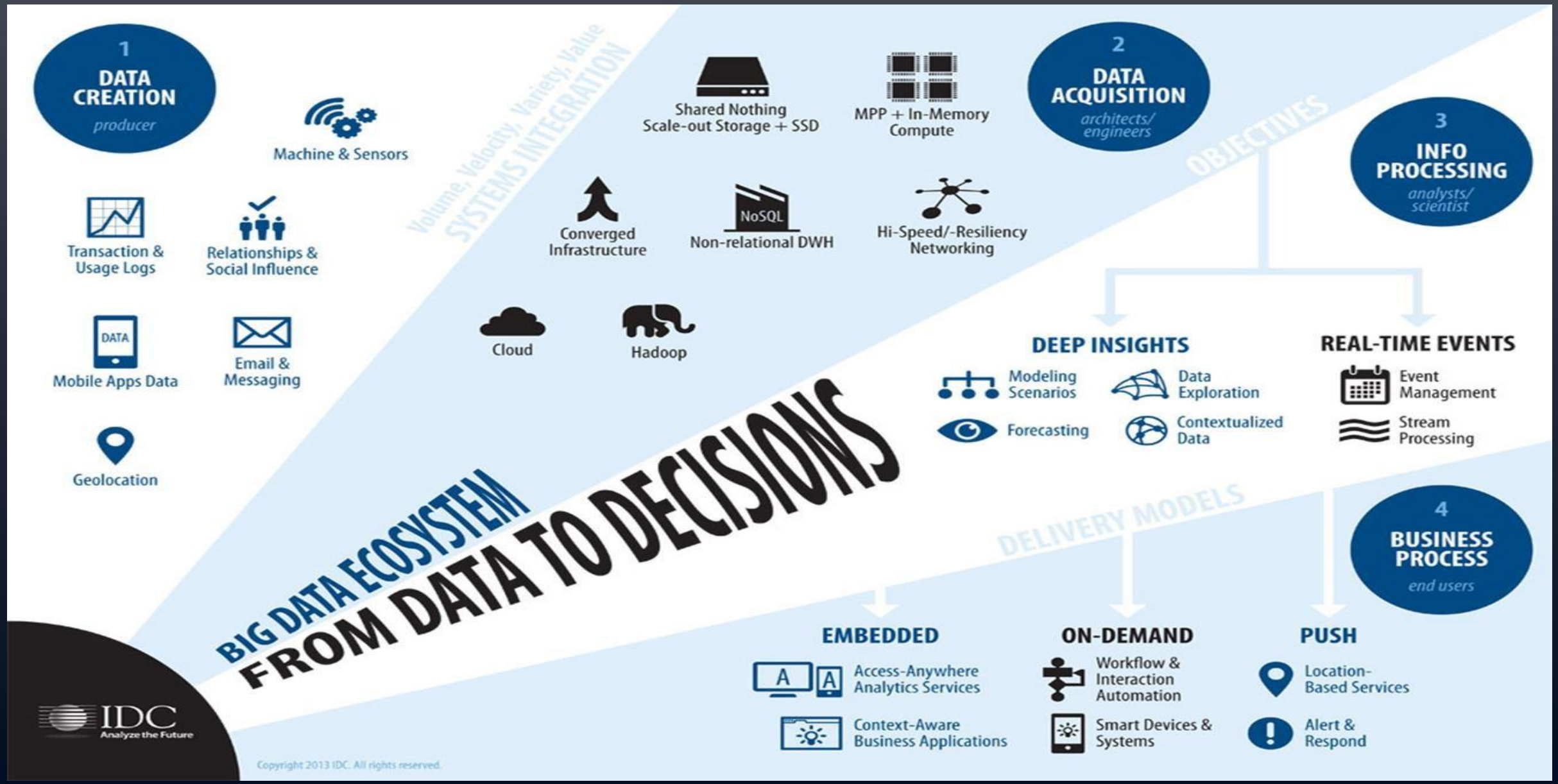
- **Data exhaust** is data that is shed as a byproduct of people's actions and movements.
  - ➢ It is the mechanism behind many services like voice recognition, spam filters, language translation, and much more.
  - ➢ When users indicate to a voice recognition program that it has misunderstood what they said, they in effect "train" the system to get better..
- Google developed the world's most comprehensive **spell checker** as an artifact of its work proposing corrections to errors in search engine queries.

The Value of Open Data

- Government are the original gatherers of information on a mass scale. Many people call for “open government data”.
- A website FlyOnTime.us shows how likely bad weather will delay flights at a particular airport, from the information amassed by the federal government.
- This shows an entity that does not collect or control information flows, like search engine or big retailer, can still obtain and use data to create value.

https://medium.com/open-data-policy-lab/open-data-index-10-insights-on-the-value-of-open-data-f810e7cb8e9

Big Data Ecosystem - From Data to Decisions



Big Data as Research & Scientific Topic

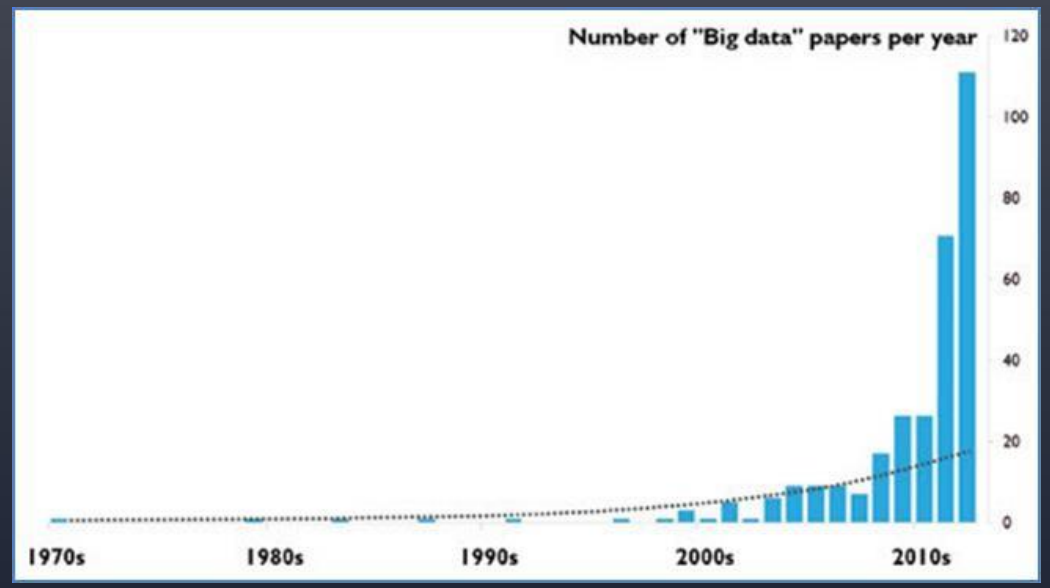


Figure 1: Time line of Big Data as topic of research. The dotted line represents the exponential growth curve best fitting the data represented by the blue bars. This shows the number of Big Data articles increasing faster than the best exponential fit.

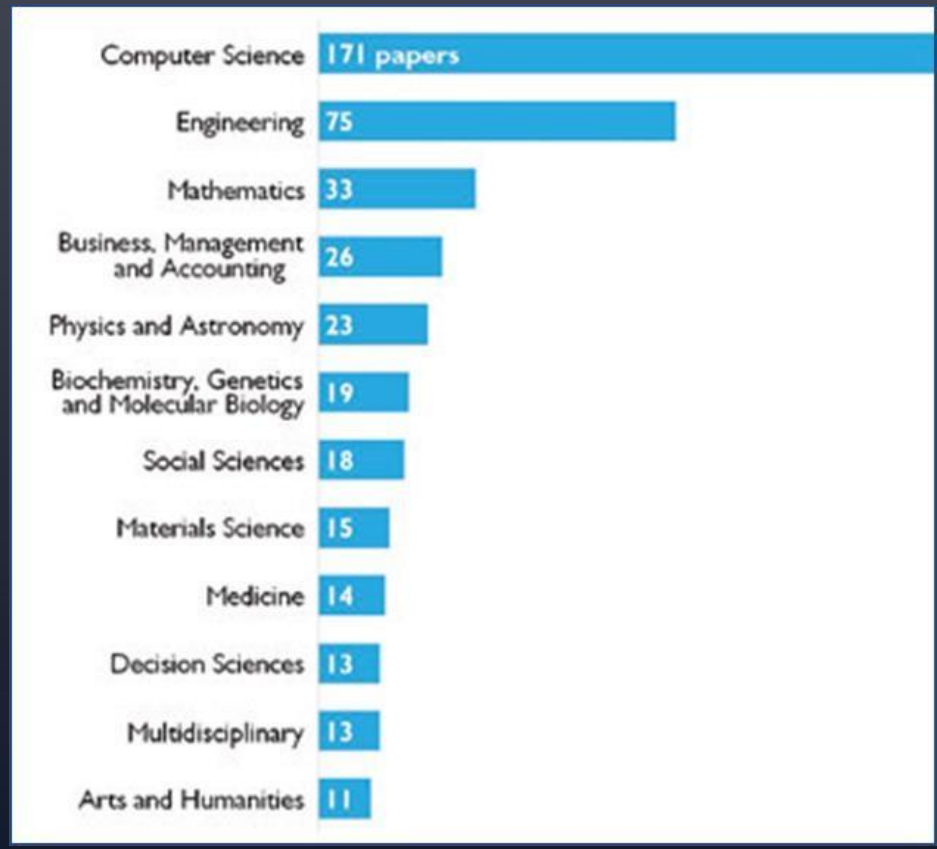


Figure 2: Subject areas researching Big Data

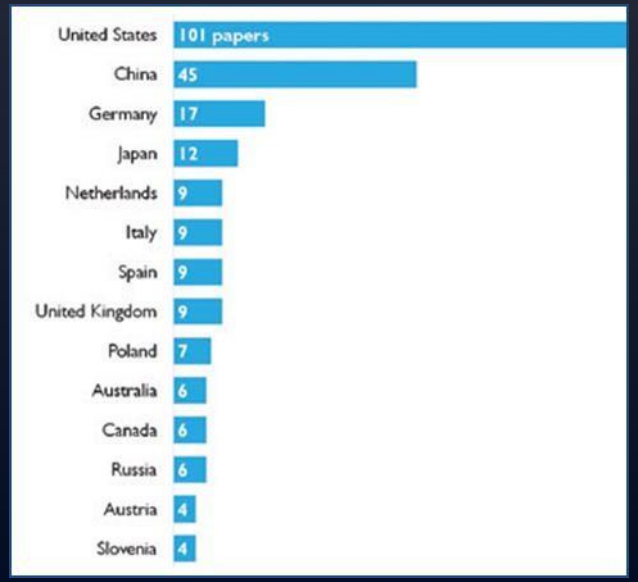
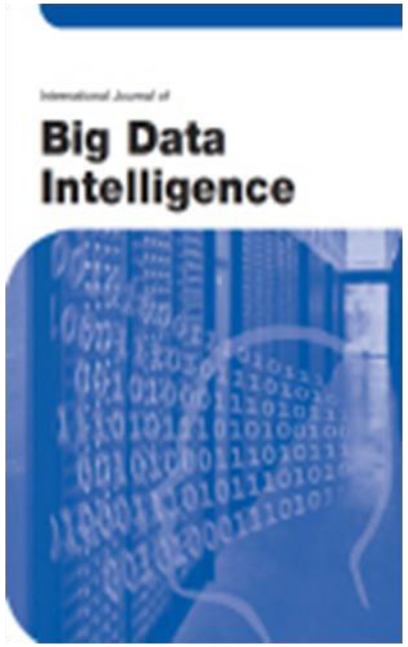
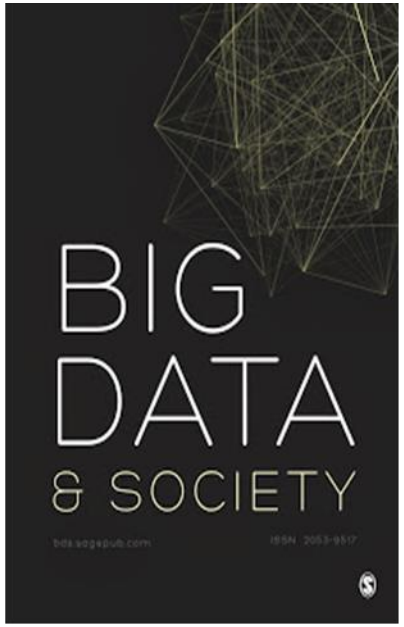


Figure 3: Geographical Distribution of Big Data papers

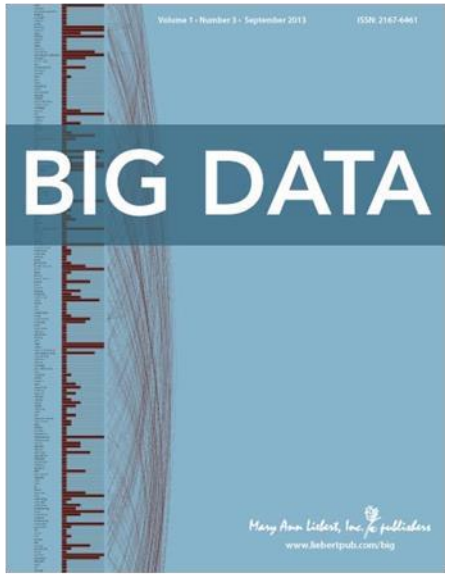
Journals

http://www.journalofbigdata.com



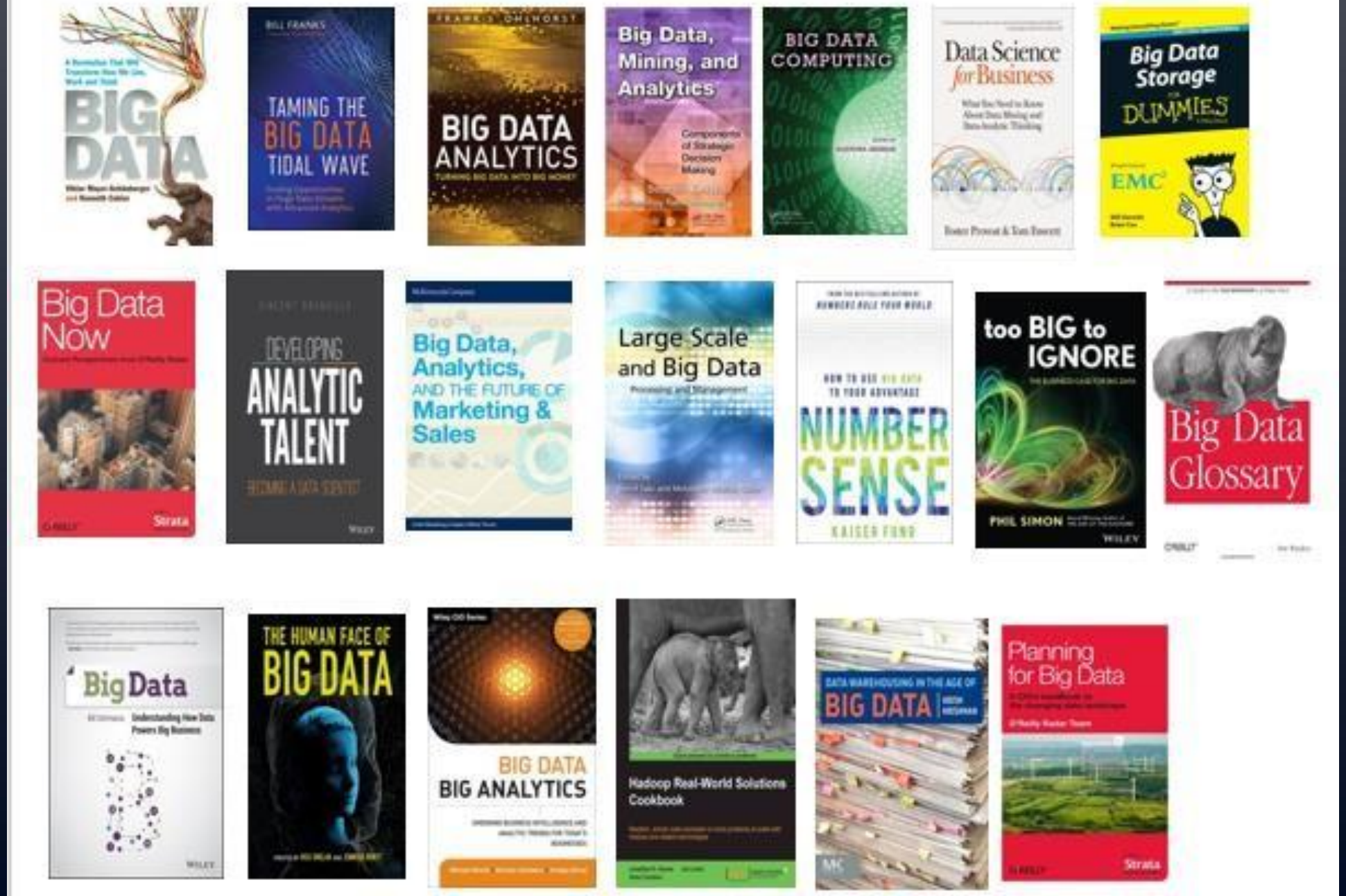


http://bds.sagepub.com/



http://www.liebertpub.com/big

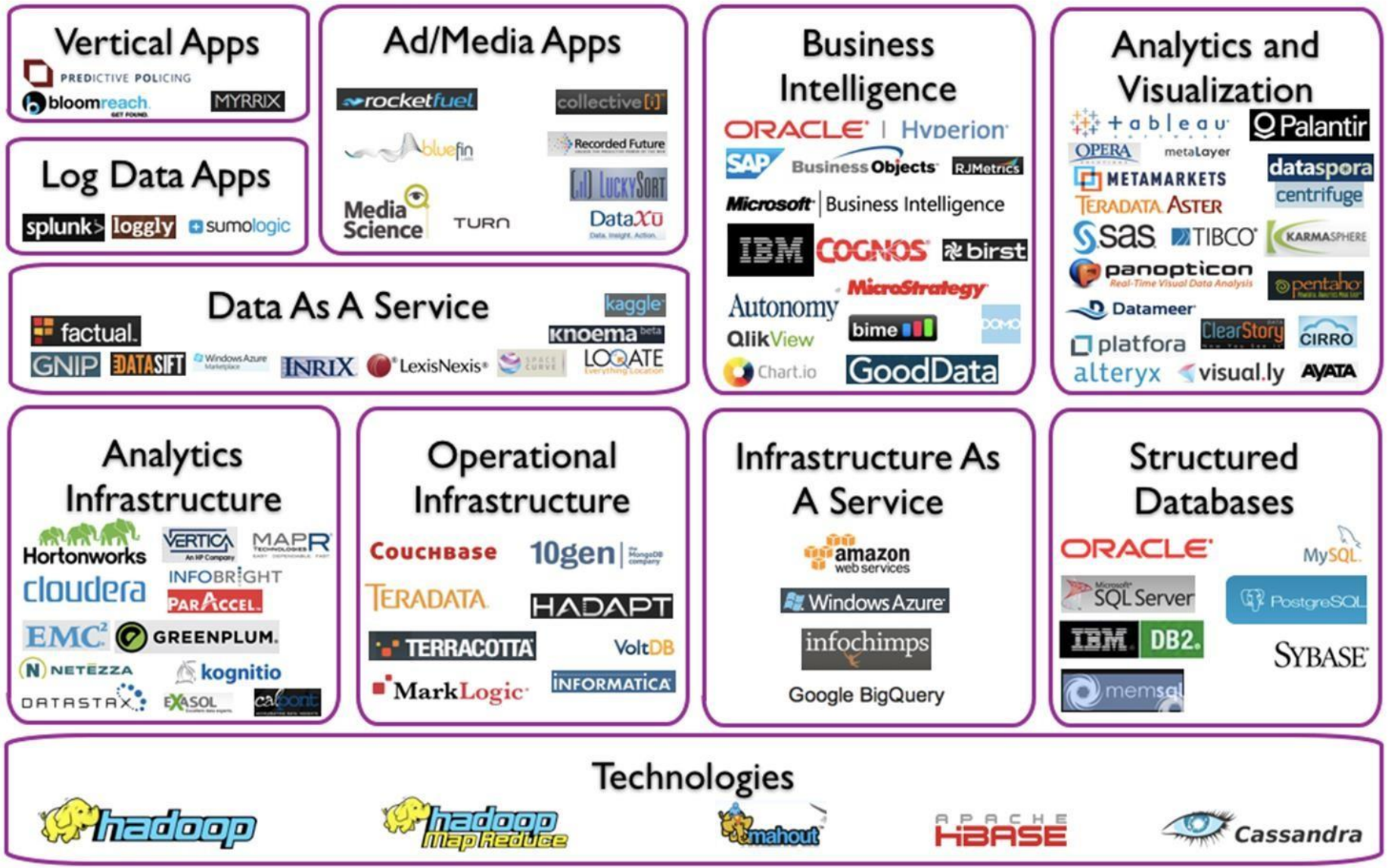
http://www.journals.elsevier.com/big-data-research/


BIG DATA
TAMING THE BIG DATA TIDAL WAVE
BIG DATA ANALYTICS
Big Data, Mining, and Analytics
BIG DATA COMPUTING
Data Science for Business
Big Data Storage for DUMMIES
EMC
Big Data Now
Strata
DEVELOPING ANALYTIC TALENT
Big Data, Analytics, AND THE FUTURE OF Marketing & Sales
Large Scale and Big Data
NUMBER SENSE
too BIG to IGNORE
PHIL SIMON
Big Data Glossary
Big Data
THE HUMAN FACE OF BIG DATA
BIG DATA BIG ANALYTICS
Hadoop Real-World Solutions Cookbook
DATA WAREHOUSING IN THE AGE OF BIG DATA
Planning for Big Data
Strata

Conferences

- IEEE International Conference on Big Data (BigData 2022)
- Data 2022
- World Data Summit 2022
- Big Data World 2022
- IEEE Big Data Service 2022
- Global Big Data Conference

Big Data Landscape



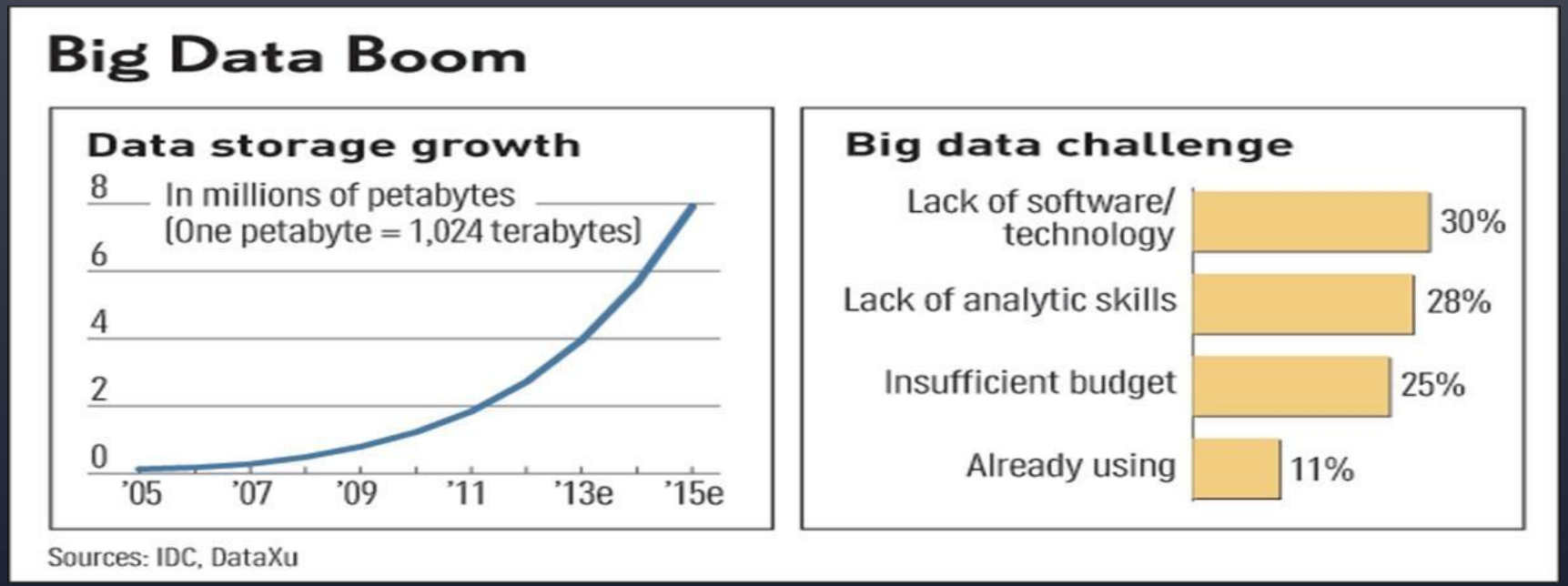
dave@vcdave.com

blogs.forbes.com/davefeinleib

Opportunities & Benefits of Big Data Utilization

- Make better fact-based decision faster.
- Improved customer experience.
- Increased sales.
- New product innovations.
- Reduced risk
- More efficient operations.
- Higher quality products and services.

Challenges in Handling Big Data

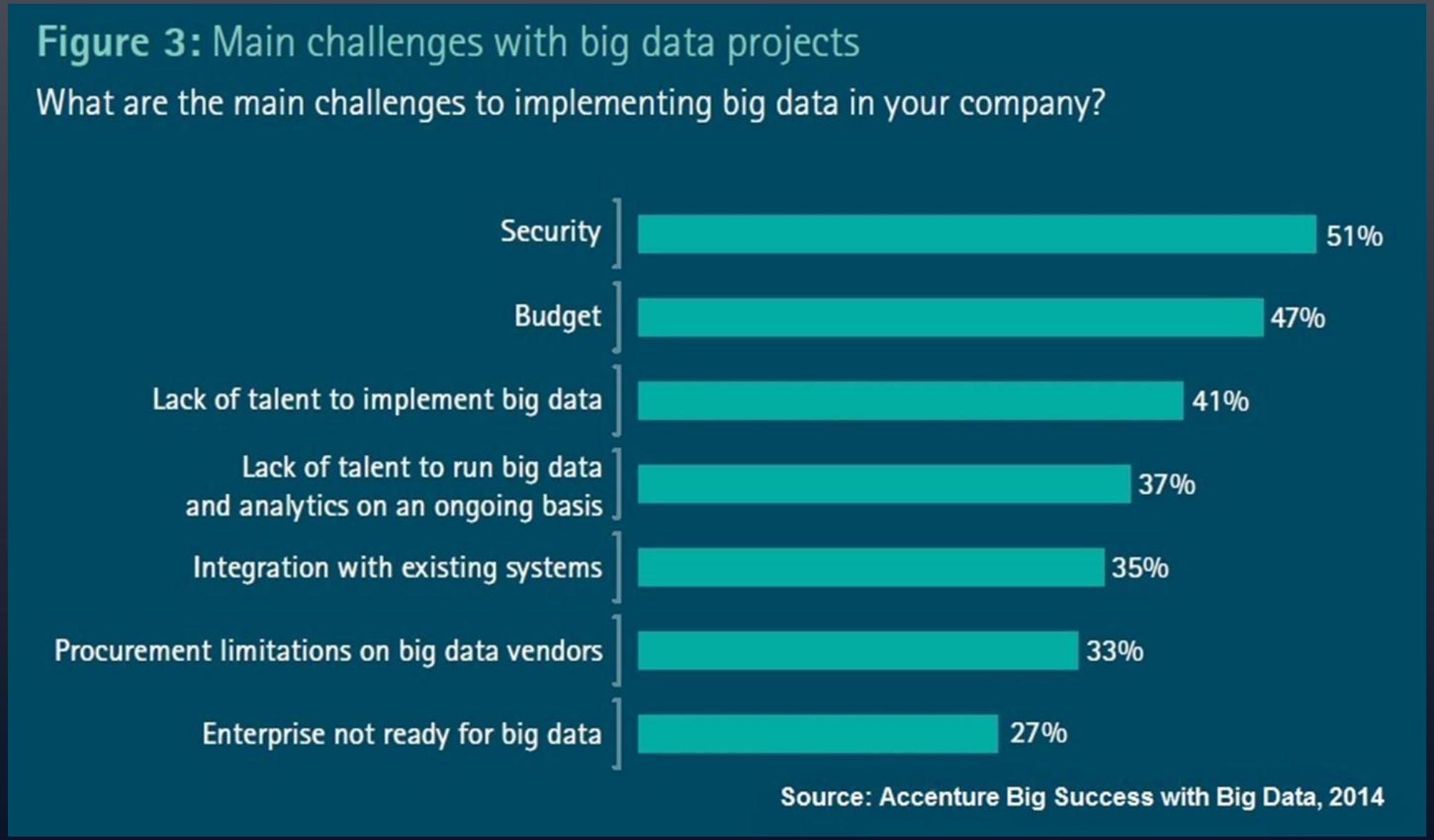


The Bottleneck is in technology

New architecture, algorithms, techniques are needed.

Also in technical skills

Experts in using the new technology and dealing with big data.



Source - http://www.forbes.com/sites/gilpress/2014/09/10/new-surveys-on-big-data-big-decisions-analysis-and-intuition/

Data Drivenness

"Without **data**, all anyone has are **opinions**. **Data elevates** the probability that you'll make **the right decision**."

W. Edwards Deming

"**Data driven** means that progress in an activity is compelled by **data** rather than by **intuition** or **personal experience**. It is often labeled as the business jargon for what scientists call **evidence based decision making**"

Wikipedia 2015-02-02

Data-driven Decision Making

- **Data-driven decision making** refers to the practice of basing decisions on the analysis of data (i.e. 'learning from data'), rather than purely on gut feeling and intuition.

Data Denial	Data Indifferent	Data Informed	Data Driven
You distrust data and avoid using it	You don't care about data and have no need for it	You use it only when it supports your opinions or decisions	You use it to shape and inform all your decisions

What Makes an Organization Data-Driven?

Prerequisite #1

An organization must be collecting data.

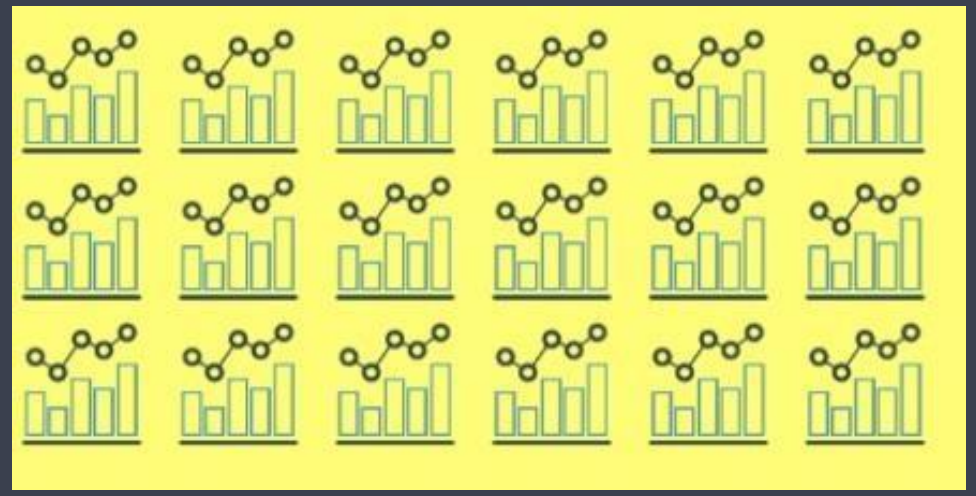
Prerequisite #2

Data must be accessible and queryable.

Prerequisite #3

People with skills to use the data, extract the right data and use that data to inform next steps.

What Makes an Organization Data-Driven?



Having lots of reports?

Having lots of alerts?

Having lots of dashboards?

Necessary but not sufficient.
Need causal explanation (why Q),
Need analysis.

Source: Carl Anderson, Creating a Data-Driven Organization

From Reporting , Alerting to Analysis

Reporting	Analysis
Descriptive	Prescriptive
What?	Why?
Backward-looking	Forward-looking
Raise questions	Answer questions
Data → information	Data + information → insights
Reports, dashboards, alerts	Findings, recommendations, predictions
No context	Context + storytelling

To be data driven – must have **analytics**

Framework for Understanding Analytics

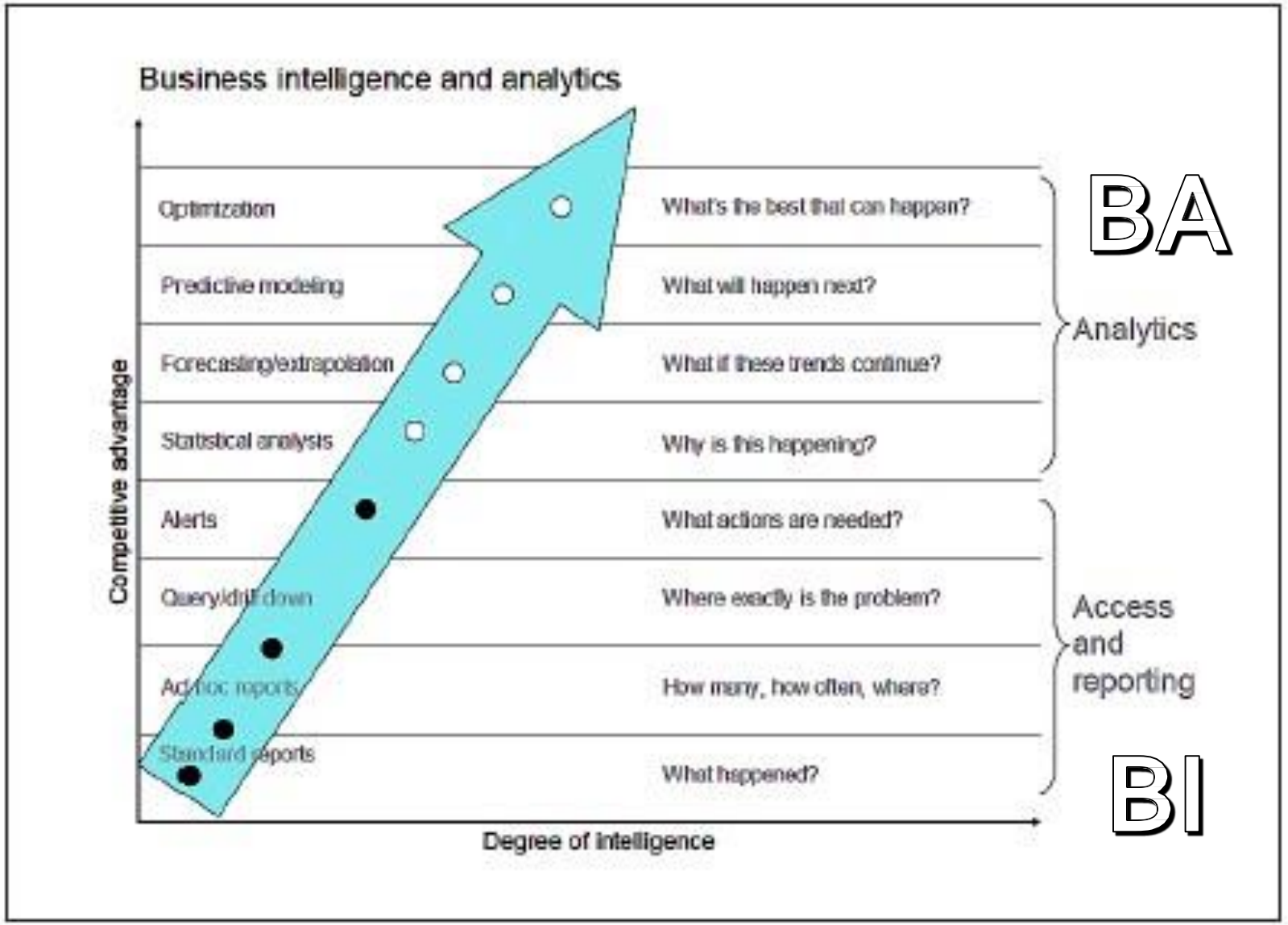
	Past	Present	Future
Information	A) What happened? Reporting	B) What is happening now? Alerts	C) What will happen? Extrapolation
Insight	D) How and why did it happen? Modeling, experimental design	E) What's the next best action? Recommendation	F) What's the best/ worst that can happen? Prediction, optimization, simulation

Only by understanding why something happened you can formulate a plan or set of recommendations (E).
E) and F) are **truly data-driven** but **if and only if the information is acted upon**

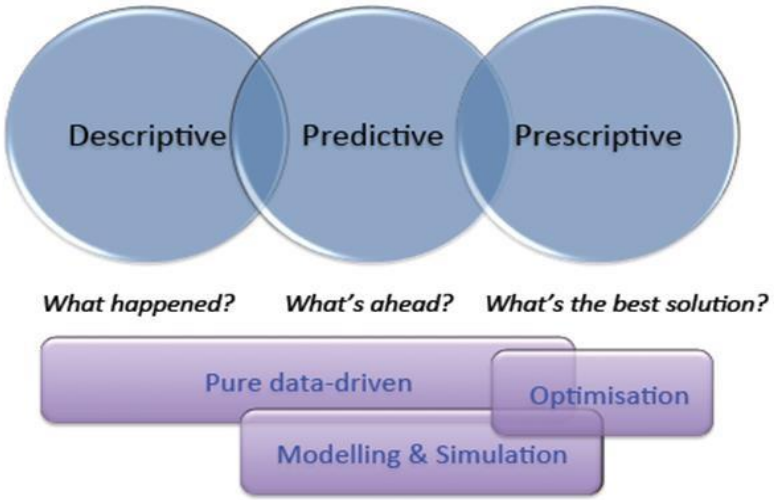
Analytics Value Chain

In a data-driven organization, the data feeds reports, stimulating deeper analysis. Analysis is placed in the hands of the decision makers who incorporate them into their decision-making process, influencing the direction that the company takes and providing value and impact.

Analytics Maturity



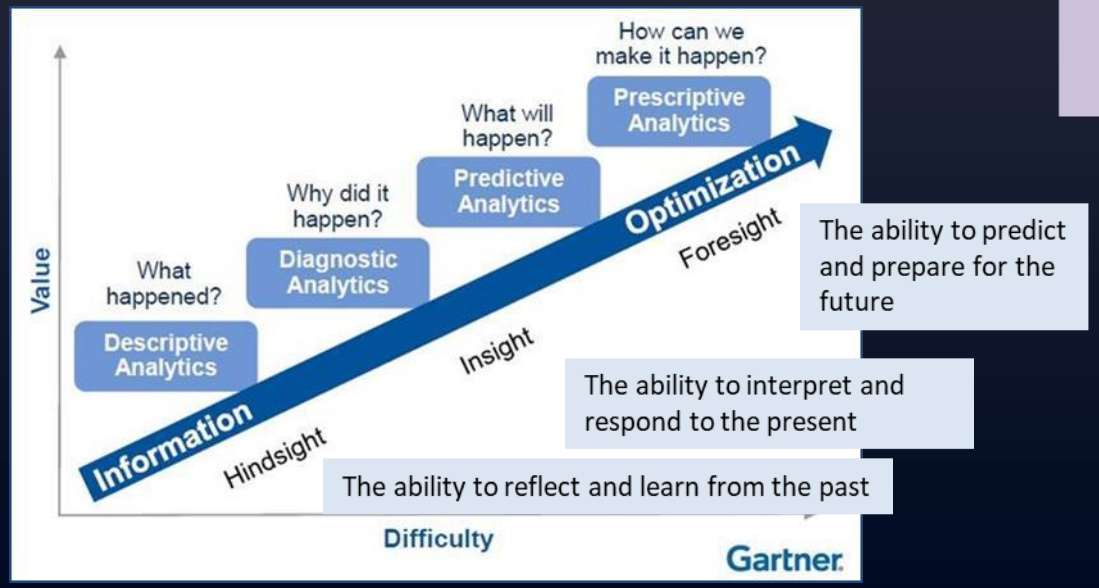
How to leverage data?



Business Analytics is the use of: data, information technology, statistical analysis, quantitative methods, and mathematical or computer-based models to help managers gain improved insight about their business operations and make better, fact-based decisions.

Scope of Business Analytics

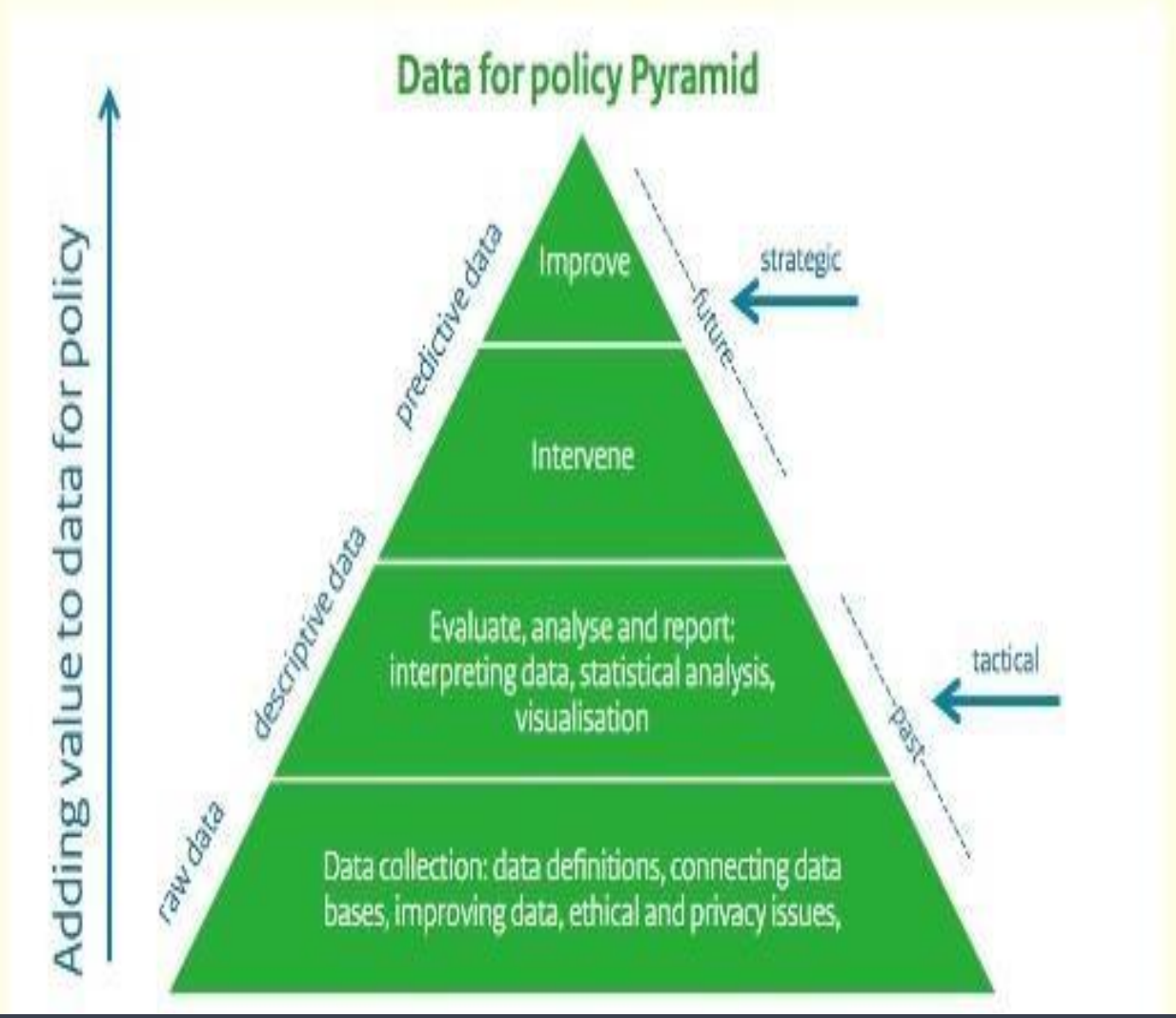
- **Descriptive analytics**
  - uses data to understand past and present
- **Predictive analytics**
  - analyzes past performance
- **Prescriptive analytics**
  - uses optimization techniques



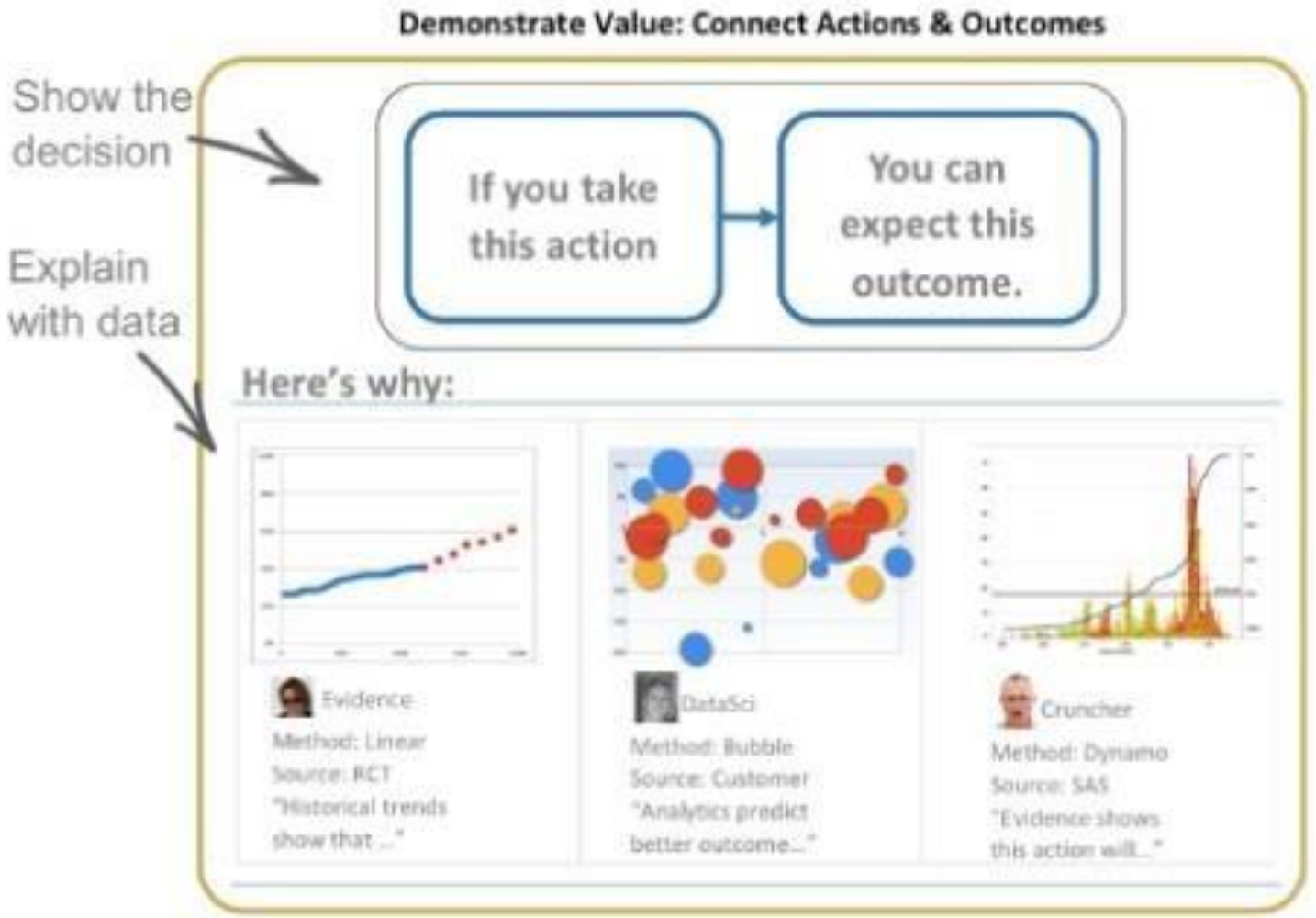
Current-sight	What is happening?	Operational Reporting
Hindsight	What has happened?	Historical Reporting
Insight	Why did it happen?	Investigative Analytics
Foresight	What will happen?	Predictive Analytics

Four Predominant Information Categories

- Policy makers want the top of the iceberg, but they need to remember the stuff beneath sea (adapted from @HetanShah):



Tie actions to outcomes



https://www.slideshare.net/CarlAnderson4/ddo-seattle

Data-driven World

Airbus 380

- 1 billion line of code
- each engine generate 10 TB every 30 min
- 640TB per Flight

twitter

12 TB of data per day

NYSE

1TB of data everyday

What we do with these amount of data?

SCIENCE -Data bases from astronomy, genomics, environmental data, transportation data, …

Humanities & Social Sciences - Scanned books, historical documents, social interactions data, new technology like GPS …

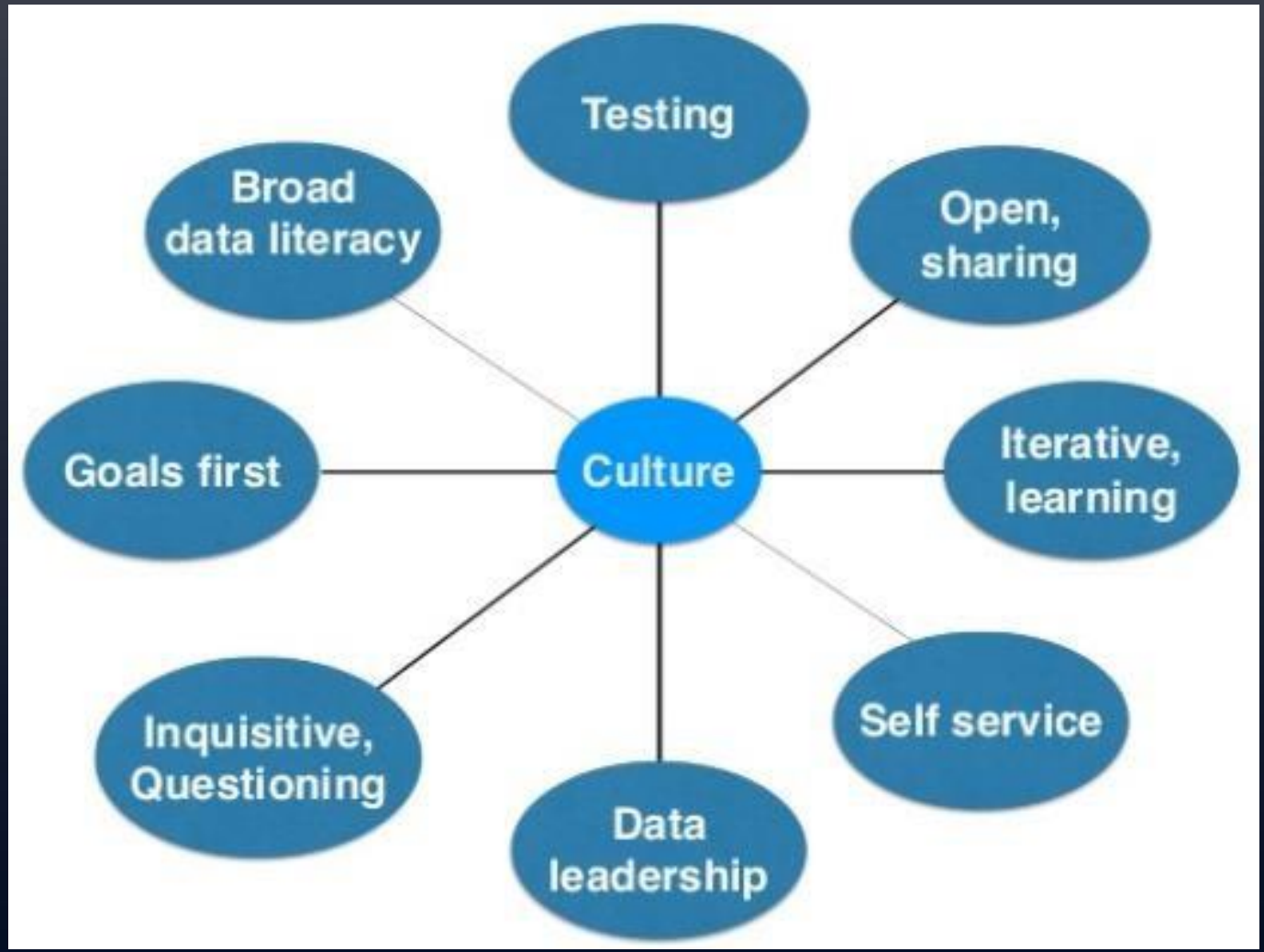
Business & Commerce
Corporate sales, stock market transactions, census, airline traffic, …

Entertainment
Internet images, Hollywood movies, MP3 files, …

Medicine
MRI & CT scans, patient records, …

Data driven culture

https://www.slideshare.net/CarlAnderson4/ddo-seattle



Recall: Data Science versus Statistics

Statistics traditionally is concerned with analyzing **primary** (e.g. experimental) **data** that have been collected to explain and check the validity of specific existing ideas (hypotheses).

- ✓ Primary data analysis or top-down (**explanatory and confirmatory**) analysis.
- ✓ '**Idea (hypothesis) evaluation or testing**' .

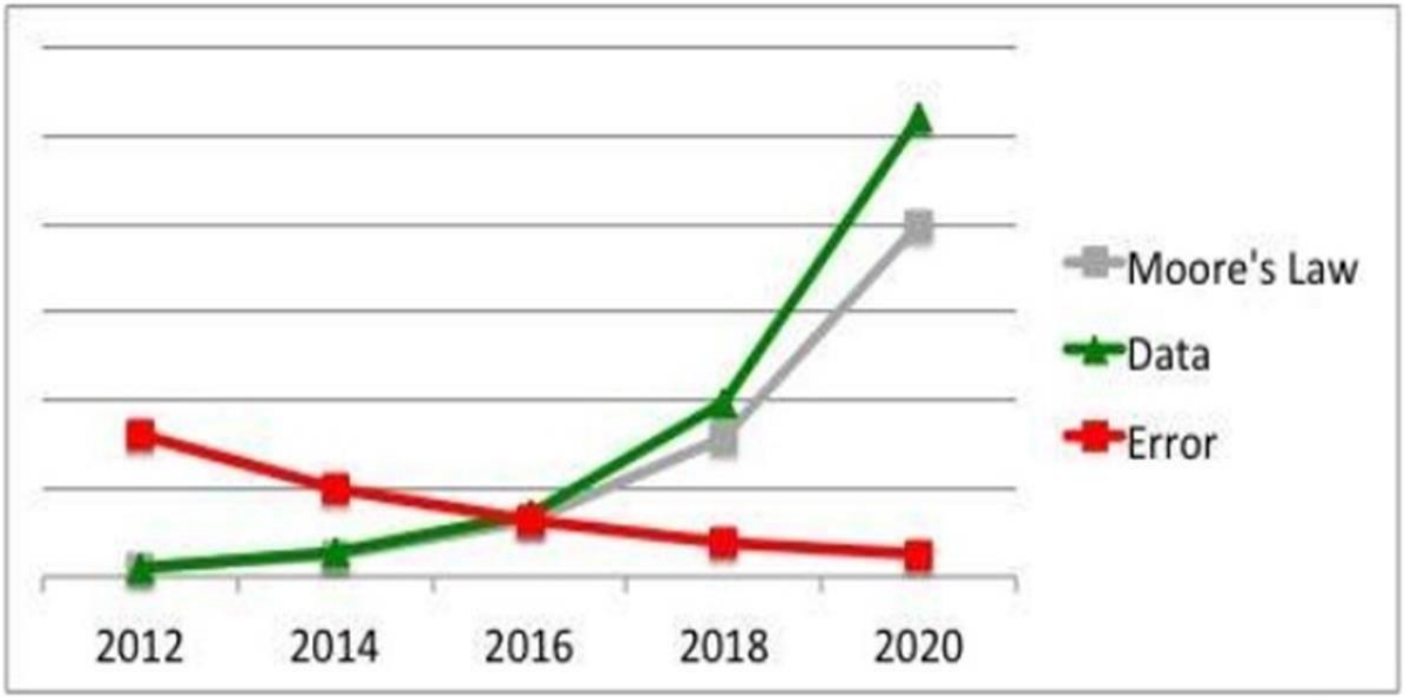
Data science (or data mining), on the other hand, typically is concerned with analyzing **secondary** (e.g. observational or 'found') **data** that have been collected for other reasons (and not 'under control' of the investigator) to create new ideas (hypotheses).

- ✓ Secondary data analysis or bottom-up (**exploratory and predictive**) analysis.
- ✓ '**Idea (hypothesis) generation**'

Why bother?

"It's not who has the best algorithm that wins. It's who has the most data."

- Banko and Brill, 2001



Source : https://amplab.cs.berkeley.edu/2013/02/07/for-big-data-moores-law-means-better-decisions

Data overtake algorithm!

More Data or Better Models

In machine learning, very often:
more data -> better outcomes

- More examples to learn from
- More possible feature types
  - —We're looking for the most useful for our task

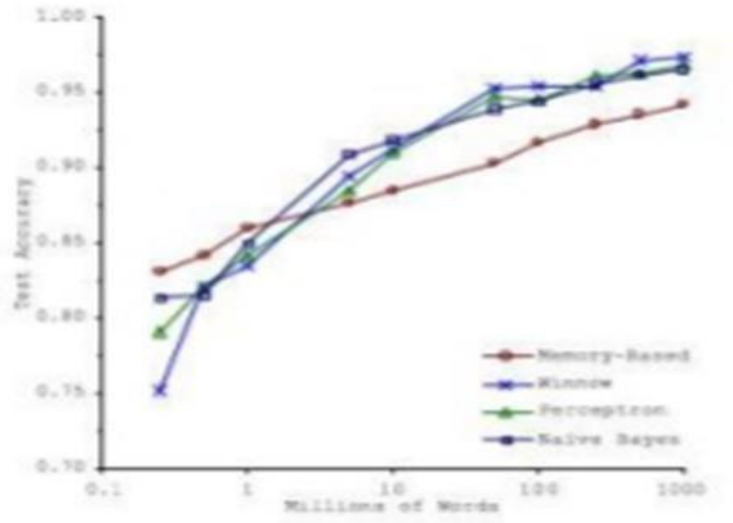
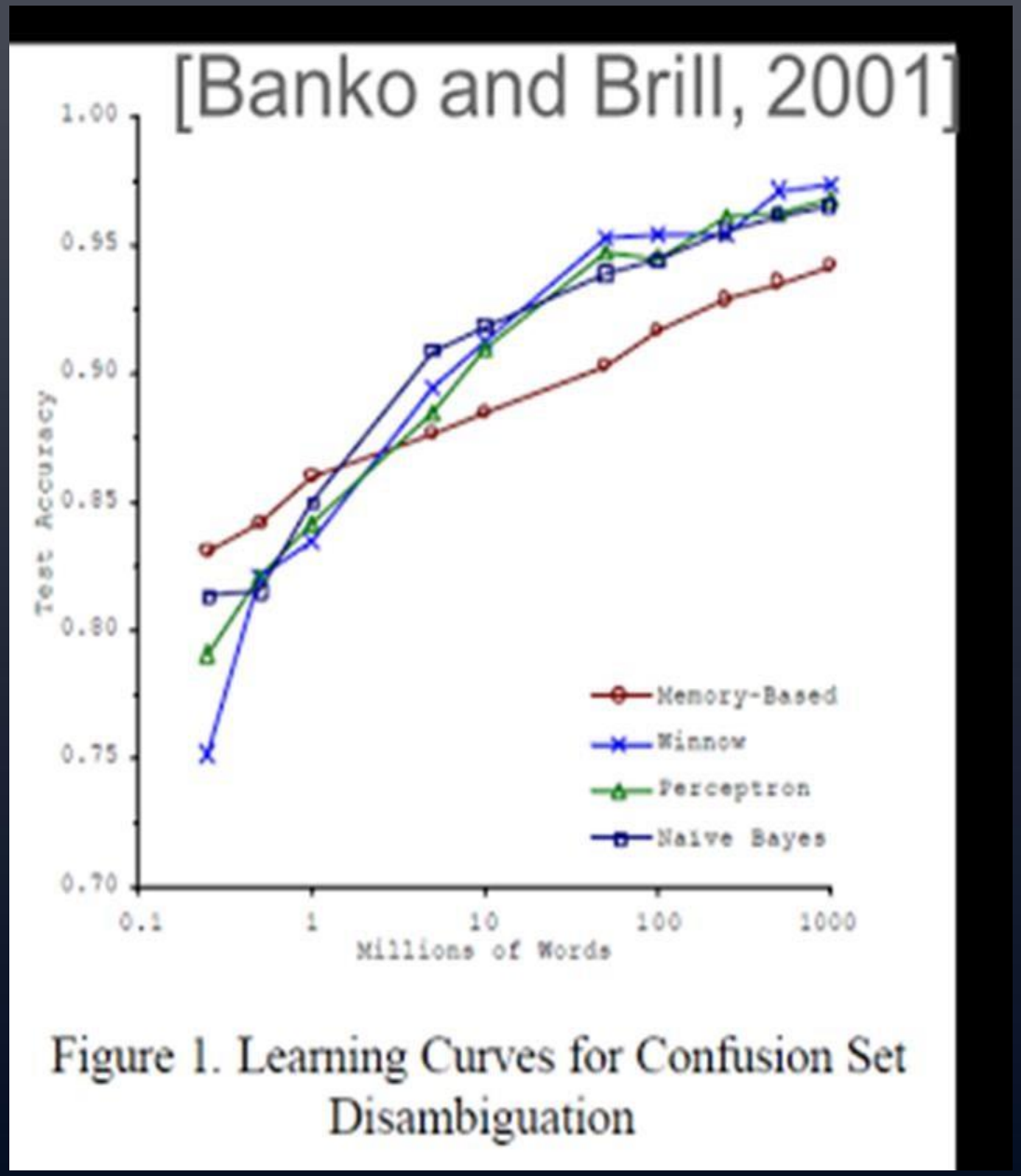


Figure 1. Learning Curves for Confusion Set Disambiguation

Banko & Brill, 2001

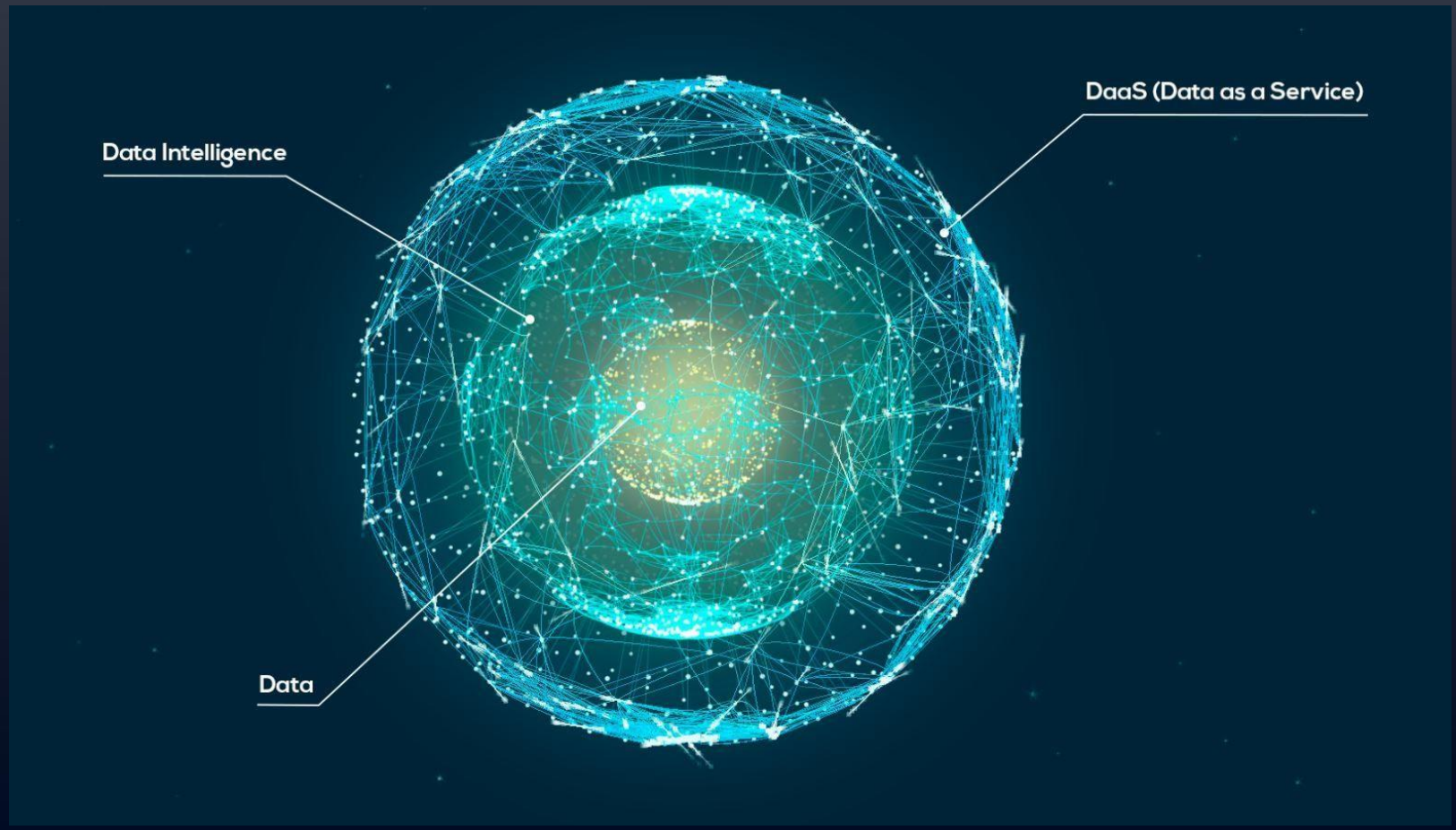


The figure shows that, for the given problem, very different algorithms perform virtually the same. however, adding more examples (words) to the training set monotonically increases the accuracy of the model.

Harnessing Big Data

OLTP : Online Transaction Processing (DBMSs)
OLAP : Online Analytical Processing (Data Warehousing)
RTAP: Real-Time Analytics Processing (Big Data Architecture & technology)

Becoming a **data-driven** company is a useful first step, but is based on building tools, abilities, and a culture that acts on data, instead of really making an internal transformation around data.
Use your data-driven experiences to move up to a higher lever and become a **data-centric** company, putting your data at the core of your organization.



Many organizations now claim to be **data-driven**, making decisions which are often fully automated, based on data rather than simply informed by the figures in reports.
The next evolution will be towards **data-centricity**.

This is where an absolute commitment to high-quality centralized data forms the core of business operations.
Tools are built around data, rather than the current status quo of building tools that act upon the organizational data siloes.
It becomes less about collecting and hoarding data – the Big Data mentality – and more about **acting on data intelligently**.

Reference: https://www.dataversity.net/predictions-2019-data-analytics-trends-to-watch/

Summary

- ✓ Think Big
- ✓ Start Small
- ✓ Scale Up
- ✓ Fail Fast & Cheaply
- ✓ Start with experimentation
- ✓ Focus on Value

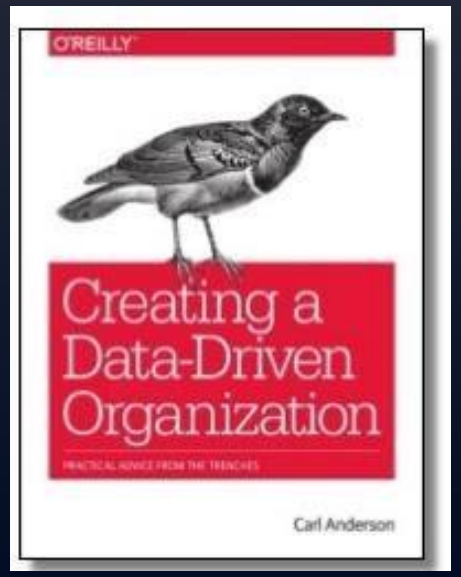
TED TALKS

Big data is better data – Kenneth Cukier, TEDSalon Berlin

https://www.ted.com/talks/kenneth_cukier_big_data_is_better_data?language=en

References

- Mayer-Schönberger, V., & Cukier, K. (2013). Big data: A revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.
- Research Trends- Special Issue on Big Data
- Lesson Link - https://www.tes.com/lessons/ye-DW-tMx48Lew/big-data-and-data-driven-decision-making



(Credit: iStock)