Role of Big Data and Cloud in Digital Economy

Dr. Muhammad Nouman Durrani and Nadeem Kafi Khan
Assistant Professors at FAST NUCES, Karachi.

14th Symposium on:
Preparing for our Digital Economy
Pakistan Academy of Engineering

Sequence of Presentation

- What is Big Data?
- Contributors
- Characteristics of Big Data
- Architecture of Big Data
- Digital Economy
- Big Data Scope in Digital Economy
- Opportunities and Challenges of Digital Economy
- Role of Big Data in Digital Economy
- Understanding Digital Economy from the perspective of Cloud computing and Big-Data

What is Big Data?

Big Data: Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, etc.

traditional computer scientists

data that will not fit in main memory.

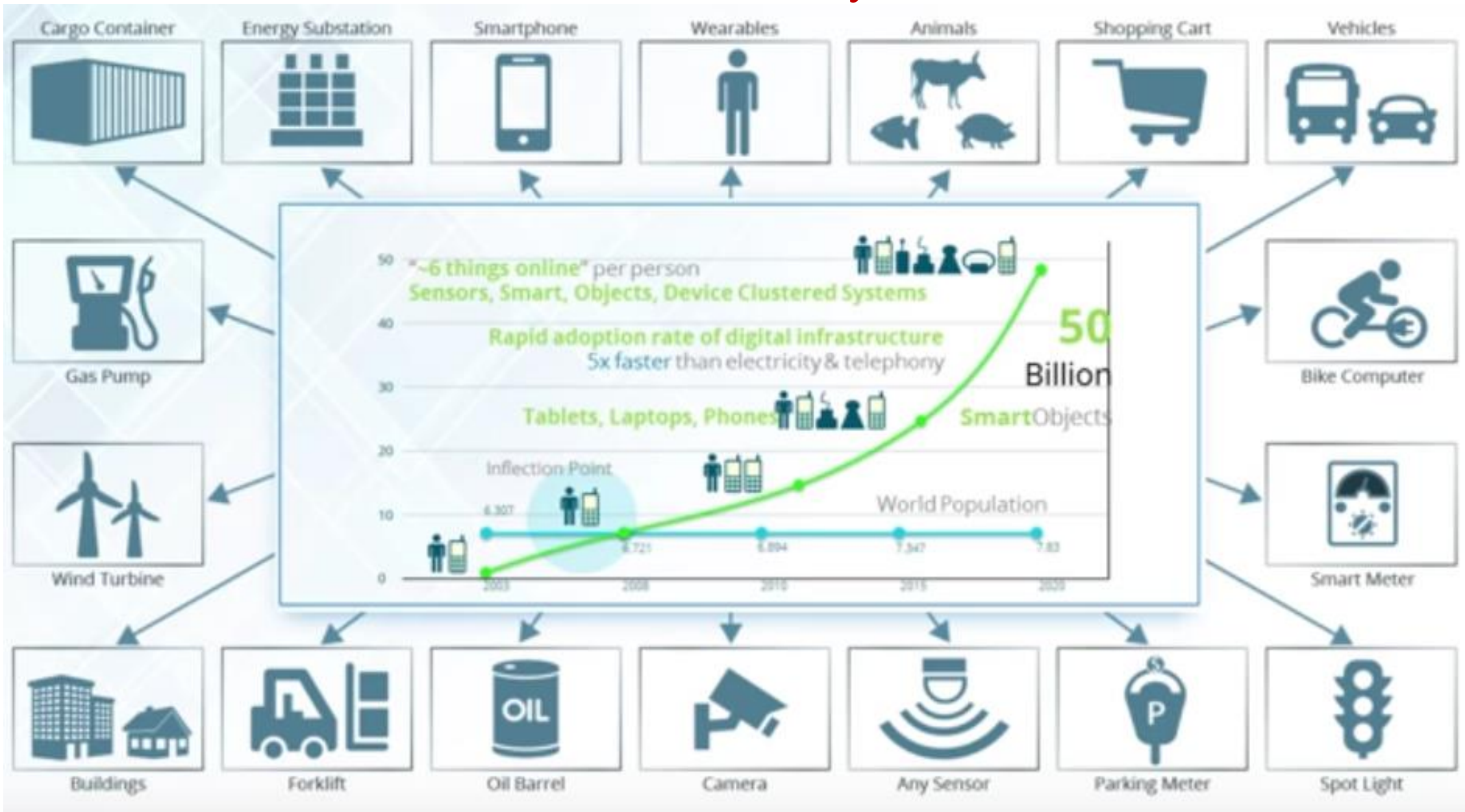
data with a *large* number of observations and/or features.

statisticians

other fields

non-traditional sample size (i.e. > 100 subjects); can't analyze in stats tools (Excel).

IoTs: 50 Billion Connected Devices by 2020



THE WORLD OF DATA

- Peta-bytes are in norm
  - Google processes 24 PB a day (2009)
  - AT&T transfers about 30 PB a day through its networks
  - Microsoft migrated 150 PB of user data from Hotmail to Outlook (2013)
  - Facebook stores about 357 PB of user uploaded images (2013)
  - eBay has 6.5 PB of user data + 50 TB/day (2009)
- How big is Internet? 672 Exabytes of accessible data (2013)

Contributors: Surveillance guys

Contributors: Social Networks

Data Generated Every Minute!

On a typical day:

- 500 million tweets
- 55 million FB status updates
- 1 Billion pieces of contents shared on FB

Contributors: Scientific Instruments

Social media and networks
(all of us are generating data)

Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Big Data Definition

- No single standard definition…

"***Big Data***" is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it…

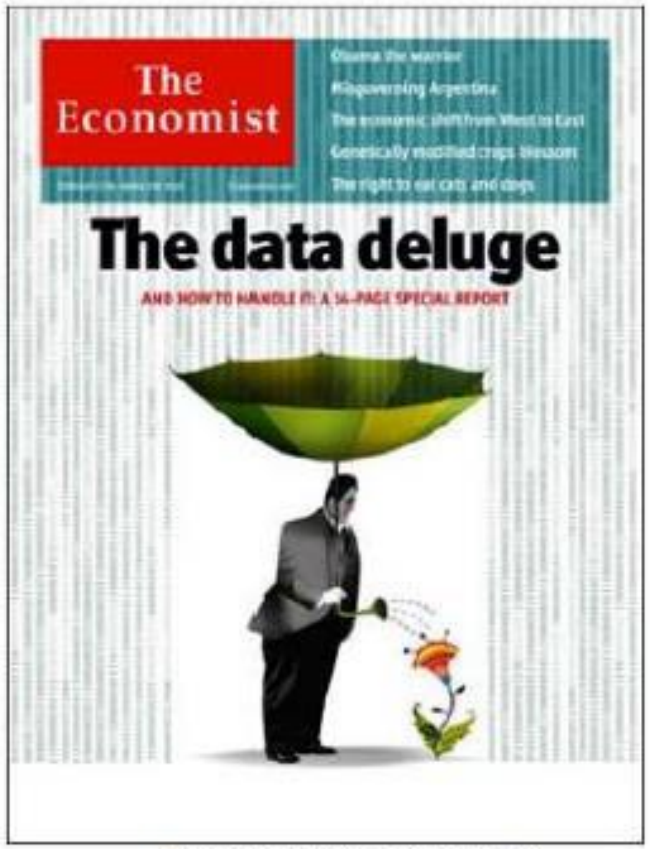
What is Big Data?
Is it only the volume that makes it Big?

- "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population." *Teradata magazine article, 2011*

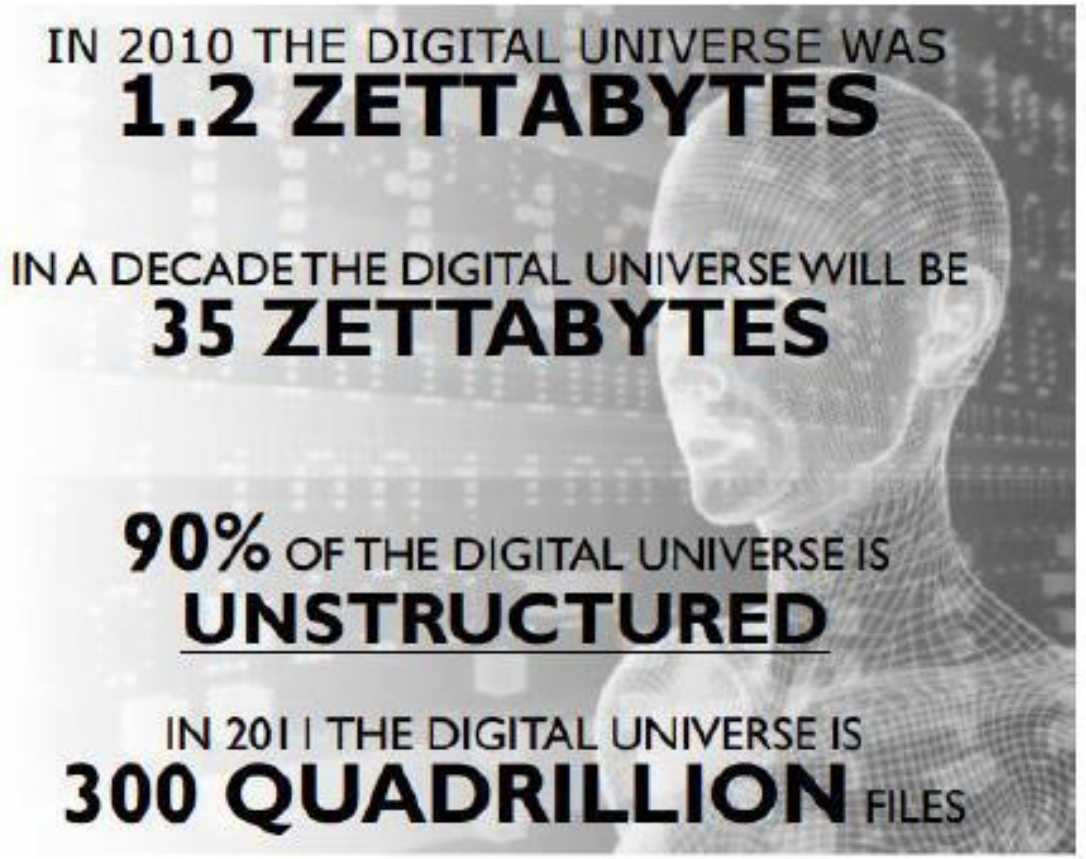
- "Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze." *The McKinsey Global Institute, 2011*

- "Big Data is any data that is expensive to manage and hard to extract value from." *Michael Franklin, Univ; of California, Berkeley*

Customer Challenges: The Data Deluge



The Economist, Feb 25, 2010

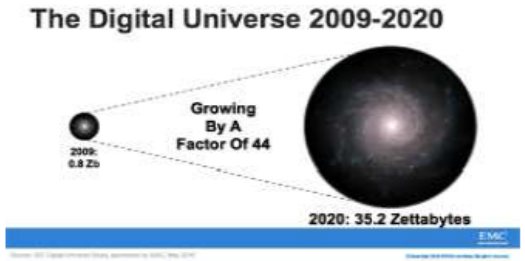


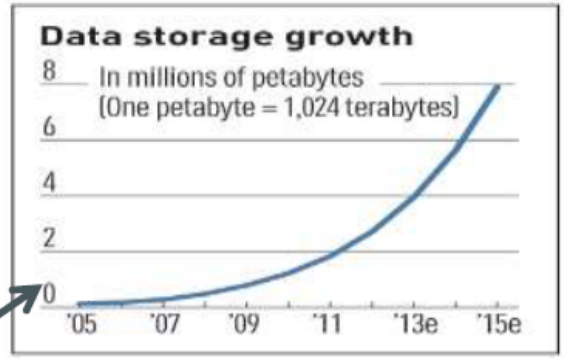
The **data deluge** refers to the situation where the sheer volume of new **data** being generated is overwhelming the capacity of institutions to manage it and researchers to make use of it.

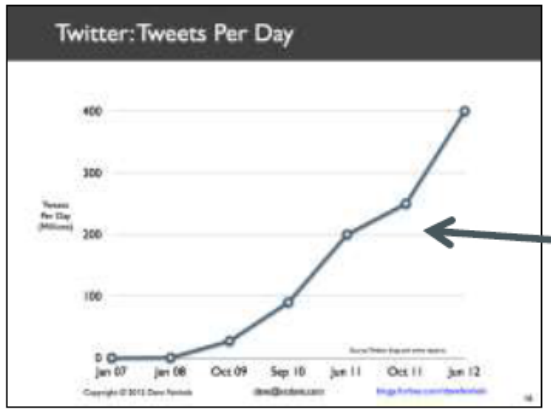
WIRED The New York Times Bloomberg Businessweek Forbes WALL STREET JOURNAL

Characteristics of Big Data:
1-Scale (Volume)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially





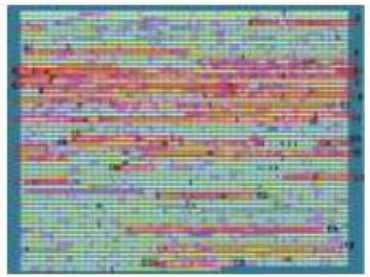


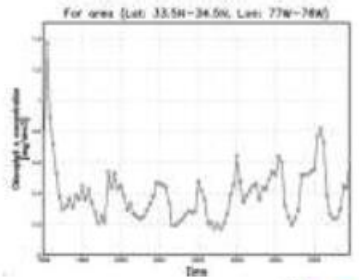
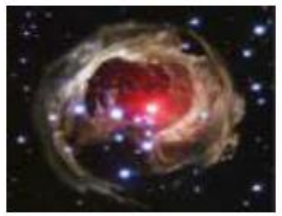
Exponential increase in collected/generated data

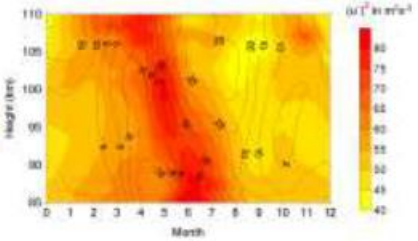
Collection, storage and distribution of data is overwhelming the traditional management techniques

Characteristics of Big Data:
2-Complexity (Varity)

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc…
- A single application can be generating/collecting many types of data



Most of the data is unstructured; e.g. web pages, emails, document files, audio and video, sensors

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

Characteristics of Big Data: 3-Speed (Velocity)

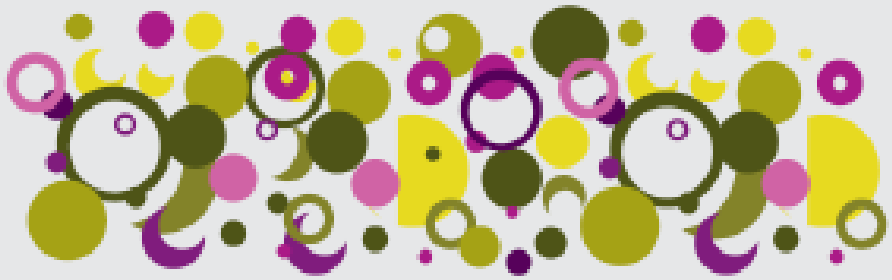
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities

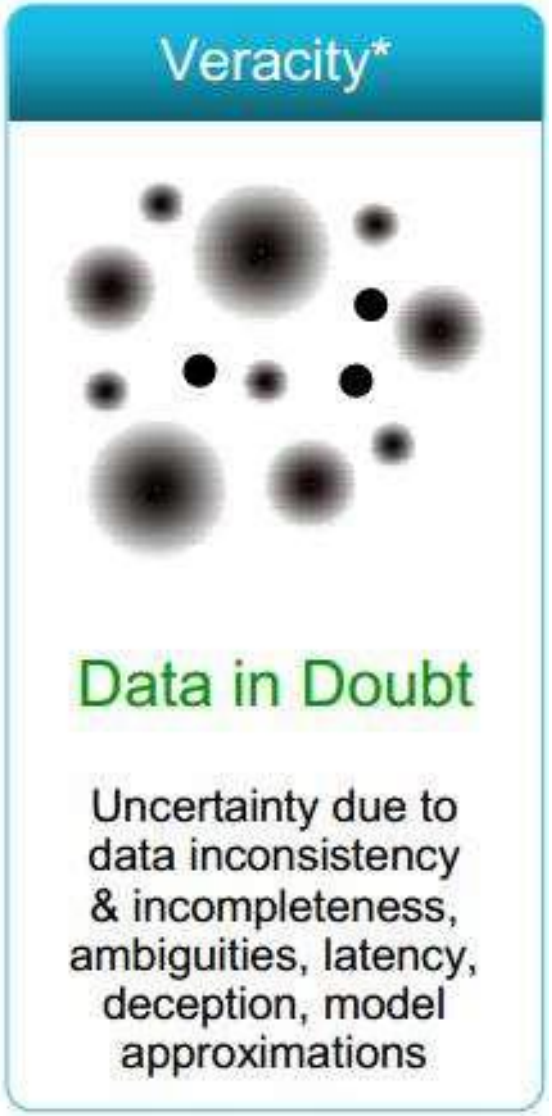
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Characteristics of Big Data

4 - Data in Doubt (Veracity)

Veracity

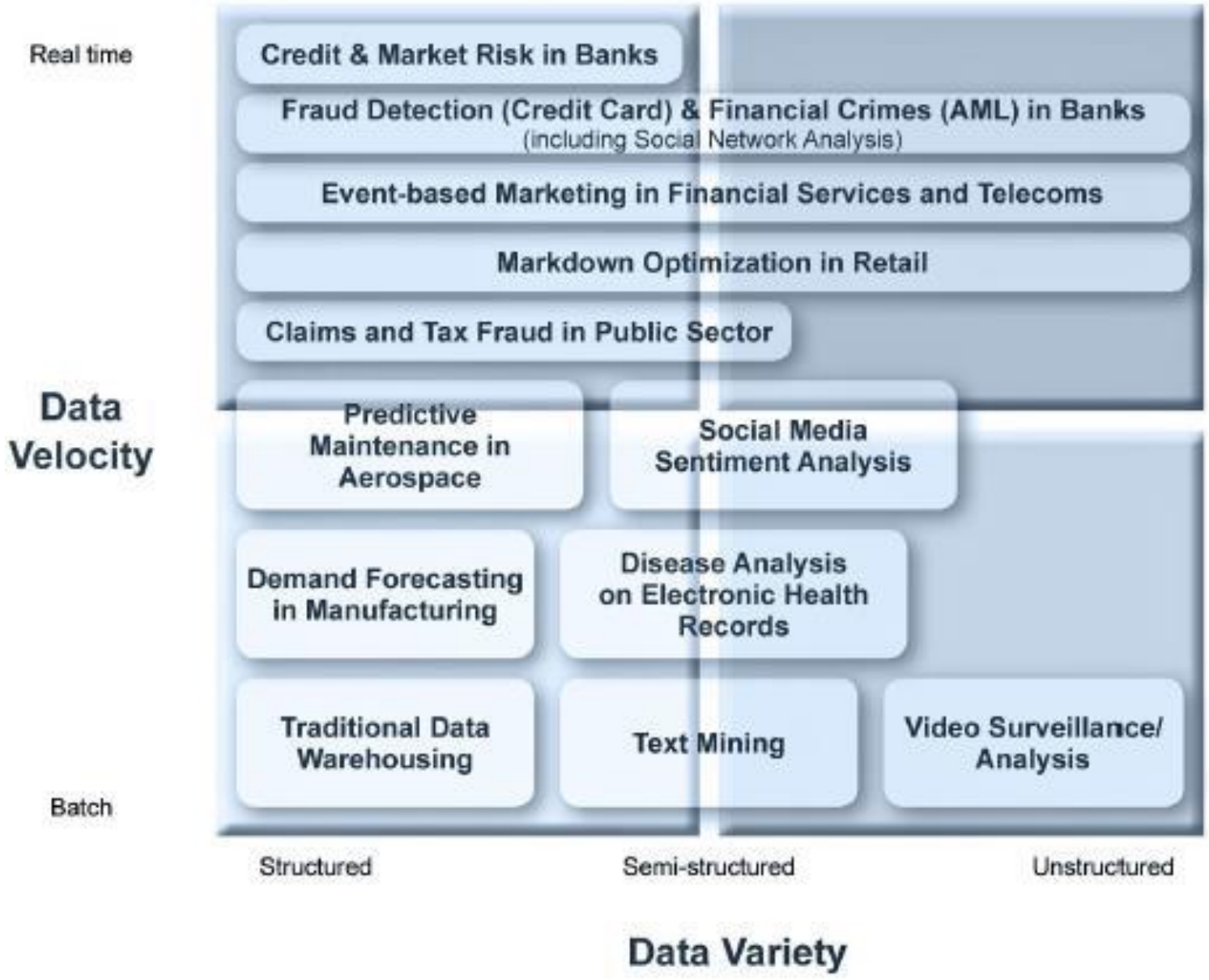




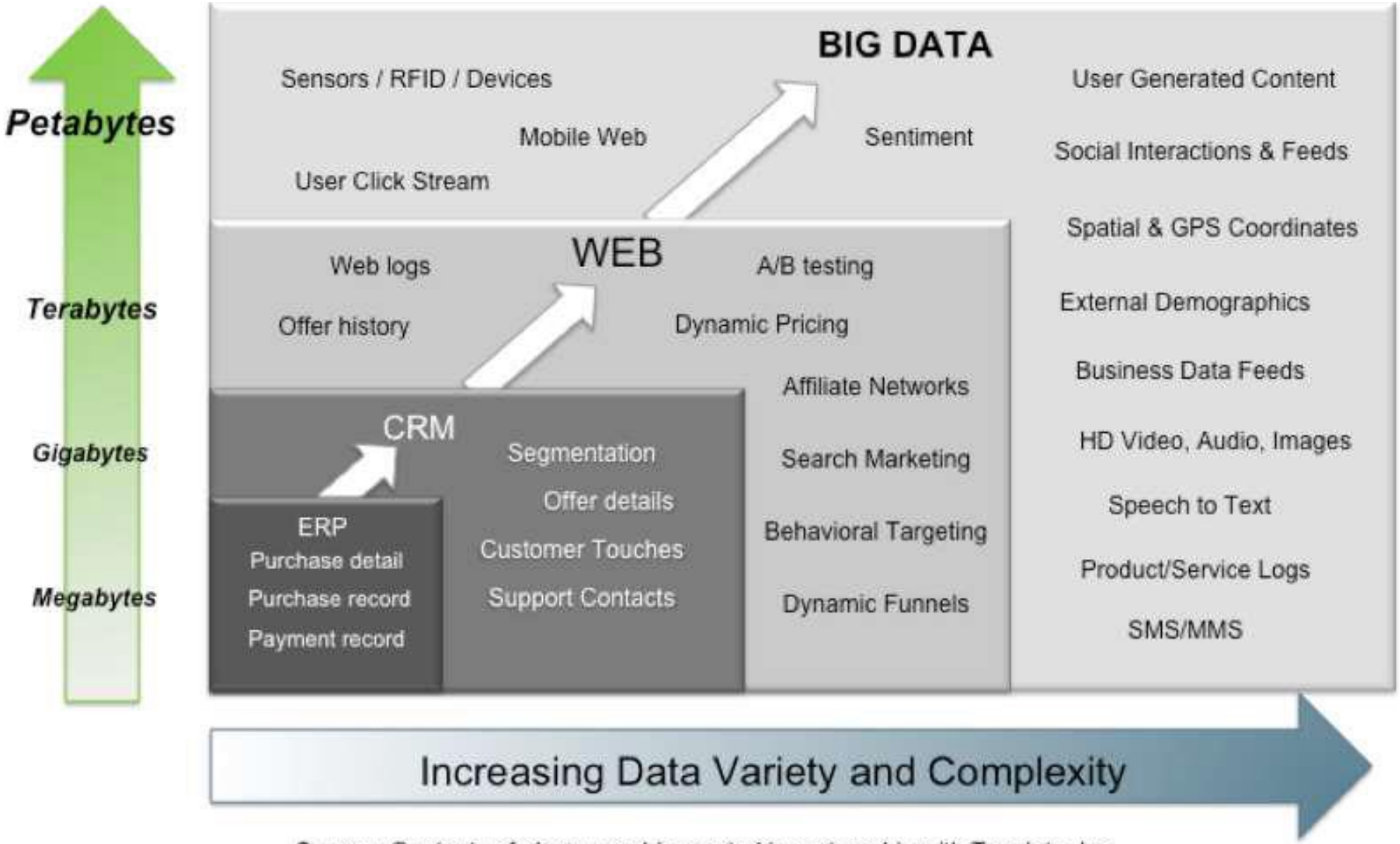
Refers to the quality/reliability of data – how much can you trust it?

- **Data cleansing methods are used to improve quality**
- **While there are tools to help automate data preparation and cleansing, they are still in the pre-industrial age.**
  - As a result, organizations must now analyze both structured and unstructured data that is uncertain and imprecise.
- *Systematic problems with sensors; errors causing data loss; etc.*

Types of Data: View (I)

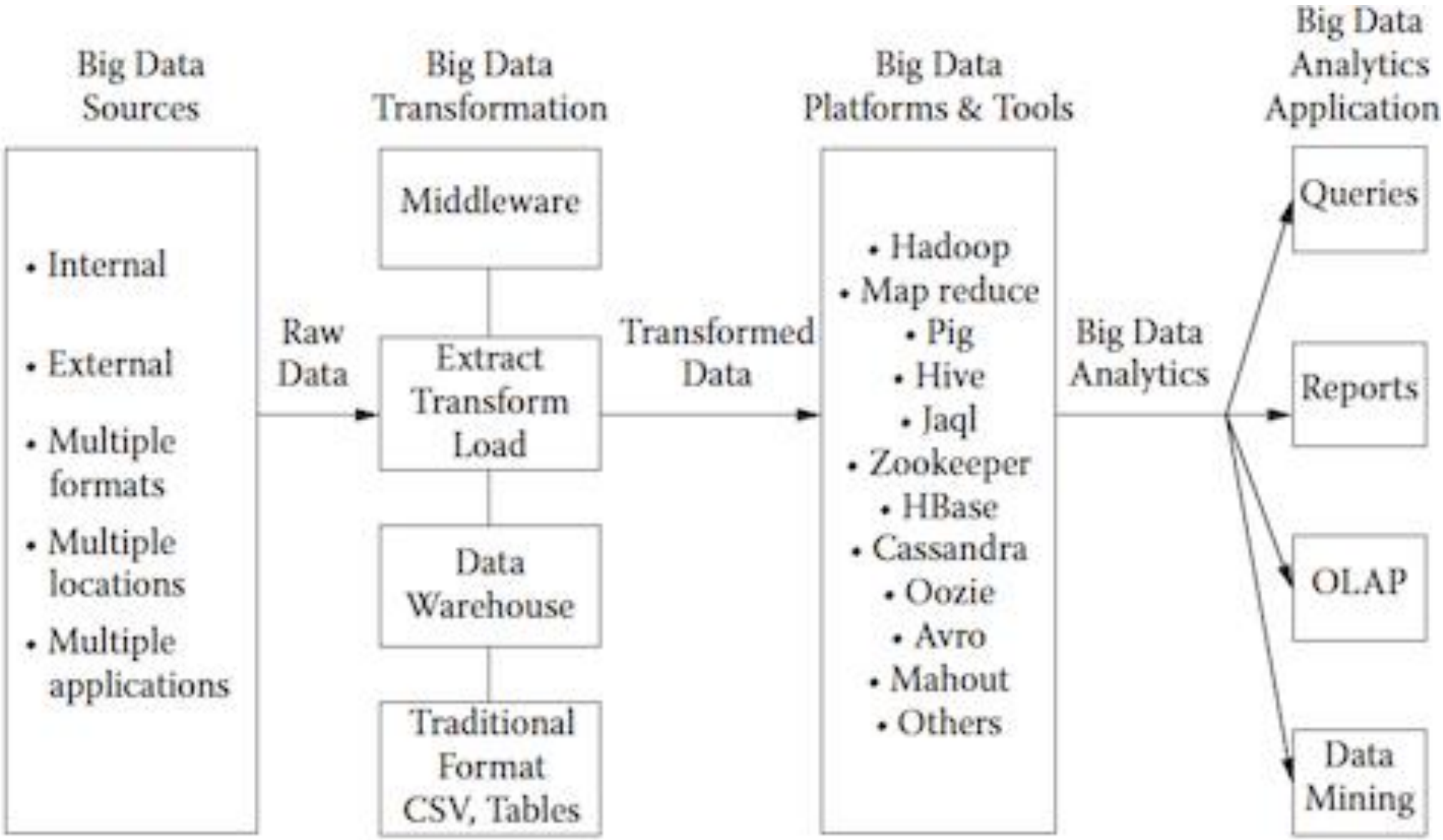


Types of Data: View (II)



***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Big Data Architecture



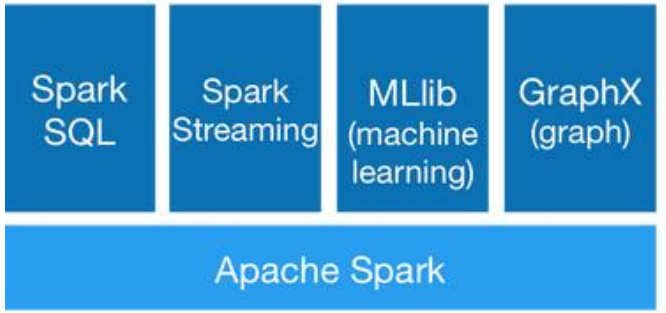
Big Data Platforms: Apache Hadoop

- The **Apache Hadoop** software library is a framework that allows distributed processing of large data sets across clusters of computers.

- Designed to scale up from single servers to thousands of machines,
  - Each machine offering local computation and storage.

http://hadoop.apache.org

Big Data Platforms: Spark

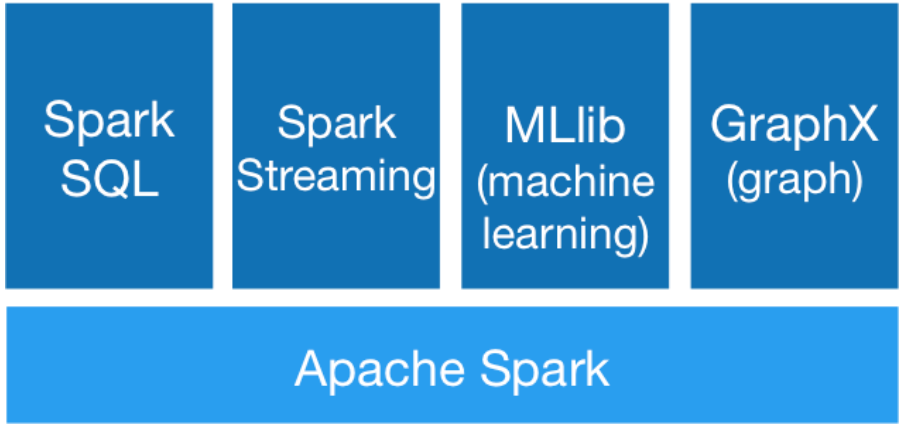


- Apache Spark is a cluster computing platform designed to be fast and general-purpose for large-scale data processing.
  - Spark offers the ability to run computations in memory, that increases the processing speed of an application.

- It is designed to cover a wide range of workloads such as batch applications, iterative algorithms, interactive queries, and stream processing, that previously required separate distributed systems

  - Spark makes it easy to combine different processing types.

Big Data Platforms (e.g. Spark) provides generality:



- Combine SQL, streaming, and complex analytics.
- Spark powers a stack of libraries including:
  - SQL for working with structured data.
  - MLlib for machine learning,
  - GraphX API for graphs and graph-parallel computation, and
  - Spark Streaming for building streaming applications.
- You can combine these libraries seamlessly in the same application.

Runs Everywhere

Spark run in its standalone cluster mode, on EC2, on Hadoop YARN, on Mesos, or on Kubernetes.

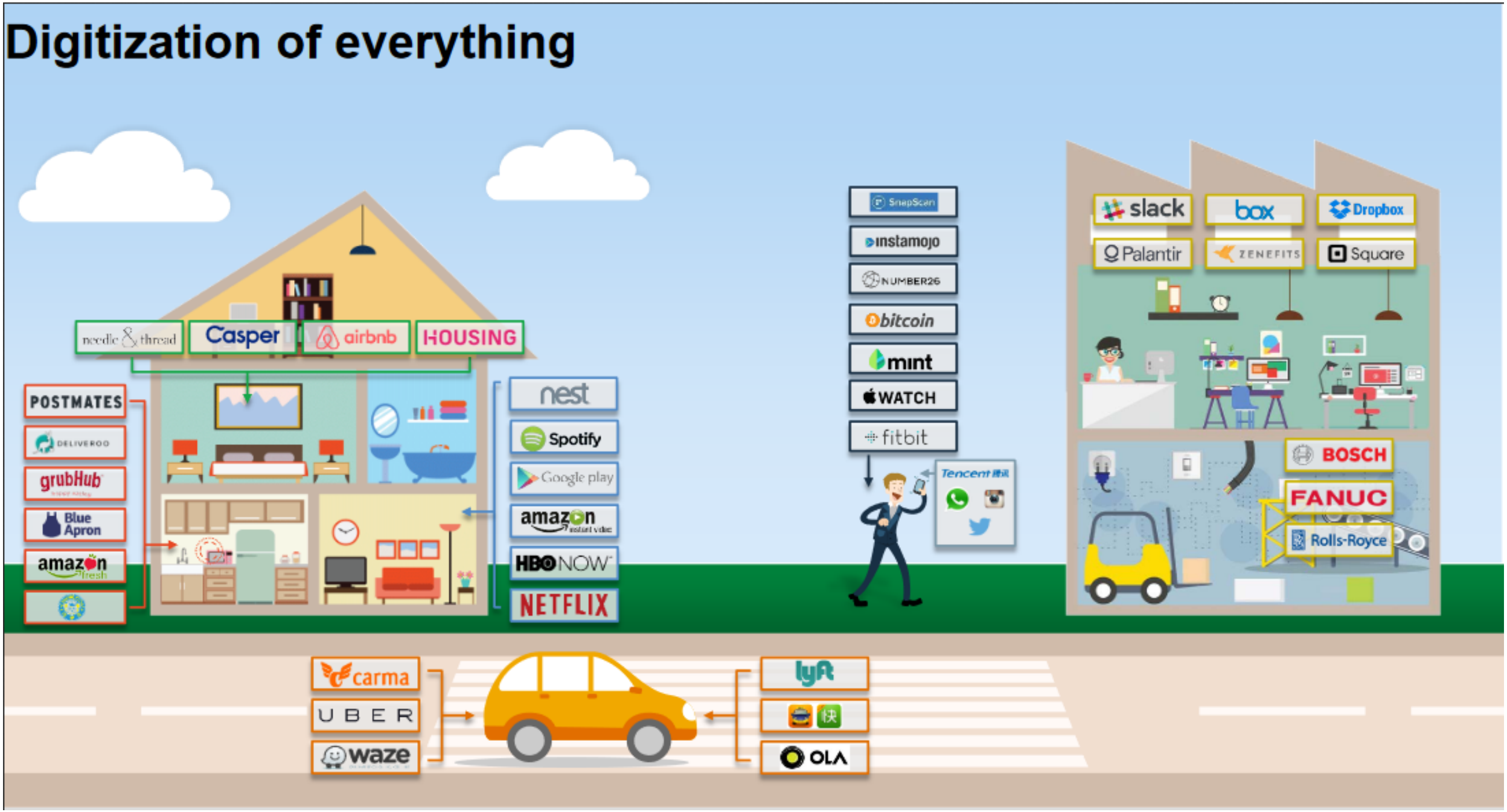
Spark access data from diverse data sources such as HDFS, Alluxio, Apache Cassandra, Apache HBase, Apache Hive, and hundreds of other data sources.

Digital Economy

- Digital Economy refers to an economy that is based on digital technologies [1].
- Economic processes, transactions, interactions and activities that are based on digital technologies [2].
- 'Digital Economy' was coined by Don Tapscott's 1995 in the best seller book: The Digital Economy.

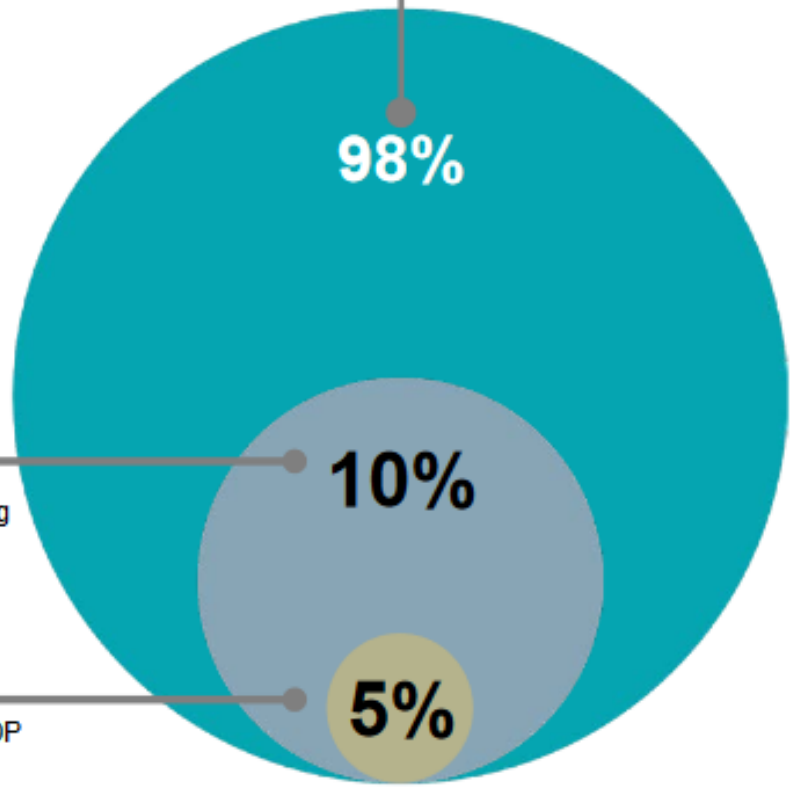
Sources:
1. https://www.techopedia.com/
2. https://en.wikipedia.org/wiki/Digital_economy

Digitization of everything



Size of the Digital Economy

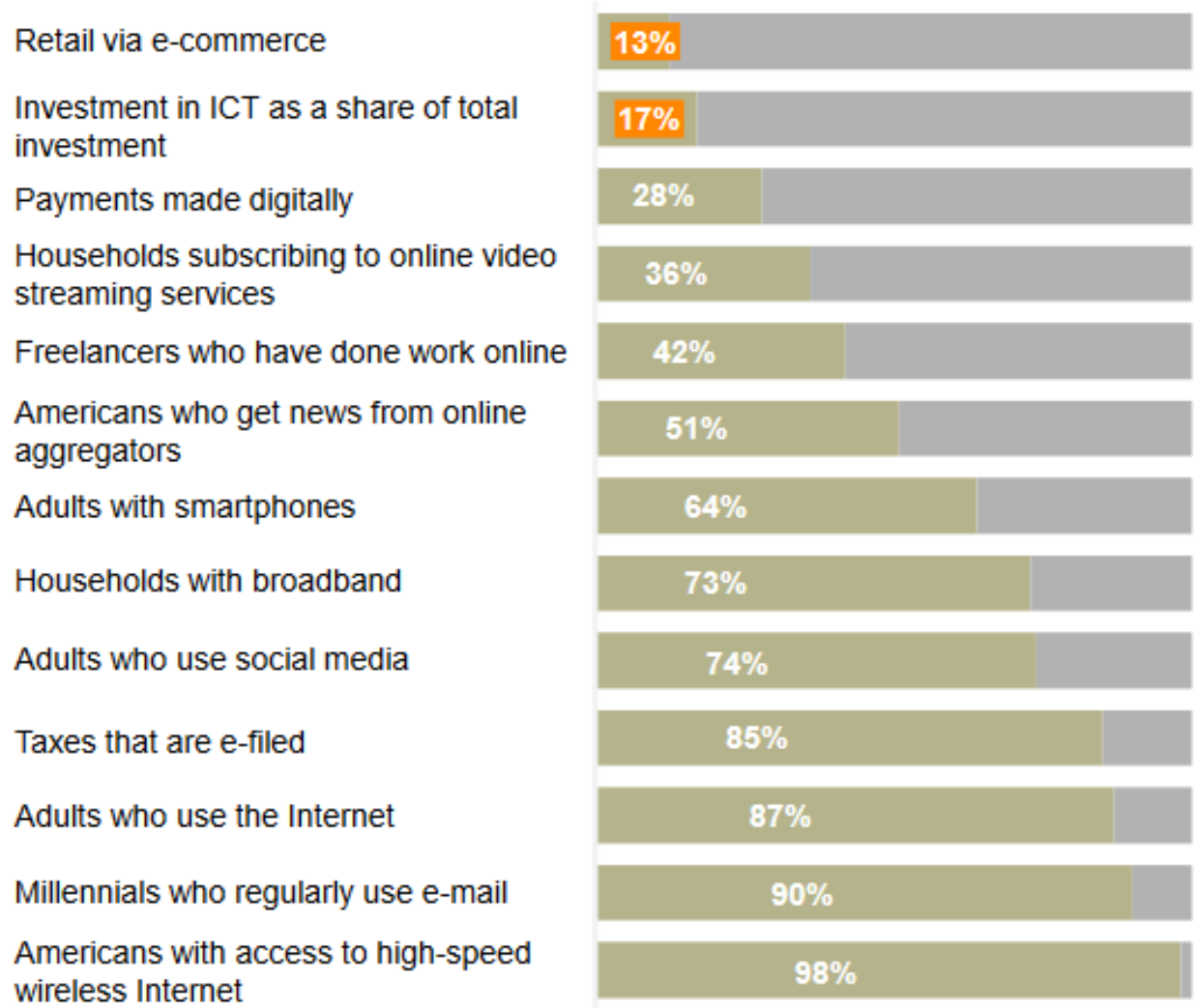
Percent of US economy impacted by digitization



10%
ICT as a share of GDP, taking into account price effects

5%
ICT Sectors as a share of GDP in official statistics

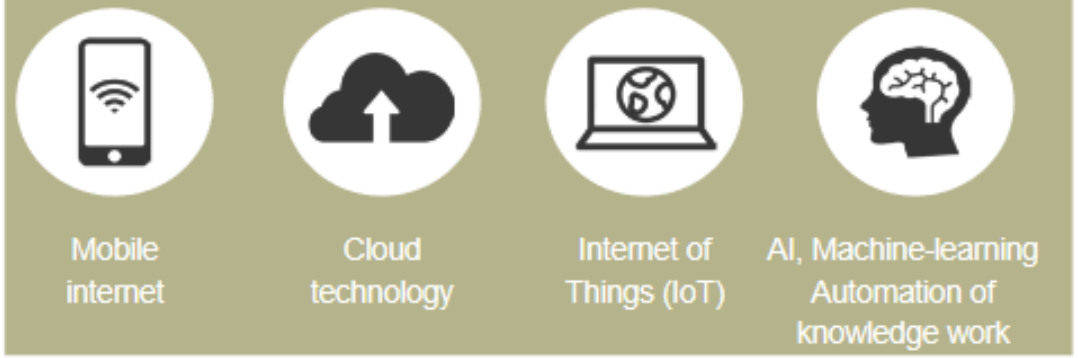
Up to 98% Examples



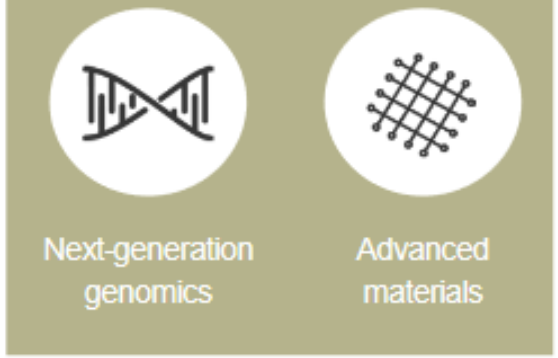
More to come as more technologies (mostly digital) continue to transform the economy

Disruptive Dozen – Selected for scope and scale over next 10 years

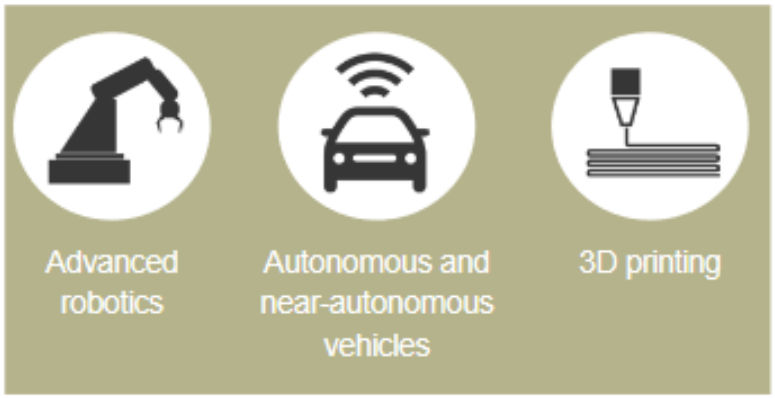
IT and how we use it



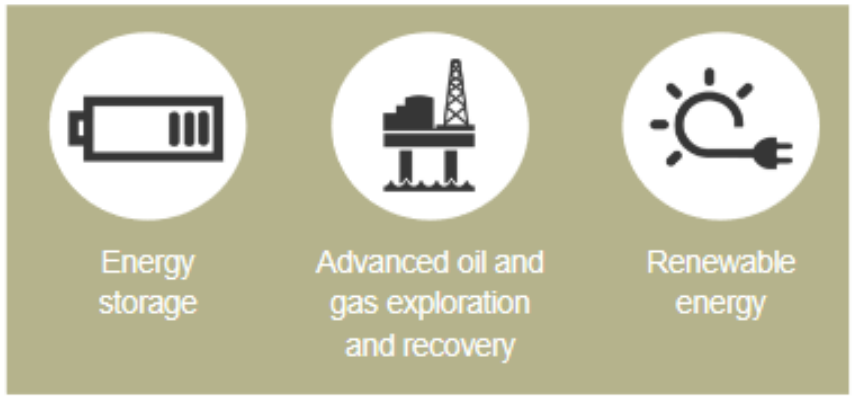
New building blocks



Machines at work



Rethinking energy comes of age



More…

- Blockchain/ Cryptocurrencies
- Human-Computer Interfaces (AR, VR etc)
- Digital IDs/Biometrics
- Quantum Computing
- …

SOURCE: McKinsey Global Institute analysis

Digital Economy Scope

- Research shows that a 10 basis-point improvement in three levers:
  - digital skills,
  - digital technologies and
  - digital accelerators

This will add 2-3 % to the 2020 GDP in the countries surveyed [1].

- Businesses that fail to get digitally connected will become excluded from the global market.

- Huge potential of the digital economy is underexploited, with 59% of enterprises being non-digital in Europe
  - Only 2% taking full advantage of digital opportunities.

1. Digital Disruption: the Growth Multiplier, Accenture Strategy (2016)

Big Data Analytics Scope in Digital Economy

- Companies leverage data to adapt products and services to:
  - Meet customer needs
  - Optimize operations
  - Optimize infrastructure
  - Find new sources of revenue
  - Can reveal more patterns and anomalies

To support big data project:

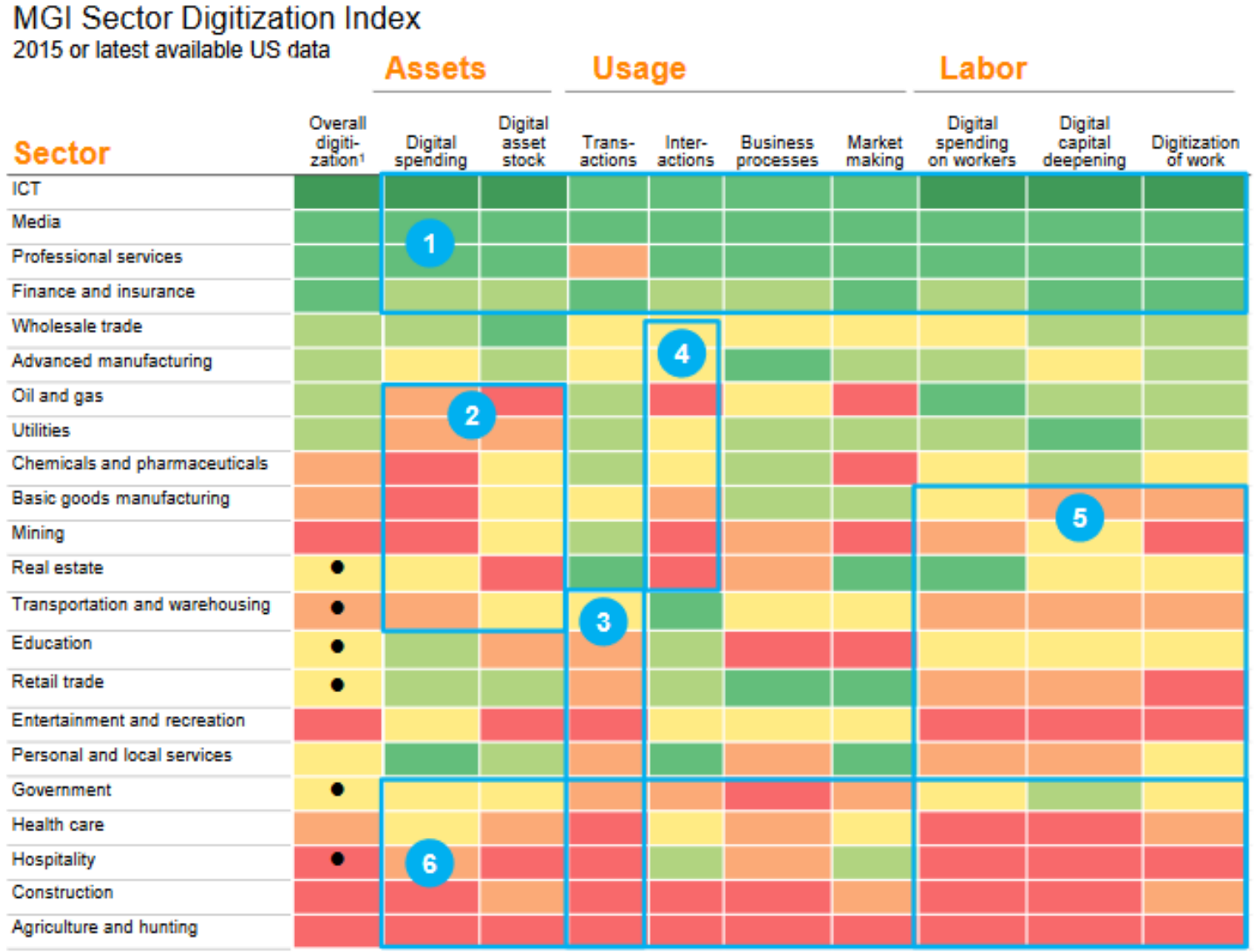
IBM estimated that 4.4 million jobs were globally created by 2015, where 1.9 million of these jobs were in the United States

Three opportunities and challenges

1. Digitization for growth, productivity and innovation
2. Digitization of Globalization
3. Digitization of work

Foundations to get right

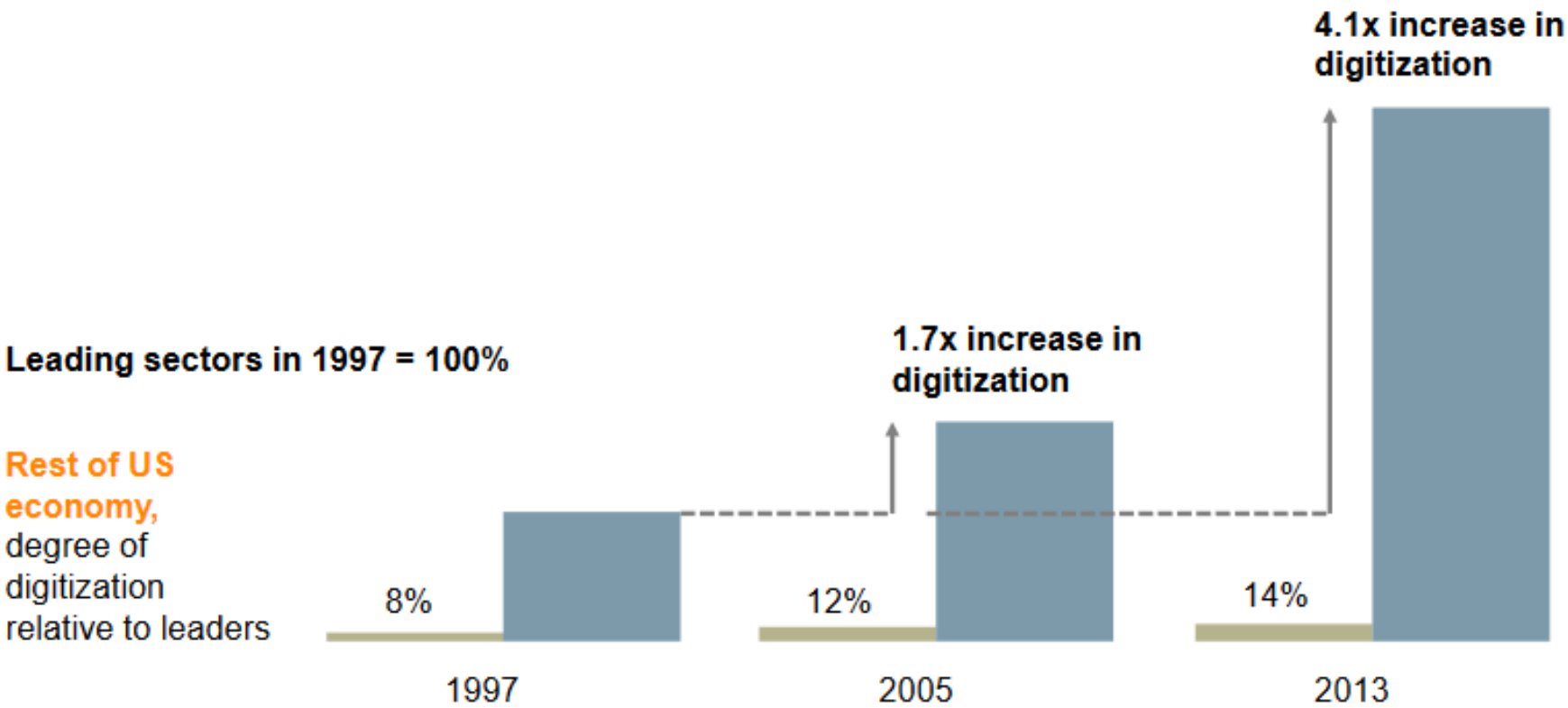
Extent of Digitization varies by sector



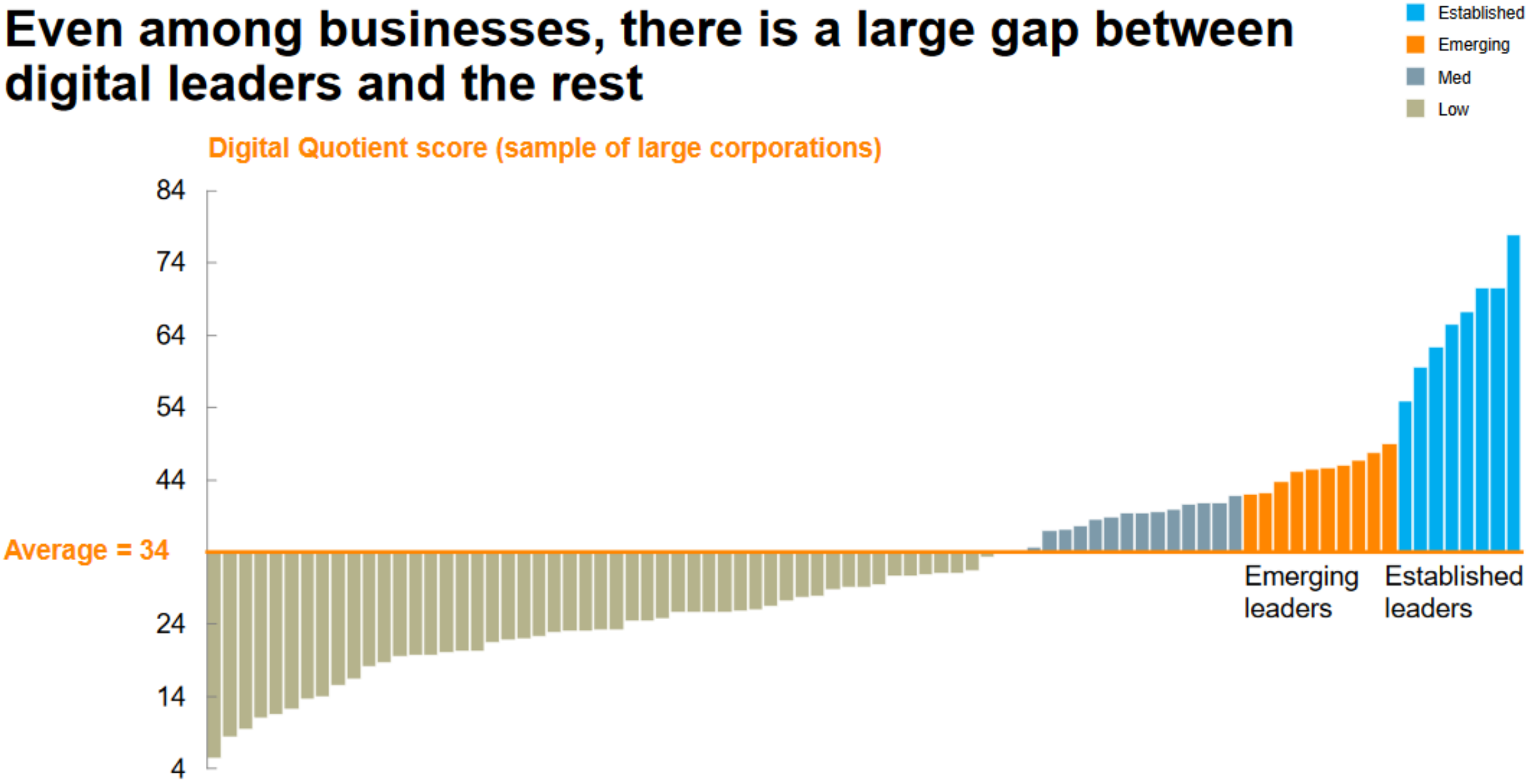
1. Knowledge-intensive sectors, highly digitized
2. Capital-intensive, potential to further digitize their assets
3. Service sectors with long tail of small firms having room to digitize customer transactions
4. B2B sectors with the potential to digitally engage and interact with their customers and users
5. Labor-intensive sectors with the potential to provide digital tools and skills to their workforce
6. Large, localized, low productivity could transform for productivity and delivery of services

SOURCE: BEA; BLS; US Census; IDC; Gartner; McKinsey social technology survey; McKinsey Payments Map; LiveChat customer satisfaction report; Appbrain; US contact center decision-makers guide; eMarketer; Bluewolf; Computer Economics; industry expert interviews; McKinsey Global Institute analysis

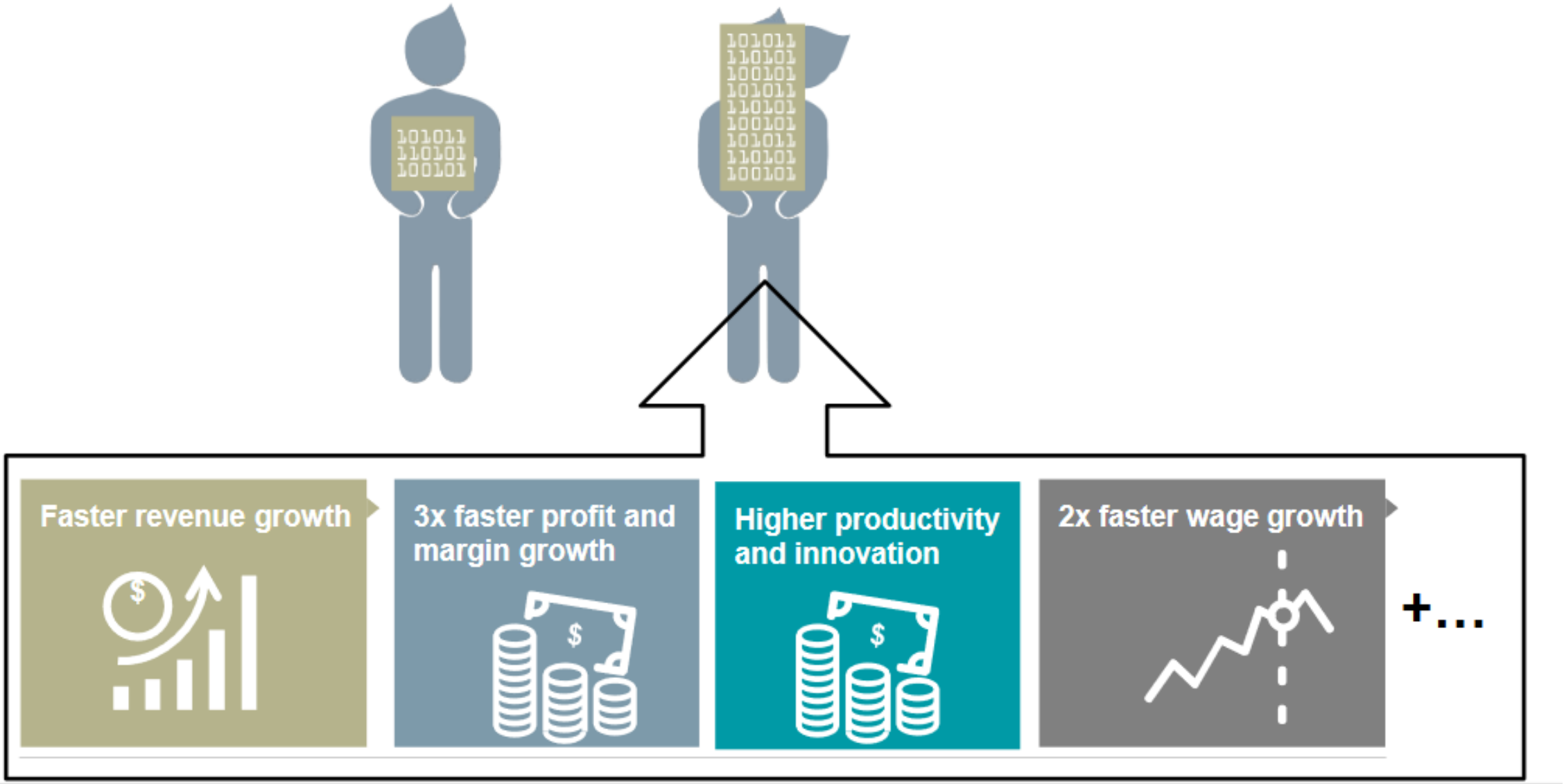
Gap widening between the most digitized and the rest



Even among businesses, there is a large gap between digital leaders and the rest



Digital “haves” and “have-mores”



Three opportunities and challenges

1. Digitization for growth, productivity and transformational innovation
2. Digitization of Globalization
3. Digitization of work

Foundations to get right

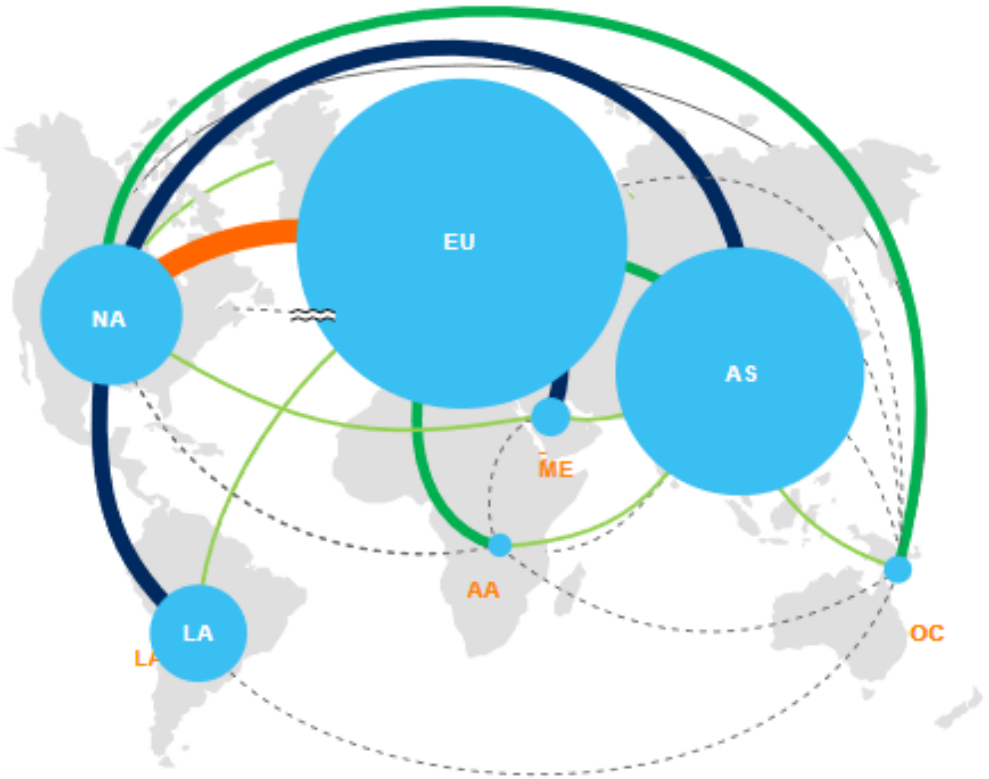
Cross-border data flows surging and connecting more countries, companies and people

Size of bubble represents intraregional used bandwidth

Used cross-border bandwidth

2005[1]

100% = 211.3 tbps



45x

Regions	Bandwidth Gigabits per second (Gbps)
NA United States and Canada	<50
EU Europe	50–100
AS Asia	100–500
LA Latin America	500–1,000
ME Middle East	1,000–5,000
AA Africa	5,000–20,000
OC Oceania	>20,000

In the 21st century, globalization will be increasingly defined by flows of data and information. Cross-border data flows are rising, a 45 times increase between 2005 and 2014, now topping global flows of trade and finance.

1. Digital Globalization : the new era of digital flows, McKinsey Global Institute (2016)

Individuals engaging globally, increasingly enabled by digital

Social networking users with at least one foreign connection
914 million

International travelers
429 million

Cross-border e-commerce shoppers
361 million

People living outside home country
244 million

Cross-border online workers
44 million

Cross-border online students
13 million

Students studying abroad
5 million

NOTE: Numbers adjusted to account for overlap between platforms and for individuals making multiple international trips in the same year.

SOURCE: Facebook; AliResearch; US Department of Commerce; OECD; World Bank; McKinsey Global Institute analysis

Three opportunities and challenges

1. Digitization for growth, productivity and innovation
2. Digitization of Globalization
3. Digitization of work

Foundations to get right

Labor Market Platforms Emerging

	Examples
Platforms matching individuals with jobs	monster [7,999 job searches/minute]; LinkedIn [364 million members]; careerbuilder EMPOWERING EMPLOYMENT [24 million visitors/month]; SKILLFUL A MARKLE INITIATIVE
Platforms for "OnDemand" work	Upwork [12.5 million users]; UBER [1 million drivers]; TaskRabbit [25,000 service providers]; samasource [6,500 workers]; freelancer.com [14,300,000 users]
Talent management tools for recruiting and training	GOOD.CO; pymetrics beta; reviewsnap Web-based Performance Management; PayScale

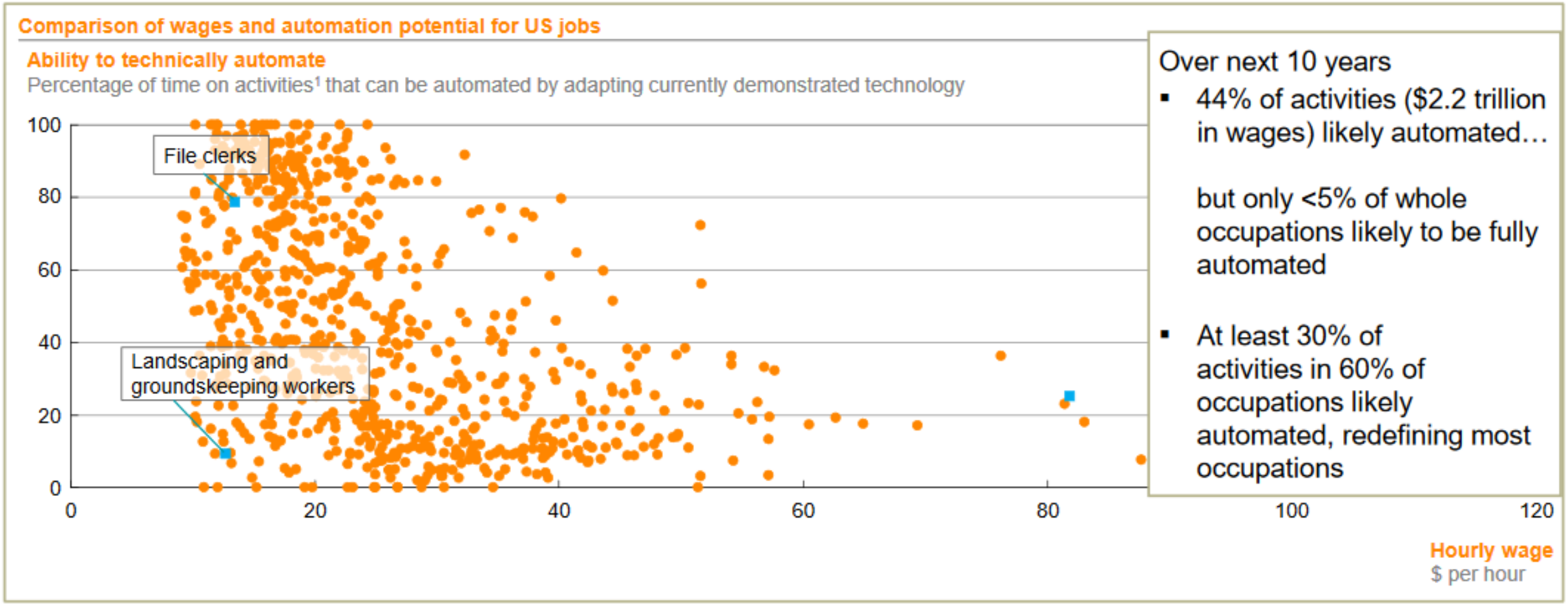
1 All companies are illustrative and not exhaustive, many span multiple functions

Labor Market Platforms Emerging

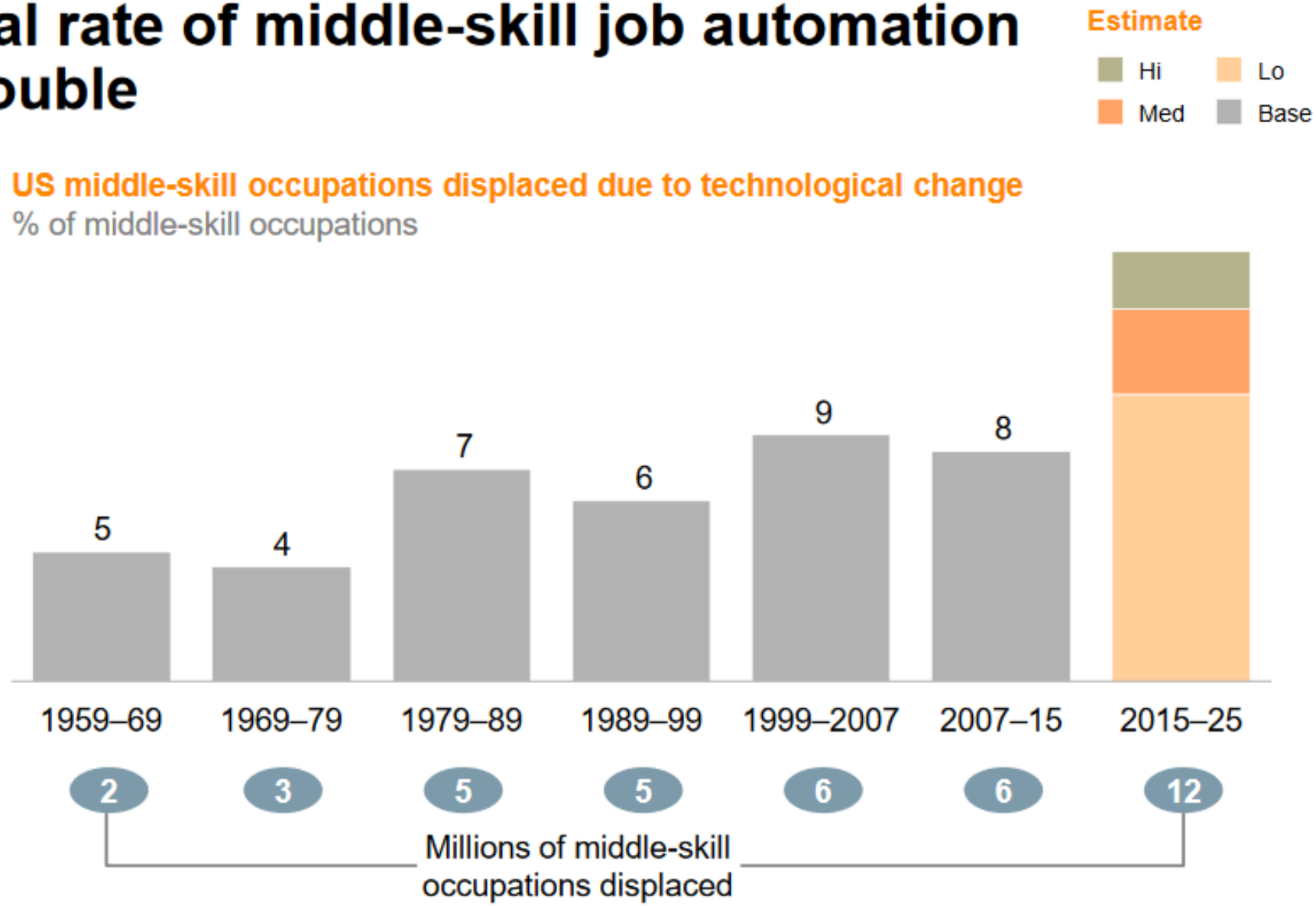
- Just engaging with customers online create growth
- Figures have shown that SMEs from many countries that have actively engaged with consumers on the internet have experienced sales growth rates that are up to 22 % higher over three years than those companies in countries with low or no internet presence.

- By not taking full advantage of digital technologies, EU businesses miss out on the chance to expand and create jobs.
- It is estimated that if all EU countries mirrored the performance of the USA or the best-performing EU countries, 400,000 to 1.5 million new jobs could be created in the EU internet economy.

Automation potential in high-wage, high-skill occupations as well as low-wage, low-skill occupations



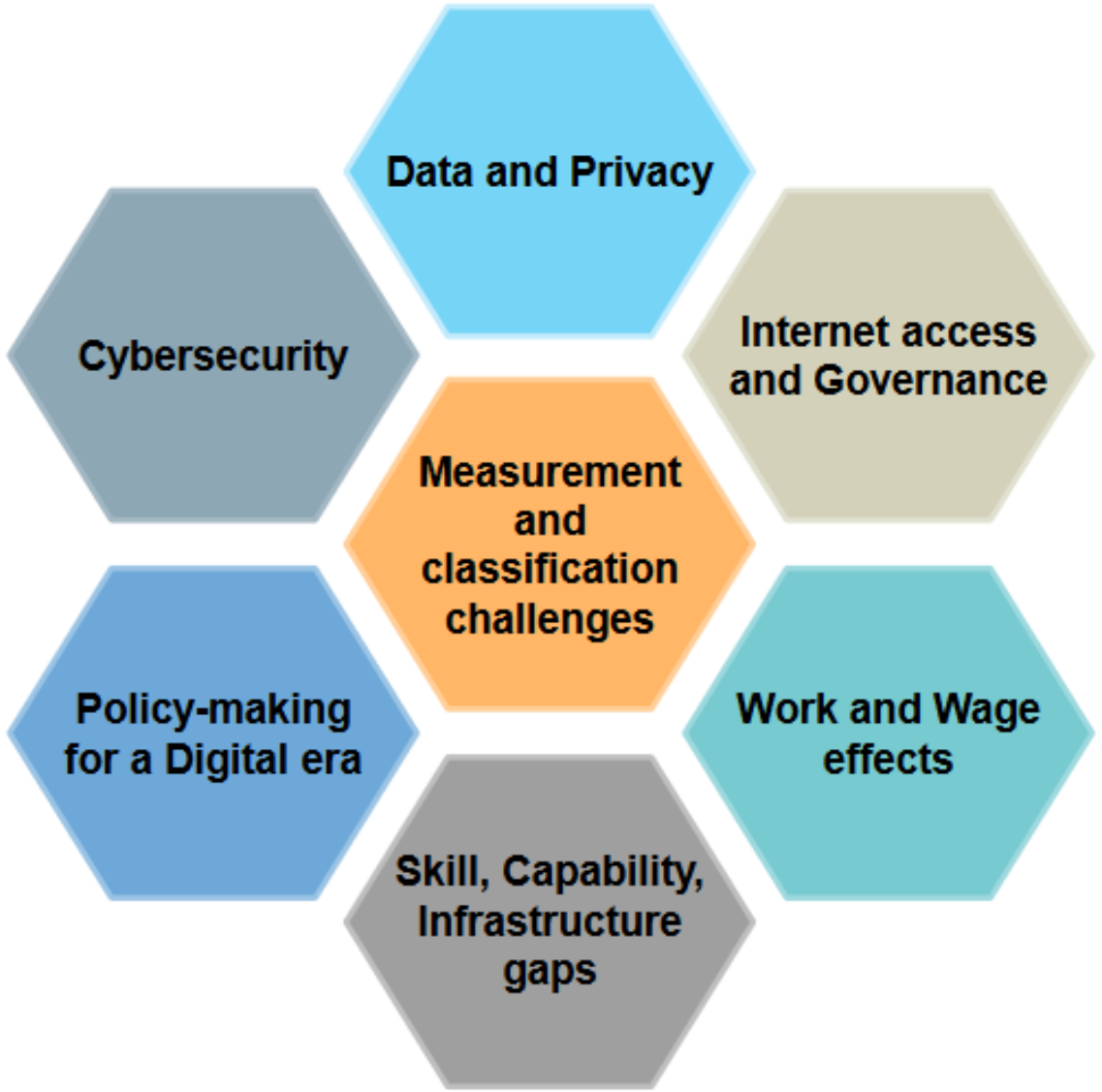
Historical rate of middle-skill job automation could double



Foundations to get right

To fully capture benefits to:

- Economy (growth, productivity, competitiveness)
- Businesses and entrepreneurs
- Individuals (workers, consumers, citizens)
- Society

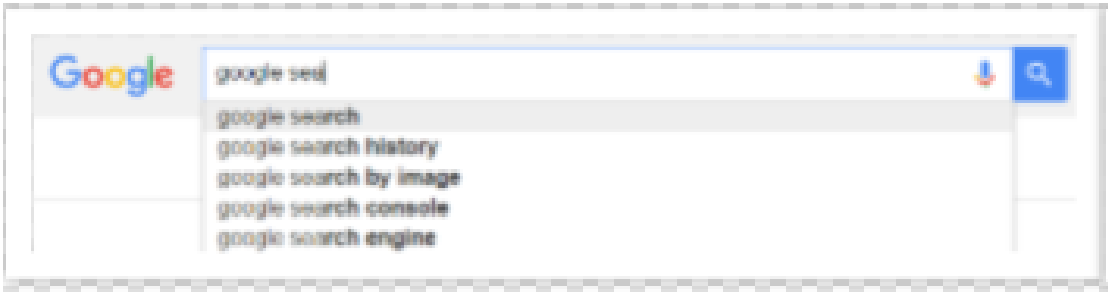


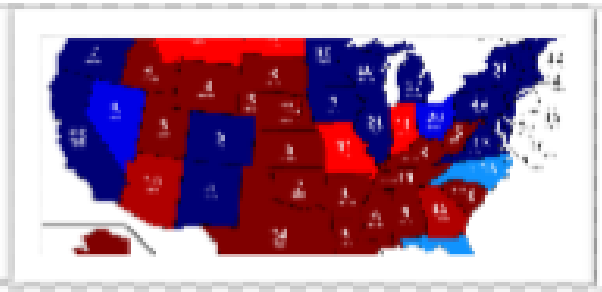
Role of Big Data in Digital Economy: Text Processing

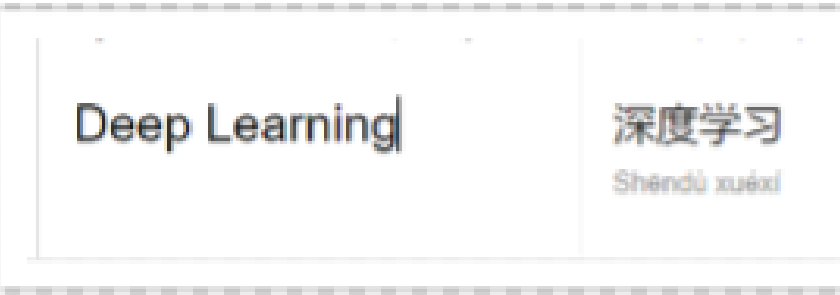
- Open Google and search for a news article on the ongoing World Cup and get hundreds of search results in return about it.

- Nate Silver analyzed millions of tweets and correctly predicted the results of 49 out of 50 states in 2008 U.S Presidential Elections.

- You type a sentence in google translate in English and get an Equivalent Chinese conversion.

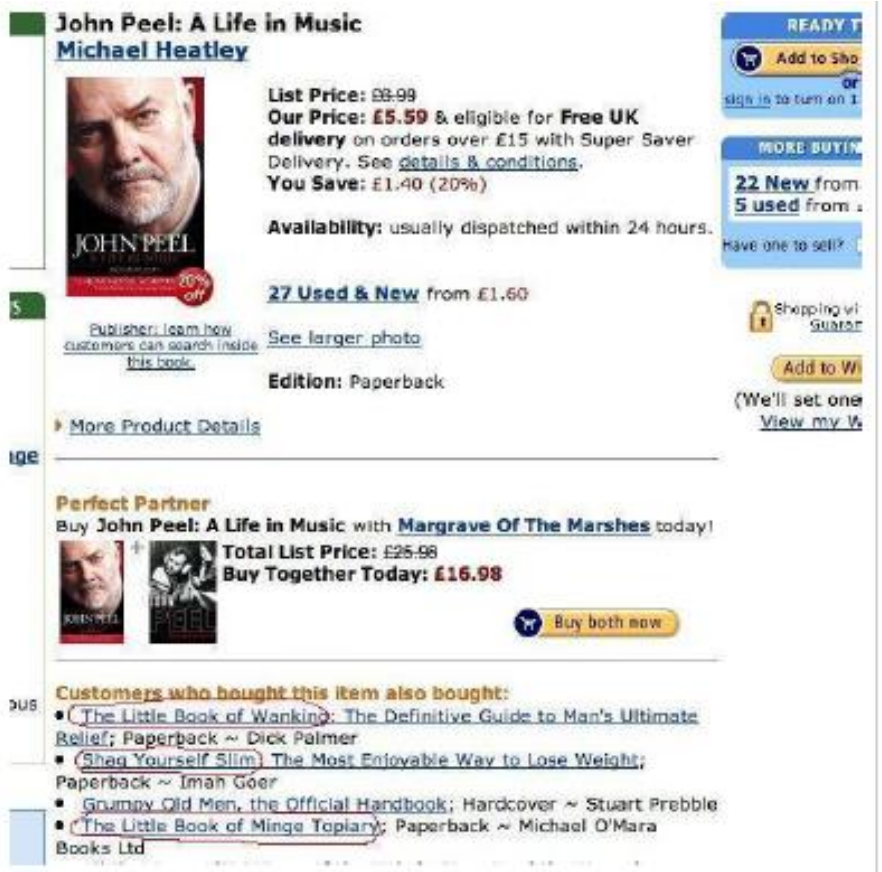






Role of Big Data in Digital Economy: Recommendation Systems

- Netflix: 2/3 of the movies watched are recommended
- Google News: recommendations generate 38% more clickthrough
- Amazon: 35% sales from recommendations
- Choicestream: 28% of the people would buy more music if they found what they liked.



Based on:

- **Past behavior**
- **Relations to other users**
- **Item similarity**
- **Context . . .**

Role of Big Data in Digital Economy: Container Flows

Big data is used to support port operators and liners strategic planning.

- Chuck et al. used Artificial Neural Networks to Predict Container Flows between the Major Ports of Asia', by considering
  - GDP,
  - interest rates,
  - the value of export and import trade,
  - the numbers of export and import containers and
  - the number of quay cranes.
- The forecasting results indicate that the prediction errors are relatively small in most selected ports
- Shipping companies use the container flow prediction model to improve strategic planning.

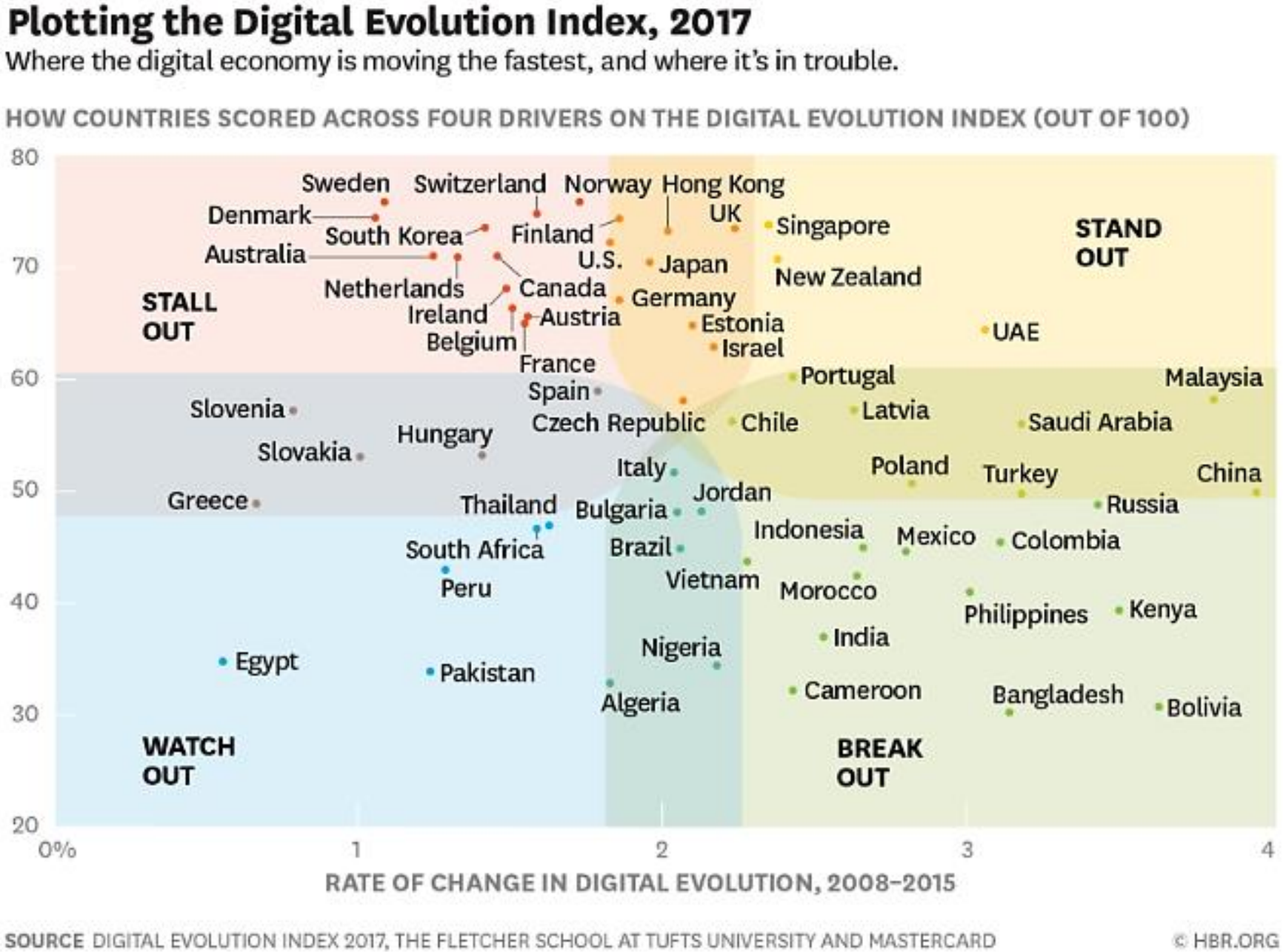
Role of Big Data in Digital Economy: Global manufacturing and Supply Chain

- Manyika et al. [1], pointed out that the big data analytics can be helpful to support global manufacturing and supply chain innovation by creating data transparency, improving human decision-making and promoting innovative business models.
- However, there is a lack of data analytics techniques available to help decision-makers and practitioners to capture and harvest the potential value of data
- Thus, a data analytic infrastructure that helps decision-makers to make use of the high volume of data to serve as inputs for decision-making is necessary.

1. 2011 Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. 2011. "Big Data: The Next Frontier for Innovation, Competition, and Productivity."
2. Tan et al. 2015 Tan, K. H., Y. Zhan, J. Guojun, Y. Fei, and C. Chingter. 2015. "Harvesting Big Data to Enhance Supply Chain Innovation Capabilities: An Analytic Infrastructure Based on Deduction Graph." *International Journal of Production Economics* 165: 223–233.

Role of Big Data in Digital Economy: Arrival timings of trucks

- Chintan Amrit et al. used big data model to predict the arrival times of 230 trucks using real traffic and weather data related to a distribution centre in Europe.
- The experimental results show that, while a big data approach delivers valuable insights.
- Factors, such as human or organizational factors, could influence the arrival time; therefore, these factors should be taken into consideration in future predictive models.



https://hbr.org/2017/07/60-countries-digital-competitiveness-indexed

Big data management challenges:

Challenge 1: Data quality

- Big data must be cleaned, secured, vetted for compliance and continuously maintained.
- Since data comes in so fast, companies find it difficult to perform all of the data preparation steps to ensure optimum data quality.
- In some cases, organizations simply store all of their incoming big data without doing much to it.

- This creates data pollution.
- Plus, inaccurate data can raise the risk of business decisions being based on erroneous information.

Big data management challenges

Challenge 2: Platform integration

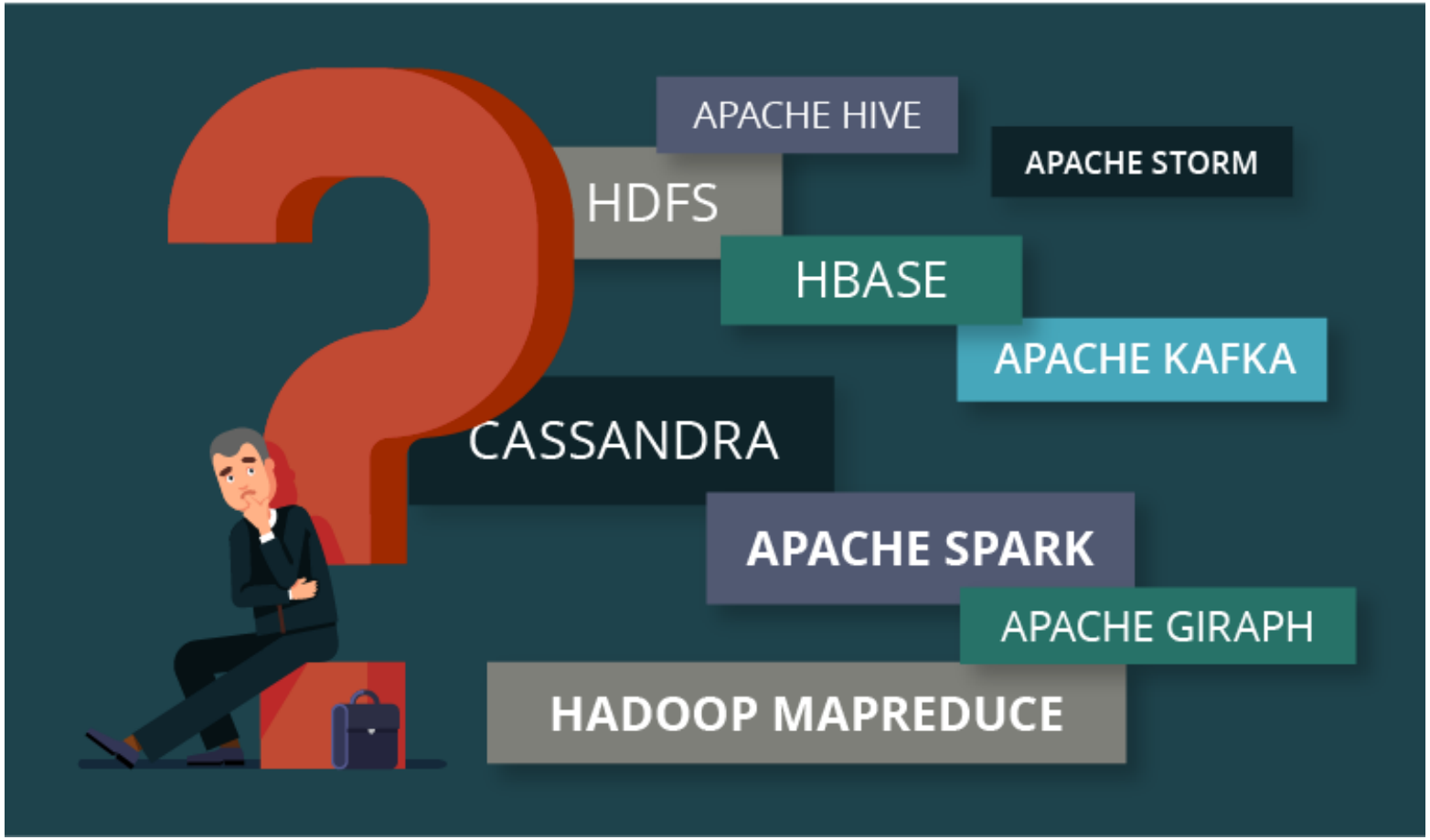
- Big data integration centers around integrating data from different business departments into a "single version of the truth" that everyone in the business can use.
- However, it is just as challenging for IT to manage big data that comes in all flavors and on many different hardware and software platforms.
- "There are a plethora of backend distributed data stores, " said Mansour Raad, senior software architect at ESRI.
- "Some of these distributed data stores are not natively supported by [our] platform....
- Depending on the data store different APIs are used to handle these situations.
  - It's not optimal.
- Accessing and storing data in unsupported data stores requires developers to constantly change their program for each data store.
- This slows development cycles and makes it much longer for customers to get insights from the data."

Big data management challenges

- **Challenge 3: Access and security**
- Who gets access to which data, and at what level of permission? For example, a document management system contains text-based documents, photos, images, drawings, and videos. Who has access to what documents and who has the right to modify the documents?
- **Solution**
- This is a policy (and sometimes political) question. It must be resolved in a sit-down meeting between IT and end users to determine who should gain access.

Challenge 4: Confusing variety of big data technologies

- It can be easy to get lost in the variety of big data technologies now available on the market.
- Do you need Spark or would the speeds of Hadoop MapReduce be enough?
- Is it better to store data in Cassandra or HBase?
- If you are exploring the ocean of technological opportunities without a clear view of what you need.



Cloud Computing:

Enabling successful Digital Transformations and Big Data Analytics

Enablers and Challenges: Digital Transformation

- Enablers
  - Agility
  - Cost and Labor Effectiveness
  - Security
  - Rapid Prototyping

- Challenges
  - Data Security and Service Quality
  - Performance and Costs
  - Migrating and integrating Existing Systems

https://hashedin.com/blog/cloud-computing-digital-transformation/

Challenges: Big Data and Big Data Analytics

- Data collection and growth
- Integration and synchronization b/w different data sources
- Preparing and validating data
- Security and Privacy of data

- Generating timely insights through Big Data Analytics

- Produce and retain new talent (recruitment, training, etc.)
- Change management as a result of innovative technologies

Combined challenges (1)

In summary, we need

- A scalable and elastic computing platform
- Global accessibility to connect people, application and devices
- Transparent operations - hiding technicality/location from users

Combined challenges (2)

Contd.

- Ease of interfacing with platform functionality and integration with existing applications

- Low initial costs (Pay-As-You-Go model ~ Utility Computing) to allow Small and Medium Enterprise (SMEs), researchers, and students.

Combined challenges (3)

Contd.

- Allow prototyping and production deployment at the same time
- Addresses Security and Privacy through industry standards
- Provides service assurance and interoperability with other providers.

What is Cloud Computing?

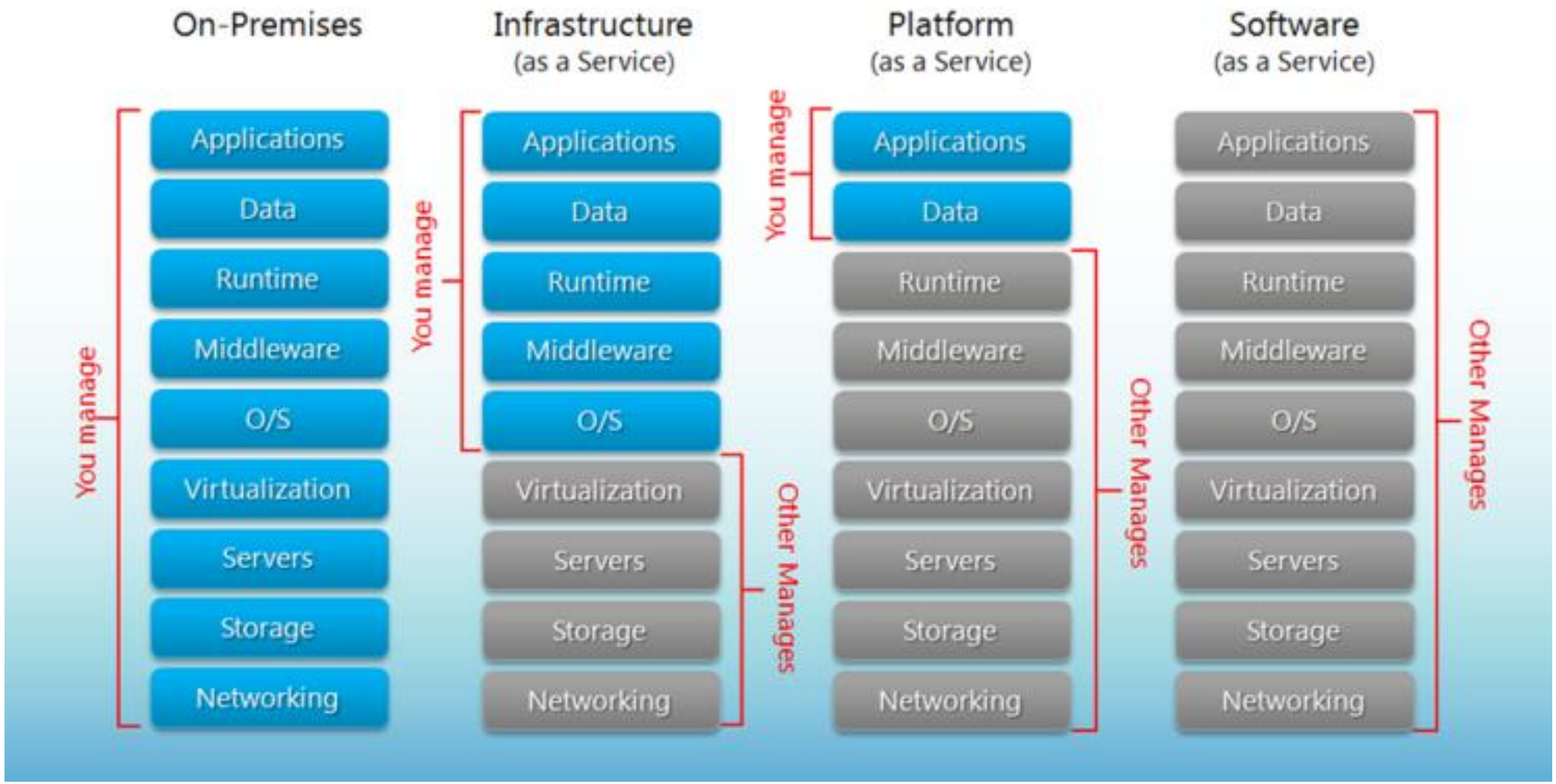
- Cloud Computing (or Cloud) refers to the use of resources - Compute, Storage, Networking, which are provisioned on-demand and managed by a Cloud OS either locally (in-house) or remotely (vendor).

- The end-uses are concerns with use of the resources and are usually charged by the hour (Pay-As-You-Go).

Cloud Service Models

- Hardware Infrastructure (IaaS – Infrastructure as a Service)
- Middleware Solutions (PaaS – Platform as a Service)
- Full software application Suites (SaaS - Software as a Service) e.g. Salesforce.com, Microsoft Office 365, Google Apps

Separation of Responsibilities

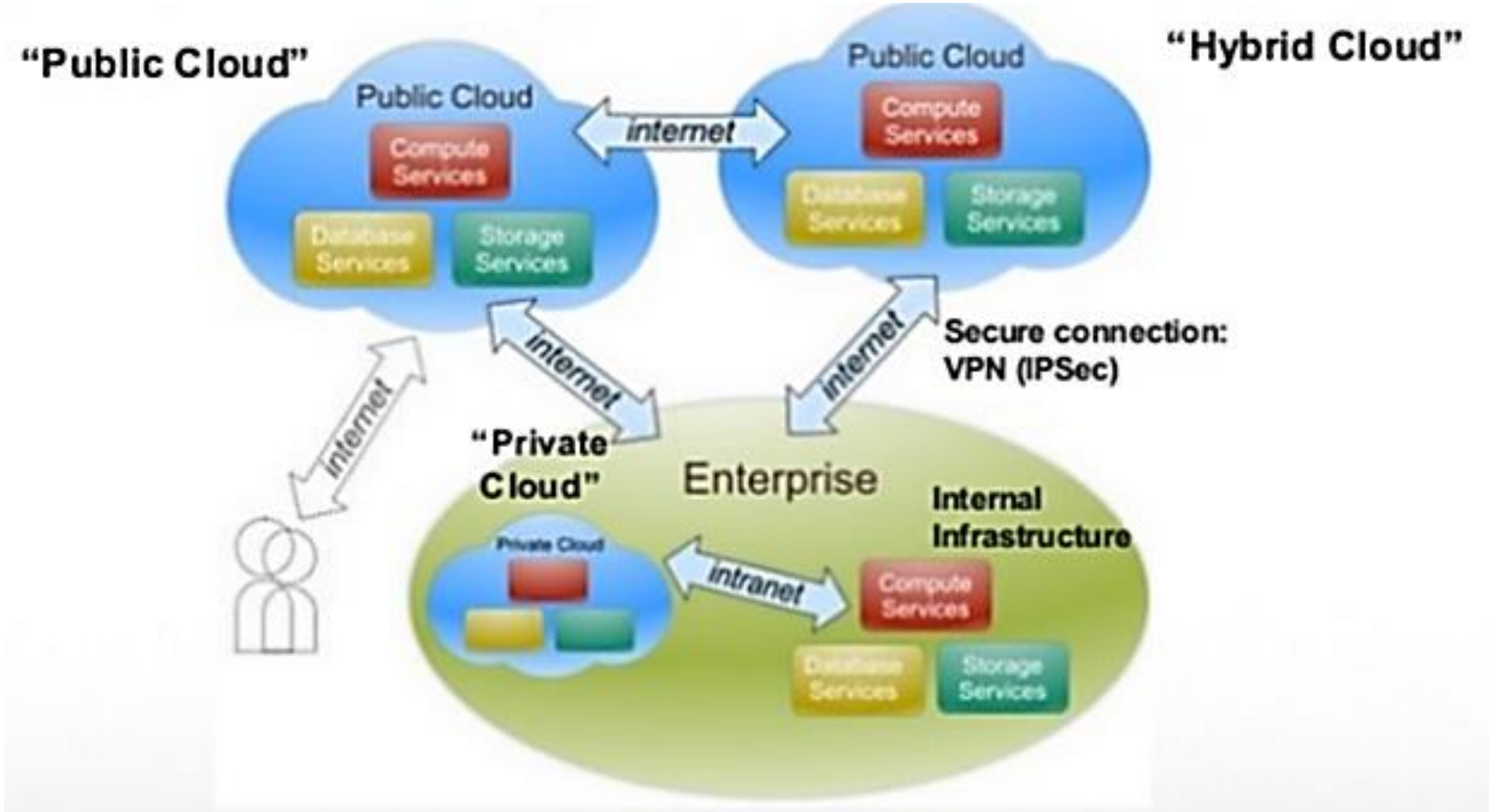


http://robertgreiner.com/2014/03/windows-azure-iaas-paas-saas-overview/

Cloud Deployment Models

- Cloud are deployed as

1. Private clouds managed locally by corporation and governments
2. Public clouds by vendor (e.g. Amazon, Microsoft, IBM, Oracle, etc.)
3. Hybrid clouds – where private clouds consume resources on-demand in public clouds



https://blogs.vmware.com/apps/2016/03/bca_top_ten_cloud_part1.html

Major Cloud Providers

- Amazon – Amazon Web Services
- Microsoft – Azure
- Google – Google Cloud Platform
- IBM – IBM Cloud
- Oracle – Oracle Cloud
- others

Digital Transformation using Cloud

- 79% of respondents say that Cloud is important/very important to it.
- 38 per cent of Cloud users say that Cloud has given their organizations a significant competitive advantage.
- 26 per cent average saving when using Cloud.
- 43 per cent report that the intention is to achieve better use of data and analytics.
- 30 per cent report it is to improve innovation abilities.

https://www.cloudindustryforum.org/content/cloud-fundamental-digital-transformation-strategies

Thanks

Q & A

Backup Slides

Challenge #3: Paying loads of money

Big data adoption projects entail lots of expenses. If you opt for an on-premises solution, you'll have to mind the costs of new ***hardware***, new ***hires*** (administrators and developers), ***electricity*** and so on. Plus: although the needed frameworks are open-source, you'll still need to pay for the development, setup, configuration and maintenance of new ***software***.
If you decide on a cloud-based big data solution, you'll still need to ***hire staff*** (as above) and pay for ***cloud services***, big data solution ***development*** as well as setup and maintenance of needed ***frameworks***.
Moreover, in both cases, you'll need to allow for future ***expansions*** to avoid big data growth getting out of hand and costing you a fortune.

- **Solution:**
- If you are new to the world of big data, trying to seek professional help would be the right way to go. You could hire an expert or turn to a vendor for big data consulting. In both cases, with joint efforts, you'll be able to work out a strategy and, based on that, choose the needed technology stack.

Challenge #4: Complexity of managing data quality

- **Data from diverse sources**
- The problem of data integration, since the data you need to analyze comes from diverse sources in a variety of different formats.
- For instance, ecommerce companies need to analyze data from website logs, call-centers, competitors' website 'scans' and social media. Data formats will obviously differ, and matching them can be problematic.
- For example, your solution has to know that skis named *SALOMON QST 92 17/18*, *Salomon QST 92 2017-18* and *Salomon QST 92 Skis 2018* are the same thing, while companies ScienceSoft and Sciencesoft are not.
- **Unreliable data**
- Nobody is hiding the fact that big data isn't 100% accurate. And all in all, it's not that critical. But it doesn't mean that you shouldn't at all control how reliable your data is.
- Not only can it contain wrong information, but also duplicate itself, as well as contain contradictions.
- And it's unlikely that data of extremely inferior quality can bring any useful insights or shiny opportunities to your precision-demanding business tasks.

- **Solution Platform integration**
- There are software automation tools available with hundreds of pre-developed APIs for a wide spectrum of data, databases, and files. You might still find yourself hand-developing an API on a case-by-case basis, but these tools can do a majority of the work.

Data veracity is the degree to which data is accurate, precise and trusted.

- **Biases:** An organization makes a decision using a calculated value that suffers from statistical bias.
- **Data Lineage:** An organization gets data from hundreds of sources. It discovers that one of the sources is extremely inaccurate but lacks the data lineage information to identify where the data has been stored in various databases.
- **Bugs:** A software bug causes data to be calculated or transformed incorrectly.
- **Noise:** A self-driving car needs to decide if a plastic bag being blown by the wind is a dangerous obstacle.
- **Abnormalities:** Two weather sensors in close proximity report dramatically different conditions.
- **Information Security:** An organization's data is changed by an advanced persistent threat.

Source: Internet

- **Solution:**
- Big data, being a huge change for a company, should be accepted by top management first and then down the ladder. To ensure big data understanding and acceptance at all levels, IT departments need to organize numerous trainings and workshops.
- To see to big data acceptance even more, the implementation and use of the new big data solution need to be monitored and controlled.

- Data quality **Solution**
- Define your business rules for data cleaning and preparation and seek out automation tools that can perform data prep tasks for you. Second, determine which data you absolutely don't need and establish data purging automation at the front of your data collection processes to jettison this data before it ever hits your network.

Information Overhead

- “People read around 10 MB worth of material a day, hear 400 MB a day, and see 1 MB of information every second” - The Economist, November 2006
- In 2015, consumption will raise to 74 GB a day - UCSD Study 2014