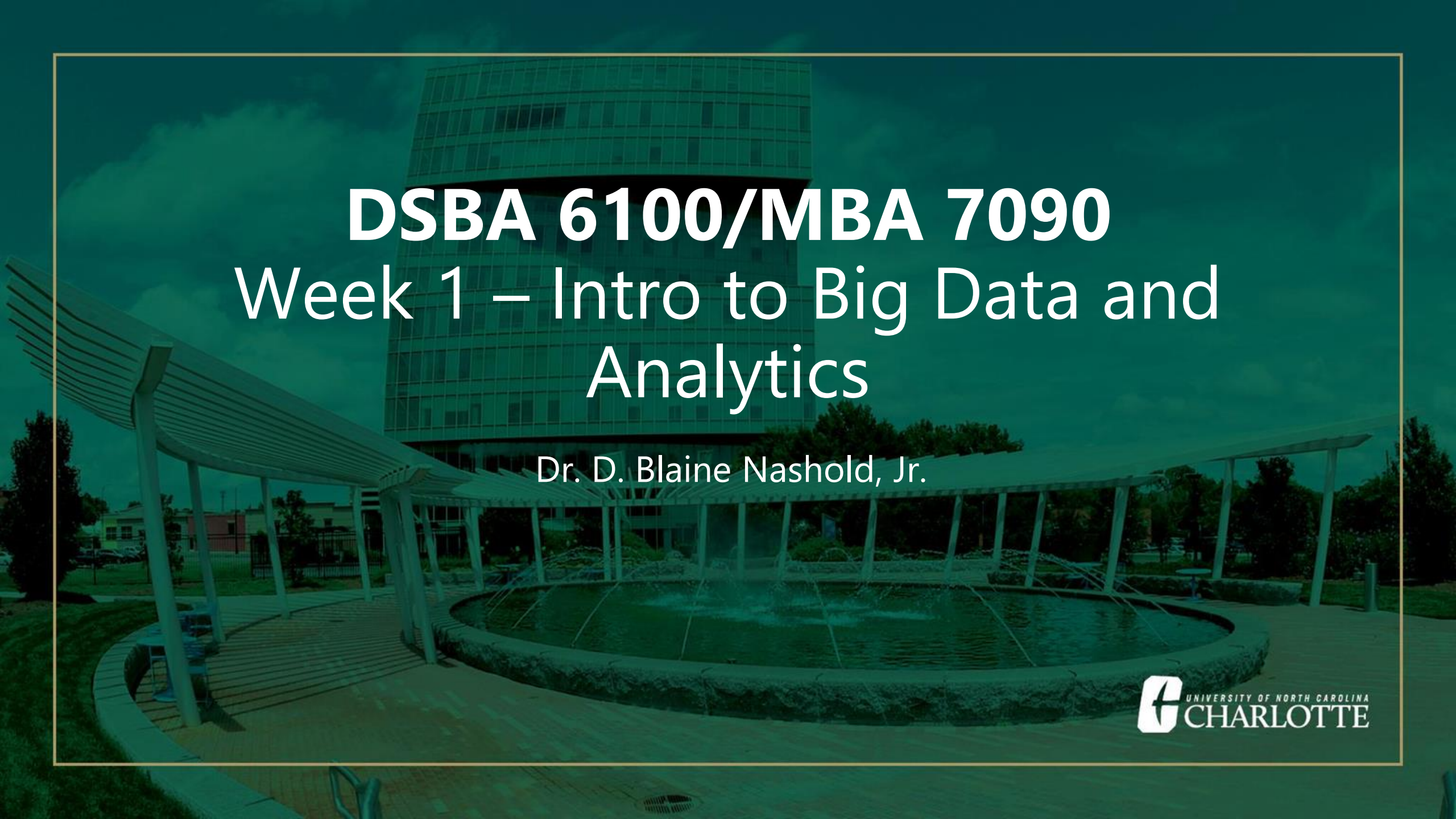


DSBA 6100/MBA 7090
Week 1 – Intro to Big Data and Analytics

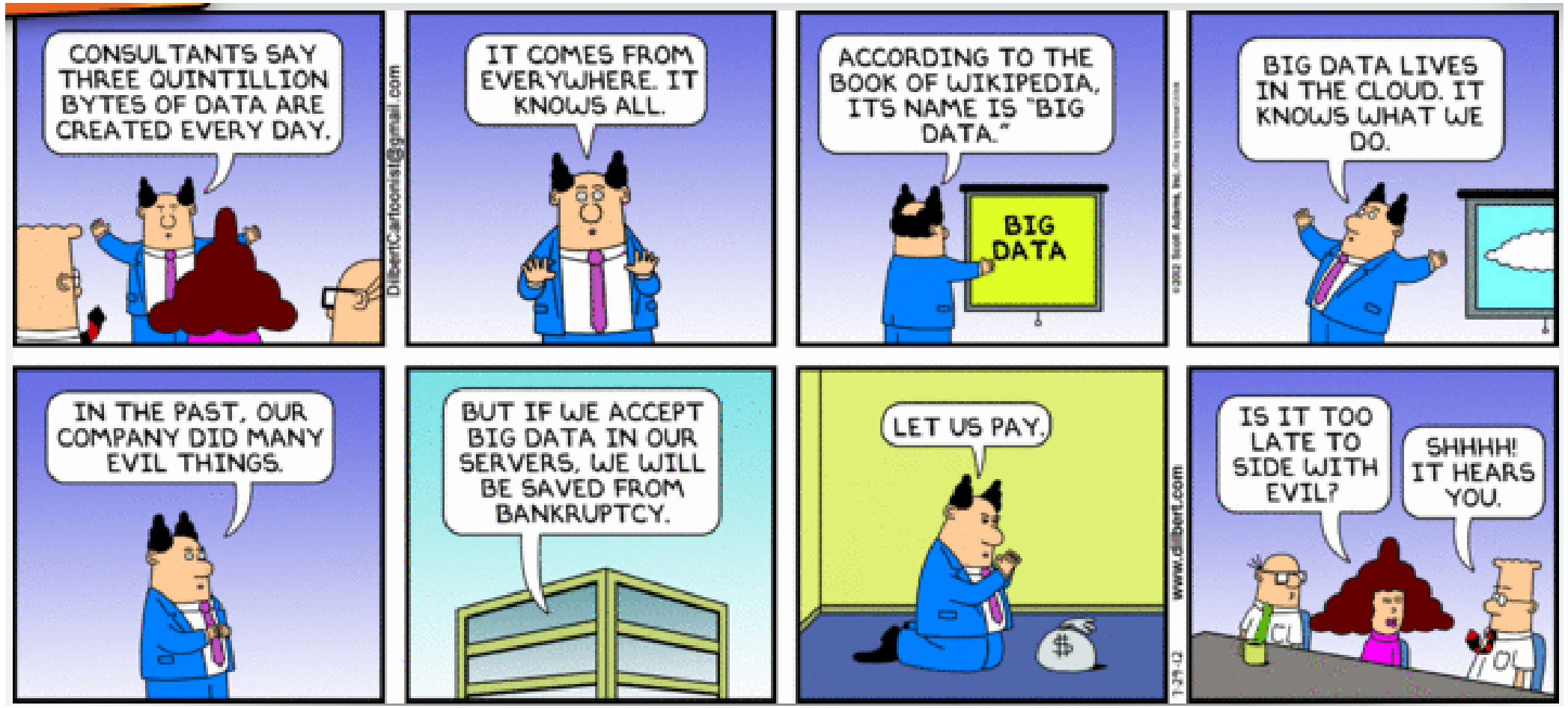
Dr. D. Blaine Nashold, Jr.

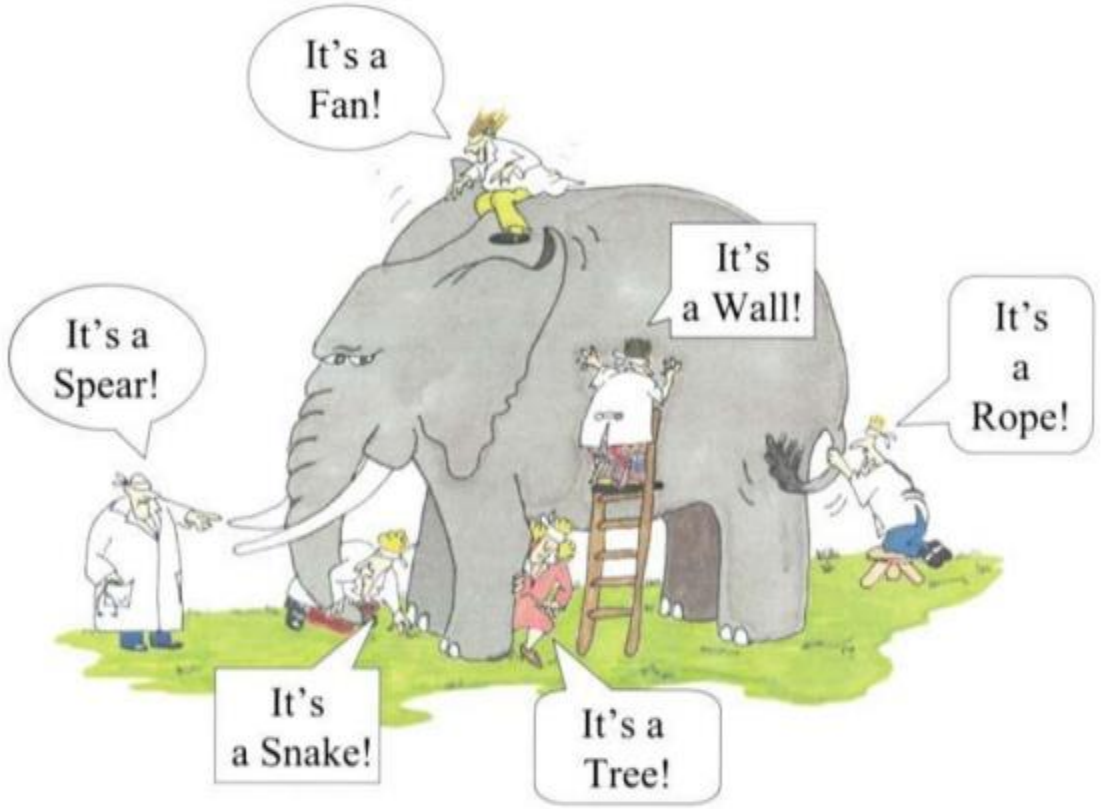
UNIVERSITY OF NORTH CAROLINA CHARLOTTE

Discussion Outline

- What is “Big Data”? What drives “Big Data”?
- Data structures and repositories in the age of “Big Data”
- “Big Data” analytics
- Problems addressed with “Big Data” and analytics
- “Big Data” ecosystem and key roles in this ecosystem

Big Data?


CONSULTANTS SAY THREE QUINTILLION BYTES OF DATA ARE CREATED EVERY DAY.
IT COMES FROM EVERYWHERE. IT KNOWS ALL.
ACCORDING TO THE BOOK OF WIKIPEDIA, ITS NAME IS "BIG DATA."
BIG DATA
BIG DATA LIVES IN THE CLOUD. IT KNOWS WHAT WE DO.
IN THE PAST, OUR COMPANY DID MANY EVIL THINGS.
BUT IF WE ACCEPT BIG DATA IN OUR SERVERS, WE WILL BE SAVED FROM BANKRUPTCY.
LET US PAY.
IS IT TOO LATE TO SIDE WITH EVIL?
SHHHH! IT HEARS YOU.
DilbertCartoonist@gmail.com
www.dilbert.com



BIG DATA

Arnon Rotem-Gal-Oz

Director of Technology Research, Amdocs

The blind men and the elephant. Poem by John Godfrey Saxe (Cartoon originally copyrighted by the authors; G. Renee Guzlas, artists http://www.nature.com/ki/journal/v62/n5/fig_tab/4493262f1.html

So seriously, what is “Big Data”?

Big Data

Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.[2] "There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem."[3]

There is no agreed upon definition for "big data." The tools of data science are as appropriate for gigabyte as they are for petabyte scale datasets. "Big data" typically refers to data on the scale of terabytes (10 to the 12th power) and petabytes (10 to the 15th power). A petabyte is a million gigabytes.

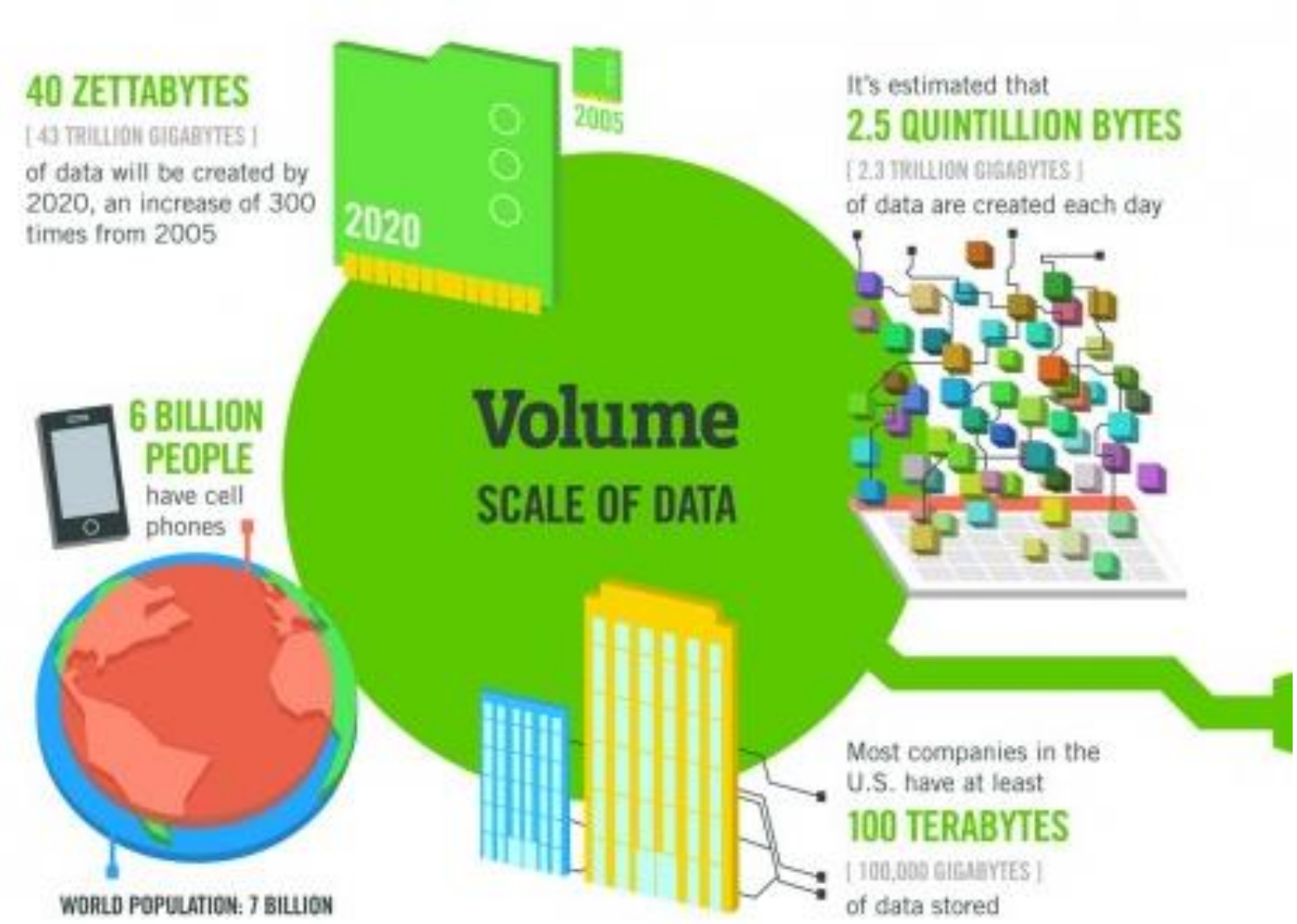
http://datascience.berkeley.edu/about/what-is-data-science/

Big Data is data whose scale, distribution, diversity, and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of business value.

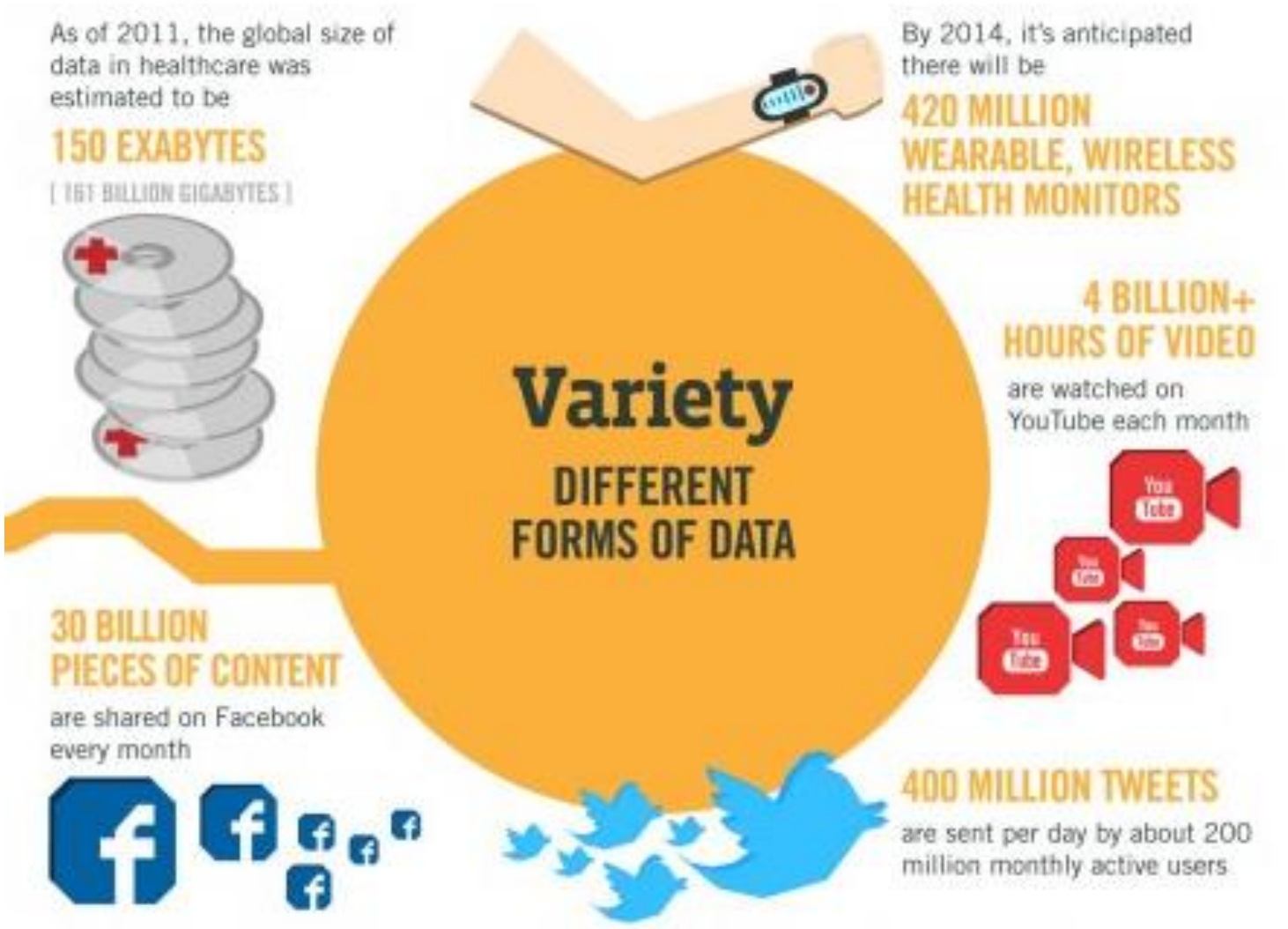
McKinsey & Co.; Big Data: The Next Frontier for Innovation, Competition, and Productivity

What is Big Data? The 5 Vs…

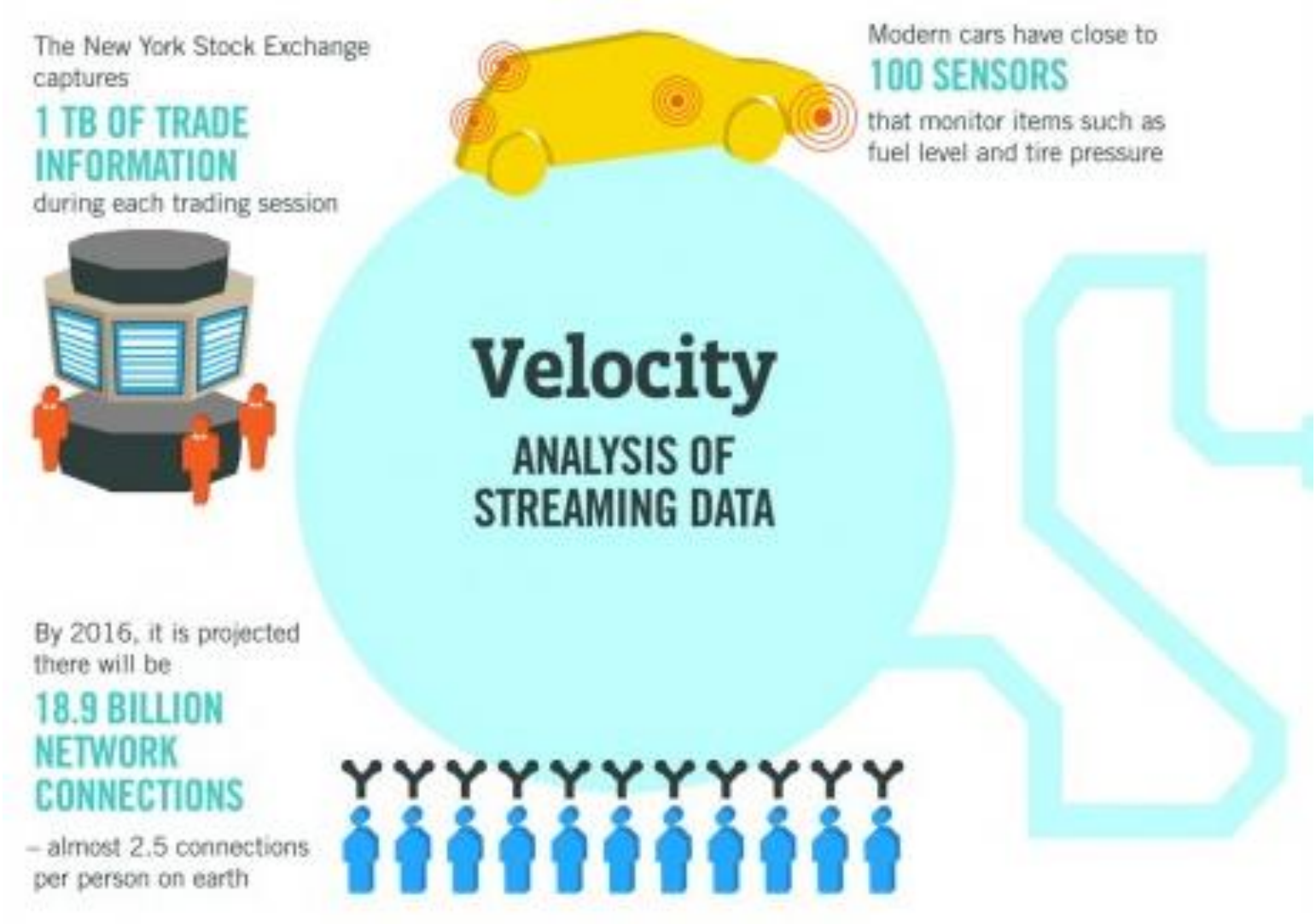
- **Volume:** Organizations have access to large amount of data. Every day, 15 petabytes of new information are being generated. As of 2010, the codified information base of the world doubles every 11 hours.
- **Variety:** Data comes in all formats; 80% of new data growth is unstructured content: generated largely by email, documents, images, video and audio, and geospatial data
- **Velocity:** Data is generated at a much faster pace than in the past. Need to switch from batch to streaming data processing, with decisions to be made in fractions of a second
- **Variability:** Data flows with highly inconsistent peaks (e.g., trending topics in social media)
- **Veracity:** Due to the volume and variety of the data, uncertainty about the truth in the data, and how one can verify the truth economically



Source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

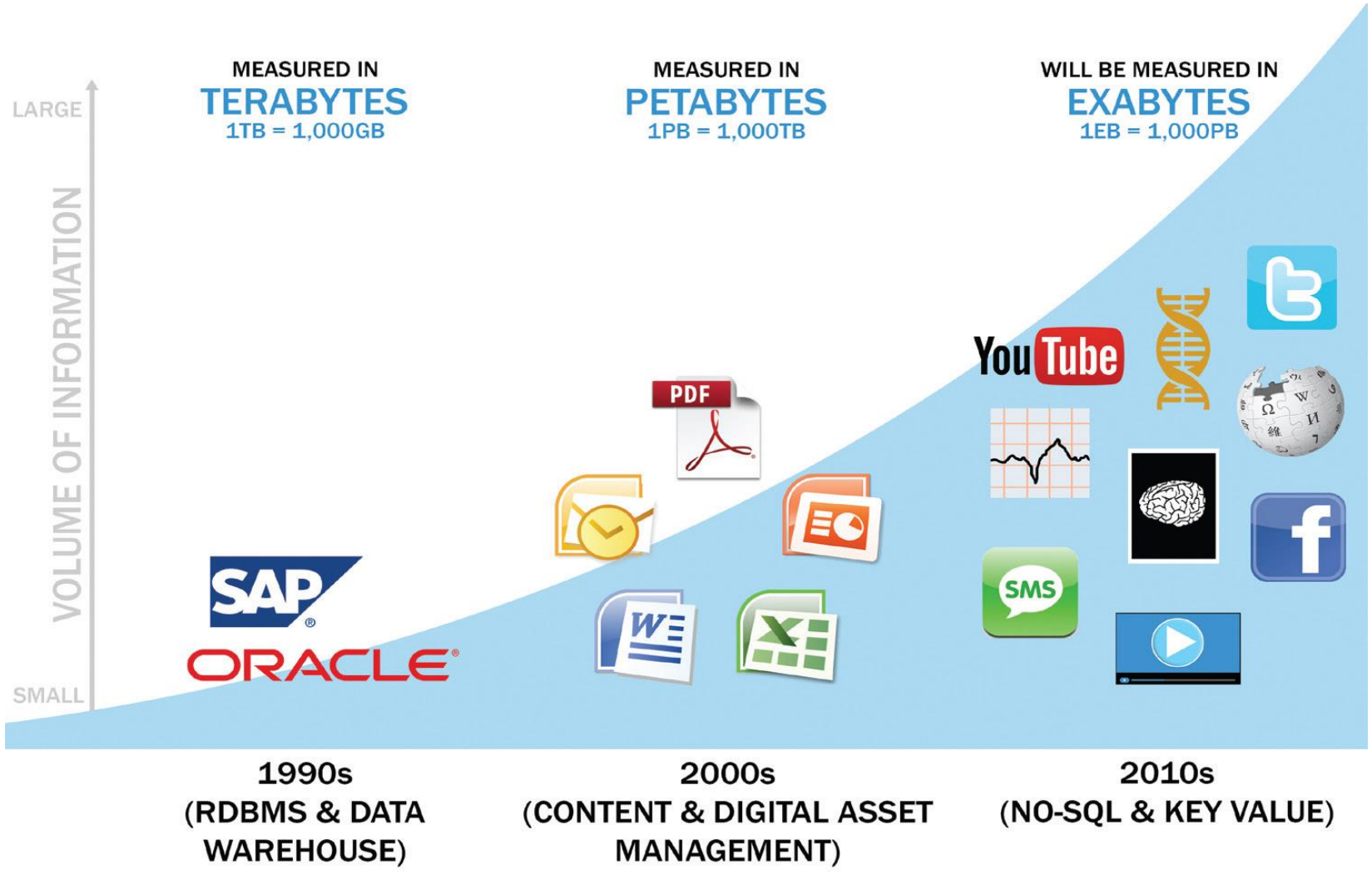


Source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data



Source: http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Evolution of Data & Sources



Data Structures in Big Data

- **Structured Data**
  - Defined data type, format, and structure
  - Transaction data, OLAP data cubes, RDBMS, CSV and Excel data
- **Semi-structured Data**
  - Text data with some discernible patterns that enables parsing fairly easily
  - XML, system logs, web logs
  - Web click-stream
  - Within semi-structured data, there is a range depending on the efforts required to understand the structure and content
- **Unstructured Data**
  - Data with no inherent structure

Spreadsheets to Analytic Sandbox

Data Repository	Characteristics
Spreadsheets and data marts ("spreadmarts")	Spreadsheets and low-volume databases for recordkeeping Analyst depends on data extracts.
Data Warehouses	Centralized data containers in a purpose-built space Supports BI and reporting, but restricts robust analyses Analyst dependent on IT and DBAs for data access and schema changes Analysts must spend significant time to get aggregated and disaggregated data extracts from multiple sources.
Analytic Sandbox (workspaces)	Data assets gathered from multiple sources and technologies for analysis Enables flexible, high-performance analysis in a nonproduction environment; can leverage in-database processing Reduces costs and risks associated with data replication into "shadow" file systems "Analyst owned" rather than "DBA owned"

Case Discussion 1 (Diagnosis)

Discuss a big data use case (from your reading or experience):

1. What is the most significant problem/challenge faced by the [firm/protagonist]?
2. Who or what is [responsible/to blame] for the crisis faced by the [firm/protagonist]?
3. Why has the [firm/protagonist] performed so well/poorly?
4. As [the case protagonist], what keeps you up at night? What are you most worried about?

What is Analytics?

- Analytics is a process that involves statistics, computer programming, and operations research to explore, visualize, discover and communicate patterns or trends in big data

- Data analytics refers to the use of analytics on any type of data or data-driven problems
  - Genomics data analysis

- Business analytics focuses more on business-data and in driving business decisions and actions
  - Retail data analysis

Types of Analytics

- **Descriptive analytics**
  - Describing what is contained in a dataset or database
  - Knowing what is happening in the organization and some underlying trends
  - Visualization is a key part of descriptive analytics
  - Example: Bar chart of customers by age
- **Predictive analytics**
  - Aims to determine what is likely to happen in the future
  - Based on statistical and machine learning techniques
  - Helps identify trends and relationships not readily observed in descriptive analytics
  - Example: Predicting if a patient will come in for readmission

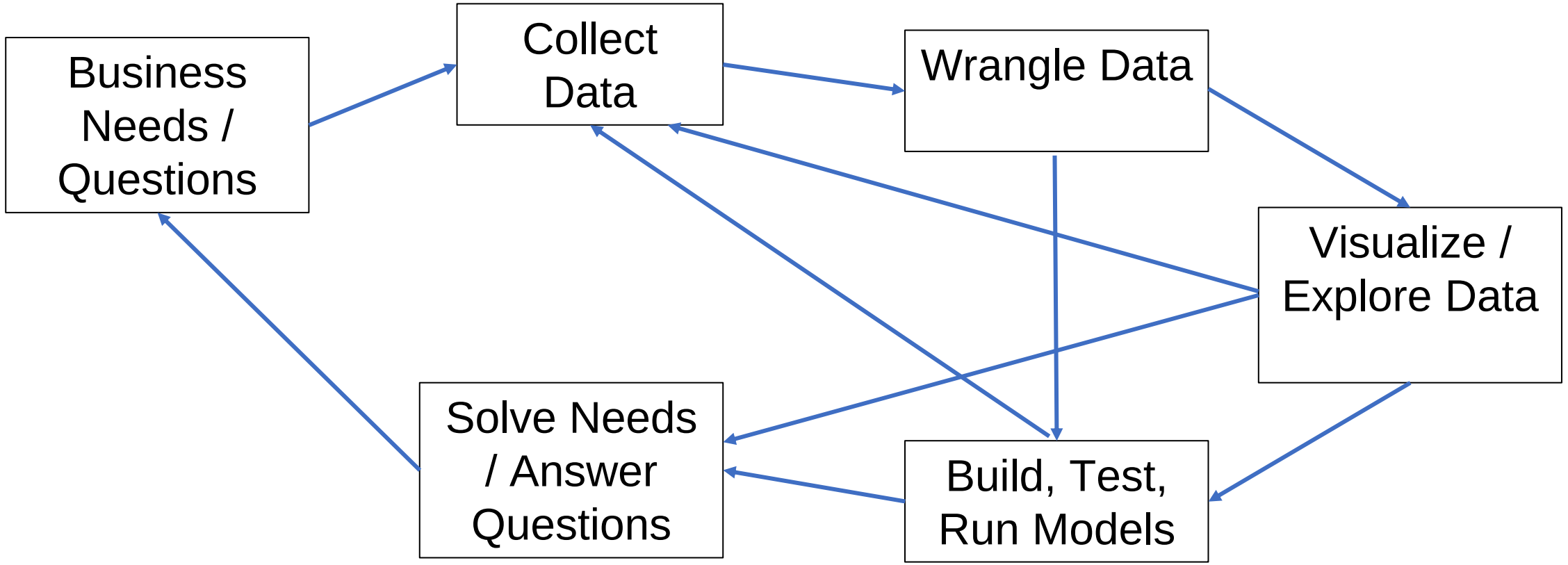
Types of Analytics

- **Prescriptive analytics**
  - Application of operations research and management science to provide a decision or recommendation for a specific action
  - Also called decision or normative analytics
  - Example: Airfare to charge

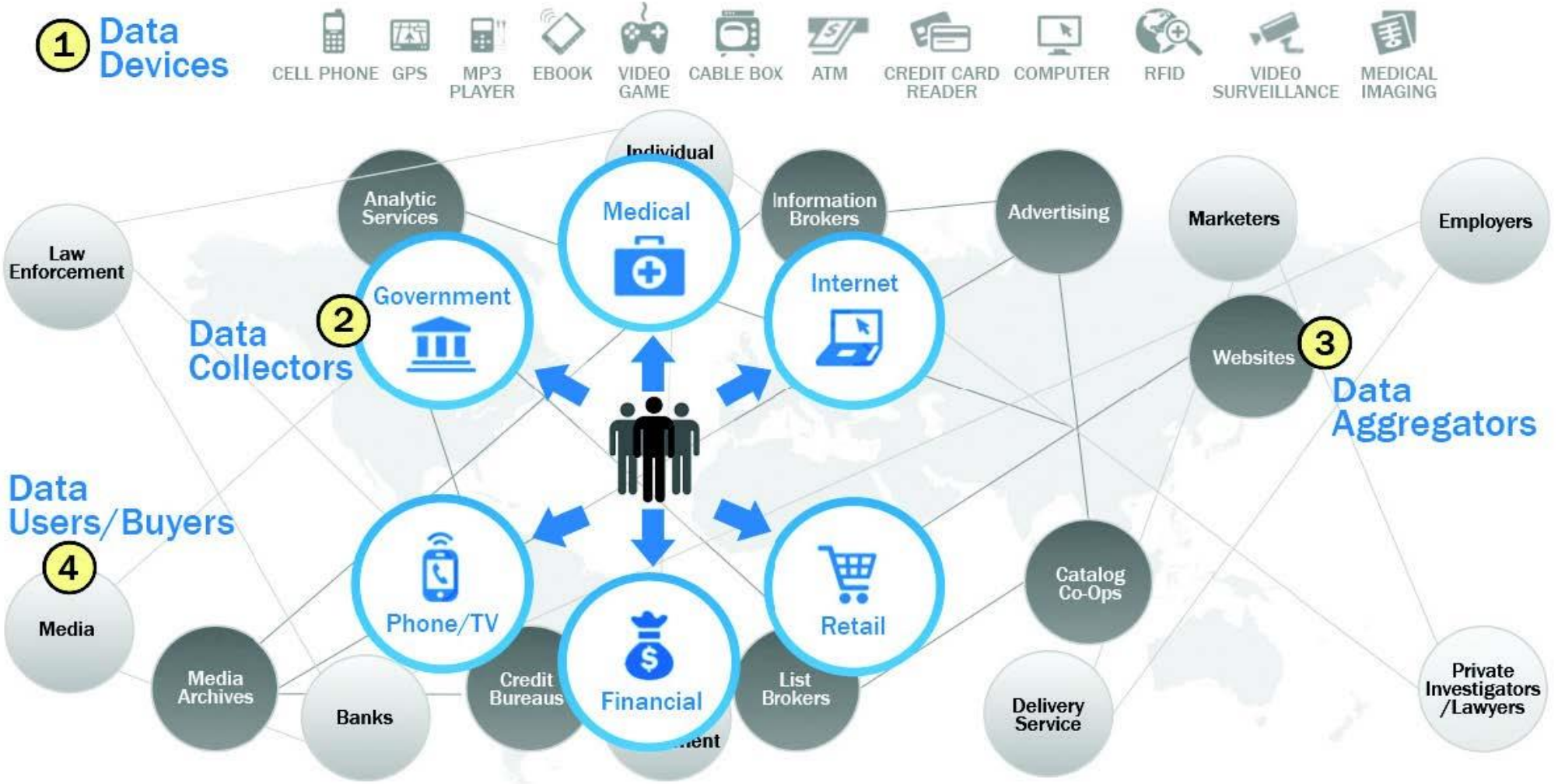
Business Drivers & Analytics Applications

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Basel II-III, Sarbanes-Oxley (SOX)

Big Data Analytics as a Process



Big Data Ecosystem



Big Data Ecosystem – Key Roles

- **Data Scientist**
  - Deep analytical talent
  - Skills to handle raw, unstructured data and apply complex analytical techniques
  - Advanced training in math, stats, and machine learning
- **Data Professional**
  - Define key questions to be answered with analytics
  - Basic understanding of the data and data scientists
  - Bring deep domain expertise and business understanding
- **Technology Enablers**
  - Support technology expertise to support analytical projects and infrastructure

Big Data Challenges

- Identifying the right data and knowing how to use the data
- Finding the right talent capable of both working with new technologies and of interpreting the data to find meaningful business insights
- Overcoming obstacles with data access and connectivity
- Responding to the rapidly changing technology landscape
- Working across functions, such as IT, engineering, finance and procurement, where ownership of data is fragmented across the organization
- Addressing security and privacy concerns of customers/citizens

Eric Spiegel, President and CEO of Siemens U.S.A. (Wall Street Journal blog)

Sources

- Chapter 1 (Introduction to Big Data Analytics) from the book "Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing, and Presenting Data" by EMC Education Services, John Wiley & Sons, 2015.
- “Big Data in Big Companies” – Thomas H. Davenport and Jill Dyche
- “A Very Short History of Data Science” – Gil Press