CSP 554
BIG DATA TECHNOLOGIES

Module 1a

Course Information

Big Data Concepts

If we have data, let's look at data.
If all we have are opinions, let's go with mine.

Jim Barksdale

CEO of Netscape
COO of FedEx
Board Member of Time Warner
Net worth: US$ 700 Million

Data Collection

In the process of data democratization… the world's data have never been more open that today

The world's data sources (e.g., social media, news outlets) often permit –restricted–access to their data

- Web Scraping: methodically scrape website content
- Application Programmable Interfaces (APIs)
  - ASK for permission and GET access to resource(s)
  - So… turn the "tap" of a data source (coding task) and store the data somewhere for analysis

Insight Creation

Scanning through a few dozen data records manually is easy.

But… what about thousands or millions of data records form multiple sources?

Humans are slow, Computers are fast!

Get the data, store it and then let the computer generate insights!

Knowledge Discovery

Automating Insight Creation

The automated nontrivial extraction of implicit, previously unknown, and potentially useful information from data

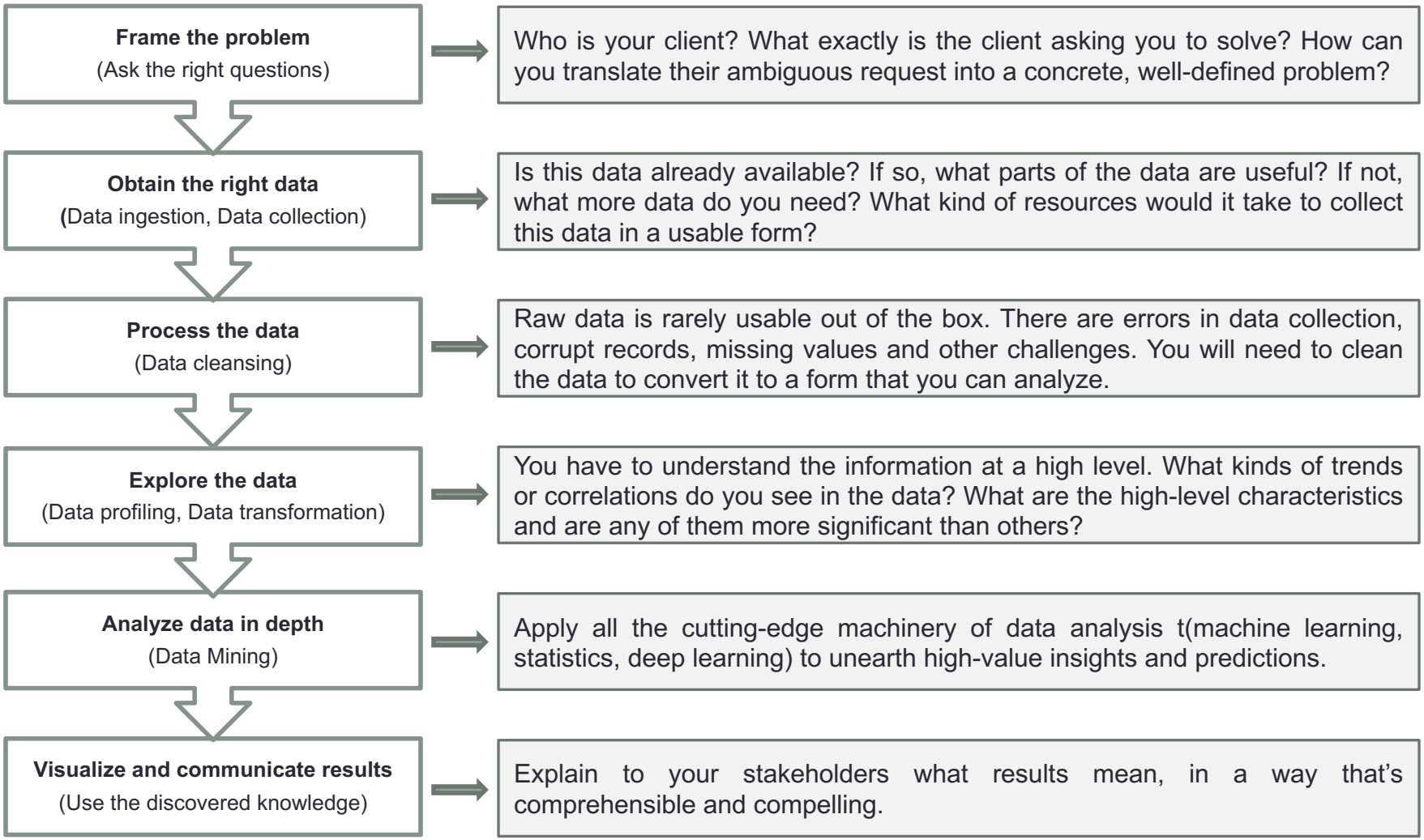
Data, in its raw form, is simply a collection of elements, from which little knowledge can be gleaned

With the development of knowledge discovery techniques the value of the data is significantly improved

The goal is to distinguish from source data, something that may not be obvious but is valuable or enlightening in its discovery

Extraction of knowledge from source data is accomplished by applying data mining methods

Knowledge Discovery Process



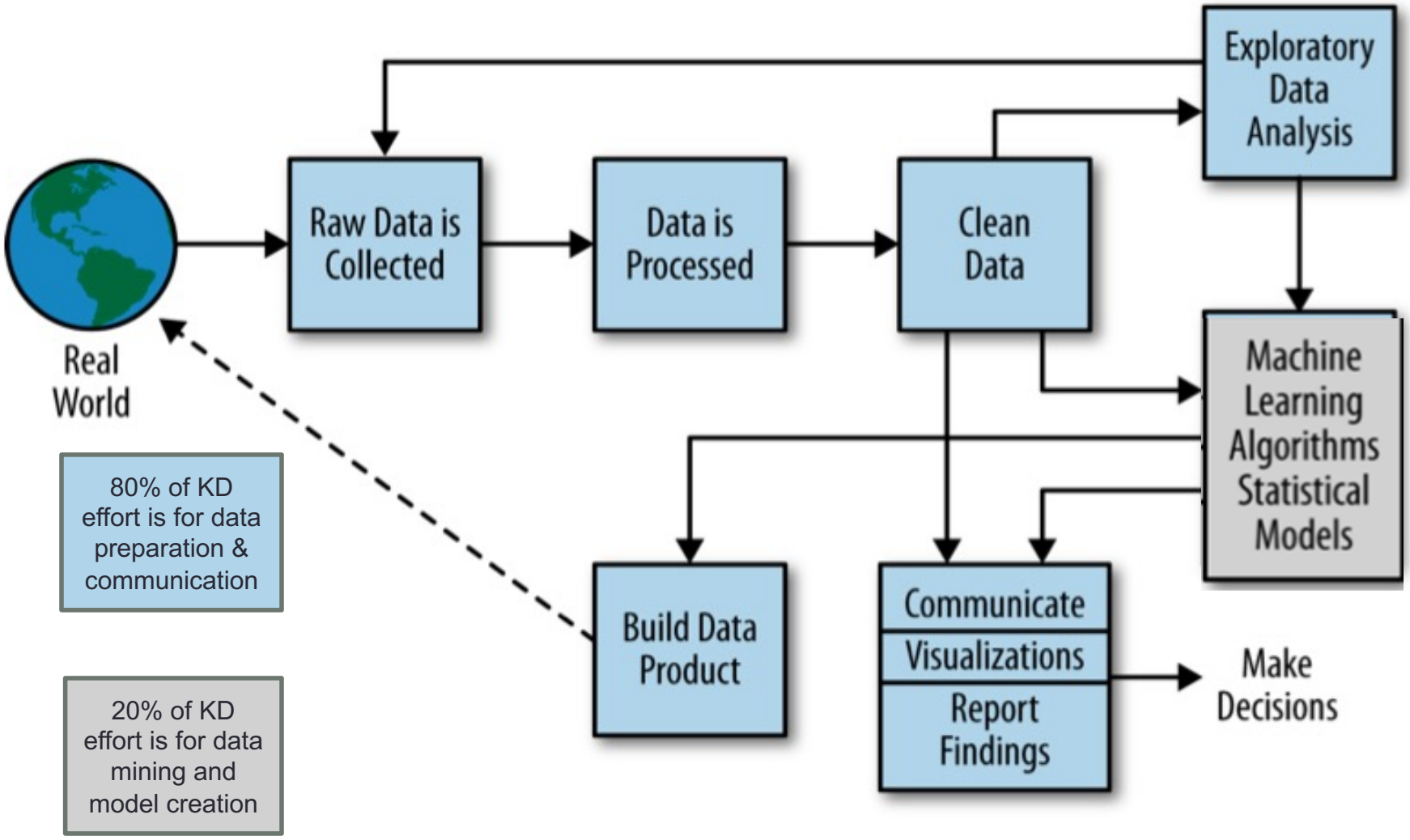
Data Mining
Analyzing Data in Depth

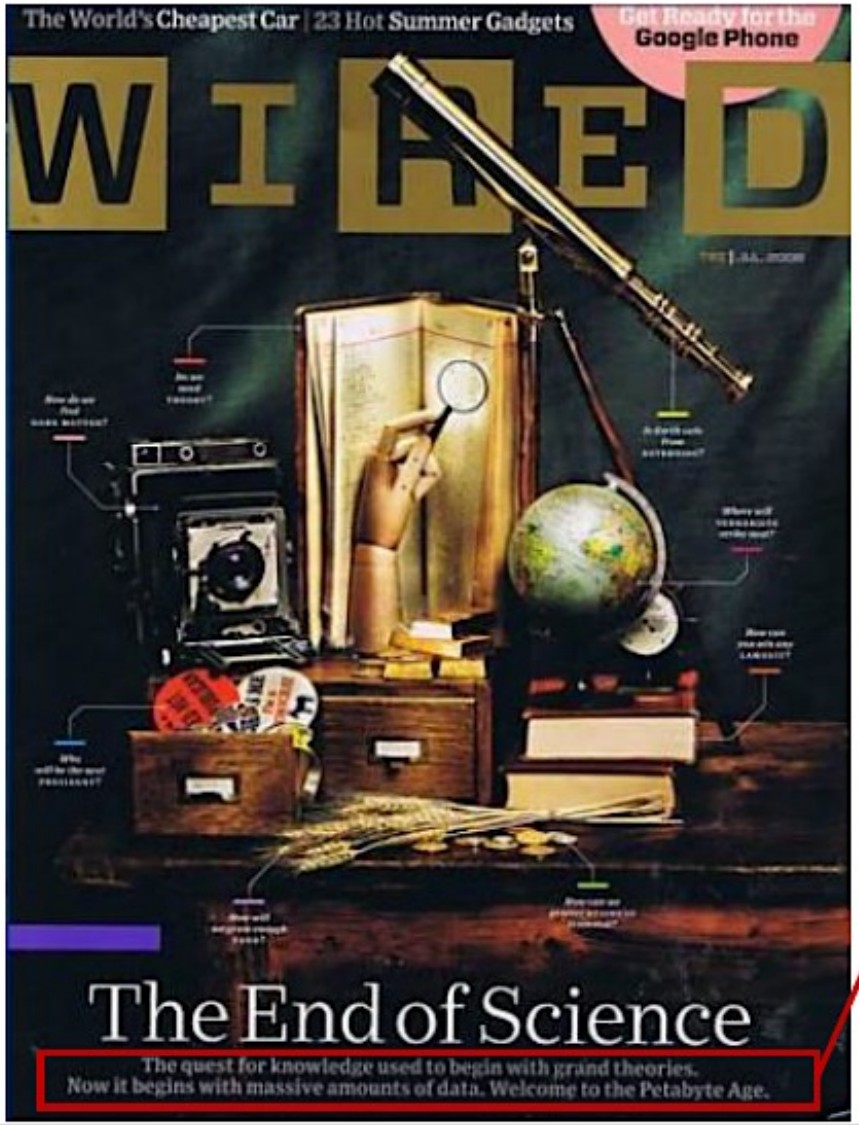
Data mining is a multidisciplinary field
drawing work from areas including:

database technology
artificial intelligence
machine learning / deep learning / neural networks
statistics
high performance computing / parallel processing
data visualization

Knowledge Discovery Process

Distribution of Data Science Effort





The quest for knowledge used to begin with grand theories
(a hypothesis).

Now it begins with massive amounts of data.

Welcome to the Petabyte Age.

Wired, Jun 2008

Big Data

2.5 QUINTILLION

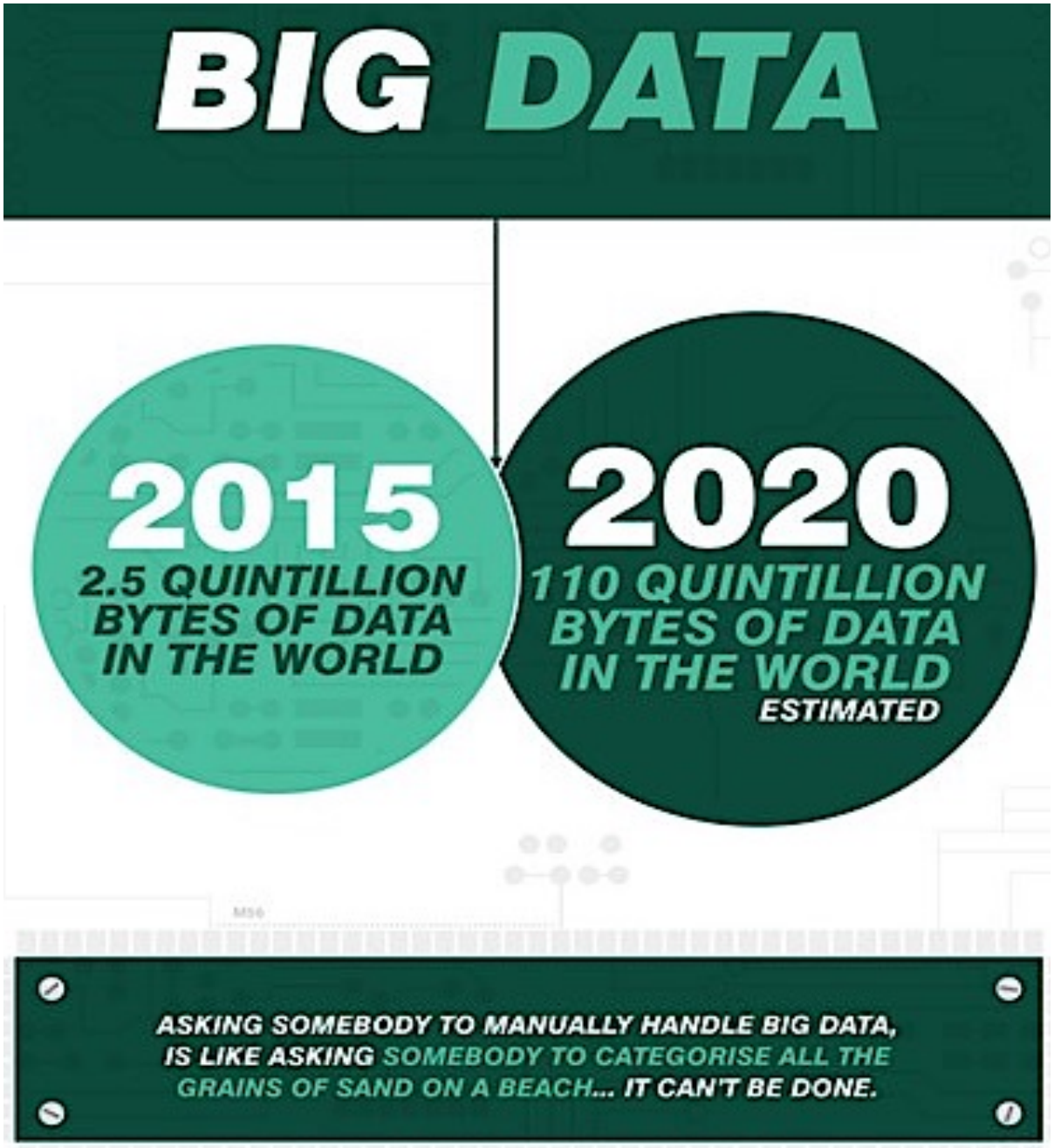
bytes of data are created every day
(that's 2,500,000,000,000,000,000 bytes.)

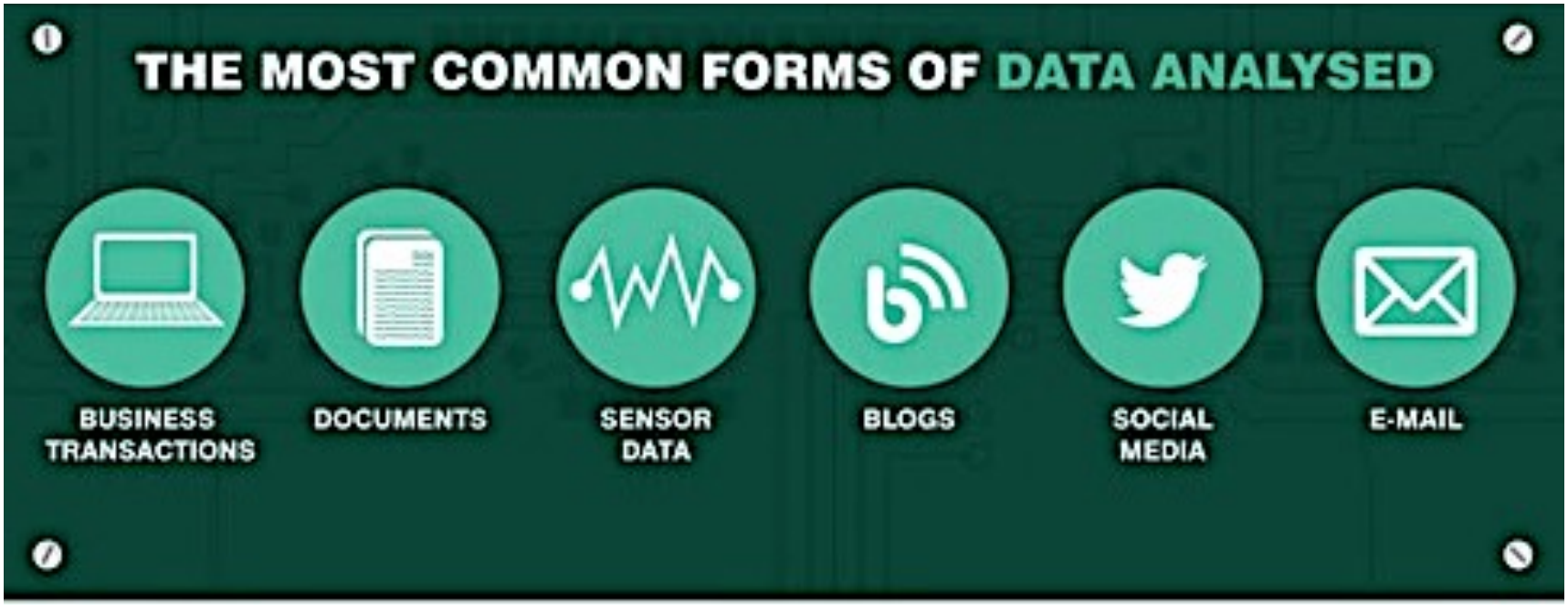
The volume of business data doubles every

this year **1.2 YEARS** + 1.2 years

BY 2020, THE DIGITAL UNIVERSE WILL EQUAL

That's 5,247GB of machine-generated data for every person on the planet


BIG DATA
2015
2.5 QUINTILLION BYTES OF DATA IN THE WORLD
2020
110 QUINTILLION BYTES OF DATA IN THE WORLD
ESTIMATED
ASKING SOMEBODY TO MANUALLY HANDLE BIG DATA, IS LIKE ASKING SOMEBODY TO CATEGORISE ALL THE GRAINS OF SAND ON A BEACH... IT CAN'T BE DONE.


THE MOST COMMON FORMS OF DATA ANALYSED
BUSINESS TRANSACTIONS
DOCUMENTS
SENSOR DATA
BLOGS
SOCIAL MEDIA
E-MAIL


30 BILLION PIECES OF CONTENT SHARED ON FACEBOOK EVERY MONTH.


175 MILLION TWEETS ARE POSTED FROM 65 MILLION ACCOUNTS EVERY DAY.

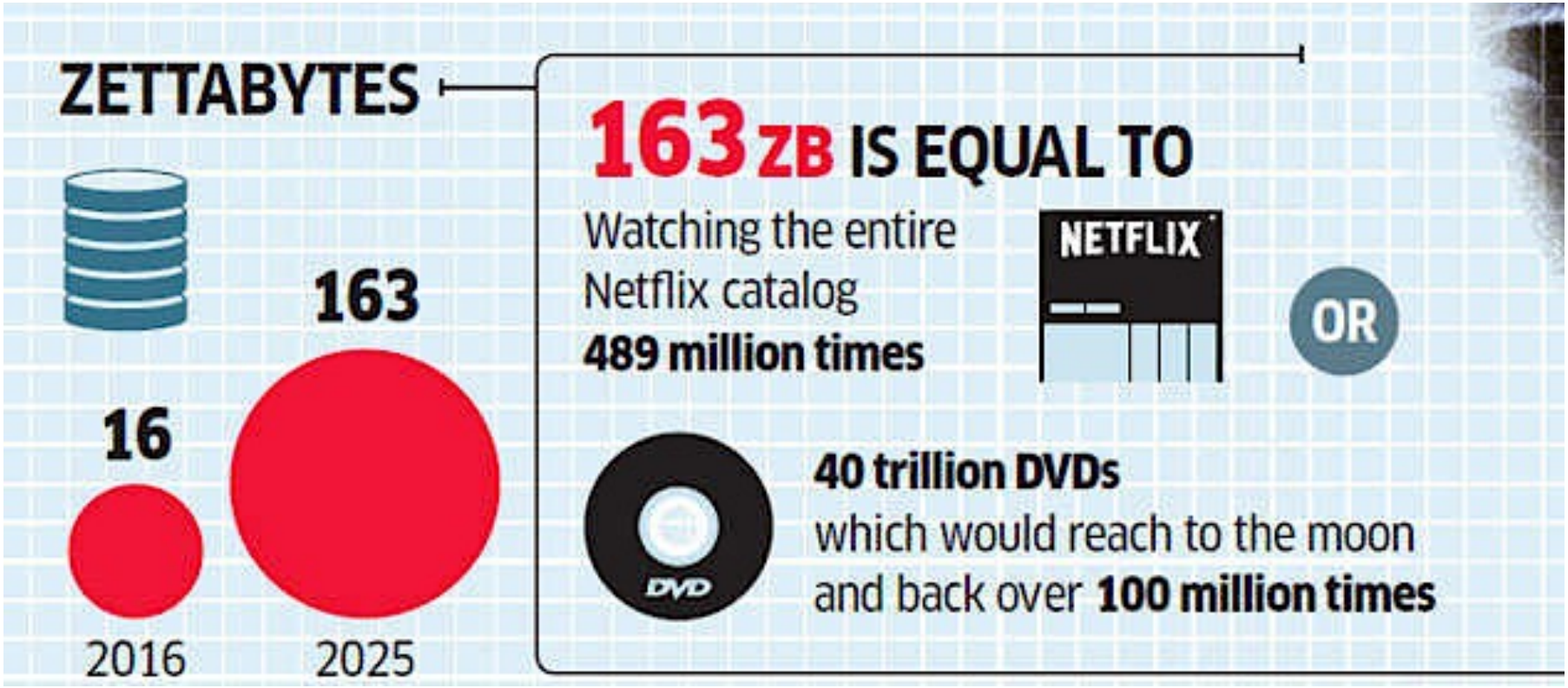

PROCESSED 20,000 TERABYTES OF DATA A DAY IN 2008.

1 ZETTABYTE =

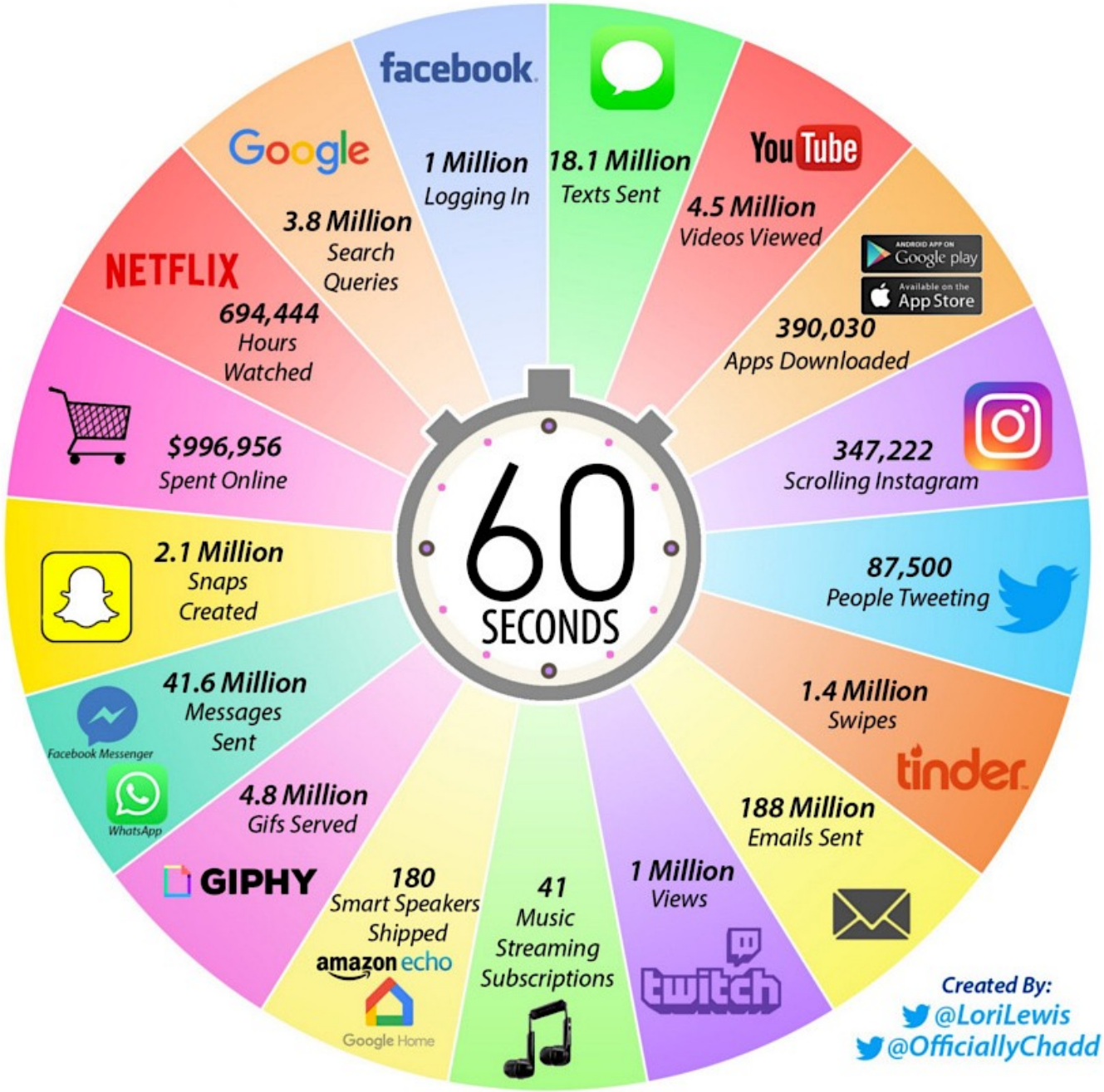
1,000,000,000,000,000,000,000 BYTES =

10^{21} BYTES

Growth of Data Volume Across Time



Growth of Data Streams Over Time



Big Data refers to datasets that are too large or complex for traditional data-processing application software to adequately deal with

The Challenge Posed by Big Data

Why Do We Need Specialized Big Data Technology?

To remain competitive, firms absolutely must uncover insights from vast and diverse data sets.

But this data is of such size and complexity that many common software packages and systems that are used to collect, process and analyze lesser amounts or simpler format data cannot be used to generate results in a reasonable time

Big data technologies are exactly those software packages, tools and systems which are architected and implemented specifically to uncover business value from large volumes of data

Big Data is…a **Volume** Problem

Sensory Data

Boeing 787 generates

40TB of data per hour

in flight.

Google's self-driving car generates **1GB of data per minute**.

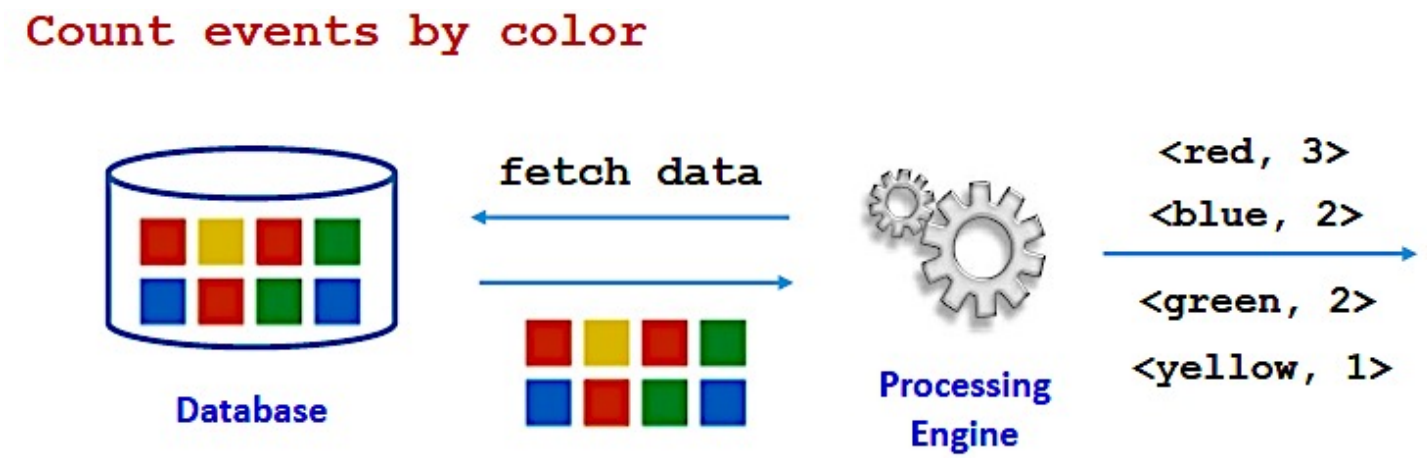
The Internet of Things

21 Billion devices by 2020 accounting for 12% of the digital universe.

Big Data is…a **Velocity** Problem

Batch Data

Assumes that the data is available when andif we want it(e.g., reading and parsing data from a file or database)

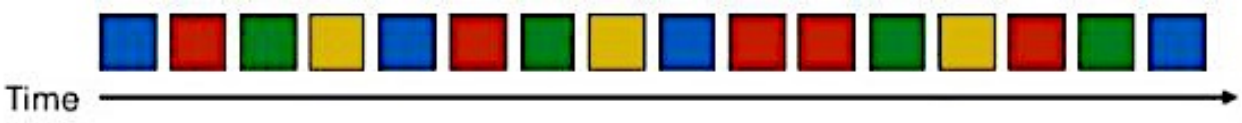


The processing engine knows the dataset in advance and controls the input rate of the data

Streaming (Realtime) Data

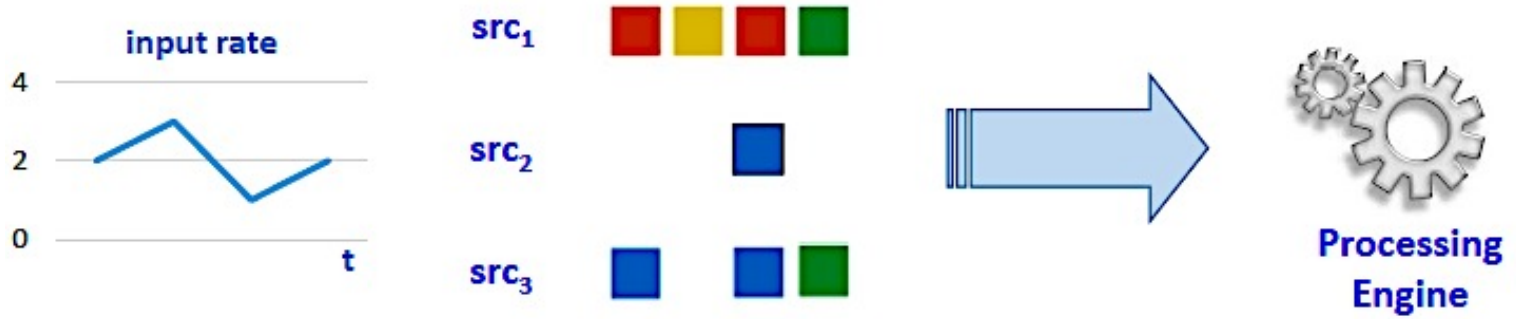
Unbounded Data -> the volume of the data is overwhelming

- Conceptually infinite sequence of data items



Push Model -> data arrives at high velocity and different rates

- Potentially multiple sources pushing data to the processing engine at different rates (data distribution changes over time)

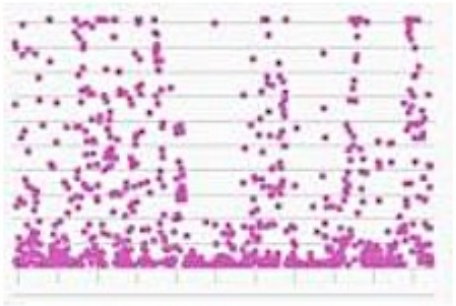


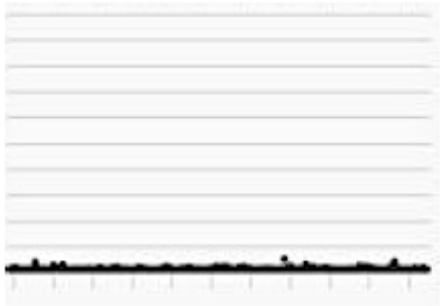
US Presidential Elections 2016

200K tweets/min

Per minute Emotions During First Debate

Clinton

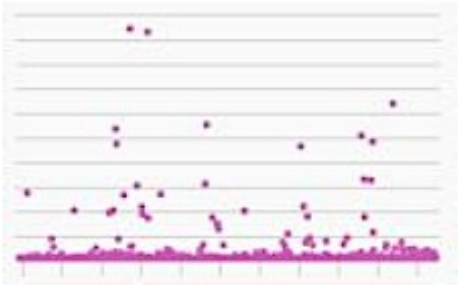


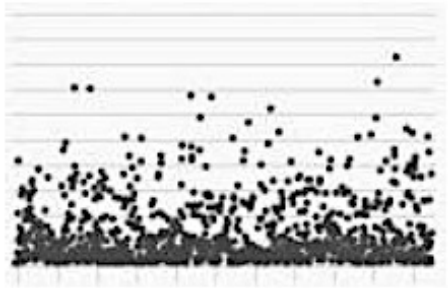


Happiness

Anger

Trump



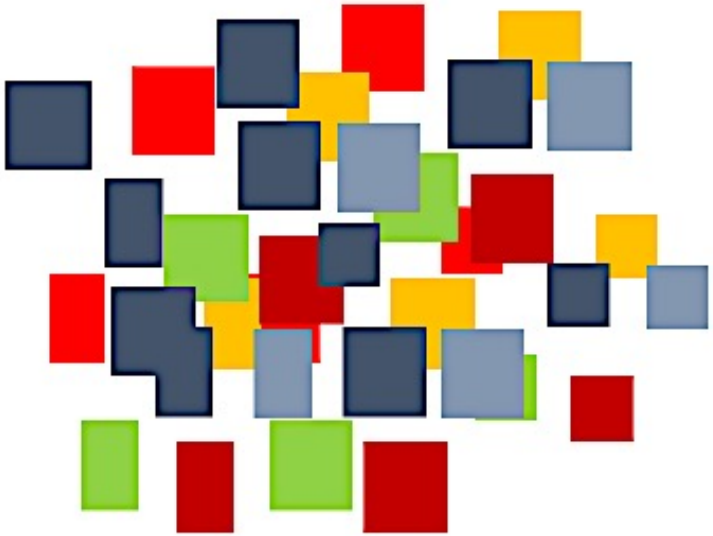


https://qz.com/810092

Big Data is…a **Variety** Problem

Unstructured Data

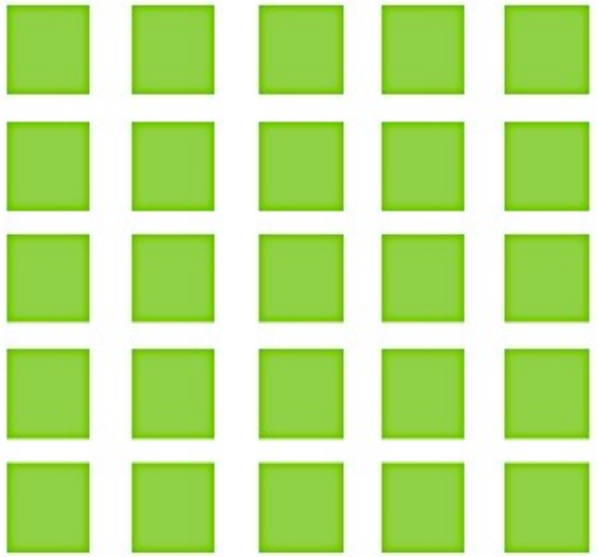
Text
Images
Sound
Video
...



keith
@KeetPotato
Follow

[ordering cake over phone]
"and what would you like the cake to say?"
[covers phone to ask wife]
"do we want a talking cake?"

Structured Data



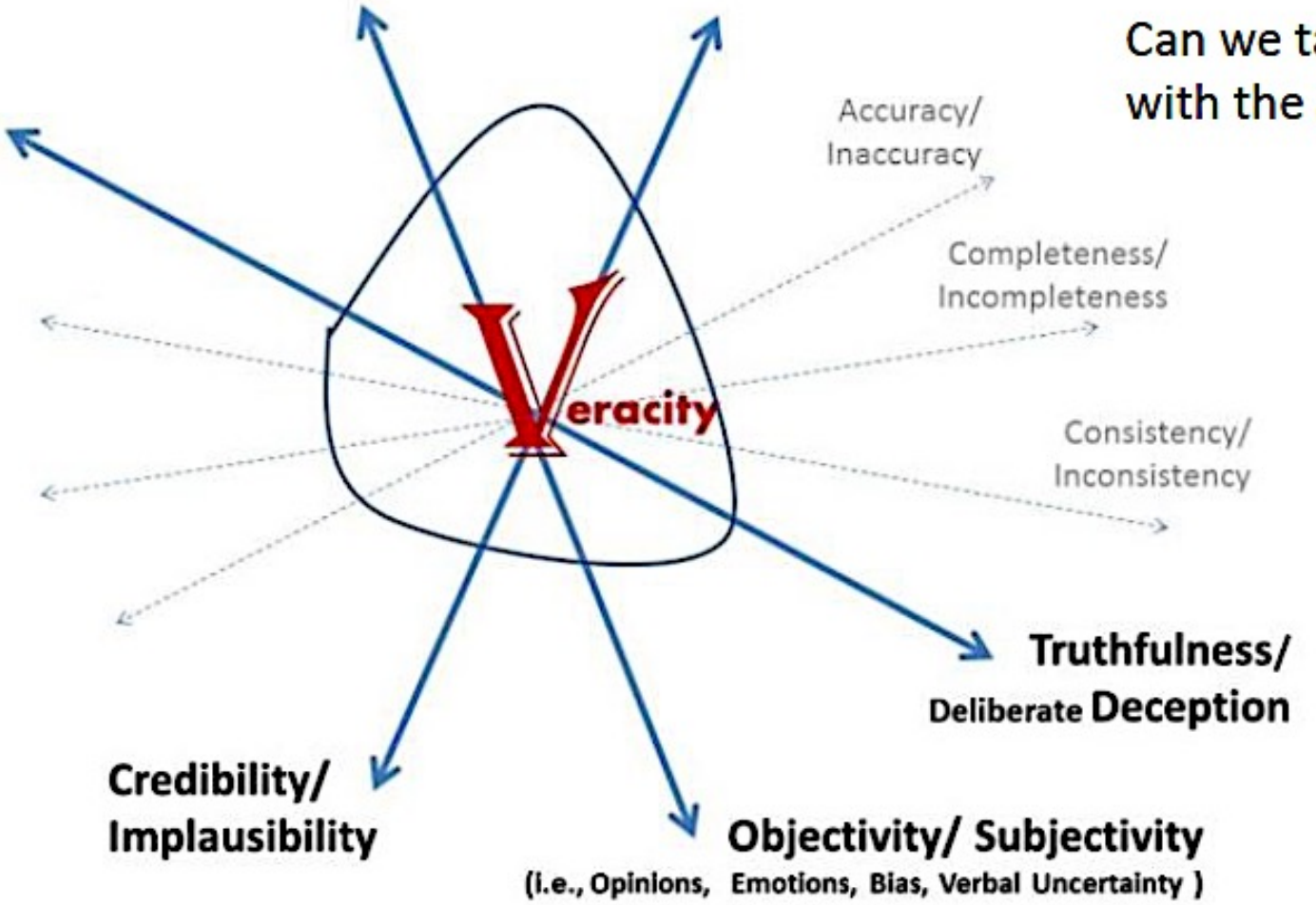
.xsl
.csv
.json
.xml
...

Almost 80% of today's business data is unstructured (text) data

Big Data is…a **Veracity** Problem

Definition of veracity:
1: conformity with truth or fact : ACCURACY
2: devotion to the truth : TRUTHFULNESS
3: power of conveying or perceiving truth

Data Quality

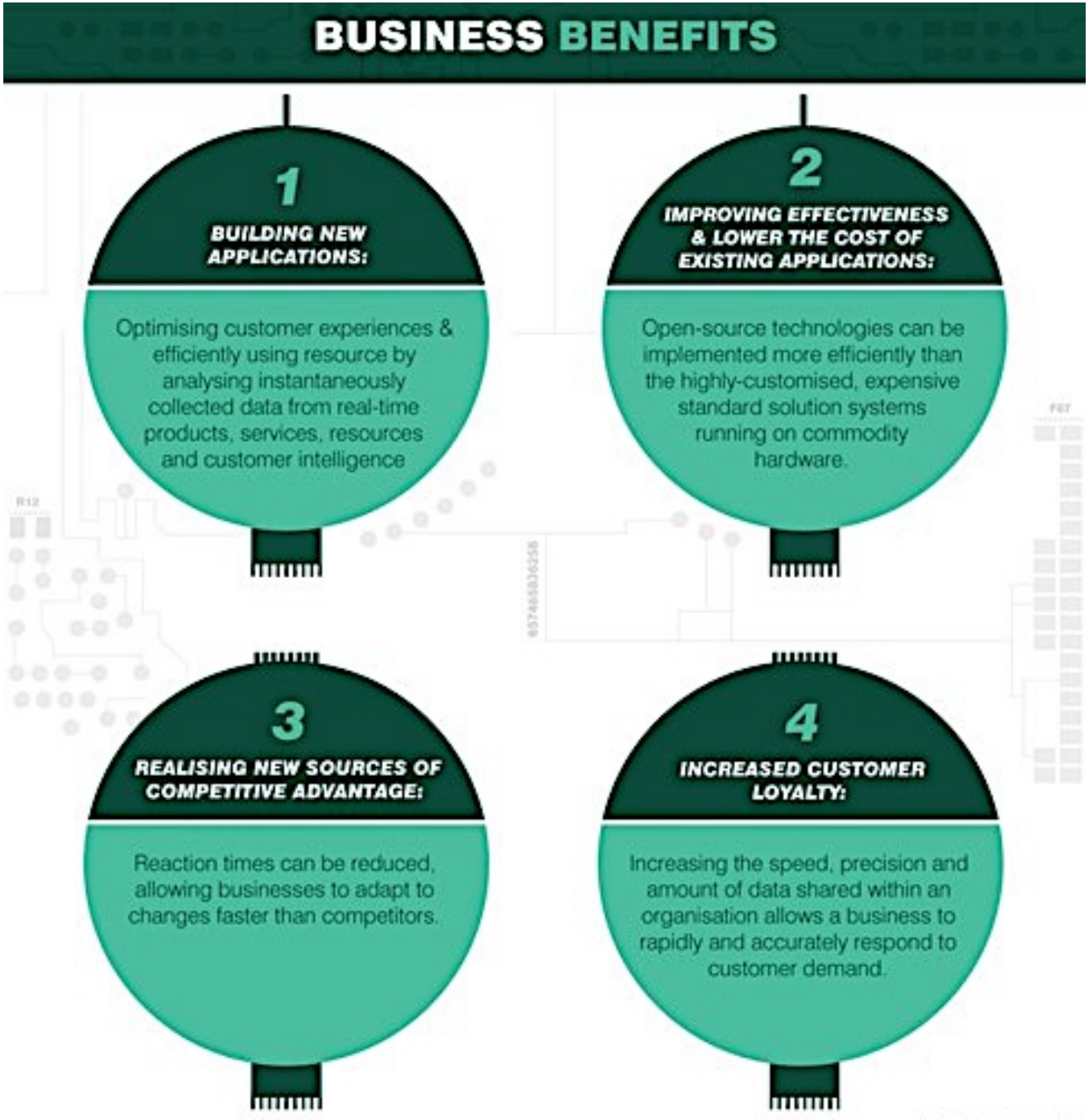


Can we take **good decisions** with the **available** data?

Can we **trust** the data regardless of the **source**?

Big Data Generates Value in Several Ways

- Creating transparency: making data accessible timely manner.
- Enabling experimentation: collecting more accurate and detailed performance data, setting up controlling experiments.
- Segmenting populations to customize actions, target promotions and advertisement.
- Replacing/supporting human decisions making with automated algorithms.
- Innovating new business models, products and services:
  - UBER, Spotify, LinkedIn, Twitter, Netflix are well-known examples of this. Big Data is affecting healthcare too


BUSINESS BENEFITS
1
BUILDING NEW APPLICATIONS:
Optimising customer experiences & efficiently using resource by analysing instantaneously collected data from real-time products, services, resources and customer intelligence
2
IMPROVING EFFECTIVENESS & LOWER THE COST OF EXISTING APPLICATIONS:
Open-source technologies can be implemented more efficiently than the highly-customised, expensive standard solution systems running on commodity hardware.
3
REALISING NEW SOURCES OF COMPETITIVE ADVANTAGE:
Reaction times can be reduced, allowing businesses to adapt to changes faster than competitors.
4
INCREASED CUSTOMER LOYALTY:
Increasing the speed, precision and amount of data shared within an organisation allows a business to rapidly and accurately respond to customer demand.

COURSE INFORMATION

Contact Information

Joseph Rosen

Technical Fellow

Enterprise Architect

Blue Cross and Blue Shield Association

jrosen@iit.edu (Emails responded to in a day or less)

C: 312-860-0860

Office Hours:

- Virtual office hours: Thursday, 5:15pm-6:15pm Central and by appointment, all via cell

Course Summary

- This course provides a rapid immersion into the area of big data and the technologies which have recently emerged to manage it.
- We start with an introduction to the characteristics of big data and an overview of the associated technology landscape
- And continue with an in depth exploration of Hadoop, the leading open source framework for big data processing.
  - We focus is on the most important Hadoop components such as Hive, Pig, stream processing and Spark as well as architectural patterns for applying these components.
- We go on to an exploration of the range of specialized (NoSQL) database systems architected to address the challenges of managing large volumes of data
- Overall the objective is to develop a sense of how to make sound decisions in the adoption and use of these technologies

Required Texts

- Tom White. 2015. *Hadoop: The Definitive Guide* (4th ed.). O'Reilly Media, Inc. (TW)
- Pramod J. Sadalage and Martin Fowler. 2012. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley.(PS)

Free Online Texts (also Blackboard)

- Jimmy Lin and Chris Dyer. 2010. *Data-Intensive Text Processing with Mapreduce*. Morgan and Claypool Publishers.
  - https://vgc.poly.edu/~juliana/courses/BigData2014/Textbooks/MapReduce-algorithms-Jan2013-draft.pdf
- Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge University Press, New York, NY, USA.
  - http://infolab.stanford.edu/~ullman/mmds/bookL.pdf

Grading Policies

Short Quizzes	15%
Assignments	15%
Project / Paper Proposal	5%
Project / Paper Draft	5%
Project / Paper	25%
Half Term Quiz	35%

Grade Distribution

A = 100 – 90%, B = <90% – 75%, C = <75% – 60% (Undergraduate only)

Quiz and Exam Policy

- Short Quizzes (every week or two, via Blackboard)
  - Each short quiz will be open notes and books and consist of multiple choice and short answer questions and should take no more than 15-20 minutes (timed) to complete.
  - Covers lectures, readings, assignments
  - The lowest short quiz score will be dropped when calculating the average short quiz score.
  - Since you can miss a short quiz with no penalty there are no make ups or exceptions (except as explicitly provided for by IIT policy).
- Half Term Exam (via Blackboard)
  - The half term exam will be open notes and books and consist of multiple choice and short answer questions, some longer essay questions and take 3 hours (timed).

Assignments, Projects/Paper Policies

- Assignments, project/paper proposals and project/paper drafts must be submitted by their announced due dates and times
- You can submit any two assignments up to one week late each
  - Contact me well ahead of time for possible accommodations
- Beyond this 5% of total points will be deducted from your assignment score for each day it is late
- Without prior negotiated accommodations the final project or paper must be submitted by 11:59pm on its due date
- Beyond this, 10% of total points will be deducted from your score for each day the paper or project is late

Project/Paper Topics

- May be reviews of research papers or labs applying big data technology.
- By the end of term you will have completed a paper or conducted an investigational project
  - Where you applied big data technology to a problem of interest to you, your community or your organization
- Research papers are done individually
- Projects are done in teams of 3-4
  - Each student registers for one of 10 project areas or to do a research paper
  - I assign groups of students who express interest in the same project area to a virtual team

BIG DATA CONCEPTS IN DEPTH

The Rise of Big Data

- Companies churn out an increasing volume of transactional data
  - Capturing trillions of bytes of information about their customers, suppliers, and operations
- Millions of networked sensors are being embedded in the physical world
  - In devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data in the age of the Internet of Things (IoT).
- As companies and organizations go about their business and interact with individuals, they are also generating a tremendous amount of data
  - Created as a by-product of other activities.
- Social media sites, smartphones, and other consumer devices including PCs have allowed billions of to contribute to the amount of big data available.
  - And the growing volume of multimedia content has played a major role in the exponential growth in the amount of big data

The Rise of Big Data

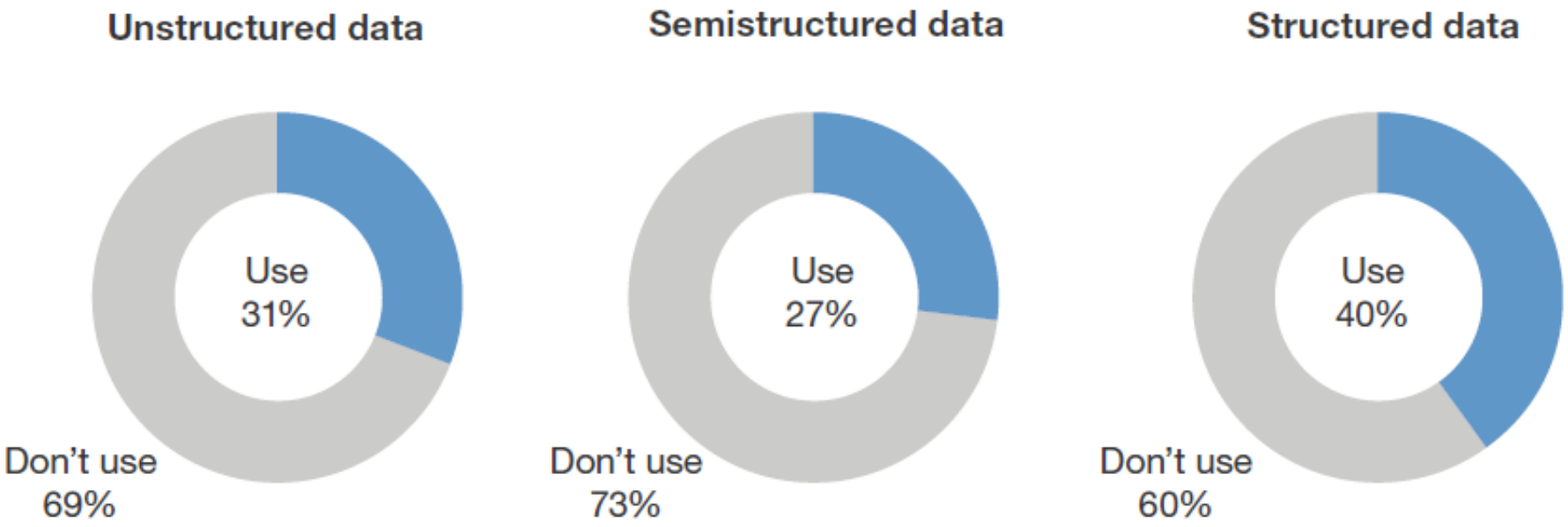
- In itself, the sheer volume of data is a global phenomenon—but what does it mean?
- Many citizens around the world regard this collection of information with deep suspicion
  - Seeing the data flood as nothing more than an intrusion of their privacy.
- But there is strong evidence that big data can play a significant economic role
  - To the benefit not only of private commerce but also of national economies and their citizens.
  - Data can create significant value for the world economy, enhancing the productivity and competitiveness of companies and the public sector and creating substantial economic surplus for consumers

The Challenge Posed by Big Data

- The proliferation of data generated by enterprises has afforded an unprecedented opportunity for businesses to know more about their customers, competitors, and business operations.
  - Internal applications, consumer web/mobile apps, and the Internet of Things (IoT)
- However, the unfortunate truth is that the potential of most data lies dormant.
- On average, between 60% and 73% of all data within an enterprise goes unused for business intelligence (BI) and analytics
  - Forrester's Global Business Technographics Data And Analytics Survey, 2015.
- That's unacceptable in an age where deeper, actionable insights, especially about customers, are a competitive necessity.

The Challenge Posed by Big Data

"Please estimate what percentage of the total size/volume of data within your company your company is currently using for business intelligence (BI)."



Base: 1,805 global technology decision-makers who know how much BI data their firm uses

What is the Value of Big Data?

- if US health care could use big data creatively and effectively to drive efficiency and quality…
  - Estimate that the potential value from data in the sector could be more than $300 billion in value every year,
  - Two-thirds of which would be in the form of reducing national health care expenditures by about 8 percent.
- In the private sector…
  - A retailer using big data to the full has the potential to increase its operating margin by more than 60 percent.
- In the developed economies of Europe…
  - A government administration could save more than €100 billion ($149 billion) in operational efficiency improvements alone

How Does Big Data Affect Our Daily Lives

Sports Predictions

Big Data has been shown to be useful in predicting the outcomes of sporting events; big data was famously used in 2012 to predict that the U.S. would win 108 medals in that years' Summer Olympics in which the U.S. ended up winning 104 medals.

Voting Prediction

Big Data has been used to predict the outcomes of elections. Statistician Nate Silver managed to predict the outcome of the 2012 presidential election with perfect accuracy.

Smartphones

When a smartphone user gets directions, asks their phone a question out loud, or any number of other functions, it is the result of analyzing big data.

How Does Big Data Affect Our Daily Lives

Personalized Advertising and Purchasing Recommendations

One of the primary uses for big data has been in the recommending of purchases and personalization of ads on websites. One study found that a person is more likely to complete Navy Seal training than to actually click a banner ad. Both customers and companies stand to benefit from more personalized and relevant ads.

Improved Traffic Flow

Several companies and cities have utilized big data to streamline the flow of traffic in their towns. Using data derived from drivers' GPS signals to react in real time to traffic conditions, weather, accidents, etc. in order to maintain smooth traffic flow.

Epidemic Detection and Prevention

Big data has recently come into use by Google and more recently by the traditional medical establishment to predict where outbreaks of potentially epidemic viruses such as the flu are most likely to appear.

Big Data Use Cases

Smarter Healthcare

Homeland Security

Traffic Control

Manufacturing

Multi-channel sales

Telecom

Trading Analytics

Search Quality

Big Data Use Cases

Big Data Exploration

Find, visualize, understand all big data to improve decision making

Enhanced 360° View of the Customer

Extend existing customer views (MDM, CRM, etc) by incorporating additional internal and external information sources

Security/Intelligence Extension

Lower risk, detect fraud and monitor cyber security in real-time

Operations Analysis

Analyze a variety of machine data for improved business results

Data Warehouse Augmentation

Integrate big data and data warehouse capabilities to increase operational efficiency

Changing Perspective

Old Model: Few companies are generating data, all others are consuming data

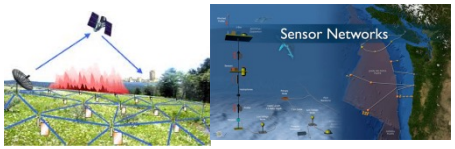
New Model: all of us are generating data, and all of us are consuming data

Who's Generating Big Data

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

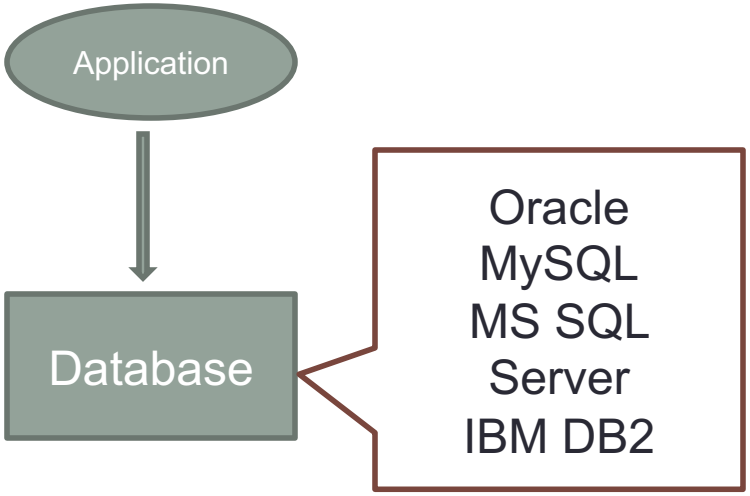
Key Enablers

- Increase of storage capacities
- Increase of processing power
- Increase of availability of data
- Availability of the public cloud

The Road to Big Data Technology

- You start out with a traditional database
  - SQL
  - Schemas
  - Relational Model

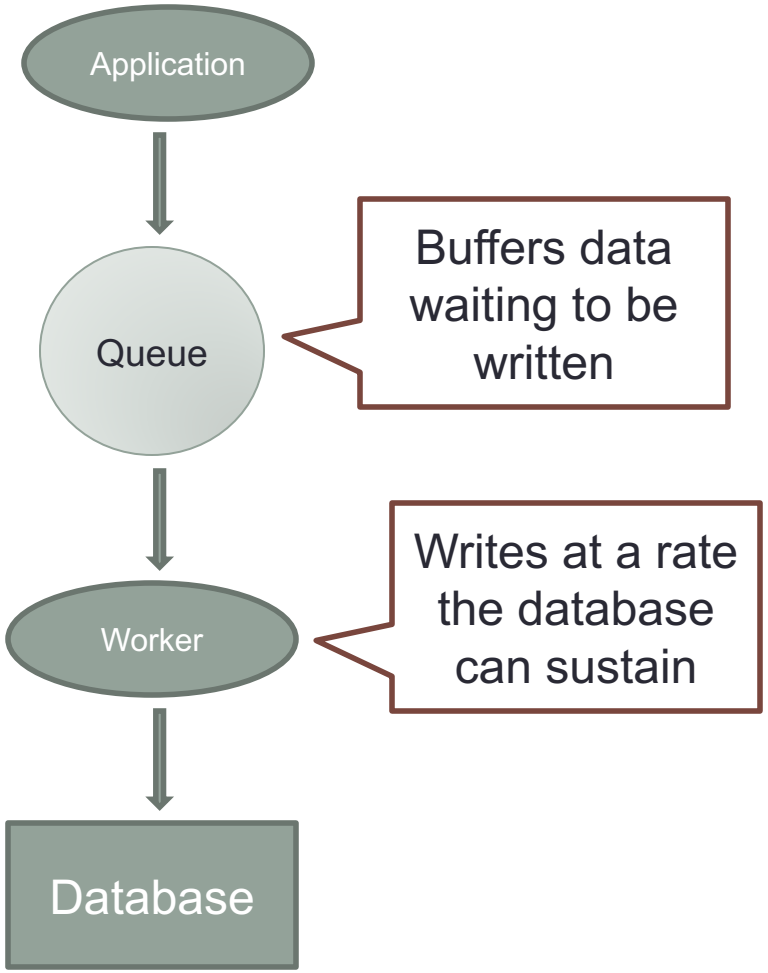
But… at some point the database can't keep up with the write rate



The Road to Big Data Technology

- So you add a queue layer to buffer writes until the database can handle them

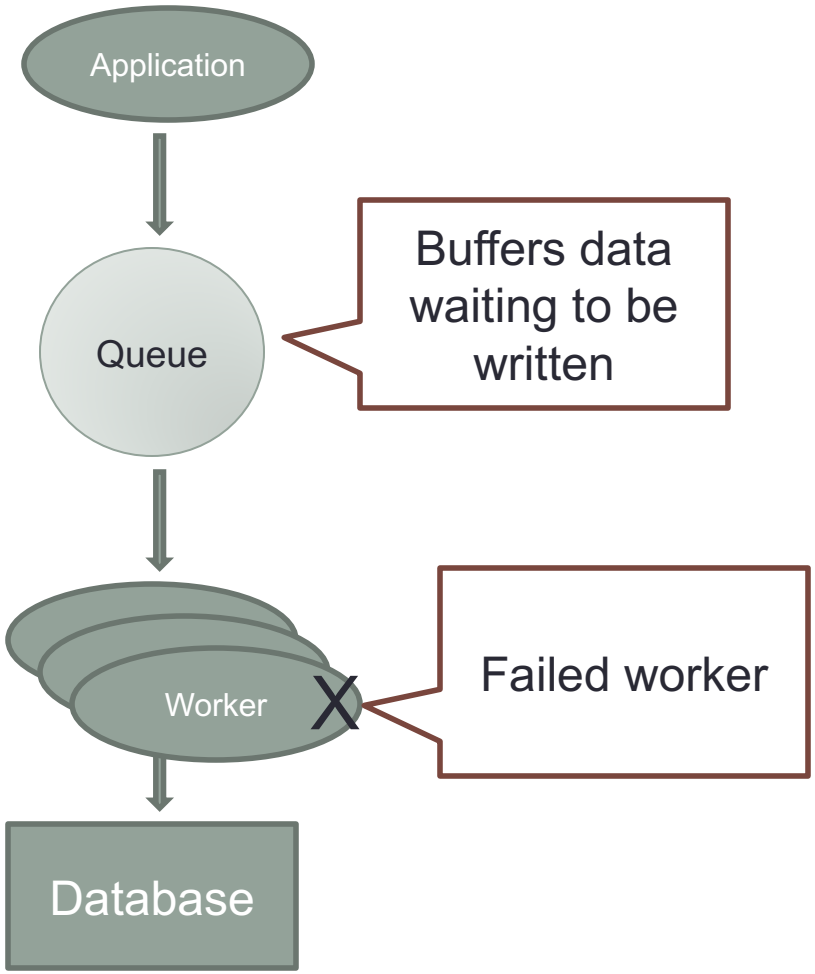
But… at some point the worker cant' keep up & the queue buffer starts to grow out of control



The Road to Big Data Technology

- So you add more workers to operate in parallel

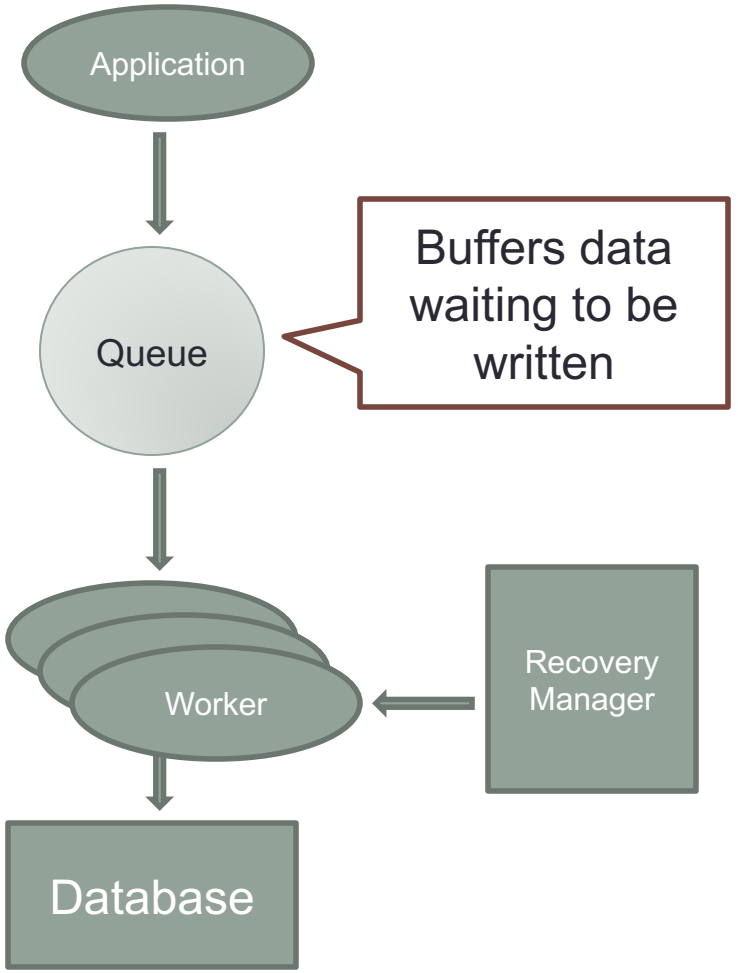
But… at some point one or another worker fails



The Road to Big Data Technology

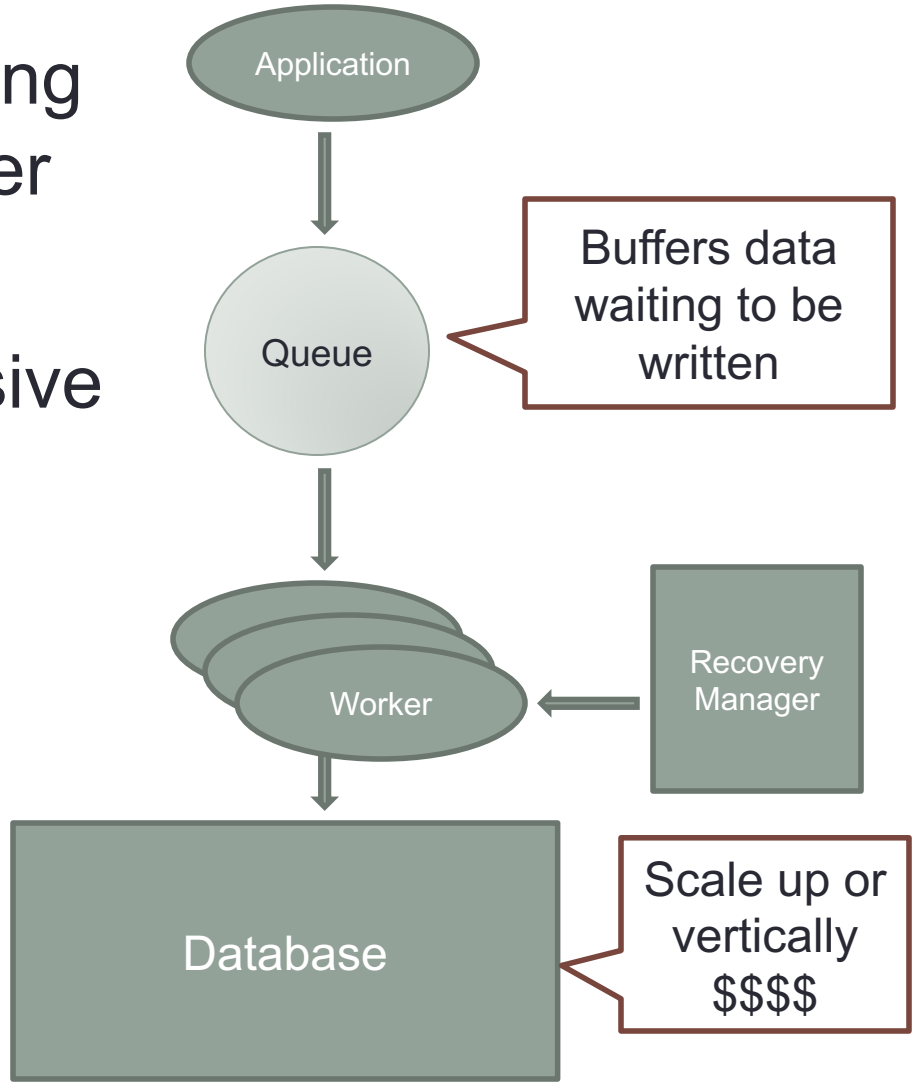
So you need to design recovery mechanisms

But… at some point the database can't process writes fast enough



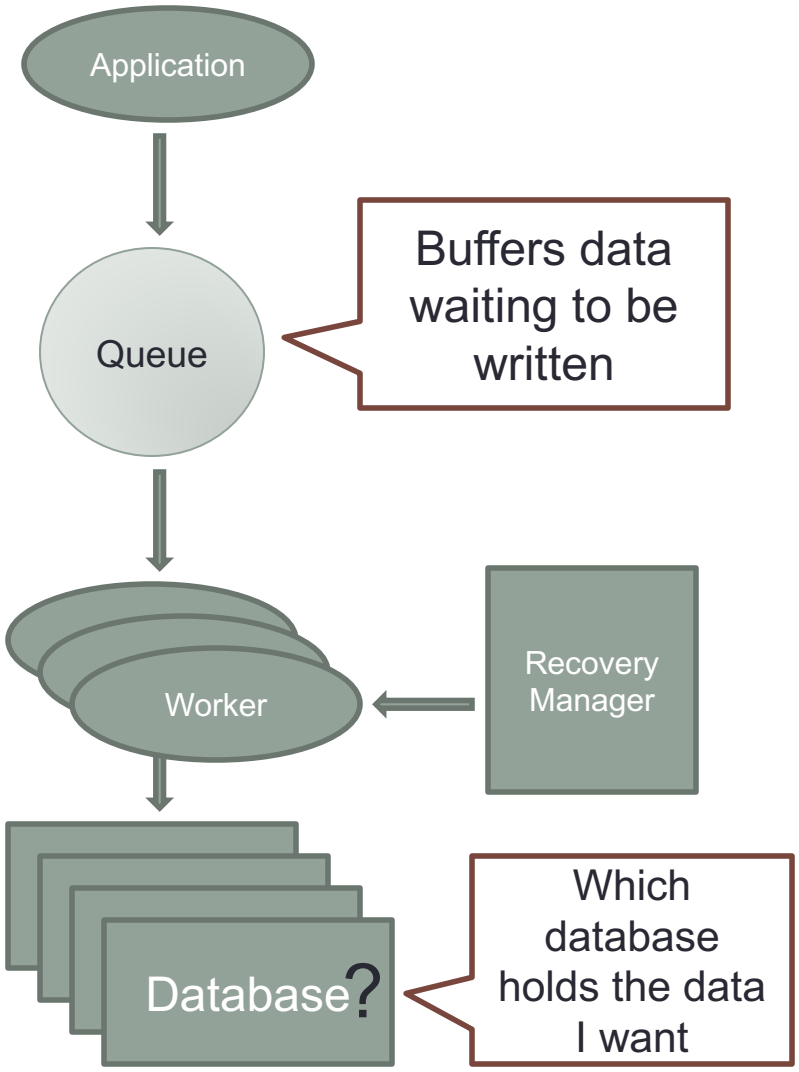
The Road to Big Data Technology

- So you think about buying a bigger database server
- But… this is too expensive



The Road to Big Data Technology

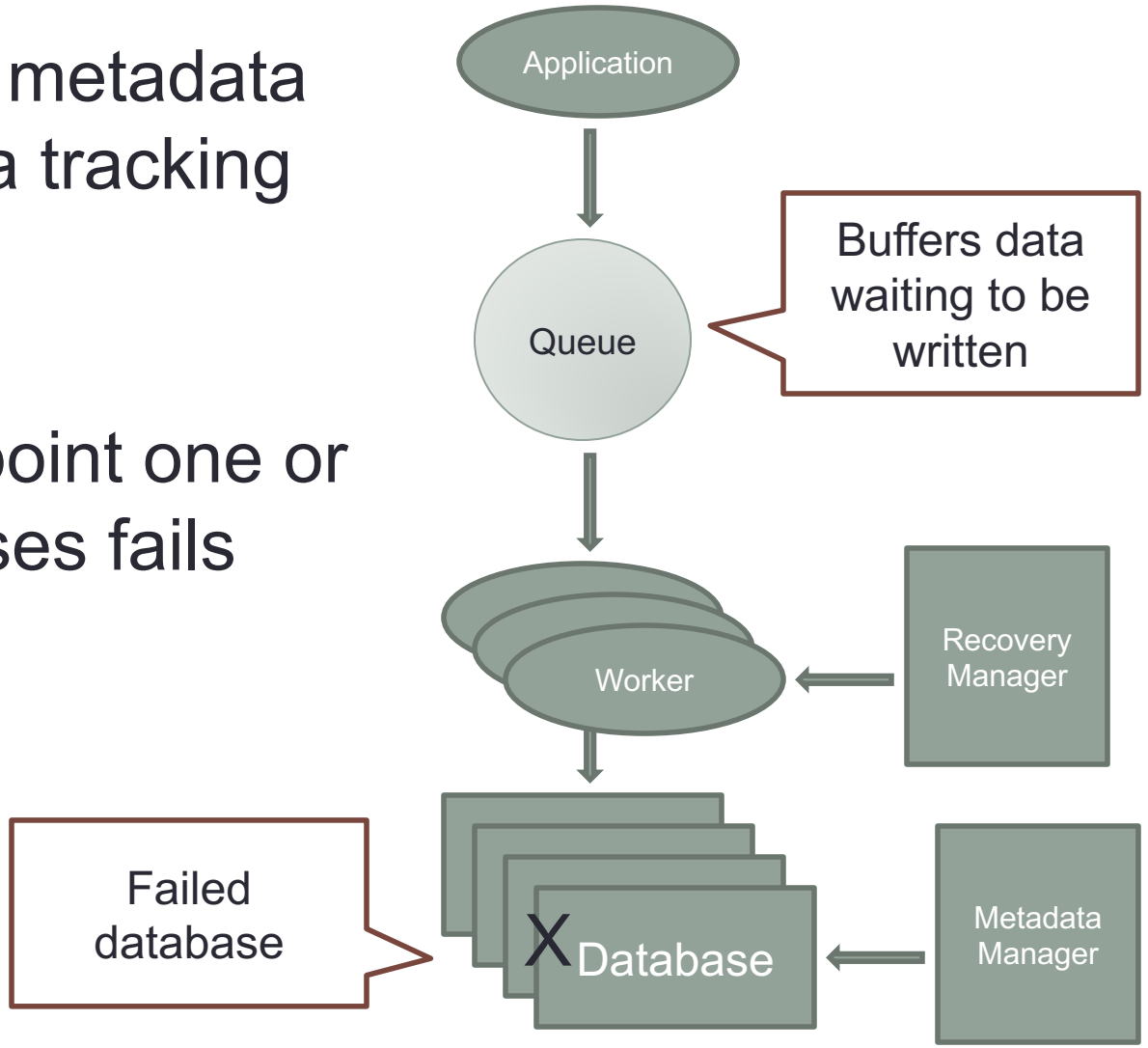
- So you divide your data across multiple smaller cheaper database servers
- But.. It becomes more difficult to manage the data



The Road to Big Data Technology

So… you add a metadata manager (a data tracking system)

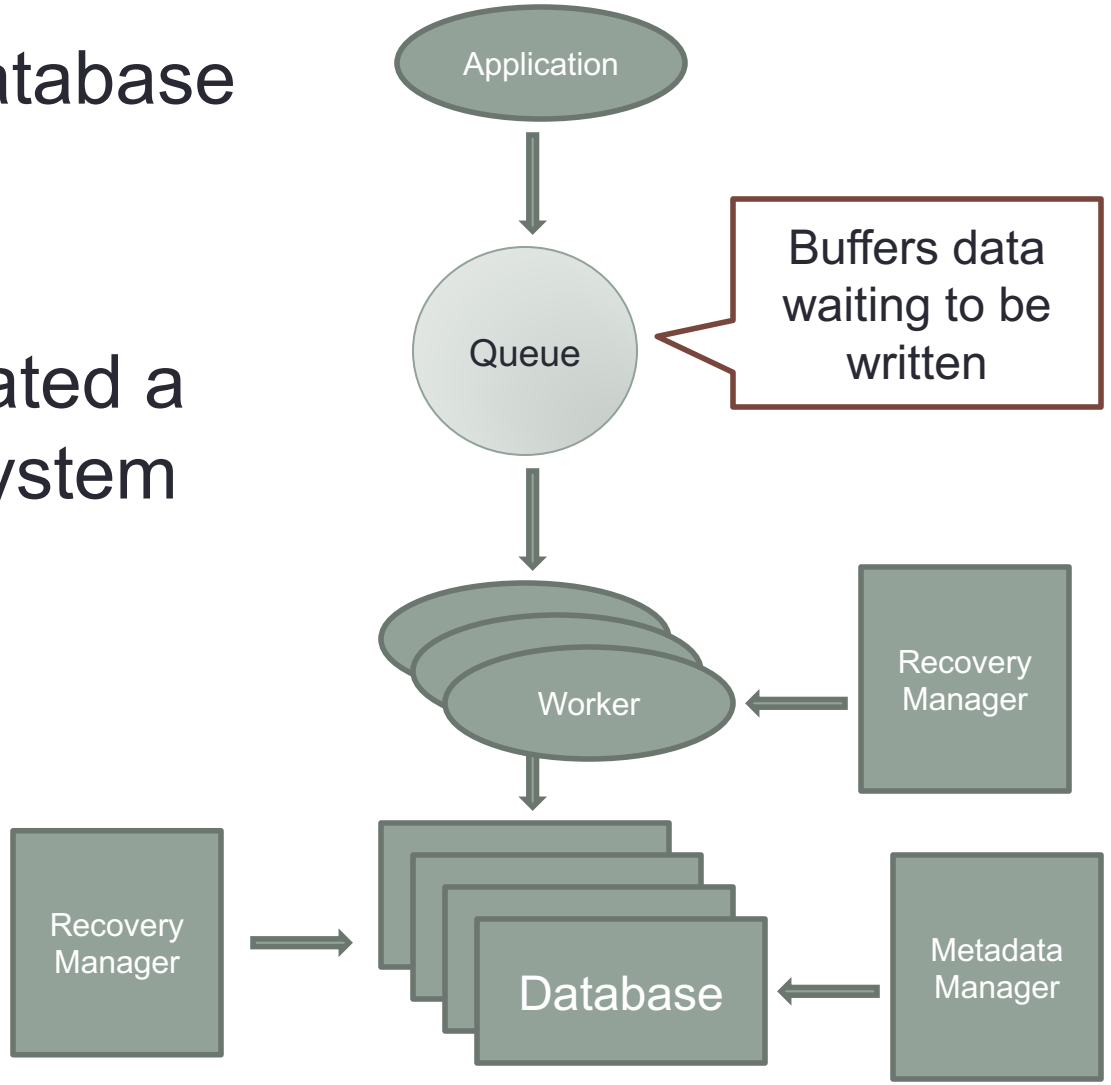
But… at some point one or another databases fails



The Road to Big Data Technology

So… you add a database recovery manager

But… now we created a complex ad-hoc system



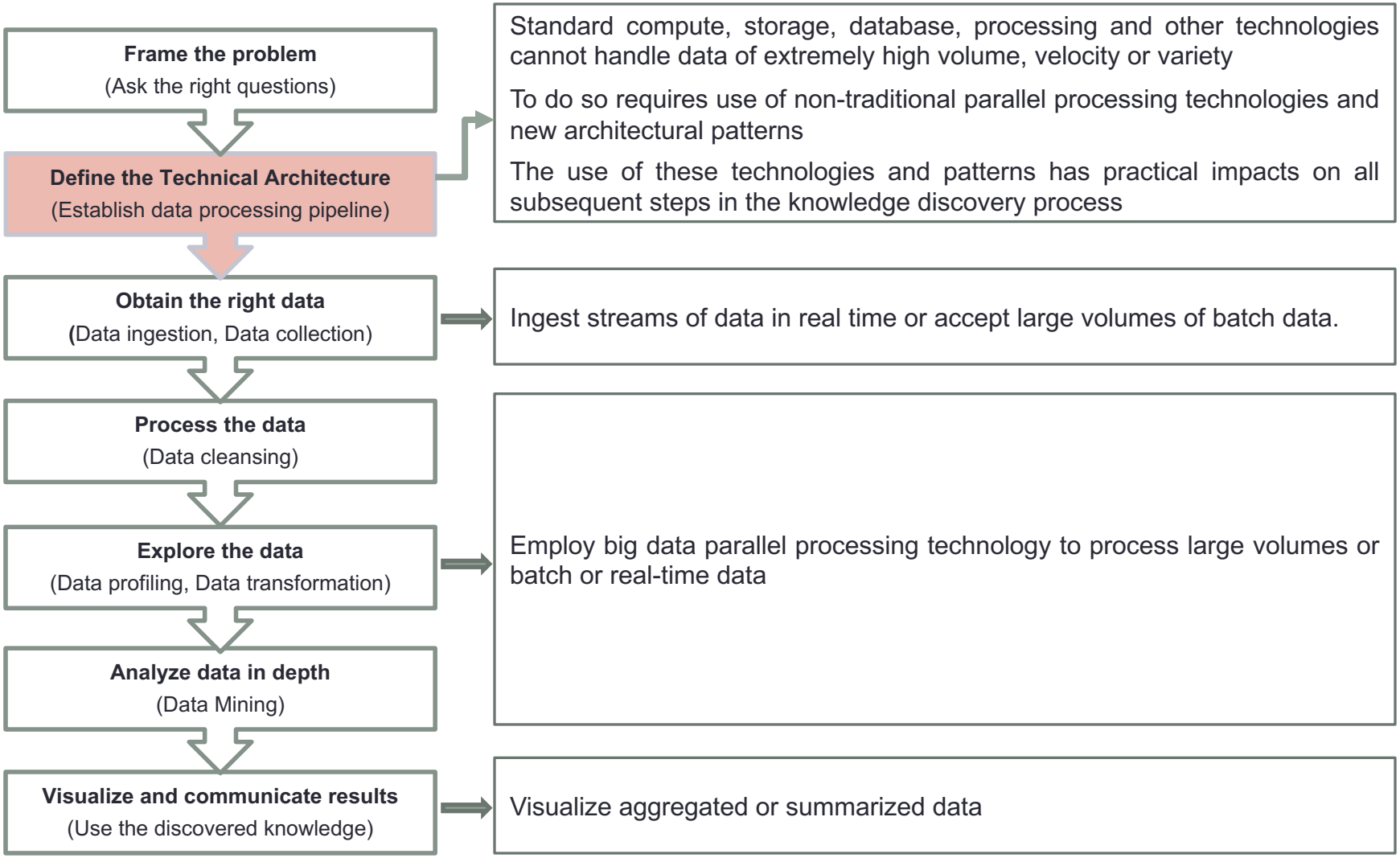
What You Really Need

Purpose Built Big Data System

- High availability and fault tolerance
- Low latency reads and writes
- Inexpensive scalability
  - Infinite compute capability
  - Infinite storage capability
  - Using commodity hardware
  - Or using cloud services
- Flexibility
  - Accepts all varieties of data
  - Allows all sorts of queries
  - Easy changes to schemas
- Ease of use
  - Easy to learn
  - Minimal maintenance
- Extensible
  - A range of software tools and templates

Knowledge Discovery Process

Impact of Big Data



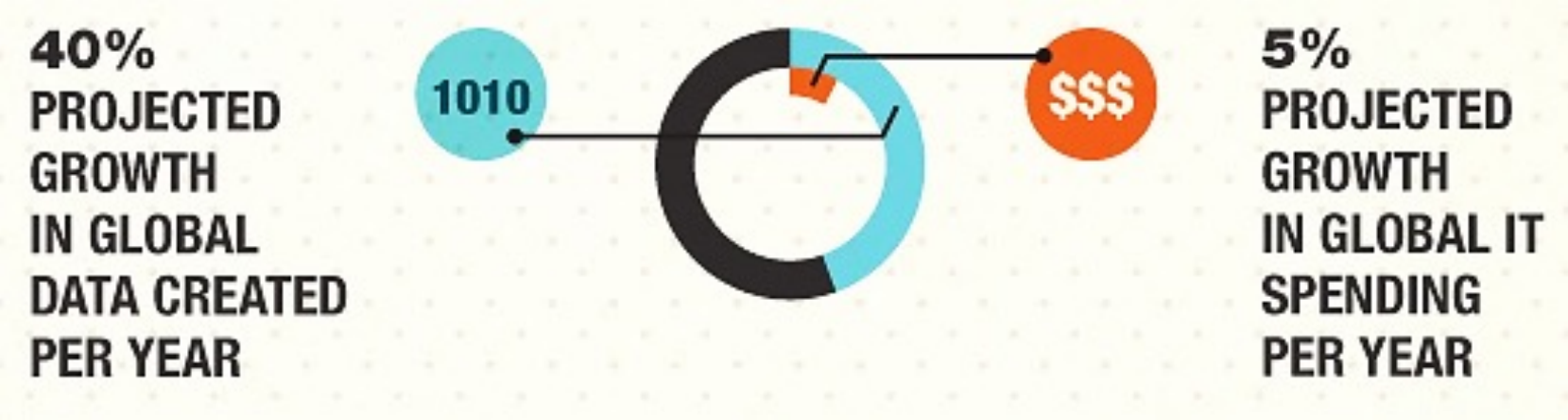
Why is Big Data a Challenge?

- Don’t collect so much data in the first place
- Or just store part of the data until later when it can be processed
- Or delete part of the data that can’t be stored or processed using the technology to hand
- Or sample the data at the rate at which is can be processed and stored
- And then discard the data that can’t be stored

Why is Big Data a Challenge?

- Sometimes collecting large volumes of data is the point
  - Large Hadron Collider experiments to discover new physics
  - Images and other data from missions to the moon or Mars
  - Twitter, Facebook, Snapchat, YouTube, Instagram and so on
- Sometimes big data happens (unexpectedly)
  - An ecommerce site starts small and then usage grows
  - Data collected about each user starts small and then is augmented with more details
- Sometimes big data is a competitive business choice
  - Understand your customer better than other similar businesses
  - Or to meet some governance, legal or regulatory requirements
- Sometimes it makes no sense to only sample data
  - Can't record or analyze every other order a customer places
- Sometimes more data is required for scientific (data analytic) reasons
  - More effective sentiment analysis or population health trending analysis
  - Good predictive algorithms applied to large data sets are often more accurate than great algorithms applied to smaller data sets

Another Challenge of Big Data



Data volume is growing 8X faster than IT budgets
=> We must find more cost efficient ways of managing big data

Organizational Challenges

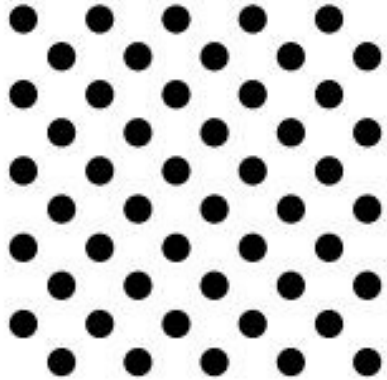
- **Legacy IT:** Many organizations are trying to manage healthcare and operational data using stagnant, closed, archaic clinical systems.
- **Data volume:** Organizations must gather and make sense of all sorts of structured,
- **Data format:** unstructured, internal, and external data from a myriad amount of sources from clinical systems to machine data (e.g., heart monitors).
- **Siloes:** Data residing in a variety of systems that support different operational processes result in error prone, largely inefficient, and lost data.
- **Lack of integration:** Disconnected systems like EMRs and clinical systems fail to provide the holistic view needed to identify gaps and opportunities.
- **Skills and tools:** Keeping up with fast-changing and advancing Big Data technologies is a daunting task when time and resources are already constrained and competition for industry and technical talent remains fierce.

Why Are Big Data Systems Different?

- The basic requirements for working with big data are the same as the requirements for working with datasets of any size.
- However, the massive scale, the speed of ingesting and processing, and the characteristics of the data that must be dealt with at each stage of the process present significant new challenges when designing solutions.
- The goal of most big data systems is to surface insights and connections from large volumes of heterogeneous data that would not be possible using conventional methods.
- In 2001, Gartner's Doug Laney first presented what became known as the "three Vs of big data" to describe some of the characteristics that make big data different from other data processing
  - Volume, Velocity, Variety

Attributes of Big Data (3 Vs)

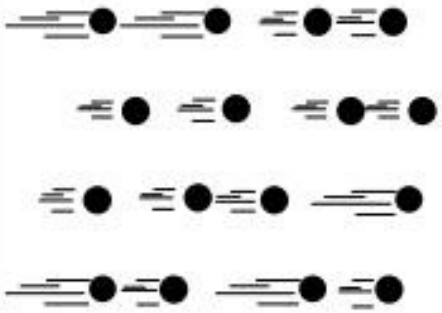
Volume



Data at Rest

Terabytes to exabytes of existing data to process

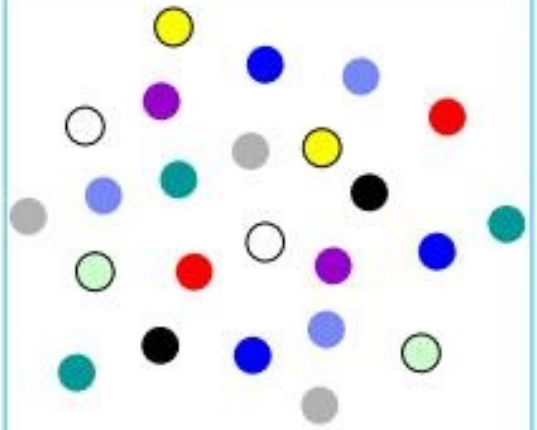
Velocity



Data in Motion

Streaming data, milliseconds to seconds to respond

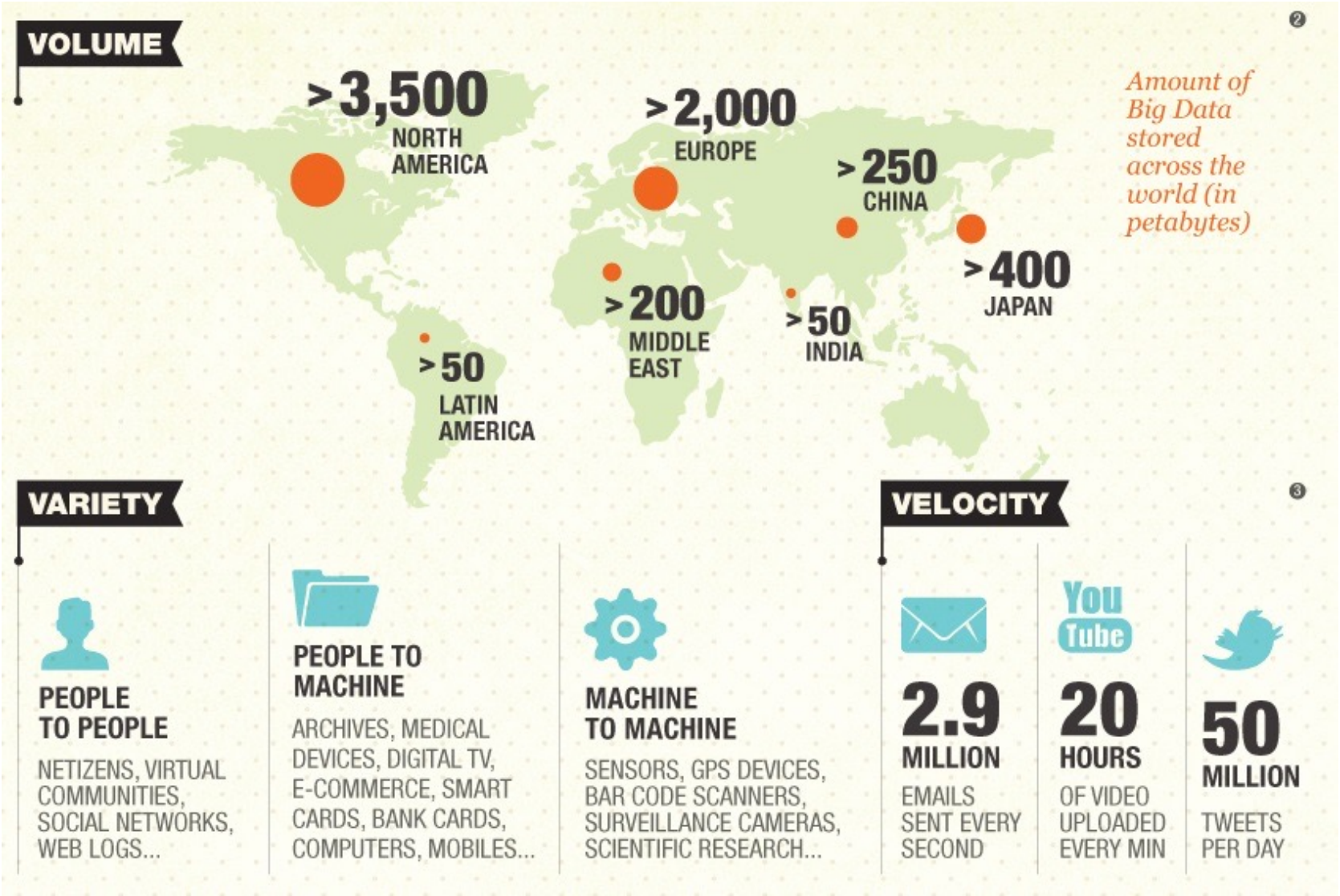
Variety



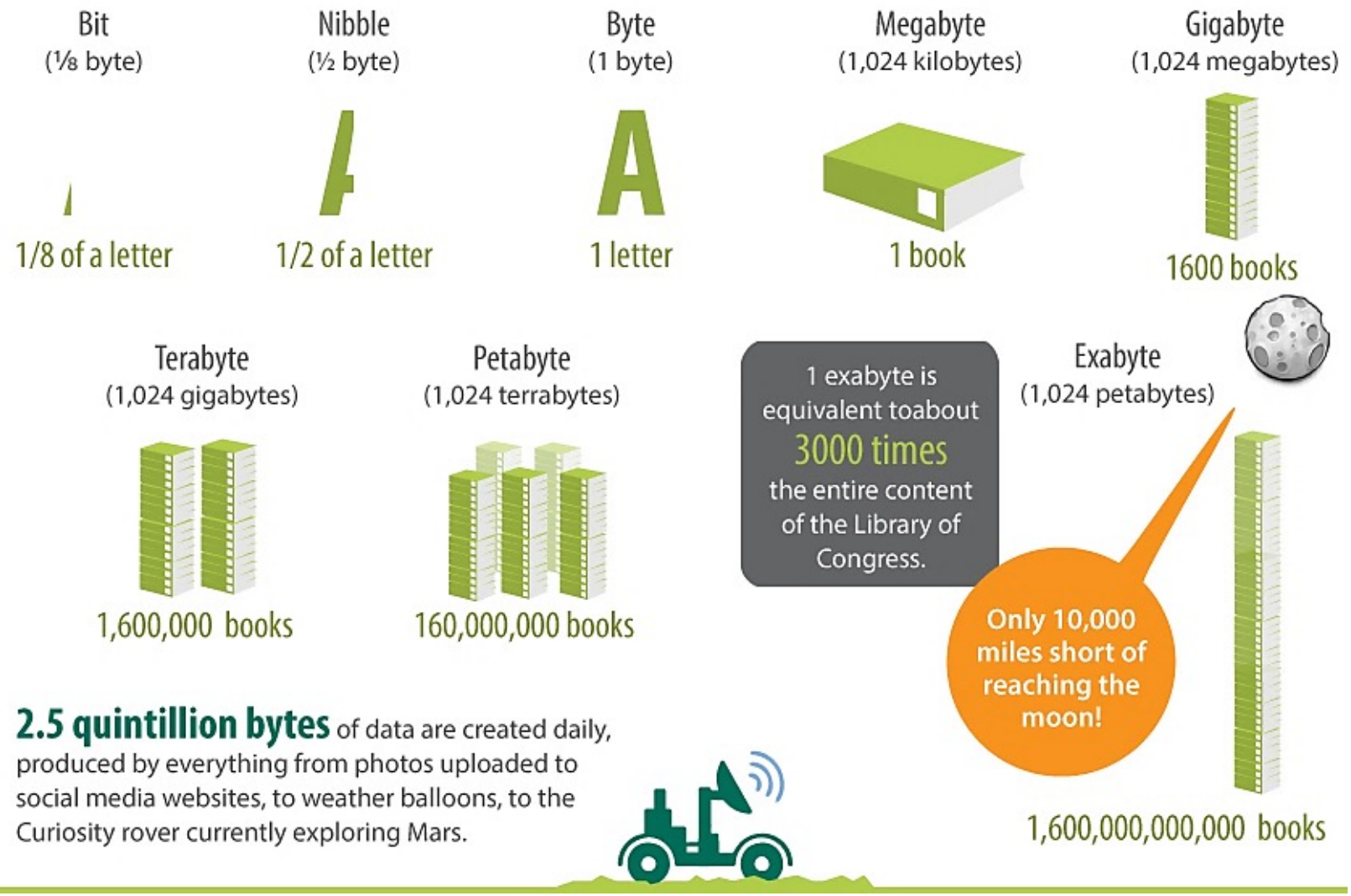
Data in Many Forms

Structured, unstructured, text, multimedia

Some “Facts” About the 3 V’s



Volume


Bit
(⅛ byte)
1/8 of a letter
Nibble
(½ byte)
1/2 of a letter
Byte
(1 byte)
1 letter
Megabyte
(1,024 kilobytes)
1 book
Gigabyte
(1,024 megabytes)
1600 books
Terabyte
(1,024 gigabytes)
1,600,000 books
Petabyte
(1,024 terrabytes)
160,000,000 books
1 exabyte is equivalent toabout 3000 times the entire content of the Library of Congress.
Exabyte
(1,024 petabytes)
Only 10,000 miles short of reaching the moon!
1,600,000,000,000 books
2.5 quintillion bytes of data are created daily, produced by everything from photos uploaded to social media websites, to weather balloons, to the Curiosity rover currently exploring Mars.

When is Big Data "Big?"

- The 1980 US Census data would be considered as Big Data in the 1980s, where the IBM 3850 Mass Storage System with a capacity of 102.2GB was the monster storage device of its day
- This would certainly not be considered as Big Data today where a personal computer can afford such capacity
- In today's environment, the size of datasets that may be considered as Big Data could range from
  - Terabytes (10^{12} bytes)
  - Petabytes (10^{15} bytes)
  - Exabytes (10^{18} bytes)
  - Depending on the industry, how data is used

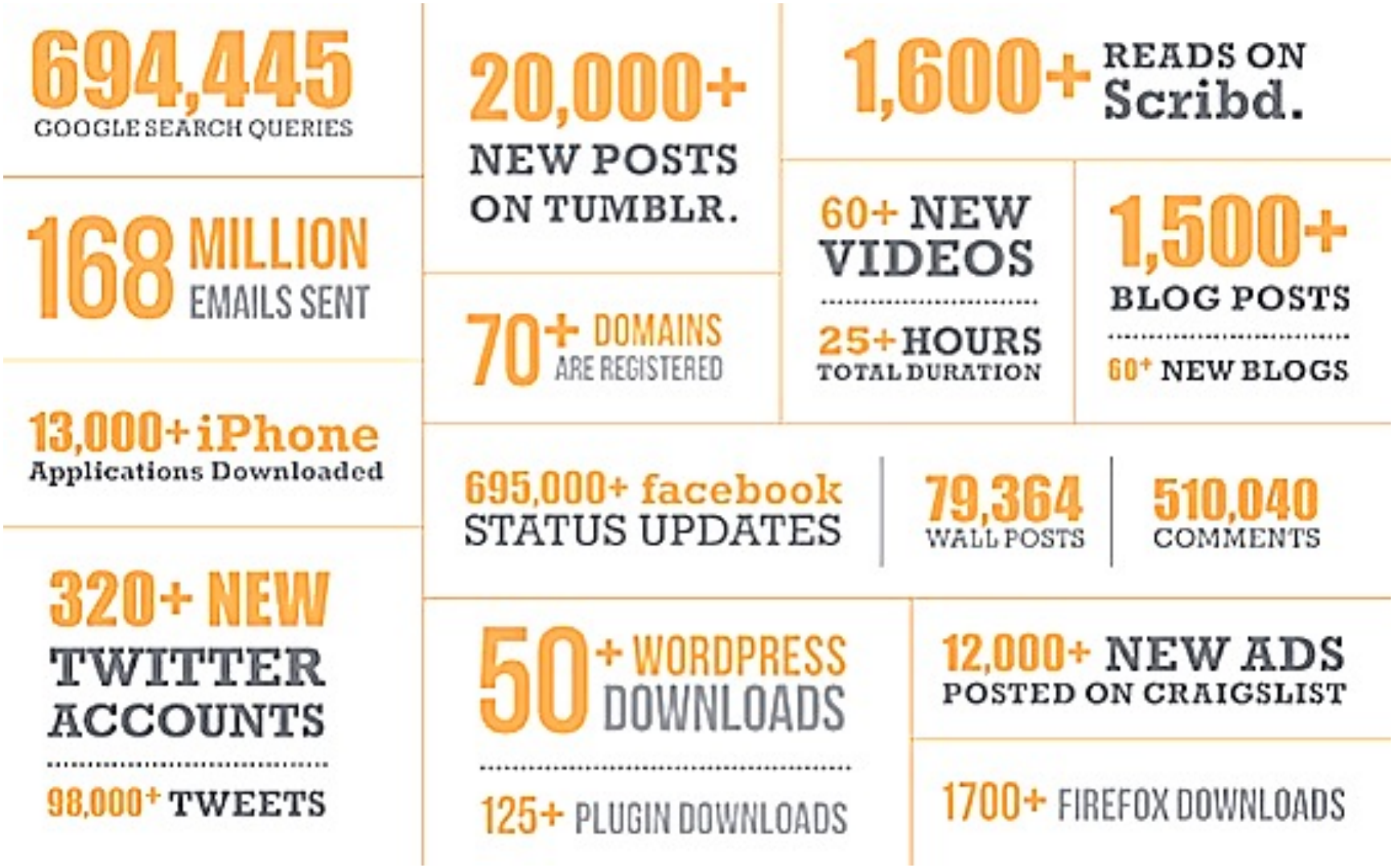
Volume

- The sheer scale of the information processed helps define big data systems.
- These datasets can be orders of magnitude larger than traditional datasets
  - Which demands more thought at each stage of the processing and storage life cycle.
- Often, because the work requirements exceed the capabilities of a single computer, this becomes a challenge of pooling, allocating, and coordinating resources from groups of computers.
- Cluster management and algorithms capable of breaking tasks into smaller pieces become increasingly important.

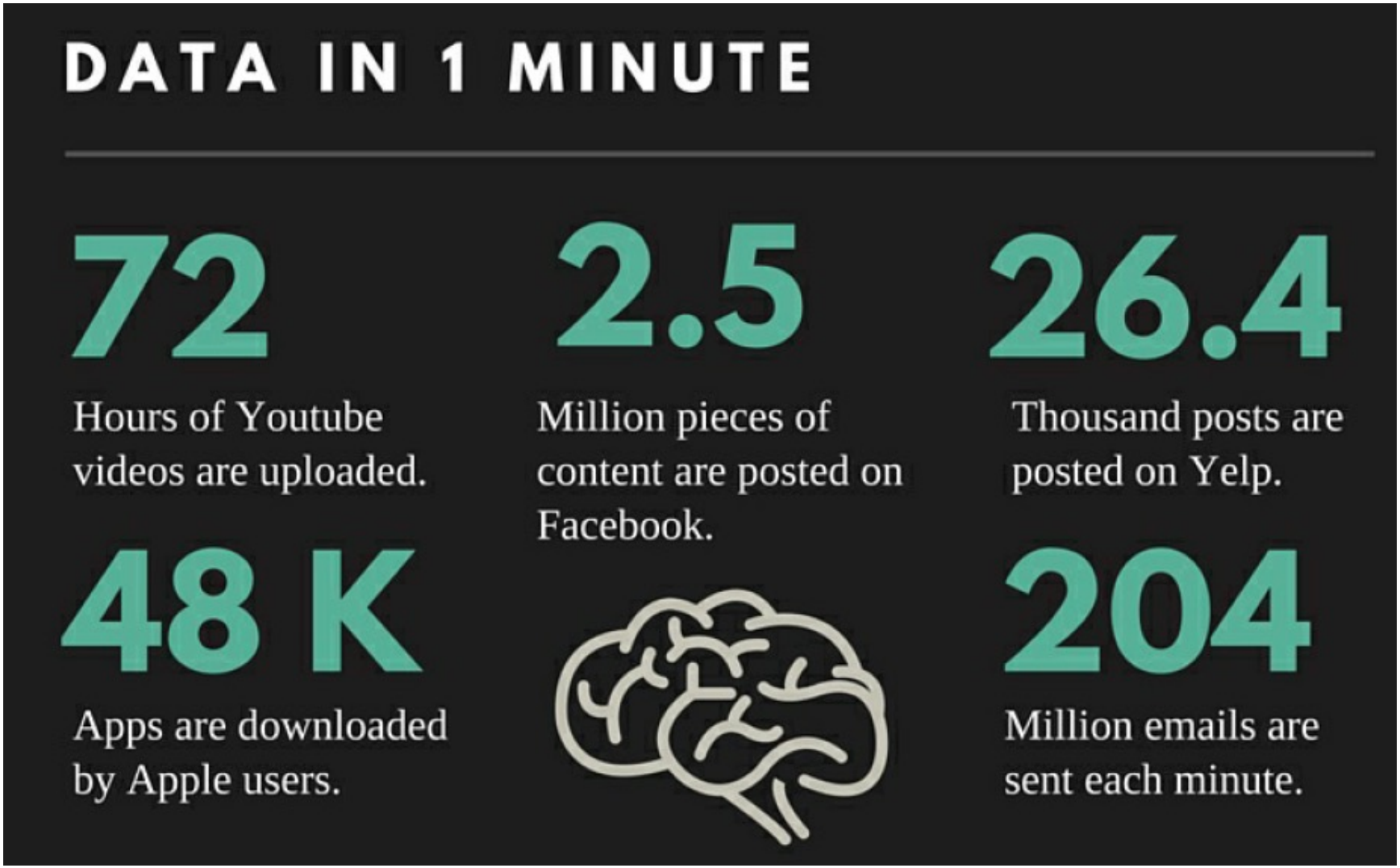
Volume

- A typical PC might have had 10 gigabytes of storage in 2000.
- Today, Facebook ingests 500 terabytes of new data every day.
- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.
- The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds
  - Containing environmental, location, and other information, including video.

Data Collected in One Minute



Some “Facts” About Big Data



Four Domains of Big Data in 2025

In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

A Brief Interlude…

How can we store all this data?

What if a data file is larger than any single SSD (solid state drive) or hard disk?

What technology and what storage architecture should be used?

How can we do so reliably?

And what might this cost?

A Brief Interlude…

We are going to explore a range of approaches to addressing these questions

But let's start with one that is gaining popularity

Public cloud hosted block storage as a service

But First Let's Think of File Systems

- When you access a file system you must do so through a server or virtual machine running an operating system
- So file storage, operating system and compute capability are all coupled together to some degree
- For example, the Linux file system is organized differently to that provided via MS Windows
- And that means that Windows files can only be read if you have an active Windows machine (or translation software)
- The storage systems that file systems are built over CAN be expanded to files of almost and size
- But this requires costly hardware along with operational maintenance, and capability for backup and restore

Now Let's Think of the Internet

- You can retrieve a pdf document, a song, a web page, an image, an HTML web page, and so forth
- Consider the providers of this content as offering storage services holding content until you request it
- And the interface to these services is the HTTP protocol
  - GET content
  - PUT content
- These services are operating system independent and leave it to the requestor to interpret the stored content
- And the requestor does not care through what means or where the content was stored

Now Let's Think of the Internet

- There is no notion of directories or subdirectories, just a host name and a path to the content of interest

  http://www.somehostname.com/documents/grades.pdf

- The host name is the name of a container for content

  www.somehostname.com

- The path identifies the specific content item

  /documents/grades.pdf

- And the interface to these services is via the HTTP (or REST) protocol

From the Internet to Object Stores

What if we were able to provide storage of content
Independent of servers, operating systems and file systems?

What if we made this storage available as a (REST) service over the Internet?

What is this service could store terabytes to petabytes of data?

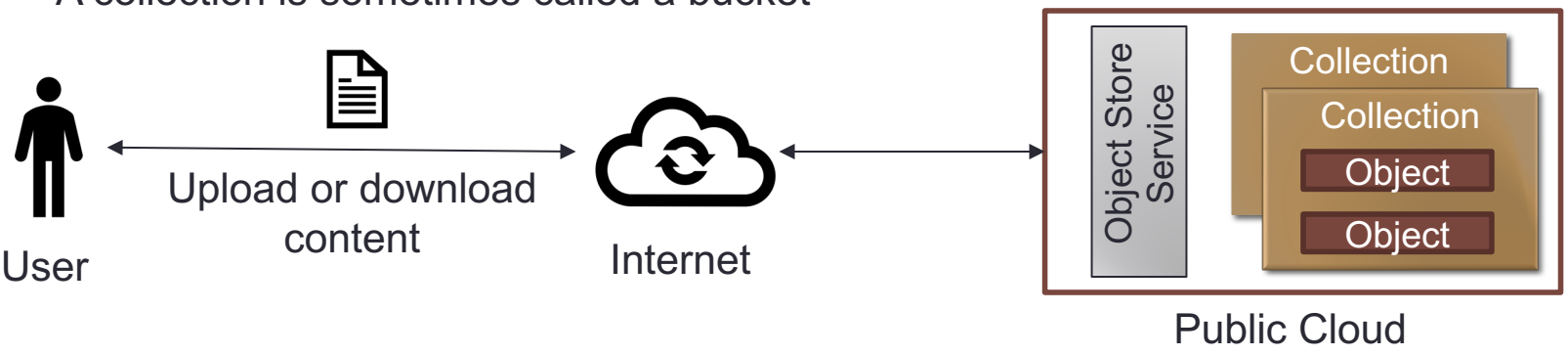
What is this service was highly available, very durable, secure?

Oh, and what if it was very cheap, too?

Then we would have an **Object Store**

From the Internet to Object Stores

- Object stores are often cloud services accessible from the Internet
- Instead of providing other's content they allow users to upload (store) and download (retrieve) content of their own
- Allows content of nearly unlimited size to be uploaded & downloaded
- In an object store each content item is stored as an object
- An object consist of a globally unique identifier (GUID), some metadata, and the content itself
  - An objects GUID is also know as its key
- An object store holds groups of related object together in collections, each having a unique name
  - A collection is sometimes called a bucket



From the Internet to Object Stores

- An object is data (think a file) tagged with a key (think file name)
- Object data is treated as just a sequence of bytes (think blob)
- A bucket is container of objects (think directory) tagged with a name
- Buckets hold objects only and not other buckets (think no sub-directories)
- A bucket name must be unique across all of AWS
- If the bucket name is “myawsbucket” it is addressed as
  https://myawsbucket.s3.amazonaws.com
- An object key must be unique within a bucket
- If the ley is “myobject” then it can be addressed as
- https://myawsbucket.s3.amazonaws.com/myobject
- Objects are “write once, read any” (WORM)
  - You can only write a whole object at a time
  - Once the object is written you can’t update or append to it

Amazon S3

- Simple Storage Service
- Secure, durable, highly-scalable object storage
- Accessible via a simple web services interface
- Store & retrieve any amount of data
- Use alone or together with other AWS services

Amazon S3 Concepts

- Buckets
  - Containers for objects stored in S3
  - Organize the Amazon S3 namespace at the highest level
  - Each bucket has a name unique across AWS
- Objects
  - Fundamental entities stored in Amazon S3
  - Consist of data & metadata
  - Data portion is opaque to Amazon S3
  - Metadata is a set of name-value pairs that describe the object
  - Object is uniquely identified within a bucket by a key (name) and a version ID

Amazon S3 Concepts

- Keys
  - Unique identifier for an object within a bucket.
  - Every object in a bucket has exactly one key
  - Combination of a bucket name, object key & version ID uniquely identify each object

Amazon S3 Concepts

Globally Unique

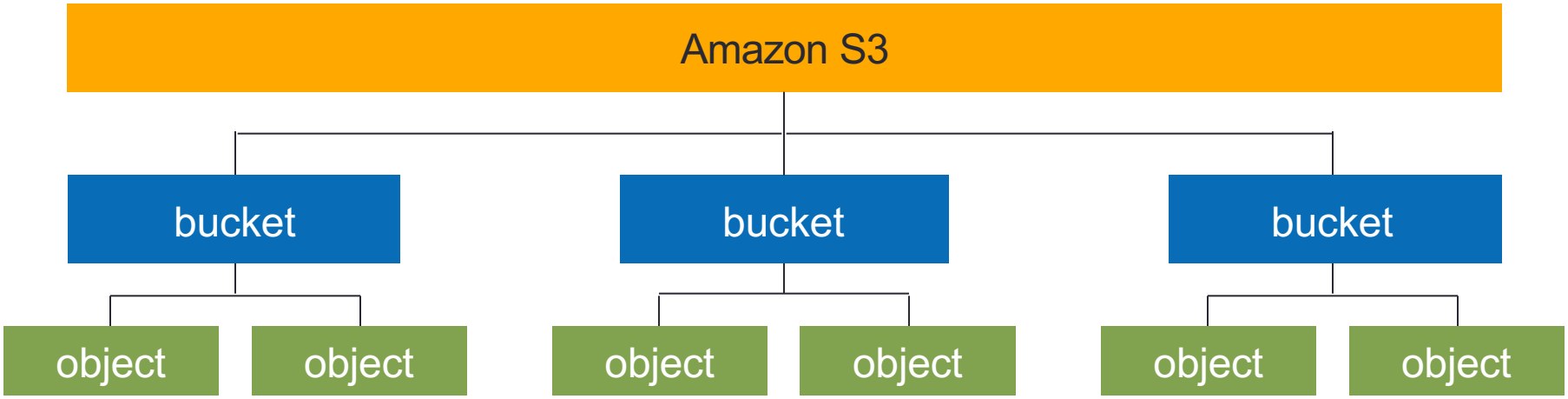
Bucket Name + Object Name (key)

Amazon S3 Concepts

Globally Unique

⇓

Bucket Name + Object Name (key)

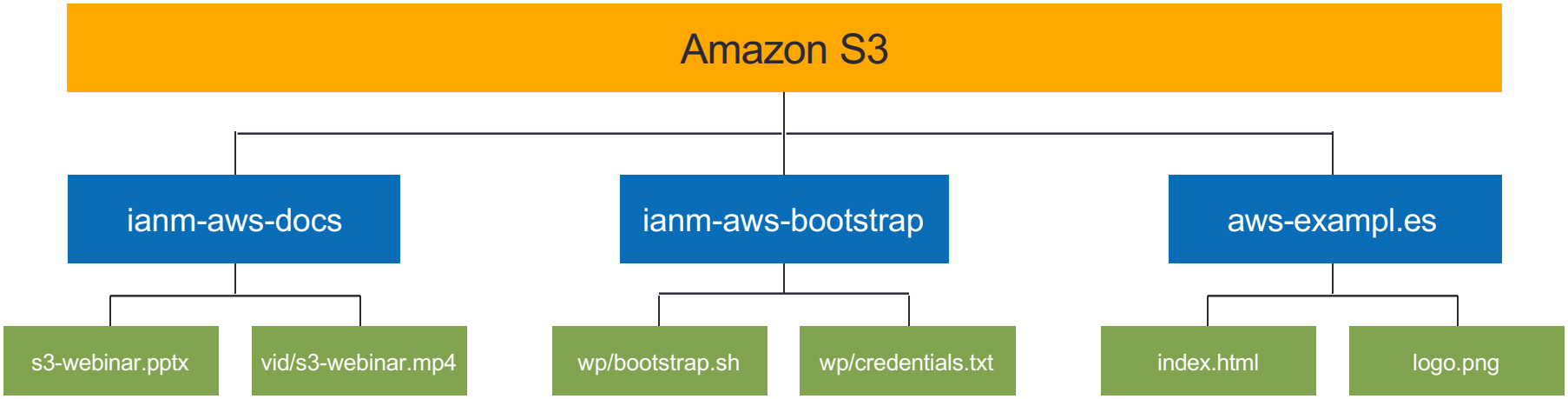


Amazon S3 Concepts

Globally Unique

⇓

Bucket Name + Object Name (key)



Amazon S3 Concepts

Object key

Max 1024 bytes UTF-8

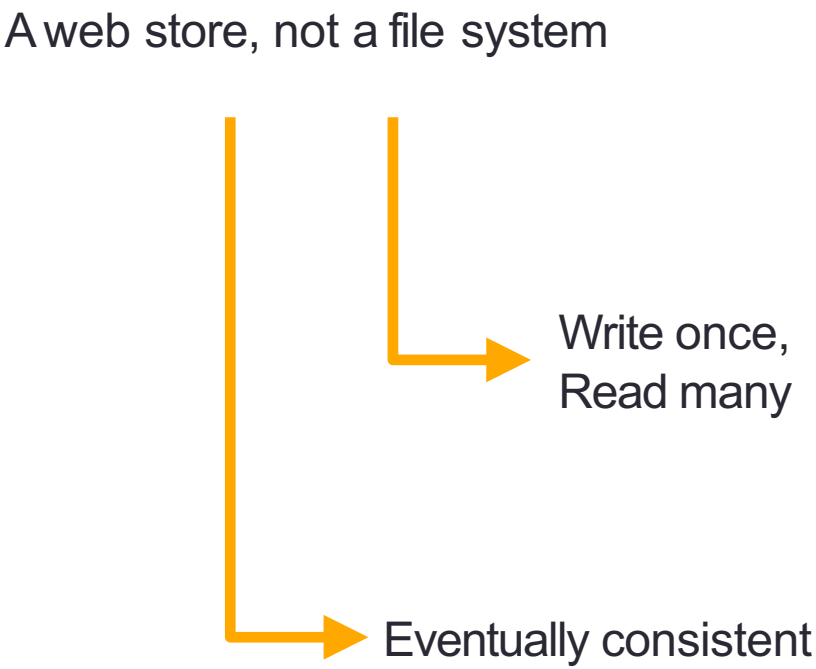
Including ‘path’ prefixes

Unique within a bucket

assets/js/jquery/plugins/jtables.js

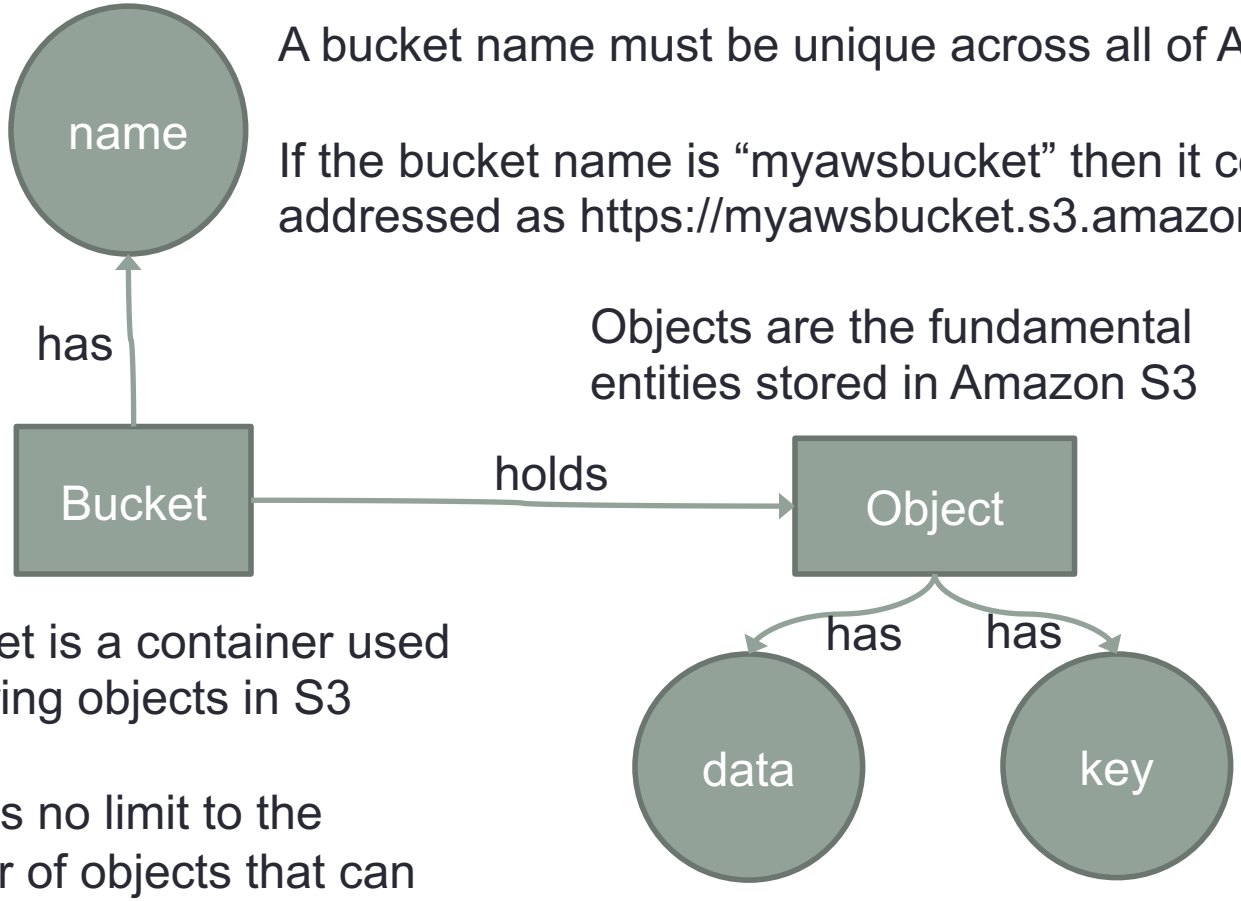
an example object key

Amazon S3 Concepts



Amazon S3 Concepts

Amazon S3 Concepts



A bucket name must be unique across all of AWS

If the bucket name is “myawsbucket” then it could be addressed as https://myawsbucket.s3.amazonaws.com

Objects are the fundamental entities stored in Amazon S3

A bucket is a container used for storing objects in S3

There is no limit to the number of objects that can be stored in an S3 bucket

An object is associated with a sequence (blob) of bytes

An object is uniquely identified within a bucket by a key (name)

If the key is “myobject” then it can be addresses as https://myawsbucket.s3.amazonaws.com/myobject

Velocity

- Another way in which big data differs significantly from other data systems is the speed that information moves through the system.
- Data is frequently flowing into the system from multiple sources and is often expected to be processed in real time
  - To gain insights and update the current understanding of the system.
- This focus on near instant feedback has driven many big data practitioners away from a batch-oriented approach and closer to a real-time streaming system.
- Data is constantly being added, massaged, processed, and analyzed in order to keep up with the influx of new information.
- This require robust systems with highly available components to guard against failures along the data pipeline.

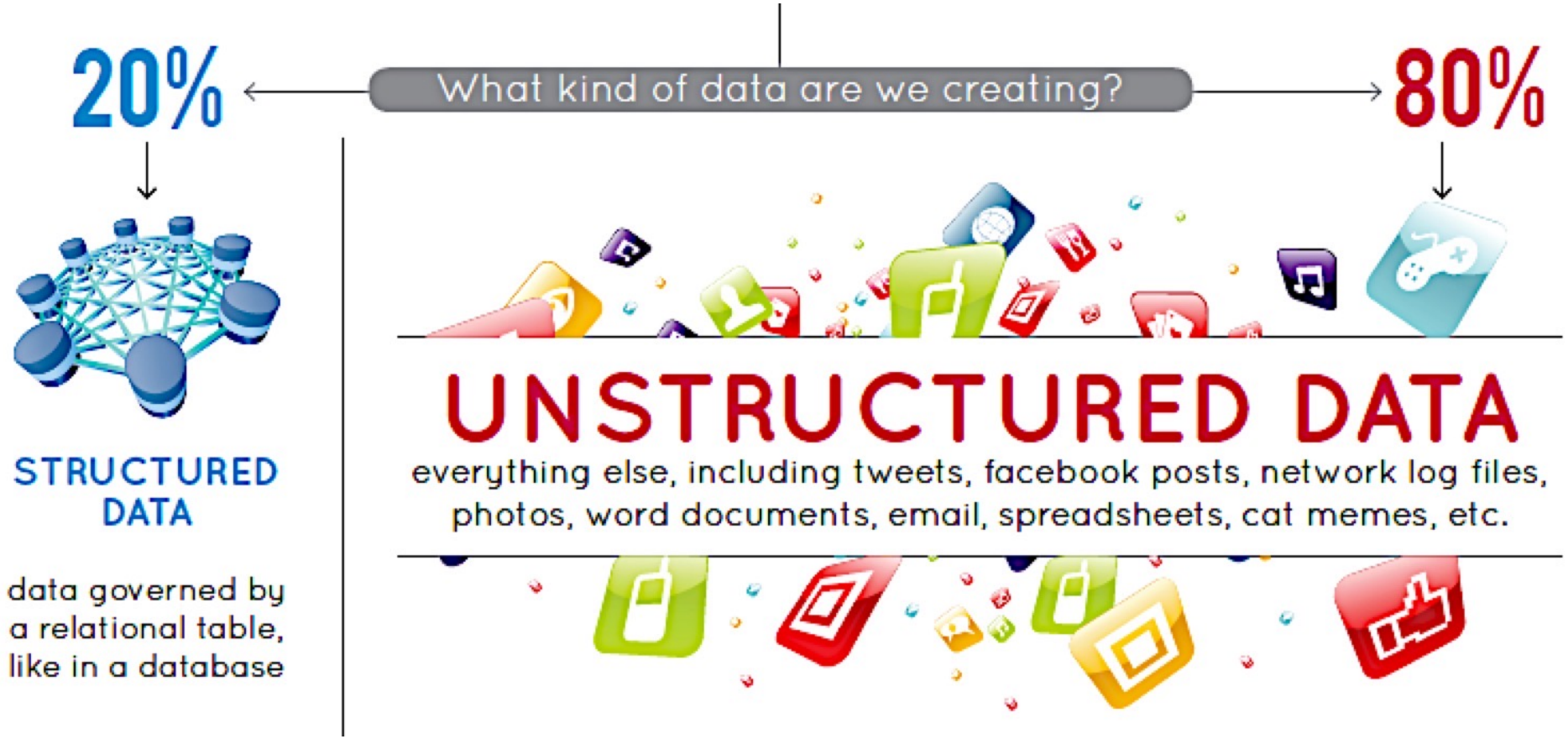
Velocity

- Some Internet of Things (IoT) applications have health and safety ramifications that require real-time evaluation and action.
- Other internet-enabled smart products operate in real-time or near real-time.
- As an example, consumer eCommerce applications seek to combine mobile device location and personal preferences to make time sensitive offers.
- Operationally, mobile application experiences have large user populations, increased network traffic, and the expectation for immediate response.

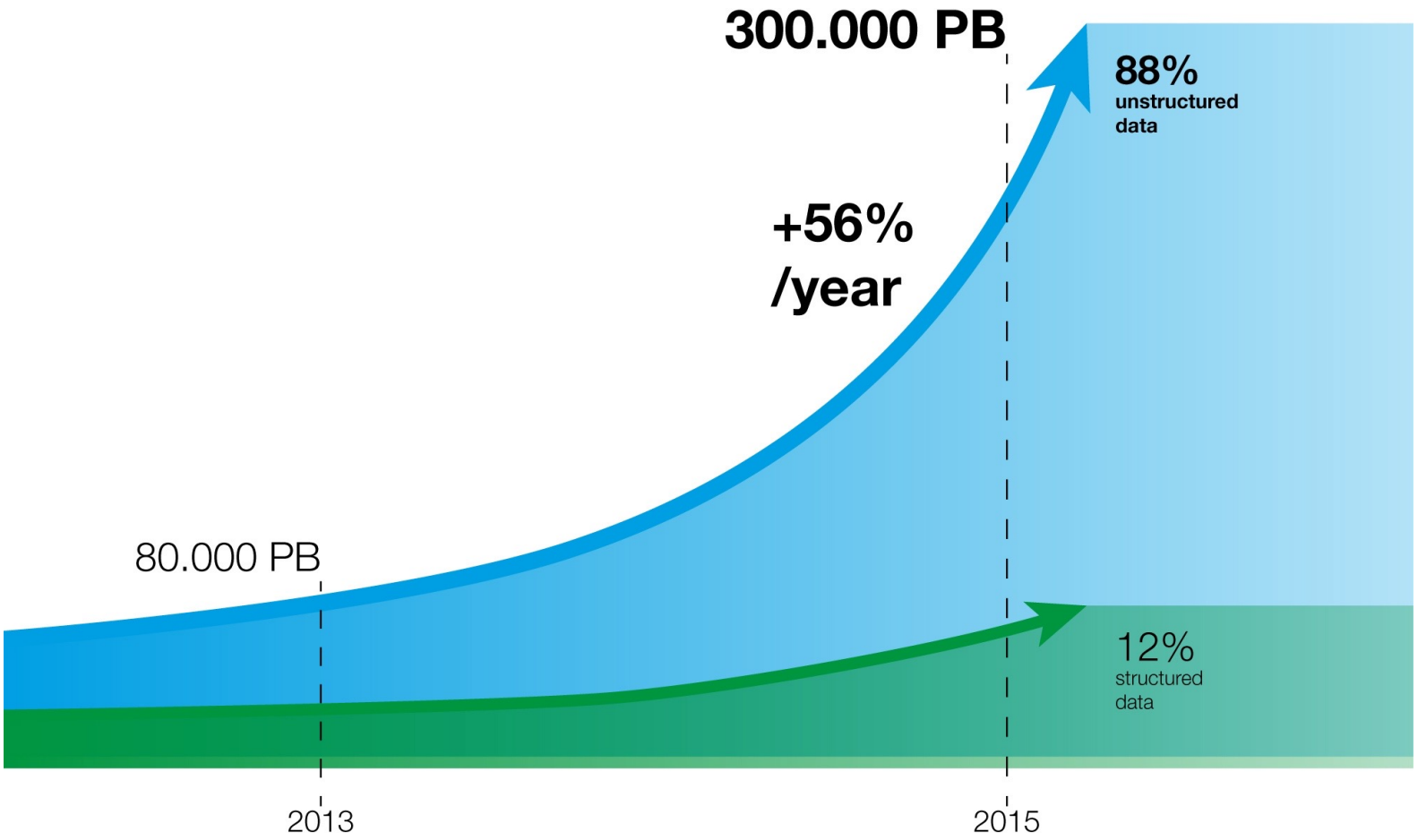
Velocity

- Clickstreams and ad impressions capture user behavior at millions of events per second
- High-frequency stock trading algorithms reflect market changes within microseconds
- Machine to machine processes exchange data between billions of devices
- Infrastructure and sensors generate massive log data in real-time
- On-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

Variety


What kind of data are we creating?
20%
80%
STRUCTURED DATA
data governed by a relational table, like in a database
UNSTRUCTURED DATA
everything else, including tweets, facebook posts, network log files, photos, word documents, email, spreadsheets, cat memes, etc.

Variety


300.000 PB
88%
unstructured
data
+56%
/year
80.000 PB
12%
structured
data
2013
2015

Variety

- Unstructured and semi-structured data types, such as text, audio, and video require additional processing to both derive meaning and the supporting metadata.
- Once understood, unstructured data has many of the same requirements as structured data, such as summarization, lineage, auditability, and privacy.
- Further complexity arises when data from a known source changes without notice.
- Frequent or real-time schema changes are an enormous burden for both transaction and analytical environments.

Structured Data

- Exemplified by data contained in relational databases and spreadsheets.
- Structured data conforms to a database model, which is largely characterized by…
  - The various fields that data belongs to (name, address, age and so forth)
  - The data type for each field (numeric, currency, alphabetic, name, date, address).
- The model also has a notion of restrictions or constraints on each field (for example, integers in a certain range)…
- And constraints between elements in the various fields that are used to enforce a notion of consistency
  - No duplicates, cannot be scheduled in two different places at the same time, etc.

Structured Data Example

Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)
Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48,0
Albania	EASTERN EUROPE	3581655	28748	124,6
Algeria	NORTHERN AFRICA	32930091	2381740	13,8
American Samoa	OCEANIA	57794	199	290,4
Andorra	WESTERN EUROPE	71201	468	152,1
Angola	SUB-SAHARAN AFRICA			
Anguilla	LATIN AMER. & CARIB			
Antigua & Barbuda	LATIN AMER. & CARIB			
Argentina	LATIN AMER. & CARIB			
Armenia	C.W. OF IND. STATES			
Aruba	LATIN AMER. & CARIB			
Australia	OCEANIA			
Austria	WESTERN EUROPE			
Azerbaijan	C.W. OF IND. STATES			
Bahamas, The	LATIN AMER. & CARIB			

Country	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)
Afghanistan	0,00	23,06	163,07
Albania	1,26	-4,93	21,52
Algeria	0,04	-0,39	31
American Samoa	58,29	-20,71	9,27
Andorra	0,00	6,6	4,05
Angola	0,13	0	191,19
Anguilla	59,80	10,76	21,03
Antigua & Barbuda	34,54	-6,15	19,46
Argentina	0,18	0,61	15,18
Armenia	0,00	-6,47	23,28
Aruba	35,49	0	5,89
Australia	0,34	3,98	4,69
Austria	0,00	2	4,66
Azerbaijan	0,00	-4,9	81,74
Bahamas, The	25,41	-2,2	25,21

Unstructured Data

- Unstructured Data (or unstructured information) refers to information that either does not have a pre-defined data model or is not organized in a predefined manner.
- Unstructured information is typically text-heavy, but may also contain data such as dates, numbers, and facts.
- Other examples include the “raw” (untagged) data representing photos and graphic images, videos, streaming sensor data, web pages, PDF files, PowerPoint presentations, emails, blog entries, wikis, and word processing documents

Unstructured Data Example

Animation

Word Doc

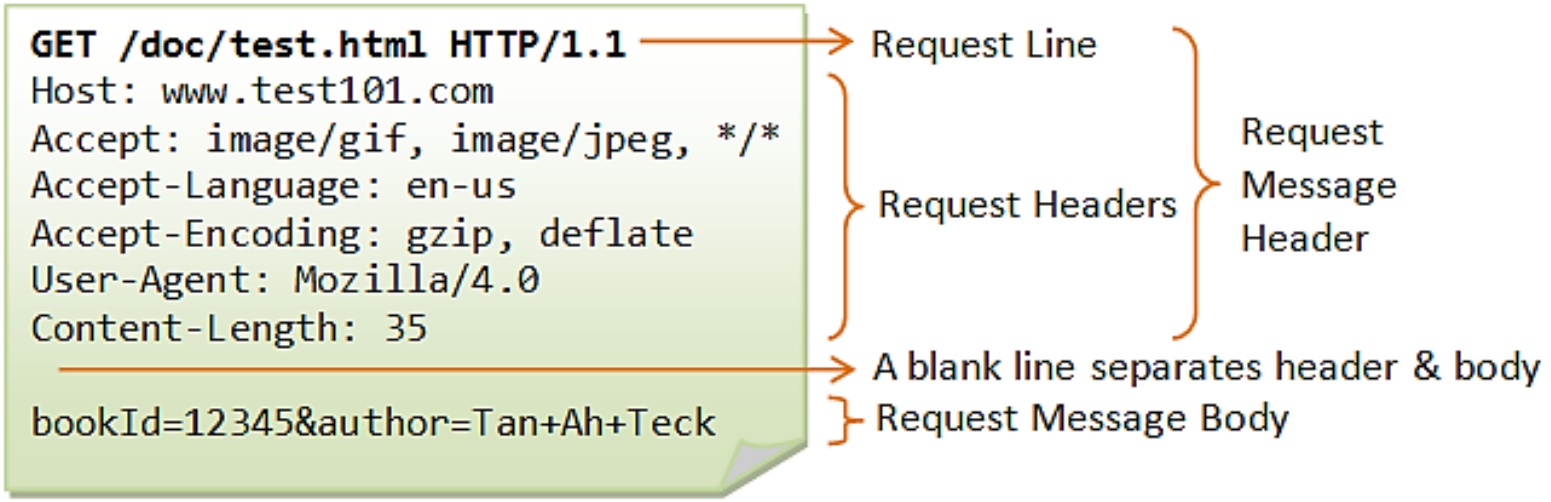
X-Ray Image

Semi-Structured Data

- Semi-structured data lies in between structured and unstructured data.
- It is a type of structured data, but lacks a strict structure imposed by an underlying data model.
- With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure from which complete semantic meaning can be easily extracted without much further processing.
- For example, word processing software now can include metadata showing the author's name and the date created, while the bulk of the document contains unstructured text.
  - Sophisticated learning algorithms would have to mine the text to understand what the text was about, because no model exists that classifies the text into neat categories
  - As an additional nuance, the text in the document may be further tagged as including table of contents, chapters, and sections.
- Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.
- Photos or other graphics can be tagged with keywords such as the creator, date, location

Semi-Structured Data Examples

HTTP Request



Medical Record

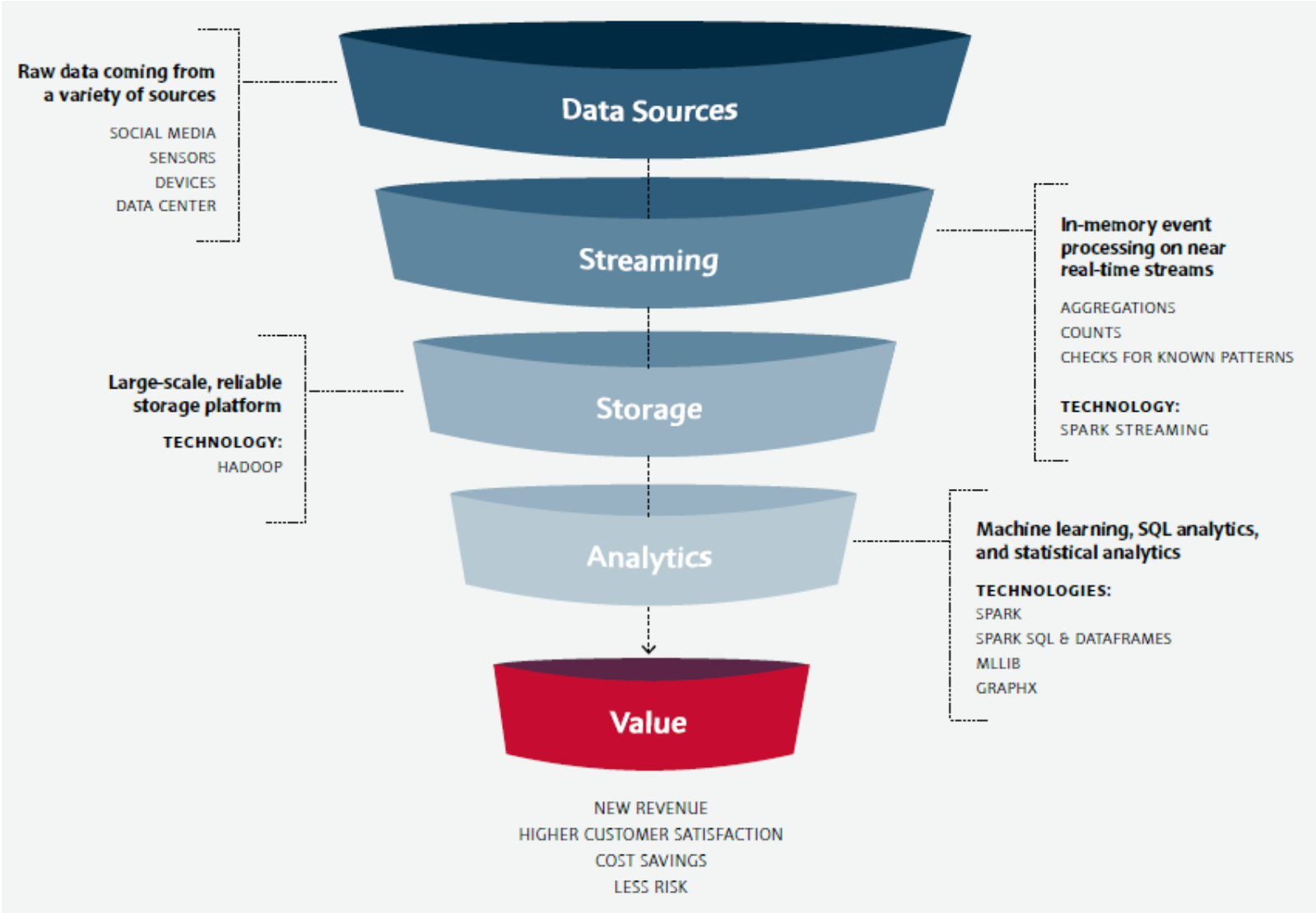
Component	Your Value
GLUCOSE	**108** mg/dL
SODIUM	**136** meq/L
POTASSIUM	**4.0** meq/L
CHLORIDE	**101** meq/L

Study Result
Impression
IMPRESSION:
1. Stable simple cystic pancreatic foci, which likely reflect side branch intraductal papillary mucinous neoplasms; follow up in 2 years is recommended per guidelines provided below.
2. Unchanged hemorrhagic/proteinaceous left renal cyst.

Other Characteristics

- Various individuals and organizations have suggested expanding the original three Vs:
- **Veracity**: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data (and consequently, the quality of the resulting analysis)
- **Variability**: Variation in the data leads to wide variation in quality.
  - Additional resources may be needed to identify, process, or filter low quality data to make it more useful.
- **Value**: The ultimate challenge of big data is delivering value.
  - Sometimes, the systems and processes in place are complex enough that using the data and extracting actual value can become difficult.

Value


Raw data coming from a variety of sources
SOCIAL MEDIA
SENSORS
DEVICES
DATA CENTER
Data Sources
Streaming
In-memory event processing on near real-time streams
AGGREGATIONS
COUNTS
CHECKS FOR KNOWN PATTERNS
TECHNOLOGY:
SPARK STREAMING
Large-scale, reliable storage platform
TECHNOLOGY:
HADOOP
Storage
Analytics
Machine learning, SQL analytics, and statistical analytics
TECHNOLOGIES:
SPARK
SPARK SQL & DATAFRAMES
MLLIB
GRAPHX
Value
NEW REVENUE
HIGHER CUSTOMER SATISFACTION
COST SAVINGS
LESS RISK

Value from Big Data in Healthcare

- **Optimize costs**
  - The McKinsey Global Institute estimates that the potential value from Big Data in healthcare could reach more than $300 billion USD a year.
- **Reducing preventable readmissions**
  - Comply with a U.S. Affordable Care Act provision that reduces payouts to Medicare Prospective Payment Systems (PPS) hospitals with excess readmissions.
  - Implementing proper discharge planning and instructions for cardiac patients could save a 350-bed hospital $486,000 USD every year.
- **Improving physician and worker performance**
  - Help improve physician performance by giving them the tools to create treatment plans based on information that validates what has
  - This supports more targeted treatments that can translate into bottom line savings.
- **Drive greater operational efficiencies**
  - Applying proven data analytics in every area of the organization leads to better decisions being made in real time that will result in more efficiently run facilities.

Value from Big Data in Healthcare

- **Improve patient care**
  - Two-thirds of executives working in federal healthcare agencies believe that Big Data will improve population health management and preventive care
- **Know patients better:** Healthcare providers can achieve a total view of a patient
  - By combining multiple data sources, such as medical device data, public health data, and socioeconomic data
- **Track patients better:** Big Data extends beyond any individual organization, which what makes it so valuable.
  - By tracking things like patient populations, current events, weather patterns, and other third-party data sources, healthcare providers can achieve greater visibility into the impact of care

Possible Definition

Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

The concept of Big Data is therefore relative to the storage and processing capability of prevalent technology of the time.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. New York, NY: McKinsey Global Institute.

Possible Definition (Famous)

Big Data is the frontier of a firm's ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers

Forrester

Possible Definition (Famous)

Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making

Gartner

Gartner, Inc. (2011, June 27). *Gartner says solving 'big data' challenge involves more than just managing volumes of data* [Press release]. Retrieved from http://www.gartner.com/ newsroom/id/1731916

Possible Definition

Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the strictures of your database architectures. To gain value from this data, you must choose an alternative way to process it

O'Reilly

Possible Definition

- “Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze
- This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data
- We don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes).
- We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase

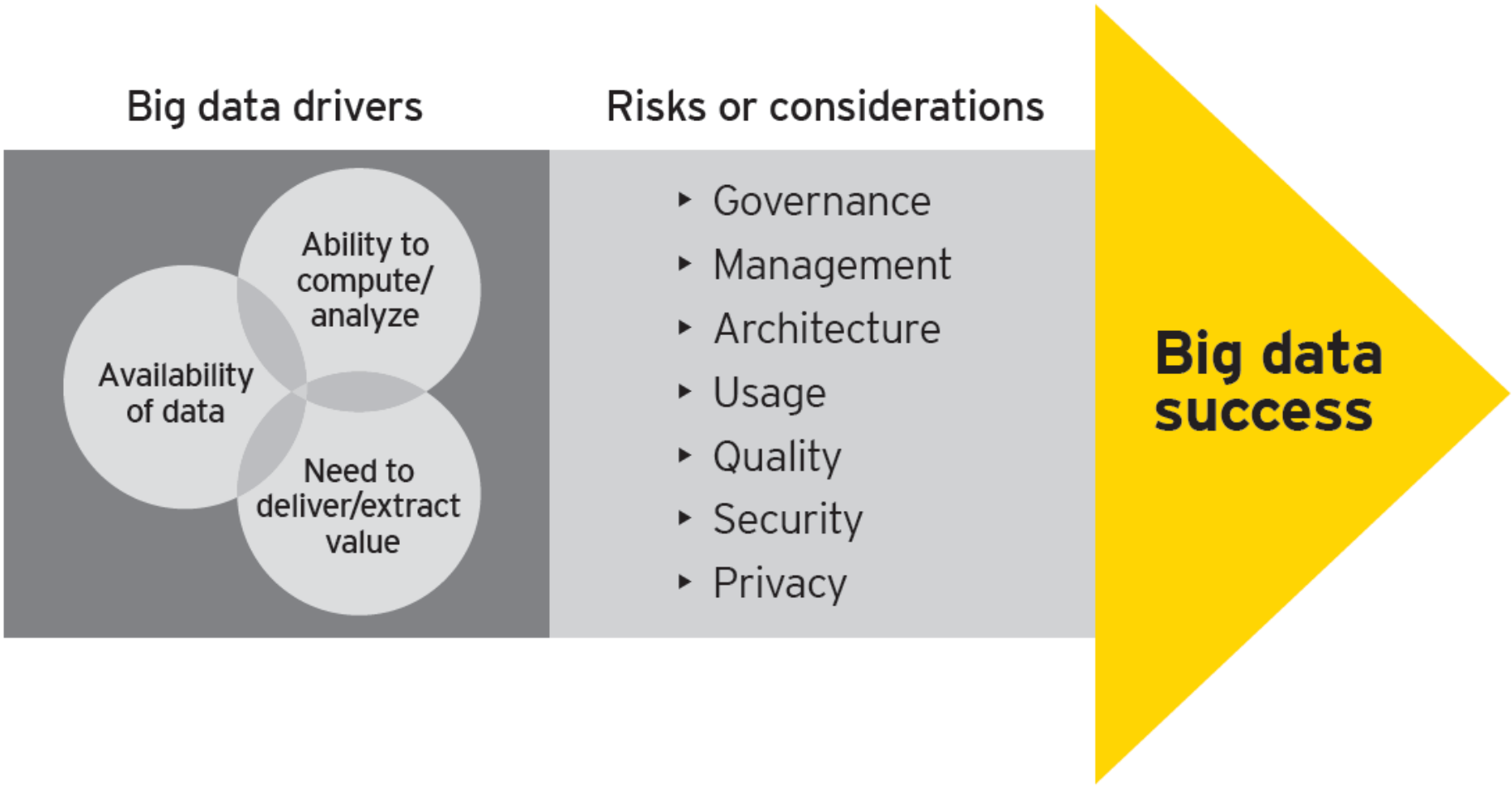
Possible Definition

- Big data is a generic term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large datasets.
- While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years.

So What Is Big Data?

- An exact definition of “big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently.
- With that in mind, generally speaking, big data is:
  - Large datasets having a range of structures
  - The category of computing strategies and technologies that are used to handle large datasets
- In this context, "large dataset" means a dataset too large to reasonably process or store with traditional tooling or on a single computer.
- The common scale of big datasets is constantly shifting and may vary significantly from organization to organization.

Big Data Drivers and Risks



Governance

- Managers will need to learn to embrace the evidence-based decision-making process.
- Organizations have to redefine their understanding of “judgements” of the outcome of big data analytics.
- Data can be of great value, but companies have to consider ownership and privacy issues before using big data results.
- In the case of medical data, it is sometimes not clear who is the owner of the data, but using the data without the right legal foundation or consent of the patient may cause big problems.
- Big data may bring about intellectual property issues, e.g. copyright and database rights infringements.
- It will be a challenge to make sure that employees are not sharing inappropriate information, or too much data outside of the organization.

Management

- Simplified access to diverse sources of data and easy-to-ingest large amounts of information may result in increasing amount of “noise” in data and decrease in the overall level of data quality.
- Many new technology market players don’t have mature enterprise-ready capabilities around implementation, support, training, etc.
- New big data methods, architecture and volume variety impose additional risks of lack of control and governance over data, and this requires additional organizational focus.
- Under the context of the complex data landscape, it is especially important to establish and maintain data lineage.
- Organizations may struggle with finding the right skills and building internal capabilities for handling big data as most technologies and methods are relatively new, and market resources are in short supply.

Architecture

- More is not always better. More data can lead to an increased number of data quality issues, and confusion and lack of consistency in business decision making
  - Especially when conflicting information is present.
- Integrated data architecture increases the challenges of data linkages and matching algorithms to distinguish items of relevance from piles of data.
- Increased complexity of architectural landscape and the growing amount of data bring new challenges around data governance and data privacy.
- Lack of capabilities, both within organizations and externally, make it hard to keep up with rapidly evolving hardware/software technology and implementation methods.

Usage

- A key challenge is to know the right business questions to ask.
- There are misunderstandings over what data is needed to make strategic or operational decisions.
- Many organizations do not have the ability to analyze data timely enough to take advantage of new insights.
- Not considering information from outside the organization (e.g., weather) that is relevant to answering bigger question is an ongoing concern.
- Organizations can get overloaded and overwhelmed by trying to handle too much data.
- The challenge of getting the right information to the right person at the right time is expanded due to the sheer size of big data.
- The costs associated with managing and monitoring the quality, credibility and integrity of big data can be prohibitive.
- There is a necessity to temper the expectation that big data will solve everything.

Quality

- A key concern is building and maintaining golden sources of data and determining which data sources should be golden.
- Understanding the data domains and the level of data quality required for each data domain creates risk.
- The need to interpret and assess unstructured data can be a challenge, and the quality of the unstructured data is often unproven.
- Structured and unstructured data may not be integrated cohesively.
- Existing information governance models will not be aligned to manage data quality for the newly acquired data.

Security

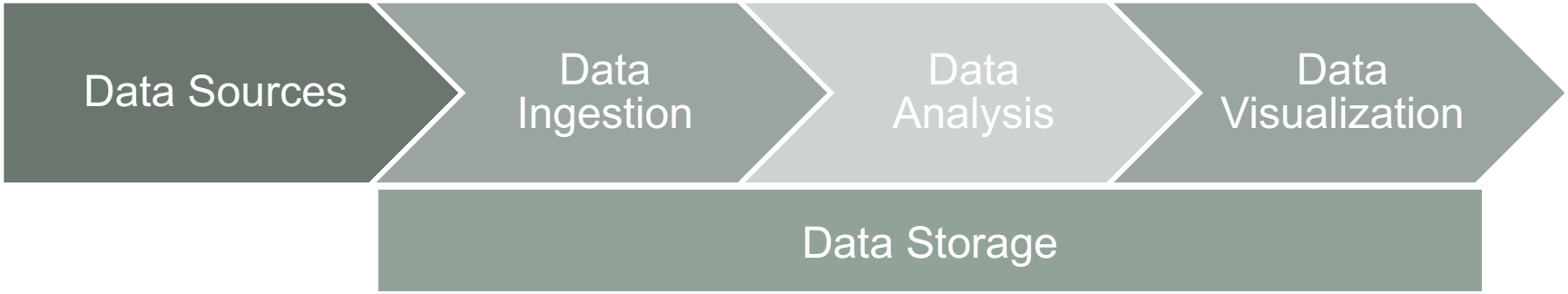
- Diverse sources of data results in distributed storage and management, compounding security vulnerabilities.
- Cloud computing puts more data in motion and causes additional security complexities.
- Lost data results in more direct impact to the end consumer. This, coupled with the behavioral data collected, can lead to more sophisticated attacks, e.g., social engineering attempts on targeted consumers.
- Increasing global regulations raise the stakes around security as the cost of dealing with data breaches continues to grow.

Privacy

- Privacy considerations around personal information have always existed, but companies have traditionally dealt with them on a limited or single location basis.
- Sensor and geolocation data may be used to identify an individual, even if no name is attached to the data, therefore increasing the need to define “personal information” that needs to be protected.
- The ability to collect new sources of information, like car-based sensor data, increases the need to evaluate opt-in/opt-out procedures by consumers.

Big Data Pipeline

- So how is data actually processed when dealing with a big data system?
- While approaches differ, there are some commonalities in the strategies and software that we can talk about generally.
- The general steps of activities involved with big data processing are:
  - Ingesting data into the system
  - Persisting the data in storage
  - Computing and Analyzing data
  - Visualizing the results



Ingesting Data into the System

- Data ingestion is the process of taking raw data and adding it to the system.
- The complexity of this operation depends heavily on the format and quality of the data sources
  - And how far the data is from the desired state prior to processing.
- One way that data can be added to a big data system are dedicated ingestion tools.
  - **Apache Sqoop** can take existing data from relational databases and add it to a big data system.
  - **Apache Flume** is designed to aggregate and import application and server logs.
  - Queuing systems like **Apache Kafka** can also be used as an interface between various data generators and a big data system

Ingesting Data into the System

- This process is sometimes called ETL, which stands for extract, transform, and load.
  - While this term conventionally refers to legacy data warehousing processes, some of the same concepts apply to data entering the big data system.
- Typical ETL operations might include
  - Modifying the incoming data to format it
  - Categorizing and labelling data
  - Filtering out unneeded or bad data, or potentially validating that it adheres to certain requirements.
- With those capabilities in mind, ideally, the captured data should be kept as raw as possible
  - For greater flexibility further on down the pipeline.

Persisting the Data in Storage

- The ingestion processes typically hands data off to components that manage storage, so that it can be reliably persisted to disk.
- While this seems like it would be a simple operation, the volume of incoming data, the requirements for availability, and the distributed computing layer make more complex storage systems necessary.
- This usually means leveraging a distributed file system for raw data storage.
  - **Apache Hadoop's HDFS** filesystem allow large quantities of data to be written across multiple nodes in the cluster.
  - This ensures that the data can be accessed by compute resources, can be loaded into the cluster's RAM for in-memory operations, and can gracefully handle component failures

Persisting the Data in Storage

- Distributed databases, especially NoSQL databases, are well-suited for this role
  - They are often designed with fault tolerant considerations
  - And can handle heterogeneous data.
- There are many different types of distributed databases to choose from
  - Depending on how you want to organize and present the data.

Computing and Analyzing Data

- Once the data is available, the system can begin processing the data to surface actual information.
- The computation layer is perhaps the most diverse part of the system
  - As the requirements and best approach can vary significantly depending on what type of insights desired.
- Data is often processed repeatedly, either iteratively by a single tool or by using a number of tools to surface different types of insights.
- **Batch processing** is one method of computing over a large dataset.
  - The process involves breaking work up into smaller pieces, scheduling each piece on an individual machine, reshuffling the data based on the intermediate results, and then calculating and assembling the final result.
  - his is the strategy used by **Apache Hadoop's MapReduce**.
- Batch processing is most useful when dealing with very large datasets that require quite a bit of computation

Computing and Analyzing Data

- While batch processing is a good fit for certain types of data and computation, other workloads require more **real-time processing**.
  - Real-time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available.
  - One way of achieving this is **stream processing**, which operates on a continuous stream of data composed of individual items.
  - Another common characteristic of real-time processors is in-memory computing
    - Which works with representations of the data in the cluster's memory to avoid having to write back to disk.

Computing and Analyzing Data

- **Apache Storm** and **Apache Spark** provide different ways of achieving real-time or near real-time processing.
- There are trade-offs with each of these technologies, which can affect which approach is best for any individual problem.
- In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly.

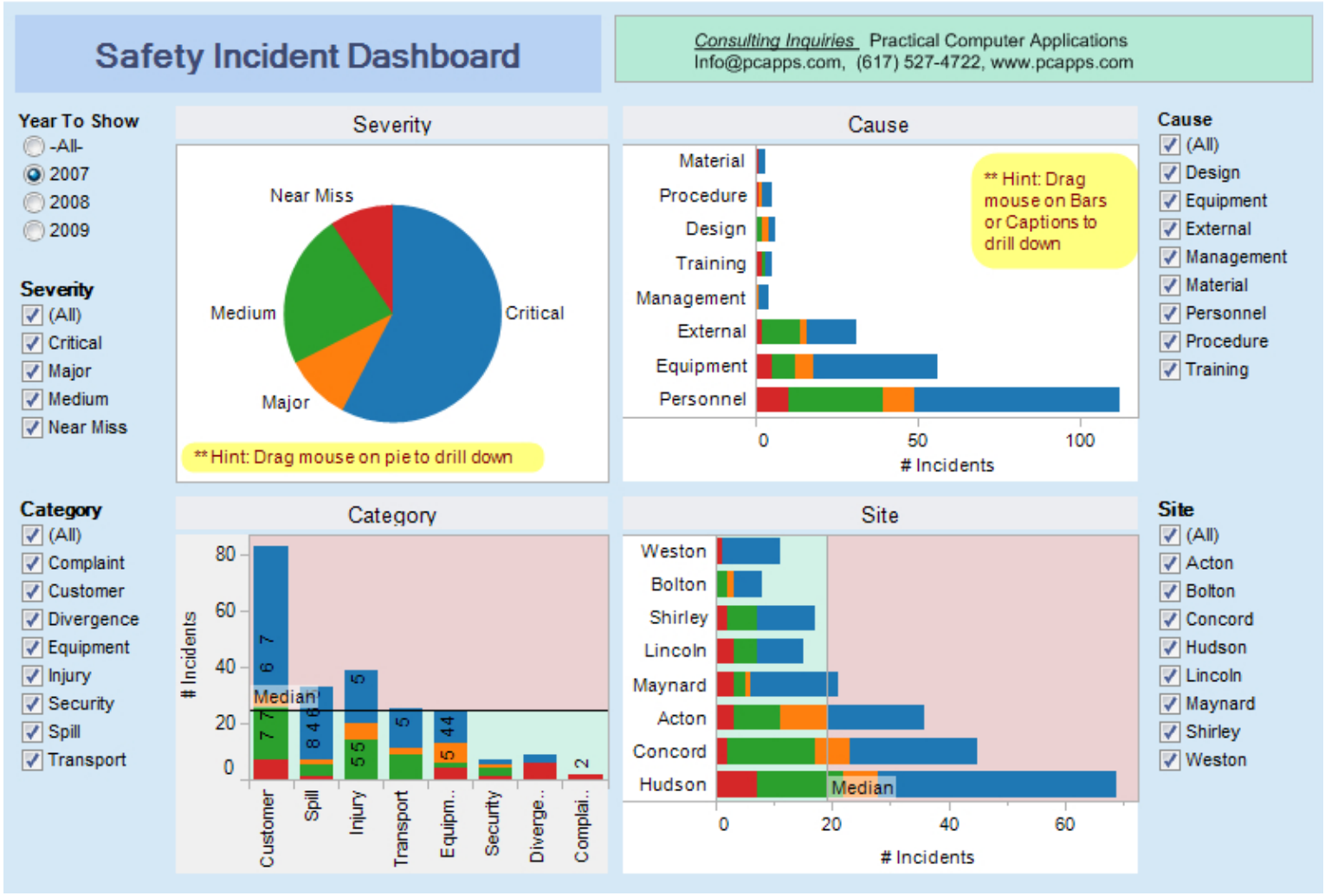
Computing and Analyzing Data

- These tools frequently plug into the above frameworks and provide additional interfaces for interacting with the underlying layers.
  - For instance, **Apache Hive** provides a data warehouse interface for Hadoop
  - **Apache Pig** provides a high level querying interface
  - SQL-like interactions with data can be achieved with projects like **Apache Impala**, **Apache Spark SQL**, and **Presto**.
- For machine learning, projects like **Apache Mahout**, and **Apache Spark's MLlib** can be useful.
- For straight analytics programming that has wide support in the big data ecosystem, both **R** and **Python** are popular choices.

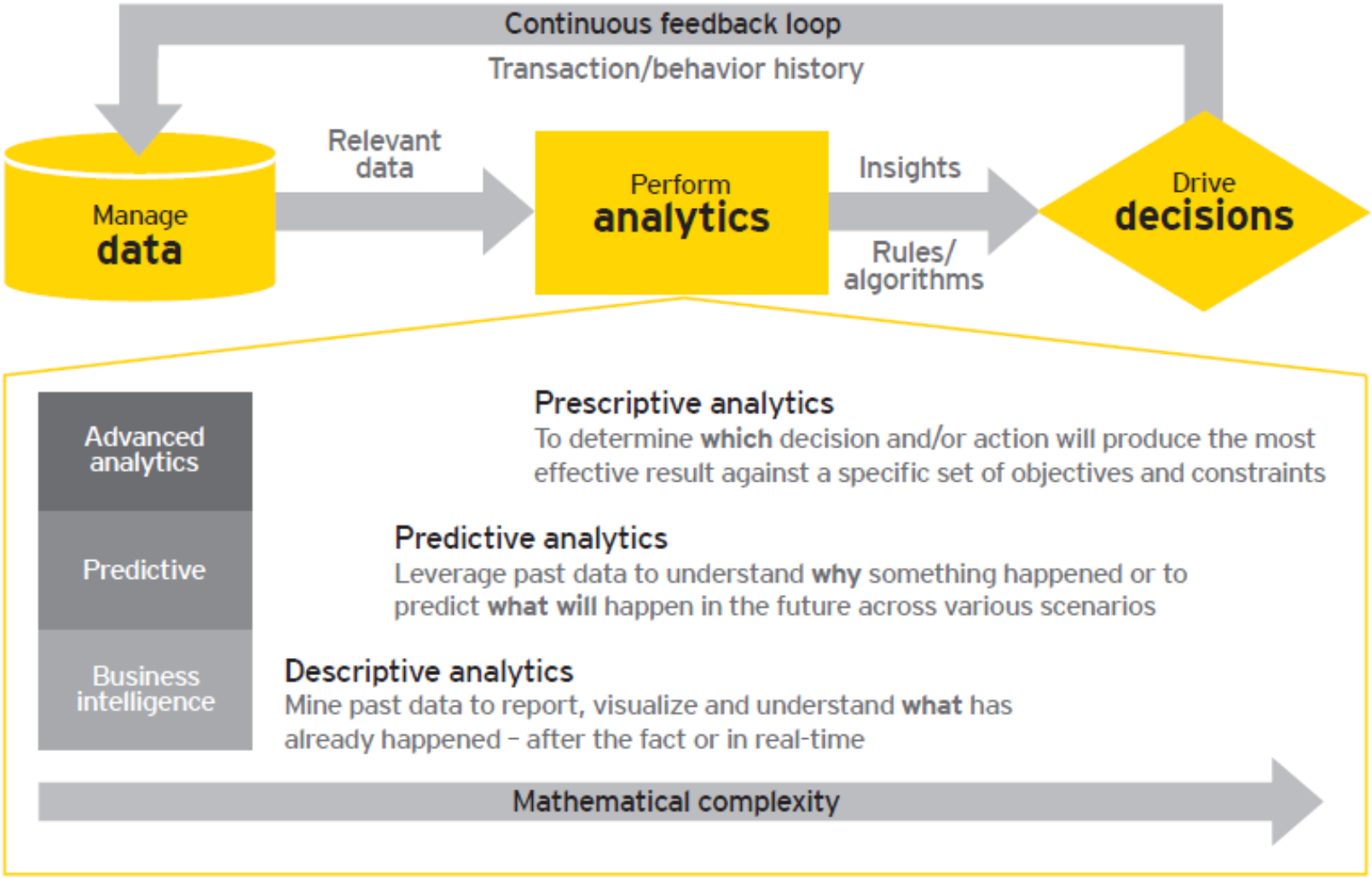
Visualizing the Results

- Due to the type of information being processed in big data systems, recognizing trends or changes in data over time is often more important than the values themselves.
- Visualizing data is one of the most useful ways to spot trends and make sense of a large number of data points.
- One visualization technology typically used for interactive data science work is a data "notebook".
  - Popular examples of this type of visualization interface are **Apache Zeppelin and Project Jupyter**
- Another approach is to support dynamic creation of ad-hoc visualizations
  - Some (commercial) examples include **Tableau** and **Spotfire**

Visualizing the Results



Another Data Pipeline Perspective



Conclusion

- Big data is a broad, rapidly evolving topic.
- While it is not well-suited for all types of computing, many organizations are turning to big data for certain types of work loads and using it to supplement their existing analysis and business tools.
- Big data systems are uniquely suited for surfacing difficult-to-detect patterns and providing insight into behaviors that are impossible to find through conventional means.
- By correctly implement systems that deal with big data, organizations can gain incredible value from data that is already available.

Big Data Glossary

- **Big data**: Big data is an umbrella 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 of data.
- **Batch processing**: Batch processing is a computing strategy that involves processing data in large sets. This is typically ideal for non-time sensitive work that operates on very large sets of data. The process is started and at a later time, the results are returned by the system.
- **Cluster computing**: Clustered computing is the practice of pooling the resources of multiple machines and managing their collective capabilities to complete tasks. Computer clusters require a cluster management layer which handles communication between the individual nodes and coordinates work assignment.

Big Data Glossary

- **Data lake**: Data lake is a term for a large repository of collected data in a relatively raw state. This is frequently used to refer to the data collected in a big data system which might be unstructured and frequently changing. This differs in spirit to data warehouses (defined below).
- **Data mining**: Data mining is a broad term for the practice of trying to find patterns in large sets of data. It is the process of trying to refine a mass of data into a more understandable and cohesive set of information.
- **Data warehouse**: Data warehouses are large, ordered repositories of data that can be used for analysis and reporting. In contrast to a *data lake*, a data warehouse is composed of data that has been cleaned, integrated with other sources, and is generally well-ordered. Data warehouses are often spoken about in relation to big data, but typically are components of more conventional systems.

Big Data Glossary

- **ETL**: ETL stands for extract, transform, and load. It refers to the process of taking raw data and preparing it for the system's use. This is traditionally a process associated with data warehouses, but characteristics of this process are also found in the ingestion pipelines of big data systems.
- **Hadoop**: Hadoop is an Apache project that was the early open-source success in big data. It consists of a distributed filesystem called HDFS, with a cluster management and resource scheduler on top called YARN (Yet Another Resource Negotiator). Batch processing capabilities are provided by the MapReduce computation engine. Other computational and analysis systems can be run alongside MapReduce in modern Hadoop deployments.
- **In-memory computing**: In-memory computing is a strategy that involves moving the working datasets entirely within a cluster's collective memory. Intermediate calculations are not written to disk and are instead held in memory. This gives in-memory computing systems like Apache Spark a huge advantage in speed over I/O bound systems like Hadoop's MapReduce.

Big Data Glossary

- **Machine learning**: Machine learning is the study and practice of designing systems that can learn, adjust, and improve based on the data fed to them. This typically involves implementation of predictive and statistical algorithms that can continually zero in on "correct" behavior and insights as more data flows through the system.
- **Map reduce (big data algorithm)**: Map reduce (the big data algorithm, not Hadoop's MapReduce computation engine) is an algorithm for scheduling work on a computing cluster. The process involves splitting the problem set up (mapping it to different nodes) and computing over them to produce intermediate results, shuffling the results to align like sets, and then reducing the results by outputting a single value for each set.
- **NoSQL**: NoSQL is a broad term referring to databases designed outside of the traditional relational model. NoSQL databases have different trade-offs compared to relational databases, but are often well-suited for big data systems due to their flexibility and frequent distributed-first architecture.
- **Stream processing**: Stream processing is the practice of computing over individual data items as they move through a system. This allows for real-time analysis of the data being fed to the system and is useful for time-sensitive operations using high velocity metrics.