TMIP

Better Methods. Better Outcomes.

Webinar Series

TMIP VISION

TMIP provides technical support and promotes knowledge and information exchange in the transportation planning and modeling community.

U.S. Department of Transportation
Federal Highway Administration

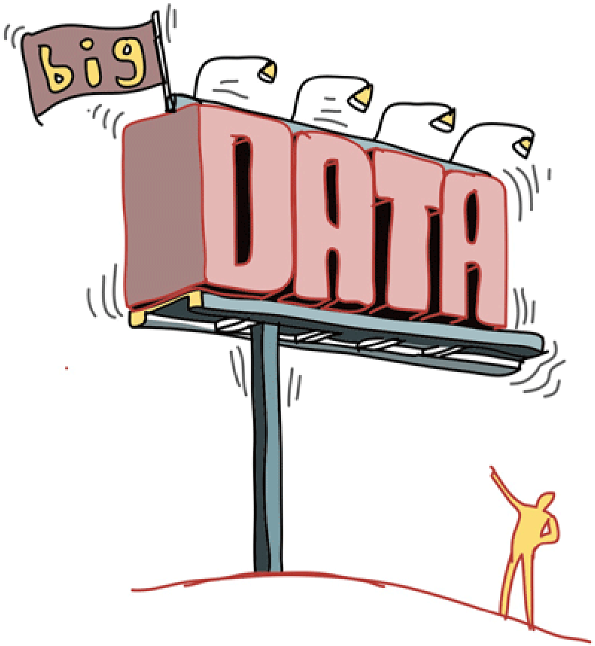
DISCLAIMER

The views and opinions expressed during this webinar are those of the presenters and do not represent the official policy or position of FHWA and do not constitute an endorsement, recommendation or specification by FHWA. The webinar is based solely on the professional opinions and experience of the presenters and is made available for information and experience sharing purposes only.

WHAT IS BIG DATA?

Big Data is

- Talked about everywhere
- Surprisingly amorphous
- Overhyped
- A very real imperative



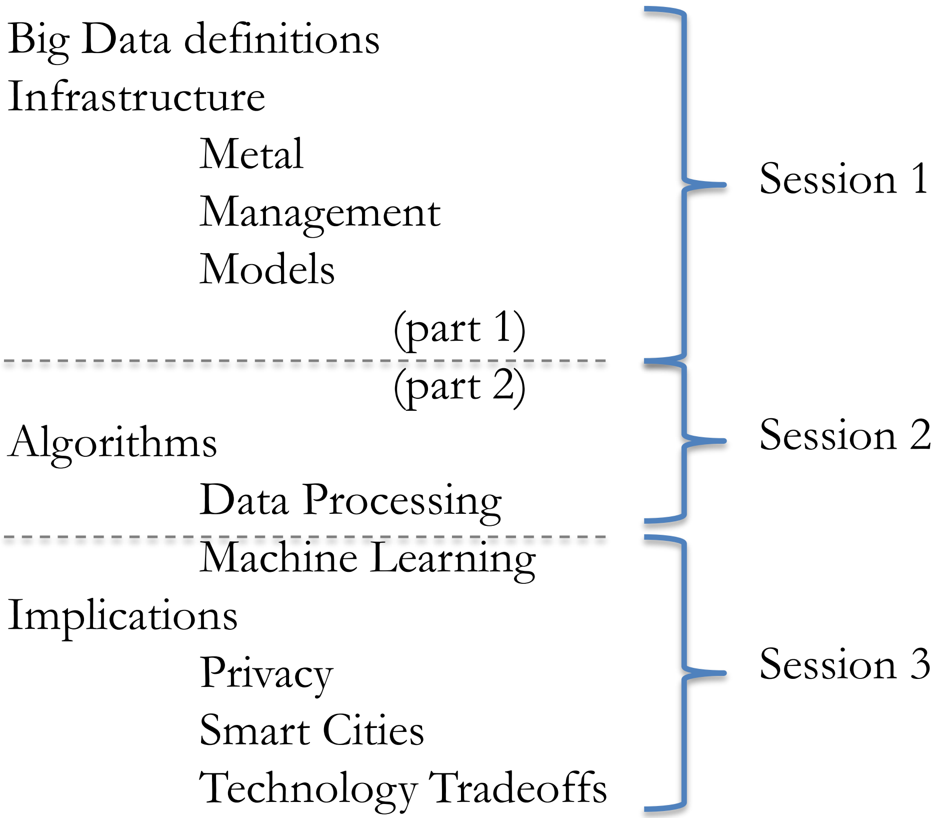
Poll Question Follow Up: Bridge Deck Inspection

BIG DATA AND TRANSPORTATION

Webinar Goals

- Background Material
  - A Historical Perspective
  - Some Definitions
  - Where Does The Data Come From
- An Approach To Organizing Big Data
  - Infrastructure
  - Algorithms
  - Implications

Webinar Roadmap



Big Data definitions

Infrastructure

- Metal
- Management
- Models
  - (part 1)

Session 1

- (part 2)

Algorithms

- Data Processing

Session 2

- Machine Learning

Implications

- Privacy
- Smart Cities
- Technology Tradeoffs

Session 3

Webinar Roadmap

Big Data definitions

Infrastructure

- Metal
- Management
- Models

(part 1)

Session 1

(part 2)

Algorithms

- Data Processing

- Machine Learning

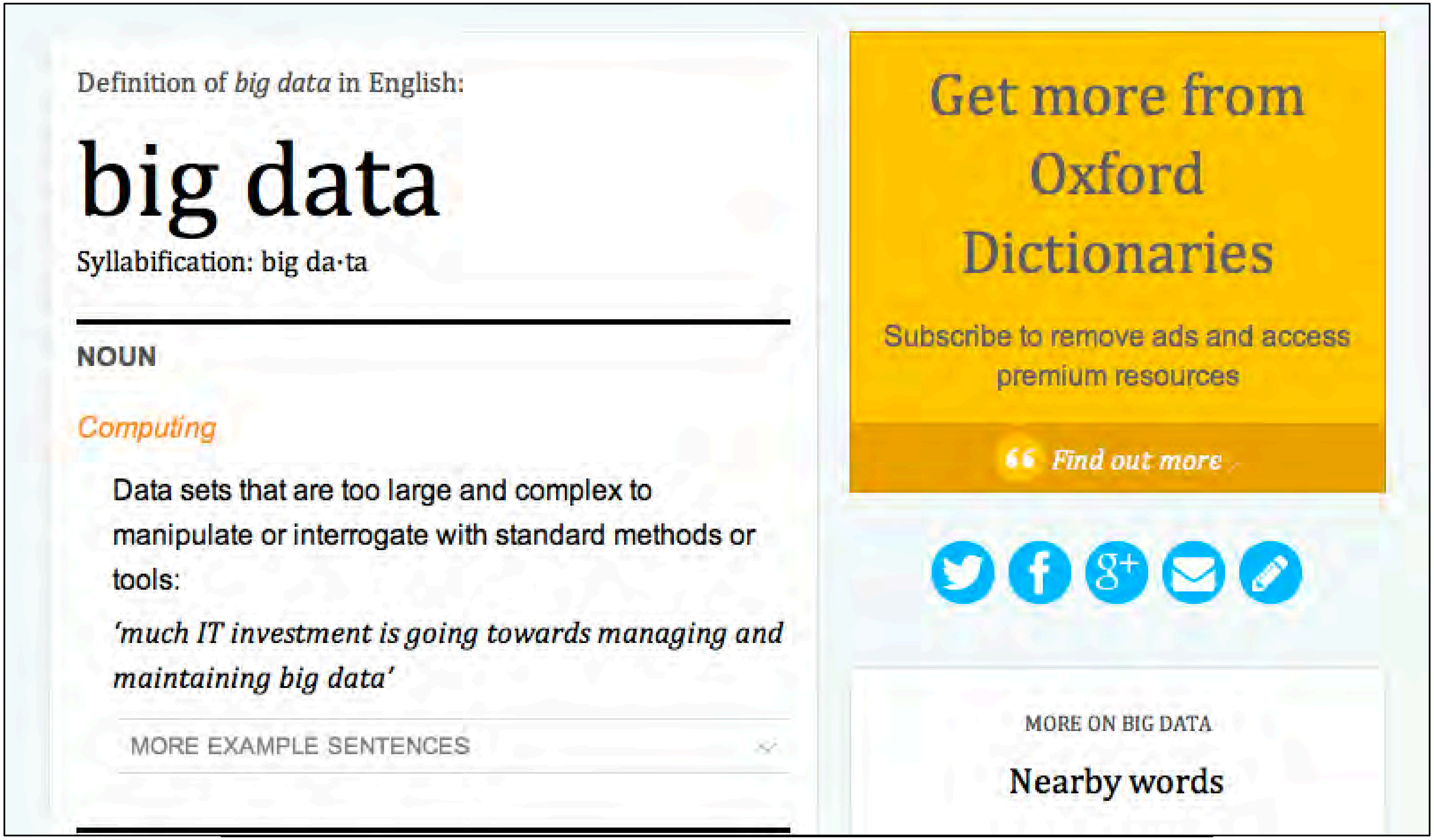
Implications

- Privacy
- Smart Cities
- Technology Tradeoffs

Background

WHAT IS BIG DATA?

According to the Dictionary



Definition of *big data* in English:

big data

Syllabification: big da·ta

NOUN

Computing

Data sets that are too large and complex to manipulate or interrogate with standard methods or tools:

'much IT investment is going towards managing and maintaining big data'

MORE EXAMPLE SENTENCES

MORE ON BIG DATA

Nearby words

The Etymology of Big Data

John R. Mashey was the chief scientist at Silicon Graphics

He had regularly used the term "Big Data" In his talks in mid and late 1990s

Before Big Data There Was Public Policy

U.S. Constitution – Article 1, Section 2, Paragraph 3

Representatives and direct Taxes shall be apportioned among the several States which may be included within this Union, according to their respective Numbers, which shall be determined by adding to the whole Number of free Persons, including those bound to Service for a Term of Years, and excluding Indians not taxed, three fifths of all other Persons. The actual Enumeration shall be made within three Years after the first Meeting

Key Policy: A U.S. Census is Required Every 10 Years

The actual Enumeration shall be made within 3 Years after the first Meeting of the Congress of the United States, and within every subsequent Term of 10 Years, in such Manner as they shall by Law direct.

Before Big Data: The 1880 Census

The 1880 Census Took 9 Years To Complete

Year	Census Population
1850	23.191 Million
1860	31.442 Million
1870	38.558 Million
1880	50.198 Million
1890	62.979 Million
1900	76.211 Million

How long would the 1890 Census take?

Tabulator Machines and the 1890 Census

Hermann Hollerith (1860 – 1929)

- Attended Columbia University School of Mines
- Invented a punch card system
  - Based on idea from Dr. John Shaw Billings
- Formed the Tabulating Machine Company
- Won Census Bureau contest
  - Unofficial census count in 2 months!
  - Paid $750,000 for rent of his machines

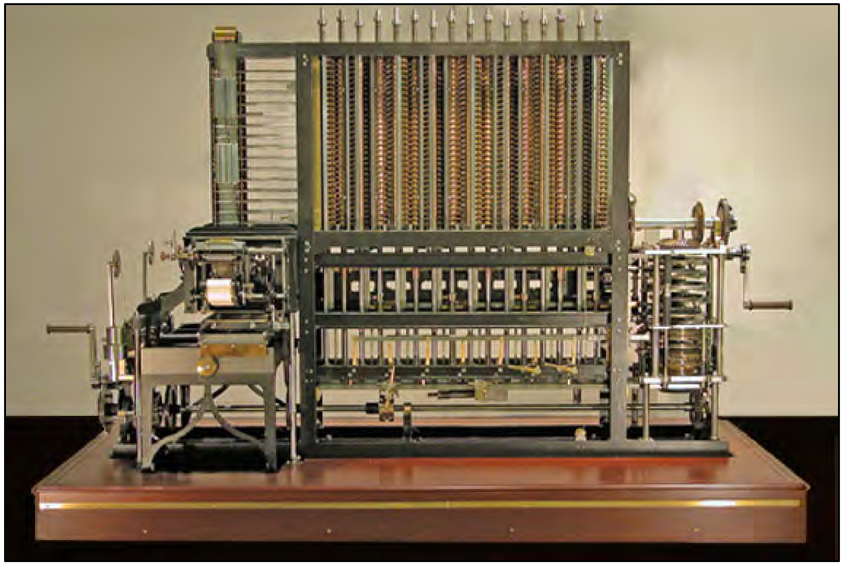
1924

- Tabulating Machine Company evolved into IBM

What is Big Data?

From a Quantitative Perspective

- According to *John Rauser (Pinterest, Amazon)*
    - Data is big data when you can't process it on one machine



The Charles Babbage Difference Engine (designed in 1849)

What is Big Data?

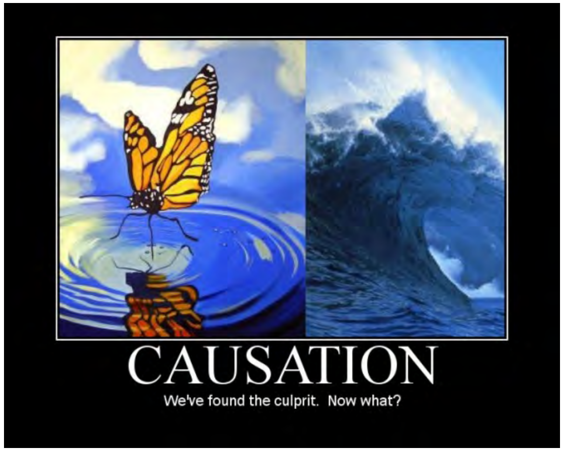
From a Quantitative Perspective

- According to *Mark Whitehorn (TheRegister.co.uk)*
  - Any data that doesn't fit well into tables and that generally responds poorly to manipulation by SQL

What is Big Data?

From an Analysis Perspective

- According to *Cukier (The Economist) & Mayer-Schoenberger (Oxford)*
  - Analytical Shift: N=Small → N=ALL
  - Analytical Shift: Causation → Correlation

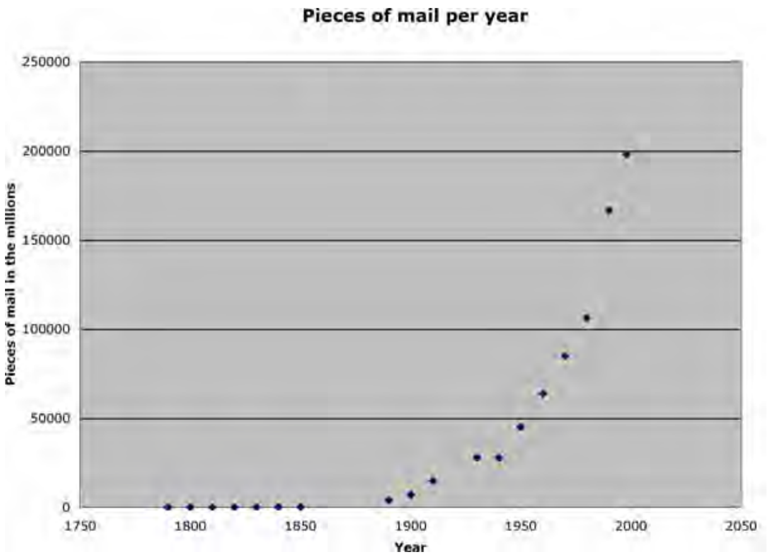


What is Big Data?

From a Data Growth Perspective

- According to the *U.S. Chamber of Congress Foundation*
  - 90% of today's data was created in the last 2 years

Growth of Postal Mail

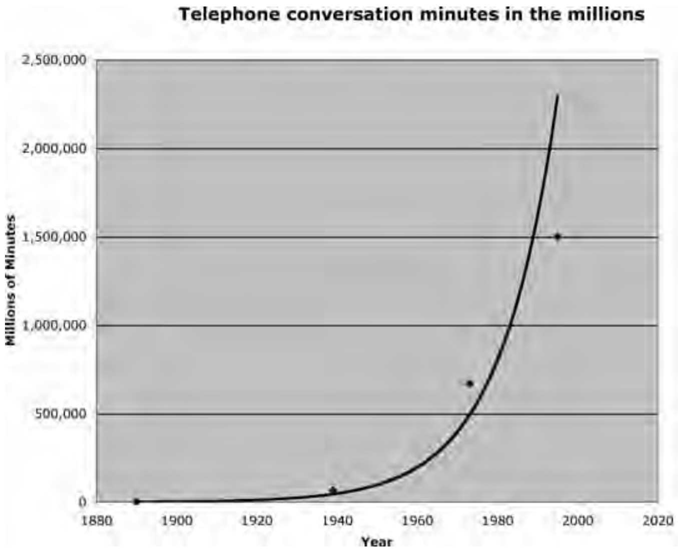


What is Big Data?

From a Data Growth Perspective

- According to the *U.S. Chamber of Congress Foundation*
  - 90% of today's data was created in the last 2 years

Growth of Telephone Minutes

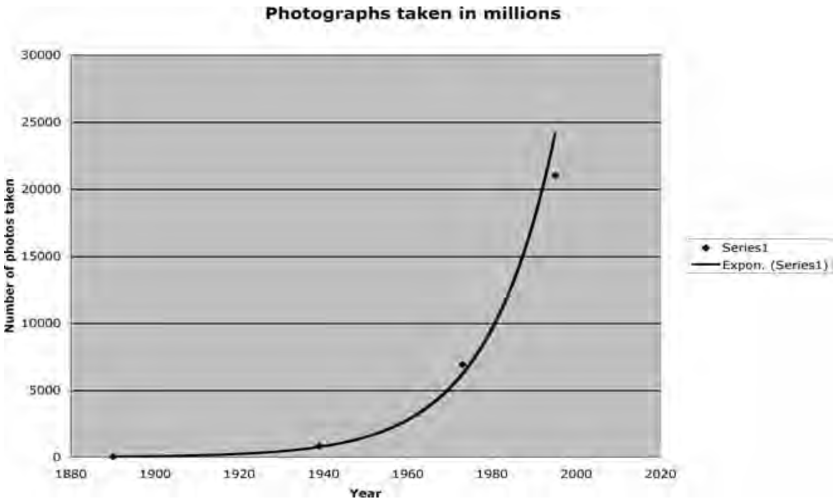


What is Big Data?

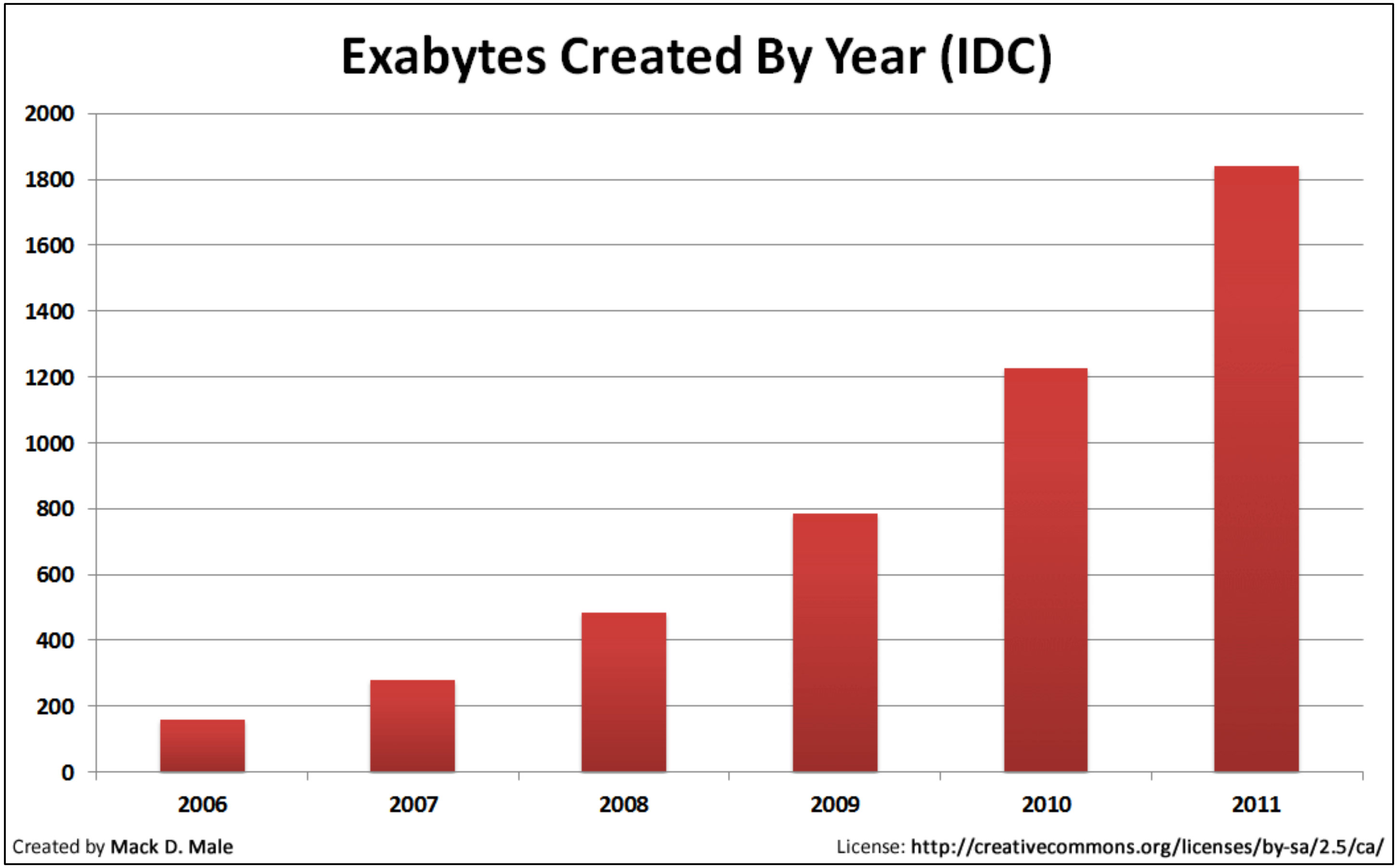
From a Data Growth Perspective

- According to the *U.S. Chamber of Congress Foundation*
  - 90% of today's data was created in the last 2 years

Growth of Photos Taken



What is Big Data?



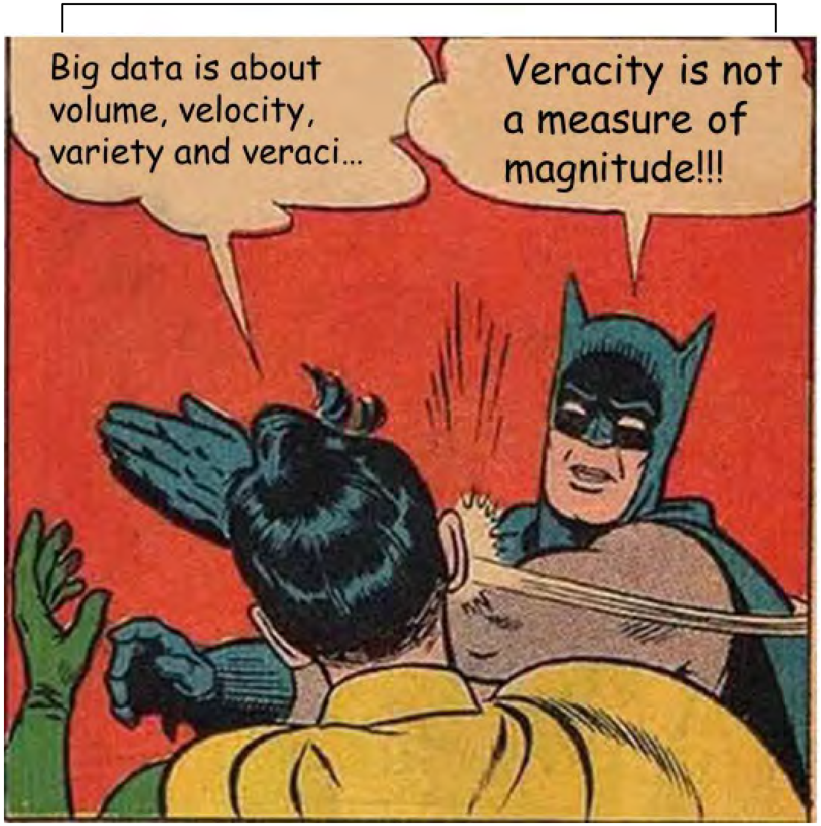
What is Big Data?

Three V's Perspective

- Volume
- Variety
- Velocity

Other V's

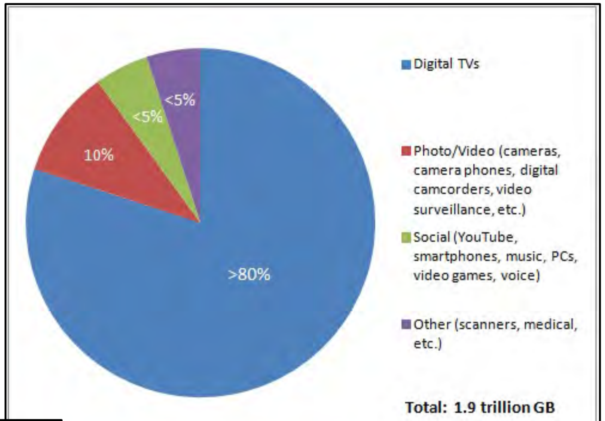
- Veracity
- Value
- Viability

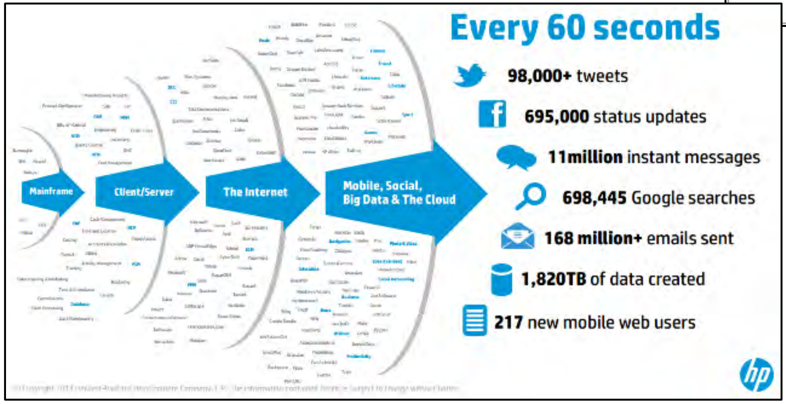

Big data is about volume, velocity, variety and veraci...
Veracity is not a measure of magnitude!!!

Where Does The Data Come From?

Explicitly From People

- 100 hours of video uploaded to YouTube per minute
- 11 billion person-hours of DVR recording in 2013

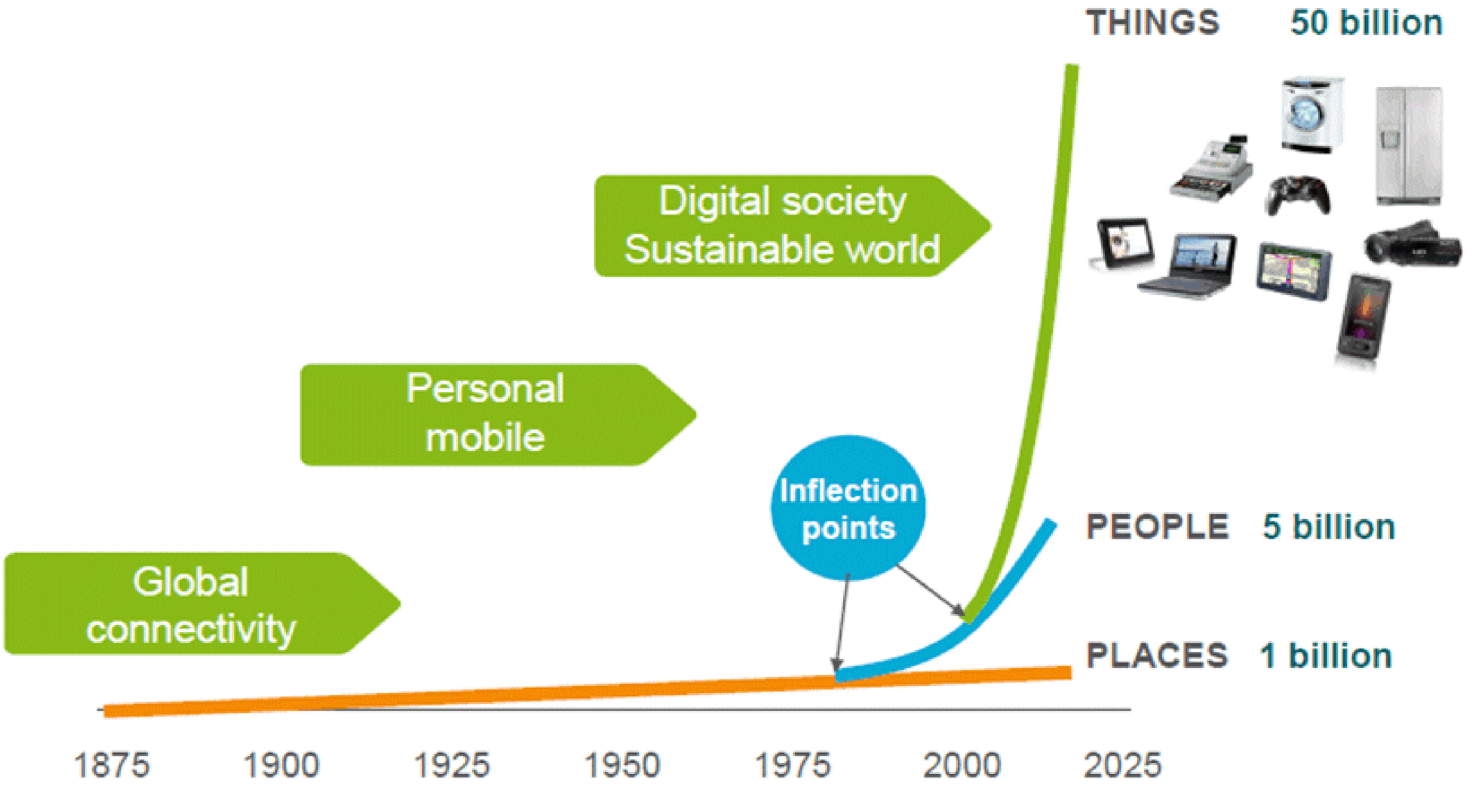




Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820TB of data created
- 217 new mobile web users

People Won’t Dominate Data for Long



Source: Ericsson AB, "Infrastructure Innovation - Can the Challenge be met?," Sept 2010

Where Does The Data Come From?

From Planes

- A Boeing jet generates 10 TB of data every 30 minutes
- A single NY → LA flight (6 hours)
  - 240 Terabytes of data
- There are 28,537 commercial flights in the US everyday
- About 6 Exabytes per day!!!

Where Does The Data Come From?

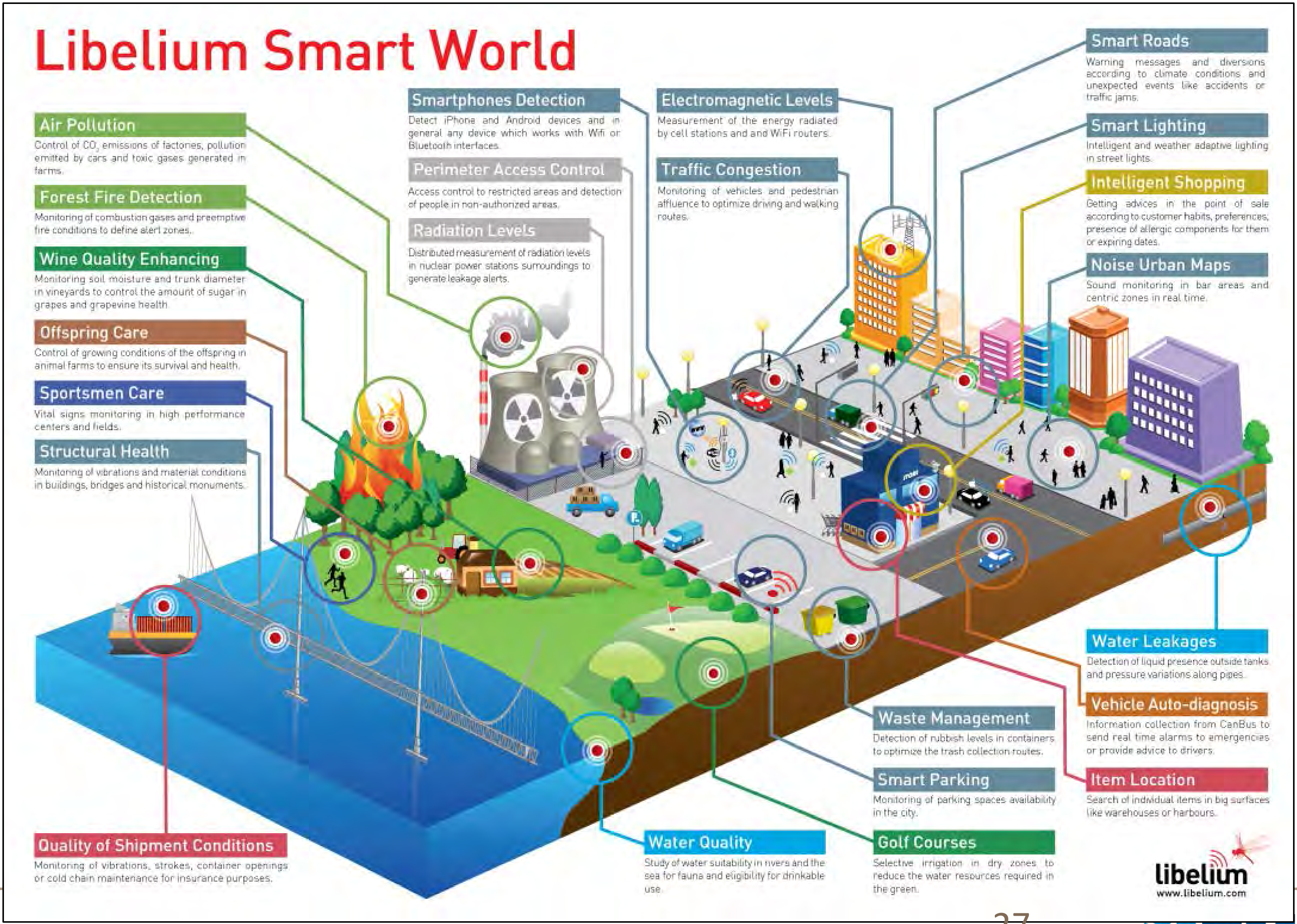
From Fast Cars

- A Formula 1 car has 200 sensors (compared to 20 sensors for a mid-level sedan)
- 30 TB of telemetry data per race

Where Does The Data Come From?

From (not so) Fast Cars

- Google's autonomous car generates 1 GB of data per second; close to an exabyte of data per year per car
  - Most of this is useless data…*currently*



Libelium Smart World

Air Pollution
Control of CO_2 emissions of factories, pollution emitted by cars and toxic gases generated in farms.

Forest Fire Detection
Monitoring of combustion gases and preemptive fire conditions to define alert zones.

Wine Quality Enhancing
Monitoring soil moisture and trunk diameter in vineyards to control the amount of sugar in grapes and grapevine health.

Offspring Care
Control of growing conditions of the offspring in animal farms to ensure its survival and health.

Sportsmen Care
Vital signs monitoring in high performance centers and fields.

Structural Health
Monitoring of vibrations and material conditions in buildings, bridges and historical monuments.

Smartphones Detection
Detect iPhone and Android devices and in general any device which works with Wifi or Bluetooth interfaces.

Perimeter Access Control
Access control to restricted areas and detection of people in non-authorized areas.

Radiation Levels
Distributed measurement of radiation levels in nuclear power stations surroundings to generate leakage alerts.

Electromagnetic Levels
Measurement of the energy radiated by cell stations and and WiFi routers.

Traffic Congestion
Monitoring of vehicles and pedestrian affluence to optimize driving and walking routes.

Smart Roads
Warning messages and diversions according to climate conditions and unexpected events like accidents or traffic jams.

Smart Lighting
Intelligent and weather adaptive lighting in street lights.

Intelligent Shopping
Getting advices in the point of sale according to customer habits, preferences, presence of allergic components for them or expiring dates.

Noise Urban Maps
Sound monitoring in bar areas and centric zones in real time.

Water Leakages
Detection of liquid presence outside tanks and pressure variations along pipes.

Vehicle Auto-diagnosis
Information collection from CanBus to send real time alarms to emergencies or provide advice to drivers.

Item Location
Search of individual items in big surfaces like warehouses or harbours.

Waste Management
Detection of rubbish levels in containers to optimize the trash collection routes.

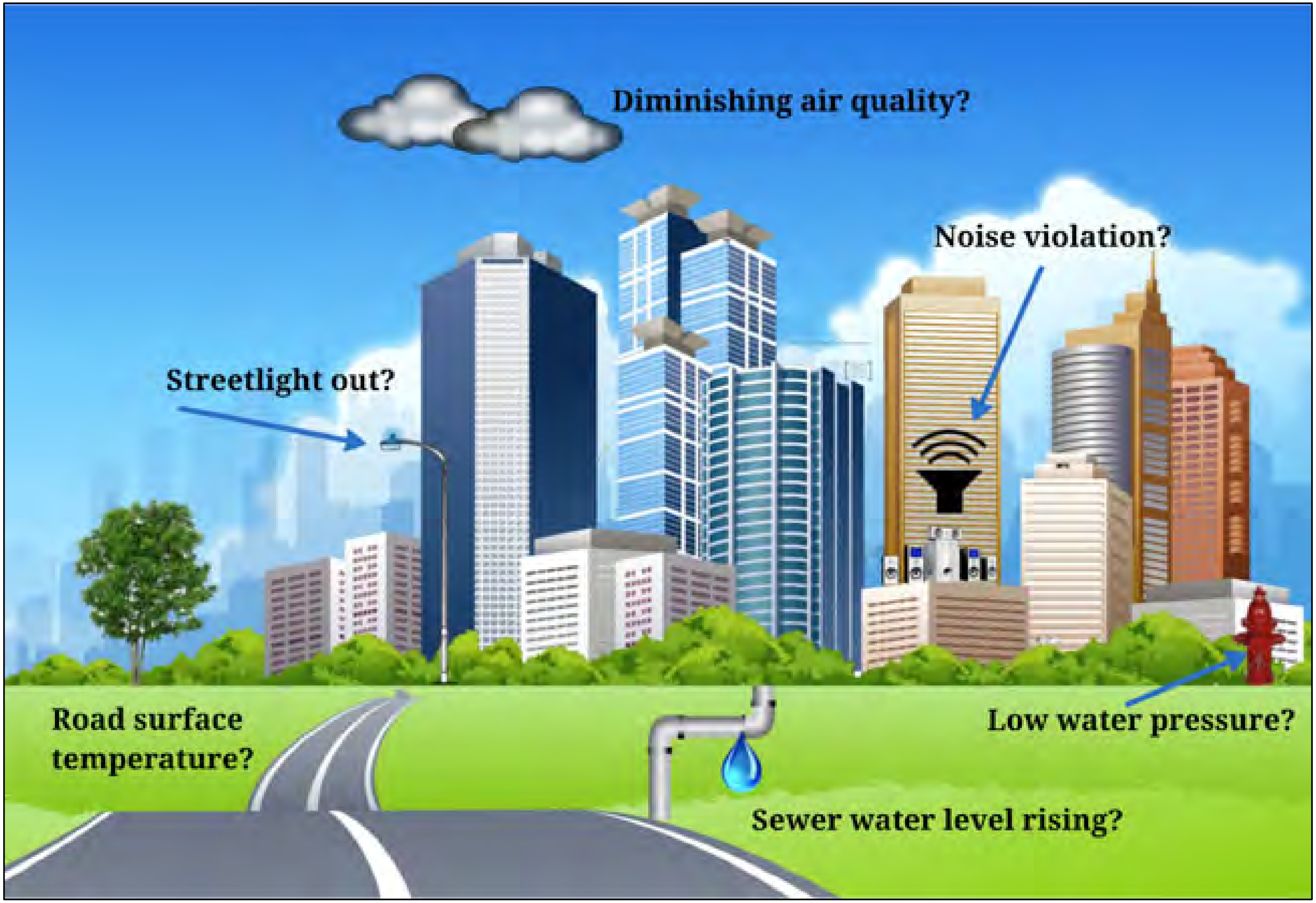
Smart Parking
Monitoring of parking spaces availability in the city.

Golf Courses
Selective irrigation in dry zones to reduce the water resources required in the green.

Water Quality
Study of water suitability in rivers and the sea for fauna and eligibility for drinkable use.

Quality of Shipment Conditions
Monitoring of vibrations, strokes, container openings or cold chain maintenance for insurance purposes.

libelium
www.libelium.com


Diminishing air quality?
Noise violation?
Streetlight out?
Road surface temperature?
Low water pressure?
Sewer water level rising?

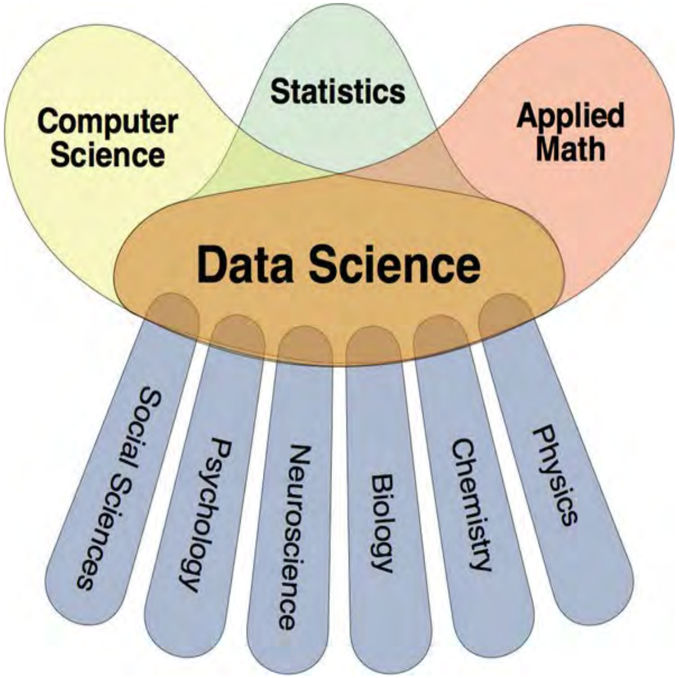
Data Science Skills

"…the skills of a "data scientist" are those of a modern statistician."

- Cosma Shalizi
  CMU Statistics Professor

From Cathy O'Neil:

- When hiring a data scientist, look for
  - Data grappling skills
  - Data visualization experience
  - Knowledge of stats
  - Experience with forecasting and prediction
  - Great communication skill



Big Data Success Stories

Retailers – Kroger, Target

- What: Personalized direct marketing
- Data: Loyalty card info, wifi location data
  - Some collected, some bought (e.g., Acxiom)

11 million direct mail flyers per quarter

- Each flyer contains 12 personalized coupons

Big Data Success Stories

Retailers – Kroger, Target

- What: Personalized direct marketing
- Data: Loyalty card info, wifi location data
  - Some collected, some bought (e.g., Acxiom)

Notional example

- Female shopper, 23 years old
- Bought cocoa-butter lotion, big purse, blue rug
- 87% chance she is pregnant

Big Data Success Stories

Healthcare – Google Flu

- What: Predicted flu outbreaks
- Data: User queries
- Google prediction: 11% with influenza

google.org Flu Trends

Hype?

- Actual results: 6% with influenza

What is Big Data?

From a Technology Perspective

- *First, it is a bundle of* ***technologies****. Second, it is a potential* ***revolution in measurement****. And third, it is a* ***point of view****, or philosophy, about how* ***decisions*** *will be—and perhaps should be—made in the future.*

— Steve Lohr, *The New York Times*

Organizing Big Data

- Infrastructure
  - Big Data is a technology-based revolution;
    technology enables the generation and processing of data
- Algorithms
  - Big Data is a revolution in measurement;
    new algorithms efficiently extend measurement capabilities
- Implications
  - Big Data is a new approach to decision-making;
    the determination and execution of these decision have profound consequences

Webinar Roadmap

Big Data definitions
Infrastructure
Metal
Management
Models
(part 1)

(part 2)
Algorithms
Data Processing

Machine Learning
Implications
Privacy
Smart Cities
Technology Tradeoffs

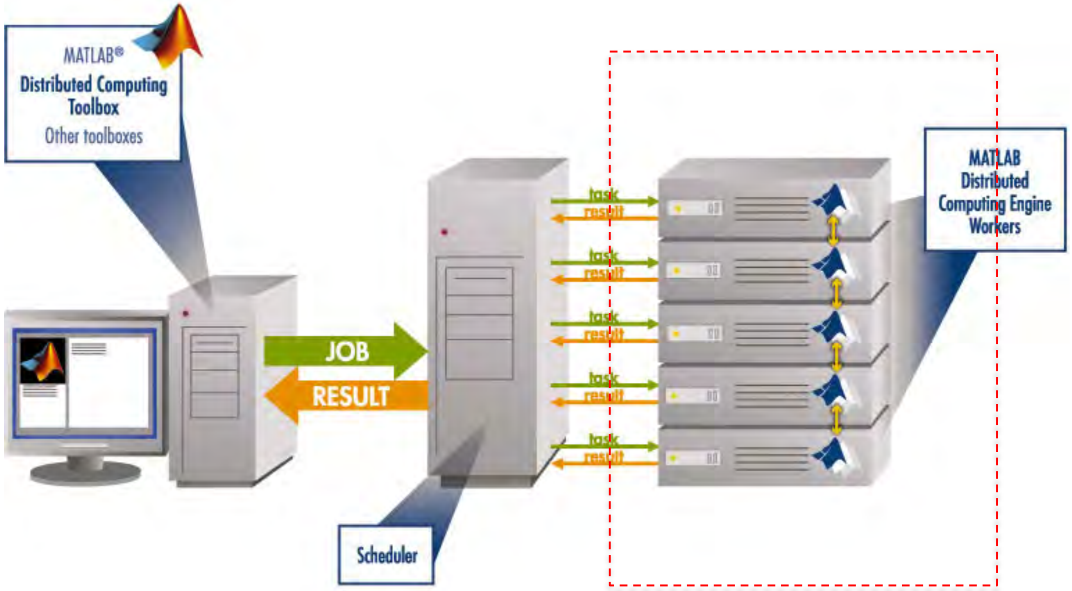
Session 1

Big Data Infrastructure

OVERVIEW

Big Data Technology Before Big Data

Parallel Processing



Processing Components

Big Data Technology Before Big Data

“Yes, Wall Street has been doing this but now the rest of the world is catching up.”

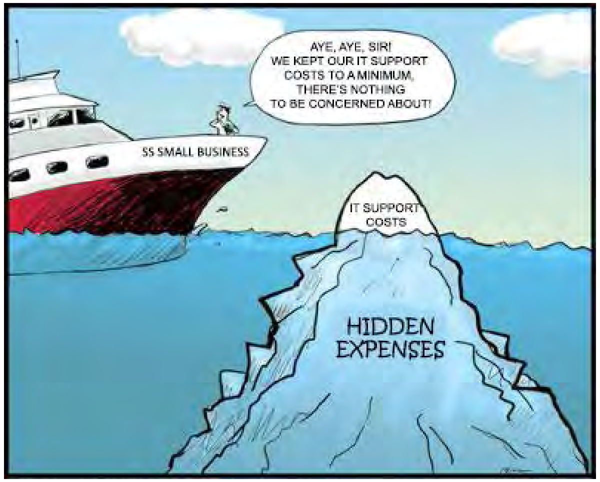
- Ivy Schmerken
  (Editor at Large, Wallstreetandtech.com)

Technology Problems

– Expensive

– Slow

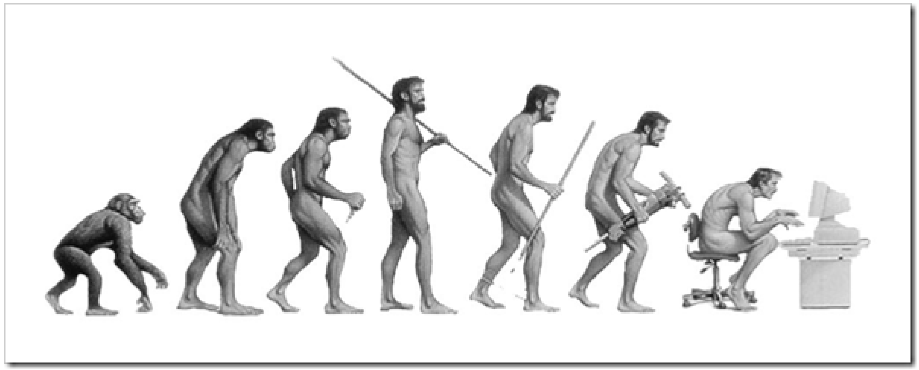
– Not Incrementally Adoptable (All or Nothing)


AYE, AYE, SIR! WE KEPT OUR IT SUPPORT COSTS TO A MINIMUM. THERE'S NOTHING TO BE CONCERNED ABOUT!
SS SMALL BUSINESS
IT SUPPORT COSTS
HIDDEN EXPENSES

Technology Evolution

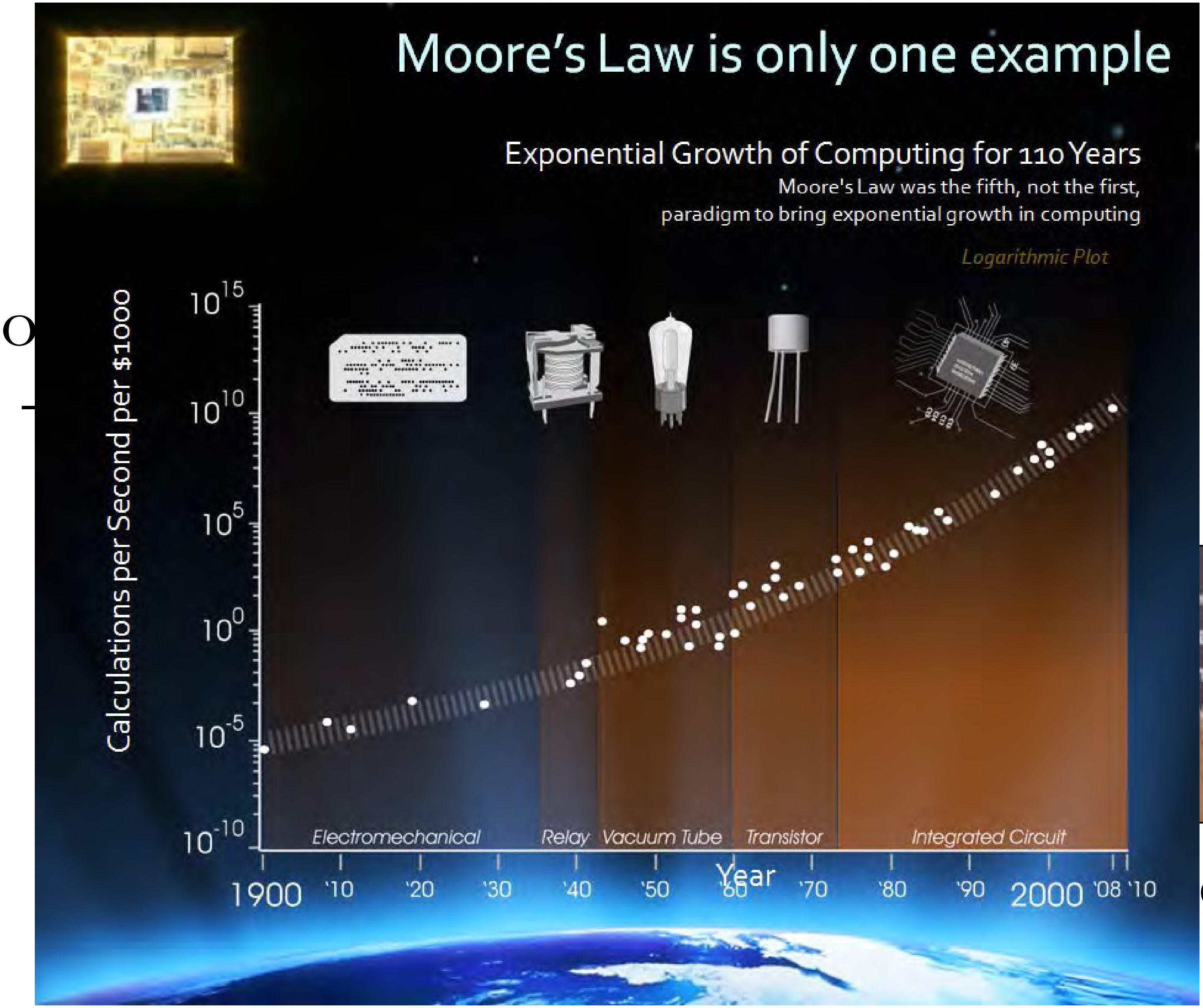
Key Factors Making Technology More Accessible

- Decreased Hardware Costs
- Decreased Software Costs
- Inexpensive Large-Scale Storage Options
- Inexpensive Parallel Processing Options



Mo

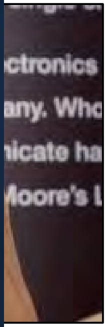
-



Moore's Law is only one example

Exponential Growth of Computing for 110 Years

Moore's Law was the fifth, not the first, paradigm to bring exponential growth in computing



ctronics
any. Who
icate ha
loore's l

e

Open Source Software

Open Source Software

- Computer source code made available with a license in which the copyright holder provides the rights to study, change and distribute the software…
  - With lots of fine print
- Richard Stallman – Founder of Gnu

Open Source

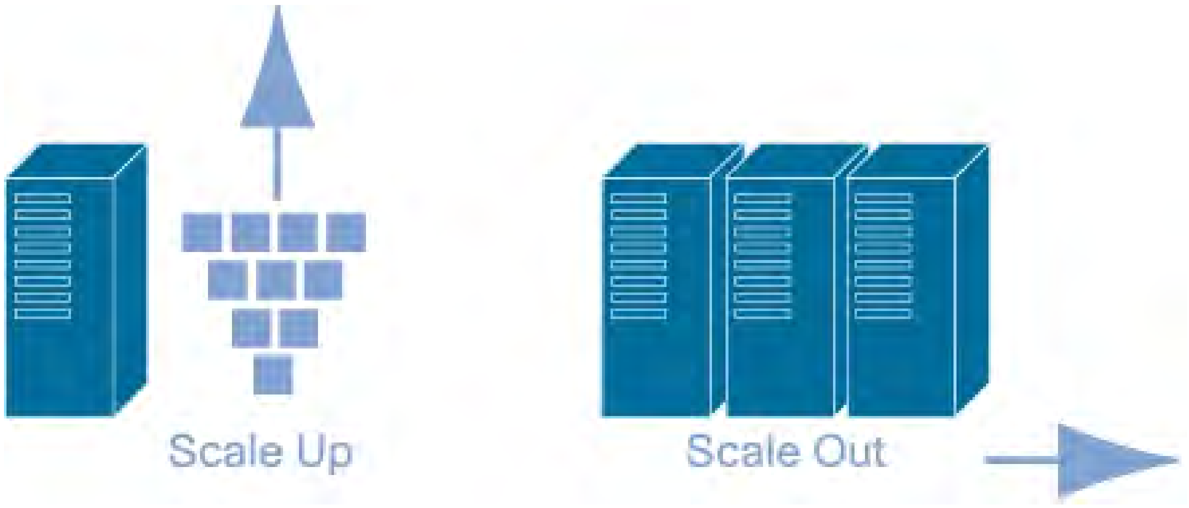
- Advantages
  - Crowd-sourced reliability
  - Cheaper adoption costs
- Disadvantages
  - Lack of reliability
  - Expensive adoption costs

"Control over the use of one's ideas really constitutes control over other people's lives; and it is usually used to make their lives more difficult. "

Richard Stallman

Inexpensive Large-Scale Infrastructure

Handling Large-Scale Data



Few Very Large Servers vs Many Smaller (commodity) Servers

Big Data: Infrastructure - Challenges

Metal: Computing Resources

- Processing, Storage, Memory, Networking

Management: Resource Provisioning

- Virtualization, Orchestration

Models: Storage, Parallel Processing

- Databases
  - Consistency, Availability, Partitioning
- Distributed File Systems
  - Data locality
- Complex Applications
  - Batch, interactive, streaming

Webinar Roadmap

Big Data definitions
Infrastructure
Metal
Management
Models
(part 1)

(part 2)
Algorithms
Data Processing

Machine Learning
Implications
Privacy
Smart Cities
Technology Tradeoffs

Session 1

Big Data Infrastructure

METAL: COMPUTING RESOURCES

Computer Basics

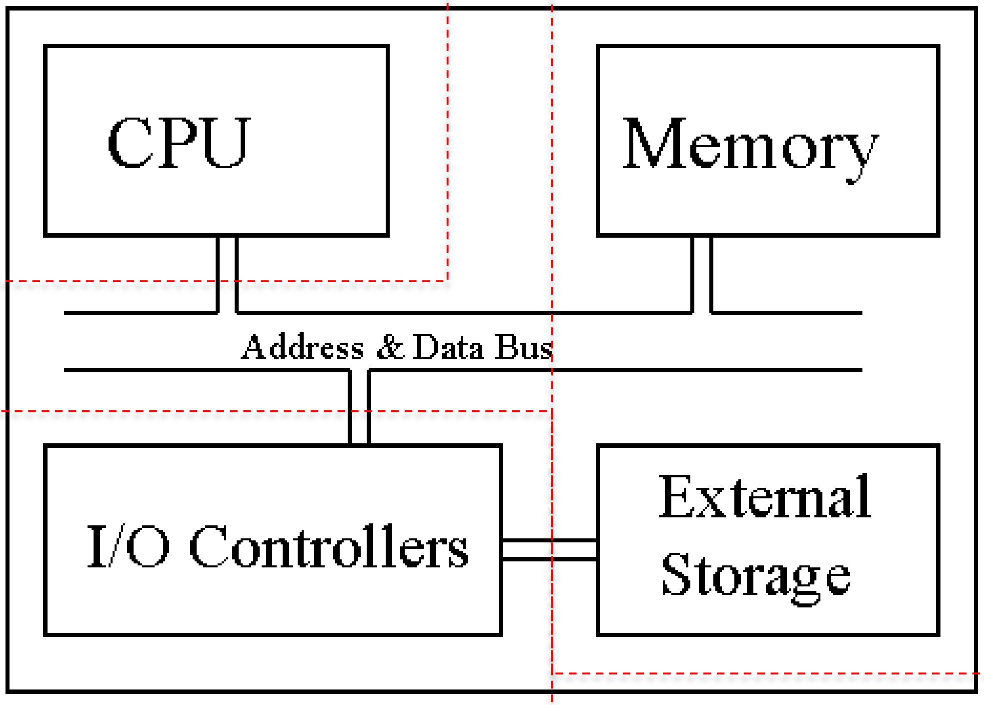
Central Processing Unit

- Given instructions for processing data (software)

Data To Be Processed

- Intermediate Data
- Persistent Data

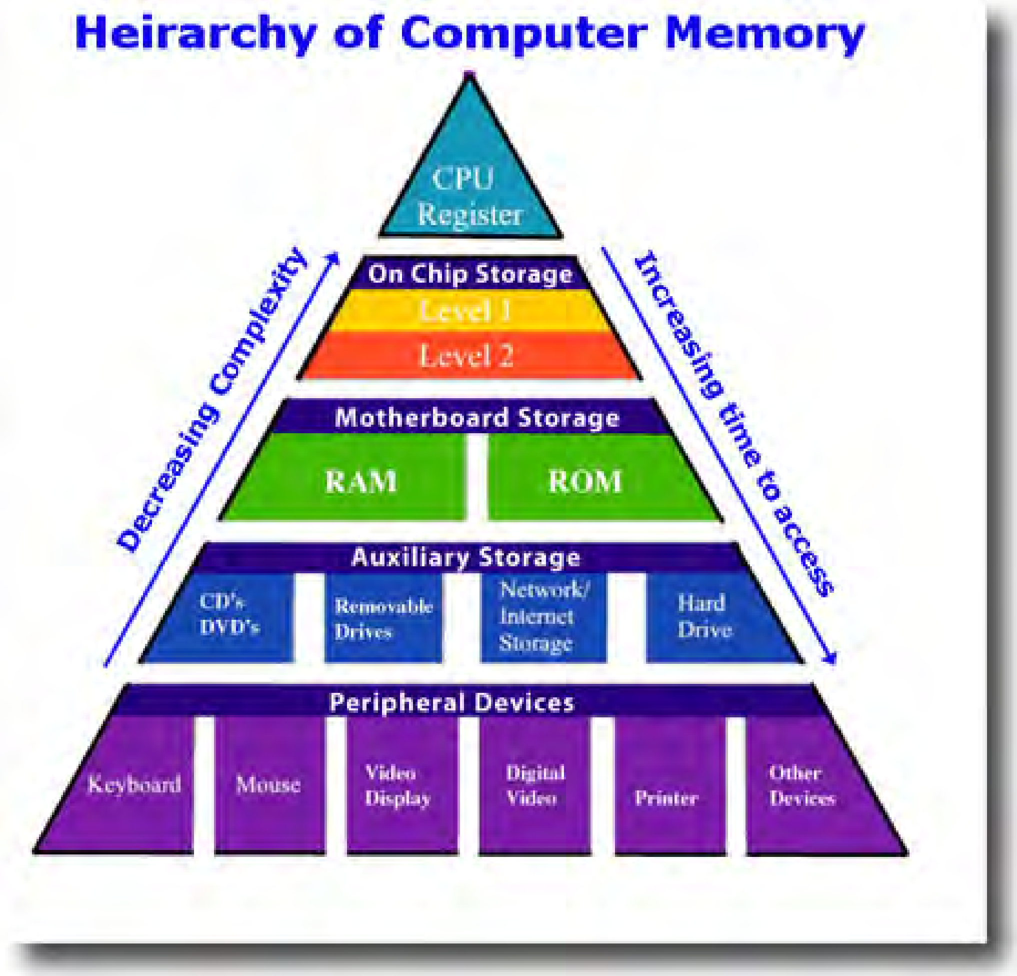
Basic Digital Computer Architecture



Input/Output

- Keyboard, Mouse, Video, …
- Network Access

Computer Basics



Computer Basics: More Resources

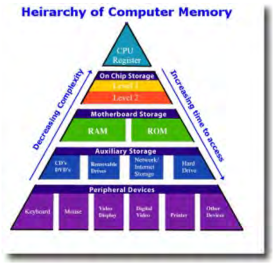
The power of the cloud

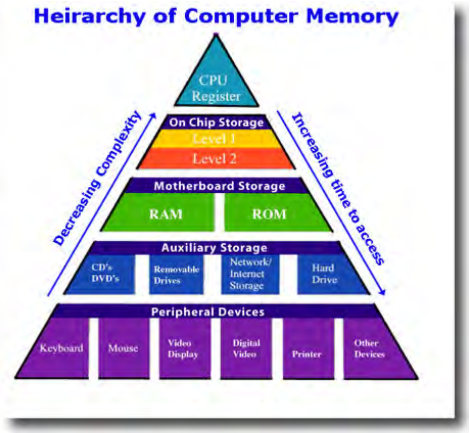
Single Machine Applications | App 1 | App 2 | App N

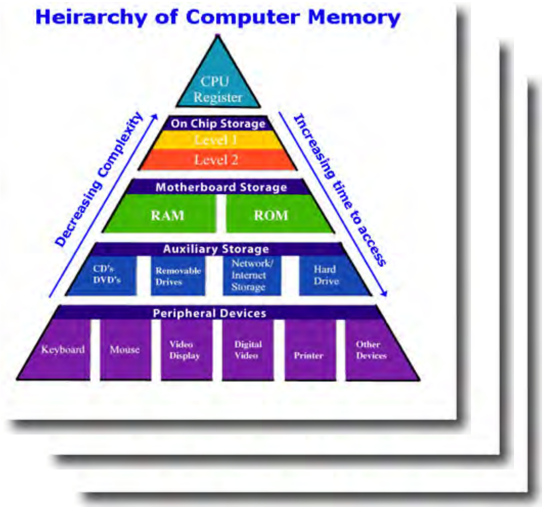
Single Machine Operating System

Single Machine Applications | 1-N

Single Machine Operating System







Desktop/Laptop
RAM: 8GB
HD: 500GB
Cores: 4
GPU: Intel HD Graphics 4000 512 MB

Server [$10Ks]
RAM: 384GB
HD: 7.3TB
Cores: 16

Server [$0.50-$3.50] / hr
RAM: 244 GB
HD: 240GB (SSD)
Cores: 32

Computer Basics: More Resources

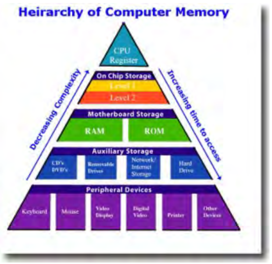
The power of the Big Data

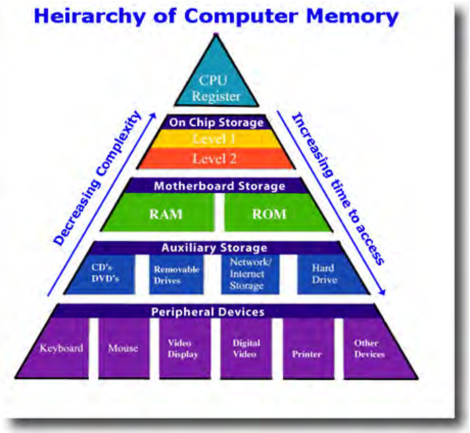
Single Machine Applications | App 1 | App 2 | App N

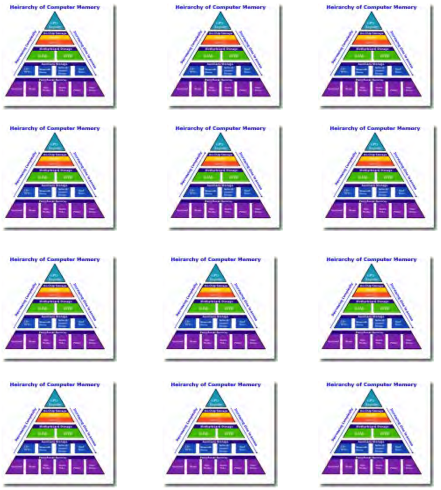
Single Machine Operating System

‘Big Data’ Applications | 1-N

‘Data Center Scale’ Operating System







Desktop/Laptop
RAM: 8GB
HD: 500GB
Cores: 4
GPU: Intel HD Graphics 4000 512 MB

Server [$10Ks]
RAM: 384GB
HD: 7.3TB
Cores: 16

Cluster/Cloud
RAM: ∞ ?
HD: ∞ ?
Cores: ∞ ?

Webinar Roadmap

- Big Data definitions
- Infrastructure
  - Metal
  - Management
  - Models

(part 1)

Session 1

(part 2)

- Algorithms
  - Data Processing

- Machine Learning
- Implications
  - Privacy
  - Smart Cities
  - Technology Tradeoffs

Big Data Infrastructure

MANAGEMENT: RESOURCE PROVISIONING

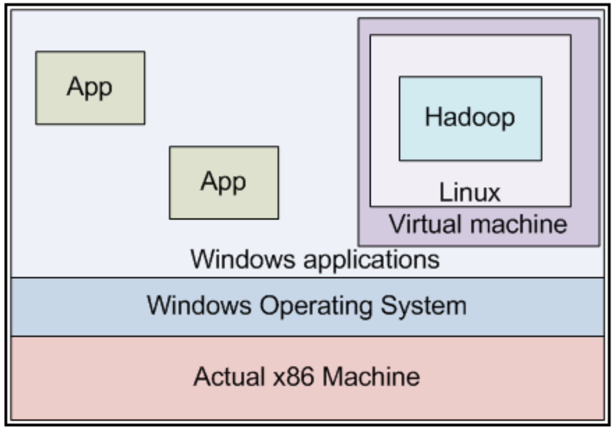
Cloud Computing

NIST Definition

- a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models.

Characteristics	Service Models	Deployment Models
• *On-demand self-service*	• *Software as a Service (SaaS)*	• *Private cloud*
• *Broad network access*	• *Platform as a Service (PaaS)*	• *Community cloud*
• *Resource pooling*	• *Infrastructure as a Service (IaaS)*	• *Public cloud*
• *Rapid elasticity*		• *Hybrid cloud*
• *Measured service*		

Virtualization



1 Physical Machine -> N Virtual Machines

Creating a machine = creating a file

Turn on a machine / freeze an image = click a button

When you spin up an EC2 instance on AWS – that's also a VM

- When you spin up 100 EC2 instances on AWS…those are also VMs

Definitions (Gartner)

- **Virtualization** is the abstraction of IT resources that masks the physical nature and boundaries of those resources from resource users. An IT resource can be a server, a client, storage, networks, applications or OSs. Essentially, any IT building block can potentially be abstracted from resource users.

Is virtualization the most efficient way to manage resources for big data?

Webinar Roadmap

Big Data definitions
Infrastructure
  Metal
  Management
  Models
    (part 1)

Session 1

    (part 2)
Algorithms
  Data Processing
  Machine Learning
Implications
  Privacy
  Smart Cities
  Technology Tradeoffs

Big Data Infrastructure: Hadoop

Hadoop is

- A data processing programming model
- A resource management framework
- A distributed file system

Hadoop is

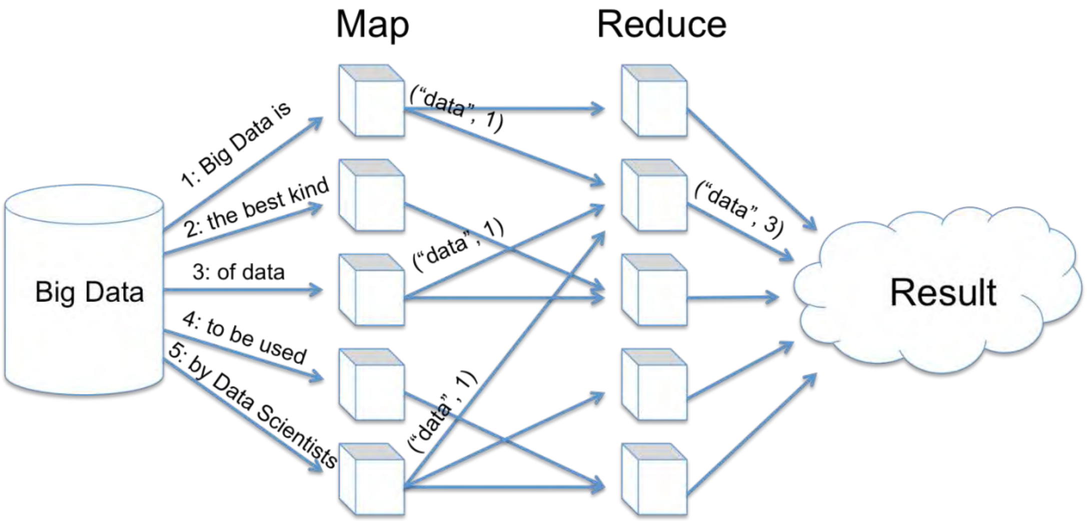
- A popular, open source parallel processing framework
- An implementation of the MapReduce algorithm

It

Data Processing Model and Resource Management

MapReduce* is

- A programming model for parallel data processing
- A cluster resource management framework

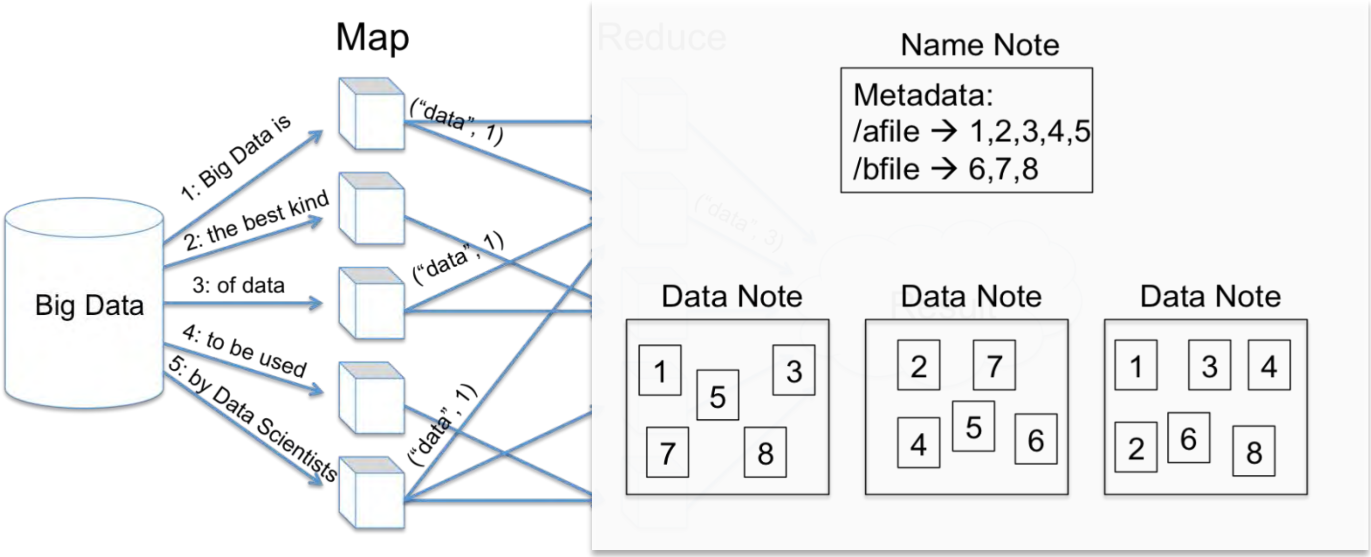


* Hadoop 1.0

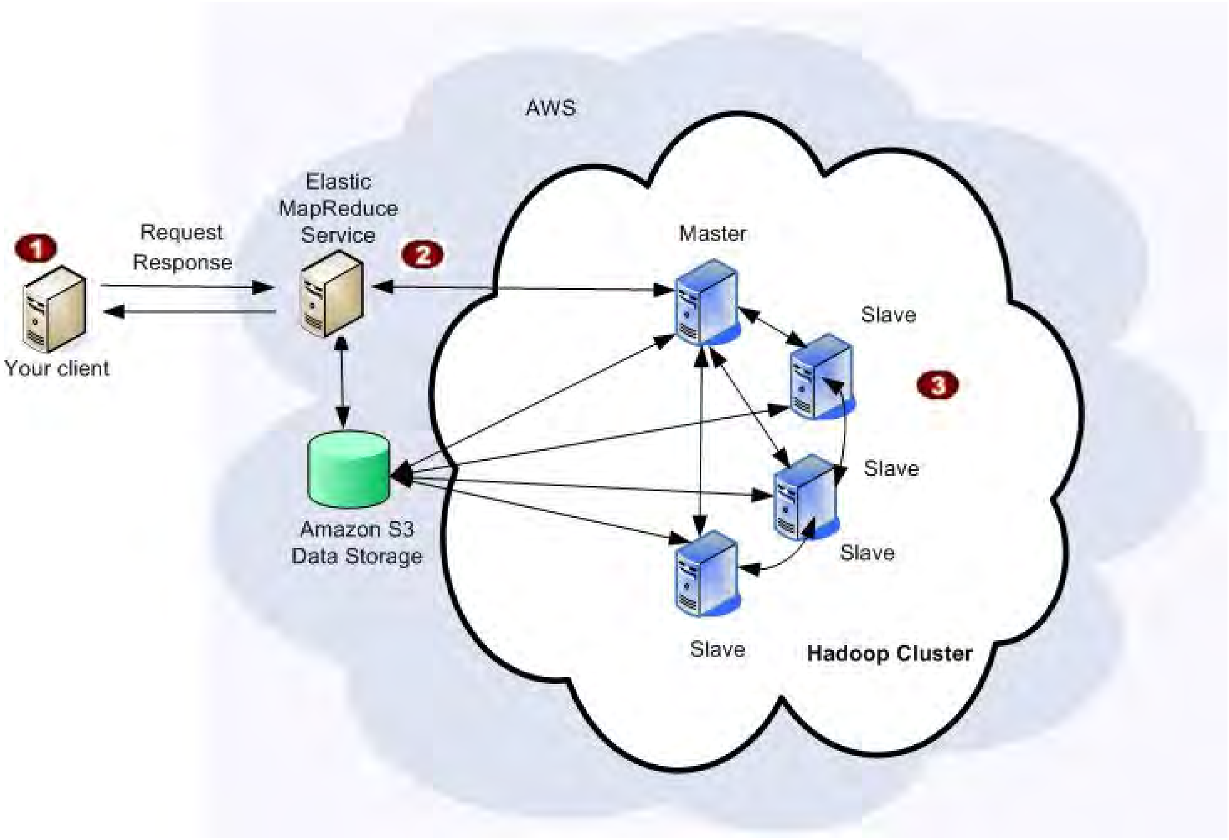
Distributed File System

HDFS is

- A redundant, reliable storage framework
- HBase is a key-value store built on HDFS



Amazon EMR Cluster – In the Cloud



As little as $0.15 / hr for a 10 node cluster

Summary

Big Data…

- Is driven by technology
- Can be organized by
  - Infrastructure
  - Algorithms
  - Implications
- Infrastructure leverages parallel processing on cluster computing systems

TMIP Updates

For future webinar announcement, please sign up for GovDelivery at http://www.fhwa.dot.gov/planning/tmip/ if you have not done so.

TMIP Contacts

If you have any questions or comments about today's presentation or TMIP, or if you are interested in sharing your experience, please contact me at:

sarah.sun@dot.gov or feedback@tmip.org.