Introduction to Big Data

Initial survey

“Big Data Era”

- ✓ 90% of worlds’ data generated over last two years

- ✓ A single jet engine produces *20TB* (10^{12}B) *of data per hour*
- ✓ Facebook has 2.89 billion users, n*140 billion links, about 4 PB of data **per day** (2021)
- ✓ *Genome of human:* sampling, biochemistry, immunology, imaging, genetic, phenotypic data
  - 1 person: 1PB (10^{15}B)
  - 1000 people: 1EB (10^{18}B)
  - 1 billion people: 1ZB *(10^{24}B)*

Big data is a relative notion: 1TB is already too big for your laptop

But What is big data anyway?

Big data: What is it anyway?

No standard definition!

- **Big data** is the **term** for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

- The **challenges** include capture, curation, storage, search, sharing, transfer, analysis, and visualization.

- The **trend** due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."

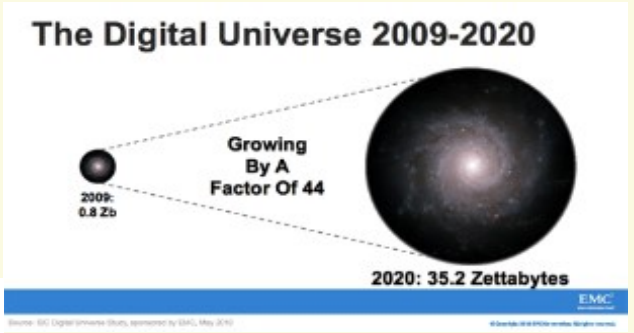
Big data: the 4 V's

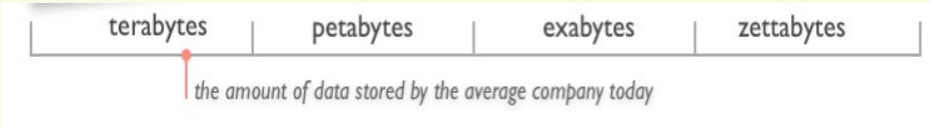
- ✓ *Volume: horrendously large*
  - PB (10^{15}B)
  - *EB* (10^{18}B)
- ✓ *Variety: heterogeneous, semi-structured or unstructured*
  - *9:1* ratio of unstructured data vs. structured data
  - collecting 95% restaurants requires at least 5000 sources
- ✓ *Velocity: dynamic, streams*
  - think of the Web and Facebook, …
- ✓ *Veracity: trust in its quality*
  - *real-life data is typically dirty!*

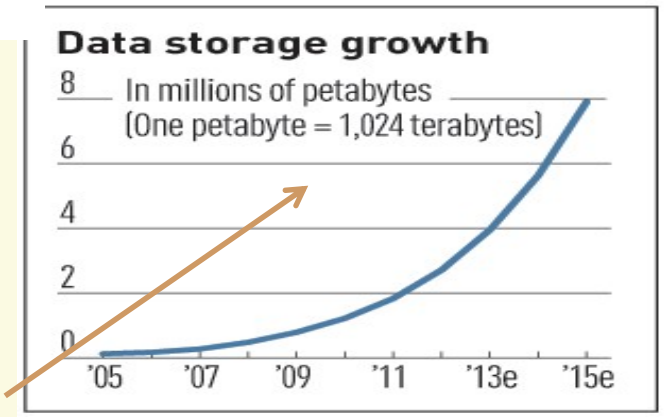
A departure from our familiar data management!

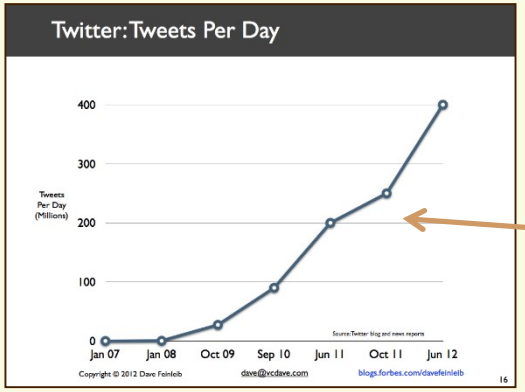
Volume (Scale)

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







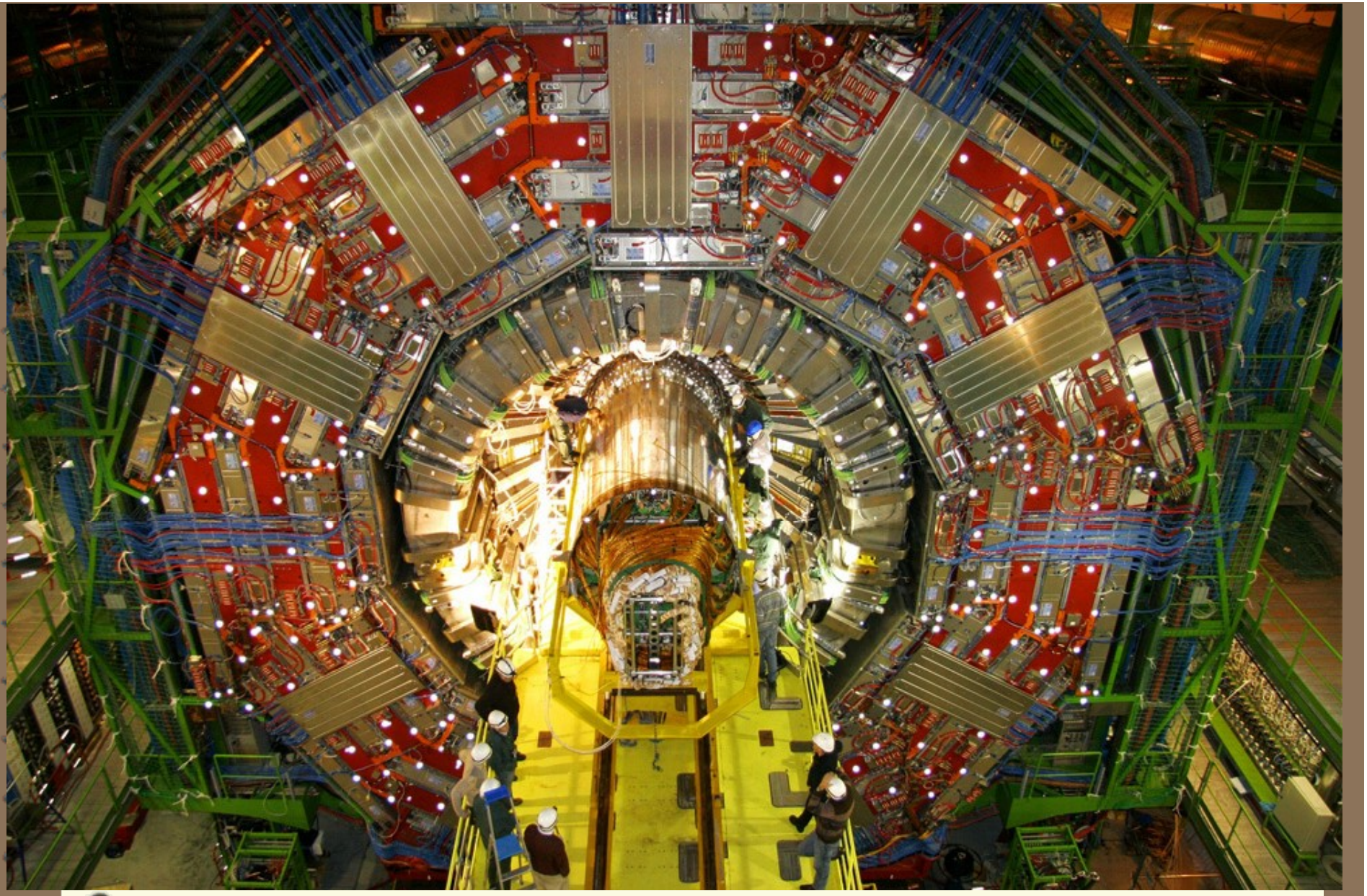


Exponential increase in collected/generated data

The Earthscope

- “The Earthscope is the world's largest science project. Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data. It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more.”

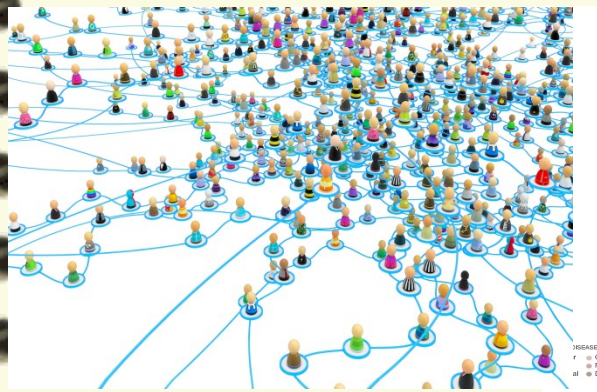
- (http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI)



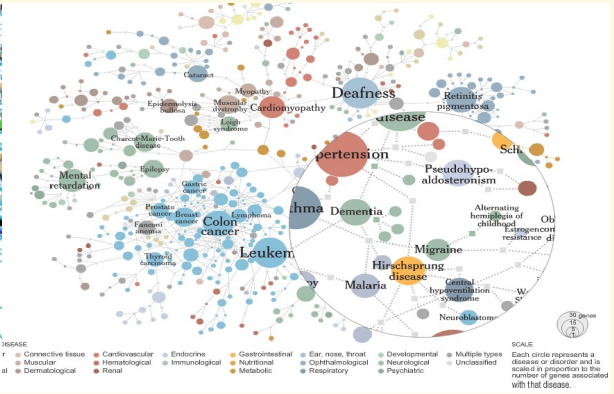
CERN's Large Hydron Collider (LHC) generates 15 PB a year

Maximilien Brice, © CERN

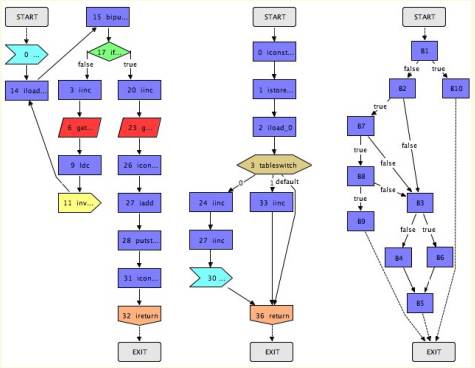
… and no data is an island



social networks

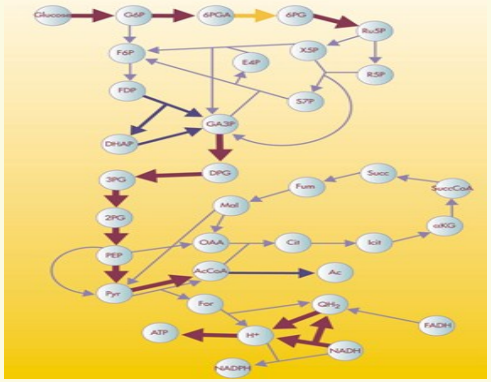


knowledge graph

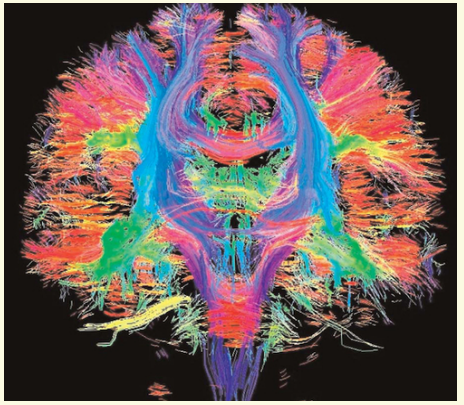


control flow graph

cyber networks

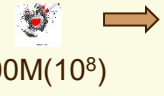


metabolic networks

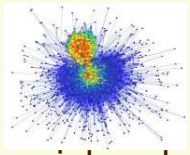


brain network

Real-life scope

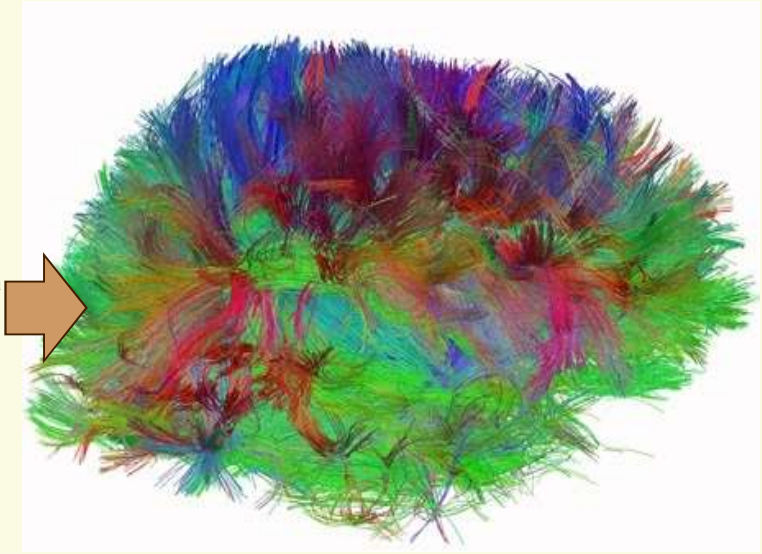


100M(10^8)



social scale
100B (10^{11})

Web scale
1T (10^{12})



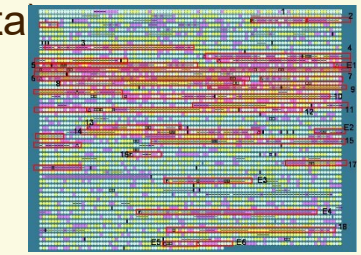
brain scale, 100T (10^{14})
(human brain project)

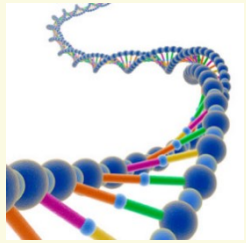
Real-life scope

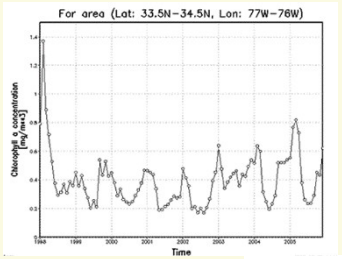
Challenge 1: Find needle in the haystack?

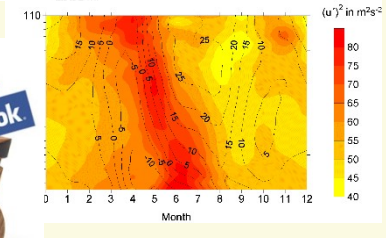
Variety (Complexity)

- ✓ Relational Data (Tables/Transaction/Legacy Data)
- ✓ Text Data (Web)
- ✓ Semi-structured Data (XML)
- ✓ Graph Data
  - Social Network, Semantic Web (RDF), …
- ✓ Streaming Data
  - You can only scan the data once
- ✓ A single application can be generating/collecting many types of data
- ✓ Big Public Data (online, weather, finance, etc)



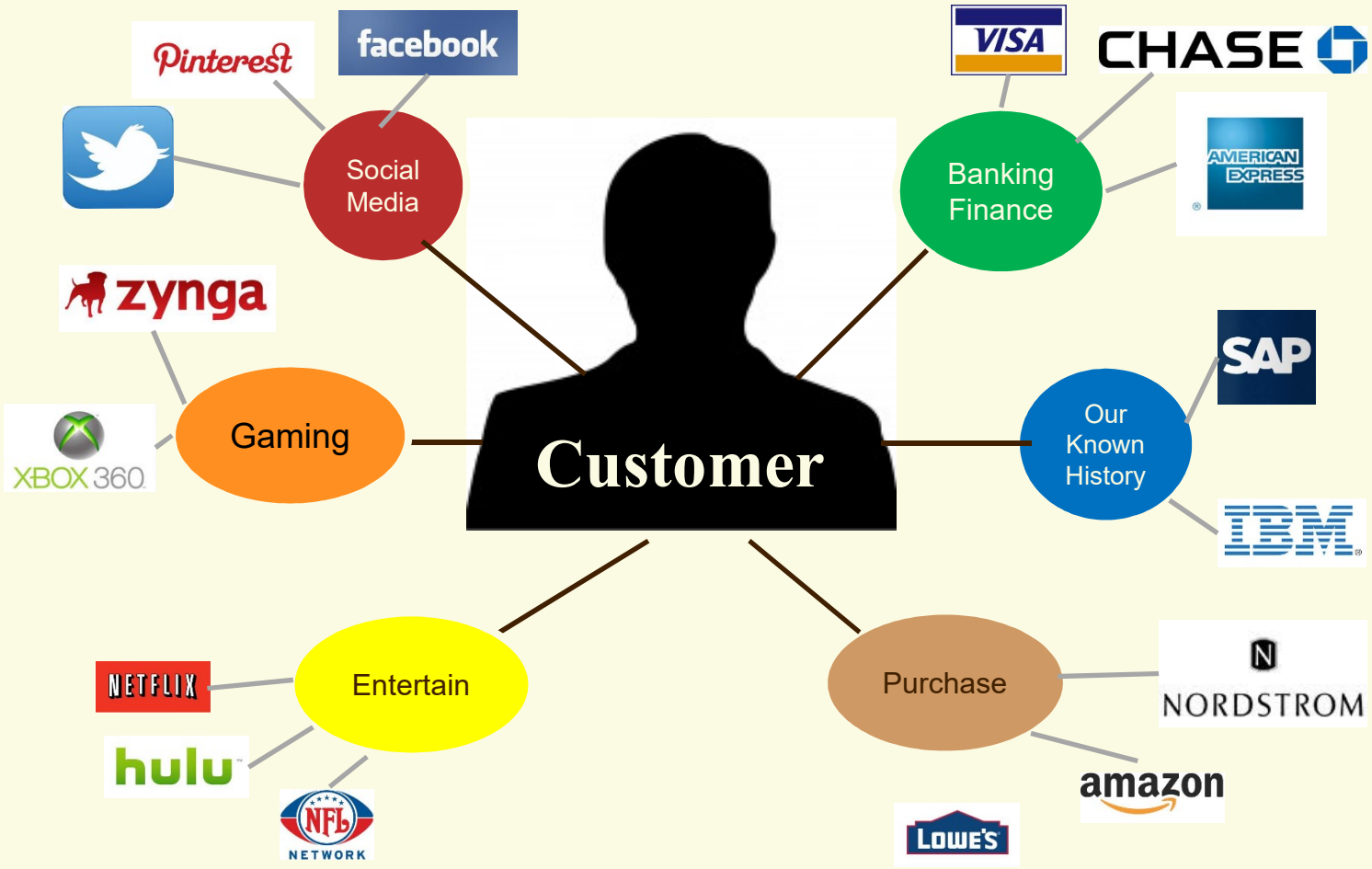




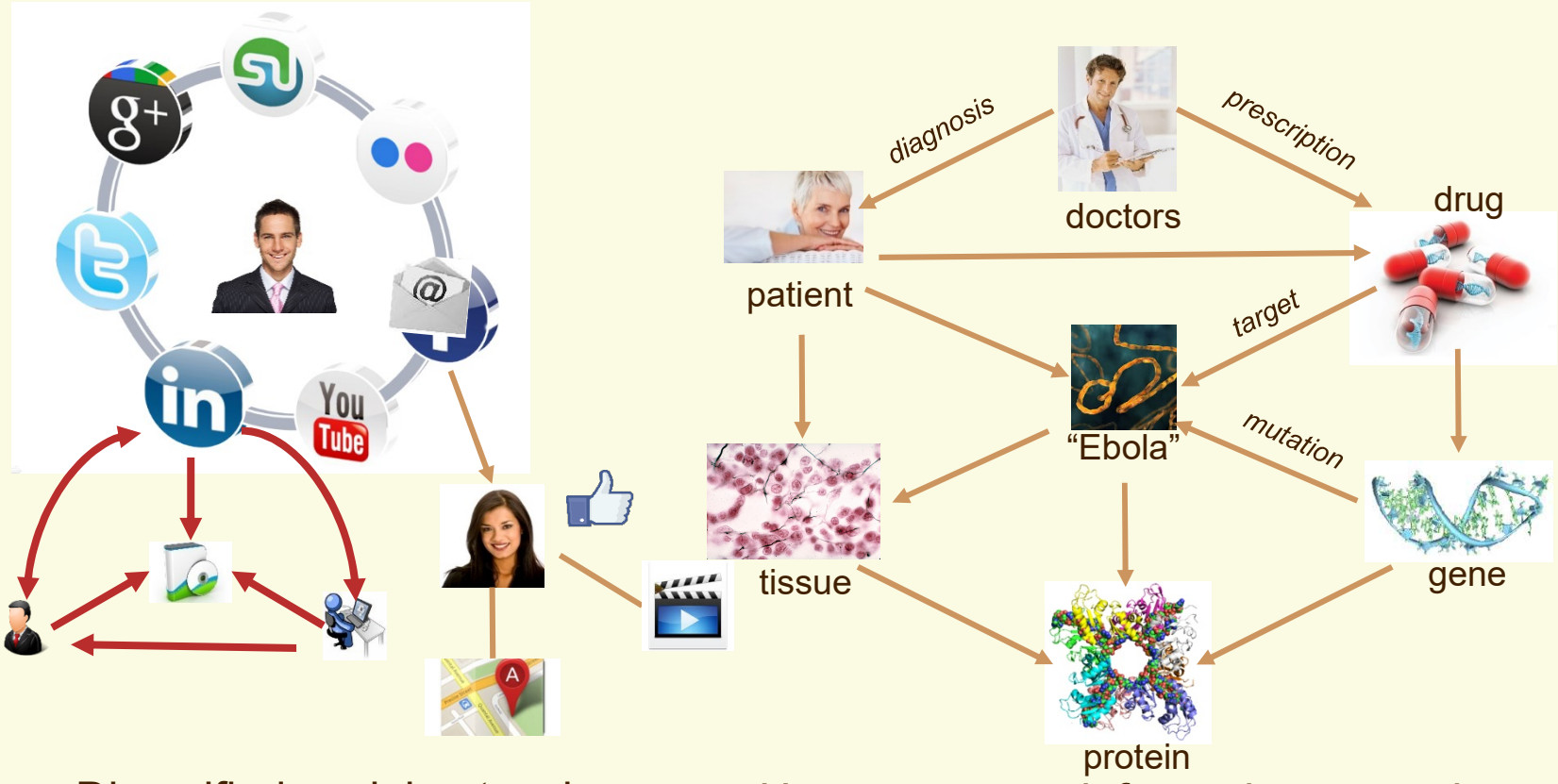


To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



A Global View of Linked Big Data



Challenge 2: "data wrangling"


IN 60 SECONDS...
1 NEW
1,600+ READS ON Scribd.
13,000+ HOURS MUSIC STREAMING ON PANDORA
12,000+ NEW ADS POSTED ON craigslist
370,000+ MINUTES VOICE CALLS ON skype
98,000+ TWEETS
320+ NEW twitter ACCOUNTS
100+ NEW LinkedIn ACCOUNTS
20,000+ NEW POSTS ON tumblr.
13,000+ iPhone APPLICATIONS DOWNLOADED
1 NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT
QUESTIONS ASKED ON THE INTERNET...
100+
40+
600+ NEW VIDEOS
6,600+ NEW flickr
50+ WordPress DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
25+ HOURS TOTAL DURATION
70+ DOMAINS REGISTERED
60+ NEW BLOGS
1,500+ BLOG POSTS
168 MILLION EMAILS ARE SENT
694,445 SEARCH QUERIES
1,700+ Firefox DOWNLOADS
29,364 WALL POSTS
510,040 COMMENTS
Google
GO-Globe.com

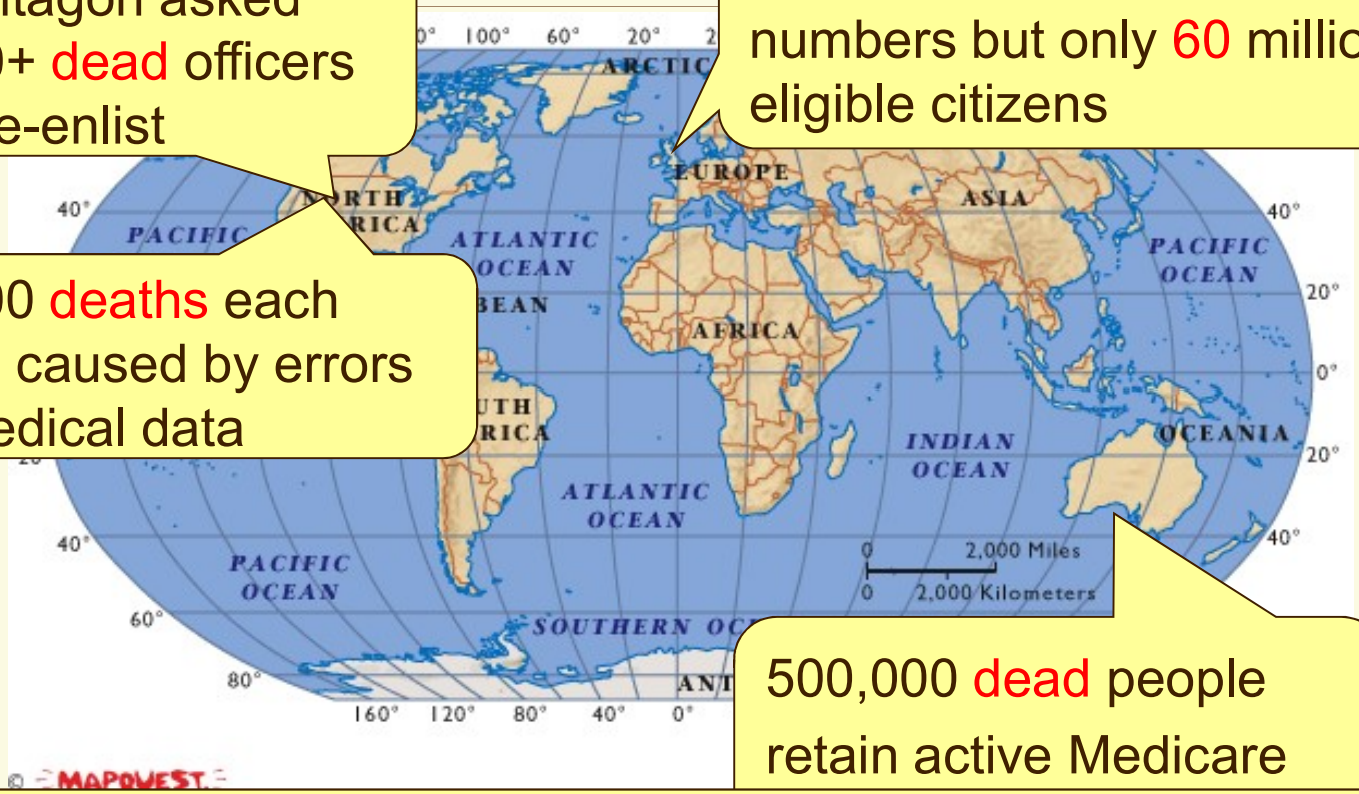
Velocity (Speed)

- ✓ Data is begin generated fast and need to be processed fast
- ✓ Online Data Analytics
- ✓ Late decisions ➔ missing opportunities
- ✓ 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
- ✓ **Challenge 3: "Drinking from a firehose"**

Data in real-life is often dirty

Pentagon asked 200+ dead officers to re-enlist

81 million National Insurance numbers but only 60 million eligible citizens



98000 deaths each year, caused by errors in medical data

500,000 dead people retain active Medicare

Data error rates in industry: 30% (Redman, 1998)

Challenge 4: Dirty data: inconsistent, inaccurate, incomplete, stale

Veracity (quality & trust)

Data = quantity + quality

When we talk about big data, we typically mean its quantity:

- ✓ What capacity of a system provides to cope with the sheer size of the data?
- ✓ Is a query feasible on big data within our available resources?
- ✓ How can we make our queries tractable on big data?
- ✓ . . .

Can we trust the answers to our queries?

- ✓ Dirty data routinely lead to misleading financial reports, strategic business planning decision ⇒ loss of revenue, credibility and customers, disastrous consequences

The study of data quality is as important as data quantity

Dirty data are costly

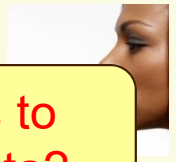
- ✓ Poor data cost US businesses $611 billion annually
- ✓ Erroneously priced data in retail databases cost US customers $2.5 billion each year
- ✓ 1/3 of system development projects were forced to delay or cancel due to poor data quality
- ✓ 30%-80% of the development time and budget for data warehousing are for data cleaning
- ✓ CIA dirty data about WMD in Iraq!

THE DATA WAREHOUSING INSTITUTE

DMReview 2000

PRICEWATERHOUSECOOPERS 2001

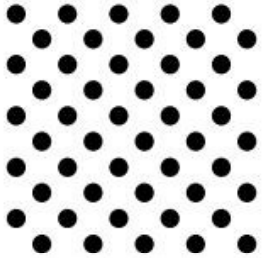
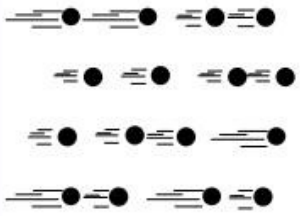
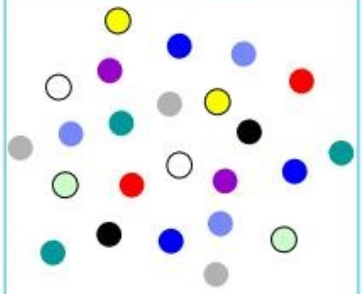
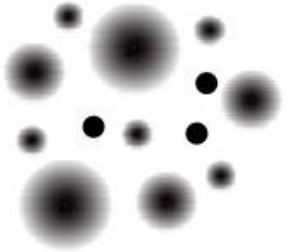
Merrill Lynch 1998



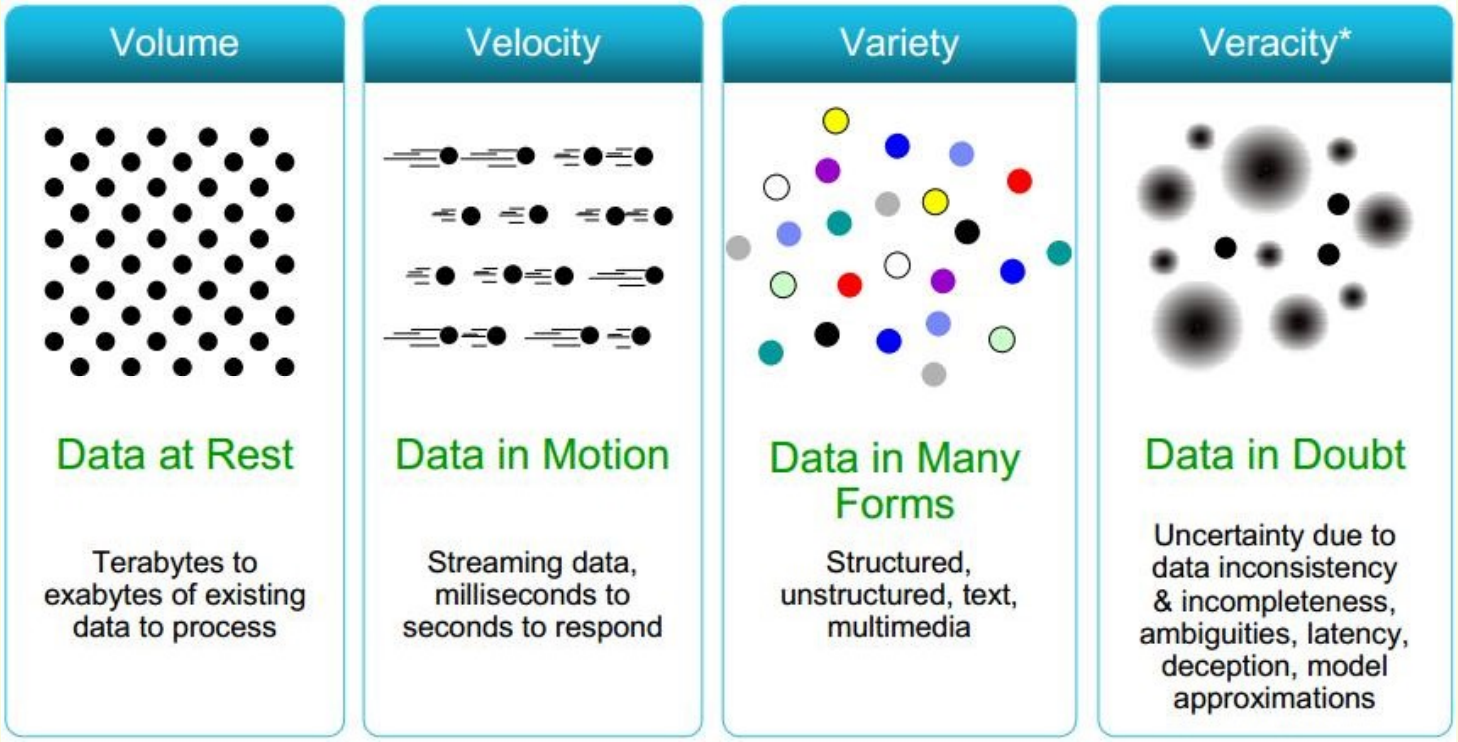
Can we trust answers to our queries in dirty data?

The scale of the data quality problem is far worse on big data!

The 4V's

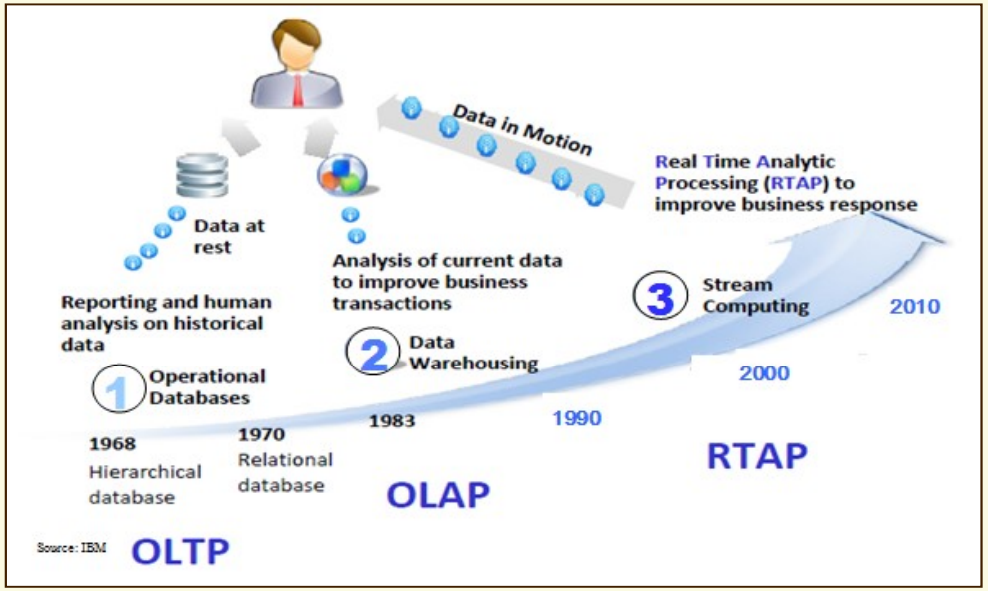
Volume	Velocity	Variety	Veracity*
			
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**
Terabytes to exabytes of existing data to process	Streaming data, milliseconds to seconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

The 4V's + n Vs…



Venue (location)
Vocabulary (semantics)
Value

Harnessing Big Data



OLTP: Online Transaction Processing (DBMSs)

OLAP: Online Analytical Processing (Data Warehousing)

RTAP: Real-Time Analytics Processing (Big Data Architecture & technology)

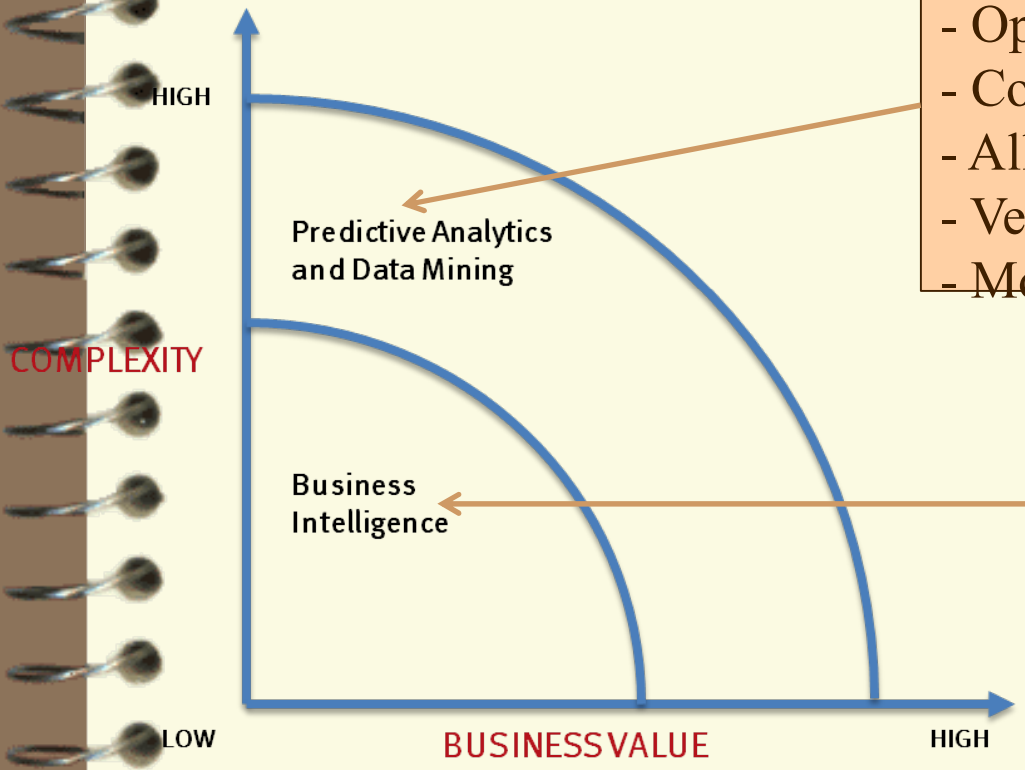
The Model Has Changed…

- **The Model of Generating/Consuming Data has Changed**

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

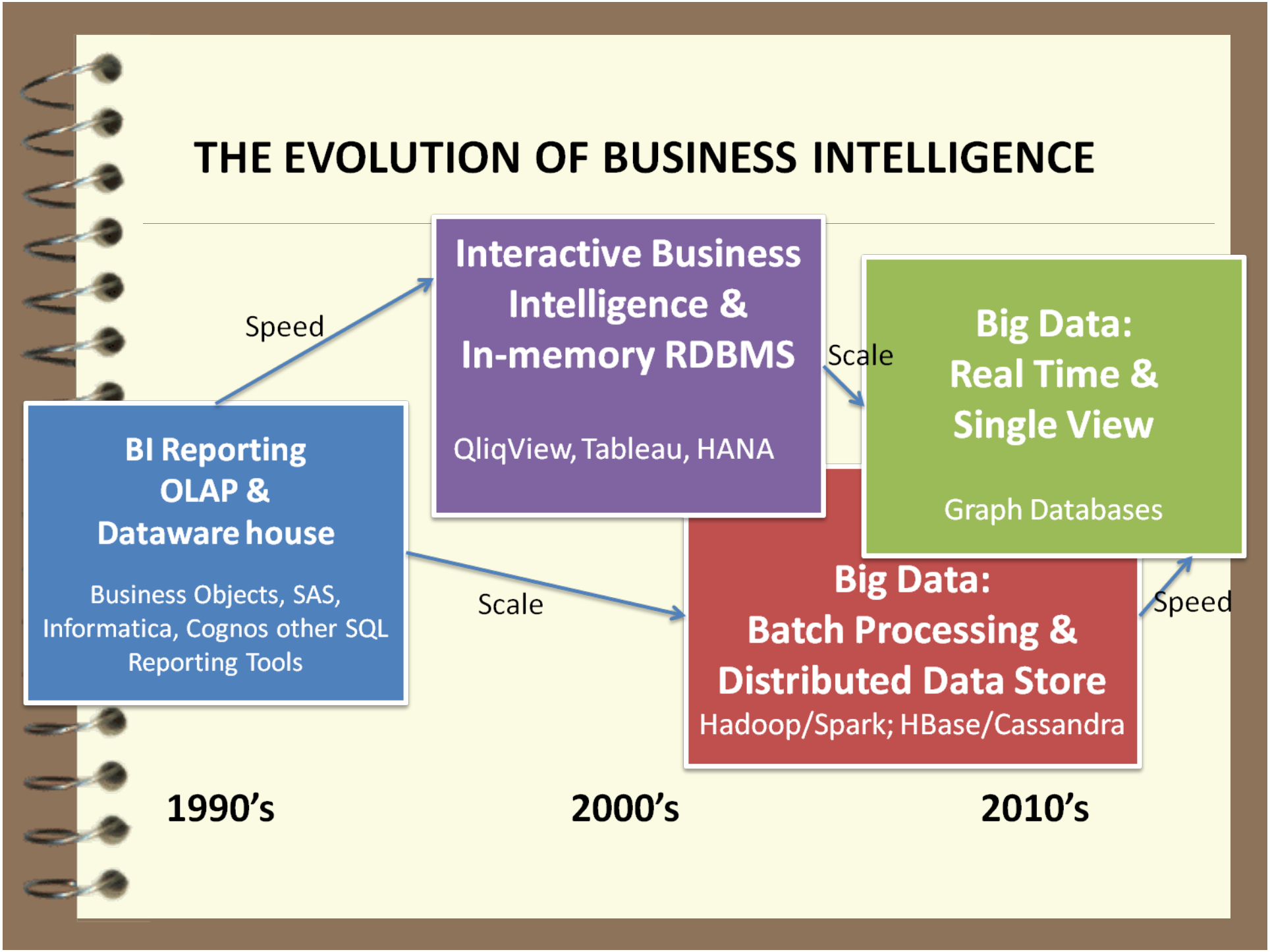
New Model: all of us are generating data, and all of us are consuming data

What's driving Big Data

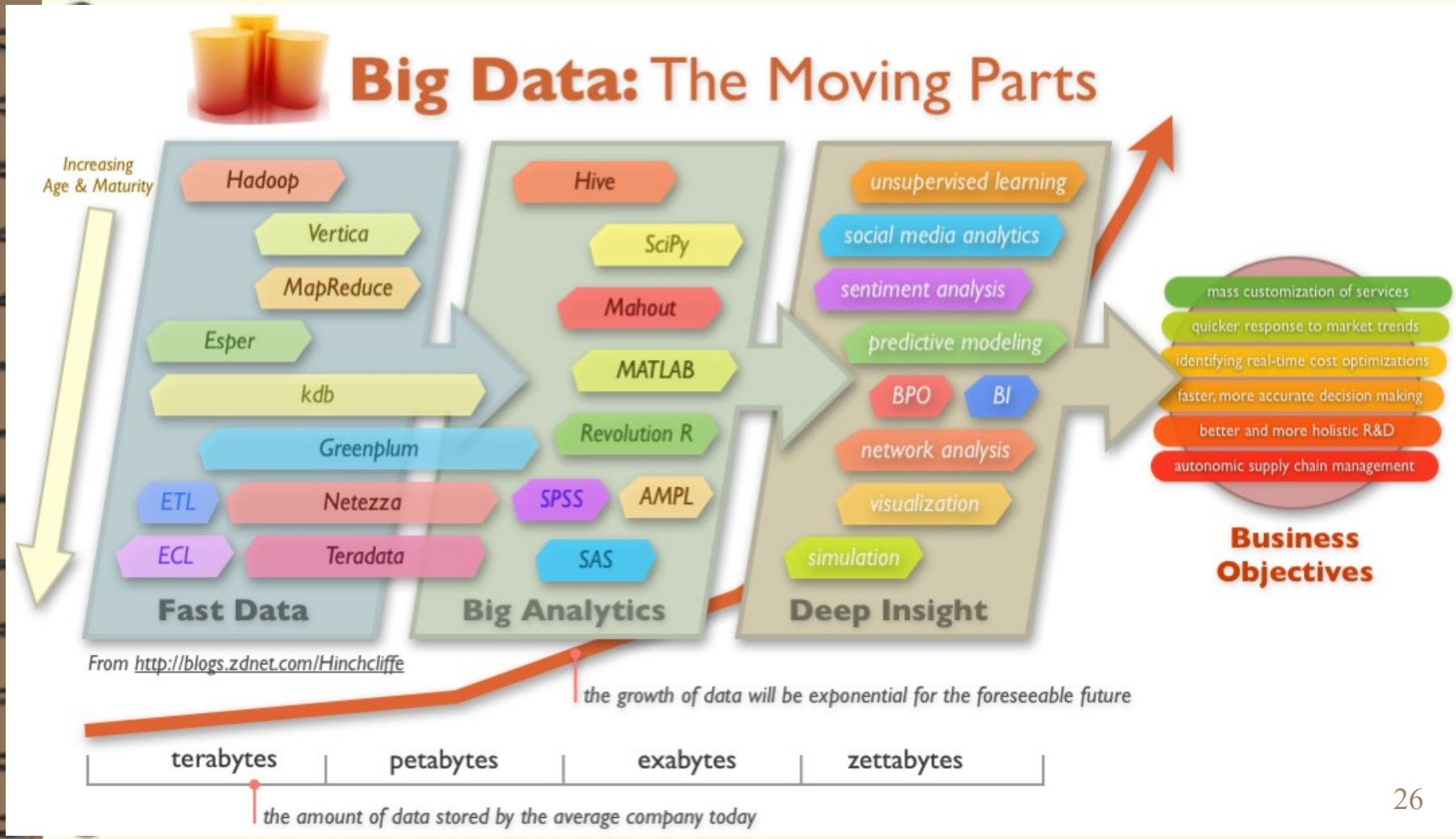


- Optimizations and predictive analytics
- Complex statistical analysis
- All types of data, and many sources
- Very large datasets
- More of a real-time

- Ad-hoc querying and reporting
- Data mining techniques
- Structured data, typical sources
- Small to mid-size datasets

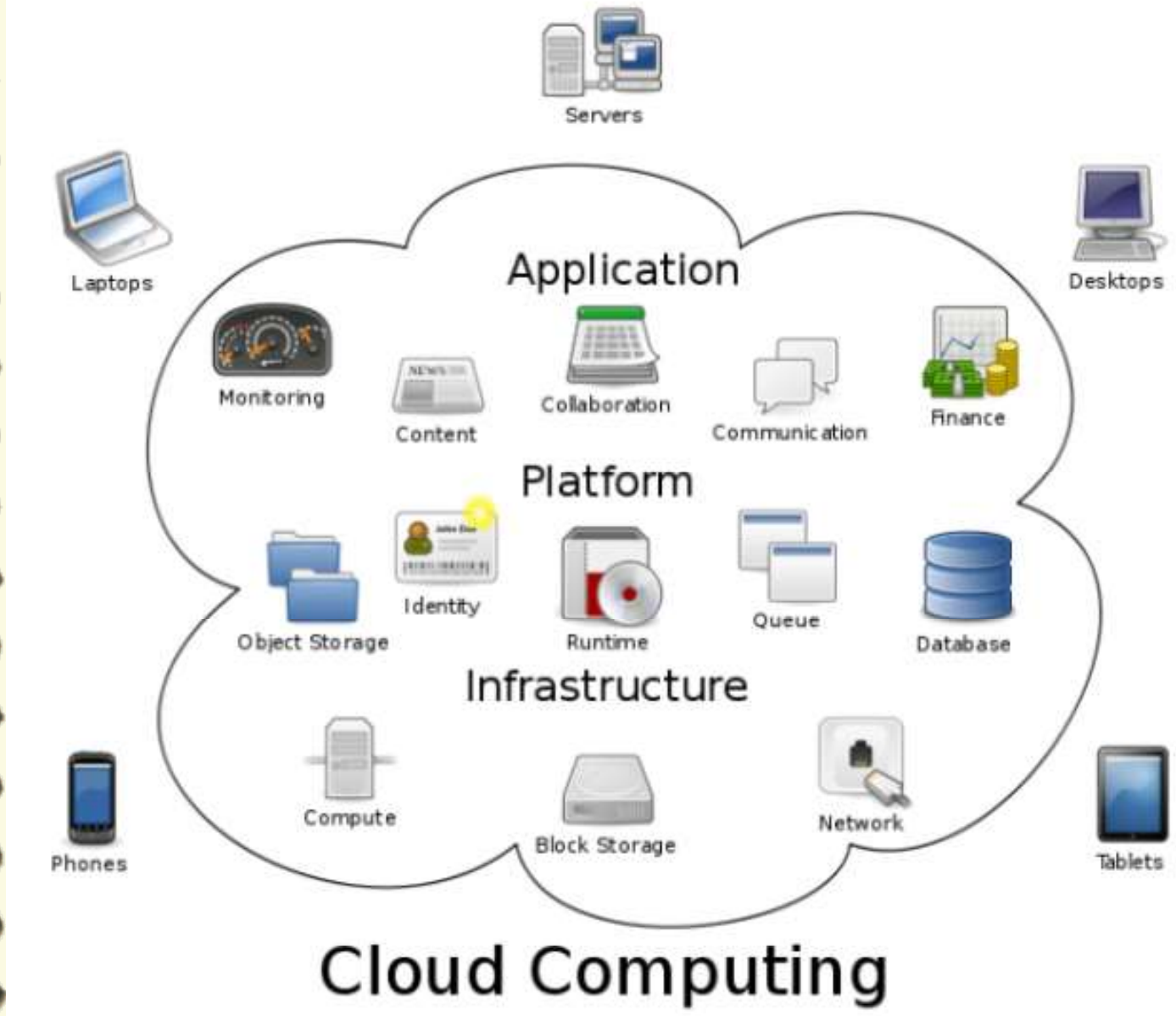

THE EVOLUTION OF BUSINESS INTELLIGENCE
Speed
Interactive Business Intelligence & In-memory RDBMS
QliqView, Tableau, HANA
Scale
Big Data: Real Time & Single View
Graph Databases
BI Reporting OLAP & Dataware house
Business Objects, SAS, Informatica, Cognos other SQL Reporting Tools
Scale
Big Data: Batch Processing & Distributed Data Store
Hadoop/Spark; HBase/Cassandra
Speed
1990's
2000's
2010's

Big Data Technology



Cloud Computing

- IT resources provided as a service
  - Compute, storage, databases, queues
- Clouds leverage economies of scale of commodity hardware
  - Cheap storage, high bandwidth networks & multicore processors
  - Geographically distributed data centers
- Offerings from Microsoft, Amazon, Google, …



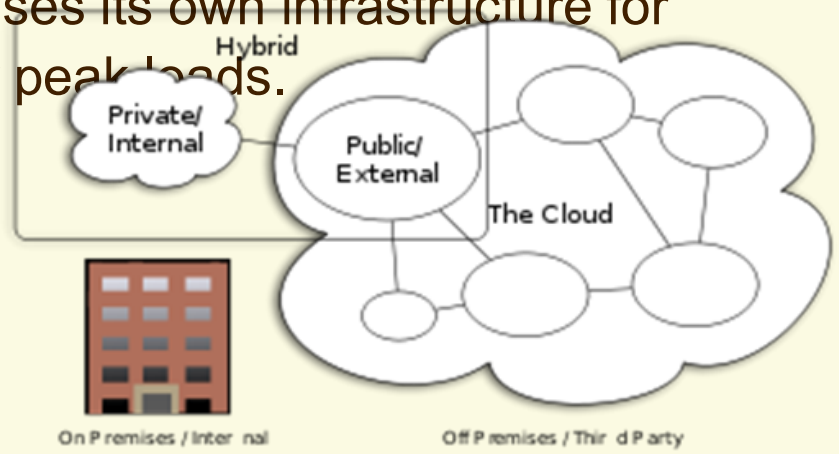
wikipedia:Cloud Computing

Benefits

- ✓ Cost & management
  - Economies of scale, “out-sourced” resource management
- ✓ Reduced Time to deployment
  - Ease of assembly, works “out of the box”
- ✓ Scaling
  - On demand provisioning, co-locate data and compute
- ✓ Reliability
  - Massive, redundant, shared resources
- ✓ Sustainability
  - Hardware not owned

Types of Cloud Computing

- **Public Cloud**: Computing infrastructure is hosted at the vendor's premises.
- **Private Cloud**: Computing architecture is dedicated to the customer and is not shared with other organisations.
- **Hybrid Cloud**: Organisations host some critical, secure applications in private clouds. The not so critical applications are hosted in the public cloud
  - **Cloud bursting**: the organisation uses its own infrastructure for normal usage, but cloud is used for peak loads.
- **Community Cloud**

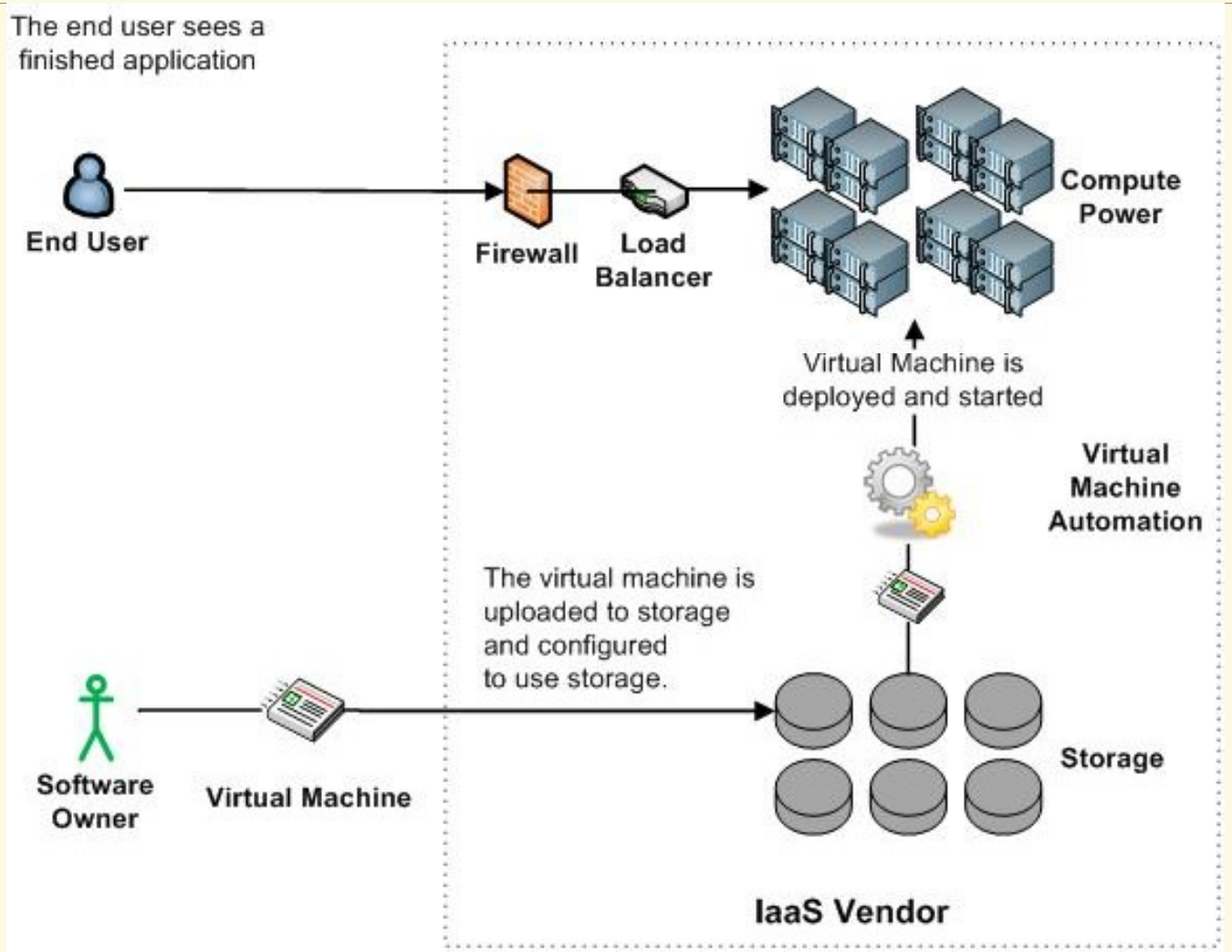


Cloud Computing Types

Classification of Cloud Computing based on Service Provided

- Infrastructure as a service (IaaS)
  - Offering hardware related services using the principles of cloud computing. These could include storage services (database or disk storage) or virtual servers.
  - Amazon EC2, Amazon S3, Rackspace Cloud Servers and Flexiscale.
- Platform as a Service (PaaS)
  - Offering a development platform on the cloud.
  - Google's Application Engine, Microsofts Azure, Salesforce.com's force.com .
- Software as a service (SaaS)
  - Including a complete software offering on the cloud. Users can access a software application hosted by the cloud vendor on pay-per-use basis. This is a well-established sector.
  - Salesforce.coms' offering in the online Customer Relationship Management (CRM) space, Googles gmail and Microsofts hotmail, Google docs.

Infrastructure as a Service (IaaS)



More Refined Categorization

- Storage-as-a-service
- Database-as-a-service
- Information-as-a-service
- Process-as-a-service
- Application-as-a-service
- Platform-as-a-service
- Integration-as-a-service
- Security-as-a-service
- Management/ Governance-as-a-service
- Testing-as-a-service
- Infrastructure-as-a-service

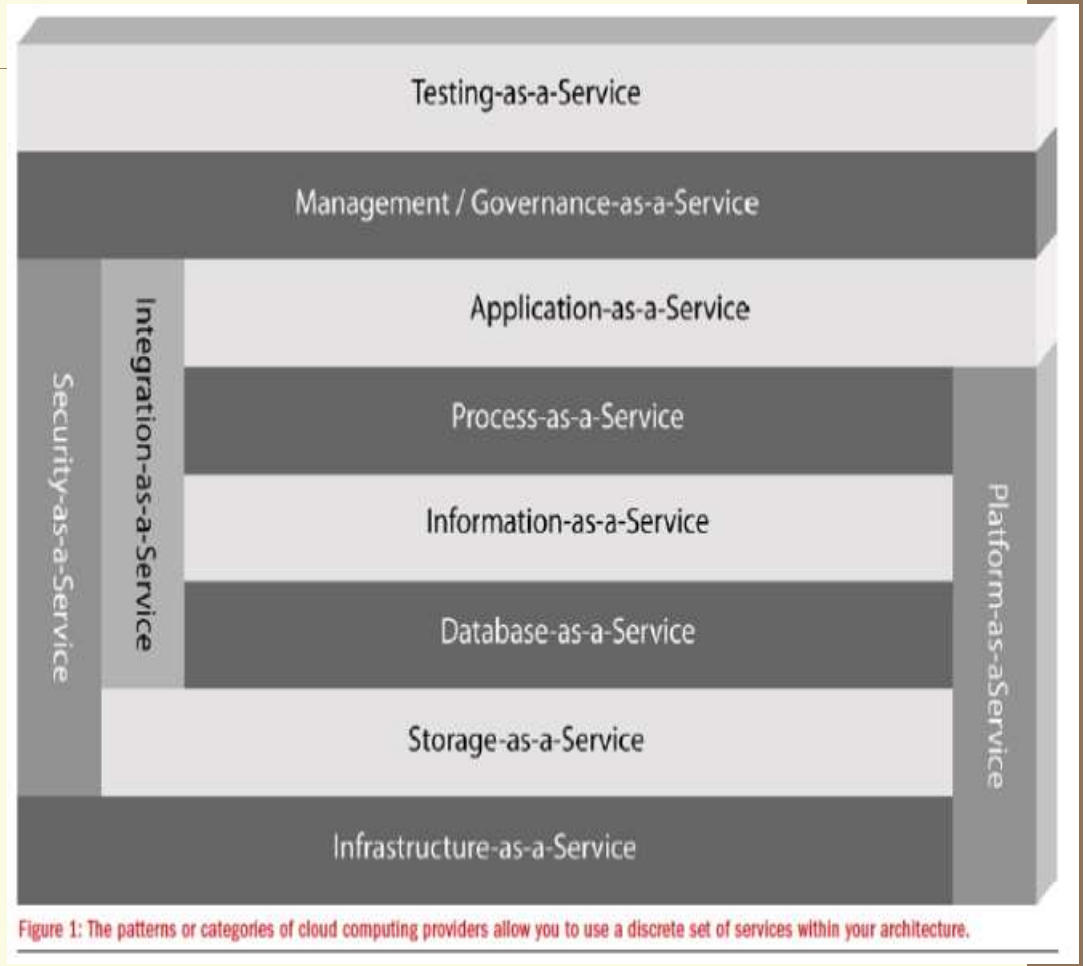


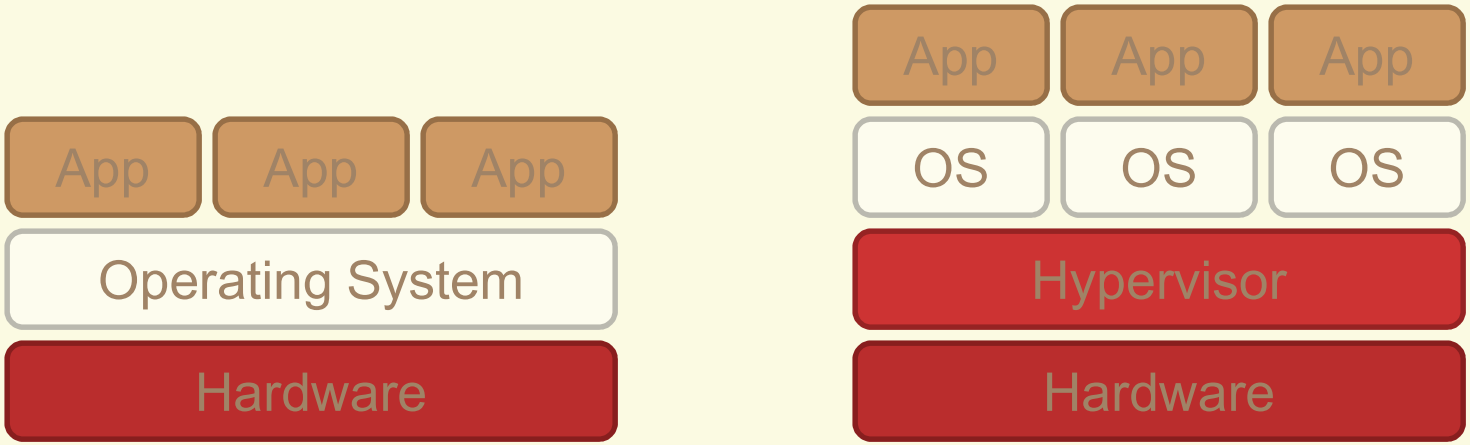
Figure 1: The patterns or categories of cloud computing providers allow you to use a discrete set of services within your architecture.

InfoWorld Cloud Computing Deep Dive

Key Ingredients in Cloud Computing

- ✓ Service-Oriented Architecture (SOA)
- ✓ Utility Computing (on demand)
- ✓ Virtualization (P2P Network)
- ✓ SAAS (Software As A Service)
- ✓ PAAS (Platform AS A Service)
- ✓ IAAS (Infrastructure AS A Servie)
- ✓ Web Services in Cloud

Enabling Technology: Virtualization



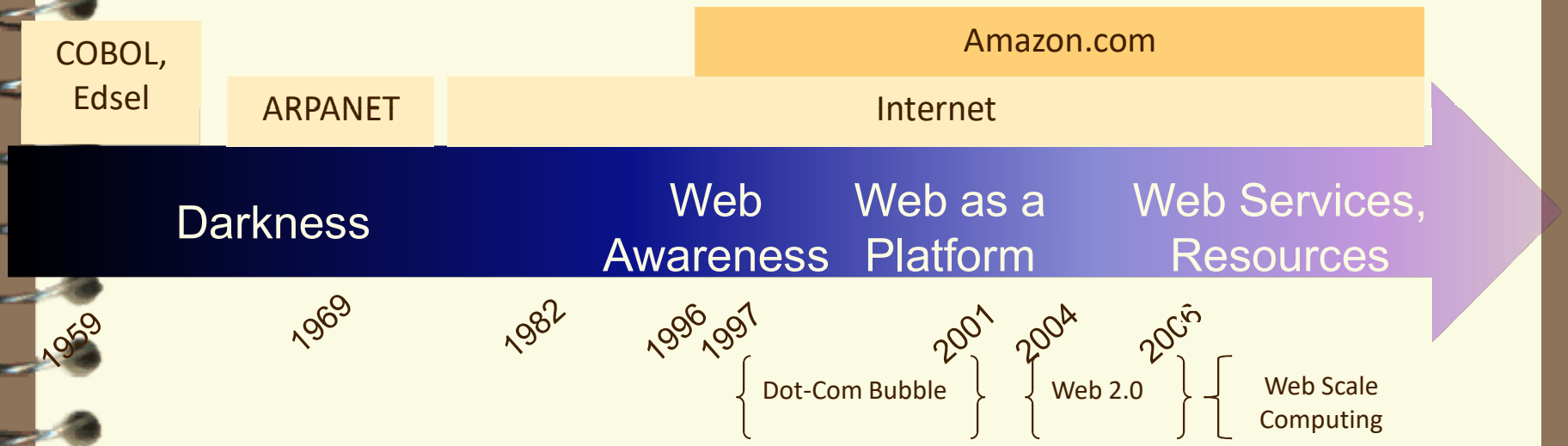
Everything as a Service

- Utility computing = Infrastructure as a Service (IaaS)
  - Why buy machines when you can rent cycles?
  - Examples: Amazon's EC2, Rackspace
- Platform as a Service (PaaS)
  - Give me nice API and take care of the maintenance, upgrades, …
  - Example: Google App Engine
- Software as a Service (SaaS)
  - Just run it for me!
  - Example: Gmail, Salesforce

Cloud versus cloud

- ✓ Amazon Elastic Compute Cloud
- ✓ Google App Engine
- ✓ Microsoft Azure
- ✓ GoGrid
- ✓ AppNexus

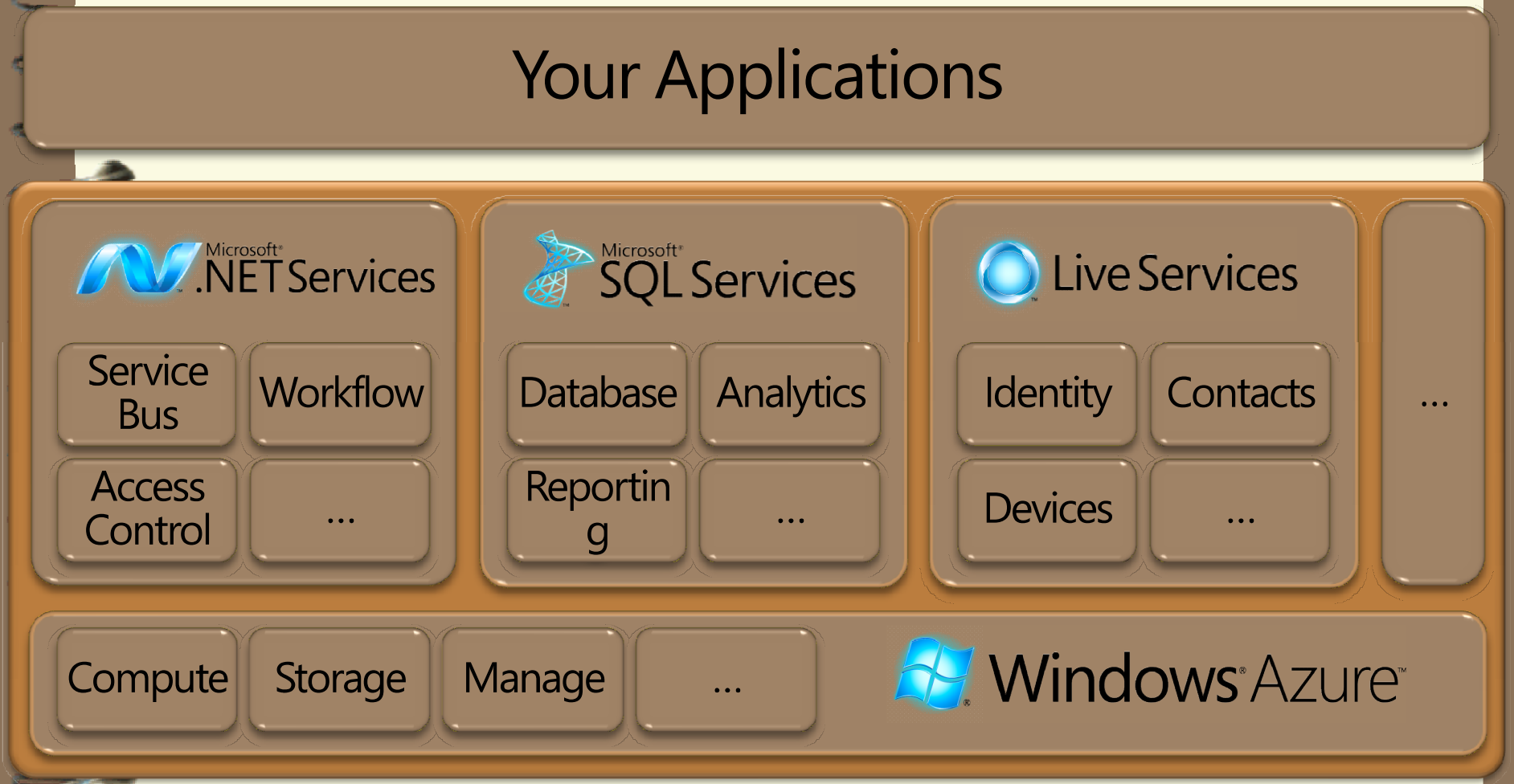
The Obligatory Timeline Slide
(Mike Culver @ AWS)



AWS

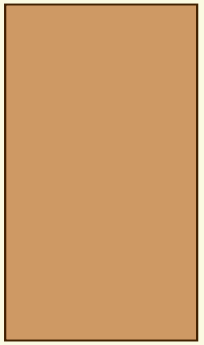
- ✓ Elastic Compute Cloud – EC2 (IaaS)
- ✓ Simple Storage Service – S3 (IaaS)
- ✓ Elastic Block Storage – EBS (IaaS)
- ✓ SimpleDB (SDB) (PaaS)
- ✓ Simple Queue Service – SQS (PaaS)
- ✓ CloudFront (S3 based Content Delivery Network – PaaS)
- ✓ Consistent AWS Web Services API

What does Azure platform offer to developers?



Google's AppEngine vs Amazon's EC2

Python
BigTable
Other API's



VMs
Flat File Storage

AppEngine:

- ✓ Higher-level functionality (e.g., automatic scaling)
- ✓ More restrictive (e.g., respond to URL only)
- ✓ Proprietary lock-in

EC2/S3:

- ✓ Lower-level functionality
- ✓ More flexible
- ✓ Coarser billing model

Why do we care about big data?

Example: Medicare

- ✓ IBM: Predict Heart Disease Through Big Data Analytics
  - traditional: EKGs, heart rate, blood pressure
  - big data analysis: connecting
    - exercise and fitness tests:
    - diet
    - fat and muscle composition
    - genetics and environment
    - social media and wellness: share information
    - …
- ✓ Google Flu Trends:
  - advance indication in the 2007-08 flu season
  - the 2009 H1N1 outbreak

Nature, 2009

A new game: large number of data sources of big volume

Big data is needed everywhere

- ✓ Social media marketing:
  - 78% of consumers trust peer (friend, colleague and family member) recommendations – only 14% trust ad
  - if three close friends of person X like items P and W, and if X also likes P, then the chances are that X likes W too
- ✓ Social event monitoring:
  - Prevent terrorist attack
  - The Net Project, Shenzhen, China (Audaque)
- ✓ Scientific research:
  - A new yet more effective way to develop theory, by exploring and discovering correlations of seemingly disconnected factors

The world is becoming data-driven, like it or not!

The big data market is BIG

- ✓ US HEALTH CARE $300 B
  Increase industry value per year by $300 B
- ✓ US RETAIL 60+%
  Increase net margin by 60+%
- ✓ MANUFACTURING –50%
  Decrease development and assembly costs by 50%
- ✓ GLOBAL PERSONAL LOCATION DATA $100 B
  Increase service provider revenue by $100 B
- ✓ EUROPE PUBLIC SECTOR ADMIN 250 B Euro
  Increase industry value per year by 250 B Euro

McKinsey Global Institute

Big Data: The frontier for innovation, competition and productivity

Why study big data?

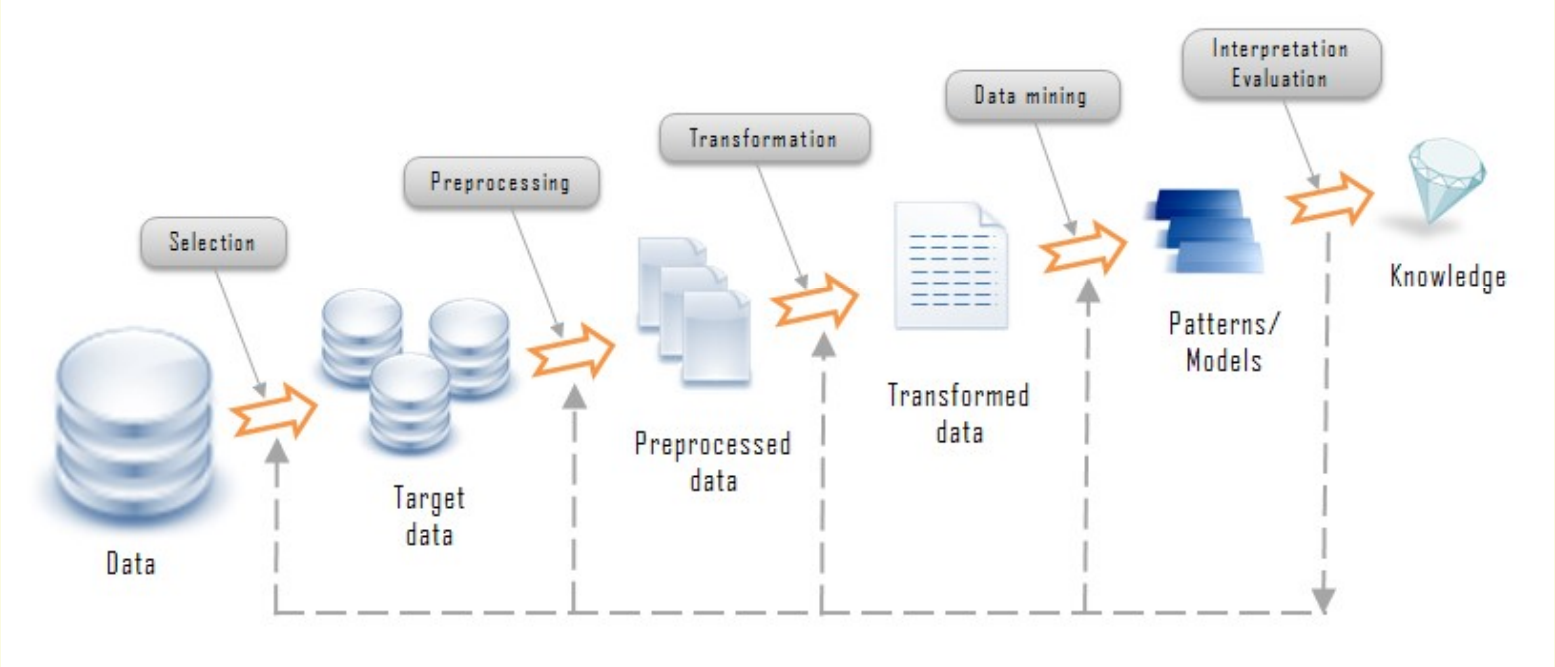
- ✓ Want to find a job?
  - Research and development of big data systems:
    ETL, distributed systems (eg, Hadoop), visualization tools, data warehouse, OLAP, data integration, data quality control, …
  - Big data applications:
    social marketing, healthcare, …
  - Data analysis: to get values out of big data
    discovering and applying patterns, predictive analysis, business inte
- ✓ Prepare you for
  - graduate study: current research and practical issues;
  - the job market: skills/knowledge in need

complexity theory, distributed databases, query answering, algorithms, data quality

Big data = Big $$$

What does this course cover?

A process of knowledge discovery



Data models, storage and management

Topic 1:
Data models, storage and management

Relational data models and DBMS:

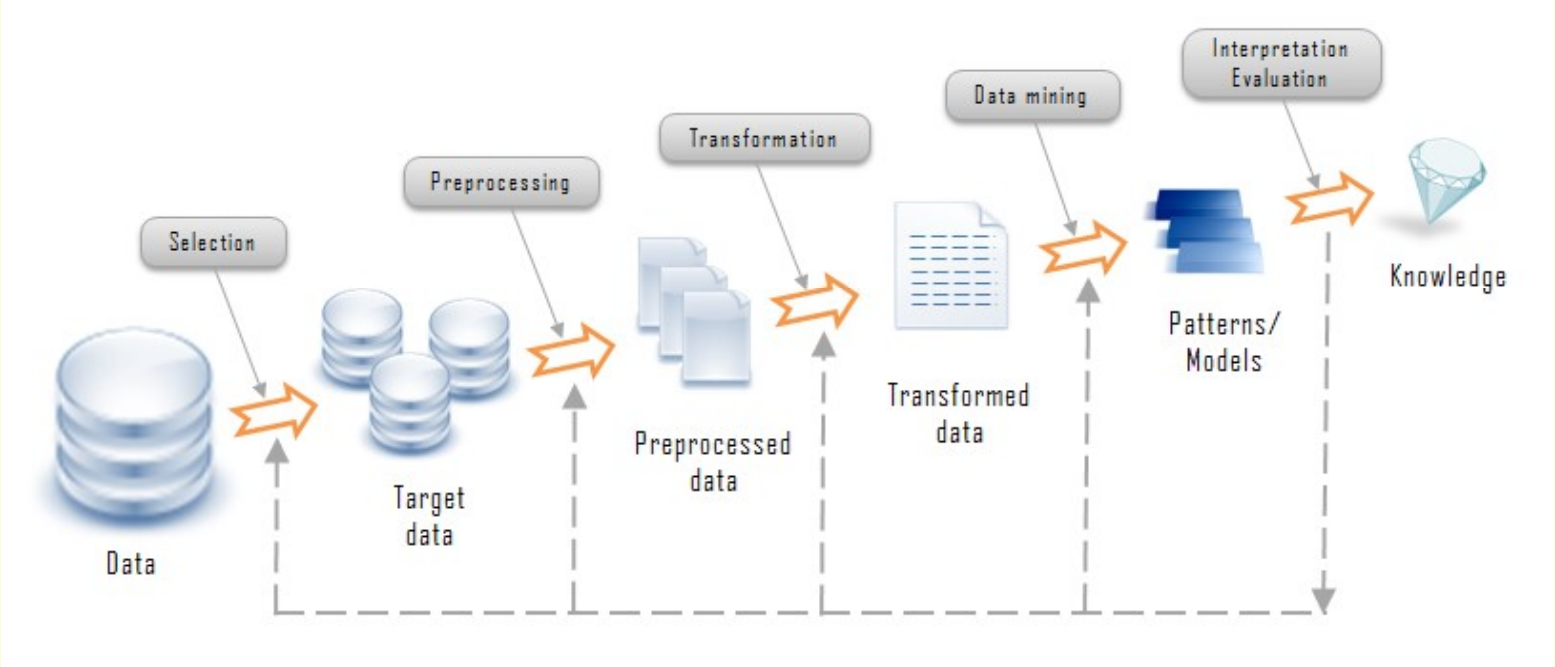
- ✓ Relational data and relation algebra
- ✓ DBMS: centralized; single processor (CPU)
- ✓ Relational databases

Challenge 1: How to represent and store Big data?

Beyond Relational databases

- ✓ Non-relational data, semi-structured data
- ✓ noSQLs, newSQLs, Key-value stores, column-stores, document stores…
- ✓ Graph data and graph databases

A process of knowledge discovery



Data models, storage and Management

Data analytics (search, mining and learning)

Topic 2:
Search Big Data

- ✓ Popular query languages
  - SQL fundamentals
  - XML, XQuery and SPARQL

Challenge 2: How to find needle in the Big Data haystack?

- ✓ Big data search algorithms: Design principles and Case study
  - Indexing and Views
  - Exact vs. Approximate search
  - Compression and summarization
  - Resource bounded search
  - Cope with data streams

Big data: Through the eyes of computation

- ✓ Computer science is the topic about

the computation of function f(x)

- ✓ Big data: the data parameter *x* is horrendously large: PB or EB

What is the challenge introduced to query answering?

Fallacies:

- ✓ *Big data introduces no fundamental problems*
- ✓ *Big data = MapReduce (Hadoop)*
- ✓ *Big data = data quantity (scalability)*

Are these true?

Topic 3:
Parallel/Distributed systems

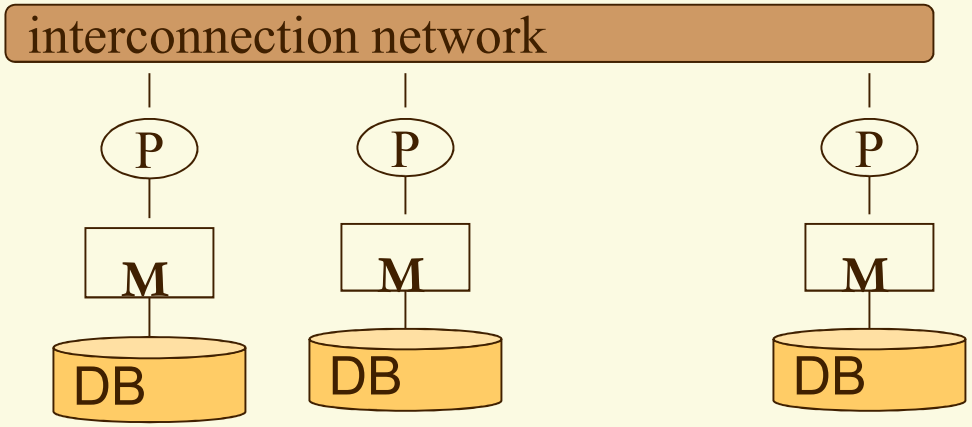
Recall traditional DBMS:

- ✓ Database: “single” memory, disk
- ✓ DBMS: centralized; single processor (CPU);

Can we do better provided with multiple processors?

Parallel DBMS: exploring parallelism

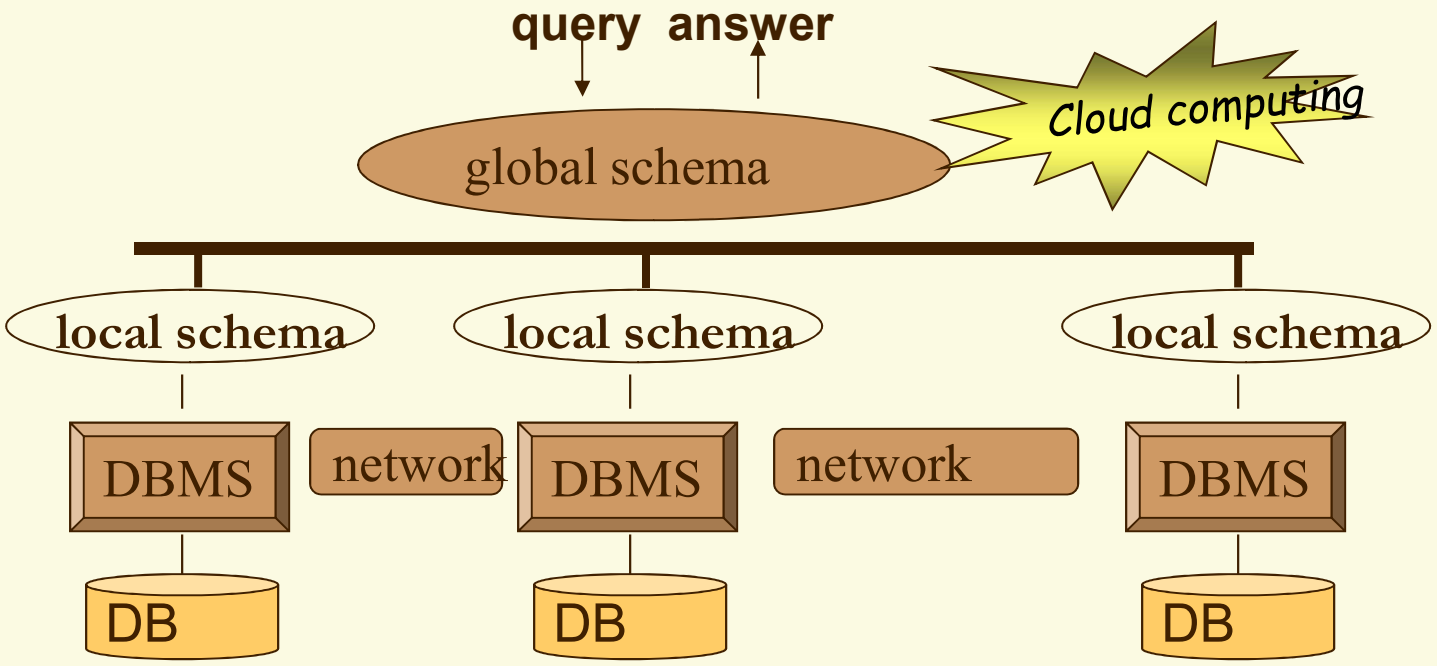
- ✓ Improve performance
- ✓ Reliability and availability



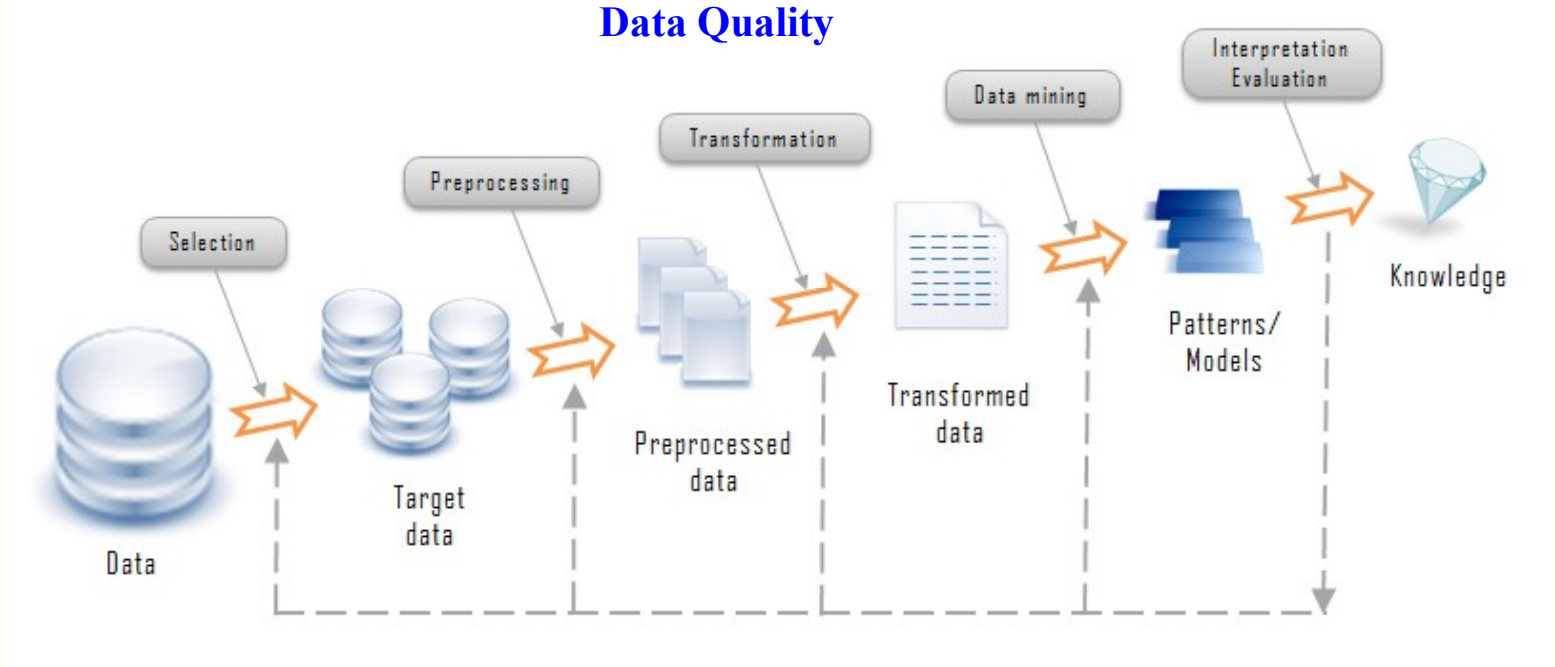
Distributed databases

Data is stored in several sites, each with an independent DBMS

- ✓ Local ownership: physically stored across different sites
- ✓ Increased availability and reliability
- ✓ Performance



A process of knowledge discovery



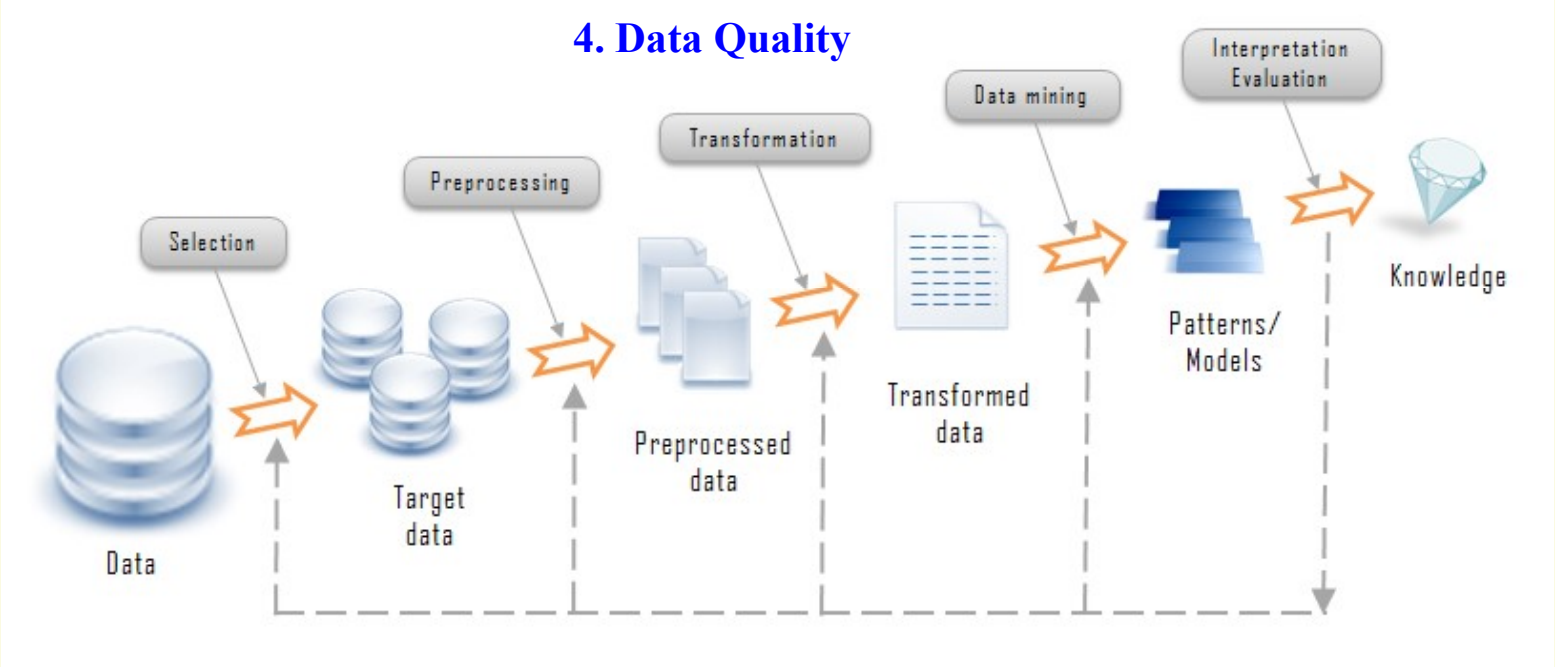
Data models, storage and Management

Data analytics (search, mining and learning)

Topic 4 & 5: data quality, security and ethics

- ✓ Data quality: Cleaning big data: error detection, data repairing, certain fixes (veracity)
- ✓ Privacy and Security (veracity)
- ✓ Data visualization

Putting together



1. Data models, storage and Management

2. Data analytics (search, mining and learning)

3. Distributed/Parallel data analysis

5. Privacy & ethics