Introduction to Big Data

Thoai Nam

Slides

- Big data, Donald Kossmann & Nesime Tatbul, ETH Zurich
- Introduction to Big Data, Ruoming Jin, Kent State University

What's Big Data?

"**Massive** amounts of **diverse**, **unstructured** data produced by **high-performance** applications."

"Data **too large & complex** to be effectively handled by standard database technologies currently founded in most organizations

12+ TBs
of tweet data every day

? TBs of data every day

25+ TBs **of log data every day**

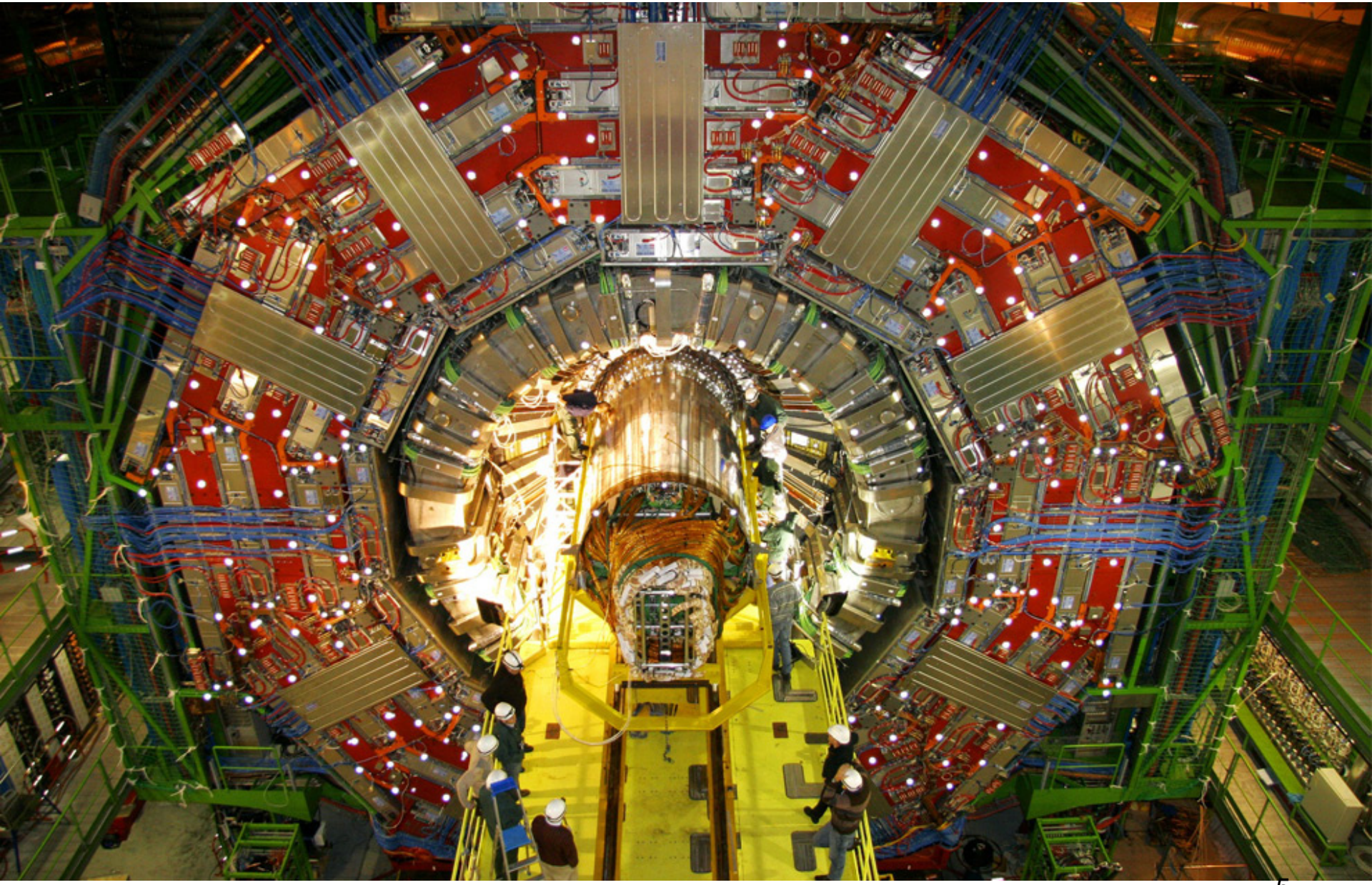
30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

2+ billion people on the Web by end 2011

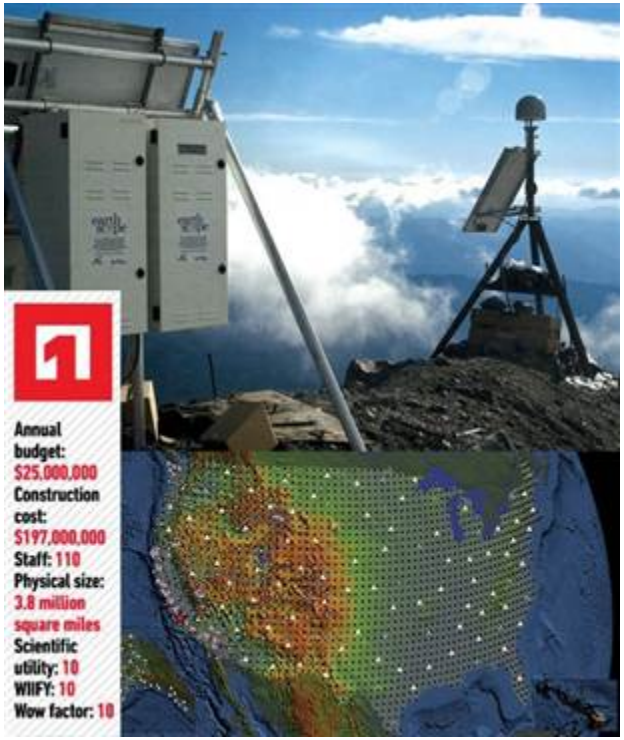
76 million smart meters in 2009… 200M by 2014



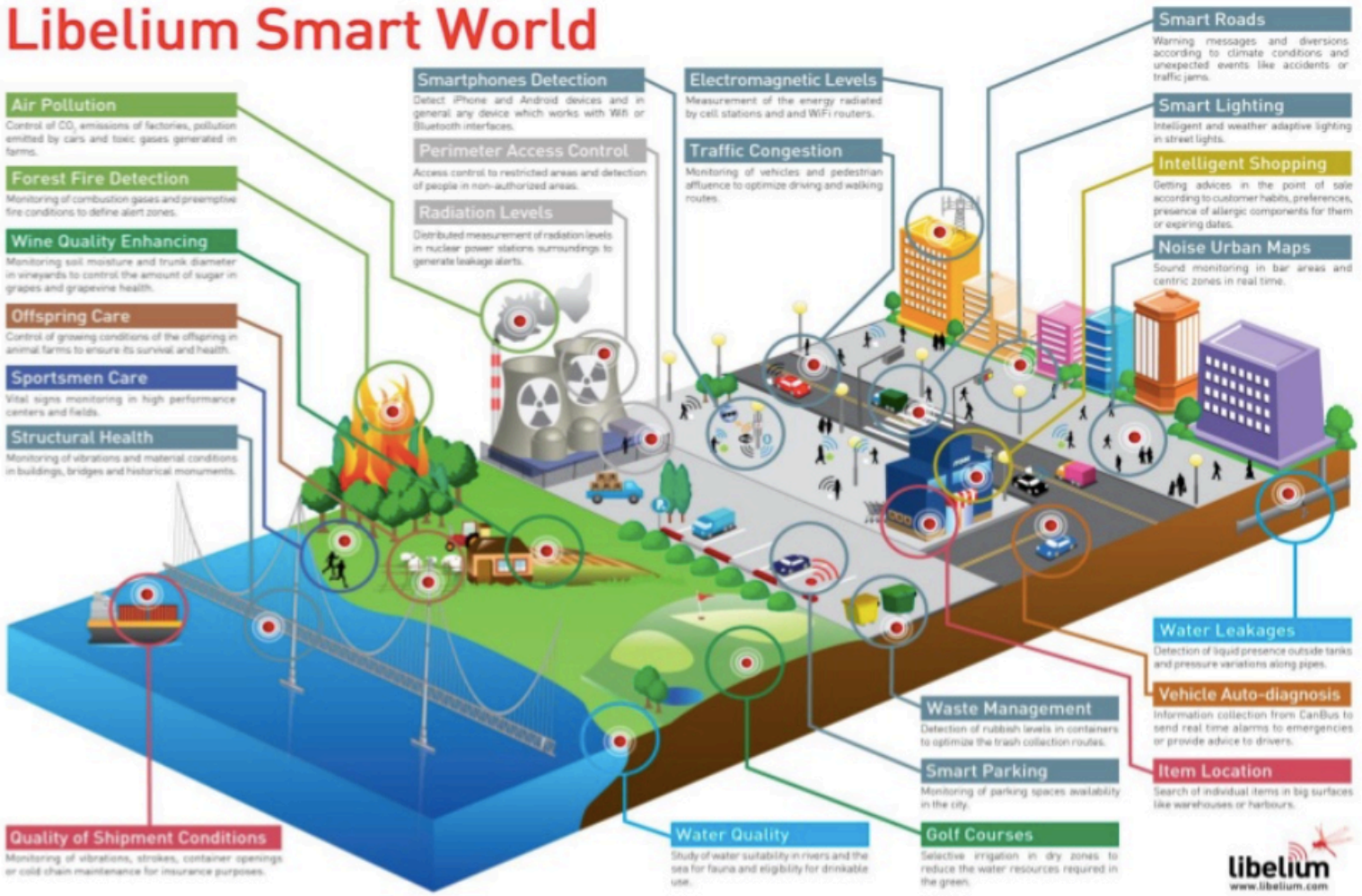
CERN’s Large Hydron Collider (LHC) generates 15 PB a year

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)



Applications

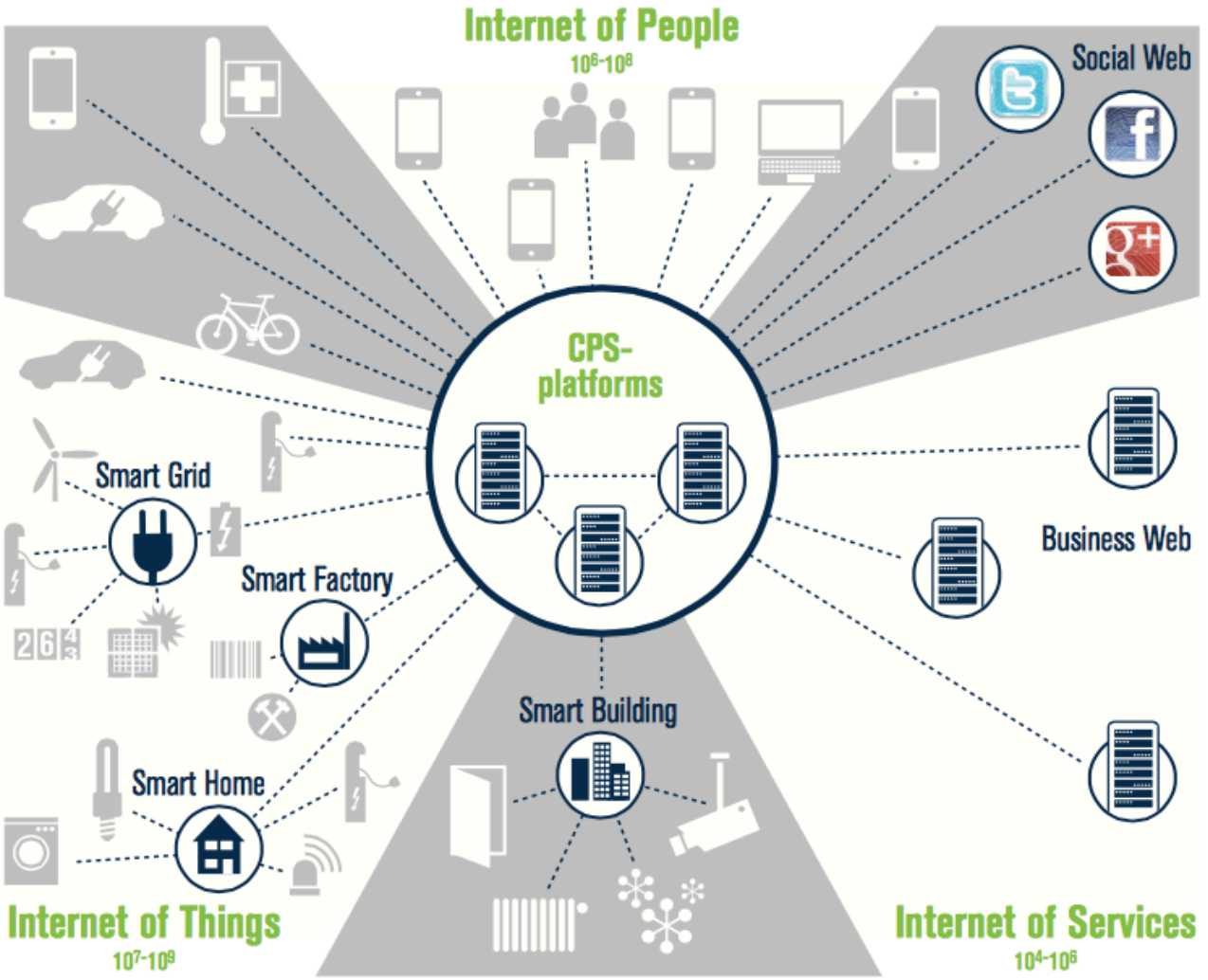


http://www.libelium.com/libelium-smart-world-infographic-smart-cities-internet-of-things/

Internet of Things (IoT)

IoT and Services

Figure 4:
The Internet of Things and Services – Networking people, objects and systems



Source: Bosch Software Innovations 2012

Internet of Things (IoT)

(Timothy Chou)

Do

Learn

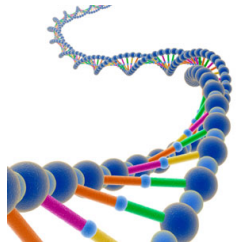
Collect

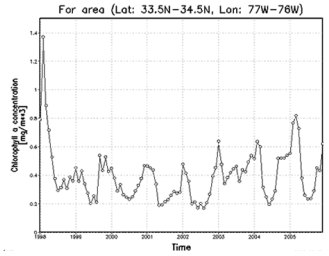
Connect

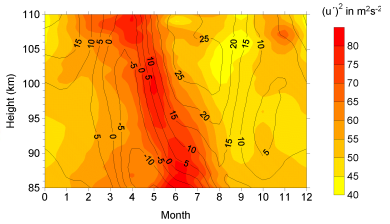
Things

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)

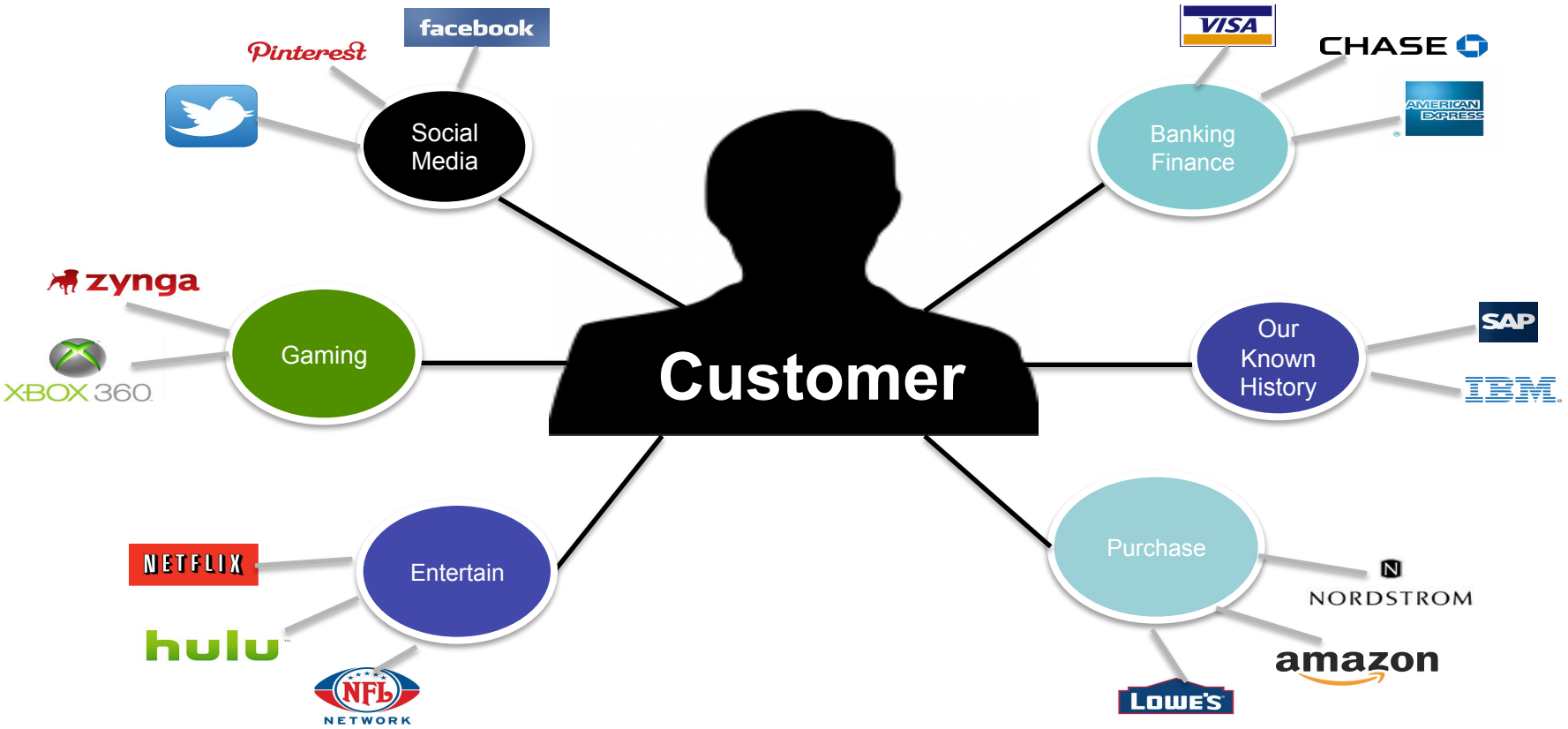






To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



a Single Customer View '***is an aggregated, consistent and holistic representation of the data known by an organisation about its customers***'

Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

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

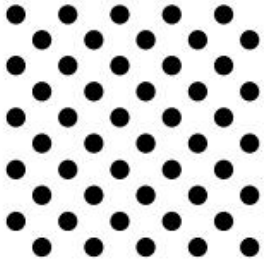
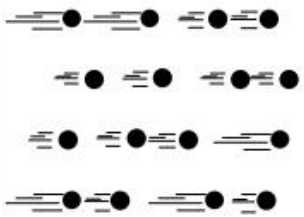
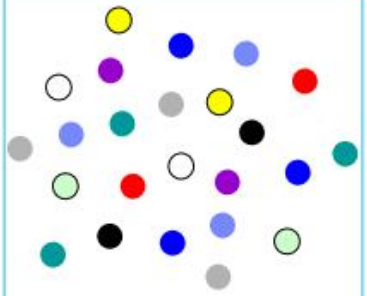
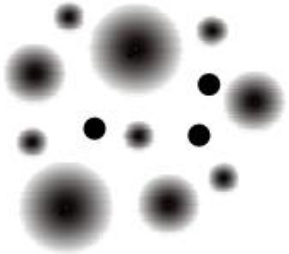
Scientific instruments
(collecting all sorts of data)

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

Some Make it 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

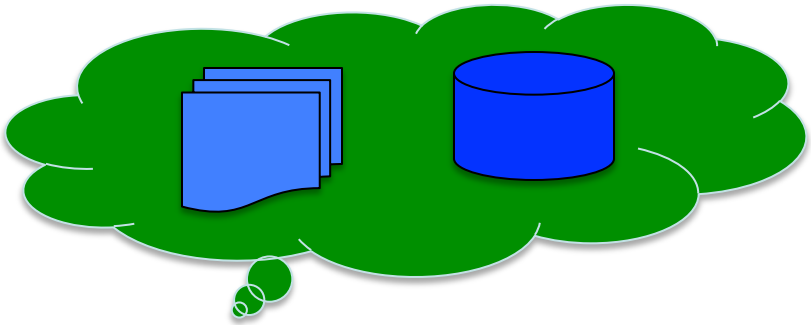
Answering tough questions

- **Problem**
  - sales for lollipops are going down
- **Data**
  - all sales data by customer, region, time, ...
- **Information**
  - lollipops bought by people older than 25
  - (but eaten by people younger than 10)
- **Knowledge**
  - moms believe: lollipops = bad teeth
- **Value**
  - dentists advertise your lollipops

Why is this difficult?

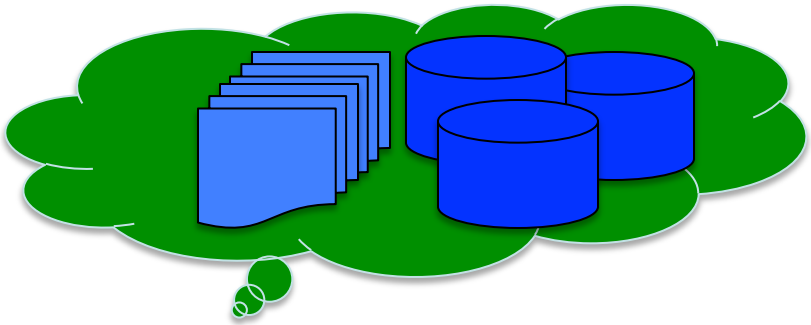
- You need more data than your data warehouse
  - you need more data that you have
  - logs, Twitter feeds, blogs, customer surveys, ...
- You need to ask the right questions
  - data alone is silent
- You need technology and organization that help you concentrate on asking the right questions

Why is this difficult? (1)

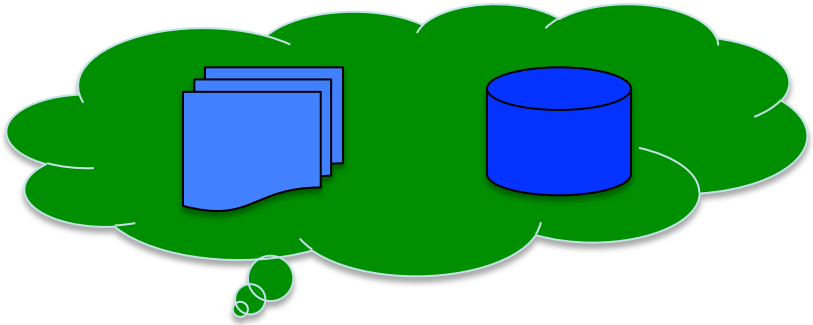


YOU! (TB)

Why is this difficult?(2)

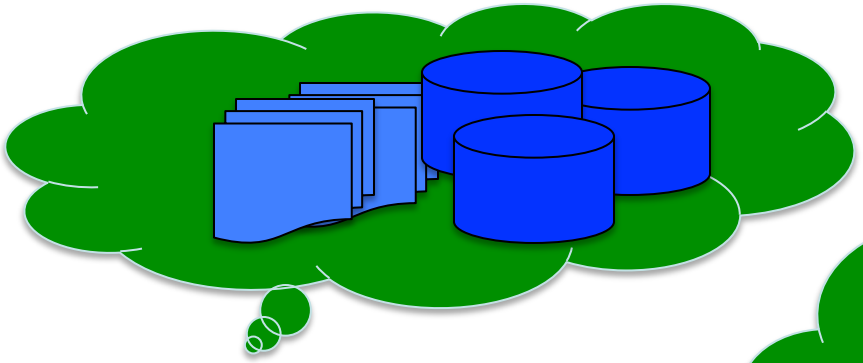


YOU! (PB)

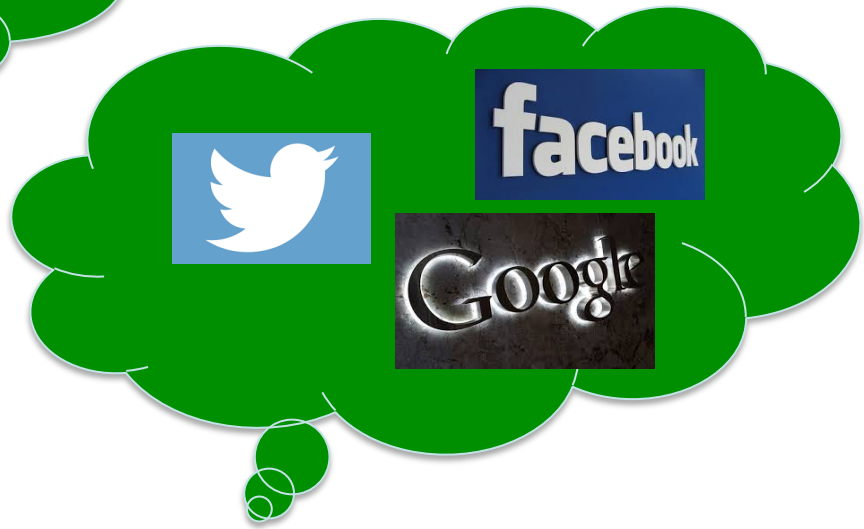


YOU FRIENDS! (TB)

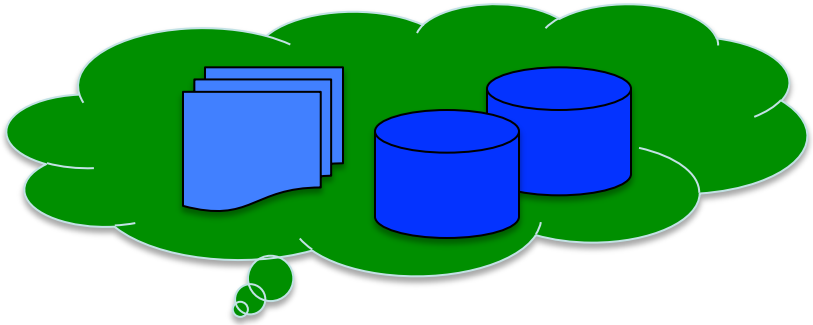
Why is this difficult?(3)



YOU! (PB)

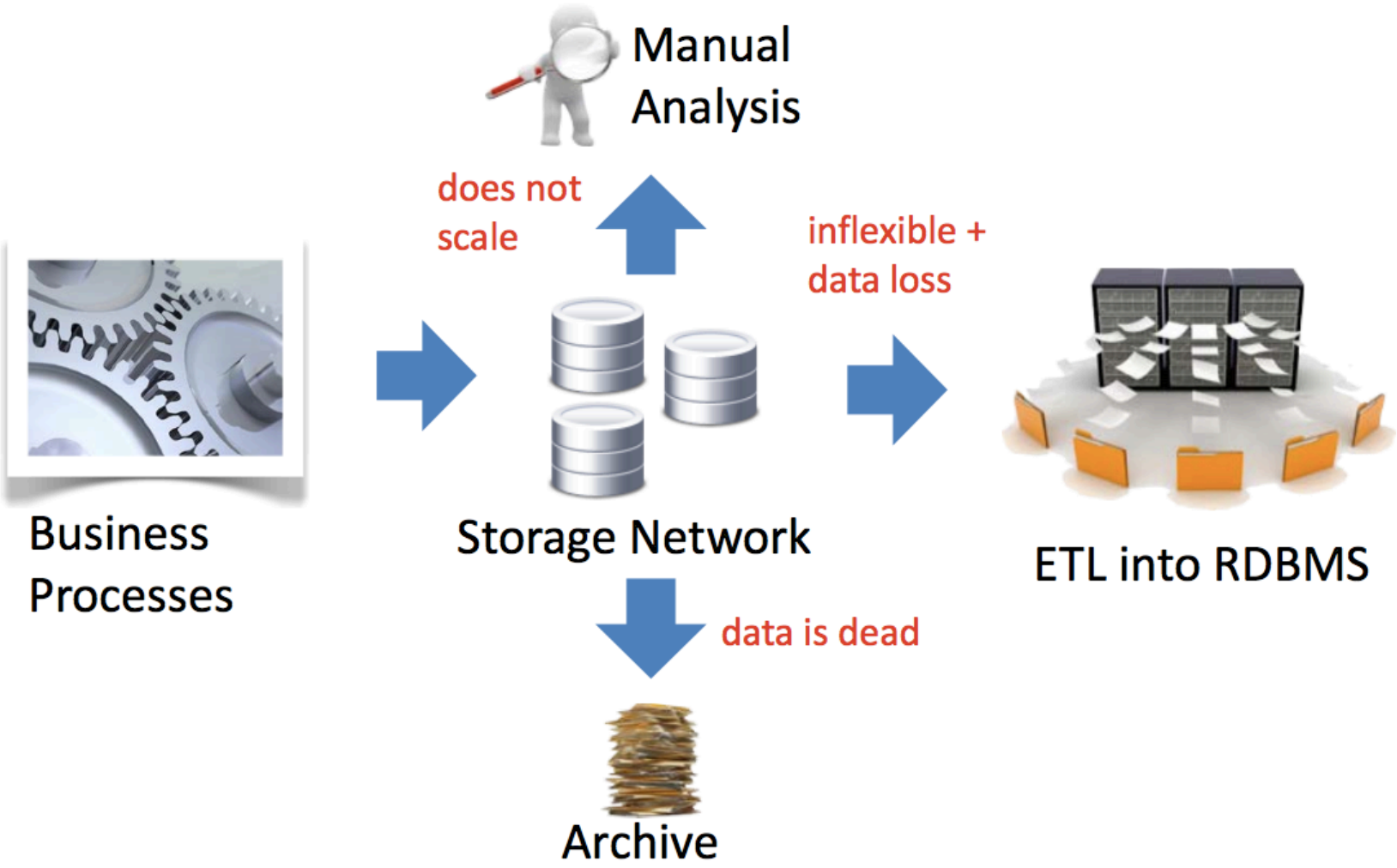


THE WORLD! (EB)



YOU FRIENDS! (TB/PB)

Limitations of state of the art



What is Big Data?

- Three alternative perspectives
  - Philosophical
  - Business
  - Technical
- (Ultimately, it is a buzz word for everybody.)

Philosophical

- What is more valuable, if you had to pick one?
  - experience or intelligence?
- Traditional (computer) science: **logic!** [intelligence]
  - understand the problem, build model / algorithm
  - answer question from implementation of model
- New science: **statistics!** [experience]
  - collect data
  - answer question from data (what did others do?)

Data Science, 4th Paradigm

- New approach to do science
  - Step 1: Collect data
  - Step 2: Generate Hypotheses
  - Step 3: Validate Hypoteheses
  - Step4: (Goto Step 1 or 2)
- Why is this a good approach?
  - it can be automated: no thinking, less error
- Why is this a bad approach?
  - how do you debug without a ground truth?

Is bigger = smarter?

- Yes!
  - tolerate errors
  - discover the long tail and corner cases
  - machine learning works much better
- But!
  - more data, more error (e.g., semantic heterogeneity)
  - with enough data you can prove anything
  - still need humans to ask right questions

Big Data Success Story

- Google Translate
  - you collect snippets of translations
  - you match sentences to snippets
  - you continuously debug your system
- Why does it work?
  - there are tons of snippets on the Web
  - there is a ground truth that helps to debug system

Big Data Farce (only a joke)

- Which lane is fastest in a traffic jam?
  - you ask people where they go and whether happy
  - (maybe, you even use a GPS device)
  - you conclude that left lane is fastest

Big Data Farce (only a joke)

- Which lane is fastest in a traffic jam?
  - you ask people where they go and whether happy
  - (maybe, you even use a GPS device)
  - you conclude that left lane is fastest
- Why is this stupid?
  - because there is no ground truth!
  - you will get a conclusion because Big Data always gives an answer. But, it does not make sense!
  - getting more data does not help either

Fundamental Problem of Big Data

- There is no ground truth
  - gets more complicated with self-fulfilling prophecies
- Hard to debug: takes human out of the loop
  - Example: How to play lottery in Napoli
    - Step 1: You visit “oracles” who predict numbers to play
    - Step 2: You visit “interpreters” who explain predictions
    - Step 3: After you lost, “analysts” tell you that “oracles” and “interpreters” were right and that it was your fault.

What is Big Data?

- Business Perspective
  - it is a new business model
- People pay with data
  - e.g. Facebook, Google, Twitter:
    - use service, give data
    - Google sells your data to advertisers • (you pay advertisers indirectly)
  - e.g, Amazone
    - pay service + give data
    - sells data and uses data to improve service

Business Perspective

- Bank
  - keeps your money securely (kind of...)
  - puts your money at work (lends it to others), interest
  - you keep ownership of data and take it when needed
- Databank
  - keeps your data securely (kind of...)
  - puts your data at work: interest or better service
  - (you keep ownership of data: hopefully to come)

Technical Perspective (?)

- You collect all data
  - the more the better -> statistical relevance, long tail
  - keeping all is cheaper than deciding what to keep
- You decide independently what to do with data
  - run experiments on data when question arises
- Huge difference to traditional information systems
  - design upfront what data to keep and why!!!
  - (e.g., waterfall model of software engineering!)

Consequences

- **Volume**: data at rest
  - it is going to be a lot of data
- **Speed**: data in motion
  - it is going to arrive fast
- **Diversity**: data in many formats
  - it is going to come in different shapes
  - (e.g., different versions, different sources)
- **Complexity**: You want to do something interesting
  - SQL will not be enough

Alternative Definition (Gartner, IBM)

- Volume: same as before
- Velocity: same as “speed”
- Variety: same as “diversity”
- **Veracity: data in doubt**
  - you do not know exactly what you have

Topics (1)

- Data Warehouses ***VOLUME***
  - The old world of Big Data
- Cloud Computing ***VOLUME***
  - The infrastructure to collect and process Big Data
- Map Reduce ***COMPLEXITY***
  - The new world of Big Data (programming model)

Topics (2)

- Semi-structured Data ***DIVERSITY***
  - The new world of Big Data (data model) – “Collect first – think later”
- Streaming Data ***SPEED***
  - Making Big Data fast
- Other Topics
  - visualization, data cleaning, security, crowd-sourcing, …
- Applications

Why now?

- Mega-trend: All data is digital, digitally born!
  - 70 years ago: computers for “+”
  - 15 years ago: disks cheaper than paper
  - 7 years ago: Internet has eyes and ears
- Because we can
  - 40 years of databases -> volume
  - 40 years of Moore’s law -> complexity
  - 2000+ years of statistics -> it is only counting
  - enough optimisms that we get the rest done, too
- Because we reached dead end with logic (?)

Because we can... Really?

- Yes!
  - all data is digitally born
  - storage capacity is increasing
  - counting is embarrassingly parallel

Because we can... Really?

- Yes!
  - all data is digitally born
  - storage capacity is increasing
  - counting is embarrassingly parallel
- But,
  - data grows faster than energy on chip
  - value / cost tradeoff unknown
  - ownership of data unclear (aggregate vs. individual)
- I believe that all these “but’s” can be addressed