DATA

Kamakshaiah Musunuru

Some important questions about data

Some important data repositories

What is data?

categories

Applications

What facebook does?

How do business use data?

What exactly you can do with data?

Pros & Cons

Security & Maintenance
- Cost of data breach!
- Reasons for data breach
- maintenance

How many times will you think/do with numbers every day after you wake-up from bed?

- Do you know how many civilians or military personnel died in yesterday's Israel's attack on Gaza?
- Do you know how much grain and food pulses are being wasted every year by Indian Government as matter of food security policy?
- Do you know under-ground water depletion levels very beneath the soil you live?
- Do you know the average melting level of ice caps at poles?
- Can you forecast the rising inflation and can you assess its impact on next year surplus?

- Do you know average quantity of shampoo that you and your family use every month?
- Have you ever reckoned or maintained an account of *the average chicken, meat, eggs, cooking oil, pulses, rice and other stationary* that your family consume?
- Do you know the burden of insolvency that your family suffer due to rising inflation in the country?
  Do you have any tangible solution ready made in hand?
- How many times you take out your wallet every day for buying something?
- Do you ever reckon amount of water and energy misused at you home and have you ever offered a controlling mechanism?

- How many headings (sub-headings) in any newspaper is managed by both text and numbers?

- How many headings (sub-headings) in any newspaper is managed by both text and numbers?
- Did you ever think why do print publishing need three different types of datum? i.e. textual, tabular and iconic/graph.

- How many headings (sub-headings) in any newspaper is managed by both text and numbers?
- Did you ever think why do print publishing need three different types of datum? i.e. textual, tabular and iconic/graph.
- Did you ever notice impact of sound and subsequent film in any commercial at your home?

Finally ....

Finally .... How many of you have the habit of obtaining, storing, sharing and using data for decisions?

Finally .... How many of you have the habit of obtaining, storing, sharing and using data for decisions?

If, so...

Finally .... How many of you have the habit of obtaining, storing, sharing and using data for decisions?

If, so... Where do you browse for data?

Finally .... How many of you have the habit of obtaining, storing, sharing and using data for decisions?

If, so... Where do you browse for data?

Assignment

Figure: world bank data repository

Figure: UNICEF

What is Data?

What is the difficulty with this term?

What is Data?

What is the difficulty with this term?

- In general, factual information, especially information organized for analysis or used to reason or make decisions.
- In Computer Sciences numerical or any other information represented in a form suitable for processing by computer.
- Values derived from scientific experiments.
- Plural of datum.
- In Communications & Information a series of observations, measurements, or facts; information and etc.

Types

Categories

- Basically
  - Textual& visual
  - Numerical & Non-numerical
  - Categorical & Non-categorical
- Advanced

Types

Categories

- Basically
  - Textual& visual
  - Numerical & Non-numerical
  - Categorical & Non-categorical
- Advanced
  - Primary & Secondary
  - Qualitative & Quantitative

Types

Categories

- Basically
  - Textual& visual
  - Numerical & Non-numerical
  - Categorical & Non-categorical
- Advanced
  - Primary & Secondary
  - Qualitative & Quantitative
- Research
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Applications

Do you know what astrophysicists are doing?

Do you know what astrophysicists are doing?

When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.

Figure: SDSS telescope in Mexico

Figure: Large Synaptic Survey Telescope

Figure: Large Synaptic Survey Telescope

What do you understand from these pictures?

What do you understand from these pictures?
Do you think I am talking about Cosmology or Astrology?

Is this limited only to ET?

Is this limited only to ET?

Wal-Mart, a retail giant, handles more than $1m$ customer transactions every hour, feeding databases estimated at more than 2.5 petabytes the equivalent of 167 times the books in America's Library of Congress (see article for an explanation of how data are quantified).

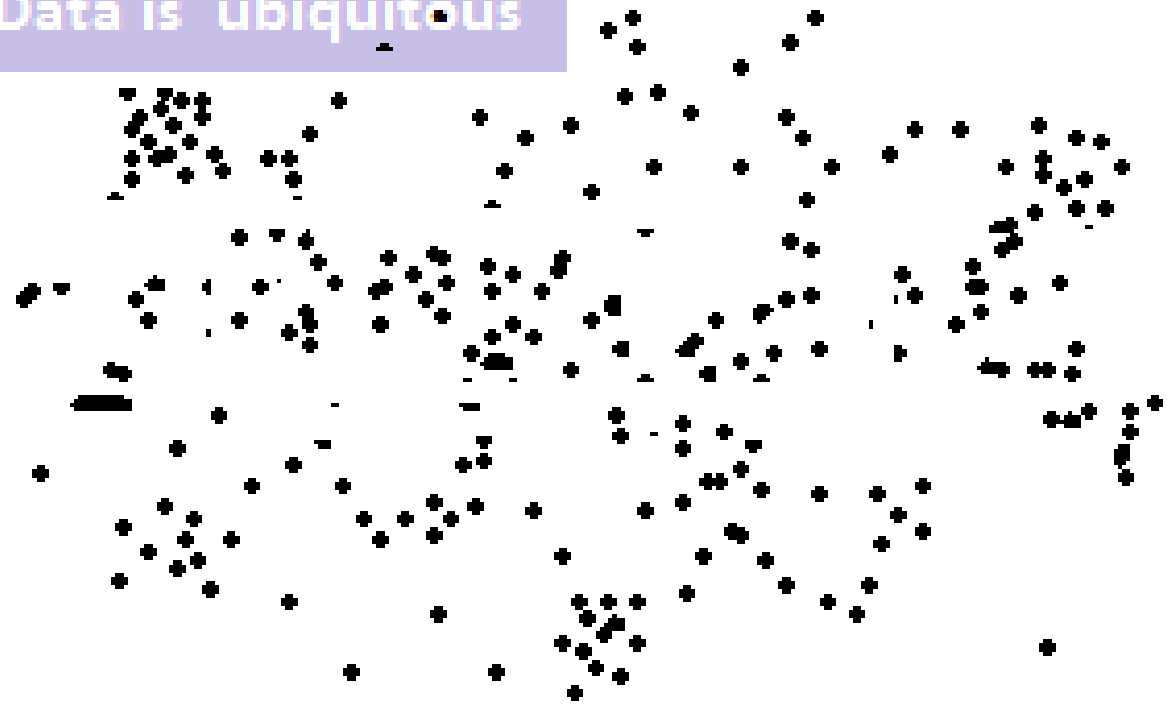
What's happening in Facebook?

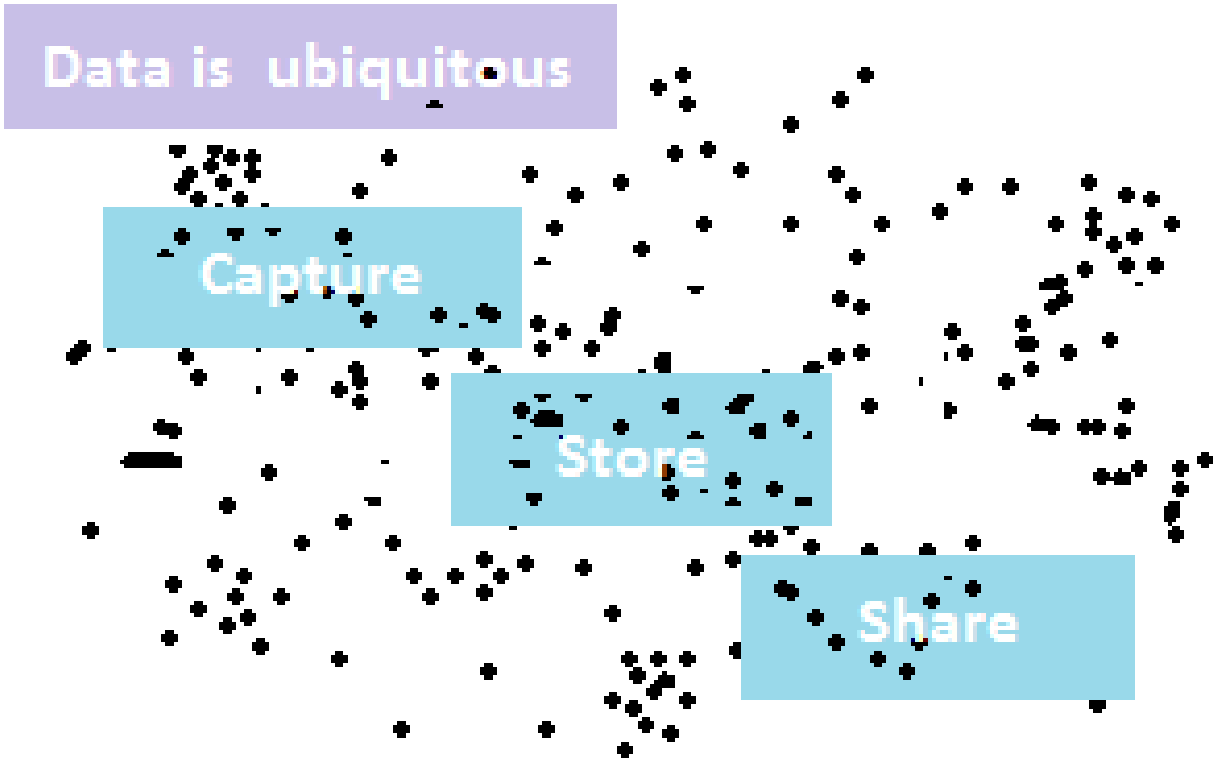
What's happening in Facebook?

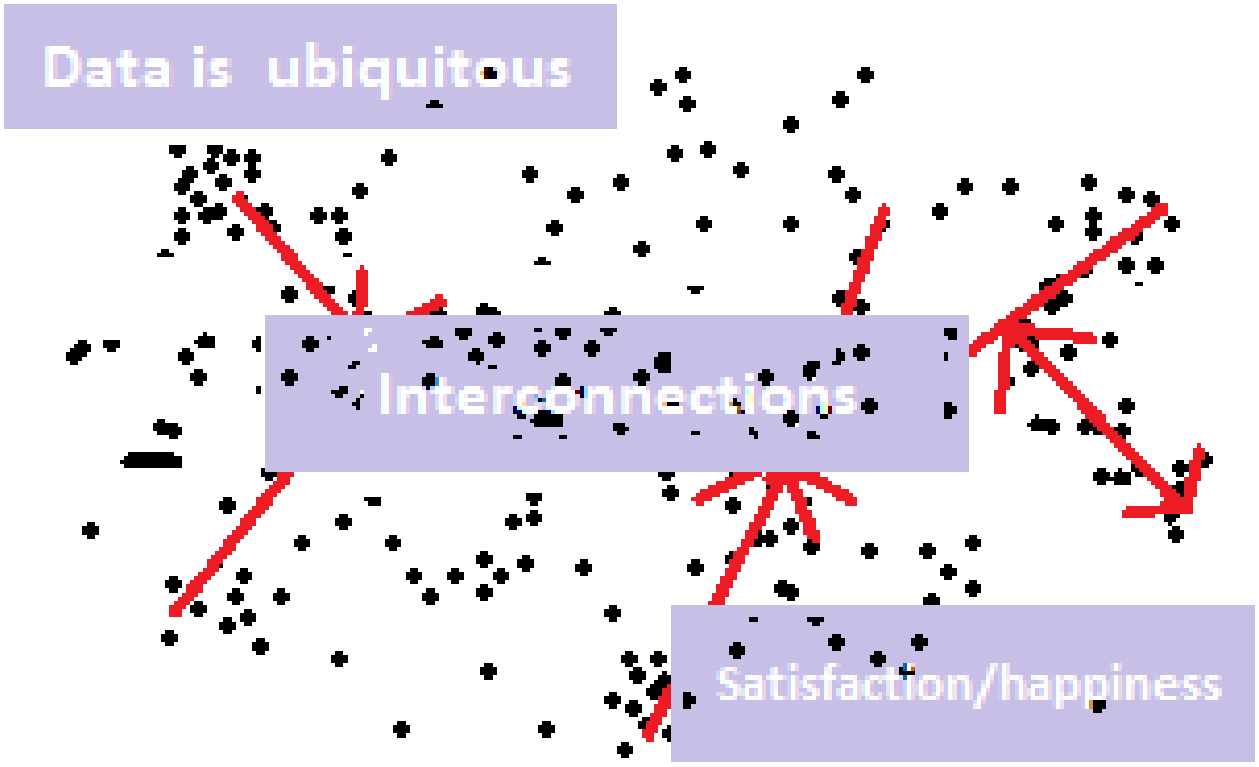
Facebook, a social-networking website, is home to 40 billion photos.

What makes Social Networking Successful?

Data is ubiquitous




Data is ubiquitous
Capture
Store
Share


Data is ubiquitous
Interconnections
Satisfaction/happiness

Hence, success is capturing, storing, sharing finally *consuming* meaningful data.

$$Success = Capture + Store + Share + Consume$$

- Microsoft's search engine *Bing*, can advise customers whether to buy an airline ticket now or wait for the price to come down by examining 225 billion flight and price records.
- *Makemytrip* capture individuals computer IP data and greets then with customized flight schedules
- Customer preferences are recalled by systems at hotels and greets them with name and profession. The hotel stay is customized by data capture and manipulation
- Banks use customer data to cousel and customize financial services (HDFC)
- Match-fixing is a matter of historical success and data analysis.

Human Genome!

Human Genome!

Decoding the human genome involves **analyzing** 3 billion base pairs which took ten years the first time it was done, in 2003, but can now be achieved in one week.

The importance of data is much more than said!

The importance of data is much more than said!

Well! What are they?

1. spot business trends
2. help prevent diseases combat crime
3. help weather forecasting
4. predict future

1. spot business trends
2. help prevent diseases combat crime
3. help weather forecasting
4. predict future finally

1. spot business trends
2. help prevent diseases combat crime
3. help weather forecasting
4. predict future finally
5. helps business managers, economists, scientist and etc. to understand, take decisions and changes human survival and standards of living

Pros & Cons

Storage

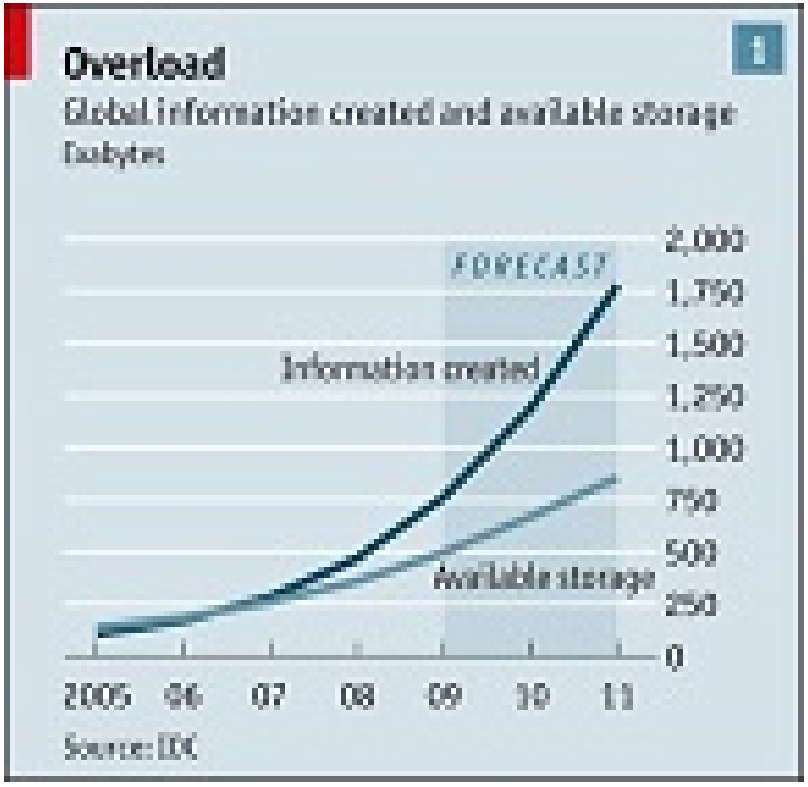


Figure: Storage problem

1. In recent years *Oracle, IBM ,Microsoft* and *SAP* between them have spent more than 15 billion on buying software firms specializing in data management and analytics.
2. This industry is estimated to be worth more than 100 billion and growing at almost 10% a year, roughly twice as fast as the software business as a whole.
3. Chief information officers (CIOs) have become somewhat more prominent in the executive suite, and a new kind of professional has emerged, the data scientist, who combines the skills of
software programmer, statistician and story teller/artist to extract the nuggets of gold hidden under mountains of data.

Hal Varian, Google's chief economist, predicts that the job of statistician will become the sexiest around.

1. Digital information is rising 10 fold every five years.
2. The computing speed and storage capacity doubling while cost decreases
   -Moore's Law

Edward Felten at Princeton University attributes this to improvements due to algorithms

Some Comments

1. "What we are seeing is the ability to have economies form around the data and that to me is the big change at a societal and even macroeconomic level", says *Craig Mundie of Microsoft*

2. "Every day I wake up and ask, how can I flow data better, manage data better, analyze data better?" says *Rollin Ford, the CIO of WalMart.*

Security & Maintenance

From Ponemons IBM-sponsored research

- It cost U.S. companies hit by data breaches last year an average of $5.4 million to cope with the after-effects up 9% from the year before.
- On average, it cost $201 per record lost, up from the $188 the year before.
- Heavily regulated industries such as healthcare, transportation, energy, financial services, communications, pharmaceuticals and manufacturing tend to have a higher per capita breach cost.
- Healthcare in general is believed to have faced the highest per-capita cost per industry at $359 and the public sector the lowest at $100.

What are the reasons for this?

Malicious and criminal attacks are cited most frequently as the root cause for data breaches globally, comprising 42% of incidents, while 30% were blamed on a negligent employee or contractor, and 29% on system glitches related to both *technology and business process* failures.

- The maintenance is another problem with data. One manager of world renowned e-retailing company boasts that "a 15-megawatt data center can use up to $360,000$ gallons of water a day." For instance; look at here
- Google had to obtain patent for its floating platform-mounted computer data center. For more details you may look at here

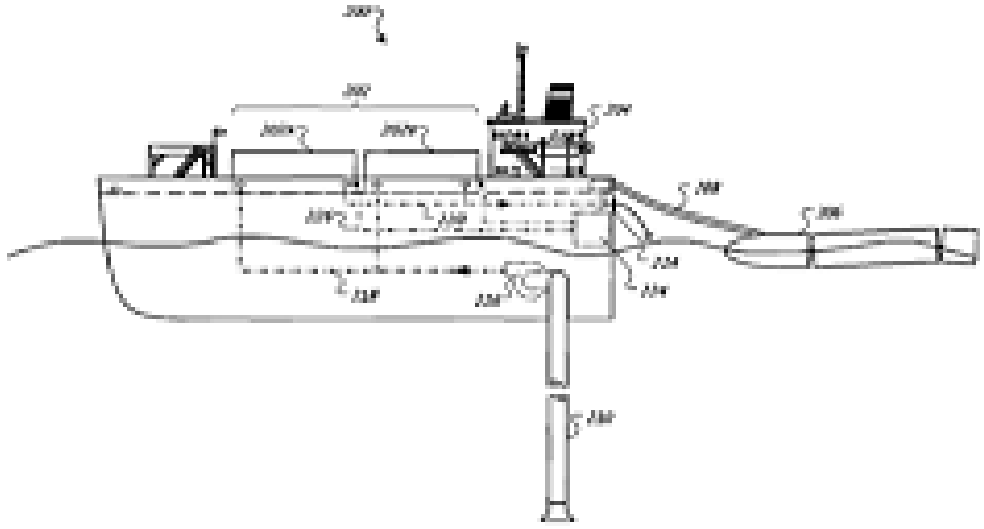


FIG. 2

This is the side view of the data center system.

Figure: Google's floating data center

Figure: Google's floating data center