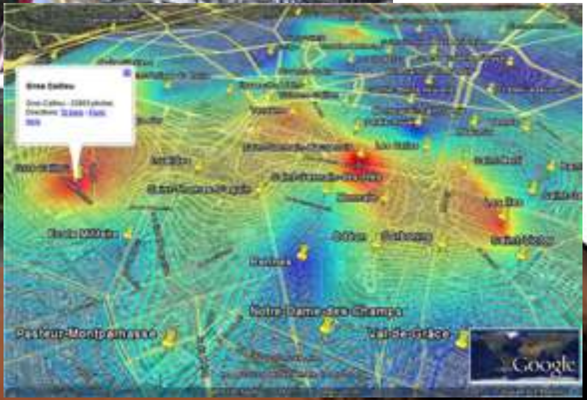
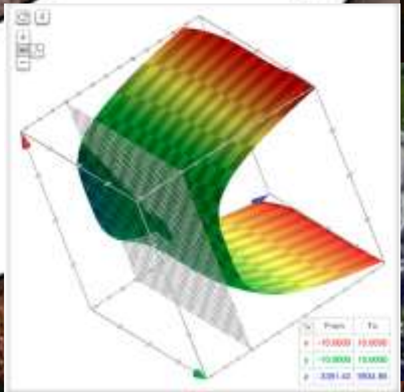
Introduction to Big Data

Srinath Perera





Outline

- Forecasts and Why Big data?
- What is Big Data?
- Parts of the Puzzle
- Quick Tutorial
- Conclusion

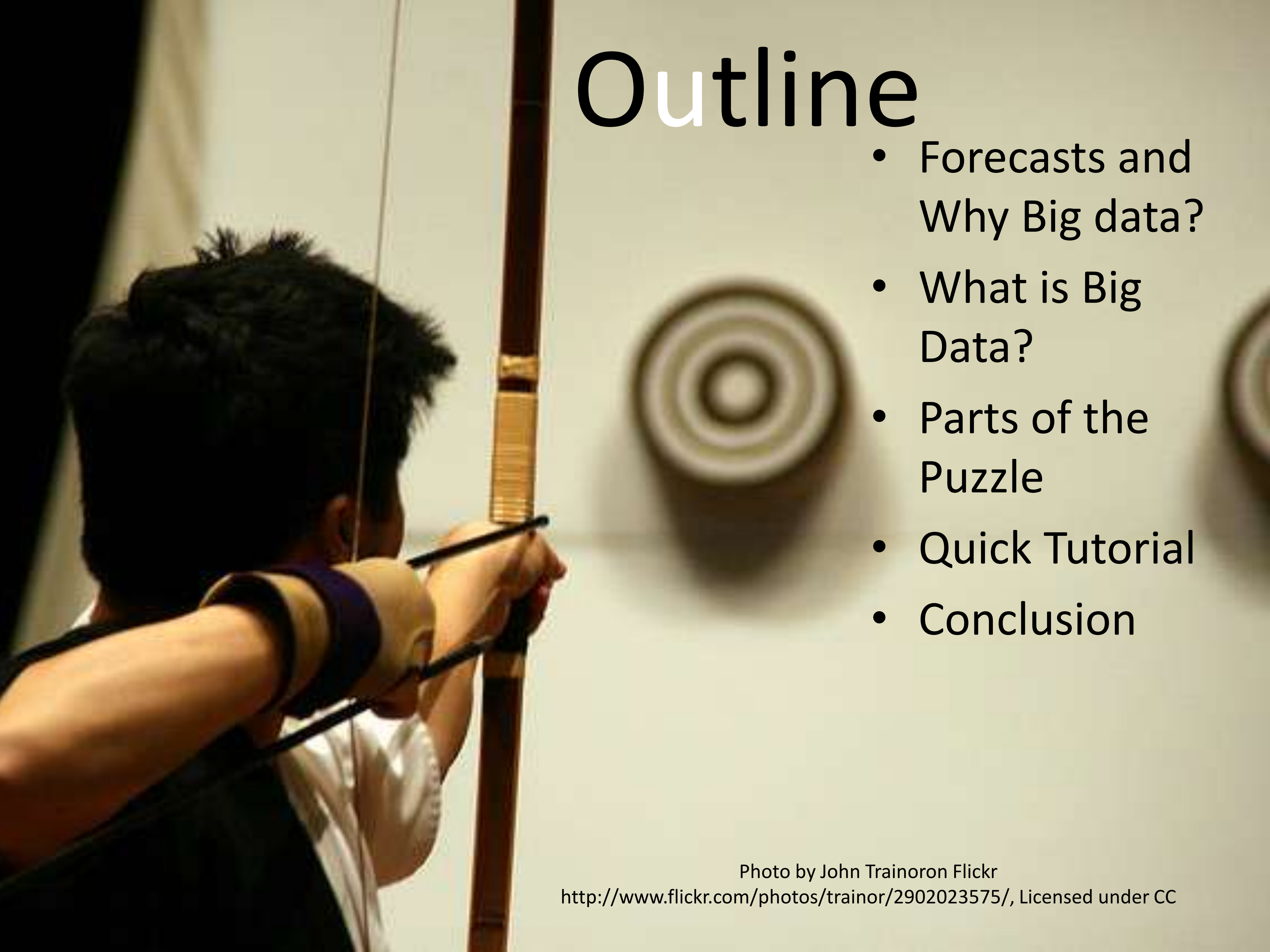
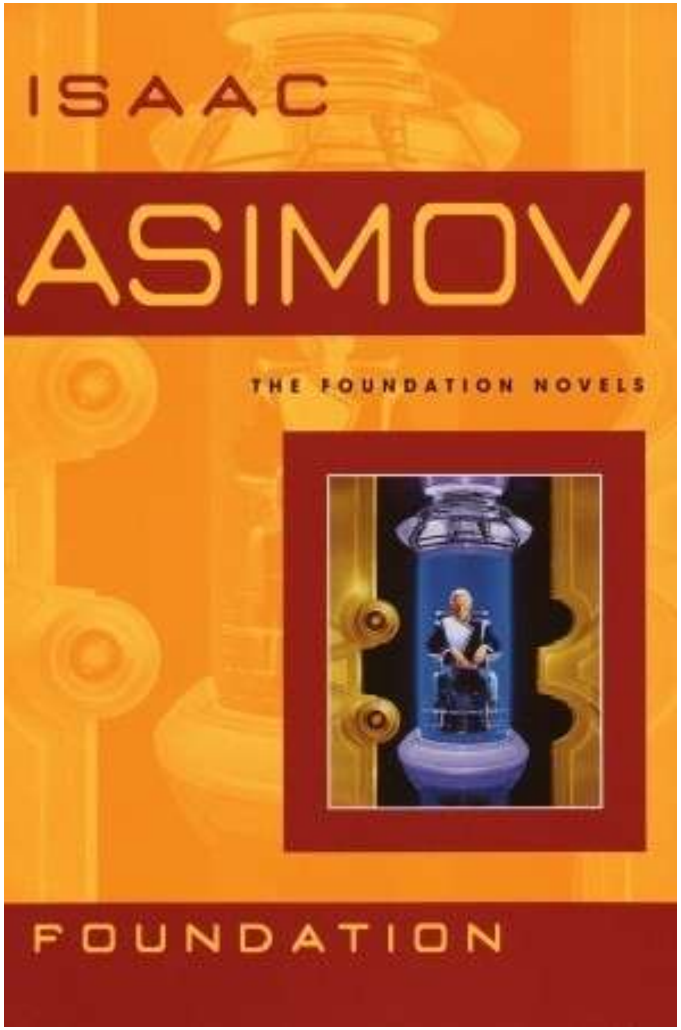


Photo by John Trainoron Flickr
http://www.flickr.com/photos/trainor/2902023575/, Licensed under CC

Asimov Foundation



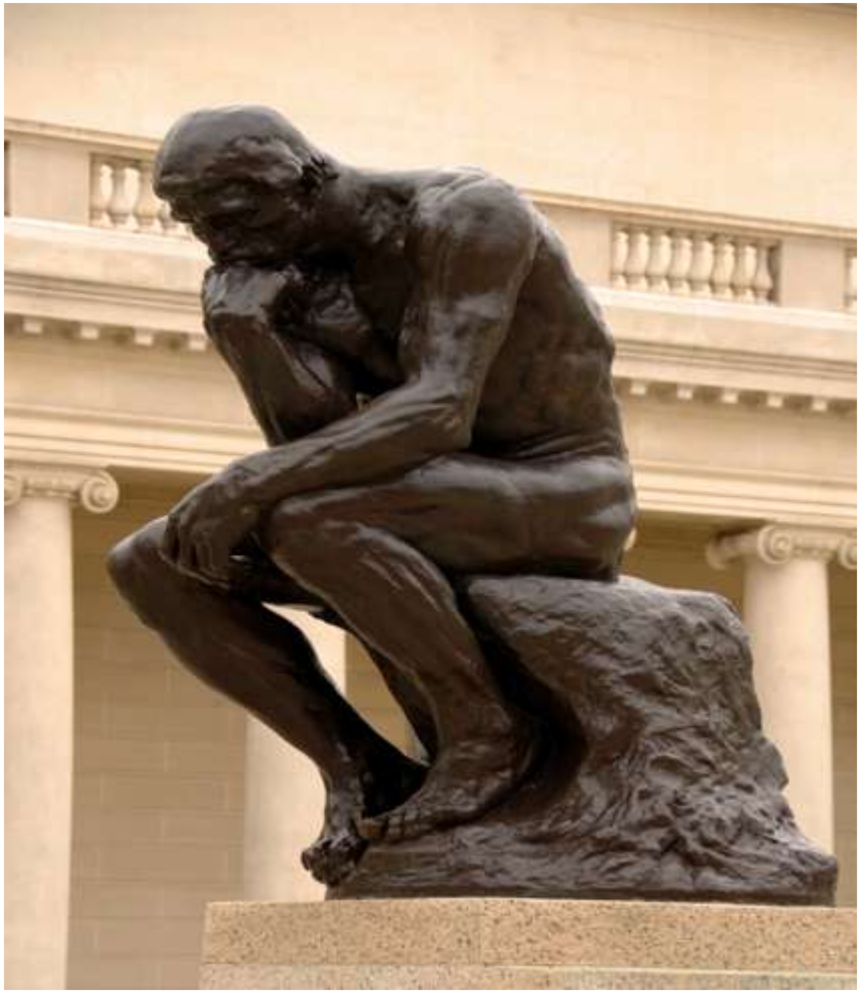
- *20 million worlds*
- *A Scientist who mathematically calculate the fate of the Universe*
- *His effort to change that Fate*
- *A Beautiful story*

Consider a day in your life

- What is the best road to take?
- Would there be any bad weather?
- What is the best way to invest the money?
- Should I take that loan?
- Can I optimize my day?
- Is there a way to do this faster?
- What have others done in similar cases?
- Which product should I buy?

People wanted to (through ages)

- To know (what happened?)
- To Explain (why it happened)
- To Predict (what will happen?)



Many Cultures had *Claim for Omniscience*

- Oracle
- Astrology
- Book of Changes
- Tarot Cards
- Crystal balls

Claim for Omniscience

Oracle
Of Delphi

To know, explain and predict!!

- Grand challenge of our time
- We have been trying to do this in many other means
- Now we trying to do this via science

Any sufficiently advanced technology is indistinguishable from magic.

---Arthur C. Clarke.

- We see a possibilities though “lot of data”

You does not seem to be convinced!!

- Why, sometimes I've believed as many as six impossible things before breakfast.

Prophecies of our time

- We can predict Weather
- We do understand language translation pretty well
- We can predict how an air plane behave good enough to fly
- Our ability on forensics
- We understand remedies for many diseases
- We can tell how astro bodies will behave (e.g. comets)

This is being Done: Weather

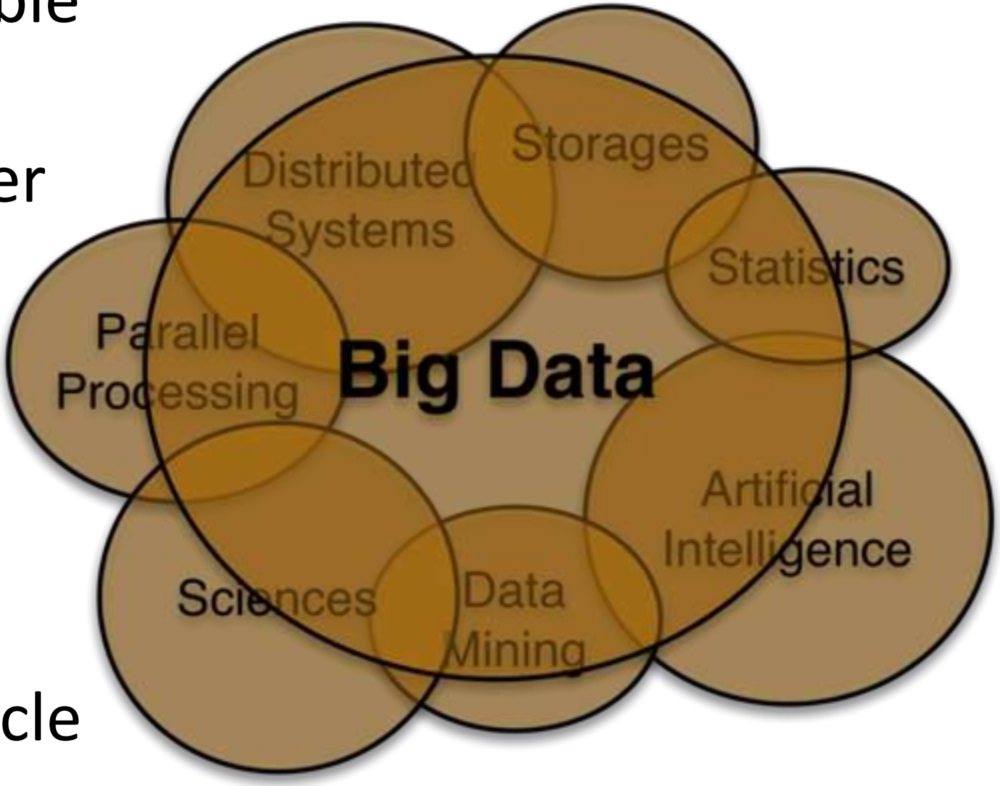
- Remarkably accurate for few days
  - SL incident
- Data :- weather radars, satellite data, weather balloons, planes etc.,
- forecast models
  - Simulation
  - Numerical method
- Challenge is computing power
  - algorithms that take more than 24 hours
  - resolution is the key ……

Democratizing Analysis

- Forecasting was done, but in limited manner and only by few (e.g. National Labs, Intelligence community)
- That is changing!!

What is Big data?

- There is lot of data available
  - E.g. Internet of things
- We have computing power
- We have technology
- Goal is same
  - To know
  - To Explain
  - To predict
- Challenge is the full lifecycle

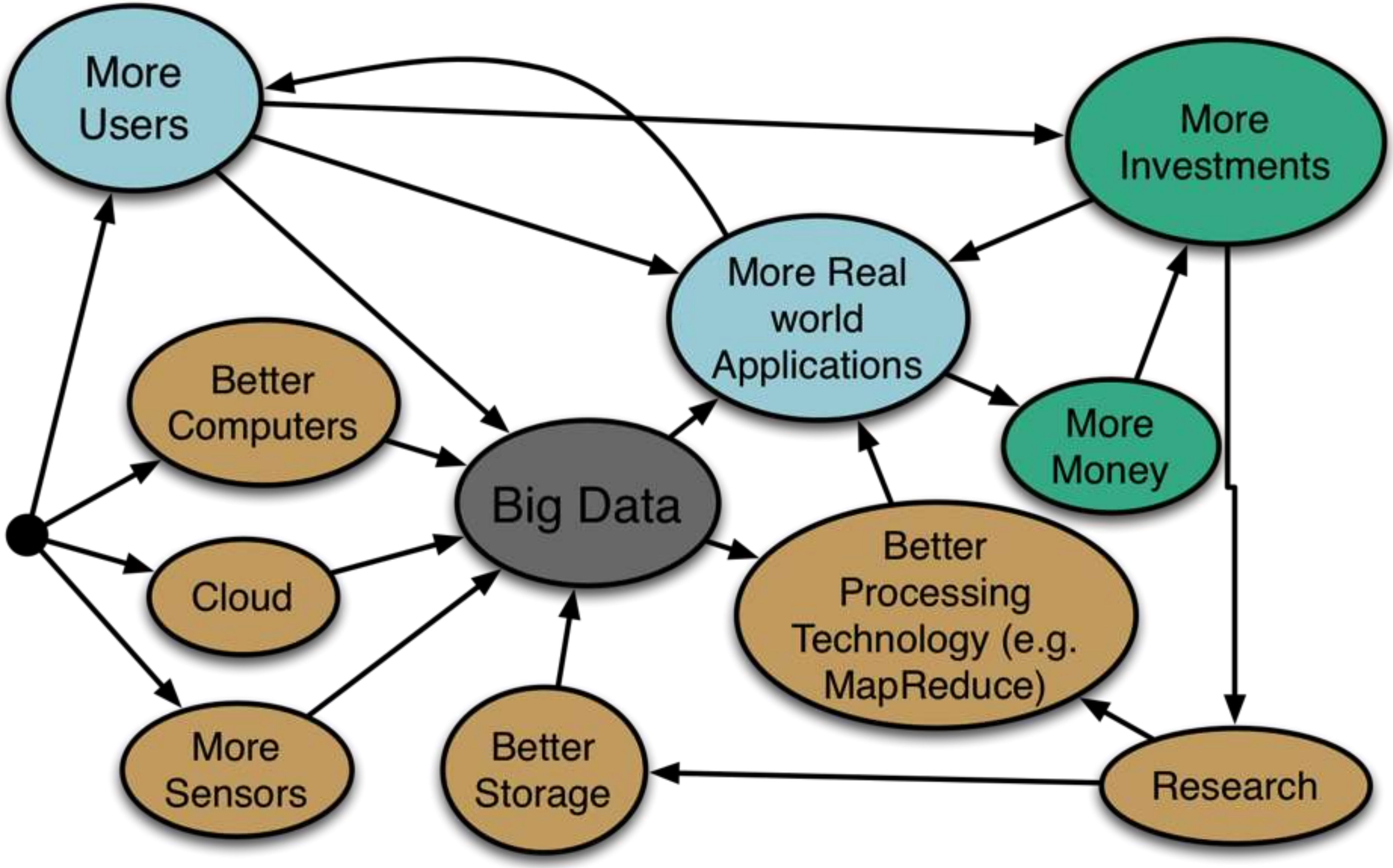


Data, the wealth of our time

"Data is a precious thing because they last longer than systems” - Tim Barnes Lee

- Access to data is becoming ultimate competitive advantage
  - E.g. Google+ vs. Facebook
  - Why many organizations try hard to give us free things and keep us always logged in (e.g. Gmail, facebook, search engine tool bars)

Drivers of Big Data



Data Avalanche/ Moore’s law of data

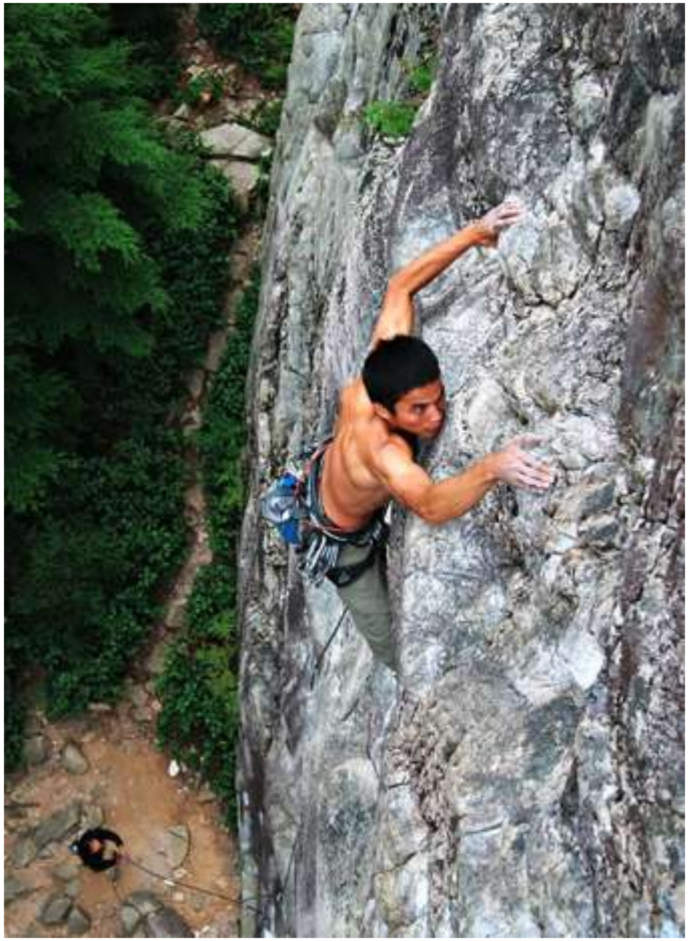
- We are now collecting and converting large amount of data to digital forms
- 90% of the data in the world today was created within the past two years.
- Amount of data we have doubles very fast

In real life, most data are Big

- Web does millions of activities per second, and so much server logs are created.
- Social networks e.g. Facebook, 800 Million active users, 40 billion photos from its user base.
- There are >4 billion phones and >25% are smart phones. There are billions of RFID tags.
- Observational and Sensor data
  - Weather Radars, Balloons
  - Environmental Sensors
  - Telescopes
  - Complex physics simulations

Why Big Data is hard?

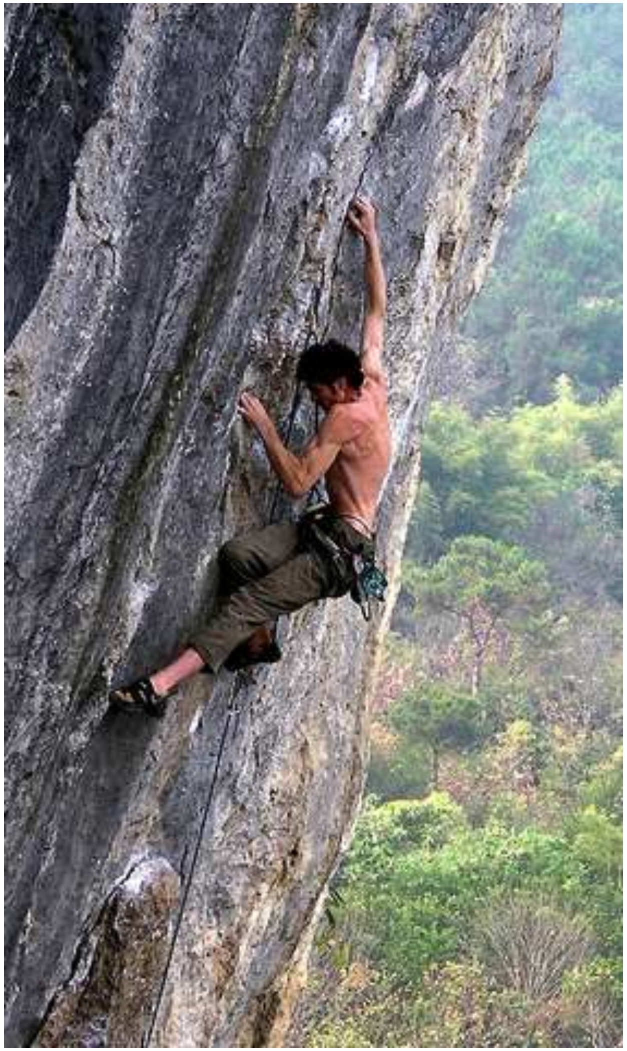
- How store? Assuming 1TB bytes it takes 1000 computers to store a 1PB
- How to move? Assuming 10Gb network, it takes 2 hours to copy 1TB, or 83 days to copy a 1PB
- How to search? Assuming each record is 1KB and one machine can process 1000 records per sec, it needs 277CPU days to process a 1TB and 785 CPU years to process a 1 PB
- How to process?
  - How to convert algorithms to work in large size
  - How to create new algorithms



http://www.susanica.com/photo/9

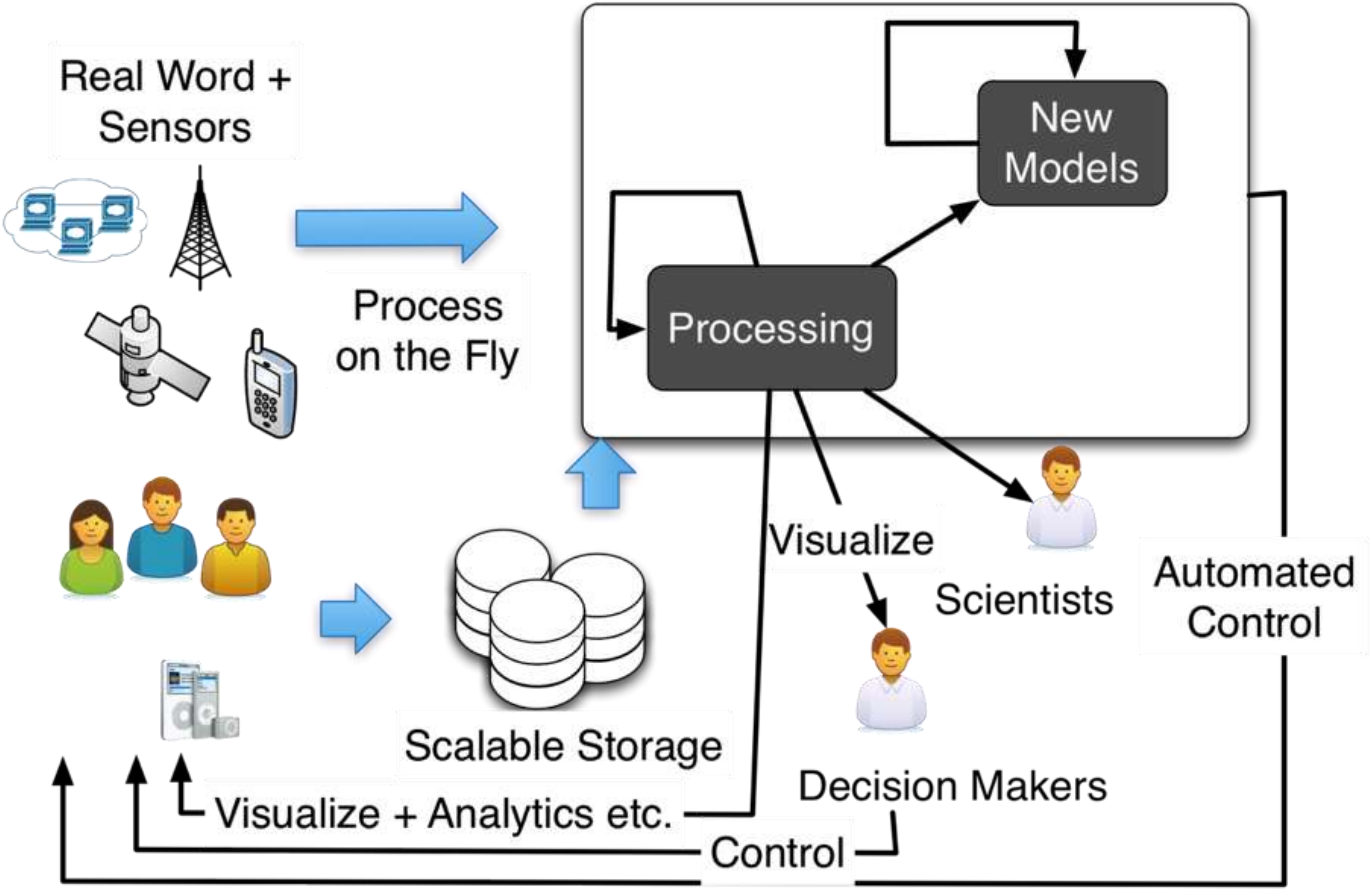
Why it is hard (Contd.)?

- System build of many computers
- That handles lots of data
- Running complex logic
- This pushes us to frontier of Distributed Systems and Databases
- More data does not mean there is a simple model
- Some models can be complex as the system



http://www.flickr.com/photos/mariachily/5250487136, Licensed CC

Big Data Architecture



Sensors

- There are sensors everywhere
  - RFID (e.g. Walmart), GPS sensors, Mobile Phone …
- Internet
  - Click streams, Emails, chat, search, tweets ,Transactions …
- Real Word
  - Video surveillance, Cash flows, Traffic, Surveillance, Stock exchange, Smart Grid, Production line …
  - Internet of Things

http://www.flickr.com/photos/imuttoo/4257813689/ by Ian Muttoo,
http://www.flickr.com/photos/eastcapital/4554220770/,
http://www.flickr.com/photos/patdavid/4619331472/ by Pat David copyright CC

Collecting Data

- Data collected at sensors and sent to big data system via events or flat files
- Event Streams: we name the events by its content/ originator
- Get data through
  - Point to Point
  - Event Bus
- E.g. Data bridge
  - a thrift based transport we did that do about 400k events/ sec

Storing Data

- Historically we used databases
  - Scale is a challenge: replication, sharding
- Scalable options
  - NoSQL (Cassandra, Hbase) [If data is structured]
    - Column families Gaining Ground
  - Distributed file systems (e.g. HDFS) [If data is unstructured]
- New SQL
  - In Memory computing, VoltDB
- Specialized data structures
  - Graph Databases, Data structure servers

http://www.flickr.com/photos/keso/363133967/

Making Sense of Data

- To know (what happened?)
  - Basic analytics + visualizations (min, max, average, histogram, distributions …)
  - Interactive drill down
- To explain (why)
  - Data mining, classifications, building models, clustering
- To forecast
  - Neural networks, decision models

To know (what happened?)

- Mainly Analytics
  - Min, Max, average, correlation, histograms
  - Might join group data in many ways
- Implemented with MapReduce or Queries
- Data is often presented with some visualizations
- Examples
  - forensics
  - Assessments
  - Historical data/ reports/ trends

http://www.flickr.com/photos/isriya/2967310
333/

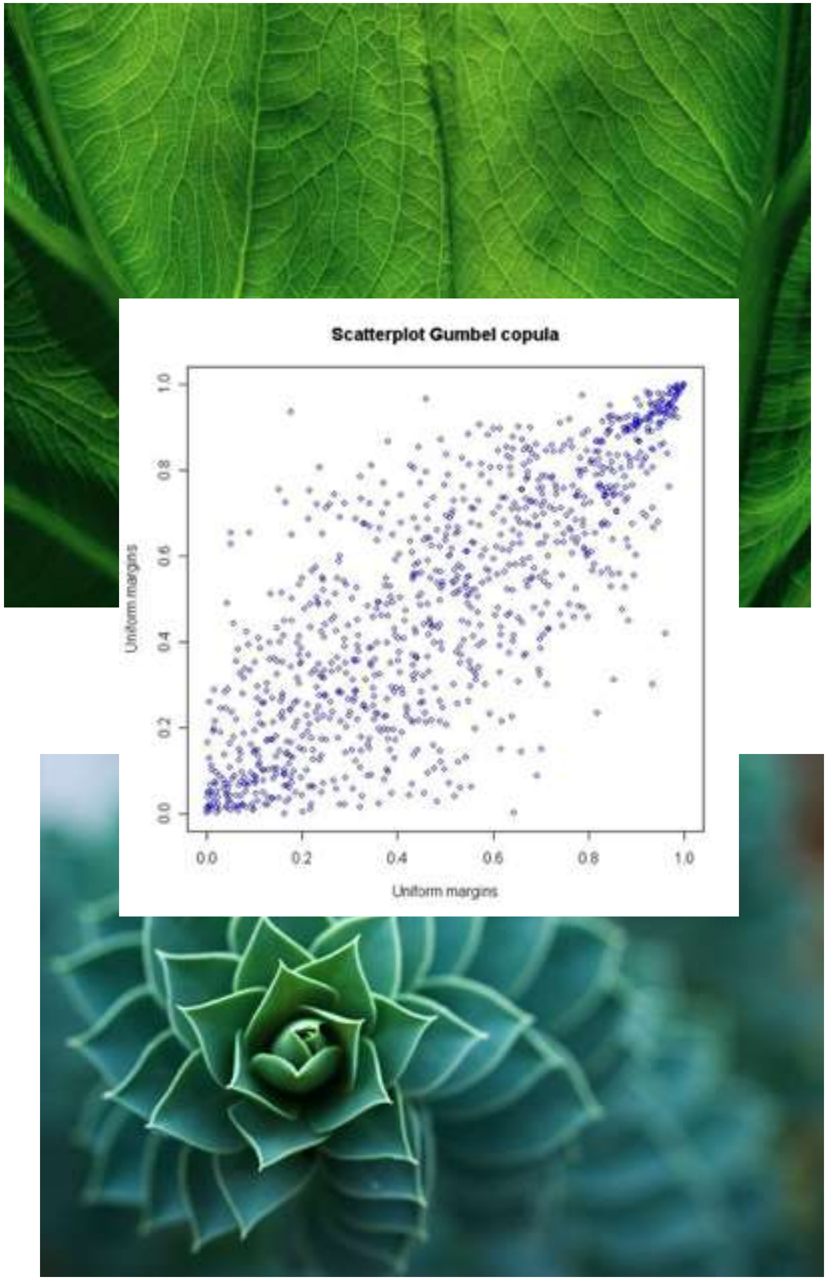
Search

- Process and Index the data. The killer app of our time.
- Web Search
- Graph Search
- Semantic Search
- Drug Discovery
- …

http://pixabay.com, by Steven Iodice

Patterns

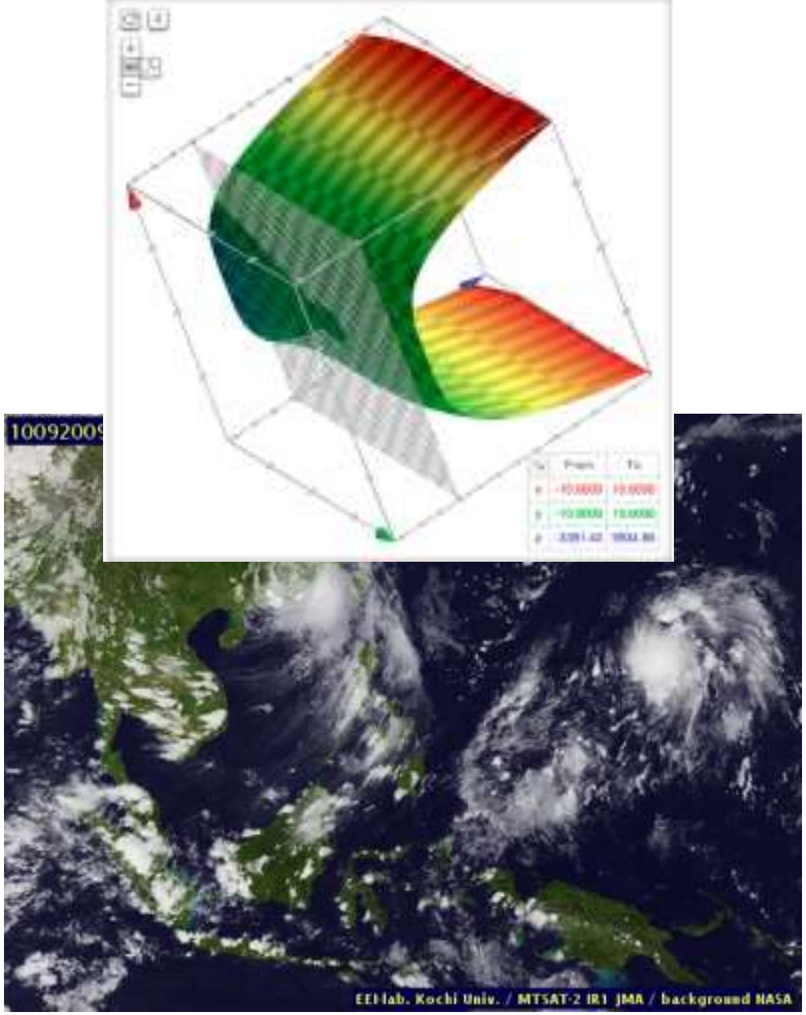
- Correlation
  - Scatter plot, statistical correlation
- Data Mining (Detecting Patterns)
  - Clustering and classification
  - Finding Similar items
  - Finding Hubs and authorities in a Graph
  - Finding frequent item sets
  - Making recommendation
- Apache Mahout



http://www.flickr.com/photos/eriwst/2987739376/ and http://www.flickr.com/photos/focx/5035444779/

Forecasts and Models

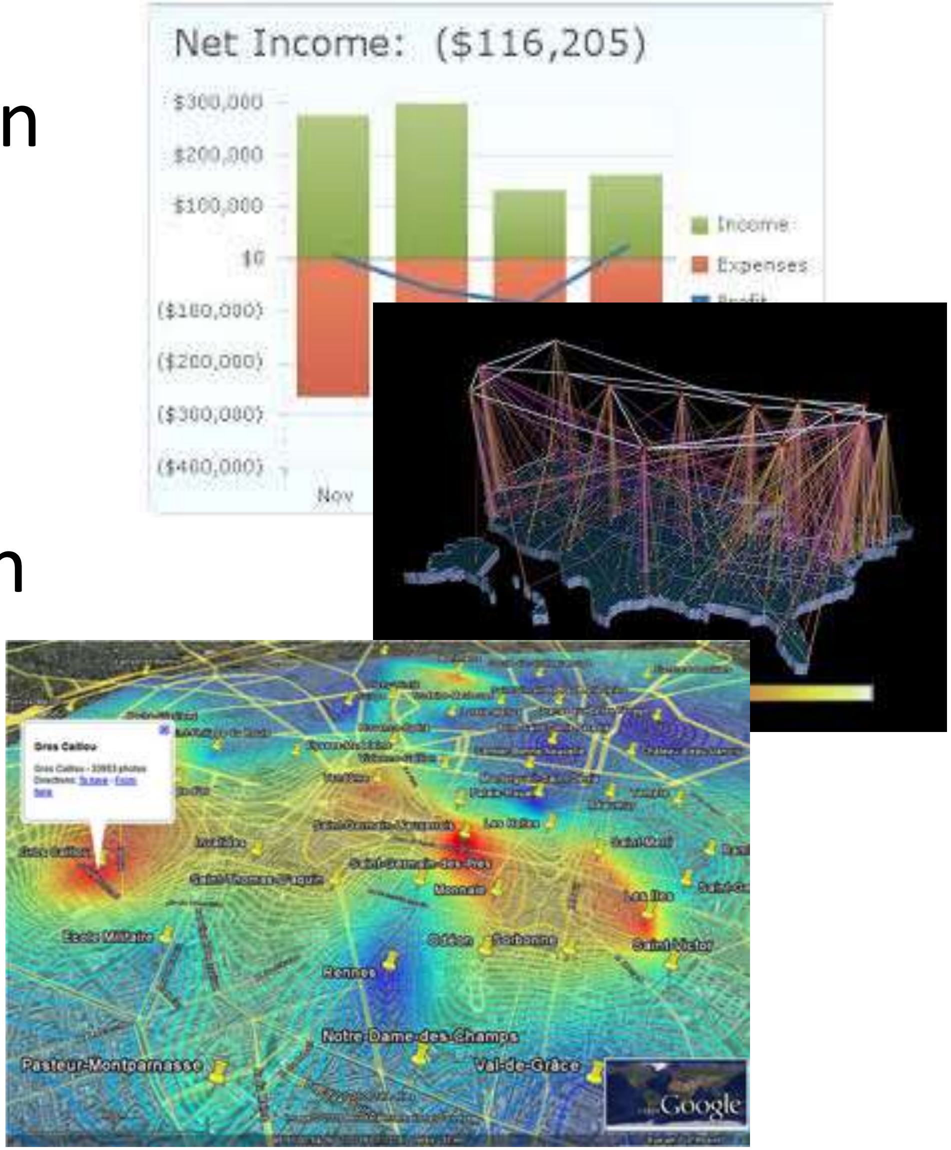
- Trying to build a model for the data
- Theoretically or empirically
  - Analytical models (e.g. Physics)
  - Neural networks
  - Reinforcement learning
  - Unsupervised learning (clustering, dimensionality reduction, kernel methods)
- Examples
  - Translation
  - Weather Forecast models
  - Building profiles of users
  - Traffic models
  - Economic models
- Remember: Correlation does not mean causality



http://misterbijou.blogspot.com/2010_09_01_archive.html

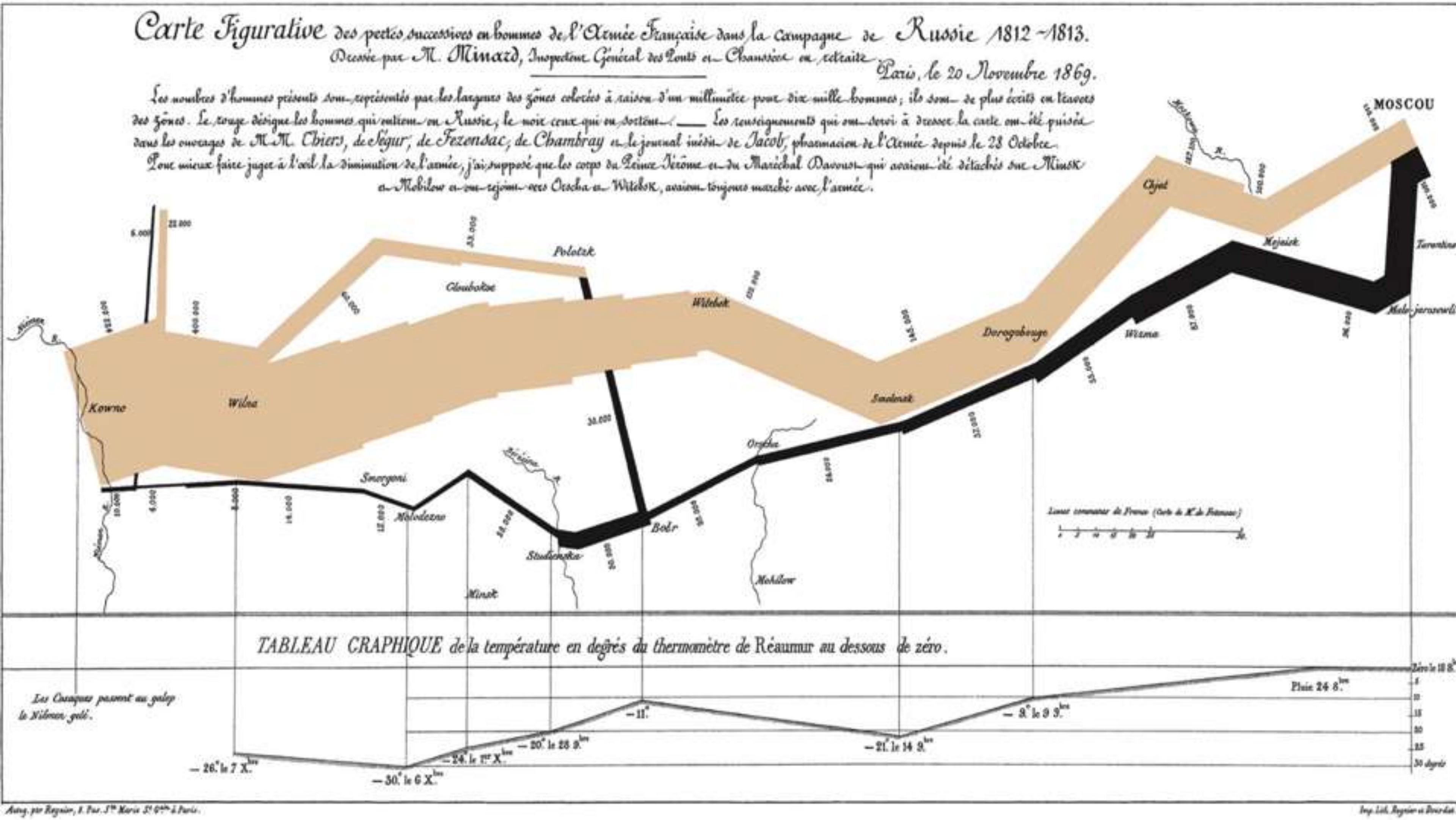
Information Visualization

- Presenting information
  - – To end user
  - – To decision takers
  - – To scientist
- Interactive exploration
- Sending alerts



http://www.flickr.com/photos/stevefaeembra/3604686097/

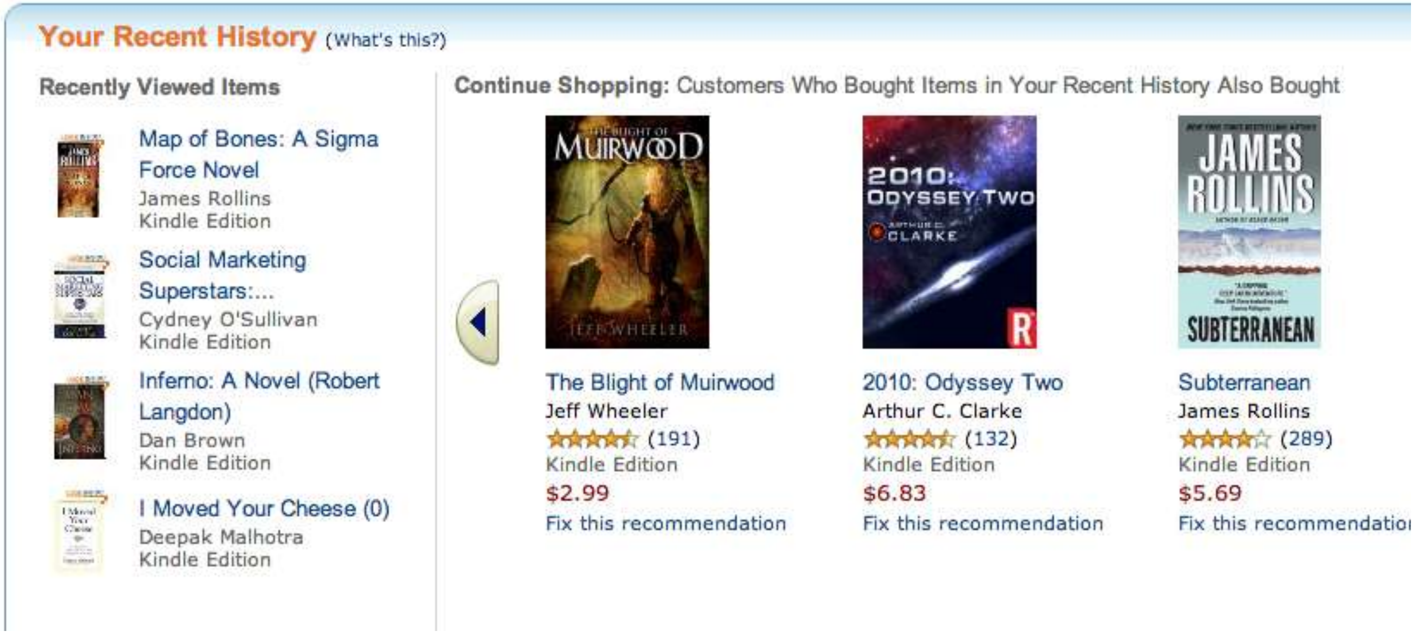
Napoleon's Russian Campaign



Show Han Rosling’s Data

- This use a tool called “Gapminder”

Usecase 1: Targeted Marketing



Your Recent History (What's this?)

Recently Viewed Items

- Map of Bones: A Sigma Force Novel
  James Rollins
  Kindle Edition
- Social Marketing Superstars:...
  Cydney O'Sullivan
  Kindle Edition
- Inferno: A Novel (Robert Langdon)
  Dan Brown
  Kindle Edition
- I Moved Your Cheese (0)
  Deepak Malhotra
  Kindle Edition

Continue Shopping: Customers Who Bought Items in Your Recent History Also Bought

- The Blight of Muirwood
  Jeff Wheeler
  ★★★★★ (191)
  Kindle Edition
  $2.99
  Fix this recommendation
- 2010: Odyssey Two
  Arthur C. Clarke
  ★★★★★ (132)
  Kindle Edition
  $6.83
  Fix this recommendation
- Subterranean
  James Rollins
  ★★★★☆ (289)
  Kindle Edition
  $5.69
  Fix this recommendatio

- Collect data and build a model on
  - What user like
  - What he has brought
  - What is his buying power
- Making recommendations
- Giving personalized deals

Usecase 2: Travel

- Collect traffic data + transportation data from sensors
- Build a model (e.g. Car Following Models)
- Predict the traffic .. Possibilities on congestion
- E.g. divert traffic, adjust troll, adjust traffic lights

Practical Tutorial

- Data collection
  - Pub/sub, event architectures
- Processing
  - Store and Process: MapReduce/Hadoop
  - Processing Moving Data: CEP
- Visualization
  - GNU Plot/ R

DEBS Challenge

- Event Processing challenge
- Real football game, sensors in player shoes + ball
- Events in 15k Hz
- Event format
  - Sensor ID, TS, x, y, z, v, a
- Queries
  - Running Stats
  - Ball Possession
  - Heat Map of Activity
  - Shots at Goal

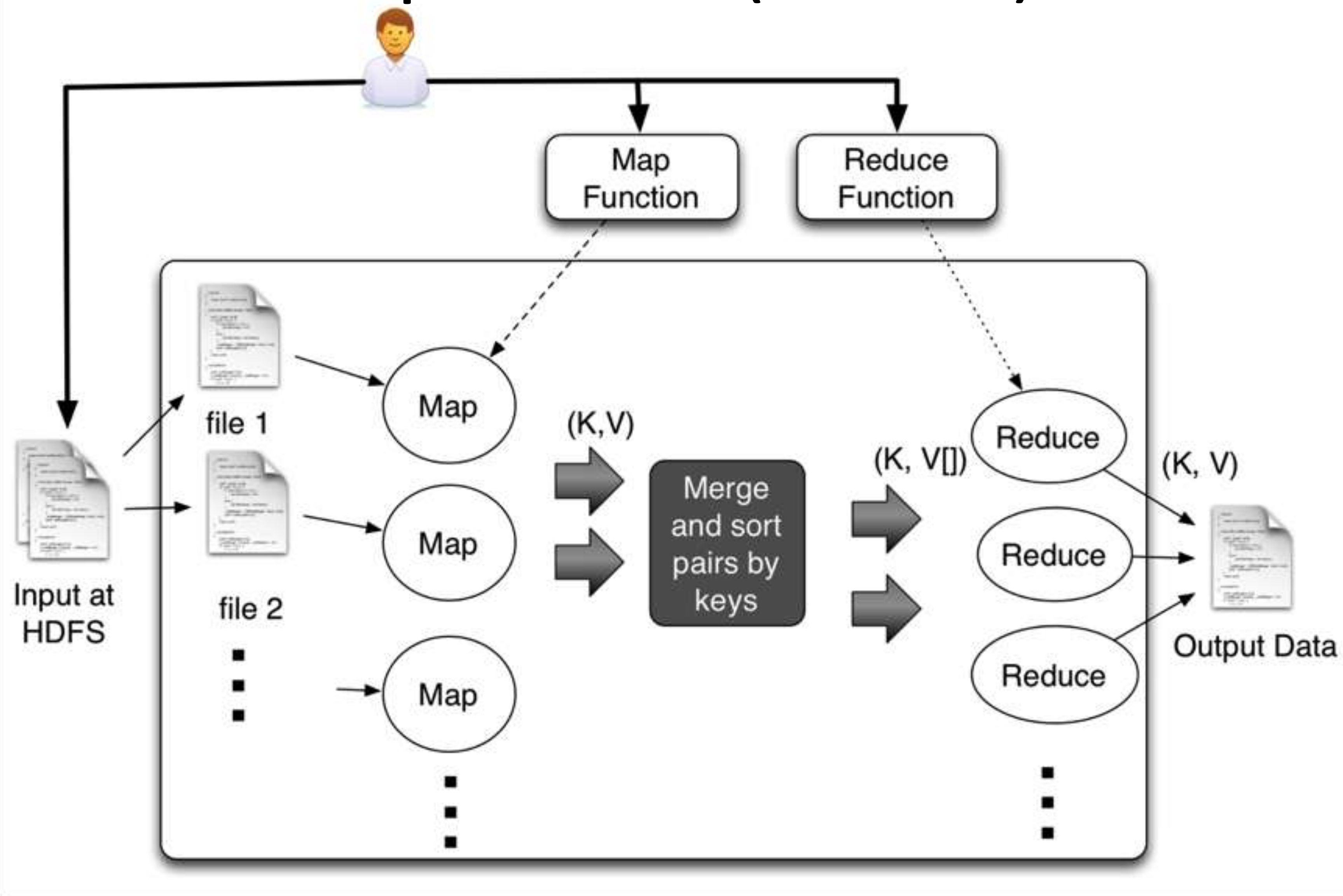
MapReduce/ Hadoop

- First introduced by Google, and used as the processing model for their architecture
- Implemented by opensource projects like Apache Hadoop and Spark
- Users writes two functions: map and reduce
- The framework handles the details like distributed processing, fault tolerance, load balancing etc.
- Widely used, and the one of the catalyst of Big data

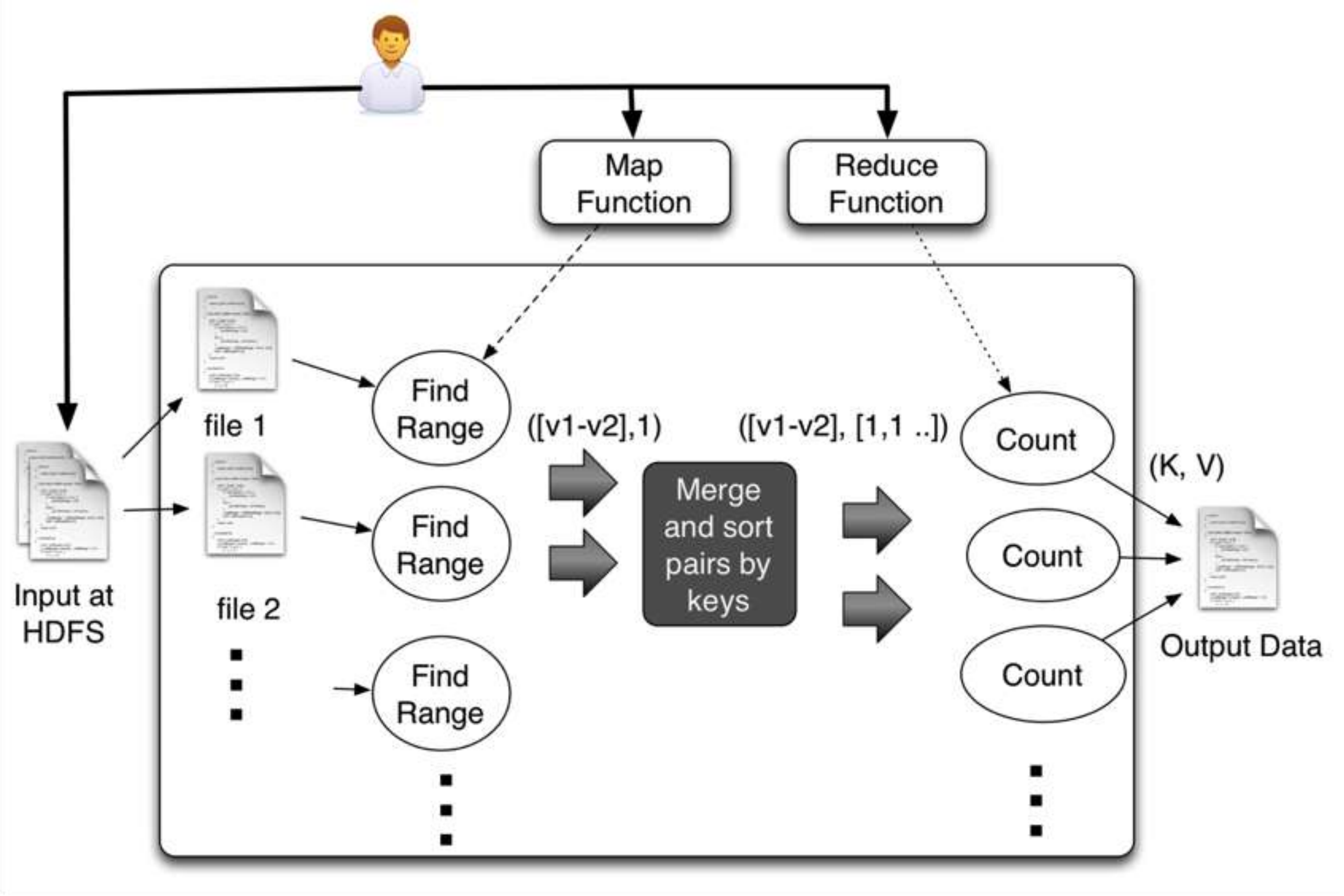
```
void map(ctx, k, v){
    tokens = v.split();
    for t in tokens
        ctx.emit(t,1)
}


void reduce(ctx, k, values[]){
    count = 0;
    for v in values
        count = count + v;
    ctx.emit(k,count);
}
```

MapReduce (Contd.)



Histogram of Speeds

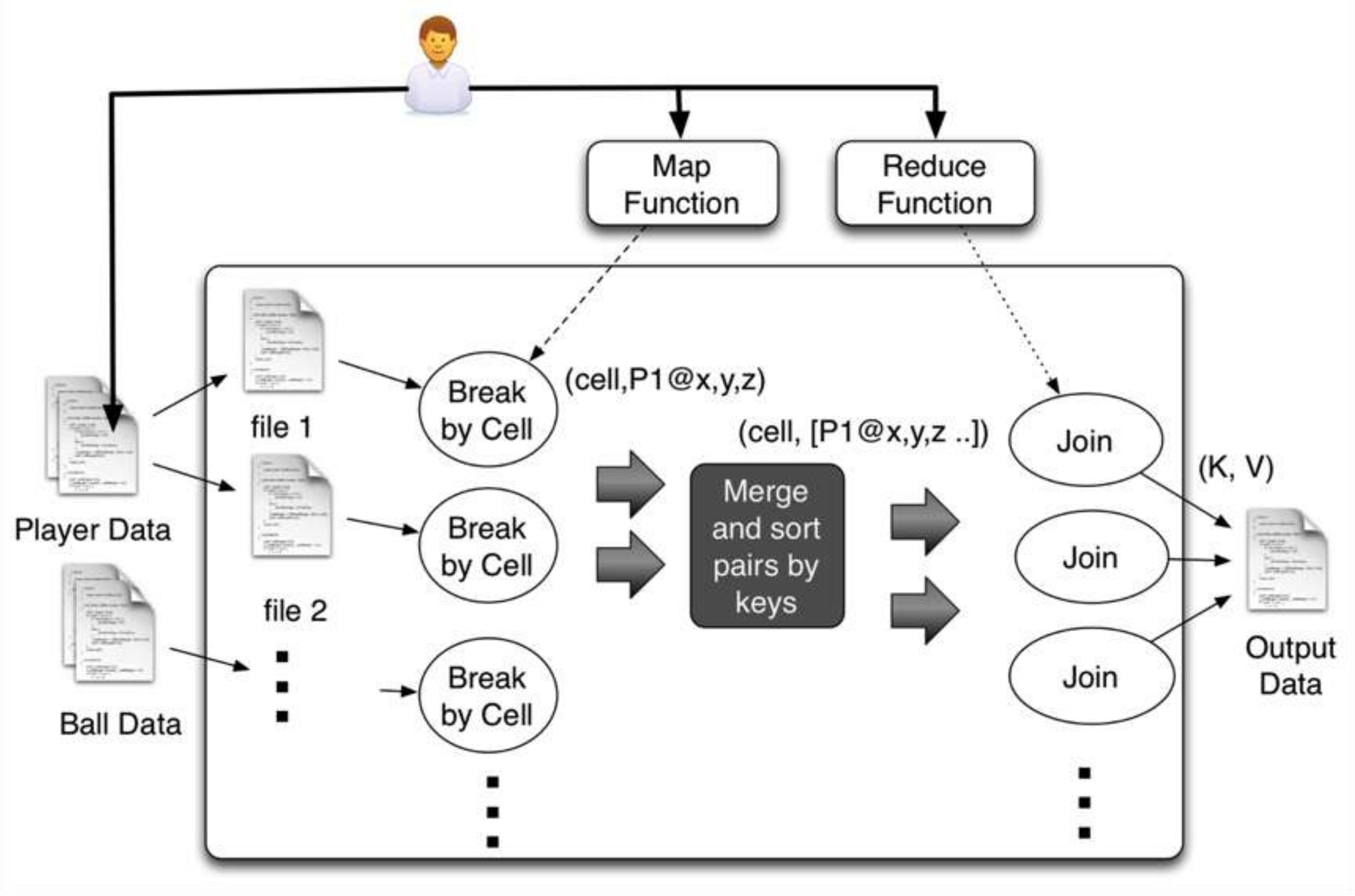


Histogram of Speeds(Contd.)

```
void map(ctx, k, v){
    event = parse(v);
    range = calcuateRange(event.v);
    ctx.emit(range,1)
}


void reduce(ctx, k, values[]){
    count = 0;
    for v in values
        count = count + v;
    ctx.emit(k,count);
}
```

How long player Spend Near the ball?



How long player Spend Near the ball?(Contd.)

```
void map(ctx, k, v){
    event = parse(v);
    cell = calcuateOverlappingCells(v.x,
        v.y, v.z);
    ctx.emit(cell,v.x, v.y, v.z);
}


Void reduce(ctx, k, values[]){
    players = joinAndFindPlayersNearBall(values);
    for p in players
        ctx.emit(p,p);
}
```

Hadoop Landscape

- HIVE - Query data using SQL style queries, and Hive will convert them to MapReduce jobs and run in Hadoop.

```
hive> SELECT country, gni from HDI WHERE gni > 2000;
```

- Pig - We write programs using data flow style scripts, and Pig convert them to MapReduce jobs and run in Hadoop.

```
A = load 'hdi-data.csv' using PigStorage(',')
    AS (id:int, country:chararray, hdi:float,
    lifeex:int, mysch:int, eysch:int, gni:int);
B = FILTER A BY gni > 2000;
C = ORDER B BY gni;
dump C;
```

- Mahout
  - Collection of MapReduce jobs implementing many Data mining and Artificial Intelligence algorithms using MapReduce

Is Hadoop Enough?

- Limitations
  - Takes time for processing
  - Lack of Incremental processing
  - Weak with Graph usecases
  - Not very easy to create a processing pipeline (addressed with HIVE, Pig etc.)
  - Too close to programmers
  - Faster implementations are possible
- Alternatives
  - Apache Drill http://incubator.apache.org/drill/
  - Spark http://spark-project.org/
  - Graph Processors like http://giraph.apache.org/

Data In the Move

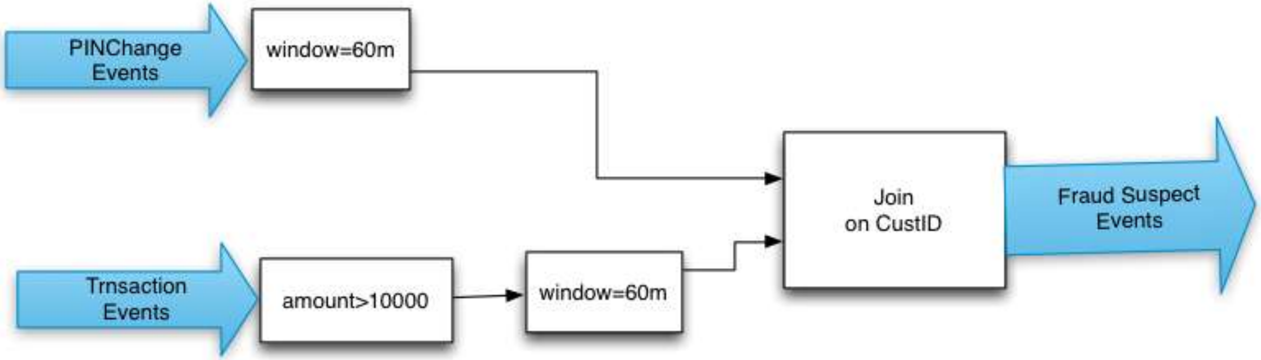
- Idea is to process data as they are received in streaming fashion
- Used when we need
  - Very fast output
  - Lots of events (few 100k to millions)
  - Processing without storing (e.g. too much data)
- Two main technologies
  - Stream Processing (e.g. Strom, http://storm-project.net/)
  - Complex Event Processing (CEP)

  http://wso2.com/products/complex-event-processor/

Complex Event Processing (CEP)

- Sees inputs as Event streams and queried with SQL like language
- Supports Filters, Windows, Join, Patterns and Sequences

```
from p=PINChangeEvents#win.time(3600) join
   t=TransactionEvents[p.custid=custid][amount>1000
0]               #win.time(3600)
return t.custid, t.amount;
```



Example: Detect ball Possession

```
from Ball#window.length(1) as b join
    Players#window.length(1) as p
        unidirectional
   on debs: getDistance(b.x,b.y,b.z,
     p.x, p.y, p.z) < 1000
        and b.a > 55
select ...
insert into hitStream

from old = hitStream ,
    b = hitStream [old. pid != pid ],
    n= hitStream[b.pid == pid]*,
    ( e1 = hitStream[b.pid != pid ]
        or e2= ballLeavingHitStream)
select ...
insert into BallPossessionStream
```

- Possession is time a player hit the ball until someone else hits it or it goes out of the ground

http://www.flickr.com/photos/glennharper/146164820/

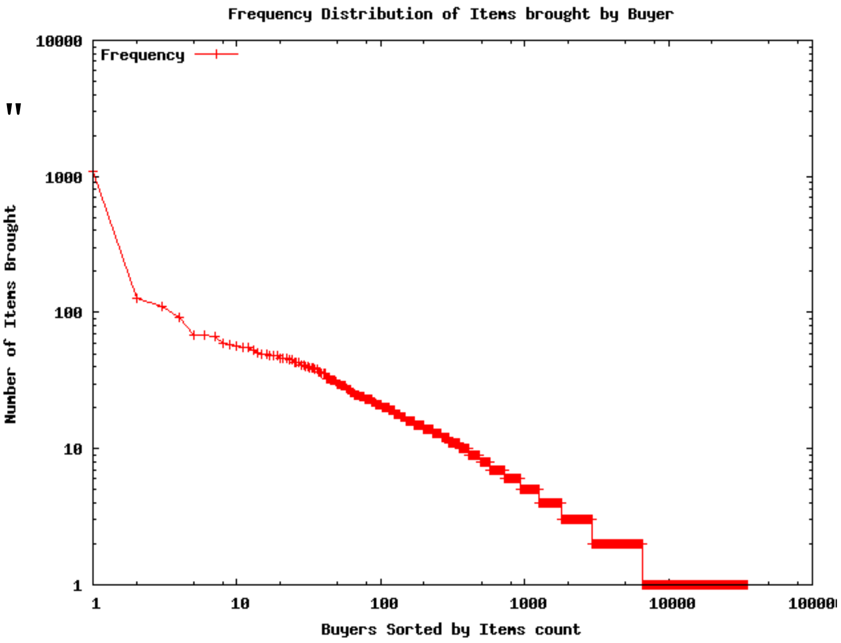
GNU Plot

```
set terminal png
set output "buyfreq.png"

set title "Frequency Distribution of Items
brought by Buyer";
setylabel "Number of Items Brought";
setxlabel "Buyers Sorted by Items count";
set key left top
set log y
set log x

plot "1.data" using 2 title "Frequency"
with linespoints
```

- Open source
- Very powerful



Big data lifecycle

- Realizing the big data lifecycle is hard
- Need wide understanding about many fields
- Big data teams will include members from many fields working together

- Data Scientists: A new role, evolving Job description
- His Job is to make sense of data, by using Big data processing, advance algorithms, Statistical methods, and data mining etc.
- Likely to be a highest paid/ cool Jobs.
- Big organizations (Google, Facebook, Banks) hiring Math PhD by a lot for this.

Dark Side

- Privacy
  - Invasion of privacy
  - Data might be used for unexpected things
- Big Brother
  - Data likely to used for control (e.g. governments)
- If technology is out there, may be it is OK. It is very hard to hide any thing, which work both ways

Challenges

- Speed (e.g. targeted advertising, reacting to data)
- Extracting semantics and handling multiple representations and formats
- Security Data ownership, delegation, permissions, and Privacy
- Making data accessible to all intended parties, from anywhere, anytime, from any device, through any format (subjected to permissions).
- Map-Reduce good enough? What about other parallel problems?
- Handling Uncertainty

Conclusions

- Lot of data, realizing data -> insight -> predications
- We do lot of predications even now.
- There is lot between data and forecasts, OK to do them.
  - Analytics
  - Visualizations
  - Patterns
  - Data mining
- If you looking to start, learn MapReduce, CEP, and GNU plot.
- Visualization is the key!!
- Learn some distributed systems and AI

Questions?