BigData

Building Scientific and Consumer Applications

Terrance Boult

Abhijit Bendale

Definition of Big Data

No single definition

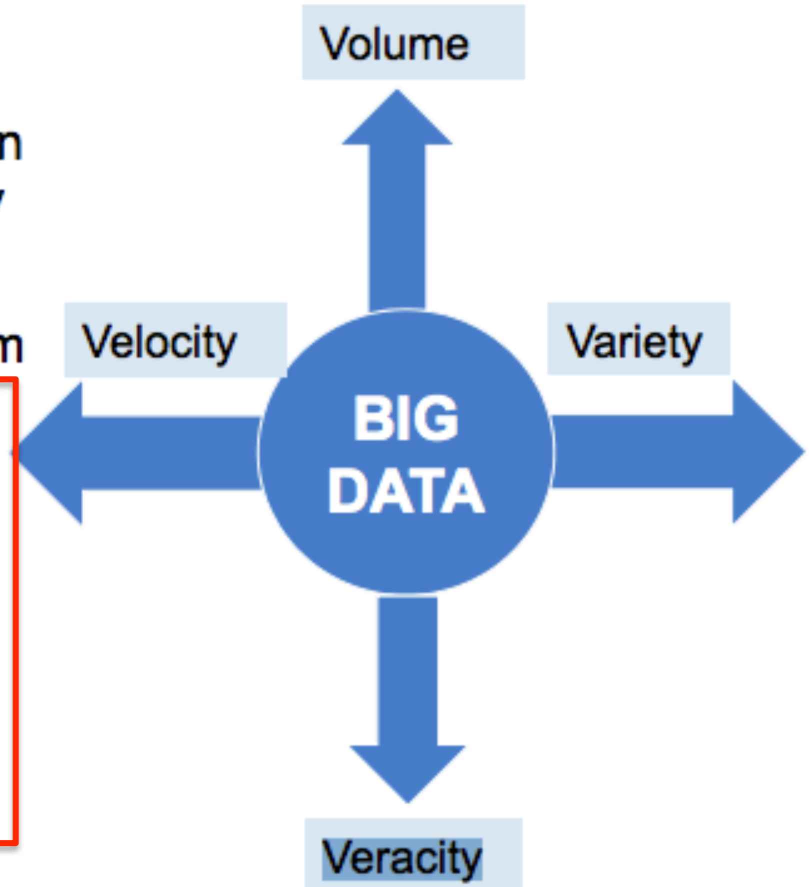
- A collection of large and complex data sets which are difficult to process using common database management tools or traditional data processing applications.

- "Big data refers to the tools, processes and procedures allowing an organization to create, manipulate, and manage very large data sets and storage facilities"

– according to zdnet.com

Big data is not just about size.
- Finds insights from complex, noisy, heterogeneous, longitudinal, and voluminous data.
- It aims to answer questions that were previously unanswered.

The challenges include capturing, storing, searching, sharing & analyzing.



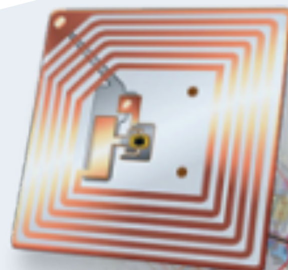
The four dimensions (V's) of Big Data

Who's Generating Big Data ?

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**

76 million **smart meters in 2009… 200M by 2014**

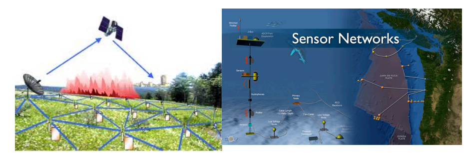
2+ billion **people on the Web by end 2011**

Who's Generating Big 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

The Model Has Changed…

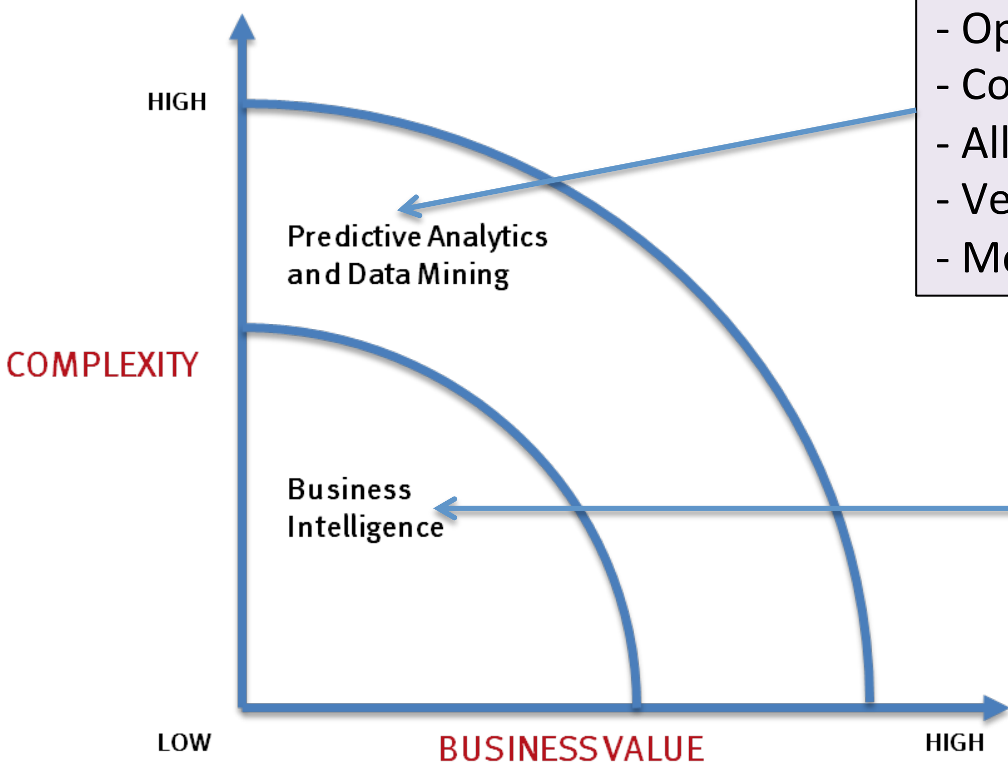
- **The Model of Generating/Consuming Data has Changed**

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

Two way street…!

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

What is driving Bigdata ?

Money to be made..!!

McKinsey Global Institute

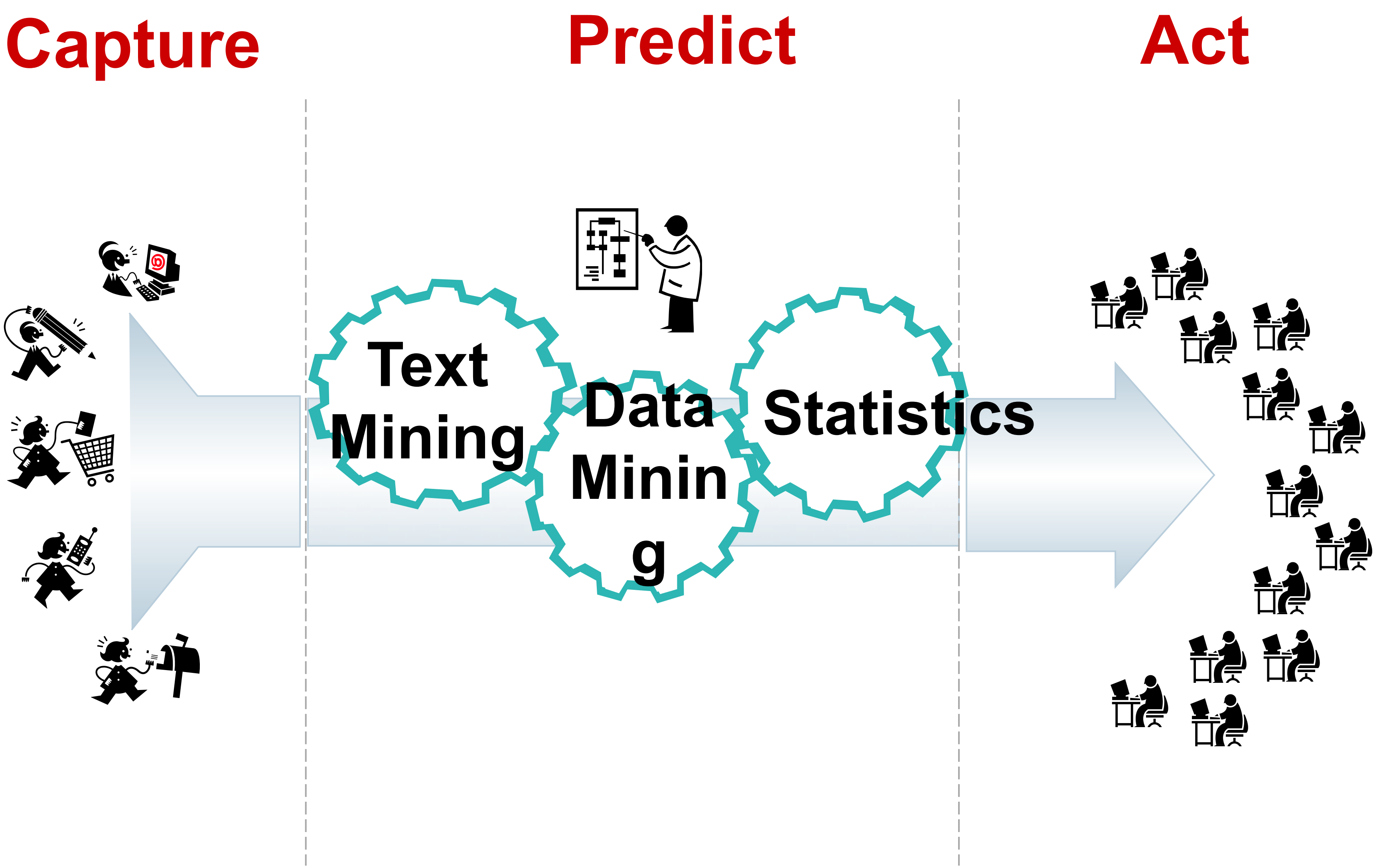
June 2011

Big data: The next frontier for innovation, competition, and productivity

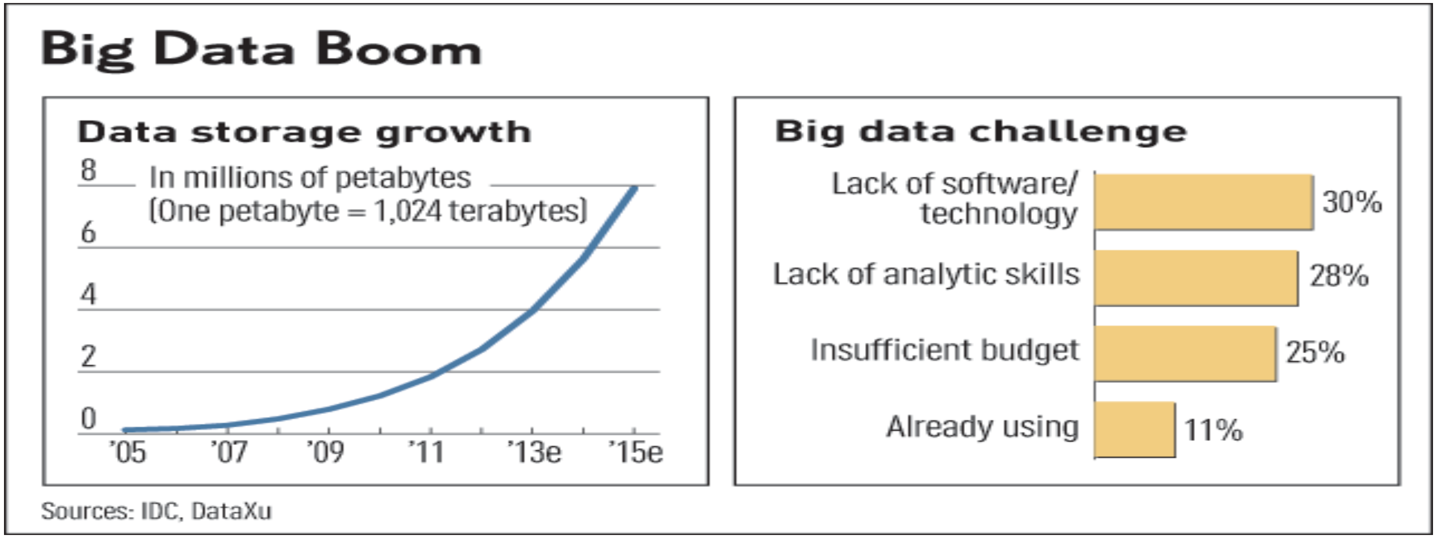
Why Big-Data?

- Key enablers for the appearance and growth of ‘Big-Data’ are:
  - Increase in storage capabilities
  - Increase in processing power
  - Availability of data

Driving Smarter Business Outcomes

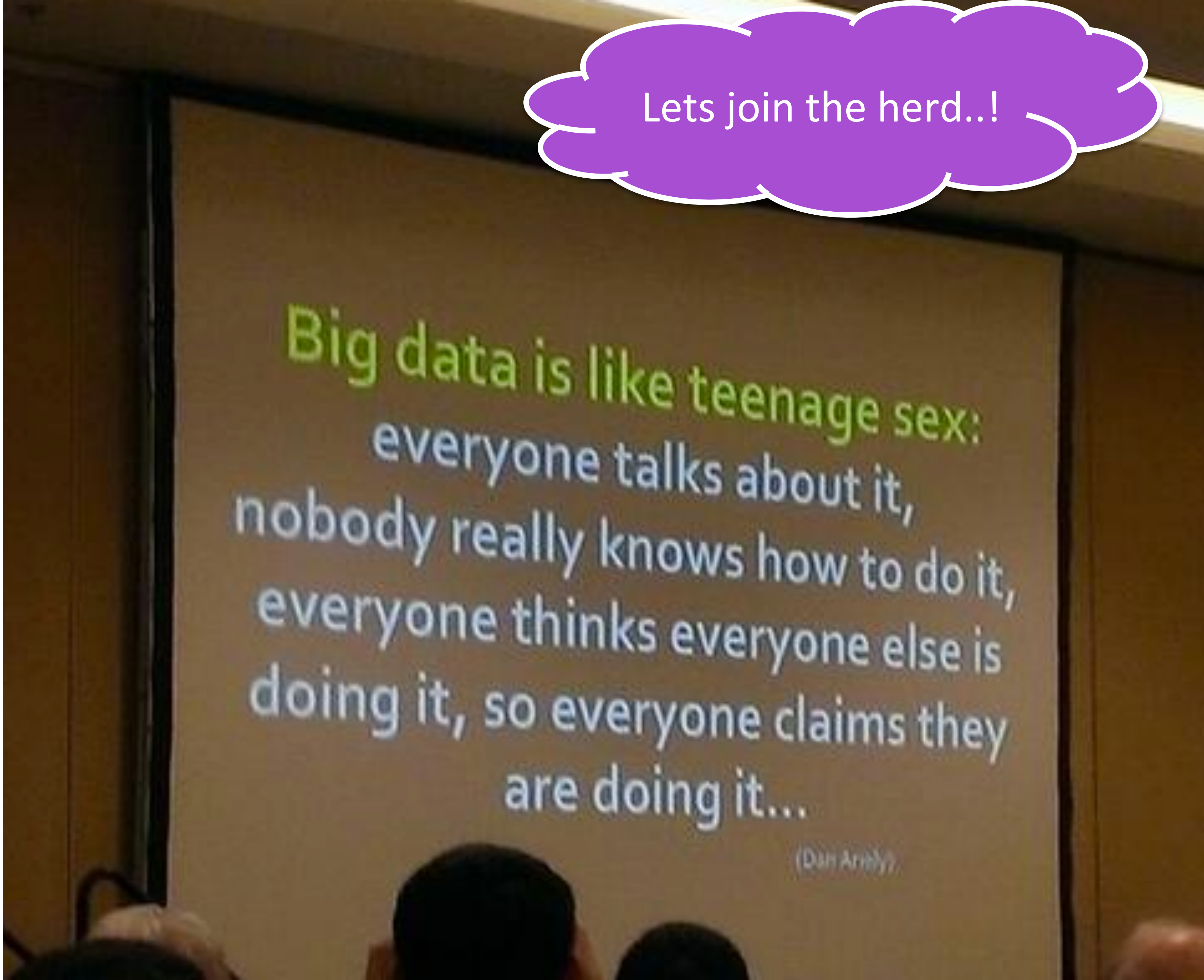


Challenges in Handling Big Data



- **The Bottleneck is in technology**
  - New architecture, algorithms, techniques are needed
- **Also in technical skills**
  - Experts in using the new technology and dealing with big data

Lets join the herd..!



Big Data Landscape

Vertical Apps
- PredPol (Predictive Policing)
- bloomreach
- MYRRIX

Log Data Apps
- splunk>
- loggly
- sumologic

Ad/Media Apps
- rocketfuel
- collective[i]
- bluefin
- Recorded Future
- LuckySort
- Media Science
- TURN
- DataXu

Business Intelligence
- ORACLE | Hyperion
- SAP BusinessObjects
- RJMetrics
- Microsoft Business Intelligence
- IBM
- COGNOS
- birst
- MicroStrategy
- Autonomy
- QlikView
- bime
- DOMO
- Chart.io
- GoodData

Analytics and Visualization
- tableau
- Palantir
- OPERA
- metaLayer
- METAMARKETS
- dataspora
- centrifuge
- TERADATA ASTER
- sas
- TIBCO
- KARMASPHERE
- panopticon (Real-Time Visual Data Analysis)
- pentaho
- Datameer
- platfora
- ClearStory
- CIRRO
- alteryx
- visual.ly
- AYATA

Data As A Service
- factual.
- kaggle
- knoema beta
- GNIP
- DATASIFT
- Windows Azure Marketplace
- INRIX
- LexisNexis
- Space Curve
- LOCATE

Analytics Infrastructure
- Hortonworks
- VERTICA
- MAPR
- cloudera
- INFOBRIGHT
- PARACCEL
- EMC²
- GREENPLUM
- NETEZZA
- kognitio
- DATASTAX
- EXASOL
- calpont

Operational Infrastructure
- COUCHBASE
- 10gen | the MongoDB company
- TERADATA
- HADAPT
- TERRACOTTA
- VoltDB
- MarkLogic
- INFORMATICA

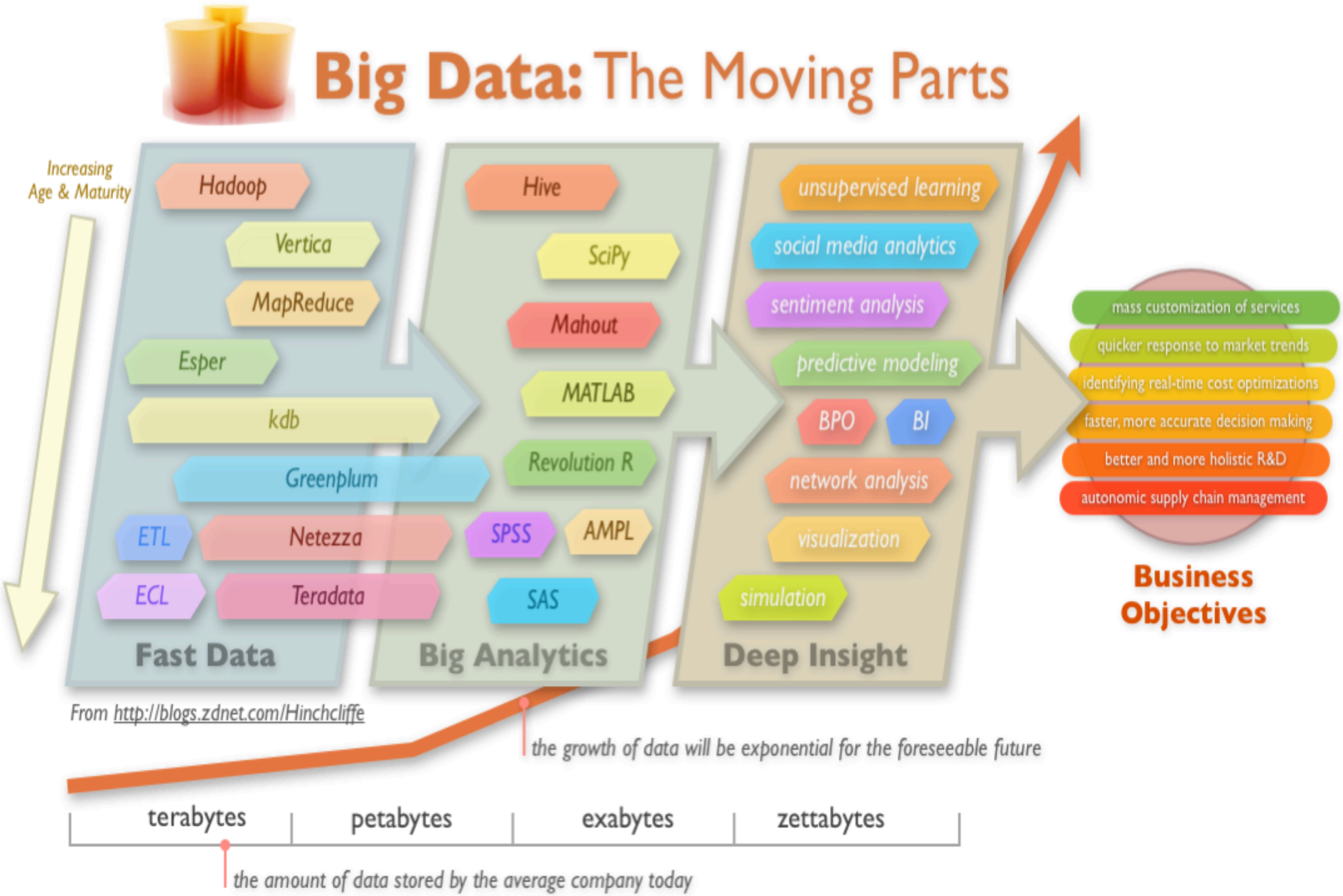
Infrastructure As A Service
- amazon web services
- Windows Azure
- infochimps
- Google BigQuery

Structured Databases
- ORACLE
- MySQL
- Microsoft SQL Server
- PostgreSQL
- IBM DB2
- SYBASE
- memsql

Technologies
- hadoop
- hadoop MapReduce
- mahout
- APACHE HBASE
- Cassandra

dave@vcdave.com

blogs.forbes.com/davefeinleib


Big Data: The Moving Parts
Increasing Age & Maturity
Hadoop
Vertica
MapReduce
Esper
kdb
Greenplum
ETL
Netezza
ECL
Teradata
Fast Data
Hive
SciPy
Mahout
MATLAB
Revolution R
SPSS
AMPL
SAS
Big Analytics
unsupervised learning
social media analytics
sentiment analysis
predictive modeling
BPO
BI
network analysis
visualization
simulation
Deep Insight
mass customization of services
quicker response to market trends
identifying real-time cost optimizations
faster, more accurate decision making
better and more holistic R&D
autonomic supply chain management
Business Objectives
From http://blogs.zdnet.com/Hinchcliffe
the growth of data will be exponential for the foreseeable future
terabytes
petabytes
exabytes
zettabytes
the amount of data stored by the average company today

What will you learn …

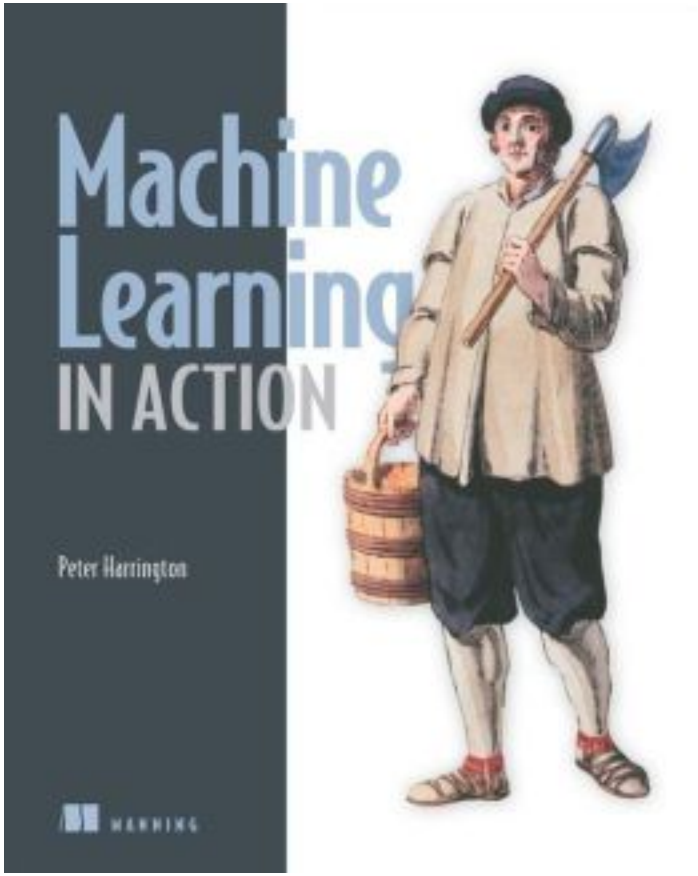
- Problem Analysis and identification
  - Machine Learning, Statistics analysis techniques to handle big data
  - Theoretical and engineering constraints
  - Parallelizing applications for high throughput
- Tools and Techniques
  - Introduction to Machine learning tools like scikits-learn, weka, R etc
  - Parallel Programming using CUDA
  - Introduction to cloud based tools like MapReduce, Hadoop, Pig, Hive etc
- Case studies from various application domains
  - Netflix case study
  - Class projects

Course Structure

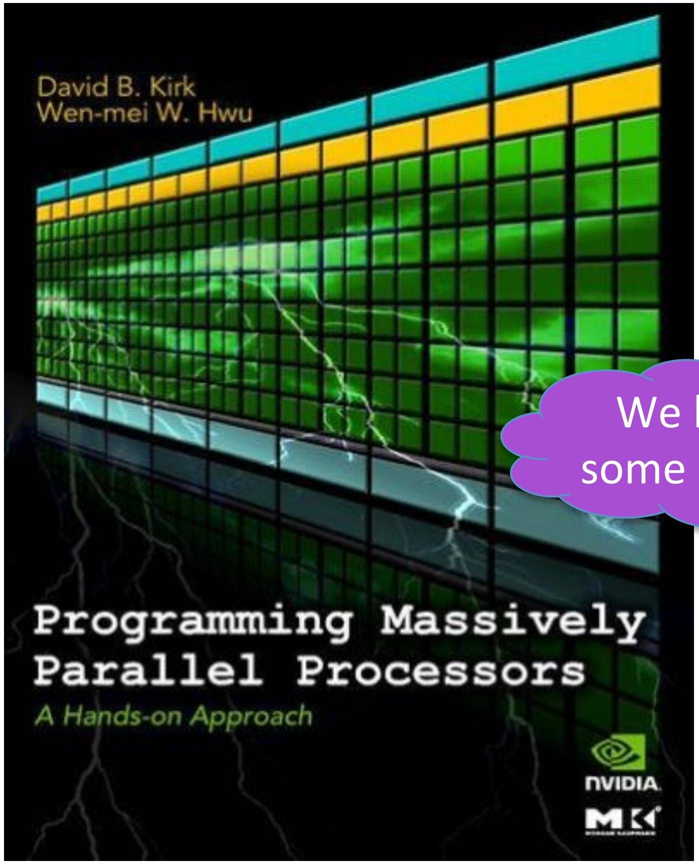
- Pre-requisites: C/C++, Python/Java/R/Matlab, OS
- Lectures (Tue Osborne, 4:45 – 7:30 pm)
- Assignments (2 undergrads, 3 graduate)
- Semester Project
- Grading
  - Assignments (30 %)
  - Project Proposal + Final Project (10 + 40 %)
  - Class participation (20 %)

Textbooks

amazon.com



“Machine Learning in Action”
Peter Harrington, Manning Publications
2012



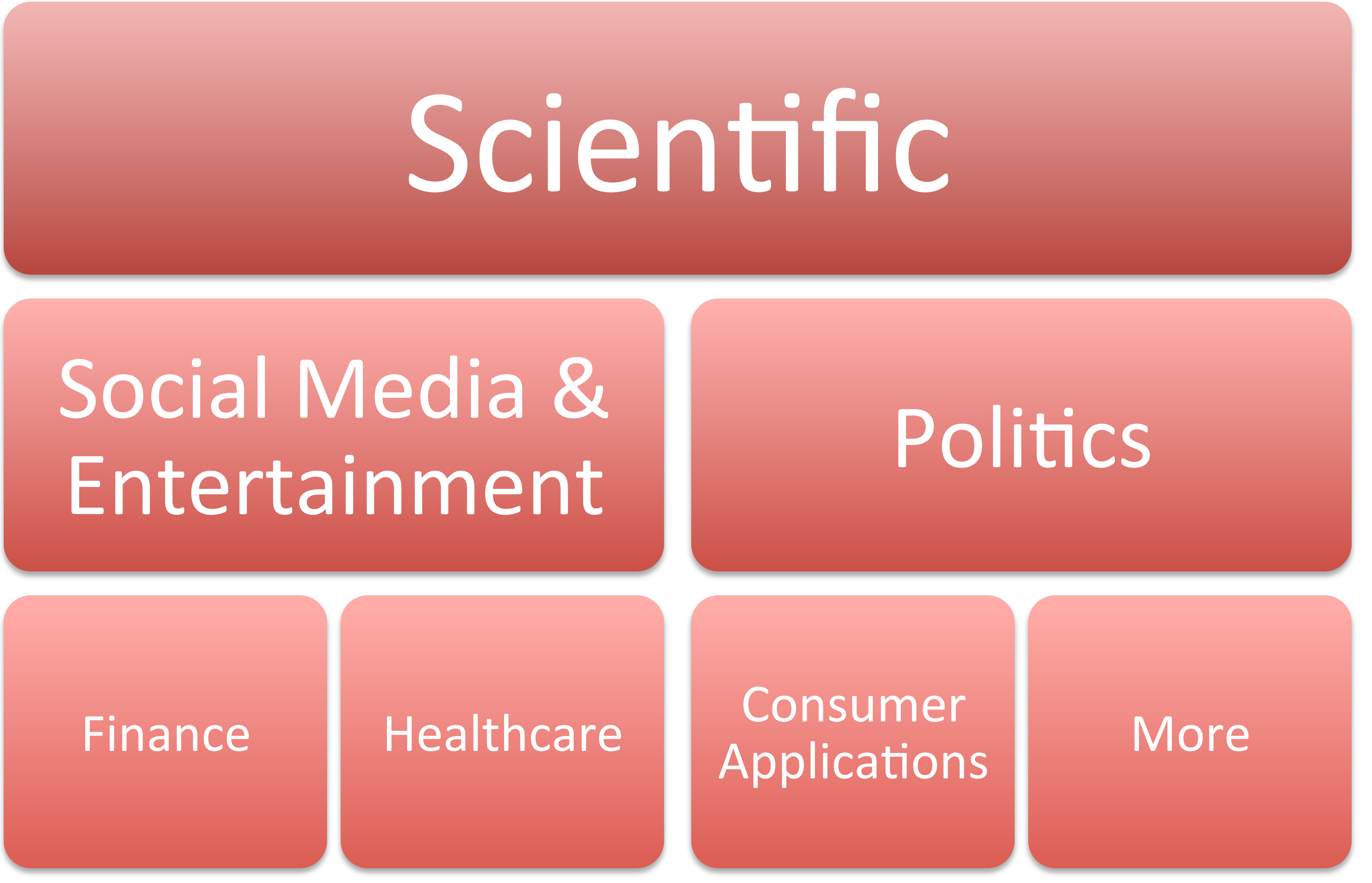
We have some copies

Programming Massively Parallel Processors:
A Hands-on Approach
David Kirk, Wen-mei Hwu
Morgan Kaufmann 2nd Edn 2012

Resources

- We have some GPUs
- Will be installed on machines on campus/ students will be able to access (more details in subsequent lectures)
- Programming
  - C/C++ for CUDA assignments
  - For class projects: C/C++, Java, Python, R etc.

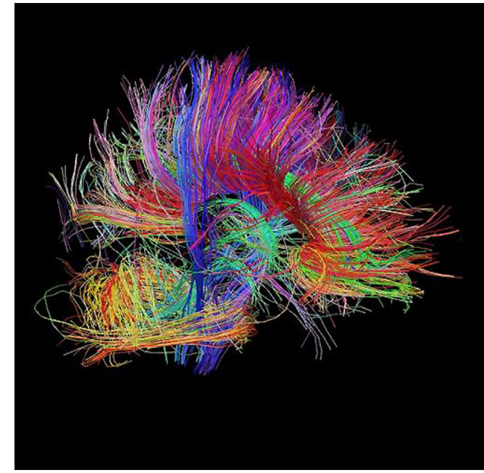
Application domains and Case Studies



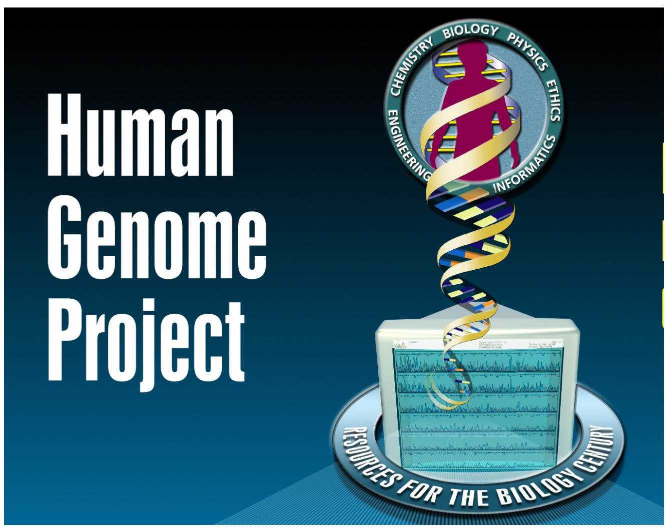
Scientific Instruments

Murchinson Widefield Array
Data produced at the rate of 16 GB/s

Scientific instruments
(collecting all sorts of data)



Human Connectome Project
12 TB total data, 100M new Photographs per day

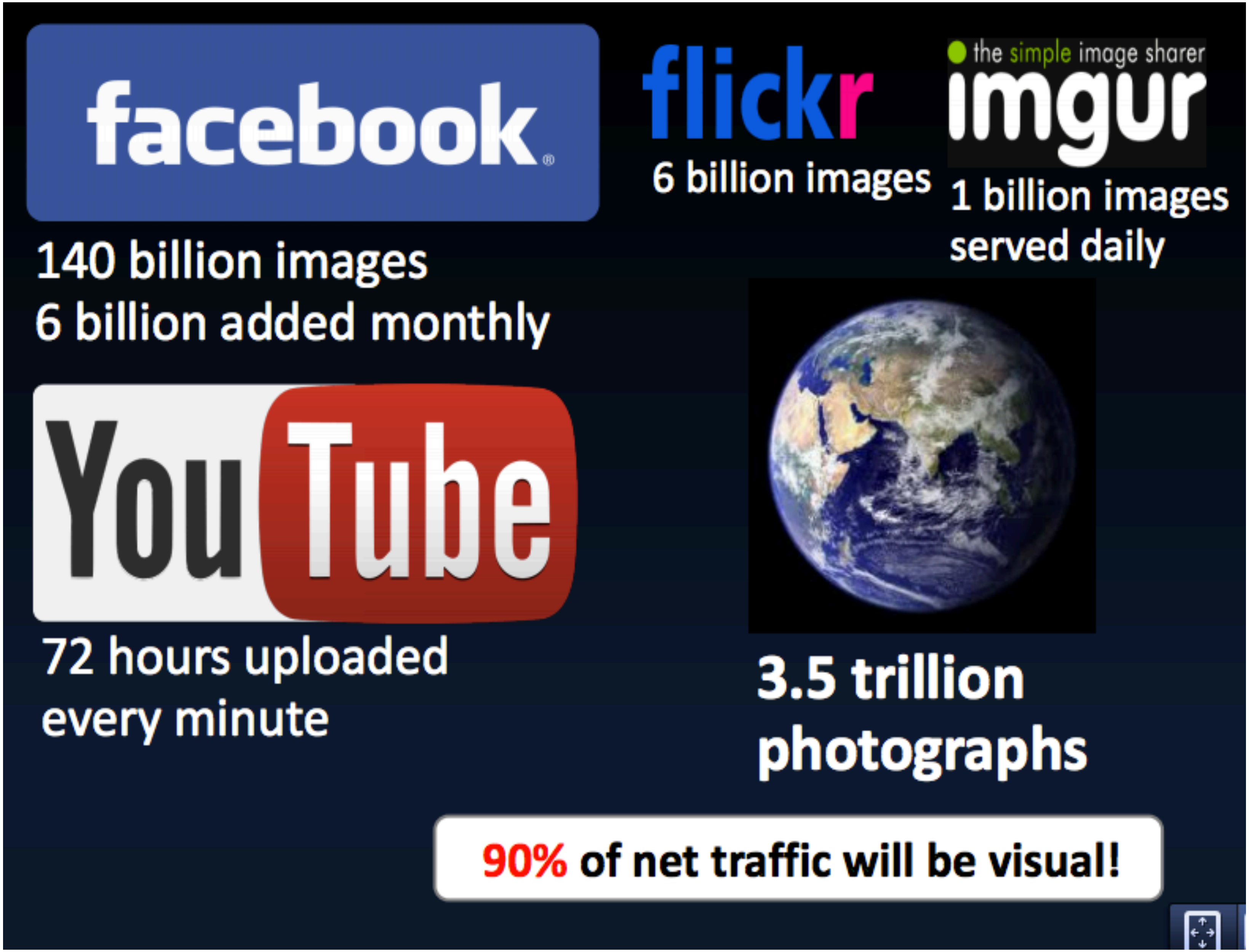


514 Petabytes, Annotations 17 TB

Large Hadron Collider:
30 PB annually, stored on 83K physical disks

Processing power of about 110 Petabytes

Social Media and Images/Video



Retailers..

Consumer Products Companies

Big Box Stores

Why are they collecting all this data?

Target Marketing

- To send you catalogs for exactly the merchandise you typically purchase.
- To suggest medications that precisely match your medical history.
- To "push" television channels to your set instead of your "pulling" them in.
- To send advertisements on those channels **just for you**!

Targeted Information

- To know what you need before you even know you need it based on past purchasing habits!
- To notify you of your expiring driver's license or credit cards or last refill on a Rx, etc.
- To give you turn-by-turn directions to a shelter in case of emergency.

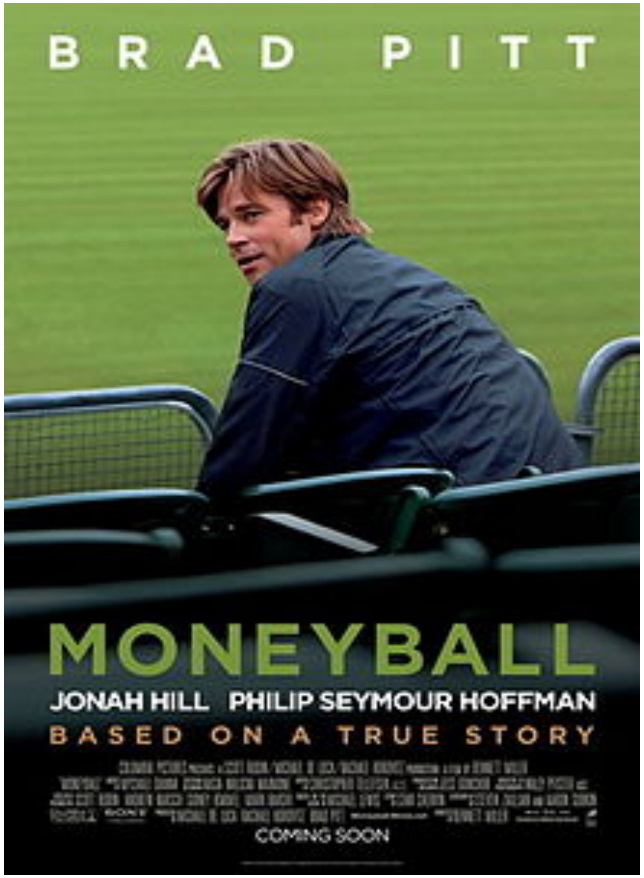
Big Data and Sports

- Moneyball: The Art of Winning an Unfair Game

Oakland Athletics baseball team and its general manager Billy Beane

- Oakland A's' front office took advantage of more analytical gauges of player performance to field a team that could compete successfully against richer competitors in MLB

- Oakland approximately $41 million in salary,
New York Yankees, $125 million in payroll that same season.
Oakland is forced to find players undervalued by the market,

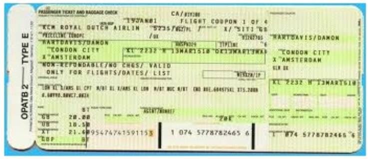


Finance

Credit Card Companies

What data are they getting?

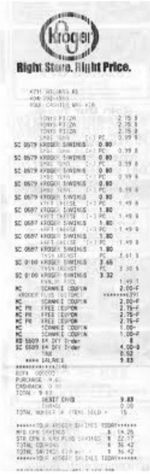
Airline ticket



Restaurant check

1	Cab Sauv Corvina	£15.25
1	Cab Sauv Glass	£4.00
26	Net Sales Total	£73.45
1	SUCK MY D[illegible] F[illegible] FACE	£0.00
1	FISH CAKES	£4.95
2	1/2 Wings Starter	£7.90
1	MELON AND PARMA HAM	£3.95
1	Calamari	£4.95
1	Garlic brd starter	£2.25
3	Meatball Starter	£17.85
1	Aub & Feta Starter	£3.95

Grocery Bill



Hotel Bill

Usage Example in Big Data

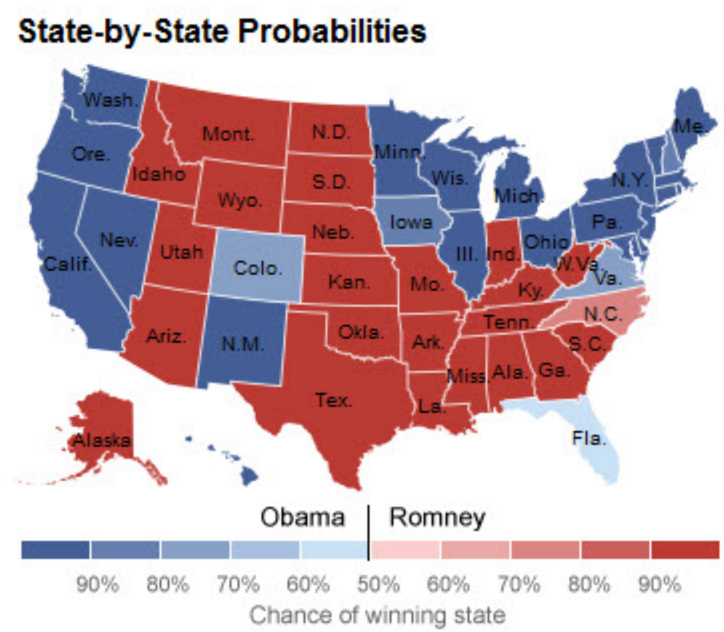
Data Analysis prediction for US 2012 Election

Drew Linzer, June 2012
332 for Obama,
206 for Romney

media continue reporting the race as very tight

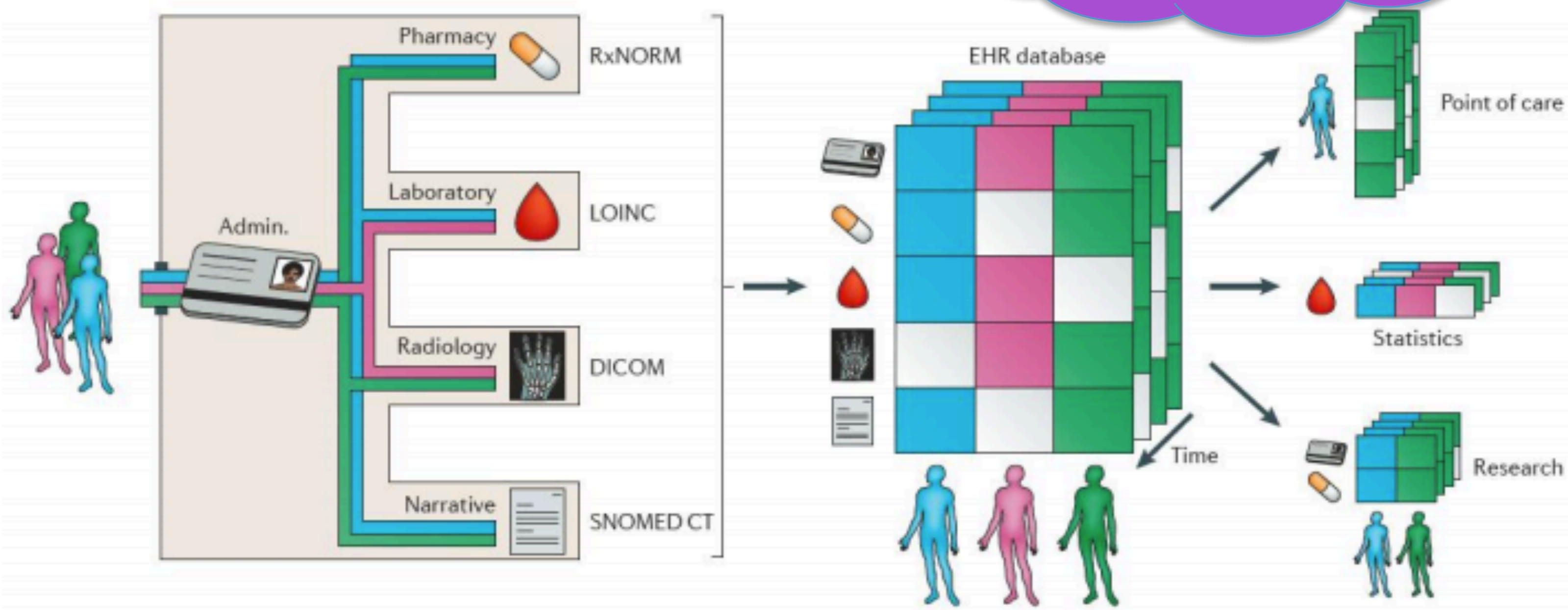
Nate Silver's, Five thirty Eight blog
Predict Obama had a 86% chance of winning
Predicted all 50 state correctly

Sam Wang, the Princeton Election Consortium
The probability of Obama's re-election
at more than 98%



Data Collection and Analysis

Healthcare Analytics

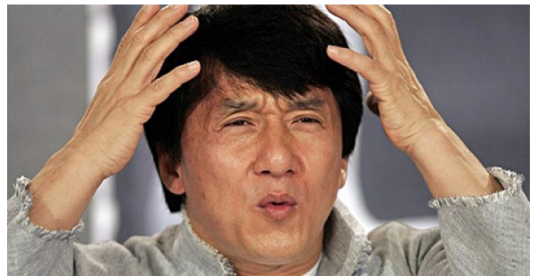


Effectively integrating and efficiently analyzing various forms of healthcare data over a period of time can answer many of the impending healthcare problems.

Jensen, Peter B., Lars J. Jensen, and Søren Brunak. "Mining electronic health records: towards better research applications and clinical care." *Nature Reviews Genetics* (2012).

How Can You Avoid *Big Data*?

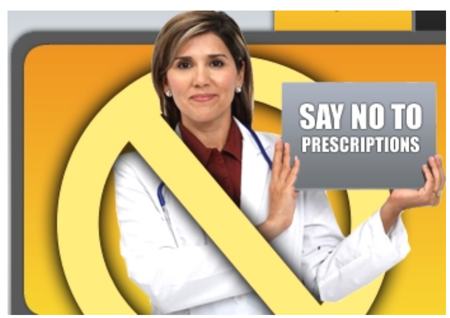
Always pay in cash

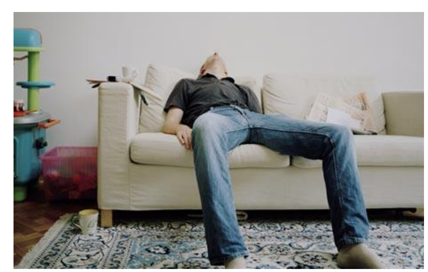


You make no sense

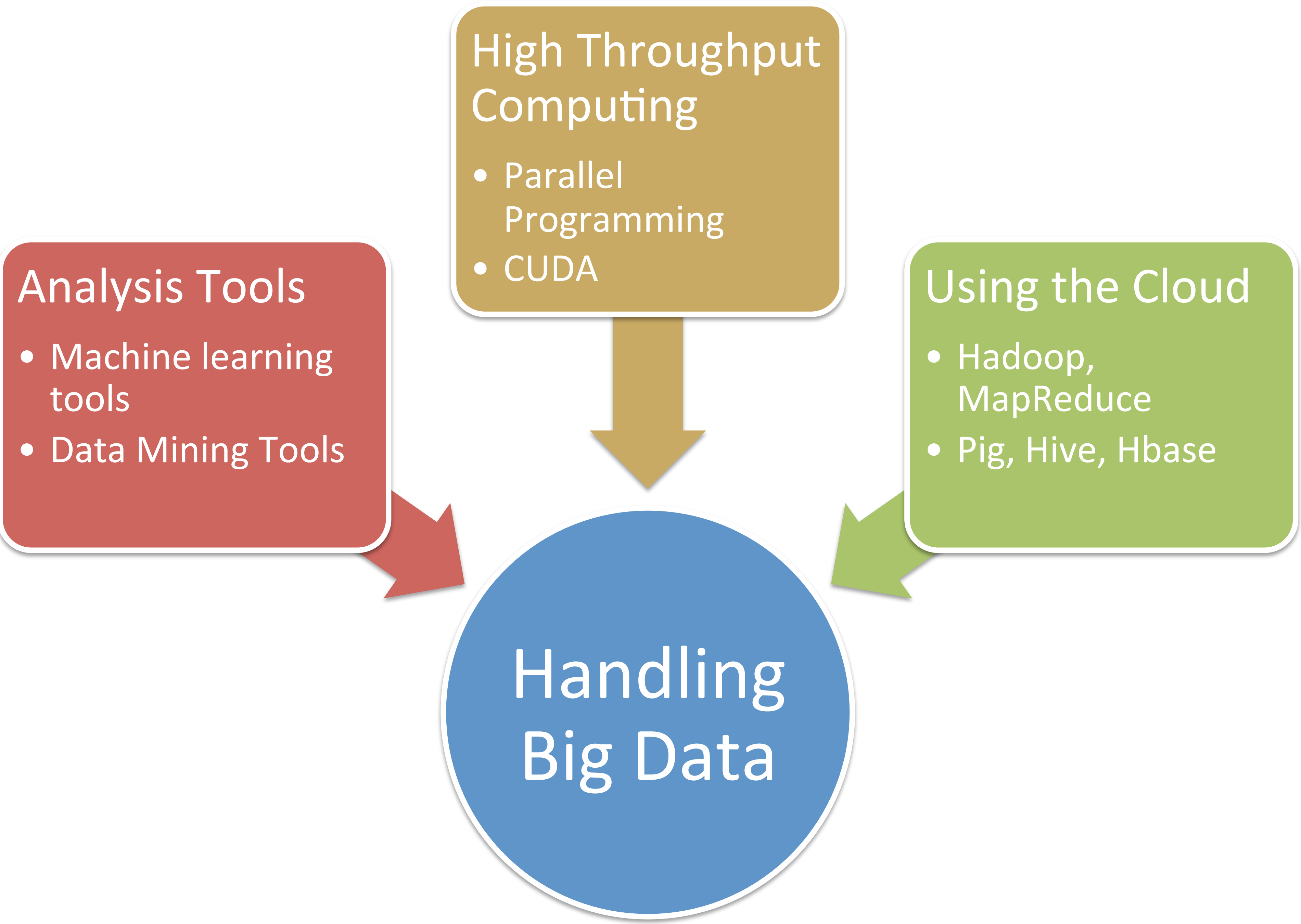
Never go online

Never use club cards





Stay home all day


High Throughput Computing
Parallel Programming
CUDA
Analysis Tools
Machine learning tools
Data Mining Tools
Using the Cloud
Hadoop, MapReduce
Pig, Hive, Hbase
Handling Big Data

scikits learn

machine learning in Python

Python

WEKA
The University of Waikato

SHOGUN 3.1.1
In $DEITY we trust all others bring data.

Python

Java

fastlab

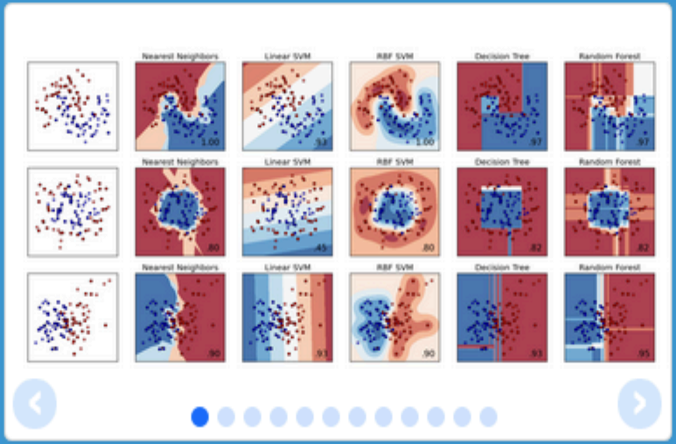
mlpack
a scalable c++ machine learning library

scikit learn

- Library of Machine Learning models
- Simple **`fit`** / **`predict`** / **`transform`** API
- Python / NumPy / SciPy / Cython

  & wrappers for libsvm / liblinear

- Model Assessment, Selection & Ensembles
- Some support for multi-core



scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying to which set of categories a new observation belong to.

Applications: Spam detection, Image recognition.
Algorithms: *SVM*, *nearest neighbors*, *random forest*, ...

— *Examples*

Regression

Predicting a continuous value for a new example.

Applications: Drug response, Stock prices.
Algorithms: *SVR*, *ridge regression*, *Lasso*, ...

— *Examples*

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: *k-Means*, *spectral clustering*, *mean-shift*, ...

— *Examples*

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency
Algorithms: *PCA*, *Isomap*, *non-negative matrix factorization*.

— *Examples*

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning
Modules: *grid search*, *cross validation*, *metrics*.

— *Examples*

Preprocessing

Feature extraction and normalization.

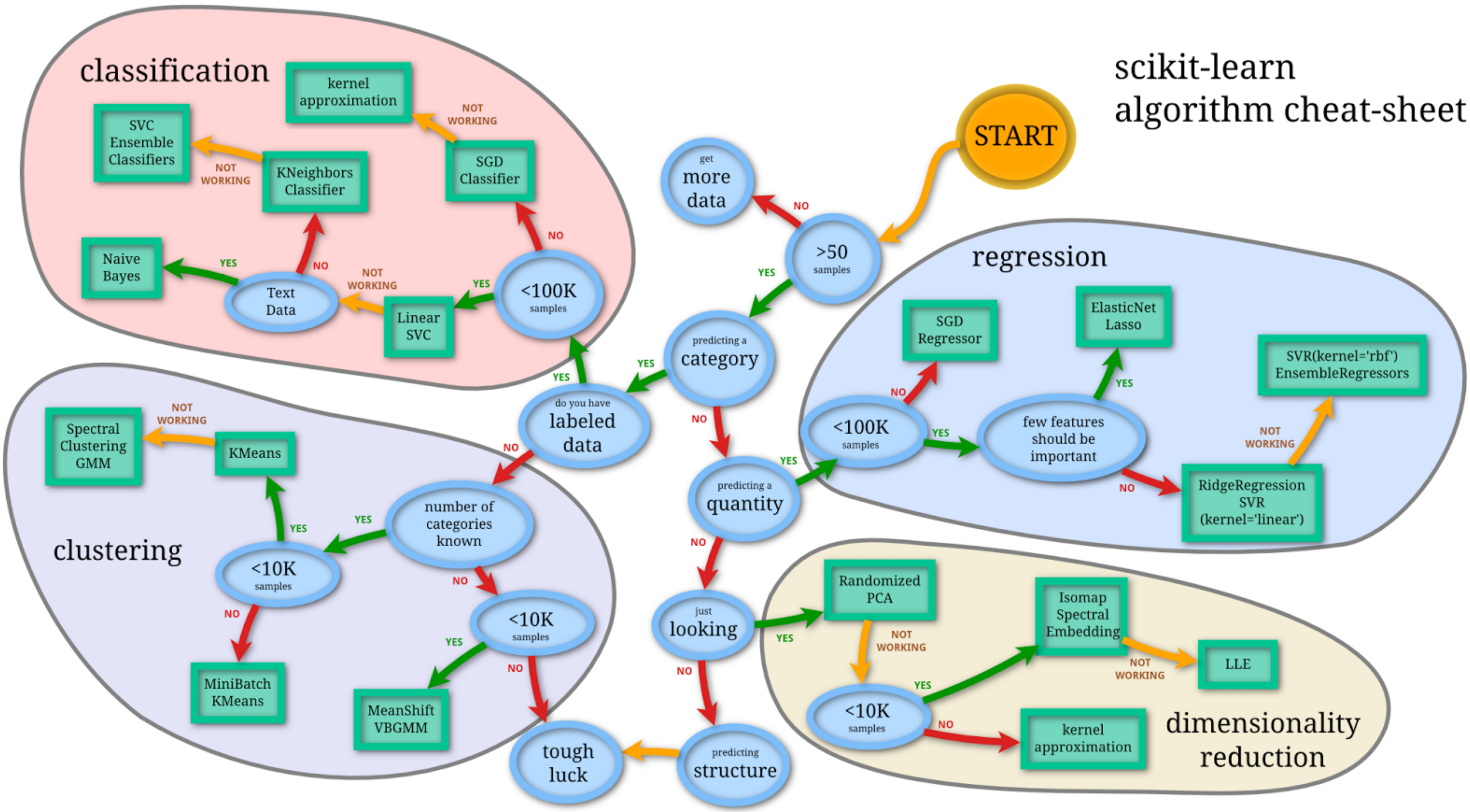
Application: Transforming input data such as text for use with machine learning algorithms.
Modules: *preprocessing*, *feature extraction*.

— *Examples*

Use off-the-shelf tools: Consistent API

```
from sklearn.svm import SVC

clf = SVC()
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
```

```
from sklearn.ensemble import RandomForestClassifier

clf = RandomForestClassifier()
clf.fit(X_train, y_train)

y_pred = clf.predict(X_test)
```

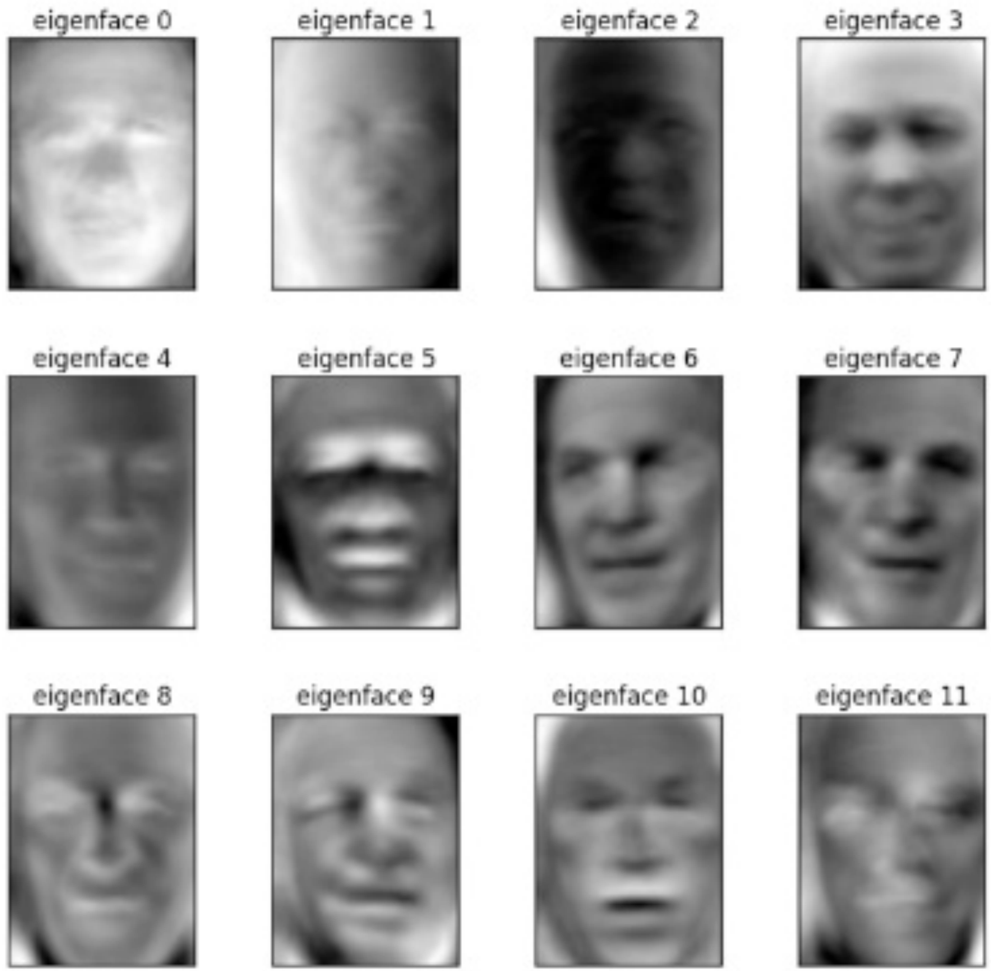


scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality reduction
get more data
>50 samples
predicting a category
do you have labeled data
predicting a quantity
just looking
predicting structure
tough luck
<100K samples
SGD Classifier
kernel approximation
Linear SVC
Text Data
Naive Bayes
KNeighbors Classifier
SVC Ensemble Classifiers
SGD Regressor
ElasticNet Lasso
few features should be important
RidgeRegression SVR (kernel='linear')
SVR(kernel='rbf') EnsembleRegressors
number of categories known
<10K samples
KMeans
Spectral Clustering GMM
MiniBatch KMeans
MeanShift VBGMM
Randomized PCA
Isomap Spectral Embedding
LLE
kernel approximation
YES
NO
NOT WORKING

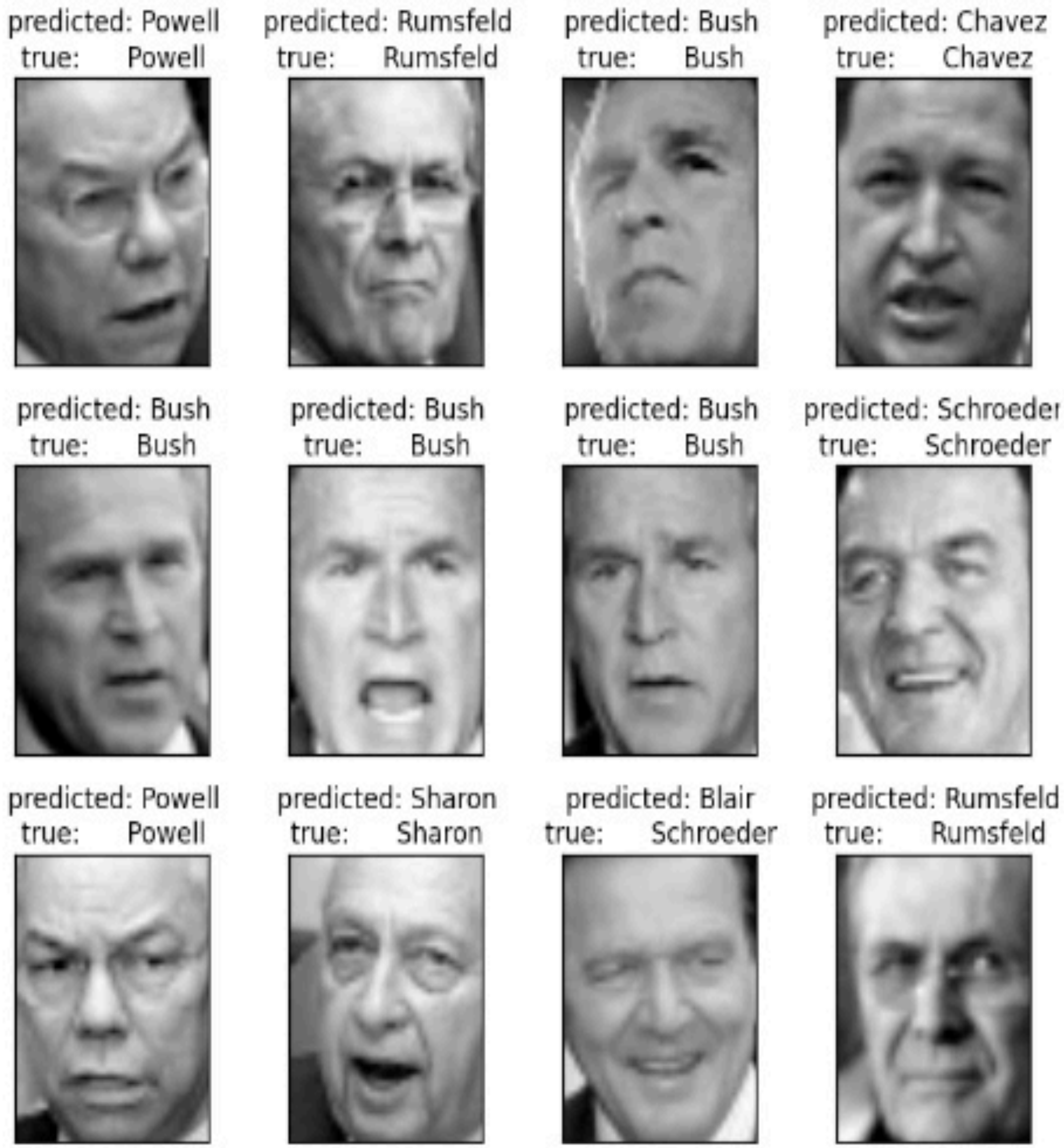
Face Recognition in 5 mins..!

```
Total dataset size:
n_samples: 1288, n_features: 1850, n_classes: 7

Extracting the top 150 eigenfaces from 966 faces
done in 0.466s

Projecting the input data on the eigenfaces orthonormal basis
done in 0.056s

Fitting the SVM classifier to the training set
done in 18.549s

Predicting people's names on the test set
done in 0.062s
                   precision    recall  f1-score   support

     Ariel Sharon       0.90      0.75      0.82        12
     Colin Powell       0.78      0.94      0.85        62
  Donald Rumsfeld       0.86      0.72      0.78        25
    George W Bush       0.89      0.96      0.92       141
Gerhard Schroeder       0.92      0.74      0.82        31
      Hugo Chavez       0.90      0.53      0.67        17
       Tony Blair       0.81      0.74      0.77        34

      avg / total       0.86      0.86      0.86       322
```

Learned Eigen Faces



Predicting…


predicted: Powell
true: Powell
predicted: Rumsfeld
true: Rumsfeld
predicted: Bush
true: Bush
predicted: Chavez
true: Chavez
predicted: Bush
true: Bush
predicted: Bush
true: Bush
predicted: Bush
true: Bush
predicted: Schroeder
true: Schroeder
predicted: Powell
true: Powell
predicted: Sharon
true: Sharon
predicted: Blair
true: Schroeder
predicted: Rumsfeld
true: Rumsfeld

High Throughput Computing

Moore's Law

Motivation

Share

- The most economic number of components in an IC will double every year
- Historically – CPUs get faster

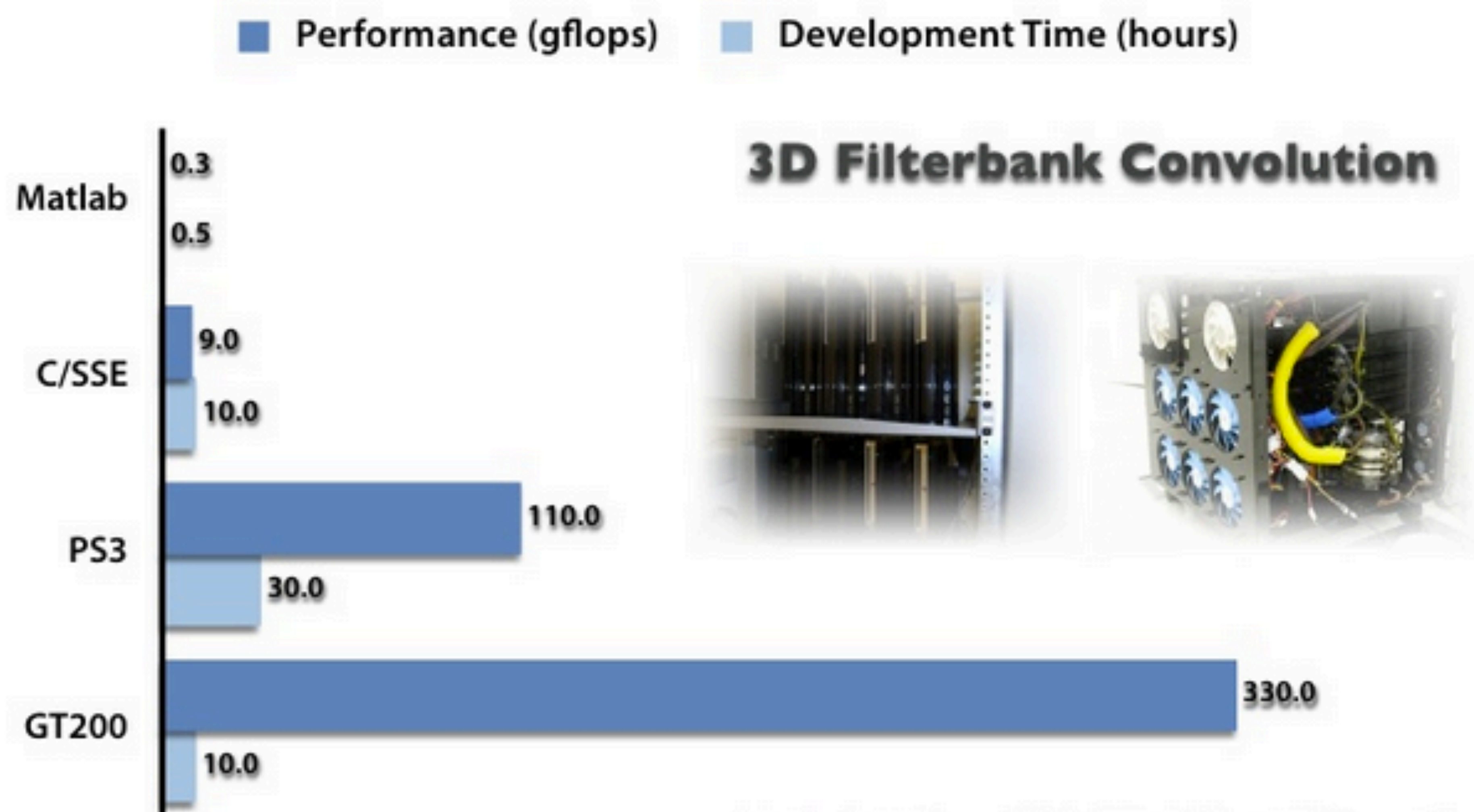
→ Hardware reaching frequency limitations

- Now – CPUs get wider

slide by Matthew Bolitho

GPUs are REALLY fast

GPU?



Nicolas Pinto, James DiCarlo, David Cox (MIT, Harvard)

Why so fast?

GPU?

Share

- Designed for math-intensive, parallel problems:

- More transistors dedicated to ALU than flow control and data cache

slide by Matthew Bolitho

Is it free?

- What are the consequences?
- Program must be more predictable:
  - Data access coherency
  - Program flow

slide by Matthew Bolitho

Parallel Computing

GPU?

- Rather than expecting CPUs to get twice as fast, expect to have twice as many!

- Parallel processing for the masses
- Unfortunately: Parallel programming is hard.

→ Algorithms and Data Structures must be fundamentally redesigned

slide by Matthew Bolitho

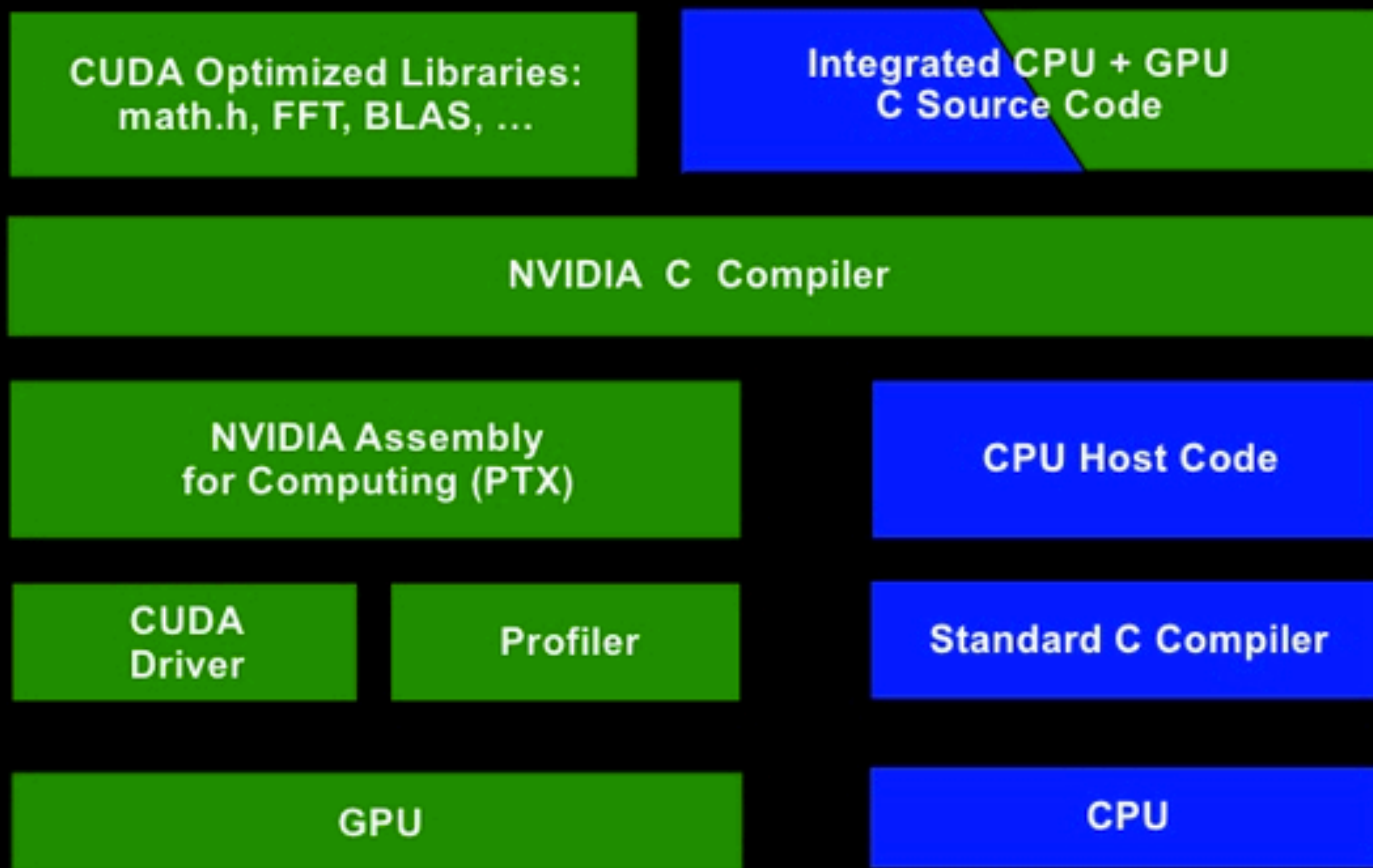
CPU vs. GPU

- CPU
  - Really fast caches (great for data reuse)
  - Fine branching granularity
  - Lots of different processes/threads
  - High performance on a single thread of execution
- GPU
  - Lots of math units
  - Fast access to onboard memory
  - Run a program on each fragment/vertex
  - High throughput on parallel tasks

- **CPUs are great for *task* parallelism**
- **GPUs are great for *data* parallelism**

slide by Mike Houston

CUDA Software Development



nvision 08
THE WORLD OF VISUAL COMPUTING

NVIDIA


BIG DATA
CLOUD

What exactly is hadoop ?

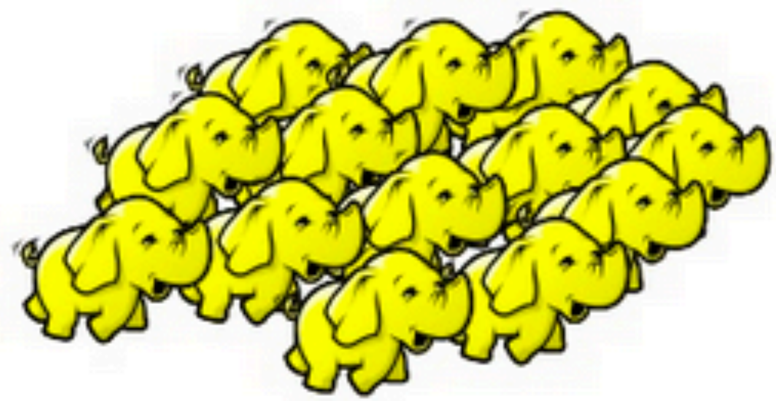
- Actually a growing collection of subprojects; focus on two right now

An overview of Hadoop Map-Reduce

Traditional Computing

(one computer)

Hadoop

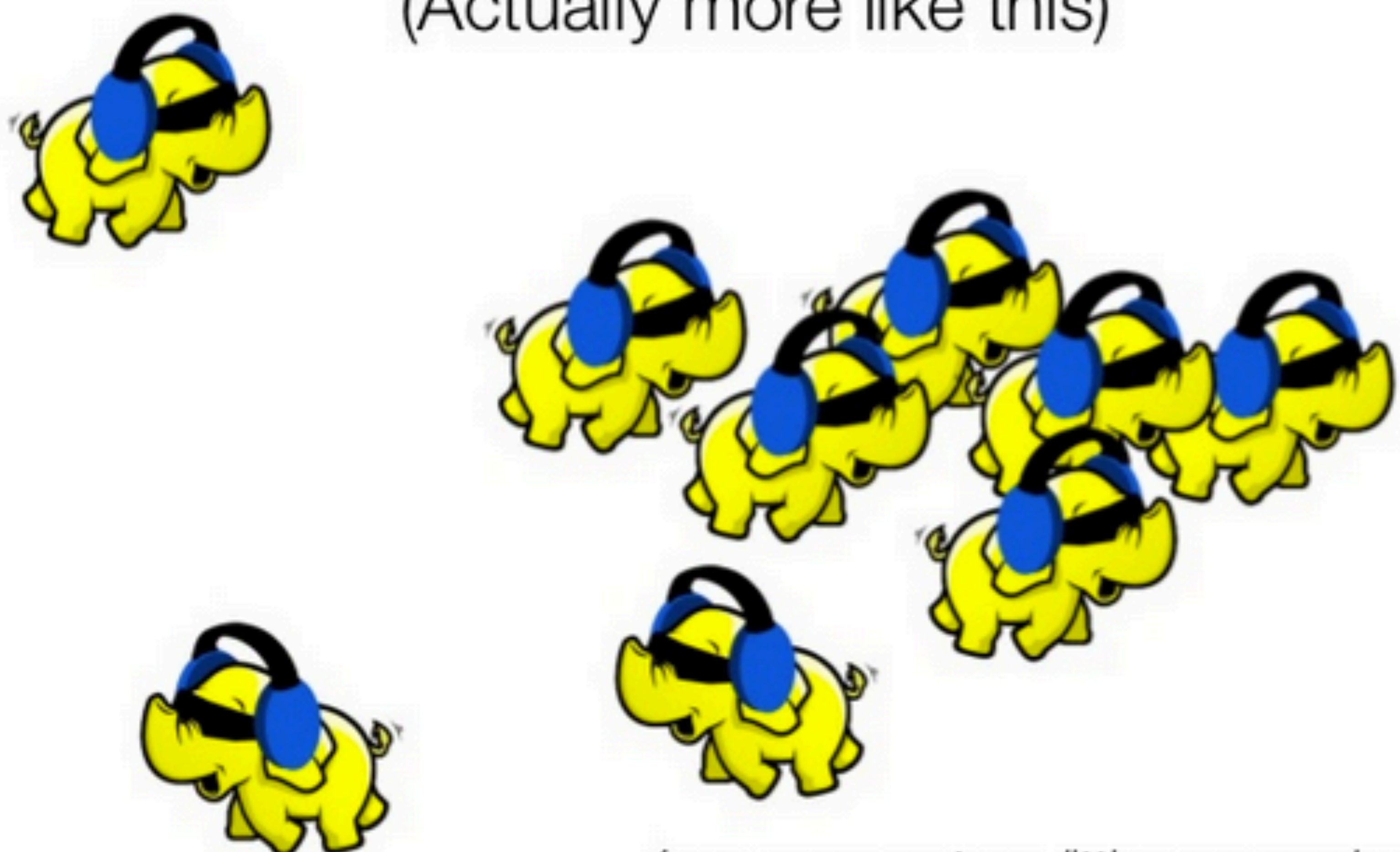


(many computers)

Share

An overview of Hadoop Map-Reduce

(Actually more like this)

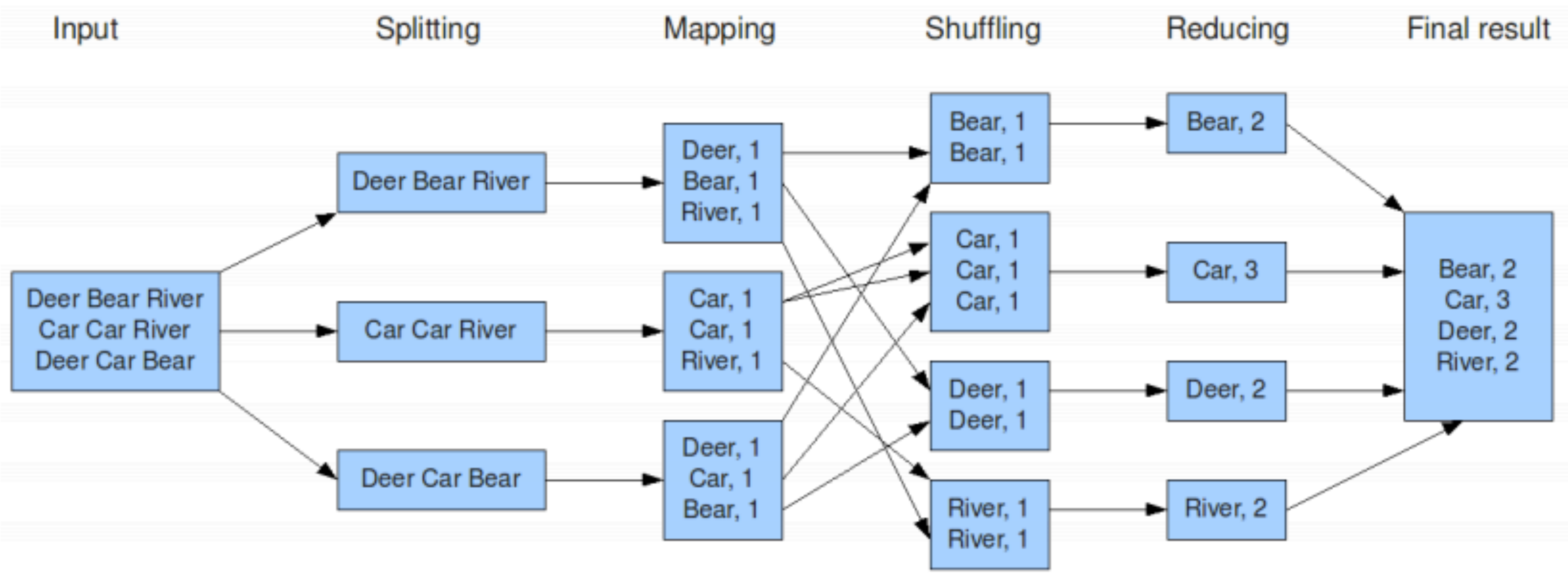


(many computers, little communication, stragglers and failures)

Map-Reduce: Three phases

1. Map
2. Sort
3. Reduce

The overall MapReduce word count process



Map-Reduce: Imagine word-count on the Web

MapReduce: Simplified Data Processing on Large Clusters

Jeffrey Dean and Sanjay Ghemawat

jeff@google.com, sanjay@google.com

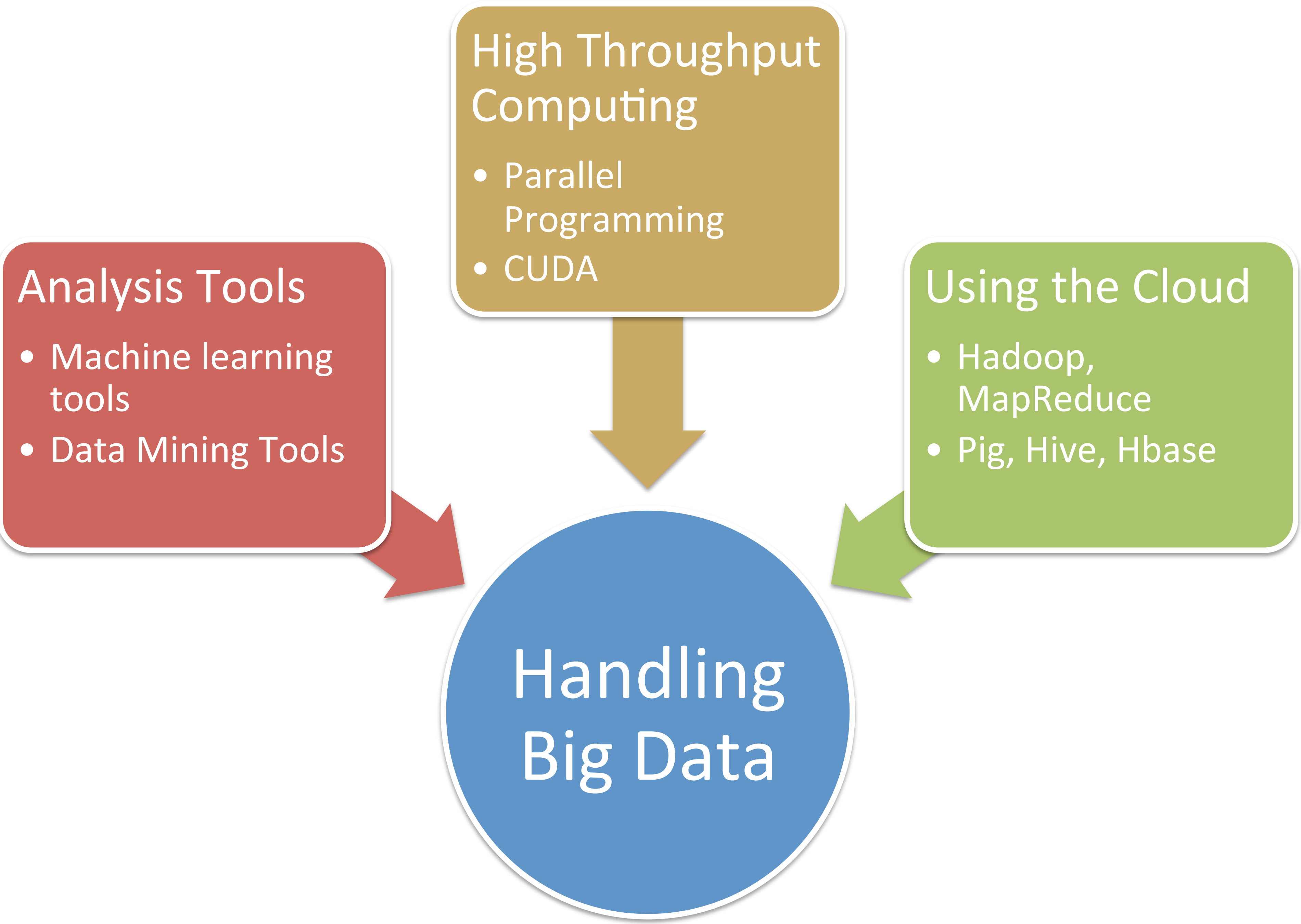
Google, Inc.

Abstract

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a *map* function that processes a key/value pair to generate a set of intermediate key/value pairs, and a *reduce* function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, as shown in the paper.

given day, etc. Most such computations are conceptually straightforward. However, the input data is usually large and the computations have to be distributed across hundreds or thousands of machines in order to finish in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures conspire to obscure the original simple computation with large amounts of complex code to deal with these issues.

As a reaction to this complexity, we designed a new


High Throughput Computing
• Parallel Programming
• CUDA
Analysis Tools
• Machine learning tools
• Data Mining Tools
Using the Cloud
• Hadoop, MapReduce
• Pig, Hive, Hbase
Handling Big Data

Always use the right tool !

Share

Office Season 2, Episode 27: Conflict Resolution)


Stay Tuned.
Coming Soon!