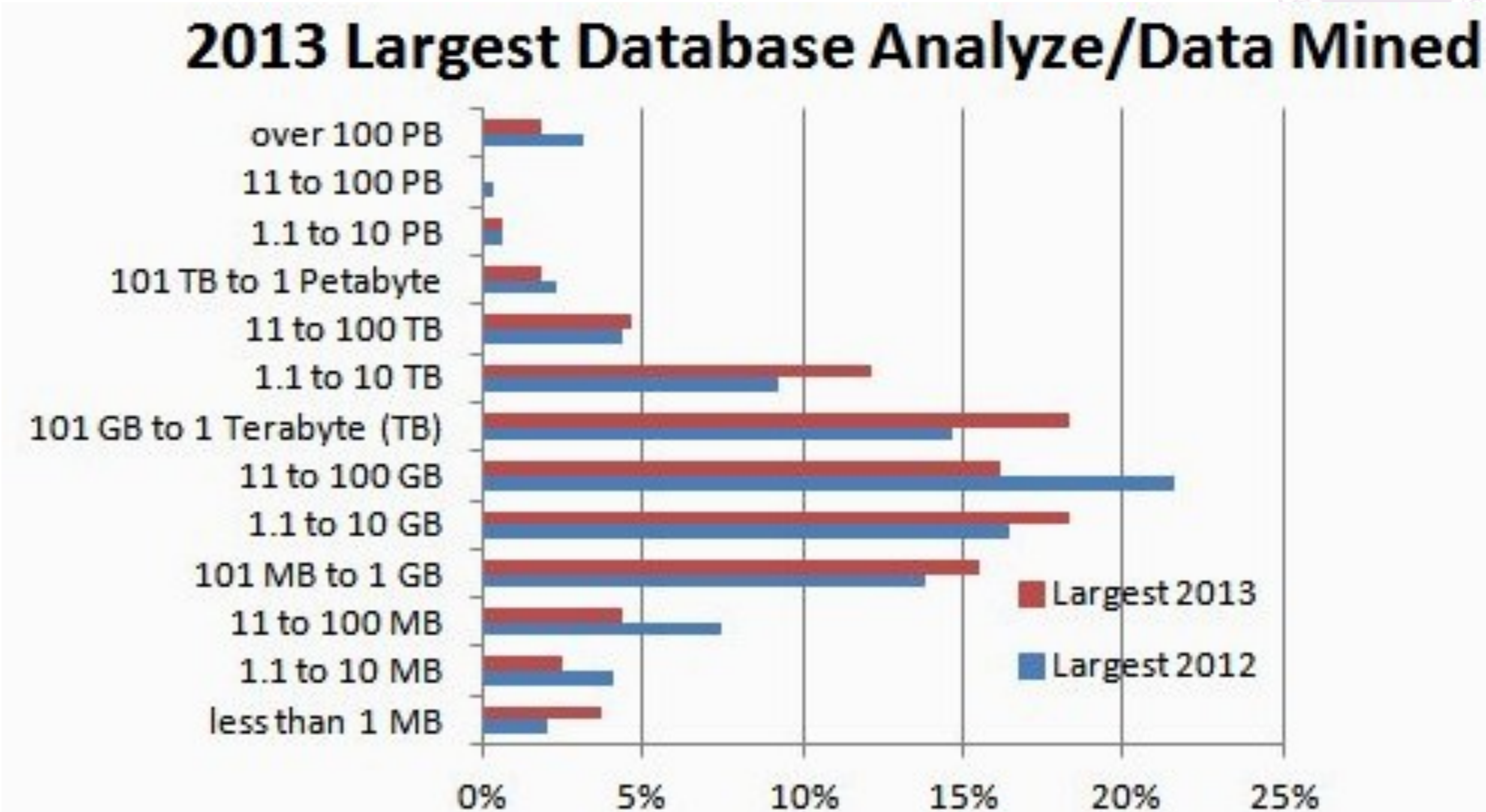
Lecture 10: Data Mining II
Handling Large-Scale Data

http://cs.nju.edu.cn/yuy/course_dm14ms.ashx

How large the data can be



[KDnuggets Poll, 2013]

What’s wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

Why large-scale data matters

Why large-scale data matters

recall from the learning theory:

$$\text{with probability at least } 1-\delta$$

$$\epsilon_g < \epsilon_t + \sqrt{\frac{1}{m}(\ln|\mathcal{H}| + \ln\frac{1}{\delta})}$$

the number of examples

Why matters

Confusion set disambiguation task

He is tallest ____ the students.
A. among
B. between

feature: the set of words in a window of the blank
memory-based: the before and the after words

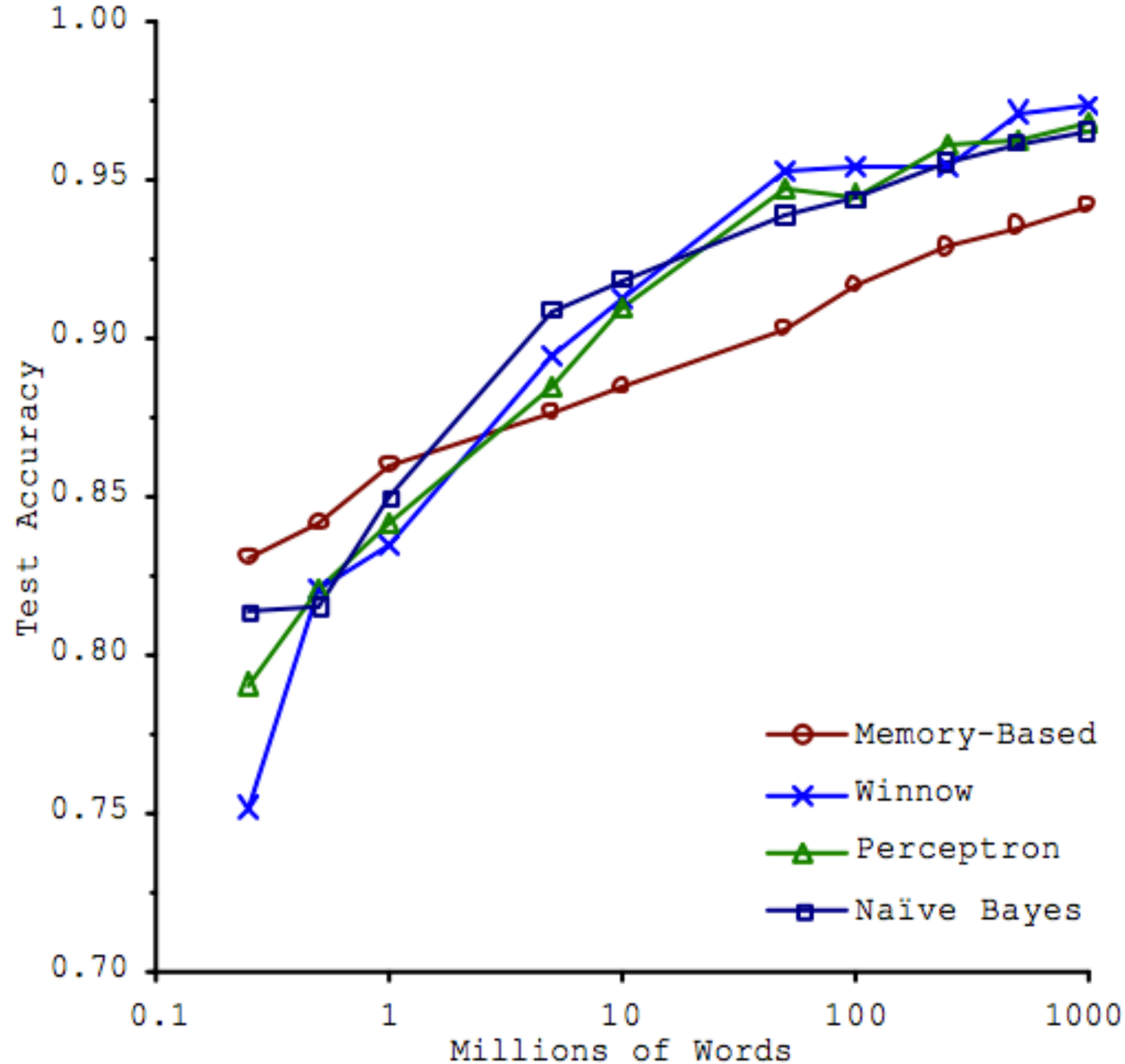


Figure 1. Learning Curves for Confusion Set Disambiguation

[Banko and Brill, ACL01]

What's wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

A: No, the data set size strongly related to the analysis quality

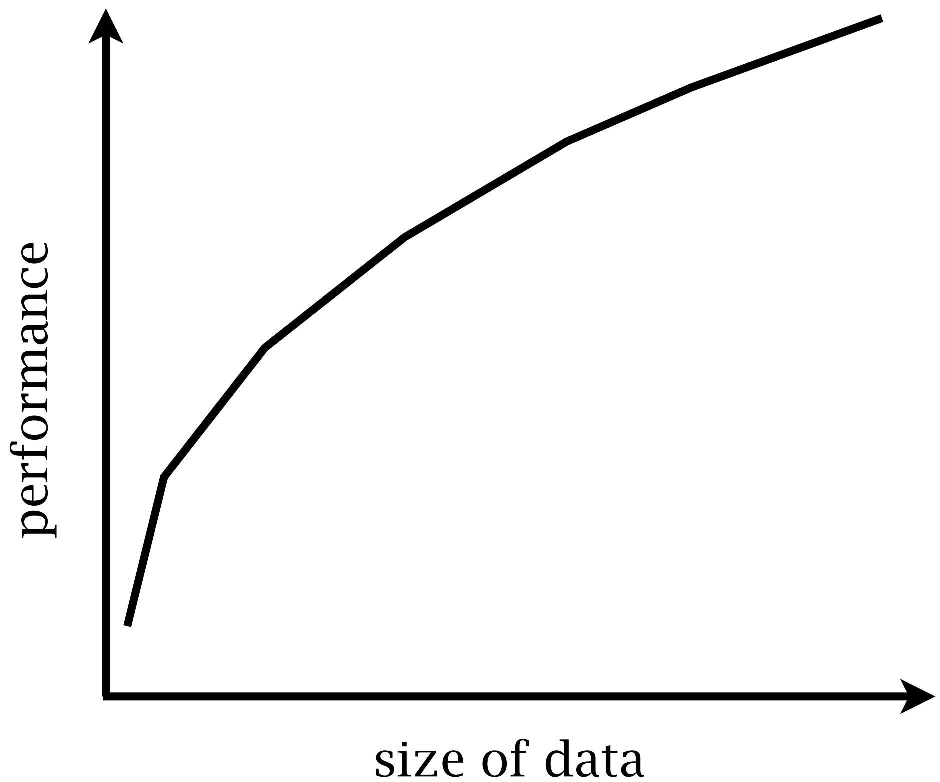
What's wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

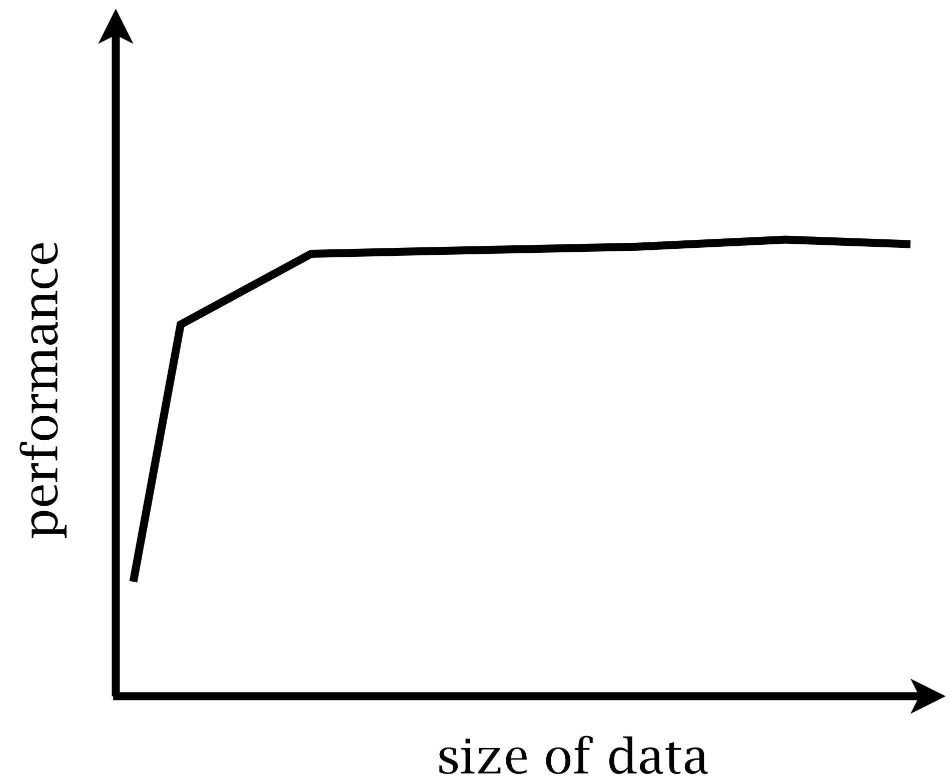
A: No, the data set size strongly related to the analysis quality

Q: Is that always true?

Sampling - check the “largeness”



real large



fake large

use a small sample of data is sufficient

What's wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

A: No, the data set size strongly related to the analysis quality

Q: Is that always true?

A: No, we should check if the data is really large.

What's wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

A: No, the data set size strongly related to the analysis quality

Q: Is that always true?

A: No, we should check if the data is really large.

Q: What's the difficulties in real large data?

What's wrong with large-scale data?

Q: Can we sample a small subset out of the data and analyze the subset?

A: No, the data set size strongly related to the analysis quality

Q: Is that always true?

A: No, we should check if the data is really large.

Q: What's the difficulties in real large data?

Time and Space

Alleviate the time difficulty

- Use simple & fast algorithms

- Accelerate algorithms
  - Online/one-pass algorithms
  - Better data structures
  - Randomization and aggregation

- Parallelize algorithms

Using simple algorithms

Algorithms that run fast

- Naive Bayes classifiers
- Decision trees
- Linear classifiers (without kernel)
  - LibSVM/LibLinear

Online/One-pass algorithms

Batch learning

build a model from a batch of examples

Online learning

examples come as a stream

Naive Bayes

Perceptron:

1. $\boldsymbol{w} = 0$

2. for each example

   if $\mathrm{sign}(y\boldsymbol{w}^\top \boldsymbol{x}) < 0$

$$\boldsymbol{w} = \boldsymbol{w} + \eta y \boldsymbol{x}$$

Online/One-pass algorithms

Batch learning

build a model from a batch of examples

Online learning

examples come as a stream

Naive Bayes

Perceptron:

1. $\boldsymbol{w} = 0$

2. for each example
if $\operatorname{sign}(y\boldsymbol{w}^\top \boldsymbol{x}) < 0$

$$\boldsymbol{w} = \boldsymbol{w} + \eta y \boldsymbol{x}$$

gradient ascent

$$\frac{\partial y\boldsymbol{w}^\top \boldsymbol{x}}{\partial \boldsymbol{w}} = y\boldsymbol{x}$$

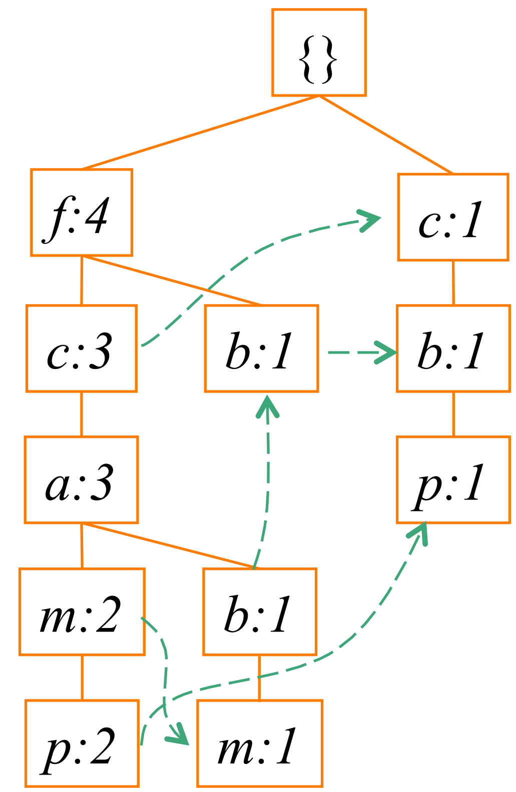
Better data structure

Finding frequent item sets

Apriori

TID	Items
T100	*f, a, c, d, g, i, m, p*
T200	*a, b, c, f, l, m, o*
T300	*b, f, h, j, o, w*
T400	*b, c, k, s, p*
T500	*a, f, c, e, l, p, m, n*

FP-Tree



from table jointing to tree structure

Better data structure

Finding nearest neighbors

brute-force: $O(n)$

kd-tree: $O(\log n)$ on average

cover-tree: $O(\log n)$ [Beygelzimer, et al. ICML'06]

hashing methods for approximate NN search

Locality sensitive hashing

LSH functions: $\mathcal{H} = \{h_{\boldsymbol{r}}\}(\boldsymbol{r} \in \mathbb{B}^n)$ where $h_r(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{r}^\top \boldsymbol{x})$

$$P(h_{\boldsymbol{r}}(\boldsymbol{x}_1) = h_{\boldsymbol{r}}(\boldsymbol{x}_2)) = 1 - \frac{\theta(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\pi}$$

Better data structure

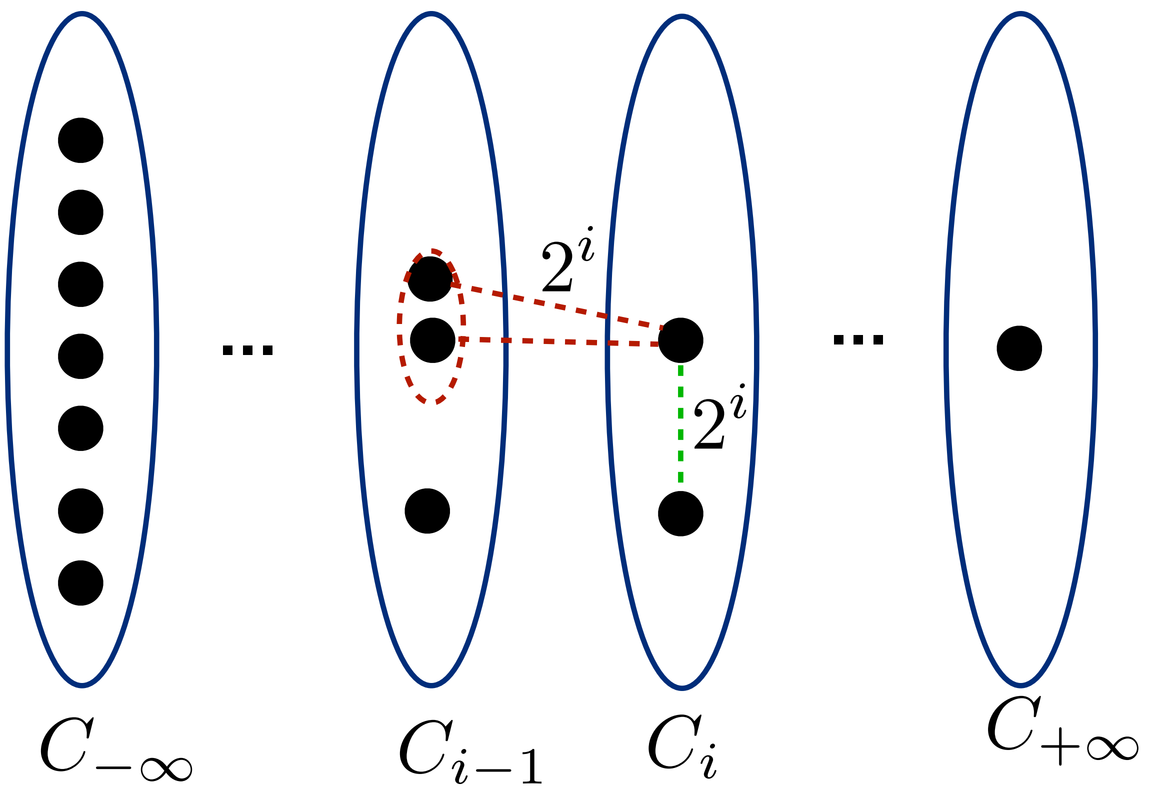
Finding nearest neighbors

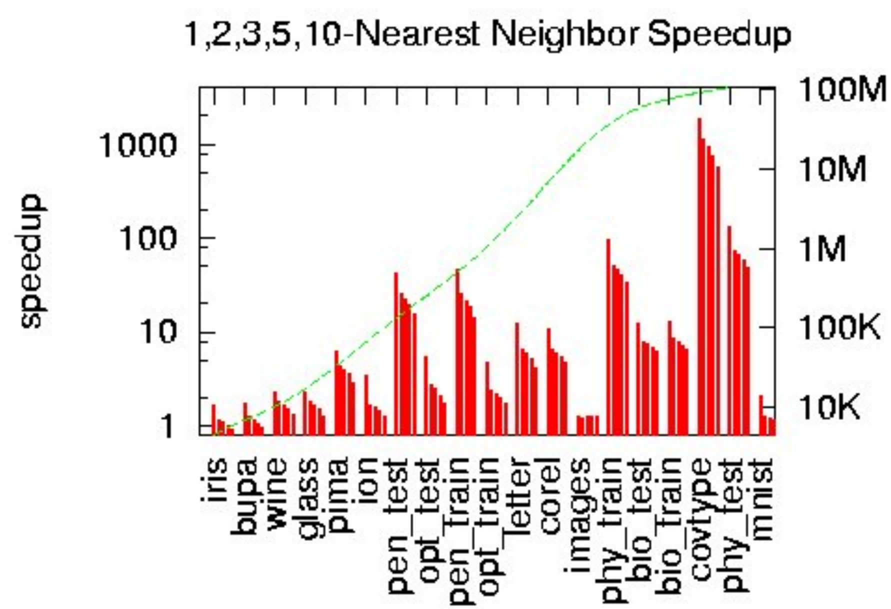
brute-force: $O(n)$

kd-tree: $O(\log n)$ on average

cover-tree: $O(\log n)$

[Beygelzimer, et al. ICML'06]





Randomization and Aggregation

gradient decent

calculate gradient over all examples

stochastic gradient decent (SGD)

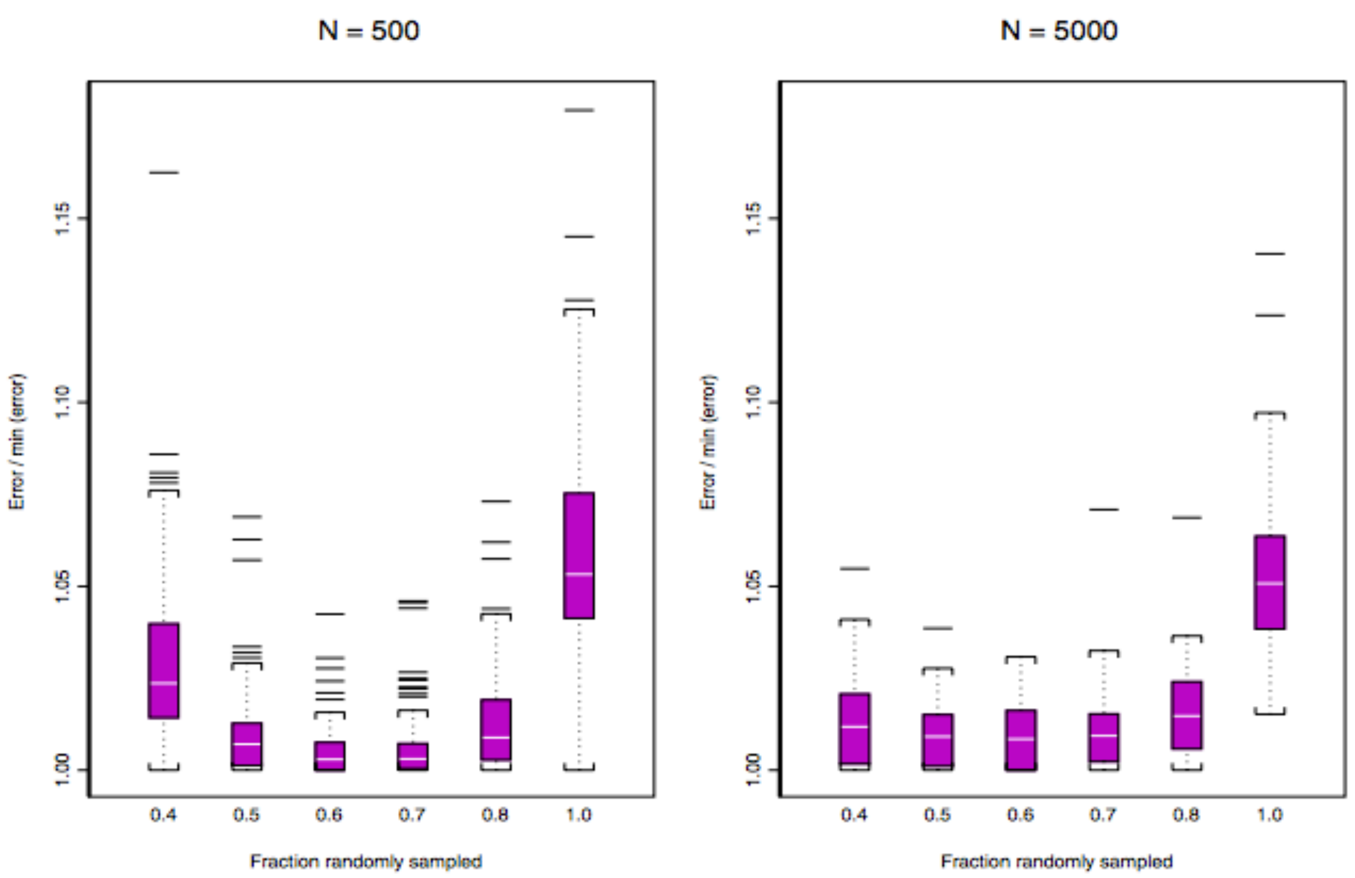
calculate gradient over some examples

average all the intermediate results to reduce variance

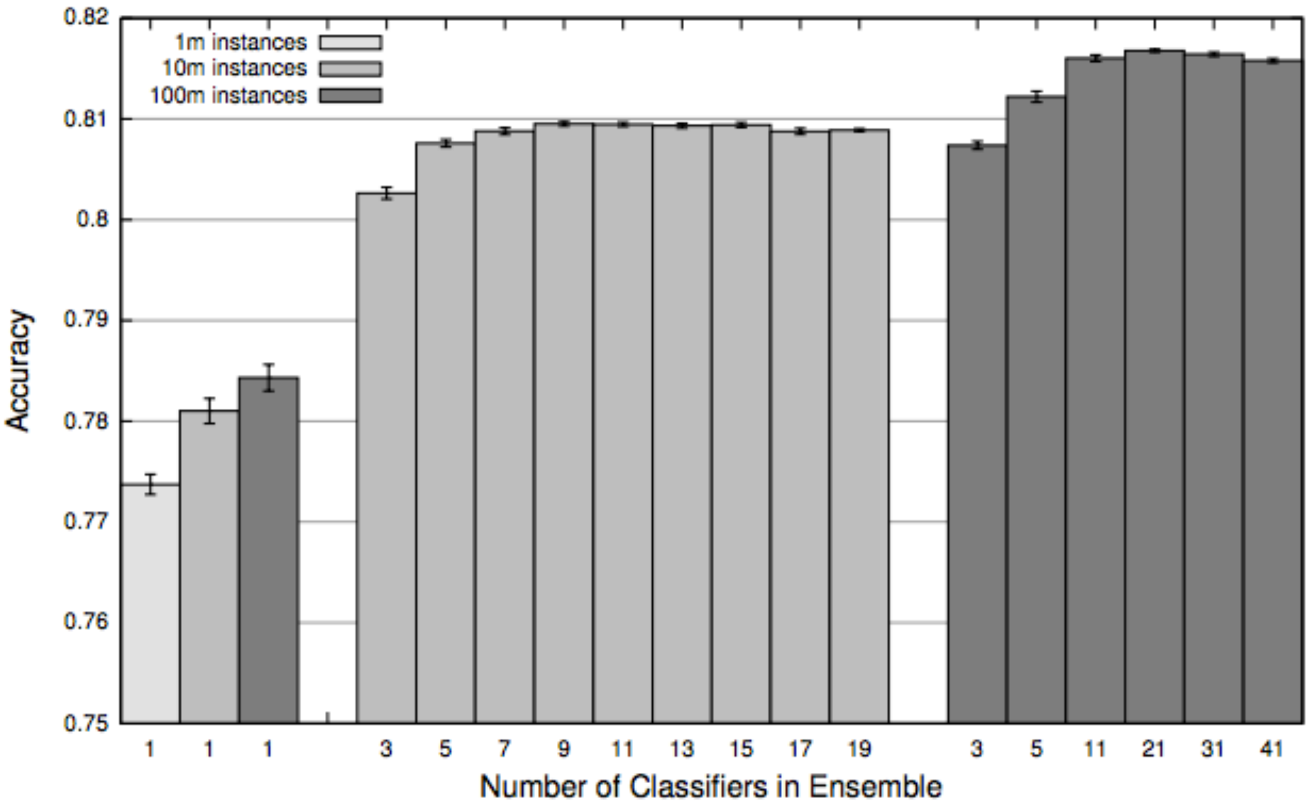
optimal model may not necessarily be optimal

Randomization and Aggregation

stochastic gradient boosting [J. Friedman, JCSDA'02]



Randomization and Aggregation



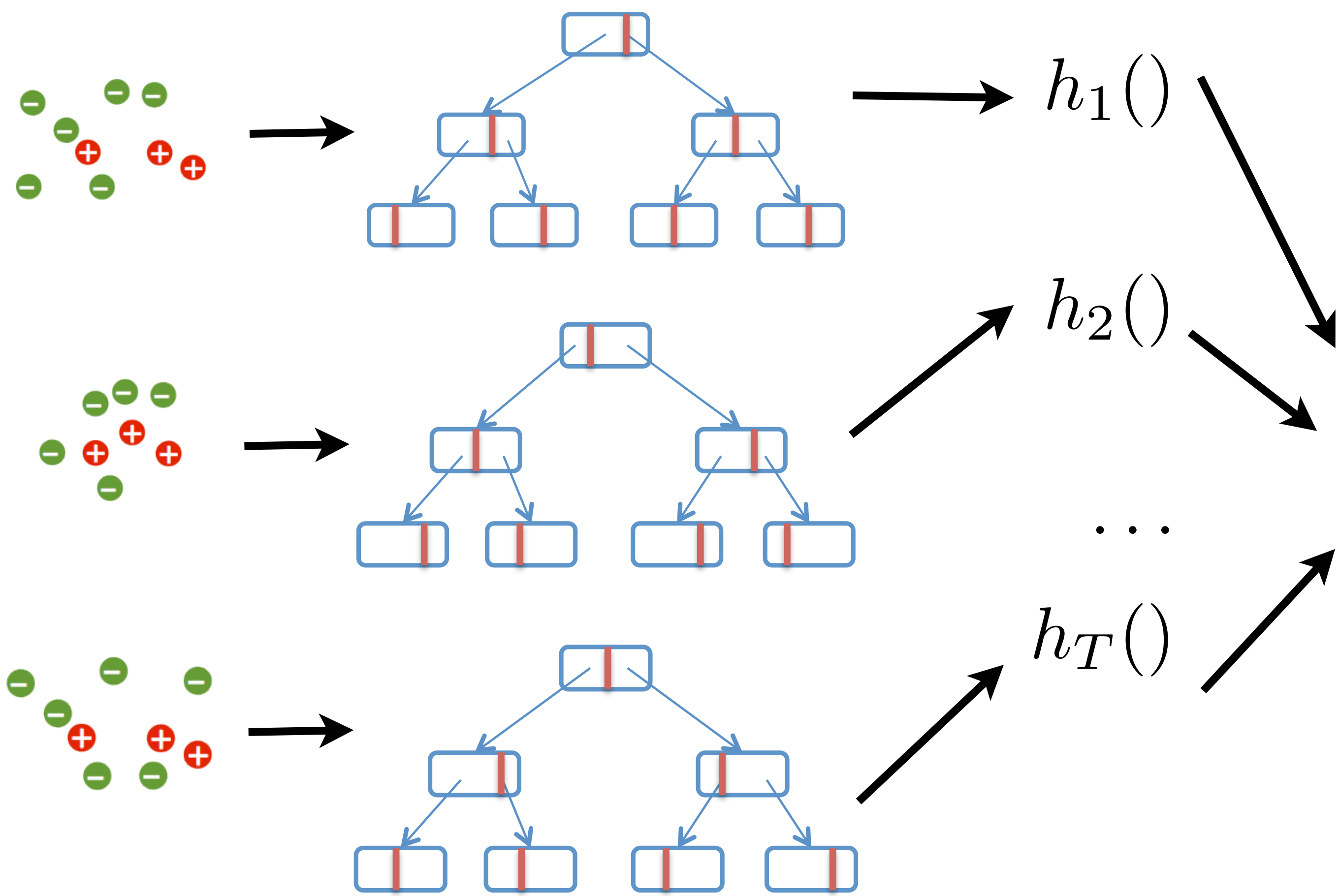
SGD classification

[Lin and Kolcz, SIGMOD'12]

Figure 2: Accuracy of our tweet sentiment polarity classifier on held out test set of 1 million examples. Each bar represents 10 trials of a particular setting, with $\{1, 10, 100\}$ million training examples and varying sizes of ensembles. Error bar denote 95% confidence intervals.

Randomization and Aggregation

Random forest



$H(\boldsymbol{x}) =$

$\frac{1}{T}\sum_{t=1}^{T} h_t(\boldsymbol{x})$ (regression)

$\arg\max_{y} \sum_{t=1}^{T} I(h_t(\boldsymbol{x}) = y)$

(classification)

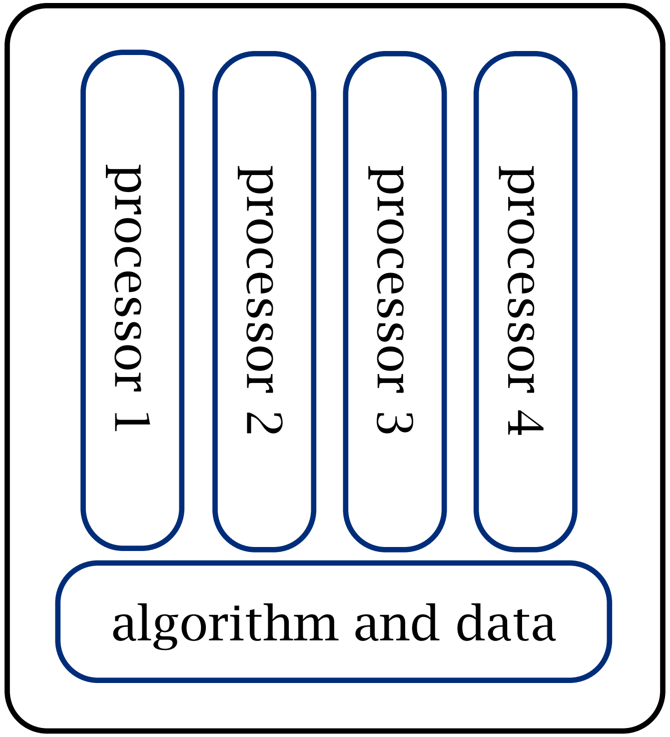
randomly sample data

randomized decision tree

subset

fast and parallelizable

Parallelization



parallel

Decision tree: select the best split points in parallel

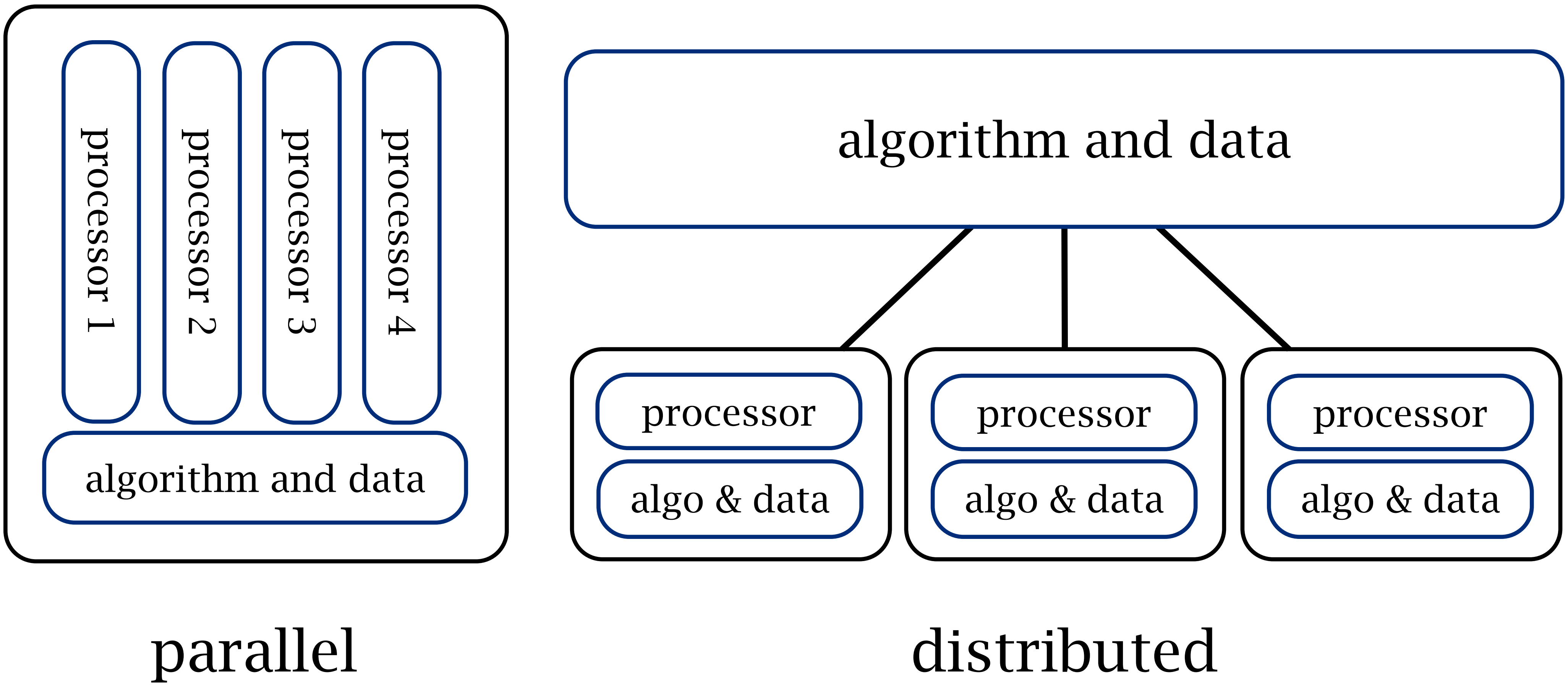
Parallel ensemble: train base learners in parallel

Alleviate the space difficulty

- Use online/one-pass/incremental algorithms
  Decision tree: C5.0
- Use distributed computing architectures

Distributed computing architectures

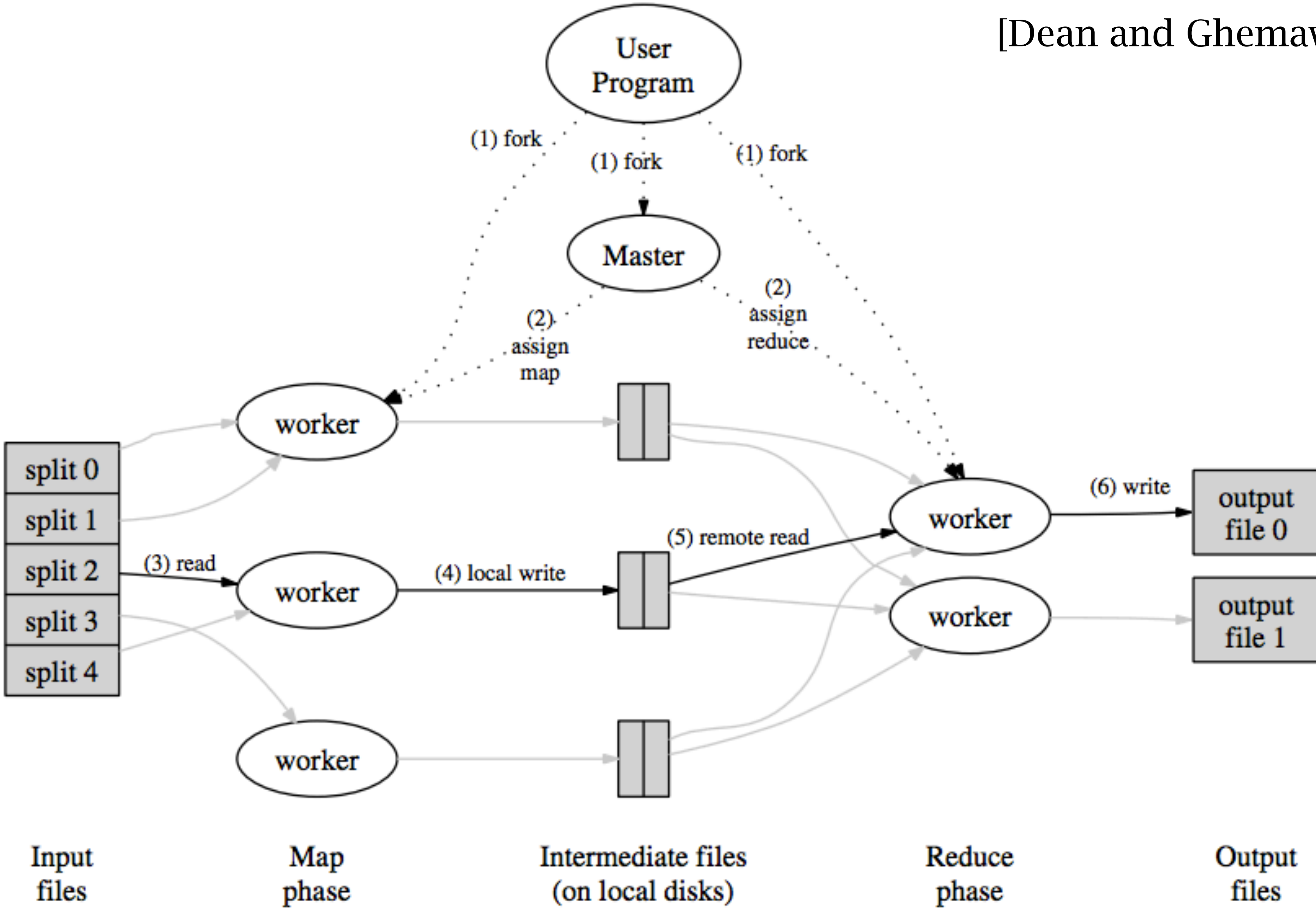
Parallel v.s. distributed computing



Distributed computing architectures

MapReduce: Simplified Data Processing on Large Clusters

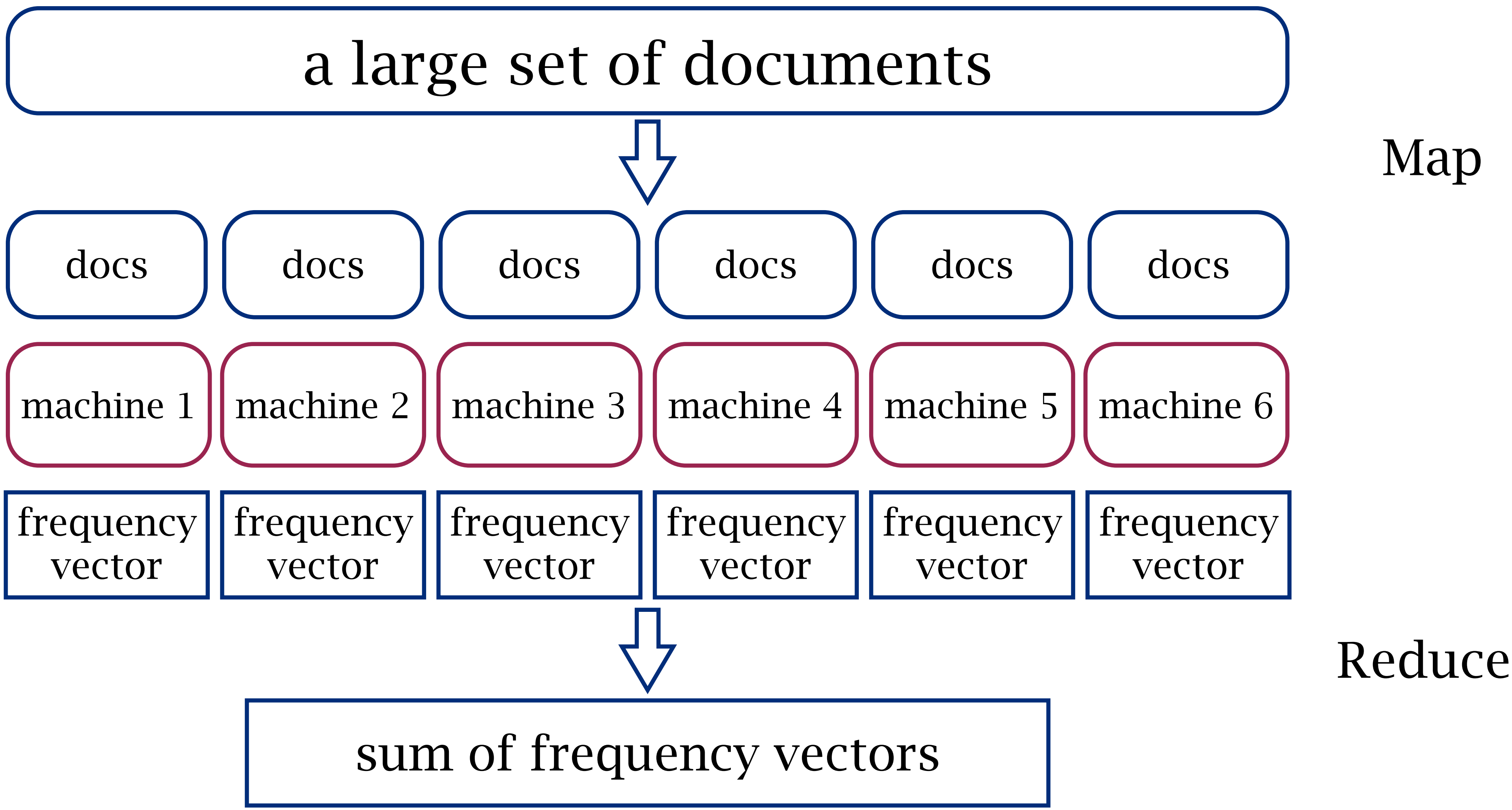
[Dean and Ghemawat, OSDI'04]



Distributed computing architectures

MapReduce

Counting word frequency

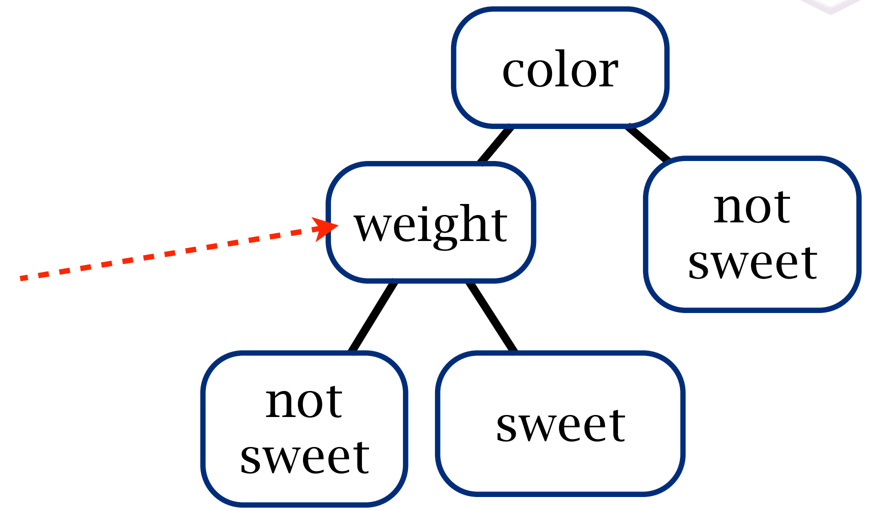


Distributed computing architectures

MapReduce

Learning decision tree

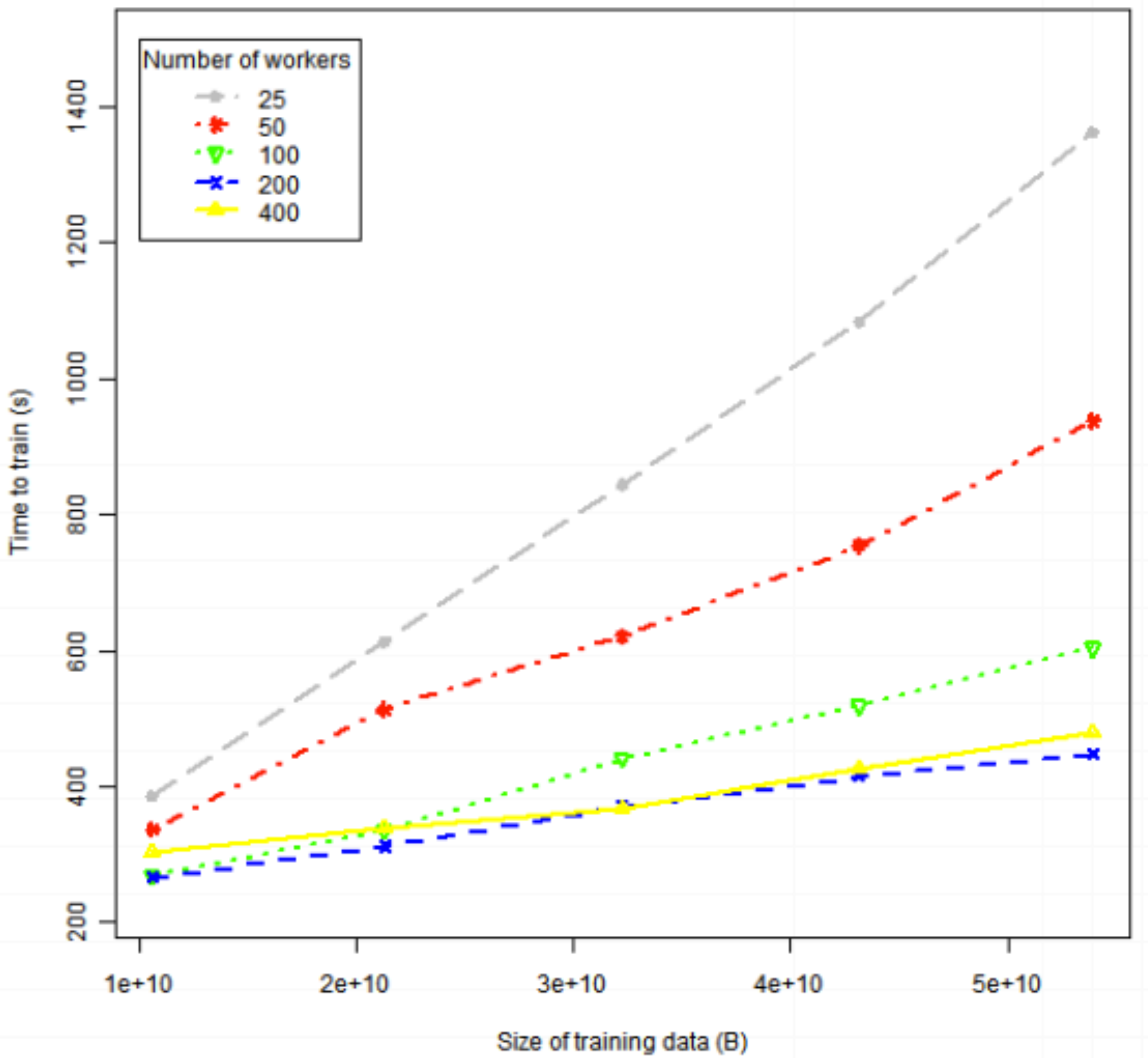
use MapReduce to find
the best split of a node



for every possible split point
map:
    split data to count the instance in each side
reduce:
    the impurity/IG of the split

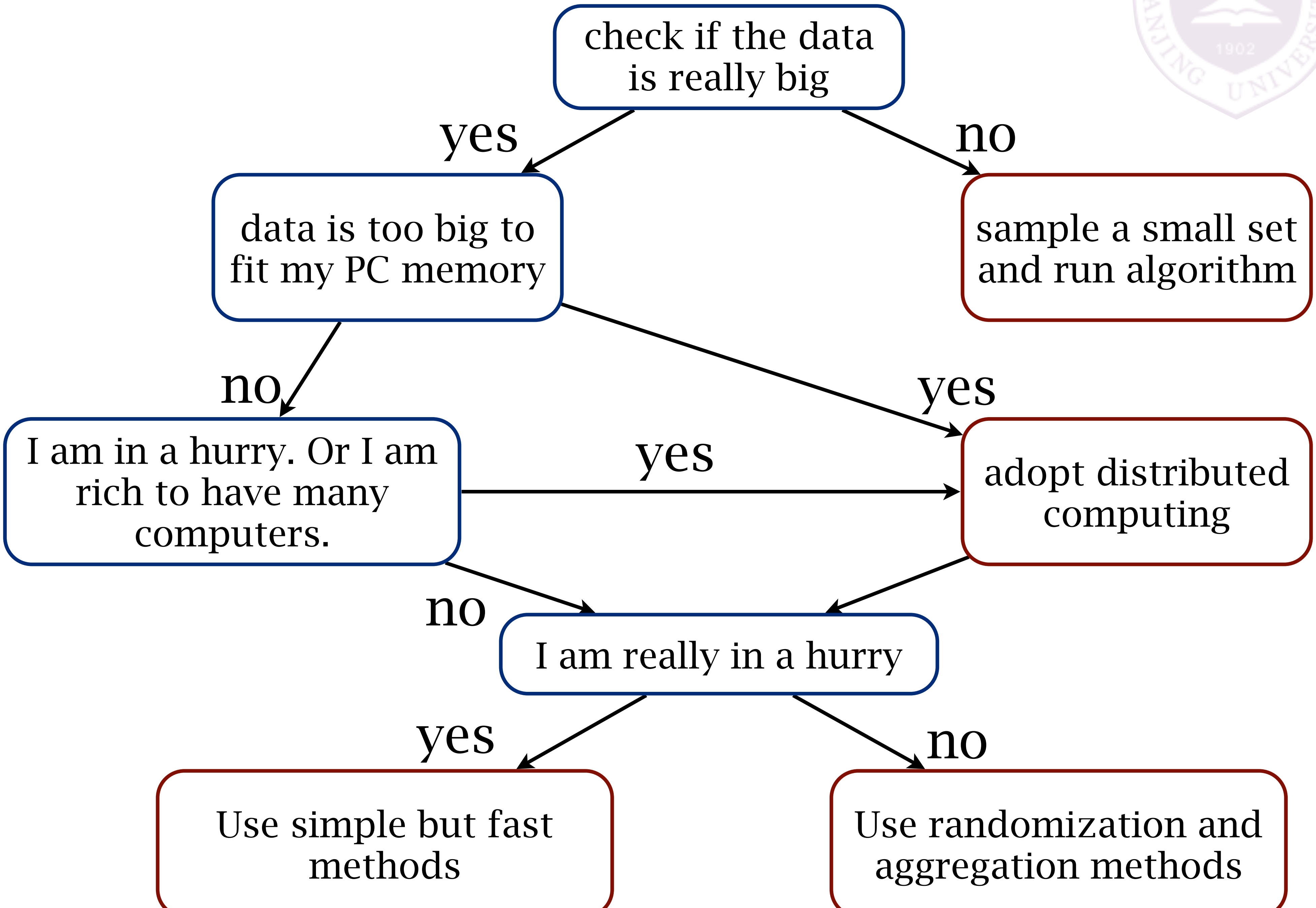
Distributed computing architectures

MapReduce   Learning decision tree



[Panda, et al., VLDB'09]

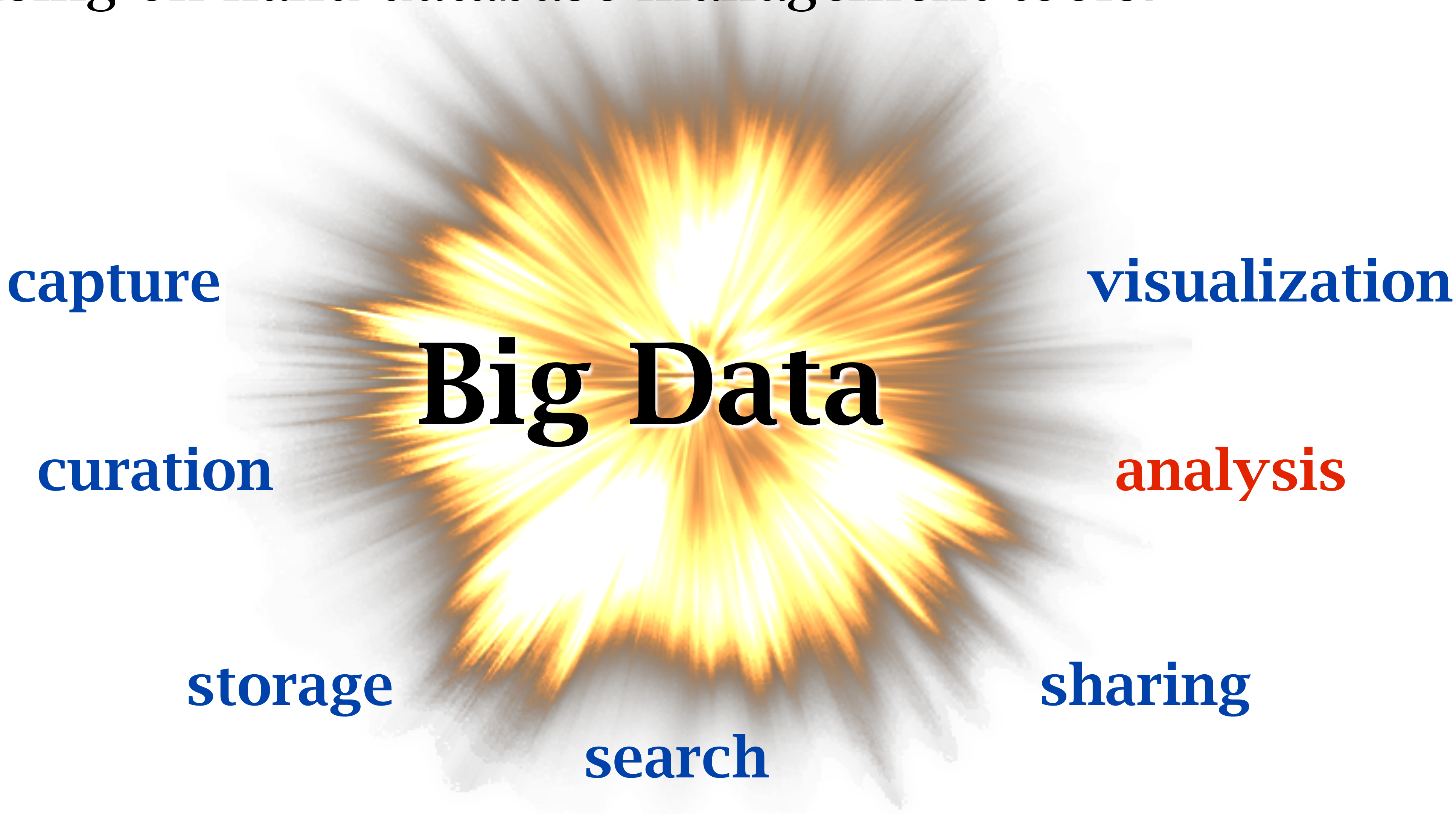
A suggestion



Big Data

What is big data?

big data is a collection of data set so large and complex that it becomes difficult to process using on-hand database management tools. [wikipedia]



Is "big data" new

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner."

mining large-scale data is not a new task

The 1st VLDB: 1975

The 1st KDD: 1995

large database
large datasets
data: GB
CPU 99MHz
RAM: 400MB

FT-Tree (KDD'04):
10 million transactions

Why big data is so hot

Login Cart

nature International weekly journal of science

Search go Advanced search

Journal home > Archive > Editor's Summary

Journal content

- **Journal home**
- **Advance online publication**
- **Current issue**
- **Nature News**
- **Archive**
- **Supplements**
- **Web focuses**
- **Podcasts**
- **Videos**
- **News Specials**

Journal information

- About the journal
- For authors
- Online submission
- Nature Awards
- Nature history

NPG services

Editor's Summary

4 September 2008

Big data: science in the petabyte era

In *Nature* this week, features and opinion pieces on one of the most daunting challenges facing modern science: how to cope with the flood of data now being generated. A petabyte is a lot of memory, however you say it — a quadrillion, 10^{15}, or tens of thousands of trillions of bytes. But that is the currency of 'big data'. We visited the Sanger Institute's supercomputing centre, and its petabyte of capacity. Wikipedia's success shows how well the 'wiki' concept of open-access editing can work. It could work too as a way of coping with the data flows of modern biology. The world's leading search engine is ten this month. Eleven years ago few would have predicted Google's domination: undaunted we ask scientists and business people to try to predict the next big thing, a Google for the petabyte era. Digital data are easily shared, and just as easily wiped or lost. The problem of keeping on-line data accessible is especially difficult for the smaller lab. In Books & Arts, Felice Frankel and Rosalind Reid champion the cause of data visualization as a way of finding meaning in an otherwise daunting data stream. From the 1700s to the mid 1950s, most 'computers' were human. Best known were the 'Harvard computers', a group of women working from the 1880s until the 1940s, at the Harvard College Observatory. Employed to classify stars captured on millions of photographic plates, some of the 'computers' made significant contributions to science. Online databases are a vital outlet for publishing the data being produced by biological research. But the data need to be properly organized. This is the role of the biocurator, but as a team of authors from 15 of the world's major online research resources explains, biocuration is now sadly neglected. An aspect of the data boom with a political dimension is the environment: how much data to collect, how much money to spend. For 'Big data' online, go to http://www.nature.com/news/specials/bigdata/ and to

- Sign up for e-alerts
- Recommend to your library
- RSS newsfeeds
- *Nature* in the news (external link)

open innovation challenges

Detecting Isocyanates in Suspended Particles

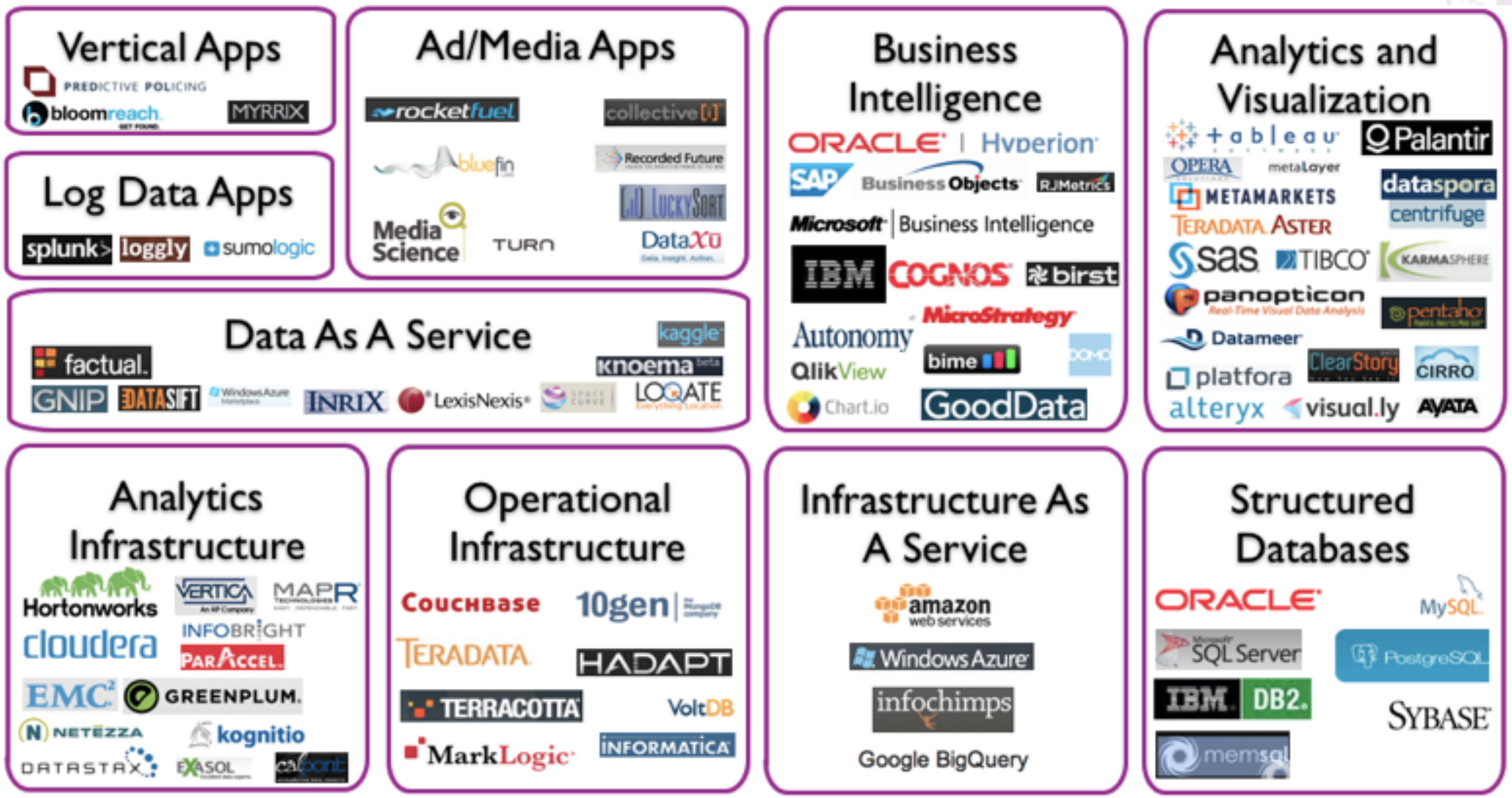
Deadline: Jan 16 2013
Reward: **$25,000 USD**

A detection technology capable of sensitive detection of isocyanates in an aqueous suspension of or...

Topical Methods to Prevent Yeast Infections

Deadline: Dec 18 2012
Reward: **$10,000 USD**

Why big data is so hot



Companies, products, and technologies included in the Big Data Landscape:
- Splunk, Loggly, Sumo Logic
- Predictive Policing, BloomReach, Atigeo, Myrrix
- Media Science, Bluefin Labs, CollectiveI, Recorded Future, LuckySort, DataXu, RocketFuel, Turn
- Gnip, Datasift, Space Curve, Factual, Windows Azure Marketplace, LexisNexis, Loqate, Kaggle, Knoema, Inrix
- Oracle Hyperion, SAP BusinessObjects, Microsoft Business Intelligence,IBM Cognos, SAS, MicroStrategy, GoodData, Autonomy, QlikView, Chart.io, Domo, Bime, RJMetrics
- Tableau Software, Palantir, MetaMarkets, Teradata Aster, Visual.ly, KarmaSphere, EMC Greenplum, Platfora, ClearStory Data, Dataspora, Centrifuge, Cirro, Ayata, Alteryx, Datameer, Panopticon, SAS, Tibco, Opera, Metalayer, Pentaho
- HortonWorks, Cloudera, MapR, Vertica, MapR, ParAccel, InfoBright, Kognitio, Calpont, Exasol, Datastax, Informatica
- Couchbase, Teradata, 10gen, Hadapt, Terracotta, MarkLogic, VoltDB,
- Amazon Web Services Elastic MapReduce, Infochimps, Microsoft Windows Azure, Google BigQuery
- Oracle, Microsoft SQL Server, MySQL, PostgreSQL, memsql, Sybase, IBM DB2
- Hadoop, MapReduce, Hbase, Cassandra, Mahout

[from Forbes]

Why big data is so hot

Office of Science and Technology Policy
Executive Office of the President
New Executive Office Building
Washington, DC 20502

FOR IMMEDIATE RELEASE
March 29, 2012

Contact: Rick Weiss 202 456-6037 rweiss@ostp.eop.gov
Lisa-Joy Zgorski 703 292-8311 lisajoy@nsf.gov

OBAMA ADMINISTRATION UNVEILS "BIG DATA" INITIATIVE:
ANNOUNCES $200 MILLION IN NEW R&D INVESTMENTS

Aiming to make the most of the fast-g
Administration today announced a "Bi
improving our ability to extract knowledge and insights from large and complex
collections of digital data, the initiative
pressing challenges.

To launch the initiative, six Federal de
than $200 million in new commitment
tools and techniques needed to acces
volumes of digital data.

National Science Foundation: In addition to funding the Big Data solicitation, and

US Geological Survey – Big Data for Earth System Science: USGS is announcing

National Science Foundation and the National Institutes of Health - Core Techniques and Technologies for Advancing Big Data Science & Engineering

Department of Defense – Data to Decisions: The Department of Defense (DoD) is "placing a big bet on big data" investing approximately $250 million annually (with $60 million available for new research projects) across the Military Departments in a series of programs that will:

Department of Energy – Scientific Discovery Through Advanced Computing: The Department of Energy will provide $25 million in funding to establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute. Led by the Energy

National Institutes of Health – 1000 Genomes Project Data Available on Cloud: