

The background of the slide is a photograph of the TU Delft campus. It shows a wide, paved pedestrian path leading through a green lawn area with several trees. To the left is a modern building with a white facade and vertical slats. In the background, a tall, dark glass skyscraper with the TU Delft logo and a clock face is visible against a blue sky with scattered white clouds. People are walking along the path.

TI2736-B

Big Data Processing

Claudia Hauff
ti2736b-ewi@tudelft.nl

**Data
streams**

**Map
Reduce**

**Iterative
algs.**

Spark

Course objectives

- **Explain** the ideas behind the “big data buzz”
- **Understand** and **describe** the four different paradigms covered in class
- **Code productively** in one of the most important big data software frameworks we have to date: Hadoop (and tools building on it)
- **Transform** big data problems into **sensible** algorithmic solutions

Final exam

75%

Nothing is mandatory! To pass: **overall grade** ≥ 5.75

Assignments

25%

Quizzes

+1

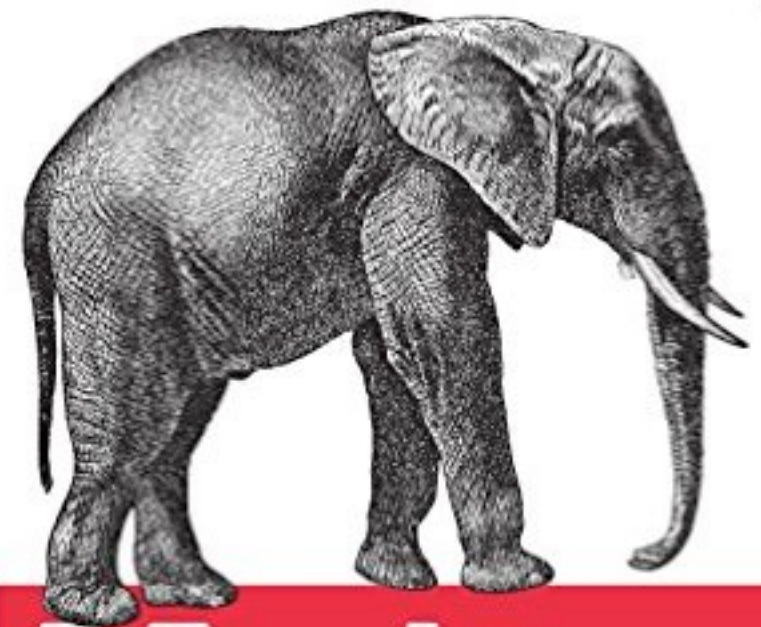
A close-up photograph of a wooden sign. The sign is made of light-colored wood with a visible grain. It features two lines of text in large, black, distressed, serif capital letters. The top line reads 'HELP' and the bottom line reads 'WANTED'. The letters are slightly weathered and have a rough, hand-painted appearance.

HELP

**If you need help with your assignments,
attend the lab sessions (week 2+).**

O'REILLY

4th Edition
Revised & Updated



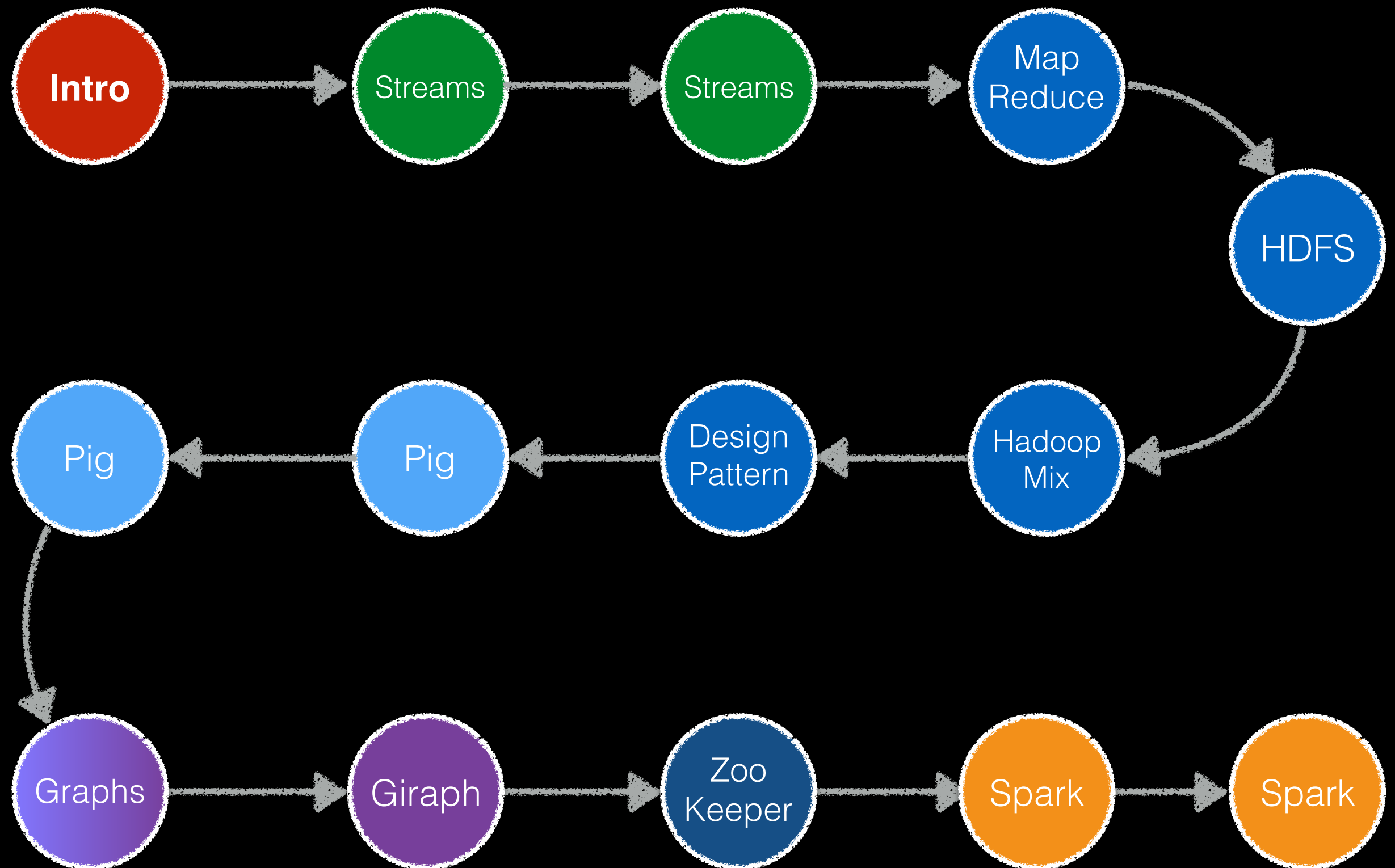
Hadoop

The Definitive Guide

STORAGE AND ANALYSIS AT INTERNET SCALE

Tom White

Big Data



Today's learning objectives

- **Explain** and **recognise** the V's of big data
- **Explain** the main differences between data streaming and MapReduce algorithms
- **Identify** the correct approach (streaming vs. MapReduce) to be taken in an application setting

What is “big data”?

- A buzz word, fuzzy boundaries

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

“Data **too large & complex** to be effectively handled by standard database technologies currently found in most organisations.”

Big Data Landscape 2016

Infrastructure

Hadoop On-Premise
cloudera, Hortonworks, MAPR, Pivotal, IBM InfoSphere, splice, bluedata, jethro

Hadoop in the Cloud
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Dubble, xplenty

Spark
databricks, GridGain, TACHYON NEXUS

Cluster Services
amazon, Microsoft Azure, Google Cloud Platform, IBM InfoSphere, CAZENA, altiscale, Dubble, xplenty

Analytics

Analyst Platforms
Palantir, AYASDI, Quid, enigma, Digital Reasoning, ORBITALINSIGHT

Analytics Platforms
Microsoft, guavus, Datameer, inter|ano

Data Science Platforms
context relevant, DataRobot, Alpine, MODE, plotly, ADATA, dataiku, DOMINO, yhat, ALGORITHMIA

Visualization
tableau, Google Cloud Platform, Roambi, QOMDATA, Qlik, CHARTIO

Applications

Sales & Marketing
RADIUS, Gainsight, bloomreach, Zeta, livefyre, blueyonder, @kahuna, Lattice, SAILTHRU, persado, infer, bsense, AVISO, ACTIONIQ, QUANTIFIND, ENGAGIO

Customer Service
MEDALLIA, ATTENSTY, CLARABRIDGE, STELLAService, NGDATA, Preact, DigitalGenius, Wise.io, eppuri, fuse|machines

Human Capital
gild, Connectifier, textio, entelo, hiQ

Legal
RAVEL, JUDICATA, Everlaw, Brevia, PREMONITION

NoSQL Databases
amazon, Google Cloud Platform, Microsoft Azure, mongoDB, MarkLogic, DATASTAX, Couchbase, SequoiaDB, redislabs, influxdata

NewSQL Databases
SAP, Clustrix, Pivotal, paradigm4, memsql, nuODB, MariaDB, VOLTDB, citusdata, deepdb, Trafodion, Cockroach LABS

BI Platforms
Power BI, amazon, Wave Analytics, Domo, GoodData, birst, platfora, atscale, looker, arcs4, SISENSE

Statistical Computing
sas, SPSS, MATLAB

Log Analytics
splunk, sumologic, kibana, CLOUD PHYSICS, loggly

Social Analytics
NETBASE, DATASIFT, tracx, bitly, synthetio, bottlenose, simple reach

Ad Optimization
MediaMath, Integral, OpenX, Adgorithms, LiveIntent, distillery, DataXu, Clippier, TAPAD

Security
CYLANCE, CounterTack, cybereason, AREA 1 SECURITY, SentinelOne, Recorded Future, Guardian Analytics, FORTSCALE, sift science, Keybase, feedzai, SCIFYD

Vertical AI Applications
facebook, Clara, KASIST, lumiata

Graph Databases
neo4j, OrientDB, InfiniteGraph

MPP Databases
TERADATA, VERTICA, NETEZZA, kognitio, dremio

Cloud EDW
amazon, Google Cloud Platform, Microsoft Azure, Pivotal, snowflake, WATERBURG DATA, Infoworks

Data Transformation
alteryx, TRIFACTA, tamr, Paxata, StreamSets, Alation

Data Integration
informatica, MuleSoft, snapLogic, BedrockData

Real-Time
amazon, METAMARKETS, confluent, DataTorrent, dataArtisans

Machine Learning
Azure Machine Learning, H2O, Dato, SKYTREE, rapidminer, CATALYTICA, deepsense.io, VISENZE, PredictionIO, glowfish

Speech & NLP
NarrativeScience, api.ai, NUANCE, Gr8space, semantic machines, cortical.io, mindfield, molu, IDIBON, yseop

Horizontal AI
IBM Watson, Cortana, sentient, VIV, nervana, nora, Hypericence, MetaMind, clarifai, DEEPTO, Geometric Intelligence

Publisher Tools
Outbrain, mixpanel, Chartbeat, yieldbot, Yieldmo

Govt/ Regulation
Socrata, OPENGOV, FN, FiscalNote, enigma, PREDPOL, mark43, OpenDataSoft

Finance
affirm, LendingClub, OnDeck, Kreditech, zena finance, LendUp, Kabbage, tidemark, INSIGHT, ZUORA, Dataminr, Lenddo, KENSHC, AIDYA, iSENTIUM, Quantopian, sentient

Management / Monitoring
New Relic, APPDYNAMICS, amazon, actifio, Numerify, splunk, DATADOG, Trocena, Anodot

Security
TANUIM, illumio, CODE42, DataGravity, CipherCloud, VECTRA, sqrl, BlueTalon

Storage
amazon, Google Cloud Platform, Microsoft Azure, panasas, nimblestorage, Qumulo

App Dev
apigee, CASK, Typesafe, CONCURRENT

Crowd-sourcing
amazon, mechanical Turk, CrowdFlower, WorkFusion

Search
hp, Autonomy, ORACLE, EXALERO, Lucidworks, elastic, ThoughtSpot, MAANA, swifttype, Algolia, SINEQUA

Data Services
UO, OPERA, Mu Sigma, DATASCIENCE, ELUCID, kaggle, dataKind

For Business Analysts
OrigamiLogic, ClearStory, CIRRO, import.io

SMB / Commerce
Google Analytics, AMPUTIDE, RJMetrics, BLUECORE, sumail, granify, Airtable, retention, custora

Education/ Learning
KNEWTON, Clever, Geclara, PANORAMA, knowtre

Life Sciences
23andMe, Counsyl, RECOMBINE, KYRUUS, FLATIRON, oozymergen, HealthTop, METABIOTA, ZEPHYR, ovio, Ginger.io, transcriptic, Glow, enihic, AICure, Atomwise

Industries
OPower, eHarmony, RetailNext, duetto, STITCH FIX, WorkFusion, BLUE RIVER, TACHYUS, Seeq, FarmLogs, SwiftKey, select, BBOXER, statmuse

Cross-Infrastructure/Analytics

amazon, Google, Microsoft, IBM, SAP, SAS, hp, Autonomy, vmware, talend, TIBCO, TERADATA, ORACLE, NetApp

Open Source

Framework
Hadoop, YARN, Spark, MESOS, TEZ, Flink, CDAP

Query / Data Flow
SLAMDATA, ARABIC DRILL, Google Cloud Dataflow

Data Access
cassandra, HBASE, mongoDB, CouchDB, riak, SCIDB, OPENSCB, nifi

Coordination
talend, Apache Zookeeper, Apache Ambari

Real-Time
STORM, Spark, APEX, Flink, TACHYON, druid

Stat Tools
R, Scala, NumPy, SciPy

Machine Learning
mlib, Aerosolve, Caffe, WEKA, FeatureFu, DIMSUM, jupyter, DL4J

Search
elasticsearch, Solr, Lucene

Security
Apache Ranger

Visualization
Zeppelin

Data Sources & APIs

Health
Apple, JAWBONE, GARMIN, practicefusion, fitbit, Withings, VALIDIC, netatmo, kinsa, Human API

IOT
UPTAKE, ThingWorx, helium, samsara

Financial & Economic Data
Bloomberg, DOW JONES, YODLEE, PREMISE, S&P CAPITAL IQ, quandl, xignite, CB INSIGHTS, mattermark, estimote, PLAID

Air / Space / Sea
PLANET LABS, spire, WINDWARD, CRUISE, SKYCATCH, Airware, DroneDeploy

Location/People/Entities
GARMIN, foursquare, InsideView, esri, STREETLINE, CARTOON, factual, PlaceIQ, Crimson Hexagon, placemeter, BASIS, Sense

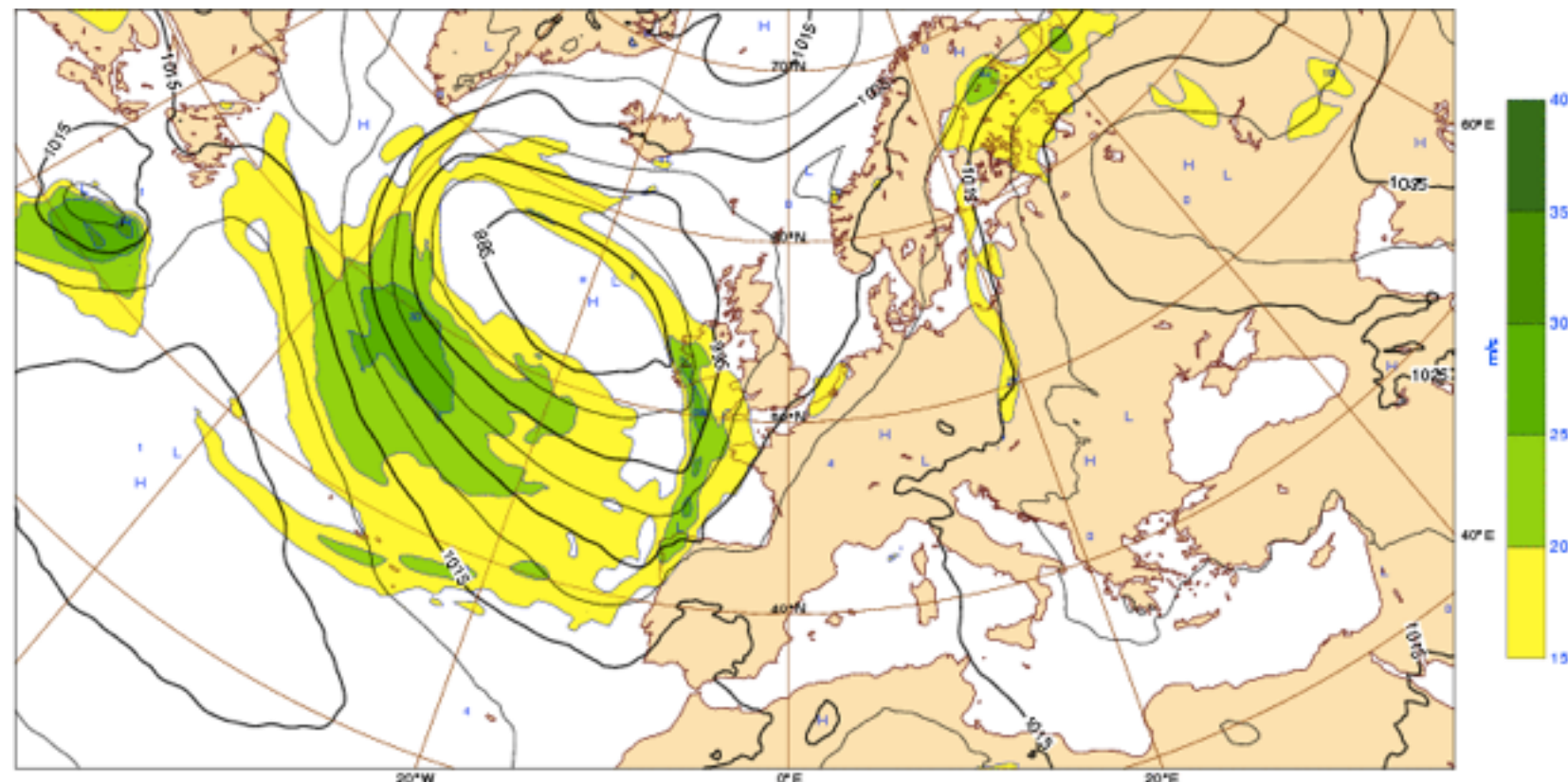
Other
qualtrics, panjiva, DATA.GOV

Incubators & Schools
GA, DataCamp, INSIGHT, DataElite, METIS, The Data Incubator

Large-scale computing is not new

Weather forecasting has been a long-term scientific challenge

- Supercomputers were already used in the 1970s
- Equation crunching



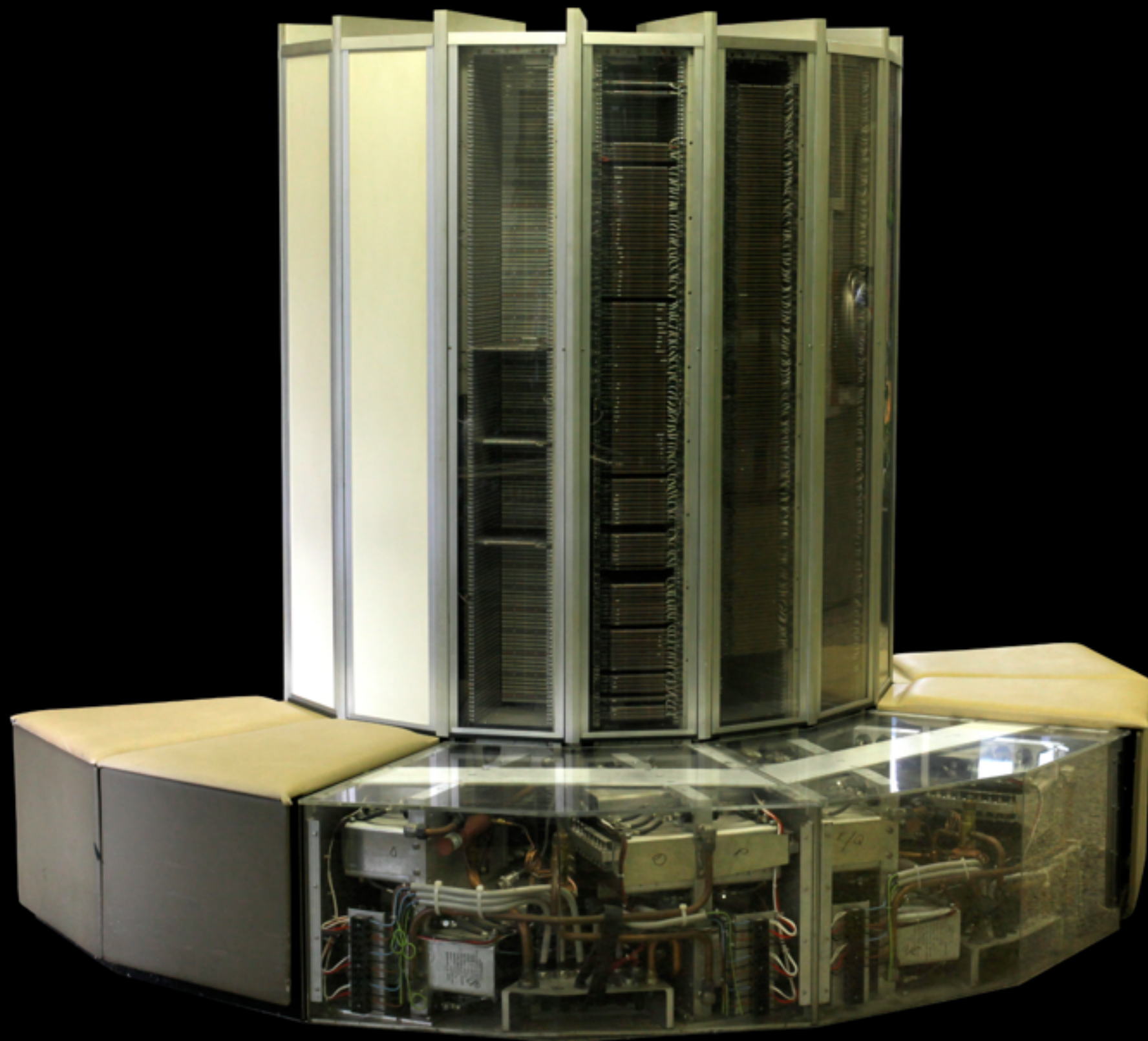
Large-scale computing is not new

Weather forecasting has been a long-term scientific challenge

- Supercomputers were already used in the 1970s
- Equation crunching

Specification	Cray-1A	IBM POWER7 System	Approx Ratio
Year installed	1978	2012	
Architecture	Vector processor	Dual Cluster of scalar CPUs	
Number of Cores	1	~49,000	49,000:1
Clock Speed	12.5 nsec (80 MHz)	0.26 nsec (3.83 GHz)	49:1
Peak perf per Core	160 MFLOPS	30 GFLOPS	190:1
Peak perf per system	160 MFLOPS	~1.5PFLOPS	9,200,000:1
Sustained performance	~50 MFLOPS	~70 TFLOPS	1,400,000:1
Memory	8 MiBytes	~106 TiBytes	13,900,000:1
Disk Space	2.5 GBytes	~3.1 PBytes	1,250,000:1






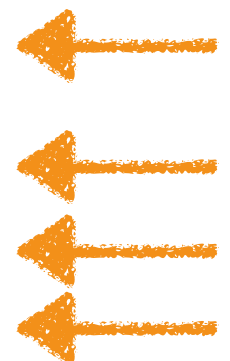
Cray-1A

Large-scale computing is not new

Weather forecasting has been a long-term scientific challenge

- Supercomputers were already used in the 1970s
- Equation crunching

Specification	Cray-1A	IBM POWER7 System	Approx Ratio
Year installed		2012	
Architecture		Dual Cluster of scalar CPUs	
Number of Cores		~49,000	49,000:1
Clock Speed		0.26 nsec (3.83 GHz)	49:1
Peak perf per Core		30 GFLOPS	190:1
Peak perf per system		~1.5PFLOPS	9,200,000:1
Sustained performance		~70 TFLOPS	1,400,000:1
Memory		~106 TiBytes	13,900,000:1
Disk Space		~3.1 PBytes	1,250,000:1





IBM Power7

Big data processing

- So-called big data technologies are about **discovering patterns** (in semi/unstructured data)
- Main focus is on **how** to make computations on big data **feasible** without a supercomputer
 - Cluster(s) of **commodity hardware**
- Q3: Data Mining course focuses on **how to discover** those patterns

Just an academic exercise?

- **Cloud computing**: “Anything running inside a browser that gathers and stores user-generated content”
- **Utility computing**
 - Computing as a **metered service**
 - A “cloud user” buys any amount of computing power from a “cloud provider” (**pay-per-use**)
 - Virtual machine instances
 - IaaS: **infrastructure as a service**
 - **Amazon Web Services** is the dominant provider

Just an academic exercise?

vCPU ECU Memory (GiB) Instance Storage (GB) Linux/UNIX Usage

General Purpose - Current Generation

GPUs

p2.16xlarge	64	188	732	EBS Only	\$14.4 per Hour
t2.micro	1	Variable	1	EBS Only	\$0.013 per Hour
t2.small	1	Variable	2	EBS Only	\$0.026 per Hour
t2.medium	2	Variable	4	EBS Only	\$0.052 per Hour
t2.large	2	Variable	8	EBS Only	\$0.104 per Hour
m4.large	2	6.5	8	EBS Only	\$0.12 per Hour
m4.xlarge	4	13	16	EBS Only	\$0.239 per Hour
m4.2xlarge	8	26	32	EBS Only	\$0.479 per Hour

You can run your own big data experiments!

Progress often driven by industry

- Development of **big data standards** & (open source) software commonly driven by companies such as Google, Facebook, Twitter, Yahoo! ...
- Why do they care about big data?

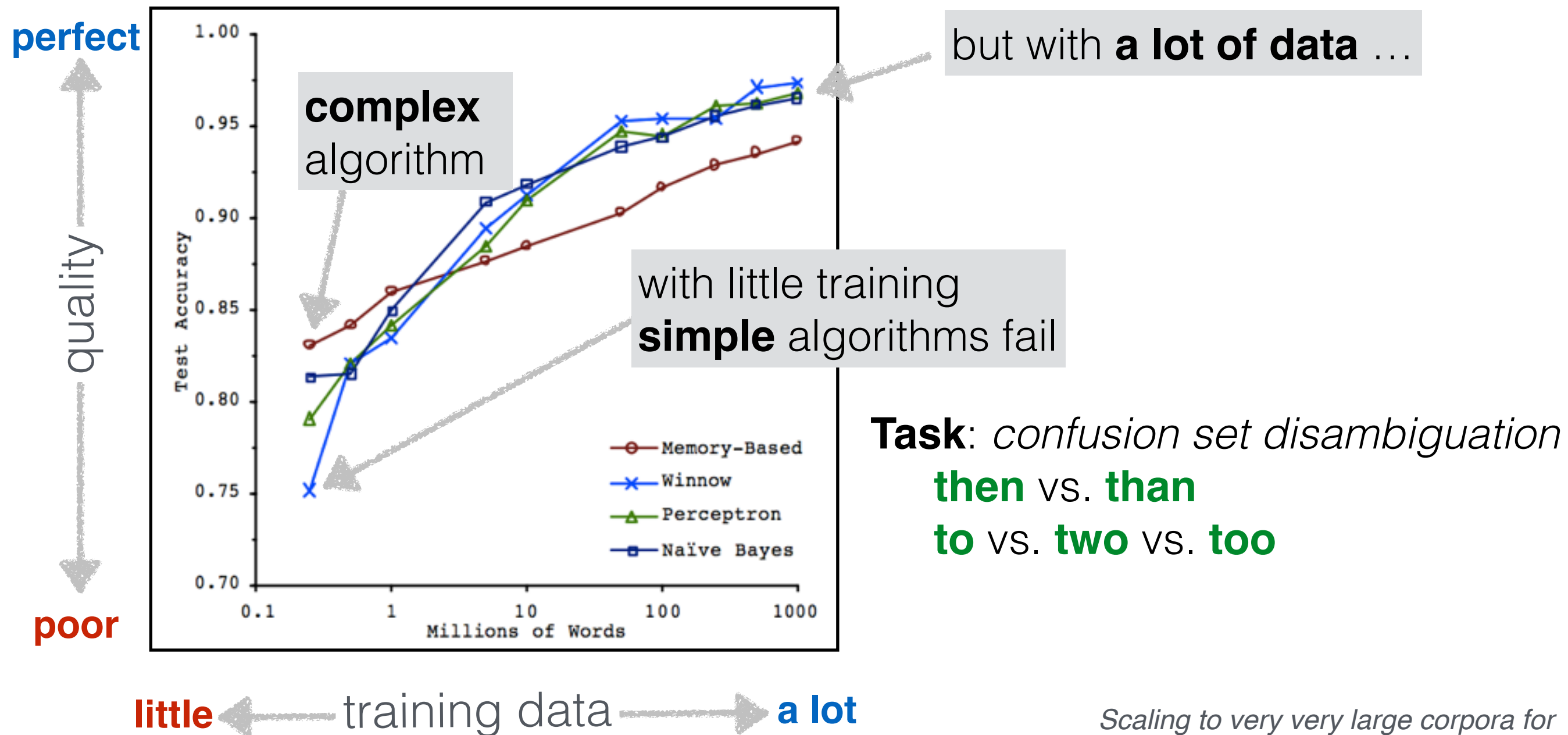


- More knowledge leads to:
 - better customer engagement
 - fraud prevention
 - new products

Big data analytics: IBM pitch

<https://www.youtube.com/watch?v=1RYKgJ-QK4I>

Big data vs. small data



*Scaling to very very large corpora for
natural language disambiguation.
M. Banko and E. Brill, 2001.*

The 5 V's

- **Volume**: large amounts of data
- **Variety**: data comes in many different forms from diverse sources
- **Velocity**: the content is changing quickly
- **Value**: data alone is not enough; how can value be derived from it?
- **Veracity**: can we trust the data? how accurate is it?

Velocity types

- **Batch processing**: running a series of computer programs without human intervention
- **Near real-time**: brief delay between the data becoming available and it being processed
- **Real-time**: guaranteed responses between the data becoming available and it being processed

Variety types

- **Structured** data
(well defined fields)
- **Semi-structured** data
- **Unstructured** data
(by humans for humans)

standard in the past

Show less (9 rows / 4 columns total) - Export data

Module Type	DIA	DIA Storage Server	
Size / Rack Configuration	2U – Up to 11 modules/per	8U – 4 modules /rack	8U – 4 modules /rack

Number
Total Nu
Total Me
Number
Storage
Data St
Operatin

[W3C home](#) > [Mailing lists](#) > [Public](#) > [public-ldp-wg@w3.org](#) > [November 2014](#)

Re: ldp wishlist for crosscloud

This message: [[Message body](#)] [[Respond](#)] [[More options](#)]

Related messages: [[Next message](#)] [[Previous message](#)] [[In reply to](#)] [[Next in thread](#)] [[Replies](#)]

From: Sandro Hawke <sandro@w3.org>
Date: Sun, 09 Nov 2014 21:48:38 -0500
Message-ID: <54602786.4090003@w3.org>
To: "henry.story@bbfish.net" <henry.story@bbfish.net>
CC: Linked Data Platform WG <public-ldp-wg@w3.org>

On 11/09/2014 06:51 PM, henry.story@bbfish.net wrote:
 >> On 9 Nov 2
 >>
 >> On 11/09/2
 >>> Hi Sandro
 >>>
 >>> thank
 >>> a long ti
 >>>
 >>>> On 9 Nov
 >>>>
 >>>> As you s
 >>>> which allows
 >>>> software..

The exponential growth in the amount of biological data means that revolutionary measures are needed for data management, analysis and accessibility. Online databases have become important avenues for publishing biological data. Biocuration, the activity of organizing, representing and making biological information accessible to both humans and computers, has become an essential part of biological discovery and biomedical research. But curation increasingly lags behind data generation in funding, development and recognition.

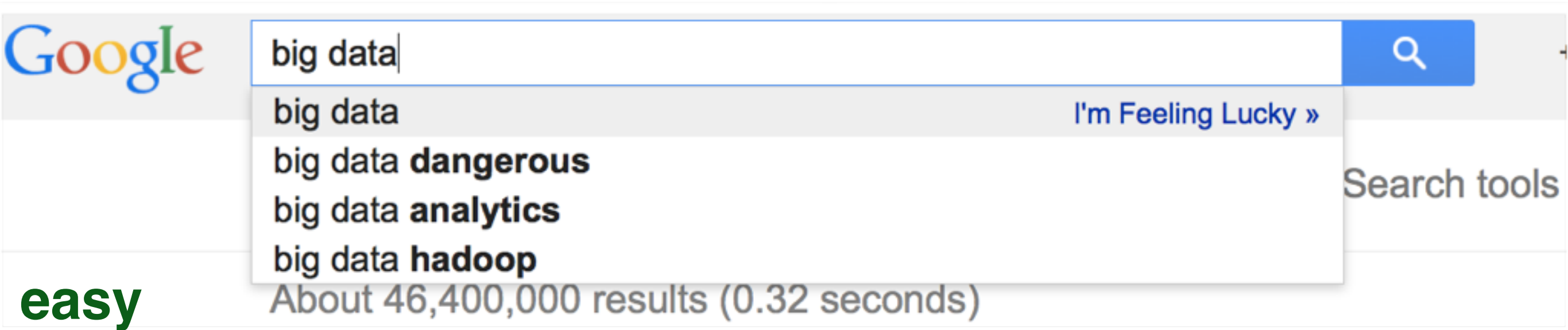
We propose three urgent actions to advance this key field. First, authors, journals and curators should immediately begin to work together to facilitate the exchange of data between journal publications and databases. Second, in the next five years, curators, researchers and university administrations should develop an accepted recognition structure to facilitate community-based curation efforts. Third, curators, rese
 next ten year
 professional (

most common today



Unstructured text

- To get value out of unstructured text we need to impose structure **automatically**
 - Parse text
 - **Extract meaning** from it (can be easy or difficult)
- Amount of data we create is more than doubling every two years, most new data is unstructured or at most semi-structured

Extracting meaning



Extracting meaning

Google +Claudia  

Tables experimental Results 1 - 10 of about 535,155 for big data. (0.31 seconds)

Web
Web Tables
Fusion Tables
Send Feedback

[Big Data Solutions | Talend](https://www.talend.com/products/big-data)
<https://www.talend.com/products/big-data>
FEATURES Job Designer Components for ... Hadoop Job Scheduler NoSQL Support
[Show less \(11 rows / 4 columns total\)](#) - [Export data](#)

FEATURES	Talend Open Studio	Talend Enterprise	Talend Platform
Job Designer	x	x	x
Components for HDFS,	x	x	x
Reporting and Dashboards			x
Big Data Profiling,			x
Indemnification/Warranty		x	x

[Data Science, Big Data, Gregory Piatetsky](http://www.odbms.org/2014/10/data-science-human-science/)
<http://www.odbms.org/2014/10/data-science-human-science/>
Region US Canada 51 Europe 27 Asia 11
[Show less \(6 rows / 4 columns total\)](#) - [Export data](#) - created: 7 Oct 2014

Region	Only or Mostly Human	Equally Human and	Only or Mostly Machine
US/Canada (51%)	66%	19%	11%
Europe (27%)	49%	30%	21%

Search tools

easy

difficult

Examples of Volume & Velocity: Twitter

- **>500 million** tweets a day
- **>300 million** active users / day
- On average >6000 tweets a second
- Peaks of >100,000 tweets a second
 - Super Bowl
 - US election (**75M** tweets in 24h)
 - New Year's Eve
 - Football World Cup (**672M** tweets in total #WorldCup)



- Messages are instantly accessible for **search**
- Messages are used in **post-hoc** analyses to gather insights



70

WORLDWIDE LHC
COMPUTING GRID
(WLCG)

ENGINEERING
SERVICES

WORLD WIDE WEB
SERVERS


Examples of Velocity: targeted advertising on the Web

- US revenues in 2013: ~\$40 billion
- Advertisers usually get paid per click
- For each search request, search engines decide
 - **whether** to show an ad
 - **which** ad to show
- Users willing at best to wait **2 seconds** for their search results
- Feedback loop via user clicks, user searches, mouse movements, etc.

Examples of Velocity: targeted advertising on the Web

- US rev
- Advert
- For ea
- **when**
- **which**
- Users search
- Feedb mouse

The screenshot shows a Google search for "big data" with the following elements:

- Search Bar:** "big data" entered, with a search button and a user profile "+Claudia".
- Navigation:** "Web" (selected), "Images", "News", "Videos", "Books", "More", "Search tools".
- Results:** "About 874,000,000 results (0.56 seconds)".
- Advertisements (Left Column):**
 - Big Data for Non-Geeks - Get the Big Data Playbook**
Ad www.sas.com/
6 Common Plays Using Hadoop.
SAS Software has 3,129 followers on Google+
What is Big Data - What is Hadoop - Big Data Insights - Hadoop Solutions
 - Big Data Whitepaper - OpenText.com**
Ad www.opentext.com/big-data-whitepaper
Leverage **Big Data** with Enterprise Information Management.
Free Paper!
OpenText has 204 followers on Google+
 - Big Value from Big Data - Pentaho.com**
Ad www.pentaho.com/Big-Data
Learn **Big Data** Trends and Use Cases with Pentaho's Free eBook Guide!
- Advertisements (Right Column):**
 - Google Bigquery**
cloud.google.com/BigQuery
Analyseer **Big Data** In De Cloud Met SQL. Zonder Servers. Probeer Het!
 - Big Data in 2014**
www.tableausoftware.com/big-data
3.7 ★★★★★ advertiser rating
7 Things you Need to Do About **Big Data** in 2014. Get the Free Article!
 - MarkLogic Big Data**
www.marklogic.com/
The New Database Revolution.
Learn More About **Big Data** Now!
 - Big Data as Fresh Data**
www.ericsson.com/big-data
Agile tools for Telecom Operators.
Read about **Big Data** as Fresh Data!
 - Bigdata**
www.gigya.com/5-Ways-Big-Data
5 Ways **Big Data** Will Transform Marketing. Get Free Whitepaper!
 - Big Data for Oil and Gas**
www.hds.com/
Free Paper: How **Big Data** Analytics
- Organic Results:**
 - Big data** is an all-encompassing term for any collection of **data** sets so **large** and complex that it becomes difficult to process using traditional **data** processing applications. The challenges include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.

www.ibm.com
 - Big data - Wikipedia, the free encyclopedia**
en.wikipedia.org/wiki/Big_data

More interactions/data on the Web

- **YouTube**: 4 billion views a day, one hour of video uploaded every second
- **Facebook**: 483 million daily active users (*Dec. 2011*); 300 Petabytes of data
- **Google**: >1 billion searches per day (*March 2011*)
- **Google** processed 100 Terabytes of data per day in *2004* and 20 Petabytes data/day in *2008*
- **Internet Archive**: contains 2 Petabytes of data, grows 20 Terabytes per month (*2011*)

Movie recommendations

Bob **likes**
X-MEN.

So does Alice.

Alice

Tom

Should we **suggest**
X-MEN to Frank?

Frank

Joe



Jane **dislikes**
X-MEN.

Movie recommendations contd.

- **Ignore the data**, use **experts** instead (movie reviewers); assumes no large subscriber/reviewer divergence
- Use all data but **ignore individual preferences**; assumes that most users are close to the average
- Lump people into **preference groups** based on shared likes/dislikes; compute group-based average score per movie
- Focus computational effort on **difficult** movies

The research field of **recommender systems** is concerned with this issue.

Movie recommendations contd.

- **Netflix Prize**: open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings (>**100 million** ratings by ~0.5 million users for ~18,000 movies)
- First competitor to improve over Netflix's baseline **by 10%** receives \$1,000,000
- Competition started in 2006, prize money was paid out in 2009 (winner was **20 minutes faster** runner up)

Thousands of (research) teams competed.
Innovation driven by industry again!

Example of Variety: Restaurant Locator

- **Task:** Given a person's location anywhere in the world, list the top five restaurants in his immediate neighbourhood.
- **Required data:**
 - World map, list of all restaurants in the world (opening hours, GPS coordinates, menu, special offers)
 - Reviews/ratings
 - Optional: social media stream(s)
- Data is continuously **changing** (restaurants close, new ones open, data formats change, etc.)

Society can benefit too...

- Accurate **predictions of natural disasters** and diseases
- Better responses to **disaster discovery**
 - Timely & effective decisions
 - Provide resources where need the most
- Complete disease/genomics databases to enable **biomedical discoveries**
- Accurate models to support **forecasting** of ecosystem developments

Idea: earthquake warnings

- Social sensors: users (humans) that use Twitter, Facebook, Instagram, i.e. portals with **real-time** posting abilities



- Challenges: how to detect when a tweet is about an **actual earthquake, which** earthquake is it about and **where** is the centre

Idea: earthquake warnings contd.

It is possible!

- **Goal**: warning should reach people earlier than the seismic waves
- Travel times of seismic waves: 3-7km/s; arrival time of a wave 100km away: **20 seconds**
- Speed of an existing system (Sakaki et al., 2010):

Twitter-based

Date	Magnitude	Location	Time	E-mail sent time	#tweets within 10 min	Announce of JMA
Aug. 18	4.5	Tochigi	6:58:55	7:00:30	35	07:08
Aug. 18	3.1	Suruga-wan	19:22:48	19:23:14	17	19:28
Aug. 21	4.1	Chiba	8:51:16	8:51:35	52	8:56
Aug. 25	4.3	Uraga-oki	2:22:49	2:23:21	23	02:27
Aug. 25	3.5	Fukushima	22:21:16	22:22:29	13	22:26
Aug. 27	3.9	Wakayama	17:47:30	17:48:11	16	17:53
Aug. 27	2.8	Suruga-wan	20:26:23	20:26:45	14	20:31
Aug. 31	4.5	Fukushima	00:45:54	00:46:24	32	00:51
Sep. 2	3.3	Suruga-wan	13:04:45	13:05:04	18	13:10
Sep. 2	3.6	Bungo-suido	17:37:53	17:38:27	3	17:43

earthquake

40

traditional warning system

A brief introduction to

Streaming & MapReduce

A brief introduction to Streaming

Data streaming scenario

- **Continuous** and **rapid** input of data (“stream of data”)
- **Limited memory** to store the data - less than linear in the input size
- **Limited time** to process each data item - sequential access
- Algorithms have **one** (**or very few** passes) over the data
- Can be approached from a **practical** or **mathematical** point of view: metric embedding, pseudo-random computations ...

We go for the practical setup!

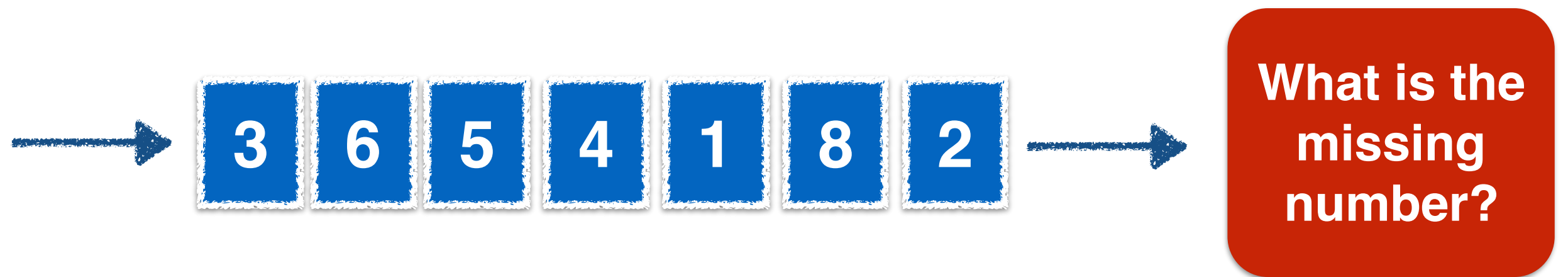
Data streaming example



stream of n numbers; permutation from 1 to n ;
one number is missing;
we are allowed **one pass** over the data

Solution 1: memorise all number seen so far;
memory requirements: n bit (impractical for large n)

Data streaming example



stream of n numbers; permutation from 1 to n ;
one number is missing;
we are allowed **one pass** over the data

Solution 2:
(closed form)

$$s = \frac{n(n+1)}{2} - \sum_{j \leq i} \pi[j]$$

subtract seen
numbers

sum of all numbers from 1 to n memory: $2 \log n$

Data streaming example



we are allowed **one pass** over the data,
can only **store three numbers**

- **Average**: can be computed by keeping track of two numbers (sum and #numbers seen)
- **Median**: sample data points - but how?

Data streaming contd.

- Typically: **simple functions** of the stream are computed and used as input to other algorithms
 - Median
 - Number of distinct elements
 - Longest increasing sequence
 - ...
- **Closed form solutions** are rare
- Common approaches are **approximations** of the true value: **sampling, hashing**

A brief introduction to MapReduce

MapReduce is an industry standard

Hadoop is the open-source implementation of MapReduce framework.

Scaling Pinterest

5:20pm - 6:10pm

By: **Yash Nelapati**, **Marty Weiner**

Infrastructure Engineer, Pinterest-- Engineer, Pinterest

Pinterest grew to one of the world's largest social networks in just a few years. The first year and half was a scalability rocket ship. We had to grow the architecture hyper fast without much sleep, and had the opportunity to try lots of things, and make LOTS of mistakes before starting to get things under control.

Stop by and ask anything.

We'll give a quick overview of our architecture, some of the new systems we're building (Pinball, Frontdoor), and talk about some of the tech we use / used for databases (MySQL, MongoDB, Casandra, etc), caching (Memcache, Redis), logging (Flume, Kafka), map reduce (EMR, Qubole, Redshift), logic (Python, Java, Go, Nutcracker), load balancing / HA (haproxy, nginx, Varnish, ZooKeeper), server management (Puppet), and others. We'll keep the presentation relatively short and open the floor for any and all questions.

Data & Infrastructure at Airbnb

1:35pm - 2:25pm

By: **Brenden Matthews**

Software Engineer at Airbnb

At Airbnb, we want to change the way people travel. To accomplish that, we need to change the way we think about infrastructure. By leveraging Mesos, we've built out our next generation of infrastructure to support several frameworks like Hadoop and Storm. Mesos paves the way for application level distributed computing, and is poised to become the chassis of distributed computing for the future.

Attendees will gain insight into building, deploying and running a Mesos cluster with several frameworks, such as Hadoop and Storm.

Scaling AncestryDNA using Hadoop and HBase

2:50pm - 3:40pm

By: **Bill Yetman**, **Jeremy Pollack**

Senior Director of Engineering at Ancestry.com-- Senior Software Engineer, Ancestry.com

What do you get when you take Bioinformatics Scientists with PhDs and mix them up with Software Engineers? Why Ancestry DNA on Hadoop and HBase! Get the whole story from both the management (Bill Yetman, Sr. Director) and developer (Jeremy Pollack, Principle Engineer/Team Lead) points of view. Find out how this unique cast of characters took academic programs and created an industrial, scalable, DNA processing pipeline (a real Big Data problem) using Hadoop and HBase. The final implementation provided a 1700% performance improvement.

QCon 2013 (San Francisco)

“QCon empowers software development by facilitating the spread of knowledge and innovation in the developer community. A practitioner-driven conference ...”

Industry is moving fast

Monday, 3 November

Architectures You've Always Wondered about

The newest and biggest Internet architectures

Real World Functional

Putting functional programming concepts to work in the real world.

The Future of Mobile

The future of mobile and performance improvements

Continuous Delivery: From Heroics to Becoming Invisible

Continuous Delivery philosophies, cultures, hiccups, and best practices.

Unleashing the Power of Streaming Data

This track explores a variety of use-cases, platforms, and techniques for processing and analyzing stream data from the companies deploying them at scale!

Tuesday, 4 November

Engineering for Product Success

Architectures that make products more successful

Reactive Service Architecture

Reactive, Responsive, Fault Tolerant and More.

Modern CS In the Real World

How modern CS tackles problems in the real world.

Applied Machine Learning and Data Science

Understand your big big data!

Deploying at Scale

Containerizing Applications, Discovering Services, and Deploying to the Grid.

Wednesday, 5 November

Beyond Hadoop

Emerging Big Data Frameworks and Technology

Scalable Microservice Architectures

This track addresses the ways companies with hundreds of fine-grained web-services (e.g. Netflix, LinkedIn) manage complexity!

Java at the Cutting Edge

The latest and greatest in the Java ecosystem

Engineering culture

Successes and failures in creating an engineering culture.

Next gen HTML5 and JS

How Web Components, the Future of CSS, and more are changing the web.

Industry is moving fast

Monday Nov 16

Architectures You've Always Wondered About

Silicon Valley to Beijing: Exploring some of the world's most intriguing architectures

Applied Machine Learning

How to start using machine learning and data science in your environment today. Latest and greatest best practices.

Browser as a platform (Realizing HTML5)

Exciting new standards like Service Workers, Push Notifications, and WebRTC are making the browser a

Modern Languages in Practice

The rise of 21st century languages: Go, Rust, Swift

Tuesday Nov 17

Containers in Practice

Build resilient, reactive systems one service at a time.

Architecting for Failure

Your system will fail. Take control before it takes you with it.

Modern CS in the Real World

Real-world Industry adoption of modern CS ideas

The Amazing Potential of .NET Open Source

From language design in the open to Rx.NET, there is amazing potential in an Open Source .NET

Wednesday Nov 18

Streaming Data @ Scale

Real-time insights at Cloud Scale & the technologies that make them happen!

Taking Java to the Next Level

Modern, lean Java. Focuses on topics that push Java beyond how you currently think about it.

The Dark Side of Security

Lessons from your enemies

Taming Distributed Architecture

Reactive architectures, CAP, CRDTs, consensus systems in practice

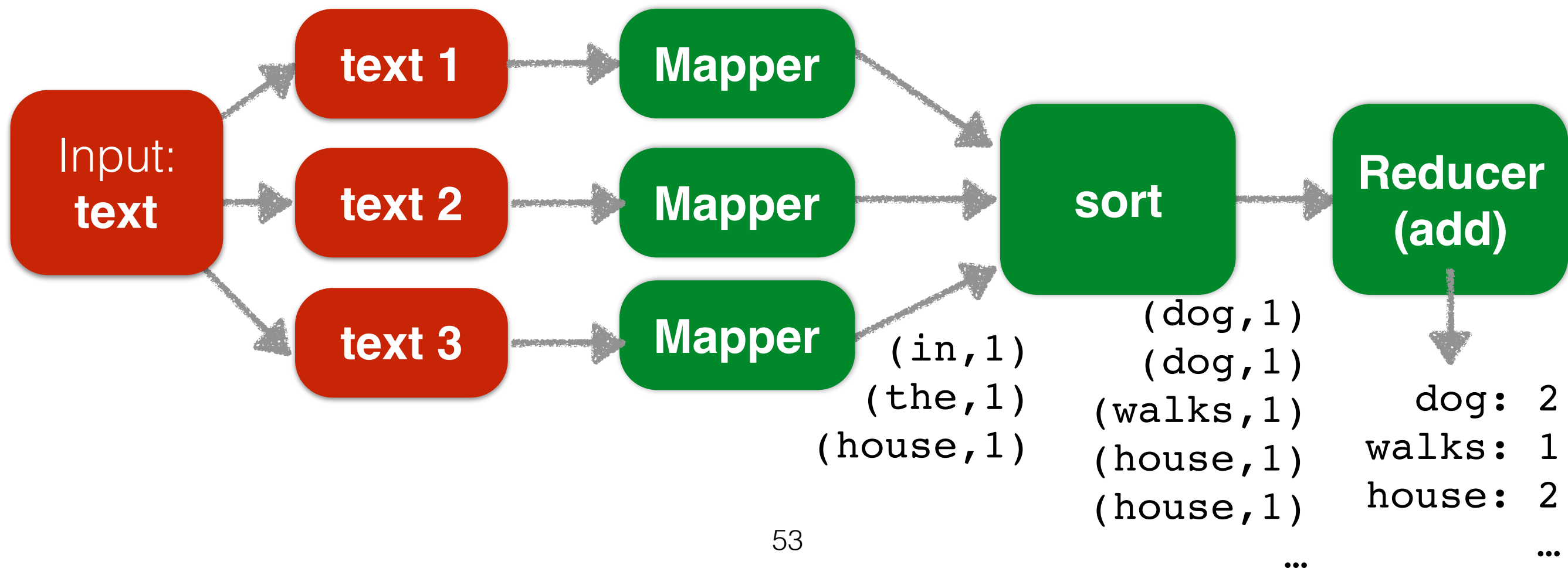
MapReduce

- Designed for **batch processing** over large data sets
- **No limits** on the number of passes, memory or time
- Programming model for distributed computations inspired by the **functional programming paradigm**

MapReduce example

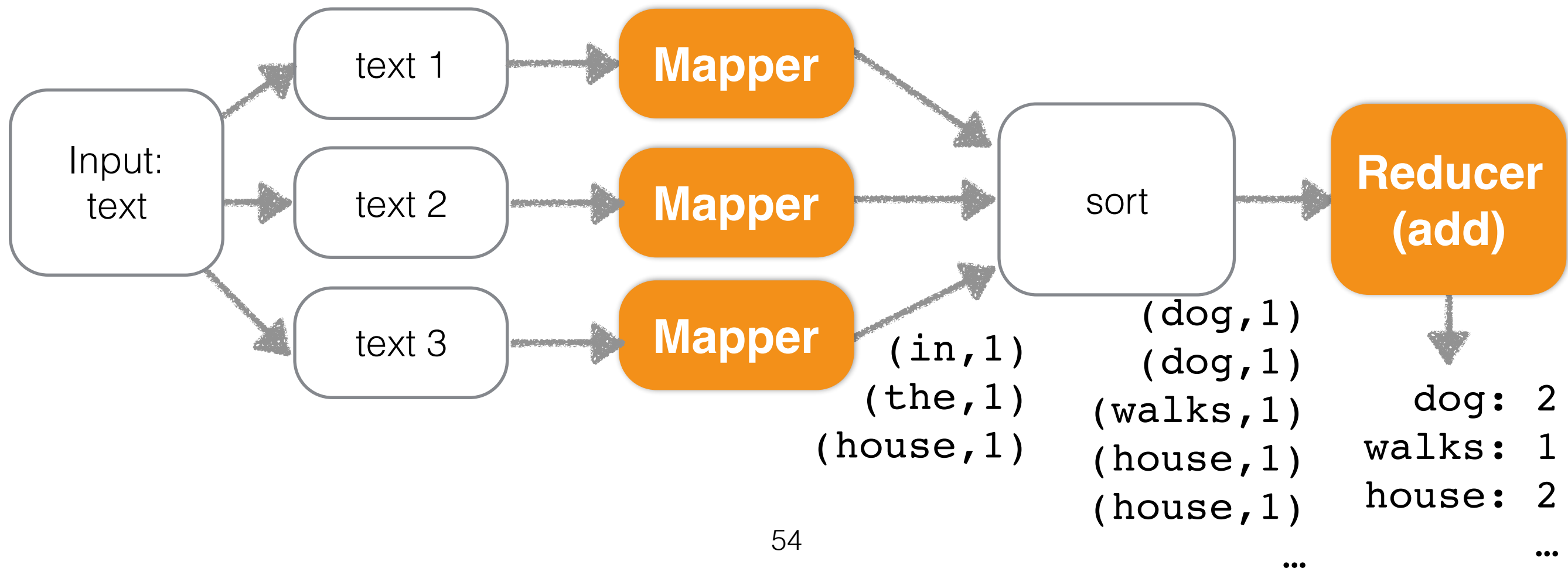
WordCount: given an input text, determine the frequency of each word. The “Hello World” of the MapReduce realm.

Input text: *The dog walks around the house. The dog is in the house.*



MapReduce example

We implement the **Mapper** and the **Reducer**.
Hadoop (and other tools) are responsible for the “rest”.



Summary

- What are the **characteristics** of “big data”?
- **Example use cases** of big data
- A brief introduction of **data streams** and **MapReduce**

Reading material

Required reading

None.

Recommended reading

Principles of Big Data: Preparing, Sharing, and Analyzing Complex Information by Jules Berman.
Chapters 1, 14 & 15.

THE END