W4121
Computer Systems for Data Science
Spring 2017

Sambit Sahu, Eugene Wu

Computer Science Department
Columbia University

1

Data

2

Data
is for serious business

3

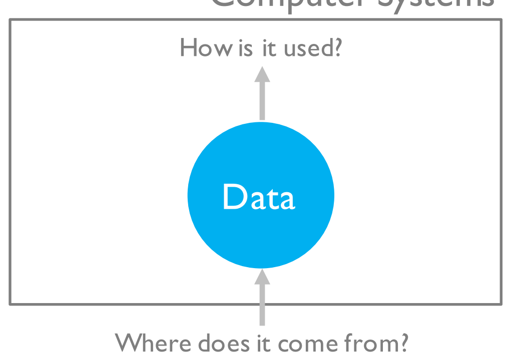
Data
is at the center of most things.

4

Data
is at the center of *everything*

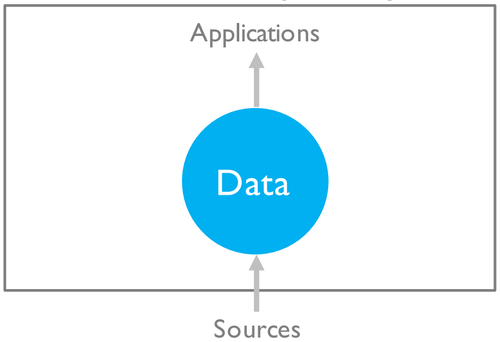
5

Computer Systems



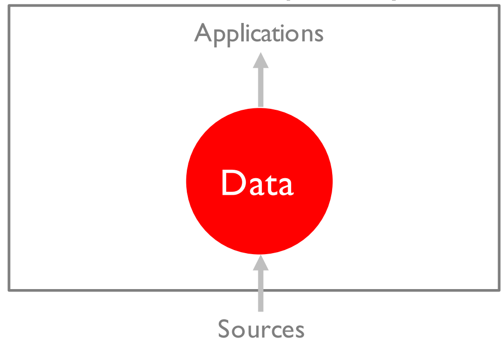
6

Computer Systems



7

Computer Systems

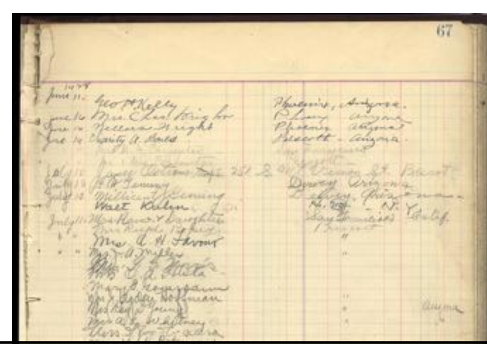


8

How did we get here?

11

Data was *Manual*

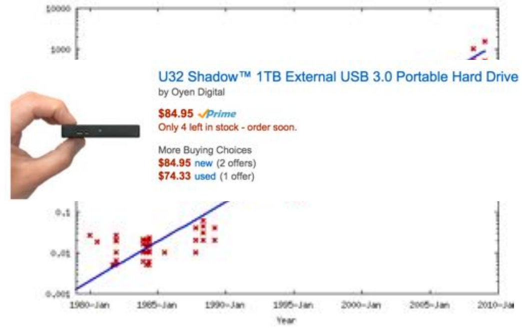


12

Data was *Expensive*

13

Data is *Cheap*



14

Data is *Automated*

Physical devices

15

Data is *Automated*

Physical devices
Software logs

16

Data is *Ubiquitous*

Physical devices
Software logs
Phones

17

Data is *Ubiquitous*

Physical devices
Software logs
Phones
GPS/Cars

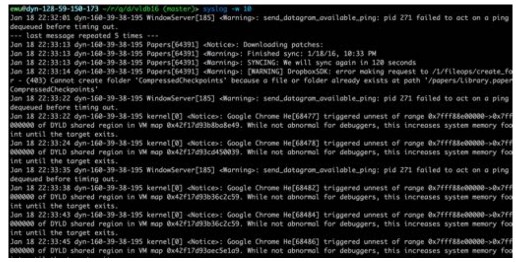
18

Data is *Everywhere*

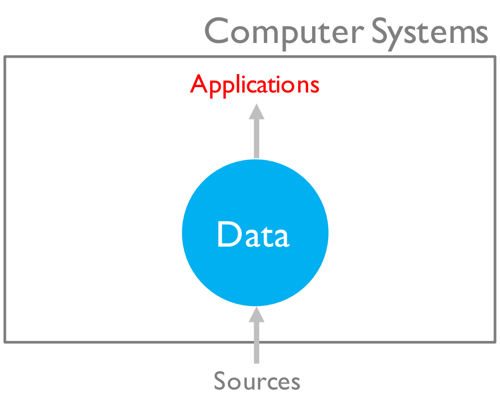
Physical devices
Software logs
Phones
GPS/Cars
Internet of *Things*

19

Data is *Temporal*



20



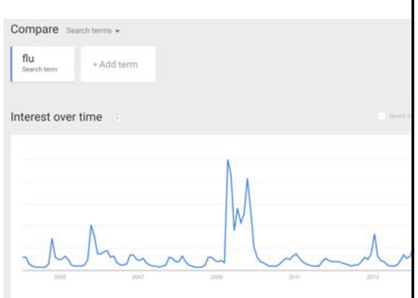
21

Data Science Applications

Health

Data Science Applications

Health



https://www.google.org/flutrends/

23

Data Science Applications

Health

Thank you for stopping by.

Google Flu Trends and Google Dengue Trends are no longer publishing current estimates of Flu and Dengue fever based on search patterns. The historic estimates produced by Google Flu Trends and Google Dengue Trends are available below. It is still early days for

https://www.google.org/flutrends/

What are we doing with data?

Health

Politics

What are we doing with data?

Health

Politics

Triumph of the Nerds: Nate Silver Wins in 50 States

What are we doing with data?

Health

Politics


TIME

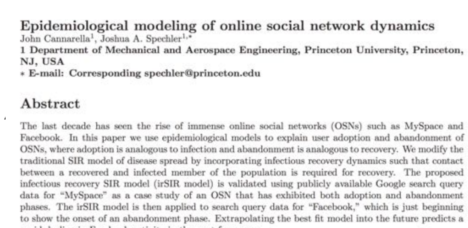
2012 ELECTION

Inside the Secret World of the Data Crunchers Who Helped Obama Win

Data-driven decisionmaking played a huge role in creating a second term for the 44th President and will be one of the more closely studied elements of the 2012 cycle

What are we doing with data?

What are we doing with data?


Epidemiological modeling of online social network dynamics

John Cannarella[1], Joshua A. Spechler[1,*]

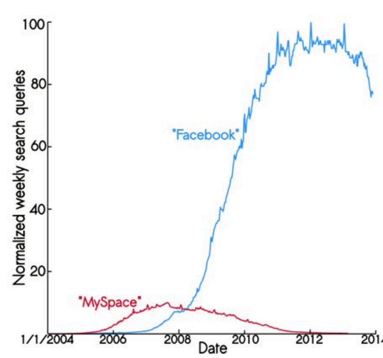
1 Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ, USA

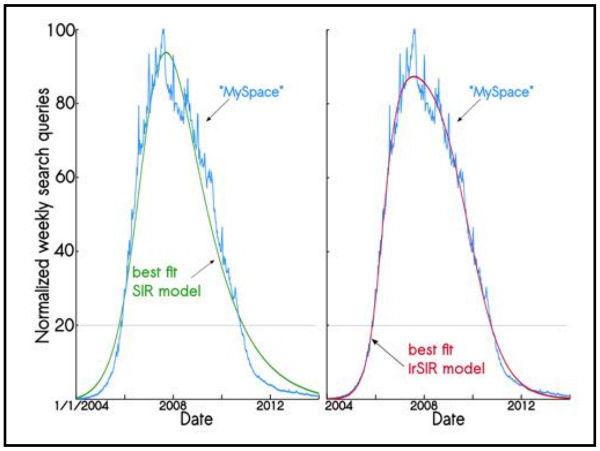
* E-mail: Corresponding spechler@princeton.edu

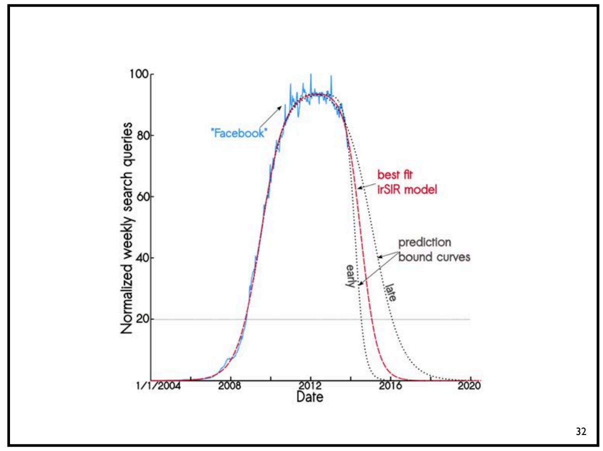
Abstract

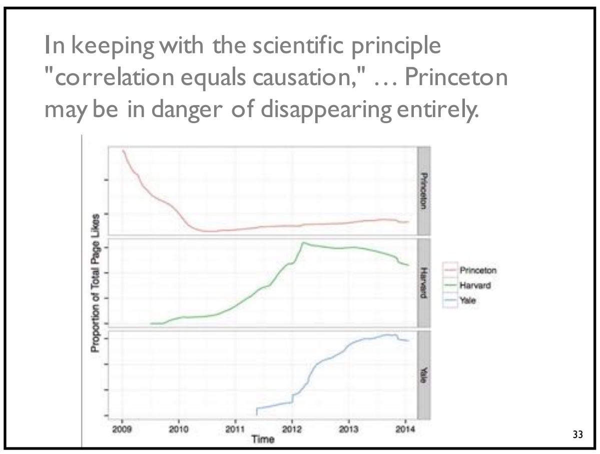
The last decade has seen the rise of immense online social networks (OSNs) such as MySpace and Facebook. In this paper we use epidemiological models to explain user adoption and abandonment of OSNs, where adoption is analogous to infection and abandonment is analogous to recovery. We modify the traditional SIR model of disease spread by incorporating infectious recovery dynamics such that contact between a recovered and infected member of the population is required for recovery. The proposed infectious recovery SIR model (irSIR model) is validated using publicly available Google search query data for "MySpace" as a case study of an OSN that has exhibited both adoption and abandonment phases. The irSIR model is then applied to search query data for "Facebook," which is just beginning to show the onset of an abandonment phase. Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years.

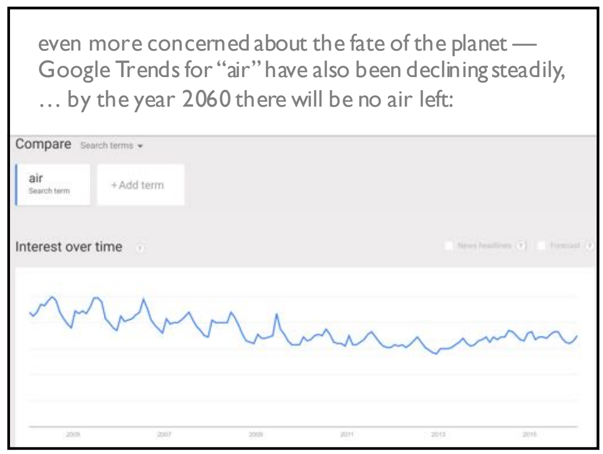
Extrapolating the best fit model into the future predicts a rapid decline in Facebook activity in the next few years

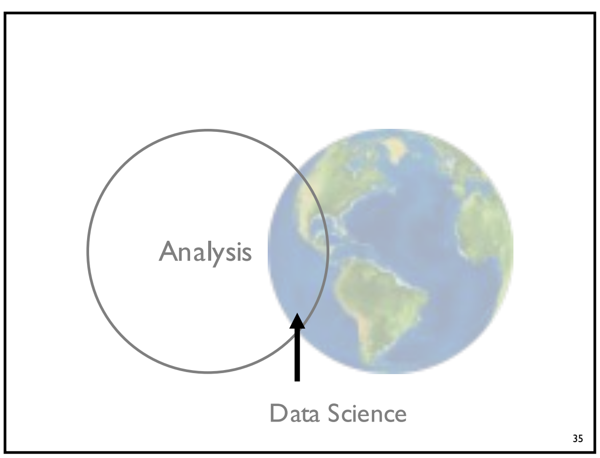


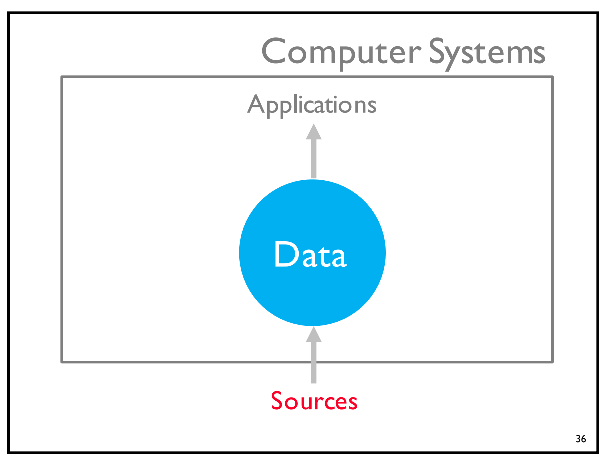

100
80
60
40
20
Normalized weekly search queries
"MySpace"
best fit
SIR model
1/1/2004
2008
2012
Date
"MySpace"
best fit
IrSIR model
2004
2008
2012
Date


100
80
60
40
20
Normalized weekly search queries
"Facebook"
best fit
IrSIR model
prediction
bound curves
early
late
1/1/2004
2008
2012
2016
2020
Date
32


In keeping with the scientific principle "correlation equals causation," … Princeton may be in danger of disappearing entirely.
Proportion of Total Page Likes
Princeton
Harvard
Yale
Princeton
Harvard
Yale
2009
2010
2011
2012
2013
2014
Time
33


even more concerned about the fate of the planet — Google Trends for "air" have also been declining steadily, … by the year 2060 there will be no air left:
Compare
Search terms
air
Search term
+ Add term
Interest over time


Analysis
Data Science
35


Computer Systems
Applications
Data
Sources
36

Autogenerated – record every…

Mouse click
Car drive
Ad impression
Webpage visit
Billing transaction
Network message
Error
Video stream

37

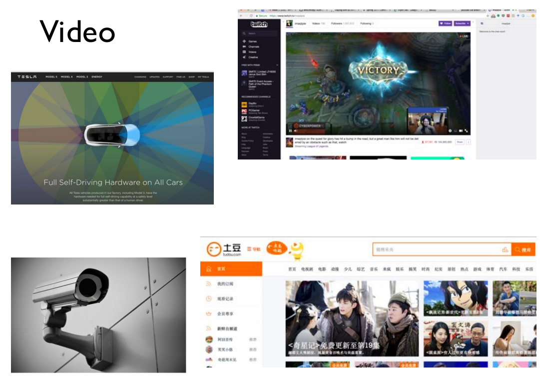
User generated

38

Internet of Things

39

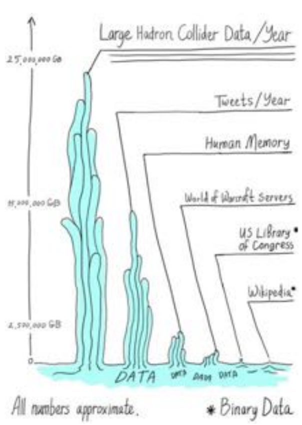
Video



40

Scientific

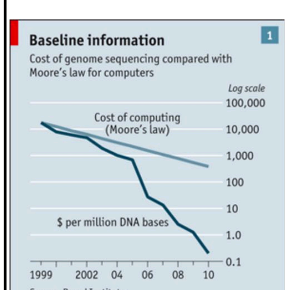
Only limited by the precision of the instrument

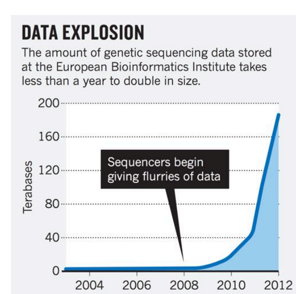


http://www.symmetrymagazine.org/article/august-2012/particle-physics-tames-big-data

41

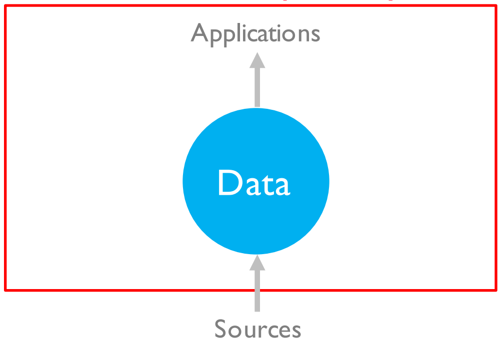
Health/Bioinformatics





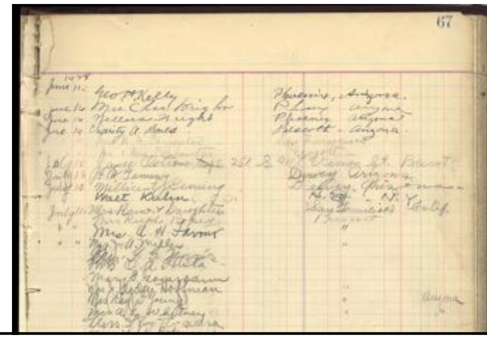
42

Computer Systems



43

Pen and Paper



44

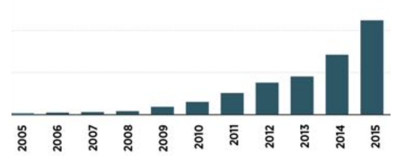
http://www.theatlantic.com/technology/archive/2013/10/computing-power-used-to-be-measured-in-kilo-girls/280633/

45

http://plyojump.com/classes/mainframe_era.php

https://en.wikipedia.org/wiki/IBM_PC_compatible

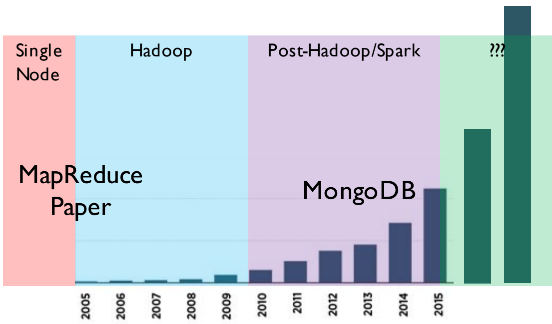
Data Volume Over Time



Note: Post-2013 figures are predicted. Source: UNECE

48

Big Data Systems Over Time



Note: Post-2013 figures are predicted. Source: UNECE

49

Course Goal

Understand fundamental principles behind large-scale data science systems

Data Processing Techniques for "big" data

Experience with modern data science tools (Spark)

50

Course Topics

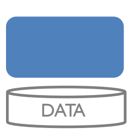
Spark/Application

Data Processing

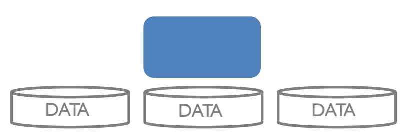
Fundamentals

51

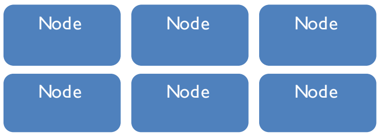
52



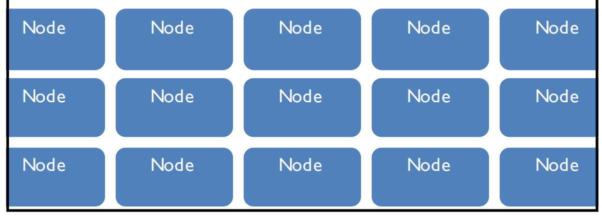
53



54

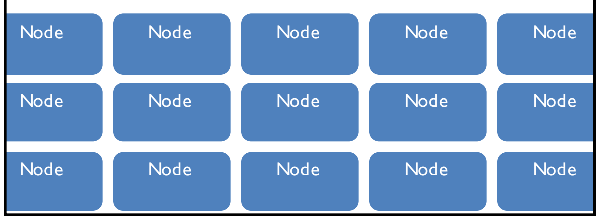


55



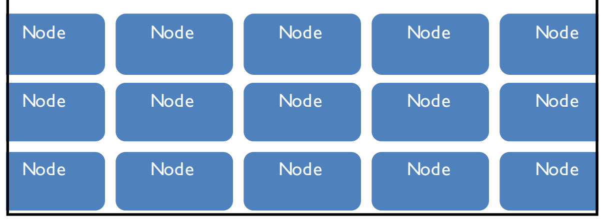
Fundamental Issues

Applies to *any* multi-node system



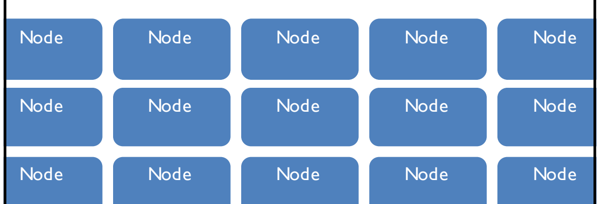
Data Processing Issues

How to lay out, index, structure data to answer queries quickly, correctly



Spark/Application Issues

How to use modern distributed computing system?



Big Data Systems in the Wild

Spark
Google Cloudflow
Azure Cloud
AWS/Redshift
Tensorflow
…

60

Course Structure Overview

Three key modules and focus areas:

Computing models
1. Batch processing with Map Reduce and higher level programming construct
2. Real-time responsive analytics with Spark and Spark Streams

Data and storage
1. Various data models and storage
2. Graph processing and big data visualization

Big data systems
1. Distribute systems and challenges in extremely large scale systems
2. Hard problems in distributed systems and trade-offs
3. Google's storage and cluster computing stack

Designing Machine Learning Systems with Big Data

61

Course Administrative Details

1. Course materials
   1. Primarily lecture notes. Additional reference readings will be provided as needed based on the lecture topics including research papers.
2. All course related submissions will be done using courseworks.
3. Important deadlines and communications will be done using Courseworks Announcement.
4. 3-4 TAs will be available to assist in the course. We will announce their contact emails.
5. Grading
   1. 60% Homework
   2. 30% Tests/Quizzes
   3. 10% Class participations (ask/answer questions)
6. Good programming background in one of the languages Python/Java

62

Logistics

Register with piazza

We will not answer direct emails

HW0 will be up soon

Very simple, but required.

63

Collaboration Policy

Read Syllabus on course site for allowed conduct

CS Dept academic honesty policies
http://www.cs.columbia.edu/education/honesty

We will not tolerate *any* cheating

64

Computation on Big Data

- Computation on huge amount of data is not a luxury – it is a necessity!
- Imagine Facebook logs for logins. FB wants to compute how many people are logging in from which continents for each hour.
- Let us see how to compute this.
- What is the big deal?
  - – How big is the data?
  - – Huge data – the data file does not fit into single server's disks…how do you compute if data does not even fit into server's storage?
  - – Data is on multiple servers – on a cluster of servers. So how do you compute and where do you compute what?
  - – How do we compute the final results?
  - – Who takes care of some machine or computing failure?
  - – How do you automate such computations spread across machines?

65

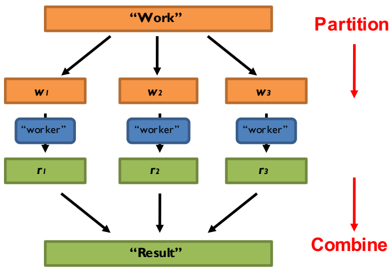
Computing Models for Big Data

We will learn two computing paradigm for big data on a cluster of machines

1. Batch processing with Map Reduce
   1. Idea is to divide and conquer the task – compute partial results on smaller chunks of data and then merge the partial results to compute final result
   2. Move computing task to where data is
- Real-time processing with Spark
  - Map Reduce is great but too slow due to lot of disk based operations
  - Spark computes with in-memory data

66

Divide and Conquer



So what is Hadoop/Map Reduce

Hadoop/Map Reduce is a computing system on a cluster of machines that provide at the minimum the following

- Storage across a cluster of machines (HDFS)
- A computation model to divide-conquer a task (map-reduce)
- A runtime to enable map-reduce style of computation

Why MapReduce not efficient for iterative computations?

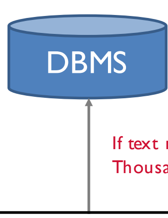
- MapReduce is an excellent computing model that scales for log processing type of computations described earlier.
- What about iterative models that use the same data again and again?
  - Every operation is to read and write to disk. So every iteration requires reading and writing to disk. Too many disk based operations for iterative computing.
  - Many machine learning based computations are iterative in nature.
- So what is the solutions? Can the data be somehow kept in memory until all the operations on it completes…
- **Spark Model**: Resilient Distributed Datasets (RDD)
  - Recent computing model that is 100x faster and more suitable for iterative and real-time analytics
  - We will learn how to write real-time analytics using Spark and Spark Streams.

Data Processing Topics

Data models
Data cleaning
  Data wrangling, Entity Resolution, Explanation
Large scale analytics
Visualizations and scaling them

How does data get into a DBMS?

Entity resolution
Data extraction
Missing data
…



Extract
Transform
Load

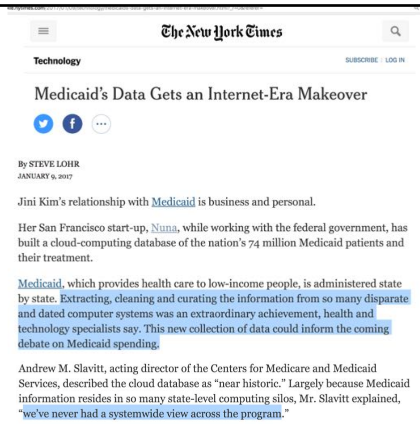
If text matches XXX, then…
Thousands of rules

How does data get into a DBMS?

Entity resolution
Data extraction
Missing data
…

Extract
Transform
Load

verizon



The New York Times

Technology

SUBSCRIBE | LOG IN

Medicaid's Data Gets an Internet-Era Makeover

By STEVE LOHR
JANUARY 9, 2017

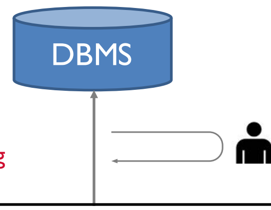
Jini Kim's relationship with Medicaid is business and personal.

Her San Francisco start-up, Nuna, while working with the federal government, has built a cloud-computing database of the nation's 74 million Medicaid patients and their treatment.

Medicaid, which provides health care to low-income people, is administered state by state. Extracting, cleaning and curating the information from so many disparate and dated computer systems was an extraordinary achievement, health and technology specialists say. This new collection of data could inform the coming debate on Medicaid spending.

Andrew M. Slavitt, acting director of the Centers for Medicare and Medicaid Services, described the cloud database as "near historic." Largely because Medicaid information resides in so many state-level computing silos, Mr. Slavitt explained, "we've never had a systemwide view across the program."

How does data get into a DBMS?

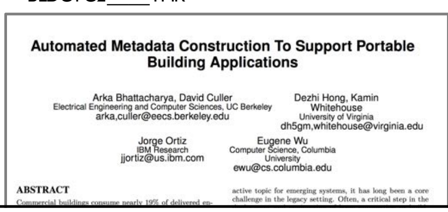


Interactive
data cleaning

How does data get into a DBMS?

Text → data records

- Building sensor IDs – no consistency, arbitrary
- BLDA1C600A_ART
- BLDC1C2____TMR



Automated Metadata Construction To Support Portable Building Applications

Arka Bhattacharya, David Culler
Electrical Engineering and Computer Sciences, UC Berkeley
arka,culler@eecs.berkeley.edu

Dezhi Hong, Kamin Whitehouse
University of Virginia
dh5gm,whitehouse@virginia.edu

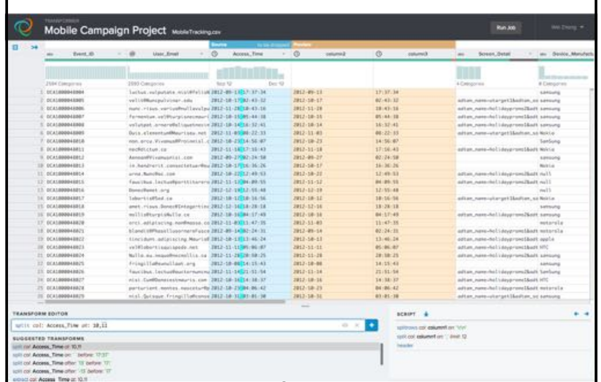
Jorge Ortiz
IBM Research
jjortiz@us.ibm.com

Eugene Wu
Computer Science, Columbia University
ewu@cs.columbia.edu

ABSTRACT

Commercial buildings consume nearly 19% of delivered en-

active topic for emerging systems, it has long been a core challenge in the legacy setting. Often, a critical step in the



Large scale analytics

Data volumes too large to even scan once

How to deal with this?

- Spend more time
- Concurrency
- Reduce data size
- Read less data
- Do less work
- Waste less time doing work

Large scale analytics

- Columnar databases
- In-memory databases
- Intermediate results
- Graph "databases"
- Sketching and sampling

Visualization

How to think about and approach visualization
Modern visualization tools
How to scale visualizations