UNIVERSITY of WASHINGTON

Aside on “Big Data”

Bill Howe, PhD
Director of Research,
Scalable Data Analytics
University of Washington
eScience Institute

Big Data: Three challenges

- Volume
  - the size of the data
- Velocity
  - the latency of data processing relative to the growing demand for interactivity
- Variety
  - the diversity of sources, formats, quality, structures

Astronomy

LSST (~100PB; images, spectra)

PanSTARRS (~40PB; images, trajectories)

SDSS (~400TB; images, spectra, catalogs)

telescopes

n-body sims

spectra

Ocean Sciences

models

stations

AUVs

gliders

cruises, CTDs

flow cytometry

ADCP

satellites

OOI (~50TB/year; sims, RSN)
IOOS (~50TB/year; sims, satellite, gliders, AUVs, vessels, more)
CMOP (~10TB/year; sims, stations, gliders, AUVs, vessels, more)

# of bytes

# of data sources

Big Data

"Big Data is any data that is expensive to manage and hard to extract value from."

Michael Franklin
Thomas M. Siebel Professor of Computer Science
Director of the Algorithms, Machines and People Lab
University of Berkeley

Key idea: "Big" is relative! "Difficult Data" is perhaps more apt!

History of the term "Big Data"

Erik Larson, 1989, Harper's magazine

"The keepers of big data say they do it for the consumer's benefit. But data have a way of being used for purposes other than originally intended."

Takeaway: private data is becoming commoditized

Predates the rise of the Internet, but foreshadows emerging topics in data science: ethics, validation, privacy

Big Data History

"E-commerce, in particular, has exploded data management challenges along three dimensions: volumes, velocity and variety."

On Volume:
"The lower cost of e-channels enables and enterprise to offer its goods or services to more individuals or trading partners, and up to 10x the quantity of data about an individual transaction may be collected—thereby increasing the overall volume of data to be managed."

On Velocity:
"E-commerce has also increased point-of-interaction (POI) speed, and consequently the pace data used to support interactions and generated by interactions"

On Variety:
"Through 2003/04, no greater barrier to effective data management will exist than the variety of incompatible data formats, non-aligned data structures, and inconsistent data semantics."

Doug Laney, "3-D Data Management: Controlling Data Volume, Velocity and Variety", Gartner, 2001

History of the term "Big Data"

"Big Data ... and the Next Wave of InfraStress"

-- *John R. Mashey, former Chief Scientist, SGI*

Takeaway: Disk capacities growing incredibly fast, disk latencies not keeping pace: trouble ahead!

A technology-oriented view of Big Data

Big Data Now

"…the necessity of grappling with Big Data, and the desirability of unlocking the information hidden within it, is now a key theme in all the sciences – arguably the key scientific theme of our times."

Francis X. Diebold
Paul F. and Warren S. Miller Professor of Economics
School of Arts and Sciences
University of Pennsylvania

Where does big data come from?

- "data exhaust" from customers
- new and pervasive sensors
- the ability to "keep everything"

Car black boxes: Privacy nightmare or a safety measure?

February 15, 2013 | By Ronald D. White

What if the black box in your new car becomes a tool to invade your privacy? What if, on the other hand, it winds up saving your life after an accident?

Those are some of the questions being raised this week over black box data event recorders in cars. Privacy advocates worried on Thursday that the data could be misused. Safety advocates argued on Friday that a watered-down version of the recorders would slow safety innovations.

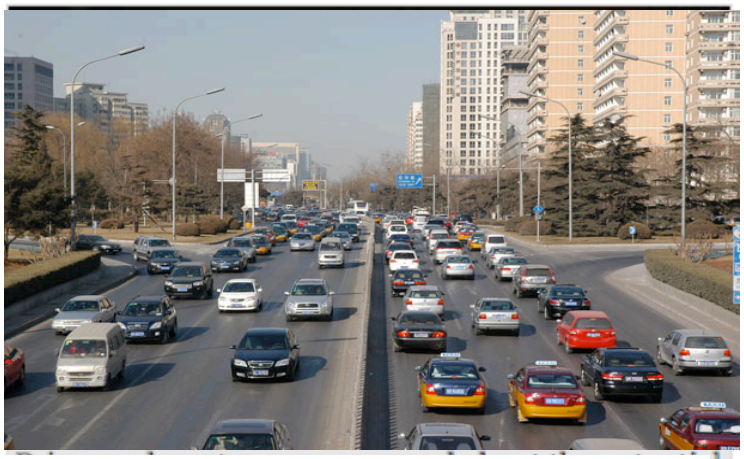
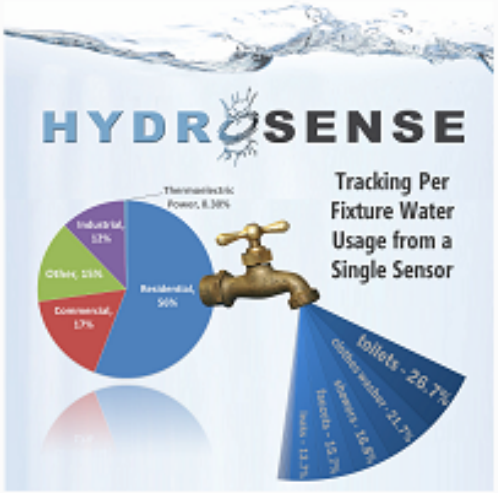


photo in public domain

Angeles...)

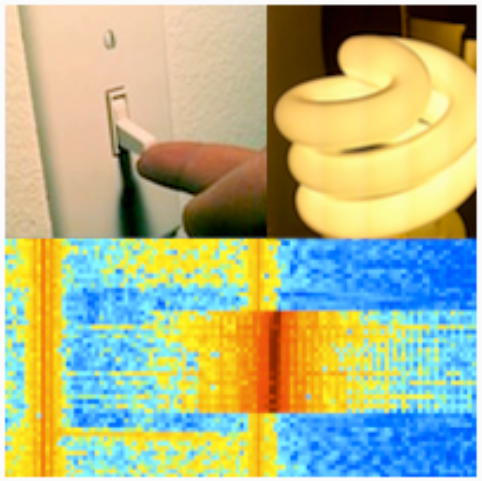


HydroSense

Water Fixture Usage with a Single Sensor

HydroSense is a pressure-based sensor that automatically determines water usage activity and flow down to the source (e.g., dishwasher, laundry, shower) from a single non-intrusive installation point.

Lead Researchers: Jon Froehlich, Eric Larson, Shwetak Patel



ElectriSense

Electrical Device Energy Usage with a Single Sensor

ElectriSense is a single plug-in sensor that provides whole home device level usage data. That is, using a single sensor plugged in anywhere in the home, ElectriSense can infer which electrical appliances are on and which off. This data could be used for numerous applications, for example, for providing home owners with itemized electrical bill that not only shows the total energy consumption but breaks the total on a per appliance basis (TV consumed 20 KWh, Lighting consumes 18 KWh and so on).

Lead Researchers: Sidhant Gupta, Shwetak Patel