Training for Big Data

Learnings from the CATS Workshop

Raghu Ramakrishnan
Technical Fellow, Microsoft

Head, Big Data Engineering
Head, Cloud Information Services Lab

What is Big Data?

- Store any kind of data
  - Files, relations, docs, logs, graphs, multimedia …
    - HDFS
- Do any type of analysis
  - SQL queries, ML, image processing, log analytics …
    - Hive, Map-Reduce, Mahout, …
- In any mode
  - Batch, interactive, streaming
    - Hive, Spark, Storm
- At any scale
  - “terabytes or more” or “more than your old system”
  - Elastic capacity, commodity hardware

Why is it Important?

Data is Now a Core Asset

- Big Data systems can be transformative
  - More data: More is observable, measurable
  - Less cost: Data acquisition, storage, compute, cloud
  - Easier: Elastic capacity, clean-as-you-go
  - Scalable analytic tools: Can get more out of data
- We can cost-effectively do things we couldn't contemplate before, the Fourth Paradigm can be brought to bear on a wide range of domains:
  - *Data-driven* science, commerce, government, social programs, manufacturing, medicine …
- Biggest bottleneck—trained people

Privacy, Security, Law

Big Data Applications

Web Scale

1.5+M read rps

4 Key Platforms

10,000+ servers

10 Geo Zones

102B emails/month

13+B Ad serves/day

11B visits/month

Web Images Video Local Finance More

Web Search

Explore Y! Palo Alto new | My Yahoo! | Preview email with the Yahoo! Toolbar [redacted] Hi, vickyb6 | Sign Out | Page Options

9/11 Remembered

From tragedy to tango

- Story behind iconic photo
- Healing at home plate
- Complete coverage

YAHOO! SITES Edit

- Mail (4)
- Autos
- Dating
- Finance (Dow Jones)
- Flickr
- Games
- Health
- Horoscopes
- Jobs
- Messenger
- Movies
- News
- omg!
- Real Estate
- Shine
- Shopping
- Sports
- Travel
- TV
- Weather (66°F)

More Yahoo! Sites

MY FAVORITES Edit

TODAY - September 06, 2011

Texas fires kill 4, destroy more than 1,000 homes

At least 180 blazes have erupted across the rain-starved state, including one catastrophic inferno near Austin.

Rick Perry calls on elite team to lead wildfire search

Obama's ratings drop to dismal low

A new poll shows record dissatisfaction with the president's performance and policies. Perry surges »

- 'Young guns' draw fire
- Poll: Freedom vs. terrorism
- Tea party shaping 2012

Madonna disses fan's gift | Old Navy's worst items | Obama's ratings sink lower | Beyoncé bares baby bump

9 - 12 of 48

NEWS WORLD LOCAL FINANCE

- Nevada sheriff: IHOP shooter had 2 more guns in van
- Carol Bartz out as Yahoo! chief; interim CEO named
- Postal Service could stop operating early next year
- Sources: Obama to propose $300 billion to jump-start jobs
- Police: Man breaks in to Celine Dion's home, raids fridge
- Colo. State student loses legs as she tries to hop train
- San Jose school honors Flight 93 pilot - SJ Mercury News
- Summer to make brief appearance, then go... - S.F. Chronic...
- Muramatsu rolls over opponents in East... - Sunnyvale Sun
- Fantasy Football · MLB · NFL · NCAAF · Golf · NASCAR

updated 08:47 pm More: News | Sept. 11 Anniversary

Markets: Dow: 11,139.29 -0.89% Nasdaq: 2,473.83 -0.26%

Enter stock symbol Get Quotes Scottrade Open An Account

TRENDING NOW

1. BCS rankings
2. Amanda Knox
3. Jessica Simpson
4. Fried bubble gum
5. Allergy season
6. Freddie Mercury
7. Anthony Michael ...
8. Bankruptcy
9. Carol Burnett
10. Aerosmith

AdCho

The All-New Nissan VERSA SEDAN

INNOVATION UPSIZED. INNOVATION FOR ALL.

From $10,990*

Click for 360 Tour

Visit NissanUSA.com - Ad Feedback

VIDEO PICKS

Go to Video

Tennis star Novak Djokovic dances with fan

Kirsten Dunst in end-of-world film 'Melancholia'

Famous runaway cow turns herself in

Check out the latest models Y! Autos

~2B User Ids

~750M Uniq Users*

175+M Users in US

285+M Mail users

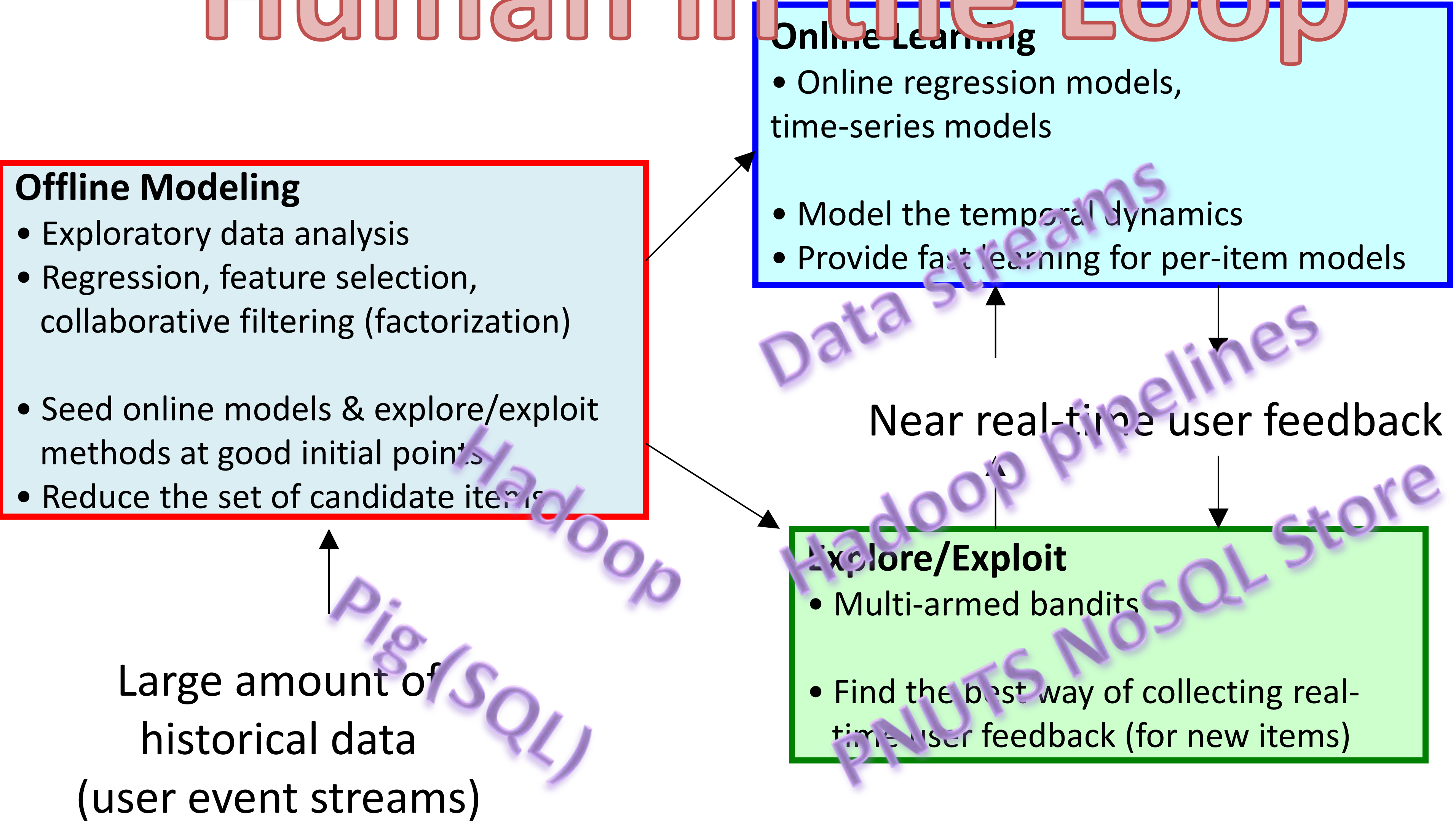
40+ Countries

All Data from July/Aug. Worldwide unless indicated to the contrary

* Yahoo!-branded sites

CORE Modeling Overview

Human in the Loop



Online Learning
- Online regression models, time-series models
- Model the temporal dynamics
- Provide fast learning for per-item models

Offline Modeling
- Exploratory data analysis
- Regression, feature selection, collaborative filtering (factorization)
- Seed online models & explore/exploit methods at good initial points
- Reduce the set of candidate items

Explore/Exploit
- Multi-armed bandits
- Find the best way of collecting real-time user feedback (for new items)

Near real-time user feedback

Large amount of historical data
(user event streams)

Data streams

Hadoop pipelines

Hadoop

Pig (SQL)

PNUTS NoSQL Store

Traditional ecology

Field work

Experiments

Theorizing

(Slide courtesy Drew Purves, MSR)

Some huge questions we can’t answer

Will forests accelerate or slow climate change?

How can we safely genetically engineer crops?

Is there enough to water for both agriculture, and industry, in the future?

How can we feed 9+ billion humans, with less water, less oil and less phosphorous?

How many species are there on Earth? How can we predict them?

What would we do if a new disease hit wheat? Or the pollinators died out?

How can we optimize supply chains to minimize environmental impact?

Will the world become more fire prone as the climate changes?

(Slide courtesy Drew Purves, MSR)

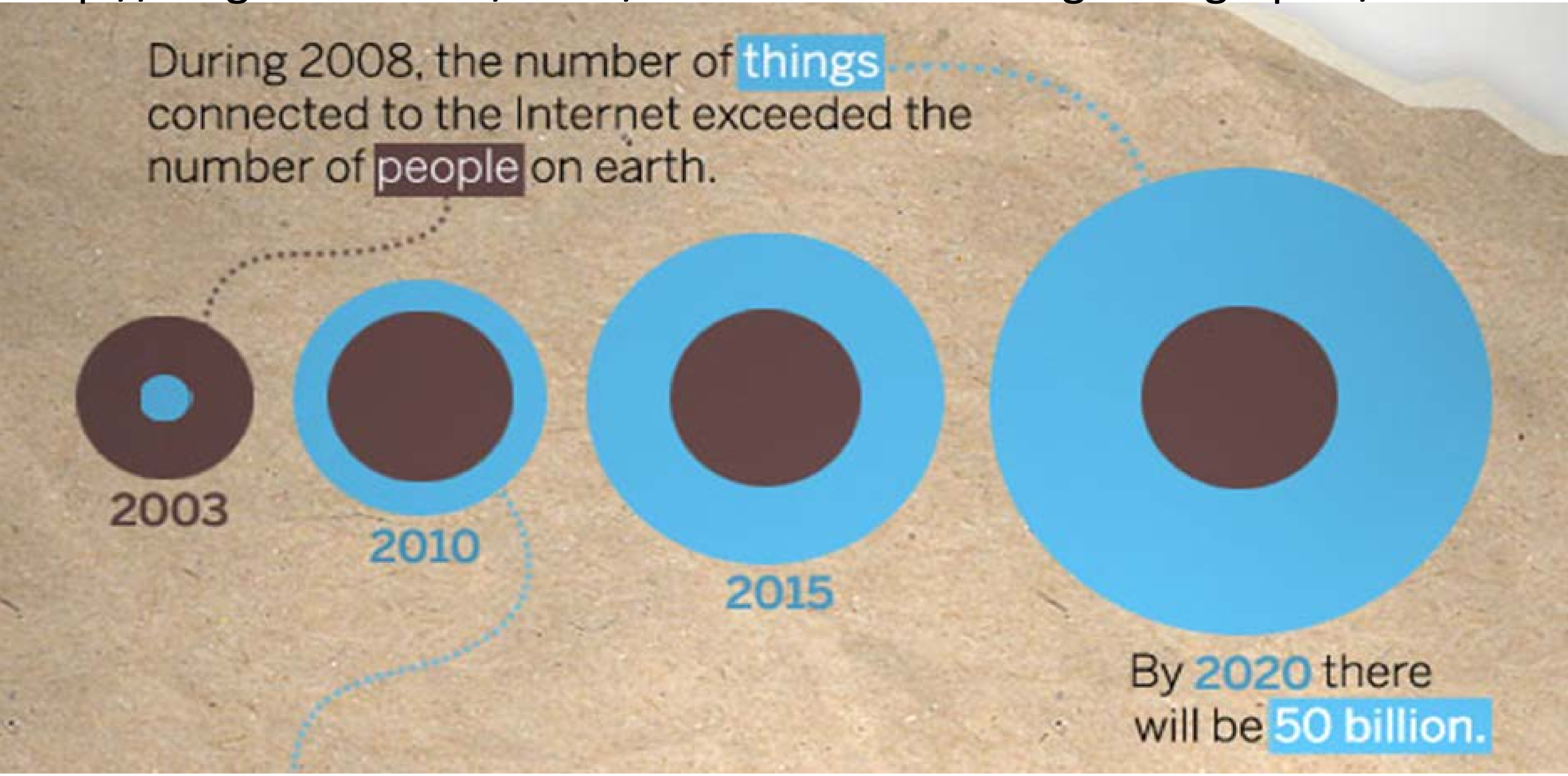
Enter a new kind of ecology?

Traditional Ecology	“Joined up” Ecology
Qualitative insights	Quantitative predictions
Driven by academic curiosity	Driven by society’s needs
Fragmented into subdisciplines, divorced from other fields of study	Integrated across subdisciplines, and with other fields of study
Divorced from policy	Connected to policy
Huge shortage of all data	Huge abundance of some data, huge shortage of other data
Computation and statistics an afterthought – a necessary evil!	Computational techniques and statistics central – and exciting!
A disparate set of software tools	An interoperable suite of software tools

(Slide courtesy Drew Purves, MSR)

IoT: Connected Devices, Continuous Observations, Instantaneous Action, Rich History

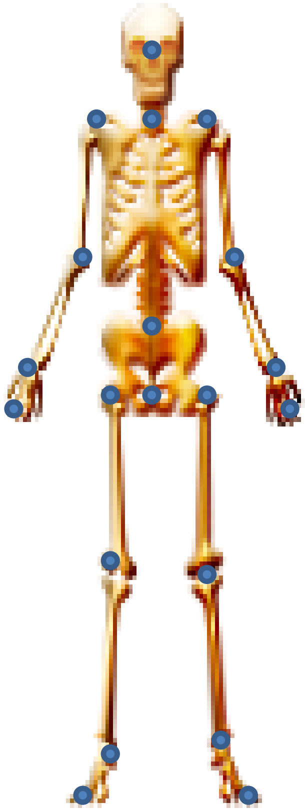
http://blogs.**cisco**.com/news/the-internet-of-things-infographic/



Gartner: *50 billion intelligent devices by 2015* and *275 Exabytes per day of data being sent across the Internet by 2020*

Data Created by People—Kinect

- The Kinect is an array of sensors.
  - Depth, audio, RGB camera…



- SDK provides a 3D virtual skeleton.
  - 20 points around the body.
  - 30 frames per second.
- Inexpensive and ubiquitous.
  - Between 60-70M sold by May 2013.

(Slide courtesy Assaf Schuster, Technion)

Big Data Training

What should we teach? To whom, and how?

“Analysis of Big Data requires cross-disciplinary skills, including the ability to make modeling decisions while trading off optimization and approximation, and being attentive to system robustness.”

Training Students to Extract Value from Big Data

(NAE report on a workshop organized by CATS)

Skills to Teach

- Data management
- Machine Learning
- Parallel computing
- Security
- Software engineering
- Statistical analysis and inference
- Visualization
- Grounding in application domain
  - Formulating real problems in terms of actionable outcomes from analyzing specific data
  - Objective evaluation of outcomes and iterative improvement

Very cross-disciplinary; ubiquitous in applicability across domains;
any given application typically requires careful use of domain expertise

Data Analysis Pipeline

- Ask a general question
- Frame it in light of available and relevant data
  - Understanding scope of inferences, sampling, biases
- Query and transform the data to prepare it
- Exploratory analysis
  - Get a feel for the data, and what insights it could offer
  - Understanding randomness, variability, uncertainty, sparsity
- Modeling
  - Dimension reduction, feature selection, prediction, classification, parameter estimation
- Model forensics, interpretation
  - Residuals, model comparison, model uncertainty
- Actionable insights
  - Integration into real-world context

Domain knowledge, intuition

Data management, computing, security

Statistics, ML

Visualization helps at all stages

Curricula, Programs

- Should Big Data and Data Science be a new discipline, or seen as a complementary skillset for other disciplines?
  - New major or minor?
  - Certificate programs? Masters programs?
  - How should curricula/programs be developed? Role of parent disciplines/departments?
- Do we need to design new courses to build effective curricula?
  - Or can we create curricula by taking collections of courses from the parent disciplines?
- Should we introduce at undergrad or grad level?
- Role of intense courses ("boot camps"), MOOCs, etc.

Curricula should be developed jointly across parent disciplines

Big Data Links

- H.V. Jagadish et al., CACM, July 2014
  - Big Data and its Technical Challenges
- D. Agrawal et al., CACM 56(6):92-101 (2013)
  - Content Recommendation on Web Portals
- **NAE Workshop Report, 2014**
  - **Training Students to Extract Value from Big Data**
- Gary Marcus, New Yorker
  - April 3 2013 http://www.newyorker.com/online/blogs/elements/2013/04/steamrolled-by-big-data.html
  - May 23 2013 http://www.newyorker.com/online/blogs/culture/2012/05/google-knowledge-graph.html
- Alan Feuer, NYT, Mar 23, 2013
  - http://www.nytimes.com/2013/03/24/nyregion/mayor-bloombergs-geek-squad.html?pagewanted=all&_r=0
- Quentin Hardy, NYT, Nov 28, 2012
  - http://bits.blogs.nytimes.com/2012/11/28/jeff-hawkins-develops-a-brainy-big-data-company/
- Steve Lohr, NYT, Feb 11, 2012
  - http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html
- Economist, Nov 2011
  - http://www.economist.com/blogs/dailychart/2011/11/big-data-0
- Forbes special report on Big Data, February 2012
  - http://www.forbes.com/special-report/data-driven.html
- T. Hey et al., (Eds) (2013)
  - The Fourth Paradigm: Data-Intensive Scientific Discovery.
- J. Manyika et al., McKinsey Global Institute, May 2011
  - Big Data: The Next Frontier for Innovation, Competition, and Productivity.
- NPR, Nov 2011
  - http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data