Making Sense of Big Data with the Berkeley Data Analytics Stack

Michael Franklin
UC Berkeley
franklin@cs.berkeley.edu

ABSTRACT
The Berkeley AMPLab is creating a new approach to data analytics. Launching in early 2011, the lab aims to seamlessly integrate the three main resources available for making sense of data at scale: Algorithms (machine learning and statistical techniques), Machines (in the form of scalable clusters and elastic cloud computing), and People (both individually as analysts and in crowds). The lab is realizing its ideas through the development of a freely-available Open Source software stack called BDAS: the Berkeley Data Analytics Stack. In the four years the lab has been in operation, we’ve released major components of BDAS. Several of these components have gained significant traction in industry and elsewhere: the Mesos cluster resource manager, the Spark in-memory computation framework, and the Shark query processing system. BDAS features prominently in many industry discussions of the future of the Big Data analytics ecosystem—a rare degree of impact for an ongoing academic project.

Given this initial success, the lab is continuing on its research path, moving “up the stack” to better integrate and support advanced analytics and to make people a full-fledged resource for making sense of data. In this talk, I’ll first outline the motivation and insights behind our research approach and describe how we have organized to address the cross-disciplinary nature of Big Data challenges. I will then describe the current state of BDAS with an emphasis on our newest efforts, including some or all of: the GraphX graph processing system, the Velox and MLBase machine learning platforms, and the SampleClean framework for hybrid human/computer data cleaning. Finally I will present our current views of how all the pieces will fit together to form a system that can adaptively bring the right resources to bear on a given data-driven question to meet time, cost and quality requirements throughout the analytics lifecycle.

Categories and Subject Descriptors
E.0 [General]

Keyword
Big Data

BIO
Michael Franklin is the Thomas M. Siebel Professor of Computer Science and Chair of the Computer Science Division at the University of California, Berkeley. He has over 30 years of experience in the database, data analytics, and data management fields as a researcher, lab director, faculty member, entrepreneur, and software developer. Prof. Franklin is also the Director of the Algorithms, Machines, and People Laboratory (AMPLab) at UC Berkeley. The AMPLab currently works with 23 industrial sponsors including founding sponsors Amazon Web Services, Google, and SAP, and received a National Science Foundation CISE “Expeditions in Computing” award, announced as part of the White House Big Data research initiative in March 2012. AMPLab is well-known for creating a number of popular systems in the Open Source Big Data ecosystem including Spark, Mesos, Shark, GraphX and MLlib, all parts of the Berkeley Data Analytics Stack (BDAS). Prof. Franklin is also a co-PI and Executive Committee member for the Berkeley Institute for Data Science, part of a multi-campus initiative to advance Data Science Environments. He is an ACM Fellow, a two-time winner of the ACM SIGMOD “Test of Time” award, and recipient of the outstanding Advisor Award from the Computer Science Graduate Student Association at Berkeley.
Academia, Startups and Mobile Advertising

Mark Corner
Fiksu Inc. and University of Massachusetts at Amherst
mcorner@fiksu.com, mcorner@cs.umass.edu

Abstract
Should more academics be making the leap to startups? I have been lucky enough to have that experience: going from a tenured professor position in mobile systems at UMass Amherst to a CTO job at a $100M+ startup in the mobile advertising space, Fiksu Inc. This talk examines some of the tradeoffs between academia and the startup world, the challenges in starting companies, and the calculus for joining them. Startups abound in the world of mobile advertising technology, which provides the economic foundation of the free-to-play internet and the continuing growth of e-commerce. We will delve into the murky world of advertising technology and discuss the unique opportunities that mobile presents.

ACM Classification
K.7 The Computing Profession

Keywords
Advertising Technology; Mobile Advertising; Startups; Academia.

Short Bio
Mark Corner is an expert in the areas of mobile and pervasive computing, mobile advertising technology, real-time bidding (RTB), large-scale systems, networking, file systems, and security. He is a tenured Associate Professor in the School of Computer Science at the University of Massachusetts Amherst and has been there since 2003. After tenure Mark put his research career on hold to be an early-stage founder and CTO at Fiksu Inc., a Boston-based mobile advertising technology company.

As a professor he has authored dozens of publications on mobile systems, holds two patents, and has been awarded millions of dollars in federal and industrial research grants. He was selected for DARPA’s prestigious Computer Science Study Panel in 2009.

He was the recipient of an NSF CAREER award in 2005, Best Paper awards at USENIX FAST 2007, ACM Multimedia 2005, and ACM Mobicom 2002. Prof. Corner also serves on the editorial board of IEEE Pervasive and IEEE Transactions on Mobile Computing. Mark holds a PhD in Electrical Engineering from the University of Michigan and BS and MS degrees from the University of Virginia. At UMass he teaches the Operating Systems course and a course on Usability.

The 50+ person team that Mark built enabled revenue growth from 0 to more than $100 million / year. He oversees the development of a massively scalable infrastructure on Ruby on Rails, Postgresql, and AWS that processes thousands of events per second from Real Time Bidding and in-app telemetry. The SDK that the team built touches more than half of all mobile devices, the RTB platform bids in excess of 500k times per second using proprietary optimization techniques, and Fiksu’s large-scale analytics system summarizes data from hundreds of billions of in-app events.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s). Copyright is held by the author/owner(s).

HotMobile’15, February 12–13, 2015, Santa Fe, New Mexico, USA.
ACM 978-1-4503-3391-7/15/02.
http://dx.doi.org/10.1145/2699343.2699368
User Engagement: The Network Effect Matters!

Ricardo Baeza-Yates, Mounia Lalmas
Yahoo! Labs Barcelona
Barcelona, Spain
{rbaeza,mounia}@acm.org

Abstract
In the online world, user engagement refers to the quality of the user experience that emphasizes the positive aspects of the interaction with a web application and, in particular, the phenomena associated with wanting to use that application longer and frequently. This definition is motivated by the observation that successful web applications are not just used, but they are engaged with. Users invest time, attention, and emotion into them.

Online providers aim not only to engage users with each service, but across all services in their network. They spend increasing effort to direct users to various services (e.g., using hyperlinks to help users navigate to and explore other services), to increase user traffic between their services. Nothing is known for users engaging across such a network of Web sites, something we call networked user engagement. We address this problem by combining techniques from web analytics and mining, information retrieval evaluation, and existing works on user engagement coming from the domains of information science, multimodal human computer interaction and cognitive psychology. In this way, we can combine insights from big data with deep analysis of human behavior in the lab or through crowd-sourcing experiments.

This work was mainly done in collaboration with Janette Lehmann and Georges Dupret from Yahoo! Labs, as well as Elad Yom-Tov from Microsoft Research.

Categories and Subject Descriptors
H.1.2 [User/Machine Systems]: Human information processing

General Terms
Human Factors, Measurement.

Keywords
User engagement, network of services, Web sites, metrics.

1. CONTENTS
User engagement is measured in many ways, through methods of self-reporting (e.g., questionnaires), observer methods (e.g., facial expression analysis, speech analysis, desktop actions, etc.), neuro-physiological signal processing methods (e.g., respiratory and cardiovascular accelerations and deaccelerations, muscle spasms, etc.), and from a web analytics perspective (through online behavior metrics that assess users' depth of engagement with a site). These methods represent various tradeoffs between scale of data and depth of understanding (for instance, surveys are small-scale but deep, whereas clicks are large-scale but shallow in understanding). However, little work has been done to integrate these various measures into a coherent understanding of engagement success.

The presentation comprises three parts:
1) First, we define user engagement, list its many characteristics as identified in the research and analytic literature, and discuss through real examples the challenges associated with measuring user engagement [1, 2].
2) Second, we describe recent data-driven approaches looking at user engagement through the development of new measures that allow for a better representation of how users engage with and across different web services, what we call networked user engagement [5, 6].
3) Finally, we describe how emerging research directions looking at affect and cognition as well as graph related measures are providing additional insights into measuring networked user engagement [3, 4].

2. CONCLUSIONS
This research line is a first step forward to fully understand user engagement in the context of the Web. This work also shows the need and importance of combining different points of view coming from different disciplines. This way of thinking is crucial to many areas, going beyond the Web and will in time lead to a new genre of computational social sciences that transcend specific applications on the Internet.
3. REFERENCES


Short Biography

Ricardo Baeza-Yates is VP of Yahoo! Research for Europe, Middle East and Latin America, leading the labs at Barcelona, Spain and Santiago, Chile, since 2006, as well as supervising the lab in Haifa, Israel, since 2008. He is also part time Professor at the Dept. of Information and Communication Technologies of the Universitat Pompeu Fabra in Barcelona, Spain, since 2005. Until 2005 he was Professor and Director of the Center for Web Research at the Department of Computer Science of the Engineering School of the University of Chile (currently in a leave of absence).

He obtained a Ph.D. from the University of Waterloo, Canada, in 1989. Before he obtained two masters (M.Sc. CS & M.Eng. EE) and the electrical engineering degree from the University of Chile, Santiago. He is co-author of the best-seller Modern Information Retrieval textbook, published in 1999 by Addison-Wesley with a second enlarged edition in 2011, as well as co-author of the 2nd edition of the Handbook of Algorithms and Data Structures, Addison-Wesley, 1991; and co-editor of Information Retrieval: Algorithms and Data Structures, Prentice-Hall, 1992, among more than 400 other publications.

He has received the Organization of American States award for young researchers in exact sciences (1993) and the CLEI Latin American distinction for contributions to computer science in the region (2009). In 2003 he was the first computer scientist to be elected to the Chilean Academy of Sciences. During 2007 he was awarded the Graham Medal for innovation in computing, given by the University of Waterloo to distinguished ex-alumni. In 2009 he was named ACM Fellow and in 2011 IEEE Fellow.
Future Direction of Digital Content

20th Anniversary Keynote Talk

Masahiro Fujita
Sony Corporation
Japan
masahirof@jp.sony.com

ABSTRACT
We review the history and trends of multimedia technologies, especially focusing on audio & video products. Our lifestyles have been changing according to the developments of these technologies. We forecast the direction of multimedia technologies and our lifestyles in future.

We can guess that the future direction of audio & video is going to higher and higher fidelity. Display devices will be larger and flexible. In addition, wearable devices will be popular as consumer products. It will be achieved by printing manufacture technologies. These devices and high fidelity content will realize immersive super reality.

While we will enjoy the traditional audio & video content with the immersive super reality, we will use more social media by which realtime & global human-human interaction will be realized. We will be able to utilize other people’s knowledge and decisions in realtime using the interactive & immersive super reality.

In addition to the social networks with humans, sensor & actuator networks will also be globally established. This will extend our perception and decision abilities. We will find new problems with the distributed sensor & actuator networks, which we did not find with our ordinary perceptions. Using interactive & immersive super reality, we will collaboratively solve personal daily problems as well as global problems.

Thus, in future we will realize collaborative community with interactive & immersive super reality. In the community we will enjoy entertainment content and will solve many problems collaboratively. This community itself will be the content in the future.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous;
H.5.1 Information Systems Applications [Multimedia Information Systems]: Video

General Terms
Human Factors

Keywords
Digital content, Display devices, Printing manufacture, Interactive & Immersive Super reality, Social media, Collaborative community

Bio
Masahiro Fujita is currently Vice President of System and Software Technology Platform at Sony Corporation, Tokyo, Japan. He received a B.A. in Electronics and Communications from Waseda University, Tokyo, in 1981, and an M.S. in Electrical Engineering from University of California Irvine, in 1989.

He joined Sony Corporation in 1981, and worked for development of a spread spectrum communication system, which was used in a receiver of global positioning system for a car navigation, and for VLBI (Very Long Baseline Interferometry) for a earth quake forecast system. From 1988, he became a graduate student of University of California, Irvine, and studied artificial neural networks for visual perception. After he returned to Sony, he started the Robot Entertainment project from 1993, and developed the entertainment robot AIBO, which started to sell in 1999. After the AIBO project, he has been in charge of development for cognitive part of a small humanoid robot QRIO. In 1998 he proposed to establish RoboCup four-legged robot league using AIBO as platform, which was one of 4 official physical robot leagues in RoboCup until 2007.

In 2008 he became a president of System Technologies Laboratories, Sony Corporation, where he led incorporating intelligent functions for Sony products and services, such as personalized recommendation, augmented reality (SmartAR), and so on.
ABSTRACT
A general theory of trust in networks of humans and computers must be built on both a theory of behavioral trust and a theory of computational trust. This argument is motivated by increased participation of people in online social networking, crowdsourcing, human computation, and socio-economic protocols; e.g., protocols modeled by trust and gift-exchange games, norms-establishing contracts, and scams/deception. We illustrate a class of interactive social protocols that relies both on trustworthy properties of commodity systems (e.g., verifiable end-to-end trusted path) and participant trust, since on-line verification of protocol compliance is often impractical; e.g., it can lead to undecidable problems, co-NP complete test procedures, and user inconvenience. Trust is captured by participant preferences (i.e., risk and betrayal aversion) and beliefs in the trustworthiness of other protocol participants. Both preferences and beliefs can be enhanced whenever protocol non-compliance leads to punishment of untrustworthy participants; i.e., it seems natural that betrayal aversion can be decreased and belief in trustworthiness increased by properly defined punishment. Similarly, risk aversion can be decreased and trustworthiness increased by feasible recovery from participant non-compliance.

A general theory of trust which focuses on the establishment of new trust relations where none were possible before would help create new economic opportunities. New trust relations would increase the pool of services available to users, remove cooperation barriers, and enable the “network effect” where it really matters; i.e., at the application level. Hence, it seems important that security research should enable and promote trust-enhancement infrastructures in human and computer networks; e.g., trust networks. Finally, we argue that a general theory of trust should mirror human expectations and mental models without relying on false metaphors and analogies with the physical world.

ACM Classification:
1.0 Computing Methodologies: GENERAL

Keywords:
Trust, trustworthy systems, social protocols