Data Science in L.A.'s "Silicon Beach": Shadow Vocation or Emergent Profession?

2015 CSU-POM Conference
CSU Northridge
Saturday, February 28, 2015

Wayne Smith, Ph.D.
Department of Management
CSU Northridge
`ws@csun.edu`

This presentation is available at:

`smithw.org/csupom.pdf`   or   `github.com/wsphd/csupom/`

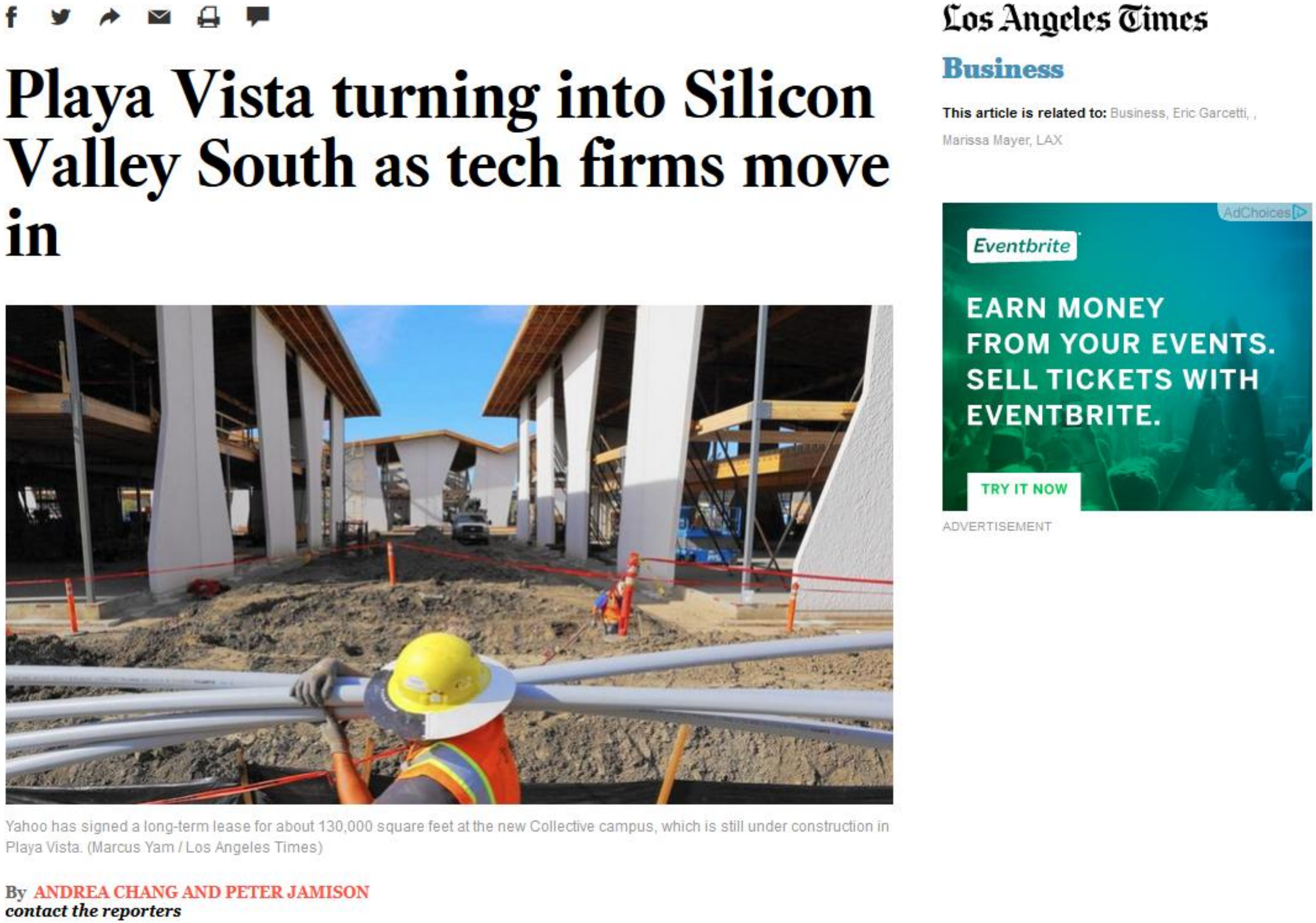
Motivation

- Background/Preliminaries

- I apologize…
  - “cloud computing”, “big data”, “[xyz] analytics”, and worst of all, “data science”
  - These are *industry* buzz “terms”, not *our* terms.
  - I prefer *decision-support systems*, *predictive modeling*, *statistical computing,* and *mathematical optimization*, but most of the folks I routinely work with prefer the term “data science” (We might agree on the terms *reproducible*, *analytics* and *visualization*).

- I am on the steering committee for the LA R/Data science Meetup group
  - There are a few other Ph.D.s heavily involved too—Szilard Pafka (primary lead), David McArthur, Rob Gould, Jeremy Miles, Tim Triche, Eduardo Arino de la Rubia, and others

- “*Data* as the *New Oil*” Opportunities and Challenges
  - Is there a perturbation (disequilibrium) in this labor market?
  - Exactly what is the production, operations, transportation, logistics of *bits* viz. *atoms?*
  - What is our institutional role (if any) in this (increasingly, non-college) community?
  - Are we witnessing the emergence of a new *data guild* in the modern service economy?



Los Angeles Times

Business

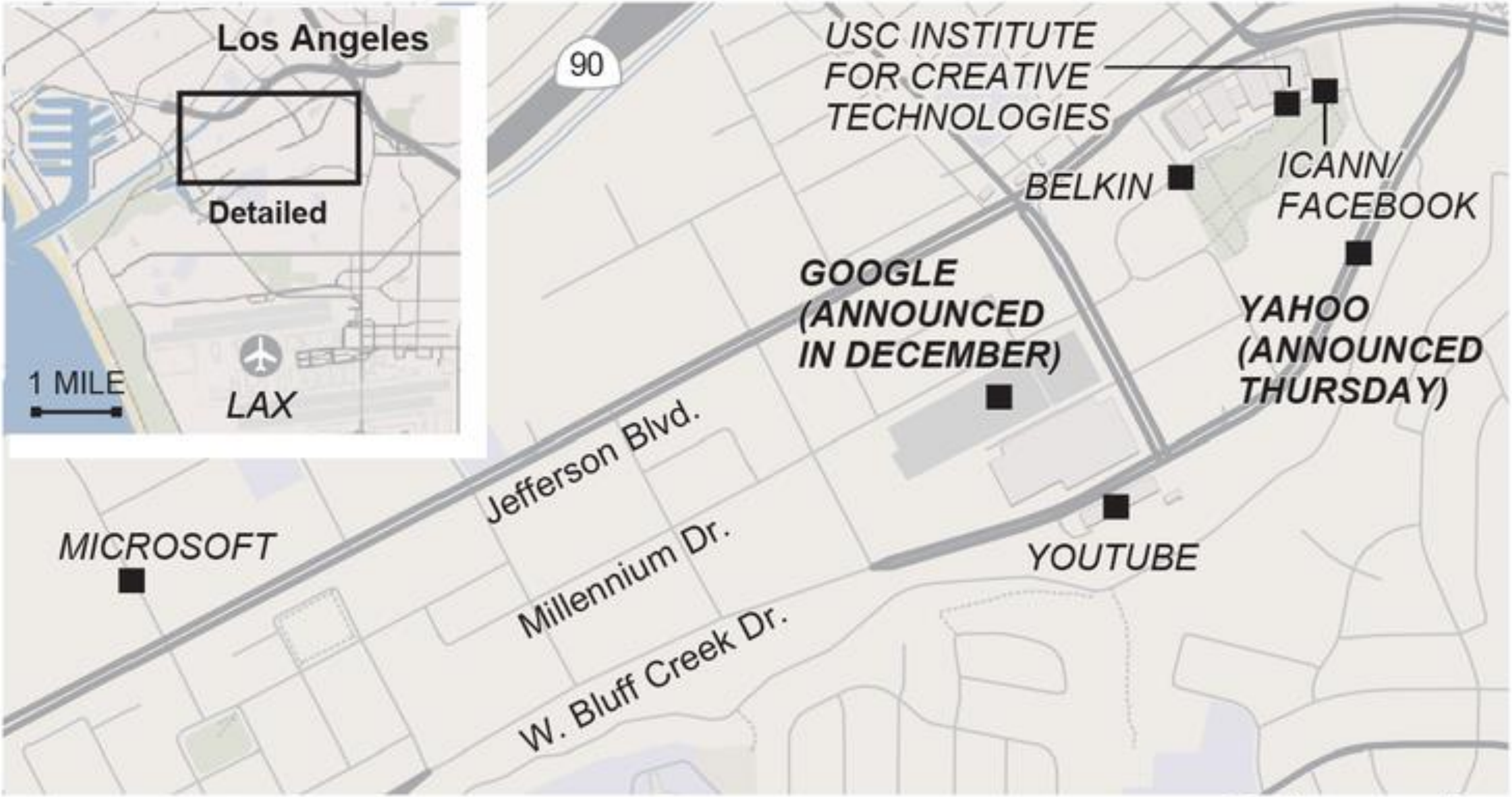
This article is related to: Business, Eric Garcetti, , Marissa Mayer, LAX

Playa Vista turning into Silicon Valley South as tech firms move in

Yahoo has signed a long-term lease for about 130,000 square feet at the new Collective campus, which is still under construction in Playa Vista. (Marcus Yam / Los Angeles Times)

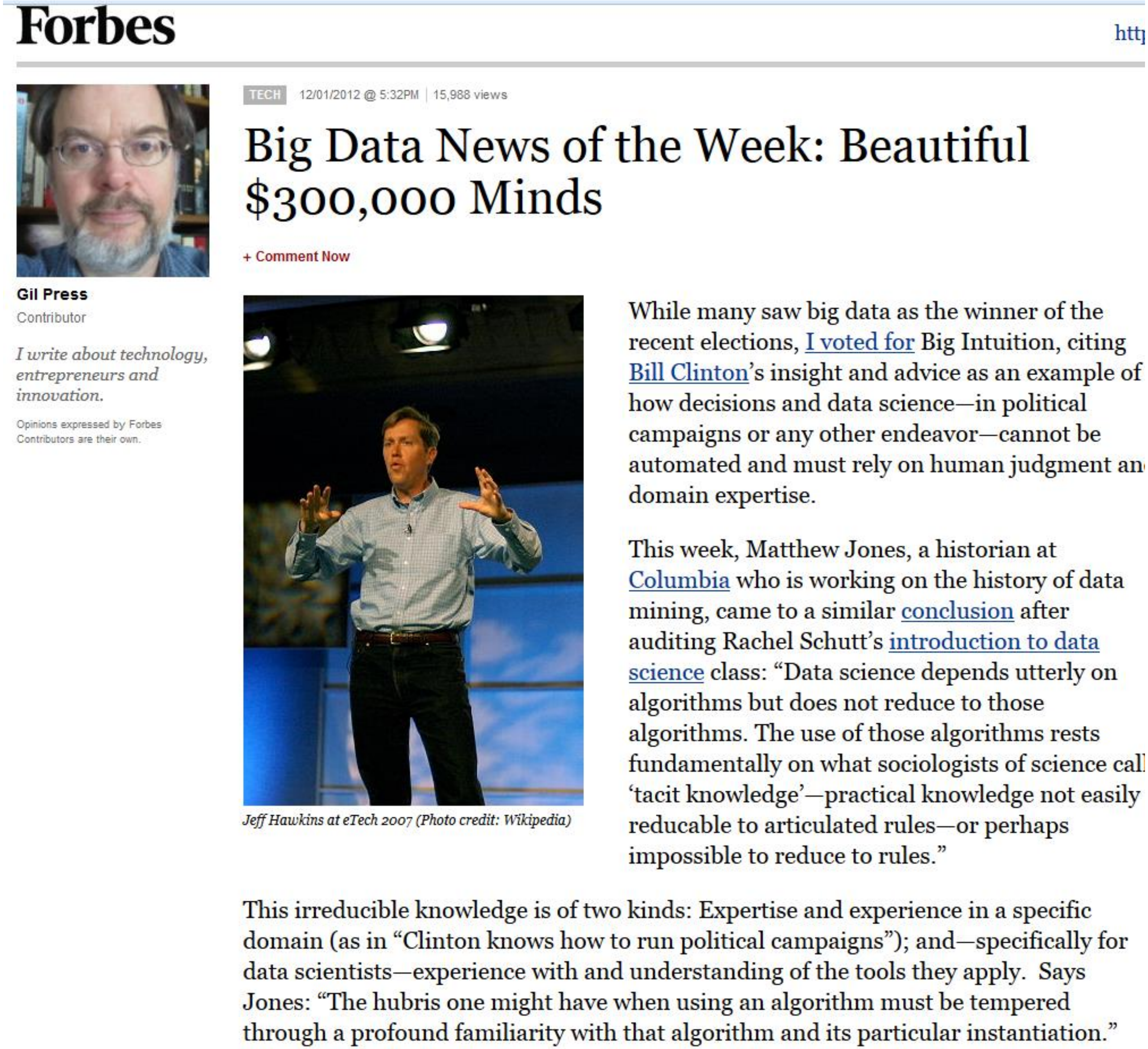
By ANDREA CHANG AND PETER JAMISON
contact the reporters

http://www.latimes.com/business/la-fi-silicon-valley-south-20150118-story.html#page=1



Sources: Mapbox, OpenStreetMap

@latimesgraphics



Forbes

htt

Gil Press

Contributor

I write about technology, entrepreneurs and innovation.

Opinions expressed by Forbes Contributors are their own.

TECH 12/01/2012 @ 5:32PM | 15,988 views

Big Data News of the Week: Beautiful $300,000 Minds

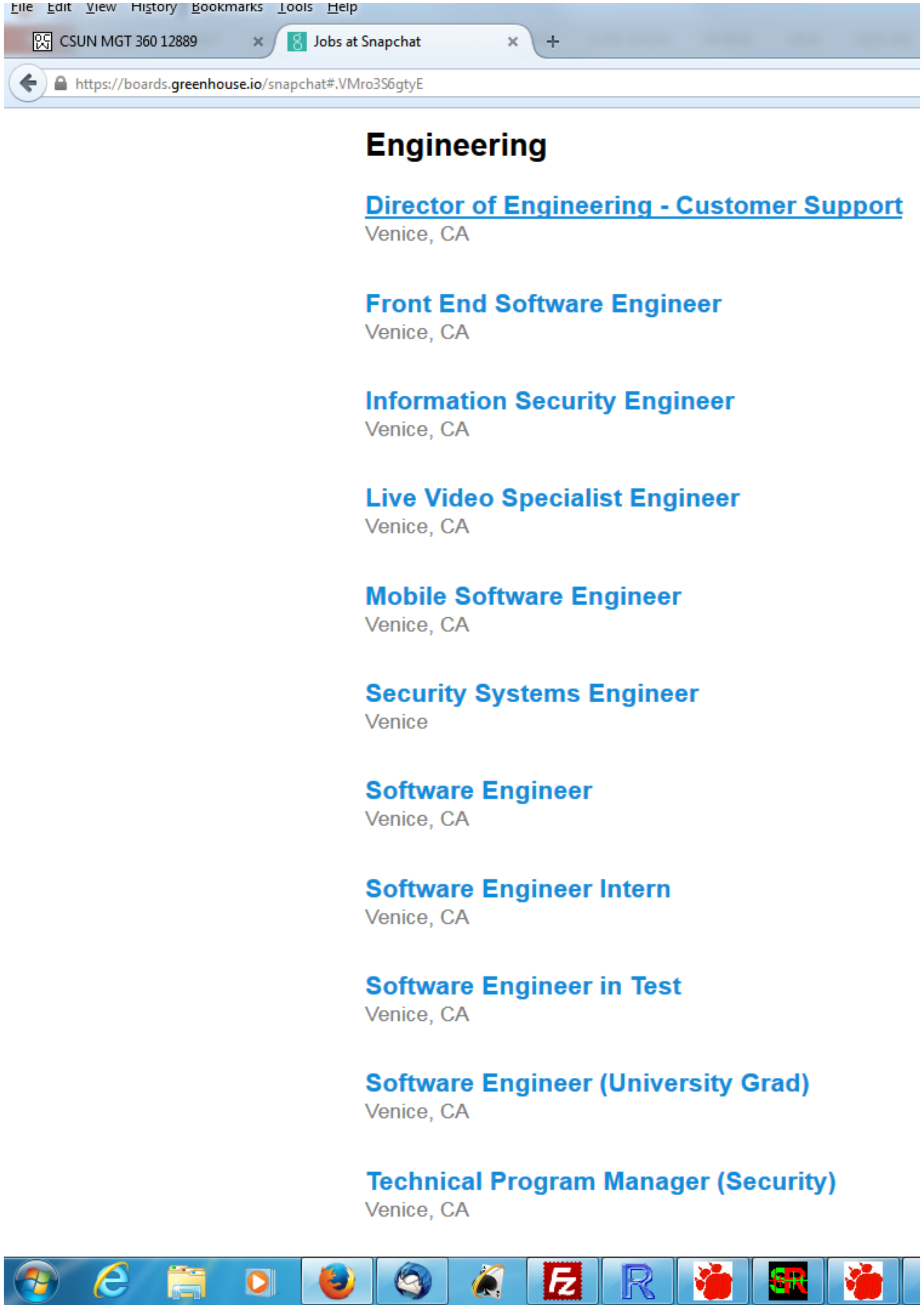
+ Comment Now

Jeff Hawkins at eTech 2007 (Photo credit: Wikipedia)

While many saw big data as the winner of the recent elections, I voted for Big Intuition, citing Bill Clinton's insight and advice as an example of how decisions and data science—in political campaigns or any other endeavor—cannot be automated and must rely on human judgment an domain expertise.

This week, Matthew Jones, a historian at Columbia who is working on the history of data mining, came to a similar conclusion after auditing Rachel Schutt's introduction to data science class: "Data science depends utterly on algorithms but does not reduce to those algorithms. The use of those algorithms rests fundamentally on what sociologists of science call 'tacit knowledge'—practical knowledge not easily reducable to articulated rules—or perhaps impossible to reduce to rules."

This irreducible knowledge is of two kinds: Expertise and experience in a specific domain (as in "Clinton knows how to run political campaigns"); and—specifically for data scientists—experience with and understanding of the tools they apply. Says Jones: "The hubris one might have when using an algorithm must be tempered through a profound familiarity with that algorithm and its particular instantiation."


File Edit View History Bookmarks Tools Help
CSUN MGT 360 12889
Jobs at Snapchat
https://boards.greenhouse.io/snapchat#.VMro3S6gtyE

Engineering

Director of Engineering - Customer Support
Venice, CA

Front End Software Engineer
Venice, CA

Information Security Engineer
Venice, CA

Live Video Specialist Engineer
Venice, CA

Mobile Software Engineer
Venice, CA

Security Systems Engineer
Venice

Software Engineer
Venice, CA

Software Engineer Intern
Venice, CA

Software Engineer in Test
Venice, CA

Software Engineer (University Grad)
Venice, CA

Technical Program Manager (Security)
Venice, CA

R and/or other tools

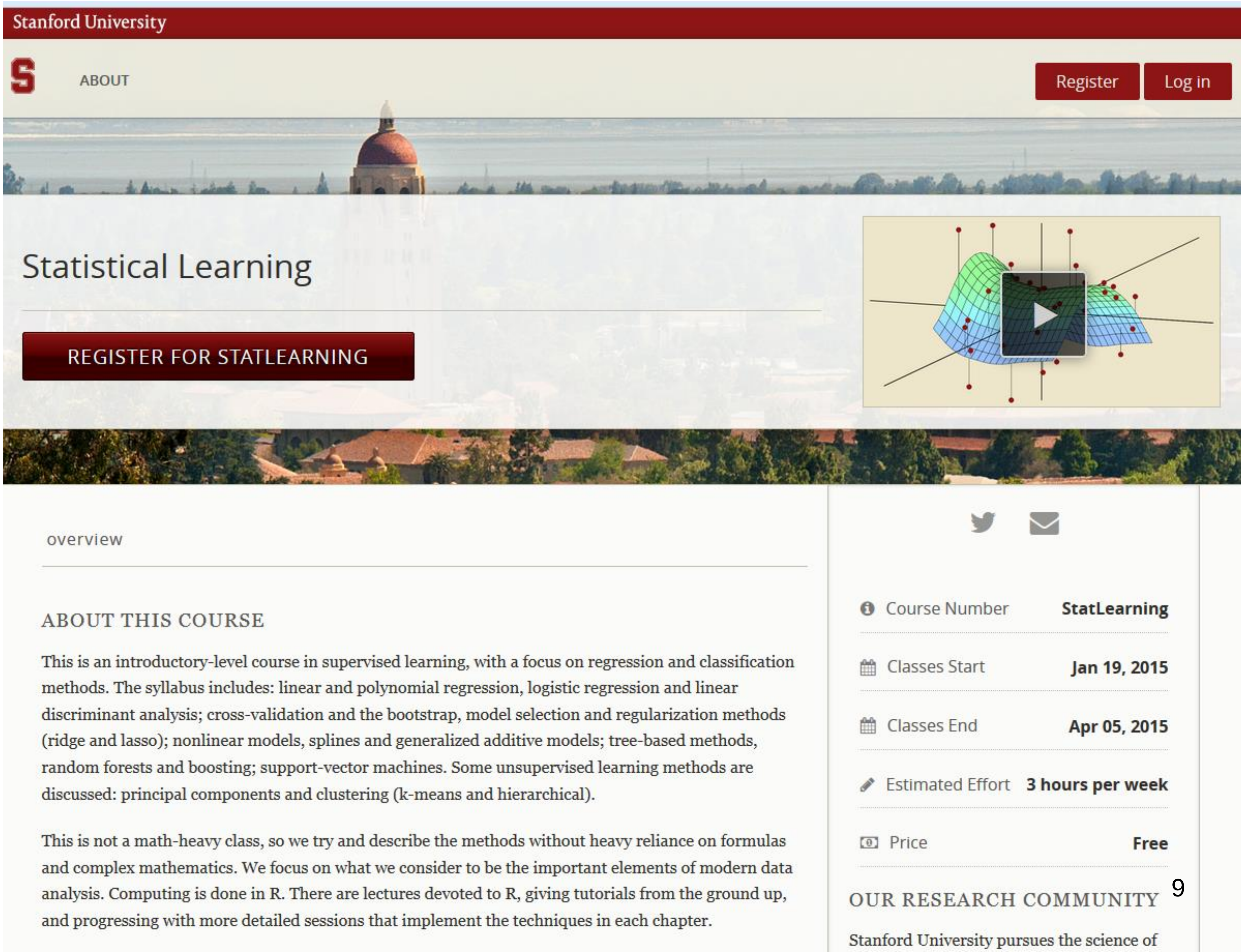
- COTS (Non-FOSS)
  - SAS, IBM SPSS, Stata, S-PLUS, BMDP, Lisrel, EQS, HLM, MPLUS, SmartPLS, …
  - Mathematica, @Risk/Excel add-ins, RATS, Matlab, Simscript, Arena, etc.
  - IBM Netezza, MS Anal. Serv., Oracle Adv. Anal., SAP HANA, Teradata Miner, etc.
  - UCINET/NVivo/Tropes, ConQuest, ArcGIS, AutoDesk, etc.
  - IBM CPLEX, MOSEK, GUROBI, Frontline, etc.
- FOSS/GNU
  - Domain-specific?
    - GRETL/xBUGS, Ggobi, Pajek, lisp-stat, Netlib/BLAS/LAPACK/FFTPACK, etc.
    - Sage, Octave, Scilab, Java/Colt/JFreeCharts, incanter, Protovis, D3, etc.
    - GNU LP, COIN-OR, NASTRAN/BLAS, NETLIB, WordNet, statconn, etc.
  - General-purpose?
    - jQuery/jqueryUI/flot/jstat, BIRT, RapidMiner, Pentaho, MongoDB, etc.
    - Apache Hadoop/Spark/Solr/CouchDB/Lucene/OpenNLP/Mahout/Velocity, etc.
    - Python — iPython/NumPy/matplotlib/statmodels/Blaze/Bokeh/scikit-learn, etc.
    - Julia — iJulia/Juno/MultivariateStats/MLBase/TimeSeries/Gadfly/JuMP, etc.

Coordination

- *Monthly Meetups* (physical, not virtual)
  - E.g., `www.datascience.la`
  - Monthly meetings (50-200 people)

  - By language/platform/environment: R, Python
  - By technology/workflow: Visualization, DW/BI, Machine Learning
  - By sector/domain: AdTech, FinTech
  - By management "level": a few directors- and mangers-only events (by invitation only)

- *Annual Conferences* (physical, not virtual)
  - E.g., "Big Data camp" at Direct TV (El Segundo)
  - `http://insidebigdata.com/2014/06/03/free-conference-big-data-camp-2014/`
  - Multiple, parallel tracks
  - Vendor support (all expenses covered fully)
  - Last three years' attendance—200 (2013), 500 (2014), 800 expected (2015)



Statistical Learning

ABOUT THIS COURSE

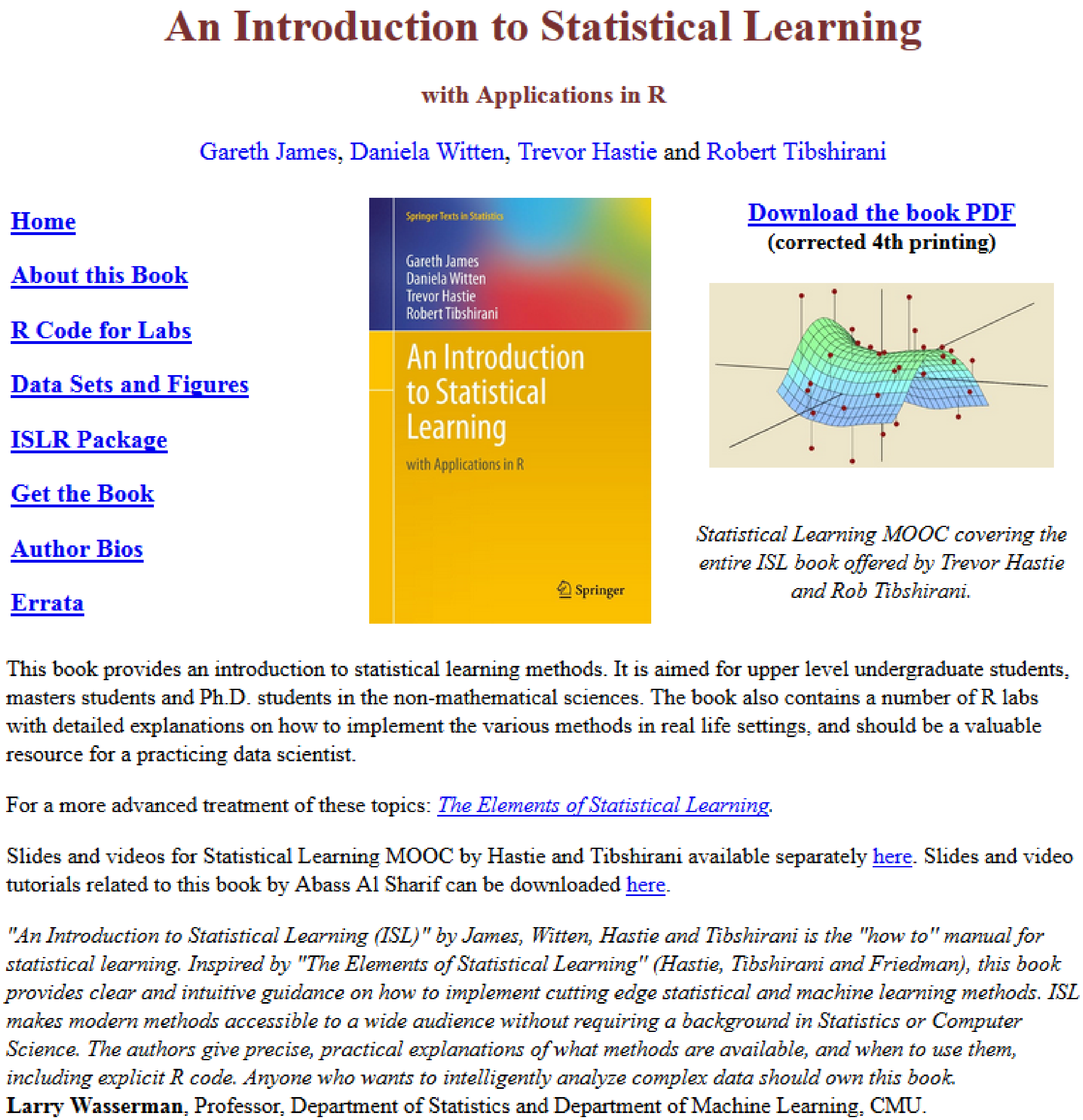
This is an introductory-level course in supervised learning, with a focus on regression and classification methods. The syllabus includes: linear and polynomial regression, logistic regression and linear discriminant analysis; cross-validation and the bootstrap, model selection and regularization methods (ridge and lasso); nonlinear models, splines and generalized additive models; tree-based methods, random forests and boosting; support-vector machines. Some unsupervised learning methods are discussed: principal components and clustering (k-means and hierarchical).

This is not a math-heavy class, so we try and describe the methods without heavy reliance on formulas and complex mathematics. We focus on what we consider to be the important elements of modern data analysis. Computing is done in R. There are lectures devoted to R, giving tutorials from the ground up, and progressing with more detailed sessions that implement the techniques in each chapter.

Course Number	StatLearning
Classes Start	Jan 19, 2015
Classes End	Apr 05, 2015
Estimated Effort	3 hours per week
Price	Free

OUR RESEARCH COMMUNITY

Stanford University pursues the science of



An Introduction to Statistical Learning

with Applications in R

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

- Home
- About this Book
- R Code for Labs
- Data Sets and Figures
- ISLR Package
- Get the Book
- Author Bios
- Errata

Download the book PDF
(corrected 4th printing)

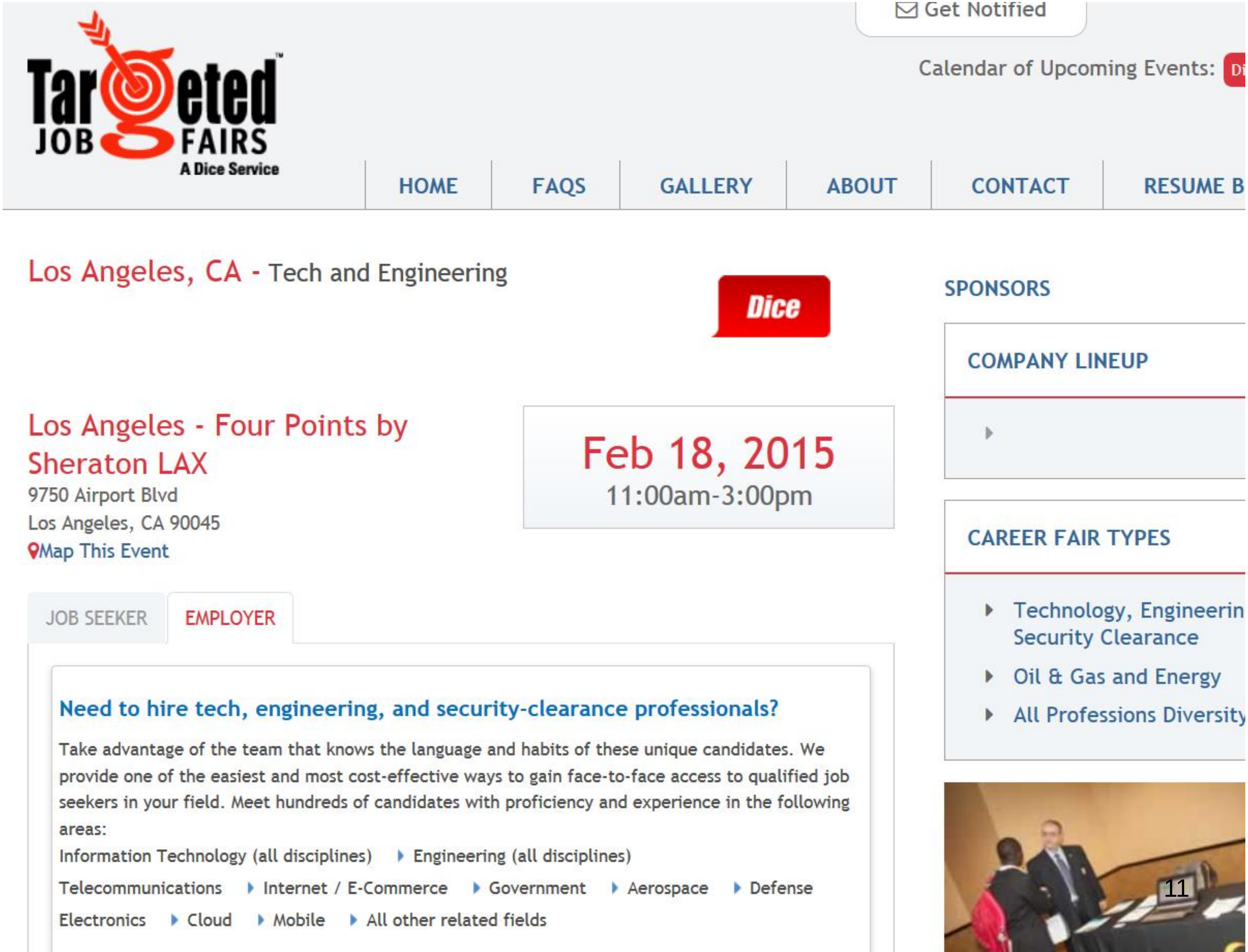
Statistical Learning MOOC covering the entire ISL book offered by Trevor Hastie and Rob Tibshirani.

This book provides an introduction to statistical learning methods. It is aimed for upper level undergraduate students, masters students and Ph.D. students in the non-mathematical sciences. The book also contains a number of R labs with detailed explanations on how to implement the various methods in real life settings, and should be a valuable resource for a practicing data scientist.

For a more advanced treatment of these topics: *The Elements of Statistical Learning*.

Slides and videos for Statistical Learning MOOC by Hastie and Tibshirani available separately here. Slides and video tutorials related to this book by Abass Al Sharif can be downloaded here.

"An Introduction to Statistical Learning (ISL)" by James, Witten, Hastie and Tibshirani is the "how to" manual for statistical learning. Inspired by "The Elements of Statistical Learning" (Hastie, Tibshirani and Friedman), this book provides clear and intuitive guidance on how to implement cutting edge statistical and machine learning methods. ISL makes modern methods accessible to a wide audience without requiring a background in Statistics or Computer Science. The authors give precise, practical explanations of what methods are available, and when to use them, including explicit R code. Anyone who wants to intelligently analyze complex data should own this book.
Larry Wasserman, Professor, Department of Statistics and Department of Machine Learning, CMU.



Los Angeles, CA - Tech and Engineering

Los Angeles - Four Points by Sheraton LAX
9750 Airport Blvd
Los Angeles, CA 90045
Map This Event

Feb 18, 2015
11:00am-3:00pm

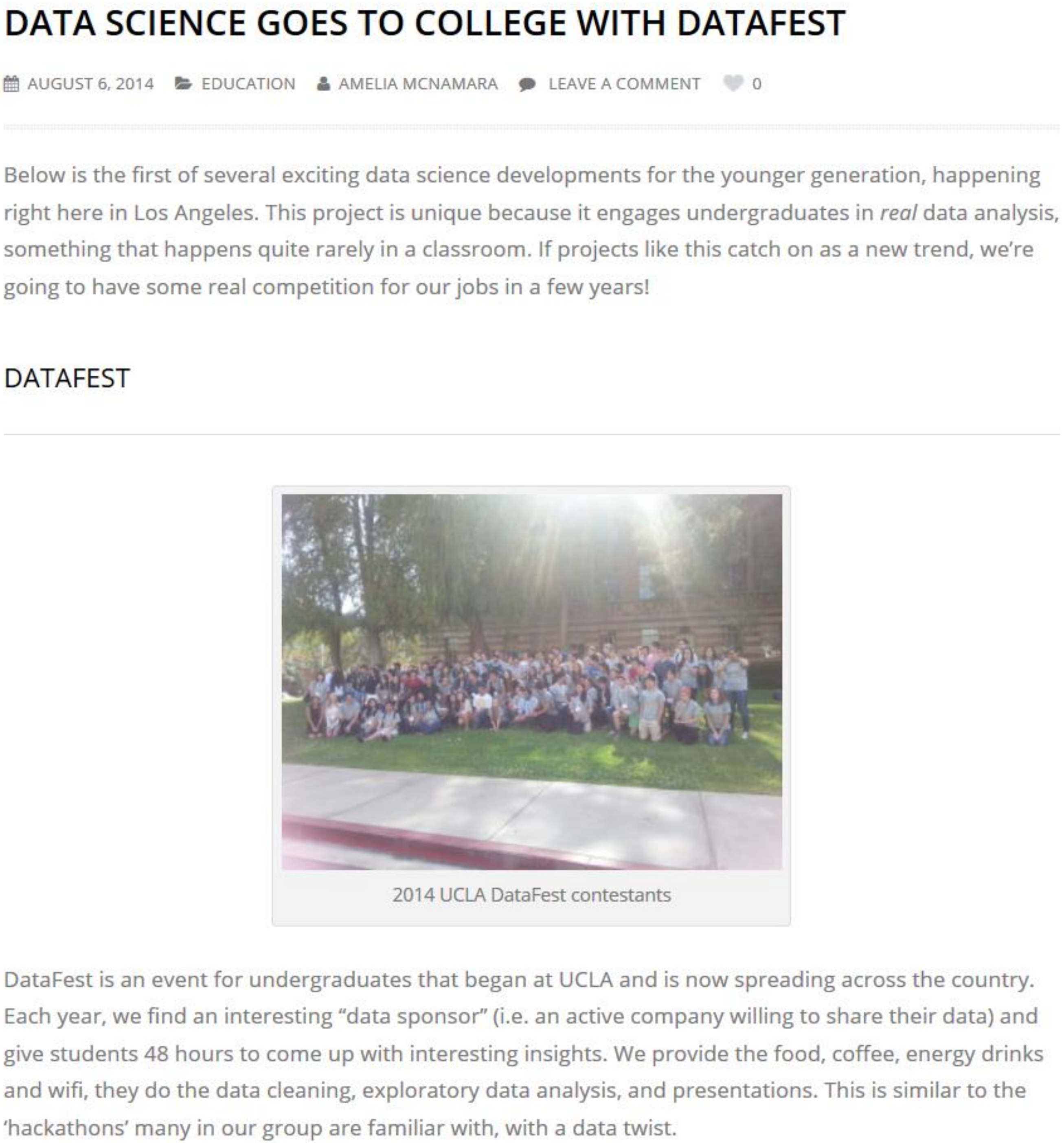
JOB SEEKER | EMPLOYER

Need to hire tech, engineering, and security-clearance professionals?

Take advantage of the team that knows the language and habits of these unique candidates. We provide one of the easiest and most cost-effective ways to gain face-to-face access to qualified job seekers in your field. Meet hundreds of candidates with proficiency and experience in the following areas:

Information Technology (all disciplines) ▸ Engineering (all disciplines)
Telecommunications ▸ Internet / E-Commerce ▸ Government ▸ Aerospace ▸ Defense
Electronics ▸ Cloud ▸ Mobile ▸ All other related fields

- Technology, Engineerin Security Clearance
- Oil & Gas and Energy
- All Professions Diversity



DATA SCIENCE GOES TO COLLEGE WITH DATAFEST

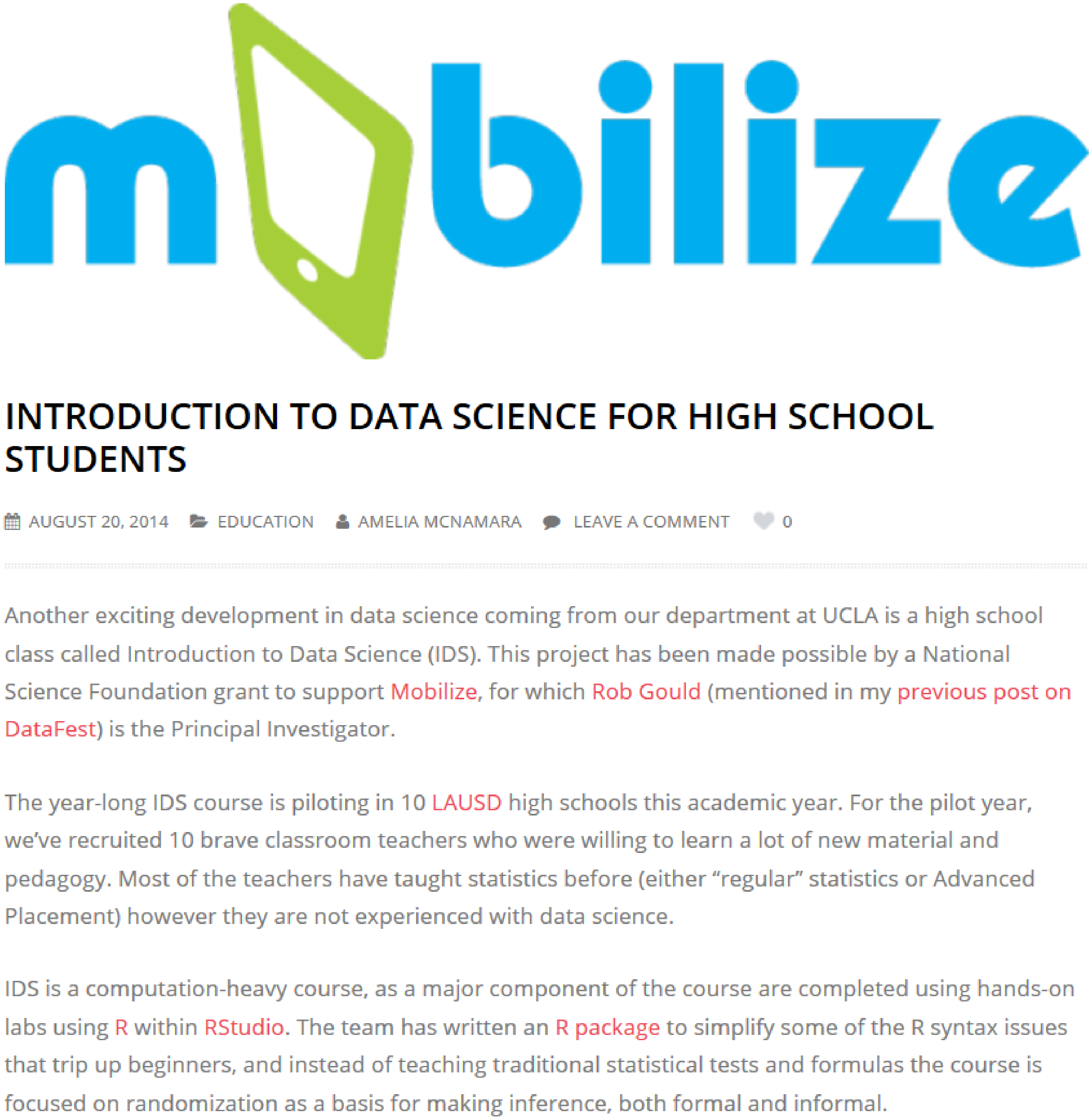
AUGUST 6, 2014 EDUCATION AMELIA MCNAMARA LEAVE A COMMENT ♥ 0

Below is the first of several exciting data science developments for the younger generation, happening right here in Los Angeles. This project is unique because it engages undergraduates in *real* data analysis, something that happens quite rarely in a classroom. If projects like this catch on as a new trend, we're going to have some real competition for our jobs in a few years!

DATAFEST

2014 UCLA DataFest contestants

DataFest is an event for undergraduates that began at UCLA and is now spreading across the country. Each year, we find an interesting "data sponsor" (i.e. an active company willing to share their data) and give students 48 hours to come up with interesting insights. We provide the food, coffee, energy drinks and wifi, they do the data cleaning, exploratory data analysis, and presentations. This is similar to the 'hackathons' many in our group are familiar with, with a data twist.



INTRODUCTION TO DATA SCIENCE FOR HIGH SCHOOL STUDENTS

AUGUST 20, 2014 · EDUCATION · AMELIA MCNAMARA · LEAVE A COMMENT · ♡ 0

Another exciting development in data science coming from our department at UCLA is a high school class called Introduction to Data Science (IDS). This project has been made possible by a National Science Foundation grant to support Mobilize, for which Rob Gould (mentioned in my previous post on DataFest) is the Principal Investigator.

The year-long IDS course is piloting in 10 LAUSD high schools this academic year. For the pilot year, we've recruited 10 brave classroom teachers who were willing to learn a lot of new material and pedagogy. Most of the teachers have taught statistics before (either "regular" statistics or Advanced Placement) however they are not experienced with data science.

IDS is a computation-heavy course, as a major component of the course are completed using hands-on labs using R within RStudio. The team has written an R package to simplify some of the R syntax issues that trip up beginners, and instead of teaching traditional statistical tests and formulas the course is focused on randomization as a basis for making inference, both formal and informal.

Relative importance of “data science”

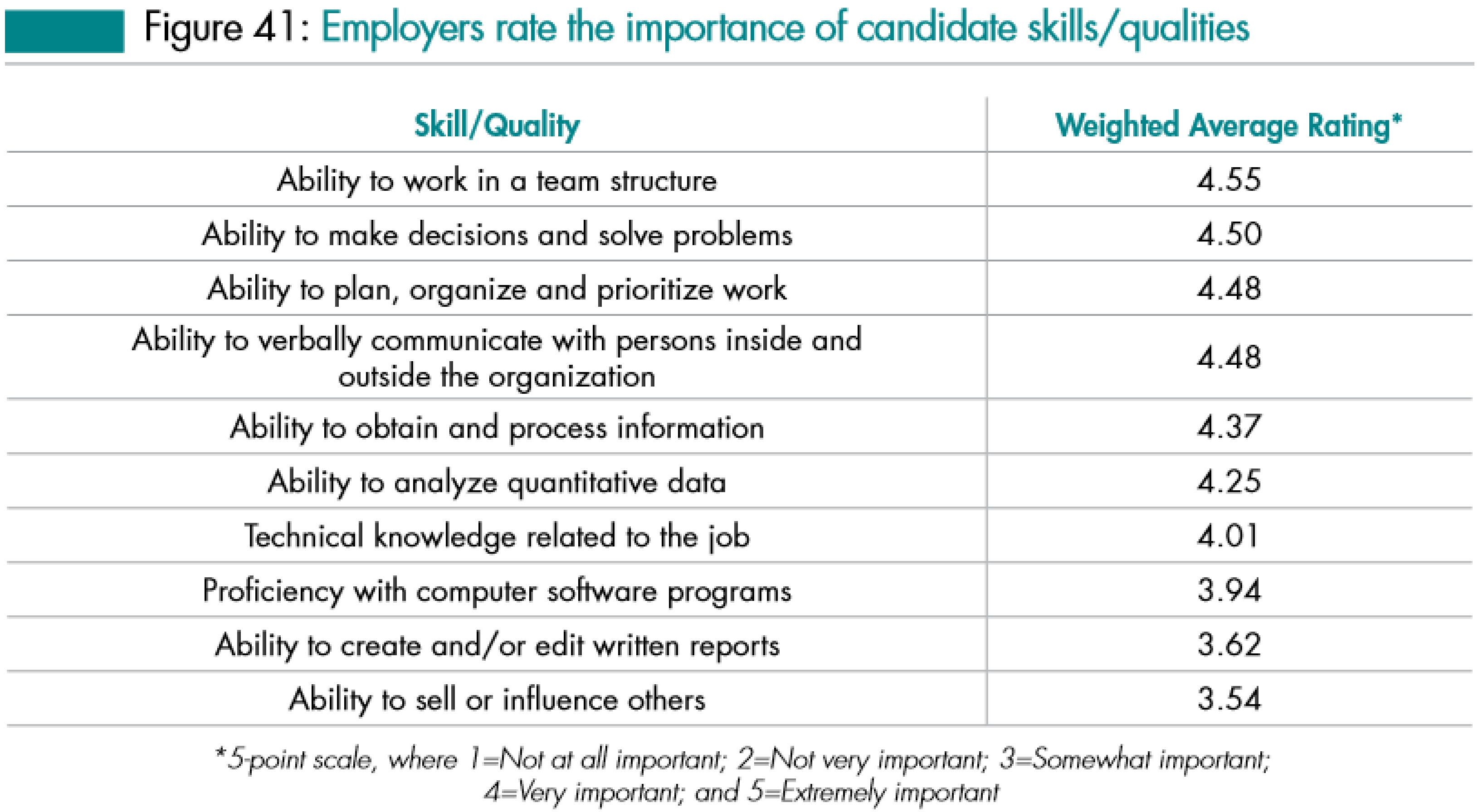
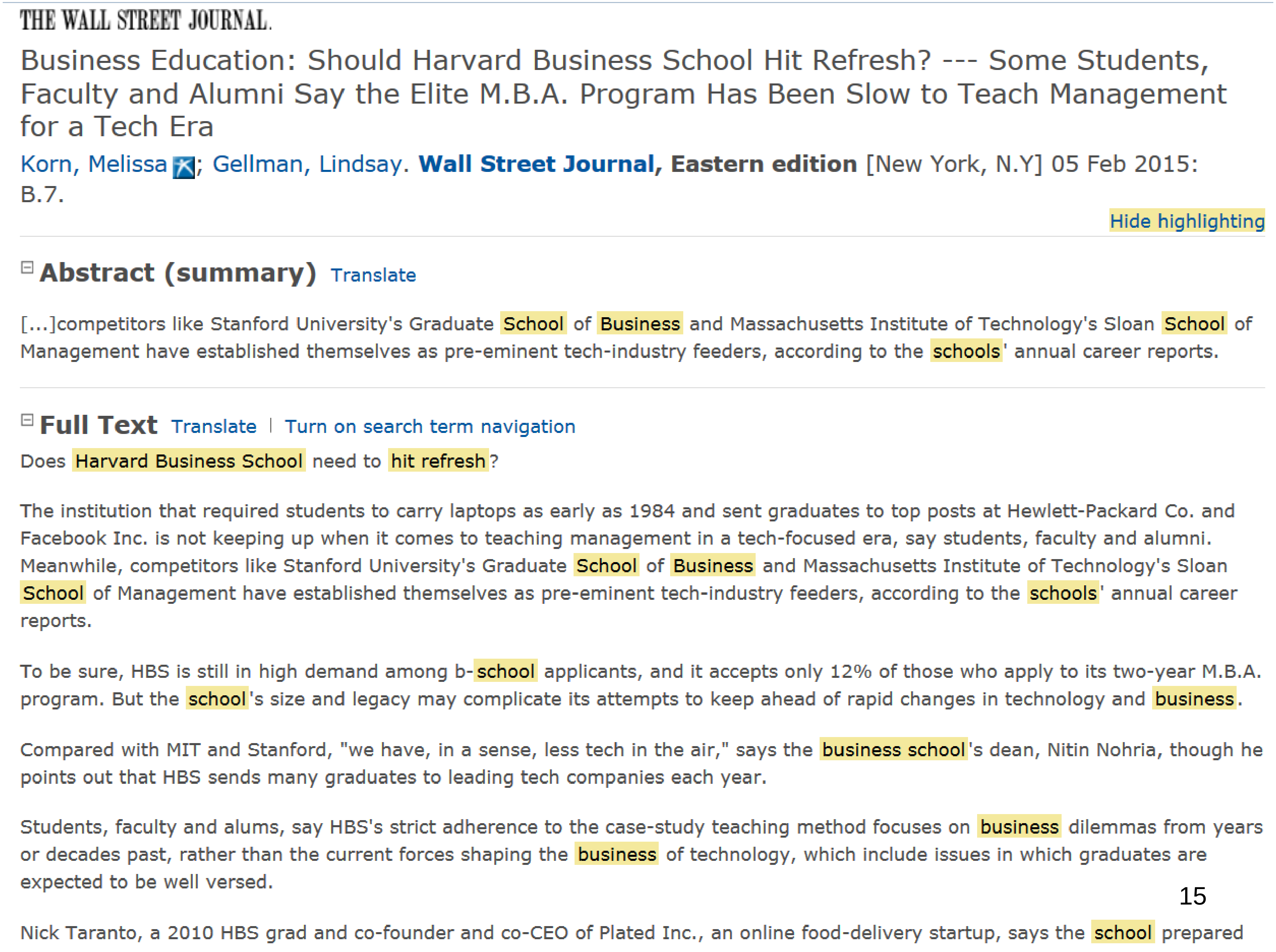


Figure 41: Employers rate the importance of candidate skills/qualities

Skill/Quality	Weighted Average Rating*
Ability to work in a team structure	4.55
Ability to make decisions and solve problems	4.50
Ability to plan, organize and prioritize work	4.48
Ability to verbally communicate with persons inside and outside the organization	4.48
Ability to obtain and process information	4.37
Ability to analyze quantitative data	4.25
Technical knowledge related to the job	4.01
Proficiency with computer software programs	3.94
Ability to create and/or edit written reports	3.62
Ability to sell or influence others	3.54

*5-point scale, where 1=Not at all important; 2=Not very important; 3=Somewhat important; 4=Very important; and 5=Extremely important

Source: Job Outlook 2014, National Association of College and Employers



THE WALL STREET JOURNAL.

Business Education: Should Harvard Business School Hit Refresh? --- Some Students, Faculty and Alumni Say the Elite M.B.A. Program Has Been Slow to Teach Management for a Tech Era

Korn, Melissa; Gellman, Lindsay. **Wall Street Journal, Eastern edition** [New York, N.Y] 05 Feb 2015: B.7.

⊟ Abstract (summary) Translate

[...]competitors like Stanford University's Graduate School of Business and Massachusetts Institute of Technology's Sloan School of Management have established themselves as pre-eminent tech-industry feeders, according to the schools' annual career reports.

⊟ Full Text Translate | Turn on search term navigation

Does Harvard Business School need to hit refresh?

The institution that required students to carry laptops as early as 1984 and sent graduates to top posts at Hewlett-Packard Co. and Facebook Inc. is not keeping up when it comes to teaching management in a tech-focused era, say students, faculty and alumni. Meanwhile, competitors like Stanford University's Graduate School of Business and Massachusetts Institute of Technology's Sloan School of Management have established themselves as pre-eminent tech-industry feeders, according to the schools' annual career reports.

To be sure, HBS is still in high demand among b-school applicants, and it accepts only 12% of those who apply to its two-year M.B.A. program. But the school's size and legacy may complicate its attempts to keep ahead of rapid changes in technology and business.

Compared with MIT and Stanford, "we have, in a sense, less tech in the air," says the business school's dean, Nitin Nohria, though he points out that HBS sends many graduates to leading tech companies each year.

Students, faculty and alums, say HBS's strict adherence to the case-study teaching method focuses on business dilemmas from years or decades past, rather than the current forces shaping the business of technology, which include issues in which graduates are expected to be well versed.

Nick Taranto, a 2010 HBS grad and co-founder and co-CEO of Plated Inc., an online food-delivery startup, says the school prepared

4-year Schools in this *New* "Service" Economy in CA?

- USC, M.S. (part of the C.S. Dept.)
  - http://www.cs.usc.edu/academics/masters/msdata.htm

- UCI Extension, certificate
  - http://unex.uci.edu/areas/it/data_science/

- UC Berkeley, M.S. (online) (part of the School of Information)
  - http://datascience.berkeley.edu/

- CSU Fullerton, certificate, online, accelerated
  - http://extension.fullerton.edu/professionaldevelopment/certificates/data-science

- Near Fails:
  - *UCOnline* – Outside investors want the students, faculty, and curriculum, but not the *bureaucracy.*
  - *CSUOnline* – Severe misunderstanding of the elasticity of demand for this service catchment group.

- Is CSU (mainstream) appropriate? Cross-disciplinary? Other approaches/ideas?

CCC’s in this *New* “Service” Economy?

- 15 CA Community Colleges will offer Bachelor’s Degrees beginning in 2017.

- **Airframe manufacturing technology**, Antelope Valley College
- **Industrial automation**, Bakersfield College
- **Emergency services and allied health systems**, Crafton Hills College
- **Mortuary science**, Cypress College
- **Equine industry**, Feather River College
- **Dental hygiene**, Foothill College and West Los Angeles College
- **Bio-manufacturing**, Mira Costa College
- **Respiratory care**, Modesto Junior College and Skyline College
- **Automotive technology**, Rio Hondo College
- **Health information management**, Mesa College
- **Occupational studies**, Santa Ana College
- **Interaction design**, Santa Monica College
- **Health information management**, Shasta College

Candidate Explanations (1/3)
(Technological)

- Neo-Classical Economics
  - Supply and Demand
  - "There is a *shortage* of individuals with "data scientist" skills; I can *exploit* that market disjunction*."*
- Empirical Finance
  - Loan-to-Value
  - "I might not achieve my *potential*, but at least I'm not in *debt*."

- Statistics (stochastic inference)
  - Privilege *specification* (narrow, overfitted results) over *generalization* (broader theory development)
  - "If I have more *data*, then I believe that I will have proportionally more *information*."
- Computer Science (machine learning)
  - Demonstrated capability with the *current* ("right now") platforms and languages is the valued skill.
  - "I need *computationally-intensive* and *unsupervised* (no response var.) approaches for "big data".

- Information Systems
  - Data management; data "wrangling"; Complex SQL Joins; rectangular transformations; file I/O
  - "The difficult *data queries* all come through me anyway; I'll might as well just do the *analysis* too."

Candidate Explanations (2/3)
(Sociological)

- Psychology
  - Ego-centrism, absolutism, and perhaps, nihilism
  - "You don't know me! Universities mostly just teach theories and I mostly just write papers."
- Sociology
  - Exponential Random Graph Models (agent- and ego-networks)
  - "I just need to immerse myself in the right industry, organizational, and community eco-system."

- Cognitive Science
  - *Visualization* is the new *Descriptive Statistics*
  - "Don't use messy *assumptions*, *models*, *and inference*; just display *high-dimensional graphs*."
- Political Science
  - *Rational Choice* theory (perceived minimal efficacy in structural, government change); *Elite* theory
  - "If I don't take this *job*, someone with an H1-B Visa might; besides I want power through my *skills*."

- Higher Education
  - *Just-in-time* training (short-term, practical) is different than *just-in-case* education (life, holistic).
  - "I want shallow, *just-in-time* theories for my practice (to answer the boss' questions next week)."
- K-12 Education
  - Autodidactism, Self-directed learning
  - "Technology is ubiquitous; I can learn from Linux, Raspberry Pi, Arduino projects, and 3D printing."

Candidate Explanations (3/3)
(Other/Interdisciplinary)

- Behavioral Economics
  - Extreme Discounting
  - "I am deliberately *overweighting* short-term preferences and *underweighting* long-term goals*."*

- History
  - Merchant-Craftsmen (some land) and Craft Guilds (no-land) (apprentice → journeyman → master)
  - "Trades[men] and crafts[men] have always created local, persistent, supportive *communities*."

Open Questions

- Is this observation an *itinerant*, *ephemeral fad* or a *persistent structure*?
- Is this eco-system *geographic-specific* or due to its virtual nature, *widespread*?
- *Vis-à-vis* other paid work, is this type of work off-shorable over time?
- If this sector is indeed *growing*, is it still just a *small part* of the economy anyway?
- Is this phenomenon (like some other aspects of IT) *gender-skewed*?
- Does the evidence suggest privileging *competencies* over *degrees*?
- Is my experience with these "professionals", ultimately, the triumph of *practice* over *theory* (as seen from the perspective of a prospective, potentially confused *soon-to-be-in-some-debt-from-a-student-loan* CSU student)?

Open Questions

- Without college degrees, how will these "data scientists" eventually lead/be ethical/well-rounded, etc. over a 40-year career arc?
  - What about Licensing? Certifications? CPE? "Badges"?
  - How will they get promoted? Corollary: How, if ever, will the *model building* get more rigorous?
  - Can they can just enroll into holistic B.S./B.A. (pre-M.S./M.B.A.) programs later in life?
  - If applied math, computer science, statistics, and business were all individual degrees *before*, then how could we possibly make a single "data science" degree (all of them together) *now*?

- What is the role of Ph.D.'s and academics (other than paid consulting)?
  - If we help (e.g., lectures, provide experts, networking, etc.), are we, in fact, directly or inadvertently helping young folks *not* go to College?
  - What do we tell our Deans and Provosts about our deep involvement with individuals who able-but-unwilling to (traditional, B.A./B.S.) college?

- Isn't this phenomenon, if evident, just a new version of (yet one more) vocation?
  - The CSU, as one component of the CA Master Plan, doesn't "train" individuals (non-professionals) into vocations. We "educate" individuals for broader success.
- Does any of this matter to our programs in either the short-run or long-run?
  - Your thoughts?

Further Reading
(peer reviewed)

- Brynjolffson, E., Hitt, L. M., and Kim H. H. (April, 2011), Strength in Numbers: How does data-driven decision-making affect firm performance? (working paper) [http://ssrn.com/abstract=1819486]

- Finzer, W. (2013) , "The Data Science Dilemma", *Technology Innovations in Statistics Education*, 7(2) [https://escholarship.org/uc/item/7gv0q9dc]

- Hardin, J., Hoerl, R., Horton, N., and Nolan, D. (Oct. 12, 2014), Data Science in the Statistics Curricula : Preparing Students to "Think with Data" [http://arxiv.org/ftp/arxiv/papers/1410/1410.3127.pdf]

- Horton, N., Baumer, B., and Wickham, H. (Feb. 1, 2015) , "Setting the stage for data science: integration of data management skills in introductory and second courses in statistics" [http://arxiv.org/pdf/1502.00318.pdf]

- Tambe, P. (April, 2011), Big Data Investment, Skills, and Firm Value, *Management Science* (pre-print) [http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294077]

Further Reading
(non peer-reviewed)

- Davenport, T., and Patil, D. J. (2012), "Data Scientist: The Sexiest Job of the 21st Century", *Harvard Business Review*, October.

- History of the Organization of Work
  - http://www.britannica.com/print/topic/648000

- Los Angeles Economic Development Corporation
  - *High Tech in LA* (October, 2014)
  - http://laedc.org/wp-content/uploads/2014/10/High-Tech-in-LA_20141006_FF.pdf

- McAfee, A., and Brynjolfsson, E. (2012), "Big Data: The Management Revolution", *Harvard Business Review*, October.

- McKinsey Global Institute Report (Manyika, J. et al.)
  - *Big data: The next frontier for innovation, competition, and productivity* (May, 2011)
  - http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Further Reading
(additional references)

- Journals:
  - *Journal of Data Science; Data, dikw, big data and data science*
  - *Big Data; Machine Learning; Data Mining and Data Discovery*
- Periodicals
  - Analytics Magazine (http://www.analytics-magazine.org/)
- “Academic” Conferences
  - INFORMS Big Data conference (April, 2015) (Huntington Beach, CA)
    - http://meetings2.informs.org/wordpress/analytics2015/
  - ACM SIGMOD “Management of Data” conf. (May/June, 2015) (Melborne, VIC, AU)
    - http://www.sigmod2015.org/
  - ACM KDD “Knowledge Discovery and Data Mining” conf. (Aug, 2015) (Sydney, AU)
    - http://www.kdd.org/kdd2015/
- “Practice” Conferences
  - TDWI Big Data conference (February, 2015) (Las Vegas, NV)
    - http://meetings2.informs.org/wordpress/analytics2015/
  - R Users Conference (June/July, 2015) (Denmark) (2014 was at UCLA)
    - http://www.r-project.org/useR-2015