



Data and Culture for Publishers

Roger Magoulas
Research Director
O'Reilly Media

roger@oreilly.com



- **O'Reilly Research**
- **Data**
 - **BookScan-based Retail POS Mart (Computers)**
 - **Ebooks / Ecommerce**
 - O'Reilly only
 - **Conferences**
 - **Job Post DB**
 - **Facebook and MySpace application usage**
 - **Apple iPhone AppStore ranks**
 - iBook Ranks since iPad release
 - **Top Twitter Users and Usage**
 - **US Government Data Analysis**
 - CTO Jobs Studies / HHS Jobs Trends
- **Analysis / Access / Communications**
 - **Research Portal**
 - **Sync'd**

- Research supports O'Reilly mission of changing the world by spreading knowledge of innovators
 - Quantitative and qualitative research on technology adoption
 - to support publishing / conferences and beyond
- Three people: quant, ops, data
 - many shared duties
 - access to fantastic O'Reilly social network
 - informs our perspective
- dmart – lots of value add
 - 11+ dimensions / 1K topic taxonomy / data from 2004
- emart – deal analysis
- Job data – messy
 - 15 Tb / 1.8 b rows
 - mostly for tech adoption, HHS project
- Apple iBook
 - iPad first day to figure out data we could use
- Twitter – sentiment and event analysis



- **Stats and Analysis as the 'sexy' job of the coming era**
- **More data, more types of data and big data tools**
- **Increased skills integration**
- **Cross-Discipline**
- **Machine Learning / Natural Language Processing**
- **O'Reilly Strata Conference**



- Google / Facebook / Zynga / LinkedIn
- Text, sensors
- Collaboration/Integration of data disciplines to speed and deepen analysis
 - do everything, no waiting
- Wide net for data skills and technology – physics => science; biostats => business
- Beyond stats – ML and NLP for unstructured text
 - people as the last mile

- Data Science is a meme more than an actual field, we refer to a set of skills that improve knowledge work productivity and effectiveness; the meme is based on our seeing how people with these skills have made an impact at companies like Twitter, Facebook, LinkedIn and Google
 - Google the entire search engine is an example of an applied data science applications
 - Google Insights used for analysis that showed the swine flu outbreak faster than CDC data
- No new components, what is new is the level of integration between components to provide more sophisticated insights from increasingly large data sets
 - Moving beyond reporting to analysis, insight, predictions
 - New tools: big data management, data munging
 - New Sources: web, sensors
 - New data types: unstructured, graphs, multi-media
 - New tasks: classifying, summarizing, sentiment analysis
 - New techniques: collective intelligence, machine learning, natural language processing, modeling
- Hal Varian, Google chief economist, quote from interview
- More data
 - Sensors, smart mobile devices, web-based
 - Unstructured text, graphs, images, audio, video
- Skills Collaboration/Integration
 - The integration we'll focus on is based on the data science frame we present in the next slide (data management, data munging, analysis and presentation)
- Cross discipline analysis
 - Science learning from business and business learning from science
 - Biostats – Many of the data science folks we know and follow come from biostats backgrounds (e.g., Mike Driscoll, Brian Dolan, Pete Skomoroch, Joe Adler)
 - Other examples: genetic algorithms used to run business simulations and crowd control, randomized control trials used for economics and other social science, graph theory used for social network analysis
- Strata Conference (strataconf.com)
 - The business of data
 - Focus on integrating skills, collaborative work, building a community
 - Amazing buy-in by data science folks we most respect
 - Technology tracks
 - Including pre-conference classes on machine learning and math
 - Business tracks
- Themes – we focus more on folks building their own tools than on commercial products



▪ **Tell Stories**

- **Communicate results**
 - make vivid, memorable, social

▪ **Input to Decision Processes**

- **Provide relevant information, not decisions**

▪ **Real-Time Integration**

- **Integrating data / analysis / modeling / predictions into real-time processes**
 - **Feedback for users**
 - **Self-tuning algorithms stay relevant**
- **Support database of expectations**

- We're wired to respond to and remember stories, take advantage of innate human characteristic
- Data is not a black box you buy, it's a process you follow, an input to decisions, part of an experiment-based learning culture
- The output (the why) of data science:
- Humans are wired to respond to and remember stories
 - Analytic types can sometimes get caught up telling the story of how they performed the a study or worked toward a result, that is not the story to relate (not in this context, more on technique sharing later)
 - Supercrunchers by Ian Ayres provides good examples of how to package analysis for quick cognition and retelling (more on Supercrunchers shortly)
 - Data stories can be used to help promote and reinforce a data-oriented culture, stories tend to spread quickly, helping spread the lessons from the analysis throughout an organization
 - Stories a heuristic to remember data, helps to make them social
- Decision Support
 - Think of how data analysis can help with many decision processes
 - Don't rely on results to make decisions, results should lead to better understanding or to asking more questions
- Tell a story; show anomalies (exceptions); show trends
 - don't show numbers, always show magnitude (especially when showing RoC)
- Real-Time Integration
 - How data science gets put to work in an application context
 - In many cases cloud enabled, sophisticated analytics computed on server and delivered through a relatively thin, often browser-based, client (e.g., recommendation engines)
 - Some of the most interesting data science work supports real-time analysis
 - Web analytics
 - Recommendation engines
 - Who you might know apps
 - Ad tracking and analysis
 - Anti-fraud analysis
 - Data center / operations support (trouble alerts, reconfiguring / redeploying resources based on demand, energy management, cost management)
 - Mobile device voice recognition, computer vision, translation
 - Real-Time Analysis via a message bus architecture
- db of expectations – sense and respond hallmark of all living things and now we're building computer systems around this (e.g., recommendation engines that use multiple models and reformulate 20 times per day)



- **Data Management**
 - Loading
 - Big Data
 - Parallelism
 - Sandboxes
 - Integration with Analysis
- **Data Collection**
 - Scraping / Feeds / APIs
 - Parsing
- **Data Integration**
 - Identification / Association
 - Deduplication / Conditioning
- **Data Organization**
- **Analysis / Insight**
 - Exploration
 - Visualization
 - **Collective Intelligence**
 - Teasing Insights from Crowd Behavior
 - Crowdsourcing / Mechanical Turks
 - **Machine Learning**
 - Classifying / Deduplication
 - Clustering
 - Summarizing
 - Sentiment
 - Human Review
 - **Natural Language Processing**
 - Entity Extraction
 - Disambiguation
 - **Statistics / Probability**
 - **Predictive Modeling**
- **Culture / Organizational Behavior**
 - Quantitative Culture
 - Organize to Learn / Experiment

- The geeky stuff
- How we see the space
- Conditioning not quality – a cost / benefit decision
- Analysis/Insight – exploring the cave of the unknown
- Need culture to make most of data and insight
 - Understand the message
 - Address innumeracy
 - Value results appropriately
 - Think experiments
 - Stay curious – keep asking questions



Key Components:

Page 2

The collage features several data visualization slides and photos of speakers. The slides include: 'Release 20*' showing a large number of releases; 'U.S. Facebook Users by Age Group (M75/W6)' showing a pie chart with segments for 18-24 (21%), 25-34 (24%), 35-44 (24%), 45-54 (24%), and 55-64 (7%); 'U.S. iTunes App Store Compared Unique Apps' showing a bar chart of app categories; 'U.S. iTunes App Store Share of All Books Titles' showing a bar chart of publisher shares; and 'U.S. iTunes App Store Share of All Books Titles' showing a bar chart of publisher shares. Photos include a man in a blue shirt on a stage, a man in a brown jacket in front of a bookshelf, and a woman in a blue top.

- Most pandering you'll likely see at conference
- Joe – asking questions
- Laura – math major
- Mike Hendrickson, Allen Noren, Laurie Petrycki, Sara Winge

Taxonomies



- **To make sense of data:**

- Categorize
- Orthogonal Dimensions
- Hierarchical
 - Drill Up / Drill Down
- Dynamic

- **BISAC for books**

- Not enough for dynamic topics like computers / technology

- **Taxonomies are hard!**

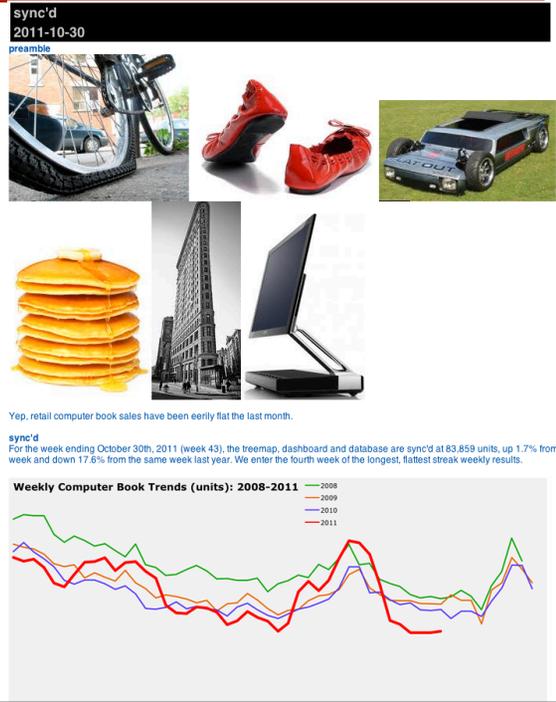
- Resources, Concentration, Ambiguity, Vigilance, Time, Madness
- Maintaining Multiple Rollups
- A Messy Process

- Linnaeus ref: categorizing fauna and flora
- BISAC great when it works
 - Dynamic example: rise of tablets, app programming
 - Ebooks, videos, one-offs, conference content, oh my
- Categorized > 25K+ books, whew!
 - triple check
 - new books / new topics / new relationships
- Ambiguity – some books hard to categorize, if multiple categorize, managing aggregate rollups (primary cat)
- Need to maintain consistency for multiple rollups
- Four rollups: topic, retail, division, cust (ecommerce)
- Machine Learning? it's possible



■ Sync'd Newsletter

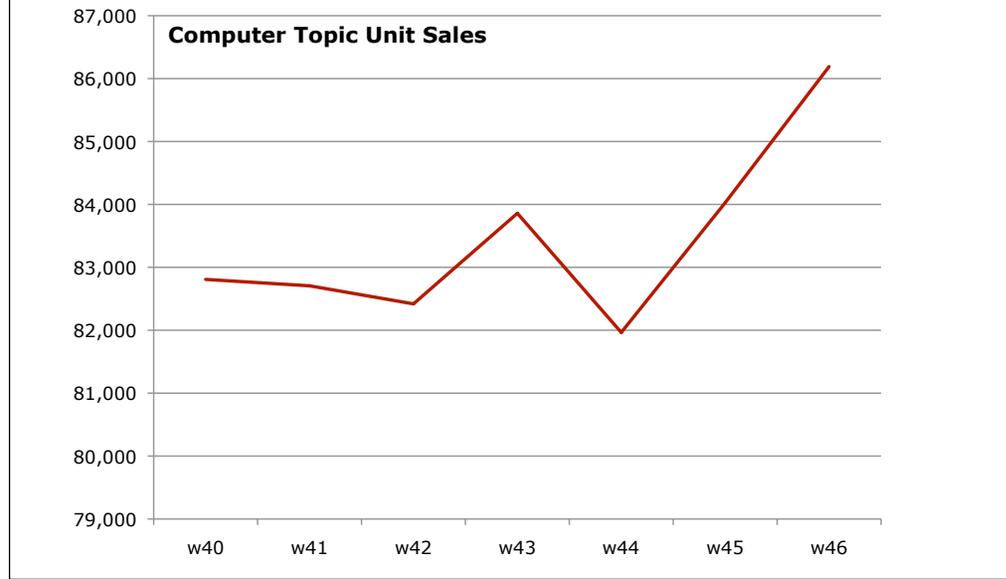
- Data + Narrative
- Anomalies
- Special Studies



- Weekly; delivered via E-mail to prompt reading
- Regular reporting
- Offbeat to keep interest



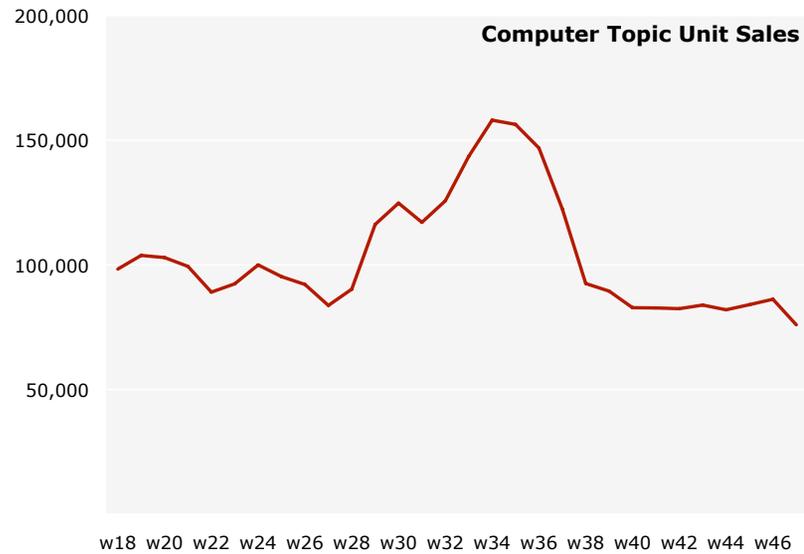
- **Magnitude Matters**
- **Context Matters**



- Lies, Damn Lies, and Statistics
- Default Excel



- **Magnitude Matters**
- **Context Matters**



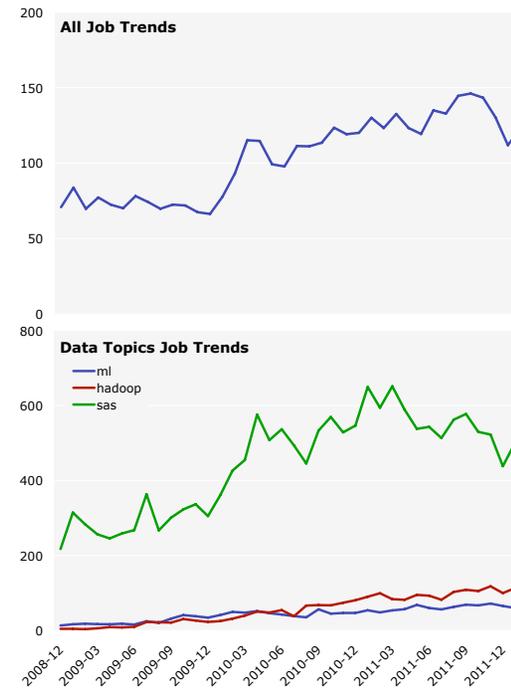
- Default Excel
- Could miss the unusually flat period

External Data - Jobs

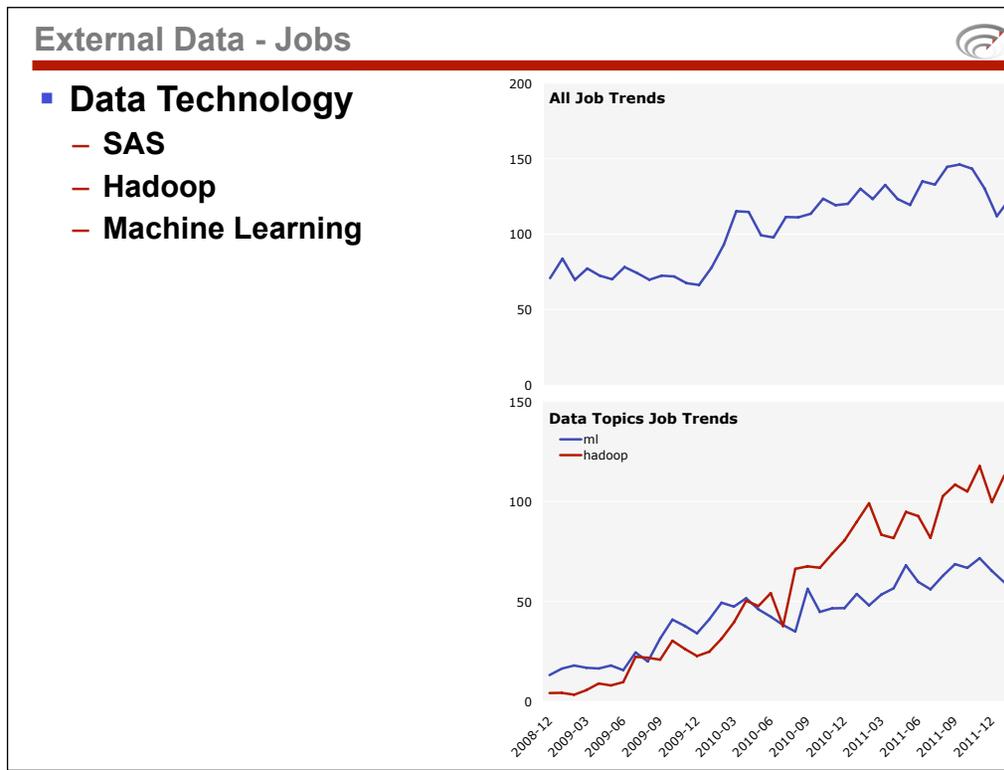


■ Data Technology

- SAS
- Hadoop
- Machine Learning



- Data a popular topic – help explain opportunity for O'Reilly
- Complements book sales data
 - better coverage for mature technologies
- SAS roughly matching the market
- Machine Learning & Hadoop smaller but growing

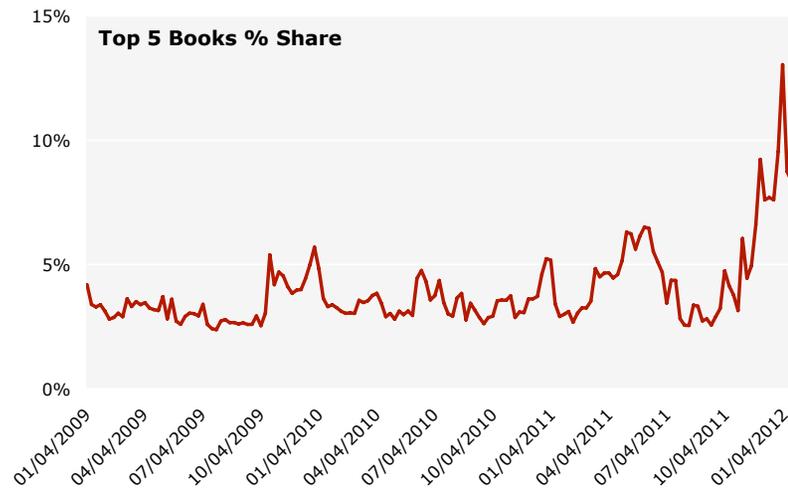


- Data a popular topic – help explain opportunity for O'Reilly
- Complements book sales data
 - better coverage for mature technologies
- SAS roughly matching the market – mature technology
- Machine Learning & Hadoop smaller but growing
- Drill down gives better sense of growth in these nascent fields
- Magnitude + rate of change



■ **Best Seller Share - Top 5 Books**

- Sustained Change Since Holiday Sales Season
- Hypothesis: Less Retail Shelf Space Focuses (Impulse) Demand to Fewer Titles or Perfect Storm

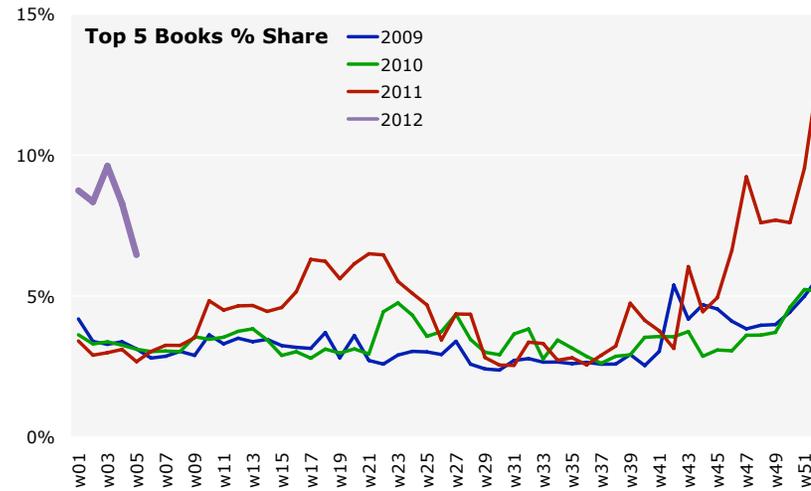


- Top books: (consumer oriented) iPad, iPhone, Kindle
 - bought to complement presents
- Monitor, consider implications to sales strategy
- B&N pushing Nook Book
- Seasonal



■ **Best Seller Share - Top 5 Books**

- Sustained Change Since Holiday Sales Season
- Hypothesis: Less Retail Shelf Space Focuses (Impulse) Demand to Fewer Titles or Perfect Storm



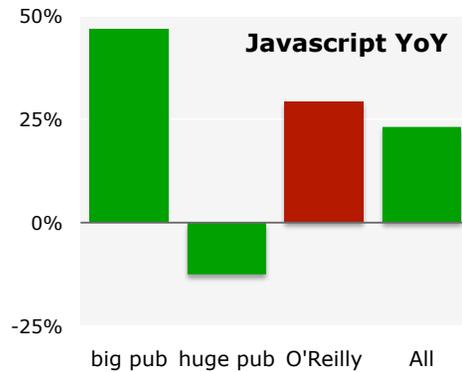
- Top books: (consumer oriented) iPad, iPhone, Kindle
 - bought to complement presents
- Monitor, consider implications to sales strategy
- B&N pushing Nook Book
- Seasonal



▪ **Publisher Efficiency by Topic - Javascript**

— **Hypothesis - Market Saturation**

- Unit Sales Up
- O'Reilly growing faster than market



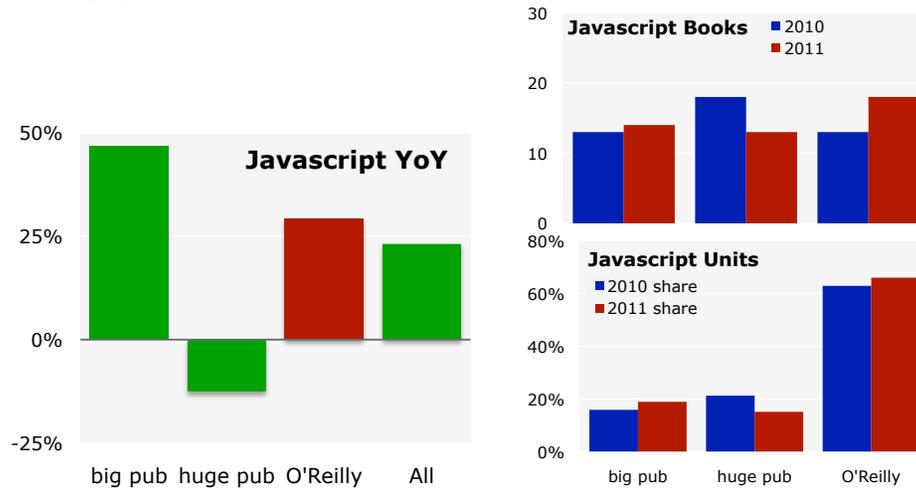
- Unit sales up in a down market
- O'Reilly growing faster than market



▪ **Publisher Efficiency by Topic - Javascript**

— Hypothesis - Market Saturation

- Unit Sales Up
- O'Reilly growing faster than market
- Dominant Share



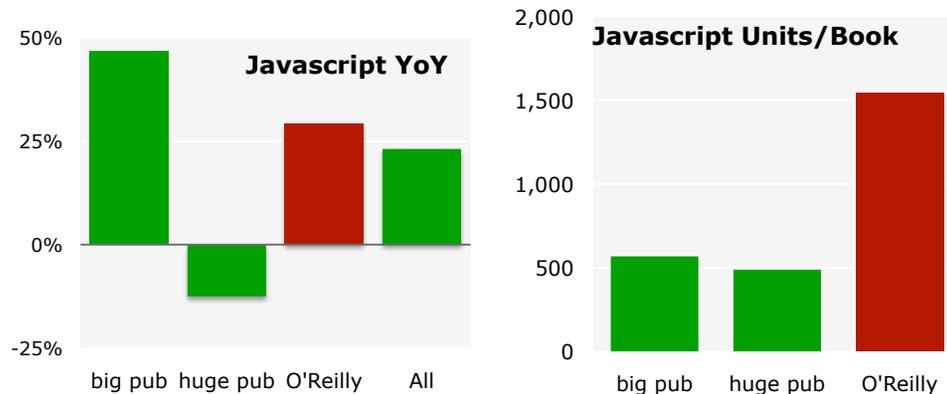
- Unit sales up in a down market
- O'Reilly growing faster than market
- Dominant share on similar publishing program
 - 2011 – rising to 66% share



▪ **Publisher Efficiency by Topic - Javascript**

— **Hypothesis - Market Saturation**

- Unit Sales Up
- O'Reilly growing faster than market
- High Share
- Efficient Publishing Program



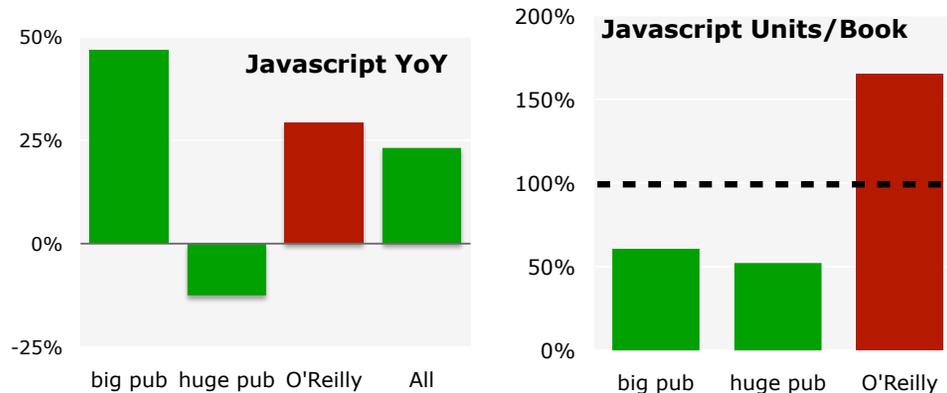
- Unit sales up in a down market
- O'Reilly growing faster than market
- Dominant share on similar publishing program
 - 2011 – rising to 66% share
- Much higher units / book ratio



▪ **Publisher Efficiency by Topic - Javascript**

— **Hypothesis - Market Saturation**

- Unit Sales Up
- O'Reilly growing faster than market
- High Share
- Efficient Publishing Program



- Unit sales up in a down market
- O'Reilly growing faster than market
- Dominant share on similar publishing program
 - 2011 – rising to 66% share
- Much higher units / book ratio
 - another view – 100% represents sales for average book in topic
 - O'Reilly well above

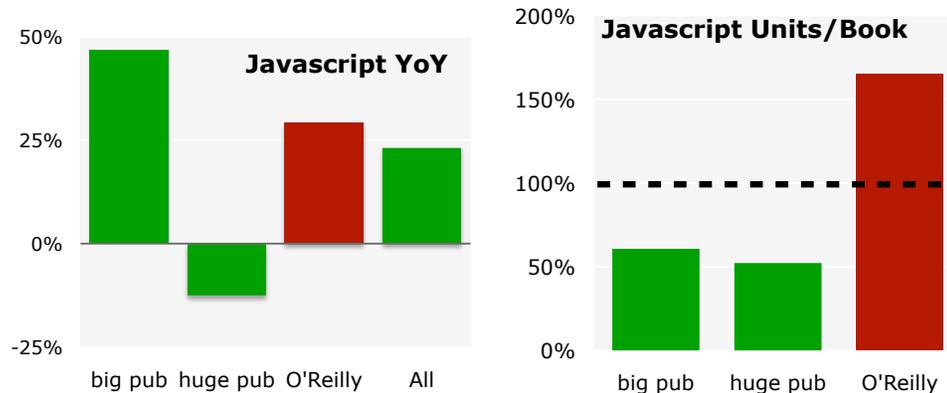


▪ **Publisher Efficiency by Topic - Javascript**

— **Hypothesis - Market Saturation**

- Unit Sales Up
- O'Reilly growing faster than market
- High Share
- Efficient Publishing Program

— **Not Saturated**



- Unit sales up in a down market
- O'Reilly growing faster than market
- Dominant share on similar publishing program
 - 2011 – rising to 66% share
- Much higher units / book ratio
 - another view – 100% represents sales for average book in topic
 - O'Reilly well above
- Consider publishing
- Note arc of analysis

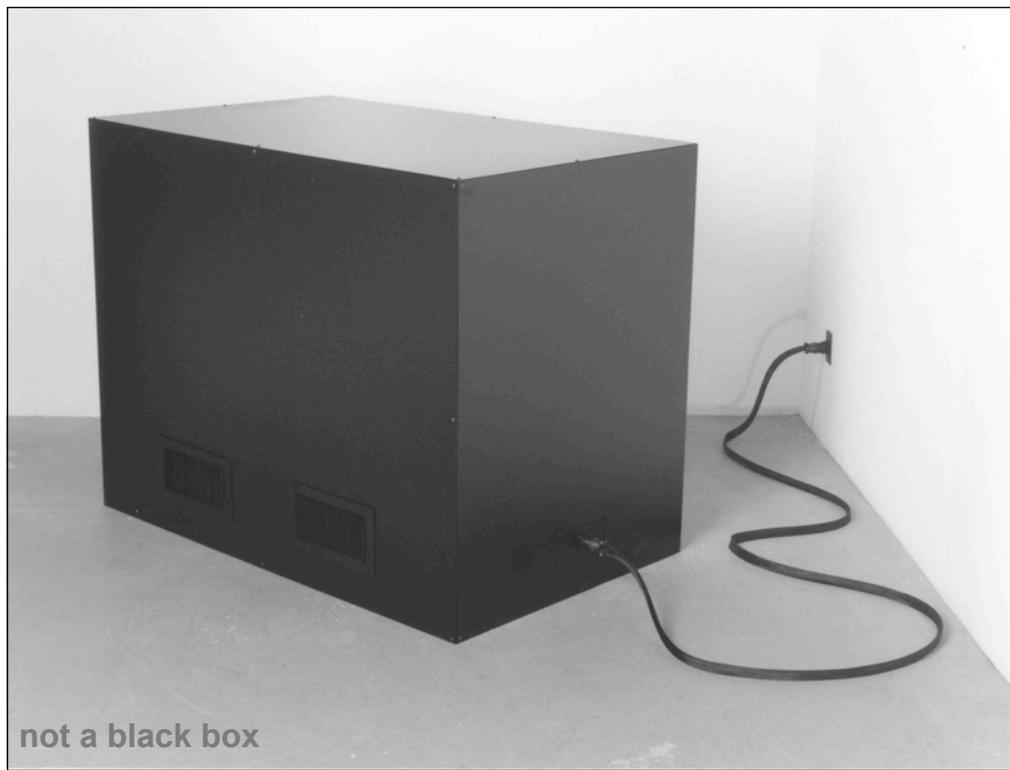
What Can You Do



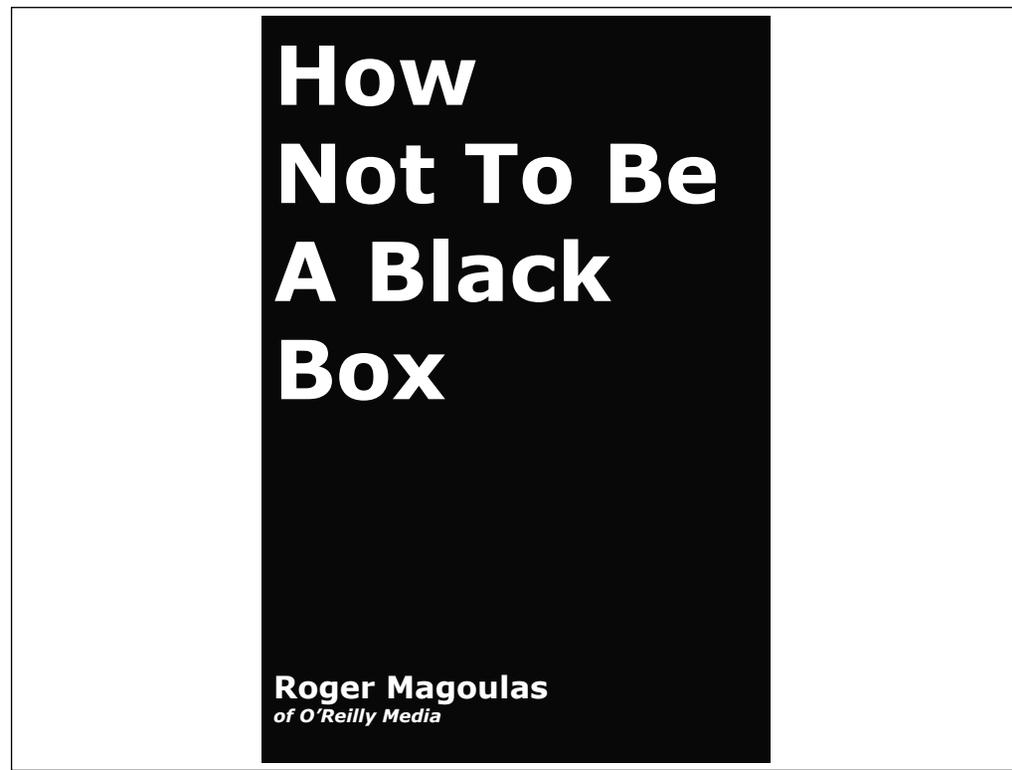
- **Get Data Savvy**
 - Find a Ben, Math Club
- **Keep Analysis Close to Data**
- **Go Outside**
- **Encourage Collaboration / Critical Vetting**
 - Internal and External
- **Experiments as Fundamental Business Process**
 - New Risk: Measuring cost of what you won't learn
- **Supply-Side Analytics**
 - Sandbox
- **Communicate with Stories**
- **Scale Up Decision Making to Match Data**

- Data Savvy – Get a book, take a class
- Let analysis requirements drive how data organized
 - learn from agile
- Critical Vetting
 - Smell Test
 - ref Jonah Lehrer teams article re: constructive criticism
- Go Outside – augment your data w/ outside sources
 - Gov (Census, BLS), Factual, Scraping
 - crowdsourcing
- Experiment
 - Test hypothesis; learn from everything – feedback loops
- Supply-Side Analytics – Let analysts explore (Google 20%)
 - Sandbox – create big data areas w/ quick spin-up and full data management support (cloud)
- Numbers w/ no story don't resonate, don't lead to action
- Occam's razor – look for simplest analysis path

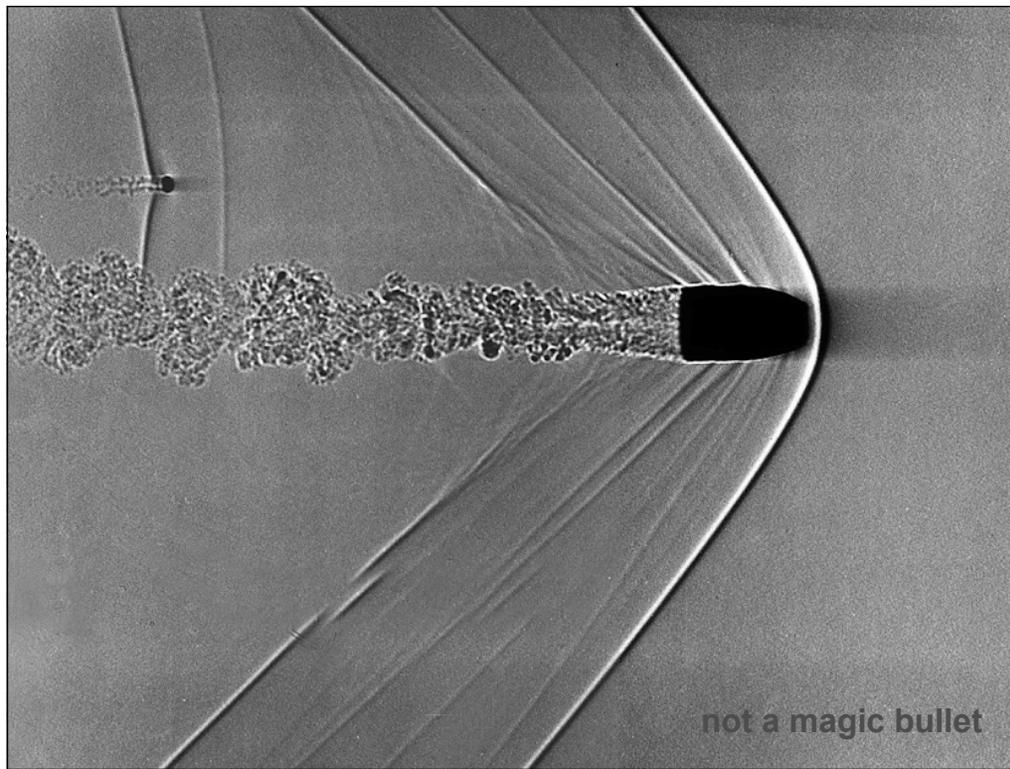
- Small team, but with enough a range of expertise, covering the data management and data insight skills required to perform an analysis and explain the results
- The integrated team is design to prevent process road blocks, and to encourage everyone to pick up the skills from others
 - Don't set the expectation that everyone can acquire and become expert at all the data science skills, but they should have enough knowledge to get basic tasks done on their own – not to have to wait if others are busy
- Online coordination tools like Google Docs allows more flexibility, and geographic independence
- Agile / Extreme Programming for training
 - Double folks up on tasks to encourage cross training
 - Encourage walk-throughs and team vetting of intermediate steps to help facilitate organization learning and expectations
- Creates example of how to organize and how to integrate skills to increase analytic productivity
- Sharing (covered in previous slides)
 - Open source style over-sharing to build skills
 - Sharing techniques and tools to get feedback, improvement, learn
 - Other recommendations covered in earlier slide: intra-company discussion, join public discussions and meet-ups, actively share
- Experimentation – learning as key goal of all processes, consider risk of missing opportunity to learn
- Supply-side analytics (as covered in previous slides)
 - give data science team time and resources to run their own, uncommissioned studies
 - Shows importance of analysis function, demonstrates data-driven culture
 - Take advantage of organizational and data knowledge accumulated in the analysis group
- Analytic Sandbox
 - Provide an easy-to-configure, quick-to-spin up facility for quickly building fast query data stores – a cloud like facility that provides fast cycling for computational analysis
 - No or easy requisition process
 - Big storage to allow experiments in data organization that can speed analysis iteration cycles
 - MapReduce can improve analytic productivity by providing fast, parallel execution of procedural logic beyond what SQL on its own can provide (e.g., logic between rows not covered by aggregate functions)
 - Hadoop or MPP databases (Aster, Greenplum, Vertica)
- Integrated Tools
 - E.g., Datameer, Mathematica, Karmasphere, Big Sheets, Splunk, Palintr
 - The tools listed all tend to perform more of the analysis functions, e.g., mixing data loading, transforming and organizing data, built-in analysis tools and built-in visualization; some of the tools have provide easy access to web based data
- Avoid becoming paralyzed by possibilities (Driscoll CIA example)



- data – doesn't make decisions

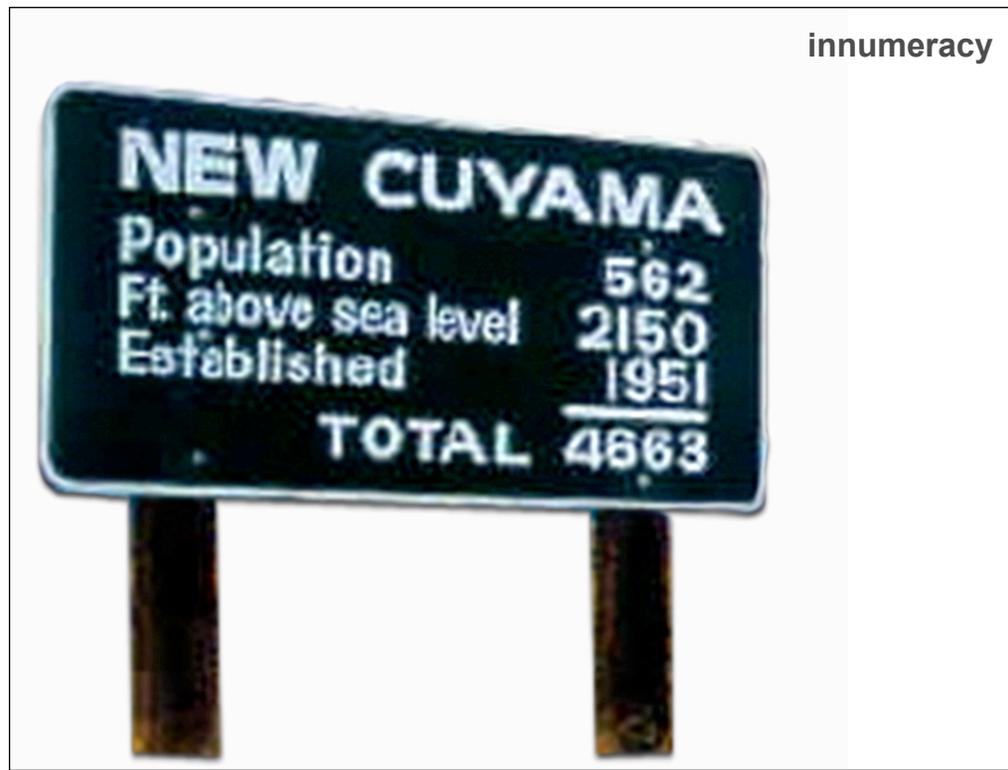


- data doesn't make decisions

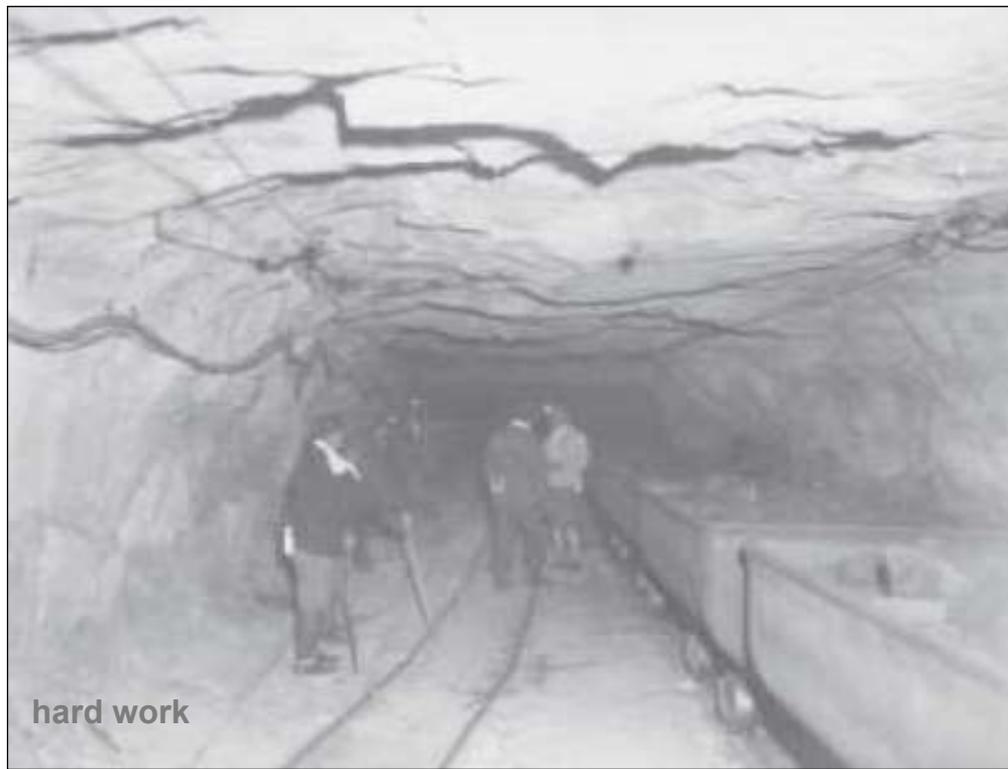


- or solve problems on its own

innumeracy



- There will be issues



- Data a process – w/ no end
- Requires resources, commitment, training, vigilance – find O'Reilly books
- Best analysis poses more questions than it answers
- Remember magnitude, direction, rate of change
- Art and Science
 - designing an experiment still an art
 - Freakonomics / Supercrunchers for inspiration
- Like many hard things – its a lot of fun
 - Ref: Improving Cognitive Functioning article, Doing things the hard way as one of five keys to increasing cognitiion
 - Others: Be creative, Constant Challenge



- stay in the game



- enlightenment and...



- bliss



Quantitative Culture



- **Functionally Integrated Teams**
 - **Responsible for all steps of analysis:**
 - Data Management / Munging
 - Analysis / Visualization / Story Telling
- **Encourage collaborative development**
 - **Cross-Function Coordination (e.g., via Google Docs)**
 - **Technical Cross-Training**
 - Use Agile and Extreme Programming Methods
- **Share processes, techniques, tool knowledge, results**
 - **Encourage integrated approach**
 - **Open source philosophy**
- **Experimentation as Fundamental Process**
- **Supply-Side Analytics**
- **Analytic Sandbox**
 - **Provide access to large, flexible, high performance data management systems**
- **Scale Up Decision Making to Match Data**

- Address problems large, enterprise scale organizations face optimizing the value of their data when they have distributed analytic silos and large, tightly controlled data stores
- Start integrating teams as example of a new way to work, in a cross-disciplinary fashion, with rapid, iterative development processes (Agile-like)
- Small team, but with enough a range of expertise, covering the data management and data insight skills required to perform an analysis and explain the results
- The integrated team is design to prevent process road blocks, and to encourage everyone to pick up the skills from others
 - Don't set the expectation that everyone can acquire and become expert at all the data science skills, but they should have enough knowledge to get basic tasks done on their own – not to have to wait if others are busy
- Online coordination tools like Google Docs allows more flexibility, and geographic independence
- Agile / Extreme Programming for training
 - Double folks up on tasks to encourage cross training
 - Encourage walk-throughs and team vetting of intermediate steps to help facilitate organization learning and expectations
- Creates example of how to organize and how to integrate skills to increase analytic productivity
- Sharing (covered in previous slides)
 - Open source style over-sharing to build skills
 - Sharing techniques and tools to get feedback, improvement, learn
 - Other recommendations covered in earlier slide: intra-company discussion, join public discussions and meet-ups, actively share
- Experimentation – learning as key goal of all processes, consider risk of missing opportunity to learn
- Supply-side analytics (as covered in previous slides)
 - give data science team time and resources to run their own, uncommissioned studies
 - Shows importance of analysis function, demonstrates data-driven culture
 - Take advantage of organizational and data knowledge accumulated in the analysis group
- Analytic Sandbox
 - Provide an easy-to-configure, quick-to-spin up facility for quickly building fast query data stores – a cloud like facility that provides fast cycling for computational analysis
 - No or easy requisition process
 - Big storage to allow experiments in data organization that can speed analysis iteration cycles
 - MapReduce can improve analytic productivity by providing fast, parallel execution of procedural logic beyond what SQL on its own can provide (e.g., logic between rows not covered by aggregate functions)
 - Hadoop or MPP databases (Aster, Greenplum, Vertica)
- Integrated Tools
 - E.g., Datameer, Mathematica, Karmasphere, Big Sheets, Splunk, Palintir
 - The tools listed all tend to perform more of the analysis functions, e.g., mixing data loading, transforming and organizing data, built-in analysis tools and built-in visualization; some of the tools have provide easy access to web based data
- Avoid becoming paralyzed by possibilities (Driscoll CIA example)



- **Publishing / Conferences / On-line / Radar / Research**



foo camp

- **Changing the world by spreading the knowledge of innovators**
- **We're essentially story-tellers**
- **Democratizing Innovation**
- **“The Future is here, it's just not evenly distributed”**
– William Gibson

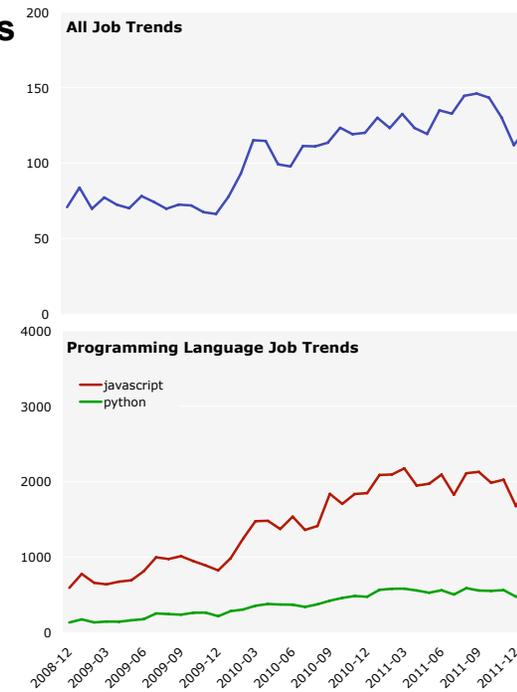
- O'Reilly and the Public Good:
 - Support for CfA; work for HHS / NIH; Explicit support for open source
- O'Reilly – more than just books; first comm'l, ad supported web site, first to use collab filtering; coined open source; coined web 2.0
 - Thought Leaders
 - ran conference that named Open Source
 - Named Web 2.0 and developed principles, including collective intelligence
 - instigated unconference movement w/ Foo camp
 - instigated DIY movement w/ Make
- democratizing innovation – MIT's Eric Von Hippel, users as greatest source of innovation; cheaper tools; global communications and sourcing give users/innovators more power
- Make magazine a manifestation of democratizing innovation
- Fundamentally we are storytellers
- who would have thought amazon would own cloud computing, apple would own music biz, people would pay for apps
- O'Reilly has unparalleled access to a great technical social network
 - events and reputation keeps us close to the community; we find out what they think is interesting; we have access to many social alpha geeks, not just nerds, many have started successful business or written wildly popular apps
 - entrepreneurial
 - subversive, disruptive, fail fast
 - DIY / hacking
 - amateur professionals
 - open source / collaborative
 - catalyst for alpha geek community
 - foster cross disciplinary mixing
 - international reach (recently in rome, milan and athens)
- Many start-ups pass by O'Reilly (incl: int'l)
- monitor variety of app platforms, facebook, myspace
 - heard about twitter when 12 users; youtube founders at Foo, 14 months before sale to google
- David Brooks – got idea for Alpa Geeks post from O'Reilly
- We do geopolitical and industrial policy analysis for gov't
- Research – quantitative and qualitative research for internal and external clients

External Data - Jobs



■ Programming Languages

- Python
- Javascript



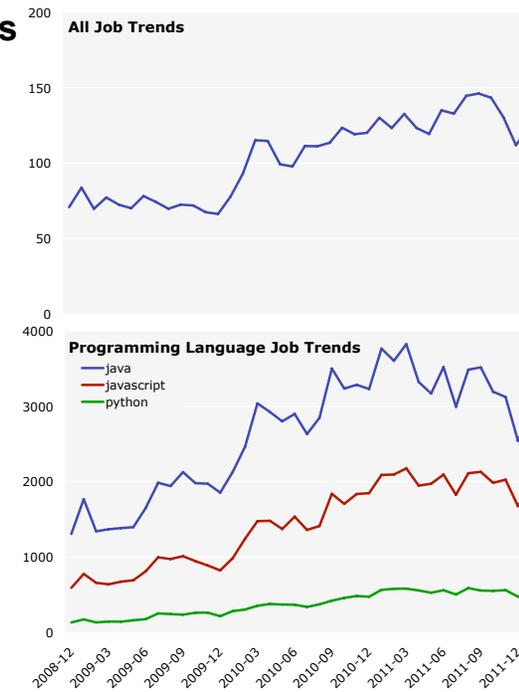
- Complements book sales data
 - better coverage for mature technologies
- Popular languages

External Data - Jobs



■ Programming Languages

- Python
- Javascript
- Java



- Complements book sales data
 - better coverage for mature technologies
- Increasingly Popular languages
- Java – mature tech shows strength