Data at O'Reilly

- **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
What's New in Data

- 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
- Google Insights used for analysis that showed the same flu outbreak faster than CDC data
- 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
- Structured text, graphs, images, audio, video
- Skills Collaboration/Integration
- The 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

The Why of Data

- **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)
- Super Crunchers by Ian Ayres provides good examples of how to package analysis for quick cognition and retelling (more on Super Crunchers 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 Science

<table>
<thead>
<tr>
<th>Data Management</th>
<th>Analysis / Insight</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Loading</td>
<td>– Exploration</td>
</tr>
<tr>
<td>– Big Data</td>
<td>– Visualization</td>
</tr>
<tr>
<td>– Parallelism</td>
<td>– Collective Intelligence</td>
</tr>
<tr>
<td>– Sandboxes</td>
<td>• Teasing Insights from Crowd Behavior</td>
</tr>
<tr>
<td>– Integration with Analysis</td>
<td>• Crowdsourcing / Mechanical Turks</td>
</tr>
</tbody>
</table>

**Data Collection**

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Machine Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Scraping / Feeds / APIs</td>
<td>• Classifying / Deduplication</td>
</tr>
<tr>
<td>– Parsing</td>
<td>• Clustering</td>
</tr>
</tbody>
</table>

**Data Integration**

<table>
<thead>
<tr>
<th>Data Integration</th>
<th>Natural Language Processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Identification / Association</td>
<td>• Entity Extraction</td>
</tr>
<tr>
<td>– Deduplication / Conditioning</td>
<td>• Disambiguation</td>
</tr>
</tbody>
</table>

**Data Organization**

<table>
<thead>
<tr>
<th>Culture / Organizational Behavior</th>
<th>Statistics / Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>– Quantitative Culture</td>
<td>– Predictive Modeling</td>
</tr>
<tr>
<td>– Organize to Learn / Experiment</td>
<td></td>
</tr>
</tbody>
</table>

- 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
- 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
• Research Portal – 3 clicks to a book
• Complete market, not just us
Research Portal

- Treemap / Topic Summary / Book Detail
Publishing Data

- **Sync’d Newsletter**
  - Data + Narrative
  - Anomalies
  - Special Studies

- Weekly; delivered via E-mail to prompt reading
- Regular reporting
- Offbeat to keep interest
Presenting Data - A Digression

- Magnitude Matters
- Context Matters

- Lies, Damn Lies, and Statistics
- Default Excel
Presenting Data - A Digression

- Magnitude Matters
- Context Matters

- Default Excel
- Could miss the unusually flat period
• 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
• 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 5 Books % Share

- **2009**
- **2010**
- **2011**
- **2012**

- **Top books:** (consumer oriented) iPad, iPhone, Kindle
  - bought to complement presents
- Monitor, consider implications to sales strategy
- B&N pushing Nook Book
- Seasonal
- Unit sales up in a down market
- O'Reilly growing faster than market
- Unit sales up in a down market
- O’Reilly growing faster than market
- Dominant share on similar publishing program
  - 2011 – rising to 66% 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
• Much higher units / book ratio
- 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
- 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

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 – doesn’t make decisions
How Not To Be A Black Box

Roger Magoulas
of O'Reilly Media

- data doesn't make decisions
• or solve problems on its own
• 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 cognition
  • Others: Be creative, Constant Challenge
• stay in the game
• enlightenment and...
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 an 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 allow 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
- 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)
O'Reilly Media

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

- 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 Hipple, 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
    - 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
- **Programming Languages**
  - Python
  - Javascript

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

**Programming Languages**
- Python
- Javascript
- Java