Big Data for Records Management
Are we there yet?

May 20, 2015

Mike Grosvenor

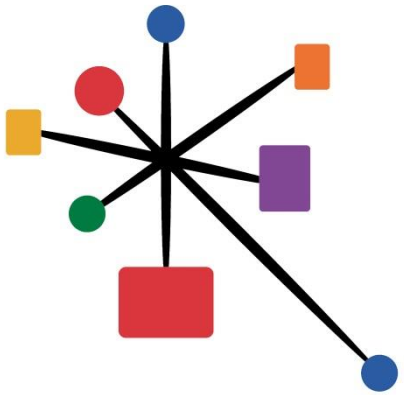
Access Sciences CELEBRATING 30 YEARS

Mike Grosvenor

- Consulting Manager, Access Sciences Corp.
  - Business Intelligence
  - Content Management
  - Data Governance
  - Regulatory Compliance
- Contact Info:
  - Email: mgrosvenor@accessscience s.com
  - Phone: 713.664.4357 (office) / 713.715.8154 (mobile)
- Linked In: http://www.linkedin.com/in/mikegrosvenor
- Twitter: @gro7or (Warning: most tweets about basketball, snacks or the weather.)

Access Sciences Overview

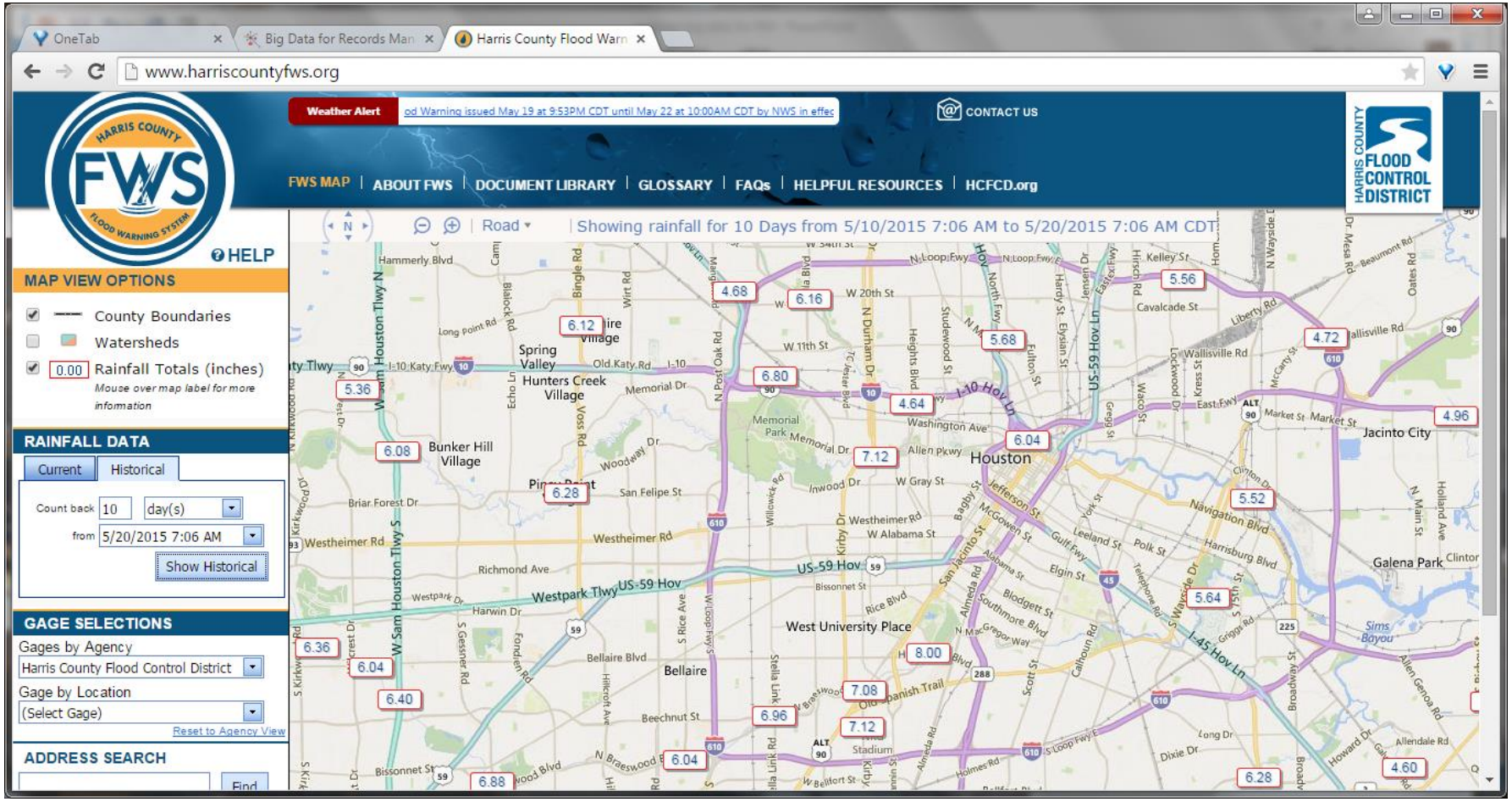
- U.S.-based, **employee-owned** company with a global reach
- Over the past 5 years Access Sciences has served **100+ clients** in over a dozen industries in 47 countries
- Client base ranges from small organizations to Fortune 100 companies **across a variety of industries**
- Named to *Houston Business Journal's* Fast Tech 50, Houston Fast 100, and multiple-year recipient of the Alfred P. Sloan Award for Business Excellence and Workplace Flexibility
- Certified Women's Business Enterprise (WBE) and Historically Underutilized Business (HUB)



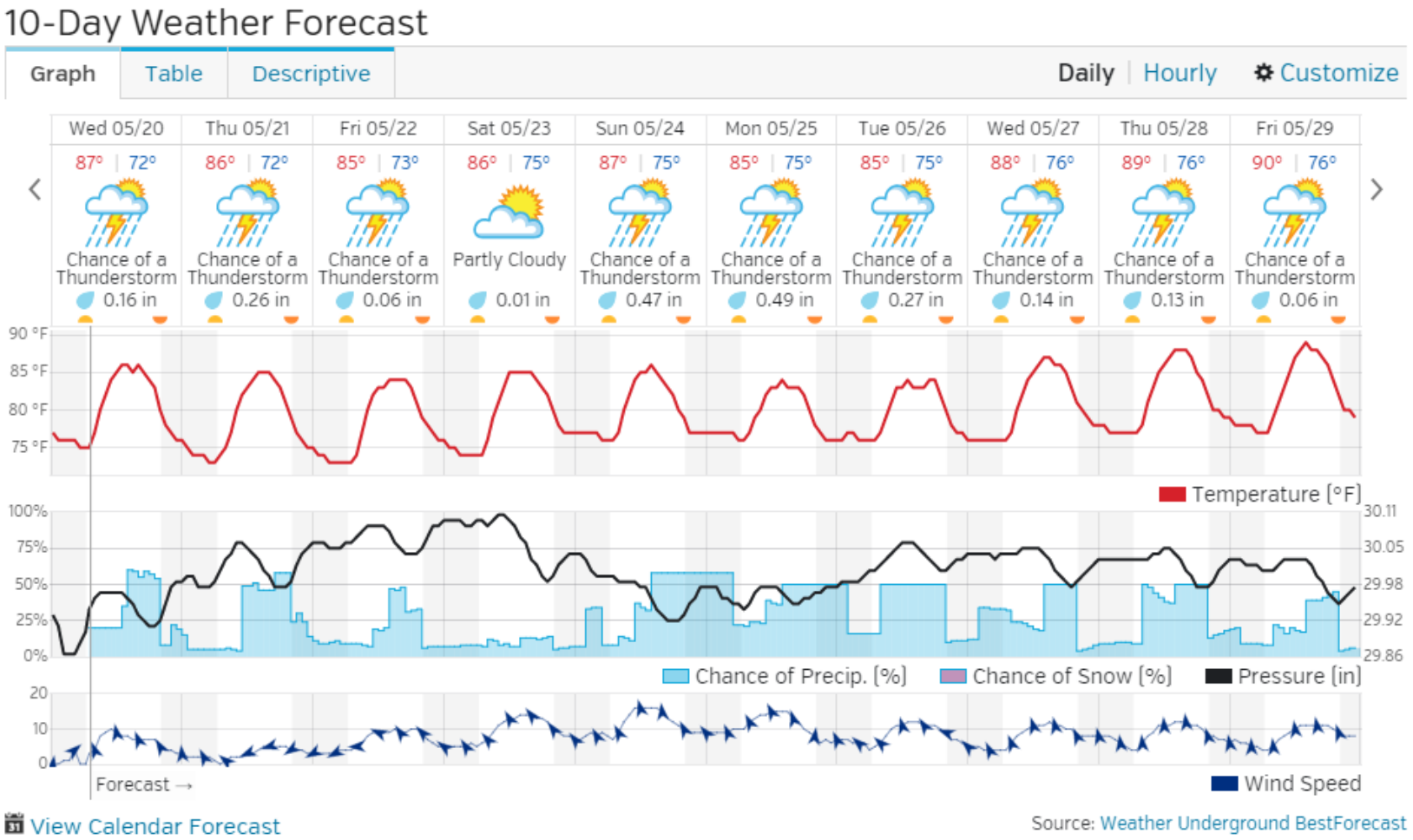
Webinar Agenda

- Simple introduction to key concepts and technology.
- Practical suggestions on identifying and promoting Big Data opportunities in your organization.
- Additional topics:
  - The increasing importance of metadata.
  - Tips on project management and stakeholder engagement.
  - Criteria for the selection of software.

Introduction (and Disclaimer)



As I was saying…



The World Of Data Is Changing

"By 2015, organizations integrating high-value, diverse, new information types and sources into a coherent information management infrastructure will outperform their industry peers financially by more than 20%."

– Gartner, Regina Casonato et al., "Information Management in the 21st Century" (2011)

The news about Big Data is becoming Big Data itself.

What Big Data is Not

- The Cloud, although 'Software-Defined Data Centers' could be part of a Big Data program.
- Hadoop, which is software that manages a distributed file system (HDFS), is the poster child for Big Data technology.
- Data Science, which combines aspects of Statistics, Programming and Business Modeling (remember when we called this 'research'?) drives much of the interest in Big Data for analysis and forecasting.

What Big Data Is (Beyond the 3 V's)

- Big Data can be whichever type of information you currently struggle to manage:
  - Real-time
  - Machine / sensor
  - Social media
  - Text / unstructured
  - Distributed / proprietary
  - External / subscription
- Big Data refers to information that would require more resources to integrate, analyze and store than would be justified by the gained value, using legacy technology alone.

Unstructured Data

- As much as 80% of corporate information*
  - Documents
  - Images
  - Log files
  - Machine Data
- Value often declines with time so you need more of it to make it useful
  - Dark Data (Blind Spot)
  - Stored but never accessed (*if even 25% has value that might be 50x what's currently being analyzed.)

Structured Data

Relational Databases are a Legacy System

- Technology adapted to specialized uses
  - Data Warehouse
    - Star Schema
    - Column DB (such as Apache Cassandra, HP Vertica, DB2 BLU*)
  - OLTP
    - In-memory DB (such as SAP HANA*, MS Hekaton, H-Store/VoltDB)
    - Memory prices falling

*denotes a DB that is both in-memory and column-based

Enhancement or Expansion requires significant development

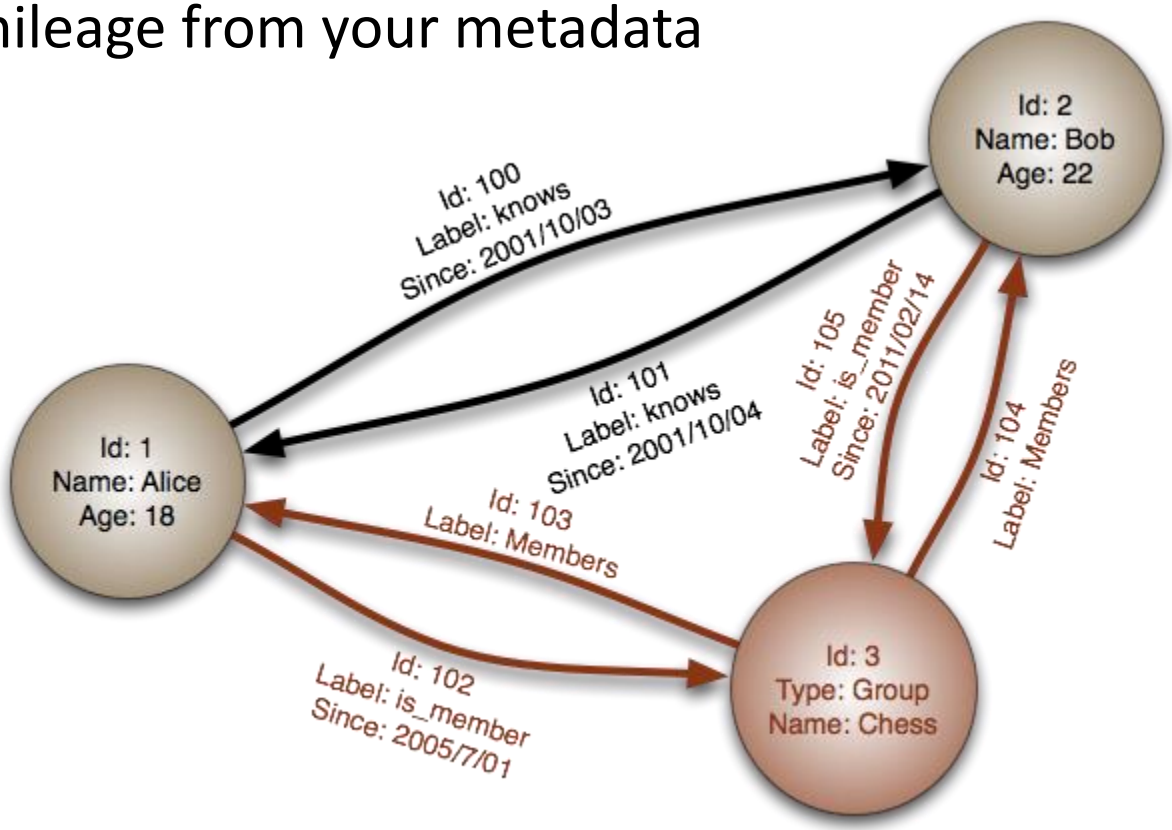
NoSQL Databases and More Technology Jargon

NoSQL has evolved to become ‘Not Only’ SQL. What’s really missing is the table join. And locking. And logs…

- Column Store – stored in clusters based on key cols, compressed otherwise, efficient disk access, scalability
- Document DB – JSON is the model, ‘denormalized’ based on access pattern, via Java, Python, .NET, Node.js, R, etc.
- Key Value – Cloudera Hbase (Hadoop), best for ‘embarrassingly parallel’ analysis. Even Google has quit using MapReduce.
- Graph DB (SparQL) & Array DB (SciQL) – newer, specialized

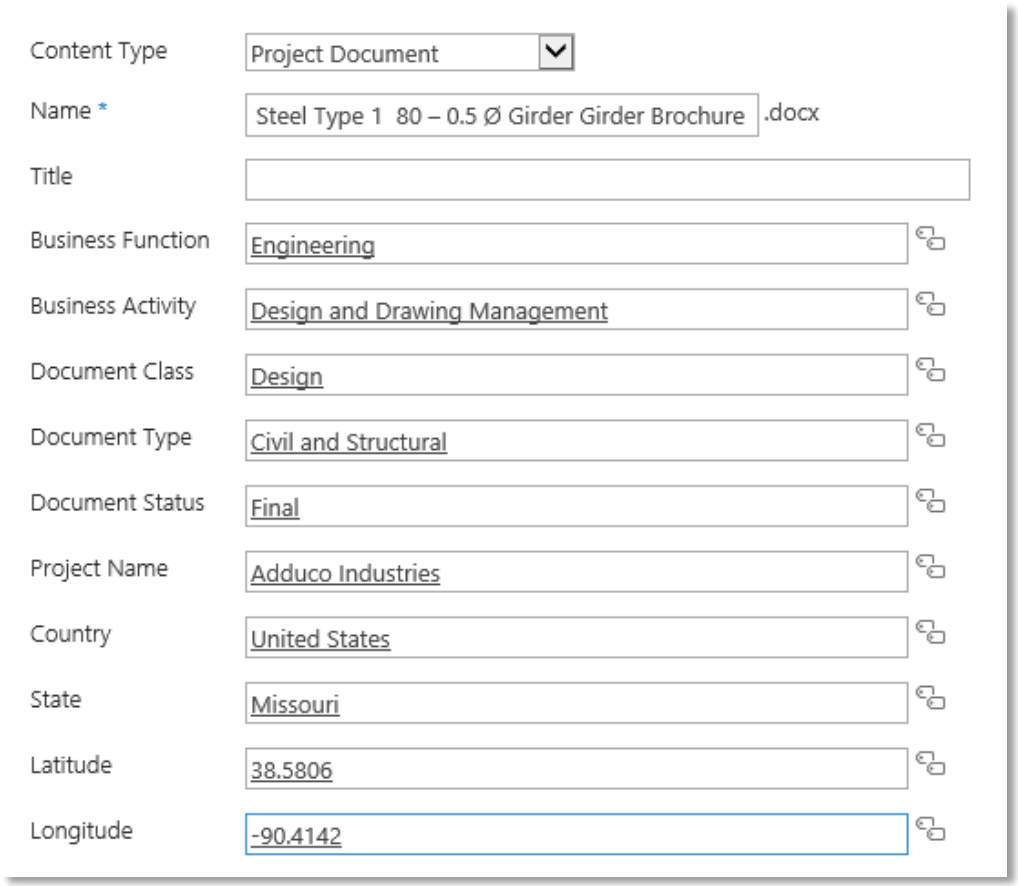
Graph Database

- Semantic relationships as important as the data itself
- More mileage from your metadata



Metadata

- Sources
  - Data Models
  - Process Flows
  - Content Management Systems
- Types (Depending on Use)
  - Facets/Tags
  - Hierarchy (folders, sites)
  - Definitions
  - Classification
  - Attributes



Focus on Metadata

Current Approach	Enhanced Approach
Promote high data quality standards.	Manage development of the metadata repository as a product.
Analyze data for quality and reconcile data issues.	Engage with IT partners as an advocate for curated, integrated data.
Identify and expose new data sources.	Interpret new system capabilities, translate to metadata strategy.
Drive resolution of data integrity issues by working across stakeholders.	Collaborate with analysts and developers to establish references across functions and departments.
Understand and document data relationships, data process flows.	Evaluate metadata for unstructured content as well as structured data for context and process alignment.

Plan Ahead but Be Agile

- Agile development practices invite Data Steward involvement to help teams respond to opportunities for innovation

- Capturing the characteristics, lineage and relationships of data serves multiple purposes that align nicely with Agile principles:
  - Communication – Team members have common reference
  - Validating decisions – Users develop business vocabulary and definitions
  - History – Subsequent iterations can build on the platform without rework
  - Governance – Outside groups can see process methodology compliance

Analogy: Looking Back While Moving Forward

- Archaeology
  - You're thinking bullwhip and fedora, but it's a trowel and sunblock.
  - Museum-quality artifacts

    Do you think there is anything worth finding?

    Exception: RISK – rare, important, sensitive, knowledge

- Urban Planning
  - Looking forward:

    Collaboration & Community

    Services & Support

    Growth
  - Put the process in place and the data will fill it. (However, "If you build it, they will come" isn't enough. Do you remember that traffic jam at the end?)

The Role of the Data Steward

- Data Stewards have experience identifying challenges in establishing common data management approaches across systems.
- Big Data requires an understanding of business problems and targeted application of technology.
- Agile development practices are common in Big Data projects and invite Data Steward involvement to help teams respond to opportunities for innovation.
- The selection of tools for Big Data challenges should fit the organizations priorities to minimize impact of the learning curve.

Big Data is a Team Sport

- There aren't many projects that don't have data components.
- The 3 P's, according to Fred Trotter (@fredtrotter)
  - Problem
  - Practice
  - Potential
- Self-directed coding enthusiasts (i.e. Hackers) can be helpful.
- The real lesson in *Moneyball* was that the right numbers give you an advantage, but don't fire the scouts just yet.
- The tools will catch up. Soon.

Stewardship Activities & Agile Principles

Stewardship Activity	Agile Principle
Metadata Repository allows work to be iterative instead of repetitive.	Agile processes promote sustainable development.
Management of metadata as an asset provides value to business projects.	Continuous attention to technical excellence and good design enhances agility.
Support collaboration among data-savvy innovators.	Build projects around motivated individuals. Give them the environment and support they need, and trust them to get the job done.
Advise data architects on integration challenges faced by business.	Simplicity--the art of maximizing the amount of work not done--is essential.
Align use of structured and unstructured data with process/workflow.	At regular intervals, the team reflects on how to become more effective, then tunes and adjusts its behavior accordingly.
Engage users to adapt to a shift in strategy or structure.	Agile processes harness change for the customer's competitive advantage.
Extend data stewardship role through engagement of additional SMEs.	The best architectures, requirements, and designs emerge from self-organizing teams.

Ad-Hoc (Actionable) Analysis

- Iterative or Agile process for improved understanding
  - Start with standard reports from DW/BI System
  - Use unstructured data from multiple sources
  - Confirm or Correct based on additional information
    - Results provide insight and/or suggest further investigation
- Correlate, don't integrate
  - Analyze data without ETL overhead
  - Tool options
    - Excel
    - Tableau
    - NoSQL (or 'Not Only' SQL)
      - Document DB
      - Graph DB
    - Event Index (e.g. Splunk)

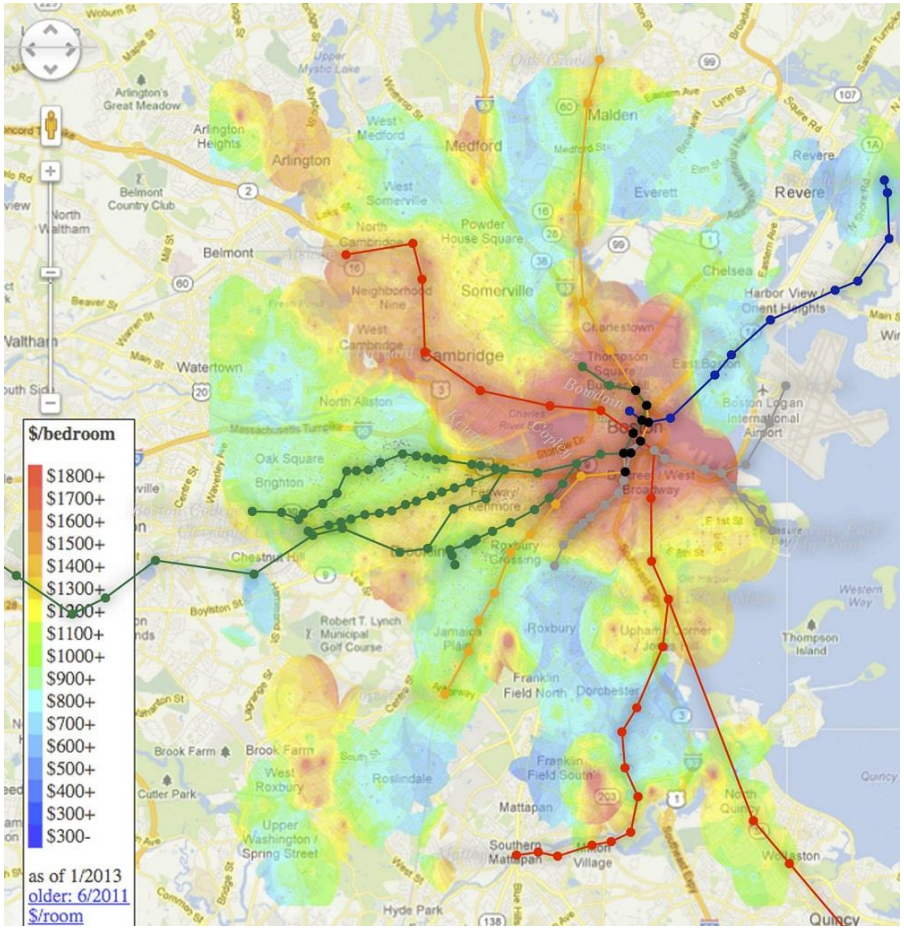
A Big Data Planning Analogy

- The Cheesecake Factory Menu
  - There's a lot of choice, and some generous serving sizes
  - It's probably better if you don't try it all by yourself

- Before you decide, consider:
  - Ingestion – just what is your appetite?
  - Storage – what are you going to do with it?

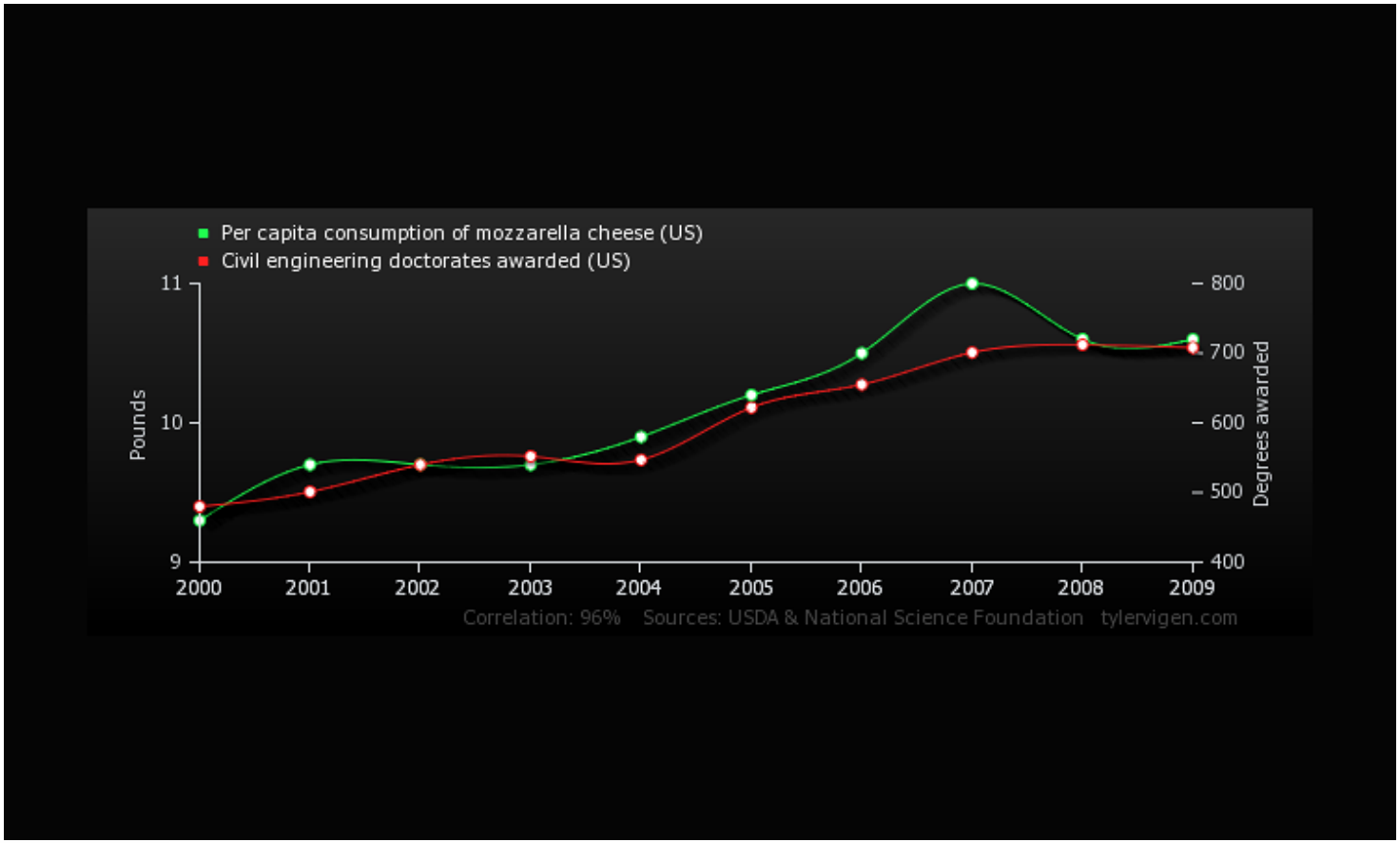
A Word of Caution

- This is a heat map of rental rates in Boston (source: wbur.org, data from 2013)
- Starts to make sense when overlayed with rail map
- Note that not all rail lines affect rental rates
- It could be that Cambridge is just more desirable anyway



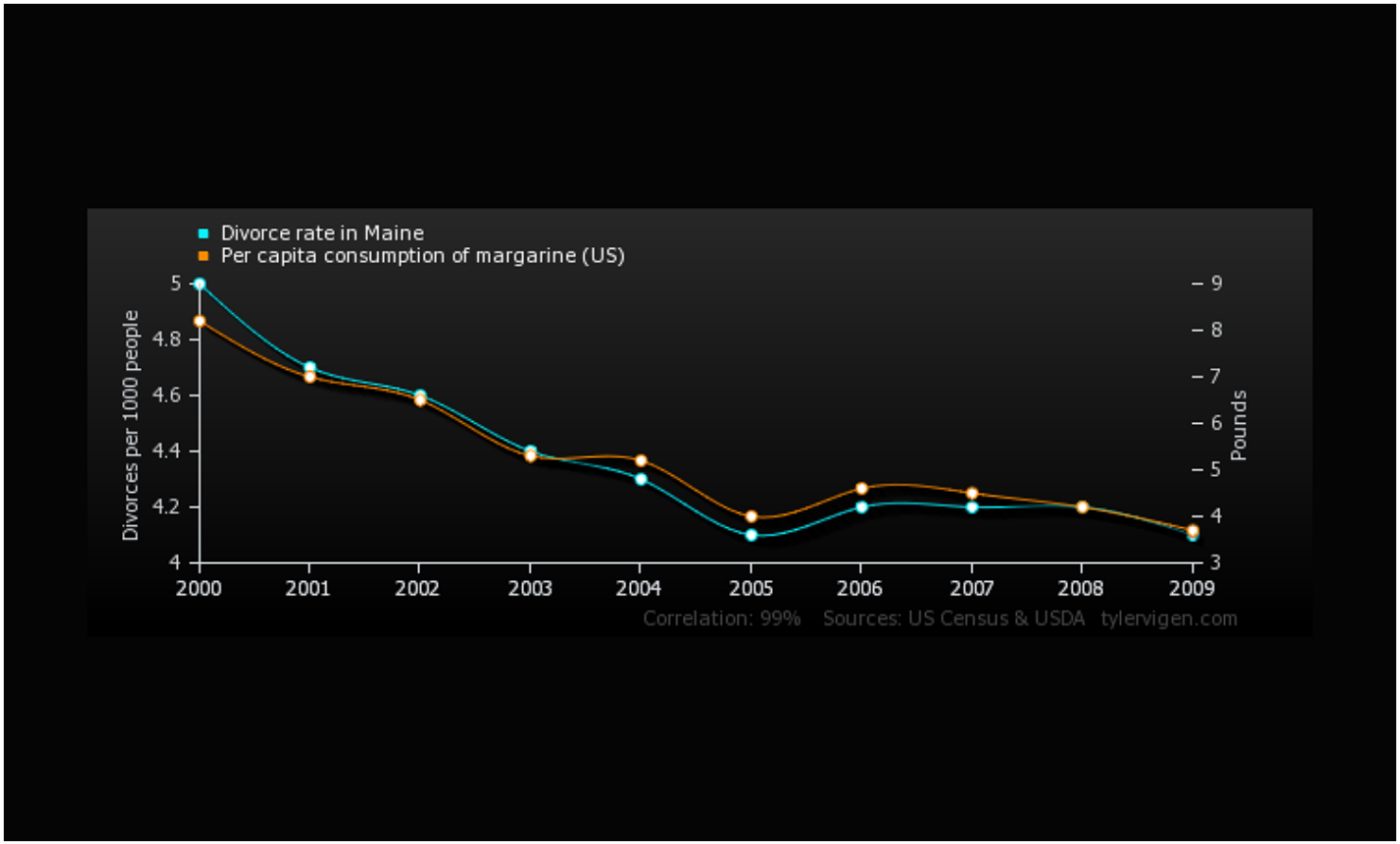
A Silly Example

- Correlation doesn't equal causation…



Example 2

- More of what not to do – BTW, these are from FastCompany



Questions?

- Thanks, and watch out for snakes!

Mike Grosvenor

- Consulting Manager, Access Sciences Corp.
  - Business Intelligence
  - Content Management
  - Data Governance
  - Regulatory Compliance
- Contact Info:
  - Email: mgrosvenor@accessscience s.com
  - Phone: 713.664.4357 (office) / 713.715.8154 (mobile)
- Linked In: http://www.linkedin.com/in/mikegrosvenor
- Twitter: @gro7or