Orbitz Worldwide is a global travel leader

- Orbitz Worldwide is a leading global online travel agency (OTA)
- Over **15 million** unique visitors per month
- Headquartered in **Chicago, IL** with major offices in the UK, France, Switzerland, Sweden, Australia and Washington DC. In addition to these locations, we also have teams in **Sunnyvale, CA**, Denver, Israel, Argentina and India.

The Goal Today

1. Provide direction on key skill sets required to enhance marketability for Big Data Analysts (conversely – this would apply for firms looking to build capabilities in Advanced Analytics)

2. Hopefully you leave feeling good/energized about being in the Analytics space! ☺

Who Moved My Cheese?**

WAKE UP!

- My Background → leading teams of "**SAS/SQL**/RDBMS" **statistical modelers**
- @ Orbitz → also oversaw Chief Data Scientist and the **Machine Learning** team
- SAS folks looking to leverage Hadoop data: that was a wake up call
- May/2012 → Prompted me to write an article on the newfound "who moved my cheese" syndrome ("***The times they are a changin' for advanced analytics***')

http://www.analytics-magazine.org/may-june-2012/572-executive-edge-the-times-they-are-a-changin-for-advanced-analytics

http://www.predictiveanalyticsworld.com/patimes/august12/ (Predictive Analytics Times – Aug 2012)***

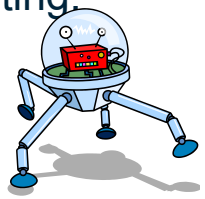
- Notes:
  1. Still early in this journey of Big Data Analytics!***
  2. References to links & firms – unsolicited and my views only – not an endorsement, just some examples or players in the space.

Machine Learning vs. Statistical Modeling

Some observations while managing "both sides of the fence":

- Much in common…
  1. Both are concerned with "data mining"
  2. Many common methods/algorithms (eg: Decision Trees, Clustering)
  3. Different jargon for similar concepts (eg: *weights/parameters, learning/fitting*)

- However, many differences in approaches/tools are stark…
  1. Support Vector Machines (SVM) is often the go-to tool for machine learners***
  2. ML focuses on predictive accuracy over interpretation of models*** (statistical modeling has a probabilistic approach with a strong emphasis on parametric assumptions, regression diagnostics, inference, hypothesis testing, interpretability of model etc..)
  3. ML folks tended to have more computer science in their DNA:



  - Often have backgrounds in CS (ML has its roots in AI)
  - More comfortable with **engineering tasks & coding/programming** (eg: Java/Python)
  - Some were familiar with **MapReduce programming**
  - ML practitioners tend to use **R** and **embrace open source**. The Big Data world is open-source dominated.

Machine Learning vs. Statistical Modeling (continued)

- Machine-Learning lends itself very well to the road ahead of Big Data…. the Big Data paradigm shift, along with open source tools, is ideally suited for ML to leverage.

- Statistical modeling is not going away -- but machine learning is rapidly increasing in relevance and prominence. It makes sense for analytical teams to complement their skill sets by incorporating machine-learning approaches in order to be better positioned for the road ahead.

- General Interest in ML has exploded recently! (eg: Stanford AI/ML course)

- What does this mean for the traditional statistical modelers?
  - **Just SAS/SQL will limit your effectiveness and opportunities** ahead **–** take a page from the ML world and transition towards the role of **Data Scientist**
  - **Cannot be complacent -- learn new tools by embracing Big Data technologies** (Hadoop & other open source tools). Despite hype around Big Data, it is generally acknowledged that Big Data and distributed computing are rapidly changing the Analytics landscape….this is not a passing fad![*]

Big Data MAGIC! ☺

- Misconception: underlying Big Data technologies will magically solve all our business problems!
- Little will change unless Analytics provides that step function lift: i.e. leverage Big Data to generate actionable insights -- ideally holistic insights -- ***enable business teams to make decisions differently*** than we did prior to the world of Big Data. **This is why YOU – the Data Scientist – are more important than ever before!**
- McKinsey Global Institute Report: "**There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions**"
- Big Data initiatives need more visible 'success stories' with clearly demonstrated ROI

"Data Scientist" : Some Key Skills for Big Data Analytics

- Big Data 3 V's as originally defined by Doug Laney/Gartner:
  - **Volume** → increase in data volume [Terabytes…Petabytes…Exabytes…Zettabytes]
  - **Velocity** → how fast data is being produced & how fast the data must be processed
  - **Variety** → the different formats of data; 80%+ is textual/semi-structured/unstructured

Source: http://api.ning.com

"Data Scientist" : Some Key Skills for Big Data Analytics

- People jumping on the "V" bandwagon: **V**alue, **V**ariability, **V**…?

- "Big" Data Misconception: Volume is NOT the biggest challenge: it is **Variety**

(eg: Voice, Video, Audio, Sensor, Social Media, Tweets, Blogs, Click-Stream, Transactional, Stock Market, Emails, RFID, machine-to-machine data, etc..)

Untapped Opportunity → Mining Unstructured Data

❖ **Key Skill** → **Text Mining/Analysis**

Data Scientist Key Skills (continued)

Reference: CHAPTER 2 of *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications* (http://www.textanalyticsworld.com/wp-content/uploads/2012/03/PracticalTextMining_Excerpt.pdf)

1. **Search and information retrieval (IR):** Storage and retrieval of text documents, including search engines and keyword search.
2. **Document clustering**: Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
3. **Document classification**: Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
4. **Web mining**: Data and text mining on the Internet, with a specific focus on the scale of interconnectedness of the web.
5. **Information extraction (IE):** Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semi-structured text.
6. **Natural language processing (NLP):** Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics.
7. **Concept extraction**: Grouping of words and phrases into semantically similar groups.

Data Scientist Key Skills (continued)

❖ **Key Skill →** **Python**

- Learn Python – scripting language -- will go a long way towards effective text analysis -- data cleansing/processing
- Python is intuitive, free (open-source) and easy to learn language (easier than R)
- http://www.udacity.com/ (free courses; Python taught in *Intro to CS* class)
- UCSC Silicon Valley Extension: http://course.ucsc-extension.edu/modules/

❖ **Key Skill →** **Engineering/Coding/Data Munging**

- Coding/hacking ability is a key skill for Big Data Analytics
- Rolling up one's sleeves and being scrappy/nimble -- being hands on
- "**Learn how to fish**": do not be limited by functionality of some templatized/ high-level GUI of a data mining tool – ability to code your algorithms and be flexible with data processing puts you well ahead **[data cleaning/prep/ validation**]
- I like this quote I came across → "A fool with a tool is still a fool" ☺

Data Scientist Key Skills (continued)

❖ **Key Skill →** **Statistical Chops**

- A great (vs. good) data scientist will straddle both sides of the fence well: predictive modeling/data mining & software engineering/coding

Data Scientist Key Skills (continued)

❖ **Key Skill →** **"Story Telling" & Communication**

- Good communication and effectively interacting with other teams have always been necessary.

- However, visualization of insights is increasingly important with Big Data – partly because of the importance & popularity of social media (eg: graph analysis for visual representation of social networks) -- so "telling the story" well is key

- Expect to see regular advances in visualization offerings. Popular/leading products in this space currently include:
  - QlikView (QlikTech)
  - Tableau
  - Spotfire (Tibco)
  - JMP (SAS)
  - SAS Visual Analytics seems really cool – I saw a demo recently
  - R packages (eg: "ggplot2" for graphics/visualization)

Data Scientist Key Skills (continued)

❖ **Key Skill →** **Curiosity & Ongoing Learning**

- My best folks have been those who do not rush into model-building but spend time exploring the data -- they enjoy diving into raw data and understanding it well first
- Cannot emphasize this skill enough: *A thirst for data exploration* makes the difference between a good & great analyst

- Push yourself to constantly learn! Stay on top of happenings in the Big Data world:
  - Online education is a terrific source!
  - Join Meetups in Big Data (http://www.meetup.com/BigDataCloud/ (ii) http://www.hivedata.com/)
  - LinkedIn groups on Big Data & Analytics
  - Kaggle competitions
  - Blogs on Data Science & Big Data
  - Finger on pulse of latest tools/methods/technologies for Big Data Analytics

Data Scientist Key Skills (continued)

❖ **Key Skill →** **Learn R!**

- R is now the data mining tool of choice! (eg: Kaggle winners have used R)
- Latest KDnuggets poll:
  http://www.kdnuggets.com/2012/08/poll-analytics-data-mining-programming-languages.html
  - Majority of **data miners prefer to use R**
  - **R & Python most popular** programming languages
  - **Python -- highest growth rate** in 2012
  - **Most popular language used along with R was Python** (and vice versa).

- R has very good **integration with Hadoop**, an area where established commercial statistical tools have been playing catch-up over the past year.
- If you want to **work startups**: many startups and smaller firms do not have deep pockets and are embracing open source tools such as the R programming language
- R is a leading language for developing new statistical methods -- **it is a platform for statistical innovation and collaboration** across both the corporate world and academia -- and has a large & vibrant user community
- In my opinion, the stronghold of established commercial players seems to be potentially threatened -- **open source tools like R are better suited for Big Data** and will slowly but surely continue to take share away from commercialized statistical packages

Data Scientist Key Skills (continued)

*** Key Skill →** **Learn R!** (continued)

- Traditional statistical vendors have recognized that R is a force to be reckoned with. Many of these vendors have developed hooks into R so users can interface with the R language
- Based on the resumes I've been reading, the next generation of data miners is clearly flocking to R as their go-to tool. Professors in general are comfortable with R; they tend to often have students use R as part of their curriculum.
- Open-source analytics tools and platforms have arrived!

NOTE: R has often not been widely adopted as a standard in the corporate world because of concerns of not being "enterprise ready" -- but even that is changing as firms such as **Revolution Analytics** focus on the enterprise capabilities for R – and as adoption of R continues at its rapid pace.

Just as firms like Cloudera & Hortonworks work on "Hadoop for the Enterprise", expect to see firms like Revolution Analytics make good strides on "R for the Enterprise"

http://www.revolutionanalytics.com/products/r-for-apache-hadoop.php

http://www.revolutionanalytics.com/products/enterprise-deployment.php

Data Scientist Must-Have Skills (continued)

- **Key Skill →**

Learn R!

➢ FYI: Some R resources sources for those new to R:

- Download R: http://www.r-project.org/
- Recommended to download **R Studio**: http://www.rstudio.com/ide/download/
- Free R **webinars** on the Internet (eg: "Intro to *R for SAS and SPSS users*")
- **Books** on R (A popular one for users coming from SAS is "*R for SAS & SPSS users*"): http://www.r-project.org/doc/bib/R-books.html
- **YouTube** Videos on R (eg: http://www.youtube.com/user/Tutorlol)
- R **Manuals**: http://cran.r-project.org/manuals.html
- Subscribe to **R Blogs** (http://www.r-bloggers.com/ is a hub that aggregates across blogs)
- Available R packages to use: http://cran.r-project.org/web/packages/
- CRAN Task Views: Browse packages by Topic (http://cran.r-project.org/web/views/)

[Courtesy Jigsaw Academy Blog for some helpful references]

Some Thoughts Regarding SAS & Big Data Analytics

The *Little SAS Book* has started the careers of many in this field***

- ✓ SAS is not going anywhere soon – it is the Big Gorilla in Analytics (90%+ Fortune 500 use SAS) – and will continue to be a leader
- ➢ **Most Analytics jobs openings look for SAS skills**. Learn SAS! ☺ ***
- ➢ SAS is generally the best way to get your foot in the Analytics door
- ➢ **Recently introduced SAS Access Engine for Hadoop*****
- ➢ Introduced **High Performance Analytics (HPA)** offering with impressive potential:
  - Step function improvement in processing times
  - Supports modeling **against complete data sets, eliminating the need for sampling** and ensuring more accurate predictive models
  - SAS Visual Analytics seems very impressive

- ➢ My opinion -- 2 areas where SAS can play its hand differently:
  1. **Pricing** (especially for small firms)
  2. **Pushing SAS in university courses** -- subsidize/free software – else the next generation of data scientists will not champion SAS in their journey/career

Online Offerings for Data Science & Advanced Analytics

- Online Education is a booming area!
  - "***Online Education Degrees Now Dwarf Traditional Universities":*** **http://techcrunch.com/2012/08/09/online-education-degrees-now-dwarf-traditional-universities/**

- **Khan Academy** (http://www.khanacademy.org/) led the way and inspired a **revolution**

- Some good sites for learning Python/R/Statistics/Machine Learning
  - www.coursera.com (ML by Andrew Ng, 'Intro to Data Science', etc.)
  - www.udacity.com → ("CS 101" – Python used to code a search engine/web crawler)
  - www.statistics.com (courses in R, Statistics, Data Prep, etc.)

- **edX**: Web portal (www.edx.org) launched in May 2012 by **Harvard** & **MIT** with $60 million in funding from the two schools. The University of California, **Berkeley** has started making its online courses available on edX.

http://finance.yahoo.com/news/elite-colleges-transform-online-higher-124855202.html

Graduate Programs in Data Science & Analytics

- **Note: Columbia University** has put together its first course with "data science" in the title. In July 2012, the school launched the *Institute for Data Sciences & Engineering*" (also using primarily R & Python)
- Graduate Programs in Big Data Analytics & Data Science: http://whatsthebigdata.com/2012/08/09/graduate-programs-in-big-data-and-data-science/
- Big Data Programs: 6 Big Name Schools

http://www.datanami.com/datanami/2012-04-10/six_big_name_schools_with_big_data_programs.html

Some examples:

- **Northwestern University:** *M.S. in Analytics* (also, *M.S. in Predictive Analytics* -- online)
- **Stanford University**: *Graduate Certificate in Data Mining Massive Datasets*
- **North Carolina State University**: *M.S. in Analytics*
- **DePaul University**: *M.S. in Predictive Analytics*
- **University of California, Berkeley**: *Intro to Data Science* course
- **University of California, San Diego** UCSD): Data Mining certificate
- **Carnegie Mellon Univ**: *MIS Management degree with a BI & Data Analytics concentration*
- **Bentley University**: *M.S. in Marketing Analytics*
- **Syracuse University**: Graduate *Certificate of Advanced Studies in Data Science*
- **University of San Francisco**: *M.S. in Analytics*
- **Stevens Institute of Tech**: *M.S. in Business Intelligence and Analytics*

Some Trends in Big Data Analytics

- **Unstructured-Data explosion** -- text processing/mining & data manipulation
- **R adoption** at a growing rate -- as well as the move to make R 'Enterprise Ready'
- Increasing maturity & **adoption of open source;** Big Data world is open-source dominated (R, Hadoop, MapReduce, Python, NoSQL databases, etc.)
- Elevating Advanced Analytics in organization: **Chief Analytics Office** (CAO) role will become more common
- High-Quality (& often free) **online analytics education ("*Massive Open Online Courses*" – *MOOCs*)** as well as offline university programs in Data Science/Analytics
- **Visualization** of Insights

Some Trends in Big Data Analytics (continued)

- **Data Scientists** in great demand:
  - Hands-on applied math/stats/machine learning
  - Computer science/software engineering/coding
  - Business Acumen/ Common Sense
  - Curiosity & a thirst for data exploration: finding needles in a haystack
  - Communication ~ telling a story effectively
  - Data Scientists increasingly found in Product Development groups (vs. Marketing say)
  - Expect trend of more "*Data Scientist*" business cards and job titles

- **Analytics Service/Solutions firms** -- well positioned to capitalize on Analytics demand

- Firms providing an **abstraction layer for self-service Analysis of Hadoop** data
  - Platfora (www.platfora.com)
  - Datameer (http://www.datameer.com/)
  - Karmasphere (https://karmasphere.com/)
  - Hadapt (http://hadapt.com/)
  - ClearStory Data (http://www.clearstorydata.com/)

- Other Trends: Streaming/Real Time Analytics, PMML, Cloud (eg: Redshift), etc.

Job Trends in Big Data Analytics
"Data Science" most in demand

- Job Trends from Indeed.com as of September 2012 (source: www.analyticbridge.com)

Concluding Thoughts

- Hal Varian, Google's Chief Economist: *"The sexy job in the next ten years will be <statisticians/data scientists>… The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill."* ***

- It has never been cooler to be a nerd! ☺

- A BIG shortage of talent for Big Data Analytics -- **extremely** tight labor pool
  - VC firms' recruiting teams for their portfolio companies
  - EMC/Greenplum hiring own data scientists – training & certification in Data Science & Big Data Analytics

- **TAKEAWAY**: Change your title to "Data Scientist" and get a raise!

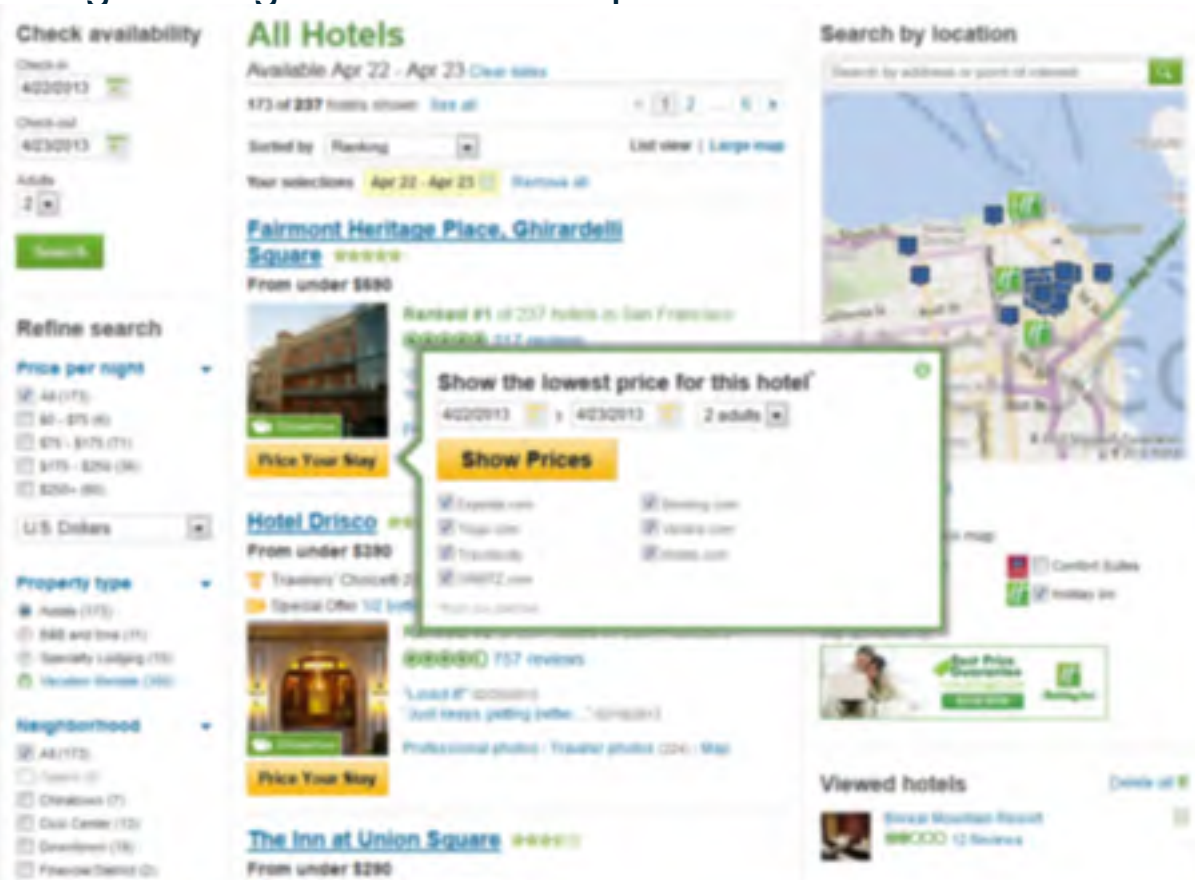
***

Orbitz Use Case: Predictive Model incorporating Text Mining

Improving a Revenue-Per-Click (RPC) Model by Mining Textual User Reviews of Hotels

Case study: Trip Advisor Predictive Model Using Text Mining

- Business problem: calculate Revenue Per Click (RPC)
  - ➢ Optimizing bidding on hotels at Tripadvisor.com



Predictive Modeling Approach for Bid Optimization

- Predictive modeling approach
  - Building Revenue Per Click (RPC) model to calculate the click value at www.tripadvisor.com for all hotels individually
  - RPC model enables smart bidding

Illustrative data:

	Flat CPC	Modeled RPC	What to do?
Hotel A	$ 0.50	$ 0.20	Reduce CPC
Hotel B	$ 0.50	$ 0.75	Increase CPC

RPC Model Variables

- RPC model input variables
  - User review (text)
  - Hotel properties including location, brand, chain, star-rating, demographic, and referral source etc.

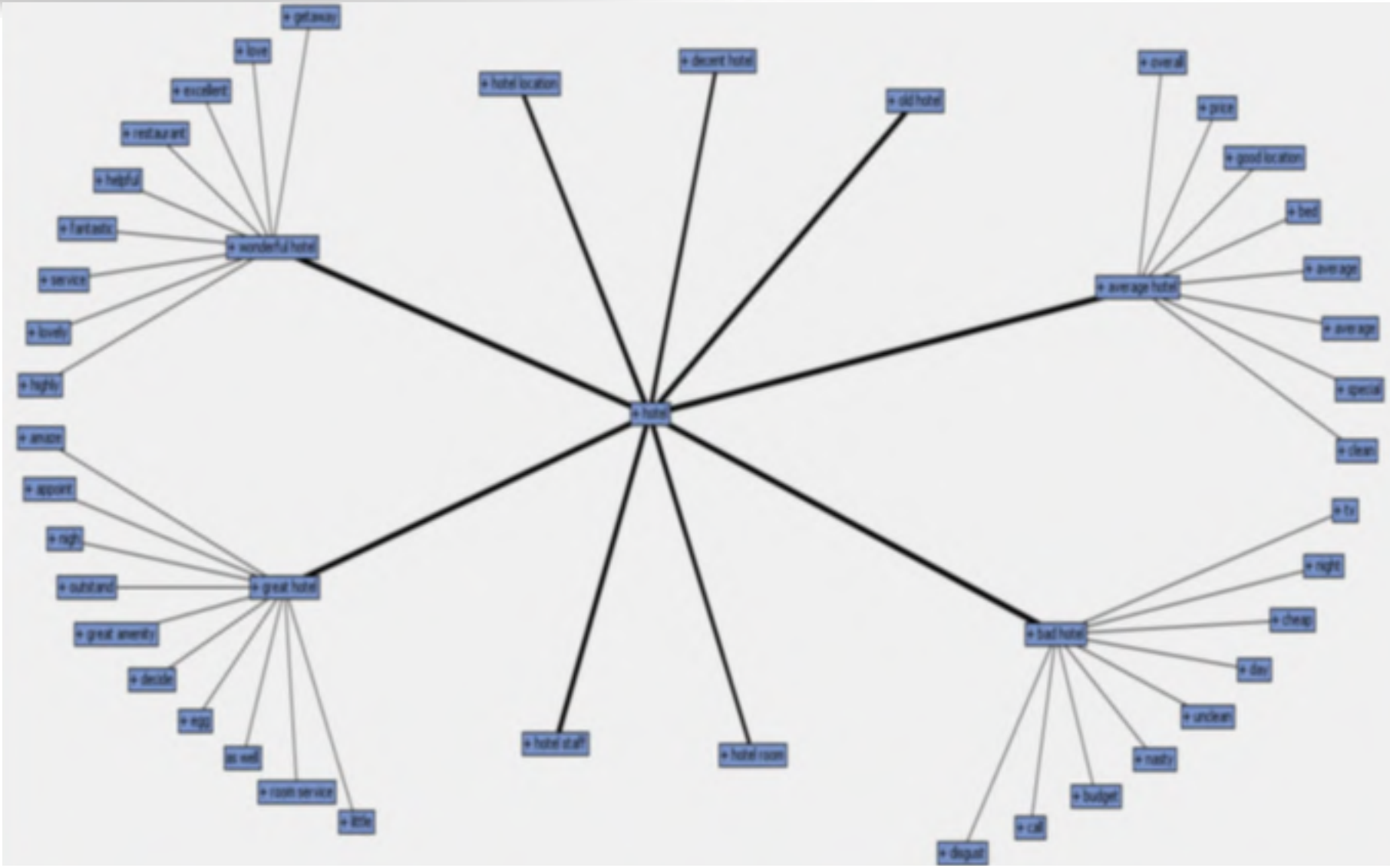
- Review score vs. review text*
  - Structured vs. non-structured
  - consumers relied much more on text review than on review score
  - Review score is often inflated, often not correlated well to the text review

* “Social Media and Lodging Performance” Kelly A. McGuire at SAS in (http://blogs.sas.com/content/hospitality/)

Sample Hotel Reviews at Orbitz.com

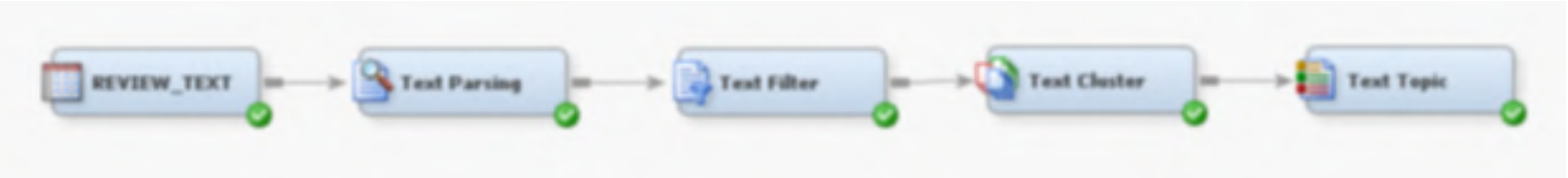
- Positive review
  - "We will definitely stay here again. Staff was incredibly friendly and helpful. The room was exactly what we needed! Total atmosphere was surreal! Great around the holidays! …"
- Neutral review
  - "Hotel is located on Van Ness and Geary...not a great place in the city, but easy to get to everything else … "
- Negative review
  - "I found the hotel decor in need of updating. My room was generally clean with the decor (carpet, furniture, bed) in need of updating. It also had an unpleasant odor … "
  - "I was thiking to surprise to my wife, but was a very bad experience. The neiborhood is very dangerous. The parking in the hotel is very expensive, they never told us about it…"

Concept Linking for Hotel



Text Mining Tool in SAS Enterprise Miner

- Text mining process



- Text Parsing – Start/Stop list, Stemming, Synonym, Multi-Term list
- Text Filter – retain only relevant and valuable info
  - Term filter and document filter
  - Term weighting and frequency weighting
  - Concept linking
- Text Cluster – put documents into disjoint clusters
- Text Topic – assigns each document to topics

Singular Value Decomposition (SVD)

- Weighted Term-Document Frequency Matrix

$$A = U_{t\times t}\Sigma_{t\times d}V_{d\times d}$$

where matrices U and V have orthonormal columns, and Σ is a diagonal matrix of singular values

- Low Rank Approximation

$$A \approx U\ \mathrm{diag}(\sigma_1,\ldots,\sigma_k,0,\ldots,0)V$$

Model performance comparison

- Improved efficiency by 40% using review text variables!