The analytics landscape:
A personal view

Charles Elkan
December 20, 2011
What is analytics?

- Big data, business intelligence (BI), decision support (DSS), data warehousing, unstructured data, knowledge discovery in databases (KDD), information visualization, map-reduce.

- analytics = convert data into intelligence + capture value
  = statistics + optimization

- statistics = machine learning = data mining

- optimization = microeconomics + operations research
Outline

1. Structured data (predictive, visual)
2. Unstructured data
3. The business of analytics
4. A research *and* business opportunity
A basic distinction

I. Structured data
   Tables in databases
   Nodes and links in networks

II. Unstructured data
   Text
   Videos
   Tables in web pages
   XML
I. Structured data

• A data warehouse is a cost center, not a profit center.
• How can structured data be a profit center?
  1. Predictive analytics
  2. Visual analytics
1. Predictive analytics

- So, what can we do with structured data?
- Answer: Make predictions, then take actions.
- Example:

  **Improved Disk-Drive Failure Warnings**
  
  Gordon F. Hughes, Fellow, IEEE, Joseph F. Murray, Kenneth Kreutz-Delgado, Senior Member, IEEE, and Charles Elkan

- But, what are the costs and benefits of alternative actions?
- And, who pays which costs?
Cost-sensitive learning

- Cross-domain theory of making optimal decisions given predictions:

  ![Google Scholar search results for cost-sensitive learning](Google_scholar_results.png)

  [PDF] The foundations of cost-sensitive learning
  C Elkan - International Joint Conference on Artificial Intelligence, 2001 - Citeseer
  This paper revisits the problem of optimal learning and decision-making when different misclassification errors incur different penalties. We characterize precisely but intuitively when a cost matrix is reasonable, and we show how to avoid the mistake of defining a cost...
2. Visual analytics

- So, what can we do with structured data?
- Answer: Find and display patterns; prompt human insight.
Patterns of human metabolism
Information visualization

• “state of the art analytic tools to identify biomarkers”
II. Unstructured data
A case study

CHALLENGES:
• Facing a highly publicized global recall, Toyota needed a way to understand its quality data – yet had an exponentially growing number of questions and a fraction of time to react

RESULTS:
• Gave close to 1,000 users, from quality engineers to executive dashboard users the ability to analyze quality data from heterogeneous sources
• Allowed users to design reports and dashboards in minutes
• Delivered analytics on 6 years of structured and unstructured data from more than a dozen sources with 110 analytical dimensions, and 250 analytical components
• Will eliminate hundreds of thousands of hours of end-user wait time per year
LaVerne Council, CIO of Johnson & Johnson:

“... allow anyone to ask a question ... folks that have given us access to their email ... data mining for answers to that question ... help us solve a very hairy issue for one of our products ... one of the associates had completed his thesis in college on that very topic ... they weren’t in the same company ... we were able to really come back with answers.”
A grand vision

- “Open source intelligence (OSI)”
Using Software to Sift Digital Records

By NATHAN KOPPEL

“The biggest pain point in litigation is the amount of money spent on attorneys reviewing documents,” said Jonathan Redgrave, a Washington, D.C., lawyer ...

... used so-called predictive coding software made by Recommind Inc. to respond to a government investigation ...

Attorneys at the firm started by reviewing a relatively small set of records to identify the important characteristics. ...

That knowledge was then coded into software, which was used to scan a larger universe of electronic records. Attorneys then reviewed the most relevant records to make the final determination about whether they should be disclosed.

... Recommind licenses its software for costs ranging from about $650 per gigabyte of data analyzed to several million dollars annually for unlimited data.
III. The business of analytics

- Analytics applications are valuable.

**Heritage Provider Network Announces the Official Launch of the $3 million Dollar Heritage Health Prize**

MARINA DEL REY, Calif., April 4, 2011 /PRNewswire/ -- Dr. Richard Merkin, President and CEO of Heritage Provider Network, announced today the official launch of the $3 million dollar Heritage Health Prize, the world's largest predictive modeling contest. ... Heritage Health Prize Advisory Board members include Arvind Narayanan, Stanford University; Charles Elkan, UC San Diego and Netflix prize judge and Claudia Perlich, MediaSixDegrees, winner and organizer of several data mining competitions.

For more information on the HPN Health Prize go to: [www.heritagehealthprize.com](http://www.heritagehealthprize.com)
Analytics **companies** are valuable

**Oracle Endeca Deal Echoes HP's Autonomy Purchase**

Oracle plans to use Endeca’s technology to boost unstructured data access and analysis for Web commerce and business intelligence applications.

By **Doug Henschen** *InformationWeek*
October 18, 2011 03:50 PM

Oracle announced Tuesday that it plans to acquire Endeca Technologies in a deal that will help it bring unstructured data into e-commerce transactions and business intelligence analyses. It's no coincidence that the deal was announced just weeks after Hewlett-Packard finalized its $10-billion-plus acquisition of Autonomy.
Are valuations bubble-icious?

- HP compared to Autonomy:
  
  Sales: $128B versus $963M
  Income: $12B versus $343M
  Value: $50B versus $11B

- Forrester: “The Autonomy IP is stagnant. There hasn’t been a major release in five years.”

- Zero recent patents for the core analytics.
IV. A research and market opportunity

What Is Autonomy, Without Its Marketing?
Posted by Leslie Owens on August 26, 2011

AUTONOMY ARCHITECTURE OVERVIEW

At the heart of Autonomy’s stack is the Intelligent Data Operating Layer (IDOL) — its brand name for search and content processing technology.

... The IDOL IP is stagnant. There hasn’t been a major release of IDOL in over 5 years.
Disruption from below

• New platform for diverse data
  Cloud-based
  Multiply the user base 10x:
  • Easy to use
  • Fun to use

• Opportunity: Add “secret sauce” to open-source software
  Newer artificial intelligence
  Patented artificial intelligence
• A role model for cloud-based ease of use: Box.net
• $650M valuation, but no intelligence.
Disruption from below

- Cloud-based software as a service (SaaS)
- Easy to use, fun to use
- Newer AI, patented AI

- Open-source foundation:
  Lucene and Solr as backend
  Tika for importing unstructured data
Newer artificial intelligence

- Sentiment analysis
- Topic models for organizing content
- Recursive neural nets for deep understanding

www.socher.org/index.php/Main/ParsingNaturalScenesAndNaturalLanguageWithRecursiveNeuralNetworks
Newer AI: Fewer topics, better fit

(a) LDA

(b) DCMLDA

Figure 2. Mean per-document log-likelihood on the S&P500 dataset for DCMLDA and fitted LDA models.
Patented AI: Sentiment analysis

- ... labels designate level of quality, such as interestingness, appropriateness, timeliness, humor, style of language, obscenity, sentiment

- ... a classifier means effective to automatically associate a quality value to items of data, wherein said quality value is indicative of the qualitative nature of said items of data
Today in the New York Times

Apple Wins Partial Victory on Patent Claim
By NICK WINGFIELD 12 minutes ago
The ruling by a U.S. agency, involving a set of important smartphone features, could force changes in Google's Android phones, including HTC's, above.
SQUID

- Sentiment analysis
- Question answering
- Unstructured data organization
- Interactive insight
- Diverse entity extraction

- But what will be most beneficial and profitable?

- Historical answer: Specific vertical applications.
Profit lies in verticals, I...
Financial groups hit by flood of new rules

By Brooke Masters in London

Financial services firms worldwide are being hit with an average 60 regulatory changes every working day, a 16 per cent increase over last year, and no let up is in sight, a study has found.

Regulators around the world announced 14,215 changes in the twelve months to November, up from 12,179 for the same period a year earlier, according to new research by the Thomson Reuters governance, risk and compliance unit.

The study tracks everything from the passage of new laws and short-selling bans to the issuance of consultation papers and speeches that contain policy announcements; in short everything compliance officers are expected to keep abreast of. The rules range from global packages like the Basel III bank capital reforms to local rules in individual US states.
Discussion

• Acknowledgement: Most images are due to other authors.