Data Scientist: The Sexiest Job of the 21st Century
by Thomas H. Davenport and D.J. Patil

Back in the 1990s, computer engineer and Wall Street “quant” were the hot occupations in business. Today data scientists are the hires firms are competing to make. As companies wrestle with unprecedented volumes and types of information, demand for these experts has raced well ahead of supply. Indeed, Greylock Partners, the VC firm that backed Facebook and LinkedIn, is so worried about the shortage of data scientists that it has a recruiting team dedicated to channeling them to the businesses in its portfolio.

Data scientists are the key to realizing the opportunities presented by big data. They bring structure to it, find compelling patterns in it, and advise executives on the implications for products, processes, and decisions. They find the story buried in the data and communicate it. And they don’t just deliver reports: They get at the questions at the heart of problems and devise creative approaches to them. One data scientist who was studying a fraud problem, for example, realized it was analogous to a type of DNA sequencing problem. Bringing those disparate worlds together, he crafted a solution that dramatically reduced fraud losses.
Data Mining

• A set of principles, concepts, and techniques that structure thinking and analysis of data

• Extracts useful information and knowledge from large volumes of data by following a process with reasonably well defined steps

• Changes the way you think about data and its role in business
Data Opportunities

- Volume of data
- Variety of data
- Powerful computers
- Better algorithms
Learning Goals

• Approach business problems data-analytically

• Interact competently on the topic of data mining for business intelligence

• Hands-on experience mining data
Roles in Data Mining

• “Data Scientist” (Geek?)
  ▪ can do the actual modeling
  ▪ applied statistician × computer scientist

• Collaborator in a data-centric project
  ▪ can translate from business to the execution

• Managing a data-mining project
  ▪ understanding the potential
  ▪ ability to evaluate a proposal and execution
  ▪ ability to interface with a broad variety of people

• Strategist, Investor, …
  ▪ envision opportunities, come up with novel ideas, evaluate the promise of new ideas, design data science projects / companies conceptually
Business data mining is a process

science + craft + creativity + common sense

a process
Data Mining Process

- Business understanding
- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment
Outline

• Business Understanding
• Data Understanding
• Data Preparation
• Modeling
• Evaluation
• Deployment
This is NOT a course about…

• Statistics

• Database Querying
  • SQL

• Data Warehousing

• Regression Analysis
  • Explanatory vs Predictive modeling
Data Mining versus...

- Data Warehousing / Storage
  - Data warehouses coalesce data from across an enterprise, often from multiple transaction-processing systems

- Querying / Reporting (SQL, Excel, QBE, other GUI-based querying)
  - Very flexible interface to ask factual questions about data
  - No modeling or sophisticated pattern finding
  - Most of the cool visualizations

- OLAP – On-line Analytical Processing
  - OLAP provides easy-to-use GUI to explore large data collections
  - Exploration is manual; no modeling
  - Dimensions of analysis preprogrammed into OLAP system
Data Mining versus...

• Traditional statistical analysis
  • Mainly based on hypothesis testing or estimation / quantification of uncertainty
  • Should be used to follow-up on data mining’s hypothesis generation

• Automated statistical modeling (e.g., advanced regression)
  • This is data mining, one type – usually based on linear models
  • Massive databases allow non-linear alternatives
Answering business questions with these techniques.

- Who are the most profitable customers?
  - Database querying

- Is there really a difference between profitable customers and the average customer?
  - Statistical hypothesis testing

- But who really are these customers? Can I characterize them?
  - OLAP (manual search), Data mining (automated pattern finding)

- Will some particular new customer be profitable? How much revenue should I expect this customer to generate?
  - Data mining (predictive modeling)
Thanks!
Questions?