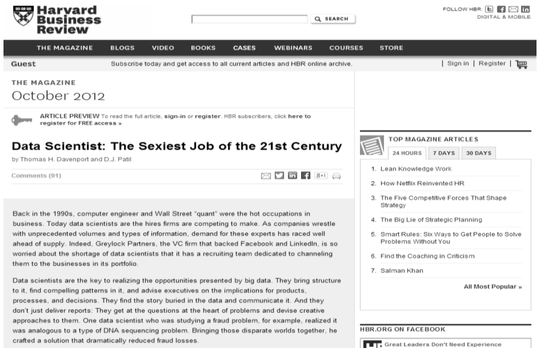
Data Mining for Business Analytics

Lecture 1: Introduction to Data Mining

Stern School of Business
New York University
Spring 2019

The Good



Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE
October 2012

ARTICLE PREVIEW To read the full article, **sign-in** or **register**. HBR subscribers, click **here to register for FREE access »**

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

Comments (91)

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.

TOP MAGAZINE ARTICLES

24 HOURS | 7 DAYS | 30 DAYS

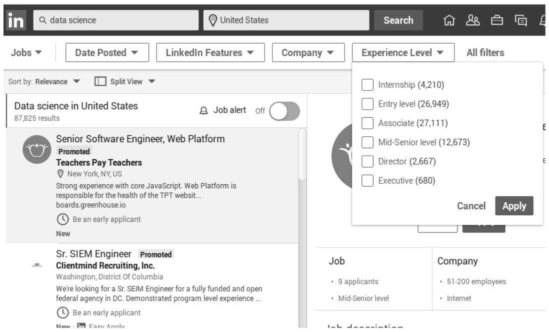
1. Lean Knowledge Work
2. How Netflix Reinvented HR
3. The Five Competitive Forces That Shape Strategy
4. The Big Lie of Strategic Planning
5. Smart Rules: Six Ways to Get People to Solve Problems Without You
6. Find the Coaching in Criticism
7. Salman Khan

All Most Popular »

HBR.ORG ON FACEBOOK

Great Leaders Don't Need Experience

The Good



data science | United States

Sort by: Relevance

☐ Internship (4,210)
☐ Entry level (26,949)
☐ Associate (27,111)
☐ Mid-Senior level (12,673)
☐ Director (2,667)
☐ Executive (680)

Data science in United States
87,825 results

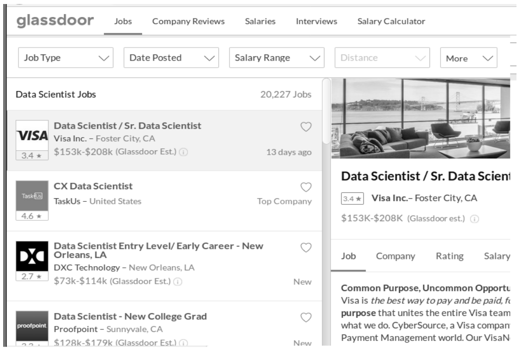
Senior Software Engineer, Web Platform
Promoted
Teachers Pay Teachers
New York, NY, US
Strong experience with core JavaScript. Web Platform is responsible for the health of the TPT websit...
boards.greenhouse.io
Be an early applicant
New

Sr. SIEM Engineer Promoted
Clientmind Recruiting, Inc.
Washington, District Of Columbia
We're looking for a Sr. SEIM Engineer for a fully funded and open federal agency in DC. Demonstrated program level experience ...
Be an early applicant
New · Easy Apply

Job
- 9 applicants
- Mid-Senior level

Company
- 51-200 employees
- Internet

The Good



Data Scientist Jobs 20,227 Jobs

Data Scientist / Sr. Data Scientist
Visa Inc. – Foster City, CA
$153k-$208k (Glassdoor Est.)
3.4 ★ 13 days ago

CX Data Scientist
TaskUs – United States
4.6 ★ Top Company

Data Scientist Entry Level/ Early Career - New Orleans, LA
DXC Technology – New Orleans, LA
$73k-$114k (Glassdoor Est.)
2.7 ★ New

Data Scientist - New College Grad
Proofpoint – Sunnyvale, CA
$128k-$179k (Glassdoor Est.)
New

Data Scientist / Sr. Data Scien
3.4 ★ **Visa Inc.**– Foster City, CA
$153K-$208K (Glassdoor est.)

Common Purpose, Uncommon Opportu
Visa is *the best way to pay and be paid, f*
purpose that unites the entire Visa team
what we do. CyberSource, a Visa compan
Payment Management world. Our VisaN

The Good

- Data Mining is pervasive

The "Bad"

- No Free Lunch

The “Bad”

- Effective Data Science requires / builds on a ***SET*** of skills:
  - Analytical thinking
  - Technical skills
  - Creativity
  - Communication
  - Domain Knowledge **(!)**

The “Ugly”

- We will be doing some math

The “Ugly”

- We will be doing some math
- We will be doing some Programming
  - Highly sought-after skill !
  - **BUT REMEMBER:**
    - Data Mining is ***not*** (just) about coding, especially in business settings!
    - We **will also** be focusing on several non-technical areas

Let’s play a game…

Data Mining Approach

“If we have data, let’s look at data. If all we have are opinions, let’s go with mine.”

-- *James Love Barksdale*

Former CEO of Netscape

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

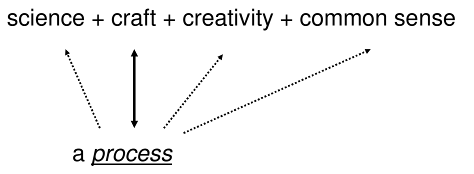
Learning Goals

- 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, …

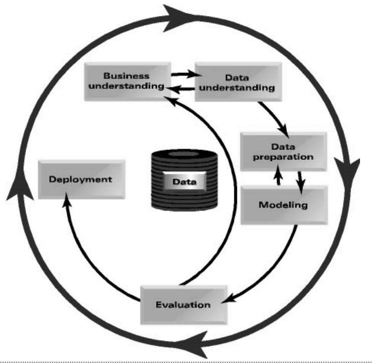
Business data mining is a process



science + craft + creativity + common sense

a *process*

Data Mining Process



Outline

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Reasons for failing Data Mining Projects

This is NOT a course about…

- Statistics
- Database Querying
  - SQL
- Data Warehousing
- Regression Analysis
  - Explanatory vs Predictive modeling
- Big Data

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?