What is Data Science?

BERCS1102016-V000500

Download video Download transcript .srt

Data Science Danger Zone

(1/1 point)

Why is a hacker with substantive expertise a dangerous combination? Select all answers that match:

- [ ] Can apply Machine Learning to big data problems
- [x] Can end up performing statistically incorrect analyses ✔
- [x] May fail to ignore irrelevant factors (more computation) ✔
- [ ] Uses traditional research to solve problems
- [x] Can create legitimate appearing analysis without any understanding of how they got there or what they have created ✔

✔

CONTRAST: DATABASES

This comparison may seem like we are just comparing two different types of databases, relational (strongly structured) databases and NoSQL (semi-structured) databases, instead of a comparison of databases in general versus data science. Indeed, relational databases can contain terabytes of data and many people perform data science on data retrieved from relational databases. However, as we will see in Week 3's lectures, much of the information that is typically used for data science is unstructured or semi-structured, and cannot be stored in a relational database.

The comparison is complicated because the strongly structured nature of relational databases means that they are often used to store supplementary information to the data available in NoSQL stores. For example, an e-commerce website would typically store credit card information in a relational database, while it could store product recommendation information in a NoSQL store. Thus, in this environment data science would involve data retrieved from both strongly-structured and semi-structured stores.

In this keynote talk for VIZBI 2010, the EMBO workshop on visualizing biological data (http://vizbi.org), Ben Fry discusses principles of graphics design and of dynamic visualization that can improve the insight gained from data. He also presents a range of visualizations of genetics data he has created, some of which have been used to improve tools used by biologists, some of which has been exhibited in art galleries and in Hollywood movies. He also presents 'Processing', a computer language he co-developed that can easily create dynamic visualization of complex data. A shorter version of Ben's keynote can be viewed on here on YouTube.

(Optional reading) Our 2009 "Above the Clouds: A Berkeley View of Cloud Computing" Technical Report defines Cloud Computing terms, presents an economic model that quantifies the key buy vs. pay-as-you-go decision, offers a spectrum to classify Cloud Computing providers, and gives our view of the top 10 obstacles and opportunities to the growth of Cloud Computing.

(Optional reading) This interview study characterizes the process of industrial data analysis and documents how organizational features of an enterprise impact it: Enterprise Data Analysis and Visualization: An Interview Study. The study is based on semi-structured interviews with 35 data analysts from 25 organizations across a variety of sectors, including healthcare, retail, marketing and finance.

(Optional reading) This Harvard Business Review article, Data Scientist: The Sexiest Job of the 21st Century, presents a deep dive on what organizations need to know about data scientists: where to look for them, how to attract and develop them, and how to spot a great one.

Curating and Filtering Data

(1/1 point)

Why is curating/filtering data a key model component?

- Real-world data patterns are easy to model
- Real-world data is very dirty ✔
- Real-world data should never be filtered
- Real-world data does not require curating

EXPLANATION

Real-world data often contains inconsistencies or errors that have to be filtered out