Lecture 1:
Introduction to Data and Scientific Method

Heidi Perry, PhD

Hack University

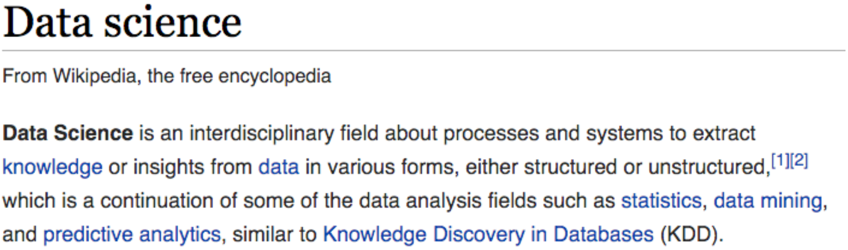
heidiperryphd@gmail.com

10/11/2016

Overview

1. Course outline
2. What is data science?
   - The "science" in data science.
3. Introduction to data
   - What is data?
4. Source Code Version Control
5. In-class exercise

What is data science?



Data science

From Wikipedia, the free encyclopedia

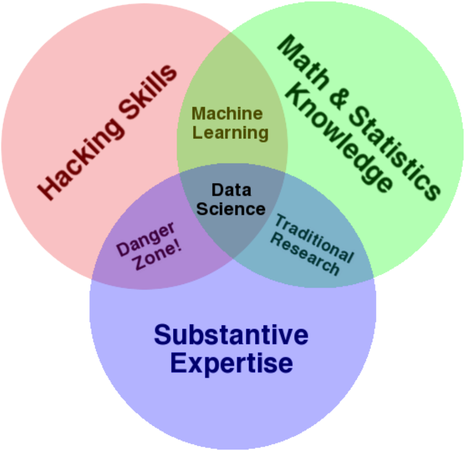
Data Science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] which is a continuation of some of the data analysis fields such as statistics, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

[2] Jeff Leek (2013-12-12). "The key word in "Data Science" is not Data, it is Science". Simply Statistics.

The key word in "Data Science" is not Data, it is Science

"But the key word in data science is not "data"; it is "science". Data science is only useful when the data are used to answer a question. That is the science part of the equation." - Jeff Leek

What is data science?



From Drew Conway

Predictive Modeling

Data scientists answer questions useing predictive modeling and classification, using some combination of traditional statistics and machine learning alogrithms.

Traditional Statistics

Traditional statistics focus on testing hypotheses, inferring results from a sample to a population, and building intrepretable models.

Machine Learning

Machine Learning focuses on predictions, with less concern about intrepretability of the model. Some algorithms (unsupervised) can even generate hypotheses.

Data Products

Communication is a key step in data analysis. The final product will be a report, presentation, interactive website, or app.

Reports and Presentations

- Write clearly and concisely.
- Tell the story of the data analysis and conclusions.
- Include details about data collection, assumptions, and transformations.
- Your work should be reproducible.

Websites and Apps

- Design for ease of use and user understanding.
- Document well, both inside the code and for the user interface.
- Control code version.
- Your work should be reproducible.

Tools

A non-comprehensive list of tools used by data scientists.

- Statistics: Regression models, A/B testing
- Programming: R, Python, or Julia (newer)
- Version control: Git, Subversion
- Data storage: PostgreSQL, SQL Server, csv files
- Big Data Tools: Hadoop, Hive, Pig, Spark, Redshift, Vertica
- Machine learning: Classification, regression, clustering, neural nets/deep learning, etc.

The Scientific Method

The Scientific Method

1. Observe
2. Propose hypothesis
3. Experiment
4. Modify hypothesis, repeat

Applied to Data Analysis

1. Exploratory analysis
2. Propose hypothesis
3. Build model
4. Modify hypothesis, repeat

The Scientific Method

“Epicycle” of analysis “The Art of Data Science” [Peng, 2015].

Data Analysis Cycle

1. State the question
2. Exploratory data analysis
3. Build model
4. Interpret
5. Communicate

Epicycle

1. Develop expectations
2. Collect data
3. Match expectations with data

Characteristics of a Good Question

1. Question should be of interest to your audience.
2. Question is not already answered.
3. The foundation of the question is plausiable.
4. Question is answerable.
5. Question is specific.
6. The answer of the question will be unambigously interpretable.

Six Types of Questions

1. Descriptive
   - Summarize characteristics about the set of data.
   - No interpretation, the results are facts about the data set.
2. Exploratory
   - Hypothesis-generating analysis looking for patterns and trends in the data.
3. Inferential
   - Pose hypothesis from exploratory analysis as a question and answer it from another set of data.
4. Predictive
   - Look for variables that predict outcomes, without being concerned about the underlying reason.
5. Causal
   - Will changing one variable will change another, on average over the population?
   - Requires randomized control trial.
6. Mechanistic
   - The "how" of the causal relationship.

What is data?

Data

Observations organized into variables.

Frequently data is organized into a table or **data matrix**, with each row representing a single **observation**, **observational unit**, **individual**, or **case**. Each column represents a characteristic of the observation called a **variable**.

Example: Hack University Courses

Course	Cohort	Instructor	# Students	Start Date	Day of Week
Dev Ops	Database	Bill McGair	8	2016-02-08	Monday
Data Science	Database	Heidi Perry	9	2016-02-09	Tuesday
Database Engineering	Database	Zeke Wander and Hobson Lane	7	2016-02-10	Wednesday
Machine Learning	Database	Hobson Lane and Zeke Wander	6	2016-02-11	Thursday
Product Management	Product	Nim Wunnan	5	2016-03-07	Monday
Data Visualization	Product		4	2016-03-08	Tuesday
Web Development	Product	Daniel West		2016-03-08	Tuesday
GIS and Spatial Analysis	Product	Joe Dickinson		2016-03-09	Wednesday
UX/Product Design	Product	Ryan Gantz	4	2016-03-10	Thursday
Data-Driven Journalism	Storytelling			2016-04-04	Monday
Visual Design and Infographics	Storytelling			2016-04-05	Tuesday
Mulitmedia Marketing	Storytelling	Zach Krahmer		2016-04-06	Wednesday
Motion Graphics	Storytelling			2016-04-07	Wednesday
Audio Podcasting	Storytelling			2016-04-08	Thursday

Types of Variables

Numerical variables

Numerical variables take numerical values, and it is sensible to add, subtract, and take averages of the values.

- **Continuous numerical** can be any number within a valid range. For example, annual income of Portland residents.
- **Discrete numerical** are numbers that can only take on discrete values. For example, the number of students enrolled in each Hack Oregon class.

Categorical variables

Categorical variables take on only a small set of defined **levels**.

- **Nominal categorical** variables have values that are categories. For example, Hack University cohort.
- **Ordinal categorical** have an inherent order. For example, (agree, neutral, disagree) on a survey.

Large data sets

- Database of your choice (PostgreSQL, MySQL, etc.)
- SFrame or whole Graphlab Create package available for free use for students

Project Proposal

1. What general question do you want to answer?
2. What are your expectations about the question?
    - It answerable? Has someone already answered it?
    - Is your intended audience interested in this question?
    - Is the data needed to answer it available?
3. Based on what you found, ask a more narrow question.
    - This will set the path for your course project.
    - Your question should be answerable with data available to your project team.
4. Summarize the data set.
    - How was the data collected?
    - How many observations? How many variables? What types of variables?

Basics of Source Code Version Control

- Repository (GitHub)
- Working Copy (`git clone <address>`)
- Trunk and branches
- Merging
- Reverting
- GitHub: fork a repository, create a pull request

Resources

Git Documentation

GitHub Guide GitHub: Syncing a fork

References

Roger Peng & Elizabeth Matsui (2015)
The Art of Data Science, Leanpub

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)
OpenIntro Statistics, OpenIntro

Podcast interview with Claudia Perlich
"Talking Machines, Episode Thirteen: Claudia Perlich" (2015)

Recommended Reading

The Art of Data Science, Chapters 1-3
OpenIntro Statistics, Chapter 1
Data Science From Scratch, Chapters 1-3

Articles for discussion:

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
Eight (No, Nine!) Problems With Big Data
Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs
Google Flu Trends gets it wrong three years running
How to Actually Learn Data Science
Data Science Inconvenient Truth
Data Science downsides

In-class work

- Install required software and create required accounts.
- Get Git up and running.
  1. Fork course repository.
  2. Checkout a working copy.
  3. Create a directory Project/YourName and initialize it with a file.
  4. Commit and push your new file.
  5. Create a pull request to add your new file to the course repository.
- Verify Python is correctly installed.
  1. Run all the cells in the Graphs notebook, check that graphs display properly.

For Thursday

Read OpenIntro Statistics, Chapter 2: Probability.