JEM221/JEM227 Data Science with R I
Topic #0

Course information
&
Introduction to Data Science

Ladislav Krištoufek

Outline

Outline

Course information

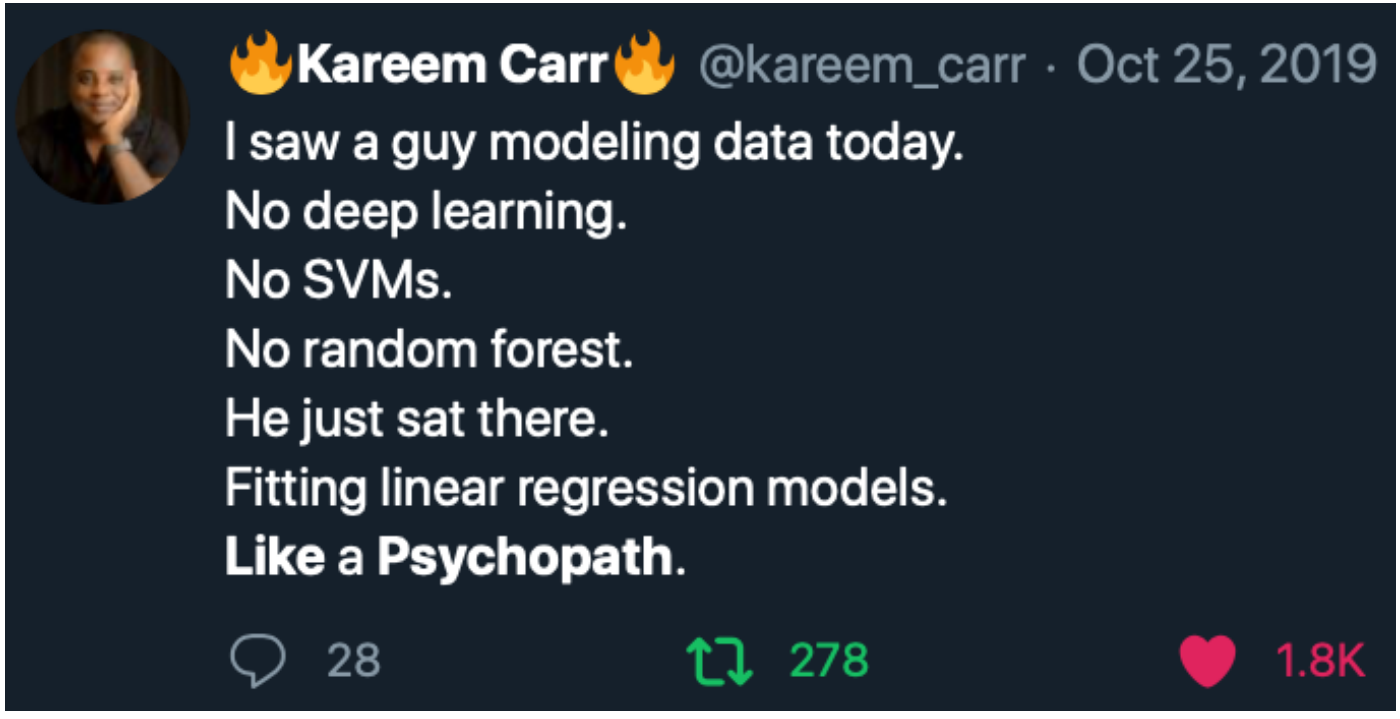
- JEM221/JEM227 Data Science with R I
- Lectures/Seminars:
  - Ladislav Krištoufek (lectures/seminars)
  - Ivan Trubelík (home assignment management & group consultations)
  - Pre-recorder lectures/seminars (links in SIS, recorded in Loom)
  - Offline consultations (not mandatory) in room 016:
    - 27 Oct (Week 4), 10 Nov (Week 6), 24 Nov (Week 8), 8 Dec (Week 10), and 22 Dec (Week 12)
    - between 09:30 and 11:00
- Contact: LK@fsv.cuni.cz

Study materials

- The course is based on and closely follows two books:
  - Toomey, Dan (2014): R for Data Science, Packt Publishing Ltd., Birmingham, UK (**T**)
  - Zumel, Nina & Mount, John (2014): Practical Data Science with R, Manning Publications Co., Shelter Island, NY, USA (**ZM**)
- Additional suggested literature:
  - Grolemung, Garret (2014): Hands-On Programming with R, O'Reilly Media Inc., Sebastopol, CA, USA (**G**)
  - Ledolter, Johannes (2013): Data Mining and Business Analytics with R, John Wiley & Sons, Hoboken, NJ, USA (**L**)
  - Ojeda, Tony et al. (2014): Practical Data Science Cookbook, Packt Publishing Ltd., Birmingham, UK (**O**)
  - Teetor, Paul (2011): R Cookbook, O'Reilly Media Inc., Sebastopol, CA, USA (**TP**)
- The mandatory books are available in the IES library. However, they are (quite) easily accessible online.
- DataCamp courses (also Python, SQL, Power BI, Git, Julia, and others).

Course aims . . .

- are:
  - to introduce you to methods outside of standard statistics and econometrics curriculum
  - to make you comfortable writing your own functions in R
  - to make you comfortable analyzing data in R
  - to enlarge your skill portfolio
  - to make you more attractive to employers
- are not:
  - to go deep into theory of the presented methods
  - to make you a proficient R coder


Kareem Carr @kareem_carr · Oct 25, 2019
I saw a guy modeling data today.
No deep learning.
No SVMs.
No random forest.
He just sat there.
Fitting linear regression models.
Like a Psychopath.
28
278
1.8K


prg.ai

Course pre-requisites

- This is a master's course.
- There are no formal pre-requisites.
- However, students are assumed to have knowledge of statistics and econometrics covered in IES bachelor's courses, specifically Statistics and Econometrics I, i.e. this course is fine for the third year bachelor's students.
- Some core knowledge of R is assumed as well (up to the level of bachelor's course Data Analysis in R).

Outline

Course schedule

- Week #1: Course information + Introduction to Data Science
- Week #2: R Basics (ZM 1, G 3-5) – meant as repetition for most and possibly as a jump over for the students with only minimal knowledge in R
- Week #3-#6: Model evaluation (ZM 5), Memorization methods (ZM 6)
- Week #7+#8: Advanced regression methods (linear, logistic, GAMs, LASSO, ridge) (ZM 7, T4-5)
- Week #9-#10: Bagging and Random Forests (ZM 9)
- Week #11-#12: Kernels and Support Vector Machines (ZM 9, T 10, L 14, BL 7)

Outline

Grading (1/2)

- The final grade consists of three components:
  - 3 skill tracks in DataCamp:
    - Skill Track “Statistics Fundamentals with R” (10 points) - by 3 December 2023 CET
    - Skill Track “Machine Learning Fundamentals in R” (15 points) - by 4 February 2024 CET
    - Skill Track “Supervised Machine Learning in R” (15 points) - by 4 February 2024 CET
  - 4 core assessments in DataCamp:
    - “R Programming” (5 points) - by 12 November 2023 CET
    - “Exploratory Analysis” (5 points) - by 12 November 2023 CET
    - “Analytic Fundamentals” (5 points) - by 12 November 2023 CET
    - “Understanding and Interpreting Data” (5 points) - by 12 November 2023 CET
    - You need to get at least 120 score to obtain 5 points for each of these four Core Assessments.
    - You can re-take the assessments twice a week during the whole semester (up till the deadline). Remember that the last one counts (not necessarily the best one).

Grading (2/2)

- 2 topical assessments in DataCamp:
  - “Statistics Fundamentals with R” (20 points) - by 4 February 2024 CET
  - “Machine Learning Fundamentals in R” (20 points) - by 4 February 2024 CET
  - To get the score, use the DataCamp score x and fit it to (x-60)/80*100% $\Rightarrow$ you need 140+ score in DataCamp to get full points from each topical assessment.
  - At least 50%, i.e. at least 10 points, from each topical assessment is a necessary (not a sufficient) condition for passing the Data Science with R I course.
  - You can re-take the assessments twice a week during the whole semester (up till the deadline, of course). Remember that the last one counts (not necessarily the best one).

Grading

Grading scale (following Dean's Provision 17/2018):

- A: $points > 90$
- B: $80 < points \leq 90$
- C: $70 < points \leq 80$
- D: $60 < points \leq 70$
- E: $50 < points \leq 60$
- F: $points \leq 50$

Outline

Harvard
Business
Review

ARTWORK: TAMAR COHEN, ANDREW J BUBOLTZ, 2011, SILK SCREEN ON A PAGE FROM A HIGH SCHOOL YEARBOOK, 8.5" X 12"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

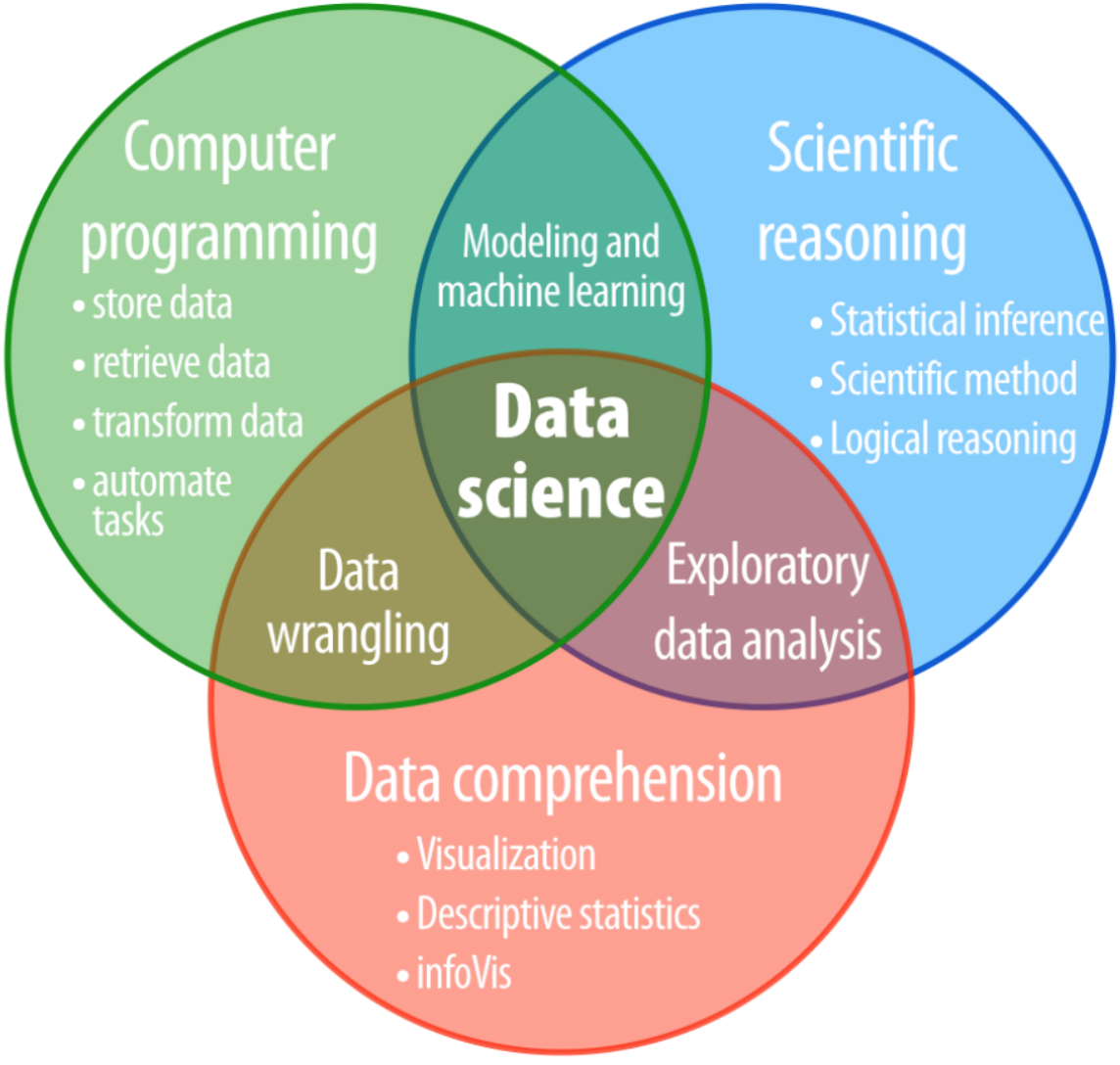
WHAT TO READ NEXT

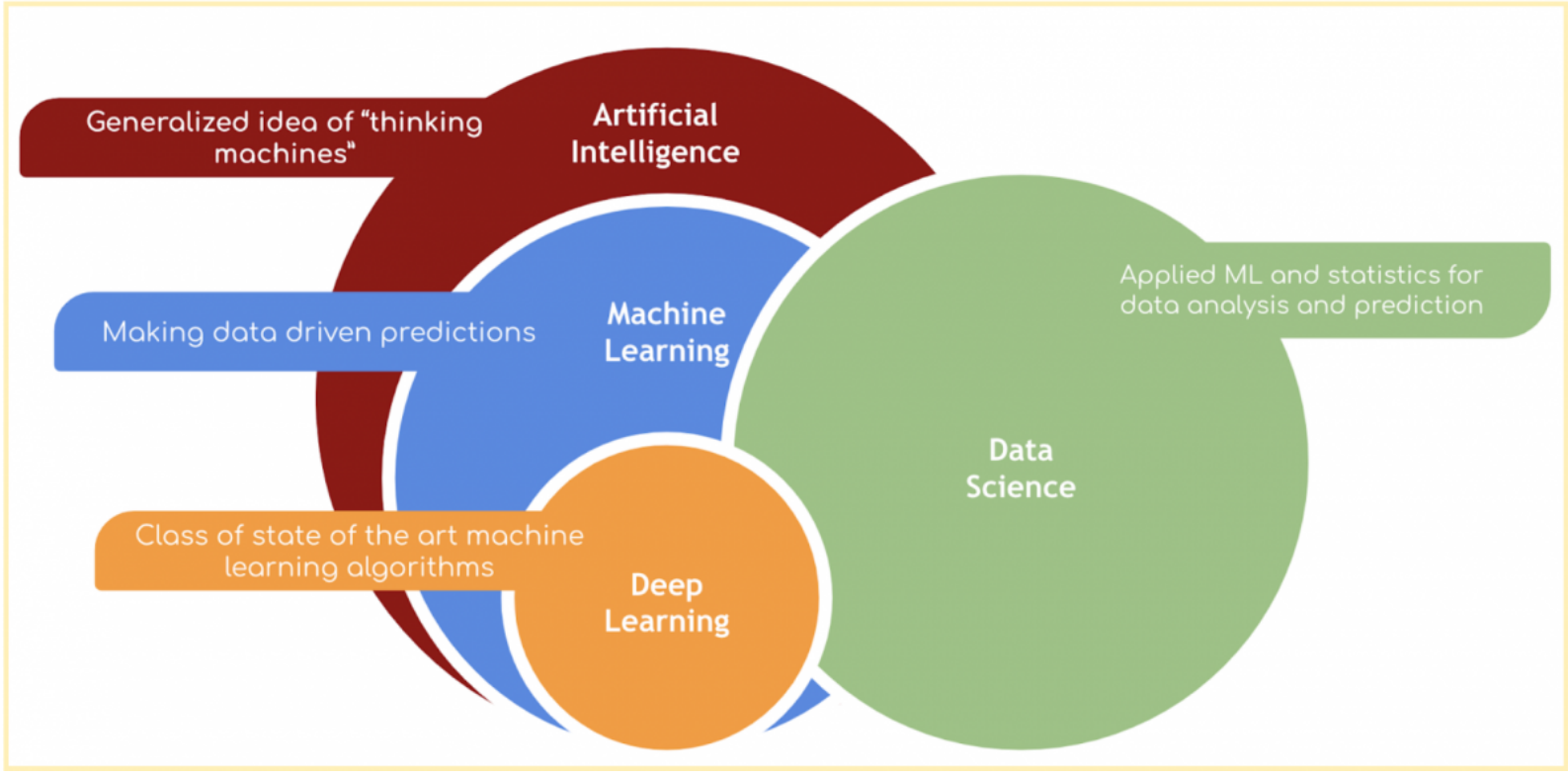
Big Data: The Management Revolution


McKinsey&Company
Big data: The next frontier for innovation, competition, and productivity
By James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers
Executive Summary (PDF–922KB)
Full Report (PDF–6MB)

What is Data Science?

According to Wikipedia,
“ Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured . . . ”


Computer programming
• store data
• retrieve data
• transform data
• automate tasks
Scientific reasoning
• Statistical inference
• Scientific method
• Logical reasoning
Modeling and machine learning
Data science
Data wrangling
Exploratory data analysis
Data comprehension
• Visualization
• Descriptive statistics
• infoVis


Generalized idea of "thinking machines"
Artificial Intelligence
Making data driven predictions
Machine Learning
Applied ML and statistics for data analysis and prediction
Data Science
Class of state of the art machine learning algorithms
Deep Learning

Data Science

- Need of entire analytics universe
- Branch that deals with data
- Different operations related to data i.e.
  - Data Gathering
  - Data Cleaning
  - Data Subsetting
  - Data Manipulation
  - Data Insights [Data Mining]

Machine Learning

- Combination of Machine and Data Science
- Machines utilize Data Science techniques to learn about the data hence called as Machine Learning
- Model Building, Model Evaluation and Validation
- 3 Types:
  - Unsupervised Learning
  - Reinforcement Learning
  - Supervised Learning
- Most popular tools are Python, R and SAS

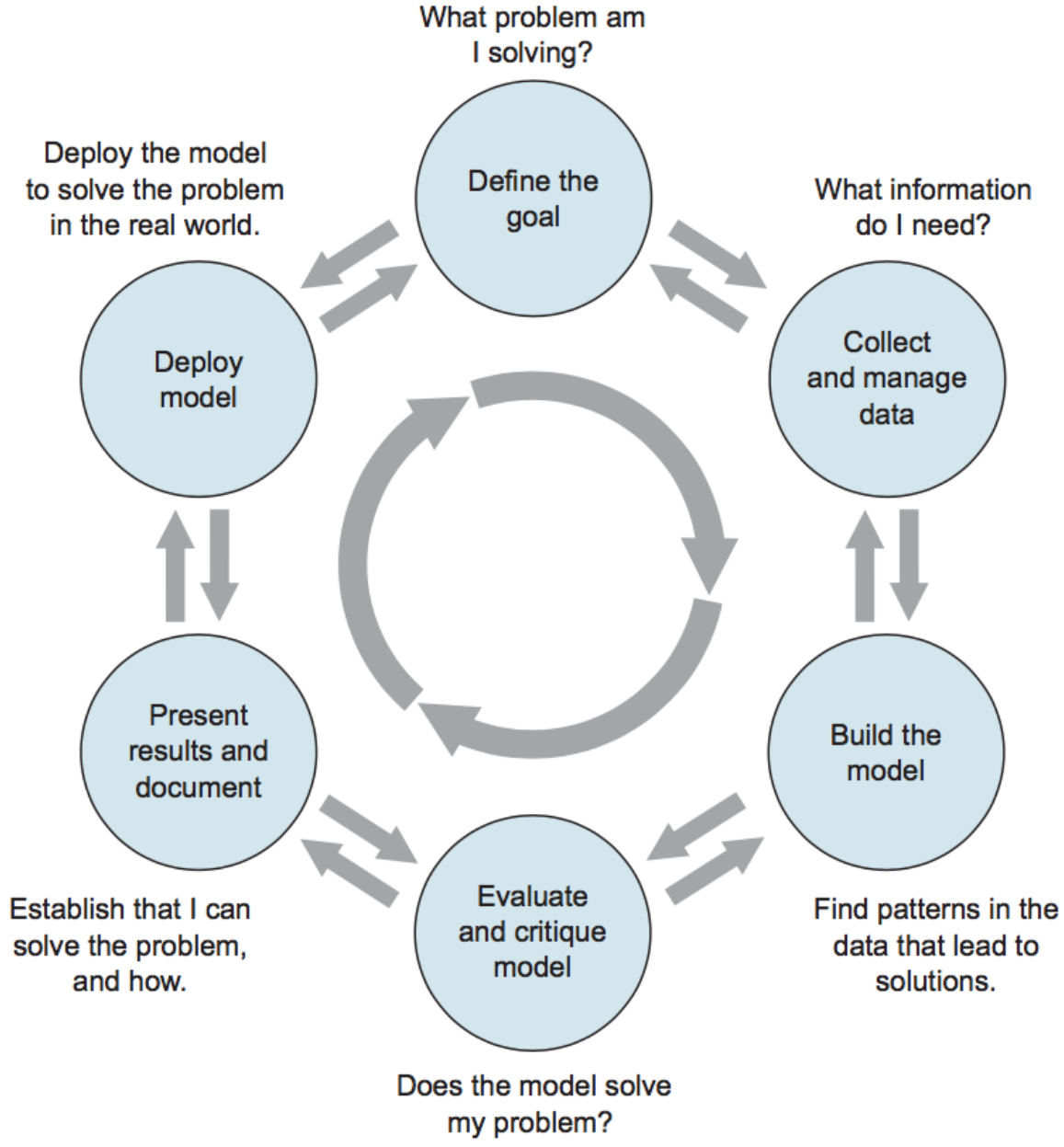
Deep Learning

- Specific branch of Machine Learning that deals with different flavours of Neural Network
- Examples
  - Simple Neural Network
  - Convolutional Neural Network
  - Recurrent Neural Network
  - Long Short Term Memory
- Mainly utilized in..
  - Object detection in Image and Video
  - Speech Recognition
  - Natural Language Processing and Understandings

Artificial Intelligence

- Big Umbrella
- Empowering machines to take decisions on their own
- As the name suggest imparting humans' natural intelligence in machines
- Thus machines have ability to understand and react according to the situation

Outline


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Find patterns in the data that lead to solutions.
Build the model
Does the model solve my problem?
Evaluate and critique model
Establish that I can solve the problem, and how.
Present results and document
Deploy the model to solve the problem in the real world.
Deploy model

Step #1 – Define the goal

- Ask a correct question.
- Is the question transferable into a hypothesis?
- Is it quantifiable?
- What specific data do you need to answer it?

Step #2 – Collect the data

- Identify the data you need.
- Can you get such data?
- Explore the data.

Step #3 – Build the model

- Identify useful methods.
- Identify correct methods:
  - classification methods
  - scoring methods
  - ranking methods
  - clustering methods
  - finding relations
  - characterization
- Apply the methods.

Step #4 – Evaluate the model

- The following questions are usually being answered:
  - Is the model accurate enough?
  - Does it perform better than a qualified guess?
  - Do the results make sense?
- If not, we need to loop back to the previous steps.

Step #5 – Present the model

- In the business world, sadly the most important part of our effort.

Step #6 – Deploy and maintain the model

- To show that you are worth having . . .
- Luckily, the world changes, the society changes, hence the models need to be maintained and updated/adjusted as well.