DATA11001

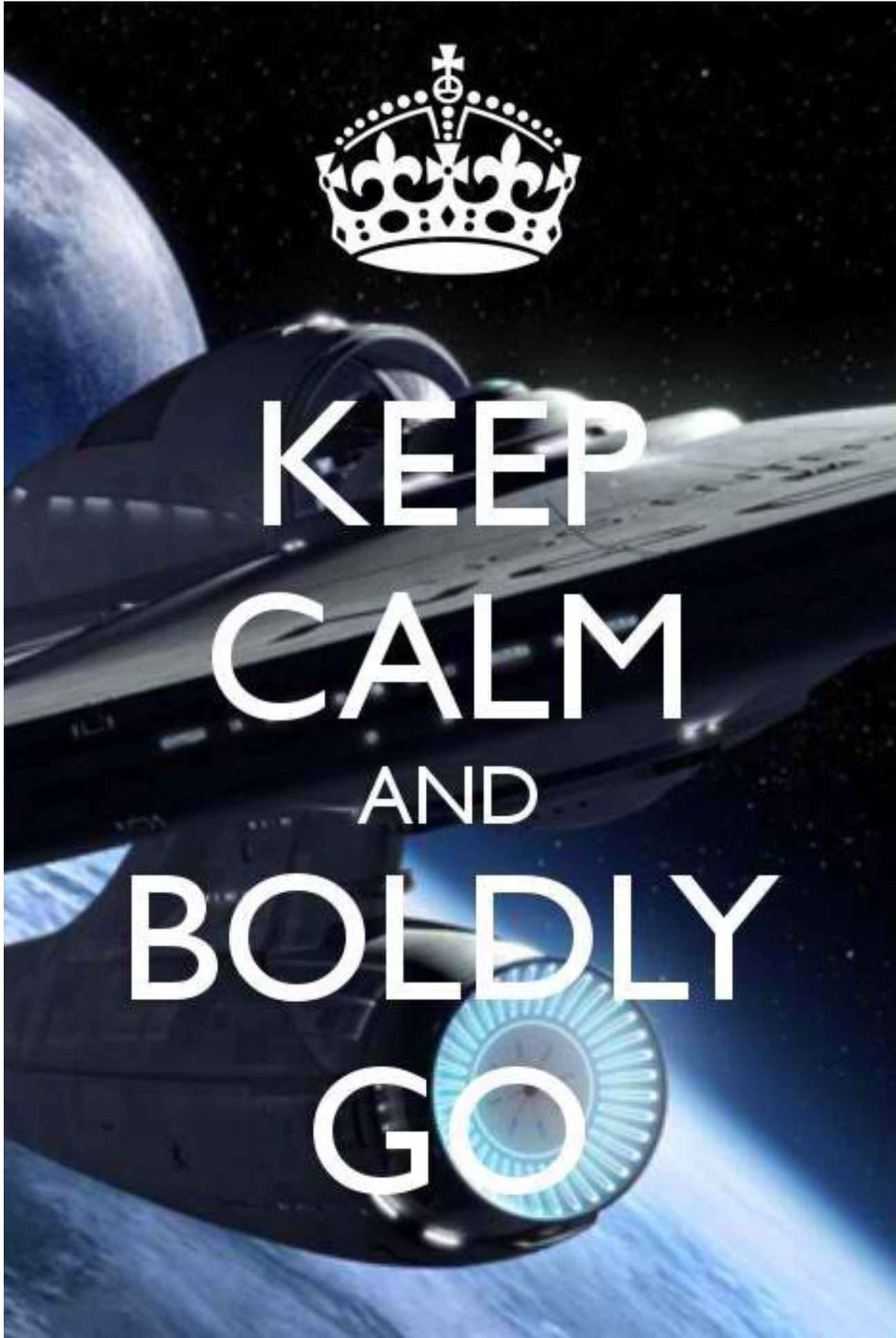
INTRODUCTION TO DATA SCIENCE

THE FINAL EPISODE: MINIPROJECTS

TODAY'S MENU

1. LOGISTICS FOR THE REST OF THE COURSE

2. MINI-PROJECTS



BASICS

- This is the last lecture except for guest lecture(s?):
  - Oct 16: **"TBA"** (*Futurice*)

- The GDPR exercise from last week will be done this week "as usual", i.e., individual solutions marked at the exercise sessions (week 41)

- There will be no other exercises

- Instead, you will be working on **"miniprojects"**

MINIPROJECTS

- Idea is to come up with an **idea** of a data science solution:
  - starting point **A** (need, data, possibilities)
  - aim **B** (added value! not just a technical solution)
  - you need to come up with:
    + **both A and B**, as well as
    + the **solution** to get from A to B

- Not good ideas:
  - take clean data and a given task (e.g., prediction) and apply a machine learning method
  - take data and visualize it without any particular aim or goal
  - copy an existing data science project

  => **try to be creative!**

EXAMPLES

- These are primarily to give you an idea of suitable topics
- IT IS NOT ADVISABLE TO ACTUALLY CHOOSE ONE OF THESE
- Instead, they are just examples to give you the flavour

- Elements of the project (will be evaluated based on these):
  - Data: sources, wrangling, management
  - Data analysis: statistics, machine learning
  - Communication of results: summarization & visualization
  - Operationalization: added value, end-user point of view

EXAMPLE #1: HOW TO BE POPULAR OF SOCIAL MEDIA

1. Create a virtual assistant that recommends ways to boost social media popularity.
2. Instagram/twitter/facebook data from open APIs. Challenges: extracting data following accepted policies. Finding other relevant data sources. Could involve big data management tasks.
3. Predicting popularity using various machine learning tools. In case of image content, e.g., deep learning.
4. For example, visualization of the data so that a user user can identify good strategies: e.g., an interactive hashtag exploration tool that highlights common choices based on co-occurrence.
5. The more popularity, the better the system. Usability also important (considering target group, e.g., teenagers wouldn't appreciate boring statistical graphics).

EXAMPLE #2: OK BUT FIRST COFFEE

1. Sometimes the coffee maker in Gurula (student space) is out of coffee just when you need it the most. It's easy enough to make more, but you'd like to know beforehand.
2. Create your own data by installing a webcam pointing at the coffee maker. Be sure not to violate anyone's privacy and be careful about controlling access to the camere feed.
3. Use computer vision techniques to predict the amount of coffee left.
4. Share information through an app (possibly a simple visualization, or just the number of cups left).
5. No coffee = bad, coffee = good. Also, possibly recommend how many cups to brew based on expected consumption, or predict peaks in demand.

EXAMPLE #3: RENT-A-FLAT

1. The demand for rental flats in Helsinki is leading to high, and increasing rental prices. However, the rents per area vary greatly. Which areas are underrated (and "underrented")?
2. Helsinki open https://dev.hel.fi/apis/open311/ . Other possible data sources can include info about important services (bars, computer stores, ...) including info about their popularity.
3. Summarization of data to generalize to future rents.
4. Visualization of GIS data, possibly over time. In addition to rental prices, the visualization can display, e.g., crime statistics, availability of services, "hipster index"
5. Information that supports individual users (with different preferences) in finding a cheap flat. Could also be used by city officials to support planning.

EXAMPLE #4: MONKIE GOES TO HOLLYWOOD

1. Propose 5 blockbuster movie ideas
2. IMDB, sales figures, reviews, ...
3. Find the most profitable combinations of the given data (e.g. combination of genre-director-lead actor, etc)
4. Visualization of sales data over movie attributes (and combinations of them), presentation of the key selling points of your proposed movies
5. Incorporate your findings into fresh movie ideas (be creative). Convince the money guys as to why choose your proposed movies, over the proposals of your competitors.

EXAMPLE #5: CLIMATE CHANGE

1. Study of meteorological and other data to let us see what the world will be like in the future
2. Plenty of open climate data available on temperature changes throughout time. Consequences: water-level rise, agricultural outcomes, weather, everyday life
3. Predict consequences given climate forecasts (under various policy implications)
4. Visualize your results to support your final statement. For example: photorealistic scenes about life at the seaside in Paris, Orange orchard in Helsinki, ...
5. Make it feel tangible but based on science (not too cheesy)

EXAMPLE #6: SHOULD I BE WORRIED ABOUT IT?

1. Consider various risks: meteorite-wise (NASA), lightnings, epidemics, cancer, car accidents, ...
2. NASA, climate data, global health data, ...
3. Predict risks in different areas (GIS). Compare the magnitudes of various risks
4. Remember: risk x damage = expected loss
5. Display risks to alert or pacify the public. Guide for travelers, ...

IMPORTANT DATES, GRADING

- Miniproject:
  - **ASAP** read instructions & **form groups**: ASAP
  - **Oct 10** register group and **topic proposals**
  - **Oct 26** 4pm-6.30pm project **pitching sessions**
  - **Oct 27** project **delivery**
  - **Oct 30** submit **feedback** on two (2) other projects (assigned automatically)

- Grading: exercises 50% + miniproject 50%

- There will be **additional exercises** available to collect more exercise points: more info about this soon!

OR, ALTERNATIVELY:

- As was stated in the beginning, you can also do an individual project + exam

- Dates:
  - **early November:** pitching session and delivery of individual projects
  - **Nov 29:** exam (separate exam)
    + you **must** register to the exam well in advance

- In this case, grading will be based on
  - exercise points (30%), miniproject (35%), exam (35%), or
  - miniproject (50%), exam (50%)

PITCHING AND DELIVERABLE

- **Pitch**: about 3-5 minute pitch about the project
  - others can give feedback (online)
  - each student will be assigned to submit written feedback for two projects: pay close attention to their pitches
  - three parallel sessions, about 15 teams per session
  - NB: this is the time reserved for the exam
  - place TBA
  - individual projects: similar but a bit later

- **Deliverable:** Demo/application/system
  1. can be a web application, mobile app (!), visualization, or a static output
  2. in addition, a blog post, Youtube video, or a report:
     + but remember the target audience of your project (no scientific report for non-scientist audiences)

PRACTICALITIES

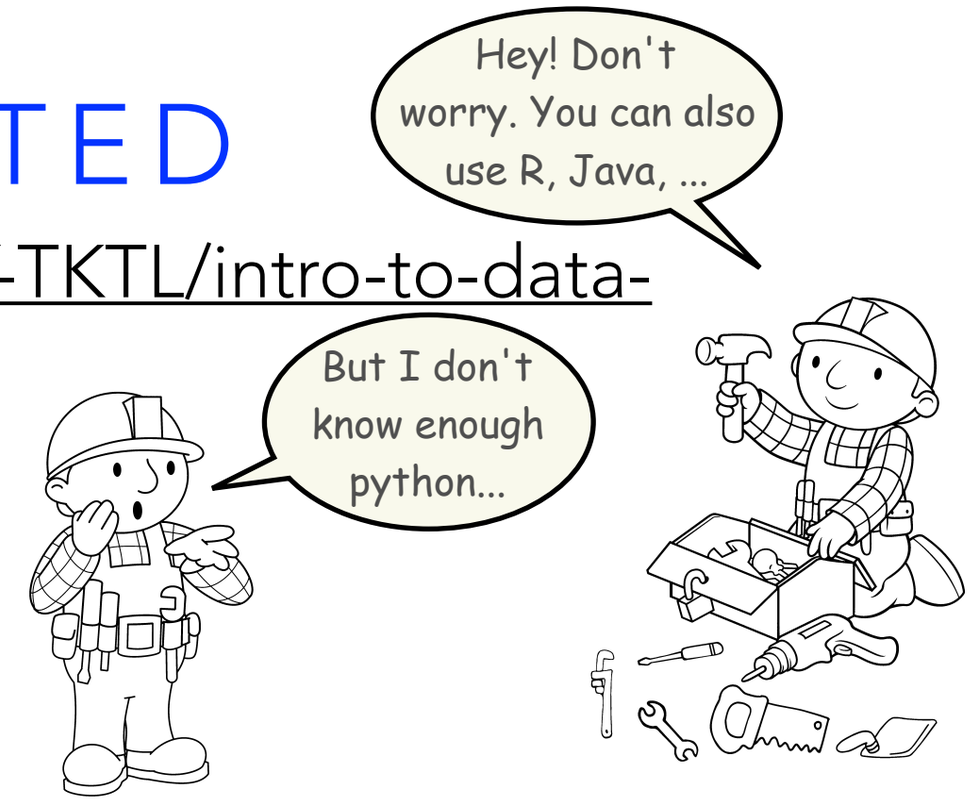
- We'll be using peergrade.io for handling the project

- First choose a **"coach"** on Doodle (see exercise page)
  - who to turn to: Chang, Johanna, or Ville-Veikko
  - also determines the pitching session

- Submit proposal on peergrade.io by **Oct 10**
  - team members
  - coach
  - topic

- We'll give quick feedback to adjust but you can keep working -- **very** unlikely that we'd reject the whole idea

HOW TO GET STARTED



- Instructions at https://github.com/HY-TKTL/intro-to-data-science-2017/wiki/Mini-projects

- Additional hints:
  - **Don't be too ambitious:** the workload per person should be similar to that of the weekly exercises x 3 (three weeks)
    + doesn't have to be the most advanced deep learning on big data with a fancy interface: linear regression on reasonable size data and a working interface is ok!
  - **Fail fast:**
    + make prototype solutions and have a contingency plan ("plan B") in case there's a problem
    + even in the worst case scenario where your project would fail to produce a concrete output, we can evaluate based on the idea and the implemented parts, so it wouldn't be the end of the world