# 02807 Computational Tools for Data Science

Anders Roy

# The Purpose of this Course

*[Why this course was originally started] There are multiple reasons for teaching this course. Many of the people who do work on machine learning and scientific computing come from a mathematical background where the focus has been on the mathematical theory rather than on the practical tools and implementations of these theories. We are teaching this course to help students get a hold of the different tools and technologies available for working with large scale data.*

# New Course Staff This Year

- Do not expect everything to be the same as last year.
- Please give us feedback during the semester, so we can adjust and improve.

# Who are we?

- Teachers
  - Anders Roy (main contact person)
  - Philip Bille
  - Inge Li Gørtz
  - Carsten Witt
  - Paul Fischer
- Teaching Assistants
  - Anders Bregendahl
  - Alexander Kvist Andreasen

# Who are you?

https://goo.gl/forms/L2PduAJjS46vTbRr1

# Practical information

- Website: http://www2.compute.dtu.dk/courses/02807/
- Main book: Mining of Massive Datasets
  - See mmds.org
  - Online version freely available here
  - Other relevant materials can be found on this page

# You have a problem?

1. Search the web

# You have a problem?

1. Search the web
2. Look up in relevant documentation

# You have a problem?

1. Search the web
2. Look up in relevant documentation
3. Ask co-students

# You have a problem?

1. Search the web
2. Look up in relevant documentation
3. Ask co-students
4. Ask TAs/teachers during class

# You have a problem?

1. Search the web
2. Look up in relevant documentation
3. Ask co-students
4. Ask TAs/teachers during class
5. Ask on Piazza

# You have a problem?

1. Search the web
2. Look up in relevant documentation
3. Ask co-students
4. Ask TAs/teachers during class
5. Ask on Piazza
6. Come by during office hours (Mondays 12.30-13.00 in 322/006)

# I already know about Python, Git, X, Y, Z, ...

1. Great for you!

# I already know about Python, Git, X, Y, Z, ...

1. Great for you!
2. Skip exercises you find too easy, instead
3. read in the book, or
4. help the beginners, or
5. do something else that doesn't waste your time.

# Lectures / classes

- Lectures Tuesdays at 13.00 in 306/033
- Afterwards exercise classes until 17.00 in 303A center area, 306 1st floor west-side area (or stay in here)
- Please tell other students to leave these areas, so we can stay together – otherwise it will be very hard for TAs/teachers to help you.

# Plan

- Week 1: The UNIX terminal and Git
- Week 2: Python brush up #1
- Week 3: Python brush up #2
- Week 4: Massively Parallel Computation (chap. 2 in MMDS)
- Week 5: Filtering and Streaming (chap. 4 in MMDS)
- Week 6: Clustering (chap. 7 in MMDS)
- Week 7: Databases
- Week 8: Locality Sensitive Hashing (chap. 3 in MMDS)
- Week 9-13: Project work in groups

# Mandatory Exercises

- 4 mandatory assignments
- Individual
- More details when first mandatory assignment is announced

# Grading

- Based on:
  - Your mandatory assignments.
  - Final project work
- No exam.
- Details will be given later.

# The UNIX Terminal and Git

# UNIX Terminal Demo

# Git demo using DTU GitLab

# Time to Work

- Final questions?
- Todays weekplan is on the website
- Go to 303A center area, 306 1st floor west-side area (or stay in here if no space left)