Data Science & Society

InfoMdss is an obligatory course in
- Applied Data Science (ADS) profiles
- Master Business Informatics (MBI)

Dr. Marco Spruit
m.r.spruit@uu.nl
Associate professor
ADS Profile Coordinator
Learning outcomes

I. Understand the role of data science and its societal impact

II. Recognise the knowledge discovery processes in applied data science

III. Identify trends and developments in big data technologies

IV. Apply selected big data technologies to solve real-world problems
Society focus

**UU Application domains:**
- Neonatology
- Business
- Epidemiology
- Geography
- Cell biology
- Ethics & Privacy
- Psychiatry
Process focus

Knowledge Discovery Process:
• = Applied Data Science
• CRISP-DM method

Data Analytics:
• Methods & Statistics
  - Traps in Big Data analysis
  - p-values, multiple testing, overfitting, etc
• Workshop tutorials!
**Big Data focus**

**Big Data:**
- Context: Book review
- Focus: Identified by experts
- 4Vs
- BD vs DWH
- SQL vs NoSQL
- Ethics & Compliance
  - Philosophical perspective?
Cloud Computing focus

**Cloud Computing:**
- Infrastructure choices:
  - Local/UU servers (control)
  - IaaS / PaaS (scalability)
  - HPC / Grid (performance)
- MS Azure DSVM
- Or... AWS, Google?
- Horizontal scaling vs vertical scaling
Domain experts *empowerment* focus

**Service Computing:**
- = Applied Data Science
  - ORTEC, CLA case studies
- “*Self-service Data Science*”
  - Empowerment of experts
  - Using pre-trained models
  - → My [research theme]....

*e.g.*
- ORTEC: Spark workflow
- Azure: Cognitive Services
Workshops in MS Azure

Deficiency assignments:
• Bash
• Python

Hadoop from Commandline:
• Wordcount in Python
• Neonatology project

Spark with Jupyter Notebooks:
• Wordcount in PySpark
• Epidemiology project
Questions