2 Dec. 2017, JADS
Discussion on Fairness in Data Science
Moderated by Dr. Maurits C. Kaptein
**Fairness:** How to avoid unfair conclusions?

**Accuracy:** How do we guarantee accuracy?

**Confidentiality:** How to not reveal secrets?

**Transparency:** How to make answers indisputable?
First focus on **Fairness**; lets assume the model output is accurate and transparent...
What is Fairness?

- Non discriminating?
- Broader definitions?
- Veil of ignorance?

Canonical example of unfair Data Science: Higher bail-out charges for African Americans based on predicted risk of recurrence

Search for examples....
Fair? (show of hands)

Decision to search a person in the airport based on membership of a terrorist group.
Fair?

Decision to search a person in the airport based on religious beliefs.
Fair?

Decision to search a person in the airport based on a set of features that does not include religious belief, but does predict it accurately.
Fair?

*Decision to search a person in the airport based on race.*
Fairness implemented:

- Is a distinct set of “non-permitted” features a sufficient approach to ensure fairness? (again, assume the models are accurate, etc.)
Fairness implemented:

- Can we identify a such a feature set?
  - How do we select these features? (legal framework, …)
  - What if the non-permitted features are directly related to permitted features?
Fair?

*Decision to change a medical treatment – to the benefit of the patient – based on race.*
Fairness implemented:

- Can our permitted feature set be context independent? Or does each problem need its own list?
  - Societal vs. personal gains?
Discussion points:

- Can fairness be formalized?
- Is the current legal framework sufficient?
- Can we enforce / regulate fairness?
- ...?