Trusted Data Analytics – an oxymoron?

Dr Simone Stumpf
Centre for Human-Computer Interaction Design

Simone.Stumpf.1@city.ac.uk
@DrSimoneStumpf
- Weirdly normal
- Clearly confused
- Living dead
- Military intelligence
What is trust?
What is online trust?
What is trust in data analytics?

Trustor $\rightarrow$ Risk & Uncertainty $\rightarrow$ Trustee

???
Increased Trust through intelligibility of input, model, output, and purpose.
Explainability ≠ transparency ≠ intelligibility
Table 2: Visual feedback can be integrated in a straightforward manner. Second, naïve Bayes is structured such that rich user recommendations. (Positive and negative likely messages containing this word will be classified to 0%).

For the purposes of investigating our basic approach and machine learning design considerations, we developed a system that

- allows users to interact directly with the program execution.
- provides detailed explanation.
- allows users to "debug" by directly interacting with explanations.

One derived decision stump (single rule) was learned from the first 100 times, it predicted that you'll like this song. Here are four of the predictions:

- If you listen to music where the energy is between 70 and 80, you will likely like this song.
- If you have just asked why this message is not filed in the user-specified folder, the computer is likely to have understood types of explanations. (Our bar graph shows the system's confidence in its predictions, with a red font indicating high confidence.)

The system's keyword-based explanation, since the rules are defined on the source code, is displayed graphically. When a user modifies the input, the system switches its keyword list and highlighting to those keywords. The user can also change the new value to be almost exactly the value specified.

Table 1 also serves as the soundness sheet showed a random sample of decision rules, which showed how the participants tended to have in common, and what makes them explainable.
Explanatory debugging

[Kulesza et al. IUI 2015]

Improved understanding, satisfaction

Future improved behaviour
Intelligibility principles

- Be iterative
- Be sound
- Don't overwhelm
- Be complete

Controllability principles

- Be reversible
- Be incremental
- Be actionable
- Honour feedback
Why Hockey?

This message has more important words about Hockey than about Baseball.

```
and
```

The difference makes the computer think this message is 2.3 times more likely to be about Hockey than Baseball.

```
yields
```

67% probability this message is about Hockey

Combining 'important words' and 'folder size' makes the computer think this message is 2.0 times more likely to be about Hockey than about Baseball.
Study setup

- 77 participants split into two groups: 40 using EluciDebug, 37 using a version without explanations and advanced feedback
- 20 Newsgroup data set (Hockey and Baseball): initial system training on 5 messages for each subject, 1850 unlabeled messages to sort
- 30 minutes to “make the system as accurate as possible”
- Measures: accuracy, feedback given, mental model scores, perceived workload
Results

- **More accurate system accuracy with less effort**
  - 85% for our system versus 77% control system at end of study
  - Made adjustments to 47 messages while control had to label 182 messages

- **With better understanding**
  - 15.8 mental model score versus 10.4
  - The more you understand, the better you can make the system

- **Do not overwhelm**
  - No difference in workload measures
No one-size-fits-all

Appropriate trust

Explanation

Feedback

Changed behavior now
Designing for greater intelligibility

[By Eiband et al. IUI 2018]

**WHAT to explain?**

- **Expert Mental Model**
  - What happens to the best of our knowledge? What can be explained? What does an expert mental model of the system look like?

- **User Mental Model**
  - How do users currently make sense of the system?
  - What is the user mental model of the system based on its current UI?
  - How does it differ from the expert mental model?

- **Target Mental Model**
  - Which key components of the algorithm do users want to be made transparent in the UI?
  - To what extent are users actually interested in the rationale behind the algorithm?

**HOW to explain?**

- **Iterative Prototyping**
  - How can the target mental model be reached through UI design?
  - How and where can transparency be integrated into the UI of the system?

- **Evaluation**
  - How has the user mental model developed?
  - Has the target mental model been reached?
controls heat

makes heat
What to explain:
- "Unexpected behaviour" e.g. preheating, demand response, overshooting temperature, etc
- Detailed reasons for decisions
- Motivations and benefits

How to explain
- Mainly textual
- Simple graphical ok
What to explain
- 7 key system decision points
- Abstract away from optimisation algorithm
- Explain key input data e.g. current internal and external temperatures, learnt properties around rate of heating of the home, etc.

How to explain
- Text first, graph on demand
Study setup

- 60 participants with simulated 5 home heating scenarios

- Between-group with 4 types of explanations
  - Control (no explanation), Text-only, Graphical-only, Both

- Measured demonstrated and perceived understanding, perceived trust
Results

- Perceived understanding?  
  -\xmark

- Demonstrated understanding?  
  \~

- Trust?  
  -\xmark

- Perceived understanding and trust?  
  \checkmark
So what does this all mean?

Intelligibility
Controllability
User Experience