Can we automate Data Science ?

Luc De Raedt

KATHOLIEKE UNIVERSITEIT LEUVEN

ERC AdG - SYNTH
Synthesising Inductive Data Models

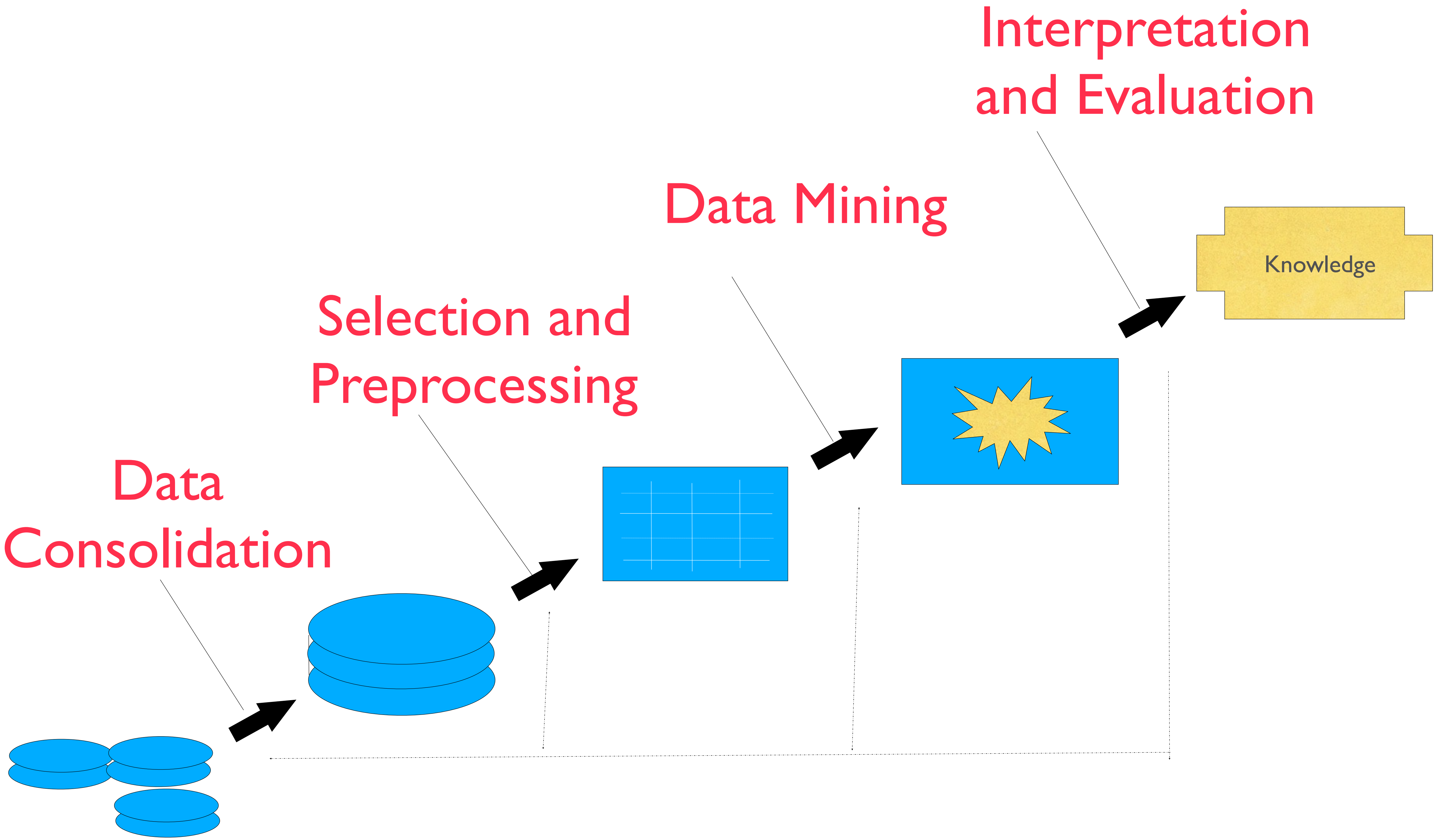
Why is automating data science interesting ?

- At least two reasons :
  - Data Science is hard
    - it could be quite practical
  - Data Scientists are intelligent
    - Artificial intelligence wants to automate intelligence

The Robot Scientist

- The robot scientists
  - Adam (functional genomics) (Ross King et al. Nature 2004)
  - Eve (drug screening (Ross King et al. Science 2009)
- Automates certain scientific tasks in the life sciences
  - construct hypotheses
  - devise experiments and carry them out using lab robots
  - interpreting the results, possibly revise the theory and repeat

The knowledge discovery process



Interpretation and Evaluation

Data Mining

Selection and Preprocessing

Data Consolidation

Fayyad et al. 96

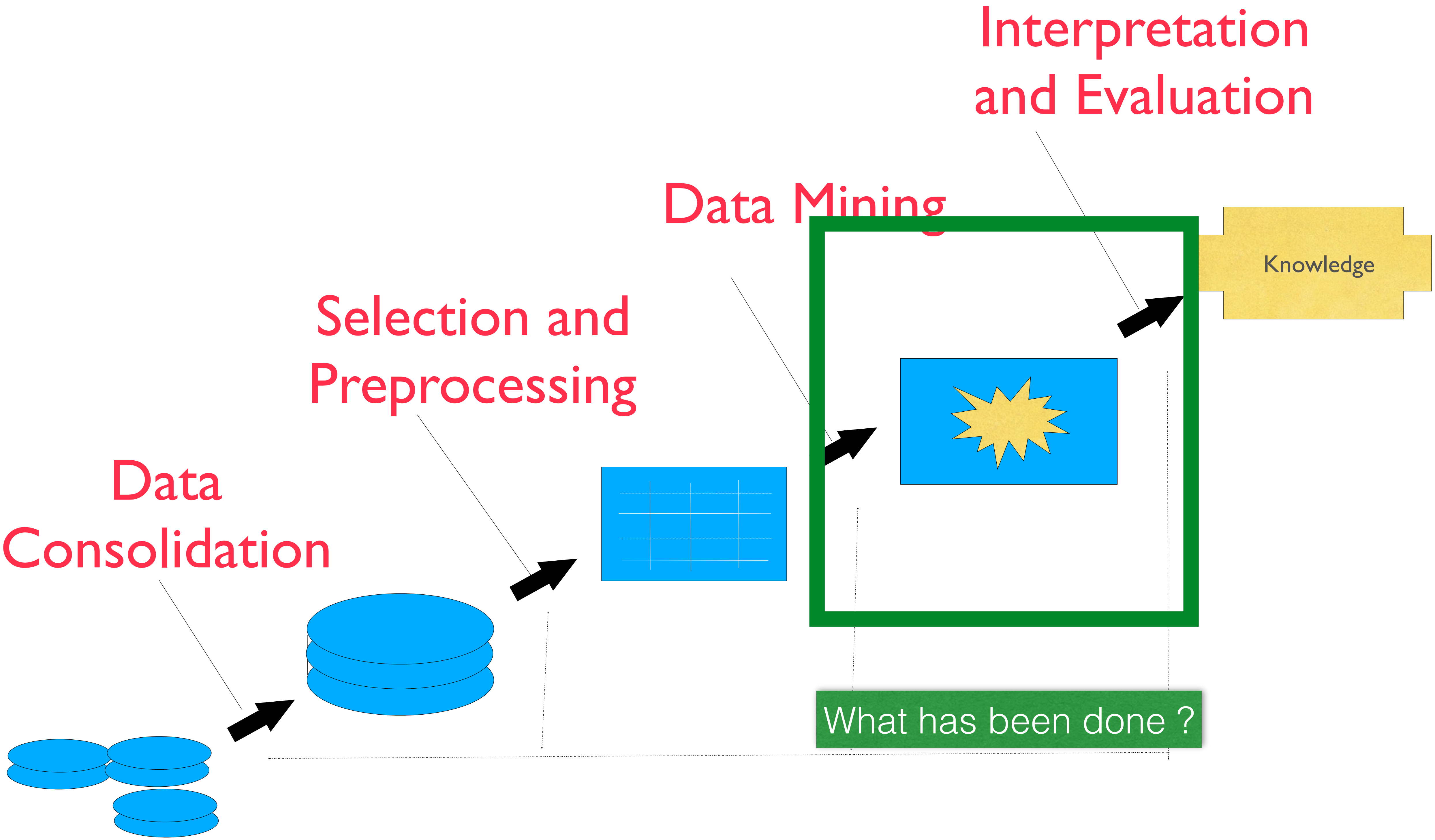
What has been done ?

- AUTO-ML workshops
    - **meta-learning** has been studied for at least 20 years
        - use decision trees or neural nets ?
    - is quite popular and very effective (e.g. Auto WEKA)
    - automatic algorithm configuration / hyper parameter tuning (e.g. Hoos et al.)
    - but it assumes that the learning task (and the portfolio of algorithms is known); focus on classification

What has been done ?

- Automated statistician (Ghahramani et al.)
  - given a data set and a task
  - generate a statistical report in natural language
  - but again assumes the task is given
  - results on e.g. time series and regression (with compositional models)
- Darpa’s current call for *Data-Driven Discovery of Models*

The KDD process



Fayyad et al. 96

What if

- we do not know the learning task ?
    - can we automatically determine the right one ?
    - can we automatically determine the type of models to consider ?
- the data still need to be pre-processed ?
    - can we automatically select the right features ? the dependent from the independent variables ?
    - can we automatically transform the data in the right form ?
    - can we retain “understandability” to the user ? no black boxes?

What if

- we do not know the learning task ?
  - can we automatically determine the right one ?
  - can we automatically determine th[illegible] of models to consider ?
- the data still need to be [illegible]ocessed ?
  - can we aut[illegible]lly select the right features ? the dependent from the independent variables ?
  - can we automatically transform the data in the right form ?
  - can we retain “understandability” to the user ?

KEY OPEN QUESTIONS

Small is beautiful

- It is not just BIG data that matters; most datasets are small;
- Democratizing data science … bring it to the (naive) end-user;
- A frequent setting — a set of (excel or mysql) tables ?
- Steps towards automatisation; full automatisation is pretty wild; no fear for data scientists losing jobs in the near future …
- Many interesting questions still open …