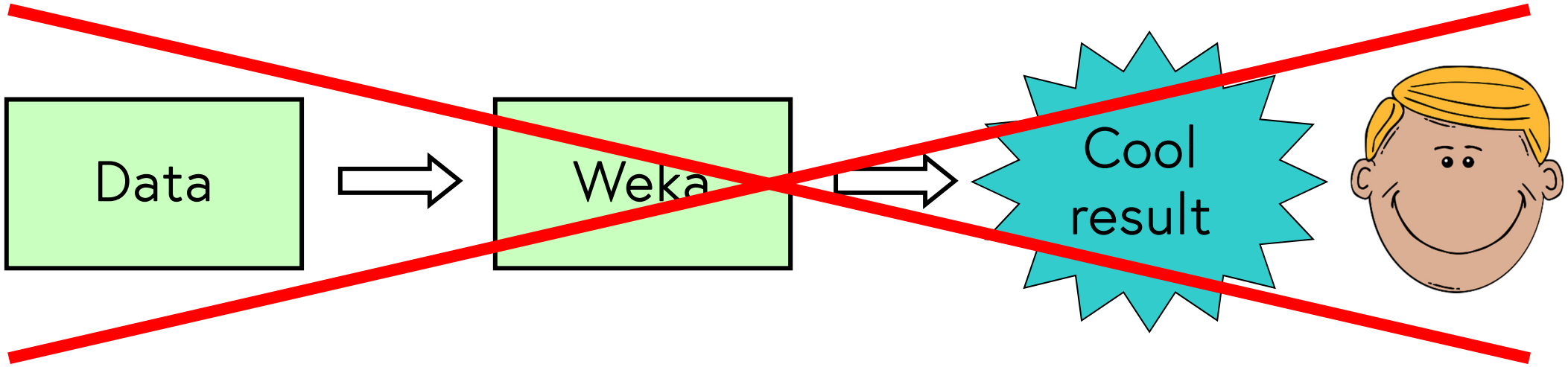
Data Mining with Weka

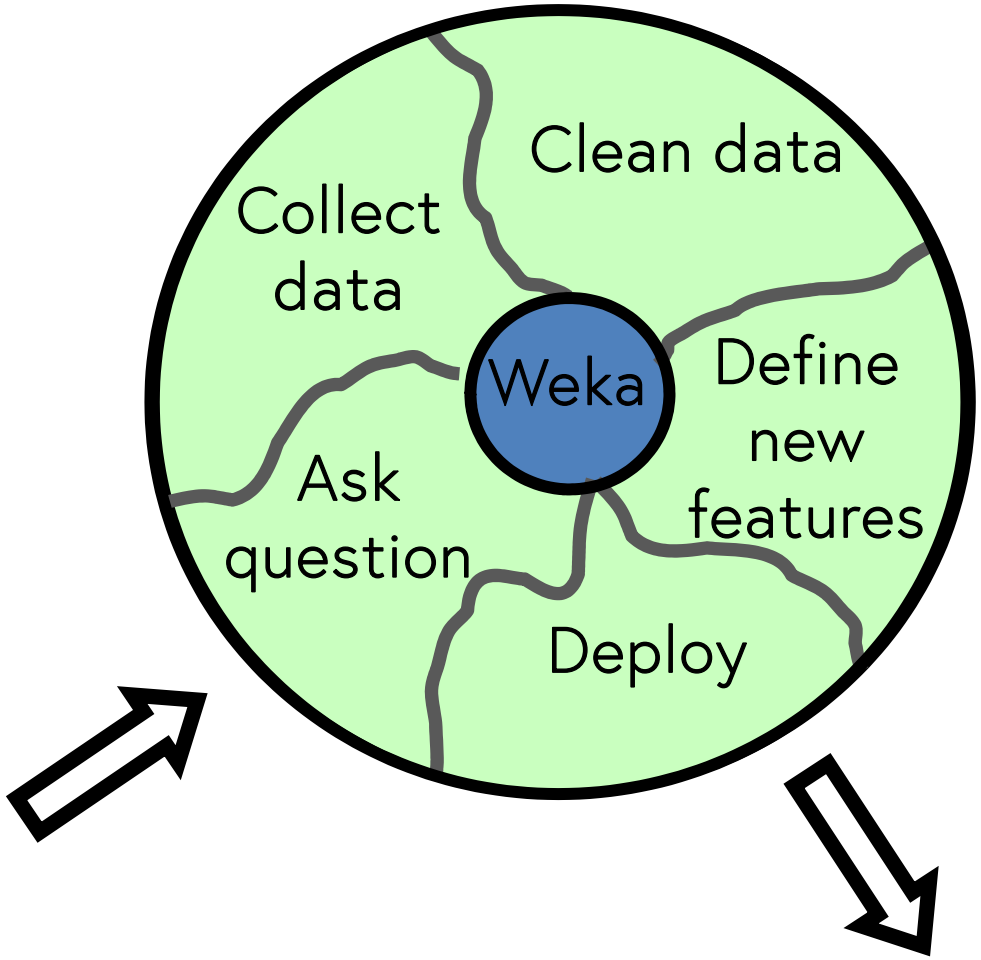
The data mining process

Ian H. Witten

The data mining process



The data mining process



The data mining process

- Ask a question
  - *what do you want to know?*
  - *"tell me something cool about the data" is not enough!*
- Gather data
  - *there's soooo much around …*
  - *… **but** … we need (expert?) classifications*
  - *more data beats a clever algorithm*
- Clean the data
  - *real data is very mucky*
- Define new features
  - *feature engineering—the key to data mining*
- Deploy the result
  - *technical implementation*
  - *convince your boss!*

The data mining process

(Selected) filters for feature engineering

- AddExpression (MathExpression)
  Apply a math expression to existing attributes to create new one (or modify existing one)
- Center (Normalize) (Standardize)
  - *Transform numeric attributes to have zero mean (or into a given numeric range) (or to have zero mean and unit variance)*
- Discretize (also supervised discretization)
  - *Discretize numeric attributes to have nominal values*
- PrincipalComponents
  - *Perform a principal components analysis/transformation of the data*
- RemoveUseless
  - *Remove attributes that do not vary at all, or vary too much*
- TimeSeriesDelta, TimeSeriesTranslate
  - *Replace attribute values with successive differences between this instance and the next*

The data mining process

- Weka is only a small part (unfortunately) …
- … and it's the easy part

  "*may all your problems be technical ones*"

  – old programmer's blessing