On TAP:

Module 3: Feature Generation

Kyle H. Ambert, PhD

Intel Big Data Solutions, Datacenter Group

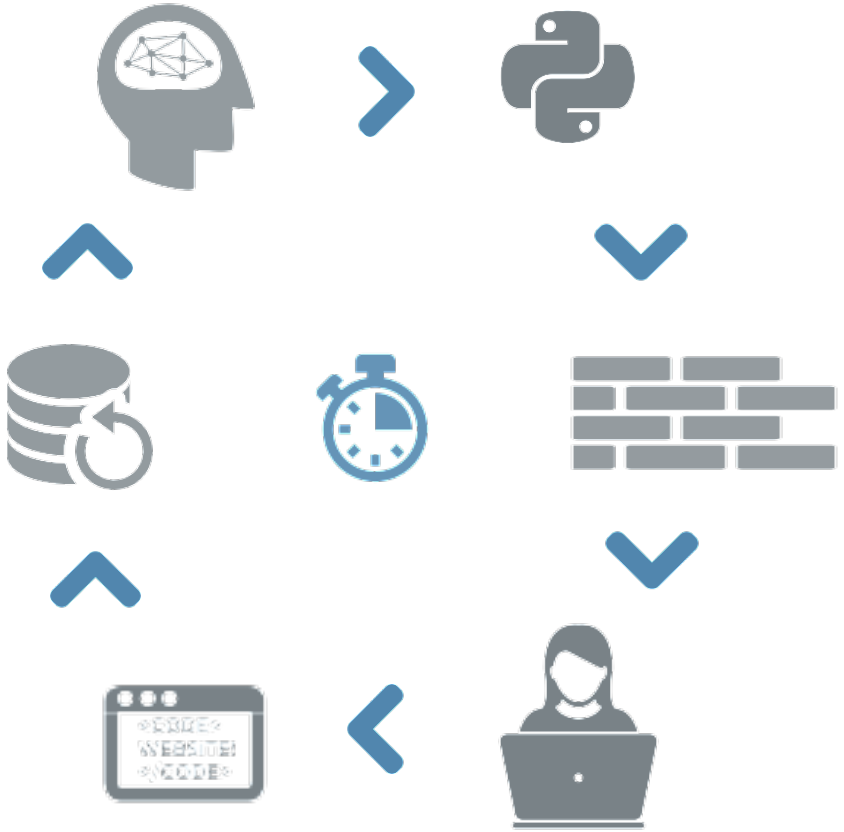
Feature Generation

- In Trusted Analytics, this frequently leverages the `add_columns` function
- As in all data science problems and platforms, the features you need will depend on the problem you're trying to solve.

Feature Generation

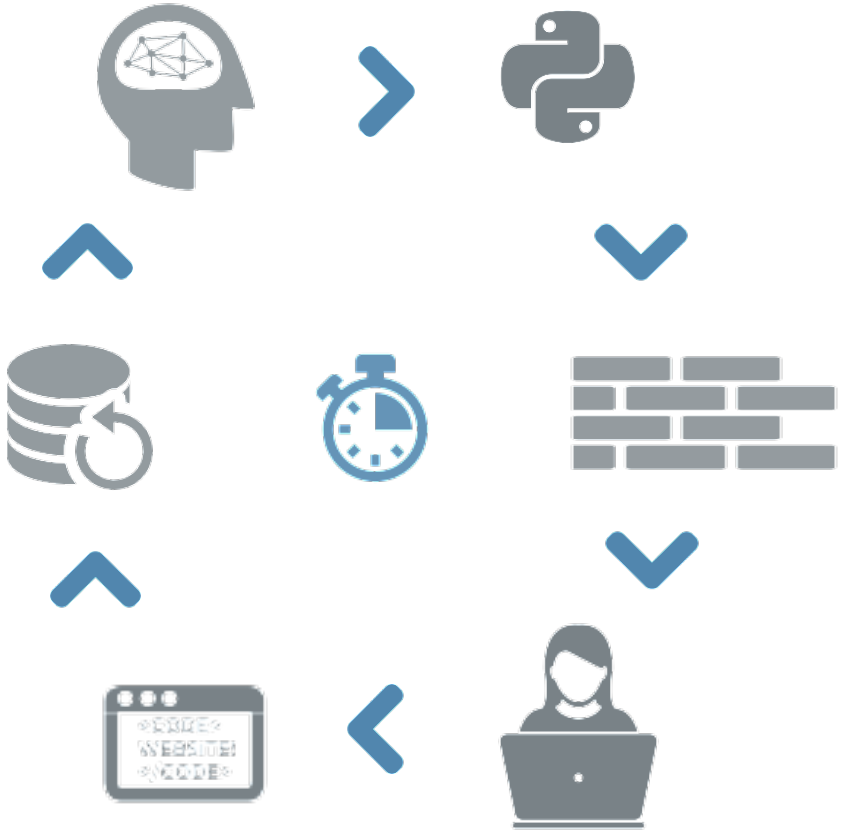
Feature Generation

- Add a column based on global computations



Feature Generation

- Add a column based on global computations
- Impute missing values



Feature Generation

- Add a column based on global computations
- Impute missing values
- Clean text data

code!

Problems:

[**1**] In the `inpat` data set there are multiple fields of the pattern 'FINAL_DIAG_CCS_LVL_n_LABELs', where *n* is a number. Use add_columns to extract these data as a single column.

[**2**] In the column updated in [1], there are sometimes codes mixed in with plain-english labels (e.g., "*Deficiency and other anemia [59.]*"). Add a second column that has removed these codes.

[**3**] Many unigrams in the labels are common, and many are potentially uninformative. Using a mixture of EDA and domain expertise, write your own stop word removal function to clean out uninformative unigrams.