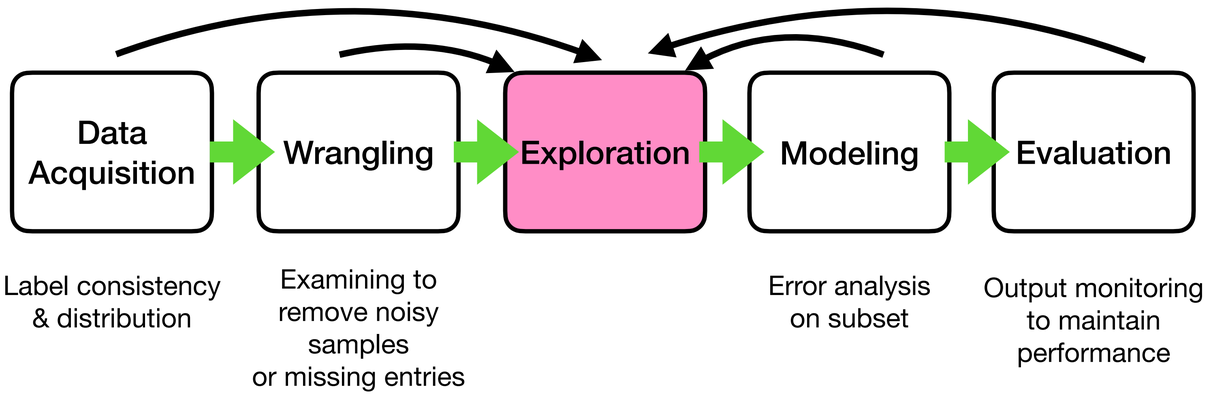
# Weedle: Composable Dashboard for Data-centric NLP in Computational Notebooks

Nahyun Kwon, Hannah Kim, Sajjadur Rahman, Dan Zhang, Estevam Hruschka

## Data-centric NLP cycle

- Data-centric NLP is highly iterative
- All steps commonly require examination & diagnosis of data



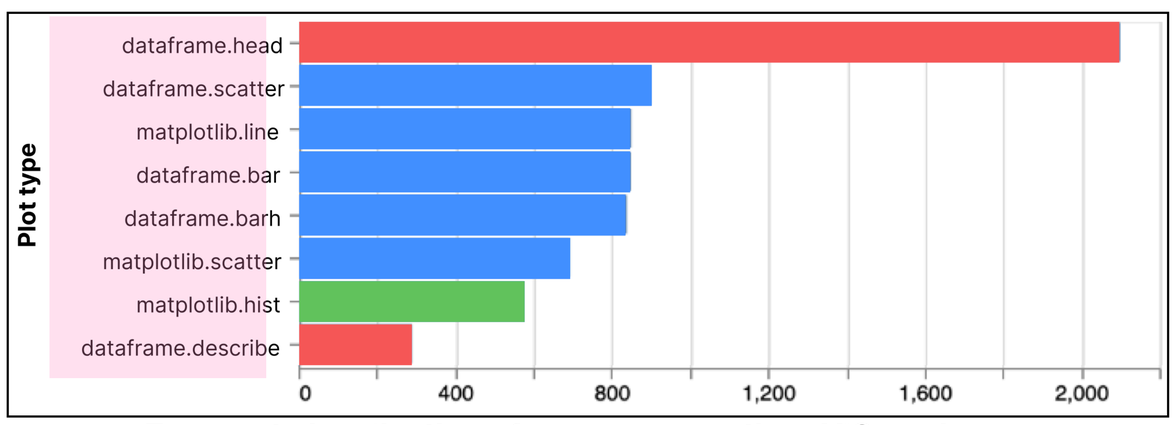

## Challenges in current EDA approaches

| Current approaches | Examples | Features | | |
|---|---|---|---|---|
| | | Text support | Customizable | Seamless |
| General EDA tools | • Tableau<br>• MS Power BI<br>• Qilk | ✗ | ○ | ✗ |
| Notebook-based EDA or Vis libraries | • dataprep.eda<br>• altair<br>• matplotlib | ✗ | ○ | ○ |
| Task-specific text exploration tools | • pyLDAvis<br>• twitter monitoring/ geographical event analysis tool | ○ | ✗ | ○ |

## Common text transformations & visualizations used in public notebooks

| Visualization | Transformation |
|---|---|
| Table view | simple overview/statistics of data, class distribution |
| Bar chart / histogram | class distribution, word count, ngram count, data item count per matching condition, document length, punctuation analysis, feature importance, embedding visualization |
| Line chart | document length, numerical trend over time |
| Scatter plot | t-SNE distribution, bivariate correlation, data item distribution, data item distribution with clustering, numerical trend over time |
| KDE plot | word count, document length |
| Pie chart | class distribution |
| Treemap | word count |

**Common transformation methods per VIS type**




**Frequent visualizations for common python VIS packages**

- Users repetitively use popular transformation & visualization methods without a centralized way that can reduce repetition
- Even with many various visualization methods available, practitioners still widely use the basic plots: bar, line, scatter plot, table view

## Weedle: seamless EDA for text with composable chart dashboard

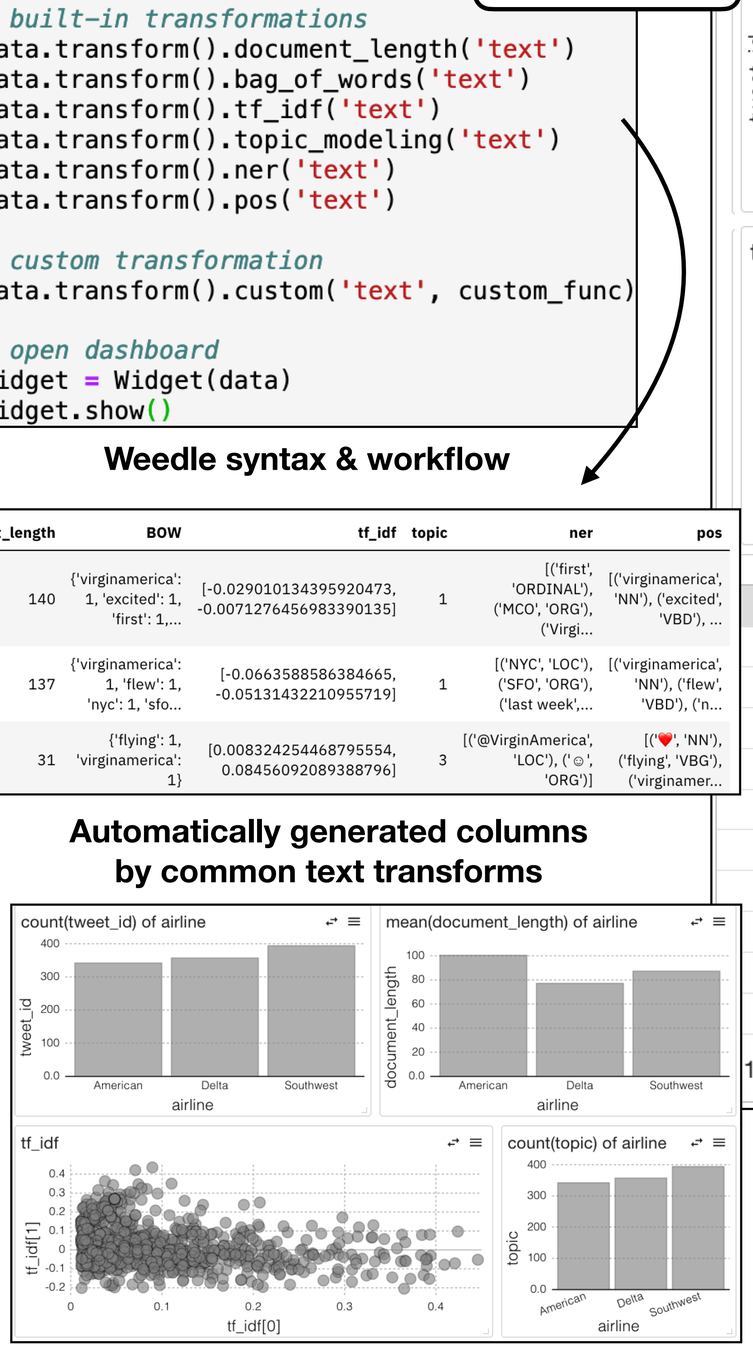
1. Seamless environment w/o context switching
2. Built-in text support
3. Composable chart dashboard
4. On-demand chart creation
5. Interactive filtering with visual components

```
from weedle import Data
from weedle import Widget
import pandas as pd

# loading data
df = pd.read_csv('./data.csv')
data = Data(df=df)

# built-in transformations
data.transform().document_length('text')
data.transform().bag_of_words('text')
data.transform().tf_idf('text')
data.transform().topic_modeling('text')
data.transform().ner('text')
data.transform().pos('text')

# custom transformation
data.transform().custom('text', custom_func)

# open dashboard
widget = Widget(data)
widget.show()
```

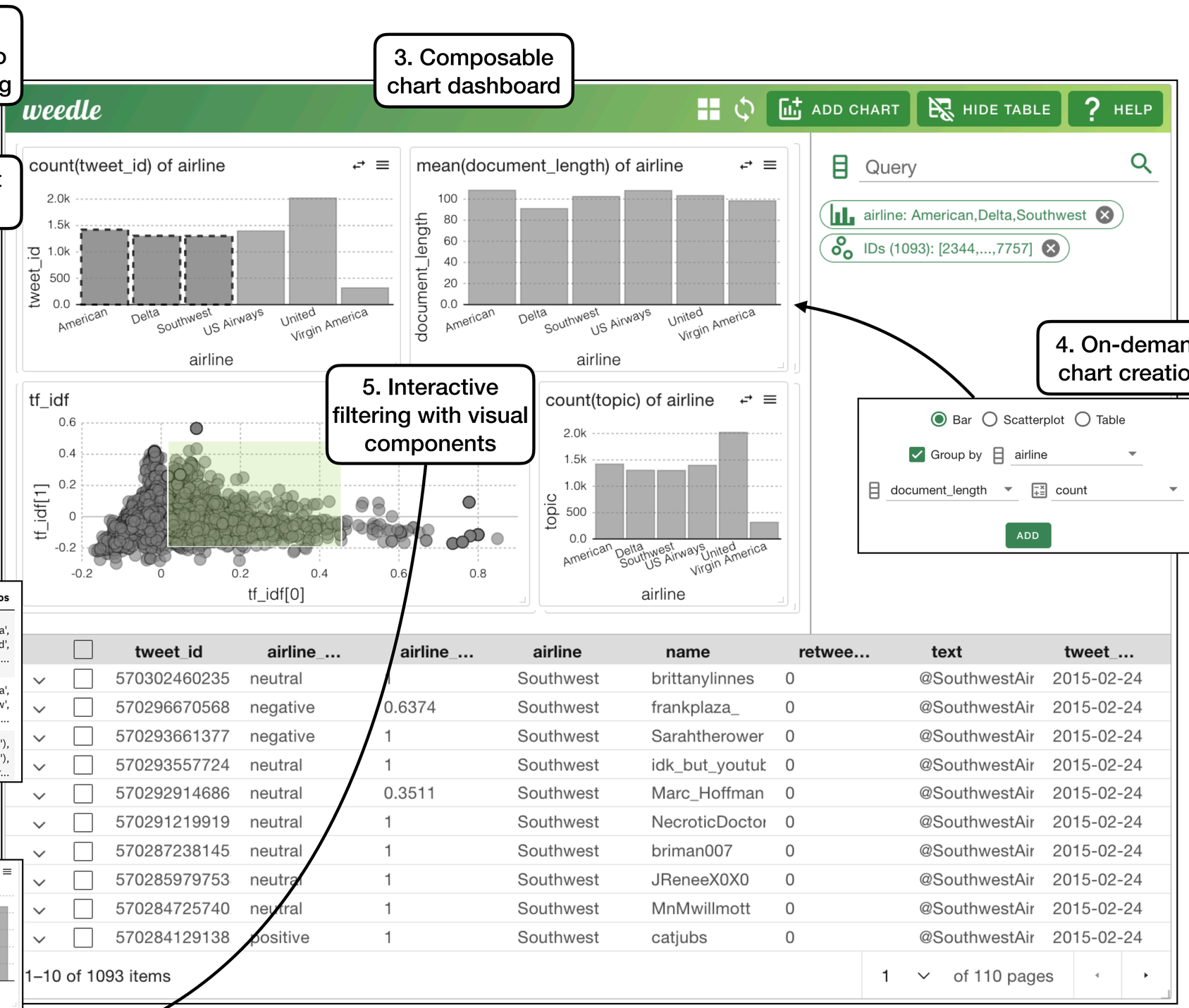
**Weedle syntax & workflow**

| document_length | BOW | tf_idf | topic | ner | pos |
|---|---|---|---|---|---|
| 140 | {'virginamerica': 1, 'excited': 1, 'first': 1,... | [-0.029010134395920473, -0.0071276456983390135] | 1 | [('first', 'ORDINAL'), ('MCO', 'ORG'), ('Virgi... | [('virginamerica', 'NN'), ('excited', 'VBD'), ... |
| 137 | {'virginamerica': 1, 'flew': 1, 'nyc': 1, 'sfo... | [-0.0663588586384665, -0.05131432210955719] | 1 | [('NYC', 'LOC'), ('SFO', 'ORG'), ('last week',... | [('virginamerica', 'NN'), ('flew', 'VBD'), ('n... |
| 31 | {'flying': 1, 'virginamerica': 1} | [0.008324254468795554, 0.08456092089388796] | 3 | [('@VirginAmerica', 'LOC'), ('☺', 'ORG')] | [('❤', 'NN'), ('flying', 'VBG'), ('virginamer... |

**Automatically generated columns by common text transforms**




**Filters update all charts**




**Customizable & interactive dashboard**