Data Curation

Never Stand Still

Computer Science and Engineering

Alireza Tabebordbar

Comp 9321

Lecturer: ***Helen Paik***

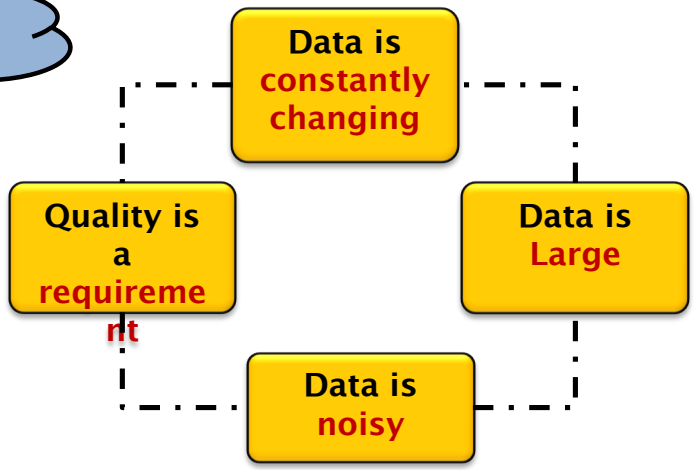
Overview

- Data Curation
- Data Curation Approaches
  - Algorithmic Data Curation
  - Rule Based Data Curation
  - Hybrid Data Curation
- Toward Automated Data Curation

Data Curation

Data Curation:

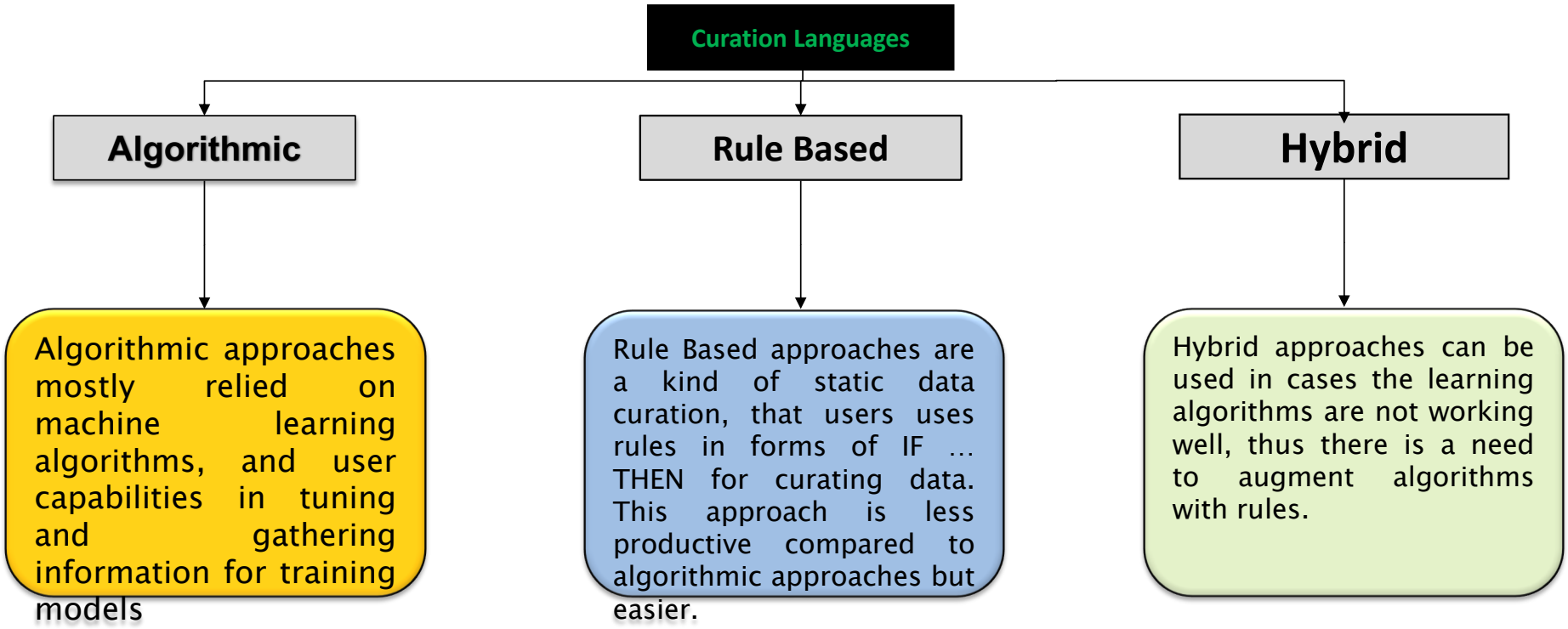
Data curation is the task of transforming raw Data into knowledge.



For Example, tweet counts in twitter:

Second ≅ 6,000 Minute ≅ 350,000 Day ≅ 500 Million Year ≅ 200 billion

Data Curation Approaches



Algorithmic Data Curation

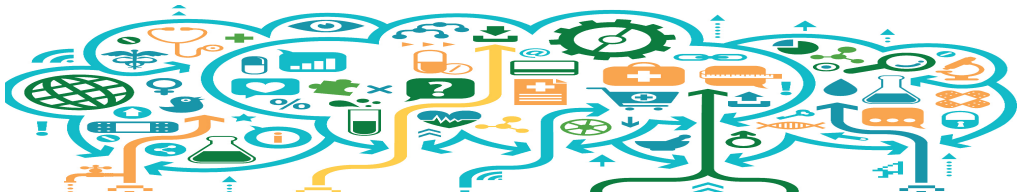
Knowledge



Processing Pipeline

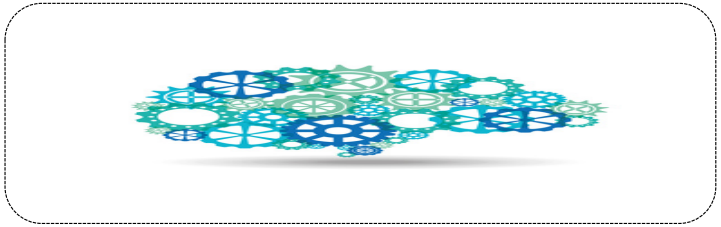
Data Processing Algorithms (e.g. SVM, Naïve Bayes)

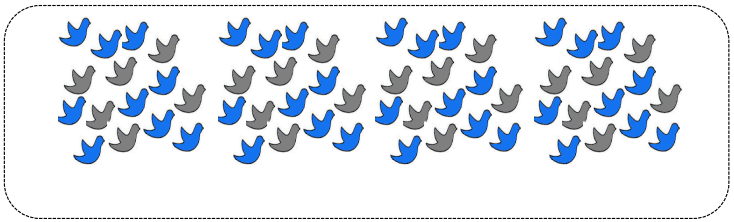
Raw Data



Algorithmic Data Curation Example

ID	Tweets	Label
1	eating is now a mental disorder in america	✓
2	men without mates face physical and mental health risks	✓
3	mental agility is the true mark of a great mind	✗





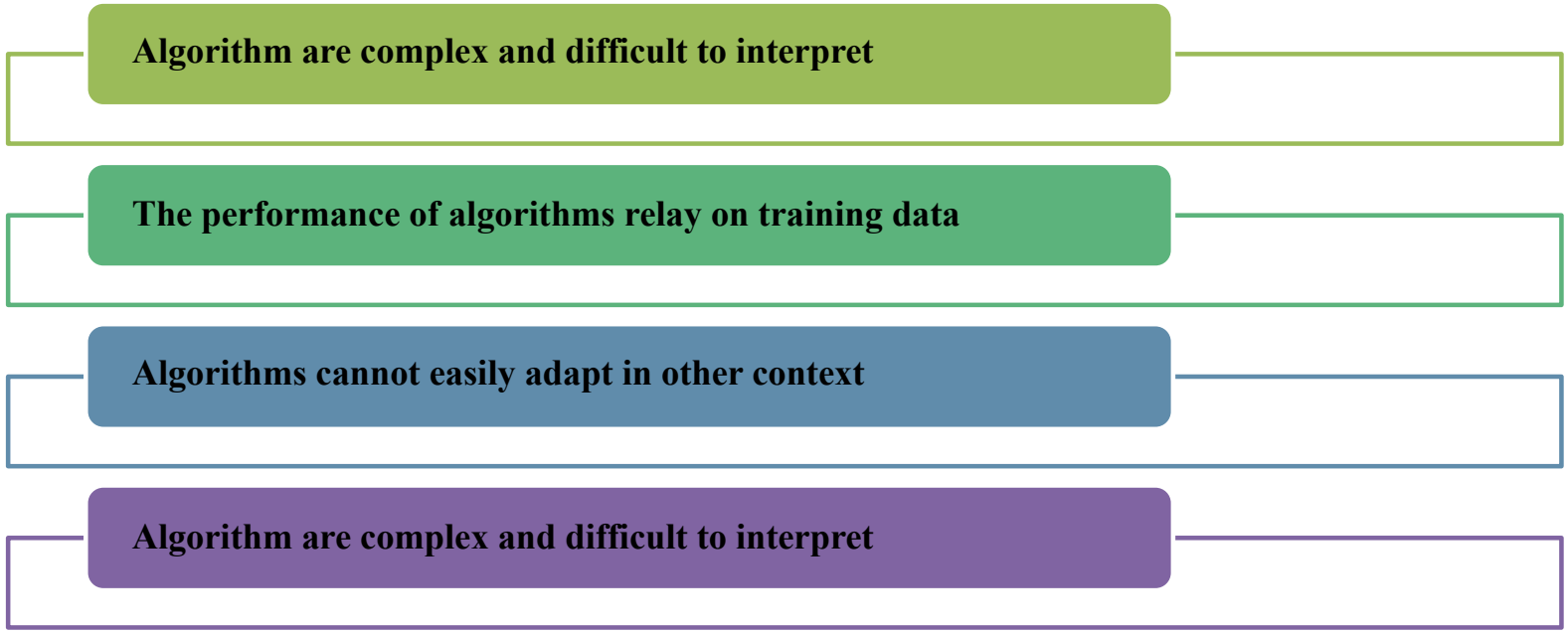
Labelling future data

Training an Algorithm using gathered data

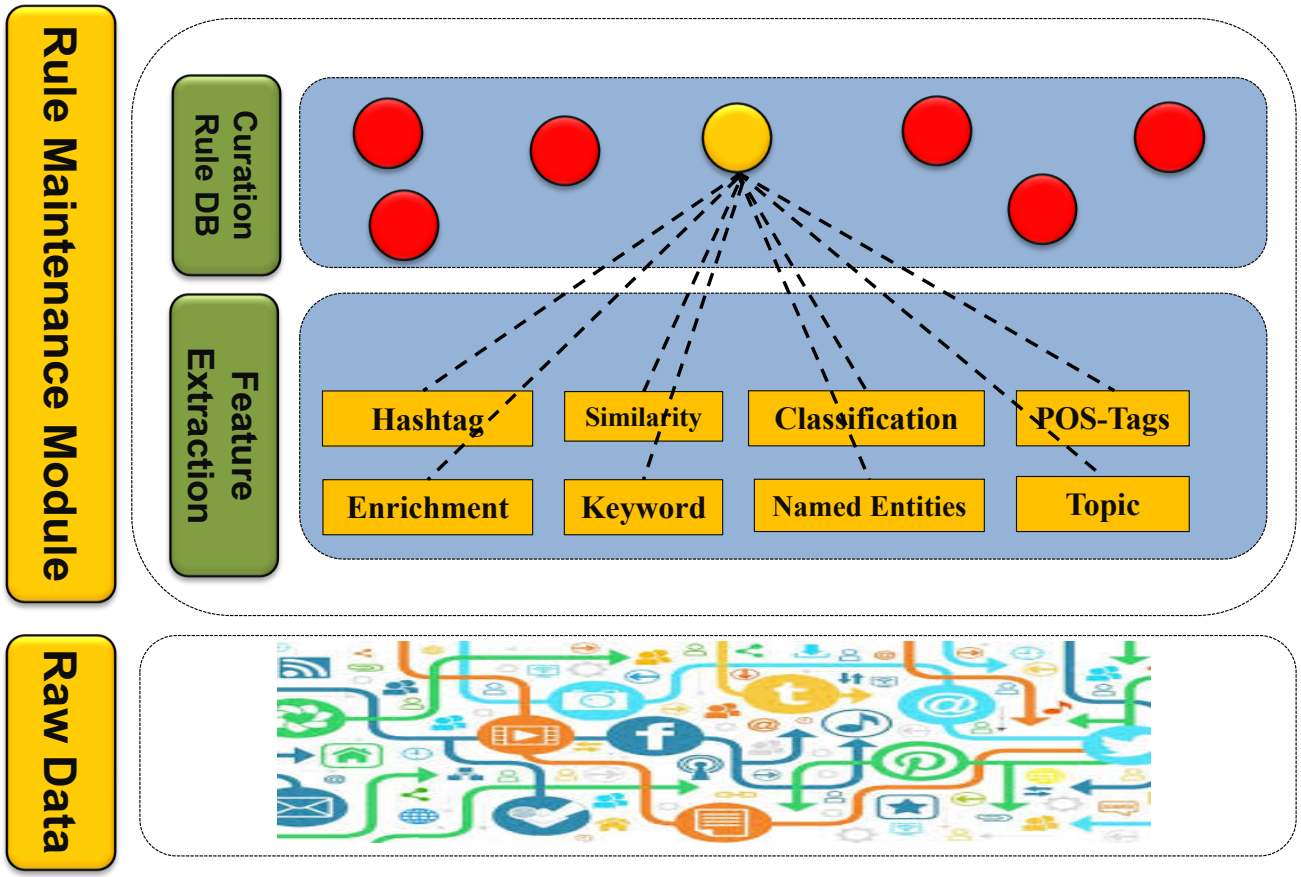
Gathering training data

Algorithmic Data Curation Challenges

- Challenges in pure algorithmic approaches:



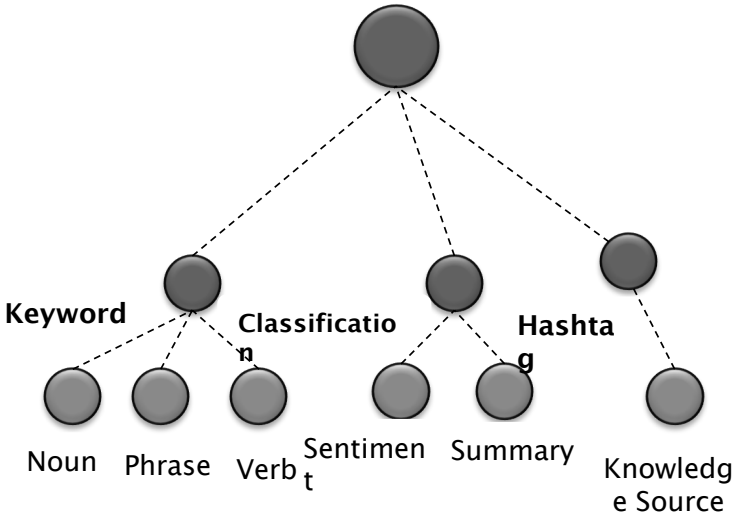
Data Curation Rule



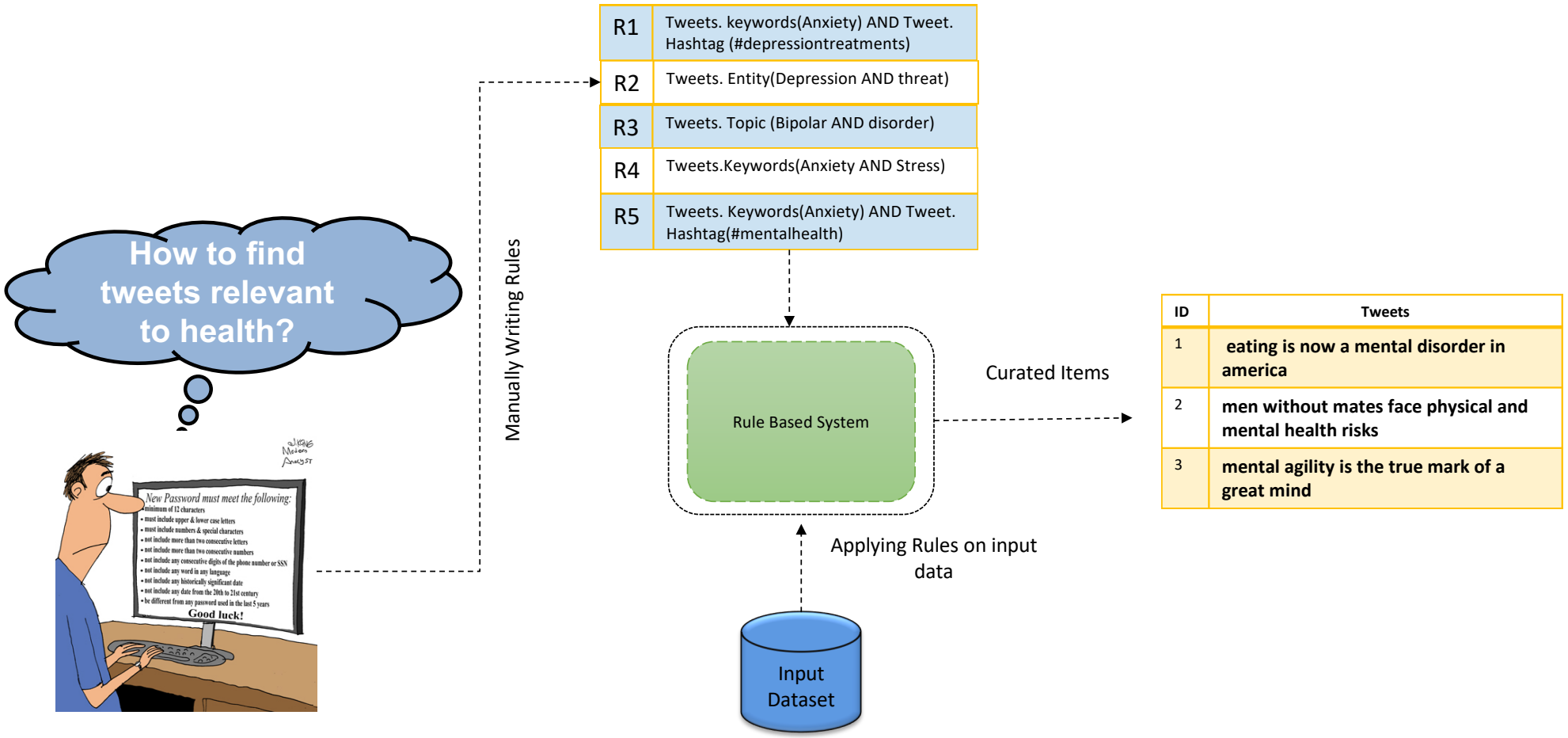
R1 : Tweet.Keyword.Contains ("autism) *AND*

Tweet.Hashtag("ADHD") *AND*

Tweet.sentimentNgative (True) = Health



Rule Based Data Curation



Rule Based Data Curation

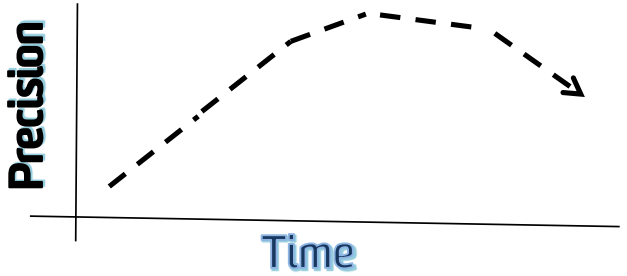
Evaluation:

- **Is the process of assessing the precision of curation rules over time.**

Adaptation:

- **The process of improving the quality of rules**

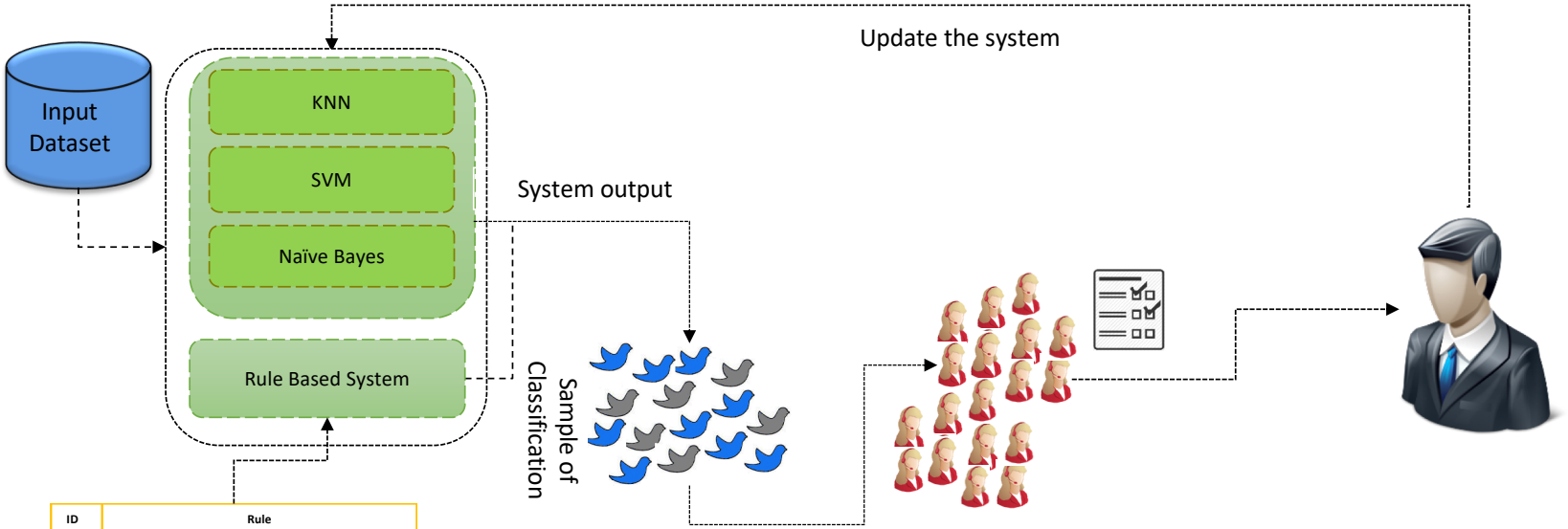
R1	Tweets. keywords(Anxiety) AND Tweet. Hashtag (#depressiontreatments)
R2	Tweets. Entity(Depression AND threat)
R3	Tweets. Topic (Bipolar AND disorder)
R4	Tweets.Keywords(Anxiety AND Stress)
R5	Tweets. Keywords(Anxiety) AND Tweet. Hashtag(#mentalhealth)



The universe of data is constantly changing. Thereby there is a need to monitor the precision of curation rules

Hybrid Data Curation

In this approach, the system relies on machine learning, rules, and crowdsourcing



ID	Rule
R1	Tweets. keywords(Anxiety) AND Tweet. Hashtag (#depressiontreatments)
R2	Tweets. Keywords(Depression AND threat)
R3	Tweets. Keywords (Bipolar AND disorder)
R4	Tweets. Keywords(Anxiety AND Stress)
R5	Tweets. Keywords(Anxiety) AND Tweet. Hashtag(#mentalhealth)

GC, P.S., Sun, C., Zhang, H., Yang, F., Rampalli, N., Prasad, S., Arcaute, E., Krishnan, G., Deep, R., Raghavendra, V. and Doan, A., 2015, May. Why big data industrial systems need rules and what we can do about it. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 265-276). ACM.

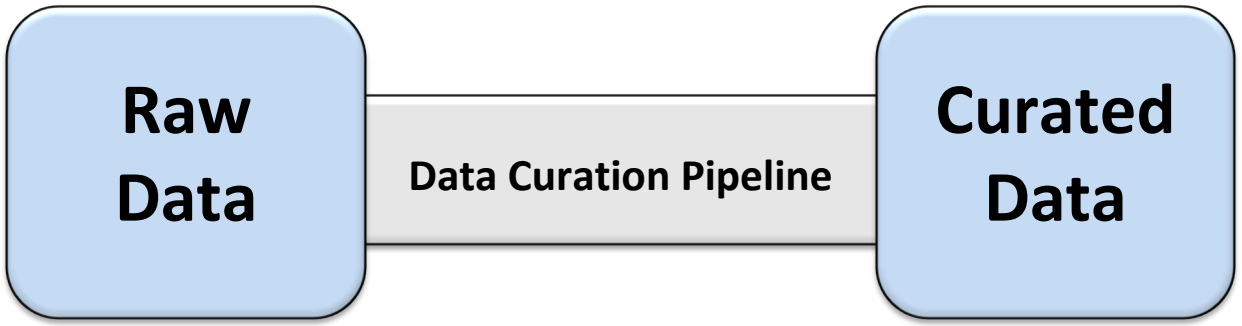
Hybrid Data Curation Challenges

We briefly can mention to following problems:

- **Expensive, as relied on analysts, and needs to be paid**
- **Time consuming, debugging learning algorithms and rules**
- **Challenging, Needs to know how to code and develop algorithms**

Automated Data Curation

- **Automated Data Curation intends to offloads analysts from data curation pipeline, and eases the process of data curation**

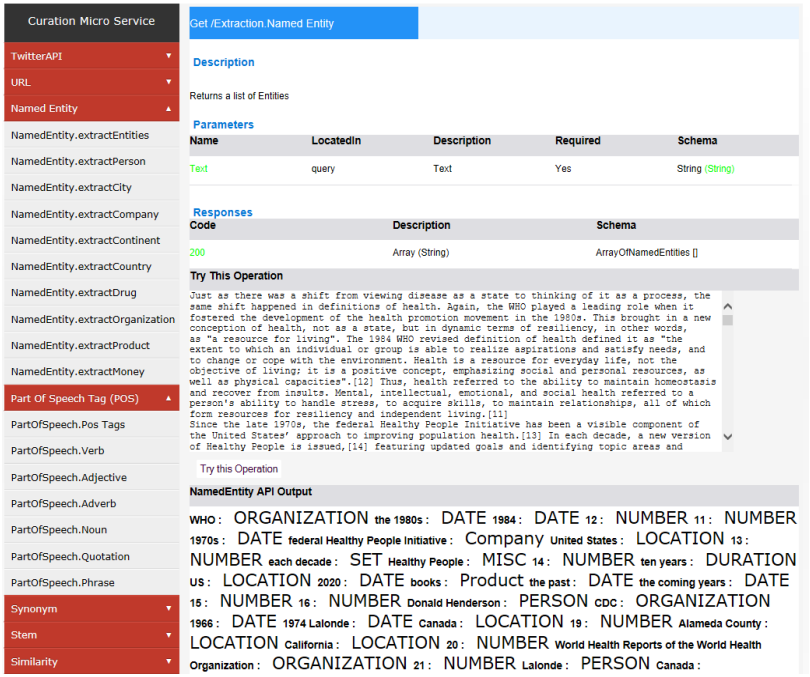


- **Makes Data curation available for masses (e.g. ordinary user without knowledge of machine learning)**
- **Makes Data curation cheaper, as requires no analysts**
- **Relieve users from challenging and time consuming curation tasks.**

On Automatic Basic Data Curation Tasks

Curation APIs:

- Named Entities
- POS-Tags
- Similarity
- URL
- Stem
- Synonym
- Keyword
- Classification



- Beheshti, Tabebordbar, Benatallah, Nouri: On Automating Basic Data Curation Tasks. WWW'2017: 165-169
- Beheshti, Tabebordbar, Benatallah, Nouri: Data Curation APIs. CoRR abs/1612.03277 (2016)