Data Curation

Alireza Tabebordbar

Comp 9321

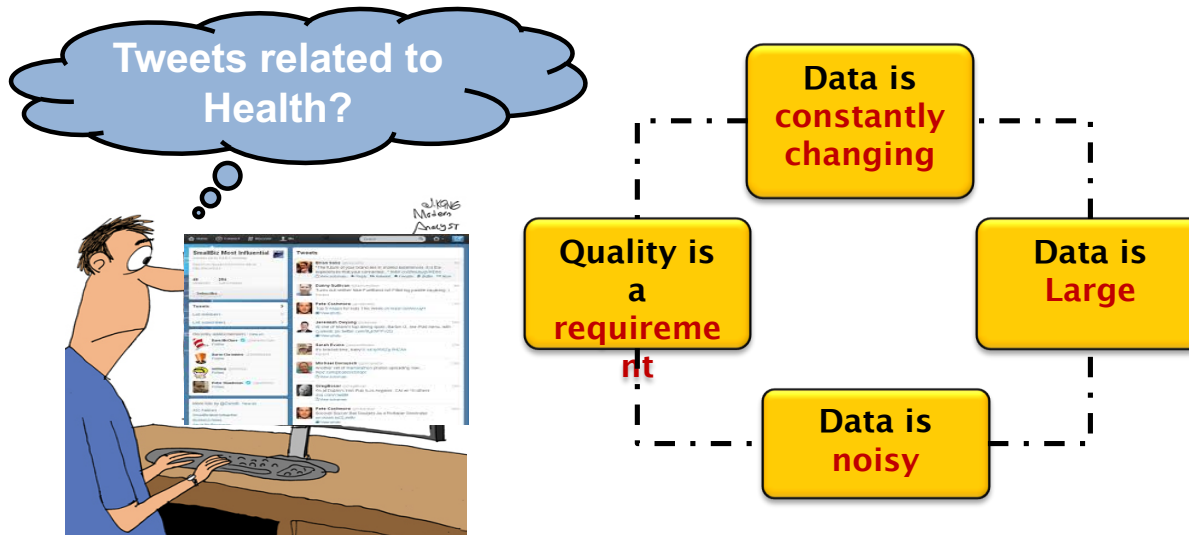*Lecturer:* **Helen Paik**

Overview

- ❑ **Data Curation**

- ❑ **Data Curation Approaches**

  - ❑ **Algorithmic Data Curation**

  - ❑ **Rule Based Data Curation**

  - ❑ **Hybrid Data Curation**

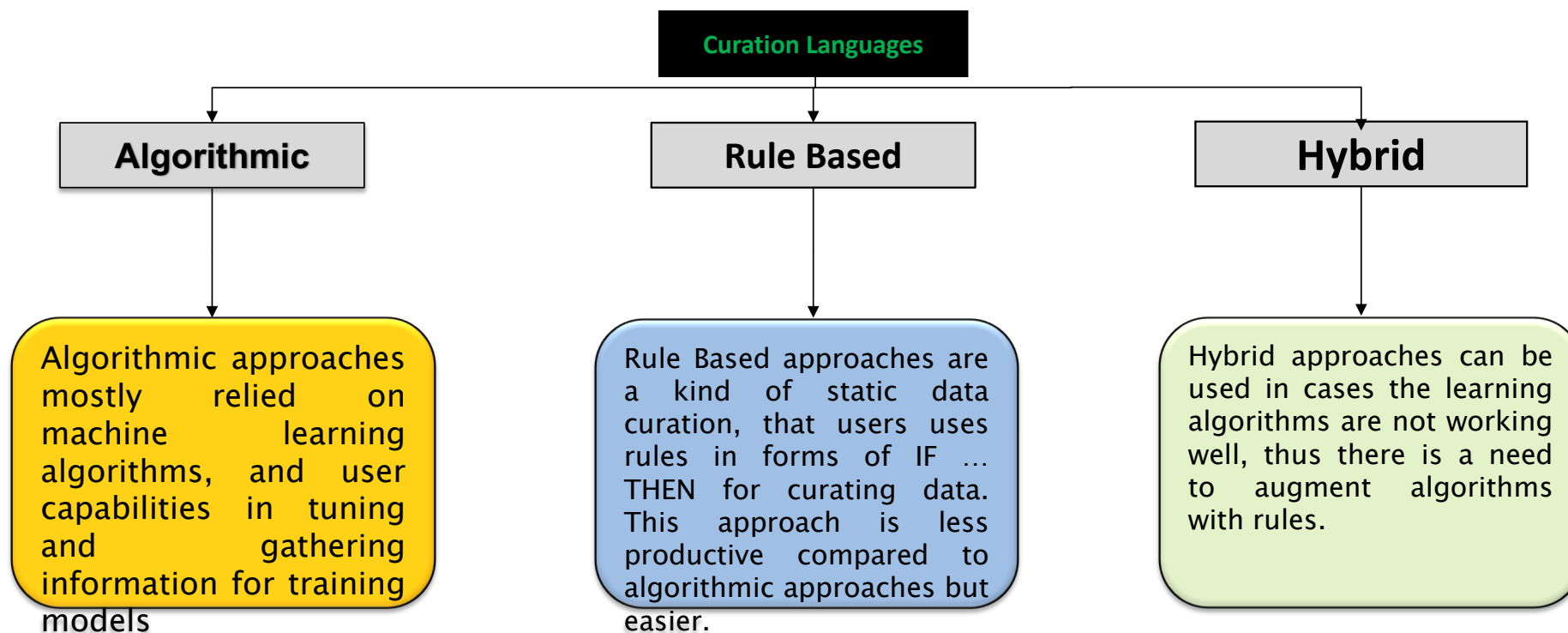- ❑ **Toward Automated Data Curation**

Data Curation

Data Curation:

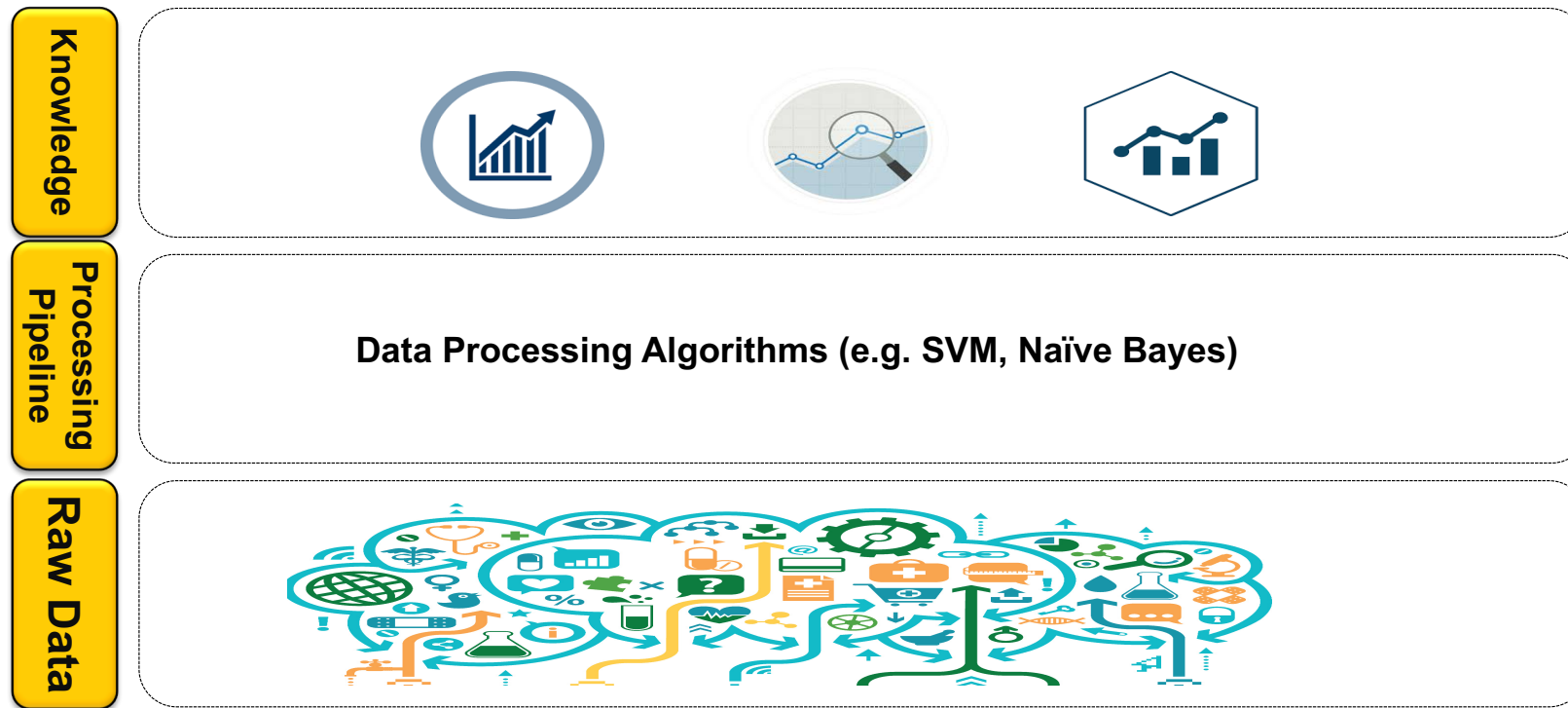Data curation is the task of transforming raw Data into knowledge.

Tweets related to Health?

Data is **constantly changing**

Quality is a **requirement**

Data is **Large**

Data is **noisy**

For Example, tweet counts in twitter:

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

Data Curation Approaches

Curation Languages

Algorithmic

Rule Based

Hybrid

Algorithmic approaches mostly relied on machine learning algorithms, and user capabilities in tuning and gathering information for training models

Rule Based approaches are a kind of static data curation, that users uses rules in forms of IF … THEN for curating data. This approach is less productive compared to algorithmic approaches but easier.

Hybrid approaches can be used in cases the learning algorithms are not working well, thus there is a need to augment algorithms with rules.

Algorithmic Data Curation

Knowledge

Processing Pipeline

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

Raw Data

Algorithmic Data Curation Example

Algorithmic Data Curation Challenges
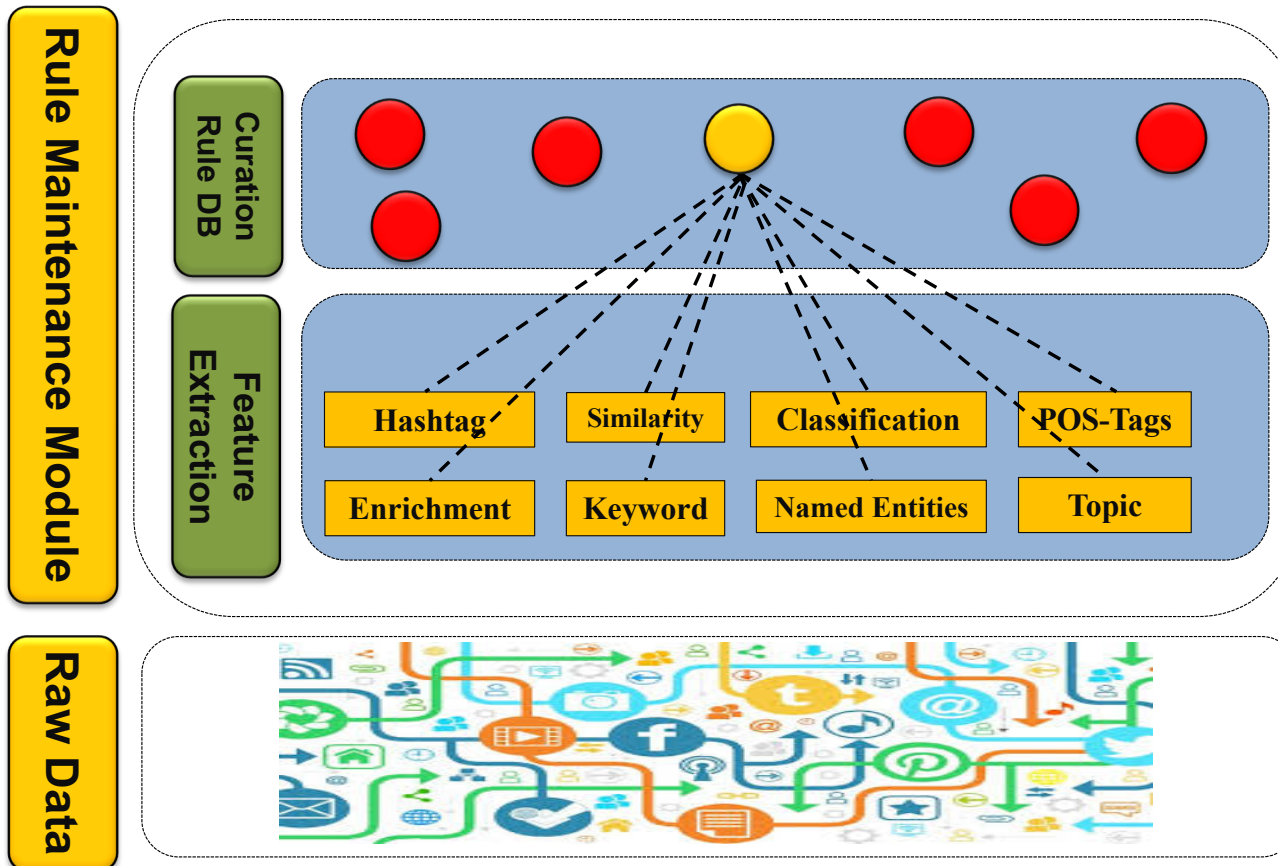
- Challenges in pure algorithmic approaches:

Algorithm are complex and difficult to interpret
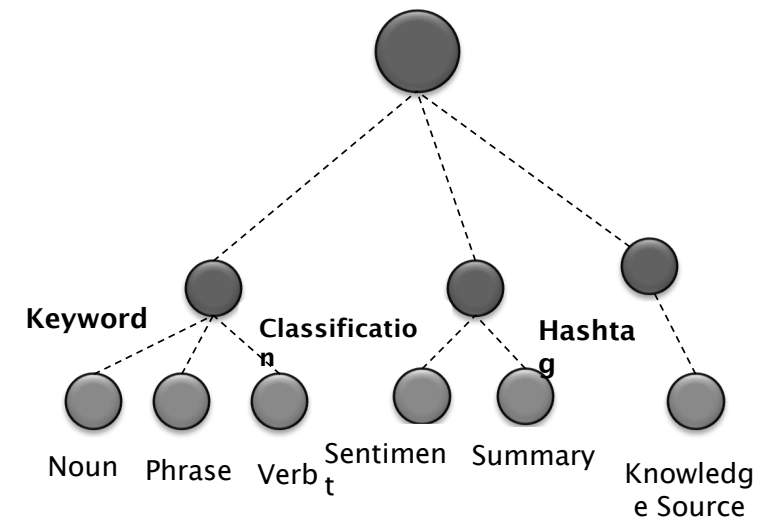
The performance of algorithms relay on training data

Algorithms cannot easily adapt in other context
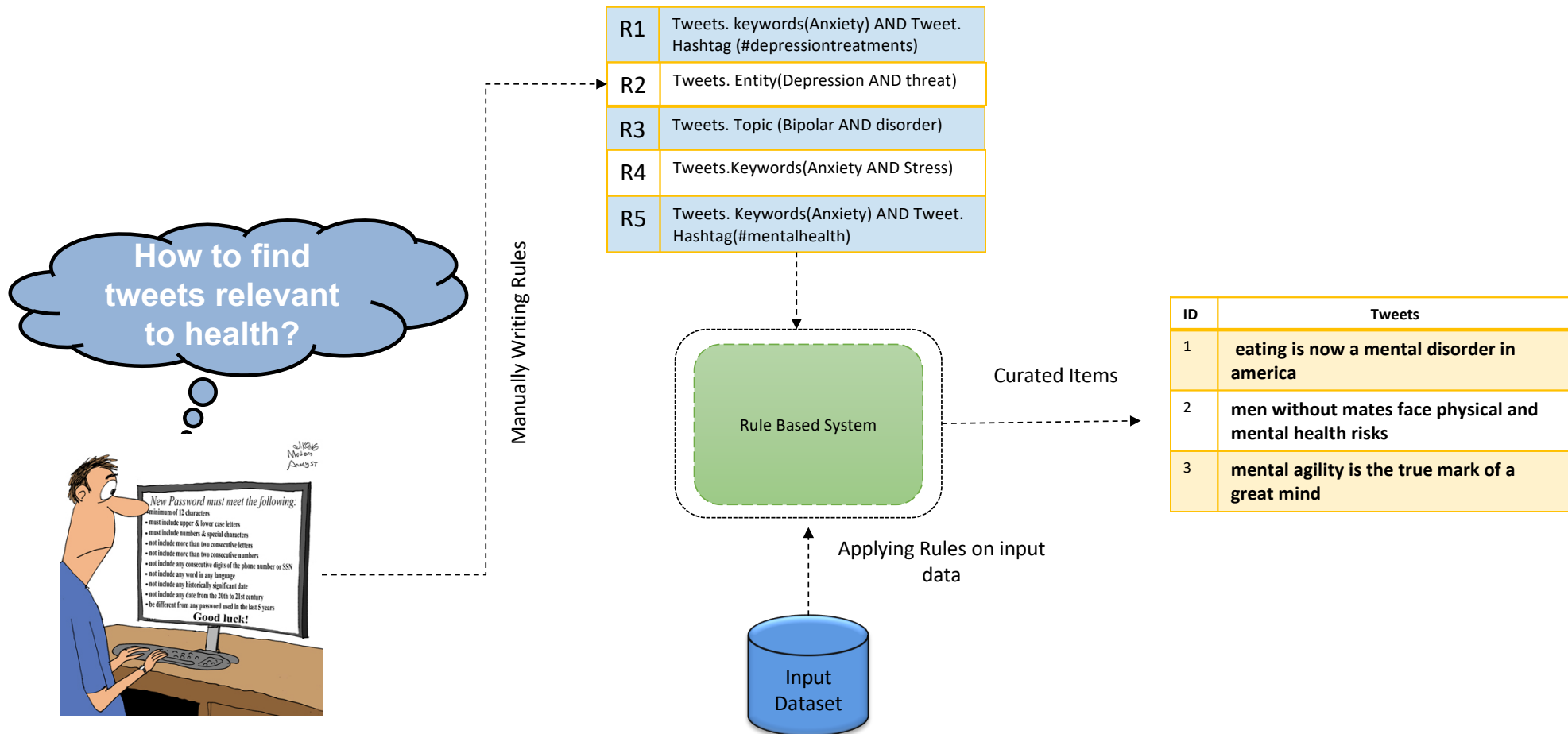
Algorithm are complex and difficult to interpret

UNSW
AUSTRALIA

Data Curation Rule

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

Tweet.Hashtag("ADHD") *AND*

Tweet.sentimentNgative (True) = Health
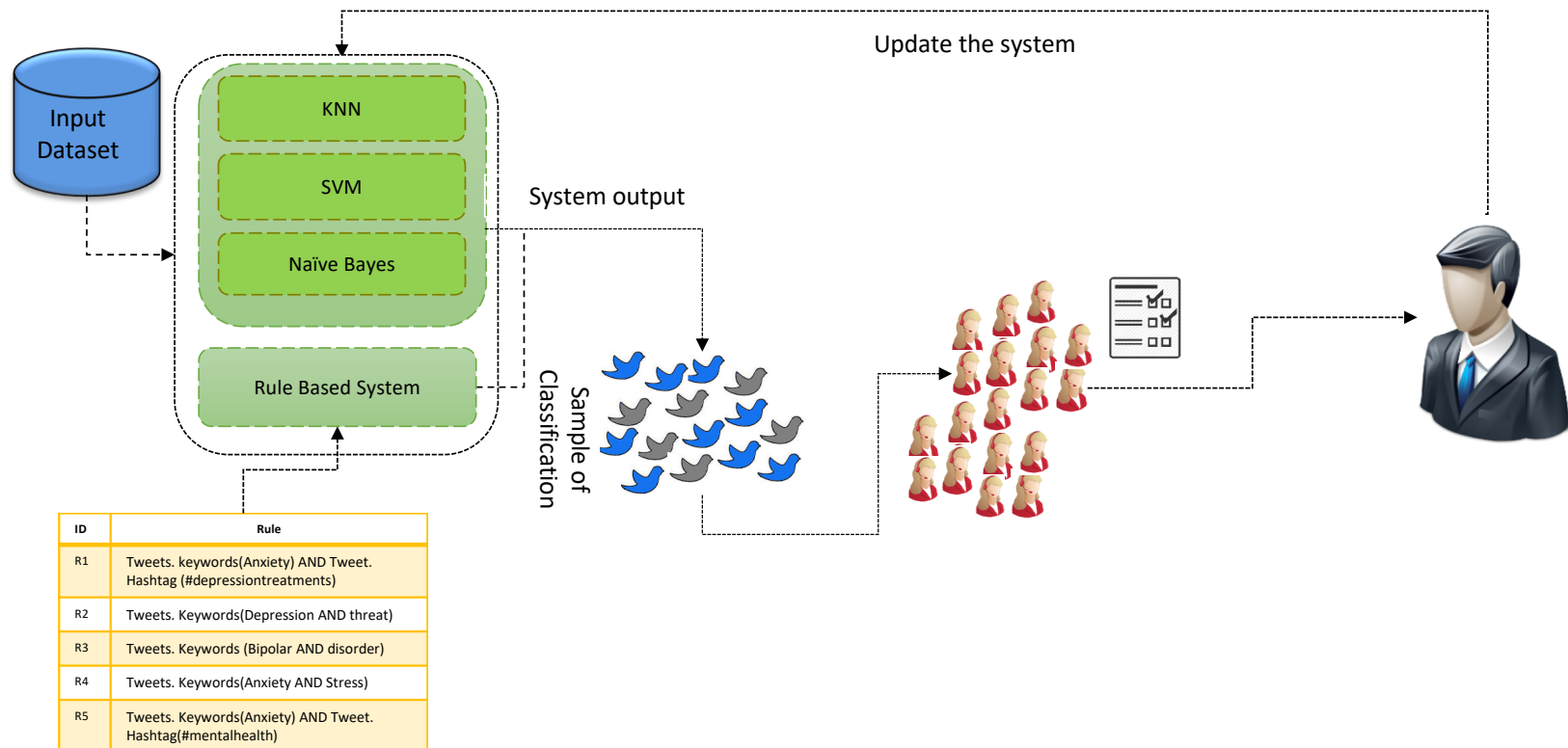
Rule Based Data Curation

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)

How to find tweets relevant to health?

Manually Writing Rules

Rule Based System

Applying Rules on input data

Input Dataset

Curated Items

ID	Tweets
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

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)

Needs to find out how well rules operate

Precision

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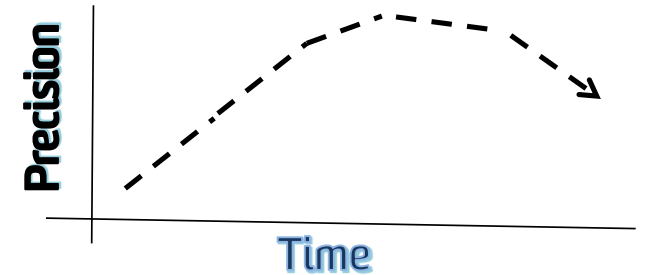
The universe of data is constantly changing. Thereby there is a need to monitor the precision of curation rules

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)

Diagram labels: Input Dataset, KNN, SVM, Naïve Bayes, Rule Based System, System output, Sample of Classification, Update the system

UNSW AUSTRALIA

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
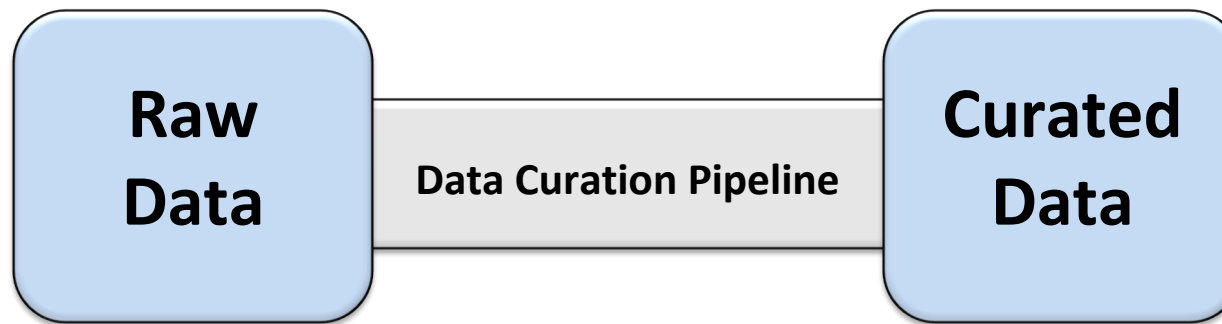
- Makes Data curation available for masses (e.g. ordinary user without knowledge of machine learning)

- 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)

UNSW
AUSTRALIA