



Data Curation

Never Stand Still

Computer Science and Engineering

Alireza Tabebordbar

Comp 9321

Lecturer: Helen Paik

Overview

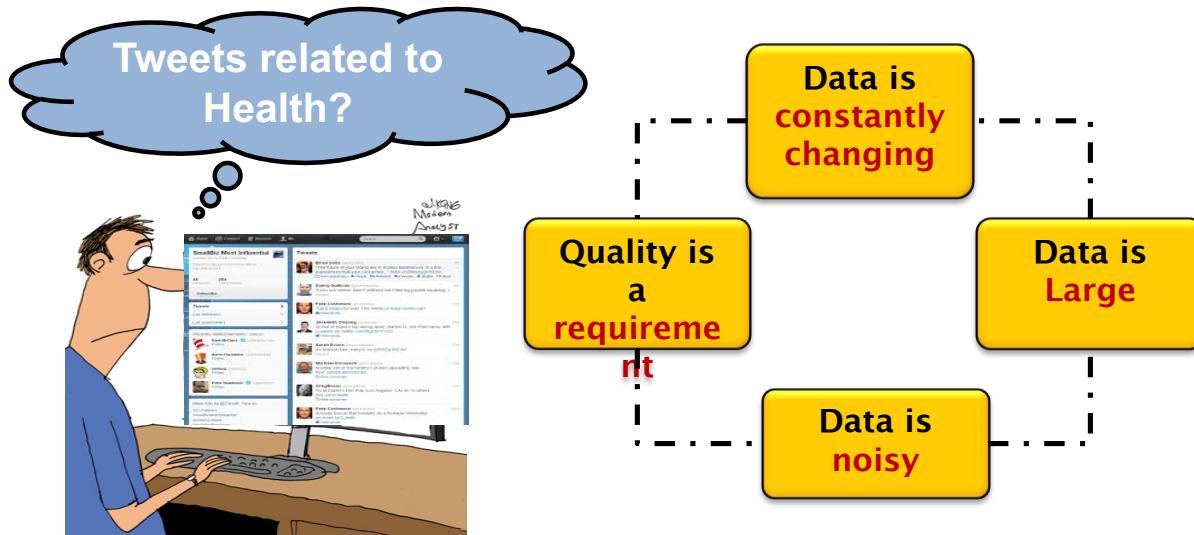
- ❑ **Data Curation**
- ❑ **Data Curation Approaches**
 - ❑ **Algorithmic Data Curation**
 - ❑ **Rule Based Data Curation**
 - ❑ **Hybrid Data Curation**
- ❑ **Toward Automated Data Curation**



Data Curation

Data Curation:

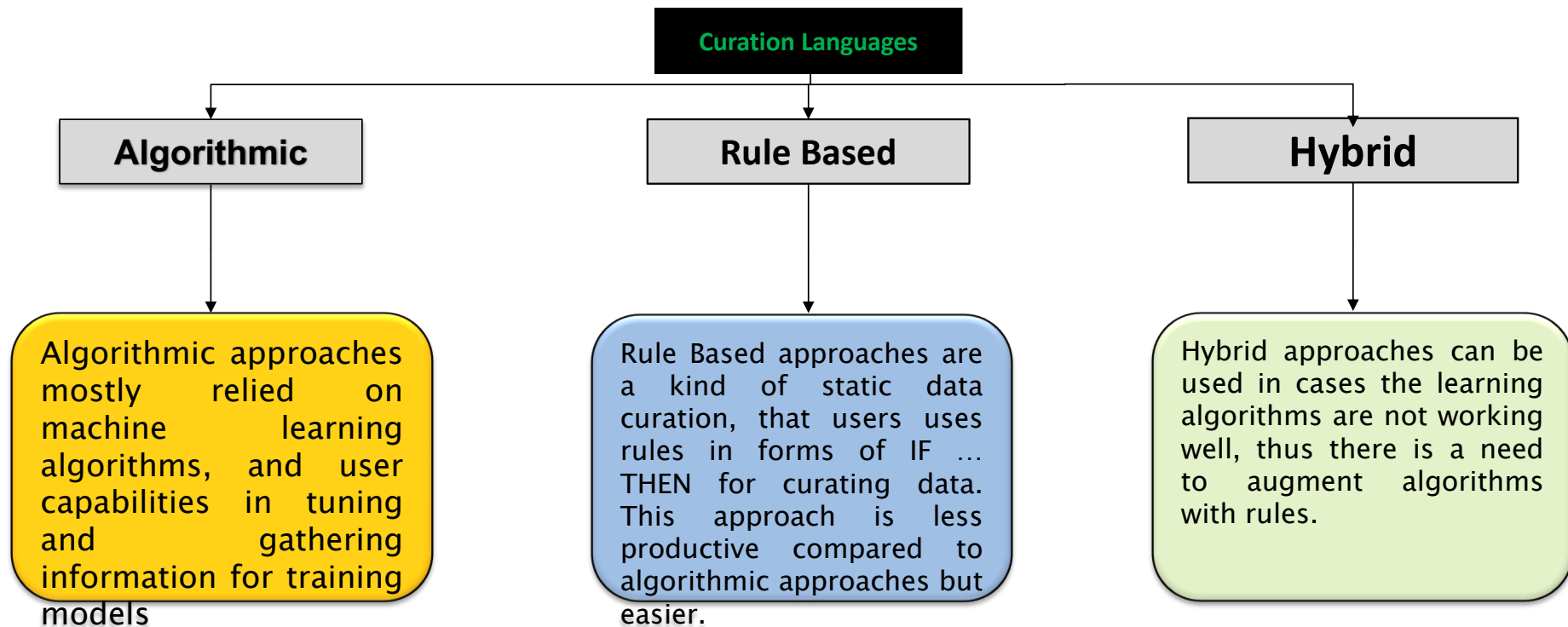
Data curation is the task of transforming raw Data into knowledge.



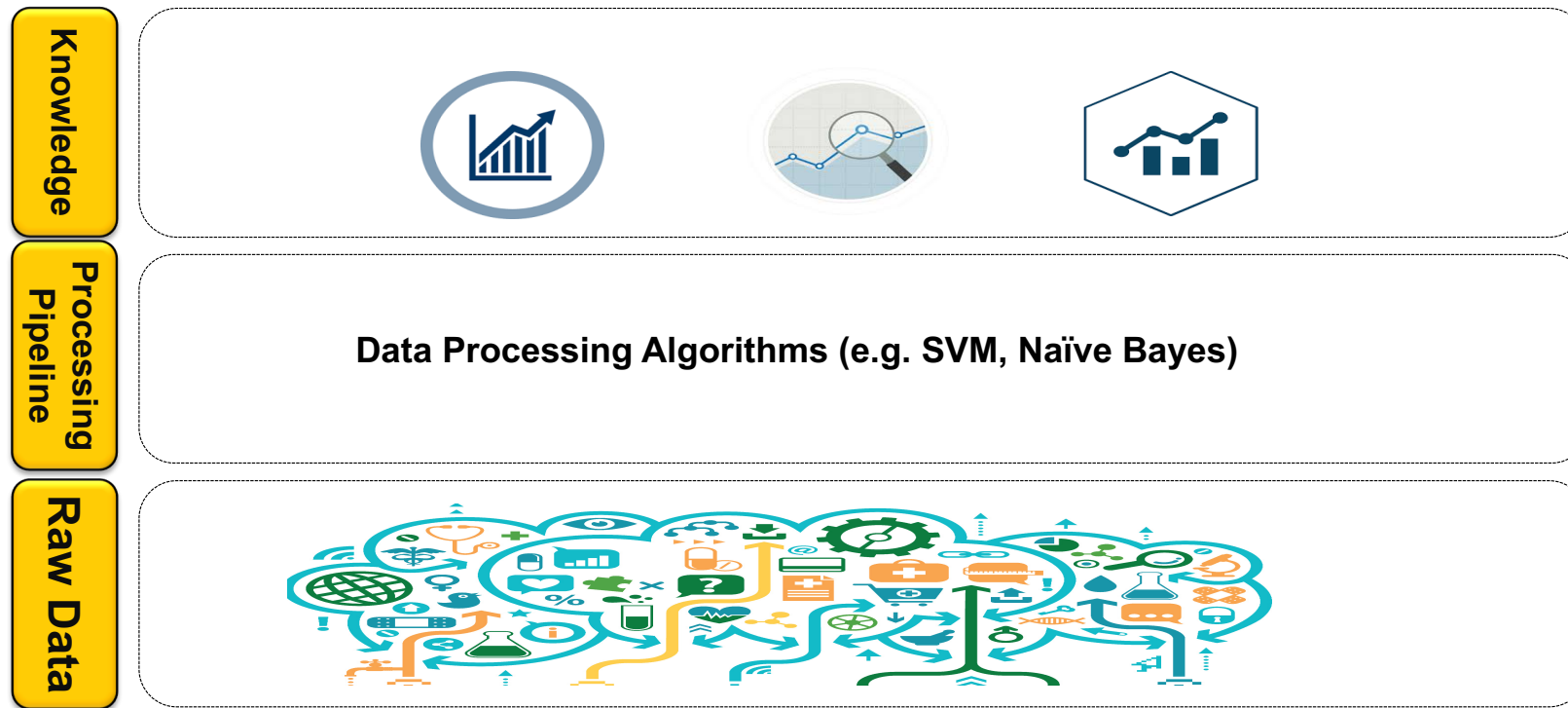
For Example, tweet counts in twitter:

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

Data Curation Approaches



Algorithmic Data Curation



Algorithmic Data Curation Example

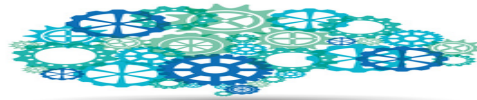


ID	Tweets	Label
1	eating is now a mental disorder in america	✓
2	men without mates face physical and mental health risks	✓
3	mental agility is the true mark of a great mind	✗

Labelling future data

Training an Algorithm using gathered data

Gathering training data



Algorithmic Data Curation Challenges

- Challenges in pure algorithmic approaches:

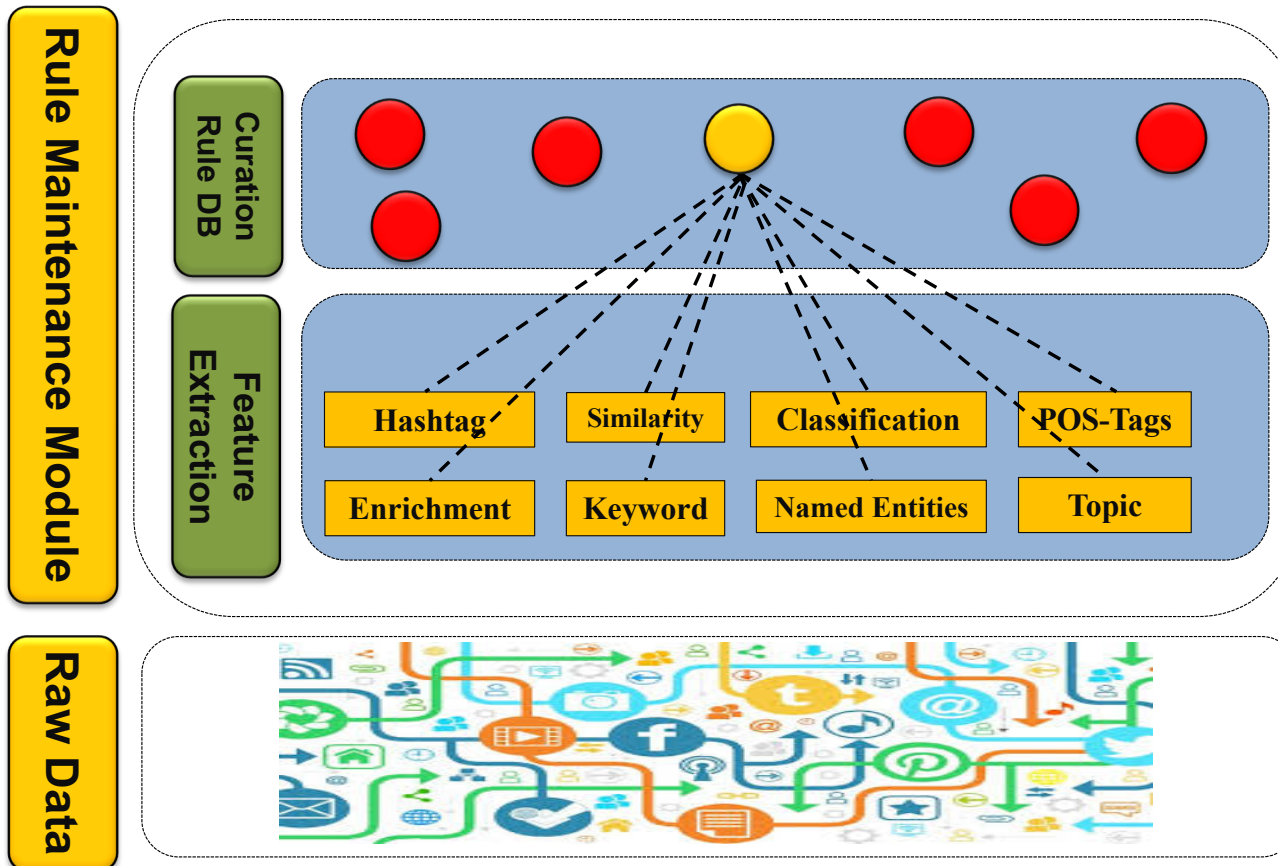
Algorithm are complex and difficult to interpret

The performance of algorithms relay on training data

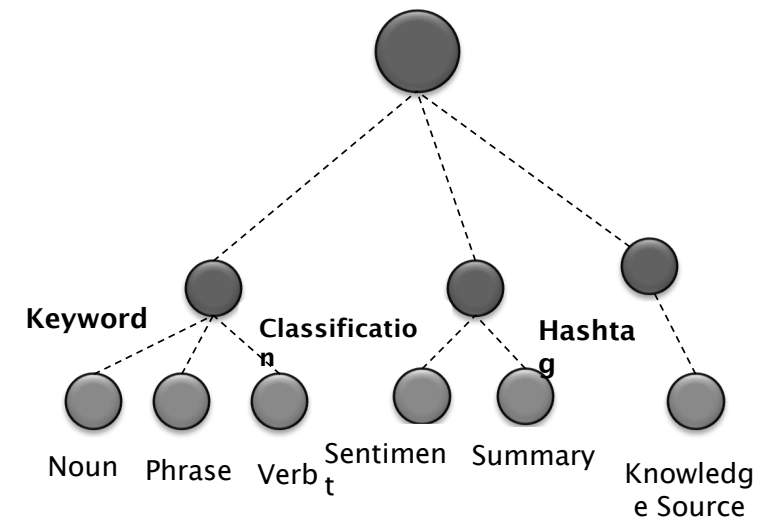
Algorithms cannot easily adapt in other context

Algorithm are complex and difficult to interpret

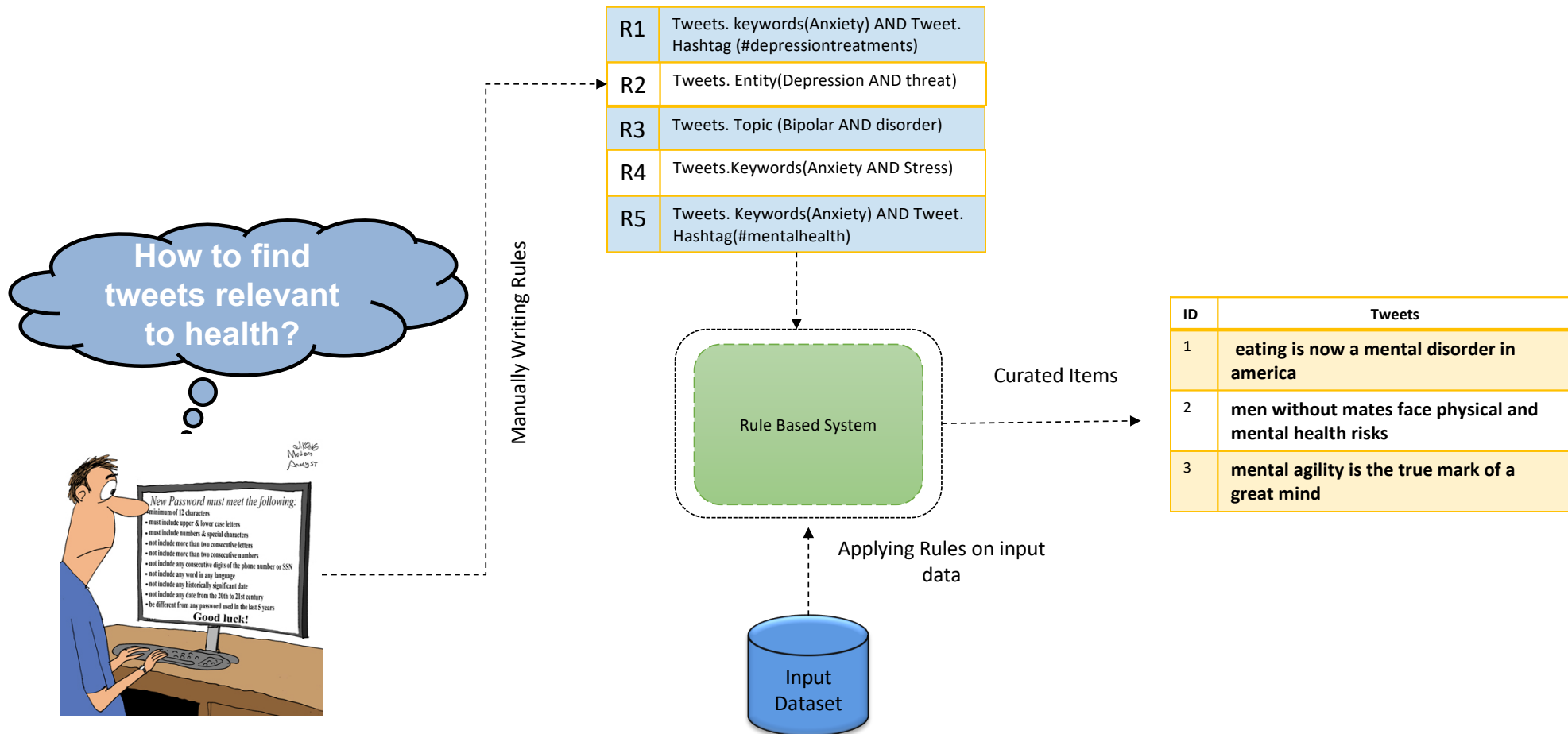
Data Curation Rule



R1 : Tweet.Keyword.Contains ("autism) AND
Tweet.Hashtag("ADHD") AND
Tweet.sentimentNgative (True) = Health



Rule Based Data Curation



Rule Based Data Curation

Evaluation:

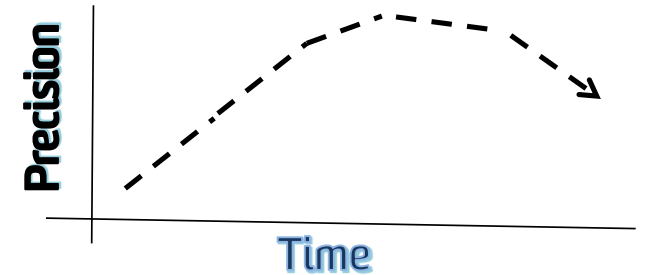
- Is the process of assessing the precision of curation rules over time.

Adaptation:

- The process of improving the quality of rules



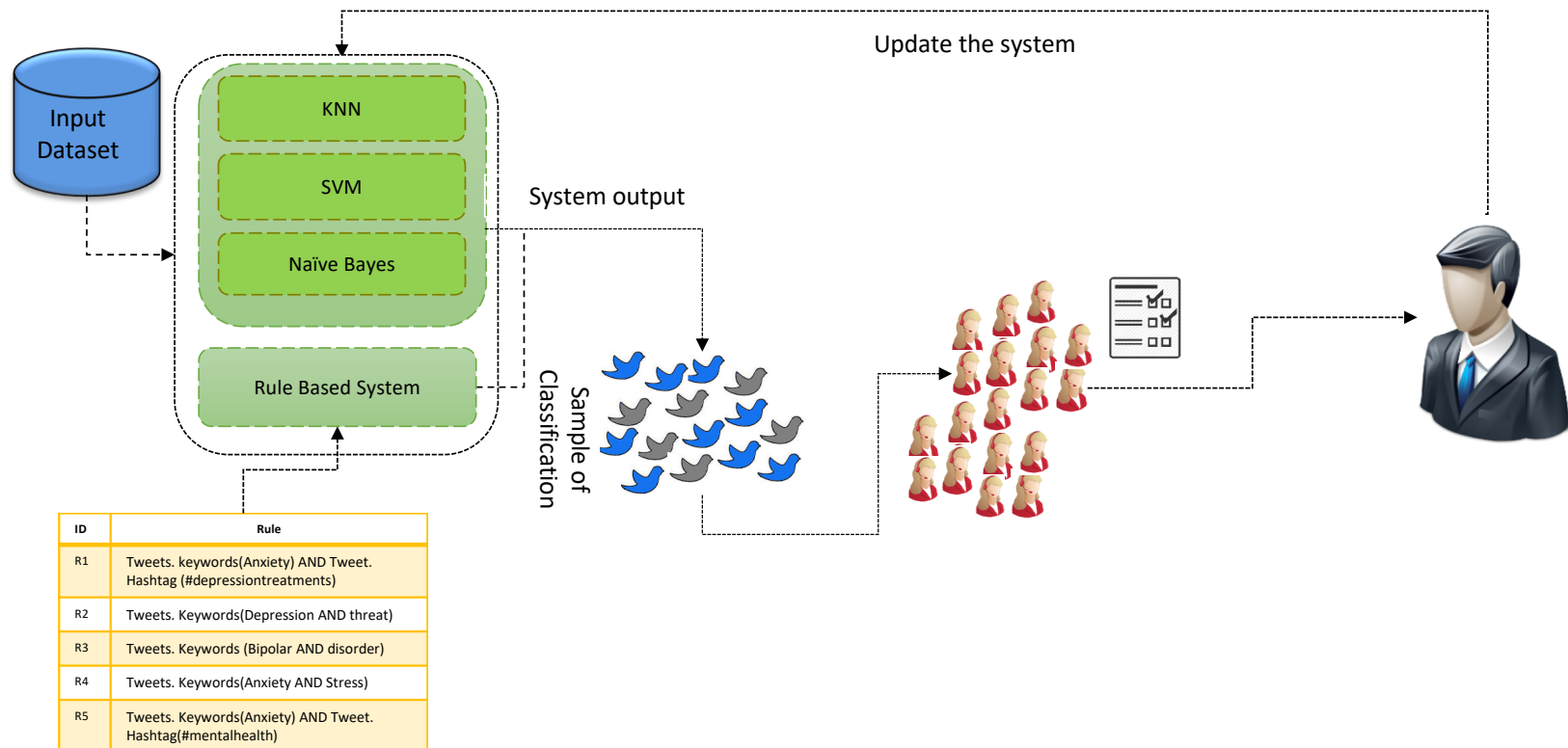
R1	Tweets. keywords(Anxiety) AND Tweet. Hashtag (#depressiontreatments)
R2	Tweets. Entity(Depression AND threat)
R3	Tweets. Topic (Bipolar AND disorder)
R4	Tweets.Keywords(Anxiety AND Stress)
R5	Tweets. Keywords(Anxiety) AND Tweet. Hashtag(#mentalhealth)



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



Hybrid Data Curation Challenges

We briefly can mention to following problems:

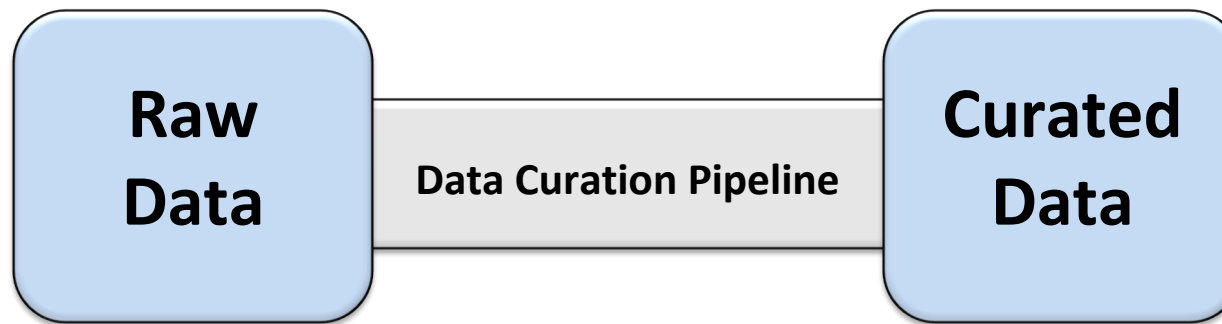
Expensive, as relied on analysts, and needs to be paid

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

Curation Micro Service
TwitterAPI
URL
Named Entity
NamedEntity.extractEntities
NamedEntity.extractPerson
NamedEntity.extractCity
NamedEntity.extractCompany
NamedEntity.extractContinent
NamedEntity.extractCountry
NamedEntity.extractDrug
NamedEntity.extractOrganization
NamedEntity.extractProduct
NamedEntity.extractMoney
Part Of Speech Tag (POS)
PartOfSpeech.Pos Tags
PartOfSpeech.Verb
PartOfSpeech.Adjective
PartOfSpeech.Adverb
PartOfSpeech.Noun
PartOfSpeech.Quotation
PartOfSpeech.Phrase
Synonym
Stem
Similarity

Get /Extraction Named Entity

Description

Returns a list of Entities

Parameters

Name	Located in	Description	Required	Schema
Text	query	Text	Yes	String (String)

Responses

Code	Description	Schema
200	Array (String)	ArrayOfNamedEntities []

Try This Operation

Just as there was a shift from viewing disease as a state to thinking of it as a process, the same shift happened in definitions of health. Again, the WHO played a leading role when it fostered the development of the health promotion movement in the 1980s. This brought in a new conception of health, not as a state, but in dynamic terms of resiliency, in other words, as "a resource for living". The 1984 WHO revised definition of health defined it as "the extent to which an individual or group is able to realize aspirations and satisfy needs, and to change or cope with the environment. Health is a resource for everyday life, not the objective of living; it is a positive concept, emphasizing social and personal resources, as well as physical capacities".[12] Thus, health referred to the ability to maintain homeostasis and recover from insults. Mental, intellectual, emotional, and social health referred to a person's ability to handle stress, to acquire skills, to maintain relationships, all of which form resources for resiliency and independent living.[11]

Since the late 1970s, the federal Healthy People Initiative has been a visible component of the United States' approach to improving population health.[13] In each decade, a new version of Healthy People is issued,[14] featuring updated goals and identifying topic areas and

Try this Operation

NamedEntity API Output

WHO: ORGANIZATION the 1980s: DATE 1984: DATE 12: NUMBER 11: NUMBER 1970s: DATE federal Healthy People Initiative: Company United States: LOCATION 13: NUMBER each decade: SET Healthy People: MISC 14: NUMBER ten years: DURATION US: LOCATION 2020: DATE books: Product the past: DATE the coming years: DATE 15: NUMBER 16: NUMBER Donald Henderson: PERSON cdc: ORGANIZATION 1966: DATE 1974 Lalonde: DATE Canada: LOCATION 19: NUMBER Alameda County: LOCATION California: LOCATION 20: NUMBER World Health Reports of the World Health Organization: ORGANIZATION 21: NUMBER Lalonde: PERSON Canada: