A Hands-on Introduction of Representation Learning

with applications in geoscience data

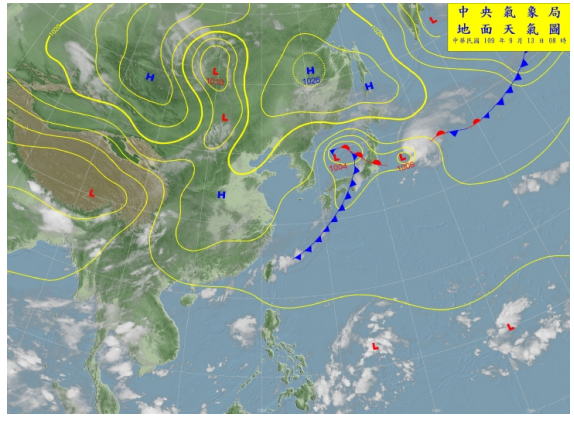
2020.Q3

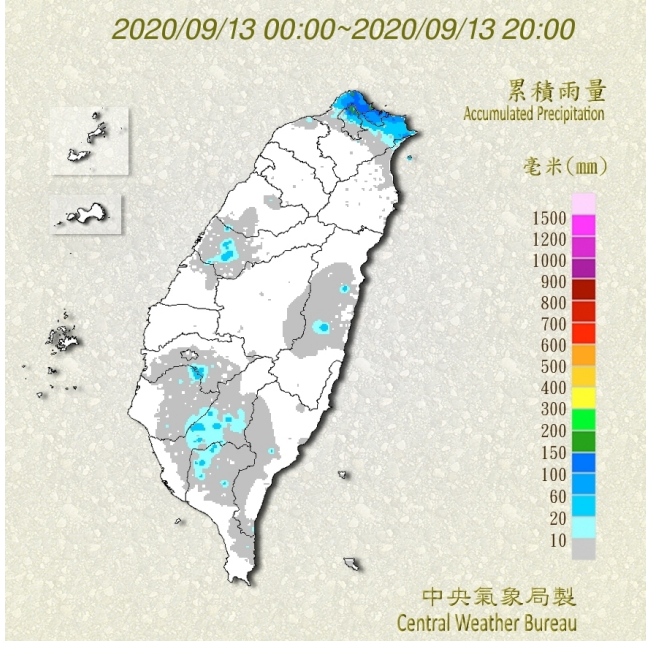
Why do We Need Representations?

- Well, mainly because no one asked me to talk about anything else.
- We are drawn by **BIG** data
    - We need to reduce the **dimensionality** of the data
    - We need **meaningful** features instead of the raw data
- Learning representation from data is an approximation to the concept of **knowledge**

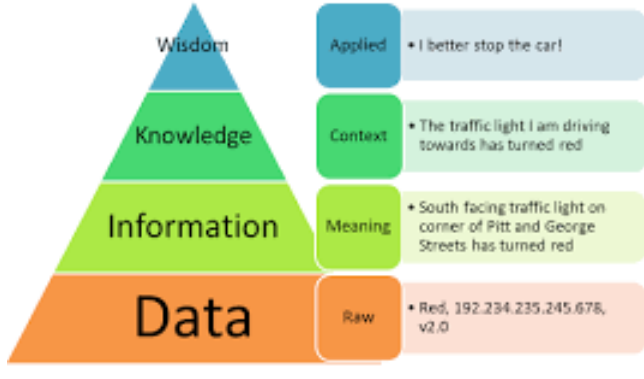
What is a Representation?

- Given the weather map below, will it rain in the following 12 hours?
- What is your reasoning?

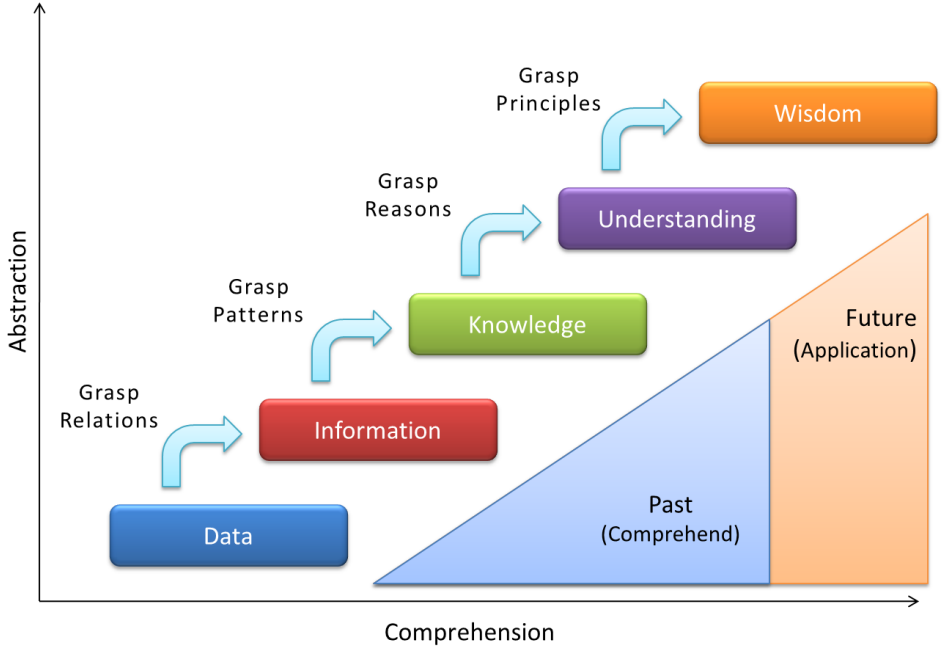



2020/09/13 00:00~2020/09/13 20:00
累積雨量
Accumulated Precipitation
毫米(mm)
1500
1200
1000
900
800
700
600
500
400
300
200
150
100
60
20
10
中央氣象局製
Central Weather Bureau

From Data to Knowledge



From Data to Knowledge



What is Representation Learning?

- From Wikipedia:

> *In machine learning,* ***feature learning*** *or* ***representation learning*** *is a set of techniques that allows a system to* ***automatically*** *discover the representations needed for feature detection or classification from raw data. This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task.*

What is Representation Learning?

- Synonyms:
    - Dimension Reduction
    - Feature Engineering
    - Manifold Learning
- **Representation Learning** is the modern name used in the deep learning era.

What do Representations Represent?

- To reduce the dimensionality
    - Compressing the data while keeping important information.
    - What is **important**?
- To learn meaningful representations
    - How to define **meaningful**?
- We need a measurable standard.

Two Major Approach

- Direct decomposition
    - We assume that we know nothing about the data itself.
    - PCA, autoencoder, ...etc.
- Latent variable approach
    - We assume there are a few determing factors behind the data.
    - FA, Variational Autoencoder, GAN, ...etc.

Topics to Cover

- PCA and FA
- Auto-Encoder
- Variational Auto-Encoder
- Generative Adversary Network (GAN)

(optional)

- Graph theory and relavant algorithms in short
- Graph representation

Things to do

1. Select the problem of your own.
    - Which dataset? Representation for what?
    - Example: NOAA-GridSat-B1 for large precipitation at station 466930

1. We will walk through the methods and example codes using the hand-riting digits dataset (https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html).

2. Apply the methods on your data and problem, and show us your finding.

References

- Feature learning | wikipedia (https://en.wikipedia.org/wiki/Feature_learning)
- An introduction to representation learning | opencourse.com (https://opensource.com/article/17/9/representation-learning)
- An overview on data representation learning: From traditional feature learning to recent deep learning (https://www.sciencedirect.com/science/article/pii/S2405918816300459)