What Can Artificial Intelligence Do in Data Assimilation?

Mia Feng

December 9, 2018

(a) UGV

(b) XiaoIce

(c) Alpha Go

Figure: AI stuff

Artificial Intelligence has been hyping up.
It is announced as an *emotional*, *creative*, and *lively* stuff.

However,

- Do AI stuff have intelligence?
- Can robots feel pain?
- Does emotional XiaoIce really empthise with you?
- Can AI become a human in the next 50 years?

Absolutely Not.

Facing all of the hype, we need to figure out what it is and what can it do.

Catalogue

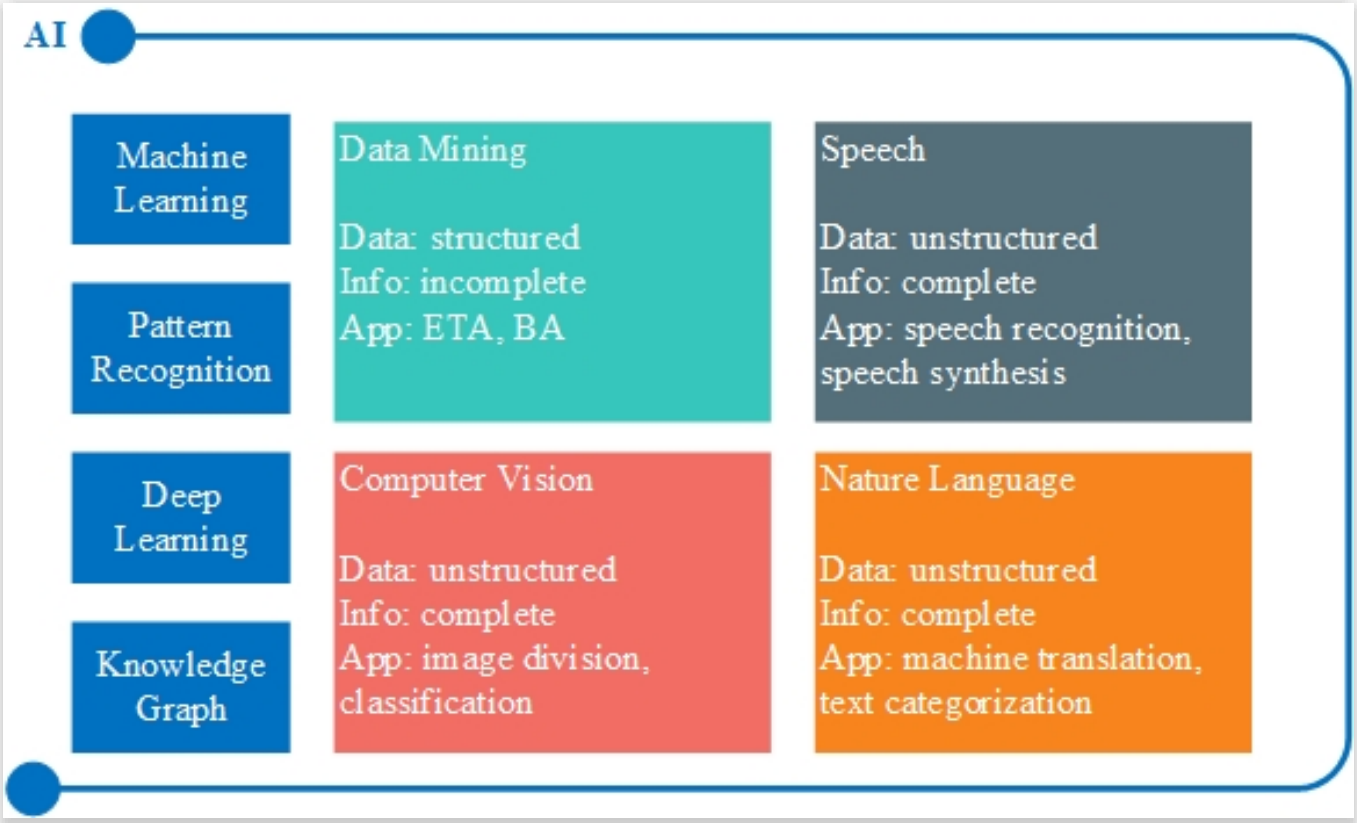
- Meet AI.
- Looking for differences: AI and D.A.
- What can AI do in D.A.?
- In what way can we get closer?

What is its name?

AI?

Methods: Machine Learning (ML), Deep Learning (DL), Pattern Recognition, Knowledge Graph.

Domains: Data Mining, Speech, NLP, CV.



What can it learn?

Discovering regularities: any, even the regulars hidden in intuitively irrelevant matters.

- The relations of entities.
- Writing poems or songs.
- Image captioning.
- Face recognition, face validation.

The rule should be *latent* but *reseasonable*, and can be *generalized*.

Case : *Baby diapers and beers*.
However, it may not work in shops, or supermarkets in China.

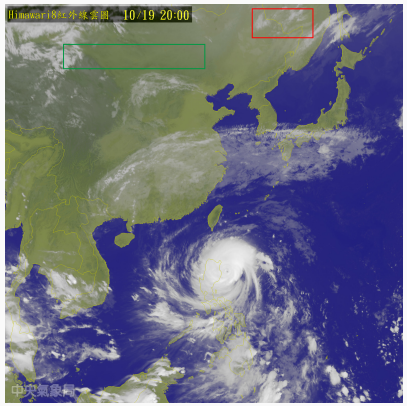
Why does it work?

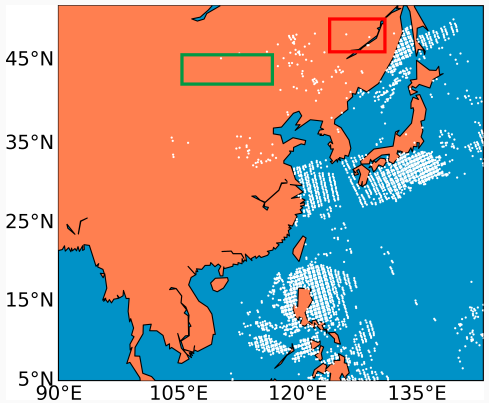
Powerful algorithms? Not really.

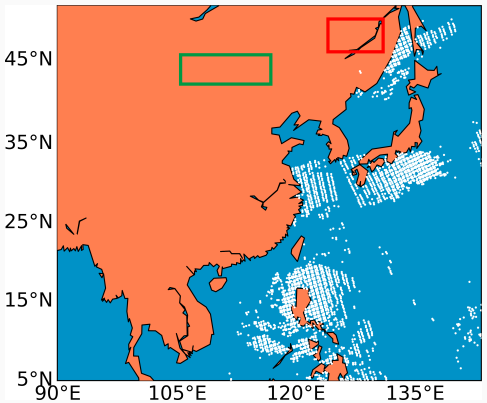
Understanding data is a top priority.

Data-driven approach:The upper bound of machine learning is determined by data and features, while algorithms and models can only help you approach it.

- Cloud detection.
- Fraud behaviour detection while topping up mobiles.







Why does it work?

Powerful algorithms? Not really.

Understanding data and representing data is a top priority.

Data-driven approach:The upper bound of machine learning is determined by data and features, while algorithms and models can only help you approach it.

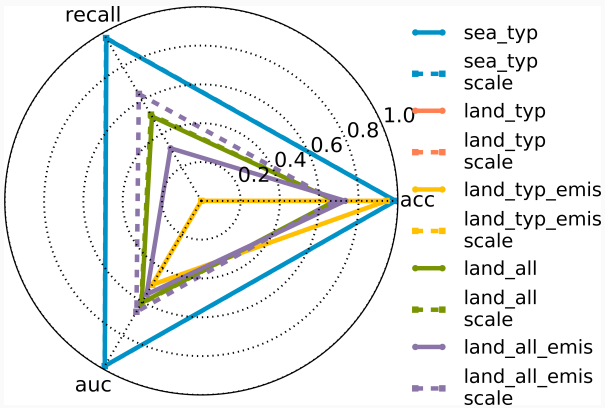
- Cloud detection.
- Fraud behaviour detection while topping up mobiles.

(a) sea

	cloudy	clear	ratio
train	9532	17841	0.35
test	2728	6587	0.29

(b) land

	cloudy	clear	ratio
train	124	7629	0.02
test	113	1879	0.06



When did it work?

LeNet (1980); AlexNet (2012), ZFNet (2013), VGGNet (2014), GoogLeNet (2014), ResNet (2015).

- GPU.–Speed
- Optimization algorithms: back propogation. –Accuracy
- Initialization like Xavier, and normalization like batch normalization–Steady
- Increasing data.–Demands
- Sklearn, tensorflow, keras, caffe etc.–Easy

Where can it be powerful?

What AI found will be exciting if clean data is represented uniquely without missing.

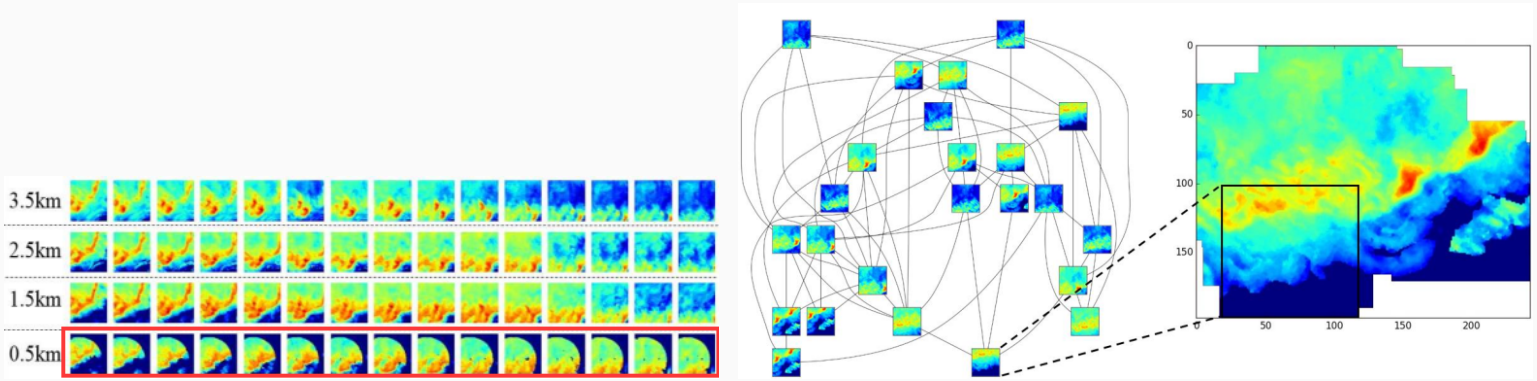
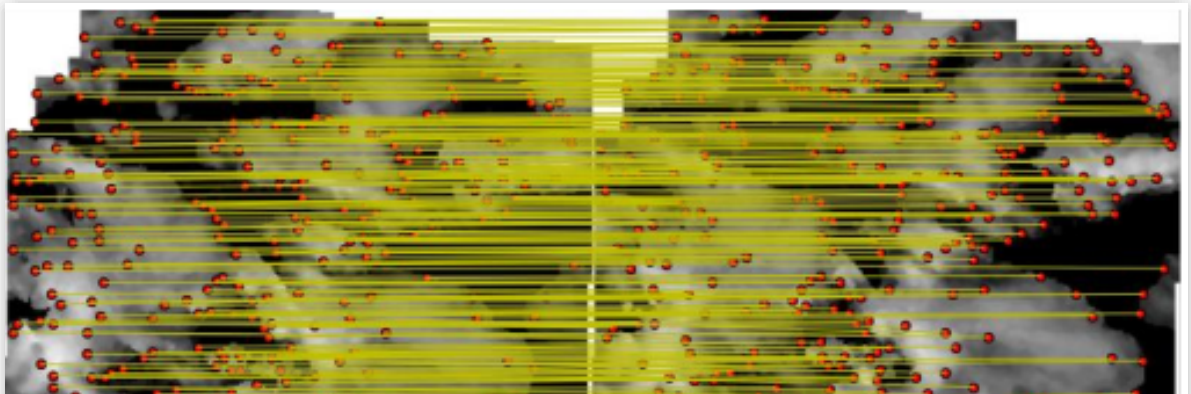


Figure: CIKM[3]

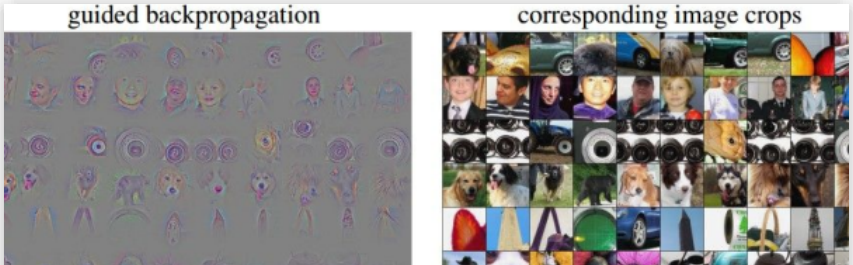


Where can it be powerful?

What AI found will be exciting if clean data is represented uniquely without missing.

- Distributed representation: enable generalization to new combinations of the values of learned features beyond those seen during training[2].
- Representation learning: identify and disentangle underlying explanatory factors hidden in the observed milieu of low-level sensory data[4].

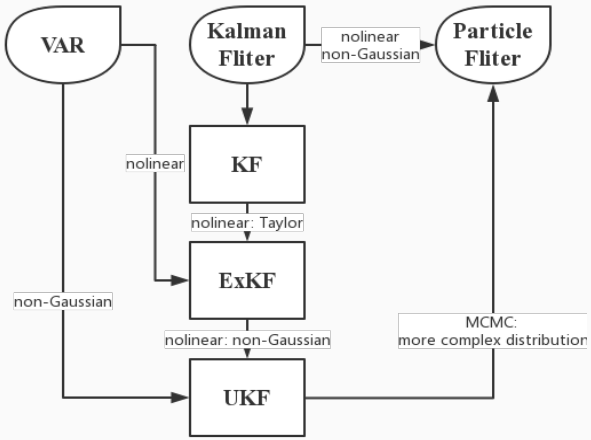
What did kernels learn[1]?

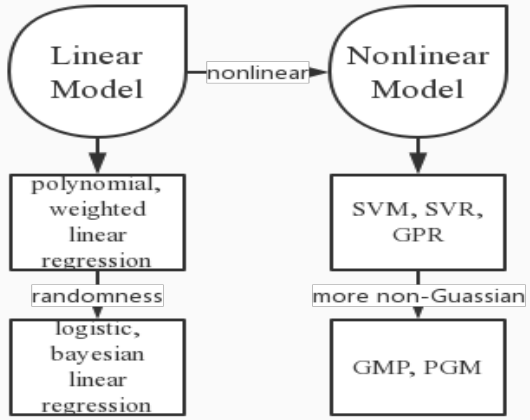


Methodologies

D.A.

The proportion of randomness and nonlinear models is increasing.





M.L.

The same.

Application Area

Look for the method of how to look for a needle in a bottle of hay.–pattern

Basically, both of them model some problems mathematically in real world and try to predict the answers, then try to reuse the models built before.

- Algorithms are used in
  M.L.: NLP, CV, Speech Recognition and Signal Processing, Object Recognition, Multi-Task and Transfer Learning, Domain Adaptation.
  D.A.: weather forecasting, ice, crop, medical treatment like ECG,
- Algorithms consist of
  M.L.: statistical models, numerical computation methods,
  D.A.: physical models (mostly published as models like WRF), statistical models, numerical computation

Data Format

Data used in

- M.L.: stuctured data mostly, pattern in them are simple.
  - One-hot vector: represent words. Impossible for representing data in D.A. mostly.
  - The relations of entities: like costumers and manufacturers.
- D.A.: spatial-temporal data, samples come from the same sampling environment are rare.
  - Infrared hyperspectral data: continuous values, numerous channels, need to be reconstructed. The number of IASI Data sampled from the same geographical coordinates at the same time is zero considering the type of satellites.
  - Salinity data: noise, tracks (Argo).

What can AI do in D.A.?

For instance,

- knowledge graph? The relation of factors?
- transfer learning? $A + B \to C$.
- adversial learning? Generalization.
- visualizing NNs? Understand it then use.
- compress NNs? Online forecasting.

Now, your turn.

In what way can we get closer?

Professional Database?

Can we have an ImageNet or a CIFAR?

- Sufficient labeled data.
- Data should be saved in a cloud system.
- Data should be accessed and transformed dynamically.

Workshops?

- Learn something?
  CS229, Deep Learning by Andrew, PRML?
- Keep pace with something?
  Study the stuff related to your research area which are published recently.

You need Learning, Communication, and Patience.


Justin Johnson Fei-Fei Li.
Cs231n: Convolutional neural networks for visual recognition.
http://cs231n.stanford.edu/.
2018.


Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.


Zhongjie Li Yichen Yao.
Cikm analyticup 2017: Short-term precipitation forecasting based on radar reflectivity images.
https://github.com/yaoyichen/CIKM-Cup-2017.
2017.


Bengio Yoshua, Courville Aaron, and Vincent Pascal.
Representation learning: a review and new perspectives.