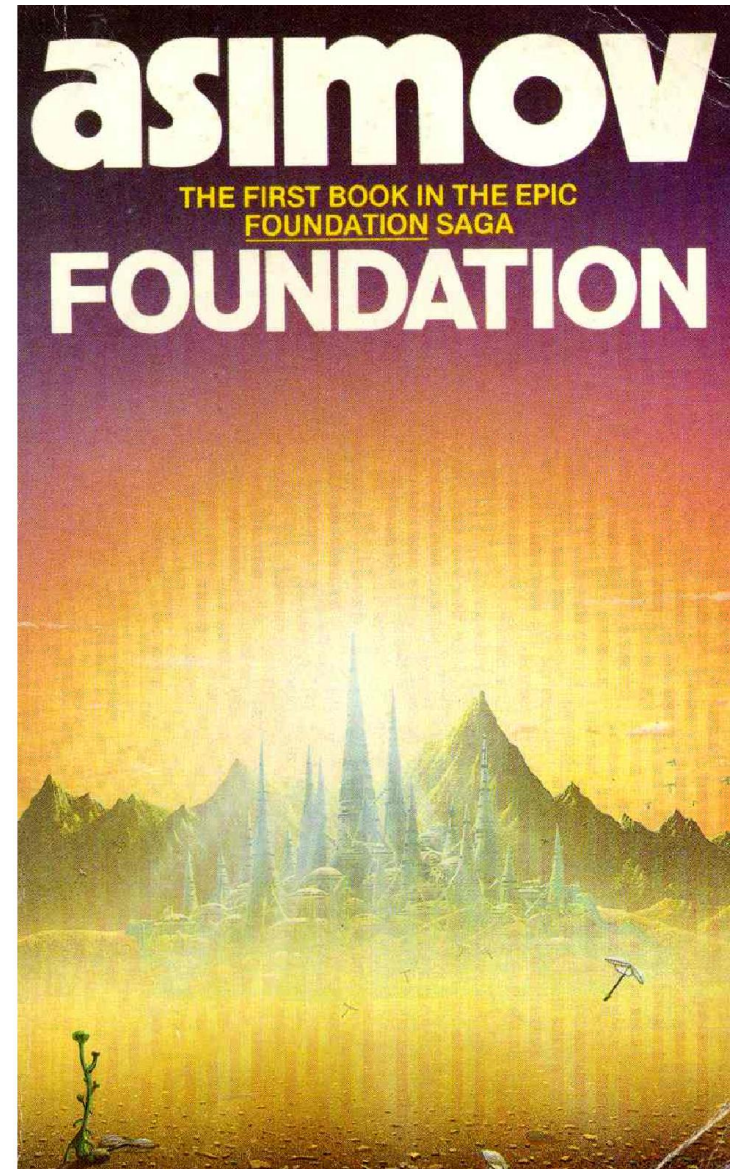
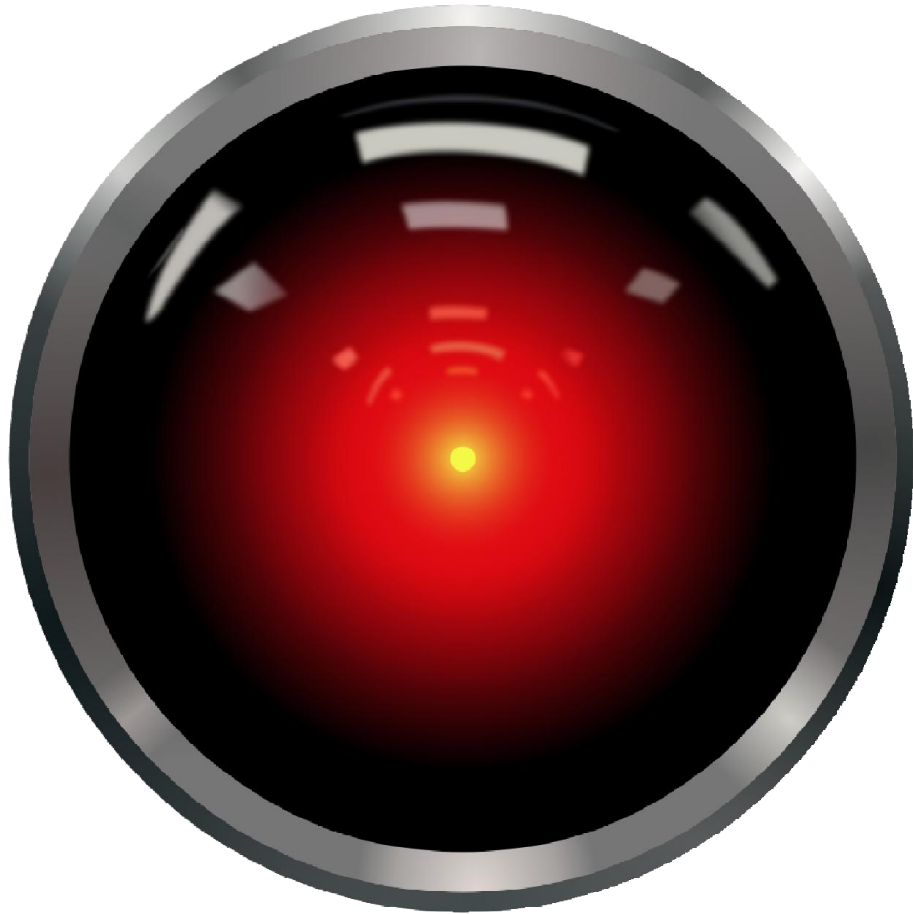


# What the future holds for the Data Science – A Machine Learning Perspective

이현봉

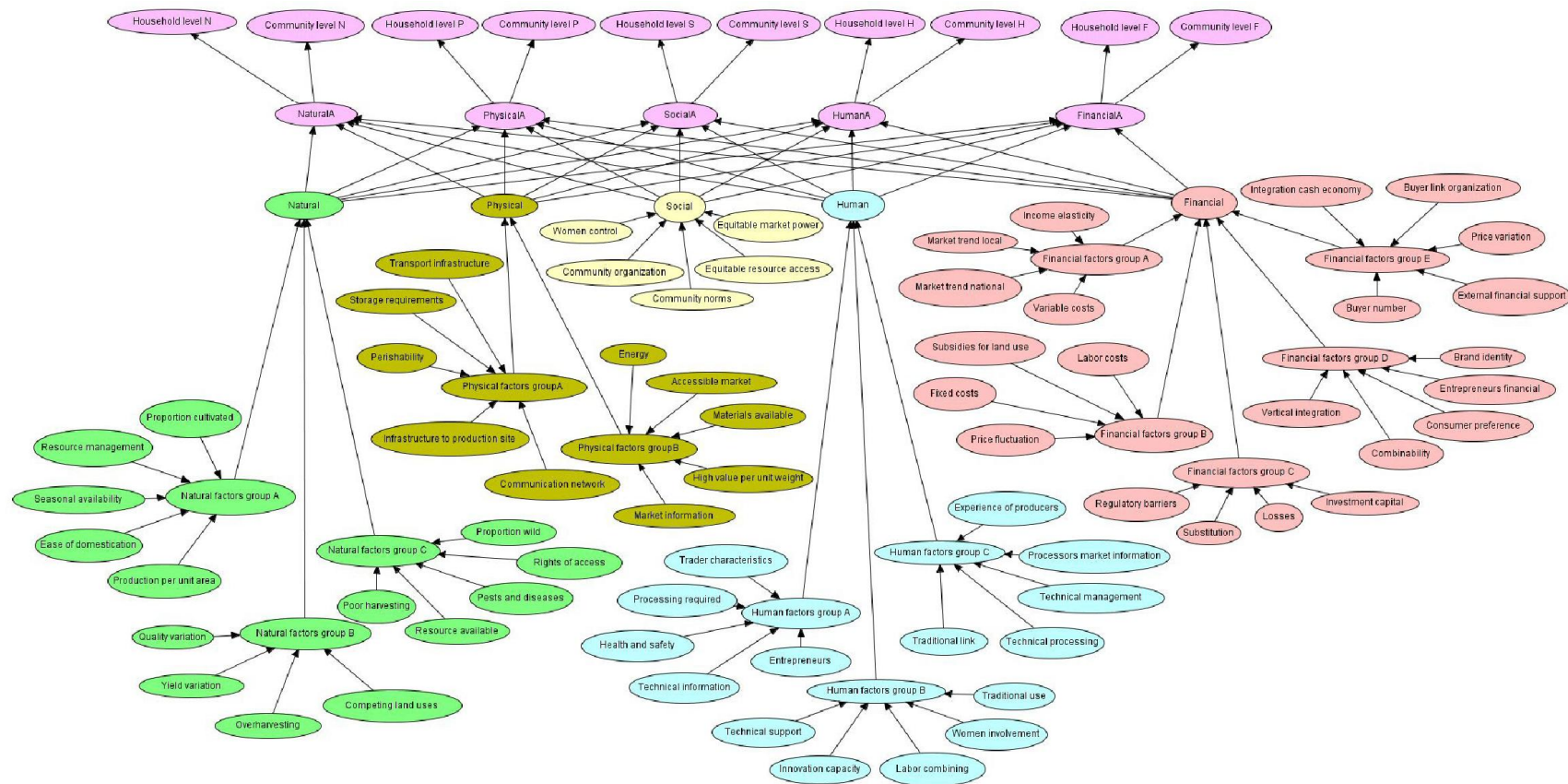
[hal.lee@teradata.com](mailto:hal.lee@teradata.com)

# How I got into Machine Learning/AI

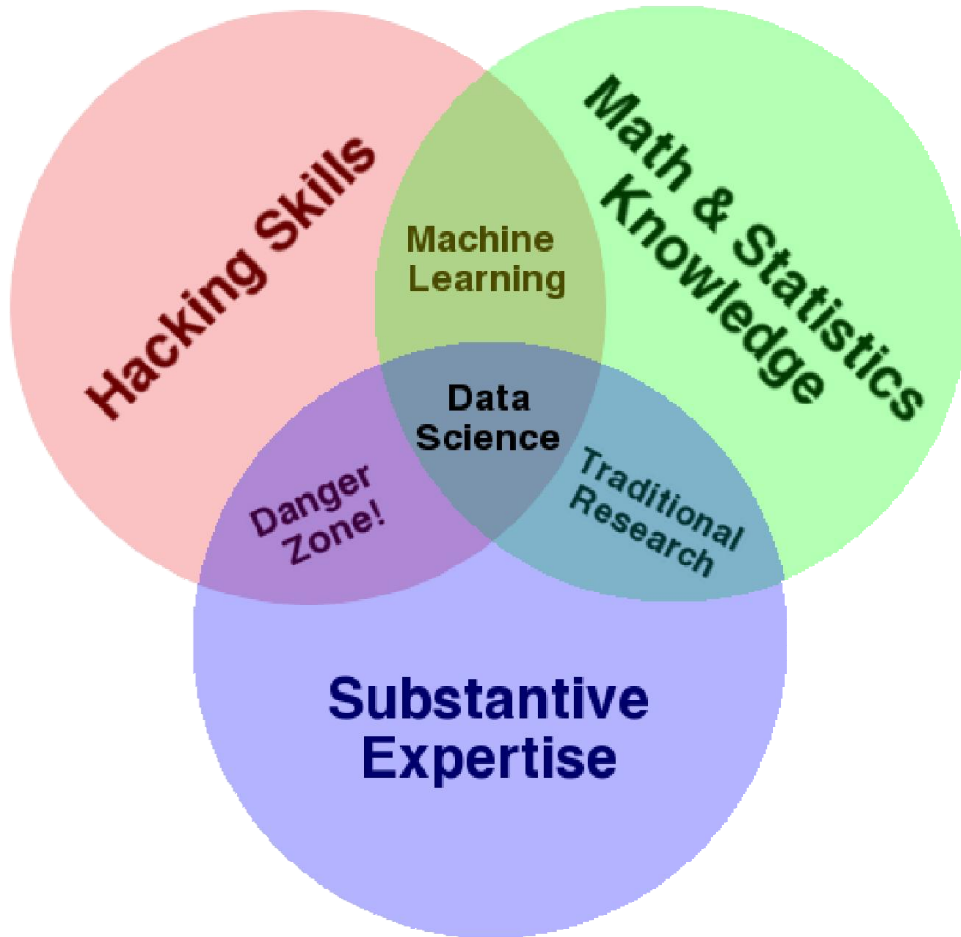


# Theory behind the "Foundation"

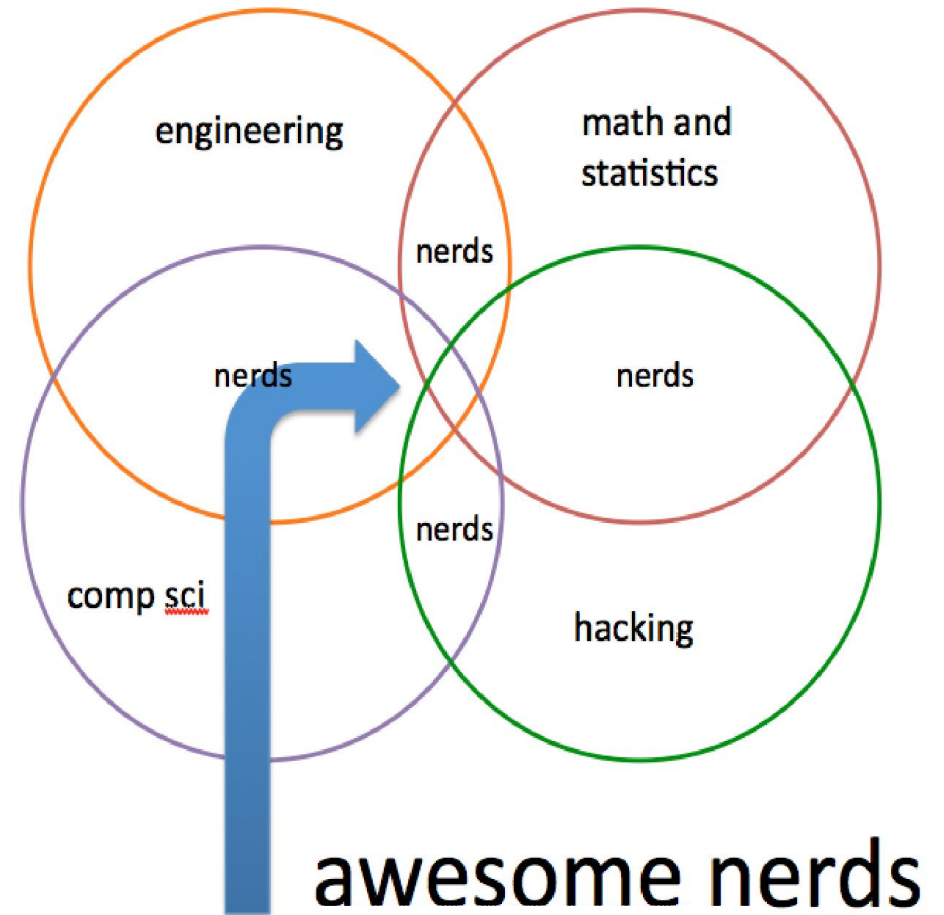
- 지구상의 모든 사람들과 역사 발전에 관여하는 모든 것들을 확률변수로 묘사
- 사람들이 죽고, 태어나고, 사람과 사람, 사람과 다른 것들간의 dependency, interaction과 influence를 Bayesian Network (or, Markov Random Field)로 묘사
- Bayesian Network 구조 자체가 진화하는 **System of Big Random Processes**
- Compute (aggregate) Probabilities and Scenarios for the future



# Data Science?



# Data Scientist?



그러면,

SAS와 R로 데이터 작업을 하지만 기계학습이나 수학 통계를 잘 모르는 나는 테크니션, operator?

바다 속 엄청나게 많은 신호와 온갖 정보를 종합해 완전 골 때리는 수학적 분석과 기계학습해서 잠수함 찾기 하는데, 데이터 사이언티스트라고 하지 않거든?



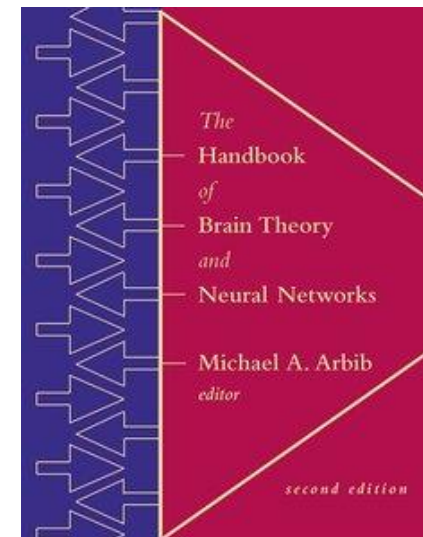
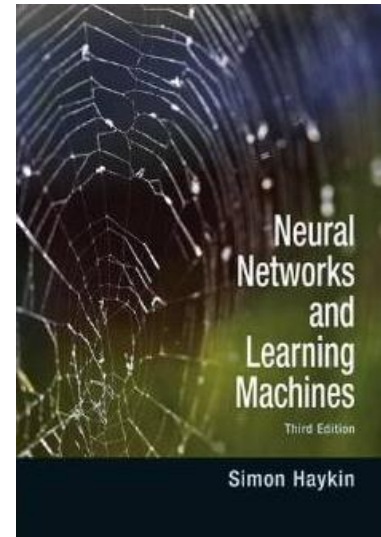
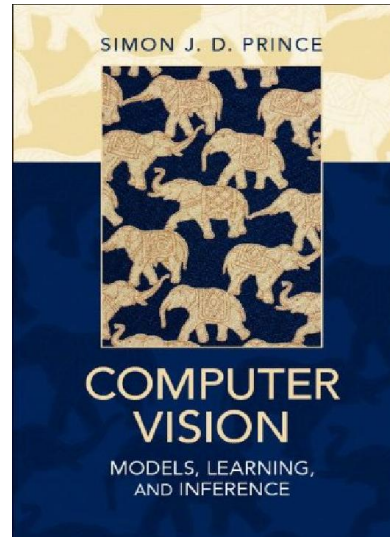
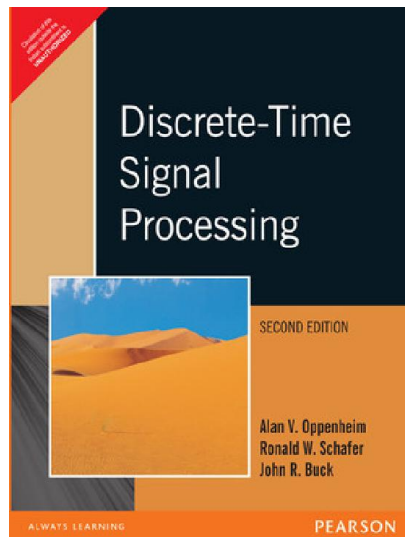
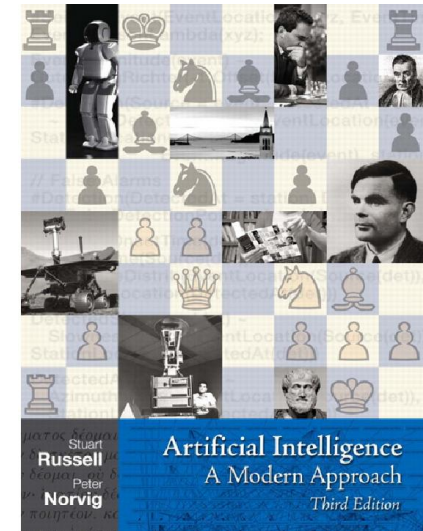
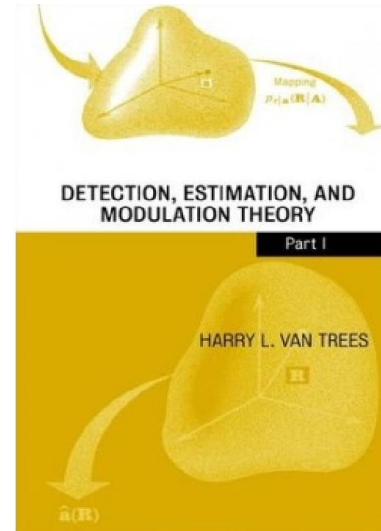
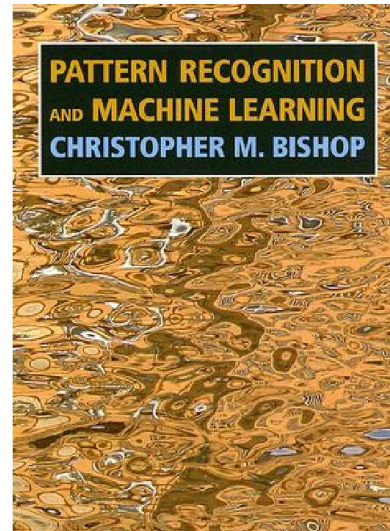
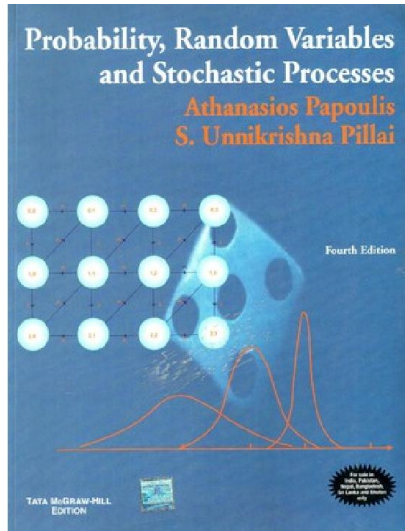
# Machine Learning?

- **인공 지능**의 한 분야로 경험을 통한 **학습**을 하여 스스로 자신을 **발전**시키려는 시스템/알고리즘을 연구 개발



시간문제?

# Machine Learning을 어떻게 학습하나?



책에 있는 모든 내용이 실제로 다 쓰임



공학에서

기계학습은 깊은 지식이 필요한 전문  
운용시스템의 한 부분으로 타이트하게  
결합되어 적용되는 경우가 일반적



데이터 사이언스에서

기계학습은 그 반대





## Question -1

- 데이터 사이언티스트가 할 일이 앞으로 많아질까?

→ 앞으로 더욱 데이터가 많이 배출될 것이기에, 이들을 그냥 버릴 것이 아니라면 **with high probability, YES.** 만약 이런 일이 안생기면 **Big Dump**가 각광받을 것.

→ 누가 무어라 해도 Data Science/Scientist는 불분명한 개념. 자칫, 얇은 지식/허풍과 동일시 될 수도. 이미 데이터를 깊이 이용하는 특정 분야에 고도로 훈련받은 **전문가**들이 있다는 것을 기억. But then, will the domain knowledge be important in the future as it is today?

→ 데이터 사이언티스트들이 전문 분야를 선택하여 집중할 것을 강요 받을 수도. **어떤 분야?** 그 쪽에 이미 전문가들이 없나? Or, Data Scientist is a generalist?

## Question -2

- AI/ML 친구들이 놀고 있지 않을 터인데 어떤 일이?

→ 이들이 배우는 것 중 하나가 전문가(Expert) 시스템. Data Munging에 많은 노력이 드는 데, 이들은 이런 것에 도사. 다만, 자주 하지 않기 때문에 참음. 만약 이런 일을 자주해야 하면 당연히 편한 방법을 찾을 것.

### 자동화를 추구

→ AI/ML 사람들은 데이터를 분석해서 리포트를 내는 것보다 데이터를 분석하는 시스템을 만드는 데에 익숙. 따라서 다양한 needs를 겨냥하는 **데이터 사이언스 상품/어플/서비스가 출현할 것이 예측됨**. 이것들은 DS의 유용성 입증과 확산을 도우는 동시에 경쟁 상대

## Question -3

- **How would Big Data & Cloud shape Data Science?**

- 빅데이터와 클라우드 플랫폼/사용은 빠른 속도로 쉬워지고 있음
- **빅 데이터 사이언스 프로젝트를 시작하기도 쉬워짐**
- 사용할 수 있는 데이터는 많지만, Signal to Noise Ratio는?
- 목적과 규모에 맞추어 적절한 플랫폼/기술로 데이터 사이언스 수행
- 데이터사이언티스트가 공부해야 할 것이 많아짐
- **Analytics As A Service / Data Science As A Service.**  
가능할까? 어떻게?

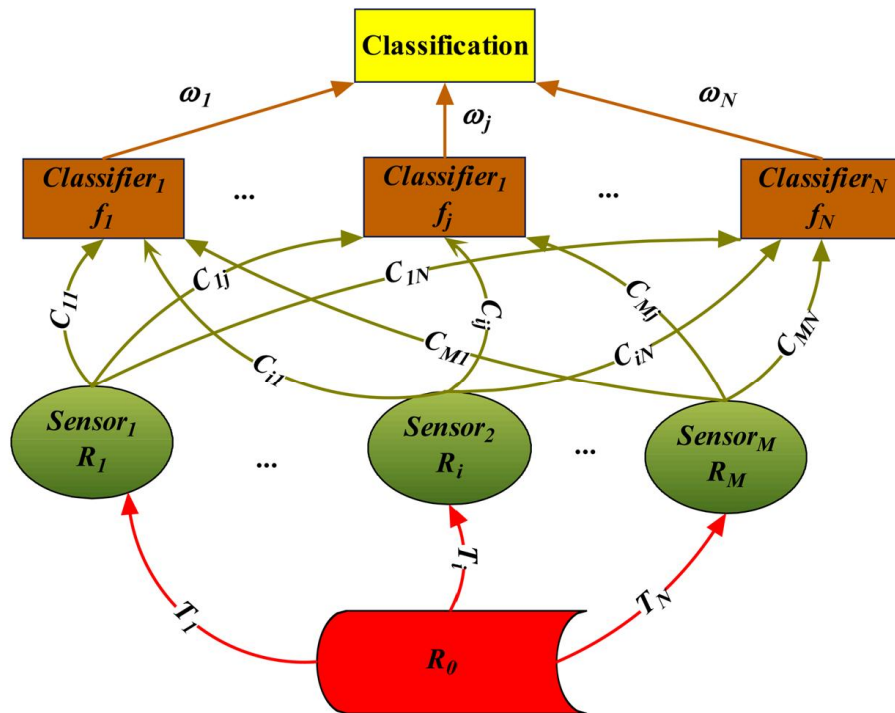


## Really thorny Question

- World is changing so fast. We are led to innovate, and “think out of the box” all the time. **How much trust would you put to a model that is based on past history?**
  - Cannot come up with good answer for that
  - How can learning model adapt itself continuously while maintaining some stability?
  - There can be a “core” pattern, or a very specific pattern that should be worth learning from the **past data**.
  - There would be people working on “Machine Learning from streaming data & on-line learning”

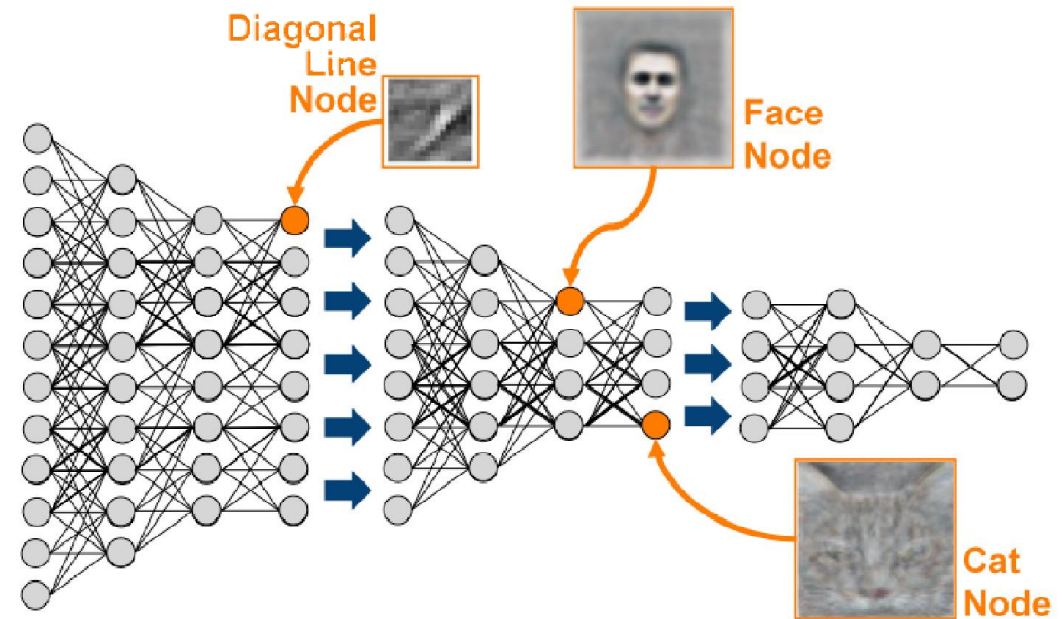
# Two driving forces in Machine Learning

## Ensemble : Collective Intelligence



Each one of us is not perfect. Let's meet and discuss the matter together. Who knows, we might come up with wiser decision

## Deep Learning



I didn't go through trouble of studying all these stuff to do "fittings and optimizations". I want to know mechanism behind general perception and emergence of intelligence. Incidentally, it's also performing fantastically in many practical tasks.

# Deep Learning for Data Scientists

- People on the Machine Perception used to say “give me good features and I will give you all the classification or whatever you want. Know what, we have hard time figuring out what’s the figure and the ground”. Damn, right.
- In the “iris” data, we have four features (Sepal.Length, Sepal.Width, Petal.Length, Petal.Width) and the label(Species). Here, we already chose/marginalized on those specific features. Hard part/choice is already done. But, are these really the best features? How did we arrive to these?
- All living organisms rely on extracting the right features and successively building up more **abstract** higher level representations. It’s manifestation of their domain knowledge, concept, and eventually their **understanding** of the world. It’s deep big (philosophical) issue.
- Many knew about great importance of the “feature engineering” but did not know how to do it systemetically until when Deep Learning showed it could be done pretty much automatically(in other words, DL takes care much of the tasks related with features)
- Currently, Deep Learning is applied to computer vision, speech recognition, topic modeling, signal processing and etc. where raw data is really low level. However, deep learning can be applied to any areas including those where data is stored in DBMS. So, data scientists should pay attention to where deep learning is heading.