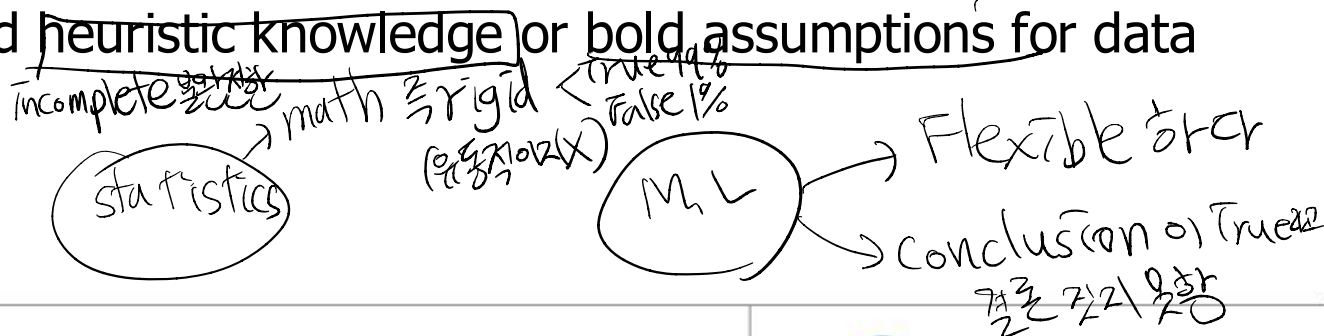# Overview of Machine Learning

What is Machine Learning?

- **Techniques to improve the performance of programs based on given data, previous results, or experiences.**
  - Methods for creating computer programs by the analysis of data sets.

- **Hmm, may we call it "data-analysis techniques"?**
  - It's not perfect, but it's pretty accurate

- **Then, How is it different from with statistics?**
  - Statistics is a branch of mathematics, but ML is not → X,Y가 득점!
  - You can add heuristic knowledge or bold assumptions for data analysis

incomplete 정보로 → math 즉 rigid < True 99% (완전하지X) False 1%

*statistics*

ML → Flexible 하다
→ conclusion이 True임
결론 갖지 못함

Applications

- **Voice/Face/Fingerprint/Iris/DNA/Signature recognition**
- **Web-search, Document & information retrieval, Machine translation**
- **Recommendation, Spam filter**
- **Credit card fraud detection, Loan application analysis**
- **Marketing, Stock market prediction**
- **Games: Chess, Go**
- **...**

How they are different

- **Artificial Intelligence, Machine Learning, and Deep Neural Network**
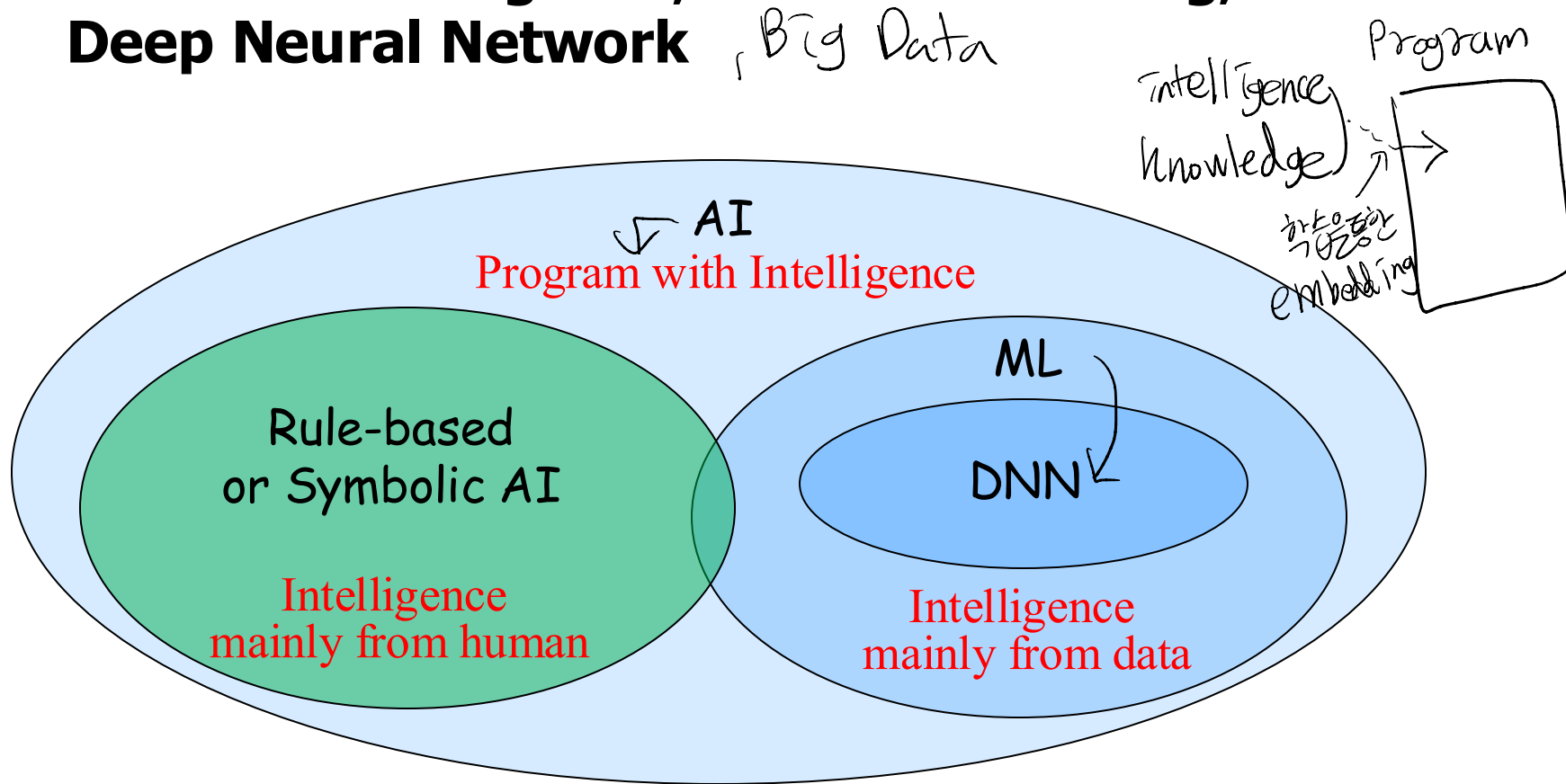
How they are different

- **Artificial Intelligence**
  - To build programs acting intelligently
  - We need to code "intelligence" into programs
  - Where do we obtain "intelligence"?
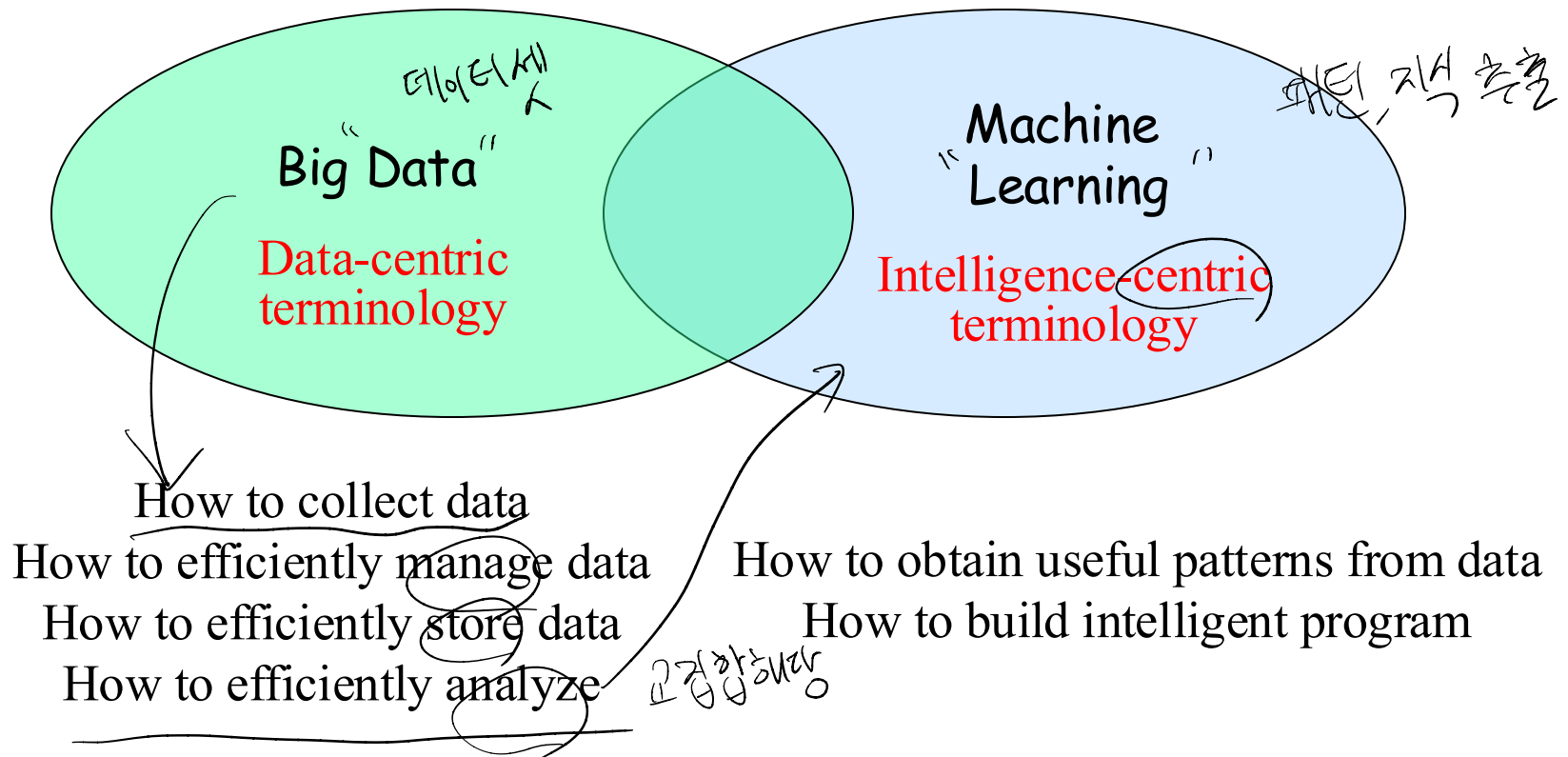
- **From human-beings**
  - They have intelligence
  - Extract "intelligence" from them by asking "how do you do that?"
  - Expert systems -> Big failure

- **From data**
  - Human-beings create data, so their intelligent behavior or decisions are resolved in the data
  - If we distillate it, we can obtain "intelligence"
  - Successful, but why now?

How they are different

- **Big Data and Artificial Intelligence**

데이터셋

Big "Data"

Data-centric
terminology

**Machine
Learning**

Intelligence-centric
terminology

패턴, 지식 추출

How to collect data
How to efficiently manage data
How to efficiently store data
How to efficiently analyze

결합/해당

How to obtain useful patterns from data
How to build intelligent program

Related Theories in ML

- **Math**
  - Probability and Statistics
    - Conditional Probability, Bayesian Rule, Independency, Maximum Likelihood, Normal Distribution, Expectation, Variance, Covariance, Covariance Matrix
  - Linear Algebra
    - Inverse Matrix, Eigen Value, Eigen Vector, Singular Value Decomposition
  - Optimization:
    - Gradient Descent, Convex Optimization, Expectation-Maximization, Quadratic Programming
- **Programming**
  - Dynamic Programming
  - Divide and Conquer
  - Computational complexity theory

Some Types of ML Problems (1)

- **Classification**
  - Voice/Face/Fingerprint/Iris/DNA/Signature recognition, Recommendation, Spam filter, Credit card fraud detection

- **Regression**
  - Loan application analysis, Marketing, Stock market prediction

- **Clustering**
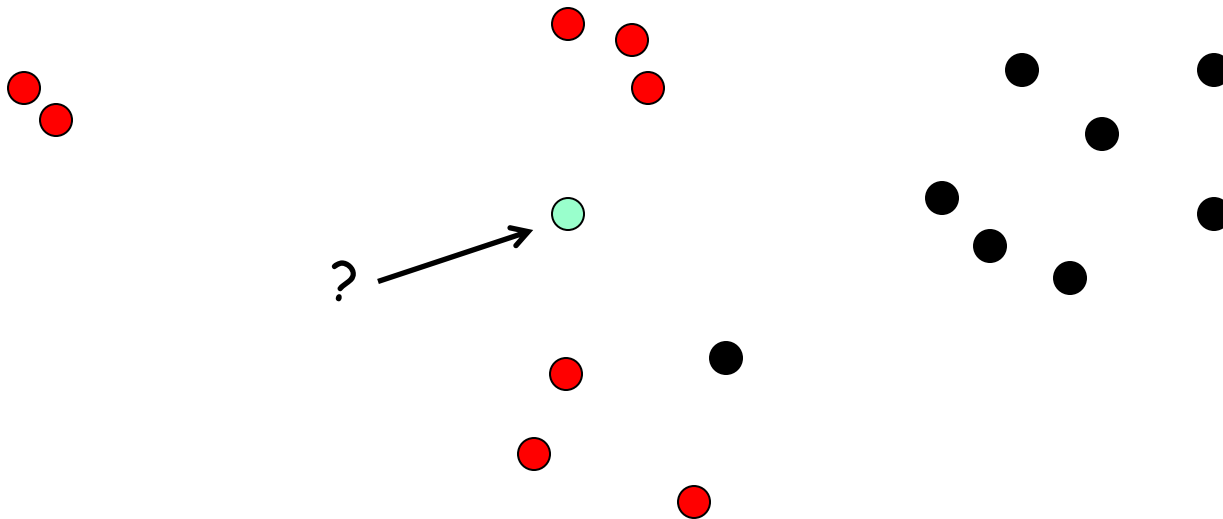 - Web-search, Document & information retrieval, Machine translation

- **Dimension Reduction**

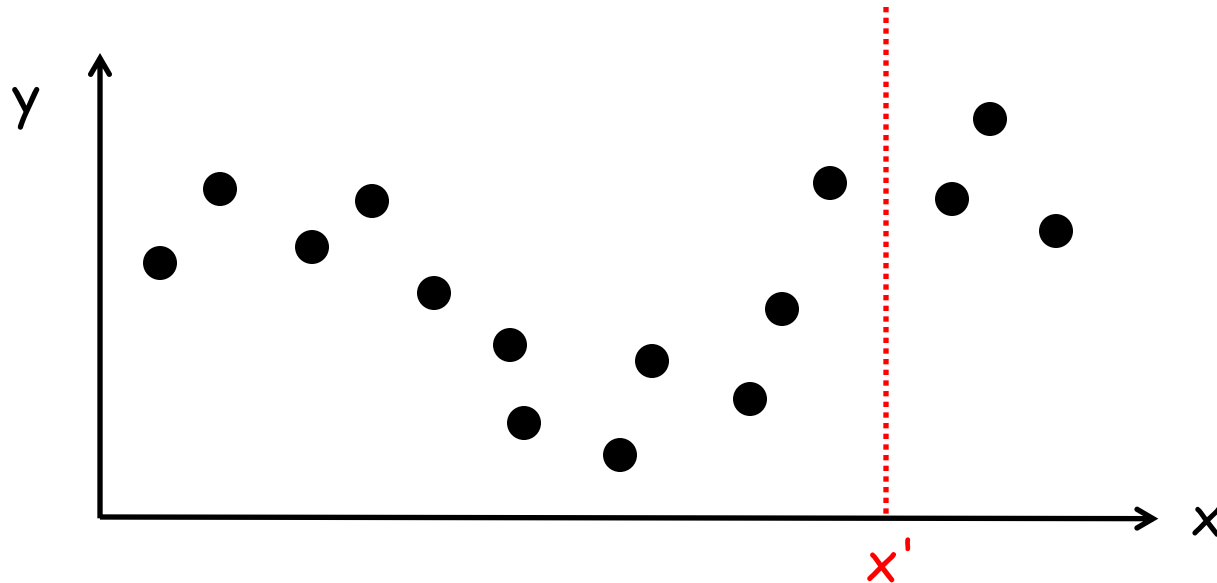# Some Types of ML Problems (2)

- **Classification**
  - A set of labeled data is given
  - Your program should find the boundary between labels
  - If a query is given, your program should answer the label

?

Some Types of ML Problems (3)

- **Regression**
  - A set of (**x**, y)'s is given. (**x** is a vector, y is a real number)
  - Your program should find the functional relation between **x** and y
  - If a query, **x'**, is given, your program should answer y for **x'**

Some Types of ML Problems (4)

- **Clustering**
  - Unlabeled data is given.
  - Your program should group the data (Finding hidden structure of data)
  - If a query is given, your program should determine the group in which the query belongs to
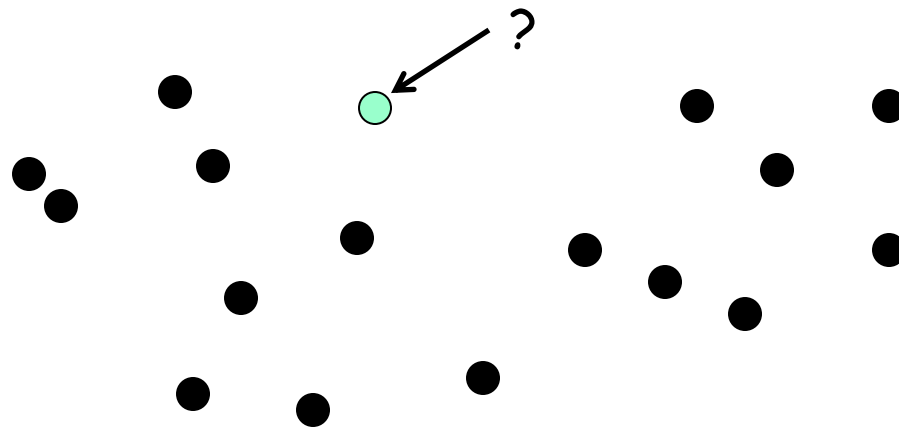
Some Types of ML Problems (5)

- **Dimension Reduction**
  - A set of unlabeled data is given.
  - Your program should reduce the dimension of data by minimizing the loss of information

Find the projection line   Remove other axises
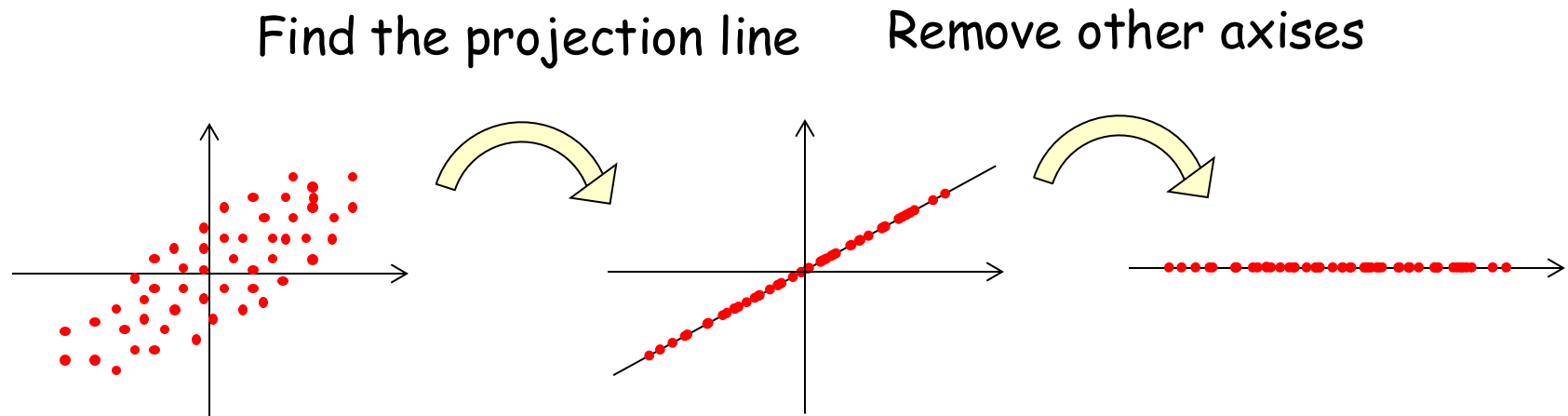
Supervised & Unsupervised Learning

- **Supervised Learning**
  - If a machine learning algorithm needs a labeled dataset
  - Most algorithms for Classification, Regression

- **Unsupervised Learning**
  - If a machine learning algorithm does not need a labeled dataset, (that is, If it only needs an unlabeled dataset)
  - Most algorithm for Clustering, Dimension Reduction