Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the participants individually and don't necessarily reflect those of blibli.com.

Blibli.com does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Machine Learning Research in blibli

Hendri Karisma

hendri.karisma@gdn-commerce.com / hendri.k.id@ieee.org

Hendri Karisma

- Sr. Research and Development Engineer at blibli.com (PT. Global Digital Niaga)
- Rnd Team for Data Science/intelligence system
- Working for Fraud Detection System. Current working in dynamic recommendation system project.

Definition of Informatics

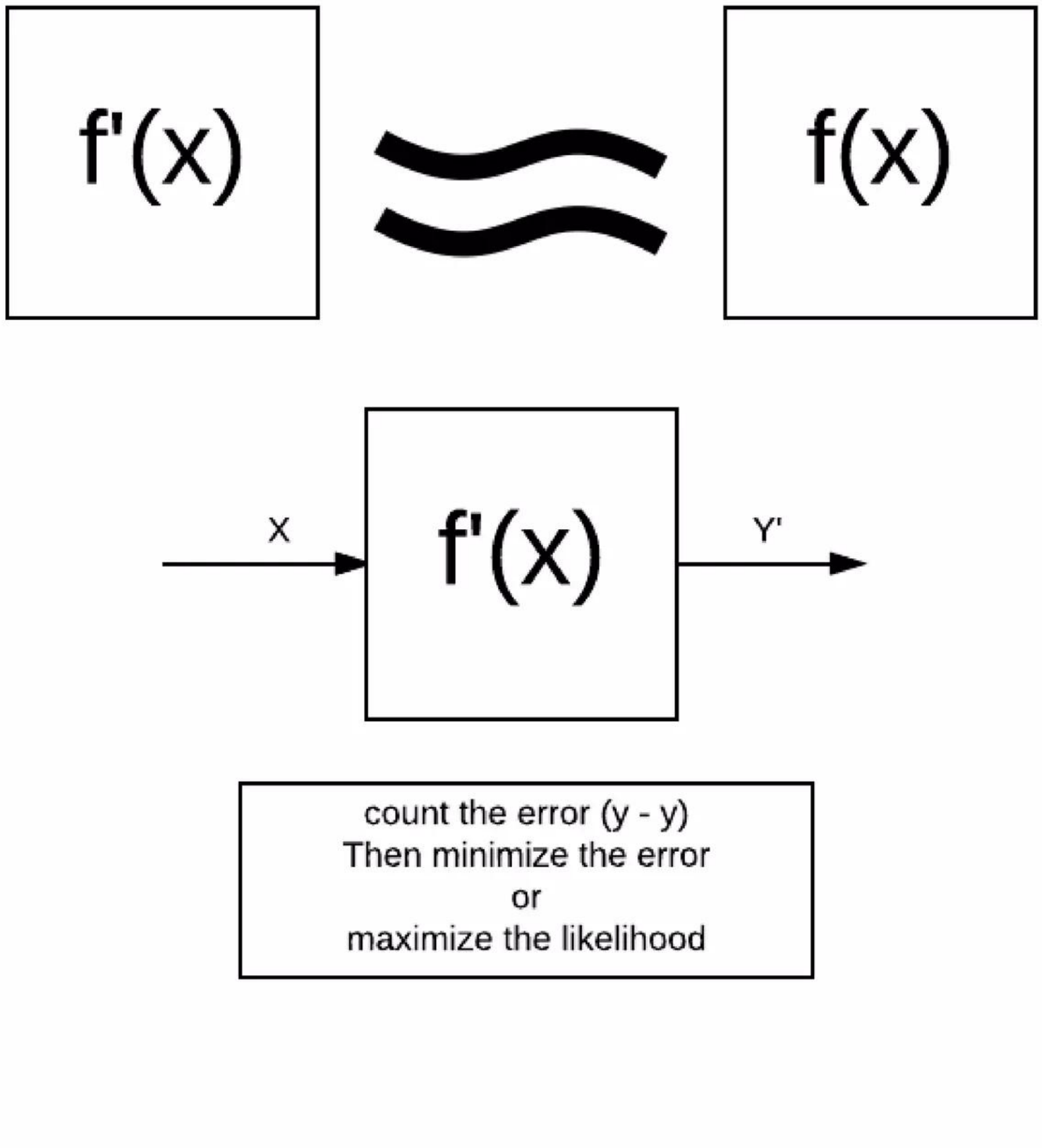
"Automation of Information" –

Prof. Dr. Ing. Iping Supriana

Machine Learning Definition

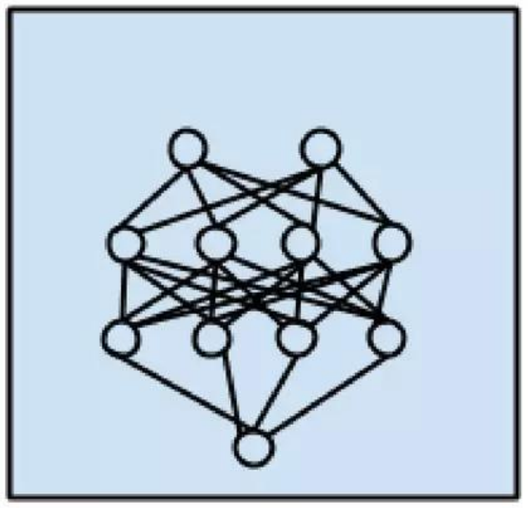
"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." – Prof. Tom Mitchel

How it works

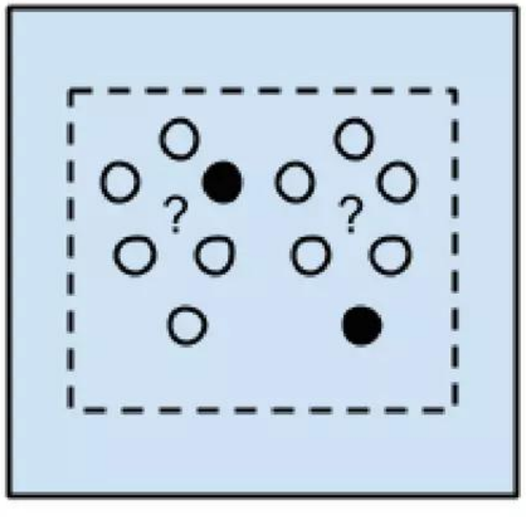

f'(x)
f(x)
X
f'(x)
Y'
count the error (y - y)
Then minimize the error
or
maximize the likelihood

Machine Learning #1

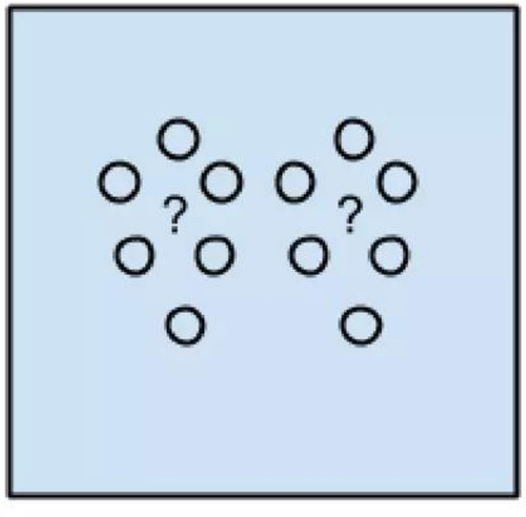
- Supervised
- Unsupervised
- Reinforcement Learning
- Semi-Supervised
- Deep Learning



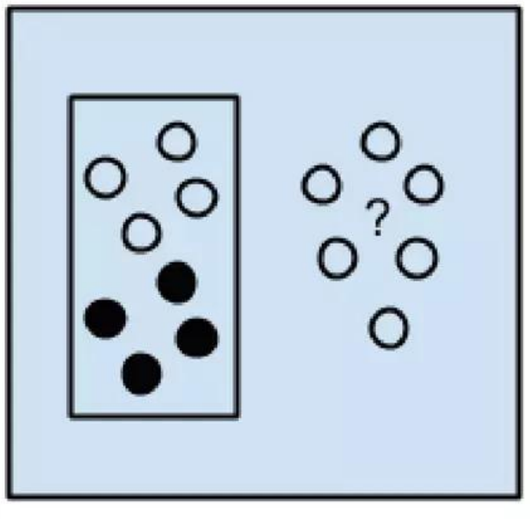
Deep Learning Algorithms



Semi-supervised Learning Algorithms

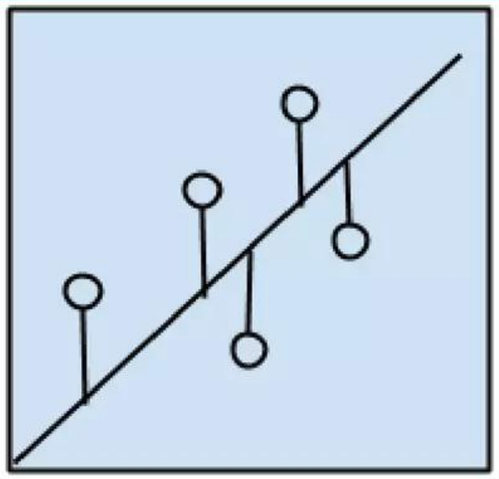


Unsupervised Learning Algorithms

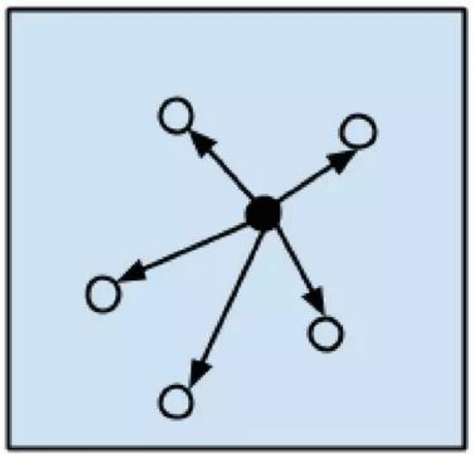


Supervised Learning Algorithms

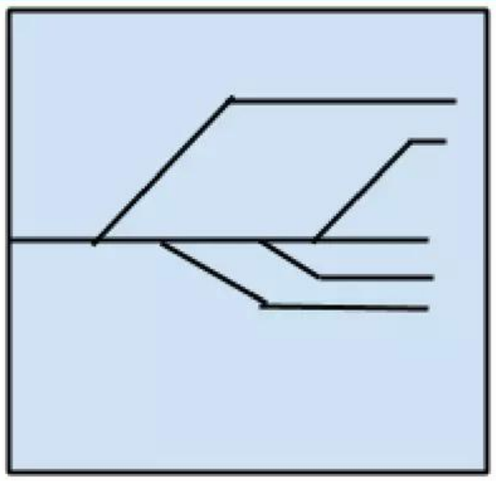
Machine Learning #2



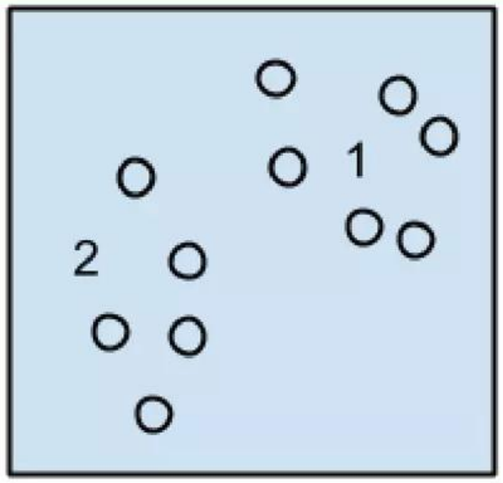
Regression Algorithms



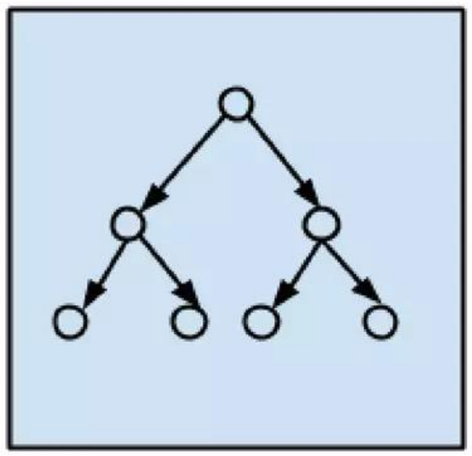
Instance-based Algorithms



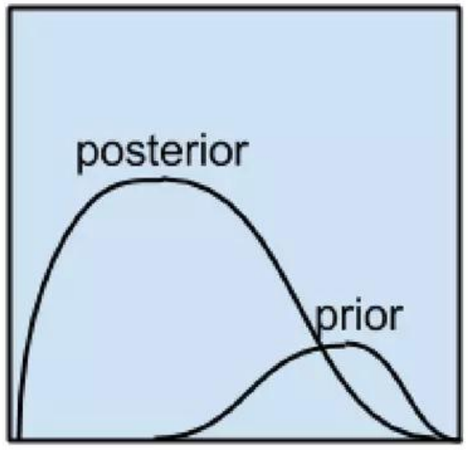
Regularization Algorithms



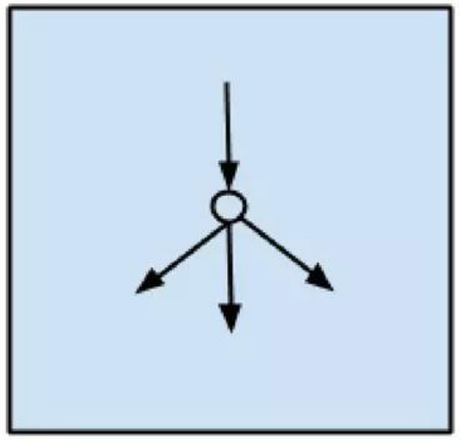
Clustering Algorithms



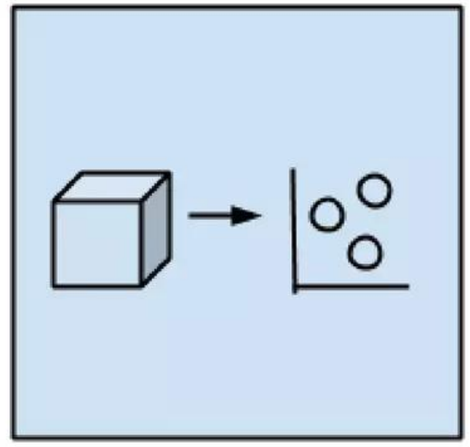
Decision Tree Algorithms



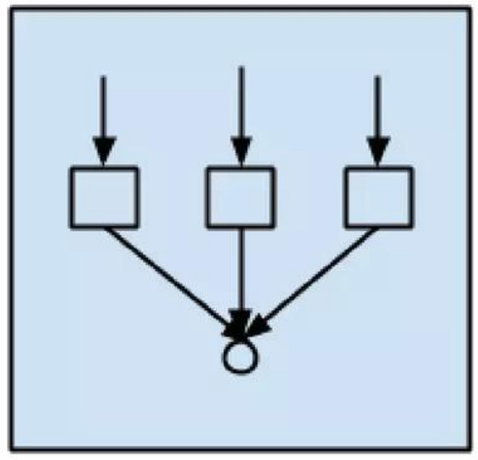
Bayesian Algorithms



Artificial Neural Network Algorithms

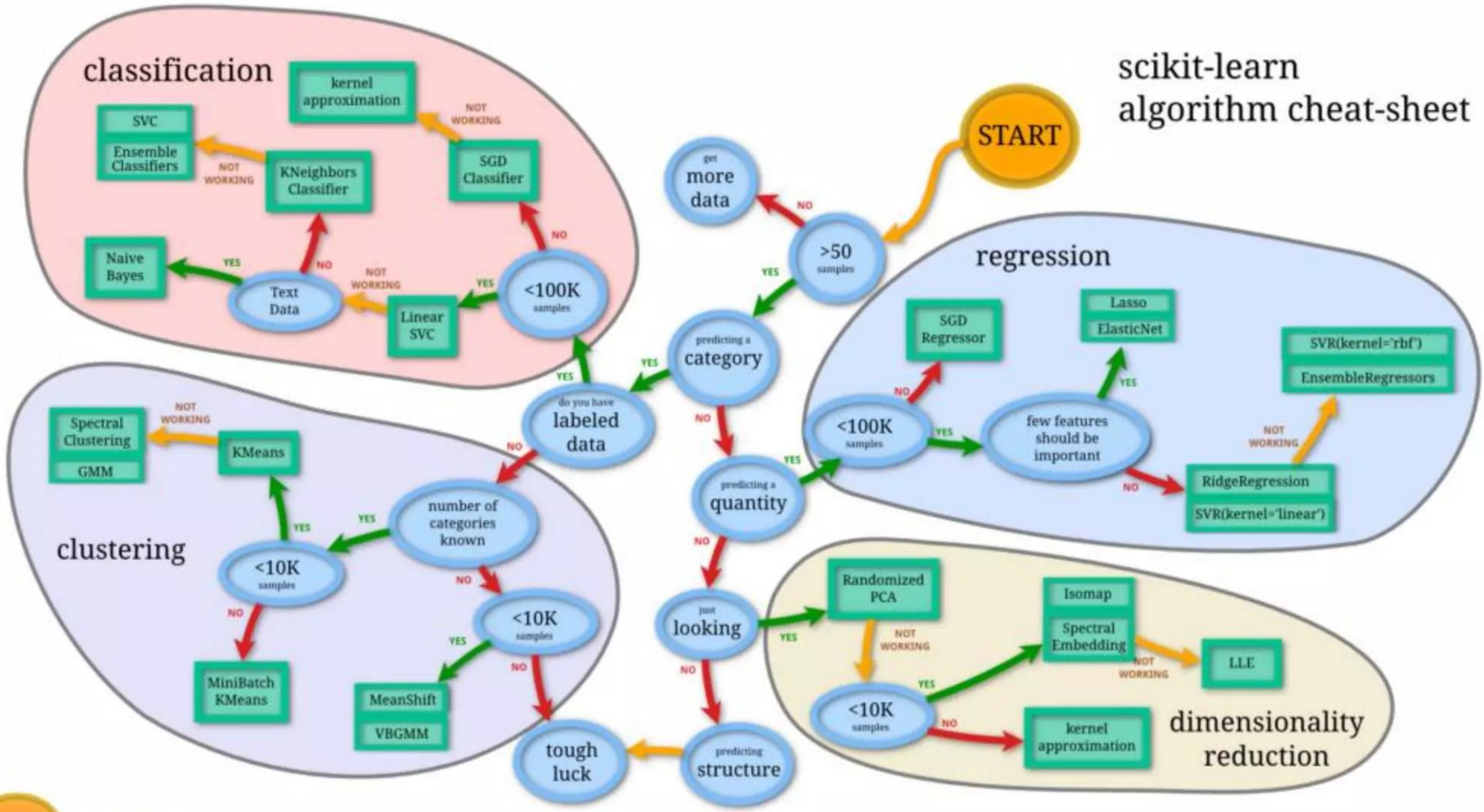


Dimensional Reduction Algorithms



Ensemble Algorithms

Machine Learning Taxonomy



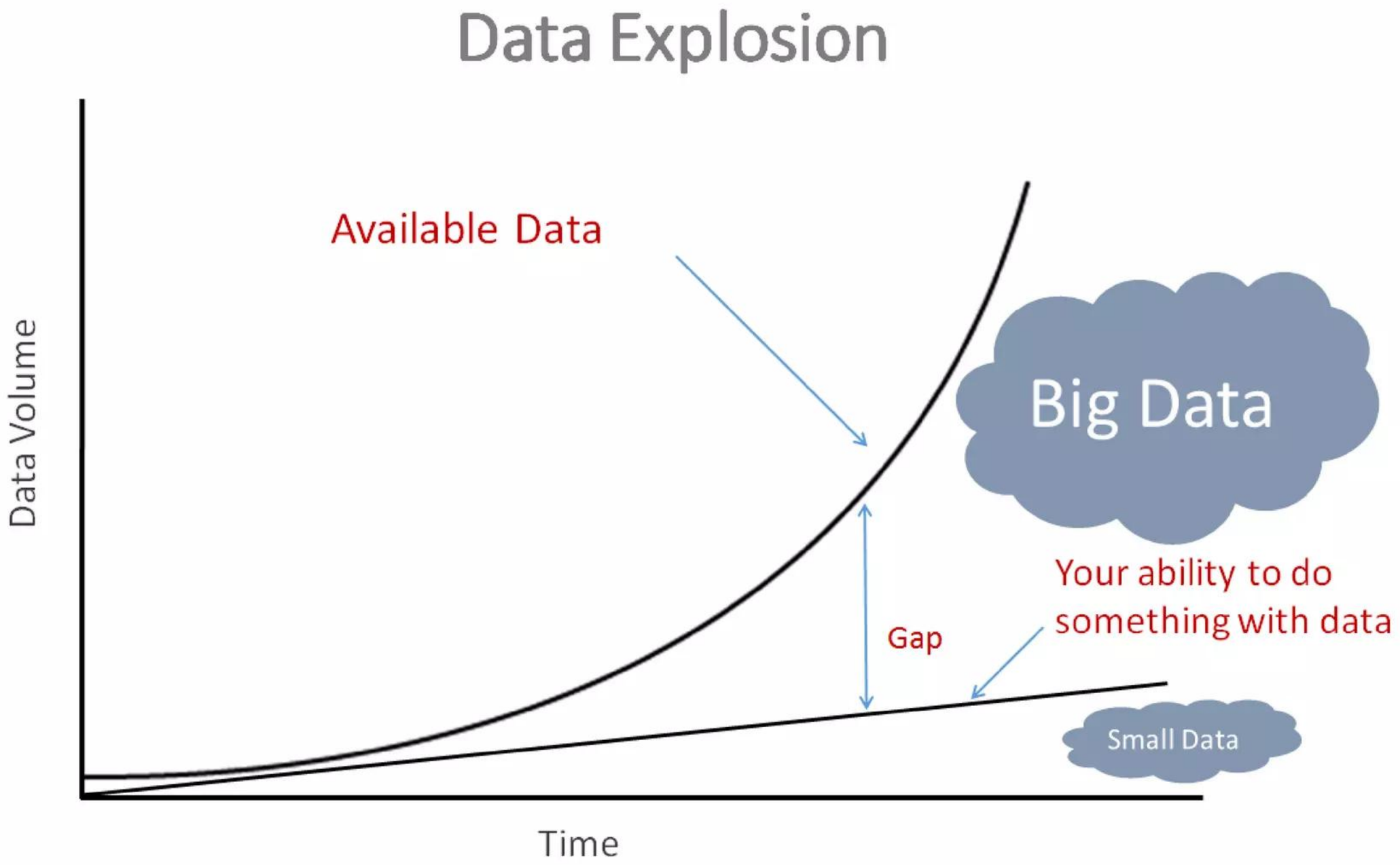
JVM Tech (tools) for ML

- weka
- Deeplearning4j (working with spark and GPU)
- H2O (working with spark and GPU, support Tensorflow, MxNet, and cafe)
- jcuDNN (JNI for wrapping nVidia cuDNN)
- Mahout
- MLlib spark

Artificial Intelligence in Industry

- **Fraud Detection System**
- **Dynamic Recommendation System and User Profiling**
- **Traveling Salesman Problem and Binpacking Problem for better warehouse management**
- **Social Media Analysis**
- Chatbot
- Company condition forecasting
- Governance simulation

The Complexity #1



The Complexity #2

- Big data : volume, variety, velocity, and veracity. (You might consider a fifth V, value.)
- Knowledge representation or the architecture of the model
- Unimplemented methods/algorithms in any libraries
- **Stack of methods**
- Data mostly unlabeled data
- Data resources (microservices)
- Features Engineering (especially from unstructured data)
- Machines (Hardware)
- High Performance Computing

Stack of Methods

- More complex methods and models
- Methods characteristic & behavior
- Methods customization
- Ex. Semi-supervised, Deep learning, features engineering
- Sample cases : our research in FDS and dynamic recommendation system

High Performance Computing

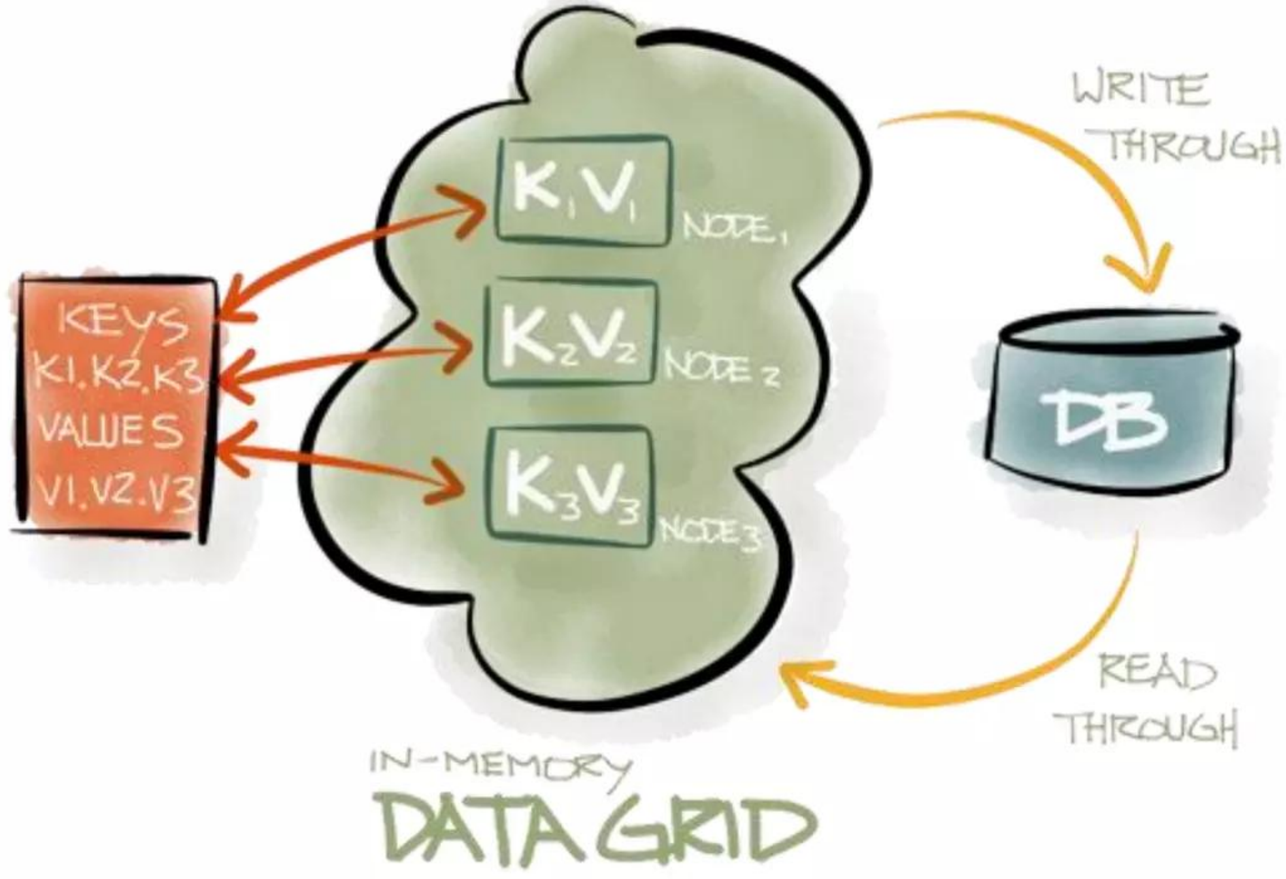
https://hadoop4usa.wordpress.com/2012/04/13/scale-out-up/

Machines

- **Cluster machine**
- **GPU machines (OpenCL and nvidia CUDA)**

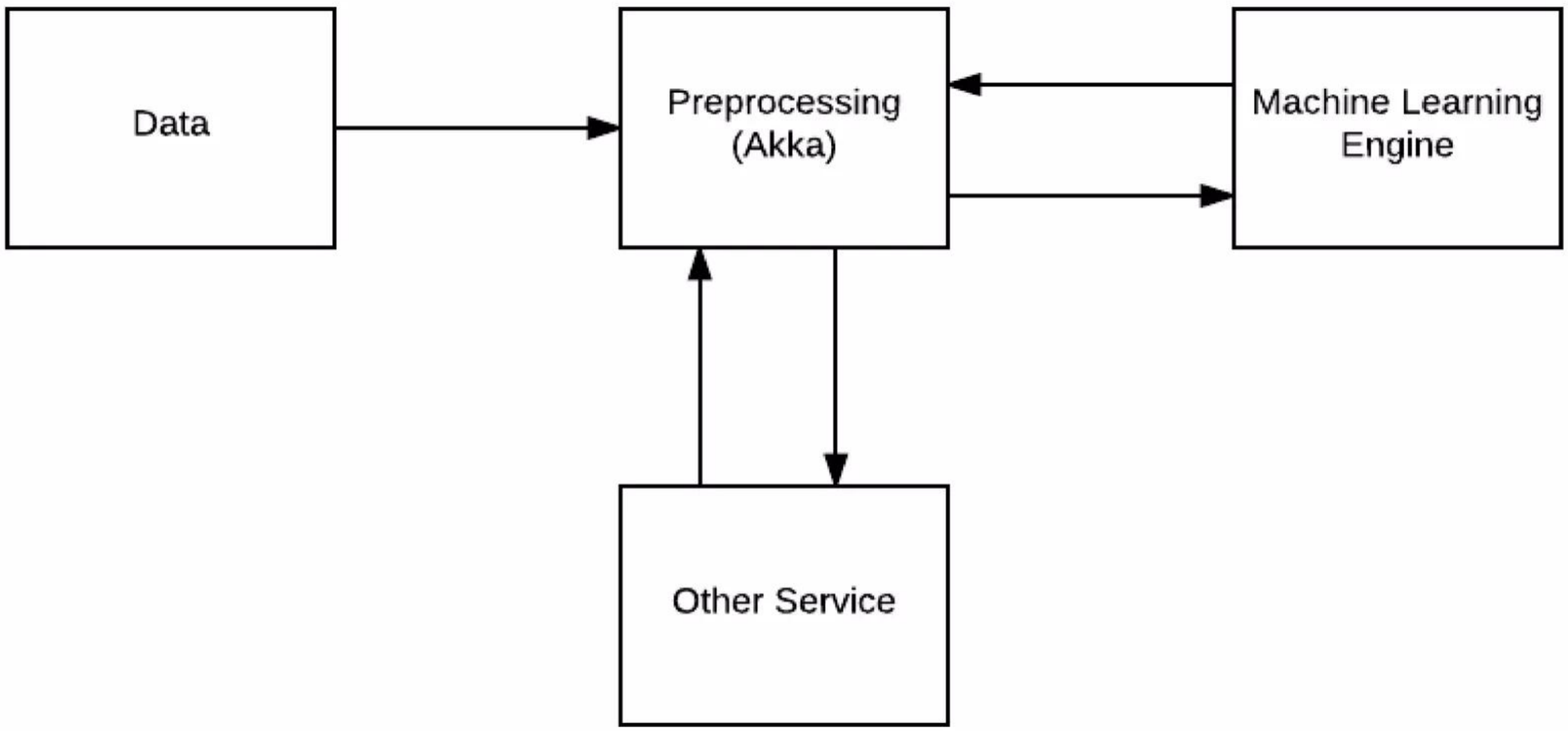
High Performance Computing #2

- **In-memory data fabric**: provides low-latency access and processing of large quantities of data by distributing data across the dynamic random access memory (DRAM), Flash, or SSD of a distributed computer system.



Machine Learning Flow

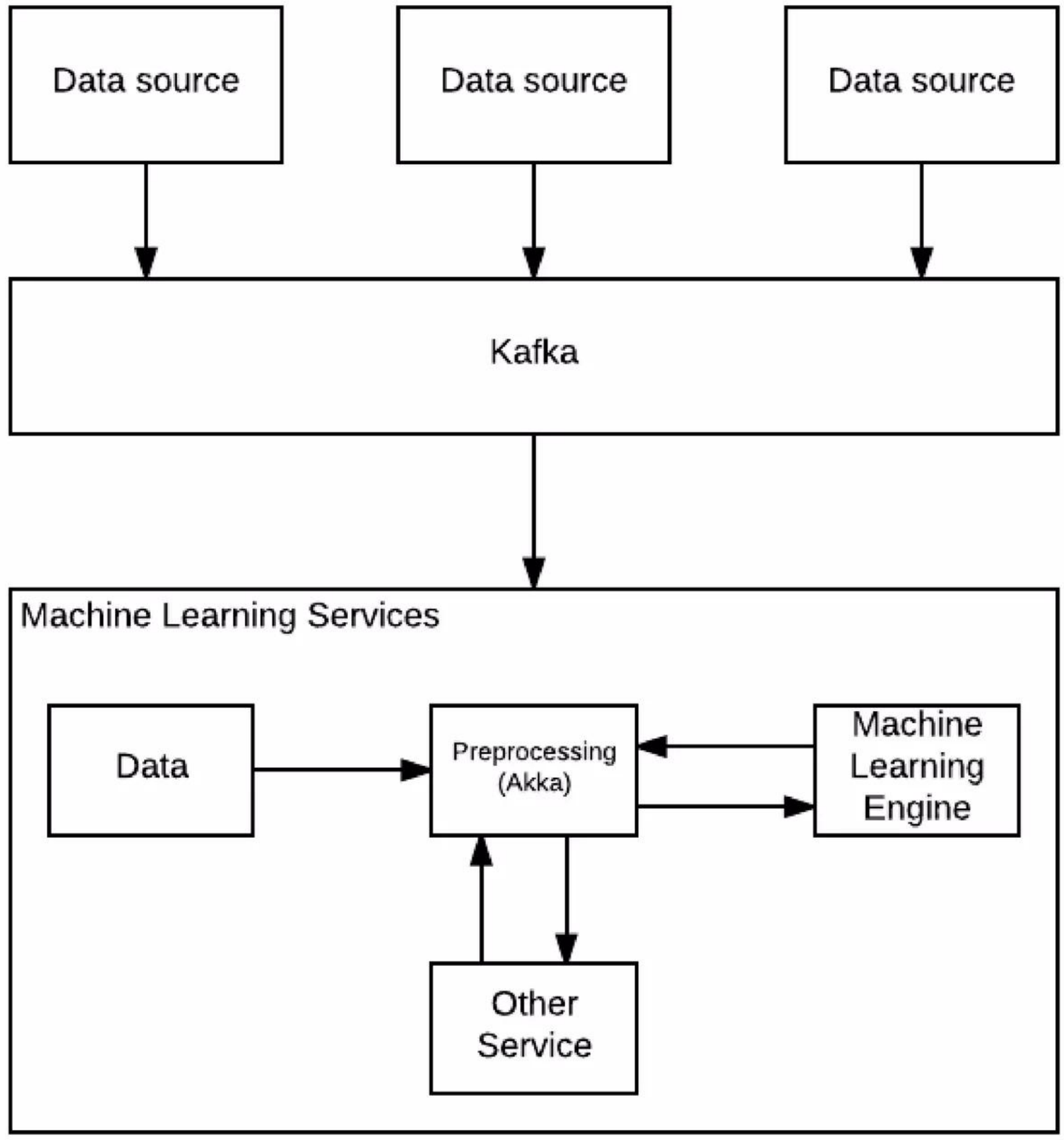
Scala + Play + Akka + Mongo



The problem is microservices

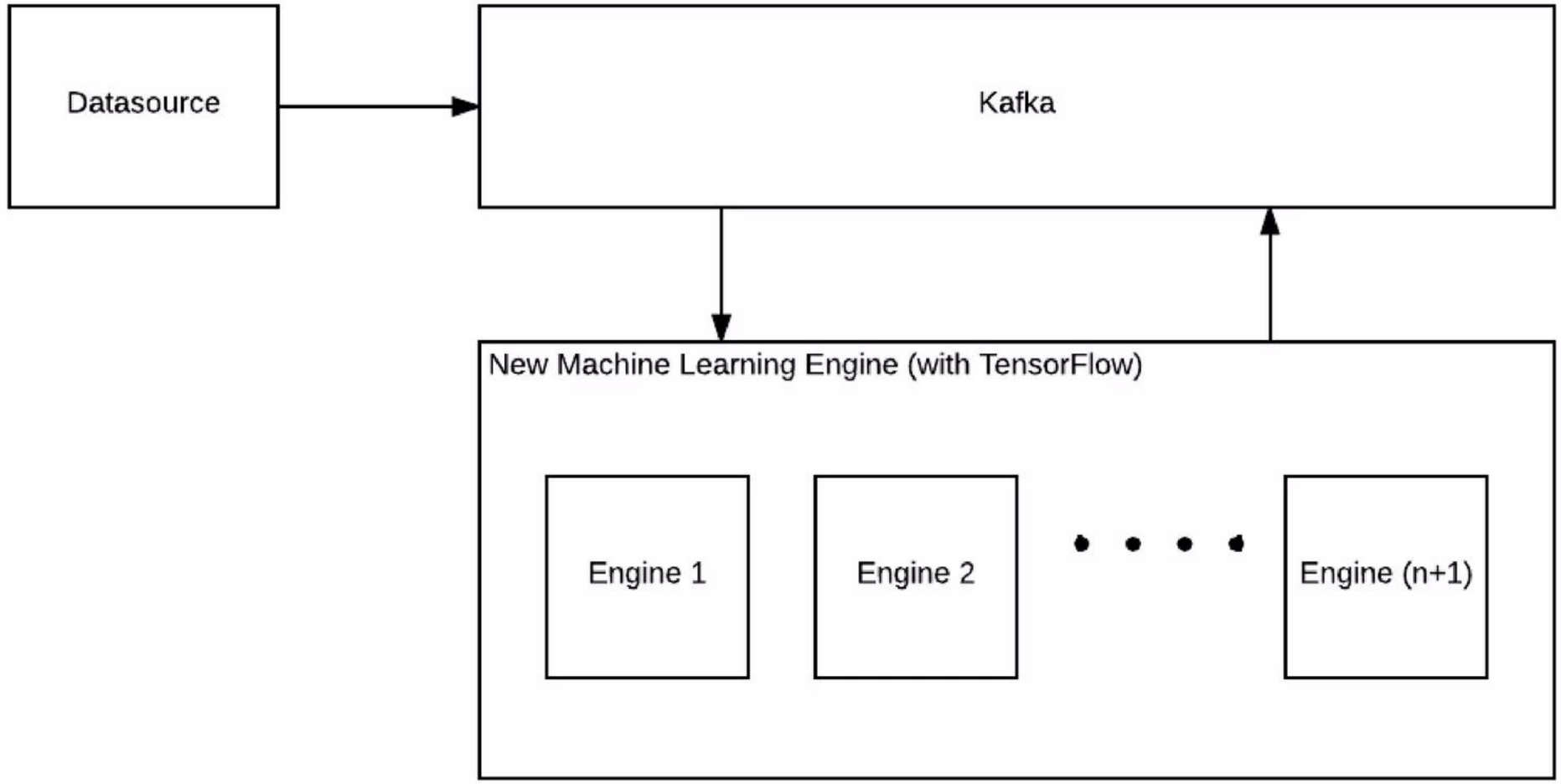
Next Machine Learning Flow

Scala + Play +
Akka + Mongo +
kafka

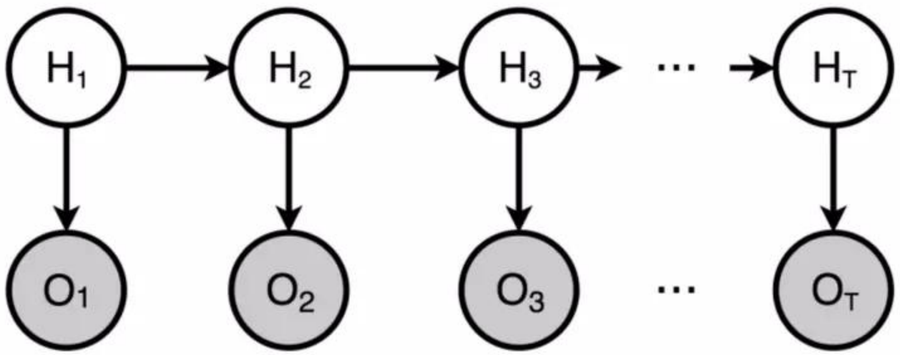


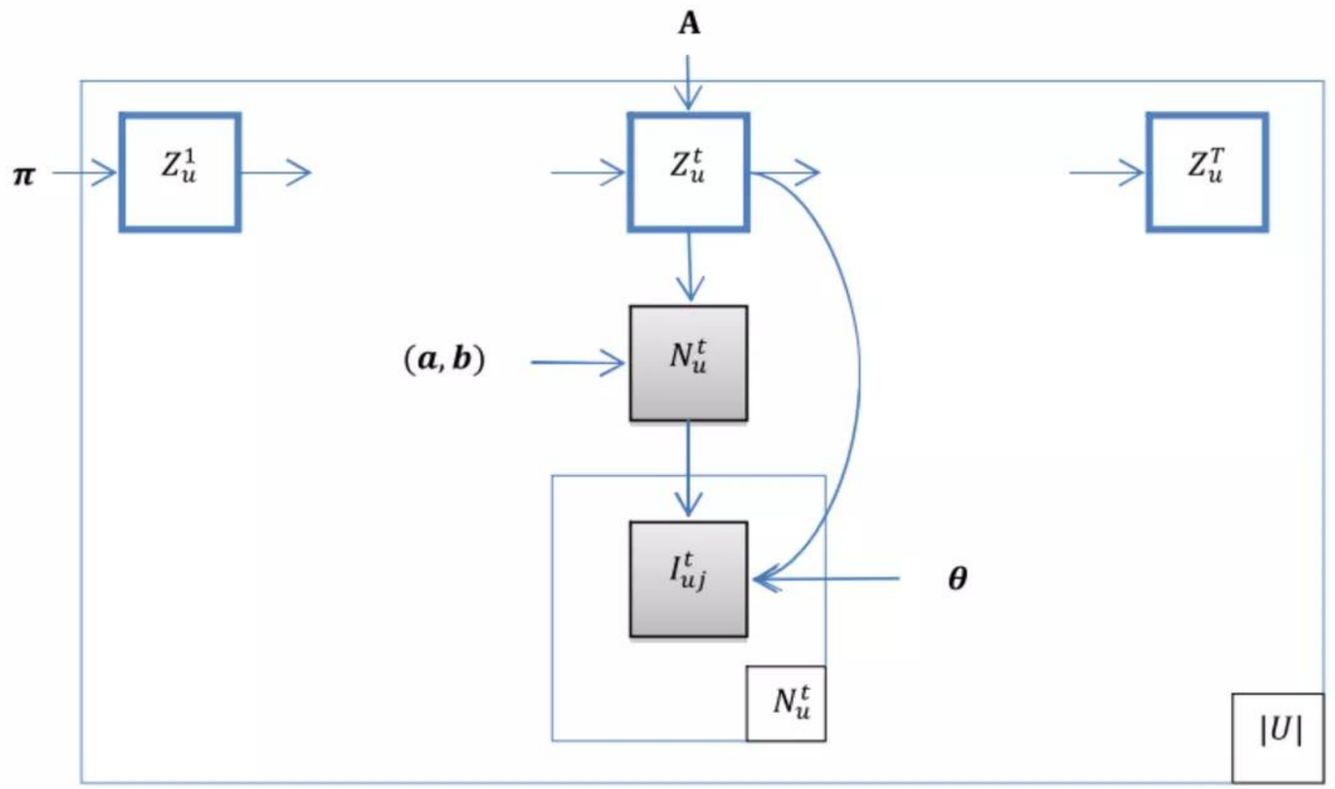
New Engine Machine Learning

- Kafka
- Tensorflow (utilize GPU)
- Spark
- Java/Scala/python



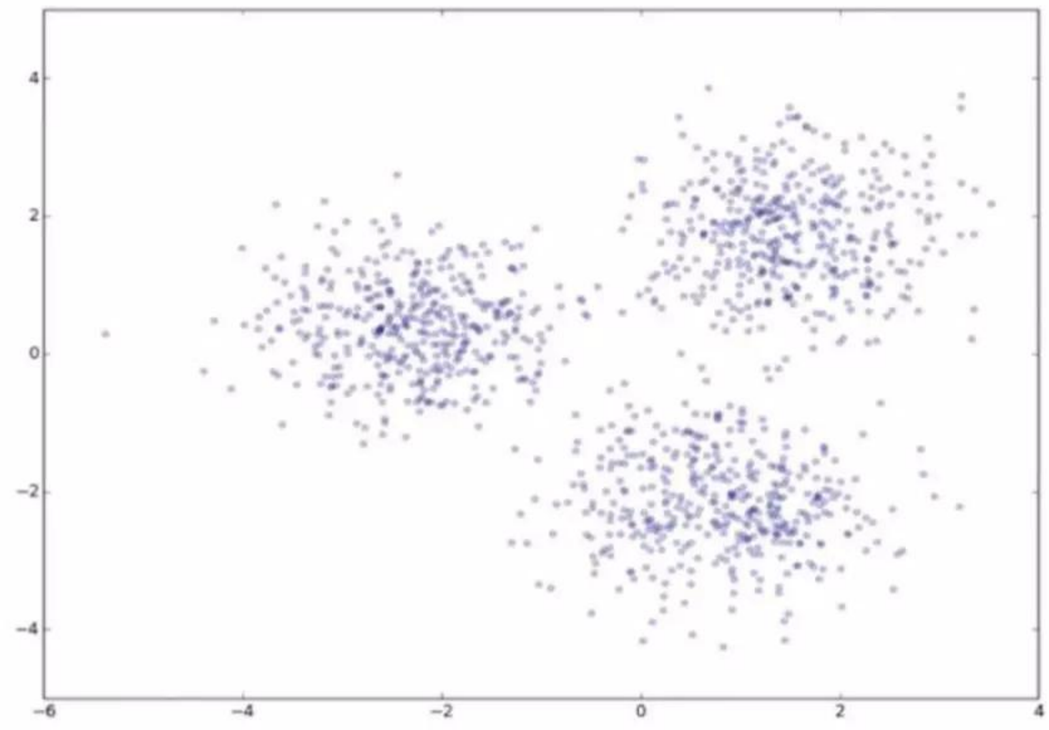
Dynamic Collaborative Filtering using HMM



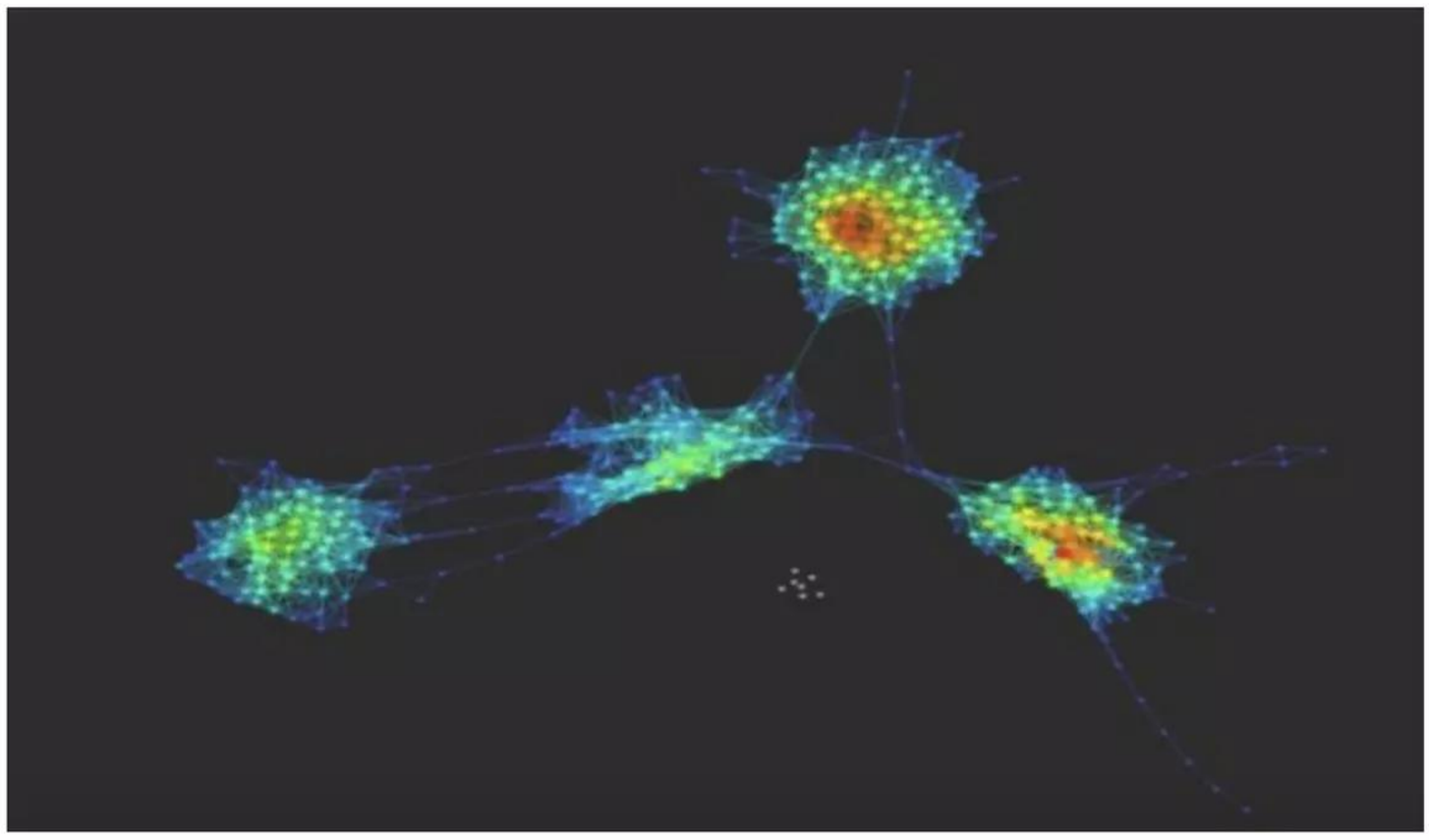


Algorithm	Control Parameters
1. Use data collected over time period 1 ... t_{trn} for training 2. ***Initialize*** $\boldsymbol{\pi}, \mathbf{A}, (\boldsymbol{a}, \boldsymbol{b}), \boldsymbol{\theta}$ to small random values 3. ***E-step***: compute $P(Z_u^t \mid I_u^{1:T})$ and $P(Z_u^{t-1}, Z_u^t \mid I_u^{1:T})$ using Equations (2) and (3). 4. ***M-step***: estimate $\boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\theta}$ using Equations (7), (8) and (10). Estimate $(\boldsymbol{a}, \boldsymbol{b})$ using Section 2.1 of (Minka 2002) 5. If expected log likelihood has not converged go to step 2 6. For each user u compute $R(i, u)$ of each item i for time period $t_{trn} + 1$ using Equation (12) a. ***Recommend top N*** items with highest $R(i, u)$	1. Set K to a value that maximizes the AIC score 2. Set length of the time period to 1 month a. Smaller if user preferences change quickly b. Larger if constrained by computing resource

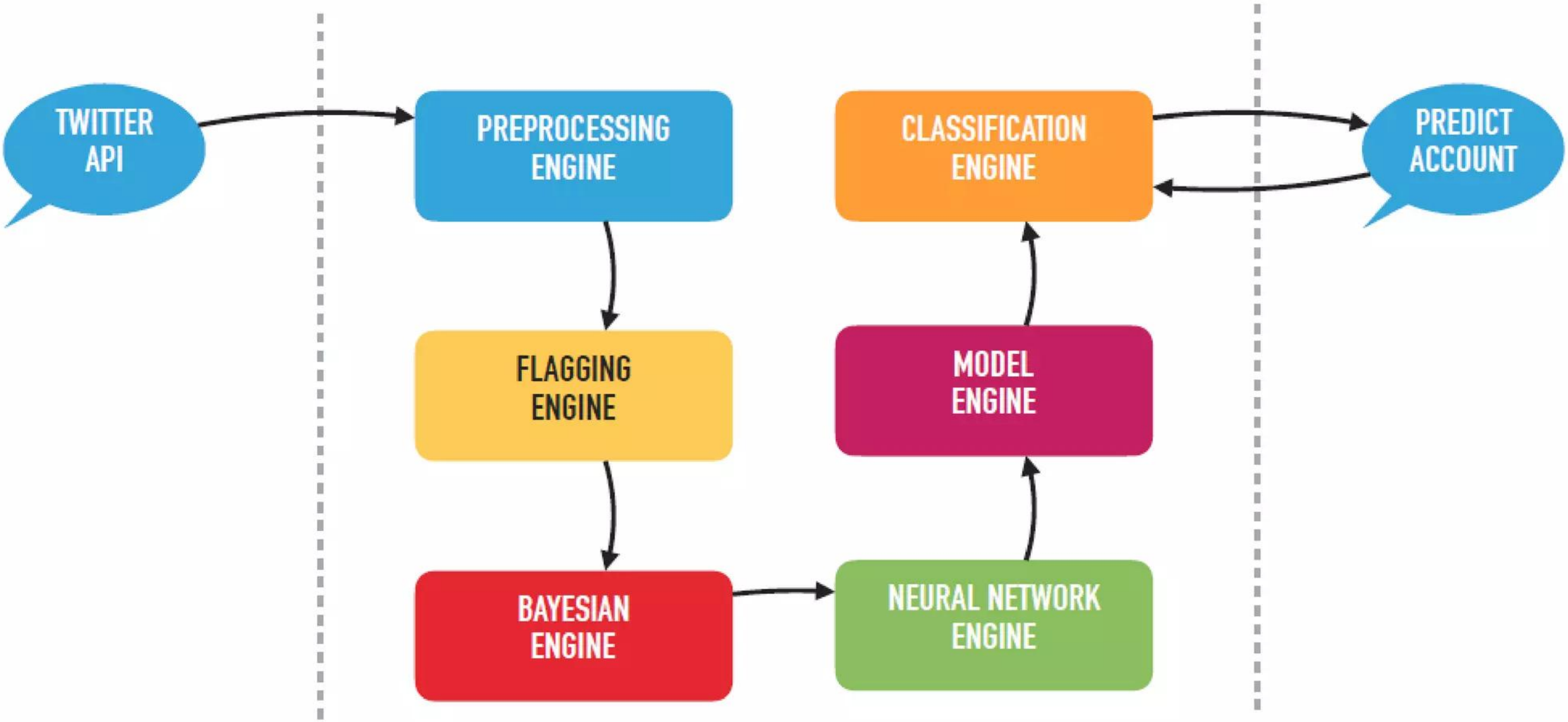
Topological Data Analysis



- Coordinate invariance
- Deformation Invariance
- Compressed Representations



Averi Project



Example

AVERI PROJECT

twitter prostitution account detection

THANK YOU

Any question?