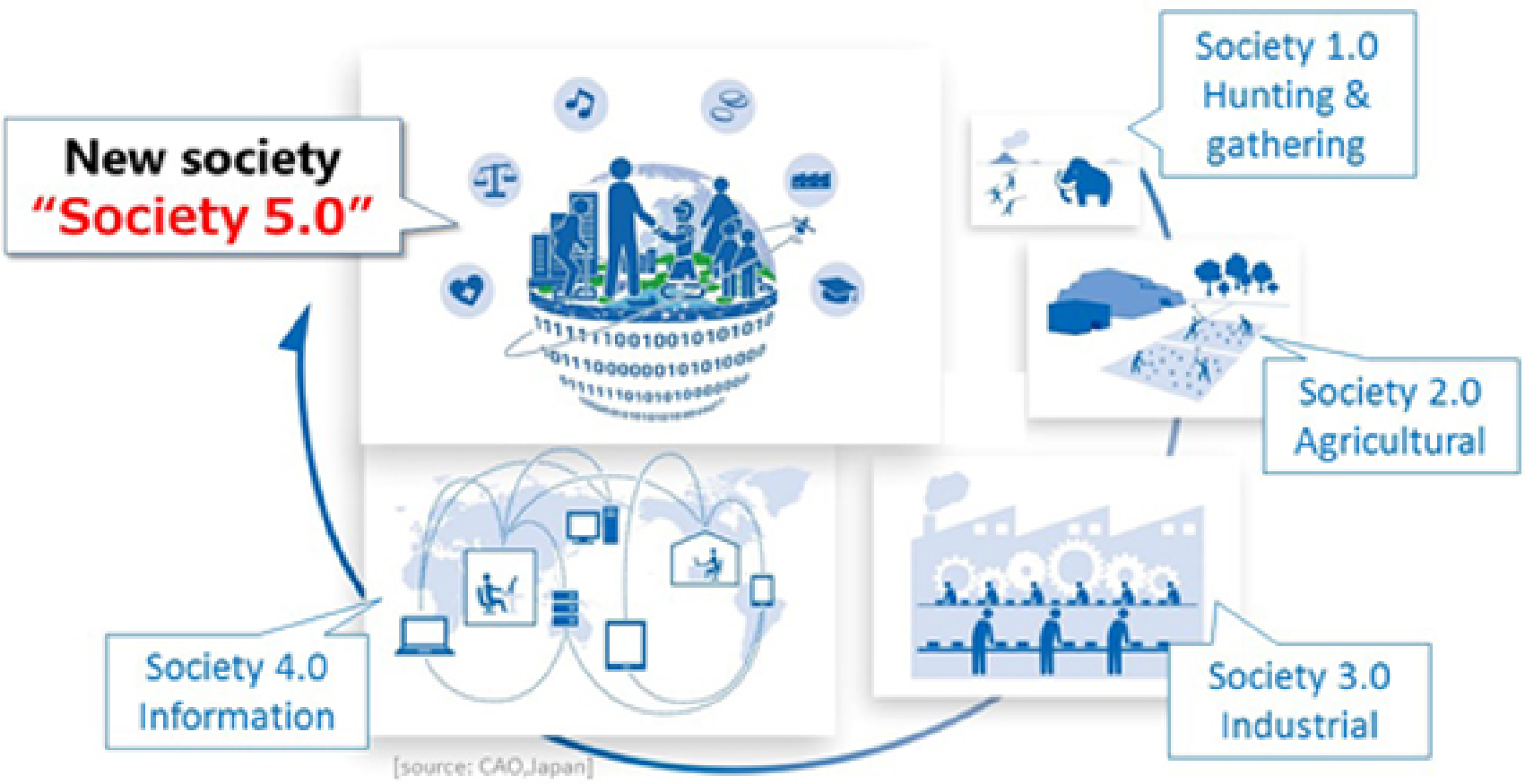
Understanding and predicting disease through AI and mathematical analysis of health and medical data

Team Leader, Health Data Mathematica Reasoning Team, MIH, RIKEN
Professor, Artificial Intelligence Medicine, Chiba University
Eiryo Kawakami, M.D. Ph.D.

February, 2020

SOCIETY 5.0 / PERSONALIZED SOLUTION

The Japanese government is promoting the policy of creating Society 5.0 aims to tackle several challenges by the digitalization across all levels of the Japanese society and the (digital) transformation of society itself.



Mission

Precision medicine / Personalized predictive medicine by deep clinical phenotyping

Physical

Real life have to be recognized as actual physical characteristics & life course.

↓

Measurement (**Deep Clinical Phenotyping Data**)

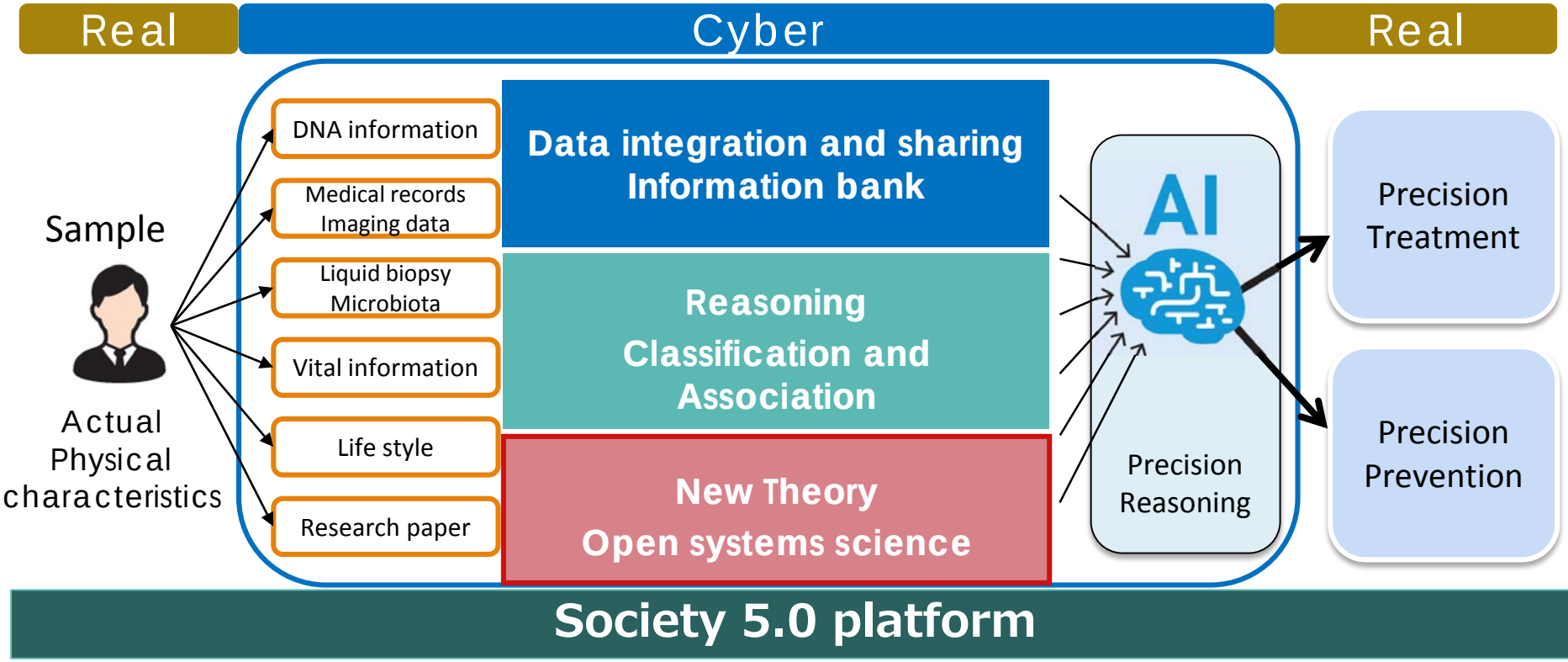
↓

Multidimensional data base ➡ Mathematically represent and redefine diseases

Cyber

Goals

■ Platform for deep phenotyping



A	B	C	D	E
Measurement Platform	Information sharing Platform	Classification Platform Remembrance Information Geometry	Association Platform Causality Synchronization Self organization	Dependability Platform New Regulatory Science

Machine learning stratification and prediction of ovarian cancer

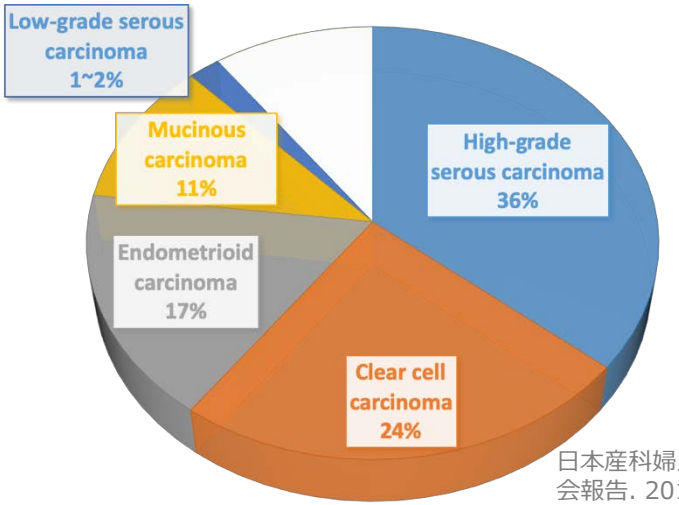
Ovarian cancer in Japan

- **10,048** new patients/year
- **4,745** death/year
- **~50%** of patients are found in late stage
- **Prognosis is bad** specifically in late stage

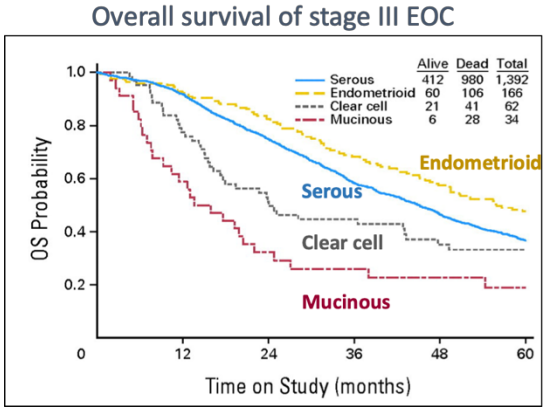
Stage		5-years relative survival rate (%)
Early	I	87.4
	II	66.4
Late	III	44.2
	IV	28.3

人口動態統計によるがん死亡データ. 2017年.
地域がん登録全国合計によるがん罹患データ. 2014年.
日本産科婦人科学会婦人科腫瘍委員会報告. 2012年患者年報
全国がん（成人病）センター協議会の生存率共同調査（2017年9月）

Histological types of ovarian cancer



日本産科婦人科学会婦人科腫瘍委員会報告. 2012年患者年報等（改変）



William E et al. J ClinOncol. 2007

Supervised prediction of ovarian cancer

32 preoperative blood markers (CA125, CA19-9, LDH, Fbg, Alb, …)

prediction

Ovarian tumor characteristics (malignancy, stage, histotype)

334 patients with EOC and **101** patients with benign ovarian tumor treated in **the Jikei University** between 2010 and 2017

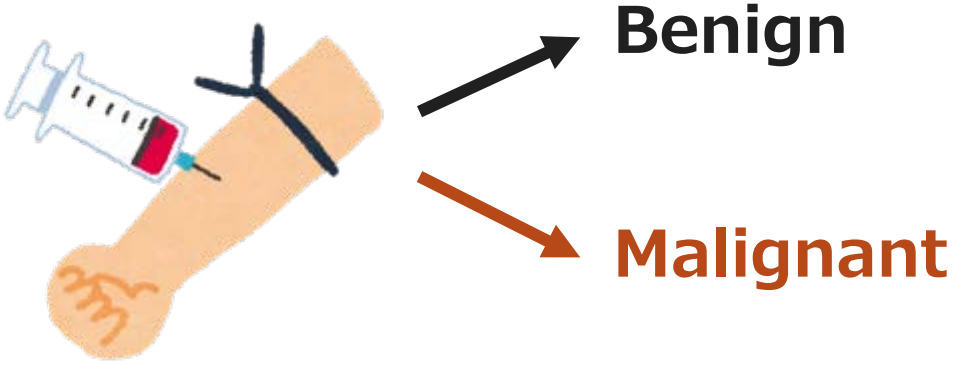
Split datasets into training and validation cohorts (2-fold cross validation)

Train classifiers with training cohort

Check accuracy with validation cohort

Precise preoperative diagnosis based on Ensemble learning

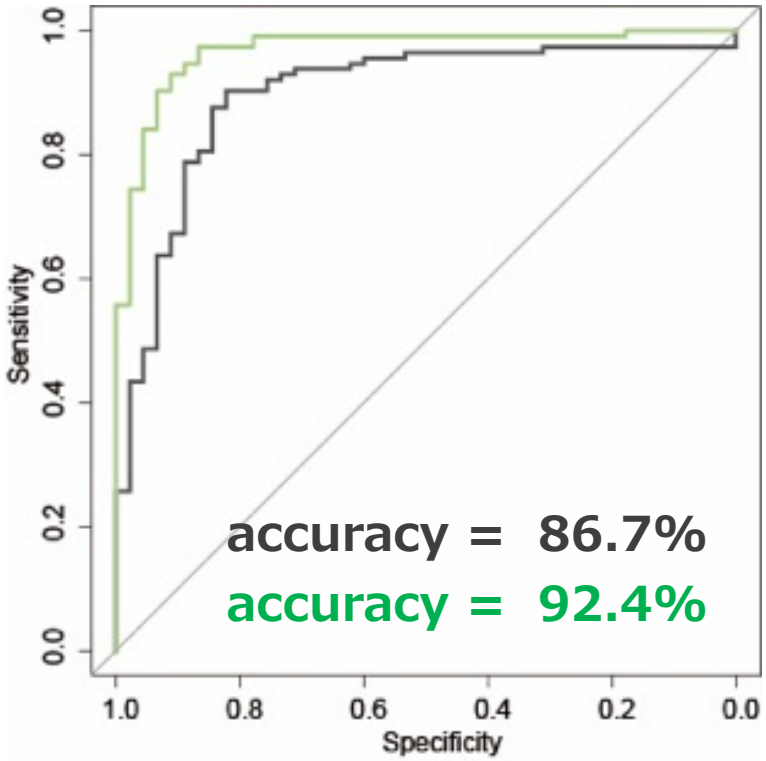
101 Benign tumor
334 Malignant cancer



Random Forest

Ensemble of thousands of decision trees

- ✓ Robust to outliers
- ✓ Not depends on type/distribution of variables
- ✓ Consider interaction between variables

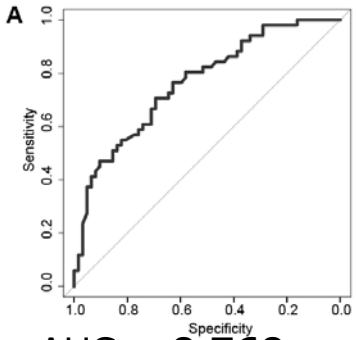


Logistic regression

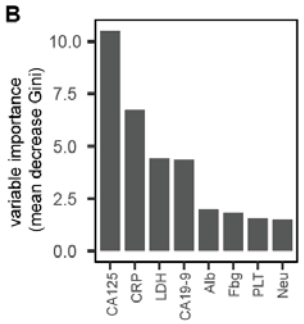
Random Forest

Are known categories really predictable?

Stratify early stage and late stage

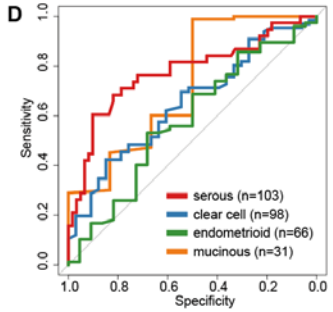


AUC = **0.760**
Accuracy = **69.0%**

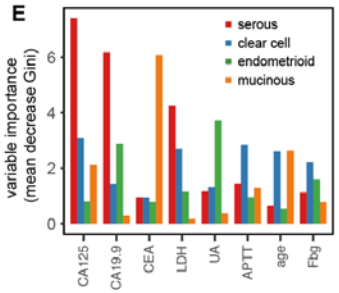


Important factors
CA125, **CRP**, and **LDH**

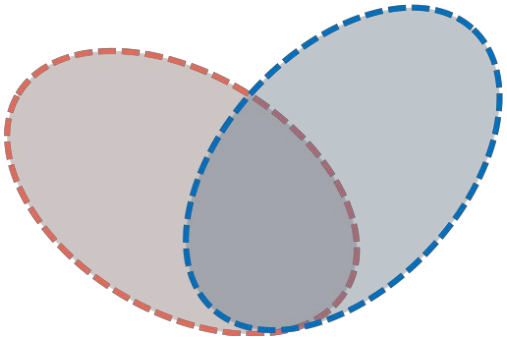
Stratify histological types



AUC = **0.597–0.785**
Accuracy = **55.6–96.0%**



Important in **serous**
CA125, **CA19-9**, and **LDH**
Important in **mucinous**
CEA



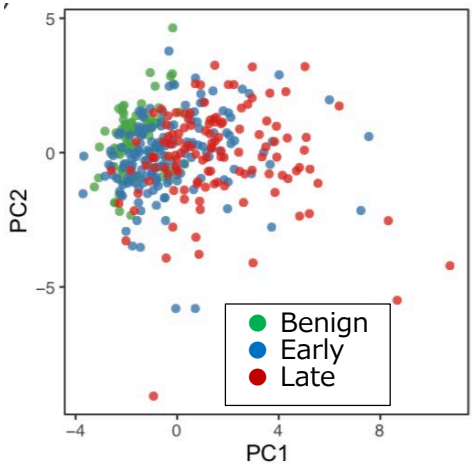
- ✓ There can be overlap between early and late stages
- ✓ Late stage cancer is sometimes mis-diagnosed due to hidden metastasis
- ✓ Supervised prediction cannot find such overlaps

Unsupervised stratification

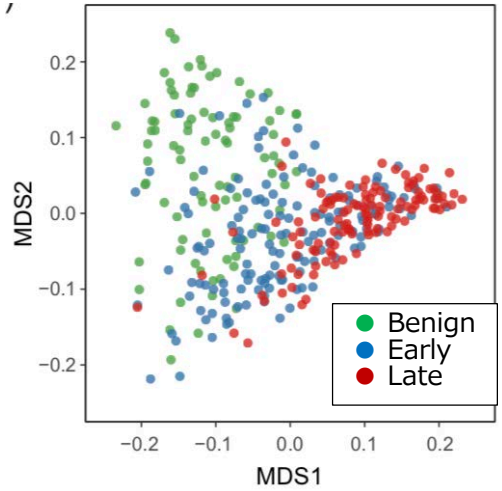
Kawakami et al.
Clin. Cancer Res. 2019

Unsupervised learning to extract difficult-to-recognize patterns

Conventional PCA



URF + MDS



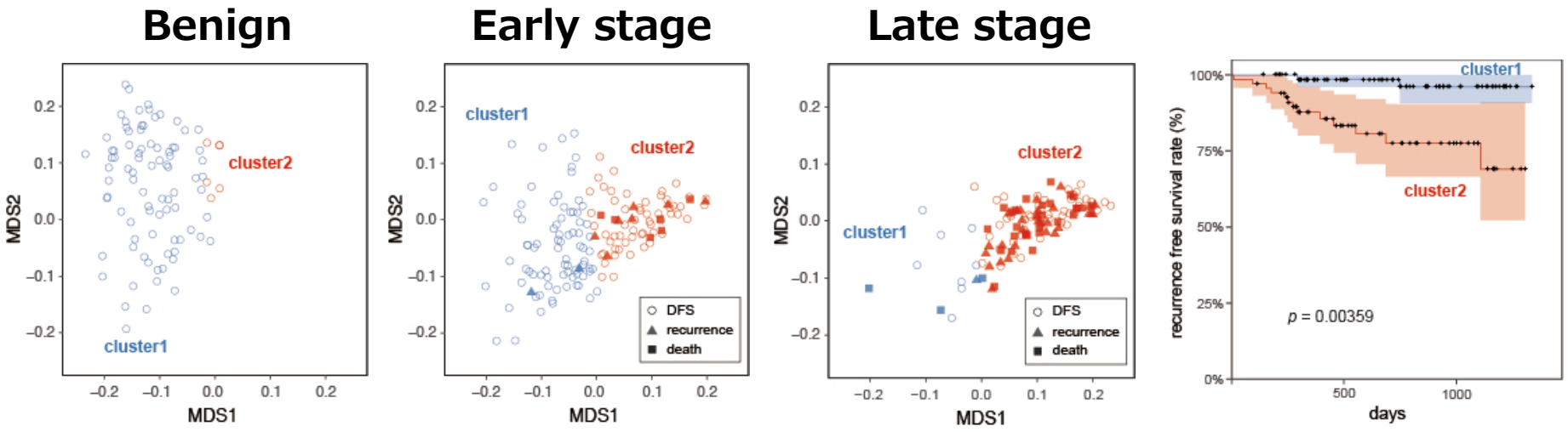
EOC and benign ovarian tumor patients with 32 preoperative blood markers

Calculate similarity of patients with unsupervised random forest (URF)

Visualize the distribution of patients with multidimensional scaling (MDS)

Extract patterns of patients based on multi-dimensional data

Novel clusters in early stage EOC



Kawakami et al. Clin. Cancer Res. 2019

Early ovarian cancer patients distributed across two clusters

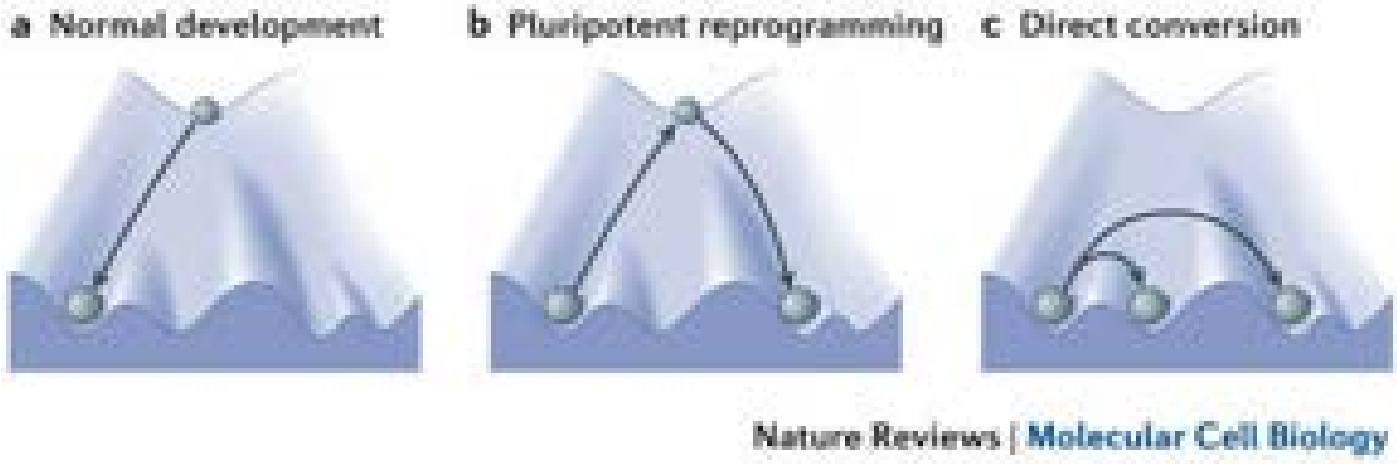
Cluster 1: Benign type

Cluster 2: Late stage type

By using unsupervised machine learning, we revealed the disease clusters that have not been recognized even by clinicians

Landscape modeling of life system

Dr. Tetsuo Ishikawa



Ladewig 2013
Nature Reviews Molecular Cell Biology

Epigenetic landscape

A concept proposed by Waddington in 1940. The model describes cell differentiation as an analogy of a dynamical system where **"state transition occurs as the ball rolls on the terrain"**.

Data-driven reconstruction of landscape

High dimensional fMRI data

Mean DMN activity

Mean FPN activity

Mean Aud activity

binarize

2^n possible states

1100101010111100011

0010100011110001110

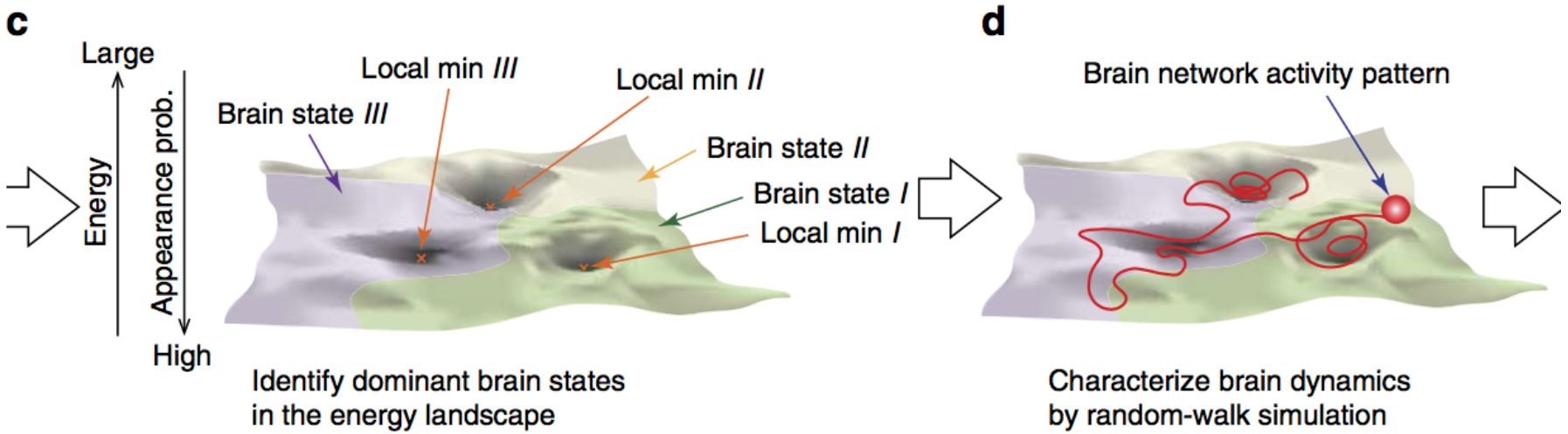
⋮

1011011000110001111

Energy of each state is calculated from observed frequency with Ising model

$$\text{probability: } P(\boldsymbol{\sigma}|\boldsymbol{h},\mathbf{J}) = \frac{\exp\left[-E(\boldsymbol{\sigma}|\boldsymbol{h},\mathbf{J})\right]}{\sum_{\boldsymbol{\sigma}'}\exp\left[-E(\boldsymbol{\sigma}'|\boldsymbol{h},\mathbf{J})\right]}$$

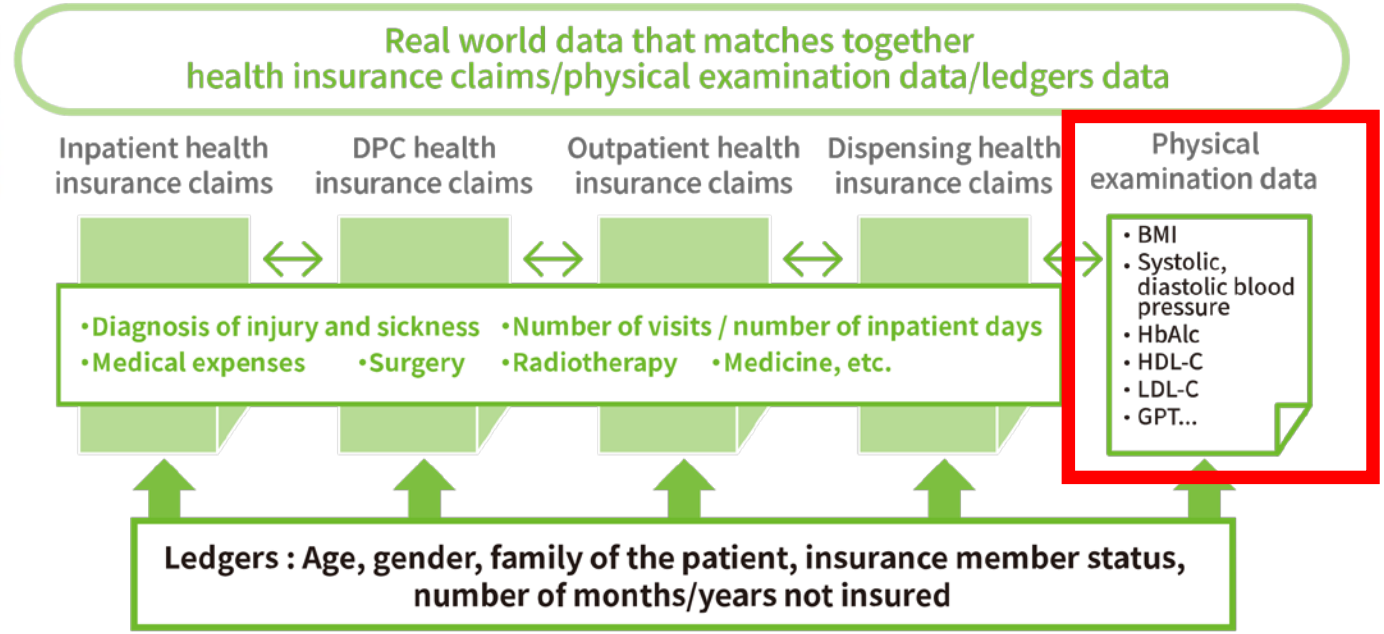
$$\text{energy: } E(\boldsymbol{\sigma}|\boldsymbol{h},\mathbf{J}) = -\sum_{i=1}^{N} h_i\sigma_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{\substack{j=1\\ j\neq i}}^{N} J_{ij}\sigma_i\sigma_j$$



Adapted from Watanabe & Rees (2017) Nature Communications

～2.5M physical examination records

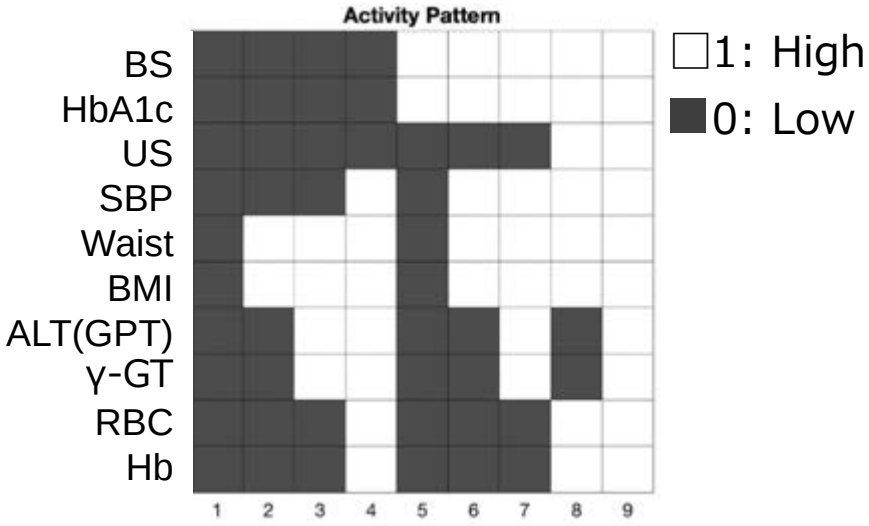
JMDC

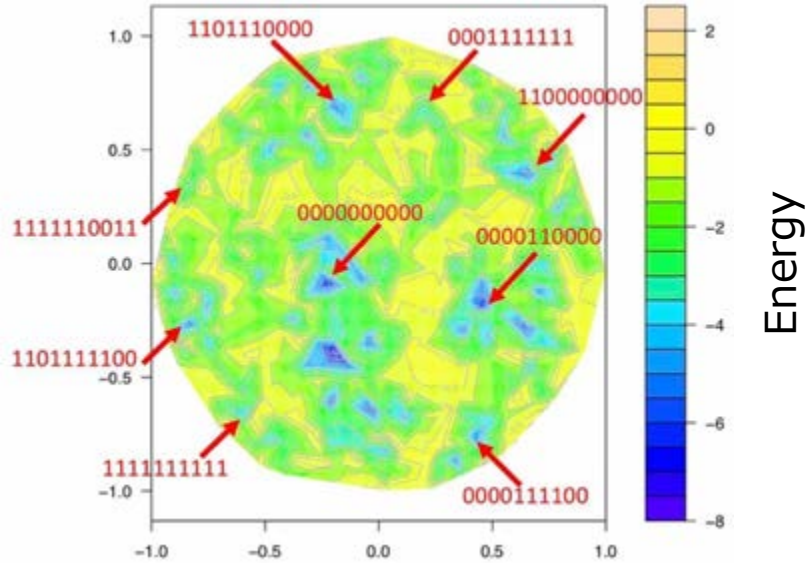


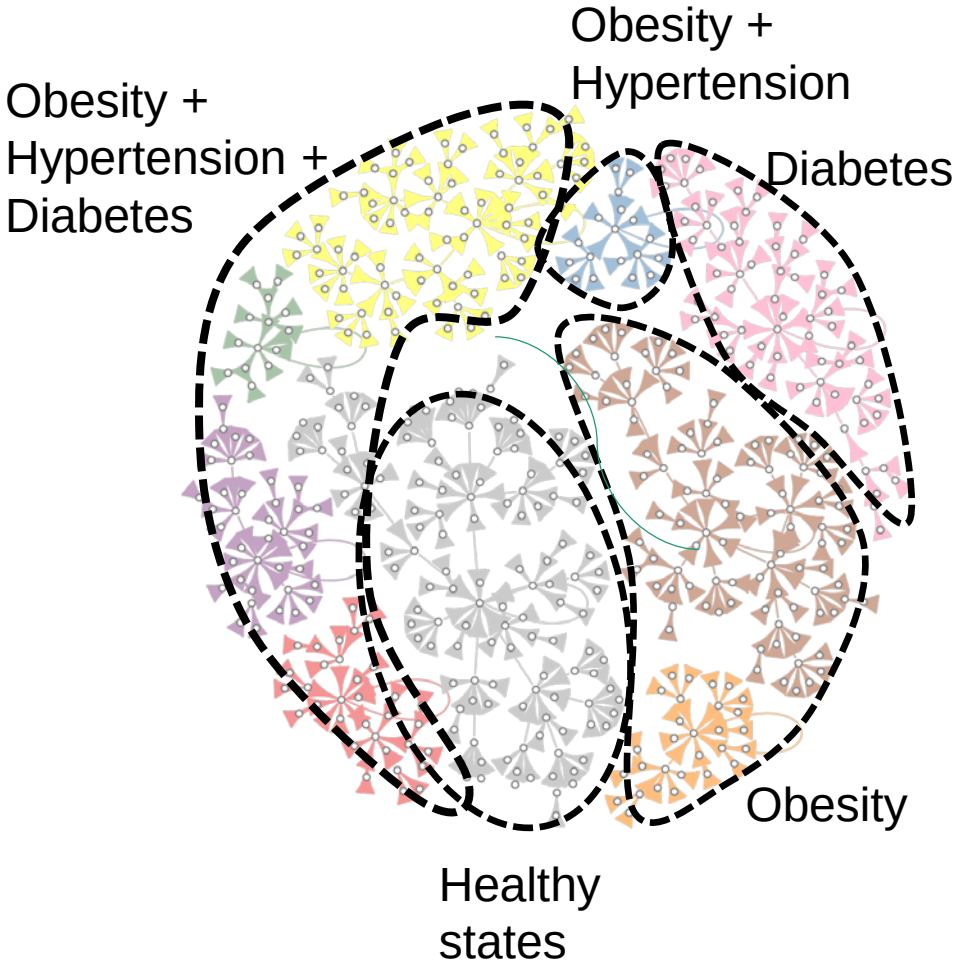
1	母数.csv	7,163,773 records	405,773 KB
2	患者.csv	331,663 records	18,787 KB
3	施設.csv	95,745 records	6,121 KB
4	レセプト.csv	26,088,779 records	2,103,235 KB
5	傷病.csv	61,444,578 records	14,967,408 KB
6	医薬品.csv	60,148,880 records	28,782,093 KB
7	診療行為.csv	179,979,079 records	35,279,923 KB
8	材料.csv	957,608 records	206,889 KB
9	健診.csv	2,581,416 records	683,131 KB

Physical examination data

Landscape of diabetes onset process

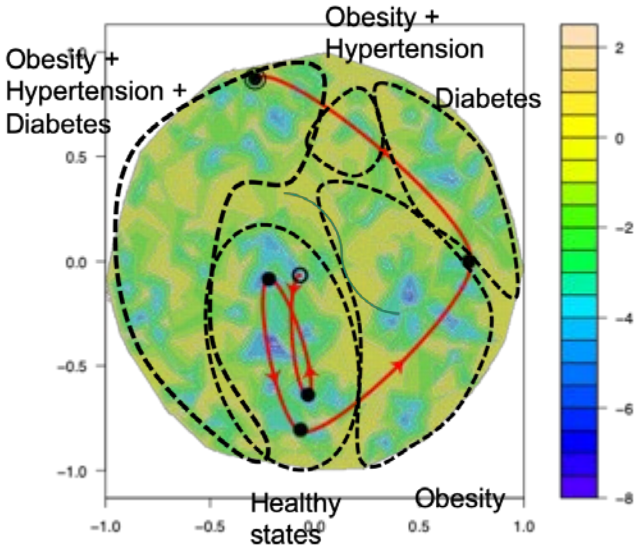




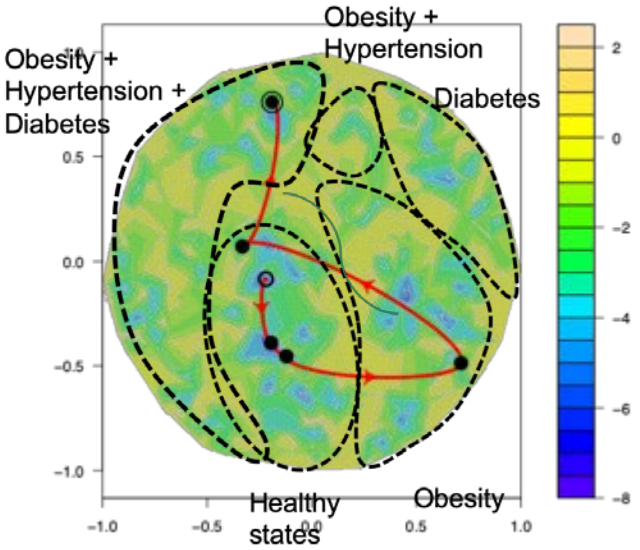


Typical diabetes onset trajectories

Patient 1

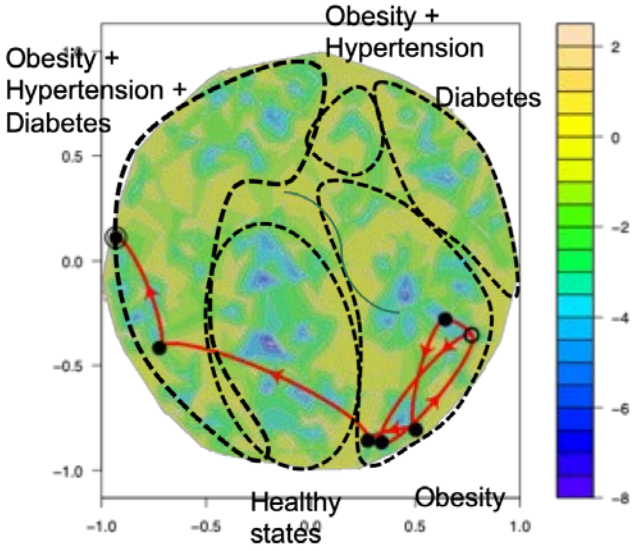


Patient 2

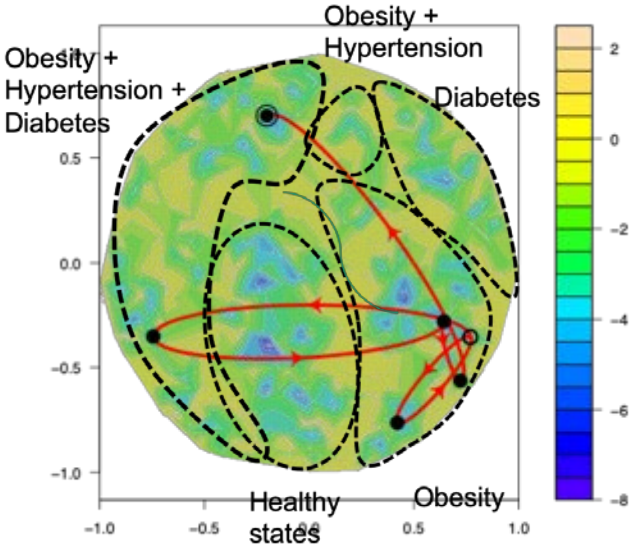


- 1st exam
- Exam
- Onset

Patient 3



Patient 4



Key messages

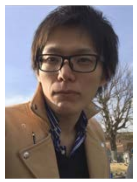
- ✓ Existing clinical classification of disease is not complete. We should re-define healthy/disease states based on clinical phenotype data
- ✓ By using unsupervised machine learning, we can extract difficult-to-recognize patterns that have not been recognized even by clinicians
- ✓ Tracking the onset and progress of diseases may lead to appropriate individualized intervention and treatment

Acknowledgement

RIKEN MIH

Tetsuo Ishikawa, Ph.D.
Particle physics
Neuroscience

Keita Koseki, M.S.
Mathematics
Medicine



Yuki Goshima, M.S.
Mathematics
Medicine

Yuichiro Ichihara
Statistics
Medicine

Tomohiro Ogino
Medicine
Mathematics

Kyohei Sano
Medicine
Deep Learning

Engineering Network
RIKEN

Makoto Yamada
Shunsuke Tagami
Hiroshi Kawasaki

AI medicine
Chiba University

Keiko Yamazaki, Ph.D.
Molecular biology
Genomics

Megumi Ohya, M.D.
Medicine
Systems biology, Deep Learning

Hidehiro Yokota, M.S.
Experimental physics
Medicine

Keisuke Chikamoto, B.S.
Economic engineearing
Medicine

Shusuke Kobayashi, B.S.
Bayesian modeling
Medicine

Yoko Kurokawa, B.S.
Physics
Medicine

Ryuya Urisaka