Hong Kong Baptist University, Faculty of Science,
60th Anniversary Distinguished Lecture

Feb. 18, 2022

Challenges in Machine Learning Research

Masashi Sugiyama

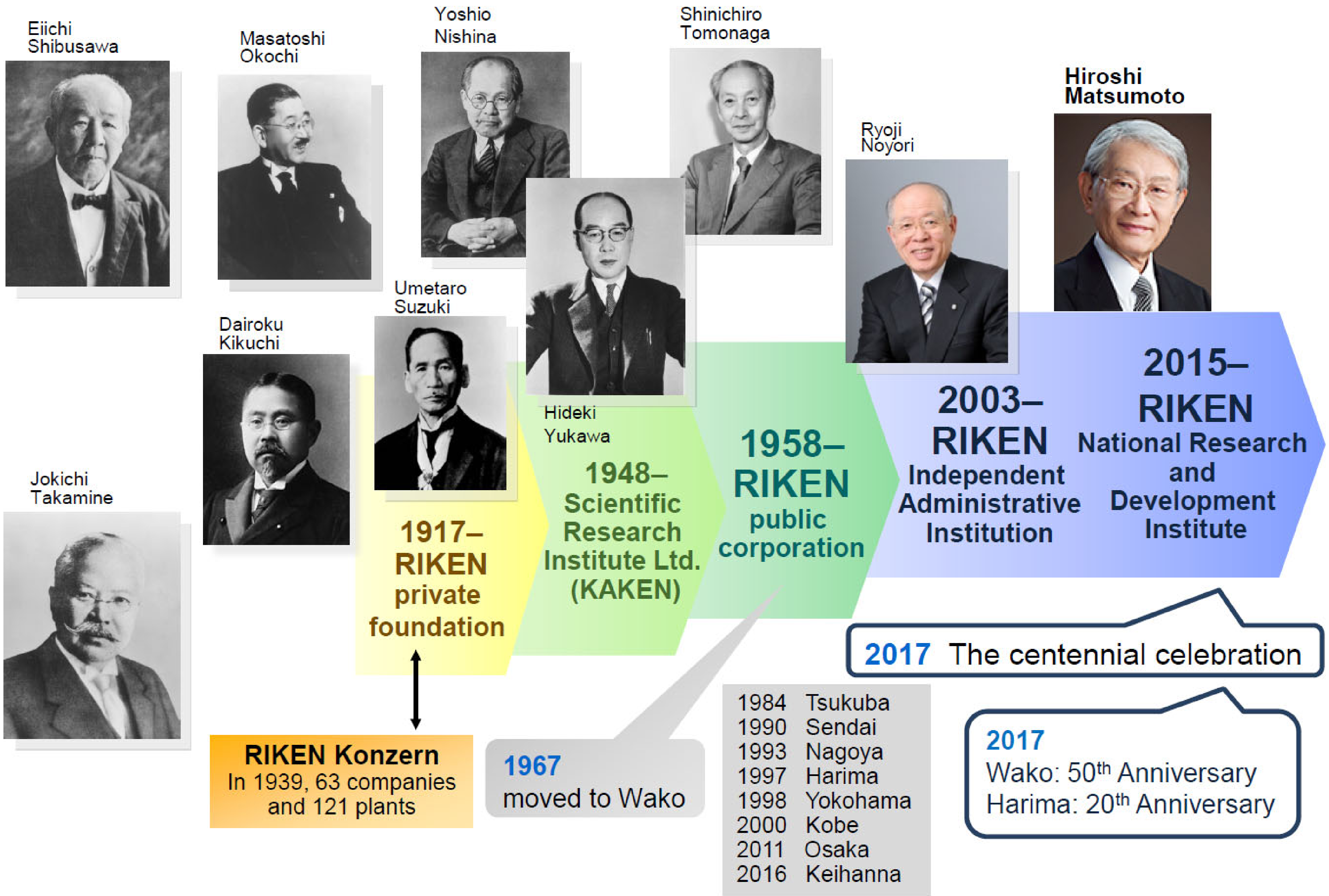
Director, RIKEN Center for Advanced Intelligence Project (AIP)

Professor, the University of Tokyo

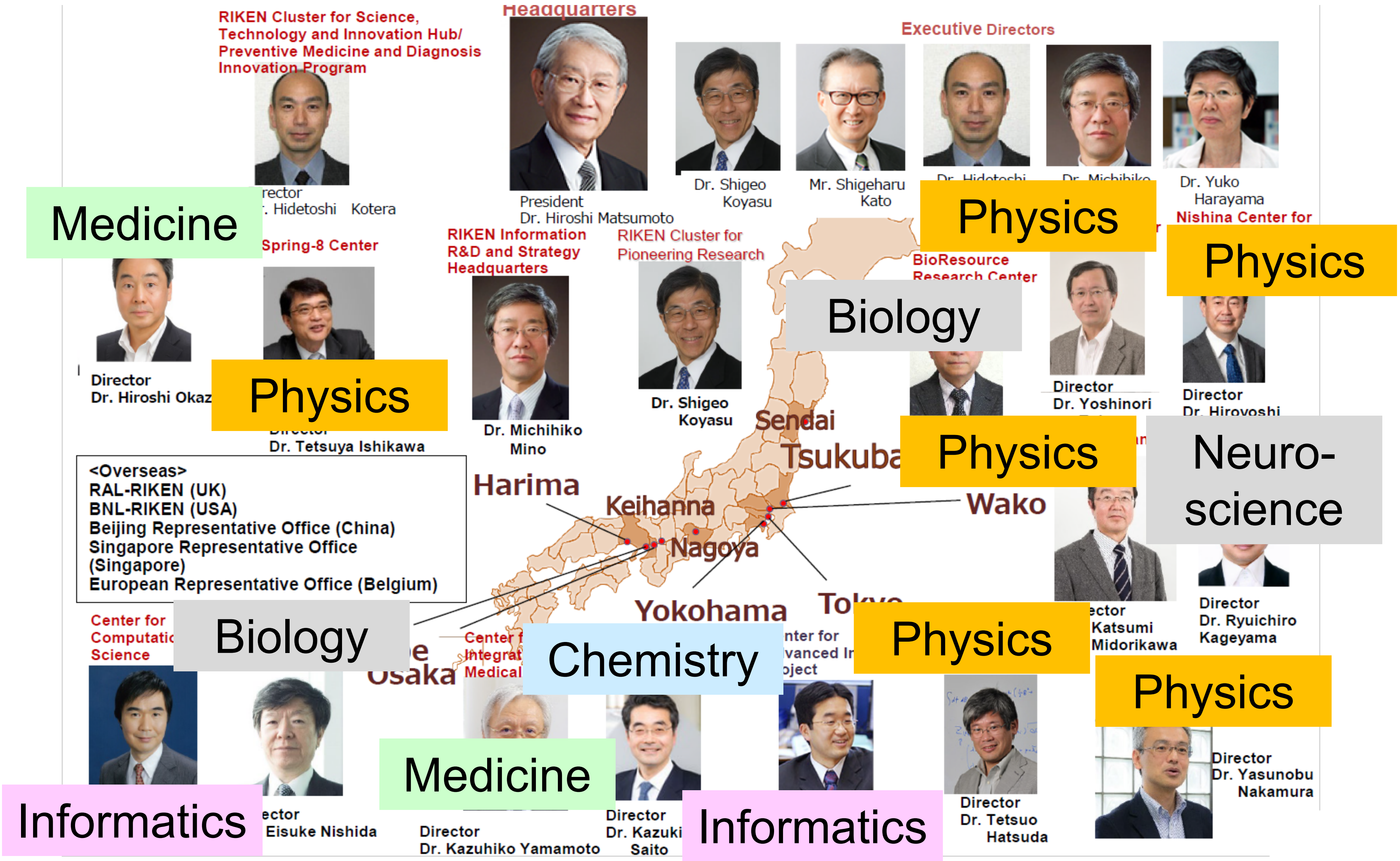
What is “RIKEN”?

- Name in Japanese: 理化学研究所
  - Pronounced as: rikagaku kenkyusho
  - Meaning: Physics and Chemistry Research Institute
- Acronym in Japanese: 理研 (RIKEN)

Brief History



Research

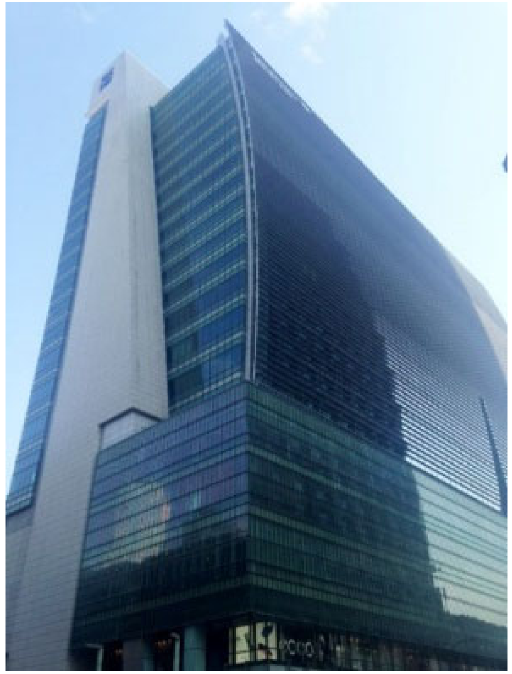

Headquarters
RIKEN Cluster for Science, Technology and Innovation Hub/ Preventive Medicine and Diagnosis Innovation Program
President
Dr. Hiroshi Matsumoto
Dr. Shigeo Koyasu
Mr. Shigeharu Kato
Executive Directors
Dr. Yuko Harayama
Medicine
Physics
Physics
Spring-8 Center
RIKEN Information R&D and Strategy Headquarters
RIKEN Cluster for Pioneering Research
BioResource Research Center
Nishina Center for
Biology
Director
Dr. Tetsuya Ishikawa
Dr. Michihiko Mino
Dr. Shigeo Koyasu
Director
Dr. Yoshinori
Director
Sendai
Tsukuba
Physics
Neuro-science
<Overseas>
RAL-RIKEN (UK)
BNL-RIKEN (USA)
Beijing Representative Office (China)
Singapore Representative Office (Singapore)
European Representative Office (Belgium)
Harima
Keihanna
Nagoya
Wako
Yokohama
Osaka
Center for Computational Science
Biology
Chemistry
Physics
Katsumi Midorikawa
Director
Dr. Ryuichiro Kageyama
Physics
Director
Dr. Yasunobu Nakamura
Medicine
Informatics
Eisuke Nishida
Director
Dr. Kazuhiko Yamamoto
Director
Saito
Informatics
Director
Dr. Tetsuo Hatsuda

What is RIKEN-AIP?

- RIKEN founded Center for Advanced Intelligence Project (AIP) in 2016, under Ministry of Education, Culture, Sports, Science and Technology (MEXT).

MEXT
MINISTRY OF EDUCATION, CULTURE, SPORTS, SCIENCE AND TECHNOLOGY-JAPAN

Main office located in the heart of Tokyo



Distributed office across Japan

In-house GPU servers

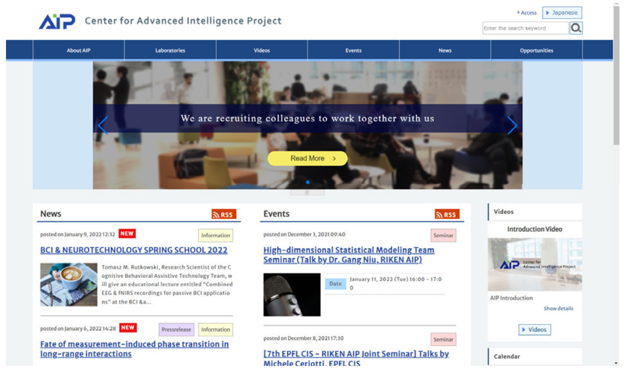
Open discussion space

AIP's 5 Missions

- Develop next-generation AI technology:
  - machine learning & optimization theory, etc.
- Accelerate scientific research:
  - cancer, material, genomics, etc.
- Solve socially critical problems:
  - natural disaster, elderly healthcare, etc.
- Study of ethical, legal and social issues of AI:
  - ethical guidelines, personal data, etc.
- Human resource development:
  - researchers, engineers, etc.

Statistics

As of Apr. 1, 2021



■ Diverse research staffs:

- 140 employed researchers (30% international, 20% female)
- 290 visiting researchers
- 60 domestic students
- 140 international interns (total)

■ Extensive collaboration:

- 3 industry collaborative centers
- 40+ industry projects
- 40+ international collaboration partners

NEC
FUJITSU
TOSHIBA

Contents

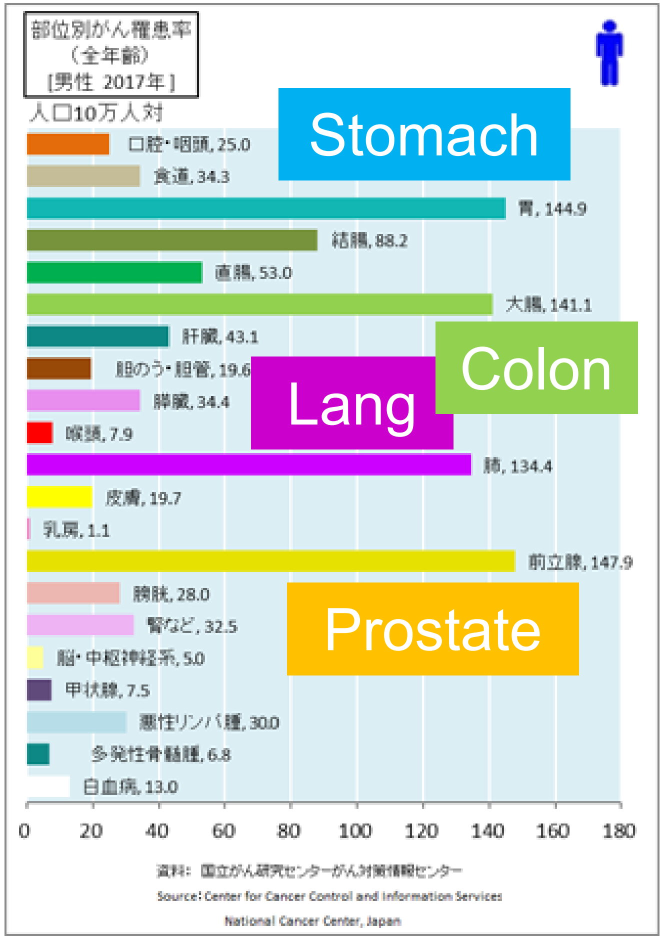
AIP's Research Challenges

- Machine learning (ML) is the core of current AI:
  - Let a computer learn like humans.
  - Successful in speech, image, language, ads,…
- However, current ML is:
  - data-hangry (requiring big labeled data for training),
  - black-box (less interpretable).
- Our challenges:
  - Develop new ML theory to overcome the limitations.
  - Explore new ML application beyond current ML.
  - Design new ML society with appropriate ethical discipline and data-circulation systems.

Contents

1-1) Prostate Cancer Diagnosis

- Prostate cancer accounts for 10% of male cancers:
  - Automatic diagnosis is desired.
- Supervised classification needs annotated pathological images:
  - Increasing doctors’ burden.
- Let’s use unsupervised deep learning for feature extraction.

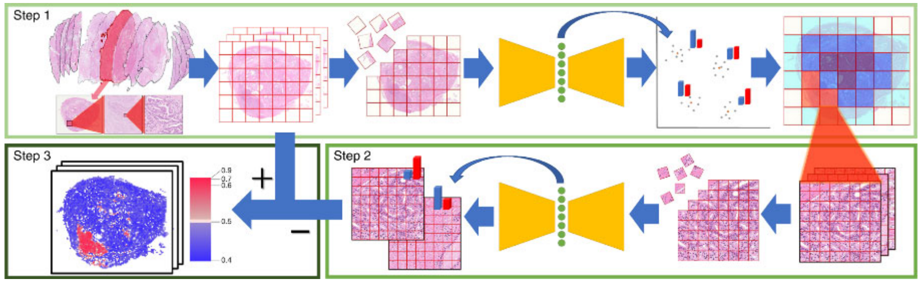


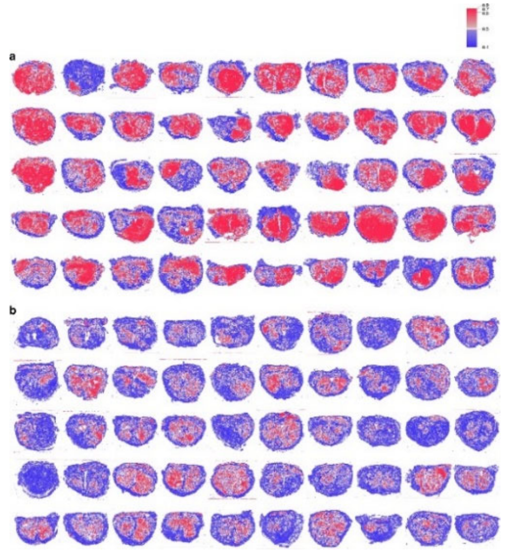
Unsupervised Deep Learning

Yamamoto et al. (Nature Communications 2019)

One of the top 50 most read Nature Communications articles in physics in 2019

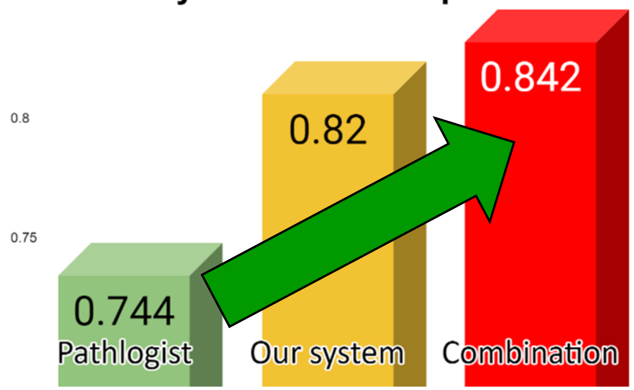
- We used 11+ billion unlabeled pathological image patches for feature extraction.





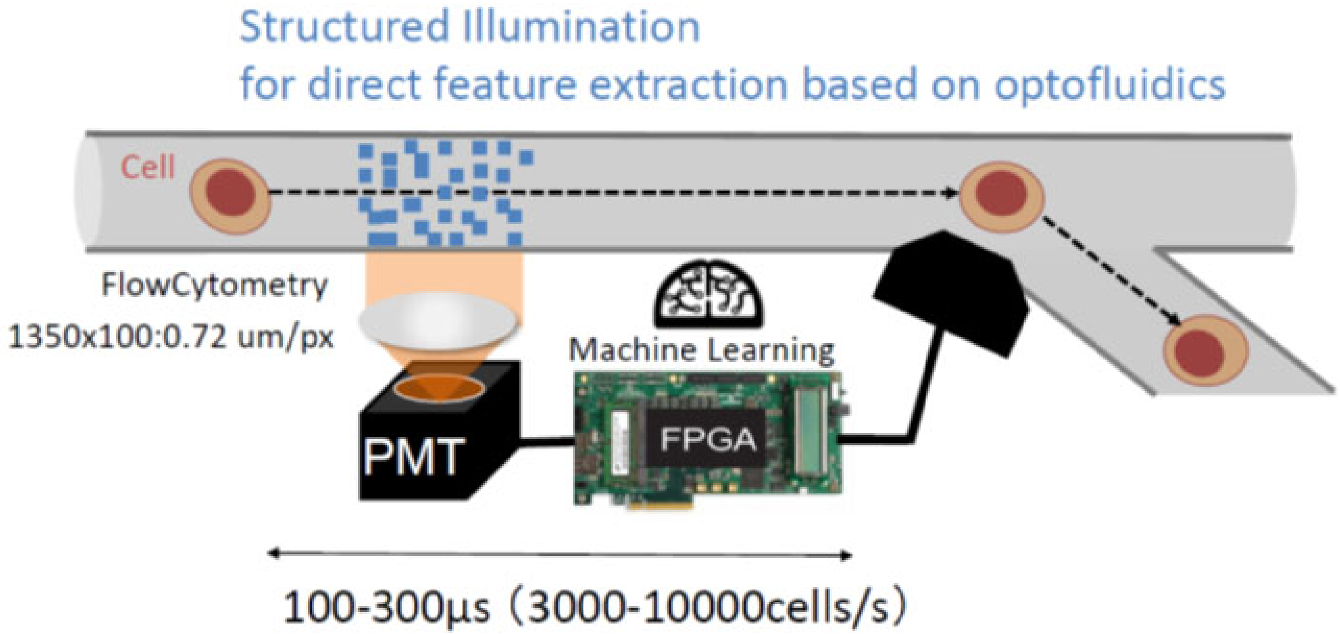
- In addition to the standard Gleason score, novel features such as interstitium change were discovered.
- Further applications in iPS cells, leukemia, and breast cancer.

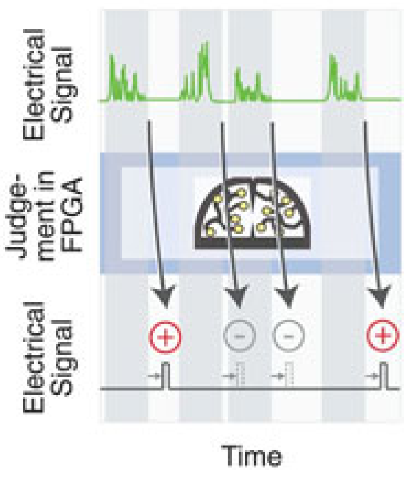
AUC for 1-year recurrence prediction



1-2) Ghost Cytometry

Ota et al. (Science 2018)

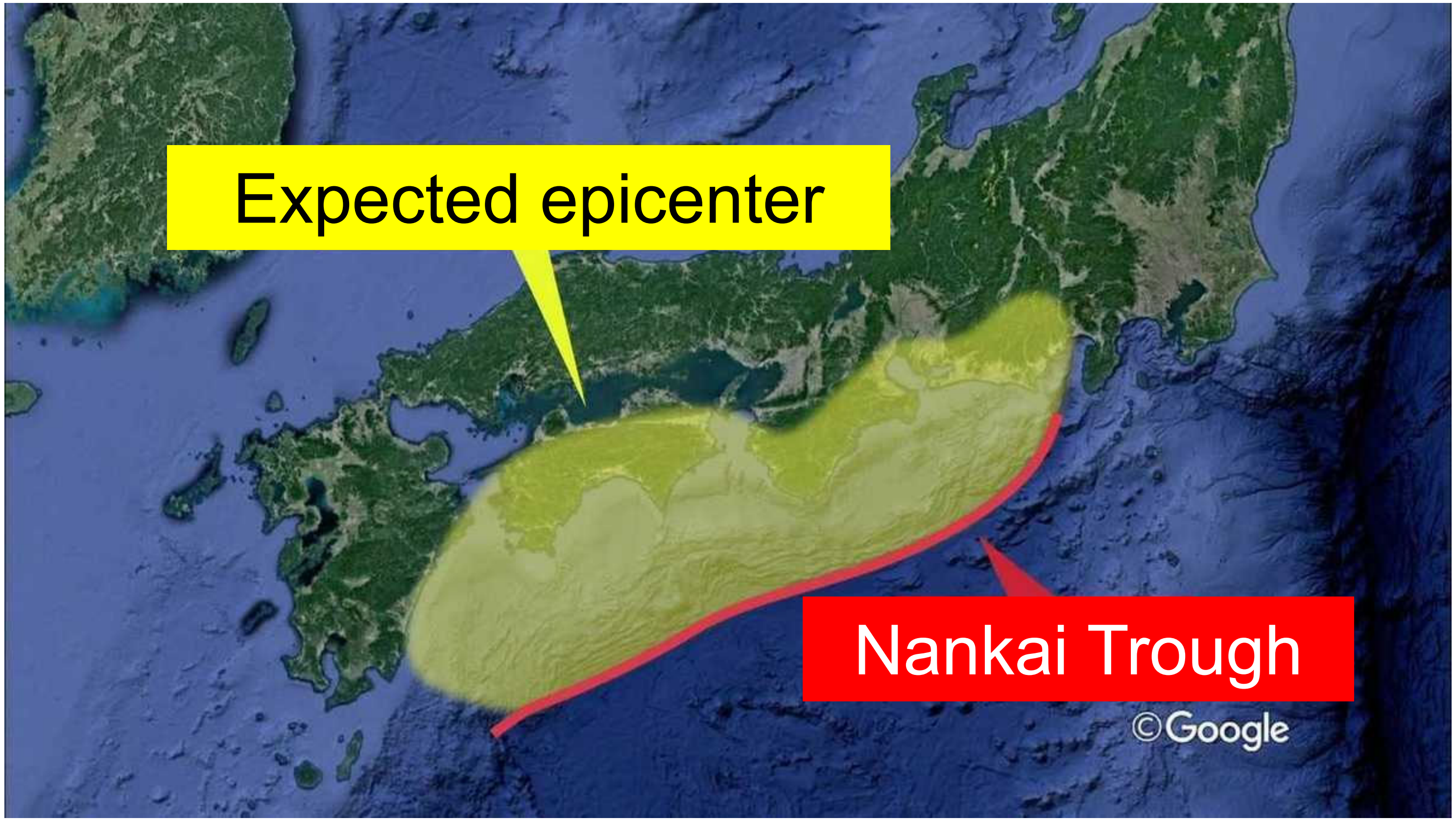




- Classify normal/abnormal cells in the flow:
  - However, deep learning inference is too slow.
- Structured illumination allows direct feature extraction, resulting in real-time classification:
  - Found a start-up for industrialization.
  - Application in tumors and iPS cells.

1-3) Earthquake Cycle Prediction

- Nankai Trough is located south of Japan, expected to cause a big earthquake in the near future:
  - Risk assessment is indispensable.

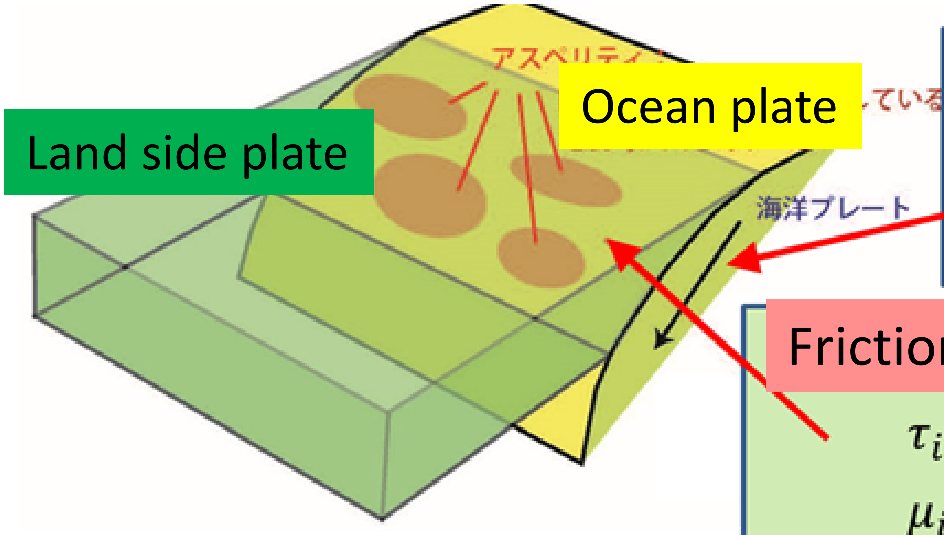


https://www.fnn.jp/articles/-/22389

Mathematical Model of Cycles

- There is a powerful mathematical model:

Equation of motion of ocean plate shear stress & land plate friction force



Shear stress (pushing force)

$$\tau_i(t) = \sum_{j=1}^{N} K_{ij}\left(v_j^{pl}t - u_j(t)\right) - \frac{G}{2\beta}\frac{du_j(t)}{dt}$$

Friction force (stopping force)

○: Friction parameter

$$\tau_i(t) = \mu_i(t)\sigma_i^{eff}$$

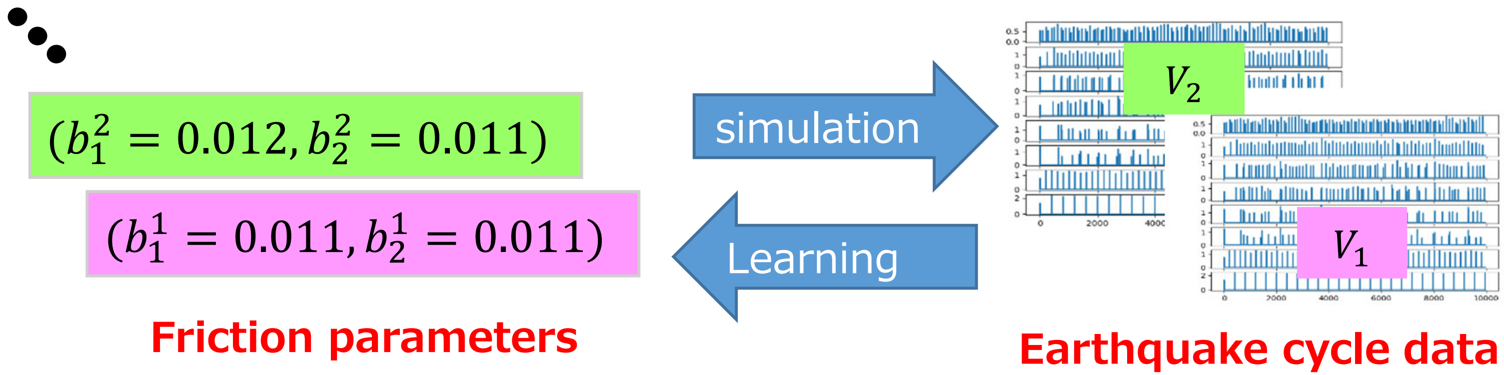
$$\mu_i(t) = \mu_* + a_i \ln\left(\frac{V_i(t)}{V_*}\right) + b_i \ln\left(\frac{V_*\theta_i(t)}{L_i}\right)$$

- Tuning of friction parameters is the key.
  - However, there are no enough supervised data.

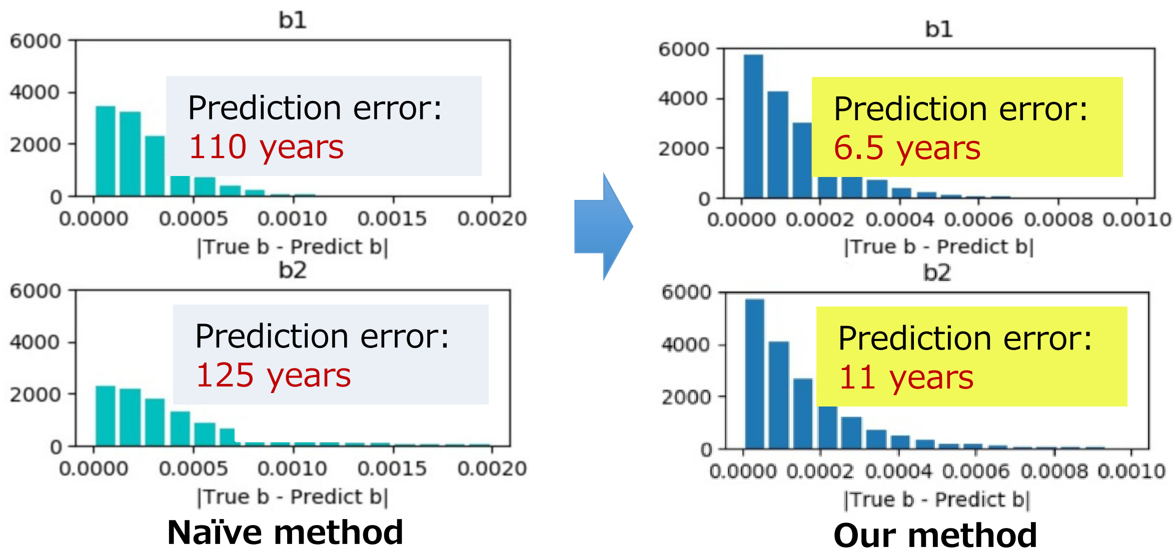
Simulation-based Machine Learning

Hachiya et al. (EGU2019)

- Alternately perform
  - Simulation: Generating artificial data by induction.
  - Learning: Training a model with artificial data.



- Prediction of earthquake cycles is highly improved.



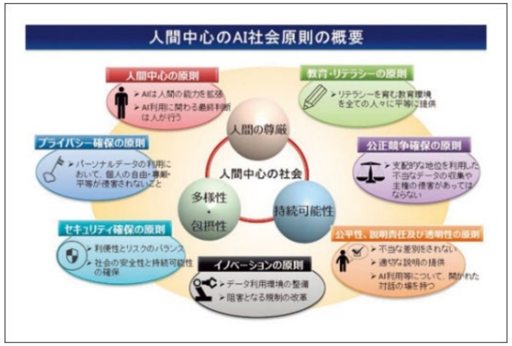
Contents

1. Introduction
2. Research at RIKEN-AIP
   - A) ML Application
   - B) ML Society
   - C) ML Theory
3. Research at IIL Team
4. Future Challenges

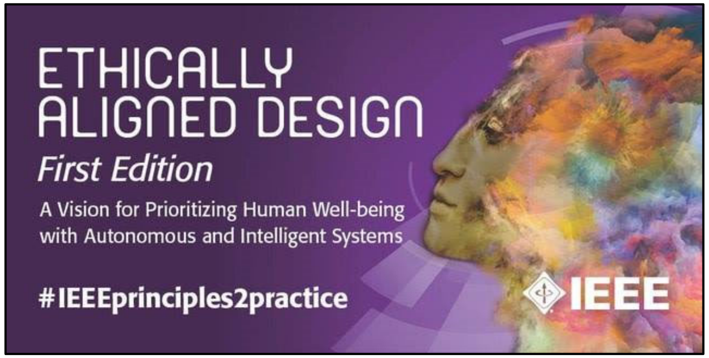
2-1) AI Ethical Guidelines

- We have contributed to the discussions on privacy, fairness, security, etc.:
  - Japanese Society for AI:
    - Ethical Guidelines (2017).

  - Ministry of Internal Affairs and Communications:
    - AI R&D Guidelines, proposed to OECD (2017),
    - AI Utilization Guidelines (2019).

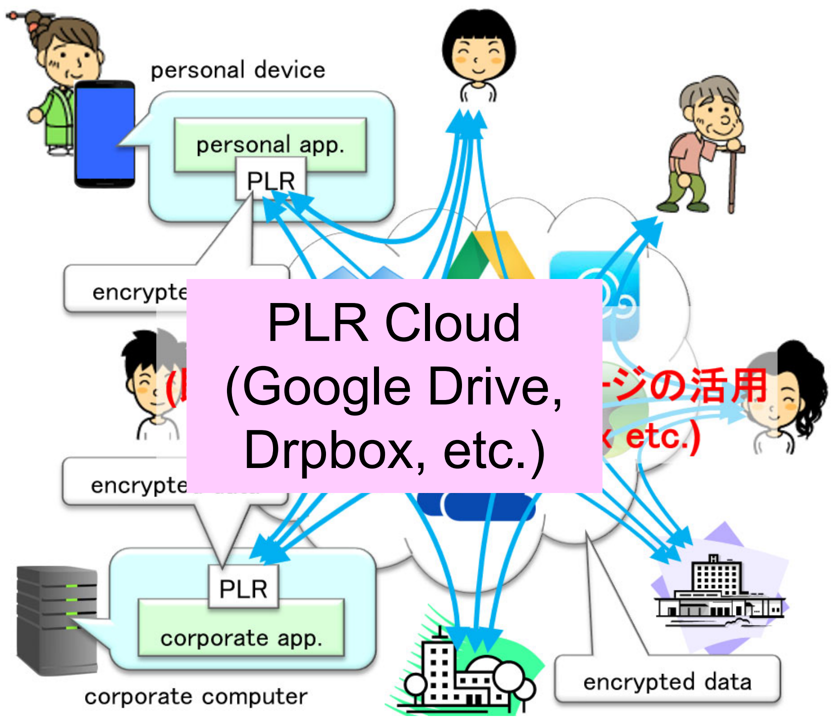


  - Cabinet Office:
    - Social Principles of Human-centric AI, proposed to G20 (2019).
  - IEEE:
    - Ethically Aligned Design (2019).



2-2) Personal Life Repository

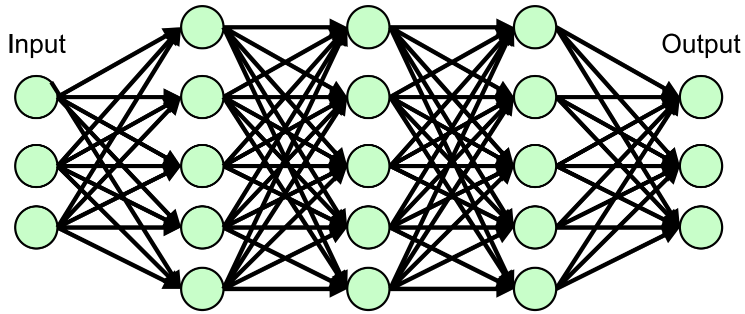
- How should personal information be managed?
  - Company-based or government-based?
- We propose an individual-based system:
  - Data subjects control data accessibility,
  - Low-cost deployment.
- Proof-of-concept:
  - Thousands of high schoolers share their learning records with the school management system.



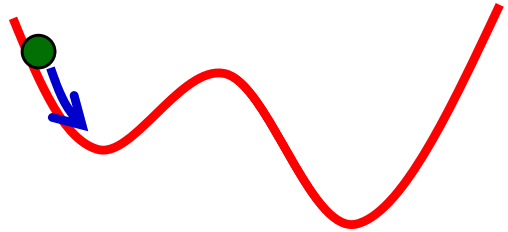
Contents

3-1) Understanding Deep Learning

- **Deep learning:**
  - Stacking many layers.
  - Hard to optimization.
  - Works excellently in practice.



- **We proved its superiority mathematically:**



  - Global optimization is possible.
    Suzuki & Akiyama (ICLR2021)

$$\widehat{L}(X_k) - \int \widehat{L}(x)\mathrm{d}\pi_\infty(x) \lesssim \exp\left(-\Lambda^*_\eta k\eta\right) + \frac{c_\beta}{\Lambda^*_0}\eta^{1/2-\kappa}$$

$$\mathrm{d}X_t = -\nabla\left(\widehat{L}(X_t) + \frac{\lambda}{2}\|X_t\|^2_{\mathcal{H}_K}\right)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}\xi_t$$

  - Better prediction for high-dimensional data.
    Suzuki (NeurIPS2020), Nitanda & Suzuki (ICLR2021), Suzuki & Nitanda (NeurIPS2021)

$$R_{\mathrm{lin}}(\mathcal{F}_\gamma) \gtrsim n^{-\frac{2\tilde{\beta}+d}{2\tilde{\beta}+2d}-\kappa'}$$

$$R_{\mathrm{lin}}(\mathcal{F}_\gamma) := \inf_{\widehat{f}:\text{linear}} \sup_{f^\circ\in\mathcal{F}_\gamma} \mathbb{E}_{D_n}[\|\widehat{f} - f^\circ\|^2_{L_2(P_X)}]$$

  - Universal approximator (INN).
    Teshima et al. (NeurIPS2020)

$$\mathbb{E}_{D^n}\left[\mathbb{E}_{W_k}[\|f_{W_k} - f^\circ\|^2_{L_2(P_X)}|D_n]\right] \lesssim n^{-\frac{\gamma}{\alpha_1-3\alpha_2+1}} + \Xi_k$$

3-2) Causal Inference

Correlation vs. Causality:

- The number of Nobel prize winners can be predicted by Chocolate consumption.
- But, eating more chocolate does not increase the number of Nobel prize winners.

Randomized controlled trial:

- Split the subjects into two group.
- Treat only one and see what happens.
- Ethically problematic (e.g., vaccines)

Messerli (2012)

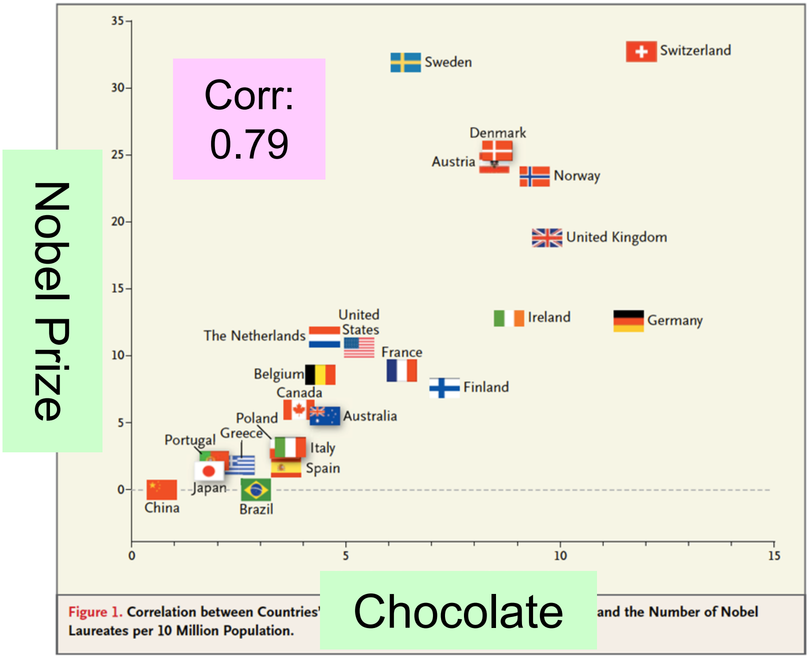
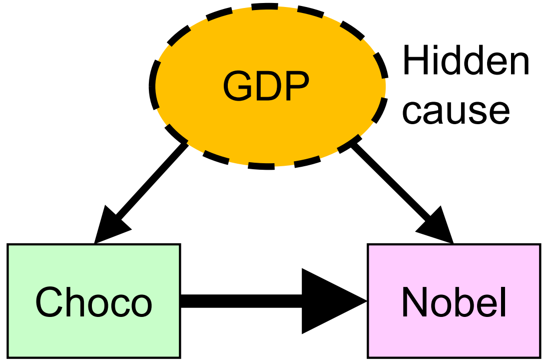


Figure 1. Correlation between Countries' [Chocolate] and the Number of Nobel Laureates per 10 Million Population.

Causal Inference in the Presence of Hidden Cause

- In causal inference, how to handle hidden cause is a big challenge!



- We developed the first method to estimate the entire structure in the presence of hidden cause:
  - Speech separation technique is employed to separate hidden cause.

Maeda & Shimizu (AISTATS2020, UAI2021)

Contents

Imperfect Information Learning Team

- **Members**:
  - Gang Niu (Research Scientist): Learning theory
  - Shuo Chen (Postdoc): Metric learning
  - Jingfeng Zhang (Postdoc): Adversarial learning
  - Jiaqi Lyu (Postdoc): Weakly supervised learning
  - Many great Visiting Scientists,
    Junior Research Associates, Part-Timers,
    and Interns over the world!
- Prof. Bo Han (HKBU) was an intern/postdoc:
  - Now the most important Visiting Scientist in our team!





Research at IIL Team

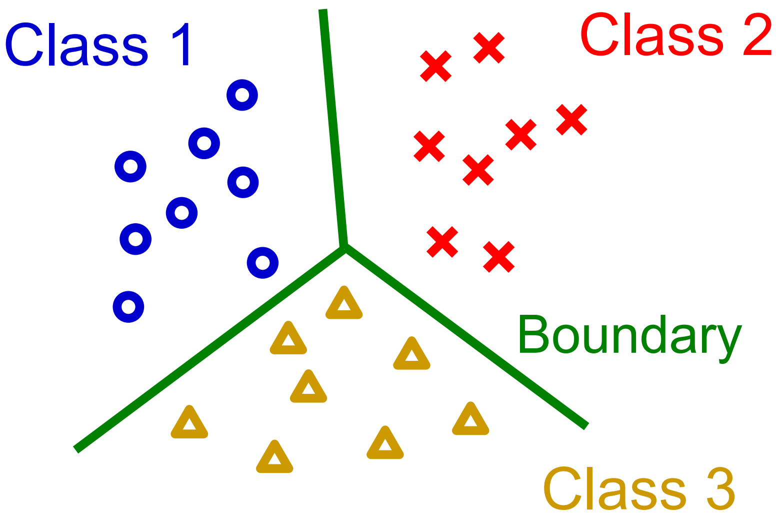
- Goal: Develop novel ML theories and algorithms that enable reliable learning from limited information.
  - Label noise: human error, sensor error.
  - Insufficient information: weak supervision.
  - Data bias: changing environments, privacy.
  - Attack: adversarial noise, distribution shift.

Contents

1. Introduction
2. Research at RIKEN-AIP
3. Research at IIL Team
    - A) Noisy-Label Learning
    - B) Weakly Supervised Learning
    - C) Transfer Learning
4. Future Challenges

Supervised Classification

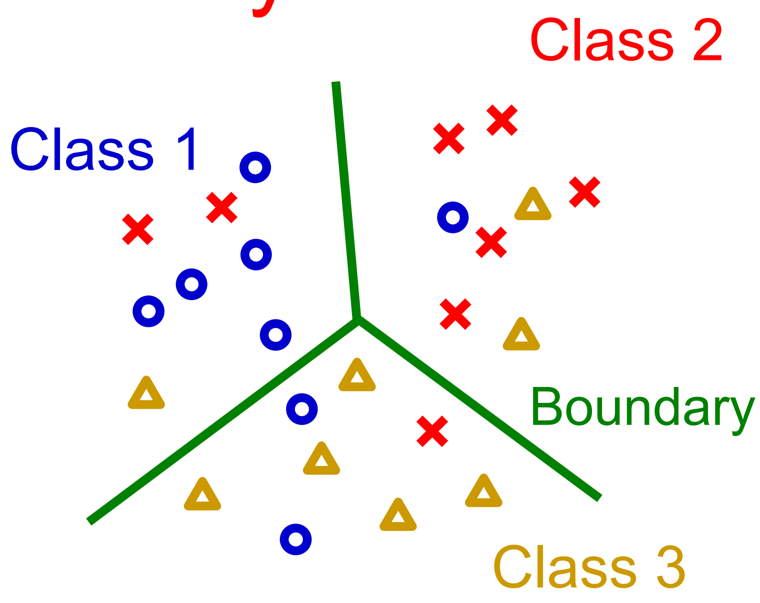
- Supervised classification with clean labels:



Training error minimization is statistically consistent and work well in practice.

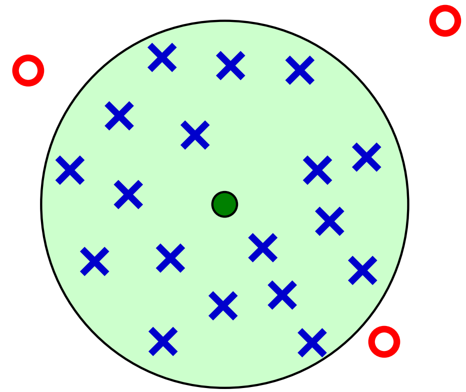
- However, real-world labels are noisy possibly due to human error:

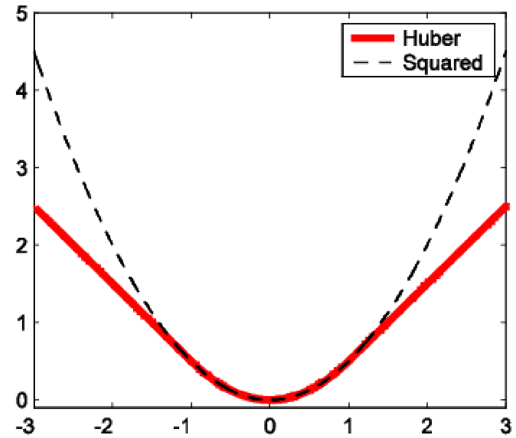
Training error minimization is no longer consistent and does not work well in practice.

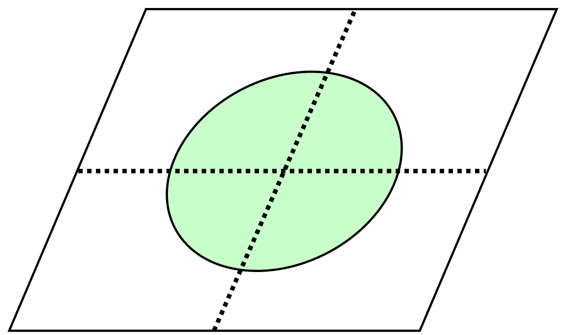


Generic Approaches

- **Unsupervised outlier removal**:
  - Substantially difficult
- **Robust loss, regularization**:
  - Not robust enough
- We want to go beyond the limitations of existing approaches!
  - Noise transition correction
  - Clean sample selection







Noise Transition Correction

- Noise transition matrix $\boldsymbol{T}$:
  - Clean-to-noisy flipping probability.

$\boldsymbol{T}^\top =$

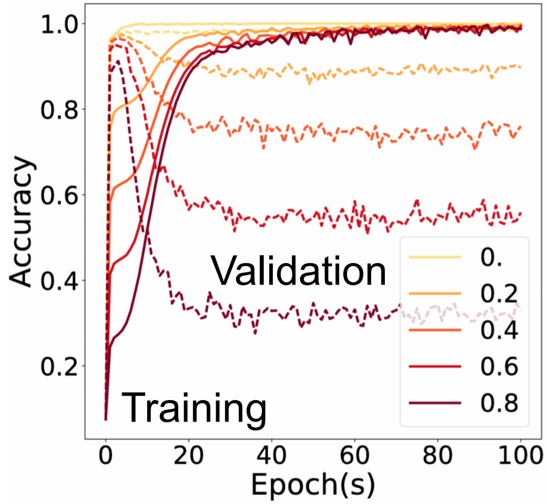
1	0.1	0.5
0	0.8	0.5
0	0.1	0

- Major approaches: Patrini et al. (CVPR2017)
  - Loss correction by $\boldsymbol{T}^{-1}$ to eliminate noise.
  - Classifier adjustment by $\boldsymbol{T}^\top$ to simulate noise.
- We want to estimate $\boldsymbol{T}$ only from noisy data:
  - Use human cognition as a “mask” for $\boldsymbol{T}$. Han et al. (NeurIPS2018)
  - Learn $\boldsymbol{T}$ and a classifier dynamically. Xia et al. (NeurIPS2019)
  - Decompose $\boldsymbol{T}$ into simpler components. Yao et al. (NeurIPS2020)
  - Regularize $\boldsymbol{T}$ to be estimable. Zhang et al. (ICML2021), Li et al. (ICML2021)
  - Extension to input-dependent noise $\boldsymbol{T}(\boldsymbol{x})$.

Xia et al. (NeurIPS2020), Berthon et al. (ICML2021)

Clean Sample Selection

Arpit et al. (ICML2017)

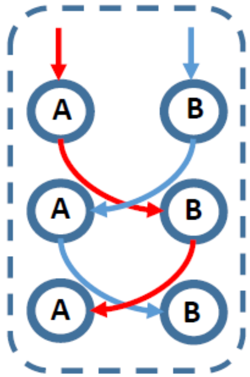


Memorization of neural nets:

- Stochastic gradient descent fits clean data faster, but naïve early stopping does not work well.

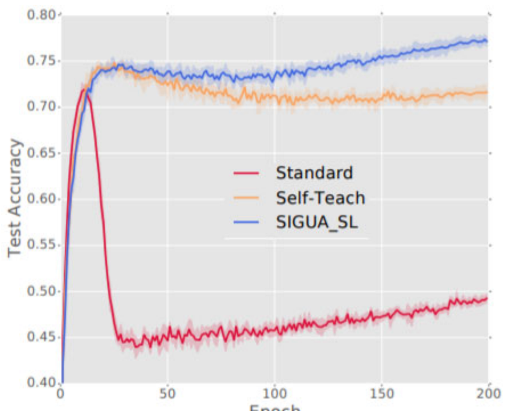
"Co-teaching" with two neural nets:

- Teach small-loss data each other. Han et al. (NeurIPS2018)
- Teach only disagreed data. Yu et al. (ICML2019)
- Gradient ascent for large-loss data. Han et al. (ICML2020)



Very robust in experiments:

- Works well even if 50% of labels are randomly flipped.

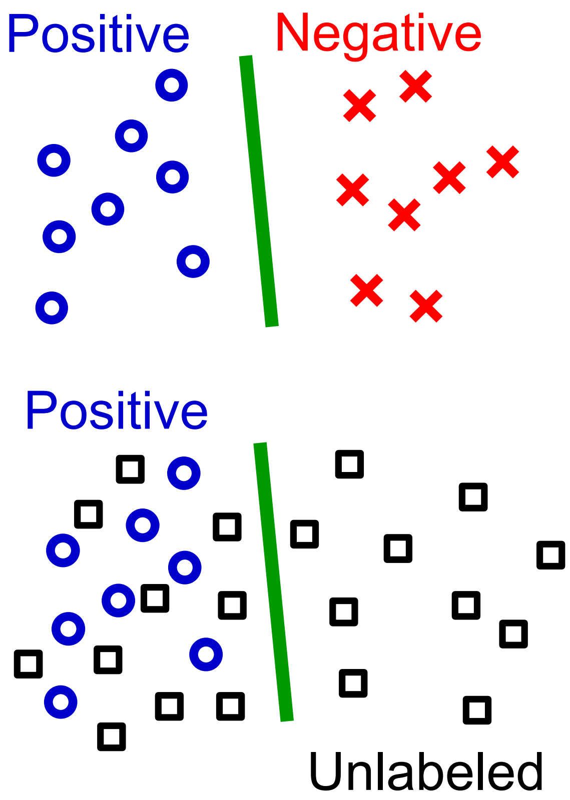


Contents

1. Introduction
2. Research at RIKEN-AIP
3. Research at IIL Team
   - A) Noisy-Label Learning
   - B) Weakly Supervised Learning
   - C) Transfer Learning
4. Future Challenges

Weakly Supervised Learning

- Fully supervised data is expensive to collect.
- Weakly supervised data can be collected easily:
  - Ex.) Click prediction in online ads: It is easy to automatically collect
    - Clicked ads (positive),
    - Unclicked ads (unlabeled).



- Learning only from P and U data is possible!

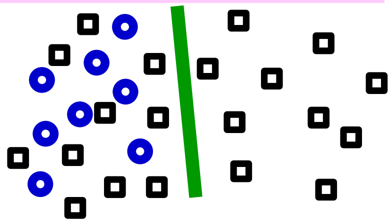
du Plessis et al. (NIPS2014, ICML2015, MLJ2017), Niu et al. (NIPS2016), Kiryo et al. (NIPS2017), Hsieh et al. (ICML2019)

  - Regard U data as noisy N data and correct the loss.
  - Statistically consistent. $\mathcal{O}_p\left(1/\sqrt{n}\right)$

Various Extensions

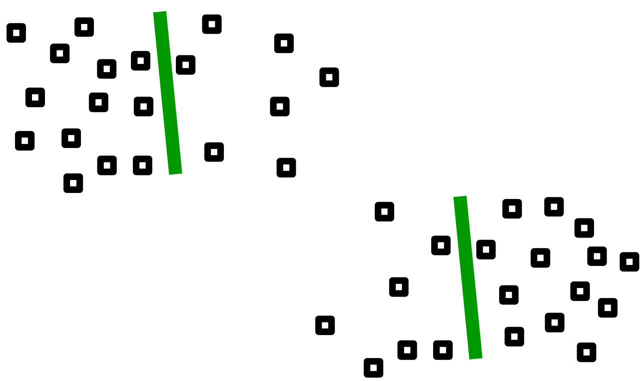
- Learning from weakly supervised data is possible in many different forms!

Positive-Unlabeled



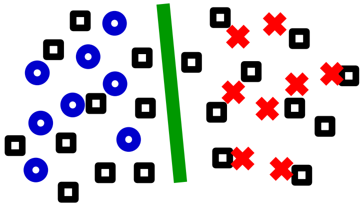
du Plessis et al. (NIPS2014, ICML2015, MLJ2017)
Niu et al. (NIPS2016),, Kiryo et al. (NIPS2017)
Hsieh et al. (ICML2019)

Unlabeled-Unlabeled



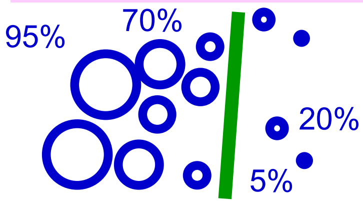
du Plessis et al.,(TAAI2013)
Lu et al. (ICLR2019, AISTATS2020)
Charoenphakdee et al. (ICML2019)
Lei et al. (ICML2021)

Semi-Supervised



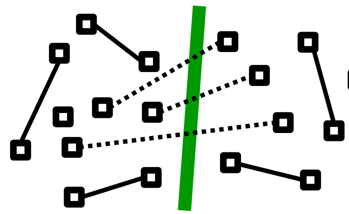
Sakai et al. (ICML2017, ML2018)

Positive-confidence



Ishida et al. (NeurIPS2018)
Shinoda et al. (IJCAI2021)

Similar-Dissimilar

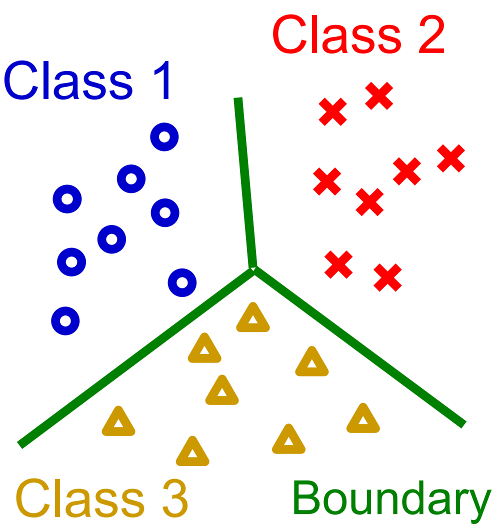


Bao et al. (ICML2018)
Shimada et al. (NeCo2021)
Dan et al. (ECMLPKDD2021)
Cao et al. (ICML2021)
Feng et al. (ICML2021)

- All are loss-correction based and consistent. $\mathcal{O}_p\left(1/\sqrt{n}\right)$
- Any loss, classifier, and optimizer can be used.

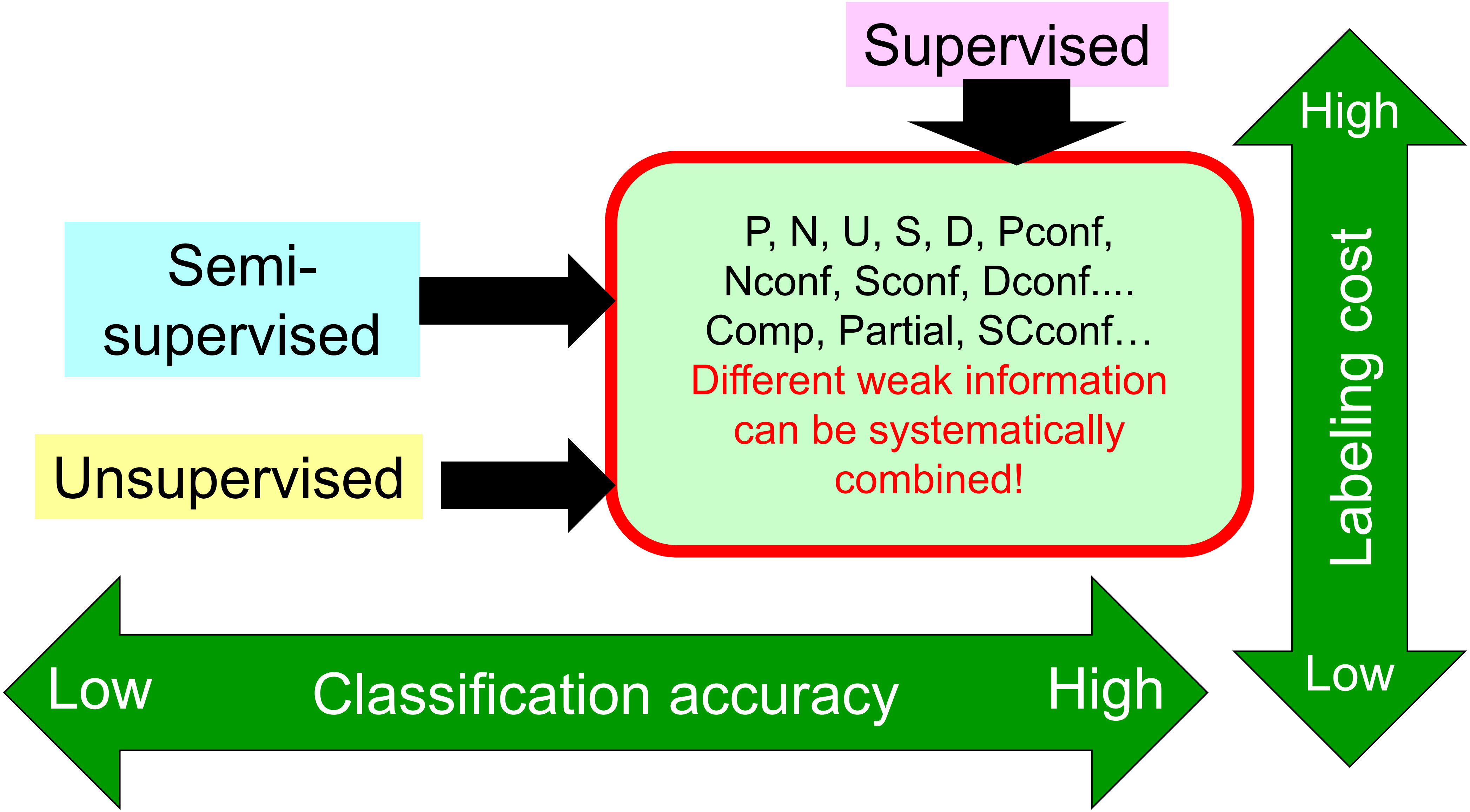
Multiclass Methods

- Labeling patterns in multi-class problems is extremely painful.



- Multi-class weak-labels:
  - Complementary labels: Ishida et al. (NIPS2017, ICML2019) Chou et al. (ICML2020)
    Specify a class that a pattern does not belong to (“not 1”).
  - Partial labels: Specify a subset of classes that contains the correct one (“1 or 2”). Feng et al. (ICML2020, NeurIPS2020) Lv et al. (ICML2020)
  - Single-class confidence: Cao et al. (arXiv2021)
    One-class data with full confidence (“1 with 60%, 2 with 30%, and 3 with 10%”)
- Systematic loss correction is possible! $\mathcal{O}_p\left(1/\sqrt{n}\right)$

Weakly Supervised Learning

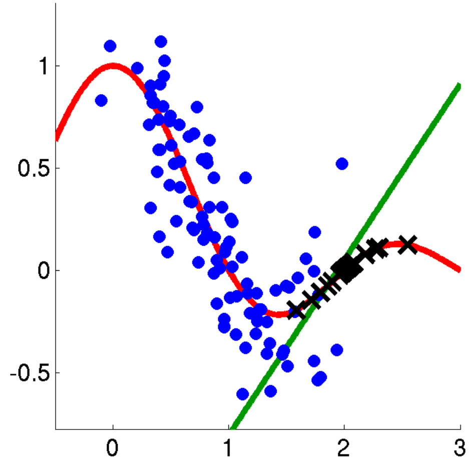


Sugiyama, Bao, Ishida, Lu, Sakai & Niu,
Machine Learning from Weak Supervision
MIT Press, 2022.

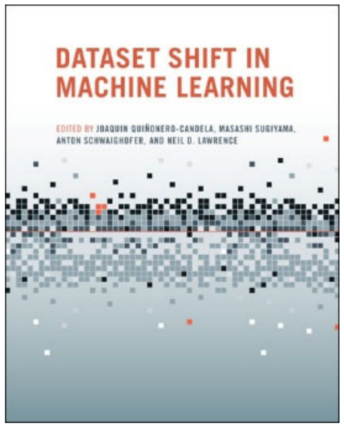
Contents

Transfer Learning

- Training and test data often have different distributions, due to
  - changing environments,
  - sample selection bias (privacy).

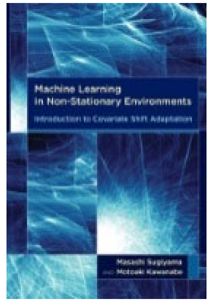


- Transfer learning (domain adaptation):
  - Train a test-domain predictor using training data from different domains.



Quiñonero-Candela et al. (MIT Press 2009)

Classical Approach for Transfer Learning



Sugiyama & Kawanabe
(MIT Press 2012)

- Two-step adaptation:
  1. Importance weight estimation:

$$\widehat{w} = \underset{w}{\mathrm{argmin}}\, \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)} \left[D\left(w(\boldsymbol{x},y), \frac{p_{\mathrm{te}}(\boldsymbol{x},y)}{p_{\mathrm{tr}}(\boldsymbol{x},y)} \right) \right]$$

  2. Weighted predictor training:

$$\widehat{f} = \underset{f}{\mathrm{argmin}}\, \widehat{\mathbb{E}}_{p_{\mathrm{tr}}(\boldsymbol{x},y)} [\widehat{w}(\boldsymbol{x},y)\ell(f(\boldsymbol{x}),y)]$$

- However, estimation error in Step 1 is not taken into account in Step 2.
  - We want to integrate these two steps!

Joint Weight-Predictor Optimization

- **Covariate shift:** Only input distributions change.

$$p_{\mathrm{tr}}(\boldsymbol{x}) \neq p_{\mathrm{te}}(\boldsymbol{x}) \qquad p_{\mathrm{tr}}(y|\boldsymbol{x}) = p_{\mathrm{te}}(y|\boldsymbol{x})$$

Shimodaira (JSPI2000)

- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Unlabeled test data: $\{\boldsymbol{x}_i^{\mathrm{te}}\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x})$
- Minimize a risk upper bound jointly w.r.t. weight w and predictor f: $J_{\ell_{\mathrm{tr}}}(f, w) \geq R_{\ell_{\mathrm{te}}}(f)^2$

Zhang et al. (ACML2020, SNCS2021)

$$\widehat{f} = \underset{f}{\mathrm{argmin}} \min_{w \geq 0} \widehat{J}_{\ell_{\mathrm{tr}}}(f, w)$$

$\widehat{J}_\ell$: Empirical approximation of J_ℓ

$$R_\ell(f) = \mathbb{E}_{p_{\mathrm{te}}(\boldsymbol{x},y)}[\ell(f(\boldsymbol{x}), y)]$$

$$\ell_{\mathrm{te}} \leq 1, \ell_{\mathrm{tr}} \geq \ell_{\mathrm{te}}$$

  - Theoretical guarantee:

$$R_{\ell_{\mathrm{te}}}(\widehat{f}) \leq \sqrt{2} \min_f R_{\ell_{\mathrm{te}}}(f) + \mathcal{O}_p(n_{\mathrm{tr}}^{-1/4} + n_{\mathrm{te}}^{-1/4})$$

Dynamic Importance Weighting

- General changing distributions: $p_{\mathrm{tr}}(\boldsymbol{x}, y) \neq p_{\mathrm{te}}(\boldsymbol{x}, y)$
- Suppose we are given
  - Labeled training data: $\{(\boldsymbol{x}_i^{\mathrm{tr}}, y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{tr}}(\boldsymbol{x}, y)$
  - Labeled test data: $\{(\boldsymbol{x}_i^{\mathrm{te}}, y_i^{\mathrm{te}})\}_{i=1}^{n_{\mathrm{te}}} \overset{\mathrm{i.i.d.}}{\sim} p_{\mathrm{te}}(\boldsymbol{x}, y)$
- For each mini-batch $\{(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}, \bar{y}_i^{\mathrm{tr}})\}_{i=1}^{\bar{n}_{\mathrm{tr}}}, \{(\bar{\boldsymbol{x}}_i^{\mathrm{te}}, \bar{y}_i^{\mathrm{te}})\}_{i=1}^{\bar{n}_{\mathrm{te}}}$, importance weights are estimated by matching losses by kernel mean matching:

Fang et al. (NeurIPS2020)

Huang et al. (NeurIPS2007)

$$\frac{1}{\bar{n}_{\mathrm{tr}}} \sum_{i=1}^{\bar{n}_{\mathrm{tr}}} r_i \ell(f(\bar{\boldsymbol{x}}_i^{\mathrm{tr}}), \bar{y}_i^{\mathrm{tr}}) \approx \frac{1}{\bar{n}_{\mathrm{te}}} \sum_{j=1}^{\bar{n}_{\mathrm{te}}} \ell(f(\bar{\boldsymbol{x}}_j^{\mathrm{te}}), \bar{y}_j^{\mathrm{te}})$$

- Extremely simple, but highly powerful!

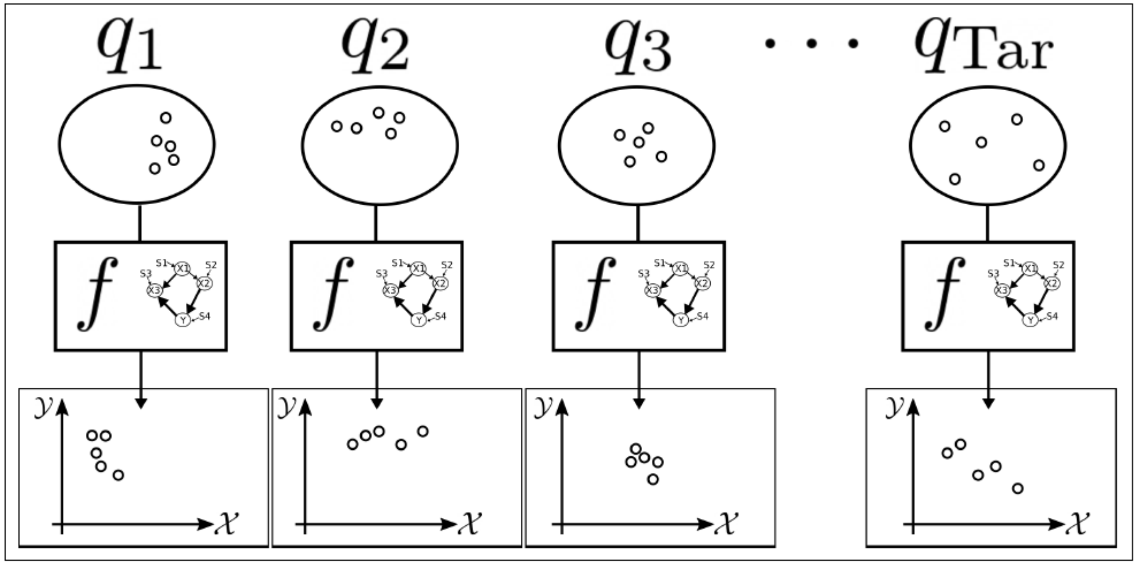
Mechanism Transfer

- Is transfer learning possible when data distributions are seemingly very different?
- Yes, if data generation mechanisms are shared:
  - Use invertible neural networks (INNs) to invert the data generation mechanism. Teshima et al. (ICML2020)
  - INNs are universal approximators. Teshima et al. (NeurIPS2020)

Independent components

"Mechanism"

Observed data



Contents

1. Introduction
2. Research at RIKEN-AIP
3. Research at IIL Team
4. Future Challenges

Challenges in Reliable ML

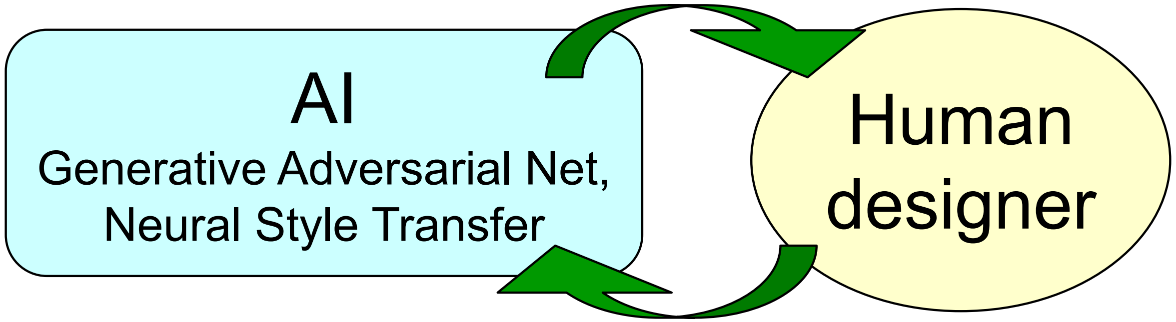
- Reliability for expectable situations:
  - Model the corruption process explicitly and correct the solution.
    - How to handle modeling error?
- Reliability for unexpected situations:
  - Consider worst-case robustness (“min-max”).
    - How to make it less conservative?
  - Include human support (“rejection”).
    - How to handle real-time applications?
- Exploring somewhere in the middle would be practically more useful:
  - Use partial knowledge of the corruption process.

History of AI and Future

- **Classic AI:**
  - 1960s: symbolic, logical AI
  - 1980s: Expert systems
- **Neuro-inspired AI:**
  - 1960s: 1-layer perceptrons
  - 1980s: Multilayer perceptrons
- **Statistical machine learning:**
  - 2000s: Statistics, Bayes, convex optimization, kernels
- **Deep learning:**
  - 2010s: Stochastic gradient, gigantic deep models
- **Next-generation AI:**
  - Integration of elements
  - Human-like AI?

Next-Generation AI

- Is Human-like AI ultimate?
  - Future AI needs not be autonomous.
  - Future AI may learn together with humans.



- AI needs to be inclusive to human society:
  - Technology
    X
    Human creativity, culture, and ethics.

Fashion show at UTokyo in Mar. 2019
(with Prof. Aihara and Emarie)

https://www.fashion-press.net/collections/11006

Thank You!

多謝！

ありがとう！