75th 1944-2019 The Institute of Statistical Mathematics

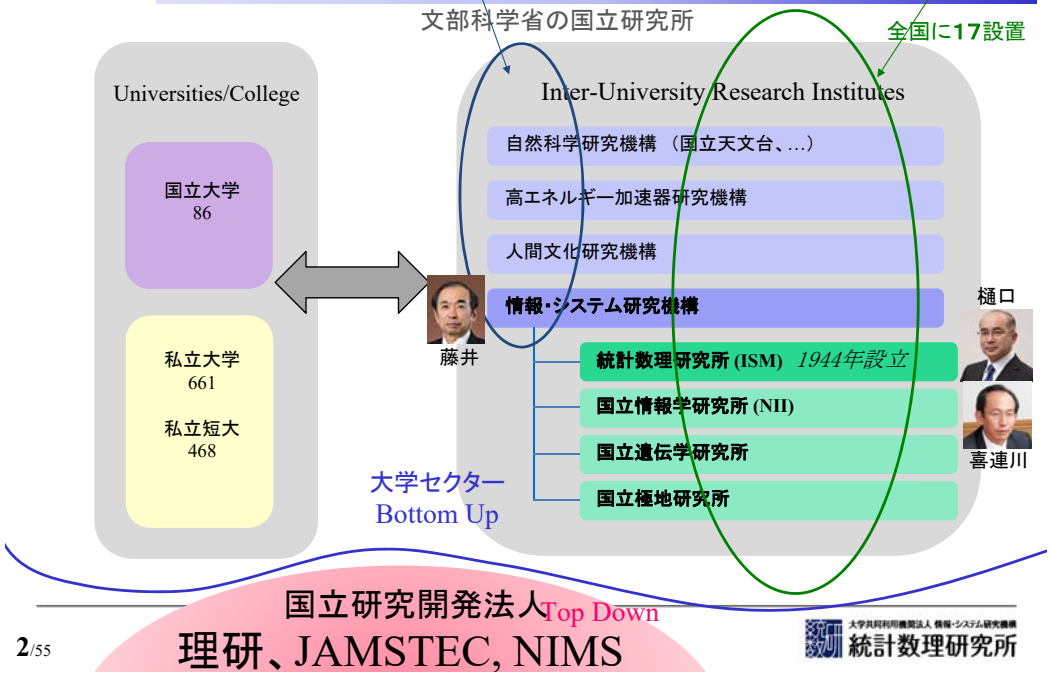
人工知能とデータサイエンティストの役回り

樋口知之 （情報・システム研究機構 統計数理研究所）

一般社団法人 データサイエンティスト協会 顧問

統計数理研究所

大学共同利用機関法人と大学共同利用機関



統計数理研究所

アウトライン

1. 平成の30年間
2. 統計学、機械学習、深層学習
3. 帰納法の弱点
4. 人材育成

統計数理研究所

アウトライン

1. 平成の30年間
2. 統計学、機械学習、深層学習
3. 帰納法の弱点
4. 人材育成

統計数理研究所

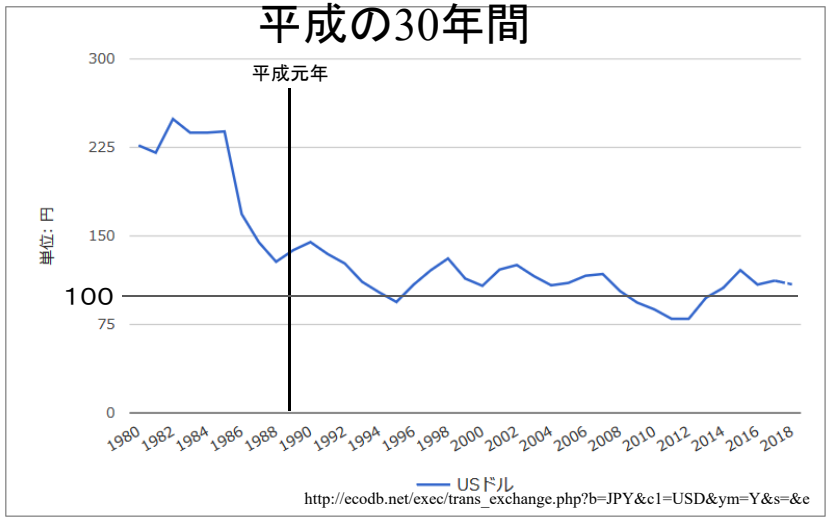
２０年、３０年、そして５０年

平成の30年間

日経平均の30年間の推移

（図のソース： 2017/12/29日経Webニュース）

円ドルレートと時価総額



Manufacturing companies to IT companies

Top of market capitalization of stock price

Compared with 10 years ago

	2007, May		2017,May		
469	Exxon Mobil	1		796	×7.6
	GE	2			×4.3
	Microsoft	3	Microsoft		×1.8
	Citi Group	4	amazon		×16.8
	Petro China	5	facebook		----
	AT&T	6	Berkshire Hathaway		
	Royal Dutch Shell	7	Johnson & Johnson		
	Bank of America	8	Exxon Mobil		×0.7
	Industrial and Commercial Bank of China	9	Tencent 腾讯		×42.5
216		10	阿里巴巴	297	----

billion US$ billion US$

Original News Source：NikkeiWeb2017/June/02

Modern Seven Sisters

Microsoft amazon facebook Tencent 腾讯 阿里巴巴

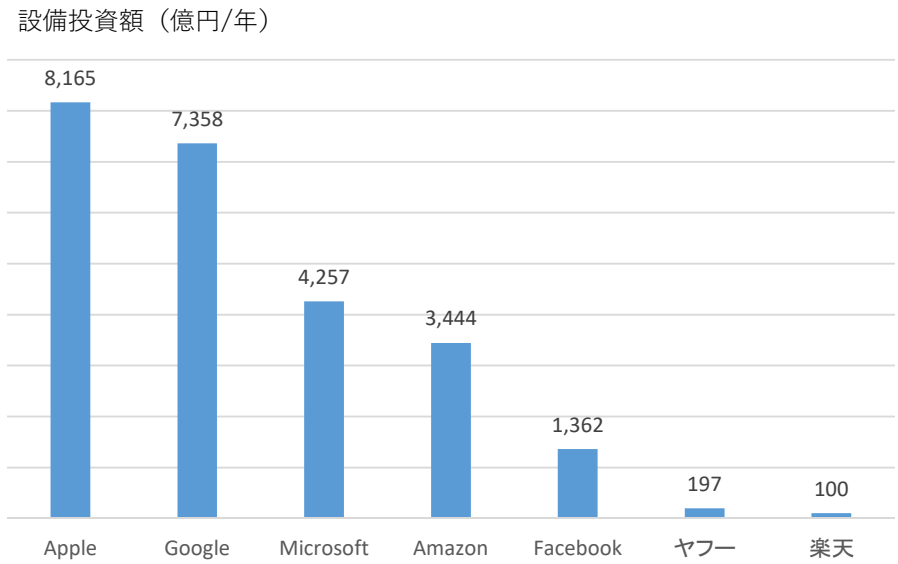
http://blog.livedoor.jp/sagittariun/archives/11995435.html

IT主要5大企業の収益源

5大IT企業の収益内容

（ソース　https://www.businessinsider.jp/post-33925
http://www.visualcapitalist.com/chart-5-tech-giants-make-billions/）

投資額の格差（USのAI×データ企業との比較）



各社のFY2013 IR資料：有形固定資産の取得、100円/$で換算

安宅和人「"シン・ニホン":AI×データ時代における日本の再生と人材育成」を参照
(http://www.meti.go.jp/committee/sankoushin/shin_sangyoukouzou/pdf/013_06_00.pdf)

検索＋スマホ世代が変える社会

知識がモノの形で伝達→知識が脳内で構造化

↓

知識がフラット化

価値観の転換
- Sharing
- エコシステム
- Citizen Science

アントレプレナーブーム

人工知能学会の栄枯盛衰

1986年創設

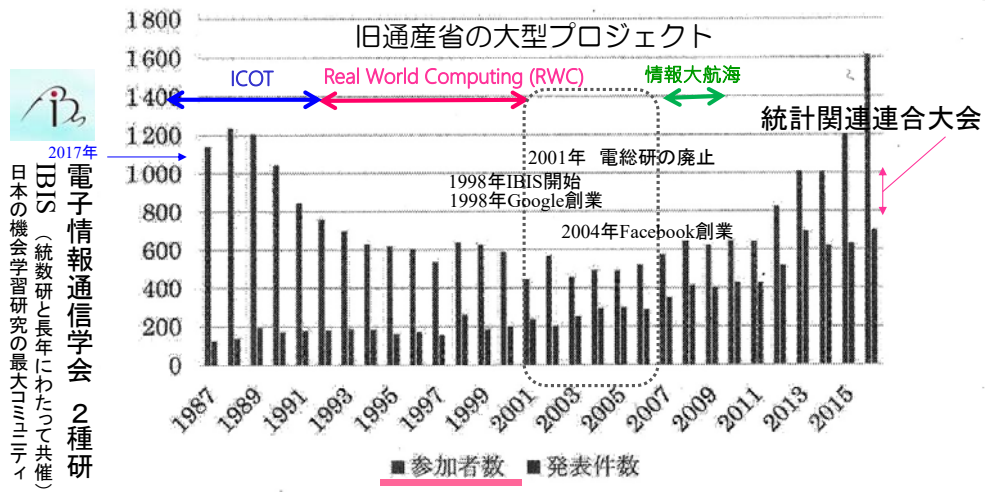
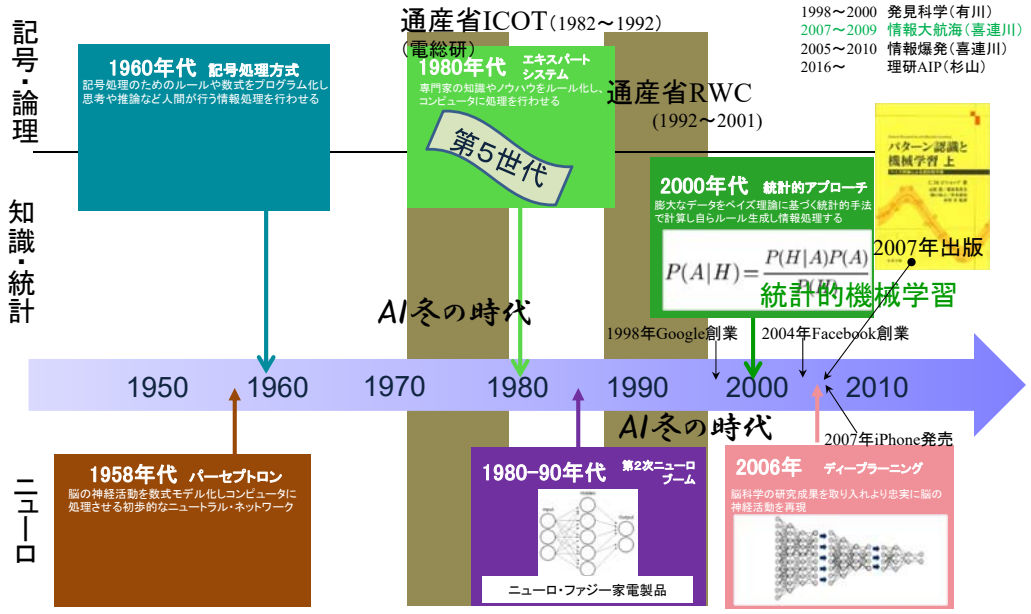


図４　全国大会参加者数と発表件数の推移

2016年10月に、それまで最高であった1992年10月の会員数を超える

人工知能研究および研究開発プロジェクトの歴史

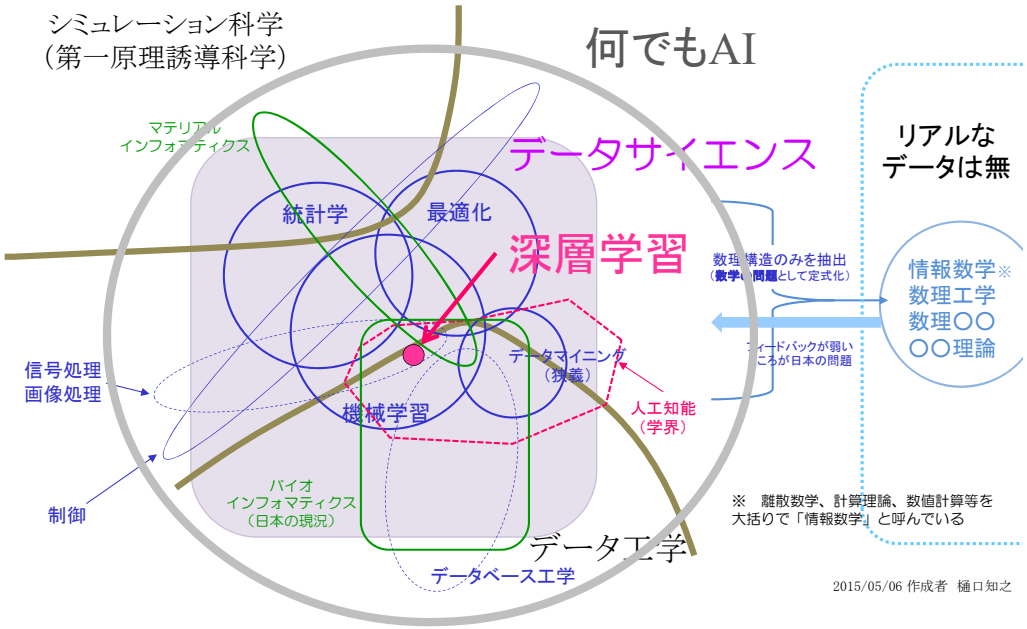


＊http://blogs.itmedia.co.jp/itsolutionjuku/2015/07/post_105.html を参考として作成
＊統数研・福水教授のスライド内容を一部借用

アウトライン

1. 平成の30年間
2. 統計学、機械学習、深層学習
3. 帰納法の弱点
4. 人材育成

データに関連した数理分野の俯瞰図



Prosperity of AI (To be precise, "The Third Neural network Boom")

- In the past seven years, the AI field has developed dramatically
  - IBM Watson (Question Answering System) wins Human Quiz King in 2011.
    No part of deep learning.
    Knowledge base of Big data analysis result + expert system
  - Overwhelming victory with Deep Learning in many contests, such as **images and sounds** competitions from 2011
    a. 2011 Image recognition of traffic sign; (erroneous recognition rate 0.56%, second place 1.16%)
    b. 2012 General object recognition for **ImageNet** (misrecognition rate 15.3%, second place 26.2%)
  - Increased attention to deep learning. IT companies focus on research
    - Google: Google Brain Project, Google Deep mind
    - Facebook: Facebook AI Research
    - Baidu
  - 2018 Google AlphaGO wins the world pro

Modification of Slide by Dr. Fukumizu@ISM

Triplet defining statistical machine learning

DL is one of the technique in statistical Machine Learning

Purpose in the meaning of machine learning

Supervised learning
Classification
Regression

Unsupervised learning
Clustering

DNN $w_{i,j}^{[k+1]}$

Decision Tree | SVM+Gaussian Kernel | Neural Network | Random Forest

Original Source
http://tjo.hatenablog.com/entry/2014/01/06/190456

Data Latent variable Parameter

$$f(|\{y_i\},\{x_j\},\{a_k\})$$

Machine Learning
(Purpose | Model | Algorithm)

Various Input Data

Result of estimation on regression/classification

Higuchi

State Space Model, Gaussian Process, Support Vector Machine, Neural Network DNN

SML is defined by a triple of purpose, model, and learning algorithm. The purpose is what you want from the data. A model is a mathematical expression used for that purpose. DNN is also its special form. A learning algorithm refers to a procedure for determining parameters included in a model from data. If models and algorithms are different, the results are always different even if the purpose is the same.

Modified from an original slide by Prof. T. Matsui

Three main types of Machine Learning methods

- *Supervised Learning*

Classification

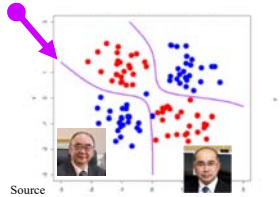
Given Dataset

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_{1000}, y_{1000})\}$$

Label


, Higuchi


, Nishimura



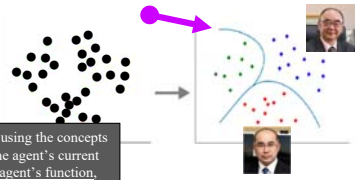
Source
http://tjo.hatenablog.com/entry/2014/01/06/190456

- *Unsupervised Learning*

Clustering

Given Dataset

$$\{x_1, x_2 \ldots, x_{1000}\}$$



Reinforcement learning refers to goal-oriented algorithms. Reinforcement learning can be understood using the concepts of agents, environments, states, actions and rewards. An agent takes actions. The environment takes the agent's current state and action as input, and returns as output the agent's reward and its next state. We can know the agent's function, but we cannot know the function of the environment. It is a black box where we only see the inputs and outputs.

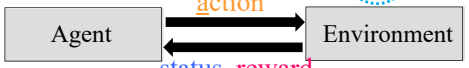
- *Reinforcement Learning*

Extrapolation

$$\{(a_1) \rightarrow (s_2, r_2) \rightarrow (a_2) \rightarrow (s_3, r_3) \rightarrow (a_3) \rightarrow (s_4, r_4) \rightarrow, \ldots \infty\}$$

(action)

(status, reward)



大学共同利用機関法人 情報・システム研究機構 統計数理研究所

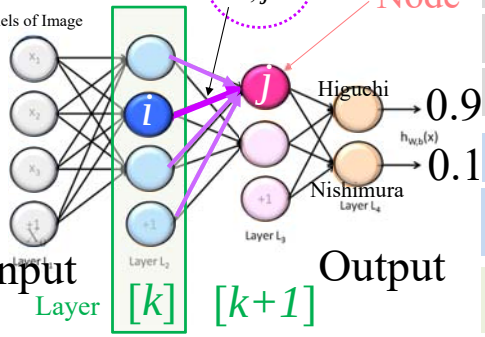
Multilayer Neural Network and Deep NN

The word of Deep Learning (DL) is used like Deep Neural Network (DNN)

$$f(x_j^{[k+1]}) = \frac{1}{1+\exp\left\{\sum_{i=1}^{I([k+1])} w_{i,j}^{[k+1]} x_i^{[k]} + b_j^{[k+1]}\right\}}$$

weight $w_{i,j}^{[k+1]}$

Node



A) DNN has a multilayered structure with the input layer at the far left, the output layer at the far right, and multiple layers at the middle.

B) DNN usually has more than 10 layers.
Ex: There are as many as 1,000 layers.

C) One layer has many terminals called nodes. Nodes are linked to nodes of the neighboring layer each other.

1) Node *i* is coupled with node *j* with a weighting factor *w_ij*.

2) The output of node *j* at the (*k*+1)-th layer is the value obtained by nonlinearly converting the sum value from all the nodes of the next layer ([*k*]-th layer).

Put the value of the pixel of the image into the input layer. In the output layer, the probability of someone's input is output.

大学共同利用機関法人 情報・システム研究機構 統計数理研究所

深層学習の強み：人工知能の「度量衡」

長さを計る基準を「度」，体積は「量」，
重さは「衡」と定め，…

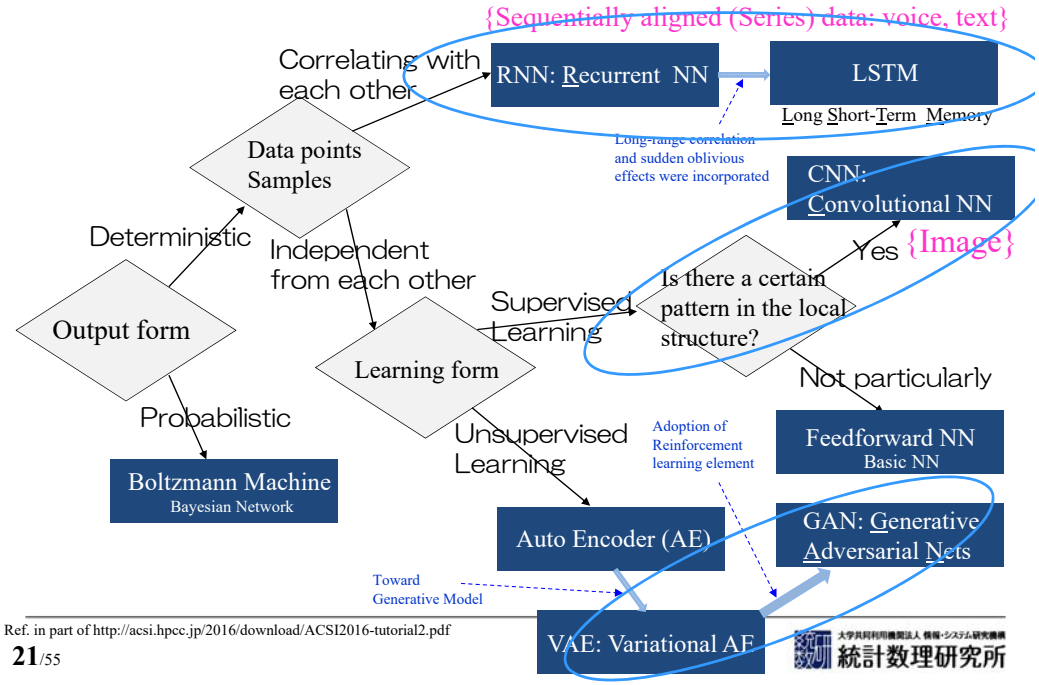
2018年 人工知能学会誌　特集「AIとデータ」に招待論文を寄稿

http://100percent-magazine.net/2015/09/08/759

① パラメータ学習がバックプロパゲーションと確率的勾配降下法で統一
② 計算プラットホームが汎化
  a. 専用計算機の整備（比較的廉価）
    - GPGPUクラスタ
  b. オープン・無償な開発プラットフォームの整備
    - オープンソース： TensorFlow, Caffe, Keras, Chainer
③ オープンかつリアルタイムな成果公開
  ✓ 査読プロセスをとおさず、プレプリサーバ（arXiv）にアップ

理論的には第二次AIブームから大きな発展はこれから……

Classification of D N N



Ref. in part of http://acsi.hpcc.jp/2016/download/ACSI2016-tutorial2.pdf

Automatic coloring (black and white photography → color) tool

(2016, May 31)

白黒→カラー化のトリック

Source: http://www.itmedia.co.jp/news/articles/1605/31/news119.html

- *Supervised Learning*

With DL, it is also possible to color black and white photos. The actual procedure is as follows. First, from a large amount of color photographs on the Internet, a black-and-white photograph is created by several methods. By doing this, you can get a huge amount of sets of black-and-white photos and color photos. We assign black-and-white photograph to an input of DL, color photograph to an output, and then perform supervised learning.

Automatic coloring 2

白黒→カラー化

Source: http://www.itmedia.co.jp/news/articles/1605/31/news119.html

Work at the cash register and Passport Control

pic.twitter.com/WFI6SnV591
http://gigazine.net/news/20150401-bakery-factory/

パン屋さんのレジの様子

The work required at each shop need only take pictures from various angles for one product. Even products that do not have bar codes can be automated.

The task of object recognition by human eyes has been replaced by machines, and that system is deeply penetrated into our social life.

羽田空港・成田空港の入国ゲートの様子

Prevent overlooking by doctors

内視鏡画像中のがん細胞
のAI同定

Example of polyp detection

Source:
http://www.itmedia.co.jp/news/articles/1707/10/news095.html

2017, July

- ✓ The aim is to prevent overlooking due to lesions and sites of occurrence difficult to detect with the naked eye, technological disparities of doctors.
- ✓ The established system at the National Cancer Center (NCC), Japan, uses AI employing deep learning to process its own algorithm for an image analysis at high speed.
- ✓ About 5,000 endoscopic images diagnosed at NCC are used for learning parameters in DL.

The future of medical care depends on medical stakeholders and AI

Home About Working with the NHS For Patients Data & Security Transparency & Independent Reviewers FAQs

2017, Feb

Helping clinicians get patients from test to treatment, faster

The results so far...

27 February 2017

We set up DeepMind Health to put the UK's most advanced technology at the service of patients, nurses and doctors.

In order to make use of the power of this deep learning for the medical field, DeepMind recently launched company DeepMind Health in UK (about 2 years ago). DeepMind is a company that produced Alpha-Go and Alpha-zero, which won the Go professional. UK is very active in using big data concerning health and medical care.

Amazing advancements of Speech recognition and text processing

Smart Speaker

https://smarthacks.jp/mag/23875

Amazon and Google don't have an intention to make a profit by selling these products. Through these products, the purpose is to store the living space information of various people as a whole big data. And, in order to create the next information service from big data, to dominate the market and gain profits, they are daringly making "manufacturing".

By adopting a deep neural network suitable for processing sequence data with long correlation called LSTM in November 2016, the automatic translation function has improved significantly.

Automatic translation

The judge becomes a job of AI

Precog in Minority Reports (2002)
Robot Police in Elysium (2013)

- Utilizing AI is not limited to lawyer work, and its introduction is also under way in the hearing process of the trial. There are many mistakes due to misunderstanding of the clerk, etc. in the record to be recorded. It can be expected that the burden will be reduced.
- It is said that AI judges can realize by learning past case data. The judge is a suitable job for AI in that it absorbs the data of enormous cases in the past and *can be designed* to make an objective judgment rather than human.

Bank work disappears at AI

1) Window operations such as deposit transfer

It is almost compatible with ATM already. Virtual banking personnel will now respond to taxpayers who need manual labor.

2) Loan and foreign exchange business

The machine automatically judges the analysis and rating of the company to be financed and can also present conditions such as interest rates. AI system for foreign exchange can check documents and exchange money.

3) Securities / investment trust sales business

Algorithmic trading that automates stock ordering is mainstream. A choice of investment targets and establishment of investment period by AI are also advanced.

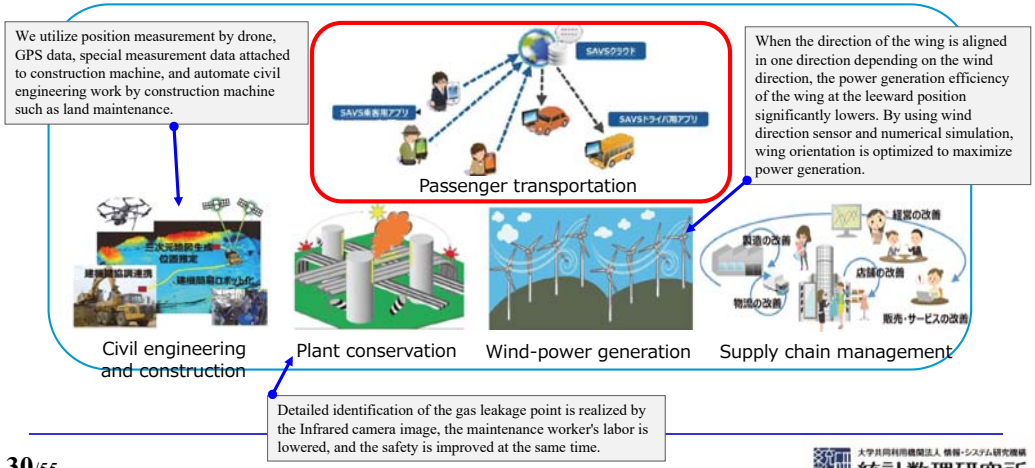
4) Trust business

In addition to administrative management departments and operations departments in real estate and securities, advice tasks such as testamentary trusts and inheritance countermeasures are replaced with AI.

NEDO(New Energy and industrial technology Development Organization) R&D project for AI and robotics

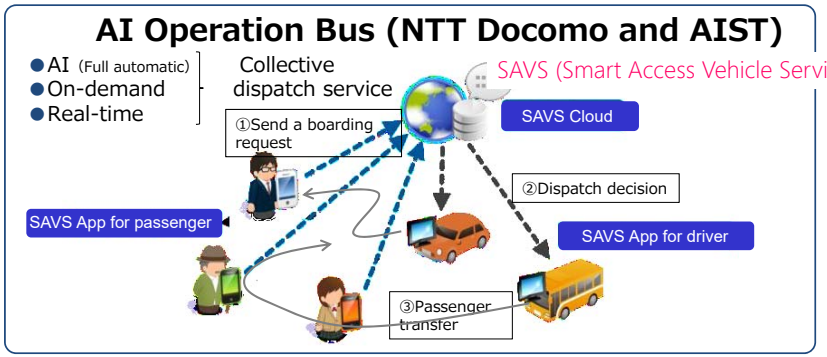
NEDO has started the project "Integration technology development as the core of next generation AI and robot" project aiming for early social implementation of AI technology from last year.

Social implementation of AI technology to 5 areas



Yokohama MaaS (Mobility as a Service) Demonstration experiment of AI operation bus

- In order to realize on-demand multipart traffic in urban areas, it is necessary to have a huge amount of movement demand unique to urban areas, and data that serves as a basis for allocation/strategy in the large change.
- In the "AI Operation Bus" demonstration experiment, we acquire vehicle dispatch data and data on the user in urban areas, as well as user data such as convenience and comfort. By using the acquired data, we can realize dispatch/operation processing by AI corresponding to the movement demand in urban areas.



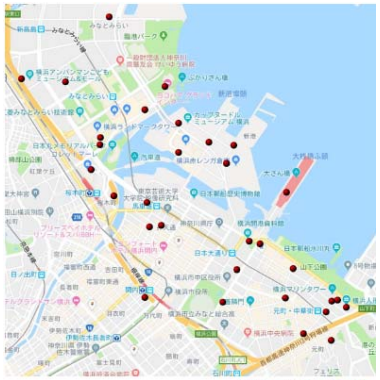
横浜MaaS実証概要＋輸送数推移

2018年12月10日
株式会社ＮＴＴドコモ
IoTビジネス部 先進ビジネス推進

横浜MaaS実証概要

「便利な移動手段により商業施設の利用が高まる」
「魅力ある商業施設のため交通需要が増加する」　の好循環を促す
“便利な移動＋サービス”を連携させるMaaSプラットフォームによる実証

【実証実験概要】
- ◆ 実施時期：2018年10月5日～12月10日
- ◆ 利用方法：
  - ・専用スマートフォンアプリ等による配車要求
  - ・協賛する商業施設の情報やクーポンをスマホに配信
- ◆ 輸送能力：同時に50人（10台程度のタクシー車両）
- ◆ モニタ：観光客を中心とする来街者（無料）
- ◆ 実証の特徴
  - ・オンデマンド/ダイナミックルート/シェアリング交通
  - ・500程度の地場商業施設と連携(実証期間中に倍増)



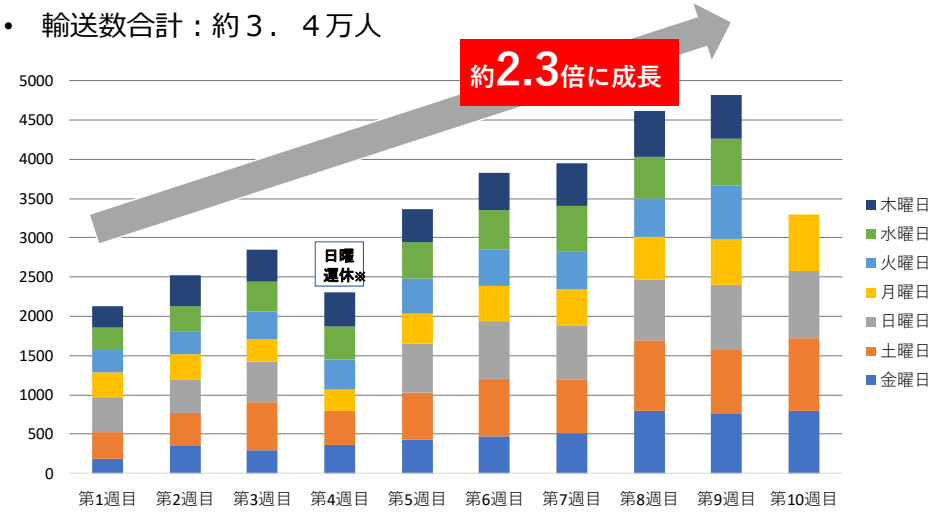
みなとみらい21、関内エリアに
31ヶ所の乗降ポイントを設置
（サイネージ型予約端末「まちかど端末」は28か所設置）

【提供価値（仮説）】

観光客	商業施設	自治体
便利な移動・お得な旅行	来街者に合わせた情報発信	災害時の来街者の影響把握

【速報】輸送数推移（週毎の利用数）

- 開始から９週間の輸送数は約２３０%の成長であり、その後も伸び続け、ニーズの存在が認められた。
- 輸送数合計：約３．４万人



2018年10月5日(金)～2018年12月10日(月)（66日間）の総輸送状況
※10月28日(日)は横浜マラソンで終日運休

Deep Learning with Reinforcement Learning

- Google AlphaGO wins the world pro

AlphaGO

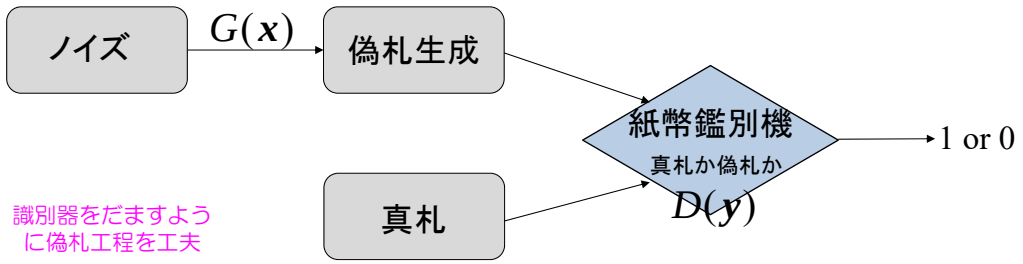
将棋のプロ vs. AI

The biggest feature of **Alpha Zero** is that it learned three games with the same program. When it repeat self-game, it will initially show random hands and hands, but gradually learn from the outcome of win and lose and choose more advantageous hands. The time required for learning until reaching a certain strength is about 9 hours for chess, about 12 hours for shogi and 13 days for go. It has become strong until they beat other AIs of the world champions level. Traditional artificial intelligence used **practical game record for learning**.

いたちごっこ：GANの基本アルゴリズムの比喩的解説

$D(y)$：真札である確率

$G(x)$：偽札を生成するモデル（偽札工程）



識別器をだますように偽札工程を工夫

真札・偽札を正しく識別

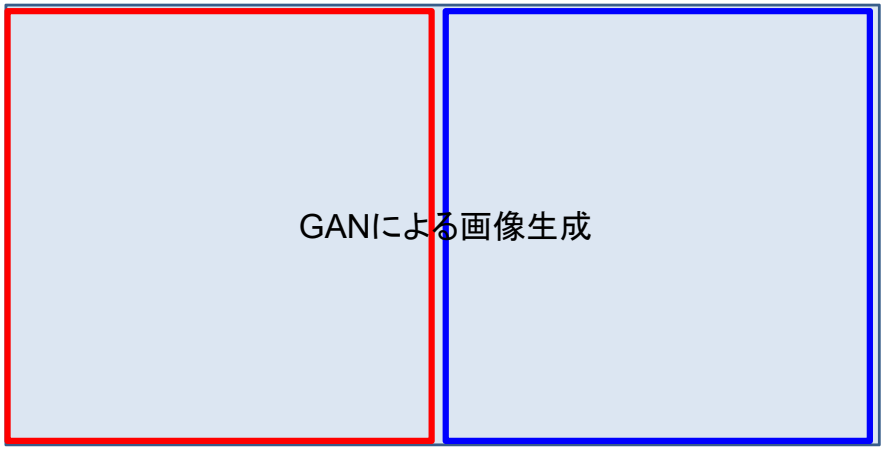
学習アルゴリズム

$$\min_G \max_D \mathrm{E}_{y\sim p_{\mathrm{data}}(y)}\left[\ln D(y)\right]+\mathrm{E}_{x\sim p_x(x)}\left[\ln\left(1-D(G(x))\right)\right]$$

GANを使った報告は、うまくいった例しか出さないケースが多々あり！

Deep Convolutional Generative Adversarial Network（DCGAN)

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
Alec Radford, Luke Metz, Soumith Chintala,
(2016)

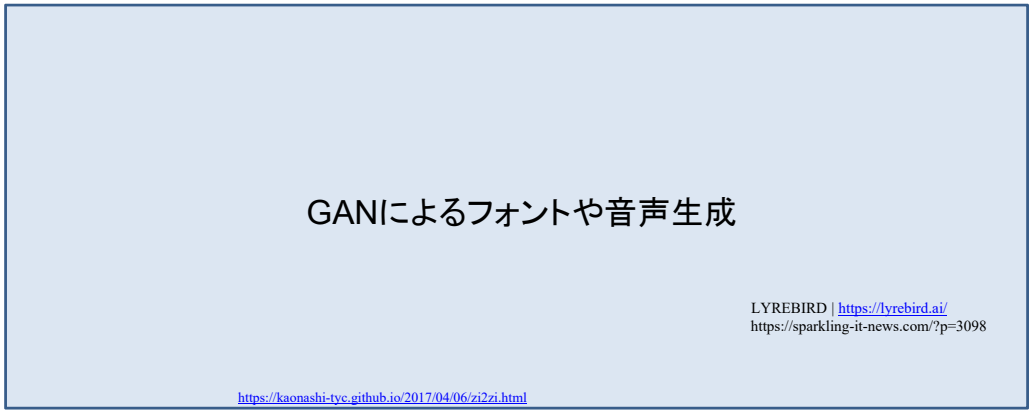


From https://ishmaelbelghazi.github.io/ALI/

http://www.whichfaceisreal.com/index.php

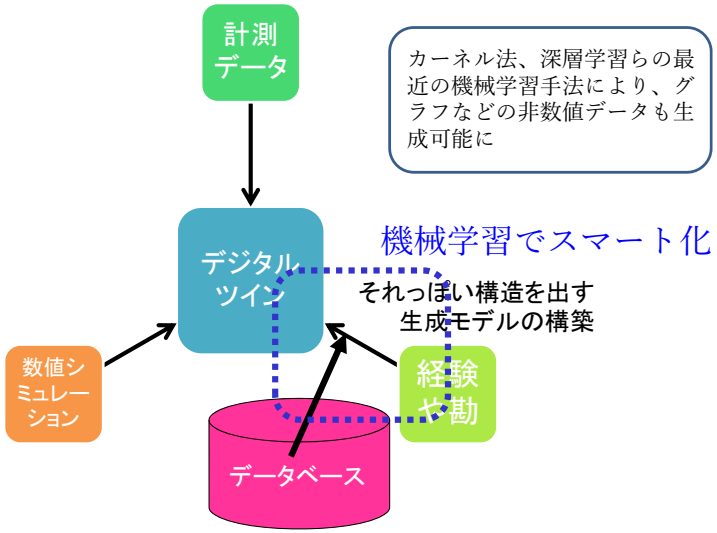
GAN的技術の先にあるのは....

■ 真贋判別の機能を具備する器機には破壊的影響あり



2017年4月6日　　2017年4月26日

暗黙知（経験と勘）生成モデル



カーネル法、深層学習らの最近の機械学習手法により、グラフなどの非数値データも生成可能に

機械学習でスマート化

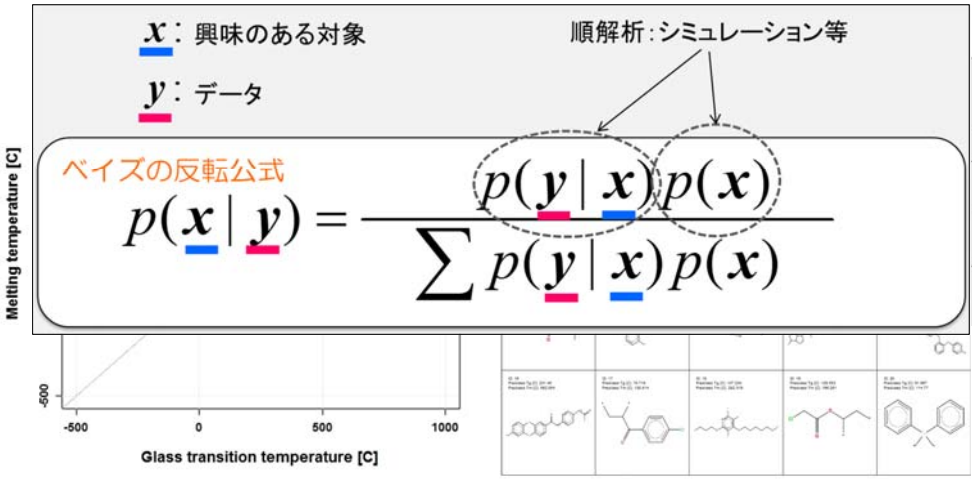
それっぽい構造を出す生成モデルの構築

NEW FUNCTIONAL POLYMERS

Higher glass transition temperatures and higher melting points

PolyInfo

吉田亮教授@統計数理研究所



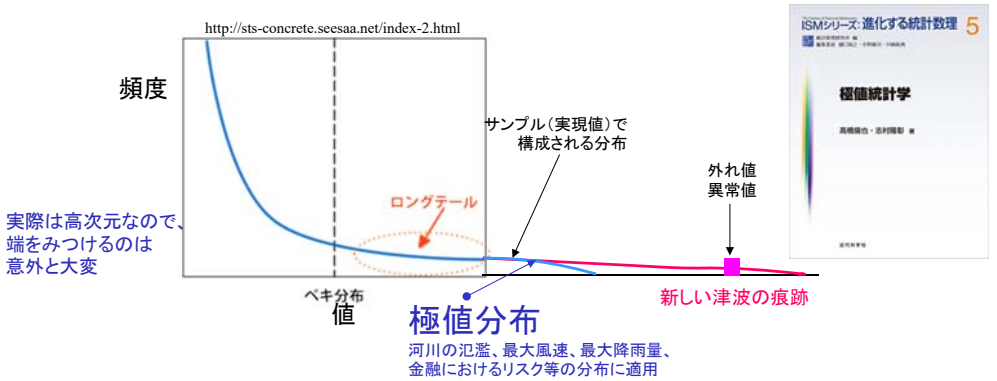
Orange: existing polymers with experimentally measured T_g and T_m
Blue: polymers computationally created by *iqspr* (circle size: the magnitude of likelihood)

アウトライン

1. 平成の30年間
2. 統計学、機械学習、深層学習
3. 帰納法の弱点
4. 人材育成

帰納法の弱点１：不（未）観測とデータの質

サンプリングの問題

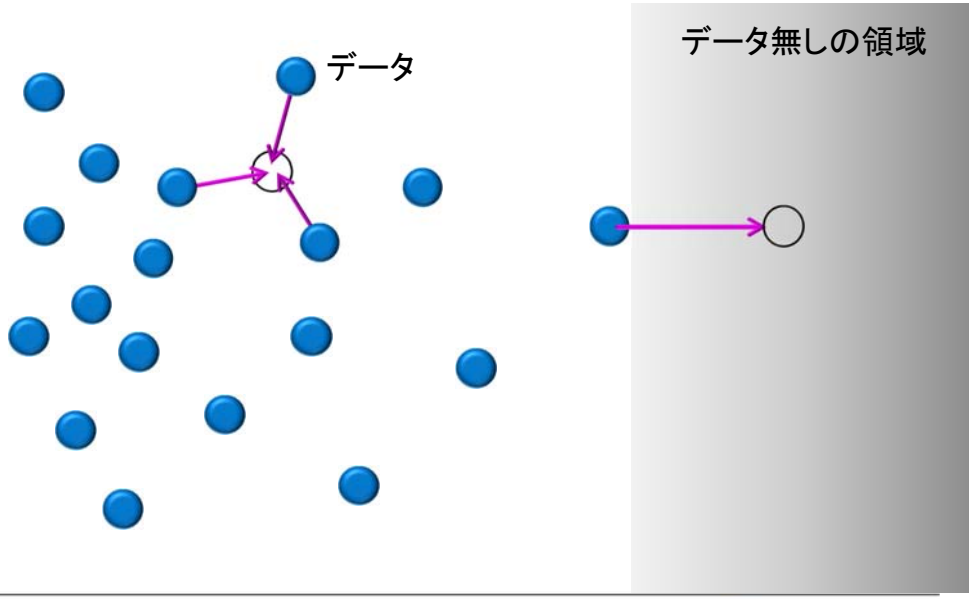


『端にこそイノベーションの卵』

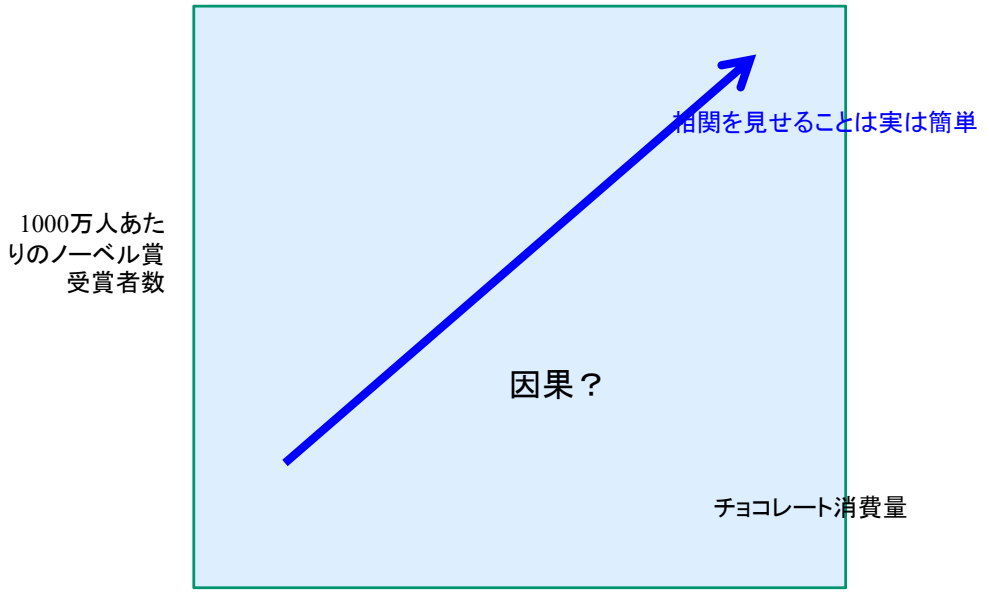
- 新発見、ひらめき
- クレーム（PL法対応） ←エラー、故障、不正、侵入

✓ データの偏り
✓ そもそも観測できていない？

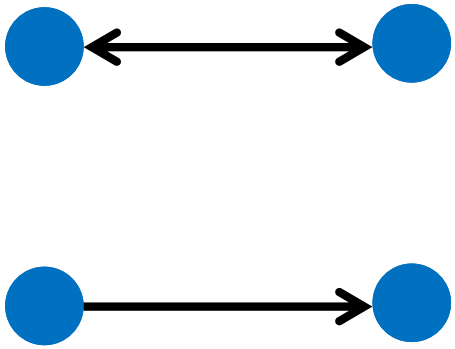
内挿と外挿



帰納法の弱点２：因果と相関



相関と因果



帰納法の弱点3：偽陽性と偽陰性のバランス

画像誤認識

Google の痛い失敗
2015年 Googleのタグ付け性能

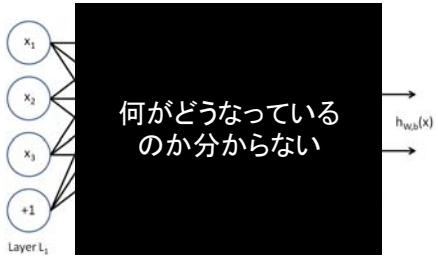
ロンドン警視庁サイバー班の悩み
ソース　2017/12/15 ネットニュース
http://news.livedoor.com/article/detail/14056063/

テスラ　半自動運転モードで追突
2018年1月22日（米国時間）、
ロサンゼルス郡の高速道路405号線

医療診断：　偽陰性を避けるため、偽陽性を甘受
自動運転：　偽陽性や誤検知を避けるため、情報をあえて無視する制御

データのみからルールを定める帰納法の限界
事例ベース、（セミ）ブラックボックス

DNNはブラックボックス →　グレイボックスへ



http://deeplearning.stanford.edu/wiki/images/4/40/Network3322.png

交通標識の
AI誤判断

Figure 2. Outline of the *DeepFace* archi

- ✓ 結果をみても何が特徴なのかわかりづらい。そもそも人海戦術。
- ✓ 途中の計算がうまくいっているのかどうかが分からない。

グレイボックス

要因を読み解く必要性は何か？

- 可読性が大切　→
  何が要因で、どんなメカニズム？
- 個別化（ニーズや利用状況にあわせること）が容易

■ブラックボックスでは機能しない

Fairness
Bias as a technical matter

- ✓ Interpretability
- ✓ Accountability
- ✓ Reproducibility
- ✓ Transparency

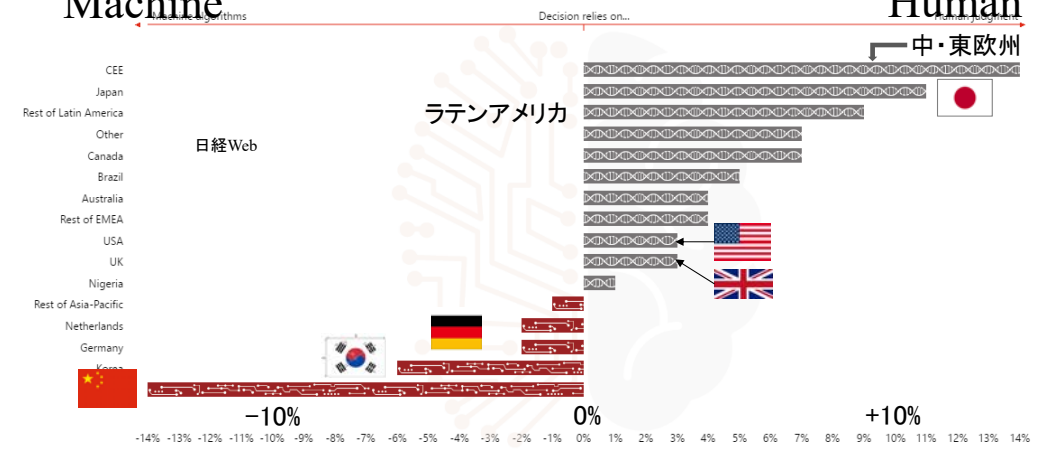
Foundation of Statistics
- Data imbalance
- Sampling strategy
- Generative model

25日に施行
タ保護規則（GDPR）」

志向性の格差

- A good mix augments human judgment with machine algorithms to create better outcomes.
- What type of analysis will inform your next big decision?
- データの利活用が最も進んでいるUSでは、経営層は最終的には人間の判断を優先する傾向がある。
  我が国はデータ利活用人材の教育や人数、雇用されている企業規模が小さいにもかかわらず、
  データよりも人間の判断を優先する傾向が強い。

Machine　　Human



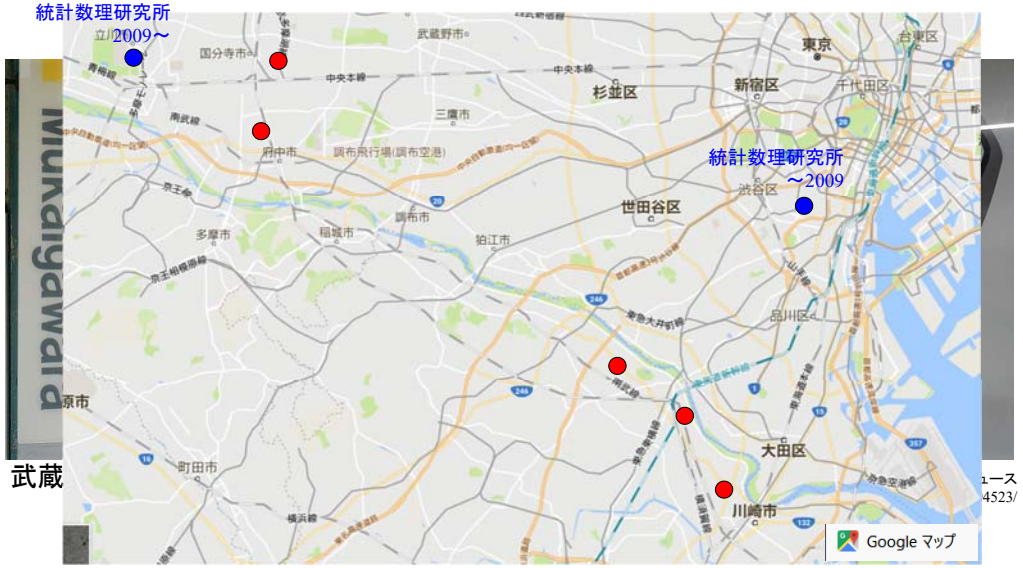
http://www.pwc.com/us/en/advisory-services/data-possibilities/big-decision-survey.html

（統数研・神谷特任准教授のスライド(2017)から改編）

アウトライン

1. 平成の30年間
2. 統計学、機械学習、深層学習
3. 帰納法の弱点
4. 人材育成

激しい生き残り競争



US jobs rankings 2016: actuaries drop to 10th

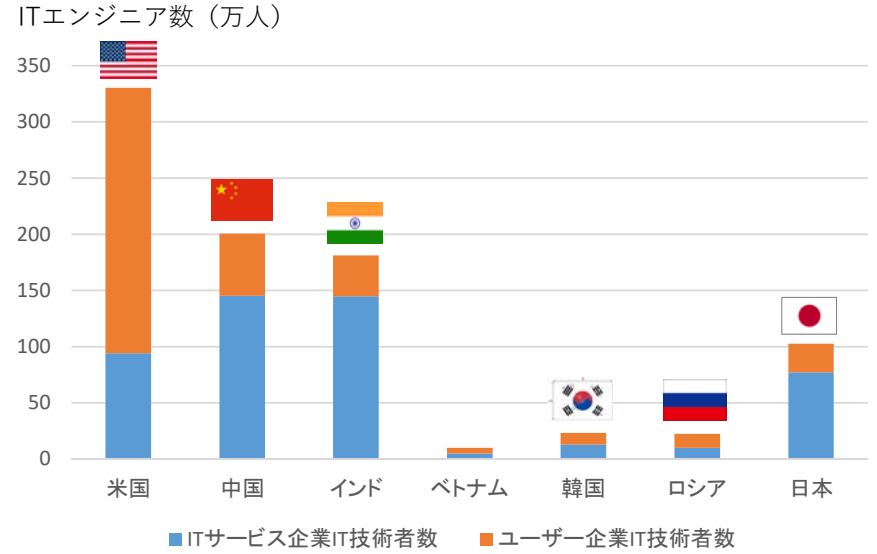
Latest annual Jobs Related report based the rankings on four core criteria: environment, income, outlook and stress

Top 10 best jobs and how they fared in 2015 in brackets, with annual median salary

1. Data scientist (+5) – $128,240
2. Statistician (+2) – $79,990
3. Information security analyst (N/A) – $88,890
4. Audiologist (-2) – $73,060
5. Diagnostic medical sonographer (N/A) – $62,540
6. Mathematician (-3) – $103,720
7. Software engineer (+1) – $97,990
8. Computer systems analyst (+1) – $82,710
9. Speech pathologist (+2) – $71,550
10. Actuary (-9) – $96,700

- See more at: http://www.theactuary.com/news/2016/04/us-jobs-rankings-2016-actuaries-drop-to-10th-place/#sthash.bhVnrjl4.dpuf

人材数の格差：ベンダー or ユーザー企業



IPA（情報処理推進機構）「グローバル化を支えるIT人材確保・育成施策に関する調査」

（統数研・神谷特任准教授のスライド(2017)から改編）

データは副産物。そして抽象物。

Feature Vector (Descriptor)の選択

End-to-end Learning

Data cleansing, Data Editing, Data Curating

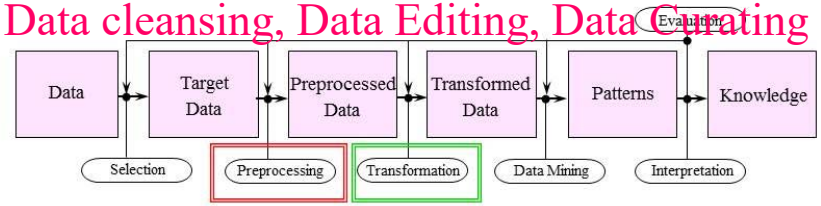
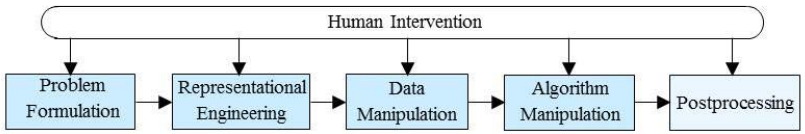


図 1: Fayyad等による知識発見のプロセス



データ処理の流れ 図 2: **Langley**による知識発見のプロセス

発見科学研究 成果報告集