統数研記念式典

人工知能の歴史、発展、社会への影響

甘利俊一

理化学研究所　栄誉研究員
東京大学　名誉教授

人工知能の衝撃

人間の知的能力を超えるのか？
第4次産業革命

人工知能の歴史：
記号と論理ー並列分散（ニューラルネット）

動力技術

機械技術

材料技術

生命技術―遺伝子編集

情報技術―知能・心

技術は止まらない！

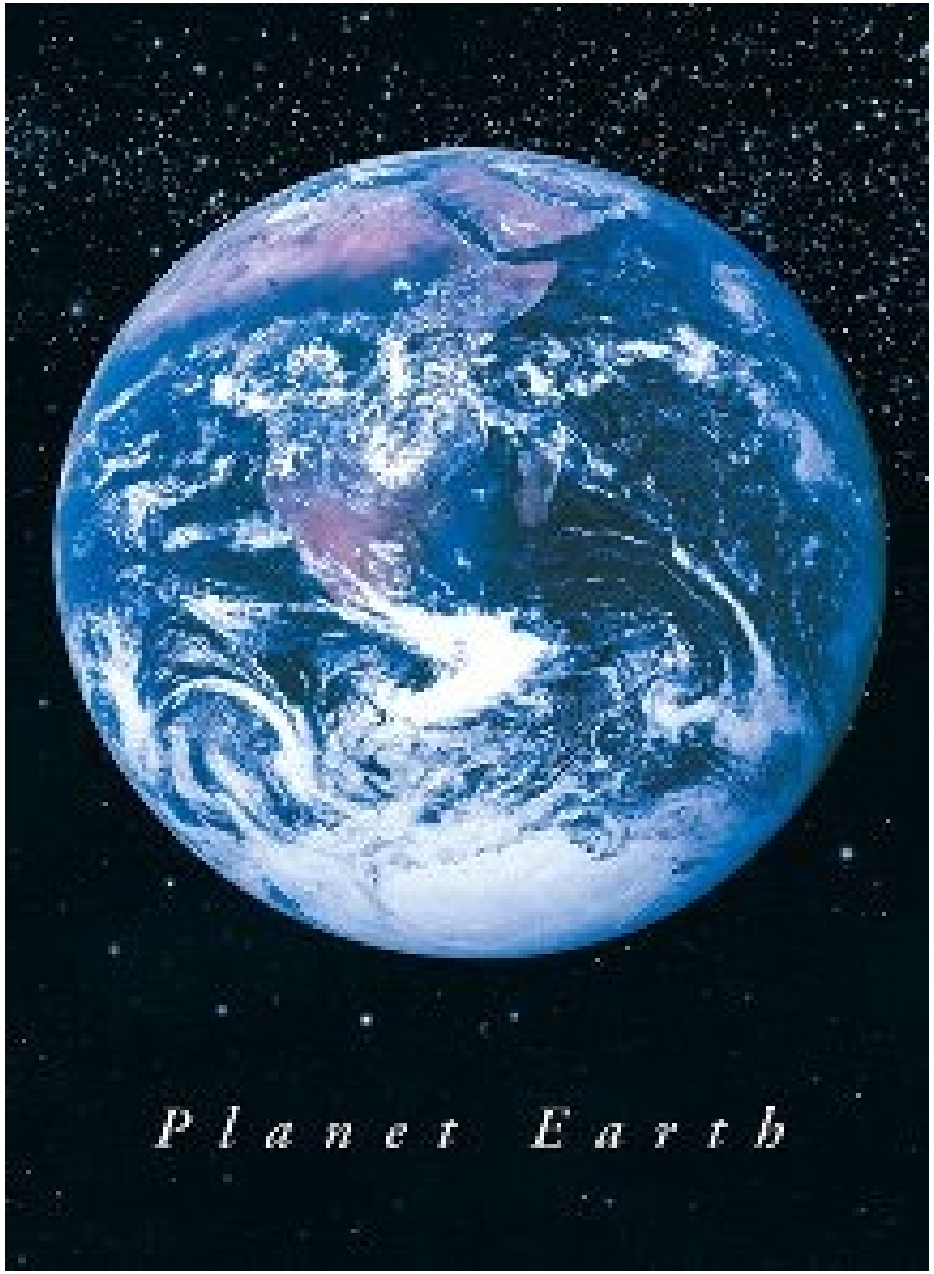
脳――人間とは何か

思考・言語・意識
人間・社会・文明

宇宙誌と脳　―　脳ができるまで

ビッグバン	(138億年前)	物理学・化学
生命	(36億年前)	生命科学
脳・神経系	(5億年前)	神経科学・情報科学
文明・社会	(20万年前？)	脳科学・情報科学・人間科学


ビッグバン（138億年前）
物質の法則
エネルギー・物質 ー 天体 ー 分子(秩序)
物理学、化学

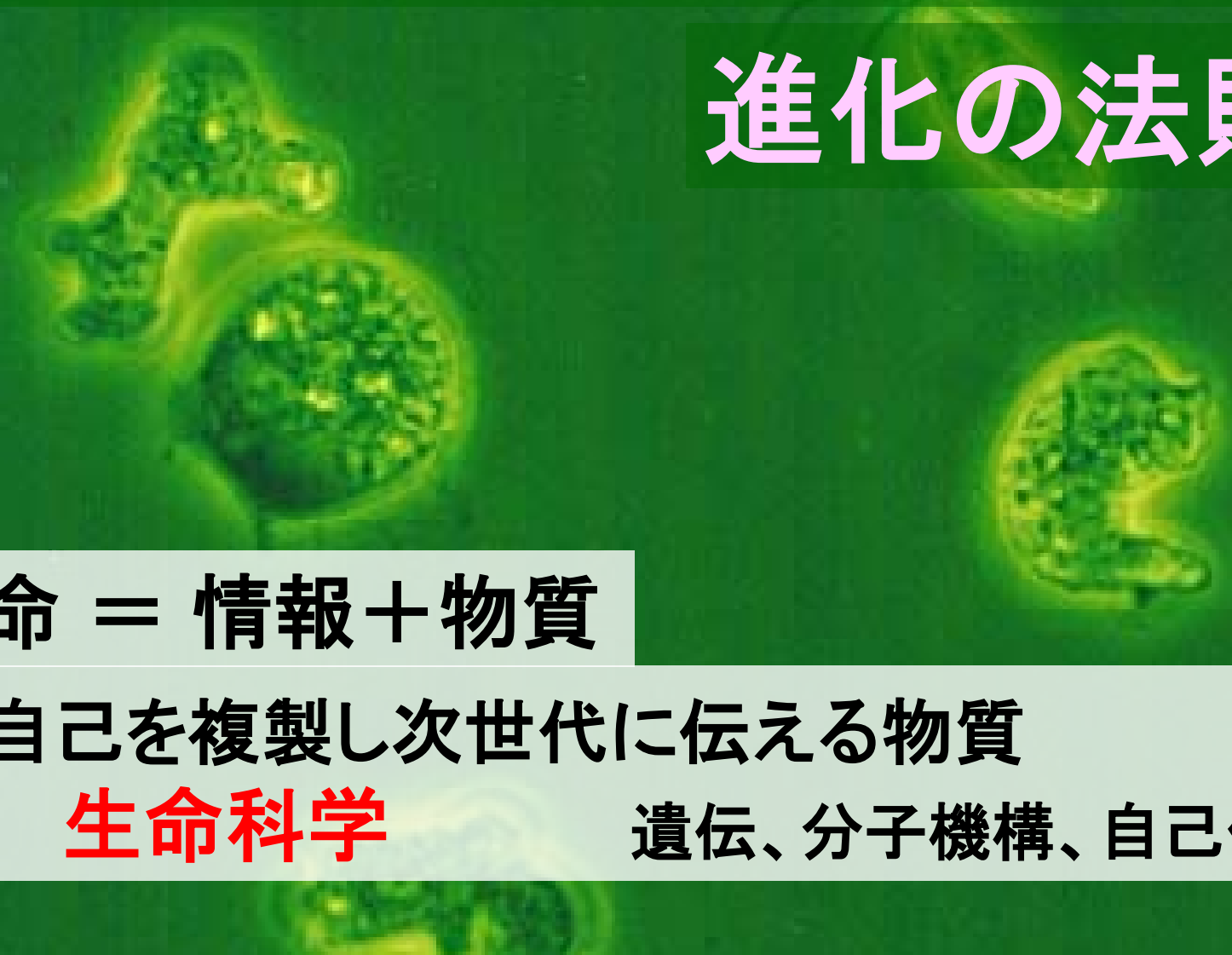


46億年前：
地球の誕生

火の玉、全球凍結

生命の誕生（36億年前）

進化の法則



生命 ＝ 情報＋物質

＝自己を複製し次世代に伝える物質

生命科学　　遺伝、分子機構、自己保存

DNA

多細胞生物

環境の情報を利用、
記憶･学習、判断･行動

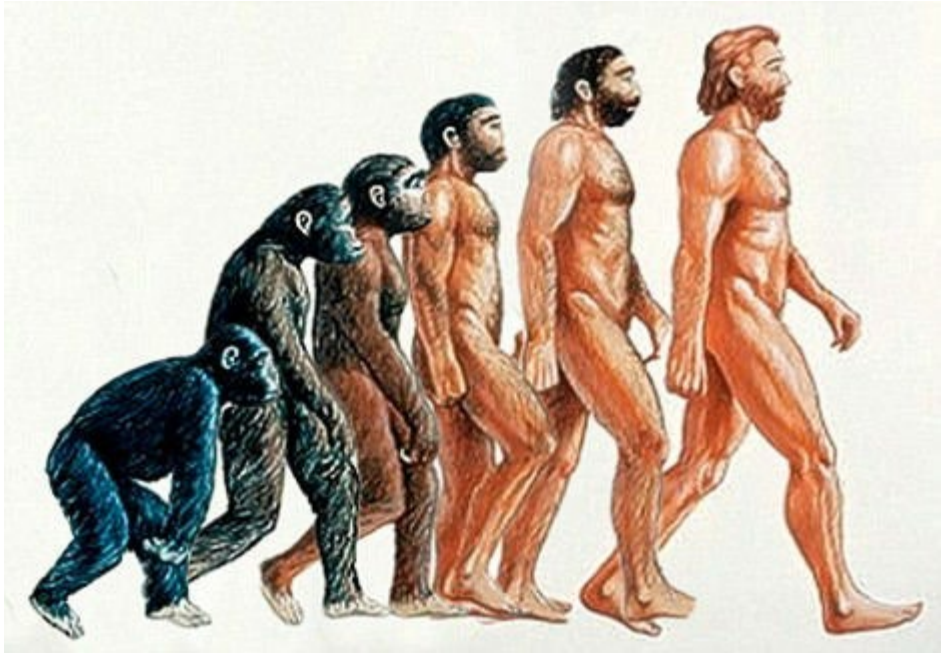
脳＝情報

脳･神経系（5億年前）：神経科学

類人猿、そして人間

社会に生きる生命；文化と社会

人類の登場（700万年前）



心、意識
文明

猿人、原人
旧人（ネアンデルタール、50万年-3万年前）
新人（ホモ・サピエンス、20万年前-現在）

心の理論
ロボットに意識はあるか

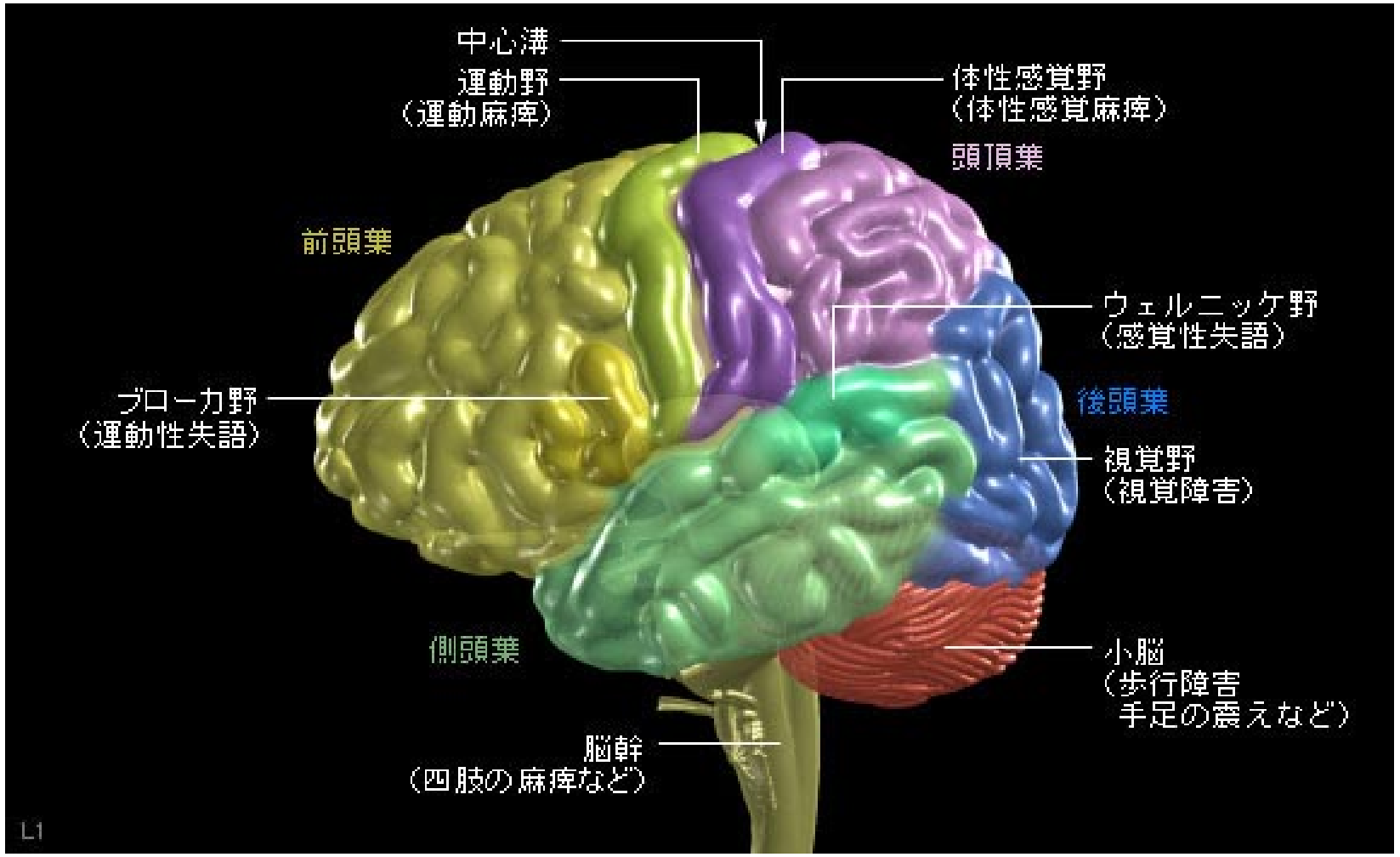
物質の法則：　宇宙

生命の法則：　情報＋物質：　進化

文明の法則：　こころ＋情報＋物質：
社会･文化

脳：大脳、海馬、小脳、脳幹

脳科学：ミクロ―マクロ、理論：神経回路網



人工知能と脳のモデル：

―歴史の要約

第一次ブーム

1956~　　AI　　脳モデル

Dartmouth 会議　　Perceptron

記号と論理　　学習する普遍計算機構

知的推論、ゲーム　　線形分離可能

実用的でない!!

暗黒期　(1965後半~1970's)　　stochastic descent learning (1967) for　MLP

第2次ブーム

1970~ AI
エキスパートシステム
(MYCIN, DENDRAL)

1980~ BT (神経回路)
MLP (backprop)
連想記憶モデル、ダイナミックス
一兆円産業か？

沈静化

確率Bayes推論
chess (1997)

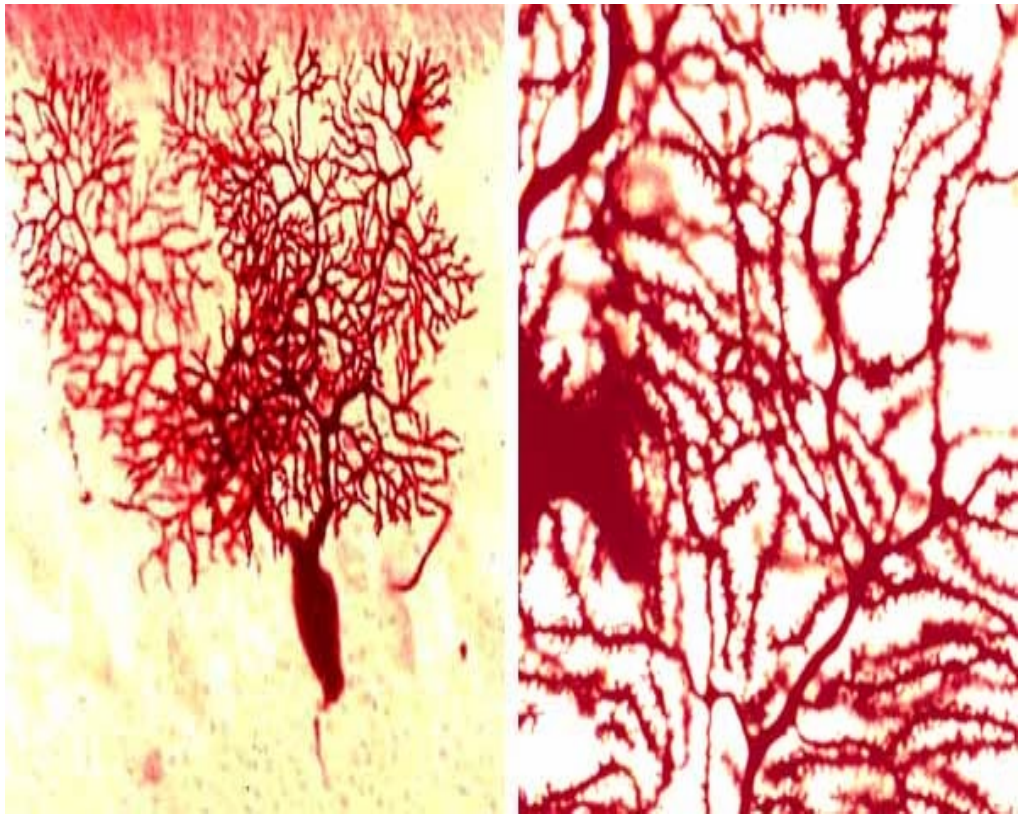
第3次ブーム　2010~　脳型の人工知能（融合）

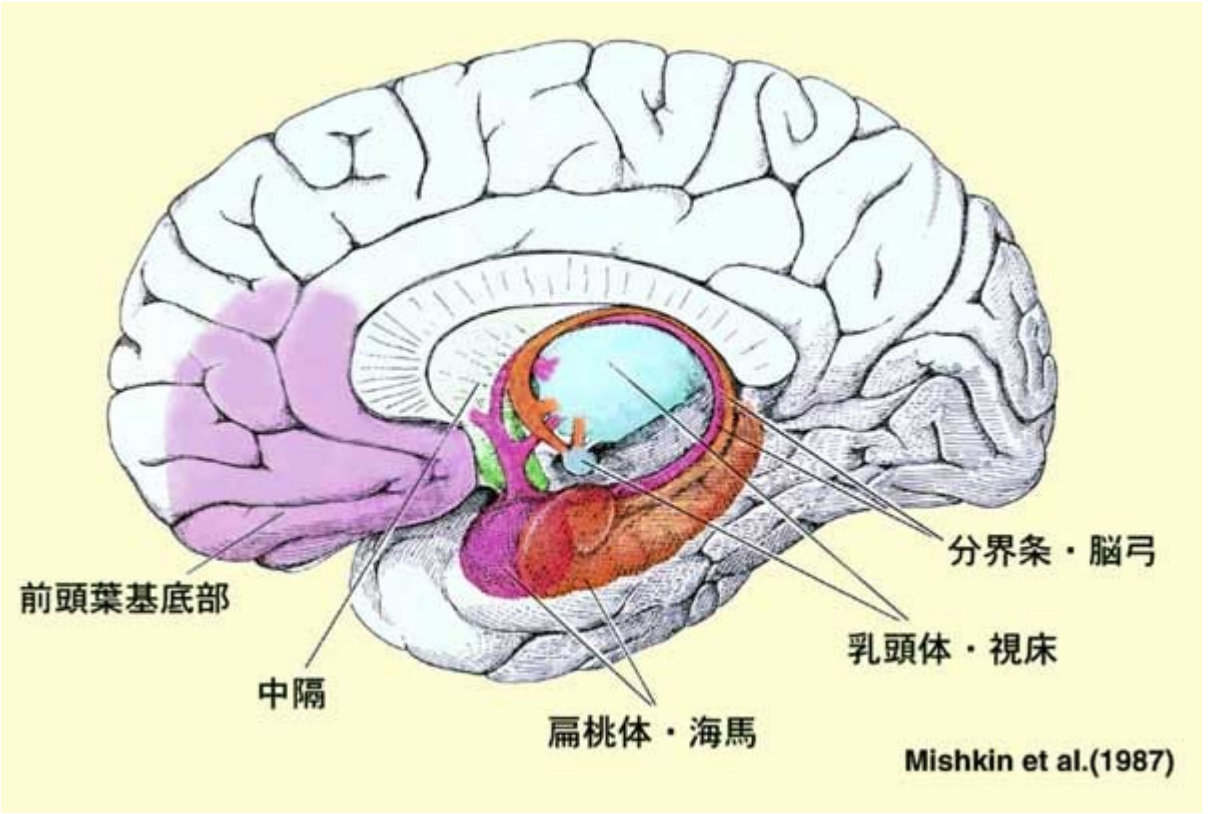
深層学習　Deep learning
（畳み込み多層回路（福島）＋確率勾配降下：日本でなぜ実現しな かったか）
確率推論 (graphical model; Bayesian; WATSON)

深層学習の勝利 ――人間以上の識別能力
パターン認識: vision, auditory, sentence analysis
　囲碁：強化学習
時系列とダイナミックス、動的パターン；言語処理

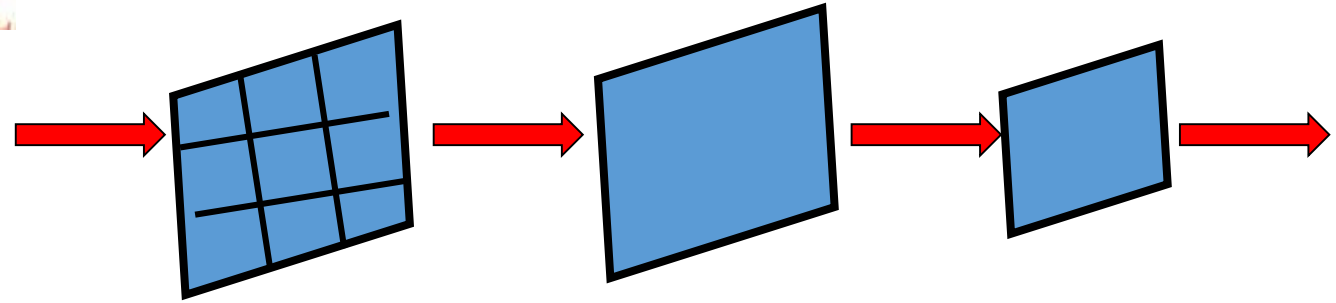
記号と論理　VS　パターンとダイナミックス、学習－融合

ニューラルネットと脳

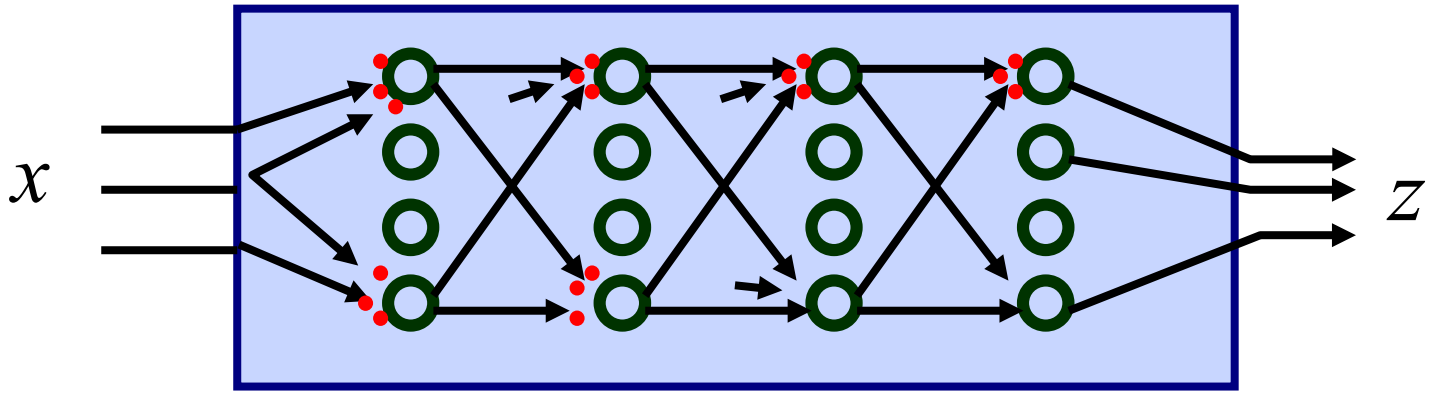




多層パーセプトロン
　表現　ダイナミックス　万能性



層状学習回路網
multilayer perceptron



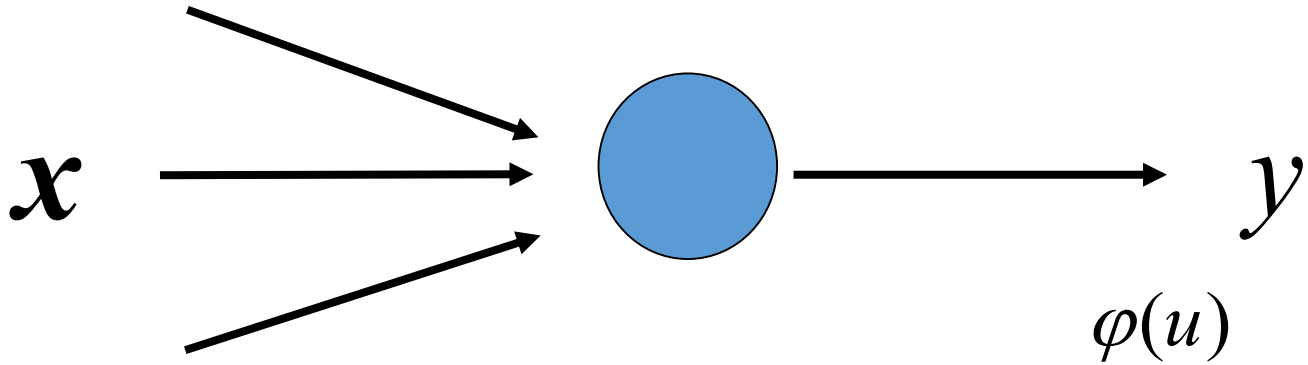
パーセプトロン　Perceptron
バックプロパゲーション　Backpropagation

$$L(\boldsymbol{x},W)=\left|y-\boldsymbol{g}(\boldsymbol{x},W)\right|^2$$

$$\boldsymbol{w}\rightarrow\boldsymbol{w}+\Delta\boldsymbol{w},\quad \Delta\boldsymbol{w}=-c\frac{\delta L(\boldsymbol{x},W)}{\delta W}$$

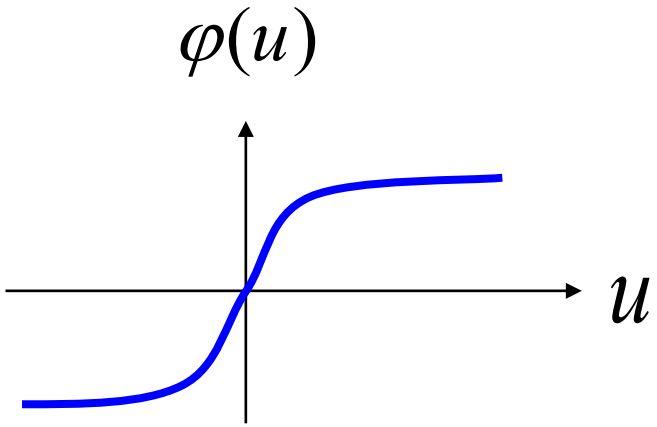
ニューロンの数理モデル

$$y = \varphi\left(\sum w_i x_i - h\right) = \varphi\left(\boldsymbol{w} \cdot \boldsymbol{x}\right)$$

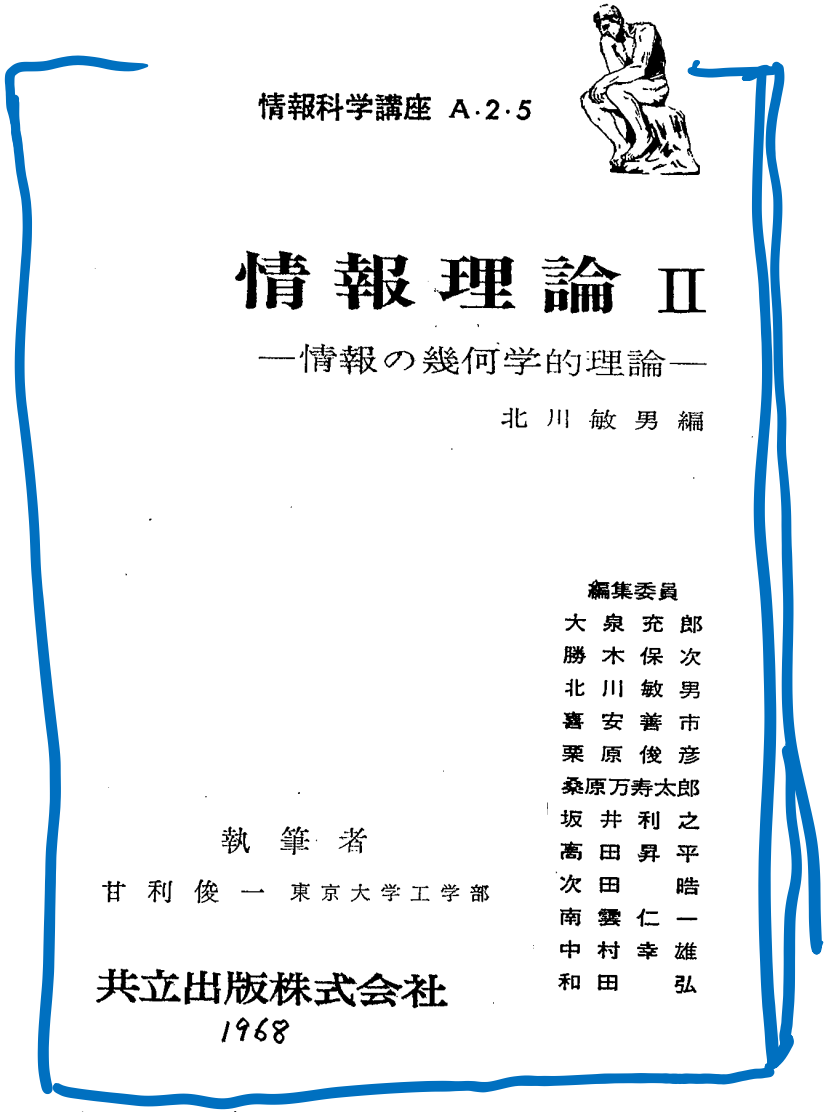


多数決素子

興奮・抑制



最初のMLPの確率勾配降下学習法 (1967;1968)



Information Theory II

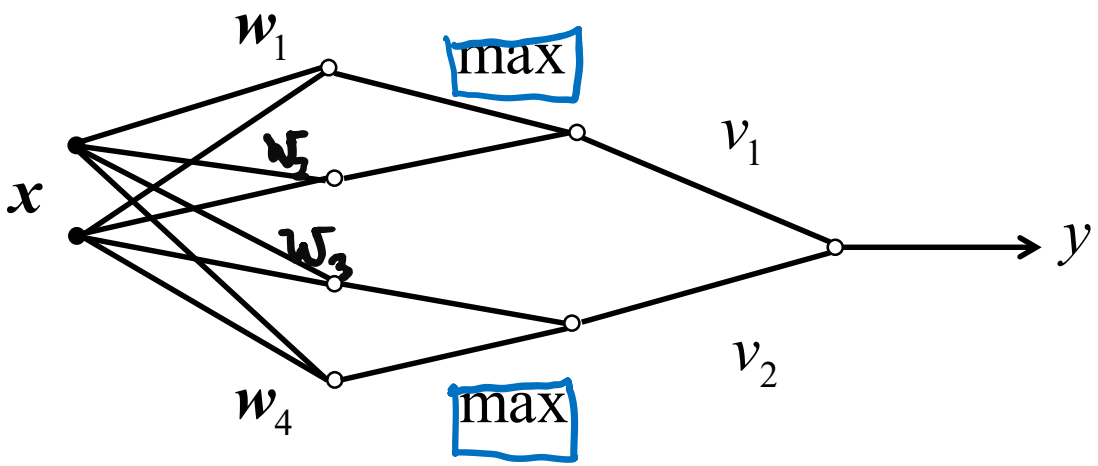
--Geometrical Theory of Information

Shun-ichi Amari
University of Tokyo

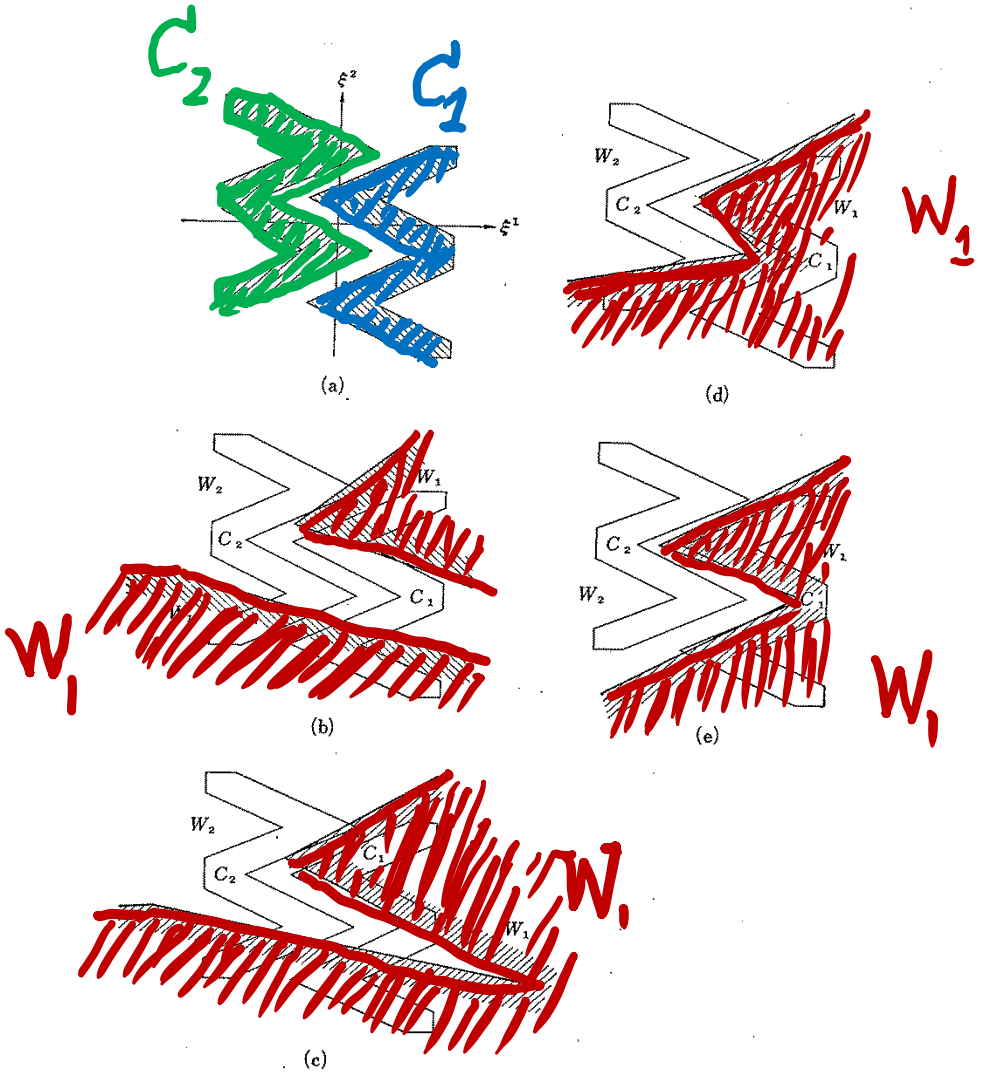
Kyouritu Press, Tokyo, 1968

$$f\left(\boldsymbol{x},\boldsymbol{\theta}\right)=v_1\max\left\{\boldsymbol{w}_1\cdot\boldsymbol{x},\boldsymbol{w}_2\cdot\boldsymbol{x}\right\}+v_2\min\left\{\boldsymbol{w}_3\cdot\boldsymbol{x},\boldsymbol{w}_4\cdot\boldsymbol{x}\right\}$$

アナログニューロン、シグモイド関数



線形分離不可能パターン分類



深層学習の勝利

パターン認識（画像、音声,...）
囲碁・ゲーム
テキスト生成
言語翻訳

深層学習の問題点　1

大量のデータ、計算力：　実験式：原理を発見しない

入力を基に正解

現象の予測

日蝕の予測

ケプラーの法則、ニュートン力学（新概念）

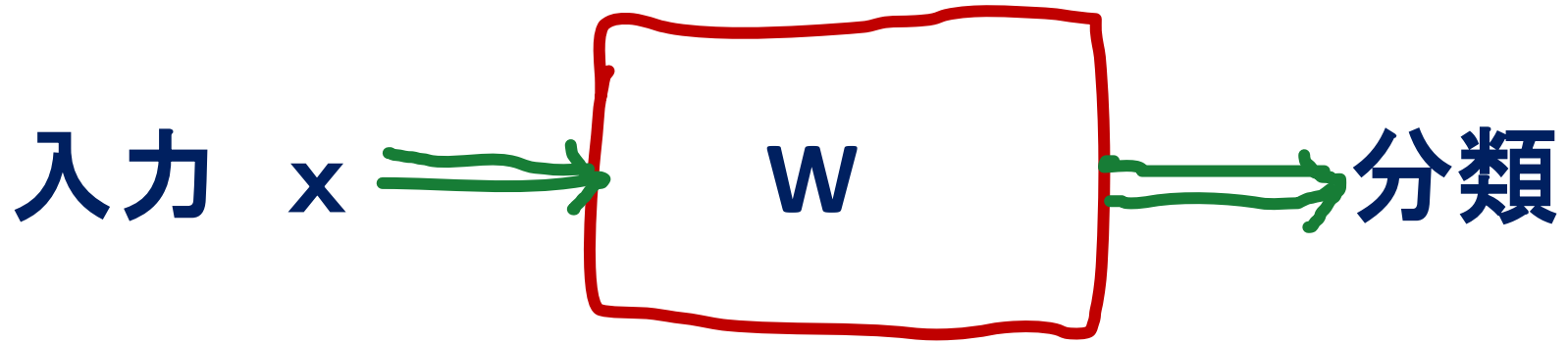
原理の創出・理解ー人間

科学

深層学習の問題点　２

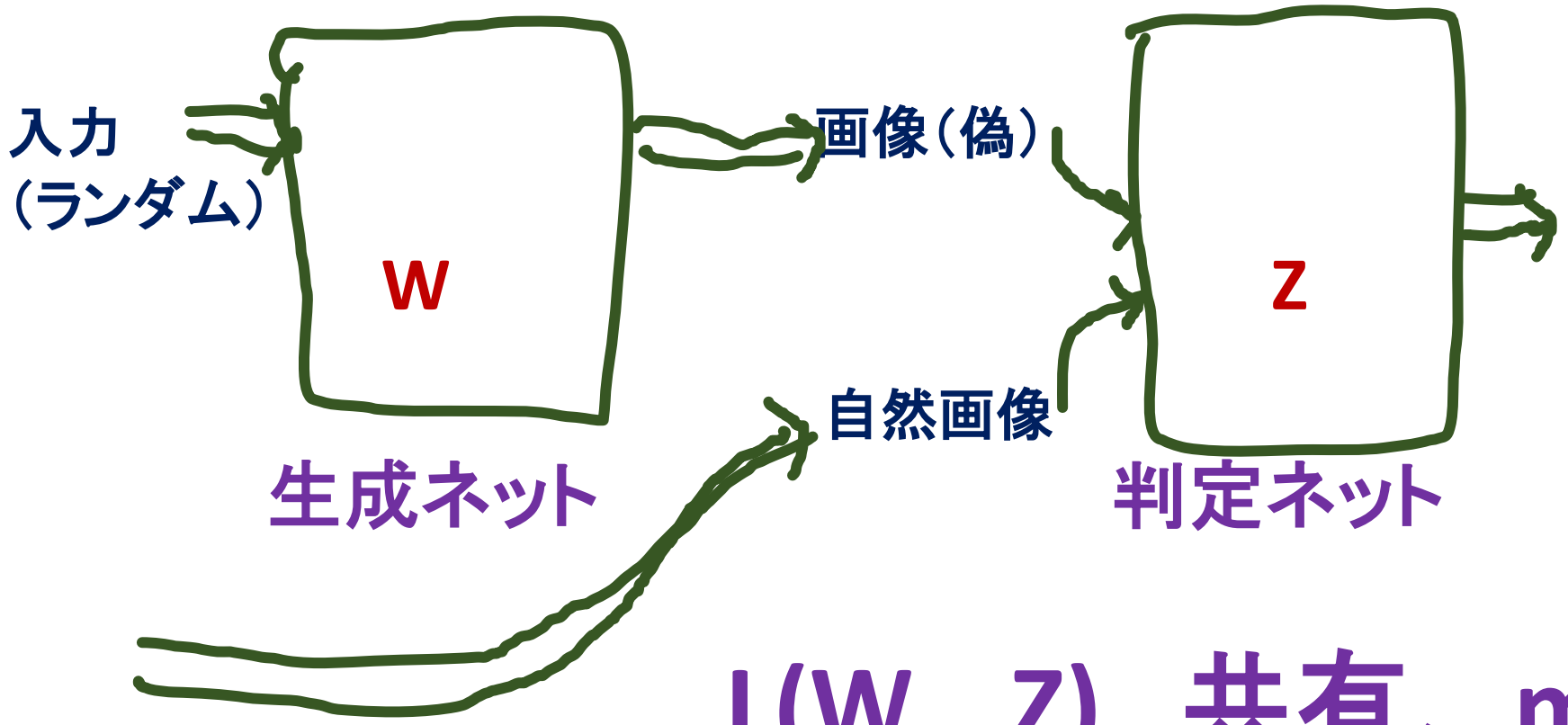
何故1000層も必要か　情報表現
敵対的例題と脆弱性

敵対的例題



xの学習

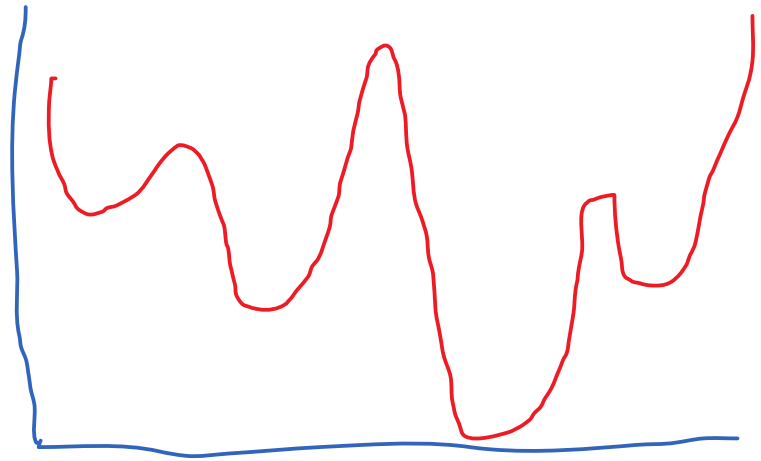
敵対的ネットワーク：　真贋判定回路



L(W,　Z)　共有、minimax

深層学習の問題点　３

局所解と大域解



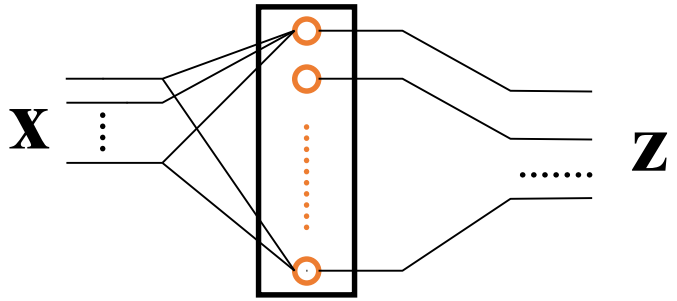
大規模系の特徴

ランダム行列 A の固有値の分布
　ほとんどが鞍点（極小解なし!!）

大規模回路
　極小解は最小解の付近に集まる

統計神経力学

1-layer network

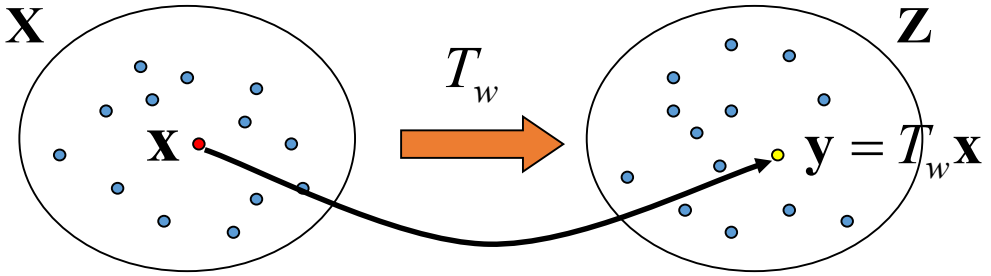


$$x_i = \pm 1$$

$$z_i = \pm 1$$

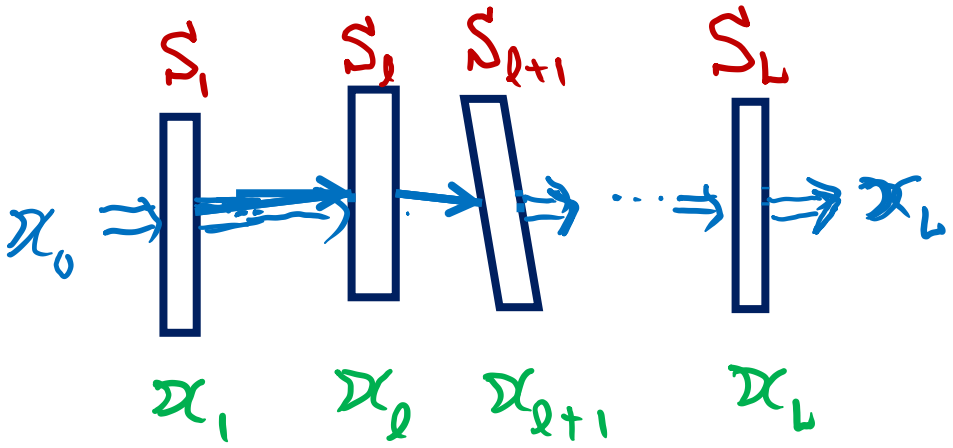
$$\mathrm{sgn}\left(\sum_{j=1}^{n} w_{ij} x_j\right)$$

$$\mathbf{z} = T_w \mathbf{x} = \mathrm{sgn}(W\mathbf{x})$$



深層回路

$$x_{\underset{l+1}{i}} = \varphi\left(\sum w_{ij}\, x_{\underset{l}{i}} + w_{0i}\right)$$



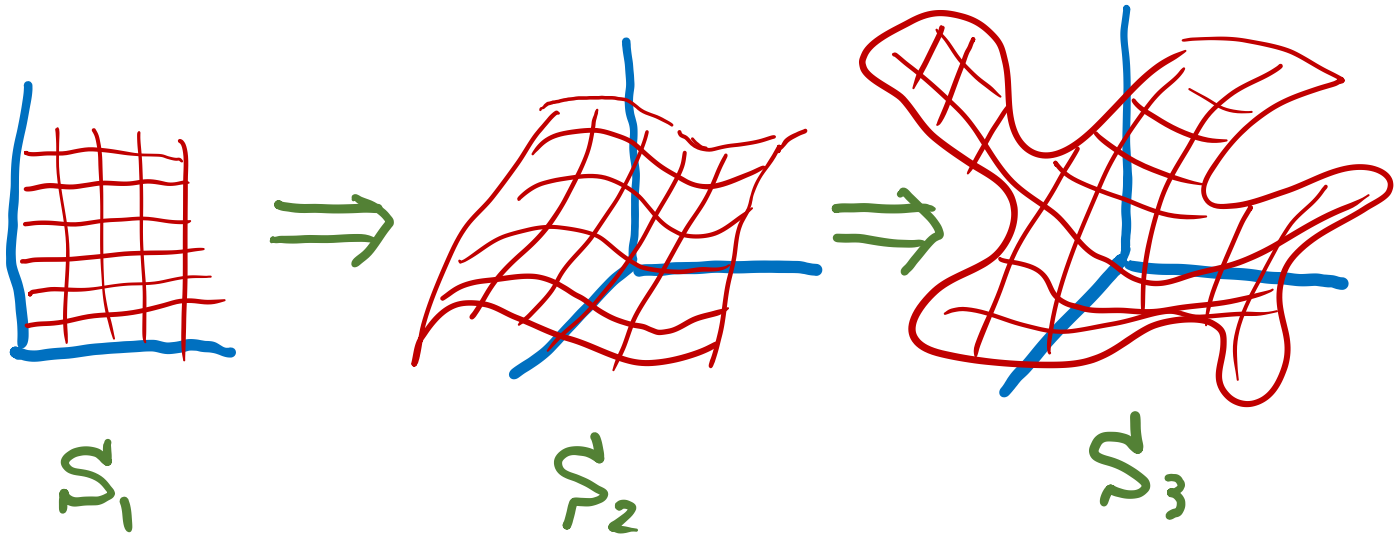
$$A_{\underset{l+1}{}} = F\left(A_{\underset{l}{}}\right)$$

$$A_{\underset{l}{}} = \frac{1}{n_l}\sum x_{\underset{l}{i}}^{\,2}$$

$$w_{ij} \sim N\left(0,\ \sigma^2 / \sqrt{n}\right)$$

$$w_{0i} = b \sim N\left(0,\ \sigma_b^2\right)$$

引き戻し計量（リーマン計量・距離）



$$ds^2 = \sum g^l{}_{ab}\, dx^a dx^b = \frac{1}{n_l} d\mathbf{x}^l \cdot d\mathbf{x}^l$$

$$g^l{}_{ab} = \mathbf{e}^l{}_a \cdot \mathbf{e}^l{}_b$$

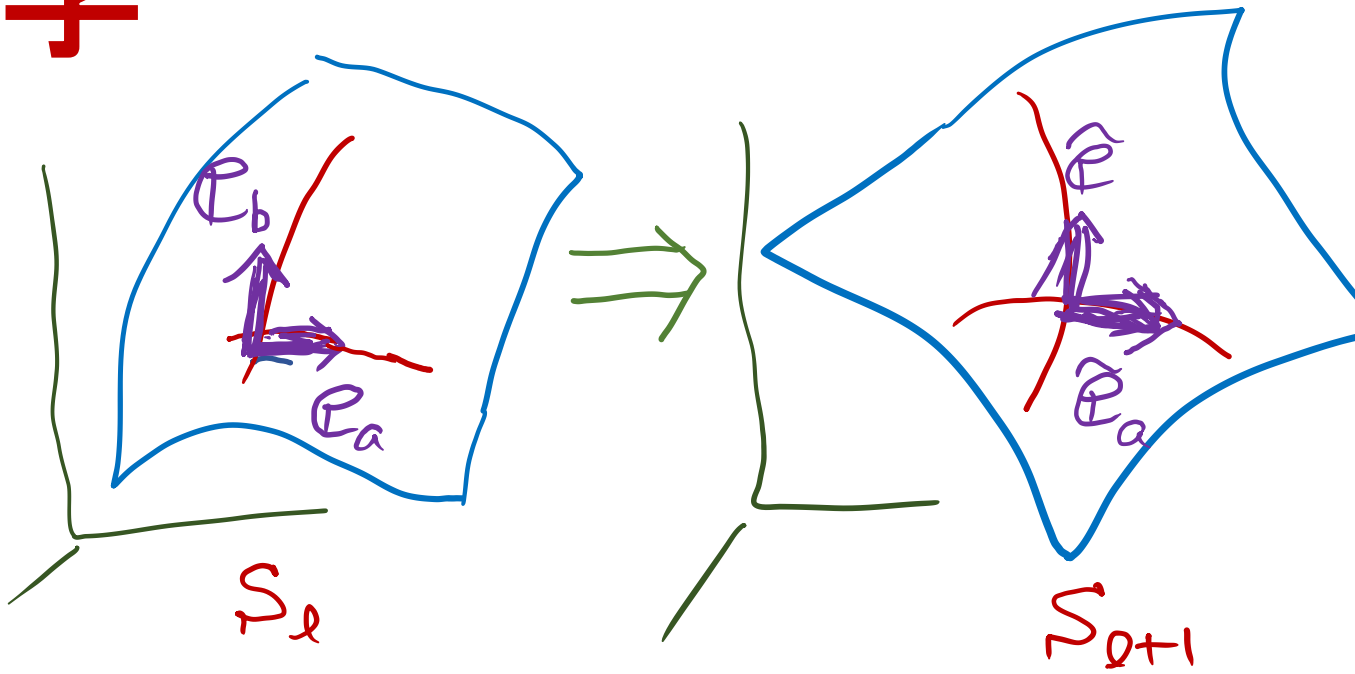
リーマン計量の力学

$$\tilde{y}_\alpha = \varphi(\sum w_{\alpha k} y_k + b_\alpha) = \varphi(u_\alpha)$$

$$d\tilde{y}_\alpha = \sum B_k^\alpha dy_k \qquad \tilde{\mathbf{e}}_a = B\mathbf{e}_a$$

$$B = (B_k^\alpha) = (\varphi'(u_\alpha) w_k^\alpha)$$

$$\tilde{g}_{ab} = \sum B_k^\alpha B_j^\alpha g_{kj} = <\tilde{\mathbf{e}}_a, \tilde{\mathbf{e}}_b>$$



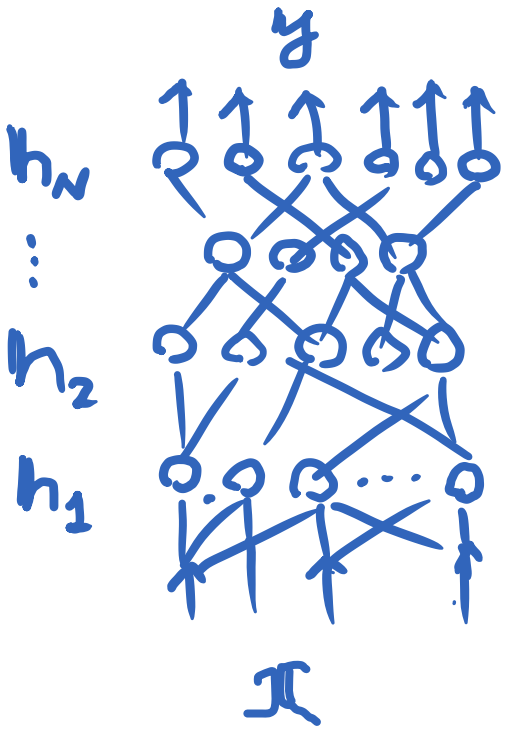
$$\mathrm{E}[\varphi'(u_\alpha))^2 w_k^\alpha w_j^\alpha] = \mathrm{E}[\varphi'(u_\alpha))^2]\mathrm{E}[w_k^\alpha w_j^\alpha]$$

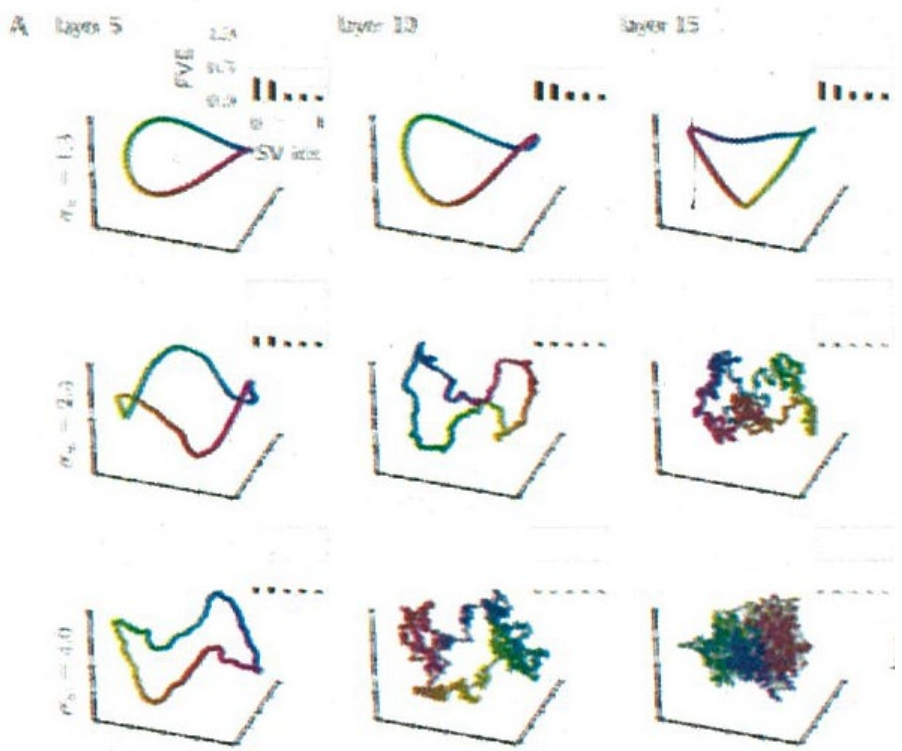
平均場近似不使用

$$\chi_1(A) = \int \sigma^2 \{\varphi'(\sqrt{A}v)\}^2 Dv = \frac{1}{2\pi} \frac{\sigma^2 A + \sigma_b^2}{\sqrt{1 + 2(\sigma^2 A + \sigma_b^2)}}$$

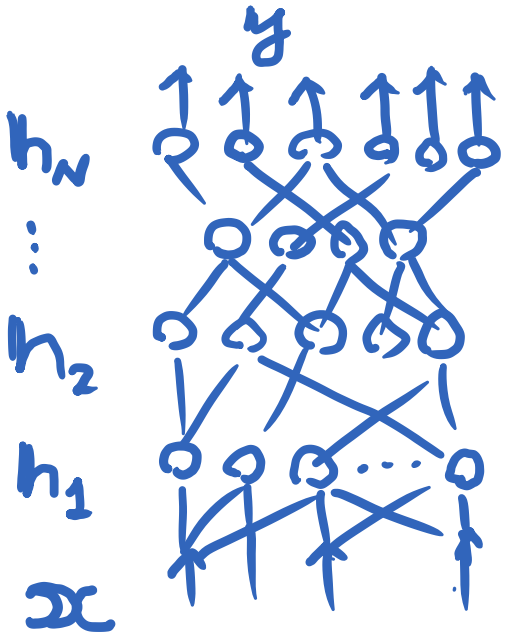
Poole et al (2016)
Deep neural networks

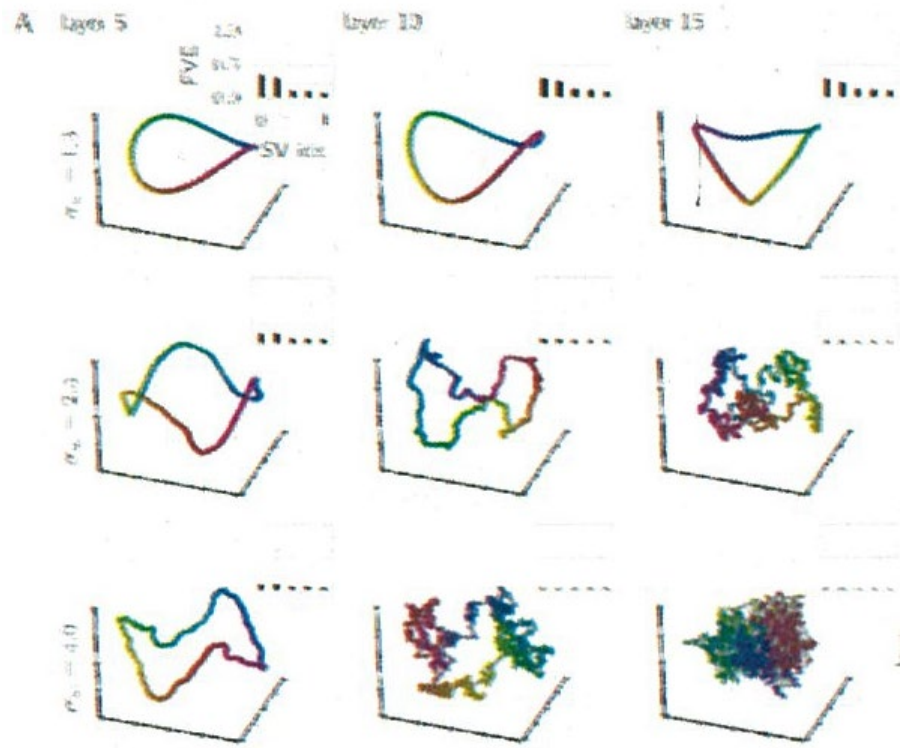
フラクタル
敵対的例題





Poole et al (2016)
Random deep neural networks





Natural Gradient

$$\max \qquad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta})$$

$$\left|d\boldsymbol{\theta}\right|^2 = \varepsilon \qquad \mathrm{KL}[\mathrm{p}(\mathrm{x},\theta):\mathrm{p}(\mathrm{x},\theta + d\theta)] = \varepsilon$$

$$\tilde{\nabla} l = G^{-1}(\boldsymbol{\theta})\nabla l$$

$$\Delta\boldsymbol{\theta}_t = -\eta_t \tilde{\nabla} l(x_t, y_t; \theta_t)$$

Fisher information

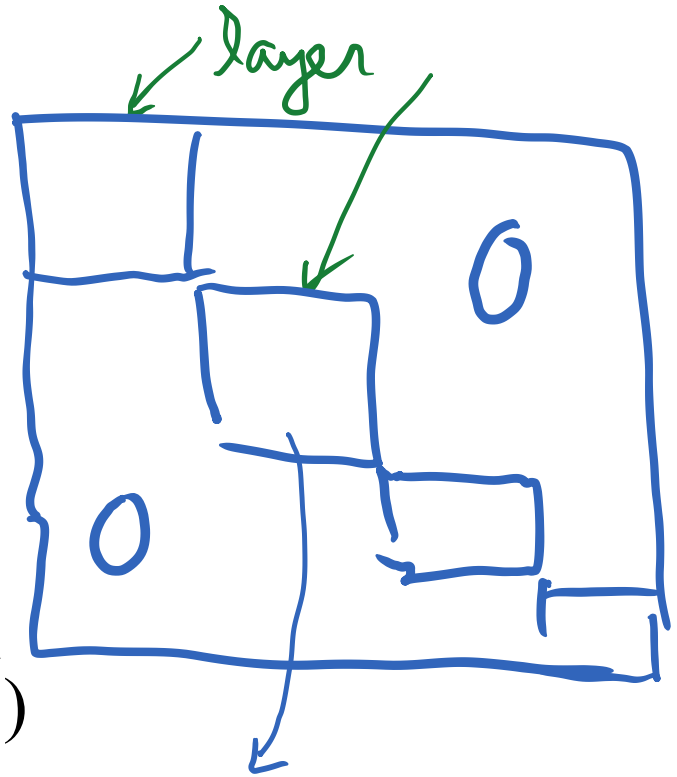
$$G = E_x\left[\frac{\partial \varphi}{\partial W_m}\frac{\partial \varphi}{\partial W_l}\right]$$

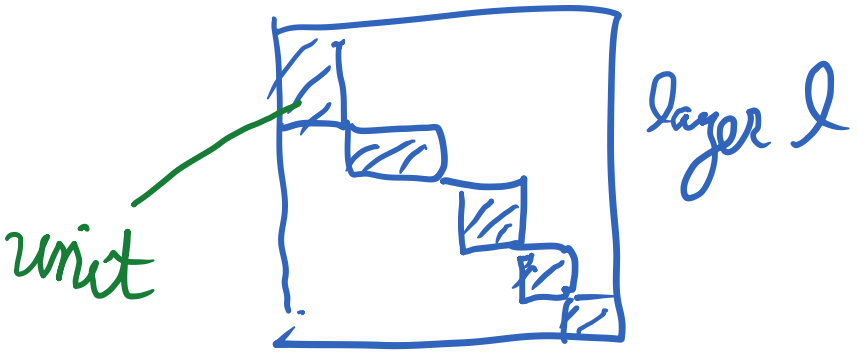
$$\frac{\partial \varphi^l}{\partial W_m} = \varphi' W \frac{\partial \varphi^{l-1}}{\partial W_m} = B\frac{\partial \varphi^{l-1}}{\partial W_m} = BB\ldots B\frac{\partial \varphi^{m+1}}{\partial W_m}$$

$$G(W_l, W_m) = \prod \chi_1 E_x\left[\varphi'\left(\overset{l}{\boldsymbol{w}}_i\right)^2 \overset{l-1}{\boldsymbol{x}}\ \overset{l-1}{\boldsymbol{x}}\right] + O_p(1/\sqrt{n})$$

$$G(W_l, W_m) = 0 \sim O_p(1/\sqrt{n}), \qquad l \neq m$$

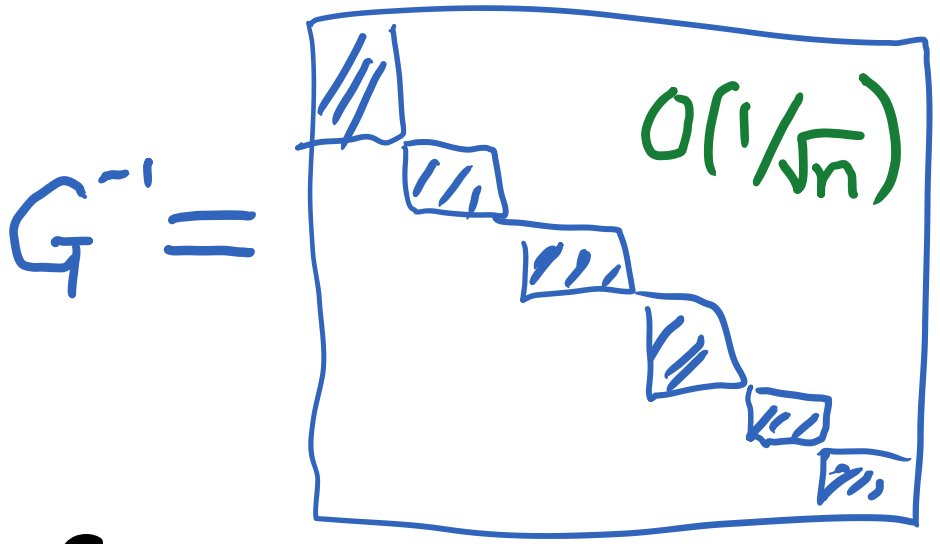
$$G\left(\overset{l}{\boldsymbol{w}}_i, \overset{l}{\boldsymbol{w}}_j\right) = 0 \sim O_p(1/\sqrt{n}), \quad i \neq j$$



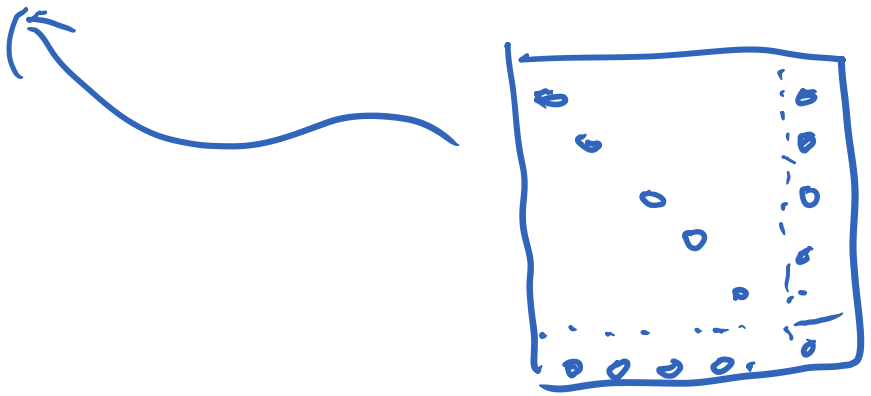


Unitwise natural gradient

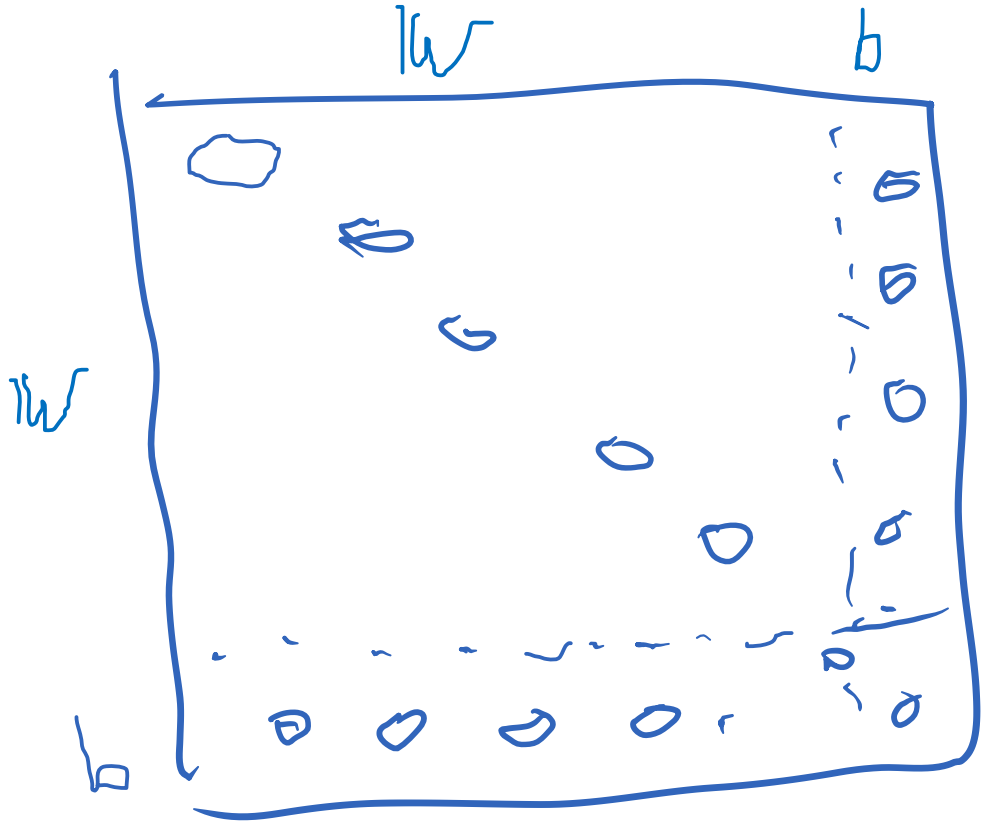
$$\Delta W = -\eta G^{-1} \nabla_W l$$



Y. Ollivier; Marceau-Caron



G=



+ww

Ollivier、Marceau-Caron：quasi-diagonal natural gradient

No simulations: Oh No!!

Natural gradient is no more dream!

TANGO: Ollivier recursive formula for $G^{-1}\nabla l$

Neural tangent kernel

数理脳科学は脳の基本原理を探求する

単純な基本モデル用いる：数理的探索（現実とは違う）

→　計算論的神経科学
　　（脳はいかにこの原理を実現したか）

→　AI：技術による原理の実現　（脳とは違う）

脳は基本原理をどう実現したか

進化によるランダムサーチ
使える材料の制約
歴史的な制約

ごたごたの設計の中で精妙な実現：超複雑

人工知能は何をどう実現するか？

人工知能は脳に何を学ぶのか： 心
意識と無意識のダイナミックス

記号 --- 興奮パターン
論理的推論 --- 並列ダイナミックス

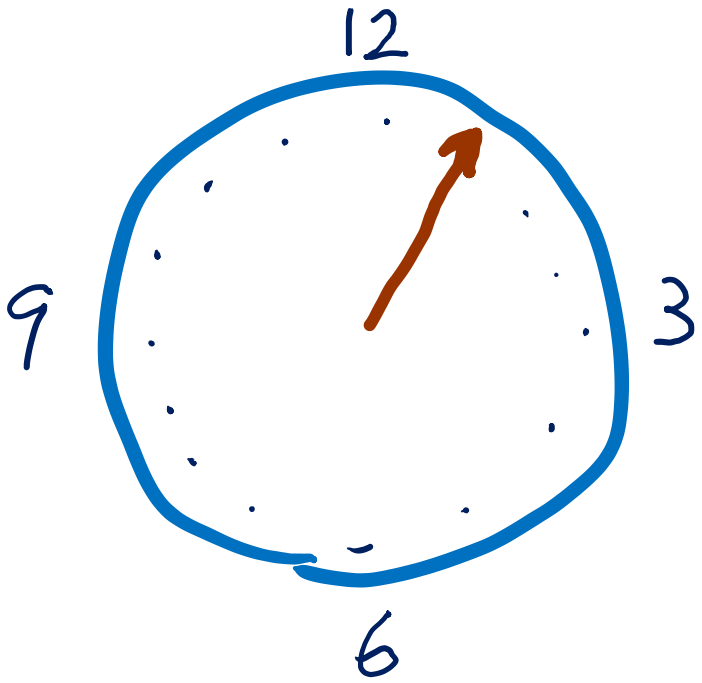
AI NN

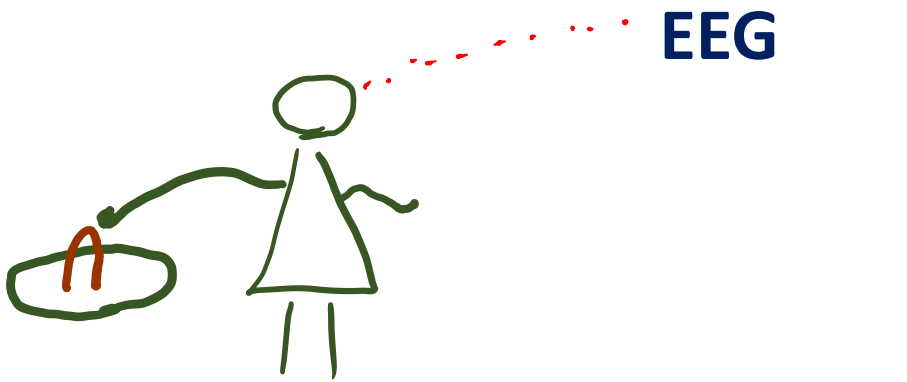
意識の発生

共同作業、自分の意図を自分で知る

言語：　論理的思考、数学

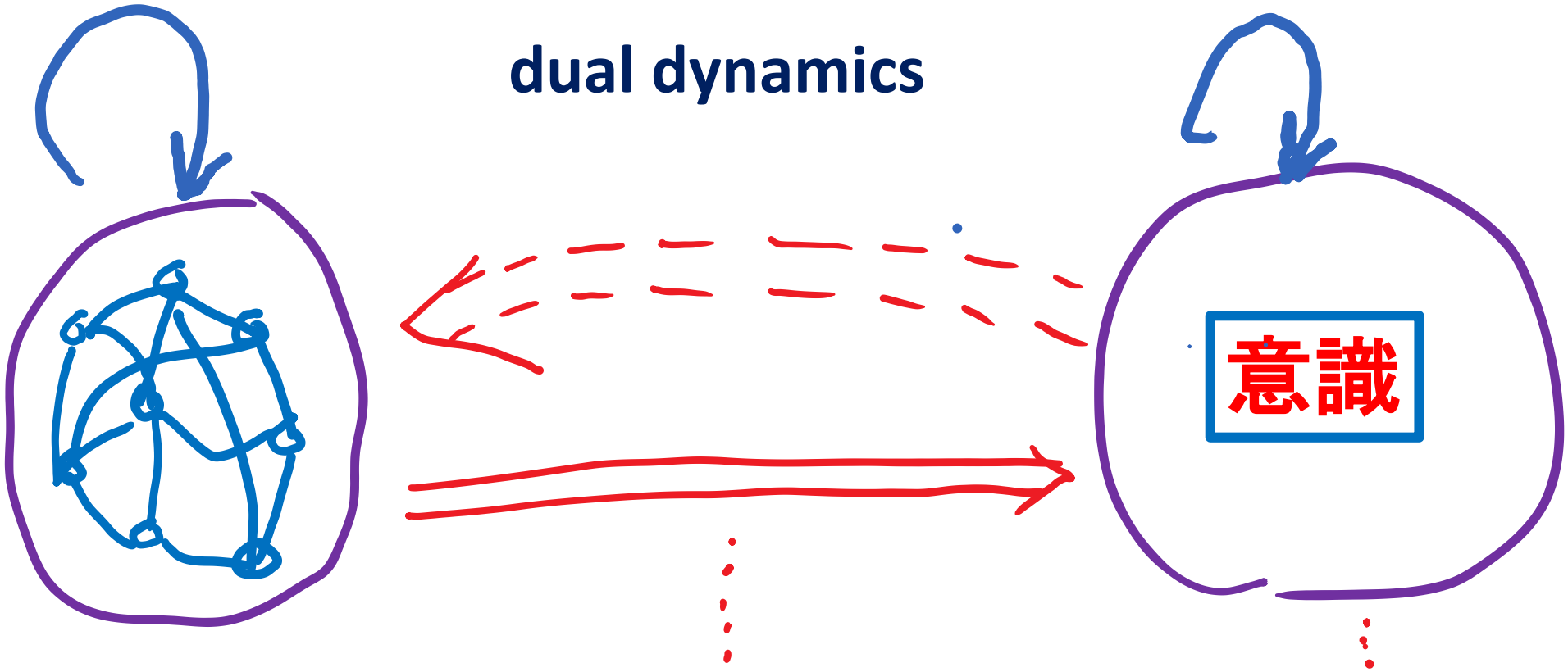
Libet の実験：自由意志





When!

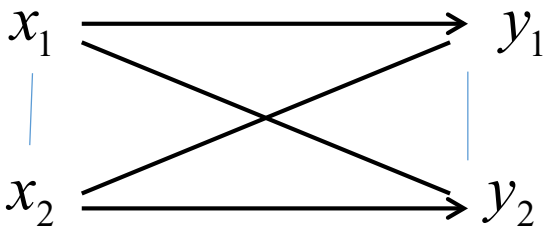
予測（先付け）と後付け　Prediction and Postdiction

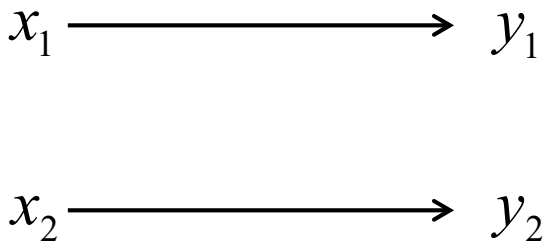


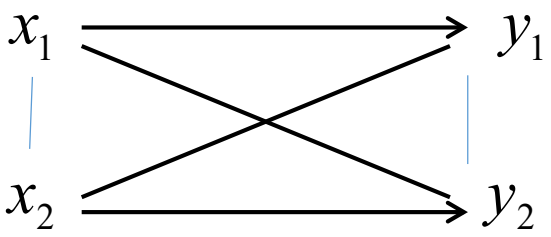
ダイナミックス　　意思決定と行動　　反省、正当化、論理

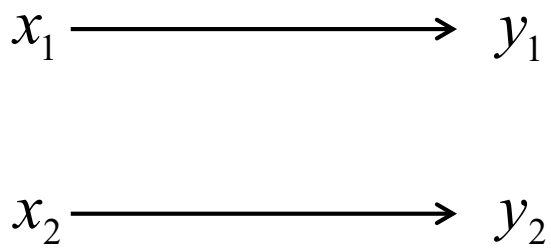
意識の情報統合理論　Tononi　　IIT

システムの情報理論――情報幾何









full model: $S_F = \{p(\mathbf{x},\mathbf{y})\}$ **Disconnected model:**

$$S_{dis} = \{q(\mathbf{x},\mathbf{y})\} \quad q(\mathbf{y}\,|\,\mathbf{x}) = \Pi q(y_i\,|\,x_i)$$

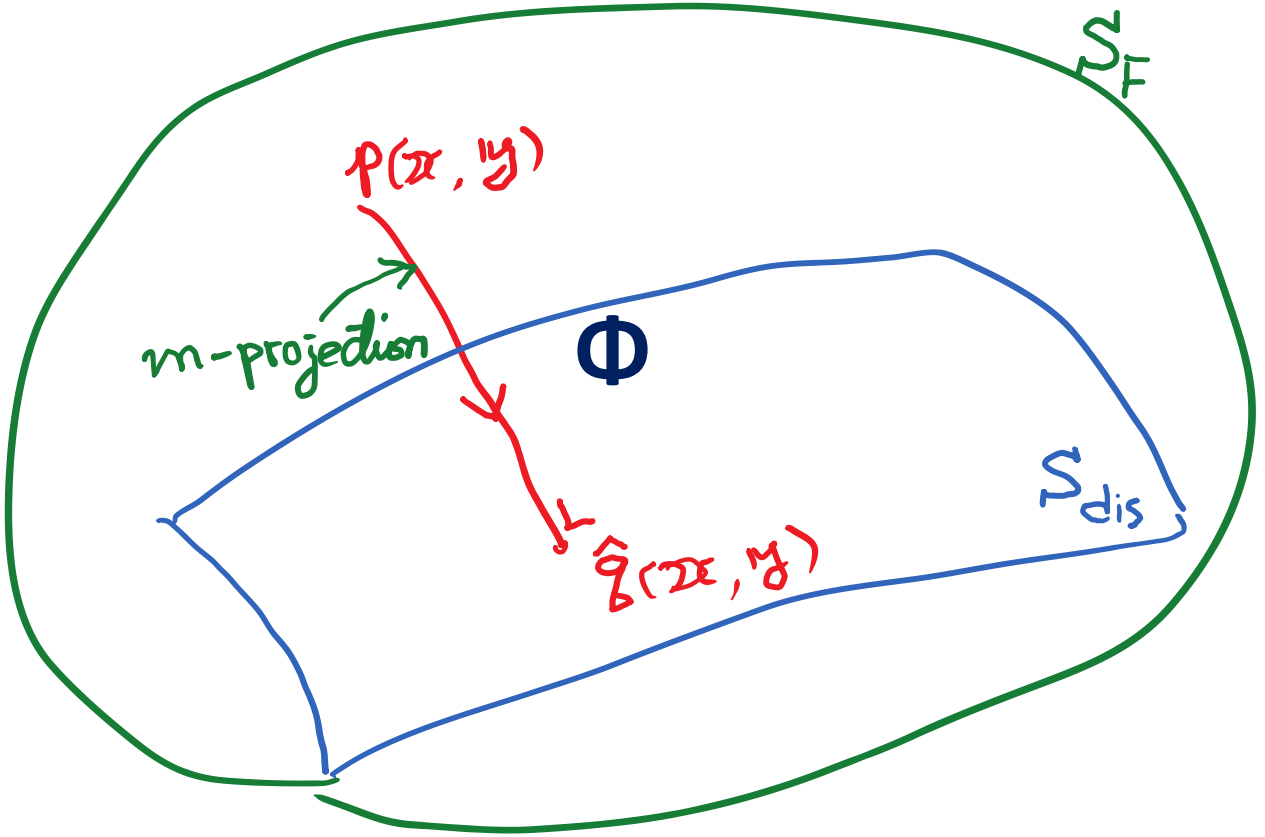
measure of interaction : N. Ay

information integration : Tononi

Barrett and Seth

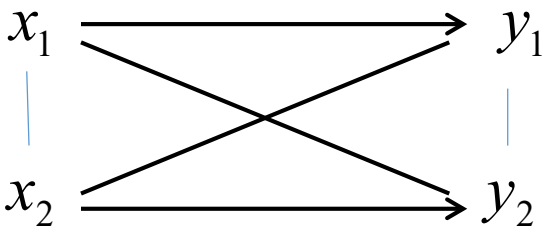
Many other Φ

Integrated Information Theory G. Tononi



Necessary condition; sufficient?

Markov Condition



(1→2) branch deleted: Markov condition: $x_1 \to x_2 \to y_2$

$$p(x_1, y_2 \mid x_2) = p(x_1 \mid x_2)\, p(y_2 \mid x_2)$$

$X_1 - X_2 - Y_2$

$X_2 - X_1 - Y_1$

$$S_{dis}: \quad \text{all } x_i \to y_j \quad (i \neq j) \quad \text{deleted}$$

人工知能が脳に学ぶべきこと: 数理的理解

判断;制御;認知;記憶

意識と心の役割; 後付け

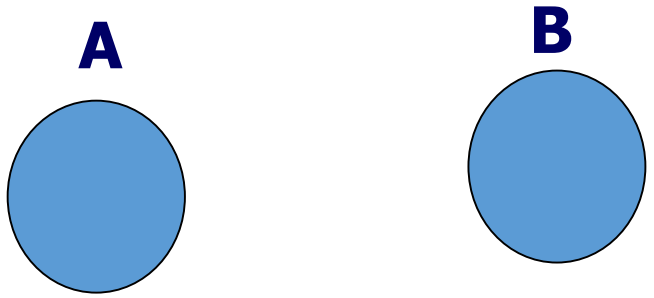
連想式記憶システム:知識体系

心の理論

葛藤する心

究極のゲーム
（ultimatum game）

10万円



A:　配分を決定する----　7万円–3万円
B:　同意または拒否

この時の脳の働き、各領野の確執
利益、公正、その後のこと、評判
二人か社会ゲームか

心を持ったロボットがつくれるか？

人の心の動きを理解する

ロボットが心を持つように見える
（感情移入）

ロボットが心を持てるのか？

人間の心：進化の産物

意識、意図、論理、感情

種の生存と個の不合理　：使命感
人間は不合理；　芸術、喜び、愛、苦悩

ただ一度の、かけがいのない人生

ロボット（金融システム）は合理的

人工知能と倫理

人工知能の安全性、制御可能性

人工知能と戦争；人工知能の金融支配；
支配の道具、格差

暴走：　人間の暴走を範として

社会への影響：技術は止まらない：制御できるか

失業問題：人口減　：AIは仕事を奪うか？　より高度な仕事

格差の拡大：

　　ベーシックインカムと人類の家畜化：働く喜び

人工知能と技術的特異点　2045

人工知能が人間を超えるとき
　人工知能が研究し、技術を進める

　人間は素晴らしいが、愚かである。

人間はどんな知能システムを作るのか？
　　社会の進化と支配

人工知能と未来社会の設計

深層学習を超えて

科学研究、技術開発

社会、文明　その脆弱性・崩壊

我々はは何をなすべきか？

日本のAIの進むべき道：政府の戦略

ブームは終わる

超大国　　文化国家

物量作戦はだめ

理論とアイデア

中小企業を含む現場との交流；産業の情報化