

工学アカデミー：AIと倫理

AIの歴史、現状、 未来文明の展望

帝京大学先端総合研究機構特任教授

甘利俊一

二つの知能システム

自然知能と人工知能：脳とAI

脳は進化の結果出来上がった

——進化の宇宙誌

AIは脳にヒントを得て技術で実現した

ビッグバン (138億年前)

物質の法則

エネルギー・物質 — 天体 — 分子(秩序)

物理学、化学

生命の誕生（38億年前）

進化の法則

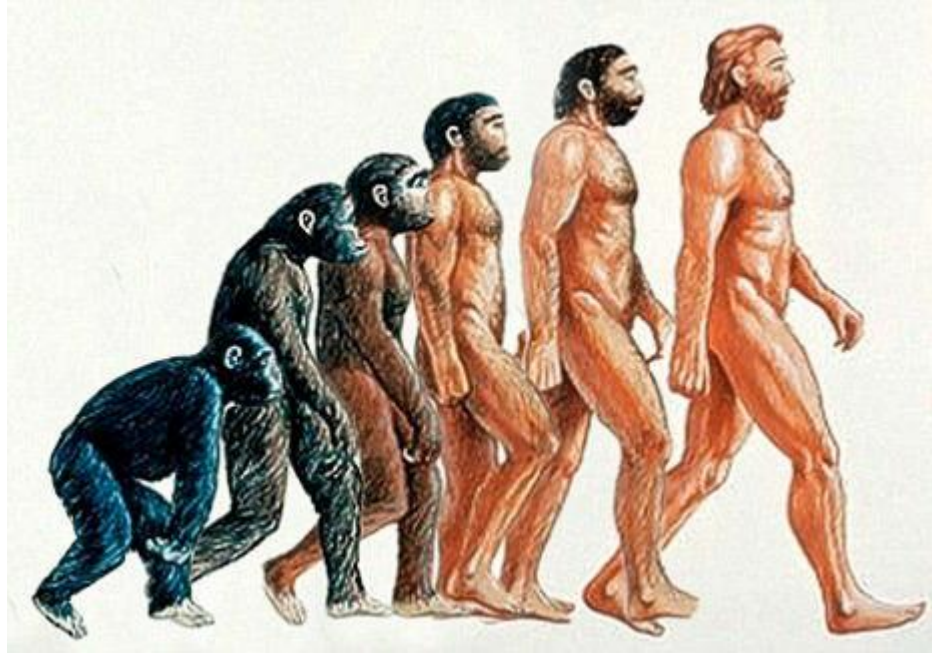
生命 = 情報 + 物質

= 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

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)

人類の進化

社会生活

心、意識、感情
共感

言語と論理、知的機能

ロボットに意識は持てるか？



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：

AI時代の社会・文明

人工知能と脳のモデル：一歴史

第一次ブーム

1956~

AI

Dartmouth 会議
記号と論理

脳モデル

Perceptron

学習する普遍計算機構

暗黒期 (1965後半~1970's)

第二次ブーム (1970~1980) コネクショニズム

多層回路の誤差逆伝搬学習、連想記憶、その他

日本は先進国だった

暗黒期 (1965後半~1970's)

1960, 70年代

日本は神経回路網モデル・理論の先進国だった

- 1967 多層回路網の確率的勾配降下法 (Rumerhaltら) 1976
- 1971 ランダム回路のダイナミクス 多安定、振動 (Wilson-Cowan)
- 1977 特徴検出細胞の自己組織化
(抑制性シナプスの学習; BMC 1981)
- 1972 連想記憶モデル (Hopfield, 1982)
- 1977 神経場のダイナミクス

理論と実験の交流—伊藤正男

第3次ブーム

2010~

脳型の人工知能

深層学習 Deep learning

畳み込み多層回路（福島） + 確率勾配降下（甘利）：

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

大規模言語モデル：chat GPT

記号と論理 VS パターンとダイナミックス

現在のAI

深層神経回路—学習

強化学習—囲碁、 α ホールド

生成AI 拡散モデル

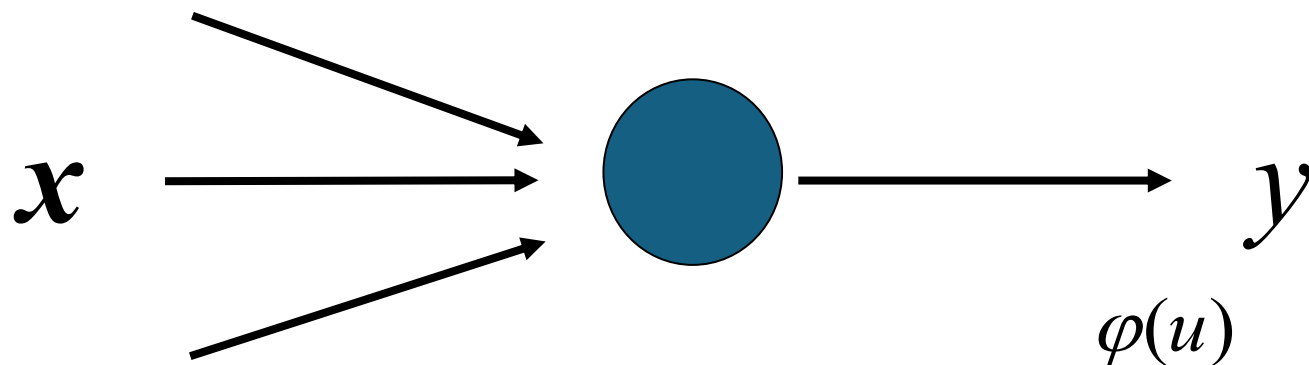
LLM 単語のベクトル表現 word2vec

transformer—抽象表現—autoencoder

attention -- 連想記憶、論理的推論

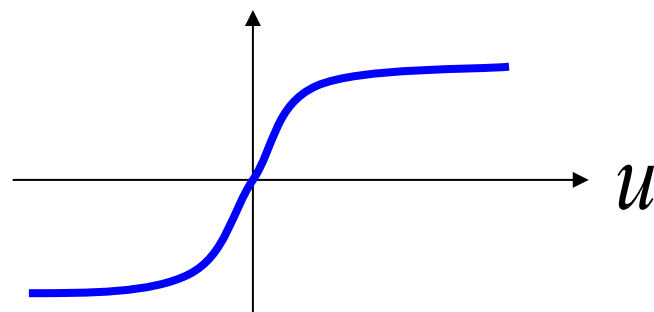
ニューロンの数理モデル

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

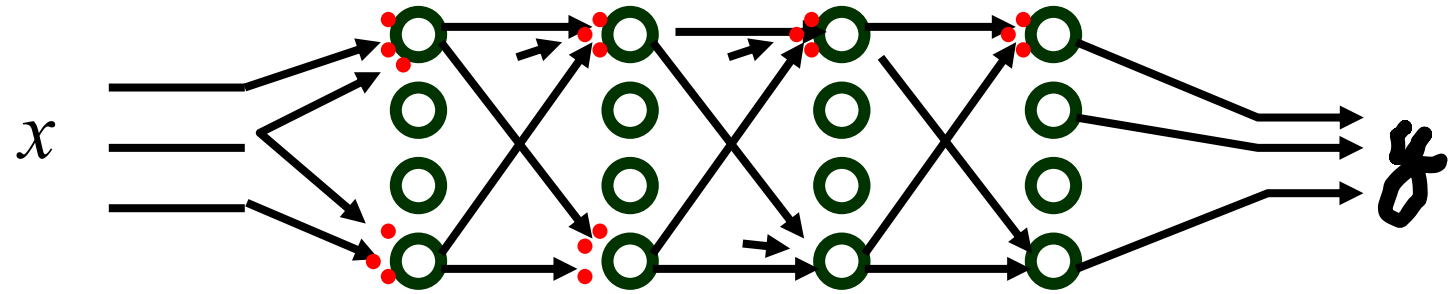


多数決素子

興奮・抑制



深層回路の万能性



$$y = f(x, \theta)$$

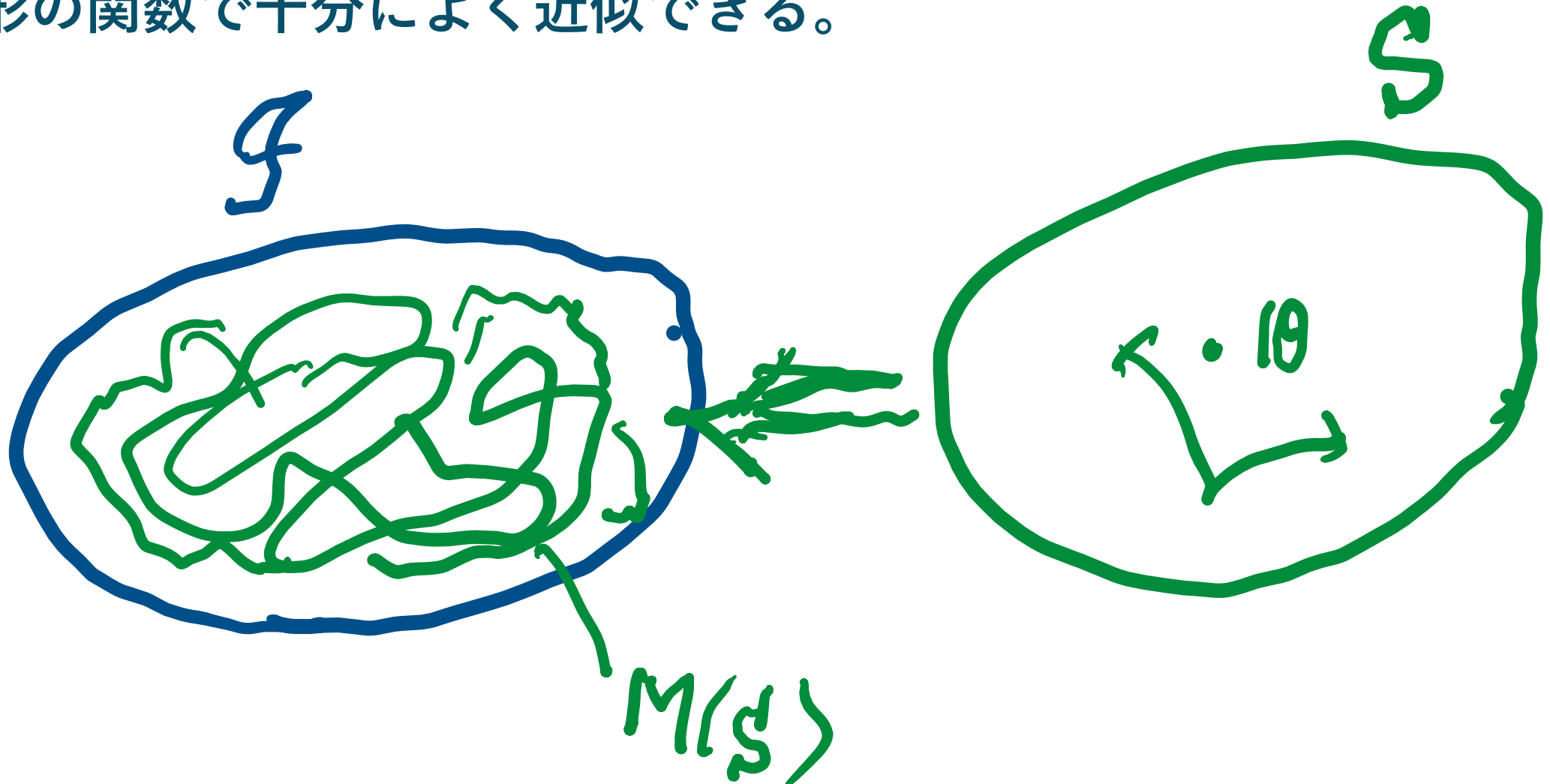
定理：任意の $f(x)$

は、うまい θ を見つけて、 $f(x, \theta)$ で十分よく近似

ニューラルネットの万能性

： 普遍近似定理

どんな関数 $g(x)$ でも、 P が十分に大きければ、 $f(x, \theta)$ の形の関数で十分によく近似できる。

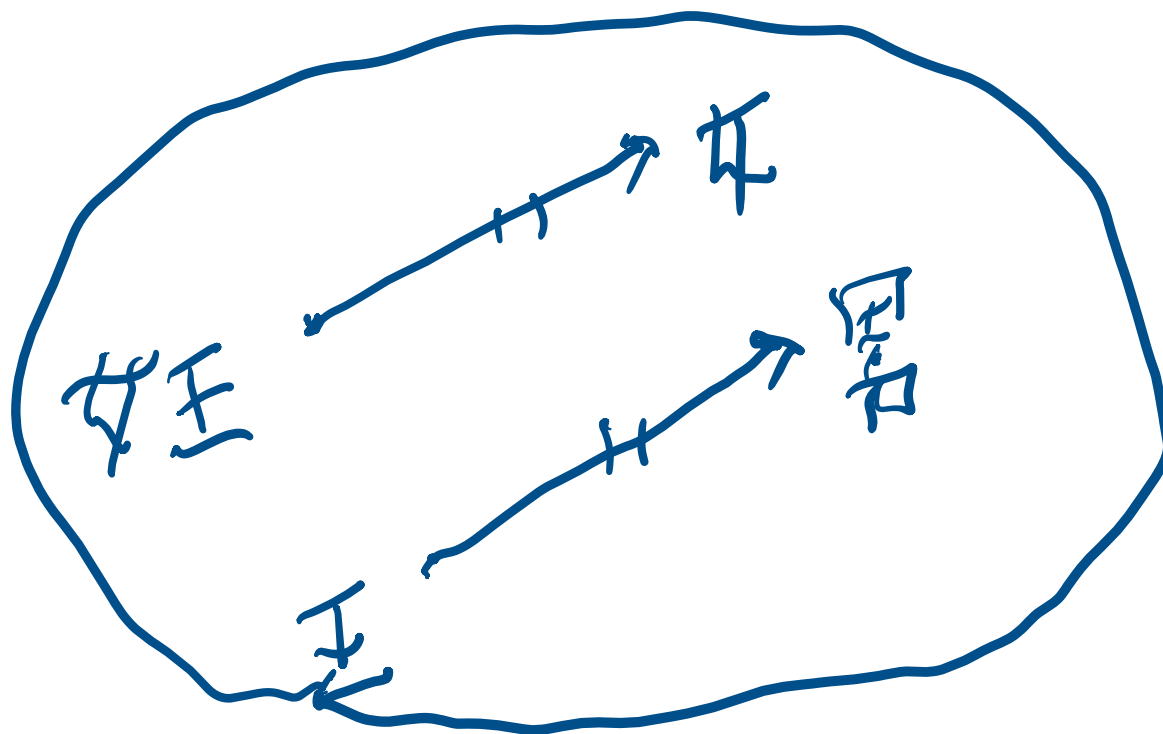


単語のベクトル表現：共起性

人間—ベクトル x

$$x = (x_1, x_2, \dots, x_n)$$

女王—女性 = 王 = 男性



共起性

$$x = (x_1, x_2, \dots, x_n)$$



TransformerとAttention

抽象表現と連想記憶

$$x_1 \rightarrow y_1$$

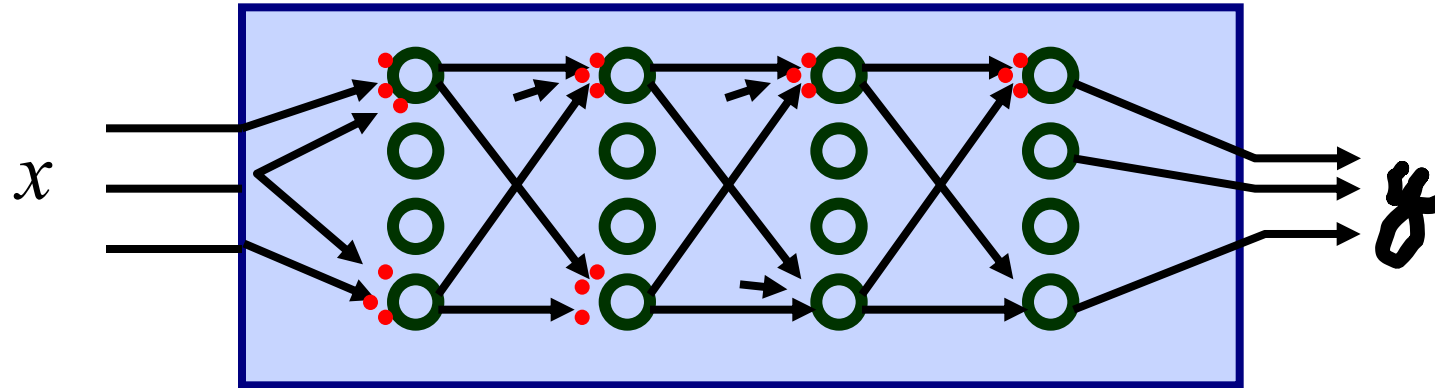
$$x_2 \rightarrow y_2$$

.....

$$x_n \rightarrow y_n$$

層状学習回路網と確率勾配降下学習法

multilayer perceptron



パーセプトロン Perceptron

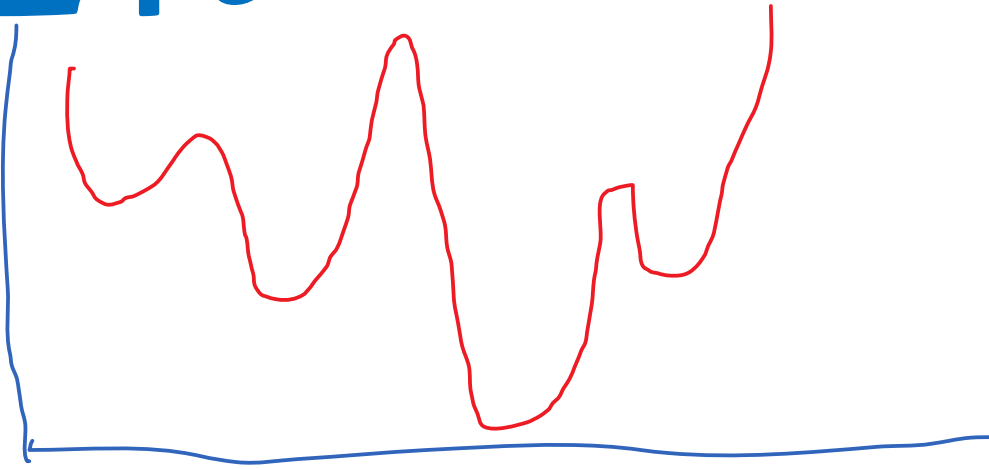
バックプロパゲーション Backpropagation

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

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

深層学習：理論の遅れ

パラメータ数 巨大：創発
大きいことはいいことだ！



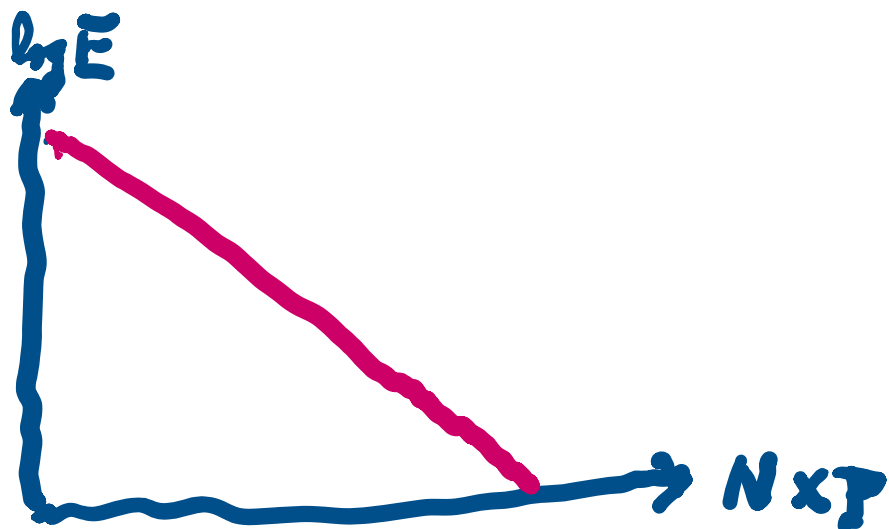
情報表現の獲得

大域解と局所解：高次元

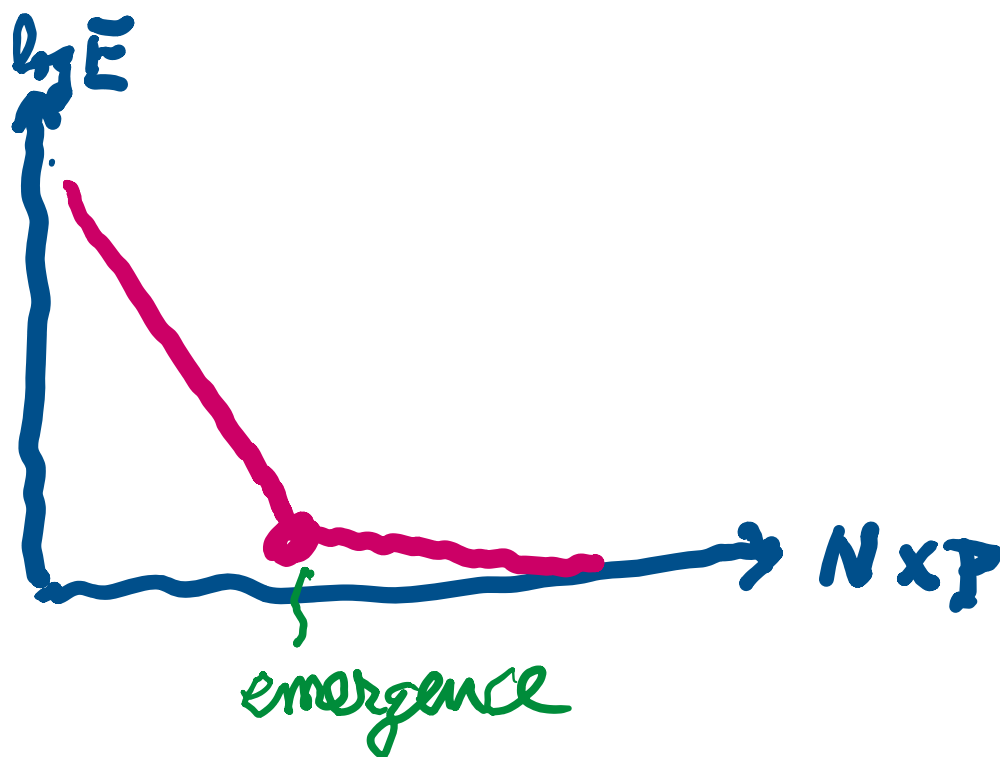
訓練誤差と汎化誤差—二重降下

經驗則

Scaling law



Emergence
創発



学習の理論的課題：汎化誤差 $y = f(x, \theta) + \varepsilon$

N: 例題数

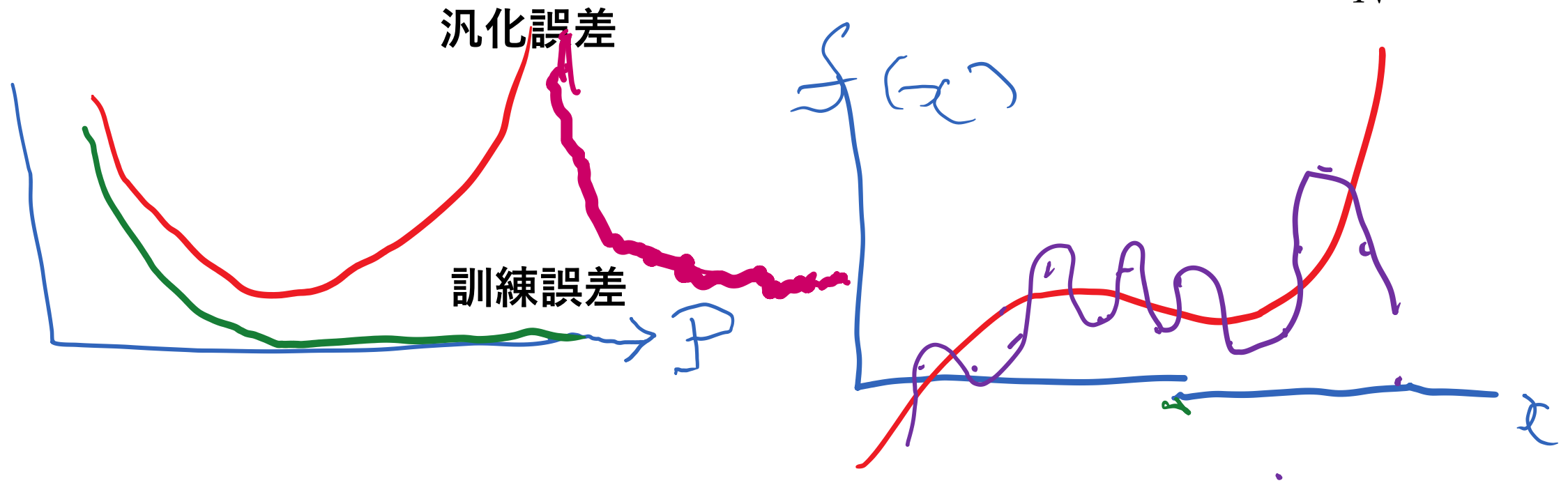
P: モデルのパラメータ数

過学習

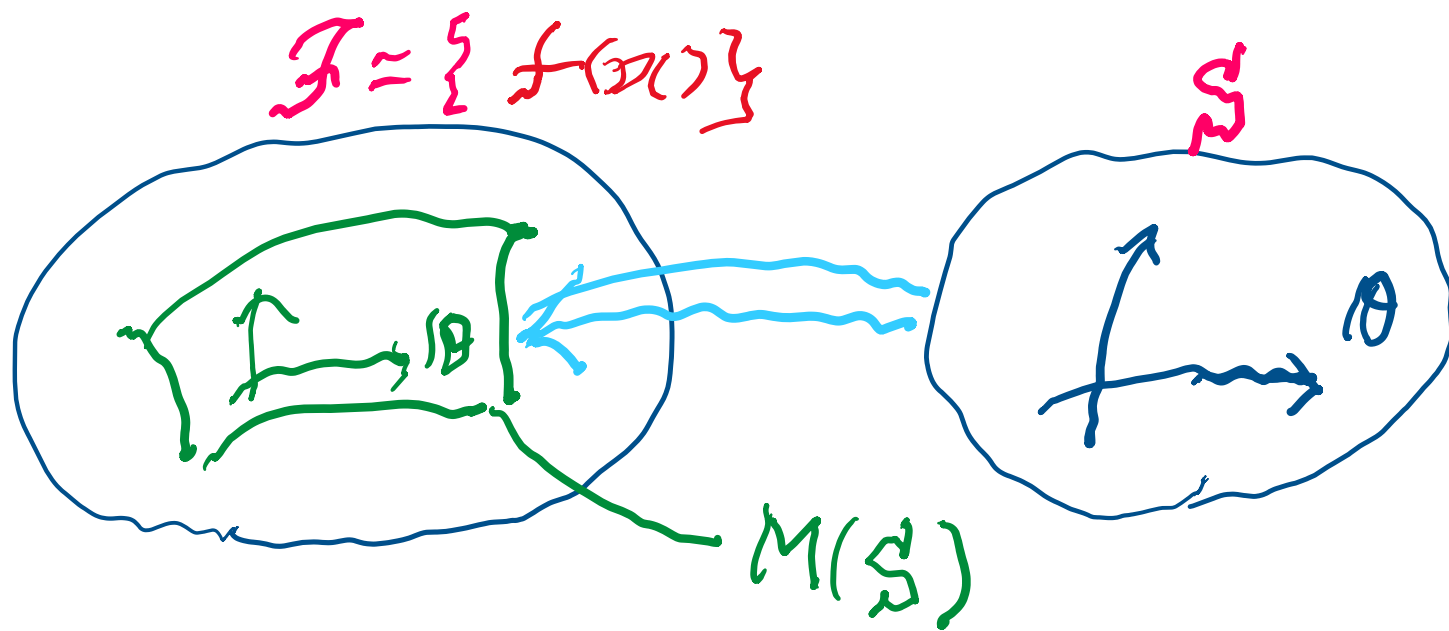
$$L_{emp} = \frac{1}{N} \sum |y_i - f(x_i, \theta)|^2$$

$$L_{gen} = E[|y - f(x, \theta)|^2]$$

$$L_{gen} \approx L_{emp} + \frac{P}{N}$$



パラメータ空間 $S = \{ \theta \}$ と関数空間 $\mathcal{F} = \{ f(x) \}$



$$\dot{f}(x, \theta) \approx \frac{\partial f}{\partial \theta} \cdot \dot{\theta}$$

$$\dot{\theta} = -\eta \frac{\partial l(x, y, \theta)}{\partial \theta}$$

脳は情報の基本原理を 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？

数理脳科学：脳の基本原理の探求

単純な基本モデルを用いる：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）

数理脳科学と人工知能

天空の科学か地上の科学か

脳の仕組みの解明

人工知能

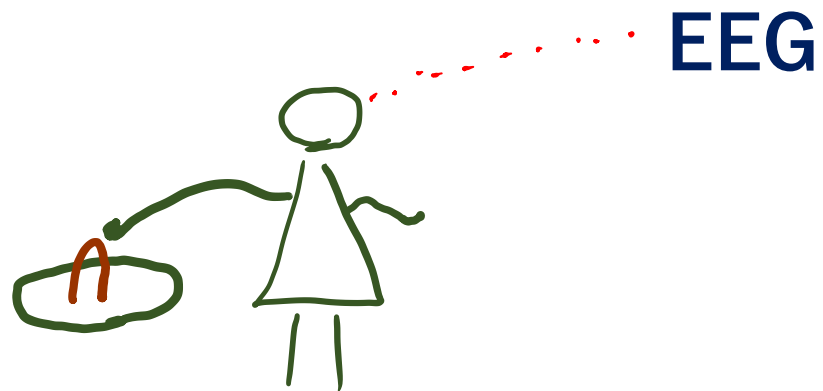
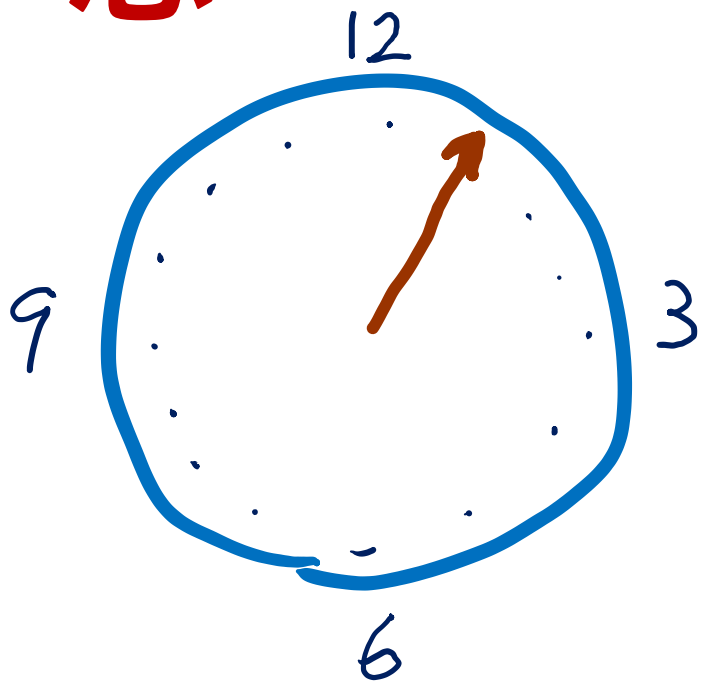
意識の発生：心

共同作業、自分の意図を自分で知る

社会と進化：共感、喜び、悲しみ、意欲、
好奇心、心の理論

世界観：自己の位置づけ

Libet の実験：自由意志

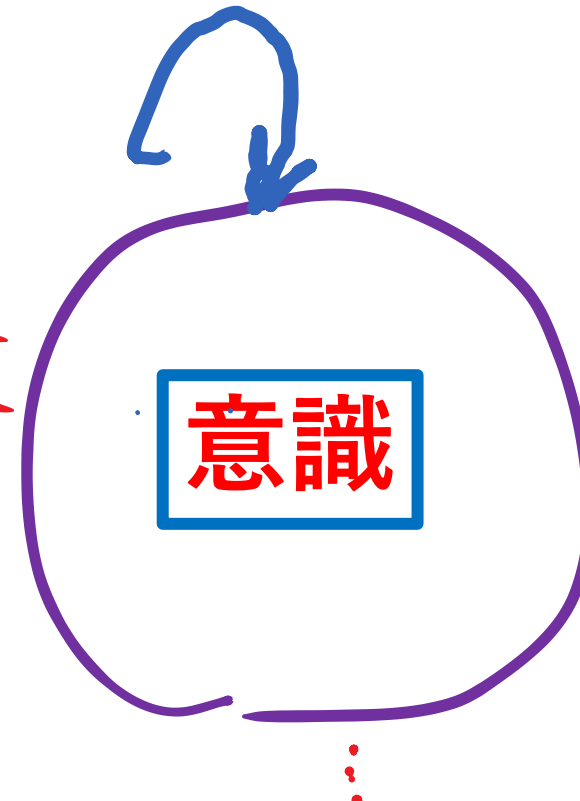
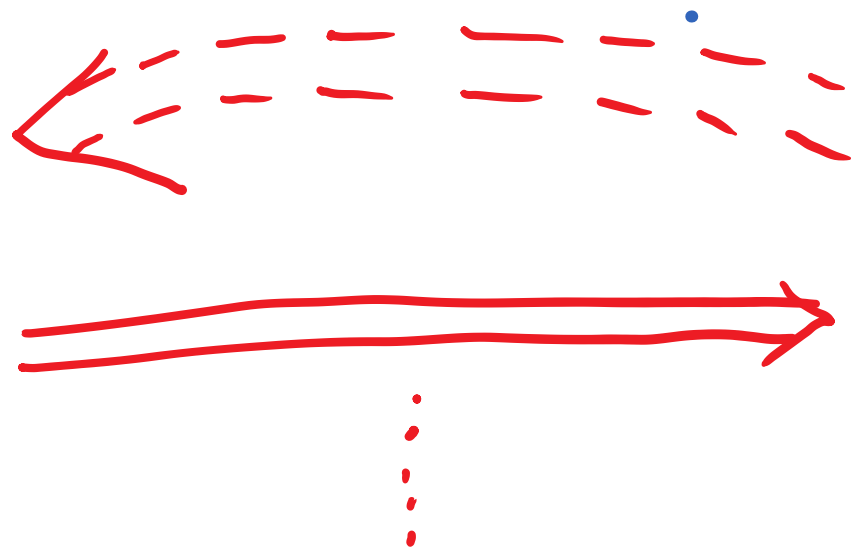
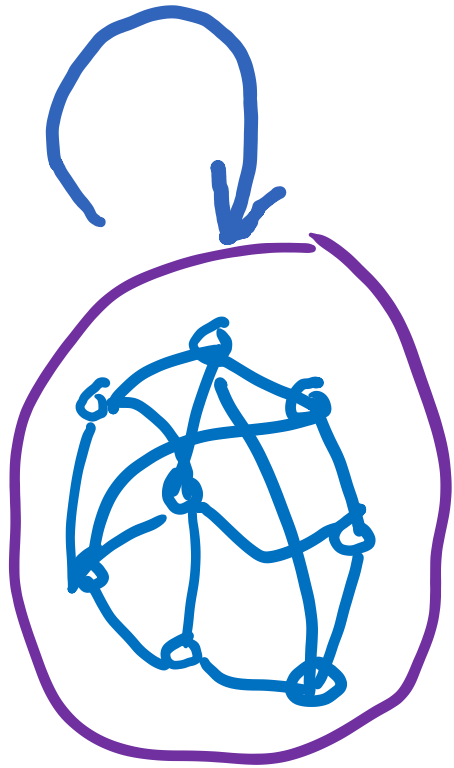


When!

予測（先付け）と後付け

Prediction and Postdiction

dual dynamics



ダイナミクス

意思決定と行動

反省、正当化、論理

ロボットは意識と心を持てるのか？

人間の心：社会における進化の産物

意識、意図、論理、感情、使命感、好奇心
世界観と自己、

種の生存と個の不合理的

人間は不合理的； 喜び、悲哀。愛、苦悩、なまけ、努力

ただ一度の、かけがいのない人生

ロボットは合理的

意識とポストディクシオン

深層学習 大規模言語モデル

意識を持つ機械 : postdiction ロボット

2重のダイナミックス、意識の介入

質問： ChatGPTは意識を持っているのでしょうか？ 論理的な思考はするのですか。

回答：

いいえ、ChatGPTは意識を持っていません。ChatGPTは大規模なデータセットから学習したパターンを基にして、ユーザーの質問や指示に対して文章を生成するAIです。感情や自我を持つことなく、単にプログラムされたアルゴリズムに従って動作しています。

論理的な思考に関しては、ChatGPTは学習データに基づいて一定の論理的なパターンを再現することができますが、それはあくまで統計的な予測に過ぎません。人間のように深い理解や意識を伴った論理的思考ではありません。

脳とAI

補完、補外、連想、論理推論

理解とは一腑に落ちる

天体の運動と予測

Keplerの法則

Newton 力学、

新しい概念 質量、加速度、引力

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
軍事利用

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：仕事＝遊び

アマチュア＝プロ

人工知能と未来社会の設計

自由、平等、協調

深層学習を超えて：人工知能
社会、文明 その脆弱性・崩壊

我々は何をなすべきか？

新しい時代に；全人的な活動

科学する、芸術する、スポーツする：鑑賞する

遊び心が重要、まず遊べ

生を謳歌する

自己の可能性を開花させる

楽観的なシナリオ

AIが生産を担う
十分なベーシックインカム

人類は遊んで暮らす、幸福か？
人類の家畜化

人間：働くこと、遊ぶこと；意欲と達成感
仕事と遊びの一体化

アマチュア科学者、アマチュア農業、
アマチュア園芸家、アマチュアスポーツマン、アマチュア芸術家

そこへ至る苦難の道——文明崩壊

AIの軍事利用

AIの覇権争いと格差の増大

AIによる人類の思考力減退

AIの悪用とポピュリズム

心を持ったAI

教育の重要性

意欲、探求心、思いやり、達成感

働く喜び = 遊ぶ喜び

アマチュア科学者

アマチュア芸術家

農芸、スポーツマン

AI時代の教育：知識の伝達ではない

AIを使いこなす人

AIに使いこなされる人——思考力の減退

教師の生きざま：仲間との連帯

ルイ アラゴン；

教えるとは未来を語ること、学ぶとは希望を胸に刻むこと