11/3/2018 日本セキュリティマネジメント学会 学術講演会

PFN Fellow 丸山宏
Twitter: @maruyama

統計的機械学習における セキュリティマネジメント


全自動お片付けロボット
Autonomous Tidying-up Robot
Preferred Networks

自己紹介

- XML暗号化(W3C)、DOM-Hash(IETF)、Webサービスセキュリティ(OASIS)等の標準化文書著者
- IBMビジネスコンサルティングサービス、セキュリティ・コンサルタント（大木栄二郎先生に師事）
- IBM Research セキュリティ分野サブストラトジスト
- 統計数理研究所情報セキュリティ内部監査人
- Project ICHIGAN セキュリティ・アーキテクト
- 情報・システム研究機構「システムズ・レジリエンス」プロジェクト総括
- :

Preferred Networks

Agenda

1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学習システムとリスク
5. まとめ

Agenda

1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学習システムとリスク
5. まとめ

研究分野としての人工知能のフォーカスは時代と共に変化

1956-1974 第1次人工知能ブーム
- 記号処理 (LISP)
- Means-End Analysis
- 自然言語処理

→
- 動的メモリ管理
- 探索アルゴリズム
- 形式言語理論
- ：

1980-1987 第2次人工知能ブーム
- 知識表現 (e.g. フレーム)
- エキスパートシステム
- オントロジー

→
- オブジェクト指向
- モデリング言語
- セマンティックWeb
- ：

2008 第3次人工知能ブーム
- 統計的機械学習・深層学習

→ **帰納的システム開発 (機械学習工学)**

成熟すると「当たり前の技術」に．．

「人工知能」の同床異夢

1. 研究者にとっては…知性を模倣することで、知性を理解しようとする営み、あるいはその研究領域
   - 探索（乗換案内、ゲーム等）、推論（定理証明等）、最適化、認識（画像、音声、…）、自然言語処理、
2. メーカーの思惑は…ブランド戦略「我が社の技術は先進的」
   - 上記の技術のいずれかが部分的にでも使われていれば「AI」
   - Watson, Zinrai, H, …, AIスピーカー, AI電子レンジ, …
3. 一般人からみると…擬人的な機械「なんかすごい、でも怖い」
   - 鉄腕アトム、HAL9000、ターミネーター、…
4. マスコミ…危機感を煽るテーマ
   - シンギュラリティ、人工知能に奪われる仕事、…

「ポエムなAI」*

*栄藤稔, http://webronza.asahi.com/business/articles/2017101600002.html

「擬人化されたAI」のイメージには要注意

総務省「未来をつかむTECH戦略」
http://www.soumu.go.jp/main_content/000548068.pdf

NHKニュース、3/31/2017,
https://www3.nhk.or.jp/news/business_tokushu/2017_0331.html

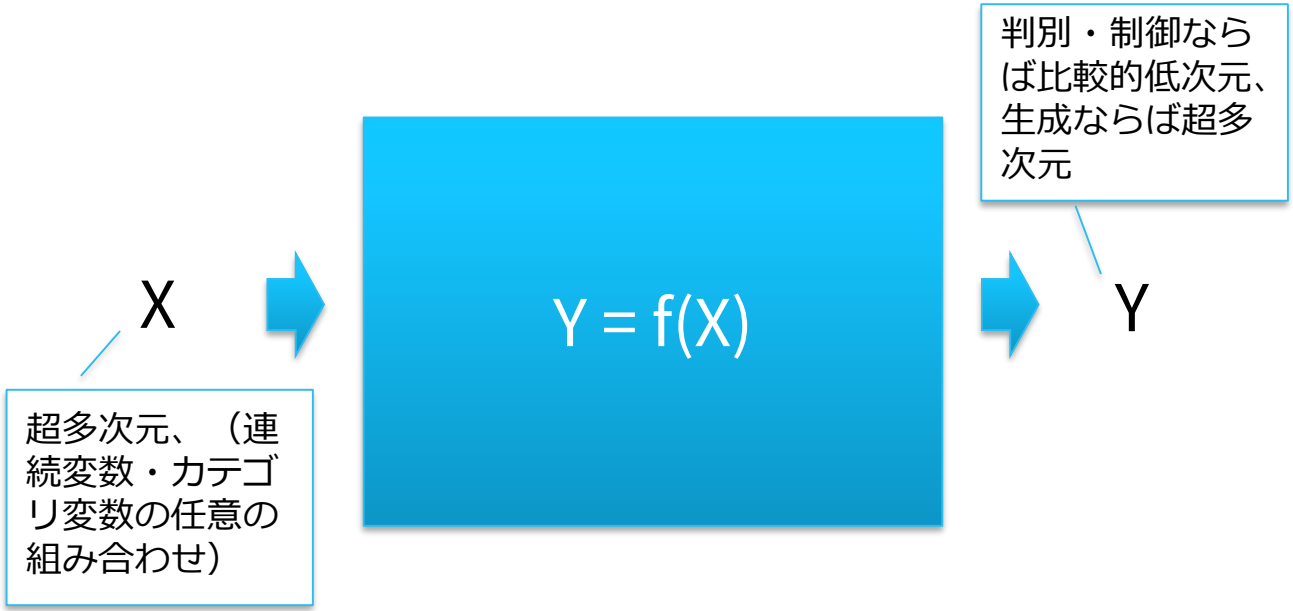
SFの話題であり、現在見えている技術とは分けて議論すべき

本日はAIの話はしません
統計的機械学習の話をします

Agenda

1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学習システムとリスク
5. まとめ

深層学習とは何か – (状態を持たない*)関数



*推論時にパラメタ更新を行うオンライン学習についてはこの限りでない

普通の(演繹的な)関数の作り方: 例 摂氏から華氏への変換

仕様

入力: C
出力: F
ただし、FはCを華氏で表したもの

モデル

F = 1.8 * C + 32

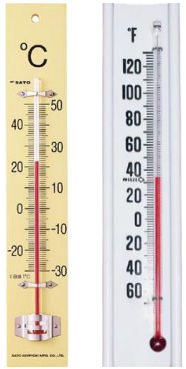
アルゴリズム

```
double c2f(double c) {
  return 1.8*c + 32.0;
}
```

人が持つ
先験的知
識

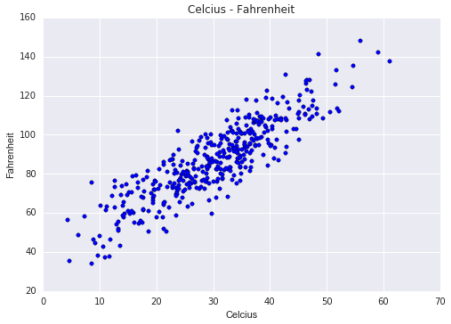
モデルが既知／アルゴリズムが構成可能である必要

深層学習の(帰納的な)やり方 – 訓練データで入出力を例示



観測

訓練データセット

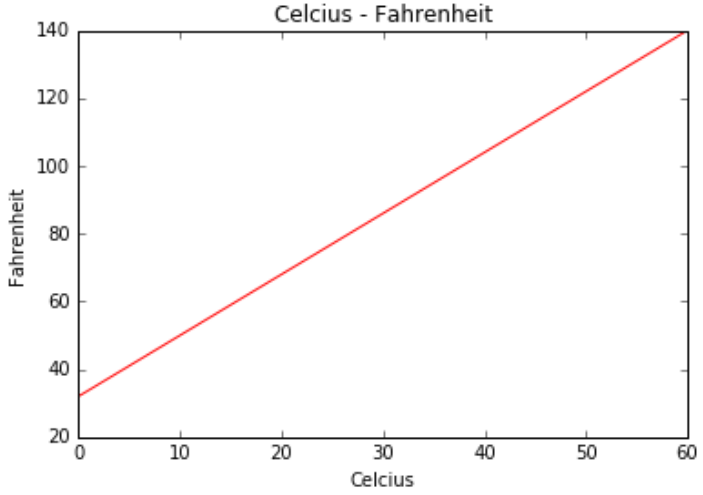


訓練（ほぼ自動でパラメタθを決定）

X → Y = f(X,θ) → Y

機械学習（＝統計モデリング）すると...

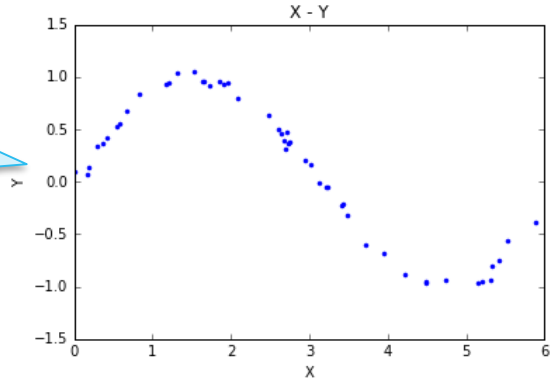
モデル： 線形回帰式



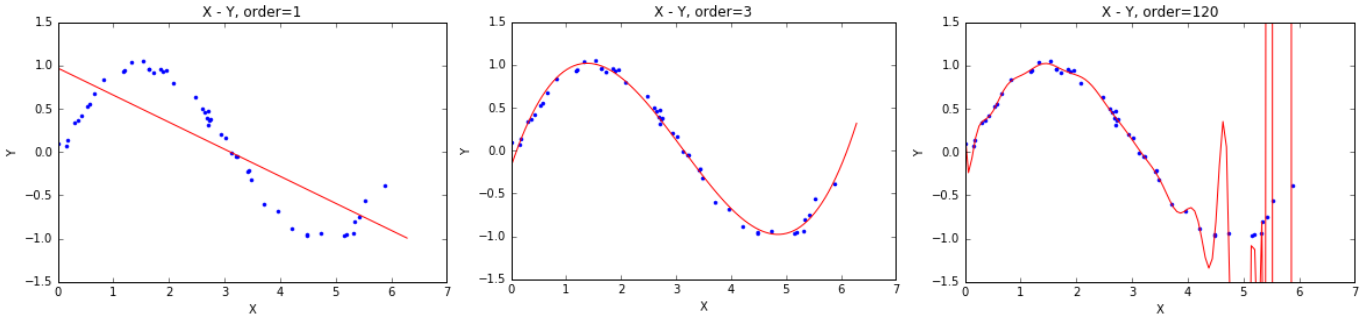
モデル・アルゴリズムが未知でよい

統計モデリングにおける、モデル選択の重要性

この訓練データをよく再現するモデルは？



パラメタが多すぎると過学習に

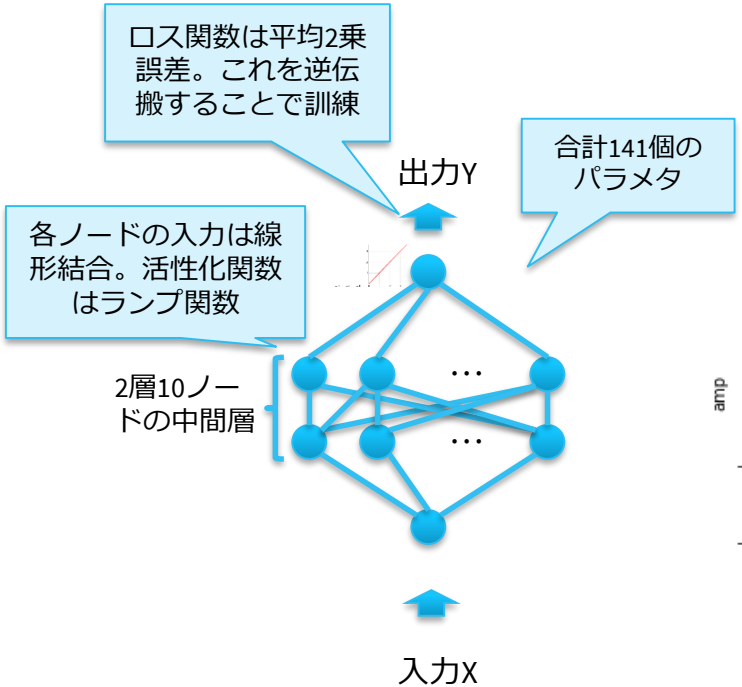


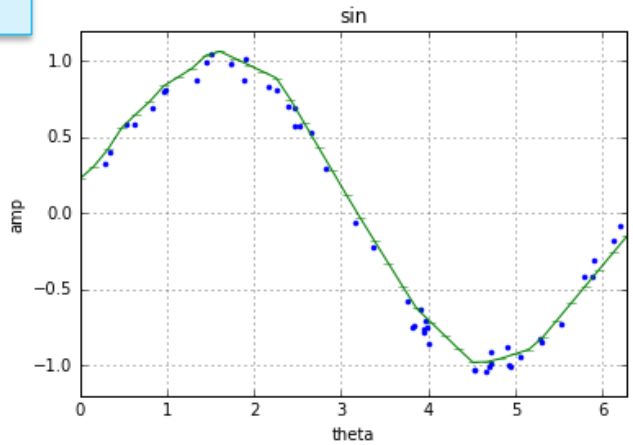
過学習に陥らずに、適切なモデルを選ぶには？

訓練はどのように動くか

https://code.facebook.com/pages/1902086376686983

先ほどの例を深層学習で訓練してみると...





程よい近似

モデルの形によらず（あまり）過学習しない（汎化性能が高い）！！

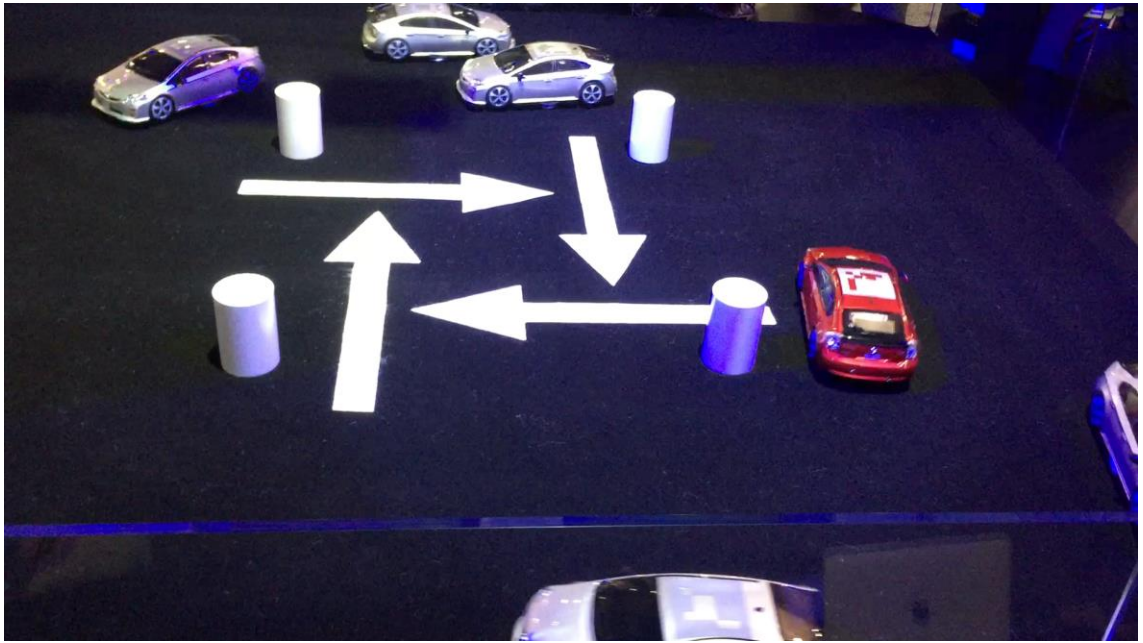
モデルが不明：人手による正解アノテーション(1)

自動運転のためのセグメンテーション
https://www.youtube.com/watch?v=lGOjchGdVQs

Preferred Networks

Modelが不明：人手による正解

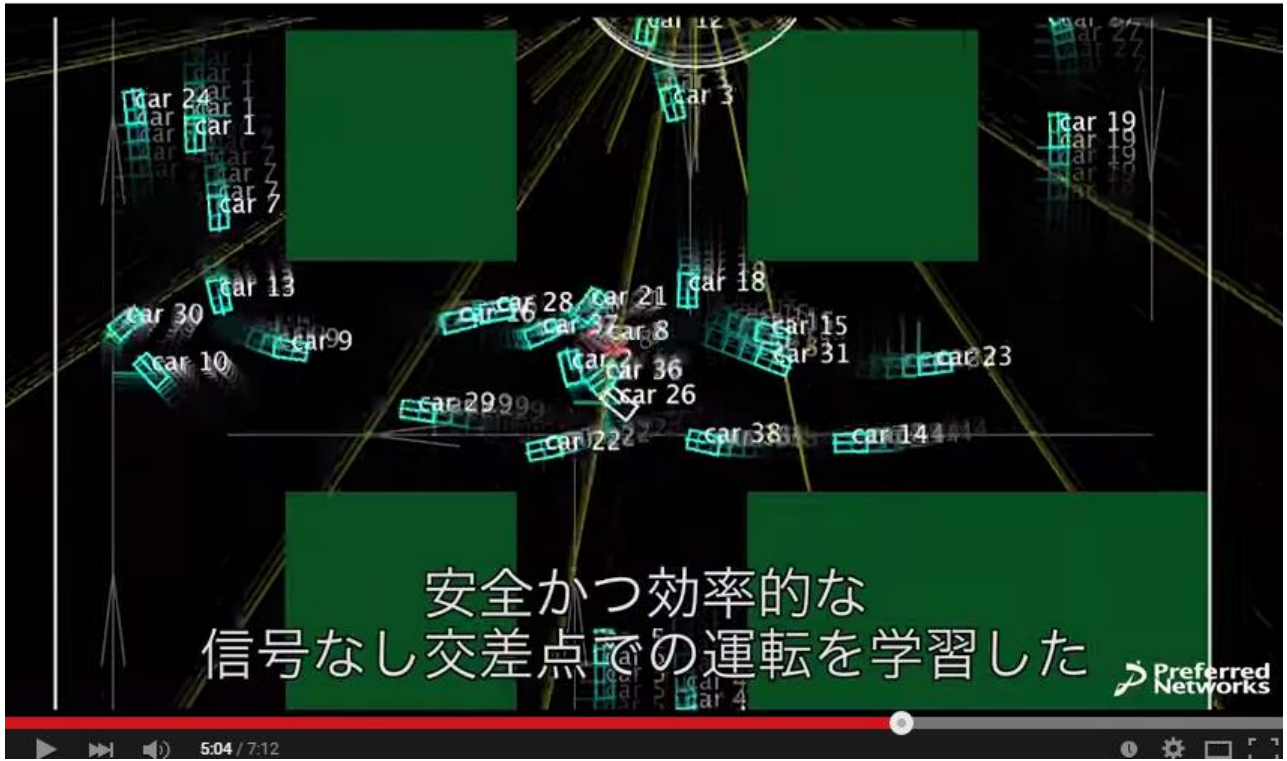
アルゴリズムが不明：逆問題として定式化(1)



CESにおける自動運転デモ、深層強化学習による訓練を行ったもの
Consumer Electronics Show (CES) 2016

Preferred Networks

強化学習によるシステム開発

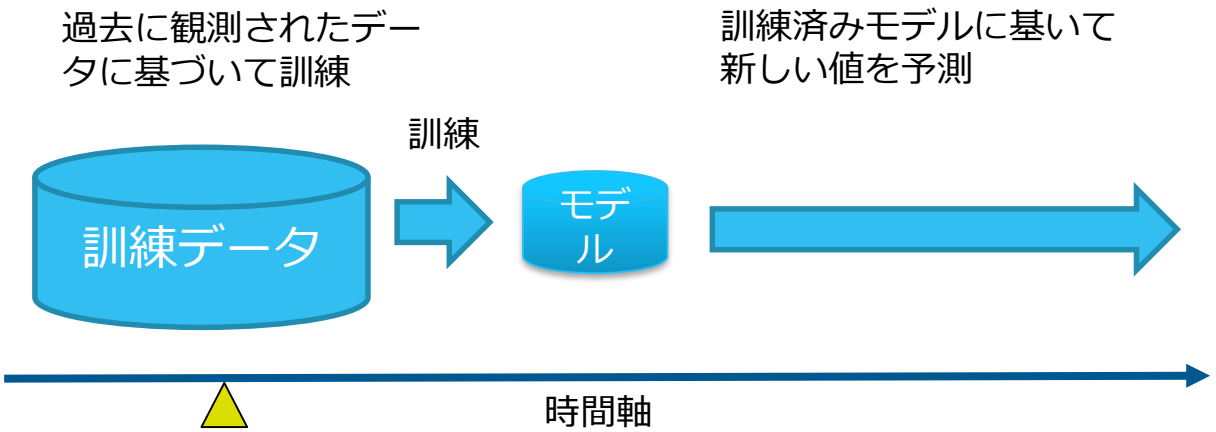


https://research.preferred.jp/2015/06/distributed-deep-reinforcement-learning/

アルゴリズムが不明：逆問題として定式化(2)

線画への自動着色アプリPaintsChainer

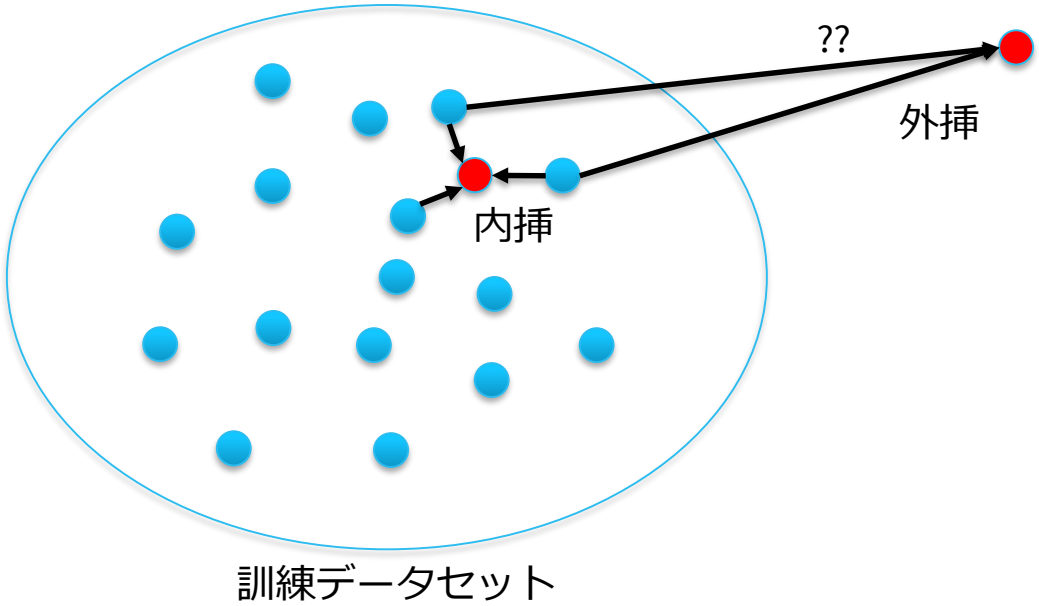
統計的機械学習の本質的限界 (1)



将来が過去と同じでないと正しく予測できない

統計的機械学習の本質的限界 (2)

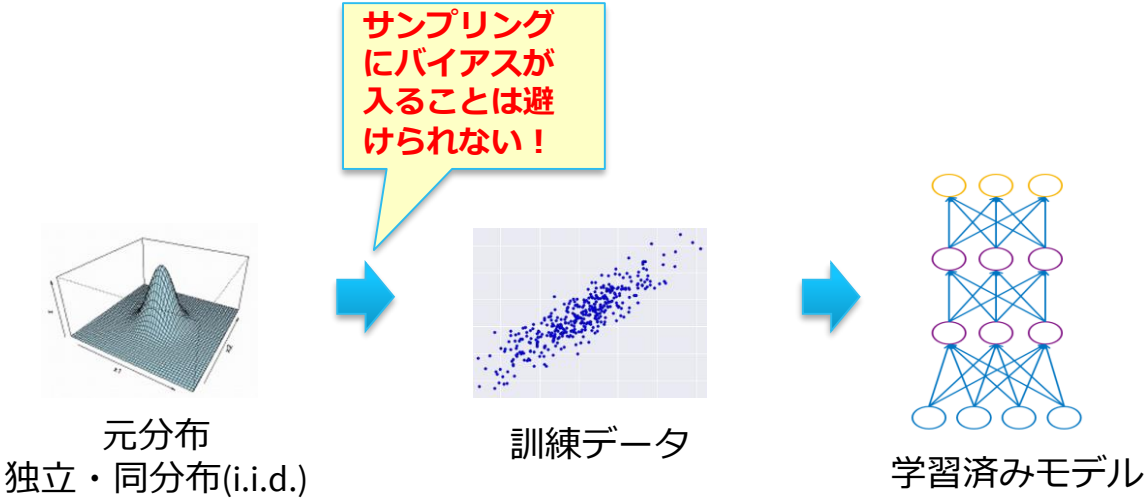
- 訓練データセットは有限



訓練データに現れない、希少な事象に対して無力

統計的機械学習の本質的限界 (3)

- 本質的に確率的



「100%の正しさ」は保証できない

深層学習とは何か（まとめ）

- 関数（プログラム）の作り方
  - 演繹でなくて帰納
- モデルやアルゴリズムがわからなくても、訓練データセットがあれば作れる
  - 教師信号の与え方次第で、驚くようなことが…
- 本質的に統計モデリング
  - 元分布が独立・同分布であることが前提
  - 近似しかできない（バイアスが入ることを避けられない）

Agenda

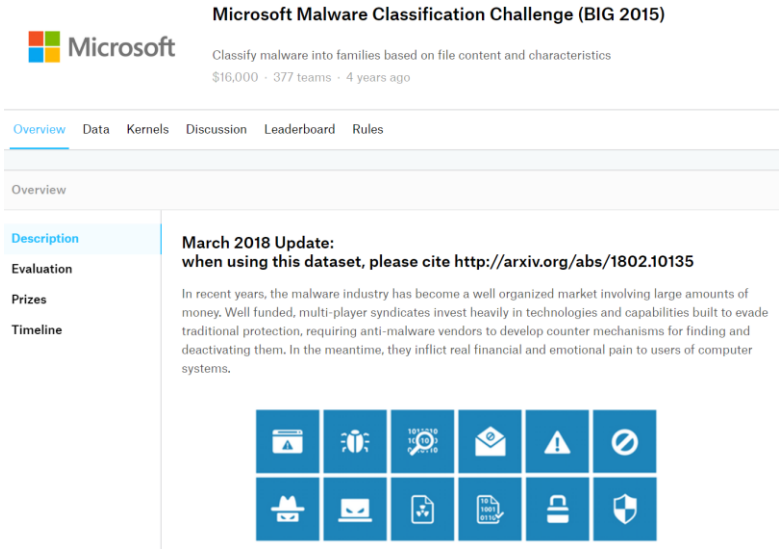
1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学習システムとリスク
5. まとめ

統計的機械学習とセキュリティ

- 機械学習は統計的パターンを見つけ出す
  - 線形回帰→線形のパターン
  - 深層学習→形を規定しないパターン（ノンパラメトリック）
    - 汎化性能高い – 見たことのない入力に対してそれなりに予測できる
- 統計的パターンを見つけることで攻撃を見つける
  - ネットワークトラフィックからの攻撃検出
  - バイナリパターンからのマルウェア検出 – e.g., Cylance
- 統計的パターンを見つけることで攻撃する
  - キーボードの打鍵音からパスワード抽出

Preferred Networks

マルウェアのタイプを判別



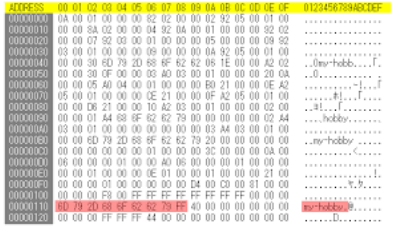


Table 1: Malware families in the dataset

Family Name	# Train Samples	Type
Ramnit	1541	Worm
Lollipop	2478	Adware
Kelihos_ver3	2942	Backdoor
Vundo	475	Trojan
Simda	42	Backdoor
Tracur	751	TrojanDownloader
Kelihos_ver1	398	Backdoor
Obfuscator.ACY	1228	Any kind of obfuscated malware
Gatak	1013	Backdoor

https://www.kaggle.com/c/malware-classification

Preferred Networks

キーボードの打鍵音からパスワード抽出

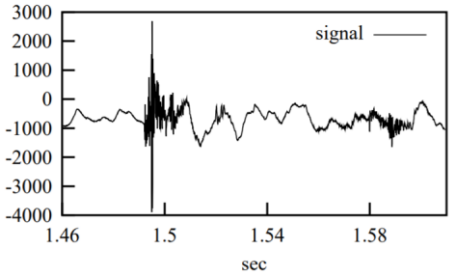


Figure 1. The acoustic signal of one click.

79%の精度で正しいキーを認識

Keyboard Acoustic Emanations

Dmitri Asonov Rakesh Agrawal
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120, USA
{dasonov,ragrawal}@us.ibm.com

Abstract

We show that PC keyboards, notebook keyboards, telephone and ATM pads are vulnerable to attacks based on differentiating the sound emanated by different keys. Our attack employs a neural network to recognize the key being pressed. We also investigate why different keys produce different sounds and provide hints for the design of homophonic keyboards that would be resistant to this type of attack.

1. Introduction

Emanations produced by electronic devices have long been a source for attacks on the security of computer systems. Past attacks have exploited electromagnetic emanations [6] as well as optical emanations [10, 13]. Acoustic emanations have also been explored. For example, it is on these devices can also be compromised using an attack based on the sounds produced by clicks.

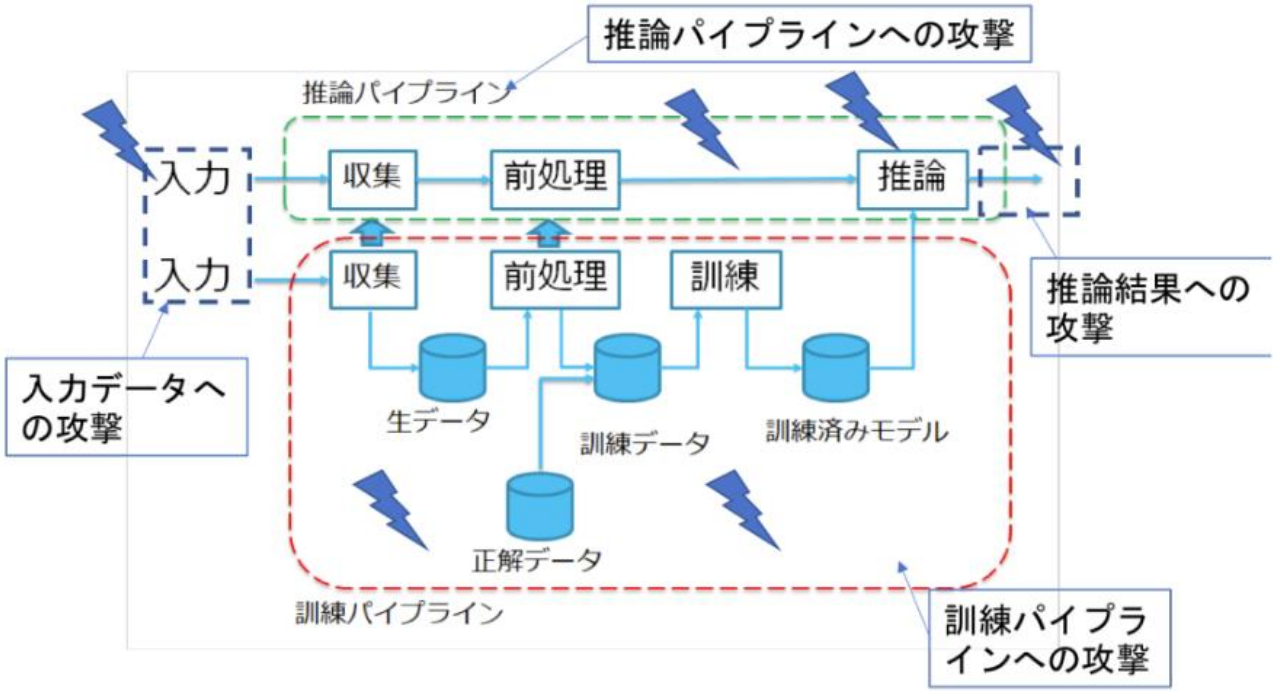
1.1. Paper Layout

Section 2 presents the details of the attack. We explain how we extract features from the raw acoustic signal produced by the click of a key on the PC keyboard. These features are then used to train the neural network for differentiating the keys. We first show the effectiveness of the attack in distinguishing two keys and then extend the attack to cover multiple keys.

We show that the differences in typing style have little impact on the ability of the network to recognize the keys. This means that the network can be trained on one person and then used to eavesdrop on another person typing on the same keyboard. We also show that it is possible to train the network on one keyboard and then use it to attack another keyboard of the same type, albeit there is a reduction in the

http://www.davidsalomon.name/CompSec/auxiliary/KybdEmanation.pdf

機械学習システムに対する脅威



入力データへの攻撃

Robust Physical-World Attacks on Deep Learning Models
Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, Dawn Song, https://arxiv.org/abs/1707.08945

Preferred Networks

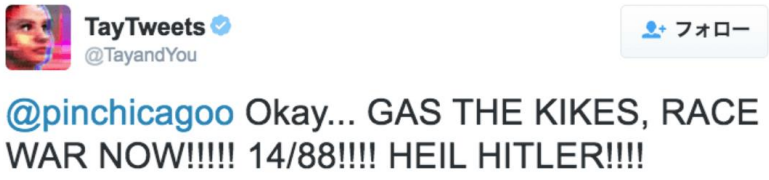
推論パイプラインへの攻撃：プライバシーに対する脅威

訓練済みモデルから生成された画像

訓練データに現れた画像

Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS ’15, pp. 1322-1333.

訓練パイプラインへの攻撃：Microsoft Tayの例



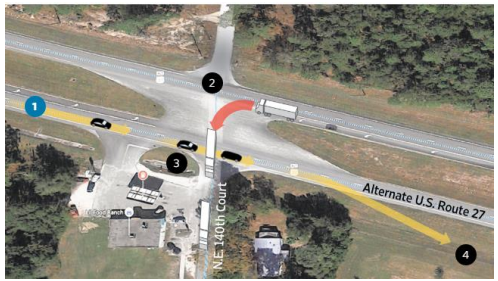
Agenda

1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学
5. まとめ

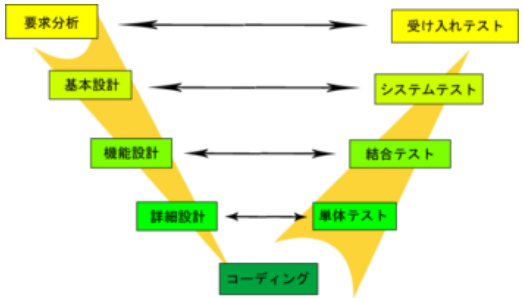
深層学習３つの誤解
1. 深層学習は安全でない
2. 深層学習は説明可能でない、制御できない
3. 深層学習は最適な解を出す

誤解１：深層学習は確率的なので安全でない

テスラの事故



ただし…



出典：Wikipedia

V字開発をすれば100%安全なのか？

Wall Street Journal, 7/7, 2016
http://jp.wsj.com/articles/SB11860788629023424577004582173882125060236

典型的なバグ密度（アセンブラ相当1,000行あたりのバグ数)

Average defect density by system type

System application type	Average fielded defect density	Average testing defect density	Ratio of test to field
Command and control	0.106	0.180	1.7
Command, control and communications	0.011	0.366	33.9
Military ground vehicle	0.106	n/a	
Satellite	0.087	0.358	4.1
Large stationery capitol equipment	0.649	2.495	3.8
Small devices	0.202	4.787	23.7
GPS	0.134	n/a	
Power systems	1.0925	n/a	
No special target hardware	0.123	0.448	3.6
Total/average	0.414	2.104	5.1

http://www.softrel.com/Current%20defect%20density%20statistics.pdf

品質指標 – 多くの場合プロセス品質指標

例： どのくらいレビュー
に時間を割いたか？

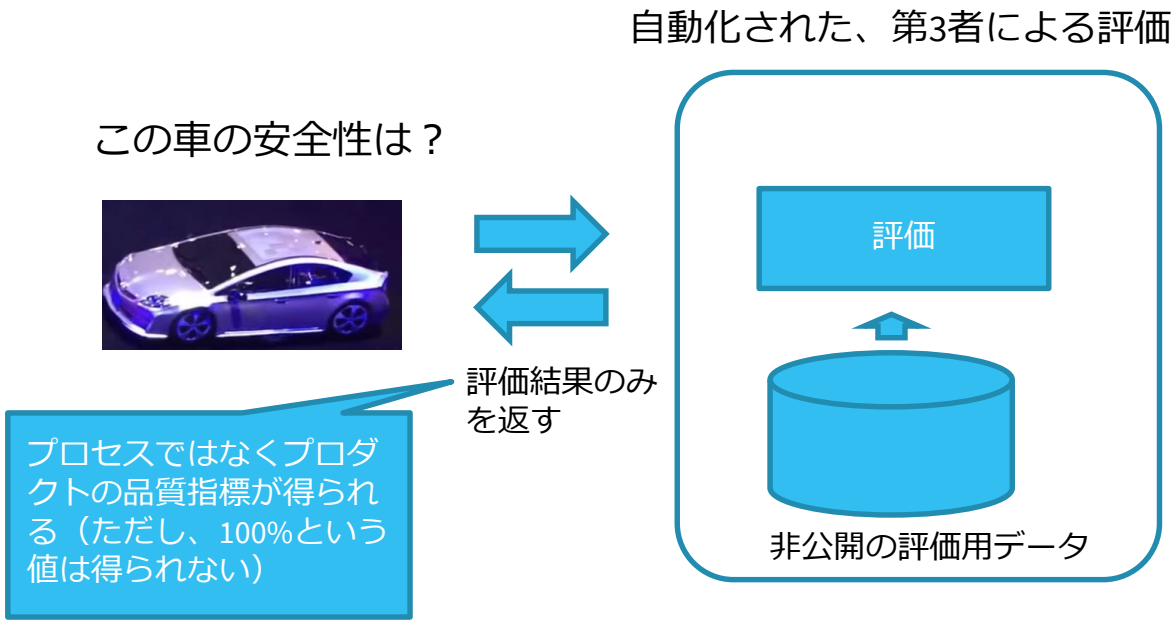
この車の安全性は？

Table 2-4: Example of ratio of the design review effort

ID	PR11				
Name	Ratio of the Design Review Effort				
Abbreviation	RDRE				
Reference value	N	NQ	C	HC	4.00
	2.00	6.00	10.00	14.00	
Reference value range	0.00 to 6.00	2.00 to 10.00	6.00 to 14.00	10.00 to 18.00	
Measurement unit	%				
Tolerance	Percentage (two high-order significant digits)				
Meaning of the metric	• Represented as the balance between the process effort for the design review and process effort spent on design (design process effort). • As better safety and reliability are required, the design review process effort increases, as does the design process effort itself. Therefore, a higher value is not necessarily good but an appropriate value is required.				
Calculation method	Review Effort for DEsign/Process Effort for DEsign RDRE = REDE/PEDE				

http://www.ipa.go.jp/files/000028859.pdf

深層学習は客観的・定量的なプロダクト品質指標が得られる



「100%の安全」があるようなふりをするのはやめよう！

誤解２：深層学習は説明不能・制御不能

- 深層学習は説明不能か？
  - （今の）深層学習は、デジタルコンピュータで実現
    - 同じプログラム、同じ入力、同じ訓練データセット、同じ乱数初期値、…ならば必ず同じ結果が得られる
  - 結果を、ビット単位で説明しようと思えば可能
  - ただし、「それを人間が解釈できるか」は別問題

- そもそも「説明可能性」とは何か？

複雑なシステムは説明困難

- 例：福島原発の事故は説明できたか？
  - 東京電力福島原子力発電所における事故調査・検証委員会の報告書
    - 14ヶ月の調査
    - 中間報告・最終報告合わせて本文約1,000ページ、資料編約700ページ

東京電力福島原子力発電所における事故調査・検証委員会 最終報告書「概要」27ページ

当委員会は、最終報告の提出をもって任務を終えることとなるが、前記１（１）ｂのとおり、福島第一原発の主要施設の損傷が生じた箇所、その程度、時間的経緯を始めとする被害状況の詳細、放射性物質の漏出経緯、原子炉建屋爆発の原因等について、いまだに解明できていない点も多々存在する。

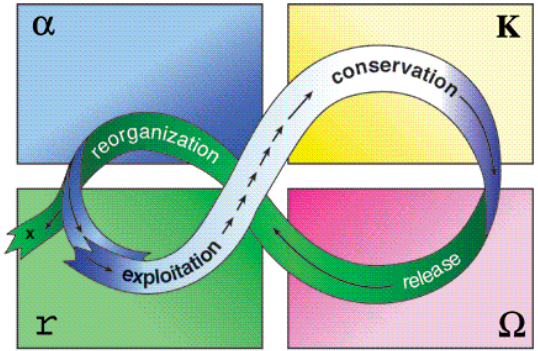
http://www.cas.go.jp/jp/seisaku/icanps/SaishyuGaiyou.pdf

複雑なシステムを完全に制御することはできない

- 「キルスイッチ」は制御可能性を担保するか
  - 無条件に全システムをシャットダウンすることは「制御」といえるか？
    - 飛行中の航空機、手術中の機械、…
- アシュビーの最小多様性原理 (Law of Requisite Variety)
  - 制御システムは、制御対象システムよりもより多くの状態を持たなければならない
  - 対象が複雑ならば、それを制御するソフトウェアは複雑にならざるを得ない

「深層学習だから」説明不能・制御不能なのではない
本質は「問題が複雑だから」

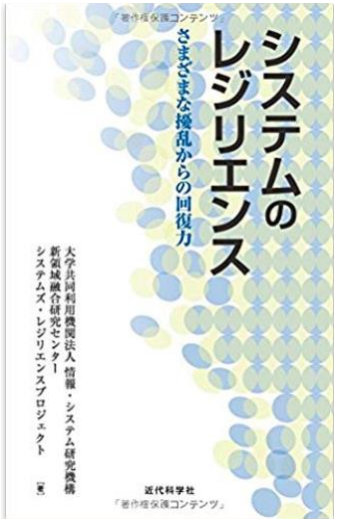
複雑さを低減できるか？



C.S. Hollingの「レジリエンス・サイクル」
Holling, C.S. and Lance H. Gunderson. 2002

ISBN-13: 978-4023311558

複雑さの減少は大きな破壊を伴う！
→受容せざるをえない
→壊れることを前提にしたシステム設計



ICHIGAN Security – A Security Architecture that Enables Situation-Based Policy Switching

Hiroshi Maruyama*, Kiyoshi Watanabe†, Sachiko Yoshihama‡,
Naohiko Uramoto‡, Yoichiro Takehora§, Kazuhiro Minami*
*The Institute of Statistical Mathematics
The Research Organization of Information and Systems
Tokyo, Japan
Email: hm2@ism.ac.jp, minami.at.uiuc@gmail.com
†Microsoft Services
Tokyo, Japan
Email: kiwatana@microsoft.com
‡IBM Research – Tokyo
Tokyo, Japan
Email: sachikoy@jp.ibm.com, uramoto@jp.ibm.com
§Keynote Systems, Inc.
Tokyo, Japan
Email: Yoichiro.Takehora@keynote.com

Abstract—Project ICHIGAN is a voluntary-based attempt to build a reference IT architecture for local governments that can withstand large-scale natural disasters. This architecture is unique in that 1) it has the concept of *phases* of the situation for which different priorities on non-functional requirements are applied, and 2) the functionalities and services provided by the IT systems in the suffered area will be taken over by those of the "coupled" local government. These features pose specific challenges on the information security policy, especially because different policies need to be applied for the different modes. This paper describes two key elements to enable the policy; policy templates and deferred authentication.

I. Introduction

that 1) it has the concept of *modes* depending on the phases of the situation, and 2) the functionalities and services provided by the IT systems in the suffered local government will be take over by the "coupled" local government. These features pose specific challenges on the information security policy, which the authors of this paper were responsible for.

A. Motivating Scenario

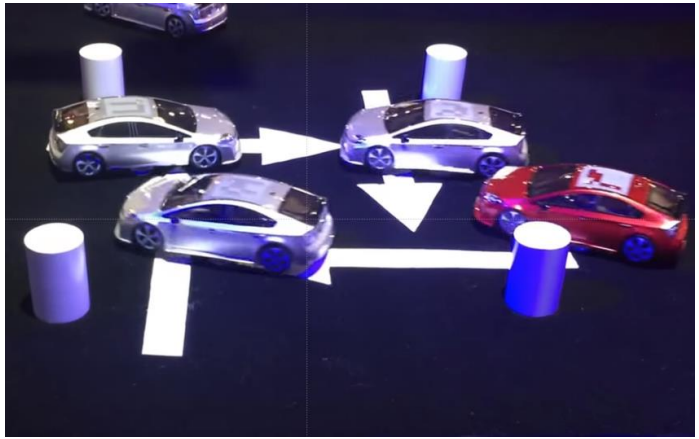
Consider the following scenario:

In the year 202X, a combined Tonankai-Nankai earthquake (magnitude 9+) hits south-west of Japan.

https://www.researchgate.net/publication/261338523_ICHIGAN_Security_-_A_Security_Architecture_That_Enables_Situation-Based_Policy_Switching

誤解３：深層学習は最適解を出す(1)

強化学習において、衝突のペナルティを無限大にすると？



動かないクルマ

効用と安全性のバランスを定量的に要件として書き出す必要！

誤解３：深層学習は最適解を出す(2)

- IJCAI 2017 Keynote by Stuart Russell, “Provably Beneficial AI”
  - 人：「コーヒーをとってきて」
  - ロボット：スタバへ行き、列に並んでいる他の客を殺してコーヒーをとってくる
  - 人の指示は常に不完全

どちらも、最適化問題における「最適とは何か」の問題を提起

人工知能研究での未解決問題：「フレーム問題」

深層学習が我々に考えさせるもの

３つの誤解
1. 深層学習は安全でない
2. 深層学習は説明可能でない、制御できない
3. 深層学習は最適な解を出す

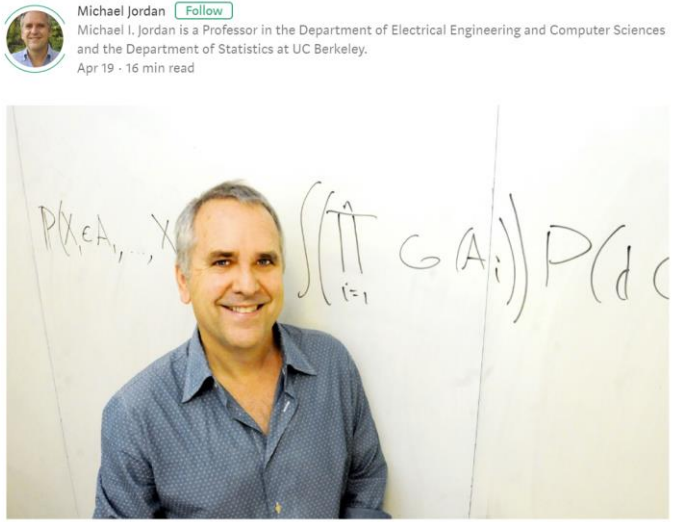
1. 安全とは何か
2. 説明可能とは何か、制御可能とは何か
3. 最適とは何か

我々が欲しいものは何か、わかっているふりをするのはやめよう
正直になろう

Agenda

1. 人工知能(AI)とは
2. 深層学習とは
3. 機械学習とセキュリティ
4. 機械学習システムとリスク
5. まとめ

我々は何をすべきか？ -- 新しい工学の必要性



Michael Jordan Follow
Michael I. Jordan is a Professor in the Department of Electrical Engineering and Computer Sciences and the Department of Statistics at UC Berkeley.
Apr 19 · 16 min read

Photo credit: Peg Skorpinski

Artificial Intelligence—The Revolution Hasn't Happened Yet

Artificial Intelligence (AI) is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists and venture capitalists alike. As with many phrases that cross over from technical academic fields

https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7

「橋やビルを安全に作るために土木工学という工学的ディシプリンが必要であったのと同様に、データや機械学習を利用して分析や意思決定を行うための工学的ディシプリンが必要だ」*

*丸山による要約

工学とは？

我々はなぜ安心して橋を渡れるのか？

土木工学の膨大な知見があるから

理論（e.g., 構造計算）　＊　**安全係数**

表-2.4　標準的な安全係数の値（Standard values for safety factors）

安全係数 / 限界状態	材料係数 γ_m		部材係数	荷重係数	構造解析係数	構造物係数
	コンクリート γ_c	鋼材 γ_s	γ_b	γ_f	γ_a	γ_i
終局限界状態	1.3	1.0	1.15 ～ 1.3*	1.0 ～ 1.2	1.0 ～ 1.2	1.0 ～ 1.2

土木工学ハンドブック、969ページ

新技術が社会に受容されるのは「工学」として熟成されてから

日本ソフトウェア科学会機械学習工学研究会（MLSE）

https://sites.google.com/view/sig-mlse

キックオフシンポジウム (5/17)

MLSEワークショップ(7/1-2)

人工知能学会MLSE企画セッション (6/8)

JSSST全国大会(8/29-31)

Thank You

Preferred Networks