

# データドリブンを支える プラットフォーム

2019年10月22日

ヤフー株式会社

角田 直行



# 自己紹介

角田 直行 (かくだ なおゆき)

CTO 技術戦略本部TI室 室長

- 2005～ 中途入社  
地図、路線、検索、検索エンジン、  
検索プラットフォームを開発
- 2012～ データ事業部門の技術統括
- 2017/6 Green500 2位獲得
- 2019/4～ 技術戦略部門へ

ヤフーが作ったスーパーコンピュータが、省エネ性能で世界第2位に！——スパコン「kukai」開発の担当者に聞く

2017年9月4日

インタビュー キャリア



ヤフーが、ディープラーニングを活用したデータ解析の高度化やサービスのパーソナライズ精度向上のために極秘で開発を進めていたスーパーコンピュータ「kukai」。

<https://linotice.tumblr.com/post/164962607234/20170904>

# Agenda

- AI・データ事例
- データプラットフォーム変遷
- 今後の課題

本発表前に  
お伝えしたいこと

# 自治体との災害情報連携

宮坂学 Manabu Miyasaka @miyasaka

もし最寄りの自治体のホームページが繋がりにくかったり落ちている場合は、ヤフーで自治体名を検索してキャッシュサイトにアクセスしてみてください。  
自治体サイトとほぼ同様の情報にアクセスできる場合があります。  
ただし、更新タイミングがずれる可能性があります。

検索: 江戸川区 ハザードマップ 江戸川区役所 江戸川区 天気

総合トップ 江戸川区ホームページ  
www.city.edogawa.tokyo.jp/

18:46 4G

search.yahoo.co.jp

JAPAN

検索: 江戸川区

ウェブ 画像 動画 リアルタイム 知恵袋 求人 地

江戸川区 ハザードマップ 江戸川区役所 江戸川区 天気

江戸川区 粗大ごみ

総合トップ 江戸川区ホームページ  
www.city.edogawa.tokyo.jp/

東京都江戸川区の公式サイトです。暮らし・手続きに関する区民向け情報のほか、市政情報、産業・事業者向け情報、最新の ...

暮らし・手続き 子育て・教育  
健康・医療・福祉 えどがわ地域情報局

「江戸川区」のキャッシュサイト

公式サイトがつながりにくい場合は、ヤフーによる以下のキャッシュサイトをお試しください

ヤフーによるキャッシュサイトはこちら  
http://www.city.edogawa.tokyo.jp.cache.yimg.jp/

東京都江戸川区の災害情報 Yahoo!天気・災害  
2019年10月12日 18時42分 現在

YAHOO! JAPAN 防災速報

いいね! 2.4万 ツイート Yahoo! JAPAN - ヘルプ

災害の情報をいち早くお知らせ  
**防災速報** 無料!

簡単! ダウンロードして防災対策

緊急地震速報や豪雨予報、避難情報などをいち早くお知らせ。  
受け取れる情報についてさらに詳しく

● スマートフォンの方はこちら  
アプリをダウンロードしてください。

iPhone版 Android版

● パソコンやケータイの方はこちら  
あなたが設定した地域の情報を、パソコンやケータイにメールでお知らせします。\*メール版のご利用はYahoo! JAPAN IDが必要です。

メール版 簡単登録へ

■ お知らせ

- 「自治体からの緊急情報」を池田市など22地域にも配信 (2019/09/30)
- 「自治体からの緊急情報」を堺市など30地域にも配信 (2019/08/29)

知っているとお九死に一生!?  
災害別の防災コラムを読む

千代田区 大阪市北区

津波予報 発表中  
津波警報 海岸から離れ、高い場所へ逃げて  
〇月〇日 0:00

地震情報  
震度4  
最大震度5強: 東京湾  
今後の情報に注意してください。  
〇月〇日 0:00

豪雨予報  
非常に激しい雨 (55mm/h以上)  
〇月〇日 0:00

千代田区の通知履歴

<https://twitter.com/miyasaka/status/1182956582264401920>

Yahoo!防災速報

# 災害への支援募金

## 令和元年台風第15号による千葉県災害への支援募金 (Yahoo!基金)

災害・復興支援



領収書  
発行なし

♥ 寄付総額 (概算) 21,174,840 円  
👤 寄付人数 26,771 人  
🕒 残り日数 44 日

寄付する

T-POINT

Tポイントを使って1ポイントから寄付できます。



## 令和元年台風19号緊急災害支援募金 (Yahoo!基金)

災害・復興支援



領収書  
発行なし

♥ 寄付総額 (概算) 119,875,981 円  
👤 寄付人数 139,190 人  
🕒 残り日数 75 日

寄付する

現在の継続寄付人数：555人

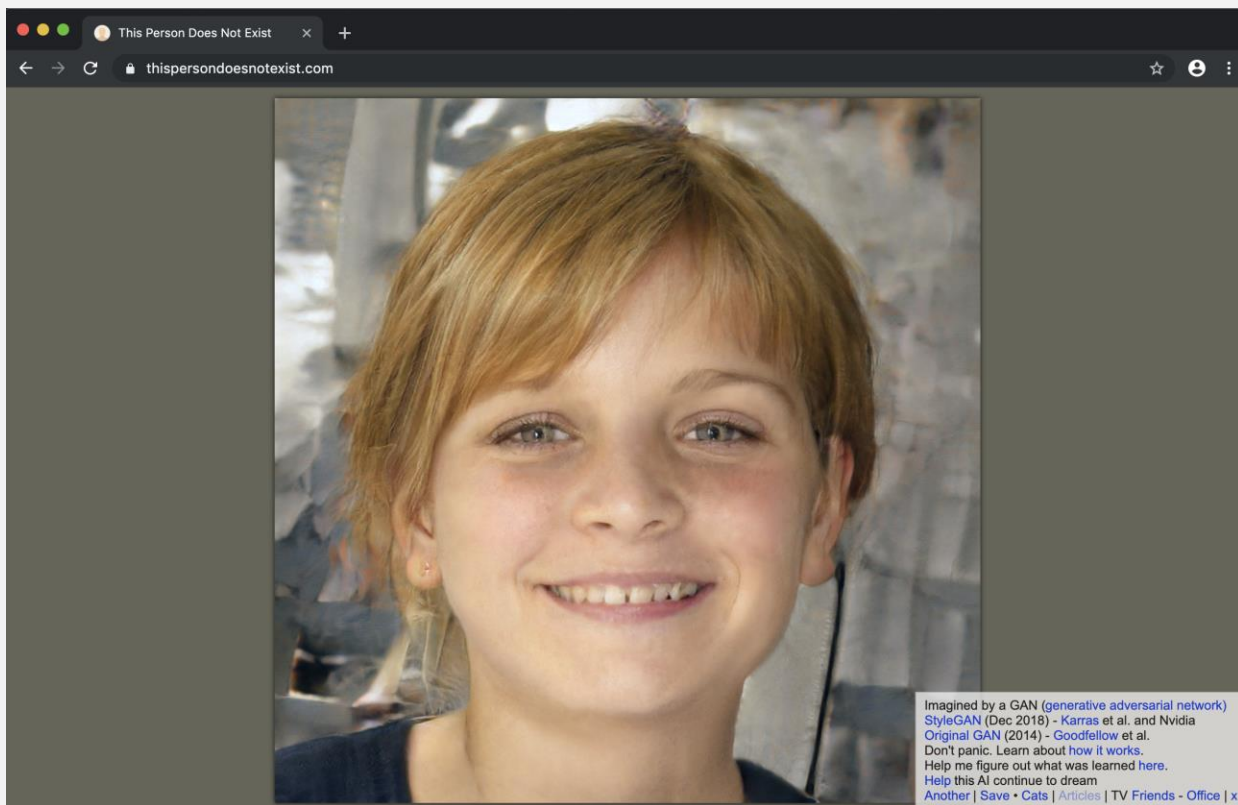
毎月の継続的な応援が大きな支えになります。

T-POINT つながる募金

<https://donation.yahoo.co.jp/>

# AI・データ事例

# 画像、音声



<https://thispersondoesnotexist.com/>



<https://www.youtube.com/watch?v=nOLu17nPQWU>



# VideoIn Ads

 **Matthew Brennan**  
@mbrennanchina フォローする

Wow! Worth watching this. China's largest video platform **#Tencentvideo** (97M paying China subscribers) will begin inserting extra ads into movies/series that didn't exist in the original. **#computervision**



836,014回再生済み 0:00 / 0:45

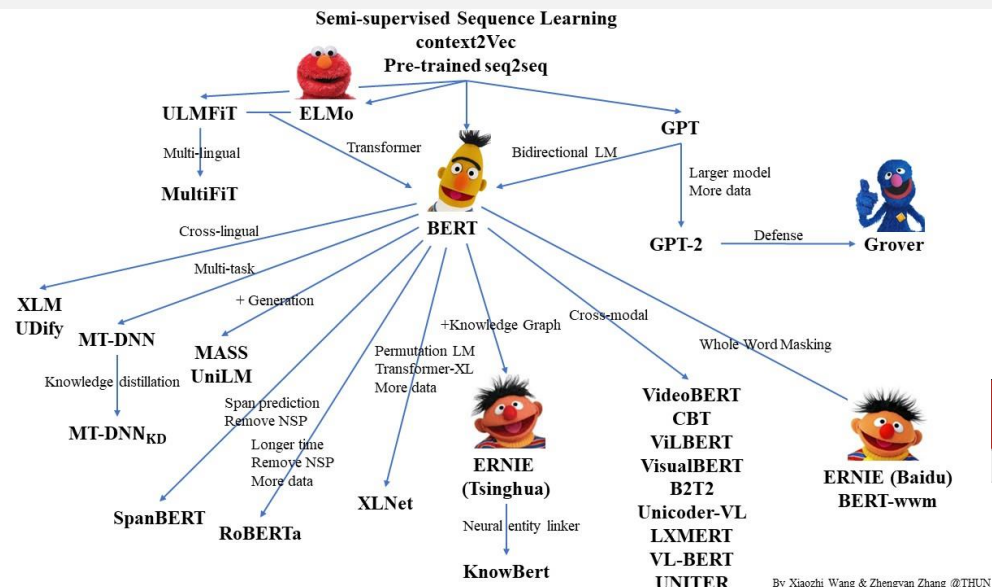
7:29 - 2019年10月15日

2,515件のリツイート 4,180件のいいね



<https://twitter.com/mbrennanchina/status/1184114082804158464>

# 自然言語処理



By Xiaozhi Wang & Zhengyan Zhang @THUN

<https://github.com/thunlp/PLMpapers>

arXiv.org > cs > arXiv:1907.06616

Search...  
Help | Ad

Computer Science > Computation and Language

## Facebook FAIR's WMT19 News Translation Task Submission

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov

(Submitted on 15 Jul 2019)

This paper describes Facebook FAIR's submission to the WMT19 shared news translation task. We participate in two language pairs and four language directions, English <-> German and English <-> Russian. Following our submission from last year, our baseline systems are large BPE-based transformer models trained with the Fairseq sequence modeling toolkit which rely on sampled back-translations. This year we experiment with different bitext data filtering schemes, as well as with adding filtered back-translated data. We also ensemble and fine-tune our models on domain-specific data, then decode using noisy channel model reranking. Our submissions are ranked first in all four directions of the human evaluation campaign. On En->De, our system significantly outperforms other systems as well as human translations. This system improves upon our WMT'18 submission by 4.5 BLEU points.

<https://arxiv.org/abs/1907.06616>



小猫遊りよう (たかにやし・りょう)  
@jaguring1

フォローする

お、48分前に性能向上があったみたいだ。AIの言語理解のために作られたベンチマーク「GLUE」において、グーグルがモデル「ALBERT」を投入。

人間のスコアを超えてからの流れ  
MT-DNN → XLNet → RoBERTa → adv-RoBERTa → ALBERT  
[gluebenchmark.com/leaderboard](http://gluebenchmark.com/leaderboard)

Rank	Name	Model	URL	Score	CoLA	SBT2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	Ax
1	ALBERT Team	Google Language4BERT (Ensemble)		88.4	88.1	87.1	88.4912	82.5920	74.2963	87.3	87.9	88.2	88.2	87.8	88.2
2	Microsoft DMS.AI & UMD	Adv-RoBERTa (ensemble)		88.3	88.0	86.8	83.5908	82.4922	74.8963	87.1	86.7	86.8	86.7	88.0	86.1
3	Facebook AI	RoBERTa		88.3	87.8	86.7	82.3888	82.2919	74.3962	86.8	86.2	86.8	86.2	88.0	86.7
4	XLNet Team	XLNet-Large (ensemble)		88.4	87.8	86.8	83.0907	81.8911	74.2963	86.2	86.8	86.8	86.3	86.4	87.8
5	Microsoft DMS.AI & MSR.AI	MT-DNN-Lossless		87.8	88.4	86.5	82.7963	81.1967	73.7889	87.9	87.4	86.0	86.3	88.0	82.8
6	GLUE Human Baseline	GLUE Human Baseline		87.1	86.4	87.8	86.3888	82.7828	68.8984	82.0	82.8	81.2	83.8	86.9	-

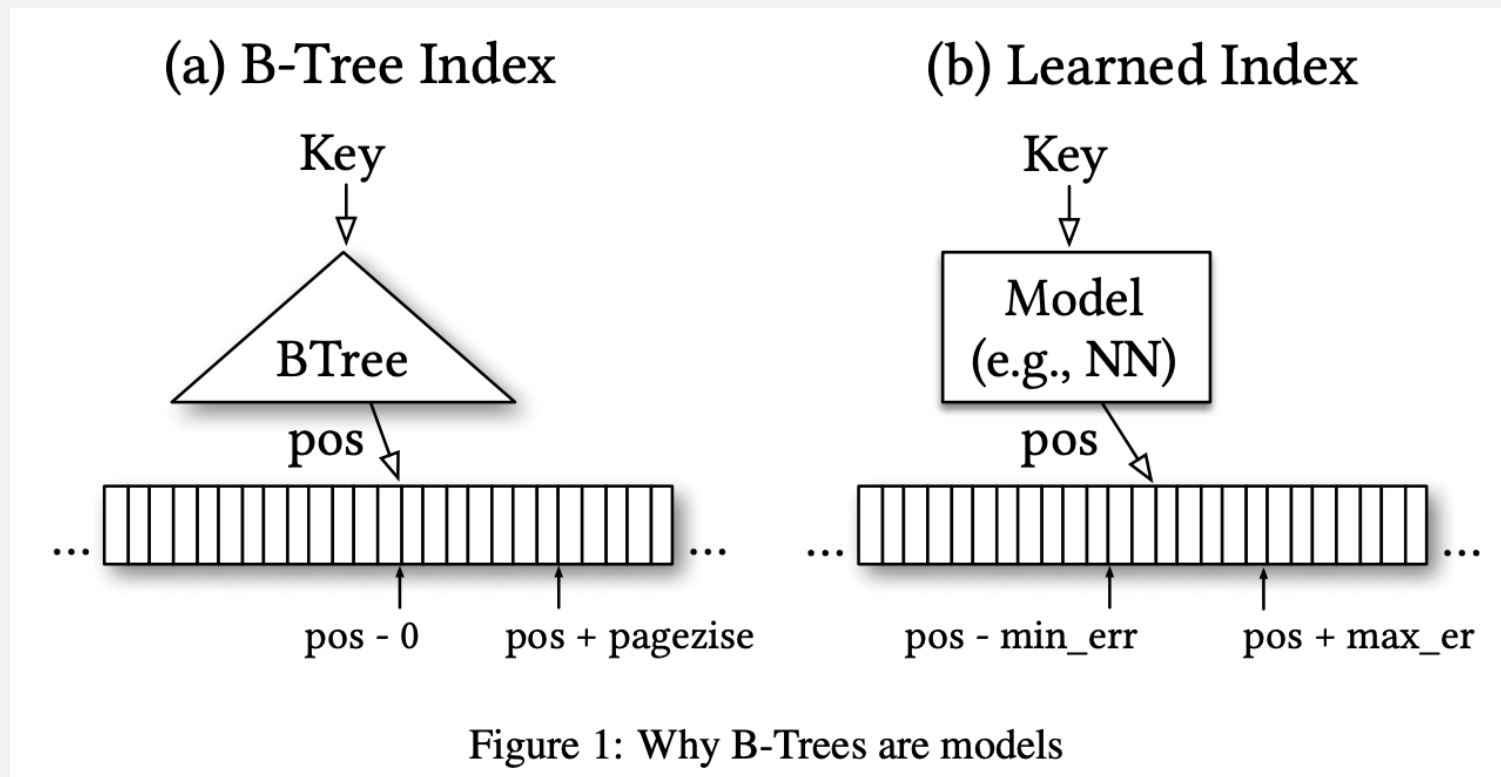
12:49 - 2019年9月16日

27件のリツイート 75件のいいね



# B-TreeをNNで置き換え

B-Treeより省メモリで70%高速



<https://arxiv.org/pdf/1712.01208.pdf>



Voice Recognition Engine **YJVOICE**



Article Title's Auto Generation



Smart Image Cropping



Personalization, Deduplication of Articles

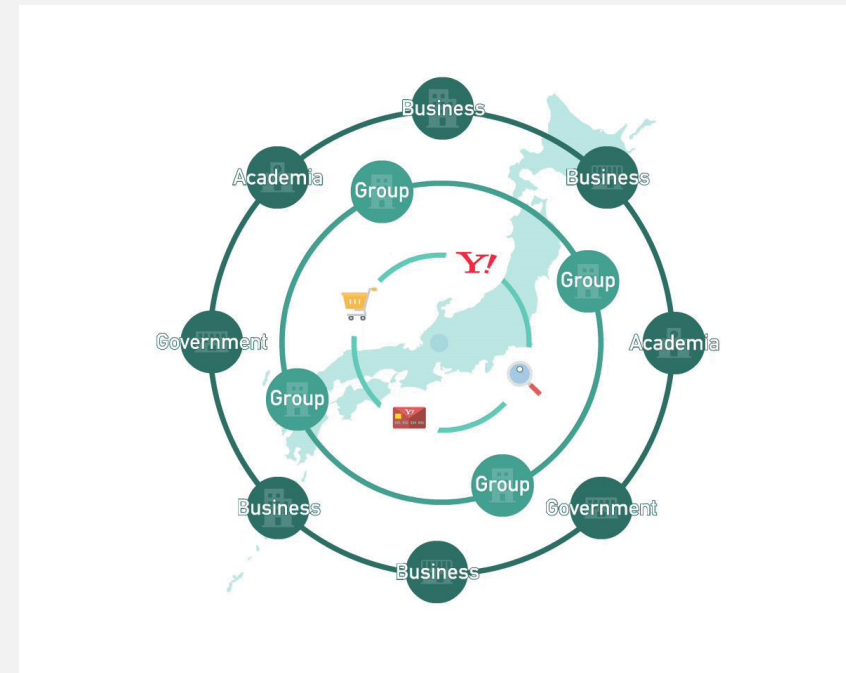


Forecasting PageViews

# データで得た知見を開放



 DATA FOREST



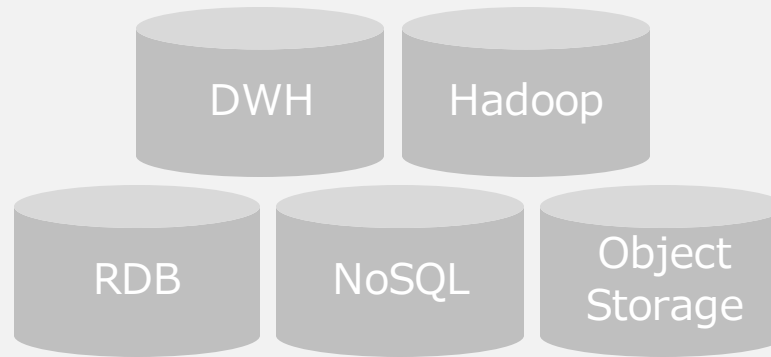
<https://dataforest.yahoo.co.jp/>

# 大規模データ処理プラットフォームの必要性

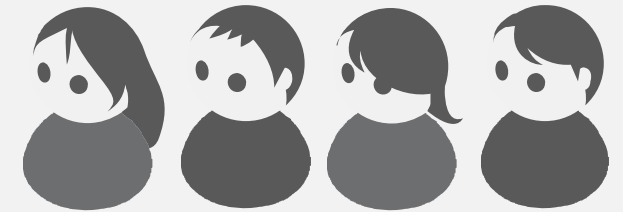
100以上のサービス



ビッグデータ



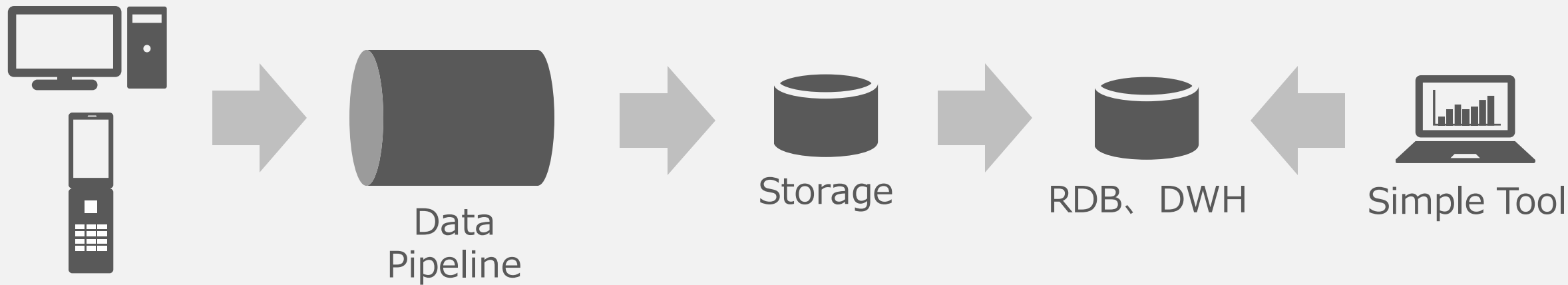
データサイエンティスト



# データプラットフォーム 変遷

# ~200X

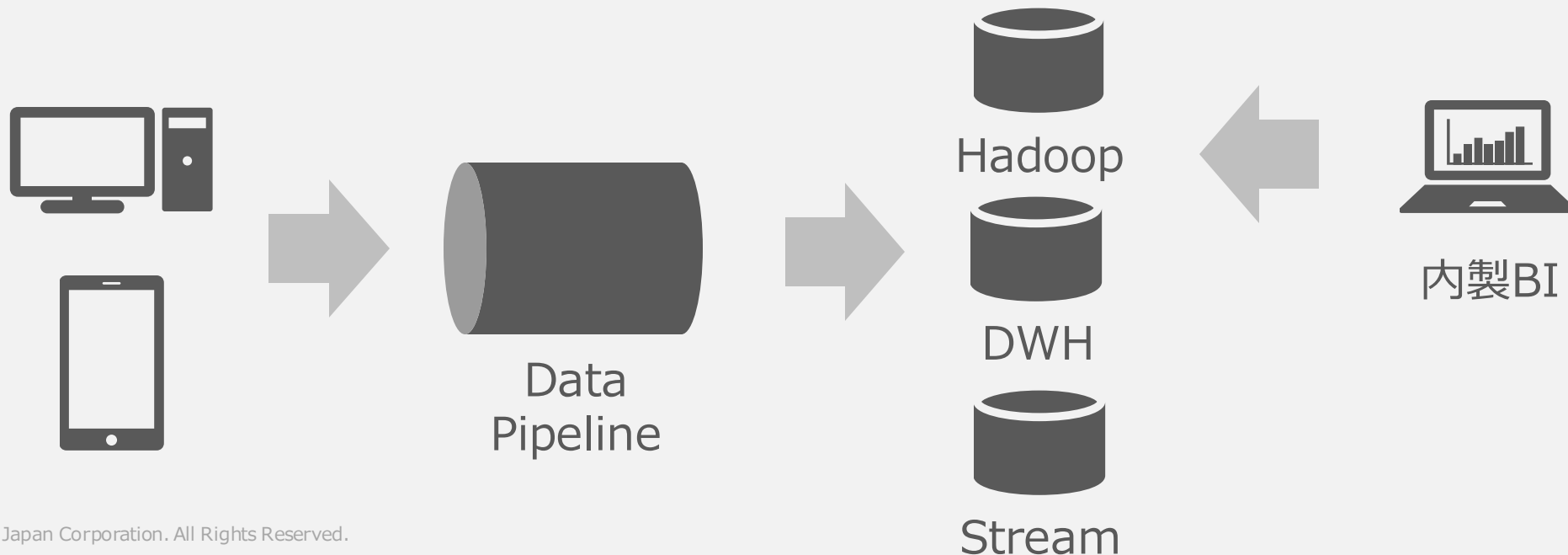
- PC & ケータイ向けサービスがメイン
- 数TB
- アクセスログ中心





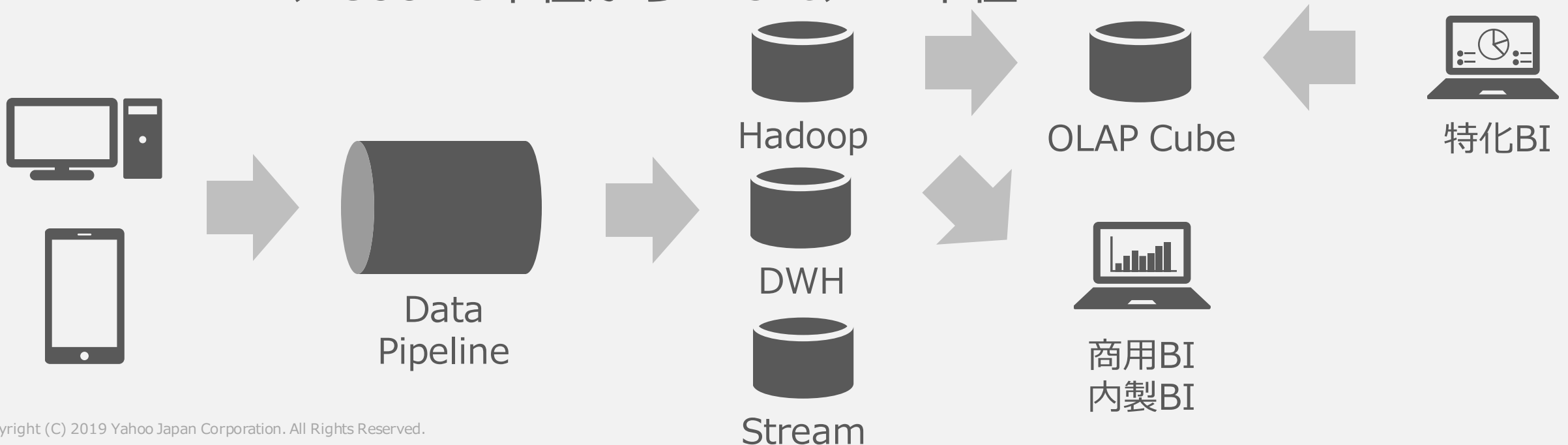
# 2010~2013

- スマホシフト
- 数PB
- リアルタイム用途にストリーム処理PFの導入



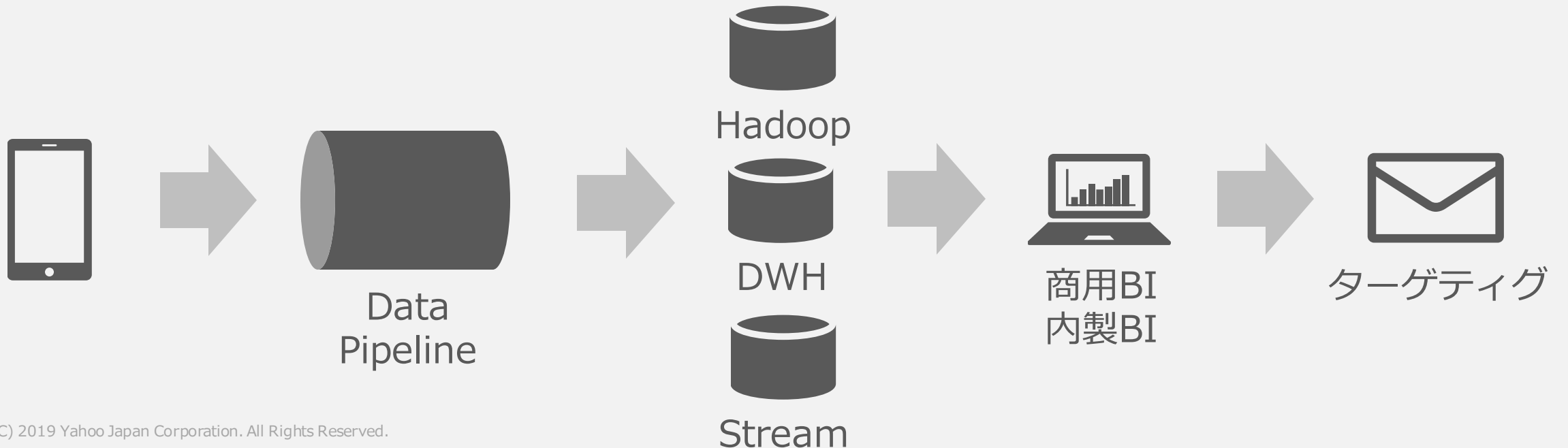
# 2014~2017

- 全サービスに統一形式の詳細ロギング
- 数十PB
- PV、Cookie単位からEvent、ID単位へ



# 2018~

- スマホユーザー解析にフォーカス
- 百数十PB
- カスタム解析、A/Bテスト、ターゲティング施策を容易に



# Lesson Learned

- 入り口(データ生成、ロギング)が最も重要
  - サービスの施策や戦略によって欲しいデータが変化する
  - 実装コストを最小限に抑えること
- アジャイルに進めるのは困難
  - サービス側に再実装コストがかかる
  - 過去にさかのぼって再処理するコストが高い
  - 集計結果の些細な差が大きな影響を及ぼす恐れがある(例:広告)
  - 横断分析ができるようフォーマットやタイミングを統一する
  - 複雑かつトラフィックの大きいユースケースを最初に手掛ける
- 保持コストはタダではない
  - 価値の高い/低いデータを把握しておく

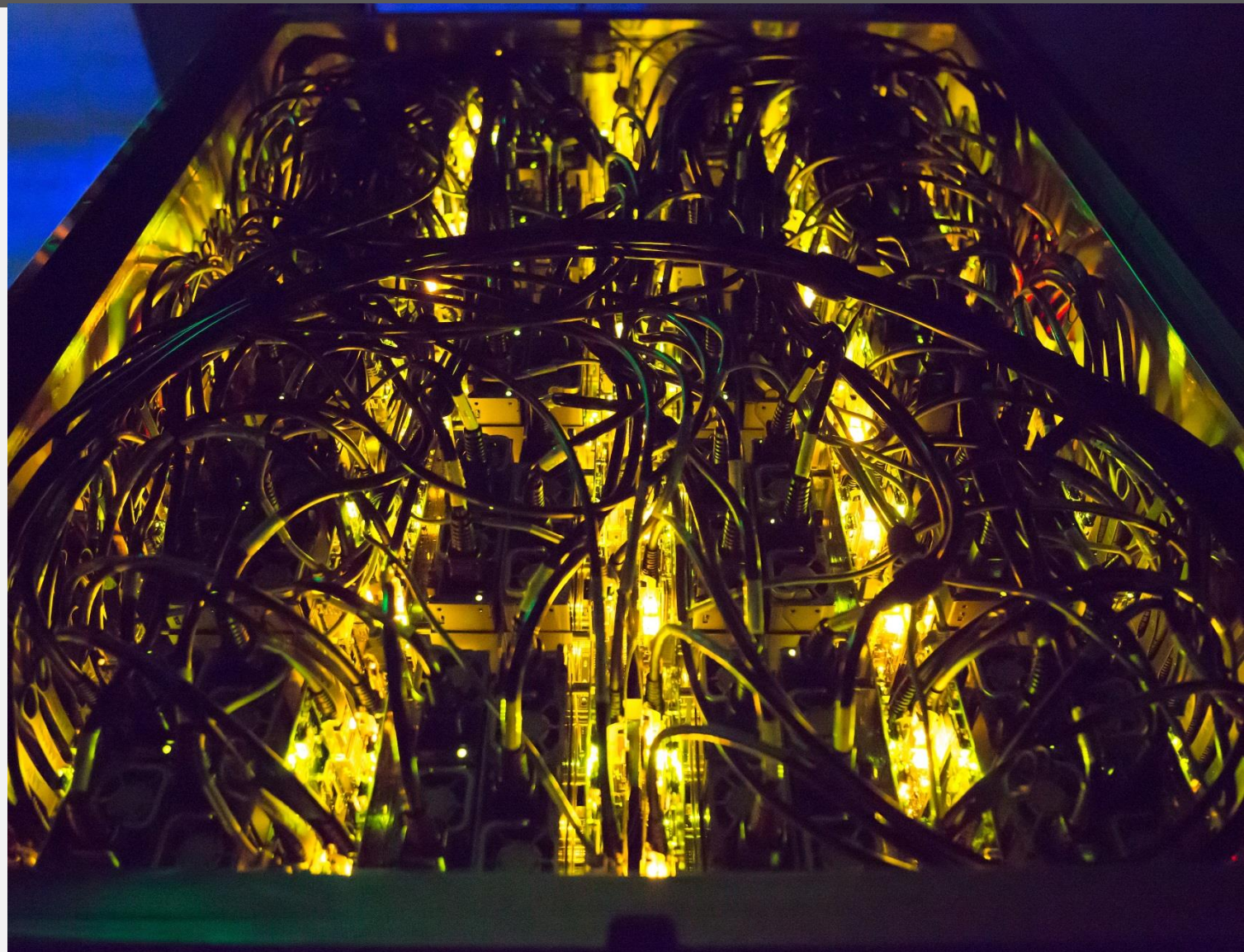
# Deep Learning向けスパコンの開発

Deep Learning用途に  
大規模GPU環境の必要性

運用コストを考慮し、  
消費電力効率を重視して設計

ベンチマークソフトを  
バイズ最適化で自動チューニング

Green500(2017/6)にて2位獲得



# Challenges

# Challenges

- プライバシー、セキュリティ
- SysML
- AutoML

# プライバシーポリシー

## プライバシーポリシー改定のお知らせ

2019年10月1日より、ヤフー株式会社（以下、Yahoo! JAPAN）は持株会社体制へ移行します（※1）。この体制変更に伴い、お客様とのプライバシーに関するお約束である「プライバシーポリシー」を改定します。（新しいプライバシーポリシーは[こちら](#)）

### **i** 本お知らせの要約

- Yahoo! JAPANは、お客様のプライバシー保護とそのため情報セキュリティ対策を第一に考えています。
- 新体制のもと、Yahoo! JAPANとグループ企業（※2）全体でよりよい体験をお客様へお届けしていけるよう、Yahoo! JAPANのプライバシーポリシーを改定します。
- グループ企業における新たな体験を目的としたデータ連携は、お客様の同意なしに開始することはありません（デフォルトオフ）。なお、改定前のプライバシーポリシー等ですでに同意いただいている内容には影響ありません。

### スコアの運用見直しに関して

Yahoo!スコアについても、2019年10月1日より初期状態でスコアの作成を行わないよう変更します。詳しくは[こちら](#)

<https://privacy.yahoo.co.jp/notice20190901.html>



# セキュリティ強化

## パスワードレス認証

ヤフー、新たなウェブ認証の規格「FIDO2」のFIDO Certified（認定）を国内企業で唯一取得  
パスワードを使わない安全なログイン環境の実現に前進

ヤフー株式会社（以下、ヤフー）は、生体認証などの次世代認証の標準化を提唱する業界団体FIDO（ファイド）アライアンスの新たな規格「FIDO2」の認定を、2018年8月に世界で初めて開催された「FIDO2」認定テストにおいて、このたび取得したことをお知らせします。国内企業では唯一の認定取得となります。



<https://about.yahoo.co.jp/pr/release/2018/09/27a/>

## 全サービスTLS 1.2対応

### セキュリティ強化のお知らせ



古いブラウザ、パソコン、スマートフォンなどでは、  
2018年10月中旬までにYahoo! JAPANの  
全ウェブサービスが順次ご利用いただけなくなります。

<https://security.yahoo.co.jp/news/tls12.html>

# SysML

## System and Machine Learning

- システムとサイエンスの情報交差点
- 機械学習に適したシステム(MLOps)
- システムの機械学習適用
  
- 機械学習工学研究会(MLSE)に  
運営委員として参加

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

arXiv.org > cs > arXiv:1904.03257

Search...

Help | Adv...

Computer Science > Machine Learning

### SysML: The New Frontier of Machine Learning Systems

Alexander Ratner, Dan Alistarh, Gustavo Alonso, David G. Andersen, Peter Bailis, Sarah Bird, Nicholas Carlini, Bryan Catanzaro, Jennifer Chayes, Eric Chung, Bill Dally, Jeff Dean, Inderjit S. Dhillon, Alexandros Dimakis, Pradeep Dubey, Charles Elkan, Grigori Fursin, Gregory R. Ganger, Lise Getoor, Phillip B. Gibbons, Garth A. Gibson, Joseph E. Gonzalez, Justin Gottschlich, Song Han, Kim Hazelwood, Furong Huang, Martin Jaggi, Kevin Jamieson, Michael I. Jordan, Gauri Joshi, Rania Khalaf, Jason Knight, Jakub Konečný, Tim Kraska, Arun Kumar, Anastasios Kyrillidis, Aparna Lakshmiratan, Jing Li, Samuel Madden, H. Brendan McMahan, Erik Meijer, Ioannis Mitliagkas, Rajat Monga, Derek Murray, Kunle Olukotun, Dimitris Papailiopoulos, Gennady Pekhimenko, Theodoros Rekatsinas, Afshin Rostamizadeh, Christopher Ré, Christopher De Sa, Hanie Sedghi, Siddhartha Sen, Virginia Smith, Alex Smola, Dawn Song, Evan Sparks, Ion Stoica, Vivienne Sze, Madeleine Udell, Joaquin Vanschoren, Shivaram Venkataraman, Rashmi Vinayak, [Markus Weimer](#), Andrew Gordon Wilson, Eric Xing, Matei Zaharia, Ce Zhang, Ameet Talwalkar

(Submitted on 29 Mar 2019 (v1), last revised 1 May 2019 (this version, v2))

Machine learning (ML) techniques are enjoying rapidly increasing adoption. However, designing and implementing the systems that support ML models in real-world deployments remains a significant obstacle, in large part due to the radically different development and deployment profile of modern ML methods, and the range of practical concerns that come with broader adoption. We propose to foster a new systems machine learning research community at the intersection of the traditional systems and ML communities, focused on topics such as hardware systems for ML, software systems for ML, and ML optimized for metrics beyond predictive accuracy. To do this, we describe a new conference, SysML, that explicitly targets research at the intersection of systems and machine learning with a program committee split evenly between experts in systems and ML, and an explicit focus on topics at the intersection of the two.

<https://arxiv.org/abs/1904.03257>

# SysML事例: スパコンを自動チューニング

## ベイズ最適化によるパラメータ探索

CONFERENCE (INTERNATIONAL)

### Bayesian Optimization of HPC Systems for Energy Efficiency

Takashi Miyazaki, Issei Sato (UTokyo), [Nobuyuki Shimizu](#)

International Supercomputing Conference (ISC 2018), 2018/6

Category:

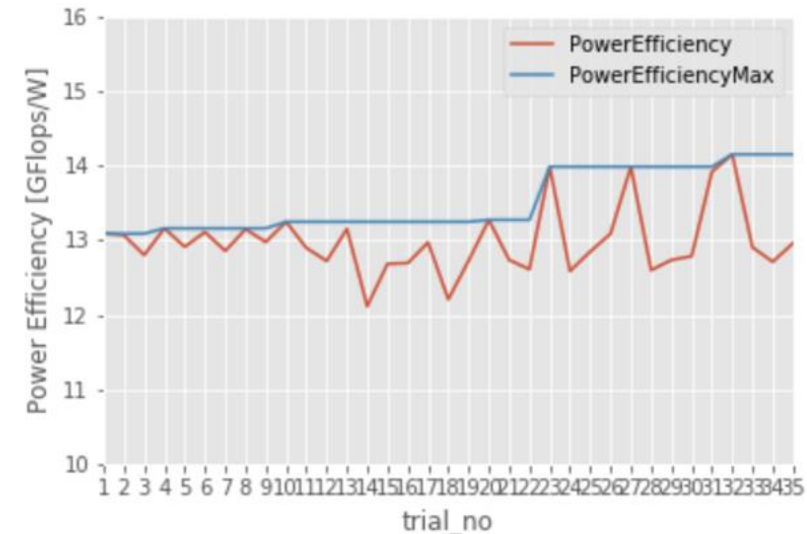
機械学習 (Machine Learning)

その他の取り組み (Misc.)

Abstract:

Energy efficiency is a crucial factor in developing large supercomputers and cost-effective datacenters. However, tuning a system for energy efficiency is difficult because the power and performance are conflicting demands. We applied Bayesian optimization (BO) to tune a graphics processing unit (GPU) cluster system for the benchmark used in the Green500 list, a popular energy-efficiency ranking of supercomputers. The resulting benchmark score enabled our system, named "kukai", to earn second place in the Green500 list in June 2017, showing that BO is a useful tool. By determining the search space with minimal knowledge and preliminary experiments beforehand, BO could automatically find a sufficiently good configuration. Thus, BO could eliminate laborious manual tuning work and reduce the occupancy time of the system for benchmarking. Because BO is a general-purpose method, it may also be useful for tuning any practical applications in addition to Green500 benchmarks.

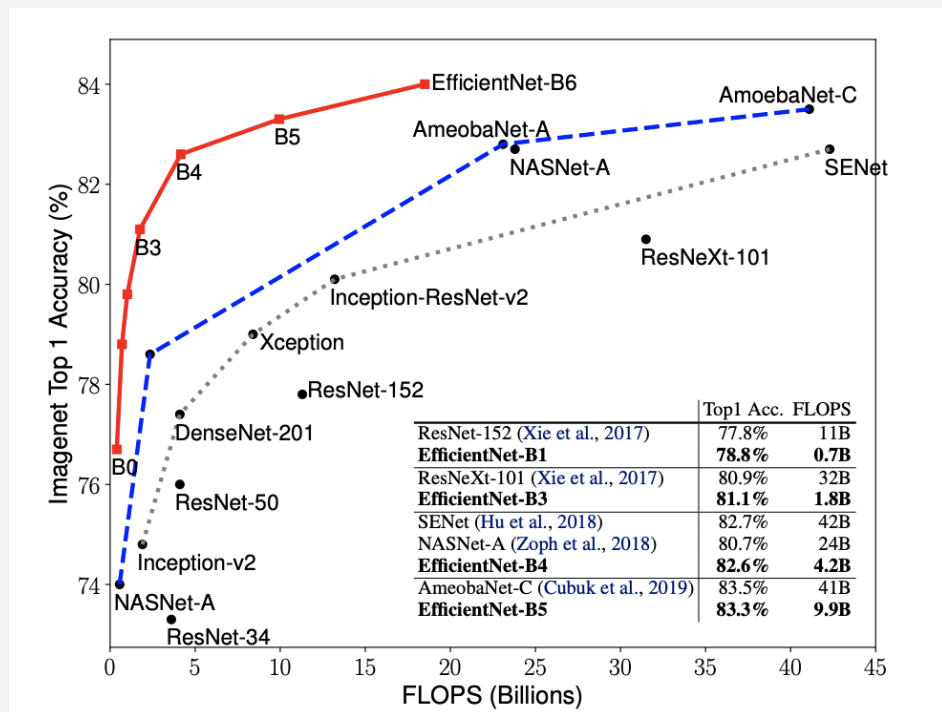
[https://research-lab.yahoo.co.jp/ml/20180624\\_miyazaki.html](https://research-lab.yahoo.co.jp/ml/20180624_miyazaki.html)



Parameter	BO	Values	Description
N	✓	Positive Integer	Matrix Size
NB	✓	$1 \leq NB \leq N$	Block Size
P		$P \cdot Q = \#Processes$	Process Grid Size
Q		$P \cdot Q = \#Processes$	Process Grid Size
NBMIN	✓	$1 \leq NBMIN \leq NB$	Recursive Stopping Condition
NDIV		Positive Integer	Number of Panels in Recursion
PFACT		3 choices	Number of Panel Fact
RFACT		3 choices	Number of Recursive Panel Fact
BCAST	✓	6 choices	Broadcast Type
DEPTH		$0 \leq DEPTH$	Lookahead Depth
SWAP		3 choices	Swapping Algorithm
GPU_CLK	✓	1189 ~ 1328	GPU Clock Frequency
CPU_CLK		1200 ~ 1700	CPU Clock Frequency

# AutoML

## 画像認識



<https://arxiv.org/pdf/1905.11946.pdf>

## 機械翻訳

Model	Params	BLEU	SacreBLEU (Post, 2018)
Gehring et al. (2017)	216M	25.2	-
Vaswani et al. (2017)	213M	28.4	-
Ahmed et al. (2017)	213M	28.9	-
Chen et al. (2018)	379M	28.5	-
Shaw et al. (2018)	213M	29.2	-
Ott et al. (2018)	210M	29.3	28.6
Wu et al. (2019)	213M	29.7	-
<b>Evolved Transformer</b>	<b>218M</b>	<b>29.8</b>	<b>29.2</b>

<https://arxiv.org/pdf/1901.11117.pdf>

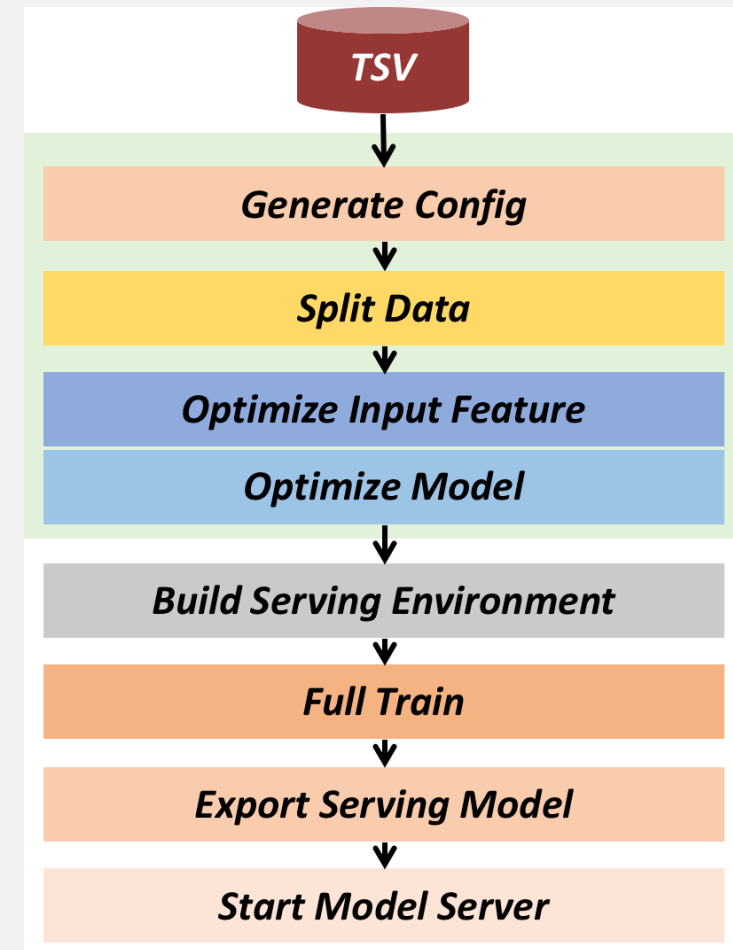
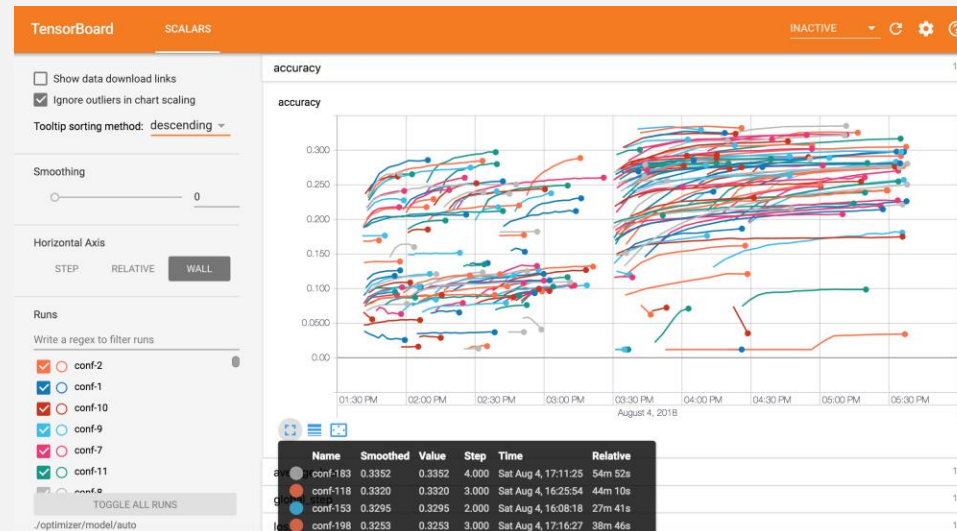
# 内製AutoML: Hilbert

ヘッダ付きTSVを用意するだけで全て自動化

- Config 生成
- データセット分割(Train/Eval/Test)
- 入力データの組み合わせ最適化(Feature Engineering)
- ハイパーパラメータ最適化
- モデルタイプ最適化

一部サービスで適用

TensorFlow World  
(2019/10/28-31)  
にて発表



# まとめ

# まとめ

- ビッグデータブーム以前よりプラットフォーム構築に取り組んでいた
- ニーズや変化に応じ、再設計/再構築
- 今後、データを最大限活かすAIに注力

# Thanks!

## Q&A