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」。

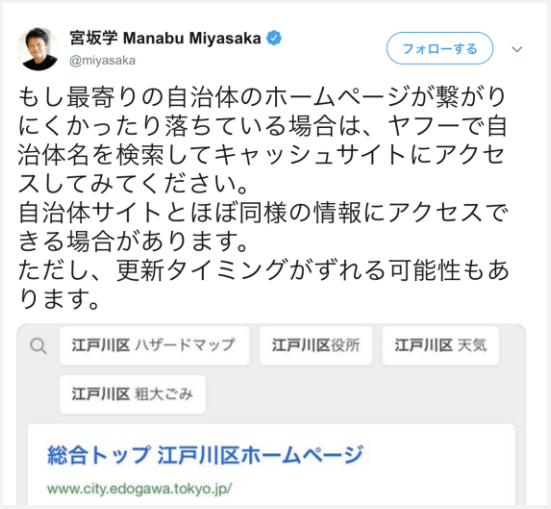
https://linotice.tumblr.com/post/164962607234/20170904

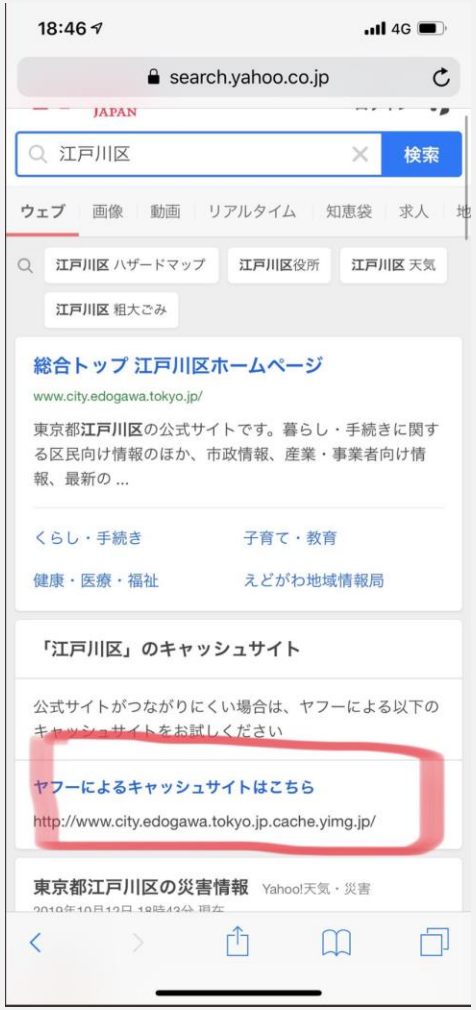
Agenda

- 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





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

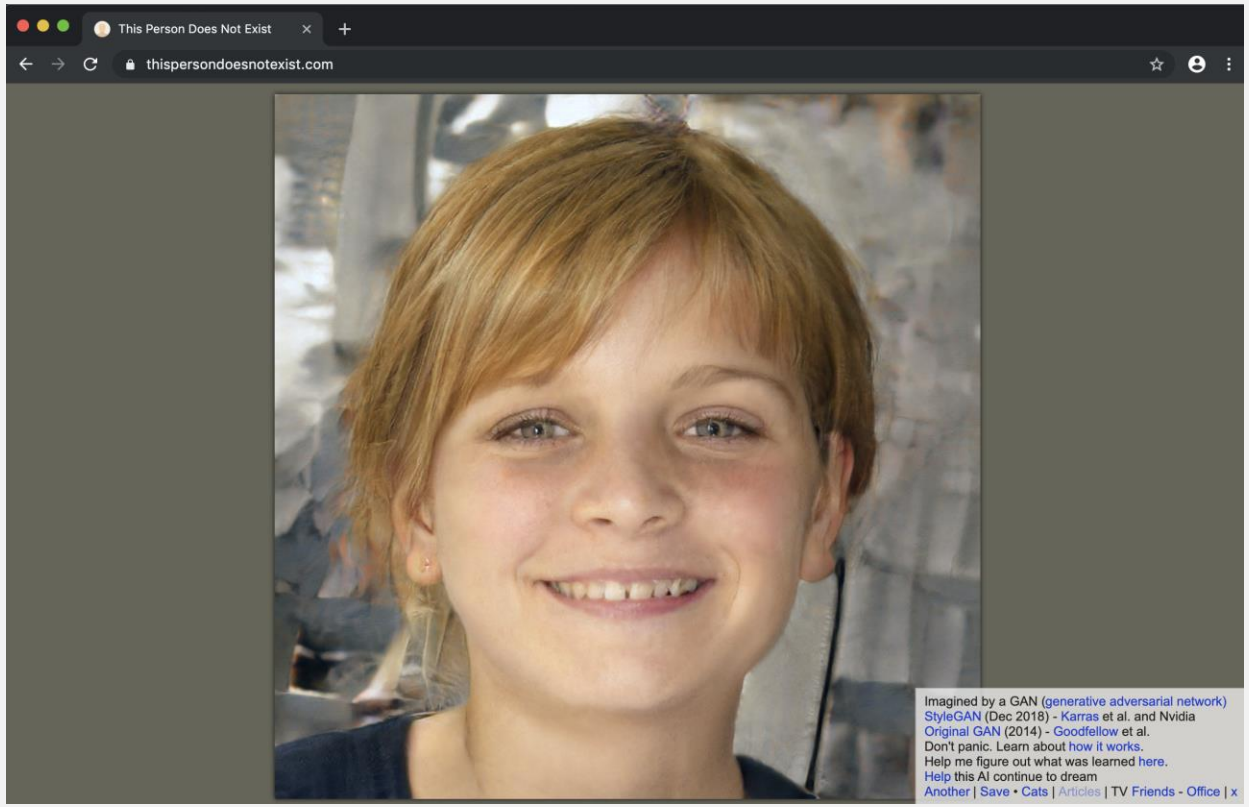
Yahoo!防災速報

災害への支援募金

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

AI・データ事例

画像、音声



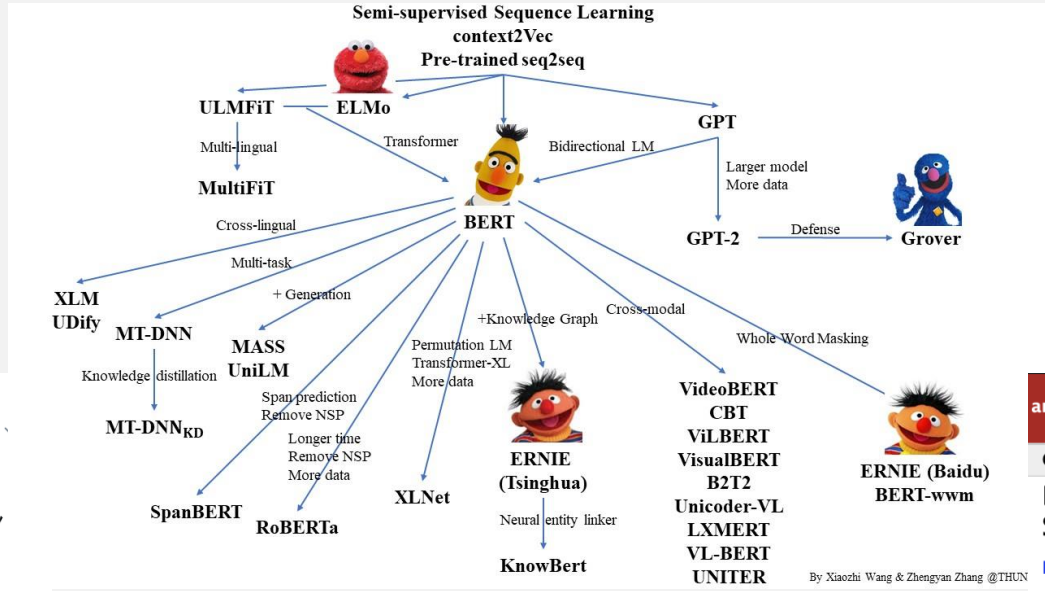
https://thispersondoesnotexist.com/

https://www.youtube.com/watch?v=nOLuI7nPQWU

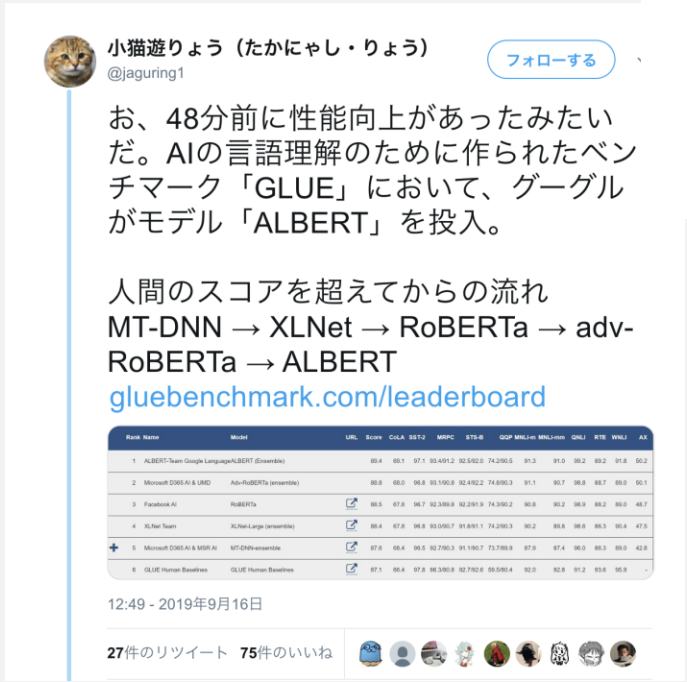
VideoIn Ads

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

自然言語処理



https://github.com/thunlp/PLMpapers



https://twitter.com/jaguring1/status/1173685273252622336

https://arxiv.org/abs/1907.06616

B-TreeをNNで置き換え

B-Treeより省メモリで70%高速

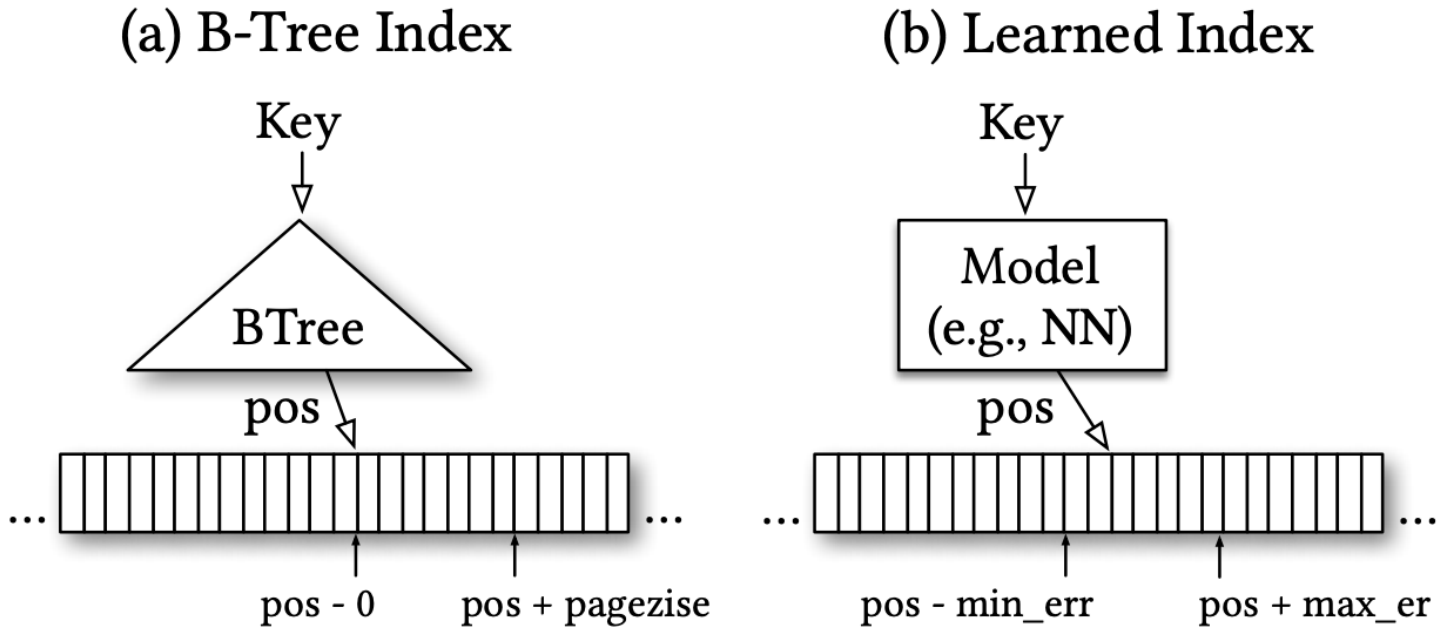


Figure 1: Why B-Trees are models

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

Voice Recognition Engine **YJVOICE**

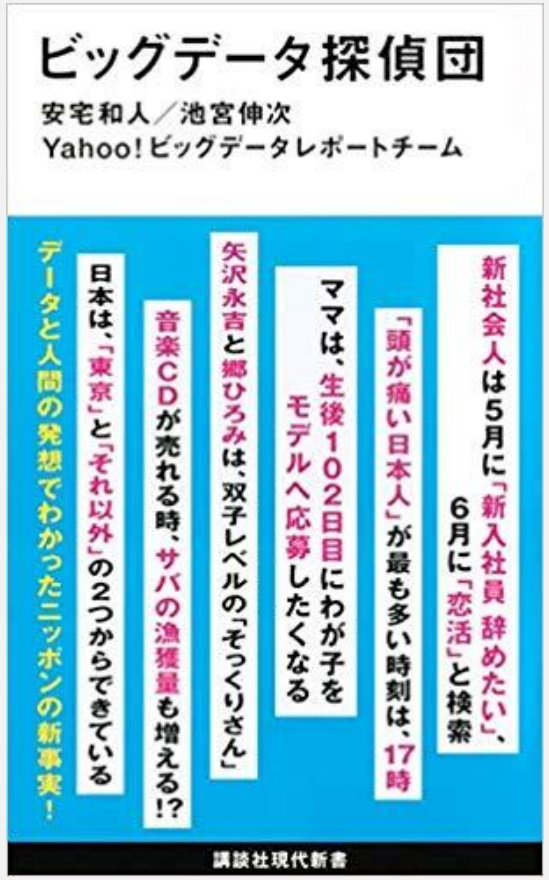
Article Title's Auto Generation

Smart Image Cropping

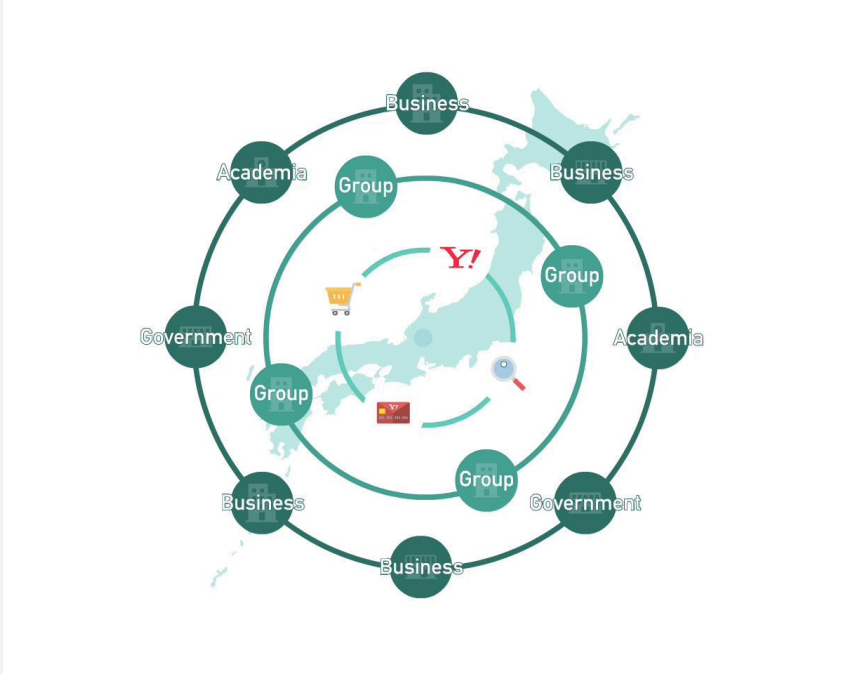
Personalization, Deduplication of Articles

Forecasting PageViews

データで得た知見を開放



DATA FOREST

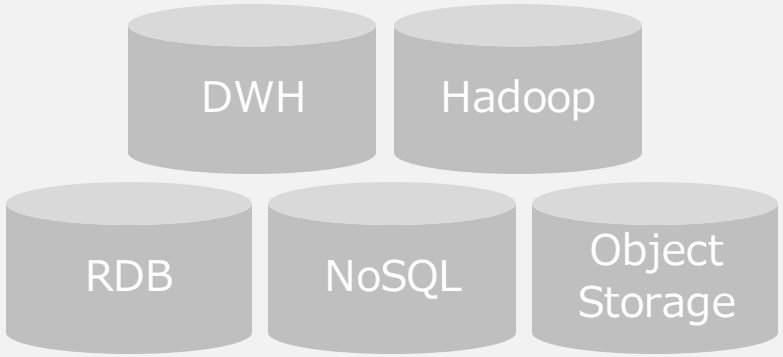


https://dataforest.yahoo.co.jp/

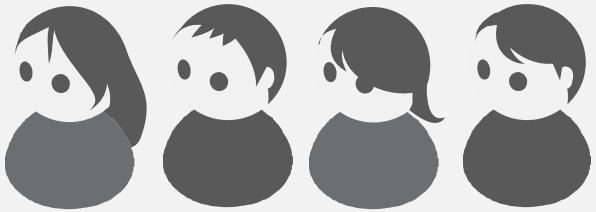
大規模データ処理プラットフォームの必要性

100以上のサービス

ビッグデータ



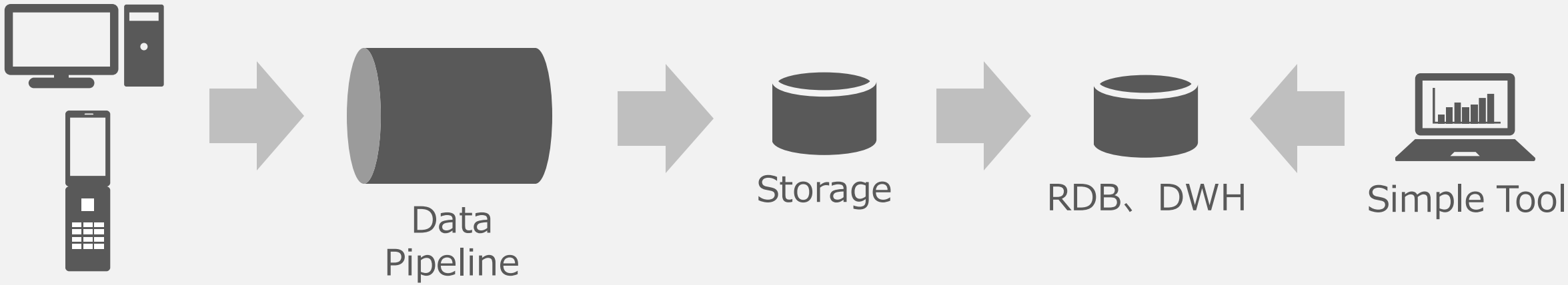
データサイエンティスト



データプラットフォーム変遷

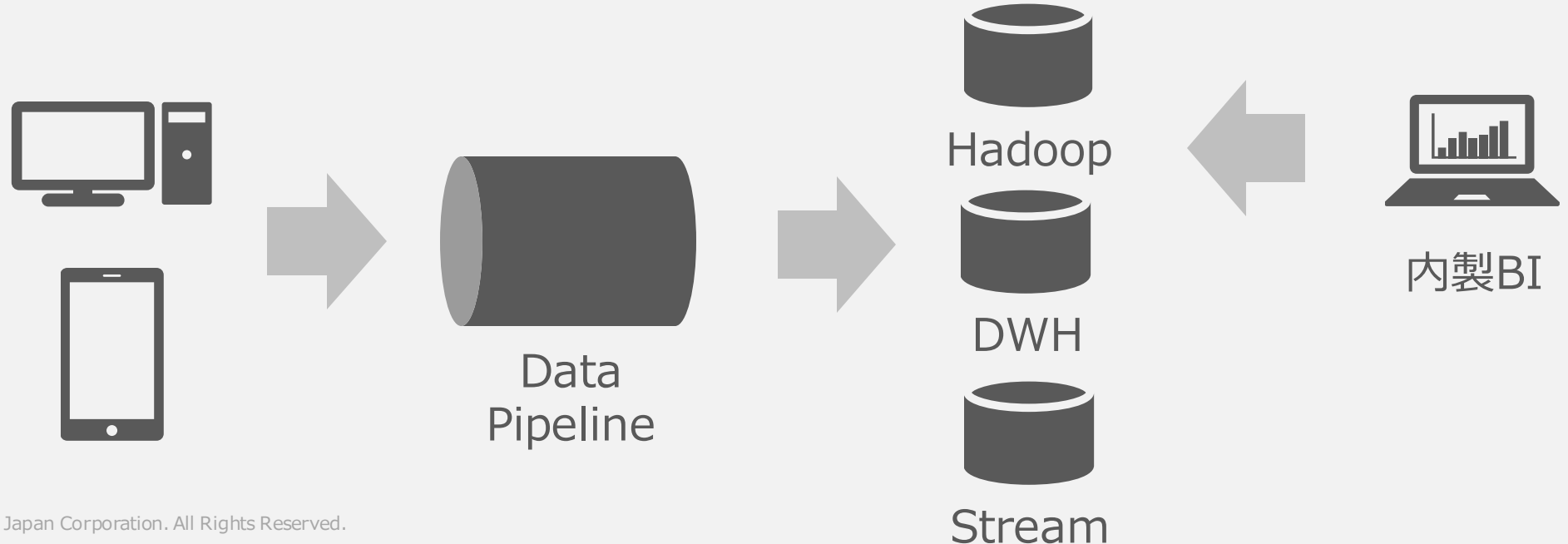
～200X

- 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



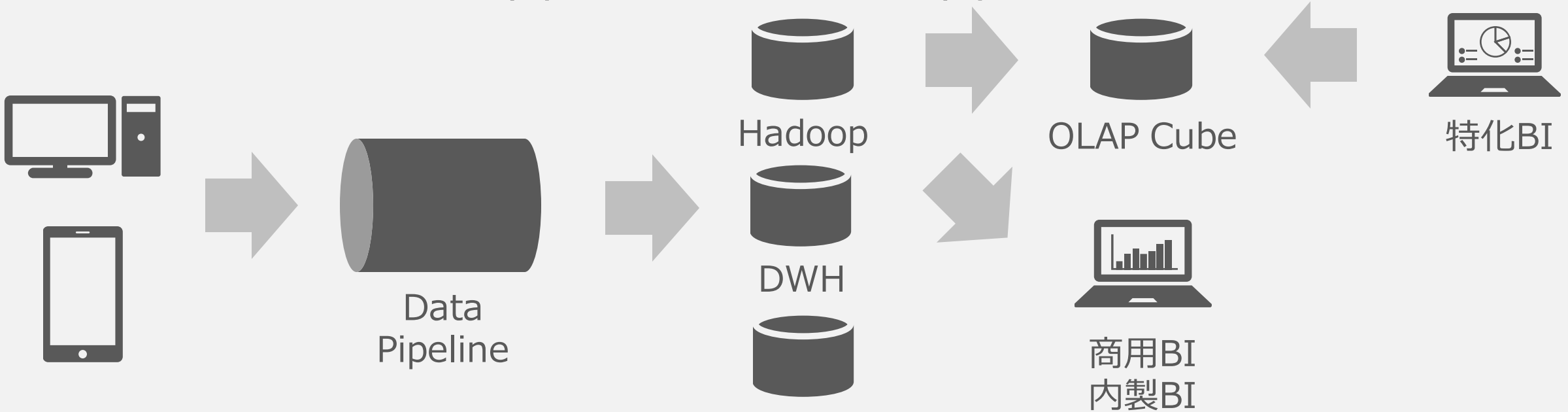
2010~2013

- 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



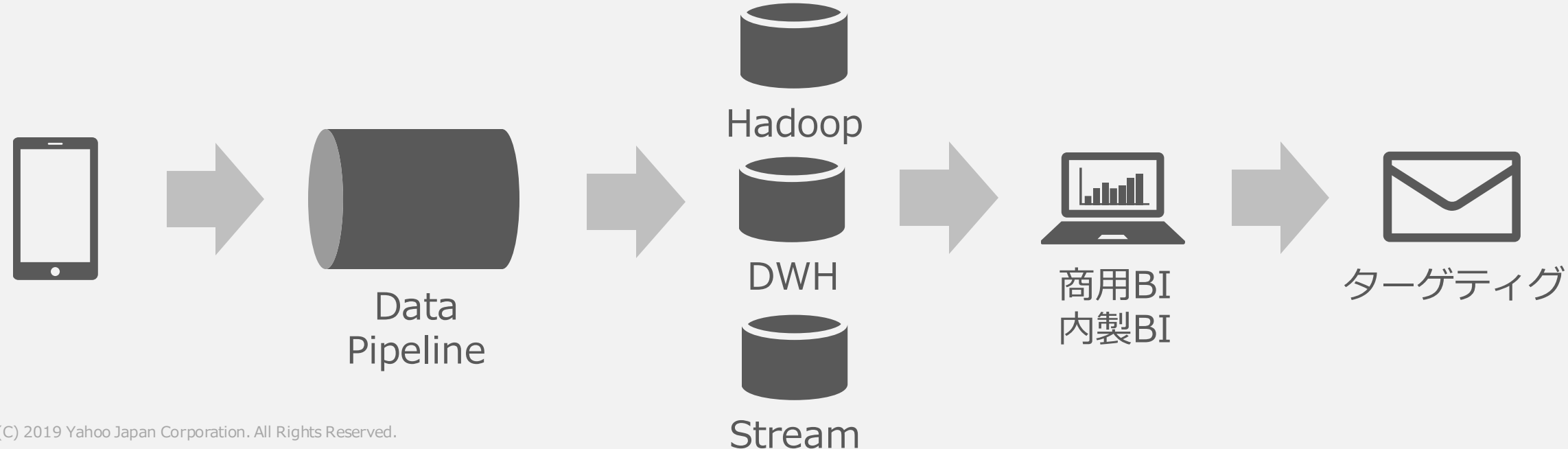
2014~2017

- 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

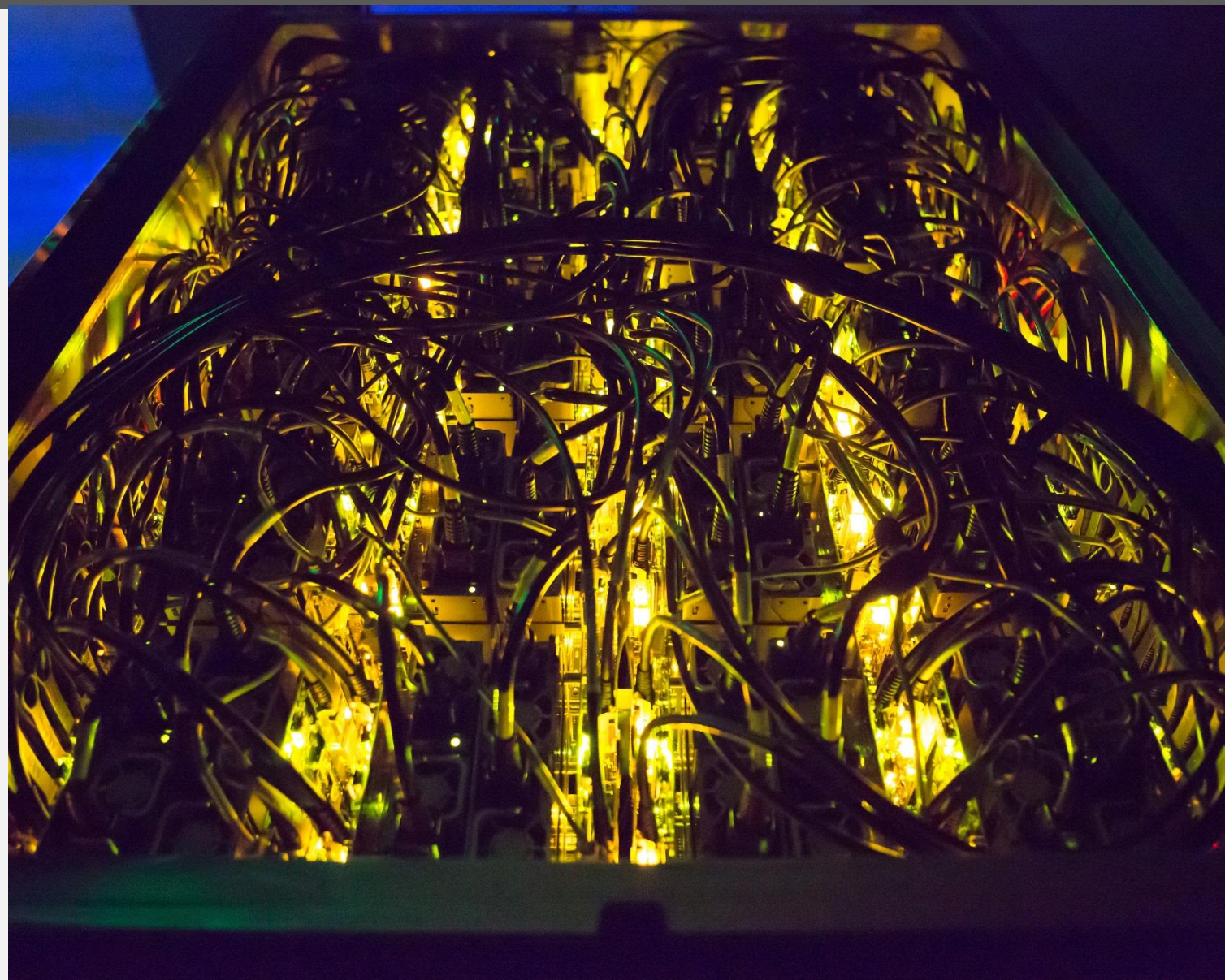
Deep Learning向けスパコンの開発

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
ベ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）

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の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

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

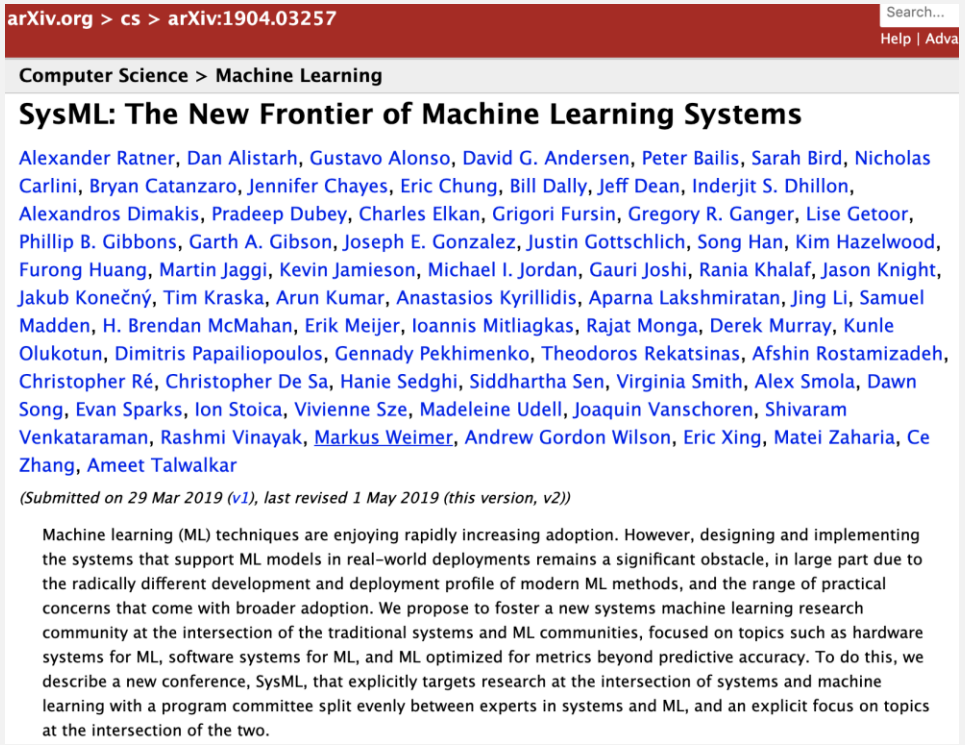
SysML

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 | Adva

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

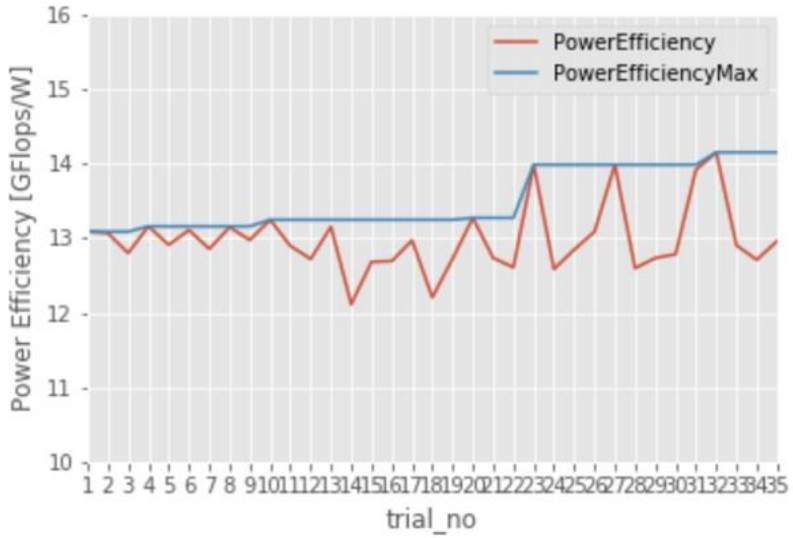
Category:

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.

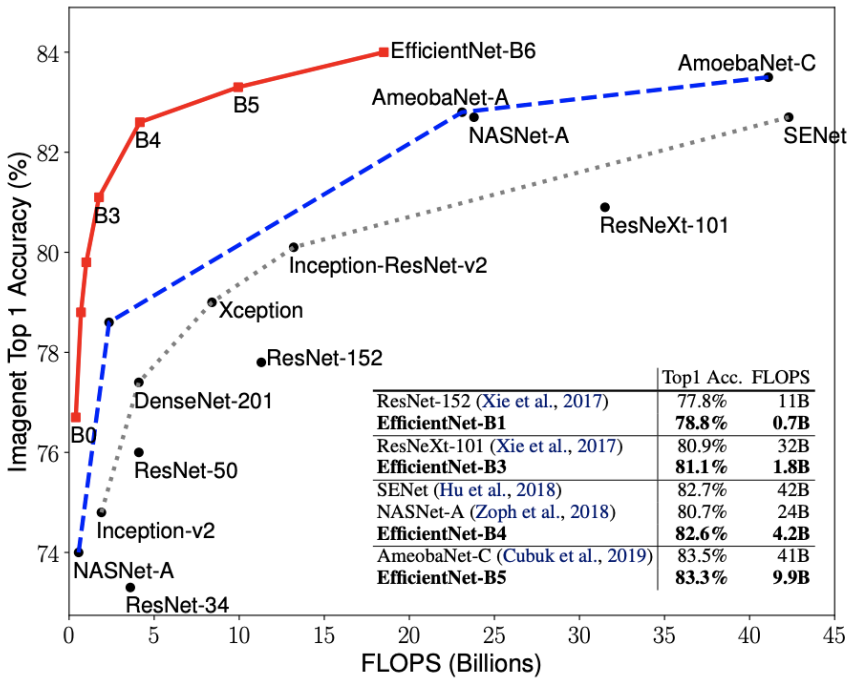
https://research-lab.yahoo.co.jp/ml/20180624_miyazaki.html



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	✓	$1189 \sim 1328$	GPU Clock Frequency
CPU_CLK		$1200 \sim 1700$	CPU Clock Frequency

AutoML

画像認識



	Top1 Acc.	FLOPS
ResNet-152 (Xie et al., 2017)	77.8%	11B
EfficientNet-B1	**78.8%**	**0.7B**
ResNeXt-101 (Xie et al., 2017)	80.9%	32B
EfficientNet-B3	**81.1%**	**1.8B**
SENet (Hu et al., 2018)	82.7%	42B
NASNet-A (Zoph et al., 2018)	80.7%	24B
EfficientNet-B4	**82.6%**	**4.2B**
AmeobaNet-C (Cubuk et al., 2019)	83.5%	41B
EfficientNet-B5	**83.3%**	**9.9B**

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

機械翻訳

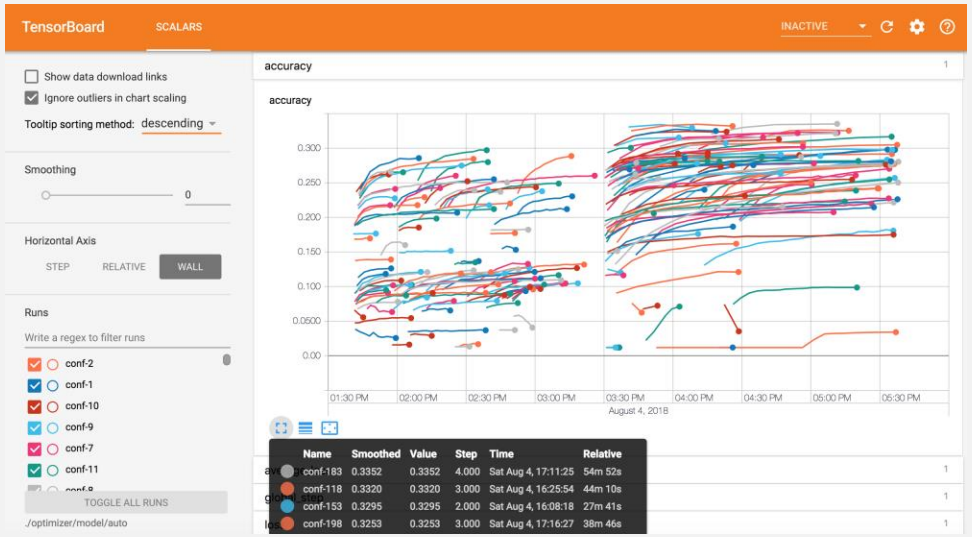
Model	Params	BLEU	SacreBLEU (Post, 2018)
Gehring et al. (2017)	216M	25.2	-
Vaswani et al. (2017)	213M	28.4	-
Ahmed et al. (2017)	213M	28.9	-
Chen et al. (2018)	379M	28.5	-
Shaw et al. (2018)	213M	29.2	-
Ott et al. (2018)	210M	29.3	28.6
Wu et al. (2019)	213M	29.7	-
Evolved Transformer	218M	**29.8**	**29.2**

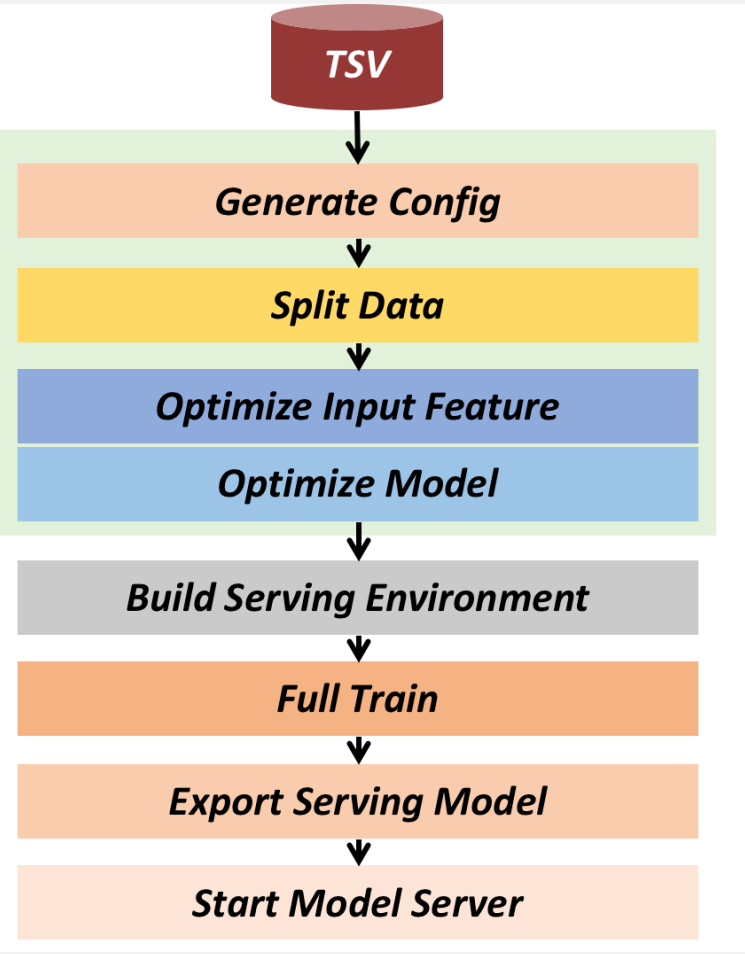
https://arxiv.org/pdf/1901.11117.pdf

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