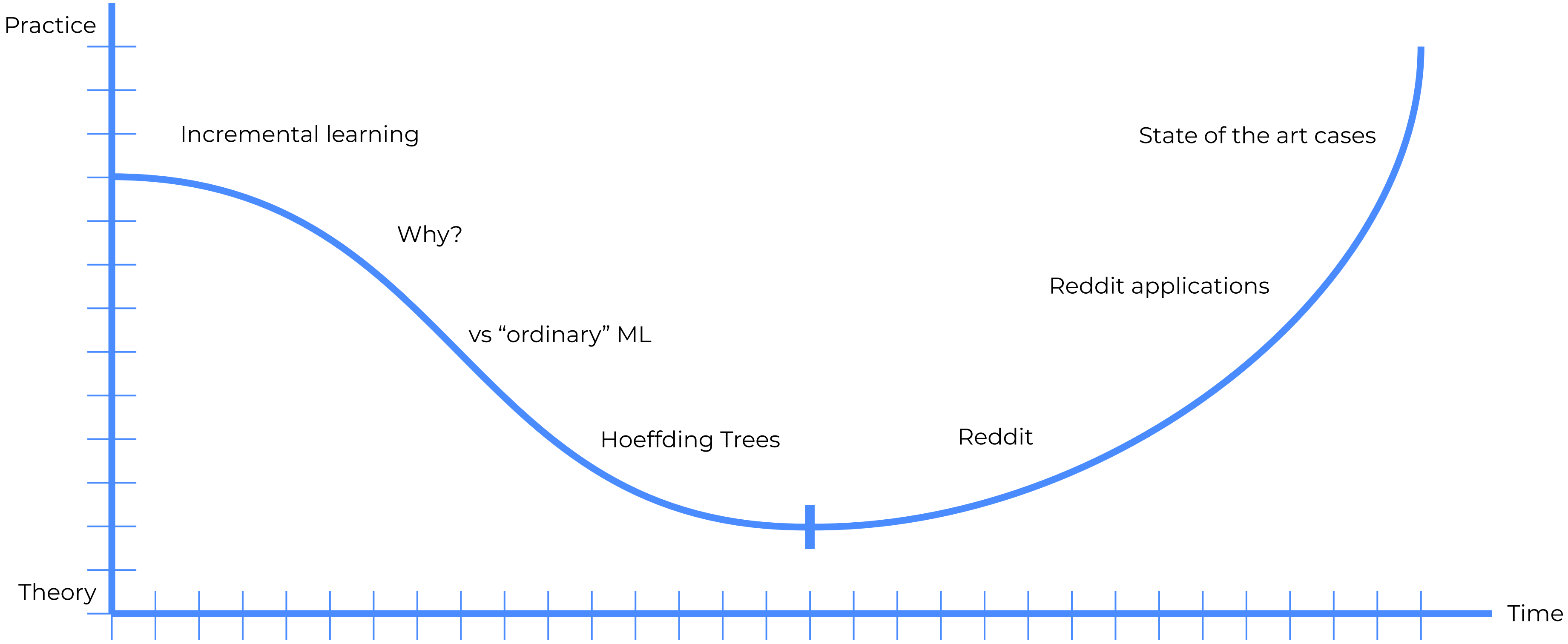
IncRedimental

a talk about incremental machine learning and applications of Reddit dataset

Jan Sawicki

What will we talk about?



VIRAL

Virality of Information on Reddit Analysed Live

Jan Sawicki

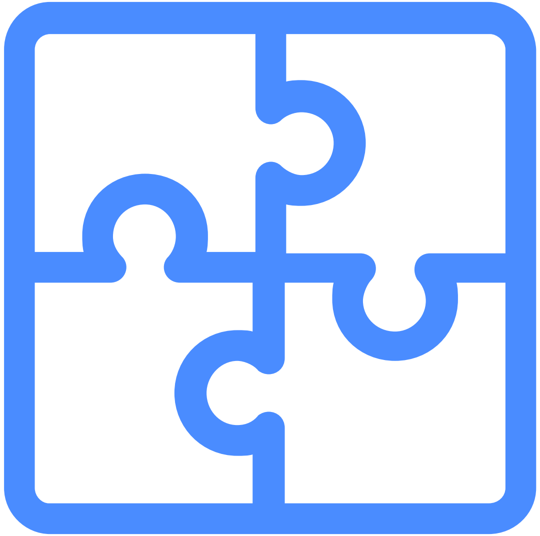
The Idea

- What make things go trendy?
- Is sexieness universal regardless of people interest/domain?
- How to predict if something goes viral?
- How do virality determinants change over time?

The domain(s) of the research

Natural Language Processing

Machine learning



?

Big Data

VIRAL

Virality of Information on Reddit Analysed Live

Jan Sawicki

VIRAL

Virality of Information on Reddit Analysed Live

Jan Sawicki

Agenda

Online learning

aka incremental
learning

Reddit

TUDFE
(The Ultimate Dataset
For Everything)

01

Incremental learning

A completely different approach to machine learning

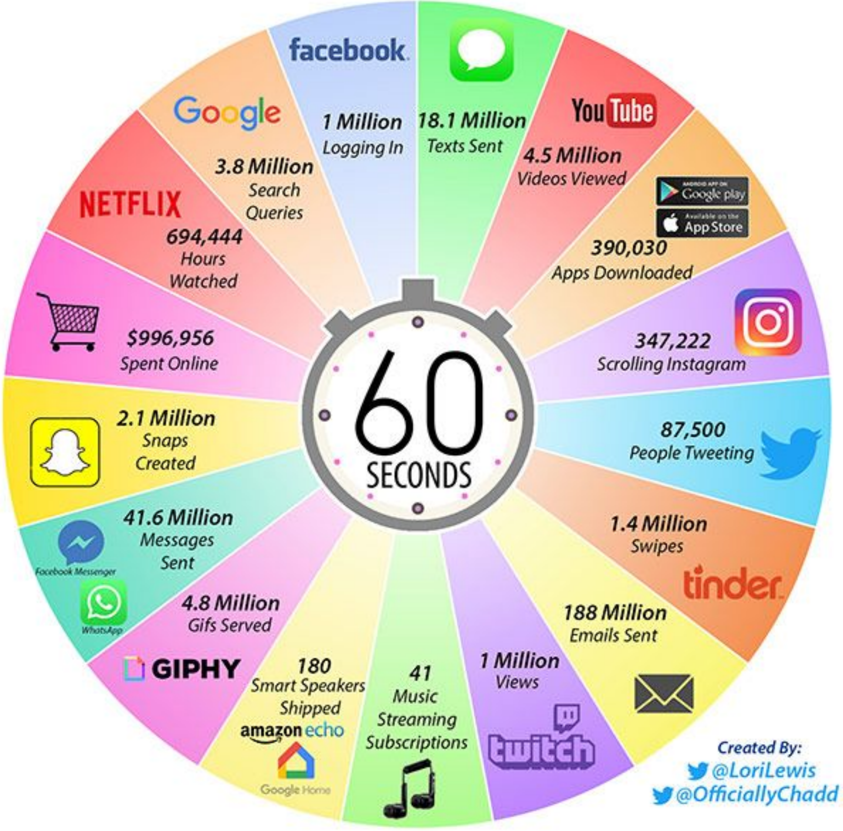
Panta rhei

2018 This Is What Happens In An Internet Minute

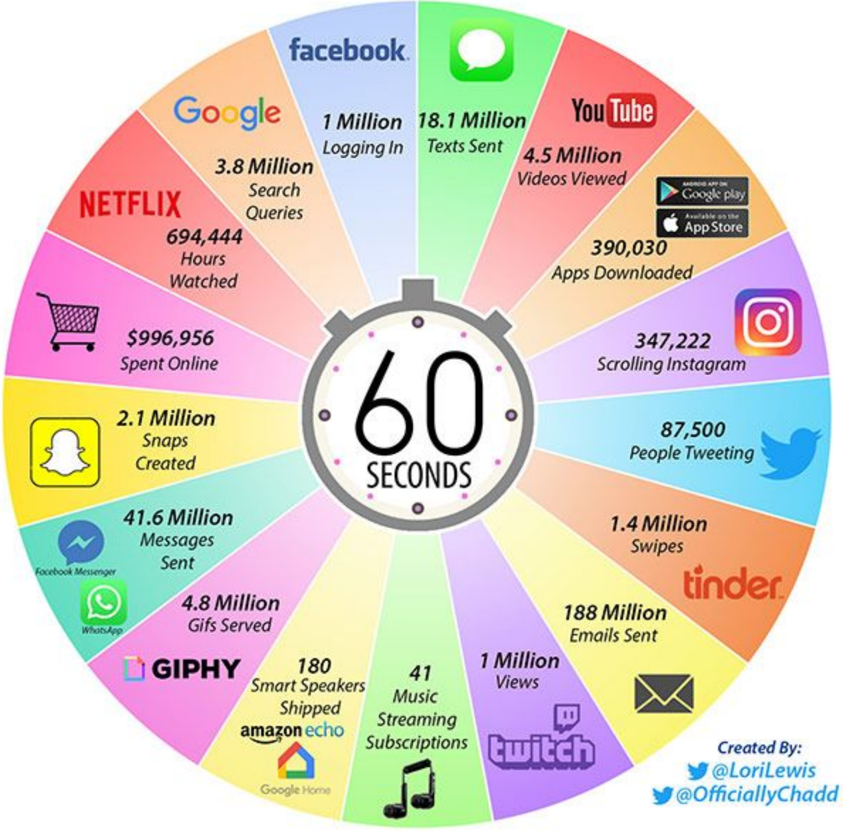
Created By:
@LoriLewis
@OfficiallyChadd

2018 This Is What Happens In An Internet Minute

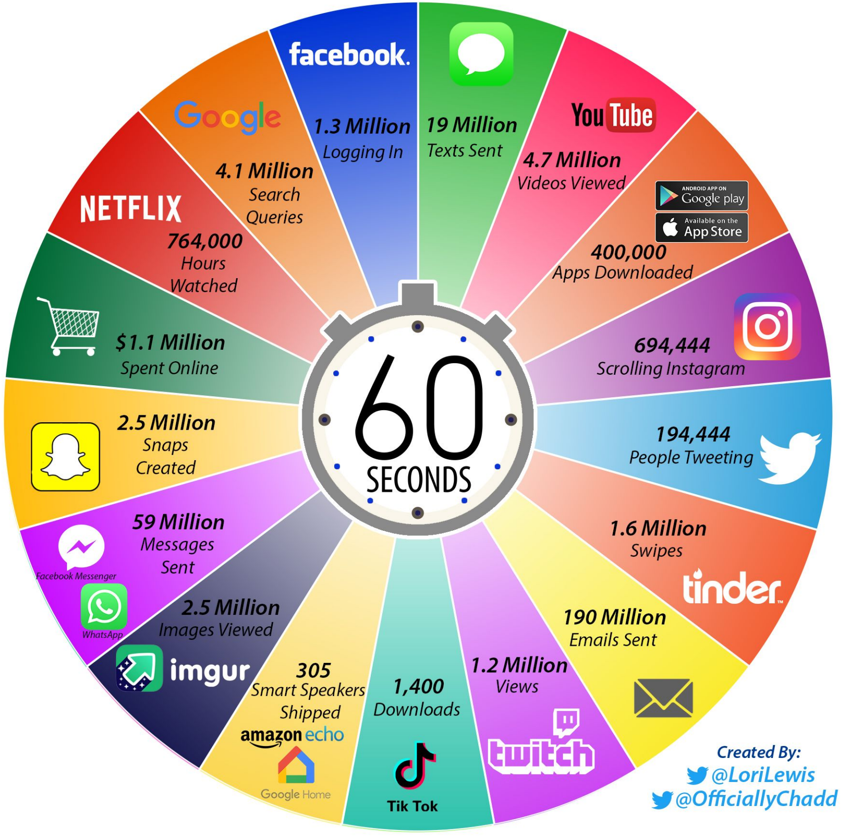
2019 This Is What Happens In An Internet Minute



2019 This Is What Happens In An Internet Minute



2020 This Is What Happens In An Internet Minute



Incremental Learning

“Incremental learning refers to the situation of continuous model adaptation based on a constantly arriving data stream”

Gepperth, Alexander, and Barbara Hammer. "Incremental learning algorithms and applications." In European symposium on artificial neural networks (ESANN). 2016.

Online Learning

“Incremental learning refers to the situation of continuous model adaptation based on a constantly arriving data stream”

+

The model does not “store” the previous data

Saffari, Amir, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. "On-line random forests." In 2009 ieee 12th international conference on computer vision workshops, iccv workshops, pp. 1393-1400. IEEE, 2009.

Why incremental?

Dataset availability

Not everything is available at our whim

Concept drift

Things change. A lot.

Evolution

Models are more ‘flexible’

Learning time (?)

No more countless hours of GPU computation and still getting 50% accuracy

Yang, Qing, Yudi Gu, and Dongsheng Wu. **"Survey of incremental learning."** In 2019 Chinese Control And Decision Conference (CCDC), pp. 399-404. IEEE, 2019.

Read, Jesse, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. **"Batch-incremental versus instance-incremental learning in dynamic and evolving data."** In International symposium on intelligent data analysis, pp. 313-323. Springer, Berlin, Heidelberg, 2012.

Shen, Wei-Min. **Efficient Incremental Induction of Decision Lists. Can Incremental Learning Outperform Non-Incremental Learning?**. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1996.

Why not incremental?

Reproducibility

How do we reproduce if it is still running?
How to reproduce if it is

Data processing

How to process if we do not know the input? (e.g. embeddings)

Parameters tuning

How to tune for something we do not know?

Stability-plasticity dilemma

How to adjust if we don't know what to adjust for?

Yang, Qing, Yudi Gu, and Dongsheng Wu. **"Survey of incremental learning."** In 2019 Chinese Control And Decision Conference (CCDC), pp. 399-404. IEEE, 2019.

Read, Jesse, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. **"Batch-incremental versus instance-incremental learning in dynamic and evolving data."** In International symposium on intelligent data analysis, pp. 313-323. Springer, Berlin, Heidelberg, 2012.

Shen, Wei-Min. **Efficient Incremental Induction of Decision Lists. Can Incremental Learning Outperform Non-Incremental Learning?**. UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY INFORMATION SCIENCES INST, 1996.

Applications

Robotics

Outlier detection

Image processing

Medical field

Michalski, Ryszard S., Igor Mozetic, Jiarong Hong, and Nada Lavrac. "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains." Proc. AAAI 1986 (1986): 1-041.

Gepperth, Alexander, and Barbara Hammer. "Incremental learning algorithms and applications." In European symposium on artificial neural networks (ESANN). 2016.

Methods

Random forest

Incremental random forest

SVM

Incremental SVM

Naive Bayes

Incremental naive Bayes

Neural network

Incremental neural network

Yang, Qing, Yudi Gu, and Dongsheng Wu. "Survey of incremental learning." In 2019 Chinese Control And Decision Conference (CCDC), pp. 399-404. IEEE, 2019.

Example: Hoeffding Tree

Equation 1: $\epsilon = \sqrt{\frac{R^2 ln \frac{1}{\delta}}{2n}}$

HoeffdingTree (S, X, G, δ)
Let HT be a tree with a single leaf l_1 (the root).
Let $X_1 = X \cup \{X_\varnothing\}$.
Let $G_1(X_\varnothing)$ be the G obtained by predicting the most frequent class in S.
For each class y_k
    For each value x_{ij} of each attribute $X_i \in X$
        Let $n_{ijk}(l_1) = 0$
For each example (x, y_k) in S
    Sort (x, y) into a leaf l using HT.
    For each x_{ij} in x such that $X_i \in X_l$
        Increment $n_{ijk}(l)$.
    Label l with the majority class among the examples seen so far at l.
    If the examples seen so far at l are not all of the same class, **then**
        Compute $G_l(X_i)$ for each attribute $X_i \in X_l - \{X_\varnothing\}$ using the counts $n_{ijk}(l)$.
        Let X_a be the attribute with highest G_l.
        Let X_b be the attribute with second-highest G_l.
        Compute ϵ using Equation 1.
        If $G_l(X_a) - G_l(X_b) > \epsilon$ and $X_a \neq X_\varnothing$, **then**
            Replace l by an internal node that splits on X_a.
            For each branch of the split
                Add a new leaf l_m, and let $X_m = X - \{X_a\}$.
                Let $Gm(X_\varnothing)$ be the G obtained by predicting the most frequent class at lm.
                For each class y_k and each value x_{ij} of each attribute $X_i \in X_m - \{X_\varnothing\}$
                    Let $n_{ijk}(l_m) = 0$.
Return HT

Inputs:
S is a sequence of examples
X is a set of discrete attributes
G(.) is a split evaluation function
δ is one minus the desired probability of choosing the correct attribute at any given node
$\mathbf{n_{ijk}}$ is the sufficient statistics needed to compute most heuristic measures

Example: Hoeffding Tree

HoeffdingTree(Stream, δ)

Input: a stream of labeled examples, confidence parameter

let HT be e tree with a single leaf (root)
init counts n_{ijk} at root
for each example (x, y) in Stream
    do HTGrow((x, y), HT, δ)

HTGROw((x, y), HT, δ)
    sort (x, y) to leaf l using HT
    Update counts n_{ijk} at leaf l
    if examples seen so far at l are not all of the same class **then**
        compute G for each attribute
    if G(best attribute) - G(second best) > $\sqrt{\frac{R^2 ln \frac{1}{\delta}}{2n}}$ **then**
        split leaf on best attribute
        for each branch
            do start a new leaf and initialize counts

A bit of theory

Theorem
If HT_δ is the tree produced by the Hoeffding tree algorithm with desired probability δ given infinite examples, DT_* is the asymptotic batch tree, and p is the leaf probability, then $\mathbf{E[\Delta_i(HT_\delta, DT_*)] \leq \delta/p}$.

Definition
The **extensional disagreement** $\mathbf{\Delta_e}$ between two decision trees DT_1 and DT_2 is the probability that they will produce different class predictions for an example:

$\mathbf{\Delta_e(DT1, DT2)} = \sum P(x) I[DT_1(x) \neq DT_2(x)]$

Definition
The **intensional disagreement** $\mathbf{\Delta_i}$ between two decision trees DT_1 and DT_2 is the probability that the path of an example through DT_1 will differ from its path through DT_2:

$\mathbf{\Delta_i(DT1, DT2)} = \sum P(x) I[Path_1(x) \neq Path_2(x)]$

I is and indicator function which return 1 (agree) or 0 (disagree)

$\mathbf{\Delta_i(DT1, DT2) >= \Delta_e(DT1, DT2)}$

Domingos, Pedro, and Geoff Hulten. "Mining high-speed data streams." In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 71-80. 2000.

Let p_l be the probability that an example that reaches level l in a decision tree falls into a leaf at that level. To sim-

Proof. For brevity, we will refer to intensional disagreement simply as disagreement. Consider an example $\mathbf{x}$ that falls into a leaf at level l_h in HT_δ, and into a leaf at level l_d in DT_*. Let $l = \min\{l_h, l_d\}$. Let $\text{Path}_H(\mathbf{x}) = (N_1^H(\mathbf{x}), N_2^H(\mathbf{x}), \ldots, N_l^H(\mathbf{x}))$ be $\mathbf{x}$'s path through HT_δ up to level l, where $N_i^H(\mathbf{x})$ is the node that $\mathbf{x}$ goes through at level i in HT_δ, and similarly for $\text{Path}_D(\mathbf{x})$, $\mathbf{x}$'s path through DT_*. If $l = l_h$ then $N_l^H(\mathbf{x})$ is a leaf with a class prediction, and similarly for $N_l^D(\mathbf{x})$ if $l = l_d$. Let I_i represent the proposition "$\text{Path}_H(\mathbf{x}) = \text{Path}_D(\mathbf{x})$ up to and including level i," with $I_0 =$ True. Notice that $P(l_h \neq l_d)$ is included in $P(N_l^H(\mathbf{x}) \neq N_l^D(\mathbf{x}) | I_{l-1})$, because if the two paths have different lengths then one tree must have a leaf where the other has an internal node. Then, omitting the dependency of the nodes on $\mathbf{x}$ for brevity,

$$
\begin{aligned}
&P(\text{Path}_H(\mathbf{x}) \neq \text{Path}_D(\mathbf{x})) \\
&\quad = P(N_1^H \neq N_1^D \vee N_2^H \neq N_2^D \vee \ldots \vee N_l^H \neq N_l^D) \\
&\quad = P(N_1^H \neq N_1^D | I_0) + P(N_2^H \neq N_2^D | I_1) + \ldots \\
&\qquad + P(N_l^H \neq N_l^D | I_{l-1}) \\
&\quad = \sum_{i=1}^{l} P(N_i^H \neq N_i^D | I_{i-1}) \;\leq\; \sum_{i=1}^{l} \delta \;=\; \delta l
\end{aligned}
\tag{2}
$$

Let $HT_\delta(S)$ be the Hoeffding tree generated from training sequence S. Then $E[\Delta_i(HT_\delta, DT_*)]$ is the average over all infinite training sequences S of the probability that an example's path through $HT_\delta(S)$ will differ from its path through DT_*:

$$
\begin{aligned}
&E[\Delta_i(HT_\delta, DT_*)] \\
&\quad = \sum_S P(S) \sum_{\mathbf{x}} P(\mathbf{x})\, I[\text{Path}_H(\mathbf{x}) \neq \text{Path}_D(\mathbf{x})] \\
&\quad = \sum_{\mathbf{x}} P(\mathbf{x})\, P(\text{Path}_H(\mathbf{x}) \neq \text{Path}_D(\mathbf{x})) \\
&\quad = \sum_{i=1}^{\infty} \sum_{\mathbf{x} \in L_i} P(\mathbf{x})\, P(\text{Path}_H(\mathbf{x}) \neq \text{Path}_D(\mathbf{x}))
\end{aligned}
\tag{3}
$$

where L_i is the set of examples that fall into a leaf of DT_* at level i. According to Equation 2, the probability that

an example's path through $HT_\delta(S)$ will differ from its path through DT_*, given that the latter is of length i, is at most δi (since $i \geq l$). Thus

an example's path through $HT_\delta(S)$ will differ from its path through DT_*, given that the latter is of length i, is at most δi (since $i \geq l$). Thus

$$
\begin{aligned}
E[\Delta_i(HT_\delta, DT_*)] &\leq \sum_{i=1}^{\infty} \sum_{\mathbf{x} \in L_i} P(\mathbf{x})(\delta i) \\
&= \sum_{i=1}^{\infty} (\delta i) \sum_{\mathbf{x} \in L_i} P(\mathbf{x})
\end{aligned}
\tag{4}
$$

The sum $\sum_{\mathbf{x} \in L_i} P(\mathbf{x})$ is the probability that an example $\mathbf{x}$ will fall into a leaf of DT_* at level i, and is equal to $(1-p)^{i-1}p$, where p is the leaf probability. Therefore

$$
\begin{aligned}
&E[\Delta_i(HT_\delta, DT_*)] \\
&\quad \leq \sum_{i=1}^{\infty} (\delta i)(1-p)^{i-1} p \;=\; \delta p \sum_{i=1}^{\infty} i(1-p)^{i-1} \\
&\quad = \delta p \left[\sum_{i=1}^{\infty} (1-p)^{i-1} + \sum_{i=2}^{\infty} (1-p)^{i-1} + \cdots \right. \\
&\qquad \left. + \sum_{i=k}^{\infty} (1-p)^{i-1} + \cdots \right] \\
&\quad = \delta p \left[\frac{1}{p} + \frac{1-p}{p} + \cdots + \frac{(1-p)^{k-1}}{p} + \cdots \right] \\
&\quad = \delta \left[1 + (1-p) + \cdots + (1-p)^{k-1} + \cdots \right] \\
&\quad = \delta \sum_{i=0}^{\infty} (1-p)^i \;=\; \frac{\delta}{p}
\end{aligned}
\tag{5}
$$

This completes the demonstration of Theorem 1. $\square$

tion 1, ensuring $\delta = 0.1\%$ requires 380 examples, and ensuring $\delta = 0.0001\%$ requires only 345 additional examples. An exponential improvement in δ, and therefore in expected disagreement, can be obtained with a linear increase in the number of examples. Thus, even with very small leaf probabilities (i.e., very large trees), very good agreements can be obtained with a relatively small number of examples per

But how do I get started?

Where do I find a dataset?

02

Reddit

The ultimate CATEGORIZED dataset for everything

How are things on Reddit?

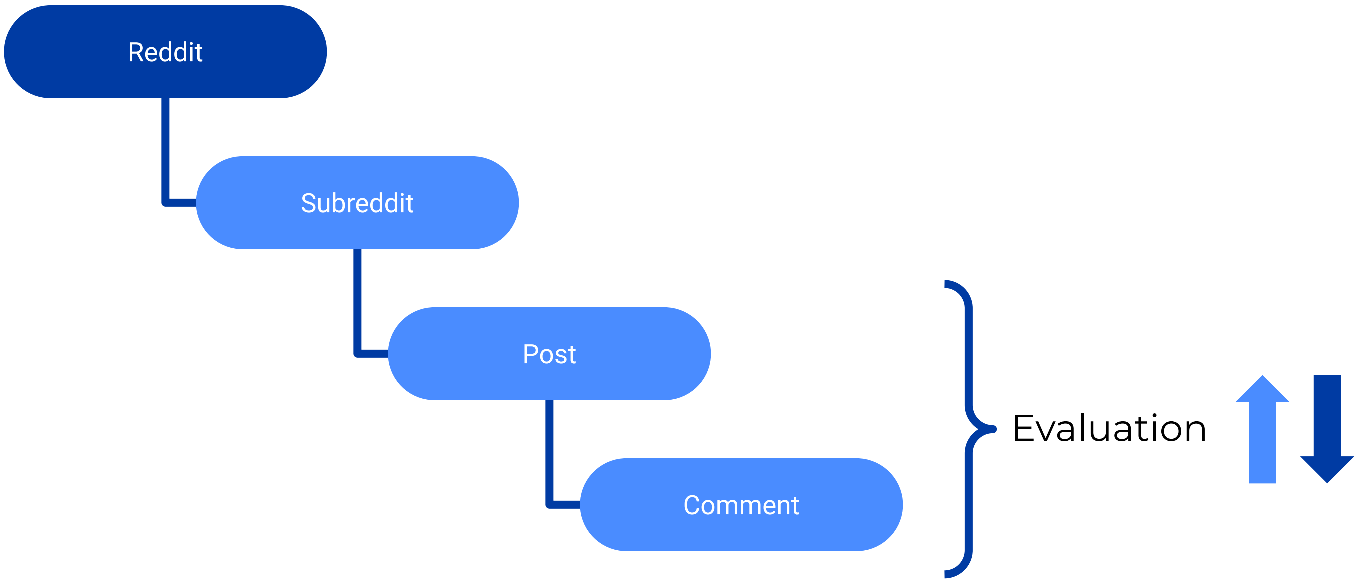
The analysis of 180 papers using Reddit (2019-2020)

Jan Sawicki

The anatomy of Reddit: An overview of academic research

- a thorough **description of the Reddit platform**
- description of **reddit-subreddit-post-comment** architecture
- analysis of sizes **discussion trees**, **scores of posts**, **social aspects**
- short comparison with **other social platforms**

Reddit structure



The charts

Main points

- Embeddings and networks
- Pushshift API instead of Reddit API
- Savvas Zannettou and Jeremy Blackburn
- Covid (obviously)
- Conversation analysis, prediction, modelling etc.
- Trend analysis uses networks

Hypothesis

Reddit is a categorized data source for any possible topic and data science task.

Experiment

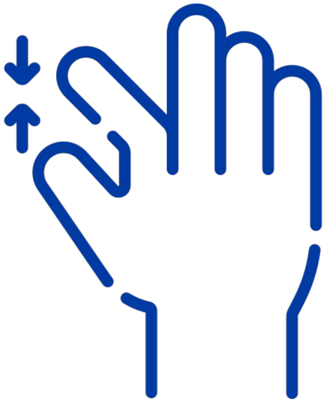
Hypothesis

Reddit is an abundant complete dataset for all possible data science tasks.

Proof (by example)

I **analyzed** manually and automatically **180** papers about Reddit from 01-01-**2019** - 10-03-**2021**

Data visualization?



Focus

“Zoomed-in” details

Context

Overall look

The Pushshift Reddit Dataset

- a whole **queryable dataset of Reddit** posts
- architecture description (**PostgreSQL**)
- data **availability**
- data **format**

Reddit ❤️ Python

Reddit API

Live feed from Reddit
(and much more!)

Pushshift API

Full archive of Reddit
with all the juicy data
we want!

Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls

- Internet **trolling** (focus on US 2016 election, Donald Trump)
- Analysing behaviours of troll
- Bots from **Russia** and **Iran**
- 10M posts from 5.5K “users” (**Twitter**, **Reddit**, **4chan**, **Gab**)
- Subreddits: uncen, funny, Bad_Cop_No_Donut, AskReddit, CryptoCurrency, PoliticalHumor, news, worldnews, gifs, aww, politics, The_Donald, racism, POLITIC, Bitcoin, copwatch, blackpower, interestingasfuck, uspolitics, newzealand,
- “Russian trolls were **pro-Trump** and Iranian trolls **anti-Trump**”
- “Russian trolls were more **efficient** and influential in **spreading URLs**”
- “automated systems to detect trolls are likely to be difficult to realize: trolls **change their behavior** over time, and thus even a classifier that works perfectly on one campaign might not catch future campaigns”
- Methods: **Hawkes Processes**, NLP (**word embedding**, hashtags analysis), **graph network** analysis

A Quantitative Approach to Understanding Online Antisemitism

- **Hypothesis** (“RQs”)
- r/**The_Donald**
- **2.6B posts** (Reddit, /pol/, Gab, and Twitter)
- Analysing **antisemitism** propaganda
- “**Meme weaponization**”
- “**Happy Merchant**” meme
- Methods: NLP (**word2vec**, **bag of words**), **changepoint analysis**, **Hawkes Processes**, graph **networks**, **SVM**, **Naive Bayes**
- “**Ethical Considerations**. During this work, we only collect publicly available data posted on /pol/ and Gab. We make no attempt to de-anonymize users and we keep the collected data in encrypted format. Overall, we follow best ethical practises as documented in “Ethical research standards in a world of big data.” ”

ELI5: Long Form Question Answering

- GENIUS idea to use r/**ELI5** for question and answers
- **“How”**, **“Why”** and **“What”** questions are most popular
- Comparisons with other QA datasets (e.g. MS MARCO v2, TriviaQA, NarrativeQA)
- Utilizing **ROUGE** metric to compare model output vs r/ELI5

How do climate change skeptics engage with opposing views?

- r/**climateskeptics**
- “Echo chambers”
- Classifying posts as “consonant” or “dissonant”
- Manual an automatic labelling
- Hypotheses
- “(...) tendency for more **senior users** to be especially **engaged** within the discussions in reaction to submissions that contain **opposing views** and dissonant information”
- “users who **engaged with opposing views** were more likely to **return** to the forum than those **engaging with attitude confirming** skeptic content”
- “most important finding of this study is, that in contrast to the classical theory of echo chambers, ‘**breaking up the echo chamber’ with information on the consequences of climate change does not seem to work**”

Similar: “No Echo in the Chambers of Political Interactions on Reddit”

“TABLE I: Comparative evaluation results for three datasets. We report micro-averaged F1 scores. “-” signifies no results are published for the given setting”

Methods	Pubmed	Reddit		PPI	
	Sup. F1	Unsup. F1	Sup. F1	Unsup. F1	Sup. F1
GCN	0.875	-	0.930	-	0.865
FastGCN	0.880	-	0.937	-	0.607
GAT	0.883	-	0.950	-	0.973
GraphSAGE-GCN	0.849	0.908	0.930	0.465	0.500
GraphSAGE-mean	0.888	0.897	0.950	0.486	0.598
RGCN-LSTM	**0.908**	0.919	0.963	0.791	0.992
RGCN-GRU	0.900	0.915	**0.964**	0.765	0.991
RGAT-LSTM	0.905	**0.921**	**0.964**	**0.806**	**0.994**
RGAT-GRU	0.902	0.913	**0.964**	0.791	0.994

- GCN are NN which can take graphs as input and perform different task like classification, labelling etc. on node level, edge level etc.
- Comparing **graph embedding + RNN** with **graph + RGNN**
- Used in tandem with **DeepWalk** (graph embedding algorithm, quire slow)
- Datasets: **Pubmed**, **Reddit** (unspecified), **PPI** (bioinformatics dataset with proteins)
- Comparing GNN in **supervised** and **unsupervised** setting
- Test network types: **GCN**, **RGCN**, **RGAT**
- “Our results demonstrate that GNN models with **recurrent units are much easier to extend to deeper models than GNN models with residual connections**. In our further analyses, we show RGNN models are more robust to noisy information from graph structure as well as local features.”

Similar:
“Grounded conversation generation as guided traverses in commonsense knowledge graphs”

Live experiment

Hypothesis

Reddit is an abundant complete dataset for all possible data science tasks.

Proof (by example)

Let's see if we can find something interesting for YOU.

The concept drift

(**Reddit evaluation process** + **IML**) + (**Scientific research**+ **IML**) = 🥀

Finis

Jan Sawicki

j.sawicki@mini.pw.edu.pl

Bibliography

Incremental Machine learning

Definition

Saffari, Amir, Christian Leistner, Jakob Santner, Martin Godec, and Horst Bischof. **"On-line random forests."** In 2009 ieee 12th international conference on computer vision workshops, iccv workshops, pp. 1393-1400. IEEE, 2009.

Gepperth, Alexander, and Barbara Hammer. **"Incremental learning algorithms and applications."** In European symposium on artificial neural networks (ESANN). 2016.

Yang, Qing, Yudi Gu, and Dongsheng Wu. **"Survey of incremental learning."** In 2019 Chinese Control And Decision Conference (CCDC), pp. 399-404. IEEE, 2019.

Joshi, Prachi, and Parag Kulkarni. **"Incremental learning: Areas and methods-a survey."** International Journal of Data Mining & Knowledge Management Process 2, no. 5 (2012): 43.

Ade, R. R., and P. R. Deshmukh. **"Methods for incremental learning: a survey."** International Journal of Data Mining & Knowledge Management Process 3, no. 4 (2013): 119.

Read, Jesse, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. **"Batch-incremental versus instance-incremental learning in dynamic and evolving data."** In International symposium on intelligent data analysis, pp. 313-323. Springer, Berlin, Heidelberg, 2012.

Hoeffding Trees

Domingos, Pedro, and Geoff Hulten. **"Mining high-speed data streams."** In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 71-80. 2000.

https://www.cms.waikato.ac.nz/~abifet/book/chapter_6.html#rfig6-4

Reddit

Medvedev, Alexey N., Renaud Lambiotte, and Jean-Charles Delvenne. **"The anatomy of Reddit: An overview of academic research."** In Dynamics on and of Complex Networks, pp. 183-204. Springer, Cham, 2017.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. **"The pushshift reddit dataset."** In Proceedings of the International AAAI Conference on Web and Social Media, vol. 14, pp. 830-839. 2020.

Zannettou, Savvas, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. **"Who let the trolls out? towards understanding state-sponsored trolls."** In Proceedings of the 10th acm conference on web science, pp. 353-362. 2019.

Finkelstein, Joel, Savvas Zannettou, Barry Bradlyn, and Jeremy Blackburn. **"A quantitative approach to understanding online antisemitism."** arXiv preprint arXiv:1809.01644 (2018).

Huang, Binxuan, and Kathleen M. Carley. **"Residual or gate? towards deeper graph neural networks for inductive graph representation learning."** arXiv preprint arXiv:1904.08035 (2019).

Zhang, Houyu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. **"Grounded conversation generation as guided traverses in commonsense knowledge graphs."** arXiv preprint arXiv:1911.02707 (2019).

... and 180 works used for "mass analysis" which are not listed

Other

Scarselli, Franco, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. **"The graph neural network model."** IEEE transactions on neural networks 20, no. 1 (2008): 61-80.

Bjork, Staffan, and Johan Redstrom. **"Redefining the focus and context of focus+ context visualization."** In IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings, pp. 85-89. IEEE, 2000.

Graphics

https://www.flaticon.com/authors/freepik

Heraclitus by Luca Giordano (https://en.wikipedia.org/wiki/Heraclitus)