

**The
Alan Turing
Institute**

**Fairness, Privacy
and Transparency by
Design for AI/ML
systems**

Dr. Taufiq Asyhari

Courtesy

- Compiled from
 - P Missier, Transparency in ML and AI, Digital Catapult London, Mar. 2018.
 - L Albertsson, Privacy by Design, Jfokus, 2018.
 - N Schmidt, Fairness in AI: How to Identify and Fix Discrimination in Machine Learning, Maryland AI Group, 2019.
 - P Bennett, et al., Fairness-aware Machine Learning: Practical Challenges and Lessons Learned, WSDM 2019.
 - A Roth, (Un)fairness in Machine Learning, University of Pennsylvania.
 - M-A Clinciu and HF Hastie, A Survey of Explainable AI Terminology, Workshop on Interactive Natural Language Technology for XAI, 2019.

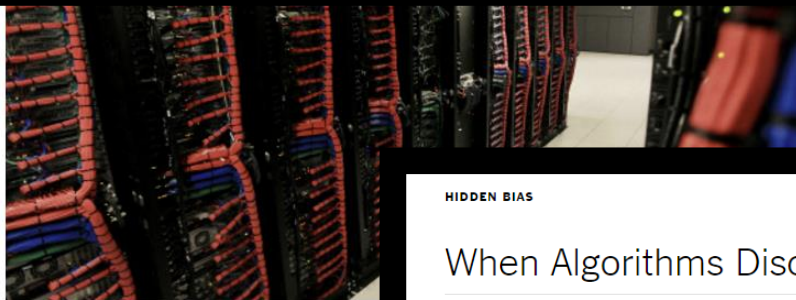


Photo by Ben Torres (Bloomberg)

Big Bad Data May Be Discriminating

August 15, 2016

AUTHORS

B Bloomberg BNA - Staff Reports

SHARING

Twitter

Facebook

Google +

LinkedIn

Email this article

Print this article

TAGS

Cybersecurity, In-House, In-House Perspective, Legal Industry, Technology

Big Data

By Kevin McGowan, Bloomberg

"Big data" is filled with promise, but it also takes care it can also drive discrimination.

"It's a bit of a black box," says the Opportunity Commission to aid employers in their hiring decisions.

Vendors that promote the use of big data, at least in the eyes of the law, risk of unlawful bias.

But others fear the algorithms will create or perpetuate discrimination.

Lipnic, lawyers and an academic say current laws are adequate to address the problem.

Global Search for Talent

Employers embrace the use of big data, but it has become a top corporate priority in Los Angeles.

Employers need to trim the list of candidates.

pool of candidates having at least the minimum job qualifications

HIDDEN BIAS

When Algorithms Discriminate



Claire Cain Miller @clairecm JULY 9, 2015

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods with high-interest loans.

How data is driving inequality

Recommend 19K

Social Surge - What's Trending



This foldable bike helmet is made from paper



Driving the most beautiful production car ever



It will now cost just \$29 to fix a cracked iPhone

OILS ARE UNEVENLY

\$

CNN Money

00:07 / 02:00

The U.S. is on the rise. But what is partly to blame.

I immediately dropped the bike and scooter and

witnessed the heist had already called the police. and charged with burglary and petty theft for the \$80.



Overview of Machine Learning / AI Systems

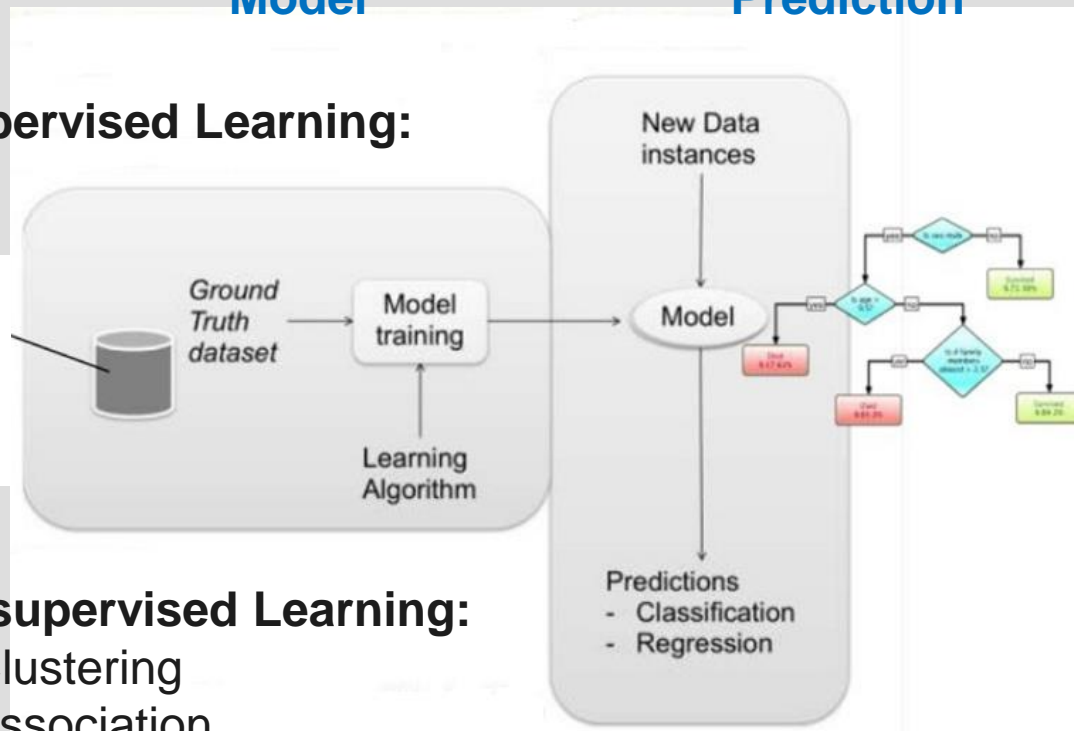
Data

Model

Prediction

Supervised Learning:

Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica



Unsupervised Learning:

- Clustering
- Association

How it matters?

- In digital age, human is impacted by decisions made by algorithmically-generated knowledge embedded within computers:
 - Scanning health risk factors
 - predicting insurance risk levels
 - automatically filtering job applicants
 - approving loans or credits
 - approving access to benefits schemes
 - personal profiling for surveillance and predicting risks of criminal



Fear of the unknown
(Clinciu and Hastie, 2019)

AI/Machine Learning within Social Contexts

- Public norms concern: fairness, privacy, transparency, accountability...
- Possible strategies
 - “traditional”: legal, regulatory, watchdog
 - Encapsulate social norms in data, algorithms, models
- Fair machine learning
- Data dependency
- Privacy-preserving machine learning



Fairness, Accountability, and Transparency in Machine Learning

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to “the algorithm made me do it.”

The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

Some Context: GDPR and Algorithmic Decision-Making

- *Article 22: Automated individual decision-making, including profiling, ...,*
- *prohibits any “decision based solely on automated processing, including profiling”*
- *which “significantly affects” a data subject.*
- Profiling is “any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person”
- Perspective: Explainable algorithm if machine learning model can be articulated and understood by human.
- Adequate explanation would provide accountability of how input features relate to predictions:
- Likelihood of loan recommendation if the applicant is a minority?
- Which features play the largest role in prediction?

Fairness

AI with Fairness Characteristics?

- Imperfect statistical models are inherently inequitable, but are not necessarily systematically discriminatory
- Oxford English Dictionary - Fairness is The quality of treating people equally or in a way that is reasonable.
- Subjectivity concerns
- Numerous mathematical definitions of fairness - some of them are contradictory
 - Anti-classification
 - Classification parity
 - Calibration

Unfairness Occurrence

- Data (input)
 - Inaccurate or insufficient data
 - Underrepresentation
- Models (output)
 - Discriminatory treatment of subpopulations
 - build or “post-process” models with subpopulation guarantees
 - Not fully causally related variable that are correlated to the protected class
- Algorithms (process)
 - learning algorithm generating data through its decisions
 - e.g. don’t learn outcomes of denied mortgages
 - lack of clear train/test division

Discrimination as Unfairness

- Legal protection for Protected Classes
 - Typically includes race, color, religion, sex, disability, familial status, national origin, and age - depending on the law
 - Employment, credit, benefits and housing are targets of regulations
- Types of Discrimination
 - Overt discrimination
 - Disparate treatment
 - Adverse impact or disparate impact

Describing Fairness: Anti-Classification

- One may not use Protected Class status when making a decision (i.e., building a learning model)
 - This corresponds to "disparate treatment"
 - Use of protected class status does not have to be explicit: Proxies for class status
- Proxies for Protected Class status
 - Complexity of the model leads to class membership being used to make decisions in a way that cannot easily be found and ameliorated.
 - If variables used in the model have a strong relationship with class status, spurious correlation or poor model building could lead to a proxy problem.
- Anti-classification may cause discrimination

Describing Fairness: Classification Parity

- Ensuring that common measures of predictive performance are equal across classes
 - Equalizing false positive and false negative rates
- Differential validity
- Examples of dangers when using classification parity:
 - Predatory lending
 - Increased crime in affected communities

Describing Fairness: Calibration

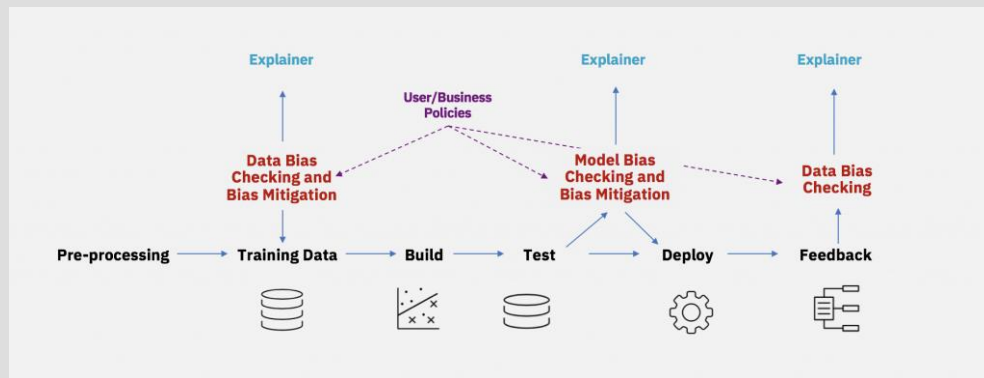
- Model outcomes, conditioned on risk, are independent of Protected Class status
- A calibrated model will predict that:
 - Protected Class members who have an estimated $X\%$ chance of defaulting truly default at a $X\%$ rate
 - Control class members who have the same estimated $X\%$ chance of defaulting truly default at the same $X\%$ rate

Less Discriminatory AI Models

- Less discriminatory modeling: Focus on model fairness over social fairness
- Same principles used to find viable less discriminatory models in traditional statistics can be used in AI models. Viable models are:
 - Similarly predictive
 - Have lower adverse impact
 - Are found through a "reasonable search"
- Searching for alternatives in AI may take extra time.
 - However, advances in computing that have allowed AI to flourish also allow a more thorough search for alternatives

Less Discriminatory AI Models: Methods

- Less Discriminatory Models may be found through:
 - Alternative feature selection: Smart searches, variable grouping, statistical methods
 - Adversarial modeling
 - Algorithm selection
 - Hyperparameter searches
 - Regularization
 - Data preprocessing



Open Source Fairness Programs: IBM's AI Fairness 360
Mitigating bias throughout the AI lifecycle

Privacy

Current Approach to Privacy and ML



User Control

Prescribe User's Rights

Know what's collected, by whom, why, opt out...



Hide Add

Why am I Seeing this

Data Protection

Anonymize



remove

"identifiable Information"

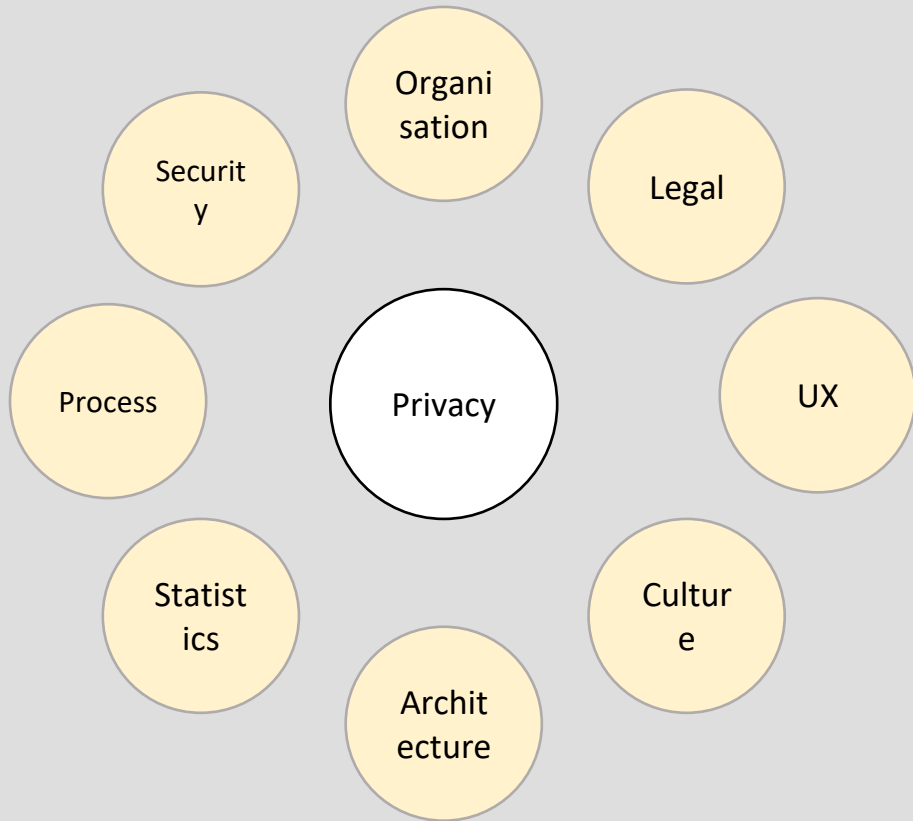
Encrypt



Encrypt at rest, in transit

Privacy by Design

- Required by GDPR
- Technical scope
 - Engineering toolbox
 - Puzzle pieces - not complete solutions
- Assuming:
 - Legal requirements
 - Security primitives

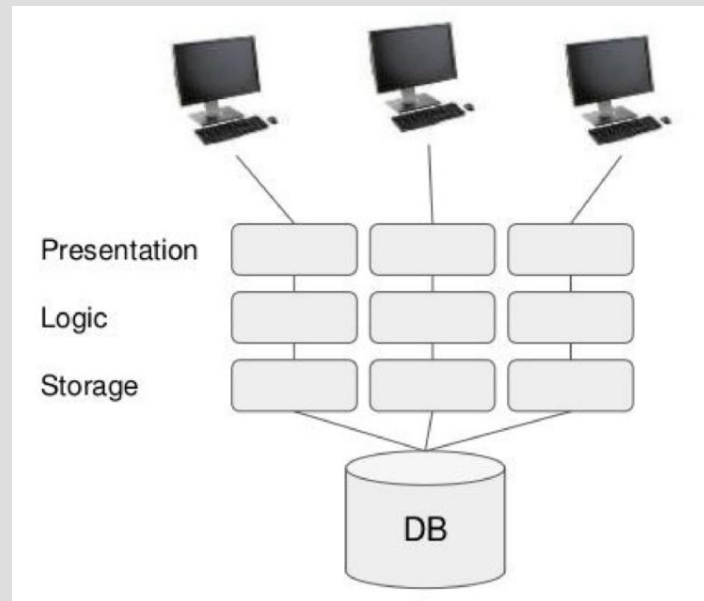


Privacy by Design

- Right to be forgotten
- Limited collection
- Limited retention
- Limited access
 - From employees
 - In case of security breach
- Consent for processing
- Right for explanations
- Right to correct data
- User data enumeration
- User data export

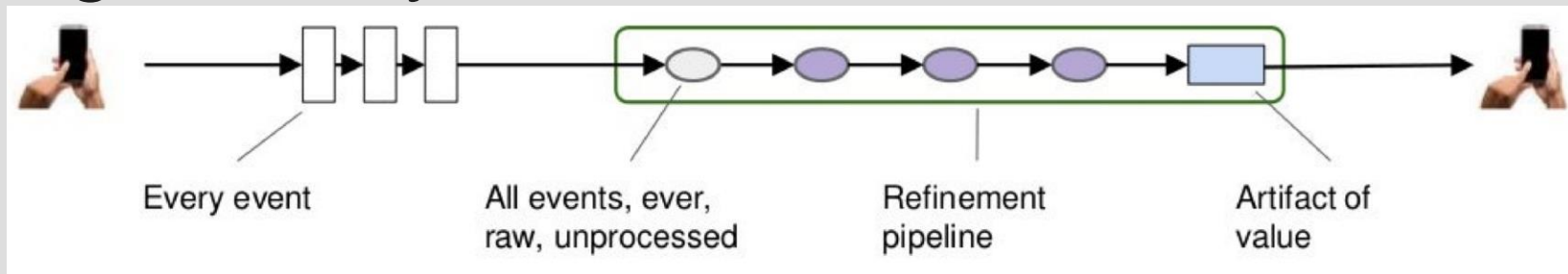
Traditional Data-Centric Systems

- Monolith
- All data in one place
- Analytics and online serving from single database
- Current state, mutable



(Albertsson, 2018)

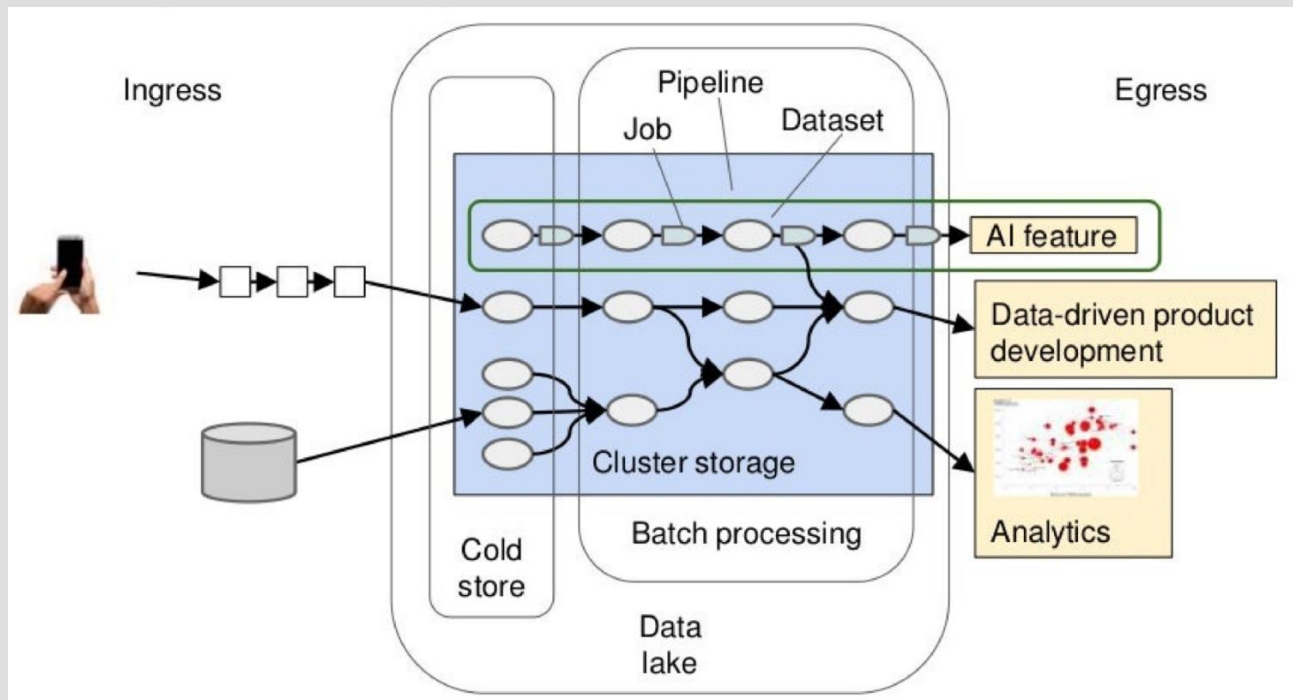
Big Data Systems



(Albertsson, 2018)

- Rationales:
 - Emerging types of data-driven AI
 - Rapid product iterations
 - Data-driven feedback
 - Democratized data
 - Robustness for scaling up business logic

Data Processing at Scale



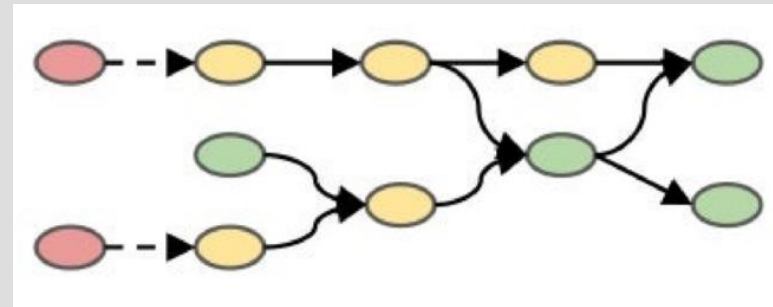
(Albertsson, 2018)

Personal information (PII) classification

- A need of field/dataset classification
- Is application content sensitive?
 - Music playlists - perhaps not
 - Running tracks, taxi rides - apparently
 - In-application messages - probably
- **Red** - sensitive data
 - GPS location
 - Messages
 - Views, likes
- **Yellow** - personal data
 - IDs (user, device)
 - Name, email, address
 - IP address
- **Green** - insensitive data
 - Not related to persons
 - Aggregated numbers
- **Grey zone**
 - Birth date, zip code
 - Recommendation / ads models?

Possible Strategy: Privacy Protection at Ingress

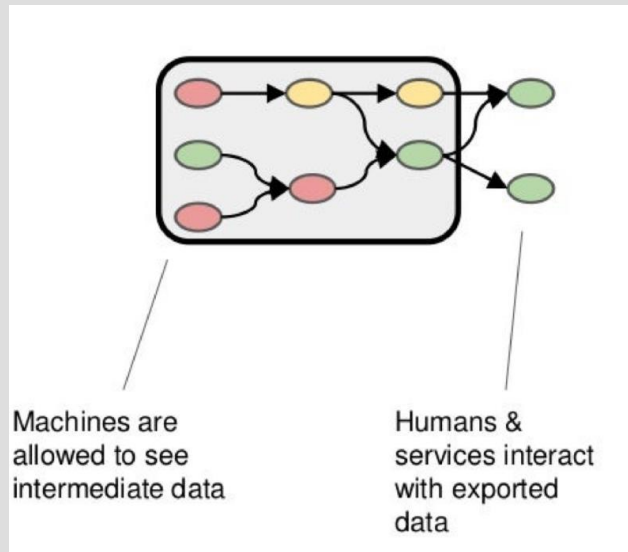
- Scramble/shuffle on arrival
- Pros:
 - Simple to implement
- Cons:
 - Limits value extraction
 - Deanonimisation possible



Adapted from (Albertsson, 2018)

Privacy Protection at Egress

- Pros:
 - Enabling processing in opaque box
 - Simpler to reason about
- Cons:
 - Strict operations required
 - Exploratory analytics need explicit egress / classification



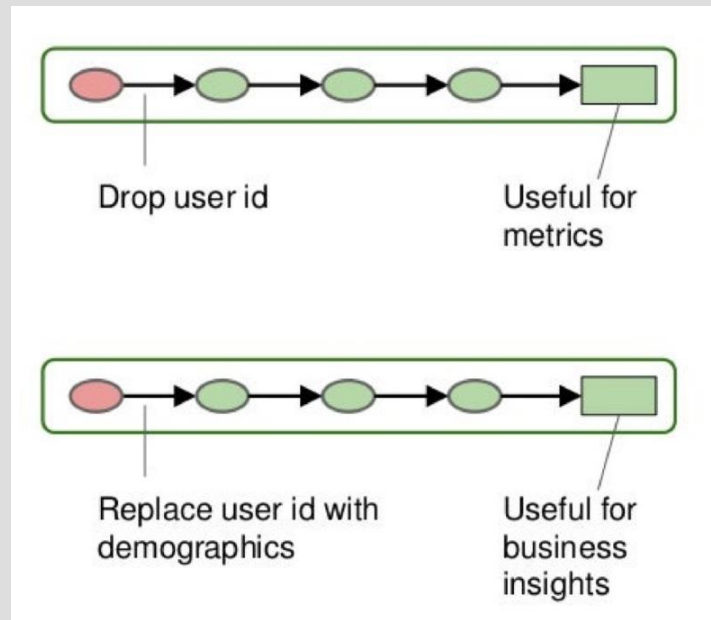
Adapted from (Albertsson, 2018)

Permission to Process

- Processing personal data requires a sanction
 - Business motive is not sufficient
- Explicit sanction
 - Consent from user
 - Necessary to perform core service
- Implicit sanction
 - Required by regulations
 - Detect money laundry, fraud, abuse
 - Bookkeeping
- Not exempt user
 - Not underage
 - No hidden identity
- Towards oblivion: PII data needs to be governed, collared, or discarded
 - Discard what you can

Discard: Anonymisation

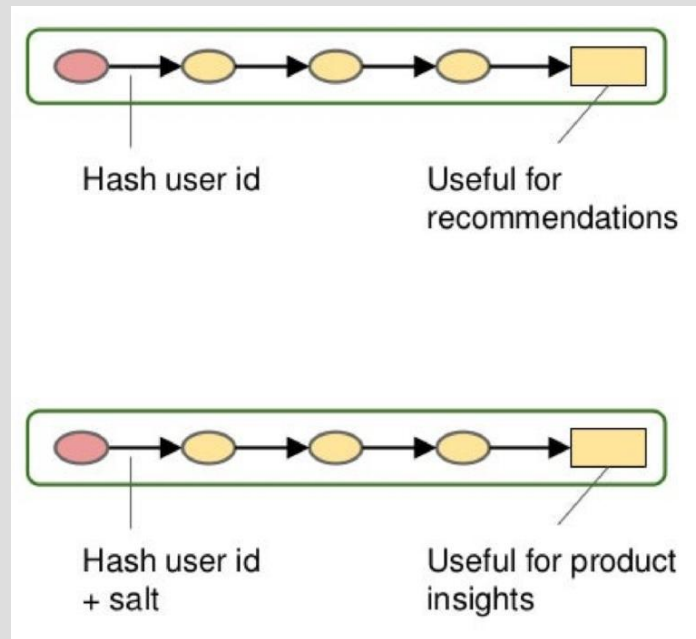
- Discard all PII
 - User id in example
- No link between records or datasets
- Replace with non-PII
 - E.g. age, gender, country
- Still no link
 - Beware: rare combination can lead to no anonymisation



Adapted from (Albertsson, 2018)

Partial Discard: Pseudonymisation

- Hash PII
- Records are linked
 - Across datasets
 - Persons can be identified (with additional data)
 - Hash recoverable from PII
- Hash PII + salt
 - Hash not recoverable
- Records are still linked
 - Across datasets if salt is constant



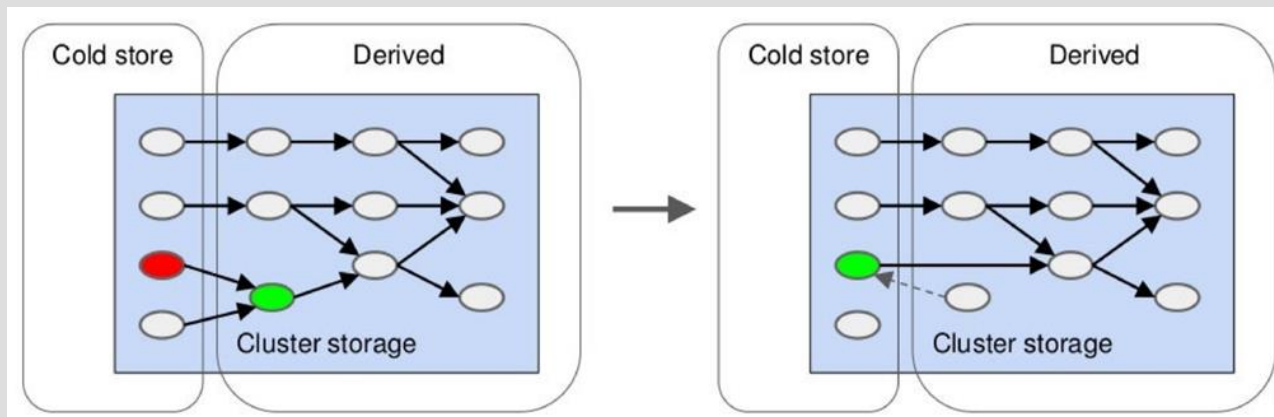
Adapted from (Albertsson, 2018)

Data Model Deadly Sins

- Using PII data as key
 - Username, email
- Publishing entity ids containing PII data
 - E.g. user shared resources (favourites, compilations) including username
- Publishing pseudonymised datasets
 - They can be de-pseudonymised with external data

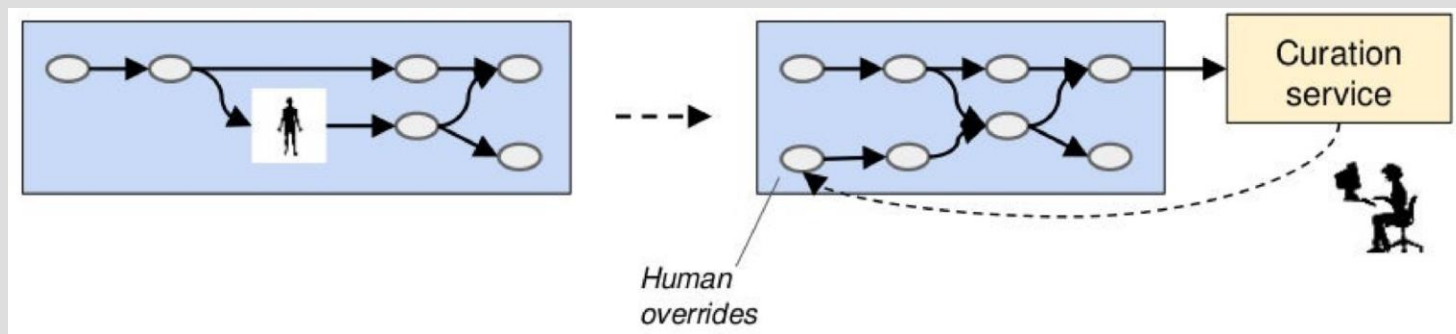
Lake Freeze

- Remove expire raw dataset, freeze derived datasets
- Workflow still works



Correcting Invalid Data: Human in the Loop

- Add human curation to cold store
 - Pipeline job merges human curation input
 - Overrides data from other sources

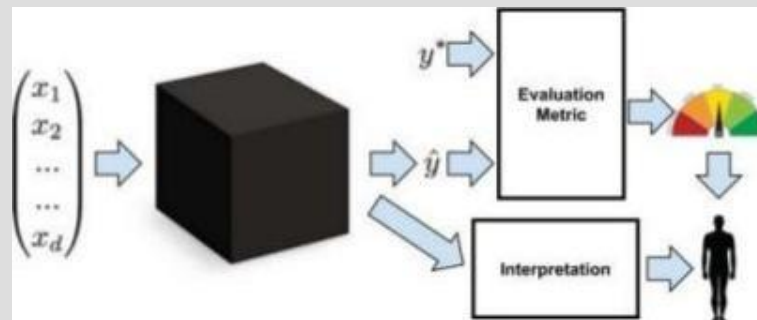


Adapted from (Albertsson, 2018)

Transparency

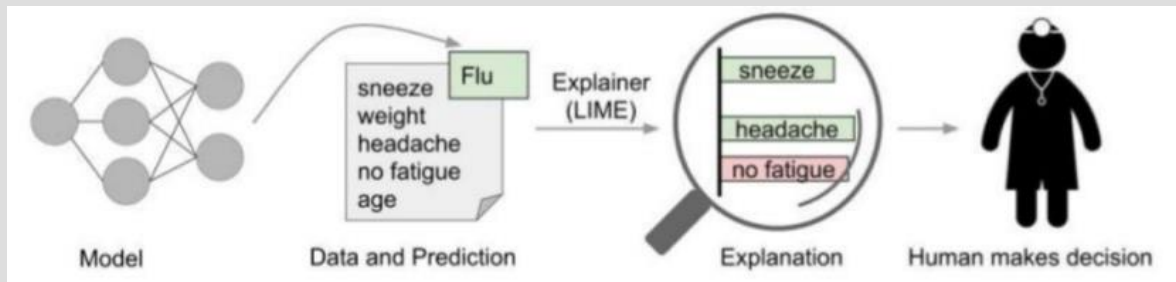
Interpretability of Machine Learning

- Transparency
 - Are features understandable?
 - Which features are more important?
- Post hoc interpretability
 - Natural language explanations
 - Visualisation of models
 - Domain-specific interpretation: "this tumor is classified as malignant because to the model it looks a lot like these other tumors"



Interpretability: Why should it be trusted?


- If the users do not trust a model or a prediction, they will not use it.



- Explaining a prediction relates to presenting textual or visual illustrations that provide qualitative understanding of the relationship between the instance's components and the model's prediction.

Features: High-Level







Example #3 of 6


True Class:  Atheism


[Instructions](#) [Previous](#) [Next](#)

Algorithm 1

Words that A1 considers important:

GOD	
mean	
anyone	
this	
Koresh	
through	

Predicted:  Atheism







Prediction correct: 


Document


From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! **GOD!**
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Algorithm 2

Words that A2 considers important:

Posting	
Host	
Re	
by	
in	
Nntp	

Predicted:  Atheism

Prediction correct: 

Document

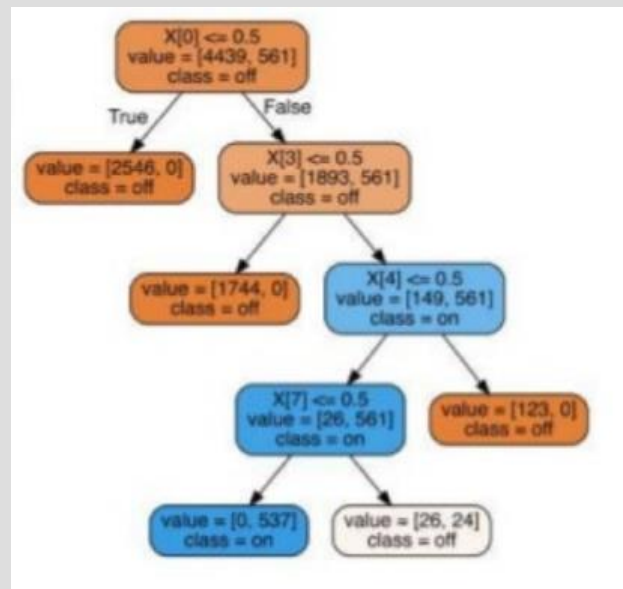
From: pauld@verdix.com (Paul Durbin)
Subject: **Re:** DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

**SVM classifier, 94% accuracy
...but questionable!**

(Missier, 2018)

Features

- Volume: how many features contribute to the prediction?
- Meaning : how suitable are the features for human interpretation?
- Raw: (low-level, non-semantic) signals such as images pixels
 - Deep learning
- Many features versus Few, high-level features.



(Missier, 2018)

Law-Obeying AI

- Operational AI systems need to obey both the law of the land and values
- Why do we need oversight systems?
 - AI systems learn continuously, i.e., they change over time
 - AI systems are becoming opaque
 - "black boxes" to human beings
 - AI-guided systems have increasing autonomy
 - Make choices "on their own."
- Development of AI oversight systems

AI Accountability

- *Asked where AI systems are weak today, Veloso (*) says they should be more transparent. "They need to explain themselves: why did they do this, why did they do that, why did they detect this, why did they recommend that? Accountability is absolutely necessary"*

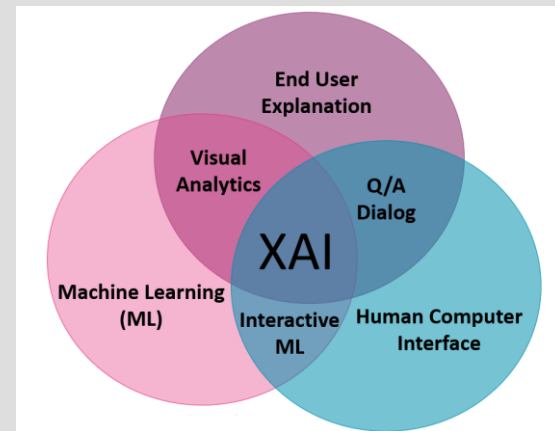
Gary Anthes. 2017. Artificial intelligence poised to ride a new wave. Commun. ACM 60, 7 (June 2017).

(*) Manuela Veloso, Head of the Machine Learning Department at Carnegie-Mellon University

Explainable AI

“Explainable AI can present the user with an easily understood chain of reasoning from the user’s order, through the AI’s knowledge and inference, to the resulting behavior” [van Lent et al., 2004].

"XAI is a research field that aims to make AI systems results more understandable to humans" [Adadi and Berrada, 2018].



(Clinciu and Hastie, 2019)

- M van Lent, W Fisher, and M Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In Proc. 16th conference on Innovative applications of artificial intelligence (IAAI'04), Randall Hill (Ed.). AAAI Press 900-907.
- A Adadi and M Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).IEEE Access, 6:52138–52160.

References

- Gary Anthes. 2017. Artificial intelligence poised to ride a new wave. Commun. ACM 60,7 (June 2017), 19-21. DOI: <https://doi.org/10.1145/3088342>
- J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," Big Data Soc., vol. 3, no. 1, p. 2053951715622512, 2016
- Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc, and Elhadad, Noemie. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In KDD, 2015
- A. Etzioni and O. Etzioni, "Designing AI Systems That Obey Our Laws and Values," Commun. ACM, vol. 59, no. 9, pp. 29-31, Aug. 2016.
- Z. C. Lipton, "The Mythos of Model Interpretability," Proc. 2016 ICML Work. Hum. Interpret. Mach. Learn. (WHI 2016), Jun. 2016.
- M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' : Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016, pp. 1135-1144.
- M van Lent, W Fisher, and M Mancuso. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In Proc. 16th conference on Innovative applications of artificial intelligence (IAAI'04), Randall Hill (Ed.). AAAI Press 900-907.
- A Adadi and M Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).IEEE Access, 6:52138–52160.
- <https://arxiv.org/pdf/1609.05807.pdf>