Making Algorithms Accountable

Katharine Jarmul

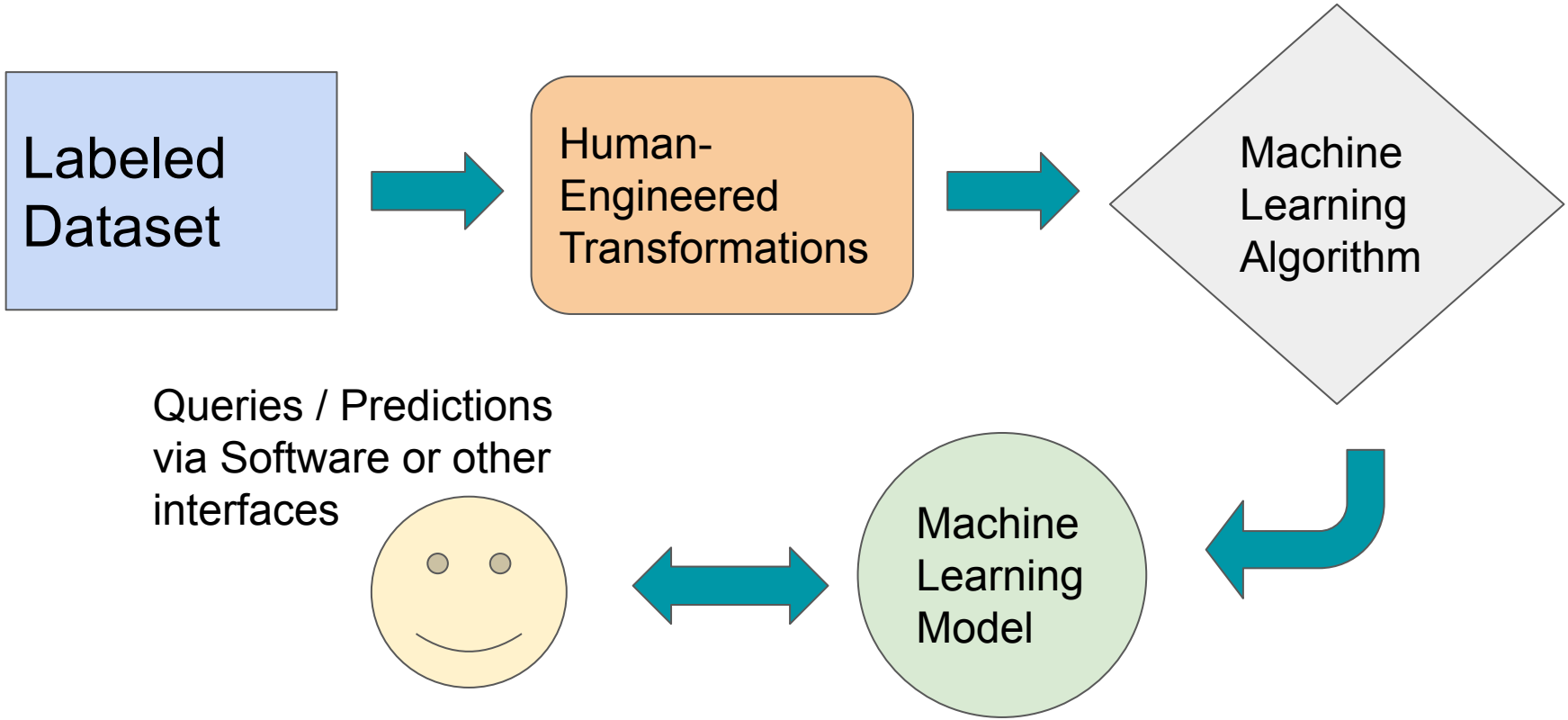
Digital Government Conference
October 22, 2019 - Helsinki

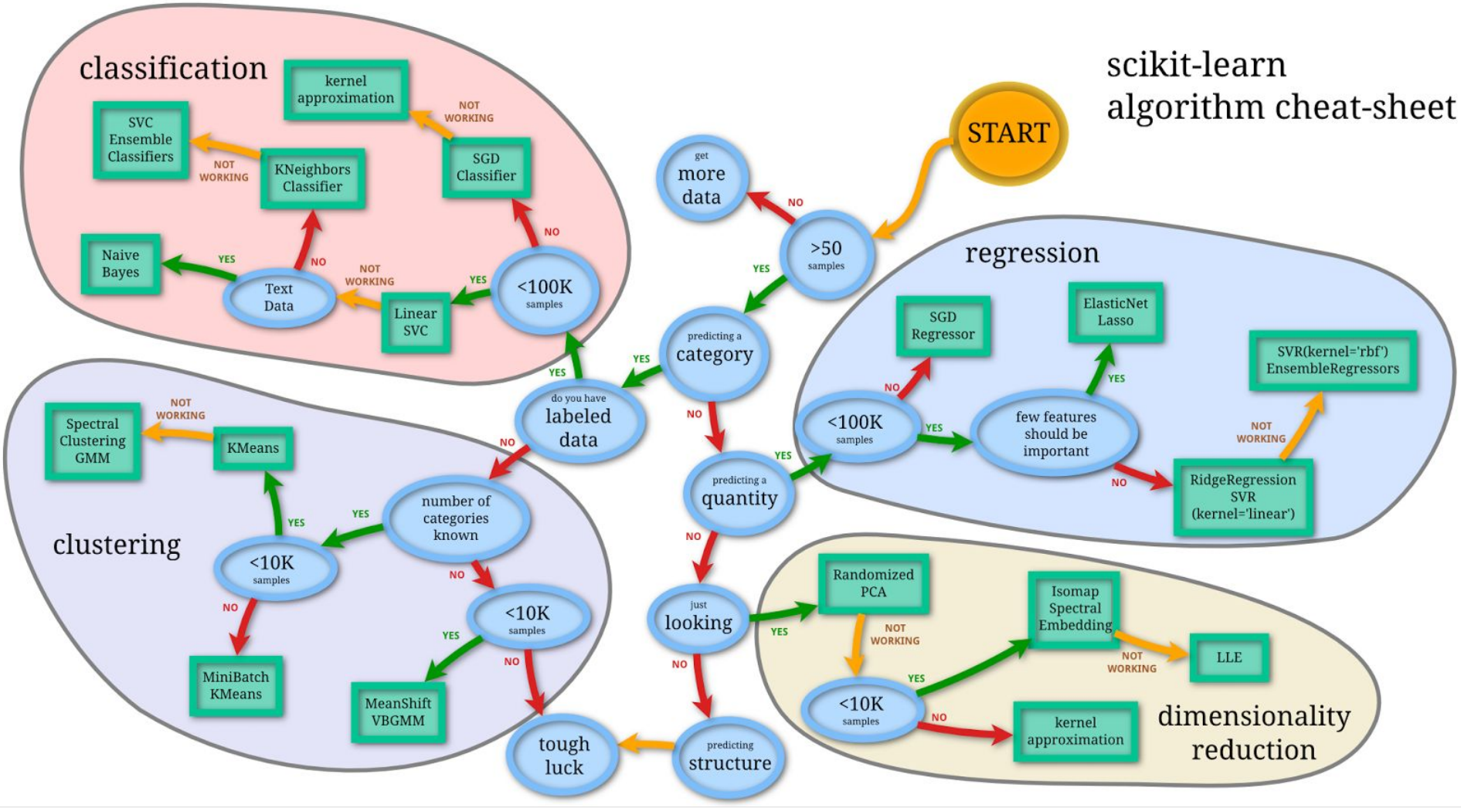
kjamistan.com

What does accountability mean?

- Responsible for one’s own actions
- Can respond to inquiry
- Is trustworthy
- Follows social norms / moral or ethical principles

What is an algorithm?




scikit-learn
algorithm cheat-sheet
START
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
NO
SVC Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
NOT WORKING
kernel approximation
NOT WORKING
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
classification
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
ElasticNet Lasso
NO
RidgeRegression SVR (kernel='linear')
NOT WORKING
SVR(kernel='rbf') EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure
tough luck
NO
number of categories known
YES
<10K samples
NO
Spectral Clustering GMM
NOT WORKING
KMeans
YES
<10K samples
NO
MiniBatch KMeans
YES
MeanShift VBGMM
NO
clustering

What’s missing from this process?

- Legal Concerns
- Privacy Concerns
- Deep understanding of domain / data
- Ethical choices / prejudice in processing or data
- Implications for user
- Understanding of socio / legal context

What would accountable machine learning look like?

- Responsible for one's own actions:
  - Predictable responses and error reporting
- Can respond to inquiry
  - Explainability and support from company/team
- Is trustworthy
  - Privacy-aware, robust against attack, unbiased
- Follows social norms / moral or ethical principles
  - Defined ethics / norms with testing / guarantees

Responsible AI

(a) Input 1 (b) Input 2 (darker version of 1)

Figure 1: An example erroneous behavior found by DeepXplore in Nvidia DAVE-2 self-driving car platform. The DNN-based self-driving car correctly decides to turn left for image (a) but incorrectly decides to turn right and crashes into the guardrail for image (b), a slightly darker version of (a).

Pei et al., *DeepXplore: Automated Whitebox Testing of Deep Learning Systems,* 2017

Explainable AI

```
In [5]:  1  eli5.explain_weights(model)
```

Out[5]: **y** top features

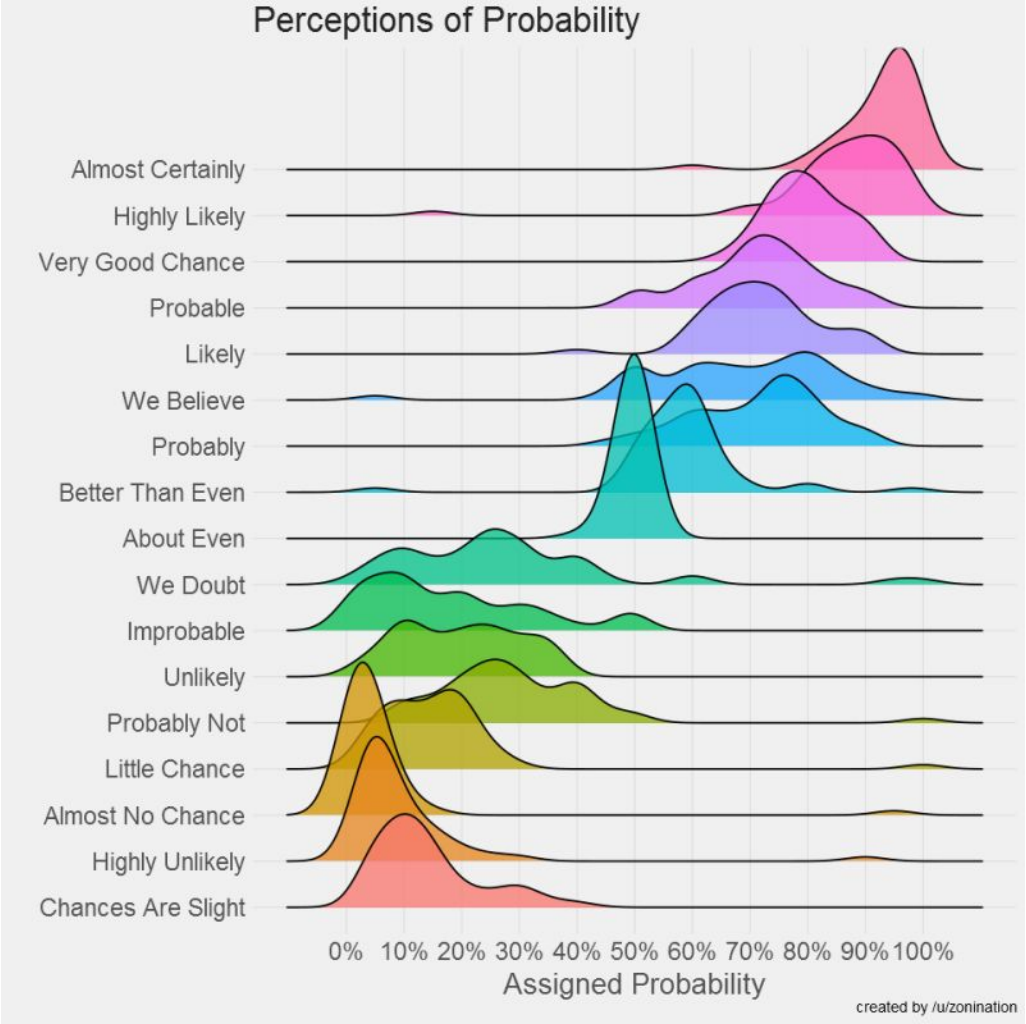
Weight[?]	Feature
+11.948	<BIAS>
+0.157	x7
+0.083	x75
+0.038	x3437
+0.036	x4787
+0.035	x3669
+0.035	x2344
+0.034	x10
+0.034	x73
+0.034	x836
+0.033	x1923
+0.033	x16
+0.030	x3391
+0.030	x98
+0.029	x2596
+0.028	x14
+0.028	x97
… 387 more positive …	
… 325 more negative …	
-0.032	x4817

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*. 2016

Understandable AI



Privacy-Aware AI

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, 2015

Robust AI

Athalye et al. *Synthesizing Robust Adversarial Examples*, 2017.

Ethical AI


SHE KNOWS PUTIN
TOO WELL
Posted by Defeat Crooked Hillary
442,065 Views

Conclusions:

- AI is built by humans, (often) for humans, and yet often undergoes very little human-centered design
- Privacy-aware, consensual, representative data collection
- Explainability, privacy-preserving techniques, robust design and ethical training are all being actively researched and there are some advances available (not perfect)
- At the end of the day, humans need to be accountable and set up a two-way communication on those processes

Slide References

- Scikit-learn algorithm map: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
- Pei et al., *DeepXplore: Automated Whitebox Testing of Deep Learning Systems*, 2017
- Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, 2016
- Perceptions of Probability: https://github.com/zonination/perceptions
- Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, 2015
- Athalye et al. *Synthesizing Robust Adversarial Examples*, 2017.
- Crooked Hillary FB Image: https://www.buzzfeednews.com/article/craigsilverman/cambridge-analytica-says-they-won-the-election-for-trump

Making Algorithms Accountable

Dr. Nakeema Stefflbauer

Digital Government Conference

October 22, 2019 - Helsinki

Making Algorithms Accountable: Critical Factors

1- Explainability

2- Privacy-preserving techniques

3- Ethical training

What is needed:

- a framework for tracking problems
- methods to reverse erroneous decisions
- Decision-makers' understanding of potential risks + impact of automation *pre- implementation*.

Product Failure due to **lack of explainability**

Facial recognition wrongly identifies public as potential criminals 96% of time, figures reveal

14-year-old black schoolboy among those wrongly fingerprinted after being misidentified

[United Kingdom]

Eight trials carried in London between 2016 and 2018 resulted in a 96 per cent rate of “false positives” – where software wrongly alerts police that a person passing through the scanning area matches a photo on the database.

Product Failure due to **lack of privacy preservation**

Smart Cities Are Creating a Mass Surveillance Nightmare + ***Hacking Risks***

[United States]

Baltimore + Atlanta government functions ground to a halt when ransomware was used successfully

Residents lost access to online bill payments, property deed transfers and court scheduling. **In Baltimore, the city was out of action for weeks & crucial data was permanently lost.**

(Likely) Failure due to **lack of privacy preservation**

France Set to Roll Out Nationwide Facial Recognition ID Program

[FRANCE]

The country's data regulator says the program breaches the European rule of consent & a privacy group is challenging it in France's highest administrative court.

It took a hacker just over one hour to break into a "secure" government messaging app this year, raising concerns about the state's security standards.

Product Failure due to **lack of ethical training**

[United Kingdom]

The British government has knowingly rolled out facial recognition software for passport photos which does not recognise darker skin-colour faces.

How might **an accountable AI system look?**

1. **Run impact tests** before public algorithm rollout, possibly to a (Food-and-Drug-Administration) FDA-type board to assess the risk of violation of existing laws, whether civil rights, human rights, or privacy laws.
2. **Maintain a civil society register for public algorithms** that contains anonymized facts about the raw training data, the algorithms that analyze it, and the decision-making models that emerge.
3. **Make privacy the default for use of personally identifiable data** such that there are very clear guidelines about how you can use that data - and how you cannot.