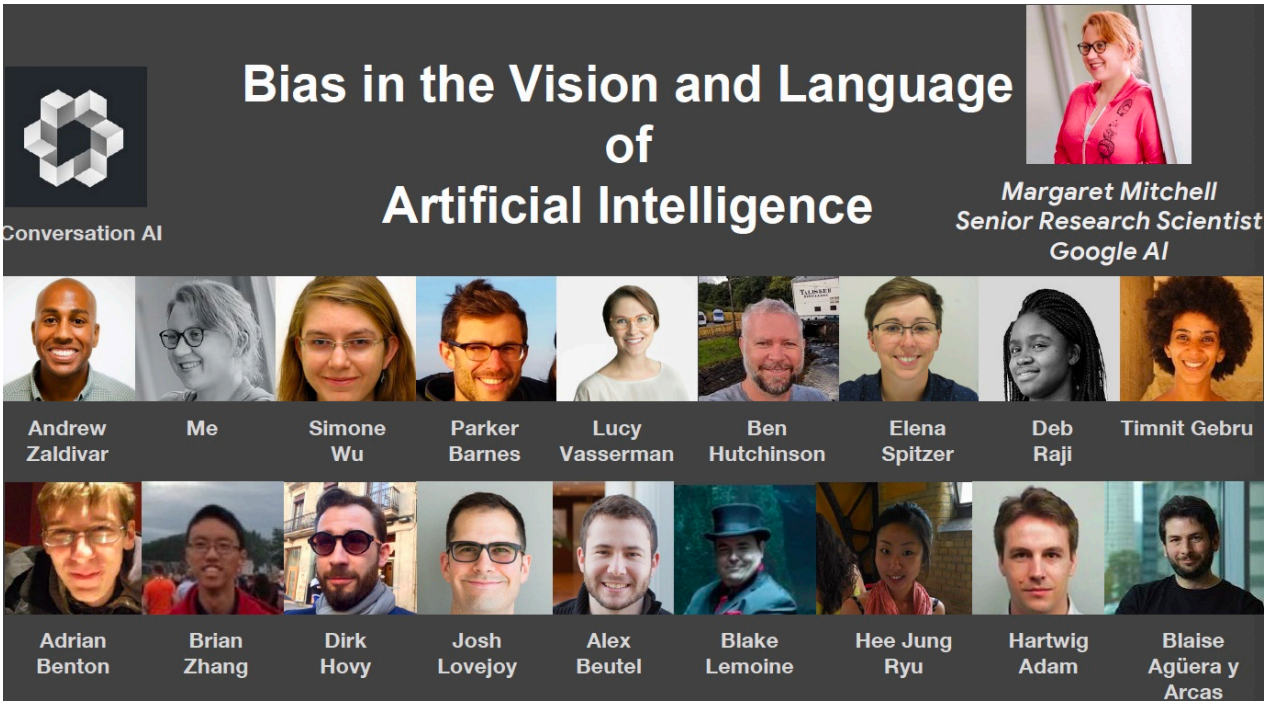
Lecture 19 Safety, Bias, and Fairness



What do you see?

Green Bananas

Unripe Bananas

What do you see?

Ripe Bananas

Bananas with spots

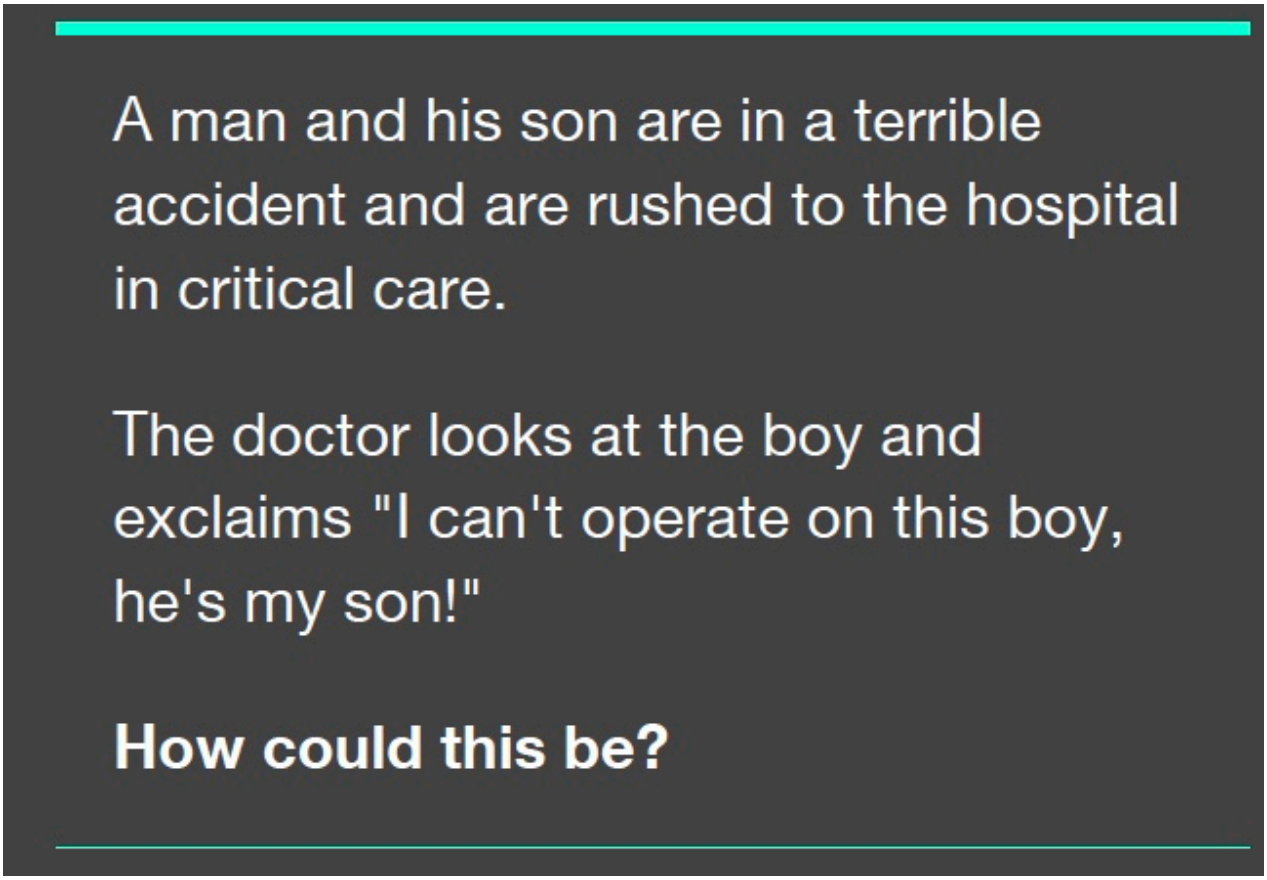
Bananas good for banana bread

Prototype Theory

分类的目的之一是减少刺激行为和认知上可用的比例的无限差异

物品的一些核心、原型概念可能来自于存储的对象类别的典型属性(Rosch, 1975)

也可以存储范例(Wu & Barsalou, 2009)



Doctor —— Female Doctor

大多数受试者忽视了医生是女性的可能性，包括男性、女性和自称女权主义者的人们。

World learning from text

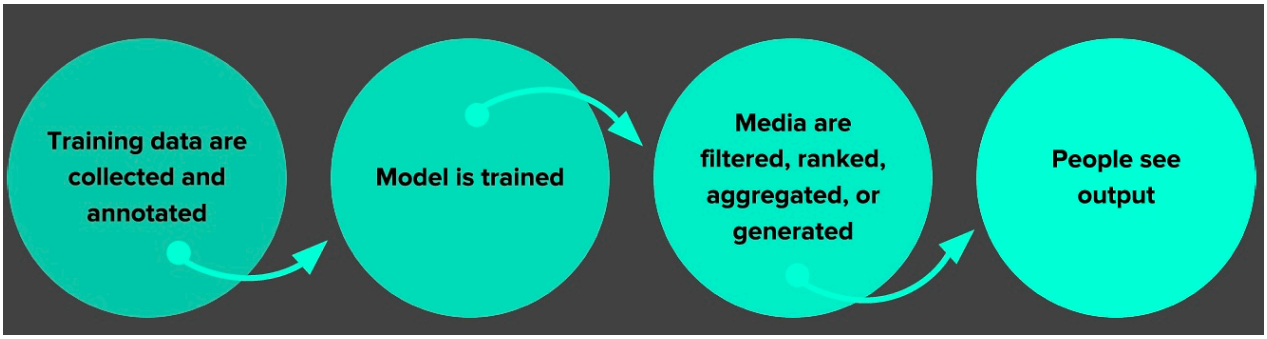
Gordon and Van Durme, 2013

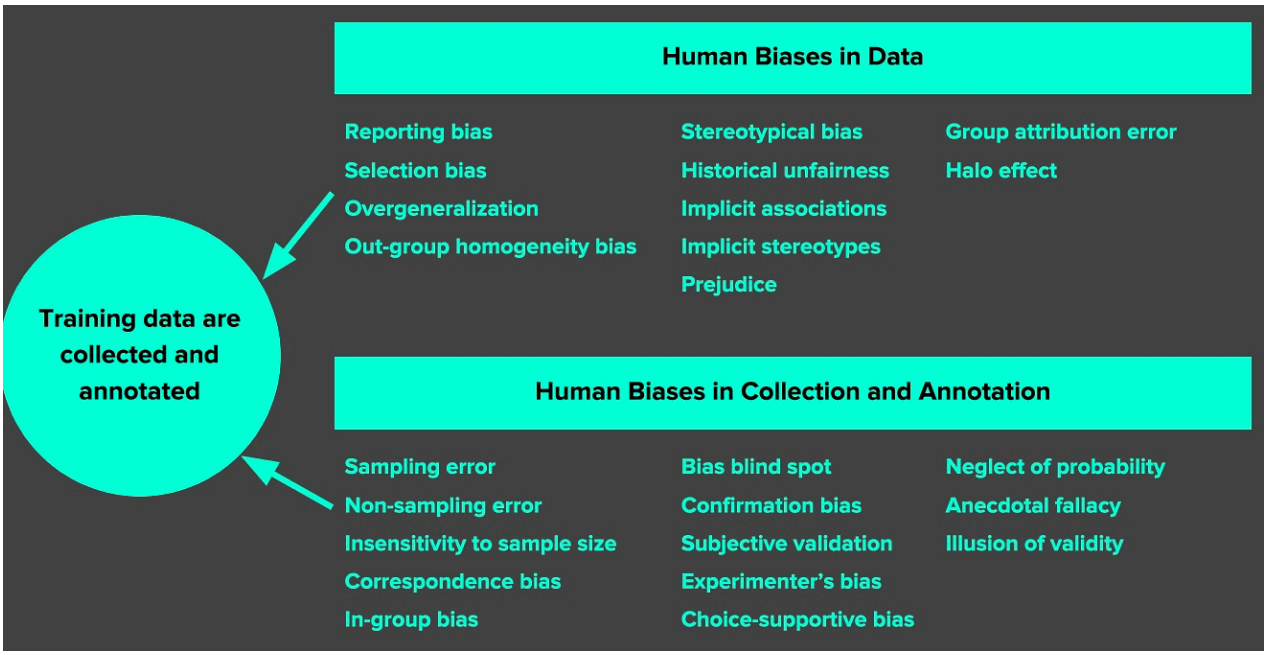
Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

- murdered 是 blinked 出现次数的十倍
- 我们不倾向于提及眨眼和呼吸等事情

Human Reporting Bias：人们写作中的行为、结果或属性的频率并不反映真实世界的频率，也不反映某一属性在多大程度上是某一类个体的特征。

更多关于我们处理世界和我们认为非凡的东西的实际情况。这影响到我们学习的一切。





- 数据
  - Reporting bias 报告偏见:人们分享的并不是真实世界频率的反映
  - Selection Bias 选择偏差:选择不反映随机样本
  - Out-group homogeneity bias 外群体同质性偏见:People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics
- 解释
  - Confirmation bias 确认偏见:倾向于寻找、解释、支持和回忆信息，以确认一个人先前存在的信念或假设
  - Overgeneralization 泛化过度:根据过于笼统和/或不够具体的信息得出结论
  - Correlation fallacy 相关性谬误:混淆相关性和因果关系
  - Automation bias 自动化偏差:人类倾向于喜欢来自自动化决策系统的建议，而不是没有自动化的相互矛盾的信息

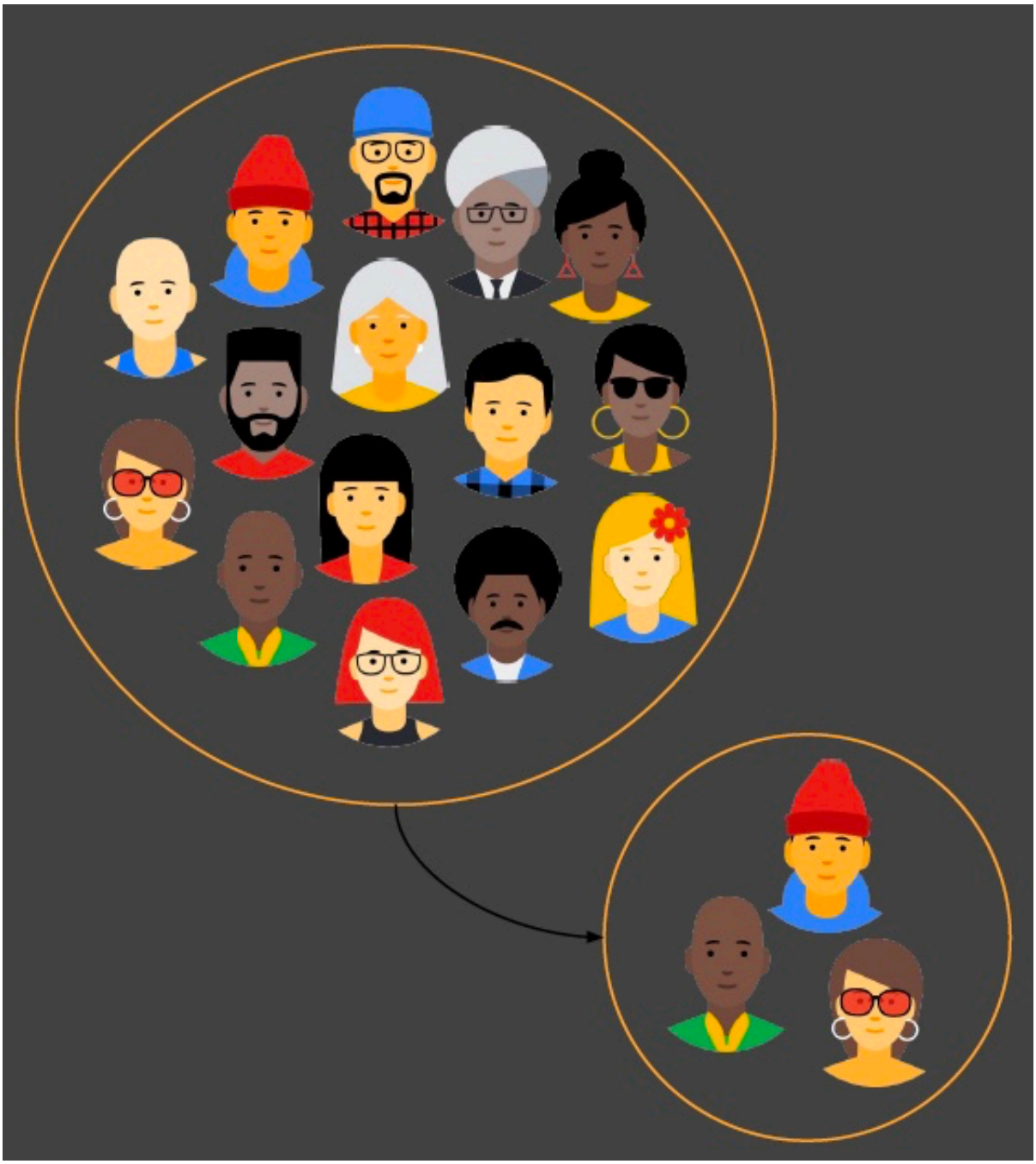
Biases in Data

Selection Bias 选择偏差:选择不反映随机样本

Out-group homogeneity bias 外群体同质性偏见：在比较态度、价值观、个性特征和其他特征时，往往群体外的成员认为比群体内的成员更相似

- 这有些难以理解：意思就是左边的四只猫之间是非常不同的，但是在狗的眼里他们是相同的

Biases in Data → Biased Data Representation

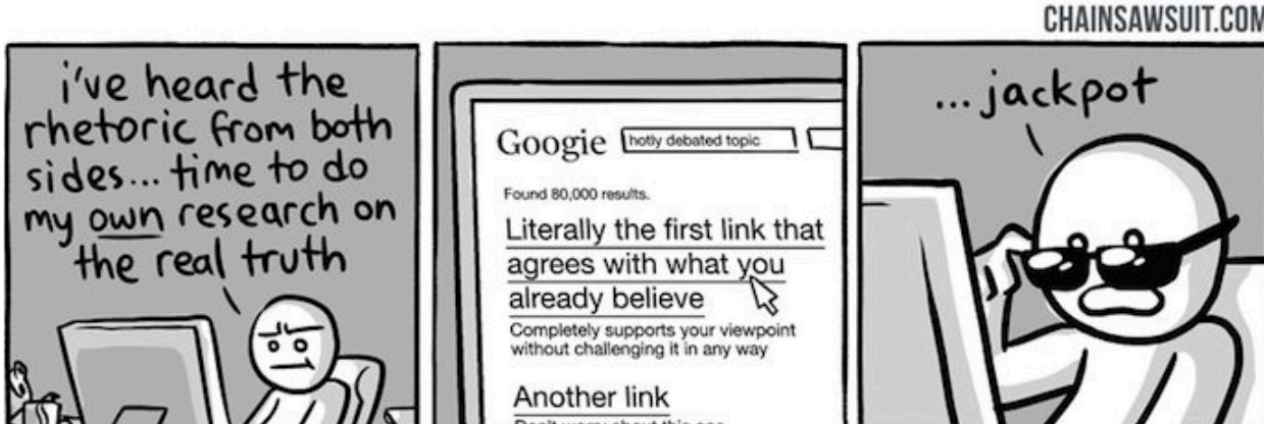


你可能对你能想到的每一个群体都有适当数量的数据，但有些群体的表现不如其他群体积极。

Biases in Data → Biased Labels

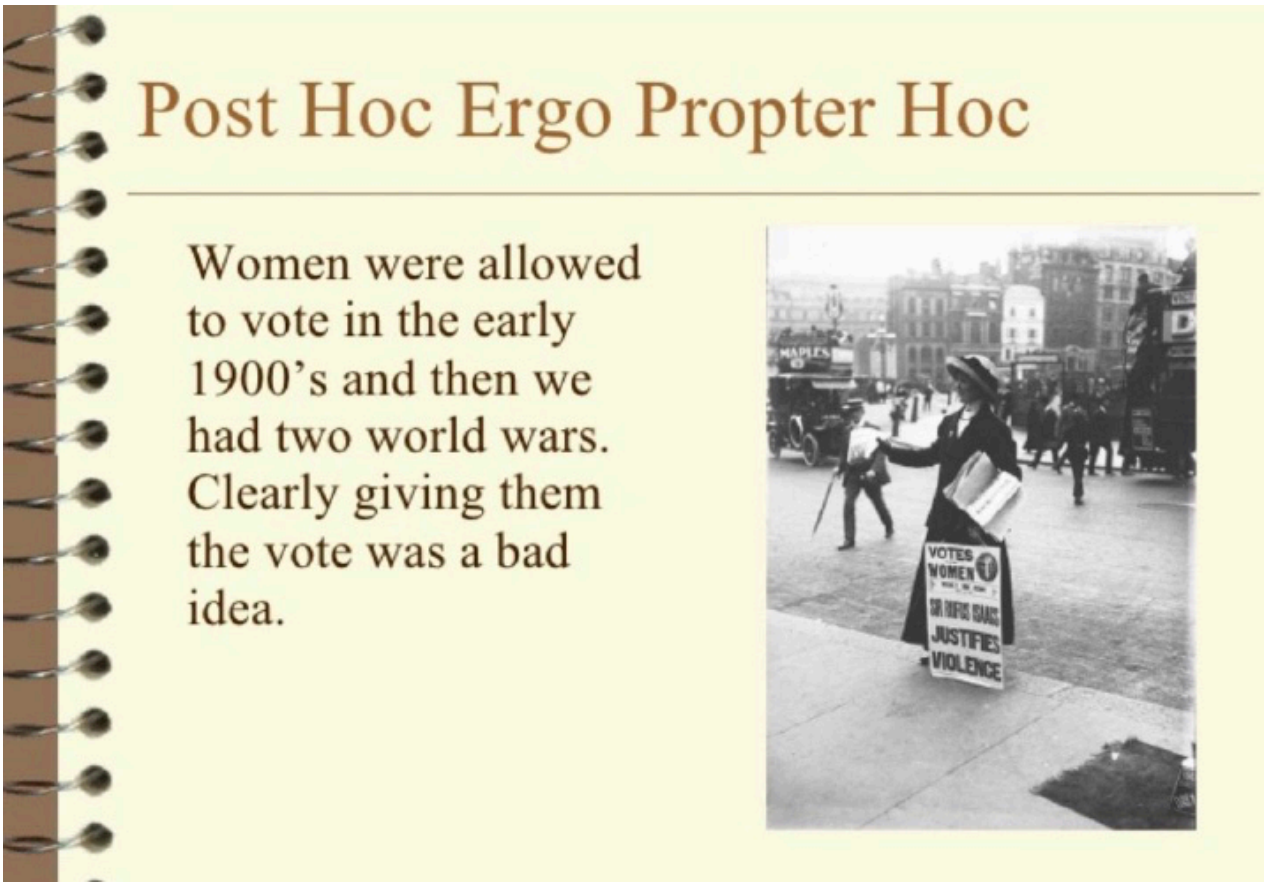
数据集中的注释将反映注释者的世界观

Biases in Interpretation



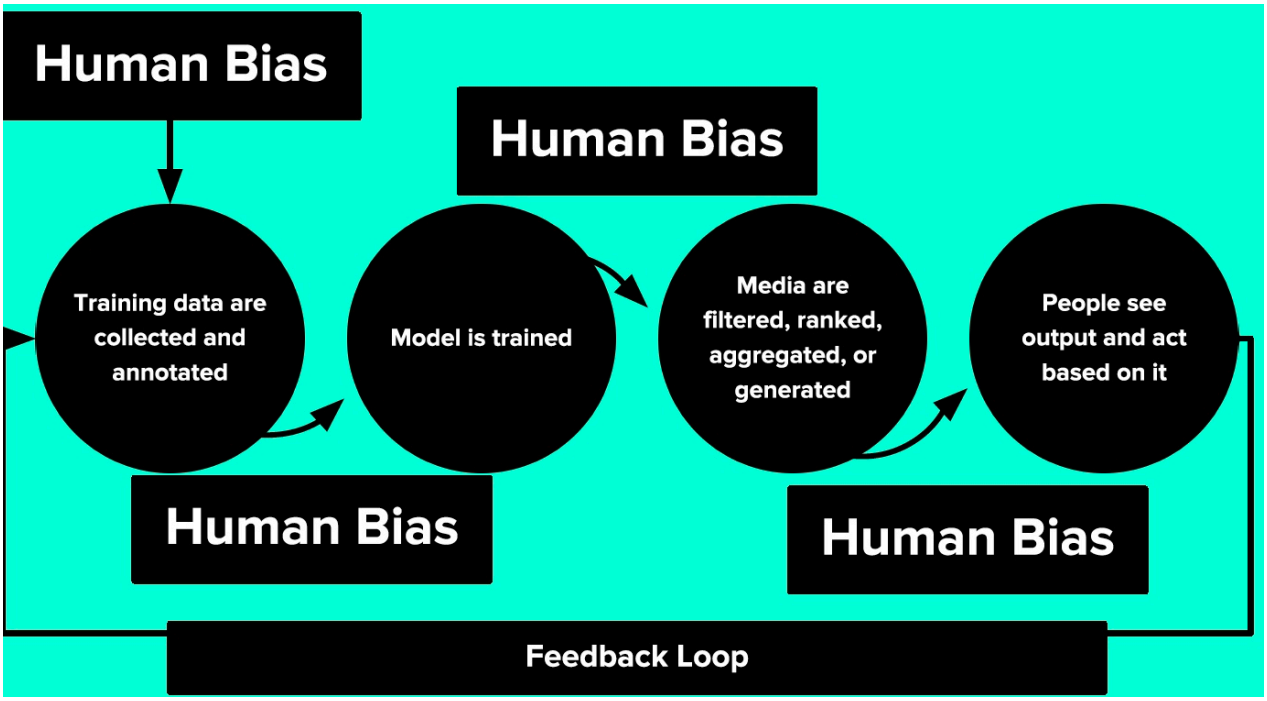
Confirmation bias 确认偏见:倾向于寻找、解释、支持和回忆信息，以确认一个人先前存在的信念或假设

Overgeneralization 泛化过度:根据过于笼统和/或不够具体的信息得出结论（相关：过拟合）



Correlation fallacy 相关性谬误:混淆相关性和因果关系

Automation bias 自动化偏差:人类倾向于喜欢来自自动化决策系统的建议，而不是没有自动化的相互矛盾的信息



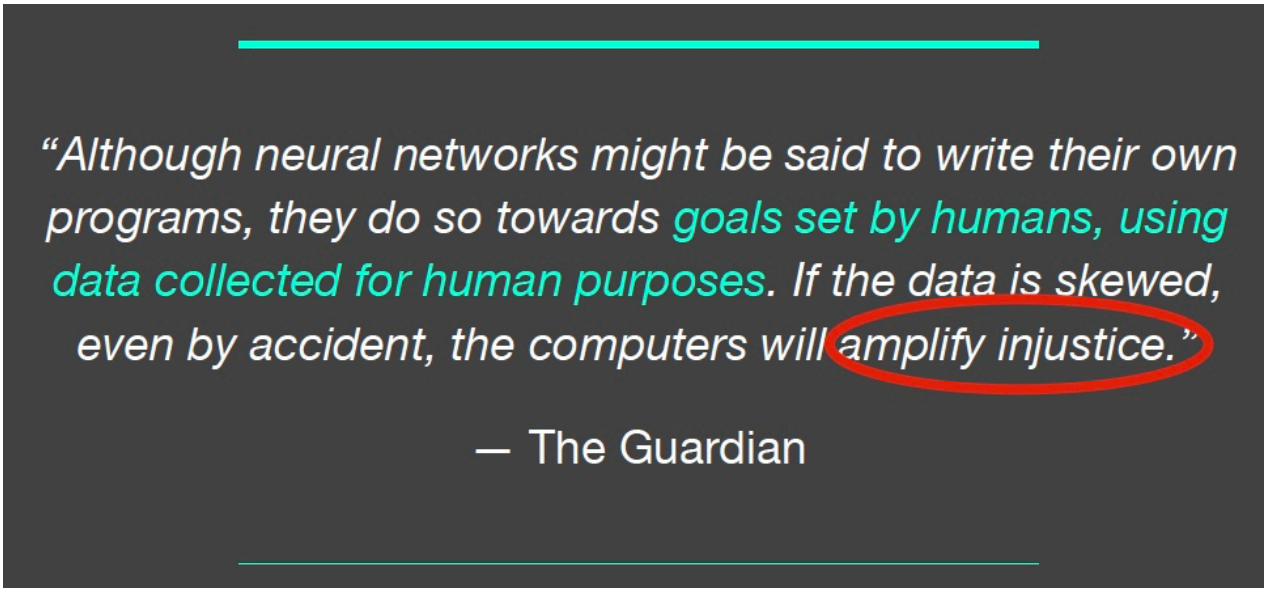
- 会形成反馈循环
- 这被称为 Bias Network Effect 以及 Bias “Laundering”

人类数据延续了人类的偏见。当ML从人类数据中学习时，结果是一个偏置网络效应。

BIAS = BAD ??

“Bias” can be Good, Bad, Neutral

- 统计以及 ML中的偏差
  - 估计值的偏差：预测值与我们试图预测的正确值之间的差异
  - “偏差”一词b(如y = mx + b)
- 认知偏见
  - 确认性偏差、近因性偏差、乐观性偏差
- 算法偏差
  - 对与种族、收入、性取向、宗教、性别和其他历史上与歧视和边缘化相关的特征相关的人的不公平、不公平或偏见待遇，何时何地在算法系统或算法辅助决策中体现出来”



- 如何避免算法偏差，开发出不会放大差异的算法

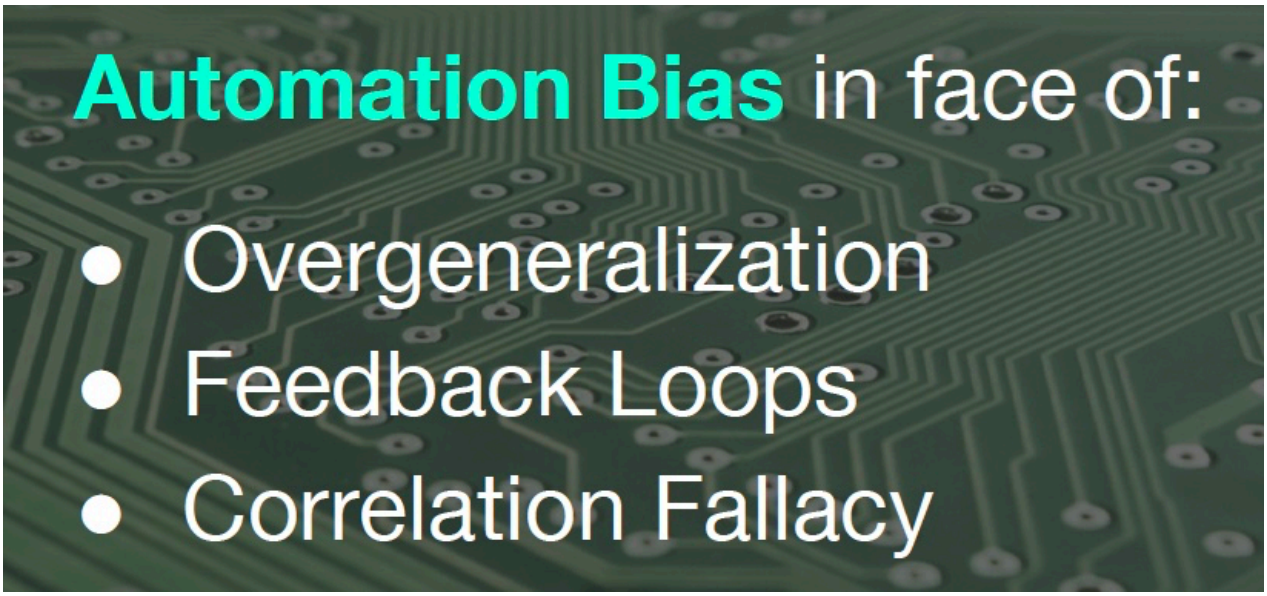
Predicting Future Criminal Behavior

- 算法识别潜在的犯罪热点
- 基于之前报道的犯罪的地方，而不是已知发生在哪里
- 从过去预测未来事件
- 预测的是逮捕的地方而不是犯罪的地方

Predicting Sentencing

- Prater (白人)额定低风险入店行窃后,尽管两个武装抢劫;一次持械抢劫未遂。
- Borden (黑色)额定高危后她和一个朋友(但在警察到来之前返回)一辆自行车和摩托车坐在外面。
- 两年后，Borden没有被指控任何新的罪行。Prater因重大盗窃罪被判8年有期徒刑。

系统默认认为黑人的犯罪风险高于拜仁



Predicting Criminality

以色列启动 Faception

Faception是第一个科技领域的率先面市的，专有的计算机视觉和机器学习技术分析人员和揭示他们的个性只基于他们的面部图像。

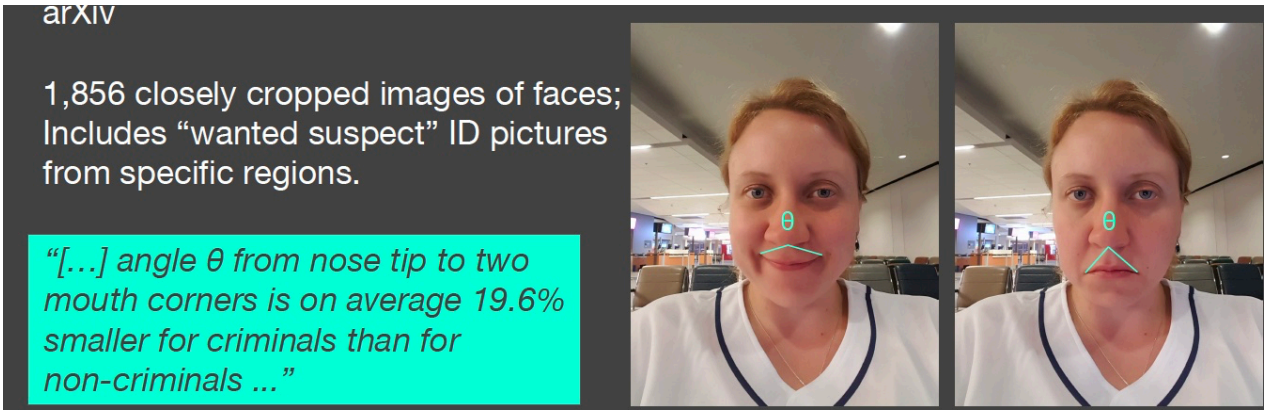
提供专业的引擎从脸的形象识别“高智商”、“白领犯罪”、“恋童癖”,和“恐怖分子”。

主要客户为国土安全和公共安全。

“Automated Inference on Criminality using Face Images” Wu and Zhang, 2016. arXiv

1856个紧密裁剪的面孔的图像,包括“通缉犯”ID特定区域的照片

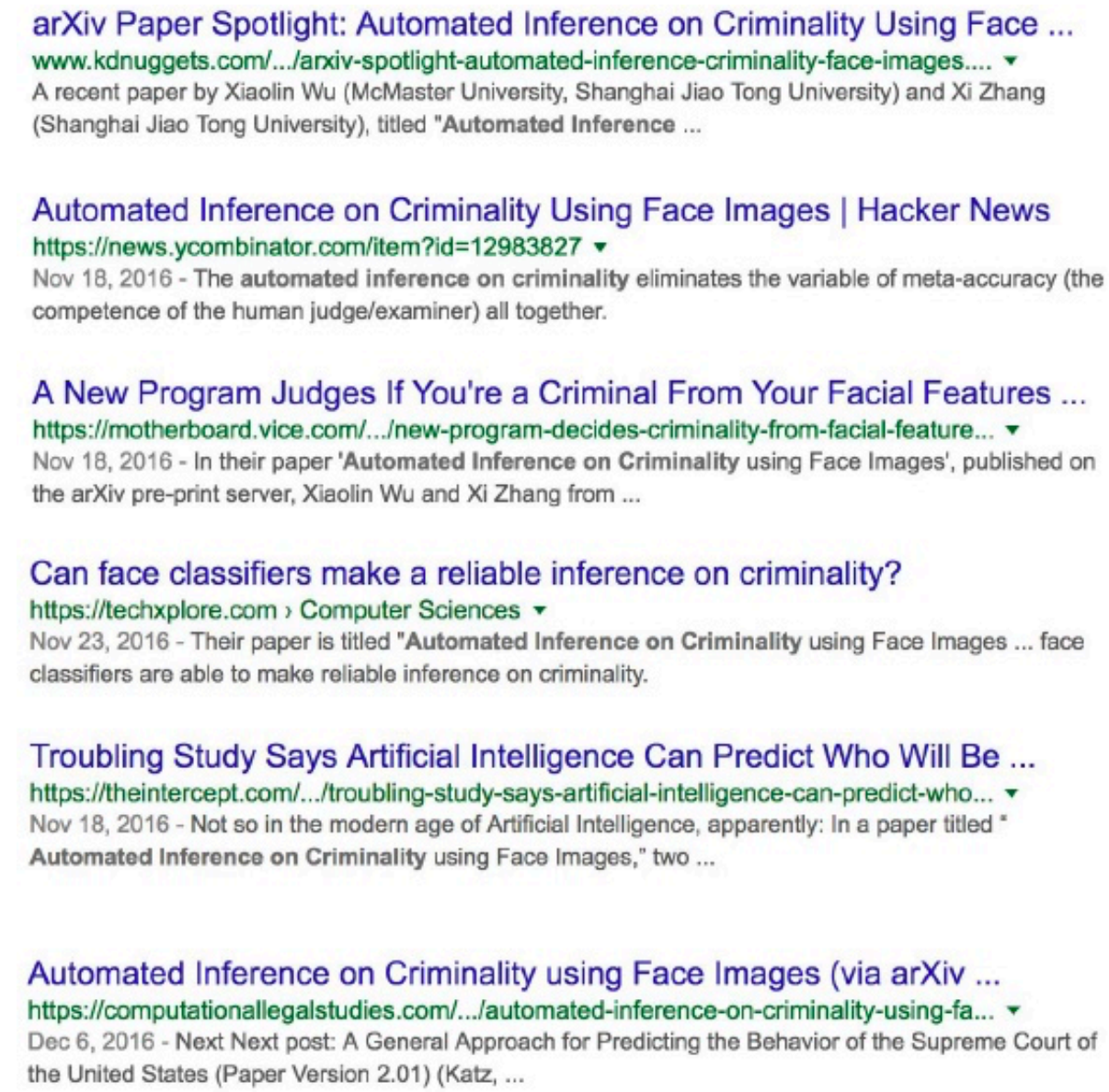
存在确认偏差和相关性偏差



See our longer piece on Medium, [“Physiognomy’s New Clothes”]

- Selection Bias + Experimenter’s Bias + Confirmation Bias + Correlation Fallacy + Feedback Loops

Predicting Criminality - The Media Blitz



(Claiming to) Predict Internal Qualities Subject To Discrimination

Predicting Homosexuality

- Wang and Kosinski, [Deep neural networks are more accurate than humans at detecting sexual orientation from facial images](), 2017.
- “Sexual orientation detector” using 35,326 images from public profiles on a US dating website.
- “与性取向的产前激素理论(PHT)相一致，男同性恋者和女同性恋者往往具有非典型的性别面部形态。”
- 在自拍中，同性恋和异性恋之间的差异与打扮、表现和生活方式有关，也就是说，文化差异，而不是面部结构的差异

See our longer response on Medium, “[Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?]()”

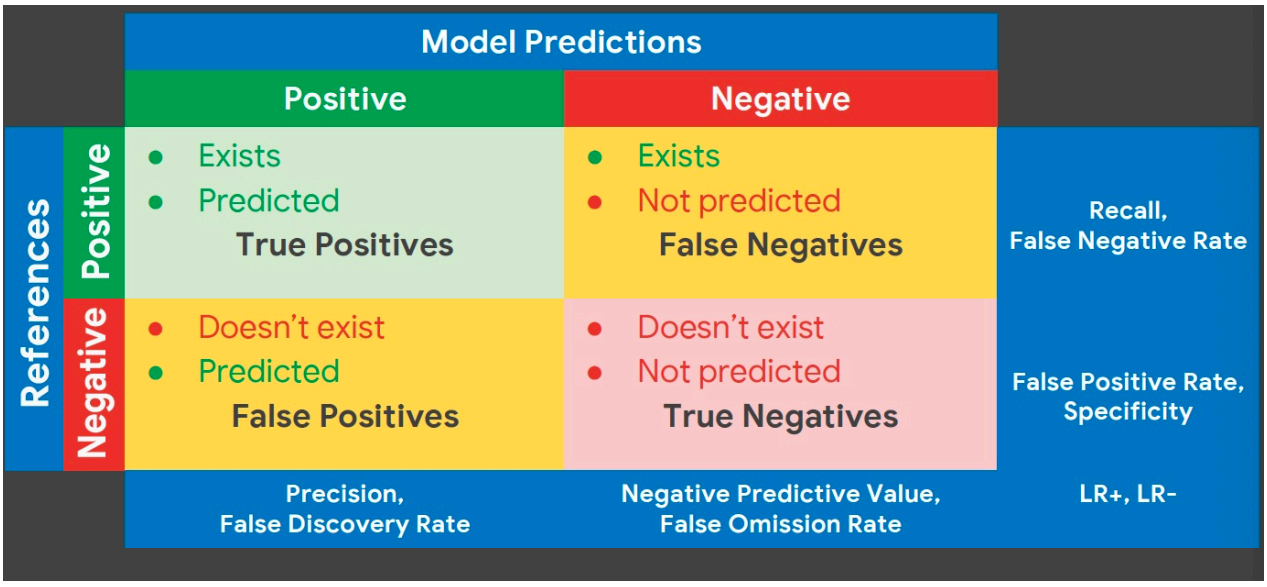
- Selection Bias + Experimenter’s Bias + Correlation Fallacy

Measuring Algorithmic Bias

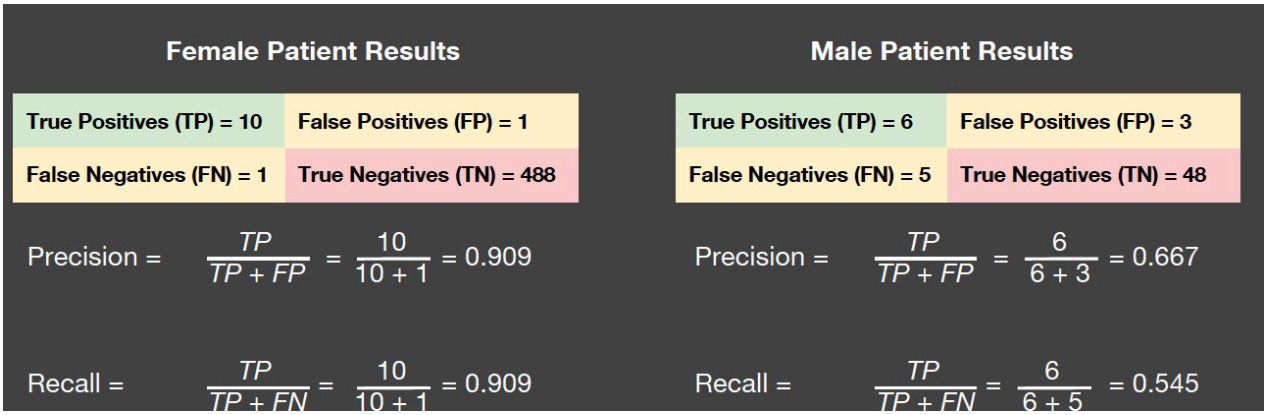
评估公平性和包容性

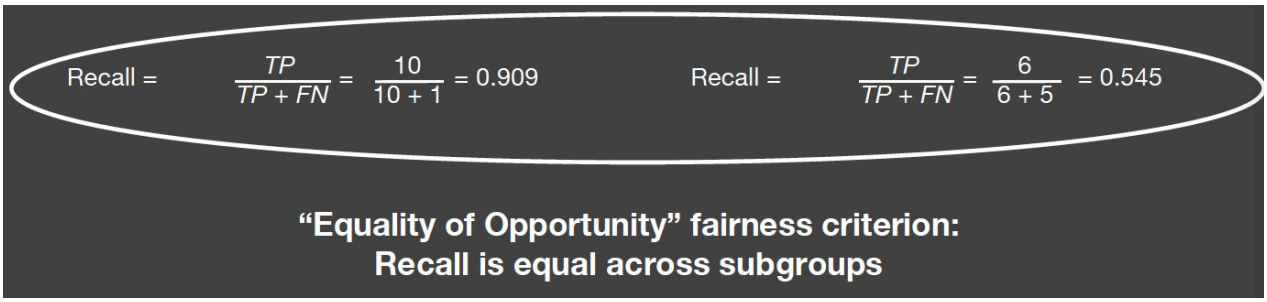
- 分类评估
  - 为每个创建（子组，预测）对。跨子组比较
  - 例如
    - 女性，面部检测
    - 男性，面部检测
- 交叉评估
  - 为每个创建（子组1，子组2，预测）对。跨子组比较
  - 例如
    - 黑人女性，面部检测
    - 白人，面部检测

Evaluate for Fairness & Inclusion: Confusion Matrix

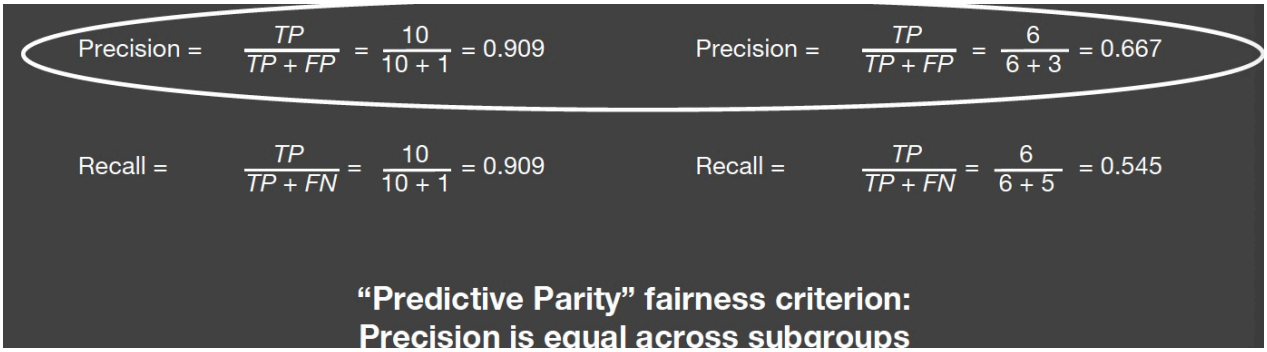


Evaluate for Fairness & Inclusion





- “机会平等”公平准则：子组的 recall 是相等的

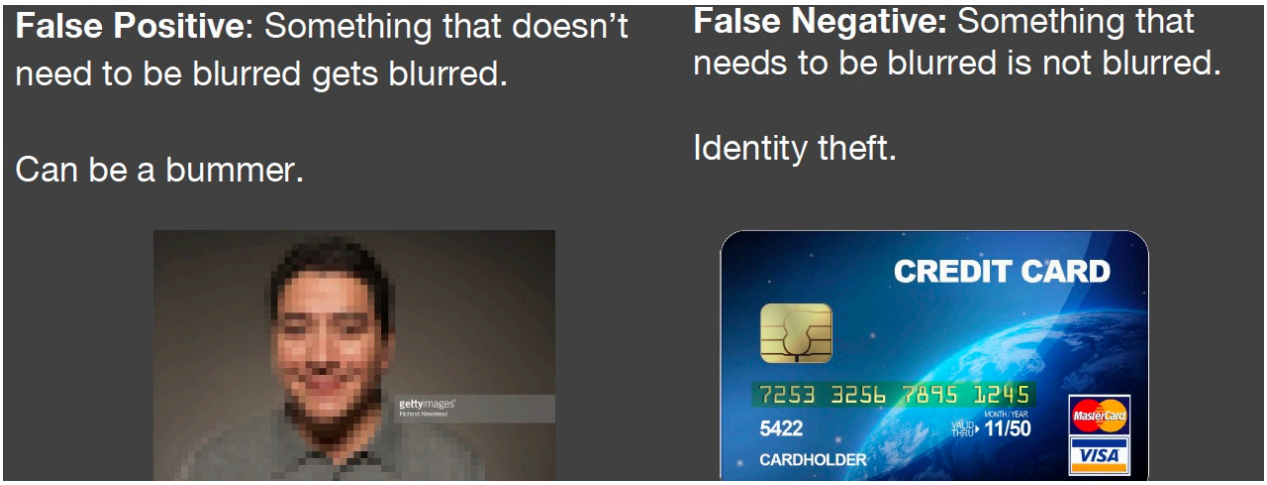


- “预测平价”公平准则：子组的 precision 是相等

选择评价指标的可接受的假阳性和假阴性之间的权衡

False Positives Might be Better than False Negatives

- Privacy in Images



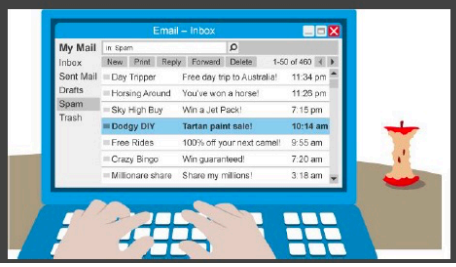
- Spam Filtering

False Negative: Email that is SPAM is not caught, so you see it in your inbox.

Usually just a bit annoying.

False Positive: Email flagged as SPAM is removed from your inbox.

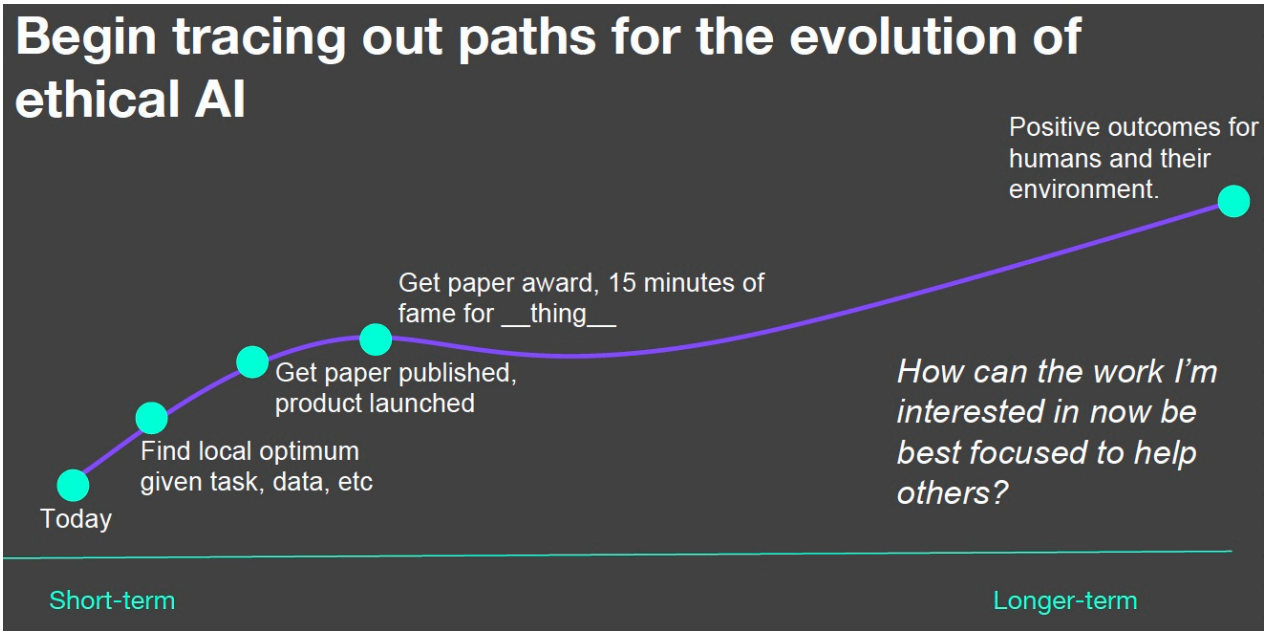
If it's from a friend or loved one, it's a loss!



AI Can Unintentionally Lead to Unjust Outcomes

- 缺乏对数据和模型中的偏见来源的洞察力
- 缺乏对反馈循环的洞察力
- 缺乏细心，分类的评价
- 人类偏见在解释和接受结果

It's up to us to influence how AI evolves.



Here are some things we can do.

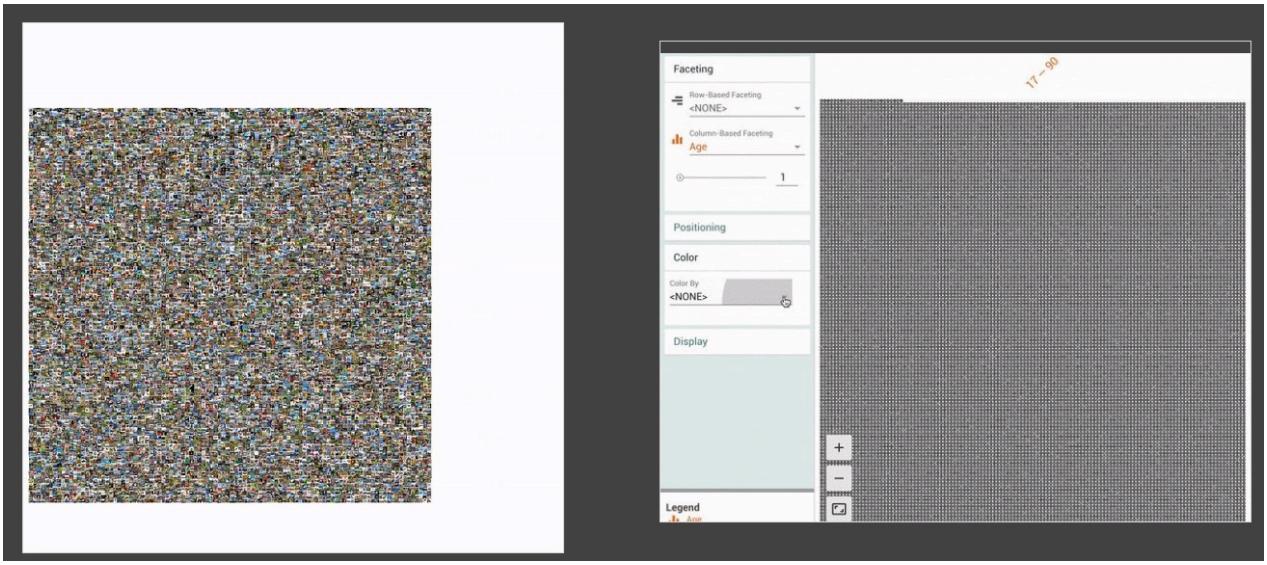
Data

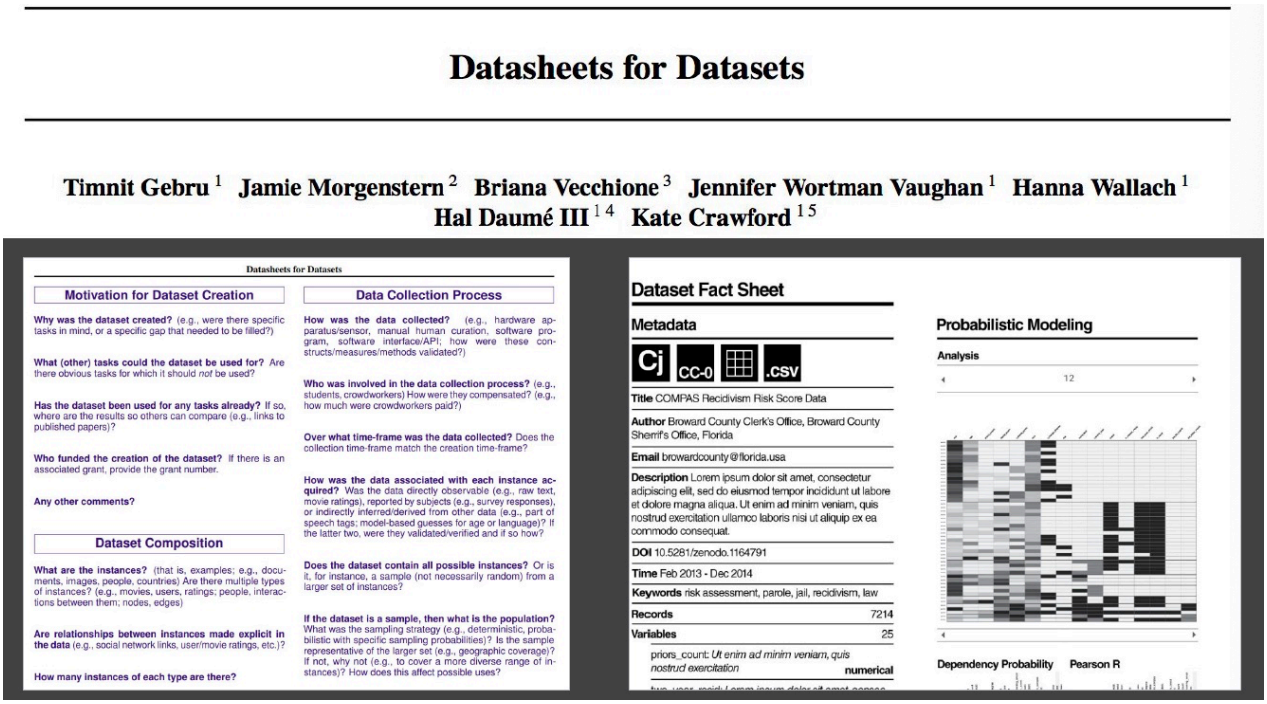
Data Really, Really Matters

- 了解您的数据：偏差，相关性
- 从类似的分布放弃单一训练集/测试集
- 结合来自多个来源的输入
- 对于困难的用例使用held-out测试集
- 与专家讨论其他信号

Understand Your Data Skews

Facets: pair-code.github.io





Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

没有一个数据集是没有偏差的，因为这是一个有偏差的世界。重点是知道是什么偏差。

Machine Learning

Use ML Techniques for Bias Mitigation and Inclusion

Bias Mitigation 偏差缓解

- 删除有问题的输出的信号
  - 刻板印象
  - 性别歧视,种族歧视,*-ism
  - 又称为“debiasing”

Inclusion

- 添加信号所需的变量
  - 增加模型性能

    - 注意性能很差的子组或数据片

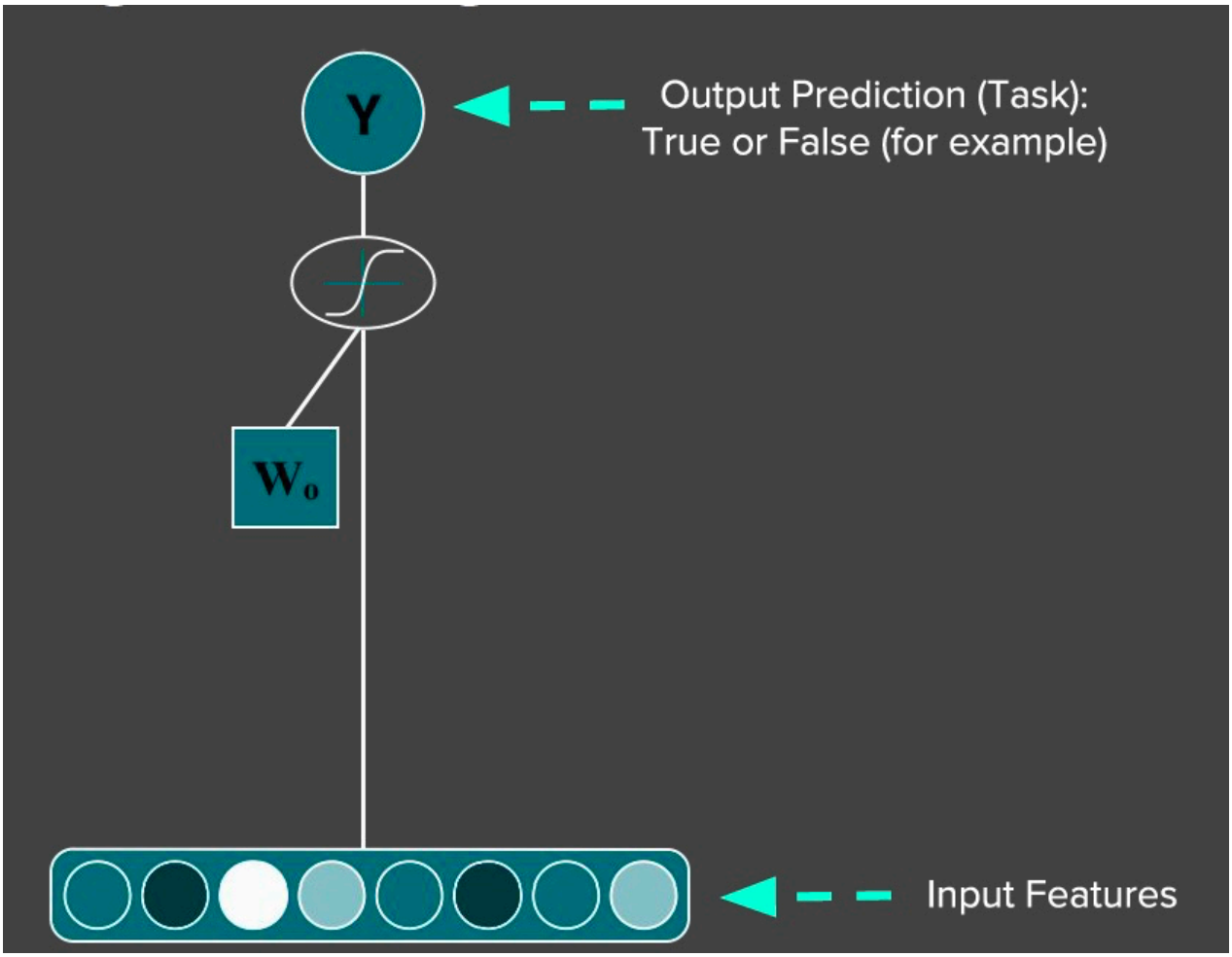
Multi-task Learning to Increase Inclusion

Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

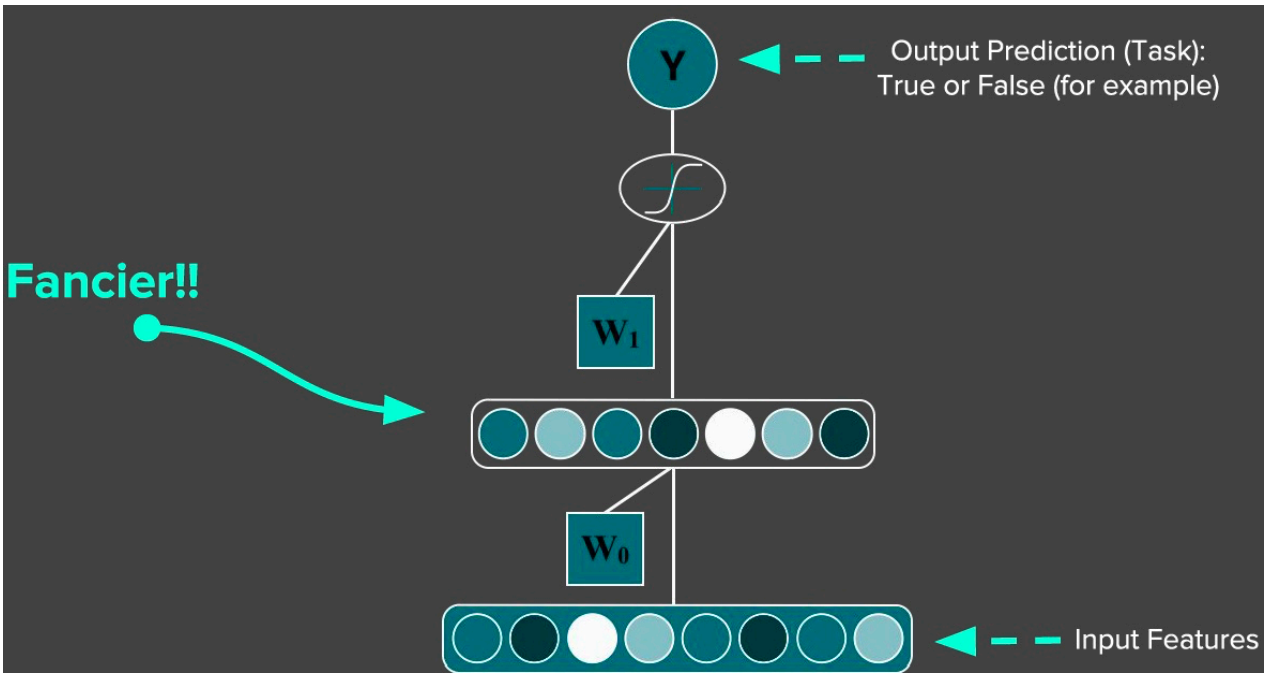
Benton, Mitchell, Hovy. Multi-task learning for Mental Health Conditions with Limited Social Media Data. EACL, 2017.

- 与宾夕法尼亚大学WWP合作
- 直接与临床医生合作
- 目标
    - 系统警告临床医生如果企图自杀迫在眉睫
    - 几个训练实例可用时诊断的可行性
- 内部数据
    - 电子健康记录
        - 病人或病人家属提供
        - 包括心理健康诊断,自杀企图,竞赛
    - 社交媒体数据
- 代理数据
    - Twitter 媒体数据
        - 代理心理健康诊断中使用自称诊断
            - 我被诊断出患有 X
            - 我试图自杀

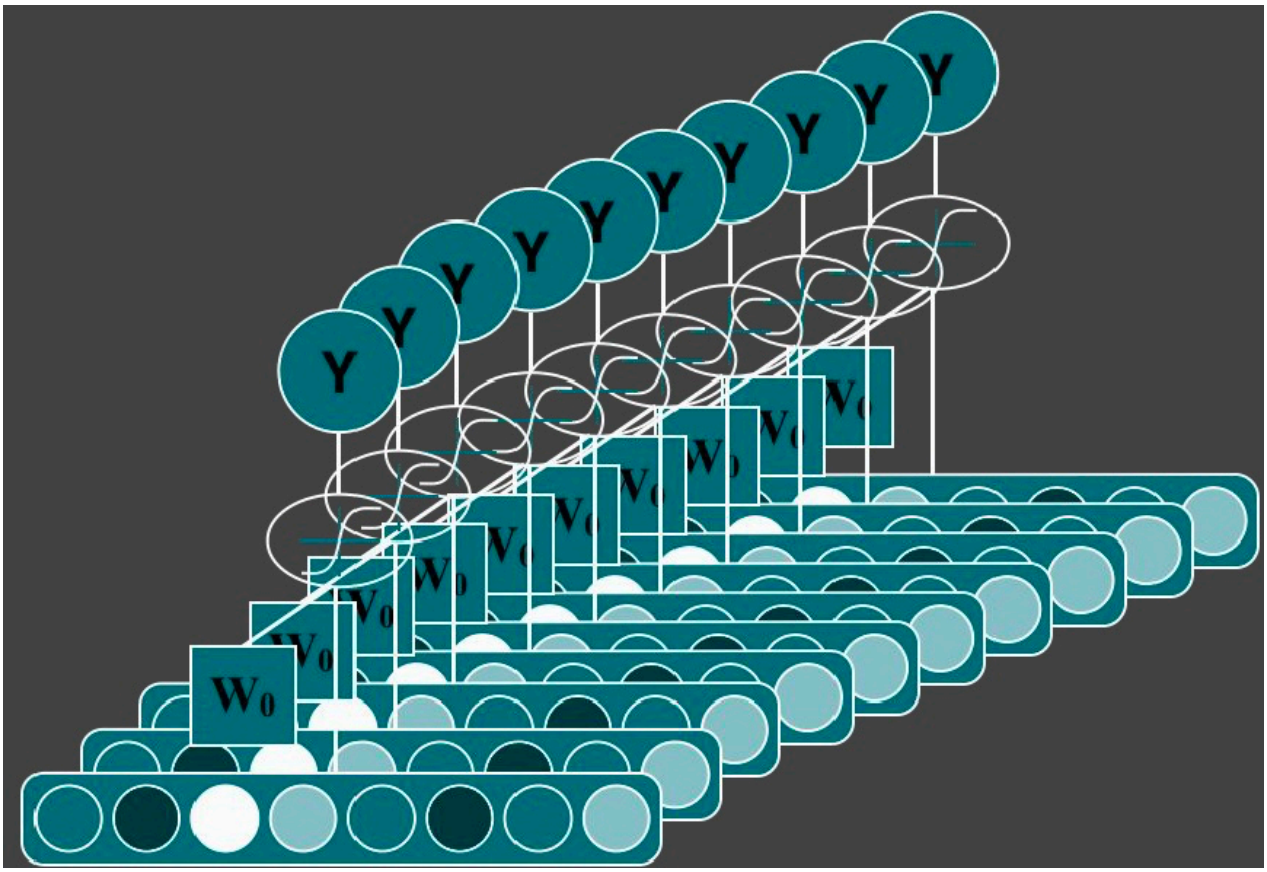
Single-Task: Logistic Regression



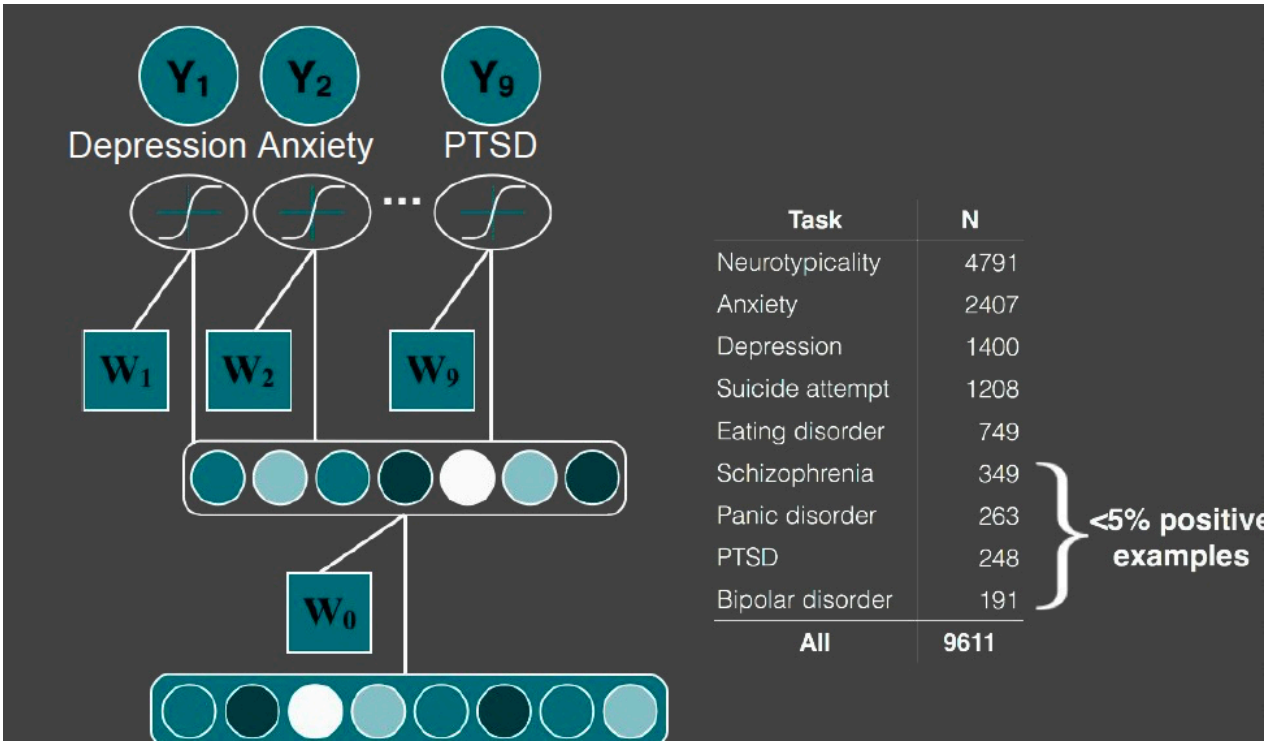
Single-Task: Deep Learning



Multiple Tasks with Basic Logistic Regression



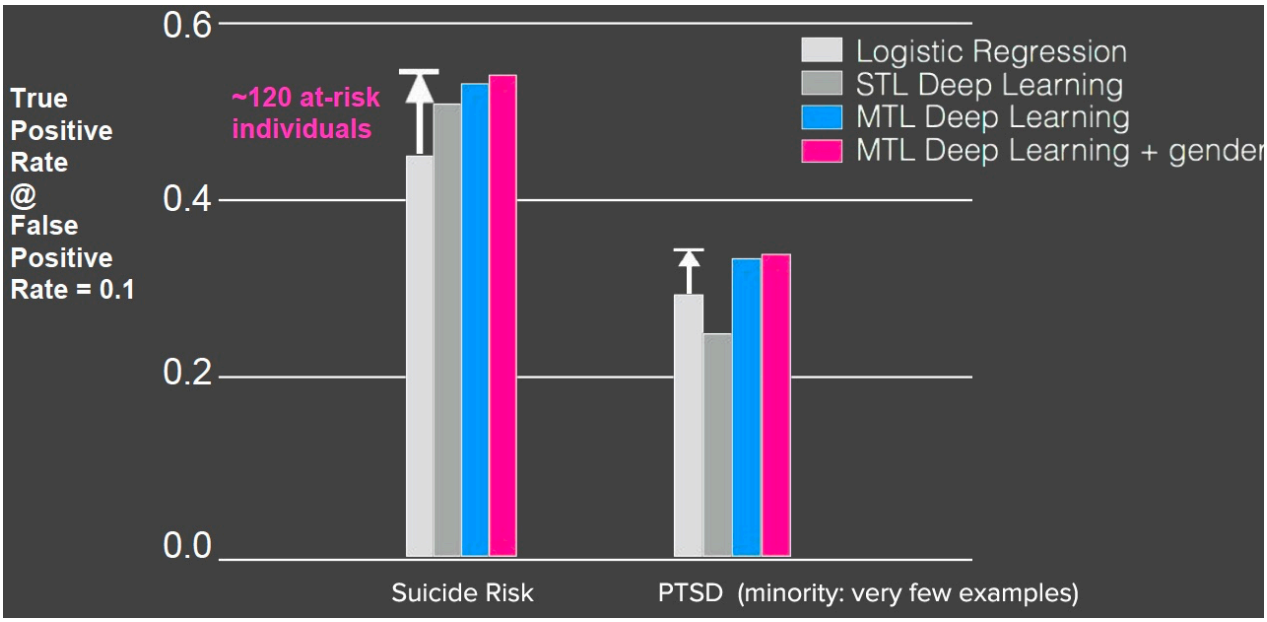
Multi-task Learning



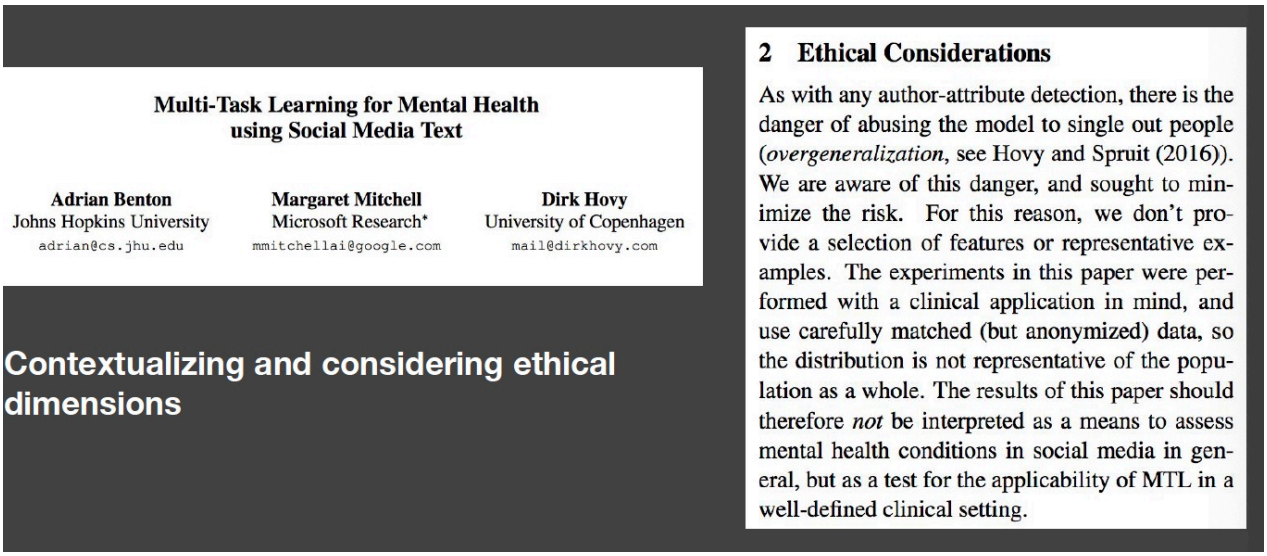
Multitask, given cormorbidity

Benton, Mitchell, Hovy. Multi-task learning for Mental Health Conditions with Limited Social Media Data. EACL, 2017.

Improved Performance across Subgroups

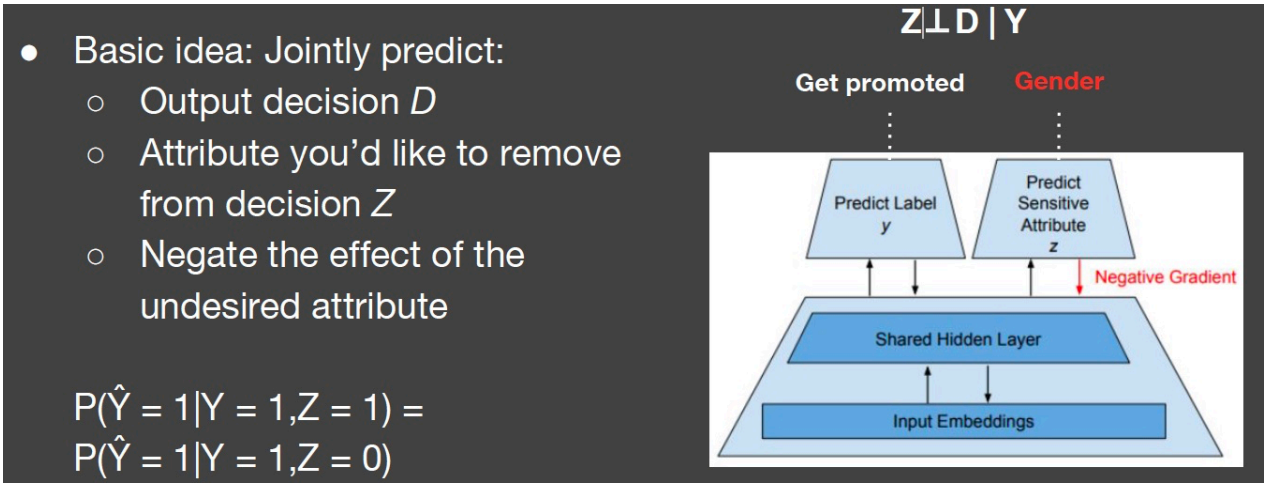


Reading for the masses....



Adversarial Multi-task Learning to Mitigate Bias

Multitask Adversarial Learning



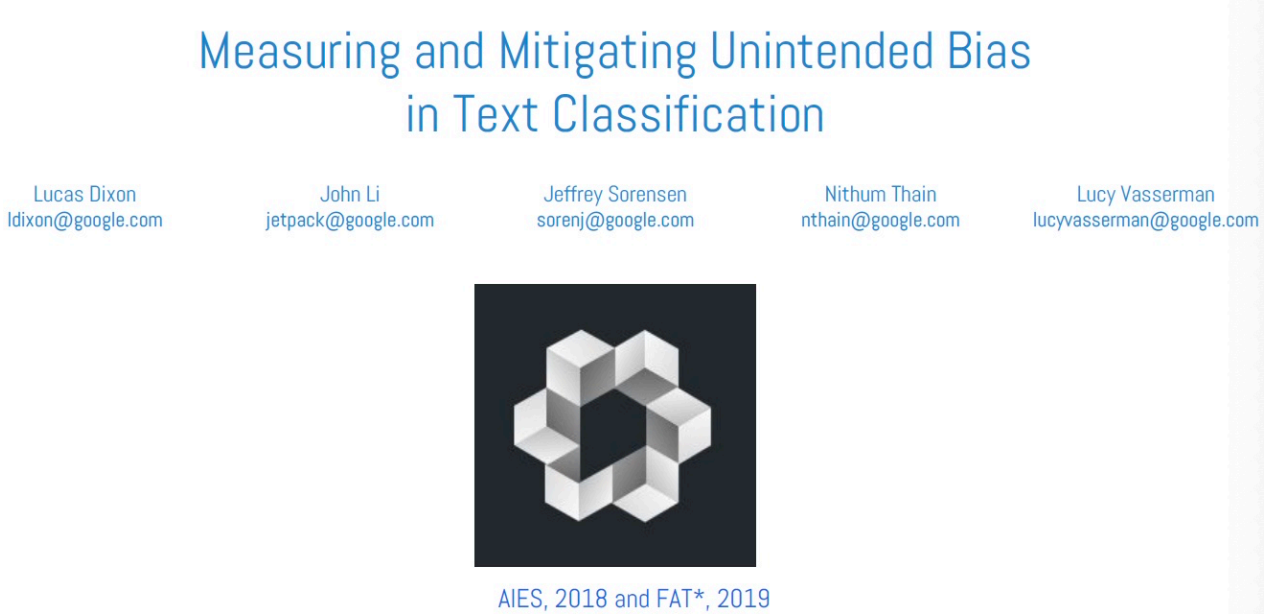
Beutel, Chen, Zhao, Chi. [Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations](). FAT/ML, 2017.

Zhang, Lemoine, Mitchell. Mitigating Unwanted Biases with Adversarial Learning. AIES, 2018.

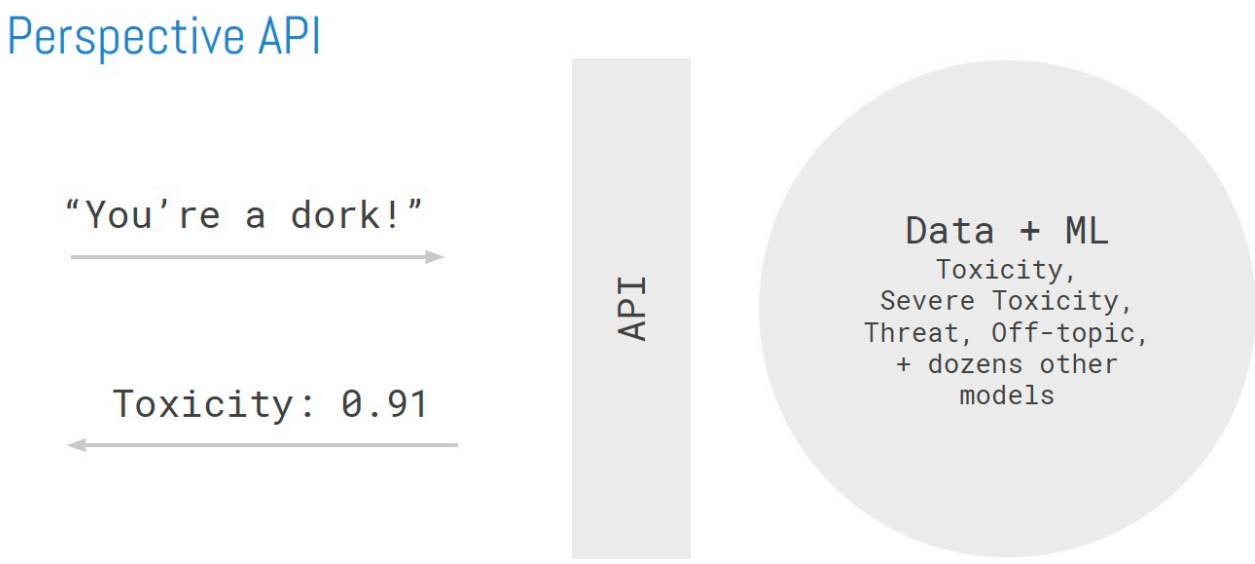
Equality of Opportunity in Supervised Learning

考虑到真正正确的决策，分类器的输出决策应该在敏感特征之间是相同的。

Case Study: Conversation AI Toxicity



- Conversation-AI
  - ML 提高在线对话的规模
- Research Collaboration
  - Jigsaw, CAT, several Google-internal teams, and external partners (NYTimes, Wikimedia, etc)



http://perspectiveapi.com/

Unintended Bias

模型错误地将经常受到攻击的身份与毒性联系起来：False Positive Bias

Sentence	model score
"i'm a proud **tall** person"	0.18
"i'm a proud **lesbian** person"	0.51
"i'm a proud **gay** person"	0.69

Bias Source and Mitigation

- 偏见造成的数据不平衡
  - 经常袭击了有毒的身份所占比例评论长度问题
- 添加维基百科文章中假定的无毒数据来修复这种不平衡
  - 原始数据集有127820个例子
  - 4620个补充的无毒例子

	Comment Length				
Term	**20-59**	**60-179**	**180-539**	**540-1619**	**1620-4859**
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Measuring Unintended Bias - Synthetic Datasets

挑战与真实数据

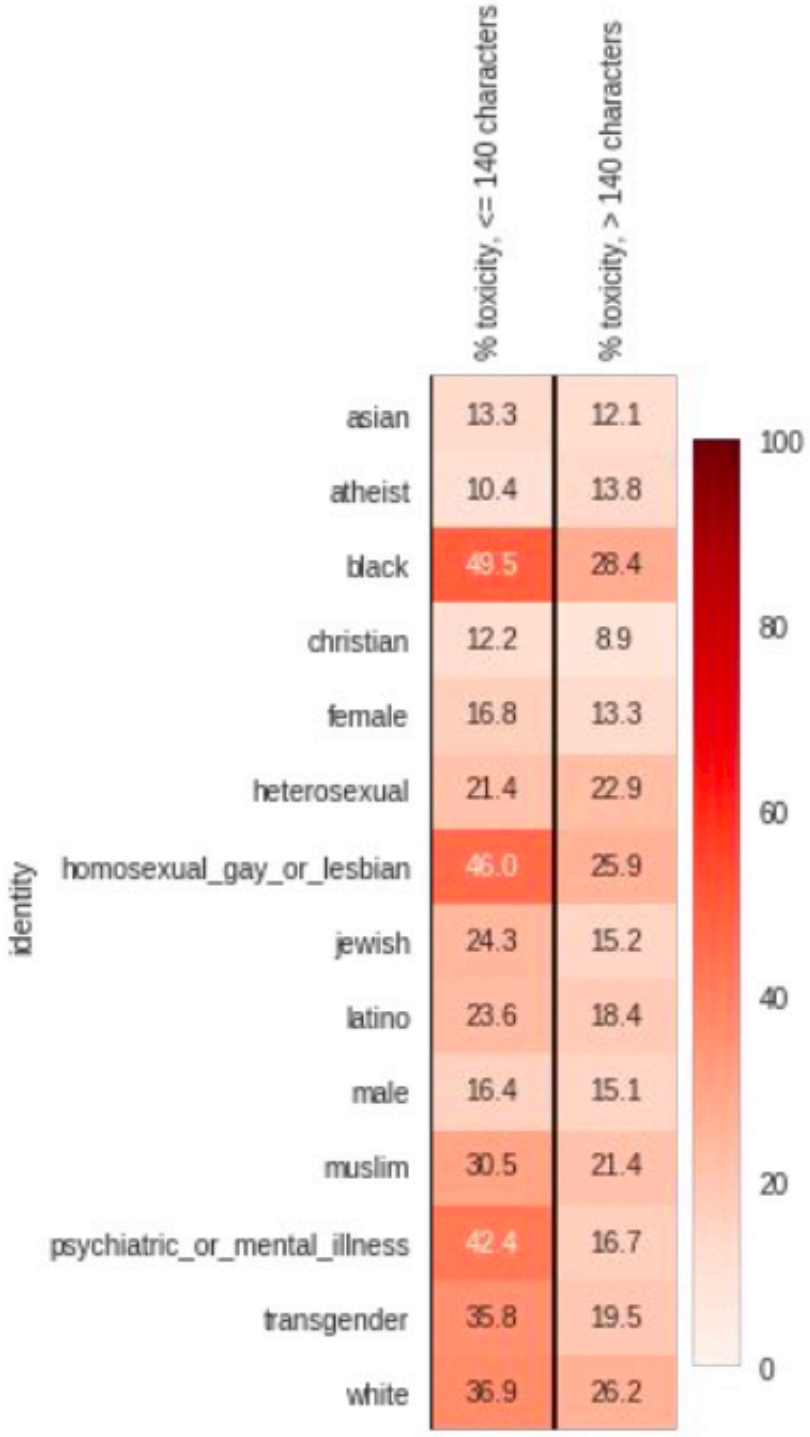
- 现有数据集是小 and/or 有错误的相关性

- 每个例子是完全独特的

Approach: "bias madlibs”: 一个综合生成的模板化数据集进行评估

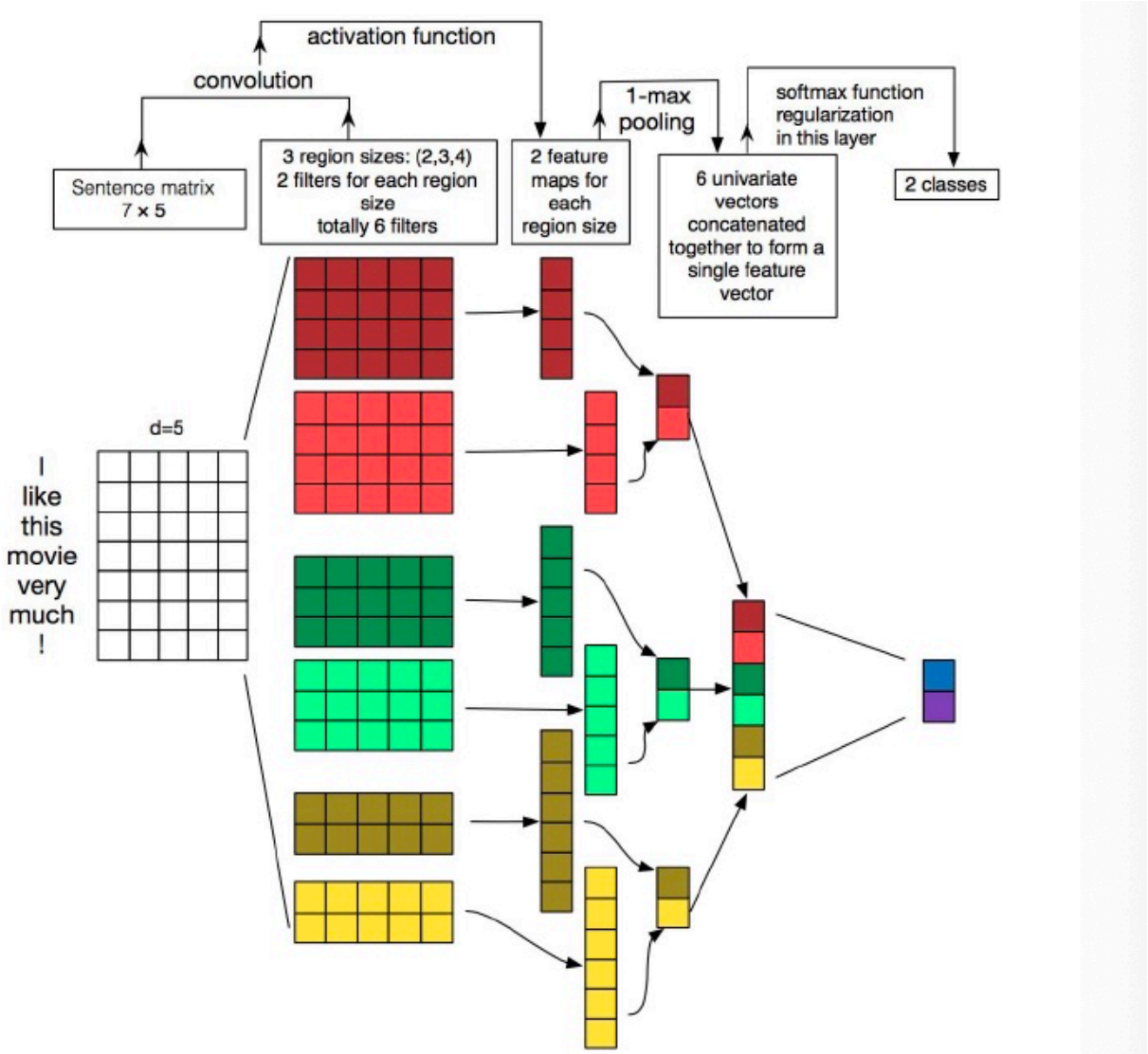
Sentence	model score
"i'm a proud **tall** person"	0.18
"i'm a proud **lesbian** person"	0.51
"i'm a proud **gay** person"	0.69
"audre is a **brazilian** computer programmer"	0.02
"audre is a **muslim** computer programmer"	0.08
"audre is a **transgender** computer programmer"	0.56

Assumptions



- 数据集是可靠的
  - 和产品相似的分布
  - 忽略注释器偏见
  - 没有因果分析

Deep Learning Model



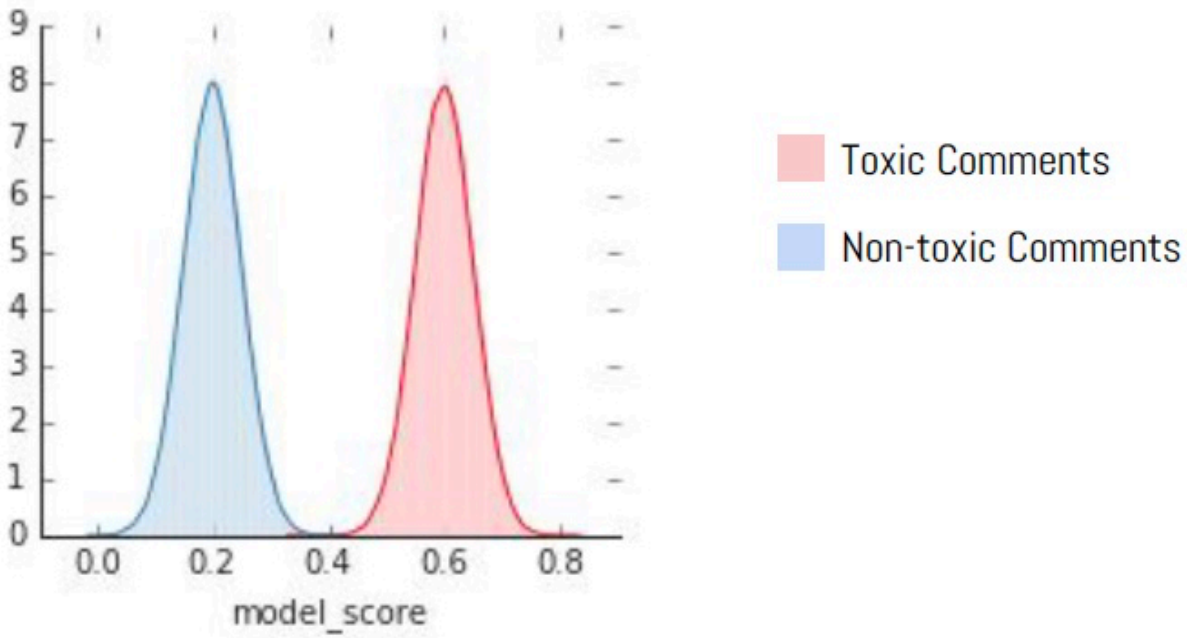
Source: Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

- CNN 架构
- 预训练的 GloVe 嵌入
- Keras 实现

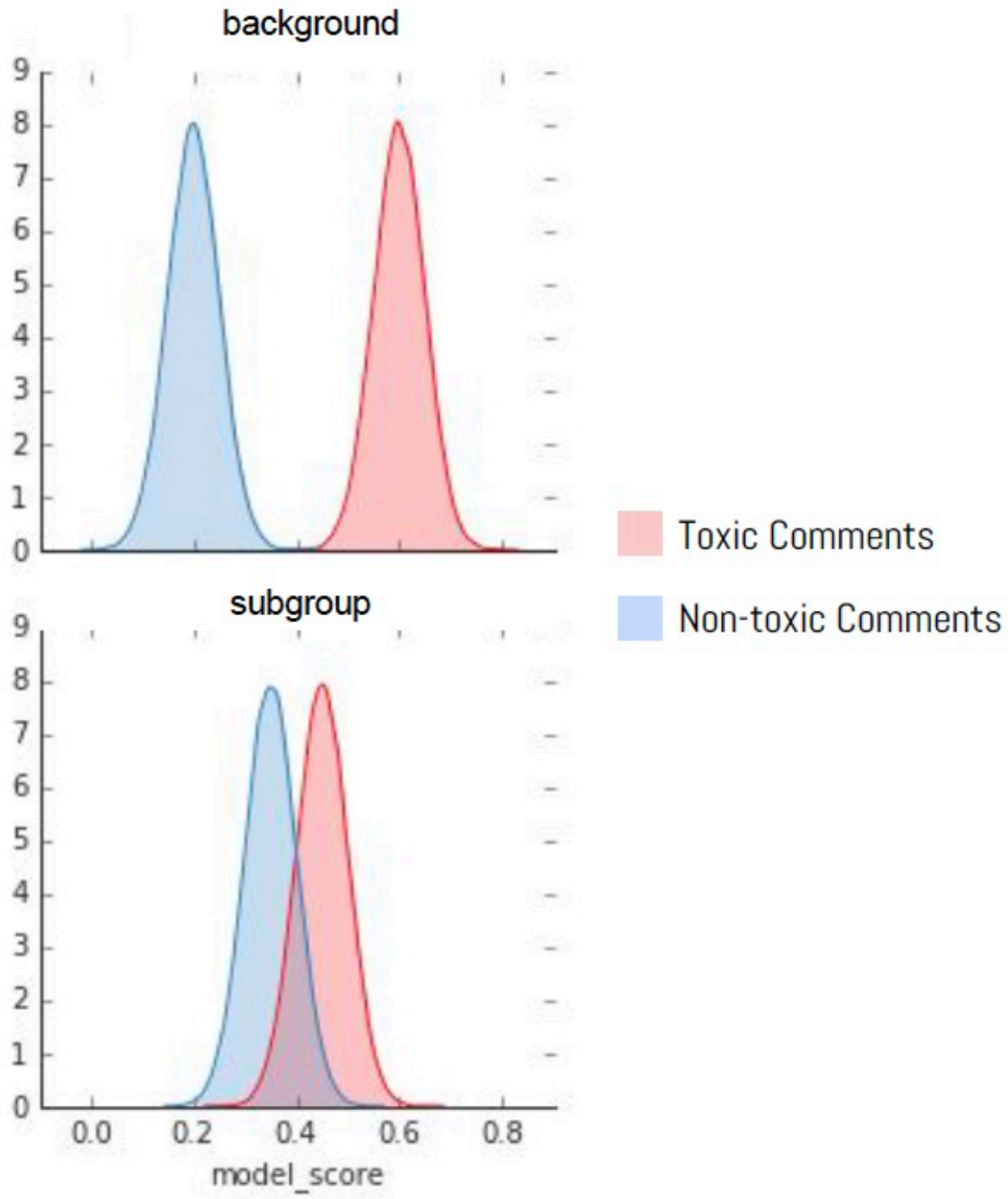
Measuring Model Performance

How good is the model at distinguishing good from bad examples? (ROC-AUC)

AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.



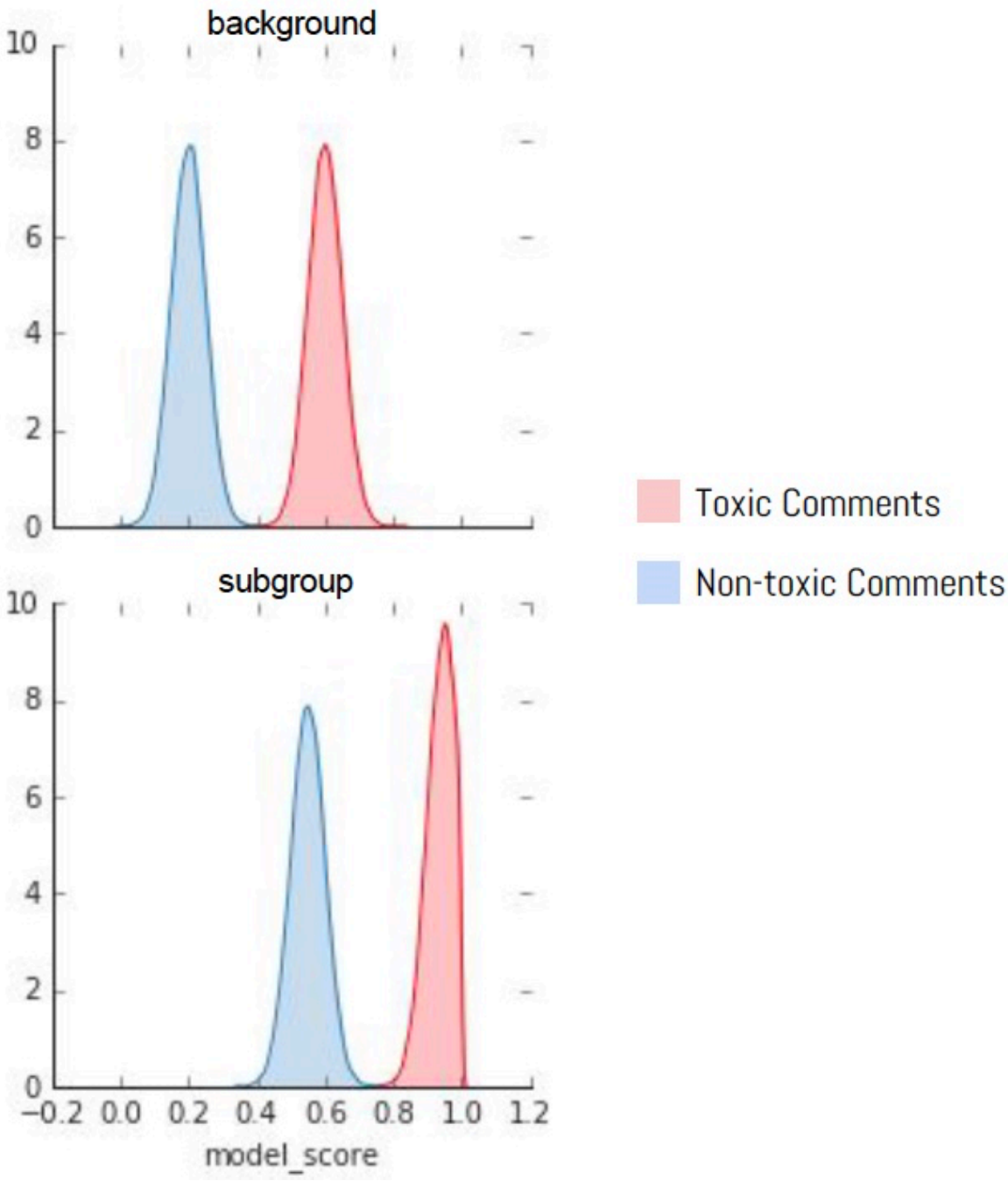
Types of Bias



Low Subgroup Performance

模型在子组注释上的性能比在总体注释上差

Metric : Subgroup AUC

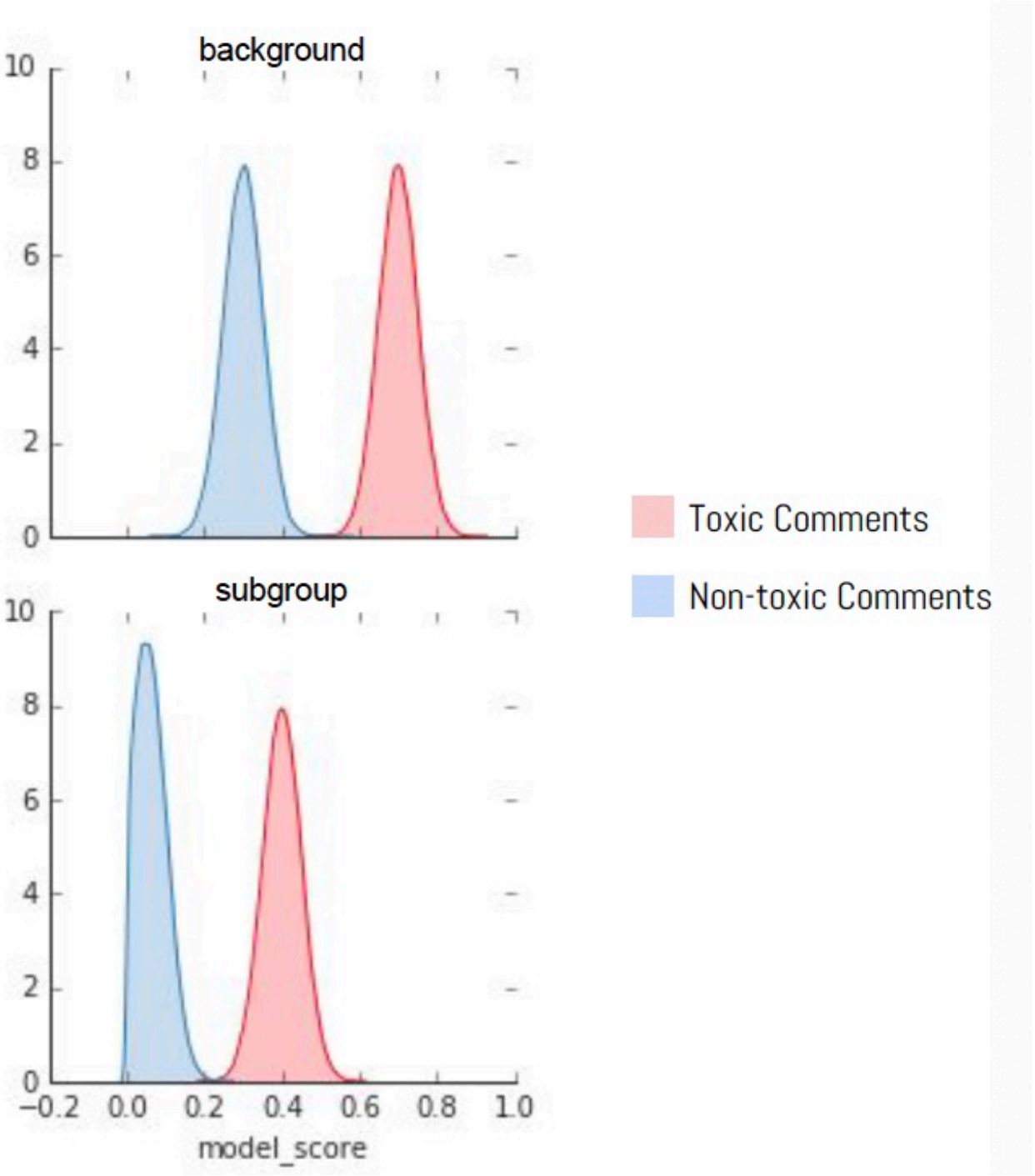


Subgroup Shift (Right)

该模型系统地对来自子组的评价打分更高

Metric: BPSN AUC

(Background Positive Subgroup Negative)



Subgroup Shift (Left)

该模型系统地对来自子组的评价打分较低。

Metric: BNSP AUC

(Background Negative Subgroup Positive)

Results

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BPSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94

0.5 0.6 0.7 0.8 0.9 1.0

Toxicity @6

Identity groups	Subgroup AUC	BPSN AUC	BPSP AUC
lesbian	1.00	0.98	1.00
gay	1.00	0.94	1.00
queer	0.99	0.98	0.99
straight	1.00	1.00	0.97
bisexual	0.98	0.98	0.99
homosexual	1.00	0.96	1.00
heterosexual	1.00	0.99	1.00
cis	1.00	1.00	0.98
trans	1.00	1.00	1.00
nonbinary	1.00	1.00	0.98
black	0.98	0.97	1.00
white	0.99	0.99	0.99

0.5 0.6 0.7 0.8 0.9 1.0

Release Responsibly

Model Cards for Model Reporting

目前还没有模型发布时报告模型效果的common practice

- What It Does
  - 一份关注模型性能透明度的报告，以鼓励负责任的人工智能的采用和应用。
- How It Works
  - 这是一个容易发现的和可用的工件在用户旅程中重要的步骤为一组不同的用户和公共利益相关者。
- Why It Matter
  - 它使模型开发人员有责任发布高质量和公平的模型。

Intended Use, Factors and Subgroups

Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.

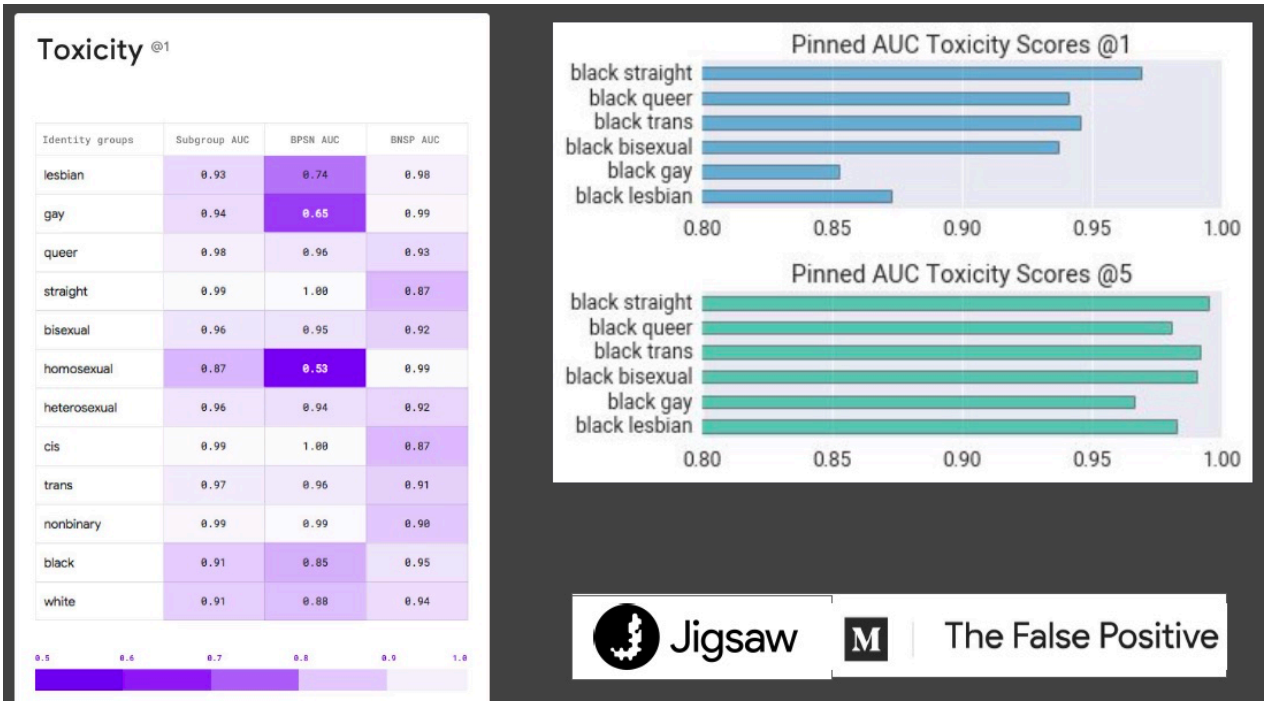
Metrics and Data

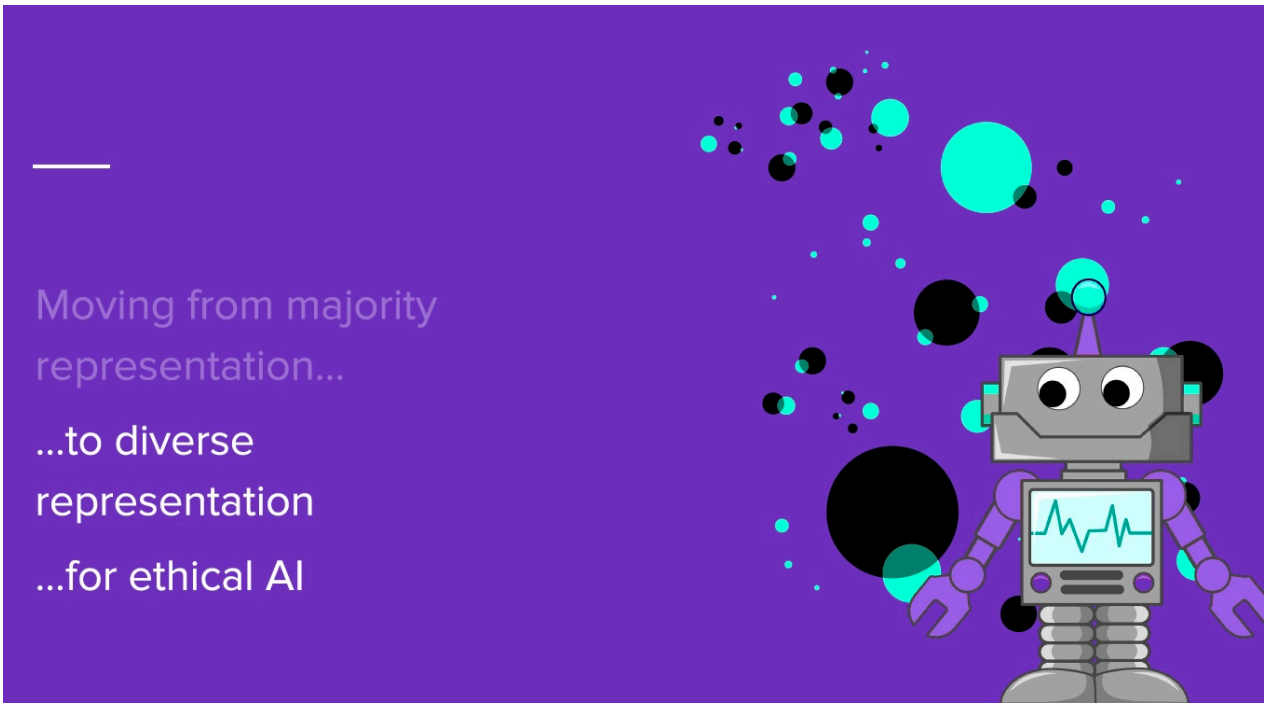
Metrics	*Pinned AUC*, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums with crowdsourced labels of whether the comment is “toxic”. “Toxic” is defined as, “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.

Considerations, Recommendations

Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & Recommendations	Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

Disaggregated Intersectional Evaluation





Reference

以下是学习本课程时的可用参考书籍：

《基于深度学习的自然语言处理》（车万翔老师等翻译）

《神经网络与深度学习》

以下是整理笔记的过程中参考的博客：

斯坦福CS224N深度学习自然语言处理2019冬学习笔记目录 (课件核心内容的提炼，并包含作者的见解与建议)

斯坦福大学 CS224n自然语言处理与深度学习笔记汇总 {>>这是针对note部分的翻译<<}