lecture19 Bias in AI——Margaret Mitchell from Google AI (guest lecture)

Prototype Theory

我们倾向于注意和谈论非典型事物。
例如“doctor”和“female doctor”。大多人，包括男人、女人和自诩女权主义者，都忽视了医生是女性的可能性。

这也影响了我们在文本学习是可以学到的东西
World learning from text. Gordon and Van Durme, 2013

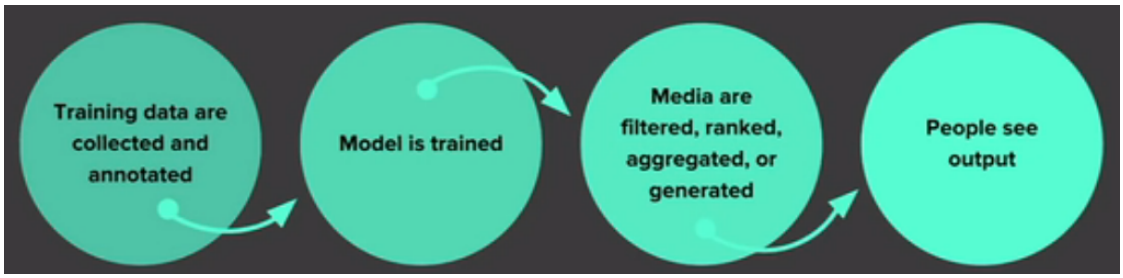
Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

“murdered”出现的次数是“blinked”的十倍。原因在于人们倾向于不提及那些典型事物，我们不倾向于提及“blinked”或者“breathing”。机器从我们发布的文本中学习到的东西也受到了人为的影响。

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。但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。

在典型机器学习范例中，第一步是收集并注释训练数据，然后训练模型，之后进行分类生成等，再之后得到输出。



这一系列活动中，即便是在数据本身都存在不同类型的人类偏见。然后，当我们收集和注释数据时，进一步的偏见被引用，像抽样错误、确认偏见等

Biases in Data

- **Selection Bias**：选择不反映随机样本
- **Out-group homogeneity bias**：这种倾向会让out-group成员看起来比in-group成员更相似。

数据中的偏见导致有偏见的数据表示和有偏见的标签

Biases in Interpretation

- **Confirmation bias** ：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**：根据过于笼统或不够具体的信息得出结论
- **Correlation fallacy**：混淆相关性和因果关系
- **Automation bias**：倾向于喜欢来自自动化决策系统的建议，而不是人类给出的建议，即使面对相互矛盾的信息。

Bias Network Effect

Predicting Future Criminal Behavior

- 算法预测可能发生犯罪的地方以部署警力
- 训练数据基于警察已经离开并逮捕过的地方。因此系统只是简单的学习这种带有偏见的模式，其他可能被发生犯罪的地区并没有被探索出来。

Predicting Sentencing预测量刑
大多数被告在监狱里填了COMPAS的调查问卷，答案输入系统生成与累犯风险相对应的分数。这些问题用于收集关于被告的社会经济地位、家庭背景、邻里犯罪、就业状况以及其他因素的数据，来预测个人犯罪或犯罪风险。
黑人比白人更容易被定罪，即使犯了同样的罪行。

Predicting Criminality

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

预测犯罪行为，尤其通过面部图像来预测。
Faception：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。

“Automated Inference on Criminality using Face Images” Wu and Zhang, 2016.他们声称即使是基于非常小的训练数据集，他们能够预测一个人是否是罪犯，准确率超过90%。
存在确认偏差和相关性偏差

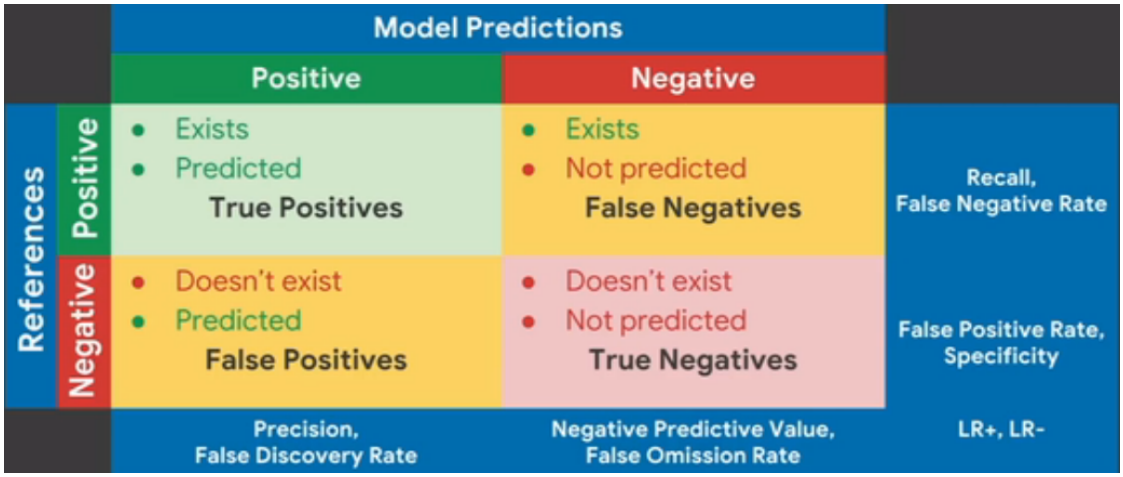
Predict Internal Qualities Subject To Discrimination

- Selection Bias + Experimenter’ s Bias + Correlation Fallacy

Wang and Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, 2017. 基于一张面部图像预测某人是否是同性恋。他们使用的培训数据来源是约会网站图片，人们自认为是straight或gay来寻找伴侣。其中存在的问题更多的与社交自我的呈现有关，而与日常生活的方式关系不大。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

Measuring Algorithmic Bias

- 在分类评估中，评估的是不同的子群而不是整个测试数据集的单个分数。
  单个分数掩盖了系统在不同类型的个人或子群实际上的表现。
  因此为每个创建(子组，预测)对
- 交叉评估



这些指标实际上很容易映射到很多不同的公平指标
机会均等：不同子组有相同的召回率
Predictive Parity预测奇偶性的公平性标准：不同子组有相同的精度

选择评价指标
False Positives Might be Better than False Negatives的例子：图像隐私、垃圾邮件过滤

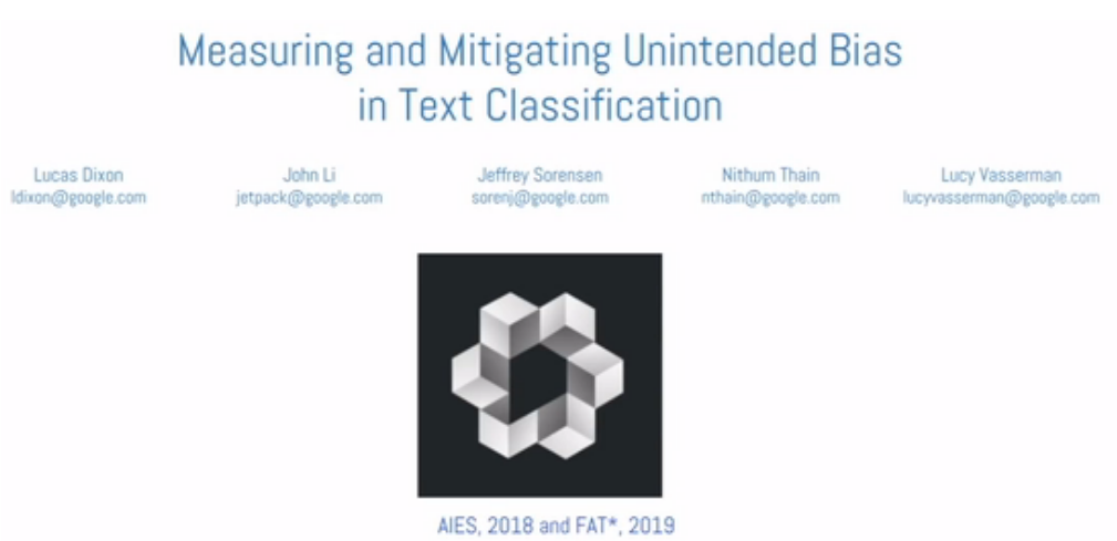
在模型中，缺乏对数据偏差来源的洞察力、缺乏对原始数据的循环反馈的深入了解、缺乏仔细的分类评估、在解释和接受结果时的人类偏见以及进一步的媒体炒作，人工智能无意地造成了不公平的结果。

THINGS WE CAN DO

Short-term：正在研究一些特定的模型，试图找到局部最优
-> 然后发表论文或者推出产品 -> 得奖或因此出名
Longer-term：如何更好的专注于帮助他人

1. Data.
    - 了解数据偏差和相关性，使数据更有代表性。
    - 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
2. Machine Learning
    - Bias Mitigation 偏差缓解。移除有问题的输出信号。也被称为去偏置de-biasing
    - Inclusion 为所需变量添加信号。将注意力添加在表现差的子组或数据片。一种相对较好的技术是多任务学习。
      例：一个可以提醒临床医生是否又即将发生的自杀倾向的系统。数据包括内部数据即由病人或病人家属提供的
      包括心理健康诊断,自杀企图的电子健康记录。与外部数据，是基于Twitter的代理数据
    - Adversarial Multi-task Learning
      一个head预测主要任务，另一个head预测[不想其影响模型预测的]东西。

Case Study: Conversation AI Toxicity



- Conversation-AI。使用深度学习来改善在线对话。perspectiveapi.com。输入一个短句，输出toxicity score
  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

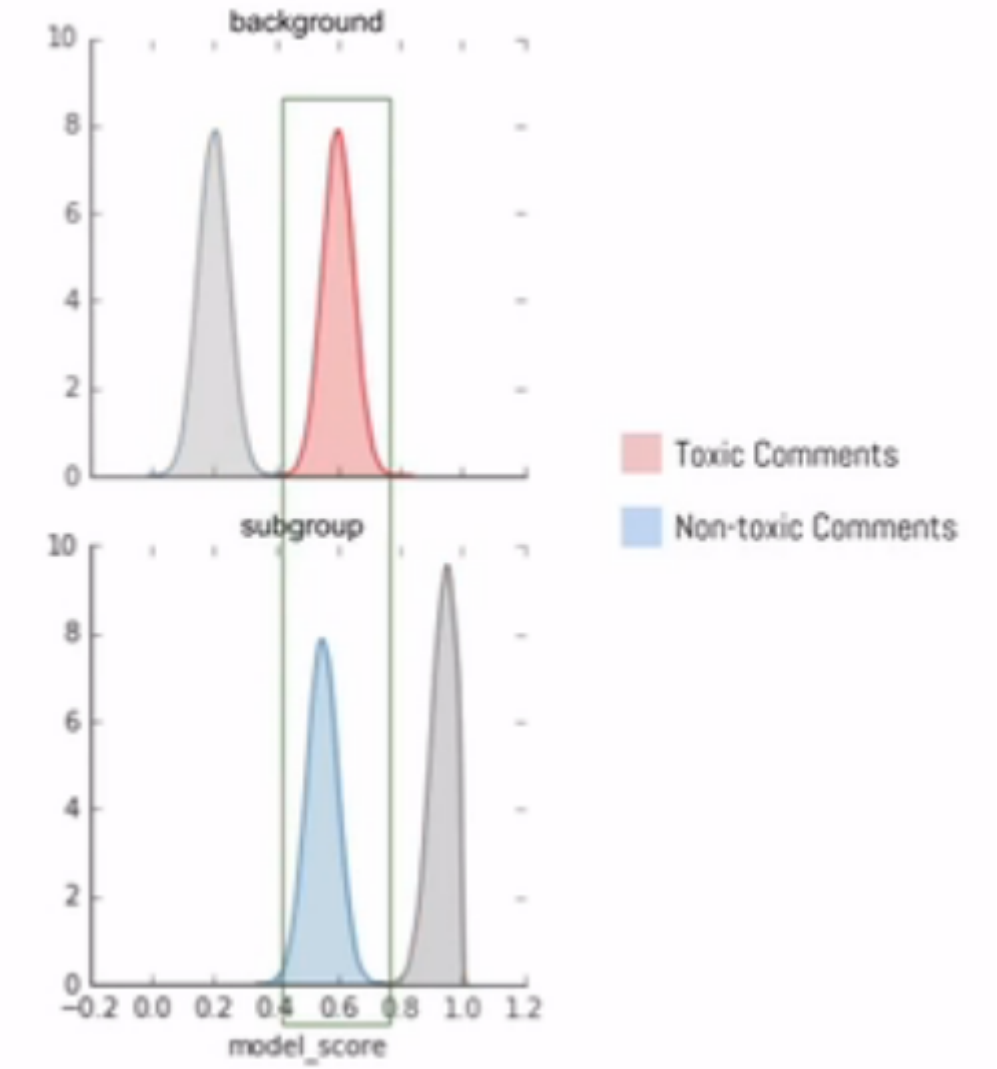
Sentence	model score
"i'm a proud **tall** person"	0.18
"i'm a proud **lesbian** person"	0.51
"i'm a proud **gay** person"	0.69

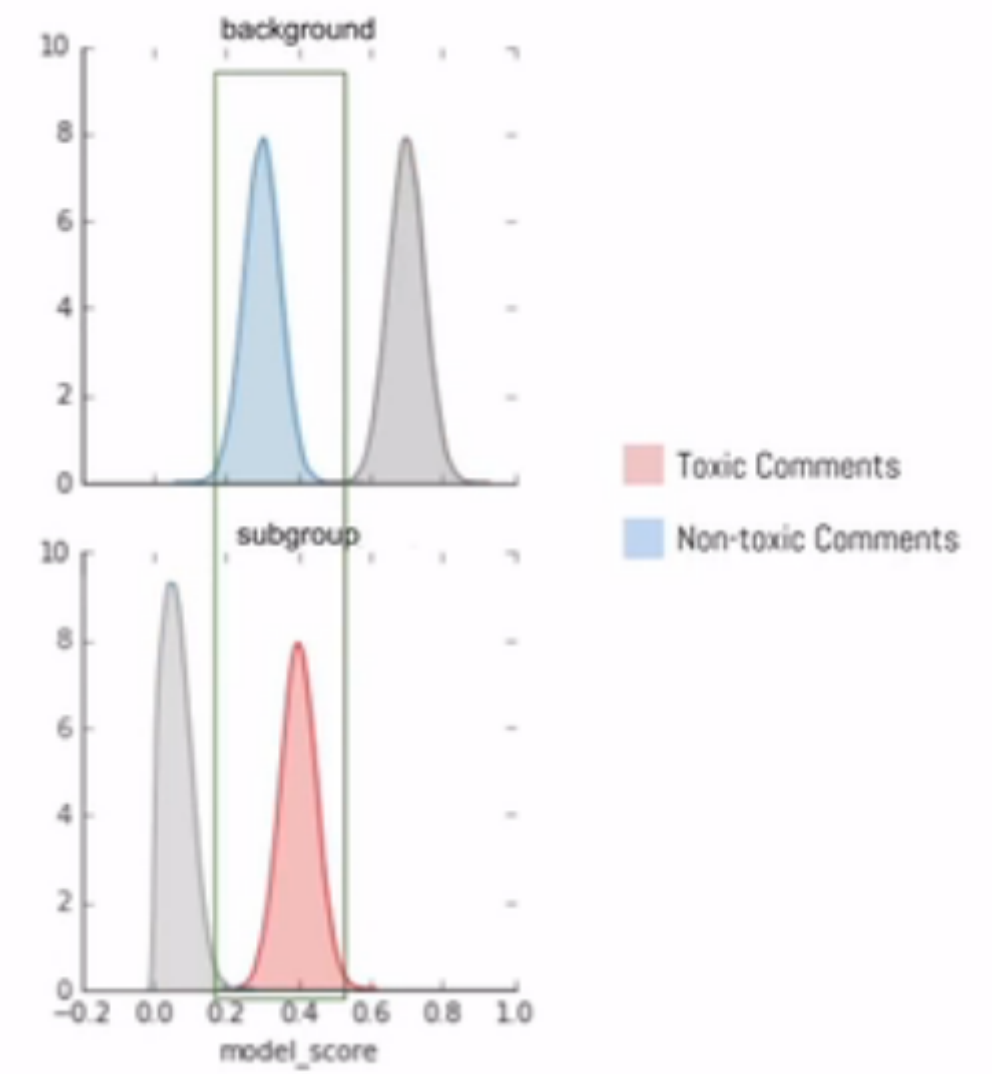
这种偏差很大程度上是由数据集失衡导致的。通过添加维基百科文章中假定的无毒数据来修复这种不平衡。

评估的挑战之一是 没有一种controlled toxicity evaluation的非常好的方法。在现实世界中，任何人都可以猜出特定橘子的toxicity是什么，但是很难获取真正好的数据来正确评估。

定义了不同类型的偏见

- low subgroup performance意味着该模型在子组上的表现比它在总体上的更差，为了衡量这一点引入指标Subgroup AUC
- subgroup shift，当模型系统的位来自较高的分的小组评分时





可以为其中每一种定义不同的指标。