

Дизайн систем машинного обучения

**13. Непредвзятость. Безопасность.
Карточки моделей**

План курса

- 1) Практическое применение машинного обучения
- 2) Основы проектирования ML-систем
- 3) Обучающие данные
- 4) Подготовка и отбор признаков
- 5) Выбор модели, разработка и обучение модели
- 6) Оценка качества модели
- 7) Развертывание
- 8) Диагностика ошибок и отказов ML-систем
- 9) Мониторинг и обучение на потоковых данных
- 10) Жизненный цикл модели
- 11) Отслеживание экспериментов и версионирование моделей
- 12) Сложные модели: временные ряды, модели над графами
- 13) Непредвзятость, безопасность, карточки моделей — Вы находитесь здесь**
- 14) ML инфраструктура и платформы
- 15) Интеграция ML-систем в бизнес-процессы

ML Model Cards

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
(mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca

ABSTRACT

Trained machine learning models are increasingly used to perform high-impact tasks in areas such as law enforcement, medicine, education, and employment. In order to clarify the intended use cases of machine learning models and minimize their usage in contexts for which they are not well suited, we recommend that released models be accompanied by documentation detailing their performance characteristics. In this paper, we propose a framework that we call model cards, to encourage such transparent model reporting. Model cards are short documents accompanying trained machine learning models that provide benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type [15]) and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains. Model cards also disclose the context in which models are intended to be used, details of the performance evaluation procedures, and other relevant information. While we focus primarily on human-centered machine learning models in the application

KEYWORDS

datasheets, model cards, documentation, disaggregated evaluation, fairness evaluation, ML model evaluation, ethical considerations

ACM Reference Format:

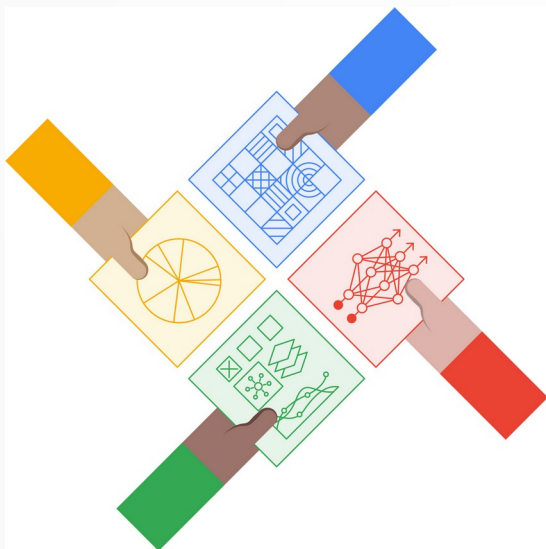
Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru. 2019. Model Cards for Model Reporting. In *FAT* '19: Conference on Fairness, Accountability, and Transparency, January 29–31, 2019, Atlanta, GA, USA*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3287560.3287596>

1 INTRODUCTION

Currently, there are no standardized documentation procedures to communicate the performance characteristics of trained machine learning (ML) and artificial intelligence (AI) models. This lack of documentation is especially problematic when models are used in applications that have serious impacts on people's lives, such as in health care [14, 42, 44], employment [1, 13, 29], education [23, 45] and law enforcement [2, 7, 20, 34].

cs.LG] 14 Jan 2019

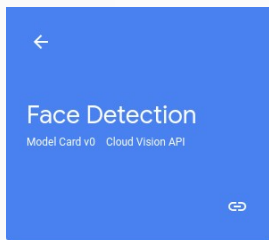
Google Model Card



- Content
- Process
- Experience
- Fairness
- Privacy

<https://modelcards.withgoogle.com/about>

Google Model Card — Face Detection



Overview

Limitations

Trade-offs

Performance

Test your own images

Provide feedback

Explore

[Object Detection](#)

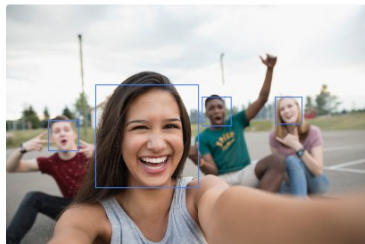
[About Model Cards](#)

Face Detection

The model analyzed in this card detects one or more faces within an image or a video frame, and returns a box around each face along with the location of the faces' major landmarks. The model's goal is exclusively to identify the existence and location of faces in an image. It does not attempt to discover identities or demographics.

On this page, you can learn more about how well the model performs on images with different characteristics, including face demographics, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION

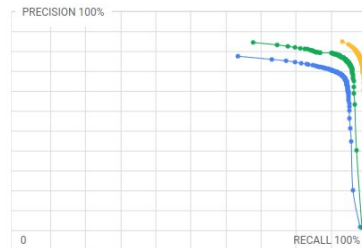


Input: Photo(s) or video(s)

Output: For each face detected in a photo or video, the model outputs:

- Bounding box coordinates
- Facial landmarks (up to 34 per face)
- Facial orientation (roll, pan, and tilt angles)
- Detection and landmarking confidence scores.

PERFORMANCE



- Open Images
- Face Detection Dataset Benchmark
- Labeled Faces in the Wild

Overall model performance, and performance sliced by different image and face characteristics, were assessed, including:

- Derived characteristics (face size, facial orientation, and occlusion)

Hugging Face Model Cards

The screenshot shows the Hugging Face interface for the `bert-base-uncased` model. At the top, the Hugging Face logo and navigation menu are visible. The model name `bert-base-uncased` is prominently displayed with a like count of 351. Below the name, various ecosystem integrations like Fill-Mask, PyTorch, TensorFlow, JAX, Rust, Safetensors, Transformers, bookcorpus, and wikipedia are listed. The model's language is English, and its license is apache-2.0. The main content area is divided into two columns. The left column contains the model's description: "BERT base model (uncased)", a pretraining objective explanation, and a disclaimer. The right column features a "Downloads last month" chart showing 21,893,255 downloads, a "Hosted inference API" section with a "Fill-Mask" example (Paris is the [MASK] of France.), and a "Compute" button.

bert-base-uncased like 351

Fill-Mask PyTorch TensorFlow JAX Rust Safetensors Transformers bookcorpus wikipedia

English arxiv:1810.04805 bert exbert AutoTrain Compatible License: apache-2.0

Model card Files Community 15

Train Deploy Use in Transformers

Edit model card

BERT base model (uncased)

Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in [this paper](#) and first released in [this repository](#). This model is uncased: it does not make a difference between english and English.

Disclaimer: The team releasing BERT did not write a model card for this model so this model card has been written by the Hugging Face team.

Downloads last month
21,893,255

Hosted inference API

Fill-Mask Examples

Mask token: [MASK]

Paris is the [MASK] of France.

Compute

<https://huggingface.co/bert-base-uncased>

<https://huggingface.co/docs/hub/model-cards>

Kaggle Intro to AI Ethics

- Model Details
- Intended Use
- Factors
- Metrics
- Evaluation Data
- Training Data
- Quantitative Analyses
- Ethical Considerations
- Caveats and Recommendations

<https://www.kaggle.com/code/var0101/model-cards/tutorial>

Итого — карточка модели

- Краткий документ для заинтересованных лиц
 - 1-2 страницы
- Управлять ожиданиями от модели
- Предотвратить нецелевое использование
 - Предполагаемое использование
 - Использованные при обучении данные
 - Границы применимости
 - Ожидаемое качество

A smooth chicken liver and pork fat pâté. Ingredients Chicken Liver (36%), Pork Fat, Water, Tapioca Starch, Pork Rind, Dextrose, Salt, Antioxidants (Ascorbic Acid, Sodium Ascorbate), Shallot Powder, Spices, Sugar, Colour (Plain Caramel), Preservative (Sodium Nitrite).

Storage Keep refrigerated. Once opened, consume within 2 days and by 'use by' date shown. Not suitable for home freezing. Packaged in a protective atmosphere. For more information about our strict welfare and quality standards visit tescoplc.com


Produced in the U.K. using chicken and pork from the U.K. for Tesco Stores Ltd., Welwyn Garden City AL7 1GA, U.K.

© Tesco 2018. SC0247H


Nutrition	
As sold	
Typical values	Per 100g
Energy	1082kJ / 261kcal
Fat	22.2g
of which saturates	7.5g
Carbohydrate	4.5g
of which sugars	2.5g
Fibre	0.5g
Protein	10.6g
Salt	1.6g

This pack contains 5 servings
*Reference intake of an average adult (8400kJ / 2000kcal)

LABEL Widely Recycled
LID & TRAY Check Locally
FILM Not Yet Recycled

200g e 

55c



5 054775 703302 >



PG103866

Безопасность ML-систем

- Обычные проблемы безопасности систем
- + Обычные проблемы безопасности компьютерных систем
- + Специфичные для ML проблемы безопасности
- Безопасность обеспечивается на всех этапах жизненного цикла ML-системы

Примеры специфичных для ML атак

- Отравление датасета →
- Извлечение исходных данных →
- Кража моделей →
- Адверсариал атаки в задачах компьютерного зрения →
- Манипулирование алгоритмами ранжирования →
- Атаки на ансамбли деревьев →
- Прохождение фильтров → →

Не только атаки, но и недосмотры

- Модель может быть плохо протестирована
- Может плохо работать для какого-то сегмента
- Может не выдерживать нагрузку
- Может не сообщать об отказах
- Любое нарушение работы модели может повлечь финансовые и репутационные риски, угрожать жизни и здоровью
- А еще дроны могут убивать не тех людей →

Стадии атаки — см MITRE ATLAS

- Reconnaissance
- Resource Development
- Initial Access
- ML Model Access
- Execution Persistence
- Defense Evasion
- Discovery
- Collection
- ML Attack Staging
- Exfiltration
- Impact



https://en.wikipedia.org/wiki/Mitre_Corporation

Adversarial Threat Landscape for AI Systems

ATLAS™

The ATLAS Matrix below shows the progression of tactics used in attacks as columns from left to right, with ML techniques belonging to each tactic below. Click on links to learn more about each item, or view ATLAS tactics and techniques using the links at the top navigation bar.

Reconnaissance	Resource Development	Initial Access	ML Model Access	Execution	Persistence	Defense Evasion	Discovery	Collection	ML Attack Staging	Exfiltration	Impact
5 techniques	7 techniques	3 techniques	4 techniques	1 technique	2 techniques	1 technique	3 techniques	2 techniques	4 techniques	2 techniques	7 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution	Poison Training Data	Evade ML Model	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities	Valid Accounts	ML-Enabled Product or Service		Backdoor ML Model		Discover ML Model Family	Data from Information Repositories	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Adversarial ML Attack Capabilities	Evade ML Model	Physical Environment Access				Discover ML Artifacts	Verify Attack	Craft Adversarial Data		Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure		Full ML Model Access								Erode ML Model Integrity
Active Scanning	Publish Poisoned Datasets										Cost Harvesting
	Poison Training Data										ML Intellectual Property Theft
	Establish Accounts										System Misuse for External Effect

Adversary Tactics and Techniques

ATT&CK Matrix for Enterprise

layout: side ▾

show sub-techniques

hide sub-techniques

Reconnaissance 10 techniques	Resource Development 7 techniques	Initial Access 9 techniques	Execution 13 techniques	Persistence 19 techniques	Privilege Escalation 13 techniques	Defense Evasion 42 techniques	Credential Access 17 techniques
Active Scanning (3)	Acquire Infrastructure (7)	Drive-by Compromise	Command and Scripting Interpreter (8)	Account Manipulation (5)	Abuse Elevation Control Mechanism (4)	Abuse Elevation Control Mechanism (4)	Adversary-in-the-Middle (3)
Gather Victim Host Information (4)	Compromise Accounts (3)	Exploit Public-Facing Application	Container Administration Command	BITS Jobs	Access Token Manipulation (5)	Access Token Manipulation (5)	Brute Force (4)
Gather Victim Identity Information (3)	Compromise Infrastructure (7)	External Remote Services	Deploy Container	Boot or Logon Autostart Execution (14)	Boot or Logon Autostart Execution (14)	BITS Jobs	Credentials from Password Stores (5)
Gather Victim Network Information (6)	Develop Capabilities (4)	Hardware Additions	Exploitation for Client Execution	Boot or Logon Initialization Scripts (5)	Boot or Logon Autostart Execution (14)	Build Image on Host	Exploitation for Credential Access
Gather Victim Org Information (4)	Establish Accounts (3)	Phishing (3)	Inter-Process Communication (3)	Browser Extensions	Boot or Logon Initialization Scripts (5)	Debugger Evasion	Forced Authentication
Phishing for Information (3)	Obtain Capabilities (6)	Replication Through	Native API	Compromise Client Software	Create or Modify System	Deobfuscate/Decode Files or Information	Forge Web
						Deploy Container	
						Direct Volume Access	

MITRE - примеры атак

[Home](#) > [Studies](#) > VirusTotal Poisoning

 KEY INFO

VirusTotal Poisoning

Incident Date: **2020** | Reporter: **McAfee Advanced Threat Research**

Actor: **Unknown** | Target: **VirusTotal**

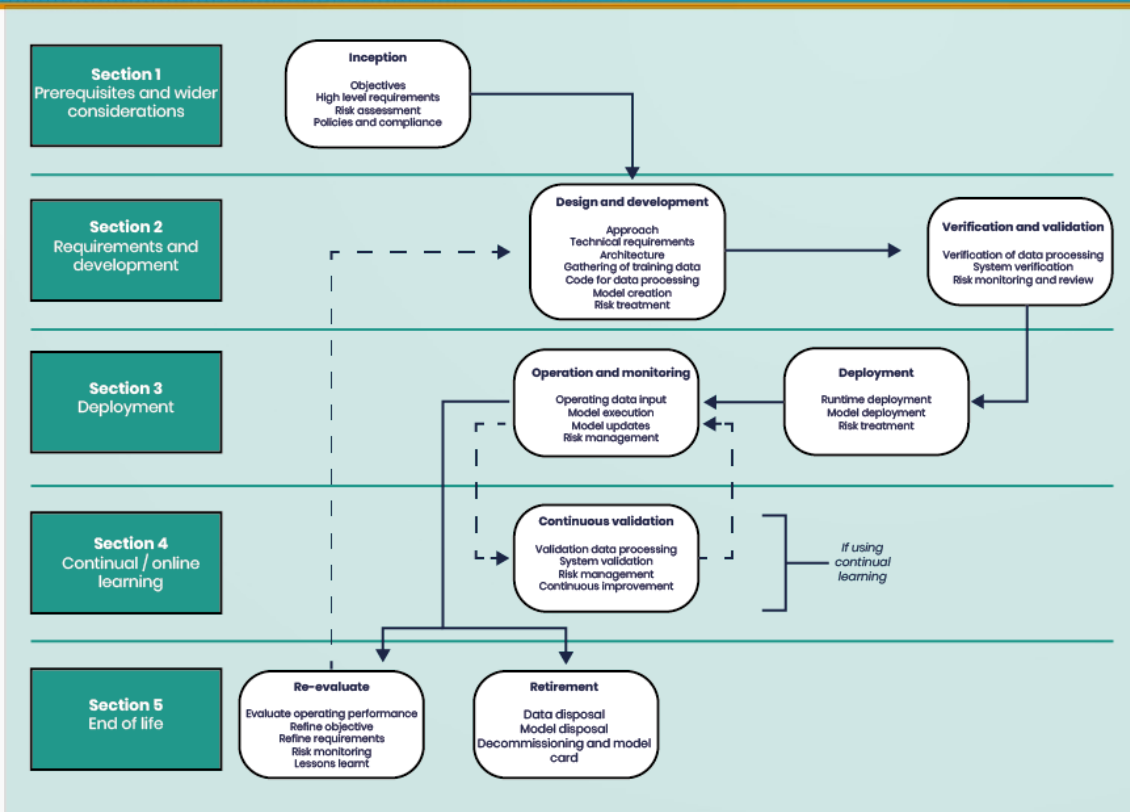
 DOWNLOAD DATA 

Summary

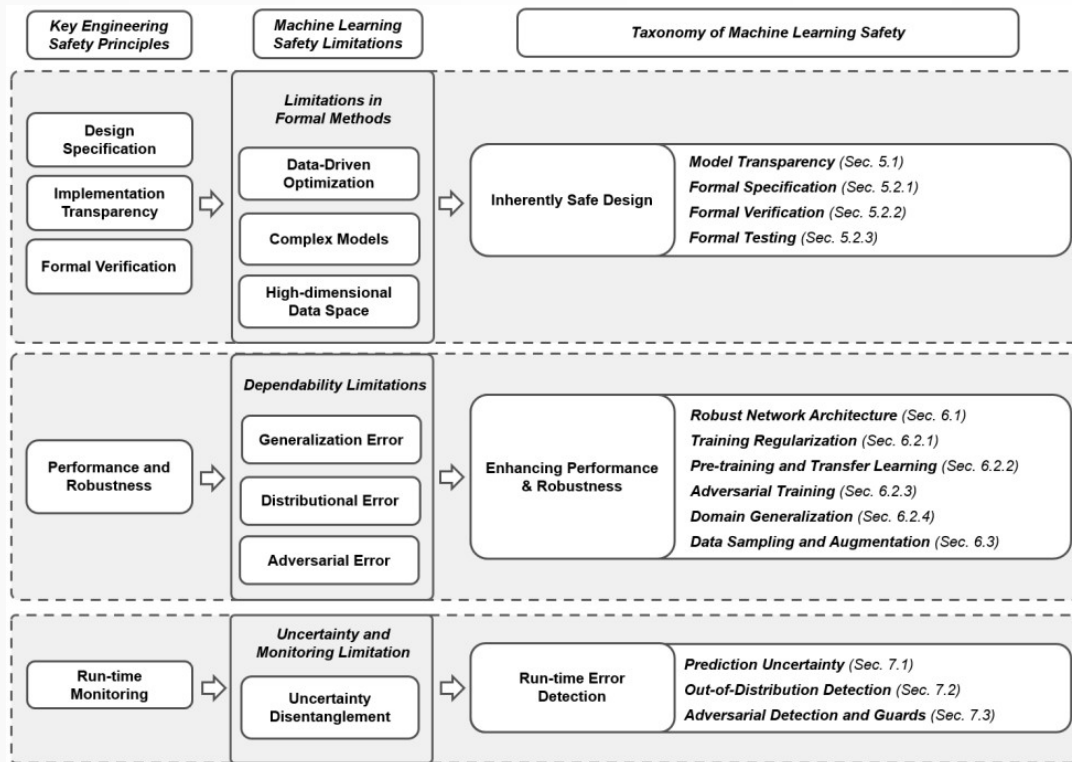
McAfee Advanced Threat Research noticed an increase in reports of a certain ransomware family that was out of the ordinary. Case investigation revealed that many samples of that particular ransomware family were submitted through a popular virus-sharing platform within a short amount of time. Further investigation revealed that based on string similarity the samples were all equivalent, and based on code similarity they were between 98 and 74 percent similar. Interestingly enough, the compile time was the same for all the samples. After more digging, researchers discovered that someone used 'metame' a metamorphic code manipulating tool to manipulate the original file towards mutant variants. The variants would not always be executable, but are still classified as the same ransomware family.

<https://atlas.mitre.org/studies/AML.CS0002>

Обеспечение безопасности — NCSC, UK

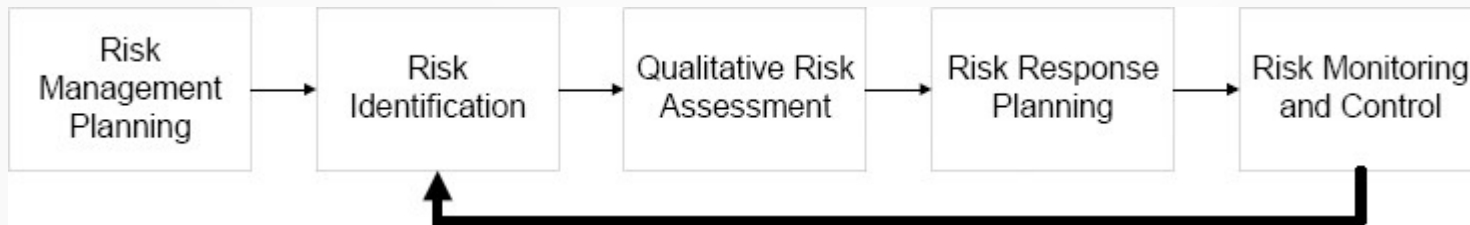


Таксономия ML Safety



Управление рисками — PMBoK

- Планирование управления рисками
- Идентификация рисков
- Качественные анализ рисков
- Количественный анализ рисков
- Планирование реагирования на риски
- Контроль рисков



Контроль рисков

- Допущение/принятие риска Accept
 - Решаем проигнорировать риск
(выгода велика, ущерб мал, нет ресурсов для других решений)
- Передача/разделение риска Transfer
 - Заплатим за то, чтобы рисковали не мы (страхование, аутсорсинг)
- Смягчение риска Mitigate
 - Минимизируем вред от риска
- Избегание риска Avoid
 - Уменьшаем вероятность реализации риска

Непредвзятость (Fairness)

- Предвзятость данных
- Этика
- Эндогенные переменные
- Fairness в современном понимании
- Инструменты и статьи

Предвзятость данных

- Статистические закономерности в данных, на которые мы не должны полагаться
- Из дискуссии в канале курса:
 - «законодательно запрещенный dataleak»
- В данных часто встречаются закономерности, на которые мы не должны полагаться (например, даталики)
- Некоторые из них еще и неэтичны — т. е. дискриминируют какую-то социальную группу

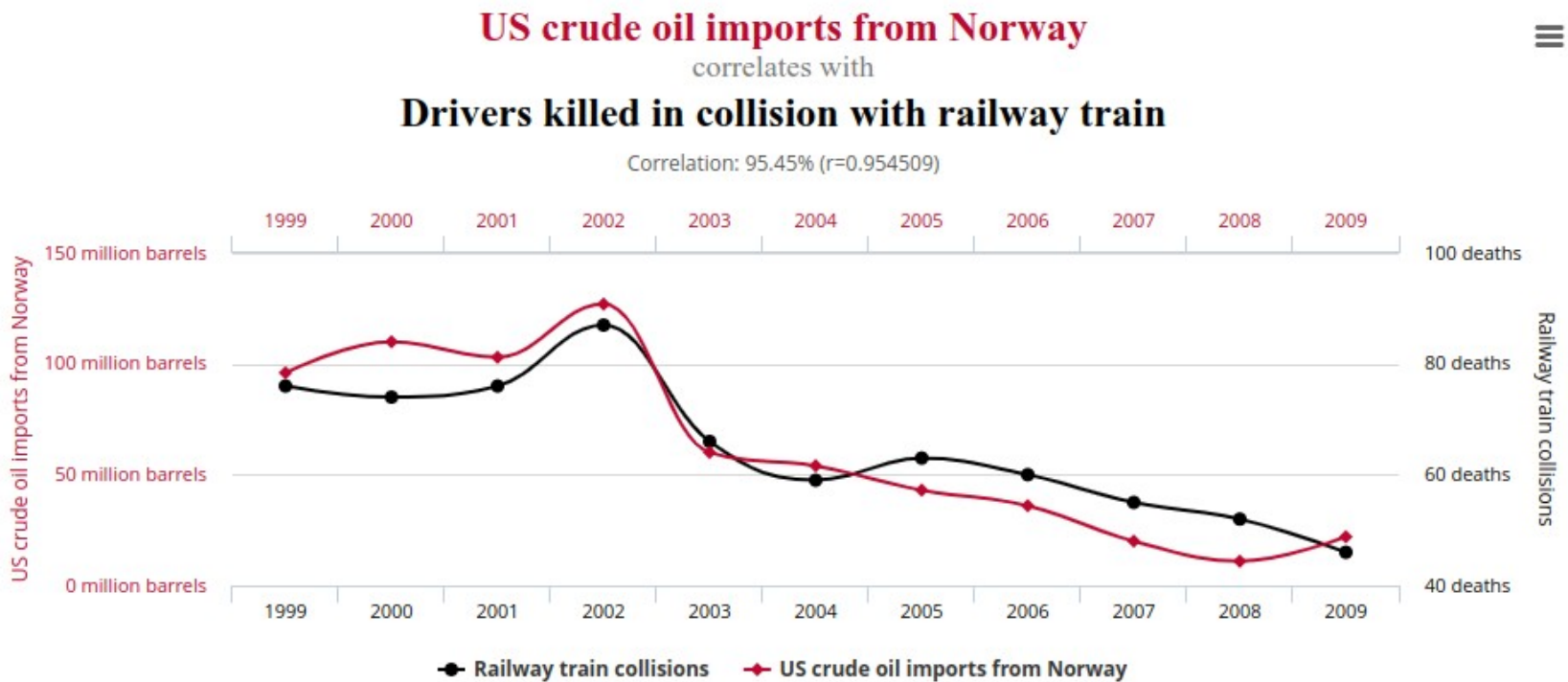
Неэтичное поведение у людей

- **Википедия:** неэтичное поведение можно кратко охарактеризовать как переход на личности, который создаёт конфликтную атмосферу и напряжённость.
- **Википедия:** первоначально смыслом слова это было совместное жилище и правила, порождённые совместным проживанием, нормы, сплачивающие общество, способствующие преодолению индивидуализма и агрессивности.

У компьютеров нет этики

- Компьютеры не совершают поступков, у них нет целей и воли
- Модель машинного обучения настолько же этична или неэтична, насколько этично или неэтично письмо или книга.
- Настоящая проблема машинного обучения — не в этике, а в плохой работе с псевдозаконмерностями в данных
- Неэтичные модели — ML-модели с псевдозаконмерностями, нарушающими сотрудничество людей
- Частный случай плохой работы с псевдозаконмерностями

Запретить импорт новрежской нефти



tylervigen.com

Как формулировали раньше

- Экзогенные переменные — не скоррелированные с ошибкой
- Эндогенные переменные — скоррелированные с ошибкой
- Причины эндогенности:
 - Пропущенные существенные переменные
 - Ошибки измерения регрессоров
 - Самоотбор
 - Одновременность
 - Серийная корреляция при наличии лагированной зависимой переменной среди регрессоров

Как формулируют сейчас

- Fairness in machine learning refers to the various attempts at correcting algorithmic bias in automated decision processes based on machine learning models.
- Decisions made by computers after a machine-learning process may be considered unfair if they were based on variables considered sensitive.
- Examples of these kinds of variable include gender, ethnicity, sexual orientation, disability and more. As it is the case with many ethical concepts, definitions of fairness and bias are always controversial.

[https://en.wikipedia.org/wiki/Fairness_\(machine_learning\)](https://en.wikipedia.org/wiki/Fairness_(machine_learning))

Итого:

- У общества есть болезненные темы, которые оно просит не трогать
- Если ваша модель будет полагаться на псевдозакономерности, связанные с этими темами, у вас будут проблемы с обществом
- Если закономерности хорошие — используйте на здоровье
- Расу, пол, возраст используют в медицинских моделях — это нормально.
- Плохо, если такие признаки увеличивают ошибку на подвыборке
- Fairness == the "true positive rate" is identical between groups
см [Equality of Opportunity in Supervised Learning](#)

Проверка ошибки на подвыборке

Sensitive feature

sex

Performance metric

Accuracy

Fairness metric

Demographic parity difference

	Accuracy	Selection rate	Demographic ...	False positive r...	False
Overall	85.4%	19.6%	18.3%	6.74%	39.5%
Male	81.6%	25.7%		9.75%	38%
Female	93.1%	7.36%		2.01%	48%

Инструменты и статьи

- Responsible AI Toolbox (The Fairness dashboard) →
- What-if Tool →
- Fairlearn → →
- ML-fairness-gym → →
- Attacking discrimination with smarter machine learning →
- The Trouble with Bias - NIPS 2017 Keynote →
- The Cost of Fairness in AI: Evidence from E-Commerce →
 - Ищите экзогенные переменные или платите налог на Fairness

Дополнительные материалы

- Model Cards for Model Reporting →
- Taxonomy of Machine Learning Safety →

Все будет в телеграм-канале