هوش مصنوعی و عدالت اجتماعی

شاهین روحانی
دانشکده فیزیک
دانشگاه صنعتی شریف
1401

مشکل

سیستم‌های هوش مصنوعی می‌توانند بین کلاس‌های جامعه تبعیض قائل شوند –

سوگیری، یک مورد خاص از مشکل کلی بی عدالتی اجتماعی است.

مشکل - به طور دقیق تر، در یک جنبه از آن: استفاده از متغیرهای پراکسی است.

یک متغیر پراکسی یک ورودی است که به راحتی اندازه گیری می شود

متغیر پراکسی به جای یک متغیر ورودی مورد نظر که یا قابل مشاهده نیست، یا شاید بسیار پرهزینه برای اندازه گیری است

استفاده می شود

SITN
science in the news
celebrating 20 years

HARVARD UNIVERSITY
The Graduate School of Arts and Sciences

Alex Najibi, OCTOBER 24, 2020

- تبعیض نژادی در فناوری تشخیص چهره
- فیس بوک چهره سفید پوستان را بهتر تشخیص می دهد

مقدمه

- هوش مصنوعی چیست

هوش مصنوعی یک چیز نیست

هوش مصنوعی به عنوان ترکیبی از موارد زیر تجربه می شود:

یادگیری ماشین (= تشخیص الگو)

داده ها

حسگرها

الگوریتم ها

زیرساخت ها

یادگیری ماشین

- هوش مصنوعی (AI)هوشی است که توسط ماشین‌ها نشان داده می‌شود، برخلاف هوش طبیعی نشان داده شده توسط حیوانات از جمله انسان.
- یادگیری ماشینی یک حوزه تحقیق هوش مصنوعی است که به درک و ساخت روش‌هایی اختصاص داده شده است که «یاد می‌گیرند»، یعنی روش‌هایی که داده‌ها را برای بهبود عملکرد در مجموعه‌ای از وظایف به کار می‌گیرند.

داده ها

- دسته بندی داده ها چیست؟ چرا انتخاب شدند؟
- چه چیزی در داده ها وجود ندارد؟ چیزی که از آن کم است چیست
- برچسب ها؟

حسگرها

- آنها در واقع چه چیزی را اندازه گیری می کنند؟ چه فرضیاتی دارند
- وابسته به چه هستند؟

الگوریتم ها

- چه کسی الگوریتم را ایجاد کرد؟ قصدشان چه بود؟
- چگونه دستورالعمل ها در طول زمان، توسط چه کسی، در پاسخ به چه؟

زیرساخت ها

- چه منابع انسانی و سخت افزاری برای ساختن اجرا و نگهداری سیستم مورد نیاز است،

اولین مورد مطالعه مراقبت های بهداشتی

آربیتراژ

- مشکل: بیمارستان ها و سیستم های مراقبت های بهداشتی بیش از حد تحت فشار هستند.
- آنها می خواهند منابع خود را به سمت افرادی که بیشتر نیاز دارند هدایت کنند

آربیتراژ

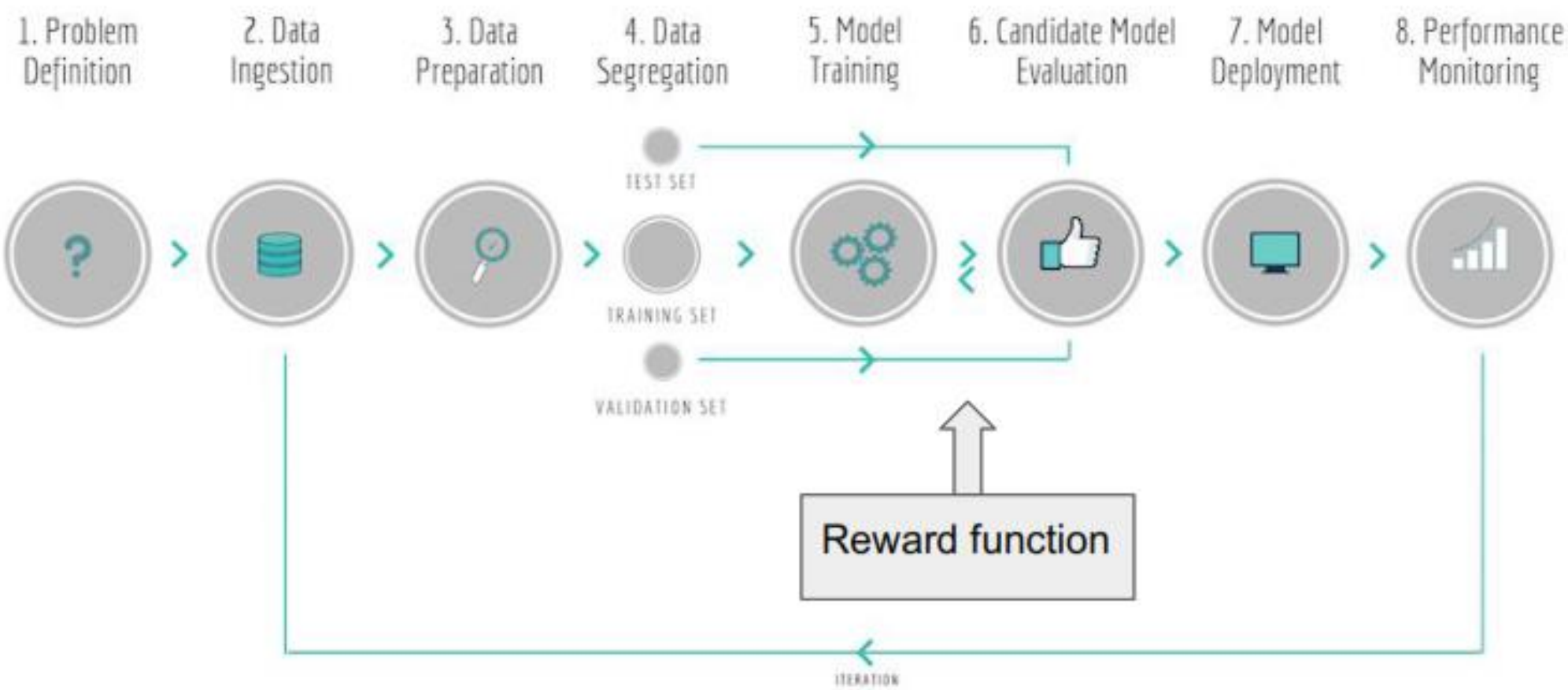
- چالش: چگونه می‌توانیم تشخیص دهیم که چه افرادی واقعاً بیمار هستند

و بنابراین چه کسی بهتر است هدف قرار گیرد

آربیتراژ

- میتواند یک چالش خوب برای یک سیستم هوش مصنوعی باشد
- بیایید ازفراگیری ماشین ML استفاده کنیم!

Using machine learning



متغیرهای پراکسی

- چه چیزی را اندازه گیری کنیم؟
- ما نمی توانیم سلامتی را اندازه گیری کنیم
  - فشار خون
  - دمای بدن
  - قند خون
  - ...

داده

- دسته بندی داده ها چیست؟ چرا انتخاب شدند؟
- دسته بندی اولیه : هزینه های بهداشت و درمان به دلیل در دسترس بودن انتخاب شده است

الگوریتم ها

- چه کسی یک الگوریتم ایجاد کرده است؟ قصدشان چه بود؟
- ایجاد شده به عنوان یک محصول تجاری،
- در نظر گرفته شده برای پیش بینی بیماران در معرض بیشترین خطر
- بیماری مزمن
- دستورالعمل ها در طول زمان، توسط چه کسی، تنظیم شده است
- پاسخ به چی ؟
- توسط موسسات و بیمارستان های متعدد، استفاده می شود

کار با هوش مصنوعی

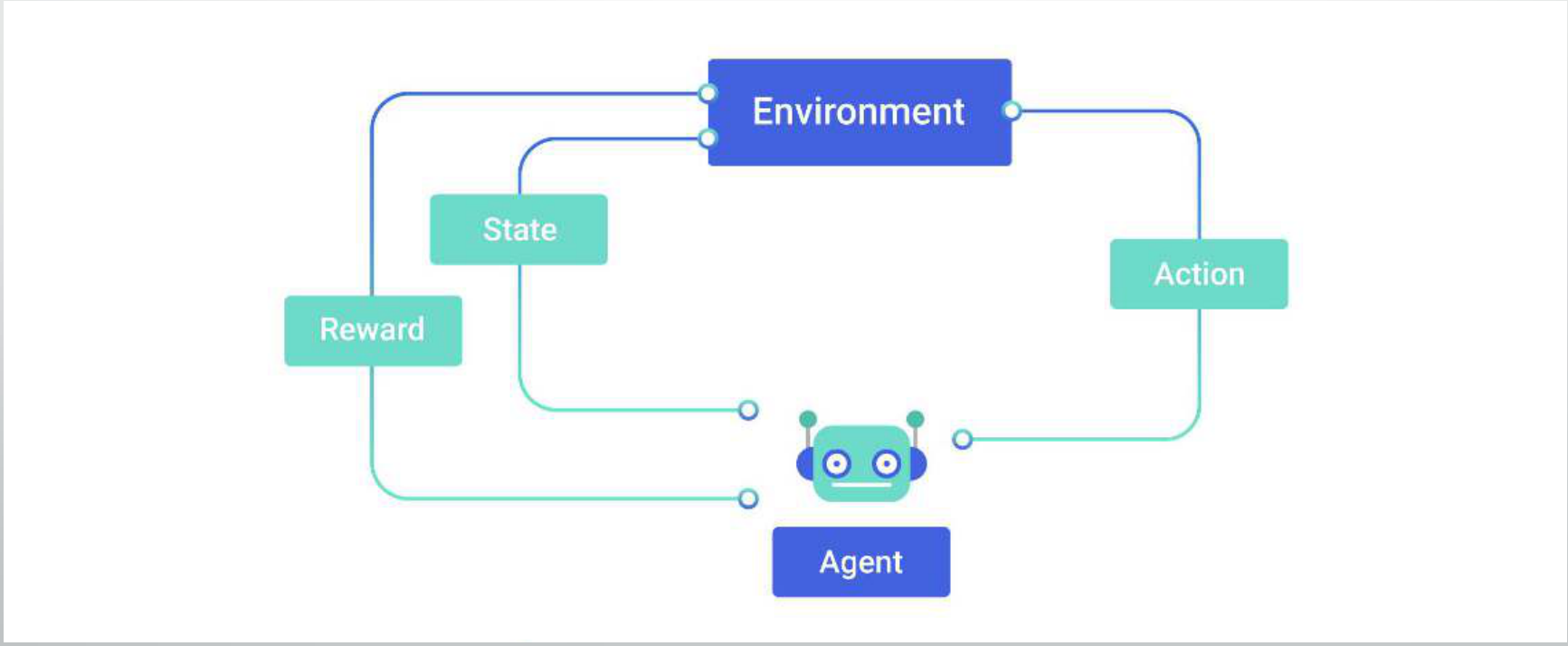
- طراحی انسان محور
- MLبرای کاربران چیست؟
  - شخصی سازی (مثلاً سفارشی کردن درخواست کمک مالی در یک فرد بر اساس فرد)
  - پیش‌بینی (مثلاً پیش‌بینی عملکرد ایمیل)
  - پردازش زبان طبیعی (مانند تبدیل تماس تلفنی پیام به متن)

چه چیز MLخوب نیست

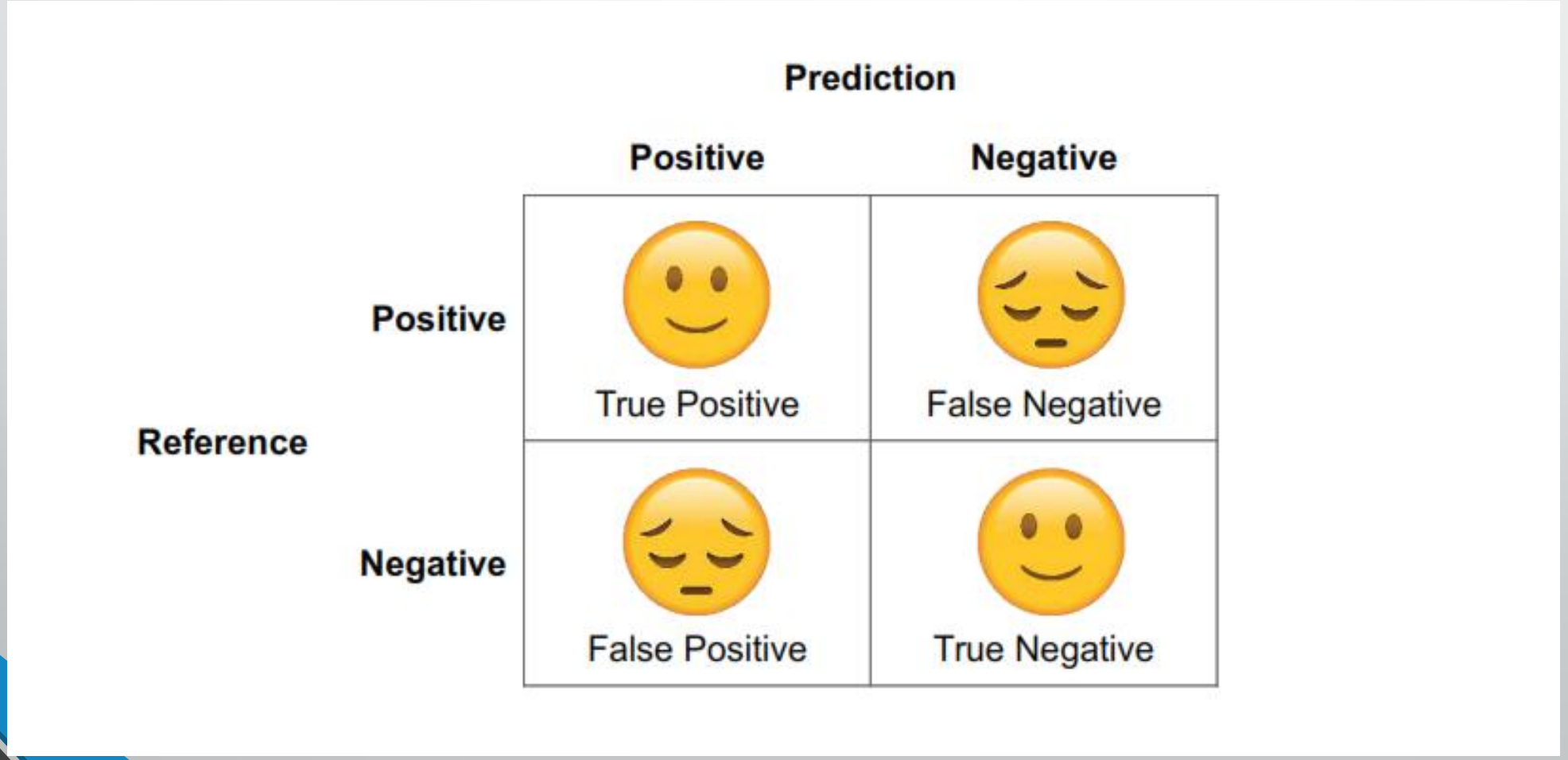
- پیش بینی پذیری (به عنوان مثال "پیش بینی فازی")
- اطلاعات ثابت یا محدود
- موقعیت‌هایی که هزینه خطا بالاست (مثلاً ایمیل های درخواست پرداخت از یک اداره دولتی، ارسال شده به افراد)

کار با ه.م.

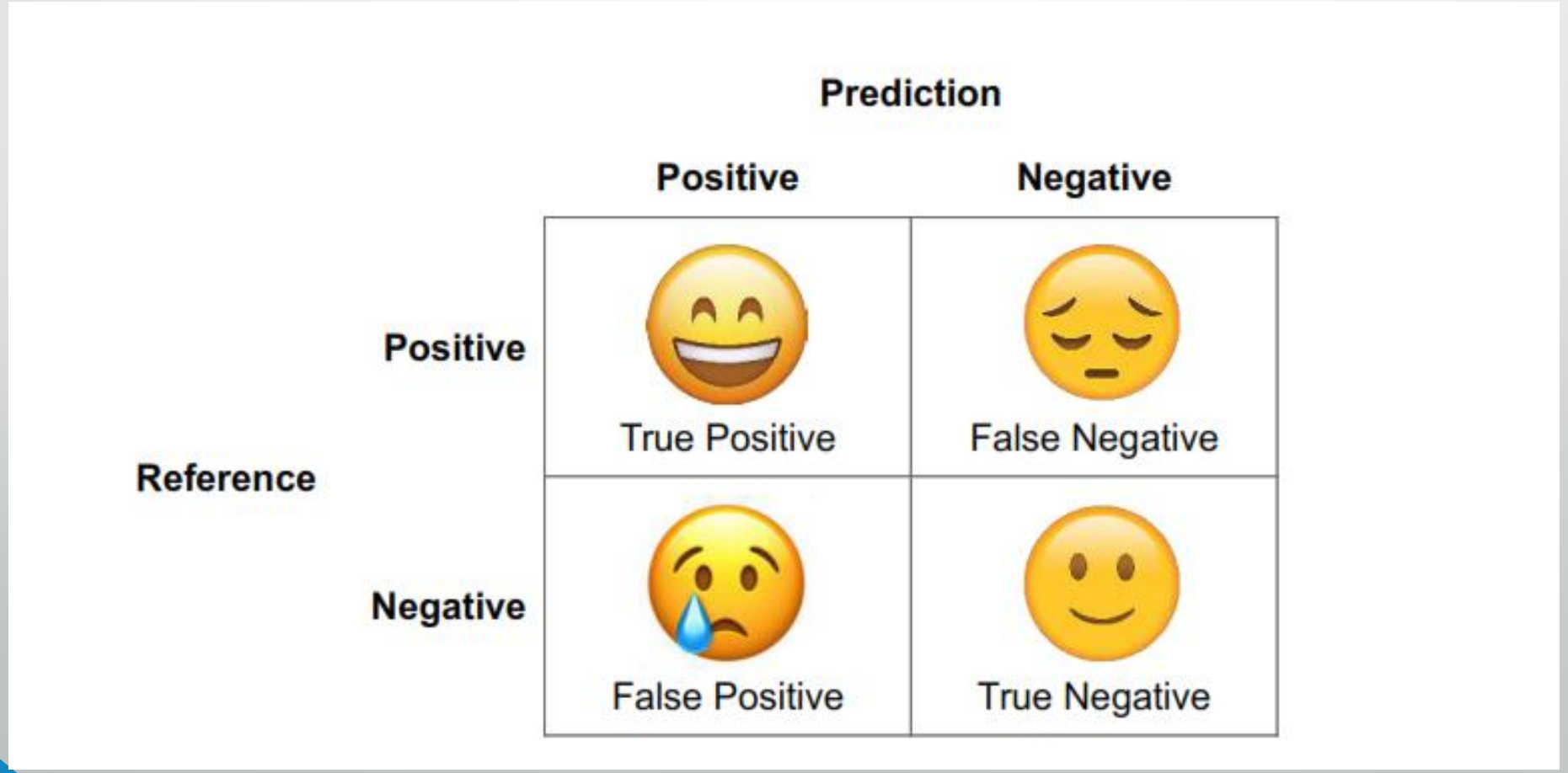
- تابع پاداش، تابع هدف، عملکرد ضرر



هوش مصنوعی عملا چیزی به جز طبقه بندی کننده نیست
نمونه طبقه بندی کننده باینری



سیستم مثال: نتایج متفاوت، هزینه های متفاوت



توصیفی یا هنجاری

Descriptive	**Normative**
Accurate when compared to human behaviour	Inaccurate when compared to human behaviour
Potentially racist/sexist/etc (like human behaviour)	*Potentially* neutral (unlike human behaviour)
Almost all ML systems are descriptive, but we often expect them to behave in normative ways.	

مثال دوم:
لیست ایمیل

1. ما یک لیست ایمیل را مدیریت می کنیم
2. مشکلی که ما با آن روبرو هستیم این است که پیش بینی اینکه کدام مسائل برای مشترکین جالب است
3. کار فعلی ما این است که ما ایمیل ها را با زیر مجموعه های کوچکی از لیست خود آزمایش می کنیم و اگر عملکرد خوبی داشته باشند،
4. سپس آنها را به لیست بزرگتر خود ارسال می کنیم

لیست ایمیل

- اما با این وجود، این بدان معنی است که بسیاری از ایمیل های ما فقط توسط 20٪ از اعضای ما باز می شوند، و اقدام تنها توسط 6٪ از اعضای ما انجام می شود.
- ما یک برنامه هوش مصنوعی می نویسیم تا مشکل را برطرف کنیم

دقت

- - نسبت مثبت واقعی از بین همه درست ها چقدر باشد ؟
- آیا ما می خواهیم به طور مطلق درست باشیم ؟
- آیا هزینه ها دقت را پرداخت می کنیم؟

فرا خواندن

- نسبت مثبت واقعی به درستی طبقه بندی شده از بین تمام مثبت های واقعی و منفی های کاذب
- یا می خواهیم بسیاری از گزینه های ممکن را داشته باشیم ؟

- Racial bias in healthcare algorithms: https://science.sciencemag.org/content/366/6464/447.full
- PAIR guidebook: https://pair.withgoogle.com/
- 3A Institute: https://3ainstitute.cecs.anu.edu.au/
- ML -- what's the process for using ML: https://towardsdatascience.com/not-yet-another-article-on-machine-learning-e67f8812ba86
- ML -- what is ML really: https://medium.com/hackernoon/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c
- Descriptive vs normative -- breaking down bias in ML: https://podtail.com/en/podcast/ai-australia/bias-in-machine-learning-systems-with-katherine-ba/
- Algorithms as administrative mechanisms: https://journals.sagepub.com/doi/full/10.1177/2053951718757253