

# توضیح پذیری بالفعل در هوش مصنوعی

سیاوش بیگدلی

علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی


۴۵٪ از سیاه‌پوست‌های تشخیص  
داده شده جرم دوباره انجام ندادند



۴۸٪ سفیدپوست‌های تشخیص  
داده شده نشد جرم دوباره انجام دادند



COMPAS: Correctional Offender Management Profiling for Alternative Sanctioning



۲۰٪ از سیاه‌پوست‌های تشخیص  
داده شده جرم دوباره انجام دادند

**تصمیم‌گیری بر مبنای داده‌های تاریخی و بدون دخالت قاضی  
چشم پوشی بر امکان تغییر در فرد، جامعه و قضاوت**

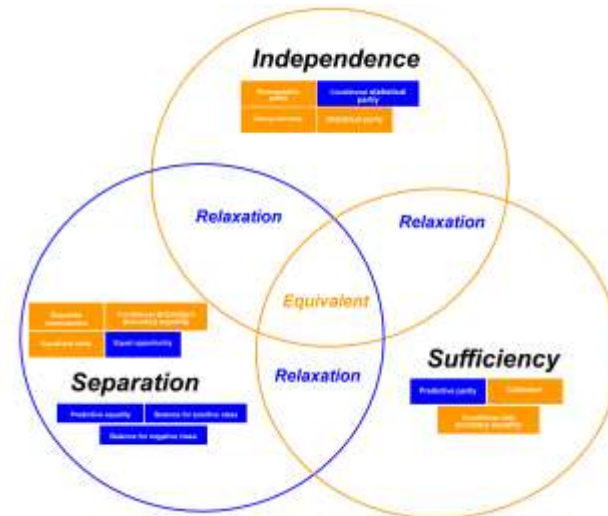
۸۱٪ سفیدپوست‌های تشخیص  
داده نشده جرم دوباره انجام دادند

# سیستم تشخیص افسردگی در داده‌های دانمارک

[Olalekan Akintande, Eike Peterson, Aasa Feragen]

Age group	Region (of Denmark),
Kommunegruppe (as configured in Denmark)	Number of address changes
Civil status (married, not married, divorced, widowed)	Sex (Male, Female and Non-binary)
Household income quartile	Living status

- بر مبنای وضعیت اجتماعی
- خطای تشخیص پزشک در مردها
- عدم دسترسی به برچسب حقیقی (ground truth)
- غیر قابل تعمیر با ابزارهای انصاف/الگوریتمی (algorithmic fairness)



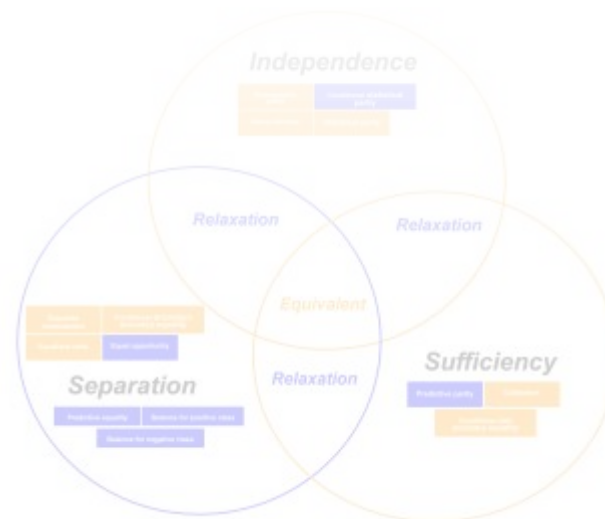
# سیستم تشخیص افسردگی بر داده‌های دانمارک

- بر مبنای وضعیت اجتماعی
- خطای تشخیص در مردها و داده‌ها

• عدم دسترسی به برجسب حقیقی (ground truth)

## تصمیم‌گیری بر مبنای داده‌های تاریخی و بدون دخالت دکتر چشم پوشی بر امکان تغییر در تشخیص

Was registered in Denmark)	Life changes
Civil status (married, not married, divorced, windowed)	Sex (Male, Female and Non-binary)
Household income quartile	Living status



If the physician should fulfill his or her moral obligation to promote the health and well-being of the patient, then the use of AI should be designed in such a way that it promotes or at least maintains the physician's decision-making autonomy.

Funer and Weising, *Physician's autonomy in the face of AI support: walking the ethical tightrope*, Frontiers in Medicine 2024



# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

همیشه یک آرایشگاه خوب در تهران معرفی کنی؟

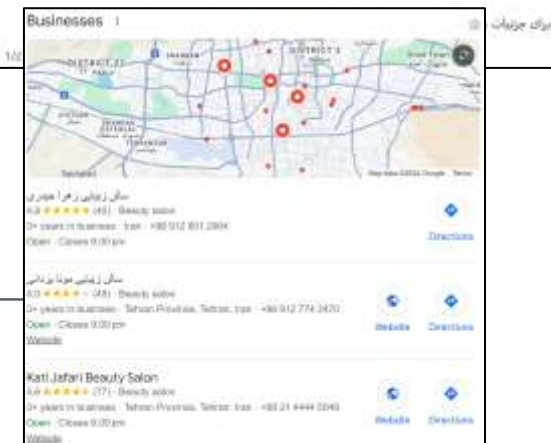
Searched 5 sites

در تهران، چندین سالن زیبایی با کیفیت بالا وجود دارند که خدمات حرفه‌ای ارائه می‌دهند. در ادامه چند مورد از بهترین‌ها را معرفی می‌کنم:

۱. سالن زیبایی لی لی رضایی: این سالن در سعادت آباد واقع شده و به دلیل خدمات حرفه‌ای از جمله میکاپ عروس، رنگ و سبزه، و مراقبت‌های پوستی شناخته شده است. آدرس: خیابان بهشت شرقی، ساختمان دینا. تلفن: پنجم شماره تماس: 02188565070

۲. سالن زیبایی مریم رانوه: یکی دیگر از سالن‌های برتر درعفرانیه است که در زمینه گریم و آرایش عروس، میکروپیگمنتیشن و خدمات مو تخصص دارد. آدرس: عفرانیه، خیابان مقدس اردبیلی، بن‌بست هفتم، شماره تماس: 02122042737

۳. سالن یووال صدف: در شهرک غرب واقع شده و خدمات گسترده‌ای از جمله میکاپ حرفه‌ای، خدمات VIP عروس، و خدمات ناخن ارائه می‌دهد. این سالن به دلیل استفاده از برندهای آرایشی معتبر و کیفیت بالای خدمات بسیار محبوب است.



شناخت فرایند آستنتاج

بازبینی و طراحی ابزار بر اساس زیربنای موجود، مثلاً تمرکز زدایی

مرکزگرا

تفویض شناخت

تغییر در پایگاه داده یا شخصی سازی مدل

کنترل فردی یا تعاونی

**خانم ۱:** بچه‌ها، یه آرایشگاه خوب برای رنگ مو و لایت می‌خوام، ولی باید طبیعی باشه و به موهام آسیب نزنه. کجا رو پیشنهاد می‌دید؟

**خانم ۲:** من "صدف" می‌رم. موادشون خیلی باکیفیت و اصل هست، موهام اصلاً خراب نشد.

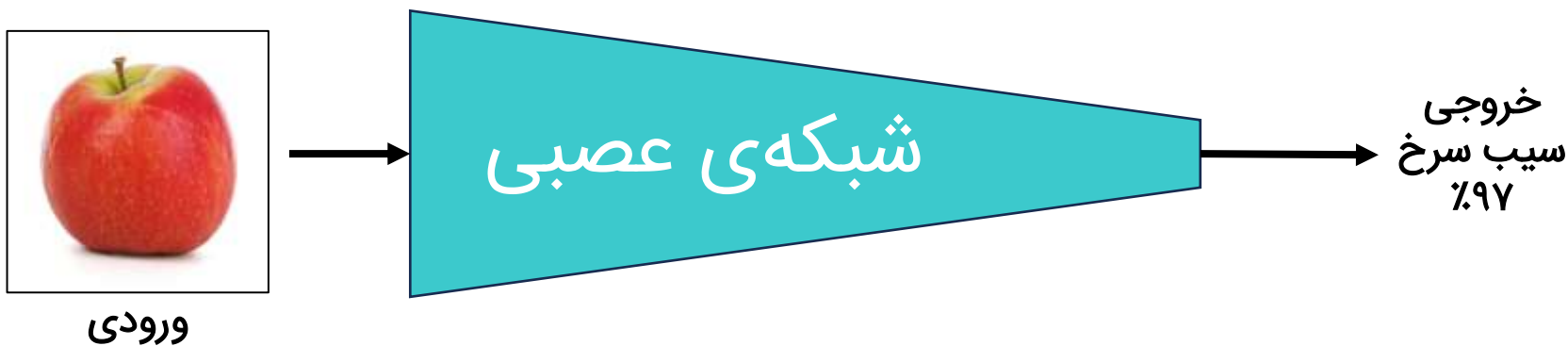
**خانم ۳:** به نظرم "مروارید" بهتره. اونجا قبل از شروع، مشاوره می‌دن و با توجه به نوع موها کار می‌کنن، برای همین خیالت راحت‌تره که آسیبی نمی‌رسه.

**خانم ۱:** خب چرا فکر می‌کنی بهتر از "صدف"ه؟

**خانم ۳:** چون من یه بار "صدف" رفتم، نتیجه خوب بود، ولی حس کردم به سلامت موهام کمتر توجه شد.

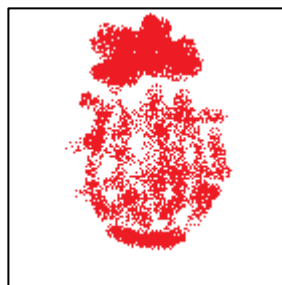
**خانم ۲:** خب بستگی داره. من تجربه خوبی ازش داشتم. انتخاب با خودته!

# هوش مصنوعی توضیح پذیر (eXplainable AI)



چه ویژگی‌های تصویر ورودی در تصمیم مدل تاثیر دارند؟

توضیح: نقشه‌ی اهمیت (saliency map)  
نشان دادن قسمت‌هایی که در تصمیم اثر داشتند و چقدر





# هوش مصنوعی توضیح پذیر (eXplainable AI)

[Nina Wang, Paraskevas Pegios]

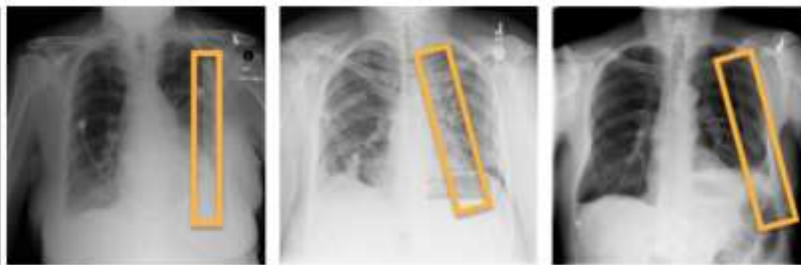
- ورودی‌های خلافِ واقع (counterfactual)
- اگر قسمتی را از ورودی حذف کنم تصمیم مدل عوض میشود؟

(a) CheXpert: Chest X-rays with 'pacemaker'.

(b) NIH: Chest X-rays with 'chest drain'.

(c) ISIC 2018: Skin lesions with 'ruler markers'.

Original

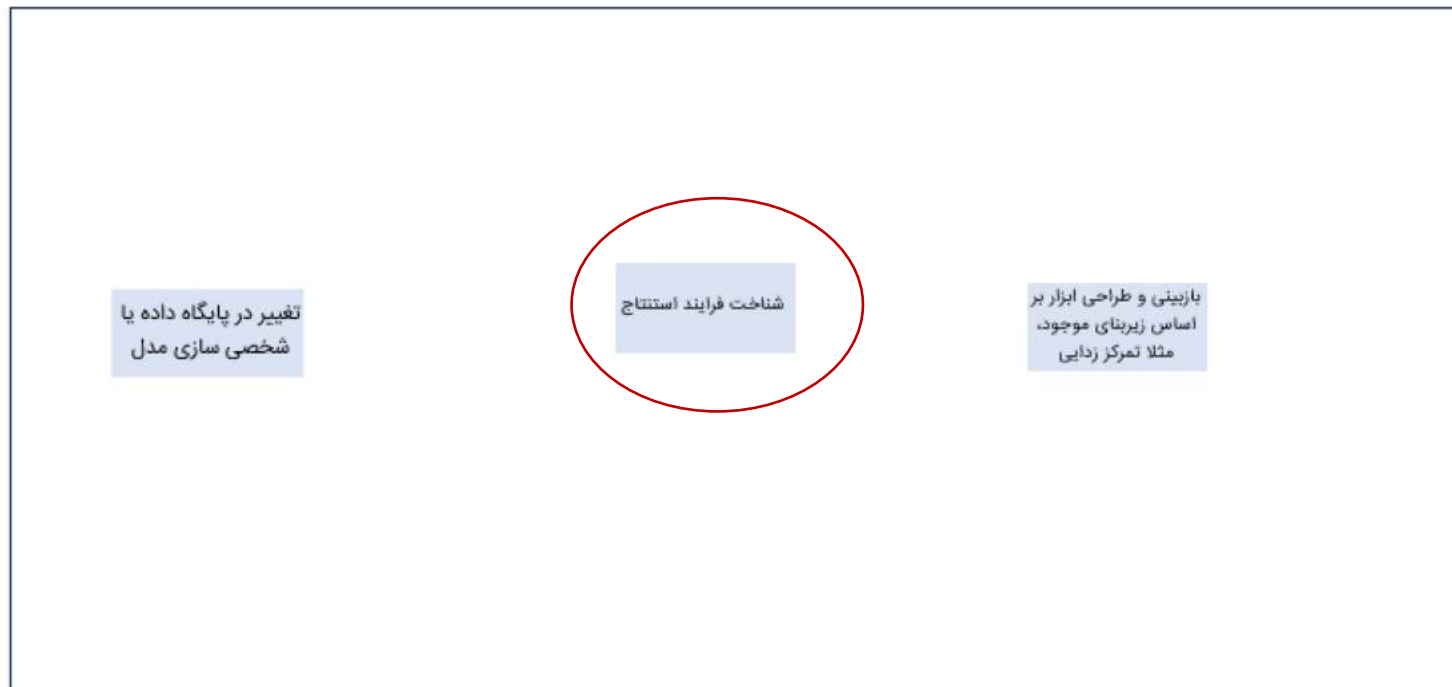


FastDiME-2+

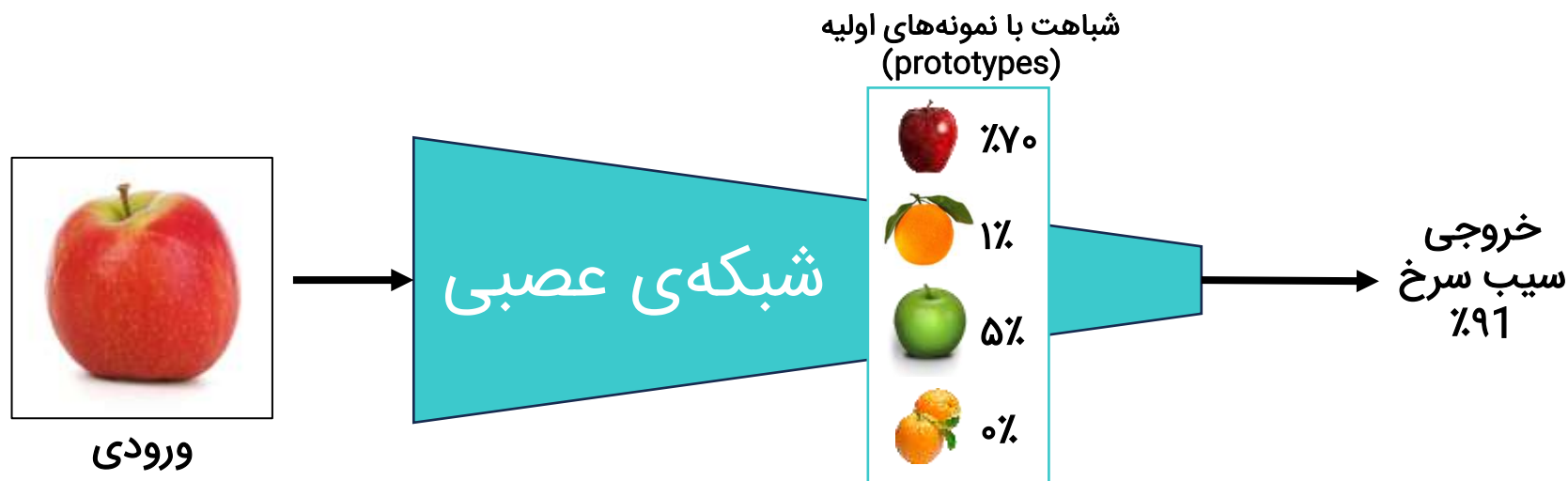


# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

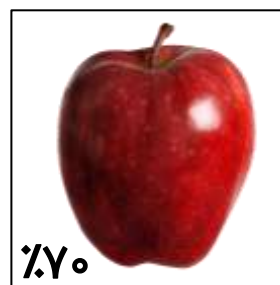


# هوش مصنوعی توضیح پذیر (eXplainable AI)



مدل بر چه اساسی این تصمیم را گرفته؟

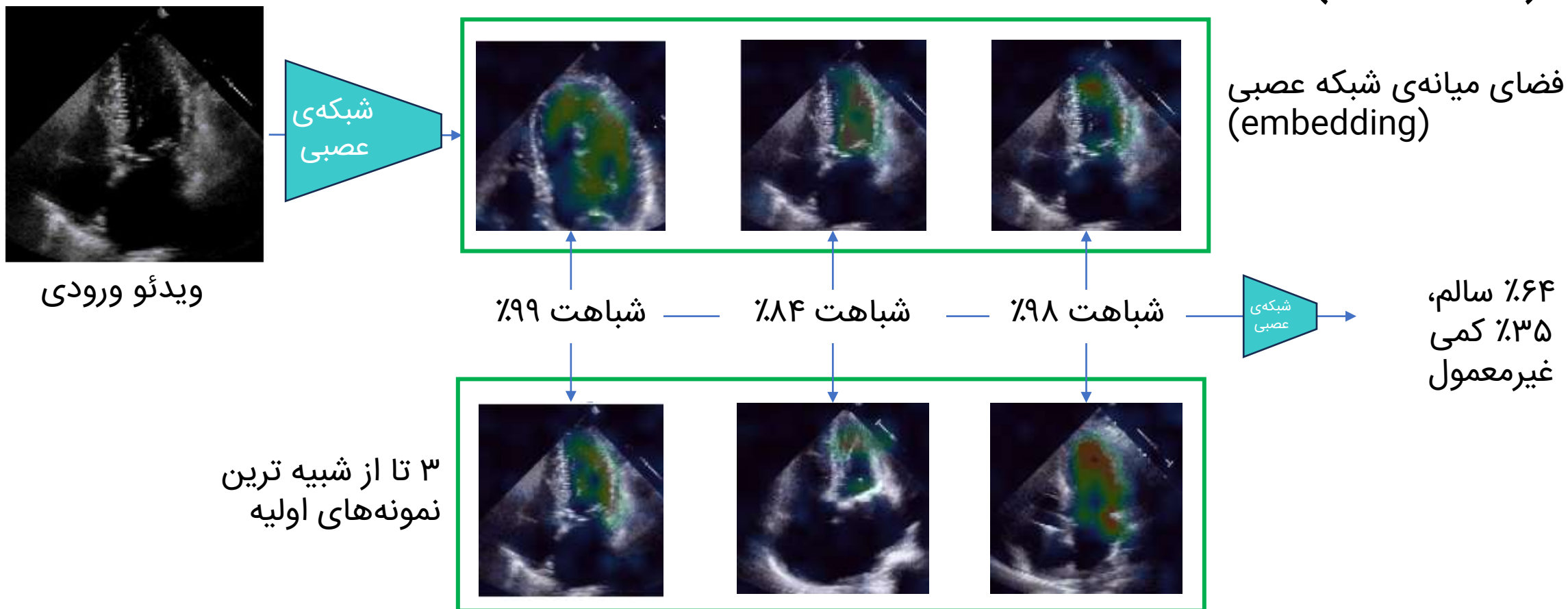
توضیح: شبهت با نمونه‌های اولیه (prototype)  
نشان دادن نمونه‌ها که در تصمیم اثر داشتند و مقدار شبهت



# هوش مصنوعی توضیح پذیر (eXplainable AI)

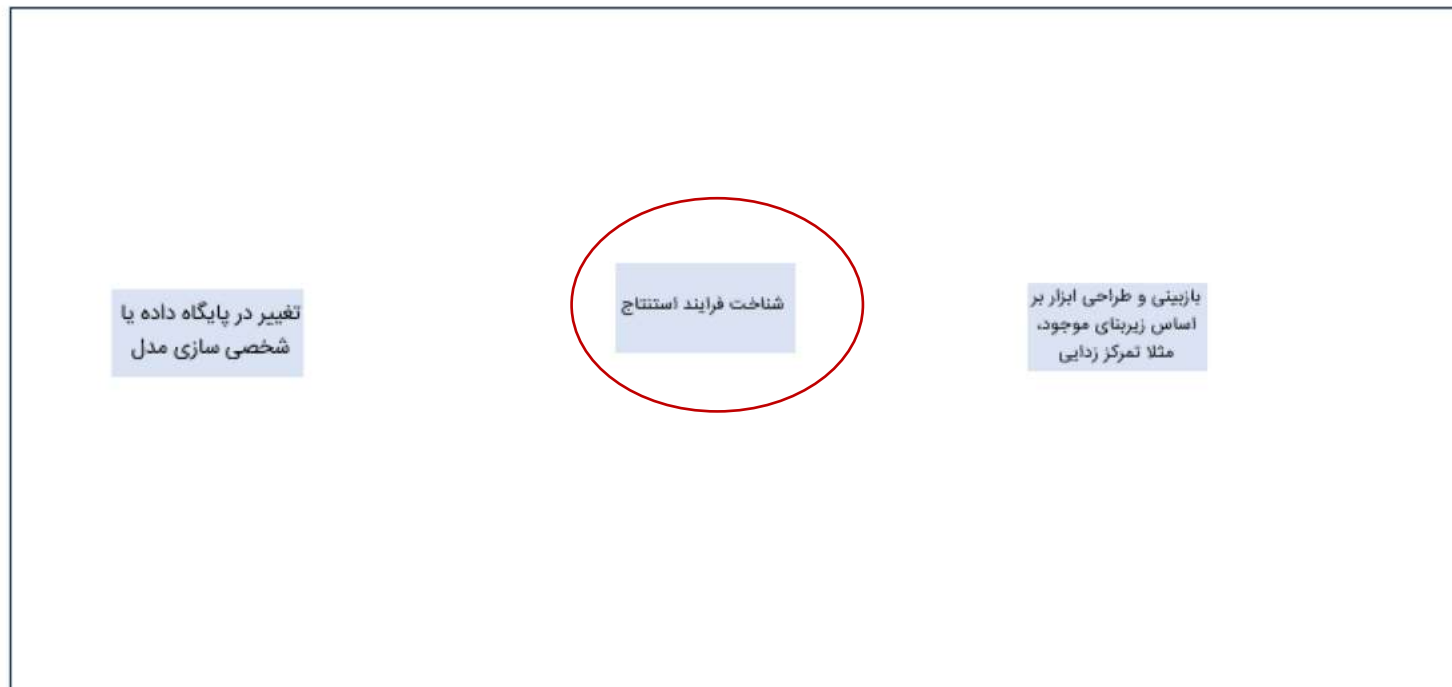
[Yeganeh Ghamari, Nina Wang]

- نمایش شباهت با نمونه‌ی اولیه و موقعیت آن در اکوگرافی قلب (ECG/EV)

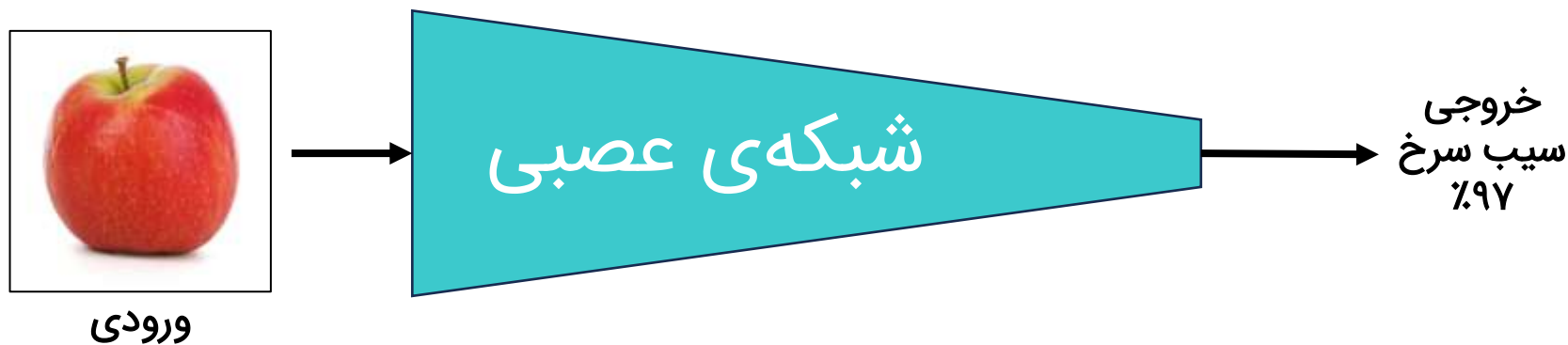


# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]



# هوش مصنوعی توضیح پذیر (eXplainable AI)



کدام داده‌های یادگیری در تصمیم مدل تاثیر دارند؟

توضیح: تاثیر نمونه‌ها (sample influence)  
نشان دادن سهم تاثیر هر نمونه در تصمیم مدل



...

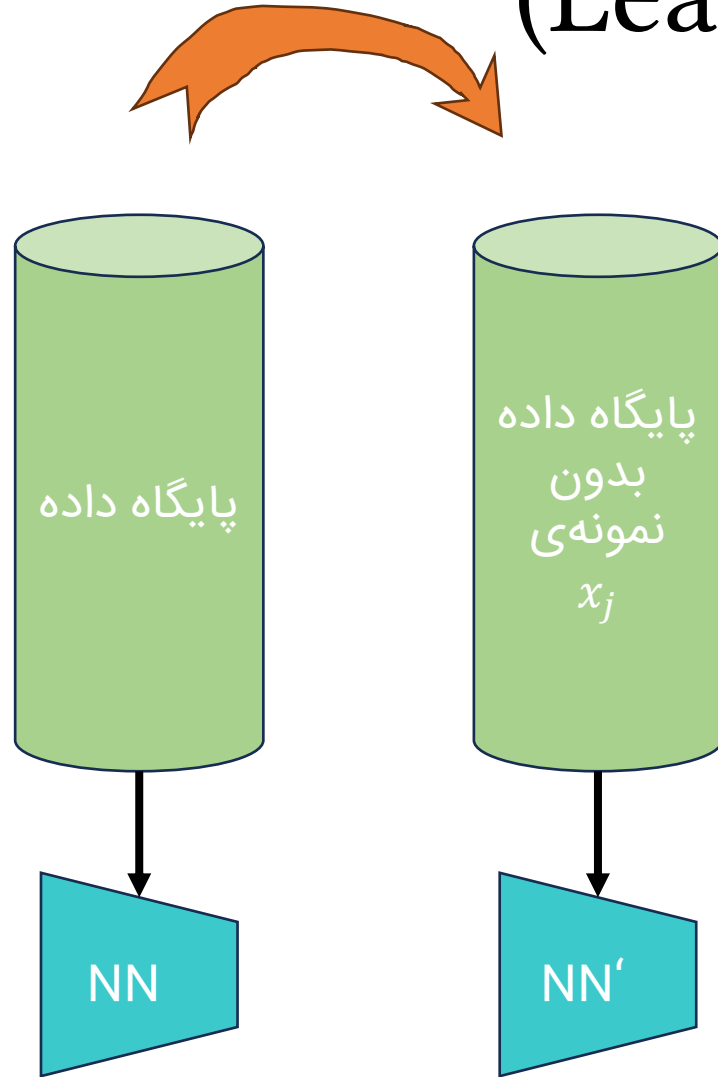


# کاربردها

- سهم مالی از مشارکت در تولید محتوی خلاقانه
- هنرمندان، نویسندگان، دلال‌های داده (از دهه‌ی ۹۰ میلادی)
  - [How Generative AI Can Impact Artists' Income: The Search for Fair Compensation | Robots.net](#)
  - [AI and Artists' IP: Unpacking Copyright Infringement Allegations in Andersen v. Stability AI Ltd. – Center for Art Law \(itsartlaw.org\)](#)
  - [How much can artists make from generative AI? Vendors won't say | TechCrunch](#)
- شرکت‌های تولید و عرضه محتوی Adobe, Getty images, Stability AI
- استناد تجربی (قضایی، پزشکی)
- تشخیص حمله‌های روی پایگاه داده
- جهت گیری و خطا در برچسب‌های پایگاه داده
- تغییر پایگاه داده [Datamodels: Predicting Predictions from Training Data](#)
  - کوچک‌سازی
  - تغییر کانتکست مدل
  - پایگاه داده خلاف واقع



# طرد نمونه (Leave-One-Out)



- نمونه‌ی مورد نظر را در بیاوریم و مدل جدیدی بسازیم

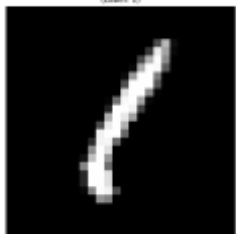
$$SI(x_0, x_j) = f_D(x_0) - f_{D-\{j\}}(x_0) \cdot$$

- نیاز به باز یادگیری مدل برای هر نمونه دارد
- هزینه‌ی بسیار بالا و غیر عملی (بجز روش‌های تخمینی)
- برداشتن یک نمونه در عمل تاثیر چندانی در مدل ندارد

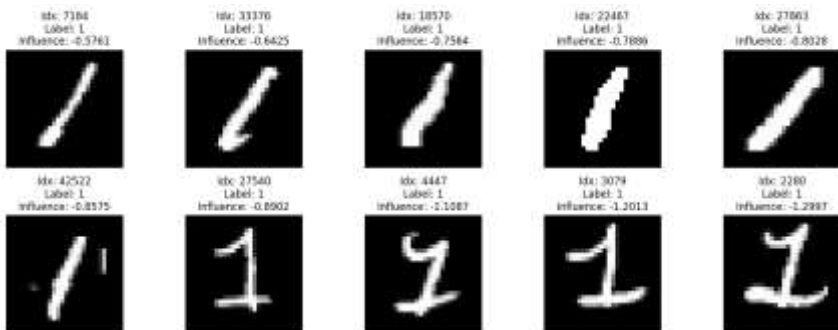
# هوش مصنوعی توضیح پذیر (eXplainable AI)

[Frida Jørgensen, Nina Wang]

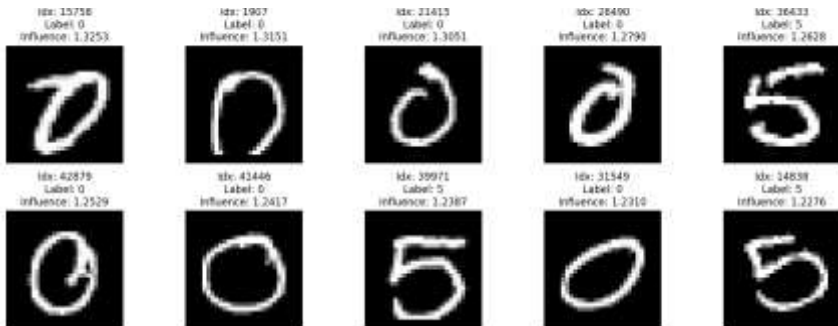
نمونه ورودی



نمونه یادگیری با نسبت مثبت



نمونه یادگیری با نسبت منفی



• تابع تاثیر (influence function)

• A : تغییر پارامترهای مدل نسبت به بهینه‌گی در نمونه یادگیری؛ مشتق درجه اول

• B : تغییر بهینه‌گی در نمونه ورودی نسبت به پارامترهای مدل؛ مشتق درجه اول

• C : تابندگی پارامترها؛ مشتق درجه دوم (Hessian)

• نسبت تغییر بهینه‌گی در داده ورودی به بهینه‌گی در داده یادگیری (IF = ACB)

# هوش مصنوعی توضیح پذیر (eXplainable AI)

[Frida Jørgensen, Nina Wang]

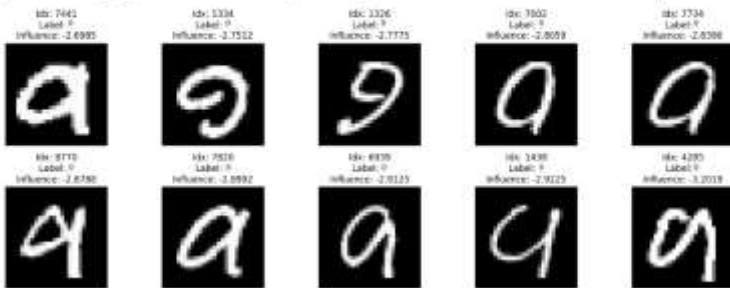
- تابع تاثیر (influence function) برای کشف داده های مضر و بد برچسب گذاری شده
- چه نمونه هایی در پایگاه داده ی یادگیری باعث خطا در تصمیم گیری عموم نمونه های تست میشوند؟



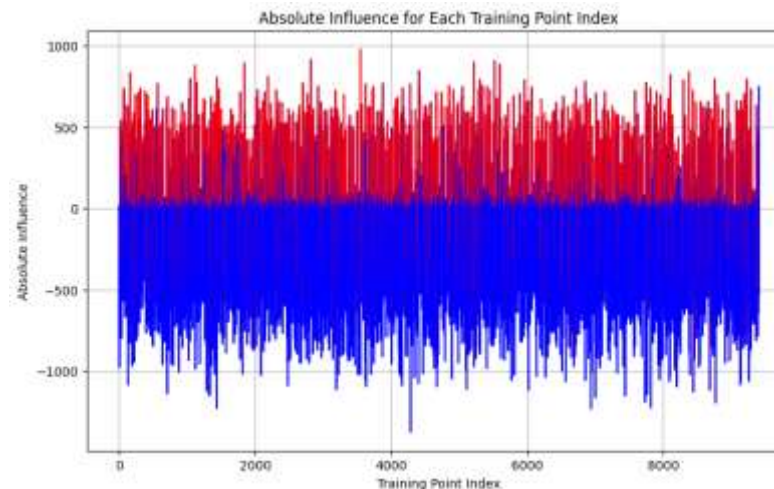
(a) Example of misclassified test point with true and predicted Label



(b) Training points with highest positive influence scores (Harmful)



(c) Training points with highest negative influence scores (Helpful)



جمع تمامی تاثیرها برای هر داده ی یادگیری

داده ی با برچسب غلط  
داده ی با برچسب درست

# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

**خانم ۱:** بچه‌ها، یه آرایشگاه خوب برای رنگ مو و لایت می‌خوام، ولی باید طبیعی باشه و به موهام آسیب نزنه. کجا رو پیشنهاد می‌دید؟

**خانم ۲:** من "صدف" می‌رم. موادشون خیلی باکیفیت و اصل هست، موهام اصلاً خراب نشد.

**خانم ۳:** به نظرم "مروارید" بهتره. اونجا قبل از شروع، مشاوره می‌دن و با توجه به نوع موها کار می‌کنن، برای همین خیالت راحت‌تره که آسیبی نمی‌رسه.

**خانم ۱:** خب چرا فکر می‌کنی بهتر از "صدف"ه؟

**خانم ۳:** چون من یه بار "صدف" رفتم، نتیجه خوب بود، ولی حس کردم به سلامت موهام کمتر توجه شد.

**خانم ۲:** خب بستگی داره. من تجربه خوبی ازش داشتم. انتخاب با خودته!

تغییر در پایگاه داده یا  
شخصی سازی مدل

شناخت فرایند استنتاج

بازبینی و طراحی ابزار بر  
اساس زیربنای موجود،  
مثلا تمرکز زدایی

توضیح پذیری بالفعل

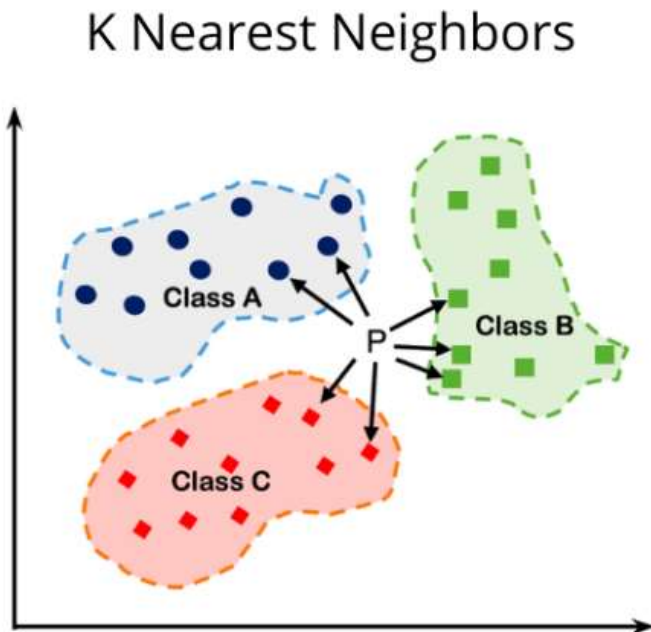
# پیشنهاد: توضیح پذیری بالفعل (ipso facto)

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

- پیدا کردن سهم تاثیر داده‌های ورودی در تصمیم گیری
- اعمال تعادل جدید در این سهم تاثیر و تغییر در تصمیم‌ها

## • یادگیری تنبل (lazy learning)

- K-نزدیک‌ترین همسایه (K-Nearest Neighbors)
- برآورد چگالی پنجره‌ای (Kernel Density Estimators)
- رگرسیون محلی (Local regression)
- محدودیت در کلی‌بینی (generalization) و ابعاد (dimensionality)



# توضیح پذیری بالفعل هوش مصنوعی

- چقدر نمونه‌ی یادگیری  $[x_j, y_j]$  در تصمیم گیری برای نمونه  $x_0$  اثر گذار است؟
- مدل جدید: تصمیم متشکل از ترکیبی از پاسخ‌ها بر اساس نمونه‌های  $x_j$
- به ازای هر نمونه یادگیری یک تصمیم برای نمونه  $x_0$  گرفته شود:  $g(x_0, x_j, y_j)$
- اجماع: جواب‌ها میانگین گرفته بشوند:  $f_D(x_0) = \frac{1}{|D|} \sum_{j \in D} g(x_0, x_j, y_j)$

$$\left. \begin{array}{ll} g(\text{apple}, \text{سیب سرخ}) = \%85 & g(\text{apple}, \text{سیب سرخ}) = \%95 \\ g(\text{apple}, \text{پرتقال بم}) = \%55 & g(\text{apple}, \text{سیب سرخ}) = \%12 \end{array} \right\} f_D(\text{apple}) = \%61$$



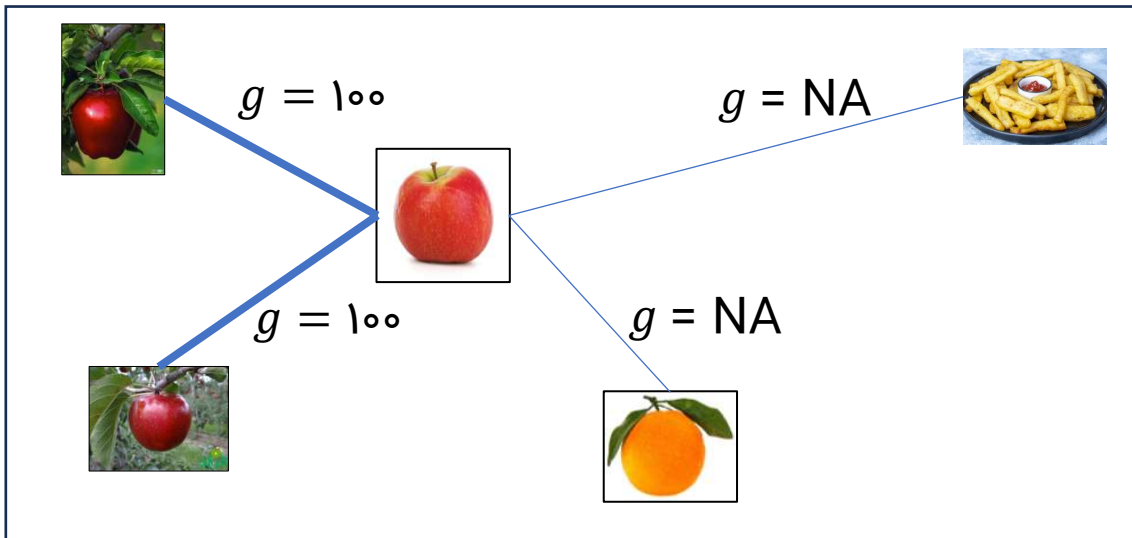
# توضیح پذیری بالفعل هوش مصنوعی

- چقدر نمونه‌ی یادگیری  $[x_j, y_j]$  در تصمیم‌گیری برای نمونه  $x_0$  اثر گذار است؟
- مدل جدید: تصمیم متشکل از ترکیبی از پاسخ‌ها بر اساس نمونه‌های  $x_j$
- به ازای هر نمونه یادگیری یک تصمیم برای نمونه  $x_0$  گرفته شود:  $g(x_0, x_j, y_j)$
- اجماع: جواب‌ها میانگین گرفته بشوند:  $f_D(x_0) = \frac{1}{|D|} \sum_{j \in D} g(x_0, x_j, y_j)$

$$\left. \begin{array}{ll} g(\text{apple}, \text{apple}) = 85\% & g(\text{apple}, \text{apple}) = 95\% \\ g(\text{apple}, \text{orange}) = 55\% & \cancel{g(\text{apple}, \text{apple}) = 12\%} \end{array} \right\} f_D(\text{apple}) = 78\%$$

# توضیح پذیری بالفعل هوش مصنوعی

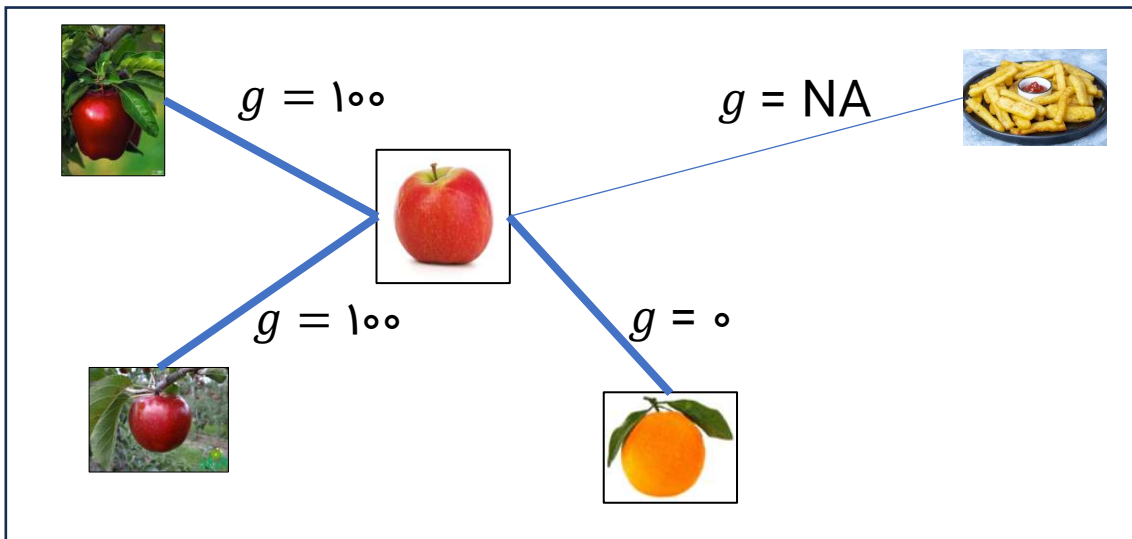
- چقدر نمونه‌ی یادگیری  $[x_j, y_j]$  در تصمیم‌گیری برای نمونه  $x_0$  اثر گذار است؟
- مدل جدید: تصمیم متشکل از ترکیبی از پاسخ‌ها بر اساس نمونه‌های  $x_j$
- به ازای هر نمونه یادگیری یک تصمیم برای نمونه  $x_0$  گرفته شود:  $g(x_0, x_j, y_j)$
- اجماع: جواب‌ها میانگین گرفته بشوند:  $f_D(x_0) = \frac{1}{|D|} \sum_{j \in D} g(x_0, x_j, y_j)$



$$f_D^{2NN}(\text{apple}) = \%100$$

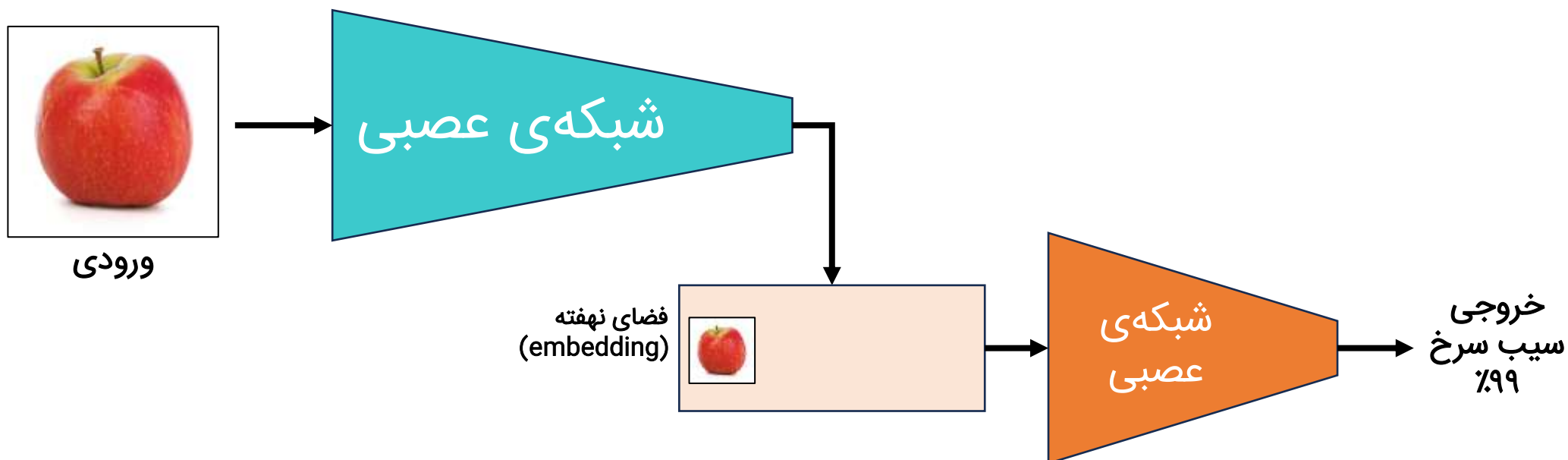
# توضیح پذیری بالفعل هوش مصنوعی

- چقدر نمونه‌ی یادگیری  $[x_j, y_j]$  در تصمیم گیری برای نمونه  $x_0$  اثر گذار است؟
- مدل جدید: تصمیم متشکل از ترکیبی از پاسخ‌ها بر اساس نمونه‌های  $x_j$
- به ازای هر نمونه یادگیری یک تصمیم برای نمونه  $x_0$  گرفته شود:  $g(x_0, x_j, y_j)$
- اجماع: جواب‌ها میانگین گرفته بشوند:  $f_D(x_0) = \frac{1}{|D|} \sum_{j \in D} g(x_0, x_j, y_j)$

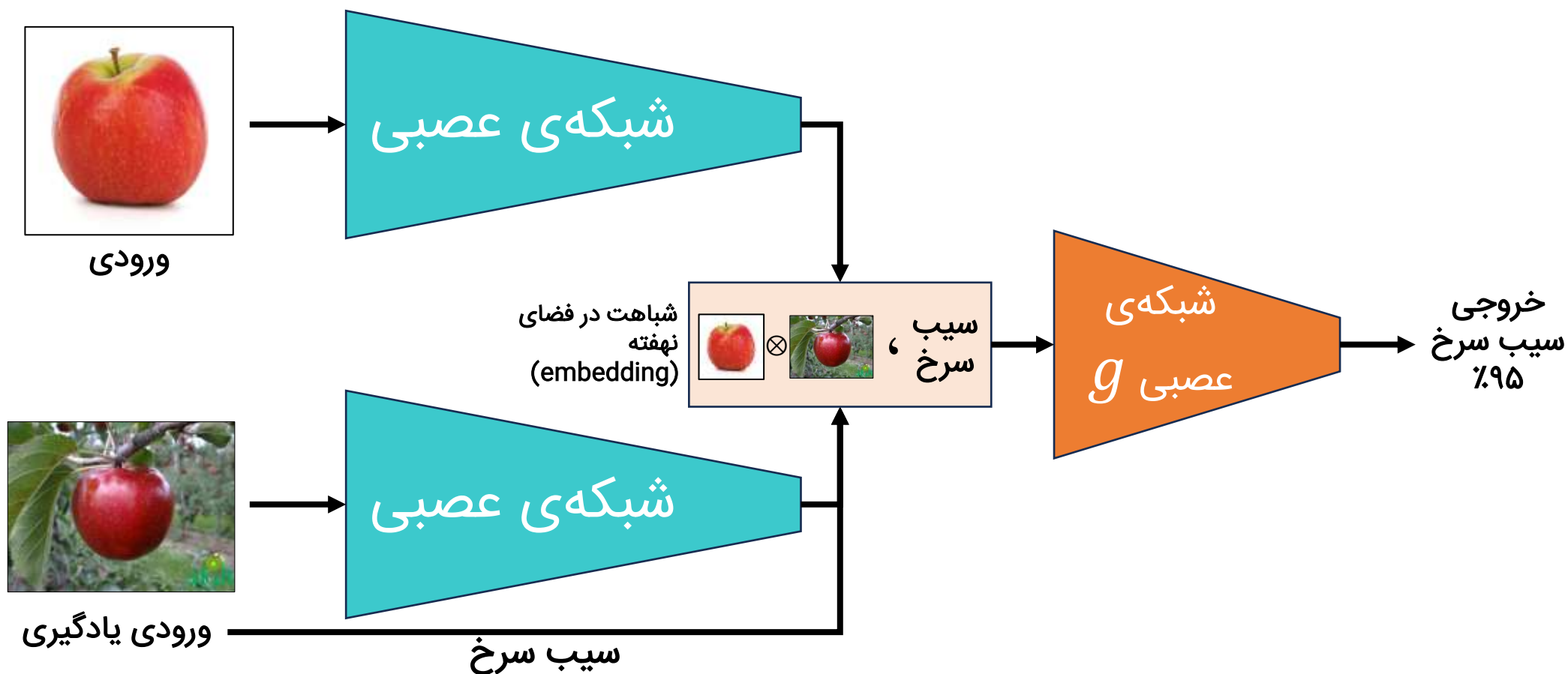


$$f_D^{3NN}(\text{apple}) = 0.66$$

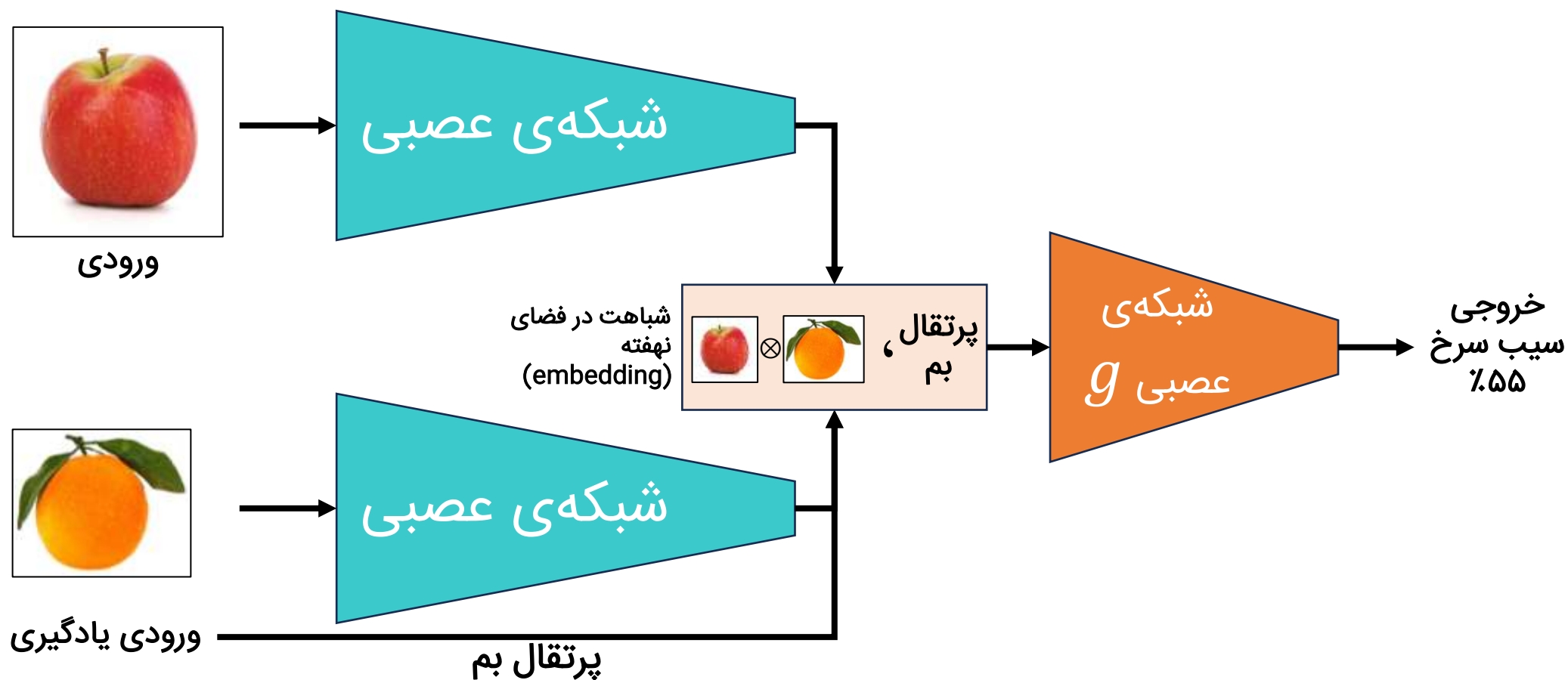
# مدل یکسره (end-to-end)



# توضیح پذیری بالفعل هوش مصنوعی

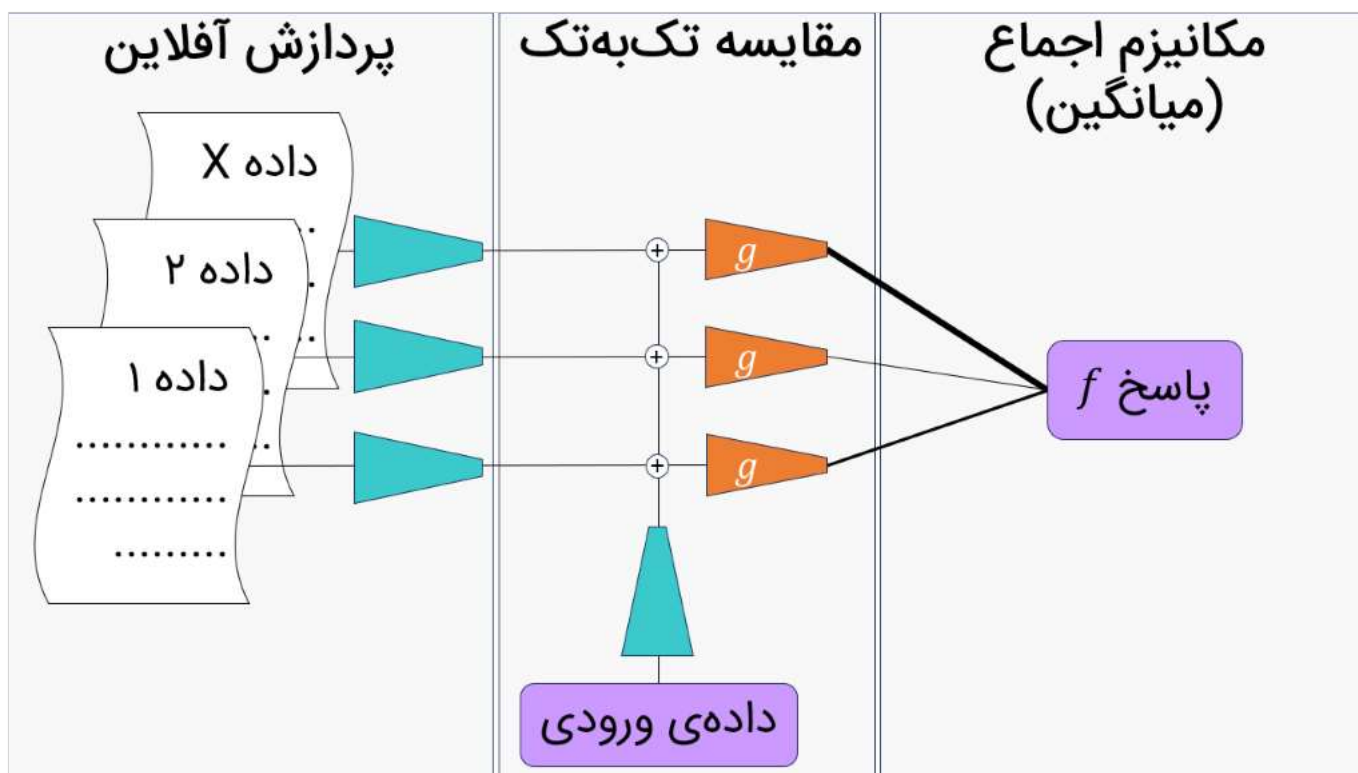


# توضیح پذیری بالفعل هوش مصنوعی



# توضیح پذیری بالفعل هوش مصنوعی

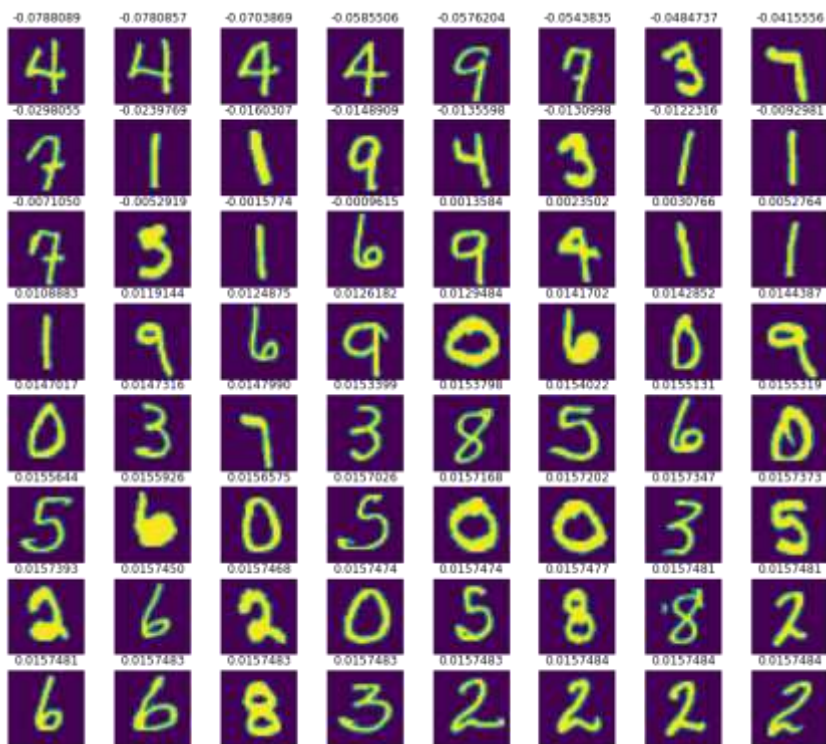
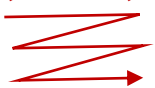
- یک بار پردازش داده‌های یادگیری
- مقایسه با انتخاب منبع
- کنترل سهم تاثیر



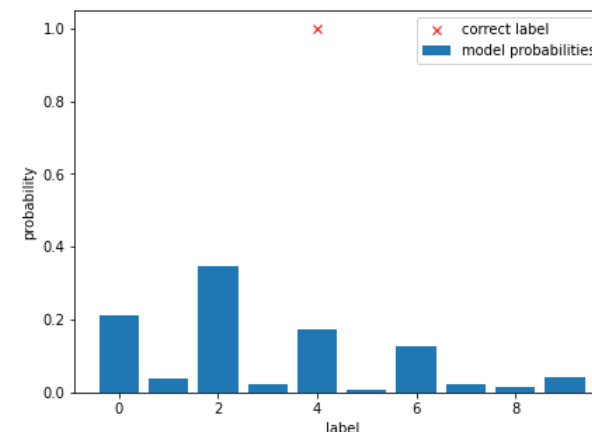
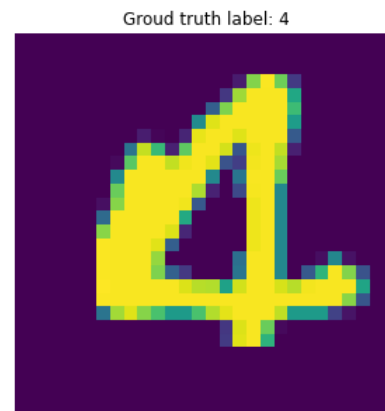


# توضیح‌پذیری بالفعل هوش مصنوعی

ترتیب تاثیر

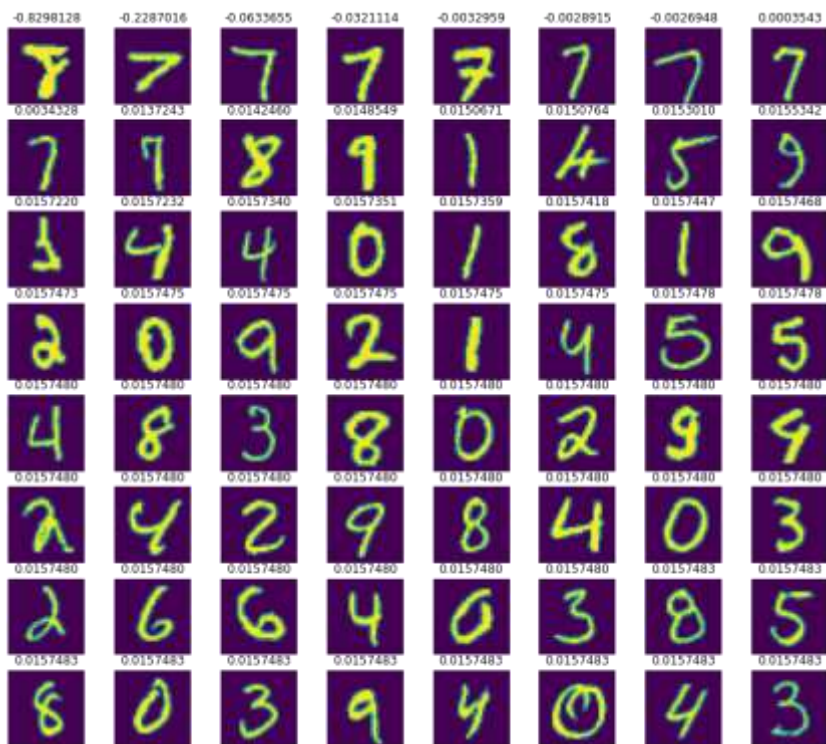
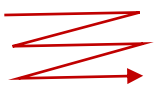


- دسته‌بندی داده‌های MNIST
- داده‌های یادگیری چه تاثیری دارند؟

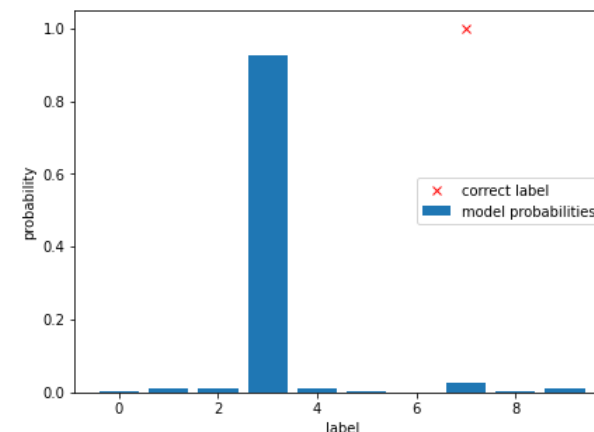
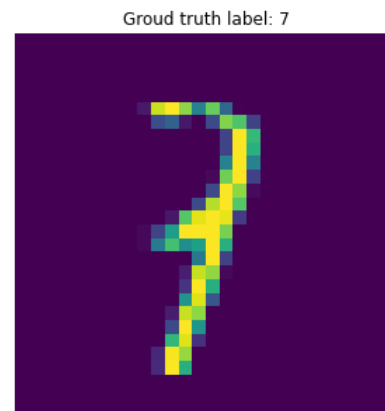


# توضیح‌پذیری بالفعل هوش مصنوعی

ترتیب تاثیر

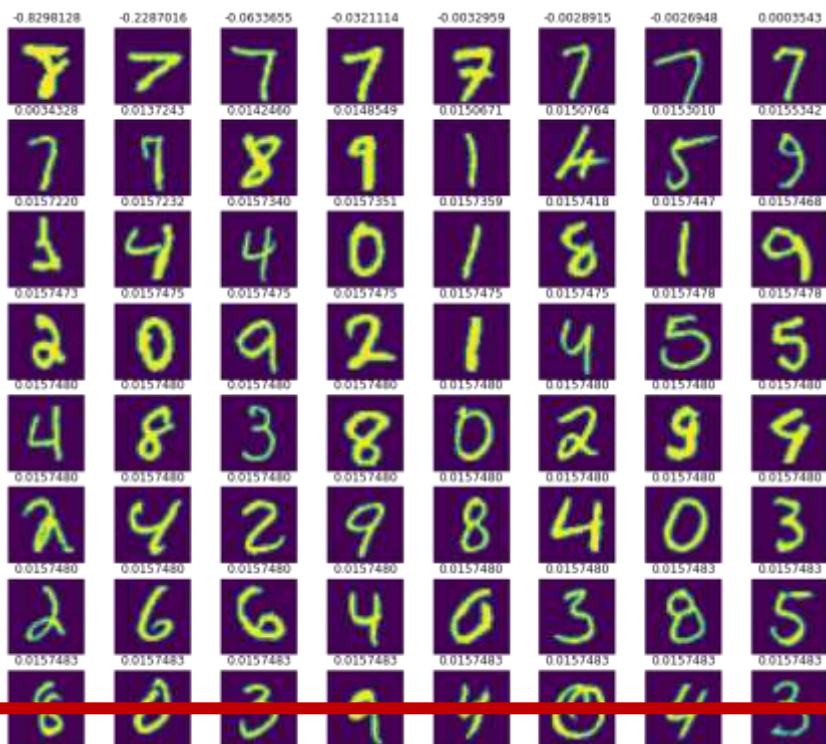
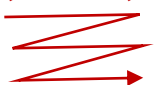


- دسته‌بندی داده‌های MNIST
- داده‌های یادگیری چه تاثیری دارند؟

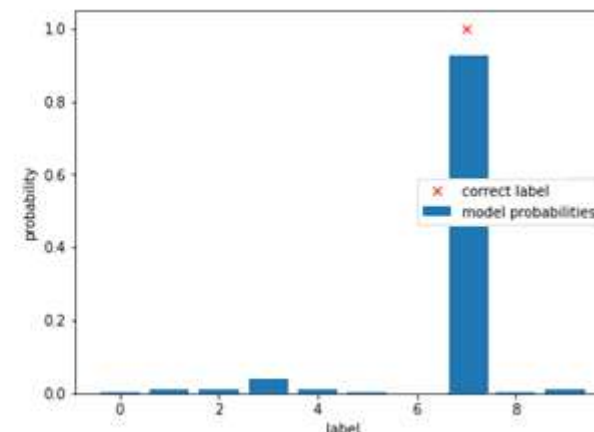


# توضیح‌پذیری بالفعل هوش مصنوعی

ترتیب تاثیر

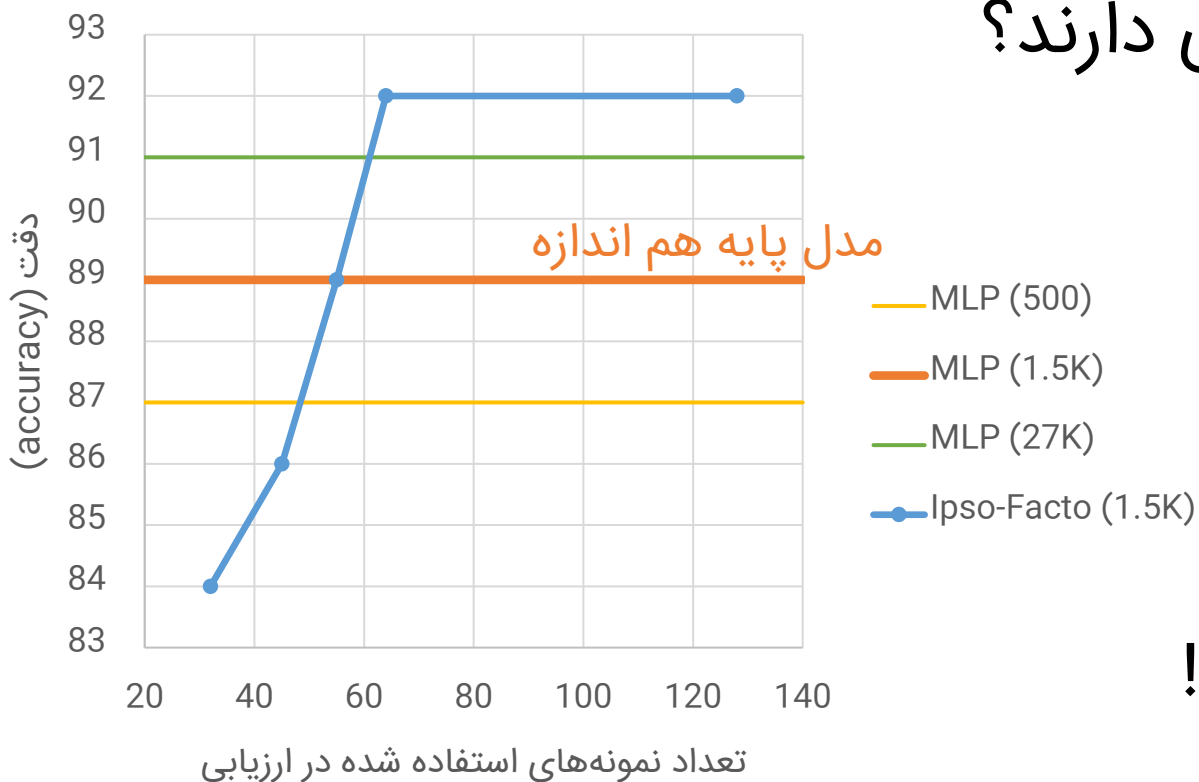


- دسته‌بندی داده‌های MNIST
- داده‌های یادگیری چه تاثیری دارند؟



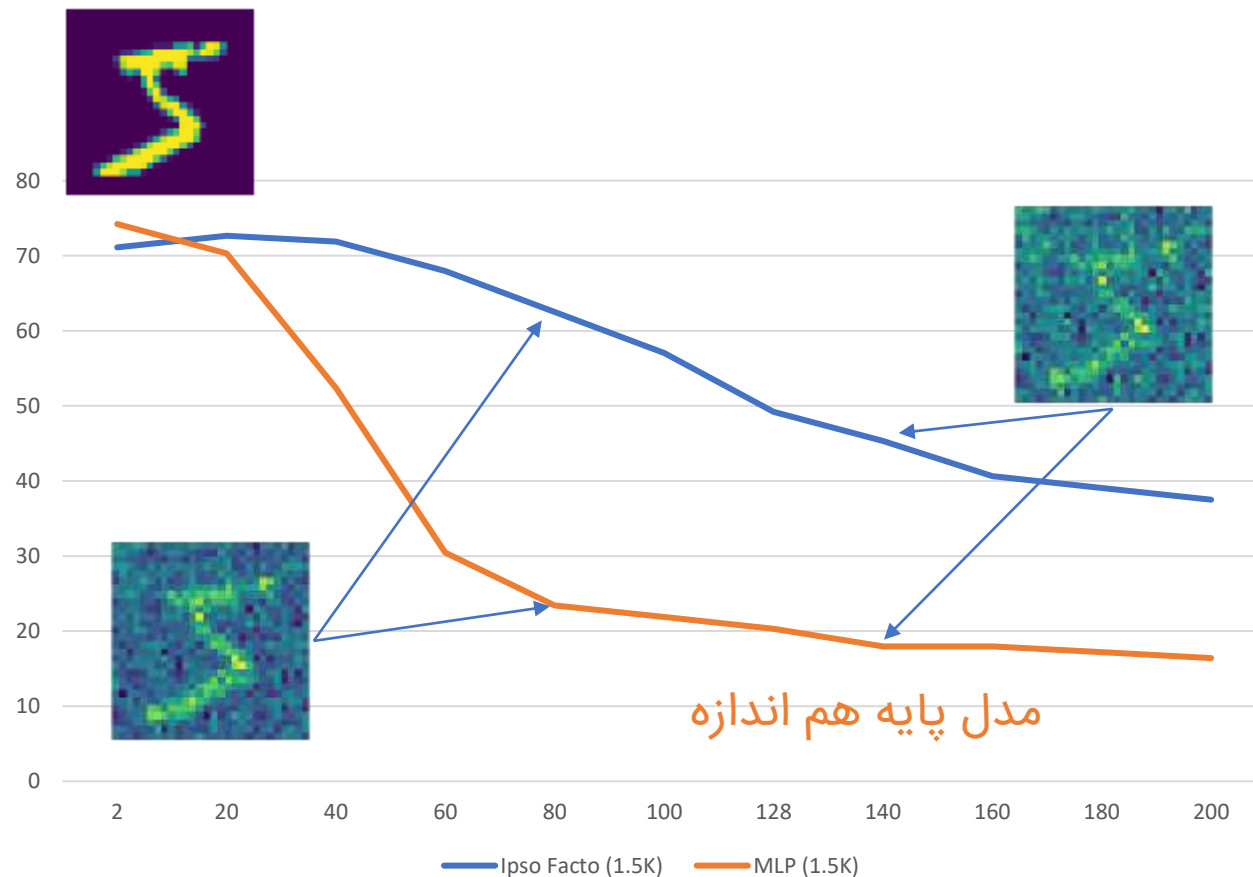
# توضیح پذیری بالفعل هوش مصنوعی

- دسته بندی داده های MNIST
- داده های یادگیری چه تاثیری دارند؟



- جدا سازی حافظه و پردازش!

# توضیح پذیری بالفعل هوش مصنوعی



- دسته بندی داده های MNIST

- قابلیت تعمیم

- تغییر دامنه با نویز

- جدا سازی حافظه و پردازش!

# کاربرد: تشخیص نمونه‌ی خارج از توزیع (OOD)

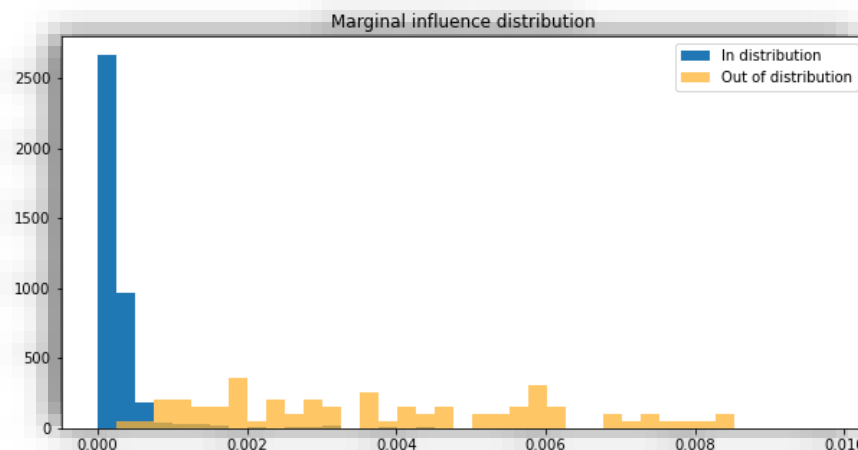
- تشخیص نمونه‌ی خارج از توزیع (Out Of Distribution)
- تاثیر  $x_j$  در تصمیم مدل بر  $x_i$  :  $SI(x_i, x_j)$
- روش آنی: تاثیر کناری نمونه (marginalized sample influence)

$$\sum_{x_i} SI(x_i, x_j) = \sum_{x_i} \Delta_{x_j} \text{Loss}(x_i) = \Delta_{x_j} \text{Loss}$$

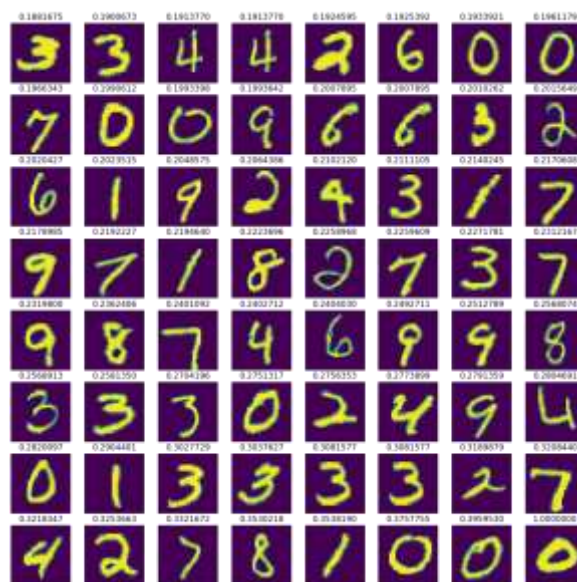


# کاربرد: تشخیص نمونه‌ی خارج از توزیع (OOD)

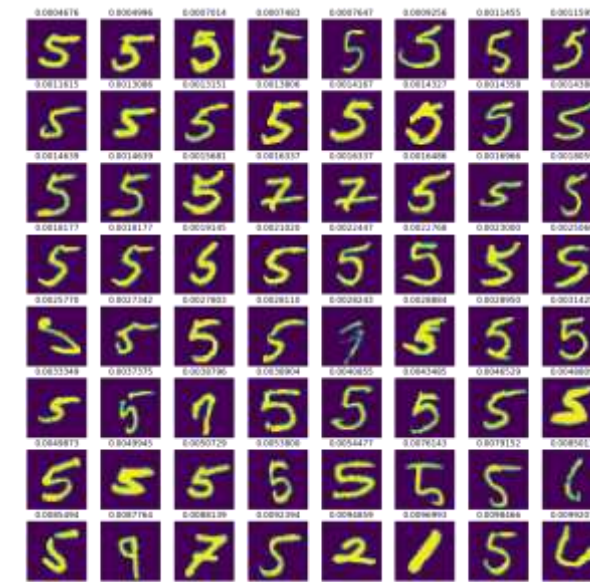
Method	OOD acc.
Parameter change rate $\left  \frac{d\theta}{d\epsilon_j} \right $	0.920
Parameter curvature $\left  \frac{dL_j}{d\epsilon_j} \right $	0.919
Max Prediction	0.916
Nearest neighbor (activation)	<b>0.960</b>
Nearest neighbor (pixels)	0.816
P-value Avg. Active. + Max pred.	0.968
P-value Avg. Active. + $\left  \frac{dL_j}{d\epsilon_j} \right $	<b>0.970</b>
Marginalized Sample Contrib. (64spe)	0.957
Marginalized Sample Importance (32 spe)	0.972
Marginalized Sample Importance (64 spe)	<b>0.977</b>
Marginalized Sample Importance (128 spe)	<b>0.977</b>



نمونه‌های با کمترین تاثیر کناری



نمونه‌های با بیشترین تاثیر کناری



[\[Detecting Extrapolation with Influence Functions\]](#)

دسته‌بندی داده‌های MNIST وقتی مدل ۵ها را در یادگیری ندیده



# توضیح‌پذیری بالفعل هوش مصنوعی

- کنترل سهم تاثیر
- شرطی کردن مدل
- تناسب‌سازی مدل بر اساس دامنه‌ی جستجو (contextualization)
- انتقال منفی (unlearning)
- کنترل سرعت و دقت
- مدل کوچک و قابل انتقال
- قابلیت تمرکز زدایی
- تخمین قطعیت و داده‌های خارج از دامنه

# توضیح پذیری بالفعل هوش مصنوعی

- محدودیت زمانی
  - پیچیدگی خطی نسبت با تعداد داده  $O(N)$
  - قابل تقلیل به لوگاریتمی  $O(\log N)$  با استفاده از الگوریتم KNN
- بهینه‌گی ترکیبات داده
  - یا نیاز به داده‌های زیاد
  - یا محدودیت در تعمیم‌پذیری (generalization)
- پیش فرض: مدل سازی سهم قابل چشم‌پوشی دارد:  $\theta_f, \theta_g \perp x_j$

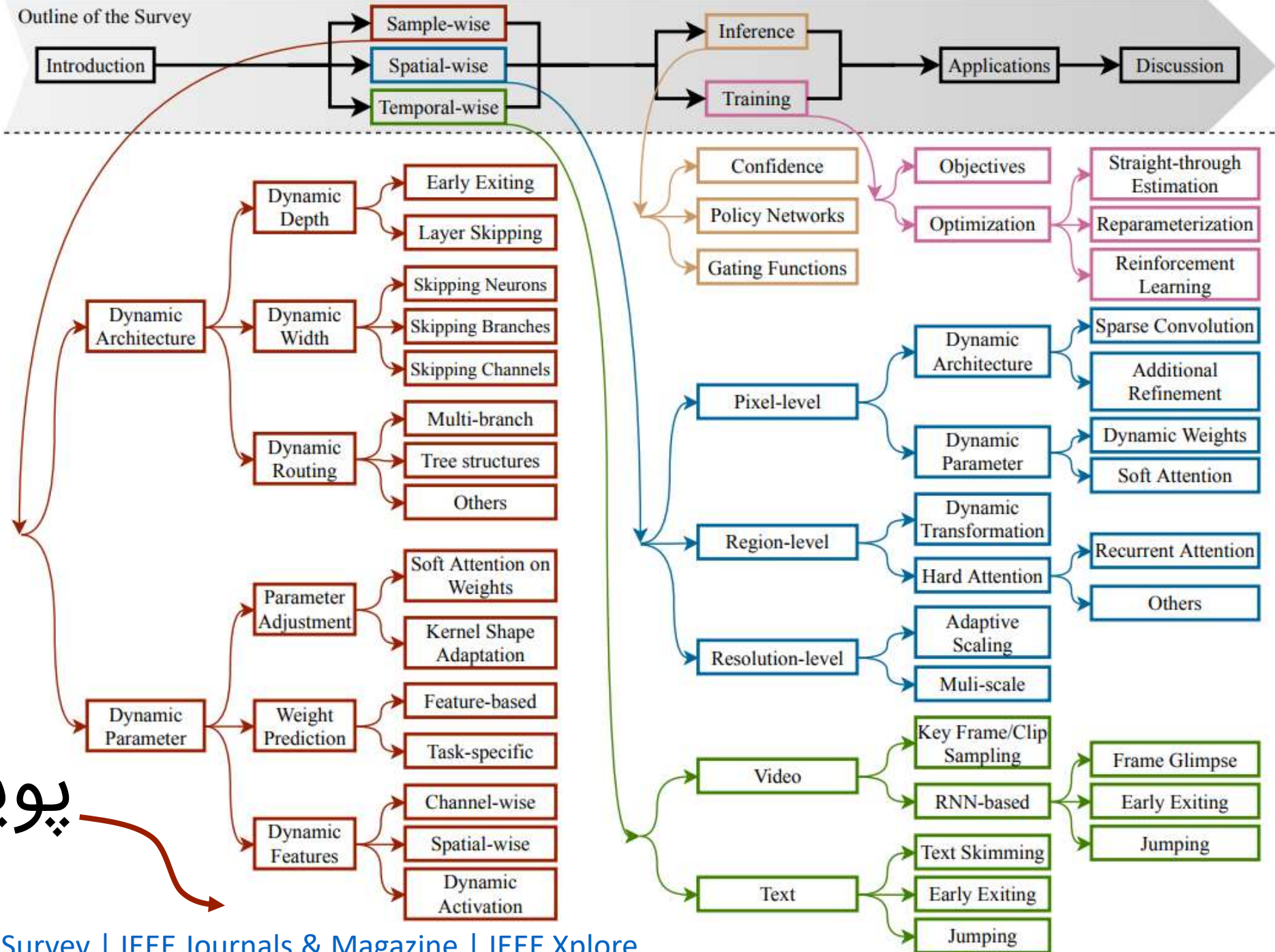
# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

تغییر در پایگاه داده یا  
شخصی سازی مدل

شناخت فرایند استنتاج

بازبینی و طراحی ابزار بر  
اساس زیربنای موجود،  
مثلا تمرکز زدایی



پویایی داده‌ها

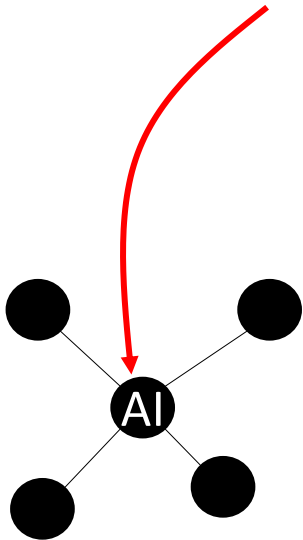
هوش مصنوعی توزیع پذیر

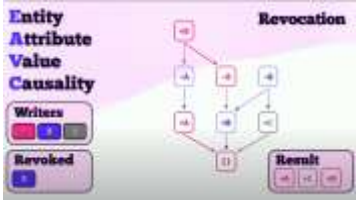
# هوش مصنوعی اجتماعی مرکزگرا

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

- همه‌ی اطلاعات و پردازش در یک مدل قرار میگیرد، مثل ChatGPT
- ارتباطات به محوریت یک عنصر (جمع‌آوری اطلاعات)
- تاثیر ناپذیری مستقیم
- جبر محتوایی
- انحصار در کنترل

محور اطلاعات، مثل مدل‌های زبانی





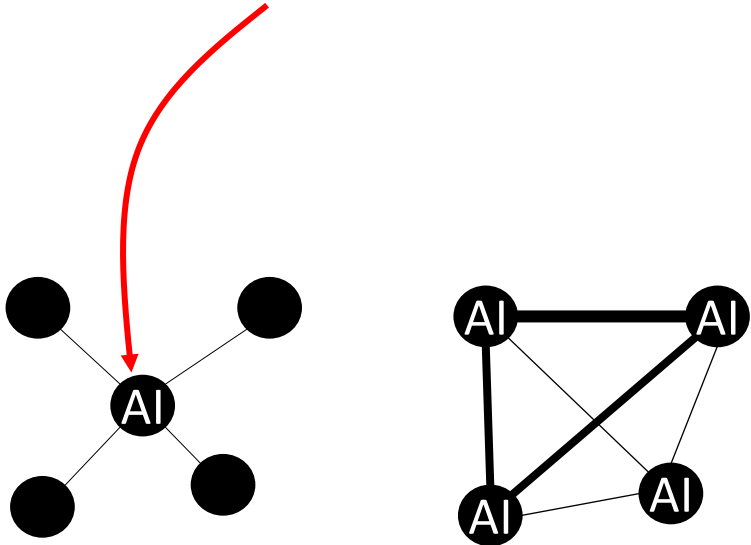
# هوش مصنوعی اجتماعی توزیع پذیر



- همه‌ی اطلاعات و پردازش در شبکه توزیع می‌شوند (distributed)
- ارتباطات ریزوماتیک (Rhizomatic)
- قدرت انتخاب

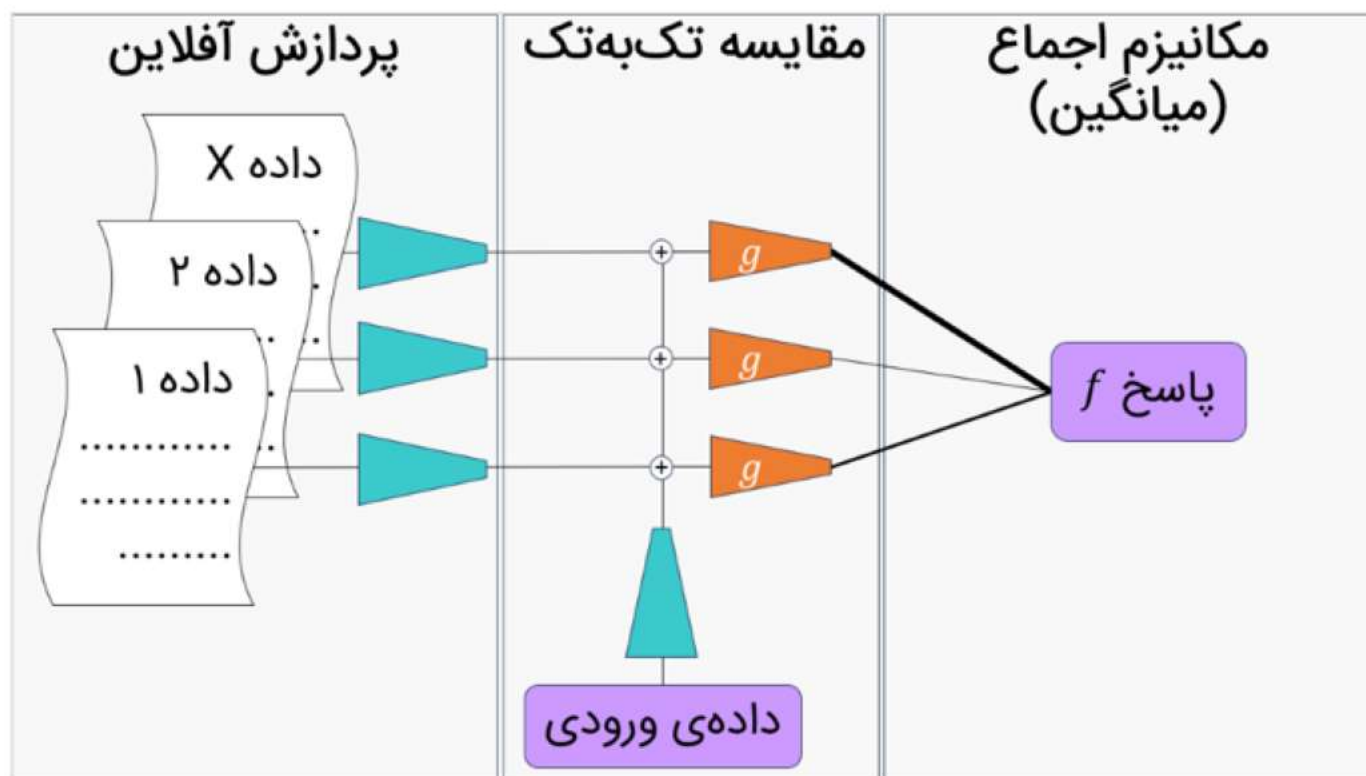
محور اطلاعات، مثل مدل‌های زبانی

- نقض اصل دوگانگی کاربر-سرویس‌دهنده
- نقض انحصار اطلاعاتی و بهره‌وری انحصاری
- قابلیت موازی‌سازی
- ...



# توضیح پذیری بالفعل هوش مصنوعی

- یک بار پردازش داده‌های یادگیری
- مقایسه با انتخاب منبع
- کنترل سهم تاثیر

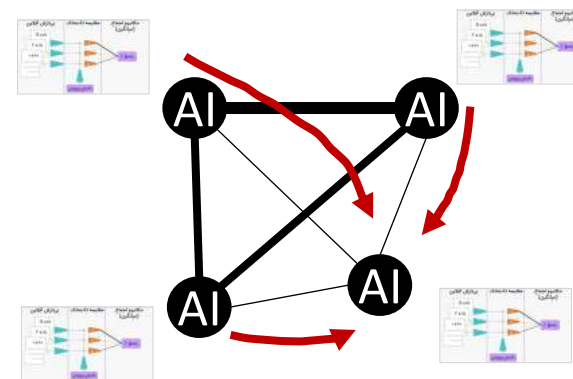
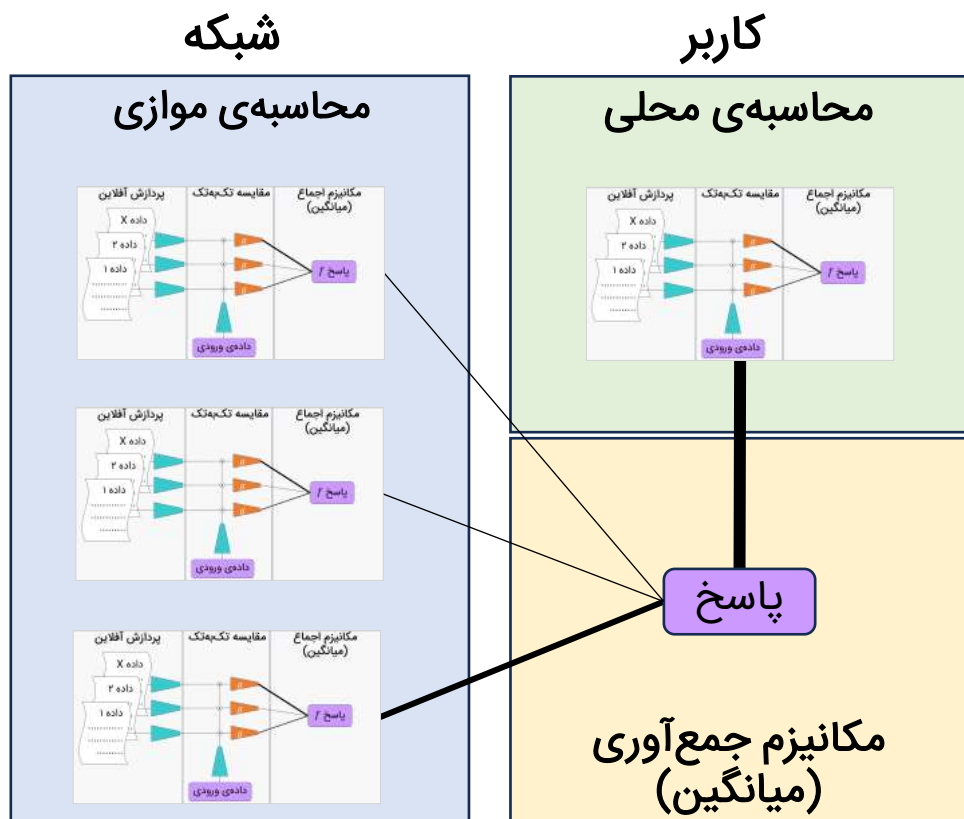




# هوش مصنوعی اجتماعی توزیع پذیر

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

- استفاده از مدل پیشنهادی روی شبکه
- پردازش موازی در هر گره (node)
- جمع‌آوری نتایج بصورت محلی
- انتخاب سهم تاثیر هر گره



# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]

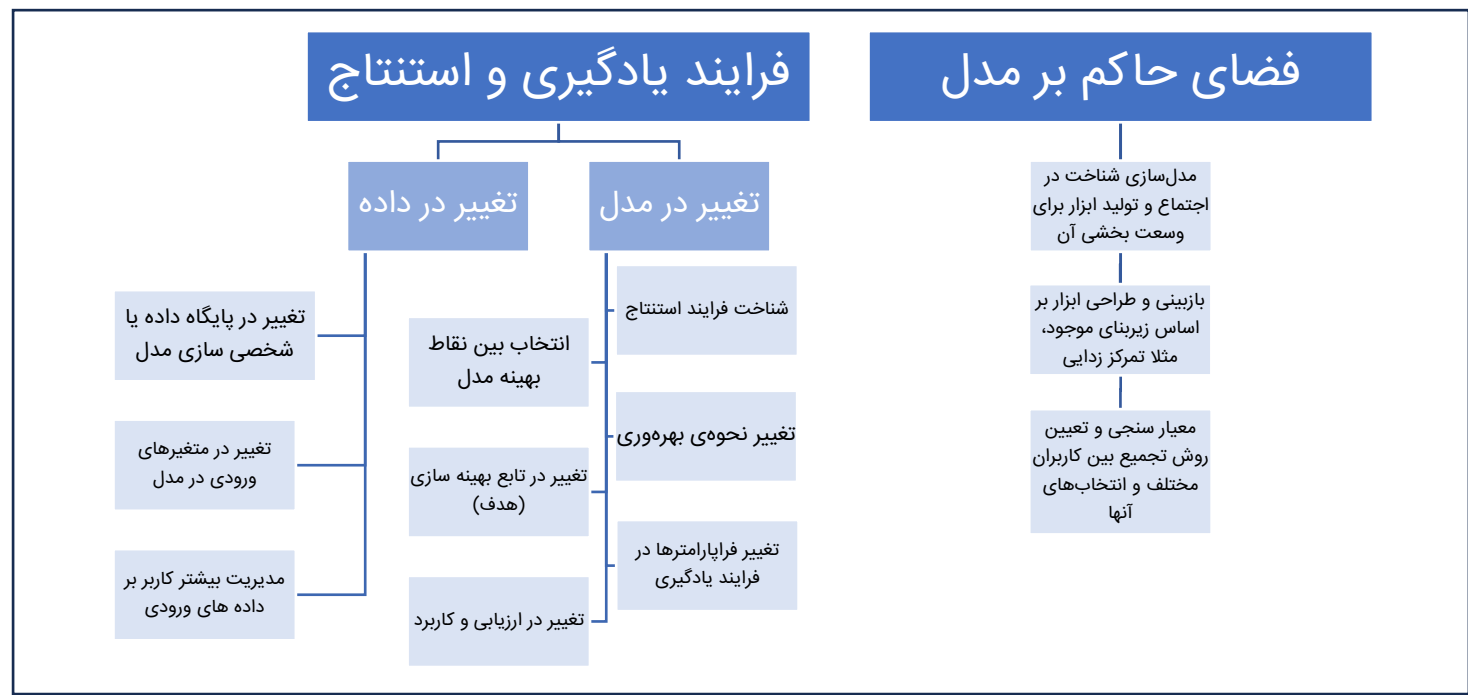
تغییر در پایگاه داده با  
شخصی سازی مدل

شناخت فرایند استنتاج

بازبینی و طراحی ابزار بر  
اساس زیربنای موجود،  
مثلا تمرکز زدایی

# نیازهای هوش مصنوعی در مقابل جریان اصلی

[علی اکبر شمس، محمد کاظم داوودی، حسن کربلایی]



## مسیر پیش رو

- جریان غالب لزوما جریان آینده هوش مصنوعی نیست
- سوال از خواست‌ها شکل گرفته
- خواست شما چیست؟ سوال شما چیست؟

[www.sociai.ir](http://www.sociai.ir)

sarbi@dtu.dk