Web Inference Basics

Ad Targeting and Surveillance Pricing

UC San Diego

A tale of two Target.com visitors

- Target logs website interactions for each shopper
- What preferences, attributes, consumer values might Target infer?

A tale of two Target.com visitors

- What preferences, attributes, consumer values might Target infer?


```
{"event_id":"evt_984a1b02-7c81-48e3-99ab-0193e18aef22",
  "timestamp": "2026-10-13T14:21:05.112Z",
    "session_context": {
    "visitor_id": "anon_visitor_0a91f278d",
    "guest_id_hashed": null,
    "is_authenticated": false,
    "target_circle_linked": false,
    "preferred_store_id": "STR_1772",
    "device": {
      "type": "mobile_web",
      "browser": "Chrome 129.0",
      "screen_resolution": "393x852"},
    "geo_location": {"ip_hash": "e3b0c...91b7852b855",
      "zip_code_inferred": "47906"}},
  "search_payload": {
    "query_raw": "budget twin sheets",
    "query_normalized": "twin sheets budget",
    "results_returned_count": 84,
    "applied_filters": {
      "price_range": {"min": 0, "max": 20.00},
      "size": ["Twin", "Twin XL"],
      "sort_order": "price_low_to_high"}
  },"interaction_telemetry": {
    "scroll_depth_percent": 68,
```

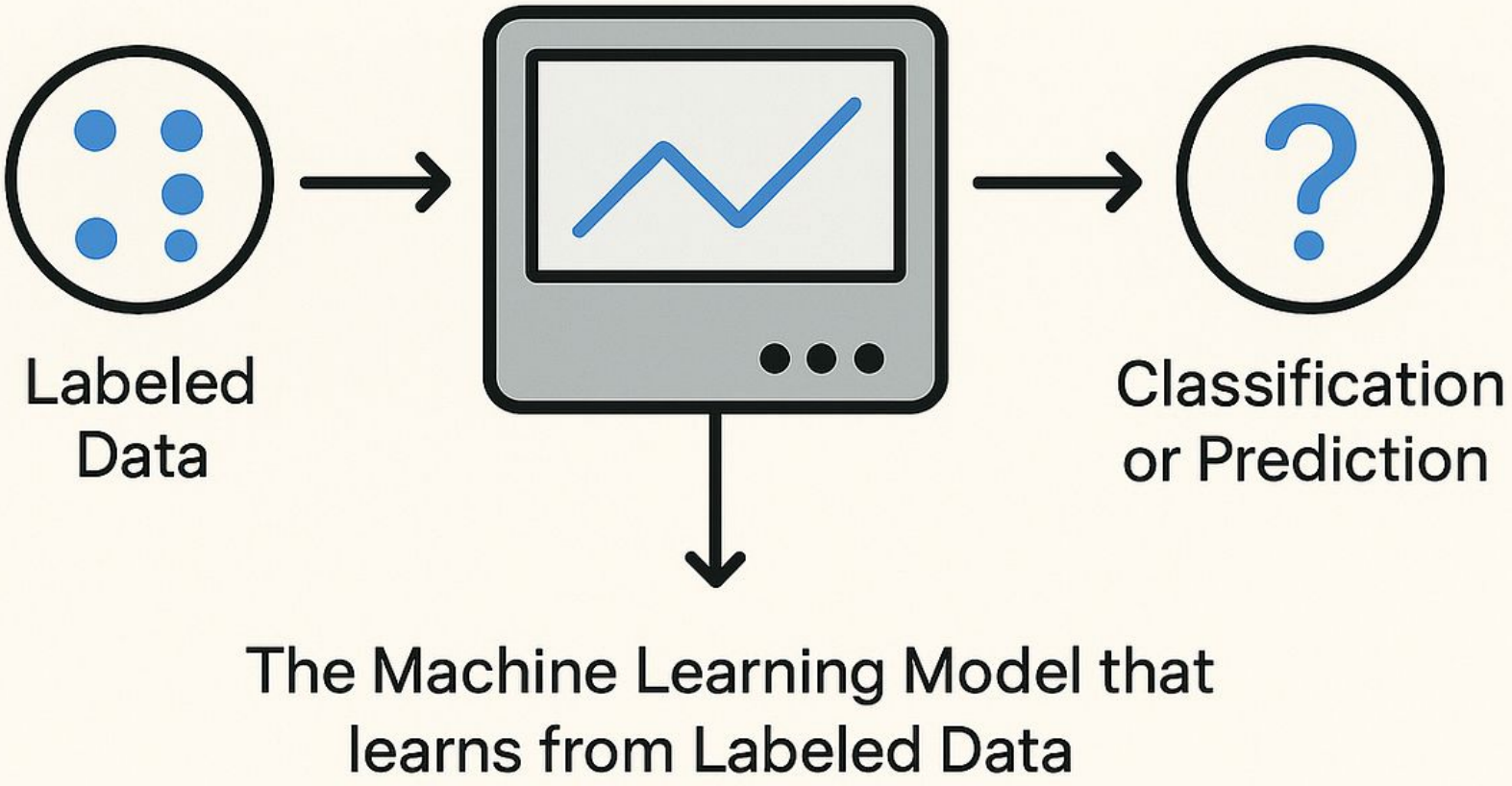
```
{"event_id": "evt_441c8f19-33d2-4e01-a129-901183aa1002",
  "timestamp": "2026-10-13T14:21:08.431Z",
  "session_context": {
    "visitor_id": "anon_visitor_771bc009e",
    "guest_id_hashed": "a3f5809121980a3e881023912a1",
    "is_authenticated": true,
    "target_circle_360_member": true,
    "preferred_store_id": "STR_0284",
    "device": {
      "type": "desktop",
      "browser": "Safari 18.1",
      "screen_resolution": "2560x1440"},
    "geo_location": {"ip_hash": "b2c1102...01bc3290",
      "zip_code_inferred": "90210"}},
  "search_payload": {
    "query_raw": "organic cotton sheets",
    "query_normalized": "organic cotton sheets 100",
    "results_returned_count": 32,
    "applied_filters": {
      "material": ["100% Organic Cotton", "GOTS Certified"],
      "brand": ["Casaluna","Threshold with Studio M"],
      "sort_order": "top_rated"}
  },"interaction_telemetry": {
    "scroll_depth_percent": 92,
```

A tale of two Target.com visitors

- What preferences, attributes, consumer values might Target infer?


```
{"event_id":"evt_984a1b02-7c81-48e3-99ab-0193e18aef22",
  "timestamp": "2026-10-13T14:21:05.112Z",
    "session_context": {
    "visitor_id": "anon_visitor_0a91f278d",
    "guest_id_hashed": null,
    "is_authenticated": false,
    "target_circle_linked": false,
    "preferred_store_id": "STR_1772",
    "device": {
      "type": "mobile_web",
      "browser": "Chrome 129.0",
      "screen_resolution": "393x852"},
    "geo_location": {"ip_hash": "e3b0c...91b7852b855",
      "zip_code_inferred": "47906"}},
  "search_payload": {
    "query_raw": "budget twin sheets",
    "query_normalized": "twin sheets budget",
    "results_returned_count": 84,
    "applied_filters": {
      "price_range": {"min": 0, "max": 20.00},
      "size": ["Twin", "Twin XL"],
      "sort_order": "price_low_to_high"}
  },"interaction_telemetry": {
    "scroll_depth_percent": 68,
```

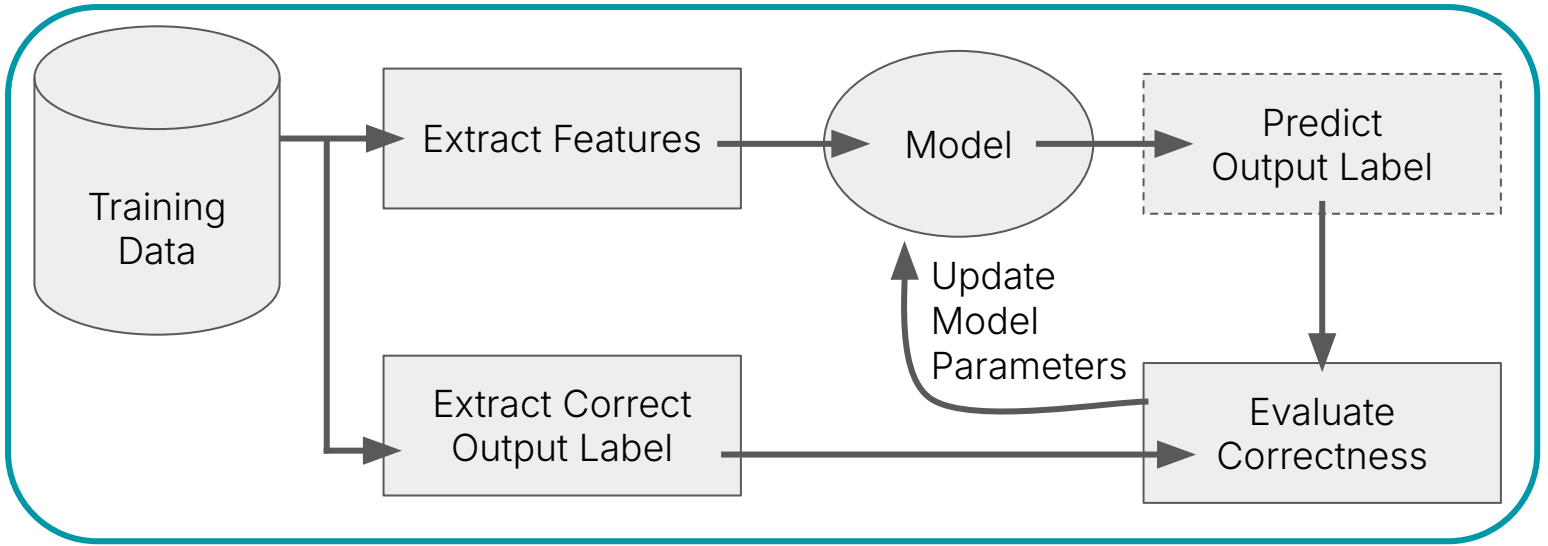
```
{"event_id": "evt_441c8f19-33d2-4e01-a129-901183aa1002",
  "timestamp": "2026-10-13T14:21:08.431Z",
  "session_context": {
    "visitor_id": "anon_visitor_771bc009e",
    "guest_id_hashed": "a3f5809121980a3e881023912a1",
    "is_authenticated": true,
    "target_circle_360_member": true,
    "preferred_store_id": "STR_0284",
    "device": {
      "type": "desktop",
      "browser": "Safari 18.1",
      "screen_resolution": "2560x1440"},
    "geo_location": {"ip_hash": "b2c1102...01bc3290",
      "zip_code_inferred": "90210"}},
  "search_payload": {
    "query_raw": "organic cotton sheets",
    "query_normalized": "organic cotton sheets 100",
    "results_returned_count": 32,
    "applied_filters": {
      "material": ["100% Organic Cotton", "GOTS Certified"],
      "brand": ["Casaluna","Threshold with Studio M"],
      "sort_order": "top_rated"}
  },"interaction_telemetry": {
    "scroll_depth_percent": 92,
```



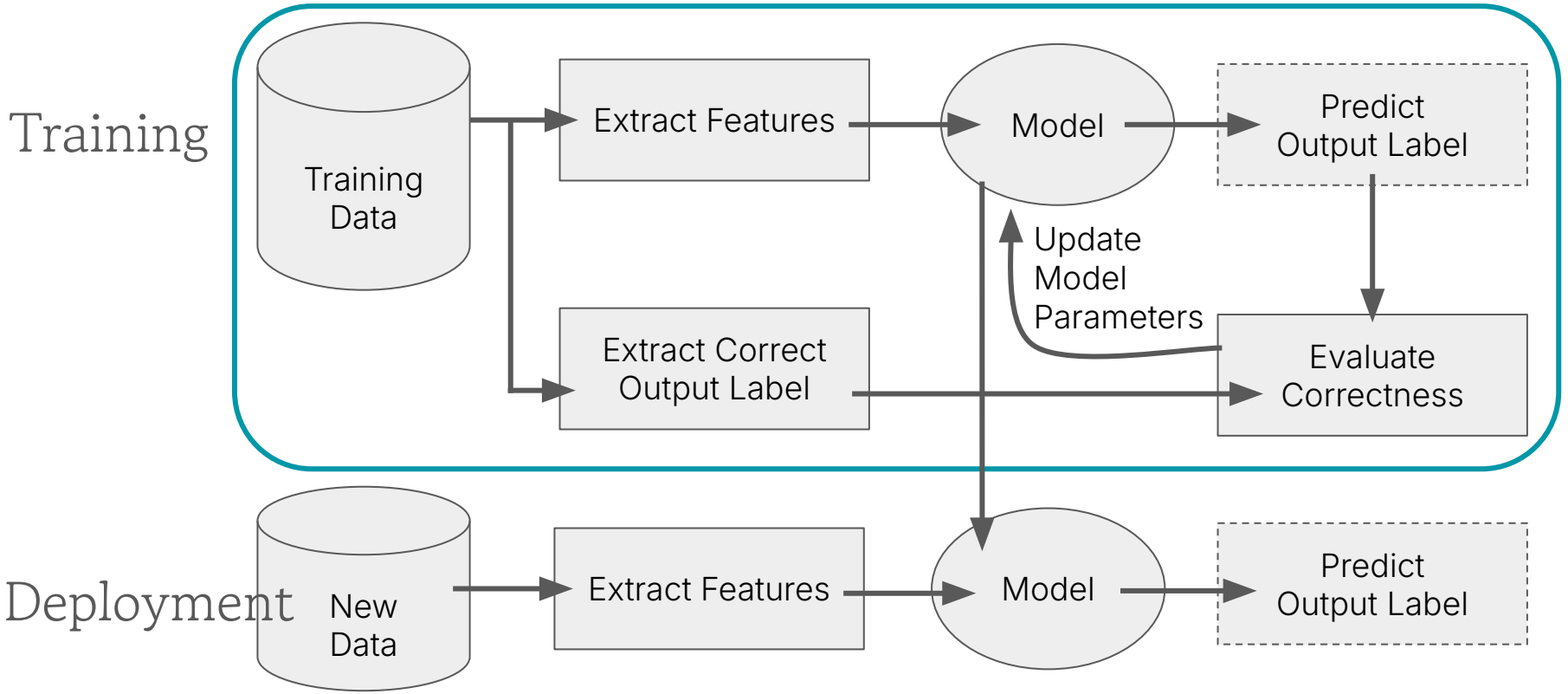
D M Kenura Dasanayaka "Machine Learning Made Simple: How Computers Learn From Data..."

How Machines Learn

Training



How Machines Learn



Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **features:** individual measurable property or characteristic of a data set, which will be used by the model to make predictions
  aka attributes, dimensions
- ideally, will be informative, discriminating, and independent

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **features:** individual measurable property or characteristic of a data set, which will be used by the model to make predictions
  aka attributes, dimensions
- ideally, will be informative, discriminating, and independent

what's an example in ad targeting?

User Profiles

- **demographic data**: age, gender, education level, relationship status, major life events
- **psychographic data:** consumer values, lifestyle choices, opinions/attitudes
- **behavioral data:** browsing history, search history, subpages viewed, purchase history, product categories viewed, the monetary value of previous orders, cart abandonment signals, time between products
- **social media/app usage:** likes, shares, follows, types of apps (e.g., fitness apps, travel planner)
- **device/browser data:** device type (phone, computer, tablet), device model, screen resolution, OS version, browser version
- **environmental data:** geolocation, commute pattern, weather

Input Data

Jeff was recently hired to work on the inferences team at Boogle. His boss asked him to develop a model to predict interest in men's and women's fashion so they can place ads more effectively.

He uses other data they have to predict gender. Jeff's model can't operate on text, so to represent this data numerically, he uses: 0 = men and 1 = women

Jeff's boss asks to meet with him to check on the project.

What feedback might Jeff get?

Data Types (for statistics)

numerical: information recorded as numbers, where there is both an ordering and consistent spacing

ordinal: information where there is an order or rank, though the exact distance between ranks may not be equal

categorical: information recorded as groups or labels, generally no sense of order, ranking, or mathematical meaning

Data Types

numerical: information recorded as numbers, where there is both an ordering and consistent spacing

ex: age, past purchase values, time between purchases

0, 1, 2, 3, ...
$100, $100.01, $100.02...

ordinal: information where there is an order or rank, though the exact distance between ranks may not be equal

ex: education, device type?

Kindergarten... PhD
Tracfone ... iPhone Duo

categorical: information recorded as groups or labels, generally no sense of order, ranking, or mathematical meaning

ex: gender, major life events, types of apps

Graduate HS
Married
Birth/adoption
New job

Data Types

numerical: information recorded as numbers, where there is both an ordering and consistent spacing

ex: age, past purchase values, time between purchases

0, 1, 2, 3, ...
$100, $100.01, $100.02...

ordinal: information where there is an order or rank, though the exact distance between ranks may not be equal

ex: education, device type?

Kindergarten... PhD
Tracfone ... iPhone Duo

⚠

categorical: information recorded as groups or labels, generally no sense of order, ranking, or mathematical meaning

ex: gender, major life events, types of apps

Graduate HS
Married
Birth/adoption
New job

Data Types

numerical: information recorded as numbers, where there is both an ordering and consistent spacing

ex: age, past purchase values, time between purchases

0, 1, 2, 3, ...
$100, $100.01, $100.02...

ordinal: information where there is an order or rank, though the exact distance between ranks may not be equal

ex: education, device type?

Kindergarten... PhD
Tracfone ... iPhone Duo

⚠

categorical: information recorded as groups or labels, generally no sense of order, ranking, or mathematical meaning

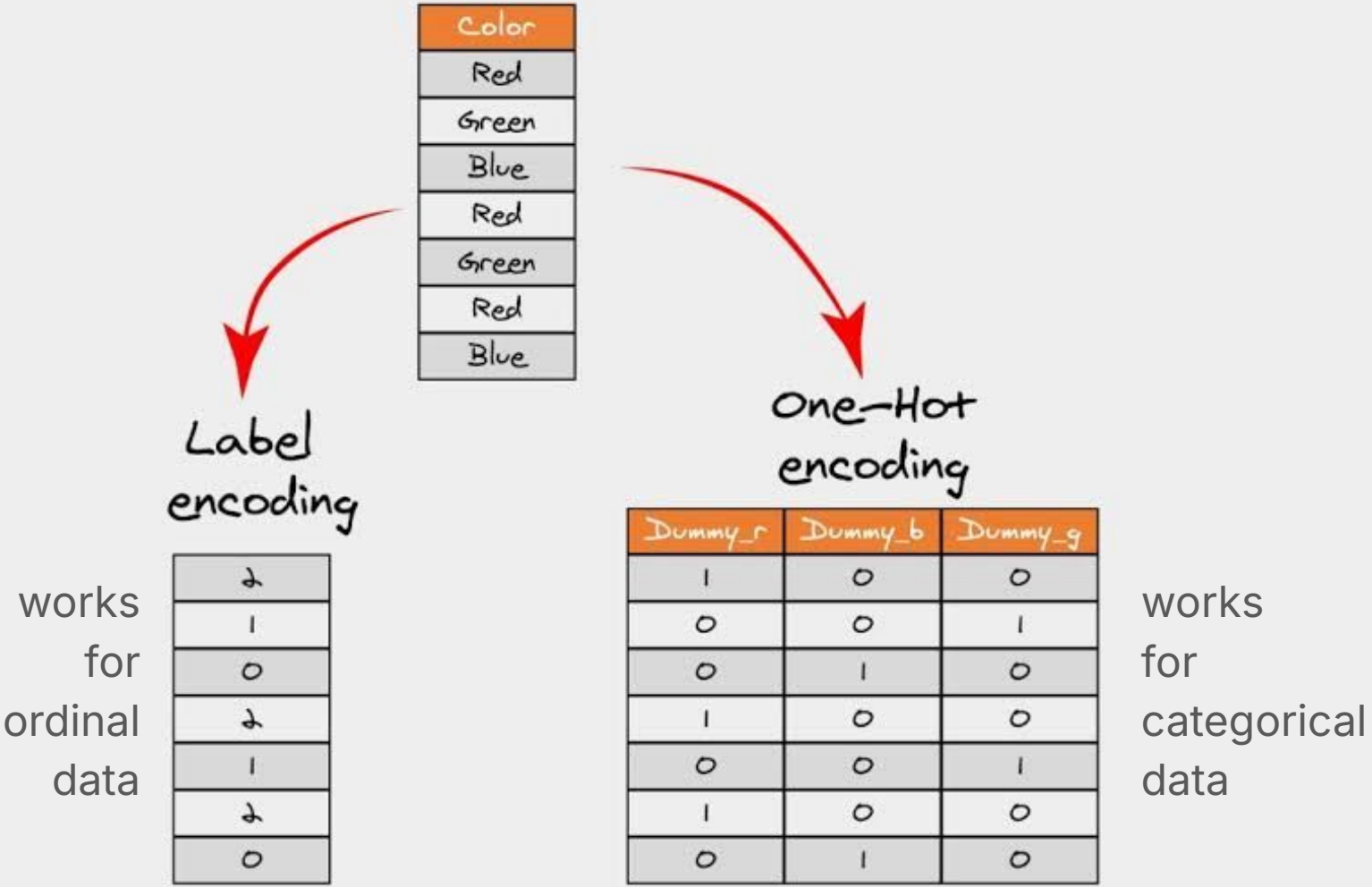
categorical data can be tricky for ML models

One-hot encoding

- **one hot encoding:** converts categorical data to quantitative data
- treats all categories as independent and equal
- Step 1: create a new feature for every category in a variable
- Step 2: use 1 | 0 to indicate the presence | absence of that category



Graduate_HS 0
Married 0
Birth_adoption 1
Death_in_family 0
New_job 0
Retire 0
Buy home 0
Major illness 0


Color
Red
Green
Blue
Red
Green
Red
Blue
Label encoding
2
1
0
2
1
2
0
works for ordinal data
One-Hot encoding
Dummy_r
Dummy_b
Dummy_g
1 0 0
0 0 1
0 1 0
1 0 0
0 0 1
1 0 0
0 1 0
works for categorical data

One-hot encoding

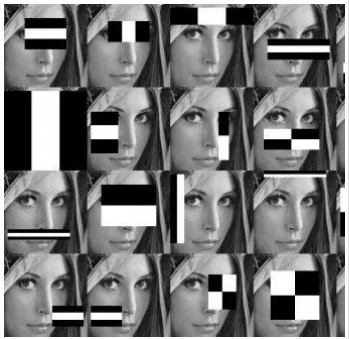
- **one hot encoding:** converts categorical data to quantitative data
- treats all categories as independent and equal
- Step 1: create a new feature for every category in a variable
- Step 2: use 1 | 0 to indicate the presence | absence of that category
- downsides:
    - high dimensionality: many categories → many columns!
    - multicollinearity: two or more features can predict another
        - drop one of the columns!

Where does this data come from?

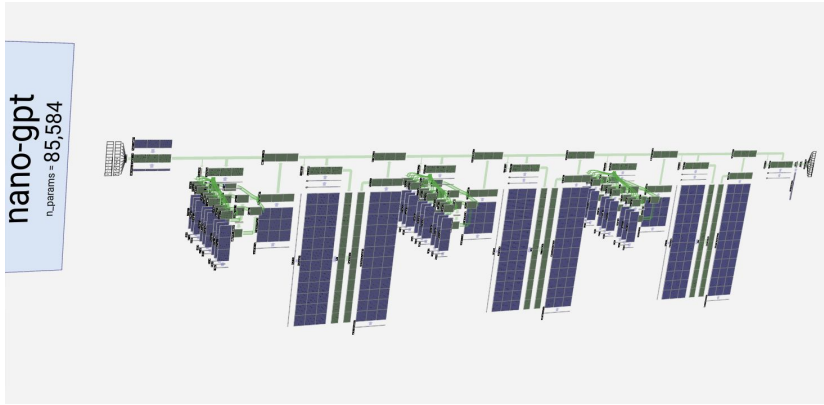
- self-reported data (e.g., membership/loyalty programs)
- prior purchase data
- cookies (Week 2)
- browser fingerprinting (Week 7)
- data brokers (Week 5)
- inferences from other data

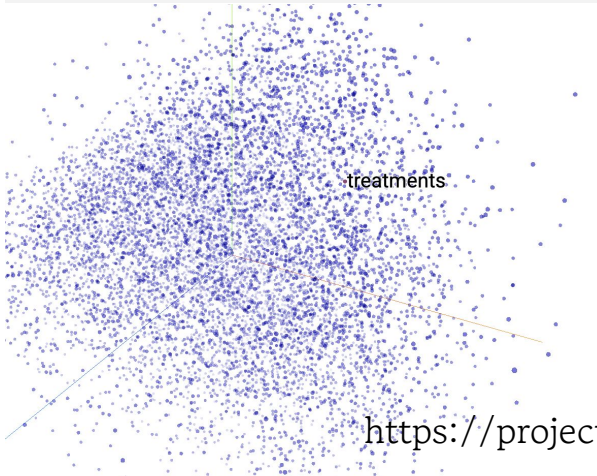
Selecting Features





Social Processes	talk, us, friend
Friends	pal, buddy, coworker
Family	mom, brother, cousin
Humans	boy, woman, group
Affective Processes	happy, ugly, bitter
Positive Emotions	happy, pretty, good
Negative Emotions	hate, worthless, enemy
Anxiety	nervous, afraid, tense
Anger	hate, kill, pissed
Sadness	grief, cry, sad
Cognitive Processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess





https://projector.tensorflow.org/

Term Frequency-Inverse Document Frequency (TF-IDF)

- measure of how characteristic a word is of a specific piece/category of text relative to other text (e.g., one website vs entire internet)
- **Term Frequency (TF):** how frequently a word appears in a specific document, normalized for the total number of words in the doc
- **Inverse Document Frequency (IDF):** how rare or distinctive a word is in the entire corpus of documents

$$\frac{\text{COUNT (}\textit{word W in document D}\text{)}}{\text{COUNT (}\textit{all words in document D}\text{)}} \times \frac{\text{COUNT (}\textit{all documents}\text{)}}{\text{COUNT (}\textit{documents containing word W}\text{)}}$$

Term Frequency-Inverse Document Frequency (TF-IDF)

Journal of Documentation
Volume 28 Number 1 1972 pp. 11-21

A statistical interpretation of term specificity and its application in retrieval

Karen Spärck Jones

Computer Laboratory, University of Cambridge, Cambridge, UK

Abstract: The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined, experiments with three test collections showing, in particular, that frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms. Results for the test collections show that considerable improvements in performance are obtained with this very simple procedure.

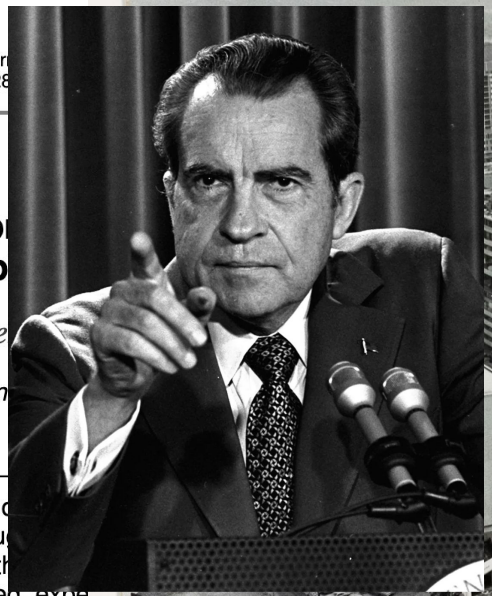
of a specific piece/category
bsite vs entire internet)
word appears in a specific
er of words in the doc
v rare or distinctive a word is

OUNT (*all documents)*

---

documents containing word W)

Term Frequency (TF-IDF)



Jour
Volume 28

**A statistical interp
and its app**

Kare

Computer Laboratory, Un

Abstract: The exhaustivity of document
usually regarded as independent. It is su
statistically, as a function of term use rat
variations in term specificity are examined, expe
particular, that frequently-occurring terms are re
argued that terms should be weighted according
less frequent, more specific, terms are of greate
Results for the test collections show that conside
obtained with this very simple procedure.

"All the News That's Fit to Print"

The New York Times

LATE CITY EDITION

Weather: Chance of showers later today, tonight. Milder tomorrow. Temp. ranges: today 36-46; Friday 43-49. Additional details on Page 62.

VOL. CXXIII ... No. 42,406 © 1974 The New York Times Company NEW YORK, SATURDAY, MARCH 2, 1974 15 CENTS

FEDERAL GRAND JURY INDICTS 7 NIXON AIDES ON CHARGES OF CONSPIRACY ON WATERGATE; HALDEMAN, EHRLICHMAN, MITCHELL ON LIST

COLSON IS NAMED

A Question of Veracity of the President Is Indirectly Raised

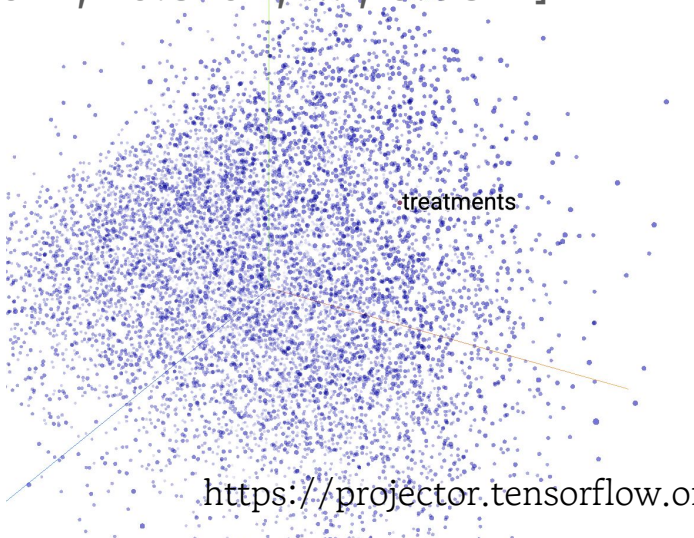
By ANTHONY RIPLEY

Word Embeddings

- low-dimensional vector representations of text
- words with similar meanings/contexts are nearby in the vector space
- much more dense/compact than one-hot encodings/bag of words

treatments = [0.2145 -0.1189, 0.0432, 0.5912, -0.3104, ... , 0.0871]

- ⚠ vector dimensions do not map onto labels/topics
- but geometric distance between this vector and those for words like *therapy*, *medication*, *surgery* will be small



https://projector.tensorflow.org/

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- **features:** individual measurable property or characteristic of a data set, which will be used by the model to make predictions
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- **features:** individual measurable property or characteristic of a data set, which will be used by the model to make predictions
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.

what's the classic measure to predict online ad performance?

Clickthrough Rate (CTR)

- **CTR:** ratio showing how often people who see an ad click it
- **impressions:** the number of people who see an ad

$$\text{CTR} = \frac{\text{clicks}}{\text{impressions}}$$

Clickthrough Rate (CTR)

- **CTR:** ratio showing how often people who see an ad click it
- **impressions:** the number of people who see an ad

$$\text{CTR} = \frac{\text{clicks}}{\text{impressions}}$$

- **BUT!** clicks are "cheap signals"

Signaling Theory

- Explains how people & animals communicate by sending information using observable actions/traits
- examples:



colorful feathers as a signal of a bird's ability to gather nutrients



college degree as a signal of a job seeker's productivity

Michael Spence "Job Market Signaling" 1973 & Amotz Zahavi "Mate Selection—A Selection for a Handicap" 1975

Signaling Theory

J. theor. Biol. (1975) **53**, 205–214

JOB MARKET SIGNALING *

MICHAEL SPENCE

1. Introduction, 355. — 2. Hiring as investment under uncertainty, 356. — 3. Applicant signaling, 358. — 4. Informational feedback and the definition of equilibrium, 359. — 5. Properties of informational equilibria: an example, 361. — 6. The informational impact of indices, 368. — Conclusions, 374.

1. INTRODUCTION

The term "market signaling" is not exactly a part of the well-defined, technical vocabulary of the economist. As a part of the preamble, therefore, I feel I owe the reader a word of explanation about the title. I find it difficult, however, to give a coherent and comprehensive explanation of the meaning of the term abstracted from the contents of the essay. In fact, it is part of my purpose to outline a model in which signaling is implicitly defined and to explain why one can, and perhaps should, be interested in it. One might accurately characterize my problem as a signaling one, and that of the reader, who is faced with an investment decision under uncertainty, as that of interpreting signals.

Mate Selection—A Selection for a Handicap

AMOTZ ZAHAVI

Institute for Nature Conservation Research, Faculty of Life Sciences, The George S. Wise Centre for Life Sciences, Tel-Aviv University, Israel

(*Received* 23 *July* 1974, *and in revised form* 2 *December* 1974)

It is suggested that characters which develop through mate preference confer handicaps on the selected individuals in their survival. These handicaps are of use to the selecting sex since they test the quality of the mate. The size of characters selected in this way serve as marks of quality. The understanding that a handicap, which tests for quality, can evolve as a consequence of its advantage to the individual, may provide an explanation for many puzzling evolutionary problems. Such an interpretation may provide an alternative to other hypotheses which assumed complicated selective mechanisms, such as group selection or kin selection, which do not act directly on the individual.

In his theory of sexual selection, Darwin (1874) tried to explain the evolution of characters such as the antlers of deer, the tail plumes of the peacock, the brilliant colouration of many birds and their fantastic displays and songs, by the cumulative effect of females preference for certain male types. He suggested that the disadvantages to male survival induced by such characters,

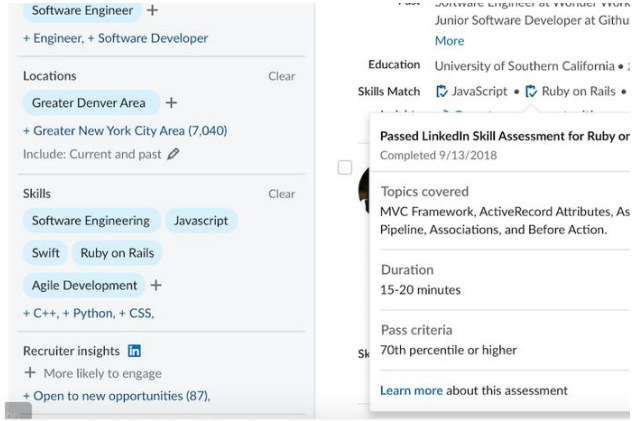
Costly Signal

Message that requires sacrifice of money/time/energy, difficult or impossible to fake



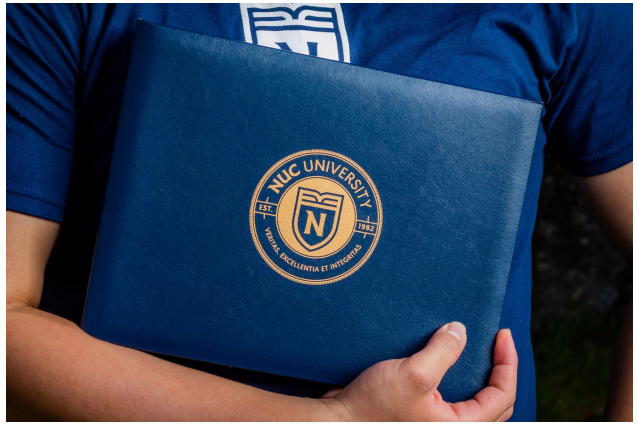
Cheap Signal

Message that costs little to nothing to send, may not reflect reality, easy to fake



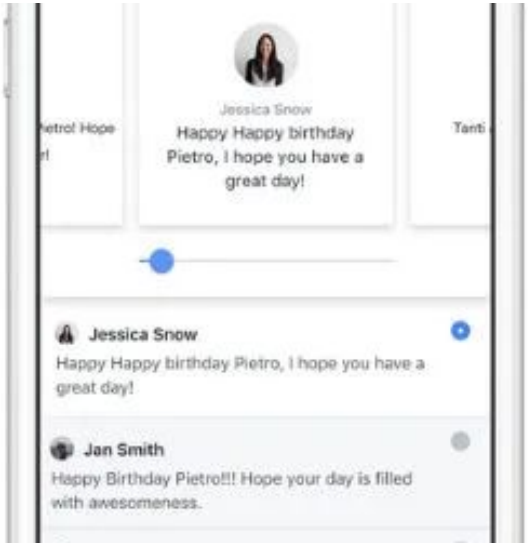
Costly Signal

Message that requires sacrifice of money/time/energy, difficult or impossible to fake



Cheap Signal

Message that costs little to nothing to send, may not reflect reality, easy to fake



Costly Signal

Message that requires sacrifice of money/time/energy, difficult or impossible to fake



Jennifer Kim et al. "Labor Saving and Labor Making of Value in Online Congratulatory Messages"

Cheap Signal

Message that costs little to nothing to send, may not reflect reality, easy to fake

CTR (clickthrough rate)

Tech & Telecom

Golf Equipment Retailer Sues Competitor for Ad-Click Fraud

April 13, 2020, 11:50 AM PDT



Peter Hayes
Reporter



A golfer tees off at a course.
Photographer: Frederic J. Brown/AFP/Getty Images

- **COURT**: D. Nev.
- **TRACK DOCKET**: No. 20-cv-00674 (Bloomberg Law Subscription)

Motogolf alleges, that in response, the defendants conspired to harm Motogolf. Id. The defendants, aware of how PPC ads work, used one or more electronic devices in various locations in or near Maine to find Motogolf's PPC ads and click on them repeatedly. Id. at 9, 13. This was done with the intent of maxing out the number of clicks Motogolf contracted for so that the ads would disappear for other viewers and Motogolf's future ad costs would increase. Id. at 9. Motogolf alleges that the defendants engaged in this conduct to intentionally gain an economic advantage over Motogolf. Id. at 6-8, 10.

Around March 7, 2019, Motogolf sent a cease-and-desist letter to Ivan Sokolovich informing him that the defendants were no longer authorized to click on Motogolf's PPC ads under any circumstances. Id. at 19. Motogolf sent a similar letter on August 13. Id. at 20. The defendants continued to click on the ads after receiving these letters. Id. at 20-21. Since the defendants began the click activity, they have exhausted Motogolf's PPC ads on at least a weekly basis. Id. at 14. This has prevented prospective customers from seeing the PPC ads and has increased the cost for Motogolf to purchase PPC ads. Id. at 15. It has also deprived Motogolf of the benefits from its PPC ads, such as useful demographic data. Id. at 23. As a result, Motogolf's share in the online golf equipment market diminished and Top Shelf's share increased. Id. at 17-19.

Cheap Signal

Message that costs little to nothing to send, may not reflect reality, easy to fake

CTR (clickthrough rate)

Costly Signal

Message that requires sacrifice of money/time/energy, difficult or impossible to fake

what might advertisers prefer?

Conversion

- **conversion:** when a user completes a desired action after seeing or clicking an ad
- examples: buying a product, signing up for a newsletter, downloading an app

Baskets

- **basket** (or shopping or market basket): term used in marketing and retail to refer to the group of products a single customer buys during one transaction
- **Average Basket Size (ABS):** number of items or products purchased in a single order, averaged over a single user or all users
- **Average Basket Value (ABV)** or Average Order Value (AOV): the total money spent in a single order, averaged over a single user or all users

Customer Lifetime Value

- **customer lifetime value (CLV or LTV):** predicted total revenue from a customer during their whole relationship with a company

$$\text{CLV} = \begin{matrix}\text{average}\\ \text{purchase}\\ \text{value}\end{matrix} \times \begin{matrix}\text{purchase}\\ \text{frequency}\end{matrix} \times \begin{matrix}\text{average}\\ \text{customer}\\ \text{"lifetime"}\end{matrix}$$



Jamie loves following sneaker trends. Recently she has been shopping at Badidas Sneakers, buying shoes every other month for the last three years. She typically purchases a single pair of shoes per visit (average purchase value of $78). Based on your knowledge of similar users, you expect she will continue to purchase Badidas for the next 2 years. Badidas' profit margin on sneakers is 40%.

how much should you be willing to spend on ads for this customer?



$$CLV = \text{average purchase value} \times \text{purchase frequency} \times \text{average customer "lifetime"}$$

$$CLV = (\$78 \times 0.40) \times 6 \times 2$$

$$CLV = \$374.4$$

- **rule of thumb:** most companies aim to spend no more than ⅓ of the CLV to acquire a customer

  Ad spend = $374/3 = $124

Break Time + Attendance



https://tinyurl.com/webtracking-attendance

super secret

Codeword:

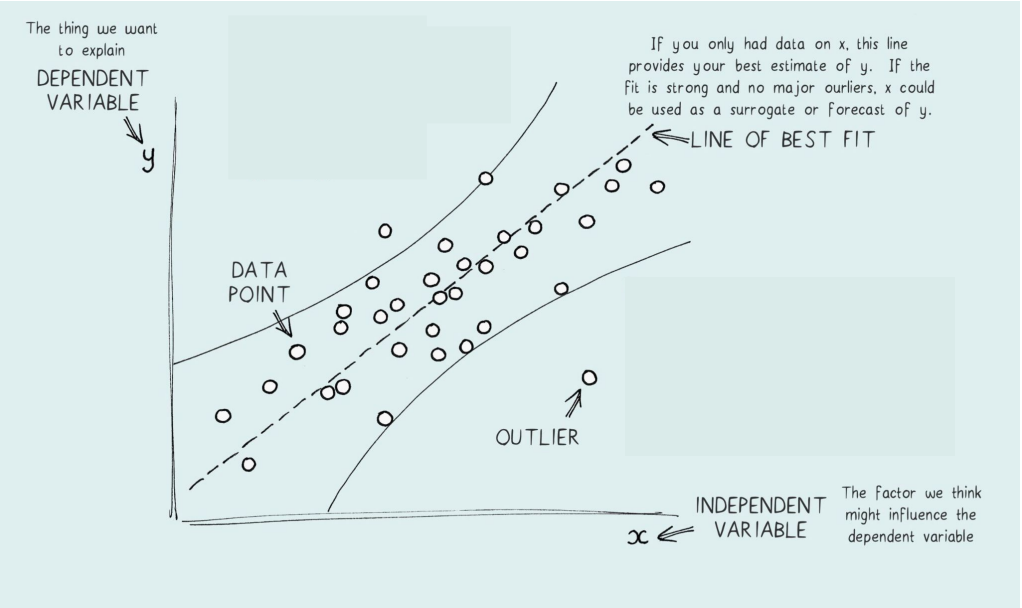
lifetimevalue?

Making Predictions

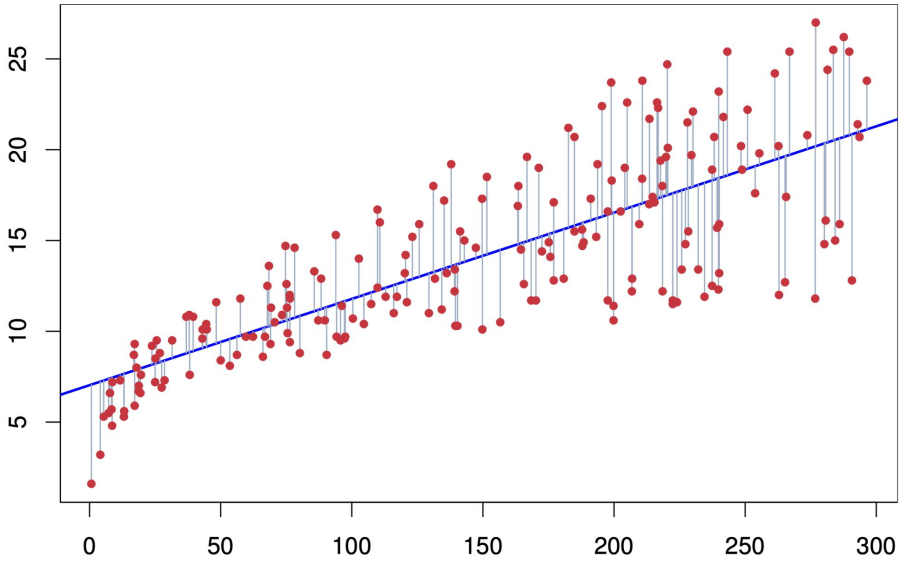
- **supervised learning:** predict a label for which you have at least some ground truth
- **unsupervised learning:** predict something for which you have no ground truth

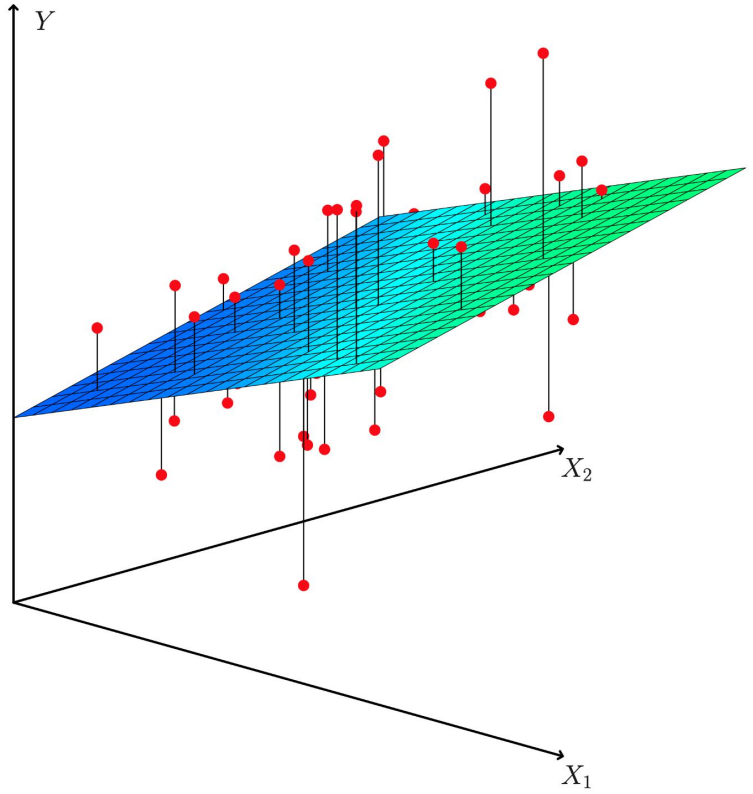
Supervised Learning

- predict a label for which you have at least some ground truth
- simplest* model: linear | logistic regression
- uses linear formula
  $y = mx + b$
- **residual:** distance from each data point to line
- start with random m, b
- iteratively update to minimize squared residuals



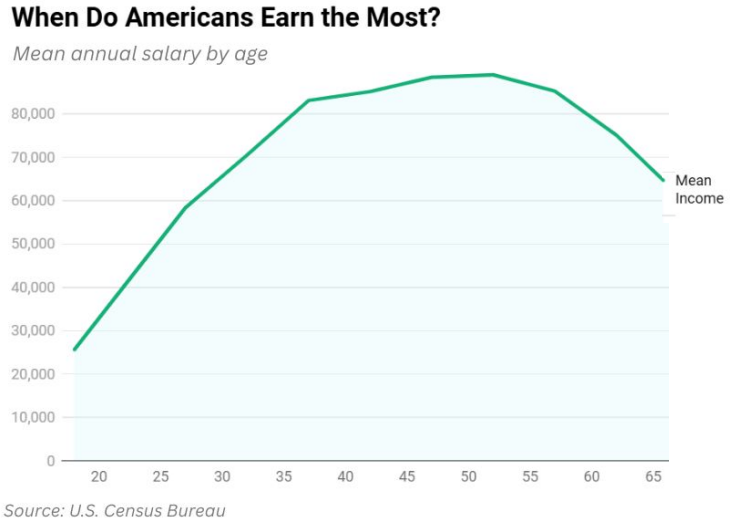
*yes, people will disagree/argue about this, ex: Naive Bayes

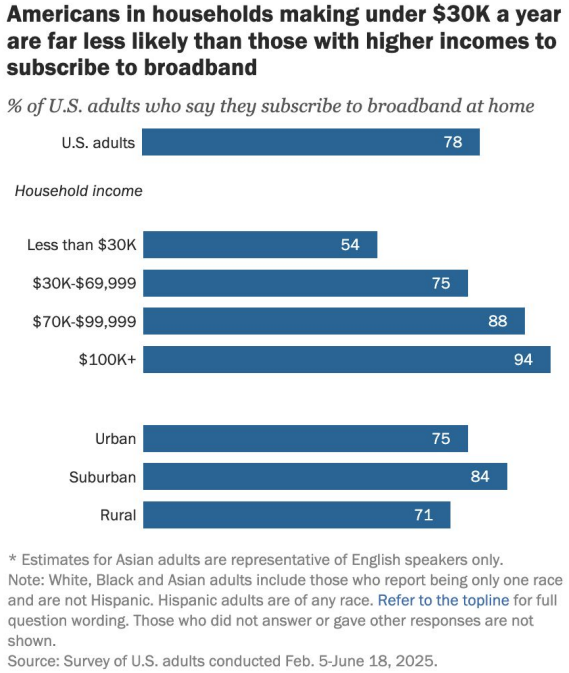

25
20
15
10
5
0
50
100
150
200
250
300


Y
X_2
X_1

Ex: Infer Unknown Attributes Like Income

- predict a label for which you have at least some ground truth
- for example: given age and internet connection speed can you predict income?



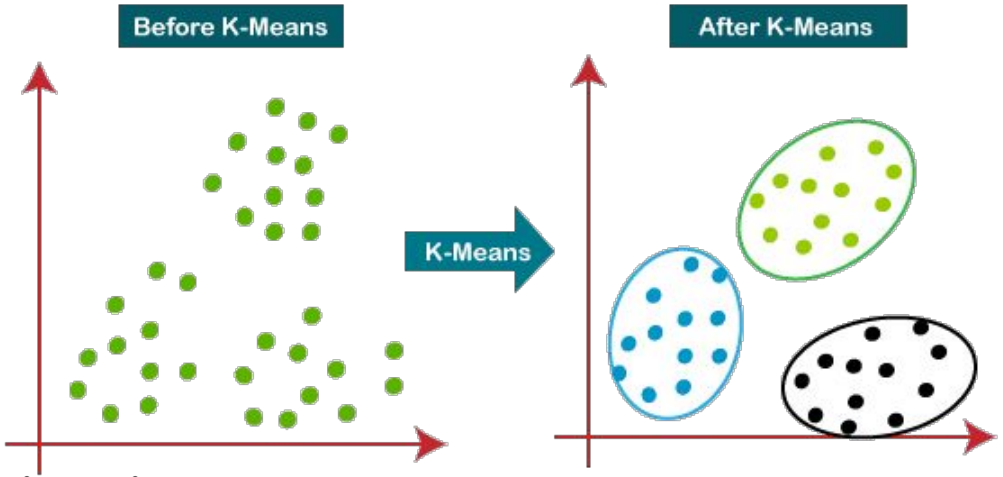




https://www.online-python.com/share/Kz41bMSsiR

Unsupervised Learning

- predict something for which you have no ground truth
- simplest* model: clustering
- groups unlabeled data into a fixed number (k) of distinct, non-overlapping clusters
- **centroid:** center of each group
- select *k* random points as centroids
- measure distances to nearest centroid
- iteratively update to minimize distance



*yes, people will disagree/argue about this, ex: Principal Component Analysis (PCA)

Ex: Infer Lookalike Audiences

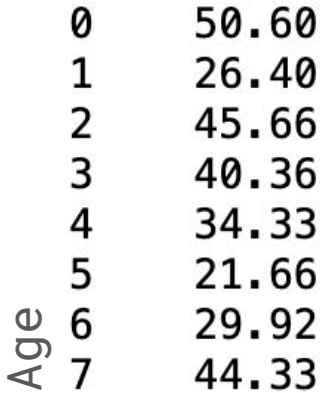
- **lookalike audience:** new people who share similar traits, interests, and behaviors with existing customers or high-value audience
- predict which customers are most similar to one another
- for example: given age, income, and interest category, can you identify similar shoppers?

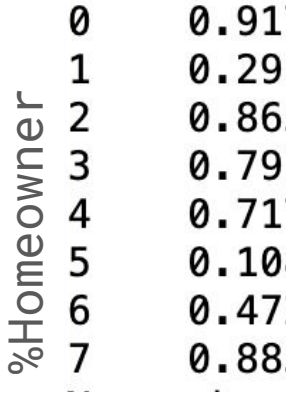


https://www.online-python.com/share/SChxdicPnu

Cluster 1 (Young Tech Enthusiasts)

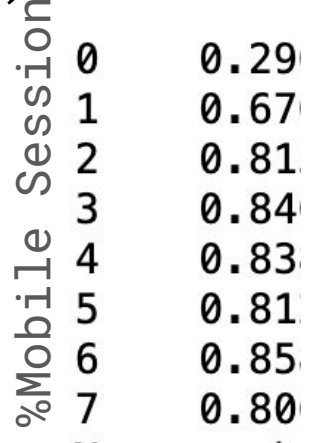
- **Demographics:** Gen Z and young adults (18–26), early-career income ($32k–$58k), renters, balanced desktop and mobile user
- **Interests:** Consumer electronics, PC hardware, emerging tech, gadgets





Cluster 2 (Suburban Homeowners & DIYers)

- **Demographics:** Older Gen X and Boomers (42–60), established income ($105k– $145k), high homeownership rate, desktop-heavy
- **Interests:** Home improvement, power tools, gardening and landscaping, home cooking



How Netflix Comes Up With Its Oddly Specific Category Names

MADELINE BUXTON

MARCH 16, 2018, 1:25 PM

Those category names barely scratch the surface of Netflix's apparent obsession with adjectives. On Twitter, users have called attention to weirder categories, including "Visually-Striking Cerebral Experimental Movies," "Gory Movies Featuring A Strong Female Lead," and "Critically Acclaimed Dysfunctional Family Dramas." Streaming fans have taken to making up names for their own desired groupings:

How Netflix Comes Up With Its Oddly Specific Category Names

The algorithm is only part of the process, though. After it identifies taste clusters, the taggers come in. Taggers are employees with the dream job of watching TV for a living. These entertainment experts name each taste cluster by finding the common descriptive thread that runs through every movie or TV show in that category. Sometimes, this is simple ("Crime TV Shows"), but other times, it requires a more nuanced take ("Visually-Striking Cerebral Experimental Movies"), especially as Netflix's library of original shows and number of taste clusters continue to expand. There were 1,300 groupings as of May 2017, and that's one of the reasons why the names have become so specific.

FAST COMPANY

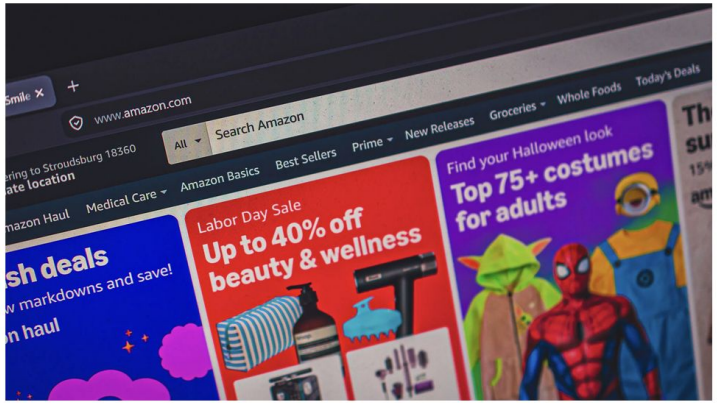
SUBSCRIBE

10-06-2026 | NEWS

‘Has flat buttocks,’ ‘cusses when scared’: Amazon keeps shockingly specific customer profiles. Here’s how to find yours

Amazon’s new ‘About you’ page lists many assumptions about shoppers—like their favorite brands, their hobbies, and the shape of their bodies.

SHARE ADD ON GOOGLE



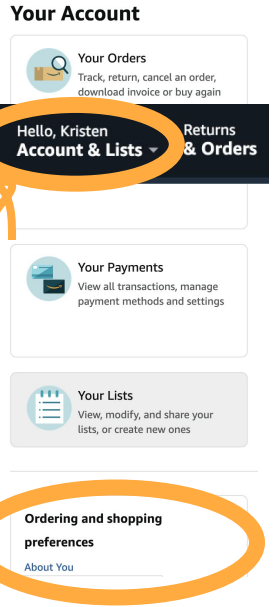
[Photo: k.dei/Adobe Stock]

BY JUDE CRAMER

Browser



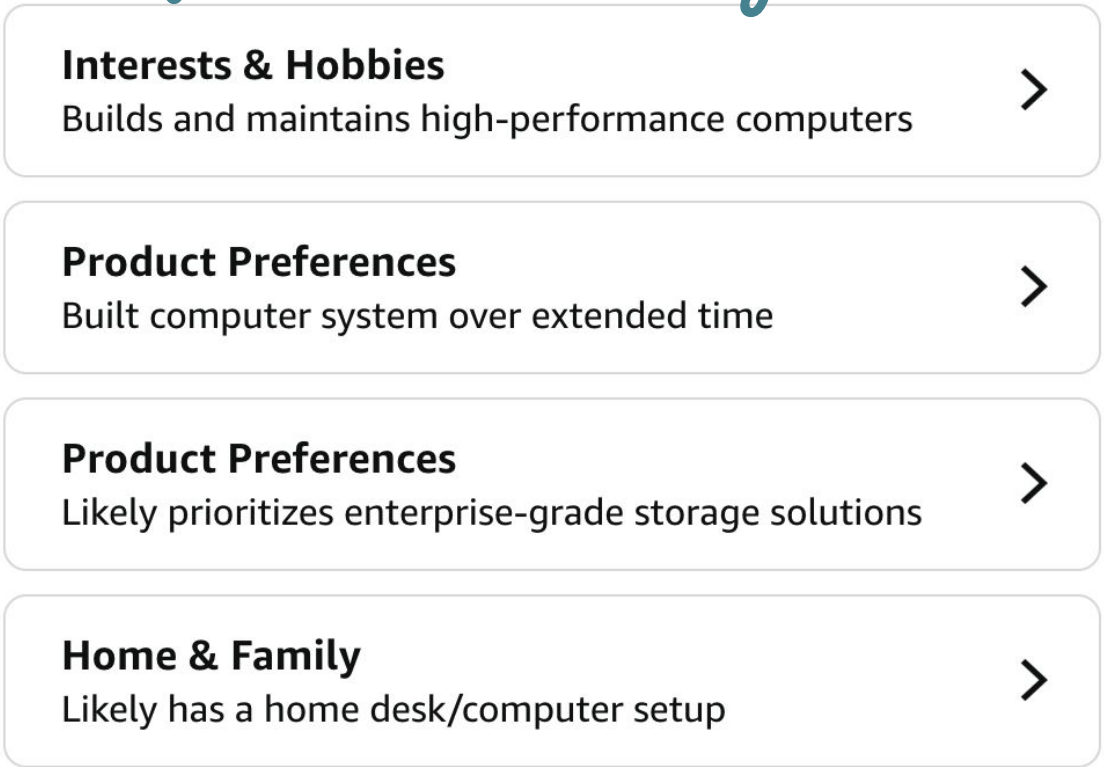
1. Amazon.com
2. Account & Lists
3. Select Account
4. > Ordering and shopping preferences
5. > About You



App

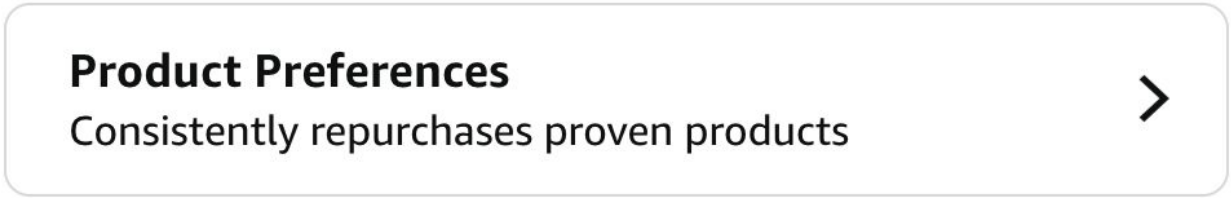
1. Open app
2. Click person icon on bottom nav bar
3. Tap your name
4. > About You

what features would you use to learn this?



how could you use these for marketing?

what features would you use to learn this?



how could you use these for marketing?

Evaluating Performance

- see how well the model will generalize using unseen data
- **train-test split:** policy of dividing any data into distinct training and testing (or held-out) subsets
  generally save 20-30% of your data for testing
- never touch your test data during training (until you're ready to deploy)

Evaluating Performance

- see how well the model will generalize using unseen data
- **train-test split:** policy of dividing any data into distinct training and testing (or held-out) subsets
  generally save 20-30% of your data for testing
- never touch your test data during training (until you're ready to deploy)
- many possible measures of performance

what's the accuracy of this model?

- UCSD Health recently licensed a model that predicts whether a mammogram indicates breast cancer
- The technicians are suspicious because it has **never** predicted cancer
- 1000 patients were tested, 10 later had cancer diagnosed by a doctor

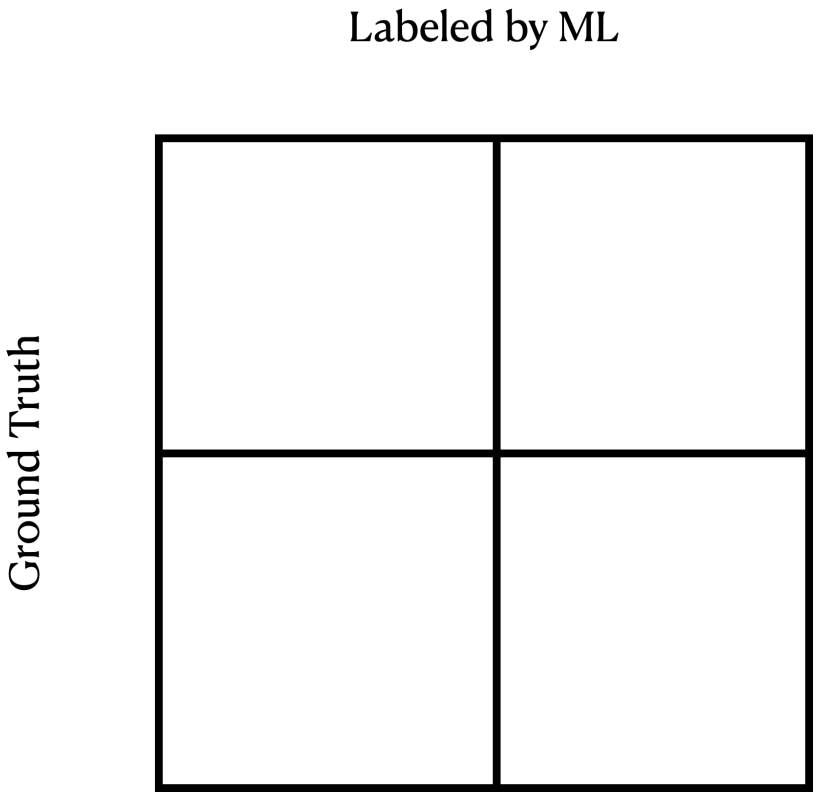
$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}}$$

The Problem with Accuracy (aka the 99% Trap)

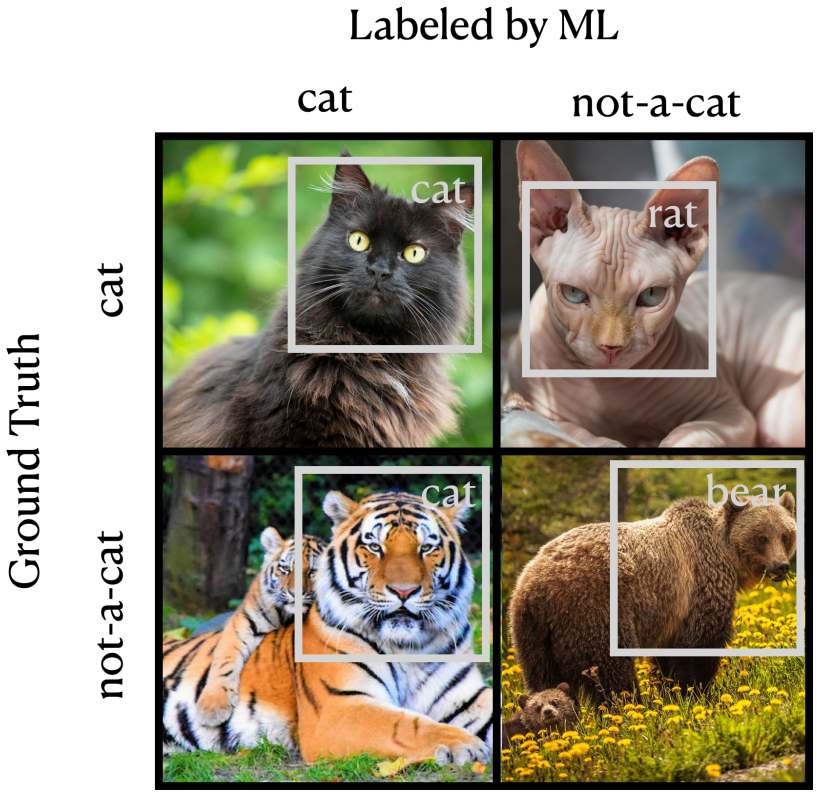
- UCSD Health recently licensed a model that predicts whether a mammogram indicates breast cancer
- The technicians are suspicious because it has **never** predicted cancer
- 1000 patients were tested, 10 later had cancer diagnosed by a doctor

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Samples}} = \frac{0 + 990}{1000} = 99\%$$

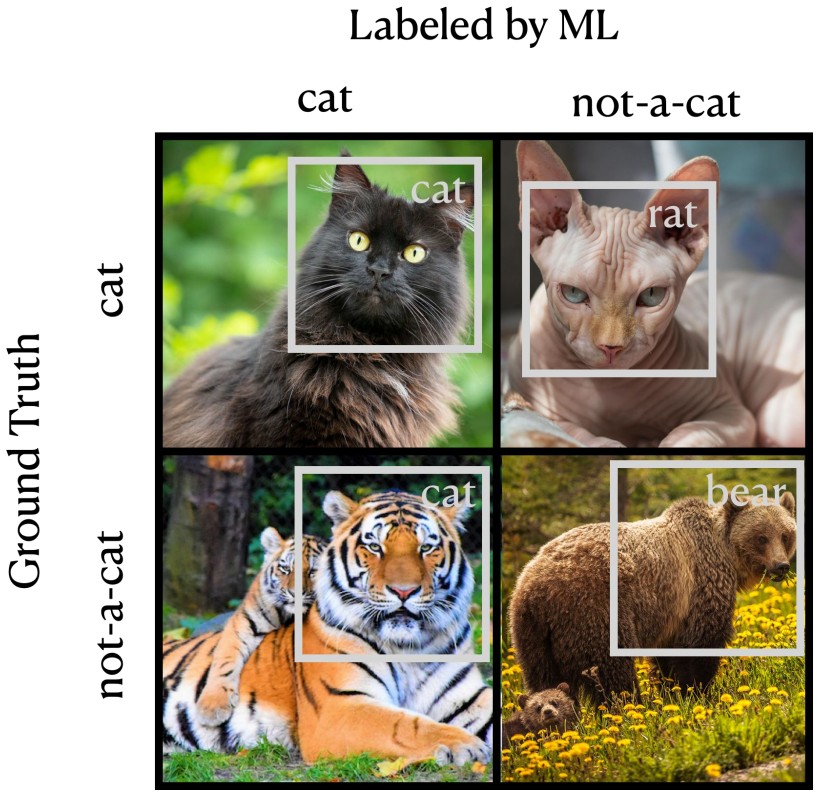
Confusion Matrix

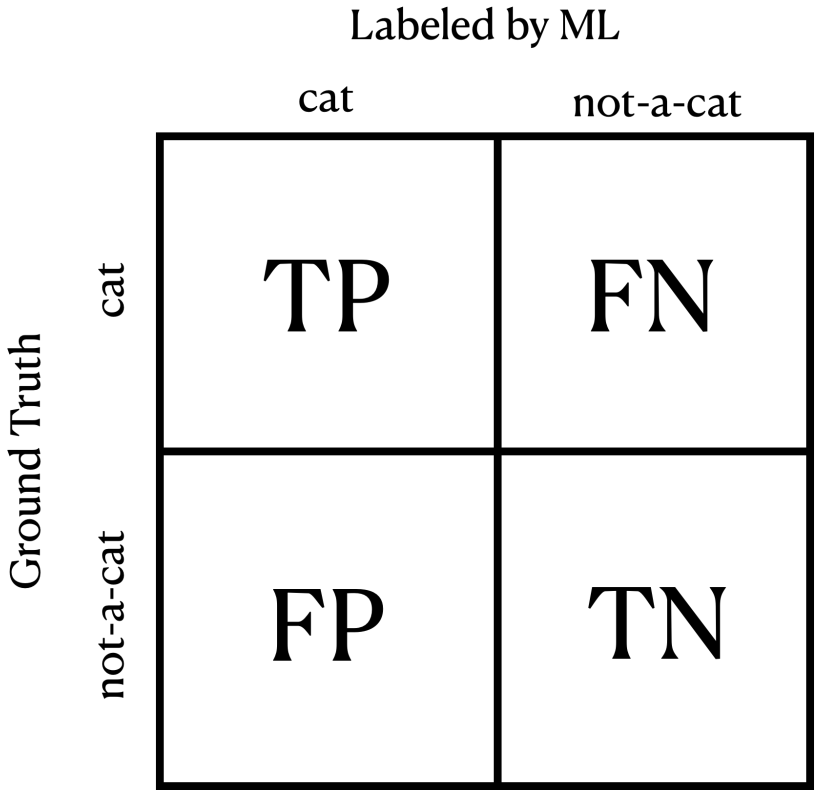


Confusion Matrix



Confusion Matrix





Evaluating Performance

- see how well the model will generalize using unseen data
- **train-test split:** policy of dividing any data into distinct training and testing (or held-out) subsets
  generally save 20-30% of your data for testing
- never touch your test data during training (until you're ready to deploy)
- many possible measures of performance: accuracy, precision, recall, F1, ROC-AUC, etc.
- **accuracy:** percentage of total predictions that are correct
  ⚠ tends to mislead if one prediction much more likely
  i.e., a health care model that always predicts `not-cancer`

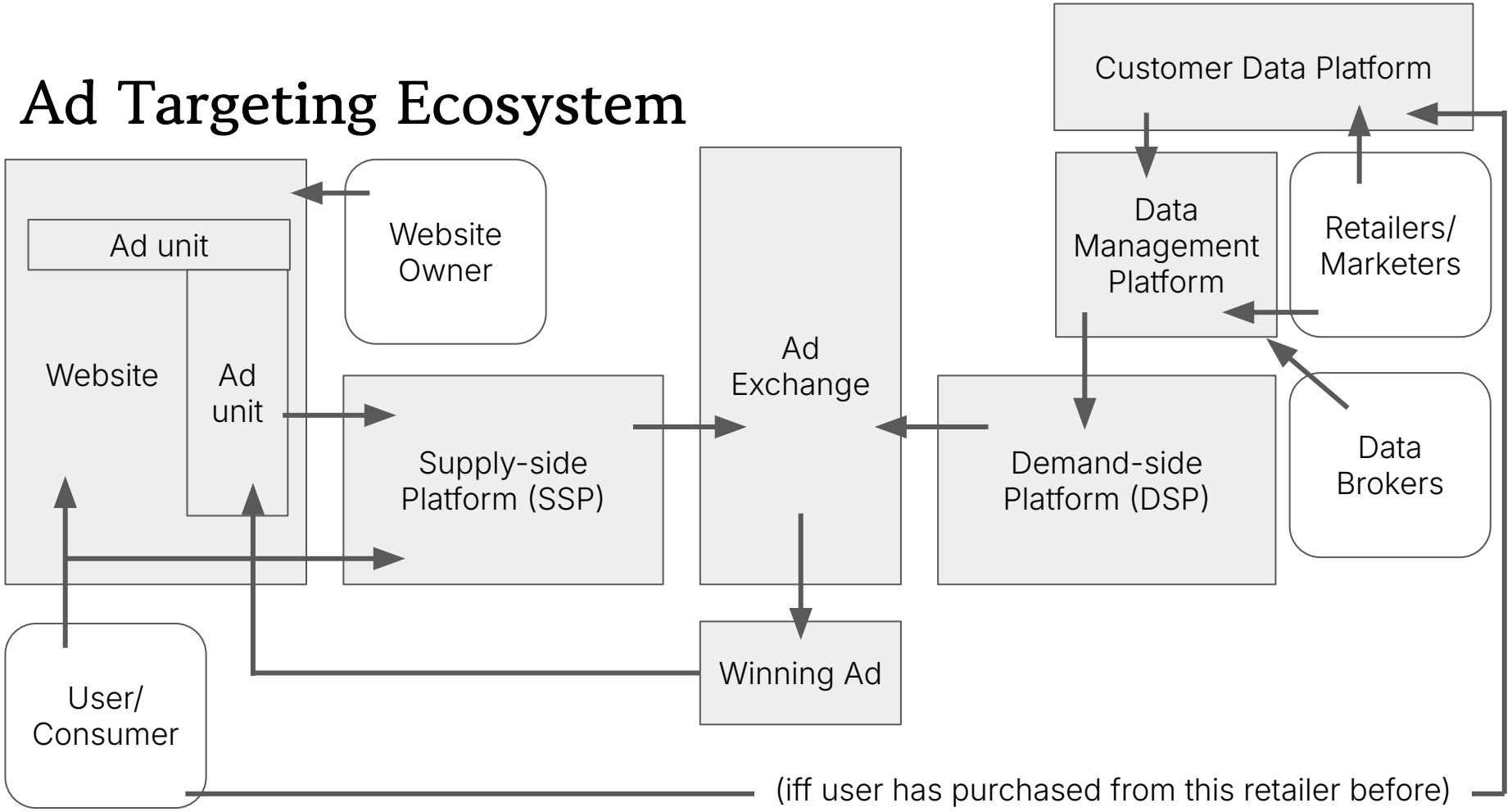
Evaluating Performance

- see how well the model will generalize using unseen data
- **train-test split:** policy of dividing any data into distinct training and testing (or held-out) subsets
  generally save 20-30% of your data for testing
- never touch your test data during training (until you're ready to deploy)
- many possible measures of performance: accuracy, precision, recall, F1, ROC-AUC, etc.
- **F1:** harmonic mean of precision and recall

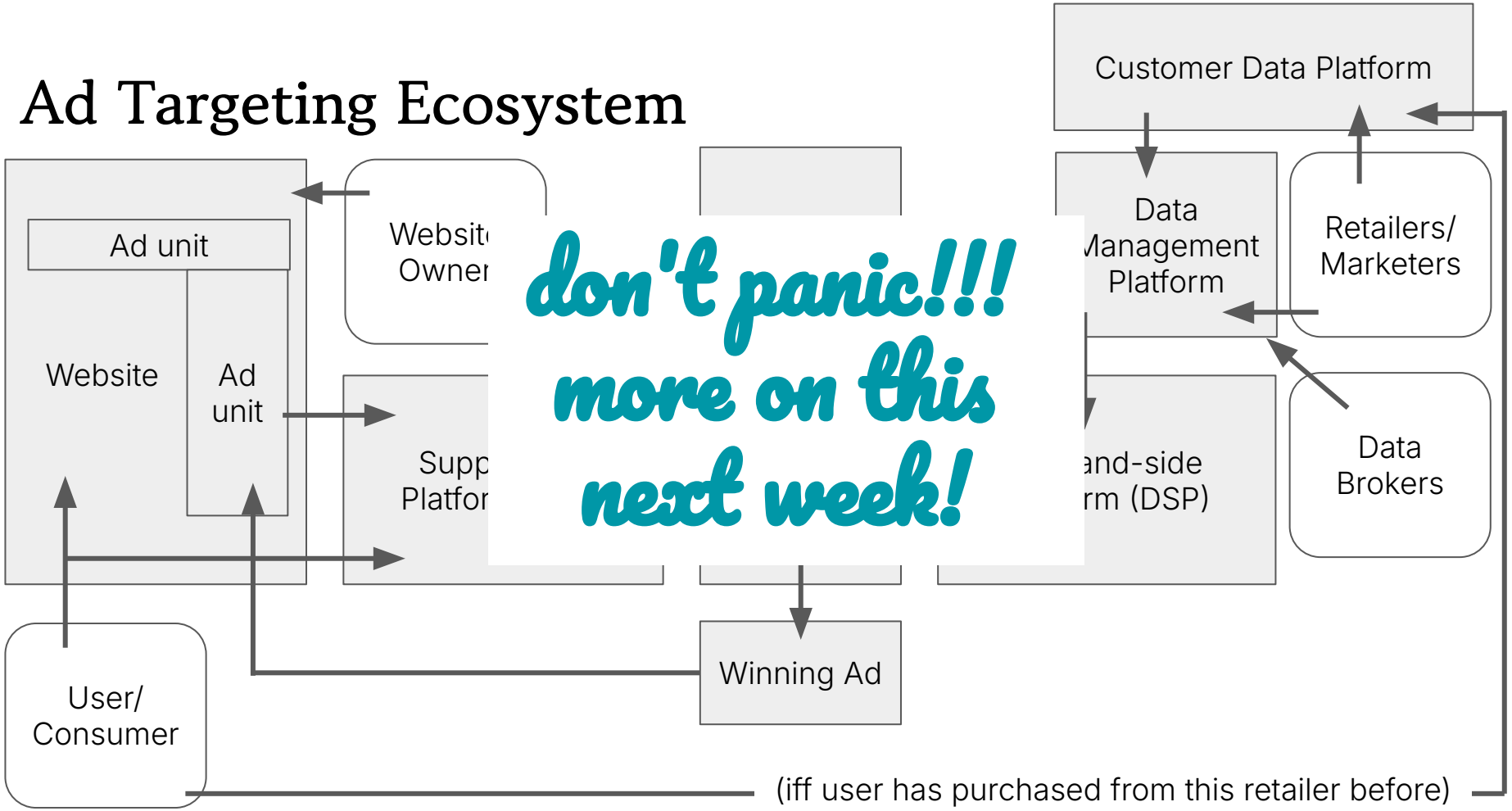
$$F1 = \frac{2TP}{2TP + FP + FN}$$

Questions?

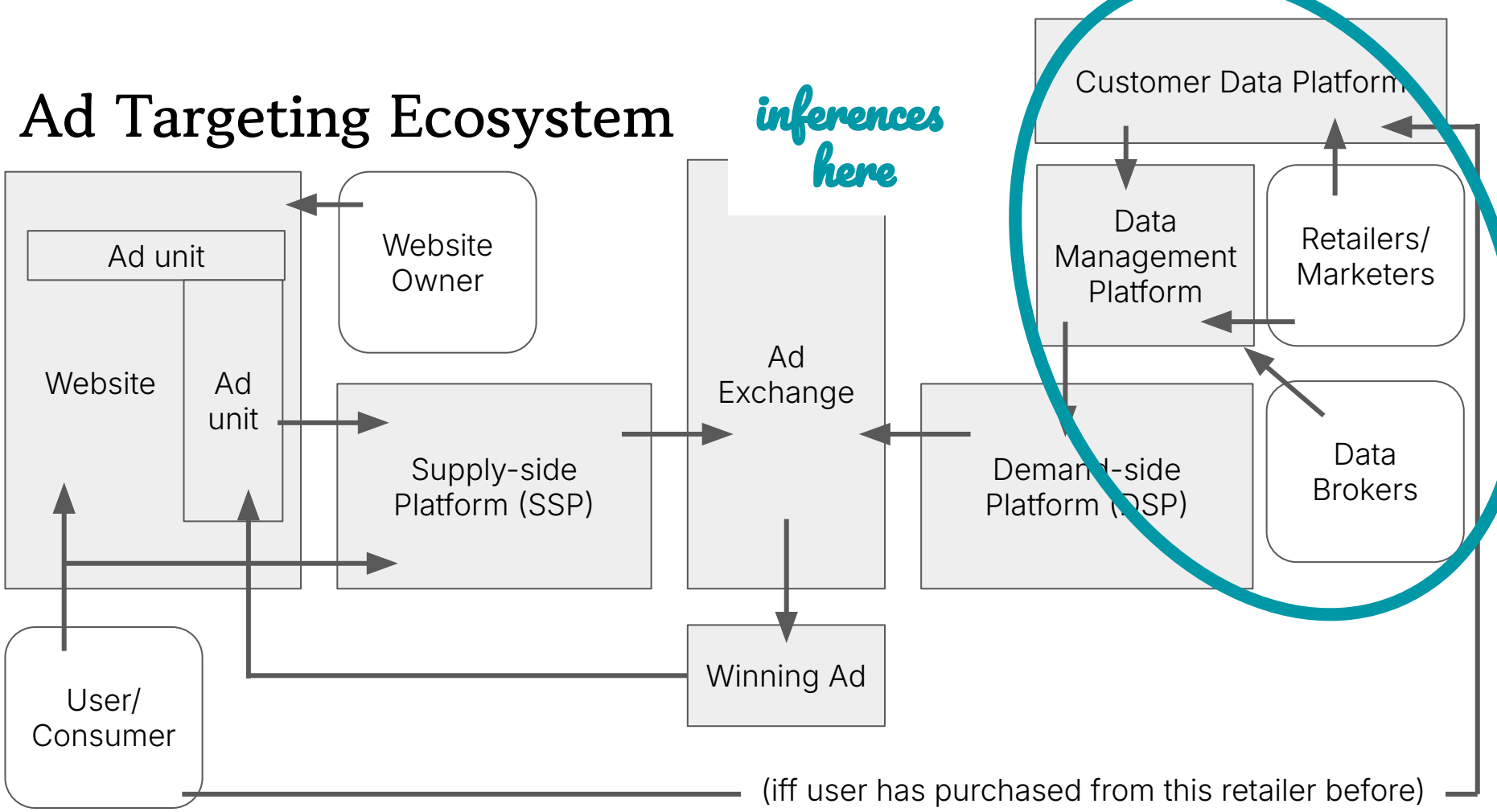
Ad Targeting Ecosystem



Ad Targeting Ecosystem



Ad Targeting Ecosystem



Exact Matching

- when an advertiser wants to target specific, known consumers
- already purchased or otherwise tracked (e.g., newsletter)
- still predict expected value to rank

Lookalike Audiences

- **most common approach**
- predict a new audience most similar to known high expected value customer

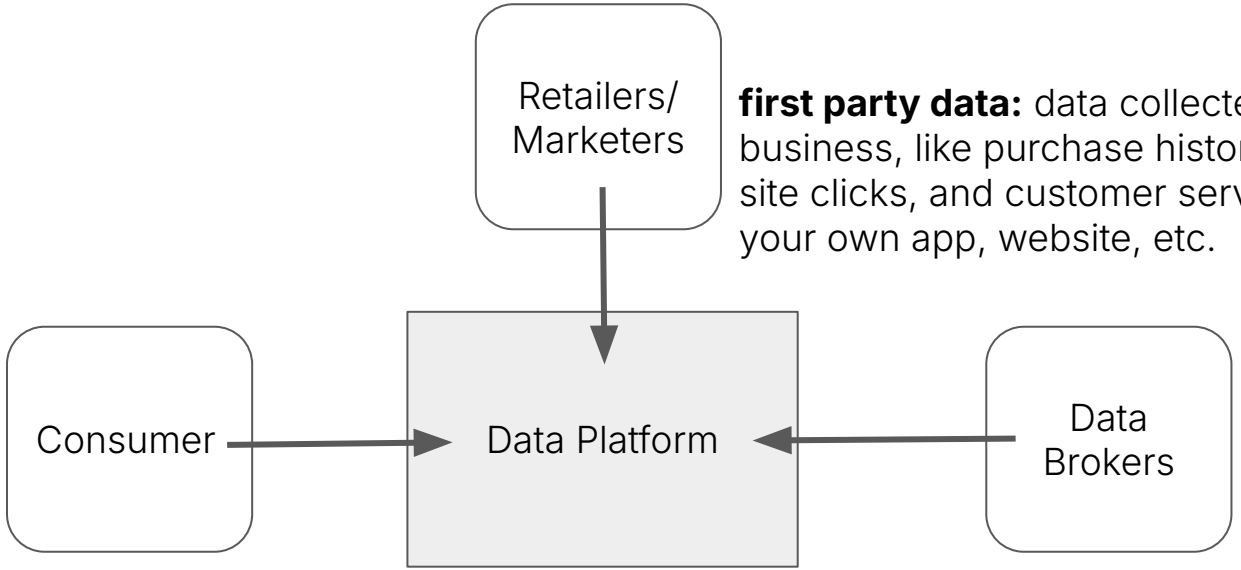
Customer Data Platform

- **most common approach**
- predict a new audience most similar to known high expected value customer

Parties

- **First party:** primary agent, in this case the retailer/marketer/business
- **Second party:** direct customer you work with
- **Third party:** any additional outside businesses or groups

Parties



first party data: data collected directly by the business, like purchase history of your products, site clicks, and customer service calls through your own app, website, etc.

zero party data: data a customer intentionally shares

third party data: data purchased from other businesses, e.g., lists of user interests or traits

Data Management Platform

- **most common approach**
- predict a new audience most similar to known high expected value customer

Other ML Applications

- creating ads

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: CTR, completed purchase, basket size, expected value
- **feature:** individual, measurable properties of input data that a model uses to learn and make predictions
- examples: specific demographic, behavior, device info w/in a user profile

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **features:** individual measurable property or characteristic of a data set, which will be used by the model to make predictions
  aka attributes, dimensions
- ideally, will be informative, discriminating, and independent
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.

Inference Basics

- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: a click on an ad, completed purchase, total "basket" size
- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns

Making Predictions

- **Infer unknown attributes (e.g., income from job category, zip code, device tier/connection quality, purchases in luxury categories/subscriptions/etc.)**
- **Directly predict total basket size etc.**
- **Lookalike audiences**

Making Predictions

- **training data**:

age, gender, education level, relationship status, major life events

what choose for input features?
what choose for output labels?

Making Predictions

- Suppose you choose:
  input: age
  output: average order value (AOV)

what choose for input features?
what choose for output labels?

Expected Value

- **expected value (EV)**:
- **customer lifetime value (CLV or LTV):** predicted total revenue from a customer during their whole relationship with a company

$$EV = \text{conversion probability} \times \text{user lifetime value}$$

Inference Basics

- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: a click on an ad, completed purchase, total "basket" size
- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns

what's an example in ad targeting?

Inference Basics

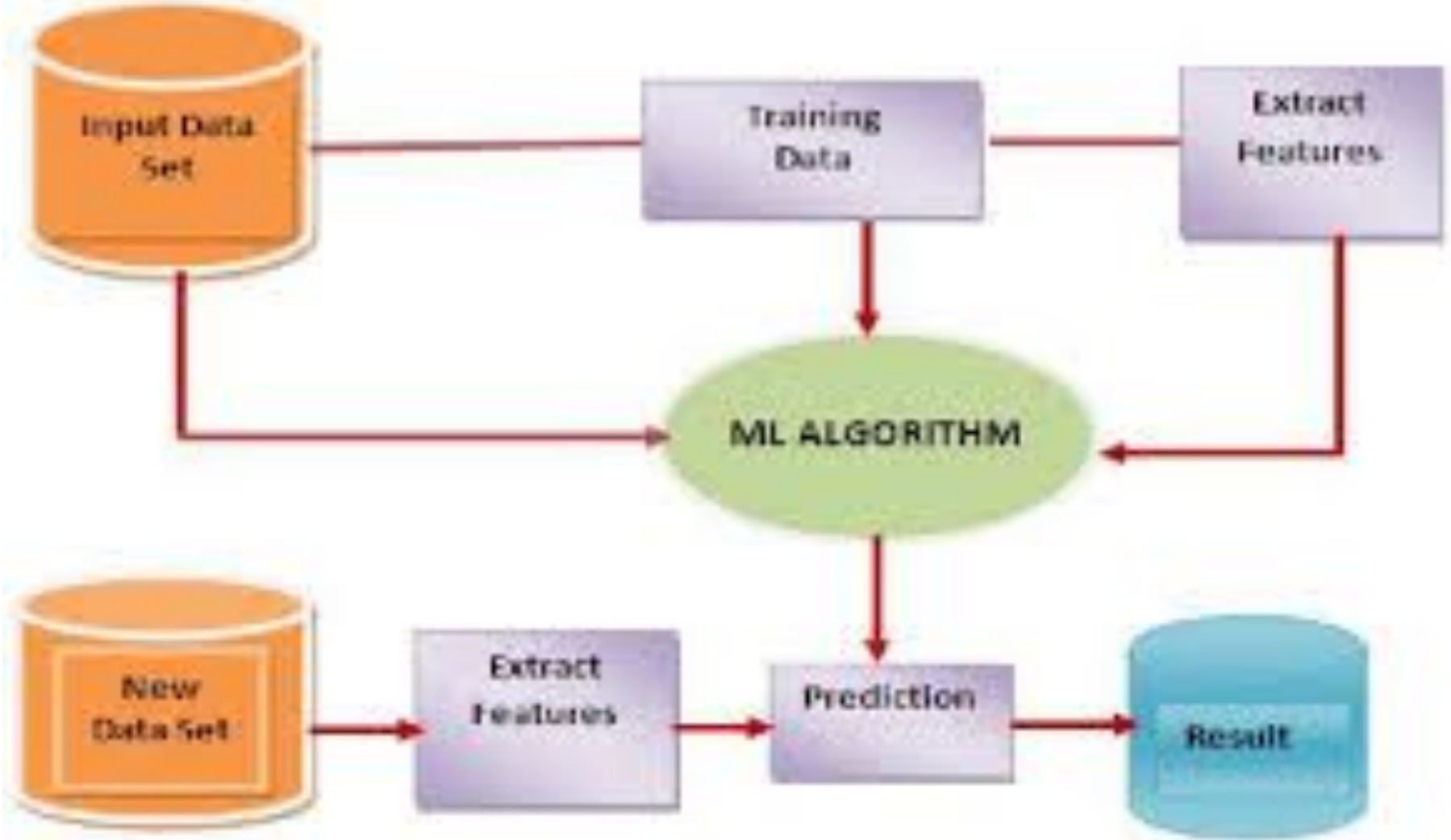
- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes both inputs and outputs (i.e., goal prediction)
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: a click on an ad, completed purchase, total "basket" size
- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- examples: data on past users along with ad clicks/purchases

How Netflix Comes Up With Its Oddly Specific Category Names

MADELINE BUXTON

MARCH 16, 2018, 1:25 PM

Those category names barely scratch the surface of Netflix's apparent obsession with adjectives. On Twitter, users have called attention to weirder categories, including "Visually-Striking Cerebral Experimental Movies," "Gory Movies Featuring A Strong Female Lead," and "Critically Acclaimed Dysfunctional Family Dramas." Streaming fans have taken to making up names for their own desired groupings:


Input Data Set
Training Data
Extract Features
ML ALGORITHM
New Data Set
Extract Features
Prediction
Result

User Profiles

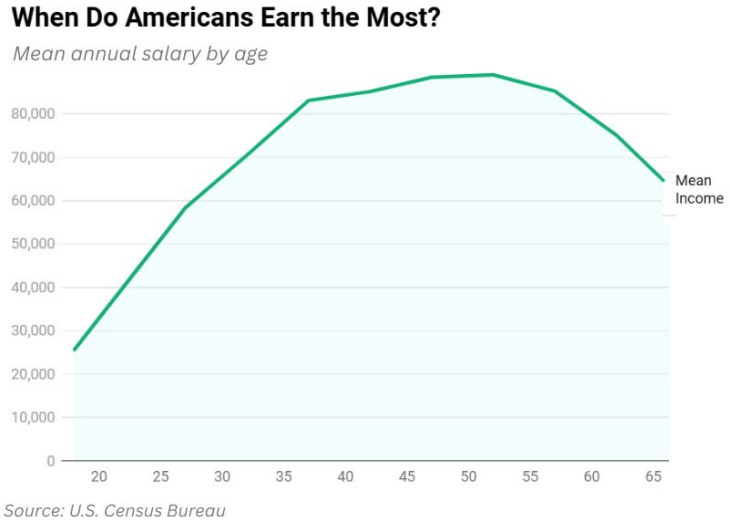
- **demographic data**:
  age: NUMERICAL [0, 1, 2, 3, 4, 5...]
  gender: CATEGORICAL [M, F, NB, ...]
  education level: ORDINAL [K, E1, E2, ... HS1, HS2, ... UG1, UG2 ... MS1]
  major life events: ???

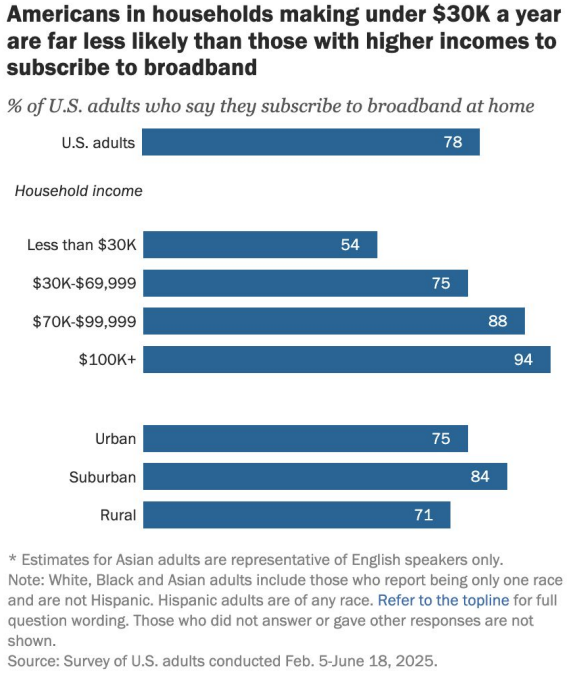
User Profiles

- **numerical:** information recorded as numbers, where there is both an ordering and consistent spacing
- **ordinal:** information where there is an order or rank, though the exact distance between ranks may not be equal
- **categorical:** information recorded as groups or labels, generally no sense of order, ranking, or mathematical meaning

Ex: Infer Unknown Attributes Like Income

- predict a label for which you have at least some ground truth
- for example: given age, device type, and past purchases can you predict income?





Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: CTR, completed purchase, basket size, expected value

Inference Basics

- **training data:** collection of information used as examples for a machine learning model to learn/recognize patterns
- often includes inputs and outputs (i.e., goal prediction) pairs
- **output label: "**answer" that a supervised model is trained to predict. also called the target variable.
- examples: CTR, completed purchase, basket size, expected value
- **feature:** individual, measurable properties of input data that a model uses to learn and make predictions

User Profiles

- **demographic data**:
  age: NUMERICAL
  gender: CATEGORICAL
  education level: ORDINAL
  major life events: CATEGORICAL?

Graduate HS
Graduate college
Married
Birth/adoption
Death in family
New job
Retire
Buy home
Major illness
....