Data Preprocessing

Data Mining & Analytics
INFO 254 / 154 - Spring ‘19

Prof. Zach Pardos

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

Preprocessing: Primer

Instance, rows, feature, attribute, column, target, label

	A	B	C
1	**Candidate**	**Interviewer**	**Rating**
2	Claudia S.	Manager A	4
3	Oliver R.	Manager A	2
4	Samuelson R.	Engineer A	5
5	Alicia M.	Engineer B	1
6	Oliver R.	Engineer B	5
7	Claudia S.	Manager B	3

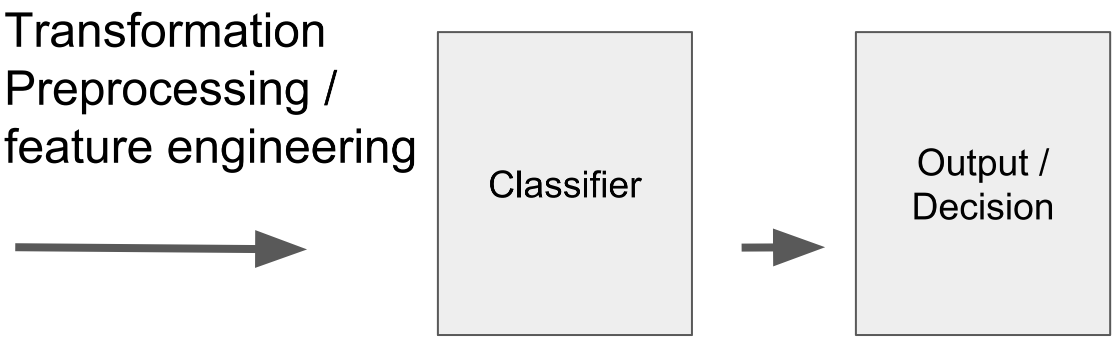
Task: choose a candidate to hire
Data: 50 candidates each interviewed by two employees

Preprocessing: Primer

Instance, rows, feature, attribute, column, target, label

Raw input data

	A	B	C
1	**Candidate**	**Interviewer**	**Rating**
2	Claudia S.	Manager A	4
3	Oliver R.	Manager A	2
4	Samuelson R.	Engineer A	5
5	Alicia M.	Engineer B	1
6	Oliver R.	Engineer B	5
7	Claudia S.	Manager B	3



Task: choose a candidate to hire
Data: 50 candidates each interviewed by two employees

Preprocessing: Primer

Instance, rows, feature, attribute, column, target, label

Theoretical Primer

Data Preprocessing as a form of:

Summarization, Kernelization, Representation Learning

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

```
{
    "votes": {
        "funny": 0,
        "useful": 2,
        "cool": 1
    },
    "user_id": "Xqd0DzHaiyRqVH3WRG7hzg",
    "review_id": "15SdjuK7DmYqUAj6rjGowg",
    "stars": 5,
    "date": "2007-05-17",
    "text": "dr. goldberg offers everything i look for in a general
 practitioner.  he's nice and easy to talk to without being patroni
zing; he's always on time in seeing his patients; he's affiliated w
ith a top-notch hospital (nyu) which my parents have explained to m
e is very important in case something happens and you need surgery;
 and you can get referrals to see specialists without having to see
 him first.  really, what more do you need?  i'm sitting here tryin
g to think of any complaints i have about him, but i'm really drawi
ng a blank.",
    "type": "review",
    "business_id": "vcNAWiLM4dR7D2nwwJ7nCA"
}
```

Preprocessing: Terminology

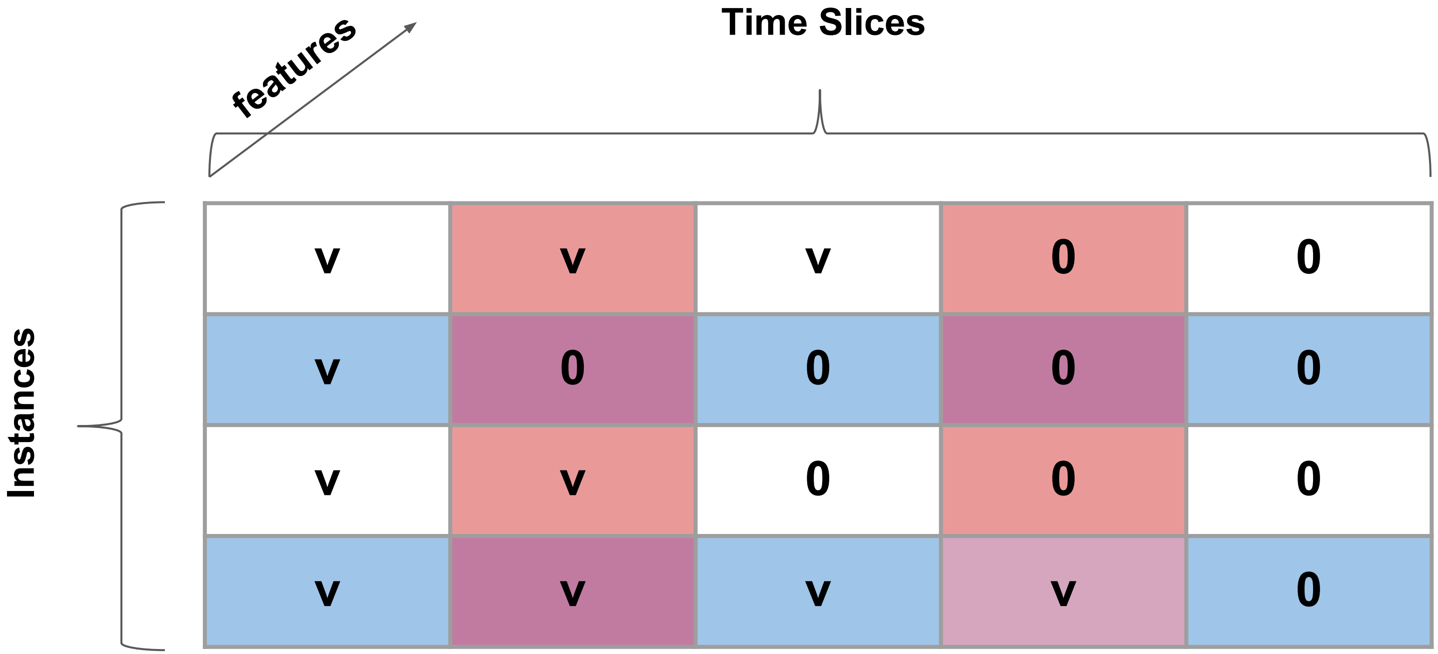
Instance, rows, feature, attribute, column, target, label

Most common form of representing data to a model

- Slight differences for: Time series, Networks

Preprocessing: Terminology

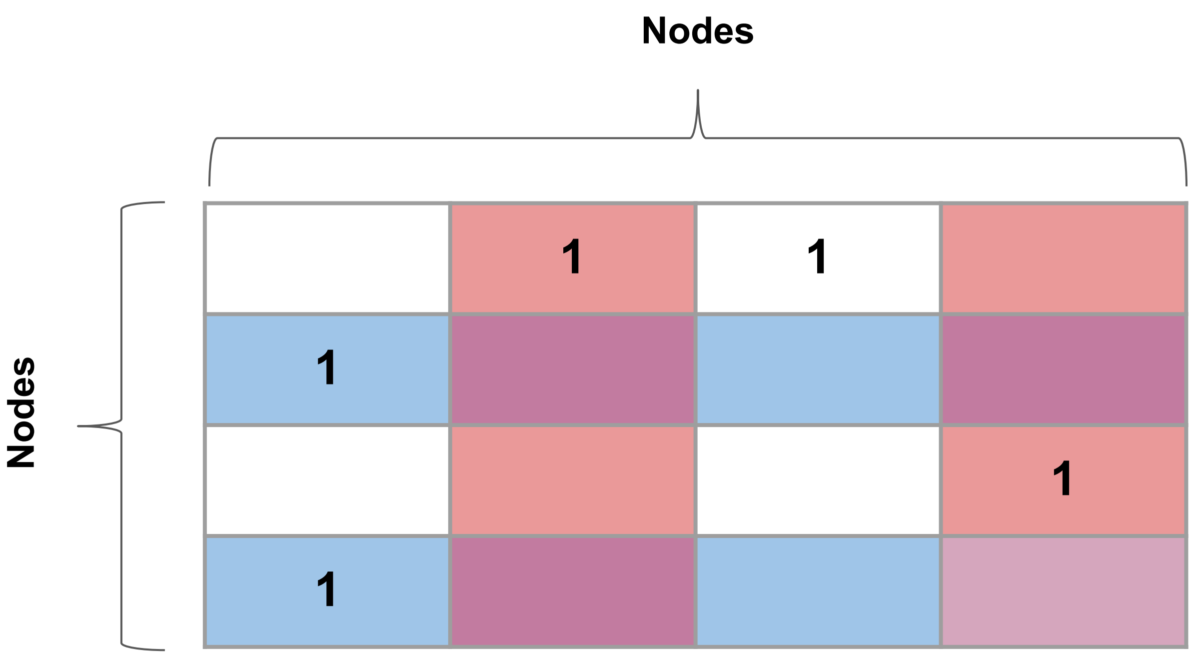
<u>Instance</u>, rows, <u>feature</u>, attribute, column, <u>target</u>, label



Time series

Preprocessing: Terminology

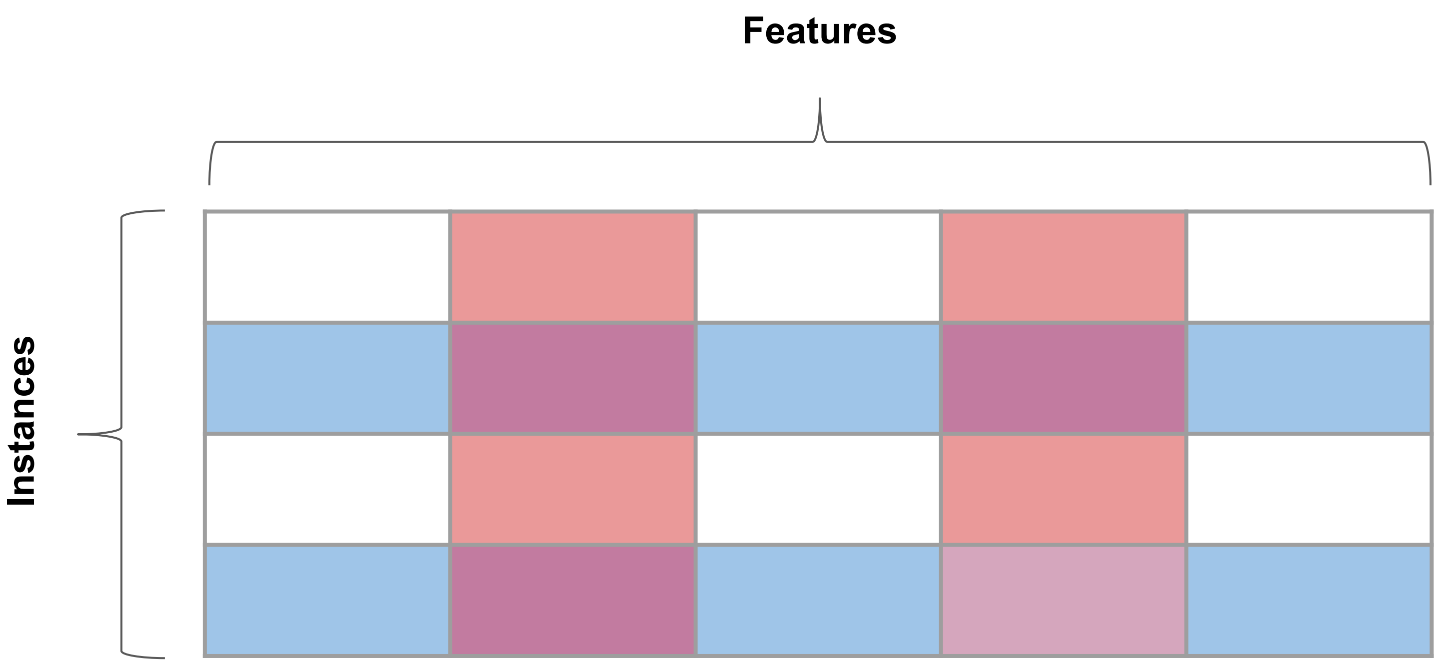
Instance, rows, feature, attribute, column, target, label



Networks (connections/relationships)

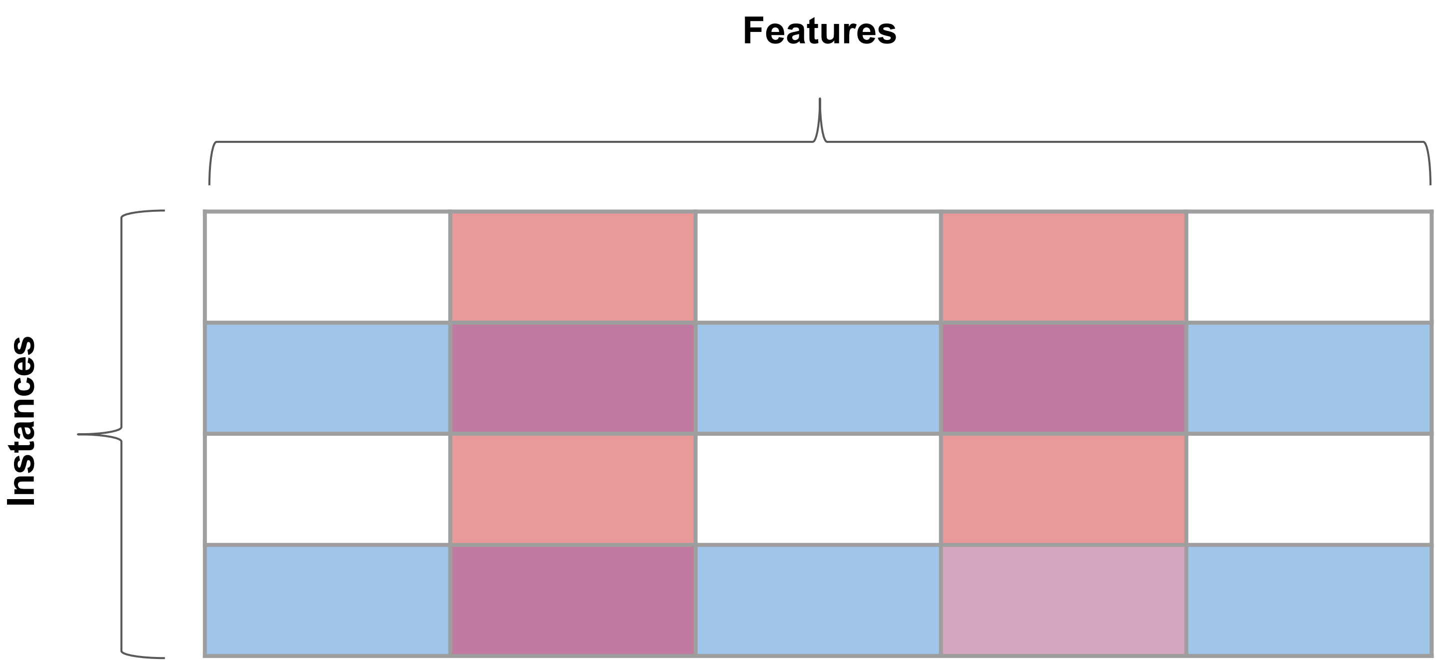
Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label



Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label



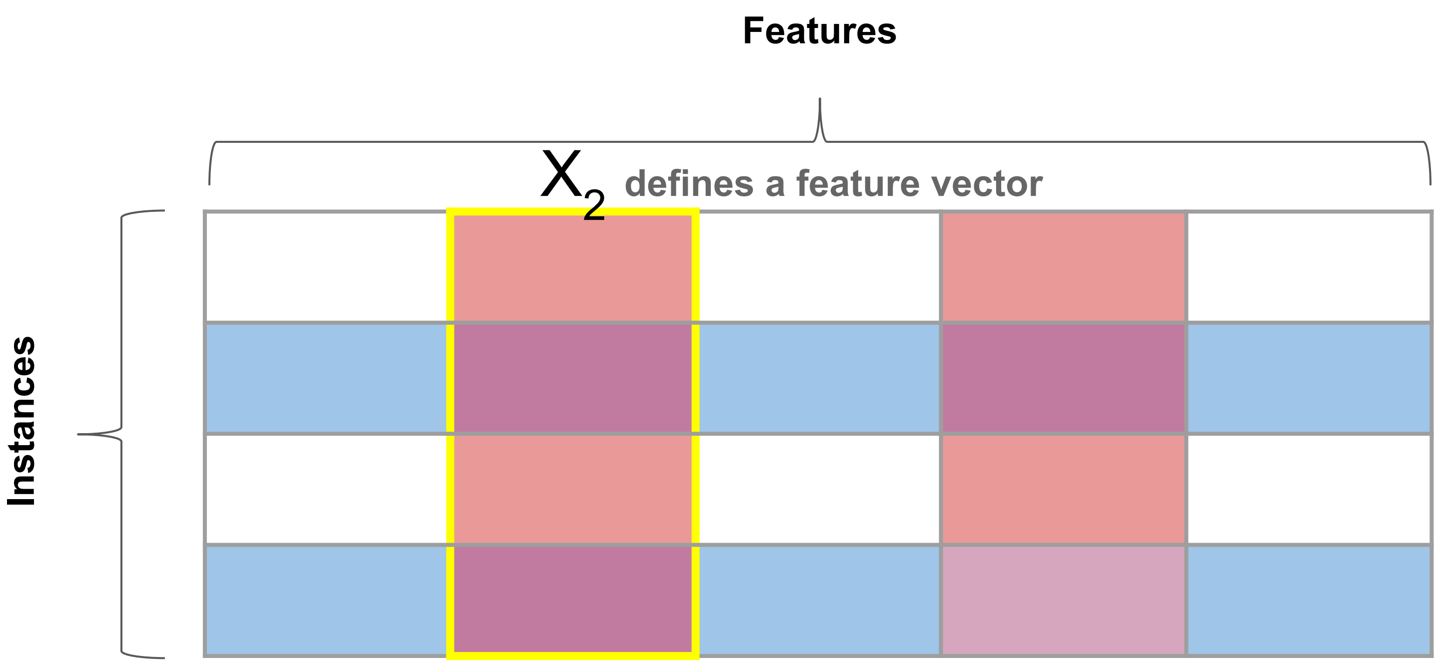
This dataset (matrix) can be expressed by: X_m^n

n # of instances

m # of features

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label



This dataset (matrix) can be expressed by: X_m^n # of instances # of features

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

This dataset (matrix) can be expressed by: X_m^n # of instances # of features

Preprocessing: Terminology

Instance, rows, feature, attribute, column, target, label

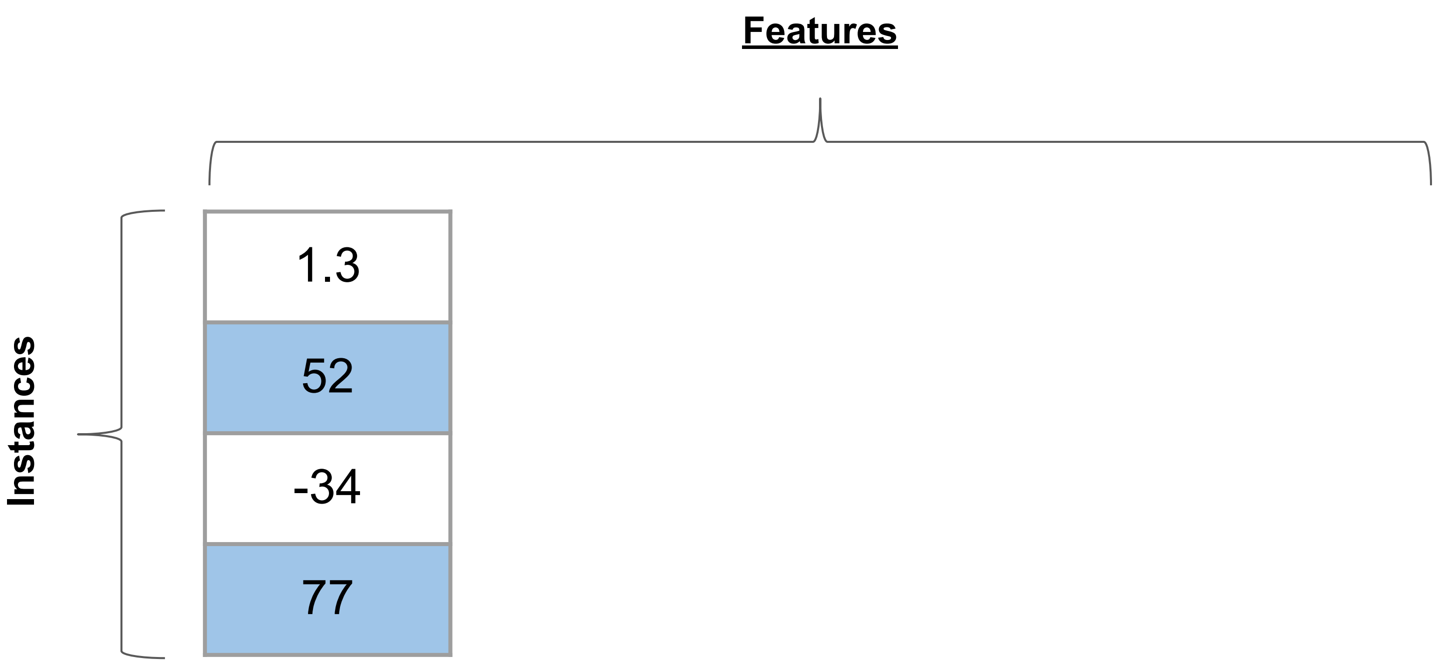
Where does the target coming from?

yelp_academic_dataset_review.json


```
{
    "review_id":"encrypted review id",
    "user_id":"encrypted user id",
    "business_id":"encrypted business id",
    "stars":star rating, rounded to half-stars,
    "date":"date formatted like 2009-12-19",
    "text":"review text",
    "useful":number of useful votes received,
    "funny":number of funny votes received,
    "cool": number of cool review votes received,
    "type": "review"
}
```

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal



What is the type is this feature?

How can it be represented to a classifier?

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

1.3	0.3180
52	0.7748
-34	0
77	1

$$\frac{X - \min(X)}{\max(X - \min(X))}$$

What is the type is this feature? *Numeric*

How can it be represented to a classifier? *As-is or normalized*

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

yes
no
no
yes

What is the type is this feature?

How can it be represented to a classifier?

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

yes	1
no	0
no	0
yes	1

What is the type is this feature? *Binary*

How can it be represented to a classifier? *0s and 1s*

Preprocessing: Features

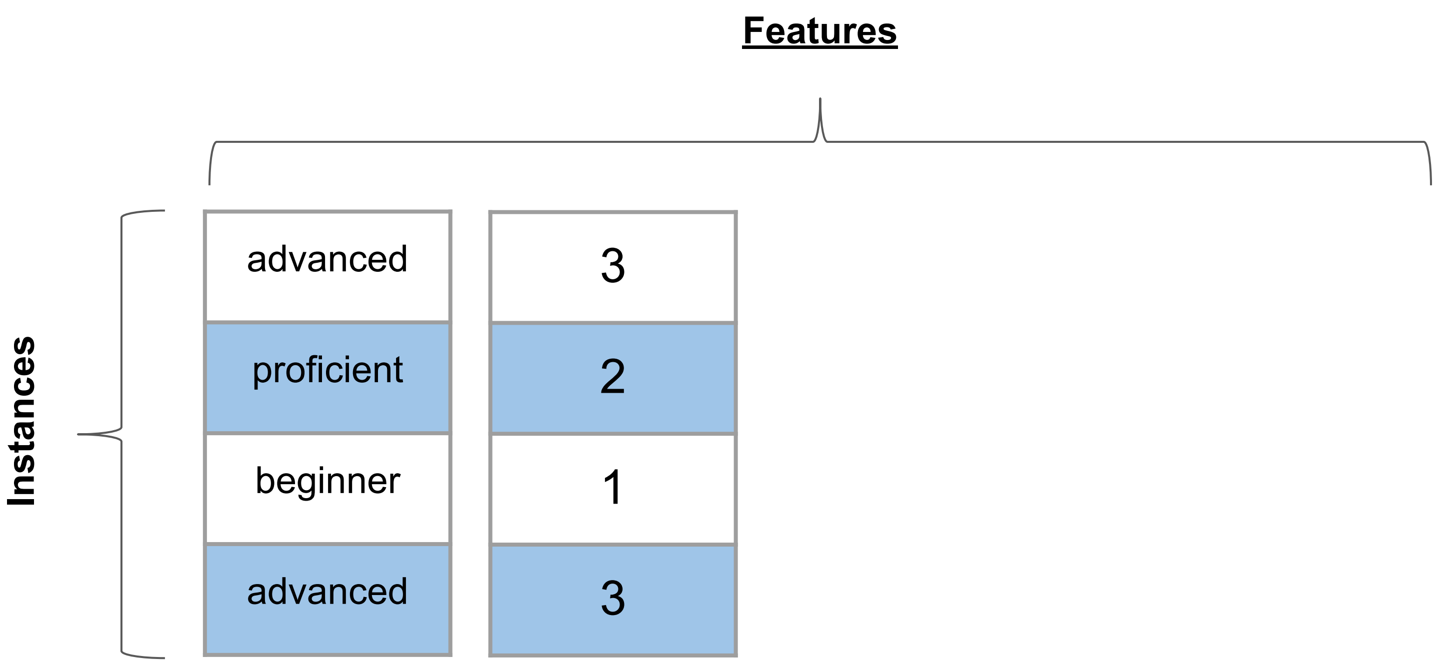
Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

What is the type is this feature?

How can it be represented to a classifier?

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal



What is the type is this feature? *Ordinal*

How can it be represented to a classifier? *Ordered numeric*

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

What is the type is this feature?

How can it be represented to a classifier?

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

labrador	1	0	0
poodle	0	1	0
beagle	0	0	1
poodle	0	1	0

What is the type is this feature? *Nominal*

How can it be represented to a classifier? *One-hot*

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

California
Colorado
Arizona
Florida

What is the type is this feature?

How can it be represented to a classifier?

Preprocessing: Features

Feature types: 1. Binary 2. Numeric 3. Ordinal 4. Nominal

Features

Instances

California
Colorado
Arizona
Florida

What is the type is this feature? *Nominal*

How can it be represented to a classifier? *One-hot, numeric proxy*

Preprocessing: Target

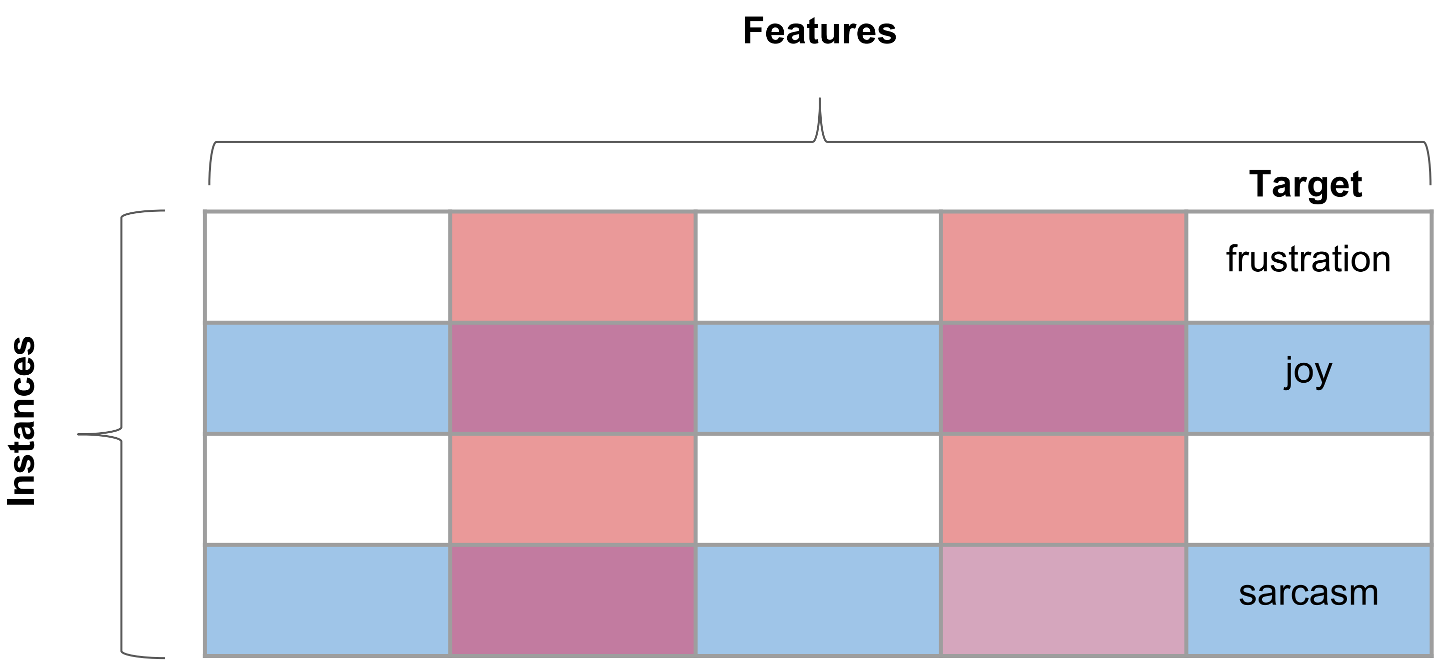
<u>Instance</u>, rows, <u>feature</u>, attribute, column, <u>target</u>, label

Where does the target coming from?

1) An existing feature that is missing from some instances

Preprocessing: Target

Instance, rows, feature, attribute, column, target, label

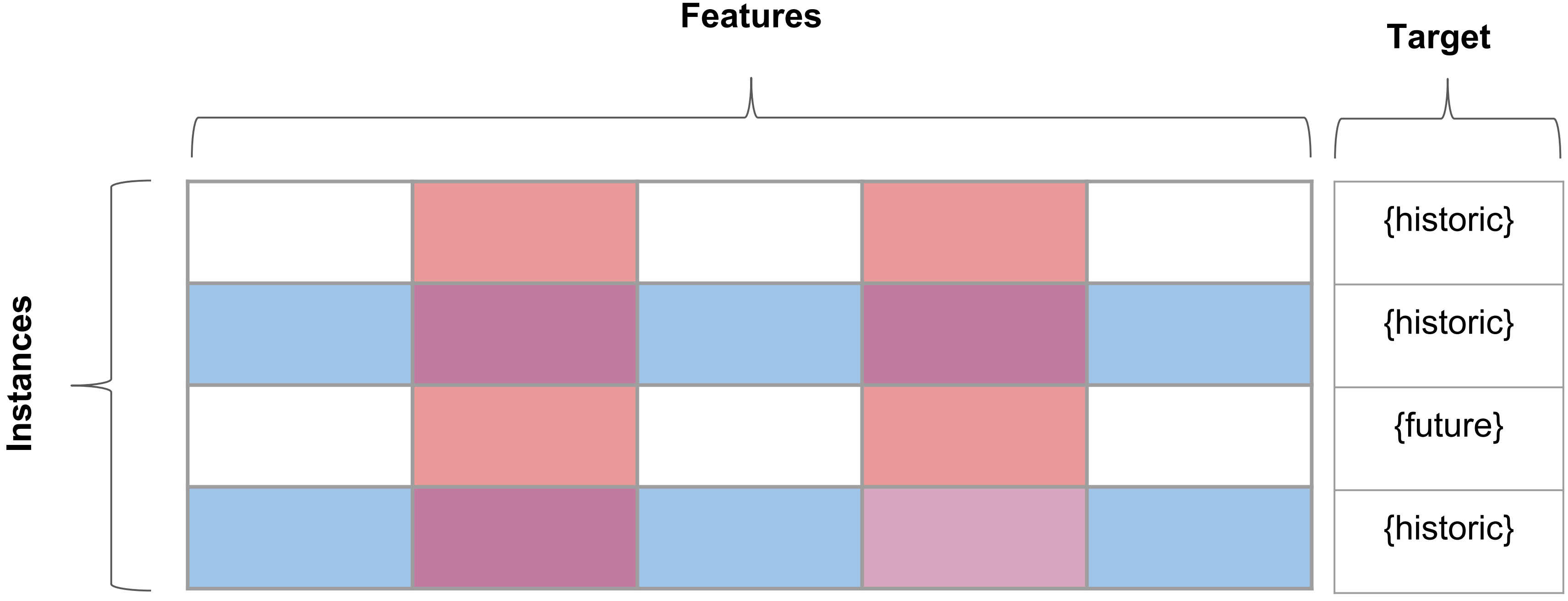


Where does the target coming from?

1) An existing feature that is missing from some instances
2) A hand labeled feature

Preprocessing: Target

Instance, rows, feature, attribute, column, target, label

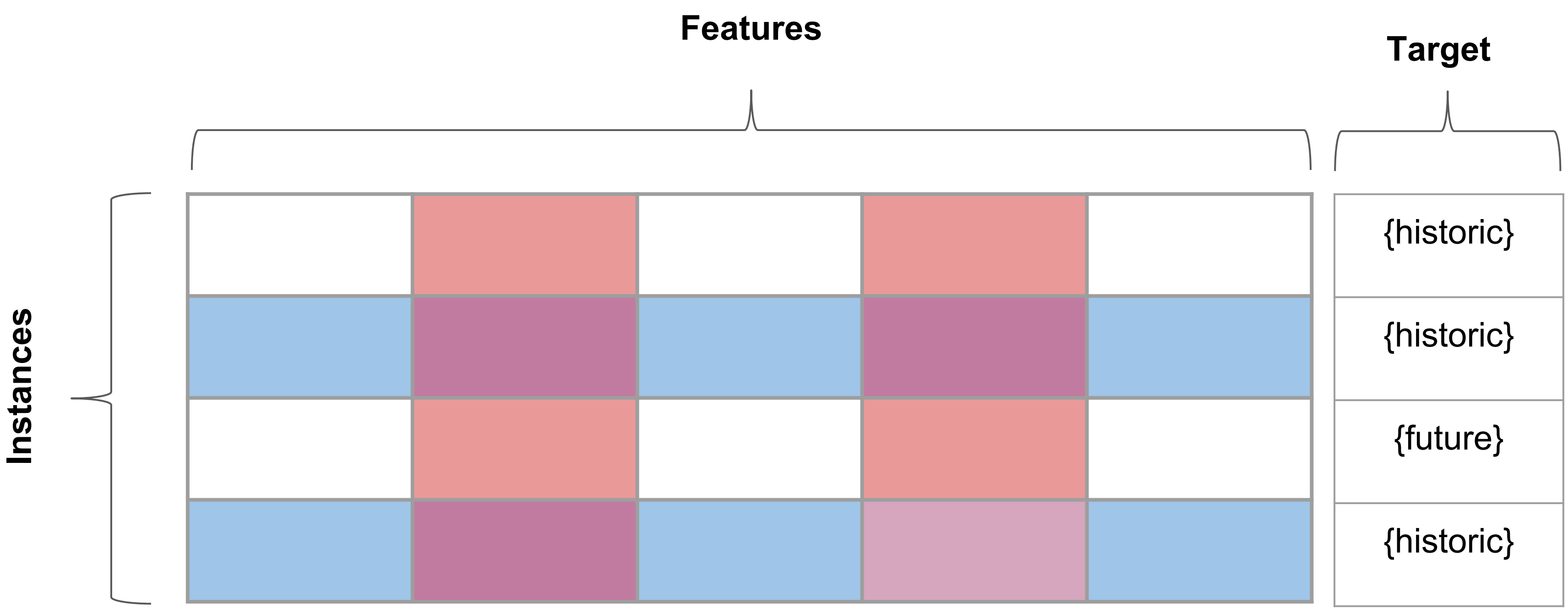


Where does the target coming from?

1) An existing feature that is missing from some instances
2) A hand labeled feature
3) A feature value that that will be known in the future

Preprocessing: Target

Instance, rows, feature, attribute, column, target, label

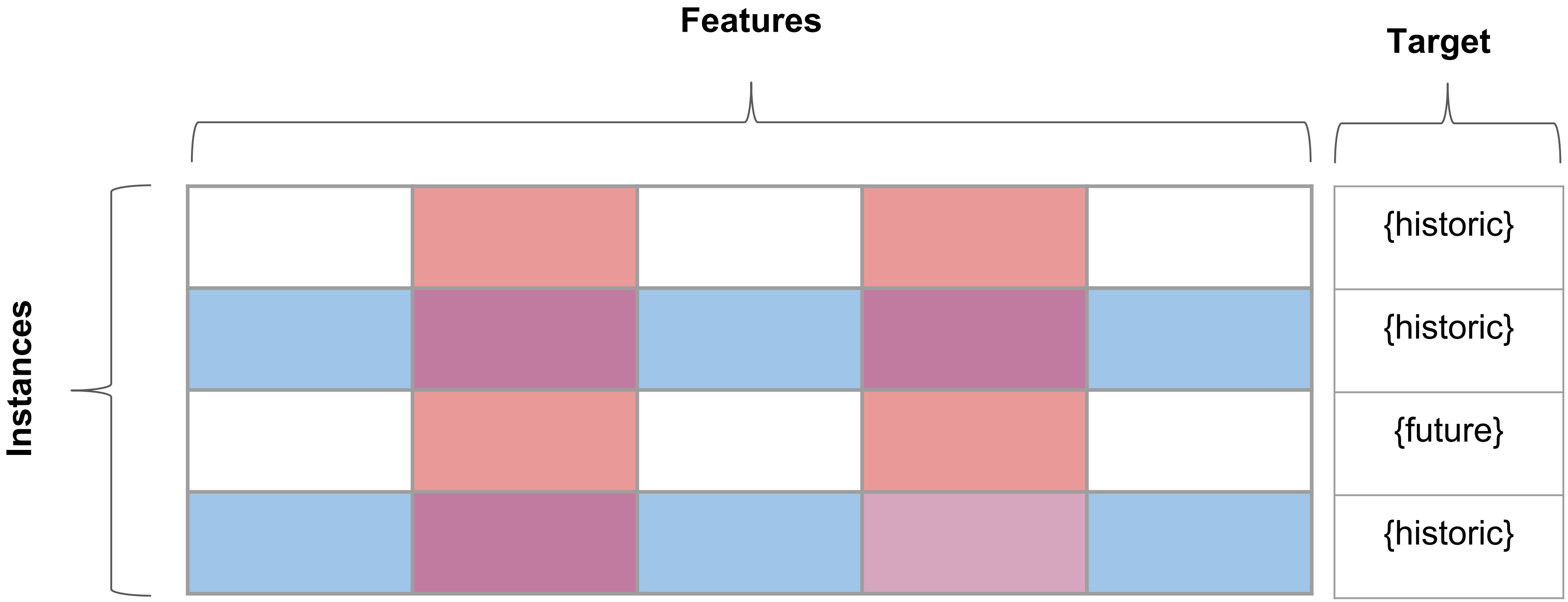


Classification: $X^n_m \rightarrow Y^n$ (the target)

Rule of thumb: The greater the variation in X, the greater (and more challenging) the generalization being sought.

Preprocessing: Target

Instance, rows, feature, attribute, column, target, label



Classification: $X_m^n \rightarrow Y^n$ (the target)

What is an Instance?

Preprocessing: Example 1

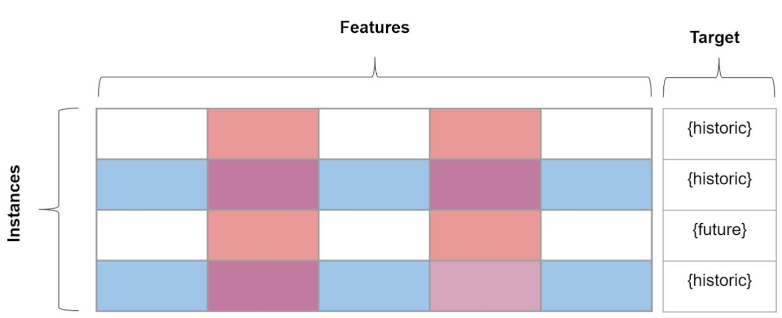
Weather Data:

Average daily temperature data for every US city for past 365 days

<u>Date</u>, <u>City</u>, <u>Temperature</u>

1/26/2016, Berkeley, 56
1/25/2016, Berkeley, 54
…….
1/26/2016, Dallas, 67
1/25/2016, Dallas, 68
…….

What is the target, features, and instances?



Preprocessing: Example 1

Weather Data:

A few potential representations to think about:

One representation could have the 365th day's temperature be the target and have 20 features which represent the historic temperatures from days 345 to 364. $X^{\#cities}_{20}$

This representation would be leaving out 344 days' worth of data (most of the dataset). To utilize more data, every contiguous 20 day period could be used as a feature set and the 21st day as the target. $X^{\#cities*345}_{20}$

Alternatively, if you wanted a single city centric dataset, the target could be the target at time T for, say, Berkeley and the features could be historic temperature data for all cities. $X^{\#345}_{20*\#cities}$

This is a region/location agnostic model.
How can we add this information to the feature set?

Preprocessing: Example 2

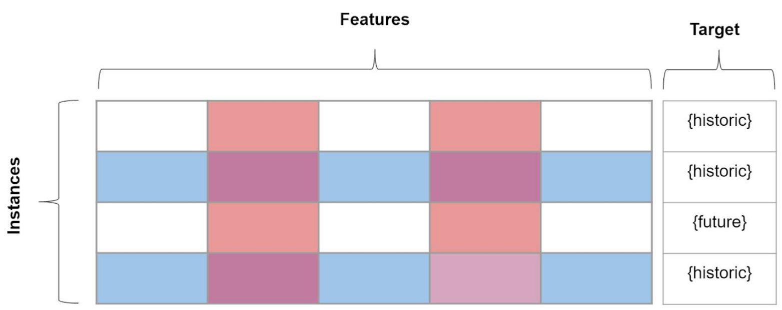
School District Data:

Average middle school and high school class grades in an academic year for each student in a district for the past 5 years

AY, Student, Teacher, School, Grade Level, Avg. Grade

2015-2016, Sammy Davis, Alabama High, 11, Hugh Laurie, B-
2014-2015, Sammy Davis, Alabama High, 10, Jeremey Irons, C+
…….
2015-2015, Alex Rodriguez, Centennial, 7, Steve McQueen, A
…….

What is the target, features, and instances?



Preprocessing: Example 3

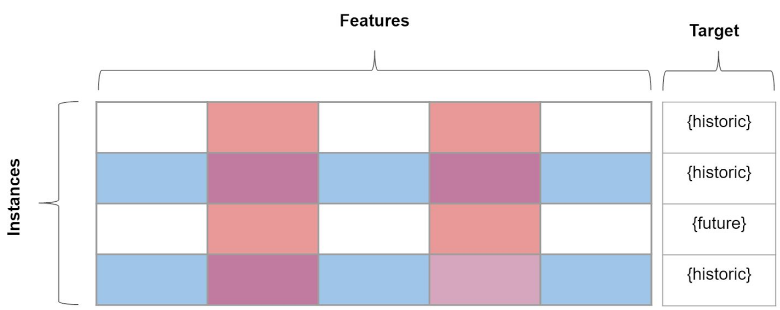
Book Text Data:

Average middle school and high school class grades in an academic year for each student in a district for the past 5 years

Book name, **Complete text**, **Year of release**, **Current Amazon Sales Rank**

A Brief History of Time, “Even if there is only..”,. 1998, 1073
Data Mining: Concepts and Techniques (3rd), “Analyzing large..”, 2011, 23899
…….

What is the target, features, and instances?



Concluding thought:

Before getting into optimizing models
- Understand the problem you're trying to solve
- Does your representation of the input data make sense for solving the problem?

See you Thursday

- Thursday quiz will be 2-5 questions (from slides/reading)
- Tutorial followed by start of Lab

Questions?

INFO 254 / INFO 154
School of Information / Spring 2019

Prof. Zach Pardos