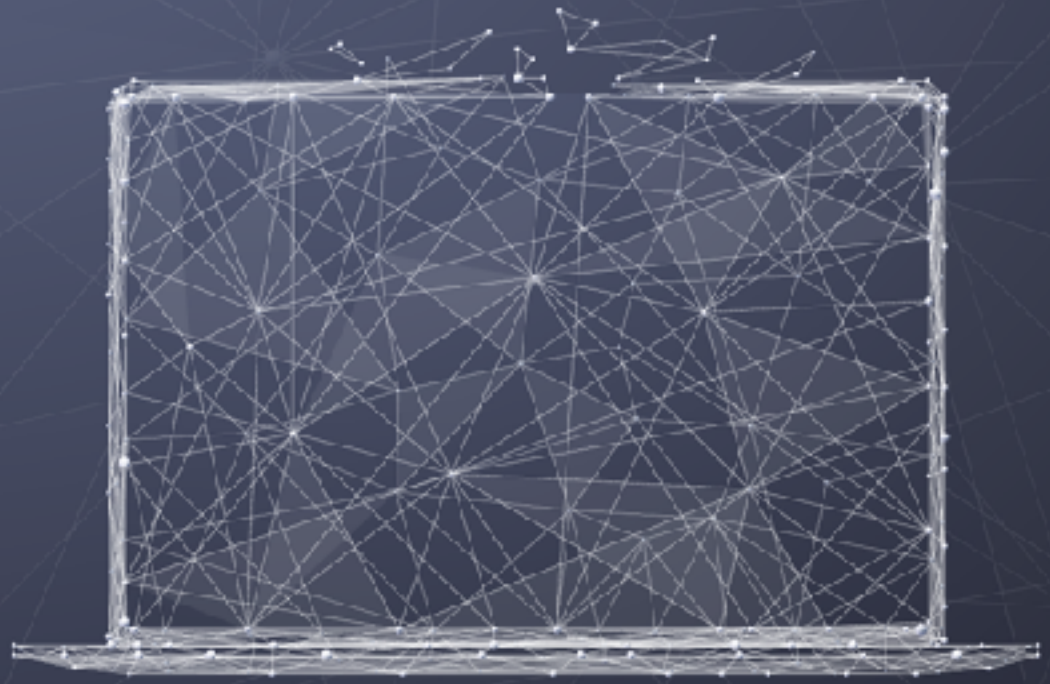
Data Science
Data Engineering I
Data quality and integration



PURDUE UNIVERSITY | College of Science

Data quality dimensions

All relevant data is recorded

The data agrees with itself

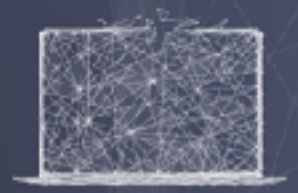
Entities are recorded once

The data is recorded correctly

The data are up to date

The data conform to their definitions

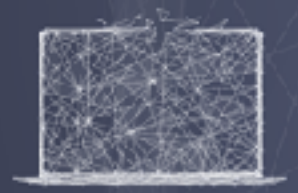
PURDUE UNIVERSITY | College of Science



Data gathering

Source data can be dirty

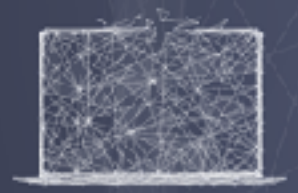
- Problems can arise due to:
  - Manual entry
  - No uniform standards for content and formats
  - Parallel data entry (may lead to duplicates)
  - Approximations, surrogates due to software/hardware constraints
  - Measurement errors



Data delivery

Transmission pipeline can cause problems

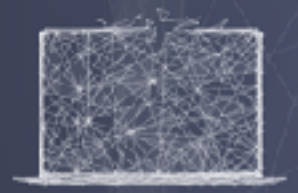
- Information may be dropped or degraded by preprocessing:
  - Inappropriate aggregation
  - Nulls converted to default values
- Loss of data:
  - Buffer overflows
  - Transmission problems
  - Problems in physical storage



Data integration

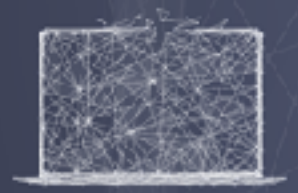
Combination of data can lead to errors

- Often data is stored in separate warehouses and then combined for analysis (e.g., across departments)
- Common source of problems
  - Heterogenous data: no common identifier, different field formats
  - Different definitions of entities/attributes
  - Time synchronization
  - Legacy data
  - Sociological factors



Common data quality problems

- Missing values
- Noise
- Outliers
- Duplicate data

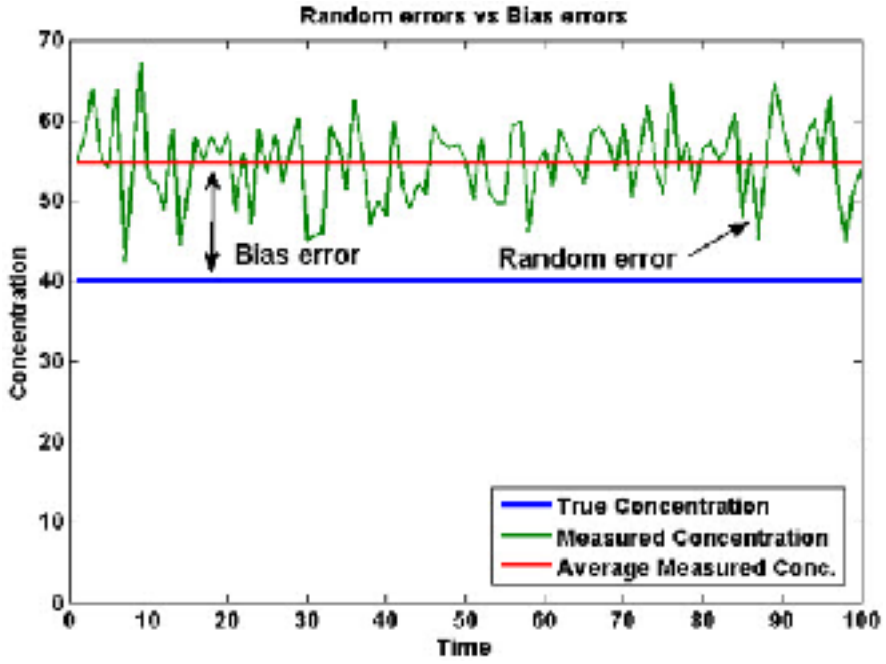


Missing values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Ways to handle missing values
  - Eliminate entities with missing values
  - Ignore the missing values during analysis
  - Estimate (ie., impute) missing values
  - Replace with all possible values (weighted by their probabilities)

Noise

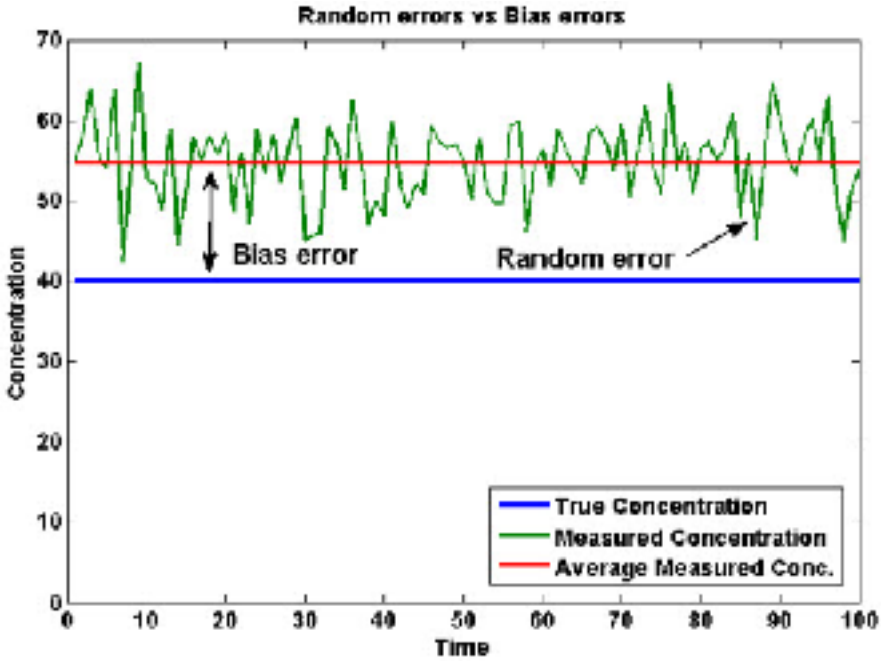
- Noise refers to measurement error in data values
- Could be random error or systematic (bias) error
- Random errors are due to inconsistencies in measurement, i.e. lack of precision (some high and some low)
- Example: different people reporting heights measured with a ruler (some may be an overestimate others may be an underestimate)



PURDUE UNIVERSITY | College of Science

Noise

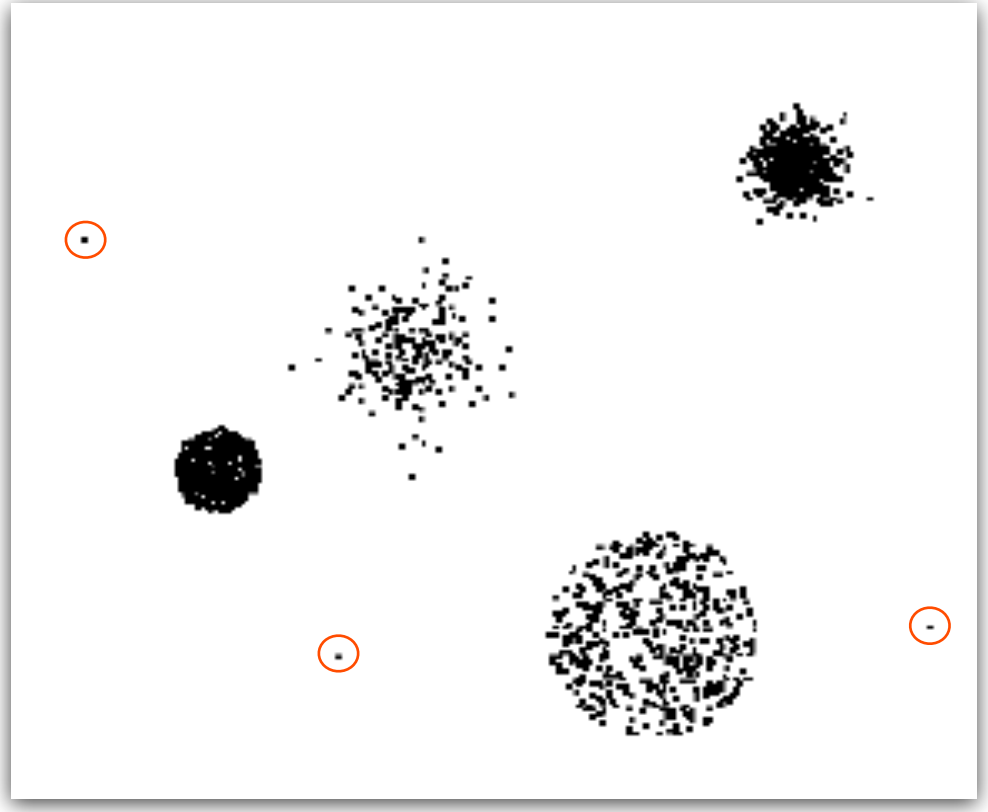
- Noise refers to measurement error in data values
- Could be random error or systematic (bias) error
- Systematic errors are due to lack of precision, i.e. methodological/personal errors that consistently deviate in one direction
- Example: heights measured with a “broken” ruler that is shorter than the numbers indicate



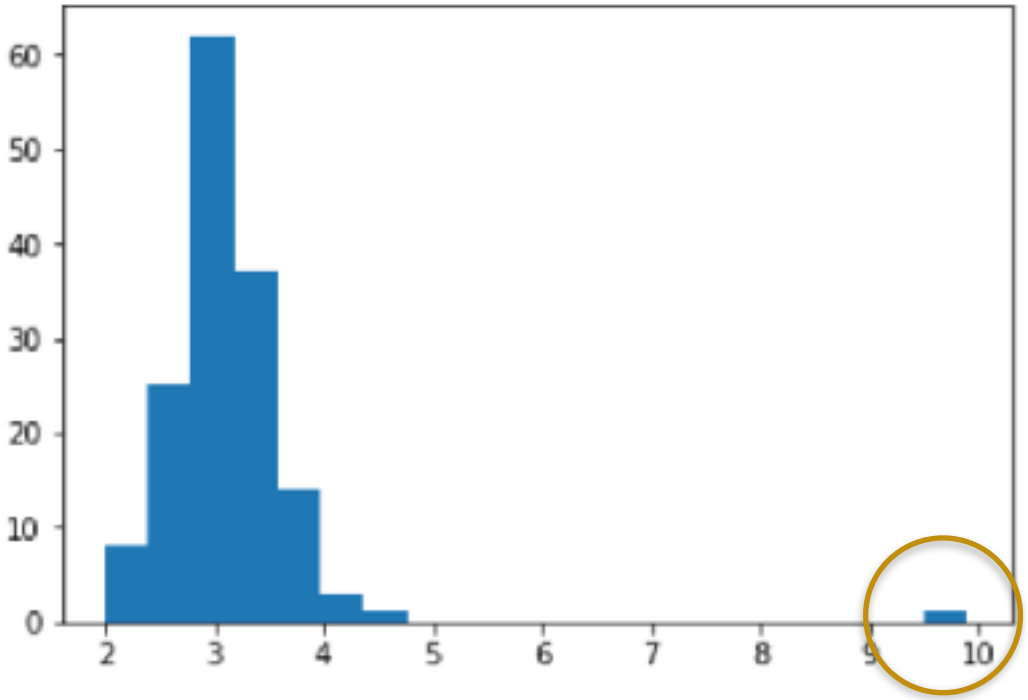
PURDUE UNIVERSITY | College of Science

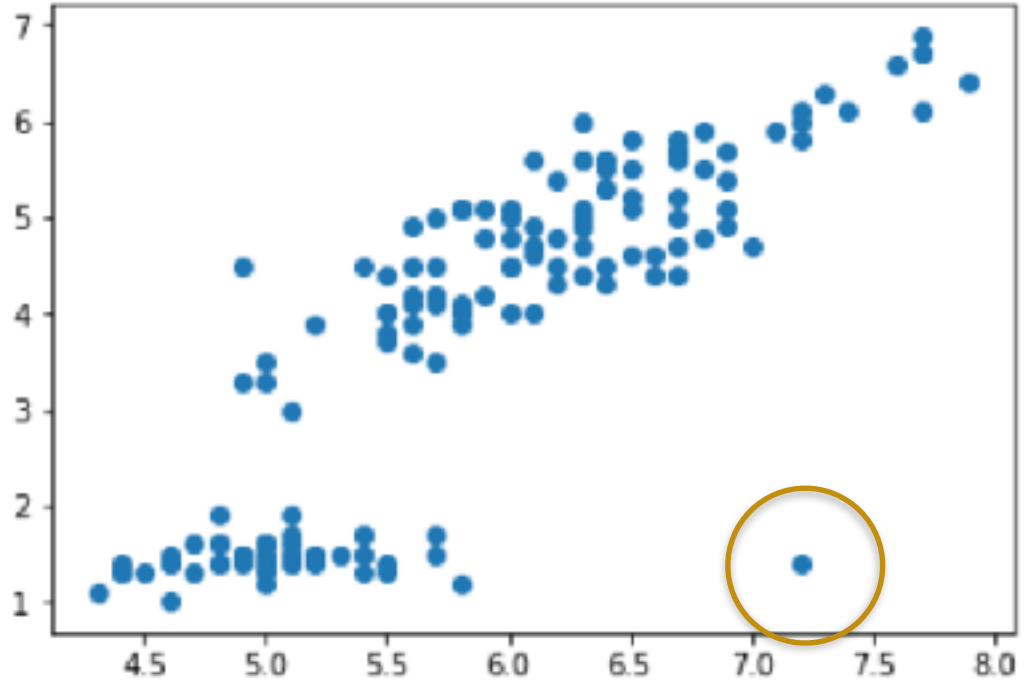
Outliers

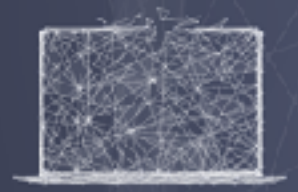
- Outliers are data objects with characteristics that are: considerably different than most of the other data objects in the data set
- Could indicate “interesting” cases (if so, highlight in output of analysis)
- Could indicate errors in the data (if so, remove or correct)



Finding outliers thru visualization



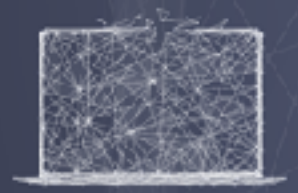




Finding outliers thru analysis

```
attrs = list(data.columns)
for i in range(4):
    avg = data.iloc[:,i].mean()
    std = data.iloc[:,i].std()
    highthresh = avg + (2*std)
    lowthresh = avg - (2*std)
    outliers = data[(data[attrs[i]]>highthresh) | (data[attrs[i]]<lowthresh)]
    print(outliers.iloc[:,i])
```

```
107    7.6
119    7.7
120    7.7
124    7.7
133    7.9
137    7.7
Name: sepal-length, dtype: float64
50    9.9
Name: sepal-width, dtype: float64
Series([], Name: petal-length, dtype: float64)
Series([], Name: petal-width, dtype: float64)
```



Finding duplicated data

```
# finding merge mistake where last three rows are added twice
data = pd.read_csv("oscar_age_female_mod.csv",sep=',')
data[data.duplicated(keep=False)]
```

	Index	Year	Age	Name	Movie
88	87	2014	44	Cate Blanchett	Blue Jasmine
89	88	2015	54	Julianne Moore	Still Alice
90	89	2016	26	Brie Larson	Room
91	87	2014	44	Cate Blanchett	Blue Jasmine
92	88	2015	54	Julianne Moore	Still Alice
93	89	2016	26	Brie Larson	Room

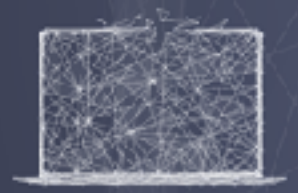
Finding duplicated data

```
# finding duplicates with different versions of name
data[data.duplicated(['Index','Year'],keep=False)]
```

	Index	Year	Age	Name	Movie
15	16	1943	25	Jennifer Jones	The Song of Bernadette
16	16	1943	25	Jen Jones	The Song of Bernadette
86	86	2013	22	Jennifer Lawrence	Silver Linings Playbook
87	86	2013	22	Jen Lawrence	Silver Linings Playbook

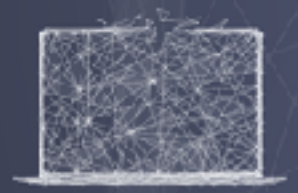
```
data.drop_duplicates(['Index','Year'],keep='first')
```

13	14	1941	24	Joan Fontaine	Suspicion
14	15	1942	38	Greer Garson	Mrs. Miniver
15	16	1943	25	Jennifer Jones	The Song of Bernadette
17	17	1944	29	Ingrid Bergman	Gaslight



Data integration

- Data is often available in multiple distinct databases and needs to be combined for analysis
- When is data integration needed?
  - To analyze data produced by different sources
  - To combine data from different websites
  - To combine legacy databases
  - Two companies merge



Merging data frames

```
# down select a few rows in different data fames
d1 = data.iloc[1:4,:]
d2 = data.iloc[8:11,:]
# concatenate the two data frames together
d3 = pd.concat([d1,d2])
```

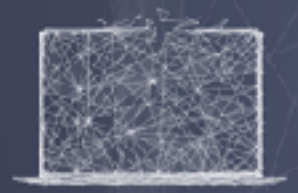
	Index	Year	Age	Name	Movie
1	2	1929	37	Mary Pickford	Coquette
2	3	1930	28	Norma Shearer	The Divorcee
3	4	1931	63	Marie Dressler	Min and Bill
8	9	1936	27	Luise Rainer	The Great Ziegfeld
9	10	1937	28	Luise Rainer	The Good Earth
10	11	1938	30	Bette Davis	Jezebel

Merging data frames

```
# compute year of birth for actresses, get unique actress records
data['YoB'] = data.Year-data.Age
actress = data.iloc[:,[3,5]].drop_duplicates()
# get first ten movies
movies = data.iloc[1:11,[0,1,3,4]]
```

	Name	YoB
0	Janet Gaynor	1906
1	Mary Pickford	1892
2	Norma Shearer	1902
3	Marie Dressler	1868
4	Helen Hayes	1900
5	Katharine Hepburn	1907
6	Claudette Colbert	1903
7	Bette Davis	1908
8	Luise Rainer	1909
11	Vivien Leigh	1913

	Index	Year	Name	Movie
1	2	1929	Mary Pickford	Coquette
2	3	1930	Norma Shearer	The Divorcee
3	4	1931	Marie Dressler	Min and Bill
4	5	1932	Helen Hayes	The Sin of Madelon Claudet
5	6	1933	Katharine Hepburn	Morning Glory
6	7	1934	Claudette Colbert	It Happened One Night
7	8	1935	Bette Davis	Dangerous
8	9	1936	Luise Rainer	The Great Ziegfeld
9	10	1937	Luise Rainer	The Good Earth
10	11	1938	Bette Davis	Jezebel



Merging data frames

```
# join actress information to movies
pd.merge(movies, actress, on='Name').sort_values('Year')
```

	Index	Year	Name	Movie	YoB
0	2	1929	Mary Pickford	Coquette	1892
1	3	1930	Norma Shearer	The Divorcee	1902
2	4	1931	Marie Dressler	Min and Bill	1868
3	5	1932	Helen Hayes	The Sin of Madelon Claudet	1900
4	6	1933	Katharine Hepburn	Morning Glory	1907
5	6	1933	Katharine Hepburn	Morning Glory	1908
6	7	1934	Claudette Colbert	It Happened One Night	1903
7	8	1935	Bette Davis	Dangerous	1908
9	9	1936	Luise Rainer	The Great Ziegfeld	1909
10	10	1937	Luise Rainer	The Good Earth	1909
8	11	1938	Bette Davis	Jezebel	1908