PRINCIPLES OF DATA SCIENCE

JOHN P DICKERSON

Lecture #5 – 9/26/2018

CMSC641
Wednesdays
7pm – 9:30pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

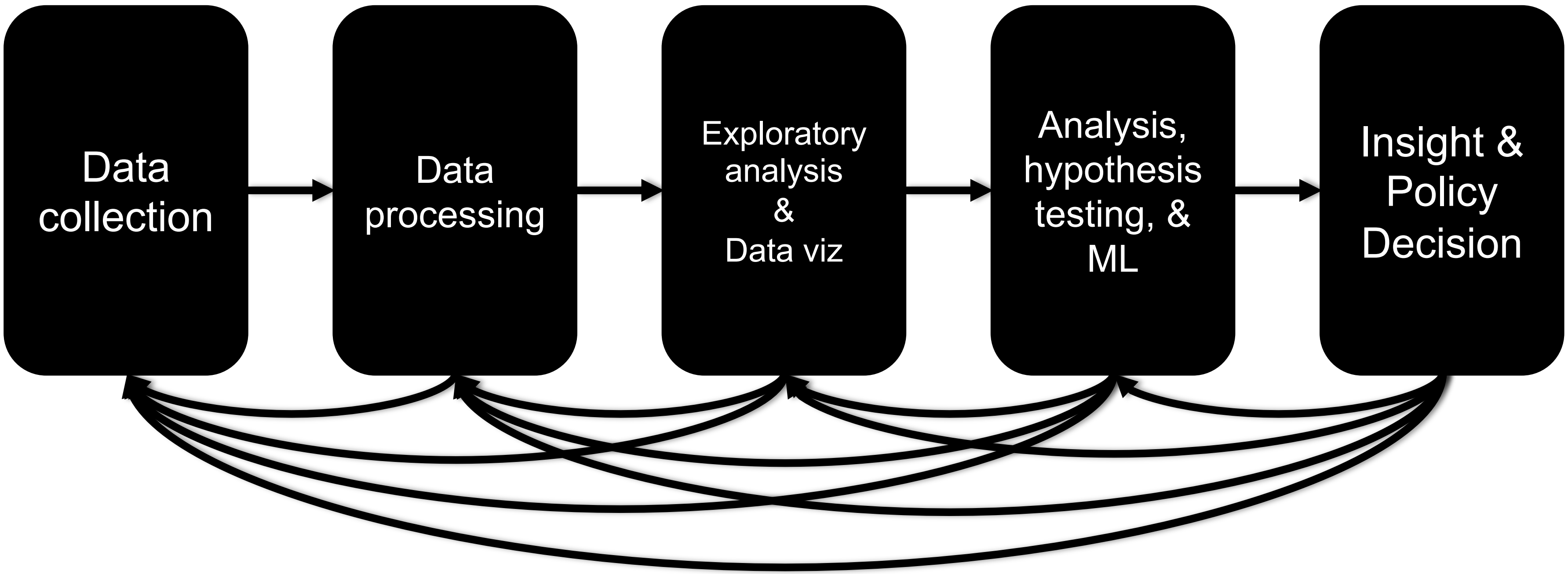
ANNOUNCEMENTS

Project 1 is out!

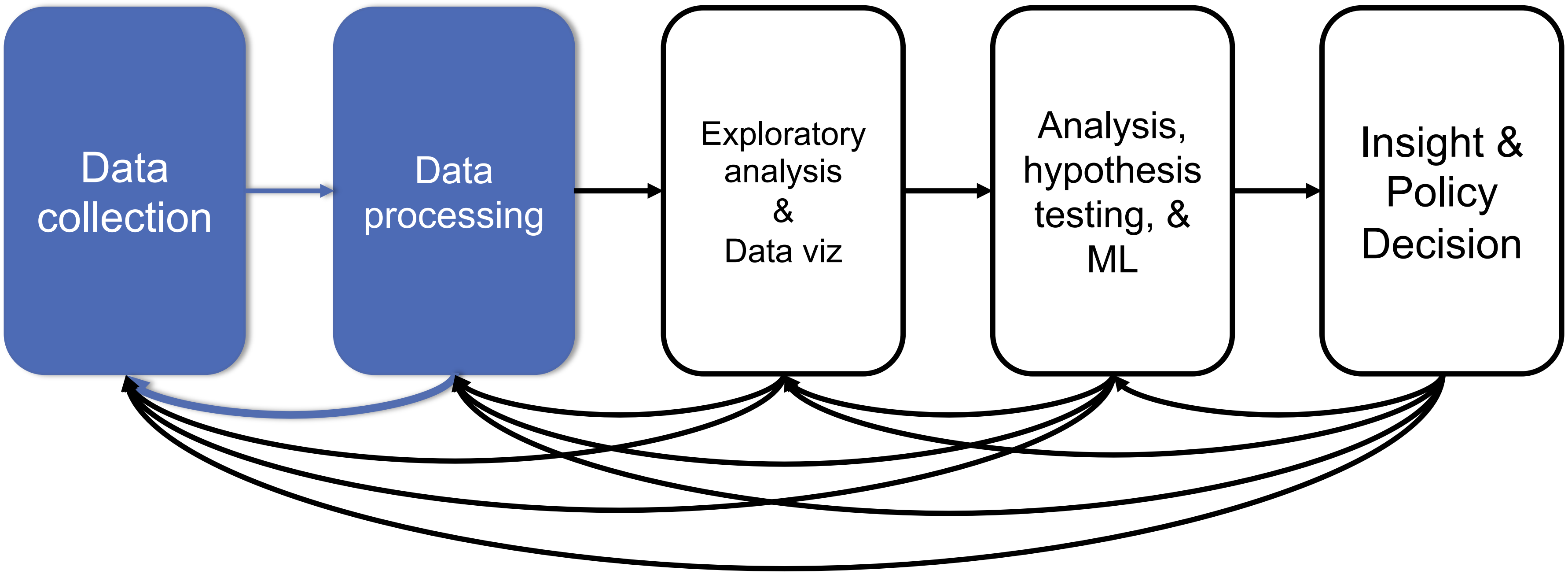
- Announced on ELMS and Piazza
- https://github.com/JohnDickerson/cmsc641-fall2018/tree/master/project1
- Due date is October 3rd

Reminder: Weekly quizzes, due on Wednesdays at noon

THE DATA LIFECYCLE

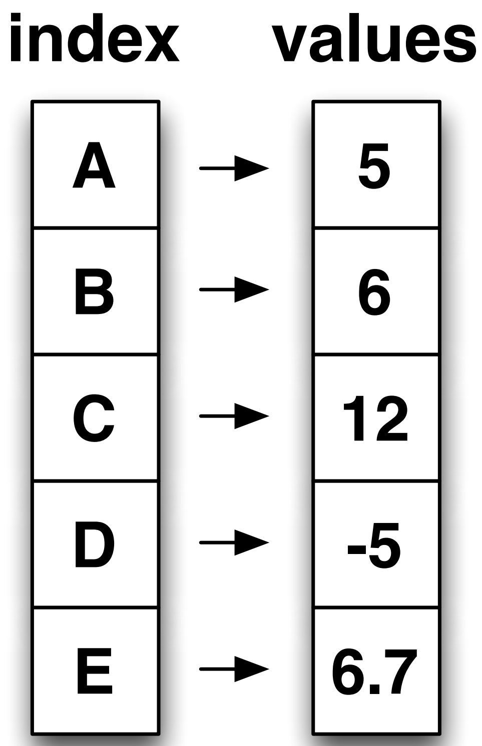


THE DATA LIFECYCLE



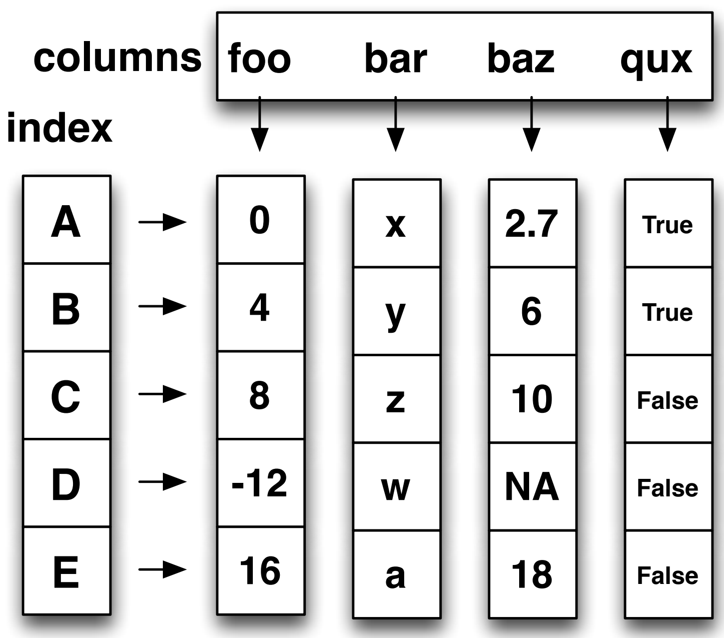
Quick wrap-up from last class: pandas/relational databases

PANDAS: SERIES



- Subclass of numpy.ndarray
- Data: any type
- Index labels need not be ordered
- Duplicates possible but result in reduced functionality

PANDAS: DATAFRAME



- Each column can have a different type
- Row and Column index
- Mutable size: insert and delete columns

- Note the use of word "index" for what we called "key"
  - Relational databases use "index" to mean something else

- Non-unique index values allowed
  - May raise an exception for some operations

RELATION

Simplest relation: a table aka tabular data full of unique tuples

Variables
(called attributes)

Labels

Observations
(called tuples)

ID	age	wgt_kg	hgt_cm
1	12.2	42.3	145.1
2	11.0	40.8	143.8
3	15.6	65.3	165.3
4	35.1	84.2	185.8

PRIMARY KEYS

ID	age	wgt_kg	hgt_cm	nat_id
1	12.2	42.3	145.1	1
2	11.0	40.8	143.8	1
3	15.6	65.3	165.3	2
4	35.1	84.2	185.8	1
5	18.1	62.2	176.2	3
6	19.6	82.1	180.1	1

ID	Nationality
1	USA
2	Canada
3	Mexico

The primary key is a unique identifier for every tuple in a relation

- **Each tuple has exactly one primary key**

FOREIGN KEYS

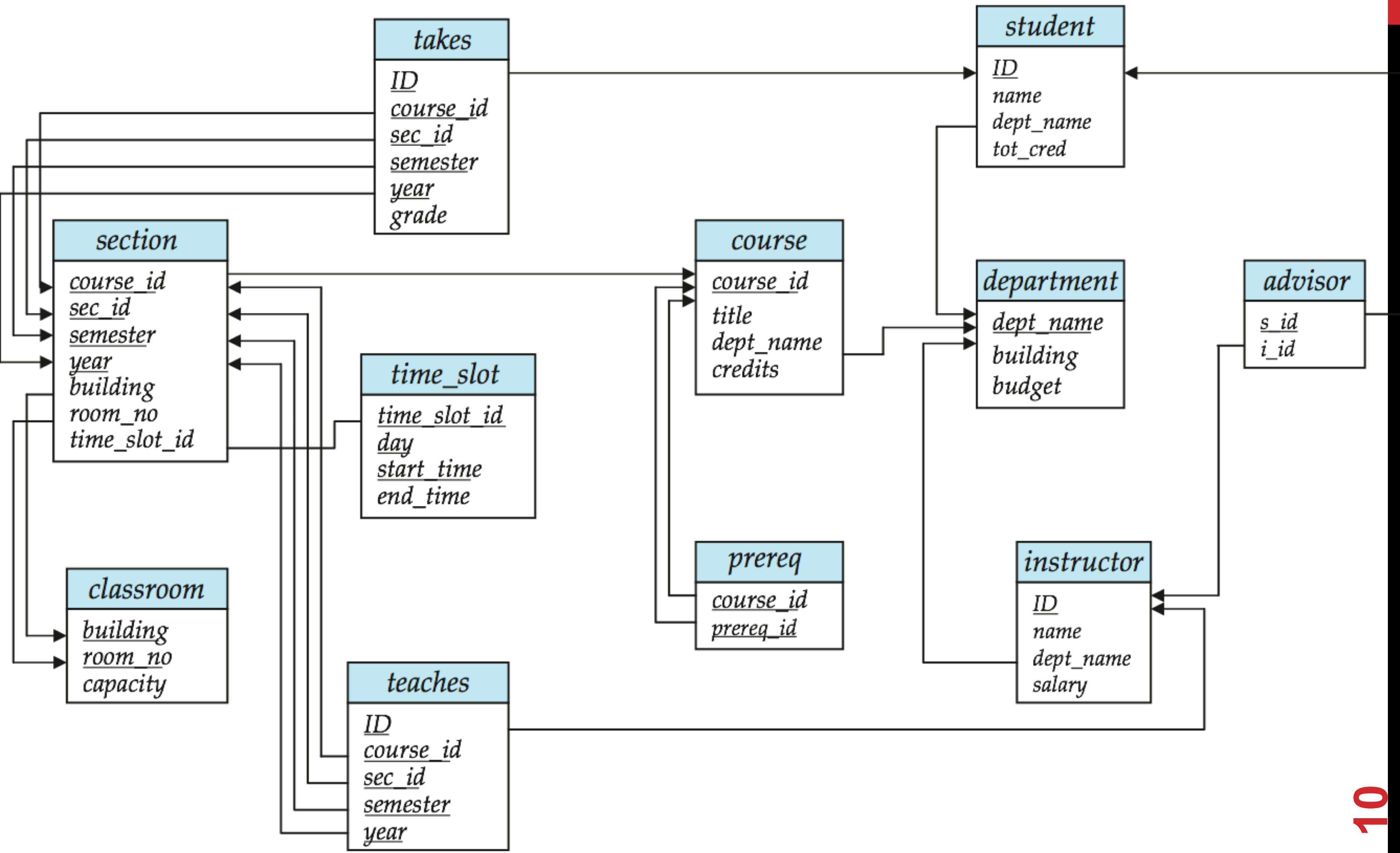
ID	age	wgt_kg	hgt_cm	nat_id
1	12.2	42.3	145.1	1
2	11.0	40.8	143.8	1
3	15.6	65.3	165.3	2
4	35.1	84.2	185.8	1
5	18.1	62.2	176.2	3
6	19.6	82.1	180.1	1

ID	Nationality
1	USA
2	Canada
3	Mexico

Foreign keys are attributes (columns) that point to a different table's primary key

- **A table can have multiple foreign keys**

SCHEMA DIAGRAMS



JOINING DATA

A join operation merges two or more tables into a single relation. Different ways of doing this:

- Inner
- Left
- Right
- Full Outer

Join operations are done on columns that explicitly link the tables together

INNER JOINS

id	name
1	Megabyte
2	Meowly Cyrus
3	Fuzz Aldrin
4	Chairman Meow
5	Anderson Pooper
6	Gigabyte

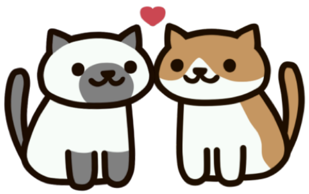
`cats`

cat_id	last_visit
1	02-16-2017
2	02-14-2017
5	02-03-2017

`visits`

Inner join returns merged rows that share the same value in the column they are being joined on (`id` and `cat_id`).

id	name	last_visit
1	Megabyte	02-16-2017
2	Meowly Cyrus	02-14-2017
5	Anderson Pooper	02-03-2017



INNER JOINS

```
# Inner join in pandas
df_cats = pd.read_sql_query("SELECT * from cats", conn)
df_visits = pd.read_sql_query("SELECT * from visits", conn)
df_cats.merge(df_visits, how = "inner",
              left_on = "id", right_on = "cat_id")
```

```
# Inner join in SQL / SQLite via Python
cursor.execute("""
               SELECT
                   *
               FROM
                   cats, visits
               WHERE
                   cats.id == visits.cat_id
               """)
```

LEFT JOINS

Inner joins are the most common type of joins (get results that appear in both tables)

Left joins: all the results from the left table, only some matching results from the right table

Left join (`cats`, `visits`) on (`id`, `cat_id`) ???????????

id	name	last_visit
1	Megabyte	02-16-2017
2	Meowly Cyrus	02-14-2017
3	Fuzz Aldrin	NULL
4	Chairman Meow	NULL
5	Anderson Pooper	02-03-2017
6	Gigabyte	NULL

RIGHT JOINS

Take a guess!

Right join
`(cats, visits)`
on
`(id, cat_id)`
???????????

id	name
1	Megabyte
2	Meowly Cyrus
3	Fuzz Aldrin
4	Chairman Meow
5	Anderson Pooper
6	Gigabyte

`cats`

cat_id	last_visit
1	02-16-2017
2	02-14-2017
5	02-03-2017
7	02-19-2017
12	02-21-2017

`visits`

id	name	last_visit
1	Megabyte	02-16-2017
2	Meowly Cyrus	02-14-2017
5	Anderson Pooper	02-03-2017
7	NULL	02-19-2017
12	NULL	02-21-2017

LEFT/RIGHT JOINS

```
# Left join in pandas
df_cats.merge(df_visits, how = “left”,
              left_on = “id”, right_on = ”cat_id”)
```

```
# Left join in SQL / SQLite via Python
cursor.execute(“SELECT * FROM cats LEFT JOIN visits ON
                   cats.id == visits.cat_id”)
```

```
# Right join in pandas
df_cats.merge(df_visits, how = “right”,
              left_on = “id”, right_on = ”cat_id”)
```

```
# Right join in SQL / SQLite via Python
☹
```

FULL OUTER JOIN

Combines the left and the right join ????????????

id	name	last_visit
1	Megabyte	02-16-2017
2	Meowly Cyrus	02-14-2017
3	Fuzz Aldrin	NULL
4	Chairman Meow	NULL
5	Anderson Pooper	02-03-2017
6	Gigabyte	NULL
7	NULL	02-19-2017
12	NULL	02-21-2017

```
# Outer join in pandas
df_cats.merge(df_visits, how = “outer”,
              left_on = “id”, right_on = ”cat_id”)
```

GOOGLE IMAGE SEARCH ONE SLIDE SQL JOIN VISUAL

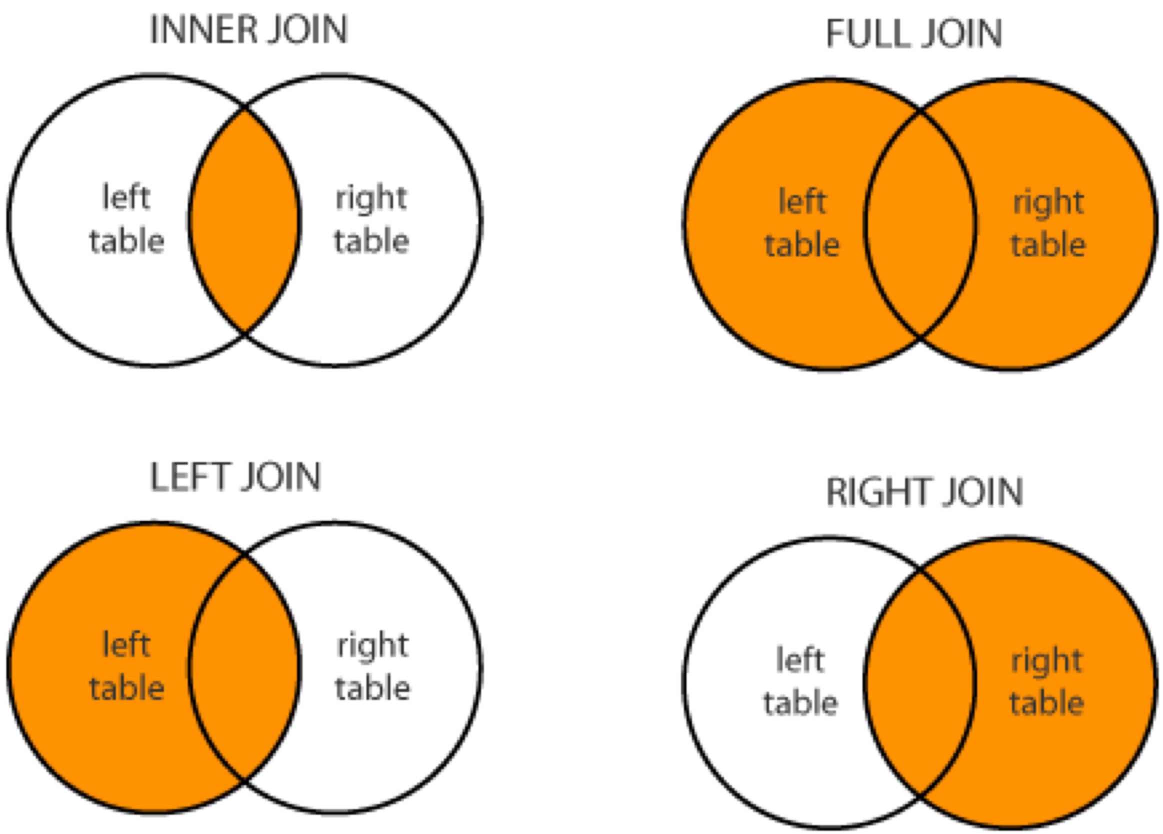


Image credit: http://www.dofactory.com/sql/join

GROUP BY AGGREGATES

```
SELECT nat_id, AVG(age) as average_age
FROM persons GROUP BY nat_id
```

ID	age	wgt_kg	hgt_cm	nat_id
1	12.2	42.3	145.1	1
2	11.0	40.8	143.8	1
3	15.6	65.3	165.3	2
4	35.1	84.2	185.8	1
5	18.1	62.2	176.2	3
6	19.6	82.1	180.1	1

nat_id	average_ age
1	19.48
2	15.6
3	18.1

RAW SQL IN PANDAS

The More You Know

If you “think in SQL” already, you’ll be fine with pandas:

- `conda install -c anaconda pandasql`
- Info: http://pandas.pydata.org/pandas-docs/stable/comparison_with_sql.html

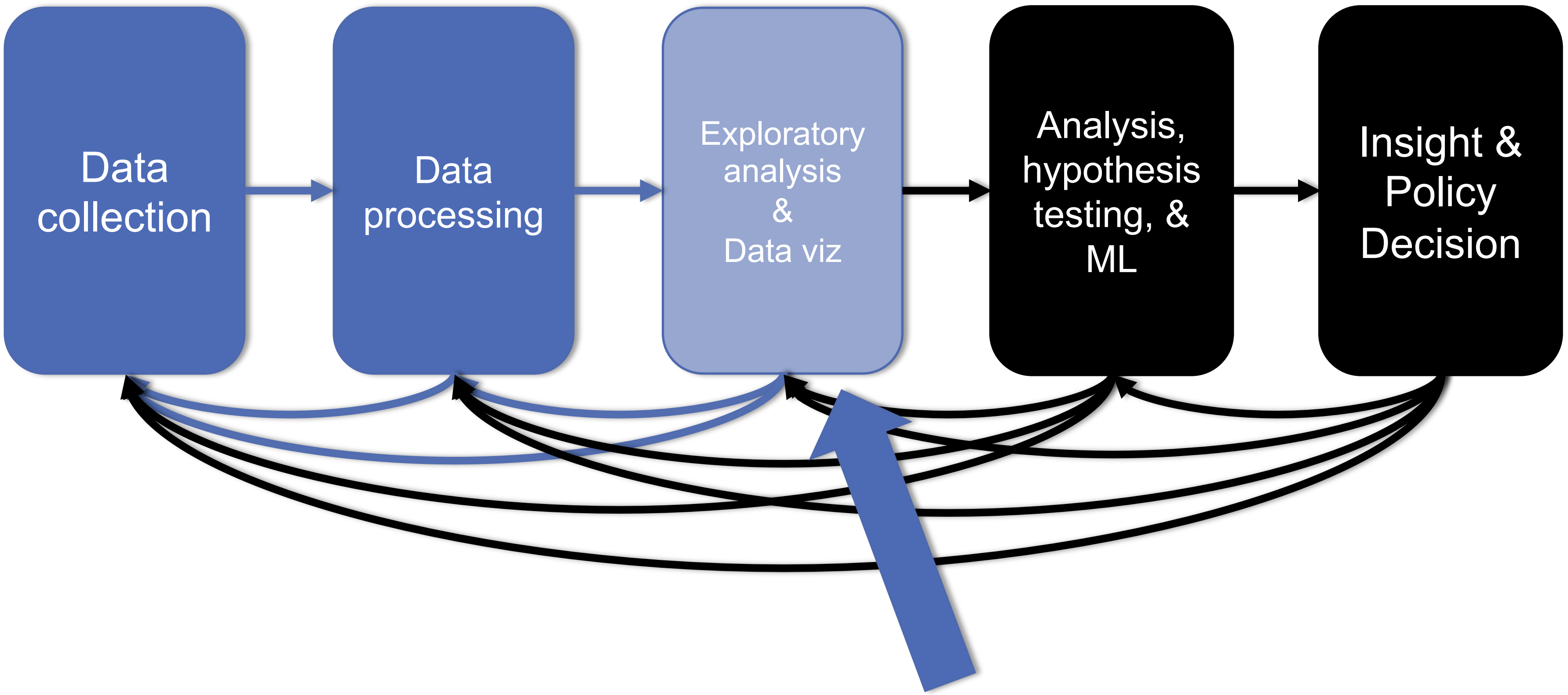
```
# Write the query text
q = """
    SELECT
        *
    FROM
        cats
    LIMIT 10;"""

# Store in a DataFrame
df = sqldf(q, locals())
```

FOR THE REST OF THIS CLASS:

EXPLORATORY ANALYSIS

TODAY'S LECTURE



Just a taste!

TODAY'S LECTURE

Missing Data …

- What is it?
- Simple methods for imputation

… with a tiny taste of Stats/ML lecturers to come.

Thanks to John Atwood and Wenjiang Fu

MISSING DATA

Missing data is information that we want to know, but don't

It can come in many forms, e.g.:

- People not answering questions on surveys
- Inaccurate recordings of the height of plants that need to be discarded
- Canceled runs in a driving experiment due to rain

Could also consider missing columns (no collection at all) to be missing data …

[JA]

KEY QUESTION

Why is the data missing?

- What mechanism is it that contributes to, or is associated with, the probability of a data point being absent?
- Can it be explained by our observed data or not?

The answers drastically affect what we can ultimately do to compensate for the missing-ness

[JA]

COMPLETE CASE ANALYSIS

Delete all tuples with any missing values at all, so you are left only with observations with all variables observed

```
# Clean out rows with nil values
df = df.dropna()
```

Default behavior for libraries for analysis (e.g., regression)

- We'll talk about this much more during the Stats/ML lectures

This is the simplest way to handle missing data. In some cases, will work fine; in others, ?????????????:

- Loss of sample will lead to variance larger than reflected by the size of your data
- May bias your sample

[JA]

EXAMPLE

Dataset: Body fat percentage in men, and the circumference of various body parts [Penrose et al., 1985]

Question: Does the circumference of certain body parts predict body fat percentage?

Given complete data, how would you answer this ?????????

One way to answer is regression analysis:

- One or more independent variables ("predictors”)
- One dependent variables (“outcome”)

What is the relationship between the predictors and the outcome?

What is the conditional expectation of the dependent variable given fixed values for the dependent variables?

LINEAR REGRESSION

Assumption: relationship between variables is linear:

- (We'll relax linearity, study in more depth later.)

Population Y-Intercept

Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

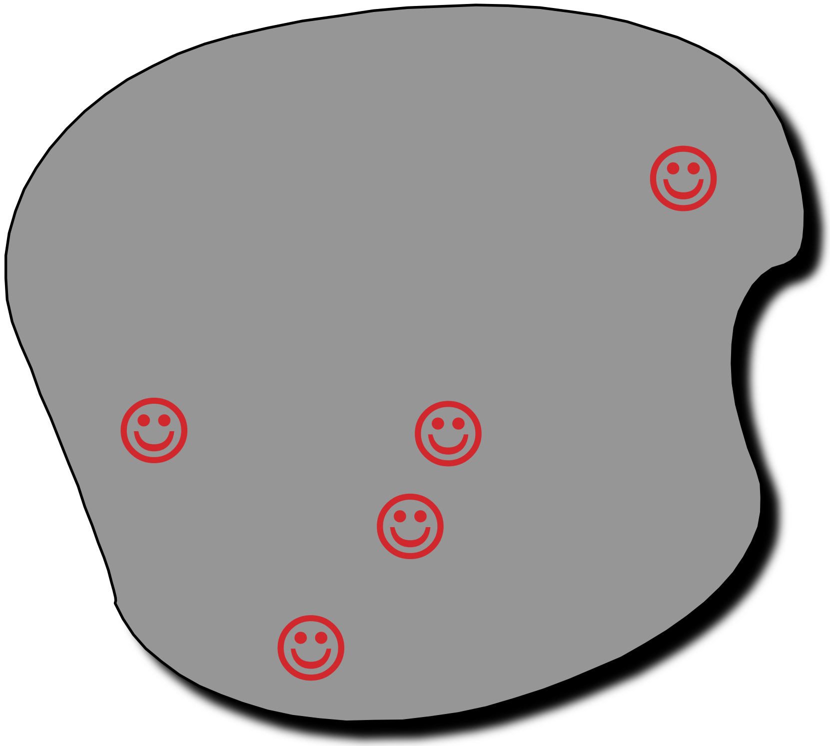
Dependent Variable (e.g., ????????)

Independent Variable(s) (e.g., ?????????)

[WF]

POPULATION & SAMPLE REGRESSION MODELS

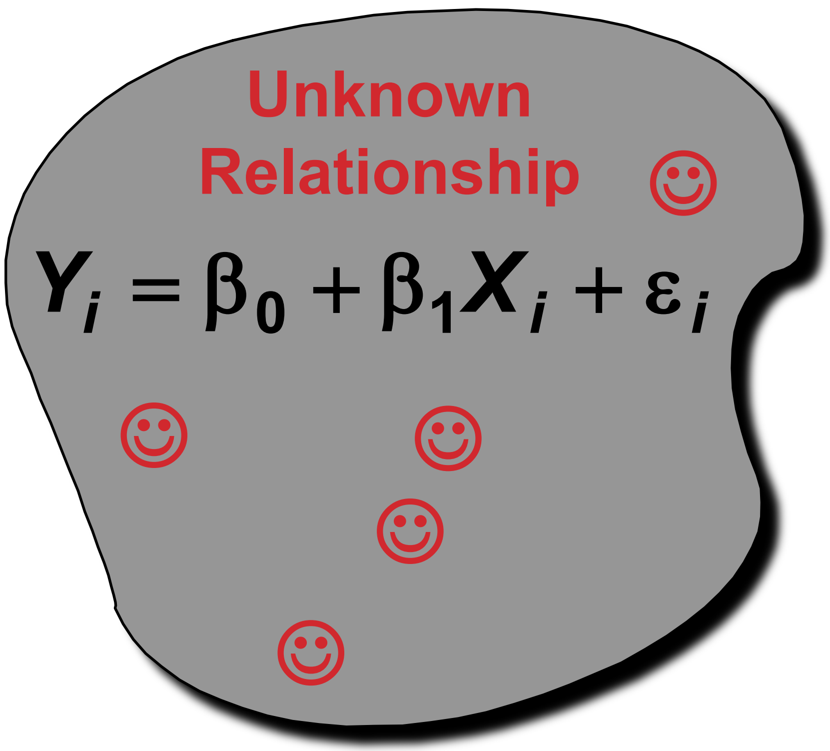
Population



[WF]

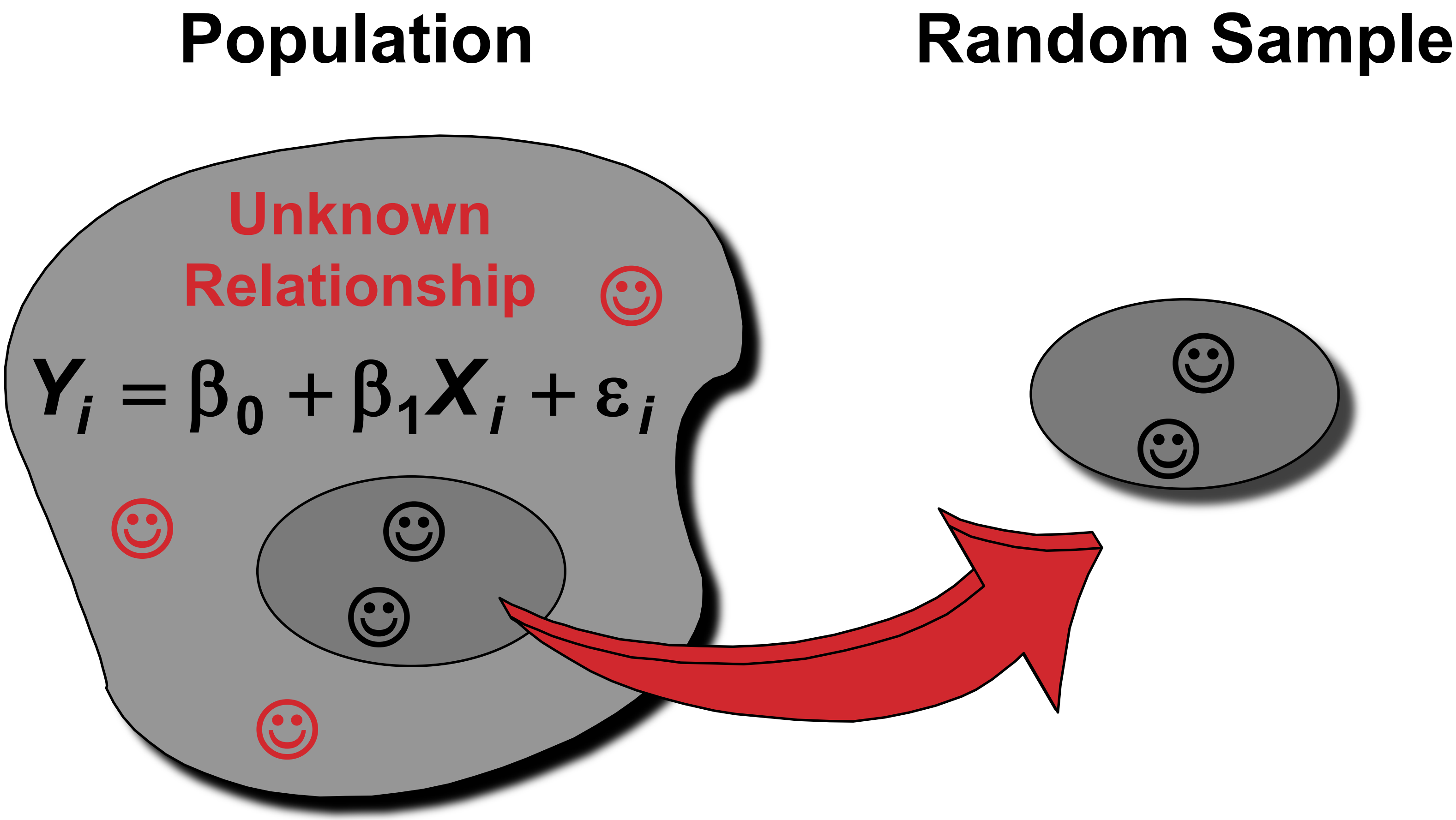
POPULATION & SAMPLE REGRESSION MODELS

Population



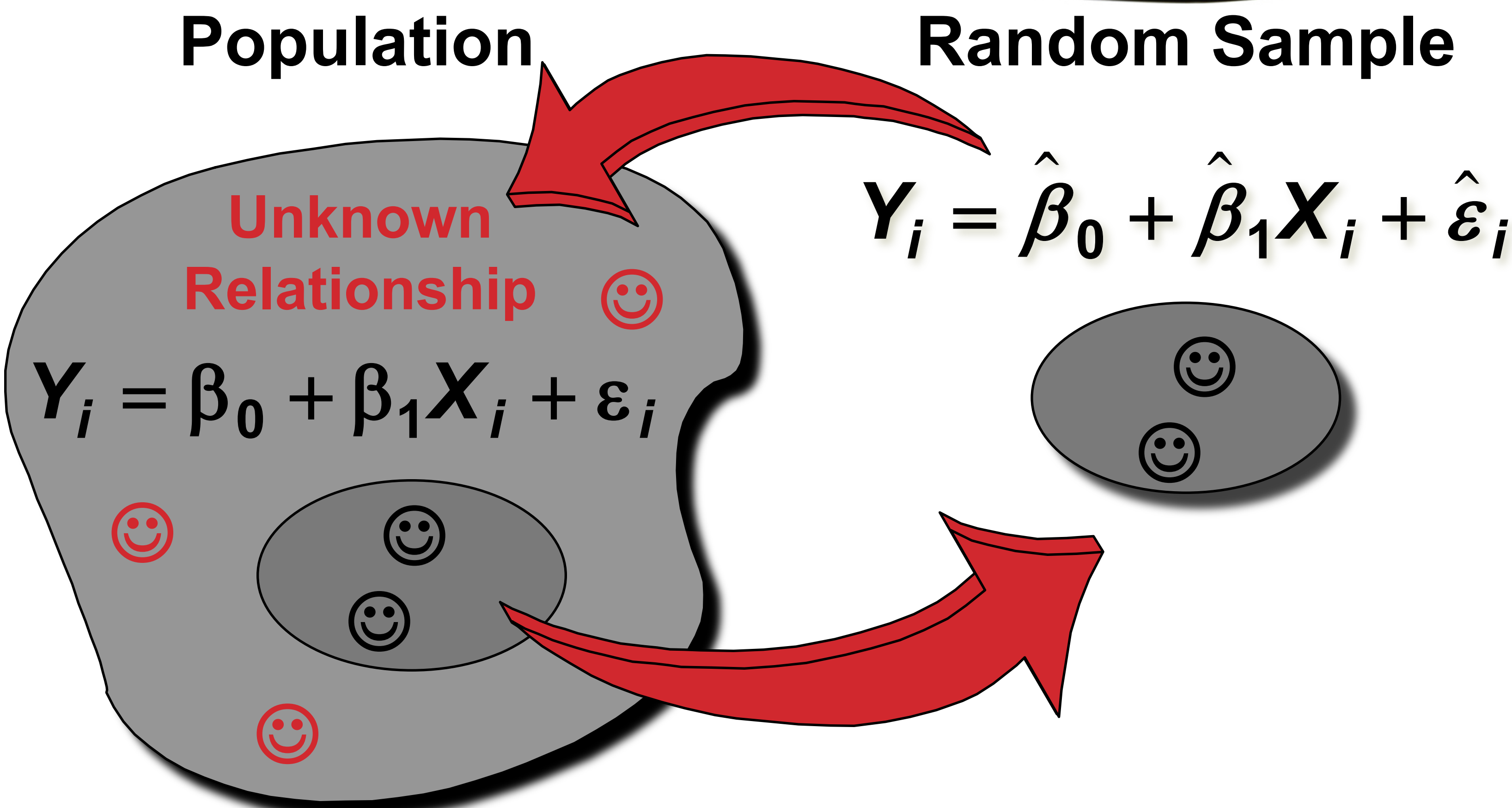
[WF]

POPULATION & SAMPLE REGRESSION MODELS



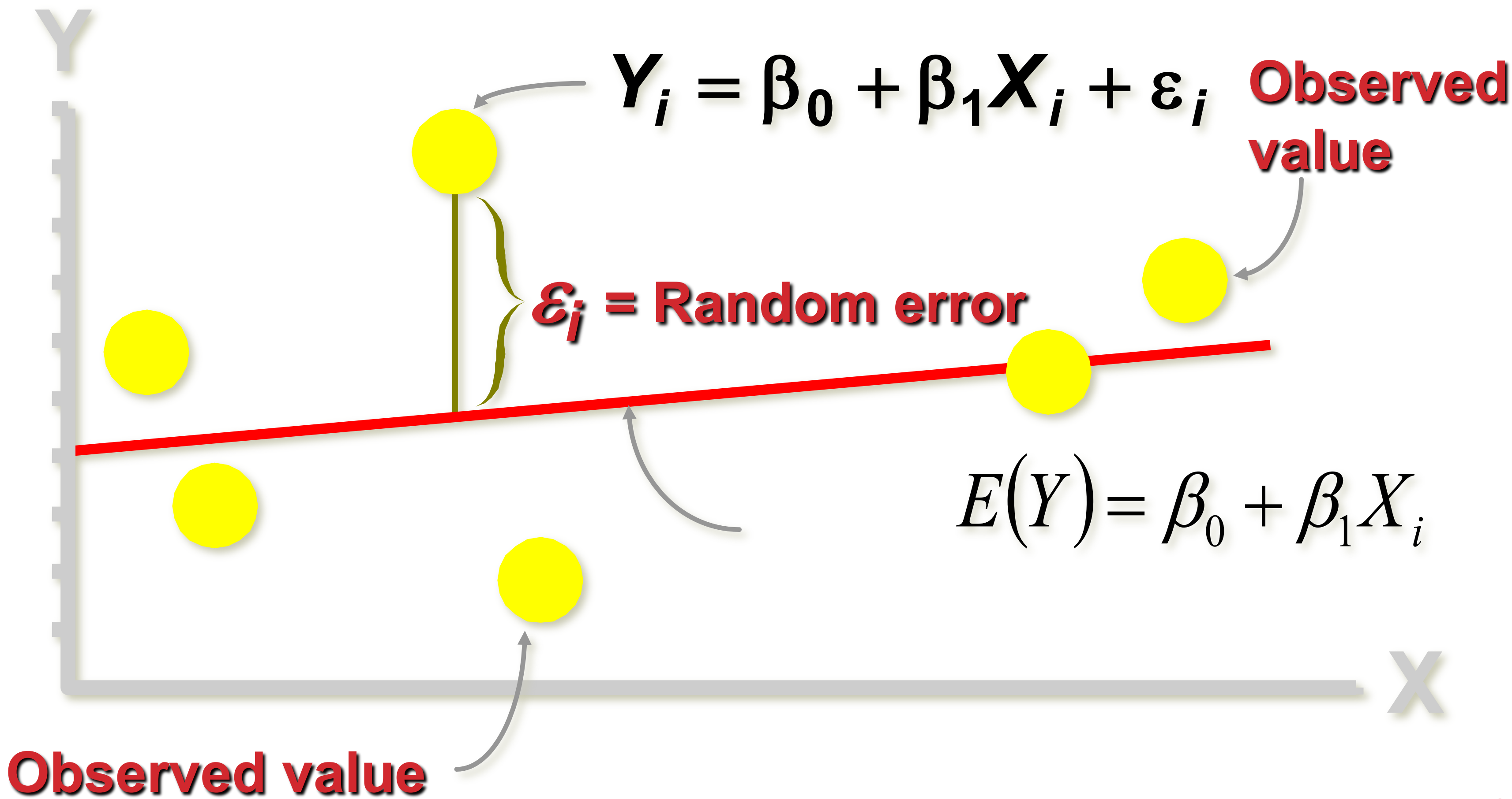
[WF]

POPULATION & SAMPLE REGRESSION MODELS



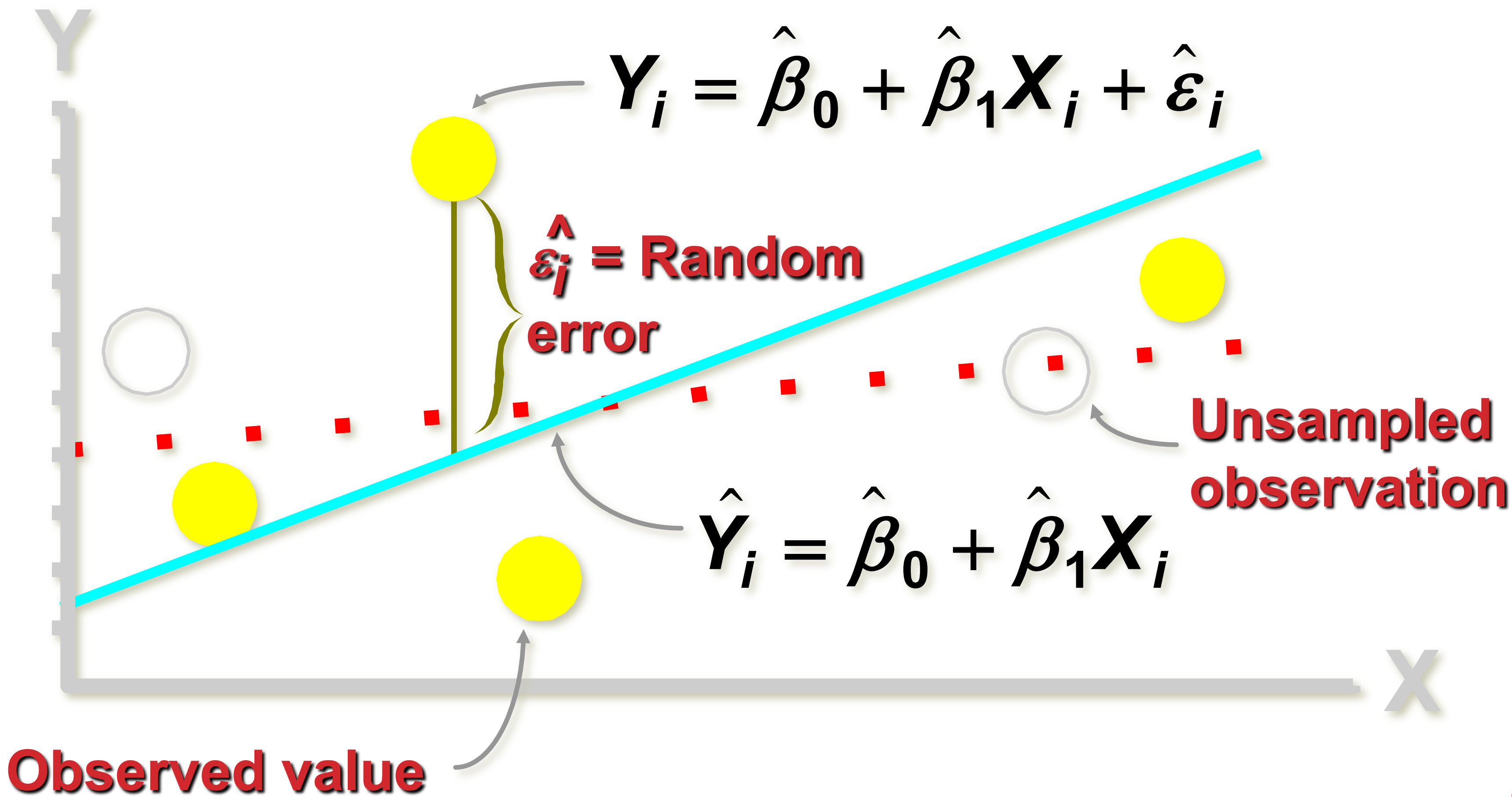
[WF]

LINEAR REGRESSION



[WF]

SAMPLE LINEAR REGRESSION MODEL



[WF]

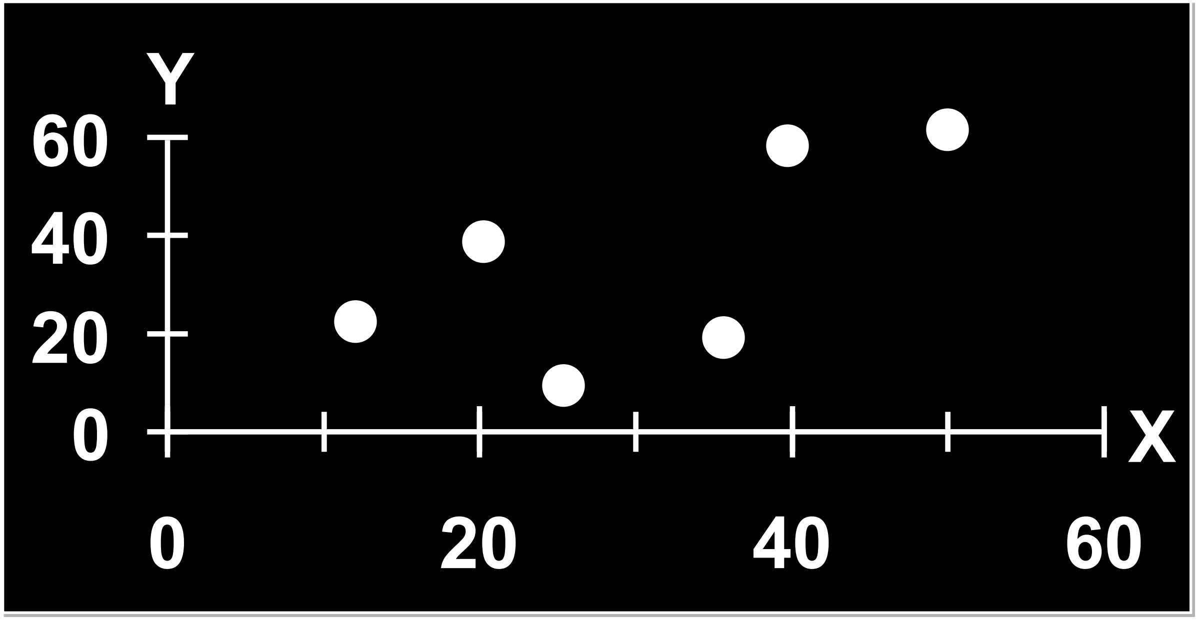
?

ESTIMATING PARAMETERS: LEAST SQUARES METHOD

SCATTER PLOT

Plot all (X_i, Y_i) pairs, and plot your learned model

If you squint, suggests how well the model fits the data



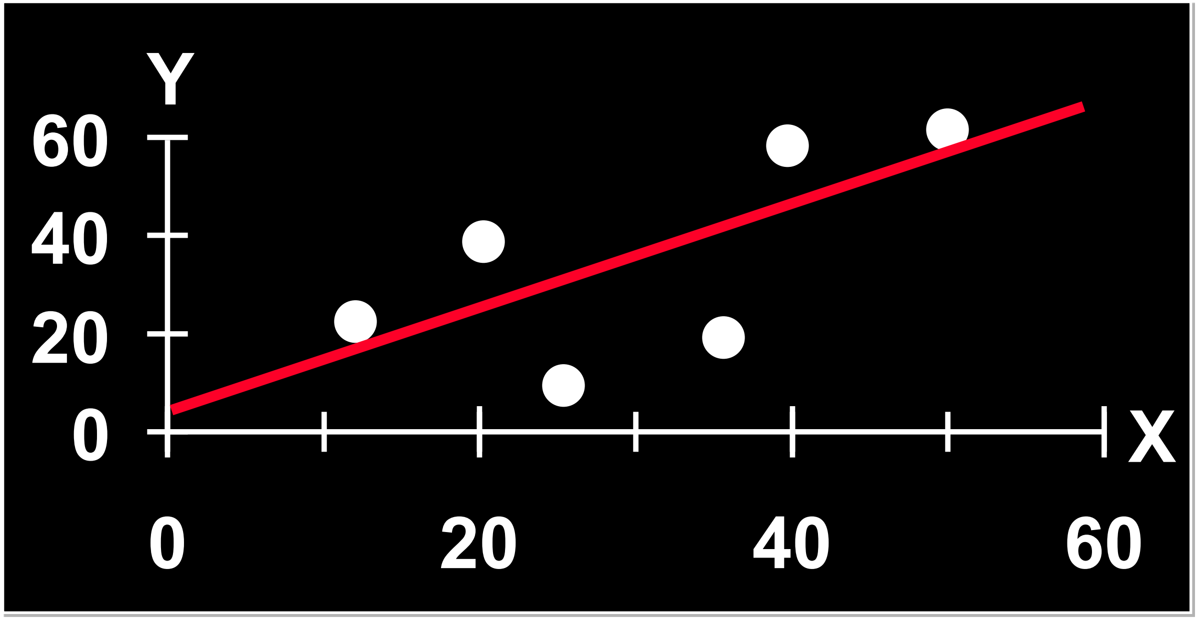
[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” …?

?????????



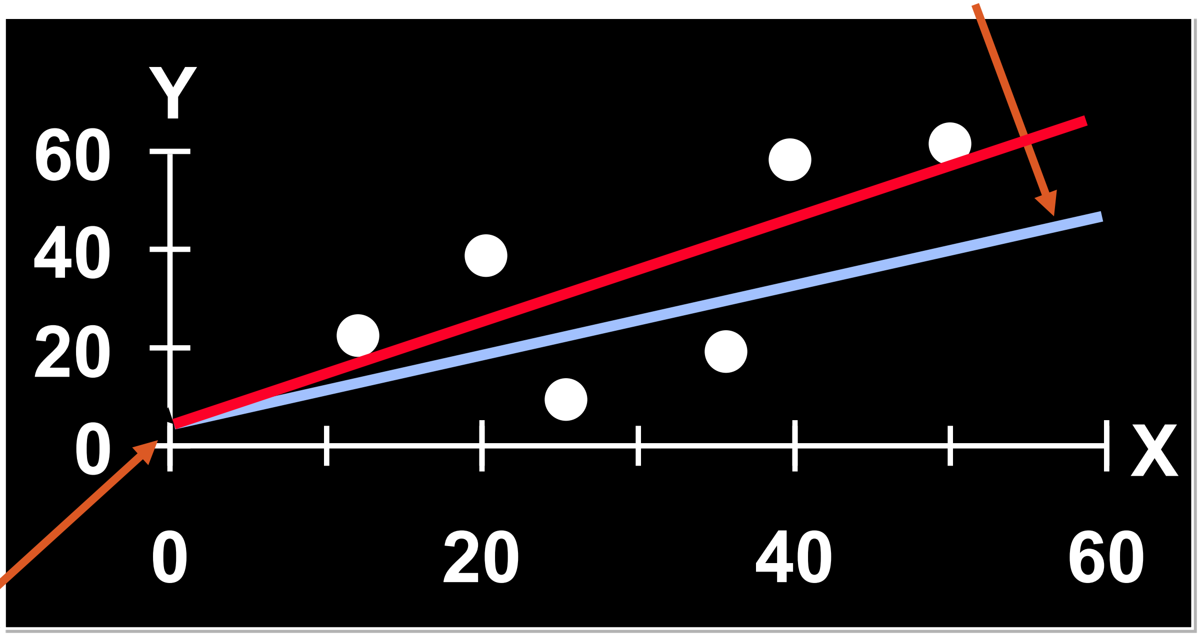
[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line "fits the best" ?????????

Slope changed



Intercept unchanged

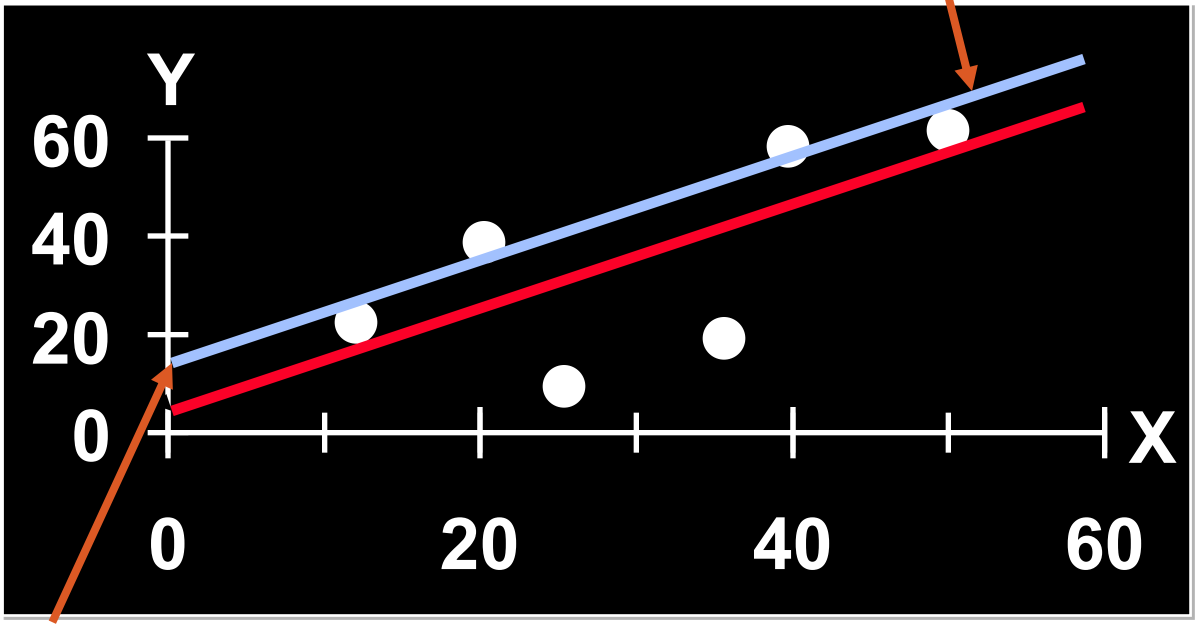
[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” ?????????

Slope unchanged



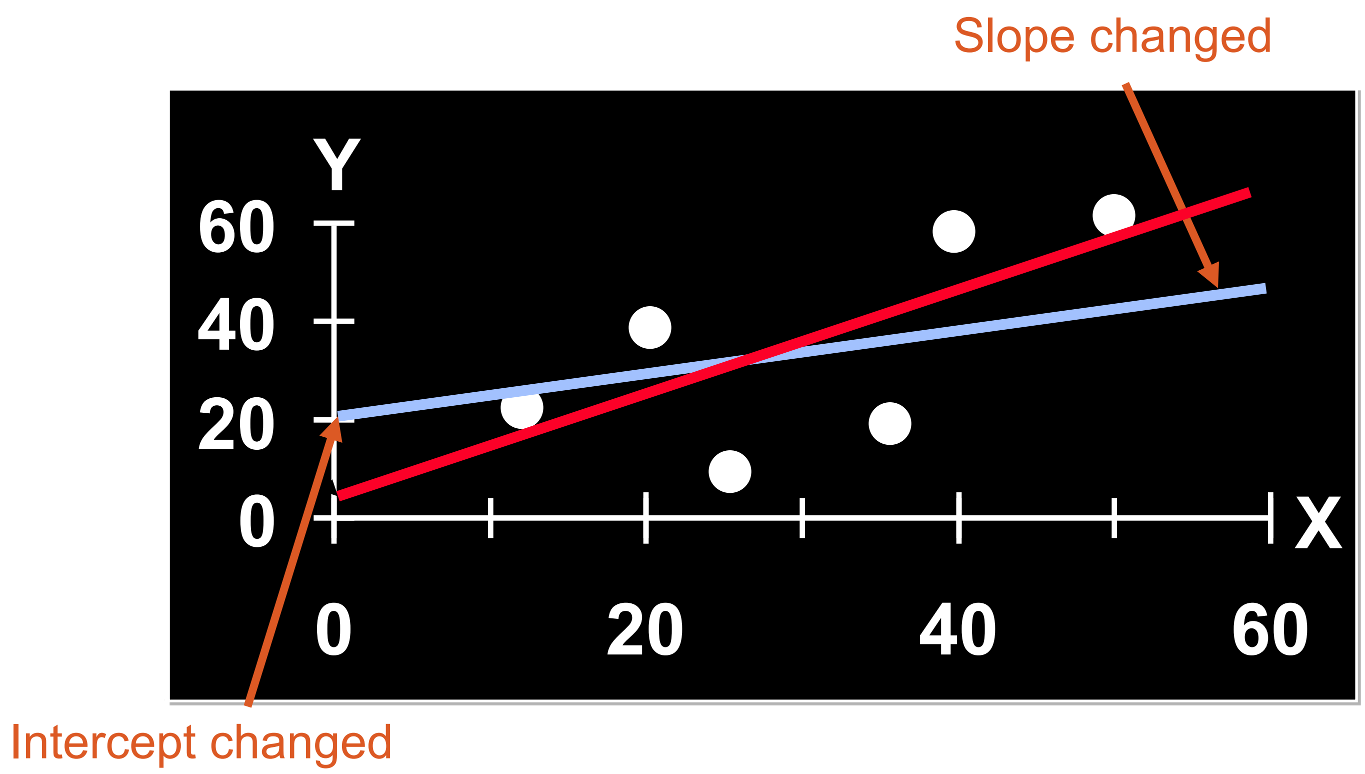
Intercept changed

[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” ?????????



[WF]

LEAST SQUARES

Best fit: difference between the true Y-values and the estimated Y-values is minimized:

- Positive errors offset negative errors …
- … square the error!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

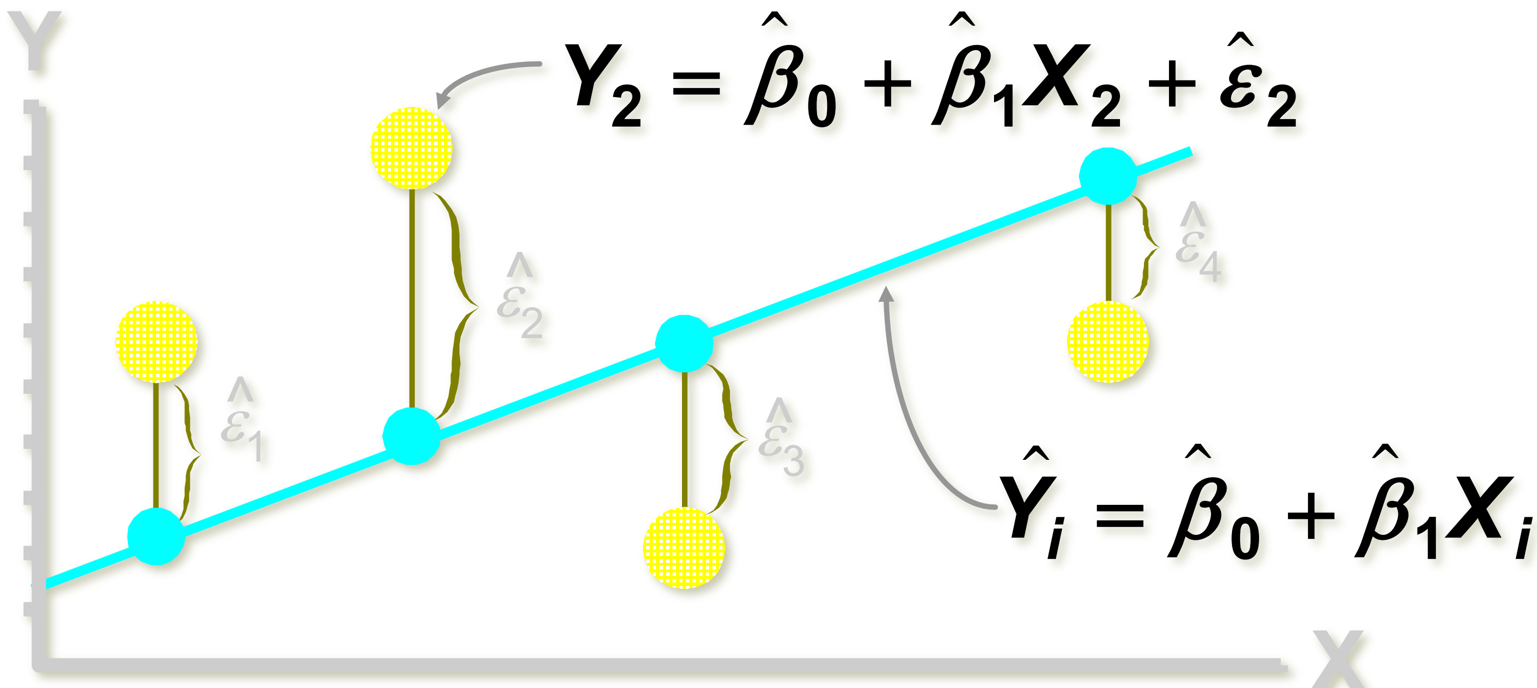
Least squares minimizes the sum of the squared errors

- Why squared? We'll cover this in more depth in March.
- Until then: http://www.benkuhn.net/squared

[WF]

LEAST SQUARES, GRAPHICALLY

$$\text{LS minimizes } \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$$



[WF]

INTERPRETATION OF COEFFICIENTS

Slope ($\hat{\beta}_1$):

- Estimated *Y* changes by $\hat{\beta}_1$ for each unit increase in *X*
- If β_1 = 2, then *Y* Is expected to increase by 2 for each 1 unit increase in *X*

Y-Intercept ($\hat{\beta}_0$)

- Average value of *Y* when *X* = 0
- If $\hat{\beta}_0$ = 4, then average *Y* is expected to be 4 when *X* Is 0

In-depth analysis of less naïve approaches to come!

[WF]

NOW, BACK TO MISSING DATA ...

EXAMPLE

Question: Does the circumference of certain body parts predict BF%?

Assumption: BF% is a linear function of measurements of various body parts and other features …

Analysis: Results from a regression model with BF% …

Predictor	Estimate	S.E.	p-value
Age	0.0626	0.0313	0.0463
Neck	-0.4728	0.2294	0.0403
Forearm	0.45315	0.1979	0.0229
Wrist	-1.6181	0.5323	0.0026

(Interpretation ???????????)

[JA]

WHAT IF DATA WERE MISSING?

In this case, the dataset is complete:

- But what if 5 percent of the participants had missing values? 10 percent? 20 percent?

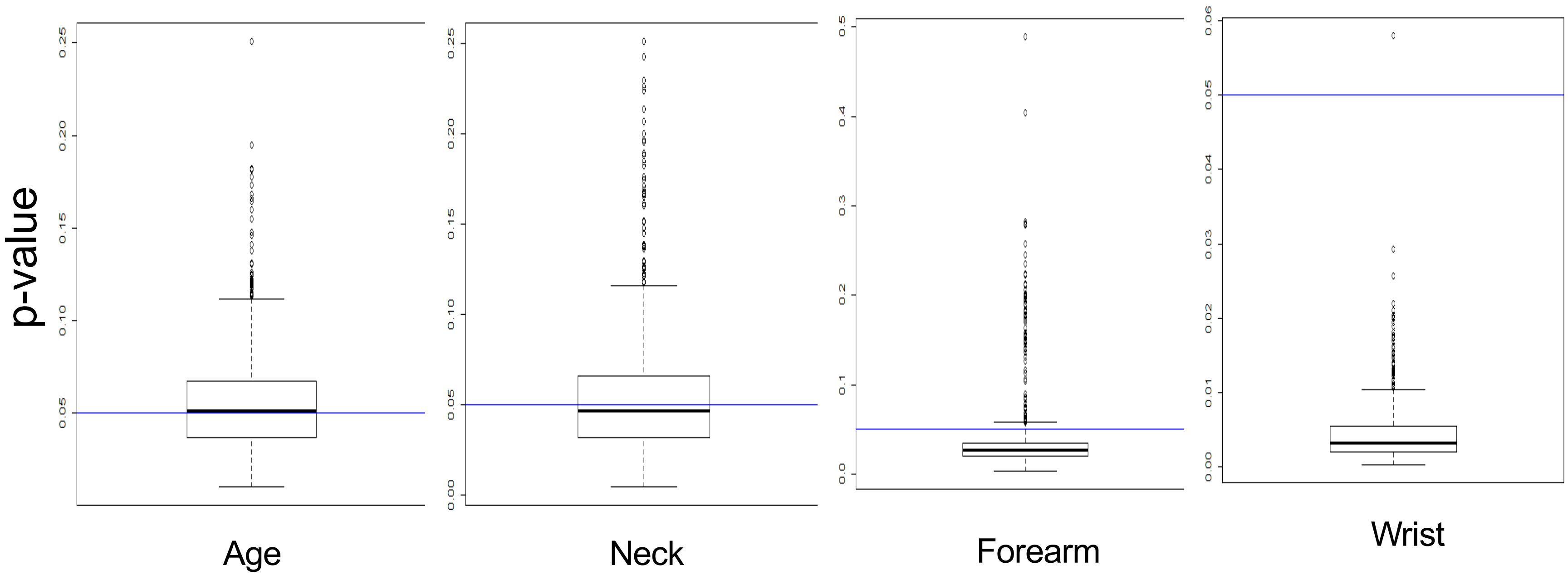
What if we performed complete case analysis and removed those who had missing values?

First let's examine the effect if we do this if when the data is missing completely at random (MCAR)

- Removed cases at random, reran analysis, stored the p-values
- p-value: probability of getting at least as extreme a result as what we observed given that there is no relationship
- Repeat 1000 times, plot p-values …

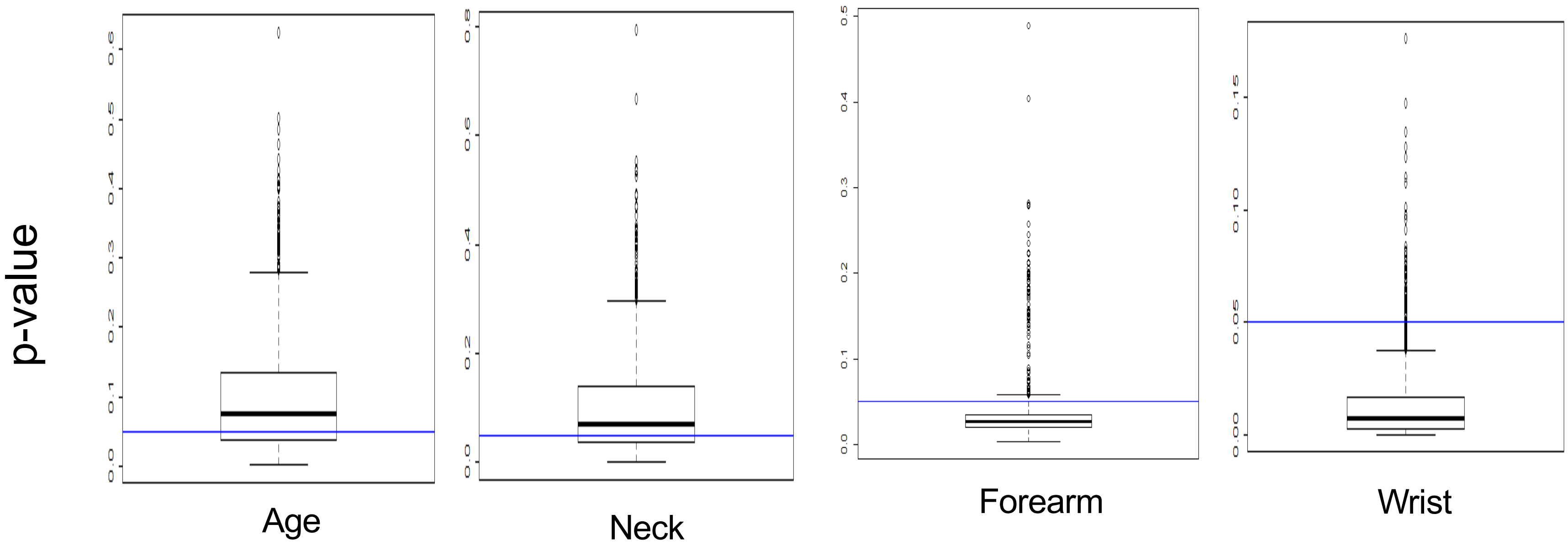
[JA]

~5% DELETED (N=13)



[JA]

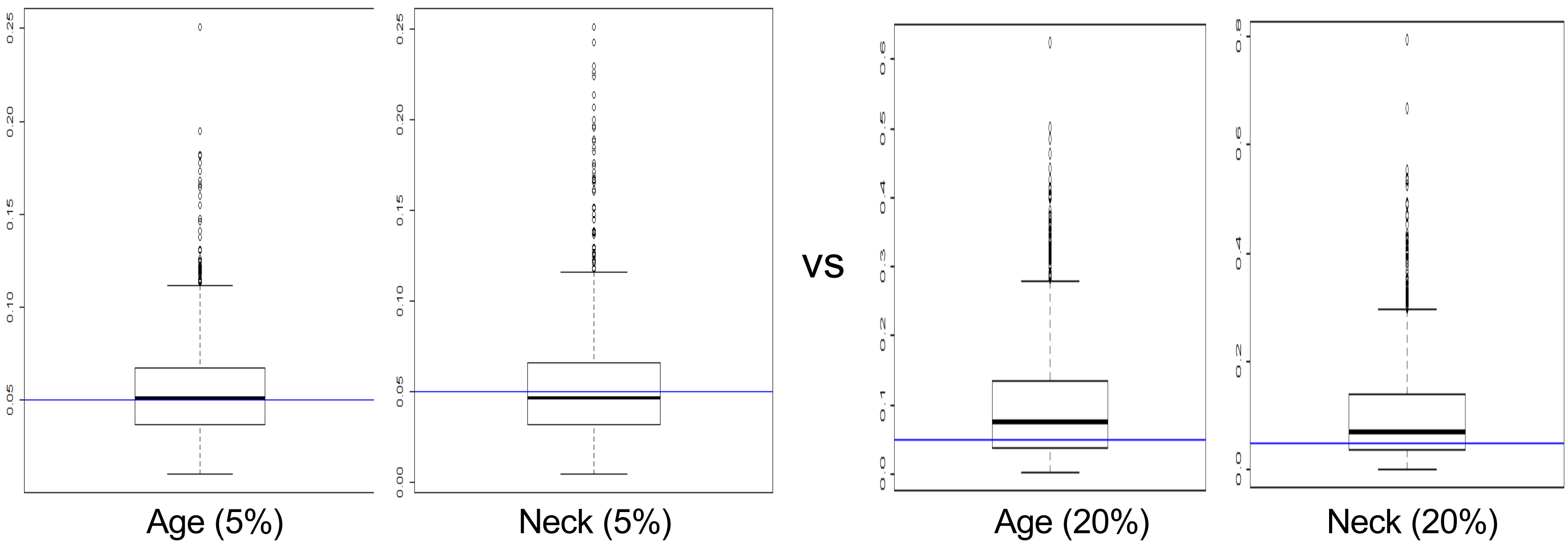
~20% DELETED (N=50)



[JA]

CONCLUSIONS SEEM TO CHANGE …

Age/Neck: fail to reject the null hypothesis usually?



Still reject Forearm/Wrist most of the time

This is assuming the missing subjects' distribtion does not differ from the non-missing. This would cause bias …

TYPES OF MISSING-NESS

Missing Completely at Random (MCAR)

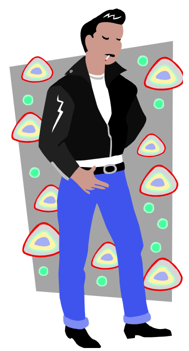
Missing at Random (MAR)

Missing Not at Random (MNAR)

[JA]

WHAT DISTINGUISHES EACH TYPE OF MISSING-NESS?

Suppose you're loitering outside of CSIC one day …



Students just received their mid-semester grades

You start asking passing undergrads their CMSC131 grades

- You don't force them to tell you or anything
- You also write down their gender and hair color

[JA]

YOUR SAMPLE

Hair Color	Gender	Grade
Red	M	A
Brown	F	A
Black	F	B
Black	M	A
Brown	M	
Brown	M	
Brown	F	
Black	M	B
Black	M	B
Brown	F	A
Black	F	
Brown	F	C
Red	M	
Red	F	A
Brown	M	A
Black	M	A

Summary:

- 7 students received As
- 3 students received Bs
- 1 student received a C

Nobody is failing!

- But 5 students did not reveal their grade …

[JA]

WHAT INFLUENCES A DATA POINT'S PRESENCE?

Same dataset, but the values are replaced with a "0" if the data point is observed and "1" if it is not

Question: for any one of these data points, what is the probability that the point is equal to "1" …?

What type of missing-ness do the grades exhibit?

Hair Color	Gender	Grade
0	0	0
0	0	0
0	0	0
0	0	0
0	0	1
0	0	1
0	0	1
0	0	0
0	0	0
0	0	0
0	0	1
0	0	0
0	0	1
0	0	0
0	0	0
0	0	0

[JA]

MCAR: MISSING COMPLETELY AT RANDOM

If this probability is not dependent on any of the data, observed or unobserved, then the data is Missing Completely at Random (MCAR)

Suppose that X is the observed data and Y is the unobserved data. Call our "missing matrix" R

Then, if the data are MCAR, P(R|X,Y) = ??????????

$$P(R|X,Y) = P(R)$$

Probability of those rows missing is independent of anything.

[JA]

TOTALLY REALISTIC MCAR EXAMPLE

You are running an experiment on plants grown in pots, when suddenly you have a nervous breakdown and smash some of the pots

You will probably not choose the plants to smash in a well-defined pattern, such as height age, etc.

Hence, the missing values generated from your act of madness will likely fall into the MCAR category

[JA]

APPLICABILITY OF MCAR

A completely random mechanism for generating missing-ness in your data set just isn't very realistic

Usually, missing data is missing for a reason:

- Maybe older people are less likely to answer web-delivered questions on surveys
- In longitudinal studies people may die before they have completed the entire study
- Companies may be reluctant to reveal financial information

MAR: MISSING AT RANDOM

Missing at Random (MAR): probability of missing data is dependent on the observed data but not the unobserved data

Suppose that X is the observed data and Y is the unobserved data. Call our "missing matrix" R

Then, if the data are MCAR, P(R|X,Y) = ??????????

$$P(R|X,Y) = P(R|X)$$

Not exactly random (in the vernacular sense).

- There is a probabilistic mechanism that is associated with whether the data is missing
- Mechanism takes the observed data as input

EXAMPLES?

?

MAR: KEY POINT

We can model that latent mechanism and compensate for it

Imputation: replacing missing data with substituted values

- Models today will assume MAR

Example: if age is known, you can model missing-ness as a function of age

Whether or not missing data is MAR or the next type, Missing Not at Random (MNAR), is not* testable.

- Requires you to “understand” your data

*unless you can get the missing data (e.g., post-study phone calls)

MNAR: MISSING NOT AT RANDOM

MNAR: missing-ness has something to do with the missing data itself

Examples: ??????????

- Do you binge drink? Do you have a trust fund? Do you use illegal drugs? What is your sexuality? Are you depressed?

Said to be "non-ignorable":

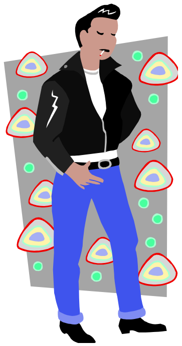
- Missing data mechanism must be considered as you deal with the missing data
- Must include model for why the data are missing, and best guesses as to what the data might be

BACK TO CSIC ...

Is the the missing data:

- MCAR;
- MAR; or
- MNAR?

???????????



Hair Color	Gender	Grade
Red	M	A
Brown	F	A
Black	F	B
Black	M	A
Brown	M	
Brown	M	
Brown	F	
Black	M	B
Black	M	B
Brown	F	A
Black	F	
Brown	F	C
Red	M	
Red	F	A
Brown	M	A
Black	M	A

ADD A VARIABLE

Bring in the GPA:

Does this change anything?

Hair Color	GPA	Gender	Grade
Red	3.4	M	A
Brown	3.6	F	A
Black	3.7	F	B
Black	3.9	M	A
Brown	2.5	M	
Brown	3.2	M	
Brown	3.0	F	
Black	2.9	M	B
Black	3.3	M	B
Brown	4.0	F	A
Black	3.65	F	
Brown	3.4	F	C
Red	2.2	M	
Red	3.8	F	A
Brown	3.8	M	A
Black	3.67	M	A

HANDLING MISSING DATA ...

SINGLE IMPUTATION

Mean imputation: imputing the average from observed cases for all missing values of a variable

Hot-deck imputation: imputing a value from another subject, or "donor," that is most like the subject in terms of observed variables

- Last observation carried forward (LOCF): order the dataset somehow and then fill in a missing value with its neighbor

Cold-deck imputation: bring in other datasets

Old and busted:

- All fundamentally impose too much precision.
- Have uncertainty over what unobserved values actually are
- Developed before cheap computation

MULTIPLE IMPUTATION

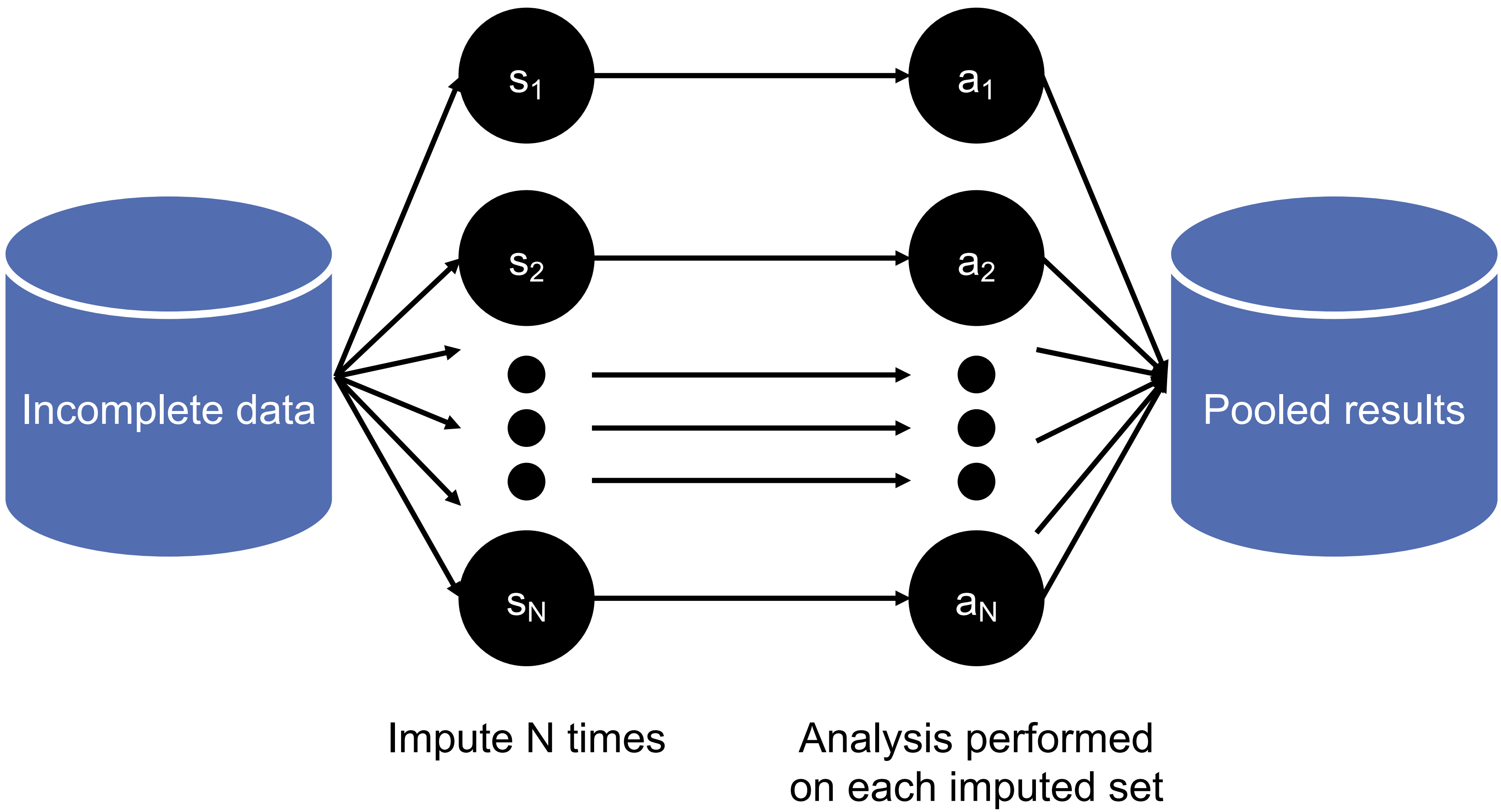
Developed to deal with noise during imputation

- Impute once → treats imputed value as observed

We have uncertainty over what the observed value would have been

Multiple imputation: generate several random values for each missing data point during imputation

IMPUTATION PROCESS



TINY EXAMPLE

X	Y
32	2
43	?
56	6
25	?
84	5

Independent variable: X
Dependent variable: Y
We assume Y has a linear relationship with X

LET'S IMPUTE SOME DATA!

Use a predictive distribution of the missing values:

- Given the observed values, make random draws of the observed values and fill them in.
- Do this N times and make N imputed datasets

X	Y
32	2
43	5.5
56	6
25	8
84	5

X	Y
32	2
43	7.2
56	6
25	1.1
84	5

For very large values of N=2 …

INFERENCE WITH MULTIPLE IMPUTATION

Now that we have our imputed data sets, how do we make use of them?   ???????????

- Analyze each of the separately

X	Y
32	2
43	5.5
56	6
25	8
84	5

Slope	-0.8245
Standard error	6.1845

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

X	Y
32	2
43	7.2
56	6
25	1.1
84	5

Slope	4.932
Standard error	4.287

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

POOLING ANALYSES

Pooled slope estimate is the average of the N imputed estimates

Our example, $\beta_{1p} = \frac{\beta 11+\beta 12}{2}$ = (4.932-.8245) x 0.5 = 2.0538

The pooled slope variance is given by

$$s = \frac{\sum Zi}{m} + (1 + \frac{1}{m}) \times \frac{1}{m-1} * \sum(\beta 1i - \beta_{1p})^2$$

Where Z_i is the standard error of the imputed slopes

Our example: (4.287 + 6.1845)/2 + (3/2)*(16.569) = 30.08925

Standard error: take the square root, and we get 5.485

PREDICTING THE MISSING DATA GIVEN THE OBSERVED DATA

Given events A, B; and P(A) > 0 …

Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) * P(B)}{P(A)}$$

Probability of seeing evidence given the hypothesis

In our case:

$$P(\mathbf{H}|\mathbf{E}) = \frac{P(\mathbf{E}|\mathbf{H}) * P(\mathbf{H})}{P(\mathbf{E})}$$

Prior probability of hypotheses

Posterior probability of the hypothesis given the evidence

Prior over the evidence

BAYESIAN IMPUTATION

Establish a prior distribution:

- **Some distribution of parameters of interest θ before considering the data, $P(\theta)$**
- **We want to estimate θ**

Given θ, can establish a distribution $P(X_{obs}|\theta)$

Use Bayes Theorem to establish $P(\theta|X_{obs})$ …

- **Make random draws for θ**
- **Use these draws to make predictions of Y_{miss}**

HOW BIG SHOULD N BE?

Number of imputations N depends on:

- Size of dataset
- Amount of missing data in the dataset

Some previous research indicated that a small N is sufficient for efficiency of the estimates, based on:

- $(1 + \frac{\lambda}{N})$-1
- N is the number of imputations and λ is the fraction of missing information for the term being estimated [Schaffer 1999]

More recent research claims that a good N is actually higher in order to achieve higher power [Graham et al. 2007]

MORE ADVANCED METHODS

Interested? Further reading:

- Regression-based MI methods
- Multiple Imputation Chained Equations (MICE) or Fully Conditional Specification (FCS)
  - Readable summary from JHU School of Public Health: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
- Markov Chain Monte Carlo (MCMC)
  - We'll cover this a bit, but also check out CMSC643!

NEXT CLASS:

SUMMARY STATISTICS &VISUALIZATION