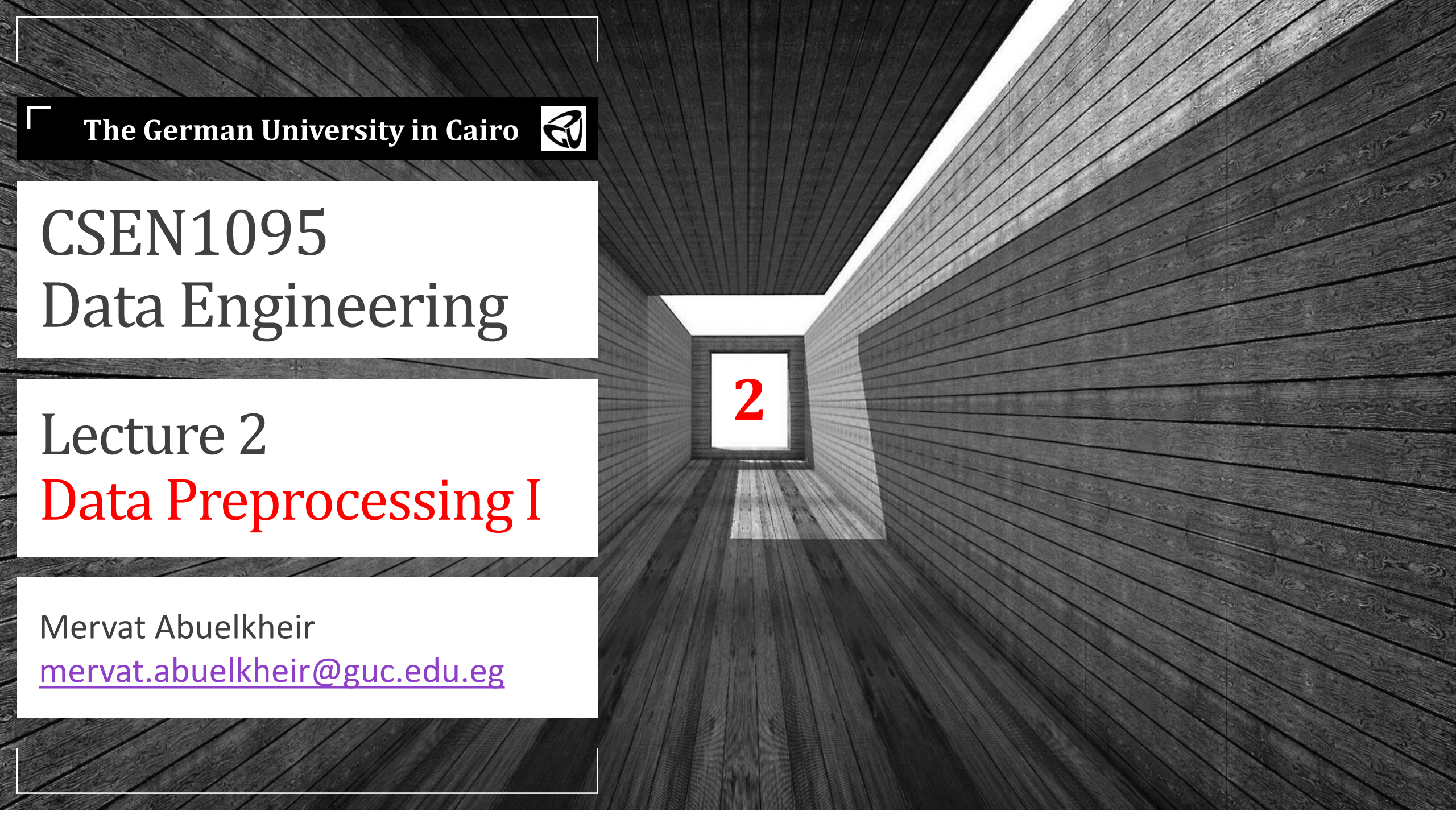
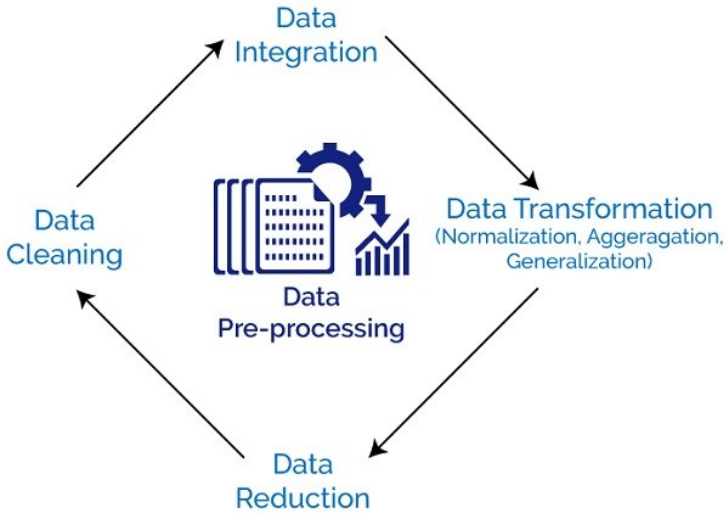
The German University in Cairo

CSEN1095
Data Engineering



Lecture 2
Data Preprocessing I

Mervat Abuelkheir
mervat.abuelkheir@guc.edu.eg



Why Preprocess Data?

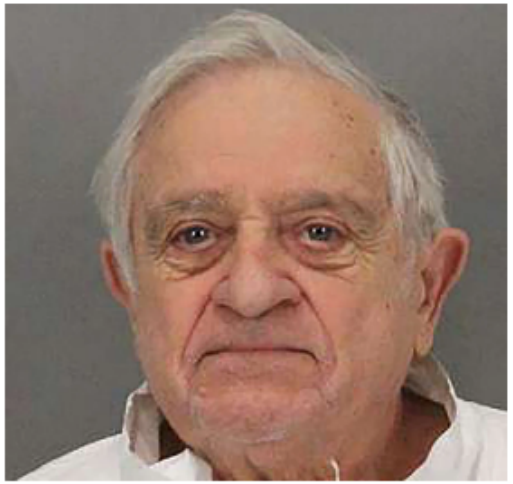
GUC

Police Use Fitbit Data to Charge 90-Year-Old Man in Stepdaughter's Killing

By Christine Hauser

Oct. 3, 2018

https://www.nytimes.com/2018/10/03/us/fitbit-murder-arrest.html



Anthony Aiello, 90, has been charged with murder in the death of his stepdaughter in San Jose, Calif., the police said.
San Jose Police Department

The last time Anthony Aiello spoke to his stepdaughter, he took homemade pizza and biscotti to her house in San Jose, Calif., for a brief visit. Mr. Aiello, 90, told investigators that she then walked him to the door and handed him two roses in gratitude.

But an unnoticed observer in the house later revealed that their encounter ended in murder, a police report said.

When Ms. Navarra's Fitbit data was compared with video surveillance from her home, the police report said, the police discovered that the car Mr. Aiello had driven was still there when her heart rate stopped being recorded by her Fitbit.

Bloodstained clothes were later found in Mr. Aiello's home, the document said. He was arrested on Sept. 25.

Mr. Aiello was "confronted" with the Fitbit information during questioning, said Brian Meeker, a San Jose police detective. "After explaining the abilities of the Fitbit to record time, physical movement, and heart rate data, he was informed that the victim was deceased prior to his leaving the house," Detective Meeker said in the report.

Mr. Aiello said that could not be true, insisting Ms. Navarra had walked him to the door, and he suggested that someone else could have been in the home, the report said.

"I explained that both systems were on internet time, and there was no deviation," Detective Meeker said.

After they finished their questions, detectives left Mr. Aiello alone in the interview room. He began talking to himself, the report said, saying repeatedly, "I'm done."

year worldwide.

As more people used the devices, it was inevitable that they would be worn by victims or suspects in crimes and potentially hold tantalizing clues or even plausible answers: Does a suspect's alibi of being at home asleep hold up? Does a victim's steady heart rate at the time of an alleged attack suggest the charge was fabricated?

Using trackers this way, of course, assumes that the devices are accurate—and not just accurate on average, but at very specific moments in time, a sort of black box for the body that reveals physiological truths that its wearer might prefer to conceal. Research on fitness trackers, however, shows they don't always perfectly mirror reality. An analysis of 67 studies on Fitbit's movement tracking concluded that the device worked best on able-bodied adults walking at typical speeds. Even then, the devices weren't perfect—they got within 10 percent of the actual number of steps a person took half of the time—and became even less accurate in counting steps when someone was resting their wrist on a walker or stroller, for example. "It's not measuring actual behavior," says Lynne Feehan, a clinical associate professor at the University of British Columbia and the lead researcher on the paper. "It's interpreting motion."

Many fitness tracker users experience moments of misinterpretation: the piano playing session that was categorized as cycling; the times during sweaty exercise when it stops picking up a heart rate. Even Fitbit's own terms of service point out that it is a consumer product with accuracy that is "not intended to match that of medical devices or scientific measurement devices."

https://www.wired.com/story/telltale-heart-fitbit-murder/

Smartwatches decipher heart rate using green LEDs that beam hundreds of times per second into capillaries through the skin. Those capillaries allow in more of the light when full of blood, and less between beats, and the device measures how much light is absorbed. That measurement is then siphoned through a proprietary algorithm to generate a heart rate figure. University of Wisconsin researchers looked at how well wrist-worn fitness trackers measured heart rate, comparing it to an electrocardiograph, the gold standard for heart monitoring. They found that the fitness trackers' heart rate deviated more from the actual rate when a subject exercised on a treadmill than when at rest. (Fitbit won't talk specifics about its accuracy, saying in a statement, "We are confident in the performance of all our devices" and that the company continues to test them.)

Tony's defense lawyers signaled that they would attack the reliability of the Fitbit data. They assembled a grab bag of disqualifications: They said Karen wore the device for only two weeks or less, and it hadn't yet normalized to her signal; they said that Fitbit, which assigns a confidence score of 0 to 3 to its data collection, at times assigns zero confidence to the data on Karen's device on the day the prosecution says she was murdered; and Edward Caden, one of the defense attorneys, said that what the prosecution calls a "spike" in Karen's heart rate is more like "a pimple." Caden even asserted that there were moments after 3:28 pm when Karen's Fitbit seems to still report heartbeat data.

Angela Bernhard, the chief trial deputy for Santa Clara County, told me in August that she expected that the defense would "be fighting to keep out a lot of the evidence that we want in" and that she intended to present the Fitbit evidence at trial. "Ultimately it's up to the judge what evidence gets brought in and what doesn't," she said. At a grand jury hearing in August, Bonham, the Fitbit executive, testified that Fitbit had turned over a voluminous Excel spreadsheet of Karen's raw heartbeat and step data. He also clarified that a confidence rating of zero means the device isn't registering a heartbeat at all, and detectives say that Karen's device showed no heartbeat and zero confidence at 3:28 pm and after. Detective Meeker testified to the reliability of Karen's device specifically: At two times in early September that Karen was visible on surveillance footage walking in stores, her Fitbit recorded movement. (Fitbit declined to comment on Aiello's case.)

Data Challenges

- Massive data
- Curse of dimensionality (high-dimensional problem in terms of features)
- Missing data values (sometimes not missing at random)
- Wrong data values (needs detection and correction)
- Sometimes data is not factual (yet not technically wrong!) and we have a complicated set of factors that affect user-provided data values

Source: CS109 Stanford's Data Science Course

Why Preprocess Data?

To improve the quality of data (*usability* and *reliability*) and make it suitable for the requirements of the intended use

- Factors of data quality
  - *Accuracy* → lack of is due to faulty instruments, errors caused by human/computer/transmission, deliberate errors …
  - *Completeness* → lack of is due to data acquired over different design phases, optional attributes
  - *Consistency* → lack of is due to semantics, data types, field formats …
  - *Timeliness* → data should be current, and available when it's needed by user
  - *Integrity* →lack of due to poor definitions of data relationships
  - *Interpretability* → how easy the data is understood

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability?
- Accuracy?
- Integrity?
- Completeness?
- Consistency?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y|−|0.0|1000

TedJ.|973−360−8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?
- Completeness?

Example 1

Two records from a pipe-delimited file:

T.Das|97336o8327|24.95|Y| – |0.0|1000

TedJ.|973 – 360 – 8779|2000|N|M|NY|1000

- Interpretability? name, phone number, revenue, indicator, gender, state, usage
- Accuracy?
- Integrity?
- Completeness?
- Consistency?

Data Preparation is itself a Challenge!

- Data quality problems are highly complex and context dependent
  - Extensive domain knowledge is needed
  - Solutions need to be chosen case by case
- No single tool can solve a majority of data quality problems!
- Do not apply data preprocessing methods manually – **AUTOMATE**
  - For large datasets
  - Reuse of code for similar issues

Major Preparation Tasks
That Improve Quality of Data

- ***Data cleaning*** → filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies
- ***Data transformation*** → normalization, discretization
- ***Data reduction*** → obtain a reduced representation of the data set that is much smaller in volume, while producing almost the same analytical results
- ***Data integration*** → include data from multiple sources in analysis, map semantic concepts, infer attributes …

Detailed Preparation Tasks

- Data cleaning
  - Impute/remove missing values
  - Detect and remove/handle noise/outliers
  - Detect and remove/handle inconsistencies
- Data transformation
  - Scale attributes with varying ranges
  - Normalize distributions
  - Encode categorical attributes
  - Discretize numerical attributes
- Data reduction
  - Partition/sample data
  - Aggregate records
  - Extract/select/fuse features
- Feature engineering
  - Scale, normalize, discretize and encode features
  - Extract/select/fuse features
  - Add new features
- Data integration
  - Handle data from multiple sources
  - Detect and remove/handle inconsistencies

Data Cleaning

GUC

Data Cleaning

- Data in the Real World Is Dirty!
  - *Incomplete*: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g. *Occupation*=“ ” (missing data)
  - *Noisy*: containing noise, errors, or outliers
    - e.g. *Salary*=“−10” (an error)
    - Intentional → Jan. 1 as everyone’s birthday?
  - *Inconsistent*: containing discrepancies in codes or names
    - *Age*=“42” and *Birthday*=“03/07/2010”
    - Rating was “1, 2, 3”, now is “A, B, C”
    - Switched fields!
    - Censored/defaulted/maxed values

Data Cleaning Involves …

… checking for and imposing **semantic** and **syntactic** validity of raw data

Through …

- filling in *missing values*
  - via missing value **imputation**
- smoothing out *noise* and identifying *outliers*
  - via **binning**, **regression**, **outlier detection**, clustering, **discretization**
- correcting *inconsistencies* in the data
  - via fuzzy joins, **regular expressions**, database profiling, metadata

But **CAREFUL!**

- Methods that fix data problems and transforms raw data to clean data must be **reproducible**
- Non-reproducible transformation cannot be distinguished from invention!
- You must preserve your raw data in their original form!
- You need a **recipe** – **working code that tracks the fully defined operations necessary to produce your clean data from your raw data**
  - Can be put into production in operational scenarios

Missing Data Values

- **Why do we care?** Why not throw away the data points with missing values?
  - Loss of information (obviously)
  - Potential of bias in analysis results for remaining data
    - **Why is data missing in the first place?**
  - **Patterns of missing data can actually mean something (missing data mechanism)**
    - **and will affect the choice of imputation method!**
- A missing value may not imply incomplete data!
  - e.g. driver's license number
- A zero may be a true value and may be a default value due to attribute constraint rules
  - e.g. zero-quantity bill versus default-zero bill that is not yet processed

Detecting Missing Values

- Scan for gaps in rows and columns
- Cross-check data schema with actual data for missing attributes
- Check data during transit for losses due to transfer/communication process
- Check history of data source
- Keep track of estimated values and error bounds (counts, means, SDs)
  - e.g. per-unit-time transactions received from a store, packets from a router, cars per area
  - Deviation when receiving new records may indicate missing records
- How to detect confounding defaults?

Types of Missing Values

- Missing Completely At Random (MCAR) – there is **no relationship** between the missing data mechanism and any values, observed or missing
  - Probability of being missing is the same for all values
  - e.g. if weighing scale ran out of battery, weight attribute has values MCAR
- Missing At Random (MAR) – there is a ***systematic*** **relationship** between the **propensity** of missing values and the *observed* data, but ***not*** the missing data
  - missingness can be explained by variables on which you have full information
  - e.g. if men are more likely to tell their *weight* than women, weight is MAR (what is the observed variable here?)
- MCAR and MAR are ***ignorable*** – enough information is available in the data to allow imputing missing values, therefore the missing data mechanism can be ignored

Types of Missing Values

- Missing Not at Random (MNAR) – there is **a relationship** between the **propensity** of a value to be missing and its values
  - Probability of being missing varies for reasons that are unknown to us
  - e.g. people with the lowest education have missing education level in education attribute, the sickest people are most likely to drop out of a medical study, a device measures some response and can only measure values above 0.5 so any value below will be missing
- MNAR is ***non-ignorable*** – the missingness mechanism has to be modeled explicitly as you deal with the missing data
- How can you assess which type of missing data you have?

		Data Explains Pattern	
		Yes	No
Missingness Pattern	Yes	MAR	MNAR
	No	---	MCAR

Handling Missing Values – Simplistic Methods

- *Fill in the missing value manually* → time consuming, not feasible for large datasets
- *Use a global constant* → replace all missing attribute values by same value (e.g. *unknown, null*)
  - Analysis task may mistakenly think that "*null*" is an interesting concept
  - Careful not to use a global constant that is a valid value of the attribute!
    - e.g. using "zero" for numerical attribute whose values can in fact include zero
- *Listwise deletion* → an object is deleted if it's missing data on *any* attribute in the analysis
  - Default, not very effective, unless:
    - Object has several attributes with missing values and only few objects have missing values – you have **statistical power**
    - **Assumes MCAR**

Handling Missing Values – Imputation

- **Goal is to replace missing values with plausible values!**
- This implies a level of **uncertainty** in the imputed values
- Do not assume in the analysis that imputed values are the real values
- **Imputation should be used with caution!**
  - Imputed values are good for aggregate analysis
  - No individual imputed value should be trusted

Handling Missing Values – Univariate Imputation

- *Mean/Average imputation* → replace missing observation with mean of non-missing observations
  - for normal (symmetric) data distributions, mean is used, skewed distribution should employ median
  - Standard error of attribute will be underestimated **(imputing the mean preserves the mean!)**
- *Use value drawn from distribution of non-missing values for the attribute* → assuming missing values follow same distribution of available values

- If missing data is 2-3% it is ***ok*** to use the above methods
- Univariate imputation is a ***poor choice*** for some ML algorithms (e.g. LR)

Handling Missing Values – Multivariate Imputation

- *Use mean or median for all samples belonging to the same class as the given object*
  - e.g. mean or median of customers in a certain age group, customers who default, …
- *Use the most probable (estimated) value* → e.g. using regression or Bayesian inference
  - Still may maintain small standard error if applied once, as parameters do not change

Handling Missing Values – Linear Regression

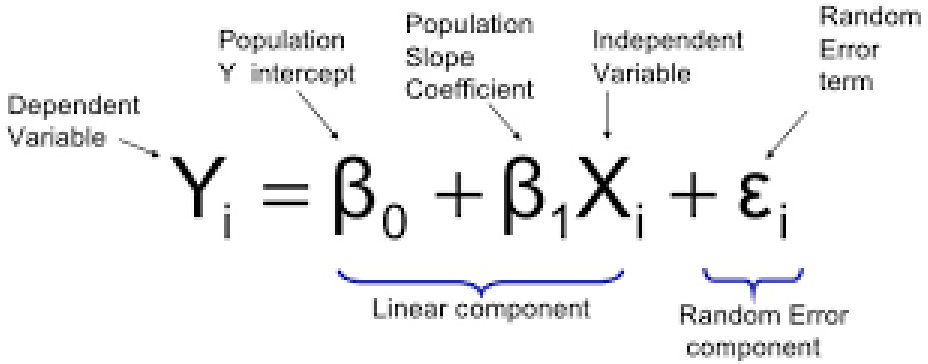
- **Regression analysis** attempts to determine the strength of the relationship between two (or more) variables:
  - A dependent variable (usually denoted by Y)
  - One or more changing variables (known as independent variables or predictors)
  - In imputation, the dependent variable is the one with missing values
- Linear Regression (LR) → dependent variable is continuous
- **Multiple Regression** → multiple independent variables

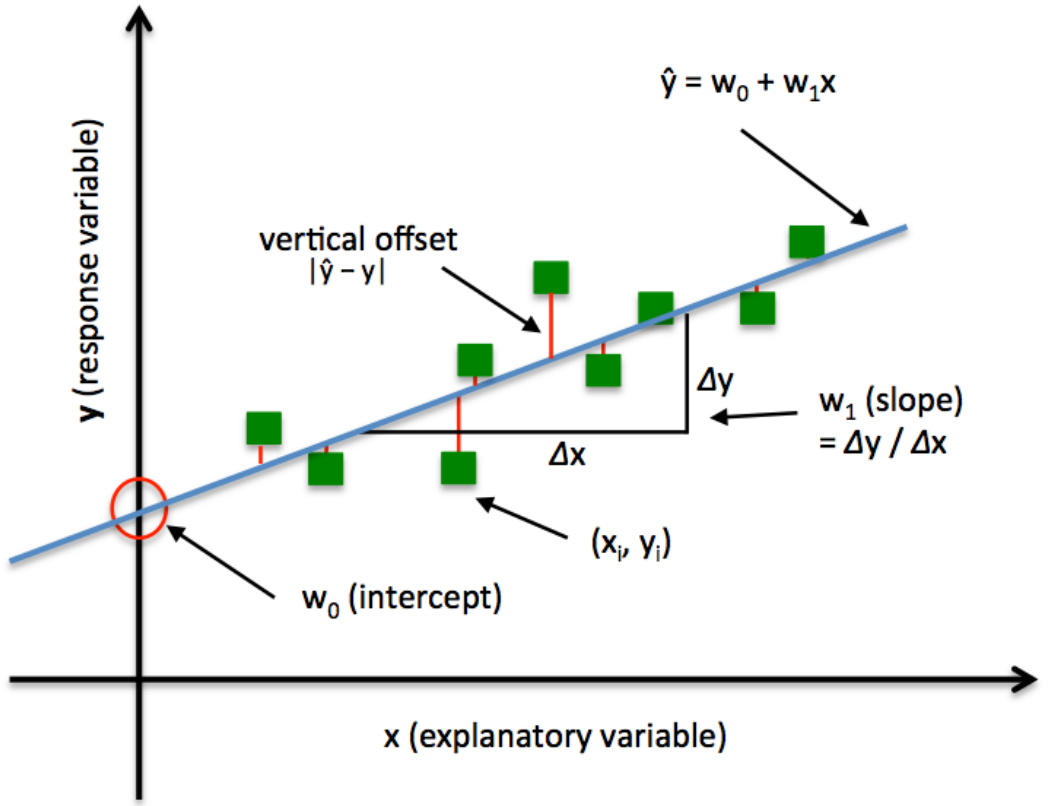
- Regression types other than linear
  - Logistic Regression → dependent variable is binary
  - Multinomial Logistic Regression → dependent variable is categorical

Brief Overview of The Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

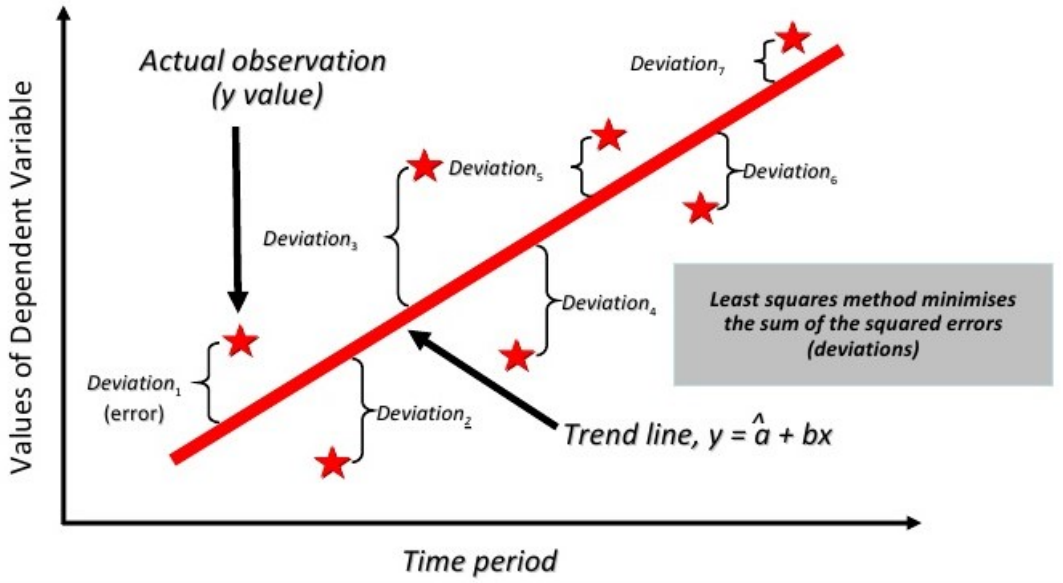
- β_0 is the intercept
- β_j is the slope of the j^{th} variable → the average increase in Y when X_j increases by one and all other Xs are held **constant**
- The β_js are the model parameters we need to find





Brief Overview of The Linear Regression Model

- **Regression error** → distance between original data points and regression line
- The distances are called the **error terms**, or the **residuals**
- How to find the best regression model
  - **Gradient Descent**



Linear Regression for Imputation – Example

- For attributes (weight, age, height), height is missing, so we use regression formula
  - $\boldsymbol{height} = \beta_0 + \beta_1 \boldsymbol{age} + \beta_2 \boldsymbol{weight}$
  - $\boldsymbol{height} = 8.33 + 0.167\boldsymbol{age} + 0.1\boldsymbol{weight}$

Weight	Age	Height	Health_Index
20	2	10	5
15	5	9	3
25	5	10	
20	4		
10	1		

Monotone missing values

Linear Regression for Imputation – Example (Cont.)

- For attributes (weight, age, height), height is missing, so we use regression formula
  - $\boldsymbol{height} = \beta_0 + \beta_1 \boldsymbol{age} + \beta_2 \boldsymbol{weight}$
  - $\boldsymbol{height} = 8.33 + 0.167\boldsymbol{age} + 0.1\boldsymbol{weight}$
- Since more than one attribute has missing values, use regression formula for attributes with missing values monotonically
  - $\boldsymbol{health\_index} = \beta_0 + \beta_1 \boldsymbol{age} + \beta_2 \boldsymbol{weight} + \beta_3 \boldsymbol{height}$

Weight	Age	Height	Health_Index
20	2	10	5
15	5	9	3
25	5	10	
20	4	10.998	
10	1	9.497	

Monotone missing values

- Regression analysis rule: for k parameters, you must have $N \geq k$ distinct data points

Multivariate Imputation by Chained Equations – MICE

1. Perform a simple (e.g. mean) imputation for every missing value in dataset
   - Those are called “placeholders”
2. Choose one variable/attribute with missing values, and set the placeholders back to missing
3. Regress this attribute on the other attributes with their observed and placeholder values
4. Impute missing values of attribute with regressed values
5. Repeat steps 2-4 for each other variable/attribute with missing values

Weight	Age	Height	Health_Index
20	2	10	5
15	5	9	3
25	5	10	v_{r3}
20	4	v_{r1}	v_{r4}
10	1	v_{r2}	v_{r5}

Multivariate Imputation by Chained Equations

Linear Regression for Multiple Imputation

- An error term is generated at random from a Gaussian distribution, so you can generate multiple estimates for the missing values with randomized error components, and construct multiple datasets for multiple imputation
- Better yet, repeat MICE (steps 2-4) for a number of cycles (default is 10) and update imputations after each cycle

Weight	Age	Height	Health_Index
20	2	10	5
15	5	9	3
25	5	10	v_{r3}
20	4	v_{r1}	v_{r4}
10	1	v_{r2}	v_{r5}

Multivariate Imputation by Chained Equations

GUC

Thank You