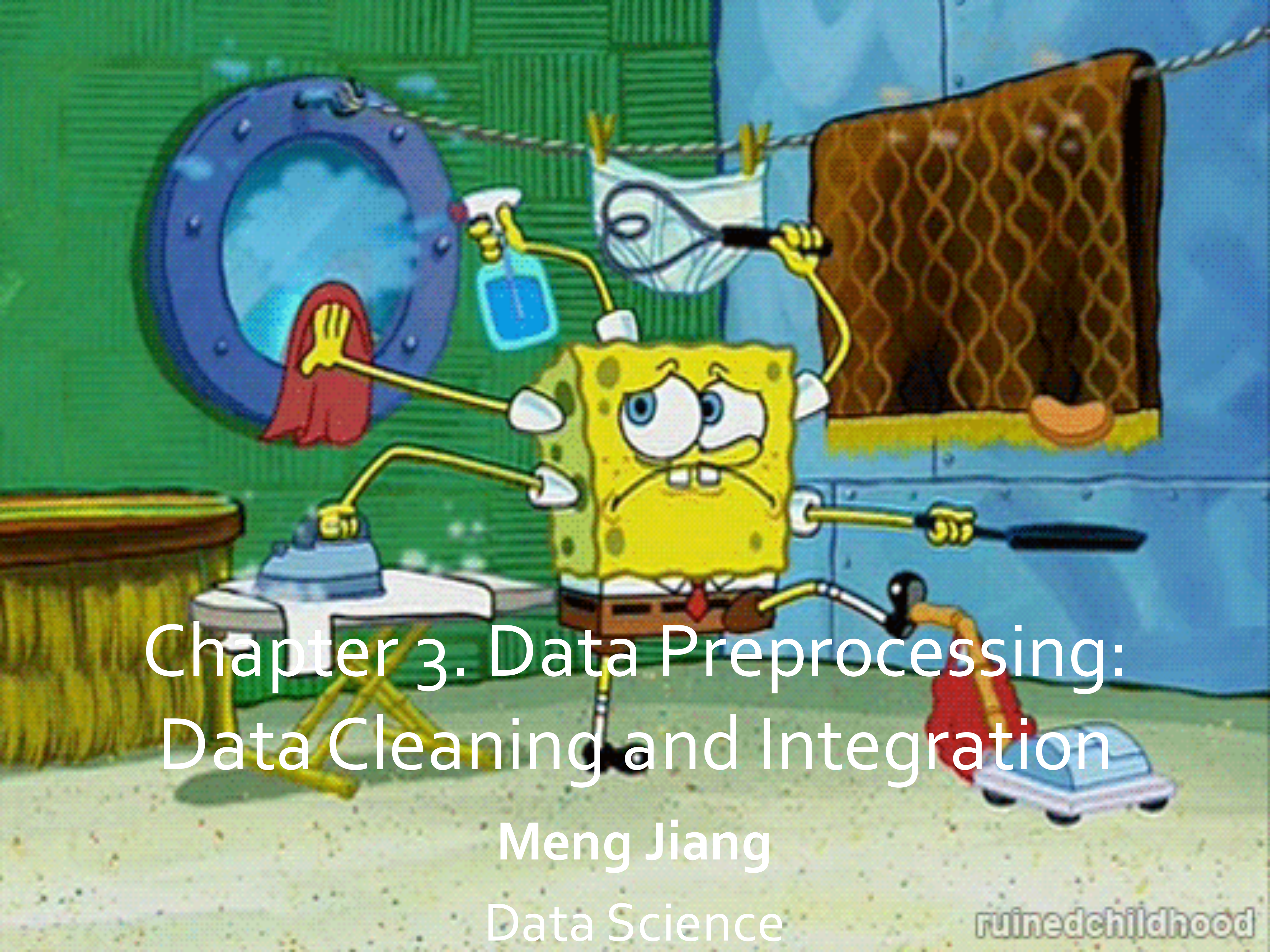
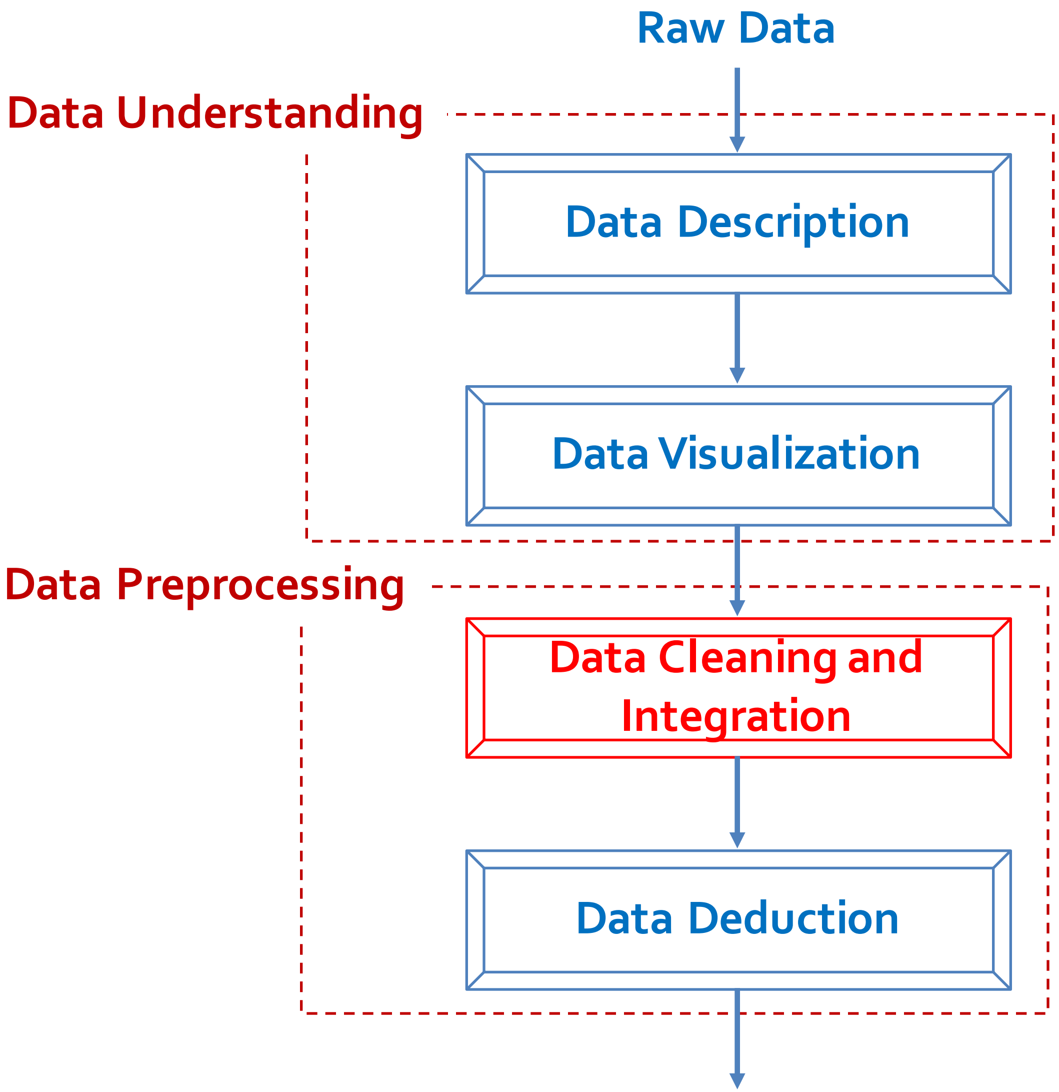


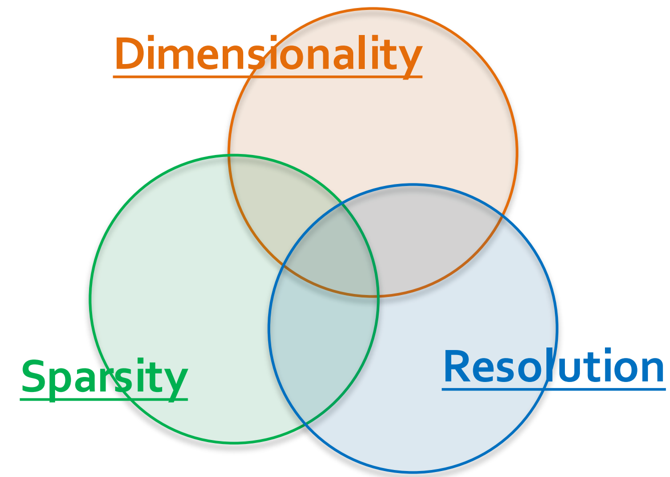
Chapter 3. Data Preprocessing: Data Cleaning and Integration

Meng Jiang

Data Science



Describing data:



Today: Data Cleaning and Integration

- **Understand data quality issues and how to handle them**
  - Describe three types of missing data
  - Describe how to handle missing data, noisy data, inconsistent data, and redundant data
- **Correlation analysis** for handling data redundancy
  - Categorical Variables: Chi-square test
  - Numerical Variables: Covariance analysis

Quality Issues in Collecting Data

- Suppose we want to build a student profile database. In the **next 5 lectures**, the instructor will ask you to write down your answers:
  - Name
  - Dorm
  - Height
  - Weight
  - Hometown
  - Major
  - Hobby
  - Intended job after graduation

Why next five? Not just one?

Quality Issues in Collecting Data

If you don't want to write down your height or weight…

If you write down you are 12 feet high…

If in the first lecture you say you are from California and in the third you submit your hometown as Florida…

If in the first lecture you say you are from California and in the third you submit your hometown as C.A. …

If you were not in class but your classmate writes down you are from California but actually you are from Florida…

Sparse data

Incomplete data

Noisy data

Inconsistent data

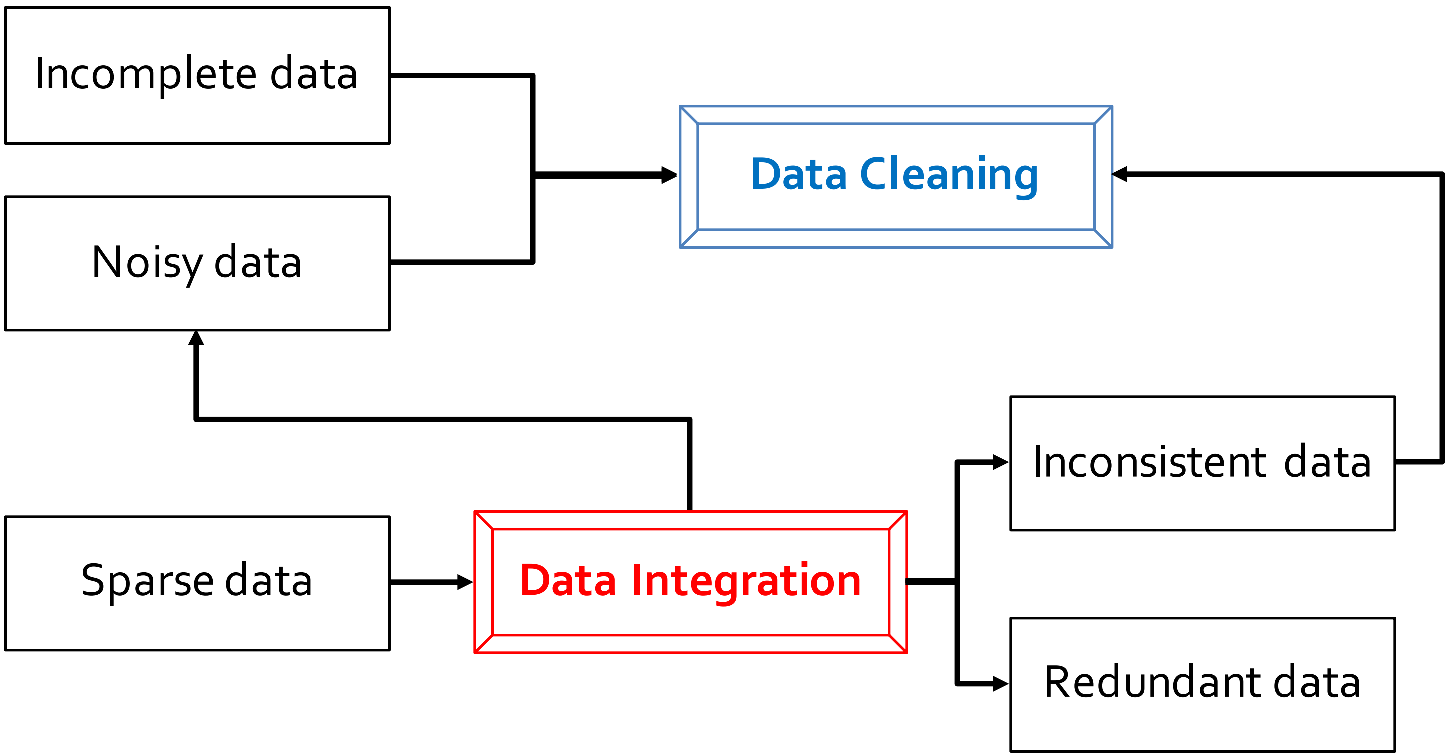
Redundant data

Quality Issues in Collecting Data

- If you don't want to write down your height or weight…
- If you write down you are 12 feet high…
- If in the first lecture you say you are from California and in the third you submit your hometown as Florida…
- If in the first lecture you say you are from California and in the third you submit your hometown as C.A. …
- If you were not in class but your classmate writes down you are from California but actually you are from Florida…

- Sparse data
- Incomplete data
- Noisy data
- Inconsistent data
- Redundant data

Quality Issues and Modules



Data Integration

- Data integration
  - Combining data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- Entity identification:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- Detecting and resolving data value conflicts
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Example: CS Institutions Data

Data Integration

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	▸ Carnegie Mellon University	18.5	150
4	2	▸ Massachusetts Institute of Technology	12.2	82
5	3	▸ Stanford University	10.9	54
6	3	▸ University of California - Berkeley	10.9	81
7	5	▸ Univ. of Illinois at Urbana-Champaign	9.9	84
8	6	▸ Cornell University	8.7	68
9	7	▸ University of Michigan	8.6	63
10	8	▸ University of Washington	8.3	56
11	9	▸ University of California - San Diego	6.9	54
12	10	▸ Georgia Institute of Technology	6.8	75
13	11	▸ University of Wisconsin - Madison	5.9	47
14	12	▸ Columbia University	5.8	47

	A	B	C
1	ihttps://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science		
2	School	Name	Early Career Pay
3	1	Stanford University	$101,000
4	2	University of Pennsylvania	$90,500
5	3	Dartmouth College	$94,700
6	4	Princeton University	$93,400
7	5	University of California - Berkeley	$97,000
8	6	Yale University	$98,000
9	7	Columbia University	$86,400
10	8	Cornell University - Ithaca, NY	$86,500
11	9	Carnegie Mellon University (CMU)	$92,200
12	10	Duke University	$80,500
13	11	University of California - San Diego (UCSD)	$84,500

Data Cleaning

- **Data in the Real World Is Dirty:** Potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - **Incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data)
    - Jan. 1 as everyone's *birthday*? (disguised missing data)
  - **Noisy:** containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - **Inconsistent:** containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"

Why We Have Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding

Types of Missing Data

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

Missing Completely At Random

- Missingness does not depend on any values of any variables in the dataset.
- Missingness instead depends on neither the values of the observed variables, nor on those of unobserved variables.

Example: The accidental dropping a test tube leading to missing lab test result

Missing At Random

- Missingness does not depend on the values of any of the missing or unobserved variables.
- Instead, missingness might depend on values of the observed variables.
- This means that the pattern of missing values is identifiable.

Example: Suppose males are less likely to respond to their income question in general, but the likelihood of responding is independent of their actual income. In this case, unbiased sex-specific income estimates can be made if we have data on the sex variable (by replacing the missing value with the sex-specific median income, for example)

Missing Not At Random

- Missingness depends on the values of the missing or unobserved variables.
- This means that the pattern is non-random, non-ignorable, and typically arises due to the variable on which the data is missing.

Example: A certain question on a questionnaire tend to be skipped deliberately by participants with certain characteristics

Example: Missing Value Types

Customer	Age	Balance
C1	25	20,000
C2	25	100,000
C3	25	15,000
C4	60	50,000
C5	60	120,000
C6	60	150,000

Missing Completely at Random (MCAR)
Missing at Random (MAR)
Missing Not at Random (MNAR)

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C1	25	20,000
C2	25	*Missing*
C3	25	15,000
C4	60	50,000
C5	60	*Missing*
C6	60	*Missing*

Customer	Age	Balance
C1	25	20,000
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	*Missing*

How to Handle Missing Data?

Ignore the tuple

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C2	25	100,000
C4	60	50,000
C5	60	120,000
C6	60	150,000

How to Handle Missing Data?

Manually fill the data

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	*Missing*
C5	60	50,000
C6	60	120,000
C7	25	150,000
C8	25	*Missing*
C9	25	100,000
…	…	…
Cn	60	120,000

Customer	Age	Balance
Cn+1	25	*Missing*
Cn+2	25	100,000
Cn+3	25	*Missing*
…	25	*Missing*
…	60	*Missing*
…	60	50,000
…	60	120,000
…	25	150,000
…	25	*Missing*
…	…	…
…	60	*Missing*

How to Handle Missing Data?

Automatically fill the data

How to Handle Missing Data?

- Fill in it *automatically* with
  - **A global constant: e.g., “unknown”, “-1”, a new class?!**

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C1	25	*-1*
C2	25	100,000
C3	25	*-1*
C4	60	50,000
C5	60	120,000
C6	60	150,000

How to Handle Missing Data?

- Fill in it *automatically* with
  - A global constant: e.g., “unknown”, “-1”, a new class?!
  - **The attribute mean**

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C1	25	*105,000*
C2	25	100,000
C3	25	*105,000*
C4	60	50,000
C5	60	120,000
C6	60	150,000

How to Handle Missing Data?

- Fill in it *automatically* with
  - A global constant: e.g., “unknown”, “-1”, a new class?!
  - The attribute mean
  - **The attribute mean for all samples belonging to the same class: smarter**

Customer	Age	Balance
C1	25	*Missing*
C2	25	100,000
C3	25	*Missing*
C4	60	50,000
C5	60	120,000
C6	60	150,000

Customer	Age	Balance
C1	25	*100,000*
C2	25	100,000
C3	25	*100,000*
C4	60	50,000
C5	60	120,000
C6	60	150,000

How to Handle Missing Data?

- Fill in it *automatically* with
  - A global constant: e.g., “unknown”, “-1”, a new class?!
  - The attribute mean
  - The attribute mean for all samples belonging to the same class: smarter
  - **The most probable value: inference-based such as Bayesian formula or decision tree**
  - …

	What?	Why?	How to Handle?
Incomplete data	√	√	√
Noisy data			
Inconsistent data			
Redundant data			

Noisy Data

- Noise:
  - random error or variance in a measured variable
- Incorrect attribute values may be due to
  - Faulty data collection instruments
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention

How to Handle Noisy Data?

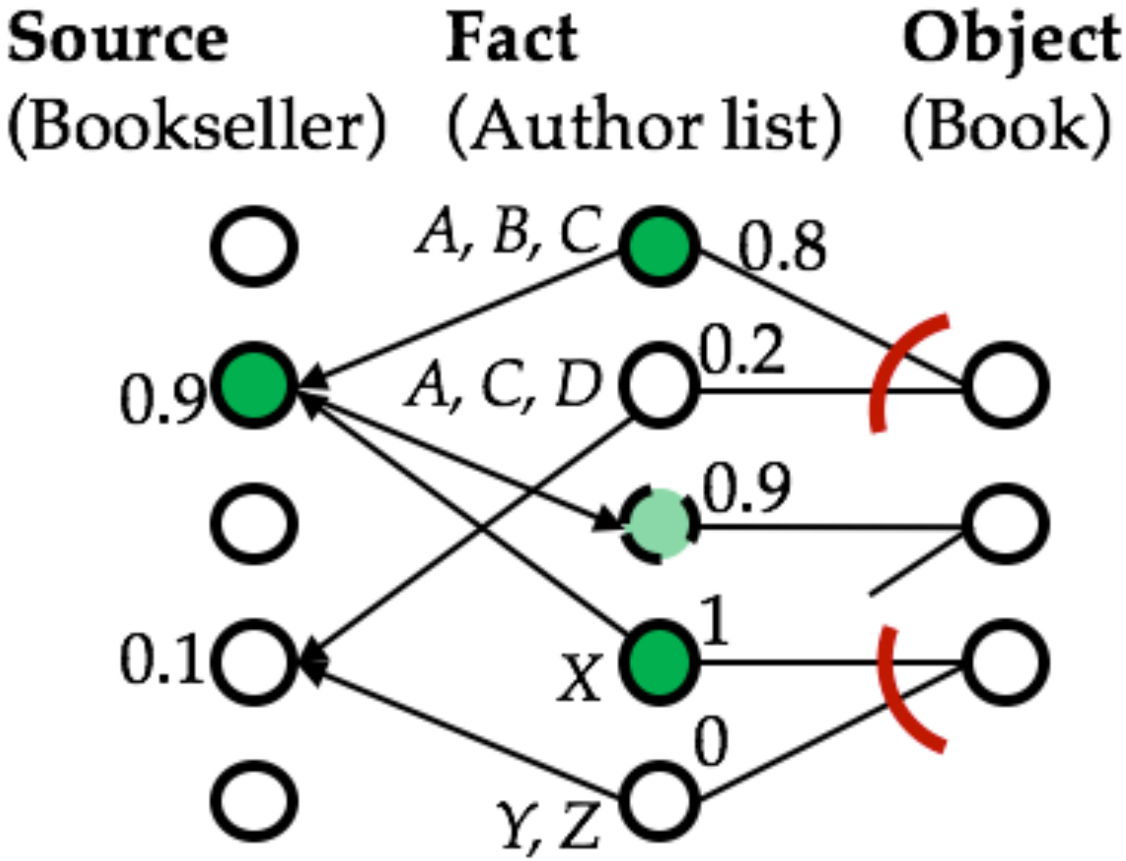
- Binning
  - First sort data and partition into (equal-frequency) bins
  - Then one can **smooth by bin means**, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - **Smooth by** fitting the data into **regression** functions
- Clustering and outlier detection
  - Detect and remove **outliers**
  - Outliers can also be detected using outlier-ness measures (e.g., Z-score)
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)

	What?	Why?	How to Handle?
Incomplete data	√	√	√
Noisy data	√	√	√
Inconsistent data			
Redundant data			

Inconsistent Data

- Data can contain inconsistent values, e.g.,
  - An address field with both ZIP code and city, but where the specified ZIP code area is not in the specified city.
  - A person's age and date of birth are inconsistent.
- Some inconsistencies are easy to detect; some may require consulting an external source.

Truth Finding Research



http://www.kdd.org/exploration_files/Article1_17_2.pdf

	What?	Why?	How to Handle?
Incomplete data	√	√	√
Noisy data	√	√	√
Inconsistent data	√	√	√
Redundant data			

What is Redundancy

- Redundant data occur often when integration of multiple databases
- Why do we have data redundancy?
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue, age

How to Handle Redundancy?

- Redundant attributes may be able to be detected by *correlation analysis (often for categorical attributes)* and *covariance analysis (often for numerical attributes)*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction			450
Not like science fiction			1050
Sum(col.)	300	1200	1500

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction	**90**		**450**
Not like science fiction			1050
Sum(col.)	**300**	1200	**1500**

How to derive "90"?

450/1500 * 300 = 90

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction	**90**	**360**	**450**
Not like science fiction	**210**	**840**	1050
Sum(col.)	**300**	1200	**1500**

How to derive "90"?

450/1500 * 300 = 90

Correlation Analysis

	Play chess	Not play chess	Sum (row)
Like science fiction	**250 (90)**	**200 (360)**	450
Not like science fiction	**50 (210)**	**1000 (840)**	1050
Sum(col.)	300	1200	1500

Correlation Analysis

- **χ^2 (chi-square) test:**

$$\chi^2 = \sum_{i}^{n} \frac{(\underset{\text{observed}}{O_i} - E_i)^2}{\underset{\text{expected}}{E_i}}$$

- **Null hypothesis:** The two distributions are independent
- The cells that contribute the most to the χ^2 value are those whose actual count is different from the expected count
  – The larger the X^2 value, the more the null hypothesis of independence is rejected, and the more likely the variables are related

	Play chess	Not play chess	Sum (row)
Like science fiction	**250 (90)**	**200 (360)**	450
Not like science fiction	**50 (210)**	**1000 (840)**	1050
Sum(col.)	300	1200	1500

Example: Chi-Square Calculation

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

We can reject the null hypothesis of independence at a confidence level of 0.001.

- It shows that like_science_fiction and play_chess are correlated.

Example: Chi-Square Calculation

Degrees of freedom (df)	χ^2 value[19]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Correlation Analysis

- Note: Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population
- Causal analysis

Effective Promotional Strategies Selection in Social Media: A Data-Driven Approach by K. Kuang, M. Jiang, P. Cui, J. Sun, S. Yang. IEEE Transactions on Big Data (TBD), 2017.

Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing by K. Kuang, P. Cui, B. Li, M. Jiang, S. Yang. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2017.

Variance for Single Variable (Numerical Data)

- The variance of a random variable *X* provides a measure of how much the value of *X* deviates from the mean or expected value of *X:*

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

  - where σ^2 is the variance of X, σ is called *standard deviation*

    μ is the mean, and **μ = E[X]** is the expected value of X
  - That is, variance is the expected value of the square deviation from the mean
  - It can also be written as:

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$$

Covariance for Two Variables

- Covariance between two variables X_1 and X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1\mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

  where $\mu_1 = E[X_1]$ is the respective mean or **expected value** of X_1; similarly for μ_2

- **Positive covariance:** If $\sigma_{12} > 0$
- **Negative covariance**: If $\sigma_{12} < 0$
- **Independence**: If X_1 and X_2 are independent, $\sigma_{12} = 0$ but the reverse is not true
  - Some pairs of random variables may have a covariance 0 but are not independent
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions) does a covariance of 0 imply independence

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?

Example: Calculation of Covariance

- Suppose two stocks X_1 and X_2 have the following values in one week:
  - (2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Question: If the stocks are affected by the same industry trends, will their prices rise or fall together?
- Covariance formula

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

- Its computation can be simplified as: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
  - $E(X_1) = (2 + 3 + 5 + 4 + 6) / 5 = 20/5 = 4$
  - $E(X_2) = (5 + 8 + 10 + 11 + 14) / 5 = 48/5 = 9.6$
  - $\sigma_{12} = (2\times5 + 3\times8 + 5\times10 + 4\times11 + 6\times14) / 5 - 4 \times 9.6 = 4$
- Thus, X_1 and X_2 rise together since $\sigma_{12} > 0$

Correlation between Two Numerical Variables

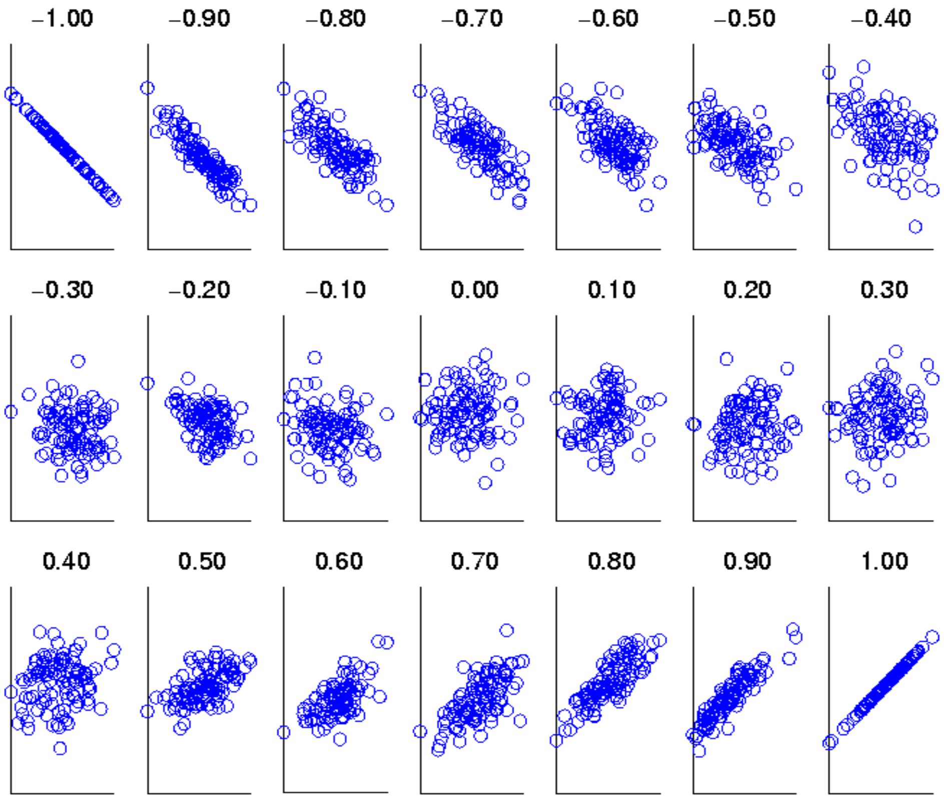
- **Correlation** between two variables X1 and X2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1{}^2 \sigma_2{}^2}}$$

- If $\rho_{12} > 0$: A and B are positively correlated (X1's values increase as X2's)
  - The higher, the stronger correlation
- If $\rho_{12} = 0$: independent (under the same assumption as discussed in co-variance)
- If $\rho_{12} < 0$: negatively correlated

Visualizing Changes of Correlation Coefficient

- Correlation coefficient value range: [–1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from –1 to 1



Covariance Matrix

- The variance and covariance information for the two variables X_1 and X_2 can be summarized as 2 * 2 covariance matrix as

$$\boldsymbol{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = E[\begin{pmatrix} X_1-\mu_1 \\ X_2-\mu_2 \end{pmatrix}(X_1-\mu_1 \quad X_2-\mu_2)]$$

$$= \begin{pmatrix} E[(X_1-\mu_1)(X_1-\mu_1)] & E[(X_1-\mu_1)(X_2-\mu_2)] \\ E[(X_2-\mu_2)(X_1-\mu_1)] & E[(X_2-\mu_2)(X_2-\mu_2)] \end{pmatrix}$$

$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

- Generalizing it to *d* dimensions, we have,

$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \boldsymbol{\Sigma} = E[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

iPython

	A	B	C	D
1	http://csrankings.org/			
2	Rank	Institution	Count	Faculty
3	1	▸ Carnegie Mellon University ◕	18.5	150
4	2	▸ Massachusetts Institute of Technology ◕	12.2	82
5	3	▸ Stanford University ◕	10.9	54
6	3	▸ University of California - Berkeley ◕	10.9	81
7	5	▸ Univ. of Illinois at Urbana-Champaign ◕	9.9	84
8	6	▸ Cornell University ◕	8.7	68
9	7	▸ University of Michigan ◕	8.6	63
10	8	▸ University of Washington ◕	8.3	56
11	9	▸ University of California - San Diego ◕	6.9	54
12	10	▸ Georgia Institute of Technology ◕	6.8	75
13	11	▸ University of Wisconsin - Madison ◕	5.9	47
14	12	▸ Columbia University ◕	5.8	47
15	13	▸ University of Pennsylvania ◕	5.6	46
16	14	▸ University of Southern California ◕	5.5	49
17	15	▸ Princeton University ◕	5.3	51
18	16	▸ University of Texas at Austin ◕	5.2	42
19	16	▸ University of Maryland - [illegible]	[illegible]	[illegible]
20	18	▸ University of California - [illegible]	[illegible]	[illegible]
21	19	▸ Northeastern University	[illegible]	[illegible]
22	19	▸ Purdue University ◕	4.8	51
23	21	▸ University of Massachusetts Amherst ◕	4.7	50
24	22	▸ New York University ◕	4.5	47
25	23	▸ Harvard University ◕	4.2	29
26	23	▸ University of California - Irvine ◕	4.2	54
27	25	▸ Rutgers University ◕	3.9	43
28	26	▸ University of California - Santa Barbara ◕	3.5	25
29	27	▸ University of Utah ◕	3.4	39
30	27	▸ Pennsylvania State University ◕	3.4	31
31	29	▸ Stony Brook University ◕	3.3	41
32	30	▸ University of California - Davis ◕	3.2	29

126 institutions

	A	B	C
1	ihttps://www.payscale.com/college-salary-report/best-schools-by-majors/computer-science		
2	School	Name	Early Career Pay
3	1	Stanford University	$101,000
4	2	University of Pennsylvania	$90,500
5	3	Dartmouth College	$94,700
6	4	Princeton University	$93,400
7	5	University of California - Berkeley	$97,000
8	6	Yale University	$98,000
9	7	Columbia University	$86,400
10	8	Cornell University - Ithaca, NY	$86,500
11	9	Carnegie Mellon University (CMU)	$92,200
12	10	Duke University	$80,500
13	11	University of California - San Diego (UCSD)	$84,500
14	12	Harvard University	$85,300
15	13	University of Washington (UW) - Main Campus	$79,600
16	14	Massachusetts Institute of Technology (MIT)	$94,100
17	15 (tie)	Brown University	$84,800
18	15 (tie)	Lehigh University	[illegible]
19	17	University of California - Santa Barbara (UCSB)	[illegible]
20	18	University of California - Santa Cruz (UCSC)	$77,400
21	19	Rice University	$81,100
22	20	New York University (NYU)	$78,200
23	21	University of California - Irvine (UCI)	$74,100
24	22	Stevens Institute of Technology	$78,800
25	23 (tie)	California Polytechnic State University (CalPoly) - San Luis Obispo	$76,600
26	23 (tie)	San Jose State University (SJSU)	$77,000
27	25	University of Virginia (UVA) - Main Campus	$78,000
28	26	University of California - Los Angeles (UCLA)	$80,600
29	27	Tufts University	$79,700
30	28	Boston College	$75,600

466 institutions

	What?	Why?	How to Handle?
Incomplete data	√	√	√
Noisy data	√	√	√
Inconsistent data	√	√	√
Redundant data	√	√	√

Extra: One Application

- Relationship prediction in heterogeneous networks
  - Can you use Chi-Square or p-value (doing correlation analysis) to select meta paths (as features) for relationship prediction?

Sun, Y., Barber, R., Gupta, M., Aggarwal, C.C. and Han, J., 2011, July. Co-author relationship prediction in heterogeneous bibliographic networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on* (pp. 121-128). IEEE.

Summary: Data Cleaning and Integration

- Understand data quality issues and how to handle them
  - Describe three types of missing data
  - Describe how to handle missing data, noisy data, inconsistent data, and redundant data
- Correlation analysis for handling data redundancy
  - Categorical Variables: Chi-square test
  - Numerical Variables: Covariance analysis

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain," Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009