

# CDS6214

## Data Science Fundamentals

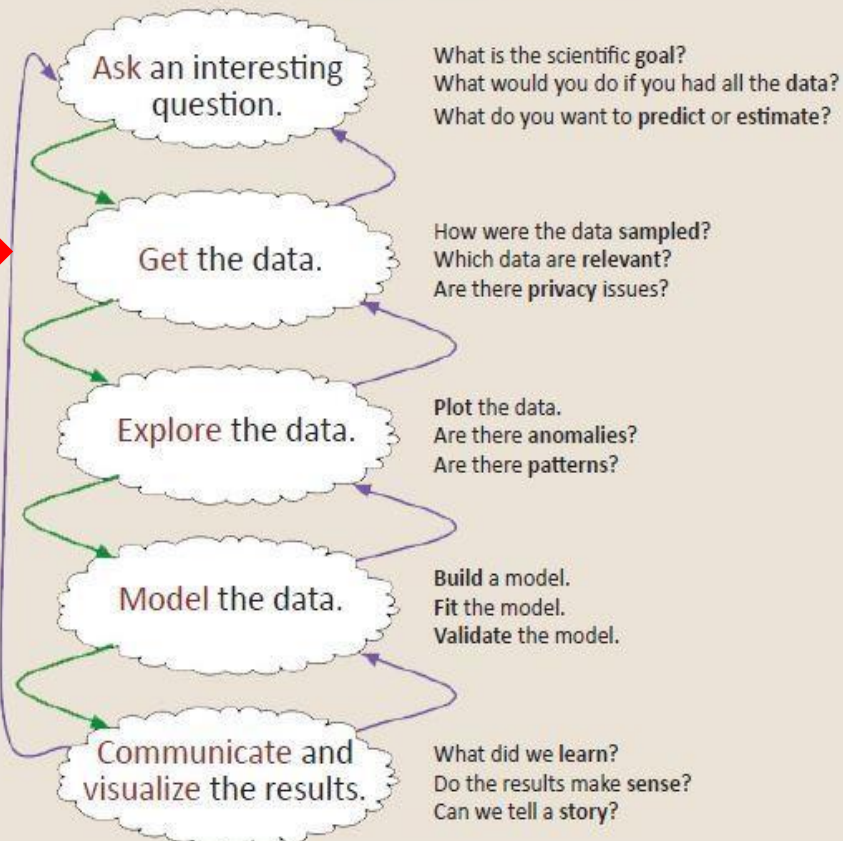
Lecture 3  
Data Wrangling

**data  
science  
pipelines**

data  
wrangling

# Data Science Process

## The Data Science Process



Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://www.cs109.org/>.

## What should we do next?

- Identify the question
- Collect and **pre-process the data**
- Explore and analyze the data
- Model the data
- Infer and visualize results

# The Issue with Data Quality

---

Here are the data quality statistics shared in an article written on 7th March 2024

- ❖ Every year, 25 30% of data becomes inaccurate leading to less effective sales and marketing campaigns Businesses lose as much as 20% of revenue due to poor data quality.
- ❖ 54% of businesses cite data quality and completeness as their largest marketing data management challenge.
- ❖ Another study by IBM estimated that the annual cost of bad data in the United States is around \$3.1T. T for Trillion!.

**data  
quality**

# Dimensions of Data Quality

---



The value of data analysis is highly dependent on the quality of the data

# (1) Completeness

---

- All necessary data have been recorded. Data can be considered complete even if optional data is missing.

**E.g.** Parents of new students at school are requested to complete a Data Collection Sheet which includes medical conditions and emergency contact details as well as confirming the name, address and date of birth of the student. Scenario: At the end of the first week of the Autumn term, data analysis was performed on the 'First Emergency Contact Telephone Number' data item in the Contact table. There are 300 students in the school and 294 out of a potential 300 records were populated, therefore  $294/300 \times 100 = 98\%$  completeness has been achieved for this data item in the Contact table.



## (2) Timeliness

---

- The data is kept up to date. It is all about having the right information at the right time.

**E.g.** Tina Jones provides details of an updated emergency contact number on 1st June 2013 which is then entered into the Student database by the admin team on 4th June 2013. This indicates a delay of 3 days. This delay breaches the timeliness constraint if the service level agreement for changes is 2 days.

# (3) Consistency

---

- The data across processes, organizations, sources are in sync with each other

**E.g.** School admin: a student's date of birth has the same value and format in the school register as that stored within the Student database.

# (4) Validity

---

- The same fields are used consistently for the same information captured.

**E.g.** Each class in a UK secondary school is allocated a class identifier; this consists of the 3 initials of the teacher plus a two digit year group number of the class: AAA99 (3 Alpha characters and two numeric characters).

**Scenario 1:** A new year 9 teacher, Sally Hearn (without a middle name) is appointed therefore there are only two initials. A decision must be made as to how to represent two initials or the rule will fail and the database will reject the class identifier of “SH09”. It is decided that an additional character “Z” will be added to pad the letters to 3: “SZH09”, however this could break the accuracy rule. A better solution would be to amend the database to accept 2 or 3 initials and 1 or 2 numbers.

## (5) Accuracy

---

- The data was recorded correctly. Sometimes this fails when there is human error or system data entry error.

**E.g.** A European school is receiving applications for its annual September intake and requires students to be aged 5 before the 31st August of the intake year. In this scenario, the parent, a US Citizen, applying to a European school completes the Date of Birth (D.O.B) on the application form in the US date format, MM/DD/YYYY rather than the European DD/MM/YYYY format, causing the representation of days and months to be reversed. As a result, 09/08/YYYY really meant 08/09/YYYY causing the student to be accepted as the age of 5 on the 31st August in YYYY. The representation of the student's D.O.B. –whilst valid in its US context –means that in Europe the age was not derived correctly and the value recorded was consequently not accurate.

# (6) Uniqueness

---

- Each record is distinct and unique.

**E.g.** A school has 120 current students and 380 former students (i.e. 500 in total) however; the Student database shows 520 different student records. This could include Fred Smith and Freddy Smith as separate records, despite there only being one student at the school named Fred Smith. This indicates a uniqueness of  $500/520 \times 100 = 96.2\%$

# Data Quality Assessment

---

1. **Identify which data items need to be assessed** for data quality i.e. data items deemed as critical to business operations and associated management reporting
2. **Assess which data quality dimensions to use** and their associated weighting.
3. For each data quality dimension, **define values or ranges representing good and bad quality data**. Apply the assessment criteria to the data items.
4. **Review the results** and determine if data quality is acceptable or not
5. Where appropriate, **take corrective actions** e.g. clean the data and improve data handling processes to prevent future recurrences
6. **Repeat the above on a periodic basis** to monitor trends in Data Quality

# Why are there errors & inconsistencies?

---

- ❖ The **diversity of data sources** brings abundant data types and complex data structures.
- ❖ Data volume is tremendous, and it is difficult to judge data quality within a reasonable amount of time.
- ❖ Data changes very fast and the “timeliness” of data may be very short.
- ❖ No unified and approved data quality standards.

# The Problem with Noise

---



- ❖ What are the impacts of having noisy data?
- ❖ Wrong insights → wrong business decisions!



# Data Cleaning / Wrangling

---

- ❖ **Data Cleaning / Data Wrangling** is...  
A process of **transforming** data from “raw” form into an appropriate format that can be conveniently **consumed** or **analysed** to **generate actionable insights**



# Data Cleaning / Wrangling

---

- ❖ Data Cleaning
- ❖ Data Preprocessing
- ❖ Data Preparation
- ❖ Data Scrubbing
- ❖ Data Munging
- ❖ Data Transformation
- ❖ ...



# Data Cleaning

---

- ❖ Often, **Data Cleaning** is a more specific process,
  - ❖ To determine inaccurate, incomplete or unreasonable data and then the detected errors are corrected by replacing, modifying or deleting the dirty data.
- ❖ A time-consuming and tedious task which cannot be ignored
- ❖ **Not a one-off process**. It is iterative as new data arrives again...

# Data Cleaning

---

- ❖ Number ONE problem in data warehousing
- ❖ Routine tasks:
  - ❖ Filling in missing data,
  - ❖ Smoothing noisy data,
  - ❖ Identifying and removing outliers,
  - ❖ Resolving inconsistencies,
  - ❖ Etc.

# Common sources of discrepancies in data

---

**Incomplete data** comes from:

- Non available data value when collected
- Different criteria between the time when the data was collected and when it is analysed
- Human/hardware/software problems

**Noisy data** comes from:

- Data collection: faulty instruments
- Data entry: human or computer errors
- Data transmission

**Inconsistent (and redundant) data** comes from:

- Different data sources, so non-uniform naming conventions
- Functional dependency
- Referential integrity violation

# Problem I: Incomplete Data

---

- ❖ Missing data, or missing values, occur when no data value is stored for the variable in an observation.
- ❖ Handling missing values
  - Ignore the tuple: usually done when class label is missing.
  - Fill in missing value manually
  - Use a global constant to fill missing value
  - Use the attribute mean/median for all samples belonging to the same class as the tuple.
  - Predict the missing value by using a learning algorithm
- ❖ Truncated fields – long texts

# Problem II: Noisy Data

---

- ❖ Noise is a random error or variance in a measured variable
- ❖ How to handle noisy data?
  - ❖ Binning methods
  - ❖ Outlier analysis
  - ❖ Regression
  - ❖ Combination of human and computer inspection

# Binning

---

- ❖ Sorted data for price (in dollars)

E.g. 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- ❖ Techniques:

- ❖ Partition into equal-size bins:

Bin 1	Bin 2	Bin 3
4, 8, 9, 15	21, 21, 24, 25	26, 28, 29, 34

- ❖ Smoothing by bin means:

Bin 1	Bin 2	Bin 3
9, 9, 9, 9	23, 23, 23, 23	29, 29, 29, 29

- ❖ Smoothing by bin boundaries:

Bin 1	Bin 2	Bin 3
4, 4, 4, 15	21, 21, 25, 25	26, 26, 26, 34



# Binning

---

- ❖ These bins can now be re-labelled based on the bin numbers (similar to “categories”)
- ❖ This way, we eliminate potentially noisy data, especially data that fluctuates a little but are naturally representative of a certain state

Bin 1	Bin 2	Bin 3
4, 8, 9, 15	21, 21, 24, 25	26, 28, 29, 34

Values: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

After binning: 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3

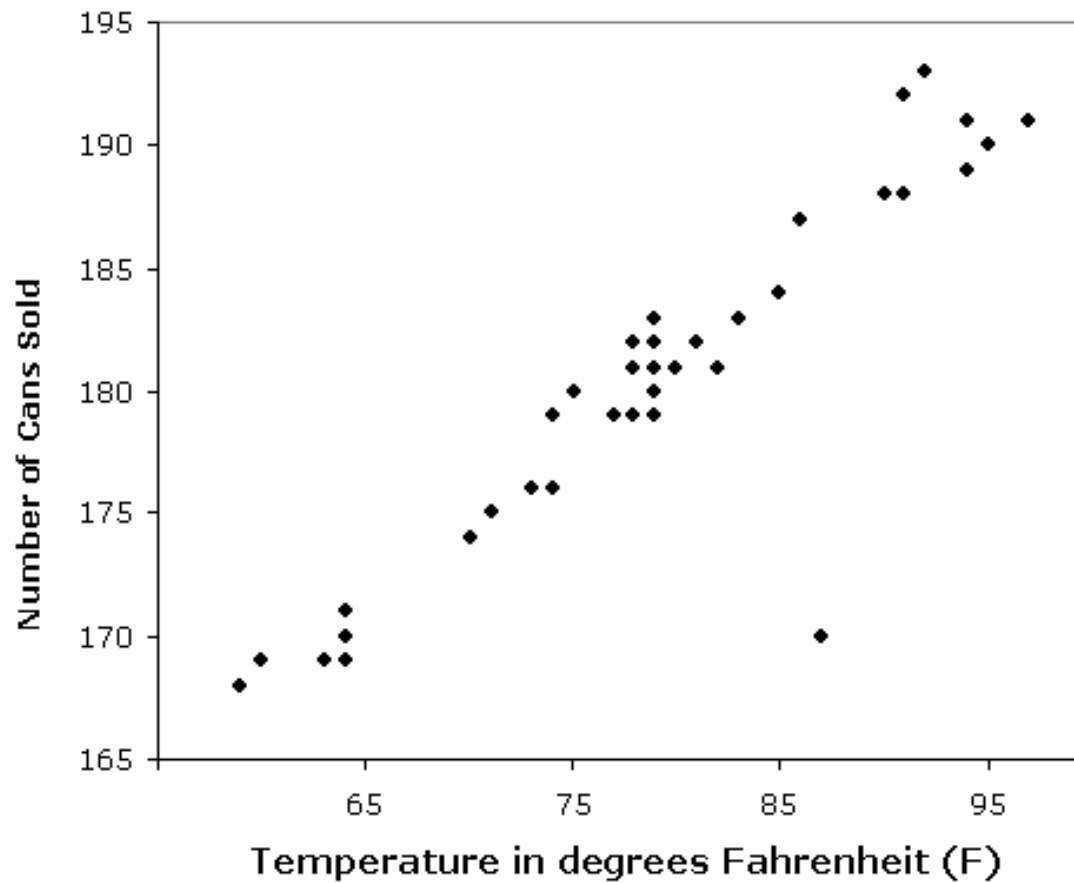
# Outliers

---

- ❖ An **outlier** is an observed data that is distant from other observations
- ❖ Outliers **are not** errors. They should be detected, but not necessarily removed. Their inclusion in the analysis is a statistical decision.
- ❖ It is possible to identify outliers by doing statistical graphs (e.g. scatter plots).

# Discuss #1: Outliers

---



# Problem III: Inconsistent Data

---

- ❖ **Naming Conventions**

- ❖ Synonyms
- ❖ Nicknames/Initials
- ❖ Abbreviations/Acronyms
  - ❖ E.g. Car, vehicle
  - ❖ New York, NY, NYC

- ❖ **Parsing Text into Fields**

- ❖ Different delimiters, line terminators
- ❖ NULL vs empty data
- ❖ Character encoding problems

- ❖ **Different Representations**

- ❖ How it is written: 2 vs TWO vs II
- ❖ Data types: integer vs float
- ❖ Units: ft vs cm; kg vs pound

# Problem III: Inconsistent Data


---

- ❖ **Primary Key Violation**
  - ❖ Two entries with the same primary key
- ❖ **Formatting Issues**
  - ❖ Dates: mon-year, mm/yy, dd/mm/yy
  - ❖ Numbers: 1000, 1000.00
  - ❖ Currency: \$, £, ¥
  - ❖ ID: 911101-01-1234 or 107123456

# Duplicate Records

---

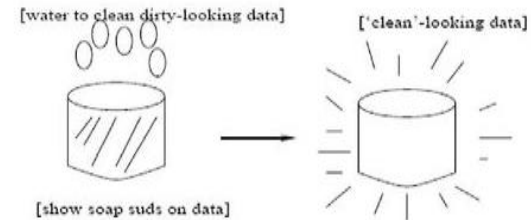
- ❖ Duplicate: Identical attribute. Updates to both records required.
- ❖ Redundancy: Exact copy of the record
- ❖ Such possibilities have to be dealt with especially **when collection comes from various sources:**

ID	Title	First	Last	AddressLine	City	Postcode	Telephone
1	Miss	Catrina	Trewin	123 Sample Road	Town	ABC 123	01980 592 999
2	Miss	Catrina	Trewin	123 Sample Road	Town	ABC 123	
3		Catrina	Trewin	123 Sample Road	Town	ABC 123	
4	Miss	C	Trewin	123 Sample Road		ABC 123	

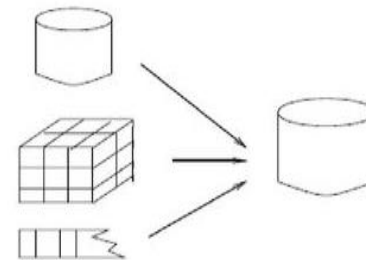
# Other Wrangling Tasks

- ❖ **Data Integration:** A process of integrating data from multiple sources as a single view
- ❖ **Data Reduction:** A process of obtaining a reduced representation of the dataset that is much smaller in volume yet produces the same (or almost the same) analytical results
- ❖ **Data Transformation:** A process involving normalization, discretization, and concept hierarchy generation

Data Cleaning



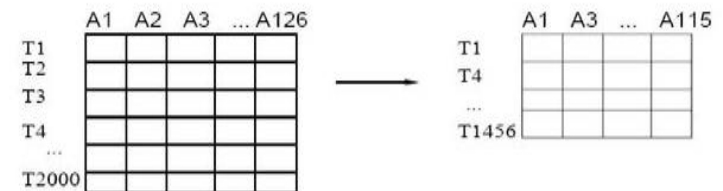
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



# Data Integration

---

- ❖ Merging of data from multiple sources and (probably) have a single view over all these sources
- ❖ Integration can be physical or virtual
- ❖ Physical: Copy the data to warehouse
- ❖ Virtual: Keep the data only at the sources
- ❖ Issues: Naming inconsistencies, aggregated attributes, redundant data



# The Need for Integration

---

❖ The main problem is the **heterogeneity among data sources**

1. **Source type heterogeneity**

❖ Systems that store the data is different

2. **Communication heterogeneity**

❖ Some systems have web interfaces, some allow direct query languages, some offer APIs...

3. **Schema heterogeneity**

❖ The structure of tables storing data can be different (even if storing the same data)

# The Need for Integration

---

## 4. Data type heterogeneity

- ❖ Storing the same data (and values) but with different data types
- ❖ E.g. Storing name as fixed length / variable length
- ❖ E.g. Storing the phone number as String or as Number

## 5. Value heterogeneity

- ❖ Same logical values stored in different ways
- ❖ E.g. Prof, Prof., Professor...
- ❖ E.g. “Right, “R”, 1....

## 6. Semantic heterogeneity

- ❖ Same values in different sources can mean different things
- ❖ E.g. Column “title” in one database means “Job Title” while in another database it refers to “Person Title”

# Entity Resolution

---

- ❖ Data coming from different sources may be different even if representing the same objects
- ❖ Entity resolution:
  - Process of figuring out which records represent the same thing
  - Linking relevant records together
- ❖ Mechanisms for entity resolution:
  - ❖ Edit Distance
    - ❖ Compare string fields
  - ❖ Normalization & Ontology
    - ❖ Using a dictionary to replace abbreviations, ontology finds synonyms
  - ❖ Clustering & Partitioning
    - ❖ Cluster records to find natural groupings

# Merging Similar Records

---

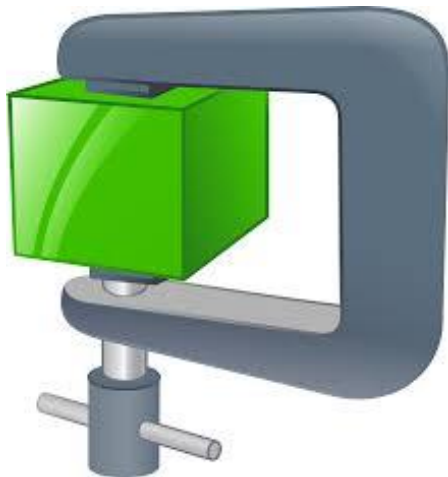
How to merge similar records???

- ❖ In some cases, e.g., misspelled synonyms, it is possible to merge results
- ❖ In other cases, e.g., conflicts, there is no easy way to find the correct values. We have to report these cases, address at the source.

# Data Reduction

---

Why do we need to reduce data?



❖ Strategies:

- **Dimensionality reduction**

- Process of reducing number of random variables or attributes under consideration

- **Numerosity reduction**

- Process of replacing original data volume by alternative, smaller forms of data representation

- **Data compression**

- Process of reducing the size of data while preserving the representation of original data (the best possible)

# How to reduce number of attributes?

---

- ❖ **Attributes Subset Selection** – Remove irrelevant or redundant attributes (or dimensions)
  - ❖ **Objective:** To find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes

# Methods for Attribute Subset Selection

- ❖ **Stepwise forward selection**
  - ❖ Starts with an empty set of attributes. Best attribute picked and iteratively the next best ones are added
- ❖ **Stepwise backward elimination**
  - ❖ Start with the full set of attributes. Iteratively eliminate the worst attribute left in set
- ❖ **Combination of both**
  - ❖ At each iteration, select the best attribute and remove the worst from the balance set

Forward Selection	Backward Elimination
Initial Attribute Set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$	Initial Attribute Set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$
Initial reduced set: $\{\}$	$\rightarrow \{A_1, A_2, A_3, A_4, A_5, A_6\}$
$\rightarrow \{A_1\}$	$\rightarrow \{A_1, A_3, A_4, A_5, A_6\}$
$\rightarrow \{A_1, A_4\}$	$\rightarrow \{A_1, A_4, A_5, A_6\}$
$\rightarrow$ Reduced Attribute Set: $\{A_1, A_4, A_6\}$	$\rightarrow$ Reduced Attribute Set: $\{A_1, A_4, A_6\}$

# Methods for Attribute Subset Selection

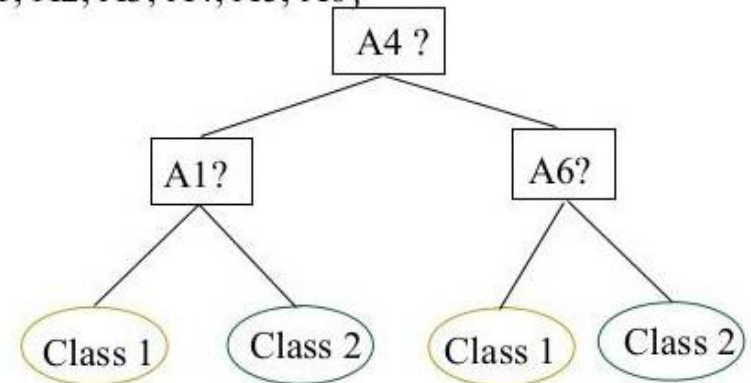
---

## ❖ Decision Tree Induction

- ❖ Decision tree algorithm is intended for classification
- ❖ Constructs a flowchart-like structure where each internal node (non-leaf) represents a test on an attribute, each branch represents an outcome to a test on an attribute and each external node (leaf) represents the prediction

Initial attribute set:

{A1, A2, A3, A4, A5, A6}



Reduced attribute set: {A1, A4, A6}



# Data Transformation

---

- ❖ Normalization
- ❖ Discretization
- ❖ Concept Hierarchy Generation

# Normalization

---

- ❖ Objective: Gives all attributes an equal weight
- ❖ Useful for classification algorithms (machine learning / neural networks)
  - ❖ More effective learning phase, speed-ups
- ❖ Useful for distance-based instance comparisons (nearest neighbours / clustering)
  - ❖ Prevent attributes with large ranges from outweighing attributes with small ranges

# Normalization Methods

---

## Min-max normalization

Normalize to a new max/min range (typically [0,1])

$v = 73600$  in  $[12000, 98000]$   
 $v' = 0.716$  in  $[0,1]$  range

## Z-score normalization

Or “zero-mean” normalization  
Perform a shift of values according to data distribution

If mean = 54000, std dev = 16000, then  
 $v = 73600$ ,  $v' = 1.225$

## Decimal scaling

Normalize by “moving” the decimal point of values to keep max value less than 1

$v'(i) = v(i)/10^k$  for smallest  $k$  such that  $\max(v') < 1$

For value in range between 1 and 1000, so  $k = 3$   
443 becomes .443

# Discretization

---

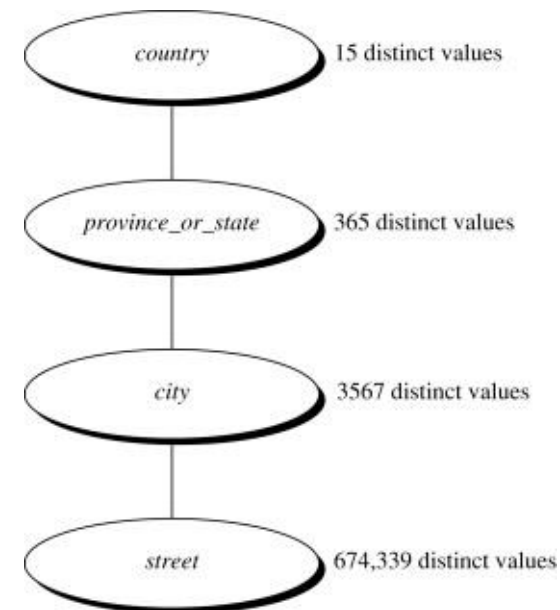
- ❖ Divide the range of continuous attribute into intervals because some data mining algorithms only accept categorical attributes
- ❖ Some techniques:
  - ❖ Binning (discussed earlier)
  - ❖ Entropy-based discretization

Use “entropy” (a statistical measure on the predictability of data) to find best way to split data into bins. Splitting is done on each partition based on maximal information gain, then repeat until splitting ends.

# Concept Hierarchy Generation

---

- ❖ Hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
- ❖ The attribute with the most distinct values is placed at the lowest level of the hierarchy
- ❖ Specification of partial/total ordering of attributes explicitly at schema level  
**street < city < state < country**
- ❖ Specification of hierarchy by explicit data grouping  
**{Urbana, Champaign, Chicago} < Illinois**



# Reading Material

---

- Endel & Piringer, “[Data Wrangling: Making data useful again](#)”, 2015
- Trifacta, “[The Opportunity for Data Wrangling in Financial Services and Insurance](#)”, 2016
- [\[YouTube\] Daniel Chen: Cleaning and Tidying Data in Pandas | PyData DC 2018](#)
- [\[PDF\] Data Wrangling with pandas Cheat Sheet](#)
- Sign up @ Kaggle: <https://www.kaggle.com/>