SELECTED TOPICS IN ENGINEERING

INTR. TO PROG. FOR DATA SCIENCE
ENGR 350

Tuesday-Thursday 10:00-12:45
ENG B05

Dr. Banu Yobaş

Thinking like …

- combine some of the best features of mathematics, engineering, and natural science.

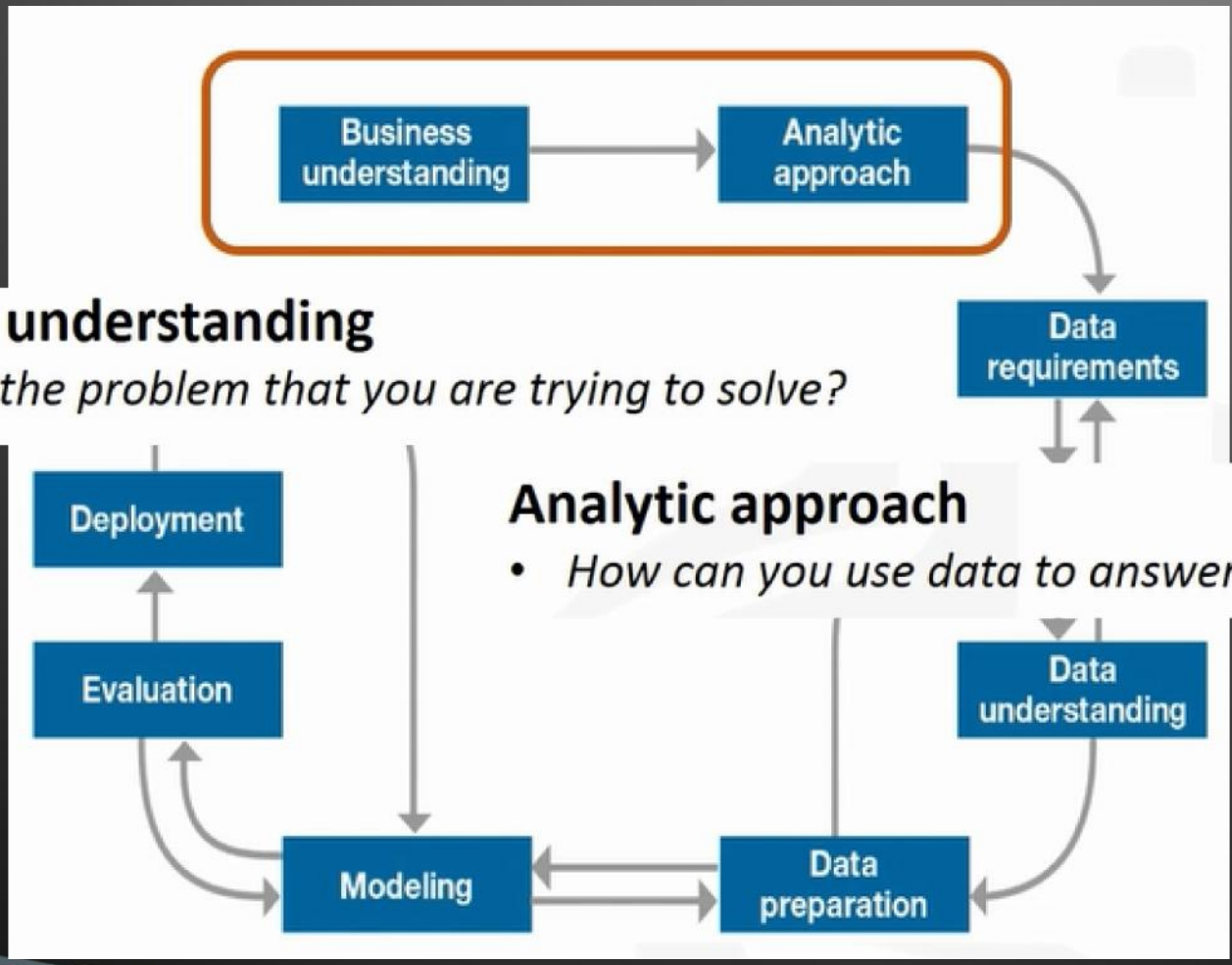
- Like mathematicians, computer scientists use formal languages to denote ideas (specifically computations).
- Like engineers, they design things, assembling components into systems and evaluating ttrade-offs among alternatives.
- Like scientists, they observe the behaviour of complex systems, form hypotheses, and test predictions.

Problem solving

means the ability to

- formulate problems,
- think creatively about solutions, and
- express a solution clearly and accurately.

Data Science Methodology



Analytic Approach shall be based on the question type

Descriptive
- Current status

Diagnostic (Statistical Analysis)
- What happened?
- Why is this happening?

Predictive (Forecasting)
- What if these trends continue?
- What will happen next?

Prescriptive
- How do we solve it?

If the question is to determine probabilities of an action
- Use a Predictive model

If the question is to show relationships
- Use a descriptive model

If the question requires a yes/no answer
- Use a classification model

The correct approach depends on the Bus.Req.

Question Types

Data Collection & Understanding

- Cleaning
- Transforming
- Use of domain knowledge in machine learning alg.

- Invalid values
- Missing data
- Duplicates ? Removal
- Formatting

Preparation

Related steps

Data Collection & Understanding

Is the data that you collected representative of the problem to be solved?

- Descriptive statistics
  - Univariate statistics
  - Pairwise correlations
  - Histogram

- to see how closely certain variables were related, thus making only one relevant for modeling

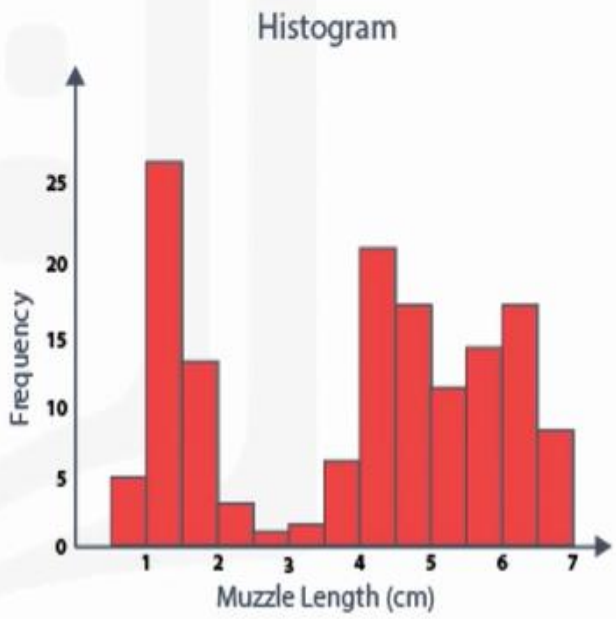
- to understand how values or a variable are distributed

Data understanding

Descriptive statistics

Data Science Methodology

Histograms are a good way to understand how values or a variable are distributed, and what sorts of data preparation may be needed to make the variable more useful in a model.



- Accuracy
- Relevance
- Accessibility
- Completeness
- Clarity
- Timelines

Data Understanding

Data Quality

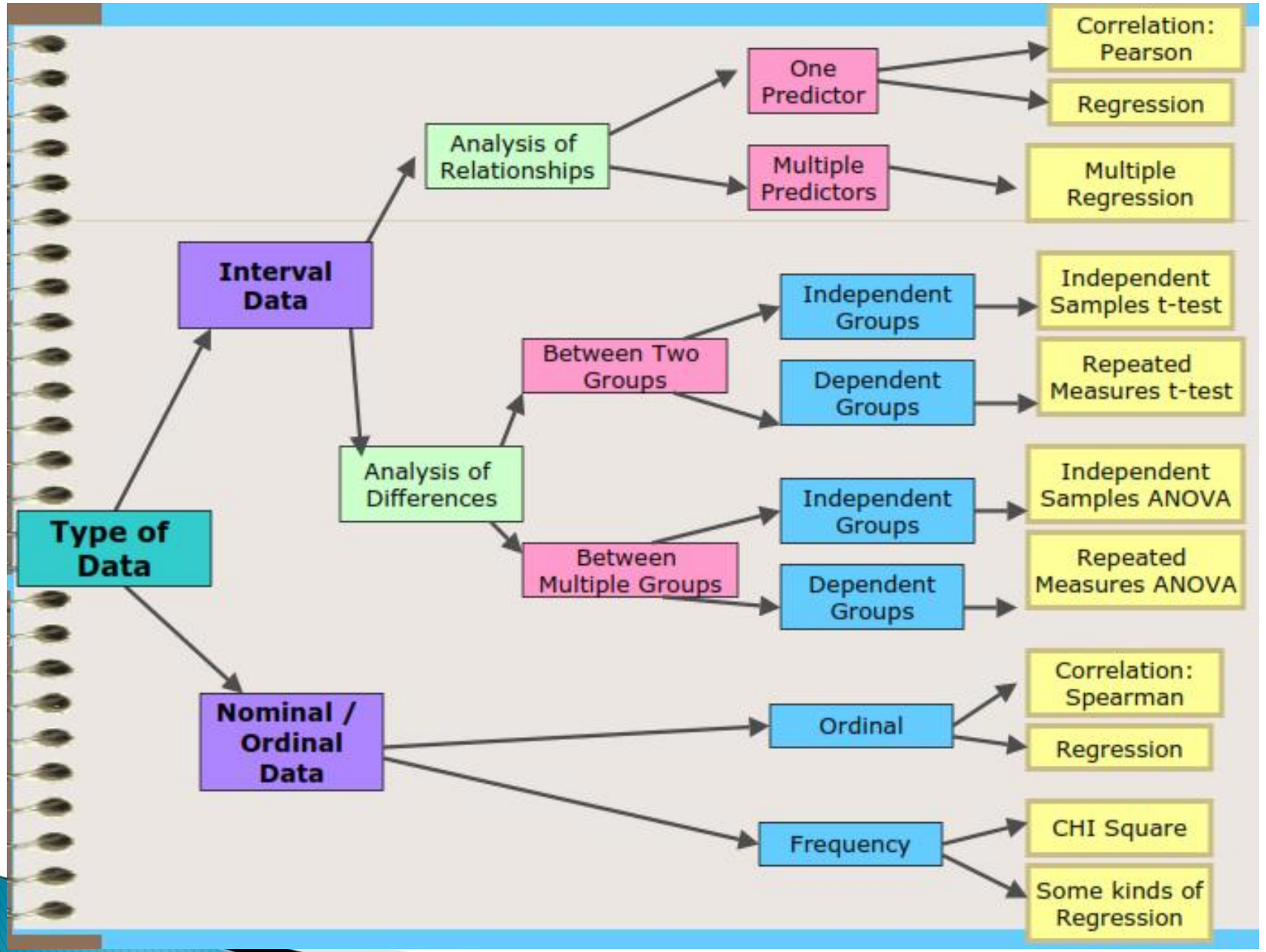
Measurement Scales

Nonmetric (qualitative)

- Nominal eg Male or Female
- Ordinal eg level of satisfaction with a course

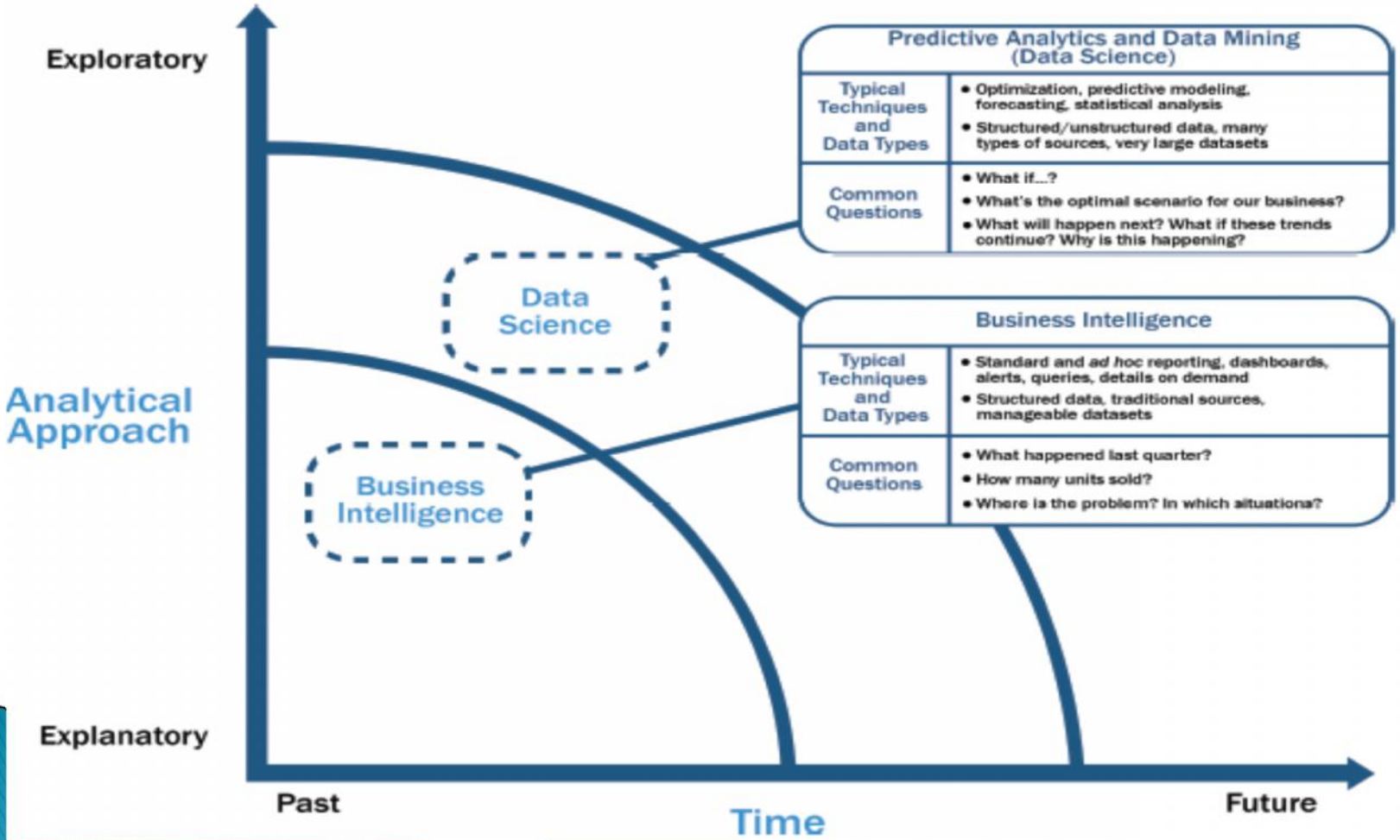
Metric (quantitative)

- Interval eg Temperature measured in ^{o}C, ^{o}F
- Ratio eg Weight, height

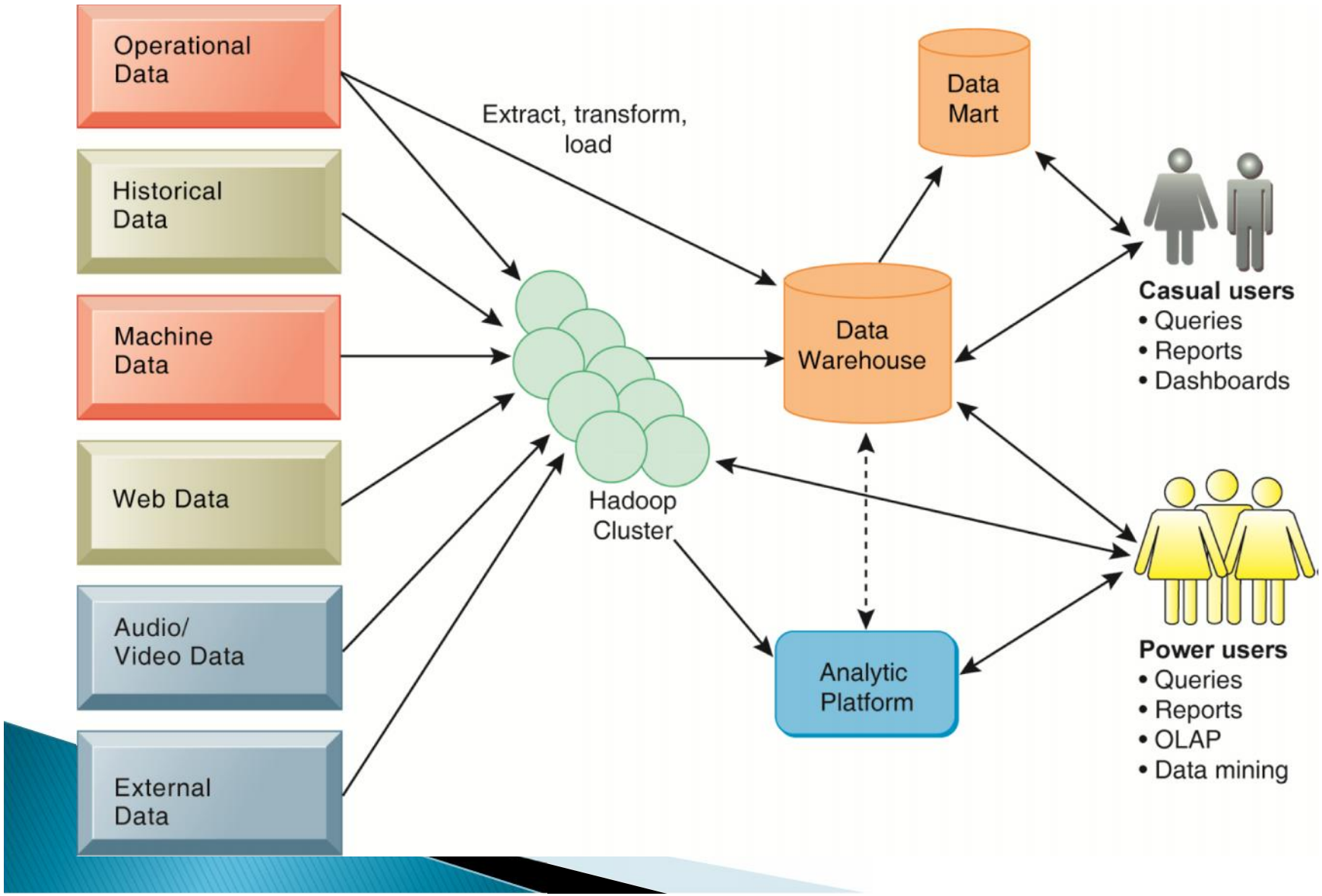

Type of Data
Interval Data
Analysis of Relationships
One Predictor
Correlation: Pearson
Regression
Multiple Predictors
Multiple Regression
Analysis of Differences
Between Two Groups
Independent Groups
Independent Samples t-test
Dependent Groups
Repeated Measures t-test
Between Multiple Groups
Independent Groups
Independent Samples ANOVA
Dependent Groups
Repeated Measures ANOVA
Nominal / Ordinal Data
Ordinal
Correlation: Spearman
Regression
Frequency
CHI Square
Some kinds of Regression

Data Science vs Business Intelligence

Source: EMC Education Services, Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley, 2015



Business Intelligence (BI) Infrastructure



Three Types of Business Analytics

Source: Thomas H. Davenport, "Enterprise Analytics: Optimize Performance, Process, and Decisions Through Big Data", FT Press, 2012

Profile of a Data Scientist

- **Quantitative**
  - mathematics or statistics
- **Technical**
  - Software engineering, machine learning, and programming skills

- **Skeptical mind-set** and **critical thinking**
- **Curious** and **creative**
- **Communicative** and **collaborative**

References

- Lisa Arthur (2013), Big Data Marketing: Engage Your Customers More Effectively and Drive Value, Wiley.
- EMC Education Services (2015), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, Wiley.
- Foster Provost and Tom Fawcett (2013), Data Science for Business: What you need to know about data mining and data-analytic thinking, O'Reilly
- Thomas W. Miller (2013) , Marketing Data Science: Modeling Techniques in Predictive Analytics with R and Python, Pearson FT Press
- Peter C. Verhoef and Edwin Kooge (2016), Creating Value with Big Data Analytics: Making Smarter Marketing Decisions, Routledge
- Stephan Kudyba (2014), Big Data, Mining, and Analytics: Components of Strategic Decision Making, Auerbach Publications
- Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. Big Data Research,2(1), 28-32
- Erevelles, S., Fukawa, N., & Swayne, L. (2016). Big Data consumer analytics and the transformation of marketing. Journal of Business Research, 69(2), 897-904.

Data Mining

A common use case for data mining is to improve sales by asking a customer who is buying a product if s/he would like another similar product as well.

If a person buys product X,
then they are likely to purchase product Y

This can be done through affinity analysis, which is the study of when things exist together

What is affinity analysis?

- Affinity analysis is a type of data mining that gives similarity between samples (objects).
  - Also used in recommender systems
- This could be the similarity between the following:
  - users on a website, in order to provide varied services or targeted advertising
  - items to sell to those users, in order to provide recommended movies or products
  - human genes, in order to find people that share the same ancestors

How to measure affinity?

- We can measure affinity in a number of ways.

For instance, we can record how frequently two products are purchased together.

- We can also record the accuracy of the statement when a person buys object 1 and also when they buy object 2.
- Other ways to measure affinity include computing the similarity between samples

Affinity Analysis

An example of affinity analysis, recommending products based on purchasing habits

We wish to find rules of the type:

If a person buys product X,

then they are likely to purchase product Y.

We can quite easily create a list of all of the rules in our dataset by simply finding all occasions when two products were purchased together.

However, we then need a way to determine good rules from bad ones. This will allow us to choose specific products to recommend.

Rules of this type can be measured in many ways, of which we will focus on two:

support and **confidence.**

Support

- **Support** is the number of times that a rule occurs in a dataset, which is computed by simply counting the number of samples that the rule is valid for.
- It can sometimes be normalized by dividing by the total number of times the premise (the if statement) of the rule is valid,
- the support measures *how often a rule exists*

in a particular dataset

Confidence

confidence measures *how accurate they are* when they can be used.

It can be computed by determining the percentage of times the rule applies when the premise applies.

We first count how many times a rule applies in our dataset and divide it by the number of samples where the premise (the if statement) occurs.

Lift

- Interest factor
- Given two items, A and B, lift indicates whether there is a relationship between A and B, or whether the two items are occuring together in the same orders simply by chance (ie: at random). Unlike the confidence metric whose value may vary depending on direction (eg: confidence{A->B} may be different from confidence{B->A}), lift has no direction.

  lift{A,B} = lift{B,A}
- the ratio of the observed support to that expected if A and B were independent

Support & Confidence & Lift

Support = (Count of product A in N transaction)/ Total Transactions (N)

Confidence (LHS-->RHS) = P(RHS | LHS)

= P(RHS∩LHS)/P(LHS)

= Support(RHS, LHS)/Support(LHS)

Lift (A,B) = Supp(AUB)/Supp(A) X Supp(B)

= P(RHS/LHS)/P(RHS)

Why use support, confidence?

- Is the rule happens by chance?
- Is the rule interesting?
- Is it profitable?

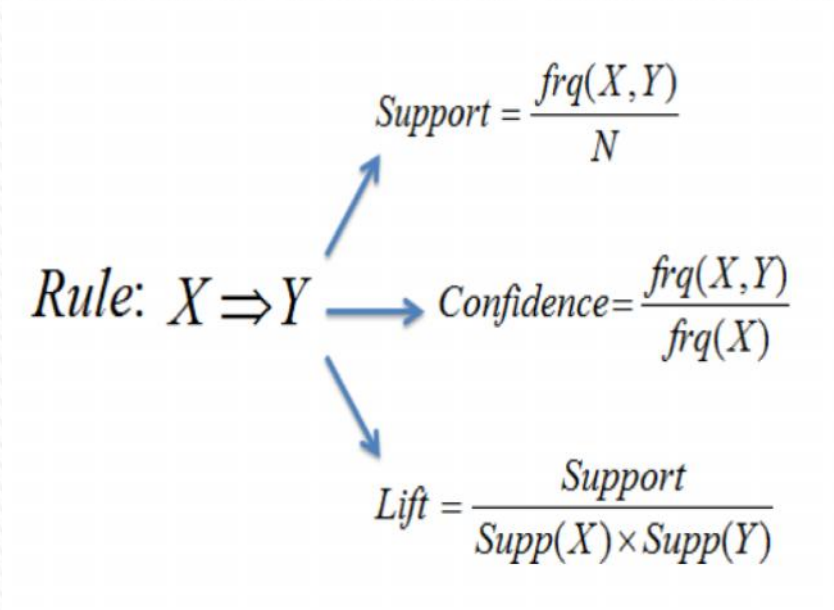
How often a rule is applicable to a given data set

- Is the inference made by the rule reliable?
- Conditional probability of Y given X

How frequently items in Y appear in transactions that contain X

Support	Confidence

Affinity Analysis OR Association Rules



Rule	Support	Confidence	Lift
$A \Rightarrow D$			
$C \Rightarrow A$			
$A \Rightarrow C$			
$B \& C \Rightarrow D$			

Source: UofT

RULE

Example

Lift

lift{A,C} = lift{C,A} = support{A,C} / (support{A} * support{C})

- Think of the denominator as the likelihood that A and B will appear in the same order if there was *no* relationship between them.
- In the example above, if C occurred in 80% of the orders and A occurred in 60% of the orders, then if there was no relationship between them, we would *expect* both of them to show up together in the same order 48% of the time (ie: 80% * 60%).
- The numerator, on the other hand, represents how often C and A *actually* appear together in the same order. In this example, that is 40% of the time. Taking the numerator and dividing it by the denominator, we get to how many more times C and A actually appear in the same order, compared to if there was no relationship between them (ie: that they are occurring together simply at random).

lift{A,C} = lift{C,A} = support{A,C} / (support{A} * support{C})
= (2/5) / (3/5 * 4/5)
= 0.2/0.48= 0.42

lift{A,D} = ?

Lift

What lift says about the relationship between A and B ?
Lift can take on the following values:

- lift = 1 implies no relationship.
  (ie: A and B occur together only by chance)
- lift > 1 implies that there is a positive relationship.
  (ie: A and B occur together more often than random)
- lift < 1 implies that there is a negative relationship.
  (ie: A and B occur together less often than random)

An example of Association Rules

Assume there are 100 customers

- 10 of them bought milk,
- 8 bought butter and
- 6 bought both of them.

bought milk => bought butter

- support =??
- confidence = ??
- lift = ??

An example of Association Rules

Assume there are 100 customers

- 10 of them bought milk,
- 8 bought butter and
- 6 bought both of them.

bought milk => bought butter

- support = P(Milk & Butter) = 6/100 = 0.06
- confidence = support/support(Milk) = 0.06/0.10 = 0.60
- lift = confidence/P(Butter) = 0.75/0.10 = 7.5

Sample Affinity application

- As an example, we will compute the support and confidence for the rule

  if a person buys apples,
  they also buy bananas.

- Data set
  - affinity_dataset.txt
- Each column (vertical row) represents each of the items:

bread, milk, cheese, apples, and bananas, respectively.

[0, 0, 1, 1, 1]

Sample Affinity application (cont.)

- Each of these features contain binary values, stating only whether the items were purchased and not how many of them were purchased.

- See how many rows satisfy our premise (bought apples)
- simply count the total for this implementation (no normalization for this example)

- Then set up some dictionaries to store the results for later use

Sample Affinity application (cont.)

Now we need to compute these statistics for all rules in our database. We will do this by creating a dictionary for both *valid rules* and *invalid rules*.

The key to this dictionary will be a tuple (premise and conclusion). We will store the indices, rather than the actual feature names. Therefore, we would store (3 and 4) to signify the previous rule

If a person buys Apples, they will also buy Bananas.

If the premise and conclusion are given, the rule is considered *valid*.

While if the premise is given but the conclusion is not, the rule is considered *invalid* for that sample.

Sample Affinity application (cont.)

- Record
  - the number of valid rules,
  - invalid rules, and
  - occurrences of each premise

- To compute the confidence and support for all possible rules,

Who benefits & How ?

(1) apples -> cheese

(2) cheese -> bananas

- A store manager can use rules like these to organize their store.

For example, if apples are on sale this week, put a display of cheeses nearby.

Similarly,

it would make little sense to put both bananas on sale at the same time as cheese, as nearly 66 percent of people buying cheese will buy bananas anyway—our sale won't increase banana purchases all that much.

Association analysis & Rules

- **Generally speaking, the best rules will have the high support, confidence and lift.**
- Association analysis results should be interpreted with caution.
- The inference made by an association rule does not necessarily imply causality. Instead, it suggests a strong co-occurrence relationship between items in the antecedent and consequent of the rule.
- Causality, on the other hand, requires knowledge about the causal and effect attributes in the data and typically involves relationships occurring over time

(e.g., ozone depletion leads to global warming)

Interpretation of values

- There may be instances where a low support is useful if you are trying to find "hidden" relationships.
- For product recommendation, a 50% confidence may be perfectly acceptable but in a medical situation, this level may not be high enough.
- The basic rule of thumb is that a lift value close to 1 means the rules were completely independent. Lift values > 1 are generally more "interesting" and could be indicative of a useful rule pattern.

Who benefits & How ?

Can be applied to other situations like:

- Click stream tracking, web mining
- Spare parts ordering and
- Online recommendation engines
- Bioinformatics
- Medical diagnosis
- Scientific data analysis