# Association rule learning

**Association rule mining**
- used to find hidden patterns = interesting relationships in transactions or co-occurrence data

**Interesting relationship**
- Interesting pattern:
  If events occur together more often than expected from their individual rates of occurrence
- 2 types
  - Frequent item sets: Collection of items that commonly occur together
    eg. {wine, diapers, soy milk}
  - Association rules: Strong relationship between two items
    eg. diapers ➞ wine
    If someone buys diapers, there's a good chance he will also buy wine

| Transaction number | Items |
|---|---|
| 0 | soy milk, lettuce |
| 1 | lettuce, diapers, wine, chard |
| 2 | soy milk, diapers, wine, orange juice |
| 3 | lettuce, soy milk, diapers, wine |
| 4 | lettuce, soy milk, diapers, orange juice |

**Transaction**
- is a set of items that co-occur in an observation - eg. in a market basket
- market basket: a common transaction in marketing
  = set of things that are purchased / considered for purchase at one time

- Idea of a "transaction":
  - is simply an observation of one or more data points that co-occur
  - **Any data points that co-occur are considered to be a transaction!!!!!!!!**
    => even if using the term "transaction" seems unusual in context
  - Example:
    If user visits multiple web pages during session => pages constitute transaction

=> **association rules can be applied to other kinds of data eg. general data frames**

**Discretize dataframe & segmention**
- association rule learning works on discrete data
- discretize columns => association rule learning can be applied
  => allows to explore segment associations

| lhs | | rhs |
|---|---|---|
| {age=19-24} | => | {Segment=Urban hip} |
| {age=19-24, income=Low} | => | {Segment=Urban hip} |

**Can be combined with other data**
- eg item profitability or customer characteristics

**Algorithms**
- Apriori algorithm
  - breadth-first search to count the support of itemsets
  - uses a candidate generation function which exploits downward closure property of support
- Eclat algorithm
  - ECLAT = Equivalence Class Transformation
  - depth-first search algorithm based on set intersection
  - suitable for sequential and parallel execution with locality-enhancing properties
- FP-growth algorithm: FP = frequent pattern

**Metrics for Interestingness**
X, Y itemsets; X ➞ Y an association rule and T the set of transactions of database

Motivation:
Select interesting rules from set of all possible rules, by putting constraints on measures of int.
Best-known constraints: minimum thresholds on support and confidence.

**Support** an itemset
- percentage of dataset that containing itemset
  => how frequently the itemset appears in the dataset
- Example:
  support({soy milk}) = 4/5, => of 5 transactions, 4 contain soyl milk
  support({soy milk, diapers}) = 3/5 => of 5 transactions, 3 contain both soy milk and diapers

**Confidence** of an association rule
- how often rule is true
  confidence( X ➞ Y ) = support(X ∩ Y) / support(X)
  = proportion of transactions that contains X which also contains Y
- Example
  Rule: {diapers} ➞ {wine}
  Confidence({diapers} ➞ {wine}) = support({diapers, wine})/support({diapers} )
  support of {diapers, wine} = 3/5; support of {diapers} = 4/5
  confidence(diapers ➞ wine) = 0.75
  => in 75% of the items in dataset containing diapers the our rule is correct

**Lift** of an association rule
- ratio of observed support to that expected if X and Y independent
  lift(X ➞ Y) = support(X ∩ Y) / (support(X) * support(Y))
  eg. lift({relish ➞ hot dogs}) = 3/5 / (4/5 * 4/5) = 50
  => combination {relish, hot dogs} occurs 50 times more often than expected
     if the two items were independent
- lift = 1: probability of occurrence of antecedent and consequent independent of each other
  No rule can be drawn involving those two events
- lift > 1: degree to which those two occurrences are dependent on one another
  Rule potentially useful for predicting the consequent in future data sets
- lift < 1: items are substitute to each other
  Presence of one item has negative effect on presence of other item and vice versa

**Connection**
Metrics tell different things => should exceed a minimum threshold for each:
Goal is to..
- Support: find item sets that occur relatively frequently in transactions
- Confidence: that show strong conditional relationships and
- Lift: that are more common than chance

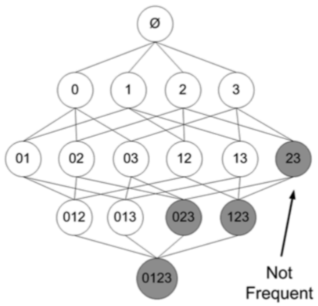
## Apriori Algorithm

**Motivation**
- Example: find all sets of items with support > 0.8
- Idea:
  - generate list of every combination of items
  - examine frequency of each itemset
  
  => Problem: computationally expensive

**Apriori principle**
- allows to reduce number of calculations to learn association rules
- reduces number of possible interesting itemsets

- Apriori principle:
  If an itemset is frequent, then all of its subsets are frequent

- Example:
  All possible itemsets of {0,1,2,3}
  If {0,1} is frequent, then {0} and {1} have to be frequent

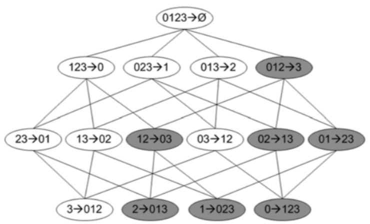



**Reversed apriori principle**
- Apriori rule itself does not help => but rule in reverse does

- Reversed apriori principle:
  If an itemset is infrequent, then its supersets are also infrequent

- stops exponential growt
  support of {2,3} computed ,then support for {0,2,3}, {1,2,3}, {0,1,2,3} needs not be computed
  => allows faster computation of frequent item sets

- Example:
  {2,3} is infrequent => {0,2,3}, {1,2,3}, {0,1,2,3} also infrequent




**Apriori Algorithm**
- finds frequent itemsets and learns association rules
  in databases containing transactions

  Algorithm:
  Input: minimum support, minimum confidence and dataset
  1. generate a list of all candidate itemsets with one item
  2. scan transaction data set for sets meeting minimum support level
  2a. sets not meeting minimum support level get discarded
  2b. combine remaining sets to make itemsets with two elements
  3. repeat 2. until all sets are tossed out

- Computation of association rules: (vs frequent itemset)
  - also give minimum level for confidence
  - Idea: generate a list of possible rules and test the confidence of each rule
  - Problem: possible to generate many association rules for each frequent itemset
    => number of rules can also be reduced to keep problem tractable
  - Principle
    If rule does not meet minimum confidence requirement, then subsets of that rule also won't
  - Example
    Rule 0,1,2 → 3 does not meet the minimum confidence,
    then any rule where left-hand side is subset of {0,1,2} will also not not minimum

# Discrete Choice

**Choice Modeling**
- Choice models are used to understand how product attributes drive customers' choices
- Used: to understand relationship between
  - attributes of products and
  - customer's choice among sets of products
- Goal: Understand how features and price affect which product a customer will choose
  Why customer chooses a specific product within a category?
- Question: How do different features and price affect the choice of product
- Choice Modeling: tries to answer question by analyzes choices to determine
  - which features of product are most attractive
  - how these features trade off against price

**Discrete Choice**
- tries to model decision process in a particular context of individual via
  revealed preferences = empirical data - eg transaction data
  stated preferences = made in a particular context or contexts
- understand and predict choice between multiple alternatives
- understand customer choice

**Discrete choice vs association mining**
- which products tend to occur together in the same shopping basket

**Setting**
- multiple alternatives: product options
- multiple attributes: product features
- multiple questions: in which to choose from multiple alternatives
  Example: conjoint analysis study
  3 alternatives with 4 attributes with different levels - one question:

**Which of the following minivans would you buy?**
Assume all three minivans are identical other than the features listed below.

| | **Option 1** | **Option 2** | **Option 3** |
|---|---|---|---|
| | 6 passengers | 8 passengers | 6 passengers |
| | 2 ft. cargo area | 3 ft. cargo area | 3 ft. cargo area |
| | gas engine | hybrid engine | gas engine |
| | $35,000 | $30,000 | $30,000 |
| **I prefer (check one):** | ☐ | ☐ | ☑ |

**Used in contexts**
- Retail purchase data: analysis of the data
- Complicated contexts: where people gather inf. from other sources beforehand - eg car sale
  Problems: difficult to reconstruct other possible options of customer
  1. no knowledge of other options
  2. no knowledge of other considered attributes
  => this context needs choice-based conjoint analysis

**Usages**
- understand how product attributes drive customers' choices
- Product in mind: Sensitivity Plots

**Motivation: Identify drivers of outcomes**
- often interested in identifying drivers of outcomes
- Example: drivers of revenue
  - use linear regression to identify & quantify the impact of several factors
  - BUT: outcome is continuous
- Not applicable: Product Choice
  - HERE: choose one of multiple options
  - do not observe a number or rating for product, but a choice of a product
  - => outcome is not continuous but u!
  - Question is not:
    „How much" = problems with outcome continuous variables
    "Which one" = problems of discrete choice analysis

  => uses multinomial logit model to analyze choice data

**Revealed and stated preference studies**
- choice modelling used in both
- RP studies:
  - use choices already made by individuals to estimate value they ascribe to items
  - individuals „reveal their preferences" (and hence their values = utilities) via their choices
- SP studies:
  - use choices made by individuals under experimental conditions to estimate values
  - individuals „state" their preferences via their choices

**Discrete choice experiment**
- = choice model
- mix of multivariate experimental design and conjoint analysis
- Steps
  1. Choose attributes and levels
  2. Choose multivariate design - eg. latin square design
  3. In that design: substitute design codes with attribute levels
  4. Do survey
  5. Analyse with multinomial logistic regression

**Advantages of DCE** vs revealed
- forces to consider trade-offs between attributes
- restricts attributes to stated attributes - controls for other factors

  => makes frame of reference explicit by explicitly giving attributes, levels & product alternatives

**Problems of revealed choice studies**
- which features user knows about
- did he consider all features known to him
- did he gather information from other sources => used different attributes

**Choice-based conjoint analysis**
- survey method (is discrete choice experiment?)
- customers asked to make choices among products with varying features and prices.
- survey choices analyzed using multinomial logit model

  => just as in analyzing real purchases

**Assumption of choice modeling: Utility = function of frequency**
Assumes that the utility that individual derives from item A over B
is function of the frequency that he chooses item A over B in repeated choices.

**Utility**
- utility is the value or benefit that individual derives from item A over item B
- utility function:
  - given set of alternatives facing individual - individual has preference ordering
  - assigns a real number to each alternative representing those preferences
  - preference order is needed to be completed and transitive
- in context of choice modeling
  - choice modeling to uses discrete choices - A over B; B over A, B & C
  - in order to infer positions of the items: A > B > C
  - on some relevant latent scale = utility

**Derived from utility theory**
- discrete choice models can be derived from utility theory
- behavior of person is utility-maximizing

  => person chooses the alternative that provides the highest utility
- utility is decomposed into:
  - variables that are observed
  - variables that are not observed

  $U_{ni} = \beta z_{ni} + \varepsilon_{ni}$

  $z_{ni}$ = vector of observed variables relating to alternative i for person n

  $\beta$ = corresponding vector of coefficients of observed variables

  $\varepsilon_{ni}$ = impact of all unobserved factors that affect person's choice

**From utility to mlogit model**
1. Begin with linear equation describing the utility of each alternative
   eg - eg. 4 wheels and price 100
2. Add an error-term that follows extreme-value distribution
   and is independent across alternatives
   => yields the utility
   utility is never observed, but assume that decisionmaker takes alternative with highest utility
3. Assuming alternative with highest utility chosen, probabilities for each option can be computed
   => involved integral and when solved gives simple formulas used by mlogit function

Defining property of choices:
Probability of choosing alternative depends on probability of other alternatives

```
v1 <- alpha * 4 + beta * 100
v2 <- alpha * 5 + beta * 150
v2 <- alpha * 2 + beta * 175
```

```
u1 <- v1 + error1
u2 <- v2 + error2
u3 <- v3 + error3
```

```
choice <- which.max(c(u1, u2, u3))
```

```
p1 <- exp(v1) / ( exp(v1) + exp(v2) + exp(v3) )
p2 <- exp(v2) / ( exp(v1) + exp(v2) + exp(v3) )
p3 <- exp(v3) / ( exp(v1) + exp(v2) + exp(v3) )
```

**Conjoint methods vs Discrete Choice Experiments**
- although often used interchangeably
- Discrete choice experiments
  have testable theory of human decision-making underpinning them = random utility theory
- Conjoint Methods:
  - decompose the value of a good (using statistical designs) from numerical ratings
  - no theory to explain about what the rating scale numbers mean
  - Disadvantages:
    - no trade-off information
      respondents do to differentiate between 'good' attributes and rate them all as attractive
    - personal scales vary from between individuals - value „2“ in „1 to 5“ differently
    - no relative measure how to compare something rated „2“ vs item rated „3“
      => no absolute ordering like a > b, b > c

**Discrete choice vs association mining**
- which products tend to occur together in the same shopping basket

**Probabilistic view**
- discrete choice models specify prob. that individual chooses option among set of alternatives
- probabilistic description of discrete choice behavior used:
  NOT: to reflect individual behavior that is viewed as intrinsically probabilistic
  BUT: to reflect lack of information that leads to describe choice in a probabilistic fashion
  => cannot know all factors affecting individual choice decisions
- Thus, models rely on stochastic assumptions to account for unobserved factors related:
  1. choice alternatives,
  2. taste variation over people (interpersonal heterogeneity)
  3. taste variation over time (intra-individual choice dynamics)
  3. heterogeneous choice sets

**Metric conjoint analysis**
- or ratings-based conjoint analysis
- respondents asked to rate single products instead of having to chose among sets of products
  => allows to use linear models instead of choice models
- Problem: more difficult for respondent to numerically rate a set of alternatives
  => reason why researches favor choice-based conjoint analysis
- choice-based needs different data structure than linear regression (each row = 1 observation)

**Data format**
- choice data uses different format - mostly „long“
- long: each profile is on its own line, column indicates question the profile was presented
  ==> each observation is described by multiple rows (eg. here: 3)
  vs linear regression: 1 row = 1 observation
  => but allows to different number of profiles in each question by including additional rows
- wide: each row = different question
- Example: first three rows describe first question was asked of respondent 1 => chose 3

| resp.id | ques | alt | carpool | seat | cargo | eng | price | choice |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | yes | 6 | 2ft | gas | 35 | 0 |
| 1 | 1 | 2 | yes | 8 | 3ft | hyb | 30 | 0 |
| 1 | 1 | 3 | yes | 6 | 3ft | gas | 30 | 1 |
| 1 | 2 | 1 | yes | 6 | 2ft | gas | 30 | 0 |
| 1 | 2 | 2 | yes | 7 | 3ft | gas | 35 | 1 |
| 1 | 2 | 3 | yes | 6 | 2ft | elec | 35 | 0 |

**Summarizing choice data**
- by computing choice counts
  => compute choice counts for each attribute before estimating choice model

```
> xtabs(choice ~ price, data=cbc.df)
price
  30   35   40
1486  956  558
```

## Fitting the Choice Model

**Fitting**
Usually using multinomial logit model (= conditional logit)

```
> m1 <- mlogit(choice ~ 0 + seat + cargo + eng + price, data = cbc.mlogit)
> summary(m1)
...
Frequencies of alternatives:
  pos 1   pos 2   pos 3
0.32700 0.33467 0.33833
...
Coefficients :
          Estimate Std. Error  t-value  Pr(>|t|)
seat7    -0.535280   0.062360  -8.5837 < 2.2e-16 ***
seat8    -0.305840   0.061129  -5.0032 5.638e-07 ***
cargo3ft  0.477449   0.050888   9.3824 < 2.2e-16 ***
enghyb   -0.811282   0.060130 -13.4921 < 2.2e-16 ***
engelec  -1.530762   0.067456 -22.6926 < 2.2e-16 ***
price35  -0.913656   0.060601 -15.0765 < 2.2e-16 ***
price40  -1.725851   0.069631 -24.7856 < 2.2e-16 ***
```

**Coefficients**
- Factors: interpreted relative to base levels of each attribute
  eg. seat7 measures attractiveness of 7 passenger relative to 6 passenger mini-vans
- negative sign: customers preferred 6 seat minivans to 7 seat minivans
- larger magnitude: stronger preference
  eg. strongly disliked electric engines (relative to base level gas) and disliked $40 K price
  => parameter are on logit scale and typically between -2 and 2
- Std. Error: how precise the estimate => more data, smaller standard error

**Intercept**
- „0 +“ in formula = model without intercept
- if intercept included:,
  - indicate preference for different positions in the question (left, right, or middle)
  - mlogit adds two additional parameters
  - called „lternative specific constants“ (ASC)
- people will choose mini-van because on left or right in survey question
  => estimated ASC not different from zero

```
                    Estimate Std. Error  t-value  Pr(>|t|)
pos 2:(intercept)  0.028980   0.051277   0.5652    0.5720
pos 3:(intercept)  0.041271   0.051384   0.8032    0.4219
seat7             -0.535369   0.062369  -8.5840 < 2.2e-16 ***
```

**Numeric predictors**
- can be used => eg. price as numeric gives better model

```
> lrtest(m1, m3)
Likelihood ratio test

Model 1: choice ~ 0 + seat + cargo + eng + price
Model 2: choice ~ 0 + seat + cargo + eng + as.numeric(as.character(price))
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -2581.6
2   6 -2582.1 -1 0.9054     0.3413
```

## Reporting a choice model

**Problem**
- coefficients are on an unfamiliar scale
- levels are interpreted relative to each other
  => better to present model using willingness-to-pay or making choice share predictions

**Willingness-to-Pay**
- allows to understand how much customers values various features
- can be computed for all levels of every attribute
  => willingness-to-pay more interpretable than attribute coefficients
- is a misnomer
- Meaning: price at which customer becomes indifferent between the to levels of factor
- Computation:
  Average willingness-to-pay for a particular level of attribute:
  Divide the coefficient for that level by the price coefficient
  Example: (Coefficient of cargo3ft) / (Coefficient of price)

- Interpretation

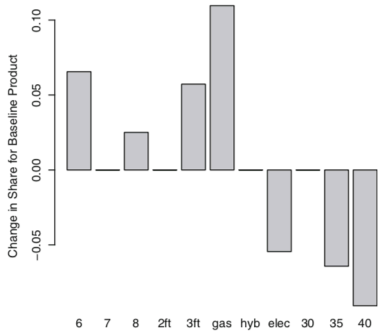
```
> coef(m3)["cargo3ft"]/(-coef(m3)["as.numeric(as.character(price))"]/1000)
cargo3ft
2750.601
```

  Customers would be equally divided between
  (Minivan with 2 ft of cargo space) & (minivan with 3 ft of cargo space that costs $2750.60 more)
  Or:
  $2750.60 is the (addtional?) price at which customers
  become indifferent between the two capacity options
  Or original „willingness to pay“ speak
  „Customer is willing to pay up to $2750.60 more for the additional 1ft (2ft vs 3ft) of space“

**Simulation of Choice Shares**
- willingness-to-pay more interpretable than attribute coefficients, but still difficult to understand
  => many analysts focus exclusively on using the model to make share predictions
- share simulator
  - allows to define a number of different alternatives
  - use model to predict how customers would choose among those new alternatives
  - similar to using predict function with different values
    => problem: no predict function in R => must write manually
- Example:
  expect choose the 7-seat hybrid with 2 ft of space at $30 K a more than 11 % of the time

```
> predict.mnl(m3, new.data)
        share seat cargo  eng price
8  0.11268892    7   2ft  hyb    30
1  0.43263922    6   2ft  gas    30
3  0.31855551    8   2ft  gas    30
41 0.07216867    7   3ft  gas    40
49 0.01657221    6   2ft elec    40
26 0.04737548    7   2ft  hyb    35
```

**Sensitivity Plots**
- often product design team has a particular product design in mind
- wants to know how share would change if they were to change their design
- Sensitivity Plot:
  shows how share would change if each of the attributes was changed one at a time
- Goal: see how share would change if different levels of the attributes were included
- Example
  - plan to build a 7-passenger hybrid with 2 ft. of cargo space and sell it at $30 K
  - Result
    7-seat design -> 6-seat design = increase share by just under 0.07
    $30K -> $35K = decrease share by about 0.06




**Multinomial choice with correlation among alternatives**
- standard logit model assumes that no correlation in unobserved factors over alternatives
  => not always suitable
- lack of correlation translates into pattern of
  substitution among alternatives that might not always be realistic in a given situatio
  => Independence of Irrelevant Alternatives (IIA) property: is name of substitution-pattern
- Red Bus/Blue Bus:
  - example in which pattern does not hold
  - consumer chooses between car and a bus - initial choice probabilities are equal:
    P(car) = P(bus)
  - second bus is introduced, and it is identical to the first bus
    => only difference is the color - first one is red and the second one is blue
  - sMNL model will evenly redistribute the probabilities to produce
    P(car) = P(red bus) = P(blue bus) = 1/3
    because the utilities are equal for both buses
  - more realistic assumption: ratio P(car) / (P(red bus) + P(blue bus)) will remain constant
    so that P(car) = 1/2 and P(red bus) = P(blue bus) = 1/3
- Solution: several models allow correlation over alt. and more general substitution pattern
  eg. Nested Logit Model
  Captures correlations between alternatives by partitioning the choice set into 'nests'

**Heterogeneous choice model**
- standard multinomial logit model estimates single set of coefficients for whole sample
  => same parameter for each individual
- BUT: different people have different preferences
  => models describing human behavior, heterogeneity is usually a good idea
- Problem of naive solution: single model for each person
  Not possible because not enough data
- Solution:
  - assume that coefficients for each person are drawn from a distribution
    => coefficients vary across indiduals
  - use multinomial logit model with random coefficients
    = hierarchical choice model = random coefficients models
    = heterogeneous choice model
- random coefficients / effects: coefficients varying across individuals
- hierarchical model: stacks together
  - upper level model: multivariate normal for coeffcients
  - lower level model: multinomial logit model for the coices
- Estimated:
  - using a multinomial logit model with random coefficients
  - in R: mlogit
  - needs vector with letters how distribution random coefs should follow across individuals

- Example
  - rpar = „random parameters“
  - twice as many parameters as m1
    1. Normal parameters:
       Describe average part worth coefficients across the population - eg. seat7
    2. Standard devation parameters
       Describing how parameters in 1 vary across the population - eg. sd.seat7

  - standard deviation parameter: indicate amount heterogeneity in parameters
    - eg. preference for 8 seats over 6 seats: much heterogeneity
      - sd.seat8 = 0.995 => larger than mean estimate seat8 = 0.39
        => suggests that some people prefer 6 seats to 8, while others prefer 8.
  - random coefficients section (of R output)
    - other way of seeing the heterogeneity
    - shows range of respondent-level coefficients
    - eg. seat8:
      - first quartile = –1.06 => indicating preference for 6 seats
      - third quartile = 0.281 => indicating preference for 8 seats
      - random coefficients as normally distributed
        hence, model assumes that majority of respondents are in the middle
        => slightly preferring 6 seats to 8
      - Resulting Action: large fraction of respondents also prefer 8 seats
        => offer a minivan with 6 seats and also minivan with 8 seats

```
Coefficients :
             Estimate Std. Error  t-value  Pr(>|t|)
seat7       -0.642241   0.070893  -9.0593 < 2.2e-16 ***
seat8       -0.390021   0.070460  -5.5353 3.106e-08 ***
enghyb      -0.926145   0.067456 -13.7296 < 2.2e-16 ***
engelec     -1.831864   0.083439 -21.9544 < 2.2e-16 ***
cargo3ft     0.550838   0.058459   9.4226 < 2.2e-16 ***
price35     -1.081310   0.070874 -15.2567 < 2.2e-16 ***
price40     -1.991787   0.085312 -23.3471 < 2.2e-16 ***
sd.seat7    -0.651807   0.101906  -6.3961 1.594e-10 ***
sd.seat8     0.995007   0.093397  10.6535 < 2.2e-16 ***
sd.enghyb    0.159495   0.137950   1.1562  0.247607
sd.engelec   0.973303   0.099850   9.7476 < 2.2e-16 ***
sd.cargo3ft  0.307194   0.131109   2.3430  0.019127 *
sd.price35  -0.260907   0.121369  -2.1497  0.031579 *
sd.price40   0.418148   0.128104   3.2641  0.001098 **
```

```
m1.rpar <- rep("n", length=length(m1$coef))
names(m1.rpar) <- names(m1$coef)
m1.rpar
 seat7    seat8 cargo3ft   enghyb  engelec  price35  price40
   "n"      "n"      "n"      "n"      "n"      "n"      "n"
```

```
random coefficients
        Min.    1st Qu.     Median       Mean    3rd Qu. Max.
seat7   -Inf -1.0818780 -0.6422410 -0.6422410 -0.2026039  Inf
seat8   -Inf -1.0611428 -0.3900209 -0.3900209  0.2811010  Inf
```

# Structural equation models

**Structural equation models**
- used to assess unobservable latent constructs = latent variables
- use measurement model that defines latent variables using one or more observed variables
- structural model imputes relationships between latent variables
- relationships estimated with independent regression equations

**Tool for testing theories**
- SEM allows building mathematical composite hypothesis reflecting a theory
- SEMs are representation of:
  - set of hypothesized relationships between observed variables and latent variables
  - into a composite hypothesis concerning statistical dependencies
- hypothesized relationships described by
  - parameters that indicate the magnitude of the relationship (direct or indirect)
  - that independent variables - either observed or latent
  - have on dependent variables - either observed or latent
- SEM allows testing of theories
  - representation of hypothesized relationships as testable mathematical models
  - allows quantification and testing of theoretical models
  => formulate theory has as SEM => test it against empirical data

**Used**
- to evaluate interconnections cannot be mapped to predictors and an outcome variable
  => eg. normal linear modeling
- to include unobserved latent variables and estimate relationships to one another or obs. data
- to estimate overall fit between observed data and proposed model with latent variables

**Related to**
- Linear modeling: estimate associations and model fit &
- Factor analysis: use latent variables

**In Marketing**
- to determine if concepts on survey match assumptions
  to assess whether items are related to theorized underlying construct => like FA
- Latent variables:
  - used to estimate the association between outcomes
    - eg. purchase behavior and underlying influencing attitudes -
    - like satisfaction and brand perception.
  - more complex models:
    - several latent variables are simultaneously associated with one another in multiple ways
    - brand perception, purchase intent, willingness to pay, and satisfaction
    - all relate to one another as latent constructs &
      relate in multiple ways to observed consumer behaviors - eg. purchases

**Motivation**
- real world hard to model
- often impossible to model every possible influence on outcome, BUT
- SEM allows improve these models by:
  - positing unobserved concepts that underlie the observed indicators
    i.e., constructs such as brand preference, likelihood to purchase, satisfaction
  - allowing to specify how those concepts influence one another,
  - assess overall congruence of model to data
  - determine if model fits data better than alternative models

**Example Problem**:
- Goal: modeling consumer's likelihood to purchase new product
- Problem:
  Likelihood influenced by many factors
  eg. prior product experience, perception of brand & features, price sensitivity, promo. effects
- Naive Idea:
  - collect survey data on stated likelihood to purchase and attitudes about brand.
  - model this as a linear relationship: purchase ~ perception
  - Problem
    - might find an effect or not,
    - BUT: model probably misses many other variable misses
    eg. perhaps effect thought was due to prior experience and not to brand
    perhaps didn't find effect because failed to account for promotional campaign
    that influences relationship
- Better
  - imperfect assessment of additional influences to improve our understanding
  - Improvement of statistical model,
    - even if incomplete, any unbiased capture of variance will improve other parts of model
    - eg.
      - only care about
        relationship between brand perception on likelihood to purchase
      - but if model includes
        - promotion and prior brand experience
          => model will capture some of the variance due to those factors
          ==> gives more realistic estimate for relationship between brand and purchase.

**SEM vs linear regression**
SEM similar to linear regression models but differ in 3 regards:
1. assess relationships with models more complex than simply predictors and outcomes
2. relationships allow for latent variables representing underlying constructs
   that are thought to be manifested imperfectly in the observed data
3. allows relationships to have multiple "downstream" effects
   eg. experience with a product - stated variable on a survey
   might relate to brand perception - latent construct expressed in several survey items
   which then relates to willingness to pay - a latent construct
   which relates observed behavior to purchase or not at a particular price point

**Creating a SEM**

- Idea:
  - create graphical path diagram of influences
  - estimating the strength of relationship for each path in model
- Paths concern two kinds of variables:
  - manifest variables: are observed - eg. have data points
  - latent variables: are thought to underlie the observed data
- Example: product involvement is latent factor
  - that underlies several other latent factors - eg. image involvement,
  - those factors in turn are observed as manifest variables on survey items

=> Structural Model: set of relationships among the latent variables
   Measurement model linkage between those elements and observed, manifest variables

**Two different SEM approaches**:

- Covariance based: CB-SEM
  - most common but more demanding
    => usually meant when talking about SEM
  - models attempt to account for as much of the total covariance in data as possible, among all observed and latent variables
  - strict assumptions:
    - data distributions: continuous data, normally distributed residuals
    - number of indicators per factor (> 3) & reliability of indicators

  => powerful if assumptions are met, otherwise…
- Partial least squares: PLS-SEM
  - more flexible approach
  - often able to fit models in situations where CB-SEM fails
  - Problem: model fit comparison not possible => no accepted measure of “goodness of fit”

**Example** Concept human intelligence

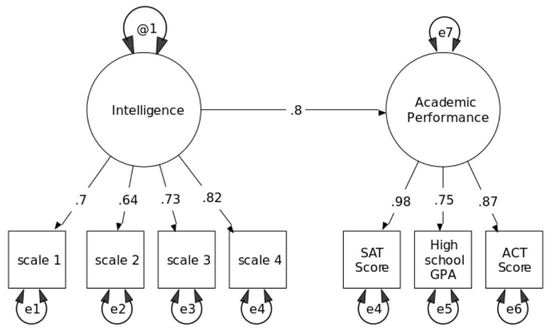
- SEM can impute relationships
  **between(!) latent variables** from(!) observable variables
  => here: intelligence -> academic performance

- Problem: cannot be measured directly as one could measure height or weight

- Instead: of measuring directly
  - develop hypothesis of intelligence
  - write measurement instruments with questions
    designed to measure intelligence according to their hypothesis

- SEM to test hypothesis using gathered data from intelligence test
  - intelligence = latent variable
  - test items = observed variables

- Model:
  - Intelligence: measured by four questions
    can predict
    Academic performance: measured by SAT, ACT, GPA
  - SEM diagrams:
    - latent variables: ovals
    - observed variables: rectangles
    - shows how error influences each intelligence question and SAT, ACT, and GPA scores, but not influence latent variables
  - SEM estimates parameters (arrows) in model to indicate strength of relationships
    => in addition to testing the overall theory,
    SEM shows which observed variables are good indicators of latent variables



# Confirmatory Factor Analysis

**Confirmatory Factor Analysis**
- popular type of SEM
- is special form of factor analysis
- to test whether measures of a construct are consistent with a theory of that construct (or factor)
- Goal: test whether the data fit a hypothesized measurement model
  => hypothesized model is based on theory or prior research
- Idea: Specify factor structure & asks
  1. "How well does the proposed model agree with the structure of the data?"
  2. "Is that model better than some other specified model?"

**Model development**
1. Develop hypothesis about what factors are underlying the measures used
2. Impose constraints on model based on these a priori hypotheses
   => forces model to be consistent with theory
   - Illustration: posited that two factors are accounting for covariance in measures
     - and factors are unrelated to one another
     - create model where the correlation between these factors is constrained to zero
     - model fit measures can then be obtained to assess
       how well proposed model captured covariance between all items or measures in model
     - If imposed constraints on model inconsistent with sample data,
       then the results of statistical tests of model fit will indicate a poor fit
     - If fit poor:
       - due to some items measuring multiple factors or
       - some items within a factor are more related to each other than others

**Exploratory factor analysis & confirmatory factor analysis**
- both:
  - to understand shared variance of measured variables
  - variance believed to be attributable to a factor / latent construct
    => but: EFA and CFA are conceptually and statistically distinct analyses
- EFA:
  - Goal: identify factors based on data and to maximize amount of variance explained
  - no specific hypotheses about:
    - how many factors will emerge,
    - what items or variables these factors will include
      => even if hypothesis exists - will not influence analysis & result
- CFA:
  - Goal: evaluates a priori hypotheses
    => largely driven by theory
  - requires hypothesis about
    - number of factors
    - if factors are correlated
    - which items/measures load onto and reflect which factors.
- Resulting contrast:
  - EFA: all loadings are free to vary
  - CFA: allows explicit constraint of certain loadings to be zero

=> EFA better for early exploratory work

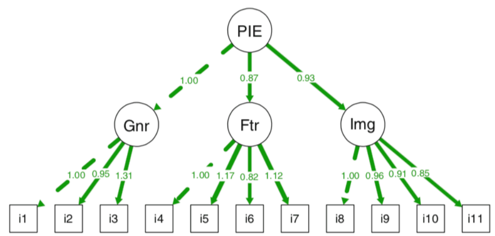
**CFA Process**
1. Specify factor structure (= structural model)
   = structure of unobserved variables + connected manifest scale items
2. Ask questions:
   A. "How well does the proposed model agree with the structure of the data?"
   => Are we able to **confirm** that **proposed model is good model for the data**
   B. "Is that model better than some other specified model?

**Latent factor model**
- unobservable factors are modeled as **latent variasbles**
- latent factors can relate to a higher-order latent constructs
- latent constructs are not directly observed
  Instead conceived to influence the survey items(not the other way around!) that **manifest** them
- Influence
  - higher order latent construct do not directly influence items on scale
  - they influence underlying factors as a higher order latent variable
  - manifest scale items are observed for each construct
- On survey: each factor is represented by a subscale comprising several items
- Factor model: model between unobserved factors
- Measurement model: model between unobserved and observed variables that manifest them

**Example** survey scale assessing: product involvement
- PIES model: "Product Involvement and Enthusiasm Scale" - latent construct (factor) model
- survey scale reflects model in which product involv. is hierarchical construct with 3 factors
- Three subscales reflecting those factors:
  - general involvement: with a product category
  - feature involvement
  - (Personal) Image Involvement

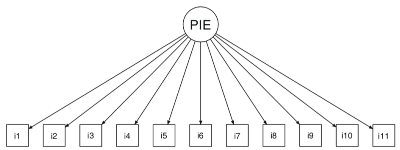
  Survey Items + PIES model with latent factors and manifest scale items + estimated coefs

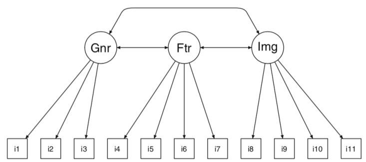
| Item | Text |
|---|---|
| *General scale* | |
| i1 | ____ are not very important to me. |
| i2 | I never think about ____. |
| i3 | I am very interested in ____. |
| *Feature scale* | |
| i4 | In choosing a ____ I would look for some specific features or options. |
| i5 | If I chose a new ____ I would investigate the available choices in depth. |
| i6 | Some ____ are clearly better than others. |
| i7 | If I were choosing a ____, I would wish to learn about the available options in detail. |
| *Image scale* | |
| i8 | When people see someone's ____, they form an opinion of that person. |
| i9 | A ____ expresses a lot about the person who owns it. |
| i10 | You can learn a lot about a person by seeing the person's ____. |
| i11 | It is important to choose a ____ that matches one's image. |




- Question:
  - „Is the PIES scheme good model for some set of survey responses for"
  - if able to confirm that PIES is good model
    then more confident in using this survey data to draw inferences about product involvement

- Alternative models to try



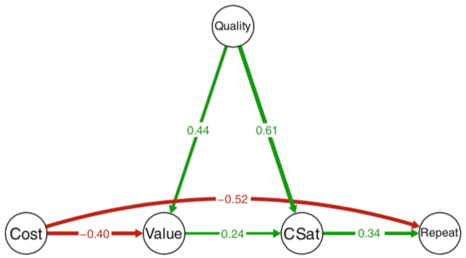
## General Models: Structural Equation Models

**Structural Equation Models**
- more general form of structural models
- where latent constructs may influence one another in more complex ways.

**CFA & SEM**
- CFA:
  - no directed arrows between latent factors
  - factors are not presumed to directly cause one another
- SEM: specifies particular factors and variables to be causal in nature
- CFA vs structural model
  - CFA: „measurement model“
  - Structural model: relations between the latent variables (with directed arrows)

**Example**
- customer satisfaction ratings and their effect on stated intention to repurchase
- cost of a product is associated with both perception of value and intent to repurchase
- perception of quality relates to both perceived value and satisfaction, which is then associated with repurchase
- data
  - responses to 15 satisfaction items
  - three items each for factors of
    Quality, Cost (fair pricing), Value, Customer Satisfaction. Repeat purchase intention

| Item | Text |
|---|---|
| *Quality* | |
| q1 | The quality of the HP printer I bought is excellent |
| q2 | HP printers are known to be highly reliable |
| q3 | I'm sure my HP printer will last a long time |
| *Cost* | |
| c1 | The HP printer was reasonably priced |
| c2 | HP sets fair prices for its products |
| c3 | The HP printers are no more expensive than others |
| *Value* | |
| v1 | I feel like I got good value for this purchase |
| v2 | The quality of the printer is worth its cost |
| v3 | I could tell my boss this purchase was good value |
| *CSat* | |
| cs1 | I am very satisfied with my newly purchase HP printer |
| cs2 | My printer is better than I expected it would be |
| cs3 | I have no regrets about having bought this printer |
| *Repeat* | |
| r1 | I would buy another HP if I had to buy another printer |
| r2 | I would buy other HP products |
| r3 | I would tell my friends and coworkers to buy HPs |




# Attribution

**Attribution (multi-touch attribution)**
Is the process of assigning credit to each marketing touchpoint = channel in customer journey in order to achieve conversion

**Goal of marketing attribution**:
- is to know the degree to which each channel contributes to the marketing success
- quantify influence each advertising impression has on a consumer's decision to convert
- allows to understand conversion path across marketing mix
- allows optimization of spend for conversions & compare value of different channels
- plan future ad campaigns by analyzing which ads were the most cost-effective

**Metrics**
CPC: cost per click
CPM: cost per thousand impressions
CPA: cost per action/acquisition
CPL: cost per lead
ROAS: return on ad spend
click-through conversion

**Understanding the customer journey**
Idea: Dissect the notion of : „From Think to Buy“
1. Descriptive statistics
2. Build basic attribution model - eg. last-click
3. Deeper investigation
   - top conversion path: most frequent customer journey leading to coversion
     => most important channel & sequences
   - time lag and path-length: in terms of interactions and days
   - how much conversions did channel create while not being last-click

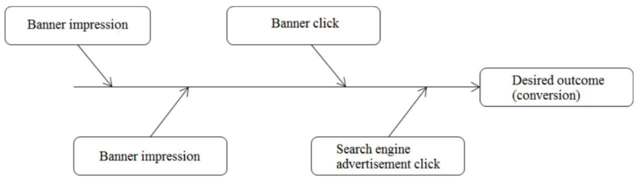
# Attribution Modelling

**Attribution Modeling**
Goal is to solve the multi-touch attribution problem

**Multi-touch attribution problem**
- Touch point
  - impression of an advertisement
  - click on ad
- MTA Problem:
  Problem of dividing credit among all advertisements a user saw before conversion
- Traditional way: assigns all credit based on rule - eg. last / first advertisement the user sees

**Customer journey**
Sequence of touchpoints of a user form start to conversion (or not)

**Example** of multi-touch attribution problem
1. customer sees a banner advertisement on a web-page
   => makes him aware and the consideration process begin
2. customer sees another advertisement of same product on a different webpage
   => desire begins & wants to know more about
3. uses google find out more & clicks on Ads on google
4. buys product on page
=> Problem: how should credit be allocated among these marketing channels




**Different Approaches to Attribution modeling**
- Traditional: heuristic / rule-based attribution models
- Algorithmic: each builds on top of mathematical theory to solve attribution problem
  - Game theory: eg. shapley-value
    =>
  - Probabilistic perspective: markov-chains
    => represents customer journey as markov graph
    -

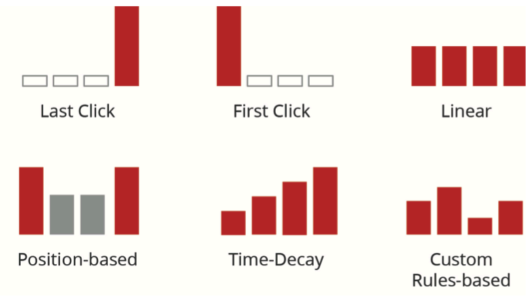
# Static / Heuristic attribution models

**Static (or heuristic) attribution models**
- assign credit between channels based on a heuristic / fixed set of rules
- most often used: last-click model
- disadvantages:
  - heuristic - human still needs to guess about performance
  - deciding which model for which business case
    => managerial decision necessary

**Last-click model**
- assigns all the credit to the last advertisement that was clicked just before the conversion.
  => **all credit** goes to **exactly one** marketing touchpoint
- user arrives directly: then credit will also be assigned to last ad user saw last
  if time-interval was not too long
- Advantage: easy to understand and easy to implement
- Disadvantage: ignores a considerable amount of available information

**Other models**
- First-Click: same as last-click, but credit goes to first touchpoint
- Linear model: attributes an equal share to all channels
- Time decay: higher impact to channels closer to conversion
- Position based: mix of first & last-click - higher attribution to first and last & some to in-between
- Custom rule based: allows to include individual assumptions

**Model Choice**
- depends on goal business and concrete ad
  => should be based on assumptions of business
- short sales cycle: last click
- brand awareness: first click more appropriate
  => selling shoes -> last click important ; selling cars - last-click not important
- each contact equally important: linear



## Dynamic / Algorithmic attribution models

**Dynamic attribution models**
- leverage data-driven approach by using different algorithmic techniques
- uses statistical modeling and ML techniques
- derive probability of conversion across all marketing touchpoints
- analyzes all customer journeys - wether leading to conversion or not to determine prob of conv.
  => in contrast to most heuristic models
- Algorithmic models: Shapley Value, Logistic / Linear Regression and Markov Chains
  => important: model interpretability
  eg. logistic regression is often appropriate due to ease of interpreting model coefficients

**Algorithmic attribution models**
- several methods can be used for algorithmic attribution
- eg. binary classification methods can be used to build models

**Logistic regression model**

$X \in \mathbb{R}$ = covariates

$A \in \{0,1\}$ = consumer saw ad or not

$Y \in \{0,1\}$ = conversion

**Consumer choice model**

$u(x,a) = \mathbb{E}(Y \mid X = x, A = a)$ where $X \in \mathbb{R}$ covariates and $A$ = Ads

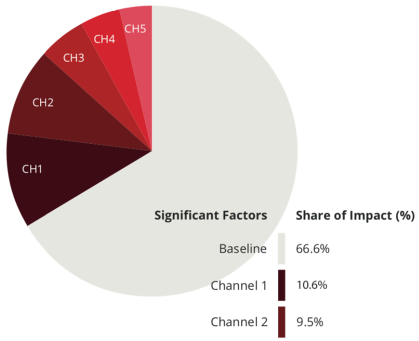
$$u = \sum_k A\beta^k \psi(x) + \epsilon$$

Covariates $X$ generally include different characteristics about
- ad served
  eg. creative, size, campaign, marketing tactic, etc.
- descriptive data about the consumer who saw the ad
  geographic location, device type, OS type, etc.

## Linear Regression for Attribution Modeling

**Linear Regression for Attribution Modeling**
- linear regression can be applied to nearly all kind of data related to the customer journey
- dependent variable: e.g. number of conversions, and
- independent variables: e.g. number of impressions per channel
  can also include offline information: seasonal data, weekdays or discount periods
- calculates impact of each potential factor on amount of conversions
  => helps to understand correlation between channels and predict future developments

**Decomposition**
- decomposition to impact of marketing channels or other variables of interest
- allows to calculate impact of each channel and how its impact changed over time
- calculation effect of each channels allows to to see what is driving conversions
  => allows to rank marketing channels
- Example: 66% baseline from direct traffic, channel 1 has higher effectiveness than channel 2



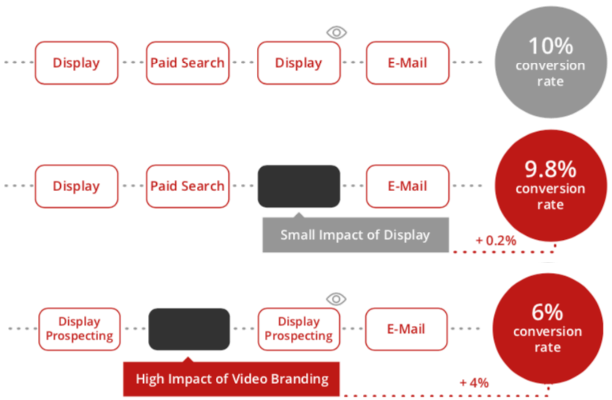
# Shapley value

**Shapley value**
- concept of cooperative game theory
- fairly assigns partial credit to each touchpoint of the marketing mix
- credit shows influence that channel had on achieving conversion
- influence increases when it has an impact over other channels

**From game theory to attribution**
- Shapley value is solution
  that solves the problem of equitably distributing the payoff of a game among players
- players an unequal contribution to that payoff
- applied to analogous situation of distributing credit for conversion among marketing channels

**Algorithm Idea**: Counterfactual Gain of touchpoint
- calculates influence of touchpoint
- Compare conversion probability between two similar sets of touchpoint sequences:
  1. sequence: users were exposed to a given touchpoint
  2. sequence: not exposed
  => the counterfactual gains of each touchpoint help to attribute its conversion credit
- Property: compares sequences
  => 1. takes order in which a touchpoint occurs into account
  2. assigns different credits based on touchpoint position

**Example** Low Impact & High Impact ad
1. Users shown sequence with display ad => 10% conversion
2. Users shown same sequence without display ad. => 9.8% conversion
===> marginal contribution of display channel = 0.2%
3. Absence of video branding => 6% conversion
====> overall contribution of Video Branding
compared to Display is significantly higher




**Shapely: Formal Context**
- is concept of cooperative game theory
- it assigns
  - a unique distribution among players of a total surplus generated by the coalition of all players
  - to each cooperative game
- Coalition game:
  Defined as
  - set $N$ of $n$ players
  - characteristic function $v : 2^N \to \mathbb{R}$
- Worth of coalition S
  $v(S)$: if S is coalition of players, then $v(S)$ is worth of coalition S
  = total expected sum of payoffs the members of S can obtain by cooperation

**Shapley value**

$\phi_i(v)$ = a **way to distribute the total gains** to the players, assuming that they all collaborate
= amount that player i gets given in a coalitional game $(v, N)$

$$= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(N-|S|-1)!}{N!}(v(S \cup \{i\}) - v(S))$$

Intuition:
- imagine the coalition being formed one actor at a time
- each actor demanding their contribution v(S∪{i}) – v(S) as a fair compensation
- and then for each actor take the average of this contribution over the possible different permutations in which the coalition can be formed

Interpretation:
Shapley value of channel i can be seen as the weighted sum of the incremental values that channel i adds to all the coalitions that don't contain this channel

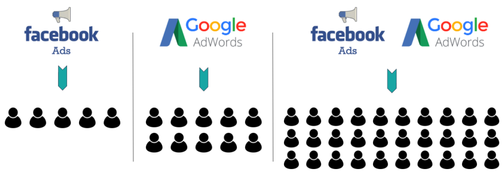
Intuition:
- Imagine that we're forming the grand coalition = coalition containing all the players
- forming it by entering each player in the coalition one player at a time
- each player receives the value by which he increases the coalition's worth

=> shapley value can then be seen as
average of the values that each player receives if the players are entered in a random order

**Properties**
- efficiency:
  - value of each channel's attribution is equal to the number of conversions it is accountable for,
  - sum of all channels' attributions equals to total number of conversions that were recorded
- individual rationality property:
  - guarantees that each channel will be accountable for at least the number of conversions that this channel can manage to generate by itself

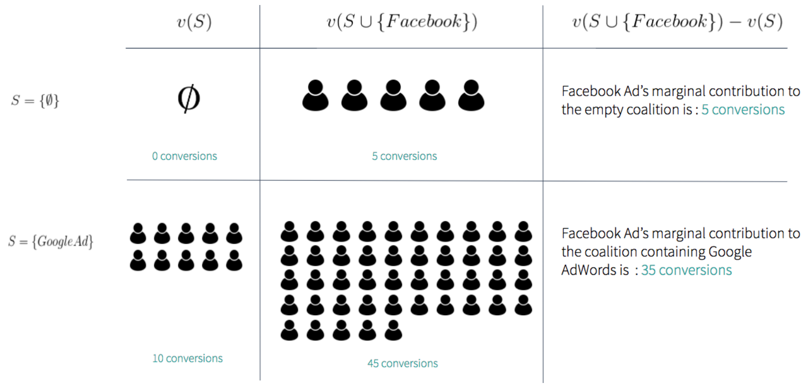
**Example**: Two marketing Channels

<u>5 customers</u> converted after clicking on a <u>Facebook ad</u>
<u>10 custumers</u> converted after clicking on a <u>sponsored Google</u> search result
<u>30 customers</u> converted only after clicking on <u>both ads</u>




<u>Use Shapley</u> to attribute proportion of total number of conversions that happened to each channel
Total conversion = 5 + 10 + 30 = 45

<u>Start with the Facebook</u>:
Only <u>two coalitions</u> that do <u>not contain this channel</u>
1. empty coalition S={ø}
2. coalition containing the channel Google AdWords S={GoogleAd}
=> these constitute the <u>summands of the Shapley equation</u>




Shapley Value: $v(GoogleAd, Facebook)$

$$v(\{GoogleAd, Facebook\}) = C(\{GoogleAd\}) + C(\{Facebook\}) + C(\{GoogleAd, Facebook\}) = 10 + 5 + 30 = 45$$

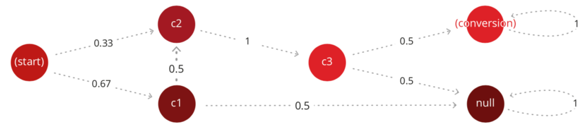
<u>Adding the weights for each summand</u>

Facebook: accounts for 20 conversions
Google: accounts for 25 conversions (after applying same steps for google)

# Markov chains for Attribution

**Markov chains for Attribution**
- represent <u>customer journey</u> as a <u>chain in directed Markov graph</u>
  - <u>vertex</u> = possible <u>state</u> = <u>channel</u>/touchpoint
  - <u>edges</u> = probability of <u>transition between the states</u> including <u>conversion</u>
- **<u>Attribution by</u>**: <u>computing model and estimating transition</u> probs for each touchpoint
- MC predict outcome <u>based on a user's movement through the states</u> of a stochastic process
- <u>state</u> = touchpoint = marketing channel user is exposed to
- <u>customer journey</u> = sequence of channels/touchpoints = Markov graphs
  - where user's conversion probability changes as he is expose to different channels over time
  - difference in conversion probability due to various touchpoints
    allows to measure a channel's impact on overall conversion
- <u>especially useful for attribution</u>
  - dependencies between channels are of particular interest
  - effects need to be quantified
- ChannelAttribution package in R

**Markov Graph for attribution**
- <u>Special states</u>:
  - <u>Start</u>: beginning of a customer journey
  - <u>Conversion</u>: successful customer journey,
  - <u>Null</u>: customer journeys that have not yet resulted in a conversion
- <u>Transition probability</u>:
  = probability that exposure to given channel results in a touchpoint with another channel
  - <u>cycles</u>: when sequence of two identical channels appears in a customer journey
- <u>Example</u>: 3 different channels : C1, C2, and C3




**Markov Property**
Present channel depends only on previous one
without incorporating previous touchpoints in the transition probability
=> memory free

**Higher Order Markov Chains**
- <u>Problem</u>: customer journey <u>should not be regarded as strictly Markovian</u>
  => due to interchannel effects.
- <u>Solution</u>: use higher-order MC
  => present state depends on more than one prior touchpoint in a customer journey

**Removal effect & Counterfactual Analysis**
- <u>removal effect of channel</u>:
  Is the <u>change in probability</u> for <u>reaching conversion state</u> from the "(start)" state
  when certain <u>channel is removed</u> from graph
  => allows counterfactual analysis
- <u>Counterfacutal Analysis</u>
  - <u>removal effect</u> represents <u>change in conversion rate</u> if a channel was not present at all
  - leverage the effect allows to <u>calculate exact contribution of each channel</u>

**Carryover & spillover effects**
- graph-based structure of Markovian chains <u>reflects sequential nature of customer journeys</u>
  => enables <u>analysis of interplay of channels</u>
- allows carryover and spillover between channels to be observed and quantified
  => only for higher-order models where probabilities based on consecutive prior states