Association Discovery

Finding Meaningful **Correlations**

Charles Parker
VP ML Algorithms, BigML, Inc

Association Discovery

- An unsupervised learning technique
  - No labels necessary
  - Useful for data discovery
- Finds "significant" correlations/associations/relations
  - Shopping cart: Coffee and sugar
  - Medical: High plasma glucose and diabetes
- Expresses them as "if then rules"
  - If "antecedent" then "consequent"
  - Significance measures
- BigML: “Magnum Opus” from Geoff Webb

Clusters

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

Clusters

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

similar

Anomaly Detection

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

Anomaly Detection

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

←anomaly

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

{customer = Bob, account = 3421}

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

{customer = Bob, account = 3421} → zip = 46140

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

{customer = Bob, account = 3421} → zip = 46140

{class = gas}

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

{customer = Bob, account = 3421} → zip = 46140

{class = gas} → amount < 100

Association Discovery

date	customer	account	auth	class	zip	amount
Mon	Bob	3421	pin	clothes	46140	135
Tue	Bob	3421	sign	food	46140	401
Tue	Alice	2456	pin	food	12222	234
Wed	Sally	6788	pin	gas	26339	94
Wed	Bob	3421	pin	tech	21350	2459
Wed	Bob	3421	pin	gas	46140	83
Thr	Sally	6788	sign	food	26339	51

Rules:

{customer = Bob, account = 3421} → zip = 46140

{class = gas} → amount < 100

Antecedent → Consequent

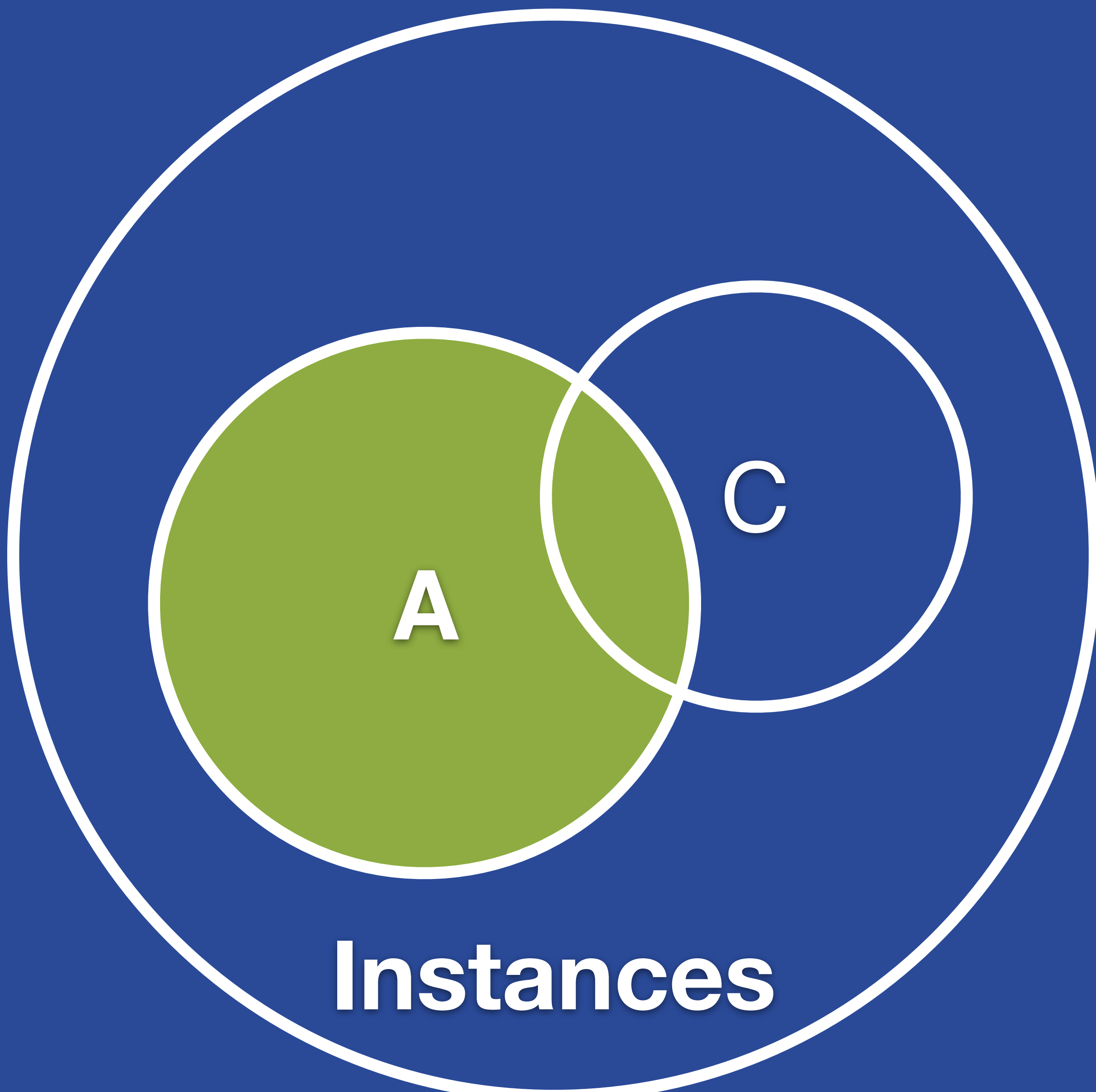
Use Cases

- Market Basket Analysis: Items that go together
- Data Discovery: how do instances relate?
- Behaviors that occur together
  - Web usage patterns
  - Intrusion detection
  - Fraud detection
- Bioinformatics
  - gene expression associated with outcomes
- Medical risk factors

What is interesting?

- In-frequent patterns can be strong, but are they interesting?
  - ❌ Vodka and caviar
  - ✅ Storms and high water sales
- Frequent patterns can be strong, but are they interesting?
  - ❌ Coffee and milk
  - ✅ High plasma glucose and diabetes
- “Frequency” isn’t the answer…
  - Depends on the data and domain
  - We need better metrics to define what is interesting

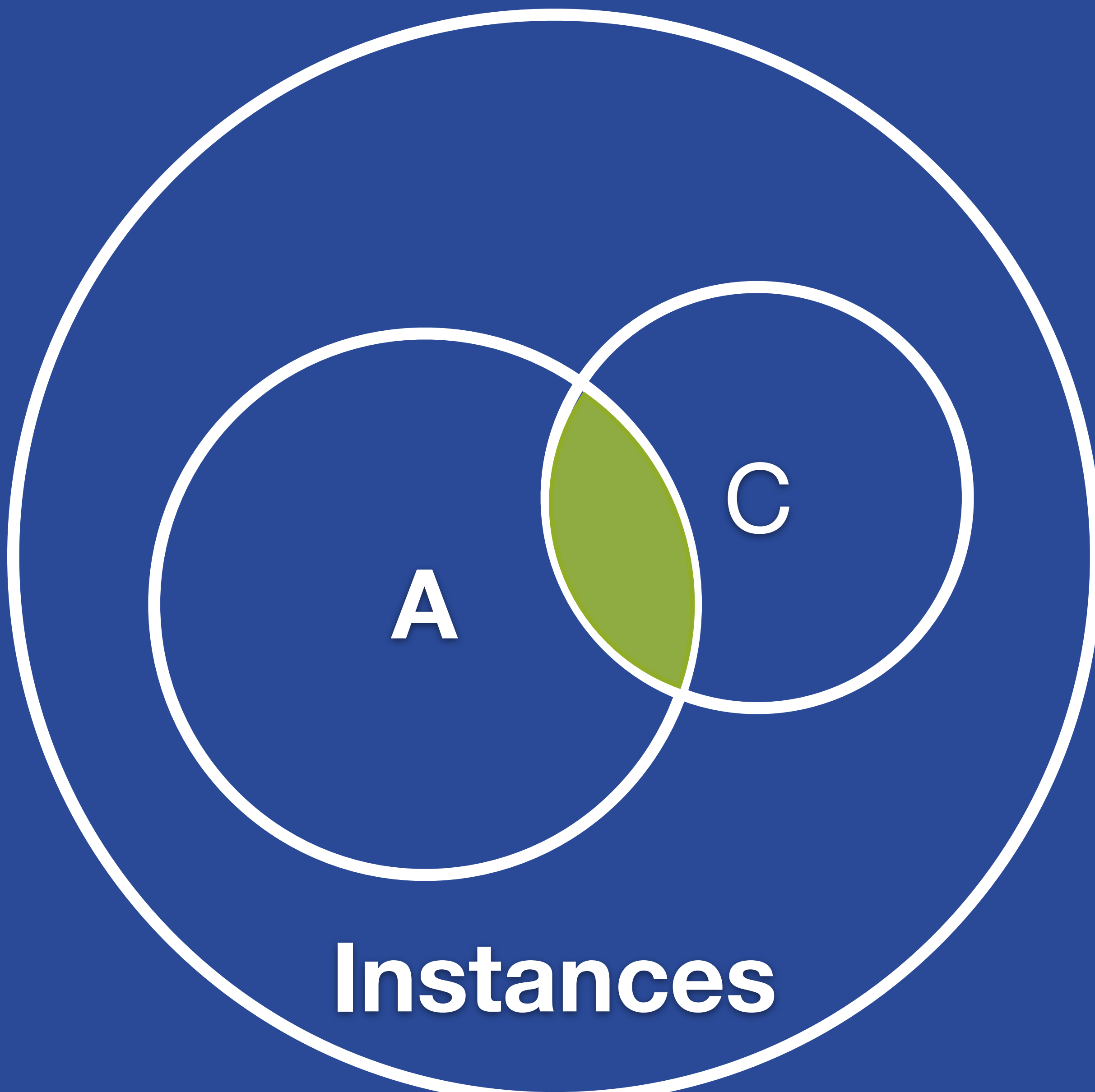
Association Metrics



Coverage

Percentage of instances which match antecedent "A"

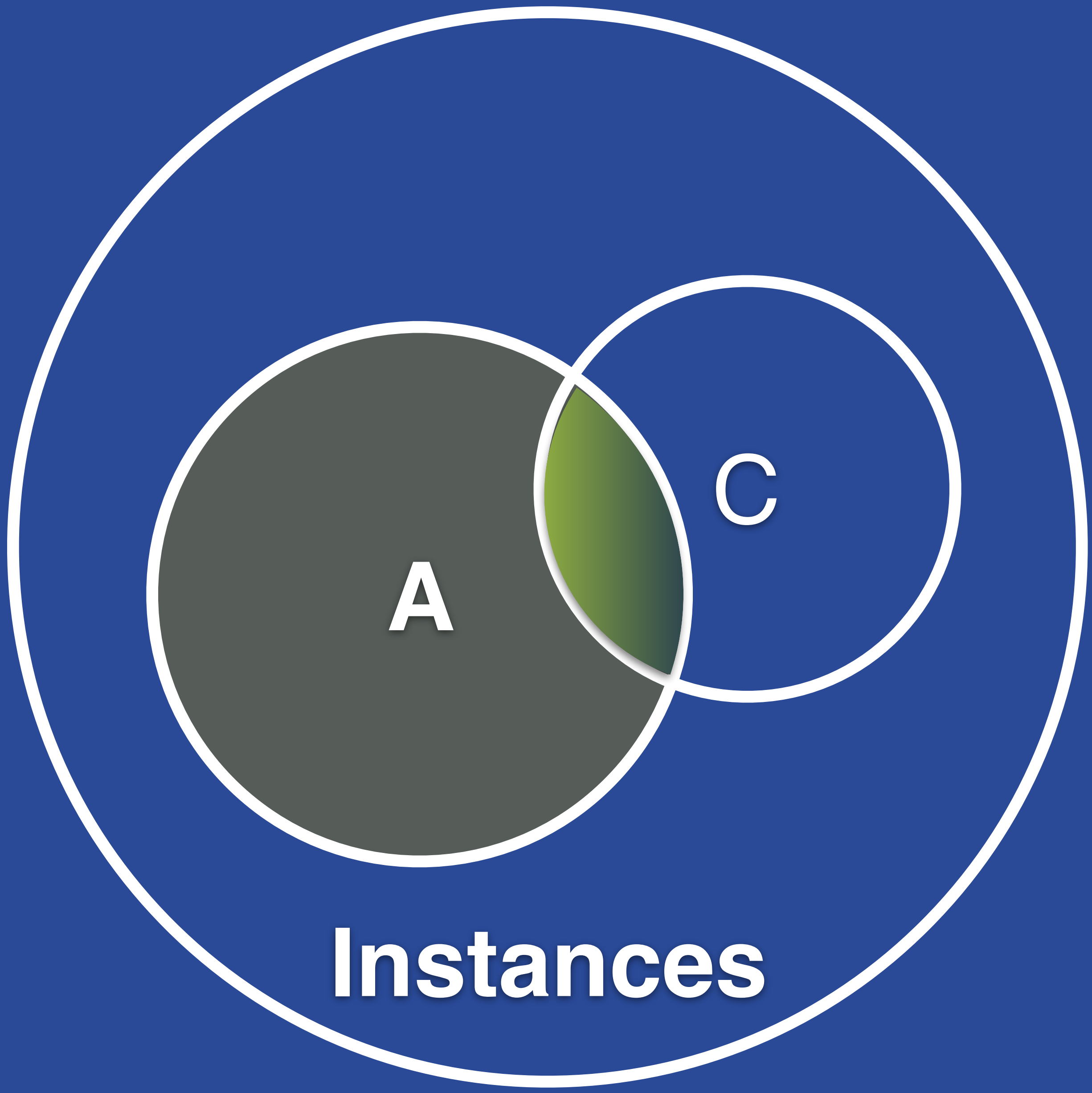
Association Metrics



Support

Percentage of instances which match antecedent “A” **and** Consequent “C”

Association Metrics



Confidence

Percentage of instances in the antecedent which **also** contain the consequent.

$$\frac{\text{Support}}{\text{Coverage}}$$

Association Metrics

0% → **Confidence** → **100%**

A *never* implies C

A *sometimes* implies C

A *always* implies C

Association Metrics

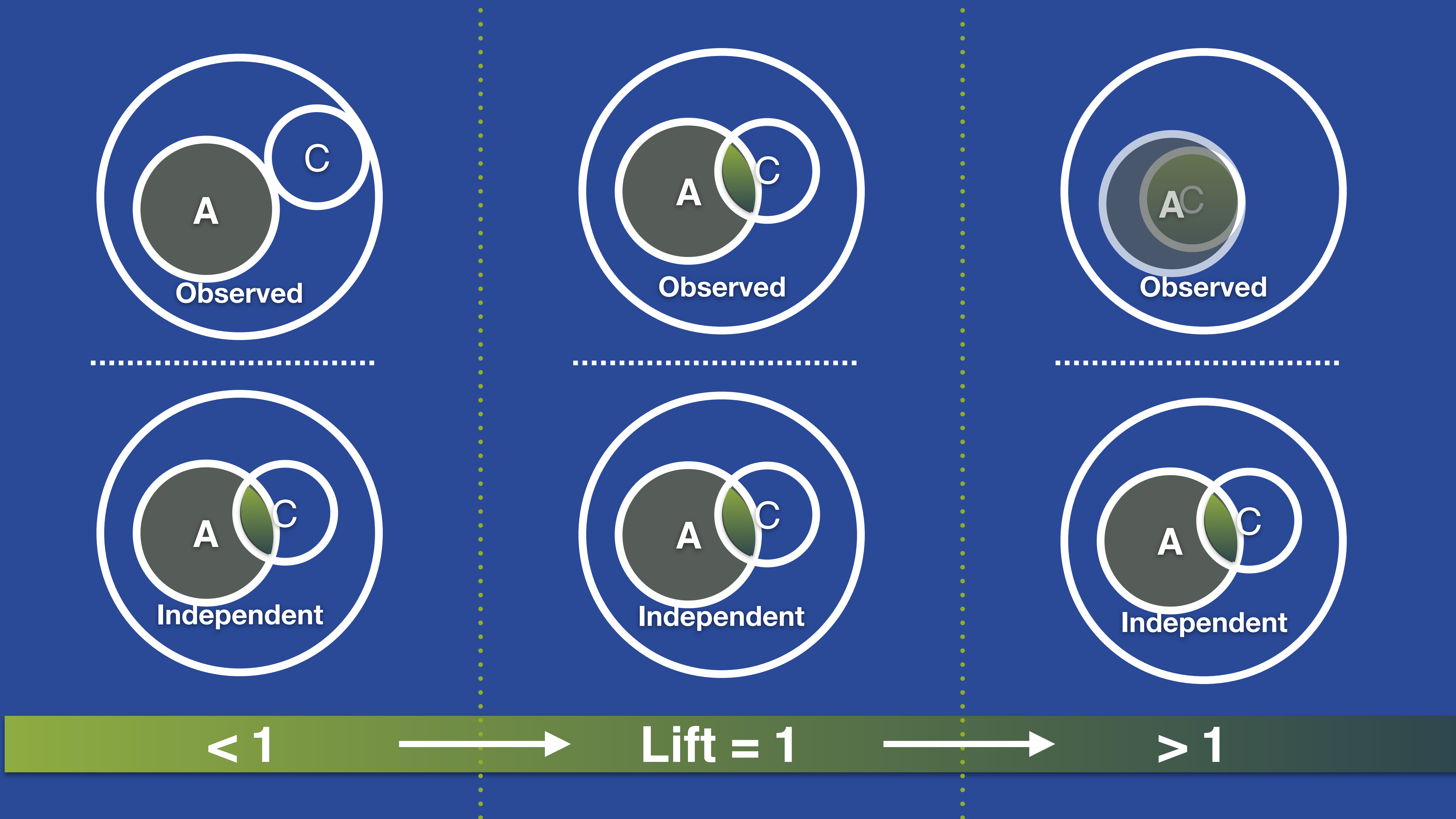
Lift

Ratio of observed support to support if A and C were statistically independent.

$$\frac{\text{Support}}{p(A) * p(C)} == \frac{\text{Confidence}}{p(C)}$$

Problem:
if p(C) is "small" then…
lift may be large.

Association Metrics

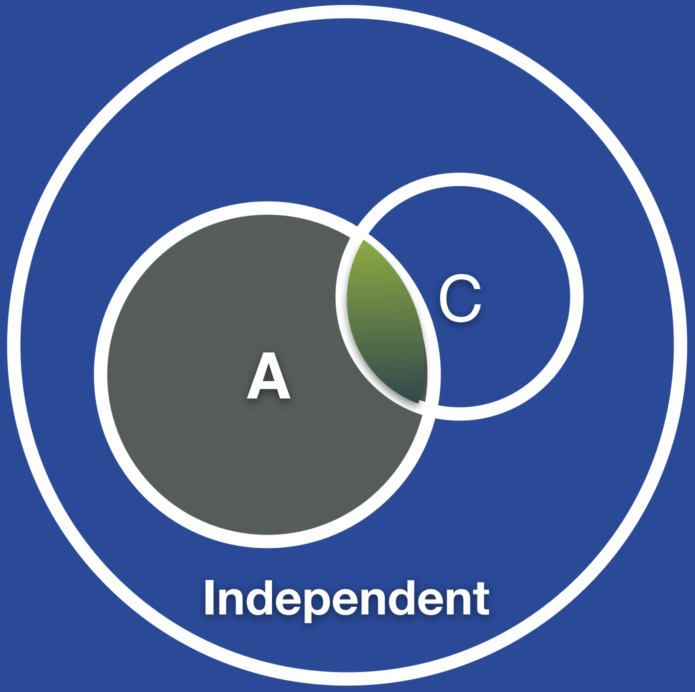


Negative Correlation

No Correlation

Positive Correlation

Association Metrics

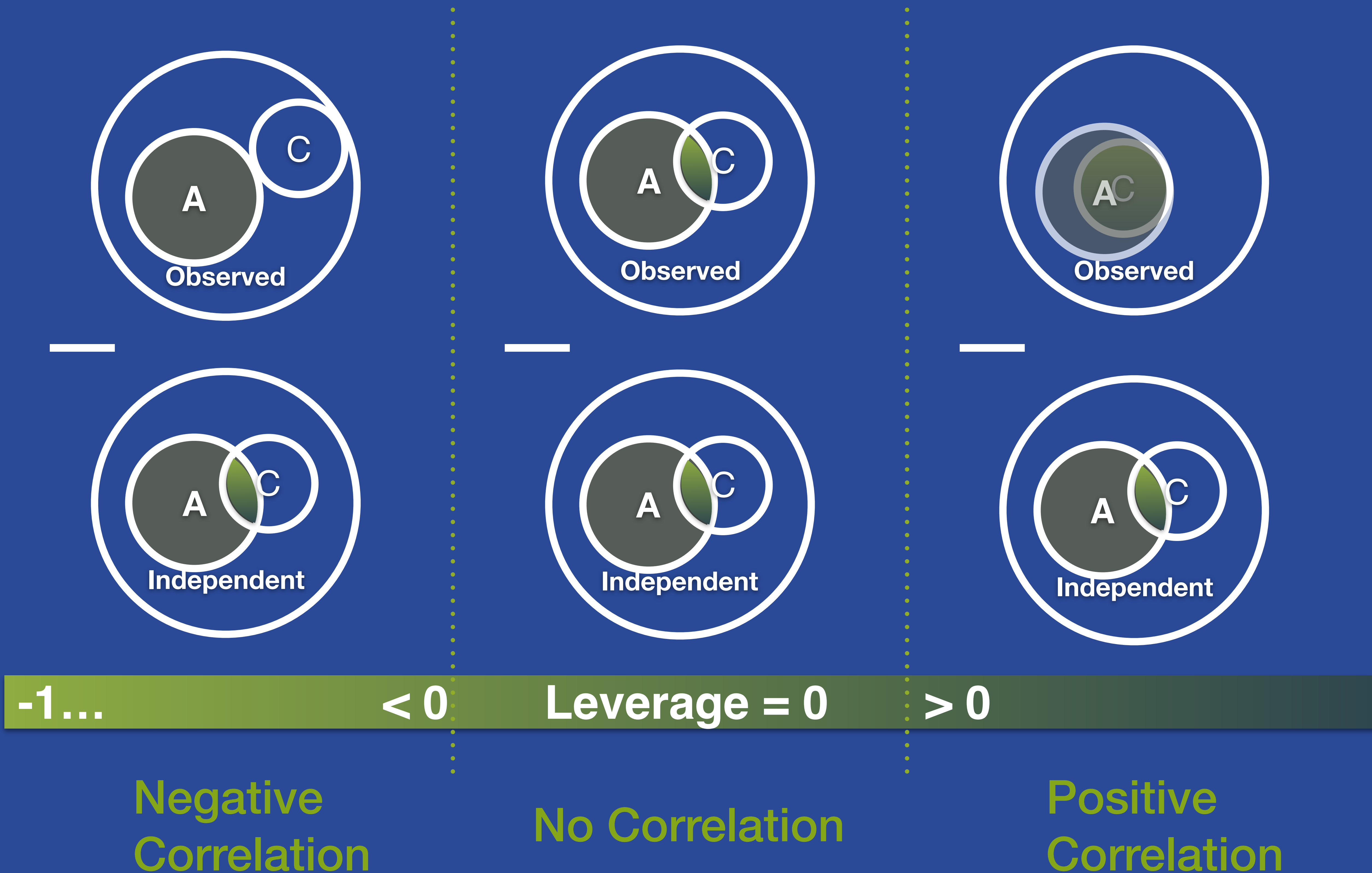


Leverage

Difference of observed support and support if A and C were statistically independent.

Support - [p(A) * p(C)]

Association Metrics



Magnum Opus

- Select measure of interest: Levarage, Lift, etc
- System finds the top-**k** associations on that measure within constraints
  - Must be statistically significant interaction between antecedent and consequent
  - Every item in the antecedent must increase the strength of association

Basic AD Configuration

1. **Search Strategy**: Support/Coverage/Confidence/Lift/<u>Leverage</u>
2. **Max Number of Associations**: 1 to 500 (default <u>100</u>)
3. **Max Items in Antecedent**: 1 to 10 (default <u>4</u>)
4. **Complement Items**: True / False
    - <u>False</u>: Coffee and…
    - True: **Not** Coffee and…
5. **Missing Items**: True / False
    - False: Loan Description contains "Ferrari" and…
    - True: Loan Description is missing and…

Data Types

1 2 3

1, 2.0, 3, -5.4 | **numeric**

A B C

true, yes, red, mammal | **categorical**

DATE-TIME

2013-09-25 10:02 | **date-time**

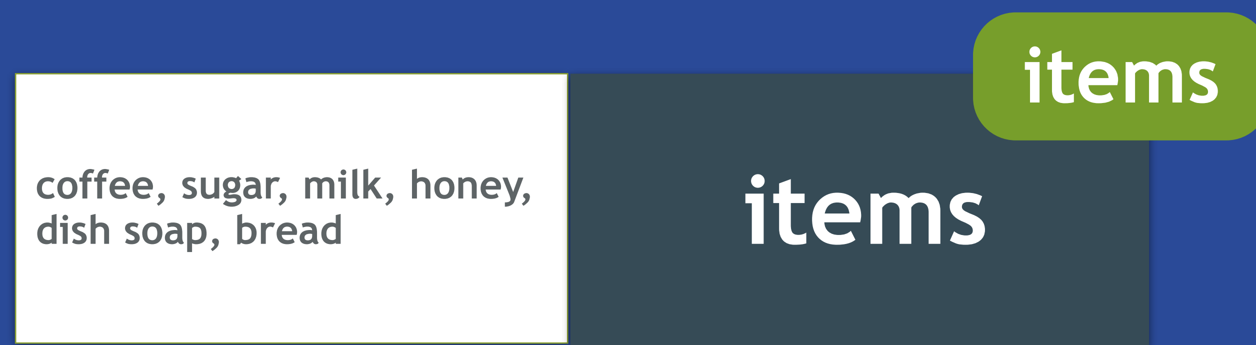
text

Be not afraid of greatness: some are born great, some achieve greatness, and some have greatness thrust upon 'em. | **text**

YYYY-MM-DD	YEAR	2013
YYYY-MM-DD	MONTH	September
YYYY-MM-DD	DAY-OF-MONTH	25
M-T-W-T-F-S-D	DAY-OF-WEEK	Wednesday
HH:MM:SS	HOUR	10
HH:MM:SS	MINUTE	02

"great" appears 2 times
"afraid" appears 1 time
"born" appears 1 time
"some" appears 2 times

Items Type



- Canonical example: shopping cart contents
- Single feature describing a list of items
- Each item separated by a comma (default)

Use Cases

- Dataset of 9,834 grocery cart transactions
- Each row is a list of all items in a cart at checkout

GOAL: Discover "interesting" rules about what store items are typically purchased together.

Association Demo #1

Use Cases

- Dataset of diagnostic measurements of 768 patients.
- Each patient labelled True/False for diabetes.

*GOAL: **Find general rules that indicate diabetes.***

Association Demo #2

Medical Risks

Decision Tree

If **plasma glucose** > 155
and **bmi** > 29.32
and **diabetes pedigree** > 0.32
and **insulin** <= 629
and **age** <= 44

then diabetes = TRUE

Association Rule

If **plasma glucose** > 146
then **diabetes** often TRUE

Association Demo #3

Your Turn!

- Starting from the 1-click Diabetes cluster (gmeans)
- Create a Batch Centoid and output as a Dataset
- Create an Association Discovery:
    - Specify the consequent as the cluster assignment
- Can you generalize any of the cluster groups?

bigml®