bigml

Machine Learning for everyone

Easily add data-driven decisions and predictive analytics to your company

The Need for Machine Learning

Talk	Text	Purchases	Data	Age	Churn?
148	72	0	33.6	50	TRUE
85	66	0	26.6	31	FALSE
183	64	0	23.3	32	TRUE
89	66	94	28.1	21	FALSE
115	0	0	35.3	29	FALSE
166	72	175	25.8	51	TRUE
100	0	0	30	32	TRUE
118	84	230	45.8	31	TRUE
171	110	240	45.4	54	TRUE
159	64	0	27.4	40	FALSE

…. but this is a simple example

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

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

text

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

“great” appears 2 times

“afraid” appears 1 time

“born” appears 1 time

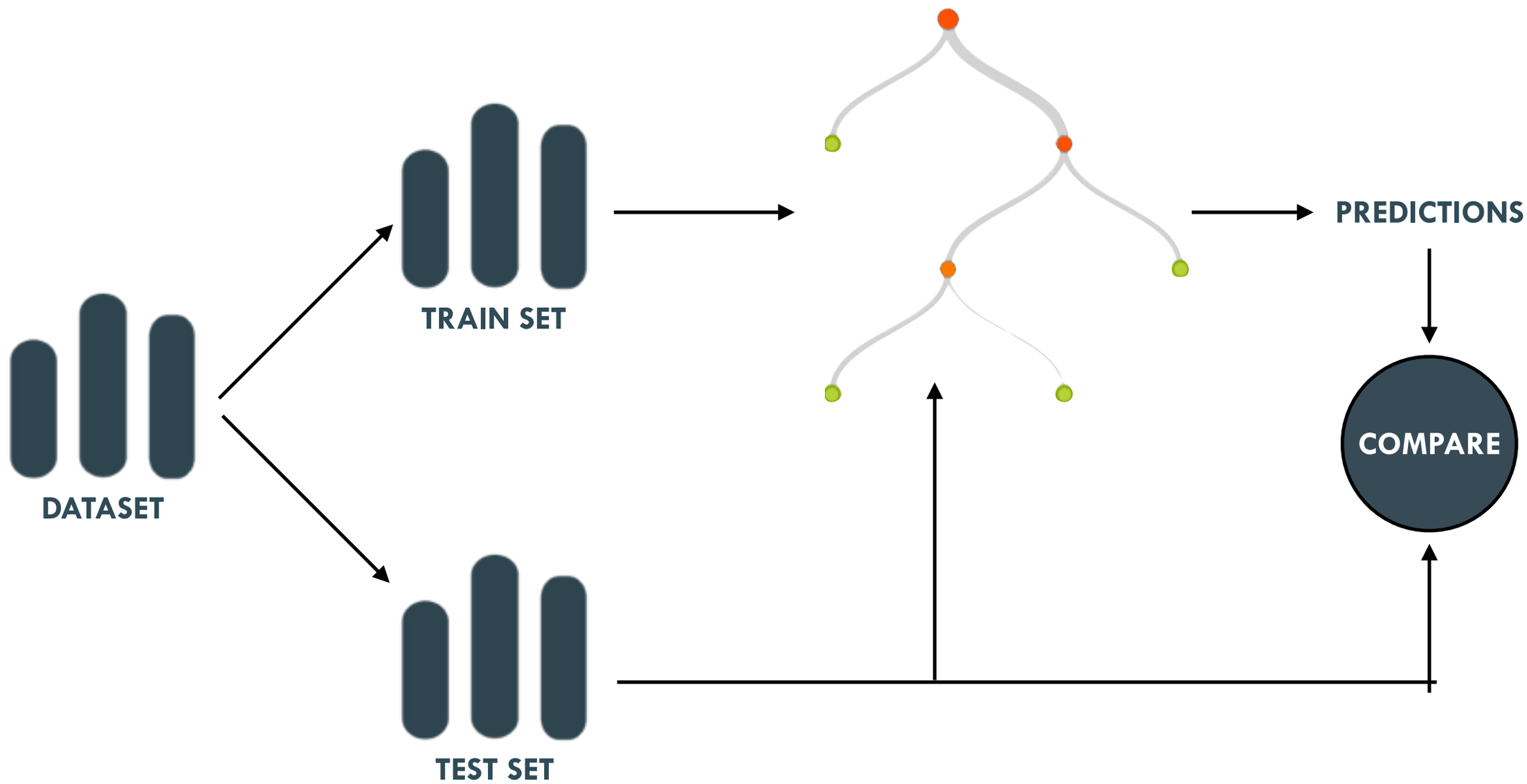
“some” appears 2 times

Text Analysis

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

great: appears 4 times

Evaluations



Ensembles

Dia	Color	Shape	Fruit
4	red	round	plum
5	red	round	apple
5	red	round	apple
6	red	round	plum
7	red	round	apple

What is a round, red 6cm fruit?

All Data: “plum”

Sample 1: “plum”

Sample 2: “apple”

Sample 3: “apple”

} **“apple”**

Bagging!
Random Decision Forest!

Supervised Learning

Classification

animal	state	…	proximity	action
tiger	hungry	…	close	run
elephant	happy	…	far	take picture

label

Regression

animal	state	…	proximity	min_kmh
tiger	hungry	…	close	70
hippo	angry	…	far	10

Multi-Label Classification

animal	state	…	proximity	action1	action2
tiger	hungry	…	close	run	look untasty
elephant	happy	…	far	take picture	call friends

Unsupervised Learning

Clustering

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
The	Sally	6788	sign	food	26339	51

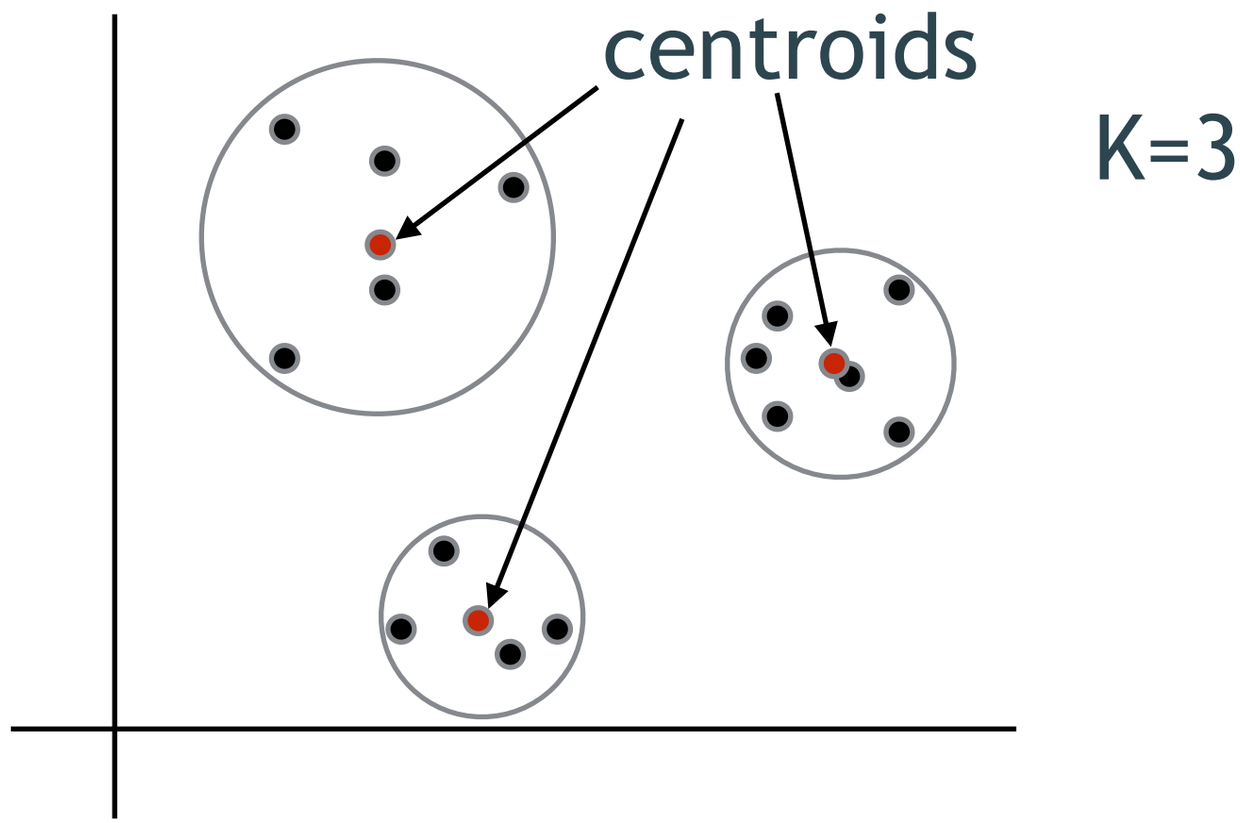
similar

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
The	Sally	6788	sign	food	26339	51

unusual

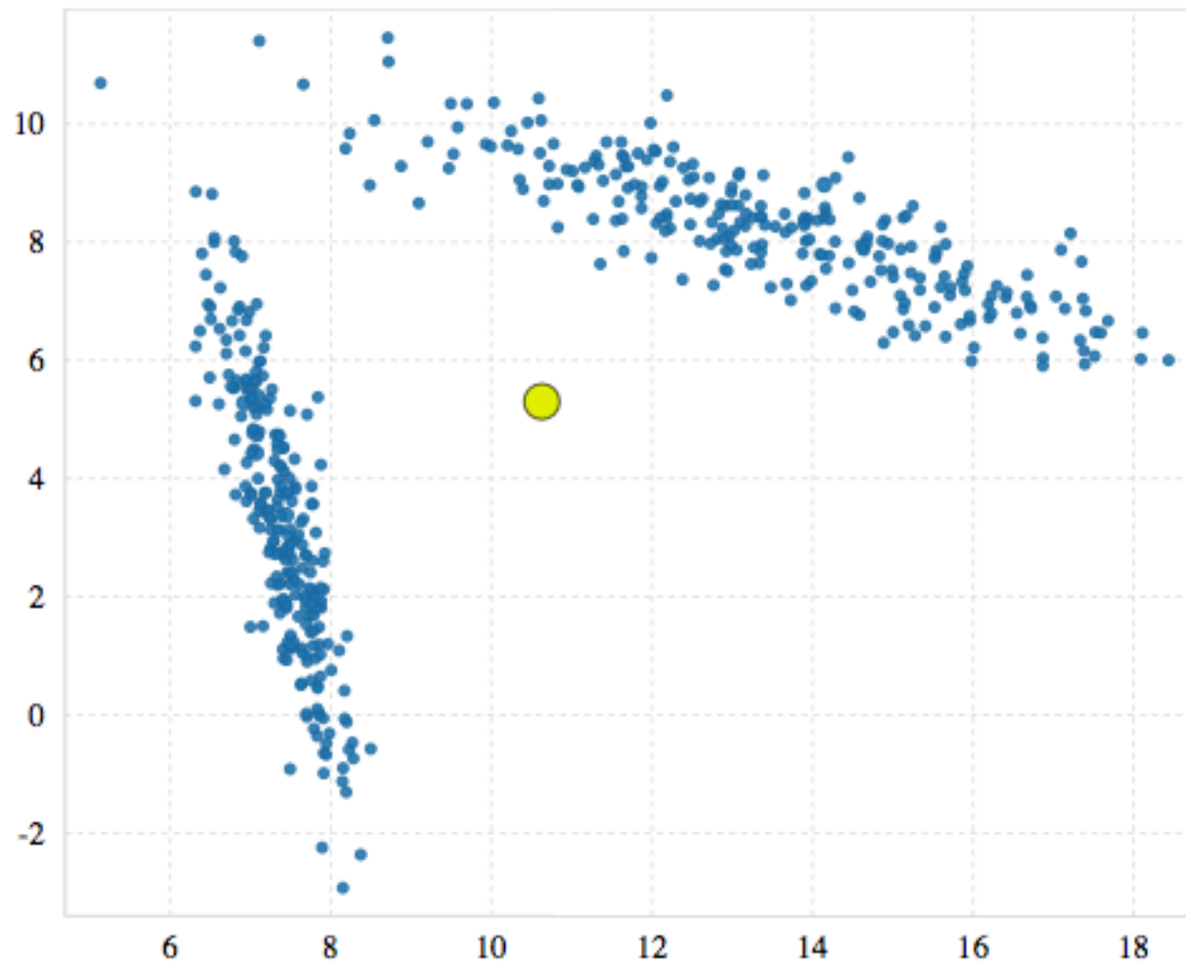
Clustering Basics



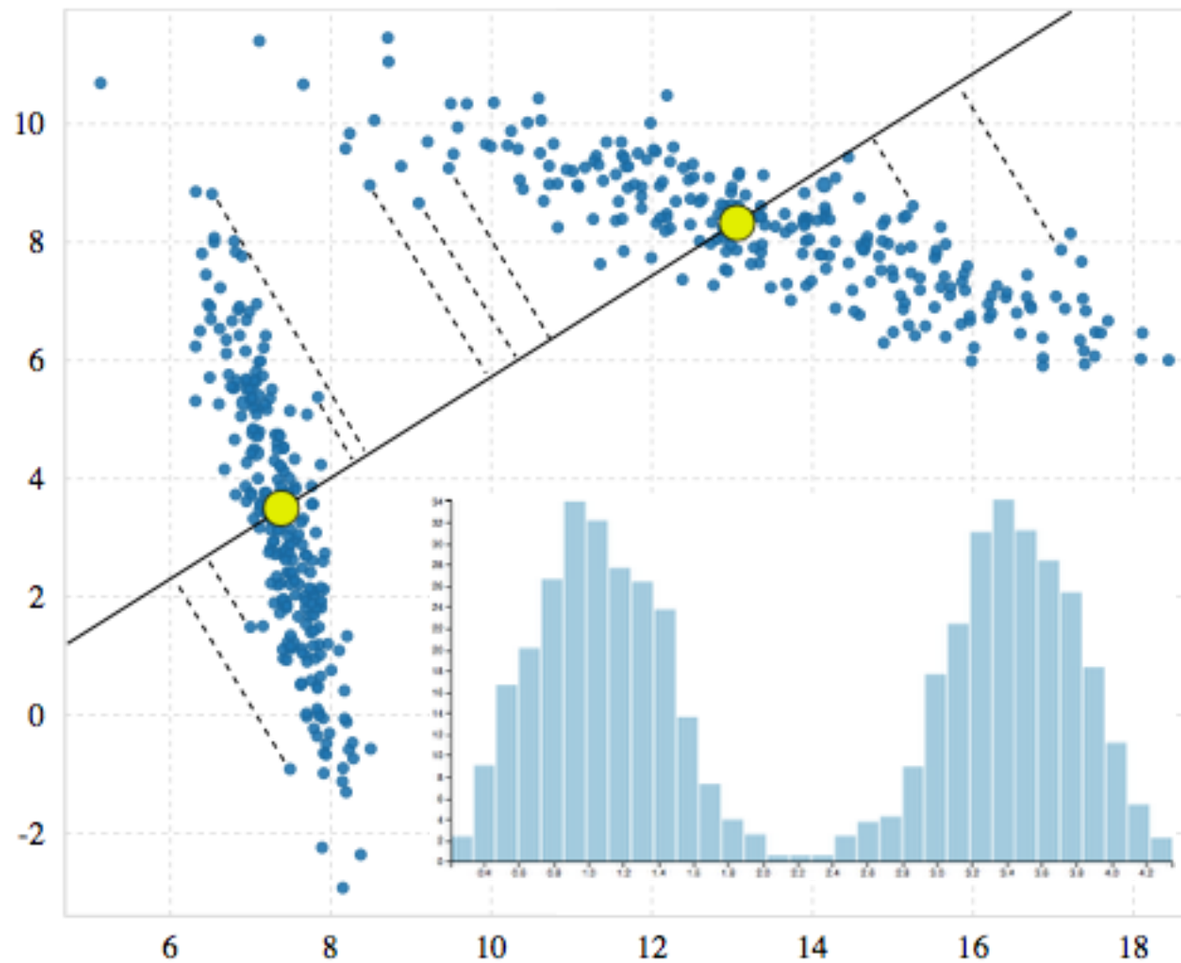
Clustering

- Customer segmentation
- Item discovery
- Data summarization / compression
- Collaborative filtering / recommender
- Active learning

Finding K: G-means



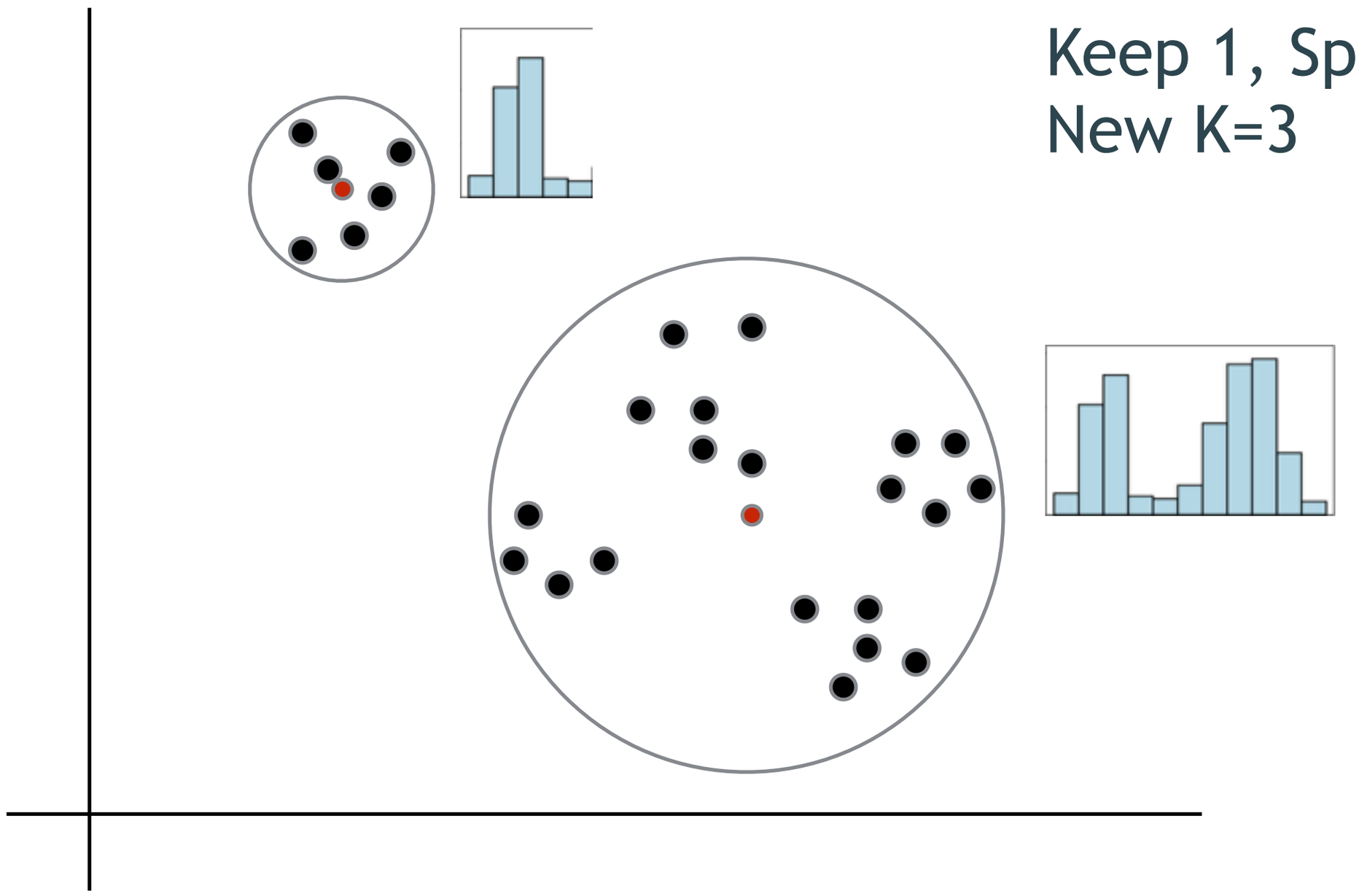
Finding K: G-means



Finding K: G-means

Let K=2

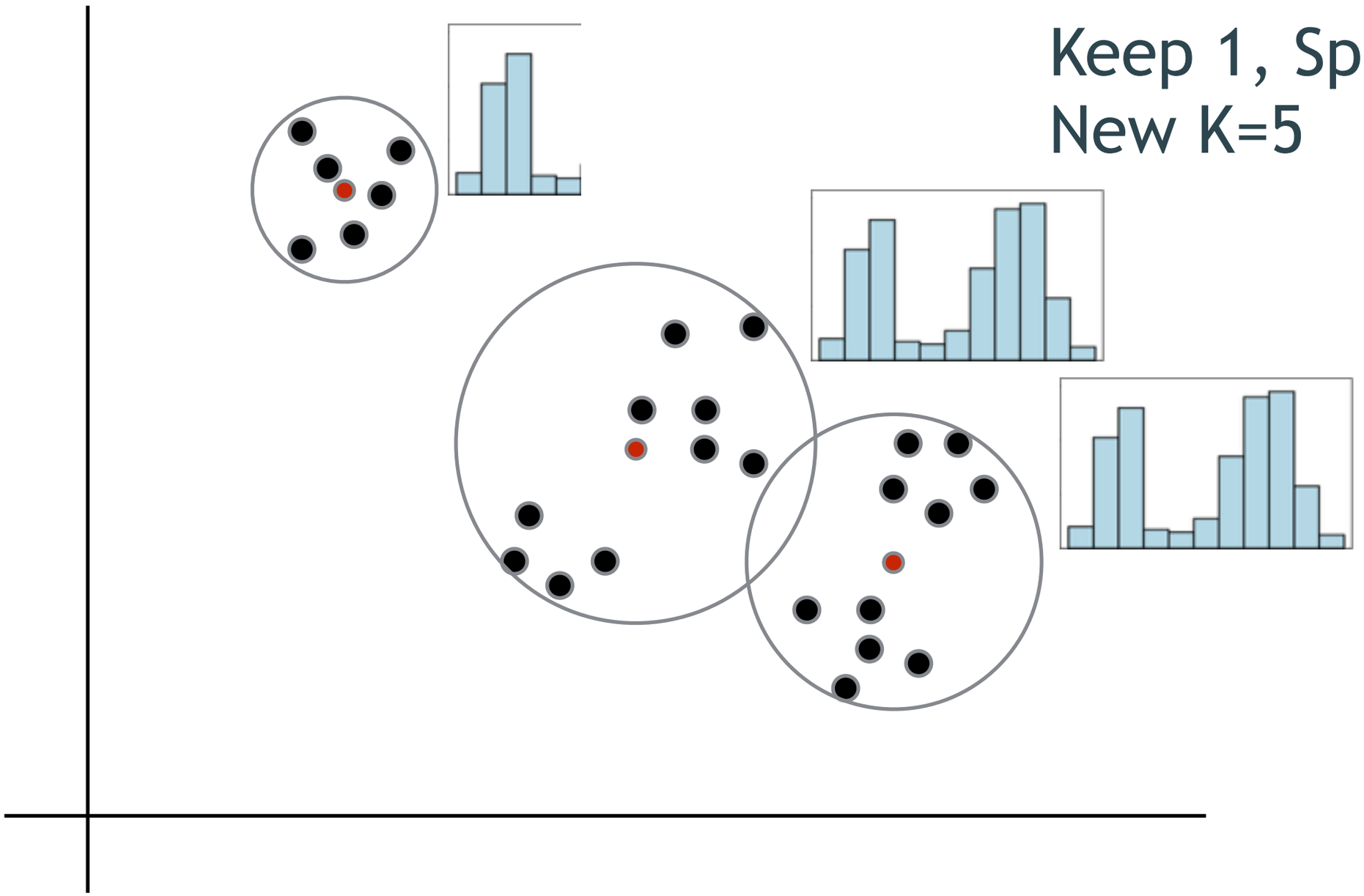
Keep 1, Split 1
New K=3



Finding K: G-means

Let K=3

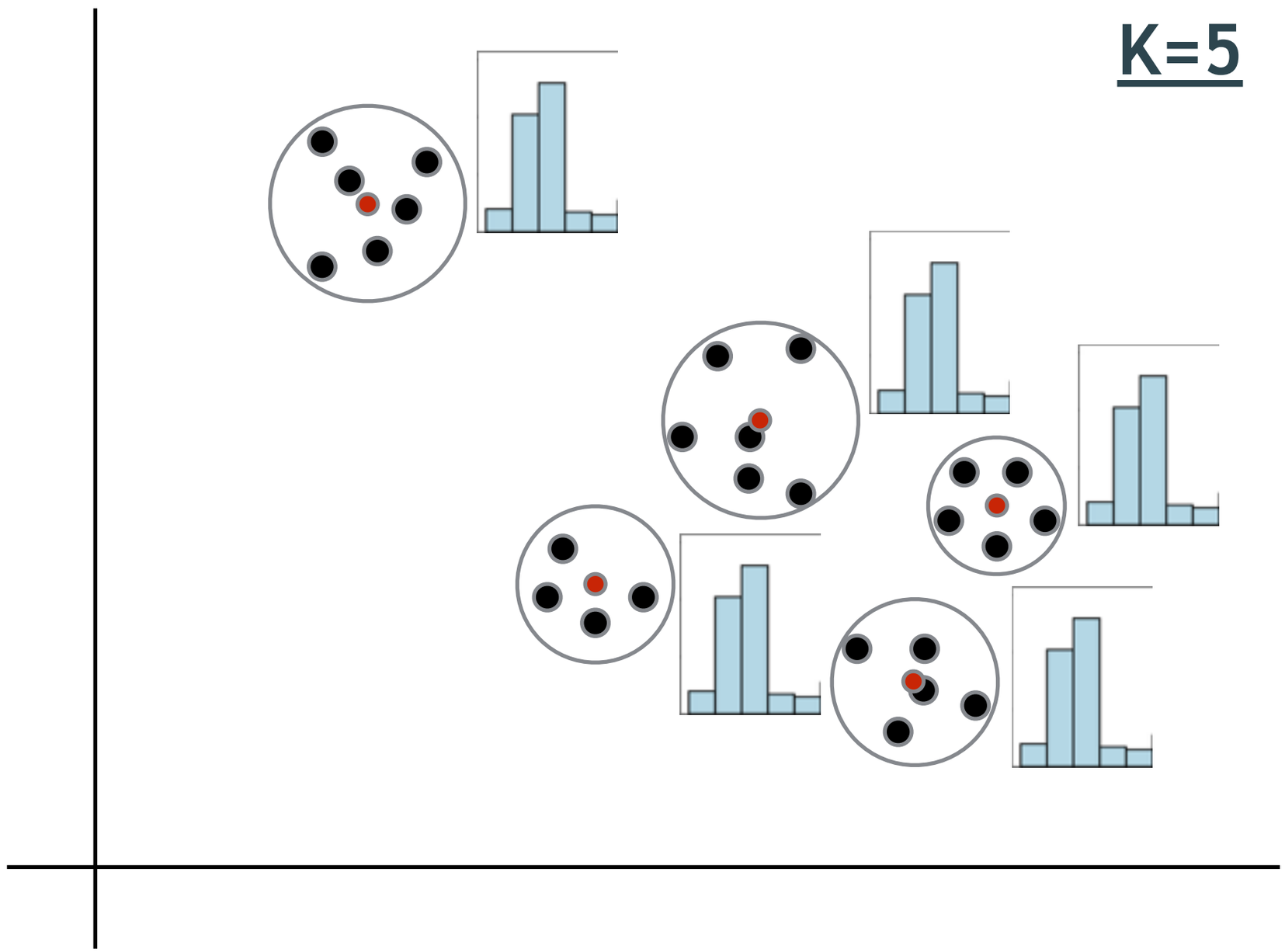
Keep 1, Split 2
New K=5



Finding K: G-means

Let K=5

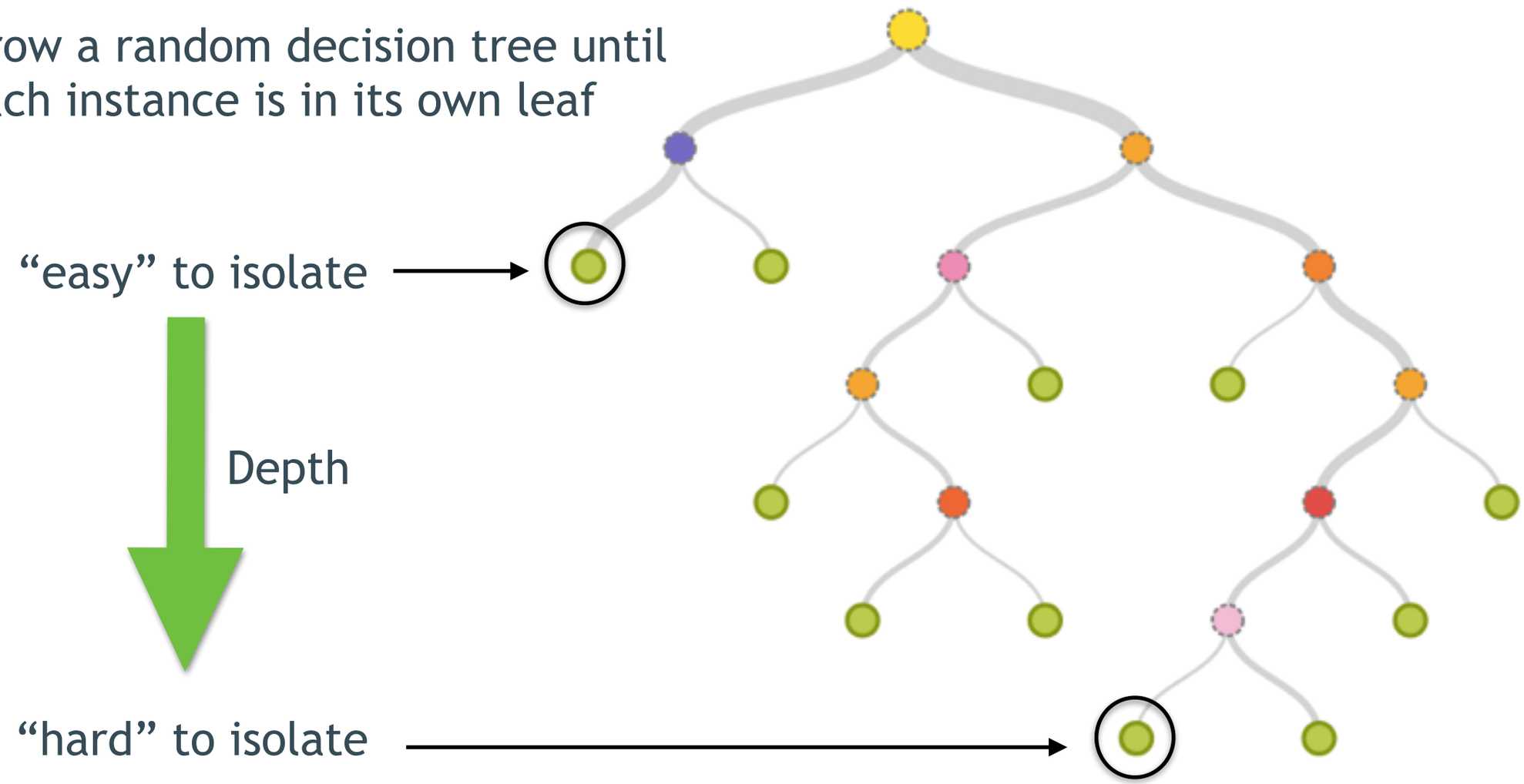
K=5



bigml Anomalies

Isolation Forest:

Grow a random decision tree until each instance is in its own leaf

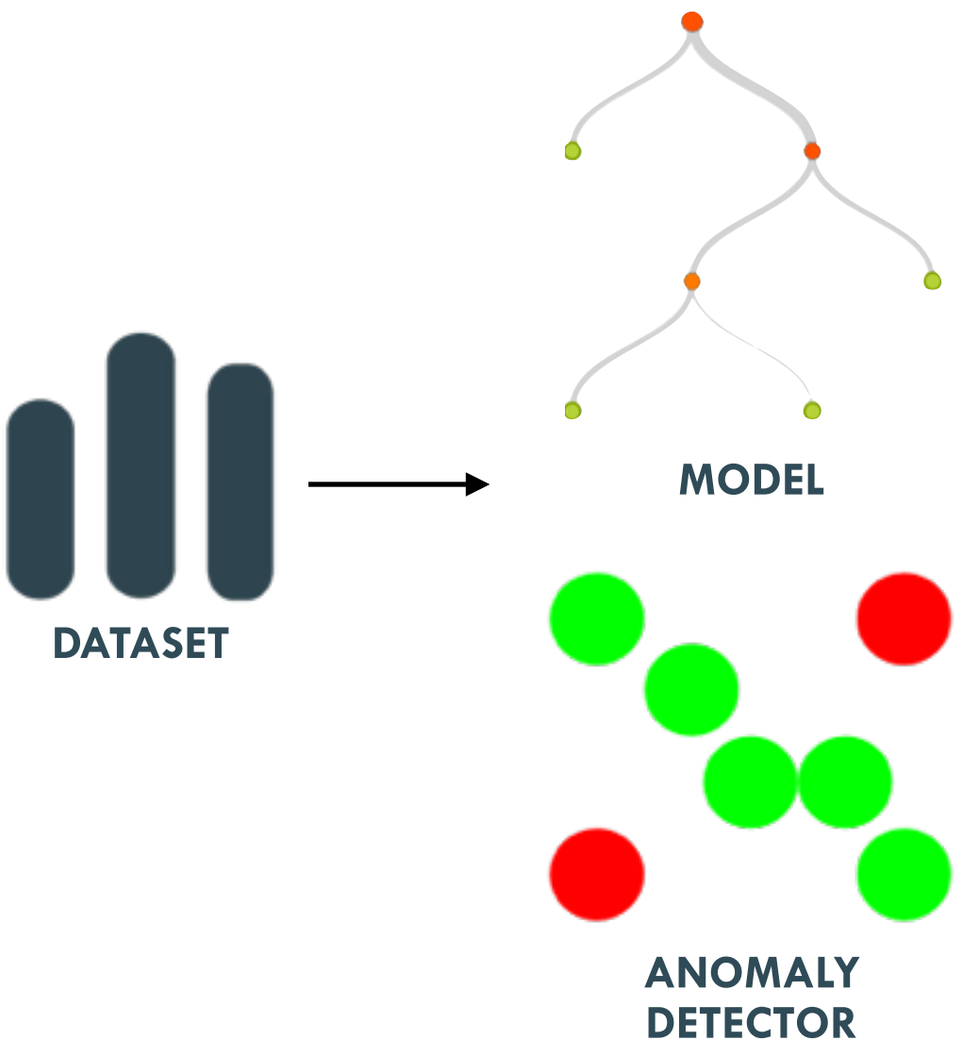


Now repeat the process several times and use average Depth to compute anomaly score: 0 (similar) -> 1 (dissimilar)

Anomaly Detection

- Unusual instance discovery
- Intrusion Detection
- Fraud
- Identify Incorrect Data
- Remove Outliers
- Model Competence / Input Data Drift

Model Competence



Prediction	T	T
Confidence	86%	84%
Anomaly Score	0.5367	0.7124
Competent?	**Y**	**N**

At Training Time

At Prediction Time

Unsupervised Learning

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
The	Sally	6788	sign	food	26339	51

rules

Rules:

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

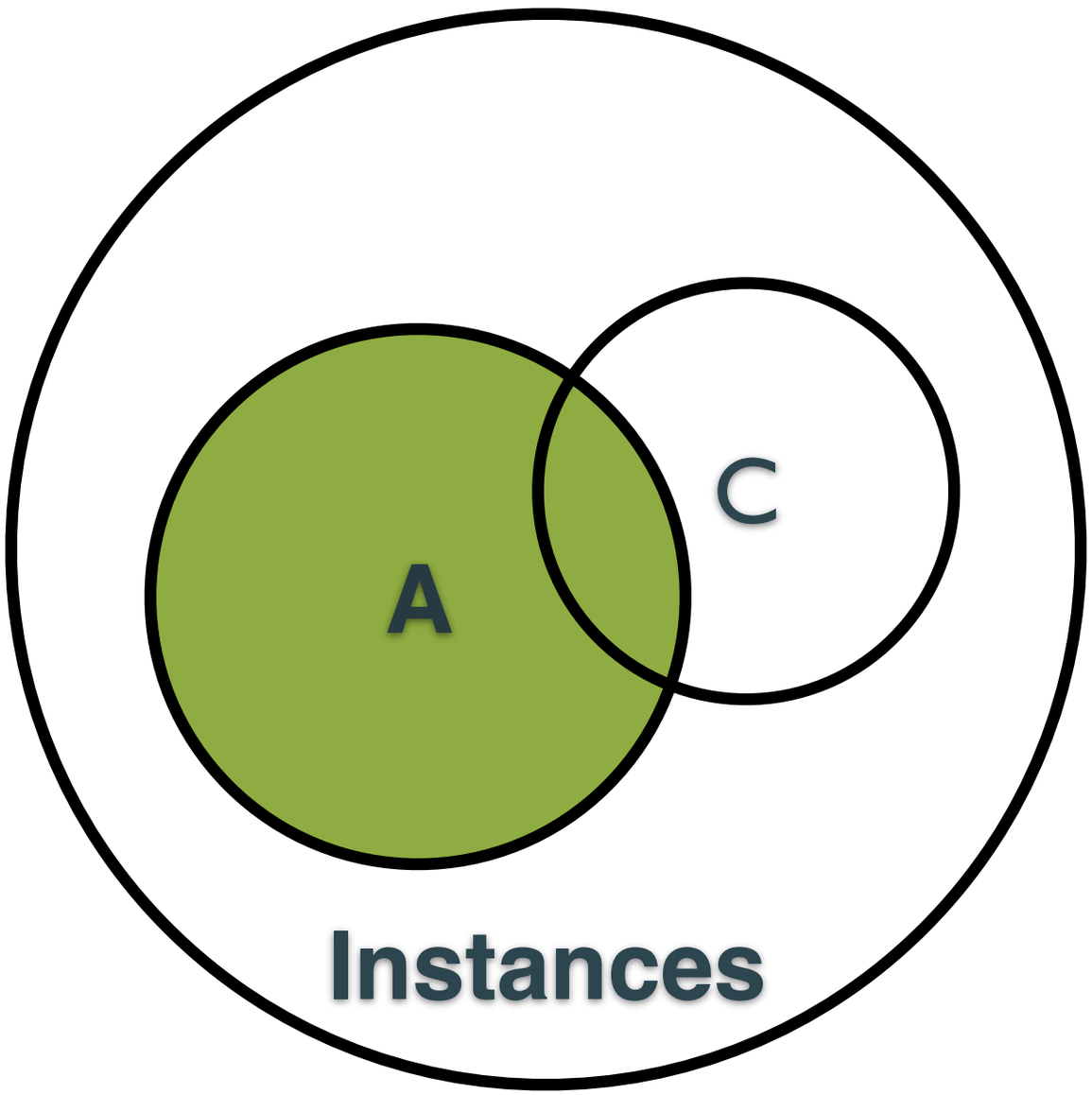
{class = gas} → amount > 80

Antecedent → Consequent

Association Discovery

- Market Basket Analysis
- Web usage patterns
- Intrusion detection
- Fraud detection
- Bioinformatics
- Medical risk factors

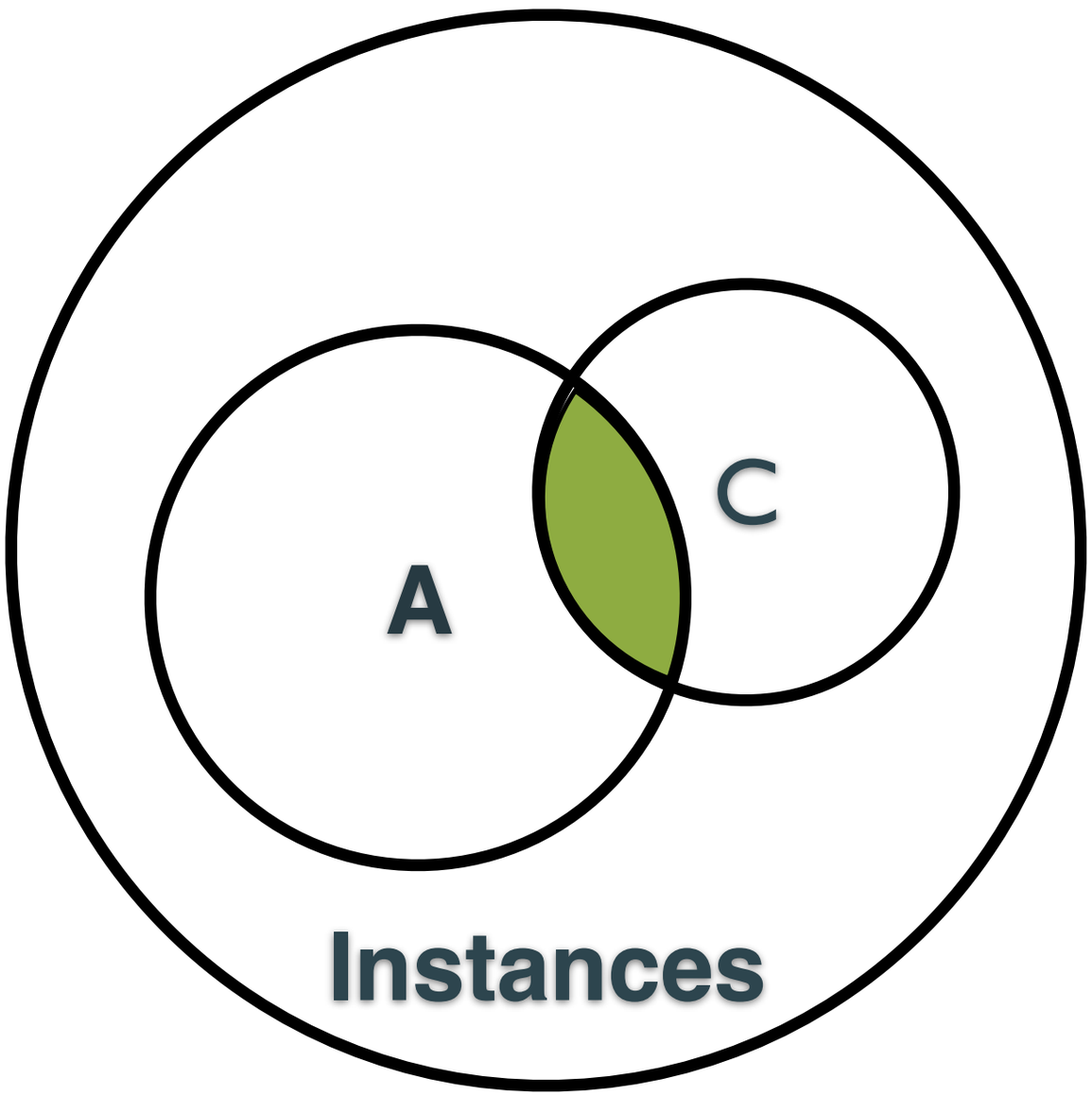
Association Metrics



Coverage

Percentage of instances which match antecedent “A”

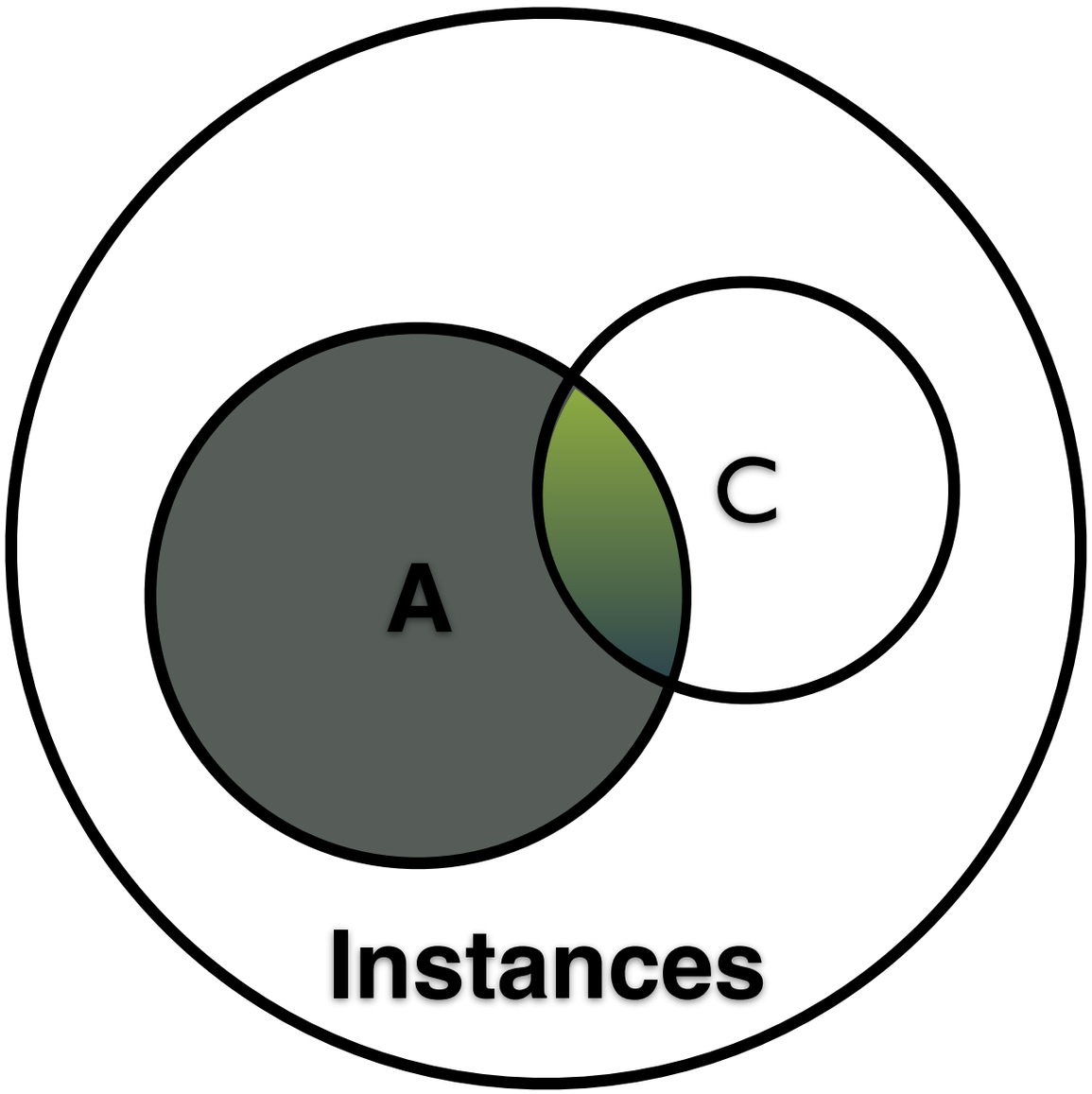
Association Metrics



Support

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

Association Metrics

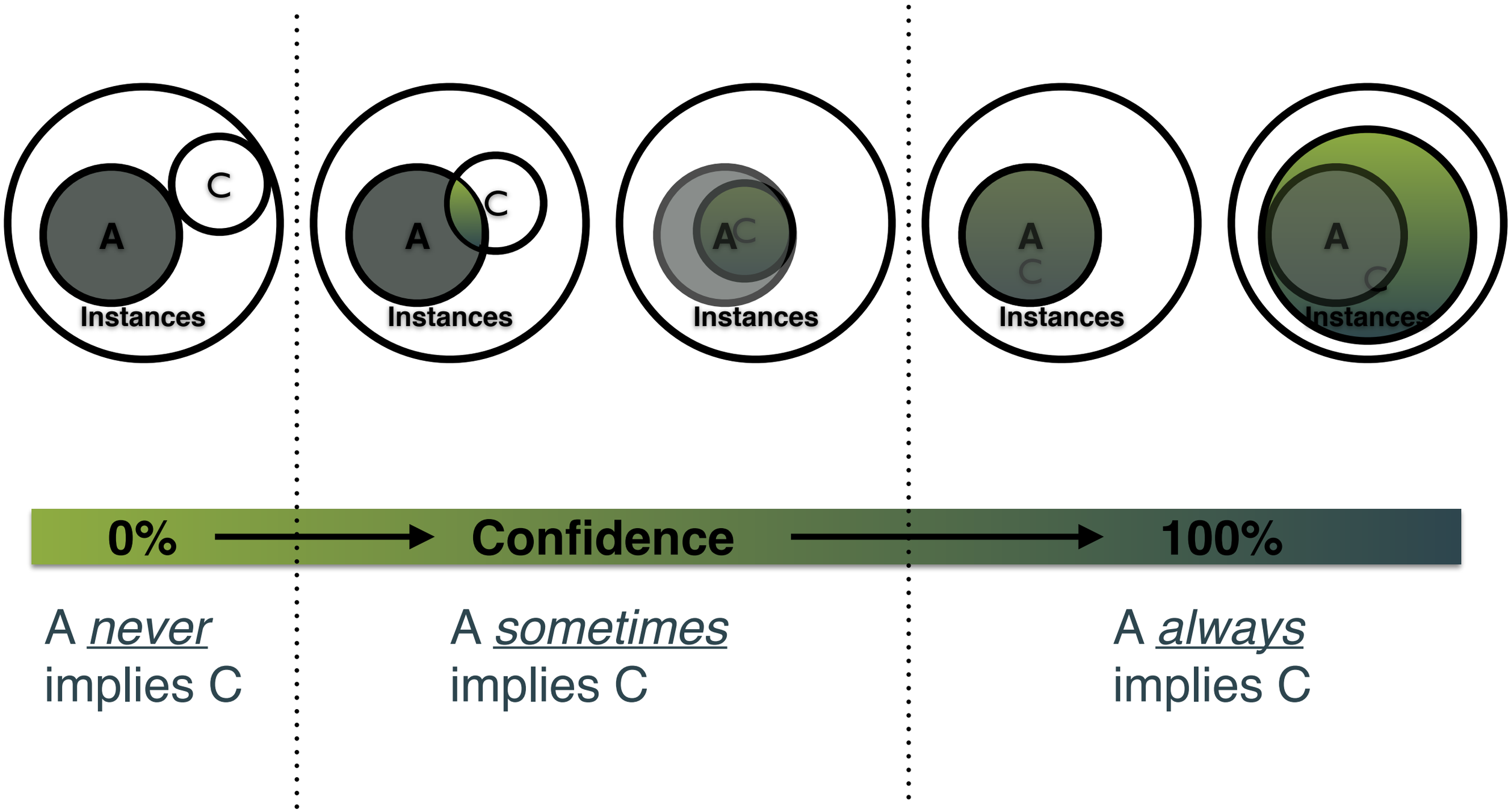


Confidence

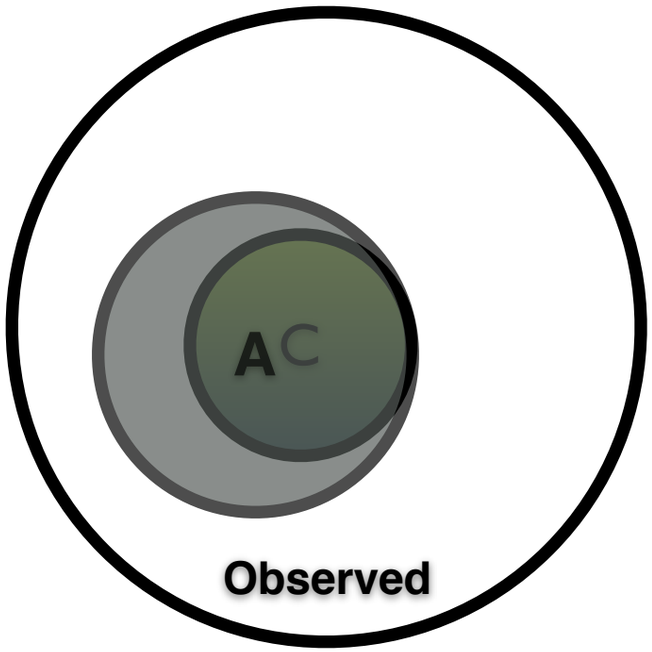
Percentage of instances in the antecedent which also contain the consequent.

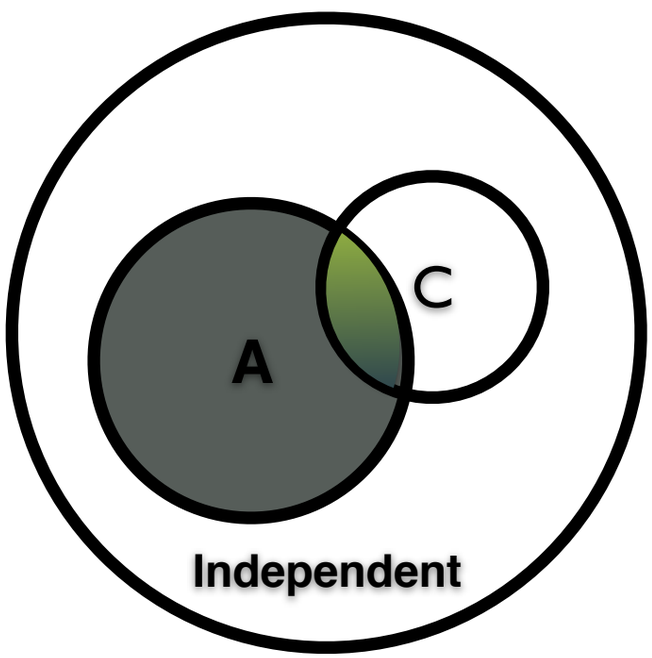
$$\frac{\text{Support}}{\text{Coverage}}$$

Association Metrics



Association Metrics



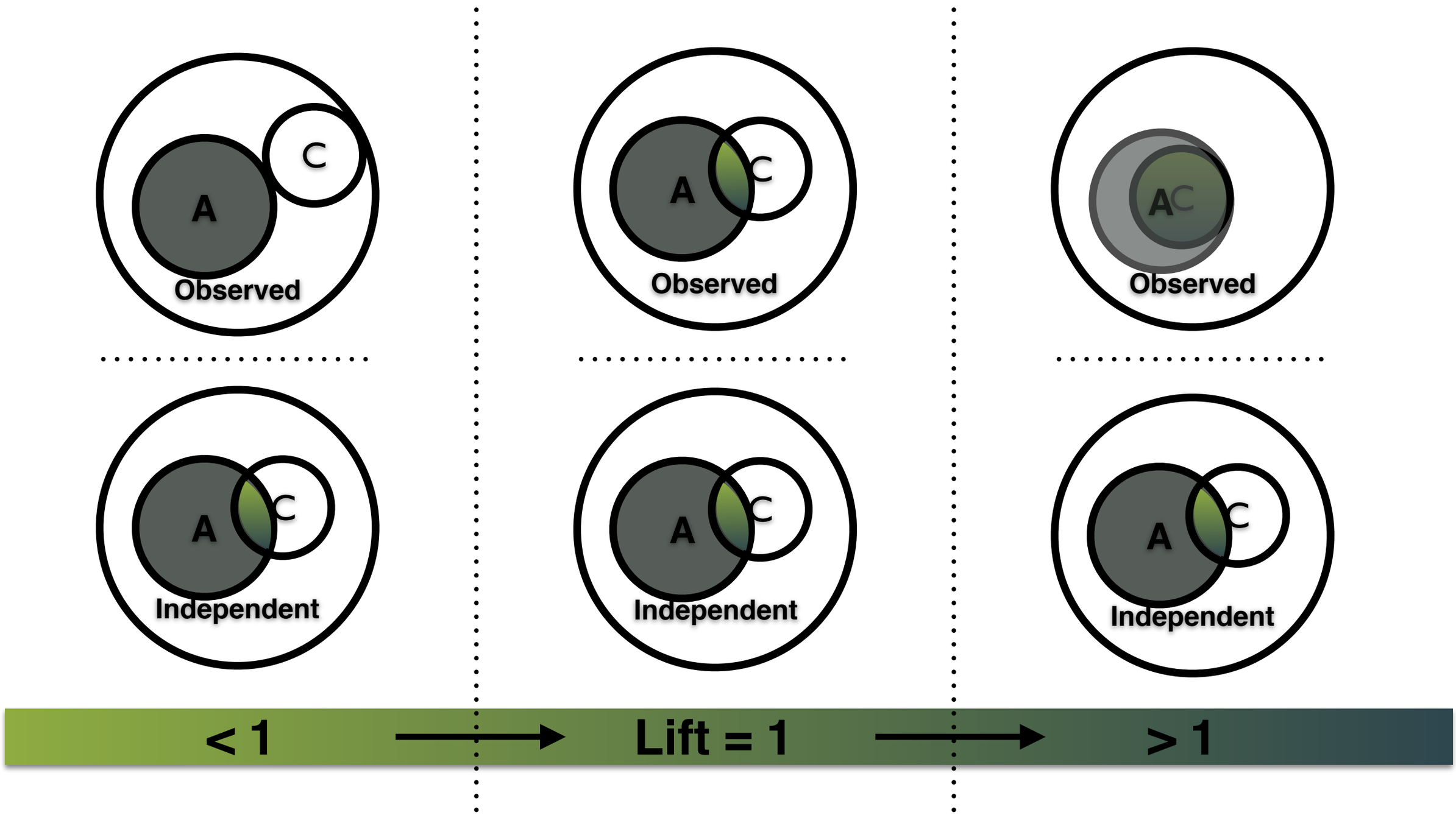


Lift

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

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

Association Metrics

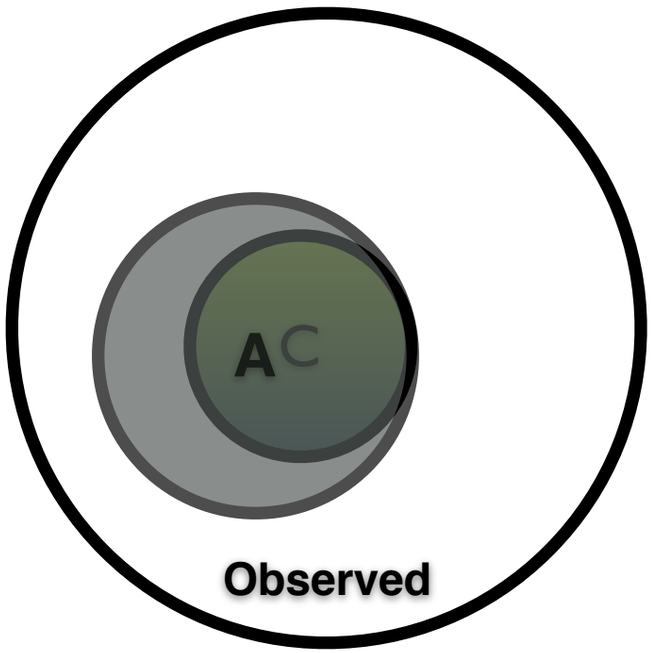


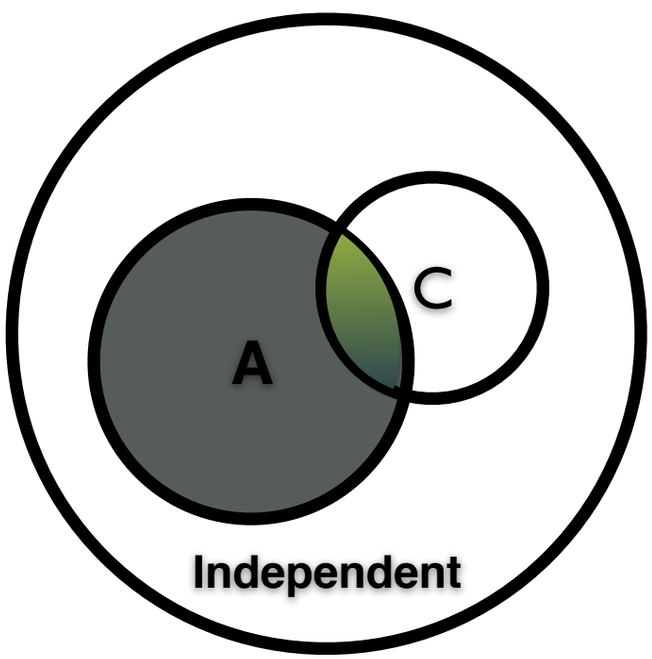
Negative Correlation

No Association

Positive Correlation

Association Metrics



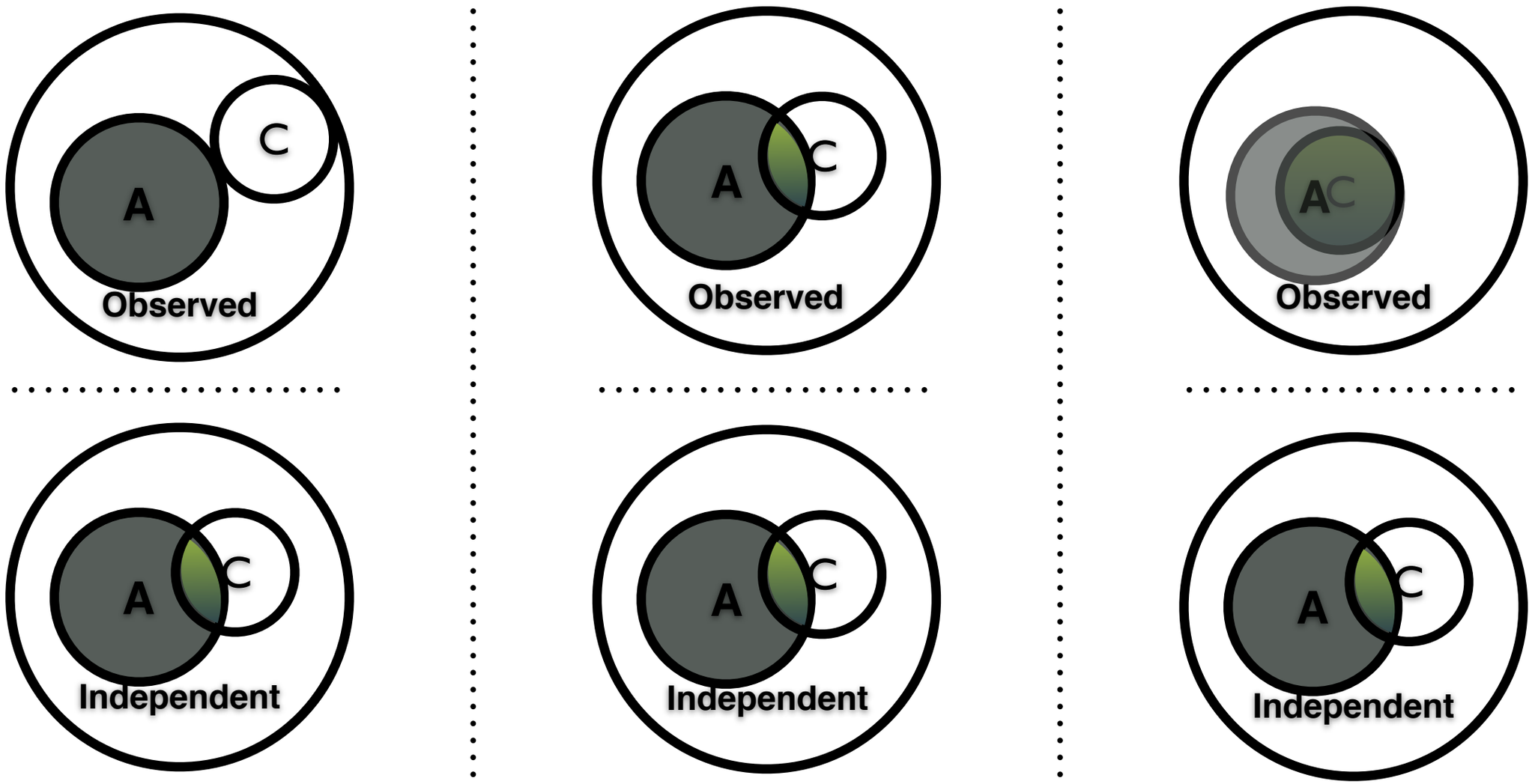


Leverage

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

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

Association Metrics



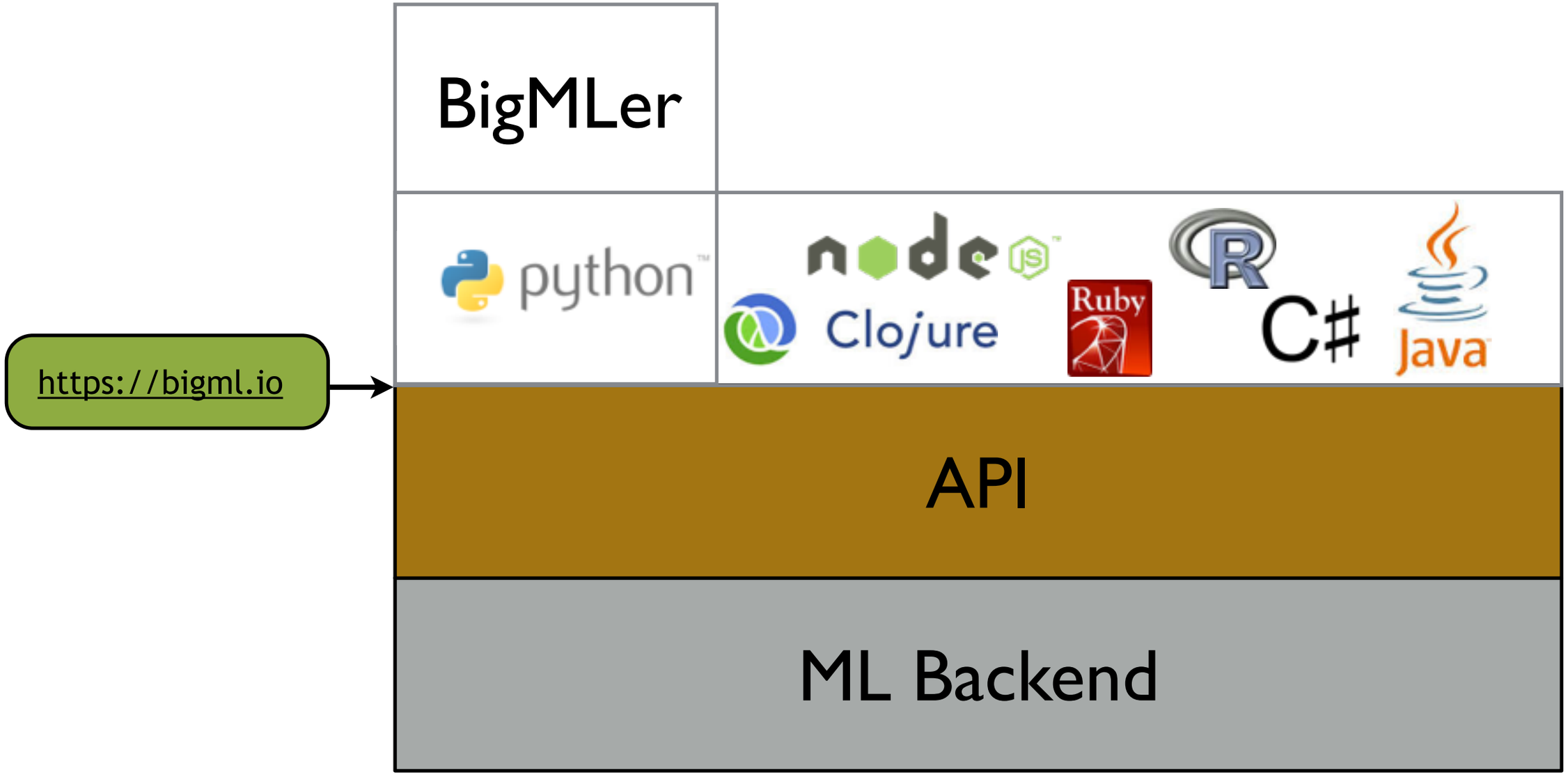
-1… < 0 | **Leverage = 0** | **> 0 …1**

Negative Correlation

No Association

Positive Correlation

API



http://bigmler.readthedocs.org/en/latest/