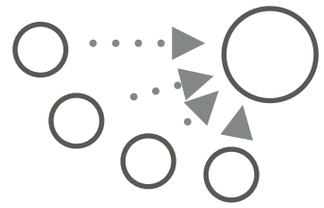


67-300 SEARCH ENGINES

LEARNING TO RANK

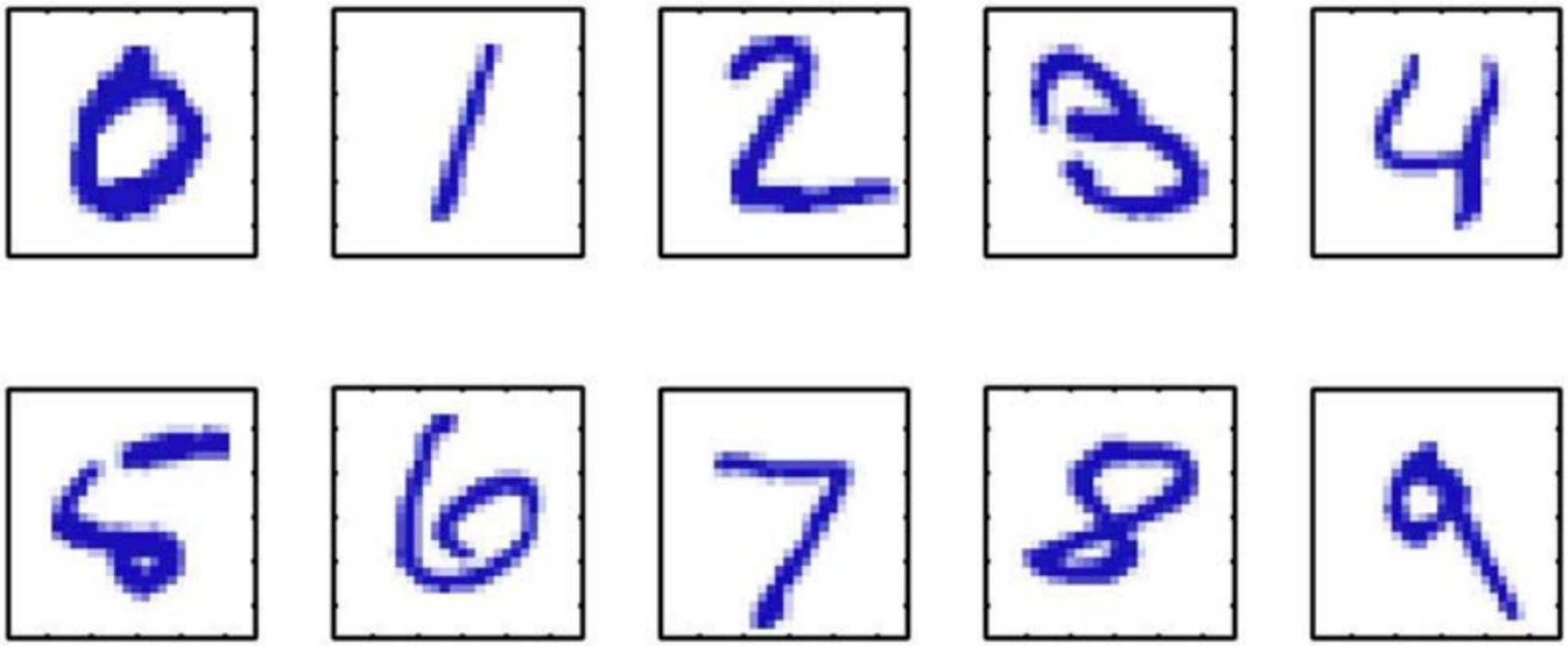
LECTURER: JOAO PALOTTI (JPALOTTI@ANDREW.CMU.EDU)
17TH APRIL 2017

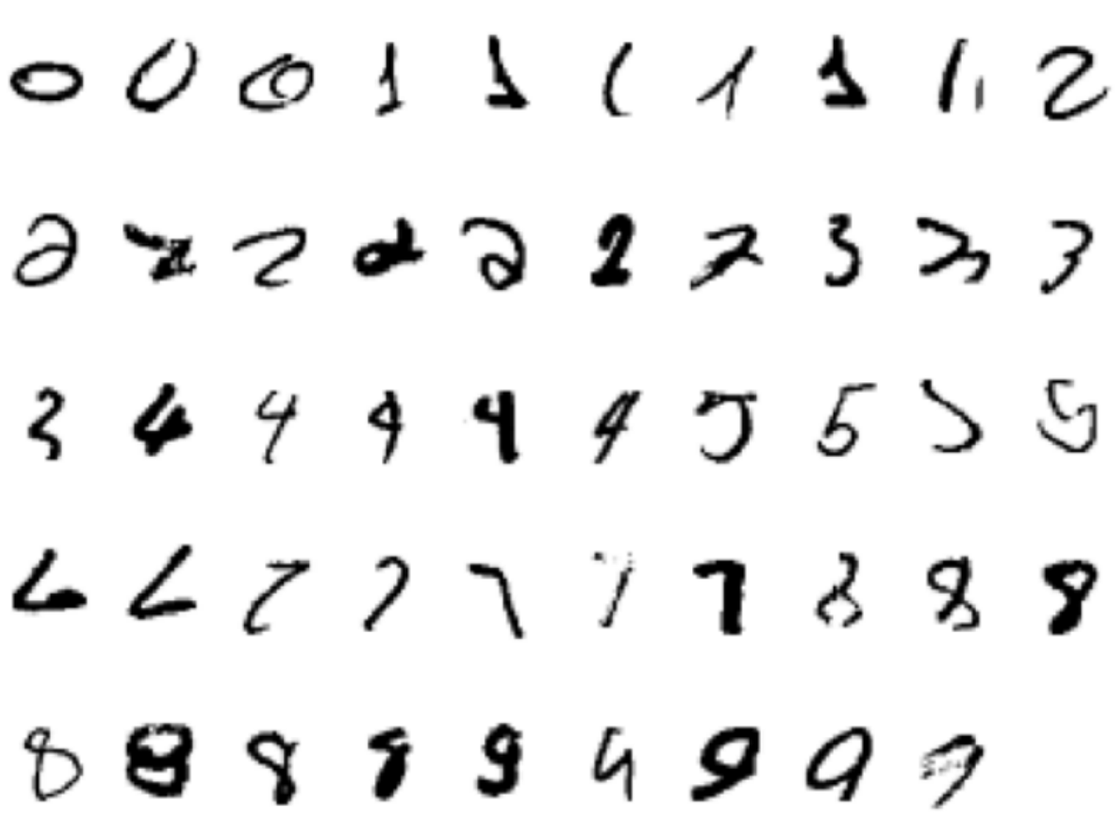
TODAY'S LECTURE IN THE STANFORD IR BOOK

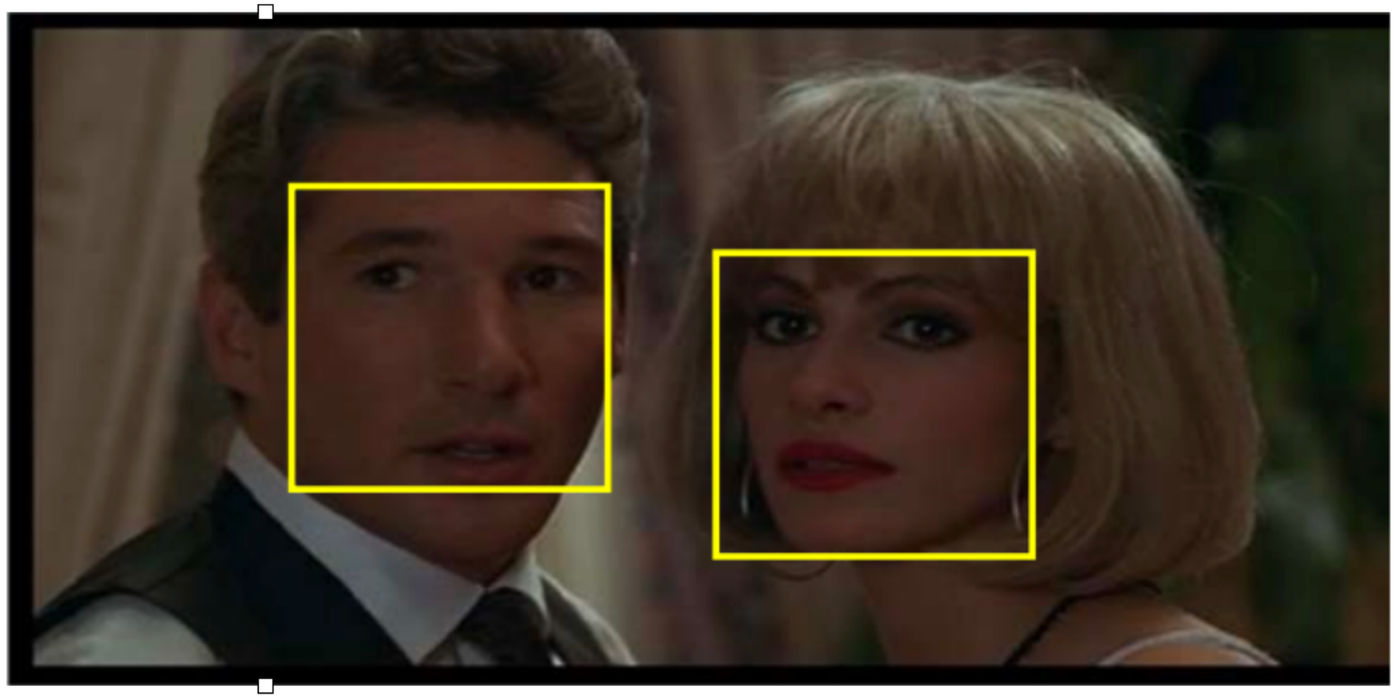
- Machine learning methods:
  - Chapter 13: Text classification for information retrieval
  - Chapter 14: Vector Space classification
  - Chapter 15: Support Vector Machines & machine learning on documents
- Chapter 15.4: Machine Learning methods in ad hoc information retrieval

LECTURE'S GOAL

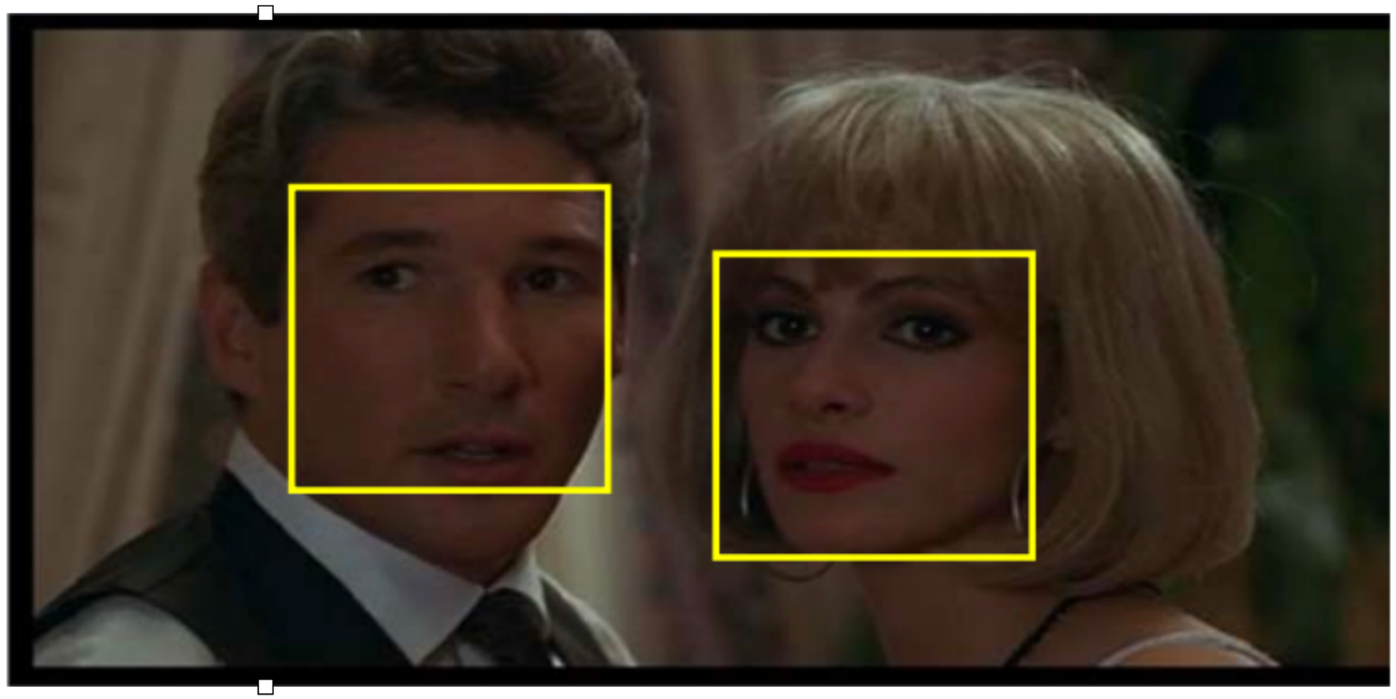
- Machine Learning Overview
- Machine Learning with Text
- Machine Learning for ad-hoc retrieval (learning to rank)

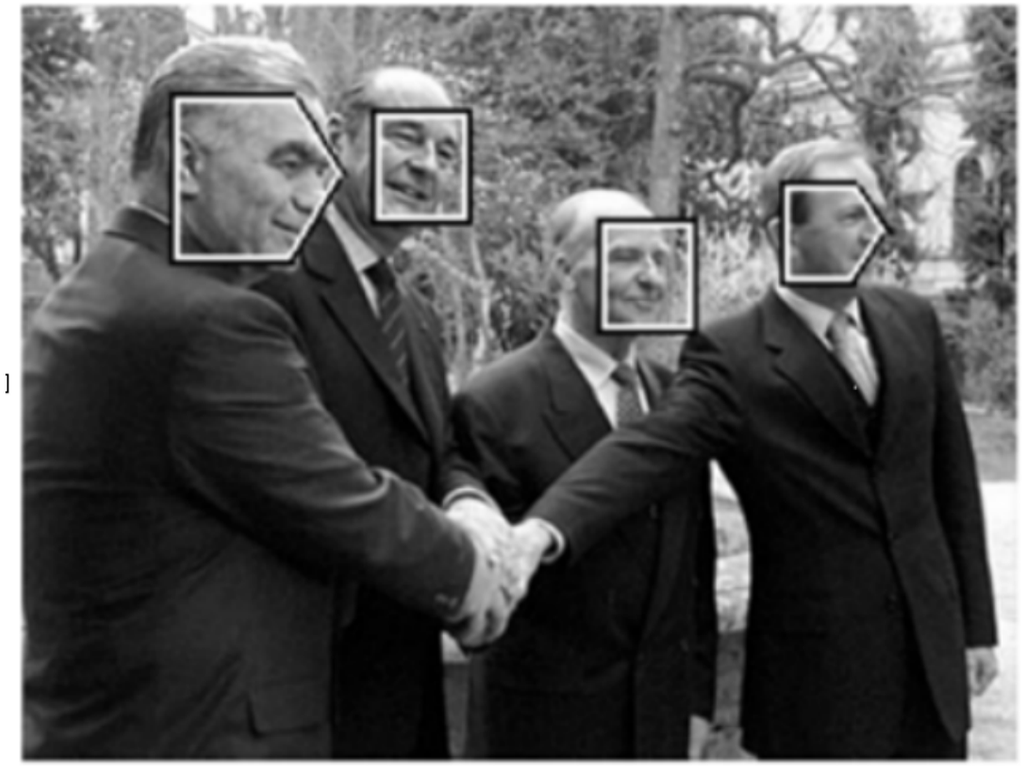




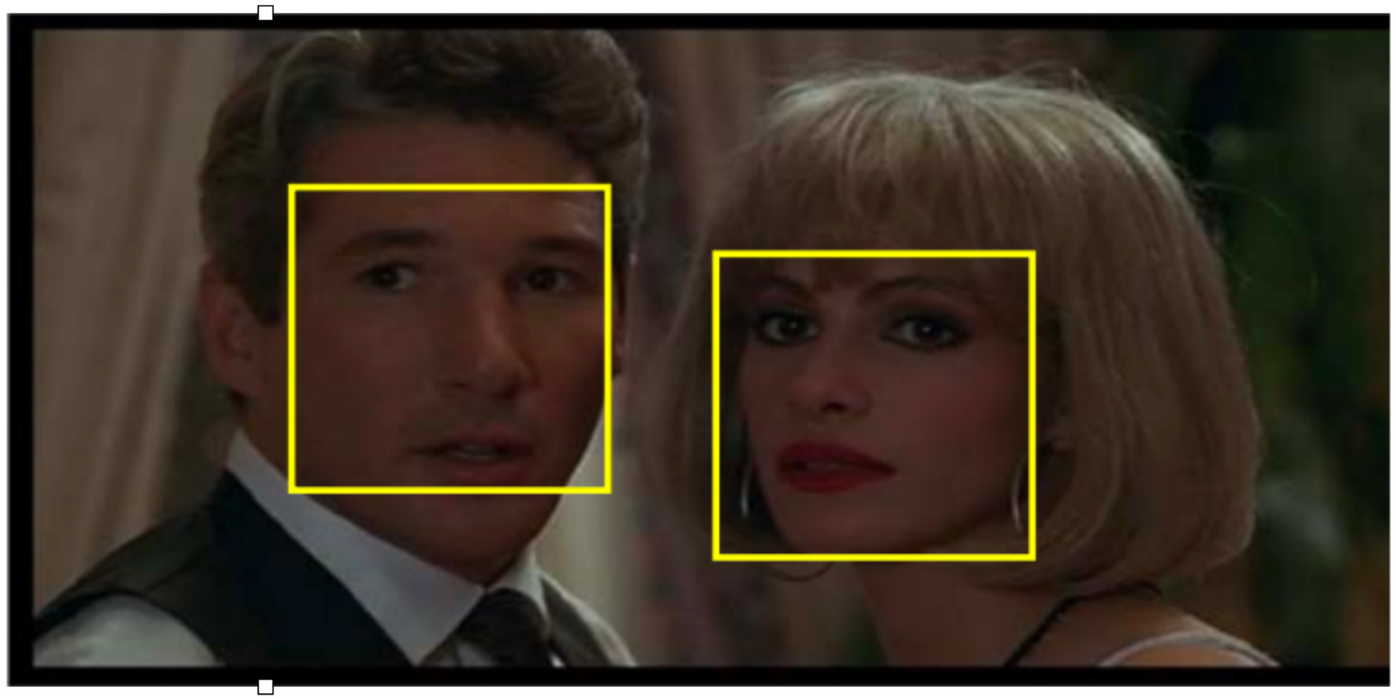


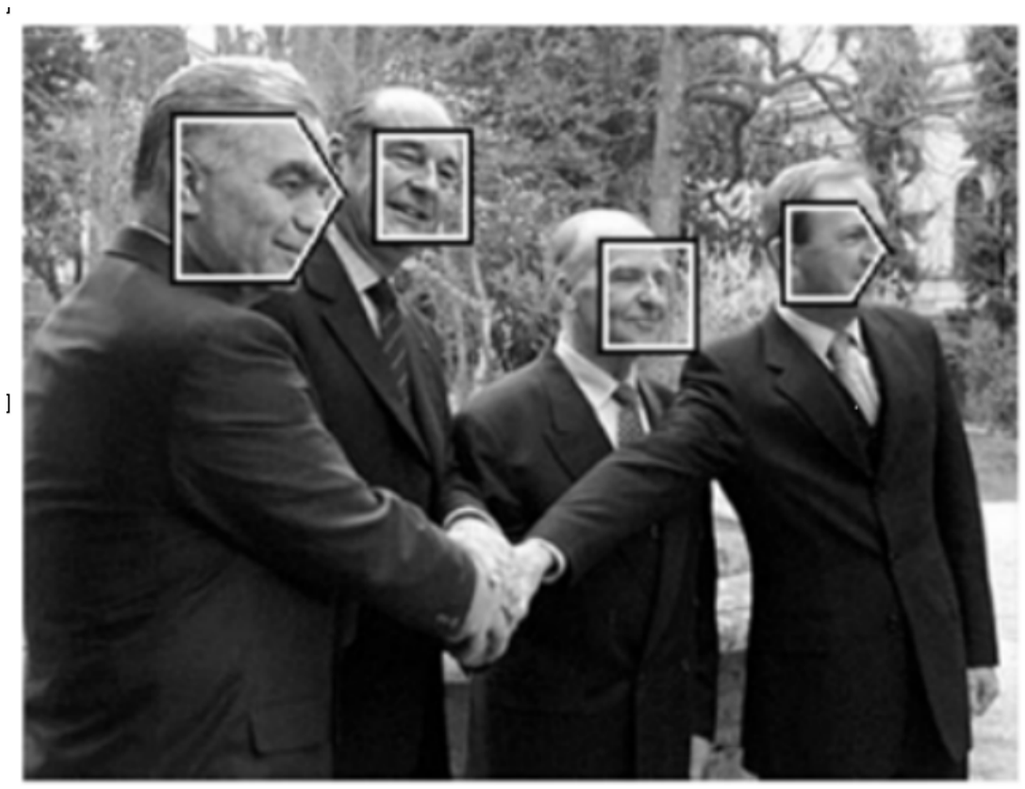
LECTURE 11 - LEARNING TO RANK





LECTURE 11 - LEARNING TO RANK





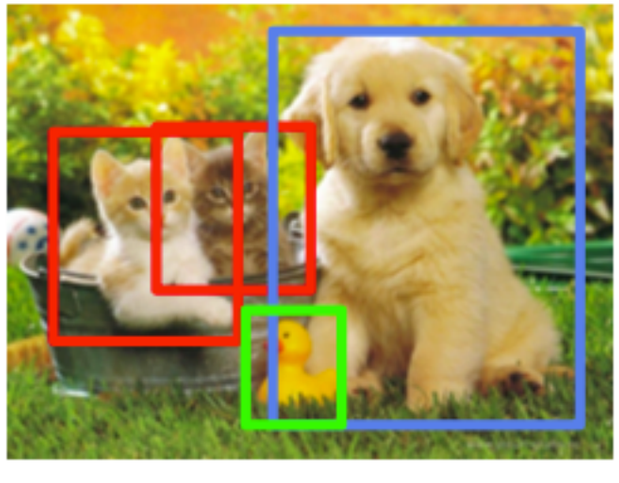
Classification

CAT

Classification + Localization

CAT

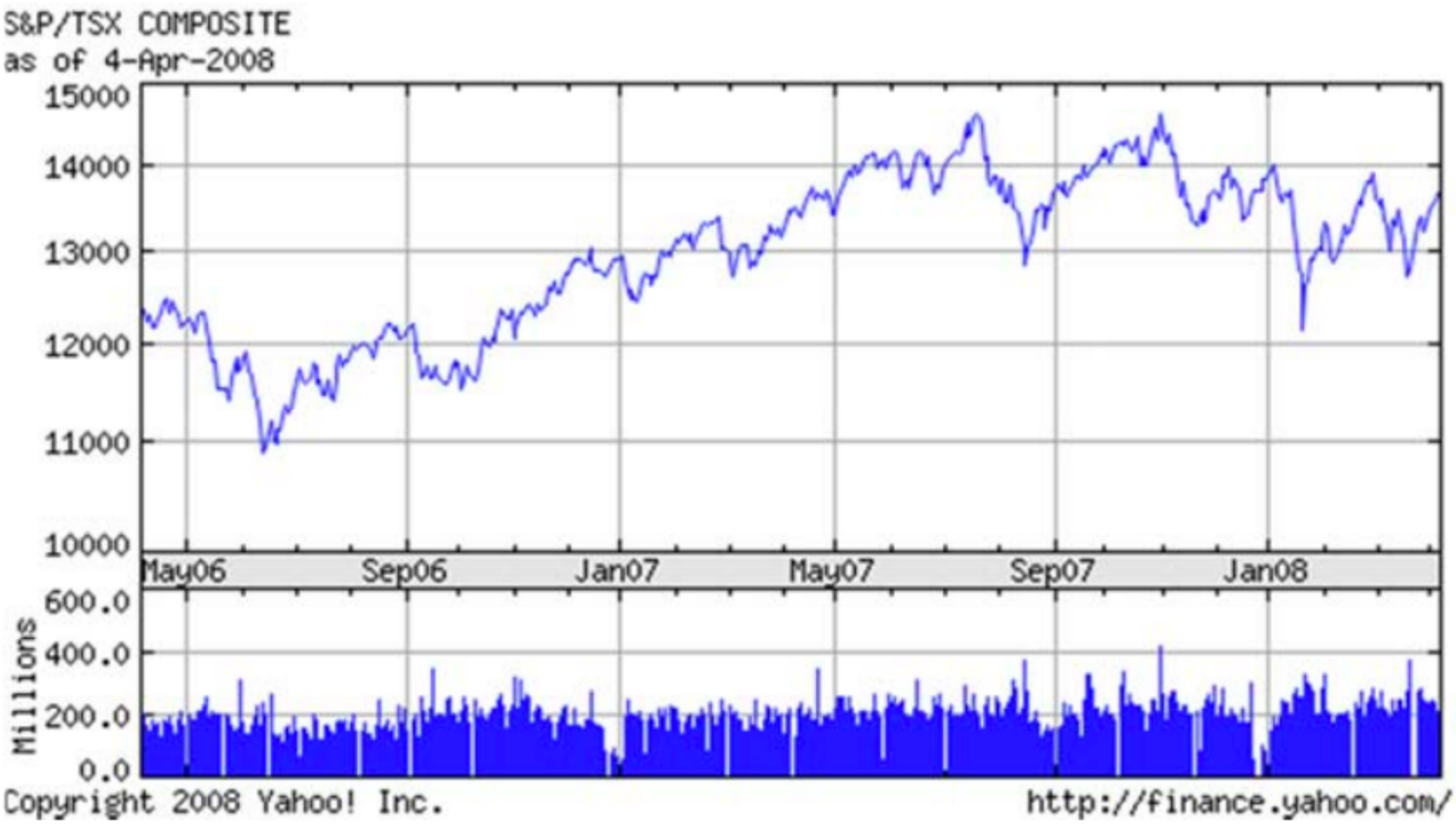
Object Detection



CAT, DOG, DUCK

Instance Segmentation

CAT, DOG, DUCK


S&P/TSX COMPOSITE
as of 4-Apr-2008
15000
14000
13000
12000
11000
10000
May06
Sep06
Jan07
May07
Sep07
Jan08
600.0
400.0
200.0
0.0
Millions
Copyright 2008 Yahoo! Inc.
http://finance.yahoo.com/

OPEC GIVE AWAY

Spam x

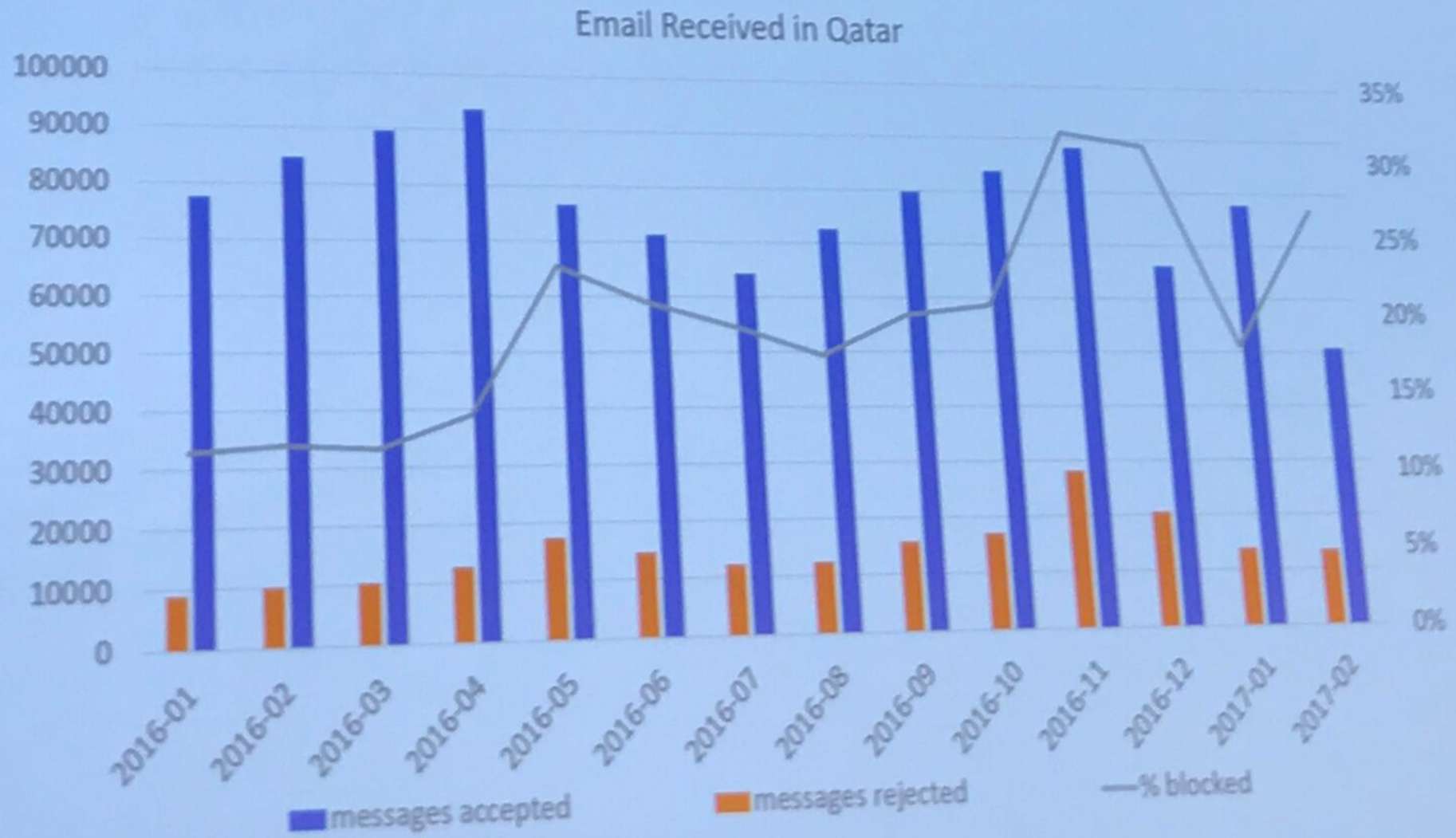
admin@rec.com 3:24 PM (3 hours ago)

to Recipients

OPEC Foreign Processing Department
> OPEC Fund for International Development (OFID)
> Martin Street, Birstall, Batley
> West Yorkshire, W17 9PJ - UK
>
>
> Attn: PRIVATE
>
> We wish to to notify you of the OFID first quater balloting final result. Your email ID emerge in our 2rd category as a winner for a cash prize of $100,000.00 (one hundred thousand US$). This is from 21 winners from email list of 10,000,000 individuals, coperate and private organisations, NGO's and public sectors selected globally in this caterory.
>
> The OPEC Fund for International Development (OFID) is a foundation owned by the Organization of Petroleum Exporting Countries (OPEC). This foundation is funded by member nations which include: Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, United Arab Emirates and Venezuela.
>
> OFID is a development organization aimed at improving lives across the world. This program tagged "Grass root Program" is part of efforts to improve international housing problems, support the research for the eradication of Ebola Virus and improve standard of living through direct participation in community development across several communities all over the world by empowering selected individuals as an engine for economic growth and social development.

SPAM?
NOT SPAM?

How much phishing do we see?









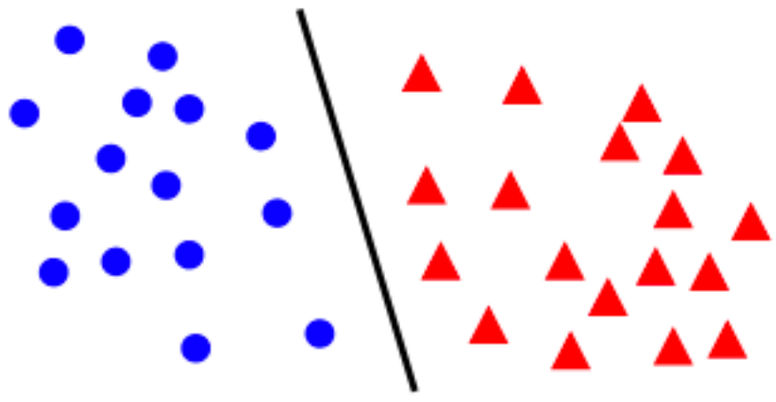




LECTURE 11 - LEARNING TO RANK

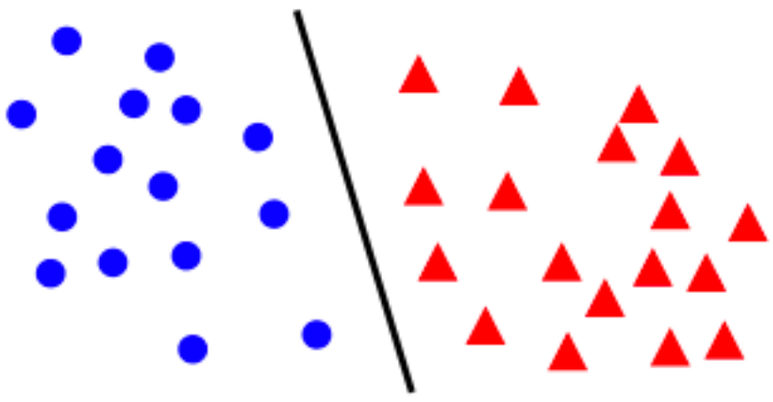
THREE MAIN MACHINE LEARNING TASKS

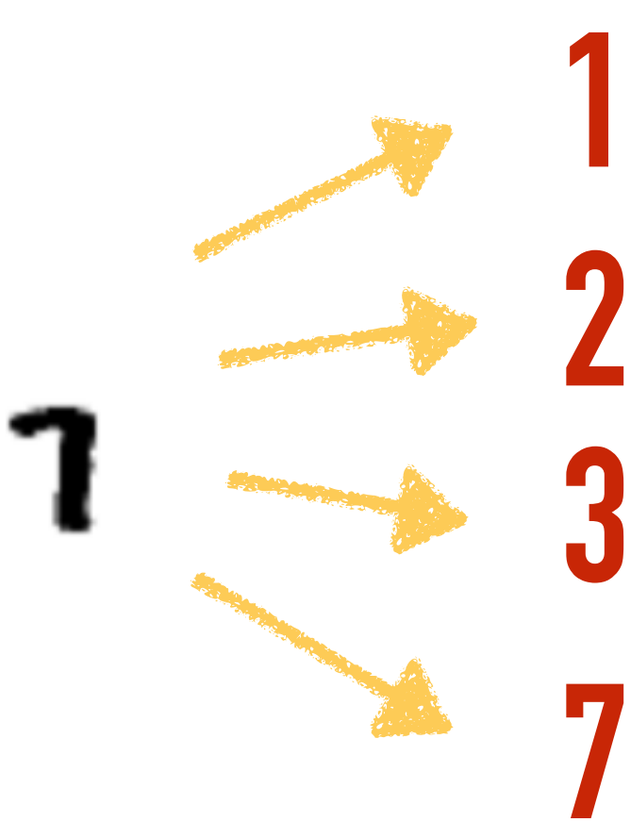
1. Classification:
   - Spam / Not Spam
   - Boy / Guitar / Beach
   - Cat / Dog



THREE MAIN TASKS IN MACHINE LEARNING

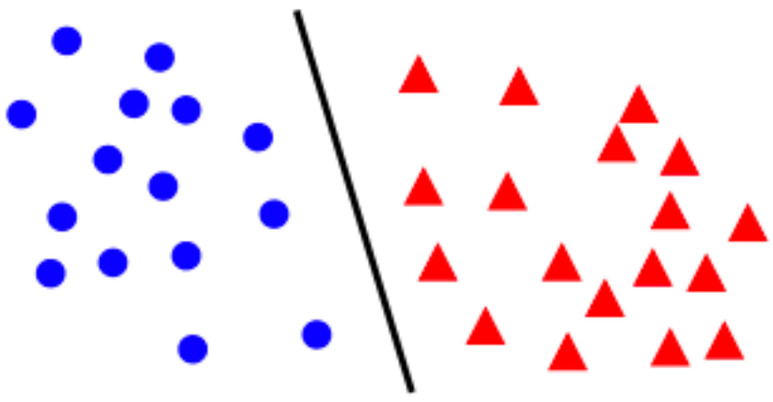
1. Classification:
   - Spam / Not Spam
   - Boy / Guitar / Beach
   - Cat / Dog

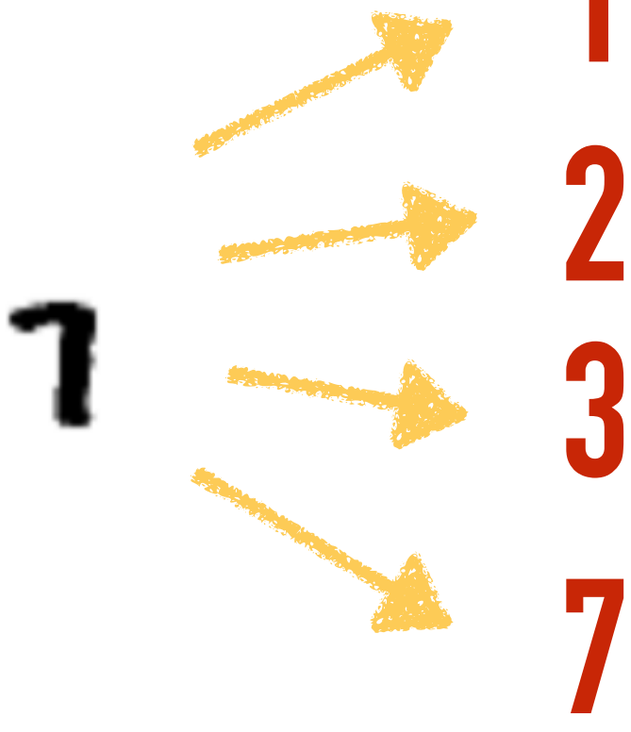




THREE MAIN TASKS IN MACHINE LEARNING

1. Classification: → Classes!!! → Supervised!!!
   - Spam / Not Spam
   - Boy / Guitar / Beach
   - Cat / Dog



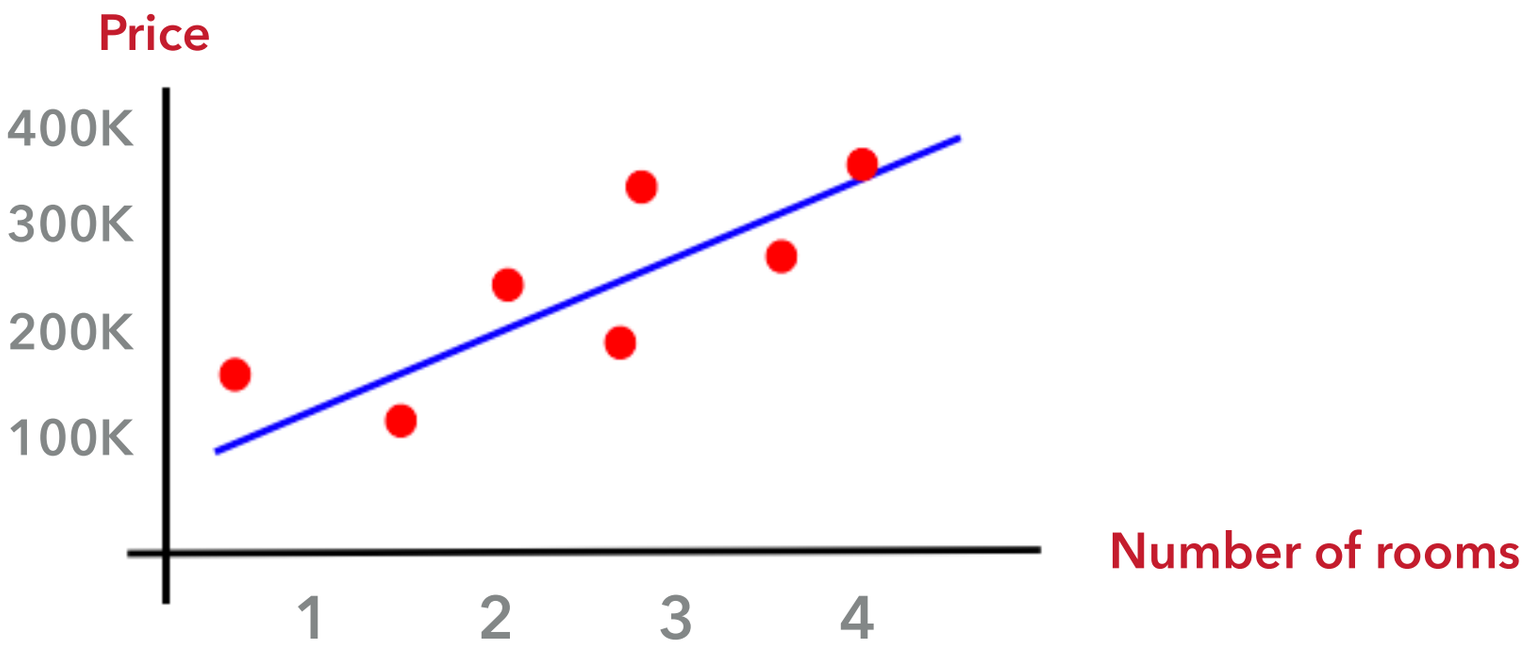


THREE MAIN TASKS IN MACHINE LEARNING

2. Regression
   - What is the price of a stock at the end of the day
   - How many iced teas can I sell today?
   - What is the value of this house?

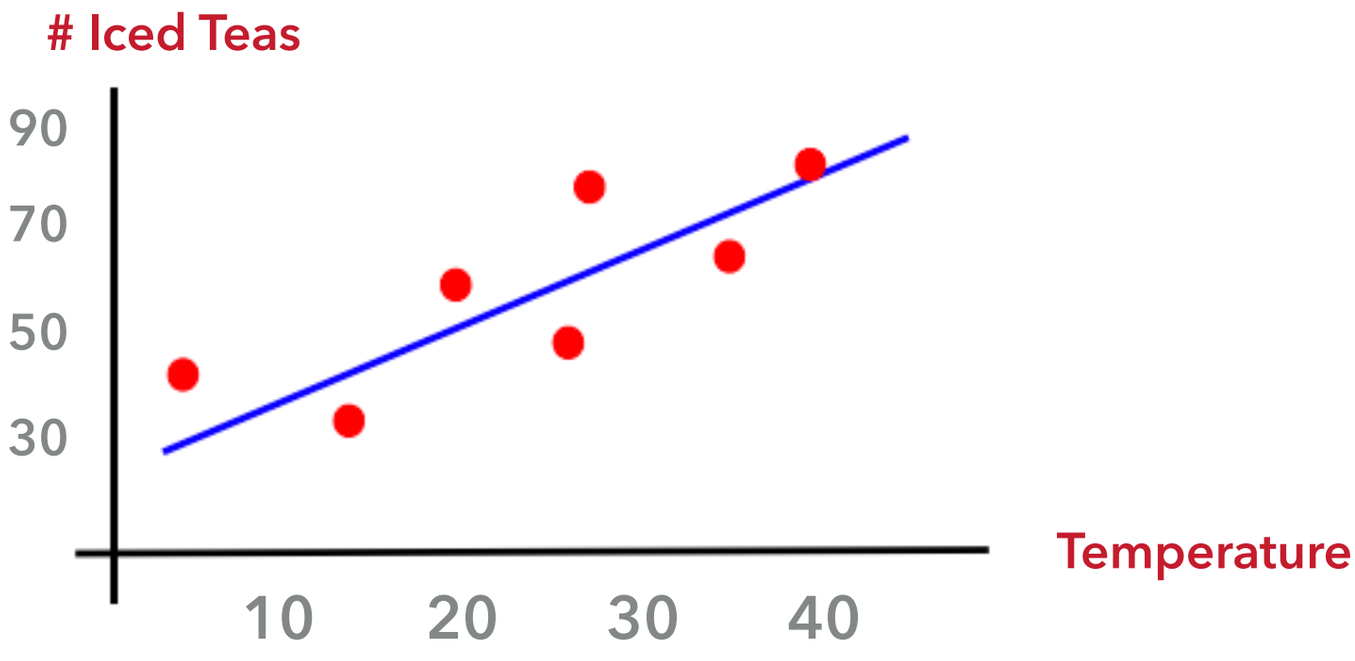
THREE MAIN TASKS IN MACHINE LEARNING

2. Regression
   - What is the price of a stock at the end of the day
   - How many iced teas can I sell today?
   - What is the value of this house?



THREE MAIN TASKS IN MACHINE LEARNING

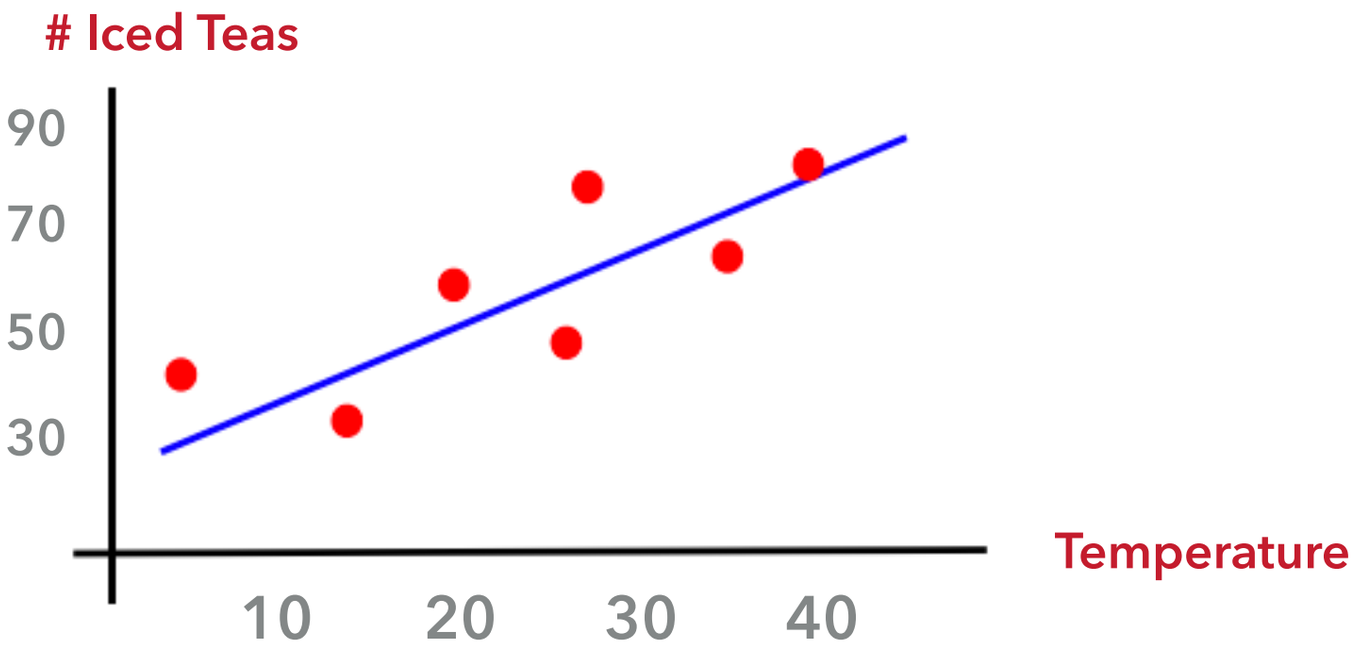
2. Regression
   - What is the price of a stock at the end of the day
   - How many iced teas can I sell today?
   - What is the value of this house?



THREE MAIN TASKS IN MACHINE LEARNING

2. Regression → Real Number!!! → Supervised!!!

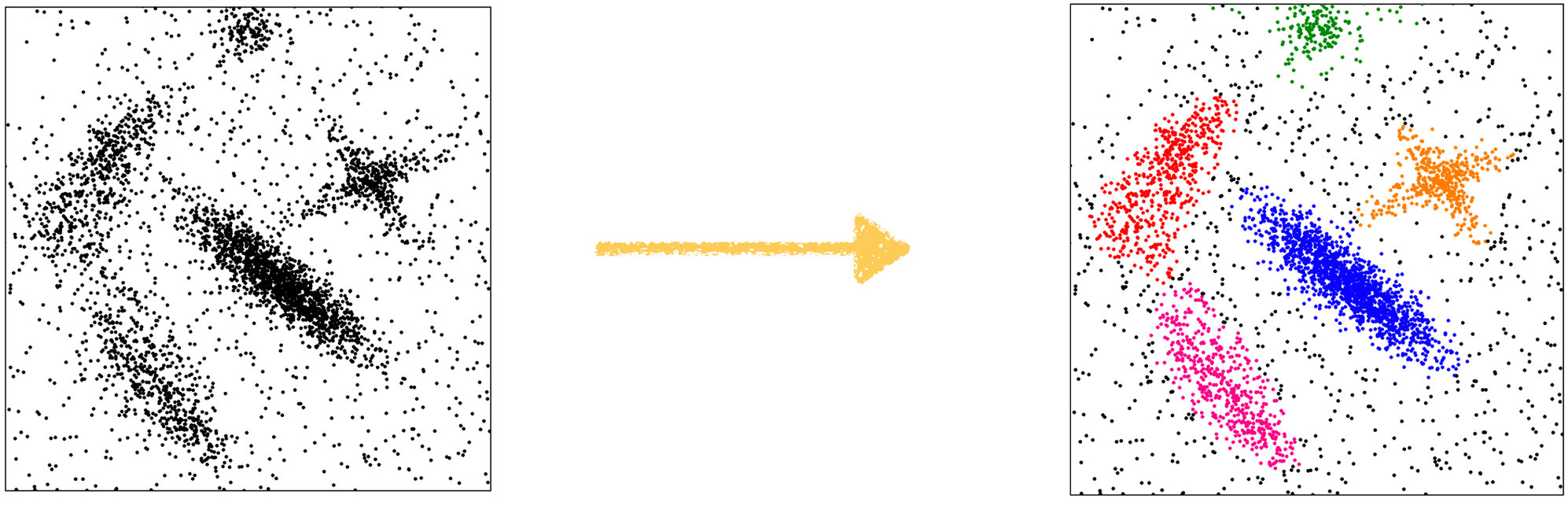
- What is the price of a stock at the end of the day
- How many iced teas can I sell today?
- What is the value of this house?



THREE MAIN TASKS IN MACHINE LEARNING

3. Clustering

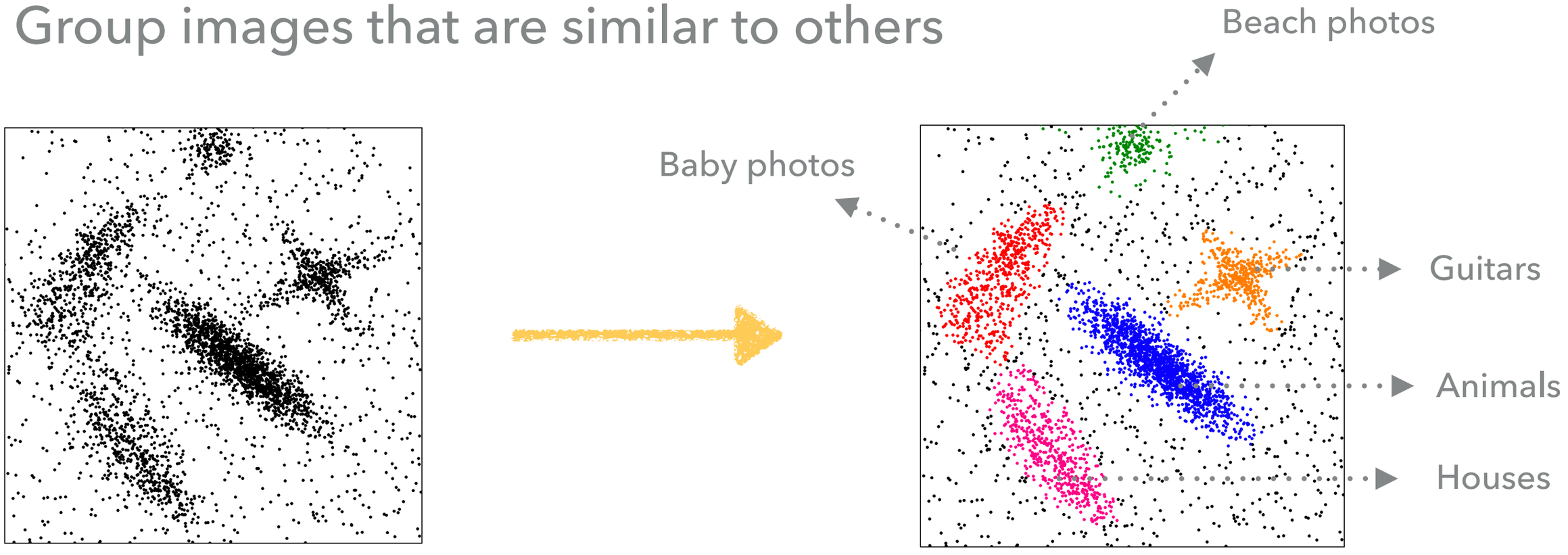
- What are the topics/subjects discussed in this dataset?
- What are the types of people that buy my product?
- Group images that are similar to others



THREE MAIN TASKS IN MACHINE LEARNING

3. Clustering

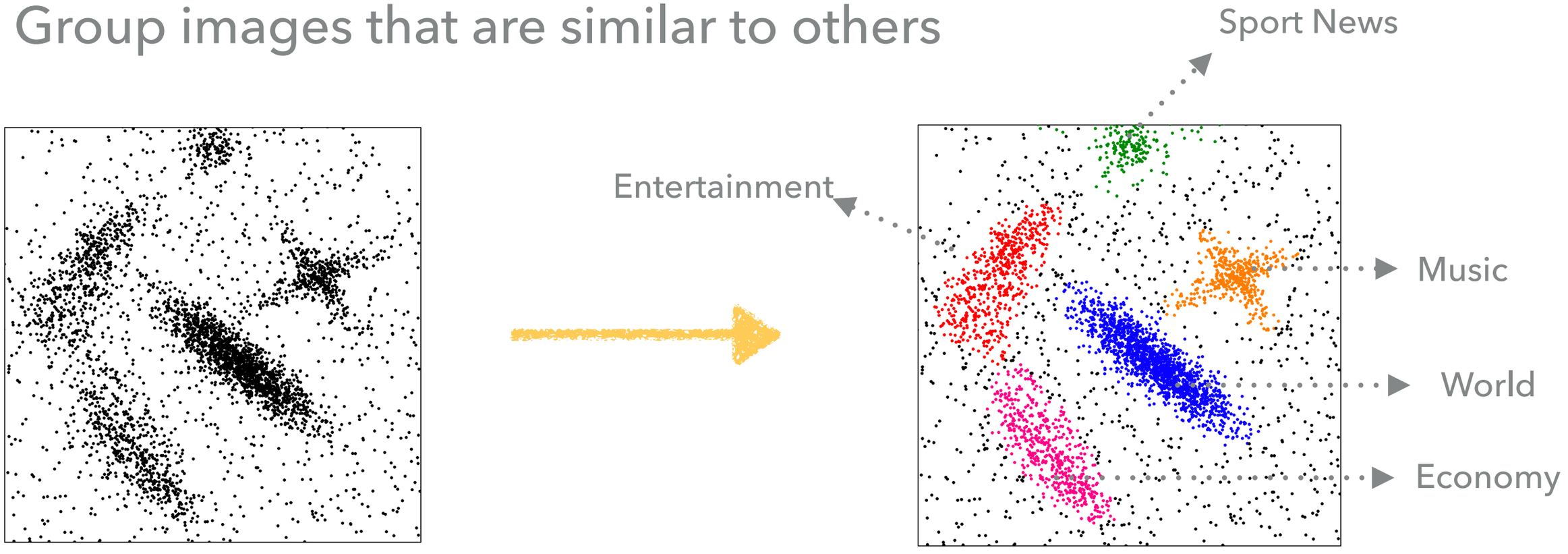
- What are the topics/subjects discussed in this dataset?
- What are the types of people that buy my product?
- Group images that are similar to others



THREE MAIN TASKS IN MACHINE LEARNING

3. Clustering

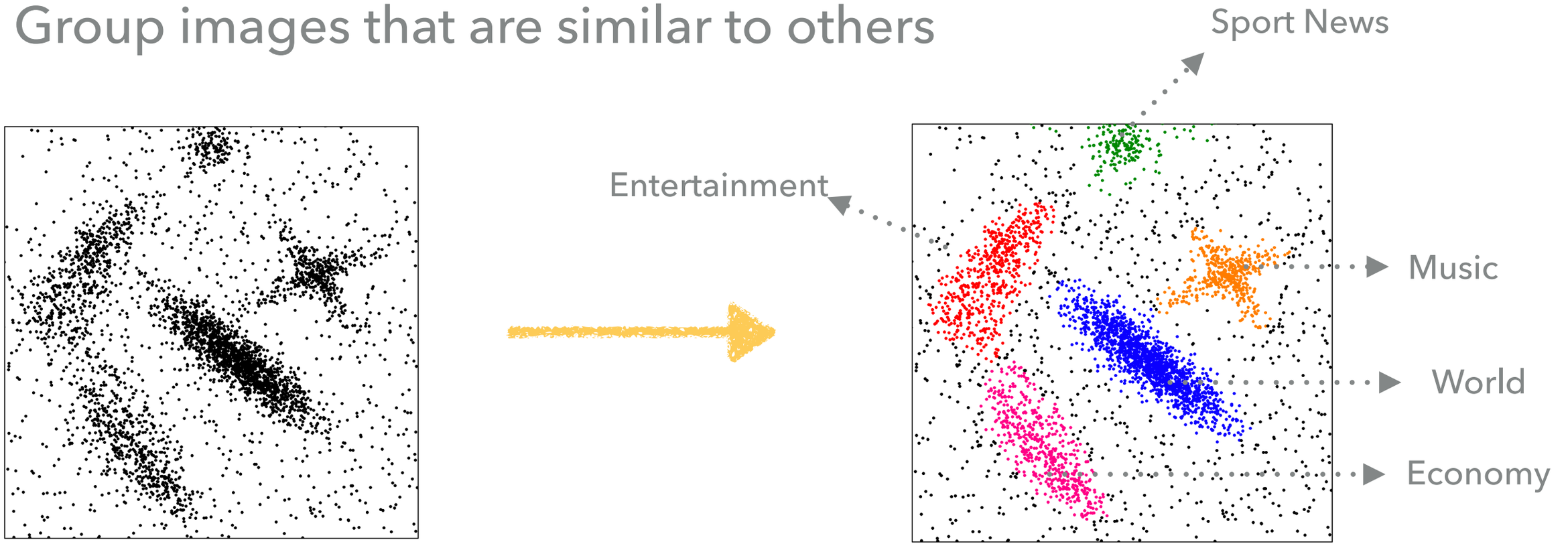
- What are the topics/subjects discussed in this dataset?
- What are the types of people that buy my product?
- Group images that are similar to others



THREE MAIN TASKS IN MACHINE LEARNING

3. Clustering → Groups!!!
   → Unsupervised!!!

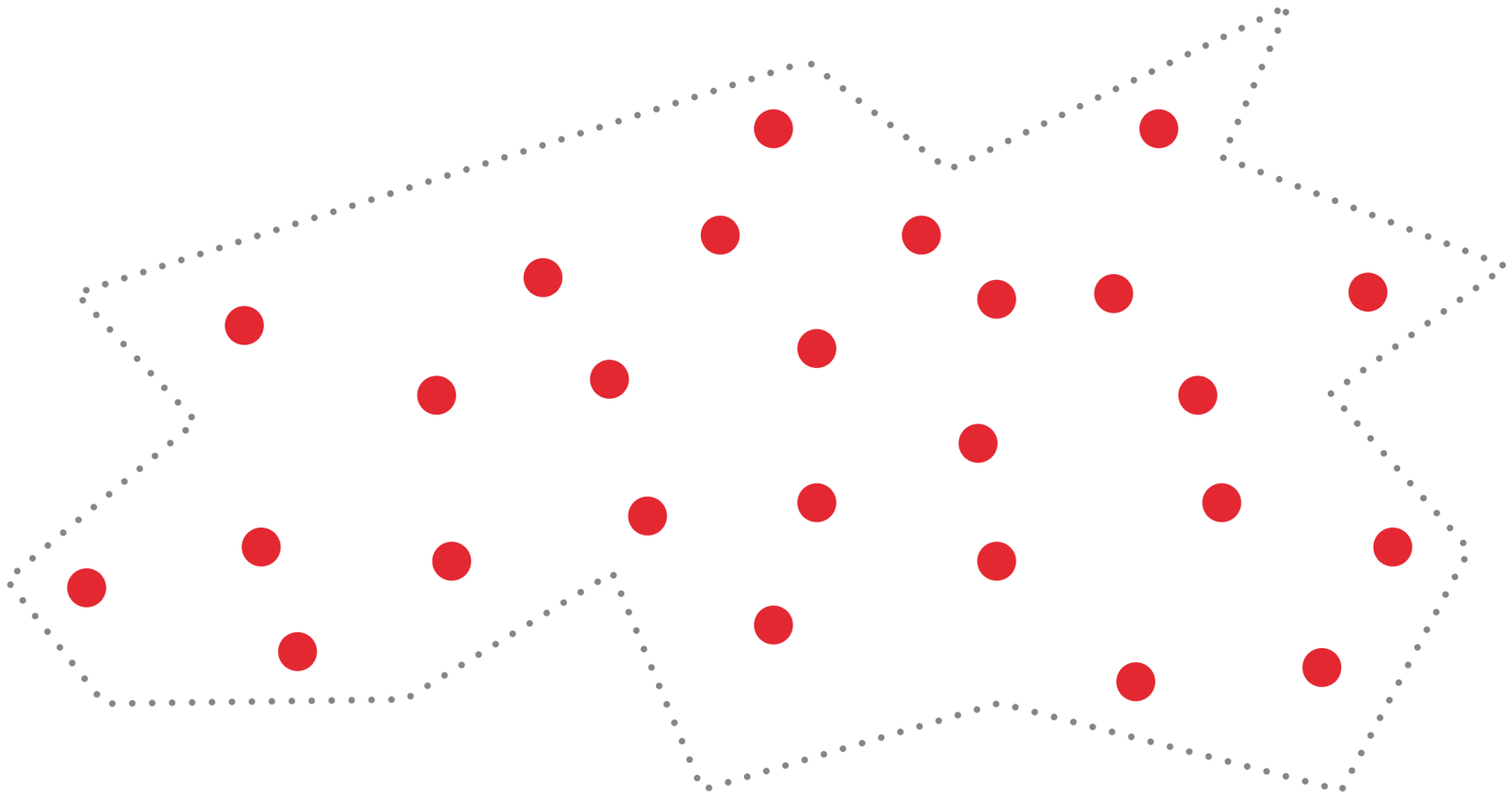
- What are the topics/subjects discussed in this dataset?
- What are the types of people that buy my product?
- Group images that are similar to others



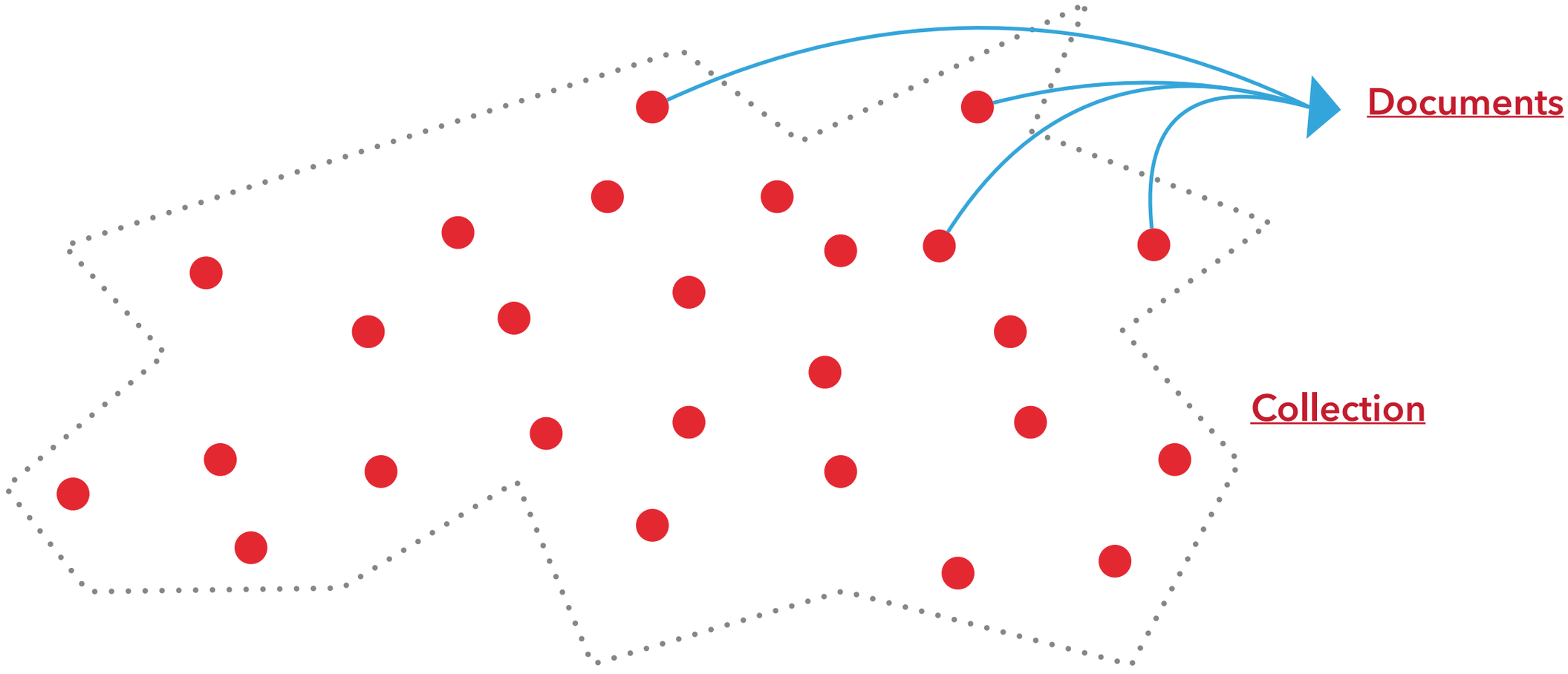
THREE MAIN TASKS IN MACHINE LEARNING (ROUGHLY)

1. Classification:
   - Find a function to tell two/more classes apart
2. Regression:
   - Find a function to estimate a value
3. Clustering:
   - Find a function to group objects

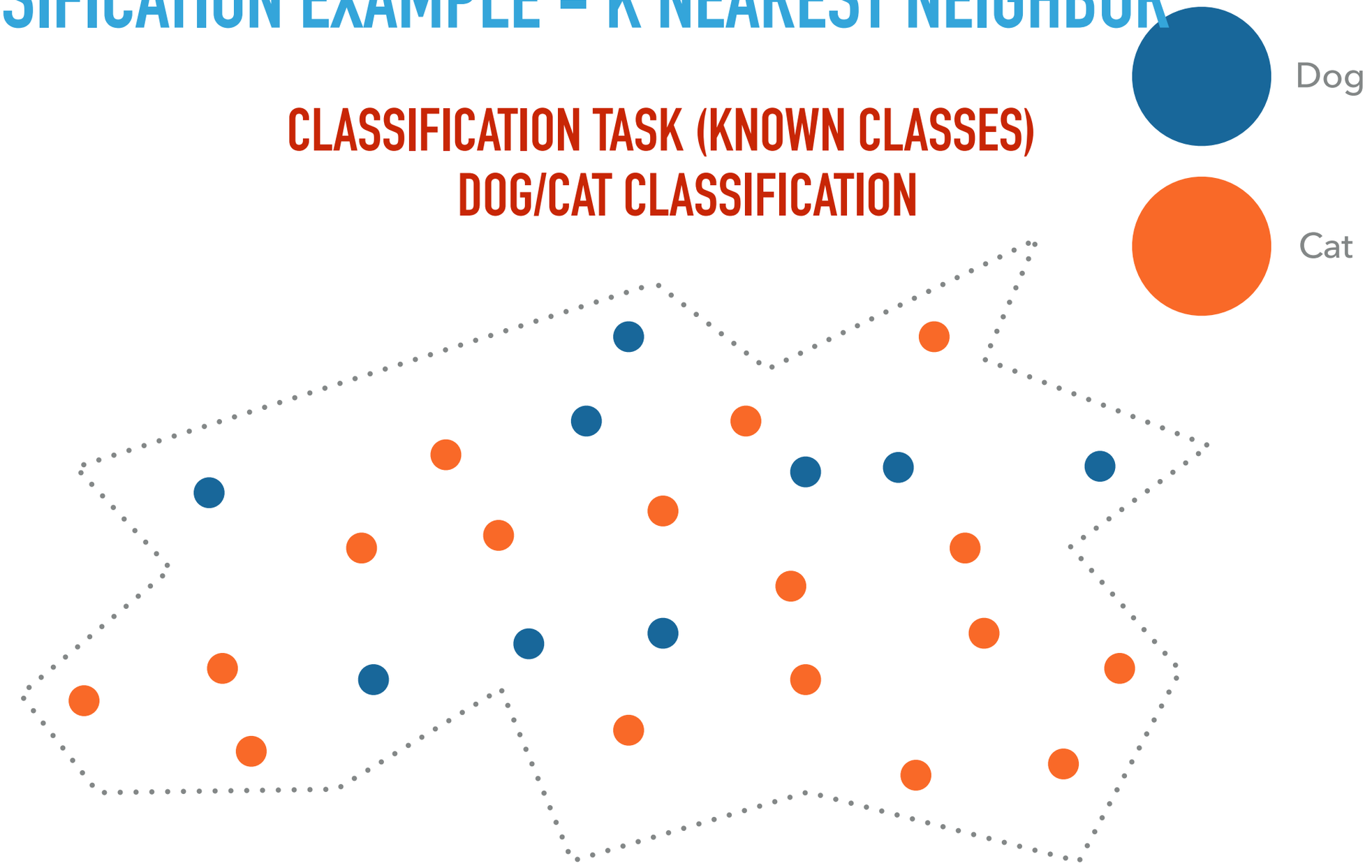
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

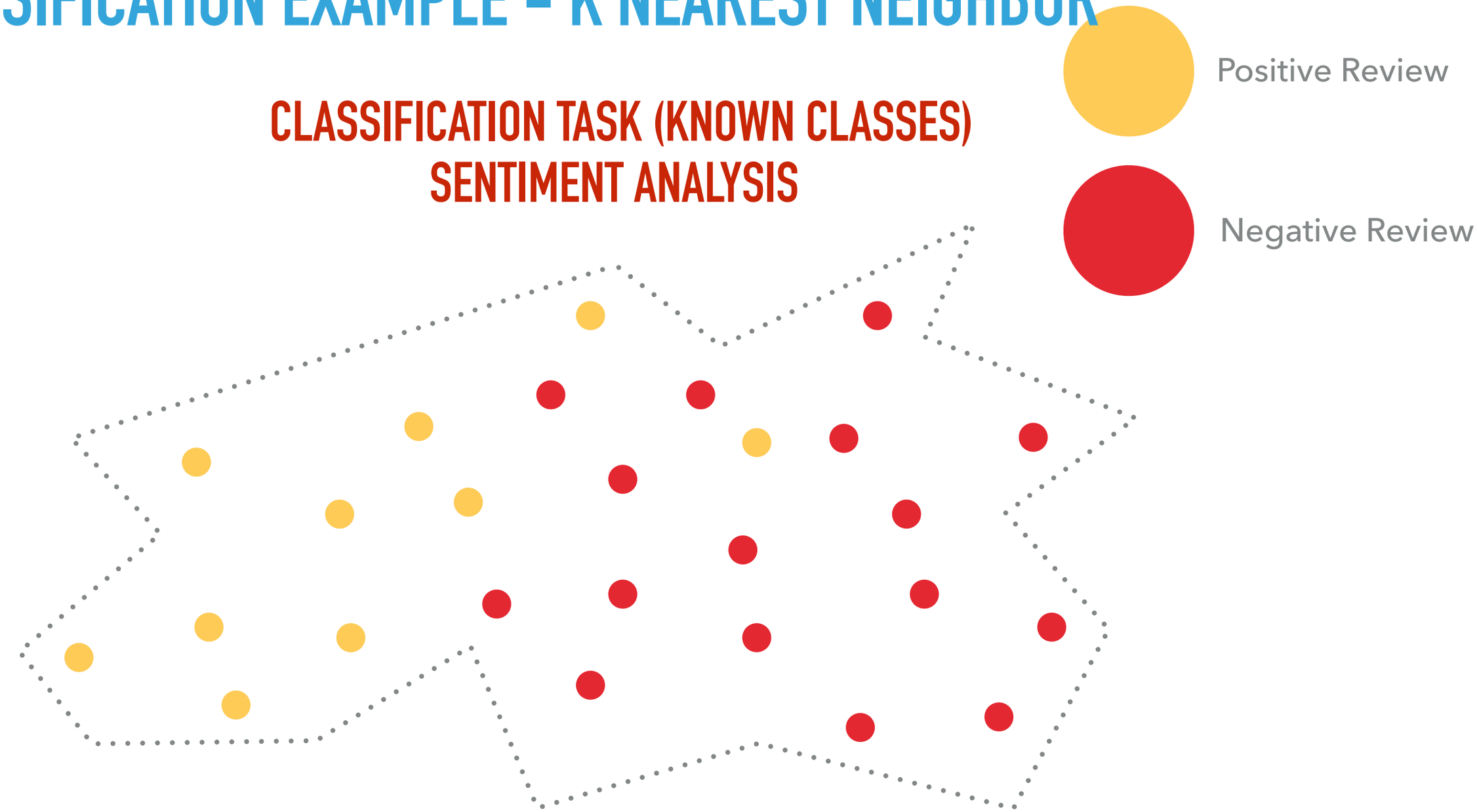


CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

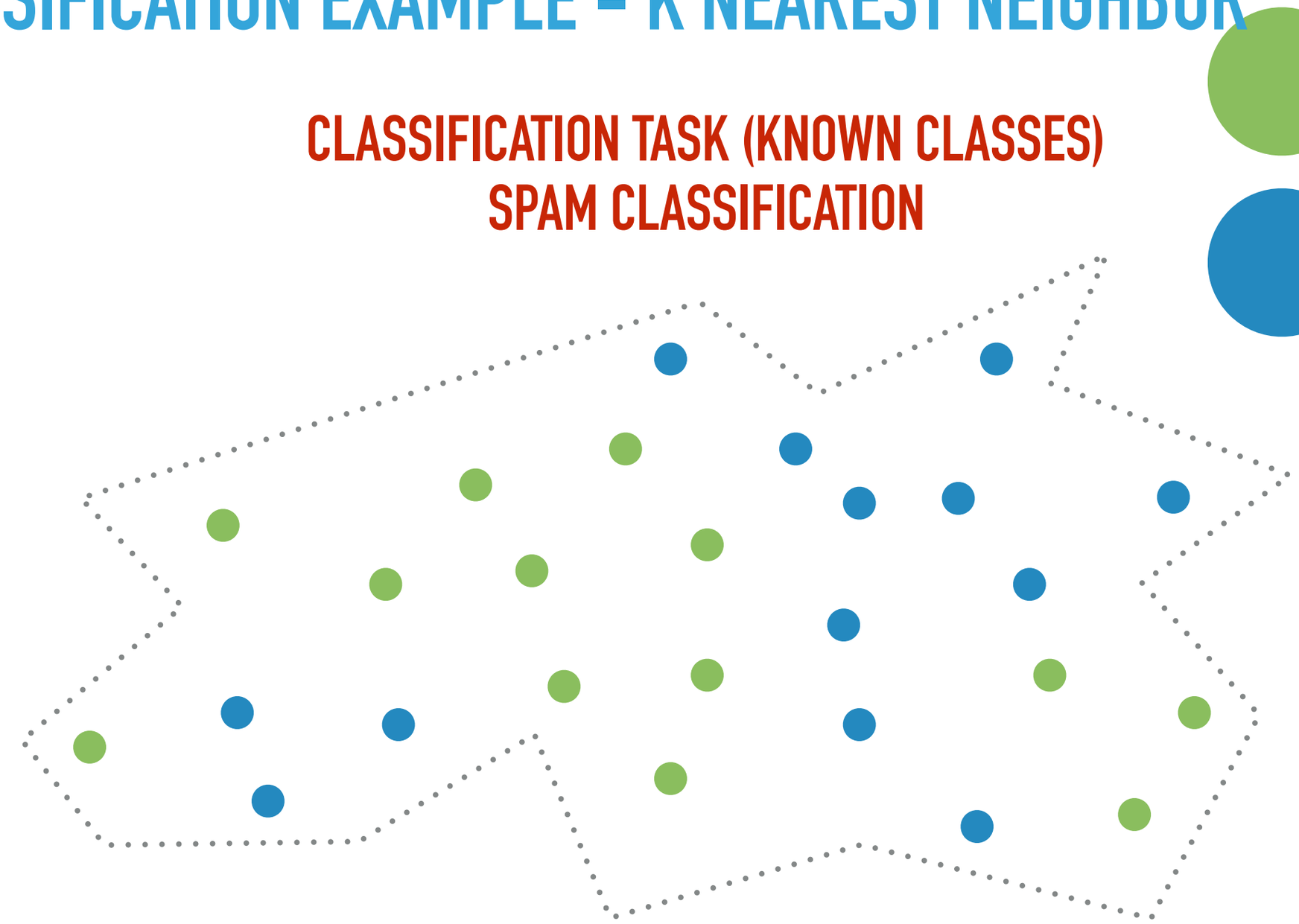


CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

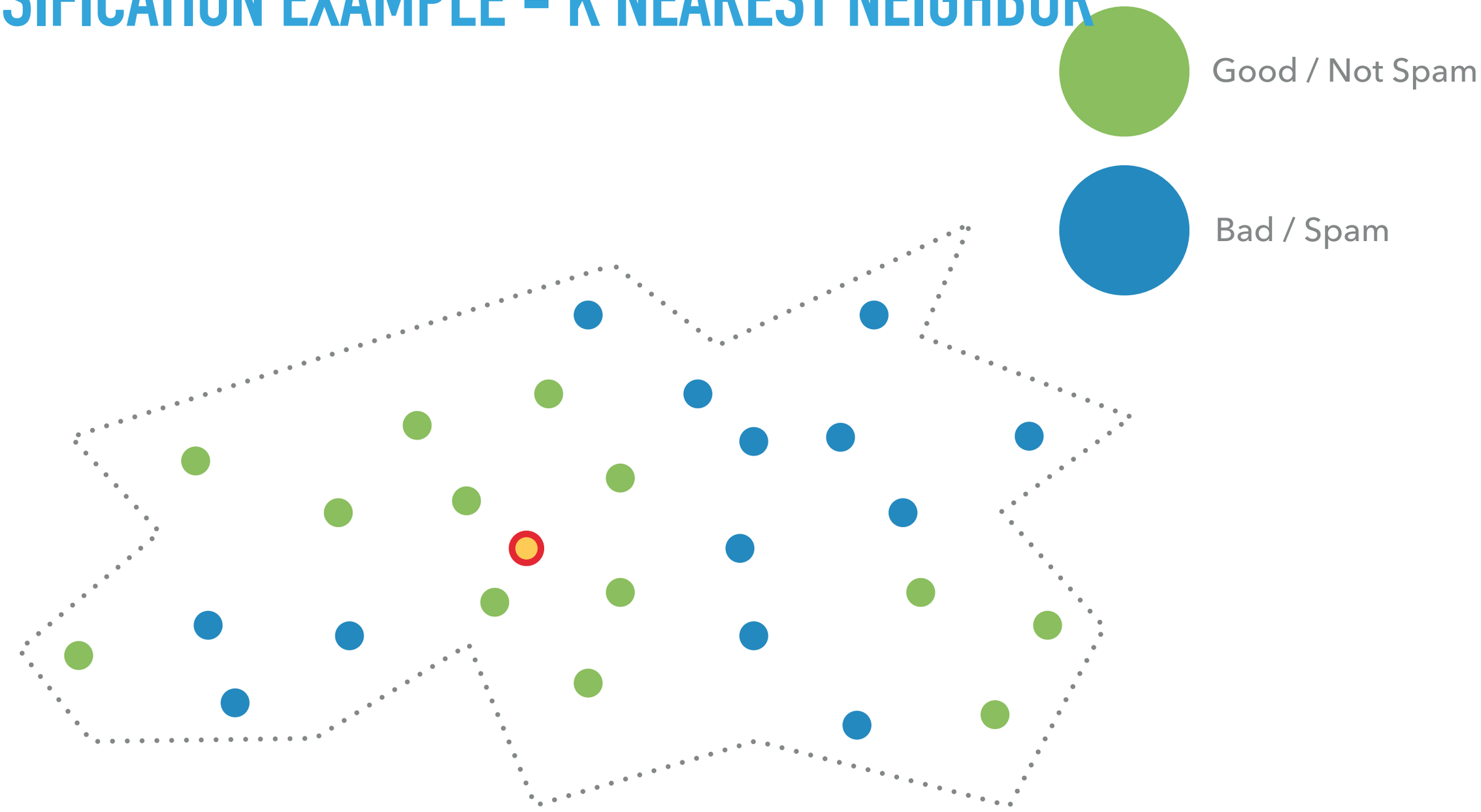
CLASSIFICATION TASK (KNOWN CLASSES)
SENTIMENT ANALYSIS



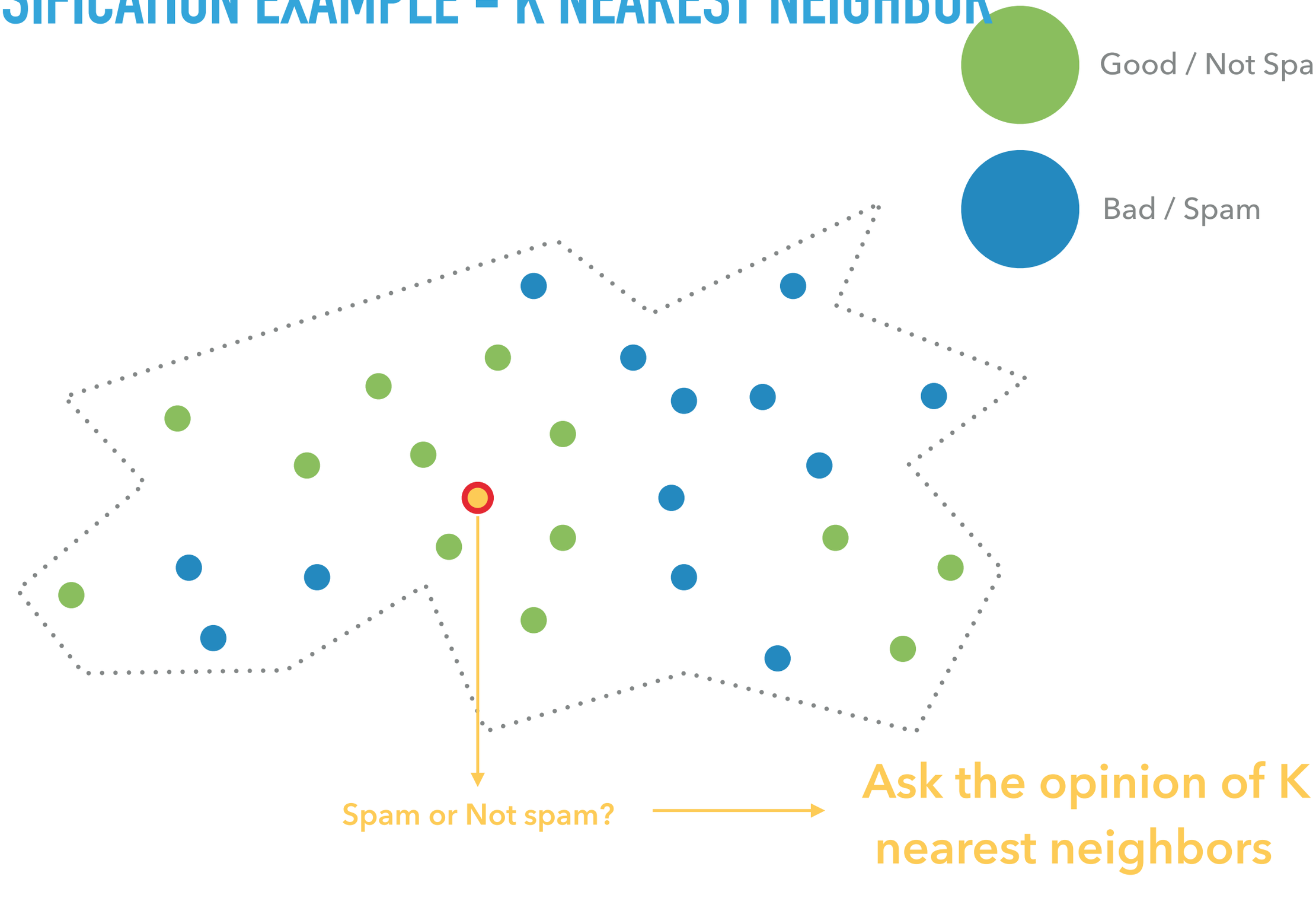
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



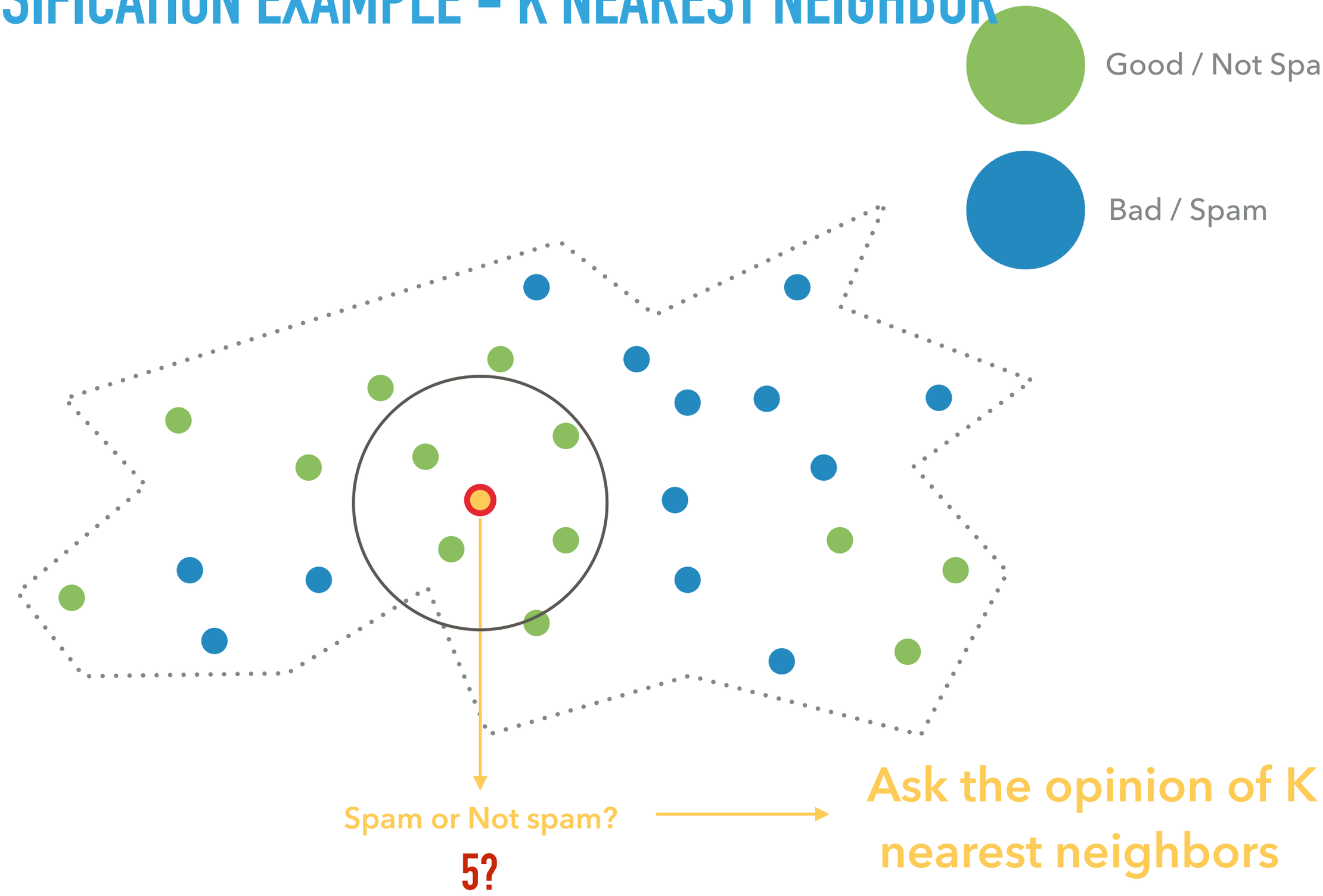
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



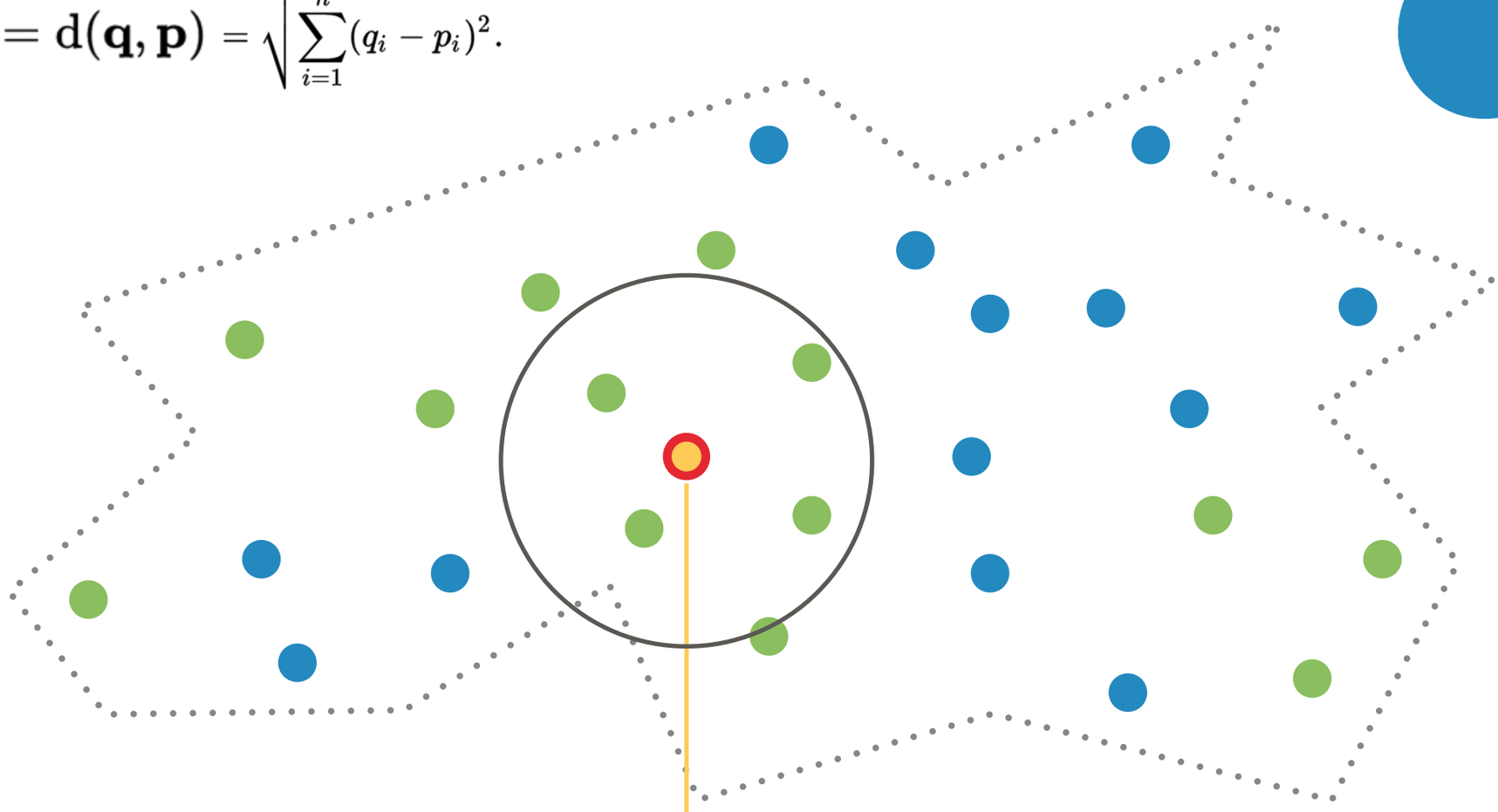
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

CALCULATE THE DISTANCE FROM THE TEST POINT (P) TO ALL OTHER POINTS (Q)

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Good / Not Spam

Bad / Spam

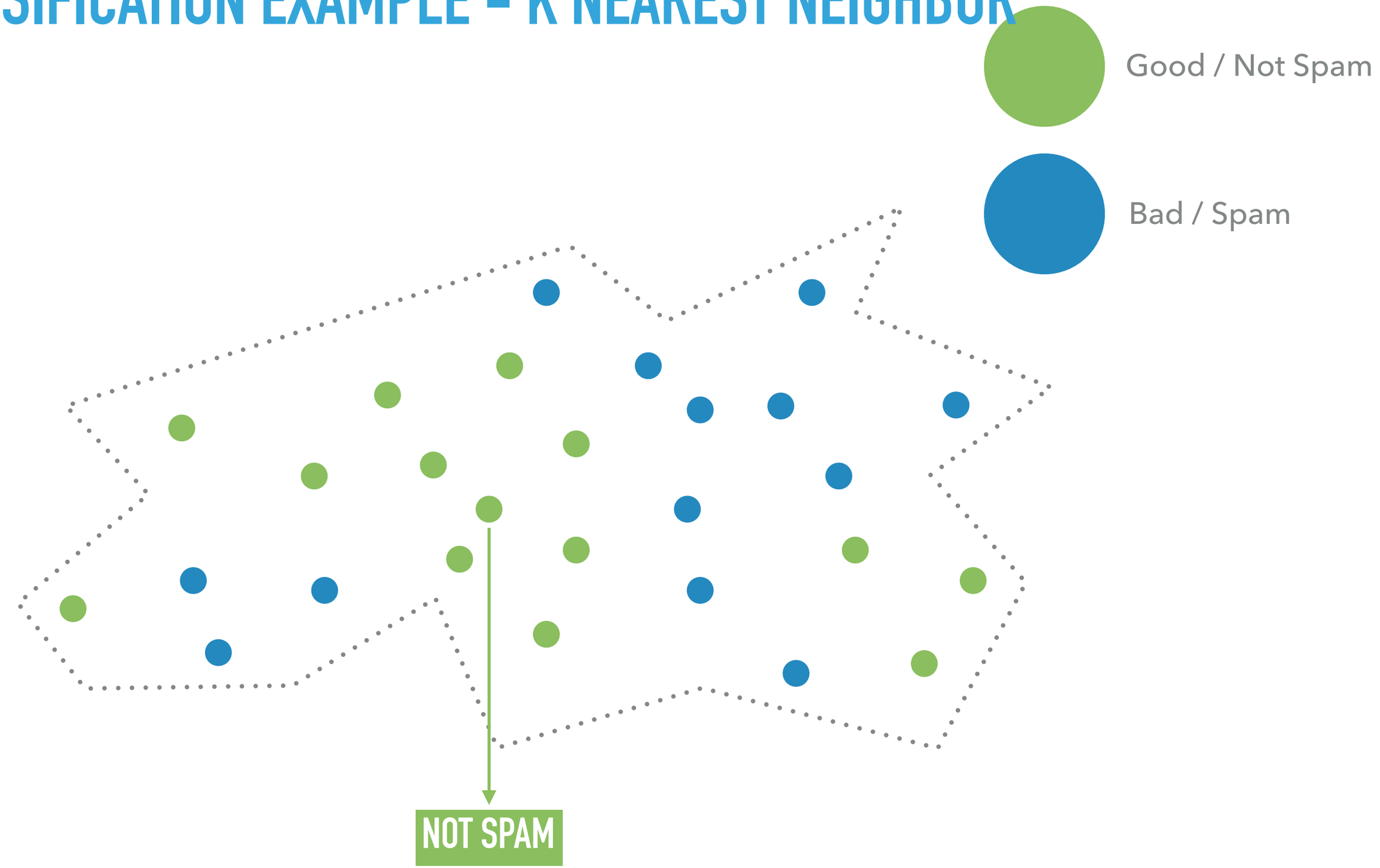


Spam or Not spam?

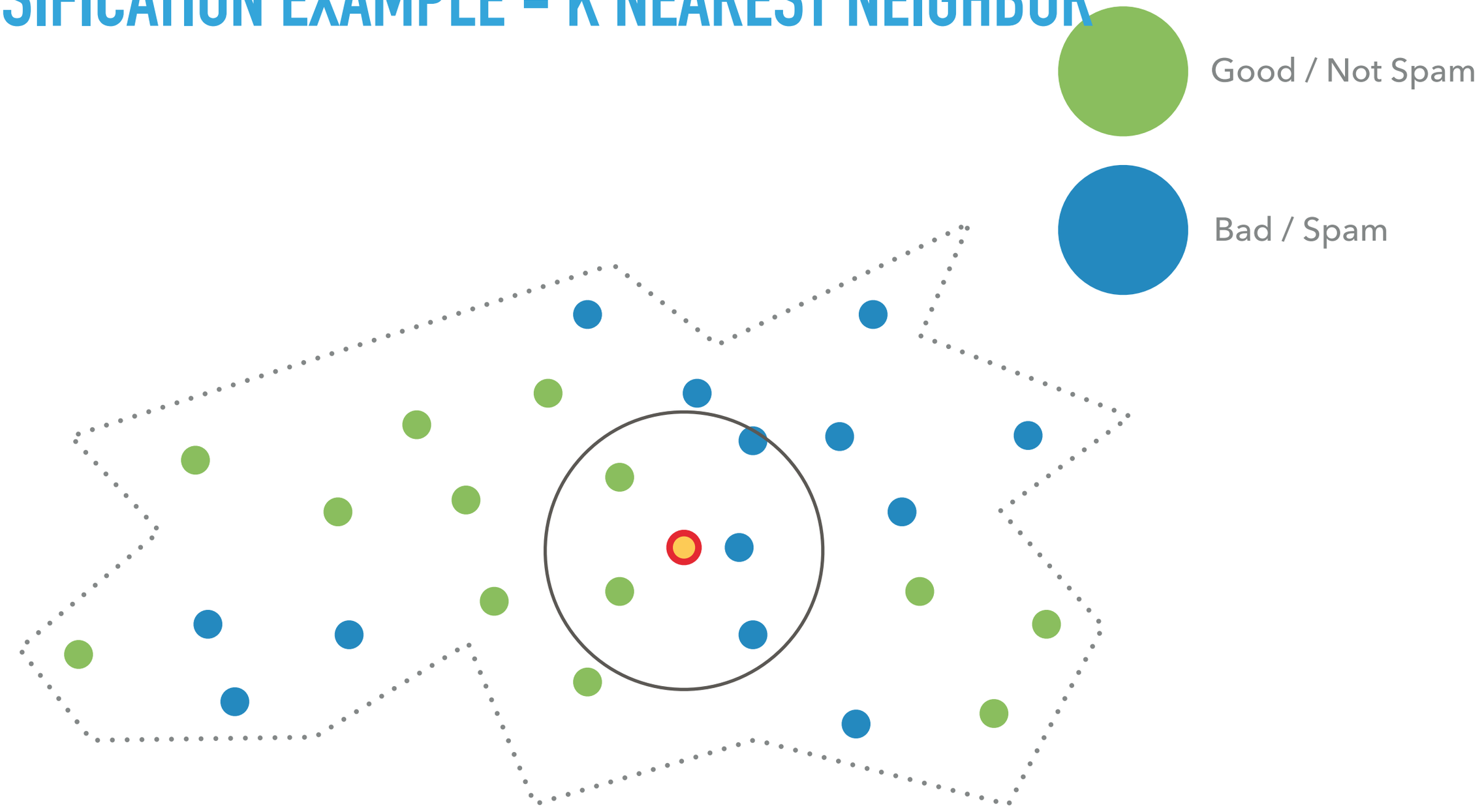
5?

Ask the opinion of K nearest neighbors

CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

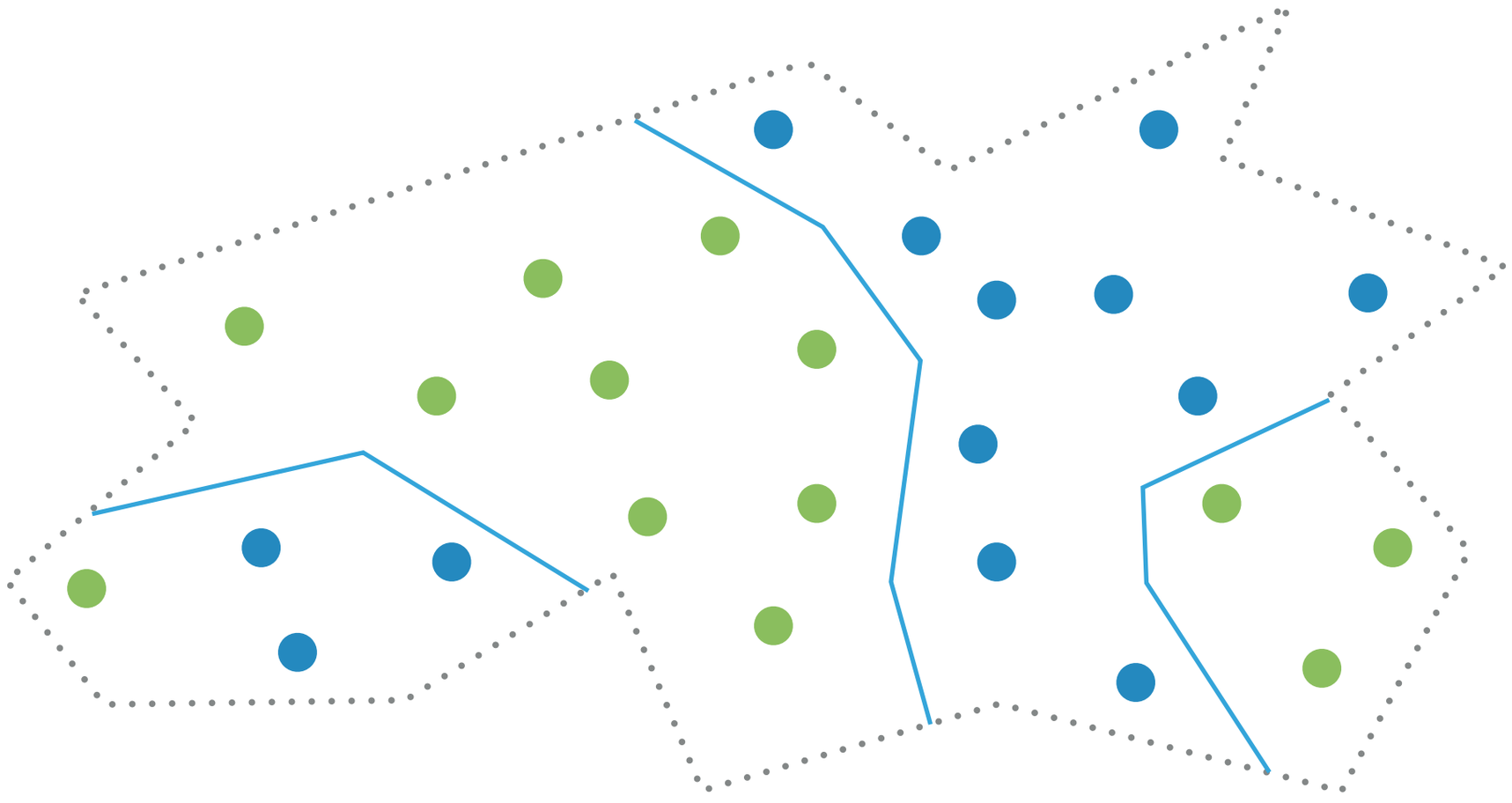


CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



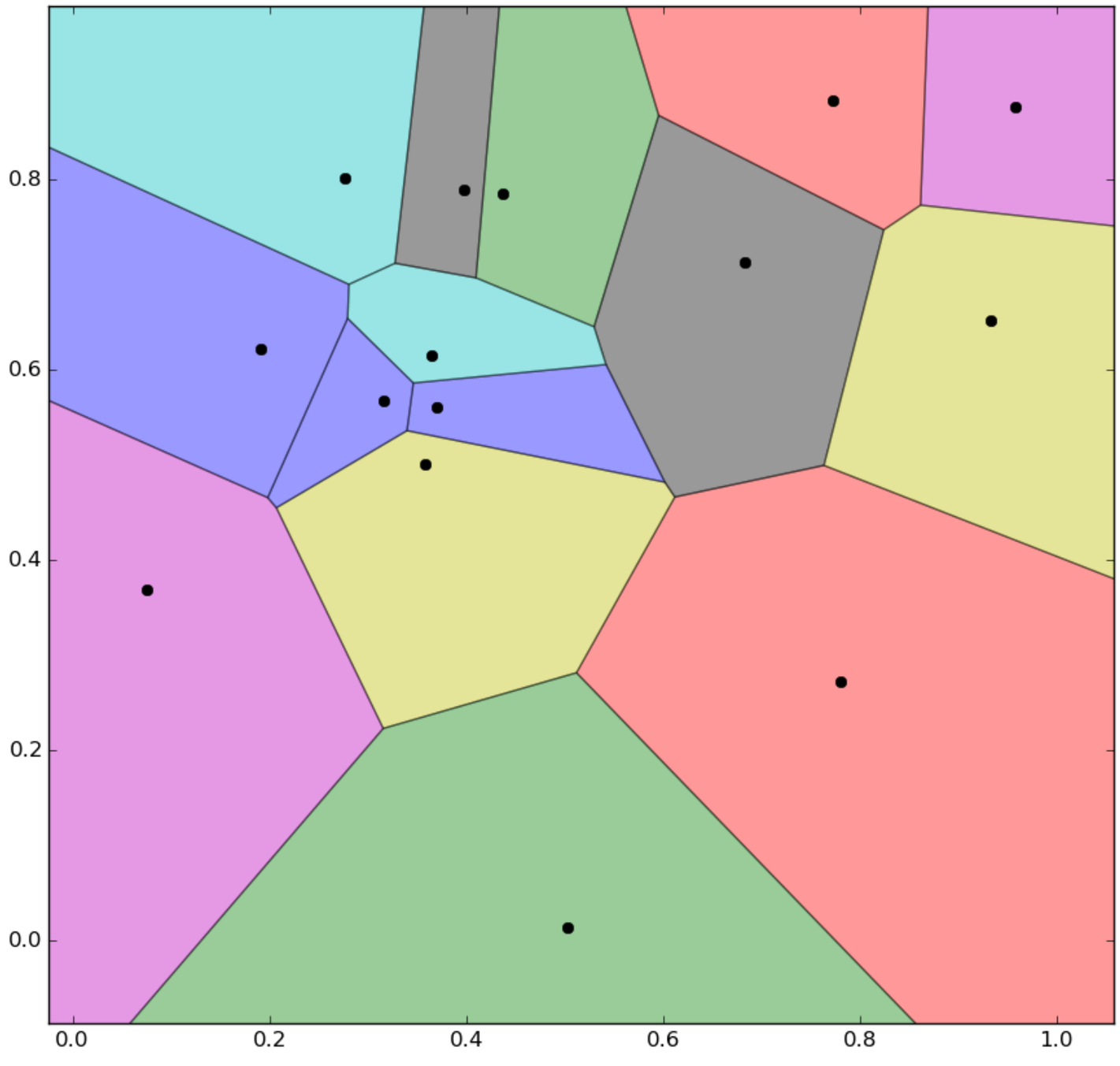
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR

VORONOI DIAGRAM



Maybe not accurate!

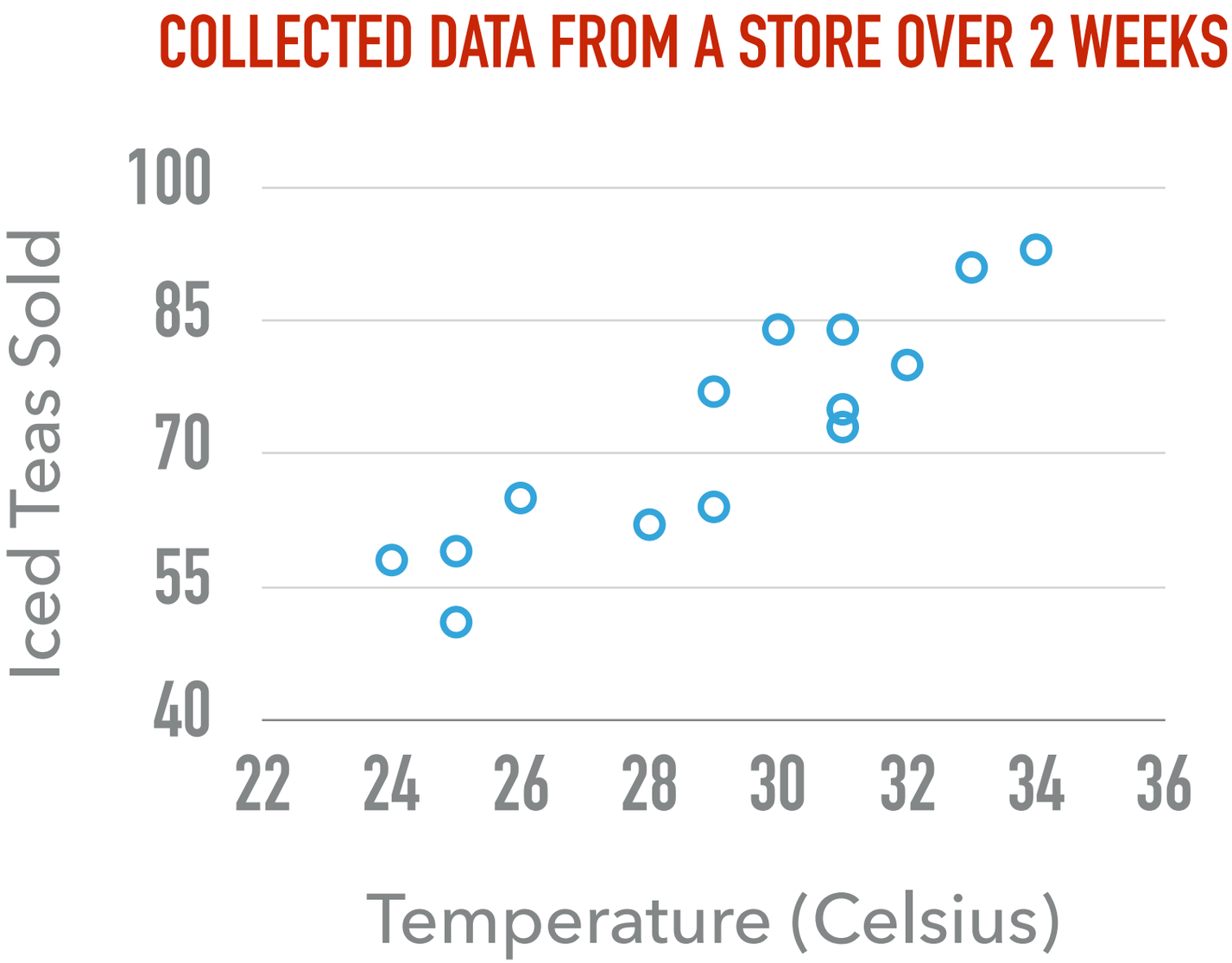
CLASSIFICATION EXAMPLE – K NEAREST NEIGHBOR



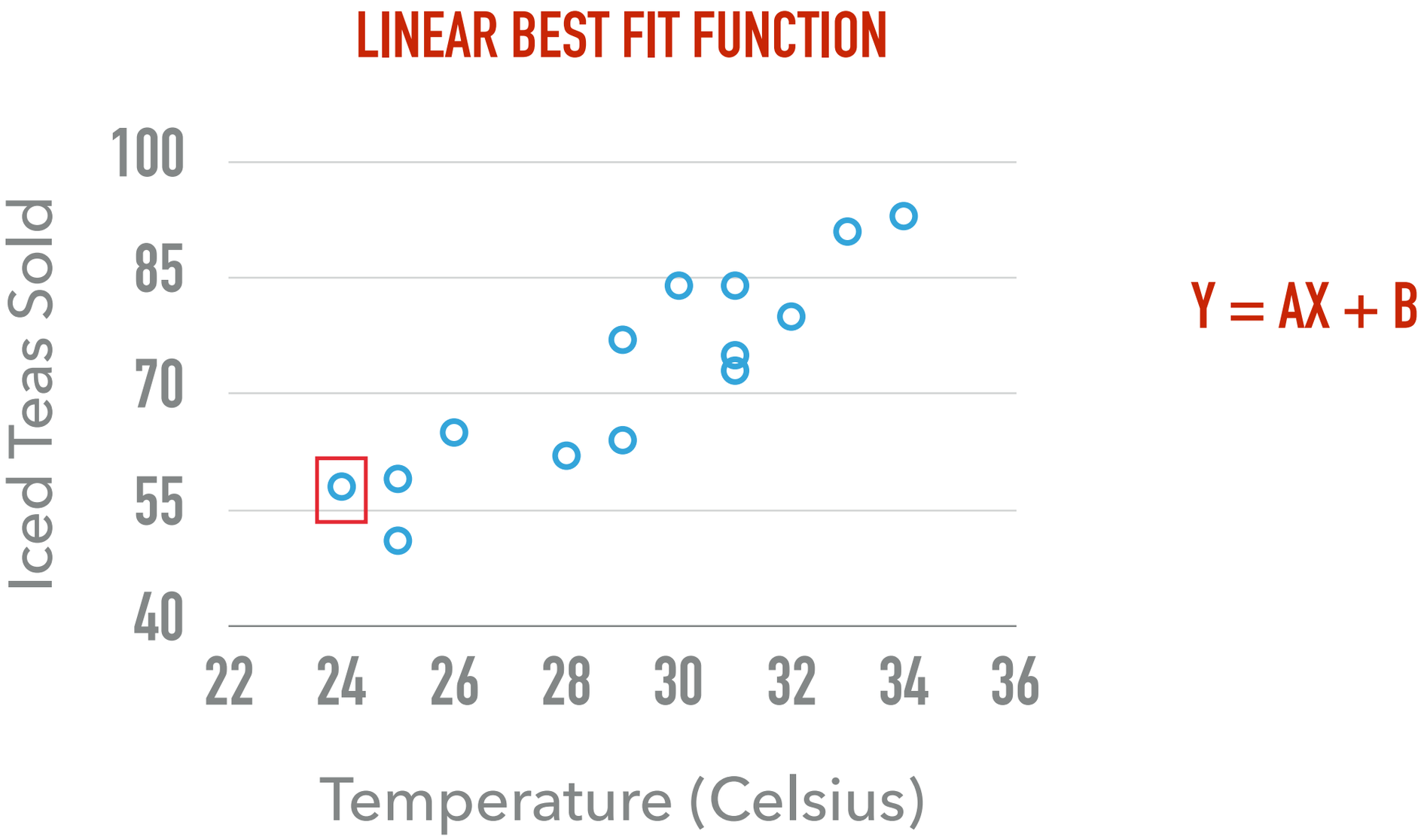
K= 1

This one is accurate!

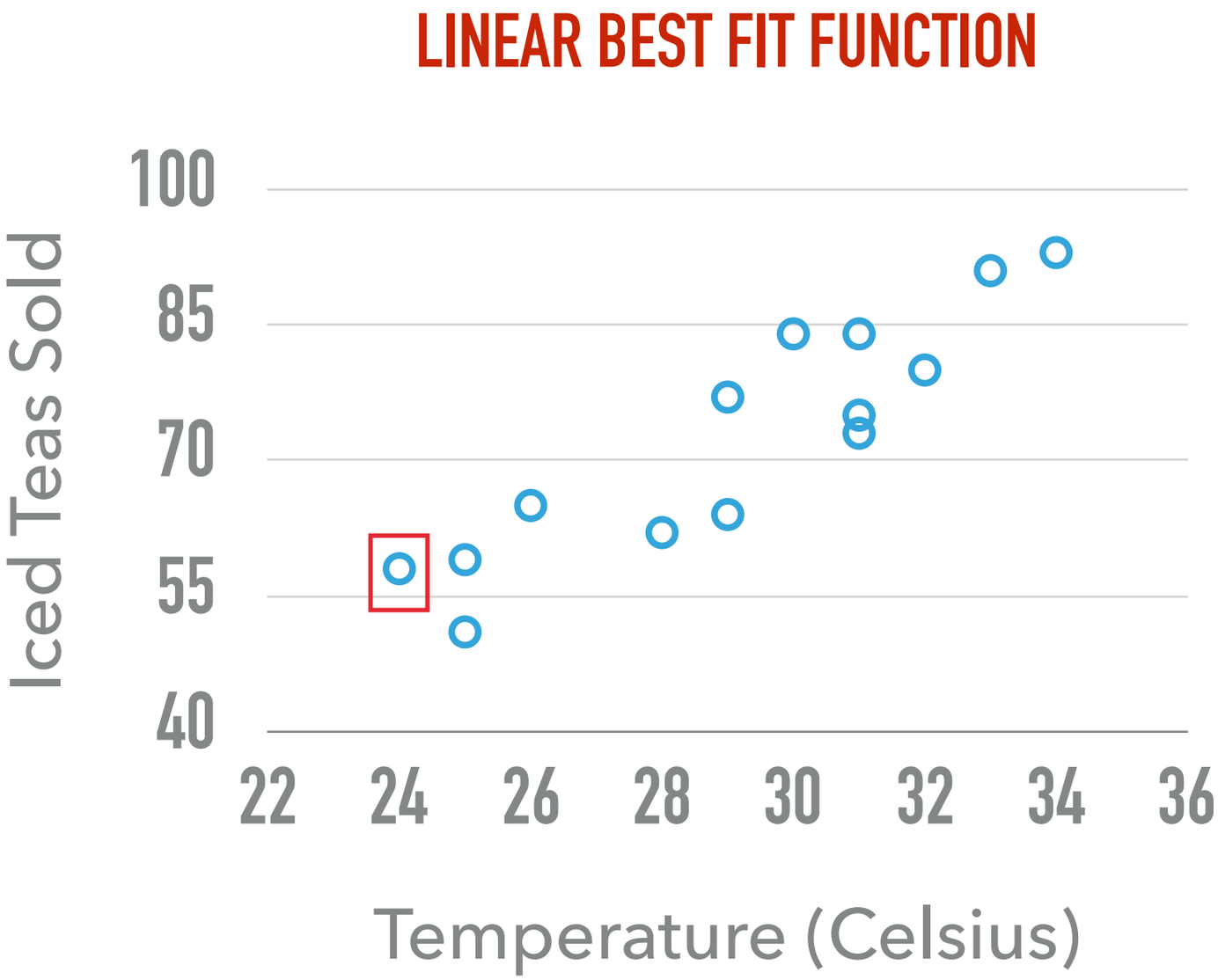
REGRESSION EXAMPLE – LINEAR REGRESSION



REGRESSION EXAMPLE – LINEAR REGRESSION



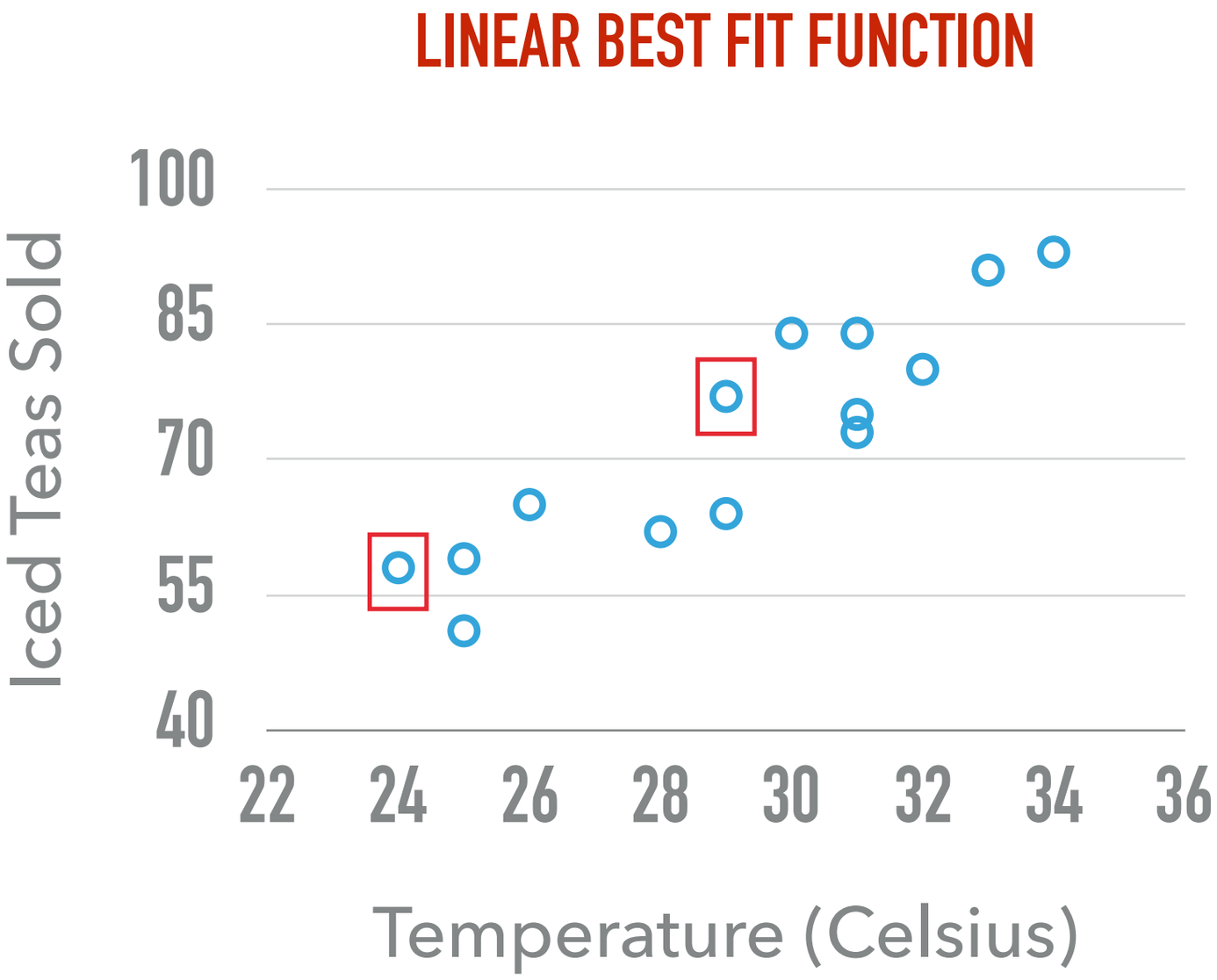
REGRESSION EXAMPLE – LINEAR REGRESSION



$Y = AX + B$

$(58) = A\,(24) + B$

REGRESSION EXAMPLE – LINEAR REGRESSION

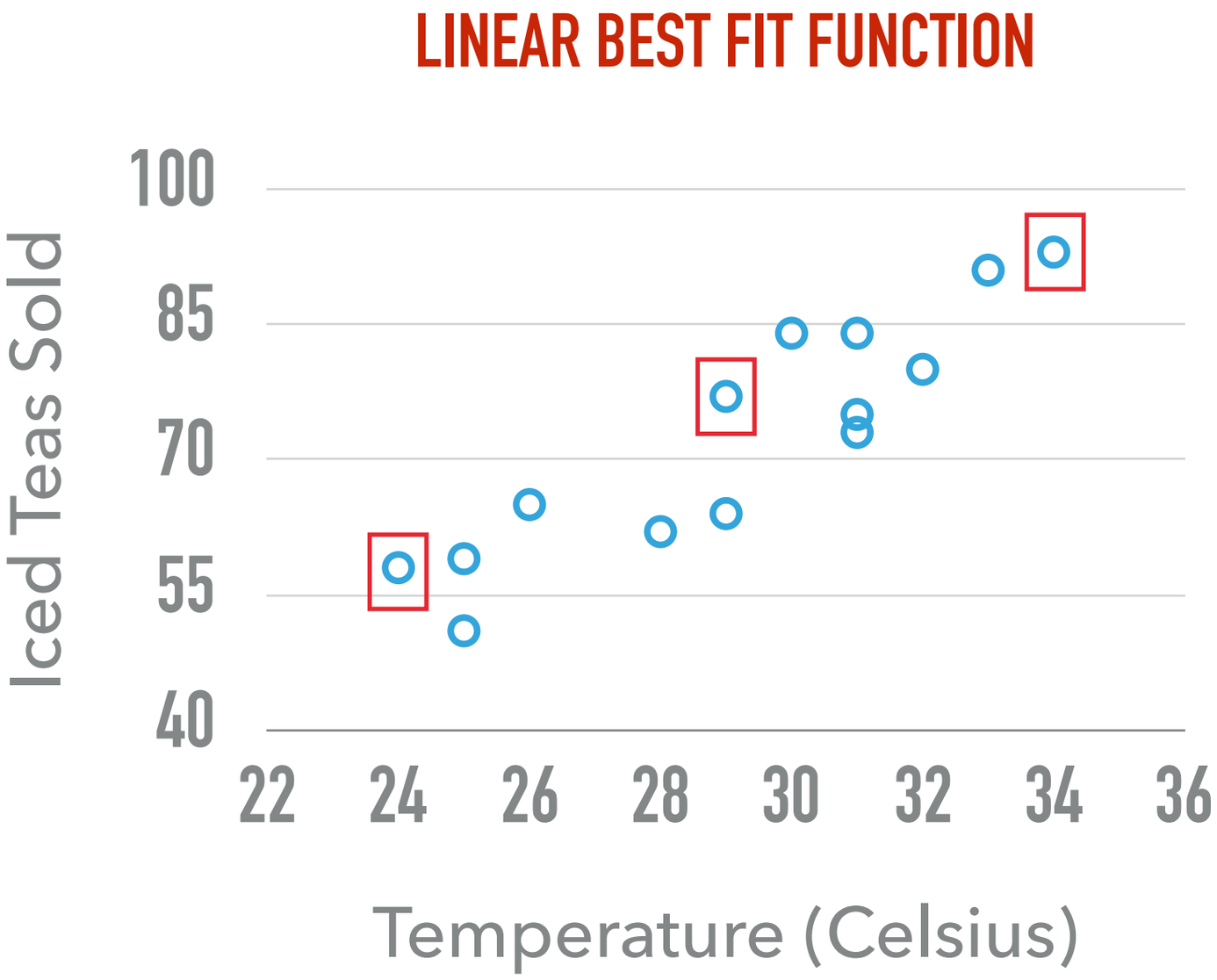


$Y = AX + B$

$(58) = A\,(24) + B$

$(77) = A\,(29) + B$

REGRESSION EXAMPLE – LINEAR REGRESSION



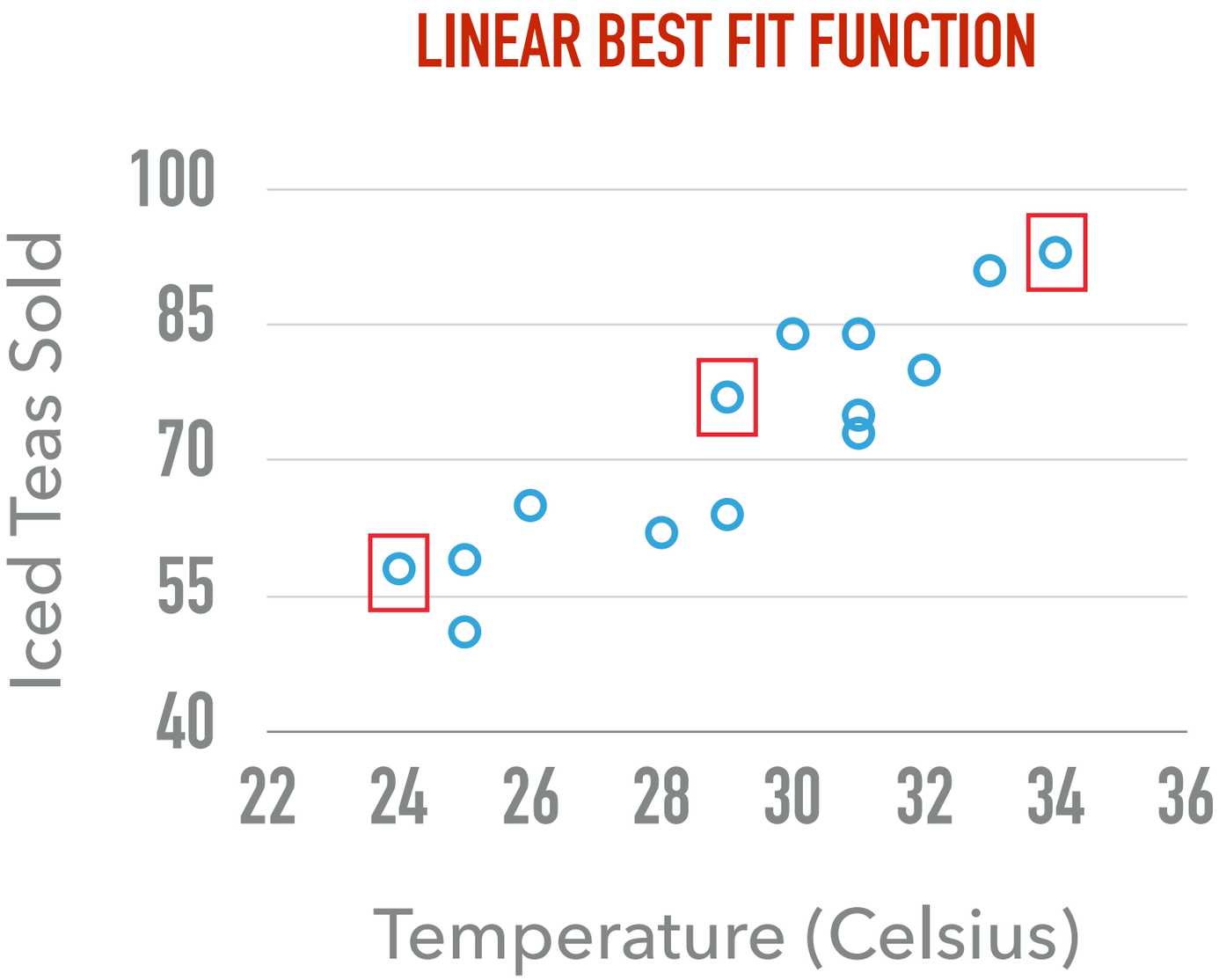
$Y = AX + B$

$(58) = A\,(24) + B$

$(77) = A\,(29) + B$

$(93) = A\,(34) + B$

REGRESSION EXAMPLE – LINEAR REGRESSION

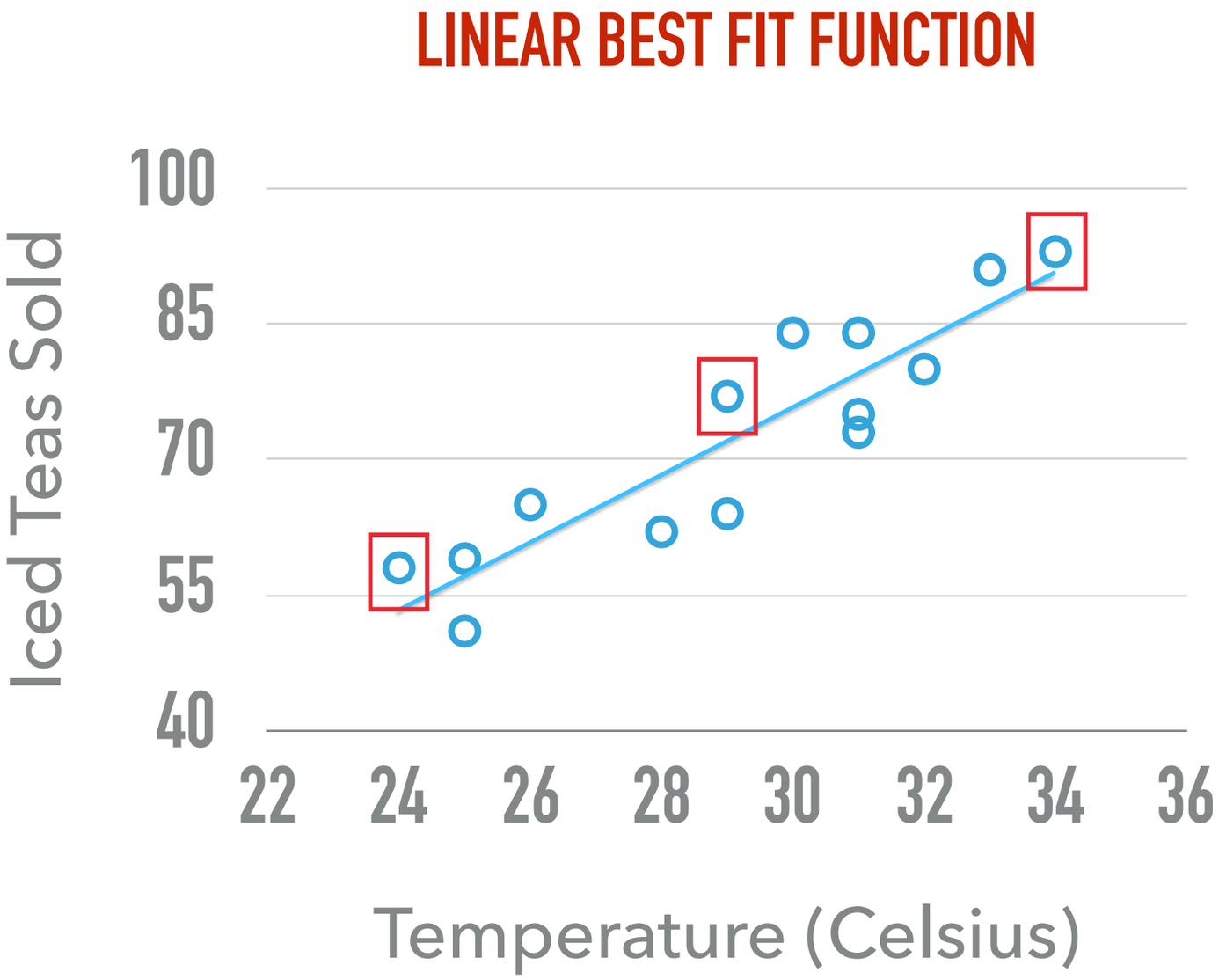


$$Y = AX + B$$

$$(58) = A\,(24) + B$$
$$(77) = A\,(29) + B$$
$$(93) = A\,(34) + B$$

What A and B values that I minimize the measured error?

REGRESSION EXAMPLE – LINEAR REGRESSION



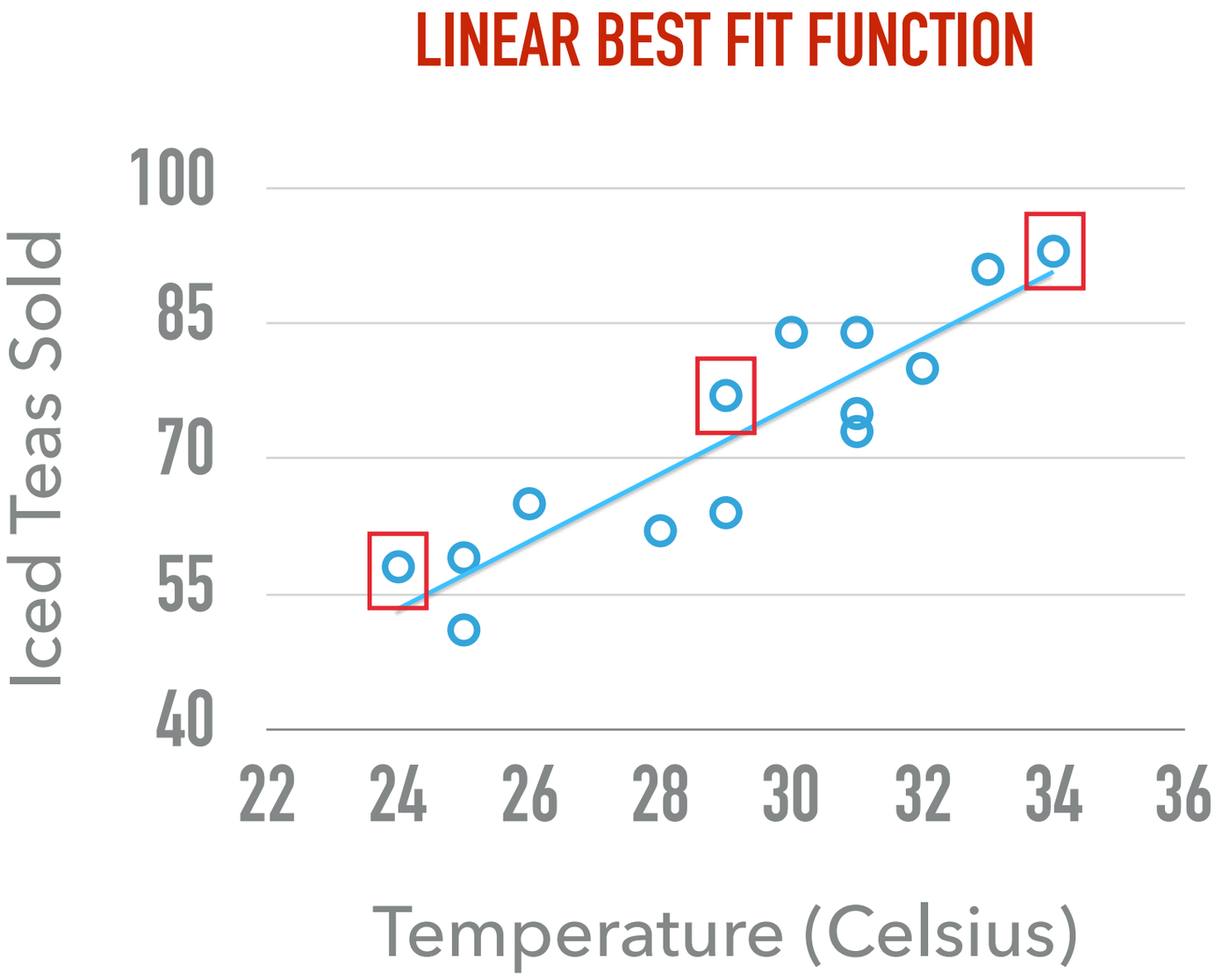
$$Y = AX + B$$

$$(58) = A\,(24) + B$$
$$(77) = A\,(29) + B$$
$$(93) = A\,(34) + B$$

What A and B values that I minimize the measured error?

$$Y = 3.7\,X - 36.4$$

REGRESSION EXAMPLE – LINEAR REGRESSION



$$Y = AX + B$$

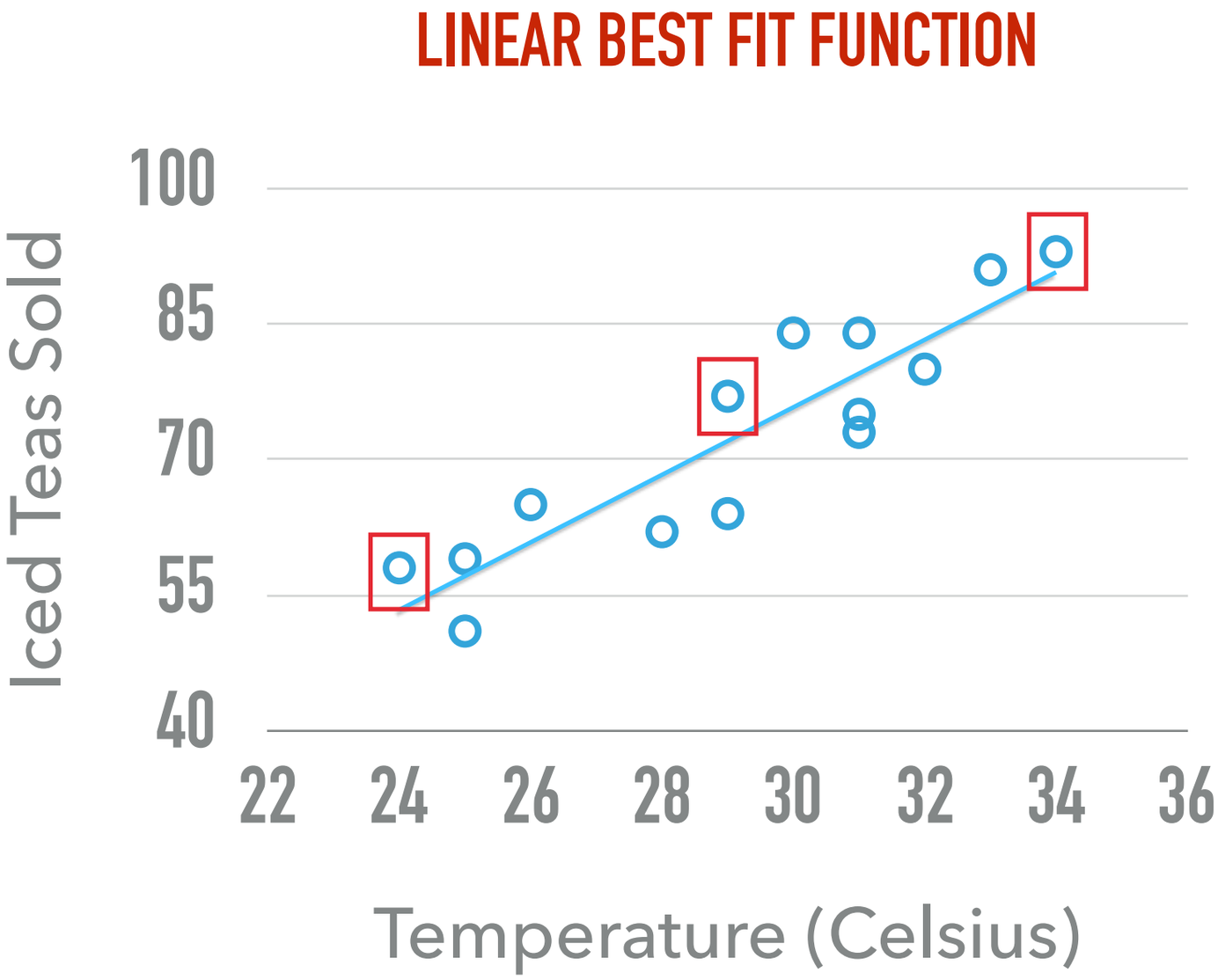
$$(58) = A\,(24) + B$$
$$(77) = A\,(29) + B$$
$$(93) = A\,(34) + B$$

New day starting with 34 C,
how many iced teas should I expect to sell today?

What A and B values that I minimize the measured error?

$$Y = 3.7\,X - 36.4$$

REGRESSION EXAMPLE – LINEAR REGRESSION



$$Y = AX + B$$

$$(58) = A\,(24) + B$$
$$(77) = A\,(29) + B$$
$$(93) = A\,(34) + B$$

New day starting with 34 C,
how many iced teas should I expect to sell today?

$3.7 \times 34 - 36.4 = 89.4$

What A and B values that I minimize the measured error?

$$Y = 3.7\,X - 36.4$$

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature
  - Popularity of store
  - Friendliness of staff
  - Average wage of population

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature X_1
  - Popularity of store X_2
  - Friendliness of staff X_3
  - Average wage of population X_4

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature X_1
  - Popularity of store X_2
  - Friendliness of staff X_3
  - Average wage of population X_4

- $y = a X_1 + b X_2 + c X_3 + d X_4$

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature X_1
  - Popularity of store X_2
  - Friendliness of staff X_3
  - Average wage of population X_4
- $y = a X_1 + b X_2 + c X_3 + d X_4$

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature X_1
  - Popularity of store X_2
  - Friendliness of staff X_3
  - Average wage of population X_4

} FEATURES

- $y = a X_1 + b X_2 + c X_3 + d X_4$

REGRESSION EXAMPLE – LINEAR REGRESSION

- Multiple variables could be used to estimate the number of iced teas sold:
  - Day temperature X_1
  - Popularity of store X_2
  - Friendliness of staff X_3
  - Average wage of population X_4
- $y = a X_1 + b X_2 + c X_3 + d X_4$

} FEATURES

FEATURE ENGINEERING

MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)

MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

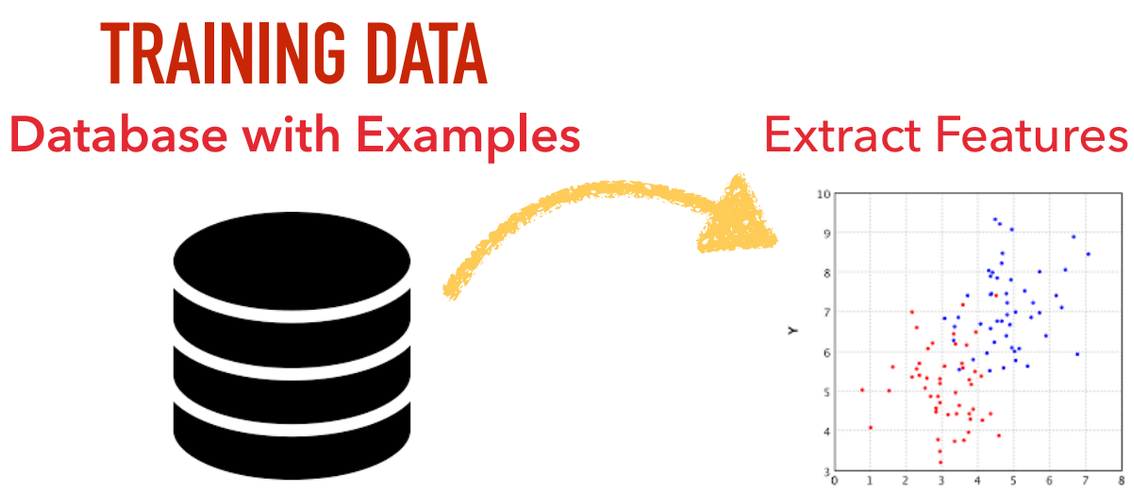
- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)

MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)

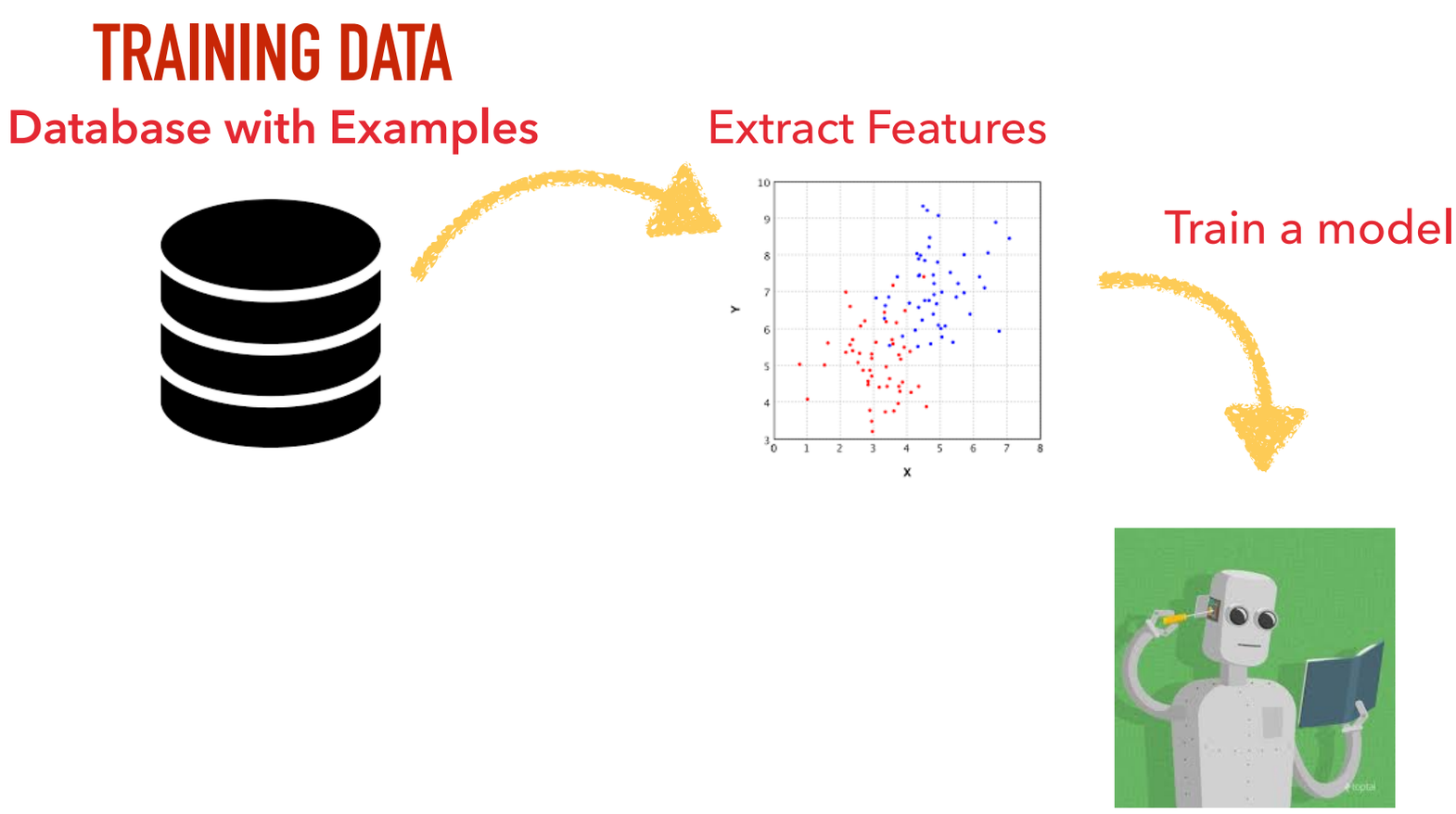
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



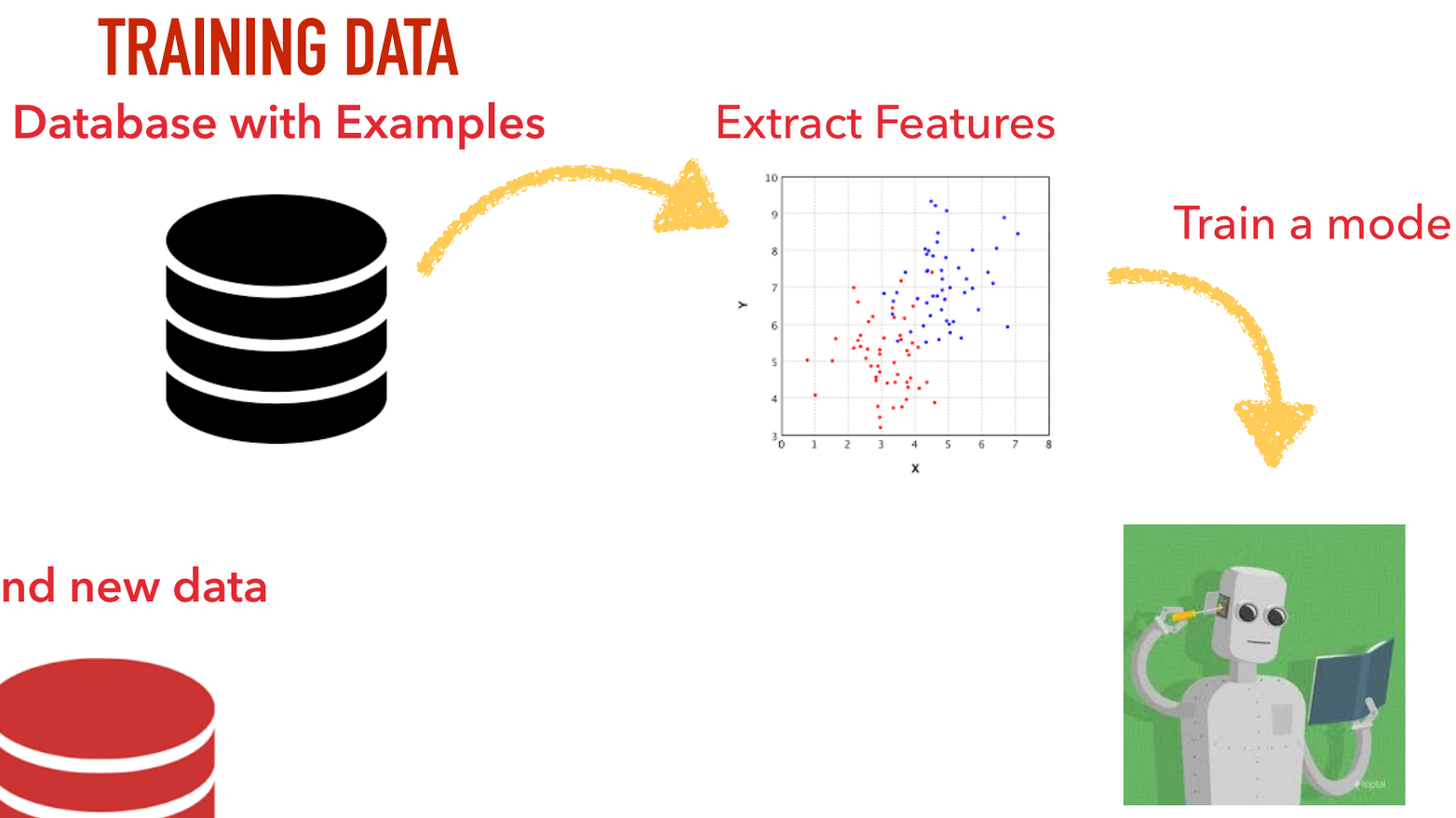
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



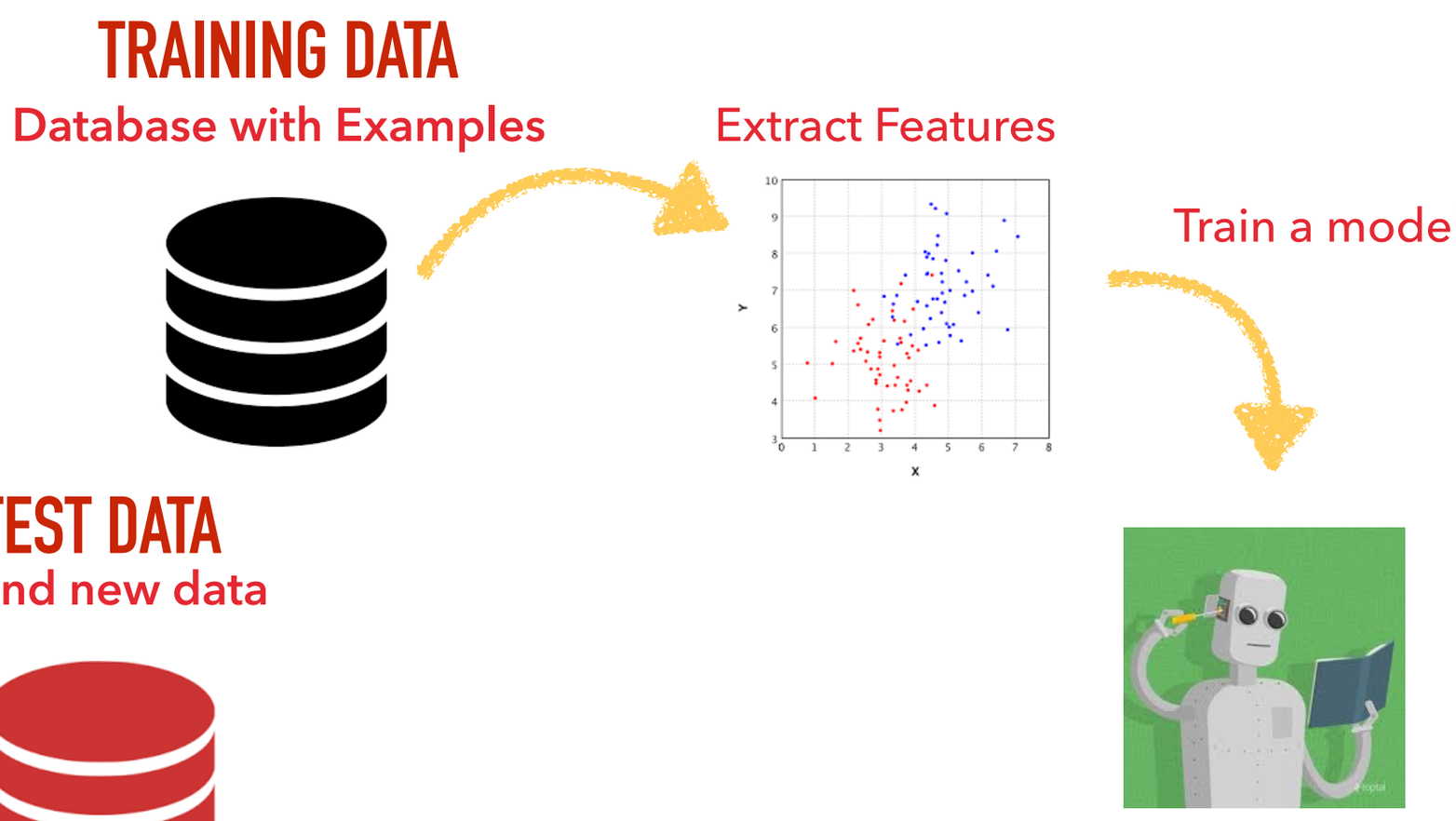
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



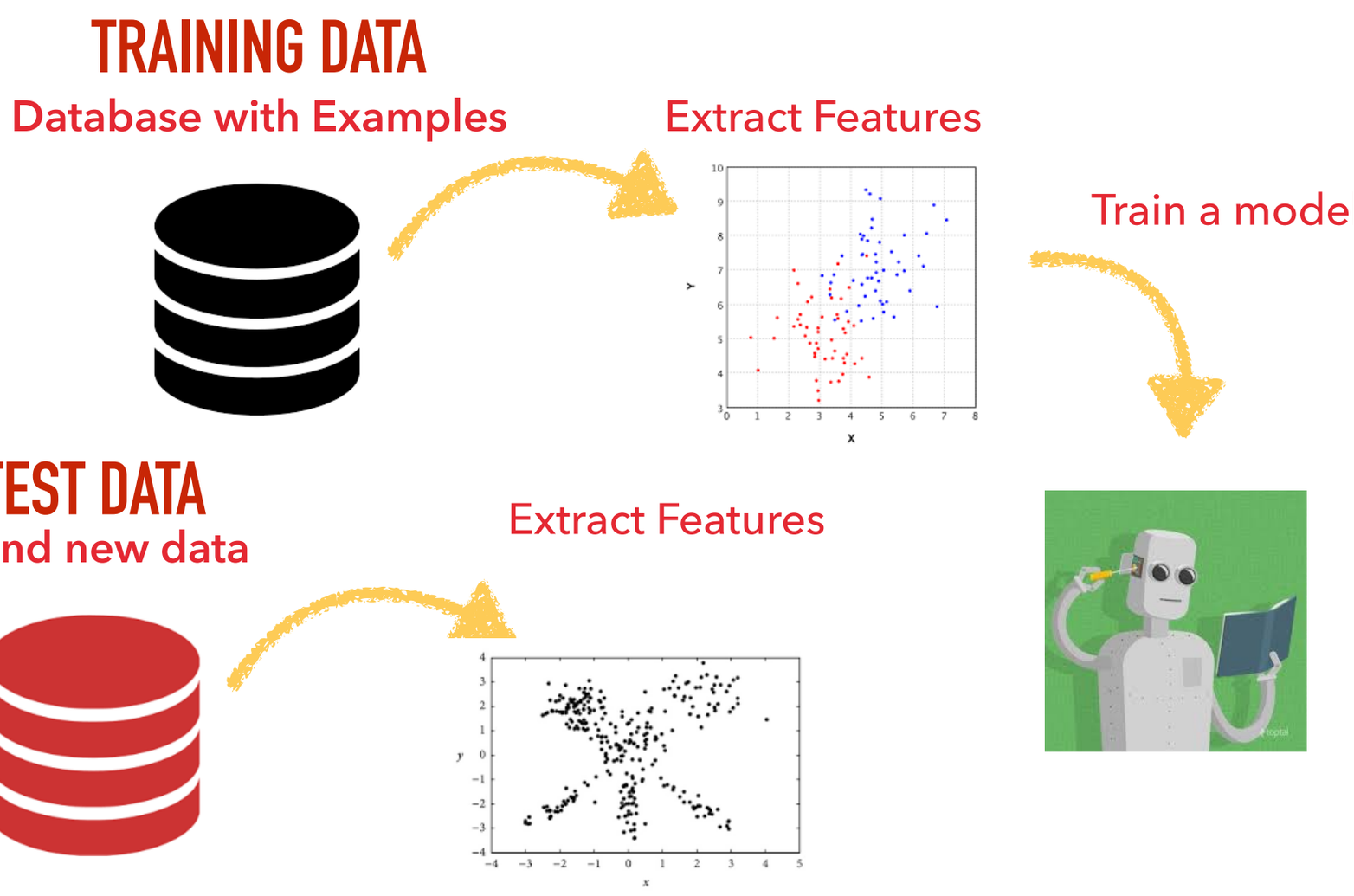
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



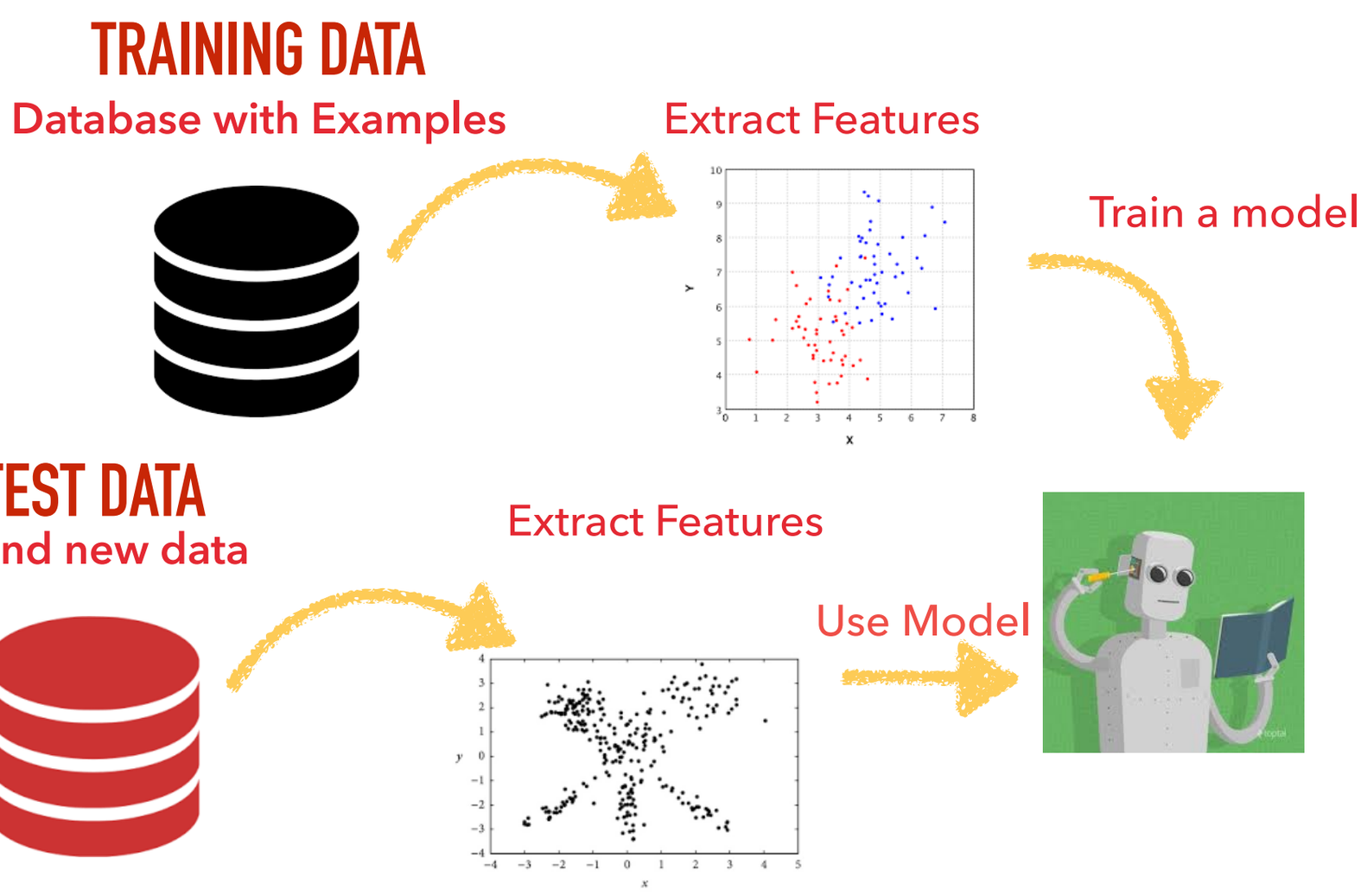
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



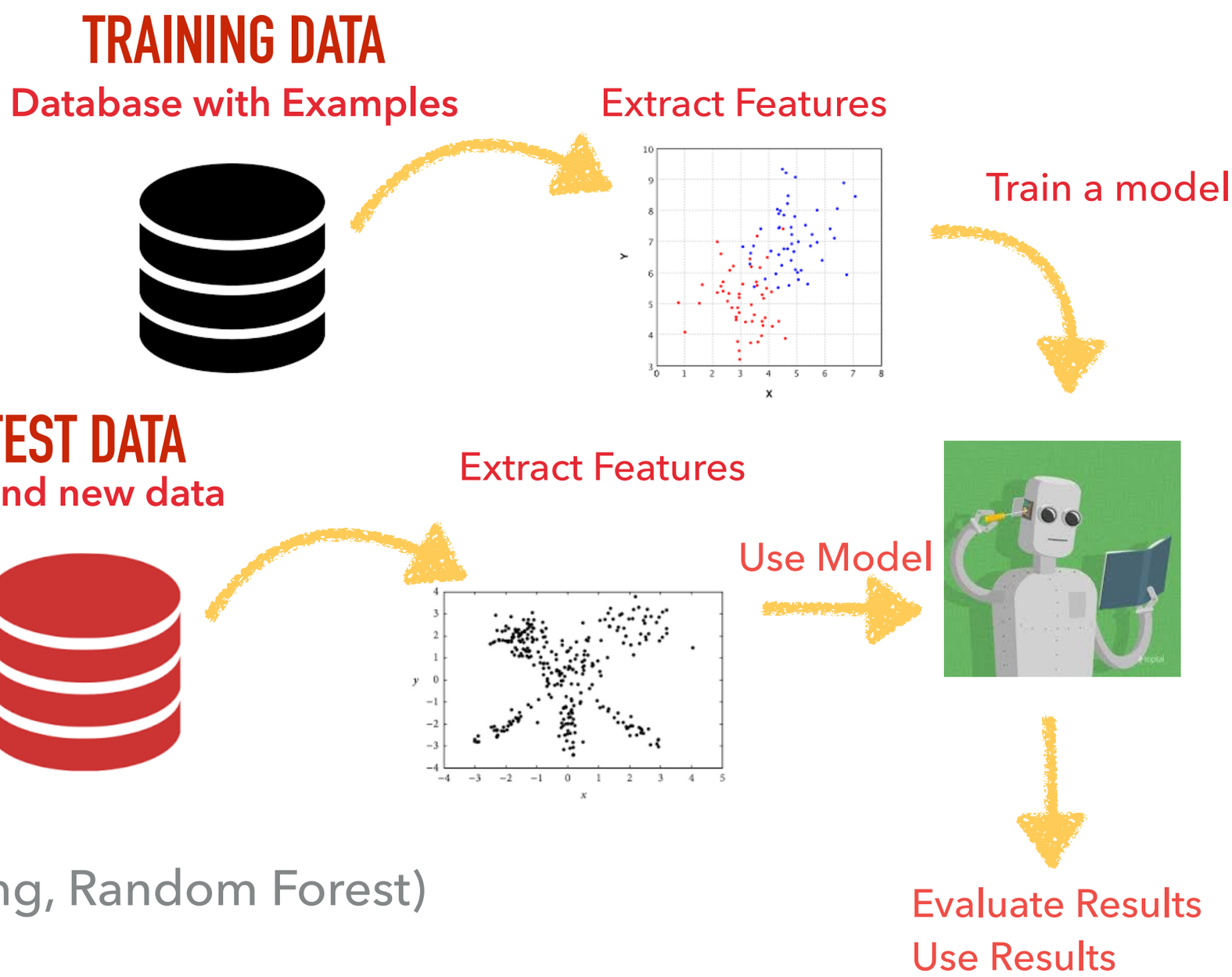
MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



MACHINE LEARNING METHODS FOR CLASSIFICATION AND REGRESSION

- K-NN
- Linear regression
- Decision trees
- Naive Bayes
- Perceptron
- Neural Networks
- Logistic Regression
- Support Vector Machines
- Ensembles (Bagging, Boosting, Random Forest)



REMEMBER FROM LAST LECTURE

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943583°W

Carnegie Mellon University (**Carnegie Mellon** or **CMU**; /kɑːrnɪgi ˈmɛlən/ or /kɑːrˈneɪgi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (4.8 km) from Downtown Pittsburgh. Carnegie Mellon has seven colleges and independent schools: the College of Engineering, College of Fine Arts, Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H. John Heinz III College of Information Systems and Public Policy, and the School of Computer Science. The university also has campuses in Qatar and Silicon Valley, with degree-granting programs in six continents.

Carnegie Mellon consistently ranks in the top 25 in the national *U.S. News & World Report* rankings.[8] It is home to the world's first degree-granting Robotics and Drama programs,[9] as well as one of the first Computer Science departments.[10] The university spent $242 million on research in 2015.[11]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 19 Nobel Prize Laureates, 19 Members of the American Academy of Arts & Sciences,[12] 72 Members of the National Academies, 114 Emmy Award Winners, 43 Tony Award laureates, 7 Academy Award Winners, 12 Turing Award winners, 4 Rhodes Scholars, and one Schwarzman Scholar.[13]

Contents [hide]

Institutional formation [edit]

Carnegie Mellon University

Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) [1] Carnegie Mellon University (1988–Present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie 1967 (merger with Mellon Institute)
Endowment	$1.709 billion (2016)[2]
President	Subra Suresh
Provost	Farnam Jahanian[3]
Academic staff	1,423[4]
Undergraduates	6,362
Postgraduates	7,141
Other students	145
Location	Pittsburgh, Pennsylvania, United States
Campus	Urban, 140 acres (57 ha)[5]
Colors	Cardinal, Black, Grey and White
Athletics	NCAA Division III UAA, ACHA, IRA 17 varsity teams[6]
Nickname	Tartans
Mascot	Scotty the Scottish Terrier[7]
Website	www.cmu.edu

Carnegie Mellon University

REMEMBER FROM LAST LECTURE

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943583°W

Carnegie Mellon University (**Carnegie Mellon** or **CMU**; /kɑːrnɪgi ˈmɛlən/ or /kɑːrneɪgi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (4.8 km) from Downtown Pittsburgh. Carnegie Mellon has seven colleges and independent schools: the College of Engineering, College of Fine Arts, Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H. John Heinz III College of Information Systems and Public Policy, and the School of Computer Science. The university also has campuses in Qatar and Silicon Valley, with degree-granting programs in six continents.

Carnegie Mellon consistently ranks in the top 25 in the national *U.S. News & World Report* rankings.[8] It is home to the world's first degree-granting Robotics and Drama programs,[9] as well as one of the first Computer Science departments.[10] The university spent $242 million on research in 2015.[11]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 19 Nobel Prize Laureates, 19 Members of the American Academy of Arts & Sciences,[12] 72 Members of the National Academies, 114 Emmy Award Winners, 43 Tony Award laureates, 7 Academy Award Winners, 12 Turing Award winners, 4 Rhodes Scholars, and one Schwarzman Scholar.[13]

Contents [hide]

Carnegie Mellon University

Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) [1] Carnegie Mellon University (1988–Present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie 1967 (merger with Mellon Institute)
Endowment	$1.709 billion (2016)[2]
President	Subra Suresh
Provost	Farnam Jahanian[3]
Academic staff	1,423[4]
Undergraduates	6,362
Postgraduates	7,141
Other students	145
Location	Pittsburgh, Pennsylvania, United States
Campus	Urban, 140 acres (57 ha)[5]
Colors	Cardinal, Black, Grey and White
Athletics	NCAA Division III UAA, ACHA, IRA 17 varsity teams[6]
Nickname	Tartans
Mascot	Scotty the Scottish Terrier[7]
Website	www.cmu.edu

Carnegie Mellon University

Institutional formation [edit]

REMEMBER FROM LAST LECTURE

INDEX1: DOCUMENT TITLES

INDEX2: DOCUMENT ABSTRACT

INDEX3: DOCUMENT HEADLINES

INDEX4: DOCUMENT BODY

Carnegie Mellon University

From Wikipedia, the free encyclopedia

Coordinates: 40.443322°N 79.943583°W

Carnegie Mellon University (**Carnegie Mellon** or **CMU**; /kɑːrnɪgi ˈmɛlən/ or /kɑːrˈneɪgi ˈmɛlən/) is a private research university in Pittsburgh, Pennsylvania.

Founded in 1900 by Andrew Carnegie as the Carnegie Technical Schools, the university became the Carnegie Institute of Technology in 1912 and began granting four-year degrees. In 1967, the Carnegie Institute of Technology merged with the Mellon Institute of Industrial Research to form Carnegie Mellon University.

The university's 140-acre (57 ha) main campus is 3 miles (4.8 km) from Downtown Pittsburgh. Carnegie Mellon has seven colleges and independent schools: the College of Engineering, College of Fine Arts, Dietrich College of Humanities and Social Sciences, Mellon College of Science, Tepper School of Business, H. John Heinz III College of Information Systems and Public Policy, and the School of Computer Science. The university also has campuses in Qatar and Silicon Valley, with degree-granting programs in six continents.

Carnegie Mellon consistently ranks in the top 25 in the national *U.S. News & World Report* rankings.[8] It is home to the world's first degree-granting Robotics and Drama programs,[9] as well as one of the first Computer Science departments.[10] The university spent $242 million on research in 2015.[11]

Carnegie Mellon counts 13,650 students from 114 countries, over 100,000 living alumni, and over 5,000 faculty and staff. Past and present faculty and alumni include 19 Nobel Prize Laureates, 19 Members of the American Academy of Arts & Sciences,[12] 72 Members of the National Academies, 114 Emmy Award Winners, 43 Tony Award laureates, 7 Academy Award Winners, 12 Turing Award winners, 4 Rhodes Scholars, and one Schwarzman Scholar.[13]

Contents [hide]

Carnegie Mellon University

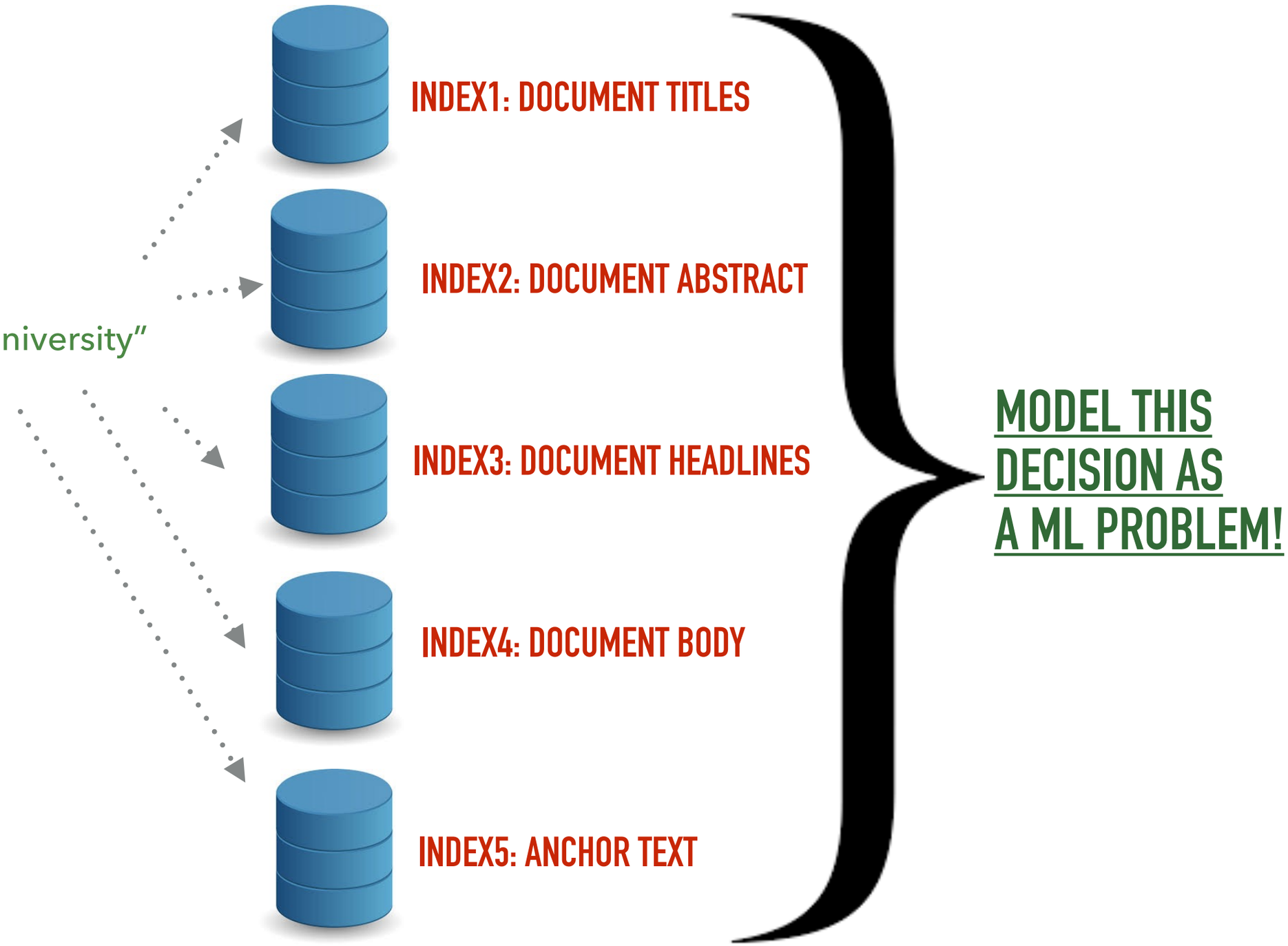
Former names	Carnegie Technical Schools (1900–1912) Carnegie Institute of Technology (1912–1967) Carnegie-Mellon University (1968–1988) [1] Carnegie Mellon University (1988–Present)
Motto	"My heart is in the work" (Andrew Carnegie)
Type	Private university
Established	1900 by Andrew Carnegie 1967 (merger with Mellon Institute)
Endowment	$1.709 billion (2016)[2]
President	Subra Suresh
Provost	Farnam Jahanian[3]
Academic staff	1,423[4]
Undergraduates	6,362
Postgraduates	7,141
Other students	145
Location	Pittsburgh, Pennsylvania, United States
Campus	Urban, 140 acres (57 ha)[5]
Colors	Cardinal, Black, Grey and White
Athletics	NCAA Division III UAA, ACHA, IRA 17 varsity teams[6]
Nickname	Tartans
Mascot	Scotty the Scottish Terrier[7]
Website	www.cmu.edu

Carnegie Mellon University

Institutional formation [edit]

REMEMBER FROM LAST LECTURE

Query: "computer science university"



PROBLEM SETTING

Query	DocID	VSM Score Title	VSM Score Abstract	VSM score Body	...	Page Rank	Relevance
“linux”	123	0.93	0.46	0.35	...	2.99	Yes
“linux"	987	0.45	0.33	0.81	...	1.32	No
“linux"	456	0.22	0.85	0.44	...	2.44	No
“qatar"	324	0.23	0.34	0.55	...	0.32	No
“qatar"	132	0.98	0.90	0.89	...	4.33	Yes
“qatar"	543	0.45	0.01	0.33	...	5.67	No

PROBLEM SETTING

Query	DocID	VSM Score Title	VSM Score Abstract	VSM score Body	...	Page Rank	Relevance
“linux”	123	0.93	0.46	0.35	...	2.99	Yes
“linux"	987	0.45	0.33	0.81	...	1.32	No
“linux"	456	0.22	0.85	0.44	...	2.44	No
“qatar"	324	0.23	0.34	0.55	...	0.32	No
“qatar"	132	0.98	0.90	0.89	...	4.33	Yes
“qatar"	543	0.45	0.01	0.33	...	5.67	No

$$y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_{400}X_{400}$$

PROBLEM SETTING

We can tell your ML model that some documents are actually much more important than others

Query	DocID	VSM Score Title	VSM Score Abstract	VSM score Body	...	Page Rank	Relevance
“linux”	123	0.93	0.46	0.35	...	2.99	1
“linux"	987	0.45	0.33	0.81	...	1.32	-1
“linux"	456	0.22	0.85	0.44	...	2.44	0
“qatar"	324	0.23	0.34	0.55	...	0.32	0
“qatar"	132	0.98	0.90	0.89	...	4.33	3
“qatar"	543	0.45	0.01	0.33	...	5.67	0

$$y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_{400}X_{400}$$

PROBLEM SETTING

$$y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_{400}X_{400}$$

- We use a learning method to learn the weights from past data:
  - High weight for w_1 means:
    - VSM score for titles is important...
  - High weight for w_2 means:
    - VSM score for abstract is important...
  - ...
    - ...
  - High weight for w_400 means:
    - High page rank values are important...

PROBLEM SETTING

$$y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_{400}X_{400}$$

- We use a learning method to learn the weights from past data:
  - High weight for w_1 means:
    - VSM score for titles is important...
  - High weight for w_2 means:
    - VSM score for abstract is important...
  - ...
    - ...
  - High weight for w_400 means:
    - High page rank values are important...

RESULT: $W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

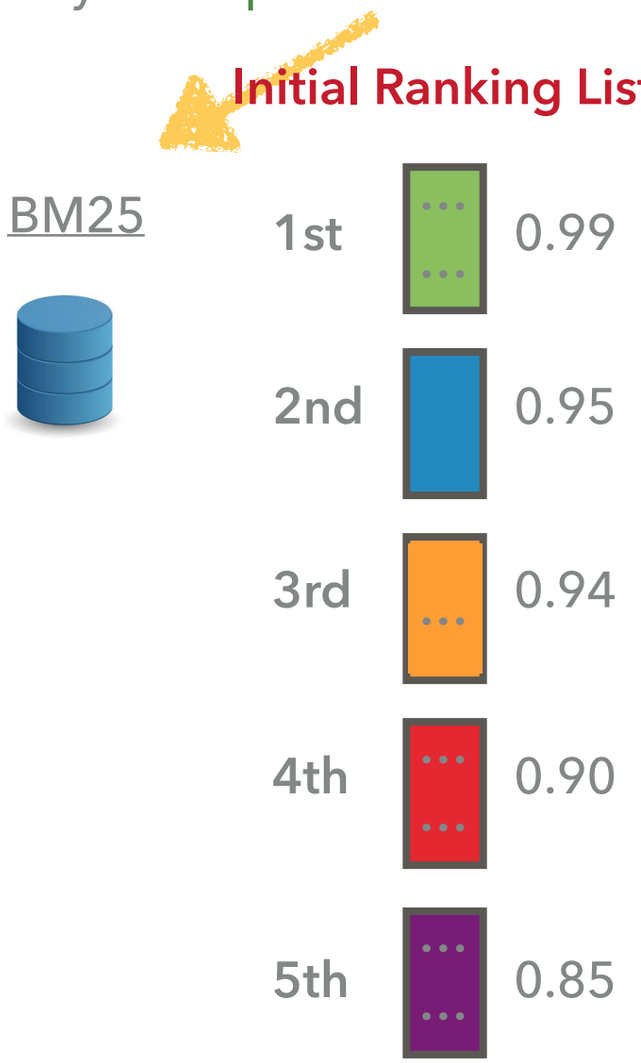
BM25

RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"



RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

Initial Ranking List

BM25

DOC TITLES	DOC ABST	DOC HEADLINES	DOC BODY	...	PAGE RANK
0.15	0.54	0.10	0.01		3.10
0.52	0.14	0.95	0.02		0.05
0.85	0.33	0.05	0.00		4.50
0.00	0.22	0.51	0.09		1.50
0.01	0.17	0.74	0.05		1.95

RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

Initial Ranking List

BM25

DOC TITLES	DOC ABST	DOC HEADLINES	DOC BODY	…	PAGE RANK
0.15	0.54	0.10	0.01		3.10
0.52	0.14	0.95	0.02		0.05
0.85	0.33	0.05	0.00		4.50
0.00	0.22	0.51	0.09		1.50
0.01	0.17	0.74	0.05		1.95

RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

Initial Ranking List

BM25

DOC TITLES	DOC ABST	DOC HEADLINES	DOC BODY	...	PAGE RANK	NEW SCORE
0.15	0.54	0.10	0.01		3.10	0.45
0.52	0.14	0.95	0.02		0.05	2.00
0.85	0.33	0.05	0.00		4.50	1.55
0.00	0.22	0.51	0.09		1.50	0.90
0.01	0.17	0.74	0.05		1.95	0.05

RANKING PROBLEM

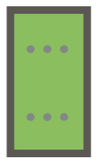
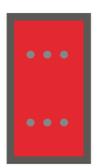
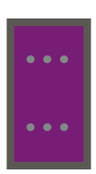
LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

Initial Ranking List

BM25

		DOC TITLES	DOC ABST	DOC HEADLINES	DOC BODY	...	PAGE RANK	NEW SCORE
4th		0.15	0.54	0.10	0.01		3.10	0.45
1st		0.52	0.14	0.95	0.02		0.05	2.00
2nd		0.85	0.33	0.05	0.00		4.50	1.55
3rd		0.00	0.22	0.51	0.09		1.50	0.90
5th		0.01	0.17	0.74	0.05		1.95	0.05

RANKING PROBLEM

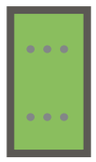
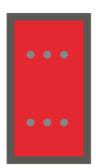
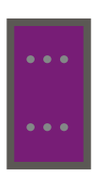
LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"

Initial Ranking List

BM25

		DOC TITLES	DOC ABST	DOC HEADLINES	DOC BODY	...	PAGE RANK	NEW SCORE
4th		0.15	0.54	0.10	0.01		3.10	0.45
1st		0.52	0.14	0.95	0.02		0.05	2.00
2nd		0.85	0.33	0.05	0.00		4.50	1.55
3rd		0.00	0.22	0.51	0.09		1.50	0.90
5th		0.01	0.17	0.74	0.05		1.95	0.05

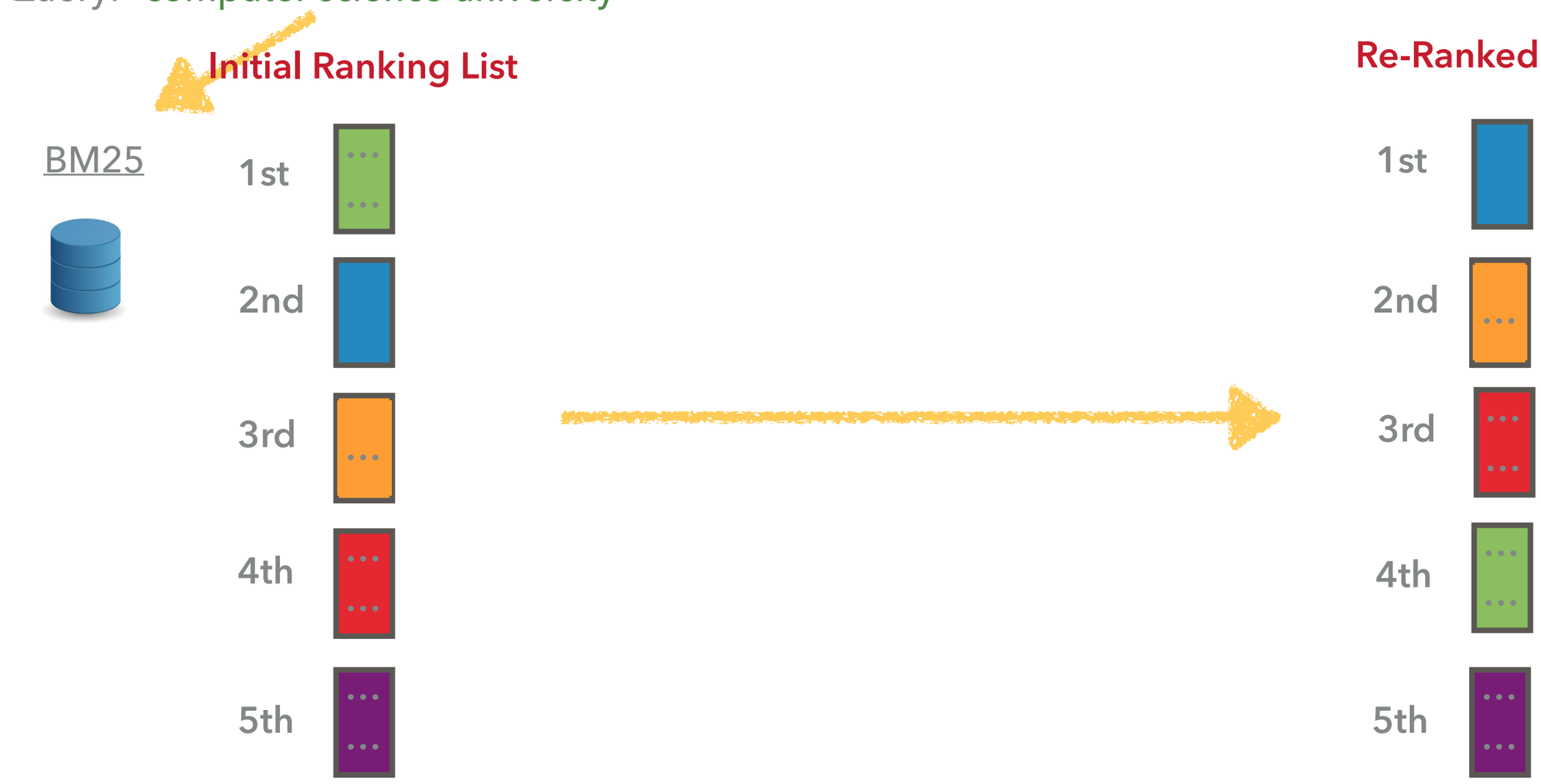
REGRESSION BASED LEARNING TO RANK

RANKING PROBLEM

LEARNT WEIGHTS

$W_1 = 0.45, W_2 = 0.99, W_3 = 0.11, \ldots, W_{400} = 1.34$

Query: "computer science university"



See: https://www.microsoft.com/en-us/research/project/mslr/

LEARNING TO RANK TYPES

- Pointwise:
  - Fit the relevance labels individually (just as we did)
- Pairwise:
  - Fit the relative orders ($d_i > d_j$)
- Listwise:
  - Fit the whole list

FEATURE ENGINEERING

- Features in LETOR dataset calculated for <query, document> pair

FEATURE ENGINEERING

- Features in LETOR dataset calculated for <query, document> pair
- Can we generate features for <query, document, user> ?

FEATURE ENGINEERING

- Features in LETOR dataset calculated for <query, document> pair
- Can we generate features for <query, document, user> ?

EXAMPLES?
PROBLEMS?

FEATURE ENGINEERING

- Features in LETOR dataset calculated for <query, document> pair
- Can we generate features for <query, document, user> ?
- Can we generate features for <query, document, group> ?

DISCUSSION

- What Machine Learning is?

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?

1. Various small tasks:

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?
    1. Various small tasks:
        - Decide whether a webpage is spam or not?
        - Decide how difficult to read a webpage is?
        - Decide what is the topic of a webpage?

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?

1. Various small tasks:

ALL THESE SUBTASKS CAN PROVIDE IMPORTANT FEATURES FOR RERANKING

   - Decide whether a webpage is spam or not?
   - Decide how difficult to read a webpage is?
   - Decide what is the topic of a webpage?

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?

1. Various small tasks:
2. Learn how to re-rank ranking lists

DISCUSSION

- What Machine Learning is?

Wikipedia: Machine learning is the subfield of computer science that, according to Arthur Samuel in 1959, gives "computers the ability to learn without being explicitly programmed."

- Why is it useful for search engines?
  1. Various small tasks:
  2. Learn how to re-rank ranking lists
     - More accurate results..
     - More personalized results...

TODAY'S LECTURE IN THE STANFORD IR BOOK

- Machine learning methods:
  - Chapter 13: Text classification for information retrieval
  - Chapter 14: Vector Space classification
  - Chapter 15: Support Vector Machines & machine learning on documents
- Chapter 15.4: Machine Learning methods in ad hoc information retrieval