Data Mining

INFS 4203/7203

Miao Xu

miao.xu@uq.edu.au

The University of Queensland, 2020 Semester 2

About the mid-term exam

- **Time**: Sep. 18th Friday, 14: 00
- **Duration**: 60 + 15 minutes
  - 60 minutes for exam, 15 minutes additionally for downloading/uploading/reading
- **Location**: online (mock exam has been provided)
- **Covered**: all lectures and tutorials from Week 1 to Week 6

  (No use of Jupyter Notebook)
- **Form:** Closed-book exam, no reference/discussion/copy to/with/from anyone/Internet/…
- **Questions**: calculation, problem solving, short answering…

CHAT policy

Thank you very much for providing feedback.

- Q&A times will be provided after **finishing** each small topic.
- Please do **not** put your question in CHAT **publicly** until explicitly asked.
- Immediate question can go **privately** to our tutor.
  - Sivangi this week
- I may still ask for short answers, but please stop when given the signal.

Last week

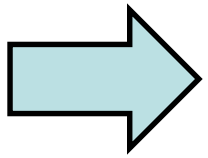
- **Prototype-based clustering** (*k*-means)
  - Easy and simple
  - Need to specify *k*, cannot deal with non-globular shape
- **Density-based clustering** (DBSCAN)
  - Can deal with non-globular shape
  - Need to specify MinPts and Eps, density-sensitive
- **Hierarchical-based clustering** (AGNES)
  - Hierarchical structure, deal with non-globular shape
  - Select within Min, Max, Average, each has its own disadvantage

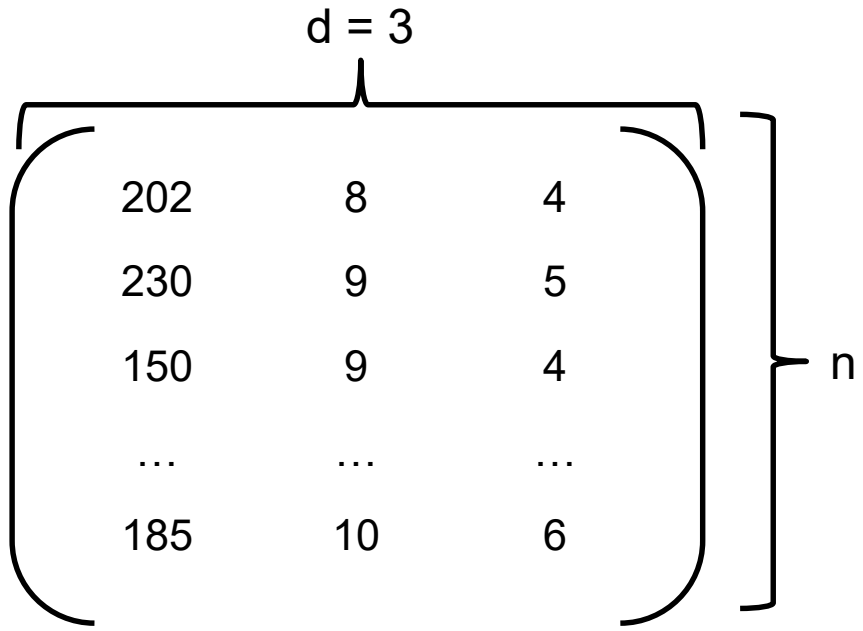
More on data

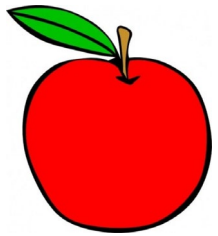
Object	Weight/g	Sweetness	Price/kg
Apple 1	202	8	4
Apple 2	230	9	5
Apple 3	150	9	4
…	…	…	…
Apple n	185	10	6

- **Data**
  - Organized into a table
  - Or a matrix
- **Each object**
  - A vector, if represented numerically
  - For d-dimensional vector, it is located in the d-dimensional space
  - Also called “data point” or simply “point”
- **How two objects are “close”**:
  - By distance

Object	Weight/g	Sweetness	Price/kg
Apple 1	202	8	4
Apple 2	230	9	5
Apple 3	150	9	4
…	…	…	…
Apple n	185	10	6

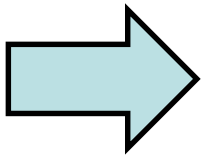


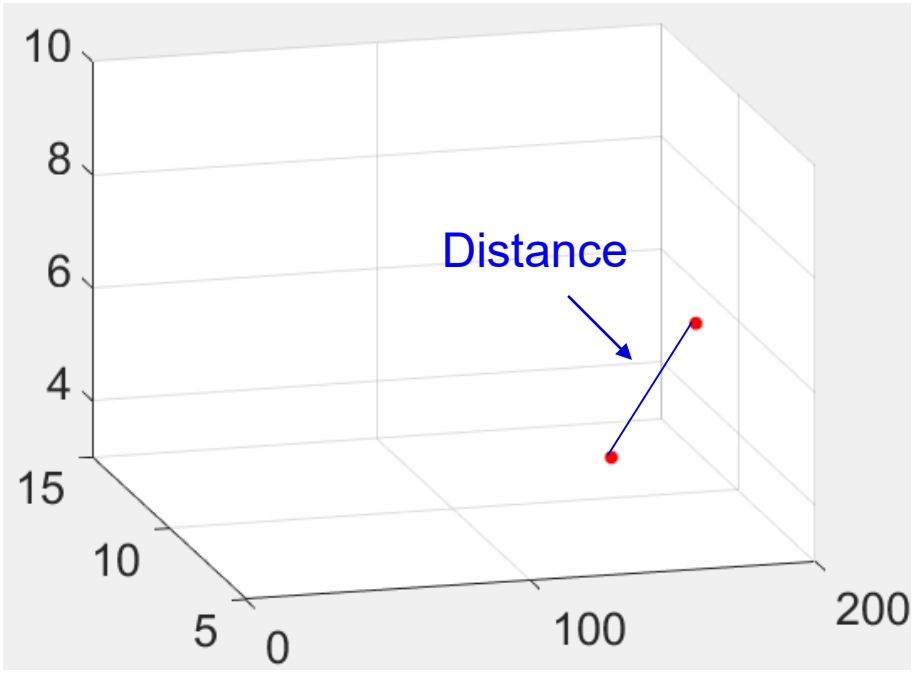




A three-dimensional vector:

[185 10 6]

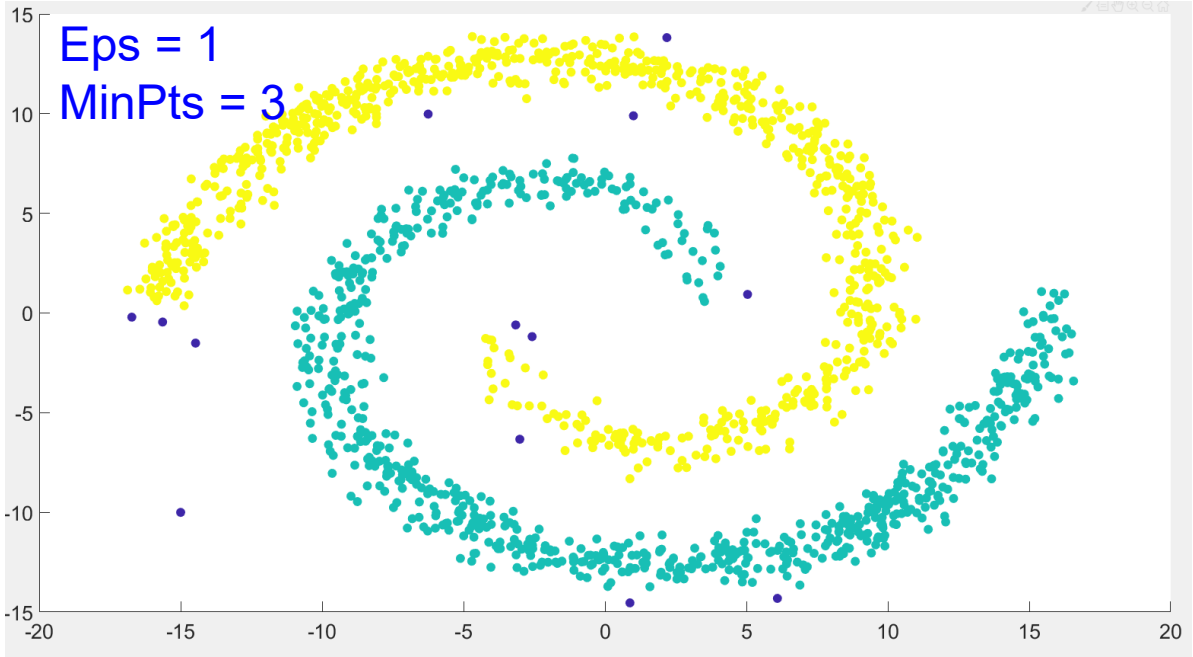




A three-dimensional vector:

[150 9 4]

- Clustering: find the majority patterns, ignore the points outside of any cluster (generally called “anomaly” or “outlier”)
- Anomaly detection: find **exceptional** cases outside of the majority pattern



Lecture 5: Anomaly Detection

Anomaly detection/outlier detection

- Anomaly:

Dictionary

Search for a

anomaly

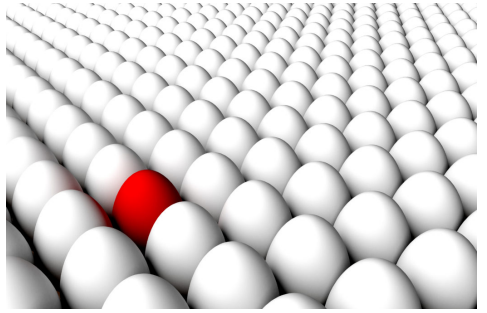
/əˈnɒm(ə)li/

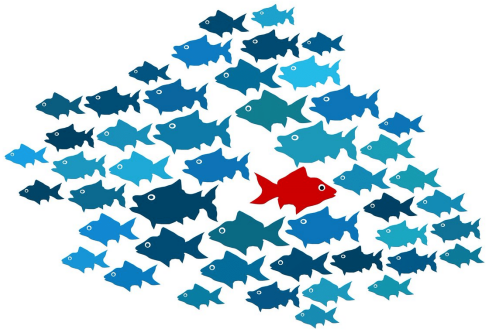
noun

1. something that deviates from what is standard, normal, or expected.
   "there are a number of anomalies in the present system"

Similar: oddity peculiarity abnormality irregularity inconsistency

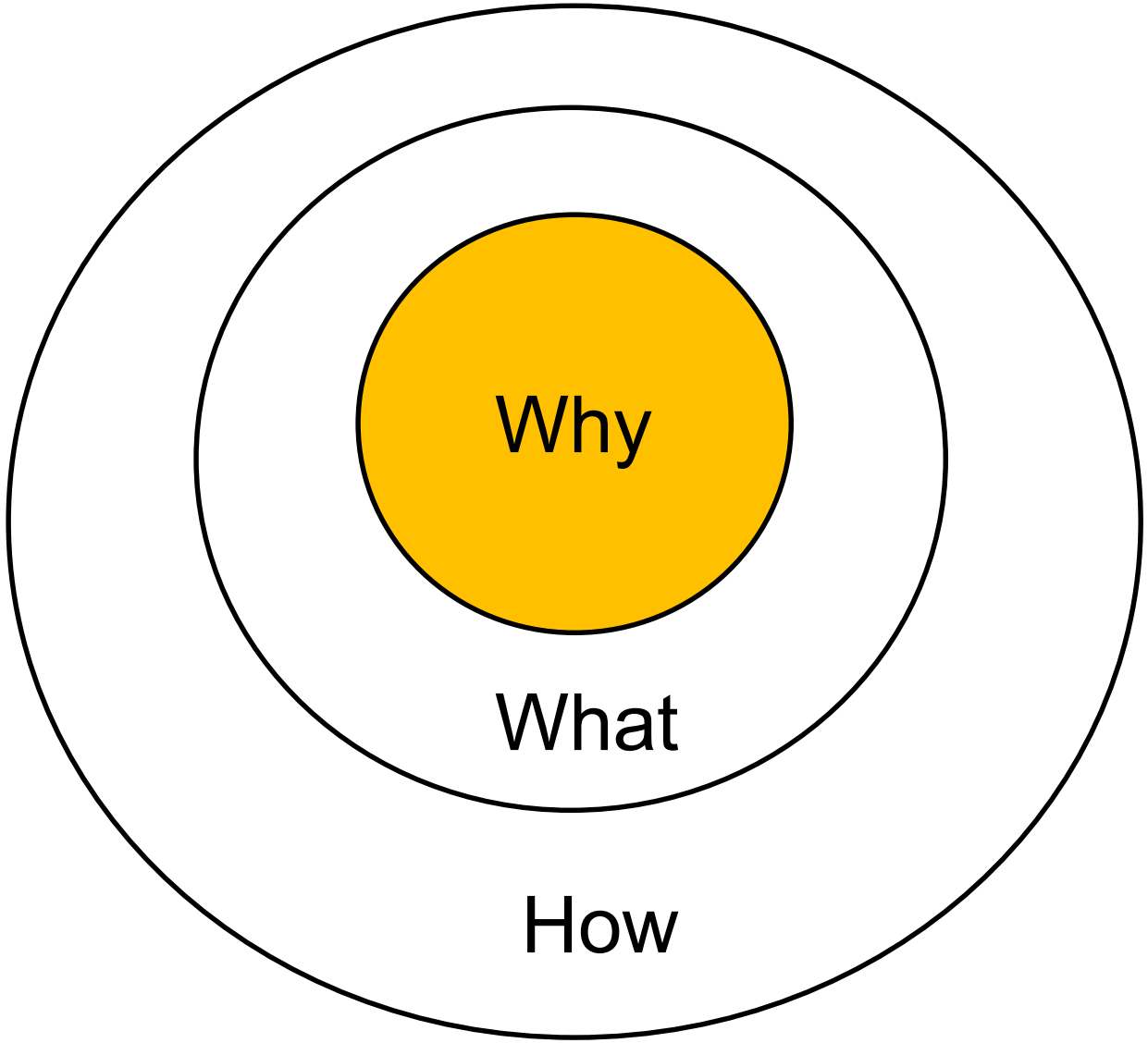
Something that deviates from what is *standard*, *normal* or *expected*.





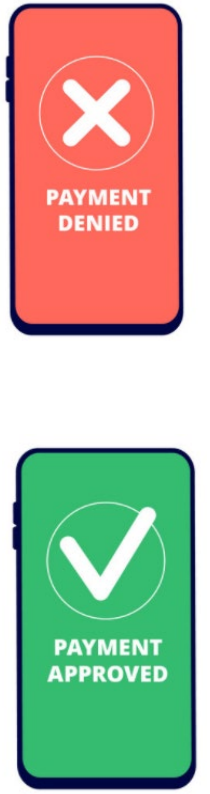
- Also called "outlier"
  - In statistics, an **outlier** is a data point that differs significantly from other observations.--Wikipedia

Will anomaly detection be useful?


Why
What
How

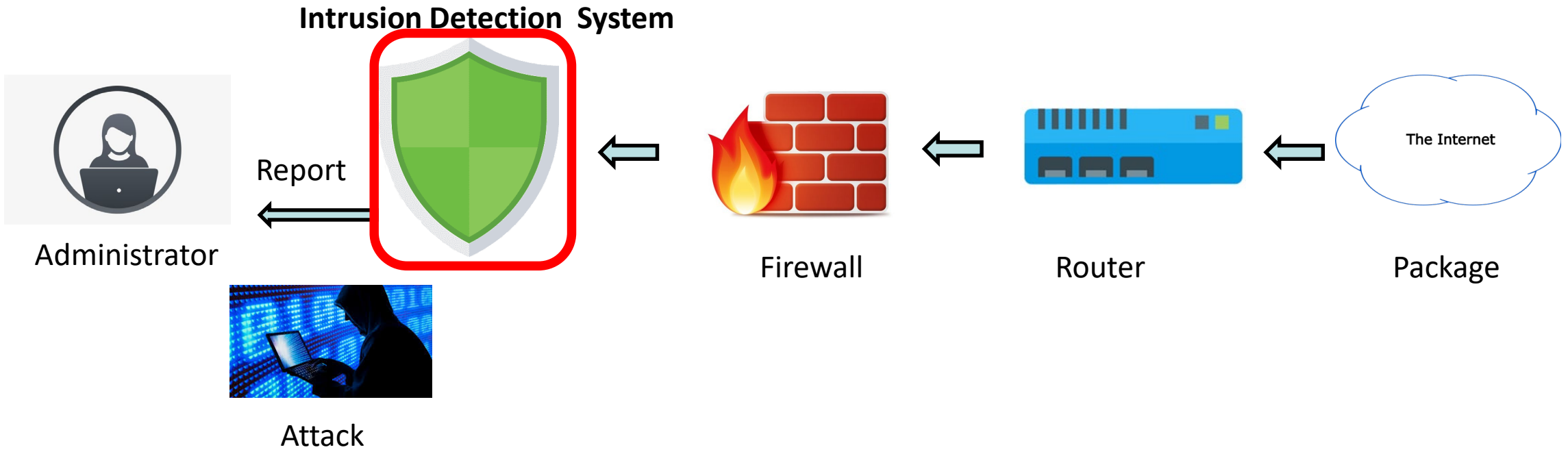
Credit card fraud detection

- **Anomaly:** <u>different</u> buying/shopping patterns/behaviors of card owner



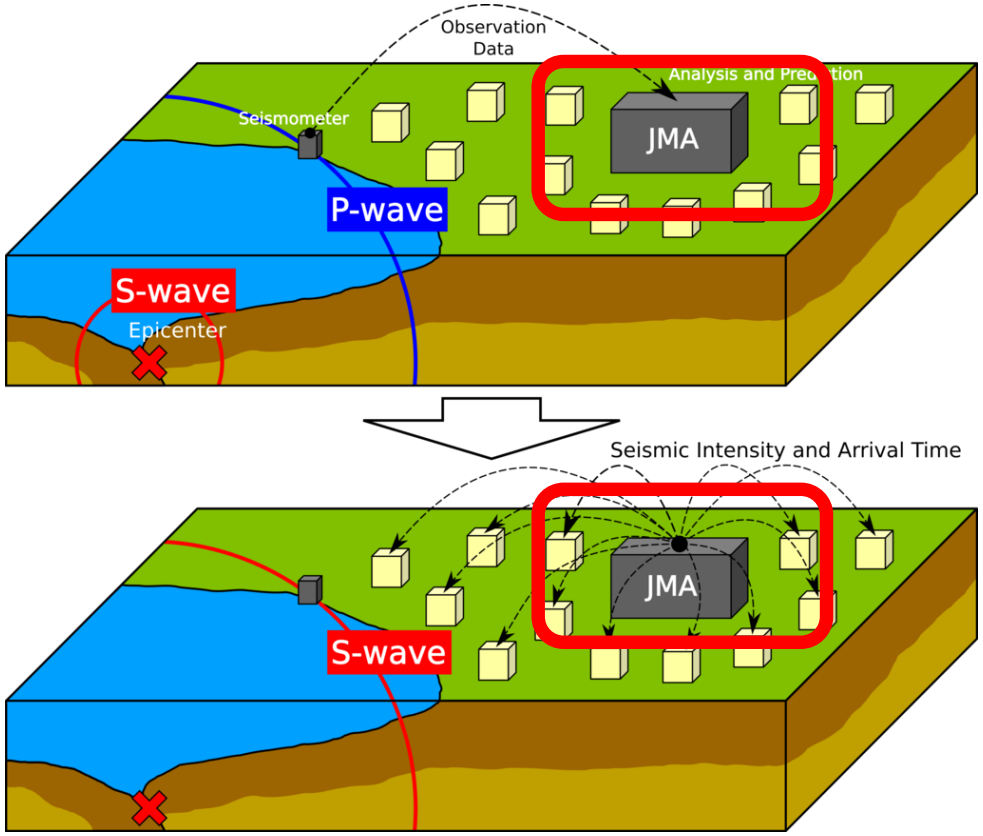
Intrusion detection

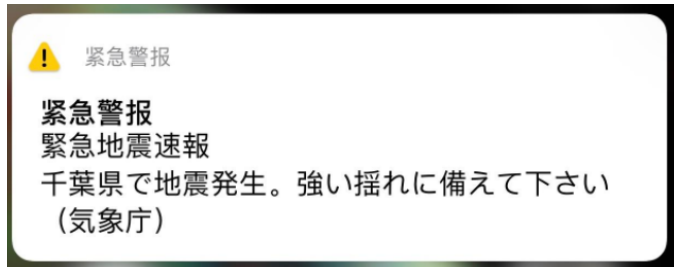
- **Anomaly:** Unusual behaviors that can be counted as attack to computer systems



Disaster prediction

- **Anomaly:** Extreme geophysical/ hydrological/ climatological/ meteorological conditions, e.g., hurricane, flood, drought, heat waves, fires, earthquake…

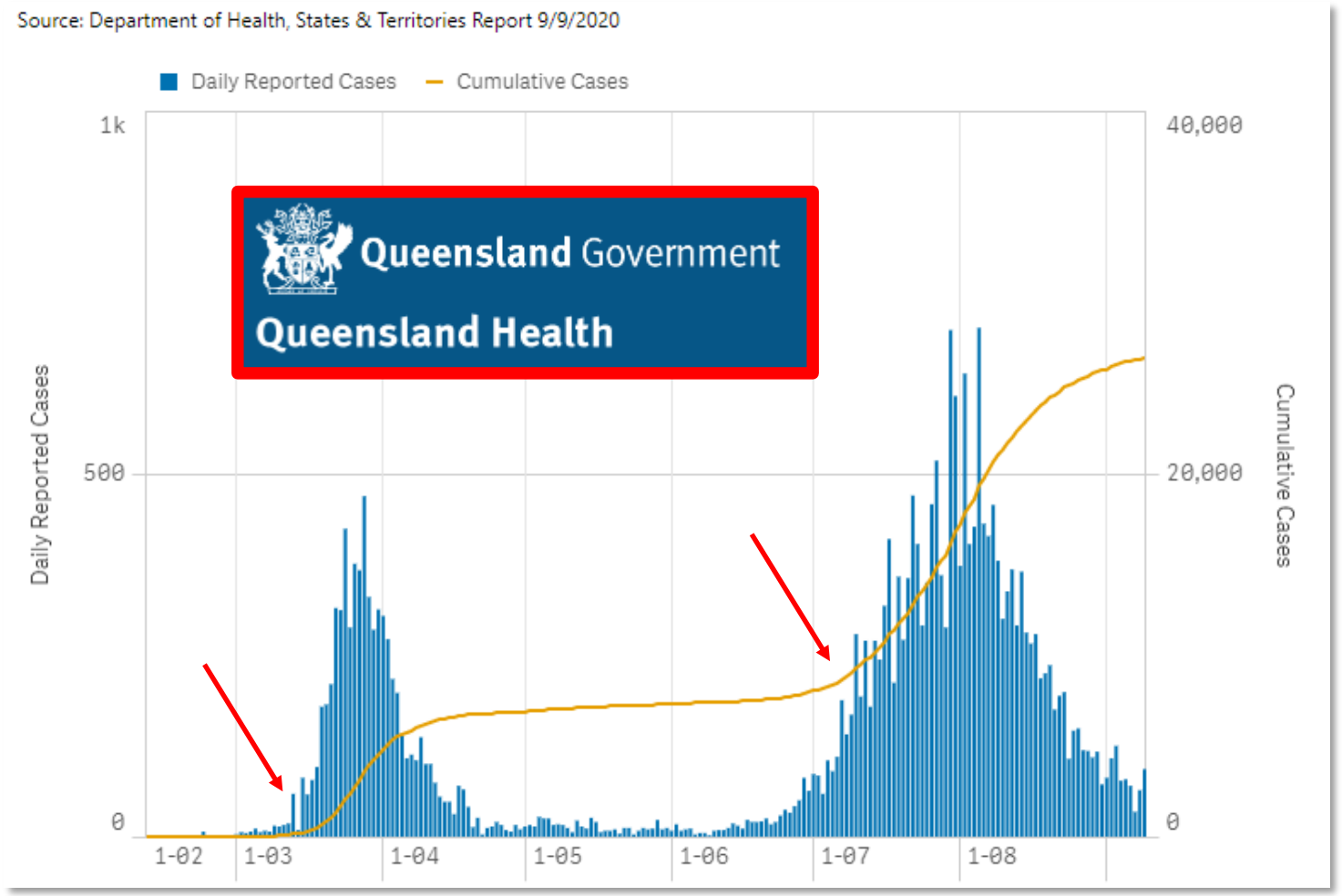




⚠ **Earthquake Early Warning**
An earthquake has occurred in Chiba Prefecture. Please prepare for a strong tremor (Japan Meteorological Agency)

Public health

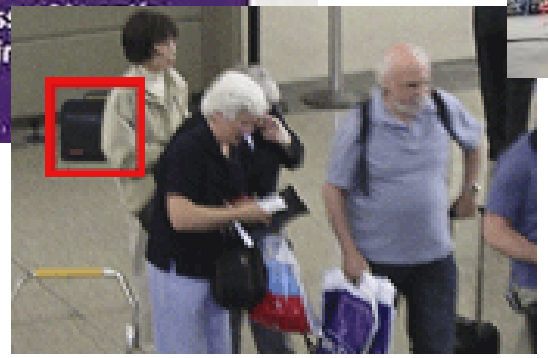
- **Anomaly:** Highly increased statistics of a particular disease



Daily and cumulative number of reported COVID-19 cases in Australia

Others

- Healthcare: unusual symptoms for a particular patient
- Public security: surveillance for potential terrorist attacks
- Industrial damage detection: faults and failures in complex industrial systems



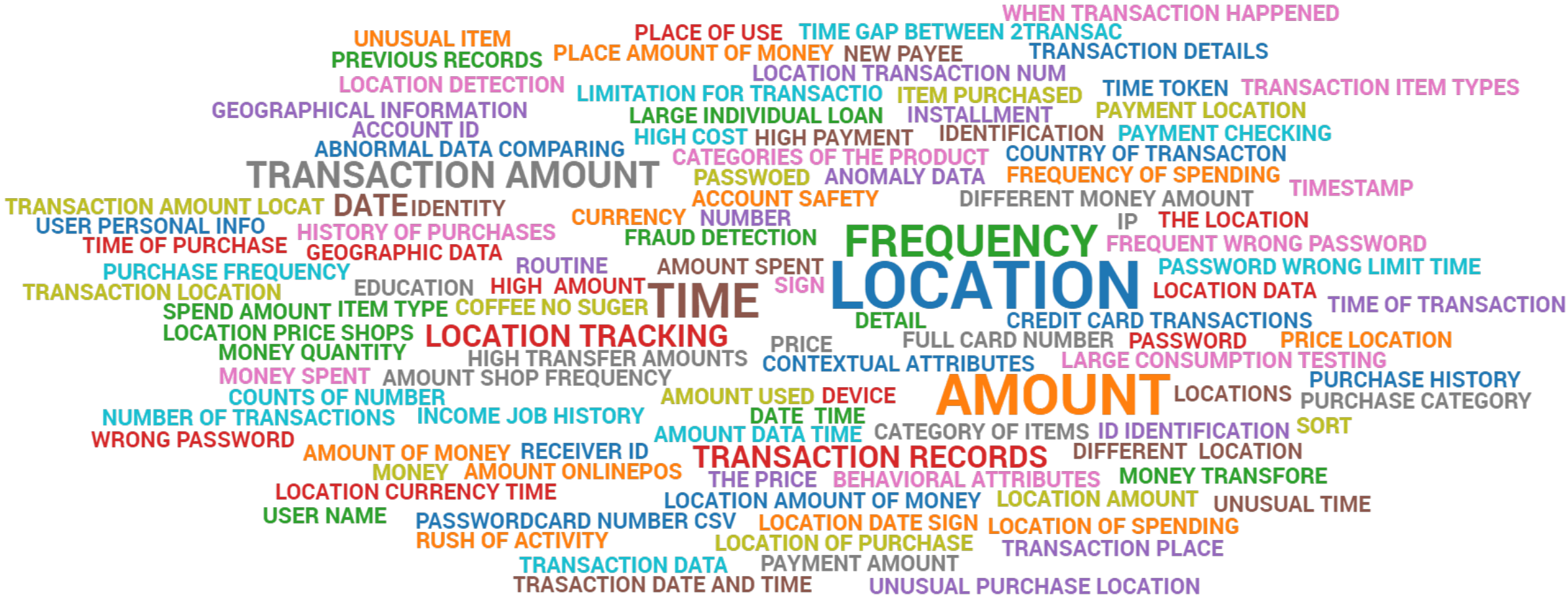
More

- Data pre-processing
- Novelty discovery (sometimes called “novelty detection”)
- …

Activity: case study

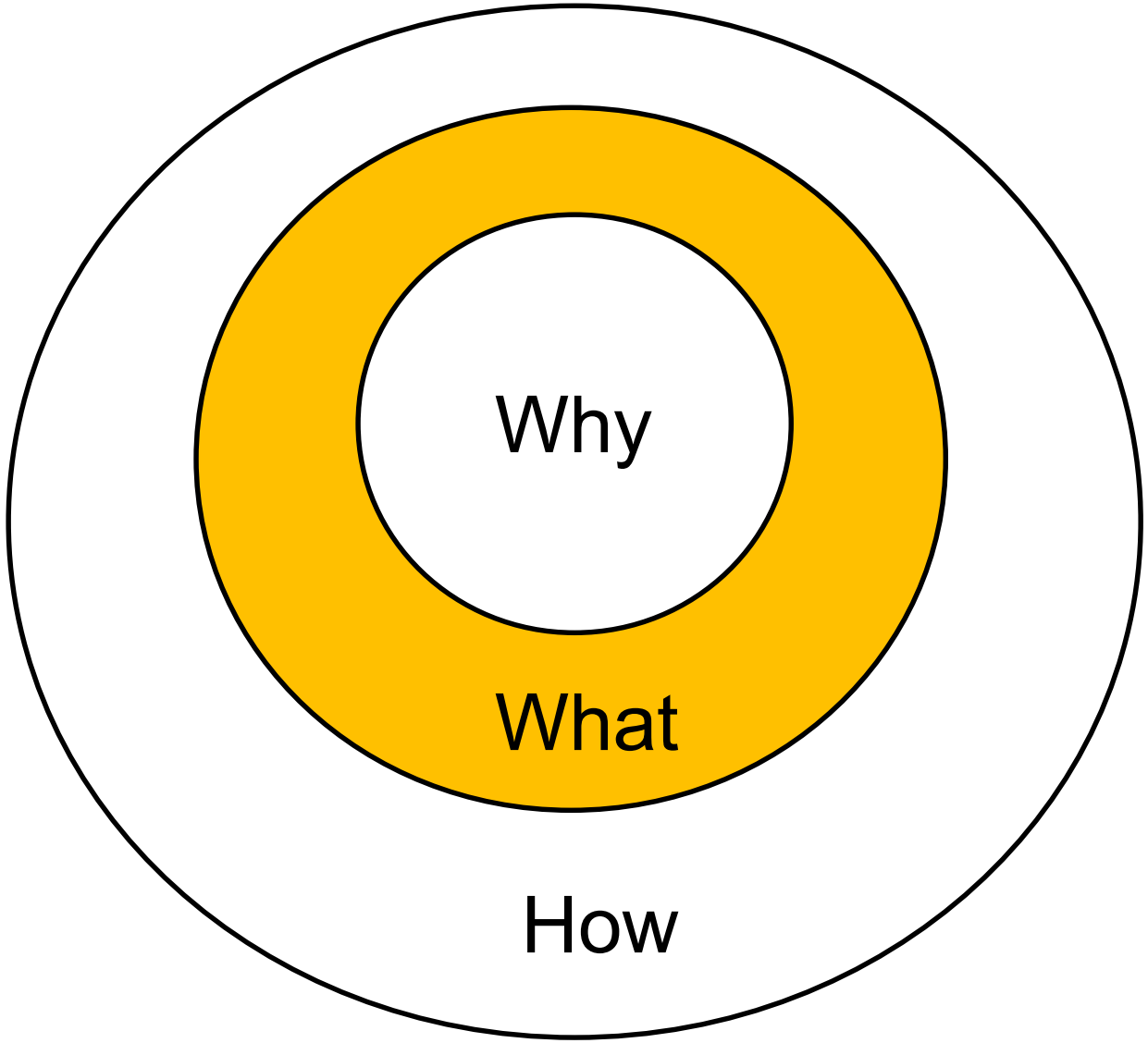
- In order to do credit card fraud detection, what kind of data should we use?
- Alternatively, what kind of features/attributes should we care about?

https://apps.elearning.uq.edu.au/wordcloud/75753


WHEN TRANSACTION HAPPENED
UNUSUAL ITEM
PLACE OF USE
TIME GAP BETWEEN 2TRANSAC
PREVIOUS RECORDS
PLACE AMOUNT OF MONEY
NEW PAYEE
TRANSACTION DETAILS
LOCATION TRANSACTION NUM
LOCATION DETECTION
LIMITATION FOR TRANSACTIO
ITEM PURCHASED
TIME TOKEN
TRANSACTION ITEM TYPES
GEOGRAPHICAL INFORMATION
LARGE INDIVIDUAL LOAN
INSTALLMENT
PAYMENT LOCATION
ACCOUNT ID
HIGH COST
HIGH PAYMENT
IDENTIFICATION
PAYMENT CHECKING
ABNORMAL DATA COMPARING
CATEGORIES OF THE PRODUCT
COUNTRY OF TRANSACTON
TRANSACTION AMOUNT
PASSWOED
ANOMALY DATA
FREQUENCY OF SPENDING
TIMESTAMP
TRANSACTION AMOUNT LOCAT
DATE
IDENTITY
ACCOUNT SAFETY
DIFFERENT MONEY AMOUNT
CURRENCY
NUMBER
IP
THE LOCATION
USER PERSONAL INFO
HISTORY OF PURCHASES
FRAUD DETECTION
FREQUENCY
TIME OF PURCHASE
GEOGRAPHIC DATA
FREQUENT WRONG PASSWORD
ROUTINE
AMOUNT SPENT
LOCATION
PASSWORD WRONG LIMIT TIME
PURCHASE FREQUENCY
EDUCATION
HIGH AMOUNT
TIME
SIGN
LOCATION DATA
TRANSACTION LOCATION
SPEND AMOUNT
ITEM TYPE
COFFEE NO SUGER
TIME OF TRANSACTION
DETAIL
CREDIT CARD TRANSACTIONS
LOCATION PRICE SHOPS
LOCATION TRACKING
PRICE
FULL CARD NUMBER
PASSWORD
PRICE LOCATION
MONEY QUANTITY
HIGH TRANSFER AMOUNTS
CONTEXTUAL ATTRIBUTES
LARGE CONSUMPTION TESTING
MONEY SPENT
AMOUNT SHOP FREQUENCY
AMOUNT
PURCHASE HISTORY
COUNTS OF NUMBER
AMOUNT USED
DEVICE
LOCATIONS
PURCHASE CATEGORY
NUMBER OF TRANSACTIONS
INCOME JOB HISTORY
DATE TIME
WRONG PASSWORD
AMOUNT DATA TIME
CATEGORY OF ITEMS
ID IDENTIFICATION
SORT
AMOUNT OF MONEY
RECEIVER ID
TRANSACTION RECORDS
DIFFERENT LOCATION
MONEY
AMOUNT ONLINEPOS
THE PRICE
BEHAVIORAL ATTRIBUTES
MONEY TRANSFORE
LOCATION CURRENCY TIME
LOCATION AMOUNT OF MONEY
LOCATION AMOUNT
UNUSUAL TIME
USER NAME
PASSWORDCARD NUMBER CSV
LOCATION DATE SIGN
LOCATION OF SPENDING
RUSH OF ACTIVITY
LOCATION OF PURCHASE
TRANSACTION PLACE
TRANSACTION DATA
PAYMENT AMOUNT
TRASACTION DATE AND TIME
UNUSUAL PURCHASE LOCATION

Which is more likely to be an anomaly?

	Date	Time	Location	Store	Amount	Weekday	Frequency	Buy what
Transaction1	10/09	13:02	St. Lucia	Subway	11.25	Y	24H	sandwich
Transaction2	11/09	5:02	Dubai	luxury mall	22,320	Y	10min	luxury brand
Transaction3	11/09	12:58	St. Lucia	Subway	11.25	Y	24H	sandwich


Why
What
How

Basic idea of anomaly detection

- Anomaly: something that deviates from what is standard, normal or expected

Input

Anomaly Detection Algorithm

Output

An **anomaly score** for each data point

Ways to determine "anomaly score" will be discussed later in the "HOW" part.

Basic tasks of anomaly detection

- Given a data set *D*, containing mostly normal (but unlabeled) data points, and a test point ***x***, compute the **anomaly score** of ***x*** with respect to *D*

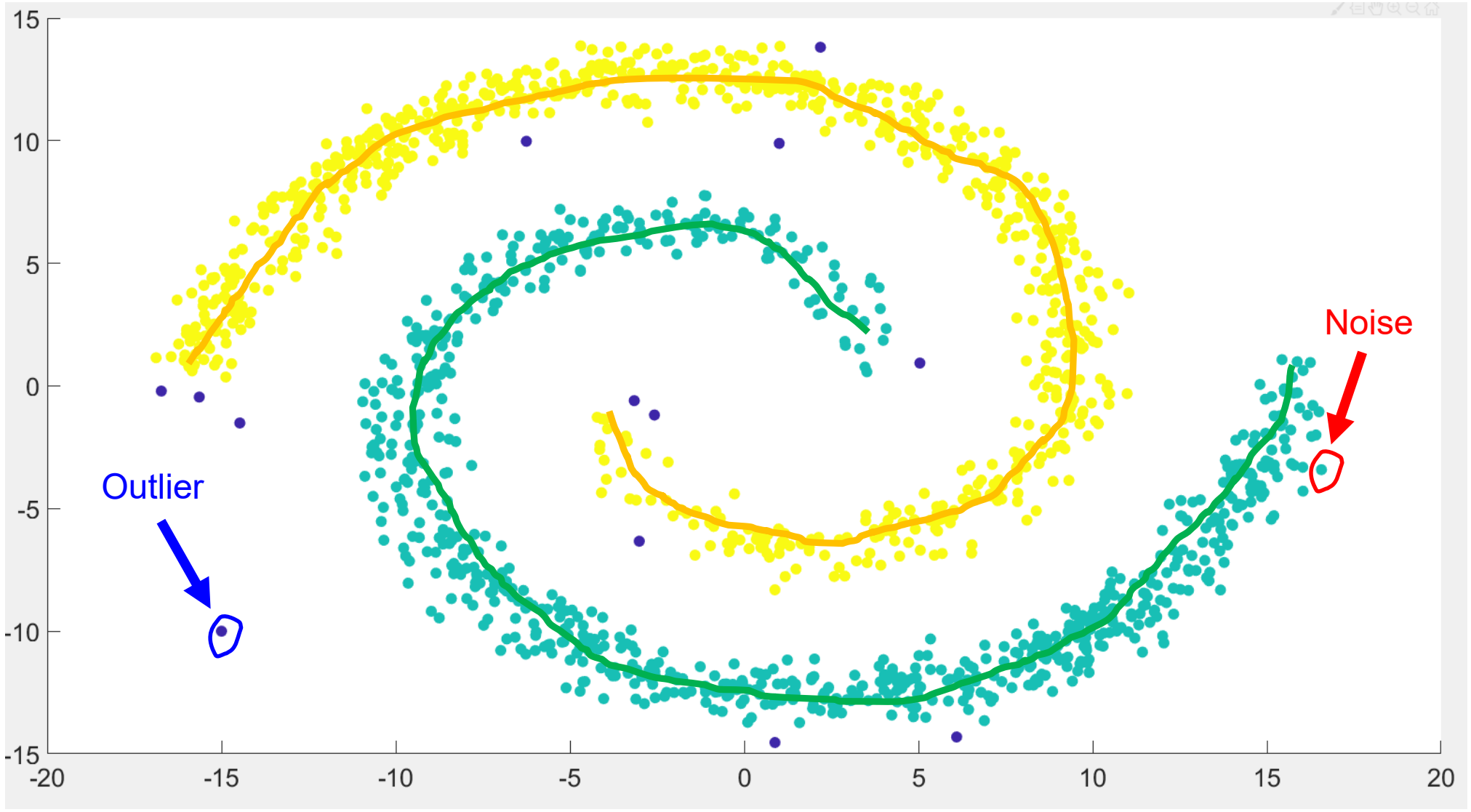
- Given a data set *D*, find all data points $\boldsymbol{x} \in D$ with anomaly scores greater/smaller than some threshold *t*

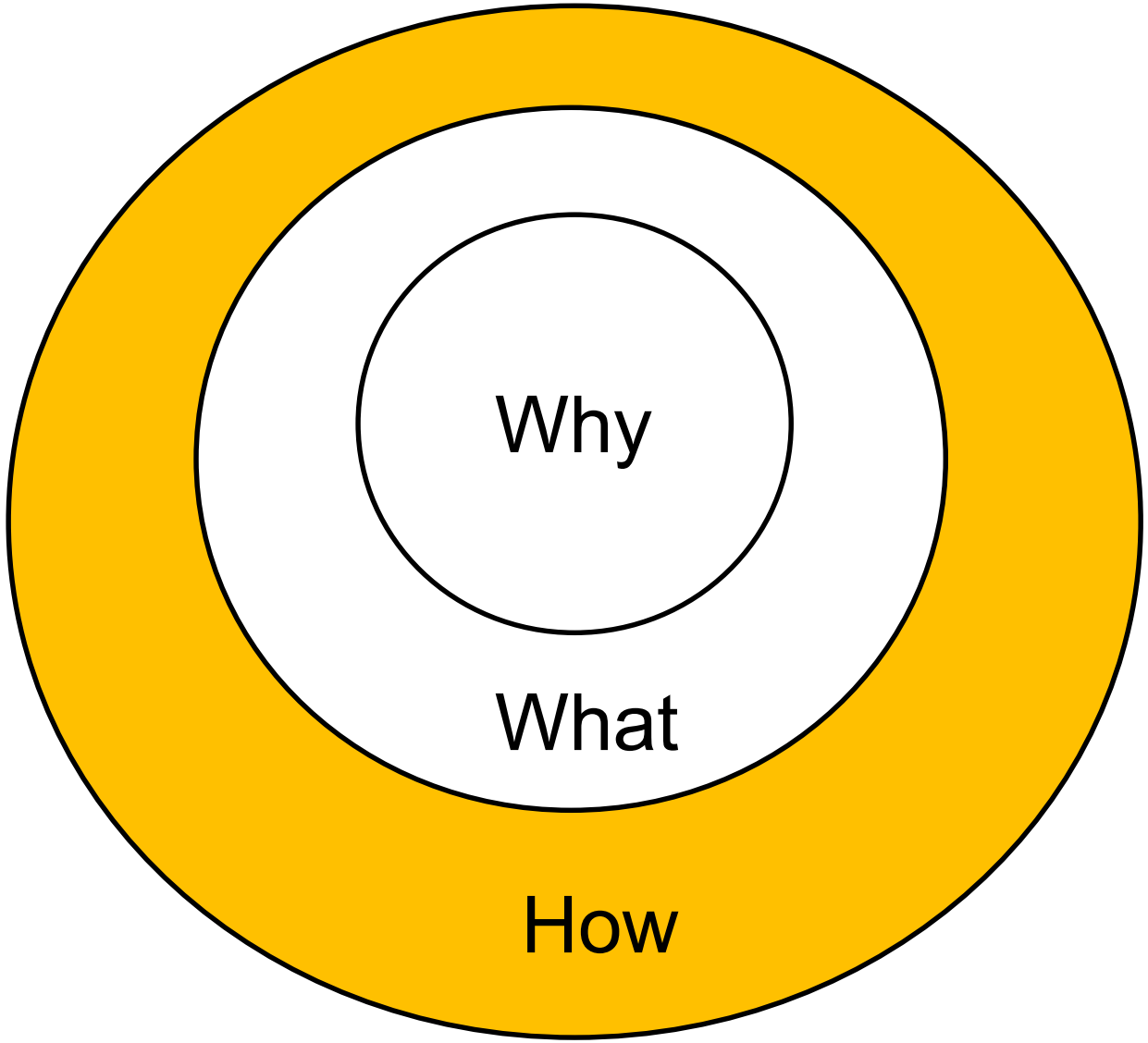
- Given a data set D, find all data points $\boldsymbol{x} \in$ D having the top-*n* largest/smalleset anomaly scores

Outlier vs. noise

- Noise
  - erroneous, perhaps random perturbation or contaminating
  - not necessarily produce unusual values or objects
  - may not be interesting (not that novel)
- Anomalies:
  - may be caused by adversaries
  - unusual values
  - Interesting (something novel)

They are related but distinct concepts!


15
10
5
0
-5
-10
-15
-20
-15
-10
-5
0
5
10
15
20
Outlier
Noise


Why
What
How

Methods

- Model-based technique
- Distance-based technique
- Density-based technique
- Cluster-based technique

Methods

- Model-based technique
- Distance-based technique
- Density-based technique
- Cluster-based technique

Model-based technique

- A *model* is created for the data, and objects are evaluated with respect to how well they *fit* the model.

- The better they fit the model
  - The less chance this point is an anomaly

Model-based approach: statistical method

- Statistical model: a probability distribution model with parameters (parametric), which are used to generate the data
- Outlier: an object has low probability with respect to a statistical model of the data
- Anomaly score: probability

We need the domain knowledge to determine what is the appropriate statistical model and how low a probability can be counted as an outlier.

Example of statistical approaches: univariate

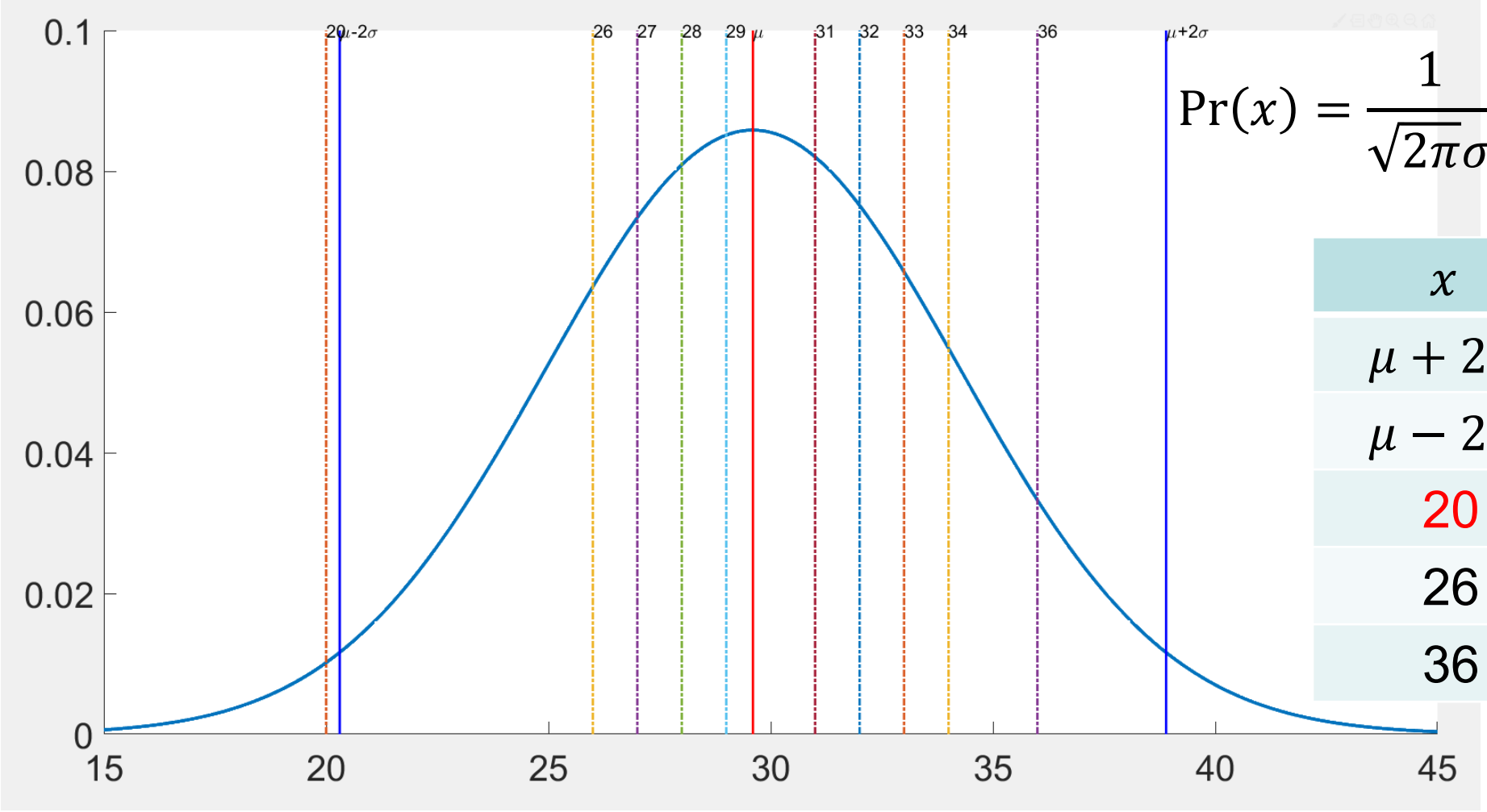
- Data:
  - 10 temperature data points collected in the summer
  - {20, 26, 27, 28, 29, 31, 32, 33, 34, 36}
- Statistical model: Gaussian model
  - μ: mean (29.6): $\mu = \mathrm{E}[x] = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - σ^2: variance (21.6): $\sigma^2 = \mathrm{E}[(x - \mathrm{E}[x])^2] = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2$
  - Gaussian distribution $x \sim N(\mu, \sigma^2)$: $\Pr(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
- Outlier: $\Pr(x) \leq \alpha = \Pr(\mu \pm 2\sigma)$

E[.] calculates the expectation.

Which data point is an outlier?

Calculate the probability

{20, 26, 27, 28, 29, 31, 32, 33, 34, 36}



$$\Pr(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

x	$\Pr(x)$
$\mu + 2\sigma$	0.0116
$\mu - 2\sigma$	0.0116
20	0.0102
26	0.0636
36	0.0333

Alternative to calculate the probability

- Measure how far away the point is from μ by σ

{20, 26, 27, 28, 29, 31, 32, 33, 34, 36}

$\mu = 29.6$
$\sigma = \sqrt{21.6}$

Outside of the range of $(\mu \pm 2\sigma)$

	$(x - \mu)/\sigma$
20	2.1
26	0.8
27	0.6
28	0.3
29	0.1
31	0.3
32	0.5
33	0.7
34	0.9
36	1.3

Example of statistical approaches: multivariate

- Data:
  - 10 (temperature, humidity*100) data collected: {(20, 31), (26, 40), (27, 45), (28, 52), (29, 60), (31, 70), (32, 71), (33, 69), (34, 72), (36, 85)}
- Statistical model
  - $\boldsymbol{\mu}$ mean: $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{x}] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x_i} = (\mathbf{29.6}, \mathbf{59.5})$
  - S covariance: $\begin{pmatrix} 21.6 & 76.2 \\ 76.2 & 288.72 \end{pmatrix}$
  - Gaussian distribution $x \sim N(\mu, \Sigma)$:

$$\Pr(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathrm{S}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

- Outlier: $\Pr(\boldsymbol{x}) \leq \alpha = 1 \times 10^{-3}$

More on the covariance matrix

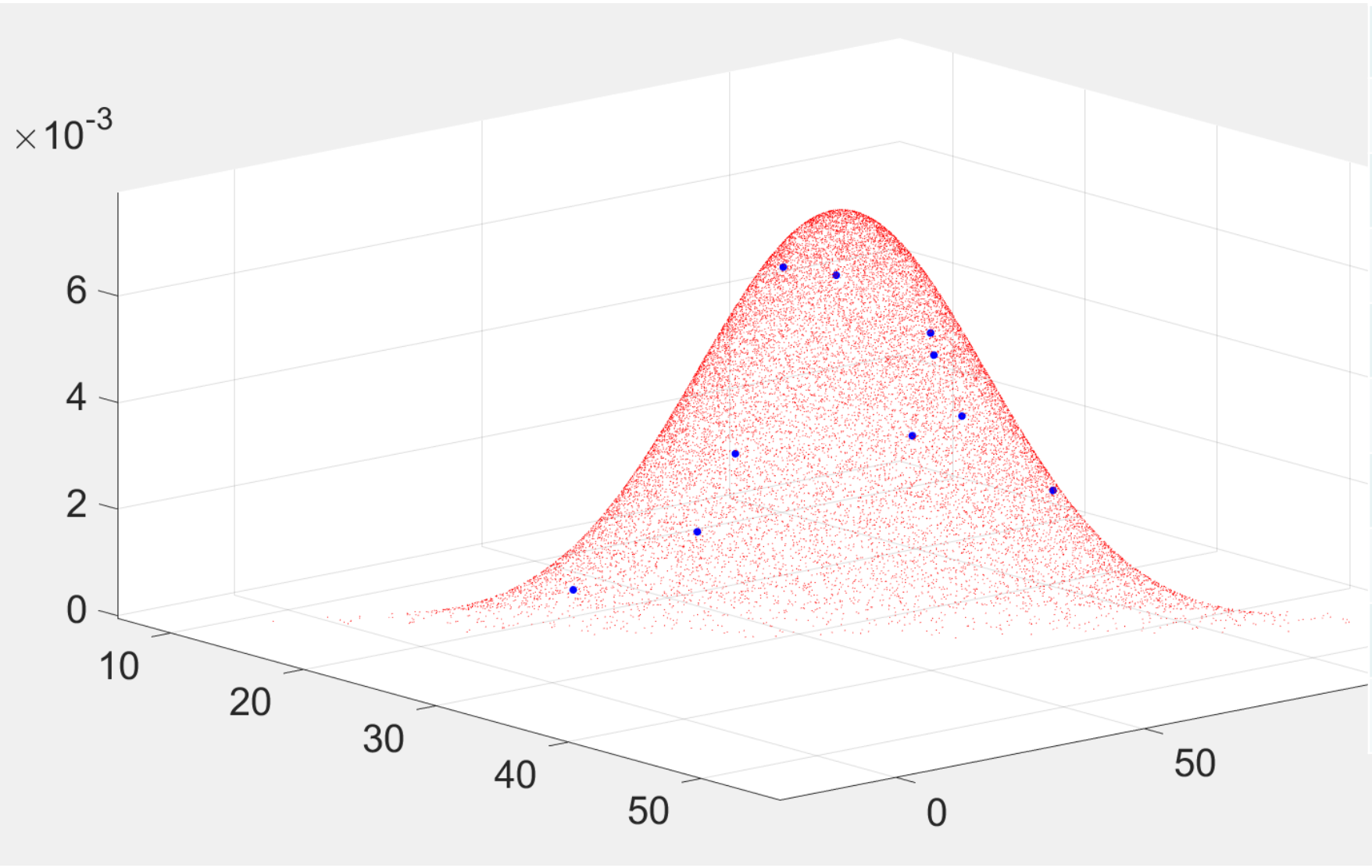
- {(20, 31), (26, 40), (27, 45), (28, 52), (29, 60), (31, 70), (32, 71), (33, 69), (34, 72), (36, 85)}
- Each point denoted as (x_1, x_2)
- S = $\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}$

$$S_{11} = \mathrm{E}[(x_1 - \mathrm{E}[x_1])(x_1 - \mathrm{E}[x_1])]$$

$$S_{12} = \mathrm{E}[(x_1 - \mathrm{E}[x_1])(x_2 - \mathrm{E}[x_2])]$$

$$S_{21} = \mathrm{E}[(x_2 - \mathrm{E}[x_2])(x_1 - \mathrm{E}[x_1])]$$

$$S_{22} = \mathrm{E}[(x_2 - \mathrm{E}[x_2])(x_2 - \mathrm{E}[x_2])]$$



X	Pr (x)
(20, 31)	0.0004
(26, 40)	0.0018
(27, 45)	0.0032
(28, 52)	0.0067
(29, 60)	0.0064
(31, 70)	0.0034
(32, 71)	0.0053
(33, 69)	0.0050
(34, 72)	0.0039
(36, 85)	0.0024

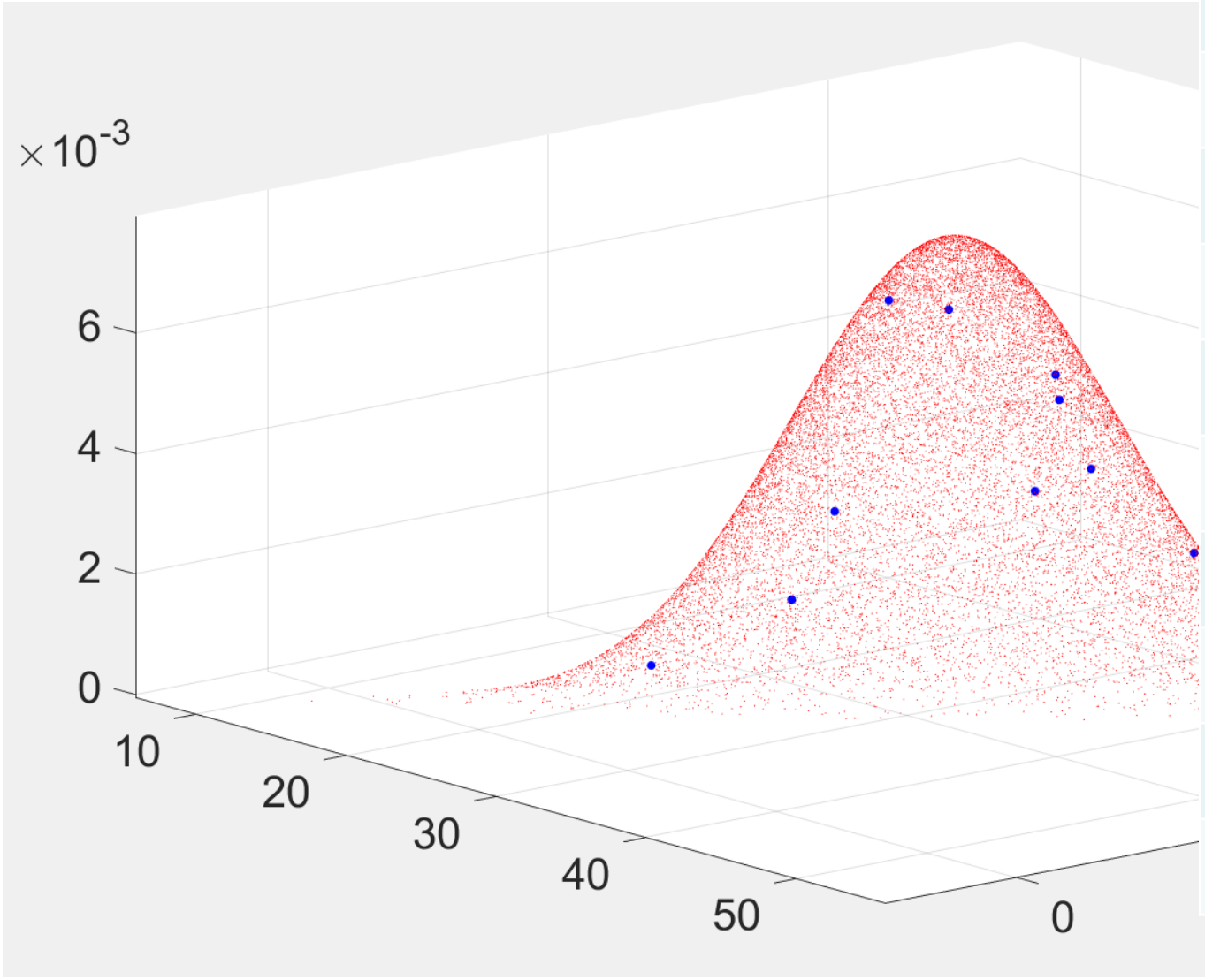
Alternative to calculate the probability

$$\Pr(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\mathrm{S}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}$$

- What really do calculation with $\boldsymbol{x}$? $(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$
- Also called "mahalanobis distance"

  (Q: is it a "metric" (Week 4) ? Please discuss on Piazza.)
- If the mahalanobis distance to "mean" is large (larger than a threshold): outlier
- Anomaly score: mahalanobis distance

More on Mahalanobis distance will be discussed next week.



X	Mahalanobis distance
(20, 31)	5.7304
(26, 40)	2.9389
(27, 45)	1.7488
(28, 52)	0.2926
(29, 60)	0.3635
(31, 70)	1.6559
(32, 71)	0.7318
(33, 69)	0.8511
(34, 72)	1.3602
(36, 85)	2.3268

Strength and weakness

- Strength
  - If you have strong knowledge about the distribution of data

- Weakness
  - Determine the distribution
  - Curse-of-dimensionality

x	$\Pr(x)$
20	0.0102
26	0.0636
36	0.0333

$\boldsymbol{x}$	$\Pr(\boldsymbol{x})$
(20, 31)	0.0004
(26, 40)	0.0018
(36, 85)	0.0024

Activity: Zoom Poll

- Anomaly detection by statistical method for

$$\{0, 12, 15, 27, 46\}$$

- Using the univariate gaussian distribution

- If $\Pr(x) \leq \alpha = \Pr(\mu \pm \sigma)$ is counted as "anomaly", which points are anomaly points?

Answer

- $\{0, 12, 15, 27, 46\}$
- Mean: $(0 + 12 + 15 + 27 + 46)/5 = 20$
- Variance $((0 - 20)^2 + (12 - 20)^2 + (15 - 20)^2 + (27 - 20)^2 + (46 - 20)^2)/4 = 303.5$
- Distance to mean measured by variance:

$\frac{0-20}{\sqrt{303.5}} = 1.148$ $\frac{12-20}{\sqrt{303.5}} = 0.459$ $\frac{15-20}{\sqrt{303.5}} = 0.287$

- $\frac{27-20}{\sqrt{303.5}} = 0.402$ $\frac{46-20}{\sqrt{303.5}} = 1.492$

Methods

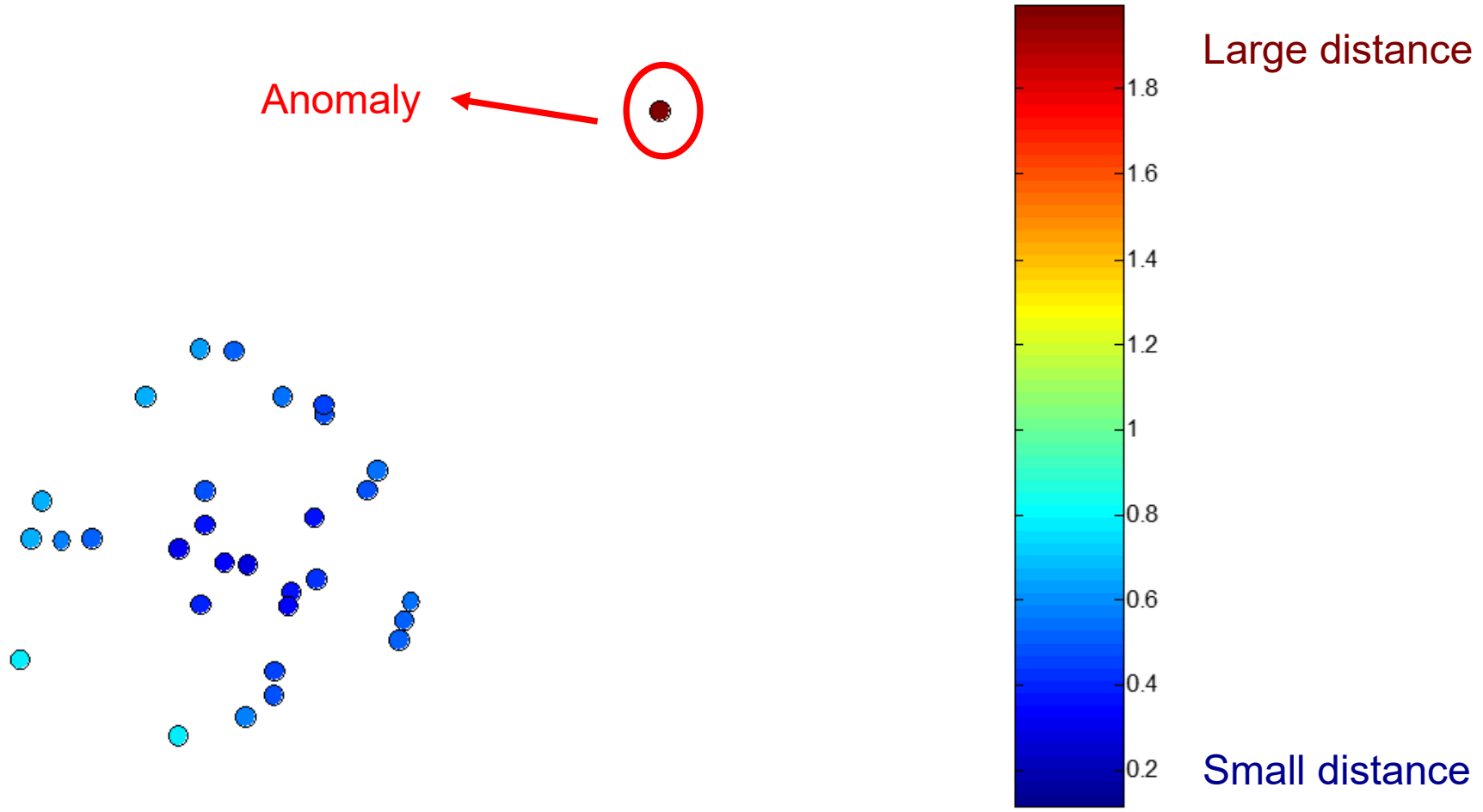
- Model-based technique
- Distance-based technique
- Density-based technique
- Cluster-based technique

Distance-based technique

- Objects are evaluated with respect to their distances to their *k*th closest point (also called *k*-nearest neighbour)
- The smaller the distance, the less chance this point is an anomaly
- Anomaly score: distance to the *k*-nearest neighbour

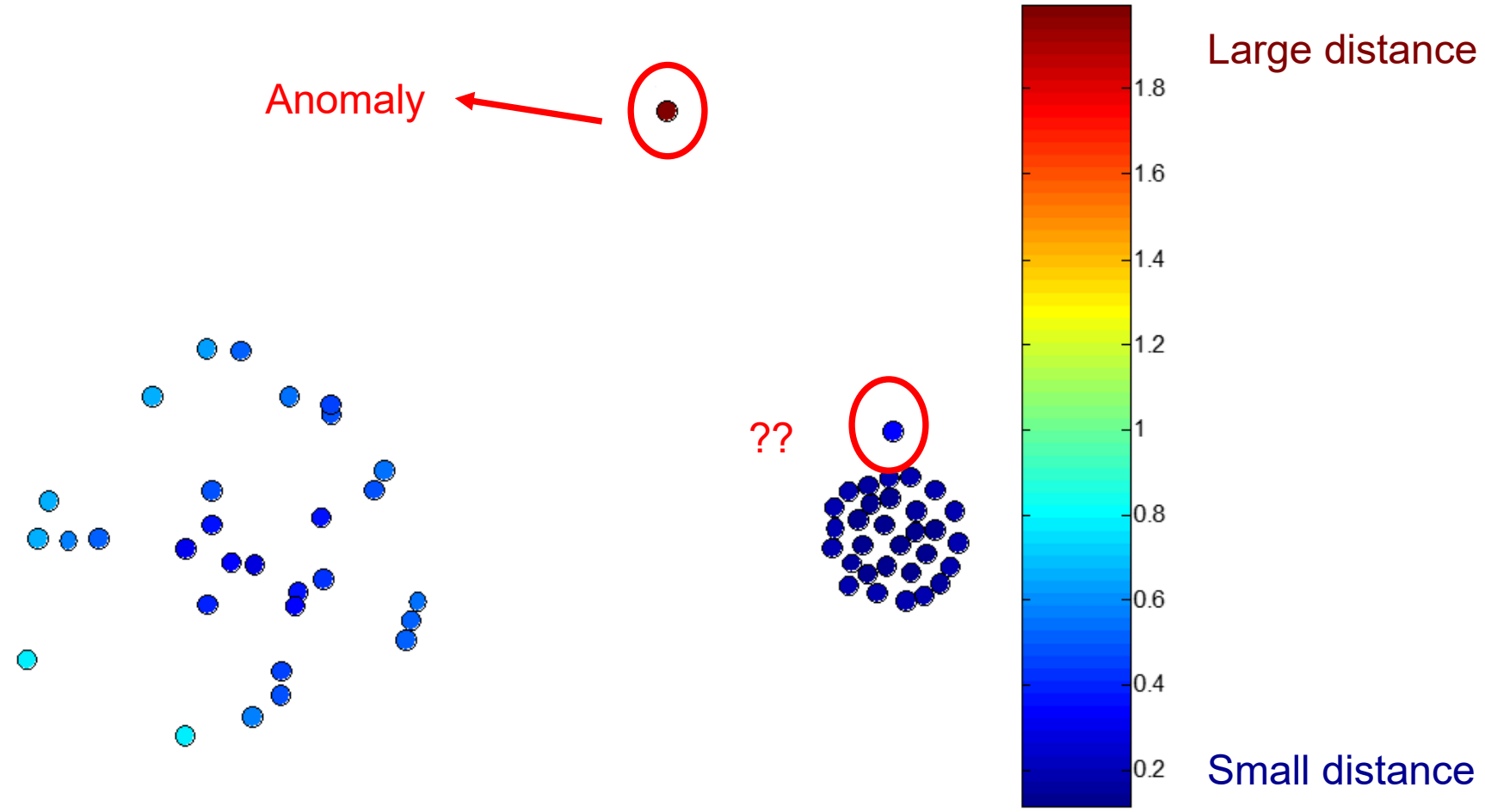
Example 1

- Which point would be identified as anomaly (k=5)?



Example 2

- Which point would be identified as anomaly (k=5)?



Strength and weakness

- Strength
  - Simple

- Weakness
  - Have to determine *k*
  - Sensitive to variations in density
  - Curse-of-dimensionality

Methods

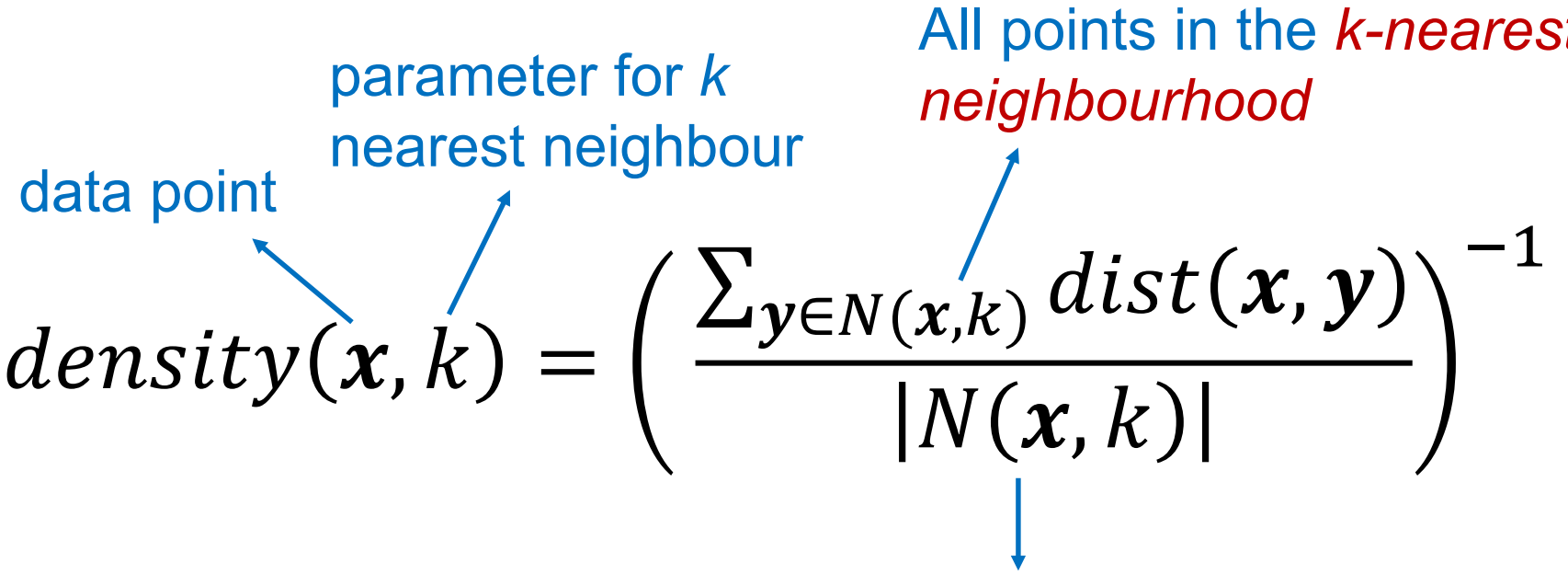
- Model-based technique
- Distance-based technique
- Density-based technique
- Cluster-based technique

Density-based technique

- Objects are evaluated with respect to their density
- The smaller the density is
  - The more likely this point is an anomaly
- Anomaly score: reciprocal of density

Density

- 1/(the average distance from any point in the ***k*-nearest neighbourhood** of $\boldsymbol{x}$ to $\boldsymbol{x}$)



$|\ N(\boldsymbol{x}, k)\ |$ gives the cardinality of the set $N(\boldsymbol{x}, k)$

- Why $|N(\boldsymbol{x}, k)|$ may not be equal to k?

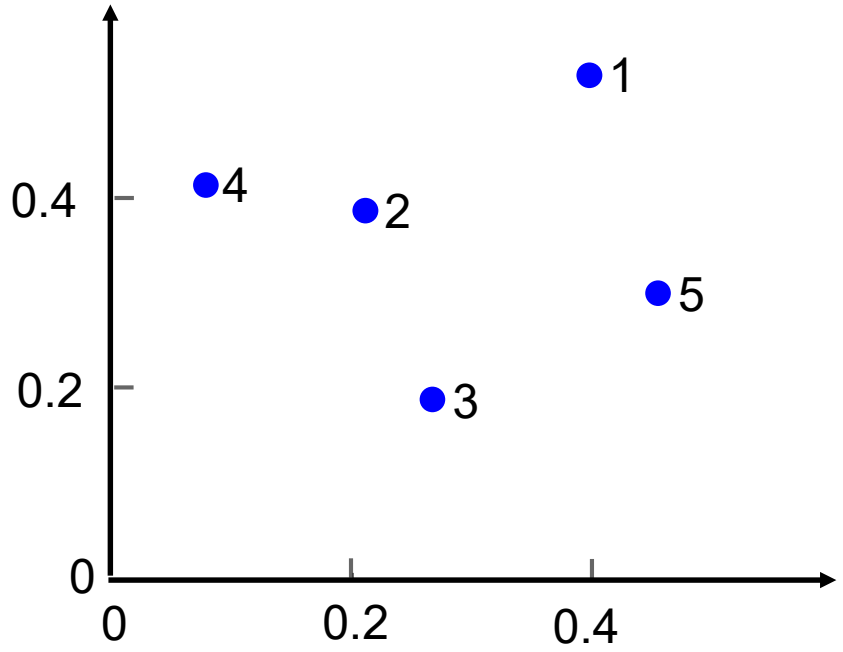
	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.235	0.342	0.235

If k = 2:

$$N(P_1, 2) = \{P_2, P_5, P_3\}$$

$$|N(P_1, 2)| = 3 \neq 2$$

Density: example



If k = 2:

$N(\mathrm{P}_1, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_2, 2) = \{\mathrm{P}_3, \mathrm{P}_4\}$
$N(\mathrm{P}_3, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_4, 2) = \{\mathrm{P}_2, \mathrm{P}_3\}$
$N(\mathrm{P}_5, 2) = \{\mathrm{P}_1, \mathrm{P}_3\}$

Distance Matrix:

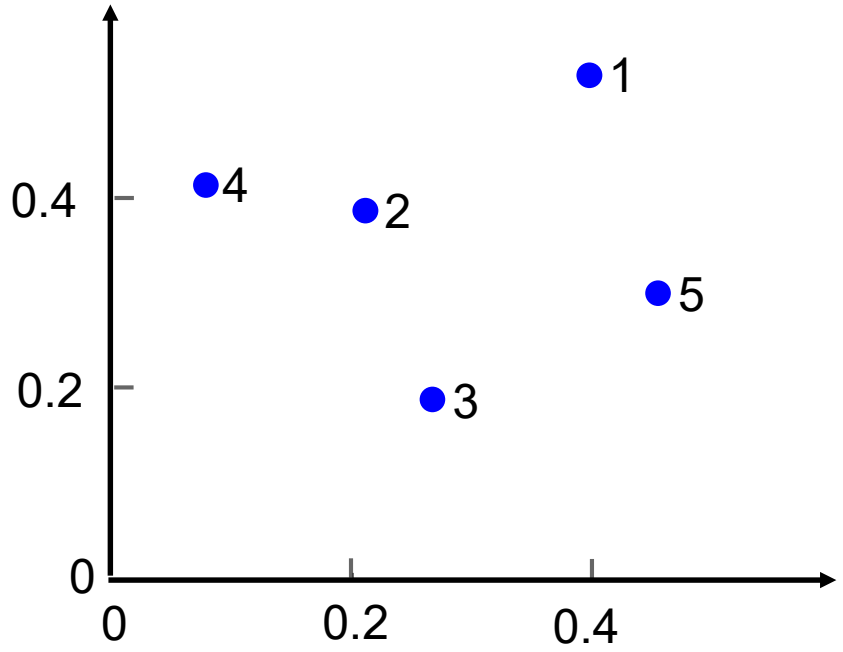
	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	0.234	0.000	0.194	0.143	0.244
P_3	0.368	0.194	0.000	0.284	0.220
P_4	0.342	0.143	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

	P_1	P_2	P_3	P_4	P_5
Density	4.264				

$$density(\boldsymbol{x}, k) = \left(\frac{\sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} dist(\boldsymbol{x}, \boldsymbol{y})}{|N(\boldsymbol{x}, k)|} \right)^{-1}$$

P_1: $\left(\frac{0.234+0.235}{2}\right)^{-1} = 4.264$

Density: example



If k = 2:

$N(\mathrm{P}_1, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_2, 2) = \{\mathrm{P}_3, \mathrm{P}_4\}$
$N(\mathrm{P}_3, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_4, 2) = \{\mathrm{P}_2, \mathrm{P}_3\}$
$N(\mathrm{P}_5, 2) = \{\mathrm{P}_1, \mathrm{P}_3\}$

Distance Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	0.234	0.000	0.194	0.143	0.244
P_3	0.368	0.194	0.000	0.284	0.220
P_4	0.342	0.143	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

	P_1	P_2	P_3	P_4	P_5
Density	4.264	5.935	4.831	4.684	4.396

$$density(\boldsymbol{x}, k) = \left(\frac{\sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} dist(\boldsymbol{x}, \boldsymbol{y})}{|N(\boldsymbol{x}, k)|} \right)^{-1}$$

$$\mathrm{P}_1: \left(\frac{0.234 + 0.235}{2} \right)^{-1} = 4.264$$

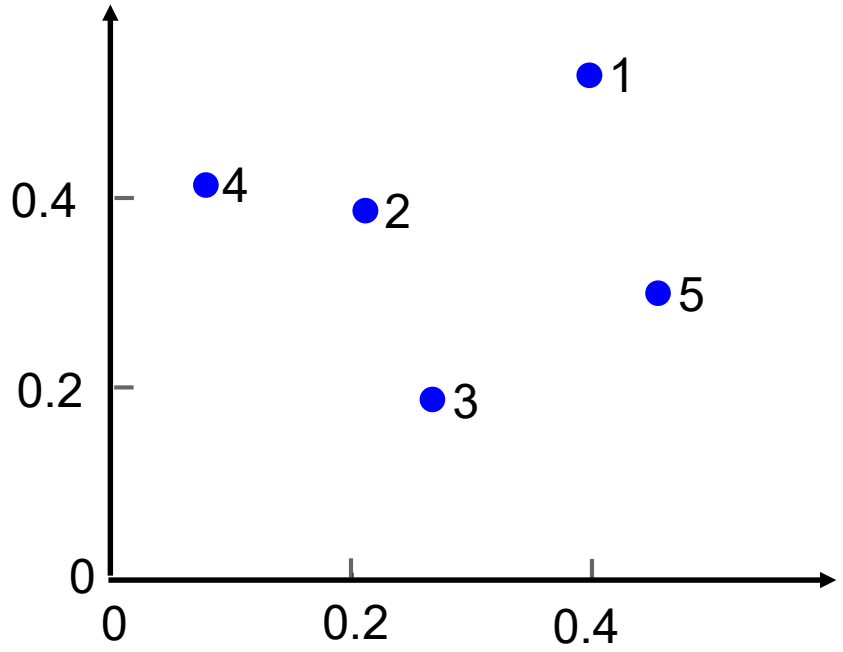
Average relative density

- The density relative to the average density within the neighbourhood

$$ave.rela.density(\boldsymbol{x}, k) = \frac{density(\boldsymbol{x}, k)}{\sum_{\boldsymbol{y} \in N(\boldsymbol{x}, k)} density(\boldsymbol{y}, \boldsymbol{k}) / |N(\boldsymbol{x}, k)|}$$

The average density in the *k-nearest neighbourhood*

Average relative density: example



If k = 2:

$N(P_1, 2) = \{P_2, P_5\}$
$N(P_2, 2) = \{P_3, P_4\}$
$N(P_3, 2) = \{P_2, P_5\}$
$N(P_4, 2) = \{P_2, P_3\}$
$N(P_5, 2) = \{P_1, P_3\}$

Distance Matrix:

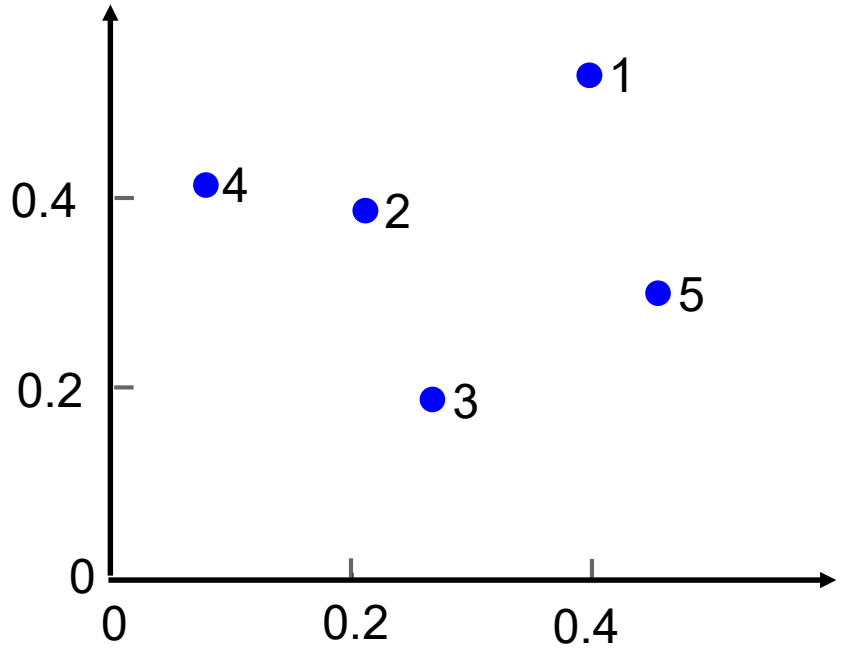
	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	0.234	0.000	0.194	0.143	0.244
P_3	0.368	0.194	0.000	0.284	0.220
P_4	0.342	0.143	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

	P_1	P_2	P_3	P_4	P_5
Density	4.264	5.935	4.831	4.684	4.396
ard	0.825				

$$ave.rela.density(P_1, 2) = \frac{density(P_1, 2)}{(density(P_2, 2) + density(P_5, 2))/2}$$

ard P_1: $\frac{4.264}{(5.935+4.396)/2} = 0.825$

Average relative density: example



If k = 2:

$N(\mathrm{P}_1, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_2, 2) = \{\mathrm{P}_3, \mathrm{P}_4\}$
$N(\mathrm{P}_3, 2) = \{\mathrm{P}_2, \mathrm{P}_5\}$
$N(\mathrm{P}_4, 2) = \{\mathrm{P}_2, \mathrm{P}_3\}$
$N(\mathrm{P}_5, 2) = \{\mathrm{P}_1, \mathrm{P}_3\}$

Distance Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	0.234	0.000	0.194	0.143	0.244
P_3	0.368	0.194	0.000	0.284	0.220
P_4	0.342	0.143	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

	P_1	P_2	P_3	P_4	P_5
Density	4.264	5.935	4.831	4.684	4.396
ard	0.825	1.248	0.935	0.870	0.967

$$ave.rela.density(\mathrm{P}_1, 2) = \frac{density(\mathrm{P}_1, 2)}{(density(\mathrm{P}_2, 2) + density(\mathrm{P}_5, 2))/2}$$

ard P_1: $\frac{4.264}{(5.935+4.396)/2} = 0.825$

Reachability distance

- Reachability distance from $\boldsymbol{x}$ to $\boldsymbol{y}$ with respect to k

$$reachdist_k(y \leftarrow x) = \max\{dist(x, y), dist_k(y)\}$$

$\boldsymbol{y}$'s distance to its *k*th nearest neighbour

Distance Matrix:

	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	0.234	0.000	0.194	0.143	0.244
P_3	0.368	0.194	0.000	0.284	0.220
P_4	0.342	0.143	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

Reachability distance matrix from row index to column index

	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	**0.235**	0.000	**0.220**	**0.284**	0.244
P_3	0.368	0.194	0.000	0.284	**0.235**
P_4	0.342	**0.194**	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

Local reachability density (lrd)

Keep the nearest neighbours:

- lrd: Calculate the density based on the reachability distance matrix
- Average relative lrd (arlrd): Calculate the average relative density based on the reachability distance matrix

If k = 2:

$$N(P_1, 2) = \{P_2, P_5\}$$
$$N(P_2, 2) = \{P_3, P_4\}$$
$$N(P_3, 2) = \{P_2, P_5\}$$
$$N(P_4, 2) = \{P_2, P_3\}$$
$$N(P_5, 2) = \{P_1, P_3\}$$

Reachability distance matrix from row index to column index:

	P_1	P_2	P_3	P_4	P_5
P_1	0.000	0.234	0.368	0.342	0.235
P_2	**0.235**	0.000	**0.220**	**0.284**	0.244
P_3	0.368	0.194	0.000	0.284	**0.235**
P_4	0.342	**0.194**	0.284	0.000	0.386
P_5	0.235	0.244	0.220	0.386	0.000

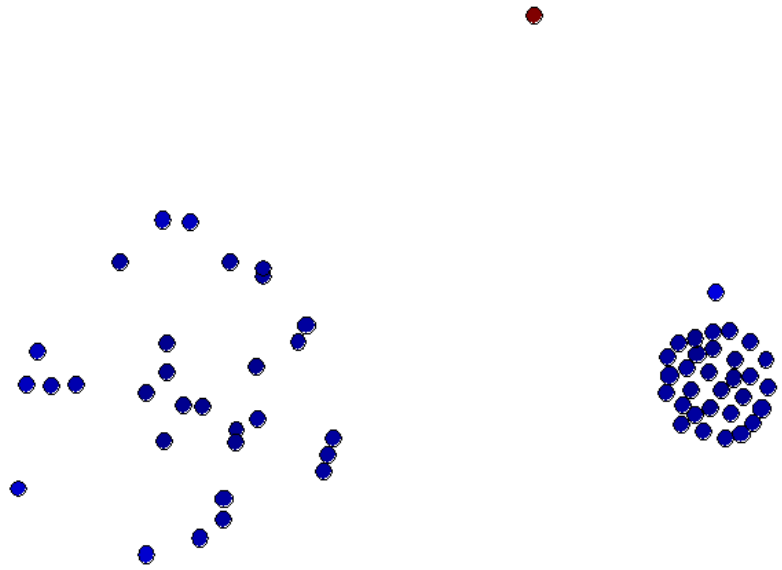
lrd and arlrd:

	P_1	P_2	P_3	P_4	P_5
lrd	4.264	3.968	4.662	4.184	4.396
arlrd	1.020	0.897	1.115	0.970	0.985
1/arlrd	0.980	1.115	0.897	1.031	1.015

$$P_2: \left(\frac{0.220+0.284}{2}\right)^{-1} = 3.968$$

Local outlier factor (LOF)

- Local outlier factor: the *reciprocal* of *average relative density* based on *reachability distance*
- Higher LOF, more likely to be anomaly



Strength and weakness

- Strength
  - Less sensitive to variations in density
- Weakness
  - Curse-of-dimensionality

Methods

- Model-based technique
- Distance-based technique
- Density-based technique
- Cluster-based technique

Cluster-based technique

- Clustering is conducted first on the data, and objects are evaluated with respect to how *strong* they belong to each cluster.

- If a point does not strongly belong to any cluster
  - This point is an anomaly

Measurement of strong

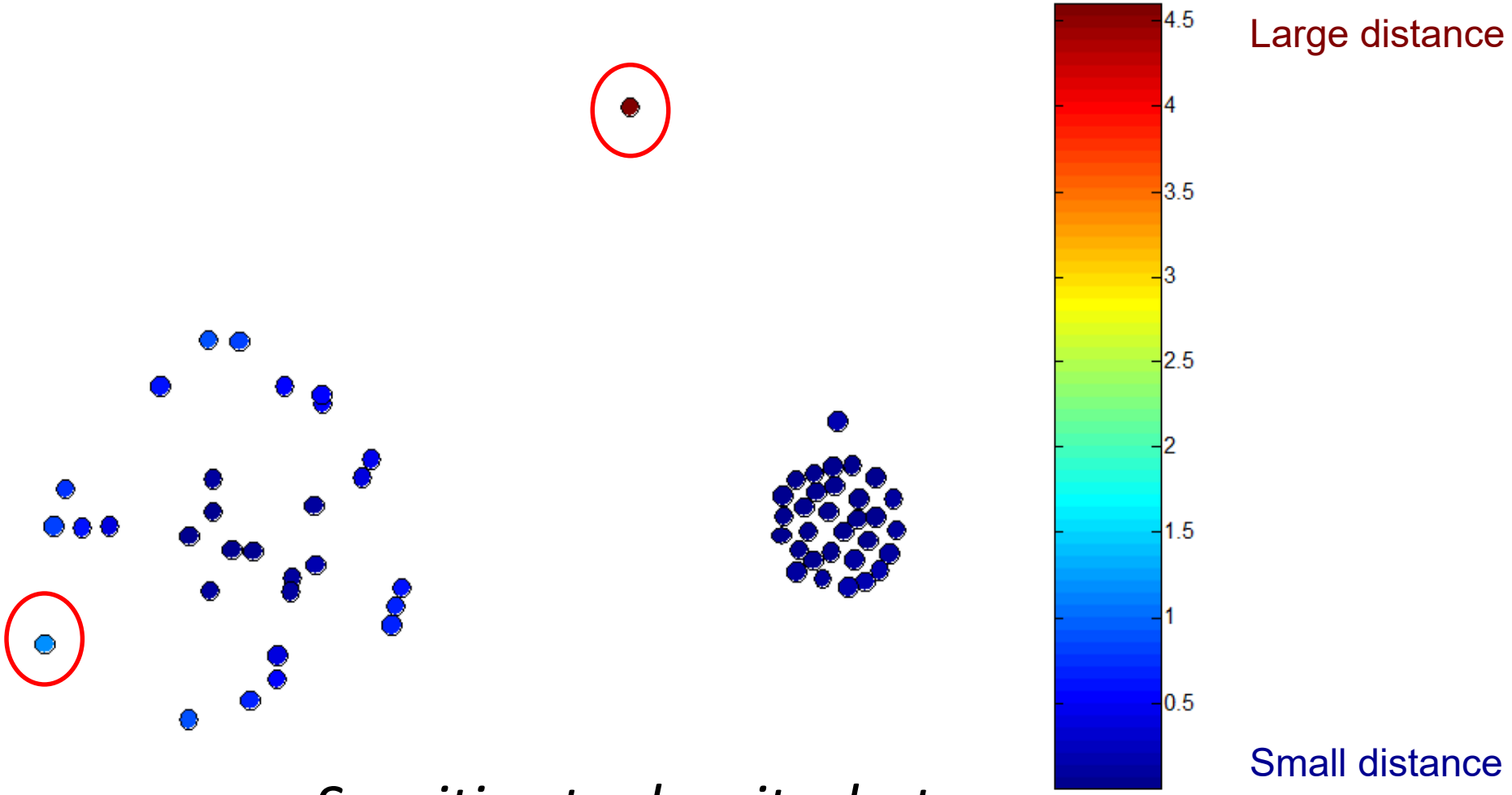
- Prototype-based clustering
  - Distance to any centroid is too large

- Density-based clustering
  - Points within the Eps circle is too few

Will these two methods be sensitive to "density"?

Example with distance

- For prototype-based clustering



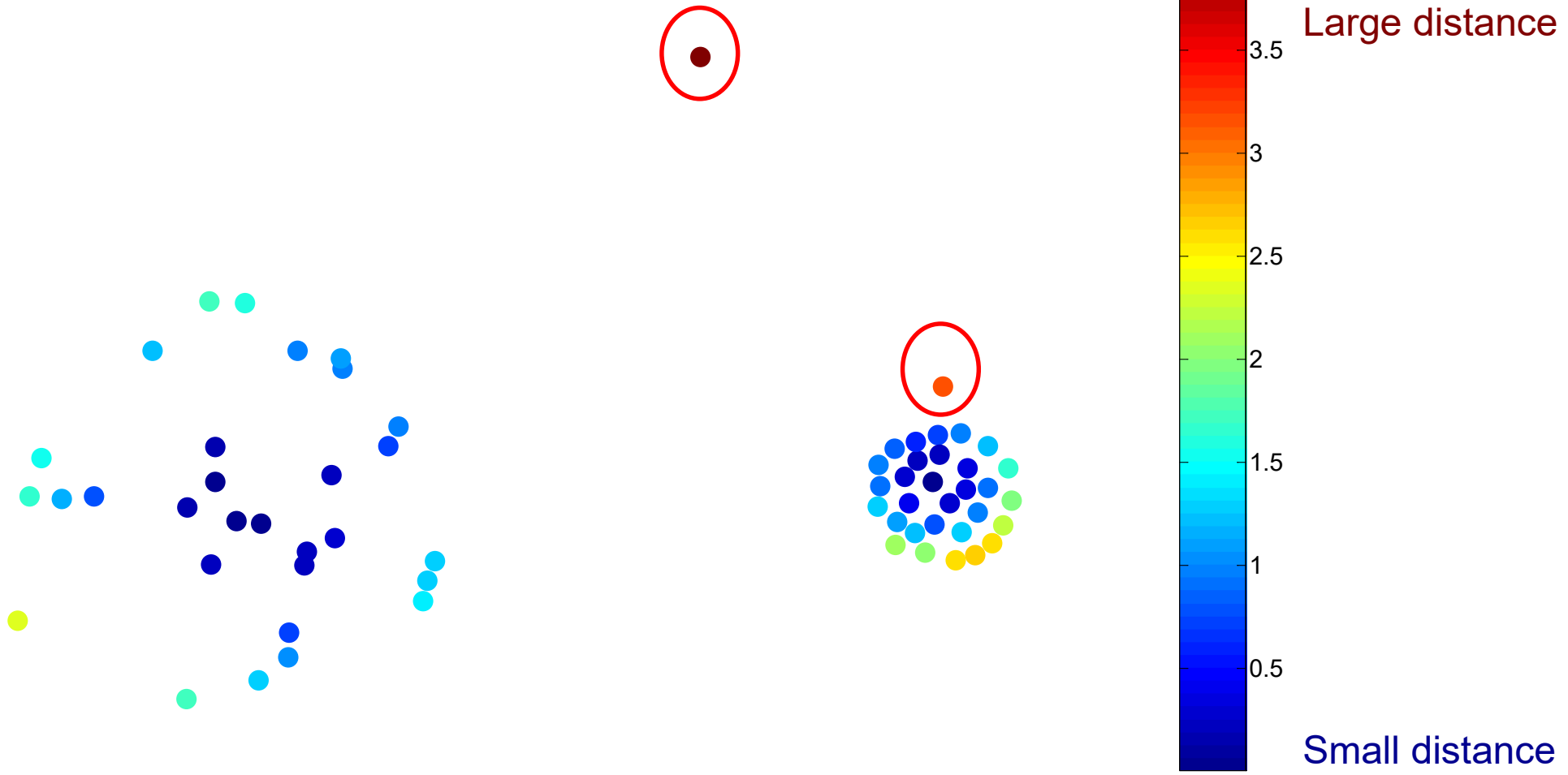
Sensitive to density, but…

Relative distance

$$relative\ dist(x, closest\_centroid) = \frac{dist(x, closest\_centroid)}{the\ average\ dist.\ from\ a\ data\ in\ the\ cluster\ to\ this\ ceontrid}$$

Example with related distance

- For prototype-based clustering



Using relative distance will make the algorithm insensitive to density.

Strength and weakness

- Strength
  - Can borrow the merit from development of clustering technique
  - One variant using relative distance is not sensitive to density

- Weakness
  - Parameter setting for clustering algorithm
  - Outlier may distort the cluster

Challenges

- Effectiveness
- Grey area
- Application-specific setting
- Understandability

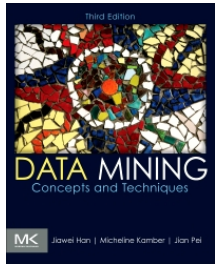
Summary

Anomaly detection: finding data points that are anomaly/outliers in data

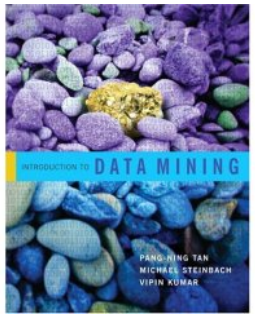
- Model-based technique
- Proximity-based technique
- Density-based technique
- Cluster-based technique

Recommended reading (not required)

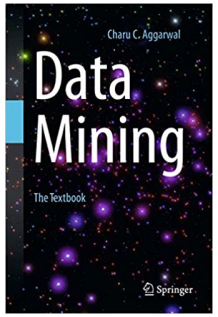
- [Han et al., 2012]
  - Chapter 12



- [Tan et al., 2005]
  - Chapter 10



- [Aggarwal, 2015]
  - Chapter 8



Next week

- Another way to do anomaly detection:
  - Collect historical data telling which are anomaly and which are not.
  - "**Learn**" from these **labelled** data to do **classification**

Cat

Dog