COMP 330/543: Outliers

Luis Guzman

Sinan Kockara

Chris Jermaine

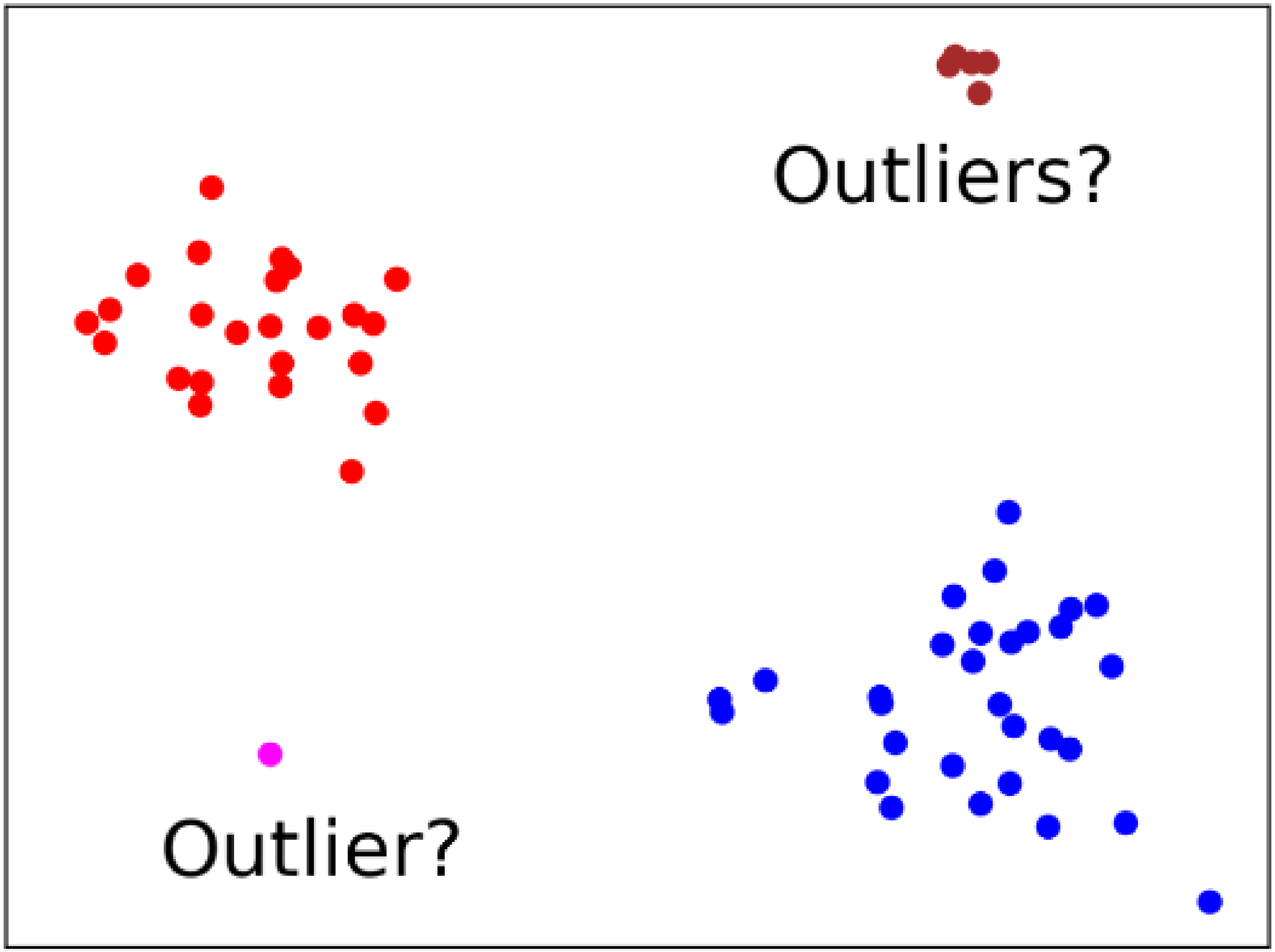
Rice University

What Are Outliers In Data Science?

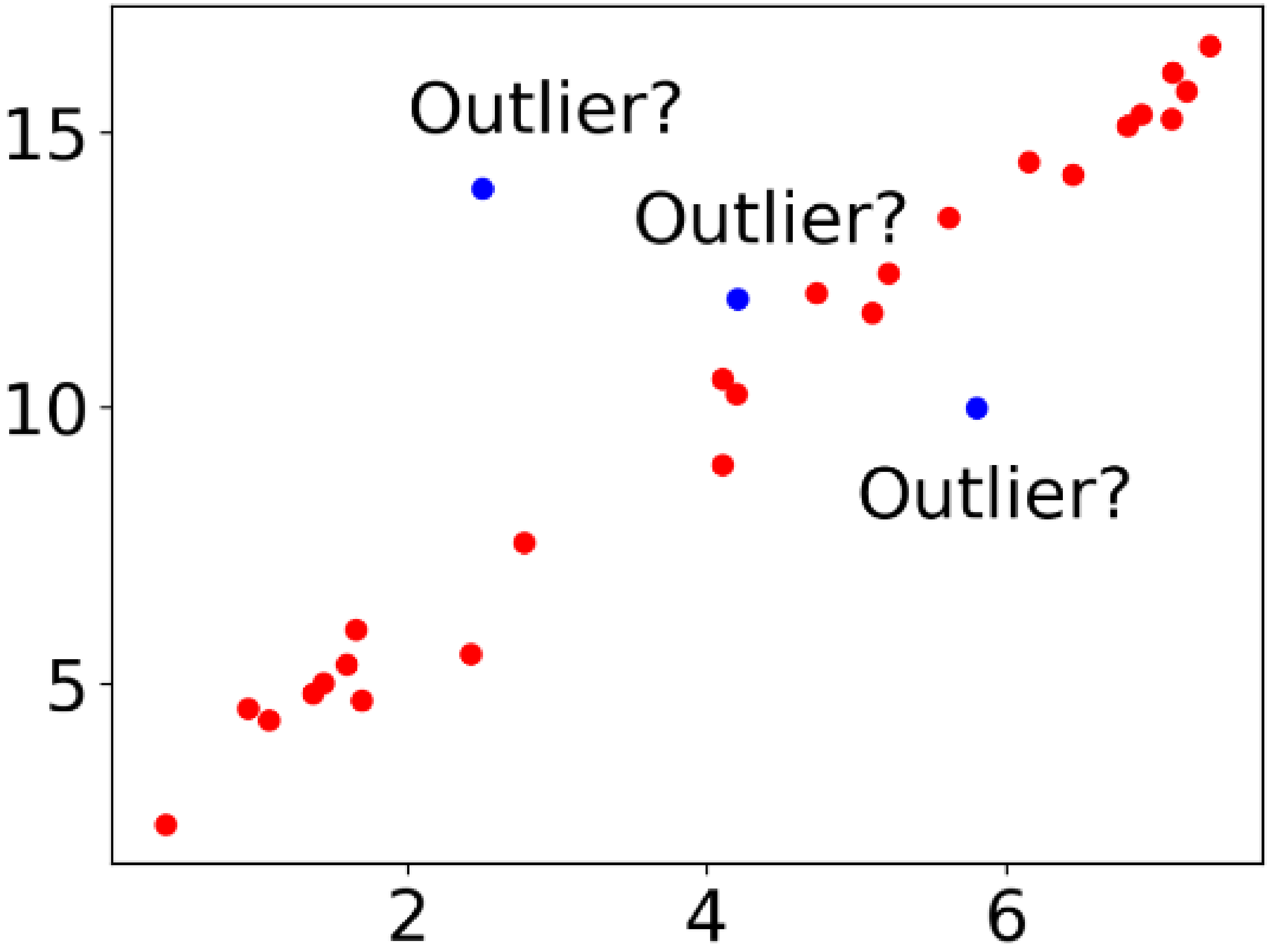
Data points that are unlike the other points, unexpected, or unusual in some way

- Weather data set: low of -12 degrees in Houston
- Sports data set: averaging 10+ rebounds, 10+ assists per game
- Stock trading data set: on day S & P 500 down 300 points, a stock up 10%

Outlier Pic: 1



Outlier Pic: 2



Outliers Not Same As Unusual Data

- Low of 12 degrees in Houston is unusual, probably not an outlier
- Low of -12 in Chicago (a bit) unusual, probably not an outlier
- It's the combination of Houston, -12 that makes this difficult to believe

Why Look For Outliers?

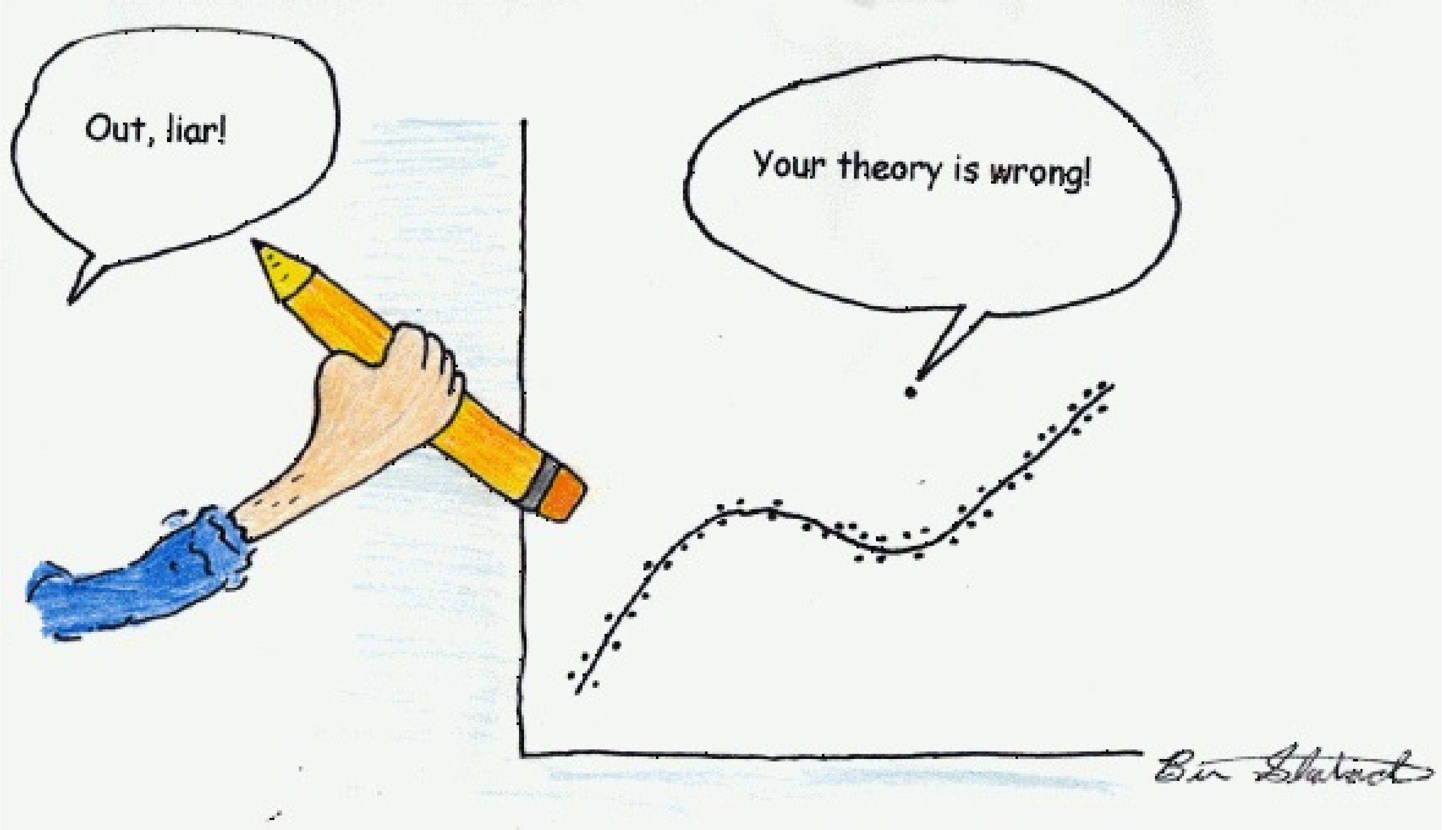
Classically, two reasons:

- To remove them from data...
- ...because outliers can hurt the learning process

Why throw out data?

- It messes up the model
- Garbage in, garbage out
- E.g., huge impact on least squares

But This Can Be Dangerous



Is it really garbage data?

- Values that defy natural laws

Why Look For Outliers?

Classically, two reasons:

- (1) Classically, to remove them from data...
- ...because outliers can hurt the learning process
- (2) To find them for further examination...
- ...because outliers might enhance understanding of data

What are some examples where we might want to find outliers?

Why Look For Outliers?

Classically, two reasons:

- (1) Classically, to remove them from data...
- ...because outliers can hurt the learning process
- (2) To find them for further examination...
- ...because outliers might enhance understanding of data

What are some examples where we might want to find outliers?

- Computer security
- Fraud detection
- Medical crisis alerts

Outlier Detection is an Unsupervised Task

Why?

Outlier Detection is an Unsupervised Task

Why?

- Supervised learning requires labels
- We don’t always know what the outliers look like
- By definition, we don’t expect them
- They are rare and unexpected

How Do We Define Outliers?

Two standard definitions:

- (1) Distance-based
- (2) Model-based

Distance-Based Outliers

Def: A point is an outlier if it is far from all other points

Outlier search often defined in terms of kNN:

- Let $d(x_i)$ be the distance to point x_i's kth NN in the data set
- Then given data set $\langle x_1, x_2, ..., \rangle$, we want to compute the set O such that...
- $|O| = m$ and $\forall(x_o \in O, x_i \in \mathrm{X} - \mathrm{O}), d(x_i) \leq d(x_o)$

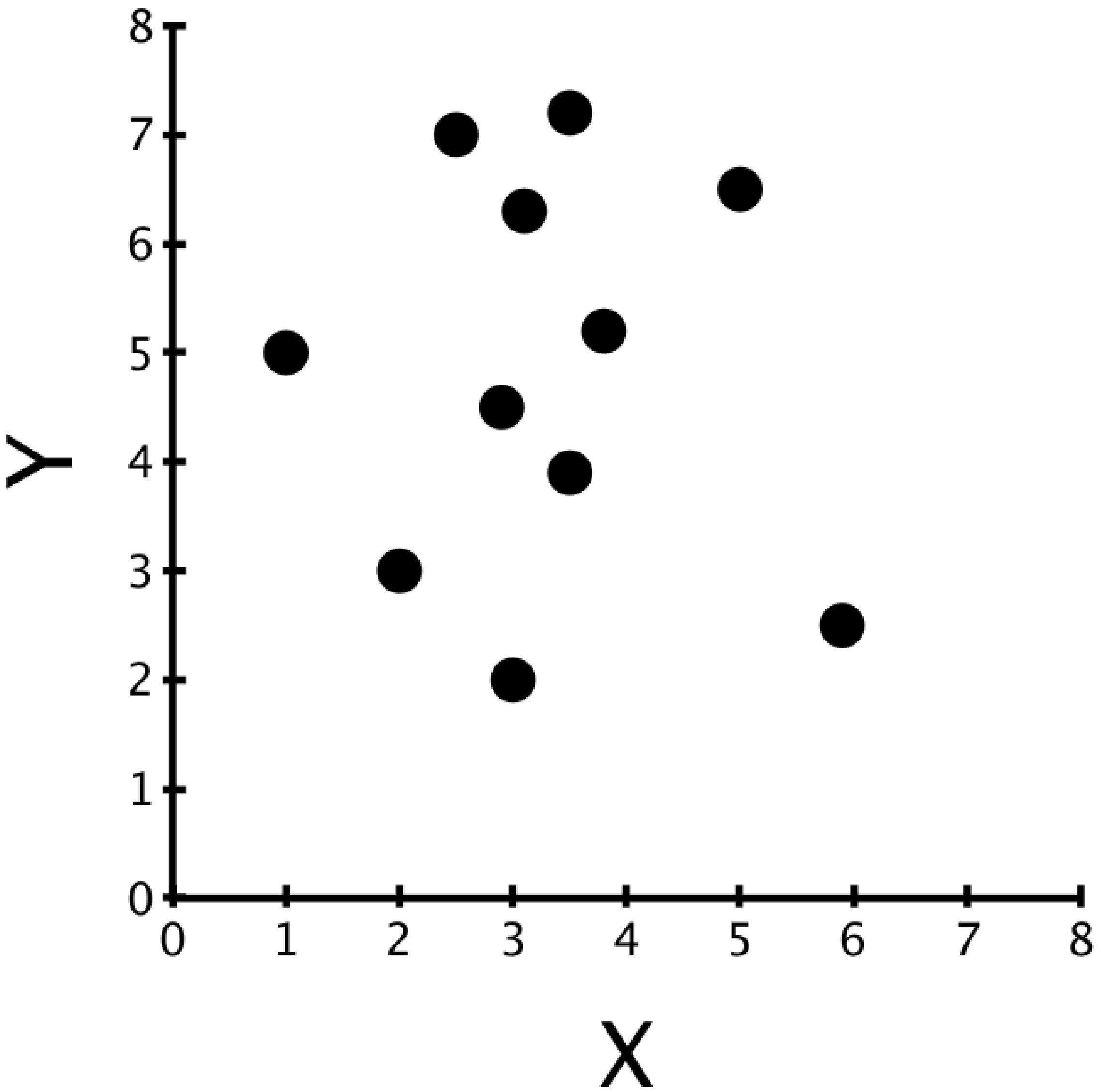
How To Find Distance-Based Outliers?

Simple algorithm:

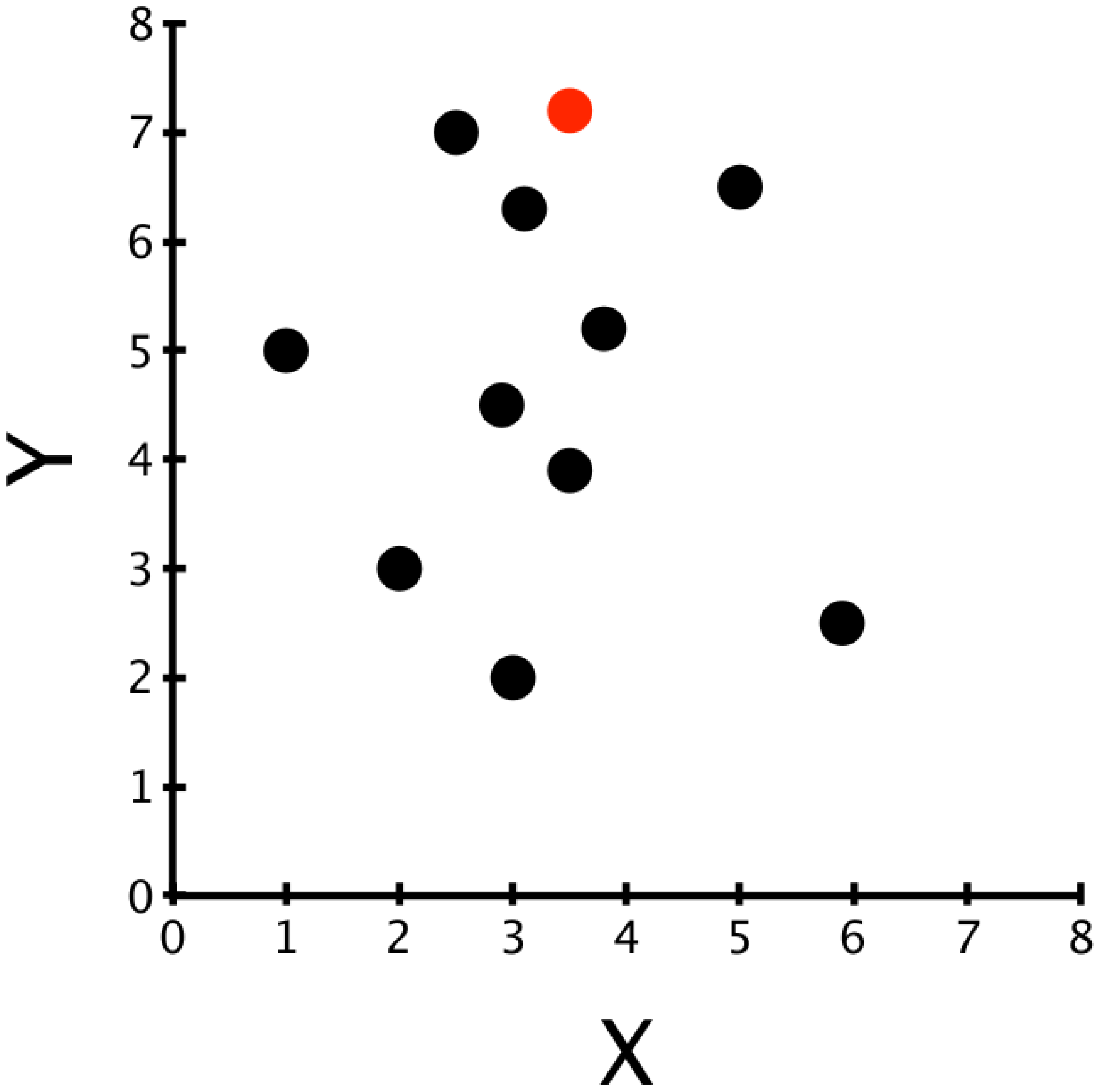
```
init **min**-priority queue $O$
**for** $x_1 \in X$:
  init **max**-priority queue $P$
  **for** $x_2 \neq x_1 \in X$:
    **insert** $dist(x_1, x_2)$ **into** $P$
    if $|P| > k$
      remove **max from** $P$
  **insert** $x_1$ **into** $O$ **with key max**$(P)$
  if $|O| > m$
    remove point **with min key from** $O$

return $O$
```

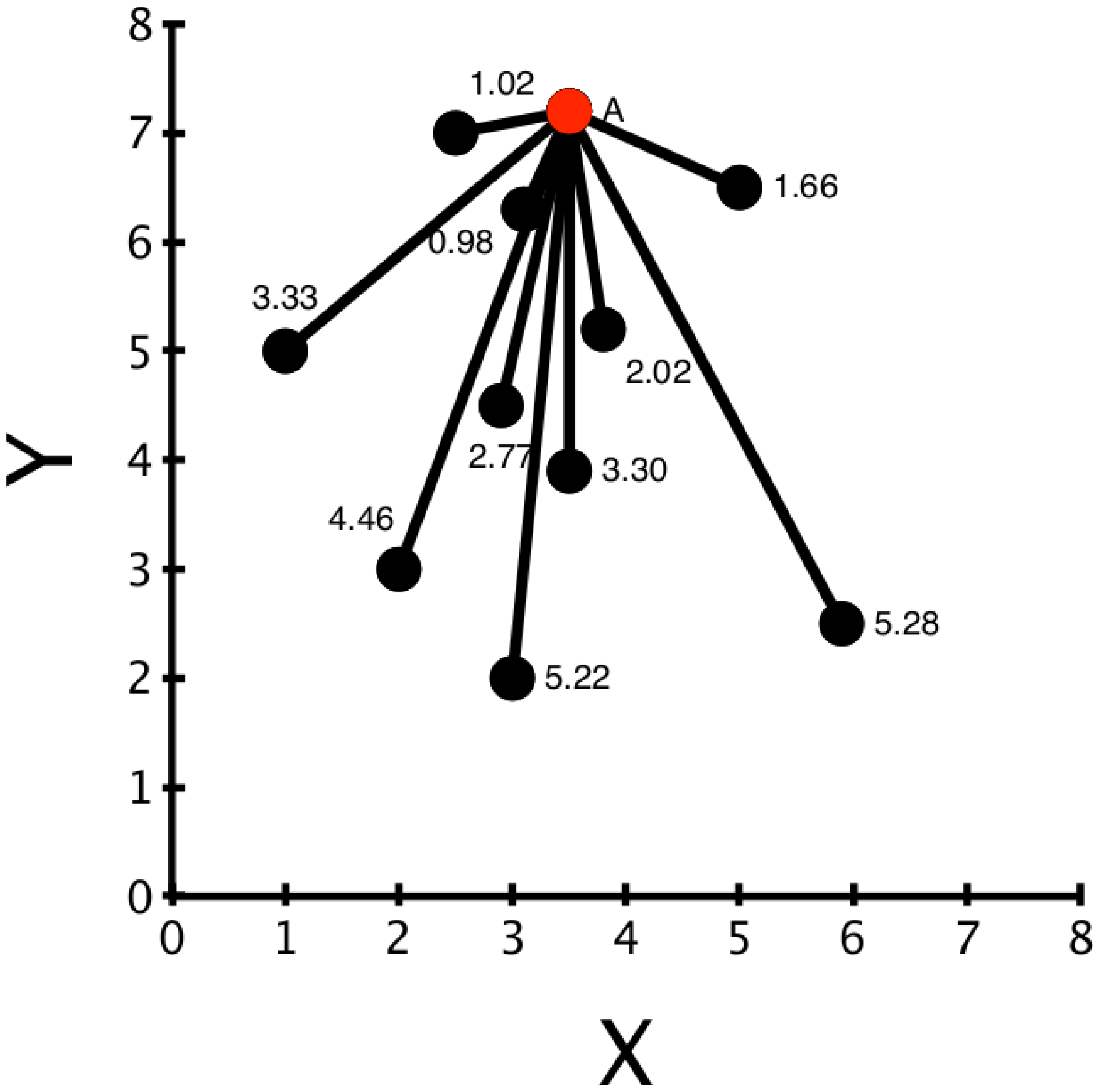
Example: 2NN



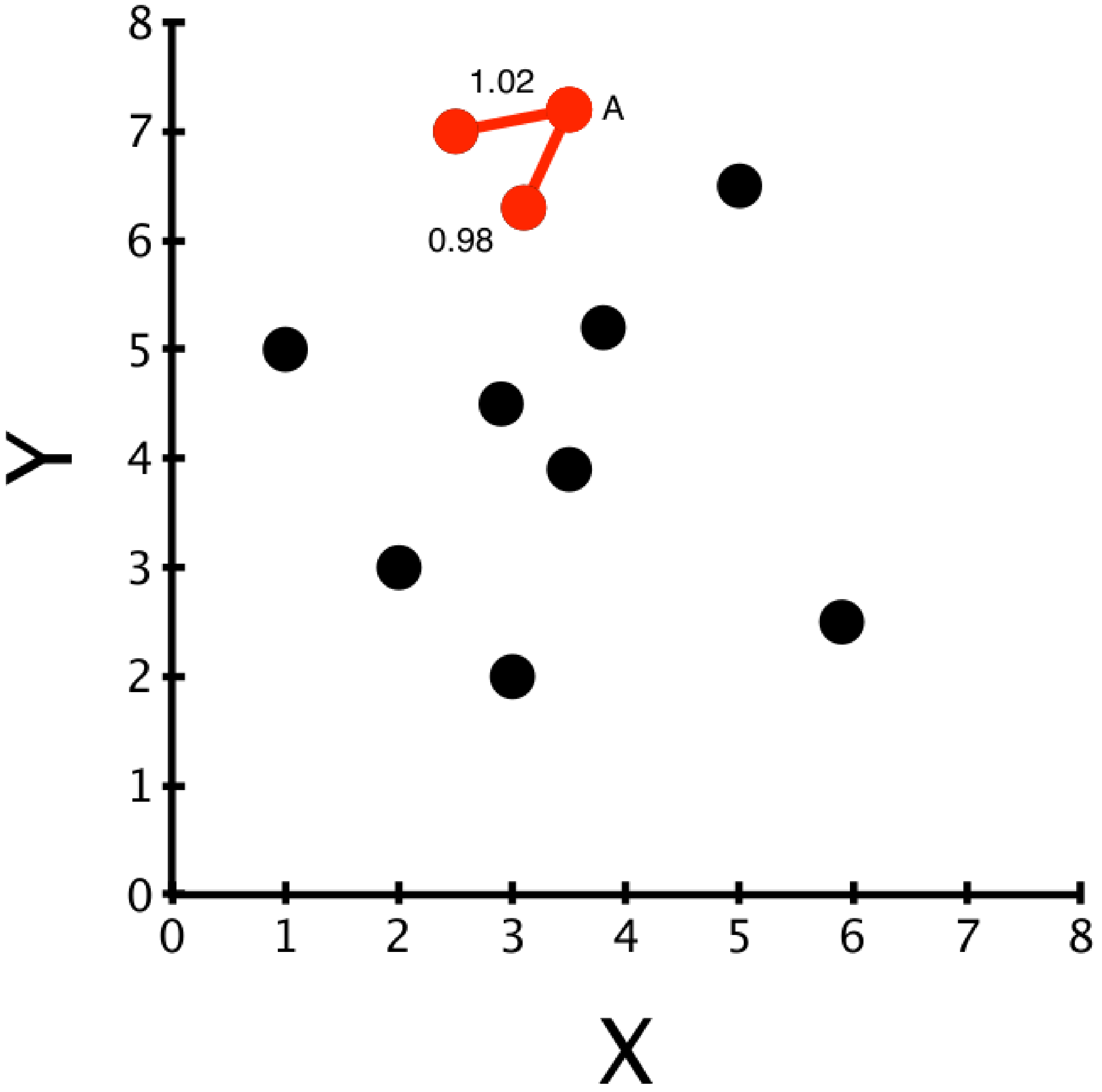
Example: 2NN



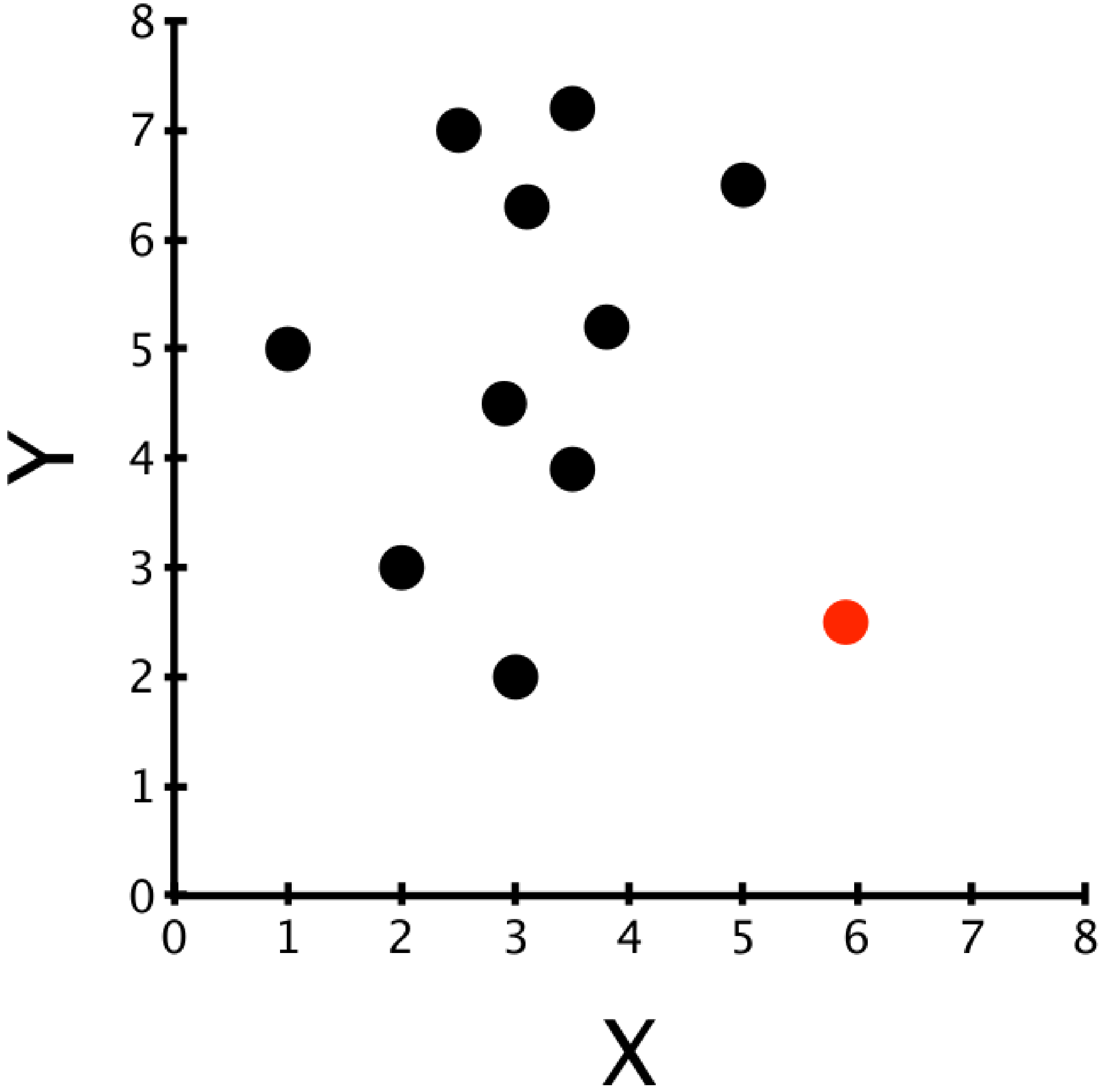
Example: 2NN



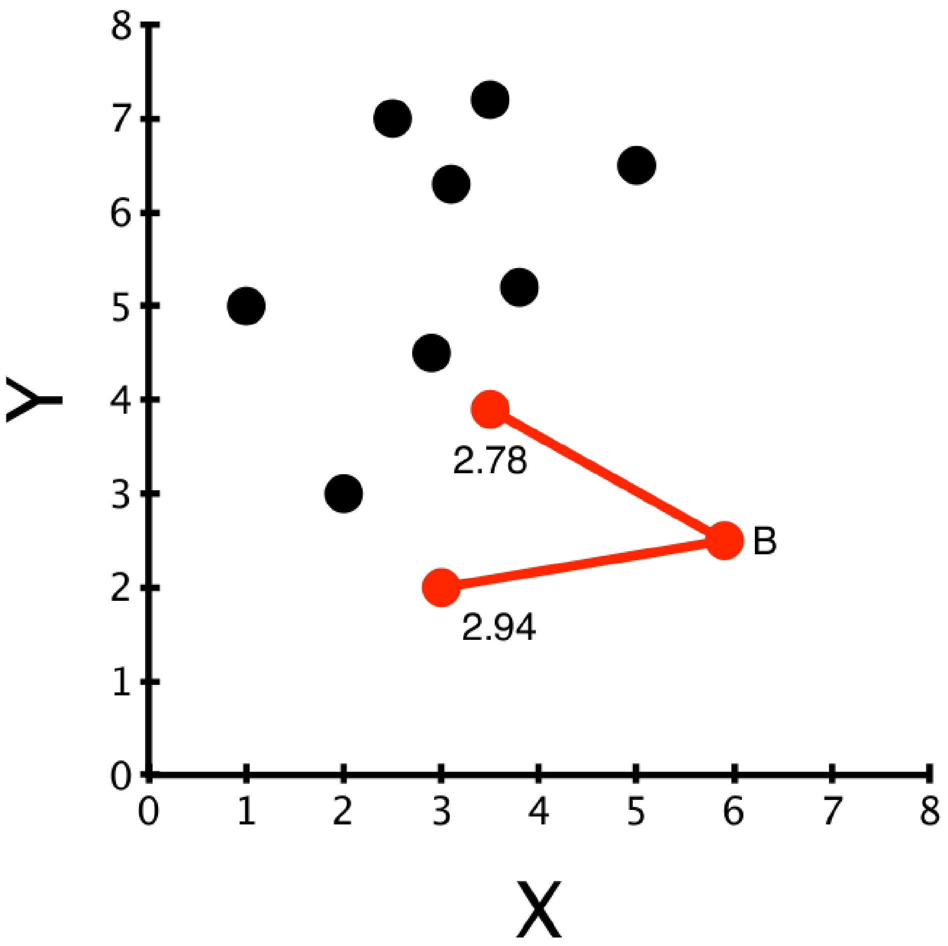
Example: 2NN



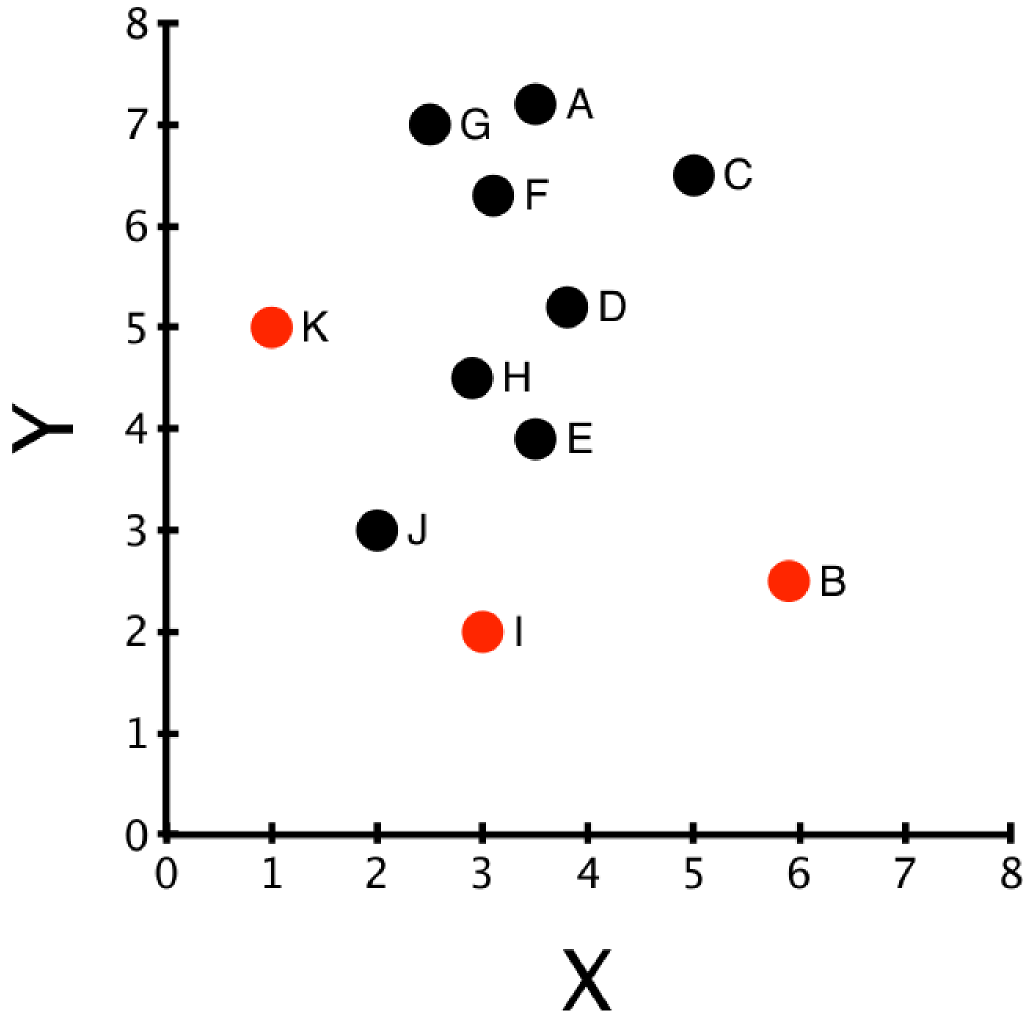
Example: 2NN



Example: 2NN



Example: 2NN $m = 3$



Point	2NN distance
A	1.02
B	2.94
C	1.77
D	1.30
E	1.33
F	0.98
G	1.02
H	1.14
I	1.96
J	1.75
K	2.24

How To Find Distance-Based Outliers?

Simple algorithm:

```
init min-priority queue $O$
for $x_1 \in X$:
  init max-priority queue $P$
  for $x_2 \neq x_1 \in X$:
    insert $dist(x_1, x_2)$ into $P$
    if $|P| > k$
      remove max from $P$
  insert $x_1$ into $O$ with key max$(P)$
  if $|O| > m$
    remove point with min key from $O$

return $O$
```

- So the algorithm works... What's the problem here?

How To Find Distance-Based Outliers?

What’s the problem here?

- Nested loop through the entire database
- Too slow for big data
- How to address?

How To Find Distance-Based Outliers?

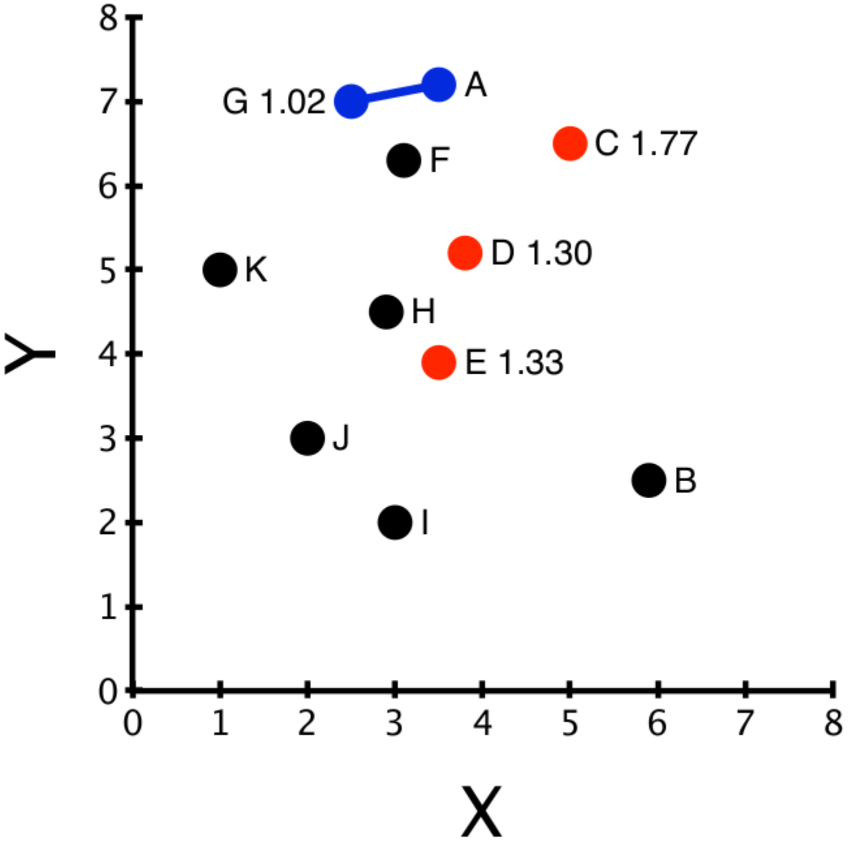
Better algorithm:

```
init min-priority queue $O$
for $x_1 \in X$:
  init max-priority queue $P$
  for $x_2 \neq x_1 \in X$:
    insert $dist(x_1, x_2)$ into $P$
    if $|P| > k$
      remove max from $P$
    if $|P| == k$ and $|O| == m$ and max$(P)$ < min$(O)$
      discard $x_1$; not an outlier
  insert $x_1$ into $O$ with key max$(P)$
  if $|O| > m$
    remove point with min key from $O$

return $O$
```

How To Find Distance-Based Outliers?

- Example: O: {(1.33, E), (1.77, C), (1.3, D)}
- Process A, find 2-NN where $dist(A, F) = 0.98$, discard A



How To Find Distance-Based Outliers?

Why does this help?

- `max`(P) is an upper bound on distance to kth NN
- So distance to kth NN can't ever be greater
- If this is not good enough to get point into top m in O
- Then can discard it early
- Can get a 100x speed up
- Still $O(N^2)$

How To Find Distance-Based Outliers?

Why does this help?

- `max`(P) is an upper bound on distance to kth NN
- So distance to kth NN can't ever be greater
- If this is not good enough to get point into top m in O
- Then can discard it early
- Can get a 100x speed up
- Still $O(N^2)$

Even better:

- Store X in randomized order
- Lower chances of getting unlucky and finding all far points first

What About the hyper-parameters

How to compute $dist(x, y)$?

- Classical method: if x, y vectors, use l_p norm of $x - y$

How to choose m?

- Very application specific
- Start small, gradually increase it
- Stop when nothing "interesting" is added

How to choose k?

- Empirically determined using the validation set
- Try $\sqrt{N}$

Model-Based Outliers

Basic idea:

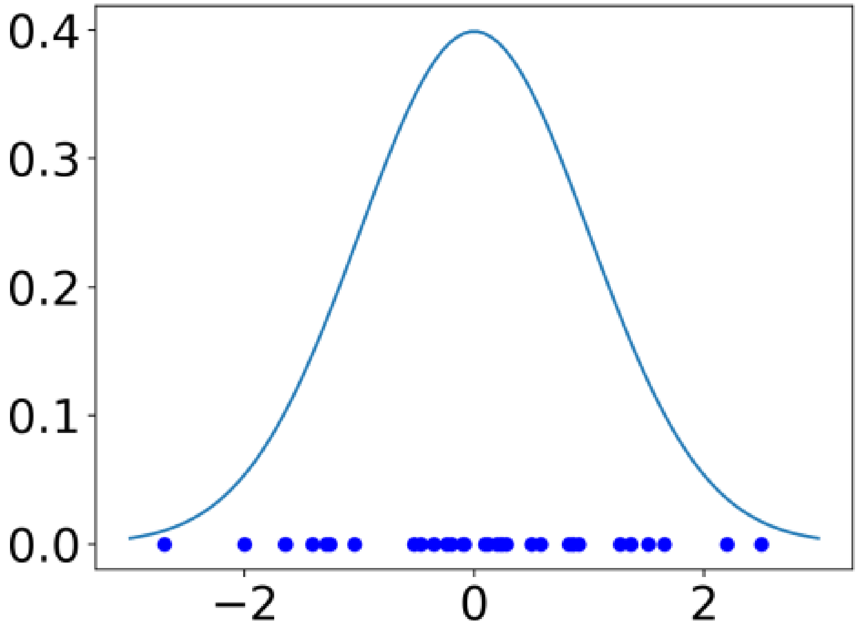
- Learn a model for what is “typical”
- Might be probabilistic
- Might be least-squares
- Then outlier is a data point with a low score according to the model

Example of Model-Based Detection

Learn Normal distribution by choosing $\{\mu\}, \{\sigma^2\}$ to maximize

$$P(x_1, x_2, ..., x_n) = \prod_i \textbf{Normal}(x_i|\mu, \sigma^2)$$

Then choose m points with smallest $\textbf{Normal}(x_i|\mu, \sigma^2)$



Questions?