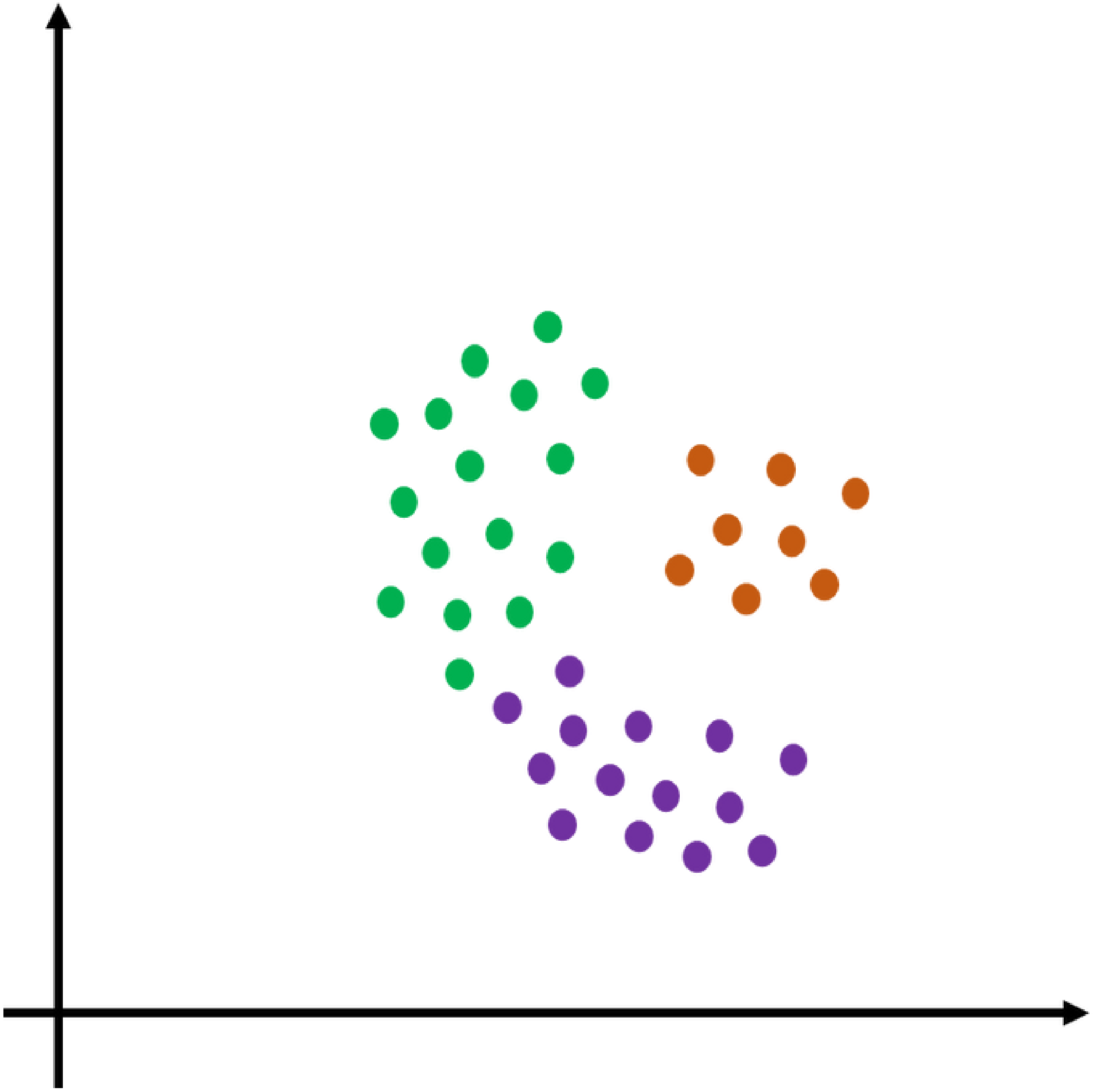
DBSCAN CLUSTERING

A BRIEF, INTUITIVE INTRODUCTION



DISCOVER HIDDEN STRUCTURE IN YOUR DATA

Dave ON DATA

Introducing Cluster Analysis

"**Cluster analysis** groups data objects based only on information found in the data that describes the objects and their relationships.

The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in other groups.

The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering"

Source: Introduction to Data Mining by Pang-Ning, Michael Steinbach, and Vipin Kumar, first edition May 2, 2005.

Because cluster analysis has no external information about groups (i.e., **labels**), it belongs to a form of machine learning known as **unsupervised learning**.

Because so much data is unlabeled, cluster analysis is a widely used tool to discover structure in data and produce new insights.

BTW - The words "groups" and "clusters" mean the same thing.

Introducing DBSCAN

The **density-based spatial clustering of application with noise (DBSCAN) algorithm** is a density-based, partial partitioning clustering technique.

Whoa! That was a mouthful. Let's break that down:

1. DBSCAN uses areas of low data point density vs. areas of high data point density to find clusters.
2. The clusters produced by DBSCAN can be arbitrarily shaped.
3. Not all data points are assigned to clusters (e.g., outliers) – unassigned data points are called **noise points**.
4. Non-noise data points will be assigned to a single cluster.

DBSCAN is a popular clustering technique.

DBSCAN is popular because of the algorithm's simplicity (it is easy to understand how DBSCAN works intuitively) and DBSCAN's ability to handle arbitrarily-shaped clusters.

A Contrived Example

The DBSCAN algorithm functions using the following definitions:

- **Core points**: These are data points inside a cluster. A data point is core if surrounded by a minimum number of other data points within a set distance. When using DBSCAN you set the distance (**eps**) and the minimum number of points within the distance (**min_samples**).
- **Border points**: A border point is not a core point, but falls within the set distance (i.e., **eps**) of a core point.
- **Noise points**: A noise point is any point that is neither a core point nor a border point.

Here's the DBSCAN algorithm:

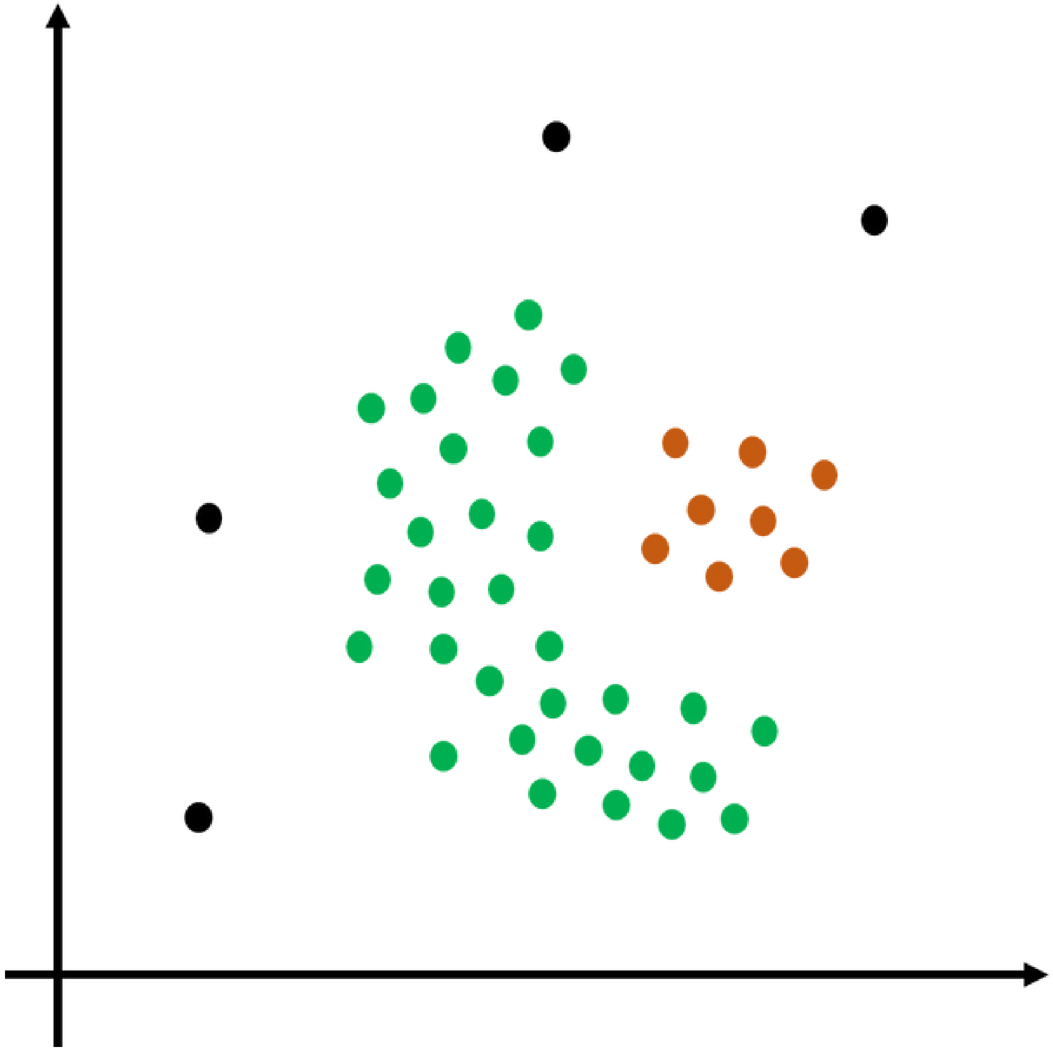
1. Label all data points as core, border, or noise points.
2. Eliminate noise points.
3. Connect all core points that are within **eps** distance of each other.
4. Make each group of connected core points a separate cluster.
5. Assign each border point to one of the clusters based on proximity to core points.

DBSCAN CLUSTERING

Consider a hypothetical dataset that needs to be clustered. To keep things simple, there are only two numeric features in the dataset.

Looking at the data, our eyes tend to see the clusters depicted.

We'll see how the DBSCAN algorithm handles this contrived dataset.

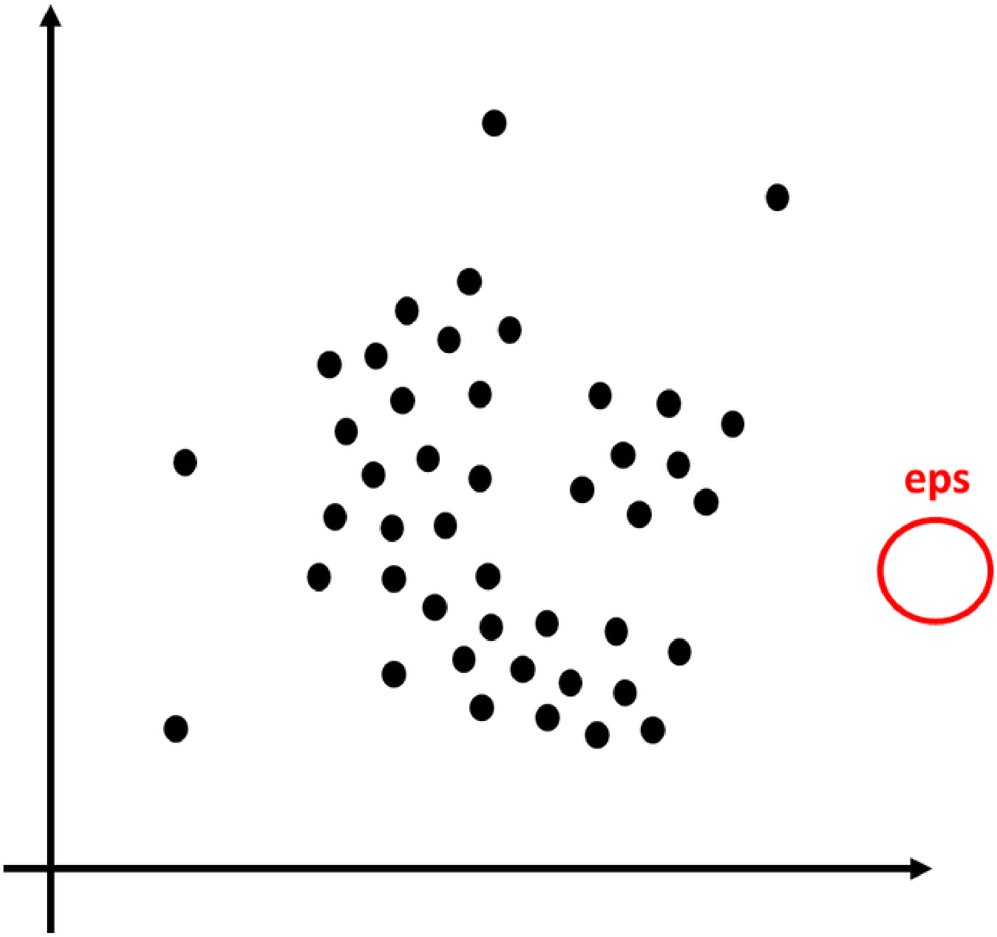


DBSCAN CLUSTERING

The first step in the DBSCAN algorithm is to classify each data point as a core, border, or noise point.

To do that, DBSCAN uses **eps** for the distance and **min_samples** for the minimum number of data points.

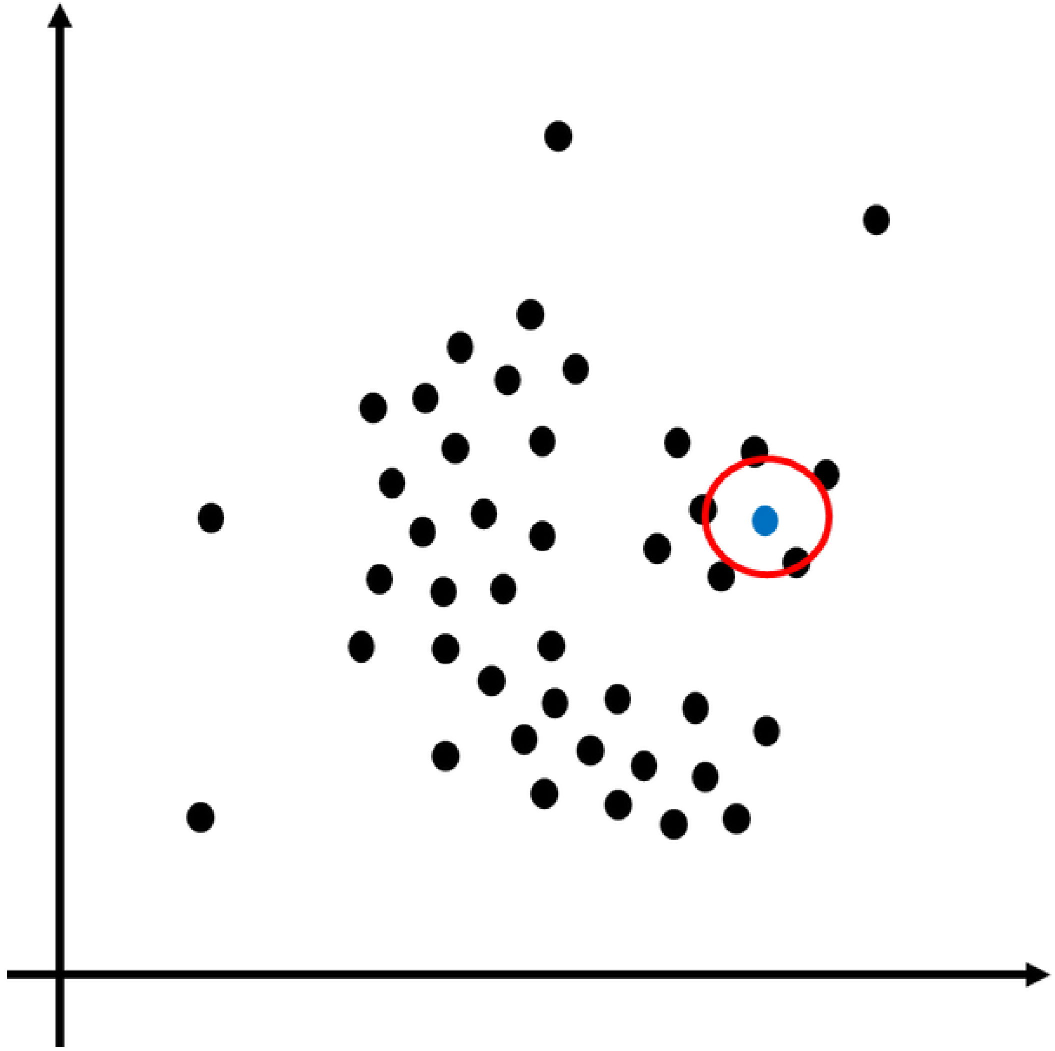
For this example, eps will be represented by the red circle, and min_samples will be 4. NOTE – min_samples includes the point currently being classified.



The DBSCAN algorithm inspects each data point in the dataset. For brevity, we'll only consider a few points to illustrate how the algorithm works.

Let's say the next point to be evaluated is circled.The total number of points within, or contacted by, the eps circle is 5.

This data point is **core**. Core points will be designated by blue.

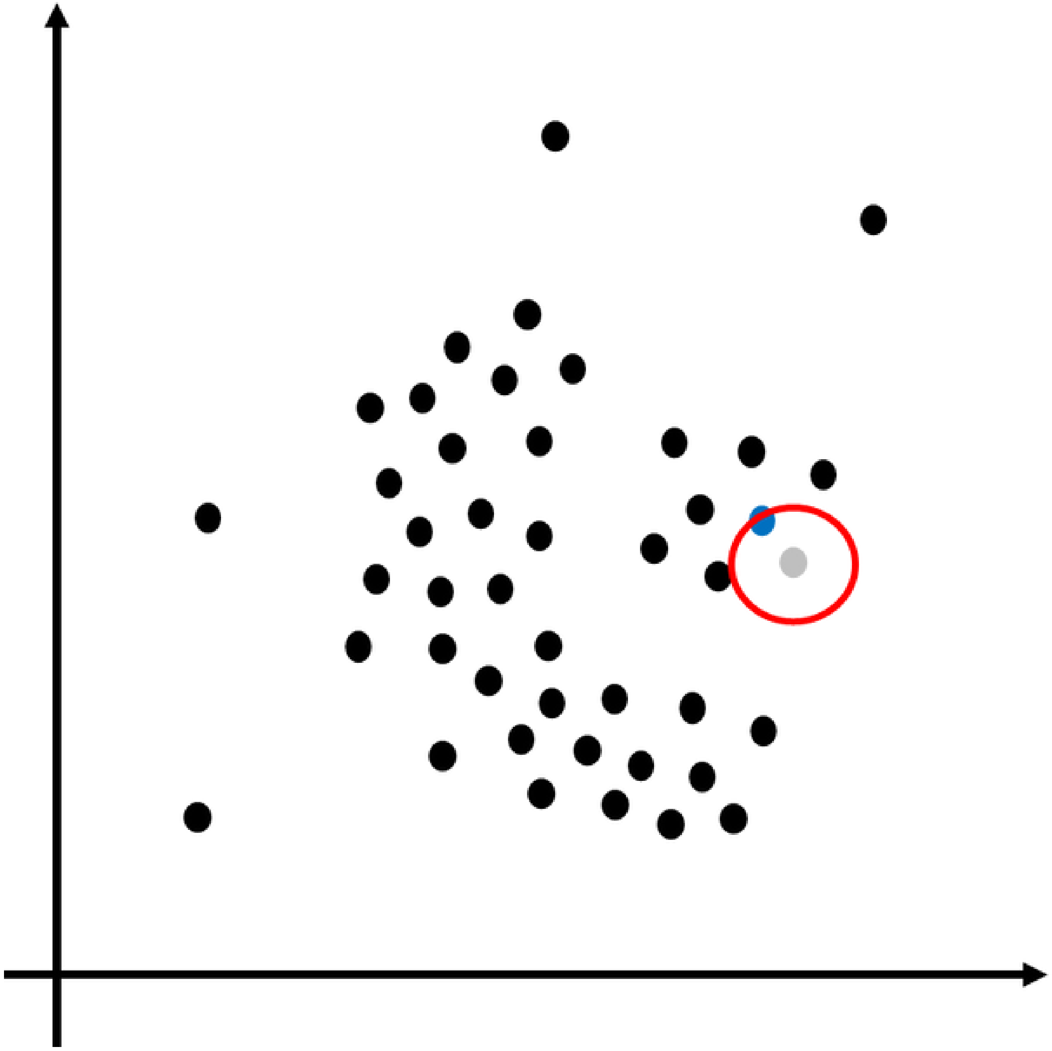


DBSCAN CLUSTERING

The next data point to be classified is in the center of the eps circle. The number of points within, or touched by, the eps circle is 3.

This point is not core but is within eps distance of a core point. Therefore, it is a **border** point.

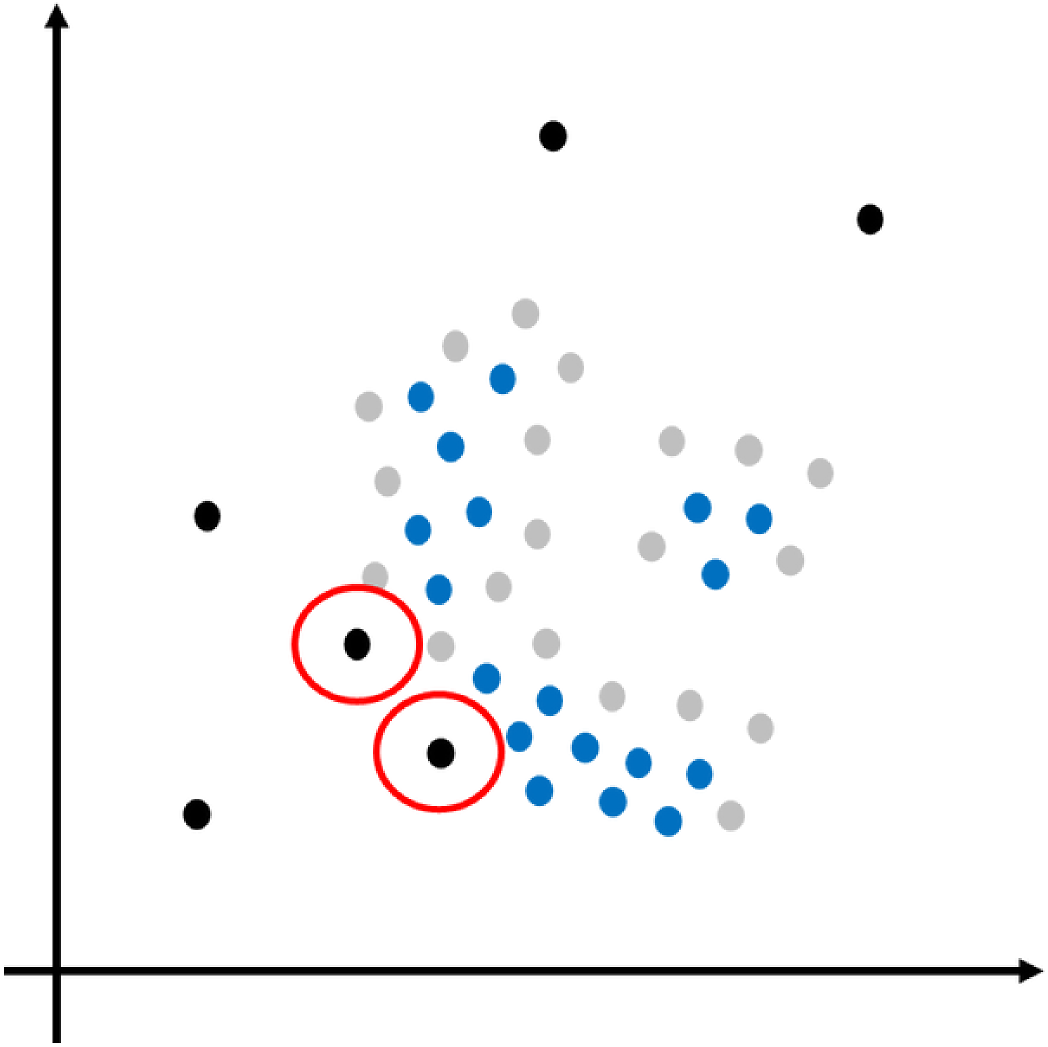
Border points will be designated by gray.



DBSCAN CLUSTERING

DBSCAN continues the process with all the data points in the data set.

What is depicted here are the results of applying the DBSCAN definitions to the dataset. Notice that two data points in addition to the four obvious noise points were identified.

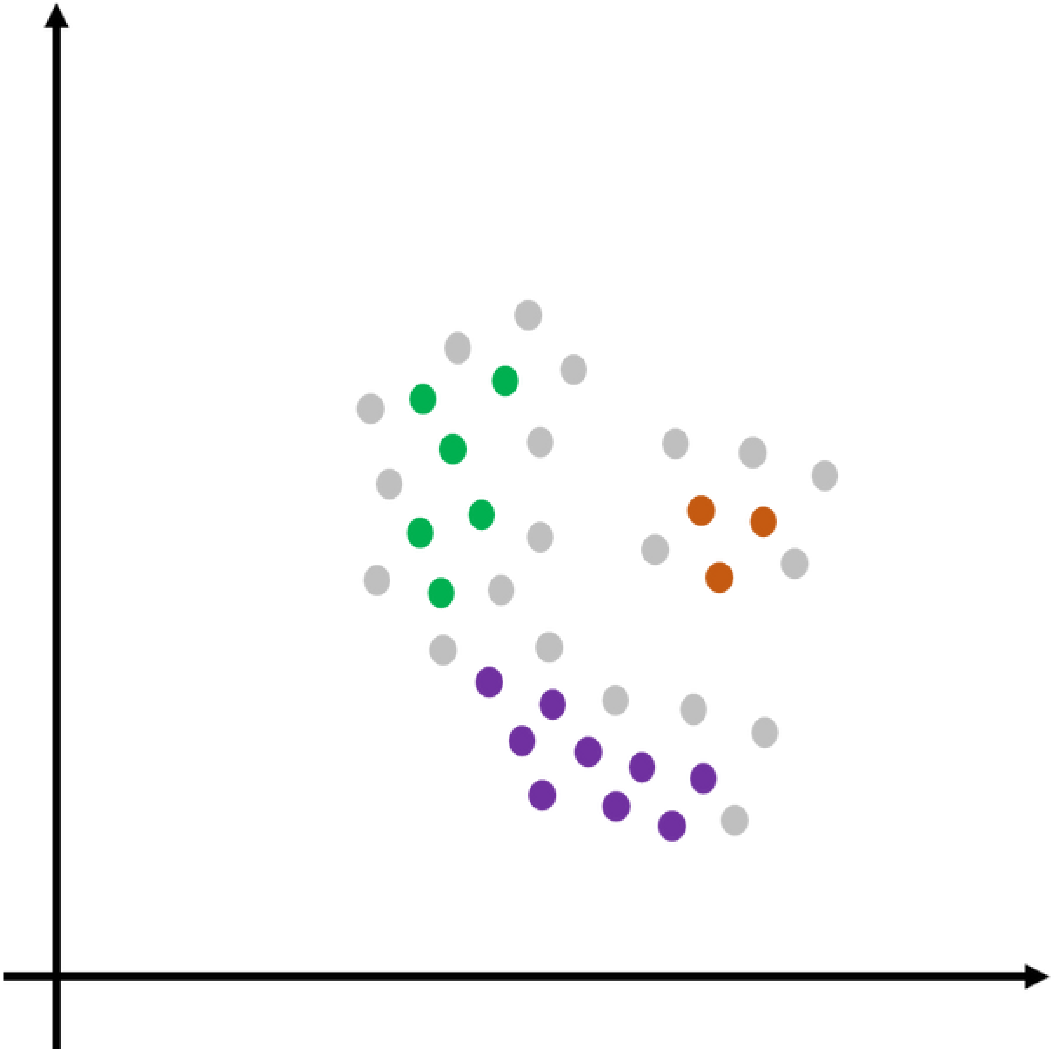


DBSCAN CLUSTERING

Next, DBSCAN eliminates noise points and then connects all core points with eps distance of each other.

Groups of connected core points become clusters.

Notice how the found clusters do not match what we tend to see?

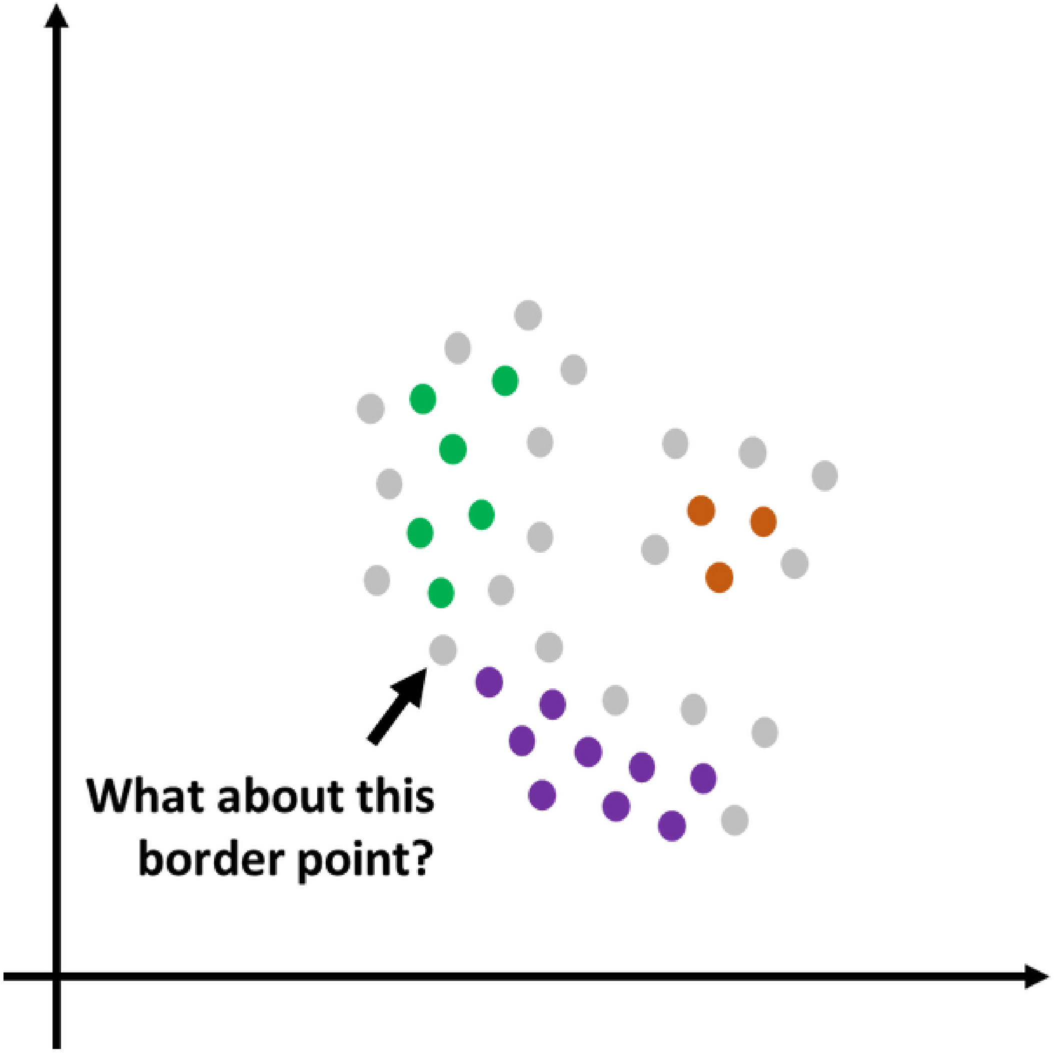


DBSCAN CLUSTERING

Last, DBSCAN assigns border points to clusters based on core point proximity. How does DBSCAN handle “ties”?

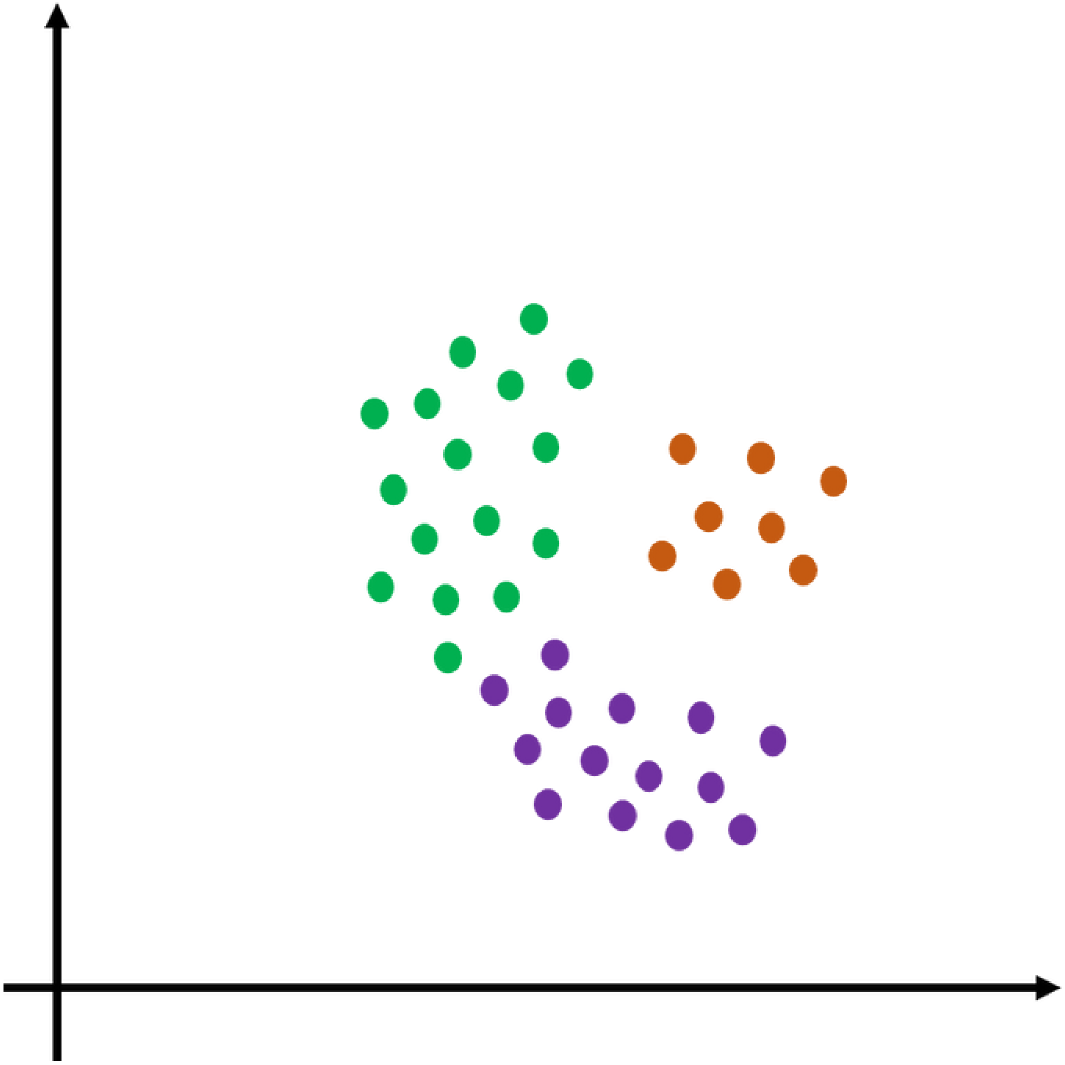
DBSCAN is greedy. When necessary, border points are assigned to the cluster that was created first.

Let’s say that the green cluster was created first.



DBSCAN CLUSTERING

Voila! DBSCAN has completed the clustering!



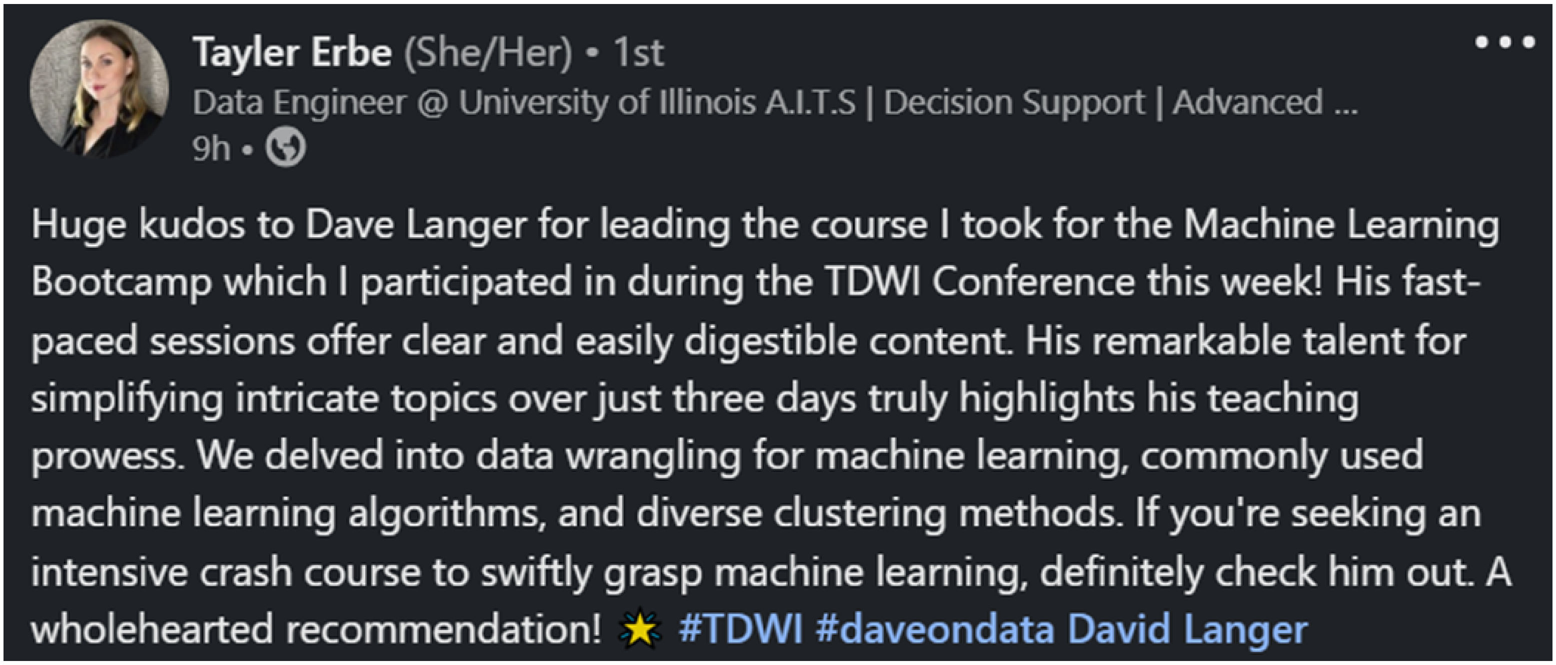
Jumpstart Your Data Science Skills

The content in this document comes from the following live training course:

- Cluster Analysis for Data Science with Python

This is one of 3 classroom experiences I will deliver as part of the TDWI-certified Machine Learning Bootcamp this February. These courses include 14 hands-on Python labs to jumpstart your ML skills.

Be sure to check with your manager. TDWI is an approved training vendor for many organizations.



Tayler Erbe (She/Her) • 1st

Data Engineer @ University of Illinois A.I.T.S | Decision Support | Advanced ...

9h •

Huge kudos to Dave Langer for leading the course I took for the Machine Learning Bootcamp which I participated in during the TDWI Conference this week! His fast-paced sessions offer clear and easily digestible content. His remarkable talent for simplifying intricate topics over just three days truly highlights his teaching prowess. We delved into data wrangling for machine learning, commonly used machine learning algorithms, and diverse clustering methods. If you're seeking an intensive crash course to swiftly grasp machine learning, definitely check him out. A wholehearted recommendation! ⭐ #TDWI #daveondata David Langer

https://bit.ly/TDWILasVegas2024MLBootcamp

How it Works

The following three live training courses are being offered at the TDWI Las Vegas conference on February 20-22:

- Machine Learning with Python Made Easy
- Data Wrangling for Machine Learning with Python
- Introduction to Cluster Analysis for Data Science with Python

Attending all three courses will earn a TDWI machine learning bootcamp certificate.

These hands-on courses are designed to quickly build machine learning (ML) skills.

Previous attendees have used these skills immediately after returning to work from the conference.

No experience with Python? No problem!

You will receive free access to a 4-hour Python Quick Start online tutorial.

14 Hands-On Labs to Build Your Skills

Feb 20th – Machine Learning with Python Made Easy

- Lab 1 – Decision Trees
- Lab 2 – Random Forests
- Lab 3 – Feature Engineering
- Lab 4 – Model Testing
- Lab 5 – Model Improvement

Feb 21st - Data Wrangling for Machine Learning with Python

- Lab 1 – Data Profiling
- Lab 2 – The Mighty pandas
- Lab 3 – Wrangling Strings
- Lab 4 – Joining Data

Feb 22nd – Cluster Analysis for Data Science with Python

- Lab 1 – K-Means Clustering
- Lab 2 – Optimizing K-Means
- Lab 3 – Optimizing DBSCAN
- Lab 4 – Dimensionality Reduction
- Lab 5 – Categorical Data

Wait! There's More!

I'm pleased to announce that TDWI will offer my text analytics course using Python in addition to the ML Bootcamp.

This course will provide you with hands-on skills to transform raw text data into a format suitable for clustering and predictive models.

Feb 23rd – Text Analytics for Data Science with Python

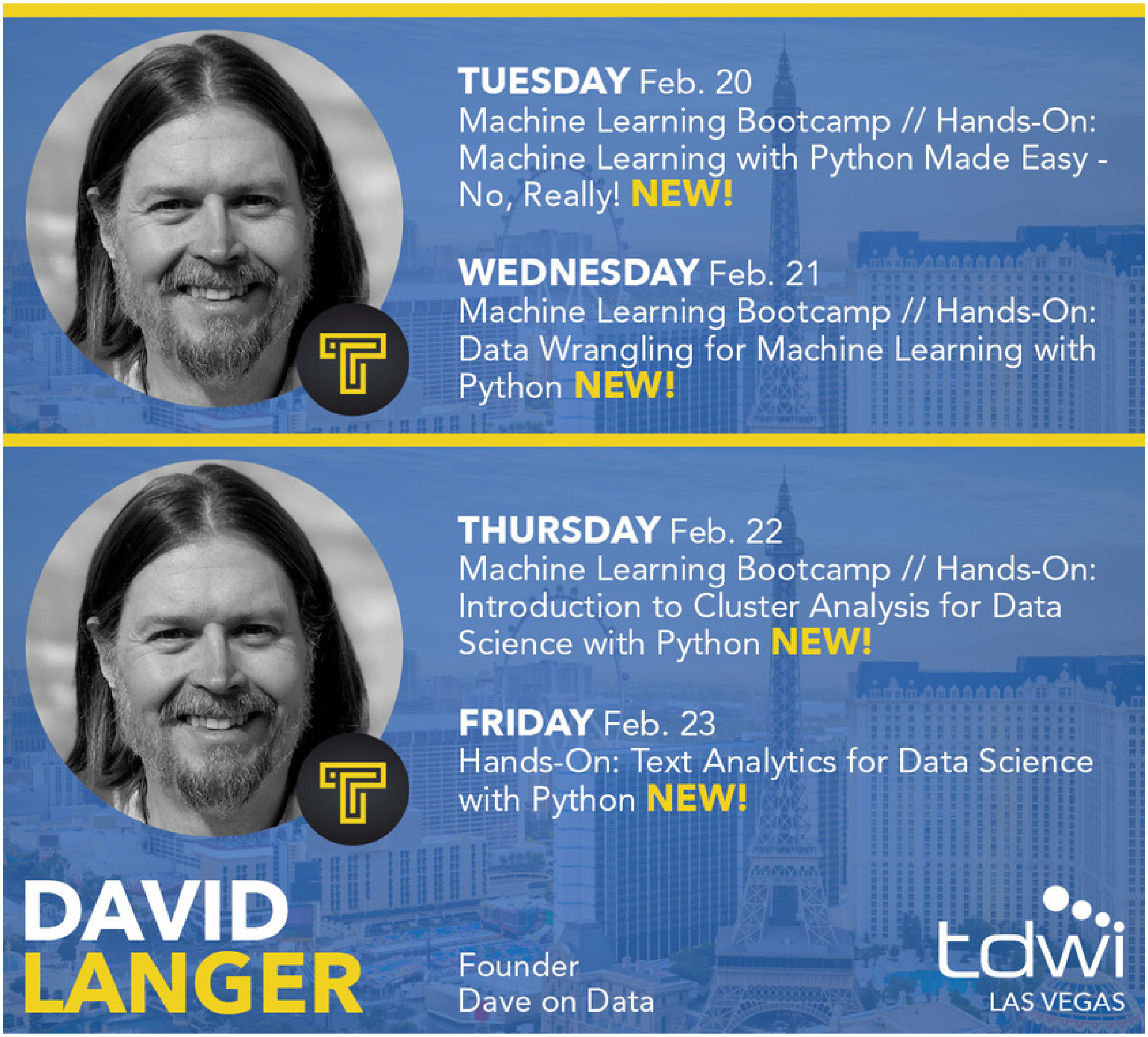
- Lab 1 – Tokenization
- Lab 2 – Token Normalization
- Lab 3 – The Vector Space Model
- Lab 4 – Clustering Documents
- Lab 5 – Classifying Documents with ML

Text data is everywhere in modern organizations (e.g., customer service chats).

Are you ready to learn the skills to analyze it?

Register Now. Seats are Limited.

Through January 19th, **you can save up to $525** off all four hands-on machine learning courses. Use promo code **LANGER** to get all the savings.



TUESDAY Feb. 20
Machine Learning Bootcamp // Hands-On: Machine Learning with Python Made Easy - No, Really! **NEW!**

WEDNESDAY Feb. 21
Machine Learning Bootcamp // Hands-On: Data Wrangling for Machine Learning with Python **NEW!**

THURSDAY Feb. 22
Machine Learning Bootcamp // Hands-On: Introduction to Cluster Analysis for Data Science with Python **NEW!**

FRIDAY Feb. 23
Hands-On: Text Analytics for Data Science with Python **NEW!**

DAVID LANGER

Founder
Dave on Data

tdwi
LAS VEGAS

https://bit.ly/TDWILasVegas2024MLBootcamp

About the Author

My name is Dave Langer and I am the founder of Dave on Data.

I'm a hands-on analytics professional, having used my skills with Excel, SQL, and R/Python to craft insights, advise leaders, and shape company strategy.

I'm also a skilled educator, having trained 100s of working professionals in live in-person classroom settings and 1000s more via live virtual training and online courses.

In the past, I've held analytics leaderships roles at Schedulicity, Data Science Dojo, and Microsoft.

Drop me an email if you have any questions:
dave@daveondata.com