Representing Insights Obtained from Data

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Plotting continuous data

Representing categorical data

Text and image data

Azure Data Studio for modeling

Power BI for visualization

Data Used in Analysis

Continuous

Categorical

All other forms of data, such as text and image data, must be converted to one of these forms

Continuous vs. Categorical Data

Continuous	Categorical
E.g. height or weight of individuals	E.g. day of week, month of year, gender, letter grade
Can take any value	Finite set of permissible values
Predicted using regression models	Predicted using classification models
Always can be sorted on magnitude	Categories may or may not be sortable

Types of Categorical Data

Binary: Only two permissible values

Multi-class: Multiple permissible values

Nominal: No ordering possible

Ordinal: Ordering possible

Text Data

```
d = “This is not the worst restaurant in the metropolis,
not by a long way”
```

Document as Word Sequence

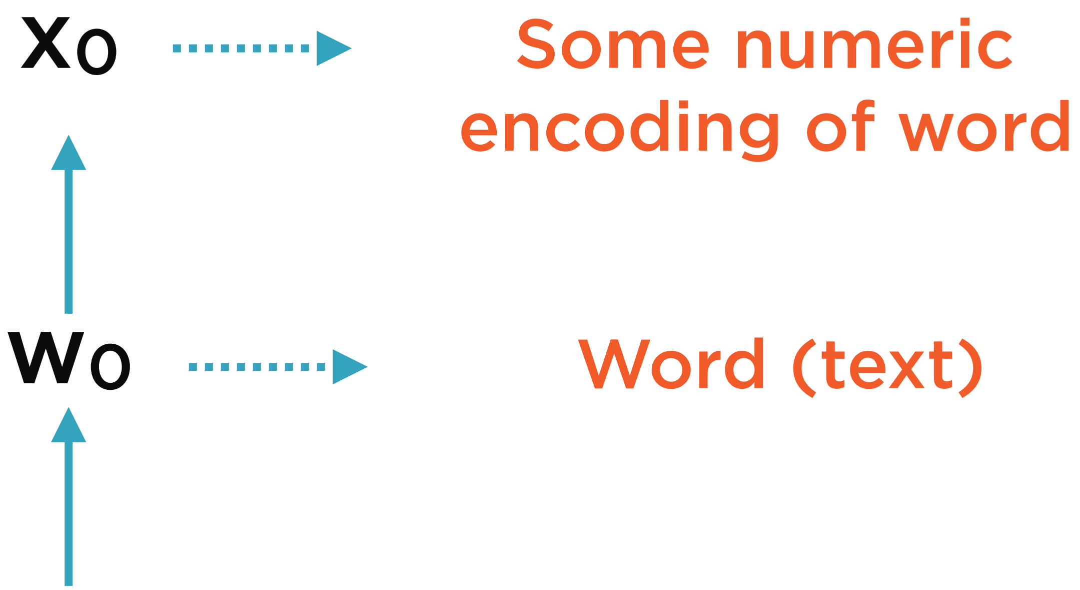
Model a document as an ordered sequence of words

```
d = “This is not the worst restaurant in the metropolis,
not by a long way”

(“This”, “is”,”not”,”the”,”worst”,”restaurant”,”in”,”the”,
“metropolis”, “not”,”by”,”a”,”long”,”way”)
```

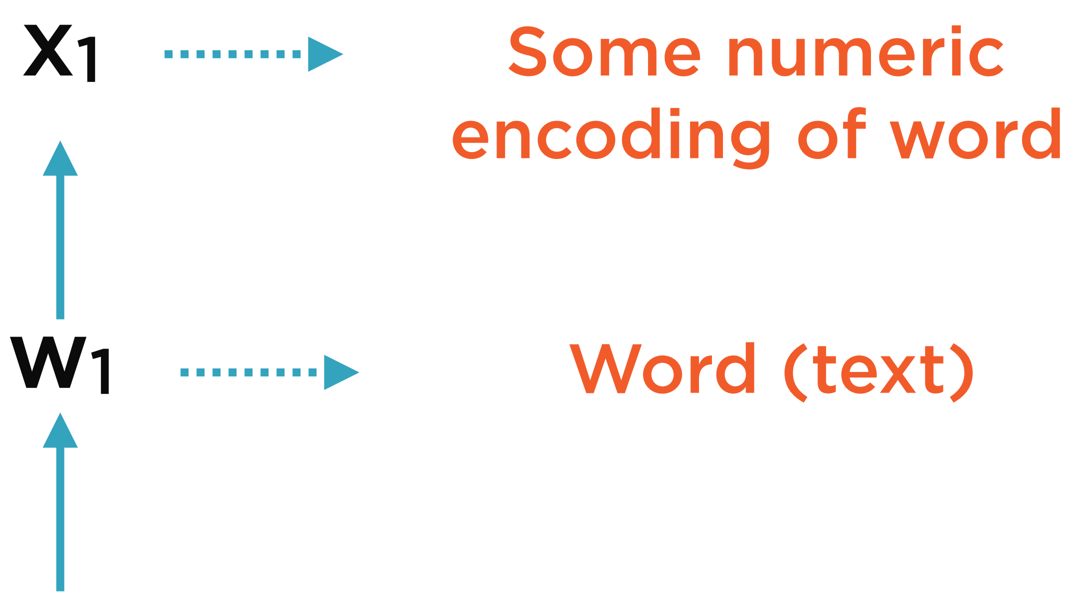
Document as Word Sequence

Tokenize document into individual words



(“This”, “is”,”not”,”the”,”worst”,”restaurant”,”in”,”the”,
“metropolis”, “not”,”by”,”a”,”long”,”way”)

Represent Each Word as a Number



x_1

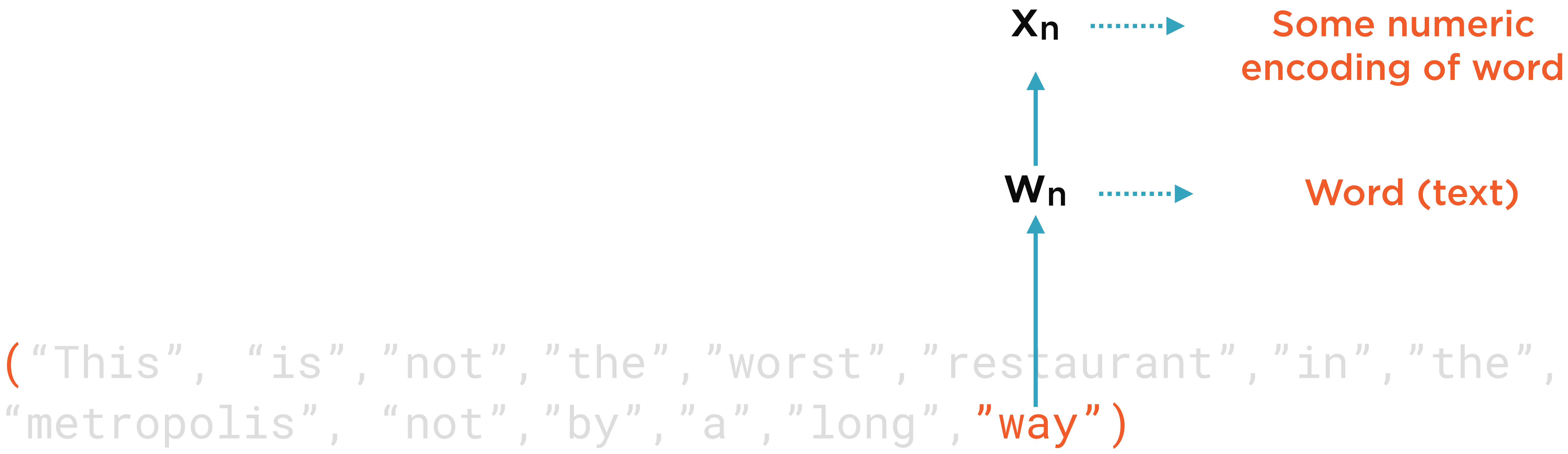
Some numeric encoding of word

w_1

Word (text)

(“This”, “is”,”not”,”the”,”worst”,”restaurant”,”in”,”the”, “metropolis”, “not”,”by”,”a”,”long”,”way”)

Represent Each Word as a Number



x_n Some numeric encoding of word

w_n Word (text)

(“This”, “is”,”not”,”the”,”worst”,”restaurant”,”in”,”the”, “metropolis”, “not”,”by”,”a”,”long”,”way”)

Represent Each Word as a Number

$$d = [x_0,\ x_1,\ \dots\ x_n]$$

Document as Tensor

Represent each word as numeric data, aggregate into tensor

Numeric Representations of Text

One-hot

Frequency-based

Prediction-based

Numeric Representations of Text

One-hot

Frequency-based

Prediction-based

Represent each word in text by its presence or absence

Numeric Representations of Text

One-hot

Frequency-based

Prediction-based

Frequency-based Embeddings

Count

TF-IDF

Co-occurrence

Frequency-based Embeddings

Count

TF-IDF

Co-occurrence

Capture how often a word occurs in a document i.e. the **counts** or the **frequency**

Frequency-based Embeddings

Count

TF-IDF

Co-occurrence

Captures how often a word occurs in a **document** as well as the **entire corpus**

Tf-Idf

Frequently in a single document

Might be important

Frequently in the corpus

Probably a common word like “a”, “an”, “the”

Frequency-based Embeddings

Count

TF-IDF

Co-occurrence

Similar words will occur together and will have similar context

Context Window

A window centered around a word, which includes a certain number of neighboring words

Co-occurrence

The number of times two words *w1* and *w2* have occurred together in a context window

Word Embeddings

One-hot

Frequency-based

Prediction-based

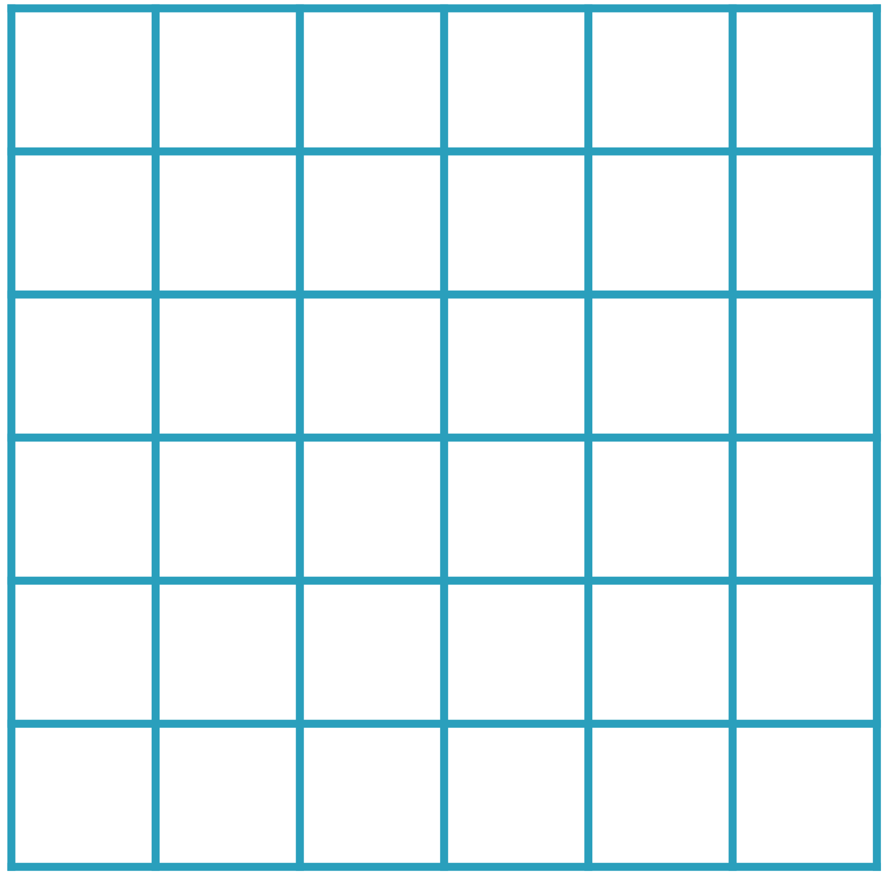
Predictions-based embeddings

Numerical representations of text which capture meanings and semantic relationships, generated using ML models

Image Data

Images as Matrices

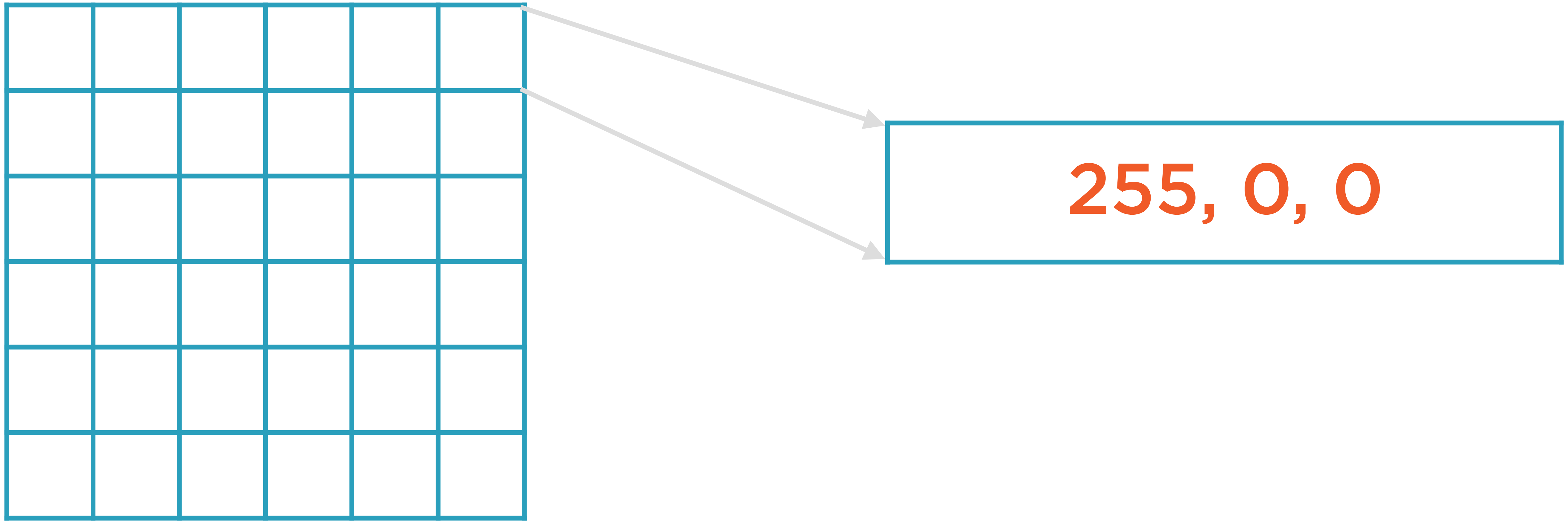
RGB Images



RGB values are for color images

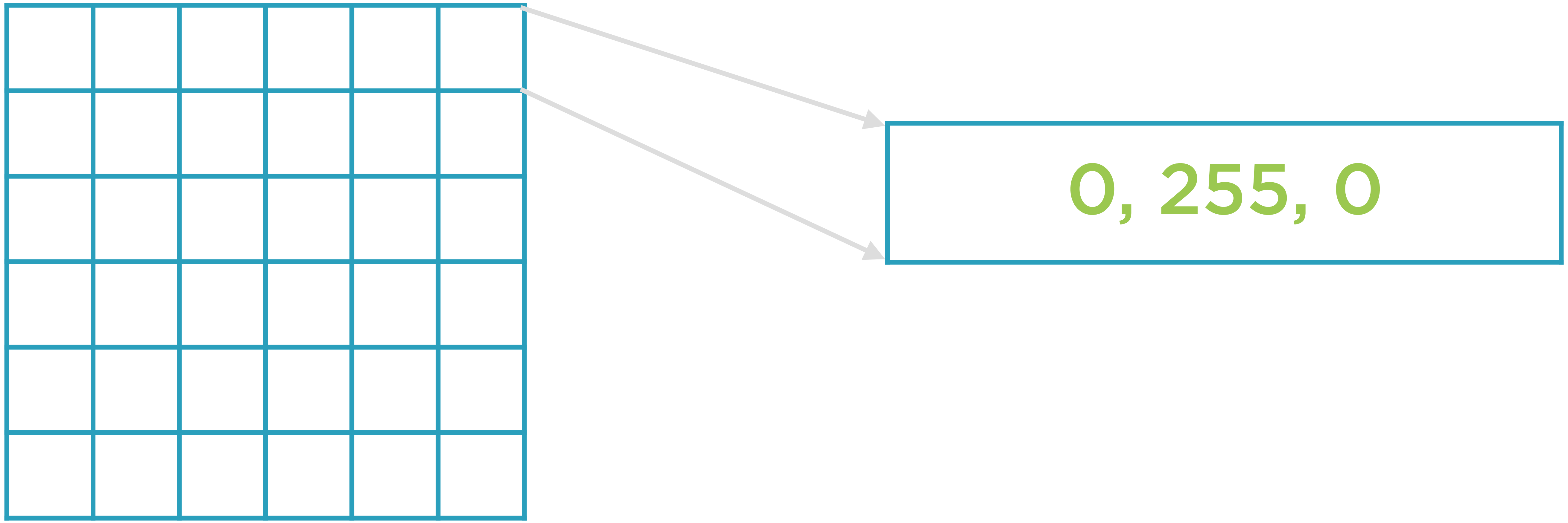
R, G, B: 0-255

RGB Images



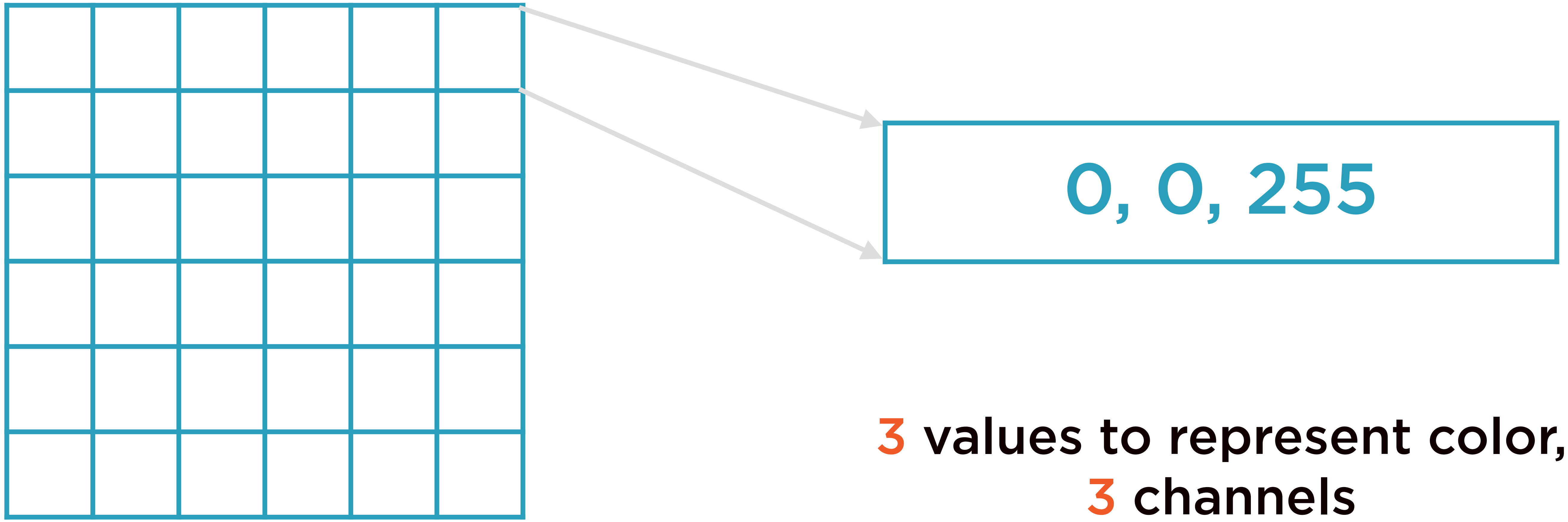
255, 0, 0

RGB Images



0, 255, 0

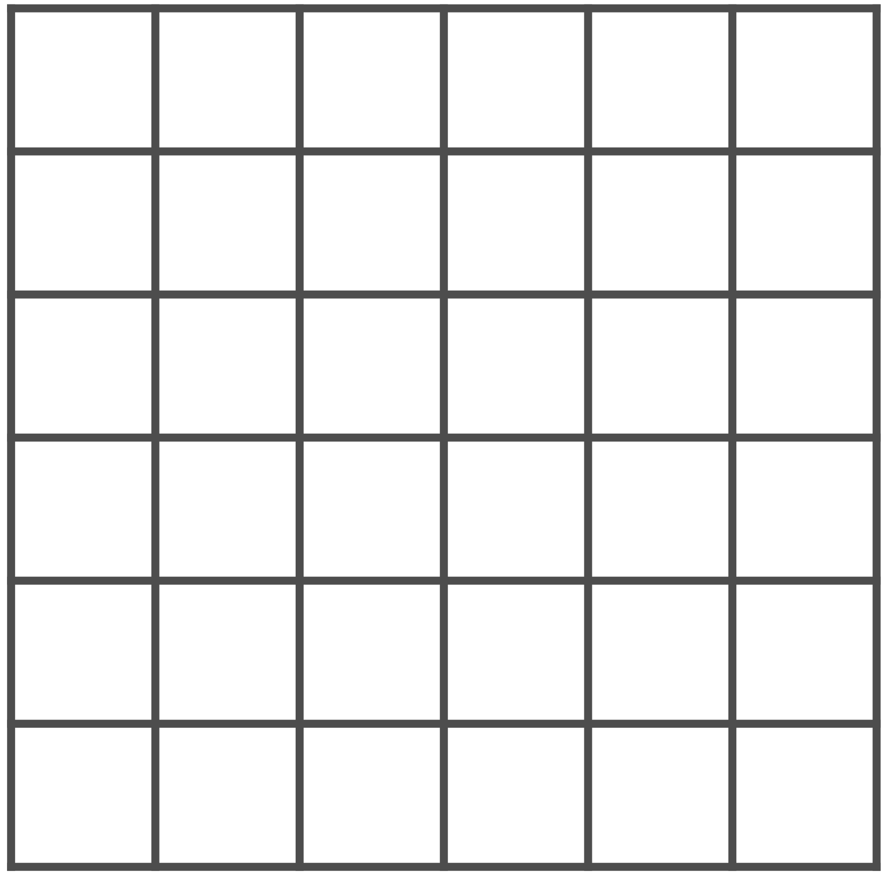
RGB Images



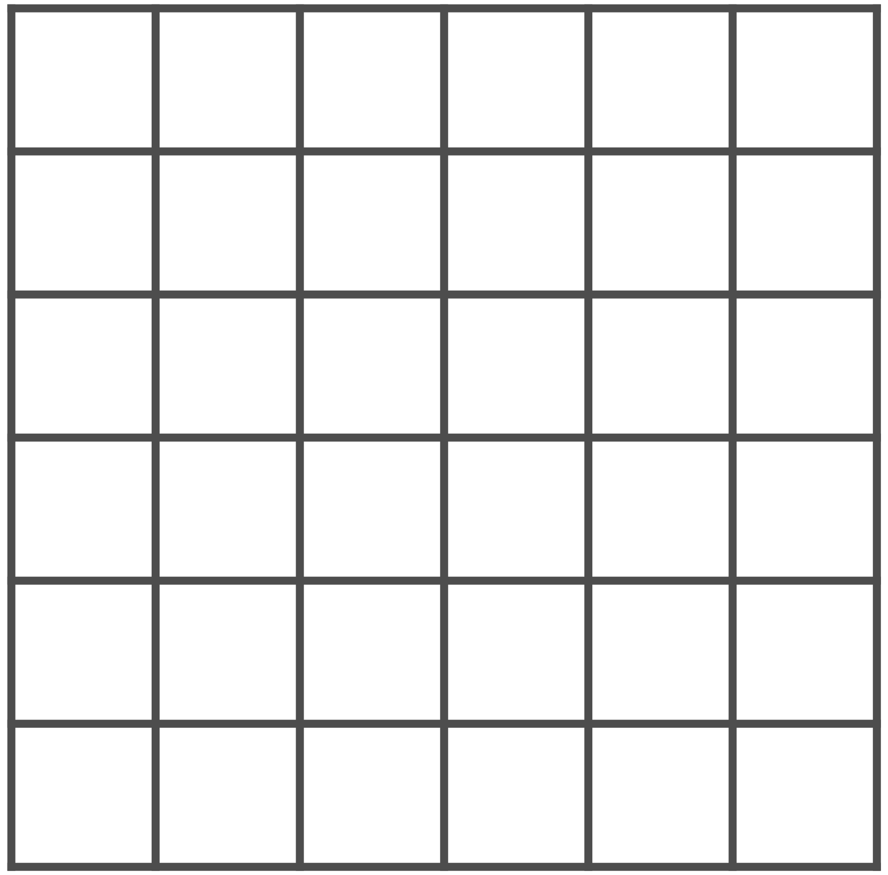
0, 0, 255

3 values to represent color,
3 channels

Grayscale Images



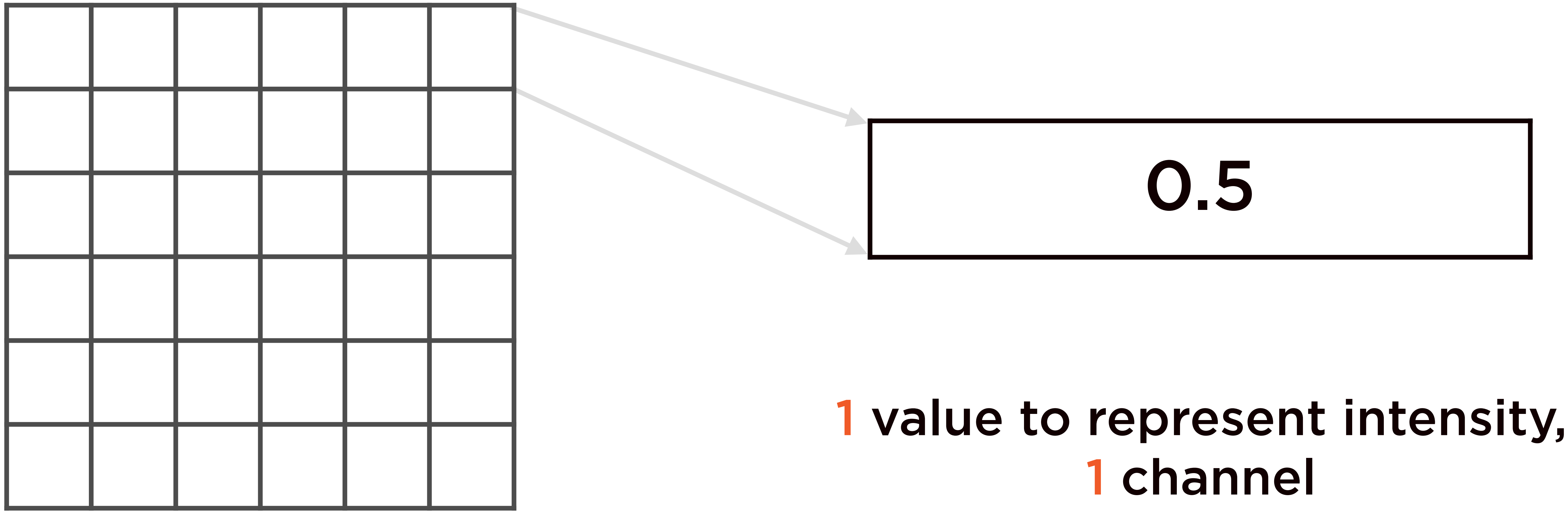
Grayscale Images



Each pixel represents only intensity information

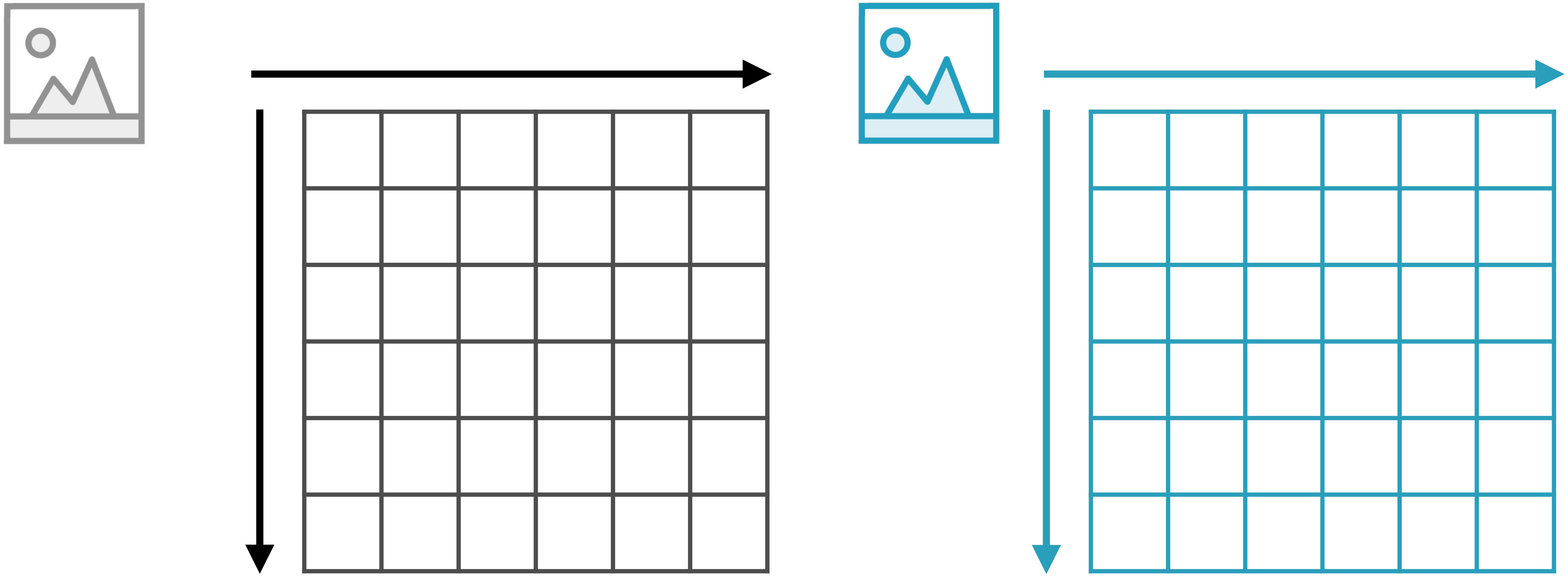
0.0 - 1.0

Grayscale Images



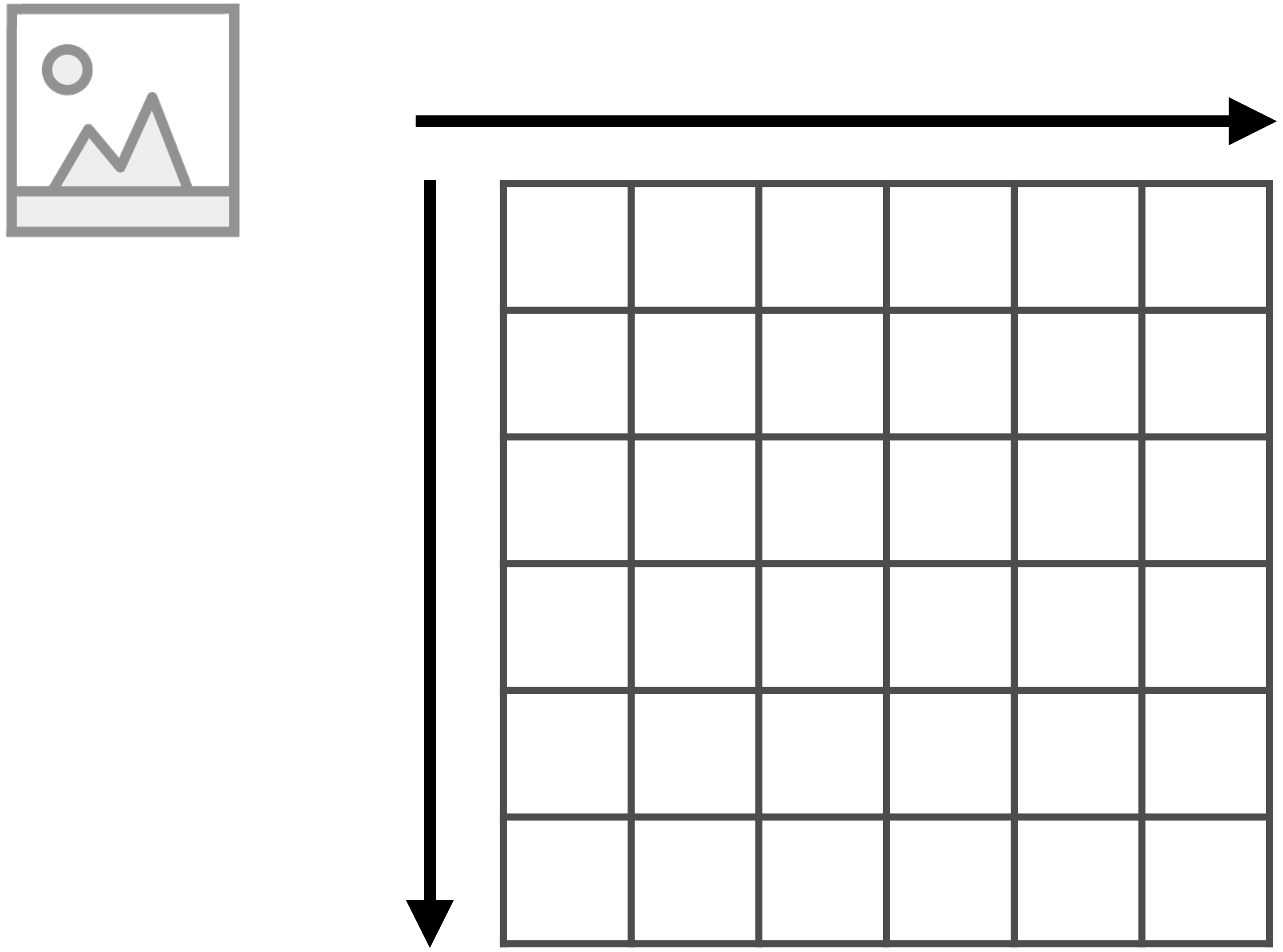
1 value to represent intensity,
1 channel

Images as Matrices

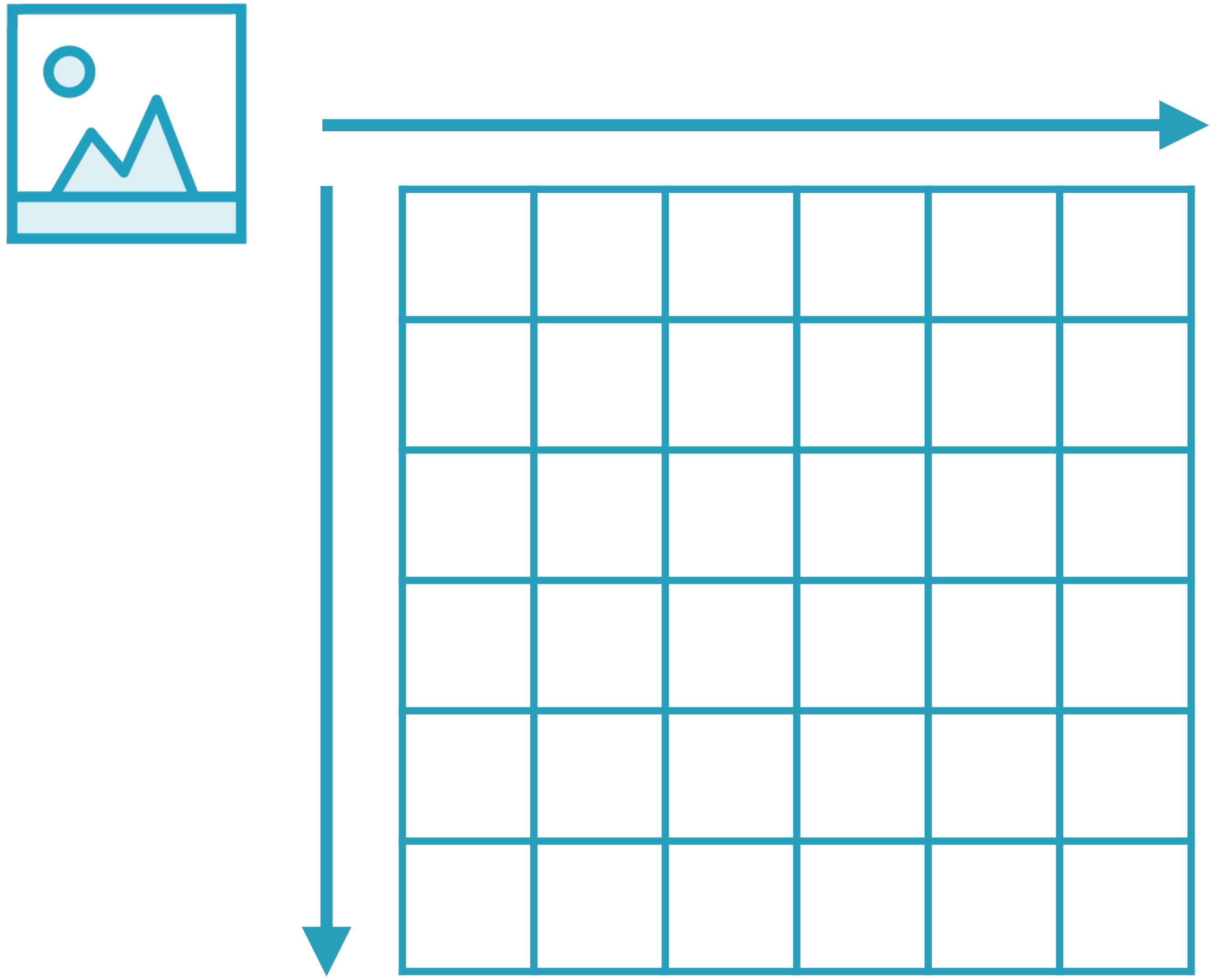


Images can be represented by a 3-D matrix

Images as Matrices



(6, 6, 1)



(6, 6, 3)

List of Images

$(10, 6, 6, \boxed{3})$

The number of channels

List of Images

$$(10, \boxed{6,\ 6}, 3)$$

The height and width of each image in the list

List of Images

$$(\boxed{10,}\ 6, 6, 3)$$

The number of images

Interacting with Azure SQL Database

SQL Server Management Studio (SSMS)

Azure Data Studio

SQL Server Management Studio

Microsoft’s very popular integrated environment for all SQL services, including SQL Server, Azure SQL Database, and SQL Data Warehouse. Old favorite of DBAs.

Azure Data Studio

Microsoft’s integrated environment for querying and visualizing data on Azure as well as on-premise. Designed for data professionals rather than DBAs.

SSMS vs. Azure Data Studio

SQL Server Management Studio	Azure Data Studio
For database professionals	For data professionals
Focus on database management	Focus on querying and visualization
Extensive wizards	Few wizards
Available only for Windows	Available for Windows, Mac and Linux
Little emphasis on command-line	For power-users of sqlcmd or Powershell

Power BI

Business analytics app with powerful visualization and data exploration capabilities; closely integrated with Microsoft and Azure data services.

Demo

Querying and visualizing data using Azure Data Studio

Demo

Visualizing data using Power BI

Summary

Plotting continuous data

Representing categorical data

Text and image data

Azure Data Studio for modeling

Power BI for visualization

Related Courses

Summarizing Data and Deducing Probabilities

Experimental Design for Data Analysis