Mechanistic Interpretability 101

Episode 1

Linear Algebra

Data Representation

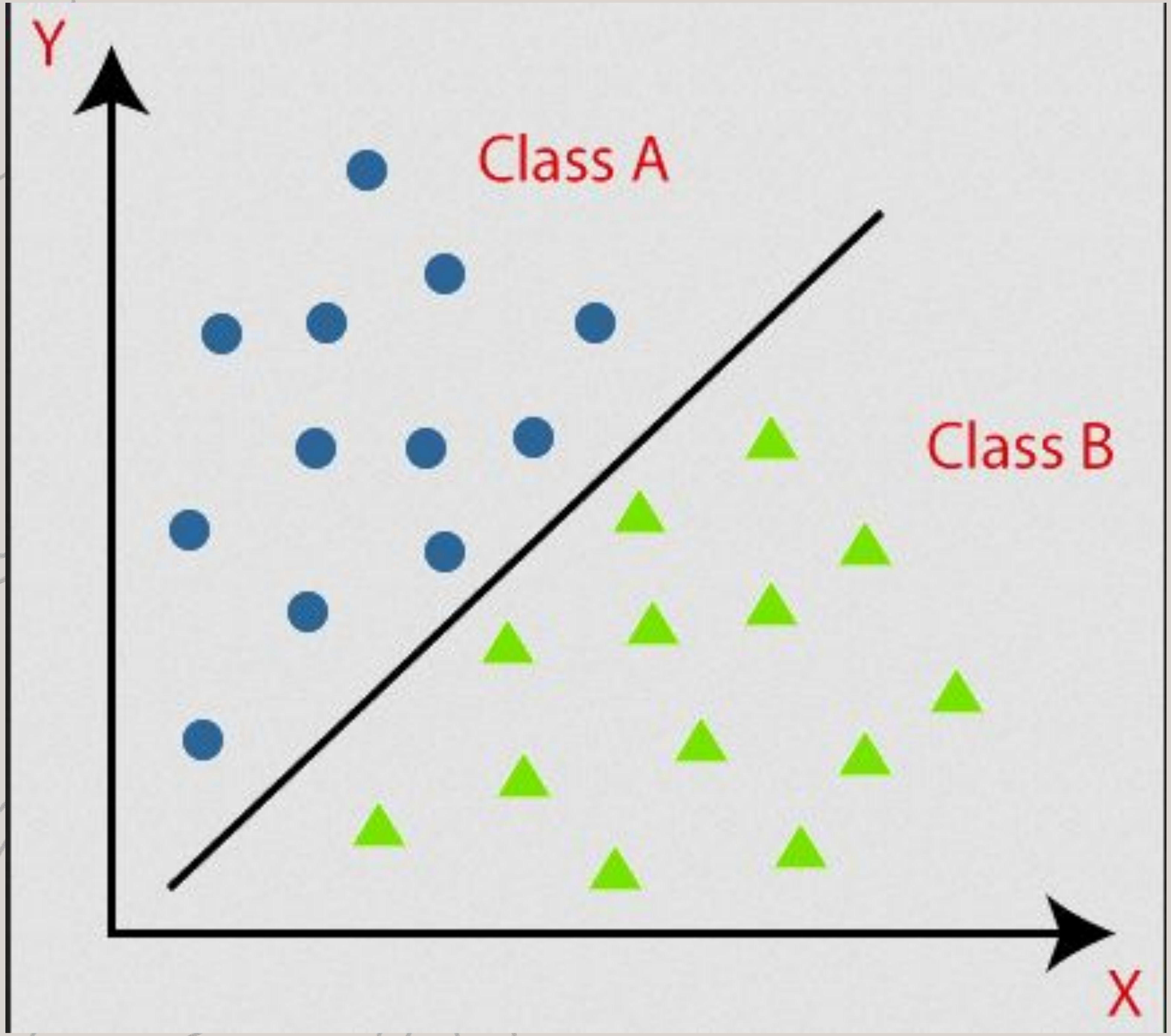


Image Source: Wikipedia

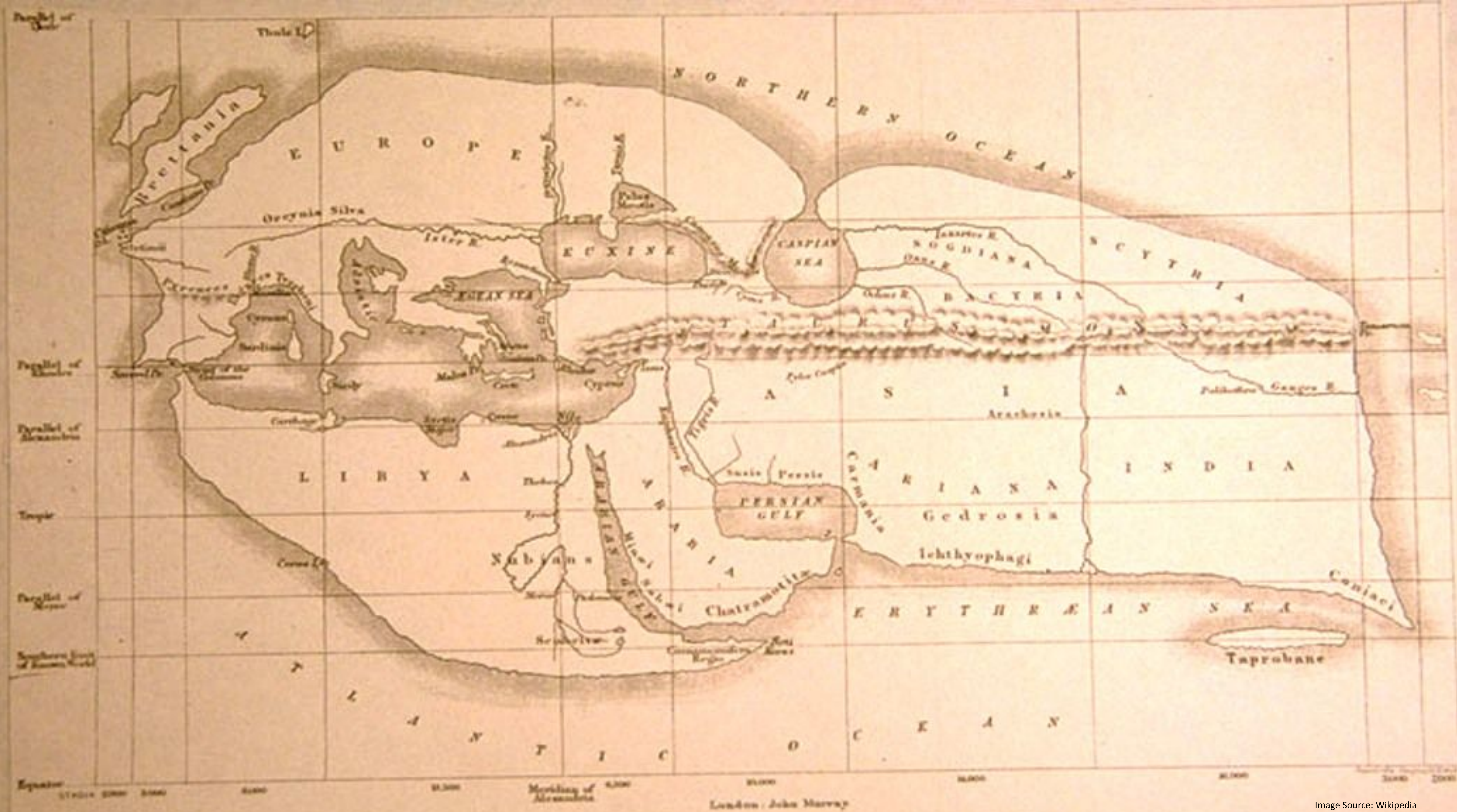
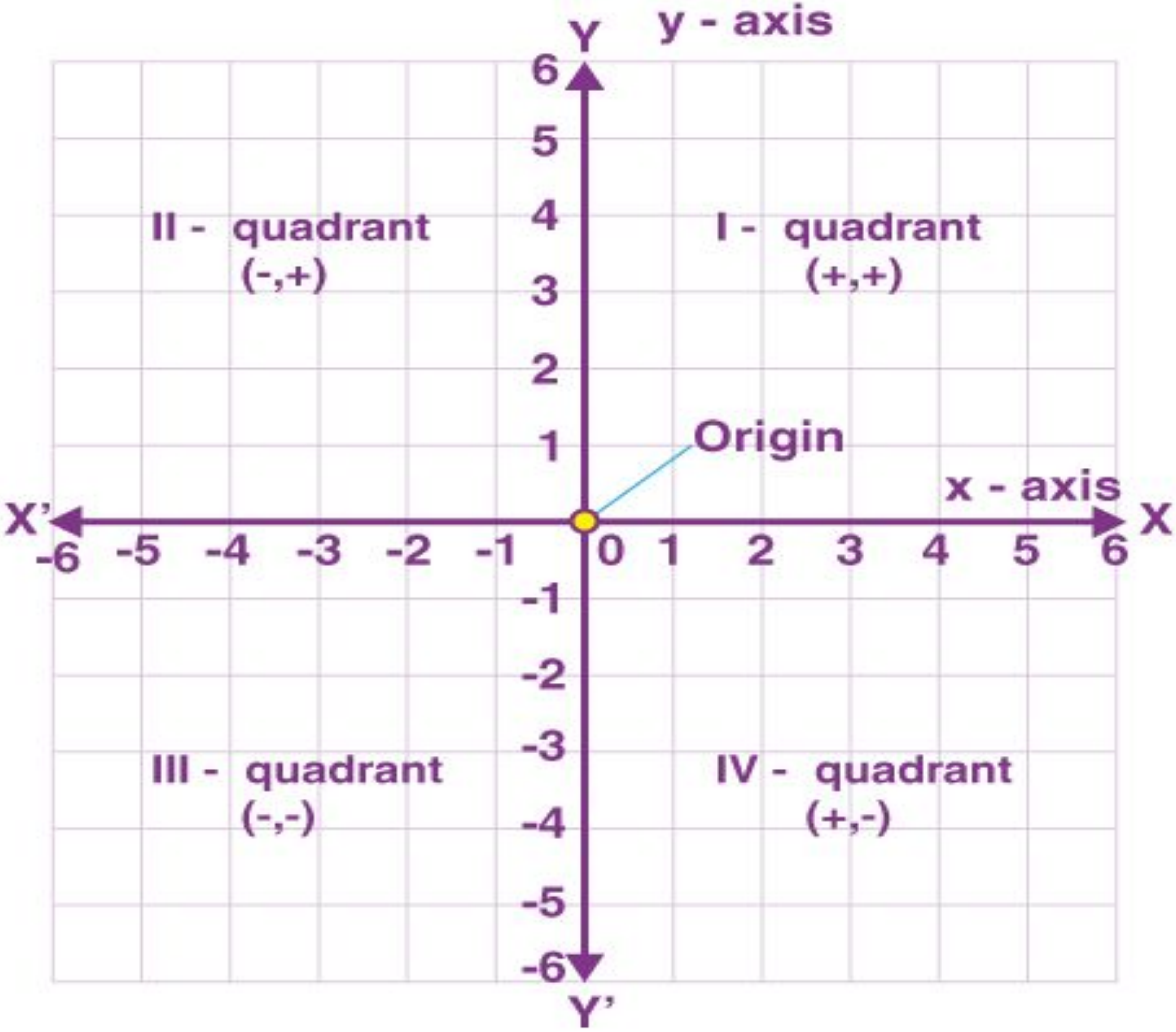


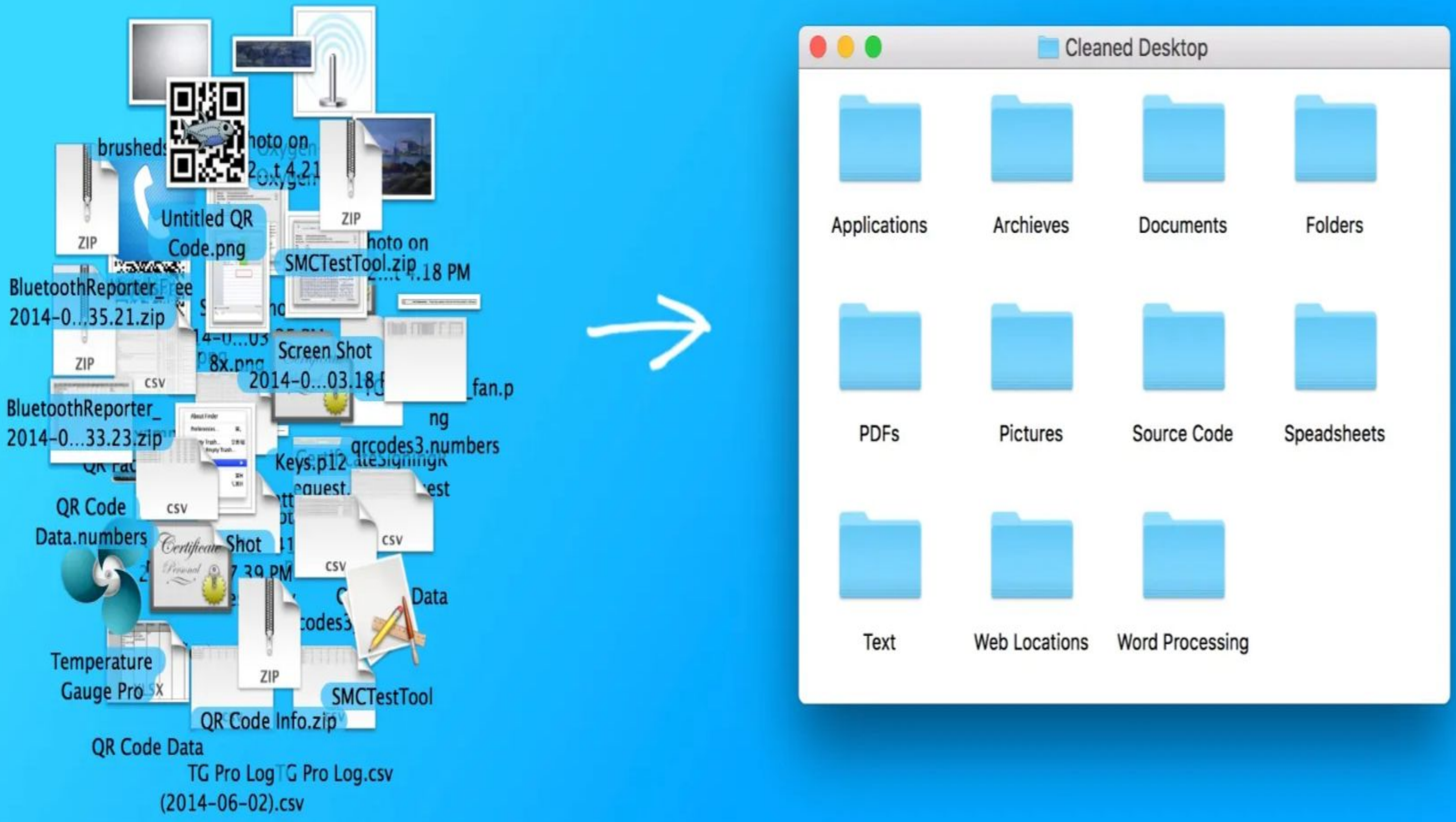
Image Source: Wikipedia



Cartesian Coordinate System

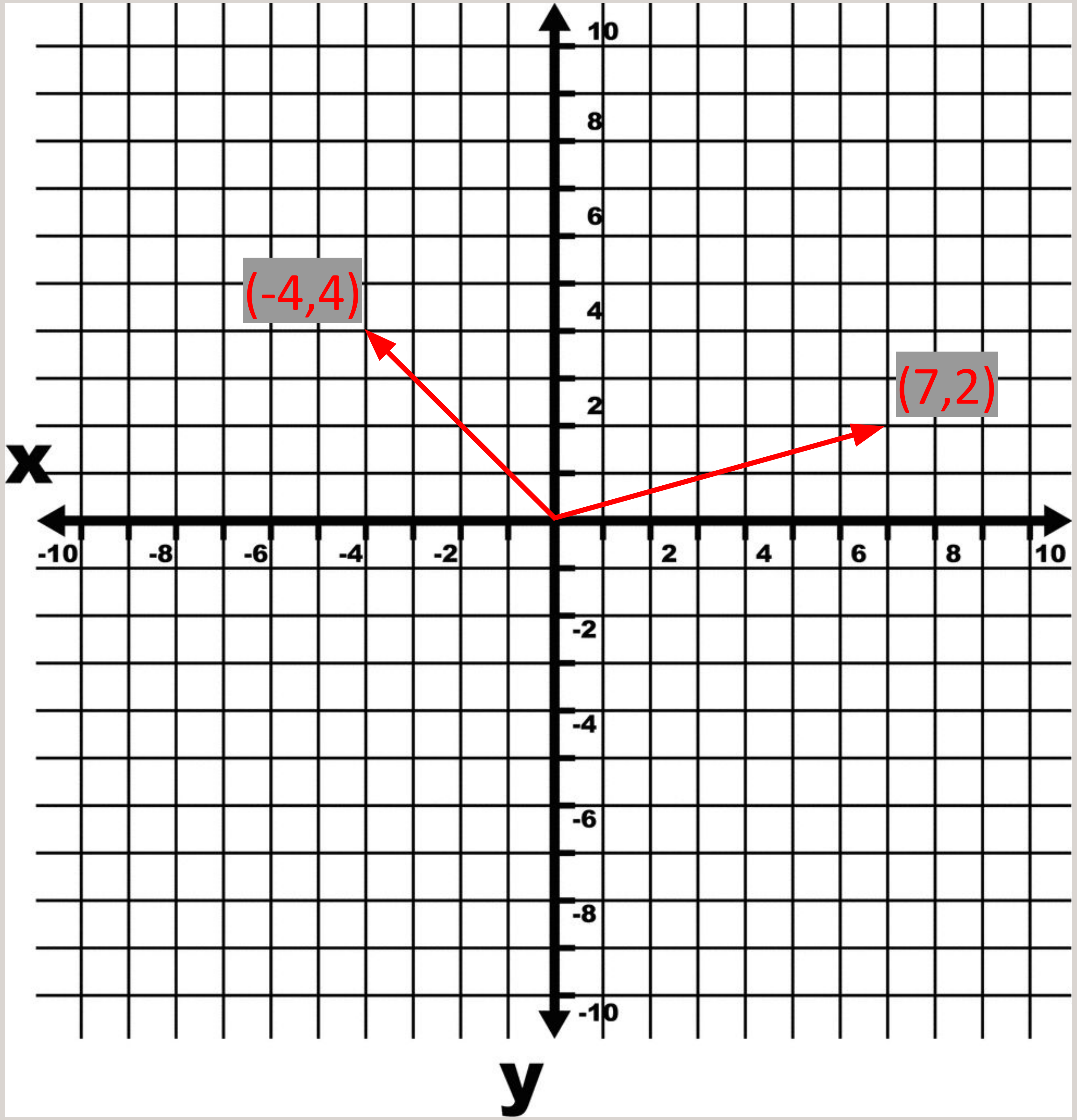
René Descartes

Image Source: Wikipedia



Untitled QR Code.png
SMCTestTool.zip
BluetoothReporter_Free 2014-0...35.21.zip
Screen Shot 2014-0...03.18
BluetoothReporter_ 2014-0...33.23.zip
qrcodes3.numbers
Keys.p12
QR Code Data.numbers
Temperature Gauge Pro
SMCTestTool
QR Code Info.zip
QR Code Data
TG Pro Log
TG Pro Log.csv
(2014-06-02).csv
Cleaned Desktop
Applications
Archieves
Documents
Folders
PDFs
Pictures
Source Code
Speadsheets
Text
Web Locations
Word Processing

Vectors and Matrices



$$M = \begin{bmatrix} 7 & -4 \\ 2 & 4 \end{bmatrix}$$

Vector Space

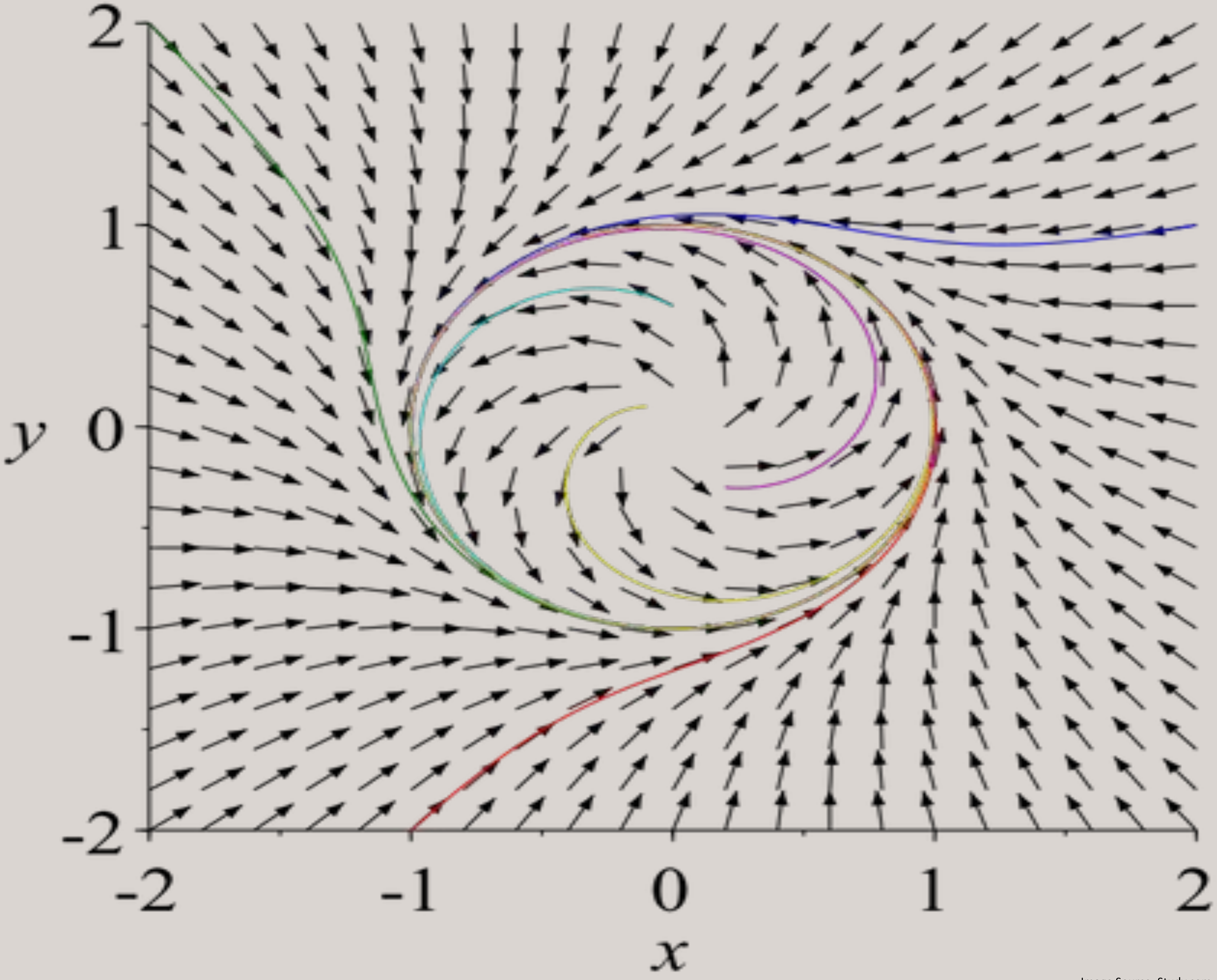


Image Source: Study.com

Vector Spaces

Addition:

1. $\mathbf{u} + \mathbf{v}$ is in V.	Closure under addition
2. $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$	Commutative property
3. $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$	Associative property
4. V has a **zero vector** $\mathbf{0}$ such that for every $\mathbf{u}$ in V, $\mathbf{u} + \mathbf{0} = \mathbf{u}$.	Additive identity
5. For every $\mathbf{u}$ in V, there is a vector in V denoted by $-\mathbf{u}$ such that $\mathbf{u} + (-\mathbf{u}) = \mathbf{0}$.	Additive inverse

Scalar Multiplication:

6. $c\mathbf{u}$ is in V.	Closure under scalar multiplication
7. $c(\mathbf{u} + \mathbf{v}) = c\mathbf{u} + c\mathbf{v}$	Distributive property
8. $(c + d)\mathbf{u} = c\mathbf{u} + d\mathbf{u}$	Distributive property
9. $c(d\mathbf{u}) = (cd)\mathbf{u}$	Associative property
10. $1(\mathbf{u}) = \mathbf{u}$	Scalar identity

Image Source: Math Stack Exchange

Image Source: 3 BLue 1 Brown

Linear Transformations

△ Folding

△ Squishing

△ Stretching

Image Source: Wikipedia

Linear Transformations

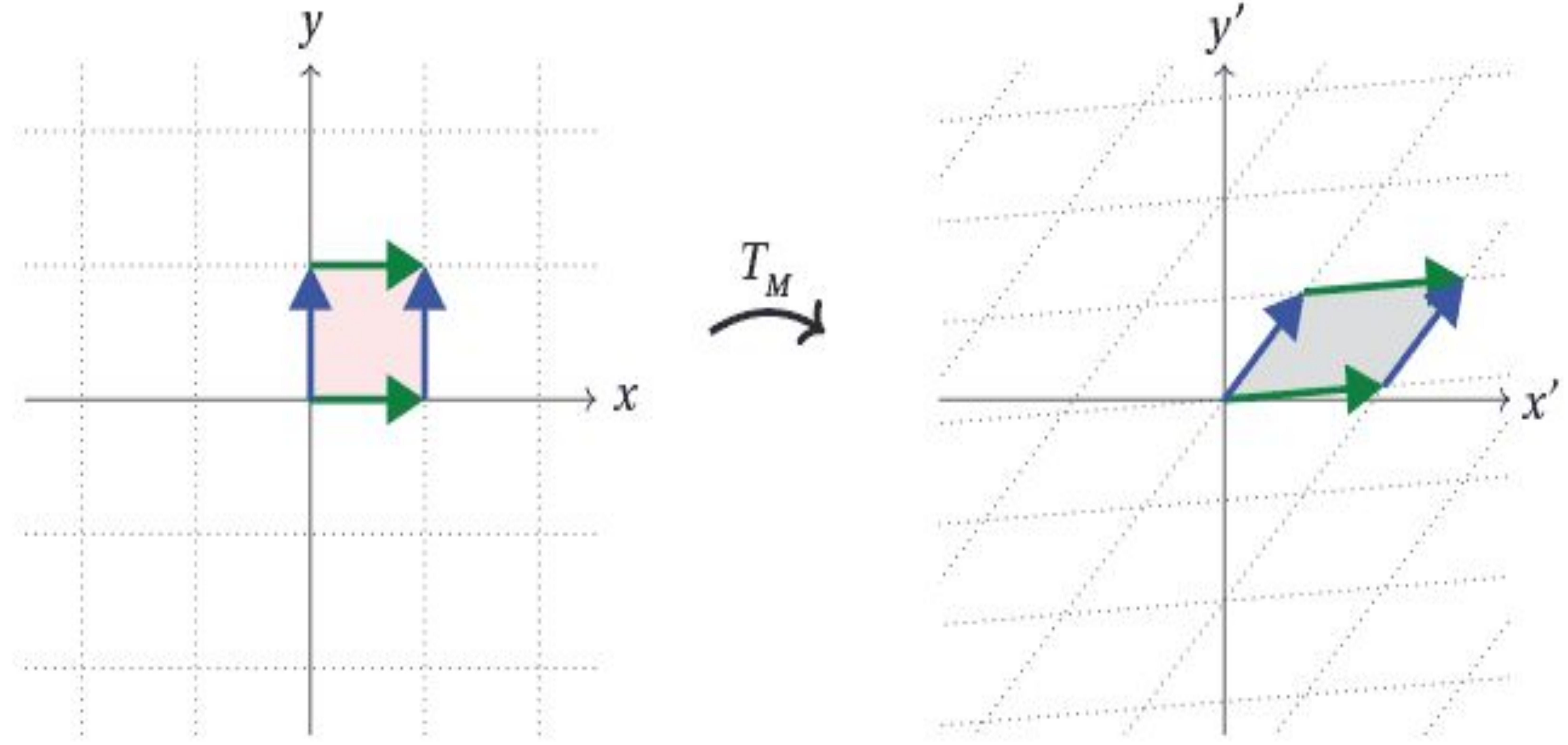
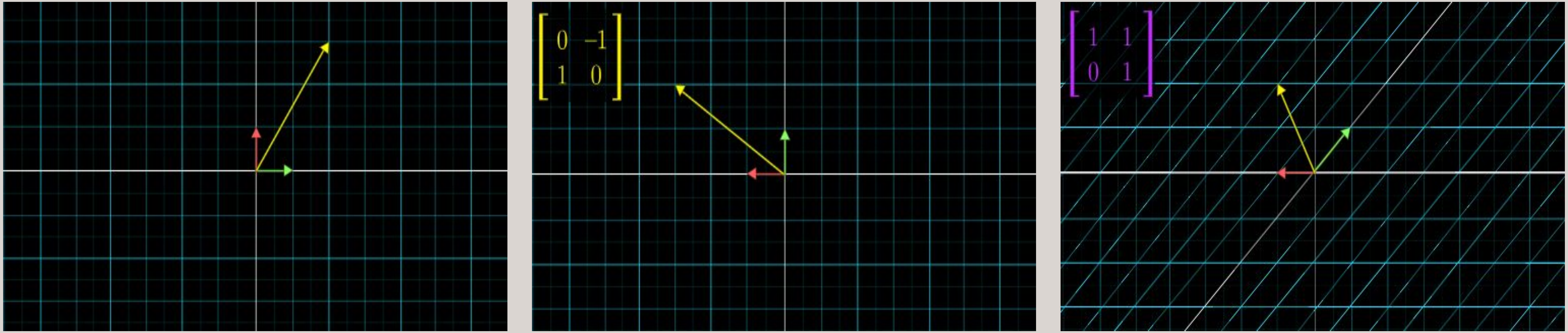


Image Source: 3 BLue 1 Brown

Matrix multiplication as Composition



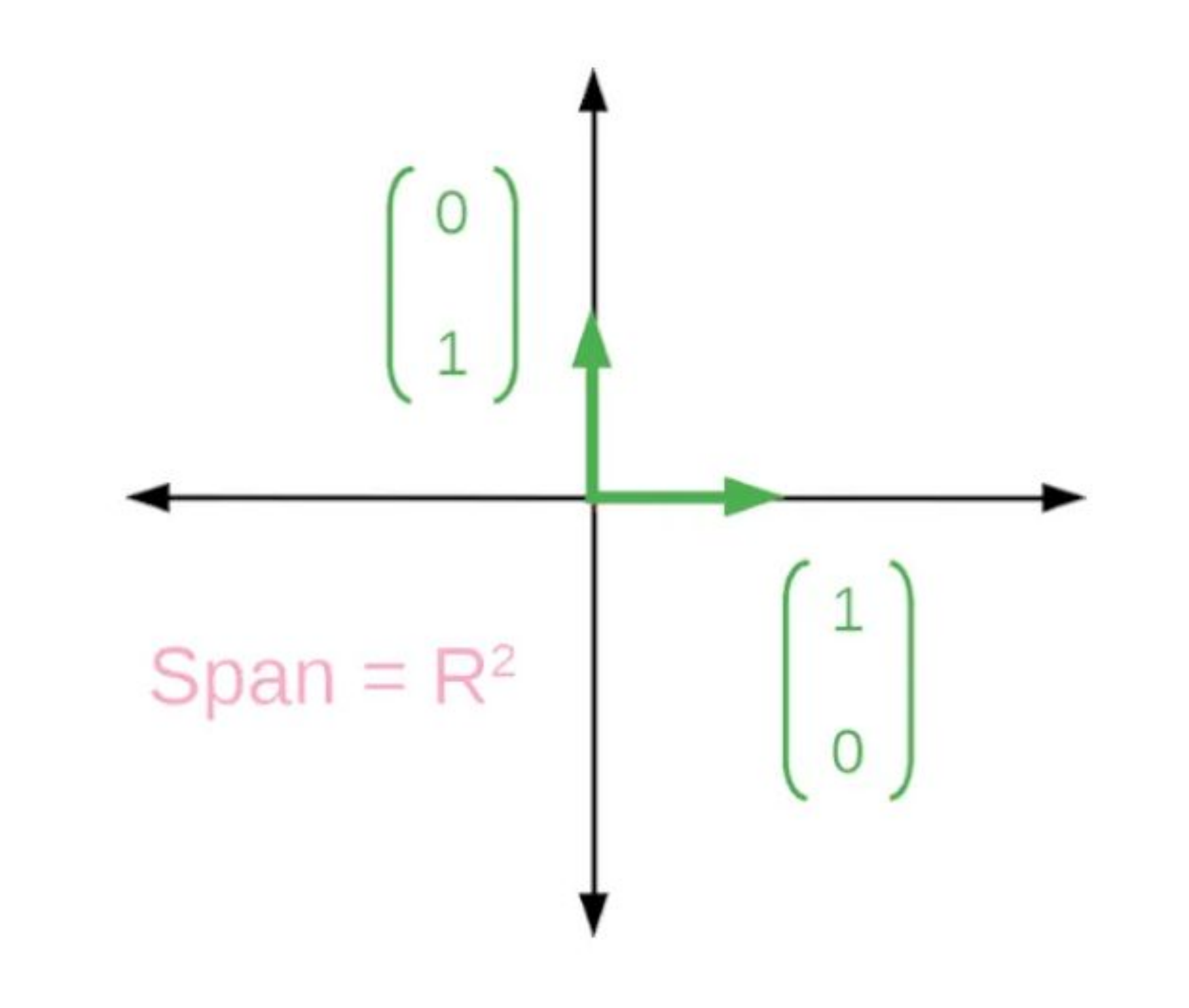
$$\underbrace{\begin{bmatrix}1 & 1\\0 & 1\end{bmatrix}}_{\text{Shear}}\left(\underbrace{\begin{bmatrix}0 & -1\\1 & 0\end{bmatrix}}_{\text{Rotation}}\begin{bmatrix}x\\y\end{bmatrix}\right)=\underbrace{\begin{bmatrix}1 & -1\\1 & 0\end{bmatrix}}_{\text{Composition}}\begin{bmatrix}x\\y\end{bmatrix}$$

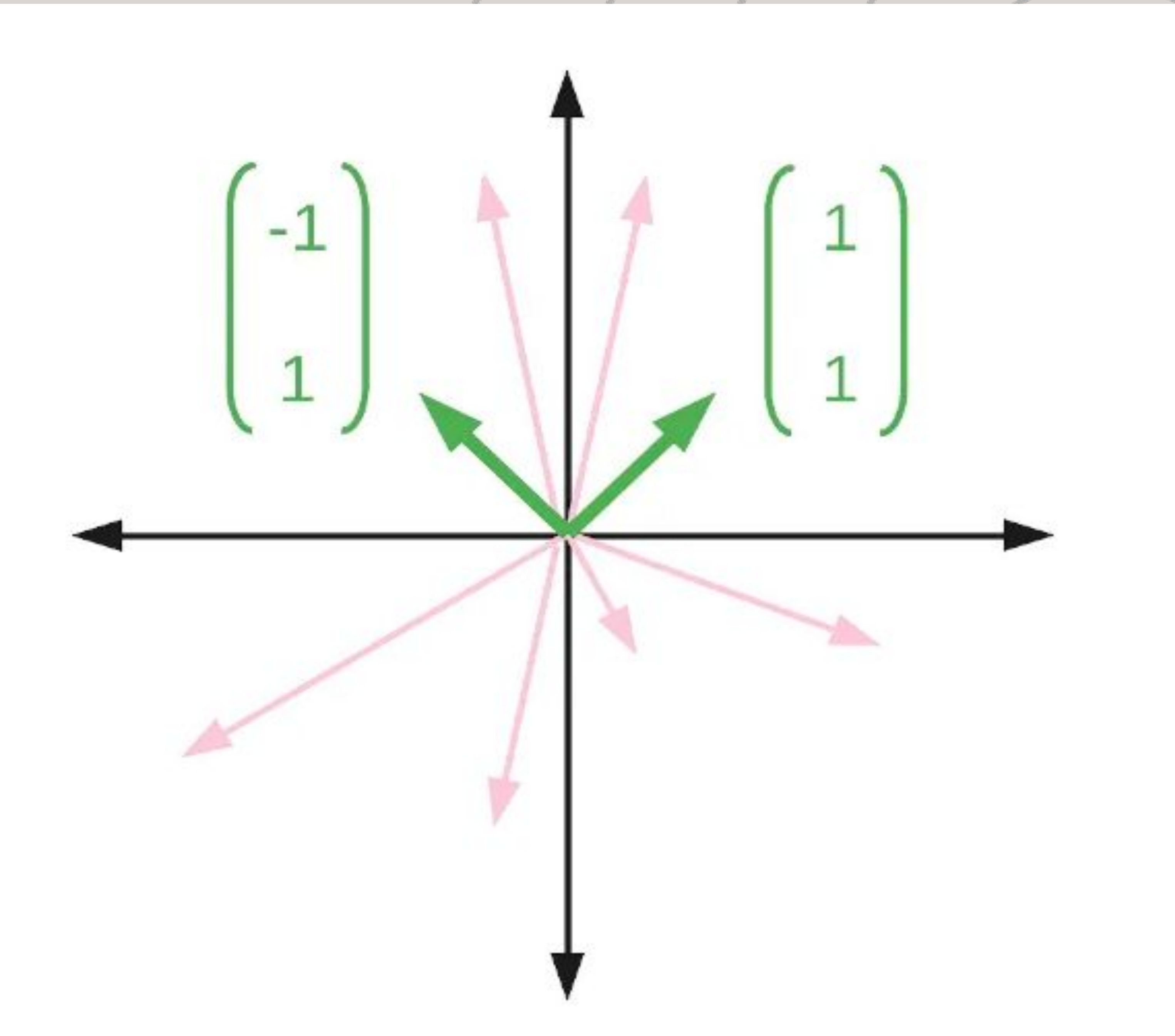
Image Source: 3 BLue 1 Brown

Concept of Basis

Photo by Steve Johnson on Unsplash

Basis





Basis: A set of n vectors, $\{v_1, v_2, \ldots v_n\}$, is a basis of some space S if:

1. **$\{v_1, v_2, \ldots v_n\}$ are linearly independent**
2. **$\{v_1, v_2, \ldots v_n\}$ span the set S. In other words, $\mathrm{Span}\{v_1, v_2, \ldots v_n\} = S$**

Rank

 = $7

 = $9

Rank

= $7

= $14

Rank

For example, the matrix A given by

$$A = \begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix}$$

can be put in reduced row-echelon form by using the following elementary row operations:

$$\begin{bmatrix} 1 & 2 & 1 \\ -2 & -3 & 1 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{2R_1+R_2 \to R_2} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 3 & 5 & 0 \end{bmatrix} \xrightarrow{-3R_1+R_3 \to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & -1 & -3 \end{bmatrix}$$

$$\xrightarrow{R_2+R_3 \to R_3} \begin{bmatrix} 1 & 2 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix} \xrightarrow{-2R_2+R_1 \to R_1} \begin{bmatrix} 1 & 0 & -5 \\ 0 & 1 & 3 \\ 0 & 0 & 0 \end{bmatrix}.$$

The final matrix (in reduced row echelon form) has two non-zero rows and thus the rank of matrix A is 2.

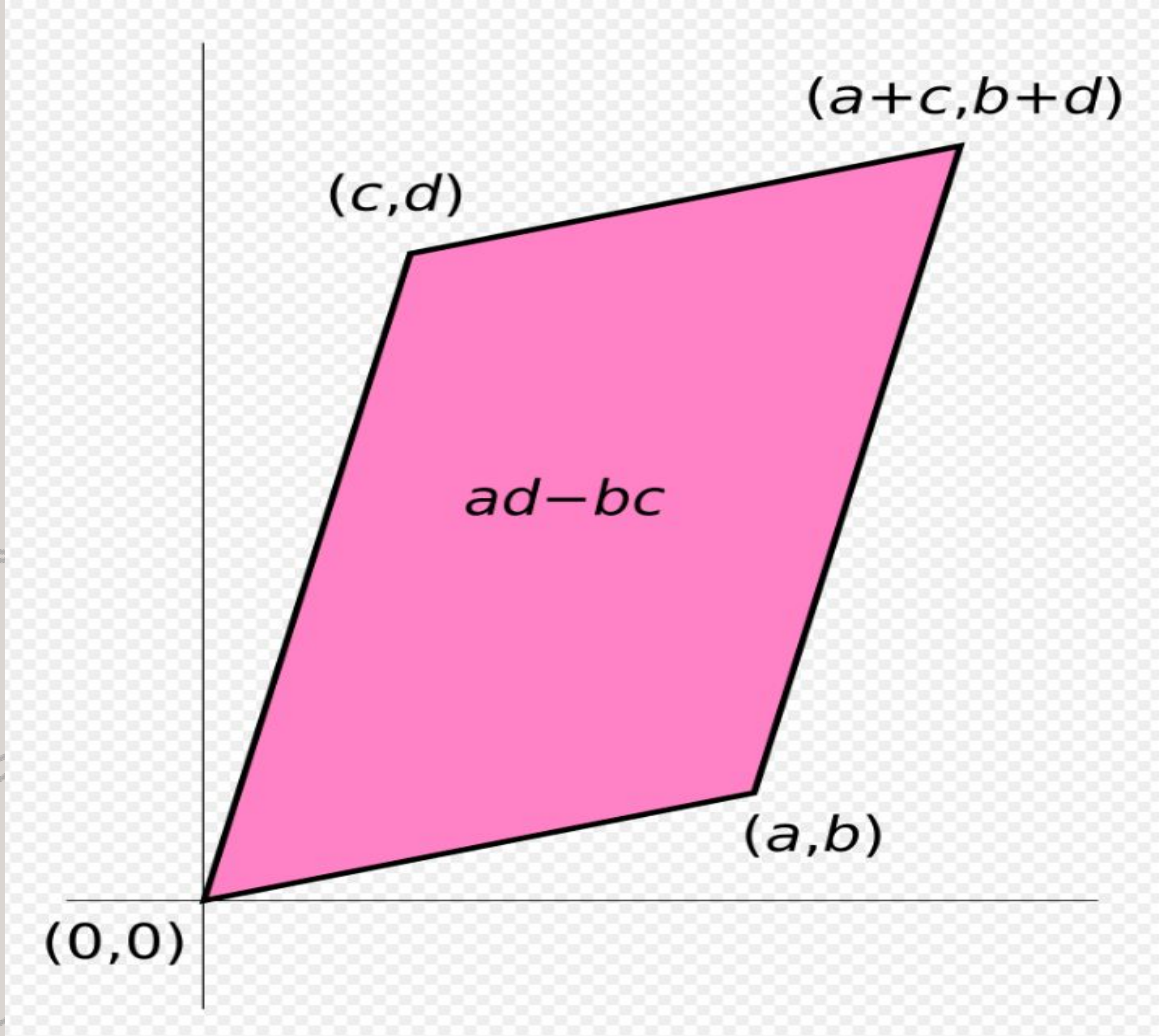
Trace

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$tr(A) = a_{11} + a_{22} + a_{33}$$

Gives Important information about:
Area Scaling Factor
Invertibility
Orientation

Determinant



$\det(A) = ad - bc$

Determinant

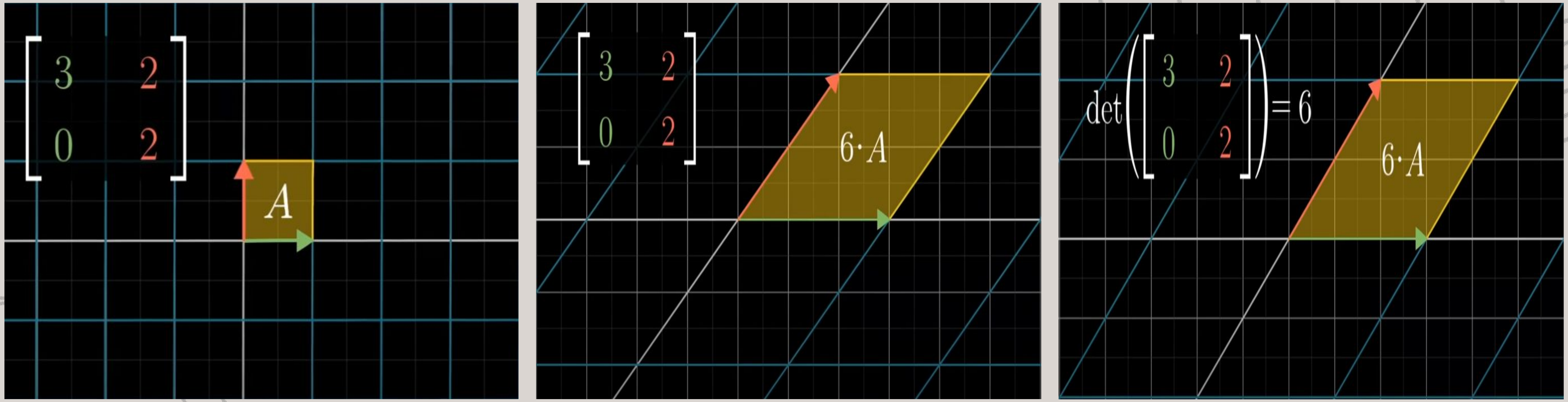


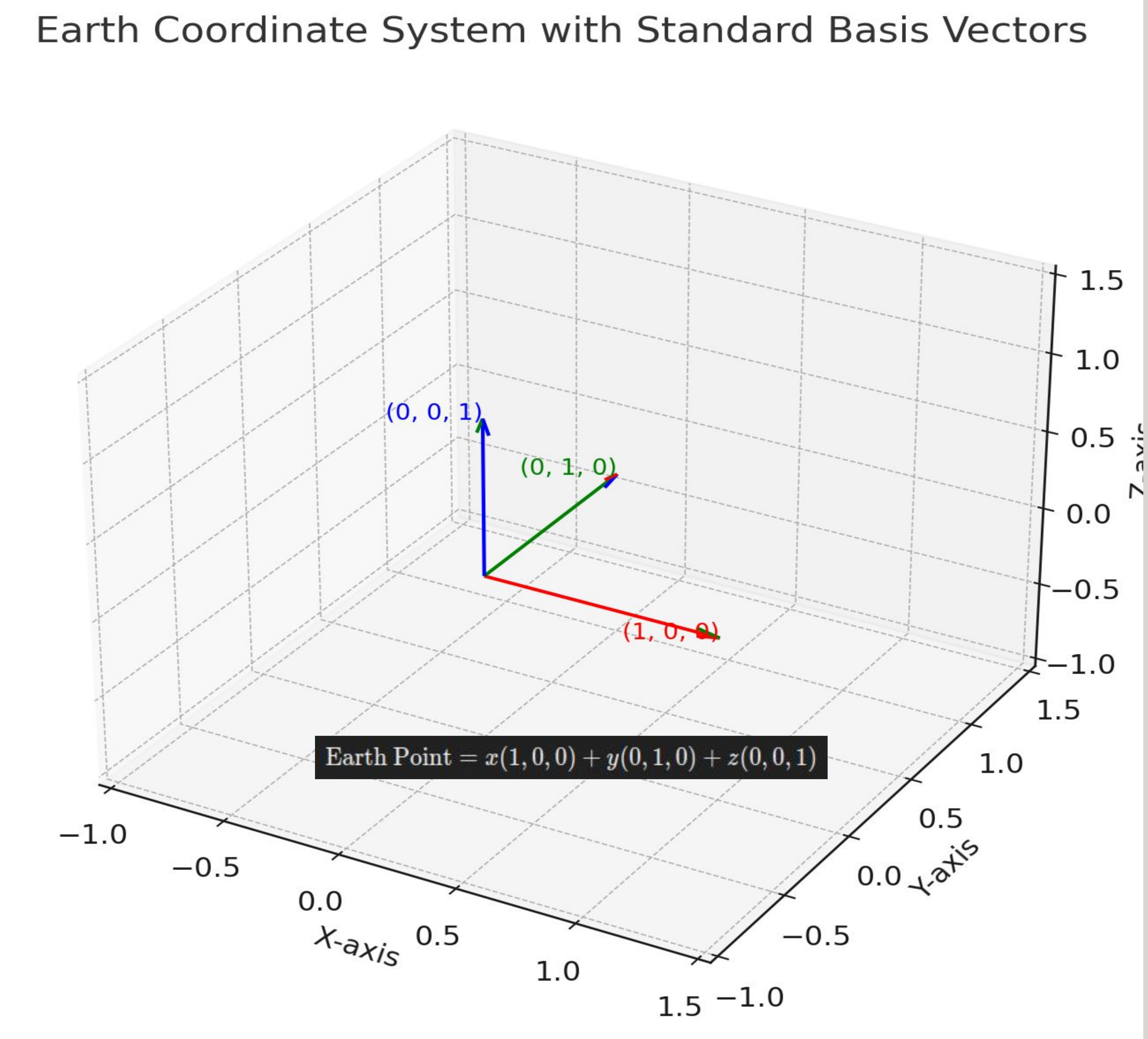
Image Source: 3 BLue 1 Brown

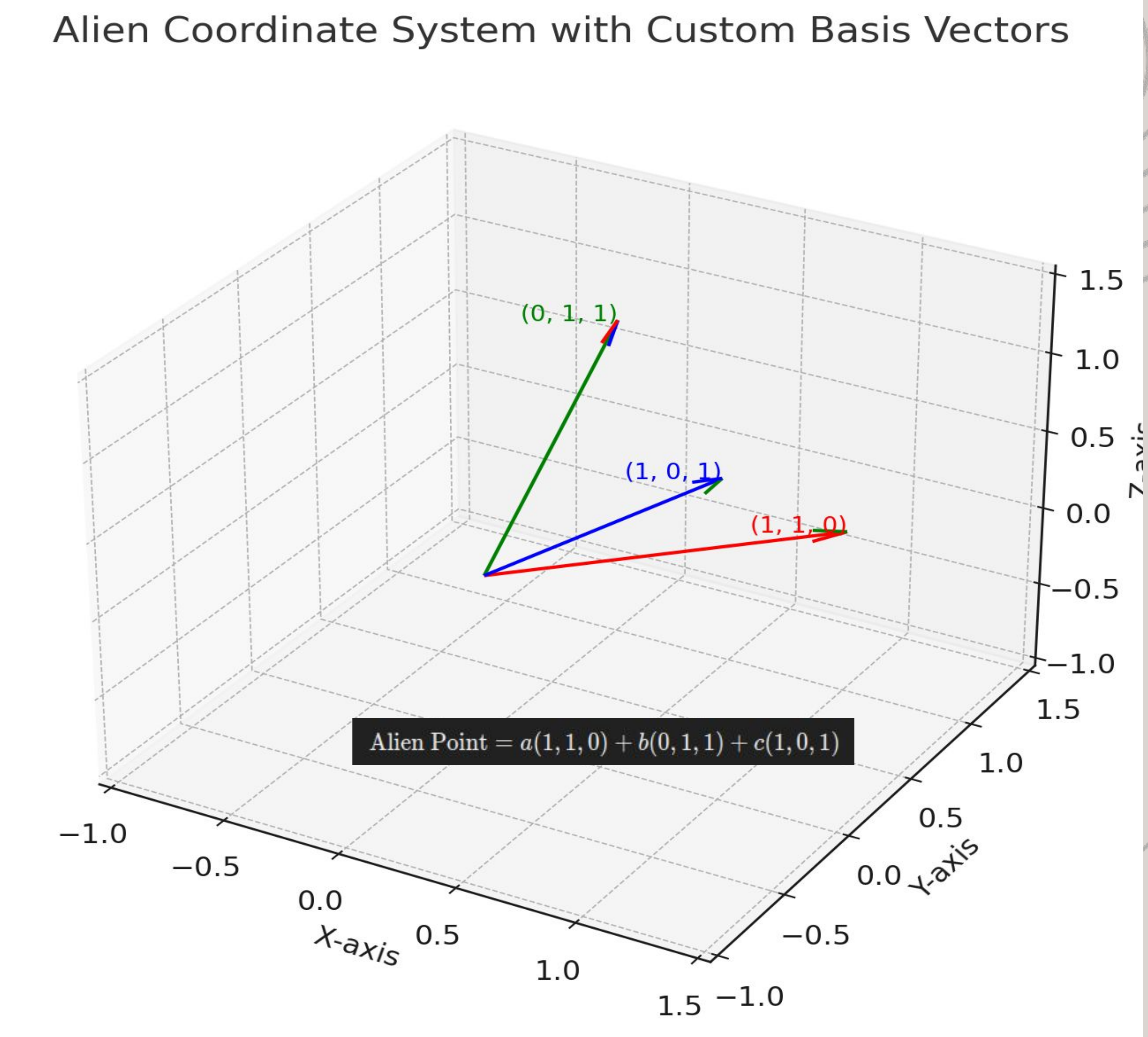
Transpose

$$\begin{bmatrix} 2 & 4 & -1 \\ -10 & 5 & 11 \\ 18 & -7 & 6 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & -10 & 18 \\ 4 & 5 & -7 \\ -1 & 11 & 6 \end{bmatrix}$$

A $\qquad$ A^T

Change of Basis





Change of Basis

$$\text{Earth Point} = x(1,0,0) + y(0,1,0) + z(0,0,1)$$

$$\text{Alien Point} = a(1,1,0) + b(0,1,1) + c(1,0,1)$$

Suppose their basis vectors in our system are described as:

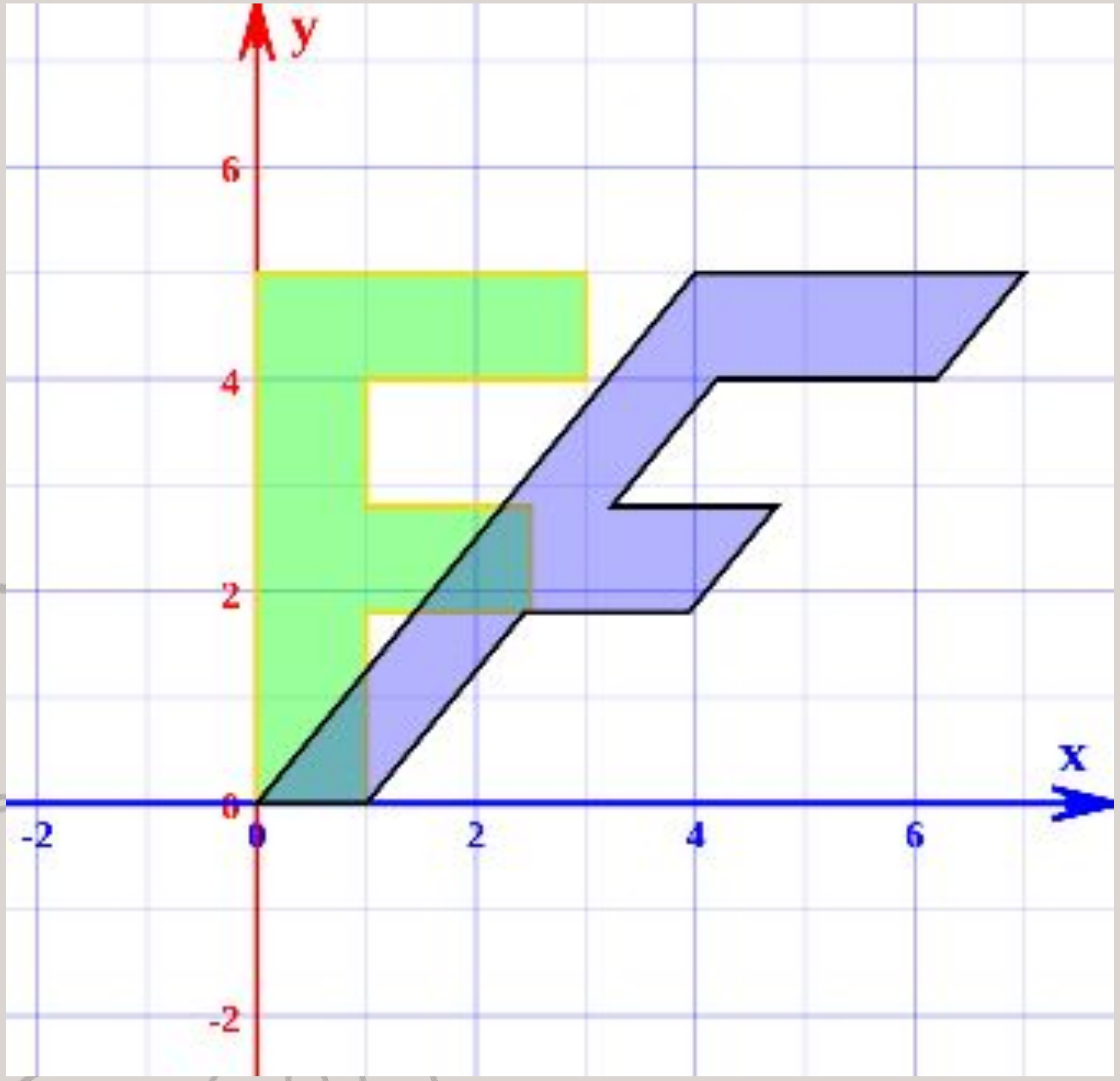
- $(1,1,0) = x_1(1,0,0) + y_1(0,1,0) + z_1(0,0,1)$
- $(0,1,1) = x_2(1,0,0) + y_2(0,1,0) + z_2(0,0,1)$
- $(1,0,1) = x_3(1,0,0) + y_3(0,1,0) + z_3(0,0,1)$

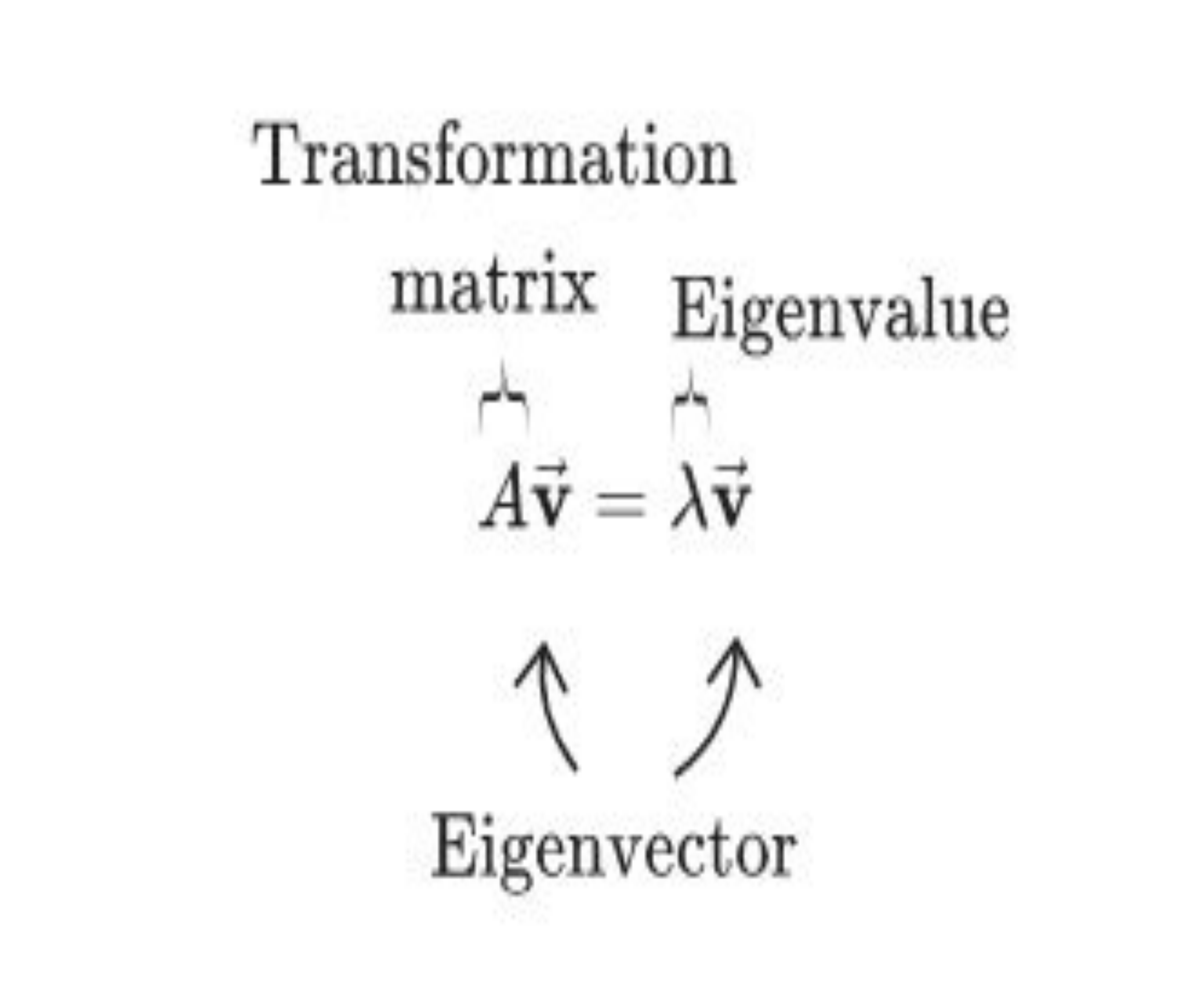
$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = B^{-1} \cdot \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Where,

$$B = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

Eigenvectors & Eigenvalues





- Data Representation
- Vectors and Matrices
- Vector Spaces
- Span
- Rank
- Trace
- Transpose
- Determinant
- Change of Basis
- Eigenvectors and Eigenvalues