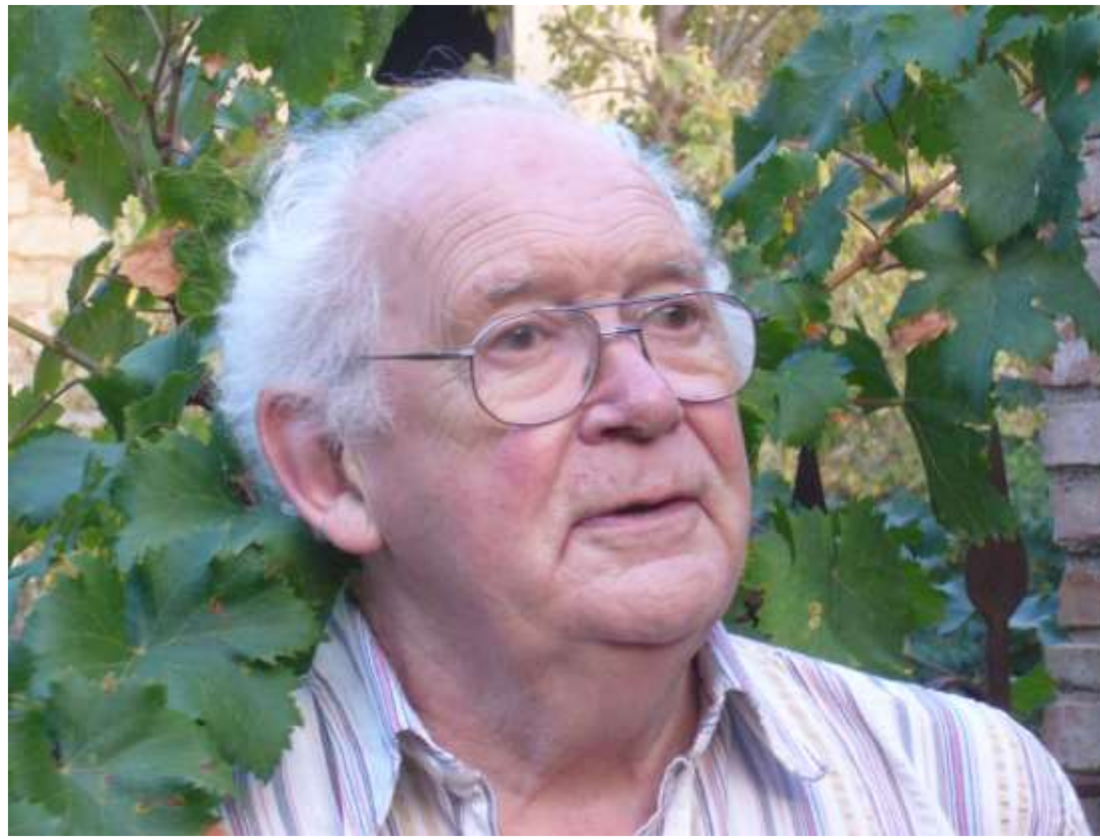
Compositional Data



John Aitchison

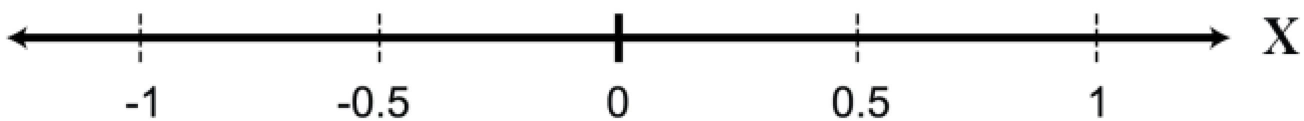
What is a domain?

Domain: the set of values for which a variable is defined

What is a domain?

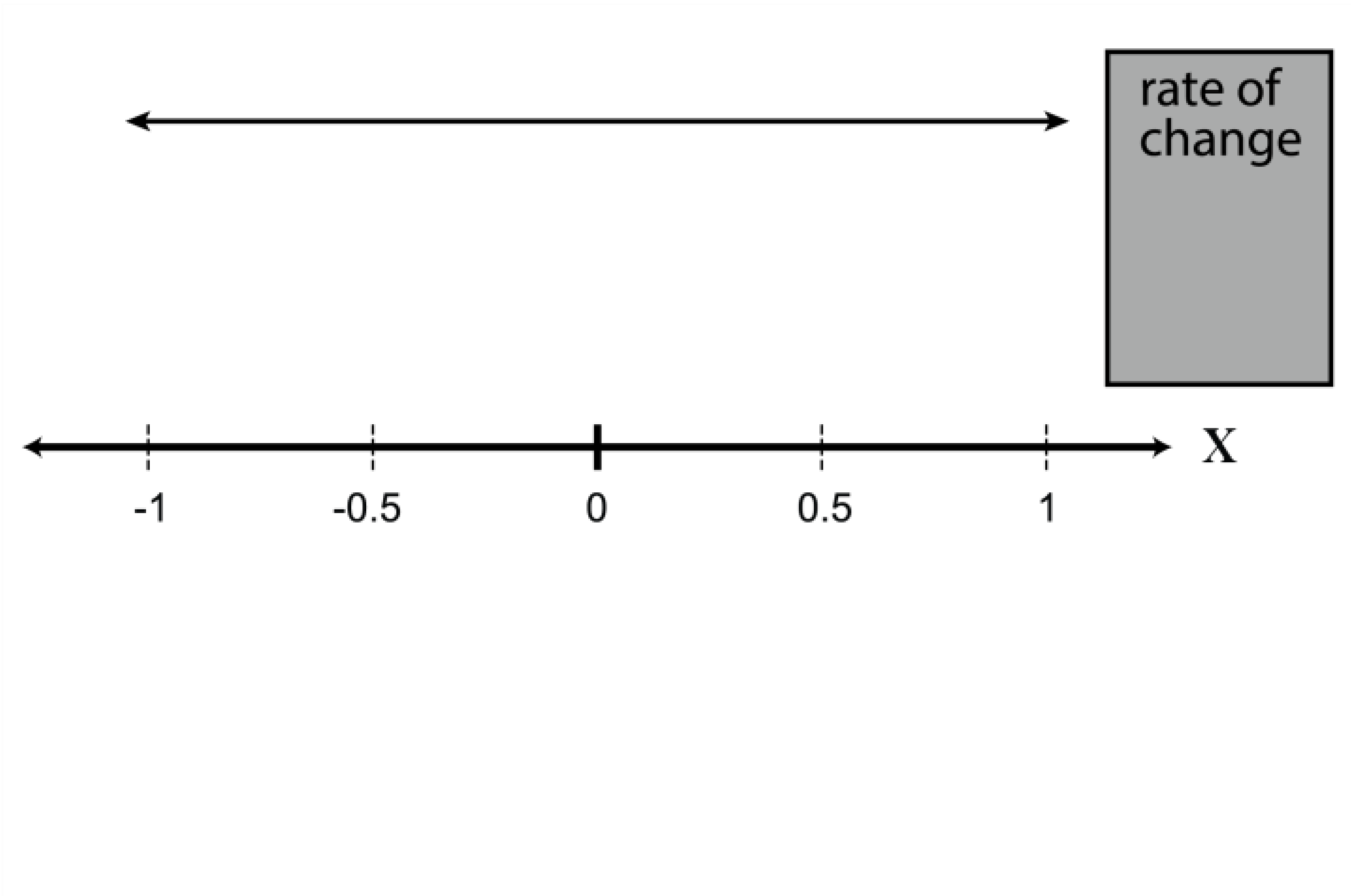
Domain: the set of values for which a variable is defined

Real numbers:



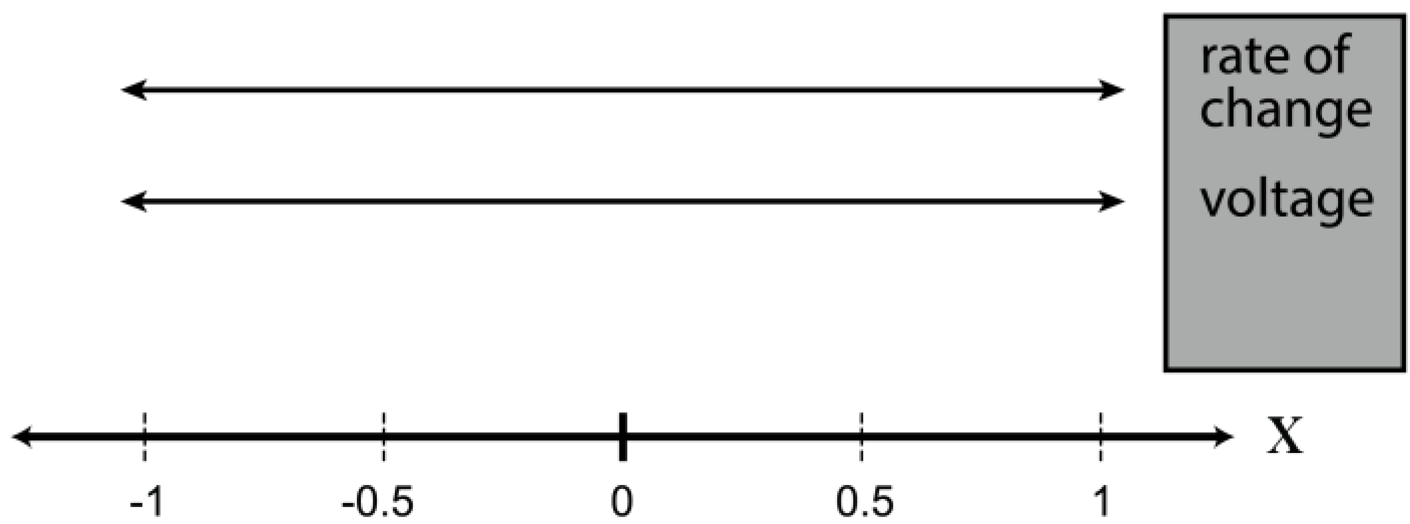
What is a domain?

Domain: the set of values for which a variable is defined



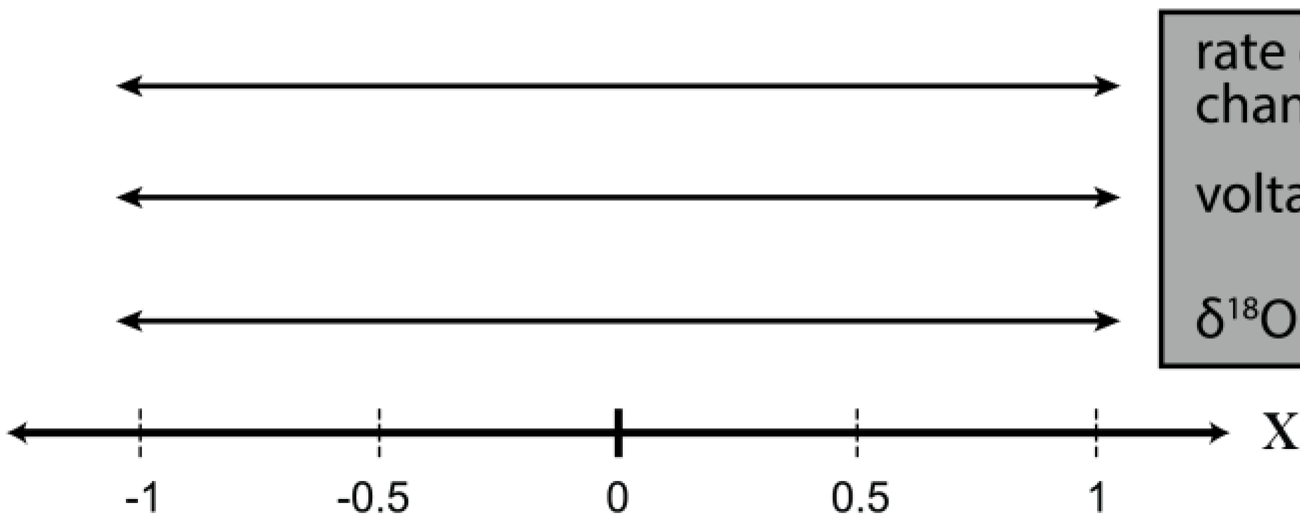
What is a domain?

Domain: the set of values for which a variable is defined



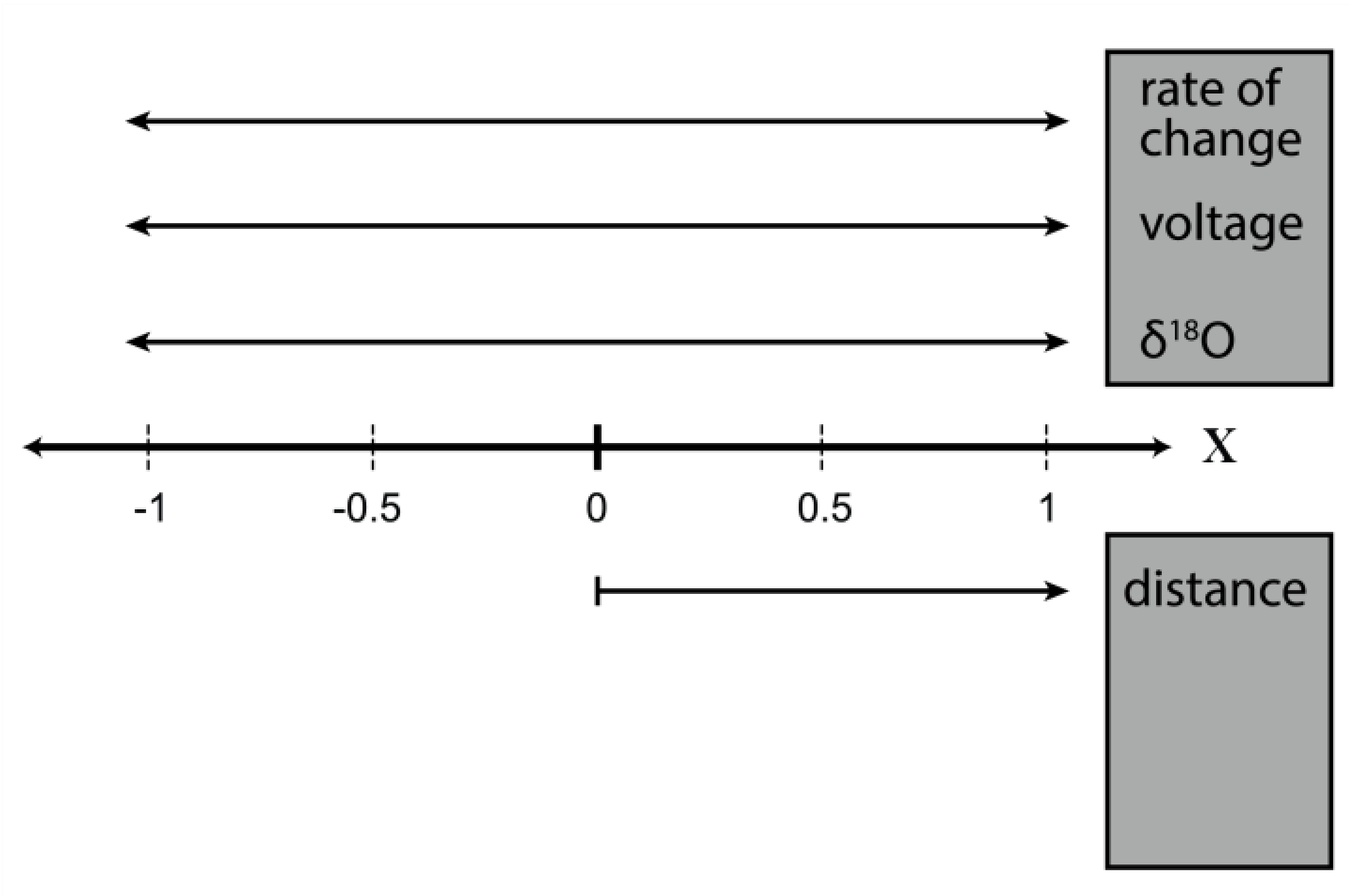
What is a domain?

Domain: the set of values for which a variable is defined



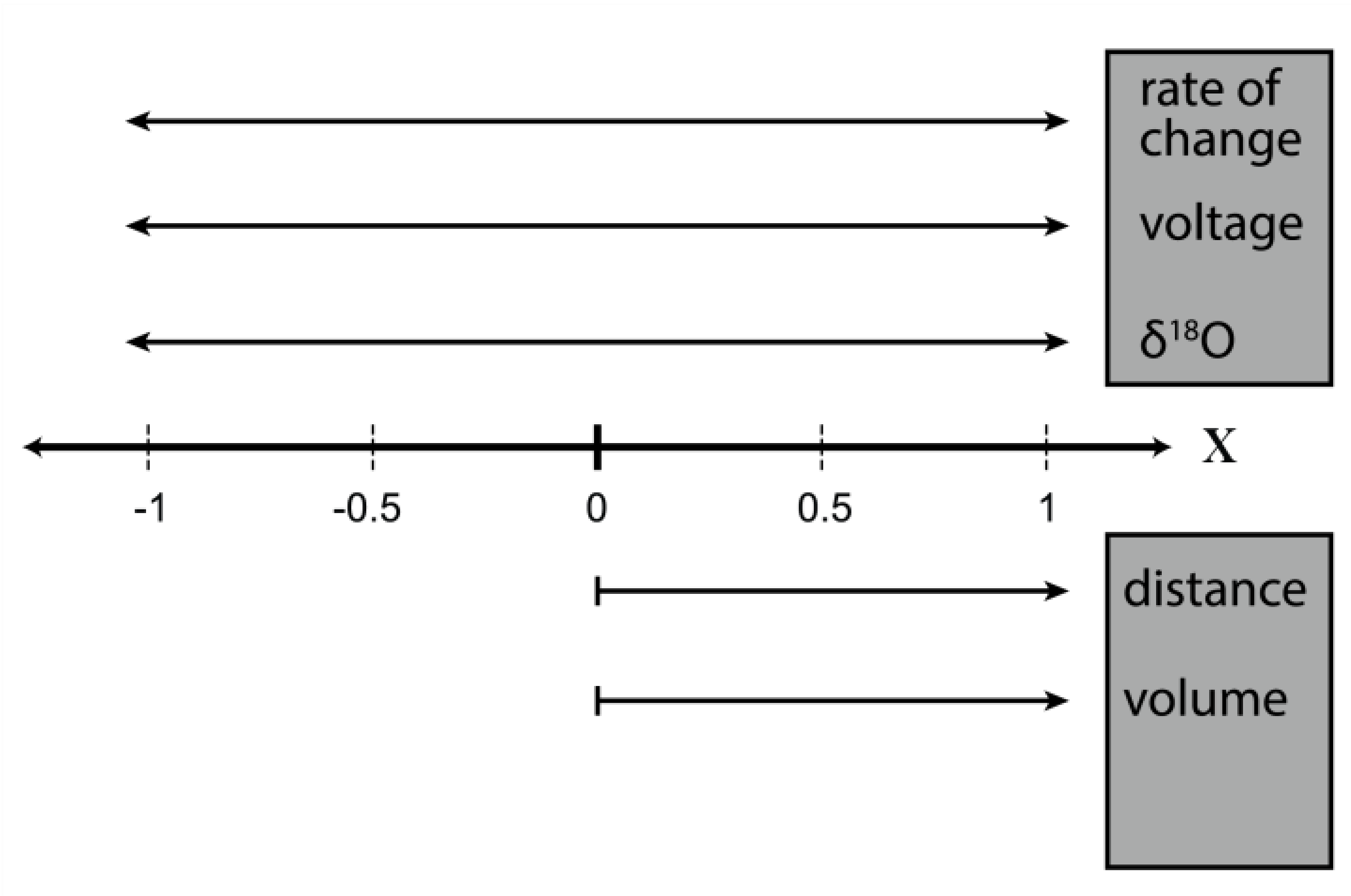
What is a domain?

Domain: the set of values for which a variable is defined



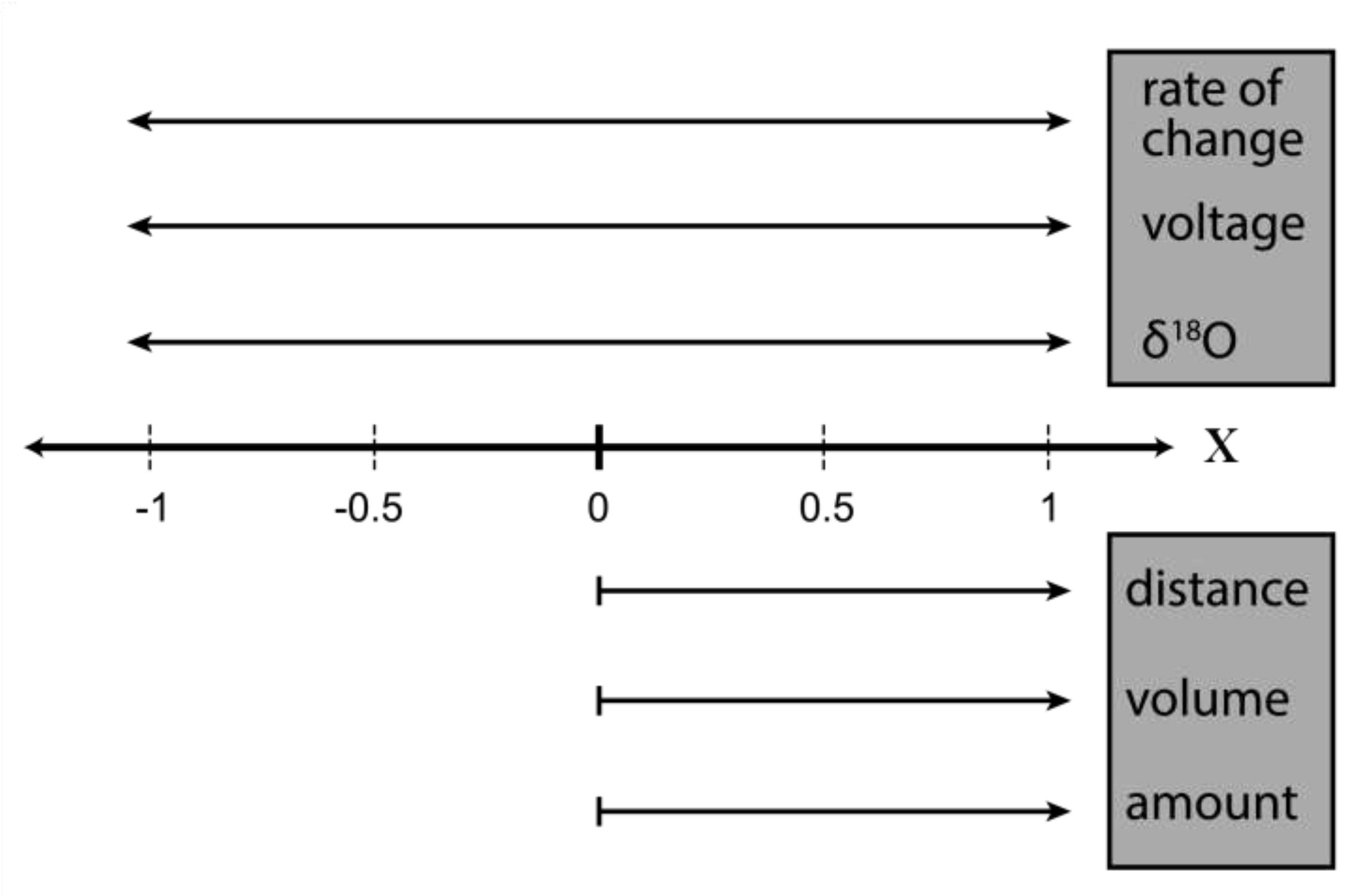
What is a domain?

Domain: the set of values for which a variable is defined



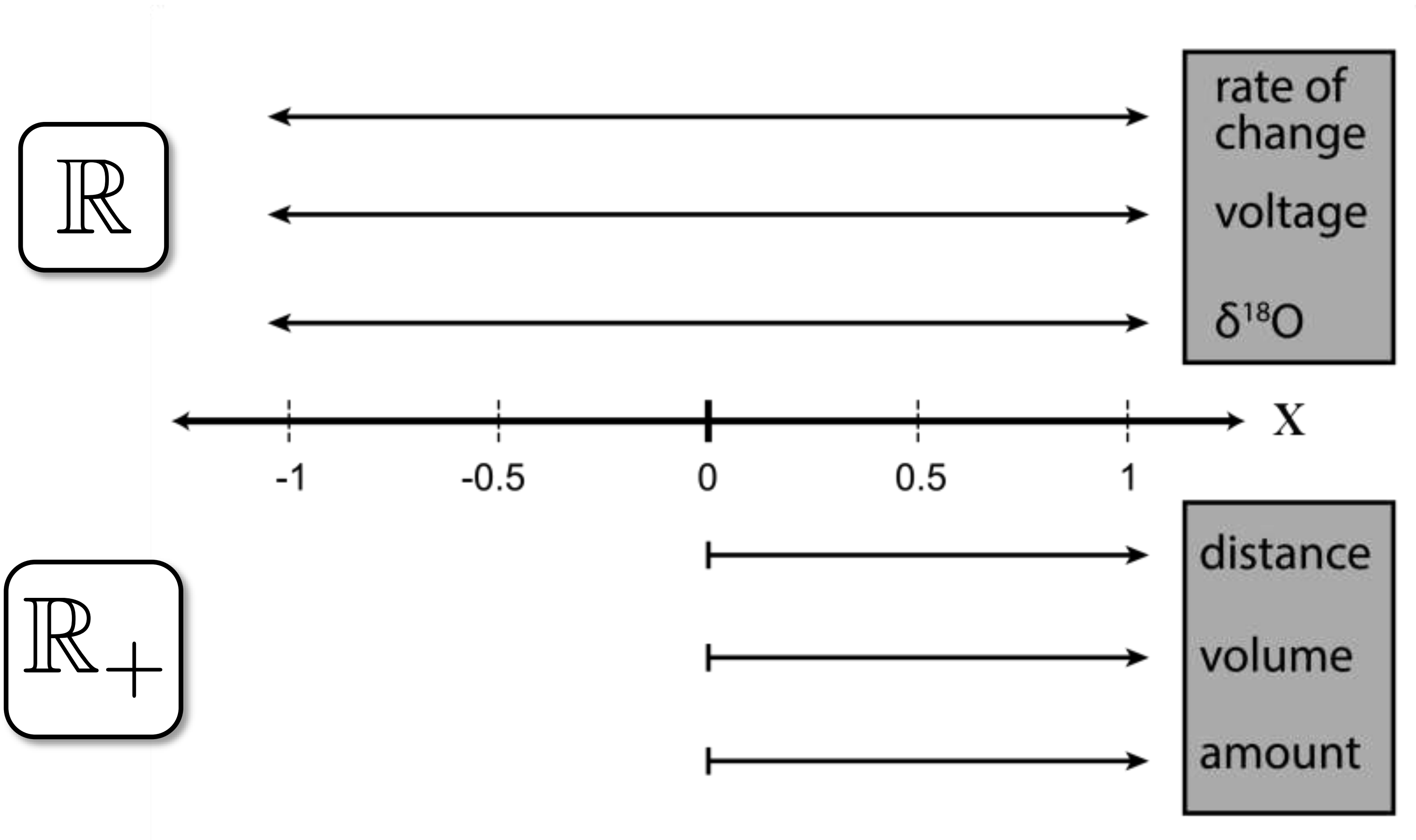
What is a domain?

Domain: the set of values for which a variable is defined



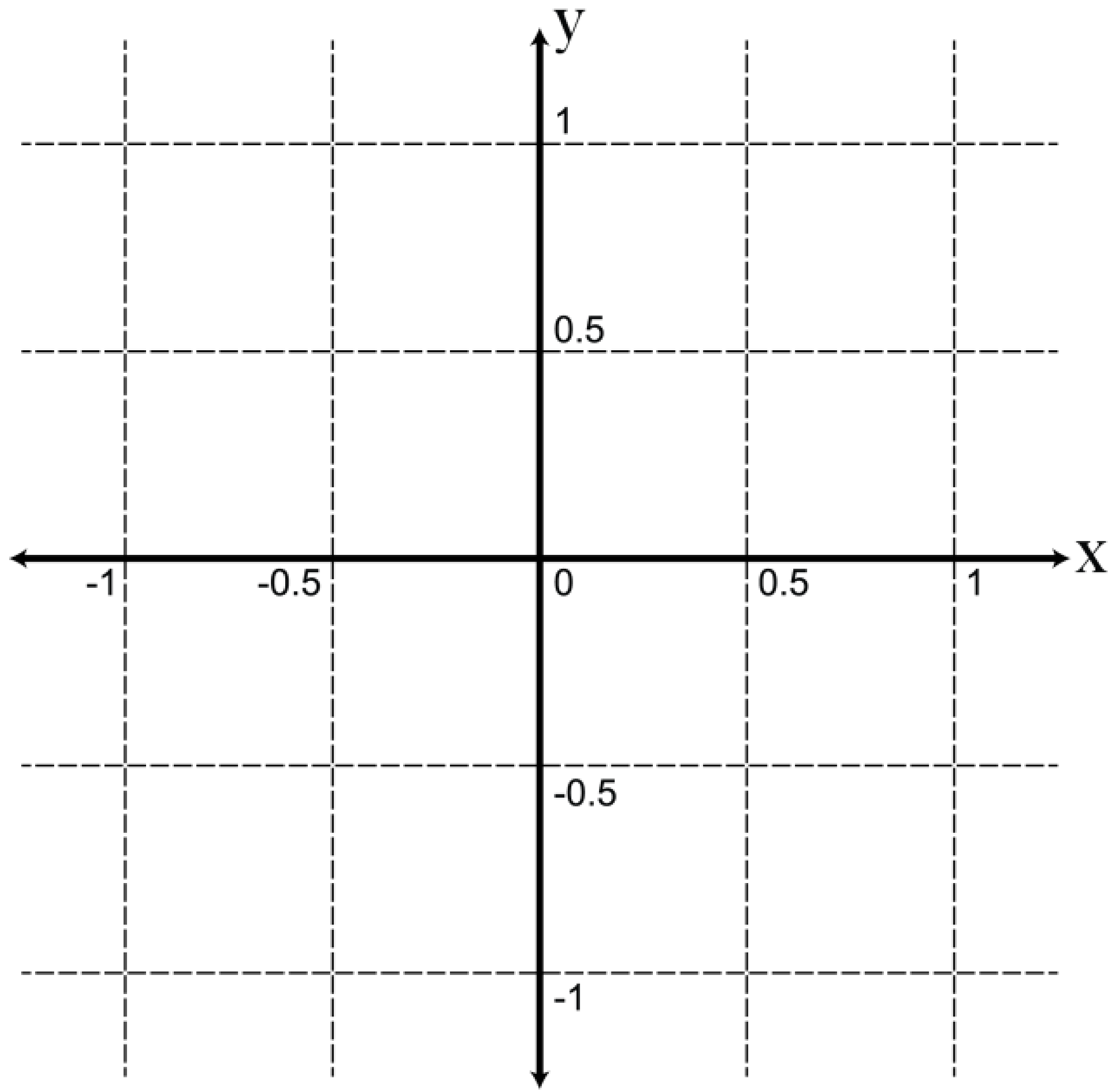
What is a domain?

Domain: the set of values for which a variable is defined



Now for two variables

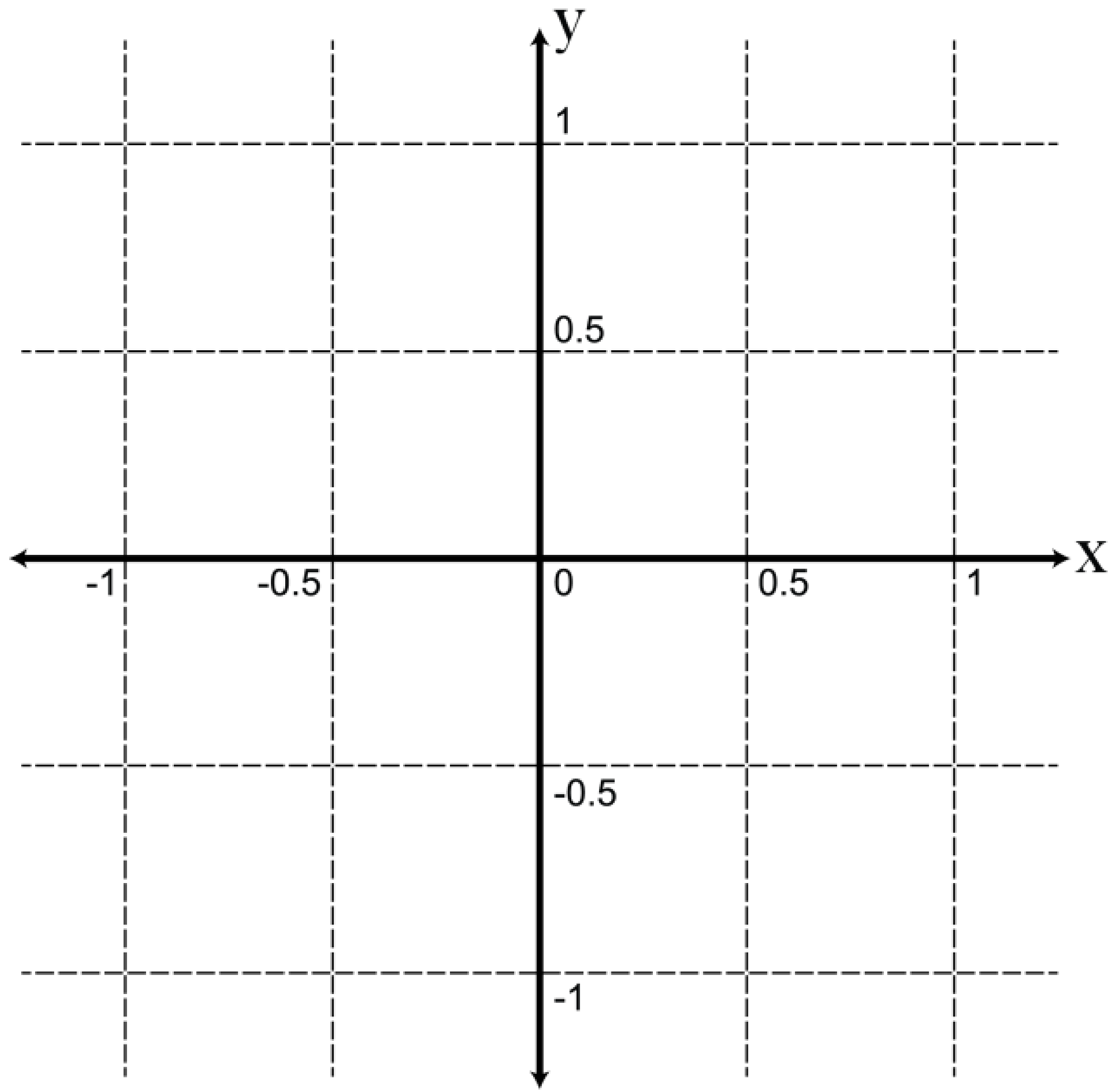
If you are measuring two variables that are defined over the **real numbers**, then you get to use the whole 2D plane when plotting x vs. y



Now for two variables

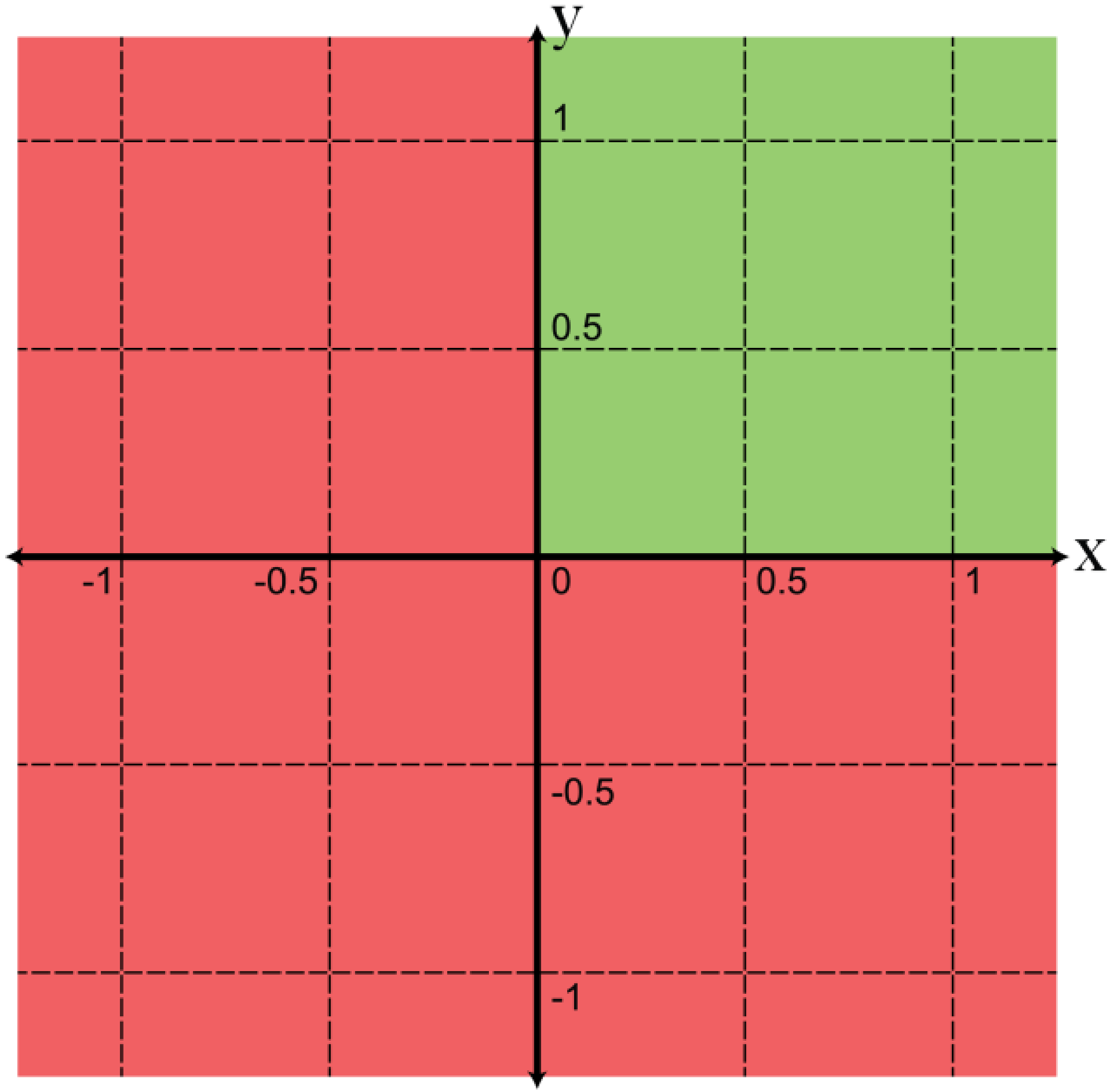
If you are measuring two variables that are defined over the **real numbers**, then you get to use the whole 2D plane when plotting x vs. y

$\mathbb{R}^2$



Now for two variables

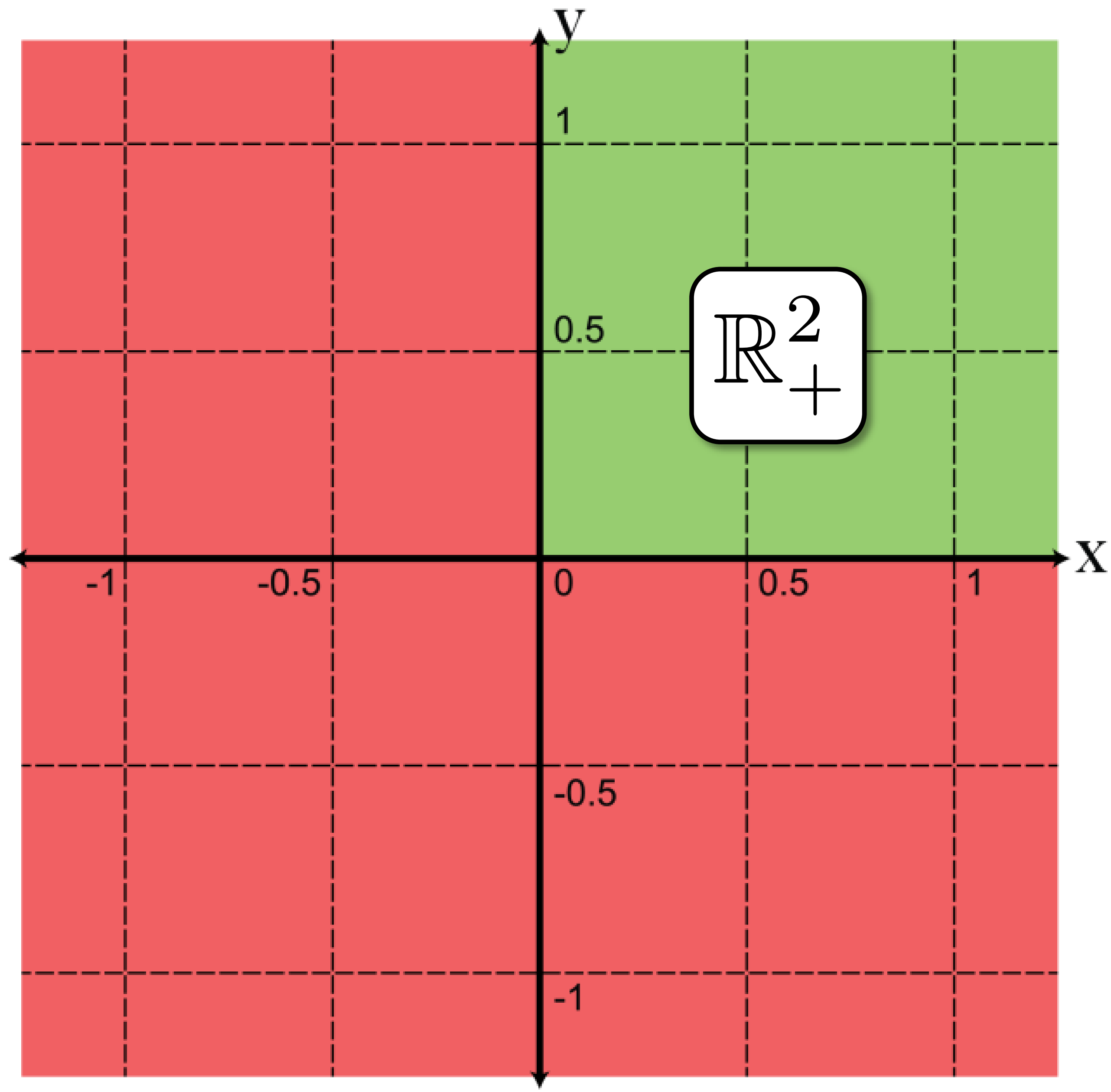
If the two variables are constrained to be **positive** real numbers, then you only get to use a portion of the real plane:



Now for two variables

If the two variables are constrained to be **positive** real numbers, then you only get to use a portion of the real plane:

$\mathbb{R}^2_+$



So, what is a composition?

- The variables you're measuring (components) all belong to $\mathbb{R}_+$ and have the same units or on the same measurement scale

So, what is a composition?

- The variables you're measuring (components) all belong to $\mathbb{R}_+$ and have the same units or on the same measurement scale
- You're not interested in the absolute abundance of any of the components—only the *relative abundances*, or proportions of a whole

So, what is a <u>composition</u>?

- The variables you're measuring (components) all belong to $\mathbb{R}_+$ and have the same units or on the same measurement scale
- You're not interested in the absolute abundance of any of the components—only the *relative abundances*, or proportions of a whole
- All compositions sum to a constant
  - e.g. 100%, or 1

Examples of compositions

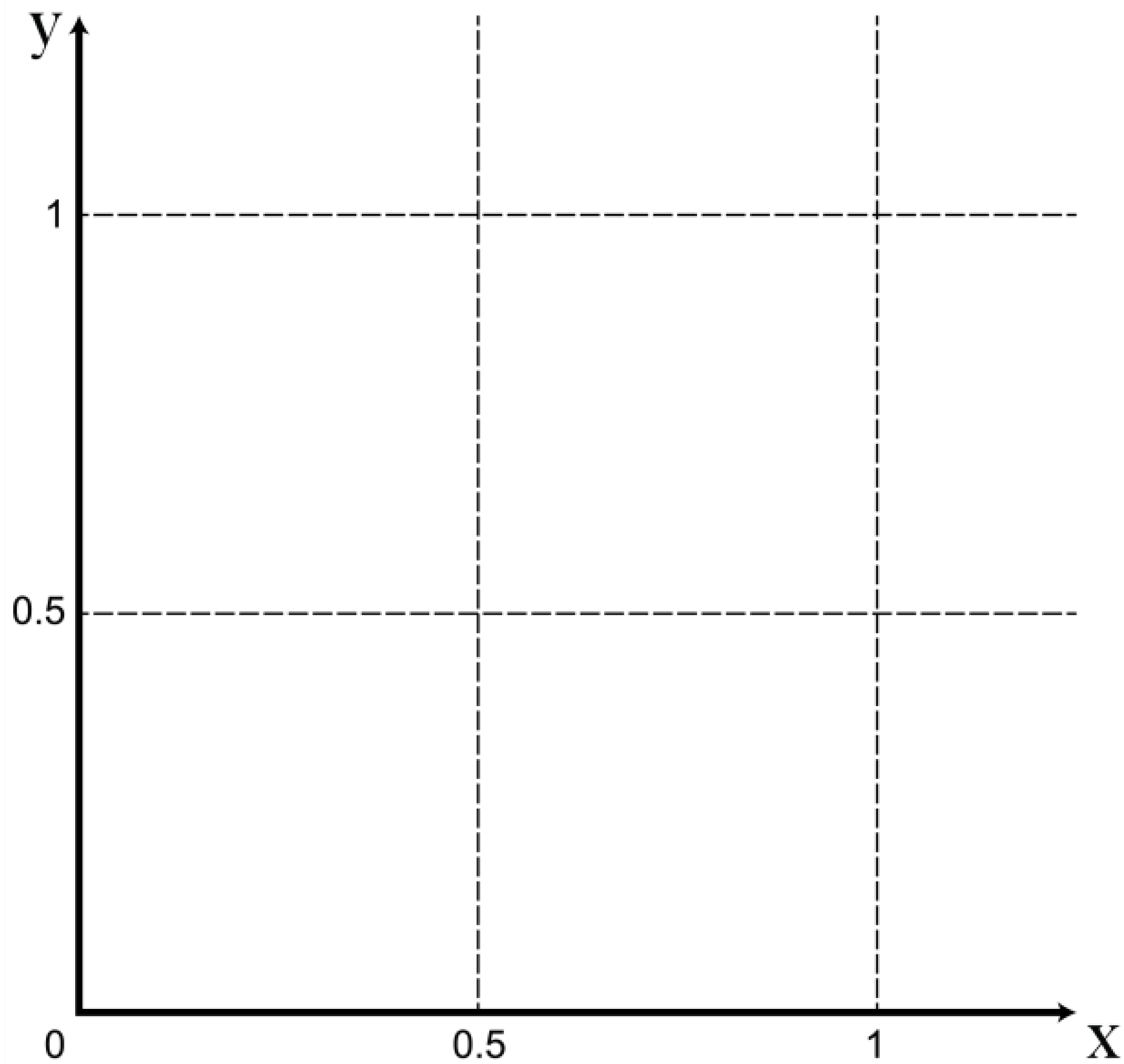
- Geochemical compositions of rocks
  - e.g. wt. % oxides, but not ppm trace elements
  - Modal abundances: cpx/plag/amph/ol
- Sediment grain sizes or compositions
  - e.g. sand/silt/clay, quartz/feldspar/lithics
- Isotopic compositions
  - $^{204}Pb/^{206}Pb/^{207}Pb/^{208}Pb$, $^{234}U/^{235}U/^{238}U$

Zoming into

$$\mathbb{R}^2_+$$

Note that the boundaries at x=0 and y=0 are absolute: nothing is allowed to be negative.

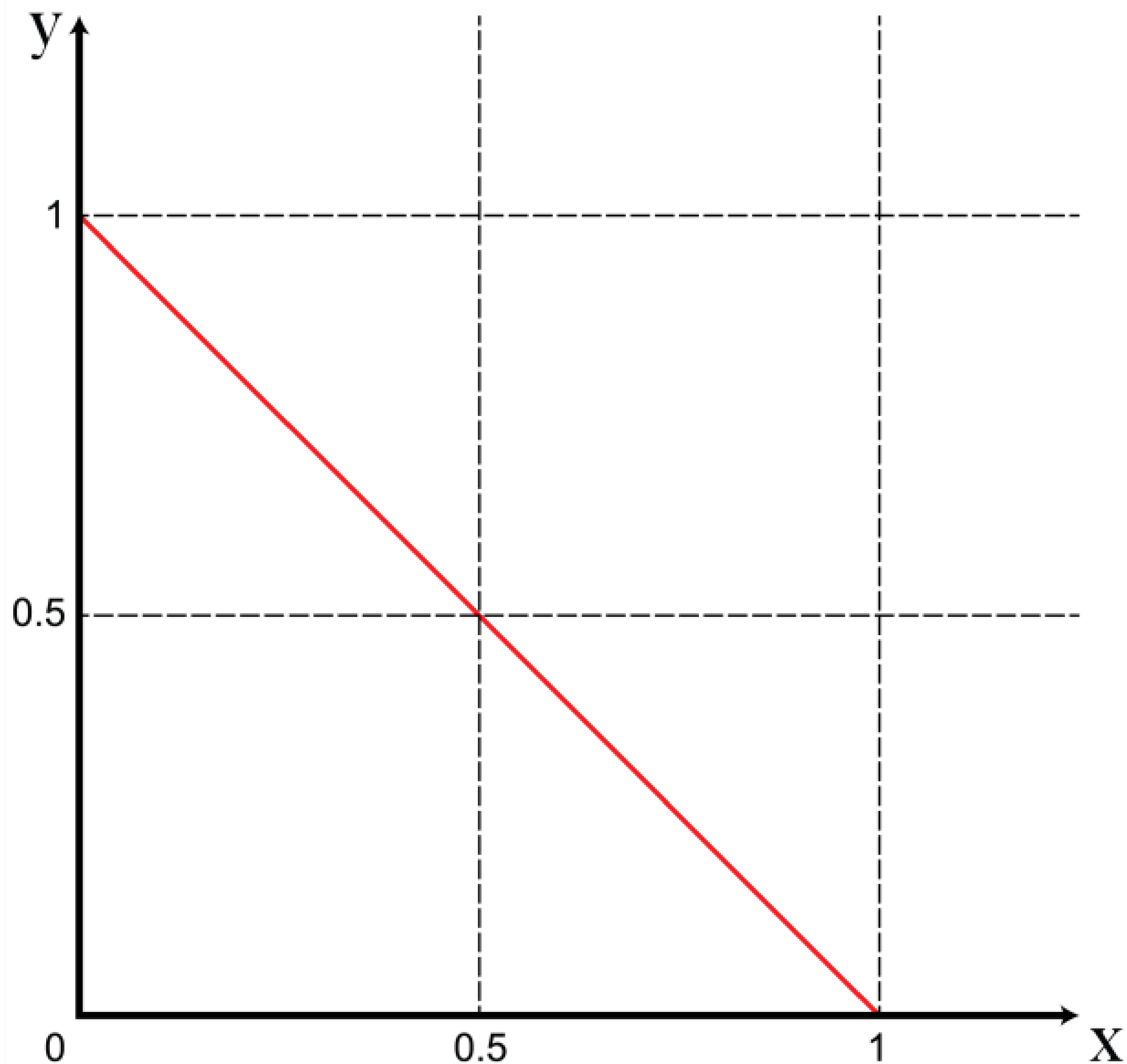
So, where do compositions fit in?



The simplex

The red line, known as the **simplex**, shows all the locations where x+y=1.

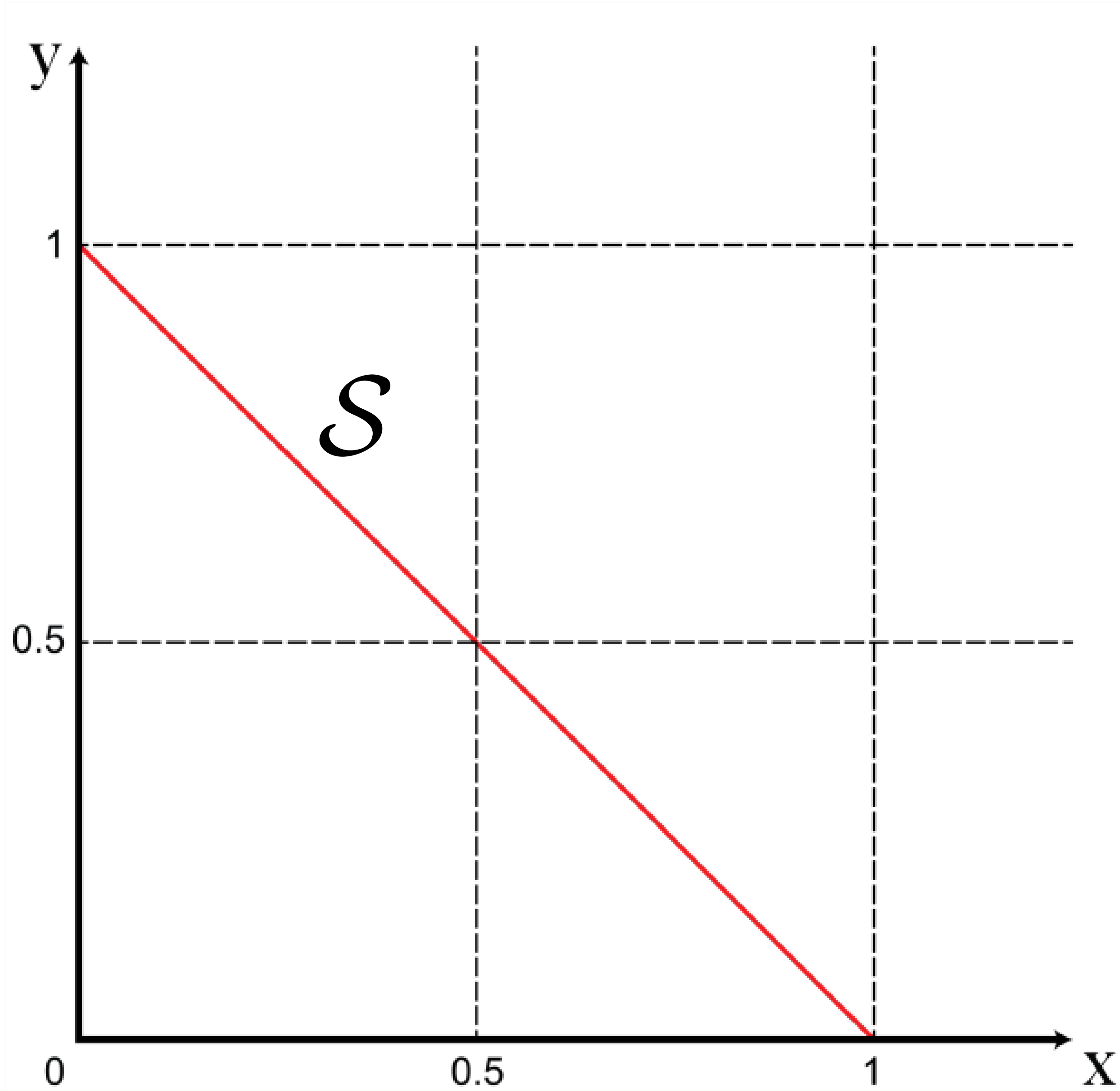
All two-component compositions (which must sum to a whole) can be represented on the line.



The simplex

The red line, known as the **simplex**, shows all the locations where x+y=1.

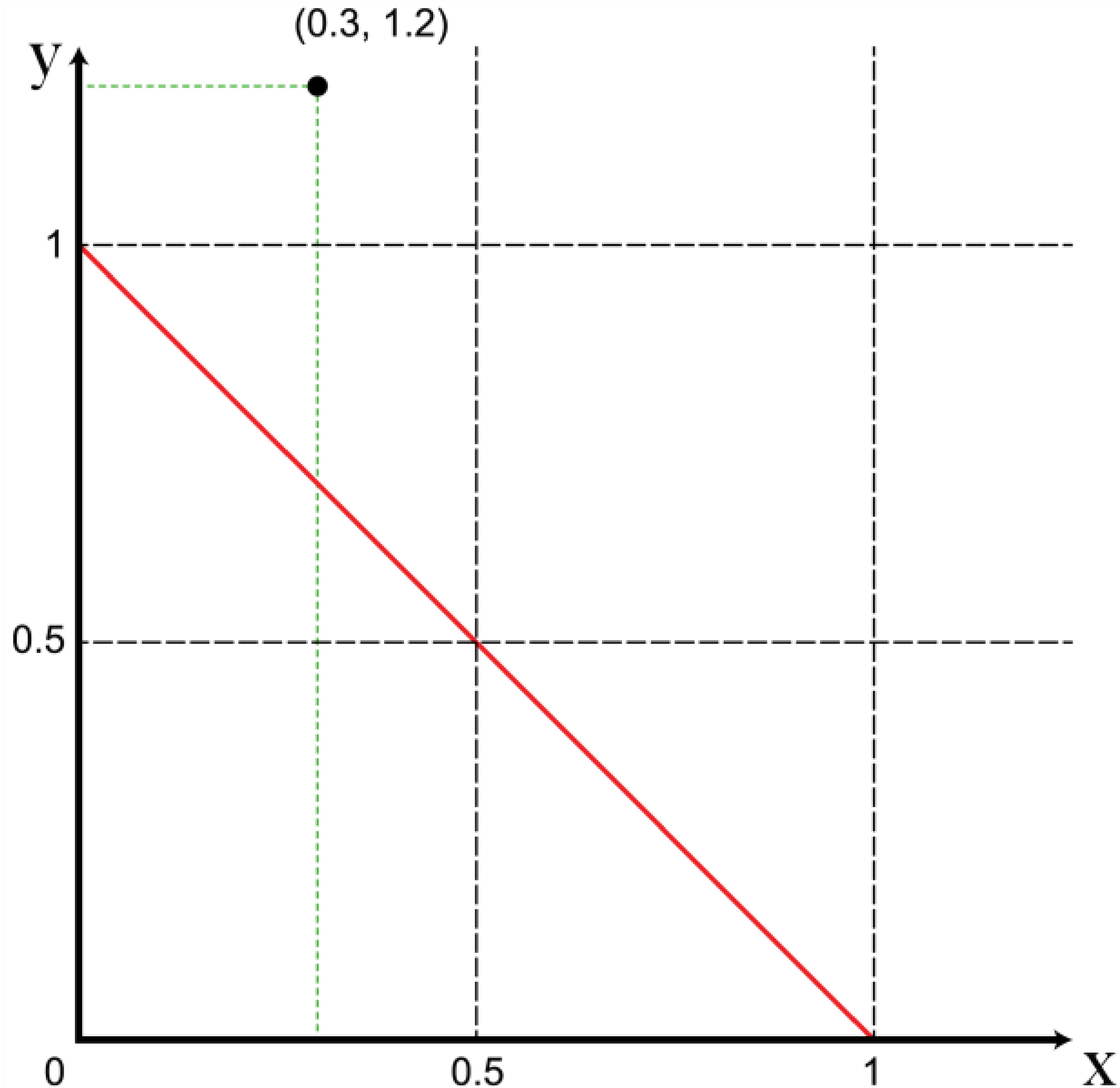
All two-component compositions (which must sum to a whole) can be represented on the line.



The simplex

However, we usually measure each of our variables in R_+.

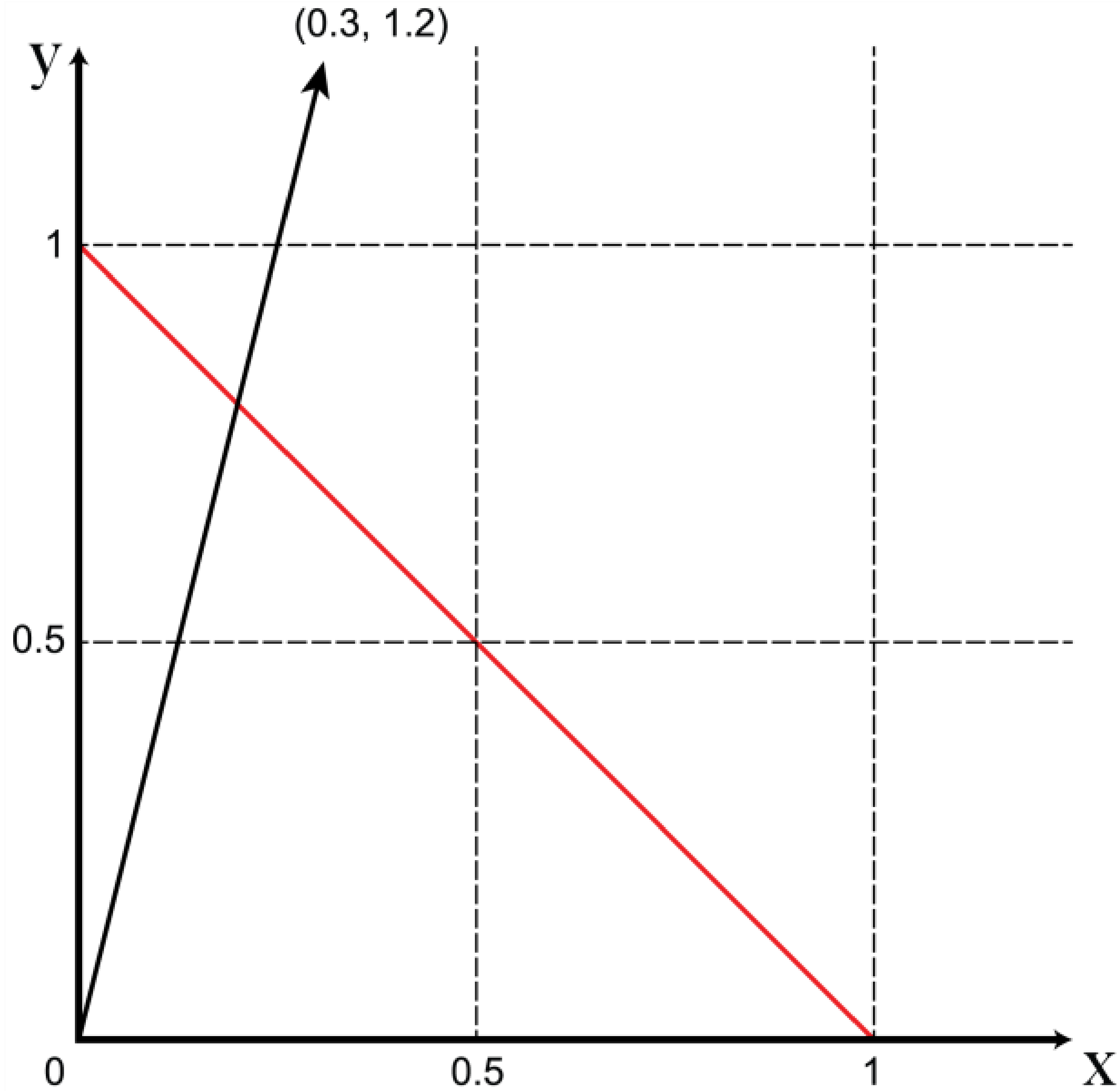
Examples include point-counts of minerals, ion beam currents, etc.



The simplex

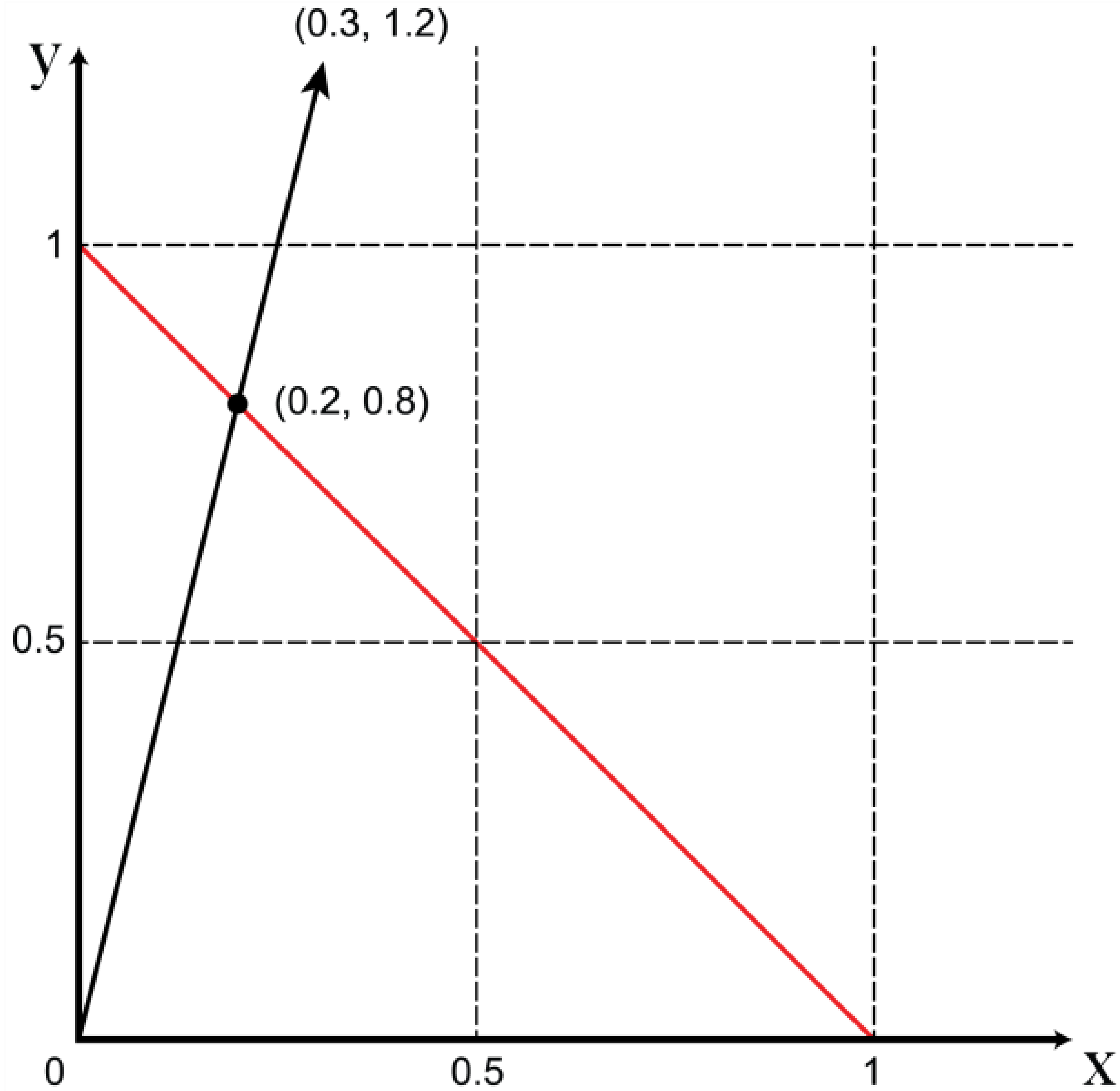
The measured variables can also be represented by a vector from the origin to the coordinates of the measurement.

This vector is called a **basis**.

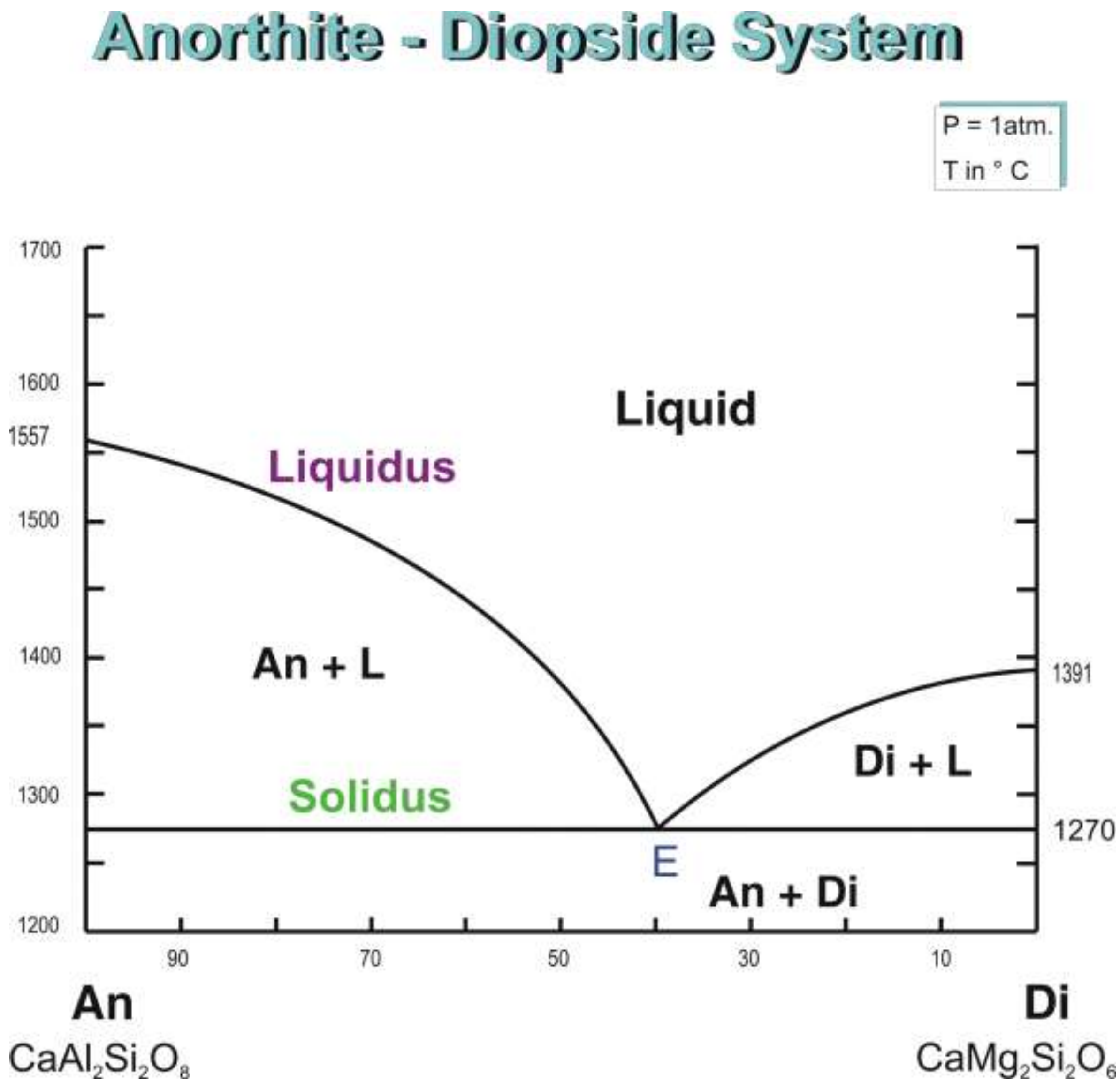


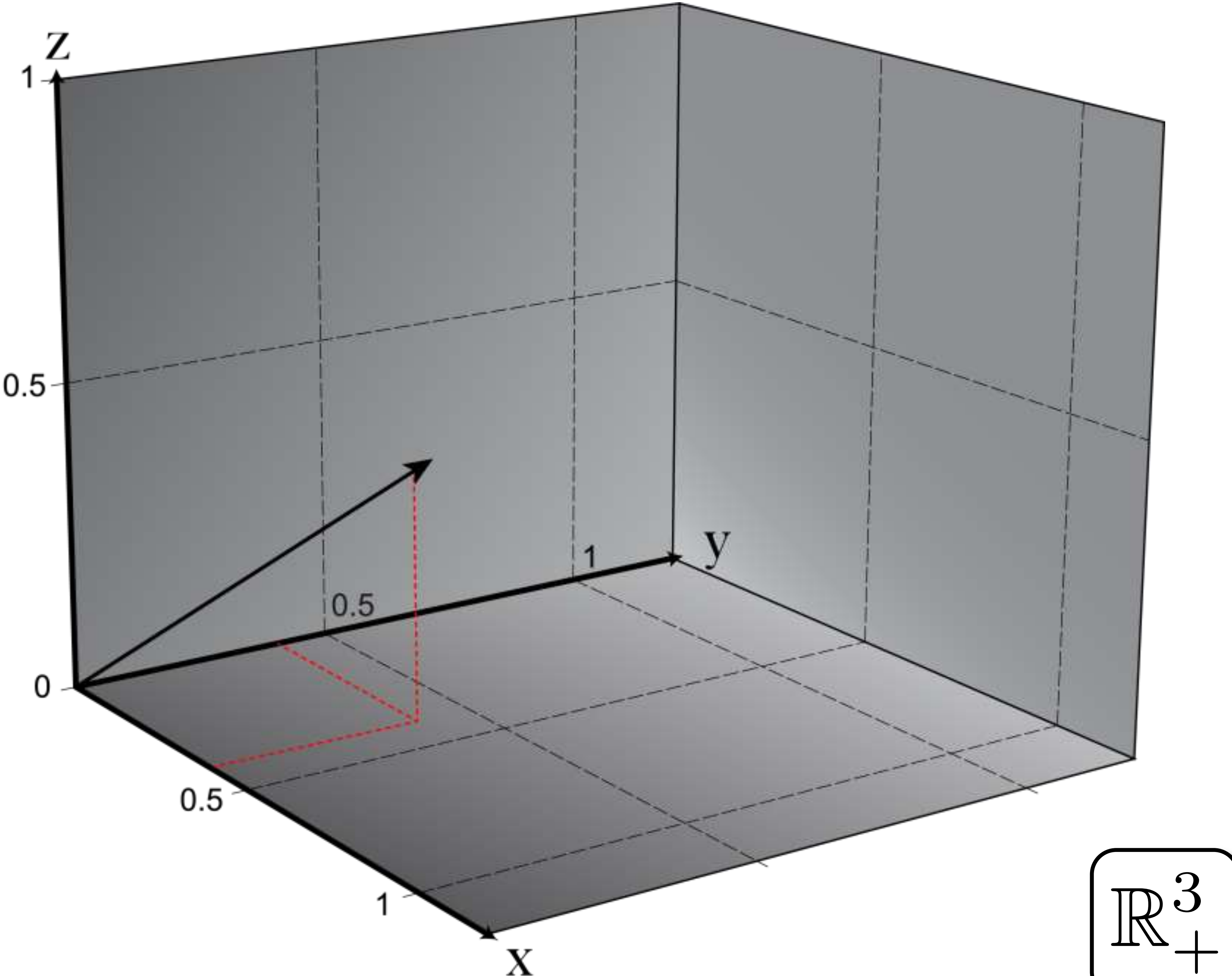
The simplex

Where the measurement vector intersects the simplex (its projection onto the simplex) is the composition, with the unit sum constraint enforced.

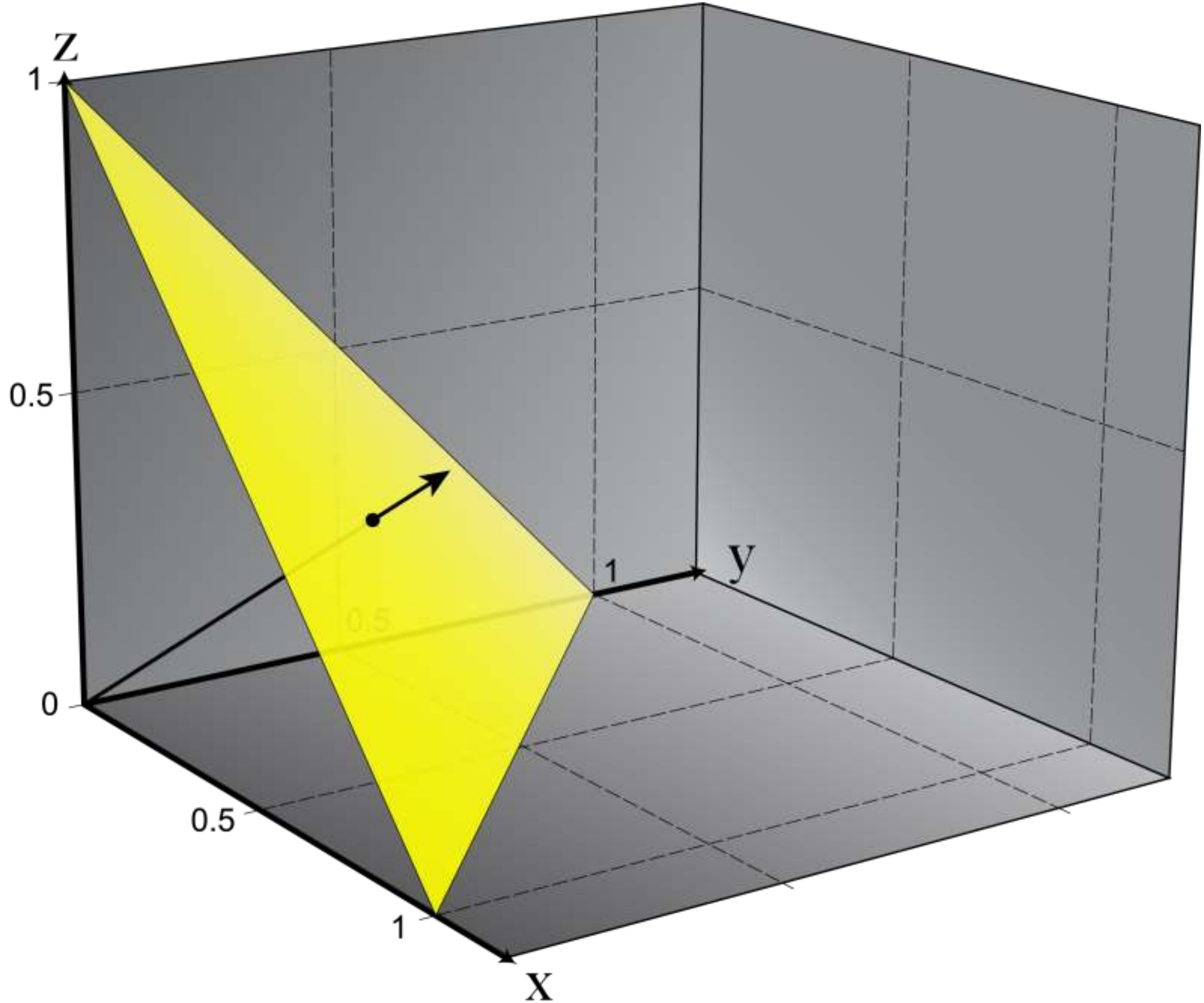


The (1D) simplex:

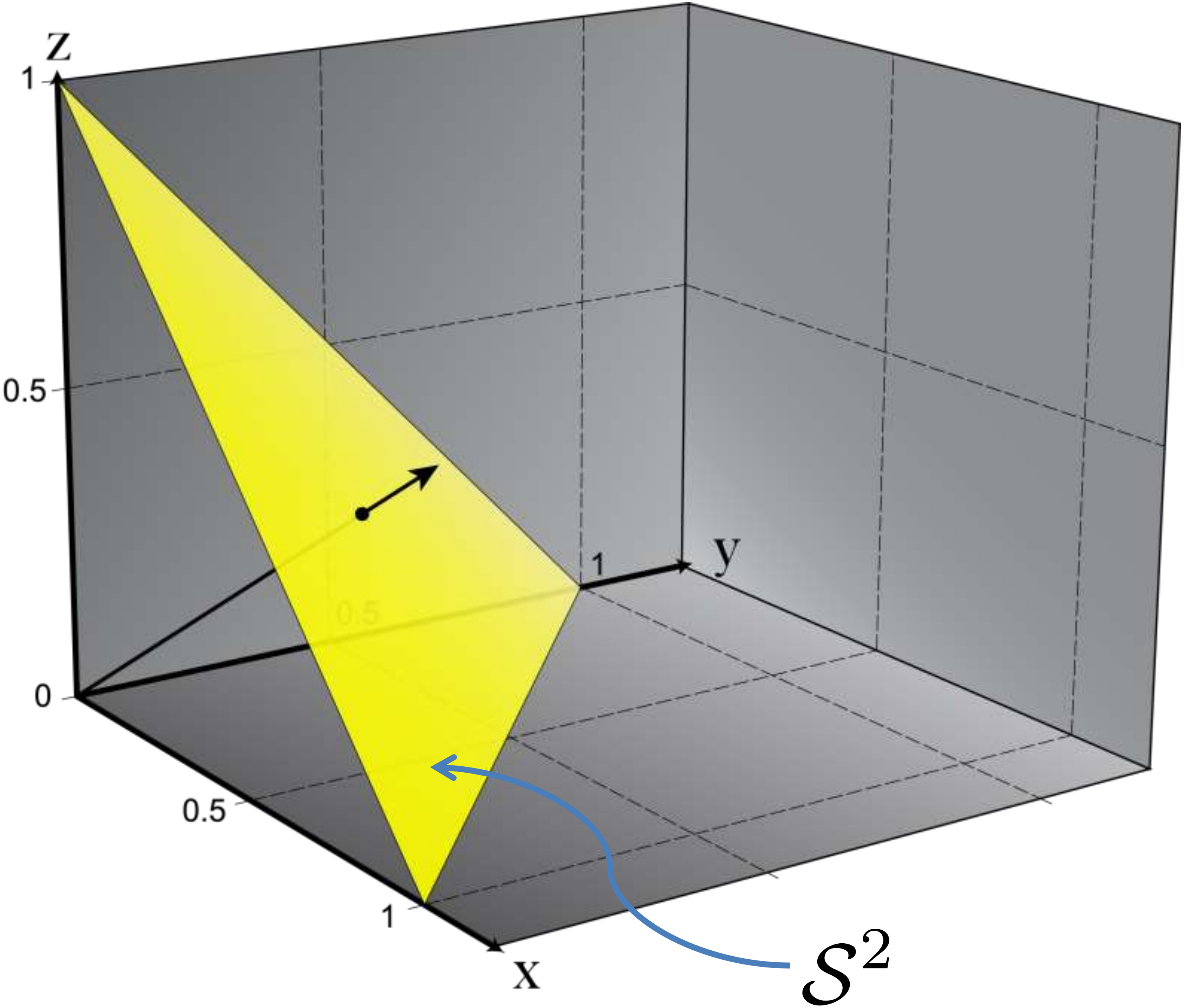




z
1
0.5
0
y
1
0.5
x
0.5
1
$\mathbb{R}^3_+$



z
1
0.5
0
0.5
1
x
0.5
1
y



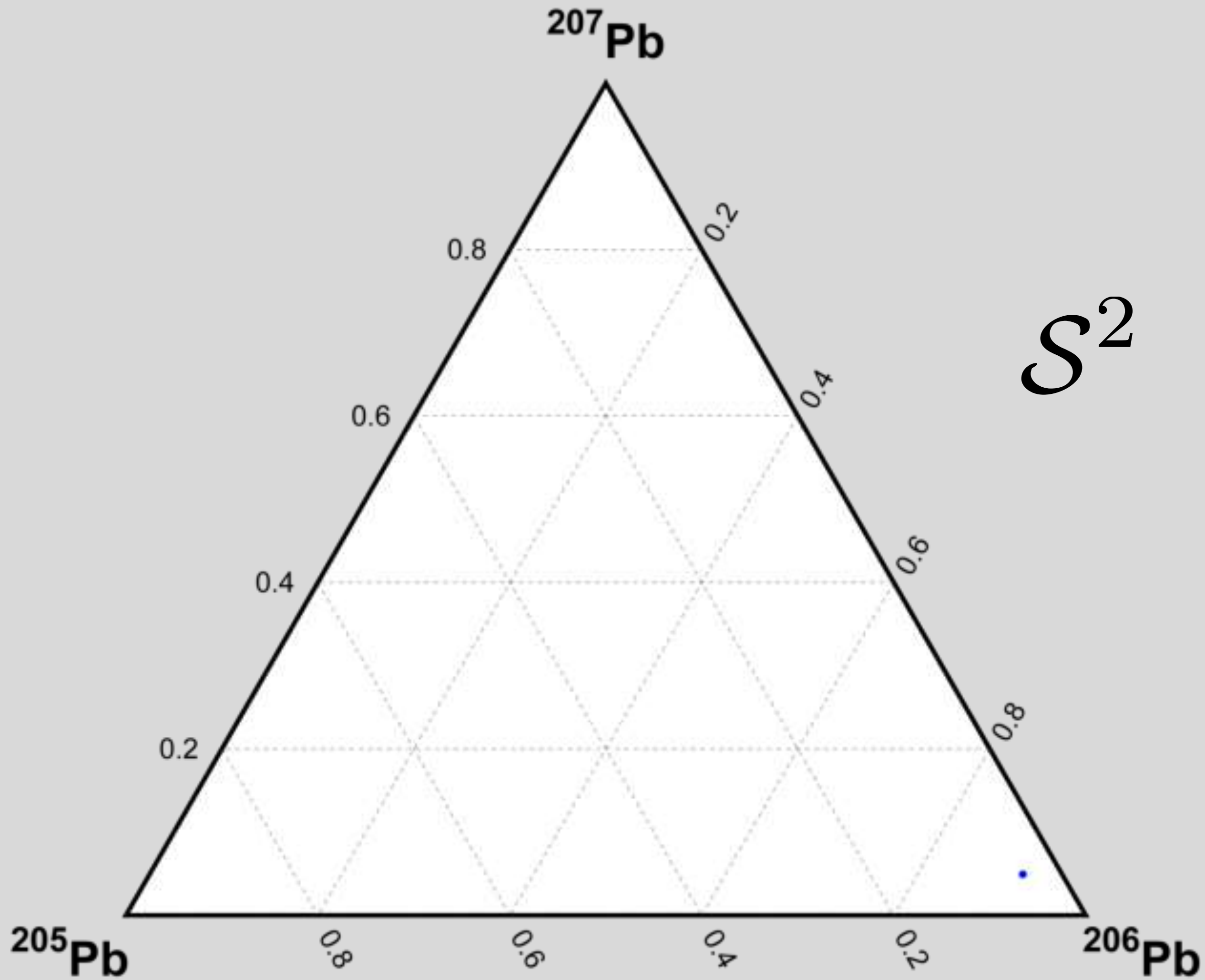
z
1
0.5
0
y
1
0.5
0.5
1
x
$\mathcal{S}^2$

Ternary plots

Express relative proportions of three components in a 2D space—they are a simplex.

Each vertex corresponds to a ‘pure’ end-member composition.

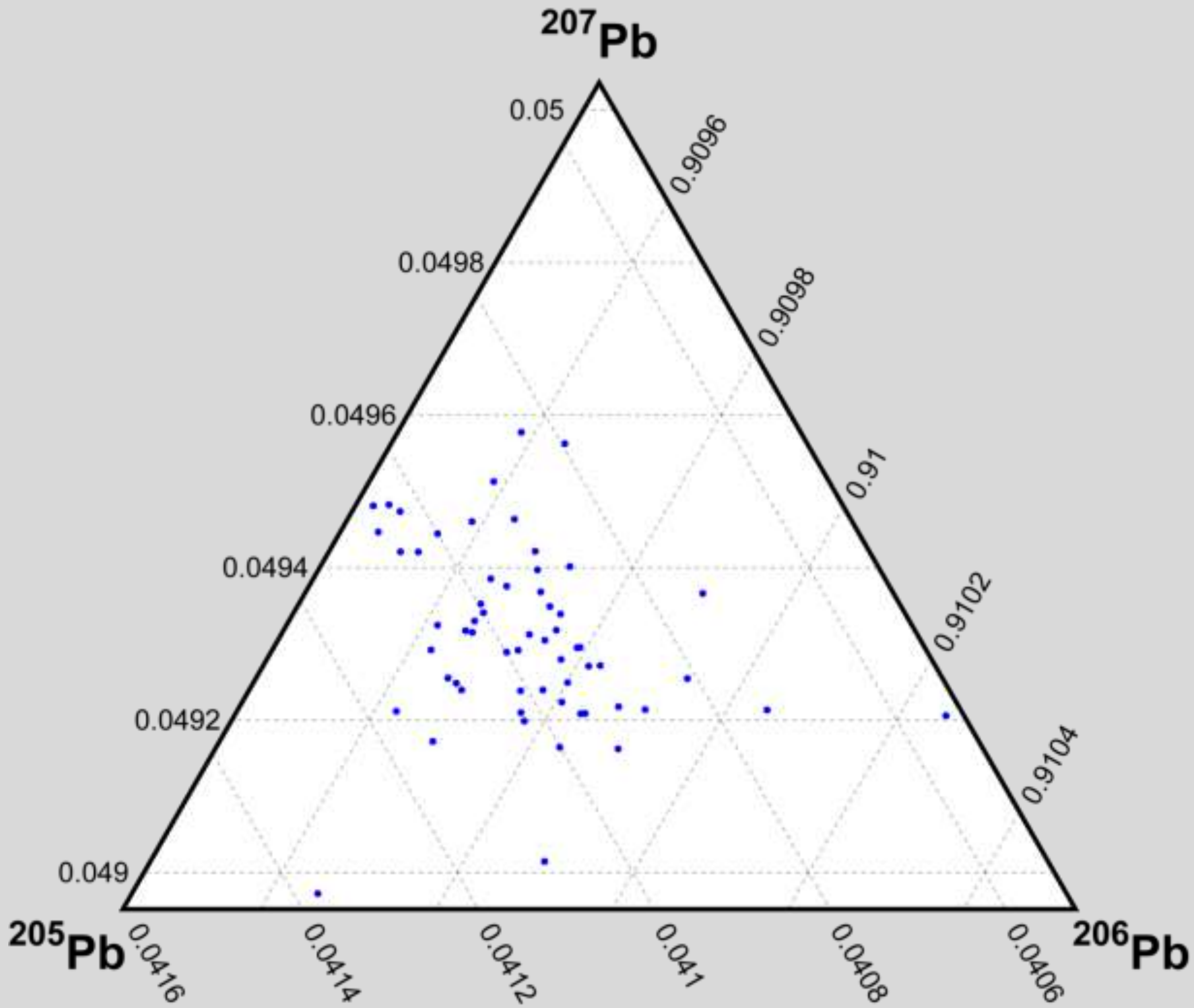
In the middle of the plot, the closer you are to a vertex, the greater the relative proportion of that component.



$\mathcal{S}^2$

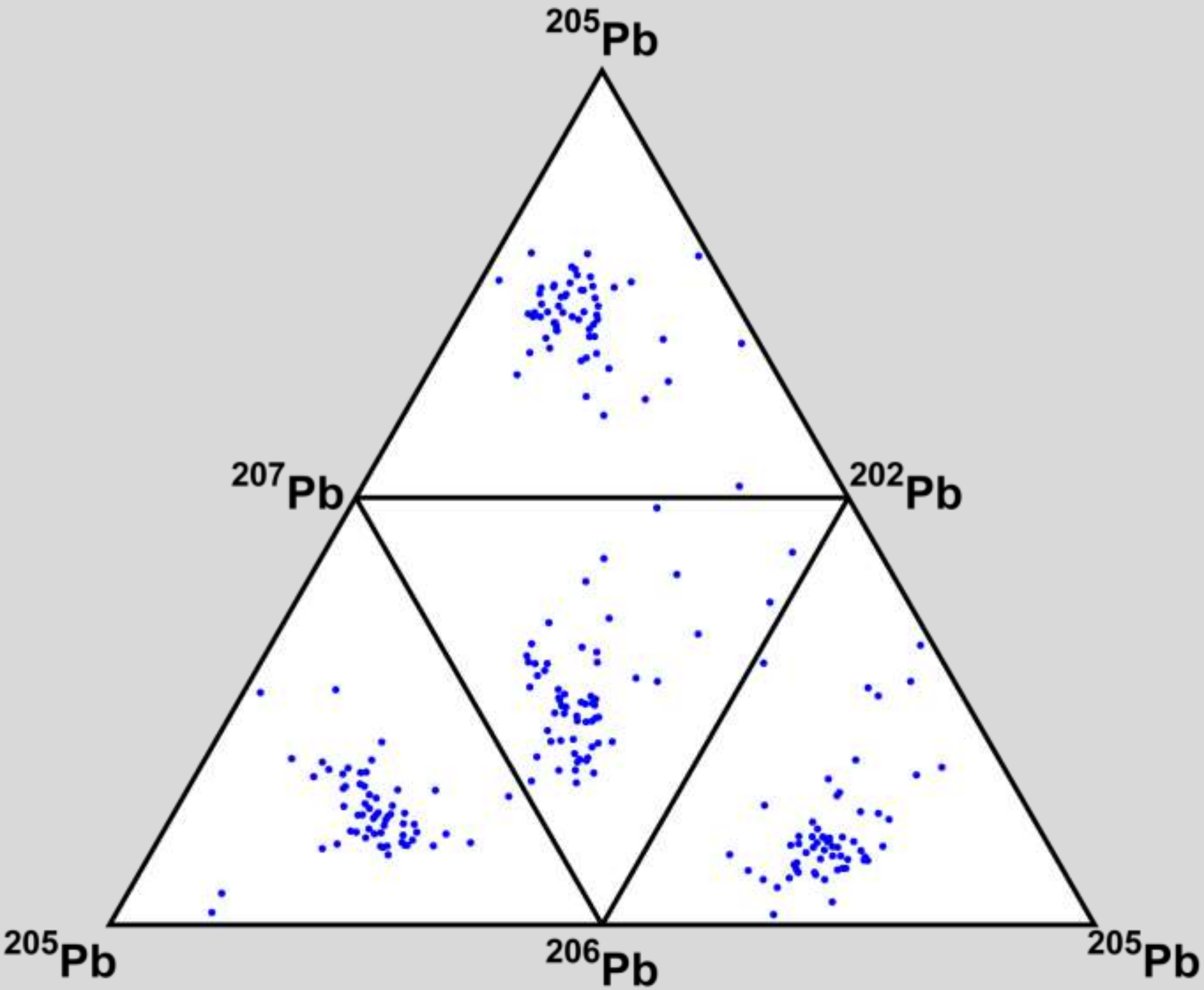
Ternary plots

Ternary plots do not need to plot the whole of compositional data space—you can zoom in to better display data.



(Quad-) Ternary Plots(?)

To visualize more than three components and stay on a (2D) page or screen, you can link multiple ternary plots together along their edges.



Compositional data presents unique problems.

- Because we recognize the compositional nature of our geochemical and geological datasets, we usually perform some kind of normalization when reporting data and performing statistical analysis.

Compositional data presents unique problems.

- Because we recognize the compositional nature of our geochemical and geological datasets, we usually perform some kind of normalization when reporting data and performing statistical analysis.
  - Weight % oxides, isotope ratios, deviations from a standard expressed in δ or ε notation

Compositional data presents unique problems.

- However, all of these approaches have drawbacks when you go to evaluate a mean and standard deviation/error/covariance matrix.

Simple test data: isotope ratios

A	B	C		B/A	C/A	A/B	B/C
3.8816	2.2237	3.5034		0.5729	0.9026	1.7456	0.6347
3.4189	3.6334	4.4136		1.0627	1.2909	0.9410	0.8232
1.8736	3.4878	6.2357		1.8615	3.3282	0.5372	0.5593
2.7661	8.5963	4.0573		3.1077	1.4668	0.3218	2.1187
2.7887	3.5317	7.1290		1.2664	2.5564	0.7896	0.4954
2.2993	2.3495	6.8411		1.0218	2.9753	0.9786	0.3434
6.9564	8.9176	1.8384		1.2819	0.2643	0.7801	4.8507
1.9362	7.4160	2.3421		3.8302	1.2096	0.2611	3.1664
2.3554	3.5661	2.7637		1.5140	1.1733	0.6605	1.2903
2.1014	1.8837	3.3307		0.8964	1.5850	1.1156	0.5656

Simple test data: isotope ratios

A	B	C		B/A	C/A	A/B	B/C
3.8816	2.2237	3.5034		0.5729	0.9026	1.7456	0.6347
3.4189	3.6334	4.4136		1.0627	1.2909	0.9410	0.8232
1.8736	3.4878	6.2357		1.8615	3.3282	0.5372	0.5593
2.7661	8.5963	4.0573		3.1077	1.4668	0.3218	2.1187
2.7887	3.5317	7.1290		1.2664	2.5564	0.7896	0.4954
2.2993	2.3495	6.8411		1.0218	2.9753	0.9786	0.3434
6.9564	8.9176	1.8384		1.2819	0.2643	0.7801	4.8507
1.9362	7.4160	2.3421		3.8302	1.2096	0.2611	3.1664
2.3554	3.5661	2.7637		1.5140	1.1733	0.6605	1.2903
2.1014	1.8837	3.3307		0.8964	1.5850	1.1156	0.5656
			mean:	**1.6416**	**1.6752**	**0.8131**	**1.4848**

Simple test data: isotope ratios

A	B	C		B/A	C/A	A/B	B/C
3.8816	2.2237	3.5034		0.5729	0.9026	1.7456	0.6347
3.4189	3.6334	4.4136		1.0627	1.2909	0.9410	0.8232
1.8736	3.4878	6.2357		1.8615	3.3282	0.5372	0.5593
2.7661	8.5963	4.0573		3.1077	1.4668	0.3218	2.1187
2.7887	3.5317	7.1290		1.2664	2.5564	0.7896	0.4954
2.2993	2.3495	6.8411		1.0218	2.9753	0.9786	0.3434
6.9564	8.9176	1.8384		1.2819	0.2643	0.7801	4.8507
1.9362	7.4160	2.3421		3.8302	1.2096	0.2611	3.1664
2.3554	3.5661	2.7637		1.5140	1.1733	0.6605	1.2903
2.1014	1.8837	3.3307		0.8964	1.5850	1.1156	0.5656
			mean:	**1.6416**	**1.6752**	**0.8131**	**1.4848**
			1/(B/A):			**1.2299**	

Simple test data: isotope ratios

A	B	C		B/A	C/A	A/B	B/C
3.8816	2.2237	3.5034		0.5729	0.9026	1.7456	0.6347
3.4189	3.6334	4.4136		1.0627	1.2909	0.9410	0.8232
1.8736	3.4878	6.2357		1.8615	3.3282	0.5372	0.5593
2.7661	8.5963	4.0573		3.1077	1.4668	0.3218	2.1187
2.7887	3.5317	7.1290		1.2664	2.5564	0.7896	0.4954
2.2993	2.3495	6.8411		1.0218	2.9753	0.9786	0.3434
6.9564	8.9176	1.8384		1.2819	0.2643	0.7801	4.8507
1.9362	7.4160	2.3421		3.8302	1.2096	0.2611	3.1664
2.3554	3.5661	2.7637		1.5140	1.1733	0.6605	1.2903
2.1014	1.8837	3.3307		0.8964	1.5850	1.1156	0.5656
			mean:	**1.6416**	**1.6752**	**0.8131**	**1.4848**
			1/(B/A):			**1.2299**	
			(B/A)/(C/A)	**0.9799**			

Simple test data: log-ratios:

A	B	C		log(B/A)	log(C/A)	log(A/B)	log(B/C)
3.8816	2.2237	3.5034		-0.5571	-0.1025	0.5571	-0.4546
3.4189	3.6334	4.4136		0.0608	0.2554	-0.0608	-0.1945
1.8736	3.4878	6.2357		0.6214	1.2024	-0.6214	-0.5810
2.7661	8.5963	4.0573		1.1339	0.3831	-1.1339	0.7508
2.7887	3.5317	7.1290		0.2362	0.9386	-0.2362	-0.7024
2.2993	2.3495	6.8411		0.0216	1.0903	-0.0216	-1.0687
6.9564	8.9176	1.8384		0.2484	-1.3308	-0.2484	1.5791
1.9362	7.4160	2.3421		1.3429	0.1903	-1.3429	1.1526
2.3554	3.5661	2.7637		0.4148	0.1599	-0.4148	0.2549
2.1014	1.8837	3.3307		-0.1094	0.4606	0.1094	-0.5699

Simple test data: log-ratios:

A	B	C		log(B/A)	log(C/A)	log(A/B)	log(B/C)
3.8816	2.2237	3.5034		-0.5571	-0.1025	0.5571	-0.4546
3.4189	3.6334	4.4136		0.0608	0.2554	-0.0608	-0.1945
1.8736	3.4878	6.2357		0.6214	1.2024	-0.6214	-0.5810
2.7661	8.5963	4.0573		1.1339	0.3831	-1.1339	0.7508
2.7887	3.5317	7.1290		0.2362	0.9386	-0.2362	-0.7024
2.2993	2.3495	6.8411		0.0216	1.0903	-0.0216	-1.0687
6.9564	8.9176	1.8384		0.2484	-1.3308	-0.2484	1.5791
1.9362	7.4160	2.3421		1.3429	0.1903	-1.3429	1.1526
2.3554	3.5661	2.7637		0.4148	0.1599	-0.4148	0.2549
2.1014	1.8837	3.3307		-0.1094	0.4606	0.1094	-0.5699
			mean log-ratio	**0.3414**	**0.3247**	**-0.3414**	**0.0166**

Simple test data: log-ratios:

A	B	C		log(B/A)	log(C/A)	log(A/B)	log(B/C)
3.8816	2.2237	3.5034		-0.5571	-0.1025	0.5571	-0.4546
3.4189	3.6334	4.4136		0.0608	0.2554	-0.0608	-0.1945
1.8736	3.4878	6.2357		0.6214	1.2024	-0.6214	-0.5810
2.7661	8.5963	4.0573		1.1339	0.3831	-1.1339	0.7508
2.7887	3.5317	7.1290		0.2362	0.9386	-0.2362	-0.7024
2.2993	2.3495	6.8411		0.0216	1.0903	-0.0216	-1.0687
6.9564	8.9176	1.8384		0.2484	-1.3308	-0.2484	1.5791
1.9362	7.4160	2.3421		1.3429	0.1903	-1.3429	1.1526
2.3554	3.5661	2.7637		0.4148	0.1599	-0.4148	0.2549
2.1014	1.8837	3.3307		-0.1094	0.4606	0.1094	-0.5699
			mean log-ratio	**0.3414**	**0.3247**	**-0.3414**	**0.0166**
			mean(B/A)$^{-1}$	**-0.3414**			
			(B/A)/(C/A)	**0.0166**			

Simple test data: log-ratios:

A	B	C		log(B/A)	log(C/A)	log(A/B)	log(B/C)
3.8816	2.2237	3.5034		-0.5571	-0.1025	0.5571	-0.4546
3.4189	3.6334	4.4136		0.0608	0.2554	-0.0608	-0.1945
1.8736	3.4878	6.2357		0.6214	1.2024	-0.6214	-0.5810
2.7661	8.5963	4.0573		1.1339	0.3831	-1.1339	0.7508
2.7887	3.5317	7.1290		0.2362	0.9386	-0.2362	-0.7024
2.2993	2.3495	6.8411		0.0216	1.0903	-0.0216	-1.0687
6.9564	8.9176	1.8384		0.2484	-1.3308	-0.2484	1.5791
1.9362	7.4160	2.3421		1.3429	0.1903	-1.3429	1.1526
2.3554	3.5661	2.7637		0.4148	0.1599	-0.4148	0.2549
2.1014	1.8837	3.3307		-0.1094	0.4606	0.1094	-0.5699
			mean log-ratio	**0.3414**	**0.3247**	**-0.3414**	**0.0166**
			mean(B/A)$^{-1}$	**-0.3414**			
			(B/A)/(C/A)	**0.0166**			
			mean ratio:	**1.4069**	**1.3837**	**0.7108**	**1.0168**

Simple test data: log-ratios:

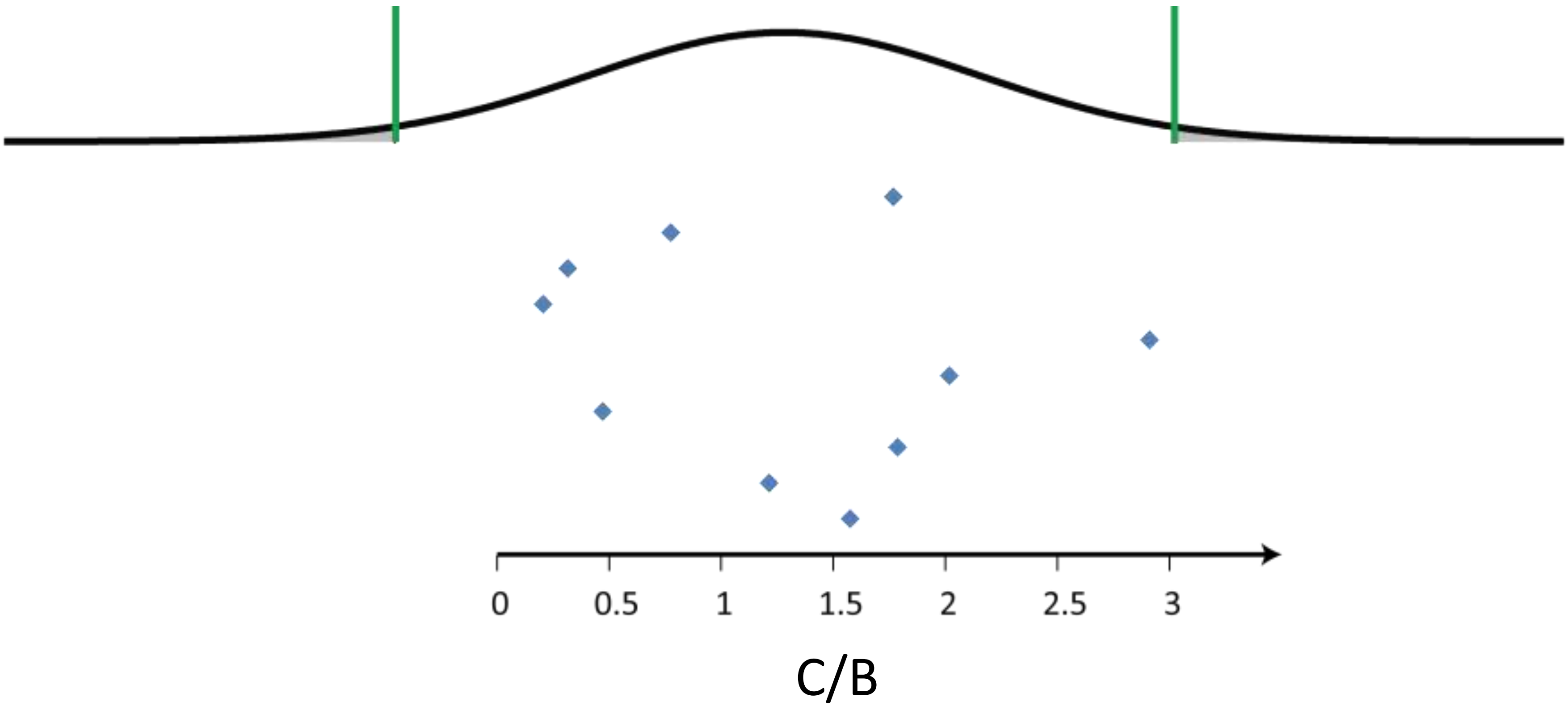
A	B	C		log(B/A)	log(C/A)	log(A/B)	log(B/C)
3.8816	2.2237	3.5034		-0.5571	-0.1025	0.5571	-0.4546
3.4189	3.6334	4.4136		0.0608	0.2554	-0.0608	-0.1945
1.8736	3.4878	6.2357		0.6214	1.2024	-0.6214	-0.5810
2.7661	8.5963	4.0573		1.1339	0.3831	-1.1339	0.7508
2.7887	3.5317	7.1290		0.2362	0.9386	-0.2362	-0.7024
2.2993	2.3495	6.8411		0.0216	1.0903	-0.0216	-1.0687
6.9564	8.9176	1.8384		0.2484	-1.3308	-0.2484	1.5791
1.9362	7.4160	2.3421		1.3429	0.1903	-1.3429	1.1526
2.3554	3.5661	2.7637		0.4148	0.1599	-0.4148	0.2549
2.1014	1.8837	3.3307		-0.1094	0.4606	0.1094	-0.5699
			mean log-ratio	**0.3414**	**0.3247**	**-0.3414**	**0.0166**
			$mean(B/A)^{-1}$	**-0.3414**			
			(B/A)/(C/A)	**0.0166**			
			mean ratio:	**1.4069**	**1.3837**	**0.7108**	**1.0168**
			1/(B/A):	**0.7108**			
			(B/A)/(C/A):	**1.0168**			

Another problem:

- The normal distribution calculated by taking the mean and standard deviation of compositional data will not stay on the simplex: the domain of a normal distribution is $\mathbb{R}$

Another problem:

- The normal distribution calculated by taking the mean and standard deviation of compositional data will not stay on the simplex: the domain of a normal distribution is $\mathbb{R}$



A better statement of the problem:

- Measures of difference are measures of distance.
- All of our statistics so far have boiled down to “all you need is S”

$$S = \sum \frac{(x_i - \bar{x})^2}{\sigma_i^2}$$

A better statement of the problem:

- Distances—d(x,X)—should have six properties (Aitchison, 1992):
  1. Positivity d(x,X) > 0 if X is not the same as x
  2. Zero difference between equivalent compositions, d(x,X) = 0 if x=X
  3. Interchangeability f(x,X) = f(X,x)
  4. Scale invariance f(ax,aX) = f(x,X)
  5. Perturbation invariance
  6. Permutation invariance

The (easiest) solution

- If we want to keep using the normal distribution, and the well-developed statistical framework that goes along with it, we need to **transform** our data out of the simplex and into the real numbers

The solution

- If we want to keep using the normal distribution, and the well-developed statistical framework that goes along with it, we need to **transform** our data out of the simplex and into the real numbers

Additive log-ratio transform:

1. Evaluate ratios of components with a common component in the denominator
   - $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, $^{208}Pb/^{204}Pb$

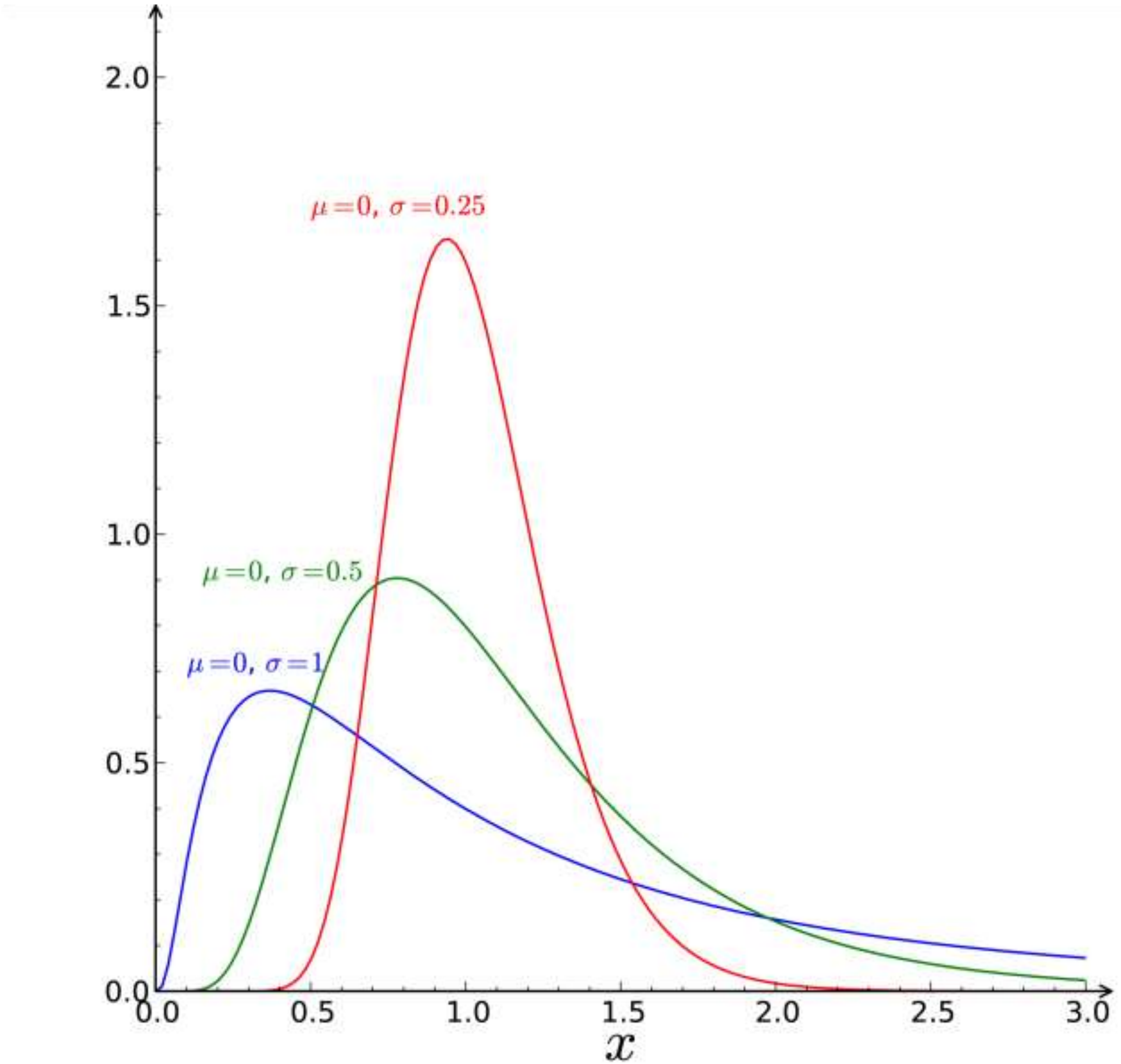
Additive log-ratio transform:

1. Evaluate ratios of components with a common component in the denominator
   - $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, $^{208}Pb/^{204}Pb$
2. Take the logarithm of each
   - $\log(^{206}Pb/^{204}Pb)$, $\log(^{207}Pb/^{204}Pb)$, $\log(^{208}Pb/^{204}Pb)$

Additive log-ratio transform:

1. Evaluate ratios of components with a common component in the denominator
   - $^{206}Pb/^{204}Pb$, $^{207}Pb/^{204}Pb$, $^{208}Pb/^{204}Pb$
2. Take the logarithm of each
   - $\log(^{206}Pb/^{204}Pb)$, $\log(^{207}Pb/^{204}Pb)$, $\log(^{208}Pb/^{204}Pb)$
3. Assume (or test that) the resulting log-ratios are normally distributed

The lognormal distribution



Consequences

- Since the additive log-ratio transformed data is normally distributed, proceed with your calculations as before, just evaluate statistics (mean, standard deviation, etc) on log-ratio data.
- When you're done, 'undo' the transform by evaluating an exponential:

$$\exp(\log(x/y)) = x/y$$

Consequences

- Log-normal distributions that are precise and far from zero look much like normal distributions, and can be assigned symmetric $\pm 2\sigma$ confidence intervals.
- Those that are close to zero and less precise have asymmetric probability distribution functions.

More consequences

- The linear regression technique that we used before does not work for data plotted as isotope ratios. This goes for linear arrays in isotope ratio space, like isochrons and mixing lines.

More consequences

- The linear regression technique that we used before does not work for data plotted as isotope ratios. This goes for linear arrays in isotope ratio space, like isochrons and mixing lines.
- The answer is to transform the data into **log-ratio space**, where x- and y-variables can be given multivariate normal distributions, then perform non-linear regression.