Background on Compositional Data

UBIQUITY OF
COMPOSITIONAL DATA

Compositional: Relating to parts of some whole

Proportions
Parts per million
Percentages

X+Y+Z=k
And all Positive

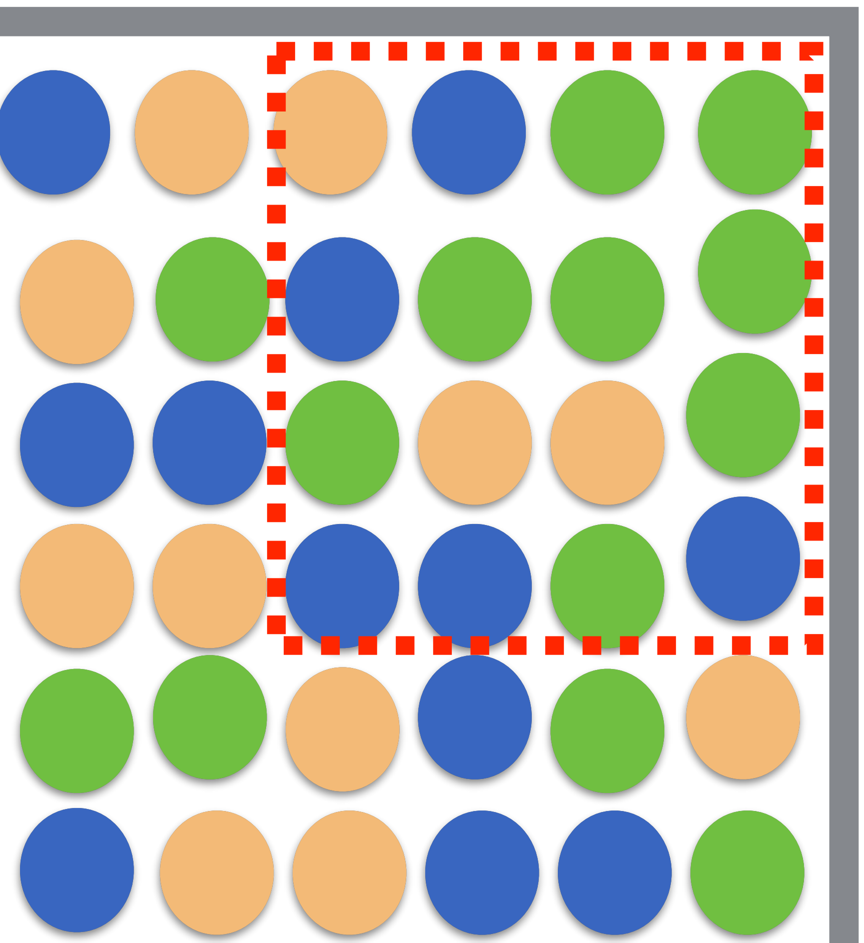
RELATIVE *DATA*

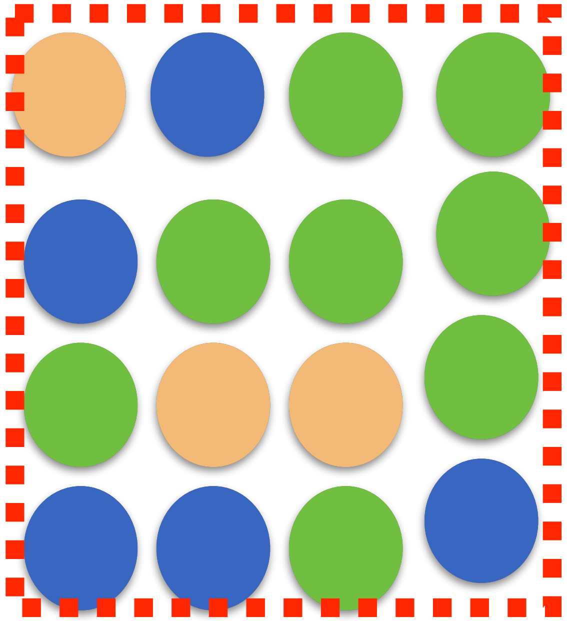
Simple Examples

- Does hongite have more calcium than struvite? (*e.g., parts per million*)
- Have I been spending more of my day in the bathroom since I ate that sandwich? (*e.g., percentage of your day*)
- Does my cow produce higher protein milk when I feed her that sandwich? (*e.g.,* proportion of calories from protein)

UBIQUITY OF COMPOSITIONAL DATA
COUNTS!

1200 Blue
1100 Orange
1300 Green





31% Blue
19% Orange
50% Green

UBIQUITY OF COMPOSITIONAL DATA
COUNTS!

Examples

- Abundance Quantification by High Throughout Sequencing
  - **Microbiome Composition (*e.g.*, counts of 16s gene)**
  - Gene expression analysis (*e.g.*, RNA-seq)
- Abundance Quantification by Flow Cytometry
- Proportion of observed mice that go on to develop a disease?
- Population of North Carolina that is pro-Trump? (*e.g., polling results)*

RESULTING
COUNT TABLE

	Species 1	Species 2	Species 3	Species 4	Species 5	Species 6	Species 7	Species 8	Species 9	Species 10
Sample 1	23	53	2	44	10	88	94	66	73	67
Sample 2	69	64	70	47	8	97	47	6	64	19
Sample 3	33	100	68	78	59	87	71	31	67	24
Sample 4	5	63	57	27	86	81	83	92	46	62
Sample 5	76	80	46	70	92	92	6	46	37	68
Sample 6	58	7	37	45	25	62	78	44	89	30
Sample 7	10	87	32	80	9	91	59	90	67	77
Sample 8	21	89	73	39	44	80	97	83	80	4
Sample 9	85	77	82	72	15	19	44	4	83	76
Sample 10	67	87	68	58	73	29	87	4	48	79
Sample 11	90	5	28	49	39	20	78	92	12	23
Sample 12	98	93	55	12	54	75	27	95	83	98
Sample 13	31	97	52	9	93	84	45	97	81	27
Sample 14	12	77	22	17	71	12	56	86	18	0
Sample 15	40	30	71	71	54	13	77	96	75	11
Sample 16	43	94	40	73	27	33	97	88	81	44

The Shape of Compositional Data (Microbiome Example)

compositional data: usual representation

definition: $\mathbf{x} = [x_1, x_2, \ldots, x_D]$ is a D-part composition

$$\begin{cases} x_i > 0, & \text{for all } i = 1, \ldots, D \\ \sum_{i=1}^{D} x_i = \kappa & \text{(constant)} \end{cases}$$

$\kappa = 1 \iff$ measurements in parts per unit
$\kappa = 100 \iff$ measurements in percent
other frequent units: ppm, ppb, ...

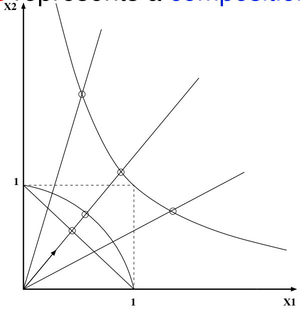
a composition is the representative in the simplex of equivalent vectors with strictly positive components

a **subcomposition** $\mathbf{x}_s$ with s parts is obtained as the closure of a subvector $[x_{i_1}, x_{i_2}, \ldots, x_{i_s}]$ of $\mathbf{x}$

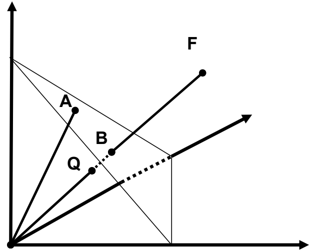
compositional data: definition

definition: parts of some whole which carry only **relative information**

Proportional vectors with strictly positive components are compositionally equivalent if they are proportional: each **equivalence class** represents a composition



compositional data in $\mathbb{R}^2$



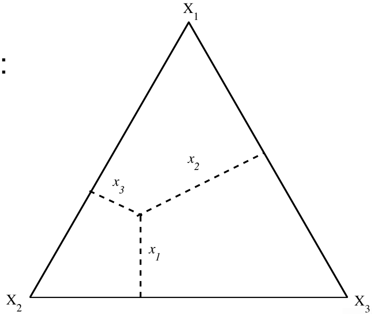
compositional data in $\mathbb{R}^3$

usual representation: subject to a **constant sum constraint**

the simplex as sample space

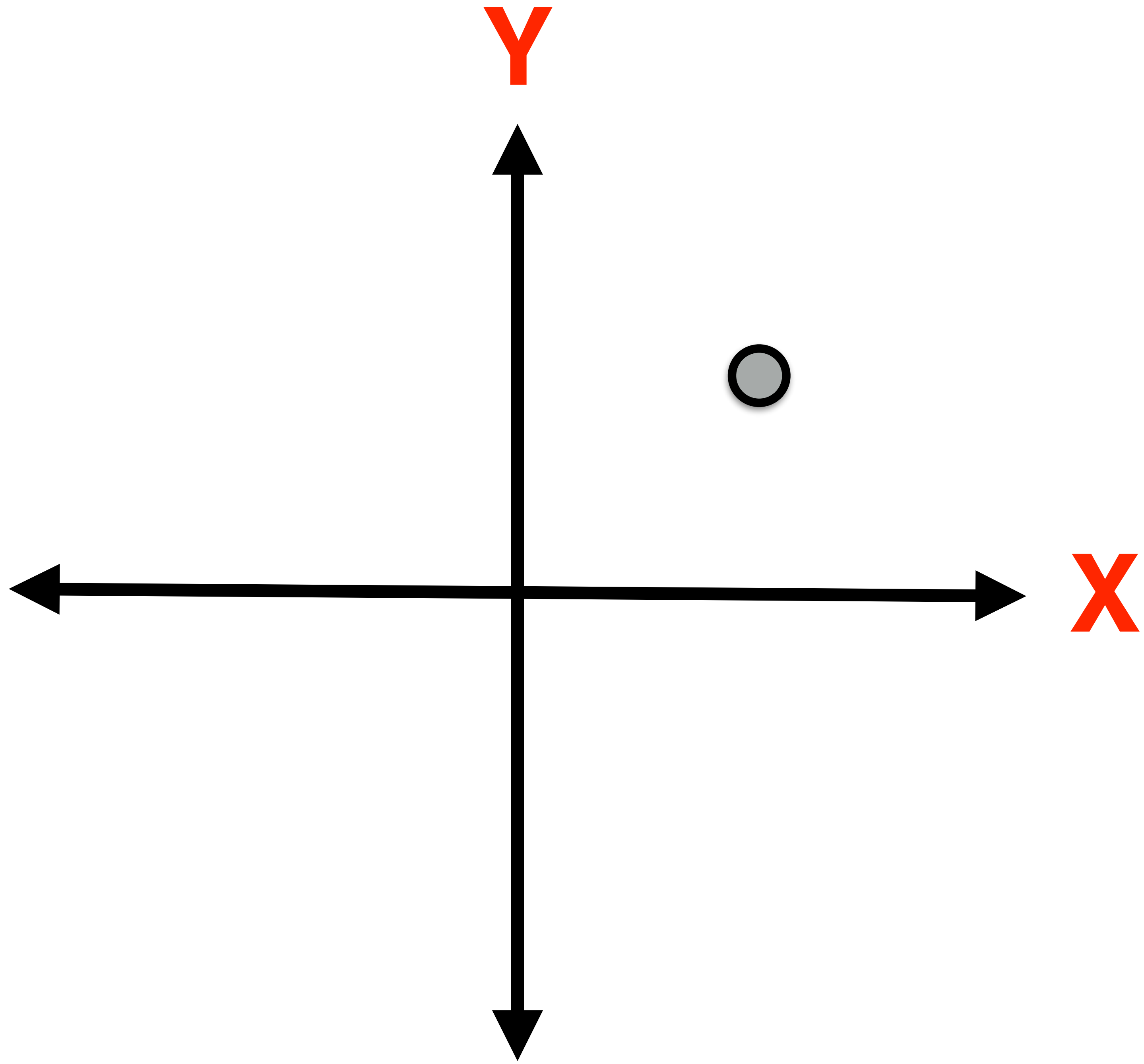
$$\mathcal{S}^D = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_D] \,\middle|\, x_i > 0; \sum_{i=1}^{D} x_i = \kappa \right\}$$

standard representation for $D = 3$:
the ternary diagram

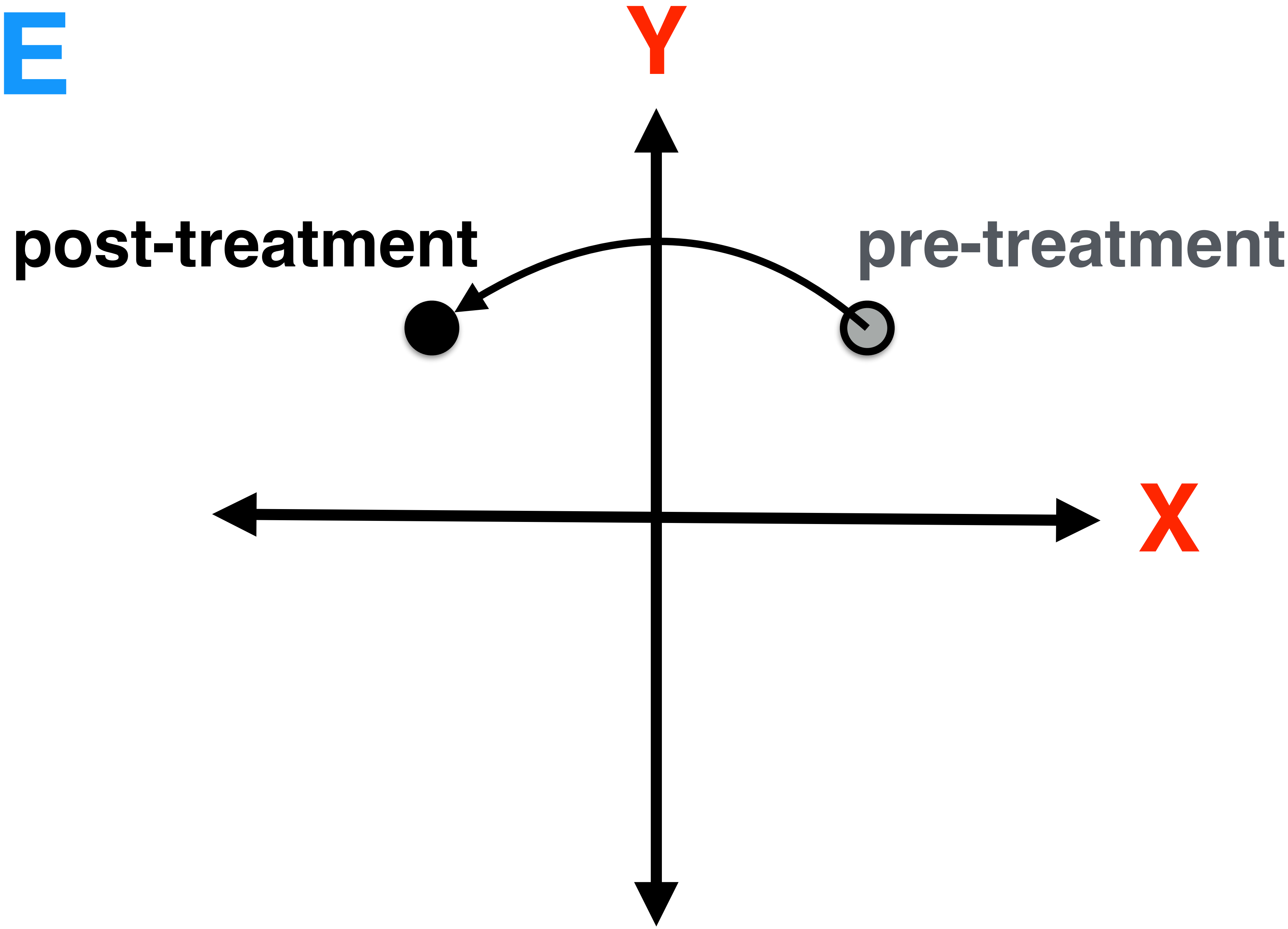


STATISTICAL

COMFORT ZONE

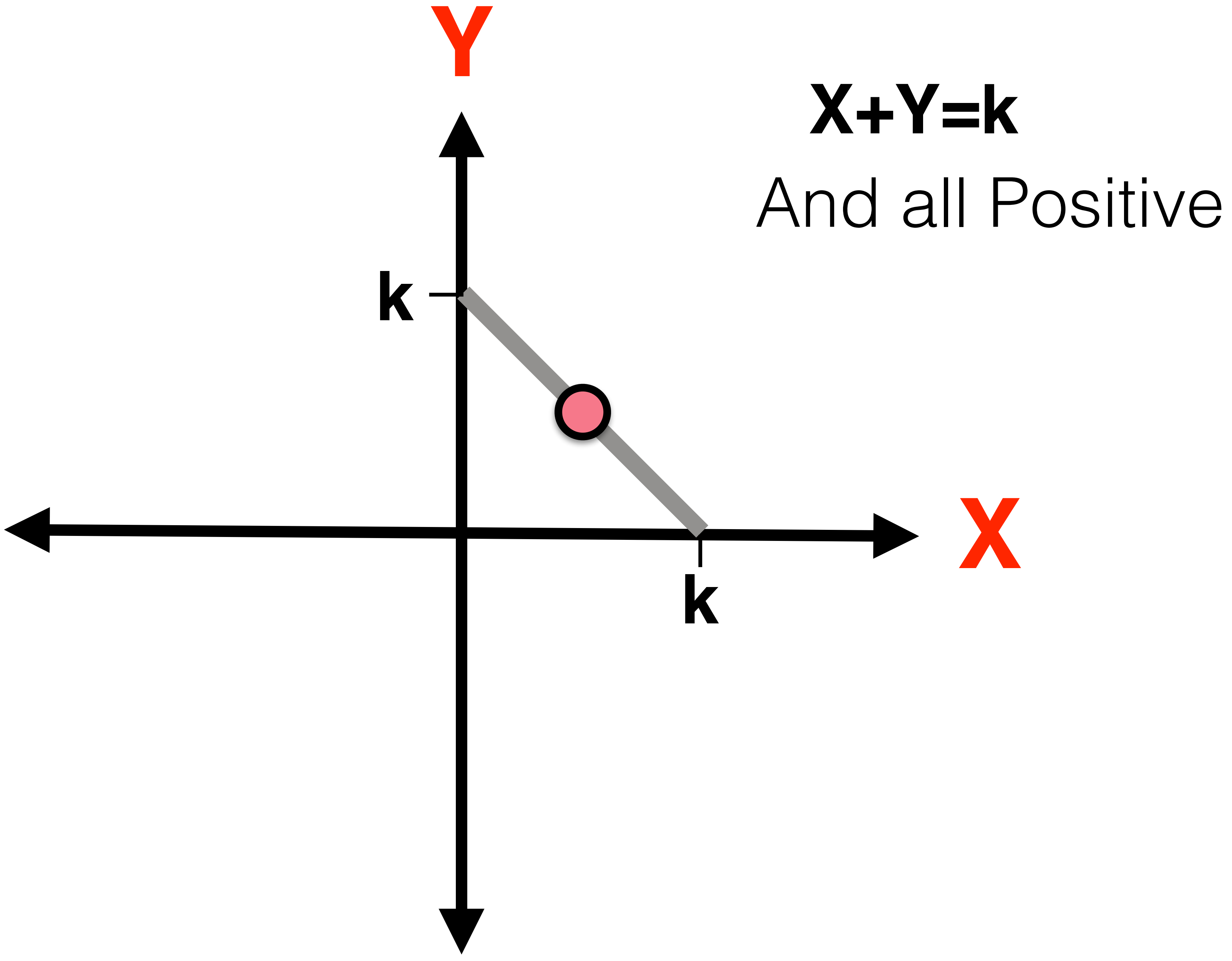


STATISTICAL
COMFORT ZONE



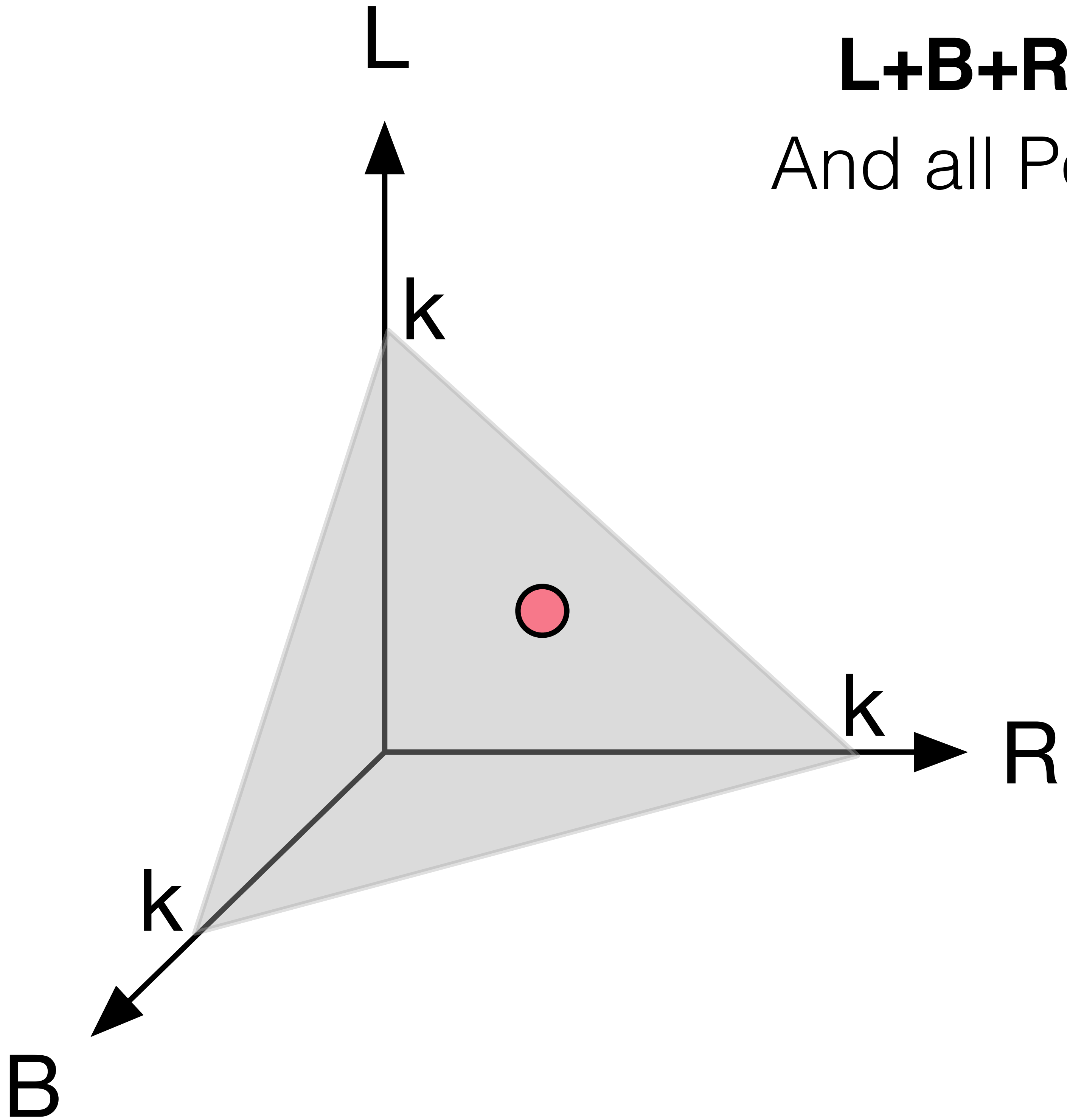
COMPOSITONAL

SIMPLEX



COMPOSITONAL

SIMPLEX

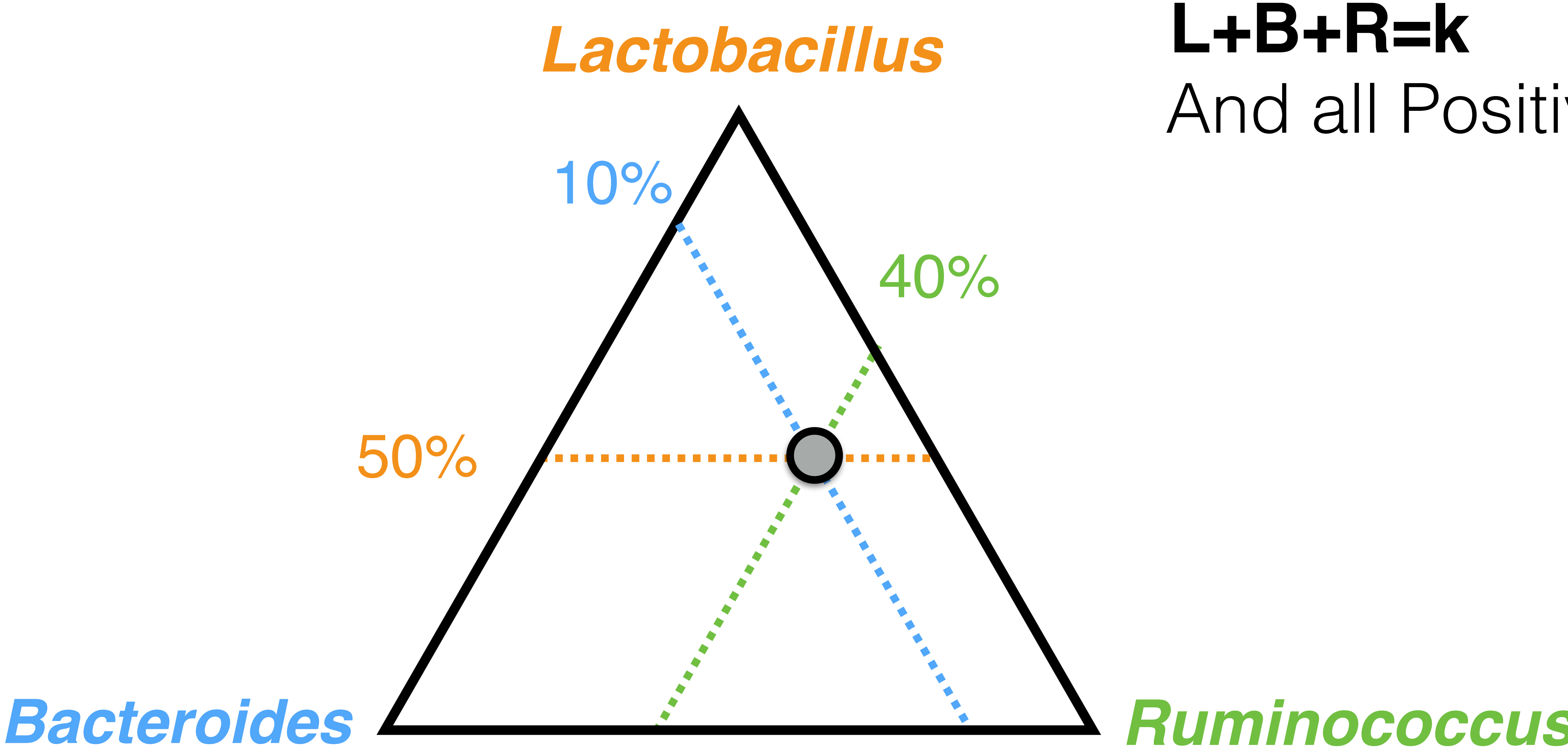


L+B+R=k

And all Positive

MODELING

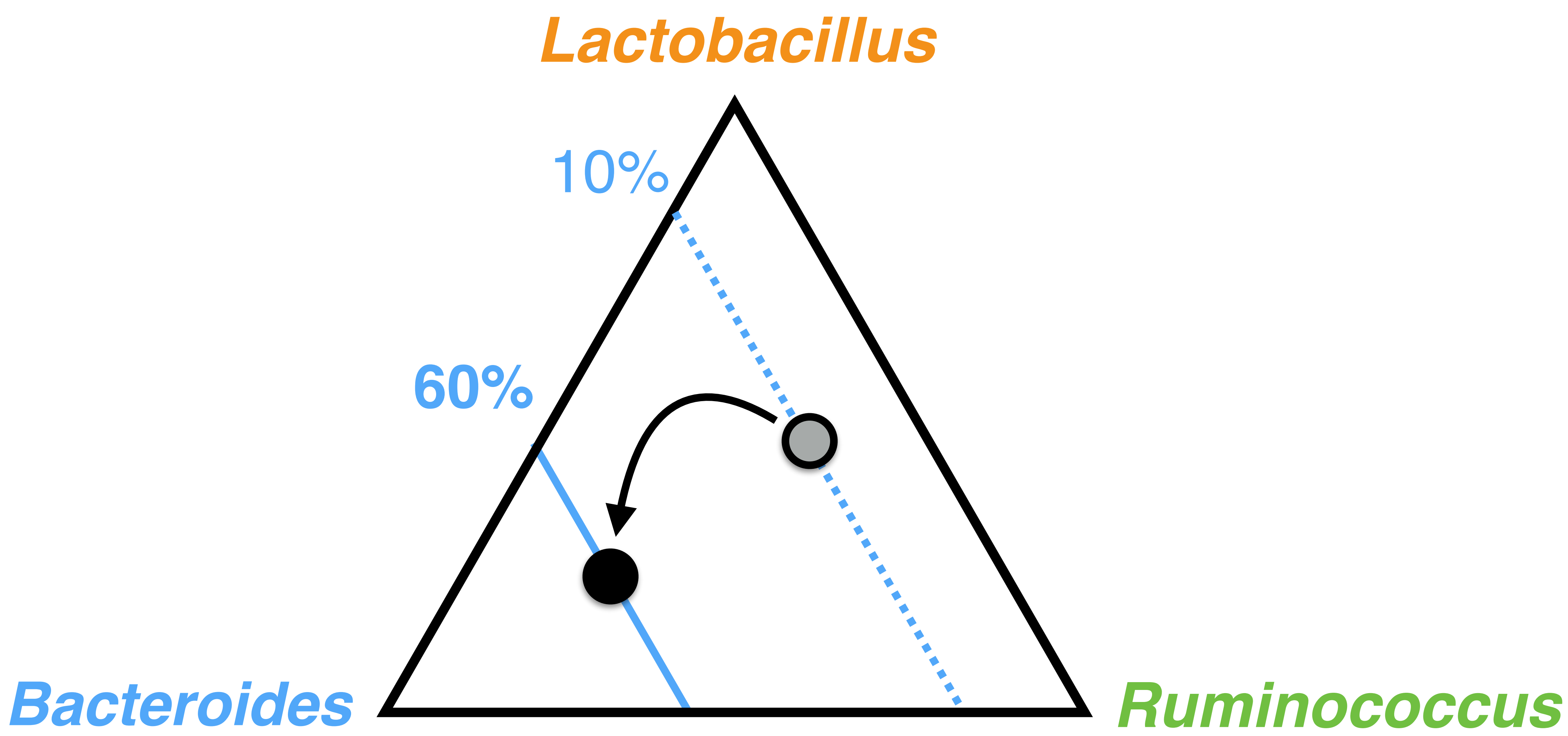
CHALLENGE



L+B+R=k

And all Positive

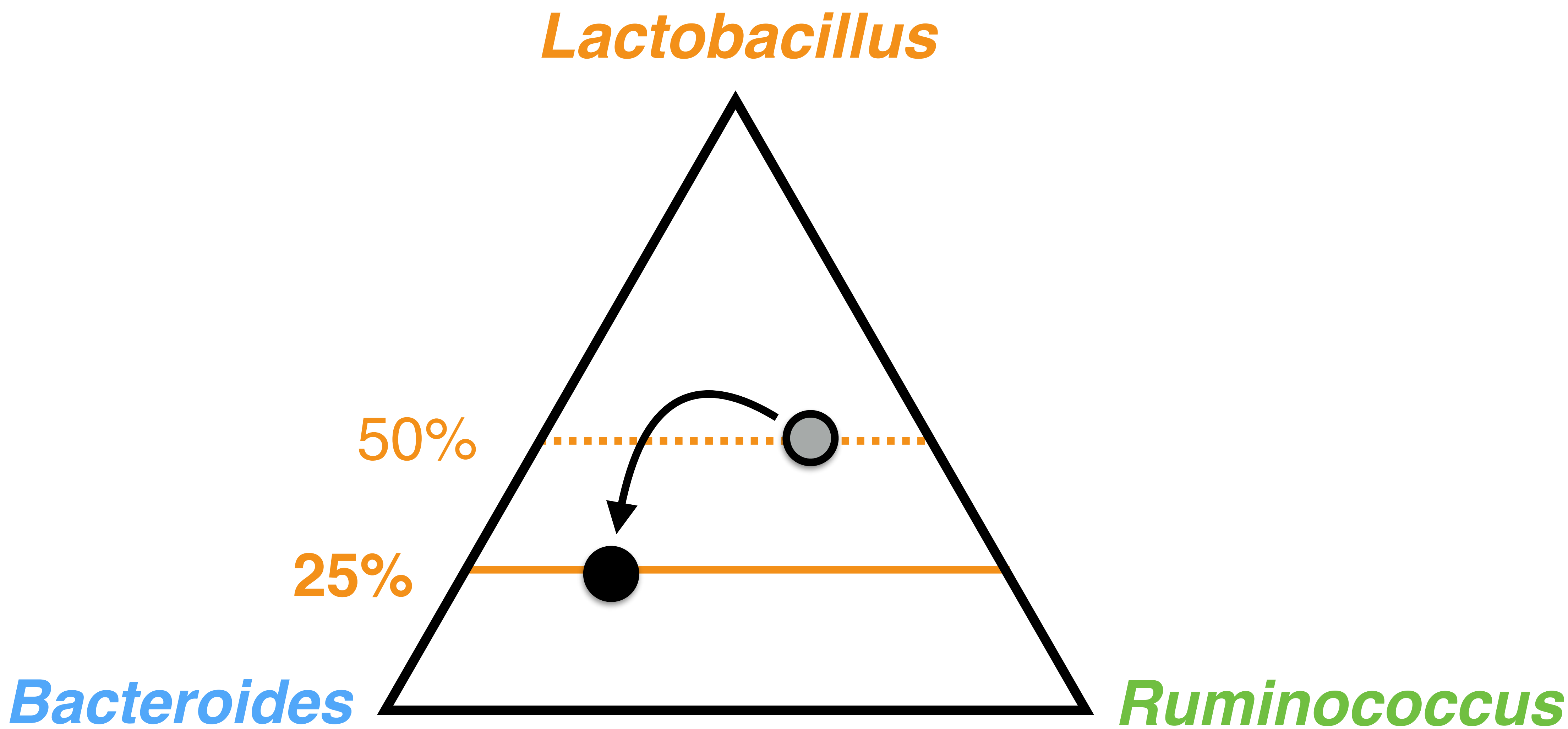
MODELING
CHALLENGE



EVENT: DELIVERY OF ***BACTEROIDES*** PROBIOTIC

MODELING

CHALLENGE

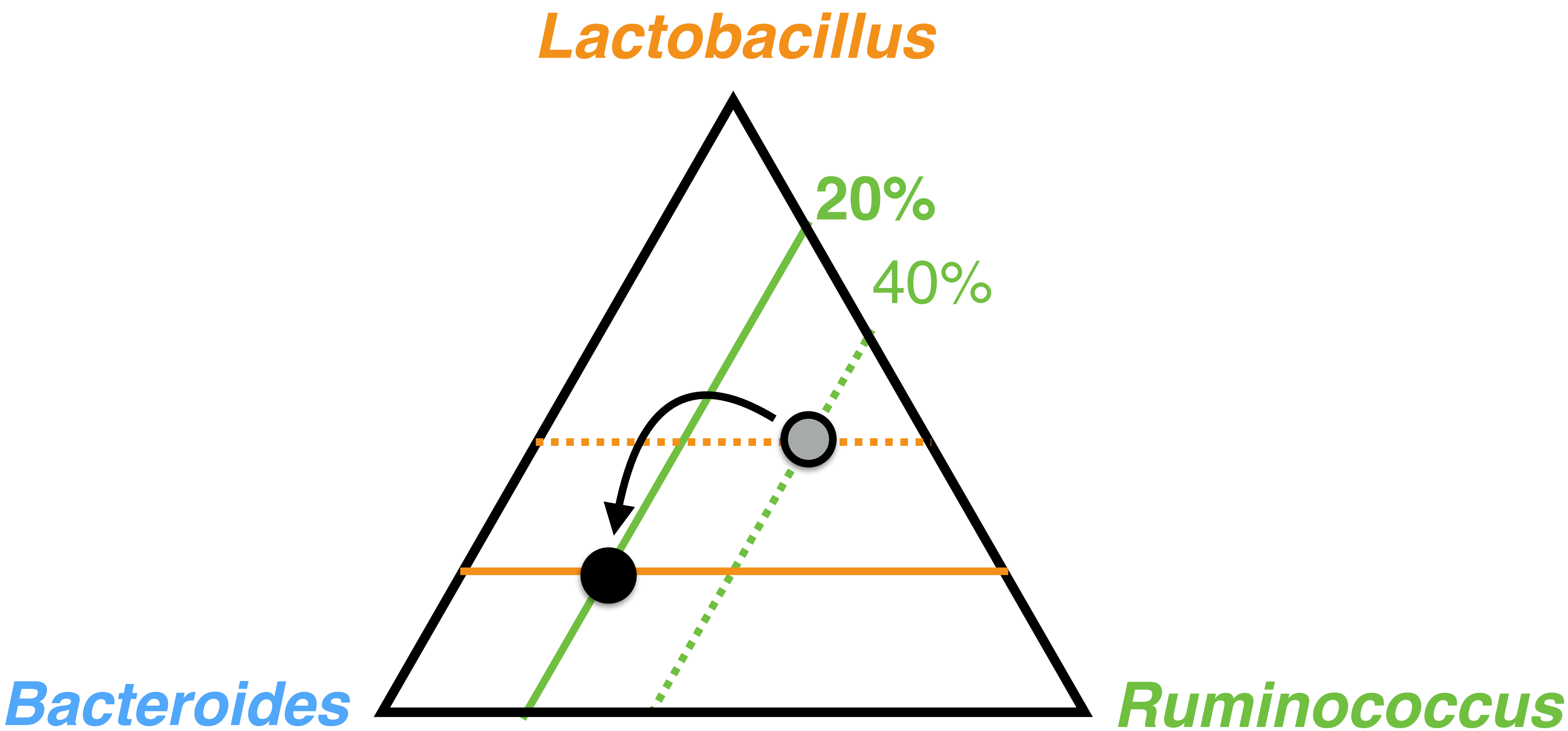


EVENT: DELIVERY OF ***BACTEROIDES*** PROBIOTIC

CHALLENGE: CHANGES IN FRACTIONS OF ***LACTOBACILLUS***

MODELING

CHALLENGE

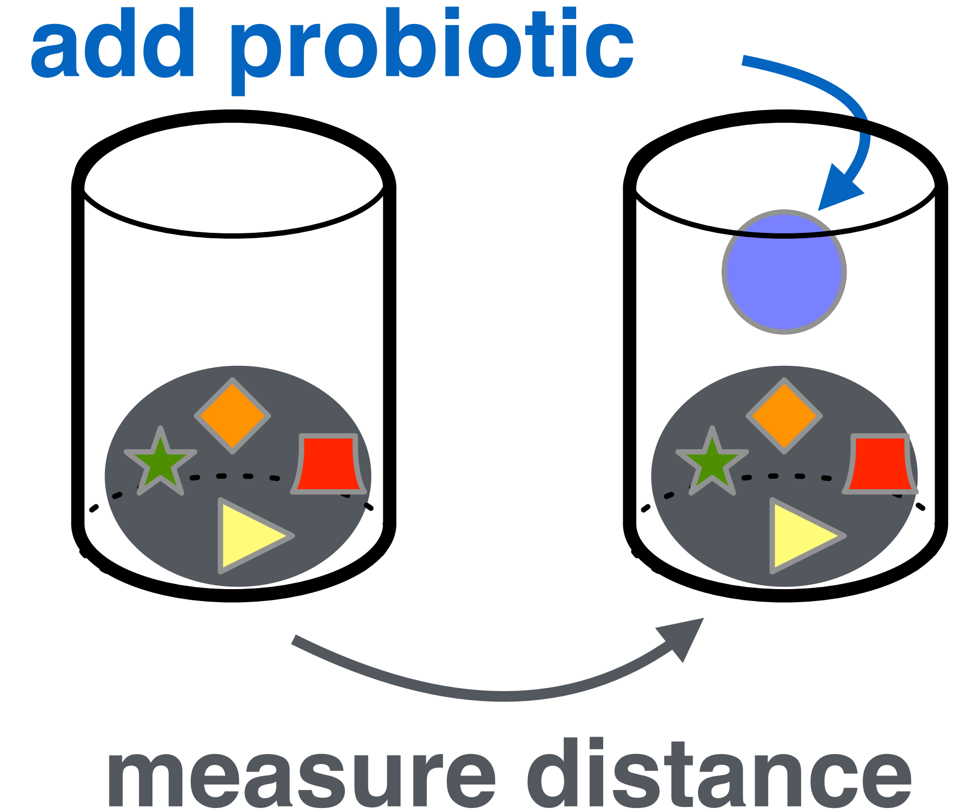


EVENT: DELIVERY OF ***BACTEROIDES*** PROBIOTIC

CHALLENGE: CHANGES IN FRACTIONS OF ***LACTOBACILLUS*** & ***RUMINOCOCCUS***

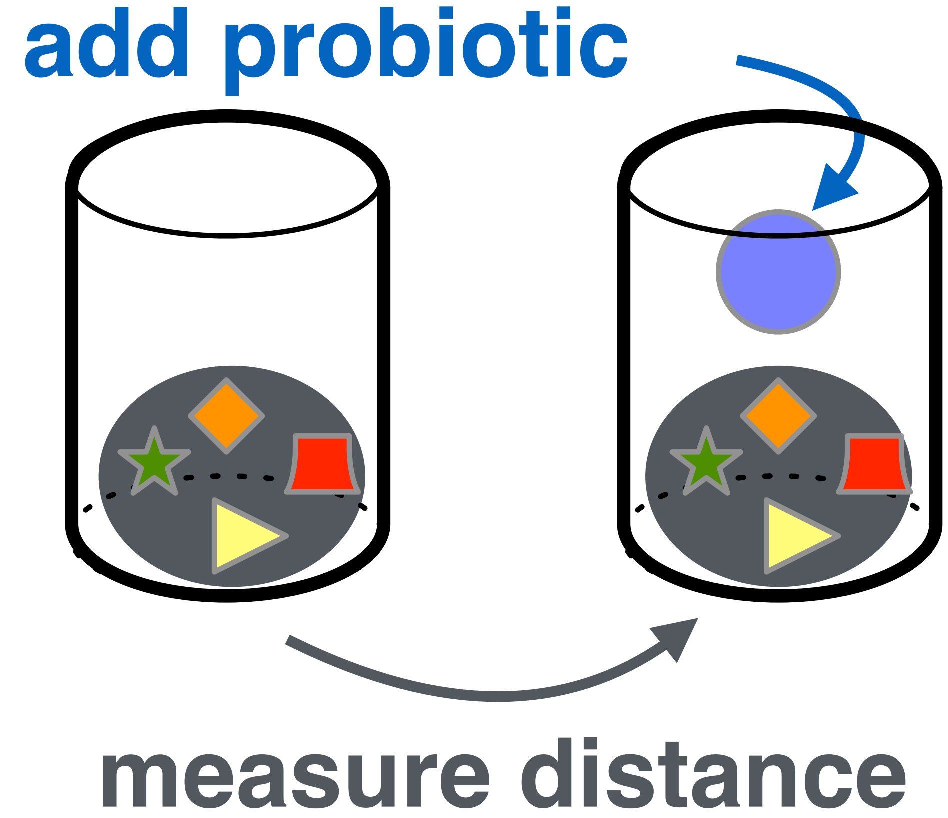
MODELING

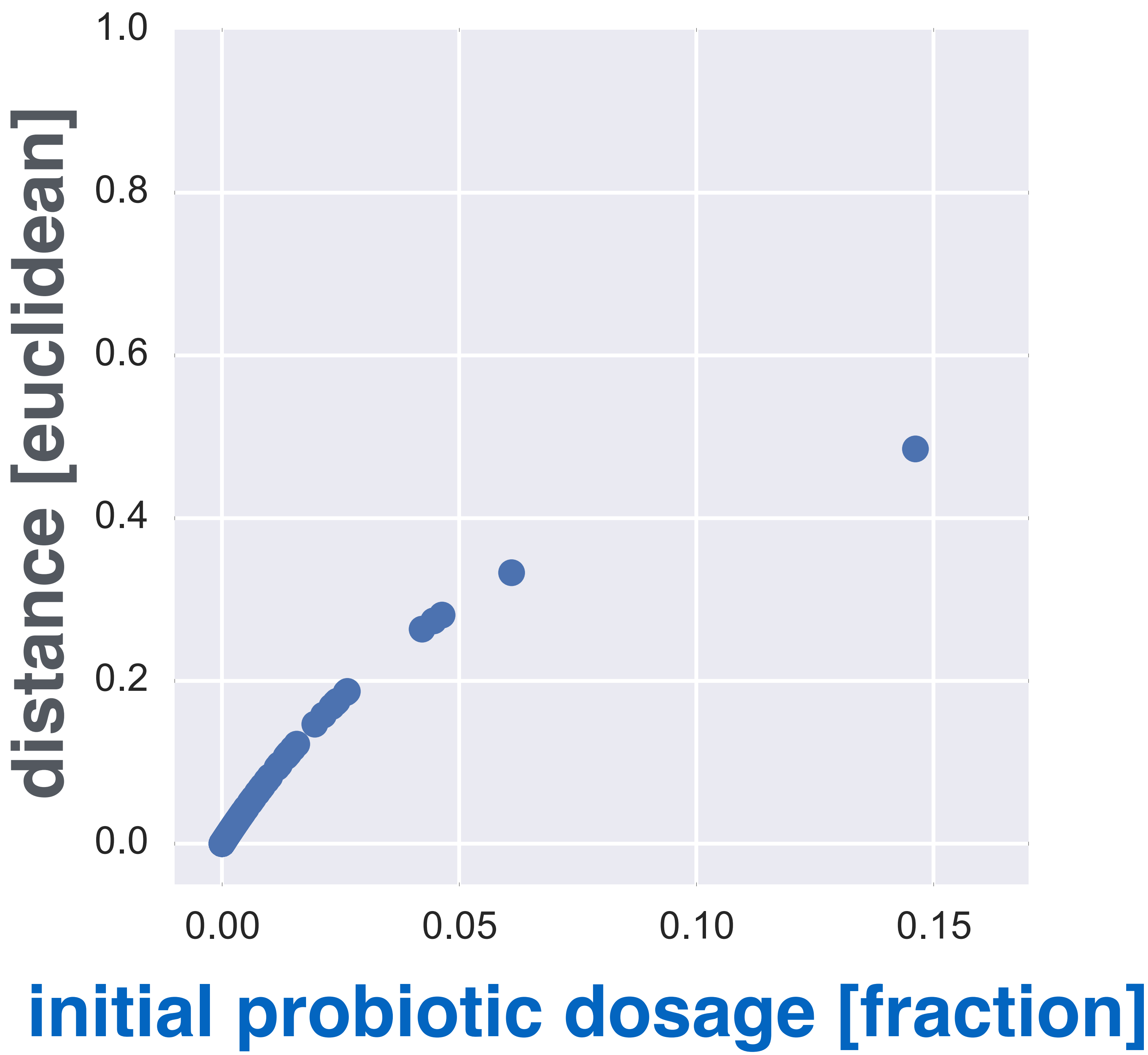
CHALLENGE



MODELING

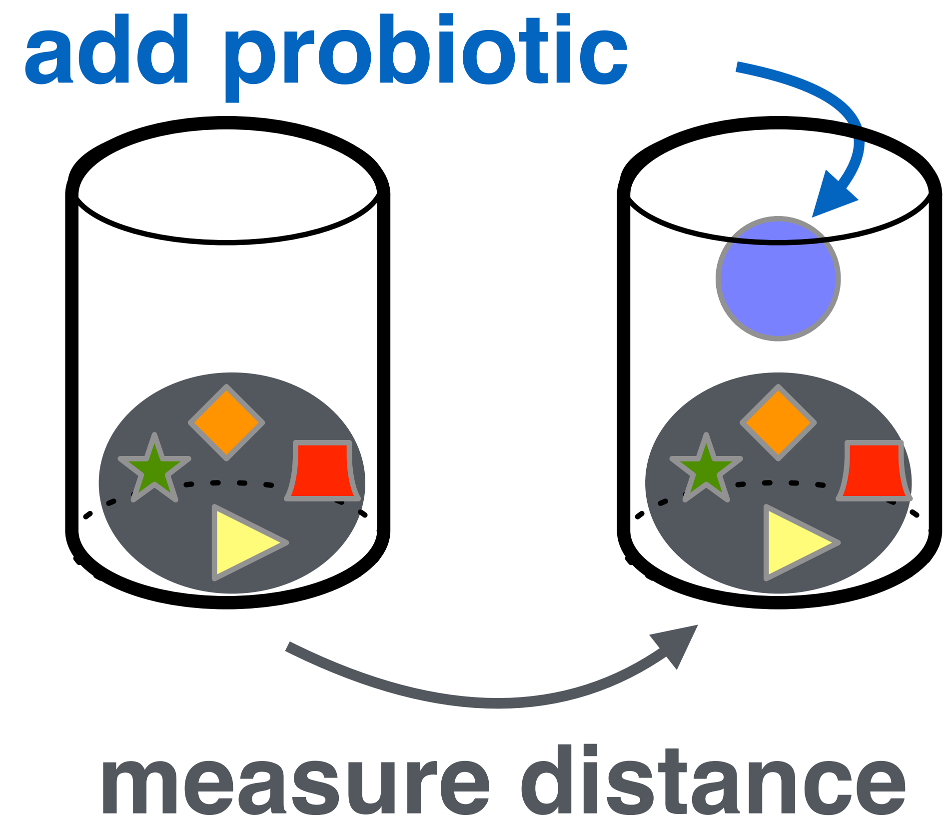
CHALLENGE

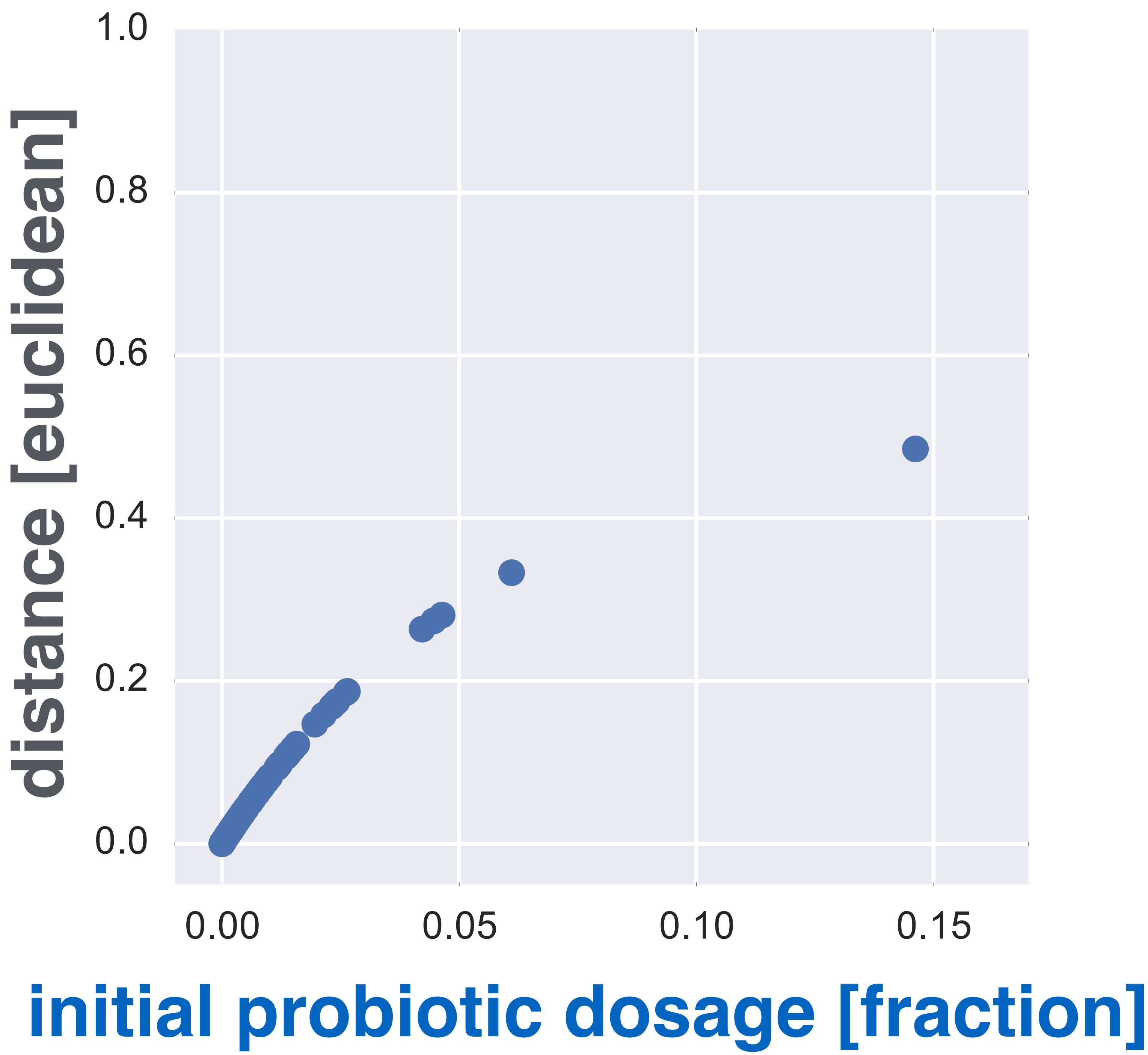




MODELING

CHALLENGE



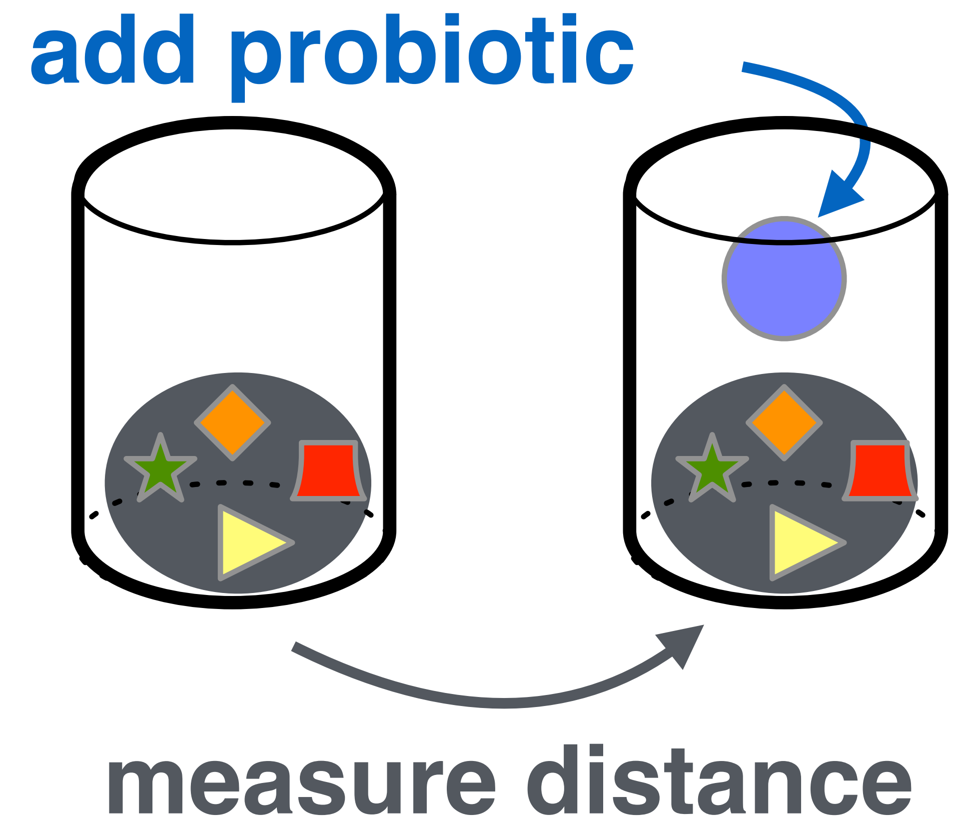


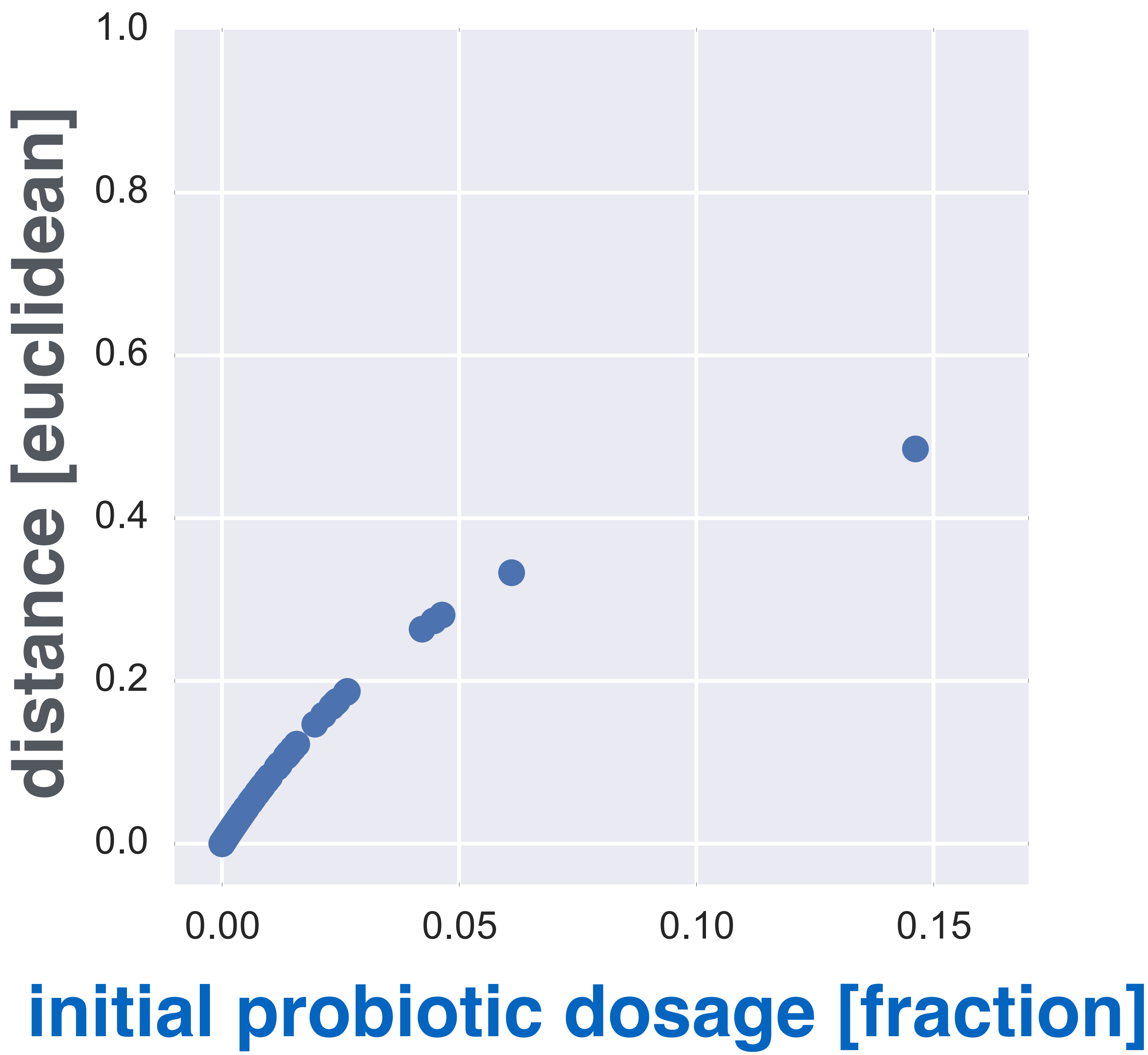
Challenges:

- Probiotic **addition alone** shifts community composition

MODELING

CHALLENGE





Challenges:

- Probiotic **addition alone** shifts community composition
- Shifts are **biased by probiotic dosage**

Compositional

Effects

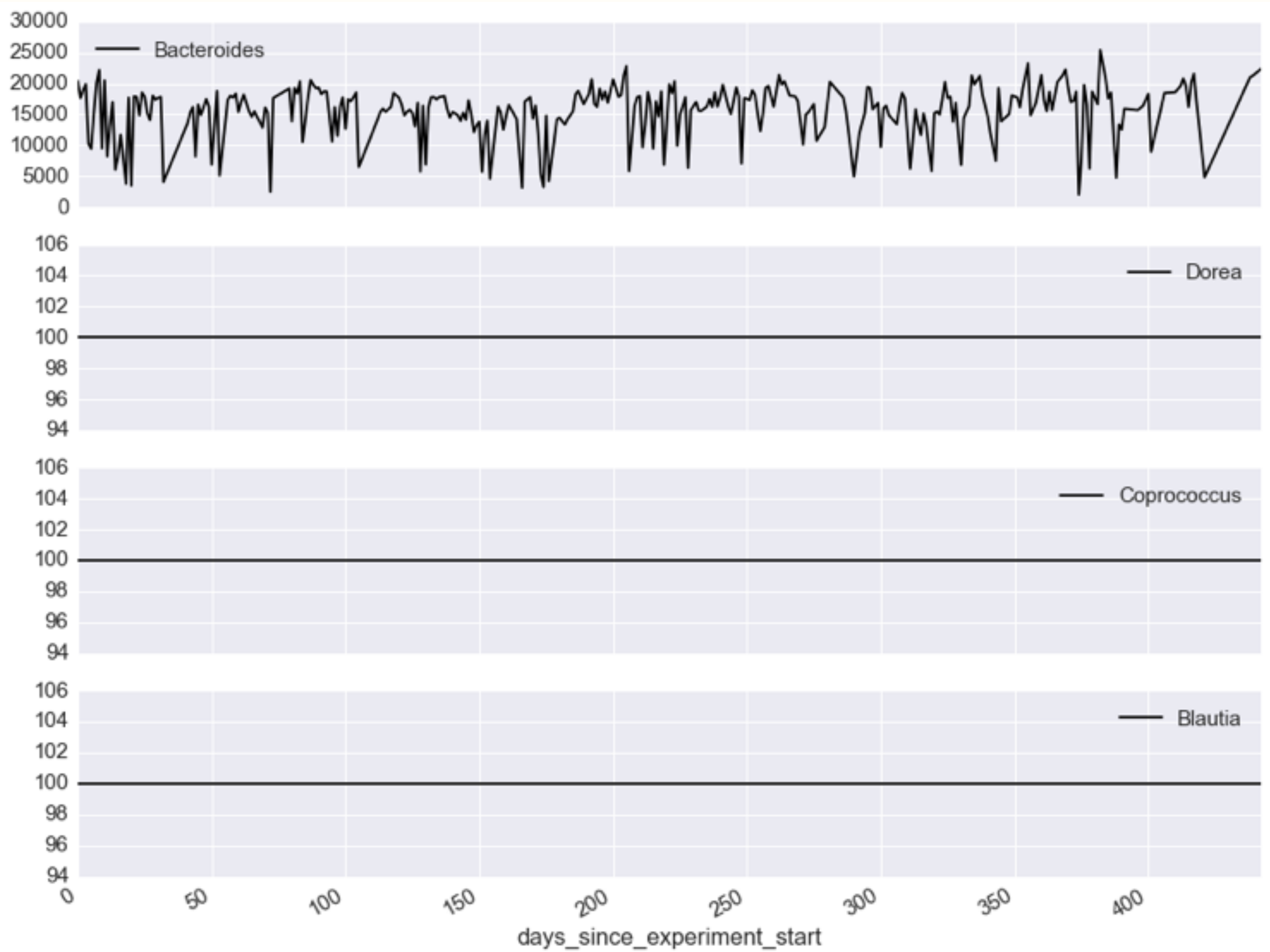
Given: $X + Y + Z = 100\%$

If: X **increases** → Y + Z must **decrease**

Not actually 3 independent variables

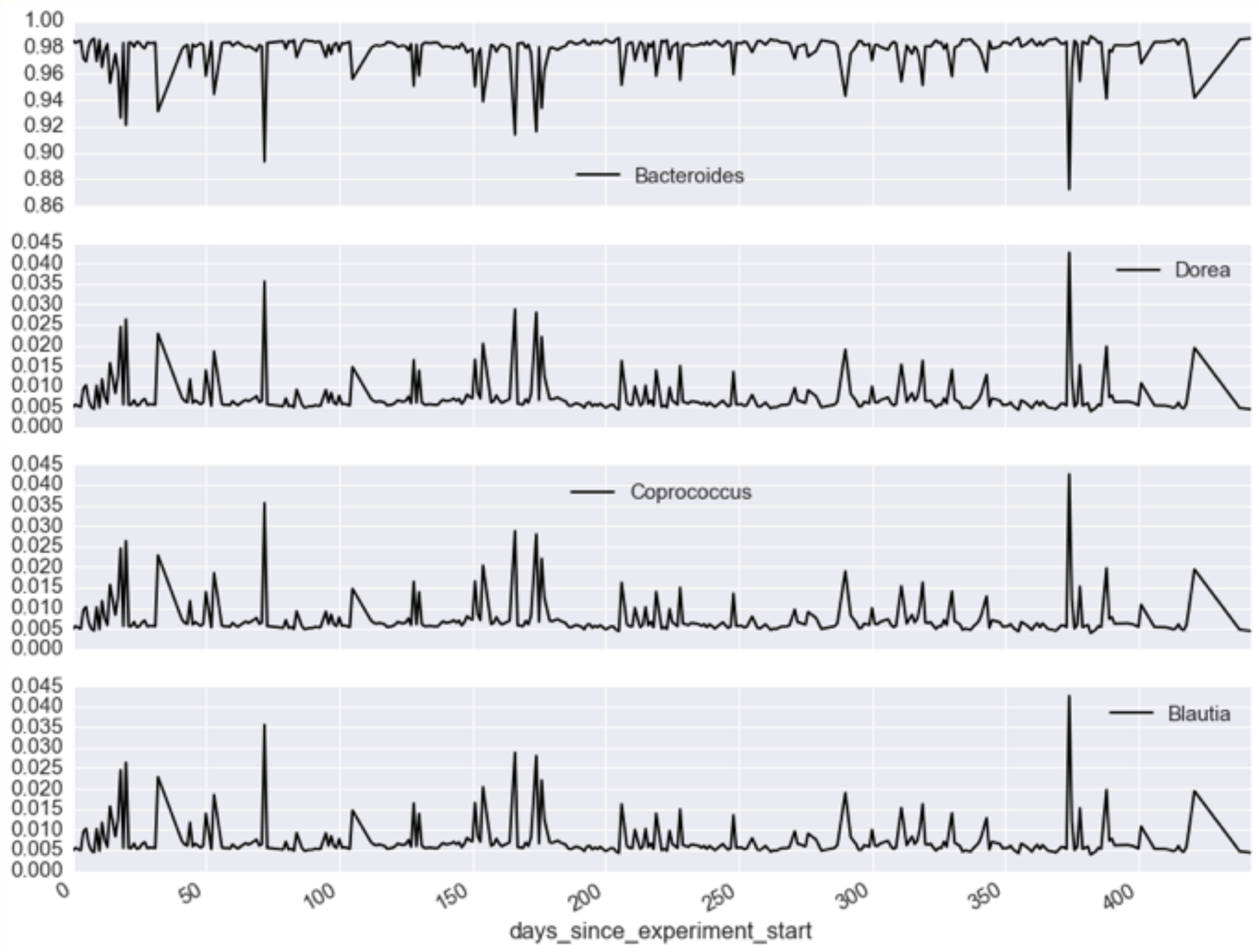
Compositional
Effects

Fake Data



Compositional
Effects

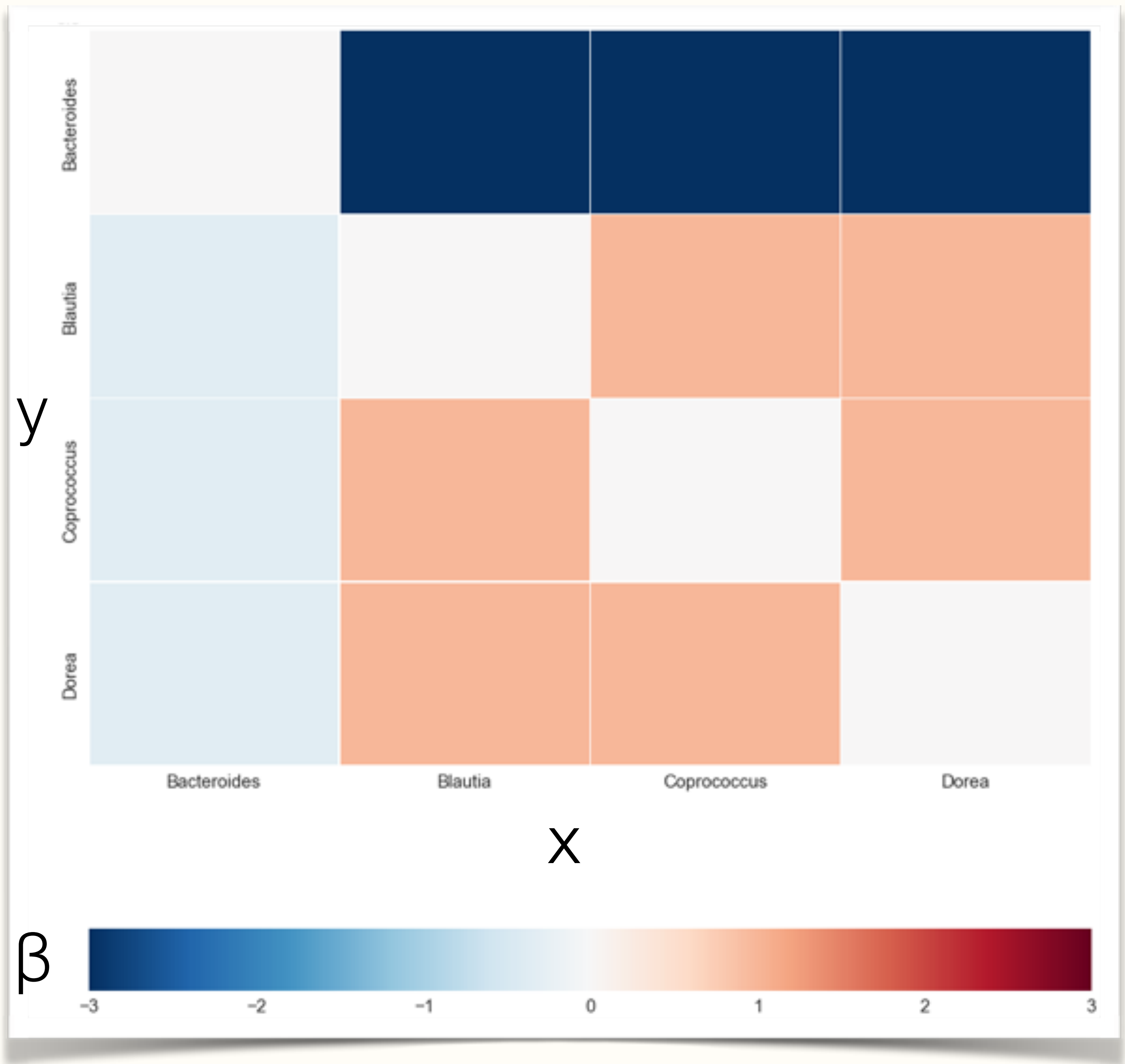
Fake Data (Normalized)



Compositional
Effects

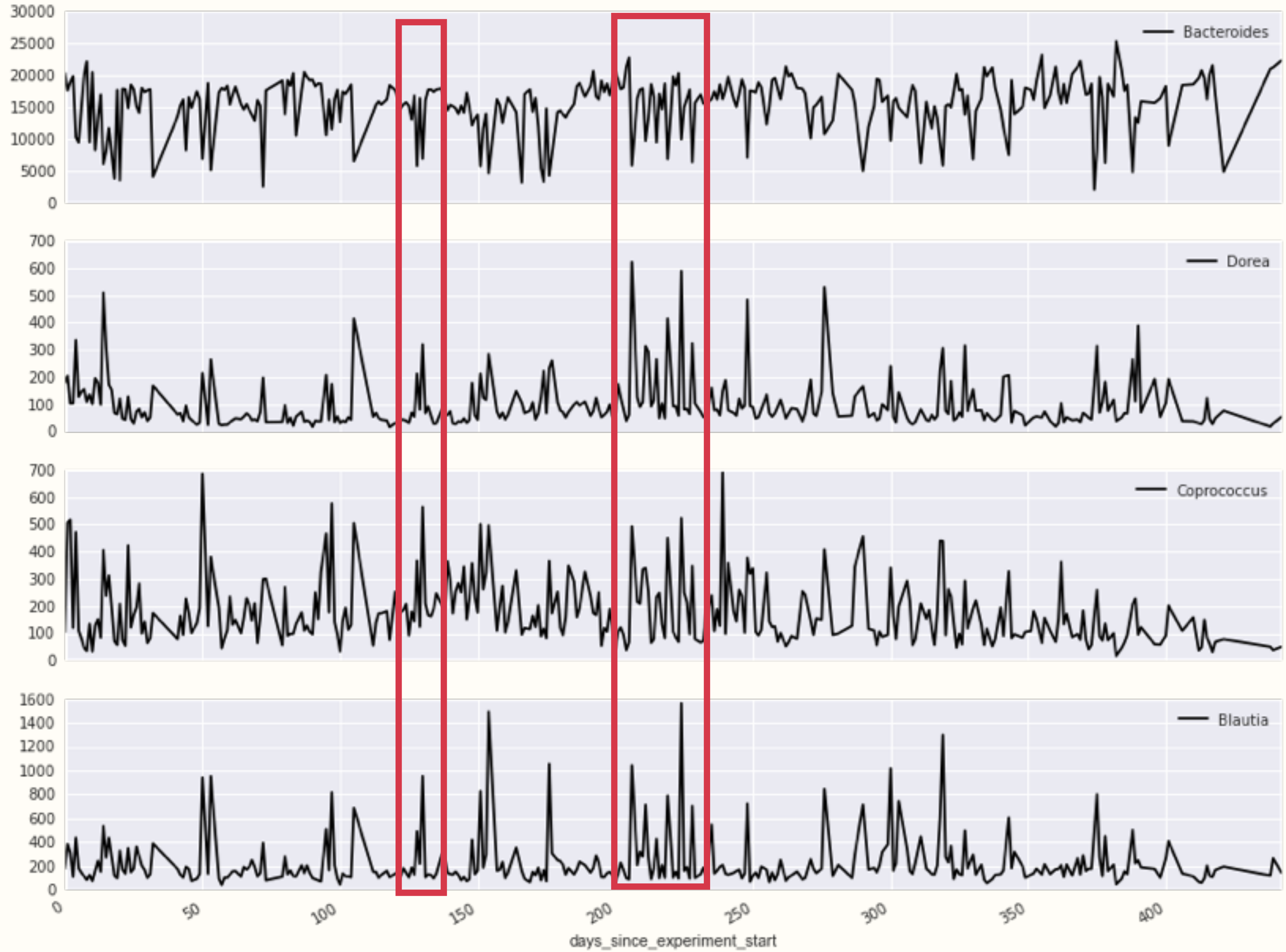
$$y_t = \beta x_t + \alpha$$

Fake (Normalized) Data



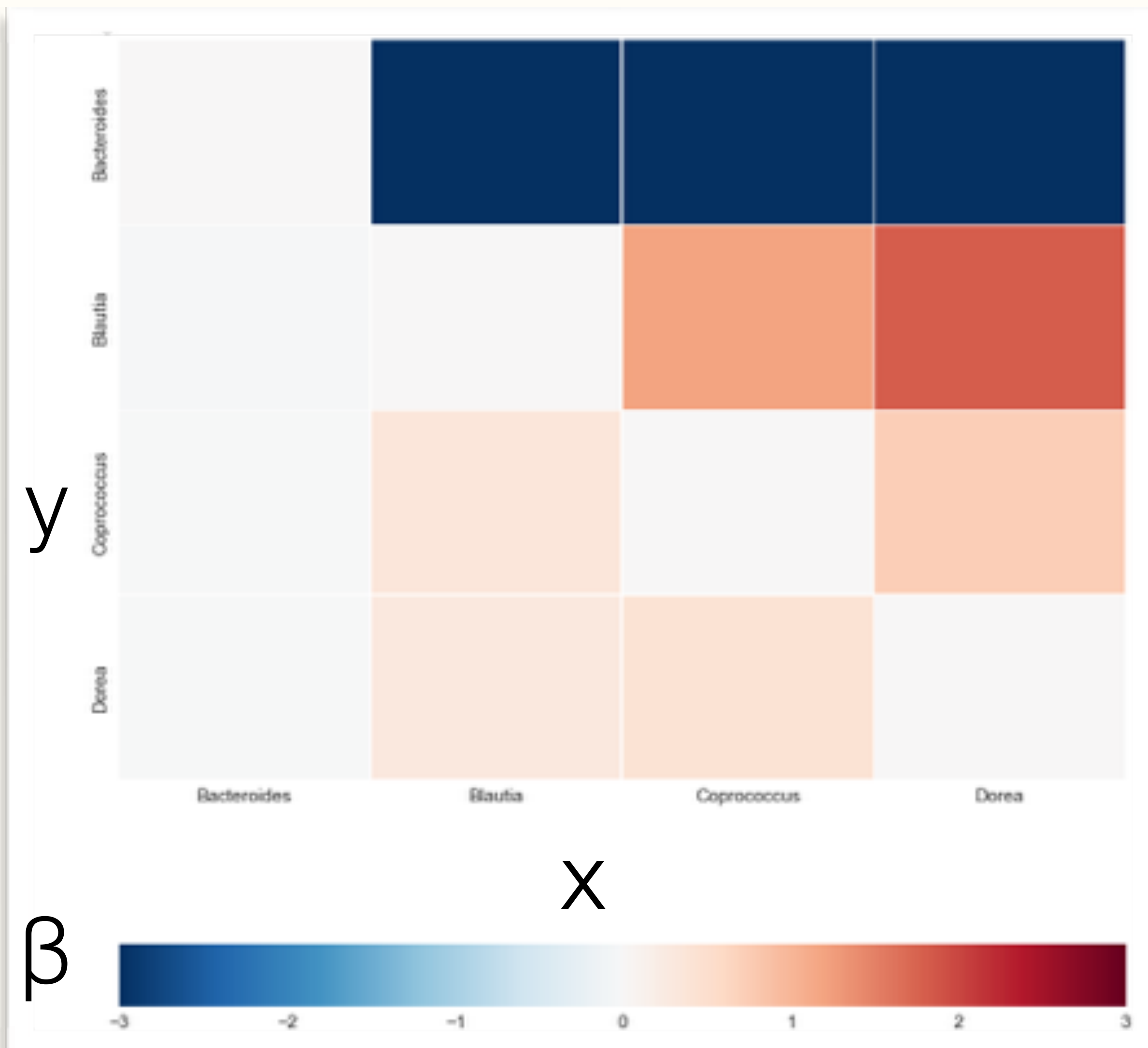
Compositional
Effects

Real Data

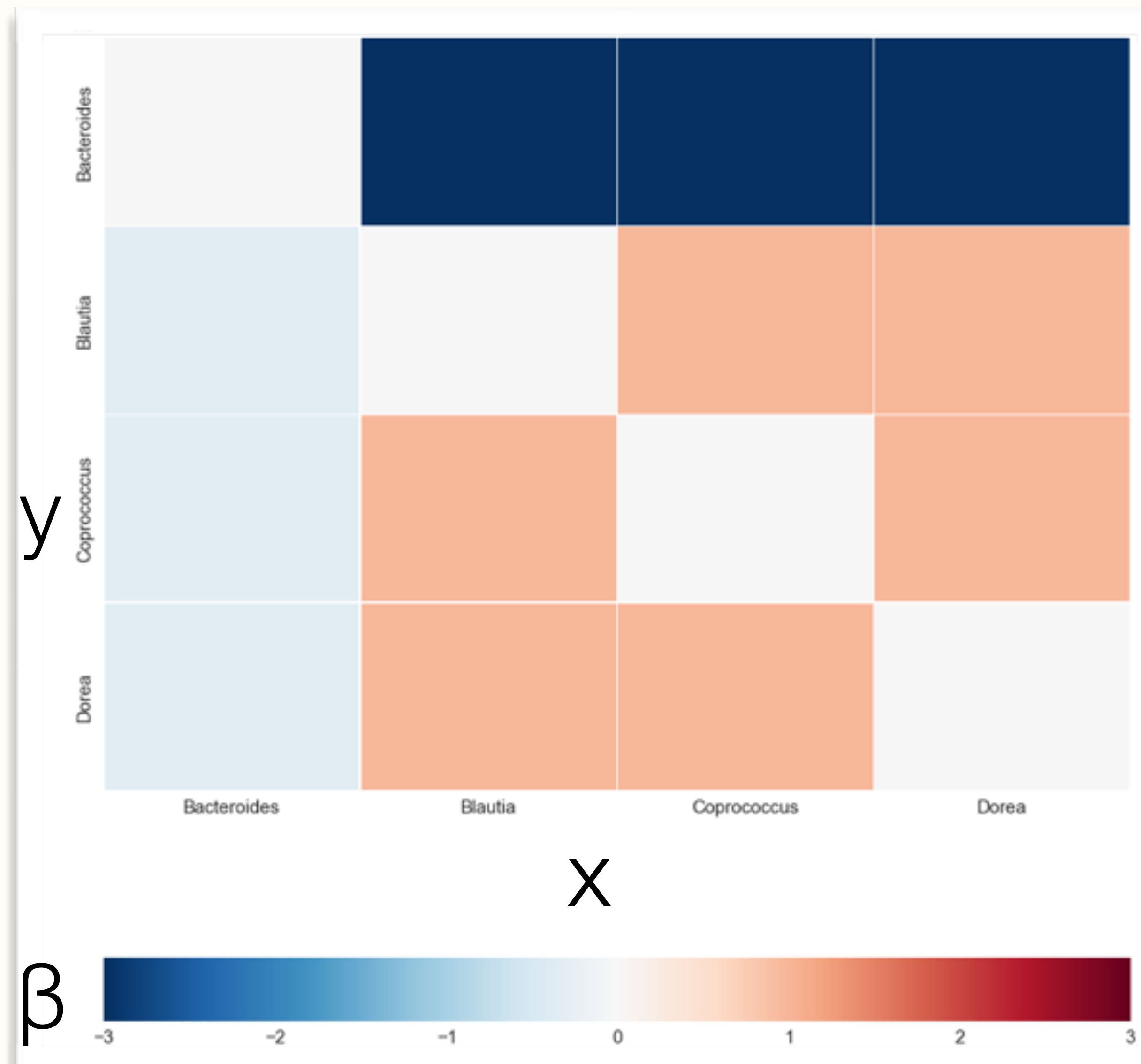


Compositional
Effects

Real Data



Fake Data



the problem: negative bias & spurious correlation

example: scientists A and B record the composition of aliquots of soil samples; A records (animal, vegetable, mineral, water) compositions, B records (animal, vegetable, mineral) after drying the sample; both are absolutely accurate (adapted from Aitchison, 2005)

sample A	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

sample B	x'_1	x'_2	x'_3
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

corr A	x_1	x_2	x_3	x_4
x_1	1.00	**0.50**	0.00	**-0.98**
x_2		1.00	**-0.87**	**-0.65**
x_3			1.00	0.19
x_4				1.00

corr B	x'_1	x'_2	x'_3
x'_1	1.00	**-0.57**	**-0.05**
x'_2		1.00	**-0.79**
x'_3			1.00

requirements for a proper analysis

- **scale invariance:** the analysis should not depend on the closure constant κ; proportional positive vectors are equivalent as compositions
- **permutation invariance:** the order of the parts should be irrelevant
- **subcompositional coherence:** studies performed on subcompositions should not stand in contradiction with those performed on the full composition

why a new geometry on the simplex?

in real space $\mathbb{R}^D$ we **add** vectors, we **multiply** them by a constant, we look for **orthogonality** between vectors, we look for **distances** between points, ...

possible because $\mathbb{R}^D$ is a Euclidean vector space

BUT the usual Euclidean geometry in real space is not a proper geometry for compositional data because

- **results might not be in the simplex** when we **add** compositional vectors, **multiply** them by a constant, or compute **confidence regions**
- **Euclidean differences are not always reasonable:** from 0.05% to 0.10% the amount is doubled; from 50.05% to 50.10% the increase is negligible

basic operations

closure of $\mathbf{z} = [z_1, z_2, \ldots, z_D] \in \mathbb{R}_+^D$, with closure constant $= \kappa$

$$\mathcal{C}[\mathbf{z}] = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^D z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^D z_i}, \ldots, \frac{\kappa \cdot z_D}{\sum_{i=1}^D z_i} \right]$$

$\mathcal{C}[\mathbf{z}]$ is the representative of $\mathbf{z}$ in $\mathcal{S}^D$

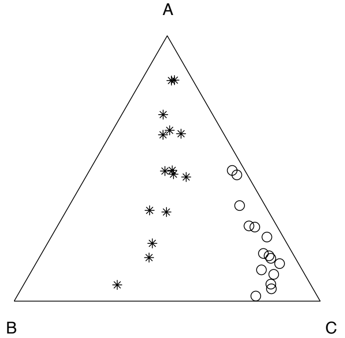
perturbation of $\mathbf{x} \in \mathcal{S}^D$ by $\mathbf{y} \in \mathcal{S}^D$

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_D y_D]$$

powering of $\mathbf{x} \in \mathcal{S}^D$ by $\alpha \in \mathbb{R}$

$$\alpha \odot \mathbf{x} = \mathcal{C}[x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha]$$

interpretation of perturbation and powering



left: perturbation of initial compositions ($\circ$) by $\mathbf{p} = [0.1, 0.1, 0.8]$ resulting in compositions ($\star$)

right: powering of compositions ($\star$) by $\alpha = 0.2$ resulting in compositions ($\circ$)

comments

- **closure = projection** of a point in $\mathbb{R}_+^D$ on $\mathcal{S}^D$
- points on a ray are projected onto the same point
  - a ray in $\mathbb{R}_+^D$ is an equivalence class
  - the point on $\mathcal{S}^D$ is a representative of the class
  - a generalization to other representatives is possible
- for $\mathbf{z} \in \mathbb{R}_+^D$ and $\mathbf{x} \in \mathcal{S}^D$, $\mathbf{x} \oplus (\alpha \odot \mathbf{z}) = \mathbf{x} \oplus (\alpha \odot \mathcal{C}[\mathbf{z}])$

vector space structure of $(\mathcal{S}^D, \oplus, \odot)$

- **commutative group structure** of $(\mathcal{S}^D, \oplus)$
  1. commutativity: $\mathbf{x} \oplus \mathbf{y} = \mathbf{y} \oplus \mathbf{x}$
  2. associativity: $(\mathbf{x} \oplus \mathbf{y}) \oplus \mathbf{z} = \mathbf{x} \oplus (\mathbf{y} \oplus \mathbf{z})$
  3. neutral element: $\mathbf{e} = \mathcal{C}\,[1, 1, \ldots, 1]$ = barycentre of $\mathcal{S}^D$
  4. inverse of $\mathbf{x}$: $\mathbf{x}^{-1} = \mathcal{C}\left[x_1^{-1}, x_2^{-1}, \ldots, x_D^{-1}\right]$

     $\Rightarrow \quad \mathbf{x} \oplus \mathbf{x}^{-1} = \mathbf{e}$ and $\mathbf{x} \oplus \mathbf{y}^{-1} = \mathbf{x} \ominus \mathbf{y}$

- **properties of powering**
  1. associativity: $\alpha \odot (\beta \odot \mathbf{x}) = (\alpha \cdot \beta) \odot \mathbf{x}$;
  2. distributivity 1: $\alpha \odot (\mathbf{x} \oplus \mathbf{y}) = (\alpha \odot \mathbf{x}) \oplus (\alpha \odot \mathbf{y})$
  3. distributivity 2: $(\alpha + \beta) \odot \mathbf{x} = (\alpha \odot \mathbf{x}) \oplus (\beta \odot \mathbf{x})$
  4. neutral element: $1 \odot \mathbf{x} = \mathbf{x}$

complete inner product space structure

inner product : $\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$

norm : $\|\mathbf{x}\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} \right)^2}, \quad \mathbf{x} \in \mathcal{S}^D$

distance : $d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}, \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}^D$

Aitchison geometry on the simplex

$(\mathcal{S}^D, \oplus, \odot)$ **is a** $(D-1)$**-dim. Euclidean space**

properties of the Aitchison geometry

distance and perturbation: $d_a(\mathbf{p} \oplus \mathbf{x}, \mathbf{p} \oplus \mathbf{y}) = d_a(\mathbf{x}, \mathbf{y})$

distance and powering: $d_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{y}) = |\alpha| d_a(\mathbf{x}, \mathbf{y})$

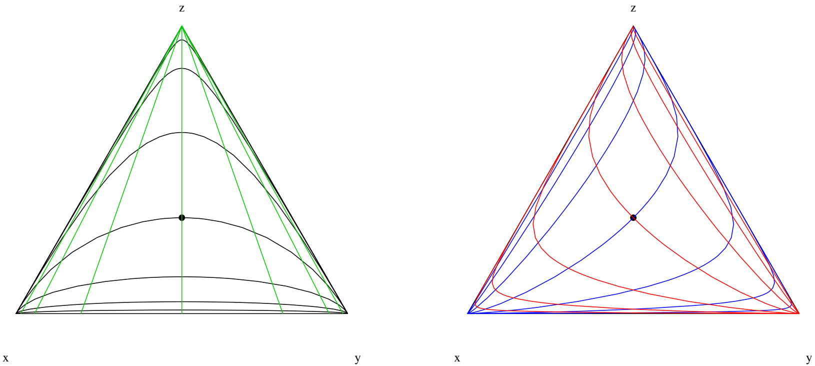
compositional lines: $\mathbf{y} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$
($\mathbf{x}_0$ = starting point, $\mathbf{x}$ = leading vector)

orthogonal lines: $\mathbf{y_1} = \mathbf{x_0} \oplus (\alpha_1 \odot \mathbf{x_1})$, $\mathbf{y_2} = \mathbf{x_0} \oplus (\alpha_2 \odot \mathbf{x_2})$,

$$\mathbf{y_1} \perp \mathbf{y_2} \iff \langle \mathbf{x_1}, \mathbf{x_2} \rangle_a = 0$$

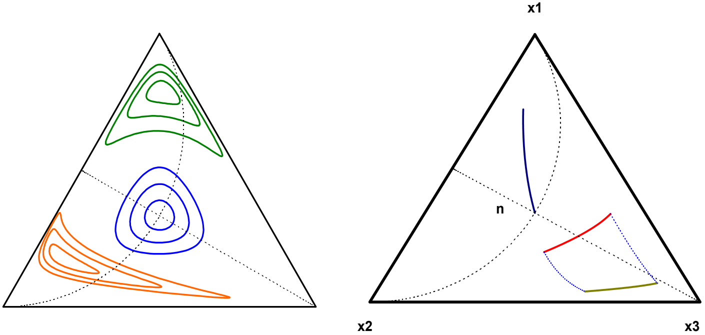
(the inner product of the leading vectors is zero)
parallel lines: $\mathbf{y_1} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x}) \quad \| \quad \mathbf{y_2} = \mathbf{p} \oplus \mathbf{x}_0 \oplus (\alpha \odot \mathbf{x})$

orthogonal compositional lines



orthogonal grids in $\mathcal{S}^3$, equally spaced, 1 unit in Aitchison distance; the right grid is rotated 45^o with respect to the left grid

ellipses and shifted segments



advantages of complete inner product spaces

- **orthonormal basis** can be constructed: $\{\mathbf{e}_1, \ldots, \mathbf{e}_{D-1}\}$
- **coordinates obey the rules** of real Euclidean space:

  $\mathbf{x} \in \mathcal{S}^D \Rightarrow \mathbf{y} = [y_1, \ldots, y_{D-1}] \in \mathbb{R}^{D-1}$, with $y_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a$
- **standard methods** can be directly applied to coordinates
- **expressing results as compositions is easy**:

  if $h : \mathcal{S}^D \mapsto \mathbb{R}^{D-1}$ assigns to each $\mathbf{x} \in \mathcal{S}^D$ its coordinates, i.e. $h(\mathbf{x}) = \mathbf{y}$, then

$$h^{-1}(\mathbf{y}) = \mathbf{x} = \bigoplus_{i=1}^{D-1} y_i \odot \mathbf{e}_i$$

PRINCIPLE OF WORKING ON COORDINATES

conclusions

- the Aitchison geometry of the simplex offers a powerful tool to analyse CoDa
- the geometry is apparently complex, but it is completely equivalent to standard Euclidean geometry in real space
- the **key** is to use a **proper representation in coordinates**

some specific references

Aitchison, J.: The statistical analysis of compositional data. Monographs on statistics and applied Probability: Chapman and Hall, London (Reprinted in 2003 with additional material by Blackburn Press), 1986.

Pawlowsky-Glahn, V. and J. J. Egozcue: Geometric Approach to Statistical Analysis on the Simplex, Stochastic Environmental Research and Risk Assessment, 15, 5, 384-398, 2001.

Aitchison, J. and J. J. Egozcue: Compositional data analysis: where are we and where should we be heading?, Mathematical Geology, 37, 7, 833-854, 2005.

Aitchison, J., Barceló-Vidal, C. , Egozcue, J. J., and Pawlowsky-Glahn, V.: A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. Proceedings of IAMG'02, Berlin, 2002.

Egozcue, J. J. and V. Pawlowsky-Glahn: Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V., (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, (ISBN: 1-86239-205-6), Special Publications, 264, 67-77, 2006.

Egozcue, J. J.: Reply to "On the Harker variation diagrams;..." by J. A. Cortés.

Mathematical Geosciences, 41, 7, 829-834, 2009.