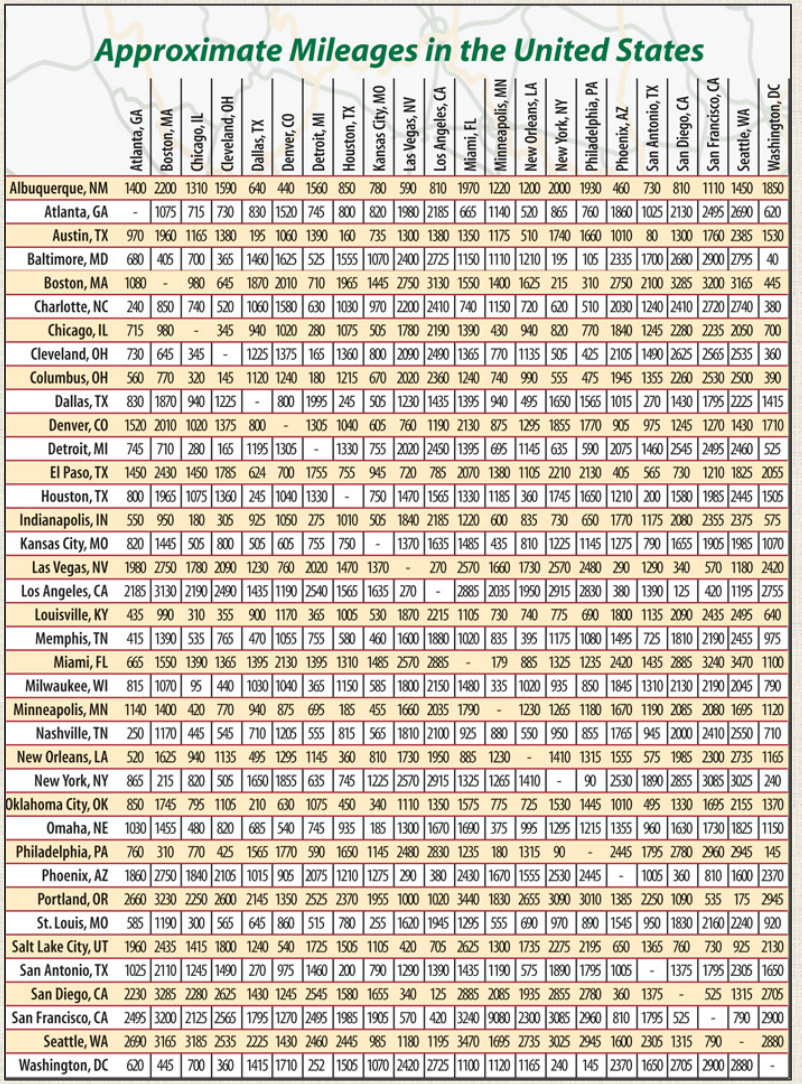
Module 1

Quantifying Metagenomic Diversity

Finding Lost Ancient Cities

Say that you know all the distances between a collection of cities, but you don't know where they all are. If you know a few of the cities' locations, can you find the others?

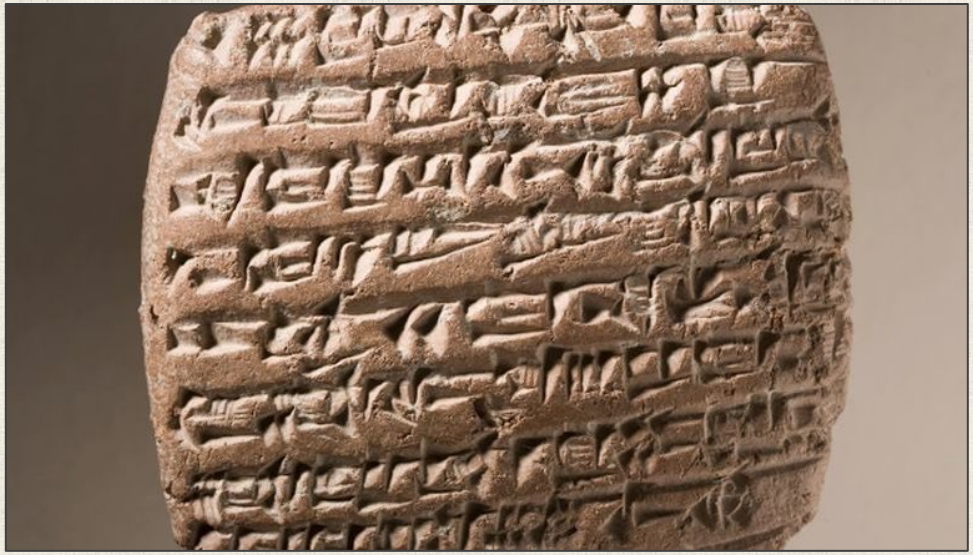


Approximate Mileages in the United States

	Atlanta, GA	Boston, MA	Chicago, IL	Cleveland, OH	Dallas, TX	Denver, CO	Detroit, MI	Houston, TX	Kansas City, MO	Las Vegas, NV	Los Angeles, CA	Miami, FL	Minneapolis, MN	New Orleans, LA	New York, NY	Philadelphia, PA	Phoenix, AZ	San Antonio, TX	San Diego, CA	San Francisco, CA	Seattle, WA	Washington, DC
Albuquerque, NM	1400	2200	1310	1590	640	440	1560	850	780	590	810	1970	1220	1200	2000	1930	460	730	810	1110	1450	1850
Atlanta, GA	-	1075	715	730	830	1520	745	800	820	1980	2185	665	1140	520	865	760	1860	1025	2130	2495	2690	620
Austin, TX	970	1960	1165	1380	195	1060	1390	160	735	1300	1380	1350	1175	510	1740	1660	1010	80	1300	1760	2385	1530
Baltimore, MD	680	405	700	365	1460	1625	525	1555	1070	2400	2725	1150	1110	1210	195	105	2335	1700	2680	2900	2795	40
Boston, MA	1080	-	980	645	1870	2010	710	1965	1445	2750	3130	1550	1400	1625	215	310	2750	2100	3285	3200	3165	445
Charlotte, NC	240	850	740	520	1060	1580	630	1030	970	2200	2410	740	1150	720	620	510	2030	1240	2410	2720	2740	380
Chicago, IL	715	980	-	345	940	1020	280	1075	505	1780	2190	1390	430	940	820	770	1840	1245	2280	2235	2050	700
Cleveland, OH	730	645	345	-	1225	1375	165	1360	800	2090	2490	1365	770	1135	505	425	2105	1490	2625	2565	2535	360
Columbus, OH	560	770	320	145	1120	1240	180	1215	670	2020	2360	1240	740	990	555	475	1945	1355	2260	2530	2500	390
Dallas, TX	830	1870	940	1225	-	800	1995	245	505	1230	1435	1395	940	495	1650	1565	1015	270	1430	1795	2225	1415
Denver, CO	1520	2010	1020	1375	800	-	1305	1040	605	760	1190	2130	875	1295	1855	1770	905	975	1245	1270	1430	1710
Detroit, MI	745	710	280	165	1195	1305	-	1330	755	2020	2450	1395	695	1145	635	590	2075	1460	2545	2495	2460	525
El Paso, TX	1450	2430	1450	1785	624	700	1755	755	945	720	785	2070	1380	1105	2210	2130	405	565	730	1210	1825	2055
Houston, TX	800	1965	1075	1360	245	1040	1330	-	750	1470	1565	1330	1185	360	1745	1650	1210	200	1580	1985	2445	1505
Indianapolis, IN	550	950	180	305	925	1050	275	1010	505	1840	2185	1220	600	835	730	650	1770	1175	2080	2355	2375	575
Kansas City, MO	820	1445	505	800	505	605	755	750	-	1370	1635	1485	435	810	1225	1145	1275	790	1655	1905	1985	1070
Las Vegas, NV	1980	2750	1780	2090	1230	760	2020	1470	1370	-	270	2570	1660	1730	2570	2480	290	1290	340	570	1180	2420
Los Angeles, CA	2185	3130	2190	2490	1435	1190	2540	1565	1635	270	-	2885	2035	1950	2915	2830	380	1390	125	420	1195	2755
Louisville, KY	435	990	310	355	900	1170	365	1005	530	1870	2215	1105	730	740	775	690	1800	1135	2090	2435	2495	640
Memphis, TN	415	1390	535	765	470	1055	755	580	460	1600	1880	1020	835	395	1175	1080	1495	725	1810	2190	2455	975
Miami, FL	665	1550	1390	1365	1395	2130	1395	1310	1485	2570	2885	-	179	885	1325	1235	2420	1435	2885	3240	3470	1100
Milwaukee, WI	815	1070	95	440	1030	1040	365	1150	585	1800	2150	1480	335	1020	935	850	1845	1310	2130	2190	2045	790
Minneapolis, MN	1140	1400	420	770	940	875	695	185	455	1660	2035	1790	-	1230	1265	1180	1670	1190	2085	2080	1695	1120
Nashville, TN	250	1170	445	545	710	1205	555	815	565	1810	2100	925	880	550	950	855	1765	945	2000	2410	2550	710
New Orleans, LA	520	1625	940	1135	495	1295	1145	360	810	1730	1950	885	1230	-	1410	1315	1555	575	1985	2300	2735	1165
New York, NY	865	215	820	505	1650	1855	635	745	1225	2570	2915	1325	1265	1410	-	90	2530	1890	2855	3085	3025	240
Oklahoma City, OK	850	1745	795	1105	210	630	1075	450	340	1110	1350	1575	775	725	1530	1445	1010	495	1330	1695	2155	1370
Omaha, NE	1030	1455	480	820	685	540	745	935	185	1300	1670	1690	375	995	1295	1215	1355	960	1630	1730	1825	1150
Philadelphia, PA	760	310	770	425	1565	1770	590	1650	1145	2480	2830	1235	180	1315	90	-	2445	1795	2780	2960	2945	145
Phoenix, AZ	1860	2750	1840	2105	1015	905	2075	1210	1275	290	380	2430	1670	1555	2530	2445	-	1005	360	810	1600	2370
Portland, OR	2660	3230	2250	2600	2145	1350	2525	2370	1955	1000	1020	3440	1830	2655	3090	3010	1385	2250	1090	535	175	2945
St. Louis, MO	585	1190	300	565	645	860	515	780	255	1620	1945	1295	555	690	970	890	1545	950	1830	2160	2240	920
Salt Lake City, UT	1960	2435	1415	1800	1240	540	1725	1505	1105	420	705	2625	1300	1735	2275	2195	650	1365	760	730	925	2130
San Antonio, TX	1025	2110	1245	1490	270	975	1460	200	790	1290	1390	1435	1190	575	1890	1795	1005	-	1375	1795	2305	1650
San Diego, CA	2230	3285	2280	2625	1430	1245	2545	1580	1655	340	125	2885	2085	1935	2855	2780	360	1375	-	525	1315	2705
San Francisco, CA	2495	3200	2125	2565	1795	1270	2495	1985	1905	570	420	3240	9080	2300	3085	2960	810	1795	525	-	790	2900
Seattle, WA	2690	3165	3185	2535	2225	1430	2460	2445	985	1180	1195	3470	1695	2735	3025	2945	1600	2305	1315	790	-	2880
Washington, DC	620	445	700	360	1415	1710	252	1505	1070	2420	2725	1100	1120	1165	240	145	2370	1650	2705	2900	2880	-

Finding Lost Ancient Cities

Say that you know all the distances between a collection of cities, but you don't know where they all are. If you know a few of the cities' locations, can you find the others?



"Ancient data, modern math and the hunt for 11 lost cities of the Bronze Age"

The Washington Post

Introduction to Metagenomics

Fill in the blank: Over half the cells in your body are _____.

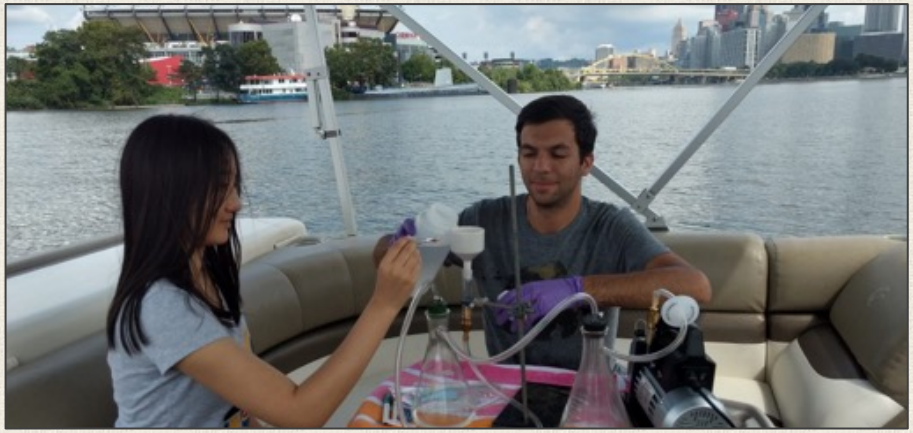
Introduction to Metagenomics

Fill in the blank: Over half the cells in your body are *bacteria* .

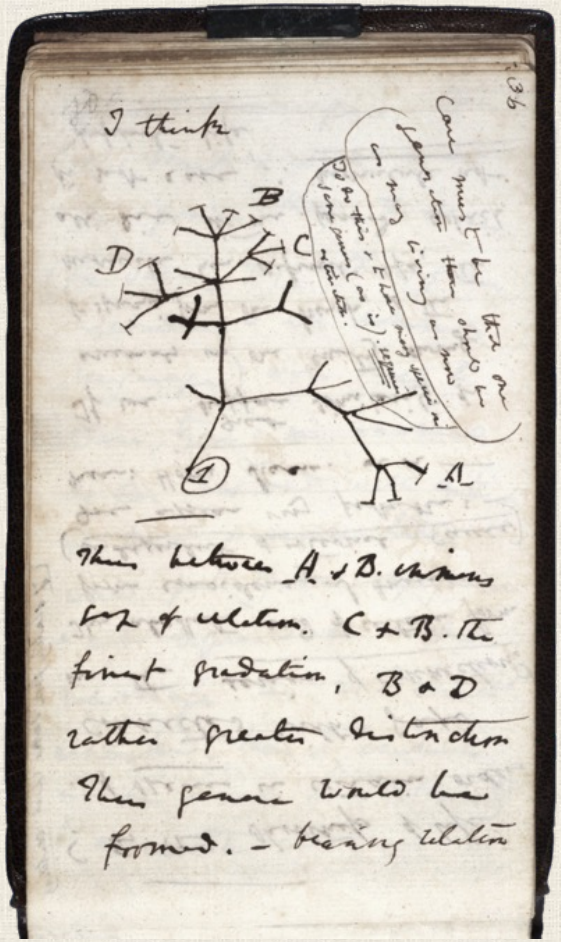
Introduction to Metagenomics

Fill in the blank: Over half the cells in your body are *bacteria* .

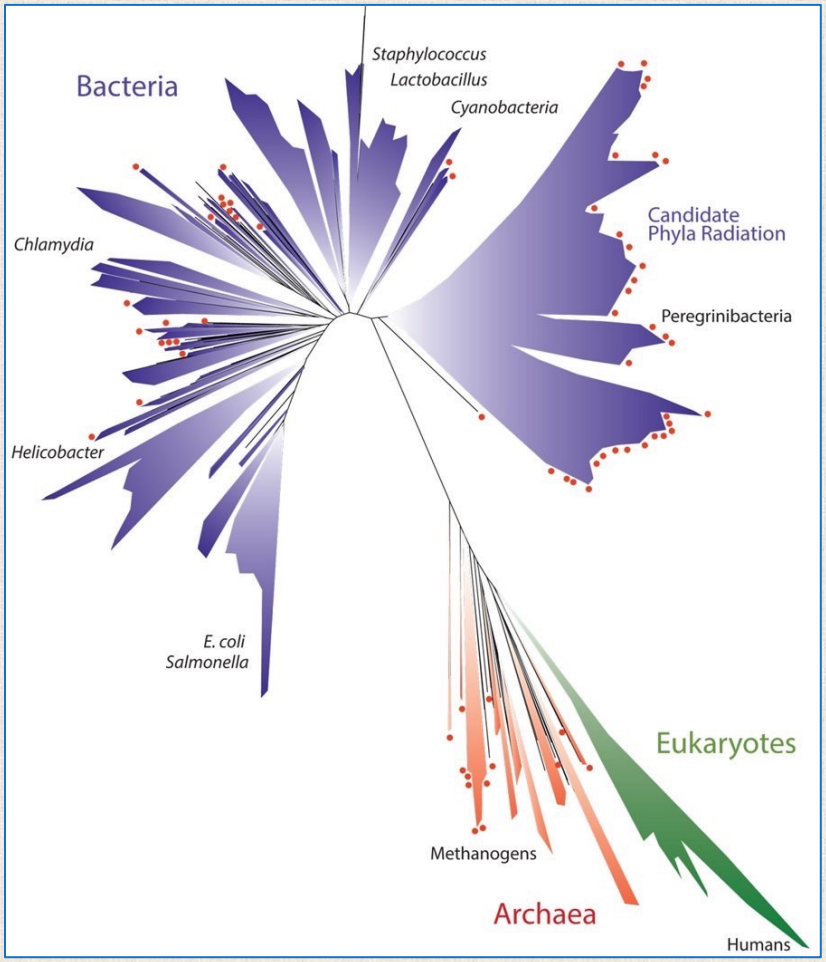
Metagenomics: The study of DNA (or RNA) recovered from an environmental sample.



"Why Do We Care About Bacteria?"



Darwin's notebook c. 1837



Hug et al., 2016
Courtesy: *Nature Biotechnology, Discovery Magazine*

??????

Wait … How Many Species Are There?

STOP: … I'm open to guesses on what you think!

Wait … How Many Species Are There?

STOP: … I'm open to guesses on what you think!

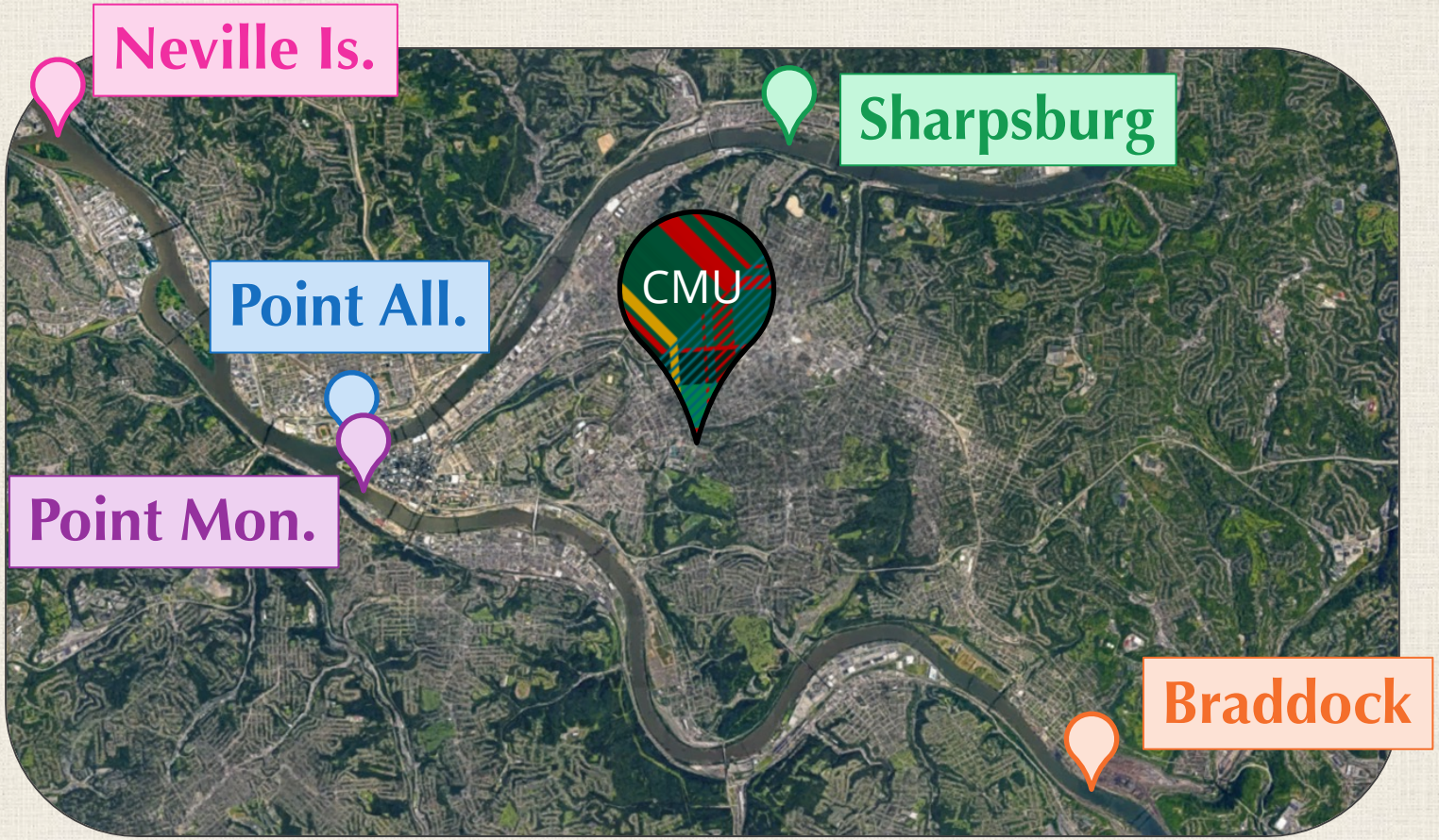
Another can of worms: what is a species?

Introduction

HOW many species are there on Earth? This is a fundamental question in science, but one that remains far from resolved. It is widely agreed that the number of described species (approximately 1.5 million species; Roskov et al. 2014) underestimates actual global richness, but the extent of this underestimation remains unclear. Projections of global biodiversity have ranged from as low as ~2 million species (Costello et al. 2012), up to ~100 million (e.g., Ehrlich and Wilson 1991; May 1992; Lambshead 1993), or even ~1 trillion (Locey and Lennon 2016).

Larsen et al., 2017

Our Five Sampling Points Over Four Seasons



DNA Sequencing Gives us a Collection of Short Strings for Each Sample

Multiple genomes from different microbes

Shatter the genome into reads

Sequence the reads

Interlude: How Are Reads Sequenced?

https://www.youtube.com/watch?v=fCd6B5HRaZ8

DNA Sequencing Gives us a Collection of Short Strings for Each Sample

Multiple genomes from different microbes

Shatter the genome into reads

Sequence the reads

DNA Sequencing Gives us a Collection of Short Strings for Each Sample

Multiple genomes from different microbes

Shatter the genome into reads

Sequence the reads

We could use whole genomes (millions of nucleotides per bacterium), but we will instead sequence **16S ribosomal RNA** (1500 nucleotides).

DNA Sequencing Gives us a Collection of Short Strings for Each Sample

Multiple genomes from different microbes

Shatter the genome into reads

Sequence the reads

STOP: If we have a collection of (possibly repeated) strings in a single sample, how can we store them computationally?

DNA Sequencing Gives us a Collection of Short Strings for Each Sample

Multiple genomes from different microbes

Shatter the genome into reads

Sequence the reads

Answer: One way is to use an array of strings, but because we may have duplicates, there is a better answer that you already know …

Recall the Frequency Map from Prep Materials!

Frequency Map Problem

- **Input:** A string *text* and an integer *k*.
- **Output:** The frequency map of *text* for all *k*-mers.

Strings	Count
"ACGGATATGACG"	4
"GAGATATACGAG"	7
"GAGAGACCACCT"	1
"TTACAGATCACG"	2
"CAGGATATCGTC"	4
"ACGGATATGACG"	8
"GAGATATACGAG"	20
(thousands more)	

QUESTION 1: HOW CAN WE DETERMINE THE DIVERSITY OF A GIVEN SAMPLE?

This is an Old Question

The diversity present within a single site/ecosystem/sample is called its **alpha diversity**.

Species	Count
Bears	4
Turkeys	7
Coyotes	1
Deer	2
Foxes	4
Groundhogs	8
Squirrels	20

This is an Old Question

The diversity present within a single site/ecosystem/sample is called its **alpha diversity**.

Species	Count
Bears	4
Turkeys	7
Coyotes	1
Deer	2
Foxes	4
Groundhogs	8
Squirrels	20

Richness of ecosystem: number of species present in the ecosystem.

So the richness of this ecosystem would be 7.

This is an Old Question

STOP: Why is richness not a great measure of diversity? How can we improve it?

Species	Count
Bears	4
Turkeys	7
Coyotes	1
Deer	2
Foxes	4
Groundhogs	8
Squirrels	20

Richness of ecosystem: number of species present in the ecosystem.

So the richness of this ecosystem would be 7.

From Richness to Evenness

Species	Count
Coyotes	1
Deer	999

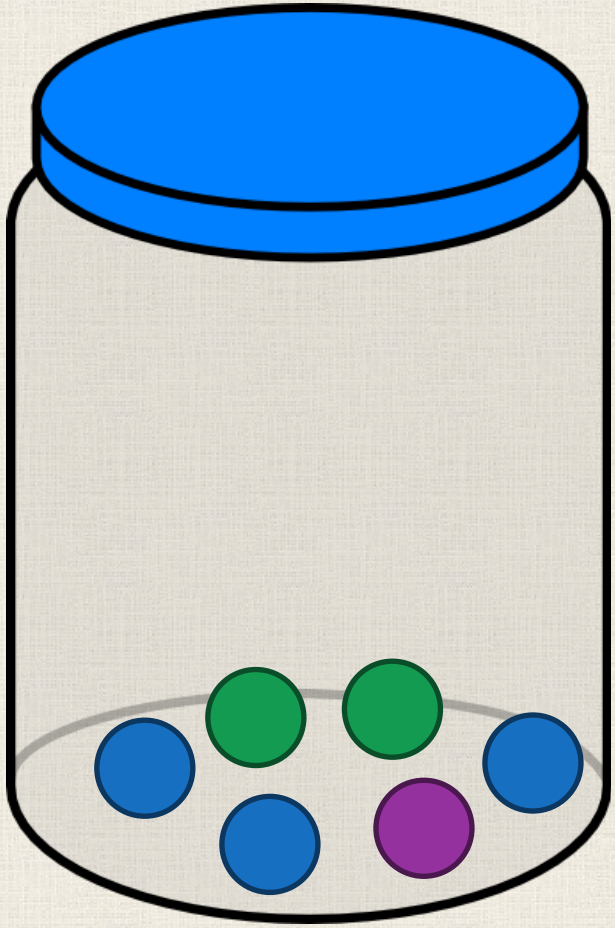
Species	Count
Coyotes	500
Deer	500

These two habitats have the same richness, but the second habitat is more diverse.

Evenness: how evenly *divided* the counts of species are in a sample.

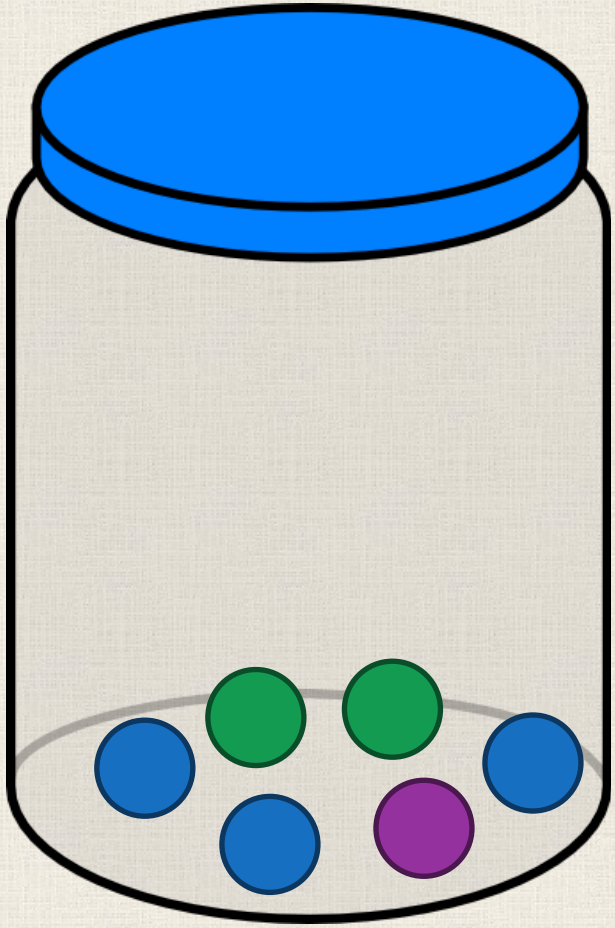
A Probabilistic Detour?

Exercise: If we draw a ball, replace it, and then draw another ball, what is the probability that we drew a green ball twice?



A Probabilistic Detour?

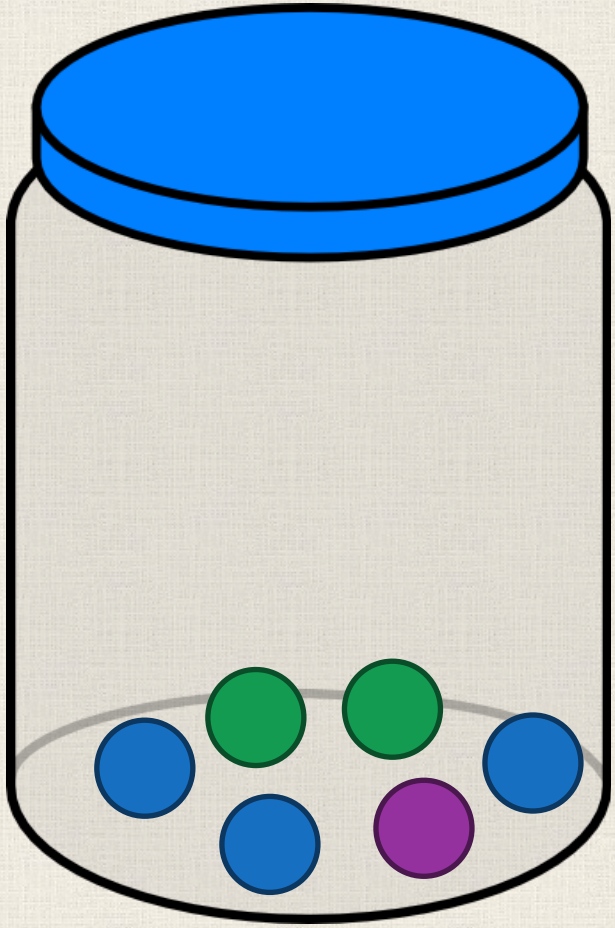
Exercise: If we draw a ball, replace it, and then draw another ball, what is the probability that we drew a green ball twice?



Answer: (2/6) * (2/6) = (4/36) = 1/9.

A Probabilistic Detour?

Exercise: What about the probability of getting a blue ball twice? What about getting a purple ball twice?

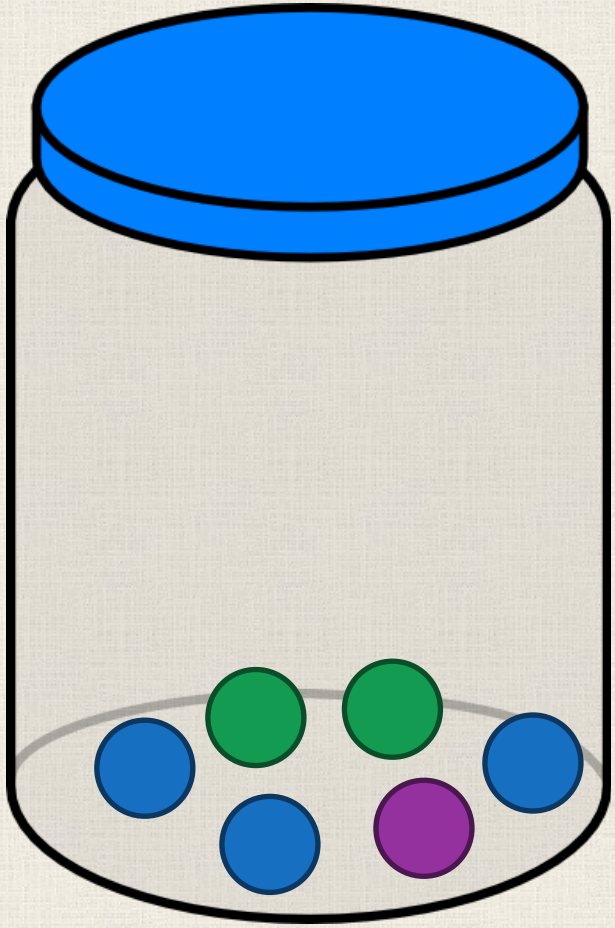


A Probabilistic Detour?

Exercise: What about the probability of getting a blue ball twice? What about getting a purple ball twice?

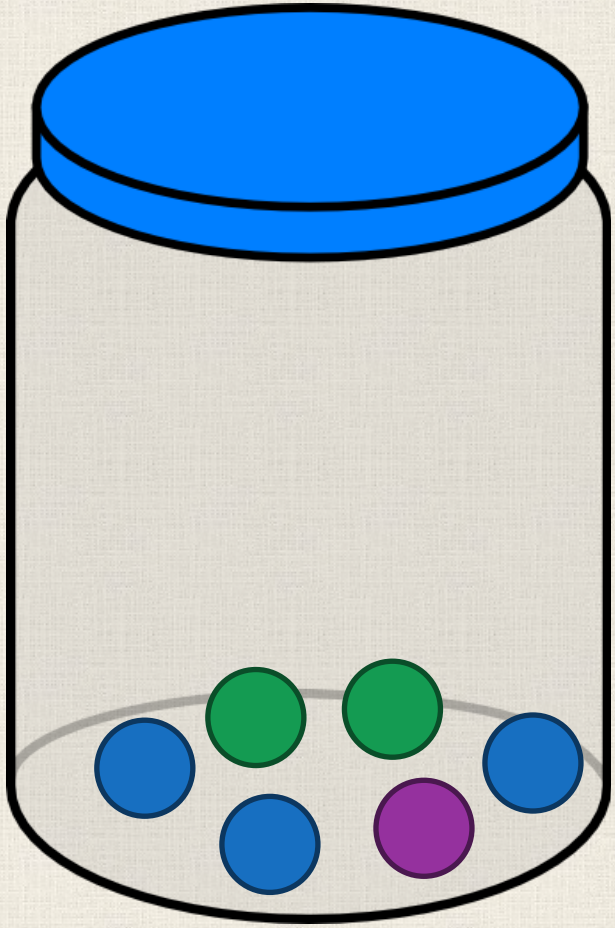
Answer: For a blue ball, it is (3/6) * (3/6) = 1/4.

For a purple ball, it is (1/6)*(1/6) = 1/36.



A Probabilistic Detour?

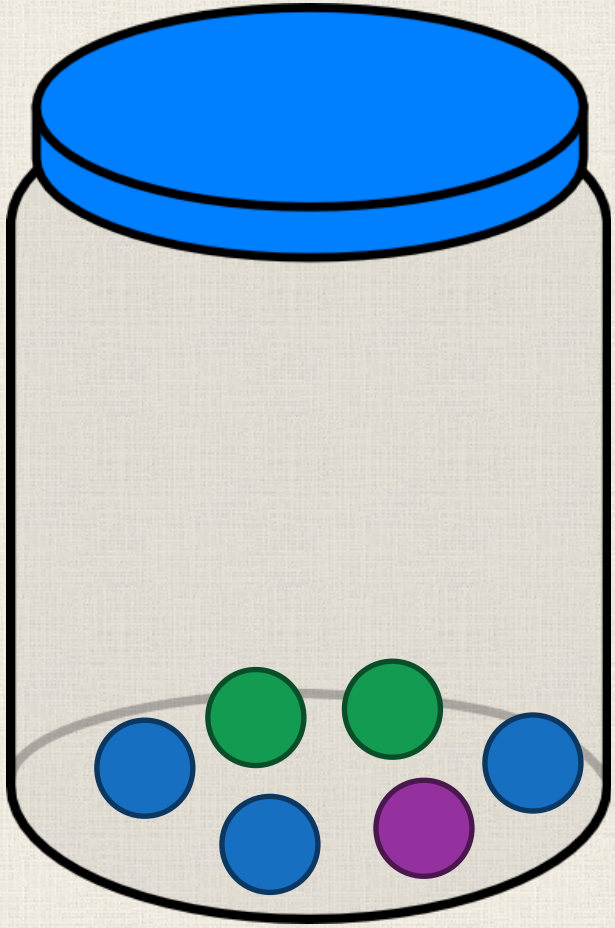
Exercise: What about the probability of drawing two balls of the same color?



A Probabilistic Detour?

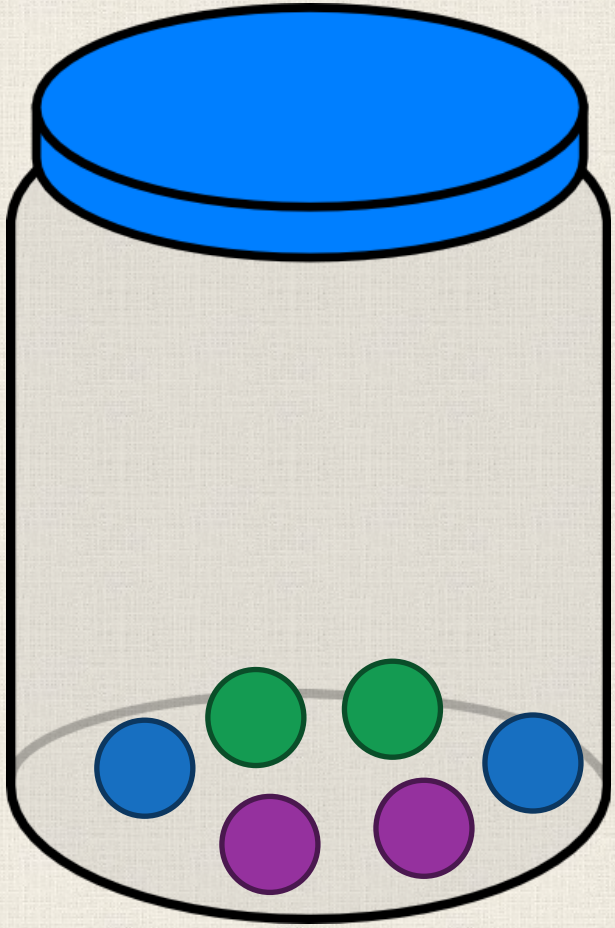
Exercise: What about the probability of drawing two balls of the same color?

Answer: It is the sum of our three probabilities: 1/4 + 1/9 + 1/36 = 7/18 = 0.3888…



A Probabilistic Detour?

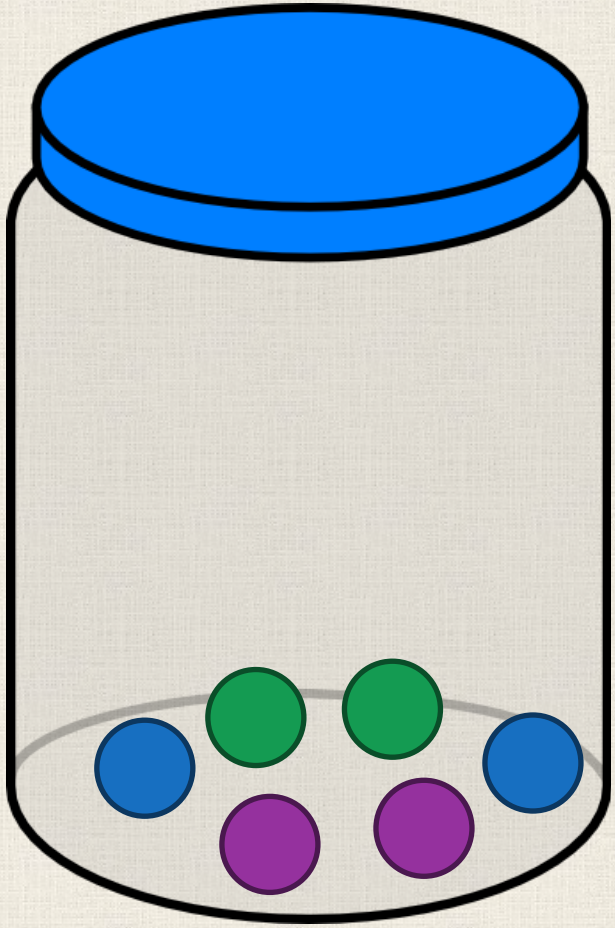
Exercise: What is the probability of drawing two balls of the same color if we instead have two balls of each color?



A Probabilistic Detour?

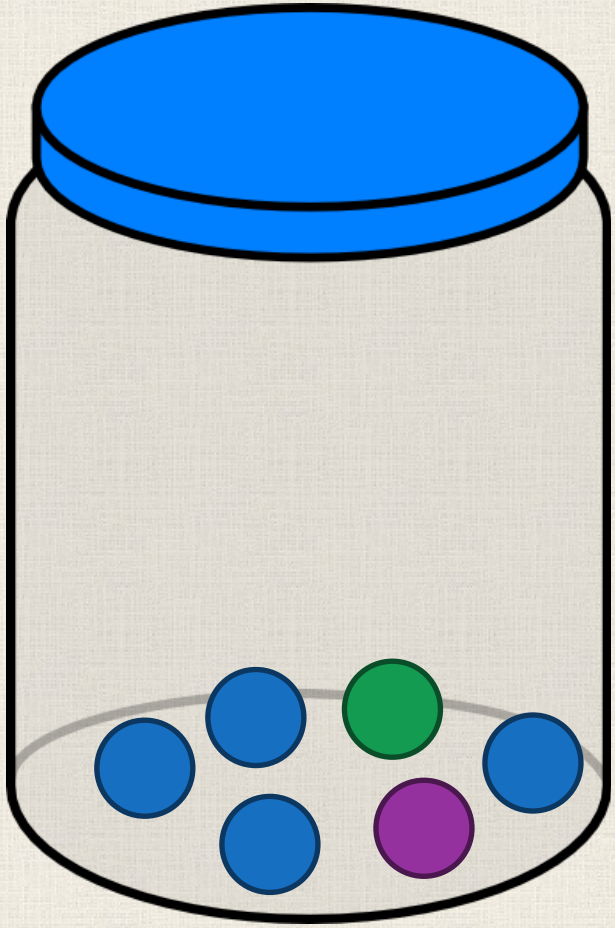
Exercise: What is the probability of drawing two balls of the same color if we instead have two balls of each color?

Answer: It is the sum $(2/6)^2 + (2/6)^2 + (2/6)^2 = 1/3 = 0.333\ldots$



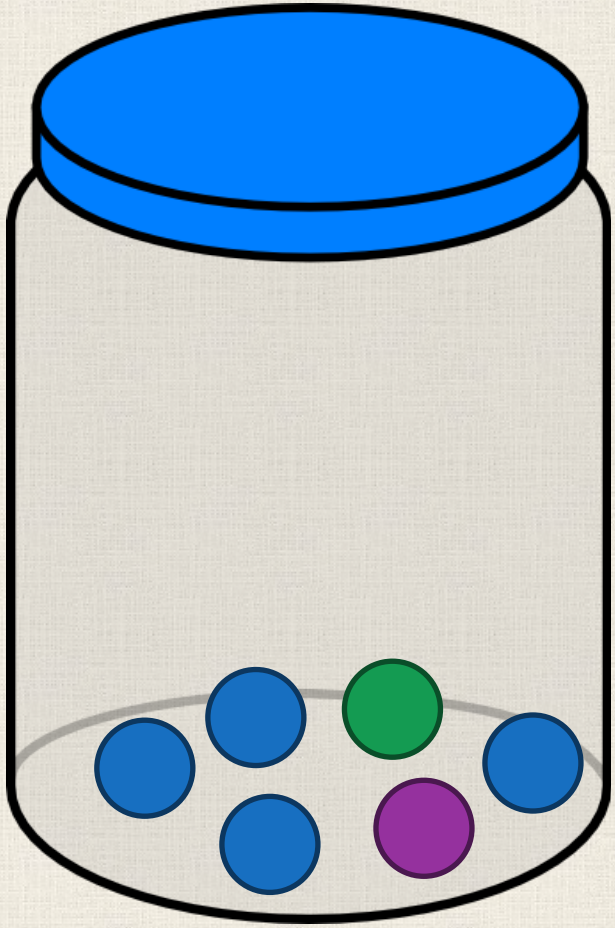
A Probabilistic Detour?

Exercise: And what if we have 1 green ball, 1 purple ball, and four blue balls?



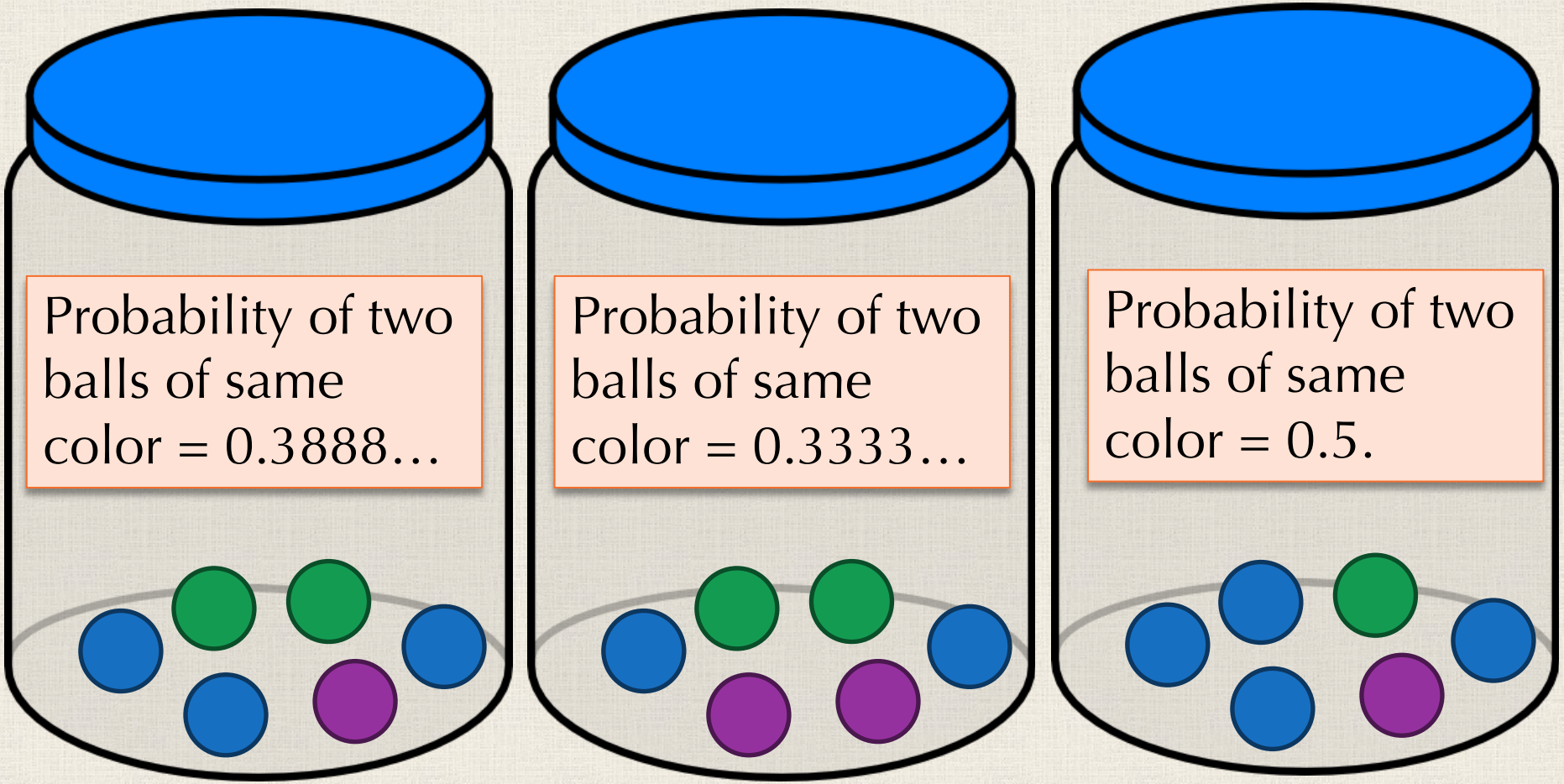
A Probabilistic Detour?

Exercise: And what if we have 1 green ball, 1 purple ball, and four blue balls?



Answer: It is the sum $(1/6)^2 + (1/6)^2 + (4/6)^2 = 1/2 = 0.5$.

Which of the Jars is the Most Even?



In an Even Sample it is *Less* Likely to Choose the Same Object Twice

Evenness: how evenly *divided* the counts of species are in a sample.

The less even a sample, the more skewed it is toward one species, and the more likely to "choose the same species twice."

Species	Count
Coyotes	1
Deer	999

Species	Count
Coyotes	500
Deer	500

In an Even Sample it is *Less* Likely to Choose the Same Object Twice

Simpson's index: The probability of choosing two objects (with replacement) of the same type.

Exercise: What is Simpson's index for each of the following samples?

Species	Count
Coyotes	1
Deer	999

Species	Count
Coyotes	500
Deer	500

In an Even Sample it is *Less* Likely to Choose the Same Object Twice

Simpson's index: The probability of choosing two objects (with replacement) of the same type.

Exercise: What is Simpson's index for each of the following samples?

Species	Count
Coyotes	1
Deer	999

Species	Count
Coyotes	500
Deer	500

In an Even Sample it is *Less* Likely to Choose the Same Object Twice

Simpson's index: The probability of choosing two objects (with replacement) of the same type.

Exercise: What is Simpson's index for each of the following samples?

$(1/1000)^2 + (999/1000)^2 = 0.998$

Species	Count
Coyotes	1
Deer	999

$(1/2)^2 + (1/2)^2 = 0.5$

Species	Count
Coyotes	500
Deer	500

In an Even Sample it is *Less* Likely to Choose the Same Object Twice

STOP: What is the smallest or largest that Simpson's index could ever be for an ecosystem?

Exercise: What is Simpson's index for each of the following samples?

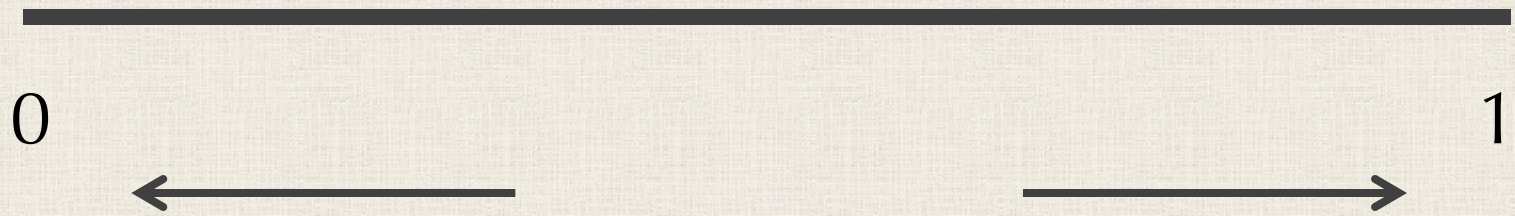
$(1/1000)^2 + (999/1000)^2 = 0.998$

Species	Count
Coyotes	1
Deer	999

$(1/2)^2 + (1/2)^2 = 0.5$

Species	Count
Coyotes	500
Deer	500

Simpson's Index Ranges from 0 to 1



Simpson's index heads toward zero as the number of "species" and their evenness increase.

Simpson's index heads toward one as we have fewer, less even "species".

QUESTION 2: HOW CAN WE COMPARE TWO DIFFERENT SAMPLES?

Comparing Two Ecosystems?

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

The extent to which two sites/ecosystems/samples differ is called their **beta diversity**.

Comparing Two Ecosystems?

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

STOP: How could we *quantify* how different these two ecosystems are?

Comparing Two Ecosystems?

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

One way of measuring distance between sites:

(# non-shared species)/(# total species)

Comparing Two Ecosystems?

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

One way of measuring distance between sites:

(# non-shared species)/(# total species) = 2/5

Comparing Two Ecosystems?

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

STOP: This is a flawed way of differentiating two ecosystems. Why?

Comparing Two Ecosystems?

Site 1

Species	Count
Coyotes	1
Deer	999

Site 2

Species	Count
Coyotes	999
Deer	1

These two ecosystems would have a “distance” of zero! But they are very different …

Distance 1: Bray-Curtis Distance

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	**0**
Deer	**2**
Groundhogs	8

Site 2

Species	Count
Bears	**0**
Turkeys	**5**
Coyotes	8
Deer	4
Groundhogs	**6**

First, sum the ***minimum*** count of each species in each row.

Here, this sum is $0 + 5 + 0 + 2 + 6 = 13$.

Distance 1: Bray-Curtis Distance

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8
Total	**21**

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6
Total	**23**

Next, take the ***average*** of the total counts at each site: $(21 + 23)/2 = 22$.

Distance 1: Bray-Curtis Distance

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8
Total	**21**

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6
Total	**23**

The **Bray-Curtis distance** is 1 minus the sum of minimum values divided by the average total:

$1 - 13/22 = 9/22$.

Distance 2: Jaccard Distance

Site 1

Species	Count
Bears	**4**
Turkeys	**7**
Coyotes	0
Deer	2
Groundhogs	**8**

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	**8**
Deer	**4**
Groundhogs	6

For this distance metric, we need to sum the ***maximum*** count of each species in each row. Here, this sum is $4 + 7 + 8 + 4 + 8 = 31$.

Distance 2: Jaccard Distance

Site 1

Species	Count
Bears	4
Turkeys	7
Coyotes	0
Deer	2
Groundhogs	8

Site 2

Species	Count
Bears	0
Turkeys	5
Coyotes	8
Deer	4
Groundhogs	6

The **Jaccard distance** is 1 minus the sum of **min** values divided by the sum of **max** values:

$$1 - 13/31 = 18/31.$$

Comparing Beta Diversity Metrics

Site 1

Species	Count
Coyotes	1
Deer	999

Site 2

Species	Count
Coyotes	999
Deer	1

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

Comparing Beta Diversity Metrics

Site 1

Species	Count
Coyotes	1
Deer	999

Site 2

Species	Count
Coyotes	999
Deer	1

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

Bray-Curtis: $1 - (1 + 1)/1000 = 998/1000$.

Jaccard: $1 - (1 + 1)/(999 + 999) = 1996/1998$.

Comparing Beta Diversity Metrics

Site 1

Species	Count
Coyotes	499
Deer	501

Site 2

Species	Count
Coyotes	501
Deer	499

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

Comparing Beta Diversity Metrics

Site 1

Species	Count
Coyotes	499
Deer	501

Site 2

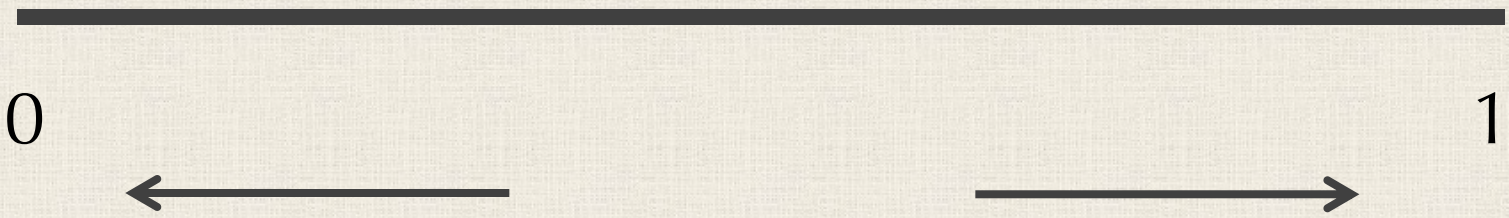
Species	Count
Coyotes	501
Deer	499

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

Bray-Curtis: $1 - (499 + 499)/1000 = 2/1000$.

Jaccard: $1 - (499 + 499)/(501 + 501) = 4/1002$.

Both Distances Range from 0 to 1



The *smaller* the distance, the more similar the habitats

The *larger* the distance, the more different the habitats

One More Distance Question

Site 1

Species	Count
Coyotes	100
Deer	900

Site 2

Species	Count
Coyotes	10
Deer	90

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

One More Distance Question

Site 1

Species	Count
Coyotes	100
Deer	900

Site 2

Species	Count
Coyotes	10
Deer	90

Exercise: Calculate the Bray-Curtis and Jaccard distances for these two sites.

Bray-Curtis: $1 - (10 + 90)/550 = 450/550$.

Jaccard: $1 - (10 + 90)/(100 + 900) = 900/1000$.

One More Distance Question

Site 1

Species	Count
Coyotes	100
Deer	900

Site 2

Species	Count
Coyotes	10
Deer	90

STOP: What is the issue we are hinting at? How could we fix it?

Bray-Curtis: $1 - (10 + 90)/550 = 450/550$.

Jaccard: $1 - (10 + 90)/(100 + 900) = 900/1000$.

One More Distance Question

Site 1

Species	Count
Coyotes	100
Deer	900

Site 2

Species	Count
Coyotes	10
Deer	90

The *ratios* of coyotes to deer at each site is the same! So you might imagine that the distance should be zero … but it's not ☹

Bray-Curtis: $1 - (10 + 90)/550 = 450/550$.

Jaccard: $1 - (10 + 90)/(100 + 900) = 900/1000$.

One More Distance Question

Site 1

Species	Count
Coyotes	100
Deer	900

Site 2

Species	Count
Coyotes	10
Deer	90

The *ratios* of coyotes to deer at each site is the same! So you might imagine that the distance should be zero … but it's not ☹

In real metagenomics experiments, if site 1 has n total counts and site 2 has m total counts with $n > m$, we randomly choose m strings from site 1.

Looking Ahead

STOP: Once we know the "distance" between every pair of samples, how might we analyze these distances? How could we visualize the distances? What patterns are we looking for?

TOWARD COMPARING MULTIPLE SAMPLES

Representing Multiple Samples

STOP: In practice, we will have far more than just two samples; we may have tens or hundreds. What variable type can store all of these samples?

Representing Multiple Samples

STOP: In practice, we will have far more than just two samples; we may have tens or hundreds. What variable type can store all of these samples?

Answer: A map of maps!

Map of Maps: Mental Image

allMaps

sample_name	Value
Allegheny_1	
Allegheny_2	•
Ohio_1	•
Ohio_2	•
Ohio_3	•
Control_1	•
Control_2	•

Key	Value
"ATGCACGCT"	8
"GGACGTACG"	1
"GTACGACAG"	2
"ATAAATTGC"	3
"GATACCAGA"	2
"ATAGGATCC"	6
"GGATATCCC"	3

This map is *allMaps*[Allegheny_1]

Map of Maps: Mental Image

allMaps

sample_name	Value
Allegheny_1	
Allegheny_2	•
Ohio_1	•
Ohio_2	•
Ohio_3	•
Control_1	•
Control_2	•

Key	Value
"ATGCACGCT"	8
"GGACGTACG"	1
"GTACGACAG"	2
"ATAAATTGC"	3
"GATACCAGA"	2
"ATAGGATCC"	6
"GGATATCCC"	3

STOP: how would we declare the *allMaps* variable?

Map of Maps: Mental Image

allMaps

sample_name	Value
Allegheny_1	
Allegheny_2	•
Ohio_1	•
Ohio_2	•
Ohio_3	•
Control_1	•
Control_2	•

Key	Value
"ATGCACGCT"	8
"GGACGTACG"	1
"GTACGACAG"	2
"ATAAATTGC"	3
"GATACCAGA"	2
"ATAGGATCC"	6
"GGATATCCC"	3

Answer: *allMaps* := make(map[string](map[string]int)).

First: Alpha Diversity of Many Samples

Multiple Richness Problem:

- **Input:** A map of frequency maps *allMaps*.
- **Output:** A map *R* such that *R*[*sample*] is the richness of *allMaps*[*sample*].

Multiple Evenness Problem:

- **Input:** A map of frequency maps *allMaps*.
- **Output:** A map *E* such that *E*[*sample*] is the Simpson's index of *allMaps*[*sample*].

Code Challenge: Solve these problems.

Next: Beta Diversity of Many Samples

allMaps

sample_name	Value
Allegheny_1	
Allegheny_2	•
Ohio_1	•
Ohio_2	•
Ohio_3	•
Control_1	•
Control_2	•

Key	Value
"ATGCACGCT"	8
"GGACGTACG"	1
"GTACGACAG"	2
"ATAAATTGC"	3
"GATACCAGA"	2
"ATAGGATCC"	6
"GGATATCCC"	3

STOP: How could we represent the beta diversity between *all pairs* of samples?

We Can Use a Matrix to Represent All Distances Between Pairs

In matrix *D*, entry *D*[*i*][*j*] corresponds to the distance (Bray-Curtis or Jaccard) between samples *i* and *j*.

(Hypothetical matrix)

Sample	0	1	2	3	4	5	6
0	0	0.91	0.76	0.21	0.26	0.60	0.78
1	0.91	0	0.05	0.24	0.91	0.47	0.51
2	0.76	0.05	0	0.37	0.97	0.41	0.03
3	0.21	0.24	0.37	0	0.92	0.77	0.55
4	0.26	0.91	0.97	0.92	0	0.23	0.41
5	0.60	0.47	0.41	0.77	0.23	0	0.76
6	0.78	0.51	0.03	0.55	0.41	0.76	0

We Can Use a Matrix to Represent All Distances Between Pairs

STOP: Why are the diagonal entries zero?

(Hypothetical matrix)

Sample	0	1	2	3	4	5	6
0	0	0.91	0.76	0.21	0.26	0.60	0.78
1	0.91	0	0.05	0.24	0.91	0.47	0.51
2	0.76	0.05	0	0.37	0.97	0.41	0.03
3	0.21	0.24	0.37	0	0.92	0.77	0.55
4	0.26	0.91	0.97	0.92	0	0.23	0.41
5	0.60	0.47	0.41	0.77	0.23	0	0.76
6	0.78	0.51	0.03	0.55	0.41	0.76	0

We Can Use a Matrix to Represent All Distances Between Pairs

Answer: The distance from any sample to itself is zero.

(Hypothetical matrix)

Sample	0	1	2	3	4	5	6
0	0	0.91	0.76	0.21	0.26	0.60	0.78
1	0.91	0	0.05	0.24	0.91	0.47	0.51
2	0.76	0.05	0	0.37	0.97	0.41	0.03
3	0.21	0.24	0.37	0	0.92	0.77	0.55
4	0.26	0.91	0.97	0.92	0	0.23	0.41
5	0.60	0.47	0.41	0.77	0.23	0	0.76
6	0.78	0.51	0.03	0.55	0.41	0.76	0

We Can Use a Matrix to Represent All Distances Between Pairs

STOP: Why is the distance matrix "symmetric" across this diagonal?

(Hypothetical matrix)

Sample	0	1	2	3	4	5	6
0	0	0.91	0.76	0.21	0.26	0.60	0.78
1	0.91	0	0.05	0.24	0.91	0.47	0.51
2	0.76	0.05	0	0.37	0.97	0.41	0.03
3	0.21	0.24	0.37	0	0.92	0.77	0.55
4	0.26	0.91	0.97	0.92	0	0.23	0.41
5	0.60	0.47	0.41	0.77	0.23	0	0.76
6	0.78	0.51	0.03	0.55	0.41	0.76	0

We Can Use a Matrix to Represent All Distances Between Pairs

Answer: $D[i][j] = D[j][i]$ because both these values represent the distance between samples i and j.

(Hypothetical matrix)

Sample	0	1	2	3	4	5	6
0	0	0.91	0.76	0.21	0.26	0.60	0.78
1	0.91	0	0.05	0.24	0.91	0.47	0.51
2	0.76	0.05	0	0.37	0.97	0.41	0.03
3	0.21	0.24	0.37	0	0.92	0.77	0.55
4	0.26	0.91	0.97	0.92	0	0.23	0.41
5	0.60	0.47	0.41	0.77	0.23	0	0.76
6	0.78	0.51	0.03	0.55	0.41	0.76	0

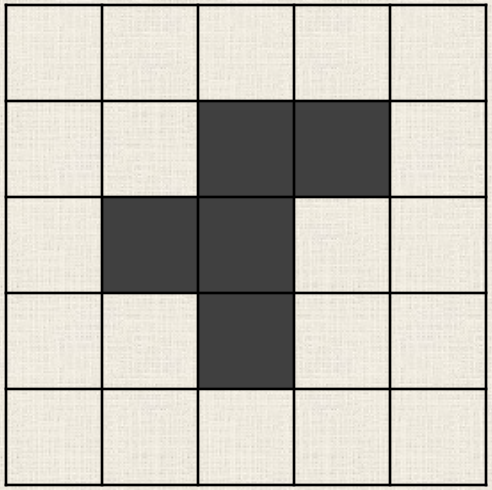
Producing a Beta Diversity Matrix

Multiple Beta Diversity Problem:

- **Input:** A map of frequency maps *allMaps* and a string *distanceMetric*.
- **Output:** A sorted list *samples* of the sample names, as well as a 2-D array *D* such that *D*[*i*][*j*] is the distance from *samples*[*i*] to *samples*[*j*] using the distance metric indicated by *distanceMetric*.

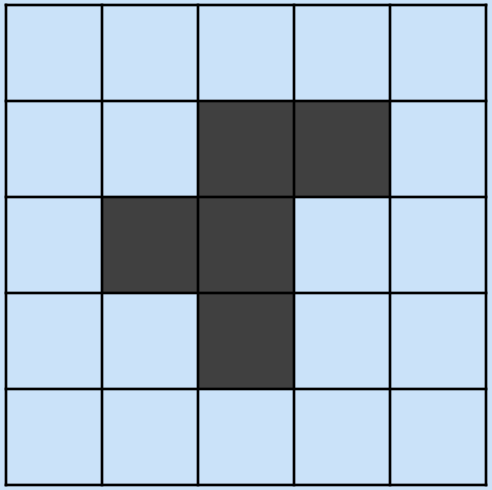
Recall: Multi-dimensional arrays

We think of a 2-D array as an array of length 5 (rows), where each element is an array of length 5 (columns).



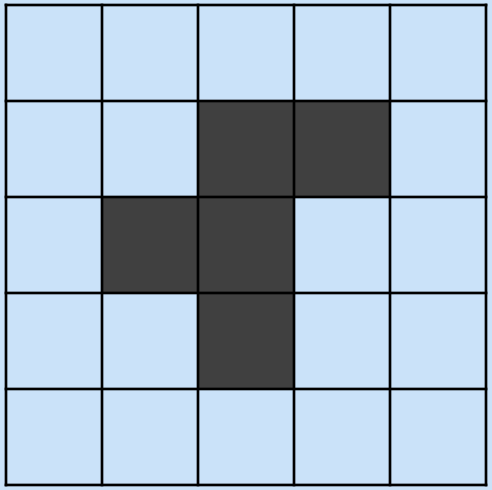
In Go, slices must be made

```
func main() {
	rPentomino := make([][]bool, 5)
	for row := range rPentomino {
		rPentomino[row] = make([]bool, 5)
	}



}
```



Setting the Cells of the R pentomino

```
func main() {
	rPentomino := make([][]bool, 5)
	for row := range rPentomino {
		rPentomino[row] = make([]bool, 5)
	}
	rPentomino[1][2] = true
	rPentomino[1][3] = true
	rPentomino[2][1] = true
	rPentomino[2][2] = true
	rPentomino[3][2] = true
	PrintBoard(rPentomino)
}
```



Returning to the Beta Diversity Matrix

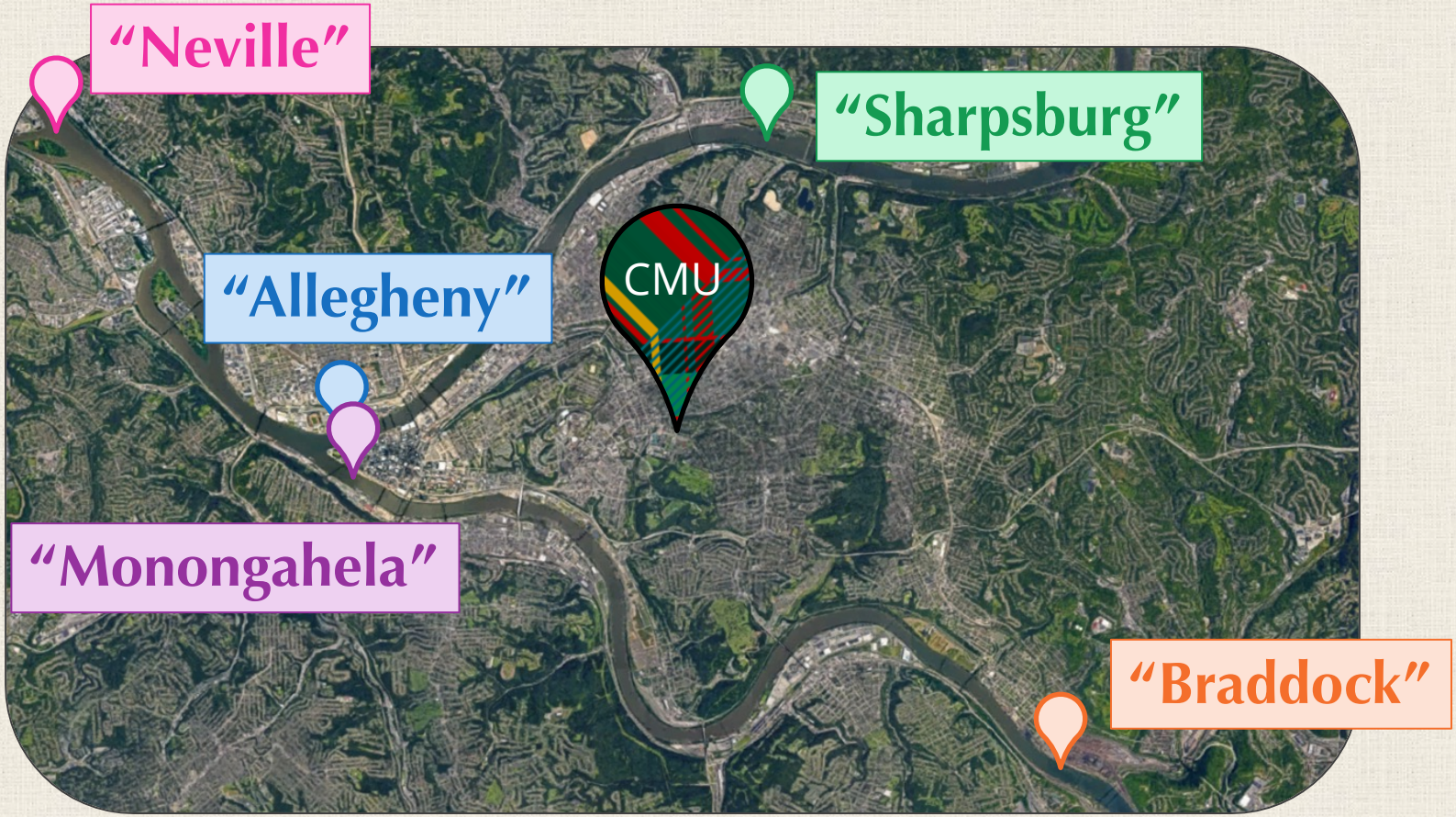
Multiple Beta Diversity Problem:

- **Input:** A map of frequency maps *allMaps* and a string *distanceMetric*.
- **Output:** A sorted list *samples* of the sample names, as well as a 2-D array *D* such that *D*[*i*][*j*] is the distance from *samples*[*i*] to *samples*[*j*] using the distance metric indicated by *distanceMetric*.

Code Challenge: Solve this problem. You may find the `sort.Strings()` function helpful …

PUTTING IT ALL TOGETHER

Recall: Our Five Sampling Points Over Four Seasons



Running Our Code on Real Data

We have:

- files corresponding to multiple samples over four seasons at five locations as well as control samples (distilled water).

We need:

- code to read strings from a file and convert strings to frequency maps. (Prep work FTW!)
- code to write the results of our alpha and beta diversity metrics to a file. (Provided!)
- code to generate plots from these files and help answer our biological questions. (R!)