ADS2 Week4 - Simulating sample data

Wanlu Liu

ZJU-UoE Institute

wanluliu@intl.zju.edu.cn
lab webisite: labw.org

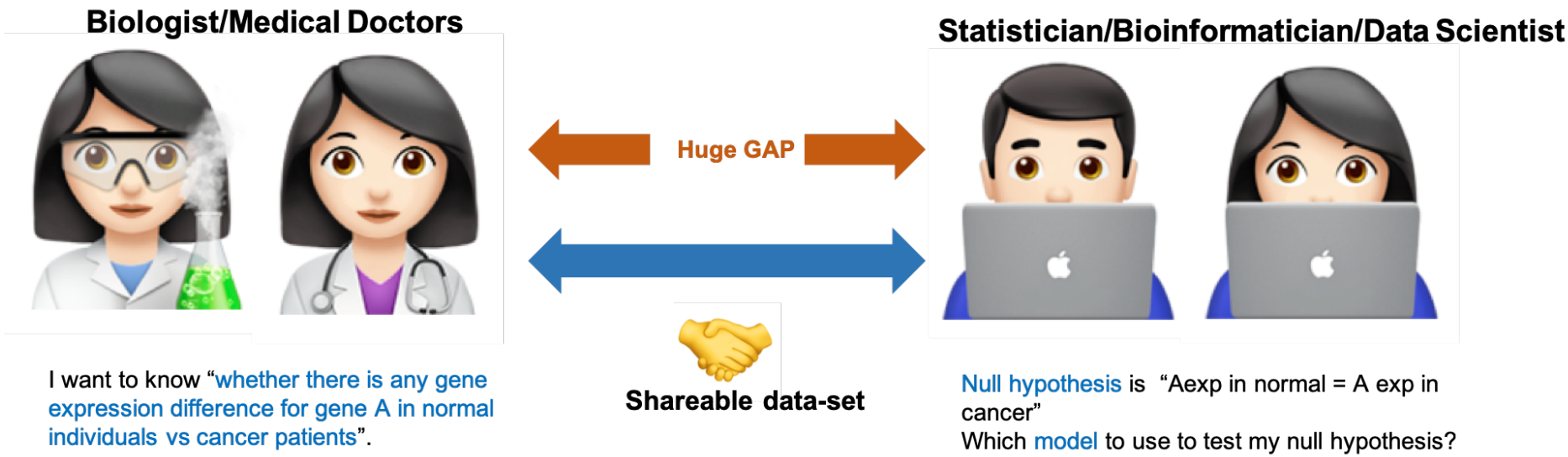
September 28, 2019

Learning Objectives - Week 4

- Describe the features of a quality shareable data-set
- Explain reasons for using synthetic data-sets, including ethical reasons
- Create synthetic data sets in R and Python
- Use synthetic data sets to test a data analysis workflow

Why shareable data-set?

Shareable data-set basically translate the 'biologist language to statistician language.



What you should deliver to the statistician?

- The raw data
- The tidy data set
- A code book describing each variable and its values in the tidy data set
- An explicit and exact recipe you used to go from step 1 to 2 and 3

'Raw Data' - Why and What

It is critical that you include the rawest form of the data that you have access to. This ensures that data provenance can be maintained throughout the workflow. Here are some examples of the raw form of data:

- The strange binary file your measurement machine spits out
- The raw sequencing data you get from sequencer
- The hand-entered numbers you collected looking through a microscope

Raw data example - raw sequencing data from RNA-seq

Raw xxx.fastq format RNA-seq data from Illumina HiSeq sequencer.

How 'Raw' is a Raw data?

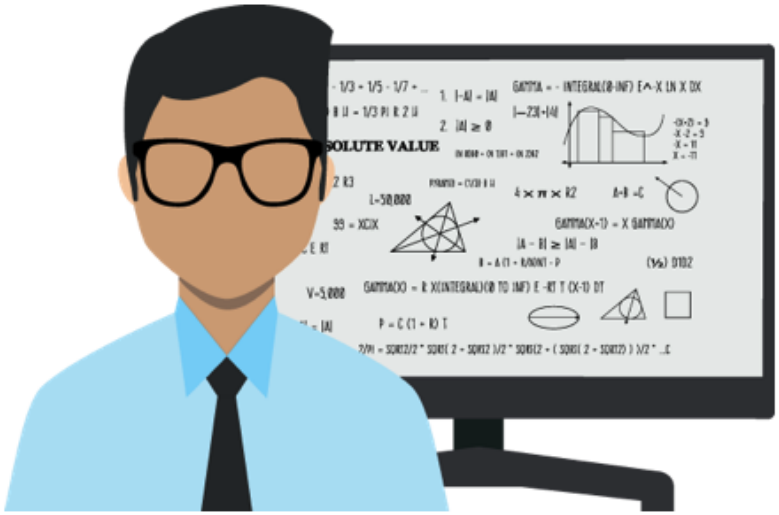
You know the raw data are in the right format if you:

- Ran no software on the data
- Did not modify any of the data values
- You did not remove any data from the data set
- You did not summarize the data in any way

What Data Scientists usually do?

What you imagine a data scientist do

Fancy modeling/machine learning/data visualization



What a data scientist really do

Data cleaning for 80% of their time

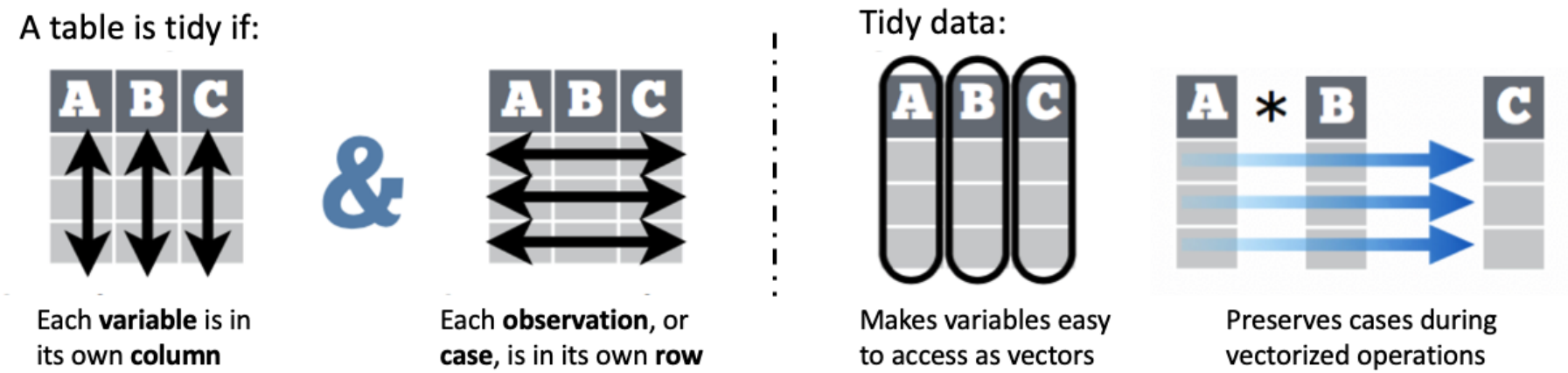
'Tidy Data' - Why and What

Tidy data is a way to organize tabular data. It provides a consistent data structure across packages/computational language (R/Python) and easy for computation/visualization. Basic principles are:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

'Tidy Data' - Principles

Tidy data is a way to organize tabular data. It provides a consistent data structure across packages/computational language (R/Python) and easy for computation/visualization. Basic principles are:

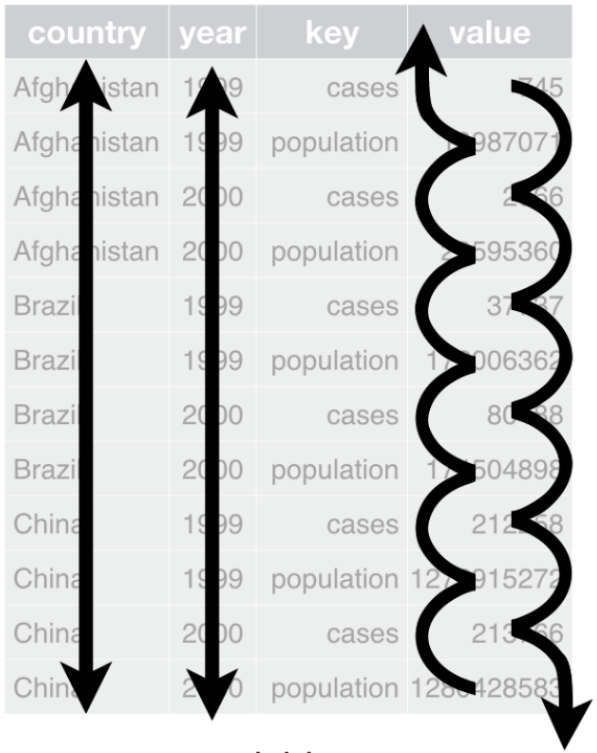


'Tidy Data' - Example (Long vs Wide)

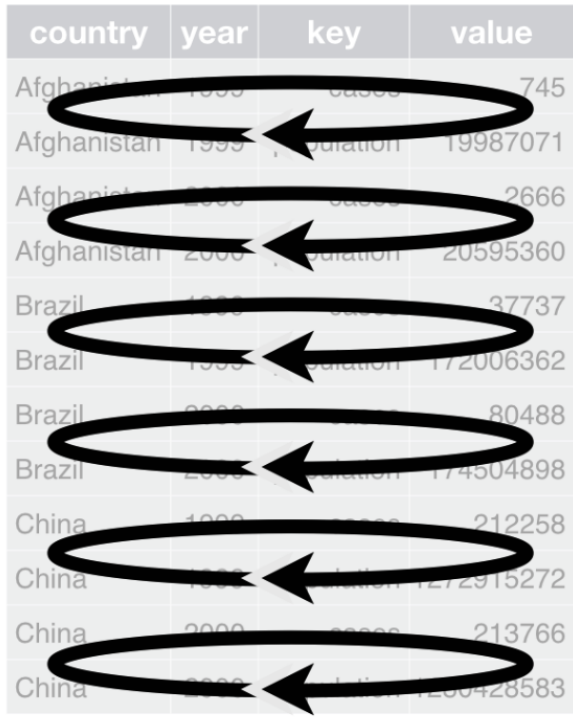
Long data formats have one observation and one measurement per row. So, multiple rows constitute a single observation.These kinds of data sets are great for plotting summary information for each group and each variable aggregated together, but hard for computing/statistics. (**NOT TIDY**)

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2



variables



observations

'Tidy Data' - Example (Long vs Wide)

Wide data has a every measurement in a single observation in a single row.This kind of data is ideal for things like scatterplots of one measurement against another, with each observation as a single data point.(**TIDY**)

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table1

variables

observations

'Code Book' - Why and What

For almost any data set, the measurements you calculate will need to be described in more detail than you can or should sneak into the spreadsheet. The **code book** contains this information. At minimum it should contain:

- Information about the variables (including units!) in the data set not contained in the tidy data
- Information about the summary choices you made
- Information about the experimental study design you used

'Code Book' - Examples

A common format for this document is a Word file. There should be a section called "Study design" that has a thorough description of how you collected the data. There should also be a section called "Code book" that describes each variable and its units. In our genomics example, the followings should be included in your code book:

- what the **unit** of measurement for each clinical/demographic variable is?
  - e.g. age in years, treatment by name/dose, level of diagnosis and how heterogeneous
- **how** you conducted your genomics experiment?
  - e.g. total RNA-seq or poly-A RNA-seq?
- **Parameters** you used to analyze genomics data?
  - annotation:UCSC/Ensembl, genome version: hg18 or hg19, mismatch number : 2 or 3 etc.
- any other information about how you did the **data collection/study design**.
  - are these the first 20 patients that walked into the clinic?
  - Are they 20 highly selected patients by some characteristic like age?
  - Are they randomized to treatments?

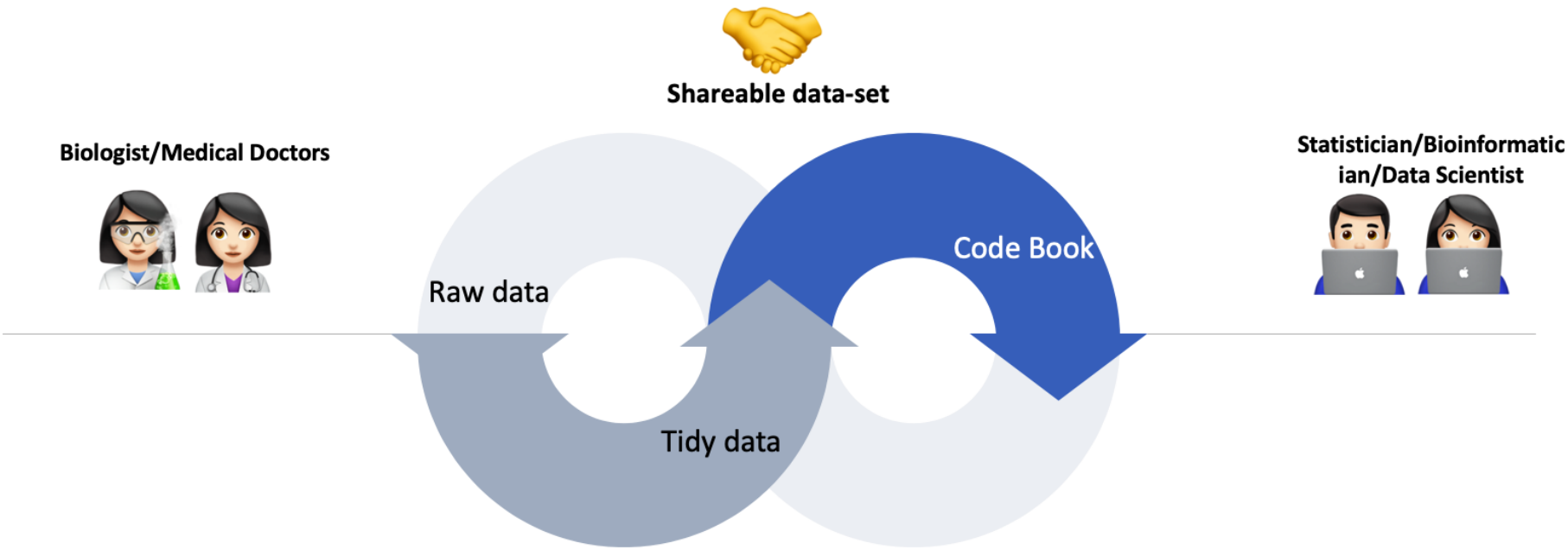
How to code variables? - Data type

When you put variables into a spreadsheet there are several main categories you will run into depending on their data type:

Statistics	**Programming**
Continuous	floating-point/numerical
Ordinal	integer/factor
categorical	integer/string/factor
binary data	Boolean
Missing	NA
Censored	NA/N.D.

Shareable data-set summary

Biologist can now communicate with Statistician.



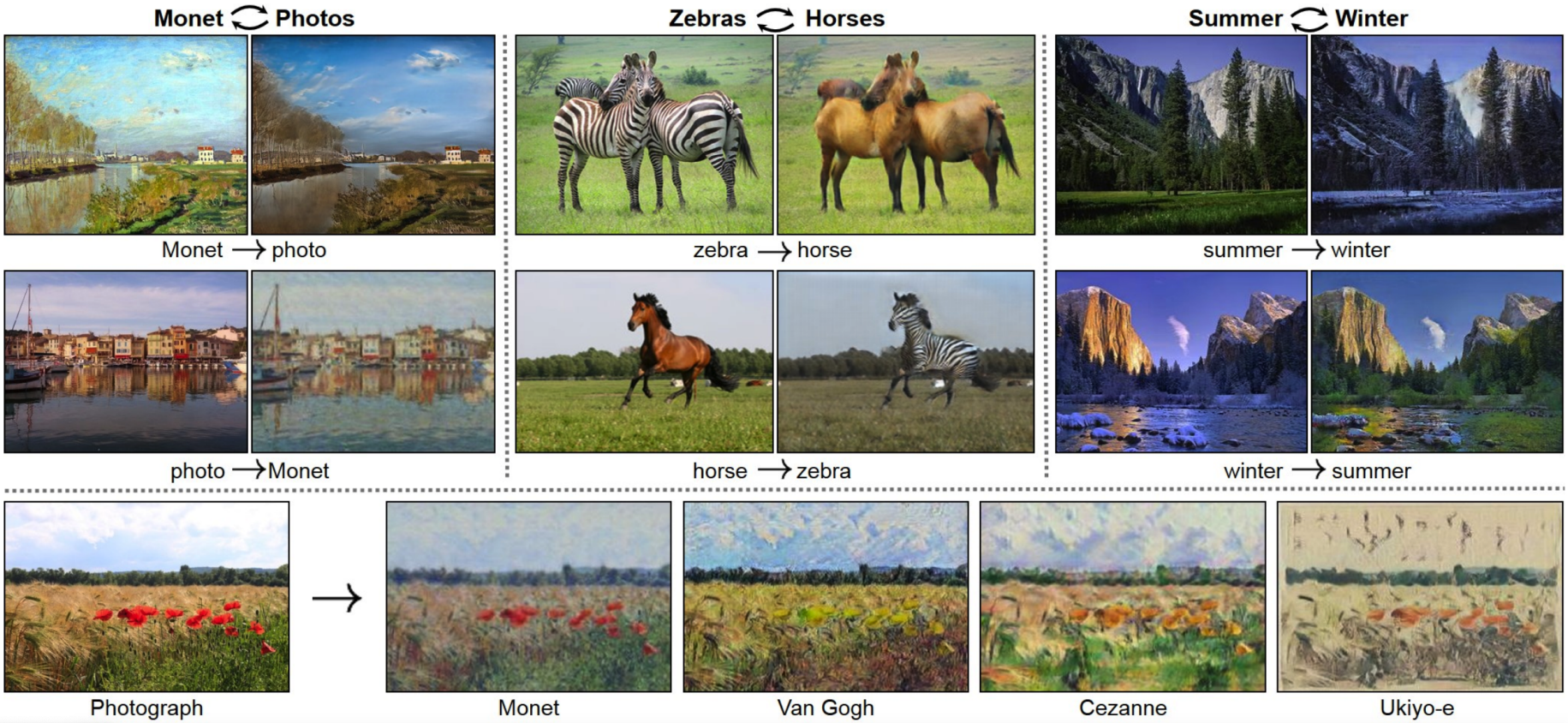
dummy or synthetic data-set - What and Why

As the name suggests, quite obviously, a **synthetic data-set** is a repository of data that is generated programmatically. Its main purpose, is to be flexible and rich enough to help statistician conduct experiments when real data is not available.

- **Machine learning**: Self driving car simulations pioneered the use of synthetic data.
- **Clinical and scientific trials**: Synthetic data can be used as a baseline for future studies and testing when no real data yet exists.
- **Research**: To help better understand the format of real data not yet recorded, develop understanding of its specific statistical properties, tune parameters for related algorithms, or build preliminary models.
- **Financial services**: Fraud protection is a major part of any financial service and with synthetic data, new fraud detection methods can be tested and evaluated for their effectiveness.
- **Healthcare**: Synthetic data enables healthcare data professionals to allow the public use of record data while still maintaining patient confidentiality.

synthetic data Example (Generative Adversarial Networks (GAN))

In deep learning field, scientist try to automatically generate facial images for neural network trianing with GAN algorithm. People then have adapted GAN algorithm to synthetically generate many other data for various applications:



CycleGAN https://junyanz.github.io/CycleGAN/

synthetic data-set - Advantages

It is often useful to build such a data-set in order to build and test a data analysis pipeline. This has several advantages:

- the pipeline can be built and tested even if the final data-set isn't available.
- The final data-set may include sensitive data (e.g. patient data), so using it for testing may pose a risk to data confidentiality or security.
- Generating synthetic data can help anticipate problems with the data collection.
- Simulated data can also be used to do things like power analysis ahead of data collection.

synthetic data-set - How

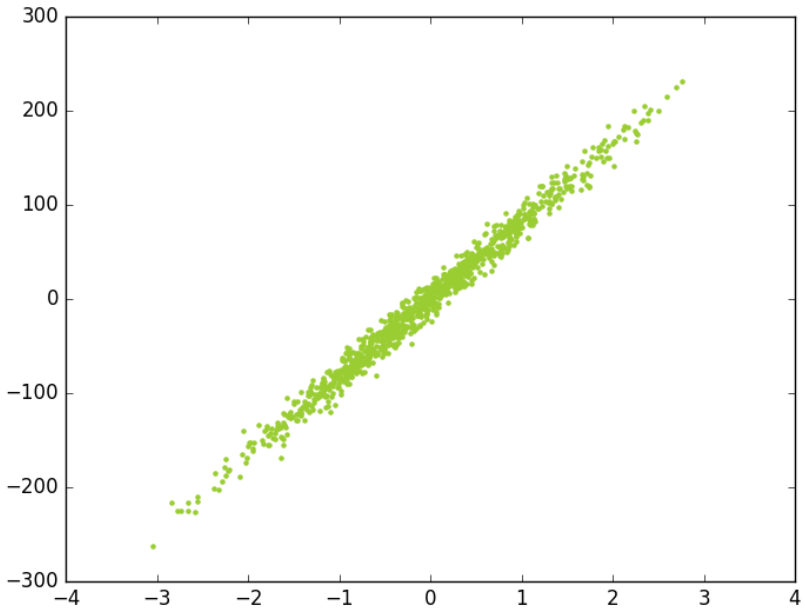
There are two major ways to generate synthetic data. They can apply to various data contexts:

- Drawing numbers from a distribution.
  - The principle is to observe real-world statistic distributions from the original data and reproduce fake data by drawing simple numbers.
- Agent based-like modeling (ABM)
  - The principle is to create a physical model that explains the observed behavior, then reproduce random data using this model.

Linear regression simulation with Python

Example (Python (sklearn))

```
>>> from sklearn import linear_model, datasets
>>> from matplotlib import pyplot as plt
>>> n_samples = 1000
>>> X, y, coef = datasets.make_regression(n_samples=n_samples,
                n_features=1,n_informative=1, noise=10,coef=True, random_state=0)
>>> plt.scatter(X,y,color='yellowgreen', marker='.')
>>> plt.show()
```



Simulate DNA sequence with Python

Example (Python)

```
>>> import random
>>> def DNA(length):
>>>     return ''.join(random.choice("A"*2+"C"*3+"G"*3+"T"*2) for _ in xrange(length))
>>> print DNA(10)
CGGCCCCGCG
>>> print DNA(50)
GTCCCCTCCGCACTTCGTACATTGGTCTTCACGTTCGGATCCCTCTACTG
>>> print DNA(100)
CGGGCCCTATTTGGGGTAAGGTGTGGAATCCGGCCGGAGTGTCCGCCCTAACTTCTGCTTTCGATGCCTCGACGTCCCGCCGGGACTTGGGCACT
```

Summary

Now you should be able to

- Understand the characteristic of shareable data-set (Raw data, tidy data, code book)
- Reveal the reasons to use synthetic data-set including ethical reasons
- Brief ideas of how to generate synthetic data-set.

References

- https://github.com/jtleek/datasharing
- https://towardsdatascience.com/synthetic-data-generation-a-must-have-skill-for-new-data-scientists-915896c0c1ae
- https://blog.aimultiple.com/synthetic-data/
- https://www.guru99.com/software-testing-test-data.html
- https://scikit-learn.org/stable/auto_examples/linear_model/plot_ransac.html#sphx-glrautoexampleslinearmodelplotransacpy
- https://en.wikipedia.org/wiki/Data_type
- Slides were generated with online LaTeX editor Overleaf

The End