Data Science Bootcamp

Course overview:

- ✓ Module 1: Introduction to Data Science & ML
- ❏ Module 2: Data Wrangling and Exploratory Data Analysis
- ❏ Module 3: Introduction to Machine Learning
- ❏ Module 4: Natural Language Processing

Now let's turn to the data we will be using...

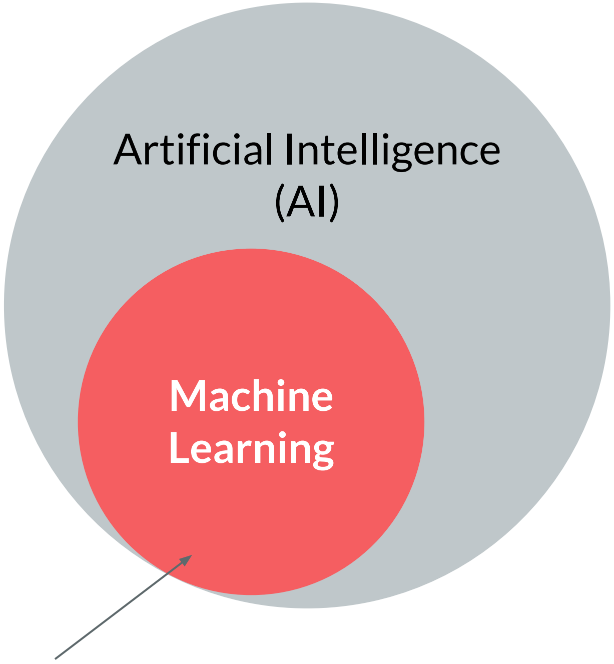
Module 1: Introduction to Data Science & ML

Module Checklist

- ✓ What is Data Science and Machine Learning?
- ❑ How do you define a research question?
- ❑ What are observations?
- ❑ What are features?
- ❑ What are outcome variables?
- ❑ Introduction to KIVA data

What is Data Science?

What is Data Science?



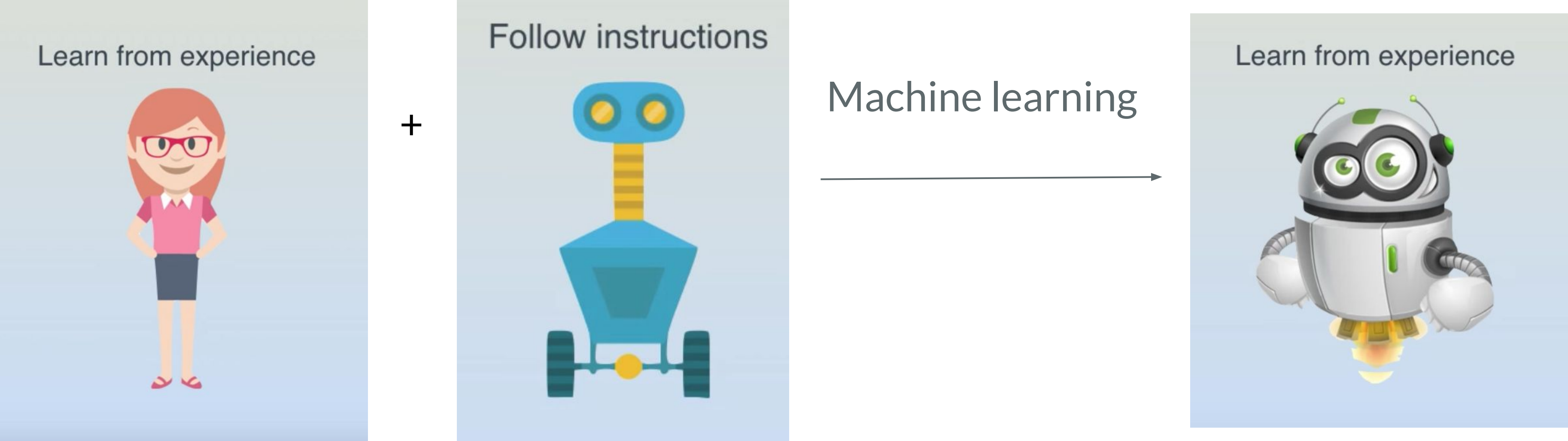
Using data science methods and sometimes big data

Machine learning is "**programming by example.**"

Instead of programming to solve a task, machine learning **seeks methods to allow a computer to come up with its own program** based on provided examples.

Sources: Ian Goodfellow, Deep Learning, http://www.deeplearningbook.org/contents/intro.html
Rob Schapire's COS 511 class, https://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf

Machine learning is a subset of AI that allows machines **to learn from raw data.**



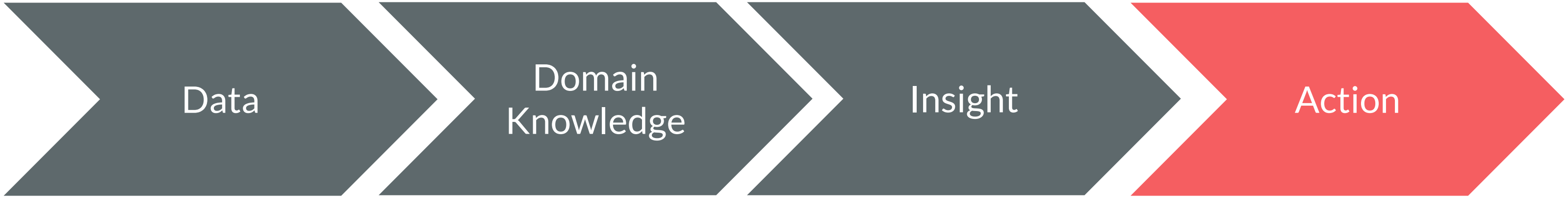
+

Machine learning

Humans learn from experience. Traditional software programing involves giving machines instructions which they perform. **Machine learning involves allowing machines to learn from raw data so that the computer program can change when exposed to new data** (learning from experience).

Source: https://www.youtube.com/watch?v=IpGxLWOIZy4

Machine learning is interdisciplinary



Machine learning is...

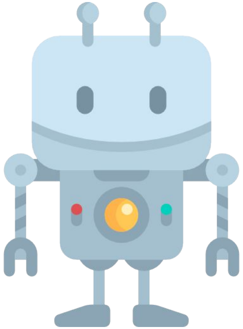
- Computer science + statistics + mathematics
- The use of data to **answer questions**

Critical thinking combined with technical toolkit

Machine learning example: Predicting malaria

Dr. Delta works to diagnose patients with malaria.
However, it takes a long time for her to see everyone.

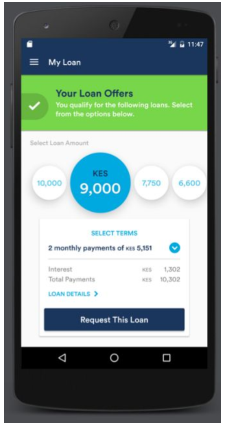
Luckily, Dr. Delta has **historical patient data about what factors predict malaria**, such as body temperature, travel history, age, medical history.



Dr. Delta can use historical data as an input in a **machine learning algorithm** to help her predict whether a new patient will have malaria.

The algorithm (the machine) learns from past data, like a human would, and is thus able to make predictions about the future.

Machine learning is a powerful tool; it can...



11:47

My Loan

Your Loan Offers

You qualify for the following loans. Select from the options below.

Select Loan Amount

10,000 | KES 9,000 | 7,750 | 6,600

SELECT TERMS

2 monthly payments of KES 5,151

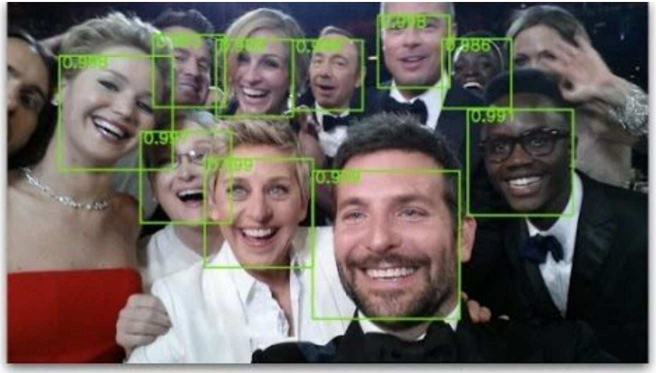
Interest KES 1,302

Total Payments KES 10,302

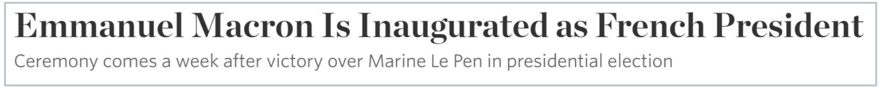
LOAN DETAILS ›

Request This Loan

Determine your credit rating based upon cell phone usage.



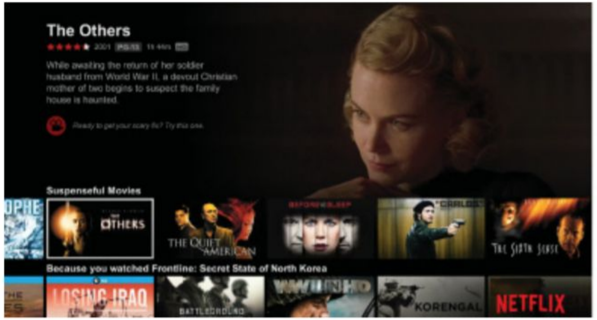
Recognize your face in a photo.



Emmanuel Macron Is Inaugurated as French President

Ceremony comes a week after victory over Marine Le Pen in presidential election

Determine the topic of a piece of text.



The Others

2001

While awaiting the return of her soldier husband from World War II, a devout Christian mother of two begins to suspect the family house is haunted.

Ready to get your scary fix? Try this one.

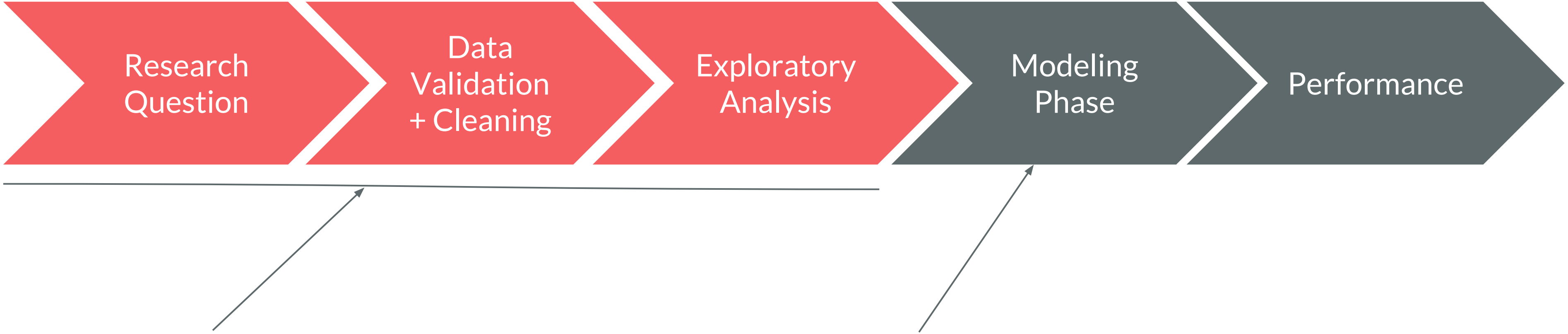
Suspenseful Movies

Because you watched Frontline: Secret State of North Korea

Recommend movies you will like.

Data Science & ML helps us answer questions. BUT.. **How do we define the question?**

Before we even get to the models/algorithms, we have to learn about our data and define our research question.



Research Question

Data Validation + Cleaning

Exploratory Analysis

Modeling Phase

Performance

~80% of your time as a data scientist is spent here, preparing your data for analysis

Machine learning takes place during the modeling phase.

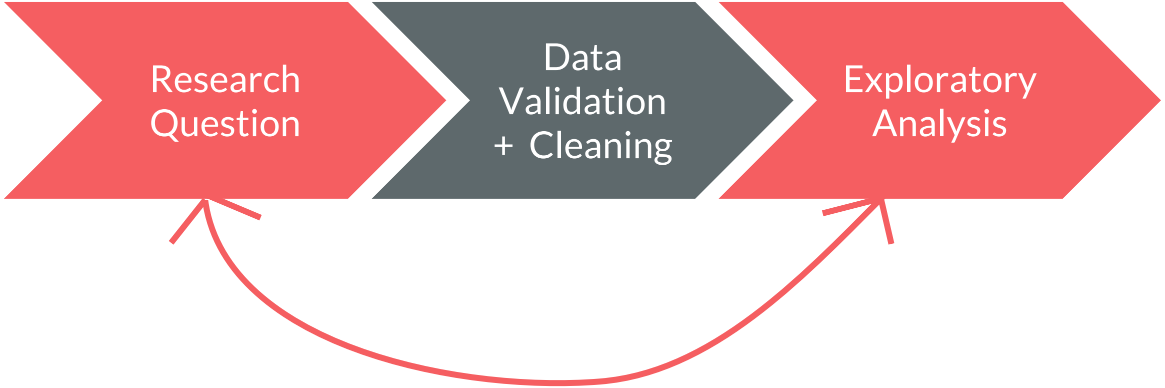
Research Question

A research question is the question we want our model to answer.

Examples of research questions:

- Does this patient have malaria?
- Can we monitor illegal deforestation by detecting chainsaw noises in audio streamed from rainforests?

We may have a question in mind **before** we look at the data, but we will often use our exploration of the data to *develop or refine* our research question.

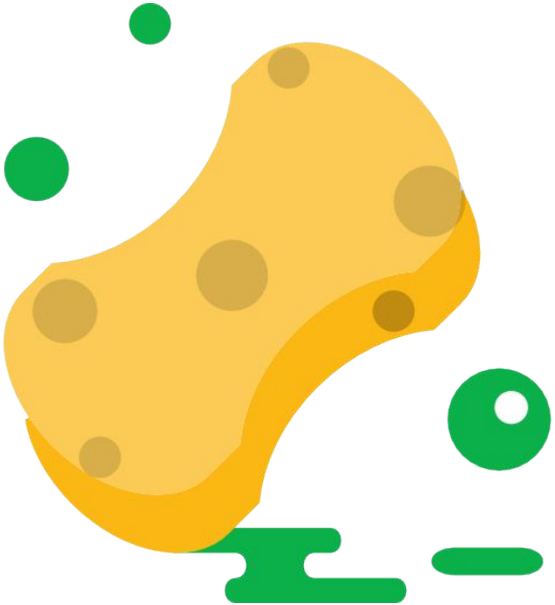


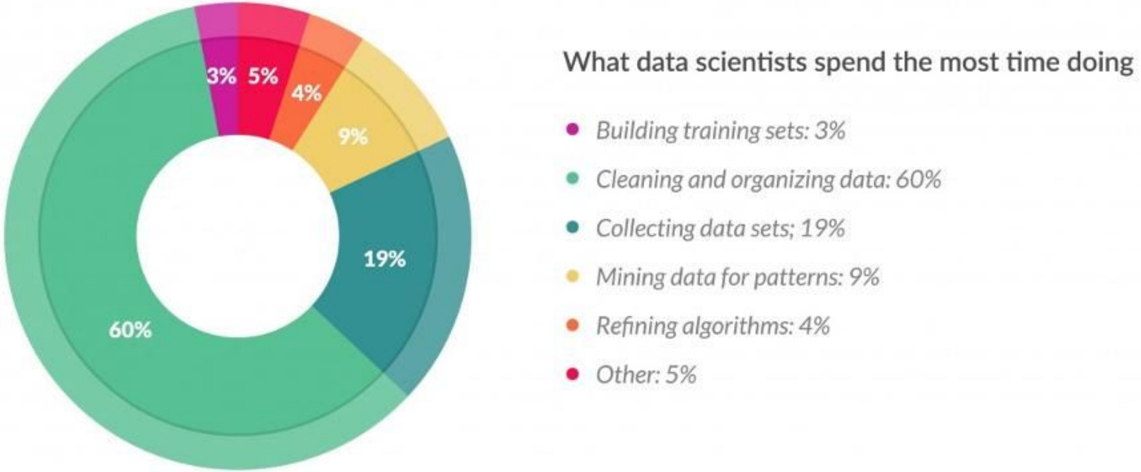
What comes first, the chicken or the egg?

Data Validation and Cleaning

Data Cleaning

"Data preparation accounts for about 80% of the work of data scientists."




What data scientists spend the most time doing
● Building training sets: 3%
● Cleaning and organizing data: 60%
● Collecting data sets; 19%
● Mining data for patterns: 9%
● Refining algorithms: 4%
● Other: 5%
3%
5%
4%
9%
19%
60%

Source: Survey of 80 data scientists. Forbes article, March 23, 2016.

Data Cleaning

Why do we need to validate and clean our data?

Data often comes from multiple sources

- Do data align across different sources?

Data is created by humans

- Does the data need to be transformed?
- Is it free from human bias and errors?

Read more: https://medium.com/@angebassa/data-alone-isnt-ground-truth-9e733079dfd4
https://www.technologyreview.com/s/608248/biased-algorithms-are-everywhere-and-no-one-seems-to-care/

Data Cleaning

Data cleaning involves identifying any issues with our data and confirming our qualitative understanding of the data.

Missing Data
Is there missing data? Is it missing systematically?

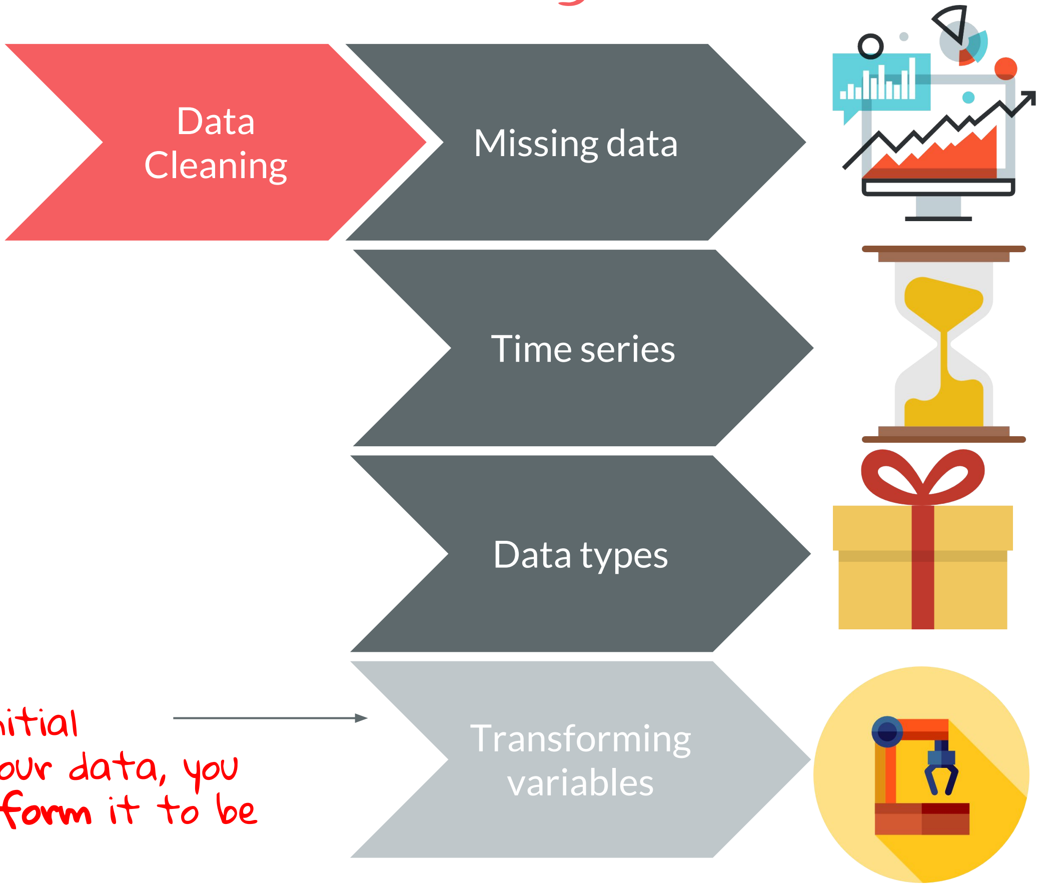
Data Type
Are all variables the right type?
Is a date treated like a date?

Times Series Validation
Is the data for the correct time range?
Are there unusual spikes in the volume of loans over time?

Data Range
Are all values in the expected range? Are all loan_amounts greater than 0?

Data Cleaning

Let's step through some examples:



After gaining an initial understanding of your data, you may need to **transform** it to be used in analysis

Data Cleaning

Missing data

Is there missing data? Is data missing randomly or systematically?

Very few datasets have no missing data; most of the time you will have to deal with missing data.

The first question you have to ask is what type of missing data you have.

Missing completely at random: no pattern in the missing data. This is the best type of missing you can hope for.

Missing at random: there is a pattern in your missing data **but not** in your variables of interest.

Missing not at random: there is a pattern in the missing data that systematically affects your primary variables.

Data Cleaning

Missing data

Is there missing data? Is data missing randomly or systematically?

Example: You have survey data from a random sample from high school students in the U.S. Some students didn’t participate:

Some students were sick the day of the day of the survey

If data is missing at random, we can use the rest of the nonmissing data without worrying about bias!

Some students declined to participate, since the survey asks about grades

If data is missing in a non-random or systematic way, your nonmissing data may be biased

Data Cleaning › Missing data

Is there missing data? Is data missing randomly or systematically?

Example: You have survey data from a random sample from high school students in the U.S. Some students didn't participate:

✓ Some students were sick the day of the day of the survey

✗ Some students declined to participate, since the survey asks about grades

If data is <u>missing at random,</u> we can use the rest of the nonmissing data without worrying about bias!

If data is missing in a non-random or <u>systematic</u> way, your nonmissing data may <u>be biased</u>

Data Cleaning → Missing data

Sometimes, you can replace missing data.

- Drop missing observations
- Populate missing values with average of available data
- Impute data

What you should do depends heavily on what makes sense for your research question, and your data.

Common imputation techniques

Use the average of nonmissing values

Take the average of observations you do have to populate missing observations - i.e., assume that this observation is also represented by the population average

Use an educated guess

It sounds arbitrary and often isn't preferred, but you can infer a missing value. For related questions, for example, like those often presented in a matrix, if the participant responds with all "4s", assume that the missing value is a 4.

Use common point imputation

For a rating scale, using the middle point or most commonly chosen value. For example, on a five-point scale, substitute a 3, the midpoint, or a 4, the most common value (in many cases). This is a bit more structured than guessing, but it's still among the more risky options. Use caution unless you have good reason and data to support using the substitute value.

Source: Handling missing data

Data Cleaning → Time series

If we have observations over time, we need to do time series validation.

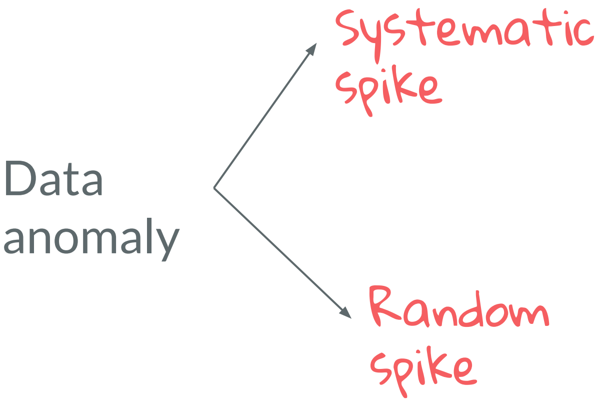
Ask yourself:

a. Is the data for the correct time range?
b. Are there unusual spikes in the data over time?

What should we do if there are unusual spikes in the data over time?

Data Cleaning > Time series

How do we address unexpected spikes in our data?



For certain datasets, (like sales data) systematic seasonal spikes are expected. For example, around Christmas we would see a spike in sales venue. This is normal, and should not necessarily be removed.

If the spike is isolated it is probably unexpected, we may want to remove the corrupted data. For example, if for one month sales are recorded in Kenyan Shillings rather than US dollars, it would inflate sales figures. We should do some data cleaning by converting to $ or perhaps remove this month.

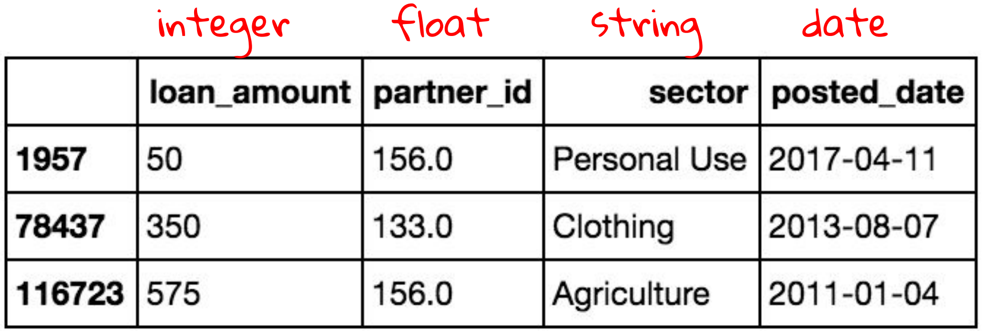
Note, sometimes there are **natural** anomalies in data that should be investigated first

Data Cleaning

Data type

Are all variables the right type?

Many functions in Python are **type** specific, which means we need to make sure all of our fields are being treated as the correct type:



integer float string date

	loan_amount	partner_id	sector	posted_date
1957	50	156.0	Personal Use	2017-04-11
78437	350	133.0	Clothing	2013-08-07
116723	575	156.0	Agriculture	2011-01-04

- **Integer:** A number with no decimal places
- **Float**: A number with decimal places
- **String**: Text field, or more formally, a sequence of unicode characters
- **Boolean**: Can only be True or False (also called indicator or dummy variable)
- **Datetime**: Values meant to hold time data.

Source: http://www.datacarpentry.org/python-ecology-lesson/03-data-types-and-format/

Data cleaning quiz!

As you explore the data, some questions arise…

Question #1

Question	Answer
There is an observation from the KIVA loan dataset that says a loan was fully funded in year 1804, but Kiva wasn't even founded then. *What do I do?*	

Question #1

Question	Answer
There is an observation that says this loan was fully funded in year 1804, but Kiva wasn't even founded then. *What do I do?*	Consult the data documentation. If no explanation exists, remove this observation.

Data Cleaning
Time series

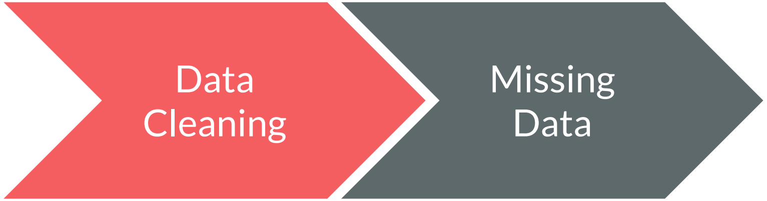
This question illustrates you should always do validation of the time range. Check what the minimum and maximum observations in your data set are.

Question #2

Question	Answer
There is an observation that states a person’s birthday is 12/1/80 but the “age” variable is missing. *What do I do?*	

Question #2

Question	Answer
There is an observation that states a person's birthday is 12/1/80 but the "age" variable is missing. *What do I do?*	We can calculate age: (e.g. 2017 - 1980 = 37)



Data Cleaning

Missing Data

This question illustrates how we may be able to leverage other fields to make an educated guess about the missing age.

Question #3

Question	Answer
The variable “amount_funded” has values of both “N/A” and “0”. *What do I do?*	

As you explore the data, some questions arise…

Question #3

Question	Answer
The variable “amount_funded” has values of both “N/A” and “0”. *What do I do?*	Check documentation if there is a material difference between NA and 0.

Question #4

Question	Answer
I’m not sure what currency the variable “amount_funded” is reported in. *What do I do?*	

Question #4

Question	Answer
I'm not sure what currency the variable "amount_funded" is reported in. *What do I do?*	Check documentation and other variables, convert to appropriate currency

A final note...

Note that our examples were all very specific - you may or may not encounter these exact examples in the wild. This is because data cleaning is very often idiosyncratic and **cannot be adequately completed by following a predetermined set of steps - you must use common sense**!

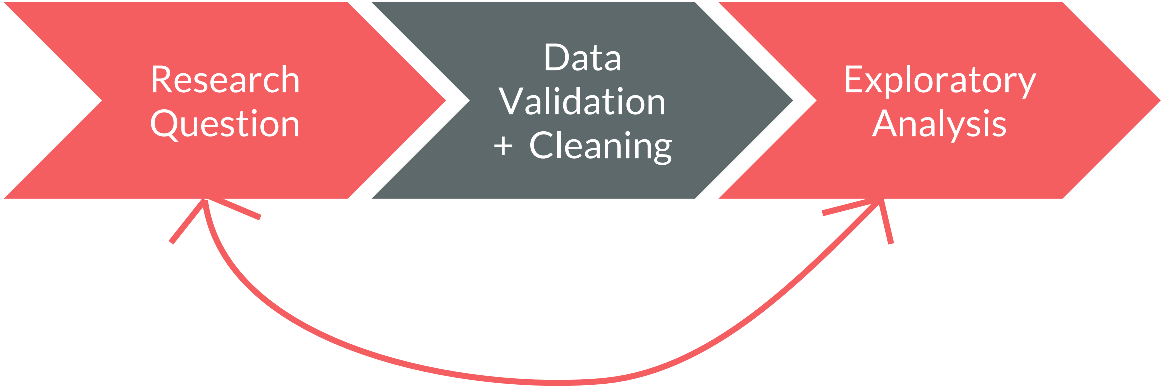
Next we turn to exploratory analysis, for which we often have to **transform our data**.

Data Cleaning
Data transformation

Exploratory Analysis

Exploratory Analysis

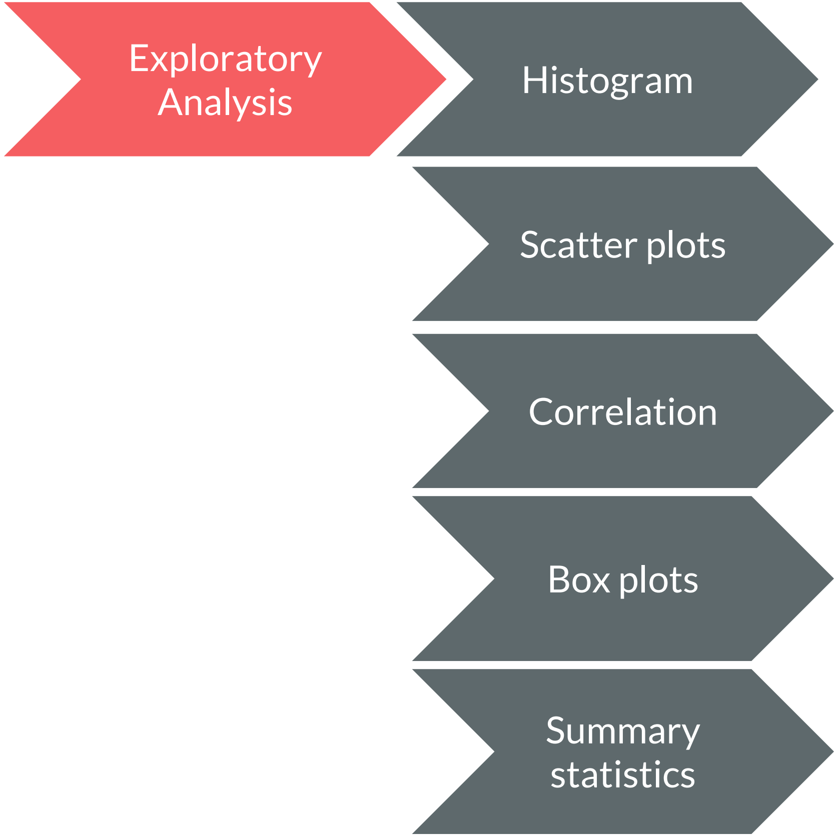
The goal of exploratory analysis is to better understand your data.



Exploratory analysis can reveal data limitations, what features are important, and inform what methods you use in answering your research question.

This is an indispensable first step in any data analysis!

Let's explore our data!



Once we have done some initial validation, we explore the data to see what models are suitable and what patterns we can identify.

The process varies depending on the data, your style, and time constraints, but typically exploration includes:

- Histogram
- Scatter plots
- Correlation tables
- Box plots
- Summary statistics
  - Mean, median, frequency

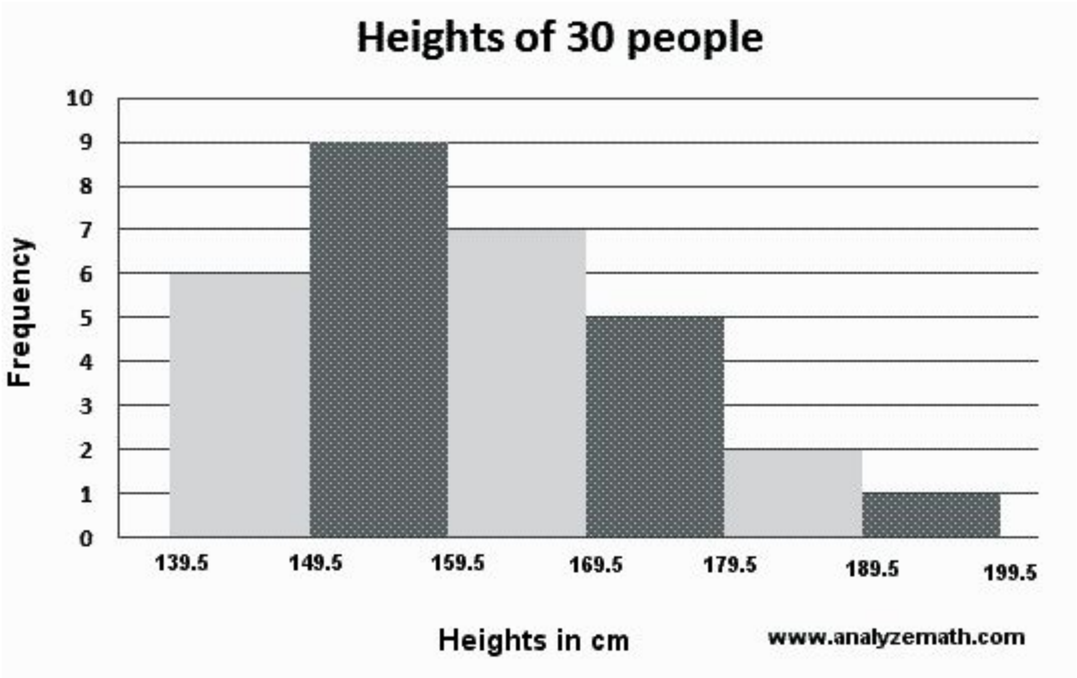
Exploratory Analysis

Histogram

Histograms tell us about the distribution of the feature.

A histogram shows the **frequency distribution** of a continuous feature.

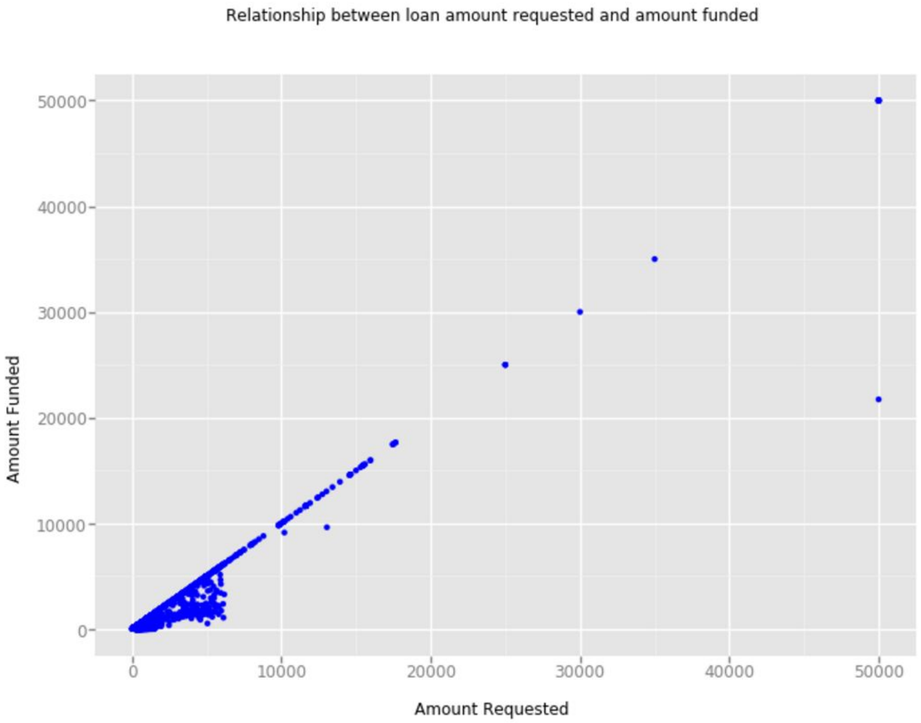
Here, we have height data of a group of people. We see that most of the people in the group are between 149 and 159 cm tall.



Exploratory Analysis

Scatter Plots

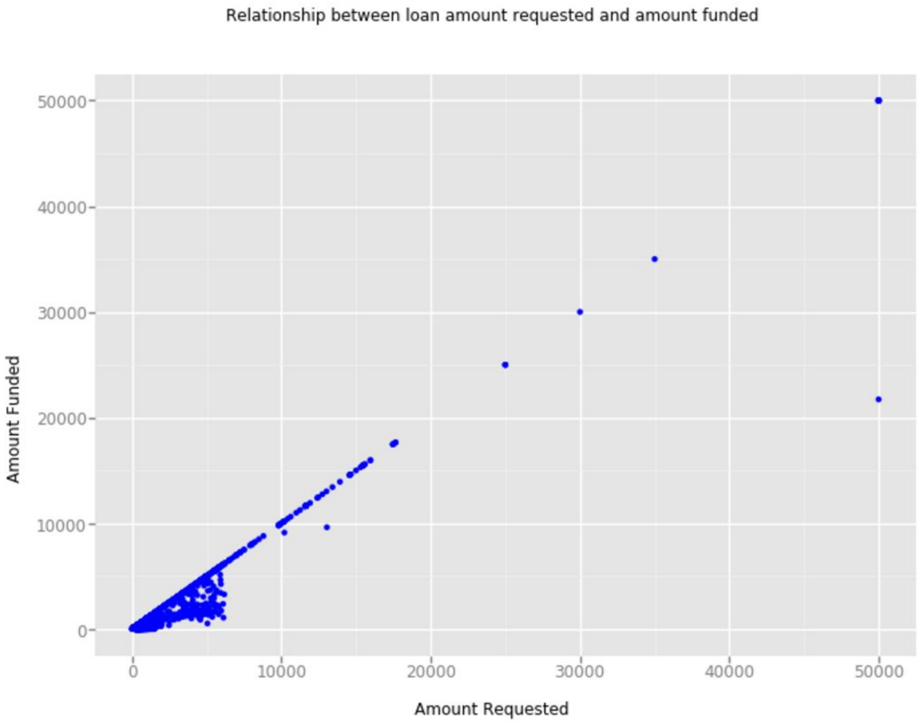
Scatter plots provide insight about the relationship between two features.



Scatter plots visualize relationships between any two features as points on a graph. They are a useful first step to exploring a research question.

Here, we can already see a positive relationship between amount funded and amount requested. ***What can we conclude?***

Scatter plot provide important data about the relationship between two features.

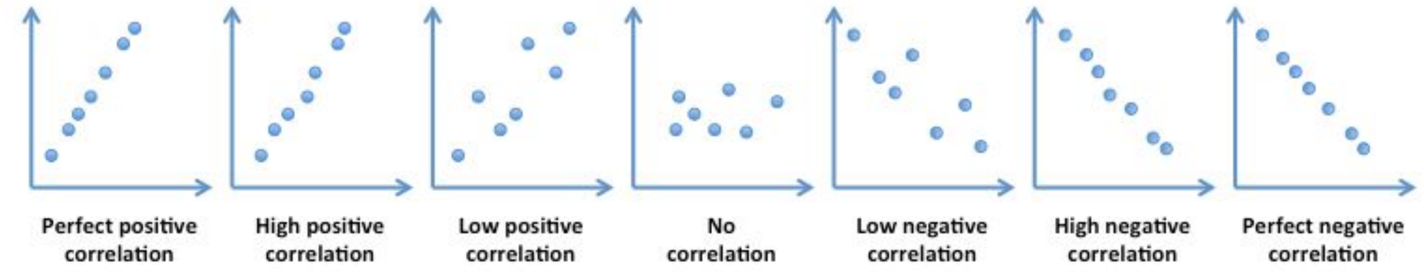


Scatter plots are an indispensable first step to exploring a research question.

Here, we can already see a positive relationship between amount funded and amount requested for a KIVA loan. ***What can we conclude?***

It looks like there is a strong relationship between what loan amount is requested and what is funded.

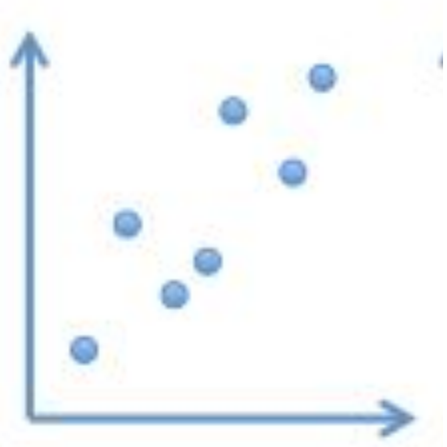
Correlation is a useful measure of the strength of a relationship between two variables. It ranges from -1.00 to 1.00



1.00 0.88 0.60 0.00 -0.55 -0.78 -1.00

See game: http://guessthecorrelation.com/

Correlation does not equal causation



Let’s say you are an executive at a company. You’ve gathered the following data:

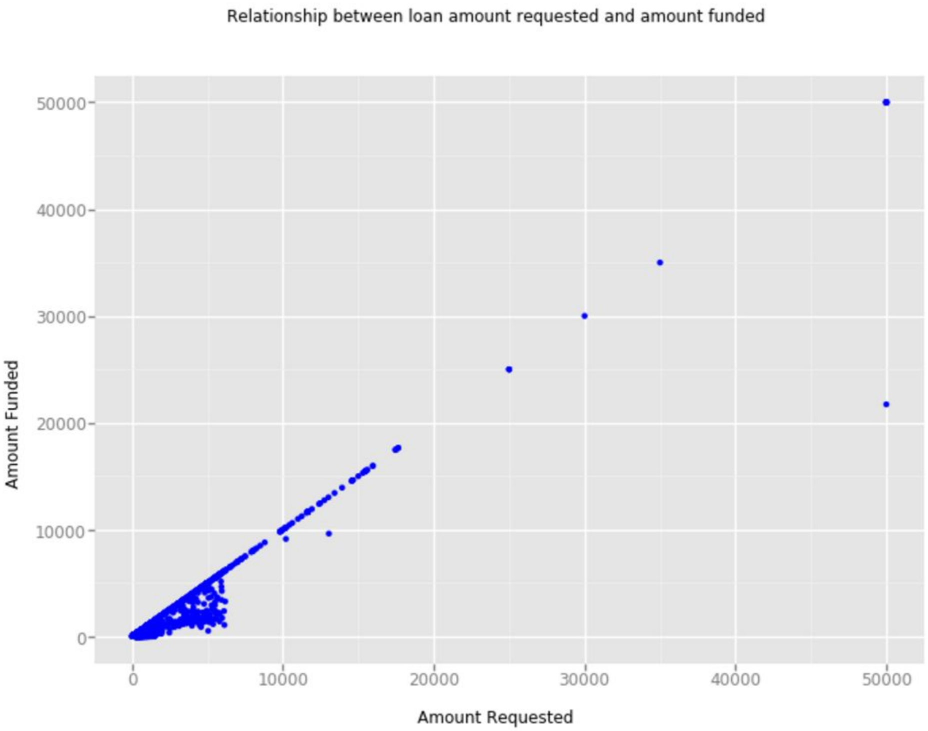
X = $ spent on advertising
Y= Sales

Based on the graph and positive correlation, you’d be tempted to say $ spent on advertising caused an increase in sales. **But hang on** - it’s also possible that an increase in sales (and thus, profit) would lead to an increase in $ spent on advertising! ***Correlation between x and y does not mean x causes y; it could mean that y causes x!***

Exploratory Analysis

Correlation

Kiva example: Correlation does not equal causation



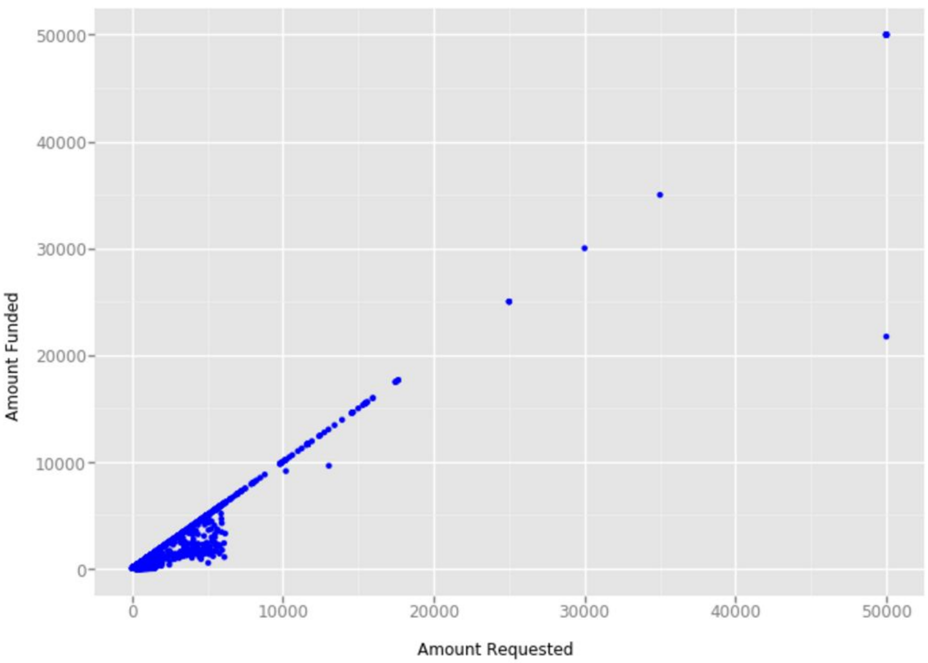
Correlation: 0.96

If you wanted to request a loan through Kiva, and were presented with this graph **only**, you might conclude that it is a good idea to request $1 million dollars.

Exploratory Analysis → Correlation

Kiva example: Correlation does not equal causation

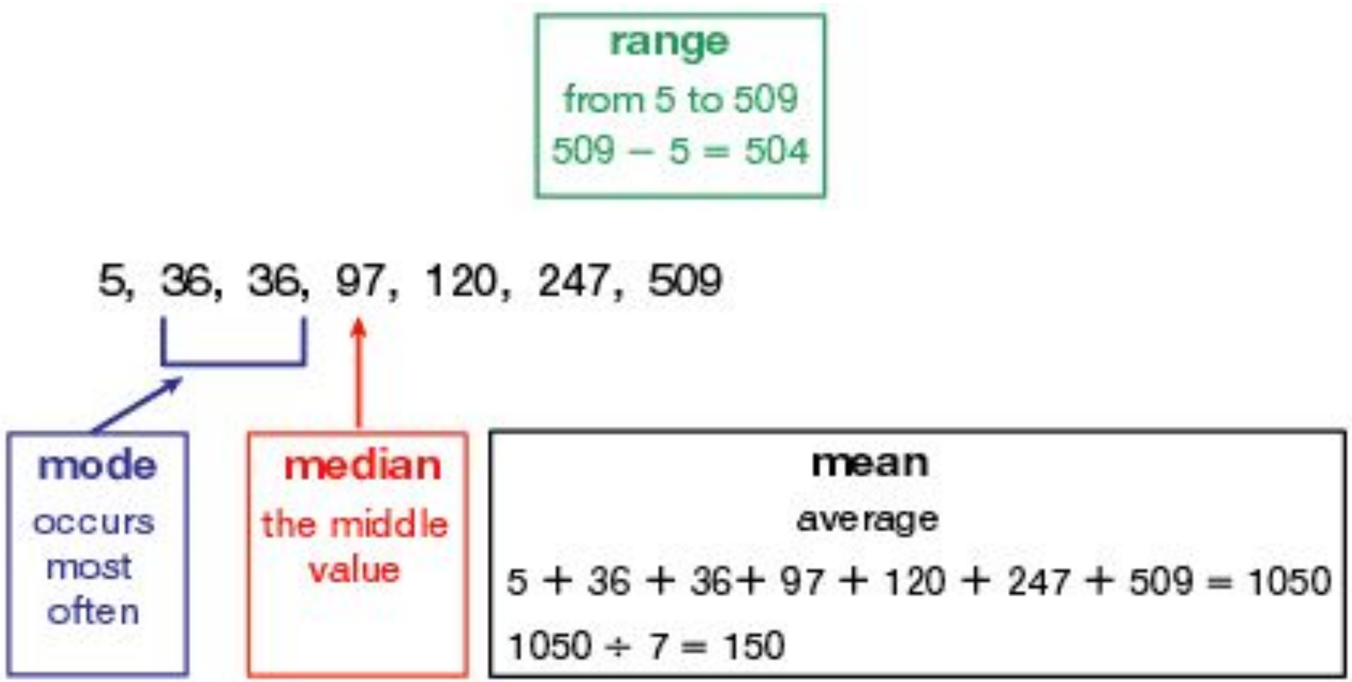
Relationship between loan amount requested and amount funded



But common sense tells us that this conclusion doesn’t make a lot of sense. Just because you request a lot doesn’t mean you will be funded a lot!

Conclusions can be invalid even when data is valid!

Mean, median, frequency are useful summary statistics that let you know what is in your data.



range

from 5 to 509

509 – 5 = 504

5, 36, 36, 97, 120, 247, 509

mode

occurs most often

median

the middle value

mean

average

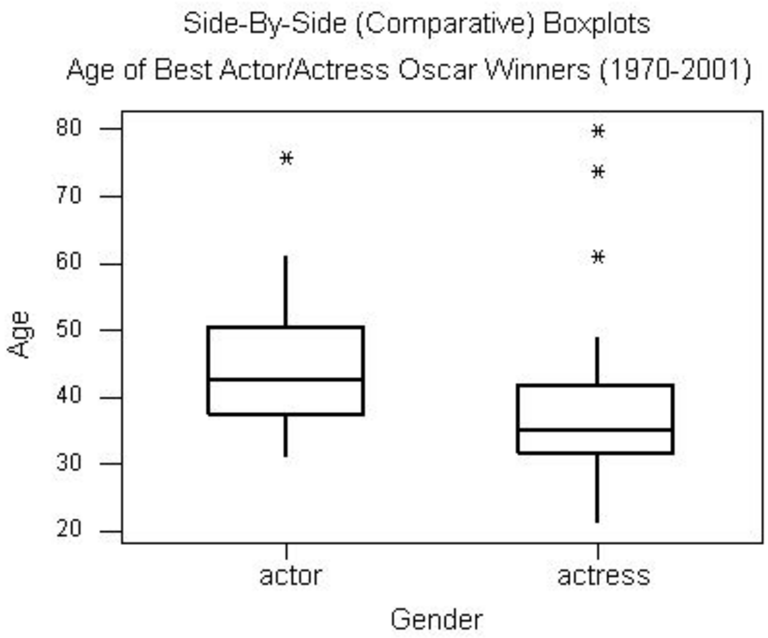
5 + 36 + 36+ 97 + 120 + 247 + 509 = 1050

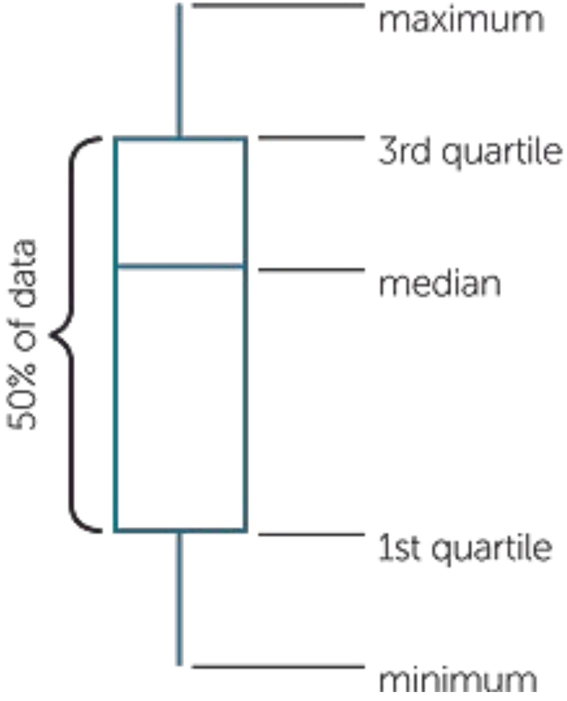
1050 ÷ 7 = 150

Exploratory Analysis

Boxplots

Boxplots are a useful visual depiction of certain summary statistics.

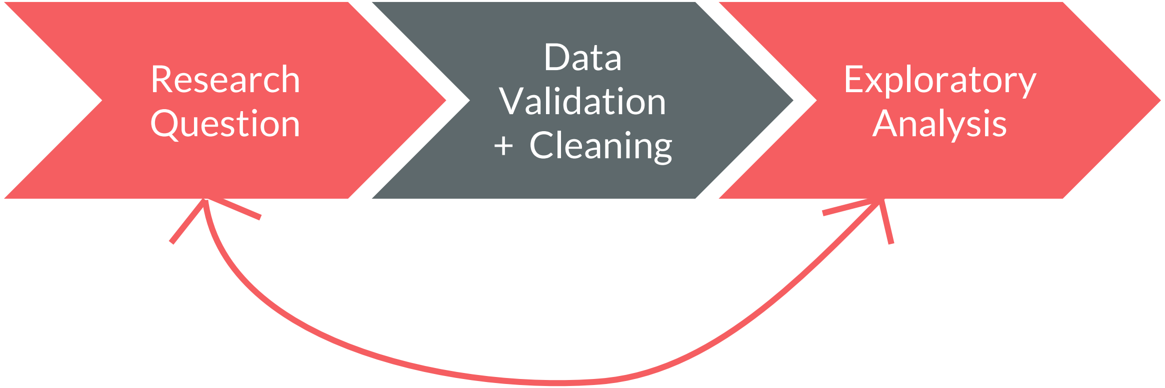




Source: University of Florida,
http://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/boxplot/

Forming a research question

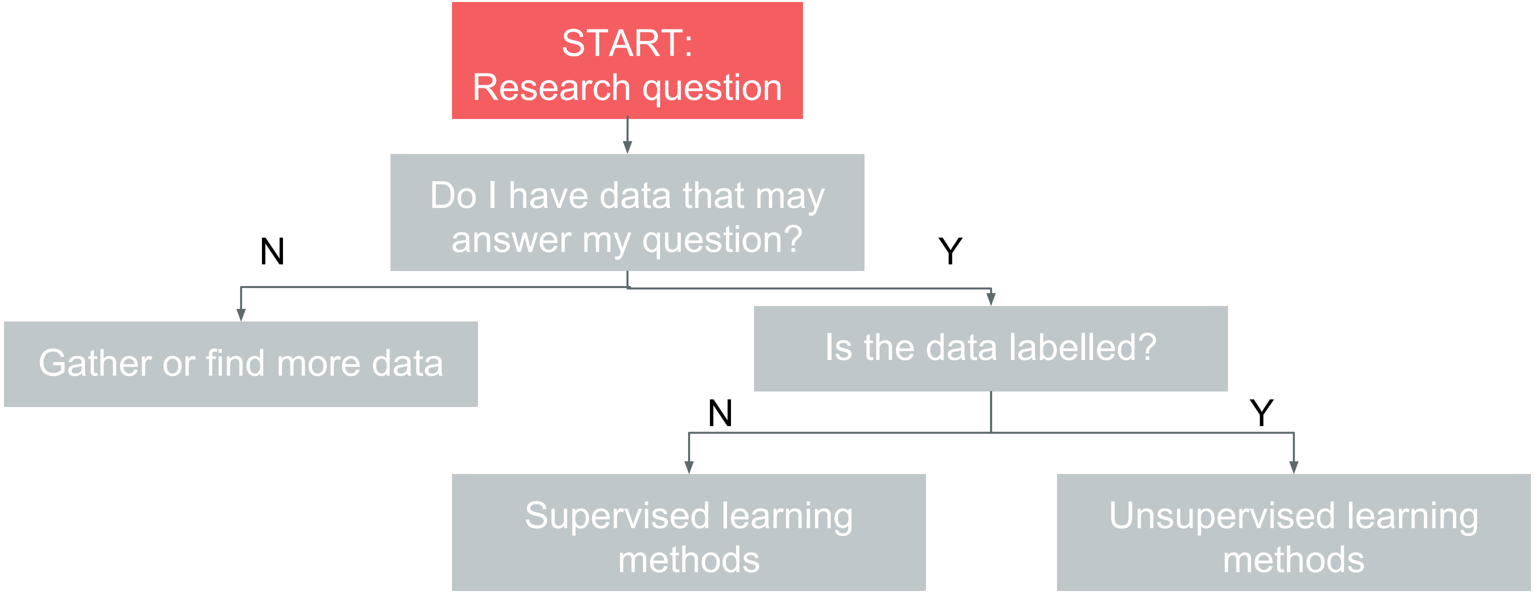
Recall: We may have a question in mind **before** we look at the data, but our exploration of the data often develops or refines our research question.



What comes first, the chicken or the egg?

How do you define the research question?

- We ask a question we expect data to answer. *What comes first, the data or the question?*



START:
Research question

Do I have data that may answer my question?

N

Gather or find more data

Y

Is the data labelled?

N

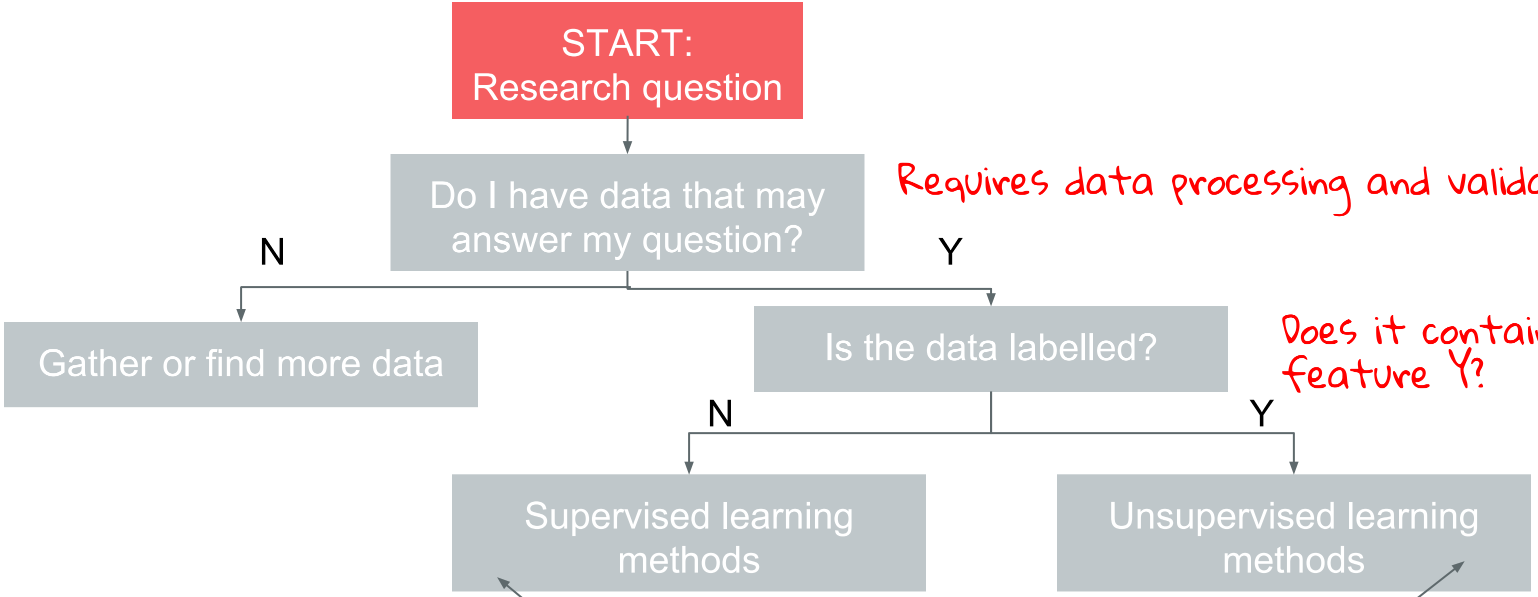
Supervised learning methods

Y

Unsupervised learning methods

How do you define the research question?

- We ask a question we expect data to answer. *What comes first, the data or the question?*



START:
Research question

Do I have data that may answer my question?

N

Y

Gather or find more data

Is the data labelled?

N

Y

Supervised learning methods

Unsupervised learning methods

Requires data processing and validation

Does it contain the outcome feature Y?

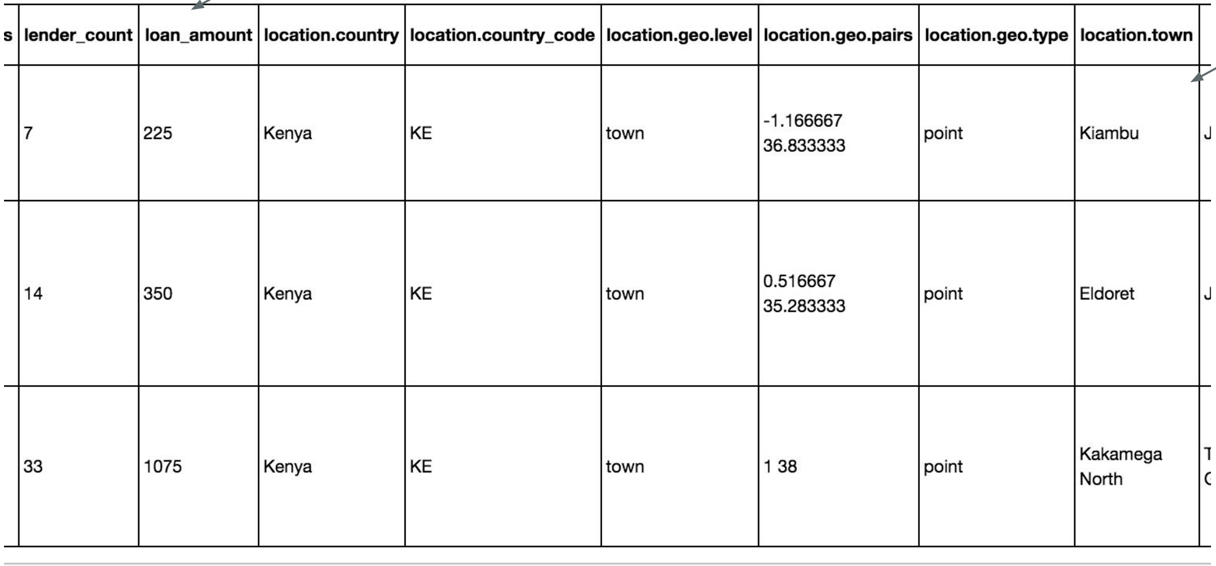
We'll cover both supervised and unsupervised methods in this course!

Research Question

Given the KIVA data below, we may find a few questions interesting.

Loan amount requested by a Kiva borrower in Kenya

Town Kiva borrower resides in.



s	lender_count	loan_amount	location.country	location.country_code	location.geo.level	location.geo.pairs	location.geo.type	location.town	
	7	225	Kenya	KE	town	-1.166667 36.833333	point	Kiambu	J
	14	350	Kenya	KE	town	0.516667 35.283333	point	Eldoret	J
	33	1075	Kenya	KE	town	1 38	point	Kakamega North	T C

One possible research question we might be interested in exploring is: Does the loan amount requested vary by town?

Research Question

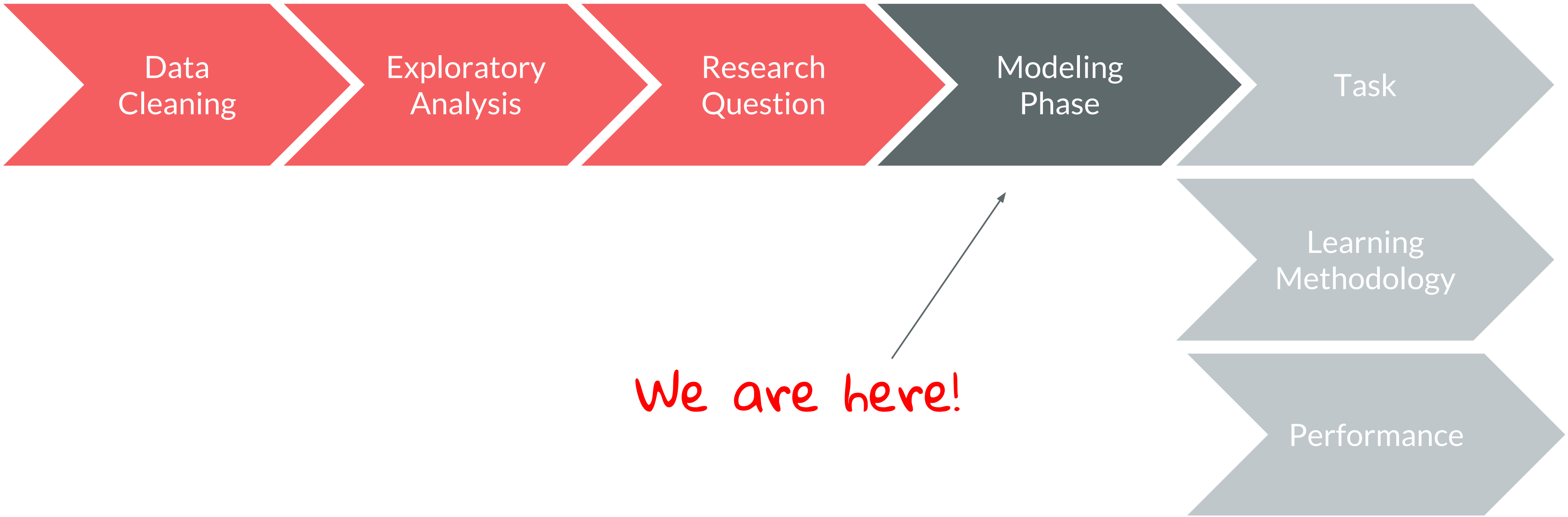
How does loan amount requested vary by town?

This is a reasonable research question, because we would expect the amount to vary because the cost of materials and services varies from region to region.

For example, we would expect the cost of living in a rural area to be cheaper than an urban city.

Looking ahead: Modeling

Now we have our research question, we are able to start modeling!



Data Cleaning

Exploratory Analysis

Research Question

Modeling Phase

Task

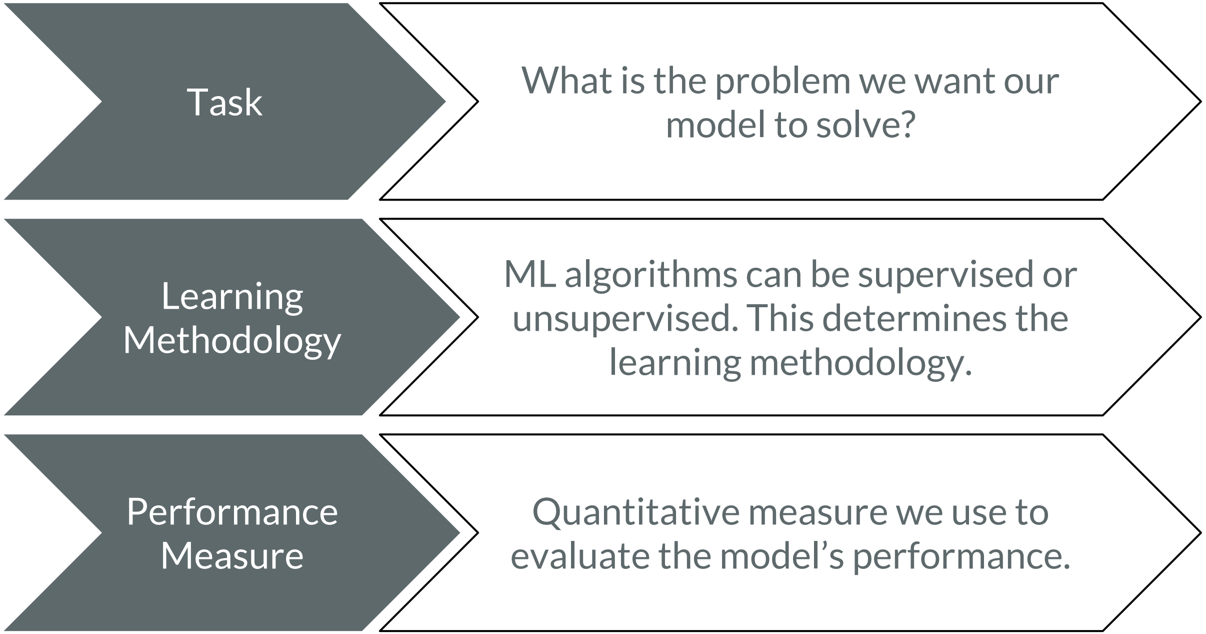
Learning Methodology

Performance

We are here!

Modeling Phase

All models have 3 key components: a task, a performance measure and a learning methodology.



Task

What is the problem we want our model to solve?

Learning Methodology

ML algorithms can be supervised or unsupervised. This determines the learning methodology.

Performance Measure

Quantitative measure we use to evaluate the model's performance.

Source: Deep Learning Book - Chapter 5: Introduction to Machine Learning

We will go over the machine learning task and learning methodology in the next lesson.

You covered this today:

- ✓ What is machine learning?
- ✓ How do you define a research question?
- ✓ What are observations?
- ✓ What are features?
- ✓ What are outcome variables?
- ✓ Introduction to KIVA data