PREDICTIVE ANALYTICS FOUNDATIONS

LECTURE 9 - 10

HOUSEKEEPING

- Questions:
  - From others asking:
    - Break problems into individual smaller tasks.
    - Print "what happened" (e.g. variable values) after each line if needed. (Read something, print it; change a variable, print it; calculate something, print it…)
- Today:
  - Datasheets and data shapes.
  - Pandas and manipulating data.
  - Simple concepts, moderately tricky code, important for the future.
  - Slicing and dicing data.

DATASHEETS

MANIPULATING DATA

- To this point, we can read data from a file and do a lot to manipulate it.
- Each time we use some data, we need to "organize" it to some degree:
  - What data do we have?
  - Which parts come where?
- It might be easier to use lots of data if it was in a standardized format.
- We can have assumptions that we rely on to make things easier.
- When we get to machine learning, we use this old, organized, data as our main source.

LARGE AMOUNTS OF DATA

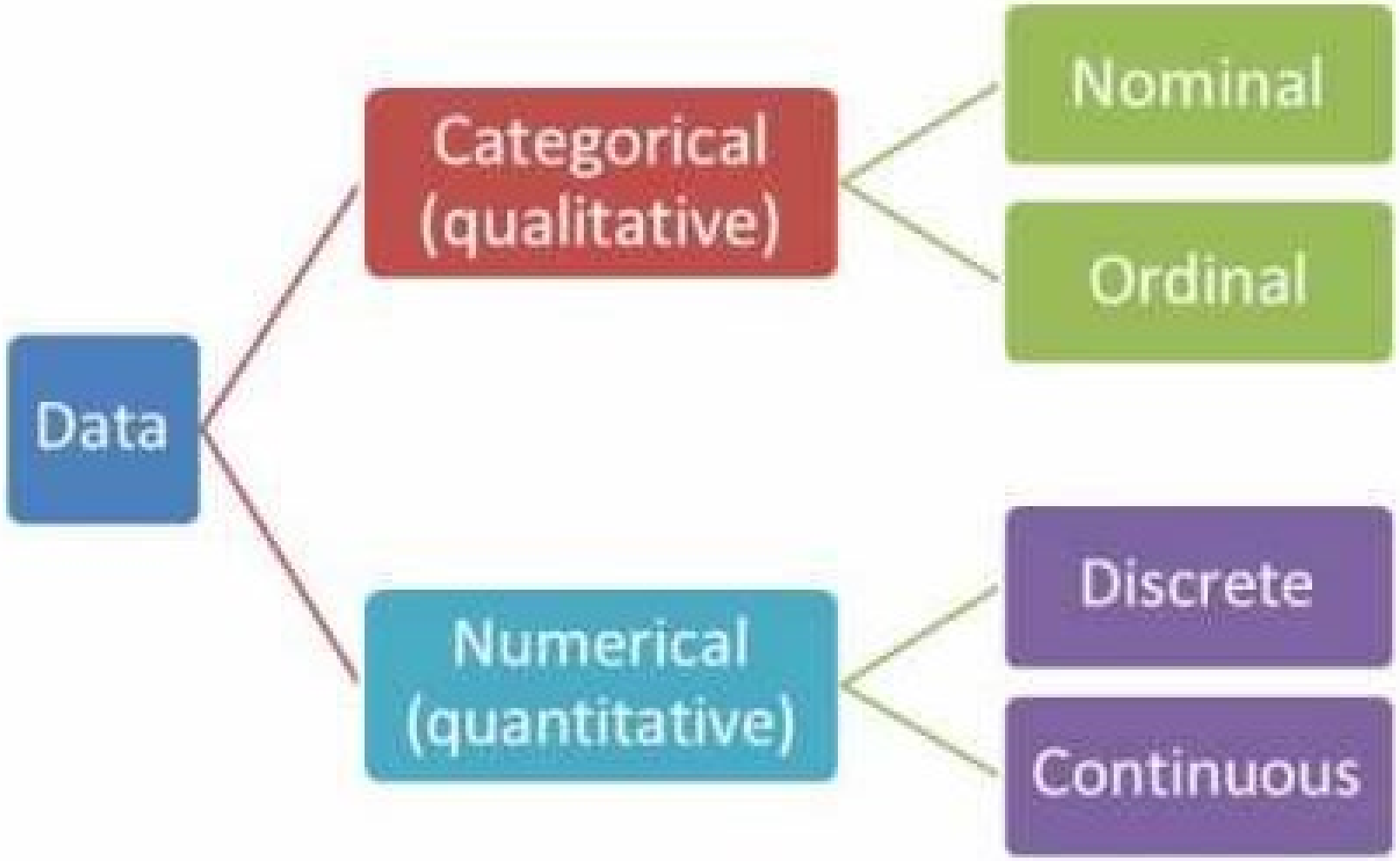
- This system of using old examples to make accurate predictions fits well with how data is stored in databases.
  - We'll look at this in more detail over the next two times, for now we just want a basic understanding of what's in a database.
- Databases are made of tables that each look like a spreadsheet.
  - Each table represents an entity – one "thing" that we track.
  - Each column represents an attribute - one value that we store for this entity.
  - Each row represents an instance – one example of that entity.
- So, each table is effectively a list of items, and what we know about them.

Loans

LoanID	Date	Amount	InterestRate	Term	Type	CustomerID	FirstName	LastName	Paid
L0001	1/15/2018	$475,000	6.20%	15	M	C0004	Wendy	Solomon	☐
L0002	1/23/2018	$35,000	7.20%	5	C	C0004	Wendy	Solomon	☑
L0003	1/25/2018	$10,000	5.50%	3	C	C0005	Alex	Rey	☐
L0004	1/31/2018	$12,000	9.50%	10	O	C0004	Wendy	Solomon	☑
L0005	2/8/2018	$525,000	6.50%	30	M	C0006	Ted	Myerson	☑
L0006	2/12/2018	$10,500	7.50%	5	O	C0007	Lori	Sangastiano	☑
L0007	2/15/2018	$35,000	6.50%	5	O	C0008	John	Smith	☐
L0008	2/20/2018	$250,000	8.80%	30	M	C0008	John	Smith	☐
L0009	2/21/2018	$5,000	10.00%	3	O	C0008	John	Smith	☐
L0010	2/28/2018	$200,000	7.00%	15	M	C0001	Eileen	Faulkner	☐
L0011	3/1/2018	$25,000	10.00%	3	C	C0002	Scott	Wit	☐
L0012	3/1/2018	$20,000	9.50%	5	O	C0005	Alex	Rey	☑
L0013	3/3/2018	$56,000	7.50%	5	C	C0009	David	Powell	☑
L0014	3/10/2018	$129,000	8.50%	15	M	C0010	Matt	Hirsch	☐
L0015	3/11/2018	$200,000	7.25%	15	M	C0003	Benjamin	Grauer	☐
L0016	3/21/2018	$150,000	7.50%	15	M	C0001	Eileen	Faulkner	☐
L0017	3/22/2018	$100,000	7.00%	30	M	C0001	Eileen	Faulkner	☑
L0018	3/31/2018	$15,000	6.50%	3	O	C0003	Benjamin	Grauer	☑
L0019	4/1/2018	$10,000	8.00%	5	C	C0002	Scott	Wit	☐
L0020	4/15/2018	$25,000	8.50%	4	C	C0003	Benjamin	Grauer	☐
L0021	4/18/2018	$41,000	9.90%	4	C	C0008	John	Smith	☐
L0022	4/22/2018	$350,000	7.50%	15	M	C0010	Matt	Hirsch	☑
L0023	5/1/2018	$150,000	6.00%	15	M	C0003	Benjamin	Grauer	☐
L0024	5/3/2018	$350,000	8.20%	30	M	C0004	Wendy	Solomon	☑
L0025	5/8/2018	$275,000	9.20%	15	M	C0007	Lori	Sangastiano	☐
(New)									☐

- Table – loans.
- Row – instance, one loan.
- Column – thing we track about the loan.
  - Including if it was paid.
  - We could know the inputs (other stuff) before giving a loan, then use these old ones to predict if it was paid.

Kinds of data



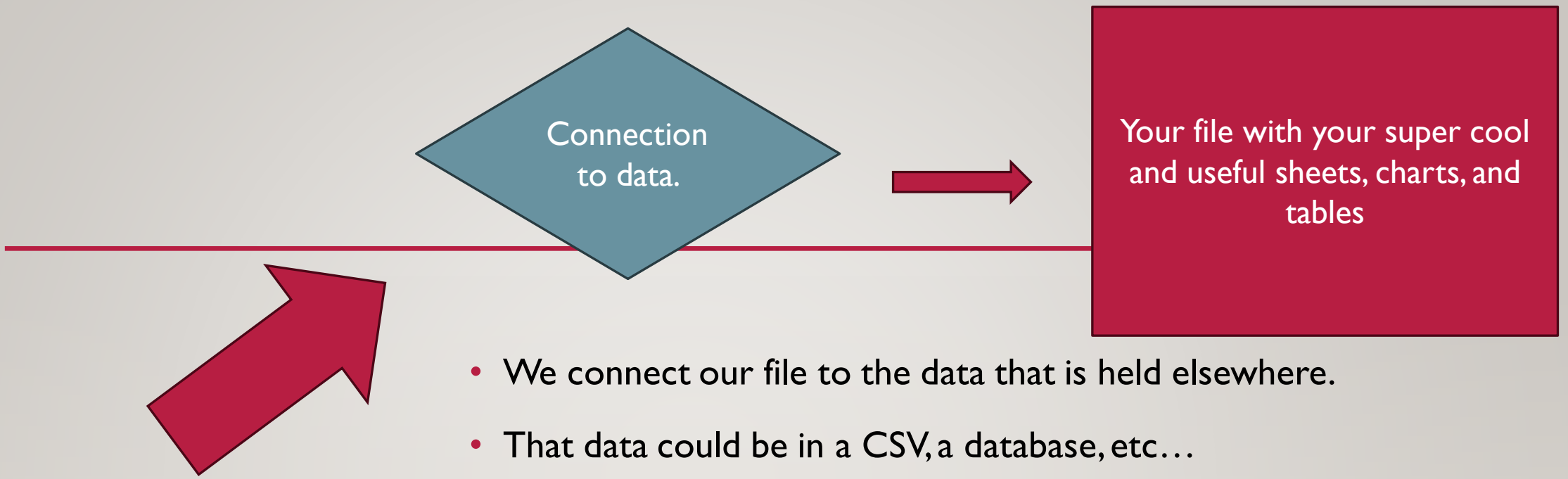
- Named categories
- Categories with an implied order
- Only particular numbers
- Any numeric value

DATASHEETS

- This format (table, row, column) is called a datasheet.
- Most databases that exist in the world look just like this.
  - We keep one table for each "item" – sale, product, customer, registration.
  - Each column is one value that we track for that thing.
  - Each row is one instance, or one specific example of that thing.
- We generally load and manipulate data that looks like this.
  - Safe assumptions for the "shape" of the data.
  - Standardized tools and commands to manipulate the data.

DATA SEPARATION

- Continuing with the idea of using a database as a source for our data…
- Databases are normally large and centralized, used by many people.
  - We can't have all our data in a spreadsheet because we don't own it.
  - Other people need to be looking at and updating that data as well.
- Solution – connect to the data that is held remotely.
  - Data stays in a database where it is, or it stays in a file like it will for us.
  - We can grab it and use it remotely in Excel, on the web, in Tableau, in another system…
- We assume that the data is structured in a datasheet format, allowing us to use those assumptions to handle any data interchangeably.



- We connect our file to the data that is held elsewhere.
- That data could be in a CSV, a database, etc…
- Our stuff is built "on" the data, we don't have it we just reach out and "grab" it from elsewhere.
  - We can use it just like we have been.
  - It isn't actually located in our workbook.
  - That data can change on its own, we update from it.
- This is how most "data tools" work in real usage.

PANDAS

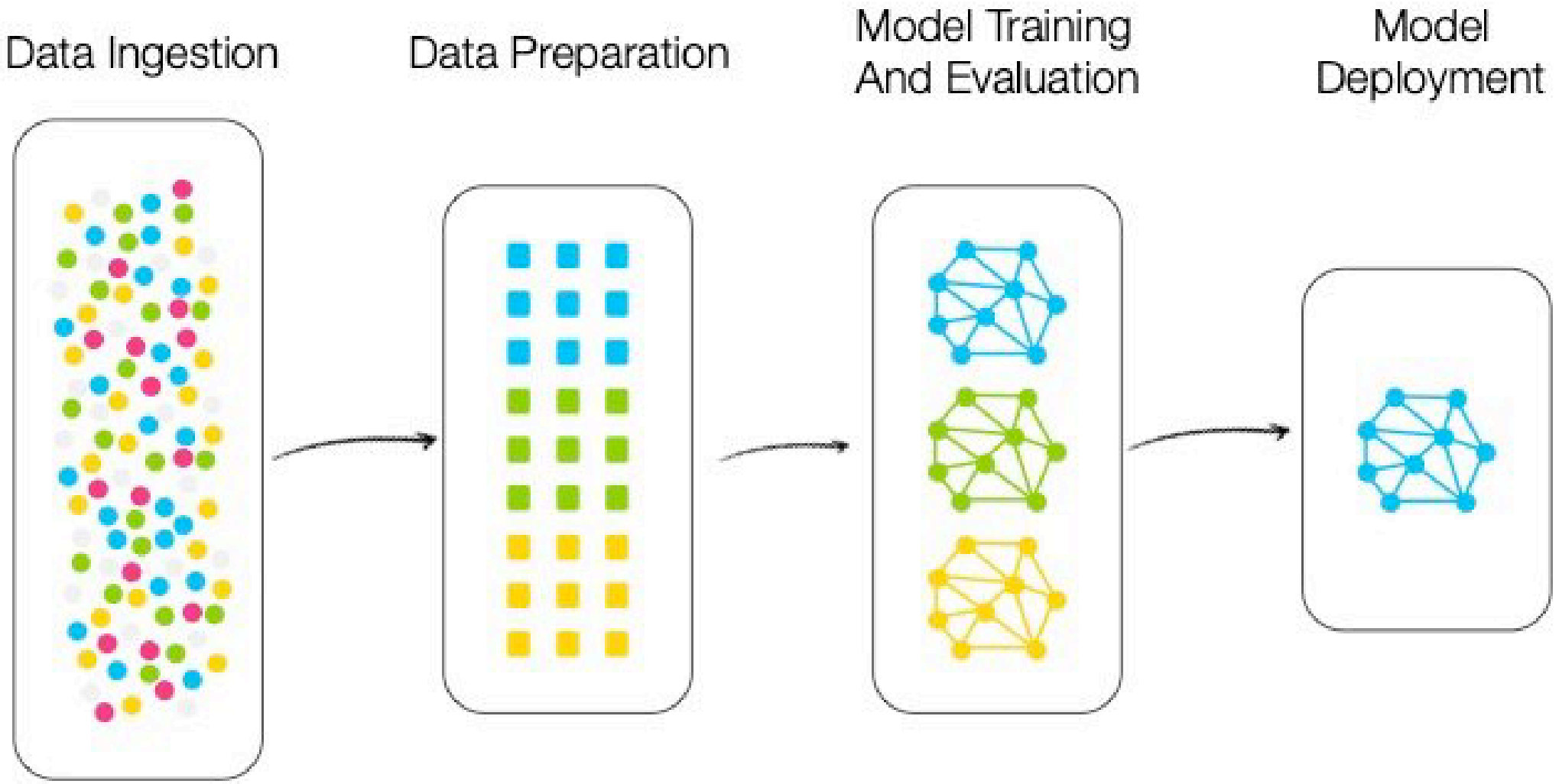
- Pandas is a library that provides a Dataframe – roughly a Python spreadsheet.
- We can manipulate our data largely as we would think about it in Excel, but through code.
- Maps easily to databases, if we are getting data from them.

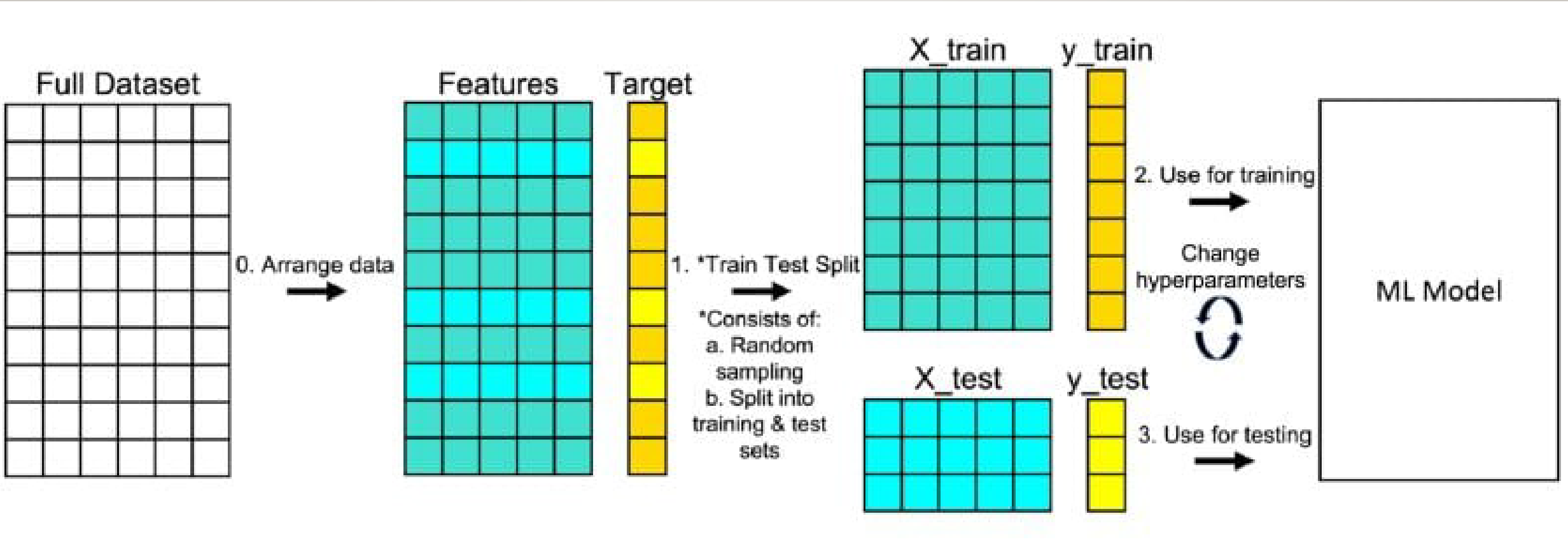
SLICE AND DICE THE DATA

- One useful thing about dataframes is that they make it easy to "slice" data.
- In ML, we commonly need to slice "vertically", or separate a column.
  - E.g. separate the column that we want to predict from the inputs.
- We can also slice "horizontally", or create groupings or samples.
  - Horizontal slices can allow us to segment the data into subsets.
  - We can compare groups against each other.
  - If we have a large amount of data, or want to do some stats calculations, we can generate samples from our "population" (population = everything, in stats).

DATA AND ANALYSIS

- Much of common data analysis that we want to do is based on Aggregation.
- Aggregation is the "totaling" of some value, here it's calculating sums, avg, count, etc…
- We can combine these aggregations with grouping and slicing of the data.
  - We can calculate any aggregate we want – min, max, average, sum, etc….
  - We can group or split our data to get the result for any subgroup.
- We can compare and contrast these groups for a powerful analysis tool.
- This works directly with the datasheet formatted data that we get from a DB.


Data Ingestion
Data Preparation
Model Training
And Evaluation
Model
Deployment


Full Dataset
0. Arrange data
Features
Target
1. *Train Test Split
*Consists of:
a. Random sampling
b. Split into training & test sets
X_train
y_train
X_test
y_test
2. Use for training
Change hyperparameters
3. Use for testing
ML Model