Chapter 2

The Recipe: The Fundamental Process of Data Scientists

Data science is the art and science of learning from data. It combines observation, creativity, and careful reasoning to help us understand how the world works — and sometimes, how it doesn't. Whether predicting the weather, diagnosing an illness, or guiding a spaceship, data science builds bridges between the messy world of real-life information and the tidy world of ideas.

At the heart of data science is the concept of a model. A model is a simplified picture of how we think something works. It's a set of instructions that says: "If this is the input, here's what the output should be." Models can take many forms — a diagram, a simulation, a story — but in data science, we usually work with a specific kind of model: a parametric model.

A parametric model is a mathematical description a process using a fixed number of adjustable parameters. For those who prefer analogies, you can think of it like a machine with knobs. Each knob controls one part of how the machine behaves. By turning the knobs just right, we can make the machine behave more like the real-world system we're trying to understand.

But how do we know where to set the knobs? That's where data comes in.

Data is the record of what actually happened — the inputs we gave a system and the results we got in return. When we compare what the model predicts to what actually occurred, we can start to see how well the model matches reality. If the match is poor, we adjust the parameters. If the match improves, we're on the right track. This back-and-forth — adjusting the model to better fit the data — is what we mean by fitting a model to data.

In this chapter, we use this whimsical idea—striking a target with a spud gun—as our running example. Though frivolous, it offers a real-world challenge: given what we can control and measure, how can we consistently hit a distant target? That question leads us naturally into the heart of data science, parametric modeling, and model fitting. The chapter gives a step-wise procedure that a data scientist follows with the purpose of addressing a specified issue.

The book explores historical scientific achievements through the lens of a data scientist. It demonstrates how past scientific pioneers created and evolved the data science procedure to solve problems of interest. Just how robust is the procedure? It applies to our frivolous spud gun as well as to ChatGPT.

2.1 The Spud Gun: From Backyard Curiosity to Competitive Engineering

In the world of homemade inventions, few devices are as entertaining—or unexpectedly educational—as the potato cannon. Built from materials like PVC pipe, starter fluid, and duct tape, these devices can launch a humble potato several hundred feet through the air. What started as backyard fun has grown into a quirky but passionate subculture of builders, tinkerers, and amateur engineers.

The precise origins of the potato cannon are hard to trace, but its popularity surged in the late 20th century alongside the spread of the internet, where instructions, photos, and videos helped enthusiasts share their designs and successes. A key innovation that made these devices accessible was the availability of PVC, invented in 1926 by Waldo Semon. It offered the right balance of strength, affordability, and ease of construction for ambitious weekend projects.

In 2009, Wired ran an article on potato cannon inventor Alan Nelsen, titled "The Grandfather of the Potato Cannon". Nelsen is often credited with helping to popularize and advance the spud gun through experimentation and public exhibitions. In that article, Wired quotes him saying:

"It's not about the potato. It's about seeing how far you can push something you made with your hands." (Source: Wired Magazine, May 2009)

Out of this playful curiosity grew competitions—events where participants gather to test the accuracy, distance, and design of their potato cannons. Some focus on raw power, launching potatoes over football fields. Others emphasize precision, such as striking a target hundreds of feet away. These gatherings highlight the creative, scientific, and often humorous spirit behind what might first appear to be a ridiculous invention.

The Pneumatic Spud Gun: Engineering with Air Pressure

A pneumatic spud gun relies on compressed air—not explosive gases—to propel a potato through a barrel. This makes it more predictable, reusable, and easier to analyze and model, which is exactly why it is ideal for a case study in data science.

A typical pneumatic spud gun includes the following elements:

1. **Pressure Chamber (Air Tank):**
    - Stores compressed air before firing.
    - Typically made from thick-walled PVC or metal pipe rated for high pressure (e.g., 2-inch diameter, Schedule 40 PVC).
    - A Schrader (tire) valve allows air to be added from a pump or compressor.
    - Must be pressure-tested and properly sealed for safety.
2. **Barrel:**
    - A long, smooth pipe slightly narrower than the potato.
    - The potato is rammed in from the muzzle end to form an air-tight seal.
    - Commonly made from 1.5- to 2-inch diameter PVC, 3–5 feet in length.
    - Connected to the pressure chamber via a reducing adapter.
3. **Valve System (Trigger):**

- Separates the pressure chamber from the barrel.
- A fast-opening valve (e.g., sprinkler valve, ball valve) ensures rapid release of pressure.
- Advanced designs may use solenoid or diaphragm valves for quicker actuation.

4. **Air Supply:**
    - A bicycle pump, air compressor, or CO_2 cartridge is used to charge the chamber.
    - Operating pressure typically ranges from 40 to 120 psi.
    - Pressure must never exceed the rated limit of the chamber materials.
5. **Frame and Accessories (Optional):**
    - May include a shoulder stock, grip handles, or stabilizing legs.
    - A pressure gauge is often installed for monitoring.

The potato is rammed into the barrel, sealing it air-tight. The pressure chamber is then filled with compressed air, separated from the barrel by the closed valve. When the valve is opened rapidly, the compressed air surges into the barrel, pushing the potato forward and launching it at high speed.

Unlike combustion-based spud guns, which rely on variable fuel-air mixtures, pneumatic launchers operate under well-controlled, measurable conditions. The performance of the launcher depends on factors such as:

- Air pressure
- Potato mass
- Launch angles (azimuth and elevation)
- Barrel length
- Valve opening speed

Knowing these inputs as the causes in a cause and effect relation makes pneumatic spud guns ideal for experimentation, measurement, and mathematical modeling.

2.2 The Data Science Process: From Questions to Models

The process of data science transforms open-ended questions into quantitative models that help us understand and control systems. The procedure typically follows six steps:

1. **Define the problem.**
2. **Propose an input-output parametric model of the system.**
3. **Identify the required data.**
4. **Collect and organize data as inputs and outputs.**
5. **Define a metric that quantifies the error between model predictions and observed outputs.**
6. **Apply an optimization routine to adjust the parameters and minimize the error.**
7. **Validate results against additional data.**

We will now apply this framework to the problem of accurately launching a potato toward a distant target.

Step 1: Define the Problem

Goal: Launch a potato such that it hits a designated target located at a known horizontal distance.

While many factors influence the potato's flight (e.g., wind, spin, drag), we focus primarily on predicting and controlling the *horizontal distance traveled* (range). We assume the azimuth angle has been properly aligned with the target, so the main variable of interest is the **elevation angle** of launch.

Step 2: Propose an Input-Output Parametric Model

We propose a physics-based model for projectile motion, focusing on predicting the *horizontal distance traveled* (range) by the potato. Since we are aiming at a fixed target, we assume the azimuth angle is already set. Thus, the critical variable in aiming is the *elevation angle*, which strongly influences the trajectory.

The inputs (also called *features*) of the model include:

- Initial pressure in the air chamber
- Elevation angle of the launcher
- Mass of the potato

The output is:

- Horizontal distance traveled (range)

We begin by considering a simplified scenario from classical physics: a projectile launched in a vacuum from the ground with no air resistance. In this idealized case, the horizontal distance R is given by the well-known formula:

$$R = \frac{v_0^2 \sin(2\theta)}{g}$$

Here, v_0 is the launch velocity, θ is the elevation angle, and g is the acceleration due to gravity. While useful conceptually, this model is too simplistic for a real potato launcher, which involves factors such as air resistance, imperfect valve dynamics, and nonuniform shapes.

To better match experimental behavior while retaining a connection to physical intuition, we now propose a more realistic model based on pressure-driven acceleration:

$$R = \alpha \left(\frac{PA}{mg}\right)^{\beta} \sin(2\theta)$$

Variables (measured for each trial):

- P: Pressure in the air chamber (in Pascals)
- m: Mass of the potato (in kilograms)
- θ: Elevation angle of the launcher (in radians)

Determined Parameters (known physical constants):

- A: Cross-sectional area of the barrel. For a barrel with inner diameter 4.0 cm,

$$A = \pi \left(\frac{0.04}{2} \right)^2 \approx 1.26 \times 10^{-3} \text{ m}^2$$

- g: Acceleration due to gravity, $g = 9.8\,\text{m/s}^2$

Undetermined Parameters (to be estimated from data):

- α: A scaling constant to account for energy losses and system inefficiencies
- β: A nonlinear exponent that adjusts how launch force affects range

The model expresses the horizontal range R as a nonlinear function of the input variables. The term $\frac{PA}{mg}$ is a dimensionless quantity representing the ratio of launch force to projectile weight. The exponent β introduces flexibility to account for system nonidealities, and the $\sin(2\theta)$ factor preserves the familiar role of θ from ideal projectile theory.

The ultimate goal is to fit the parameters α and β so that the model best matches observed data. This will be accomplished in later steps using a suitable error metric. This model strikes a balance between simplicity and physical realism. It allows us to systematically vary input parameters and predict their effects on distance, which is essential for parameter estimation.

Step 3: Identify the Required Data

To tune our model, we must collect data on:

- Launch pressure (P)
- Elevation angle (θ)
- Potato mass (m)
- Resulting horizontal distance (R)

Each launch provides a set of input conditions and an output distance.

Step 4: Collect and Organize Data

We perform several test launches, recording the pressure, angle, mass, and measured range for each. Data can be tabulated as follows:

Trial	Pressure (Pa)	Angle (degrees)	Mass (g)	Distance (m)
1	413686	45	250	38.9
2	551580	40	255	45.3
3	689475	35	245	53.5

These data become the foundation for model fitting. This dataset is called the training dataset used in step 6.

In addition to the training dataset, we create an additional dataset, the validation dataset to be used in step 7.

Step 5: Define an Error Metric

To quantify how well our model matches reality, we define an error metric — typically the *mean squared error* (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(R_i - \hat{R}_i(\alpha,\beta)\right)^2$$

Here, R_i is the actual distance observed in trial i, and $\hat{R}_i$ is the predicted distance from the model, using the current values of the parameters α and β.

What this means in plain terms: The symbol $\sum$ (the Greek letter sigma) represents a *summation*, or "add up" instruction. In this case, it tells us to compute the squared difference between the observed and predicted distances for each trial i, then add all those squared differences together, from $i = 1$ to $i = n$, where n is the total number of trials. We then divide this total by n to compute the *mean* of the squared differences — the average squared error.

In our example, there are three observations, so $n = 3$. Each observation is labeled with an index: for instance, $i = 2$ refers to the second observation. For each i, we calculate a squared difference — this represents the square of the model's error for that specific observation. The mean squared error is then computed by summing all three squared errors and dividing by 3. We'll carry out this calculation in the next step.

Why this matters: The MSE directly measures how well our chosen parameters explain the observed behavior. Large errors suggest poor parameter choices or an inadequate model structure. Small errors indicate a good fit and more reliable predictions.

Step 6: Apply an Optimization Routine to Minimize the Error

With a model structure in place and a training dataset of inputs and measured outputs, the next step is to adjust the parameters of the model to reduce the prediction error. In our case, the parameters to optimize are α and β in the equation:

$$R = \alpha \cdot \left(\frac{P \times A}{m \times g}\right)^{\beta} \cdot \sin(2\theta)$$

The prediction error is the difference between the observed horizontal distance R_{obs} and the predicted distance R_{pred}. A standard way to quantify this is the **mean squared error (MSE)**:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(R_{\text{obs},i} - R_{\text{pred},i}\right)^2$$

We start with initial parameter guesses: $\alpha = 0.7$, $\beta = 0.6$. Using these values, we calculate the predicted ranges and corresponding errors:

Trial (i)	$R_{\text{obs},i}$ (m)	$R_{\text{pred},i}$ (m)	$(R_{\text{obs},i} - R_{\text{pred},i})^2$
1	38.9	17.44	460.51
2	45.3	20.17	631.50
3	53.5	22.54	958.64

$$\text{MSE} = \frac{460.51 + 631.50 + 958.64}{3} \approx 683.55$$

After one round of parameter tuning, we try updated values: $\alpha = 0.75$, $\beta = 0.7$. With these improved values, the model predictions more closely match the observations:

Trial (i)	$R_{\text{obs},i}$ (m)	$R_{\text{pred},i}$ (m)	$(R_{\text{obs},i} - R_{\text{pred},i})^2$
1	38.9	31.93	48.51
2	45.3	37.93	54.23
3	53.5	43.51	99.61

We see that the MSE has been reduced from 683.55 to 67.45 after adjusting the parameters. This indicates that the updated parameters yield predictions that better match the experimental data. By continuing this optimization process—possibly with more data and more sophisticated optimization algorithms—we can achieve an increasingly accurate model that generalizes well to new conditions and supports practical targeting decisions for the launcher.

One method for determining how to adjust the parameters α and β is known as **gradient descent**. This technique calculates how changes in each parameter affect the error and moves them in the direction that reduces the error most rapidly. After several iterations of the gradient descent method, we might arrive at the parameters, $\alpha = 0.8$, $\beta = 0.72$. With these parameter values, the model predictions are very near their actual distances.

$$\text{MSE} = \frac{48.51 + 54.23 + 99.61}{3} \approx 67.45$$

Trial (i)	$R_{\text{obs},i}$ (m)	$R_{\text{pred},i}$ (m)	$(R_{\text{obs},i} - R_{\text{pred},i})^2$
1	38.9	37.92	0.97
2	45.3	45.29	0.0002
3	53.5	52.22	1.63

$$\text{MSE} = \frac{0.97 + 0.0002 + 1.63}{3} \approx 0.86$$

Step 7. Validate results against additional data.

The model has been fit to the data and we are satisfied with the results; it produces good outcomes when the inputs come from the dataset used to fit the data. A question remains, how well are the model's predictions using inputs that are not in the training set. To answer this question, we apply the model to the validation set and compute the mean square error. If the error is within our tolerance, we deploy the model. Otherwise, improvements are necessary.

2.3 Final Thoughts

This example illustrates how data science provides a structured, repeatable approach to solving even quirky problems like launching root vegetables with precision. The same steps apply to problems in medicine, finance, and engineering — whenever we want to use data to understand and influence the world. The following chapters describe the historical development and application of the data science procedure over more than a two millenium time frame. This has been an ongoing human collaboration across continents and time. Recently, AI has joined humans in this collaboration.

2.4 Summary Poem: The Spud Gun Modeler's Ode

In data's dance and number's sway,

We seek to learn the world's array.
Through models trimmed and finely tuned,
We aim to strike the truth, attuned.

A model's just a thoughtful guess,
A framework built to coarsely dress
Reality's chaotic threads—
A picture drawn in reason's treads.

With pressure, pipe, and spud in hand,
We launch our quest across the land.
A cannon built from backyard dreams
Becomes the heart of data schemes.

The spud gun, once a toy of glee,
Now teaches physics carefully.
With pressure known and mass in tow,
How far the starchy round will go?

We measure, fit, and test the arc,
Adjusting knobs to hit the mark.
From noisy plots, we draw a line,
And tune our math until it's fine.

Seven steps to guide our modeling path:
From problem posed to aftermath.
Define, propose, and gather well,
Then error's voice begins to tell.

We minimize with care and might,
Until prediction feels just right.
Thus data speaks, and truth is caught,
In models forged from human thought.

So wield your tools with heart and brain,
Let numbers sing and graphs explain.
For in each chart and fitted curve,
Lives insight, ready to observe.