Data Analytics Foundations

Module 2: Using spreadsheets for data analytics

DeepLearning.AI

Using spreadsheets for data analytics

Module 2 introduction

DeepLearning.AI

Module 2 outline

- **How spreadsheets organize data**
  - Importing data
  - Setting up spreadsheets
- **Spreadsheet fundamentals**
  - Sort and filter data
  - Write formulas
  - Transform data
- **Data exploration with LLMs**
  - Prompting an LLM
- **Time series data**
  - Trends, seasonality, cyclicality
  - Analytical methods

Using spreadsheets for data analytics

Solving problems with data

DeepLearning.AI

What outcomes am I interested in?

Problem	Outcome	Data
 Increase profit	➕ Sales	Sales reports
	➖ Expenses	Expense reports
 Improve patient outcomes	➕ Patient satisfaction	Patient surveys
	➖ Length of hospital stays	Admission and discharge dates

What data provides context?

Who	What	When	Where	Outcome
Customer ID	**Product**	**Purchase Date**	**Region**	**Sale**

Increasing revenue for a solar panel company

Who	What	When	Where	Outcome
Customer ID 973	**Product** Cable set	**Purchase Date** 06/15/2024	**Region** North America	**Sale** $1,600
Customer ID 111	**Product** Solar panel	**Purchase Date** 09/10/2024	**Region** Asia	**Sale** $3,200

- High paying customers?

Increasing revenue for a solar panel company

Who	What	When	Where	Outcome
Customer ID 973	**Product** Cable set	**Purchase Date** 06/15/2024	**Region** North America	**Sale** $1,600
Customer ID 111	**Product** Solar panel	**Purchase Date** 09/10/2024	**Region** Asia	**Sale** $3,200

- High paying customers?
- Specific products driving sales?

Increasing revenue for a solar panel company

Who	What	When	Where	Outcome
Customer ID 973	**Product** Cable set	**Purchase Date** 06/15/2024	**Region** North America	**Sale** $1,600
Customer ID 111	**Product** Solar panel	**Purchase Date** 09/10/2024	**Region** Asia	**Sale** $3,200

- High paying customers?
- Specific products driving sales?
- How are sales trending?

Increasing revenue for a solar panel company

Who	What	When	Where	Outcome
Customer ID 973	**Product** Cable set	**Purchase Date** 06/15/2024	**Region** North America	**Sale** $1,600
Customer ID 111	**Product** Solar panel	**Purchase Date** 09/10/2024	**Region** Asia	**Sale** $3,200

- High paying customers?
- Specific products driving sales?
- How are sales trending?
- Sales vary across regions?

Increasing revenue for a solar panel company

Who	What	When	Where	Outcome
Customer ID 973	**Product** Cable set	**Purchase Date** 06/15/2024	**Region** North America	**Sale** $1,600
Customer ID 111	**Product** Solar panel	**Purchase Date** 09/10/2024	**Region** Asia	**Sale** $3,200

- Business or individual purchaser?
  - If business, how large?
- How many orders placed?

- High paying customers?
- Specific products driving sales?
- Sales trends
- Sales vary across regions?

Using spreadsheets for data analytics

Spreadsheets for business analytics

DeepLearning.AI

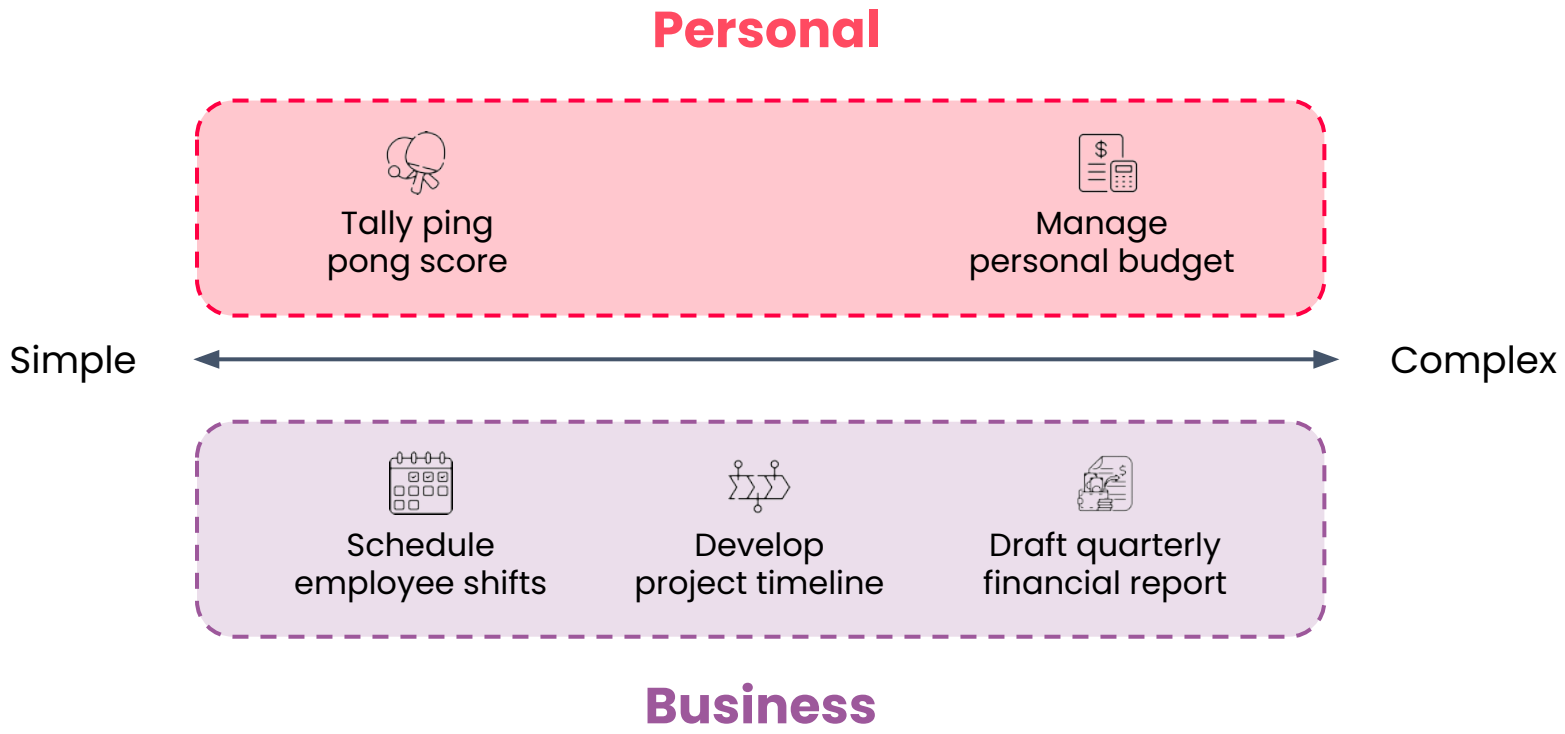
Why start with spreadsheets?

Industry standard tool

Opens in a few seconds

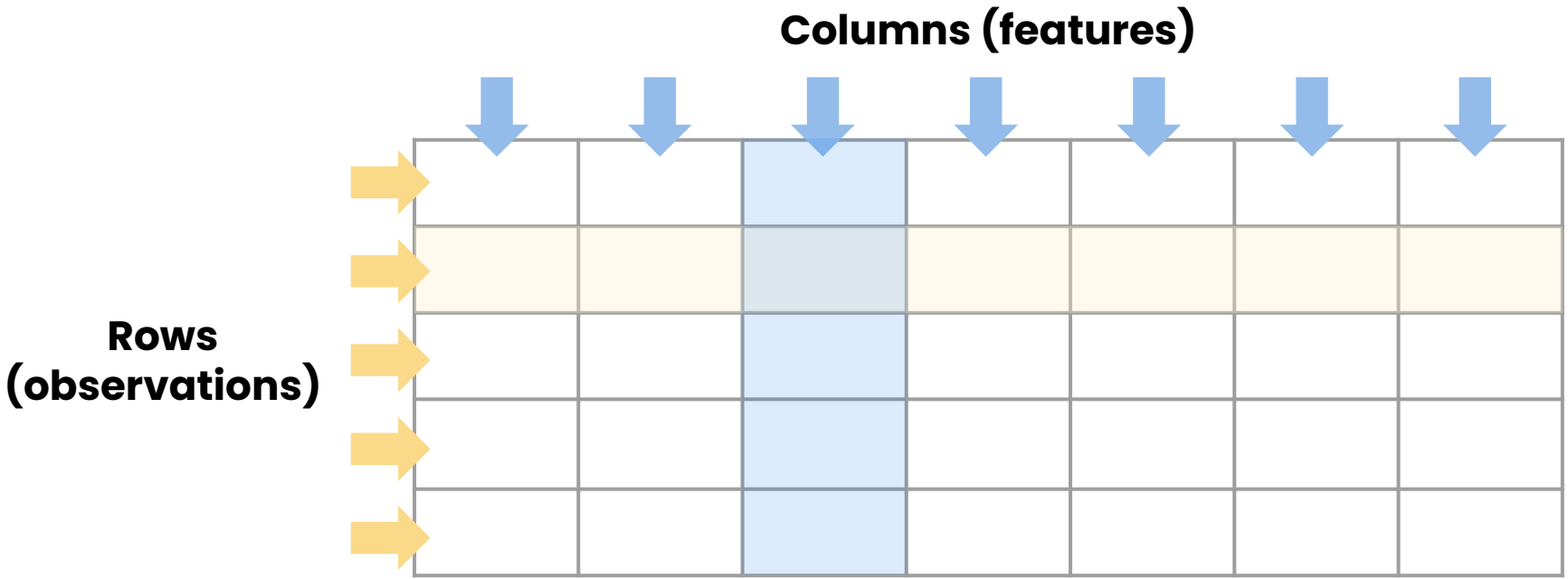
Free

Spreadsheet applications



Spreadsheet applications

⭐ Designed for working with structured data



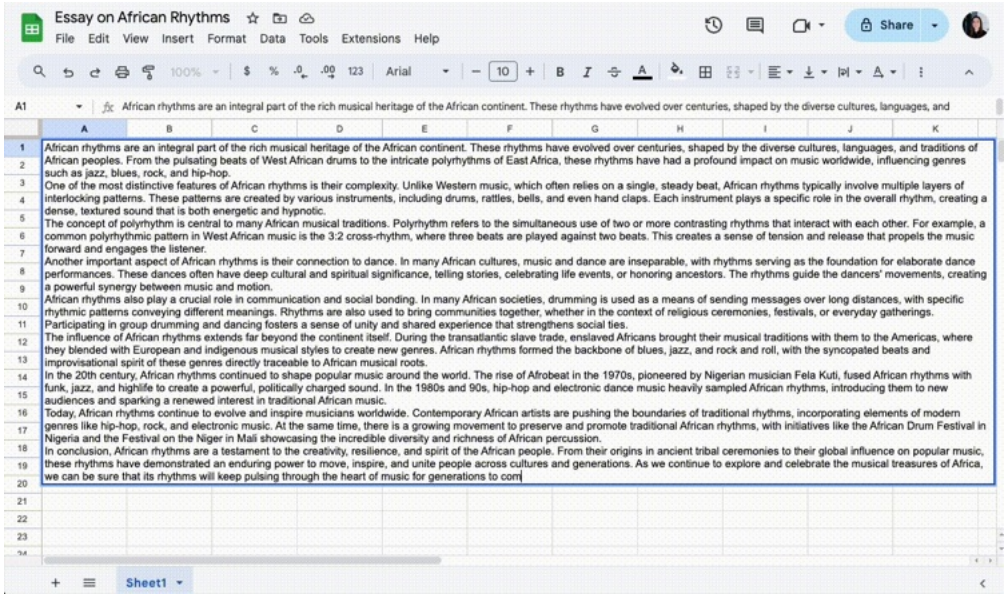
Unstructured data in spreadsheets

text images audio video

✅ collecting

✅ organizing

❌ analyzing



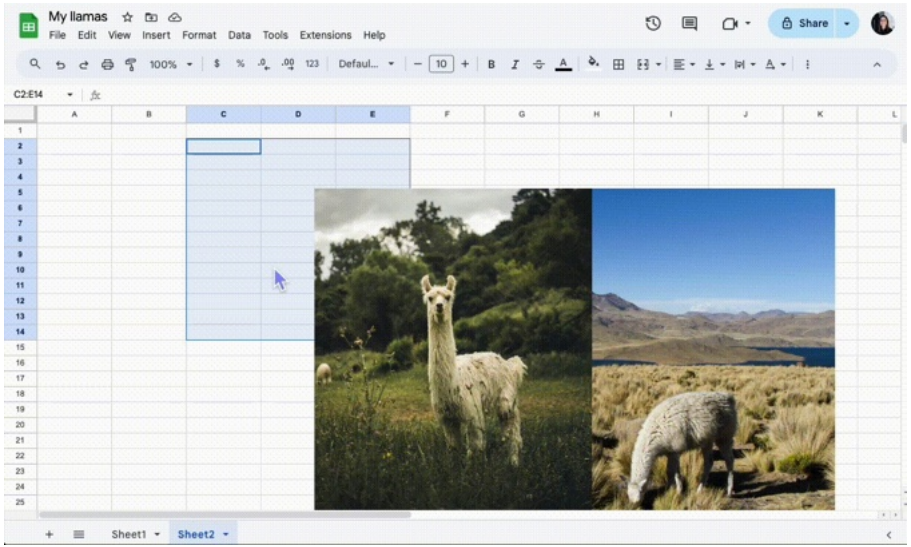
Unstructured data in spreadsheets

text images audio video

✅ collecting

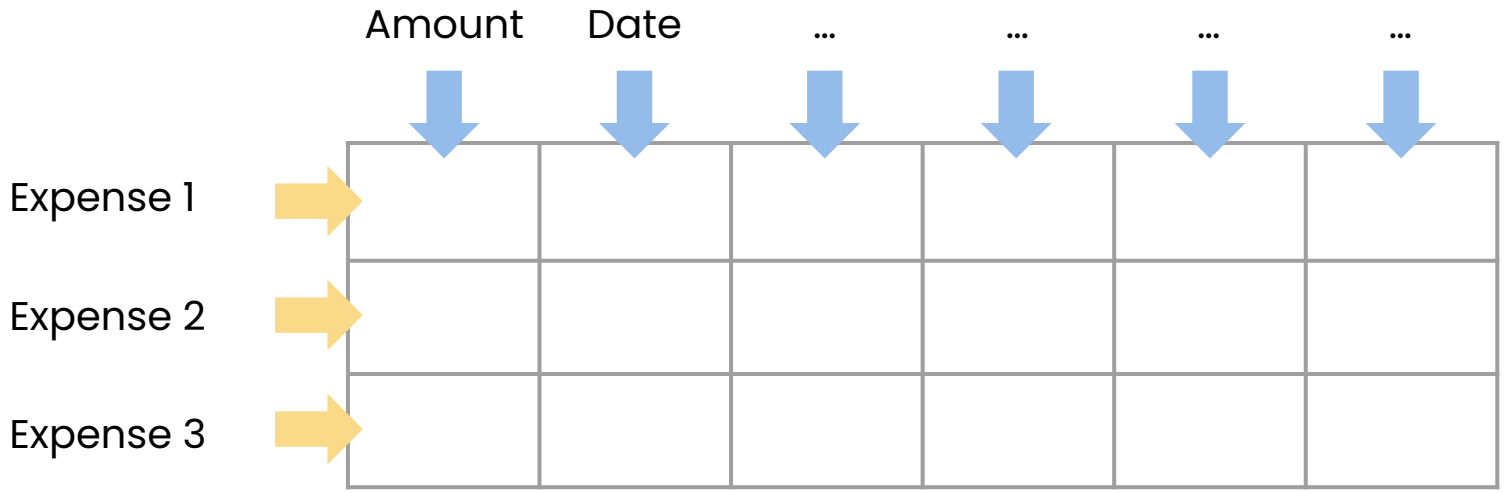
✅ organizing

❌ analyzing



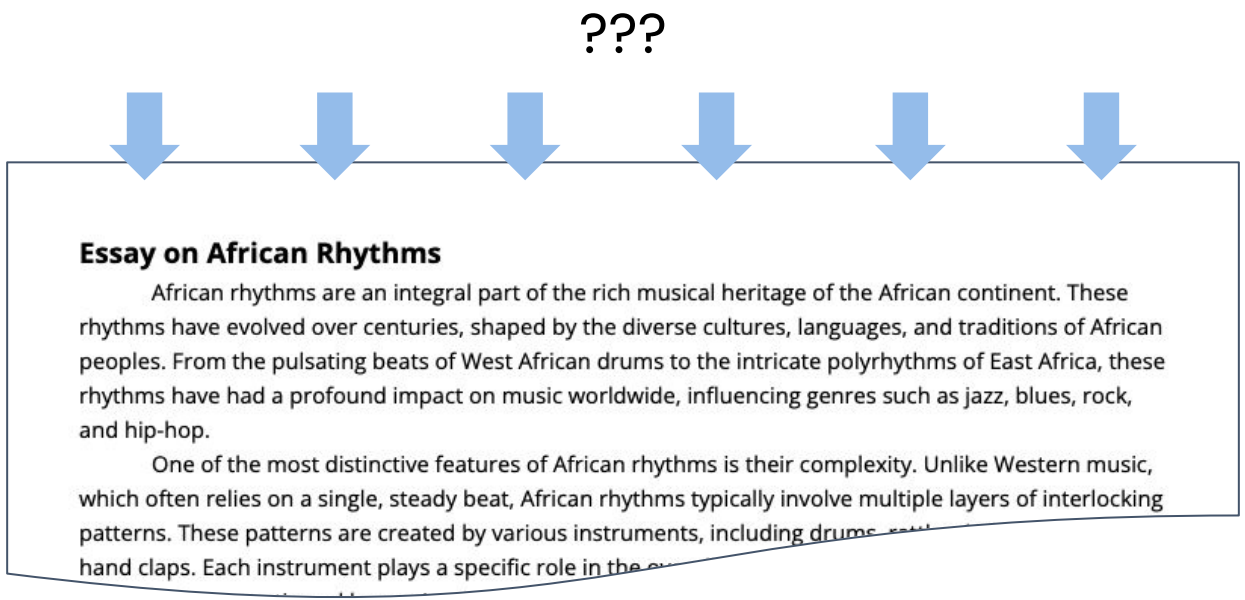
Are spreadsheets right for your use case?

1. Can your data be organized into **rows** and **columns**?



Are spreadsheets right for your use case?

1. Can your data be organized into **rows** and **columns**?



Are spreadsheets right for your use case?

1. Can your data be organized into **rows** and **columns**?

2. Are there **relationships** you want to explore between different aspects of the data?
   a. Organizing expenses by category
   b. Finding the month where you spent the most

Using spreadsheets for data analytics

Navigating Google Sheets

DeepLearning.AI

Google Sheets alternatives

Highly transferable skills

Using spreadsheets for data analytics

Importing data

DeepLearning.AI

Loading data into Google Sheets

Generate data directly in the spreadsheet

Open an existing file

Import a structured data set (.csv, .xlsx)

Using spreadsheets for data analytics

Sorting, filtering, and formatting

DeepLearning.AI

Using spreadsheets for data analytics

Getting to know your data

DeepLearning.AI

Techniques for getting to know your data

Structure:  Structured  Unstructured

Dimensions:  # of observations  # of features

Type:  Numerical  Categorical

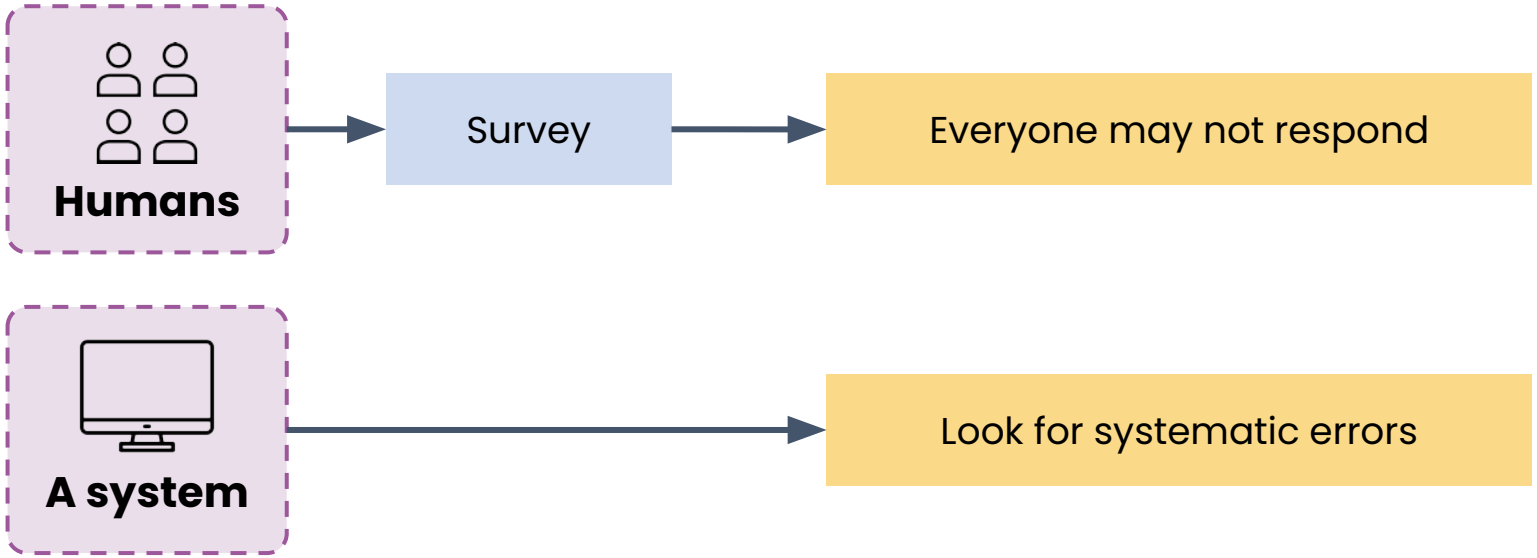
Metadata: data about your data

More metadata

Type of metadata	Example	What it tells you
File name	Hotel reservations	How you can search for the data
Author	Charles Xavier	Who produced the data
Description	About hotel rooms	Context about what information is in data
File format	.csv, .txt, .json	How you may need to interact with the data
Created/last updated date	January 20, 2024	How current the data is
Access controls	View, comment, edit	Who can access the data and how they can interact with it

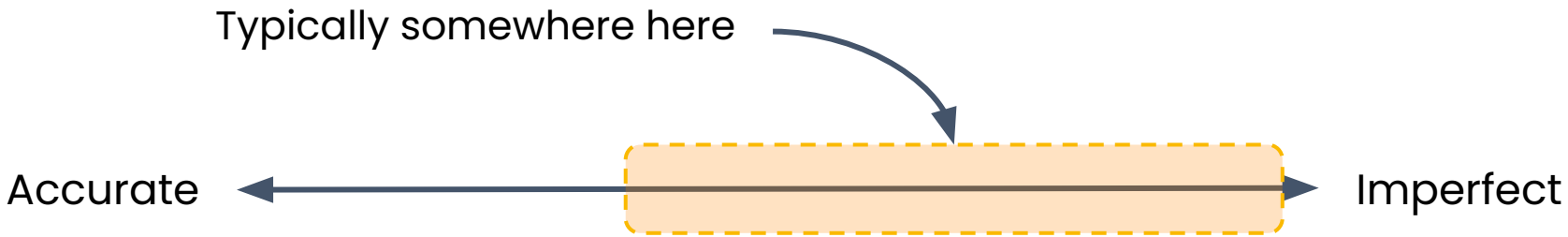
The source of your data

- How was it **collected** or **generated**?



The source of your data

- How was it **collected** or **generated**?
- What is the **quality** of the data?



Using spreadsheets for data analytics

Summary statistics – MAX, MIN, AVERAGE

DeepLearning.AI

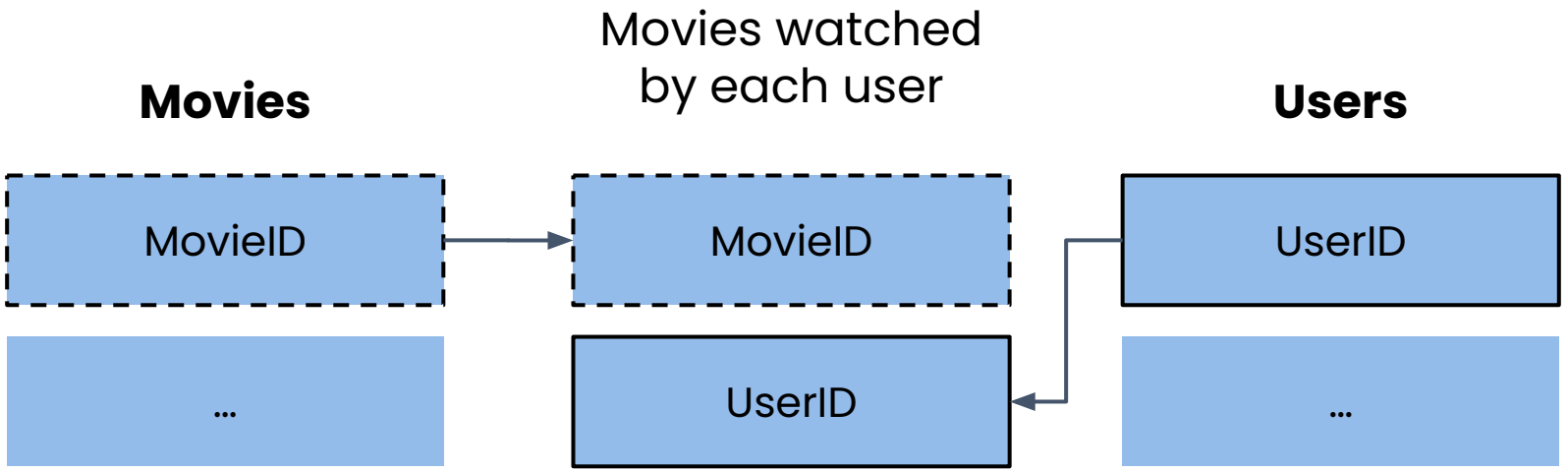
Summary statistics

Survey of active subscribers

Feature	Summary statistic	Insight
Age group	**Frequency** of each group	Type of content produced
Subscription duration	**Range** of subscription lengths	Promotional rates?

Relationship: How does Subscription duration vary across Age group ?

Relations to other data sets



Using spreadsheets for data analytics

Conditional formatting

DeepLearning.AI

Conditional formatting

Allows you to:

- Apply a particular rule across your data set
  - Numerical data
  - Categorical data
  - Dates

8
1
3
7
8
9
10
6

8
1
3
7
8
9
10
6

Visual signals for:

- Spotting trends across many data points
- Identifying positive or negative changes
- Identifying which values fall above or below a threshold

Two types of conditional formatting

Single color

Highlight cells that satisfy a particular condition

8
1
3
7
8
9
10
6

Rule:

Highlight **green** days with >= 8 sales

Color scale

Applies one of several colors to each cell based on its value

Two common types:

Sequential

Divergent

Using spreadsheets for data analytics

Summary statistics – COUNTIF

DeepLearning.AI

Q. What percent of bookings were canceled? ❌

COUNTIF ← **Conditional**

Given a range or group of cells

Tallies them up only if they meet a **condition**

Similar to a filter,

but counts rather than displays data

Formula

```
= COUNTIF ( A1:A9 , " <=0 " )
```

Range Criteria

Using spreadsheets for data analytics

Summary statistics – SUMIF, AVERAGEIF

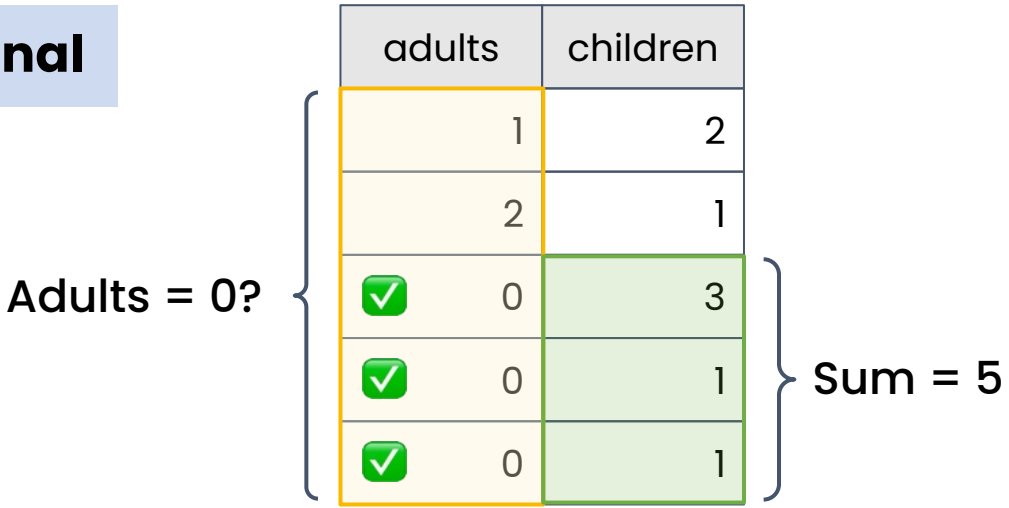
DeepLearning.AI

If the number of adults equals zero, add the number of children up.

SUMIF ← **Conditional**

Adds all the cells in a range if cells meet a condition

Check cells in one range, **add** cells in another



Formula **Range** **Condition** **Sum range**

```
= SUMIF( A1:A5 , "=0" , B1:B5 )
```

Average lead time of bookings that were and were not canceled.

AVERAGEIF

Averages all values that meet the condition

Adults = 0?

Adults	Children
1	2
2	1
✅ 0	3
✅ 0	1
✅ 0	1

Avg = 1.6

Formula

Range **Condition** **Sum range**

```
AVERAGEIF( range , condition , sum_range )
```

Using spreadsheets for data analytics

Summary statistics – COUNTIFS, SUMIFS

DeepLearning.AI

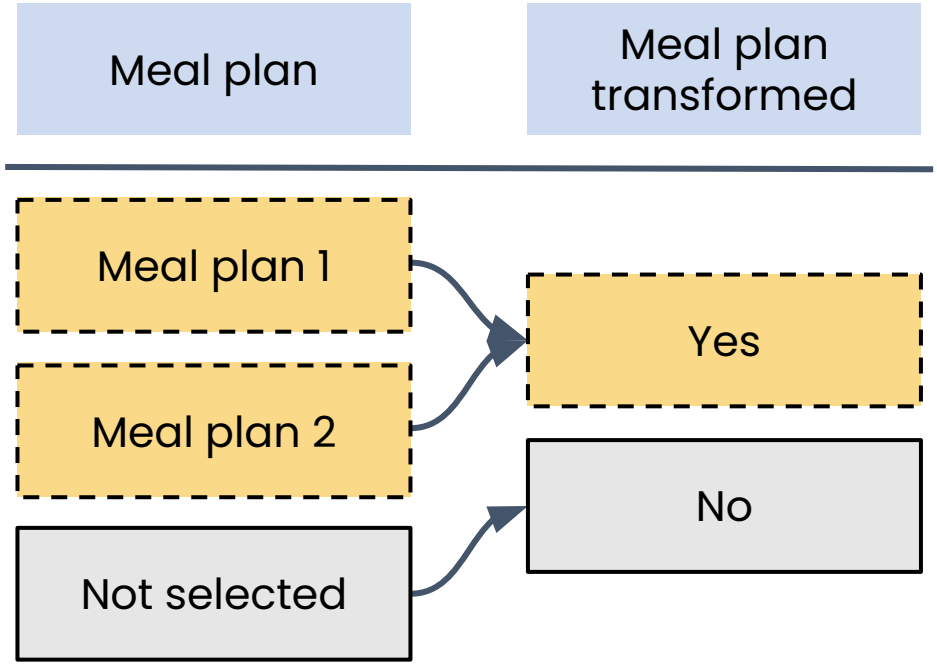
Using spreadsheets for data analytics

Data processing –
IF, IFS, RIGHT, LEFT

DeepLearning.AI

Categorical groupings

Group features with similar relationship to outcome



IF

Checks a condition, then returns values depending on whether the condition is true or false

Formula

```
=IF(A2, “=Not selected”, “No”, “Yes” )
```

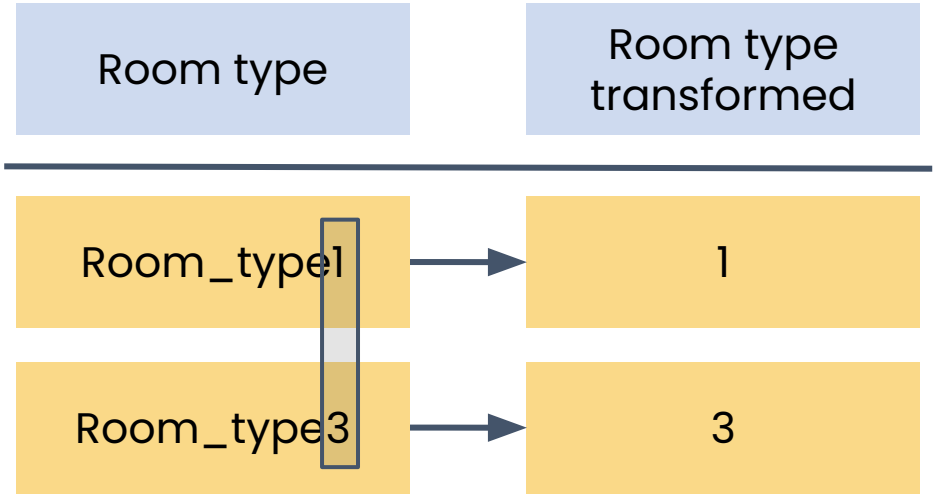
Text processing

Rename category or extract part of a category

LEFT

RIGHT

Take leftmost or rightmost characters and put them in a new cell



Formula

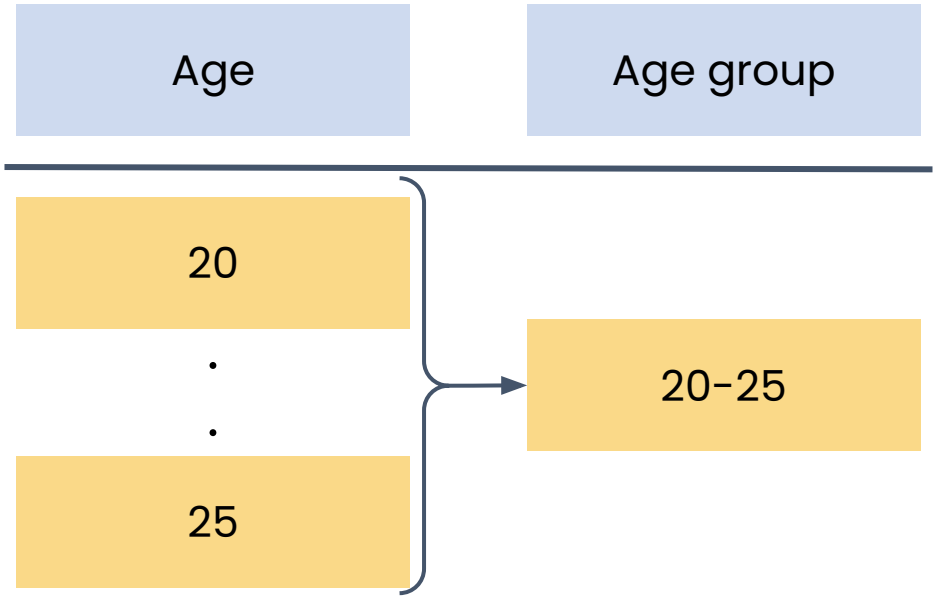
```
=RIGHT(A2, 1)
```

Binning

Group numerical data into categories

IFS

- Checks multiple conditions



Condition1 Value if true1 Condition2 Value if true2

Formula

```
=IFS(A2<18, "Under 18", A2<30, "18-29")
```

Spreadsheets functions or data processing

IF

LEFT

RIGHT

IFS

Checks multiple conditions

Condition 1 Condition 2

=**IFS**(A2=”Not selected”, “No”, A2=”Selected”, “Yes”)

Using spreadsheets for data analytics

Where does data come from?

DeepLearning.AI

Data comes from many sources

[keri, Magic 8 ball - MY SOURCES SAY NO, https://commons.wikimedia.org/, CC 2.0, (2014)]

[NASA, NOAA-M, https://commons.wikimedia.org/wiki/File:NOAA-M.jpg, (2002)]

Data comes from many sources

1. **Direct input**
   - Explicitly provided through a structured process

Feedback survey

Registration form

2. **Behavioral observation**
   - Passively observing individuals' actions

Website analytics

Mobile app usage

3. **Physical sensors**
   - Continuously monitor some phenomenon

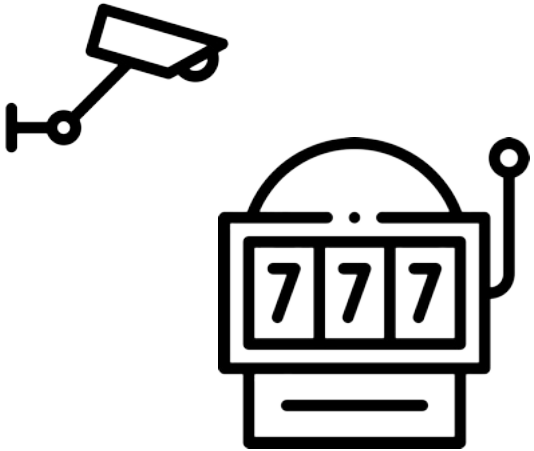
Smart thermostat

Vehicles

1st party, 2nd party, and 3rd party data

Your company
1st party

Collected by you or your company directly



1st party, 2nd party, and 3rd party data

Your company
1st party

Collected by you or your company directly

Trusted partner
2nd party

Collected as their own 1st party data

✅

1st party, 2nd party, and 3rd party data

Your company
1st party

Collected by you or your company directly

Trusted partner
2nd party

Collected as their own 1st party data

Data sales company
3rd party

Collects data to sell to multiple buyers

Other companies

1st party, 2nd party, and 3rd party data

Your company 1st party	**Trusted partner** 2nd party	**Data sales company** 3rd party
Collected by you or your company directly	Collected as their own 1st party data	Collects data to sell to multiple buyers

More control ⟷ Less control

Publicly available data

Sources

- Government agencies
- Research organizations
- Open-source databases

Usage

- Download directly
- Scrape data with Python

Great resource because

- Free to access
- Real-world data

Data privacy

Only analyze data you are authorized to access

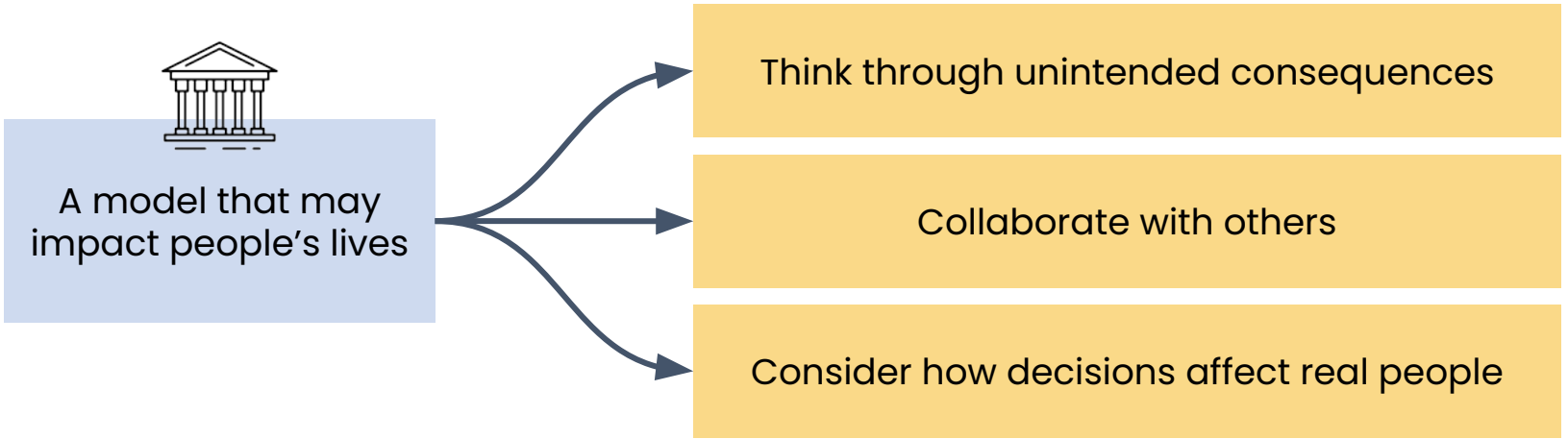
Take training

Operate within secure computing environments

Strip data of personally identifiable information

Ethical usage

Using the **data** or **insights** you are producing in a fair and beneficial way



Using spreadsheets for data analytics

Data exploration with LLMs

DeepLearning.AI

Using spreadsheets for data analytics

Introduction to time series

DeepLearning.AI

Introduction to time series

- Measuring one or more outcomes over fixed periods of time
  - Minutes
  - Hours
  - Days

- Goal: Understand how the outcomes vary over time
  - Identify increasing or decreasing trends
  - Discover unusual events
  - Forecast future outcomes

Common use cases for time series

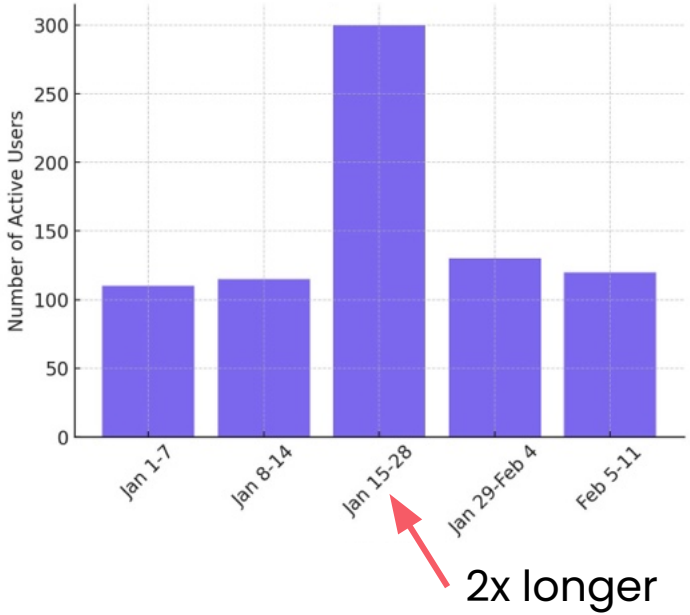
Measurement	Time interval	Purpose
 Rainfall in the Amazon	Annually	Global warming
 Active users	Weekly	User growth
 Sales	Monthly, quarterly	Revenue forecasting
 Stock market prices	By minute, hourly	Day trading
 Stock market prices	Monthly, quarterly, yearly	Long-term investing

Why have a consistent time interval?

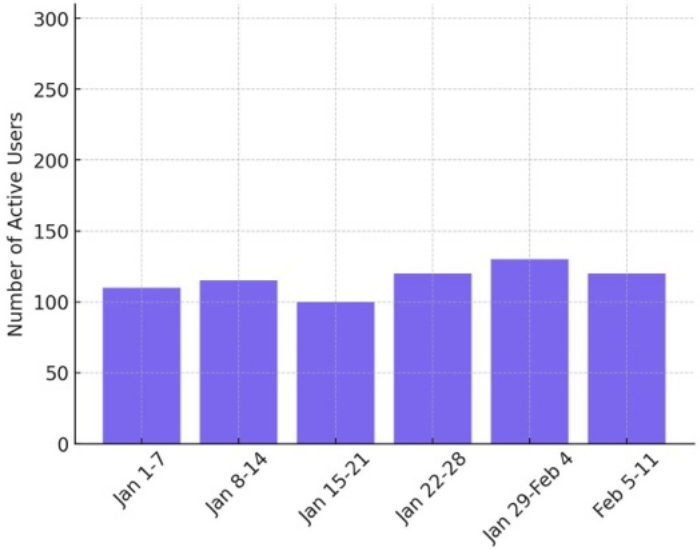
- Interpretability
- Consistent frame of reference
- Focus on comparing the outcome itself

Same data, different interval

Inconsistent interval



Consistent interval



Comparison with cross-sectional data

Type of Data	Measurement	Time Period
Time Series	1	Multiple
Cross-sectional	Multiple	1

✅ Focus is typically on each observation as a one-time event

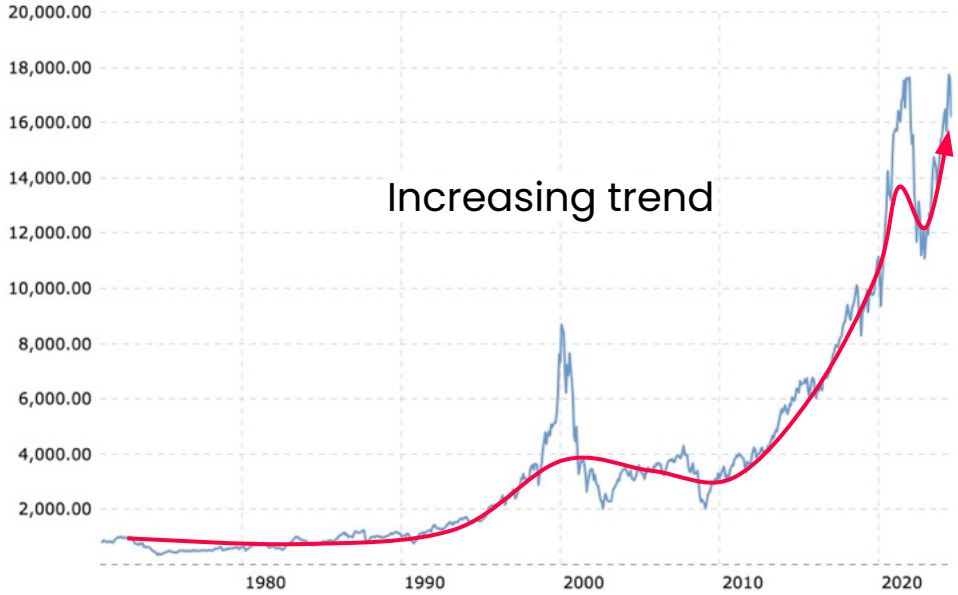
⛔ Rather than on how those events change over time

Trend | **Seasonality** | **Cyclical fluctuations** | **Irregularities**

Trend: long term direction

- Increasing
- Decreasing
- Stationary

U.S. stock market for past 55 years

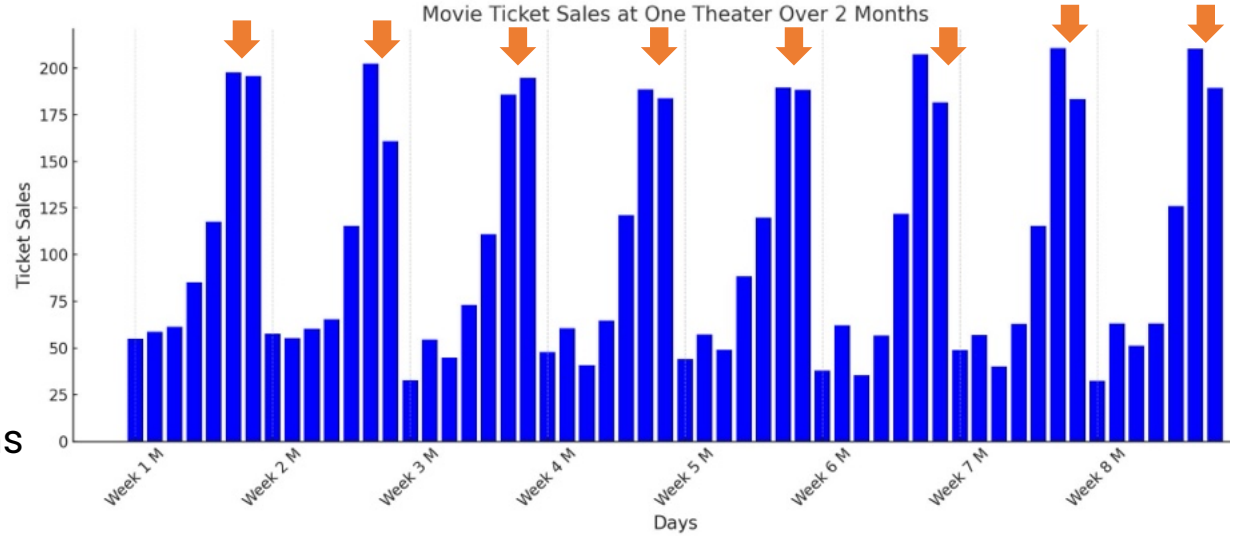


[Nasdaq Composite - 45 Year Historical Chart." MacroTrends, www.macrotrends.net/1320/nasdaq-historical-chart.]

Trend | **Seasonality** | Cyclical fluctuations | Irregularities

Seasonality: a repeating, predictable pattern that occurs at regular intervals

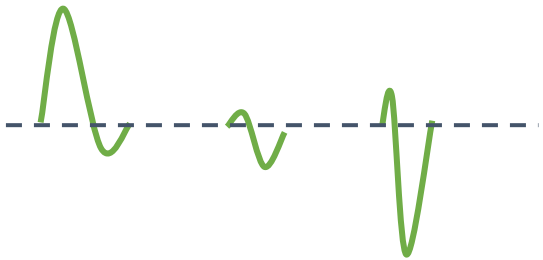
- Daily
- Weekly
- Monthly
- Yearly
- Not just weather seasons
- Can appear at multiple time intervals



Trend | Seasonality | **Cyclical fluctuations** | Irregularities

Cyclicality: repeating ups and downs that don't occur at regular intervals

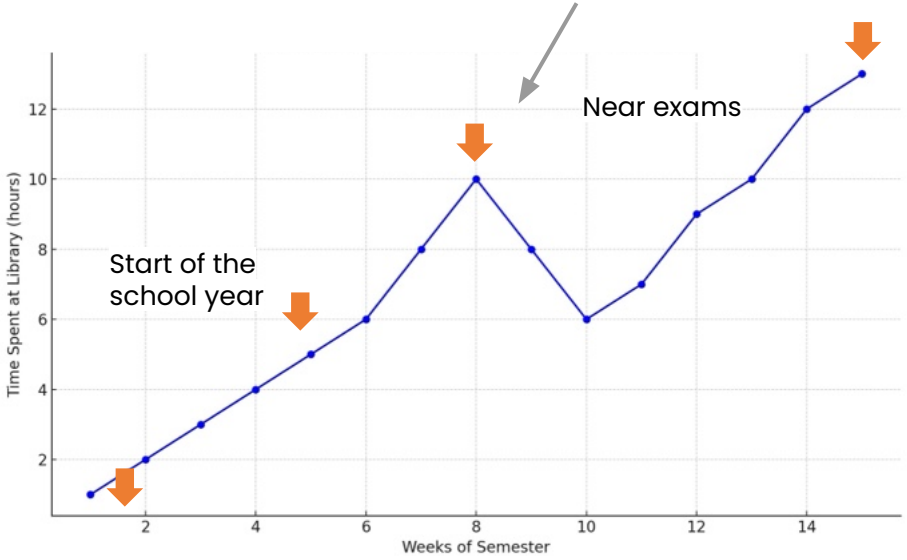
- Sizes of increases and decreases often not the same
- Harder to predict

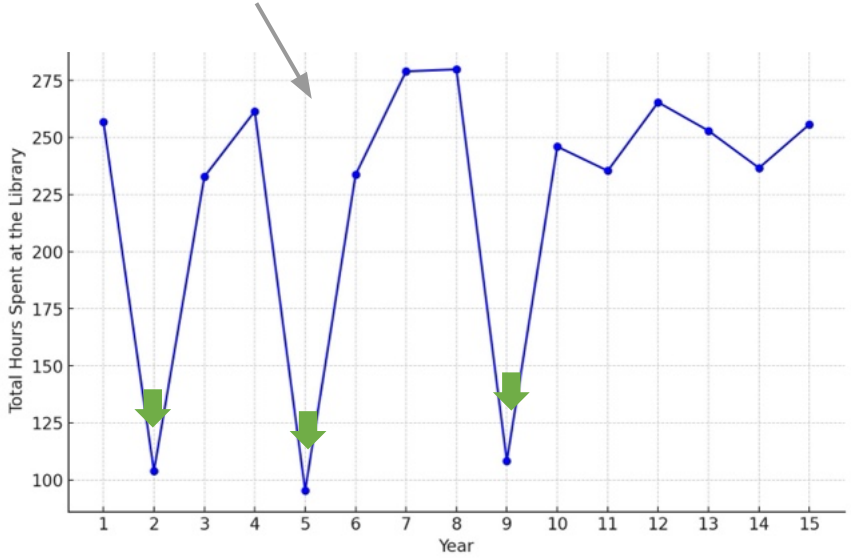


U.S. stock market for past 55 years

[Nasdaq Composite - 45 Year Historical Chart." MacroTrends, www.macrotrends.net/1320/nasdaq-historical-chart.]

Trend | **Seasonality** | **Cyclical fluctuations** | **Irregularities**



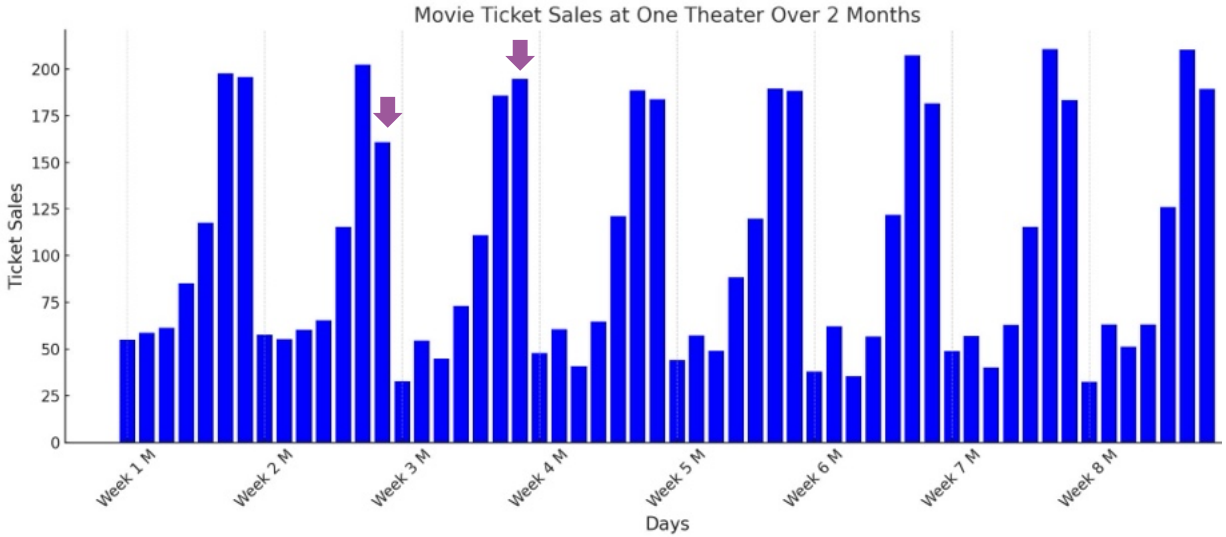


Regular ← → Unpredictable

Trend | Seasonality | Cyclical fluctuations | **Irregularities**

Irregularities: random, unpredictable fluctuations in the data

- Also called **noise**
- Makes it harder to spot the trend, seasonality, and cyclicality

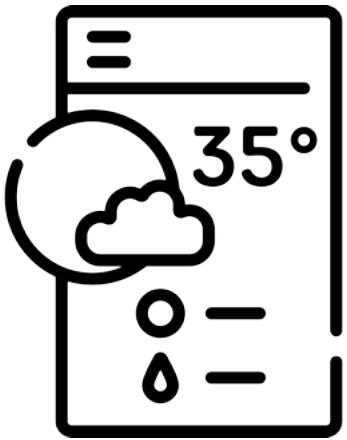


Using spreadsheets for data analytics

Real-world time series

DeepLearning.AI

Weather



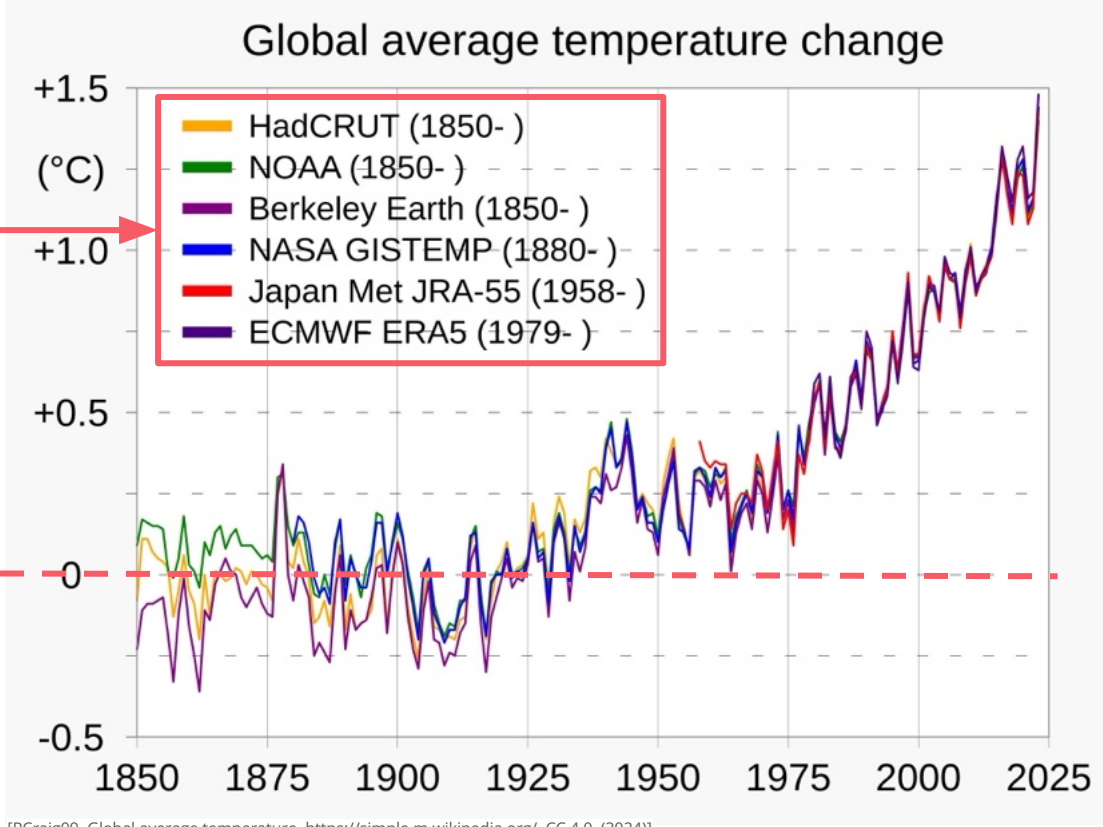
Time scale		Insights
Hours or days	→	What should I wear today?
Longer periods	→	Trends, seasonality, cyclicality

Temperature

Time series

y axis – temperature

Avg temp 1850 - 1900



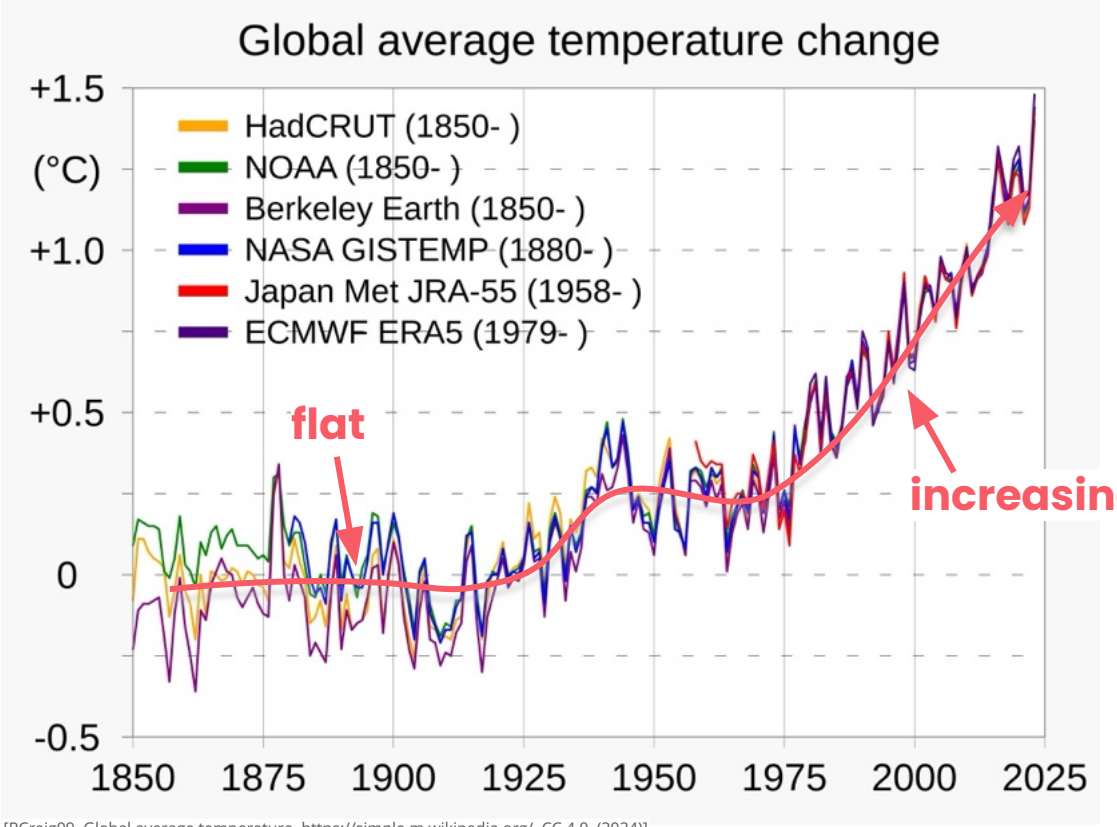
[RCraig09, Global average temperature, https://simple.m.wikipedia.org/, CC 4.0, (2024)]

x axis – year

Temperature

y axis – temperature

Trend



[RCraig09, Global average temperature, https://simple.m.wikipedia.org/, CC 4.0, (2024)]

x axis – year

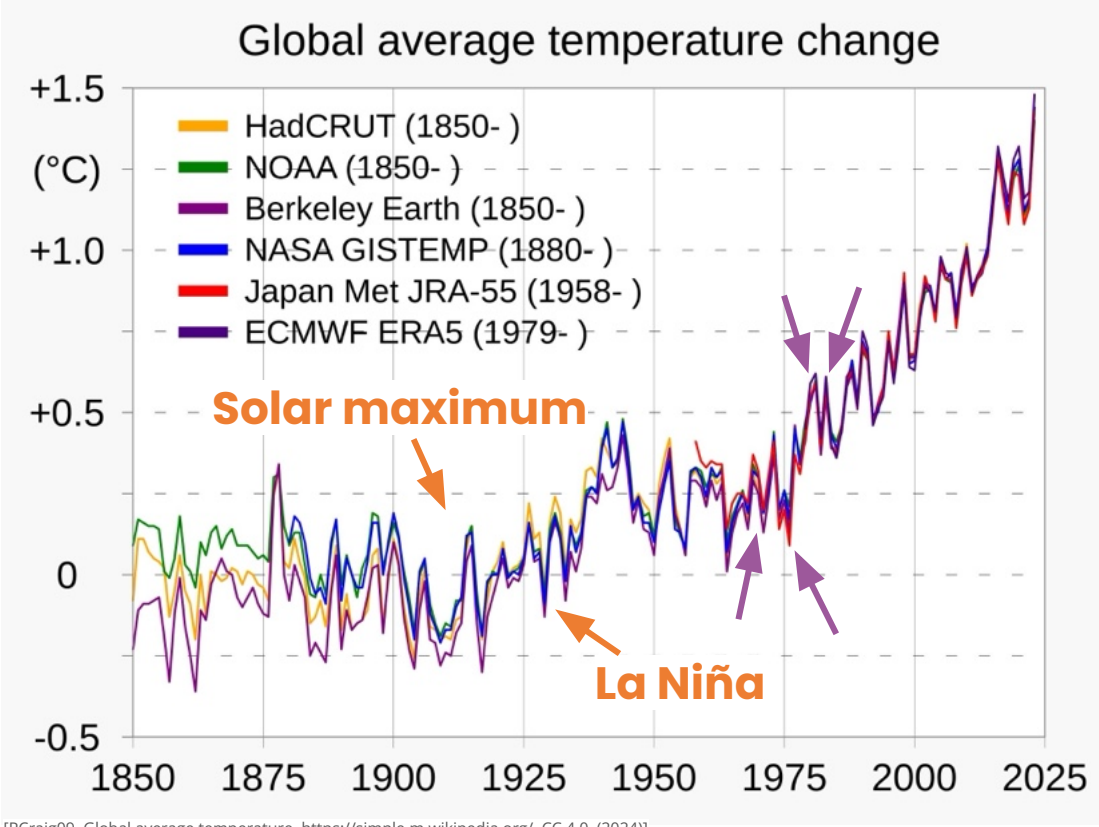
Temperature

y axis – temperature

Trend

Seasonality

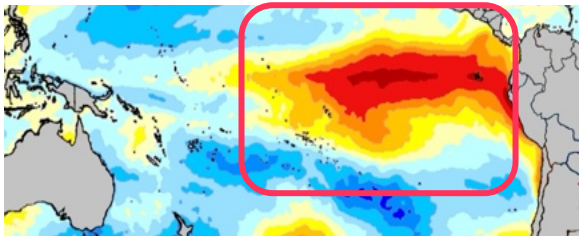
Noise



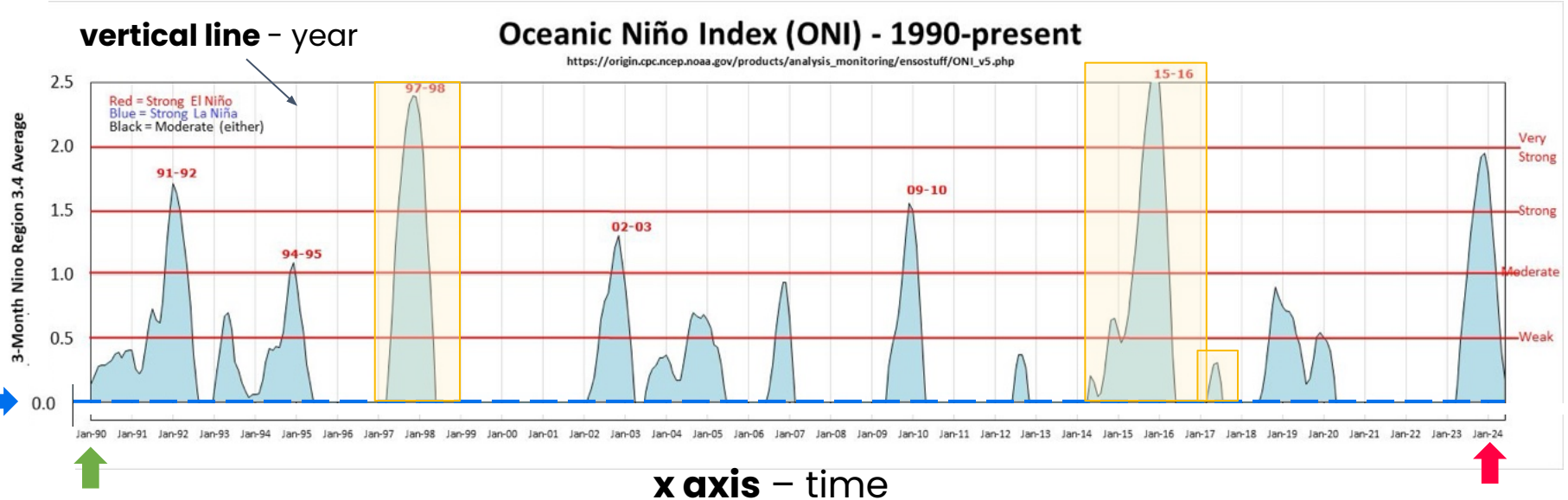
[RCraig09, Global average temperature, https://simple.m.wikipedia.org/, CC 4.0, (2024)]

x axis – year

El Nino

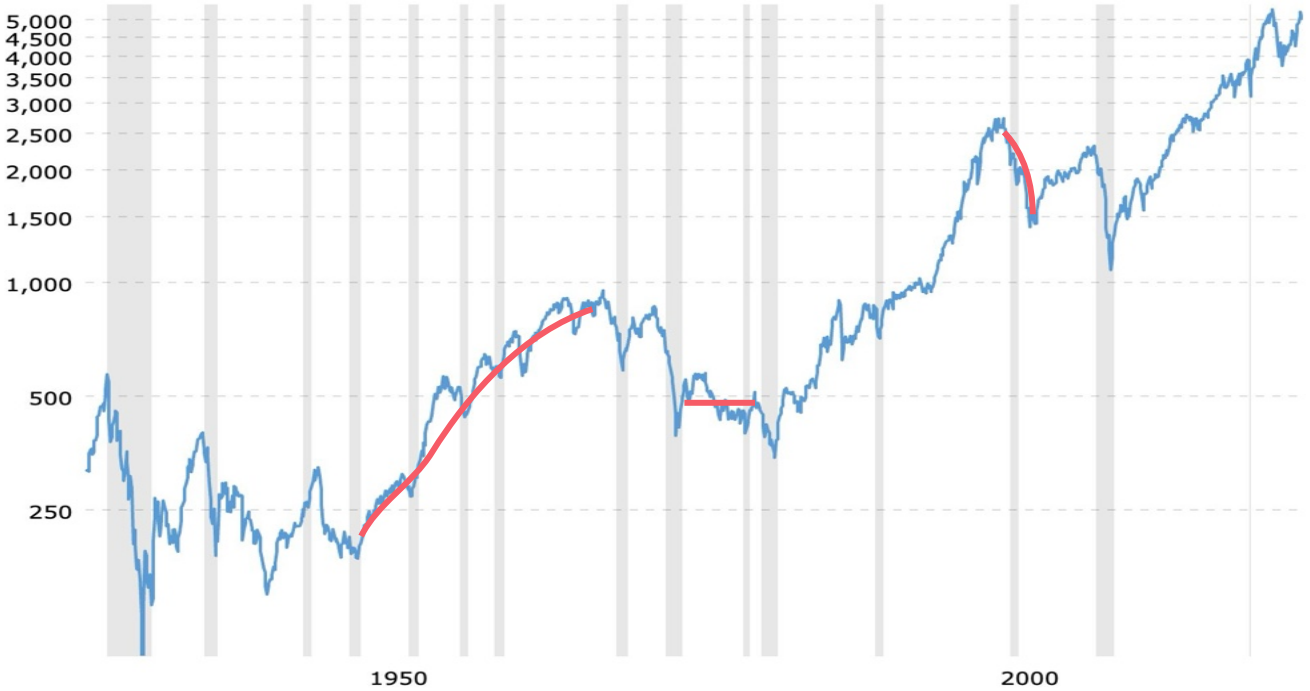


- Caused by wind patterns
- Periodic, but not at fixed intervals
- 9-12 months, but can last for years



[National Weather Service, Climate Prediction Center, Cold & Warm Episodes by Season. (n.d.)]

Stock market prices - S&P 500

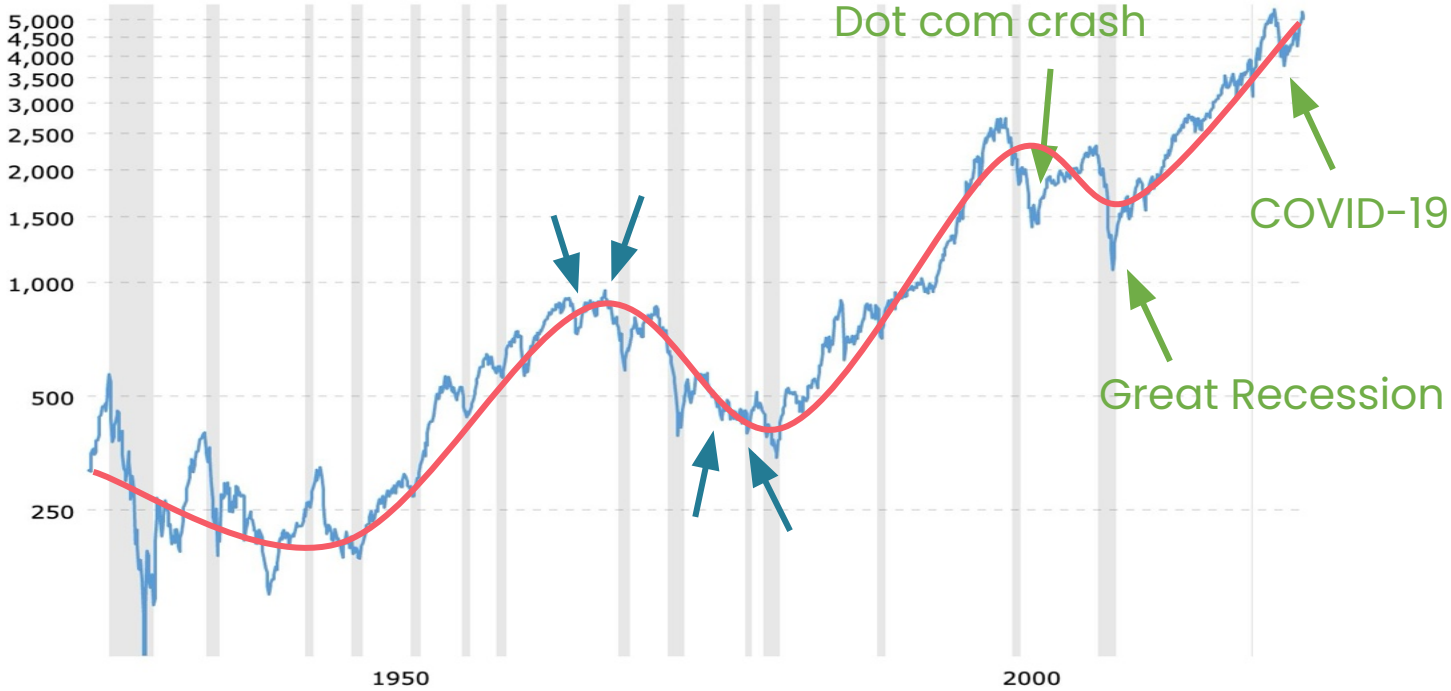


y axis – price

Trend

x axis – year

Stock market prices - S&P 500



y axis – price

Trend

Cyclicality

Noise

x axis – year

Using spreadsheets for data analytics

Moving averages

DeepLearning.AI

Average

Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6

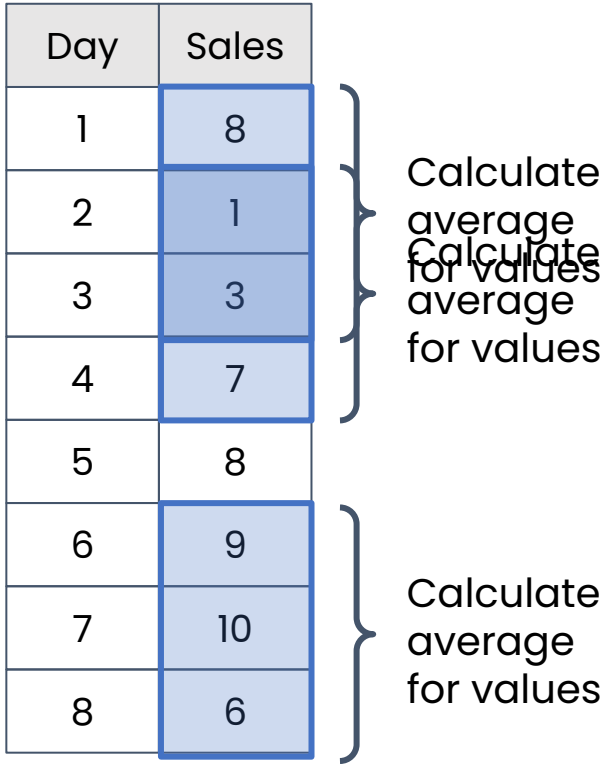
Above average

Well below average

Average 6.5

Moving average

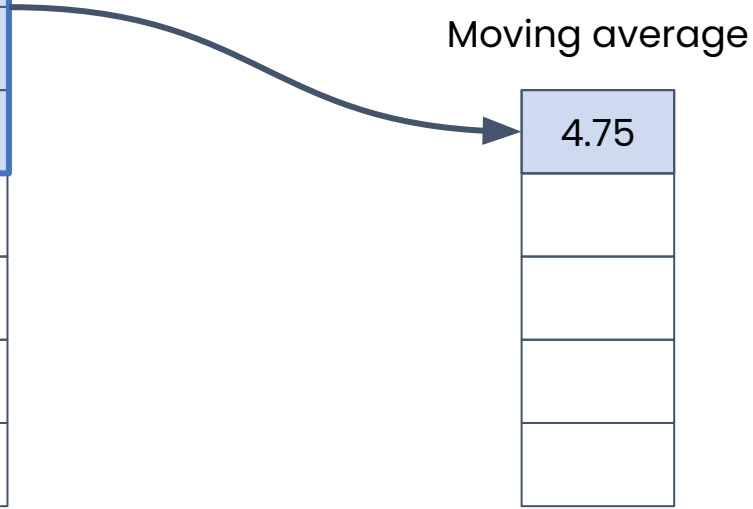
- Smooth out noisy behavior
- How to calculate a simple moving average:
  - Calculate average of **N** consecutive time periods
  - Calculates a series of values (**N**-1 shorter than data)
  - Larger values of **N** are more stable



Moving average

N = 4

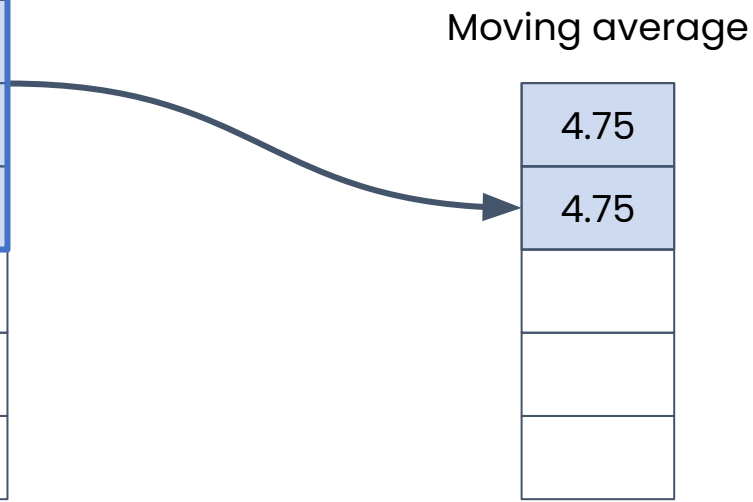
Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6



Moving average

N = 4

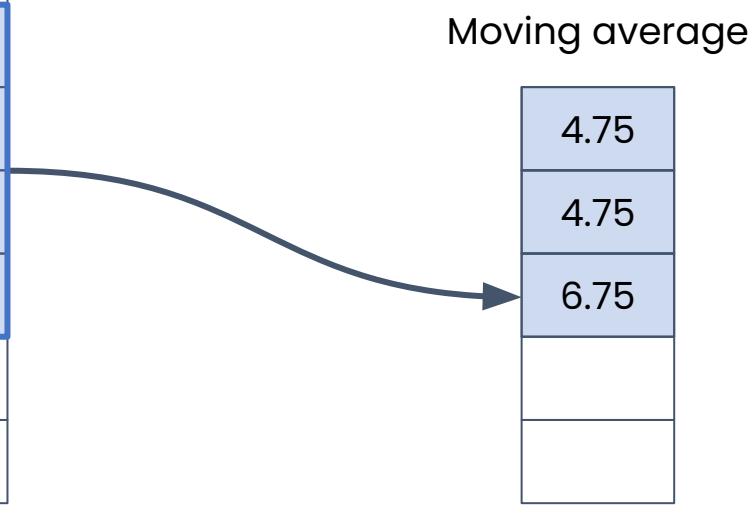
Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6



Moving average

N = 4

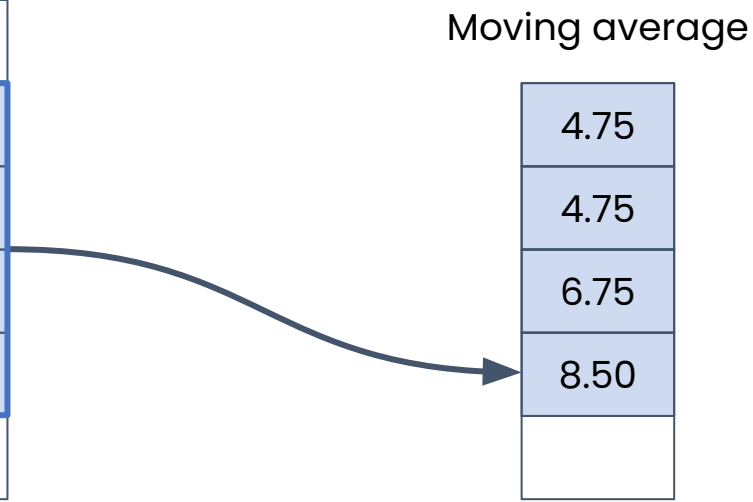
Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6



Moving average

N = 4

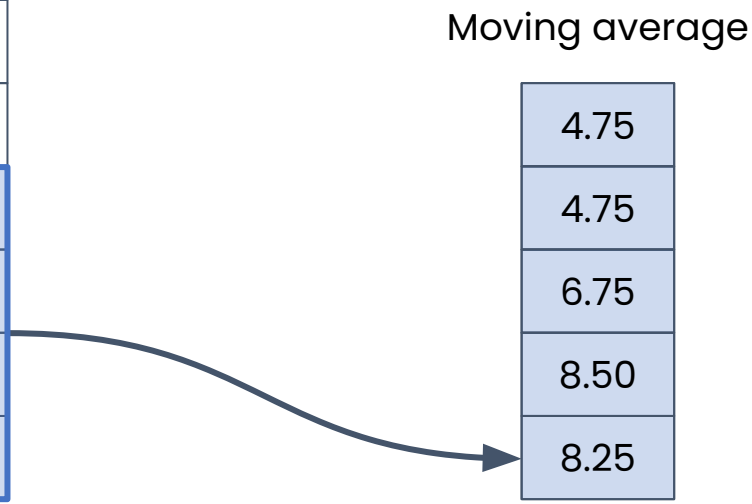
Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6



Moving average

N = 4

Day	Sales
1	8
2	1
3	3
4	7
5	8
6	9
7	10
8	6



Real-world example

- Names tend to follow cycles of popularity

In this exercise:

👶 Work with a dataset of popular US baby names

🔮 Predict whether a particular name may see a resurgence

Using spreadsheets for data analytics

Percent change

DeepLearning.AI

Percent change

To standardize these difference:

- Convert from original units to percentages:

$$\frac{Sales_{current} - Sales_{previous}}{Sales_{previous}}$$

Day	Sales
1	8
2	1

Difference: -7 sales

% change = $\frac{-7}{8}$ -87.5%

Day	Sales
20	108
21	101

Difference: -7 sales

% change = $\frac{-7}{108}$ -6%

Percent change

Day before - X_{t-1}

Current day - X_t

$$\frac{X_t - X_{t-1}}{X_{t-1}} \times 100$$

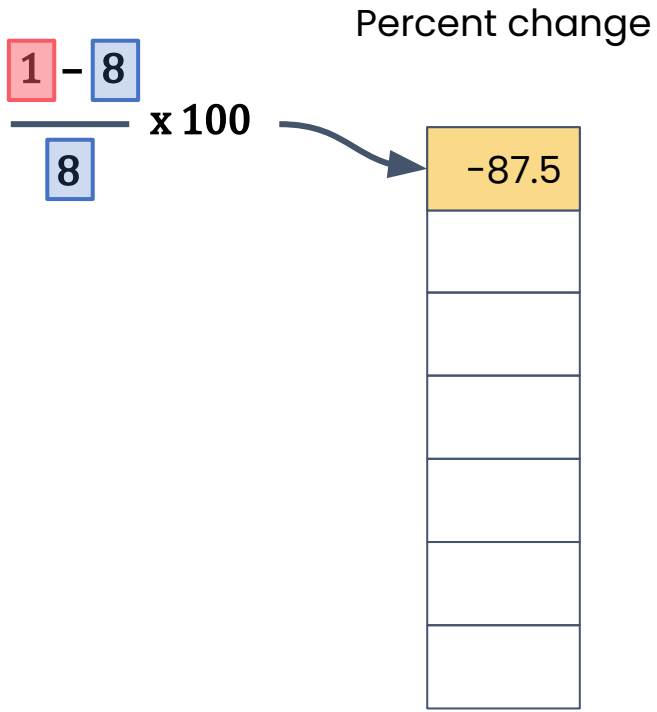
% change → positive

% change → negative

Time series analysis: Percent difference from period to period

t = 2

Day	Sales	
1	8	X_{t-1}
2	1	X_t
3	3	
4	7	
5	8	
6	9	
7	10	
8	6	



Sean Barnes

Time series analysis: Percent difference from period to period

t = 3

Day	Sales	
1	8	
2	1	X_{t-1}
3	3	X_t
4	7	
5	8	
6	9	
7	10	
8	6	

$$\frac{3 - 1}{1} \times 100$$

Percent change

-87.5
200

Time series analysis: Percent difference from period to period

t = 4

Day	Sales	
1	8	
2	1	
3	3	X_{t-1}
4	7	X_t
5	8	
6	9	
7	10	
8	6	

$$\frac{7 - 3}{3} \times 100$$

Percent change

-87.5
200
133.3

Time series analysis: Percent difference from period to period

t = 5

Day	Sales	
1	8	
2	1	
3	3	
4	7	X_{t-1}
5	8	X_t
6	9	
7	10	
8	6	

Percent change

$$\frac{8 - 7}{7} \times 100$$

Percent change
-87.5
200
133.3
14.3

Time series analysis: Percent difference from period to period

t = 6

Day	Sales	
1	8	
2	1	
3	3	
4	7	
5	8	X_{t-1}
6	9	X_t
7	10	
8	6	

Percent change

$$\frac{9 - 8}{8} \times 100$$

Percent change
-87.5
200
133.3
14.3
12.5

Time series analysis: Percent difference from period to period

t = 7

Day	Sales	
1	8	
2	1	
3	3	
4	7	
5	8	
6	9	X_{t-1}
7	10	X_t
8	6	

Percent change

$$\frac{10 - 9}{9} \times 100$$

-87.5
200
133.3
14.3
12.5
11.1

Time series analysis: Percent difference from period to period

t = 8

Day	Sales	
1	8	
2	1	
3	3	
4	7	
5	8	
6	9	
7	10	X_{t-1}
8	6	X_t

$$\frac{6 - 10}{10} \times 100$$

Percent change

-87.5
200
133.3
14.3
12.5
11.1
-40.0