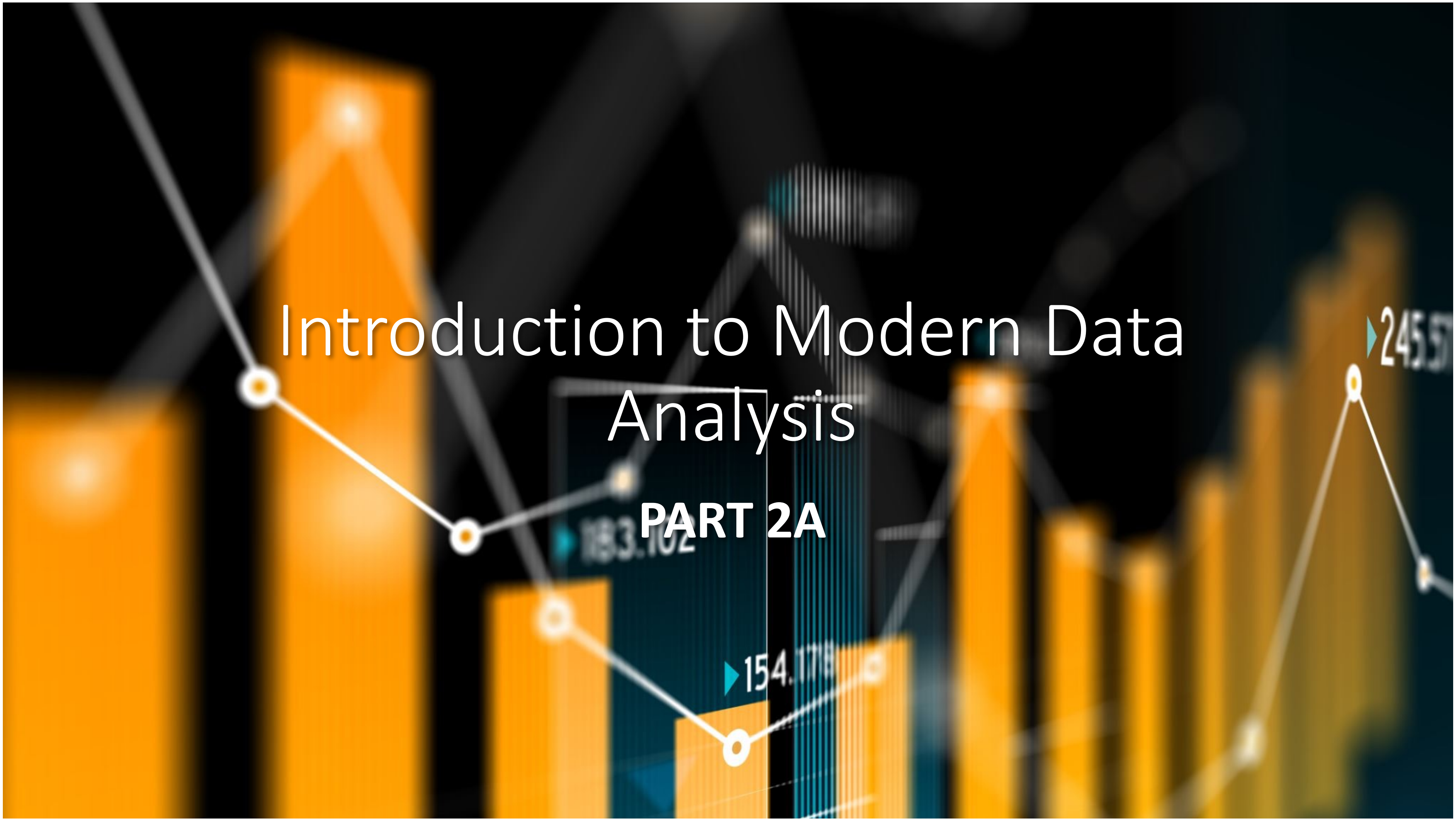
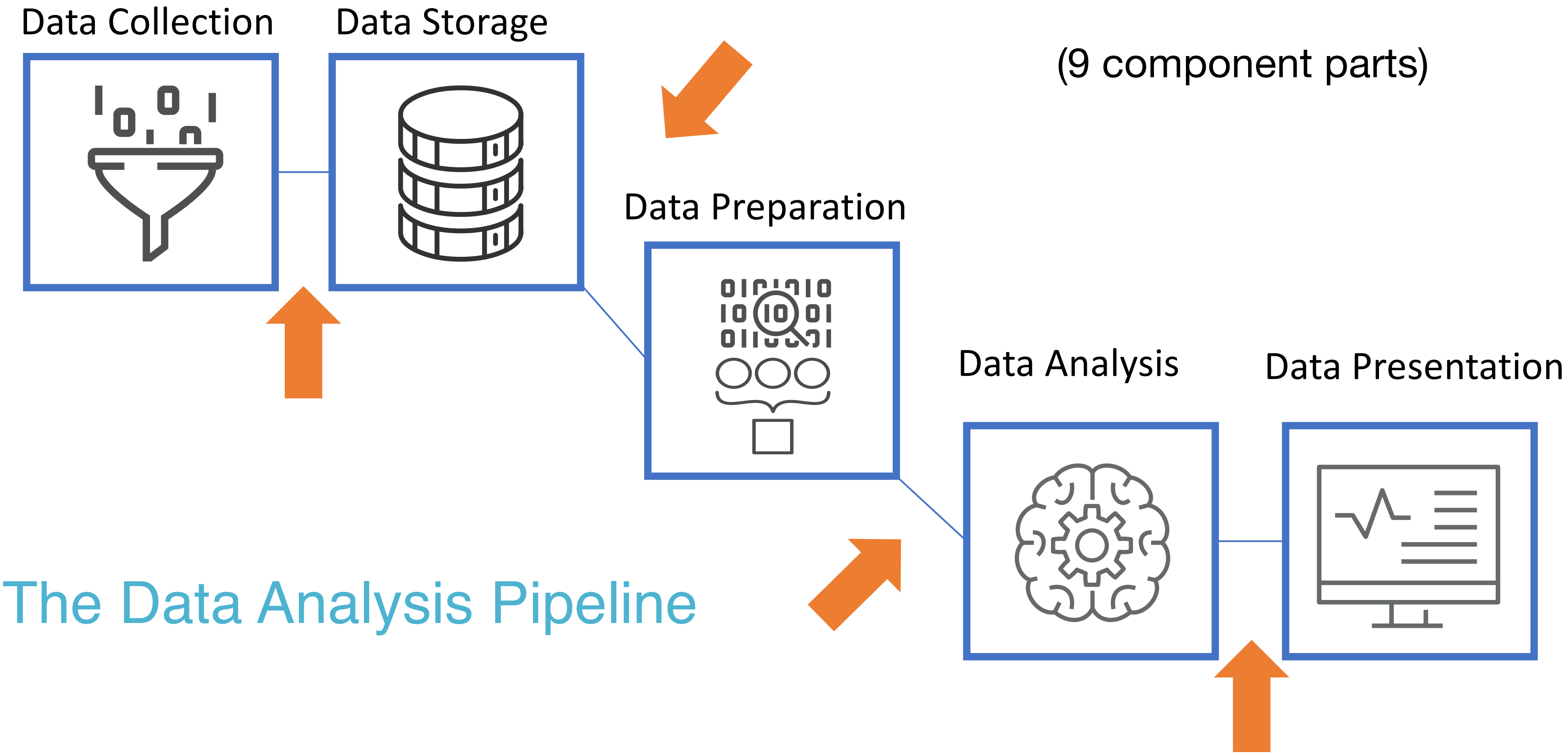


# Introduction to Modern Data Analysis

## PART 2A

# The Data Analysis Pipeline

(9 component parts)



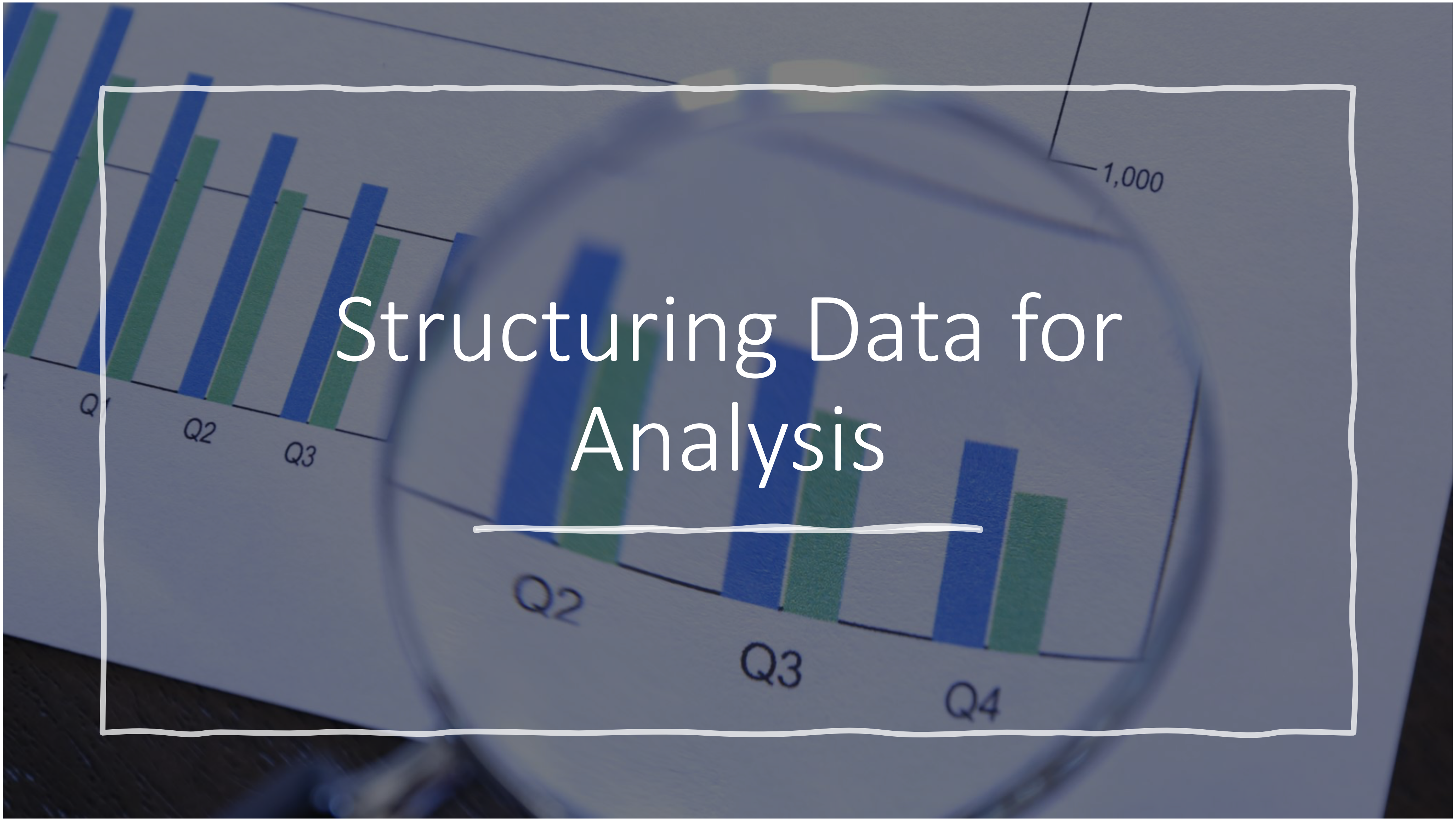
Data Collection

Data Storage

Data Preparation

Data Analysis

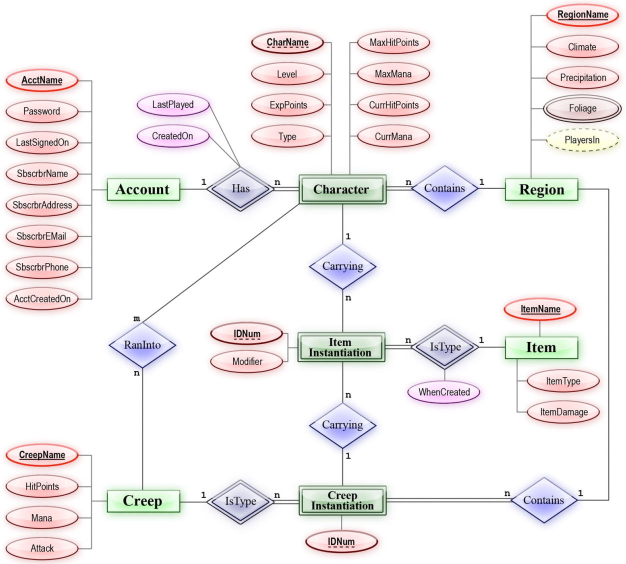
Data Presentation



# Structuring Data for Analysis

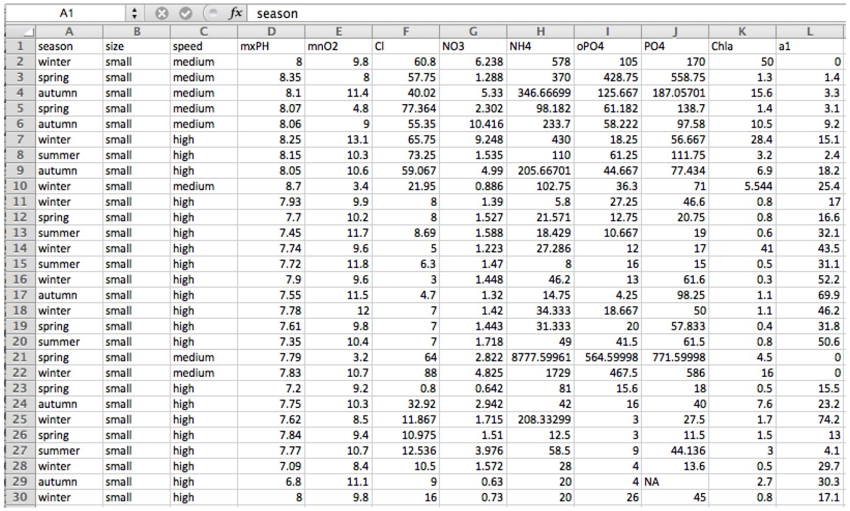
# Database vs Flat File

## Database



## Data Integrity ✓

## Flat File



A1 fx season

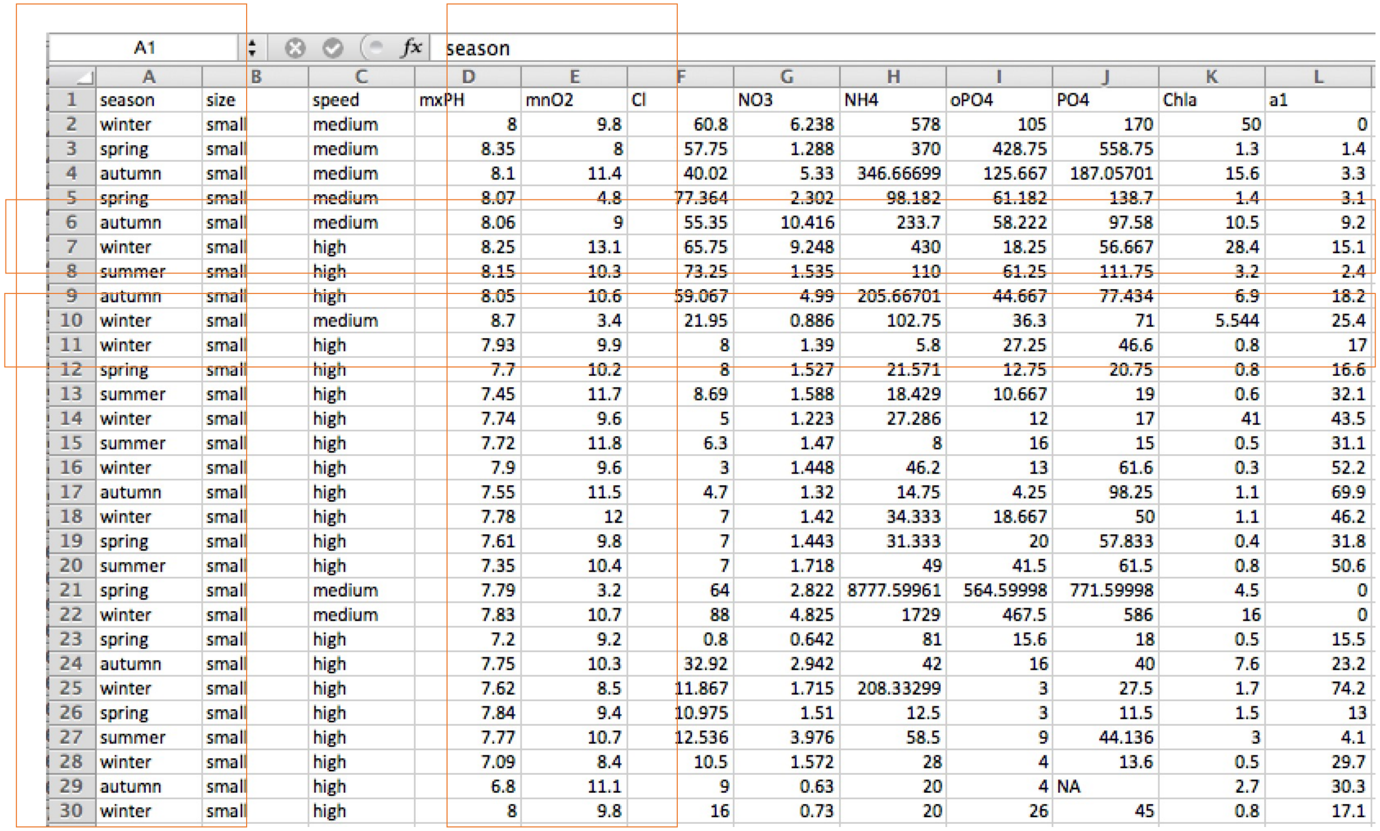
| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | season | size | speed | mxPH | mnO2 | Cl | NO3 | NH4 | oPO4 | PO4 | Chla | a1 |
| 2 | winter | small | medium | 8 | 9.8 | 60.8 | 6.238 | 578 | 105 | 170 | 50 | 0 |
| 3 | spring | small | medium | 8.35 | 8 | 57.75 | 1.288 | 370 | 428.75 | 558.75 | 1.3 | 1.4 |
| 4 | autumn | small | medium | 8.1 | 11.4 | 40.02 | 5.33 | 346.66699 | 125.667 | 187.05701 | 15.6 | 3.3 |
| 5 | spring | small | medium | 8.07 | 4.8 | 77.364 | 2.302 | 98.182 | 61.182 | 138.7 | 1.4 | 3.1 |
| 6 | autumn | small | medium | 8.06 | 9 | 55.35 | 10.416 | 233.7 | 58.222 | 97.58 | 10.5 | 9.2 |
| 7 | winter | small | high | 8.25 | 13.1 | 65.75 | 9.248 | 430 | 18.25 | 56.667 | 28.4 | 15.1 |
| 8 | summer | small | high | 8.15 | 10.3 | 73.25 | 1.535 | 110 | 61.25 | 111.75 | 3.2 | 2.4 |
| 9 | autumn | small | high | 8.05 | 10.6 | 59.067 | 4.99 | 205.66701 | 44.667 | 77.434 | 6.9 | 18.2 |
| 10 | winter | small | medium | 8.7 | 3.4 | 21.95 | 0.886 | 102.75 | 36.3 | 71 | 5.544 | 25.4 |
| 11 | winter | small | high | 7.93 | 9.9 | 8 | 1.39 | 5.8 | 27.25 | 46.6 | 0.8 | 17 |
| 12 | spring | small | high | 7.7 | 10.2 | 8 | 1.527 | 21.571 | 12.75 | 20.75 | 0.8 | 16.6 |
| 13 | summer | small | high | 7.45 | 11.7 | 8.69 | 1.588 | 18.429 | 10.667 | 19 | 0.6 | 32.1 |
| 14 | winter | small | high | 7.74 | 9.6 | 5 | 1.223 | 27.286 | 12 | 17 | 41 | 43.5 |
| 15 | summer | small | high | 7.72 | 11.8 | 6.3 | 1.47 | 8 | 16 | 15 | 0.5 | 31.1 |
| 16 | winter | small | high | 7.9 | 9.6 | 3 | 1.448 | 46.2 | 13 | 61.6 | 0.3 | 52.2 |
| 17 | autumn | small | high | 7.55 | 11.5 | 4.7 | 1.32 | 14.75 | 4.25 | 98.25 | 1.1 | 69.9 |
| 18 | winter | small | high | 7.78 | 12 | 7 | 1.42 | 34.333 | 18.667 | 50 | 1.1 | 46.2 |
| 19 | spring | small | high | 7.61 | 9.8 | 7 | 1.443 | 31.333 | 20 | 57.833 | 0.4 | 31.8 |
| 20 | summer | small | high | 7.35 | 10.4 | 7 | 1.718 | 49 | 41.5 | 61.5 | 0.8 | 50.6 |
| 21 | spring | small | medium | 7.79 | 3.2 | 64 | 2.822 | 8777.59961 | 564.59998 | 771.59998 | 4.5 | 0 |
| 22 | winter | small | medium | 7.83 | 10.7 | 88 | 4.825 | 1729 | 467.5 | 586 | 16 | 0 |
| 23 | spring | small | high | 7.2 | 9.2 | 0.8 | 0.642 | 81 | 15.6 | 18 | 0.5 | 15.5 |
| 24 | autumn | small | high | 7.75 | 10.3 | 32.92 | 2.942 | 42 | 16 | 40 | 7.6 | 23.2 |
| 25 | winter | small | high | 7.62 | 8.5 | 11.867 | 1.715 | 208.33299 | 3 | 27.5 | 1.7 | 74.2 |
| 26 | spring | small | high | 7.84 | 9.4 | 10.975 | 1.51 | 12.5 | 3 | 11.5 | 1.5 | 13 |
| 27 | summer | small | high | 7.77 | 10.7 | 12.536 | 3.976 | 58.5 | 9 | 44.136 | 3 | 4.1 |
| 28 | winter | small | high | 7.09 | 8.4 | 10.5 | 1.572 | 28 | 4 | 13.6 | 0.5 | 29.7 |
| 29 | autumn | small | high | 6.8 | 11.1 | 9 | 0.63 | 20 | 4 | NA | 2.7 | 30.3 |
| 30 | winter | small | high | 8 | 9.8 | 16 | 0.73 | 20 | 26 | 45 | 0.8 | 17.1 |

## Data Analysis ✓

# Rows vs Columns

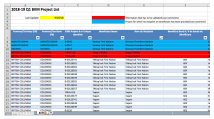
Columns contain attributes (variables, fields, etc.)

Rows contain objects*



A1 fx season

|  | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | season | size | speed | mxPH | mnO2 | Cl | NO3 | NH4 | oPO4 | PO4 | Chla | a1 |
| 2 | winter | small | medium | 8 | 9.8 | 60.8 | 6.238 | 578 | 105 | 170 | 50 | 0 |
| 3 | spring | small | medium | 8.35 | 8 | 57.75 | 1.288 | 370 | 428.75 | 558.75 | 1.3 | 1.4 |
| 4 | autumn | small | medium | 8.1 | 11.4 | 40.02 | 5.33 | 346.66699 | 125.667 | 187.05701 | 15.6 | 3.3 |
| 5 | spring | small | medium | 8.07 | 4.8 | 77.364 | 2.302 | 98.182 | 61.182 | 138.7 | 1.4 | 3.1 |
| 6 | autumn | small | medium | 8.06 | 9 | 55.35 | 10.416 | 233.7 | 58.222 | 97.58 | 10.5 | 9.2 |
| 7 | winter | small | high | 8.25 | 13.1 | 65.75 | 9.248 | 430 | 18.25 | 56.667 | 28.4 | 15.1 |
| 8 | summer | small | high | 8.15 | 10.3 | 73.25 | 1.535 | 110 | 61.25 | 111.75 | 3.2 | 2.4 |
| 9 | autumn | small | high | 8.05 | 10.6 | 59.067 | 4.99 | 205.66701 | 44.667 | 77.434 | 6.9 | 18.2 |
| 10 | winter | small | medium | 8.7 | 3.4 | 21.95 | 0.886 | 102.75 | 36.3 | 71 | 5.544 | 25.4 |
| 11 | winter | small | high | 7.93 | 9.9 | 8 | 1.39 | 5.8 | 27.25 | 46.6 | 0.8 | 17 |
| 12 | spring | small | high | 7.7 | 10.2 | 8 | 1.527 | 21.571 | 12.75 | 20.75 | 0.8 | 16.6 |
| 13 | summer | small | high | 7.45 | 11.7 | 8.69 | 1.588 | 18.429 | 10.667 | 19 | 0.6 | 32.1 |
| 14 | winter | small | high | 7.74 | 9.6 | 5 | 1.223 | 27.286 | 12 | 17 | 41 | 43.5 |
| 15 | summer | small | high | 7.72 | 11.8 | 6.3 | 1.47 | 8 | 16 | 15 | 0.5 | 31.1 |
| 16 | winter | small | high | 7.9 | 9.6 | 3 | 1.448 | 46.2 | 13 | 61.6 | 0.3 | 52.2 |
| 17 | autumn | small | high | 7.55 | 11.5 | 4.7 | 1.32 | 14.75 | 4.25 | 98.25 | 1.1 | 69.9 |
| 18 | winter | small | high | 7.78 | 12 | 7 | 1.42 | 34.333 | 18.667 | 50 | 1.1 | 46.2 |
| 19 | spring | small | high | 7.61 | 9.8 | 7 | 1.443 | 31.333 | 20 | 57.833 | 0.4 | 31.8 |
| 20 | summer | small | high | 7.35 | 10.4 | 7 | 1.718 | 49 | 41.5 | 61.5 | 0.8 | 50.6 |
| 21 | spring | small | medium | 7.79 | 3.2 | 64 | 2.822 | 8777.59961 | 564.59998 | 771.59998 | 4.5 | 0 |
| 22 | winter | small | medium | 7.83 | 10.7 | 88 | 4.825 | 1729 | 467.5 | 586 | 16 | 0 |
| 23 | spring | small | high | 7.2 | 9.2 | 0.8 | 0.642 | 81 | 15.6 | 18 | 0.5 | 15.5 |
| 24 | autumn | small | high | 7.75 | 10.3 | 32.92 | 2.942 | 42 | 16 | 40 | 7.6 | 23.2 |
| 25 | winter | small | high | 7.62 | 8.5 | 11.867 | 1.715 | 208.33299 | 3 | 27.5 | 1.7 | 74.2 |
| 26 | spring | small | high | 7.84 | 9.4 | 10.975 | 1.51 | 12.5 | 3 | 11.5 | 1.5 | 13 |
| 27 | summer | small | high | 7.77 | 10.7 | 12.536 | 3.976 | 58.5 | 9 | 44.136 | 3 | 4.1 |
| 28 | winter | small | high | 7.09 | 8.4 | 10.5 | 1.572 | 28 | 4 | 13.6 | 0.5 | 29.7 |
| 29 | autumn | small | high | 6.8 | 11.1 | 9 | 0.63 | 20 | 4 | NA | 2.7 | 30.3 |
| 30 | winter | small | high | 8 | 9.8 | 16 | 0.73 | 20 | 26 | 45 | 0.8 | 17.1 |

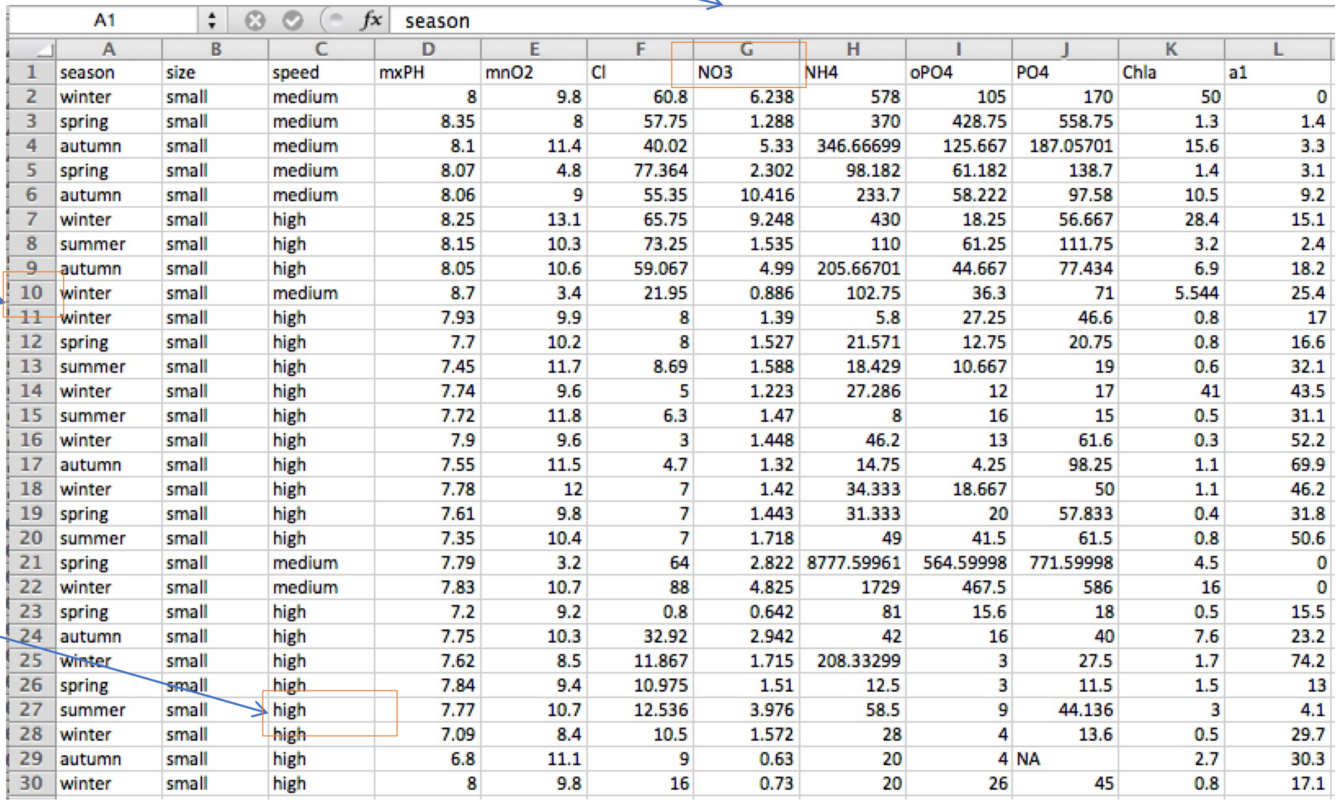


# Rows vs Columns

variable (field) name

object ID

variable (field) value (datum)

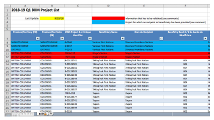


A1 *fx* season

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | season | size | speed | mxPH | mnO2 | Cl | NO3 | NH4 | oPO4 | PO4 | Chla | a1 |
| 2 | winter | small | medium | 8 | 9.8 | 60.8 | 6.238 | 578 | 105 | 170 | 50 | 0 |
| 3 | spring | small | medium | 8.35 | 8 | 57.75 | 1.288 | 370 | 428.75 | 558.75 | 1.3 | 1.4 |
| 4 | autumn | small | medium | 8.1 | 11.4 | 40.02 | 5.33 | 346.66699 | 125.667 | 187.05701 | 15.6 | 3.3 |
| 5 | spring | small | medium | 8.07 | 4.8 | 77.364 | 2.302 | 98.182 | 61.182 | 138.7 | 1.4 | 3.1 |
| 6 | autumn | small | medium | 8.06 | 9 | 55.35 | 10.416 | 233.7 | 58.222 | 97.58 | 10.5 | 9.2 |
| 7 | winter | small | high | 8.25 | 13.1 | 65.75 | 9.248 | 430 | 18.25 | 56.667 | 28.4 | 15.1 |
| 8 | summer | small | high | 8.15 | 10.3 | 73.25 | 1.535 | 110 | 61.25 | 111.75 | 3.2 | 2.4 |
| 9 | autumn | small | high | 8.05 | 10.6 | 59.067 | 4.99 | 205.66701 | 44.667 | 77.434 | 6.9 | 18.2 |
| 10 | winter | small | medium | 8.7 | 3.4 | 21.95 | 0.886 | 102.75 | 36.3 | 71 | 5.544 | 25.4 |
| 11 | winter | small | high | 7.93 | 9.9 | 8 | 1.39 | 5.8 | 27.25 | 46.6 | 0.8 | 17 |
| 12 | spring | small | high | 7.7 | 10.2 | 8 | 1.527 | 21.571 | 12.75 | 20.75 | 0.8 | 16.6 |
| 13 | summer | small | high | 7.45 | 11.7 | 8.69 | 1.588 | 18.429 | 10.667 | 19 | 0.6 | 32.1 |
| 14 | winter | small | high | 7.74 | 9.6 | 5 | 1.223 | 27.286 | 12 | 17 | 41 | 43.5 |
| 15 | summer | small | high | 7.72 | 11.8 | 6.3 | 1.47 | 8 | 16 | 15 | 0.5 | 31.1 |
| 16 | winter | small | high | 7.9 | 9.6 | 3 | 1.448 | 46.2 | 13 | 61.6 | 0.3 | 52.2 |
| 17 | autumn | small | high | 7.55 | 11.5 | 4.7 | 1.32 | 14.75 | 4.25 | 98.25 | 1.1 | 69.9 |
| 18 | winter | small | high | 7.78 | 12 | 7 | 1.42 | 34.333 | 18.667 | 50 | 1.1 | 46.2 |
| 19 | spring | small | high | 7.61 | 9.8 | 7 | 1.443 | 31.333 | 20 | 57.833 | 0.4 | 31.8 |
| 20 | summer | small | high | 7.35 | 10.4 | 7 | 1.718 | 49 | 41.5 | 61.5 | 0.8 | 50.6 |
| 21 | spring | small | medium | 7.79 | 3.2 | 64 | 2.822 | 8777.59961 | 564.59998 | 771.59998 | 4.5 | 0 |
| 22 | winter | small | medium | 7.83 | 10.7 | 88 | 4.825 | 1729 | 467.5 | 586 | 16 | 0 |
| 23 | spring | small | high | 7.2 | 9.2 | 0.8 | 0.642 | 81 | 15.6 | 18 | 0.5 | 15.5 |
| 24 | autumn | small | high | 7.75 | 10.3 | 32.92 | 2.942 | 42 | 16 | 40 | 7.6 | 23.2 |
| 25 | winter | small | high | 7.62 | 8.5 | 11.867 | 1.715 | 208.33299 | 3 | 27.5 | 1.7 | 74.2 |
| 26 | spring | small | high | 7.84 | 9.4 | 10.975 | 1.51 | 12.5 | 3 | 11.5 | 1.5 | 13 |
| 27 | summer | small | high | 7.77 | 10.7 | 12.536 | 3.976 | 58.5 | 9 | 44.136 | 3 | 4.1 |
| 28 | winter | small | high | 7.09 | 8.4 | 10.5 | 1.572 | 28 | 4 | 13.6 | 0.5 | 29.7 |
| 29 | autumn | small | high | 6.8 | 11.1 | 9 | 0.63 | 20 | 4 | NA | 2.7 | 30.3 |
| 30 | winter | small | high | 8 | 9.8 | 16 | 0.73 | 20 | 26 | 45 | 0.8 | 17.1 |

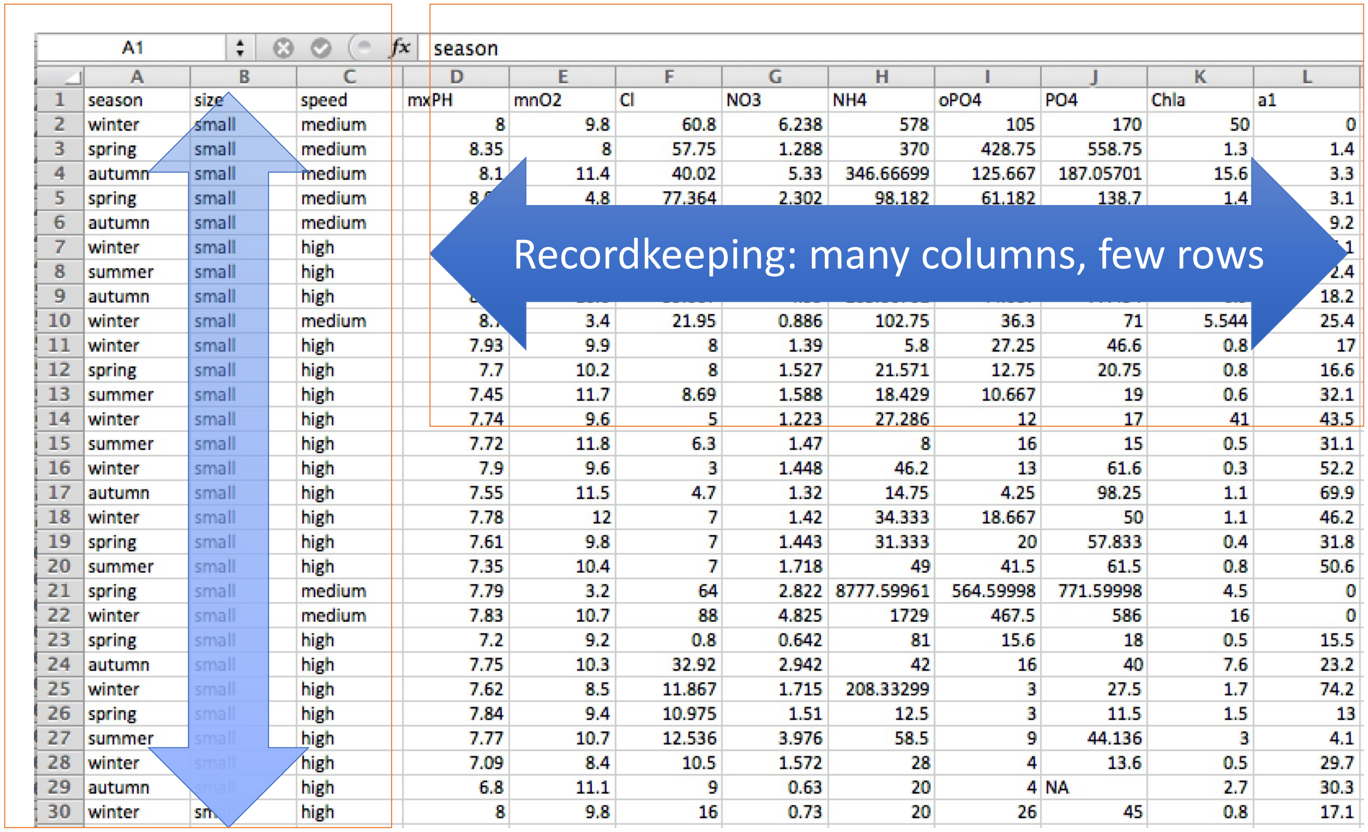
**Record-keeping**

**Research**



# Dataset Shape and Focus

Research: many rows, few columns



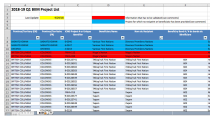
Recordkeeping: many columns, few rows

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | season | size | speed | mxPH | mnO2 | Cl | NO3 | NH4 | oPO4 | PO4 | Chla | a1 |
| 2 | winter | small | medium | 8 | 9.8 | 60.8 | 6.238 | 578 | 105 | 170 | 50 | 0 |
| 3 | spring | small | medium | 8.35 | 8 | 57.75 | 1.288 | 370 | 428.75 | 558.75 | 1.3 | 1.4 |
| 4 | autumn | small | medium | 8.1 | 11.4 | 40.02 | 5.33 | 346.66699 | 125.667 | 187.05701 | 15.6 | 3.3 |
| 5 | spring | small | medium | [illegible] | 4.8 | 77.364 | 2.302 | 98.182 | 61.182 | 138.7 | 1.4 | 3.1 |
| 6 | autumn | small | medium | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 9.2 |
| 7 | winter | small | high | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| 8 | summer | small | high | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 2.4 |
| 9 | autumn | small | high | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | 18.2 |
| 10 | winter | small | medium | [illegible] | 3.4 | 21.95 | 0.886 | 102.75 | 36.3 | 71 | 5.544 | 25.4 |
| 11 | winter | small | high | 7.93 | 9.9 | 8 | 1.39 | 5.8 | 27.25 | 46.6 | 0.8 | 17 |
| 12 | spring | small | high | 7.7 | 10.2 | 8 | 1.527 | 21.571 | 12.75 | 20.75 | 0.8 | 16.6 |
| 13 | summer | small | high | 7.45 | 11.7 | 8.69 | 1.588 | 18.429 | 10.667 | 19 | 0.6 | 32.1 |
| 14 | winter | small | high | 7.74 | 9.6 | 5 | 1.223 | 27.286 | 12 | 17 | 41 | 43.5 |
| 15 | summer | small | high | 7.72 | 11.8 | 6.3 | 1.47 | 8 | 16 | 15 | 0.5 | 31.1 |
| 16 | winter | small | high | 7.9 | 9.6 | 3 | 1.448 | 46.2 | 13 | 61.6 | 0.3 | 52.2 |
| 17 | autumn | small | high | 7.55 | 11.5 | 4.7 | 1.32 | 14.75 | 4.25 | 98.25 | 1.1 | 69.9 |
| 18 | winter | small | high | 7.78 | 12 | 7 | 1.42 | 34.333 | 18.667 | 50 | 1.1 | 46.2 |
| 19 | spring | small | high | 7.61 | 9.8 | 7 | 1.443 | 31.333 | 20 | 57.833 | 0.4 | 31.8 |
| 20 | summer | small | high | 7.35 | 10.4 | 7 | 1.718 | 49 | 41.5 | 61.5 | 0.8 | 50.6 |
| 21 | spring | small | medium | 7.79 | 3.2 | 64 | 2.822 | 8777.59961 | 564.59998 | 771.59998 | 4.5 | 0 |
| 22 | winter | small | medium | 7.83 | 10.7 | 88 | 4.825 | 1729 | 467.5 | 586 | 16 | 0 |
| 23 | spring | small | high | 7.2 | 9.2 | 0.8 | 0.642 | 81 | 15.6 | 18 | 0.5 | 15.5 |
| 24 | autumn | small | high | 7.75 | 10.3 | 32.92 | 2.942 | 42 | 16 | 40 | 7.6 | 23.2 |
| 25 | winter | small | high | 7.62 | 8.5 | 11.867 | 1.715 | 208.33299 | 3 | 27.5 | 1.7 | 74.2 |
| 26 | spring | small | high | 7.84 | 9.4 | 10.975 | 1.51 | 12.5 | 3 | 11.5 | 1.5 | 13 |
| 27 | summer | small | high | 7.77 | 10.7 | 12.536 | 3.976 | 58.5 | 9 | 44.136 | 3 | 4.1 |
| 28 | winter | small | high | 7.09 | 8.4 | 10.5 | 1.572 | 28 | 4 | 13.6 | 0.5 | 29.7 |
| 29 | autumn | small | high | 6.8 | 11.1 | 9 | 0.63 | 20 | 4 | NA | 2.7 | 30.3 |
| 30 | winter | small | high | 8 | 9.8 | 16 | 0.73 | 20 | 26 | 45 | 0.8 | 17.1 |

**Record-keeping**

✥

**Research**



# Data Preparation for Analysis

Validating, Cleaning, Augmenting, Transforming

# Data Preparation

- Data validation + verification
- Data cleaning
- Data transformation
- (Data Exploration?)

# Data Preparation

- Data validation + verification
- Data cleaning
- Data transformation
- (Data Exploration?)

Each of these steps may themselves involve data analysis and other techniques

# Data Validation + Verification

- **Verification**: Confirm that the data is correct relative to the dataset
- **Validation**: Confirm that the data correctly represents the objects
- We determine data cleaning requirements based on the results of our data verification and validation

[3, 10.43, ROUn, golden delicious]

# Data Cleaning

A question for you: **should you clean before you do exploratory analysis?**

Some possible issues:

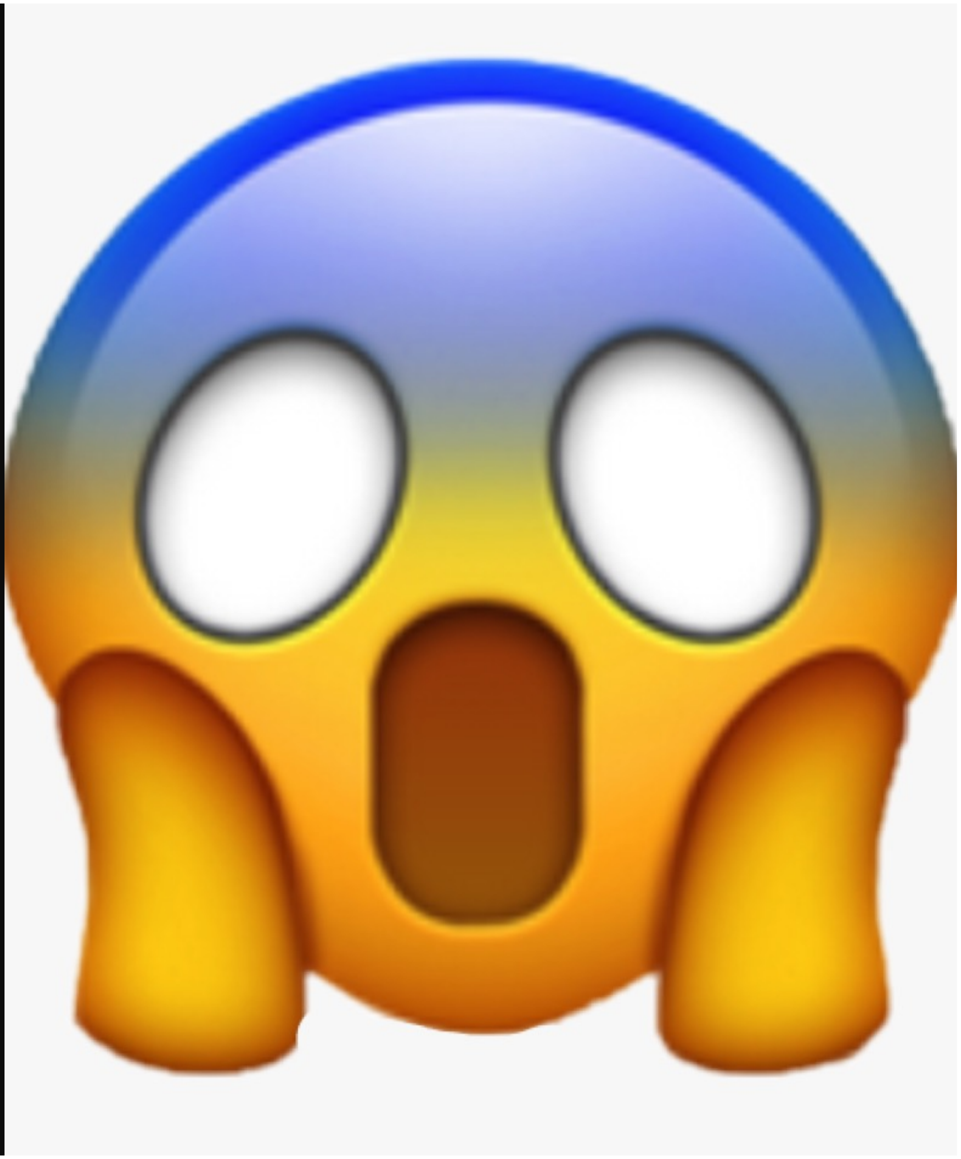
- Character encodings
- Missing Data
- Data collection or entry errors
- Systematic errors

# The Curse of Free Text Fields

- The curse of categorical data is made much worse by the curse of free text fields

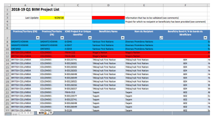
- If you have a field that is supposed to be categorical but it is a free text field, **it is no longer categorical**

- You can use machine learning techniques to help to some extent, but this is a case **where an ounce or prevention is worth a pound of cure**.



# Data Cleaning Bingo

| | | | | |
|---|---|---|---|---|
| random missing values | outliers | values outside of expected range - numeric | factors incorrectly/iconsistently coded | date/time values in multiple formats |
| impossible numeric values | leading or trailing white space | badly formatted date/time values | non-random missing values | logical inconsistencies across fields |
| characters in numeric field | values outside of expected range - date/time | DCB! | inconsistent or no distinction between null, 0,not available, not applicable,missing | possible factors missing |
| multiple symbols used for missing values | ??? | fields incorrectly separated in row | blank fields | logical iconsistencies within field |
| entire blank rows | character encoding issues | duplicate value in unique field | non-factor values in factor | numeric values in character field |



2018-19 Q1 BIIM Project List

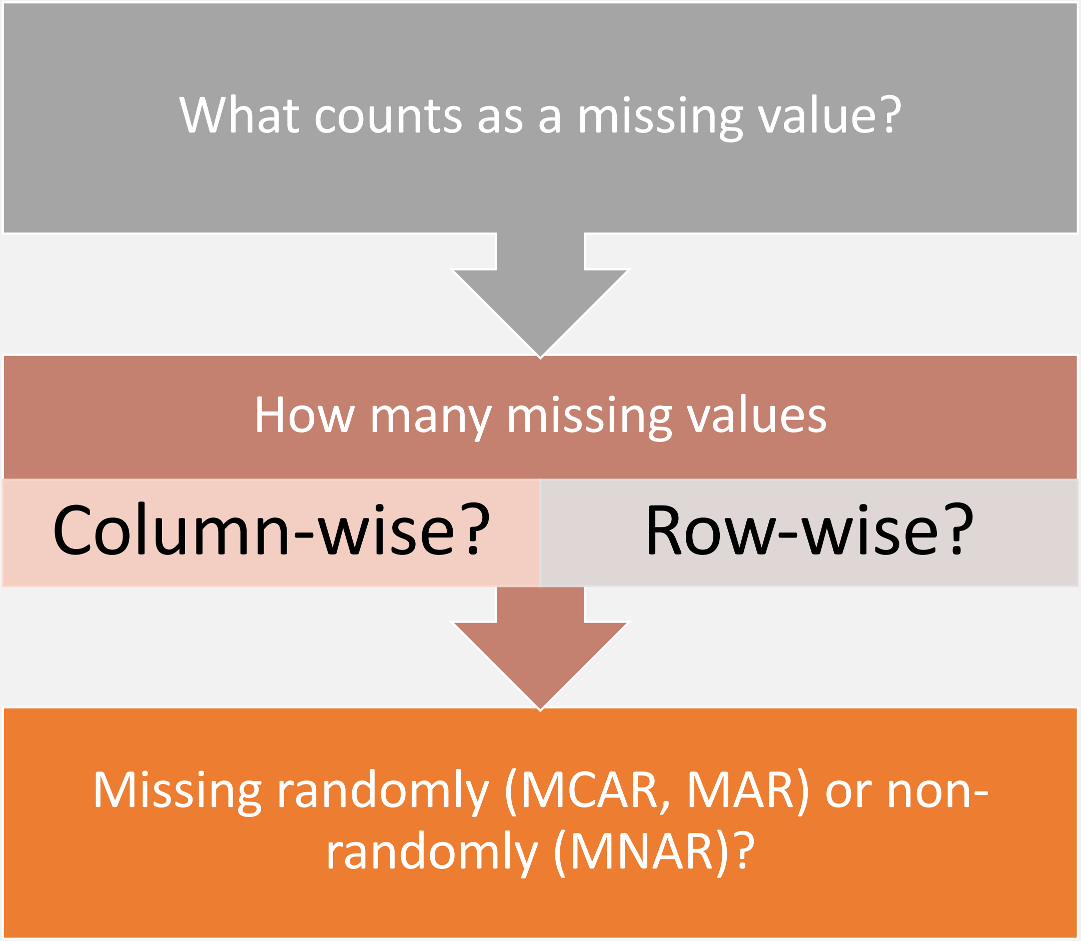
Last Update: 9/29/18

🟥 Information that has to be validated (see comments)

🟦 Project for which no recipeint or beneficiairy has been provided (see comment)

| Province/Territory (EN) | Province/Territoire (FR) | ICMS Project # or Unique Identifier | Beneficiary Name | Nom du Recipient | Beneficiry Band # / # de bande du bénéficiare |
|---|---|---|---|---|---|
| SASKATCHEWAN | SASKATCHEWAN | 6-0346 | Various First Nations | Diverses Premières Nations | |
| SASKATCHEWAN | SASKATCHEWAN | 6-0347 | Various First Nations | Diverses Premières Nations | |
| ONTARIO | ONTARIO | 4-0243 | Various First Nations | Diverses Premières Nations | |
| BRITISH COLUMBIA | COLOMBIE- | FNHA-112 | Nisga'a Nation | Nisga'a Nation | |
| BRITISH COLUMBIA | COLOMBIE- | FNHA-113 | Nisga'a Nation | Nisga'a Nation | |
| BRITISH COLUMBIA | COLOMBIE- | 9-00125741 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126301 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126302 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126303 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126348 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126349 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126925 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126356 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126357 | ʔAkisq'nuk First Nation | ?Akisq'nuk First Nation | 604 |
| BRITISH COLUMBIA | COLOMBIE- | FNHA-013 | ʔaqam | ?aqam | 602 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00123377 | ʔaqam | ?aqam | 602 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00125741 | ʔaqam | ?aqam | 602 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126348 | ʔaqam | ?aqam | 602 |
| BRITISH COLUMBIA | COLOMBIE- | 9-00126349 | ʔaqam | ?aqam | 602 |
| BRITISH COLUMBIA | COLOMBIE- | 9-0126 | ʔaqam | ?aqam | 602 |

# Cleaning: Missing Values



What counts as a missing value?

How many missing values

Column-wise?

Row-wise?

Missing randomly (MCAR, MAR) or non-randomly (MNAR)?

# Dealing with Missing Values

If percentage is very low (e.g. <= 5%) you might be able to just ignore those rows*

You can try to detect if the data is MNAR instead of MCAR/MAR using statistical tests

If missing values are MCAR/MAR you might be able to ignore them

You might be able to ‘impute’ the data using statistical modelling techniques

# MCAR, MAR, MNAR

**Missing Completely At Random (MCAR):** Genuinely no pattern to the missing values (think "due to sunspots"

**Missing At Random (MAR):** Missing values are correlated with another variable you also have.

**Missing Not At Random (MNAR**): Missing values are correlated with another variable you **don't** have

Interesting example – fields where people can select "Choose not to reply"

When does imputation make sense?

# Cleaning: Other Data Entry Errors

**Syntax errors**: Capitalization, misspellings

**Heaping**: people tend to round off measurement values (e.g. hours worked). This results in the data showing up in ‘heaps’

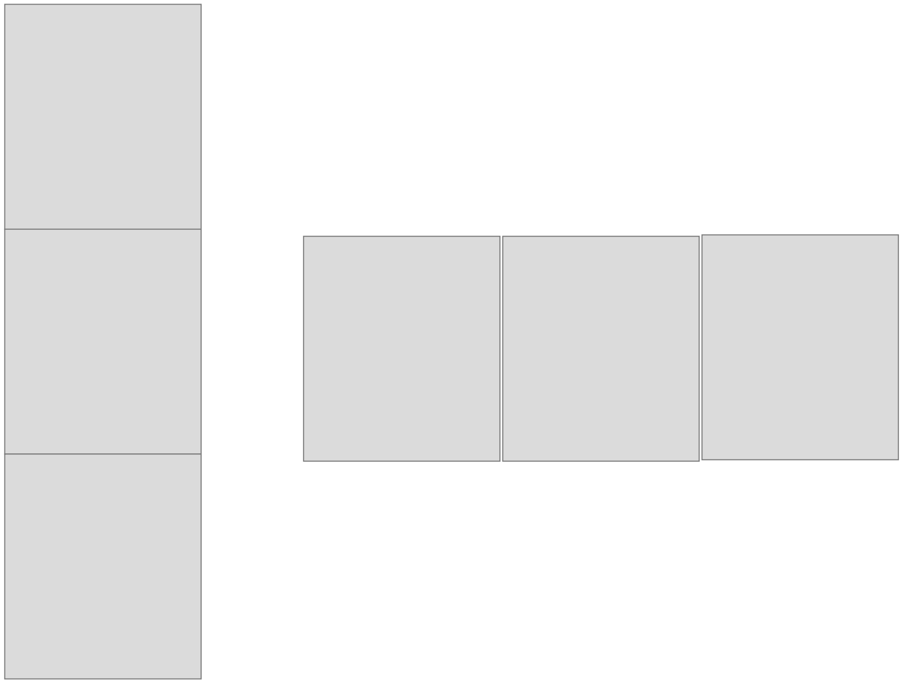
**Collector bias, sensor error**: recording what is expected rather than what is, dealing with badly calibrated sensor

# Transforming Data:

- Changing focus
- Summarizing, condensing
- Reshaping
- Adding complexity and abstraction (metrics)

# Long vs Wide Format

- A flat file with the same data can be structured in two shapes:
  - Long (Narrow) (Tall)(Stacked)
  - Wide (Unstacked)
- **Different analysis *algorithms* require particular shapes**
- Presentation of data



# Long Format to Wide Format

long

| Group# | Group-Size | Status-Check-Time |
|---|---|---|
| 1 | 14 | START |
| 1 | 12 | MIDDLE |
| 1 | 13 | END |
| 2 | 20 | START |
| 2 | 5 | MIDDLE |
| 2 | 6 | END |
| 3 | 6 | START |
| 3 | 8 | MIDDLE |
| 3 | 10 | END |

variable name

variable values

variable name
+ values

wide

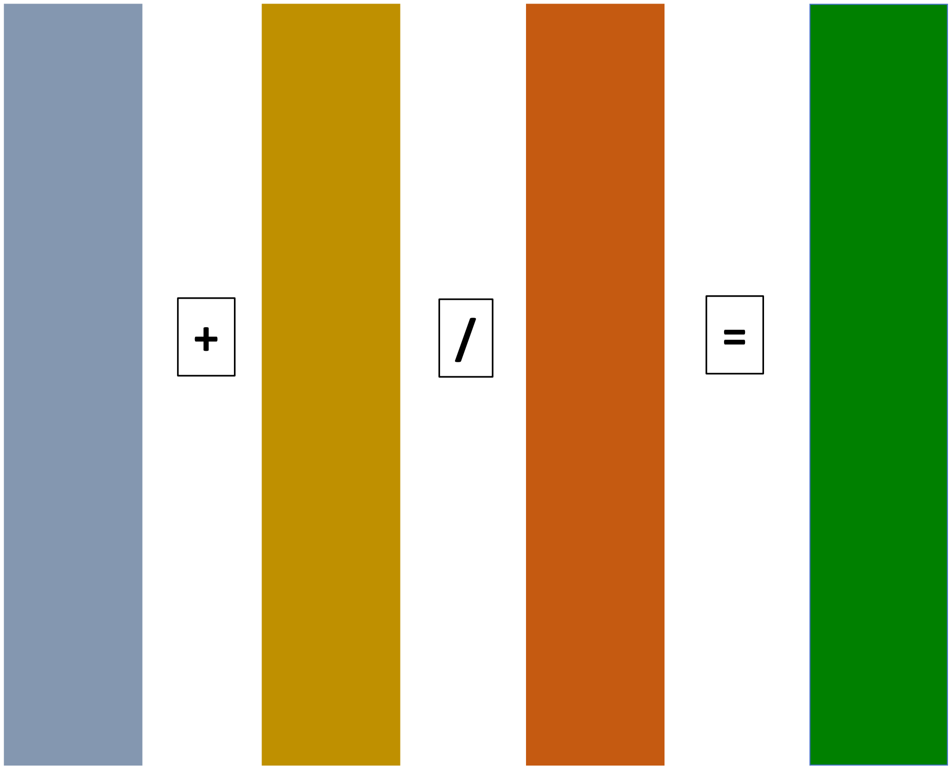
| Group# | Group-Size-START | Group-Size-MIDDLE | Group-Size-END |
|---|---|---|---|
| 1 | 14 | 12 | 13 |
| 2 | 20 | 5 | 6 |
| 3 | 6 | 8 | 10 |

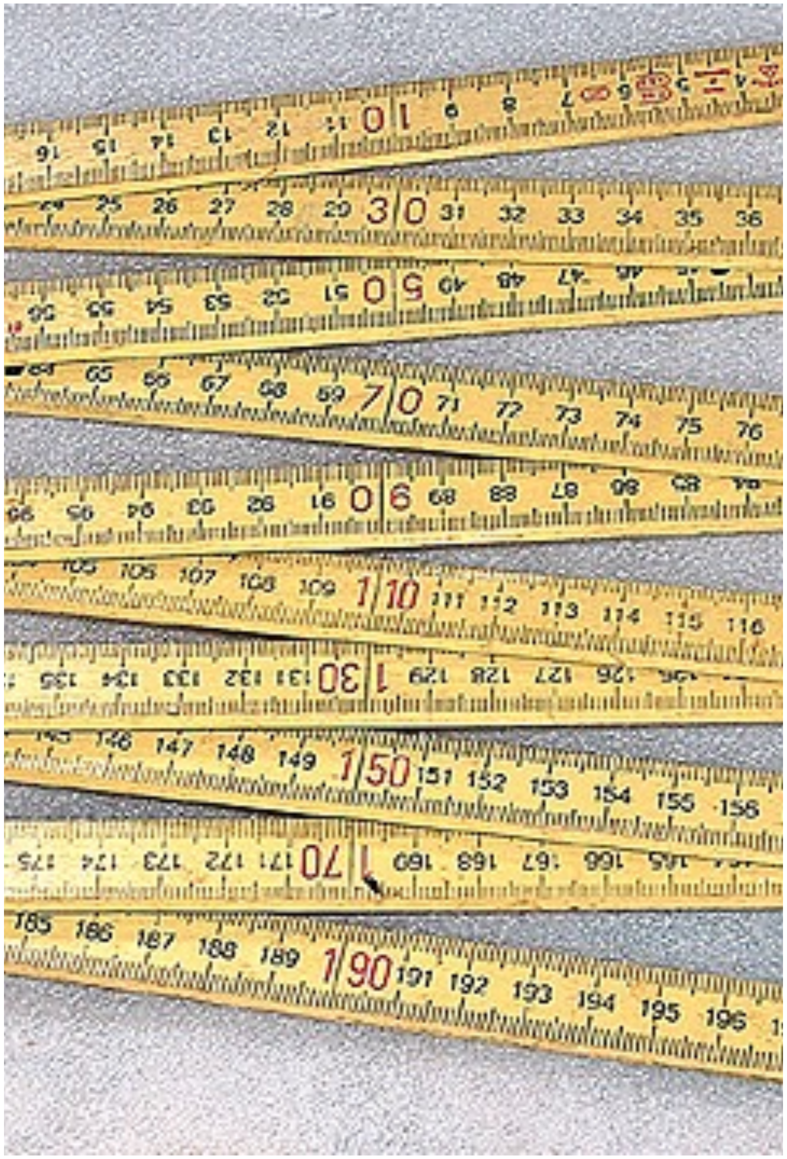
# Reshaping Data: Tools

- Reshaping your flat file by hand (or in Excel) can be *extremely* tedious! And error prone!
- This is where tools like R can be extremely helpful and time saving
- Plus – automation. Resist the 'manual' short cut!

# Adding Complexity: Metrics

- Measures:
  - Concrete properties
  - come from taking measurements
- Metrics:
  - Built up out of measures
  - Quantifies a more abstract concept





# Metrics: Good, Bad, Ugly

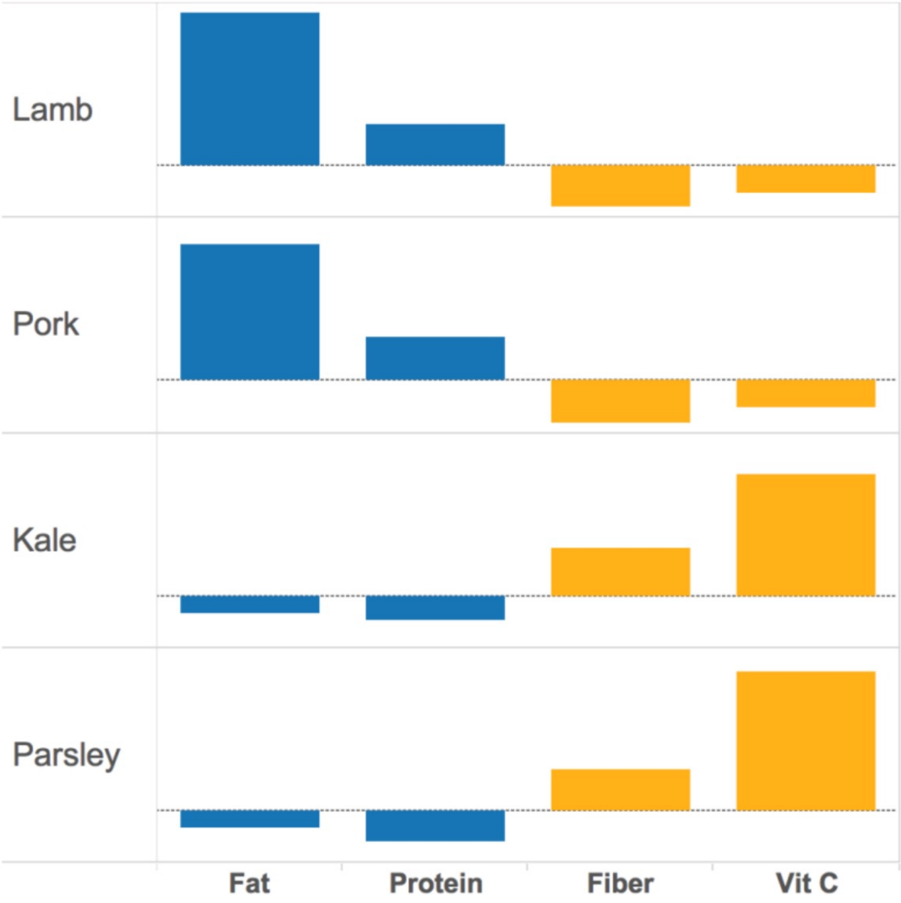
- “When a measure becomes a target, it ceases to be a good measure” ***Goodhart’s Law***
- “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor.” ***Campbell’s Law***

***(Surgeons Example)***

# Data Reduction: Principal Components Analysis (PCA)*

- In this example, presence of nutrients appears to be correlated among food items.
- In the (small) sample consisting of Lamb, Pork, Kale, and Parsley, *Fat* and *Protein* levels seem in step, as do *Fiber* and *Vitamin C*.
- In a larger dataset, the correlations are $r = 0.56$ and $r = 0.57$.
- How much could 2 variables explain?

* For categorical variables see also: MCA, FAMD
(https://drbulu.github.io/blog/factorial_methods_part1_overview/)



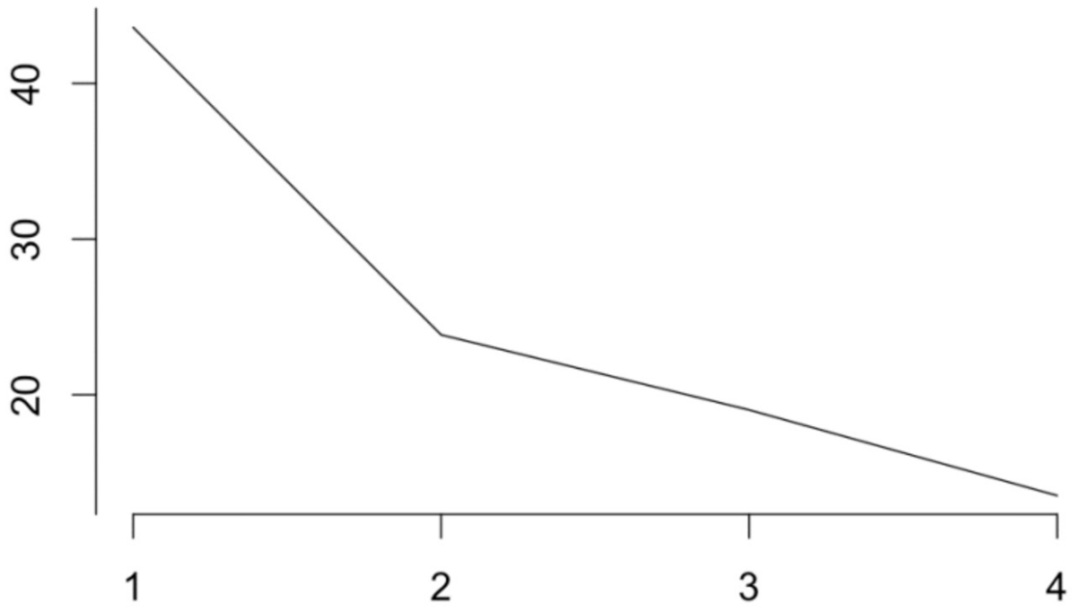
[A. Ng, K. Soo, *Numsense!*, USDA data]

# Retaining Principal Components

- The **proportion of the spread** in the data which can be explained by each principal component is shown in the scree plot.

- How many PCs are retained in the analysis?
  - keep the PCs where the cumulative proportion is below some threshold
  - keep the PCs leading to a kink

- Here, 2 PCs ≈ 68% of the spread.



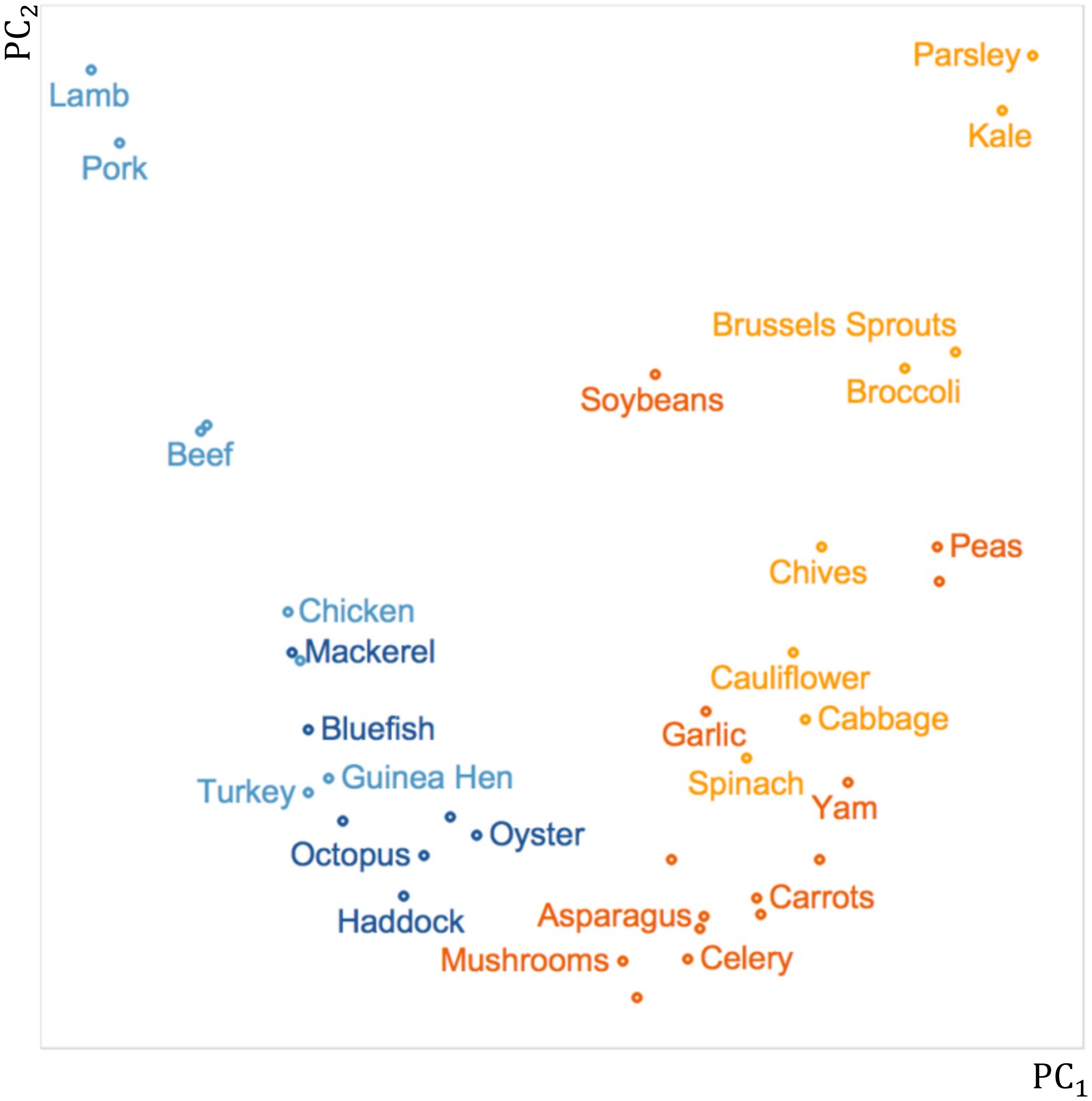
[A. Ng, K. Soo, *Numsense!*, USDA data]

$PC_1$ differentiates meats from vegetables

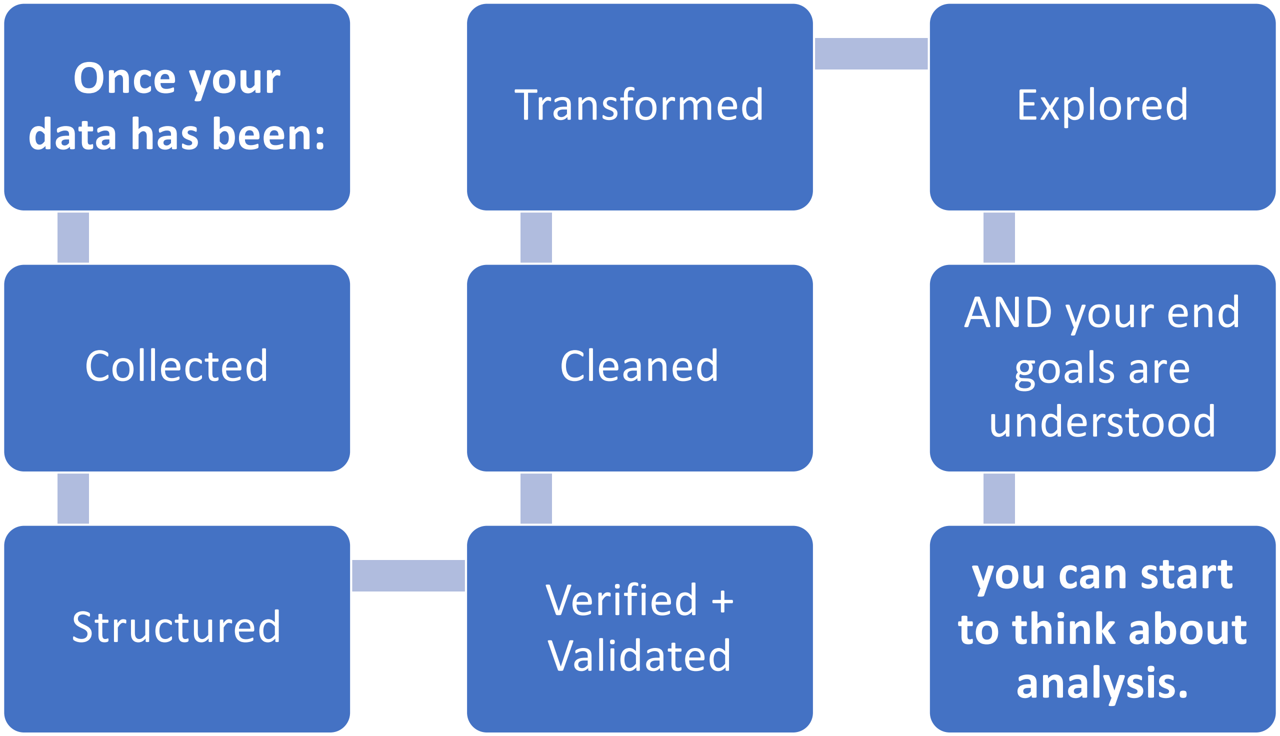
$PC_2$ differentiates **sub-categories** within meats (using *Fat*) and vegetables (using *Vitamin C*).

- Meats are concentrated on the left (low $PC_1$ values).
- Vegetables are concentrated on the right (high $PC_1$ values).
- Seafood has lower *Fat* content (low $PC_2$ values) and is concentrated at the bottom.
- Non-leafy veggies have lower *Vitamin C* content (low $PC_2$ values) and are also bunched at the bottom.

[A. Ng, K. Soo, *Numsense!*, USDA data]



# Are we there yet?



**Once your data has been:**

Collected

Structured

Verified + Validated

Cleaned

Transformed

Explored

AND your end goals are understood

**you can start to think about analysis.**