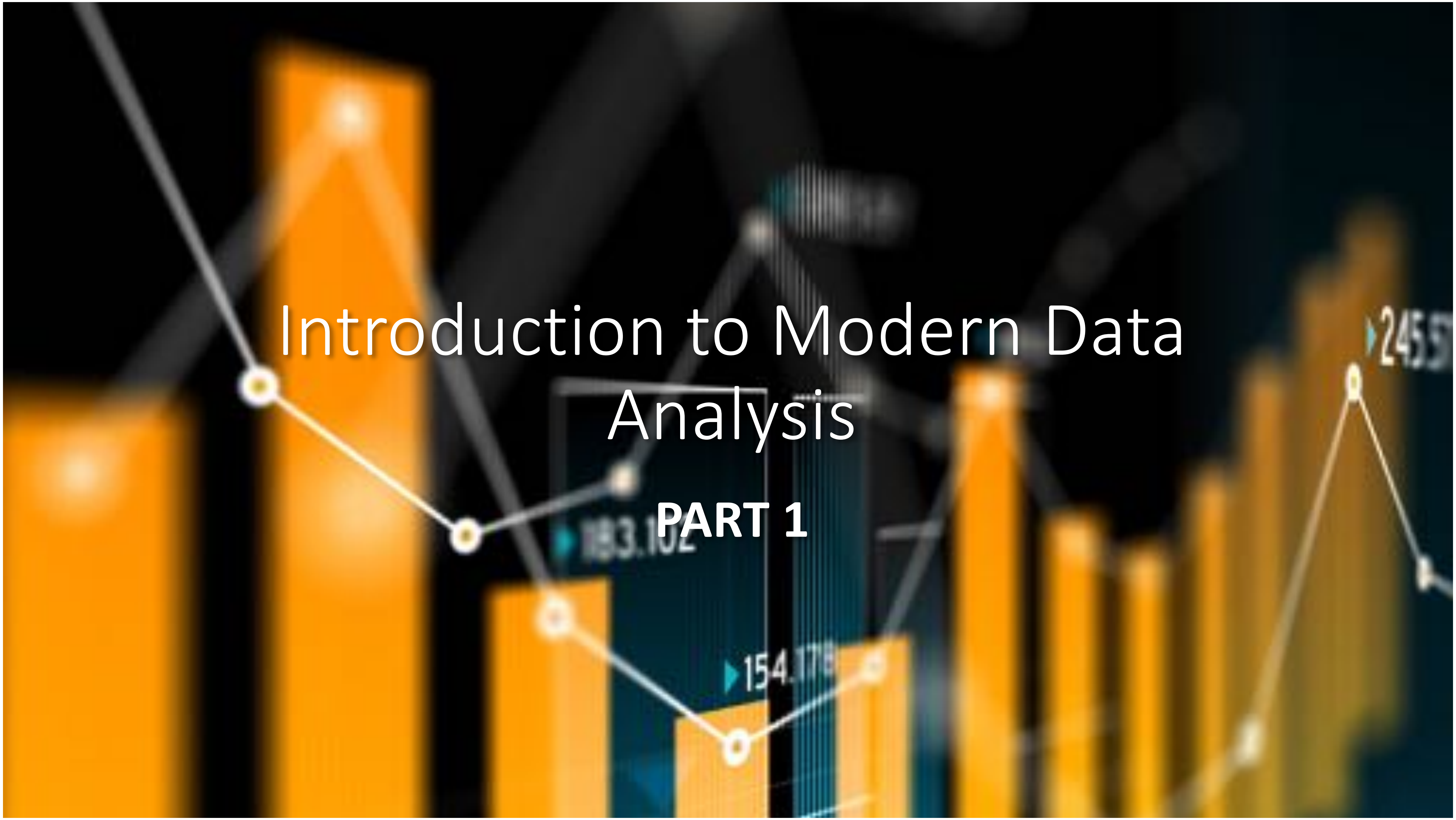


Introduction to Modern Data Analysis

PART 1

Workshop Outline

Introduction

Modern Data Analysis Teams and Technologies

- Modern Teams
- Modern Technologies
- Data Preparation

Analysis

- Machine Learning vs Statistics vs Business Intelligence
- Business Intelligence
- Machine Learning/AI
    - Intro
    - A quick tools discussion
    - Relevant Techniques

Analysis (Cont.)

- Statistics
    - A very quick tools discussion
    - Modern Statistics – Controversies and Conversations
    - Your Data, Your Questions
    - Some Relevant Statistical Concepts and Techniques

Introduction

Armchair analysis

What is (data) analysis?

Some possible answers

Finding patterns in data

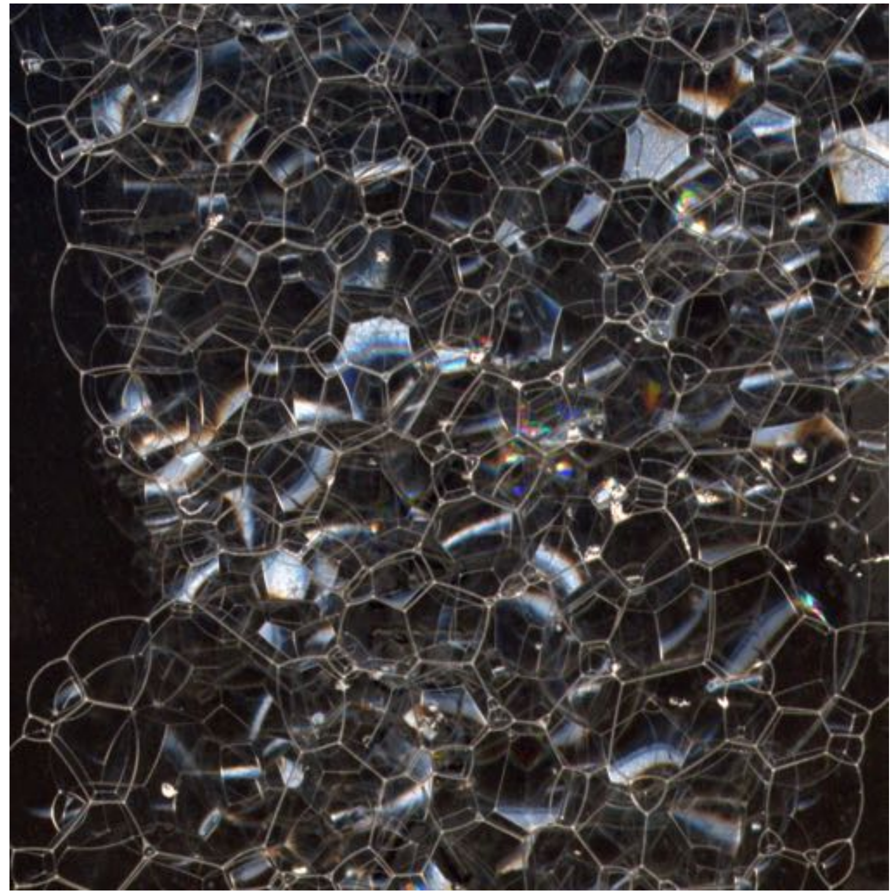
Using data to do something (answer a question, help decision-making, predict the future, knowledge discovery)

Describing or explaining your situation (your **system**)

Creating models of your data

(Testing (scientific) hypotheses?)

(Carrying out calculations on data?)



The more complicated the pattern, the more complicated the analysis.

- Typically we want to gain insight into a past, current, possible or general situation.
- For example: grant application situation, grant awarding situation
- We want to be able to: answer questions, describe what happens, explain why it happens, gain new knowledge about the situation.
- More formally: **analysis + synthesis**. A technique used for thousands of years to gain insight into our experiences.

Formal Reasoning Techniques

INDUCTIVE (INFERENTIAL), DEDUCTIVE, ABDUCTIVE, ANALOGICAL REASONING

FURTHER SPECIALIZED TECHNIQUES: THE SCIENTIFIC METHOD, STATISTICAL REASONING, MATHEMATICAL AND COMPUTER MODELLING.

EVIDENCE BASED ANALYSIS. EVIDENCE BASED ANALYSIS MAY BE MORE MORE OR LESS TECHNICAL.

Rise of analysis?

Statistics and other formal techniques (late 1800s): techniques to draw rigorous conclusions from (scientific) data to support research

Information Age (1990s): automated, electronic data collection

late 1800s — **1970s** — **1990s** — **now**

Computer Age (1970s): computers move into the public sphere

Where are we now?

Pre-Digital Age vs The Digital Age

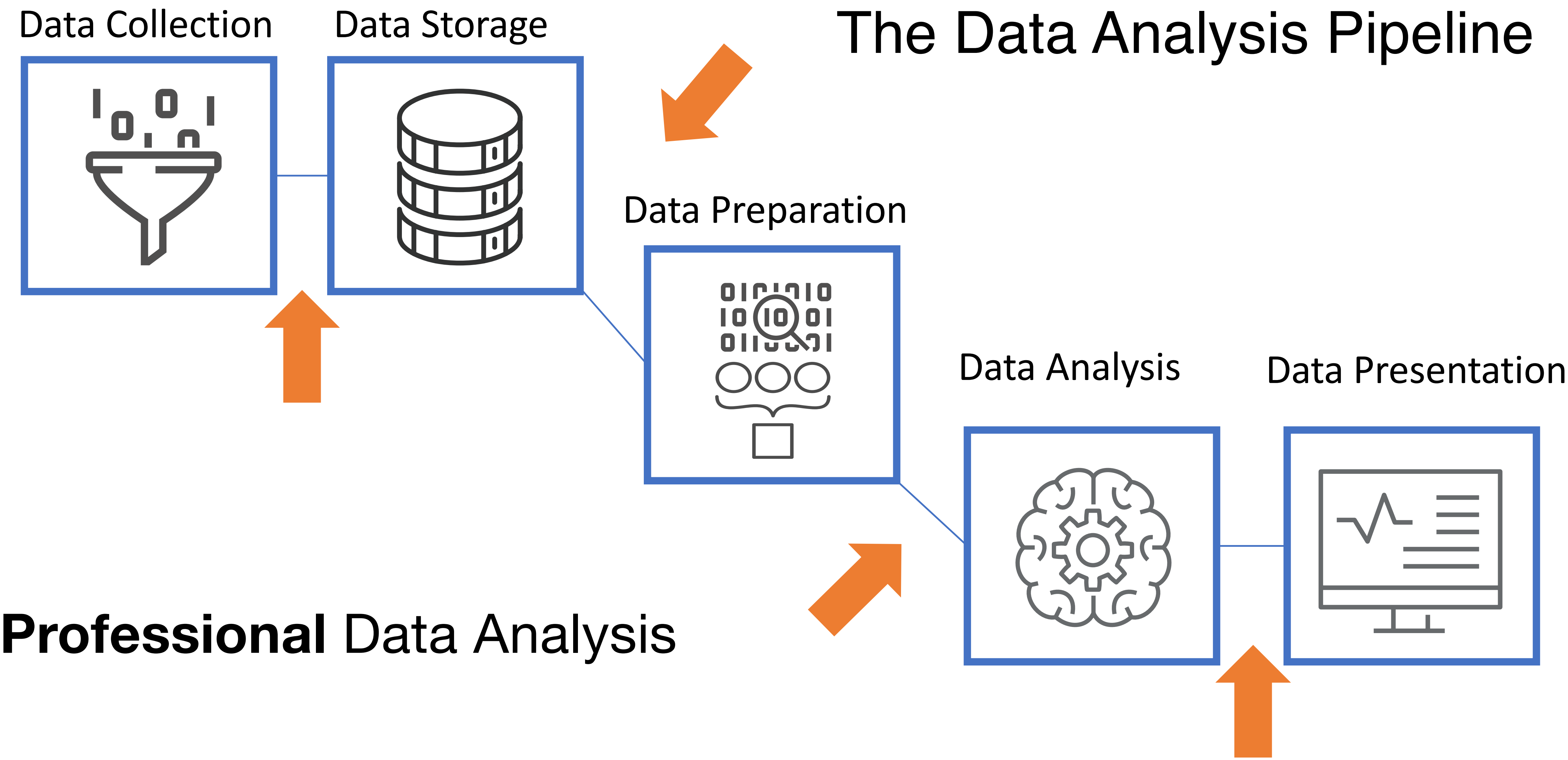
Then: Only people could carry out the activity of analysis and the components of an analysis process

Now: We can distill the essence of an analysis process into an algorithm, and automate the activity of analysis and its supporting process. We have analysis machines.

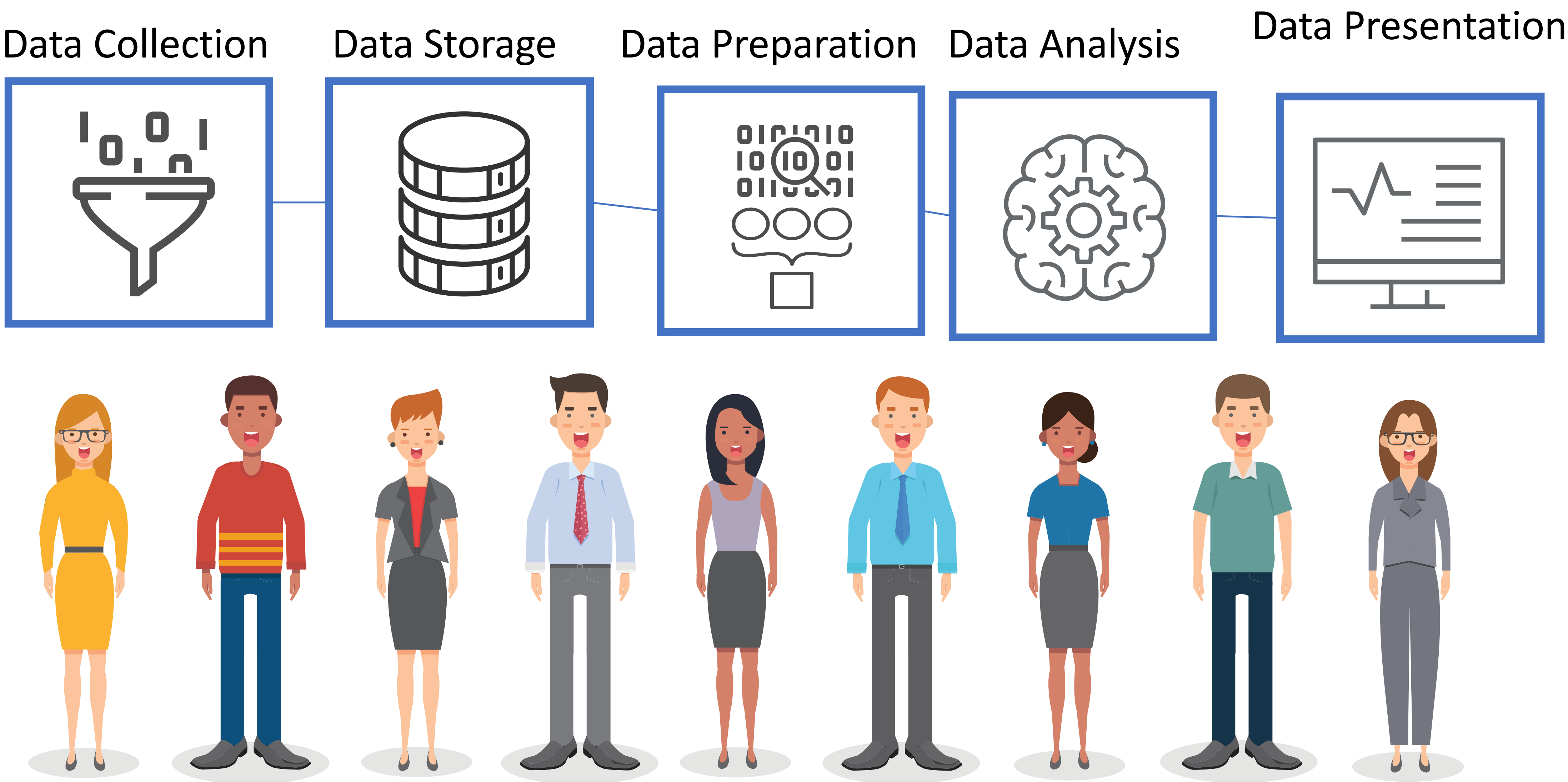
Then: A given analysis of a situation was typically seen as a one-time, one-off activity. A single person might carry out 'an analysis' and then move on.

Now: We can expect that we will probably want to repeat variations of the same analysis over and over again on new data that is streaming in on a regular basis

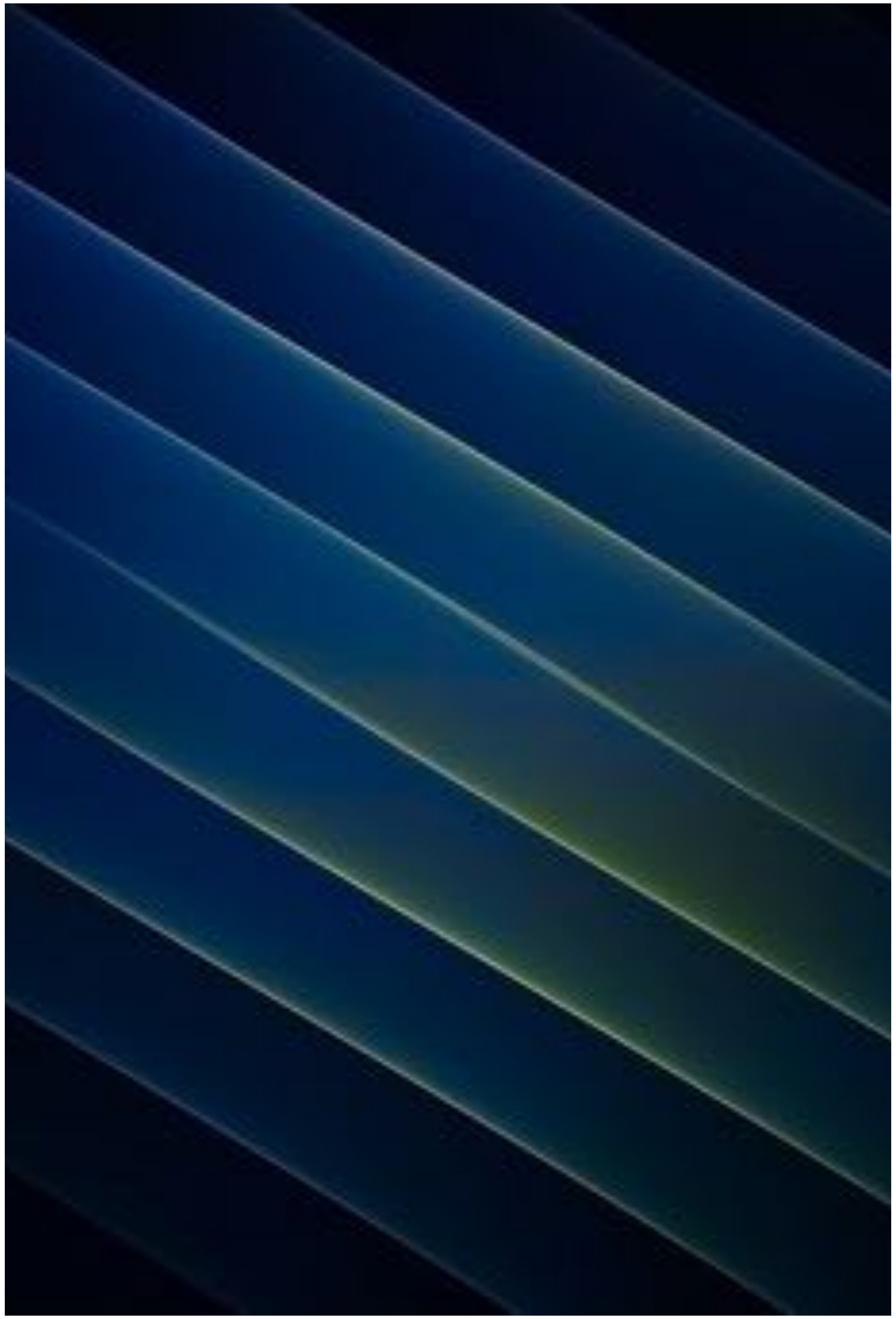
The Data Analysis Pipeline



Professional Data Analysis



Modern Data Analysis Is A Team Sport



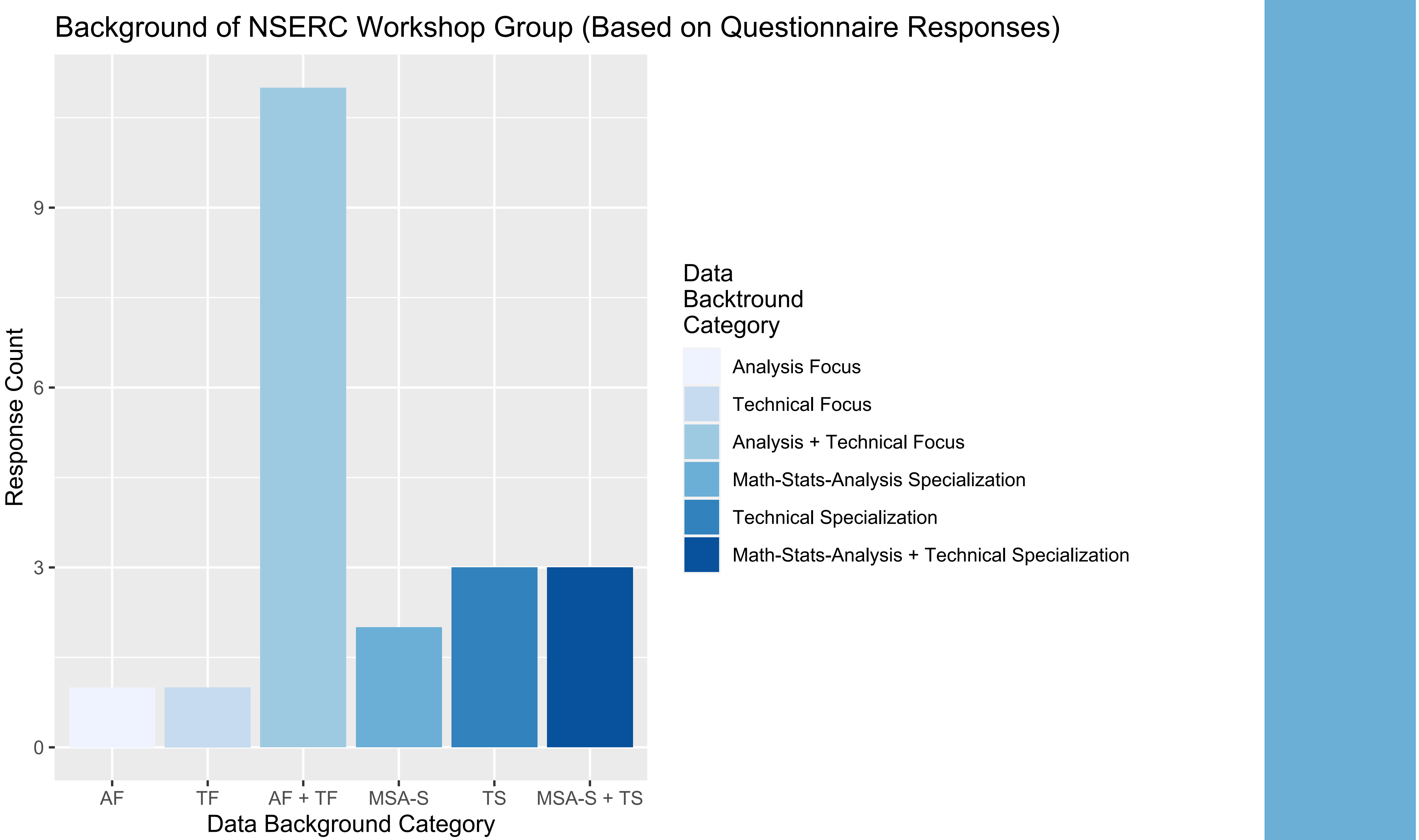
Goals For Today

Orient	Orient you towards modern data analysis
Build	Build a picture of the modern analysis landscape - what can modern data analysis DO
Understand	Understand how your work goals can be supported by different aspects of the data analysis landscape
Understand	Understand where your current interests and skillsets position you and your team within this landscape
Gain	Gain a sense of the gaps that might exist between where you are now and where you need to be to achieve analysis goals
Understand	Understand what next steps you need to take to bridge the gap, personally and in a team context
Gain	Gain awareness of resources you can draw on to take the next steps to achieve your goals

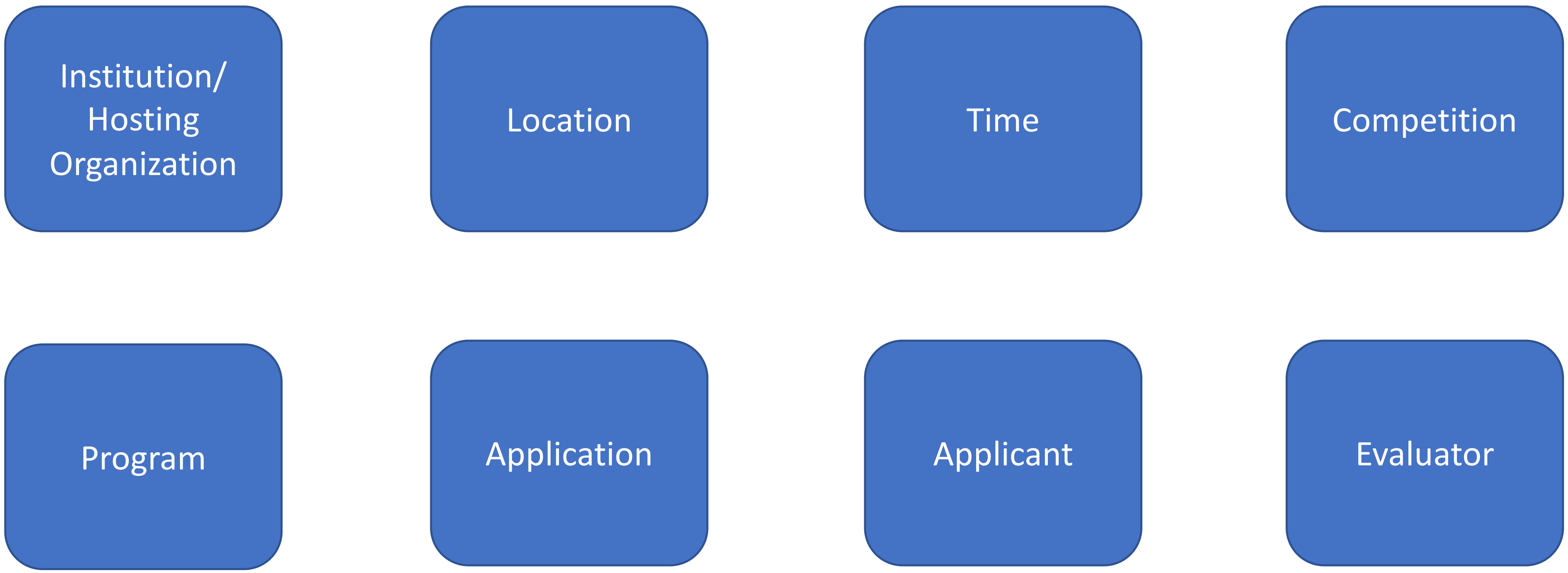
Some Useful Analogies

	Medicine	Cooking	The World of Cars
Amateur	Everyone First Aider	Home Cook	People who own cars Car Hobbyist
Semi-Pro	Paramedic	Bake-Sale Folks?	Semi-Pro Racer Gas Station Mechanic?
Professional	Doctor: GP, Specialist Nurse Hospital Director	Chef Pastry Chef Restaurant Owner	Garage Mechanic Body Shop Specialist

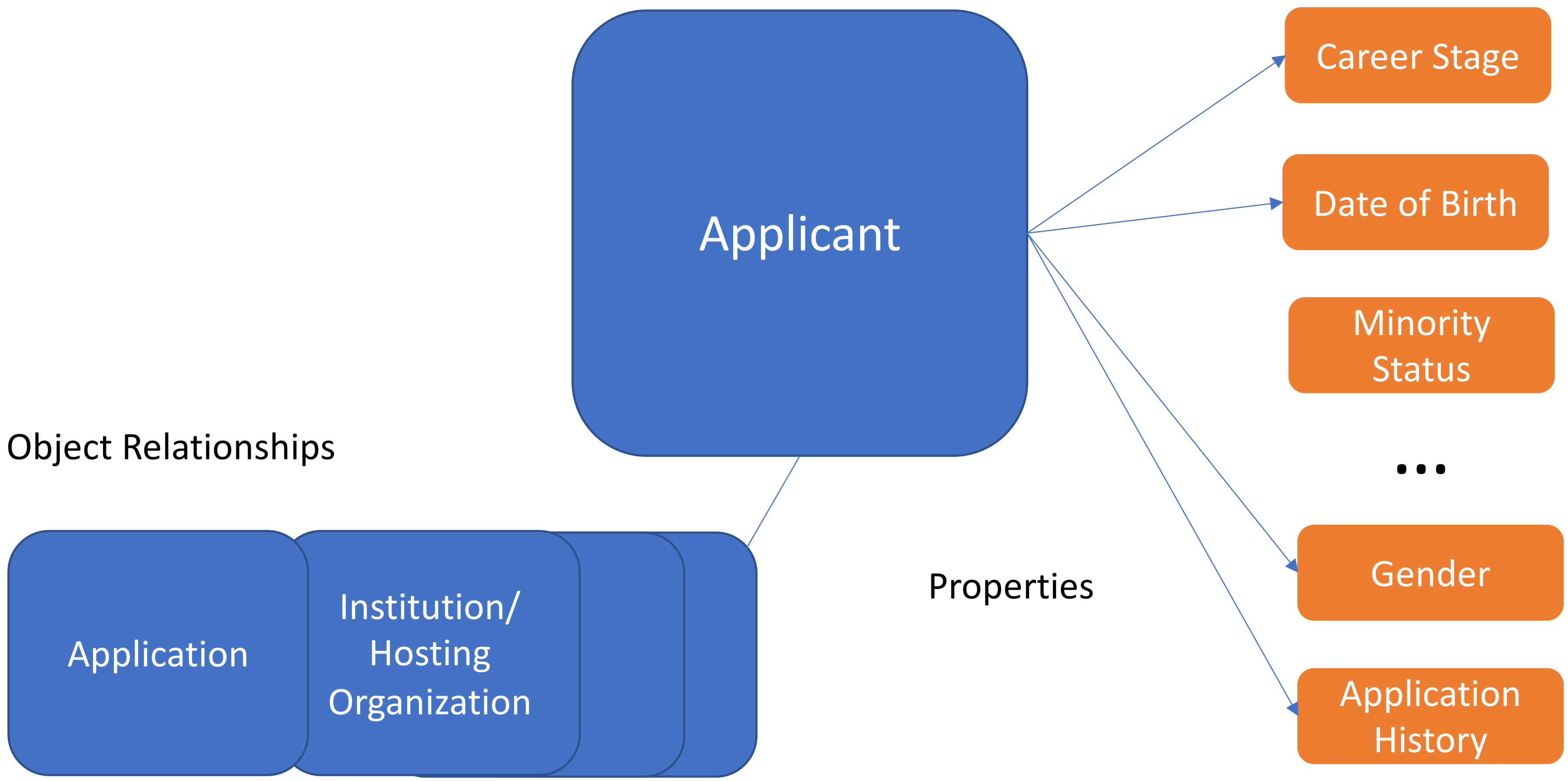
Your Team,
Your Data,
Your Questions


Background of NSERC Workshop Group (Based on Questionnaire Responses)
Response Count
0
3
6
9
AF
TF
AF + TF
MSA-S
TS
MSA-S + TS
Data Background Category
Data Backtround Category
Analysis Focus
Technical Focus
Analysis + Technical Focus
Math-Stats-Analysis Specialization
Technical Specialization
Math-Stats-Analysis + Technical Specialization

Key Objects Sketch – NSERC Grants



A Key Object with Properties – NSERC Grants



Topics from Survey: Non-Statistical

REPORTING AND DATA PREPARATION

- Any tips for generating reports from data (e.g., my team maintains an Excel "work plan" spreadsheet of our projects. Useful if I can pull reports from this data every quarter or so for senior management).
- Importing data from various sources to use in analysis (e.g. several Excel spreadsheets that are updated monthly/quarterly, websites, etc.)
- Translation of Power BI, tips and short cuts.

REPORTING AND DATA PREPARATION (CONT.)

- Best practices of data preparation for data analysis (i.e. curation, verification, setup, etc.)
- Preparing data sets for analysis

TEXT ANALYSIS + MACHINE LEARNING

- word cloud and keyword analysis. any additional methods to gain insight into data from freeform text entries
- Working with text data
- Data analysis, AI / ML

Topics from Survey: Statistical (I)

POPULATION VS SAMPLE

- statistical techniques to compare period data (e.g., year to year)
- Statistical techniques for population data
- Using significance tests when you've captured the whole population in the dataset

SIGNIFICANCE TESTS

- statistical significance tests - which to use for questions commonly asked about NSERC data and how to perform them

SIGNIFICANCE TESTS (CONT.)

- "How to determine statistical significance of findings. For example at NSERC, when is the difference between an application rate and an award rate significant?
- the statuses of grants and scholarships are successful or unsuccessful (analysis of those, how to analyse if the results from two subgroups are statistically different, what approach would you use?)

TOOLS FOR DOING STATS

- Statistical tools available in the software.

Topics from Survey: Statistical (II)

MULTI-VARIABLE, MULTI-LEVEL, MULTI-FACTOR, MULTI-PASS!

- What is the best approach for multi-variable analysis. For example at NSERC, intersectional analysis in the context of EDI could have multiple identity factors."
- Advanced analyses techniques (e.g., multi-level regression analyses)

OTHER 'COMPLEX' STATISTIC TOPICS

- Complex sampling (e.g., for large-scale surveys)

NON-PARAMETRIC TESTS AND ANALYSIS

- data that we use is often not normal (we usually give more small grants than large grants), what approach would you use for the analysis?"
- ordinal variables in surveys for example (how to analyse if the results from two subgroups are statistically different, what approach would you recommend?)

Meeting you where you are:

If you are already a statistical analysis expert (You laugh in the face of regression analysis and statistical models): Data Engineering, Data Preparation, Data Presentation, Alternatives to statistics (e.g. Machine Learning, Business Intelligence)

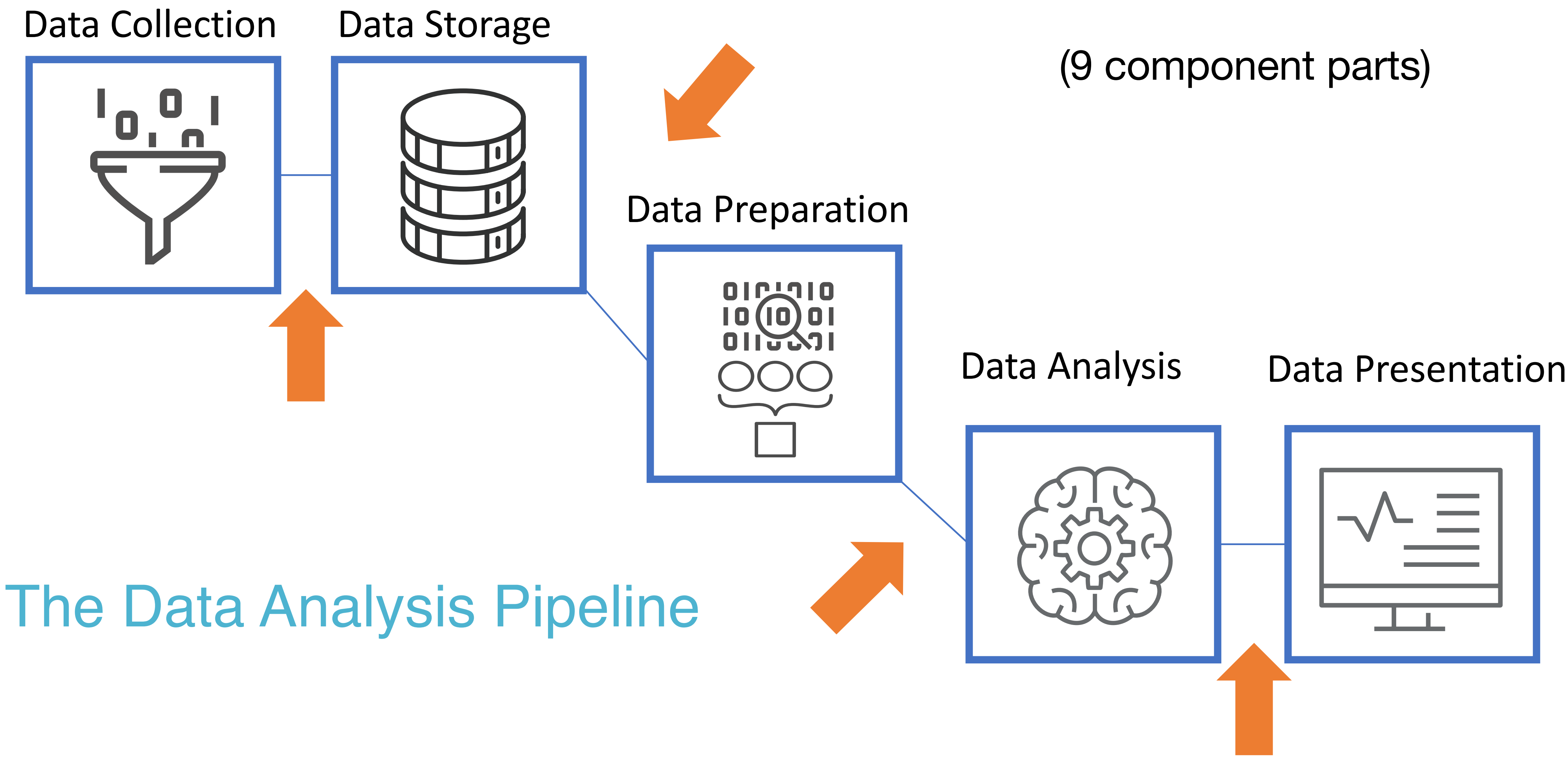
If you are already a programming expert (you eat R packages or Python modules for breakfast): New Analysis Techniques, Modern Data Engineering IT Infrastructure Practices

If you are already an IT infrastructure expert (you could stand up sophisticated cloud or on prem fully automated data pipeline systems without breaking a sweat!): An understanding of why the other members of the team are always asking you for such strange things on the IT front!

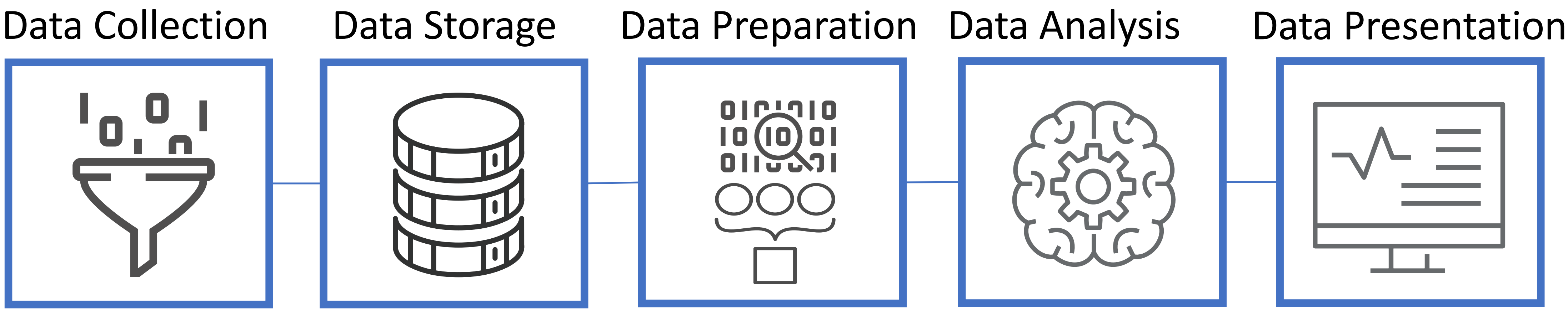
If you are already a subject matter or organization expert (you know everything there is to know about grant programs and the world of scientific grants): How the other members of the data science team can support your work, the language you need to use talk to them to make sure they give you what you want.

If you're already an expert on techniques, technologies, analysis and the subject (you could be giving this workshop): a shared orientation to modern data analysis that will help you talk to and work with other members of your data science team

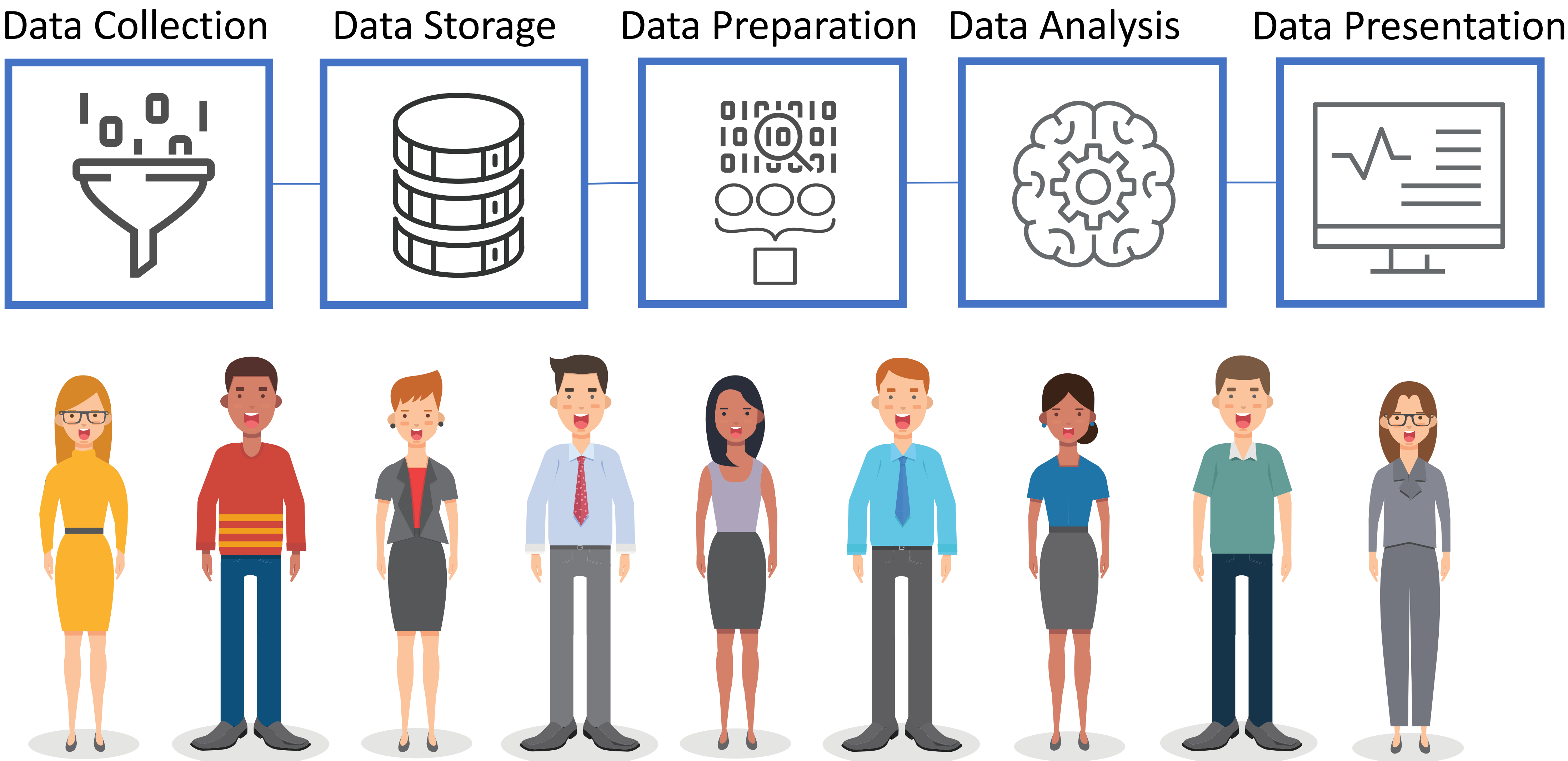
Modern Data Analysis
Teams and Technologies



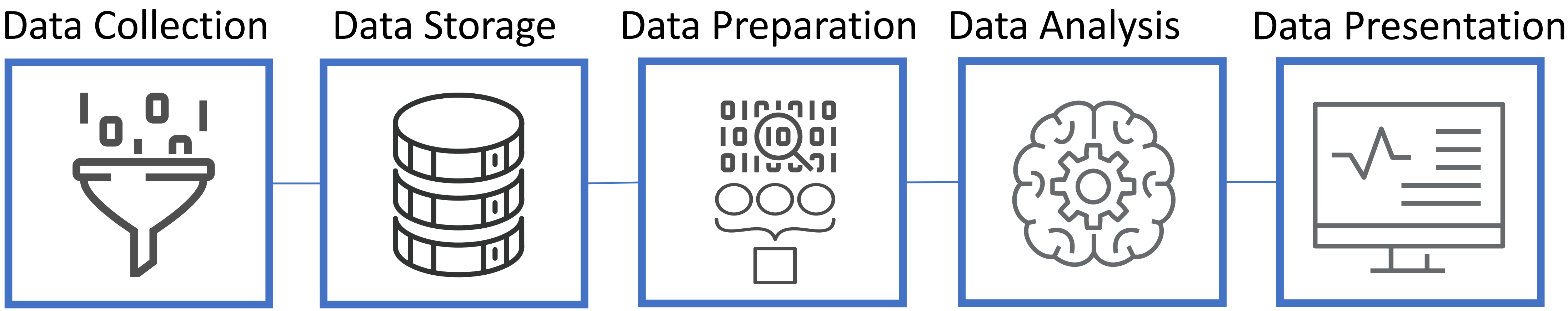
Data Collection
Data Storage
Data Preparation
Data Analysis
Data Presentation
(9 component parts)
The Data Analysis Pipeline



What Roles Support This Pipeline?



~ ratio of other team members to analysts: 10 : 1



~ ratio of analysis to other activities - 1:10

Data Team Roles (I)

Data engineering: Data infrastructure design and implementation - IT and DevOps heavy

Data collection: design of data collection strategies and implementation of data collection tools

Data architect, Data manager: Data storage and data architecture design and implementation

Data preparation: You work hand in hand with the data collection ,data managers and data analysis members of the team to get the data into a state where it is ready for analysis. To automate this you design and implement processes to carry out all of the relevant steps This is a pivotal position on the team

Analysis: You determine what analysis can work with the information you have, and can give insight that is relevant and useful. You design algorithms that can be used to automate these analyses

Data Team Roles (II)

Data Pipeline UX Expert: Interface design, user experience,

Data Communication: data visualization, data presentation

Subject Matter Expert: Knows a lot about the situation, understands what is important, what data could provide insight, how to interpret and apply the results of the analysis

Business or Organization Strategy expert: You hold the picture and know where the organization wants to head. You need to provide this information to the team.

Project Lead: You keep everyone on track and working together

Data Translator: Knows how the different pieces of the pipeline work at a high level, knows something about the subject matter. Good at connecting people and helping them talk to each other.

Where do you fit in?

Here are some preliminary questions to ask yourself:

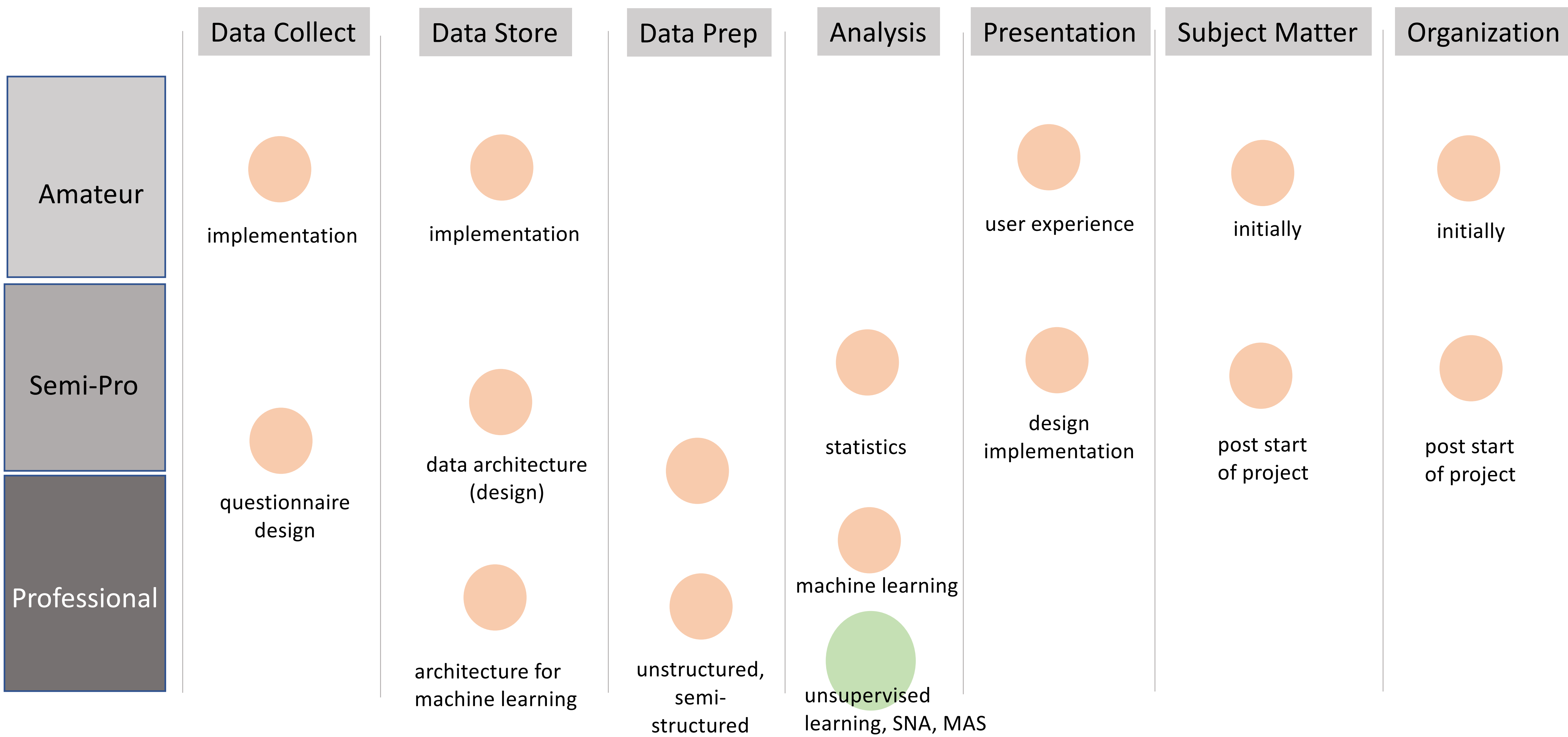
1. What part of the pipeline is most appealing to you?
2. Do you like designing OR implementing what someone else has designed? Or both?
3. Are you a generalist who likes to know a little bit of everything, or do you like to specialize and become an expert in one thing? (Are you a big picture person or a detail-oriented person)
4. Can you currently write computer programs or more generally scripts that tell computers what to do (OR do you want to be able to do so)?
5. Do you have a math or statistics background
6. Do you like working with IT technologies?
7. Do you like to facilitate communication between different members of a team
8. Do you have a deep knowledge of your organizations operations or subject matter
9. Do you have a deep knowledge of organizational goals? Do you like strategy?

Another way to think about where you fit

	Data Collect	Data Store	Data Prep	Analysis	Presentation	Subject Matter	Organization
Amateur							
Semi-Pro							
Professional							

Generalists vs Specialist: You can't do it all!

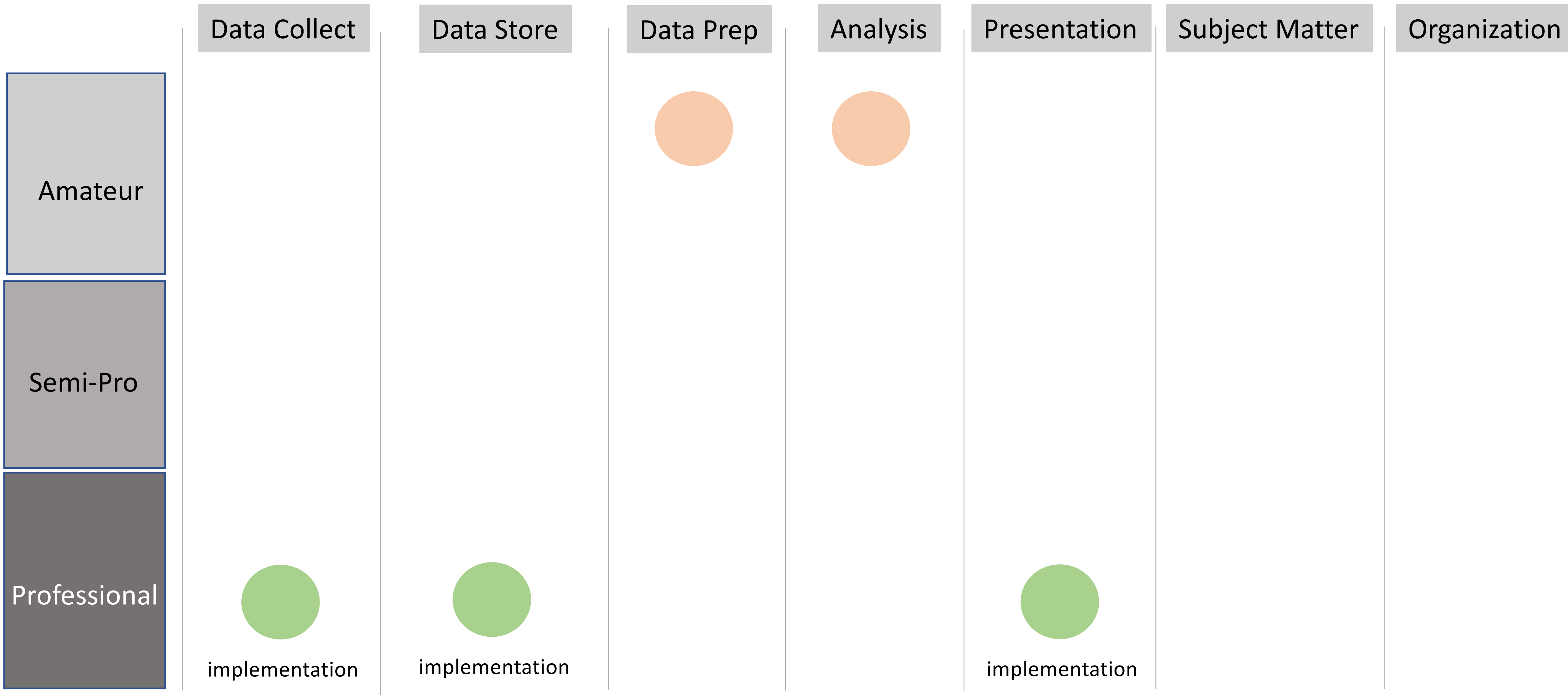
Example of a generalist: multi-purpose, can communicate across lanes,



Generalists vs Specialist: You can’t do it all!

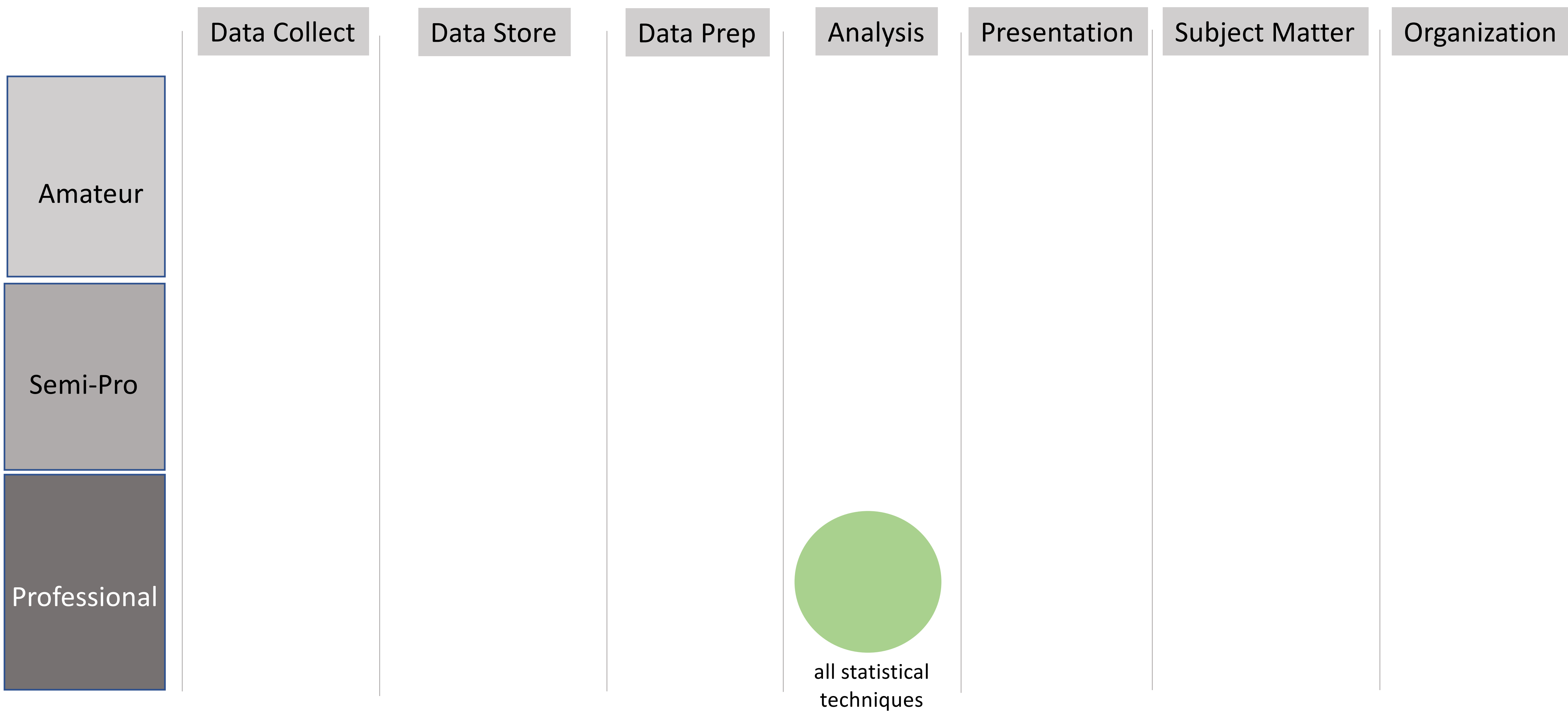
Example of a specialist (data engineer)

High quality, fast, deal with tricky situations



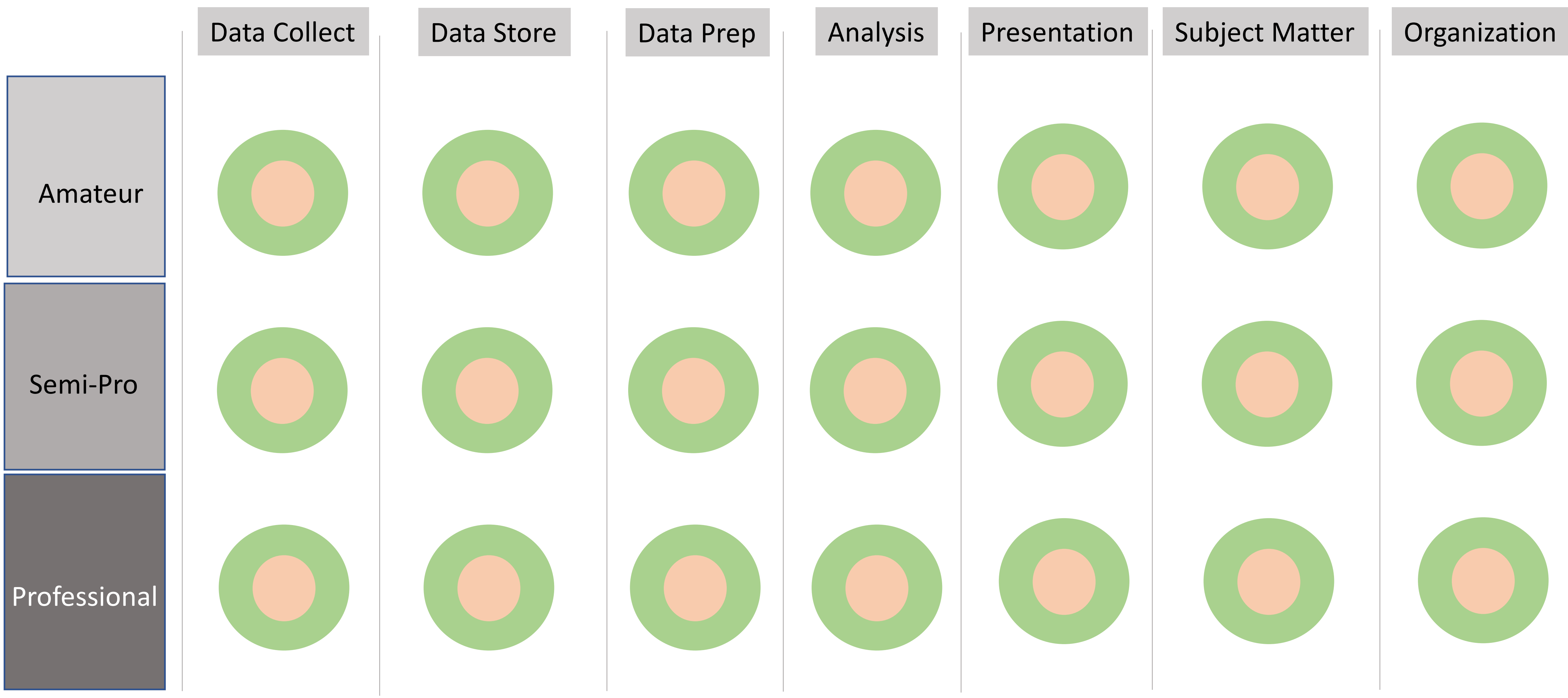
Generalists vs Specialist: You can’t do it all!

Example of a specialist (statistician)



Full coverage

A team can collectively provide you with full coverage

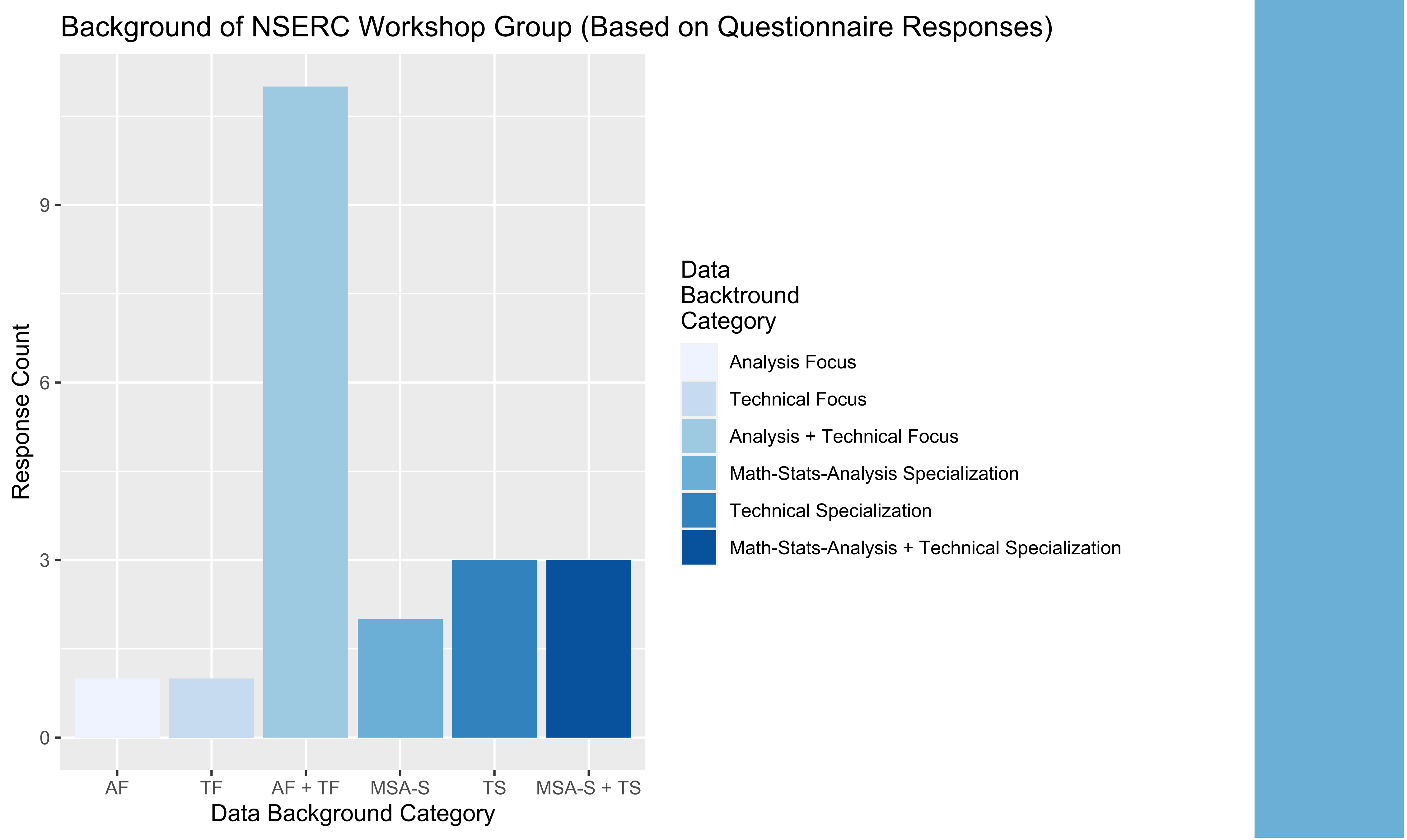


Data Analysis - Why Me???

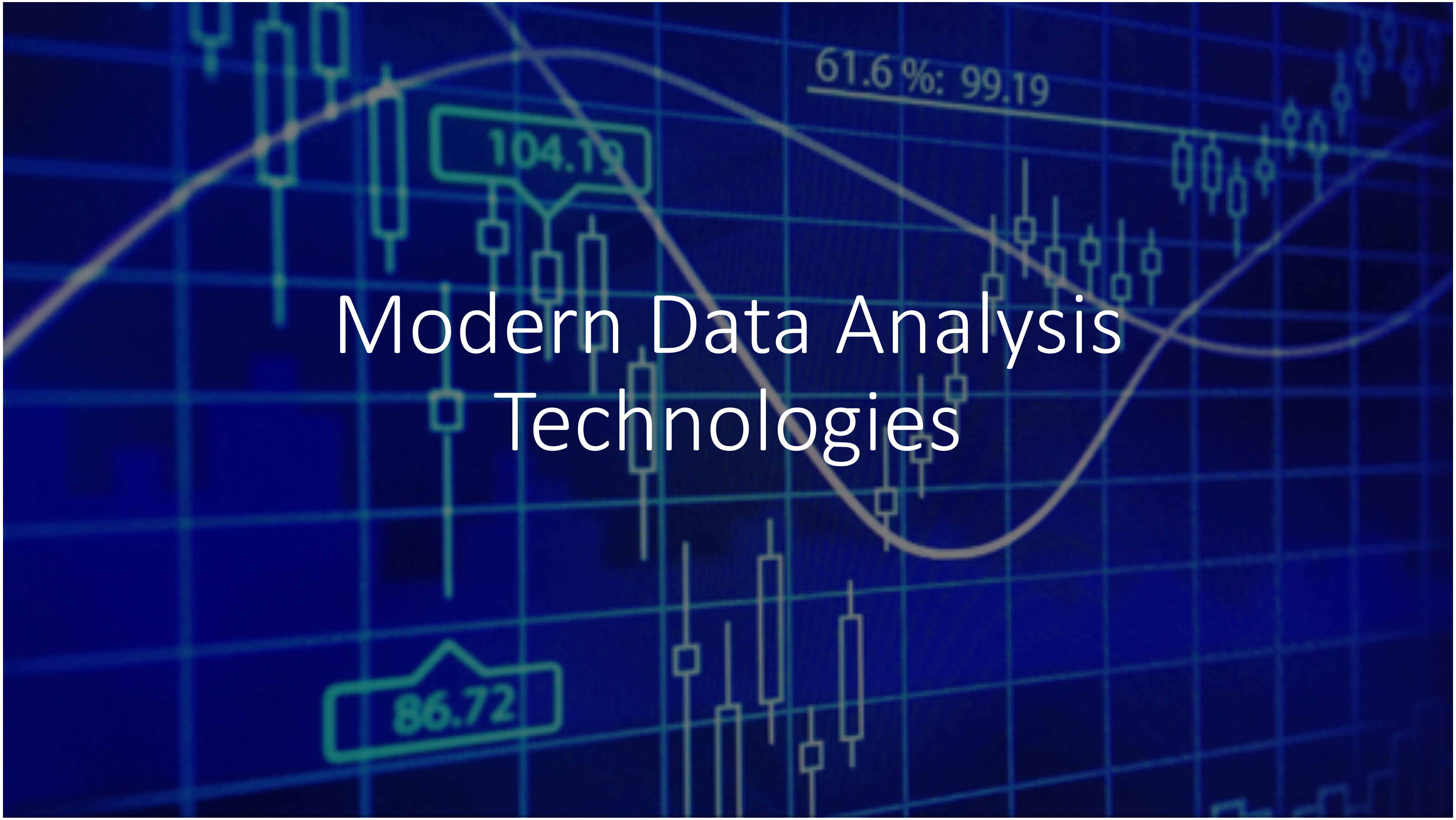
- In general, even if you are not an analyst - you must be able to talk to the analysts!
- Data engineer – person who is designing the kitchen – need to know what the cooks will be doing, who in turn need to know what people want to eat
- Data presenter – person who is doing the bodywork on the car – need it to fit on the car, who in turn must deliver a car that the driver likes
- Data Translator!!!

In the end, it’s not about the analysis!

- In a professional setting, **analysis will not be happening just for the sake of analysis**.
- In an applied setting, analysis supports business goals.
- Let’s return to our analogies for a moment.
  - Cooking analogy: In the cooking world, the person eating the food is royalty!
  - Car Hobby analogy: Yes, some people work on cars for fun. And some people work on data for fun. BUT in the end it’s about the owner/driver of the car
- Don’t lose sight of your end goal. Who knows the end goal? DON’T PRESUME THAT THIS IS YOU!
- User Centered Design

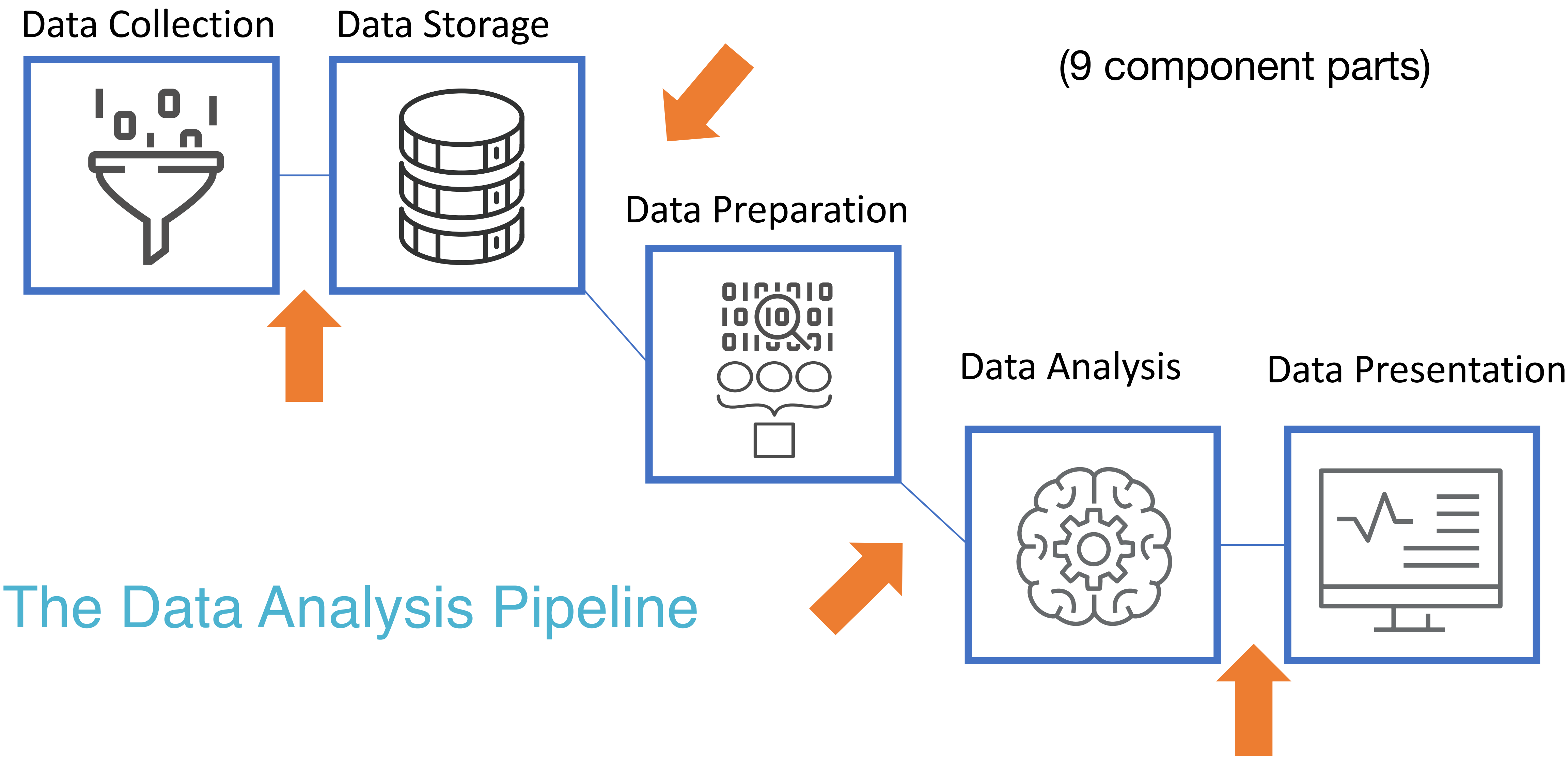

Background of NSERC Workshop Group (Based on Questionnaire Responses)
Response Count
0
3
6
9
AF
TF
AF + TF
MSA-S
TS
MSA-S + TS
Data Background Category
Data Backtround Category
Analysis Focus
Technical Focus
Analysis + Technical Focus
Math-Stats-Analysis Specialization
Technical Specialization
Math-Stats-Analysis + Technical Specialization

Modern Data Analysis Technologies

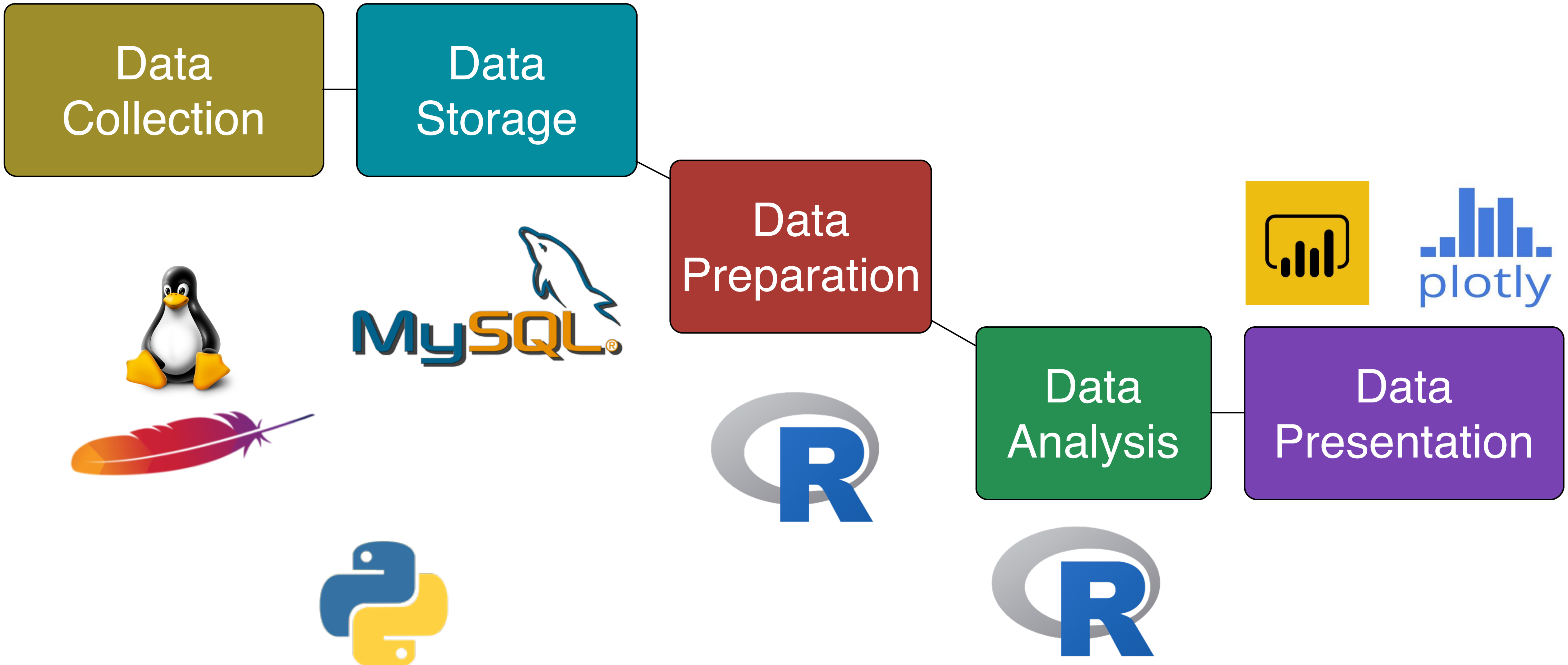


The Data Analysis Pipeline

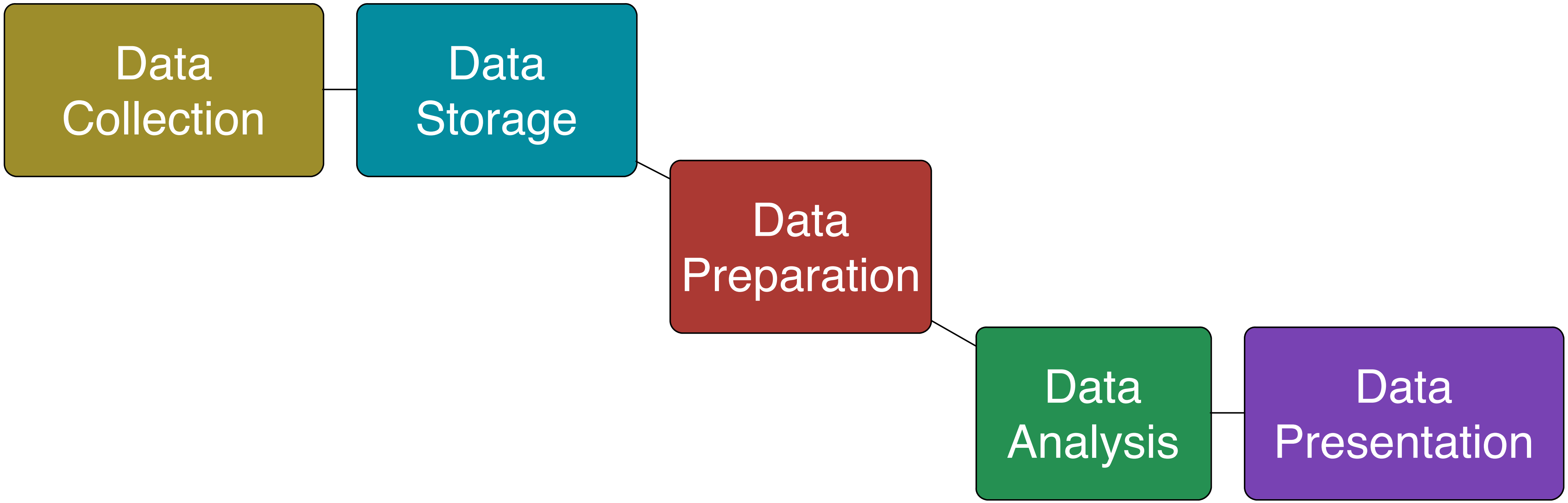
(9 component parts)

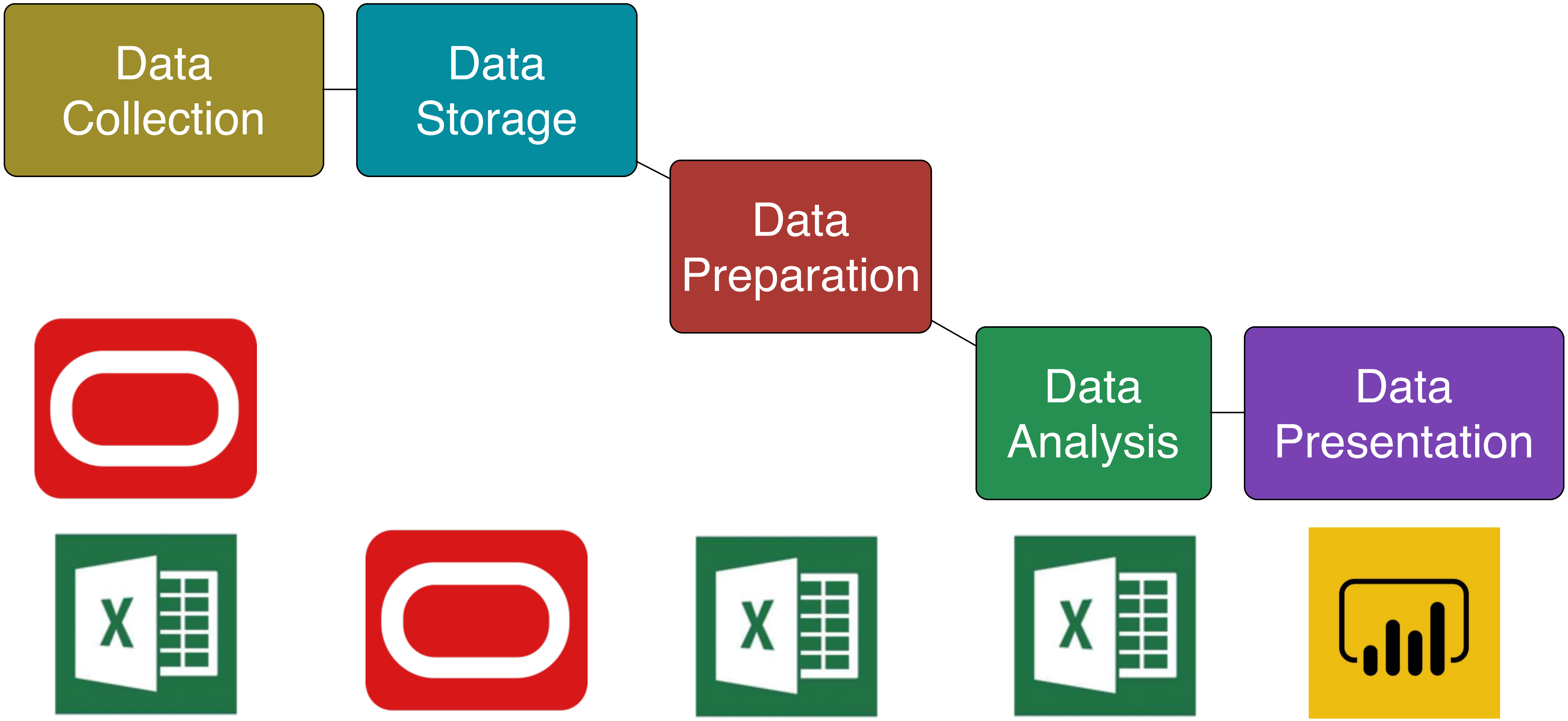


A Open-Source Driven Data Stack



Typical GoC Data Stack




Data Collection
Data Storage
Data Preparation
Data Analysis
Data Presentation

Data Pipeline Technologies: Amateur Technologies vs Professional Technologies – Compare and Contrast

	On-Premises (LAN)	Public Cloud	Private Cloud
Amateur	Shared Directory + Excel +Power Point + 'Desktop' Access	Piecemeal SaaS – e.g. Data Analysis or Presentation as a service Freemium Model	
Semi-Pro	Desktop DataScience: Desktop PowerBI SQL-Lite (Desktop) MS Access Stand-alone In-House DBMS – Read + Write		Home Brewed Solutions using Servers stood up on Cloud – e.g. AWS, GCP
Professional	Server Based End-to-End Automated Pipeline Tech: On-Premises Azure, On-Premises IBM RedHat	End-to-end SaaS data pipelines – e.g. COTS Pachyderm or more bespoke: e.g. SaaSCoder	End-to-End Cloud Data Pipeline Infrastructure (Serverless/NoServer): AWS, GCP, Azure

Understanding the Cloud Landscape

IaaS: Infrastructure as a Service

PaaS: Platform as a Service

SaaS (AIaaS, DaaS): Software (AI, Data) as a Service

Pipeline Creation Phases

1. **Research + Design**
2. **Implementation**
3. **Testing**
4. **Production + Management**
5. **Research + Design**

Agile!

Pre-Analysis: Data Collection, Structuring and Preparation

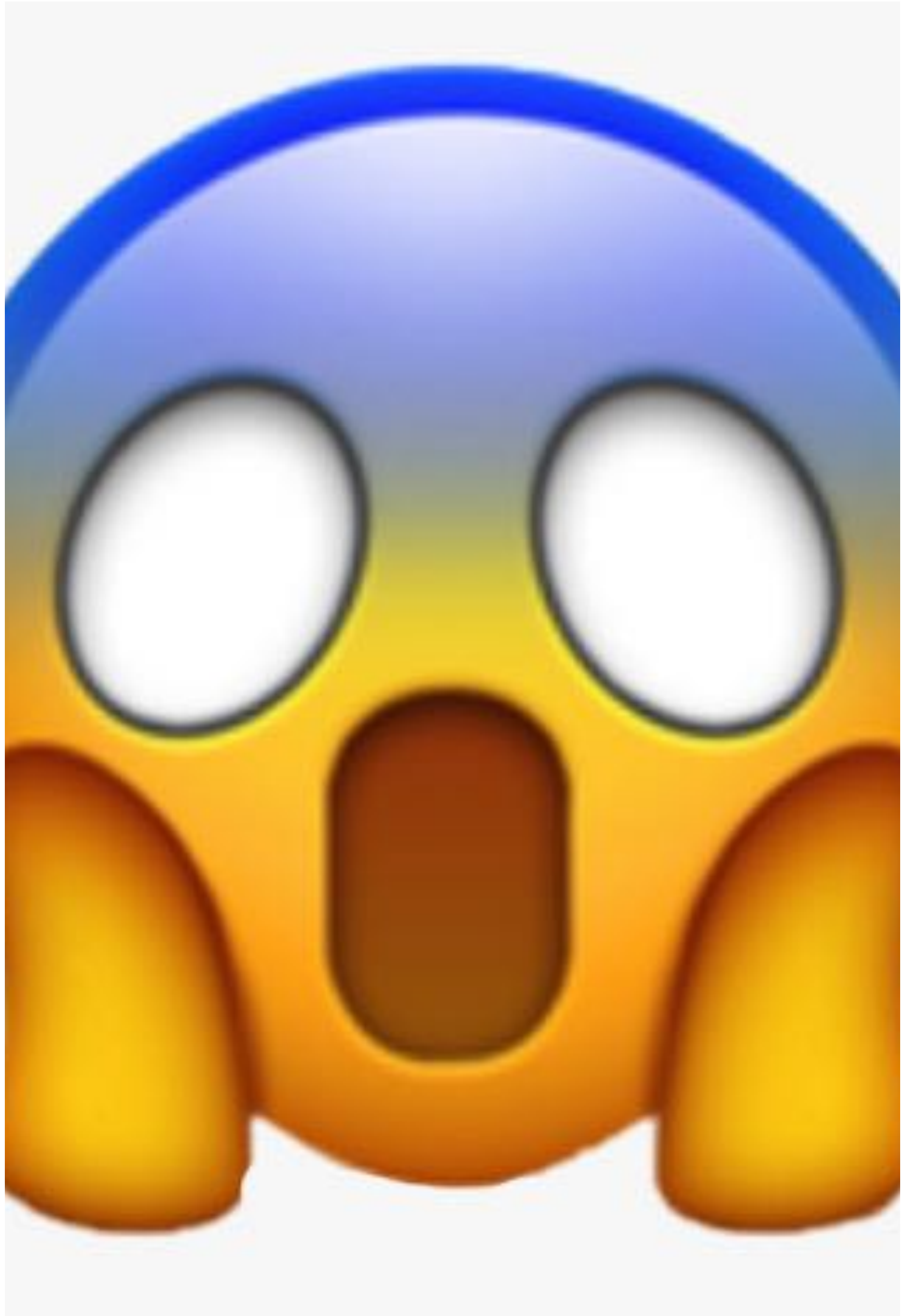
Collection: Three Main Data Sources

Recordkeeping

Research

Sensors/Monitoring

Recordkeeping: Primary Focus On Specific Entities



The Curse of Categorical Data

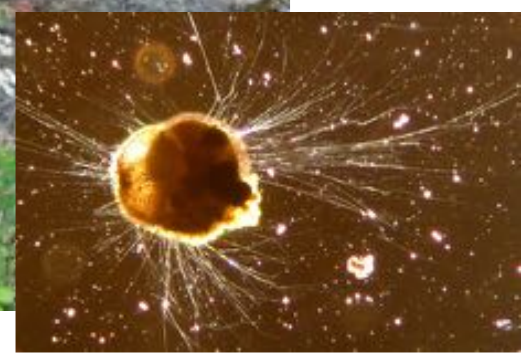
Government data tends to be very heavy on the categories and text data.

Traditional analysis methods:

- were not categorical data heavy
- did not focus on doing complex analyses with (complex) categorical data

This means we need to work harder to come up with good strategies to deal with this type of data (hint – machine learning likes categories)

Research: Focus On Generalizing



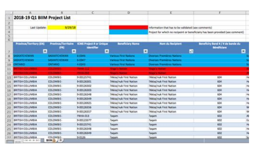
Applied Data Analysis and Science

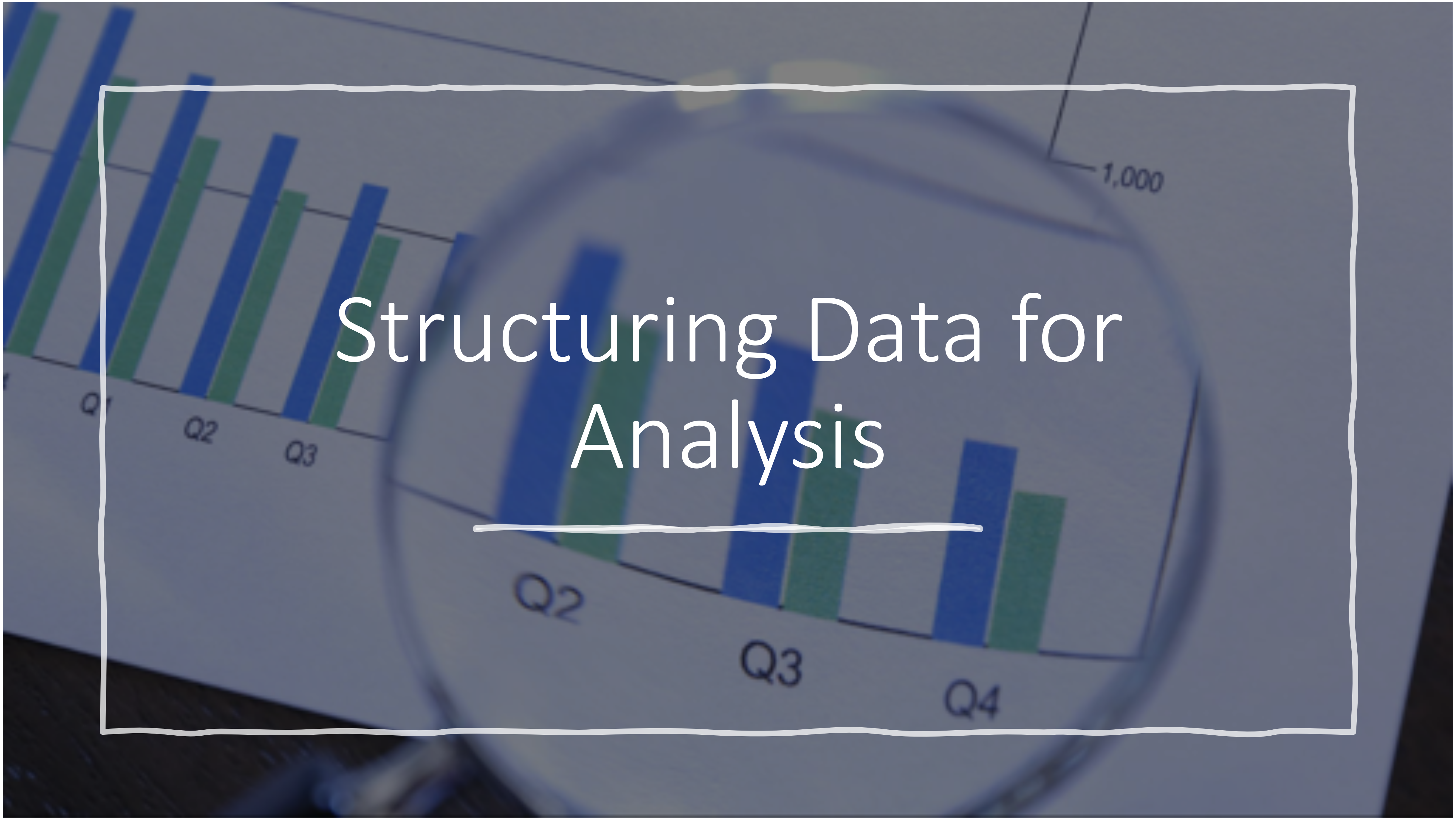
- Scientific data analysis techniques are sometimes relevant only:
  - in a *very specific experimental context*
  - *on certain types of data*
- Now that data is so much more prevalent and usable, we need to grow and adapt these techniques
- We need to break out of the 'science mindset'

Decision Support! Immediate and Focused

What Is Your Analysis Goal?

- Do you want to:
  - Carry out actions based on what is in your data (maybe not analysis?)
  - gain a deeper understanding of something **specific** (specific individuals? A specific group?)
  - come to some **general** conclusions that extend beyond the specific
- Local vs Global
- Here vs Everywhere
- **Past/Present vs Future**
- Situational Awareness vs Contingency Planning

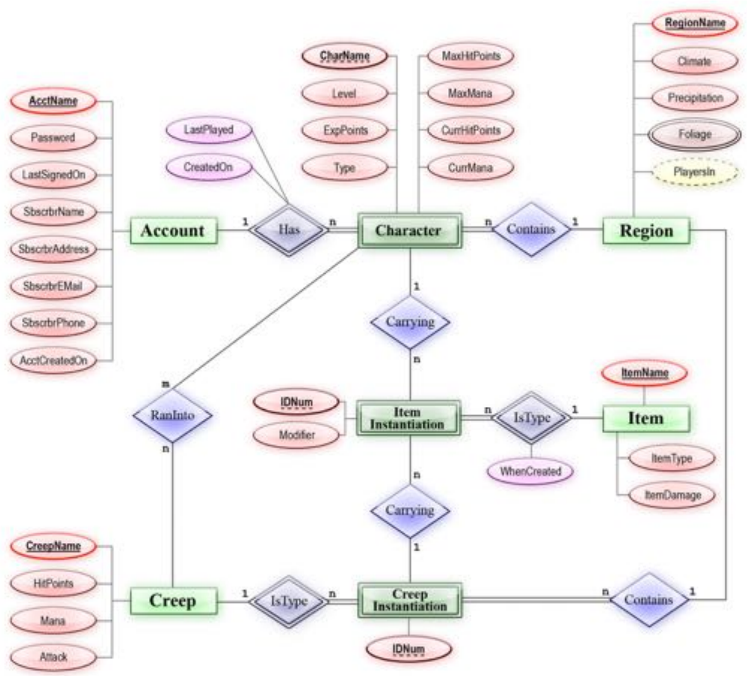




Structuring Data for Analysis

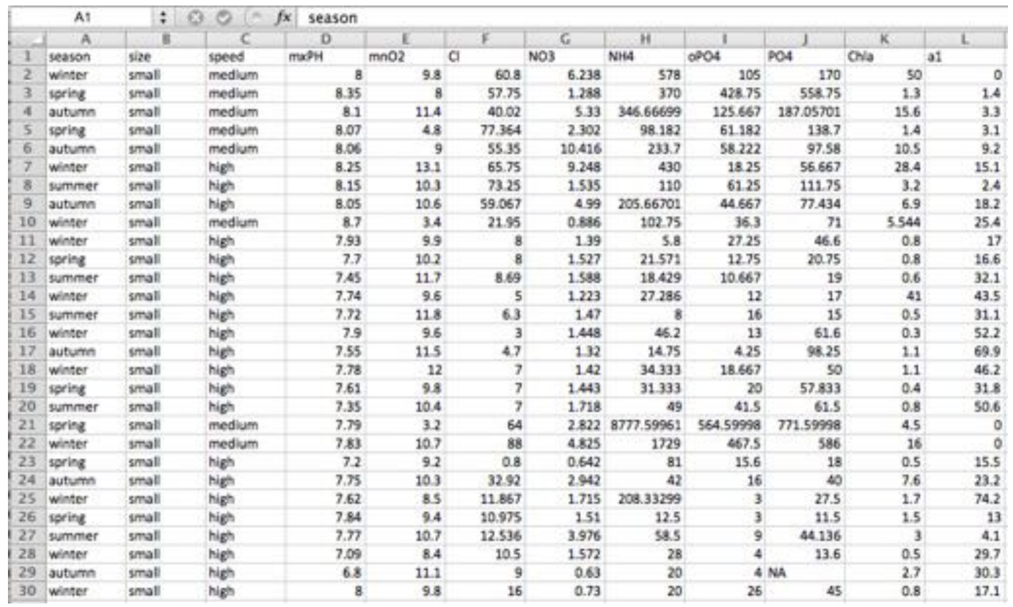
Database vs Flat File

Database



Data Integrity

Flat File



	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

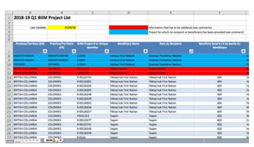
Rows vs Columns

Columns contain attributes (variables, fields, etc.)

Rows contain objects*

A1 | 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

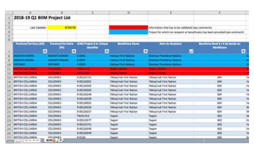
variable (field) value (datum)

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

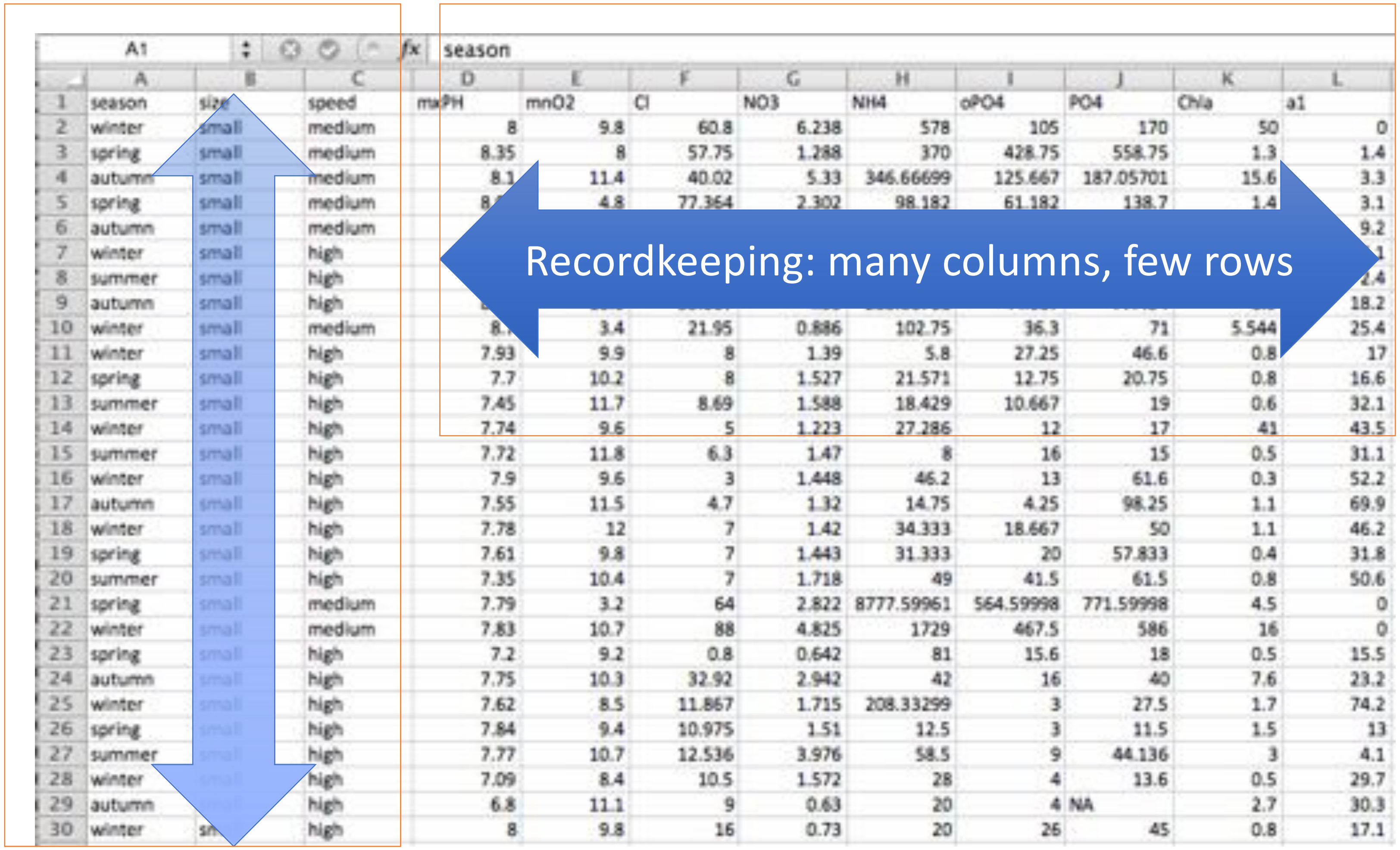
Record-keeping

Research



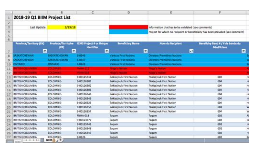
Dataset Shape and Focus

Research: many rows, few columns



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**.

The Curse of Free Text Fields

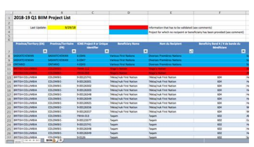
- The curse of categorical data is made much worse by the curse of free text fields
- If you have a field that is supposed to be categorical but it is a free text field, **it is no longer categorical**
- You can use machine learning techniques to help to some extent, but this is a case **where an ounce or prevention is worth a pound of cure**.

Group Name
Best Practice in Data Analysis - RGS
BI Team
CBU
Corporate Reporting
Linda Gagnon
Michael Lam
NSERC
NSERC
PPO
RGS

Plus 11 blanks!

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



Cleaning: Character Encodings

The ASCII code

American Standard Code for Information Interchange

ASCII control characters

DEC	HEX	Símbolo ASCII	
00	00h	NULL	(carácter nulo)
01	01h	SOH	(inicio encabezado)
02	02h	STX	(inicio texto)
03	03h	ETX	(fin de texto)
04	04h	EOT	(fin transmisión)
05	05h	ENQ	(enquiry)
06	06h	ACK	(acknowledgement)
07	07h	BEL	(timbre)
08	08h	BS	(retroceso)
09	09h	HT	(tab horizontal)
10	0Ah	LF	(salto de linea)
11	0Bh	VT	(tab vertical)
12	0Ch	FF	(form feed)
13	0Dh	CR	(retorno de carro)
14	0Eh	SO	(shift Out)
15	0Fh	SI	(shift In)
16	10h	DLE	(data link escape)
17	11h	DC1	(device control 1)
18	12h	DC2	(device control 2)
19	13h	DC3	(device control 3)
20	14h	DC4	(device control 4)
21	15h	NAK	(negative acknowle.)
22	16h	SYN	(synchronous idle)
23	17h	ETB	(end of trans. block)
24	18h	CAN	(cancel)
25	19h	EM	(end of medium)
26	1Ah	SUB	(substitute)
27	1Bh	ESC	(escape)
28	1Ch	FS	(file separator)
29	1Dh	GS	(group separator)
30	1Eh	RS	(record separator)
31	1Fh	US	(unit separator)
127	20h	DEL	(delete)

ASCII printable characters

DEC	HEX	Símbolo	DEC	HEX	Símbolo	DEC	HEX	Símbolo
32	20h	espacio	64	40h	@	96	60h	`
33	21h	!	65	41h	A	97	61h	a
34	22h	"	66	42h	B	98	62h	b
35	23h	#	67	43h	C	99	63h	c
36	24h	$	68	44h	D	100	64h	d
37	25h	%	69	45h	E	101	65h	e
38	26h	&	70	46h	F	102	66h	f
39	27h	'	71	47h	G	103	67h	g
40	28h	(	72	48h	H	104	68h	h
41	29h	)	73	49h	I	105	69h	i
42	2Ah	*	74	4Ah	J	106	6Ah	j
43	2Bh	+	75	4Bh	K	107	6Bh	k
44	2Ch	,	76	4Ch	L	108	6Ch	l
45	2Dh	-	77	4Dh	M	109	6Dh	m
46	2Eh	.	78	4Eh	N	110	6Eh	n
47	2Fh	/	79	4Fh	O	111	6Fh	o
48	30h	0	80	50h	P	112	70h	p
49	31h	1	81	51h	Q	113	71h	q
50	32h	2	82	52h	R	114	72h	r
51	33h	3	83	53h	S	115	73h	s
52	34h	4	84	54h	T	116	74h	t
53	35h	5	85	55h	U	117	75h	u
54	36h	6	86	56h	V	118	76h	v
55	37h	7	87	57h	W	119	77h	w
56	38h	8	88	58h	X	120	78h	x
57	39h	9	89	59h	Y	121	79h	y
58	3Ah	:	90	5Ah	Z	122	7Ah	z
59	3Bh	;	91	5Bh	[	123	7Bh	{
60	3Ch	<	92	5Ch	\	124	7Ch	|
61	3Dh	=	93	5Dh	]	125	7Dh	}
62	3Eh	>	94	5Eh	^	126	7Eh	~
63	3Fh	?	95	5Fh	_			

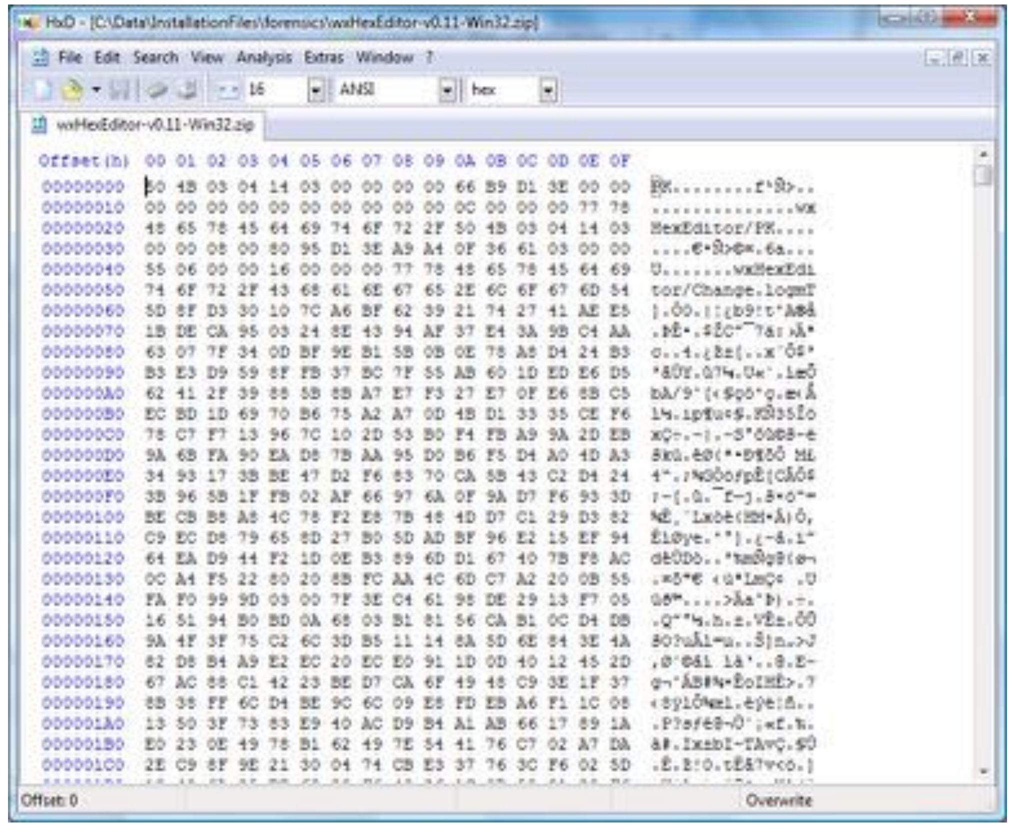
theASCIIcode.com.ar

Extended ASCII characters

DEC	HEX	Símbolo	DEC	HEX	Símbolo	DEC	HEX	Símbolo	DEC	HEX	Símbolo
128	80h	Ç	160	A0h	á	192	C0h	└	224	E0h	Ó
129	81h	ü	161	A1h	í	193	C1h	┴	225	E1h	ß
130	82h	é	162	A2h	ó	194	C2h	┬	226	E2h	Ô
131	83h	â	163	A3h	ú	195	C3h	├	227	E3h	Ò
132	84h	ä	164	A4h	ñ	196	C4h	─	228	E4h	õ
133	85h	à	165	A5h	Ñ	197	C5h	┼	229	E5h	Õ
134	86h	å	166	A6h	ª	198	C6h	ã	230	E6h	µ
135	87h	ç	167	A7h	º	199	C7h	Ã	231	E7h	þ
136	88h	ê	168	A8h	¿	200	C8h	╚	232	E8h	Þ
137	89h	ë	169	A9h	®	201	C9h	╔	233	E9h	Ú
138	8Ah	è	170	AAh	¬	202	CAh	╩	234	EAh	Û
139	8Bh	ï	171	ABh	½	203	CBh	╦	235	EBh	Ù
140	8Ch	î	172	ACh	¼	204	CCh	╠	236	ECh	ý
141	8Dh	ì	173	ADh	¡	205	CDh	═	237	EDh	Ý
142	8Eh	Ä	174	AEh	«	206	CEh	╬	238	EEh	¯
143	8Fh	Å	175	AFh	»	207	CFh	¤	239	EFh	´
144	90h	É	176	B0h	░	208	D0h	ð	240	F0h	
145	91h	æ	177	B1h	▒	209	D1h	Ð	241	F1h	±
146	92h	Æ	178	B2h	▓	210	D2h	Ê	242	F2h	
147	93h	ô	179	B3h	│	211	D3h	Ë	243	F3h	¾
148	94h	ö	180	B4h	┤	212	D4h	È	244	F4h	¶
149	95h	ò	181	B5h	Á	213	D5h	ı	245	F5h	§
150	96h	û	182	B6h	Â	214	D6h	Í	246	F6h	÷
151	97h	ù	183	B7h	À	215	D7h	Î	247	F7h	¸
152	98h	ÿ	184	B8h	©	216	D8h	Ï	248	F8h	°
153	99h	Ö	185	B9h	╣	217	D9h	┘	249	F9h	¨
154	9Ah	Ü	186	BAh	║	218	DAh	┌	250	FAh	·
155	9Bh	ø	187	BBh	╗	219	DBh	█	251	FBh	¹
156	9Ch	£	188	BCh	╝	220	DCh	▄	252	FCh	³
157	9Dh	Ø	189	BDh	¢	221	DDh	¦	253	FDh	²
158	9Eh	×	190	BEh	¥	222	DEh	Ì	254	FEh	■
159	9Fh	ƒ	191	BFh	┐	223	DFh	▀	255	FFh	

Encoding: Tools and Strategies

- Use built in options in text editors, browsers
- Command line tools: iconv, recode, vim
- Libraries in R, Python
- Hex editors
- Statistical methods, machine learning!
- (an ounce of prevention…)



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 with NSERC data – 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’

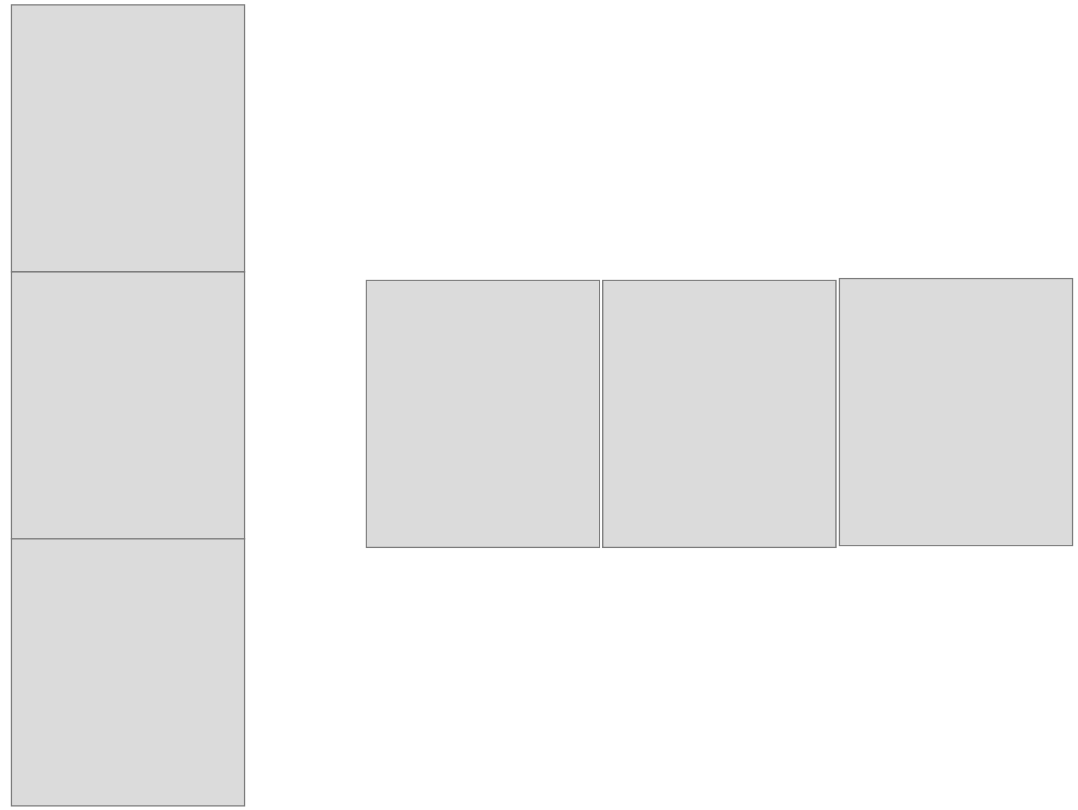
Collector bias, sensor error: recording what is expected rather than what is, dealing with badly calibrated sensor

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

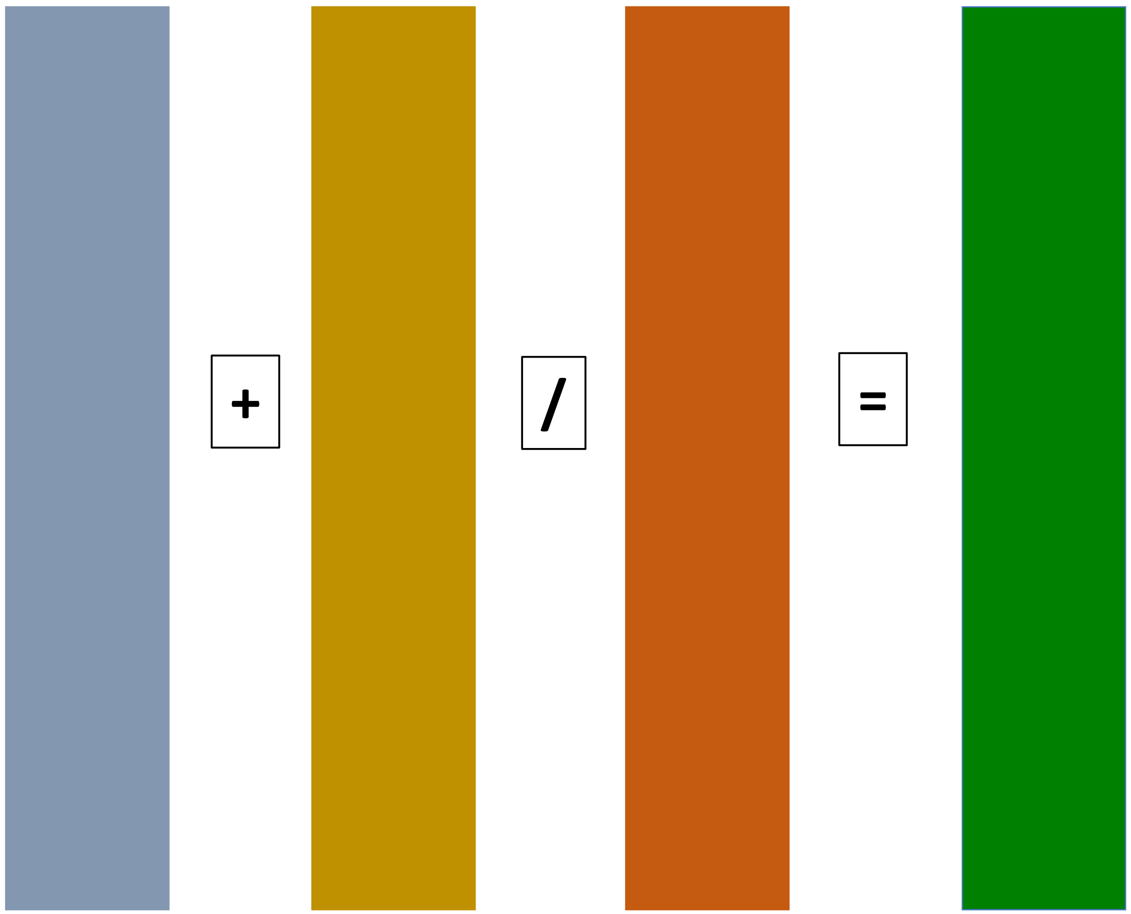
Group#	Group-Size-START	Group-Size-MIDDLE	Group-Size-END
1	14	12	13
2	20	5	6
3	6	8	10

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

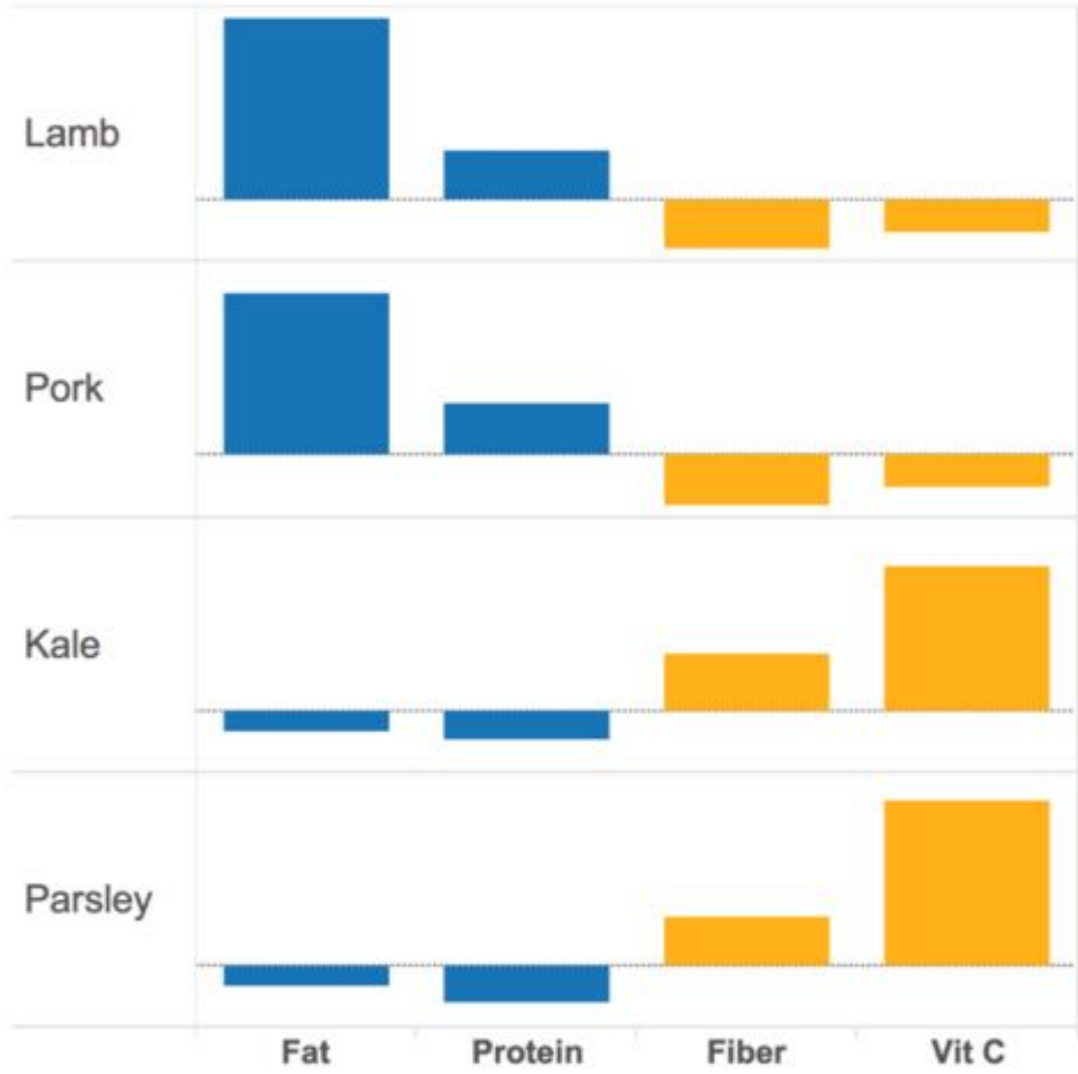
- “When a measure becomes a target, it ceases to be a good measure” ***Goodhart’s Law***
- “The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is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



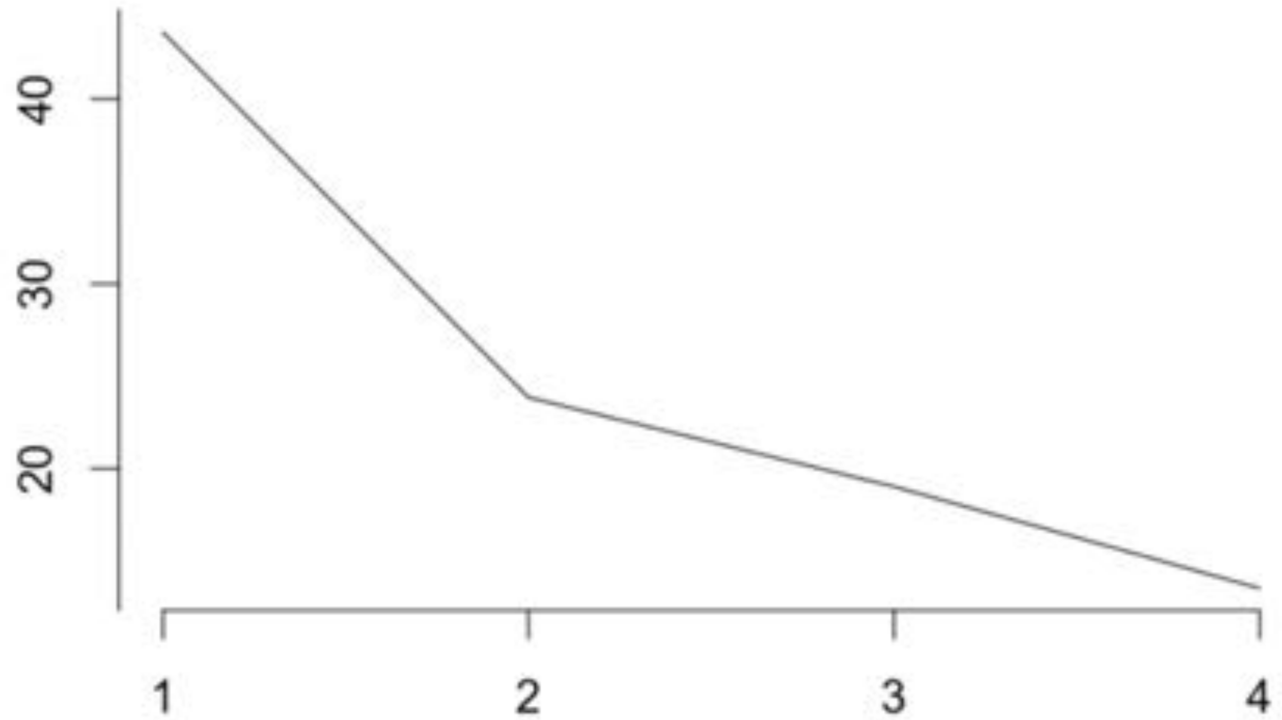
[A. Ng, K. Soo, *Numsense!*, USDA data]

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.



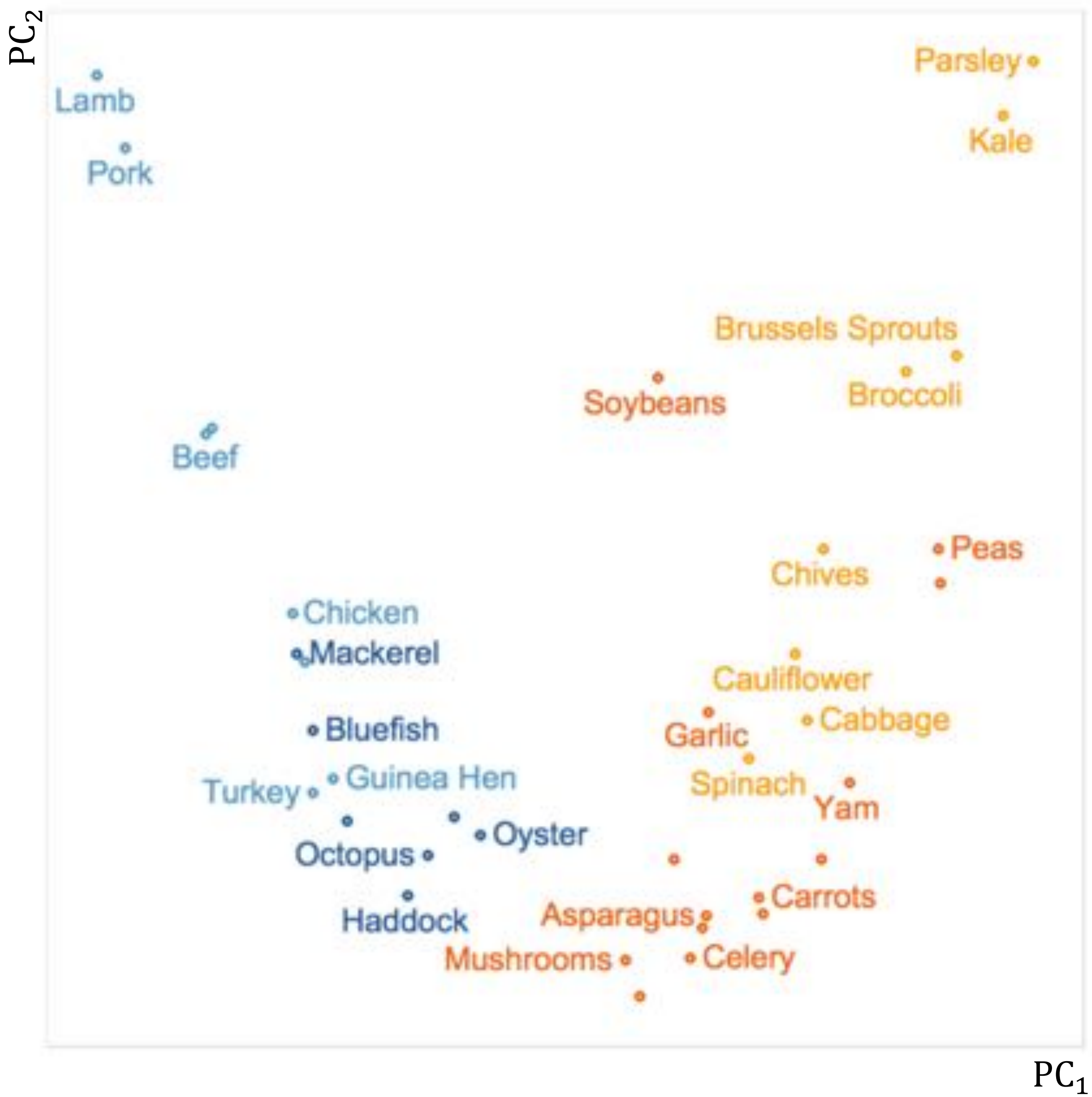
[A. Ng, K. Soo, *Numsense!*, USDA data]

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?