



Introduction to Modern Data Analysis

PART 1

Mornings Workshop Outline

Day 1:

- Analysis Context
- Modern Teams
- Modern Technologies
- Data Collection

Day 2

- Data Structuring
- Data Preparation
- Business Intelligence

• Day 3: Analysis - Statistics

- Modern Statistics –
Controversies and
Conversations
- Some Relevant
Statistical Concepts
and Techniques

• Day 4: Analysis - Machine Learning/AI

- Machine Learning
Techniques
- Focus on Text
Mining



Introduction

Armchair analysis



What is (data)
analysis?

Some possible answers

Finding patterns in data

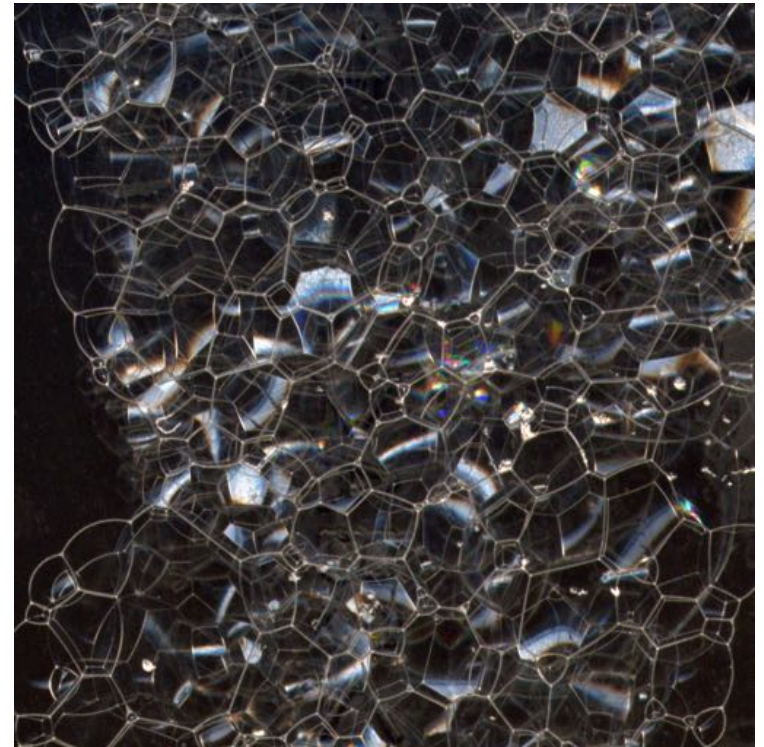
Using data to do something (answer a question, help decision-making, predict the future, knowledge discovery)

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: financial transactions over the past year, financial transactions for the upcoming year.
- 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.
- Goes beyond domain specific (e.g. financial analysis)

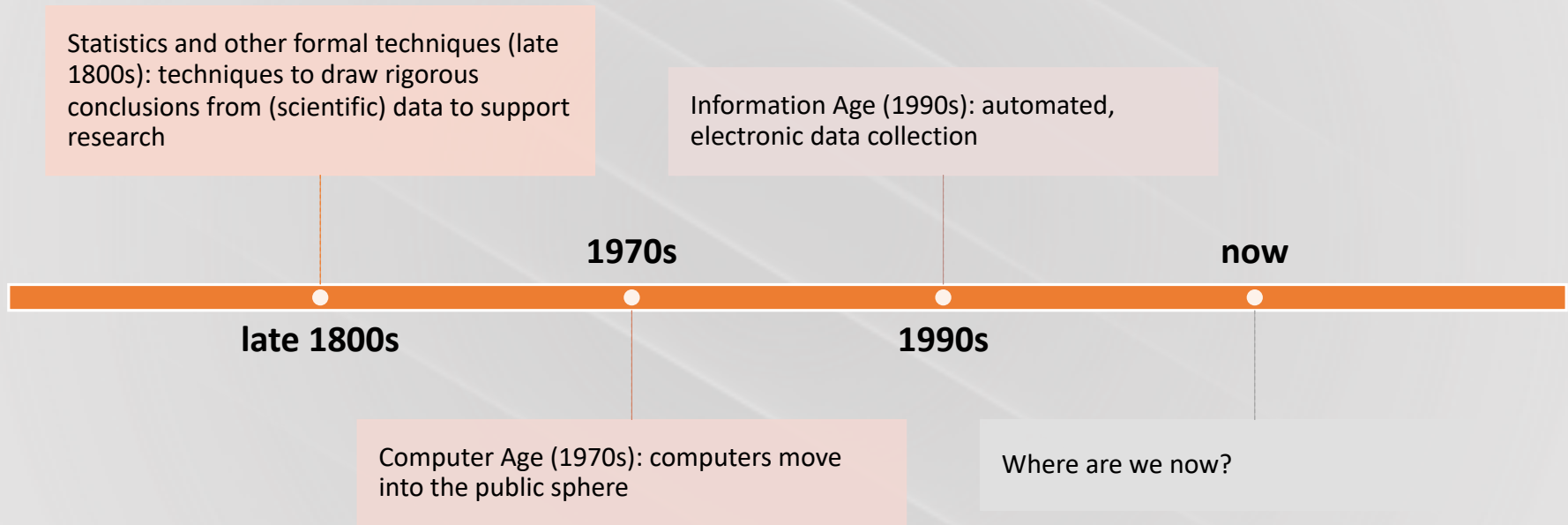
Formal Reasoning Techniques

INDUCTIVE, 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?



Analysis in the Pre- Digital Age vs The Digital Age

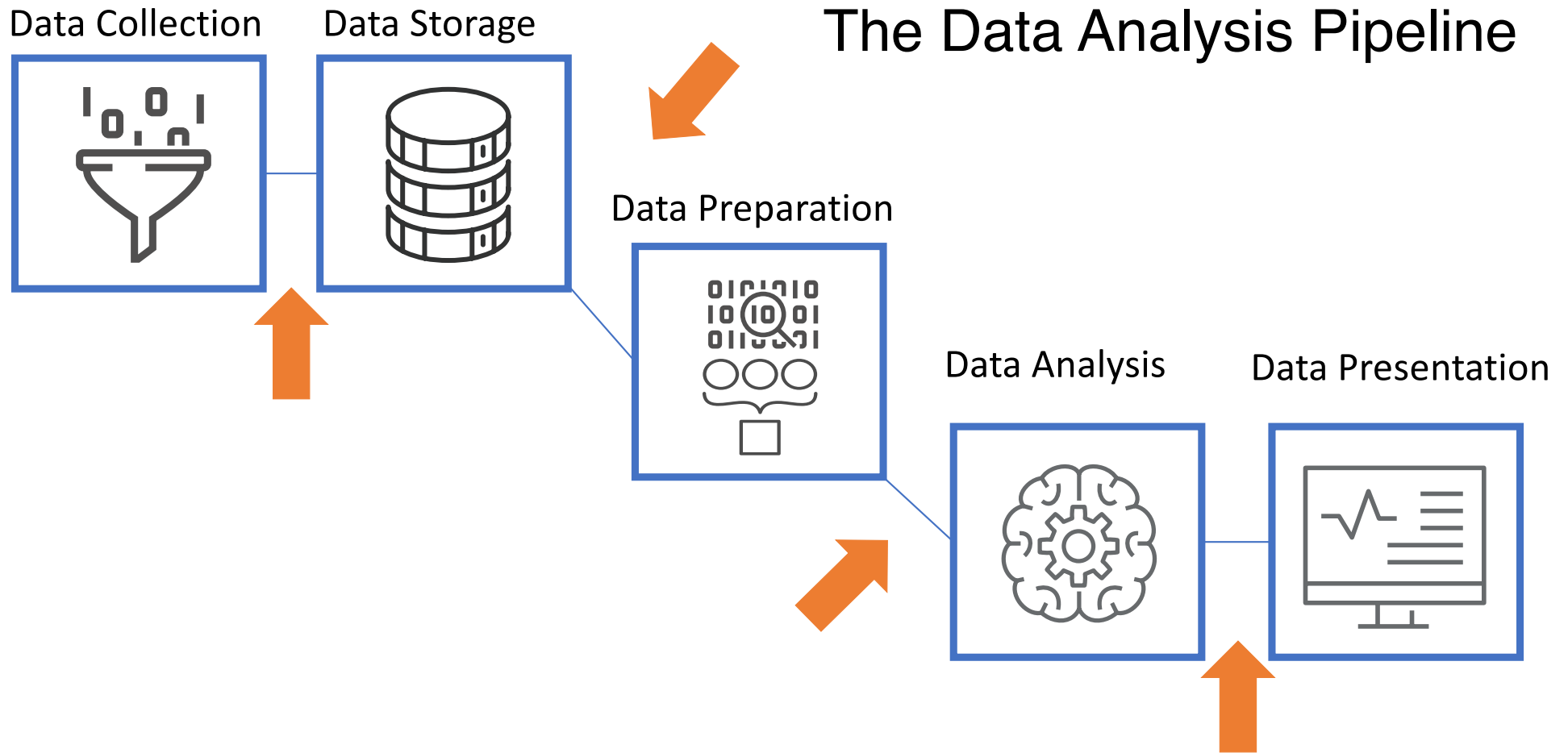
Then: Only people could carry out the activity of analysis and the components of an analysis process

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.

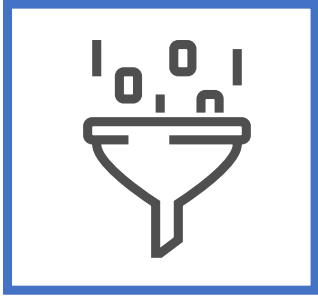
Then: A given analysis of a situation was typically seen as a one-time, one-off activity. A single person might carry out 'an analysis' and then move on.

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

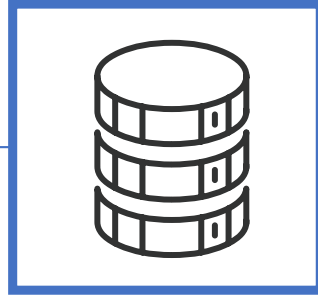
The Data Analysis Pipeline



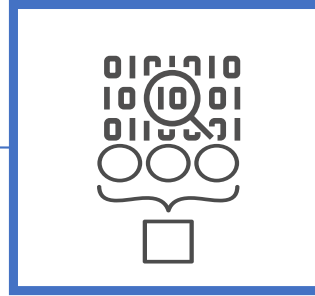
Data Collection



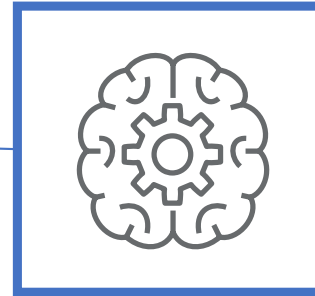
Data Storage



Data Preparation



Data Analysis



Data Presentation



Modern Data Analysis Is A Team Sport

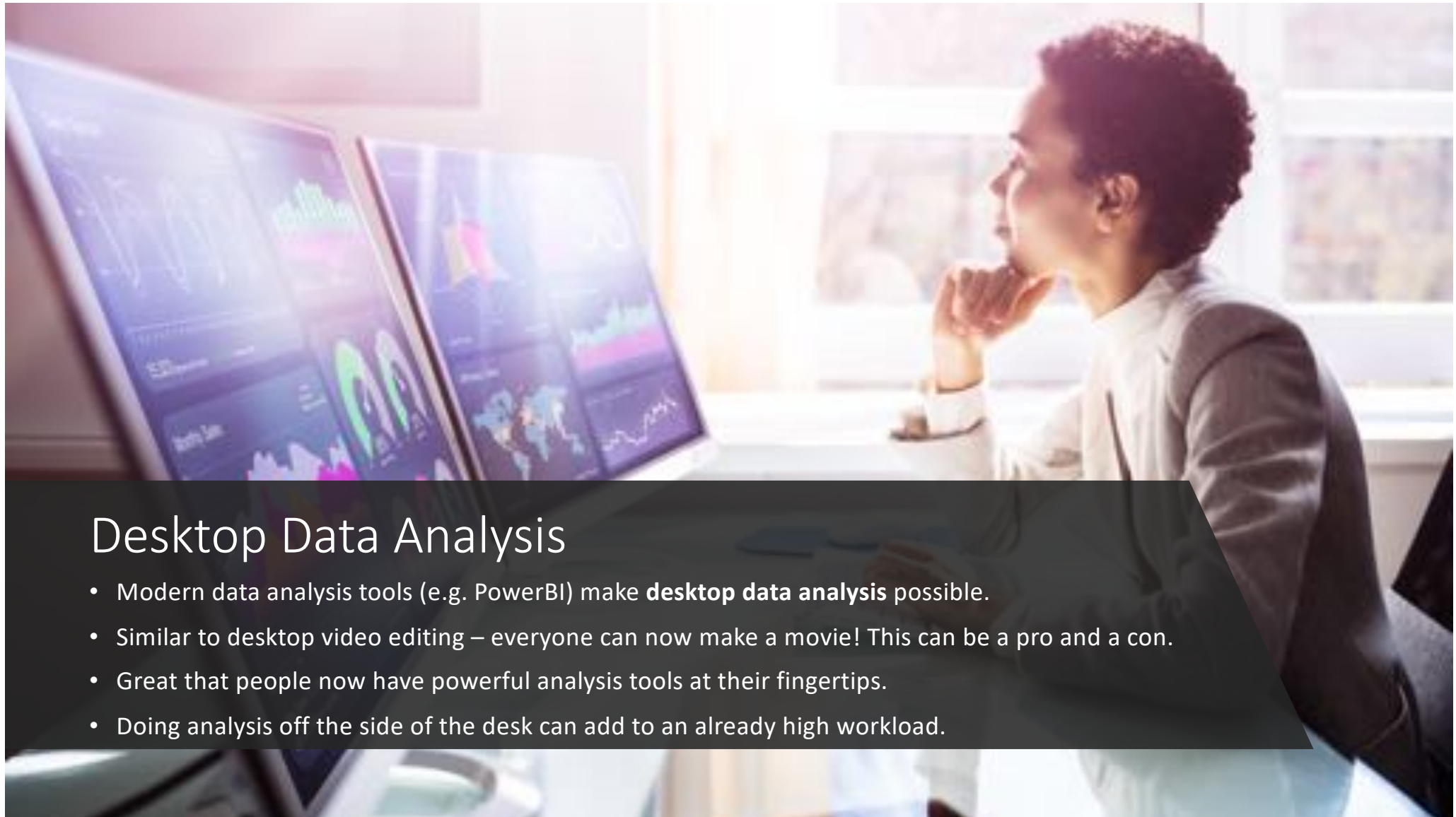


Goals For This Workshop

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

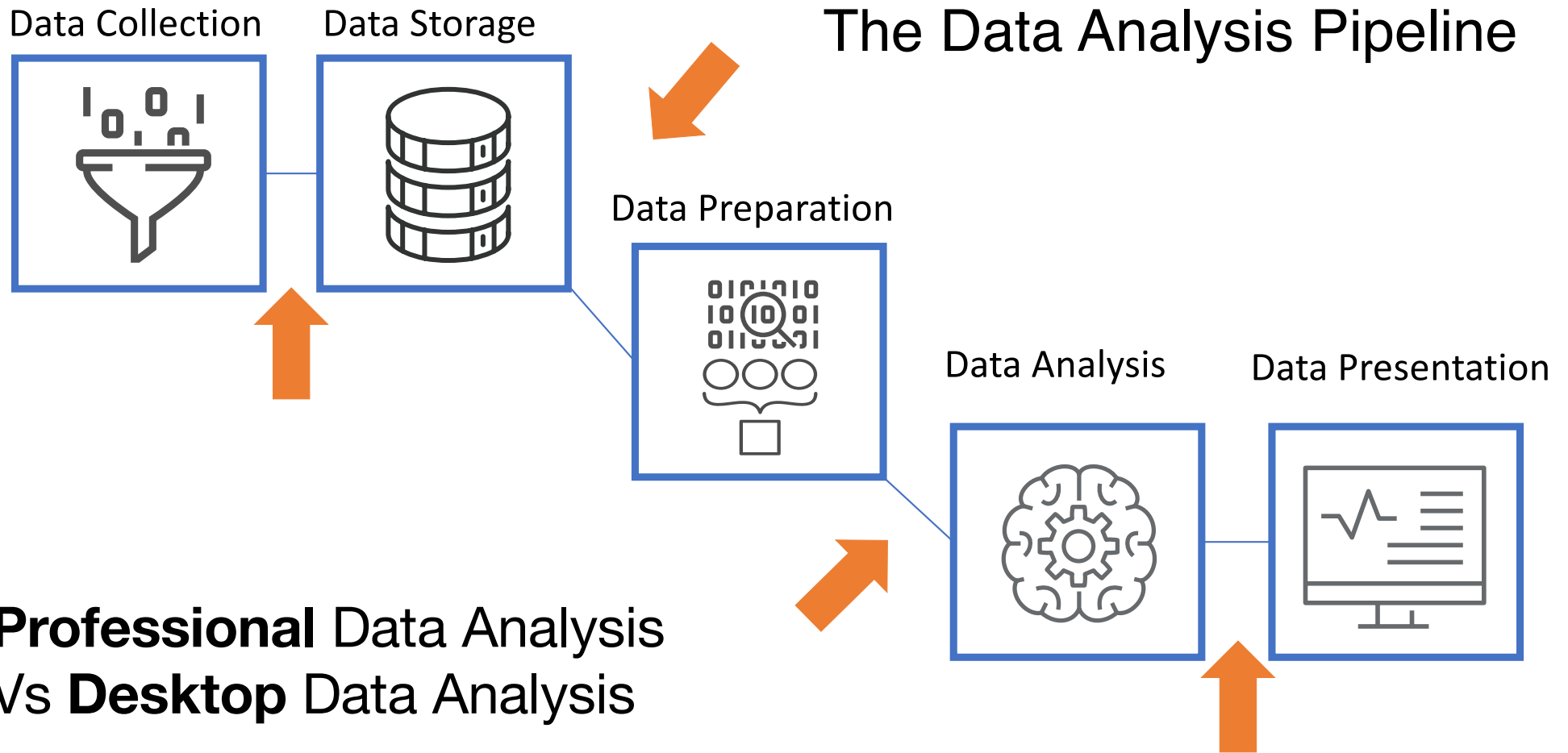
	Medicine	Cooking	The World of Cars
Amateur	Everyone First Aider	Home Cook	People who own cars Car Hobbyist
Semi-Pro	Paramedic	Bake-Sale Folks?	Semi-Pro Racer Gas Station Mechanic?
Professional	Doctor: GP, Specialist Nurse Hospital Director	Chef Pastry Chef Restaurant Owner	Garage Mechanic Body Shop Specialist



Desktop Data Analysis

- Modern data analysis tools (e.g. PowerBI) make **desktop data analysis** possible.
- Similar to desktop video editing – everyone can now make a movie! This can be a pro and a con.
- Great that people now have powerful analysis tools at their fingertips.
- Doing analysis off the side of the desk can add to an already high workload.

The Data Analysis Pipeline

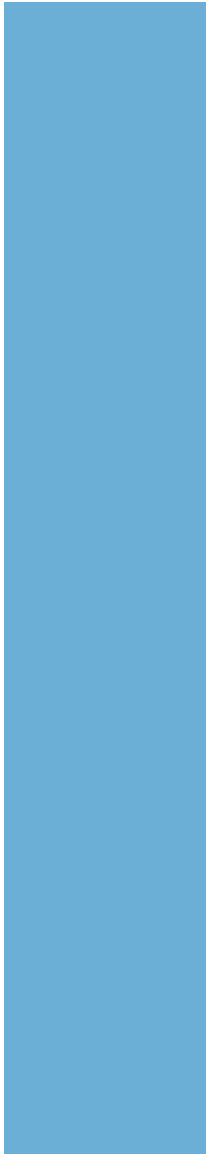
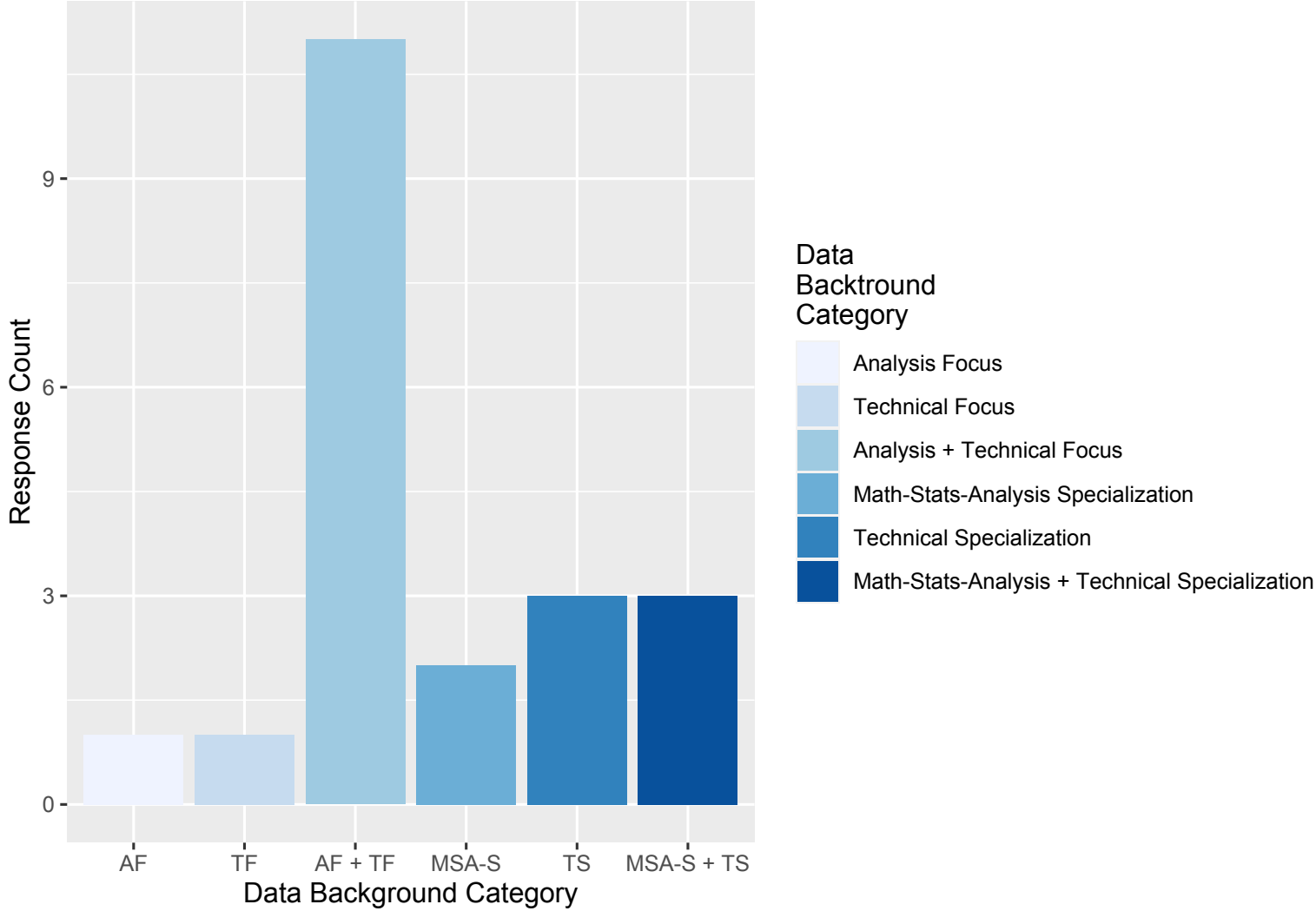


Professional Data Analysis
Vs Desktop Data Analysis

Your Team,
Your Data,
Your Questions



Background of NSERC Workshop Group (Based on Questionnaire Responses)



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)

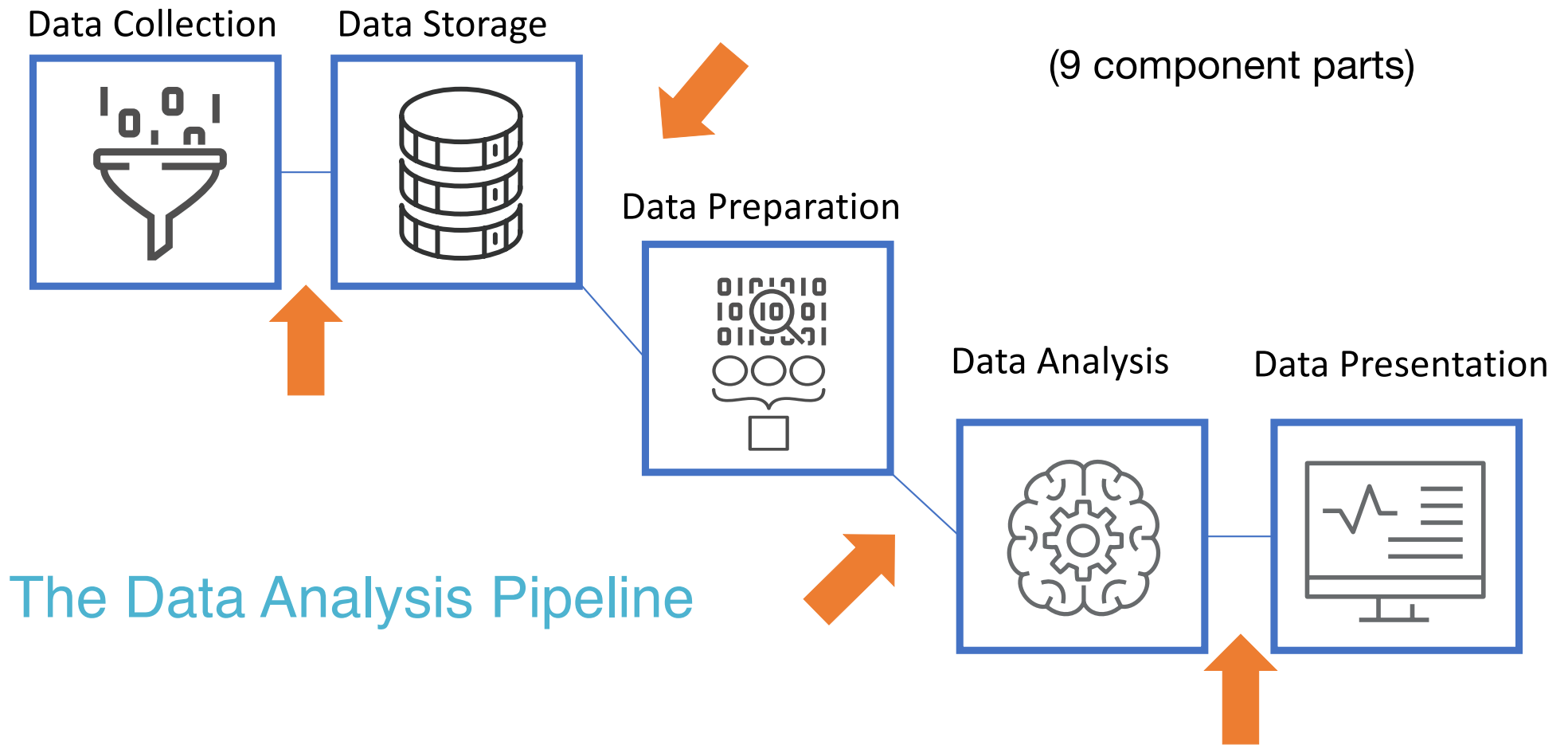
OTHER 'COMPLEX' STATISTIC TOPICS

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

NON-PARAMETRIC TESTS AND ANALYSIS

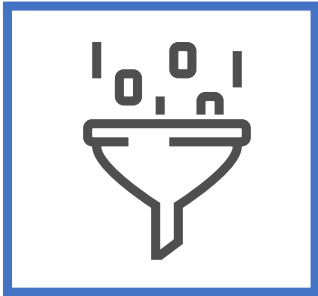
- data that we use is often not normal (we usually give more small grants than large grants), what approach would you use for the analysis?"
- ordinal variables in surveys for example (how to analyse if the results from two subgroups are statistically different, what approach would you recommend?)

Modern Data Analysis Teams and Technologies

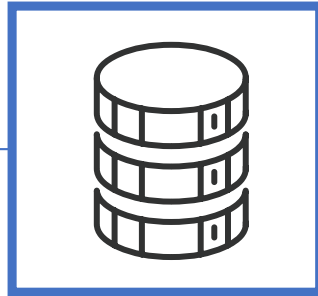


The Data Analysis Pipeline

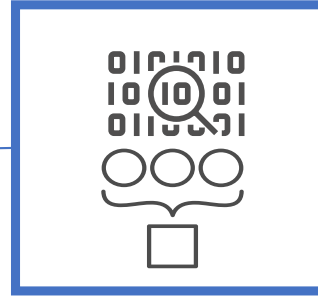
Data Collection



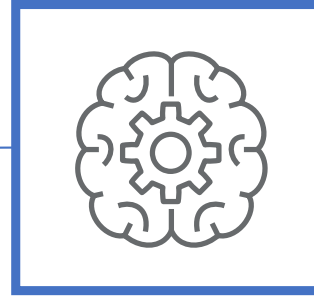
Data Storage



Data Preparation



Data Analysis

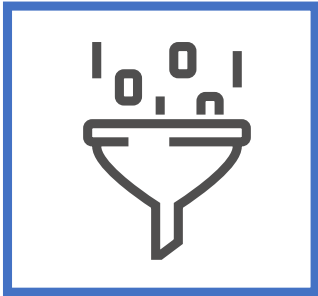


Data Presentation

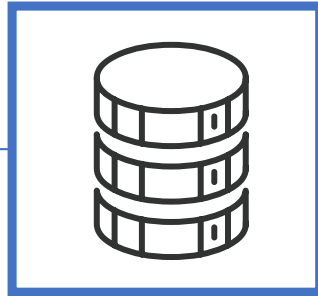


What Roles Support This Pipeline?

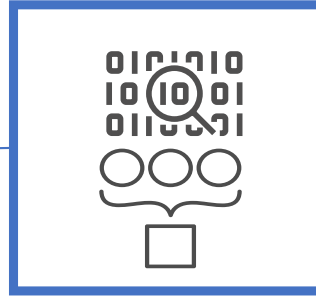
Data Collection



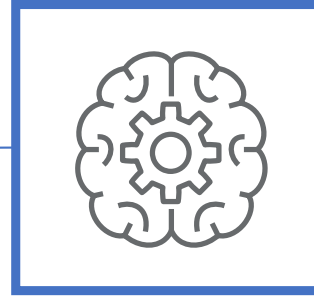
Data Storage



Data Preparation



Data Analysis

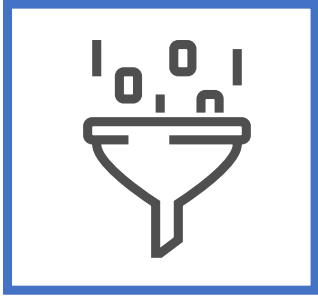


Data Presentation

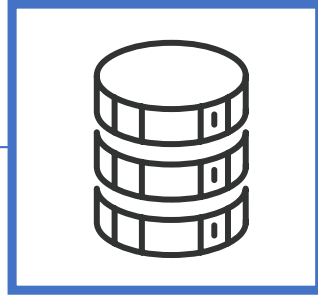


~ ratio of other team members to analysts: 10 : 1

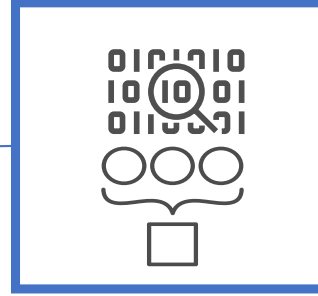
Data Collection



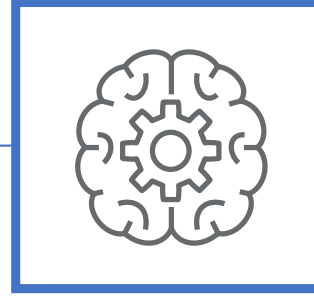
Data Storage



Data Preparation



Data Analysis



Data Presentation



~ 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.

Desktop Data Analysis Team



Let's think of desktop data analysis as 'semi-pro'



Even in an amateur or semi-pro situation, you can still build a team.



It might be less of a formal arrangement, but the different roles can all still come into play.

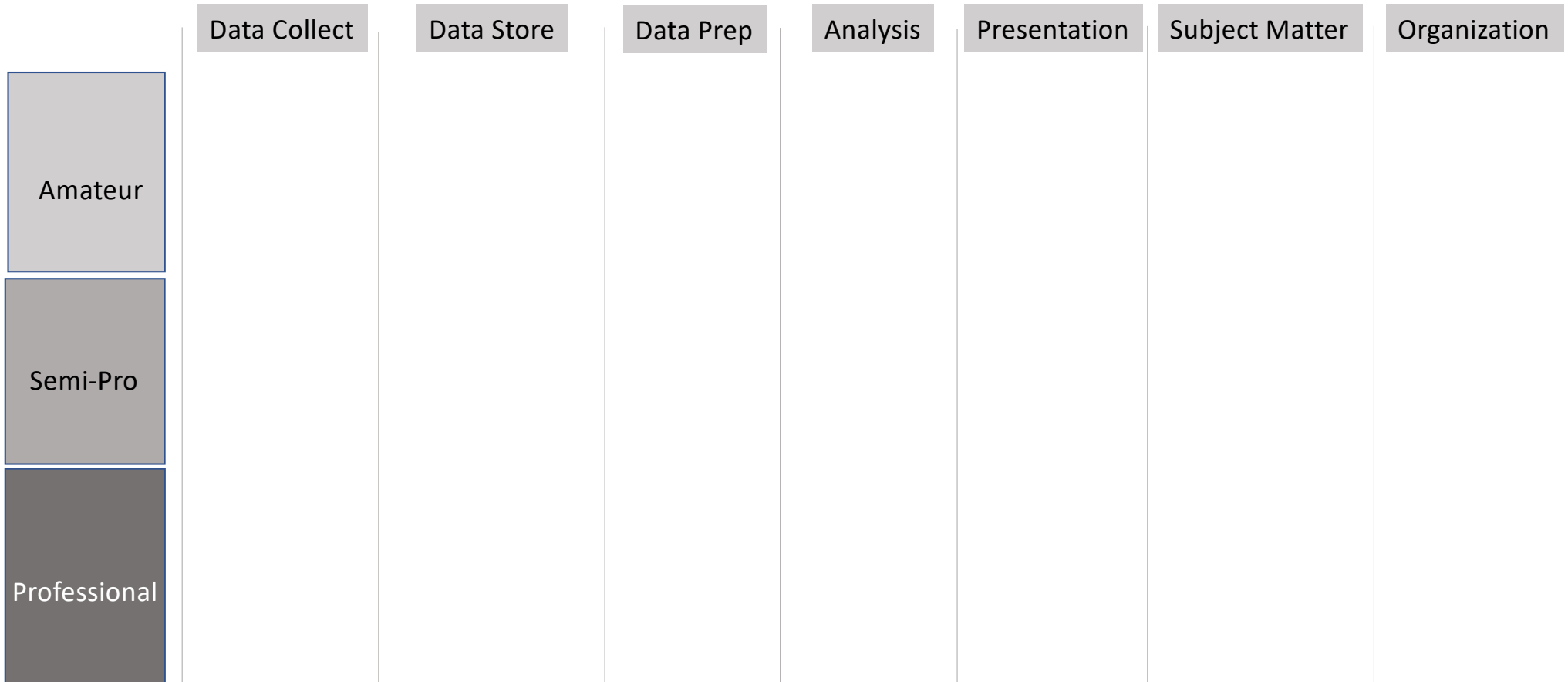


Where do
you fit in?

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?

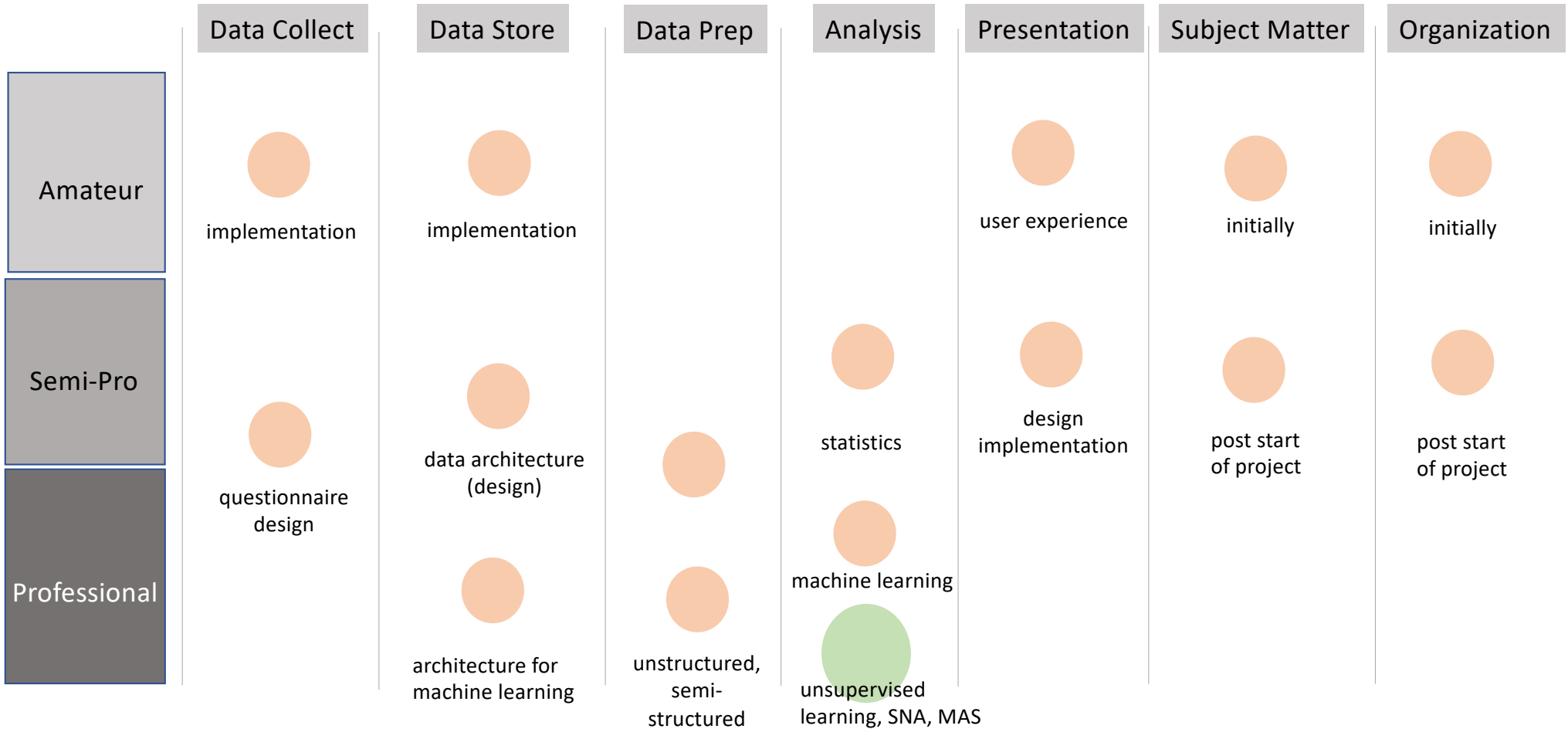
Another way to think about where you fit



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

Example of a generalist:

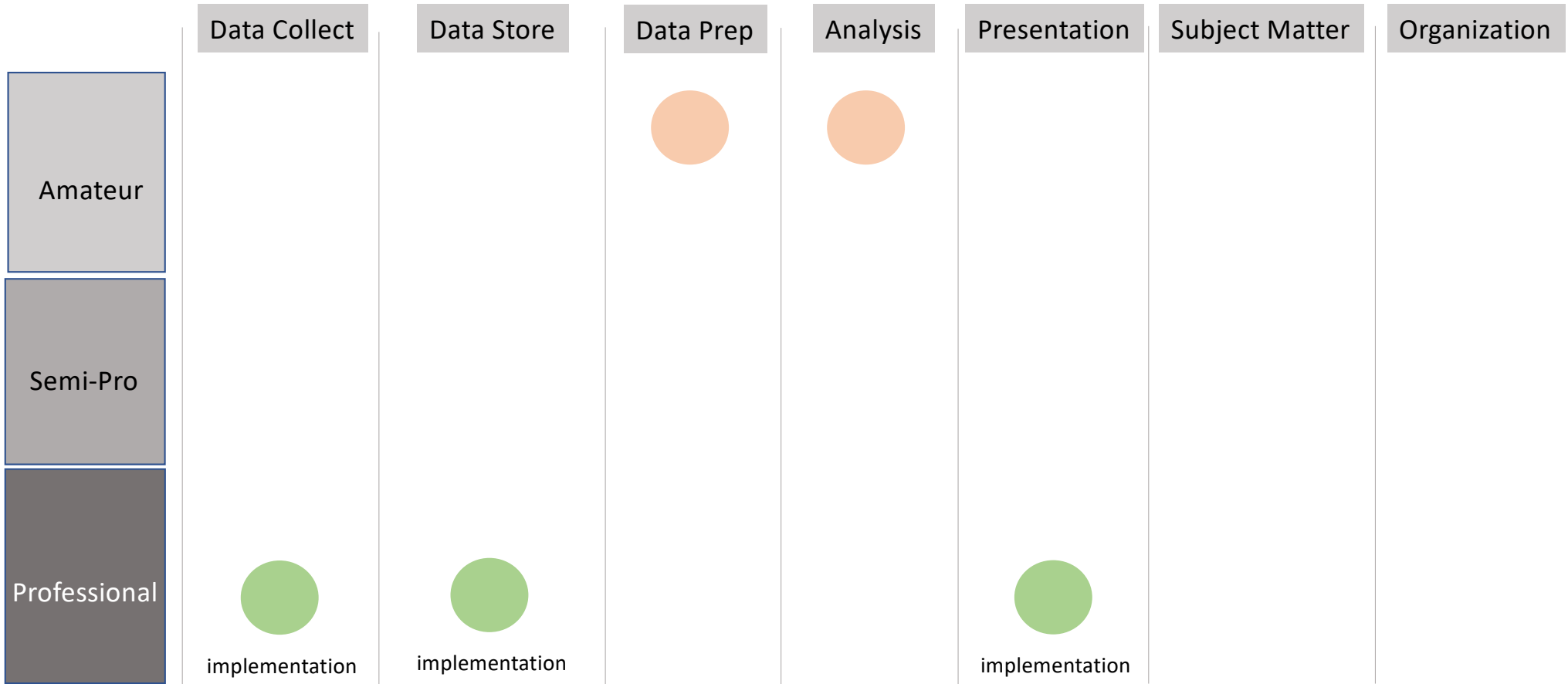
multi-purpose, can communicate across lanes,



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

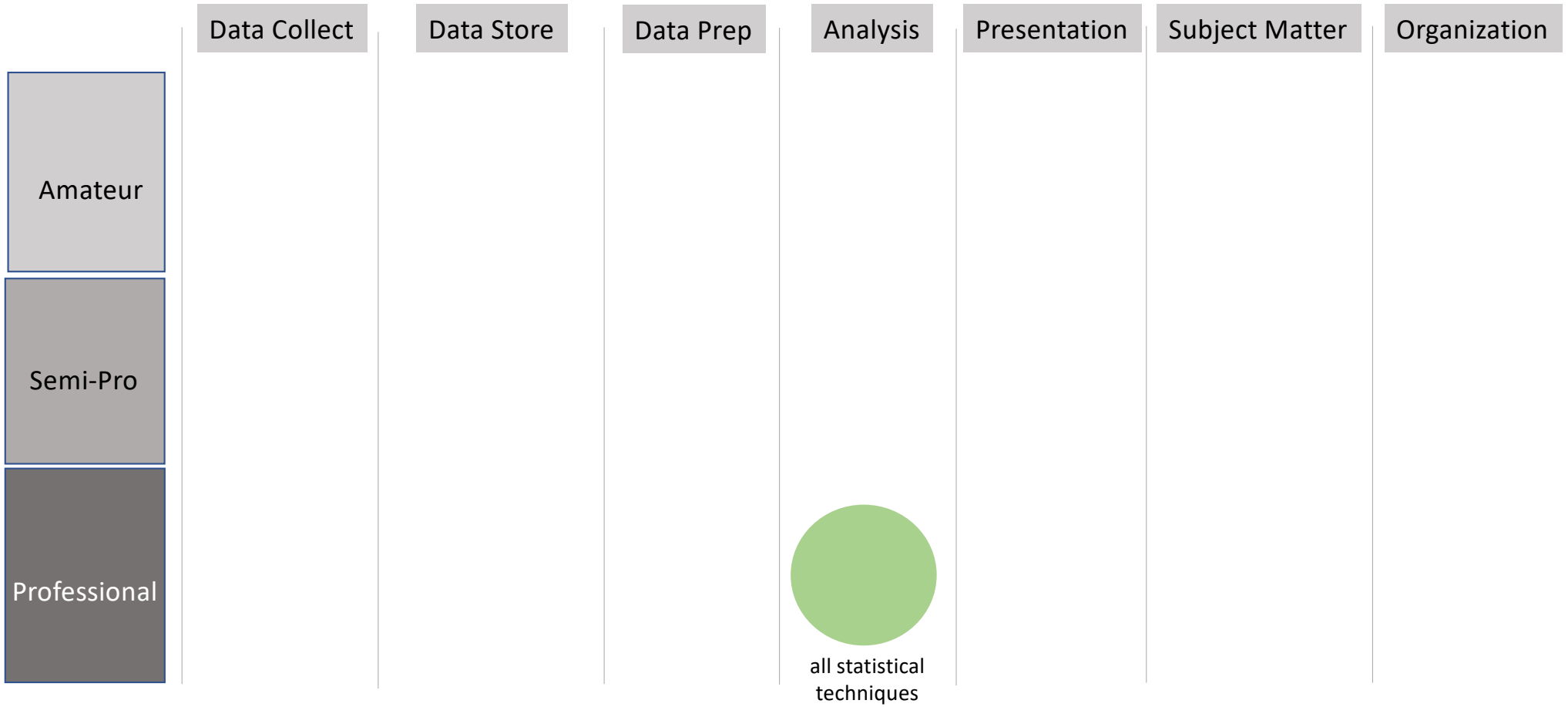
Example of a specialist (data engineer)

High quality, fast, deal with tricky situations



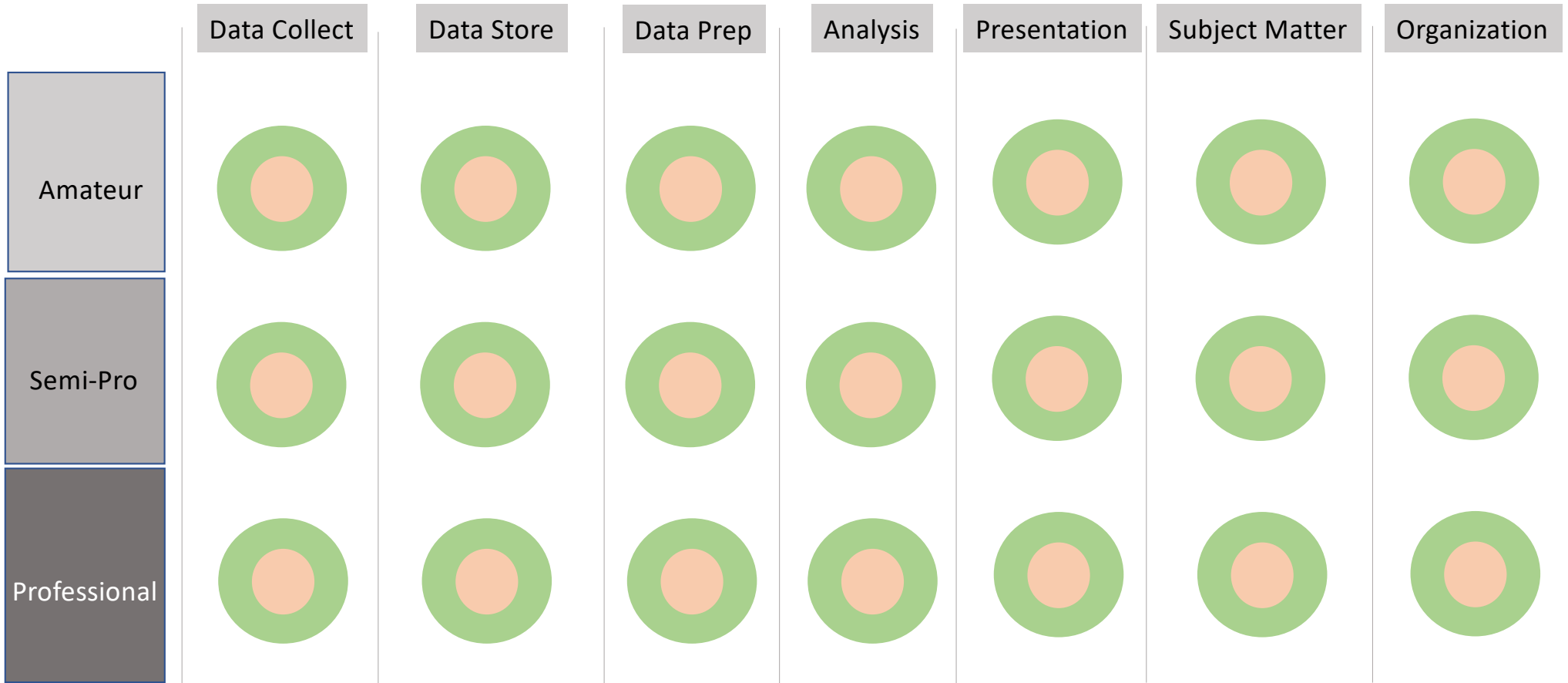
Generalists vs Specialist: You can't do it all!


Example of a specialist (statistician)



Full coverage

A team can collectively provide you with full coverage





Data Analysis - Why Me???

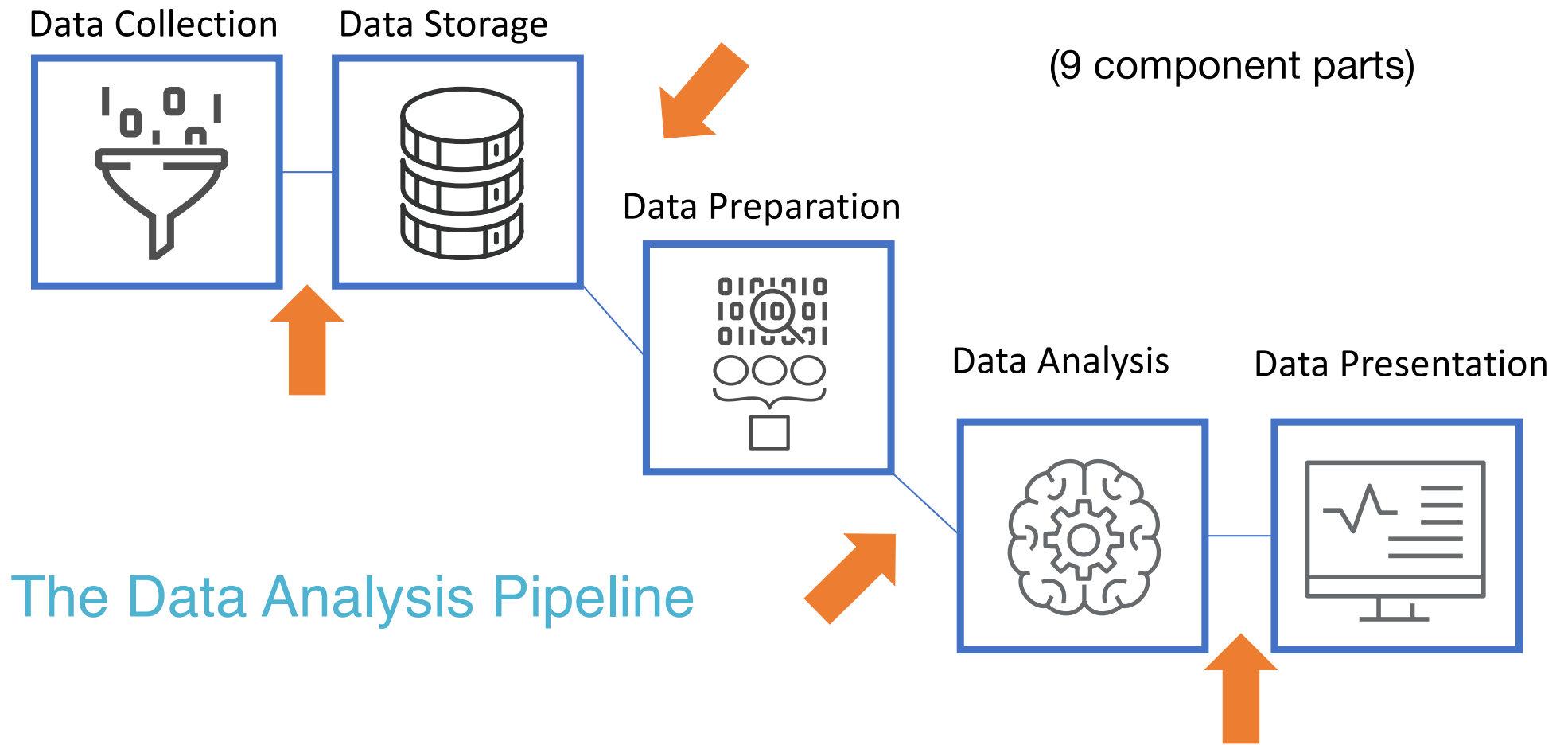
- In general, even if you are not an analyst - you must be able to talk to the analysts!
- Data engineer – person who is designing the kitchen – need to know what the cooks will be doing, who in turn need to know what people want to eat
- Data presenter – person who is doing the bodywork on the car –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

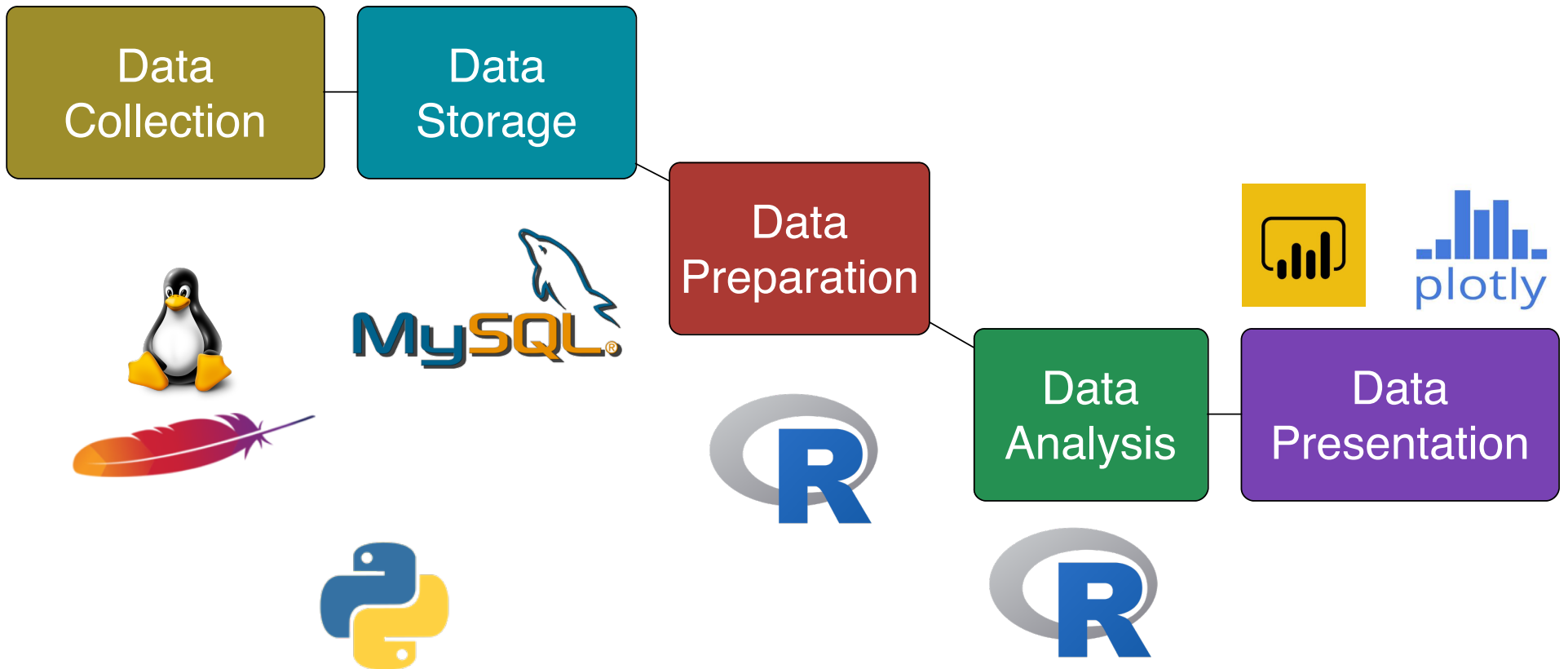
A financial candlestick chart on a blue grid background. The chart features several technical indicators: a green box highlighting a price of 104.19, another green box highlighting a price of 86.72, and a horizontal line labeled '61.6%: 99.19'. The chart shows a series of candlesticks with vertical lines representing price ranges, and a curved line representing a trend or moving average.

Modern Data Analysis Technologies

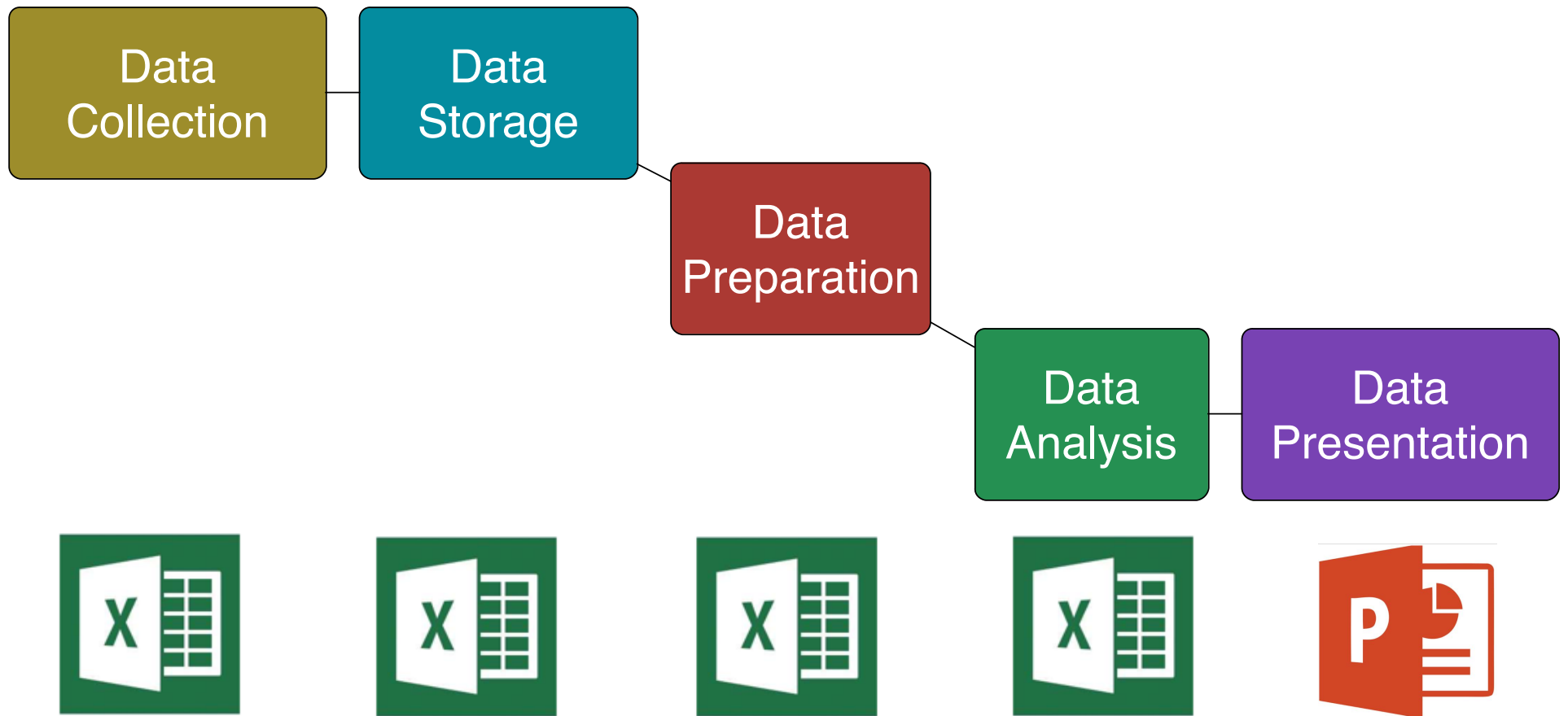


The Data Analysis Pipeline

A Open-Source Driven Data Stack



Typical GoC Data Stack



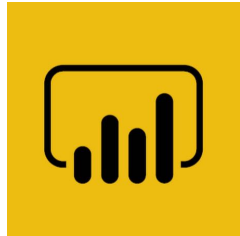
Data Collection

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

Understanding the Cloud Landscape



IaaS: Infrastructure as a Service



PaaS: Platform as a Service



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

Pipeline Creation Phases

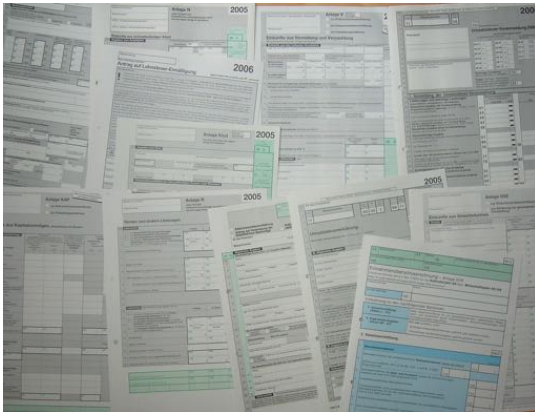
1. **Research + Design**
2. **Implementation**
3. **Testing**
4. **Production + Management**
5. **Research + Design**

Agile!



Pre-Analysis: Data Collection, Structuring and Preparation

Collection: Three Main Data Sources



Recordkeeping

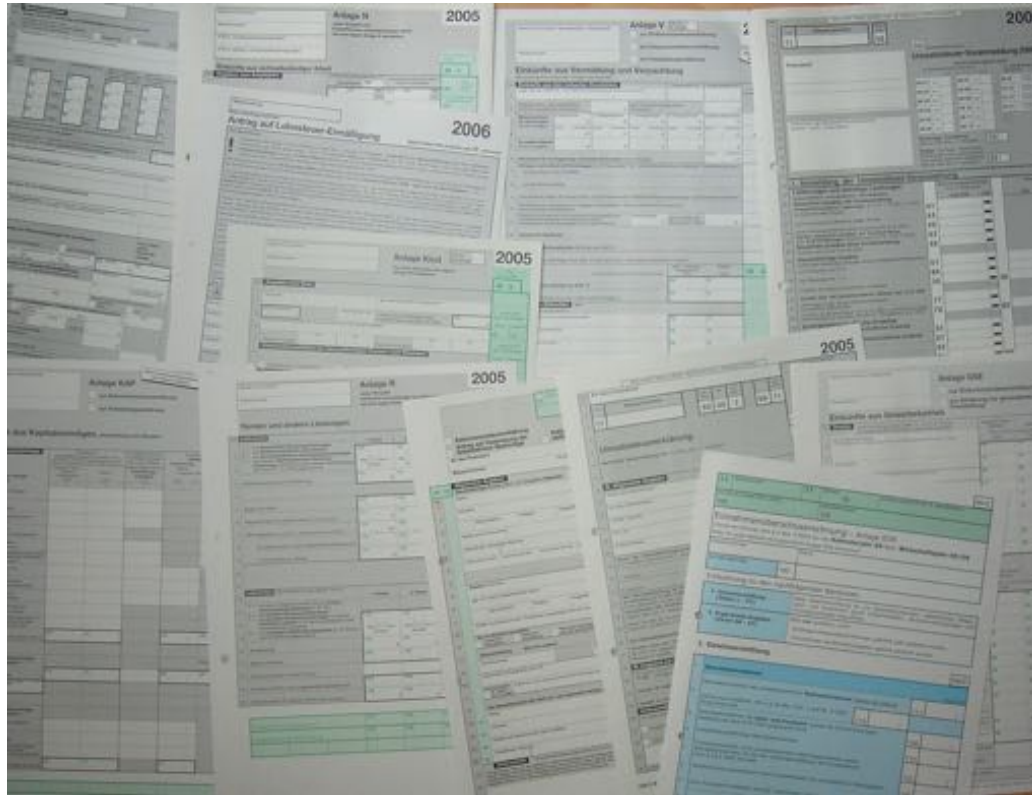


Research



Sensors/Monitoring

Recordkeeping: Primary Focus On Specific Entities





The Curse of Categorical Data

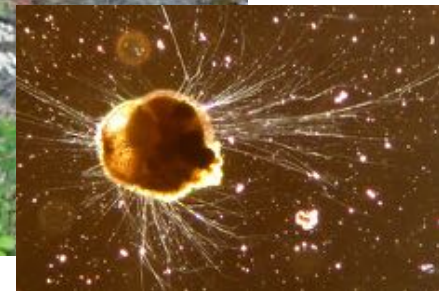
Government data tends to be very heavy on the categories and text data.

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



Column 1	Column 2	Column 3	Column 4	Column 5
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20
21	22	23	24	25
26	27	28	29	30
31	32	33	34	35
36	37	38	39	40
41	42	43	44	45
46	47	48	49	50
51	52	53	54	55
56	57	58	59	60
61	62	63	64	65
66	67	68	69	70
71	72	73	74	75
76	77	78	79	80
81	82	83	84	85
86	87	88	89	90
91	92	93	94	95
96	97	98	99	100

