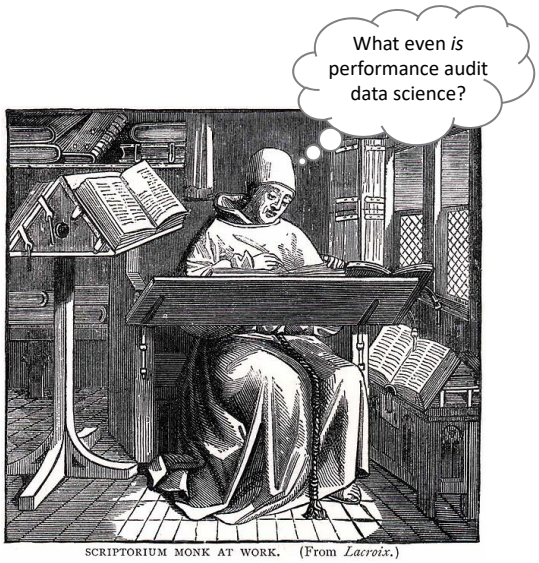
Data Science in Performance Audit IMPACT 2020

Xiaoyan Lu (ANAO)
Stephen McAteer (VAGO)
24 March 2020

Australian National Audit Office
VAGO

IMPACT 2020

Mission: write The Book on performance audit data science



SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

2

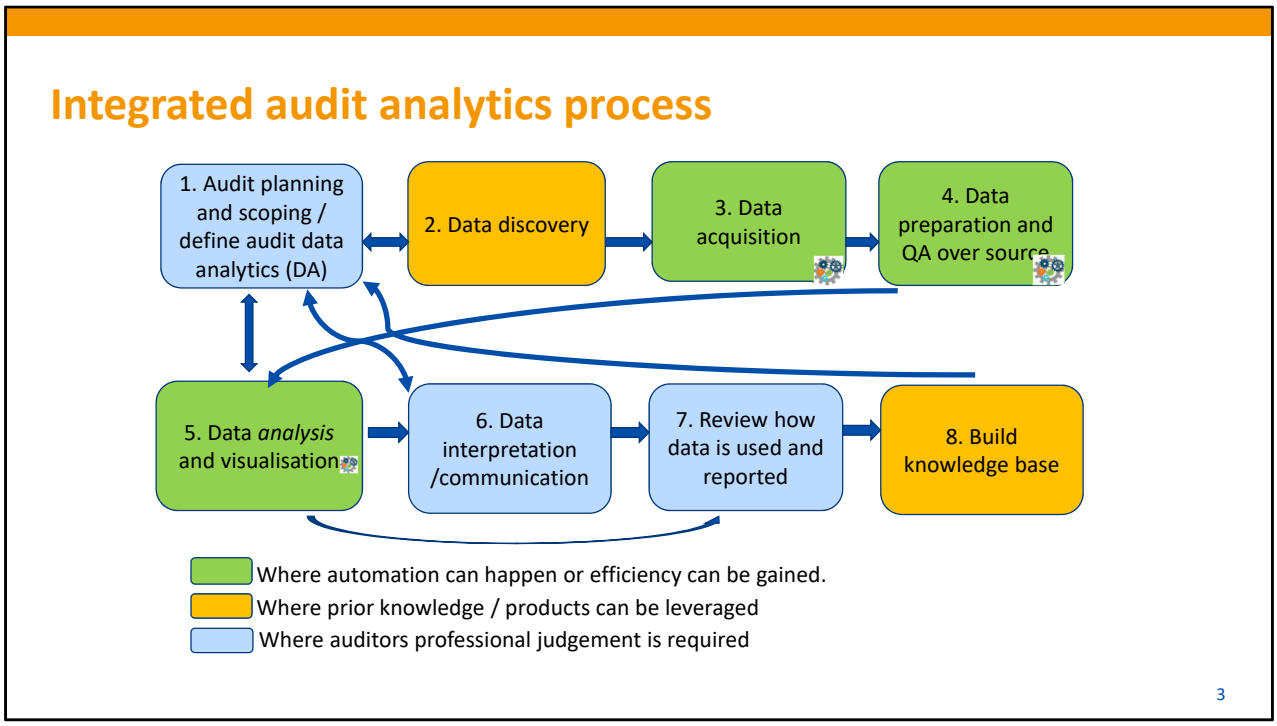
State the problem:

- Performance audit data science is still emerging
- Our job is to figure out what it is (while actually delivering results!)
- More developed areas are: reporting, decision support (incl. operations research), product development
- None of those paradigms quite fit performance audit

We will discuss:

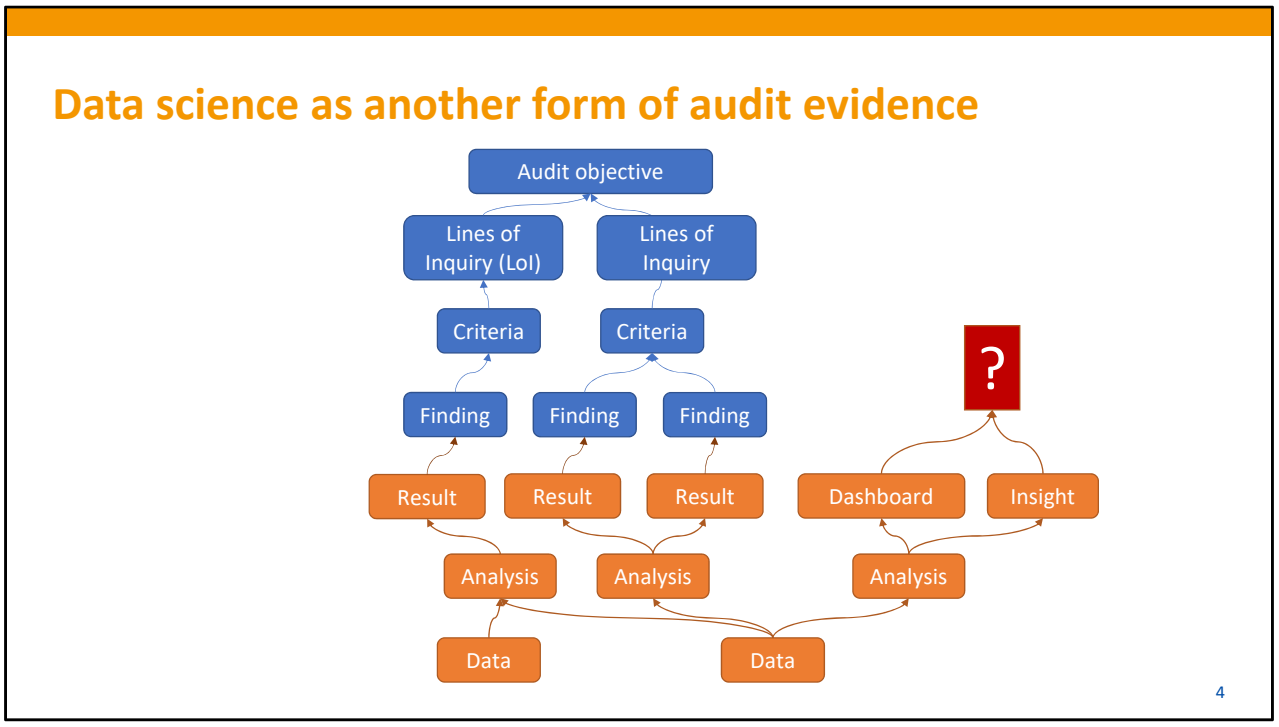
- The journey at both Victorian Auditor-Generals Office (VAGO) and Australian National Audit Office (ANAO) – quite different!
- What we have learnt so far
- What we know we don't know yet

A call to arms – help us figure it all out!



Types of services

- Contextual analysis for audit planning/scoping
- Data extraction, transformation and loading (ETL)
- Data preparation (assurance over data completeness and accuracy)
- Data analysis
- Review of entity's data related technical documentation to provide insights or assurance
- Leverage existing system assurance and/or data analytics work from other audits
- Review of performance audit team's data analysis work
- Process automation
- Data visualisation
- Data analytics capability development

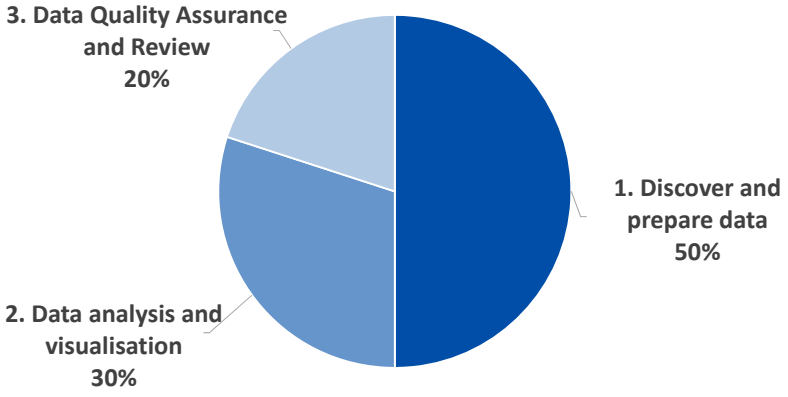


We exist to **service the audit** … the audit exists to address the **audit objective**
I call this a "value chain" (in decision support the blue stuff is different, with **organisation strategic objective** at the top)
If we don't have all links in the chain strong, there is a **risk of irrelevance** (= waste/inefficiency)
How do you ensure a strong chain? Engage **early**, engage **often**

Annual planning – influence the objective
Audit planning
- influence the Lines of Inquiry (LoI)
- understand the agency data
- consider possible analysis
- align analysis to criteria (via potential finding)
- request data (adjust approach if needed)

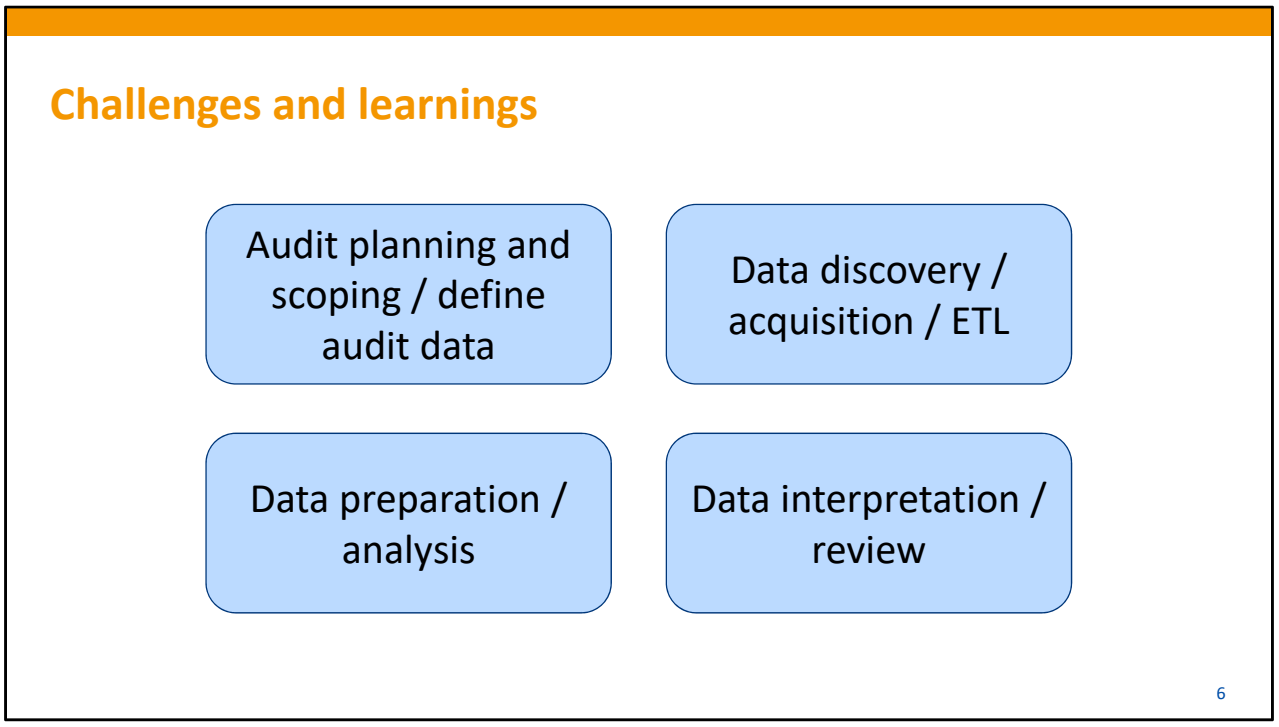
Conduct
- run analysis
- produce results (incl. vizsualization, &c.)

Time allocation for audit analytics



5

Not just analysis and visualization!

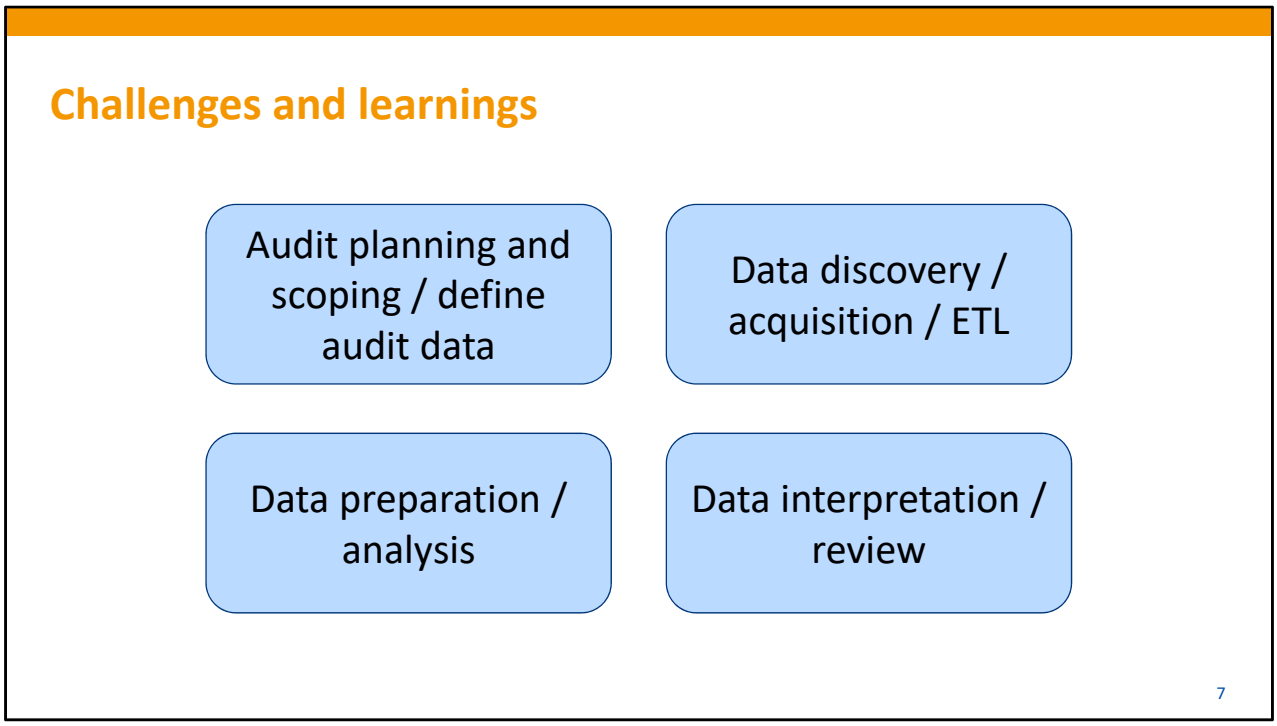


Audit planning and scoping / define audit DA

- Too many unknowns and uncertainties in this stage
- Misaligned expectations when undertaking atypical work
- **Recognising the test program may evolve so is analysis plan**
- **Be realistic about capabilities and capacities**
- **Clarify areas of misunderstandings as soon as they are identified**

Data discovery / acquisition / ETL

- Challenges to obtain quality data in a timely manner
- Unexpected large size and complexity of data provided
- No sufficient data dictionary, system documentation or data relationship model for complex datasets
- **Make decisions early for Plan B/alternative approach to manage risks on data dependency**
- **Make effective use of walkthrough meetings and communicate data requests that could be understood by all involved (no ambiguity)**
- **Leverage technology to transform data into usable formats to facilitate meaningful analysis**



Data preparation / analysis

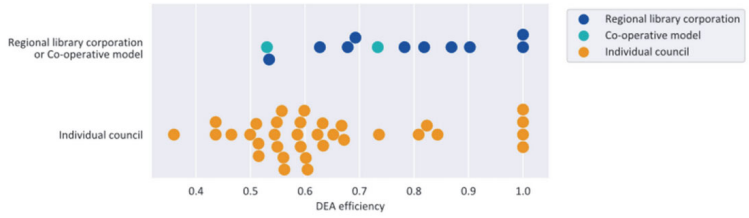
- Providing assurance over data completeness and accuracy
- Translating business rules into quantifiable data modelling
- Assumptions have to be made to enable analysis for a specific audit criteria without subject matter expertise (SME)
- **Identifying independent sources for reconciliation/verification**
- **Identifying Subject Matter Expert (SMEs) to explain business rules or verify assumptions**
- **Ensuring agreement are made with Audit Manager prior to starting the analysis work.**
- **Communicating the need to use "proxy" field is due to a data limitation.**
- **Recognising when the discrepancies can or cannot be explained by data itself**

Data interpretation / review

- Analysts are not directly contributing to the report preparation phase
- There is a risk of the data analytics output being misinterpreted
- **Agreeing on key terminology**
- **Reviewing the report to ensure data is interpreted correctly**
- **Documenting the analytical process in an easily traceable way**
- **Do leave sufficient time to review**

Case study: Library efficiency (data envelopment analysis – VAGO)

Figure 2C
Efficiency of shared service models



8

Library efficiency

- Evaluating the efficiency of a process with heterogeneous inputs and outputs
- Do different library models perform differently?
- What factors drive efficiency?
- Report: https://www.audit.vic.gov.au/report/council-libraries

Case study: do what was considered 'impossible'

Purpose:

- Use a data analytics approach to confirm the completeness and accuracy of an entity's revenue.

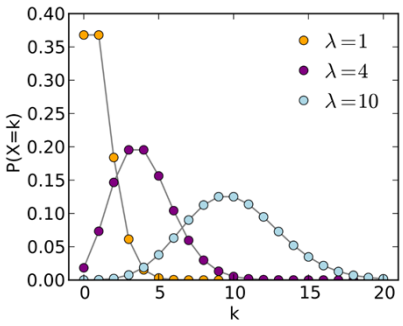
Challenges:

- The complexity of the systems involved made a controls approach less effective and the quantum of data involved and complexity of business rules made sampling and manual testing time consuming and impractical.
- No one unique identifier across all systems

Our approach and outcomes

- Translated business rules (fee charging schedule) into a series of data model
- Developed a sustainable and reliable solution to provide data assurance across multiple systems
- Designed for re-execution with minimal human intervention
- Confirmed data integrity, completeness and accuracy of calculated revenue, confirmed revenue is actually collected.
- Accurate to 99%+ to provide high data assurance
- Increase in quality and efficiency
- This approach is extended to recalculating key performance indicators

Case study: Flexible safety barrier programme evaluation (Poisson process modelling, geospatial data – VAGO)



10

Flexible safety barrier effectiveness

- Modelling of crash occurrence using Poisson modelling
- What uncertainties exist in the data?
- Are the barriers effective? We found that the programme was unlikely to achieve proposed benefits.
- Geospatial data was used to calculate before and after fatality and serious injury data.
- Report: https://www.audit.vic.gov.au/report/safety-victorias-roads-regional-road-barriers

Case study on a solution that drives efficiency: Do it once and use by all

Identified high value datasets

- AusTender data which contains 10+ years information regarding government procurement contract notices over $10k, with rules governed by the Commonwealth Procurement Rules (CPRs)
- Government procurement panel lists
- ABN lookup data

Developed an Information Report and a self-service visualisation tool

When these data combined, it can be used to explore:

- Trends at whole-of-government level and by entity over ten years (top and dominant providers)
- Compliance to CPRs
- Insights to procurement behaviors – use of panel
- Insights to contract management – use of amendments
- Used as a population to draw sample for performance audits
- What is included vs what is not included

This report can be accessed via: https://www.anao.gov.au/work/information/australian-government-procurement-contract-reporting-update-2019

Case study on a bespoke solution: Process automation

The Challenge:

- The audit team requested assistance in automating the extraction of key data from a Word format Project Summary Sheet for a project.

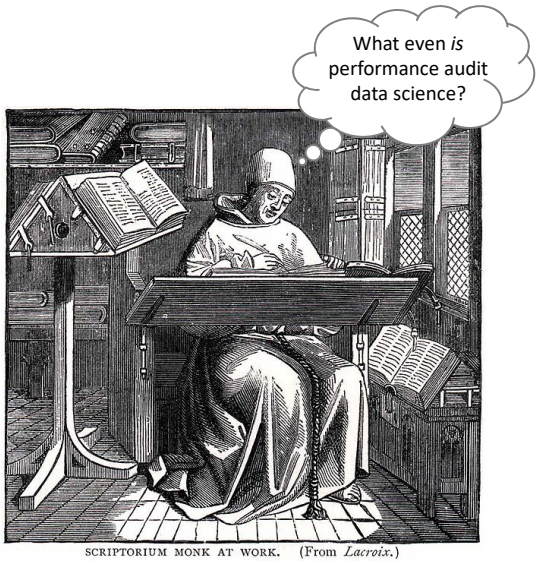
Our approach:

- Understanding the pain point of the audit team
- Using Python, a pipeline was generated to "read" through relevant documents and pull out information of interest
- The extracted data was then compile onto a structured spreadsheet.

Outcome:

- The automated solution is now handed back to the audit team with a user-guide
- A new pipeline from data extraction to data visualisation is being developed.

A call to arms: Let's write The Book together!



SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

13

Dedicated performance audit data science is small at each office.

We can collaborate to figure it out.

- Tools
- Techniques
- Processes
- Position descriptions