Technical and Non-Technical Aspects of Data Work

DATA SCIENCE ESSENTIALS

Session 1

1. Technical & Non-Technical Aspects of Data Work

Quantitative Skills

Out-of-academia context:

- apply **quantitative methods** to (business) problems in order to obtain **actionable insight**
- difficult for any given individual to have expertise in **every** field of mathematics, statistics, computer science, data science, data engineering, etc.

With a graduate degree in math/stats, for instance:

- **expertise** in 2-3 areas
- **decent understanding** of related disciplines
- **passing knowledge** in various domains

Flexibility is an ally, perfectionism... only up to a point.

Quantitative Skills

Suggestions:

- **keep up with trends**
- become **conversant in your non-expertise areas**
- know **where to find information**

In many instances (70%?), only the basics (2^{nd}–3^{rd} year mandatory courses at uOttawa, say) are sufficient to meet government/industry needs.

Focus: make sure you really **understand** the basics, stepping stones.

In the rest of the cases, more sophisticated knowledge is required.

Quantitative Skills

- survey sampling and data collection
- data processing and data cleaning
- data visualization
- mathematical modelling
- statistical methods
- regression analysis
- queueing models
- machine learning
- deep learning
- reinforcement learning
- stochastic modelling (MC simulations)
- optimization and operations research
- survival analysis
- Bayesian data analysis
- anomaly detection and outlier analysis
- feature selection/dimensions reduction
- trend extraction and forecasting
- cryptography and coding theory
- design of experiment
- graph and network theory
- text mining/natural language processing
- etc.

Software and Tools

Modern quantitative work typically involves **programming** (or the use of point-and-click software, at the very least).

But programming languages **go in and out of style.**

It is important not just to understand the syntax of a particular language, but also how computer languages and computing infrastructure work in general.

ALSO: avoid getting caught up in programming wars … they're more or less all functionally equivalent!

Software and Tools

Programming (and Related)

- Python, R, C/C++/C#, Perl, Julia, regexps (, Visual Basic?), Java, Ruby, etc.

Database Management

- SQL and variants, ArangoDB, MongoDB, Redis, Amazon DynamoDB (, Access?), Big Query, Redshift, Synapse, etc.

Data Visualization

- ggplot2, seaborn, plot.ly, Power BI, Tableau, D3.js, Google Data Studio, proprietary software, etc.

Simulations, Statistical Analysis, Data Analysis, Machine Learning

- tidyverse, scikit-learn, numpy, pandas, scipy, MATLAB, Simulink, SAS, SPSS, STATA (, Excel?), Visio, TensorFlow, keras, Spark, Scala, etc.

Typesetting and Reporting

- LaTeX, R Markdown, Adobe Illustrator, GIMP (, Word?, PowerPoint?), etc.

Software and Tools

Q: At StatCan, R or SAS?

A: Not easy to answer as StatCan is in a slow transition period. The Agency is better equipped for SAS (with "Big Data" options, such as SAS Grid).

R is [...] not as ideal for large files (e.g., Census data), so it is not an option in such cased because it is still too slow (unless you have very powerful servers). But we would prefer to use the R packages, so it's a dilemma.

TL;DR: R is our future, but SAS is still very much our present. In times of transition, **analysts/employees who know both are better positioned**.

Multiple I's Approach to Data Work

Technical and quantitative proficiency (or expertise) is **necessary** to do good quantitative work *in the real world*, but it is **not sufficient** – optimal real-world solutions may not always be the optimal academic or analytical solutions.

This might be the biggest surprise for those transitioning out of academia.

What works for one person, one job application, one project, one client, etc. may not work for another – **beware the tyranny of past success**!

The focus of quantitative work must include the delivery of **useful analyses/ products** (Multiple "I"s).

Multiple I's Approach to Data Work

- **intuition**
  understanding context
- **initiative**
  establishing an analysis plan
- **innovation**
  new ways to obtain results, if required
- **insurance**
  trying multiple approaches
- **interpretability**
  providing explainable results
- **insights**
  providing actionable results
- **integrity**
  staying true to objectives and results
- **independence**
  self-learning and self-teaching
- **interactions**
  strong analyses through teamwork
- **interest**
  finding and reporting on interesting results
- **intangibles**
  thinking "outside the box";
- **inquisitiveness**
  not only asking the same questions repeatedly

Multiple I's Approach to Data Work

Prospective employees/analysts are not solely gauged on technical know-how, but also on the ability to **contribute positively** to the workplace/project:

- communication
- team work and multi-disciplinary abilities
- social niceties and flexibility
- non-technical interests

Employers rarely chose robots when human beings are available; stakeholders are more likely to accept data recommendations from **well-rounded people**.

You should also evaluate eventual employers/clients on these axes.

Roles and Responsibilities

A data analyst or a data scientist (in the **singular**) is unlikely to get meaningful results – there are simply too many moving parts to any data project.

Successful projects require **teams** of highly-skilled individuals who understand the **data**, the **context**, and the **challenges**.

Team *size* could vary from a few to several dozens; typically easier to manage small-ish teams (with 1-4 members, say, with **role overlaps**).

Domain Experts / SMEs

- are authorities in a particular area or topic
- guide team through unexpected complications and knowledge gaps

Roles and Responsibilities

Project Managers / Team Leads

- understand the process enough to recognize whether what is being done makes sense
- provide realistic estimates of the time and effort required to complete tasks
- act as intermediary between team and clients/stakeholders
- responsible for project deliverables.

Data Translators

- have a good grasp on the data and the data dictionary
- help SMEs transmit the underlying context to the data science team

Data Engineers / Database Specialists

- work with clients and stakeholders to acquire useable data sources
- may participate in the analyses, but are not necessarily specialists

Roles and Responsibilities

Data Analysts

- clean and process data
- prepare initial visualizations
- have a decent understanding of quantitative methods (at most 1 area of expertise)
- conduct preliminary analyses

Data Scientists

- work with processed data to build sophisticated models
- focus on actionable insights
- have a sound understanding of algorithms/quantitative methods (2 or 3 areas of expertise)
- can apply them to a variety of data scenarios
- can be counted on to catch up on new material quickly

Roles and Responsibilities

Computer Engineers

- design and build computer systems and pipelines
- are involved in software development and deployment of data science solutions

AI/ML Quality Assurance/Quality Control Specialists

- design testing plans for solutions that implement AI/ML models
- help the team determine whether the models are able to learn

Communication Specialists

- communicate actionable insights to managers, policy analysts, decision-makers, stake holders
- may participate in the analyses, but are not necessarily specialists (often data translators)
- keep abreast of popular accounts of quantitative results and developments

Analysis Cheat Sheet

1. Business solutions are not always academic solutions.
2. Data and models don't always support the stakeholder's hopes, wants, needs.
3. Timely communication is key – externally and internally.
4. Data scientists need to be flexible (within reason), and willing and able to learn something new, quickly.
5. Not every problem calls for data science methods.
6. We should learn from both our good and our bad experiences.

Analysis Cheat Sheet

7. Manage projects and expectations.
8. Maintain a healthy work-life balance.
9. Respect the stakeholders, the project, the methods, and the team.
10. Data science is not about how smart we are; it is about how we can provide actionable insight.
11. When what the client wants can’t be done, offer alternatives.
12. “There ain’t no such thing as a free lunch.”

Suggested Reading

Technical and Non-Technical Aspects of Data Work

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

13. Non-Technical Aspects of Quantitative and Data Work

13.1 First Principles

- The "Multiple I" Approach
- Roles and Responsibilities
- Analysis Cheatsheet

13.3 Lessons Learned

Exercises

Technical and Non-Technical Aspects of Data Work

1. Which of the quantitative skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
2. Which of the software skills presented in this section do you possess? Which interest you? Which do you plan on learning about?
3. What data role do you hold in your organization? Which role do you think you are currently best suited for? Which role do you aspire to?
4. Have you encountered the Analysis Cheat Sheet lessons in your work? Have you encountered others?