LEARN.
CONNECT.
ELEVATE.

Data Science Capstone Project
Dr. Alejandro Gutierrez Lopez, PMP
Session 3

McGill | School of Continuing Studies

mcgill.ca
/continuingstudies

Agenda

- Present your solutions for the first project
- Review the second capstone project
- Process from problem to insight

SECOND PROJECT

Project A – Defined question

- Analysis of the firefighters dataset (and other connected datasets)

- Objective
  - Predict high risk fire areas in the city of Montreal

Project B – Open-ended question

- Using the firefighters dataset (and other connected datasets)

- You define the question that you want to solve

Initial datasets projects A & B

- Incidents
  - http://donnees.ville.montreal.qc.ca/dataset/interventions-service-securite-incendie-montreal

- Fire stations
  - http://donnees.ville.montreal.qc.ca/dataset/casernes-pompiers

Potentially relevant datasets

- Property assessment
  - http://donnees.ville.montreal.qc.ca/dataset/unites-evaluation-fonciere
- Crime
  - http://donnees.ville.montreal.qc.ca/dataset/actes-criminels
- Census Program Data
  - https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dv-vd/cpdv-vdpr/index-eng.cfm

Project A and B - Known difficulties

- Descriptions and classifications are in French
- Multiple methods to integrate different datasets (distances, aggregations, algorithms, data sources, libraries, etc.)

Project C – Open-ended question

- You define the business question and the data you want to use

Project C – Open datasets

- 311 Queries
  - http://donnees.ville.montreal.qc.ca/dataset/requete-311

Projects

Area	Project A	Project B	Project C
Scope	Defined question	Open-ended question	Open-ended project
Initial dataset	Firefighters dataset	Firefighters dataset	Open
Additional datasets	Open but review the datasets referenced in this presentation	Open	Open

Grading scheme

- We will use the same grading scheme in all the project types

Area
Business question solved with data
Use of external data
Analysis techniques
Visualization
Analysis sophistication
Visualization sophistication
Use of additional tools
Material coverage
Team performance

Important rule

- No cheating
  - Document if you are reusing existing datasets, challenges and solutions
    - What is the current state of the existing resource(s)?
    - What are your discoveries/improvements?

Teams

- Teams of three to four

Deliverables

- Workflows (Alteryx, Tableau & Python)
- Final presentation delivered during the last class (15 minutes)
- Executive report

Executive report - Format

- Maximum 10 pages of content
- Letter Calibri, 11 pts, Normal margins
- Template with necessary sections will be published on MyCourses
- Supporting images, tables and references can be added in appendix

Recommended timeline

- Week 4 – Project definition and data exploration → Deadline to define the project (November 28)
- Week 5 – Data cleaning & engineering
- Week 6 – Data modelling & solution convergence
- Week 7 – Solution improvement & documentation

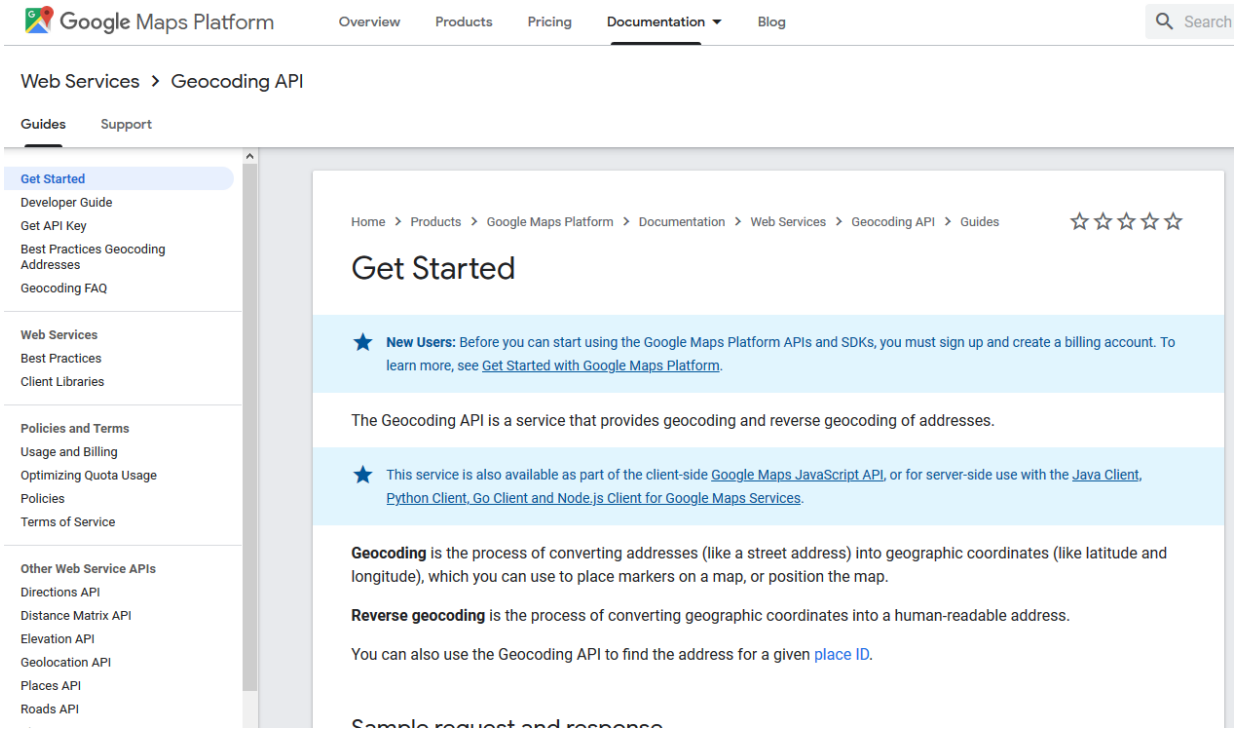
Considerations and recommendations

- Start small
- Build your solution and model iteratively
- You don’t need to use the complete datasets (years, types of calls, etc.)
- Use tools that allow you to simplify the work that needs to be done

Workflow set-up

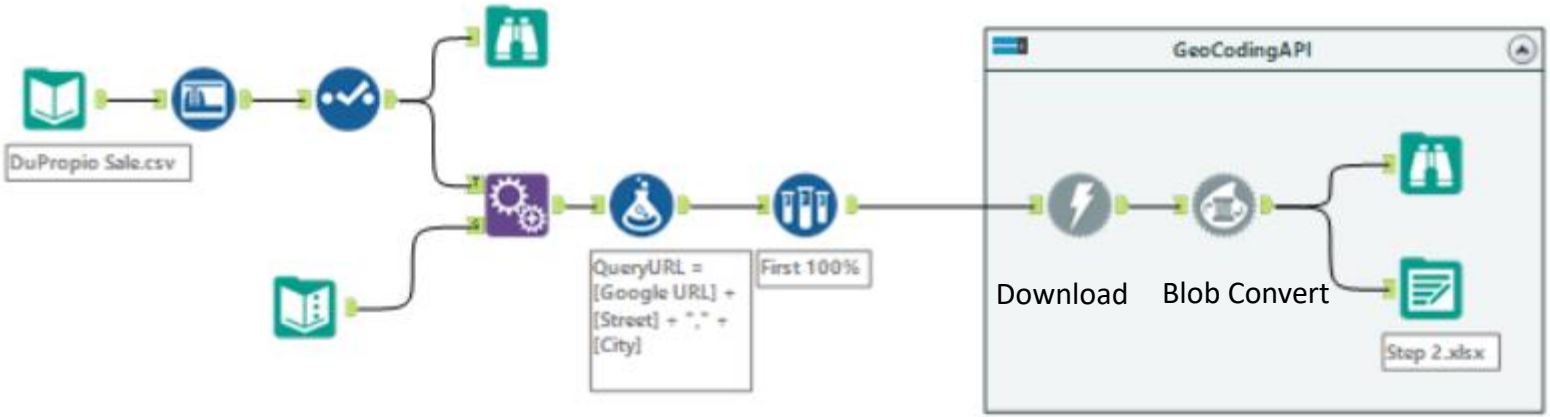
- Try to use GCP as a central repository of your data
- Enable access to the data to all team members (Python & Alteryx)

Additional Tools - GoogleMaps - Geocoding API

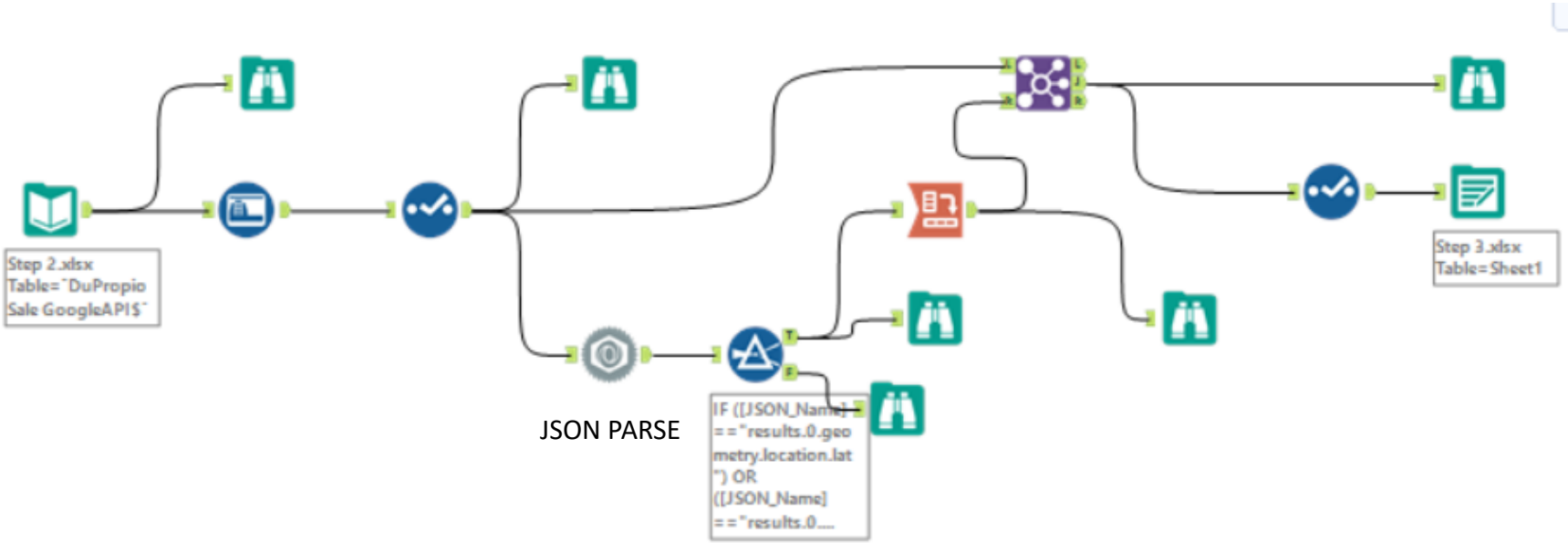


https://developers.google.com/maps/documentation/geocoding/start

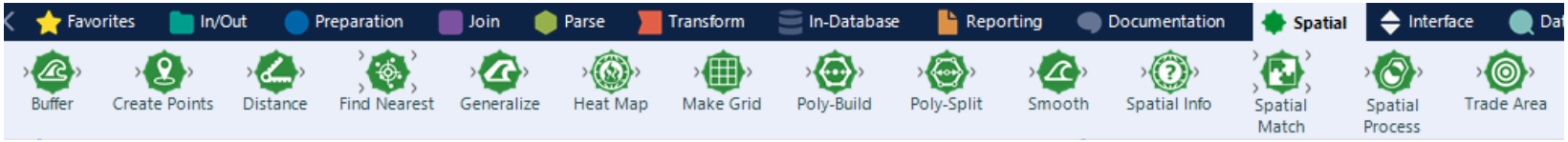
Additional Tools – Alteryx - JSON Requests (1/2)



Additional Tools – Alteryx - JSON Requests (2/2)



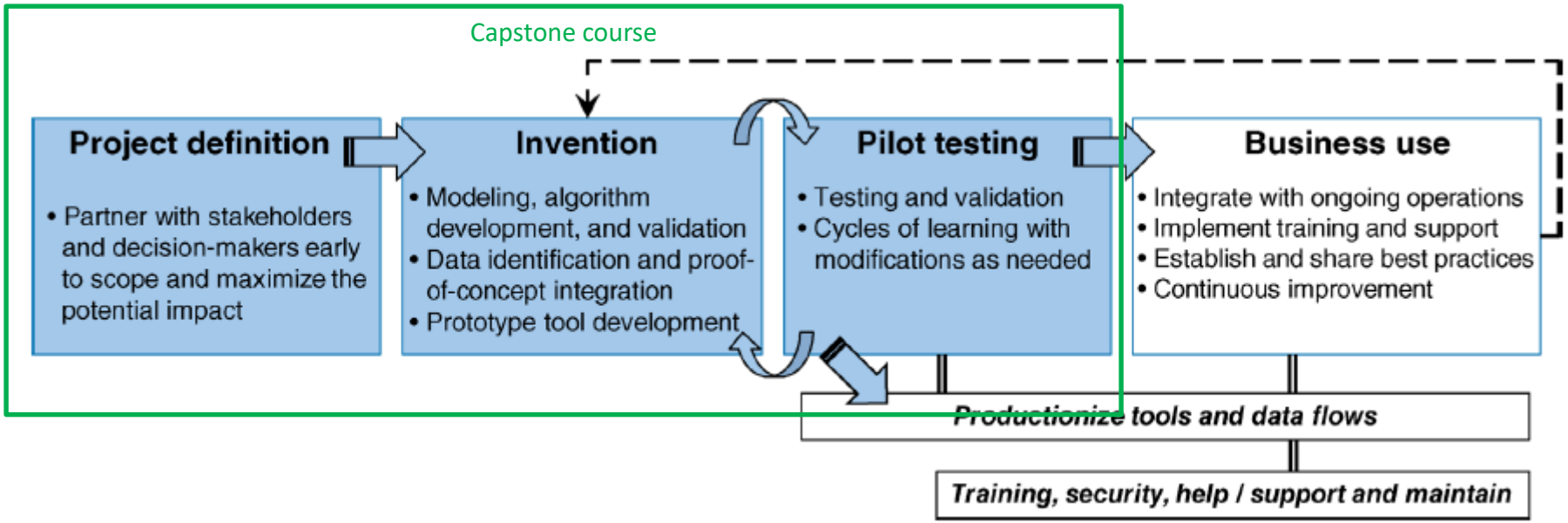
Additional Tools – Alteryx – Spatial Group



Considerations for analytics projects

- Before starting the project you might not know about the data sources or the data itself
- Difficult to predict what difficulties or problems you will find in the dataset or the information systems
- A data analysis project is more similar to a research and development project than a software development project

Life cycle of analytics projects



INFORMS Analytics Body of Knowledge

Problems (or opportunities), the starting point

- Problem is a situation in need of a repair, improvement, or replacement

What is the business problem?

- Too little revenue?
- Processes too slow?
- Too many returns of defective products?
- Too few customers?
- Low ad ROI?

Business case - **Business needs**

- Determination of what is prompting the need for action
- Situational statement documenting the business problem or opportunity to be addressed, including the value to be delivered
- Identification of **stakeholders** involved or affected
- Identification of the scope

Business case - **Analysis of the situation**

- Identification of organizational strategies, goals and objectives
- Identification of root causes(s) of the problem or main contributors of an opportunity
- Gap analysis of capabilities needed for the project versus existing capabilities of the organization

Business case - **Analysis of the situation**

- Identification of known risks
- Identification of critical success factors
- Identification of decision criteria by which the various courses of action may be assessed

Stakeholders

- An individual, group, or organization who may affect, be affected by, or perceive itself to be affected by a decision, activity, or outcome of a project

Stakeholders

- Who will use your analysis?
- Who is affected by the problem you are trying to solve?
- Who will be affected by your analysis?
- What goals do they have regarding the problem you are trying to solve?

Stakeholders

- After you complete the project
  - What is likely to be the most useful solution, if can be found?
  - Who will benefit most? Or suffer most?

What **constraints** & **risks** exist?

- Time
- Cost
- Quality
- Political
- Social
- Technical

Organizational

Think not only in terms of the project development, but after your solution is implemented

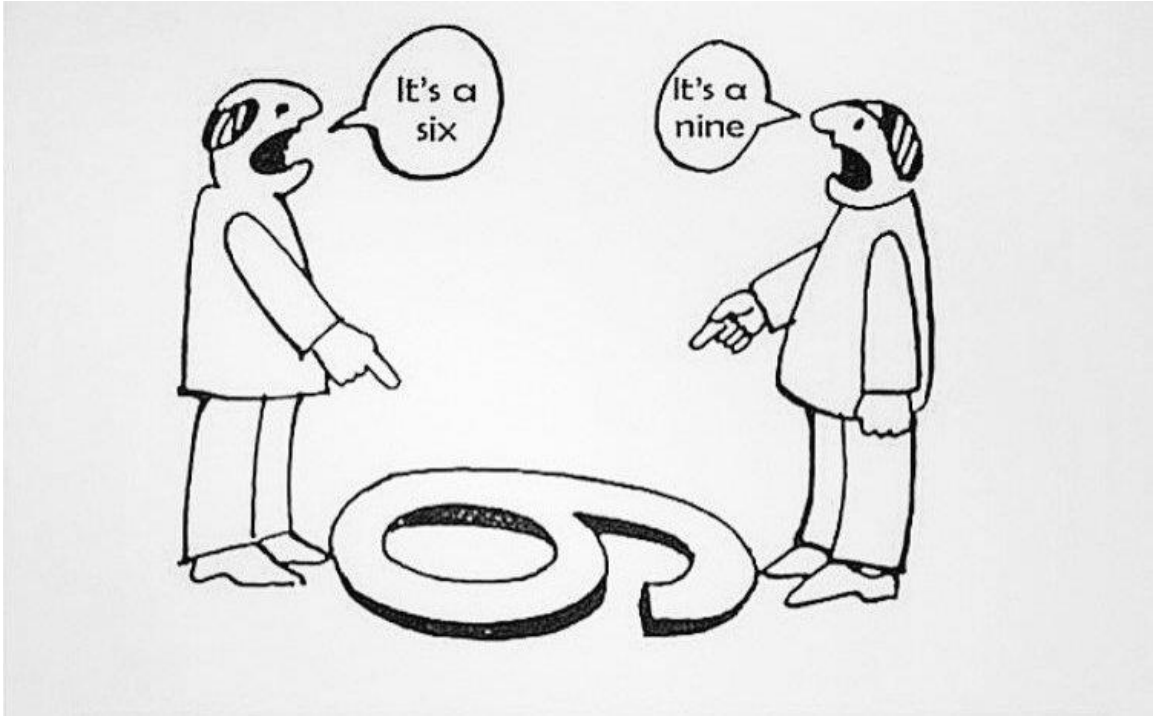
Framing a problem

- What perspectives could we use when we look at the data?

311 dataset

- Police department
- Government
- Transportation department
- Renter & landlord
- Seller & buyer

It is all about the perspective we take



Perspectives

- Can you try to look a the problem from different perspectives?
  - You can do better by considering different perspectives.
- Are data analytics projects:
  - Expense
  - Investment
  - Neutral (cost of doing business)

} Predictive model to estimate machine failure

5Ws Tool

What	What is the problem or opportunity?
Who	Who is involved, impacted, or influential?
When	When are key milestones, deadlines, decisions?
Where	Where will the work take place?
Why	Why do we expect different outcomes than before?

Thinking exercise

- You are developing a model to decide which employees should get a raise
  - Who are potential stakeholders?
  - What are potential risks?
  - What are potential constraints?
  - Who will benefit the most? Who will benefit the least?