Managing Data Science Teams

Skoltech - 06 /04/2021
Rodrigo Rivera
rodrigo.riveracastro@skoltech.ru

What to think about when leading a data science effort?

Move from: “We have some data” to “Which questions to ask?”

Who am I?

Rodrigo Rivera

Former Technical Director at Samsung, Head of Data Science at PMI, VP R&D at Lamoda and more

Entrepreneur with 1 exit

SAMSUNG NEXT

PHILIP MORRIS INTERNATIONAL

lamoda

What will happen after graduation?

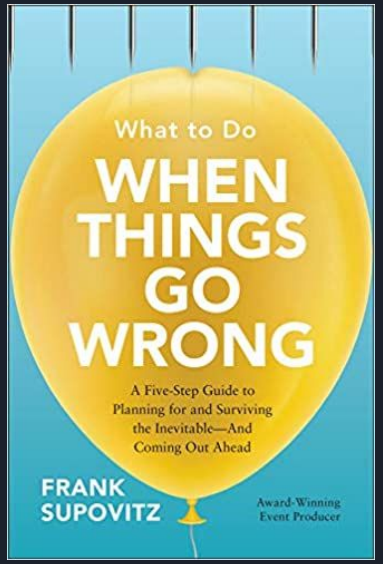
Startup VS Enterprise

And this also applies even if you stay in academia!

Once you are in the wild, you want to minimize risk

Which questions should you be asking? How should you go about it?

What are some issues that can happen during your projects?



Management is all about managing expectations

We should go from a broad perspective to the details

Managing a data science project means going from the big questions to the exact details

What if I do not have the right profiles?

Missing skills, wrong skills, lacking seniority, finding the right teams is hard!

Do we have the right communication in the team?

Working with data requires coordinating both internally and externally with other stakeholders

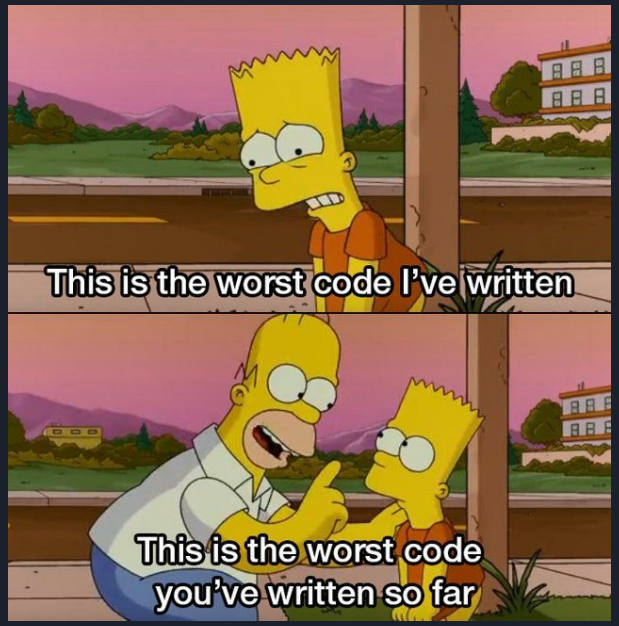
Do my deliverables have any value?

We must make sure that some can use our work regardless if it is software or publications

Is my data any good?

We are highly-dependent on the quality of data, but oftentimes we do not have any control over it

What if my team cannot reproduce the project?



Maintainability and good code is a major challenge in the community

Does my work has any bias?

Our work has societal consequences and millions can be direct or indirect users

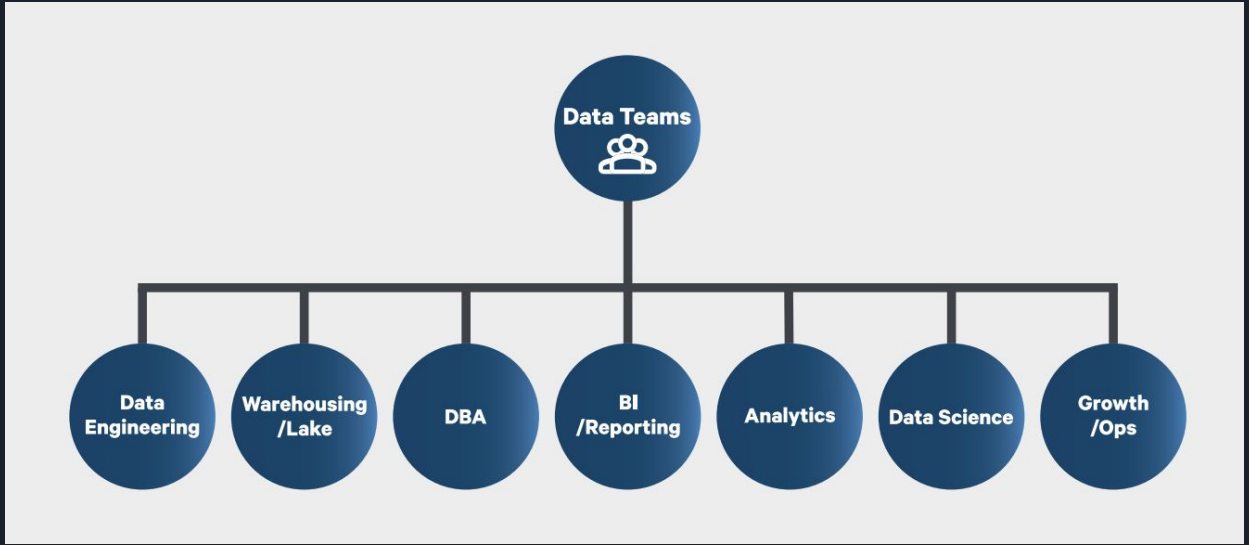
Can we go into production?

We must be ready to move from a Jupyter notebook to an actual production system

So, which questions should we ask?

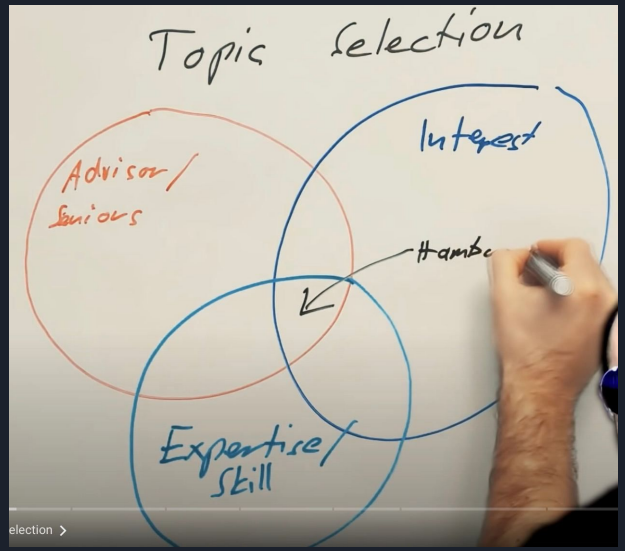
From the beginning, we must establish a basic understanding of our circumstances

Who needs to be part of the team and what skills do I need?



What can we do and where do we need to grab external support

Understanding skill-gaps also applies for our doctoral research!



Choosing the right thesis topic is also a result of understanding your skills and those around you

How are we already communicating within our group?

We must develop an understanding of the mechanisms in the group to share information

Which ethical situations might we encounter?

Apple's 'sexist' credit card investigated by US regulator

11 November 2019

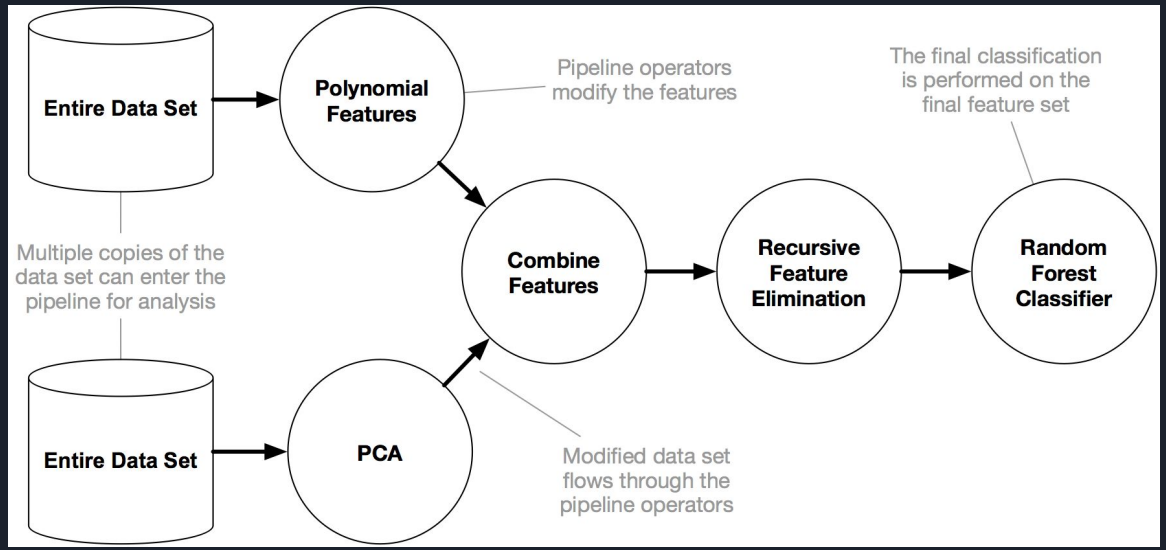
A US financial regulator has opened an investigation into claims Apple's credit card offered different credit limits for men and women.

We must understand what can be the consequences of using our models

How are we ensuring that our insights are correct?

We must think about quality controls and ways to ensure that our insights are actionable

Do we have a workflow in place for our work?

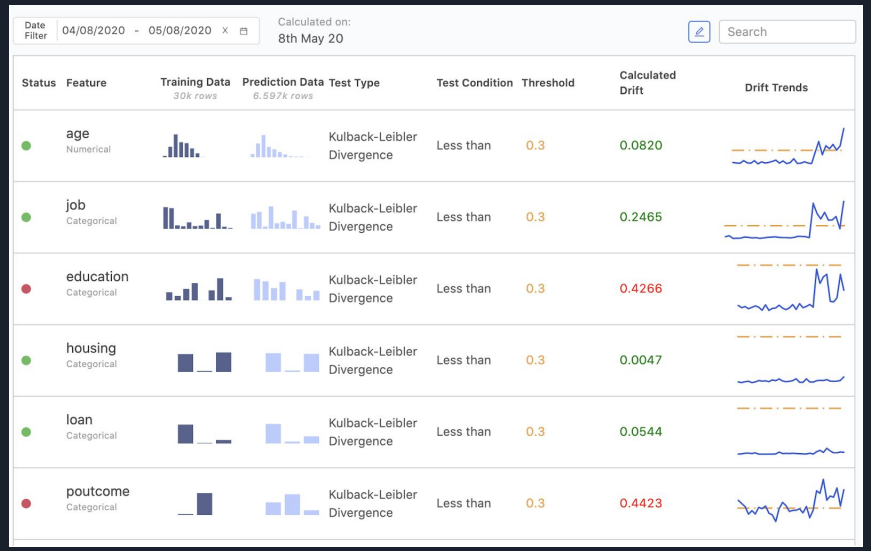


We must have tools and processes in place to keep track on how our work is used from the beginning to the end

Do we have an audit trail of the insights?

We must be able to recreate our models from previous situations

How robust is our work?



Status	Feature	Training Data *30k rows*	Prediction Data *6.597k rows*	Test Type	Test Condition	Threshold	Calculated Drift	Drift Trends
●	age Numerical			Kulback-Leibler Divergence	Less than	0.3	0.0820	
●	job Categorical			Kulback-Leibler Divergence	Less than	0.3	0.2465	
●	education Categorical			Kulback-Leibler Divergence	Less than	0.3	0.4266	
●	housing Categorical			Kulback-Leibler Divergence	Less than	0.3	0.0047	
●	loan Categorical			Kulback-Leibler Divergence	Less than	0.3	0.0544	
●	poutcome Categorical			Kulback-Leibler Divergence	Less than	0.3	0.4423	

We must have process for migration and monitoring

Overview of the questions to consider

1. Roles - What is the composition of the team?
2. Process - How the team collaborates & communicates?
3. Ethics - How to avoid issues in the use of the model?
4. Quality - Do we have accurate insights?
5. Analytics Workflow - Is the work modular and efficient?
6. Analytics and Model Management - Is there an audit trail?
7. Production Robustness - How easy to migrate & monitor?

Participants can vote at slido.com with code #116860

And let's survey the priorities!

Getting some answers!

What are some answers around our questions?

Defining the skill sets for our data science project

- Which type of project will it be? One person or 20?
- Do we need multiple roles for each person? Or will one person wear multiple hats?

Defining the skill sets for our data science project

- Data scientist: Uses visualizations and machine learning to aid in the understanding data. Has an overview of the end-to-end process
- Data Architect: Designs and maintains the architecture of data science applications. Creates relevant data models and process workflows.
- Data Engineers: Makes the data available for data science efforts. Designs, develops, and codes data applications for data capture and analysis
- Data or Business Analyst: Analyzes a large variety of data, often using visual tools, to extract information about system, service or organization performance
- IT and Infrastructure Engineer: Builds and manages IT systems
- - Machine Learning Engineer: A data scientist, who can bring code into production and deploy
- Client: Provides guidance on what can be made operational, etc

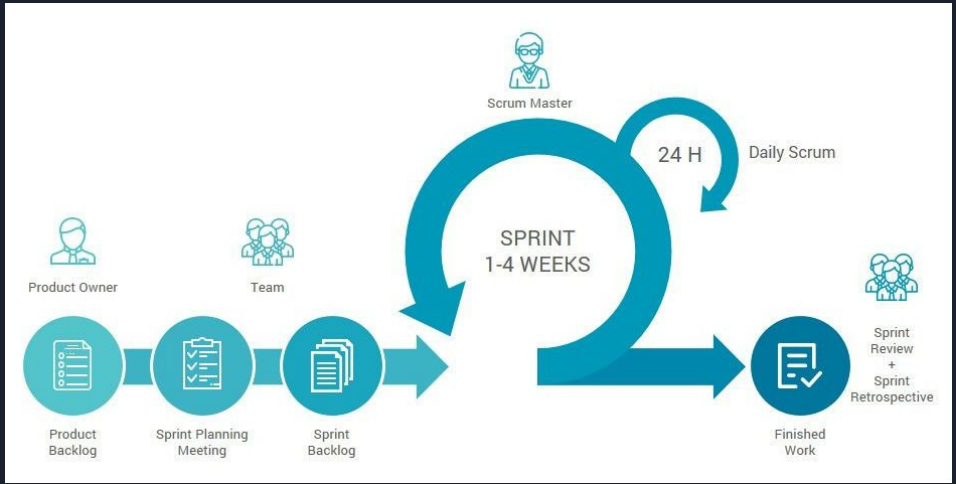
There are many other roles: AI Product Owner, Research Scientist, Tech Lead, etc

Team processes

How should we approach a task, a question or a problem?

Scrum

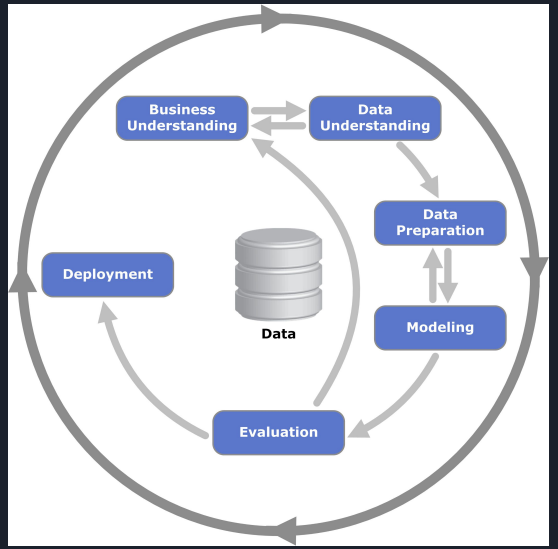
- It has clear goals and fixed deadlines
- Sometimes not suited for data



It is very popular in the software community

CRISP-DM

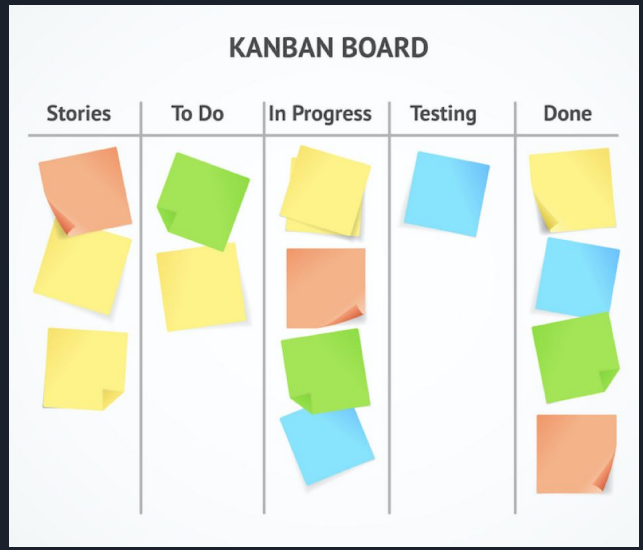
- Good concept
- Unclear how to put in practice



It is great on paper but maybe less so in practice

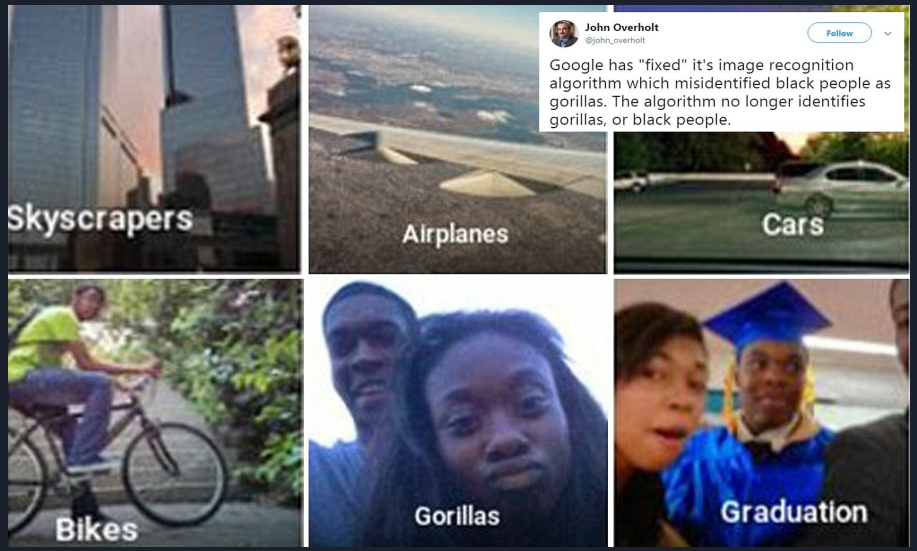
KANBAN

- Clear overview of responsibilities
- Flexible



We can understand it as a relaxation of SCRUM

Ethics: Avoid potential issues in the creation and use of the model

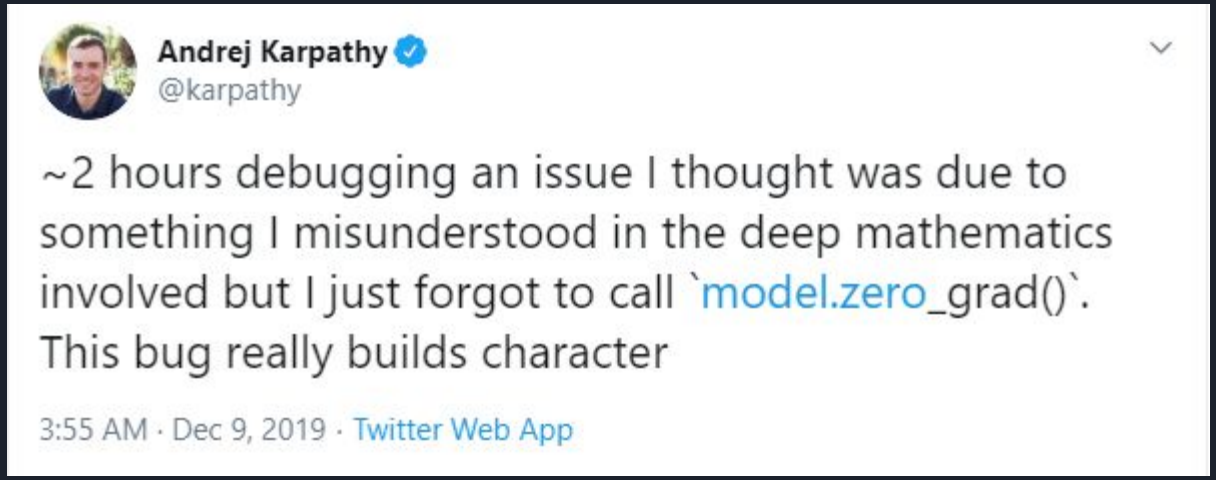


Ensure that there is no bias in the data, otherwise, it can have devastating consequences

Which questions should be asking?

Challenge	Ethical Question
- Project initiation and management related challenges	- What harm can the project do? - Who may be harmed by the project? - Do we know the consequences of relevant laws and regs? - Is ethical accountability in place? - Is ethical oversight sufficient?
- Data-related challenges	- - Are the data sources ethically sound? - Are individuals' privacy and anonymity impinged? - Can data quality issues cause unethical outcomes?
- Model- related challenges	- - Is the model fair? - Is the model transparent? - Are project objectives structured for objectivity? - Can the results be misinterpreted?

Quality: Ensuring quality is almost elusive but necessary



We deal with two challenges! Code (bugs) and data (which is also generated by code)

How can we ensure data quality?

- Did cleaning introduce errors?
- Did an enhancement break something?
- Are there bugs in the code - how did we test?
- What's the model behavior with real and test data (or AB testing)?
- What is the process to approve this model?
- What is the process to monitor model performance?
- Did we benchmark this model versus simpler ones?

Cleaning the data can destroy it or introduce bugs. And discovering bugs in data science is not easy!

What else is there for us to consider?

- Proxy bias: If the data is an indirect measurement, it may differ from the "real" behavior

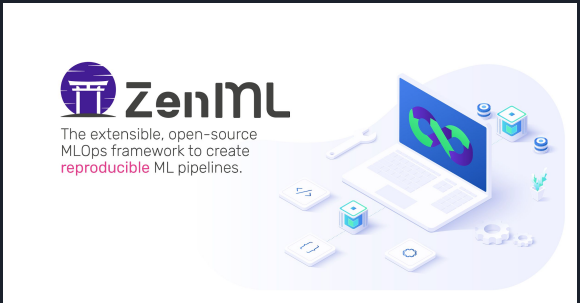
- Instrumentation bias: The "visible universe" may differ from the "whole universe"

- Confrontation bias: An analysis can generate "the outcome that you look for"

Often it is easier when you are looking for an specific answer, confirm a hypothesis

Analytics Workflow: The tools we should consider

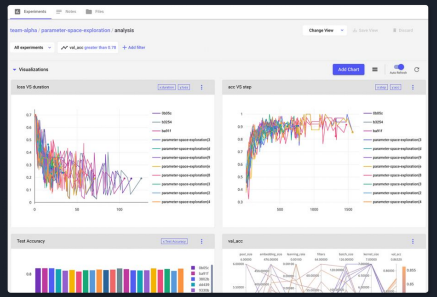
- More easily create actionable insight (do the work)
- Document the work (code + data) for future refinements
- Facilitate team coordination across the effort (ex. modular tasks)



There are many workflows in the market from GUIs to frameworks such as Kedro and ZenML, use one!

Model Management: Audit and Development

- Model versioning (track changes - code and data)
- Store models (even one-off might need to be redone or explained)
- Recreate models and help understand any issues that arise



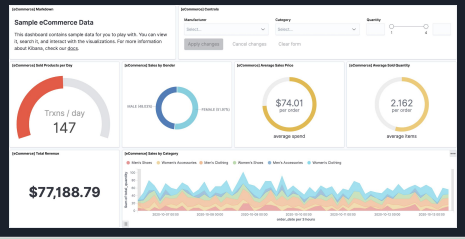
comet

neptune.ai

You need a strategy to manage your models and experiments, from stored pickle files to Neptune and Comet

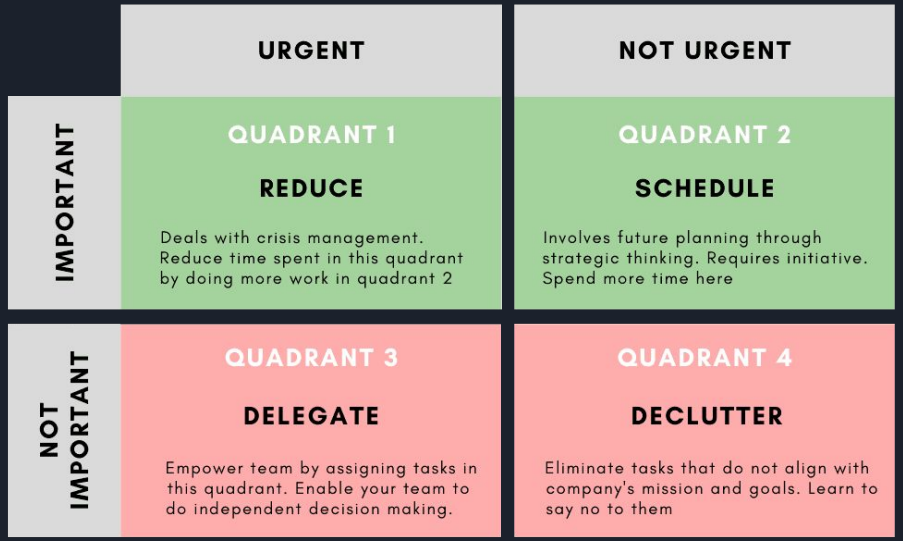
Production Robustness: Monitoring

- Migrate from R&D to production (do you need to recode R? etc)
- Monitor model performance (just like "tech ops")
  - Create alerts (ex. when models degrades)
  - Create effective dashboards (ex. show model performance)



You must think from the beginning, which data you want to track such as model shift, KPIs, etc to improve

So what are the right priorities



Set the priorities based on the project by start asking questions, an Eisenhower matrix can help you decide

Discussion

Start asking the right questions!