Developing Data Products

SC4125

Module 1

Introduction

About me: Anwitaman DATTA

https://personal.ntu.edu.sg/**anwitaman**/

Associate Professor

School of Computer Science and Engineering

NTU Singapore

Contact

anwitaman@ntu.edu.sg

N4-02A-18

Teaching material

Accompanying teaching material

These slides are accompanied with a jupyter notebook and an html deck of slides.

Data Products

"data product is a product that facilitates an end goal through the use of data" - DJ Patil (former Chief Data Scientist of the United States Office of Science and Technology Policy)

Ref: Data Jujitsu--The art of turning data into product, DJ Patil, 2012

Data Products

"products that use data to facilitate an end goal vs products whose primary objective is to use data to facilitate an end goal" - Simon O'Regan

Ref: Designing Data Products - Simon O'Regan, Towards Data Science blog, 2018

Data Products

"systems that learn from data, are self-adapting, and are broadly applicable … where an application acquires its value from the data itself, and creates more data as a result" – Combining definitions from Benjamin Bengfort, Jenny Kim and Mike Loukides.

Ref: Data Analytics with Hadoop by Benjamin Bengfort, Jenny Kim, 2016
What is data science? by Mike Loukides, 2010

Data Products: Broad Categories

- raw & derived data, algorithms & models, decision support & recommendation systems, automation

Data updated as at 1 July 2021, 1200 hrs

Case Number	PR date	Occupation	Link Reclassification	Rationale for Link Reclassification
64236	15 Jun 21	Homemaker	Community Unlinked → Community Linked (Cluster 64112)	Case is phylogenetically linked to the index case of Cluster 64112
64254	16 Jun 21	Retiree	Community Unlinked → Community Linked (Cluster 64330)	Case is linked to Cluster 64330. She resides at the same location as the index case 64330
64264	16 Jun 21	Foreign Domestic Worker	Community Unlinked → Community Linked (121 Bukit Merah View)	Case is the index of Cluster 64264. She is linked to the cases staying at 121 Bukit Merah View
[illegible]378	22 Jun 21	Unemployed	Community Unlinked → Community Linked (Cluster 64310)	Case resides in 105 Henderson Crescent
		Retiree	Community Unlinked → Community Linked (Cluster 64310)	Case resides in 105 Henderso[illegible] Crescent
				Case is the househ[illegible] of case 6444[illegible]

Figure 1.8: Weekly Reclassifications of Previously Reported Cases (next update will be on 5 Jul)

	Imported	Community Linked	Community Unlinked	Dorm Residents	Total
Total number of cases as at 21 Jun	4,754	2,233	917	54,526	**62,430**
Cases reported 22 Jun - 28 Jun	37	68	18	0	**123**
Net reclassifications of previously reported cases, incorporated 22 Jun - 28 Jun	0	15	-15	0	**0**
Total number of cases as at 28 Jun	4,791	2,316	920	54,526	**62,553**

Screenshots from Singapore MOH's Covid-19 daily report

Data Products: Broad Categories

raw & derived data, algorithms & models, decision support & recommendation systems, automation

Internal/Proprietary:

3rd Party:
- Crawling/scraping, downloadable, APIs, …
- Open/Proprietary with licencing or subscription, …

Dimension	Characteristics			
Data Source Availability	Internal	External – Closed		External – Open
Data Source Interface	Internal Interconnection	Traditional EDI	Web Services	Offline Data Dump
Data Source Pricing Model	Volume-Driven	Time-Driven	Unique	No
Data Aggregation	Resource	Database	Record	Item
Data Occurrence/Update	Stream	Event-Driven Batch		Time-Driven Batch
Data Ownership	One Legal Entity	Community		Public
Data Structure	Structured	Semi-Structured		Unstructured
Data Format	Proprietary		Open	
Intra Data Standardization	Value	Semantic	Syntax	No
Inter Data Standardization	Value	Semantic	Syntax	No
Data Currency	Forecast	Up-To-Date		Outdated
Data Completeness	High	Medium		Low
Data Accuracy	High	Medium		Low
Data Sharing	Proprietary	Free		Open

Example taxonomy of data sources
Ref: Data Source Taxonomy for Supply Network Structure Visibility, Zrenner et al, 2017.

Data Products: Broad Categories

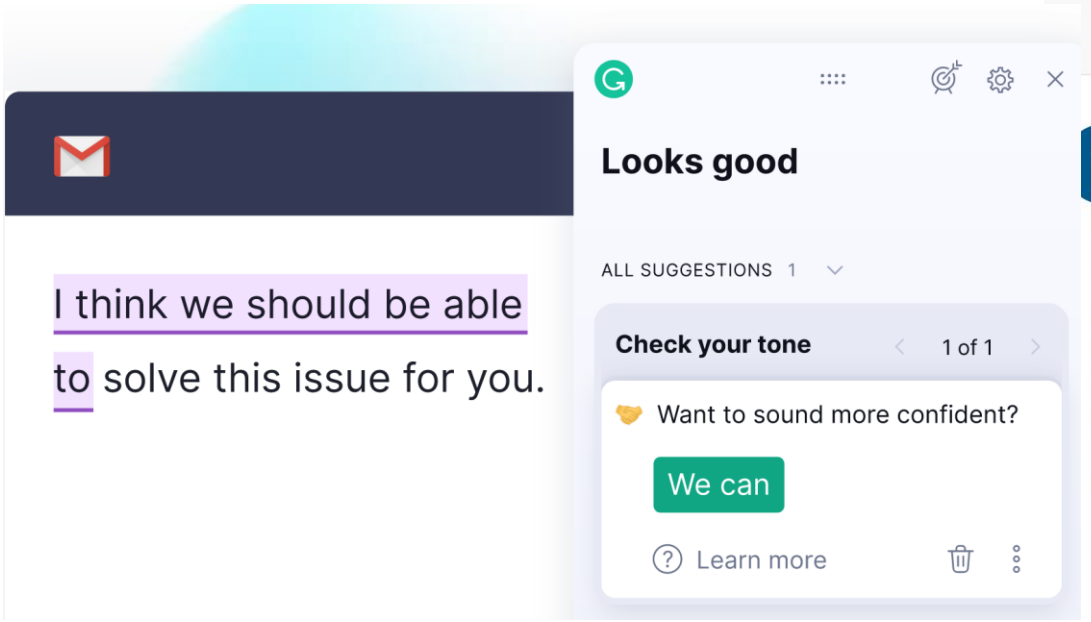
- raw & derived data, algorithms & models, decision support & recommendation systems, automation

Fig. 12 Visualising the fit of the submissions' LDA model to the submission corpus. Each of the 9 circles represent one topic, whose area is proportional to the proportions of the topics across the number of tokens in the corpus. The circles are labelled in descending order of their areas. The centres of the topic circles are laid out in two dimensions according to a multidimensional scaling algorithm that

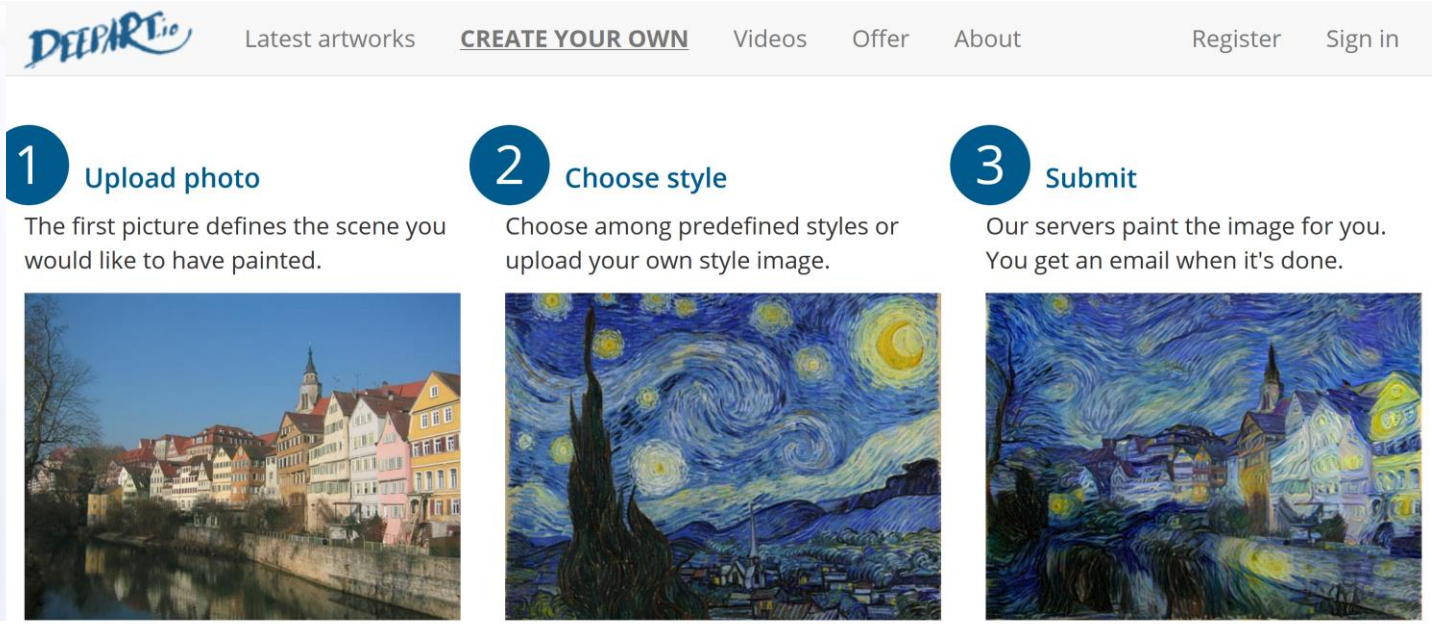
Office 365 screenshots from (i) a working draft of a manuscript studying the r/coronavirus subreddit, (ii) these lecture slides.

Data Products: Broad Categories

- raw & derived data, algorithms & models, decision support & recommendation systems, automation



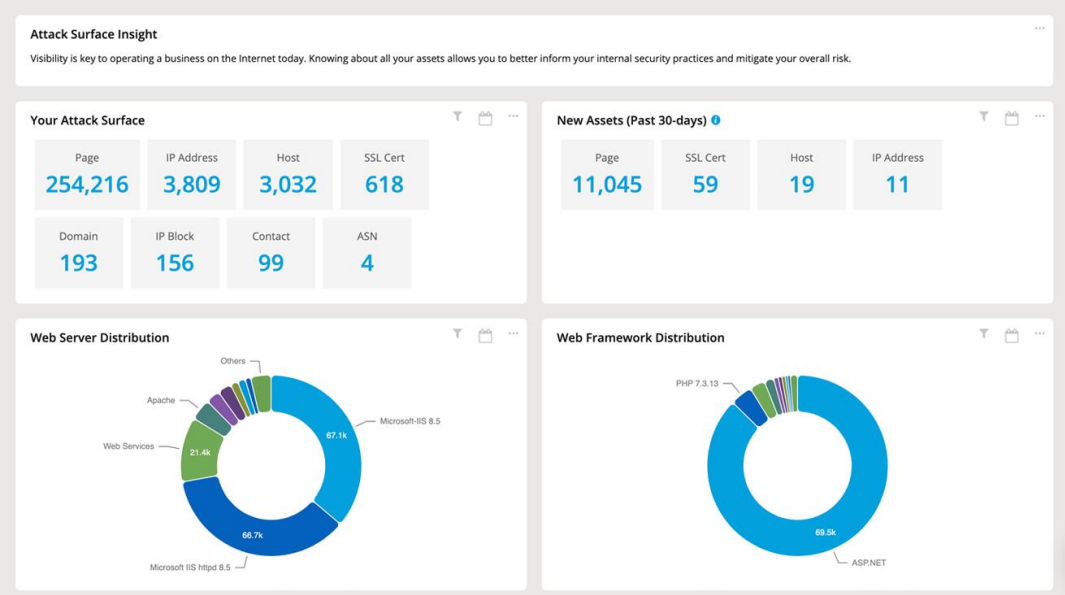
Screenshot from https://www.grammarly.com/



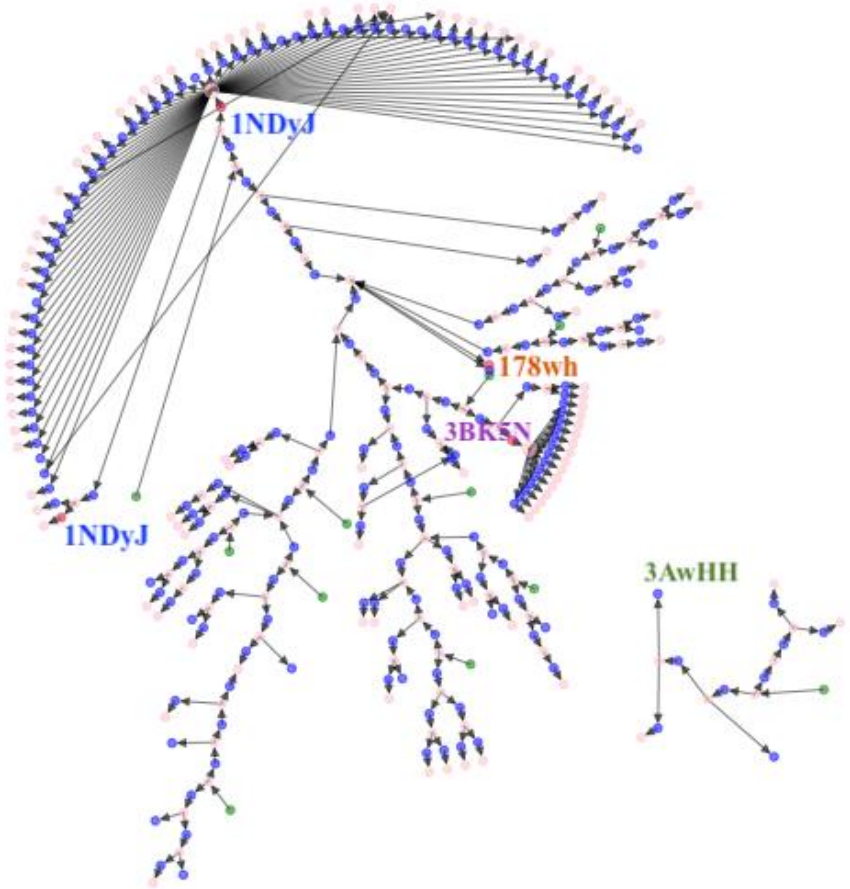
Screenshot from https://deepart.io/

Data Products: Broad Categories

raw & derived data, algorithms & models,
decision support & recommendation systems,
automation



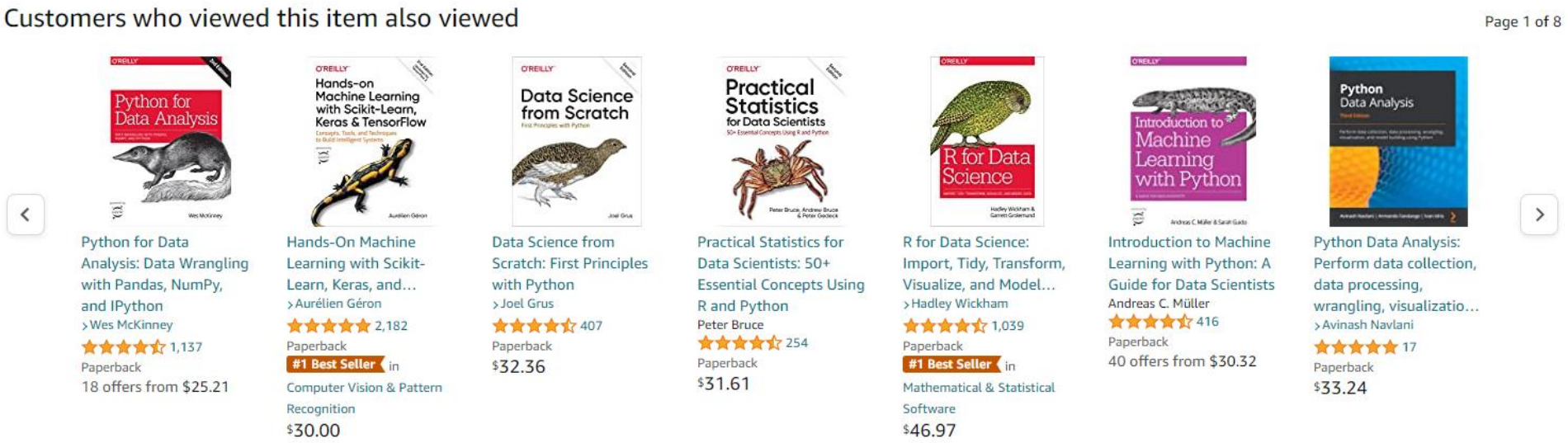
Summary of an organization's digital footprint based on OSINT
Screenshot from RiskIQ website



Tracking tainted Bitcoin flows and identifying wallet associations.
Ref: An ego network analysis of sextortionists, Oggier et al., Social Networks Analysis and Mining, 10(1), 2020

Data Products: Broad Categories

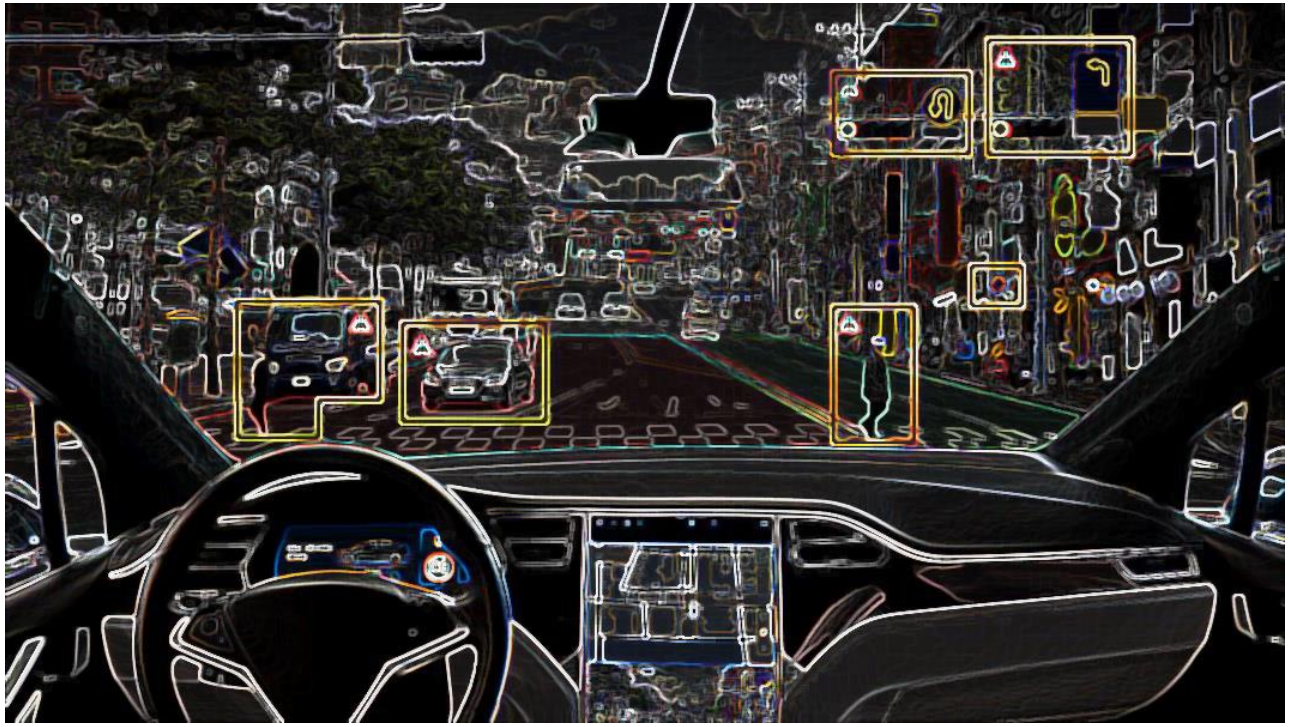
raw & derived data, algorithms & models,
decision support & recommendation systems,
automation



Screenshot from https://amazon.com

Data Products: Broad Categories

- raw & derived data, algorithms & models, decision support & recommendation systems, automation



Derived (using a “data product” similar to one mentioned earlier) from an original image from https://singularityhub.com

Wheels within wheels

Granularity

An end-system may comprise multiple sub-components, which are data products in their own right, possibly with inter-dependencies, e.g., in autonomous vehicles, there is object identification, navigation strategy, …

Under the hood

The end-user may not always explicitly see the products: It's a car! It's a data product!

Feedback reinforced

The use of the data products may generate more data/learning opportunities, leading to (potentially, many orders of magnitude of) refinements.

Iterative refinement

Start with simpler product(s) and refine, e.g., SAE levels for autonomous vehicles

Navigating faux pas, pas-à-pas!

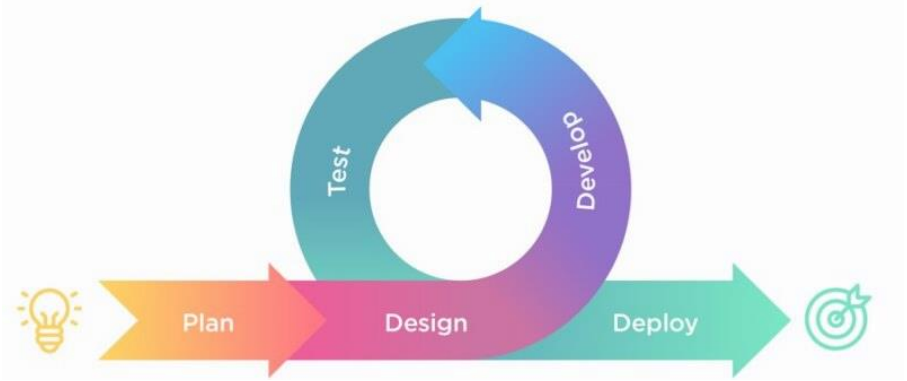
- Before committing into a full-fledged product

Is the product needed?

Is the product viable?

How long does it take to get there?

Any short cuts to get off the ground (e.g., to obtain inputs for the first two questions)?



Let's discuss:
How will you go about in developing a "find an expert" recommendation system?
Say, from within NTU's faculty.

Work around obstacles, reduce user friction

- Product design, Human-in-loop

e.g., LinkedIn's address-book import (high friction!)

vs

"people you might know"

"an interesting job for you"

vs

"recommend a job for Mr E."

engage the user: "who viewed your profile?"

Screenshot from https://quickdraw.withgoogle.com

Dealing with subjective/objective dilemmas

What is more important? Precision or Recall?

- Web-search: user-experience with a few bad results on top page?
- Trying to determine whether to apply more expensive tests for a disease detection?

What if there may not be an objective answer?

- Which jobs to recommend? How to assess (credit, recidivism) risks?

Unforeseen and unintended effects

- Bias, privacy implications, ...

Tech policy / AI Ethics

AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

Ref: https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

Feb 16, 2012, 11:02am EST

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff
Tech
Welcome to The Not-So Private Parts where technology & privacy collide

Follow

Ref: (certain caveats apply to this "story") https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

Unforeseen and unintended effects

- Impact of the system

  Effect of feedback cycle on user behaviour and future functioning/quality of service of the system itself

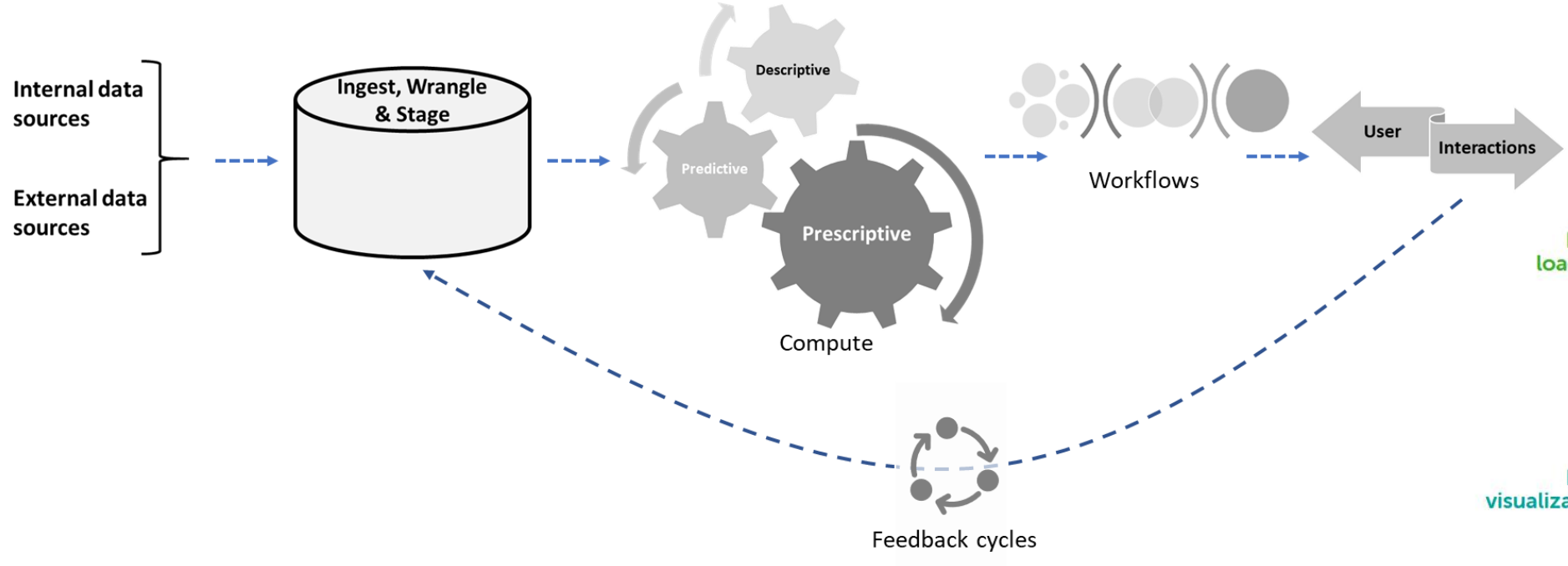
This course

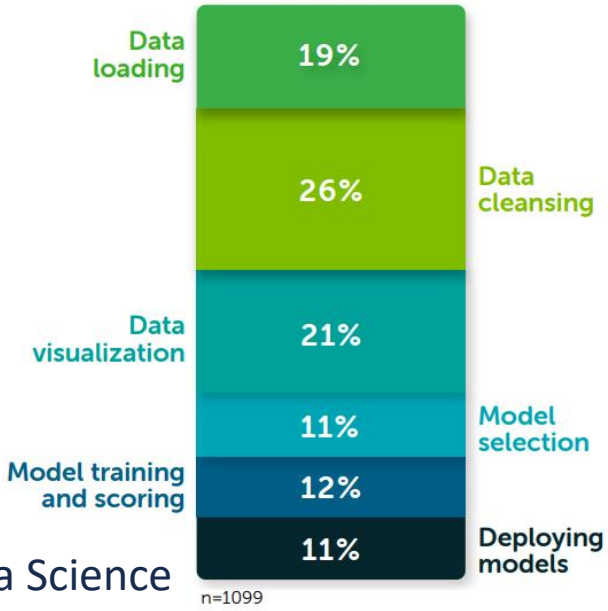
Organization, Assessments, …

What is this course (not) about?

A little bit of **principles**

Pieces of the **data product pipeline** - high level concepts, ideas and “toy” examples





Ref (for stats):
Anaconda 2020 State of Data Science

What is this course (not) about?

The proverbial **data pipeline**

- itself is fluid!
  - new data sources
  - old data sources becoming unavailable
  - the product itself changing user behaviour, triggering other changes
    - new models, new products, …
  - changes in implementation
    - need to implement new models/functionalities/user interface/products
    - (3rd party) libraries and software components becoming obsolete
    - other extrinsic reasons, e.g., business model, regulatory requirements

What is this course (not) about?

A lot of **practice**

- However, it is **NOT** about specific programming languages, algorithms, tools or technology ecosystem.
- Yet, to carry out hands-on activities, we would use specific artefacts as means to elaborate and instantiate.
  To that end, as programming language, we will principally use **Python**.
- Freedom (and responsibility) to determine and chose tools to use.
- Still, we will mostly miss a lot of the engineering issues (e.g. managing big data and the fluid pipeline!)

Quick and dirty way to grasp the big picture with predominantly open-ended hands-on exercises.

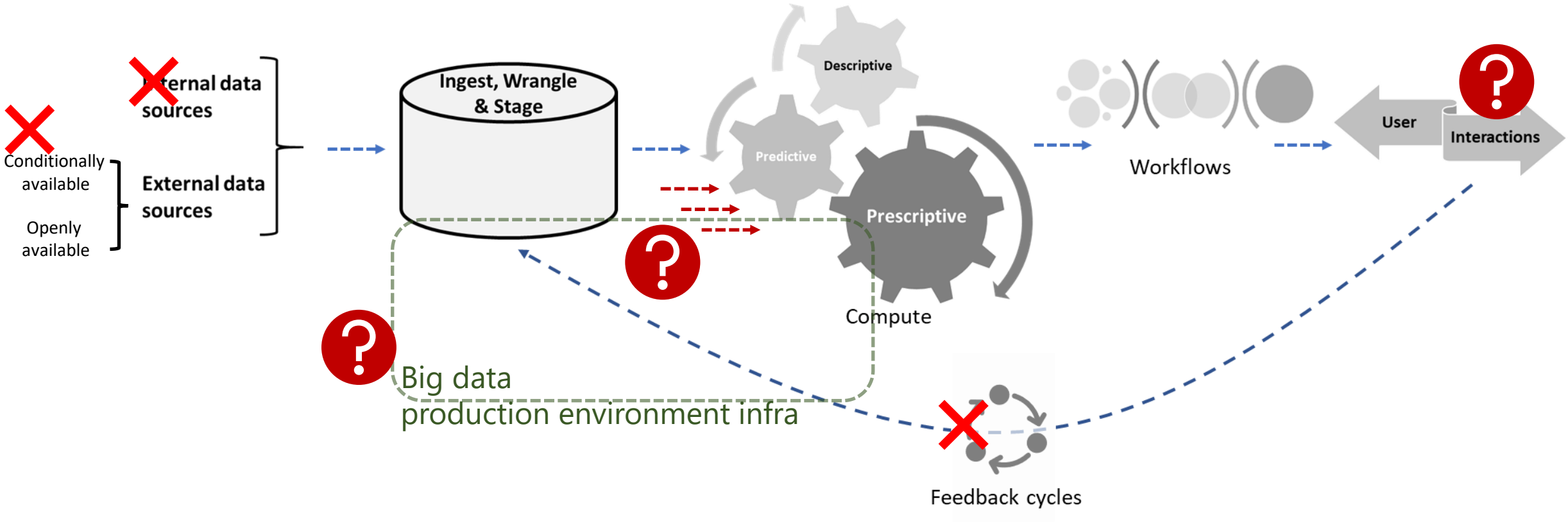
We will touch upon and leverage on most of the following. But for detailed treatment on individual aspects, refer to the dedicated SCSE courses on individual topics

- Visualization
- Machine learning
- Natural language processing
- Big data infrastructure
- Network science
- Neural networks

What this course should, but doesn't (quite) do!

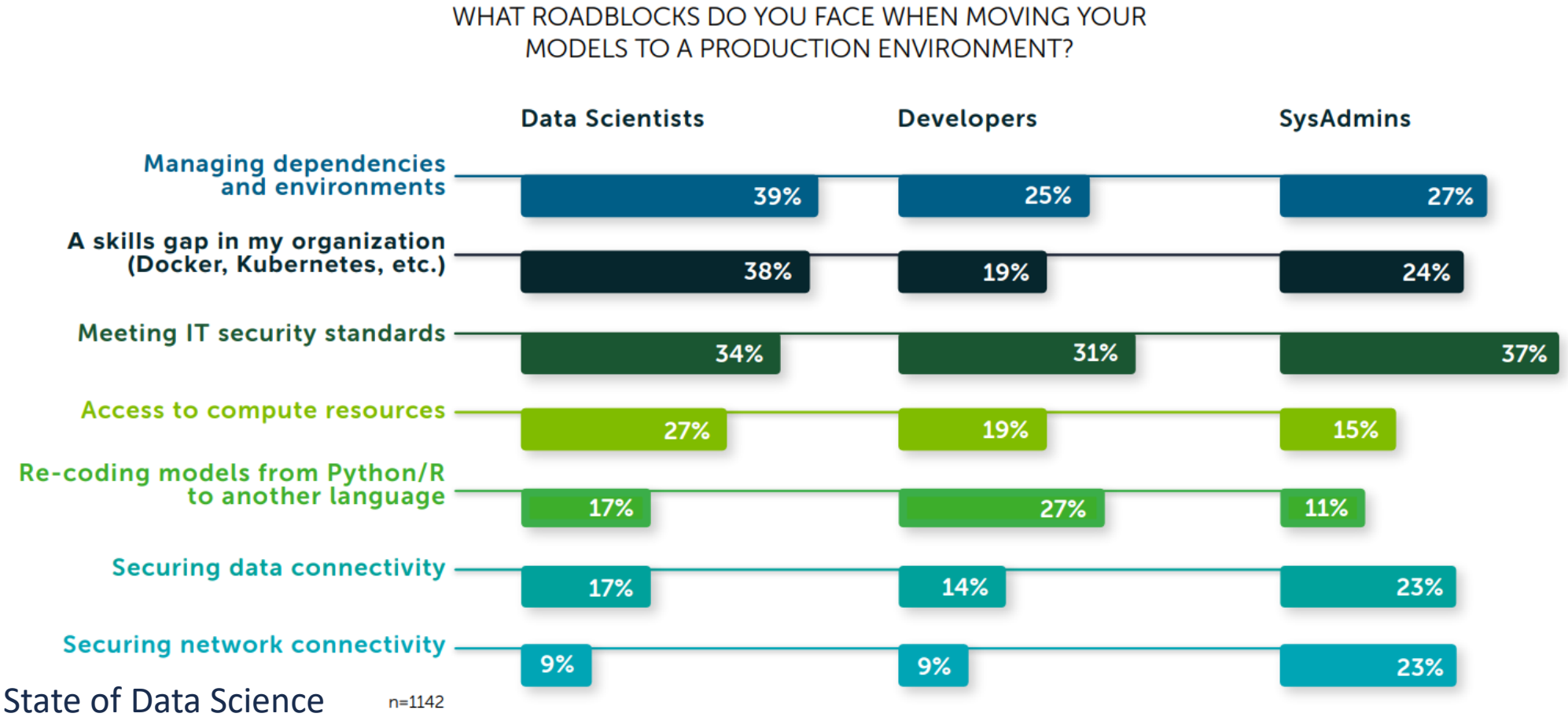
Limitations of an academic context

- Suggestions on how we can **simulate** some of these things?



Further limitations

Limitations of an academic context + scope of the current course: Overlooks production environment



Ref (for stats):
Anaconda 2020 State of Data Science

Assessment and activities

Individual

- Short quizzes [30%]
- Semi-structured assignments [30%] (with some elements of collaboration in group)
- Ungraded assignments

Group

- Open-ended capstone project [40%]
- Groups of 3 or 4 (form your own groups)

Rubric/details

- Please refer to the detailed course syllabus document

Assessment and activities

DOs

- Search 3rd party resources/discuss with other students for understanding how individual components or subtasks may be achieved. You may reuse such subcomponents (e.g., if you find a library to carry out specific munging tasks) and libraries. Provide references when you do so.
- For ungraded assignments, feel free to share and discuss your completed solution with peers or online.
- For group project, feel free to put it online in your portfolio of projects, etc. (after the semester is over)

DONTs

- Do not copy anyone else's solution, particularly for individual graded assignments (group projects ought to be distinct be default!).
- For graded individual assignments, do not share your solution with anyone (even after the course is over).

Ungraded tasks

Intermittent small tasks

Basics of ETL & Viz

Weeks 1-5:
- Data manipulation
- Extract, Transform, Load
- Visualization
- Ethics & accessibility

Standard statistical & ML tools

Weeks 5-8:
- A/B test, statistical tests
- ML and NLP tools

Selected adv. topics

Weeks 9-12:
- Big data/NoSQL
- Graph data
- Dashboards
- Data governance

recess week

Graded (Individual)

Week 5
1st task (7%)

Week 8
1st quiz (15%)

Week 9
2nd task (10%)

Week 12
3rd task (13%)

Week 13
2nd quiz (15%)

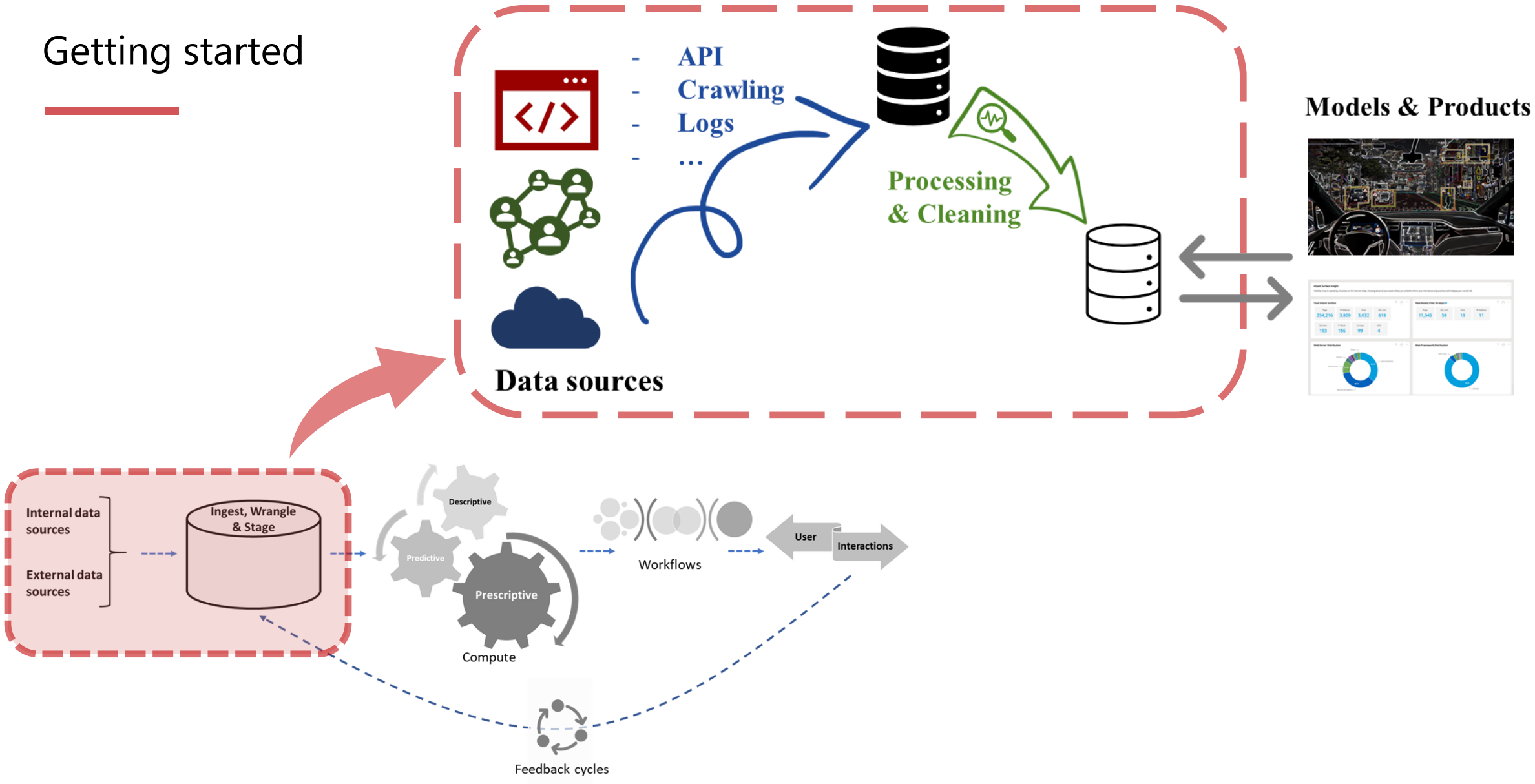
Builds on and gets increasingly complex

Graded (Team)

Week 14
Capstone (40%)

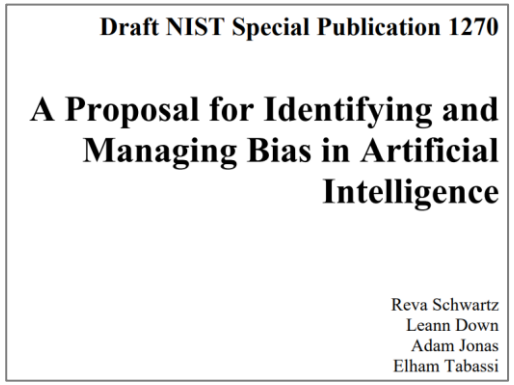
* Indicated schedule is tentative. Deadlines are collectively “negotiable” based on other course-loads, if you discuss in advance.

Getting started



Recommended readings

http://radar.oreilly.com/2012/07/data-jujitsu.html



https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270-draft.pdf

Selected References

Designing Data Products, 2018, Simon O'Regan (accessed on 18 July 2021)
https://towardsdatascience.com/designing-data-products-b6b93edf3d23

Data Analytics with Hadoop, 2016, O'Reilly Media, Inc., B. Bengfort and J. Kim,

Anaconda 2020 State of Data Science (accessed on 18 July 2021)
https://know.anaconda.com/rs/387-XNW-688/images/Anaconda-SODS-Report-2020-Final.pdf

Ungraded tasks

SQLite tasks (see accompanying Jupyter notebook)

Identify **sector specific data products** (existing, as well as potential) and map them to the categories identified in this lecture. Try it for 3-5 different sectors (of your choice) per Singapore Standard Industrial Classification 2020