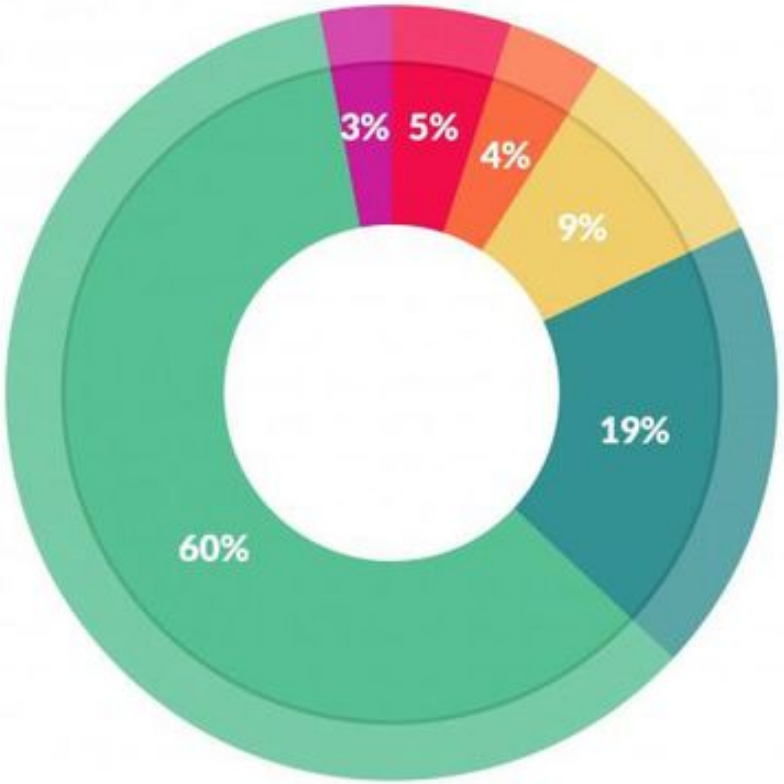
Nikola Vasiljevic and Neil Davis

FAIRification roadmap

FAIRification roadmap aspects (Rene's list)

- Strategic rational
- Short-term targets
- Mid-term targets
- Long-term targets
- Technical aspects of milestones
- Political aspects of milestones
- Economical aspects of milestones
- Community specialization
- Division of labour (departments, university IT, university libraries, national orgs)

How much time we spend analysing data?

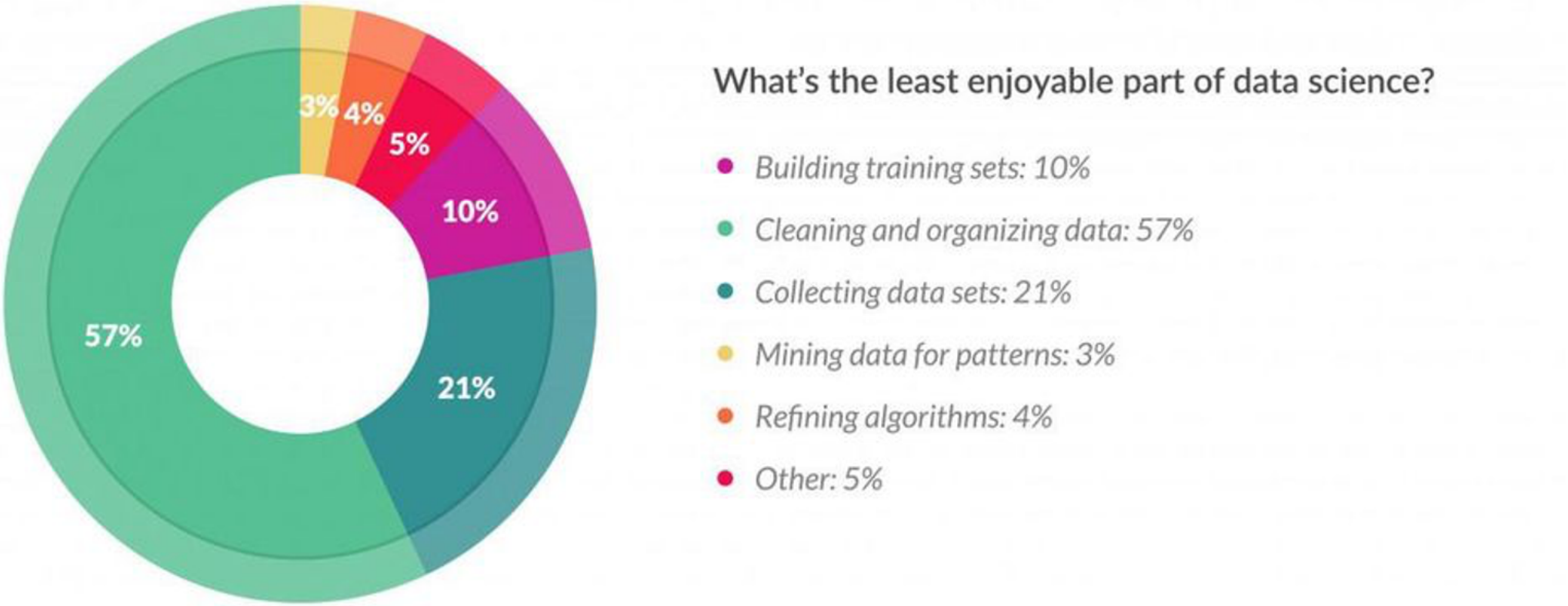


What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Forbes

What we don’t like to do?



Source: Forbes

Strategic rational

- Researchers employing data resources are invariably faced with data exploration 'from scratch’
- Re-analyzing entire datasets in search of features, which were identified by preceding inquiries
- Typically ‘waste’ 80% of resources by re-doing data processing (i.e., collecting, cleaning and organizing data)
- Research investments are duplicated
- Opportunities for building new knowledge upon previous experience are lost
- Waist max 20% resource **use min 80% on new knowledge/information generation**

Sourced from unsuccseful DFF proposal on Data Annotations: https://zenodo.org/record/3925126

Short-term targets (by end of 2020)

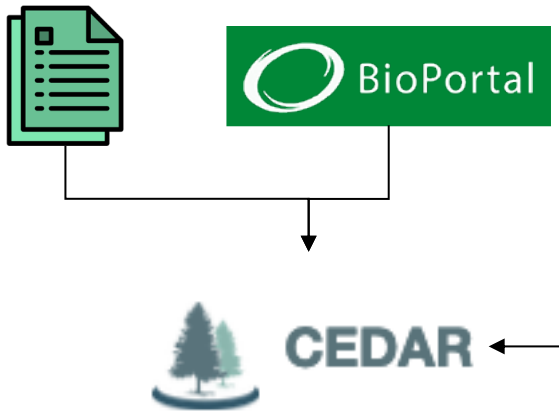
- Refurbished and publish first version of metadata templates for:
  - Controlled terminologies
  - Datastreams
  - Datasets

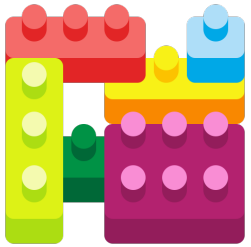
- Make several examples of instantiated templates
  - Taxonomy of topics and atmospheric variables
  - Min one model dataset metadata
  - Min one observational dataset metadata

- Make controlled terminologies public

- Provide inputs for CEDAR plugin framework development

Mid-term target





magic
'middleware'

DOI generation

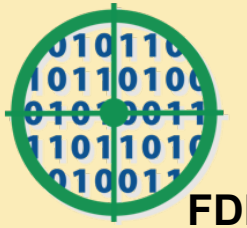
Diba

DataCite
FIND, ACCESS, AND REUSE DATA

DTU Data
- a service from DTU Library

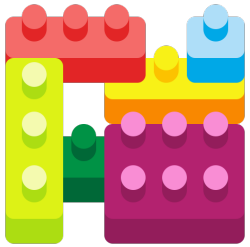
(meta)data repository

DTU Data
- a service from DTU Library

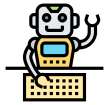


FDP

Mid-term target



magic
'middleware'



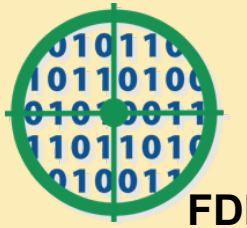
DOI generation

Diba

DTU Data
- a service from DTU Library

(meta)data repository

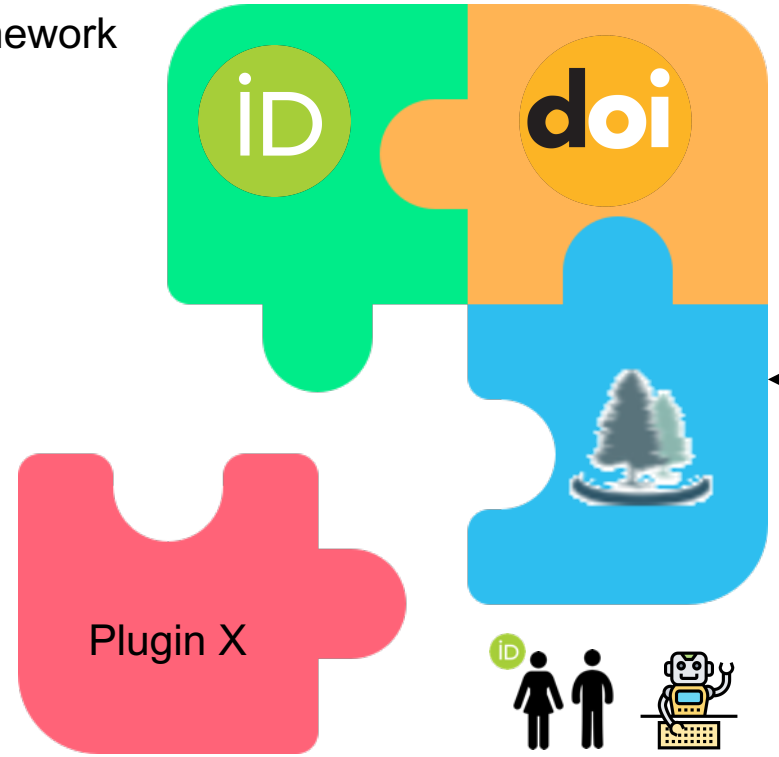
DTU Data
- a service from DTU Library



FDP

Mid-term target

CEDAR
pluging
framework



(meta)data repository

DTU Data
- a service from DTU Library

FDP

Long-term milestones

- Develop adaptable open-source Data Management (DM) platform to Preserve, Link and Distribute (meta)data:
  - Adhere/enforce the FAIR data principles
  - Tailored for Big and (un)Structured data

  - Key Features:
    - Automatically generate rich machine-actionable metadata for the entire data lifecycle:
      - Potentially build a suite of CEDAR plugins to auto-populate metadata templates (see previous slide)

    - Assign unique, resolvable and persistent identifiers (PIDs):
      - Resolve PIDs to a web page for humans
      - Resolve PIDs to a set of instructions for machines

  - Allow support for real-time annotation:
    - Annotate features of interest
    - Support for data subsetting
    - Allows for ML model training

Long-term milestones

- Improved metadata import
  - Allow for DataCite compliant metadata to be uploaded to figshare, zenodo, b2share, etc.
  - More generic JSON-LD upload
- Support for large datasets
  - Allow for data hierarchy
  - Searching within dataset
- Support for structured data
  - Allow for subsetting of structured files

Aspects

- Requires **Departamental, University and National TOP-DOWN push and sound funding**
  – *We can do as much as we can in free time we have, but it is not sustainable niether fair*

- Organizational aspect:
  – DM prototyping to be done by techy domain experts in colab with Uni IT or DeiC IT
  – DM MVP and long-term maintanance to be hand over to Uni IT teams or
  – DeiC to establish tech team for data management
  – DeiC to disseminate platform at the national level and coordinate activities among Danish universities and internationally (*you need to advertize our work and me us famous*)

- Have techy domain experts to build new features with IT team:
  – IT to maintain and make DM platform robust

- By all means do not provide us with a unextendable / rigid / unusable / ready-made / Elsavier solutions

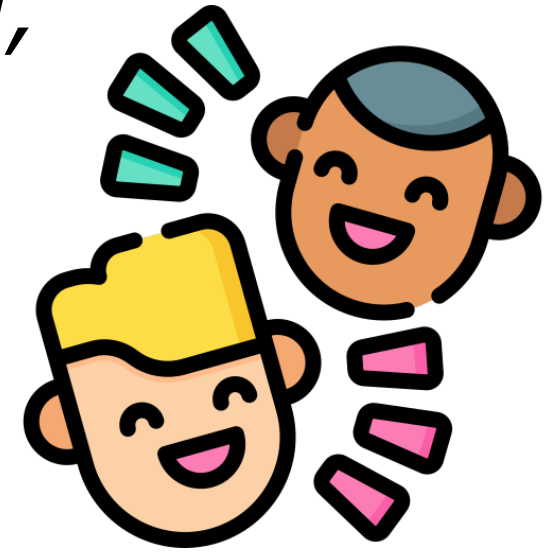
Division of labour

- Research Groups + Uni/DeiC IT with feedbacks from Uni Libraries to prototype DM platform
- Uni IT and/or DeiC IT to robust and maintain DM platform
- DeiC to lead securing funding for the above activities at national scale:
  – Call for national colab
- DeiC on the national, Uni Lib at Uni level, to make sure that research groups converge and openly share the data management work

Conclusion

"*Talk to us once you have political and strategic agenda set and you have funding to spend,*

we will very well know how to use it!"



… as you did for organizing this M4M workshop

Thank you!

niva@dtu.dk
neda@dtu.dk