Digital Research Infrastructure for Humanities

DeiC Science Forum F21

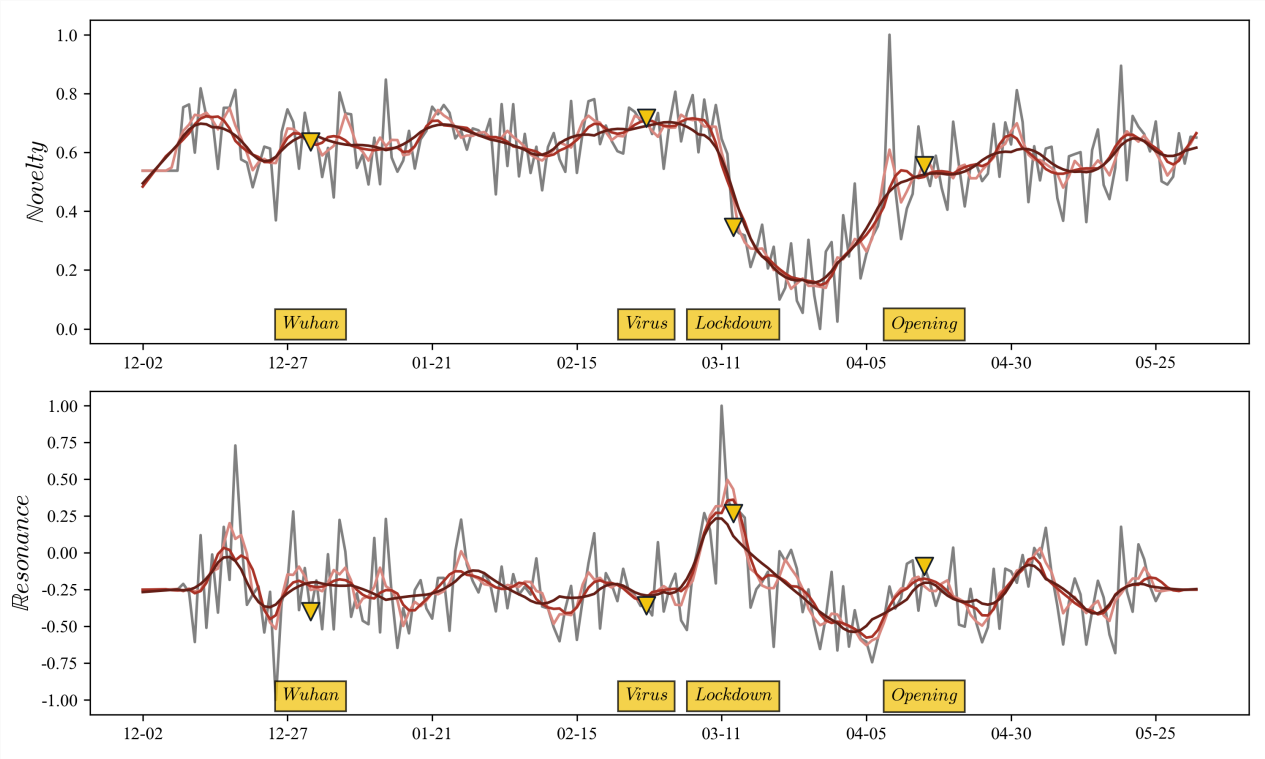
Kristoffer Nielbo (OBO Janne Nielsen)

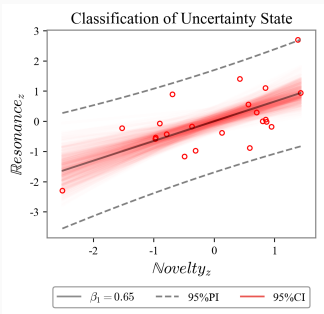
Center for Humanities Computing AArhus|chcaa.au.dk
aarhus university, denmark

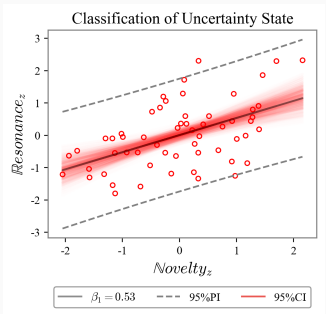
CENTER FOR HUMANITIES
COMPUTING AARHUS

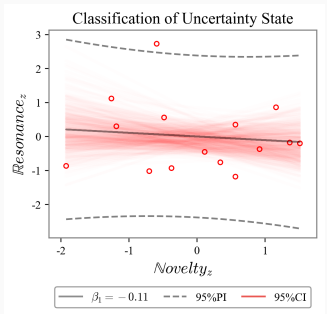
WHEN A FEW DATA POINTS ARE ENOUGH

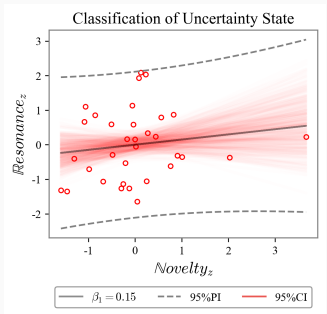
AND WHEN THEY ARE NOT...

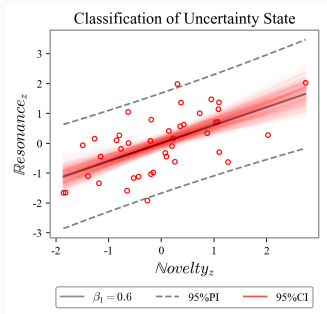




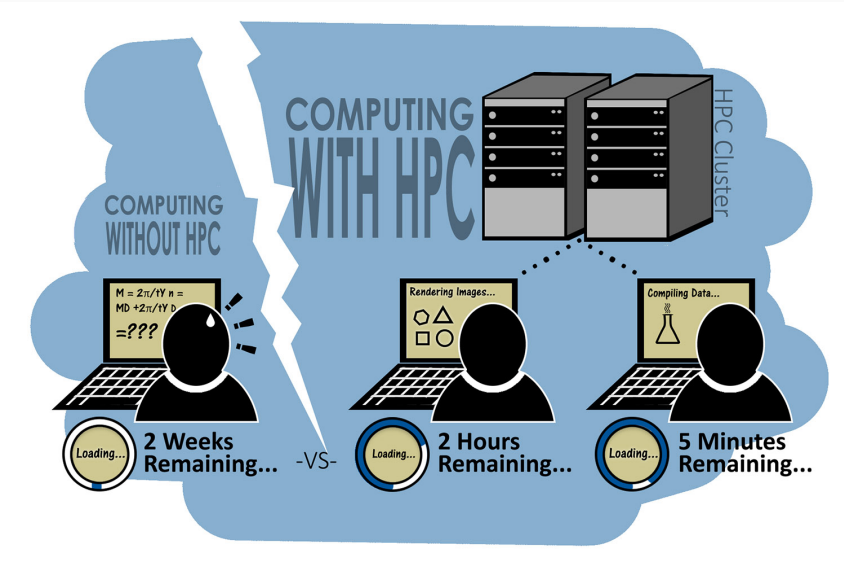








Humanities Computing



TRADITIONAL USERS

By default HPC *is not for all*: "you must learn to crawl before you can walk"

1. write code base for project
2. formulate batch script
3. submit to job queue

NEW USERS

New technologies (Big Data, AI) have created new users

- easy & interactive access
- collaborative development
- code and data sharing

for exploration, experimentation, and debugging in a sandbox-like environment

Human(ist)-in-the-Loop Technology

a need for HITL

... as task complexity increases, a need for (operational approaches to) leveraging human intelligence in the development of learning algorithms has become apparent

Type	Human Involvement	Resources	Relevance
Out-of-the-loop	not required	low	low
On-the-loop	checking	medium	medium↓
In-the-loop	required	high	medium↑

WHEN

algorithms are not understanding the input

data input is interpreted incorrectly

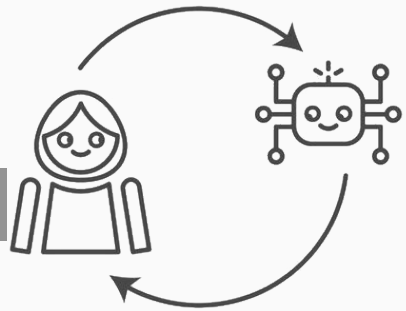
algorithms do not know how to perform the task

to make models more accurate

cost of errors is too high in development

data are rare or not available

THEN



HITL Models

HPC - PROJECTION

seamless access to more (gpu) nodes in the cloud

- maintain low barrier to entry
- interactive DREs
- diversity of tools

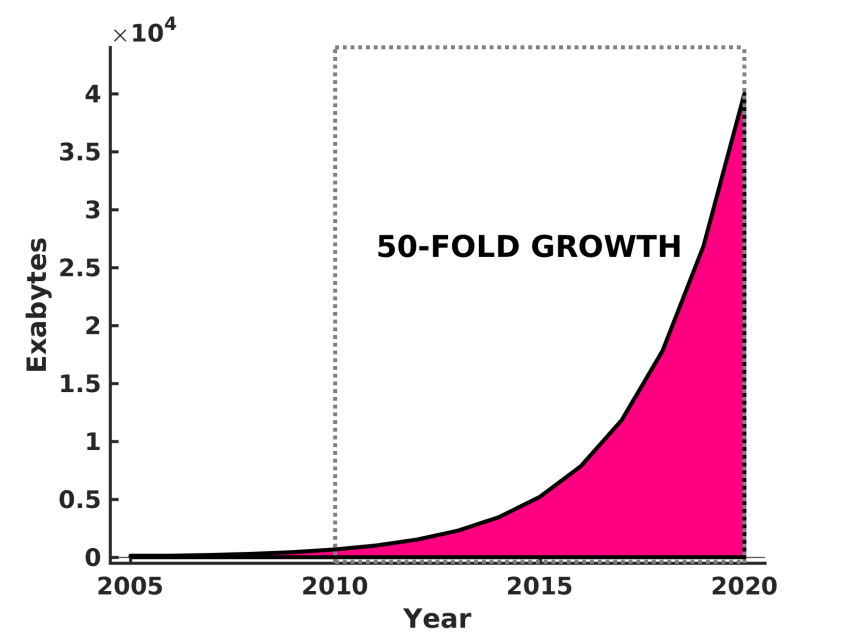
need-to-have :: **research support-as-a-service**

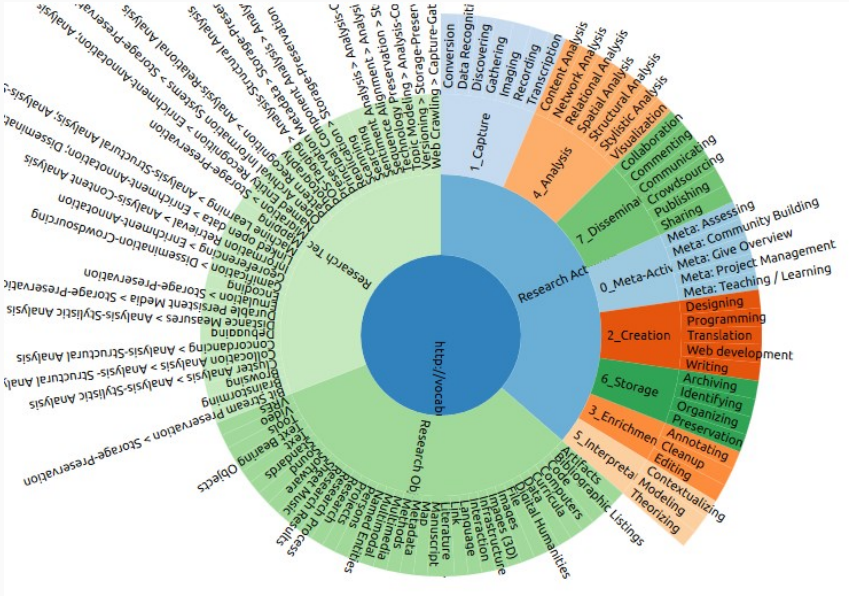
- extensive support infrastructure (front, back and inbetween)

nice-to-have :: **educational challenge**

- reliable access to cloud for teaching and training

Humanities Data





data :: **large, soft & heterogeneous**

dual problem :: data are **sensitive & restricted** access

derived problems::

- data are relatively easy to access
- no standardized procedures for risk-benefit evaluation
- research evolves at a faster scale than legal
- diversification of tools for management and analysis

operational data management standards are a must

RDMS & network - projection

DM, storage and archival solution*s*

- collaboration (inter-national)
- compliance issues (GDPR, copyright...)
- *one* solution will not fit all

need-to-have :: **data-as-a-service**

- cloud-based access to observatories, archives, collections &c
- Nordic initiative on federated learning

nice-to-have :: **internationally valid data passport**

- AAAI for cloud-based data access

THANKS

kln@cas.au.dk
knielbo.github.io
chcaa.au.dk

SLIDES

knielbo.github.io/files/kln_sf_f21.pdf