HANDLING, ANALYSIS AND STORAGE OF BIG DATA FOR HEALTHCARE

MATLAB EXPO 2021

Spencer A Thomas
Senior Research Scientist, NPL

NPL: Josephine Bunch, Ian Gilmore, Alice Harling, Greg McMahon, Bin Yan, Rory Steven, Adam Taylor, Chelsea Nikula, Tingting Fu, Efstathios Elia, Ala Al-Afeef, Alex Dexter, Teresa Murta, Robin Philip, Kenny Robinson, Amy Burton, Rasmus Havelund, Paulina Rakowska, Jean-Luc Vorng, Xavier Loizeau, Ariadna Gonzalez, Weiwei Zhou, Ammar Nasif, Marcel Niehaus, Junting Zhang.

Francis Crick Institute: Mariia Yuneva, Peter Kreuzaler, Avinash Ghanate Yulia Panina, Chandan Seth Nanda, Daria Thompson, Eileen Clark, Marion Karniely, Lucy Collinson

University of Cambridge: Kevin Brindle, Jyotsna Rao, Maria Fala

Barts Cancer Institute: John Marshall, Shreya Sharma, Joseph Hartlebury

Imperial College London: Zoltan Takats, Paolo Inglese, James McKenzie, Ala Amgheib, Liam Poynter, Rita Carvalho, Seyma Turkseven, Vincen Wu

Beatson Institute: Owen Sansom, Andrew Campbell, Arafath Najumudeen, David Gay, Madelon Paauwe, Lucas Zeiger, Saverio Tardito, David Lewis

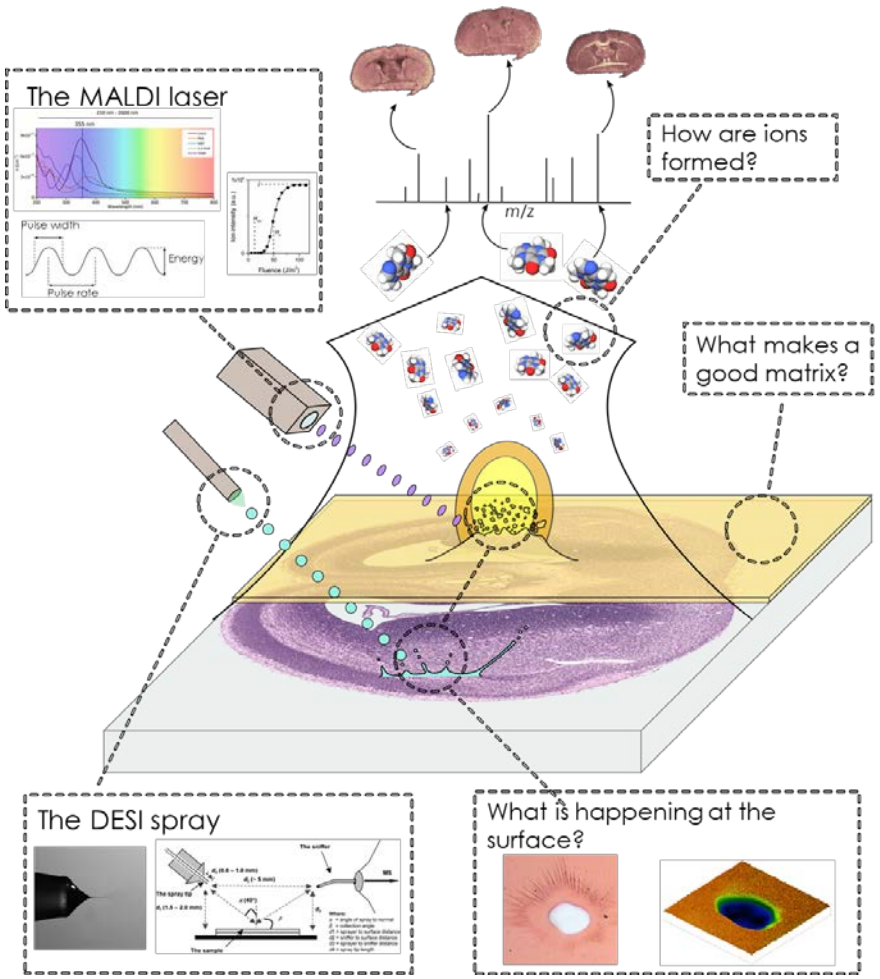
AstraZeneca: Richard Goodwin, Simon Barry, Greg Hamm, Nicole Strittmatter, Daniel Sutton, Stephanie Ling, Alan Race

Institute of Cancer Research: George Poulogiannis, Amit Gupta, Aurelien Tripp, Evi Karali, Nikolaos Koundouros, Thanasis Tsalikis

University College London: Gyorgy Szabadkai

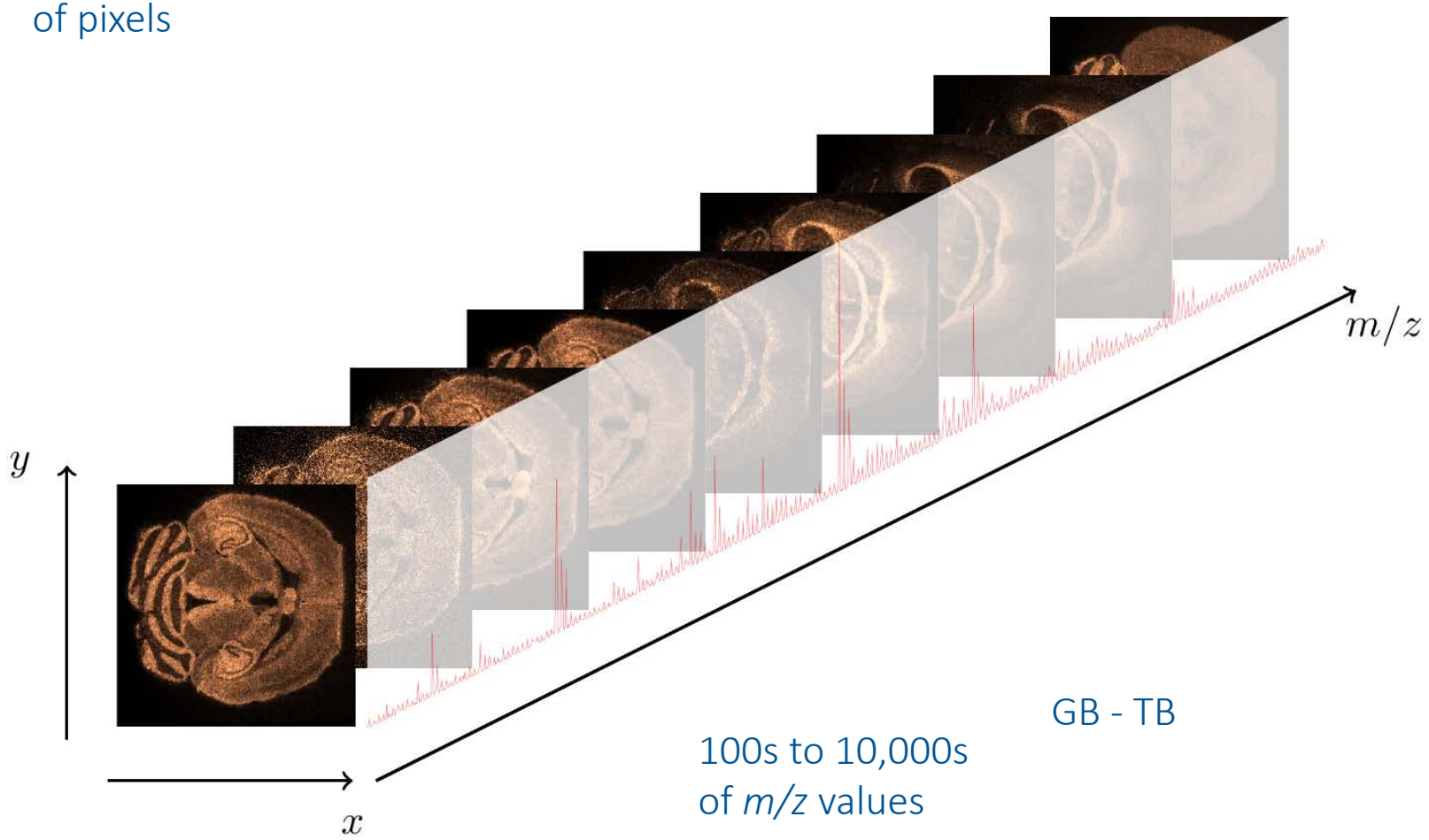
CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

MASS SPECTROMETRY IMAGING



Kenny Robinson

10,000s to millions of pixels

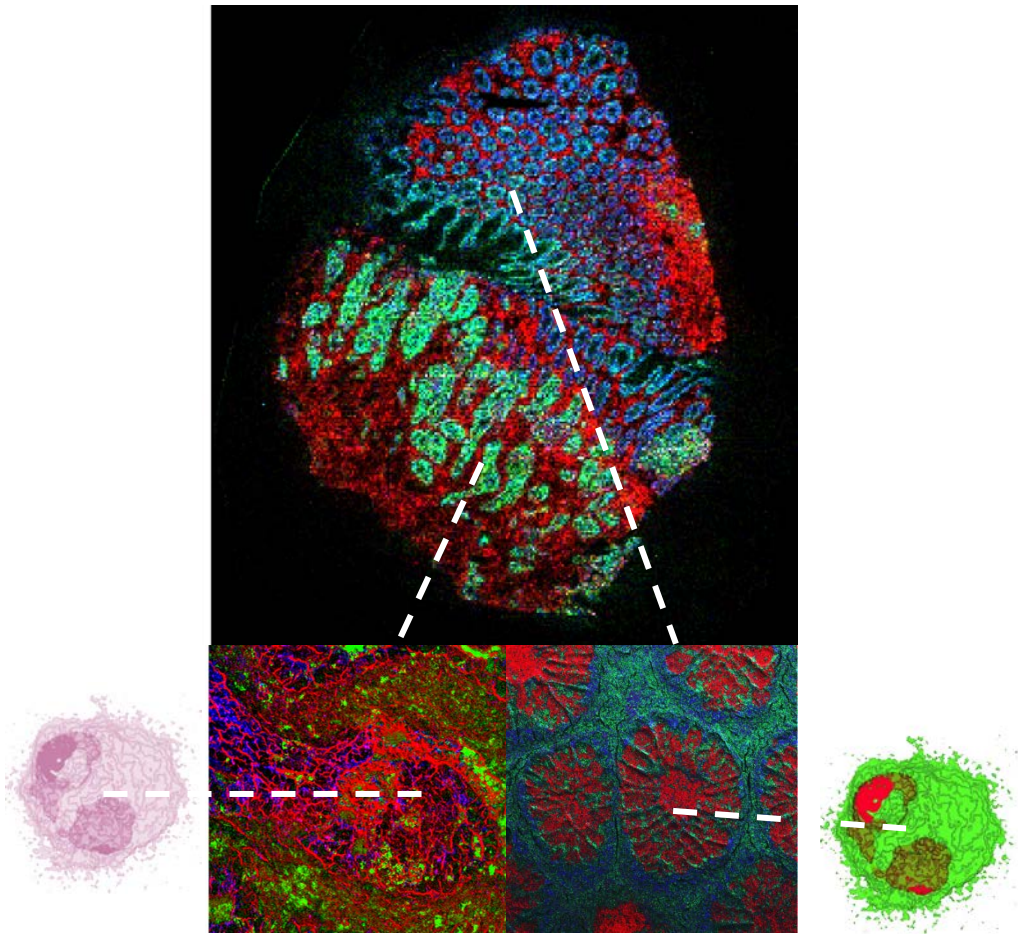


GB - TB

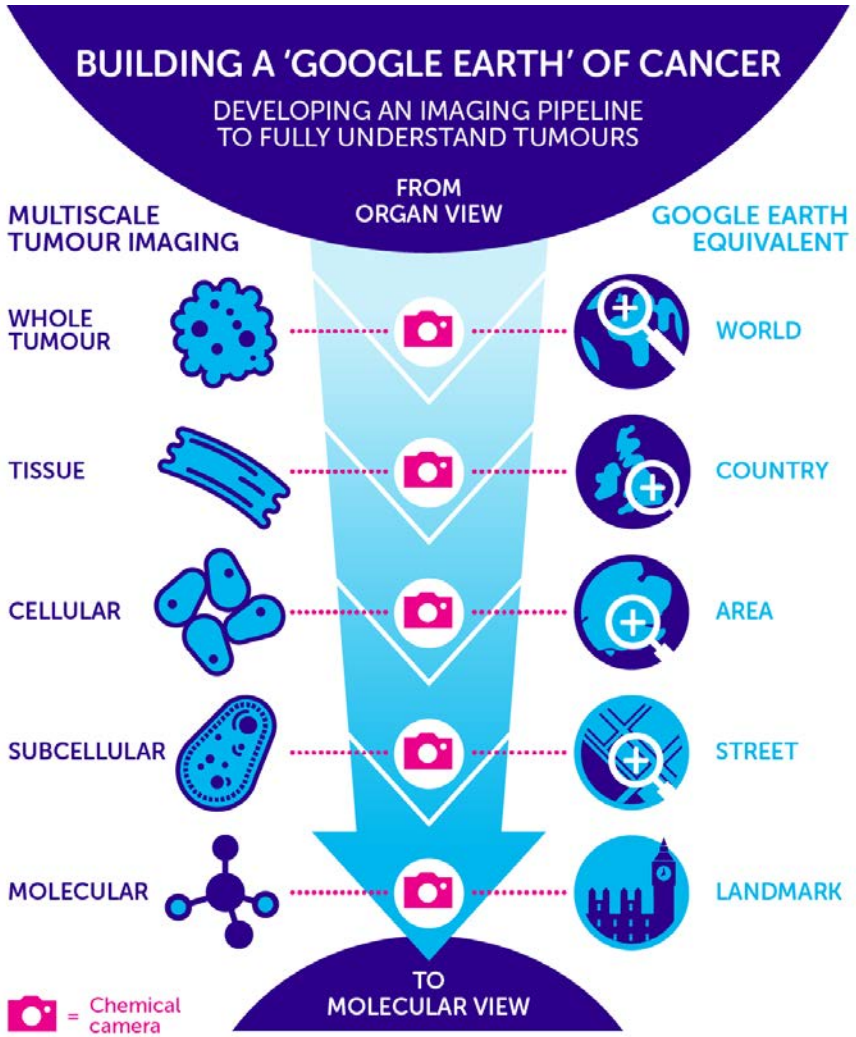
100s to 10,000s of *m/z* values

ROSETTA TEAM: CRUK GRAND CHALLENGE

National Physical Laboratory

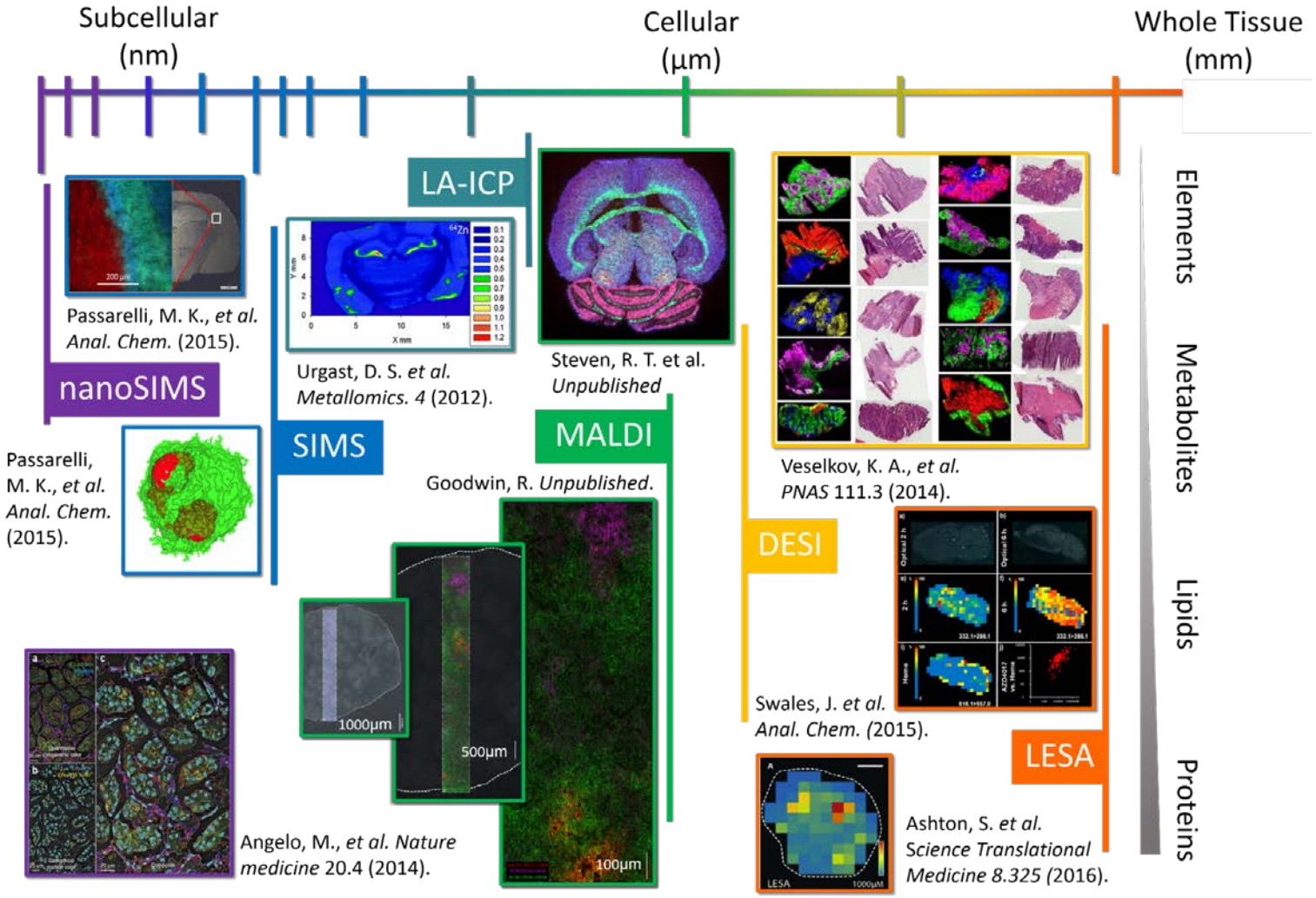


Diagnose & Treatment

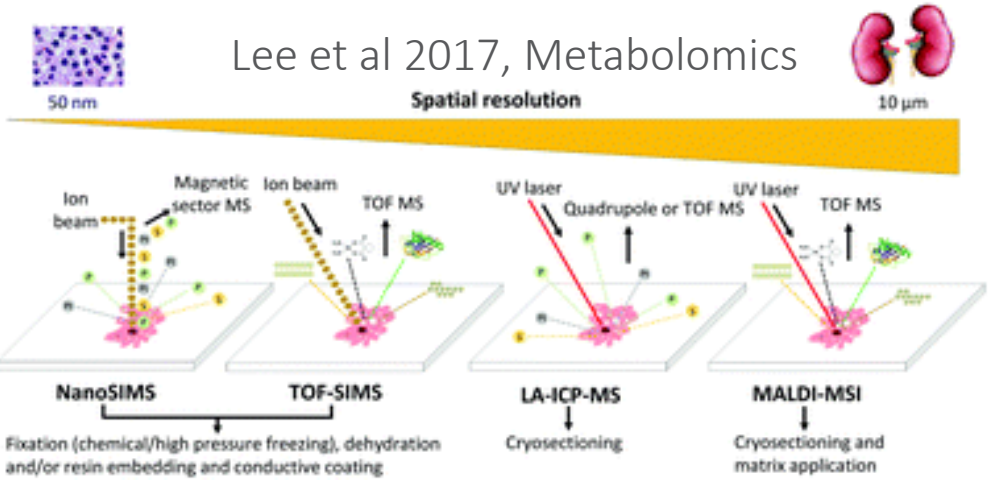


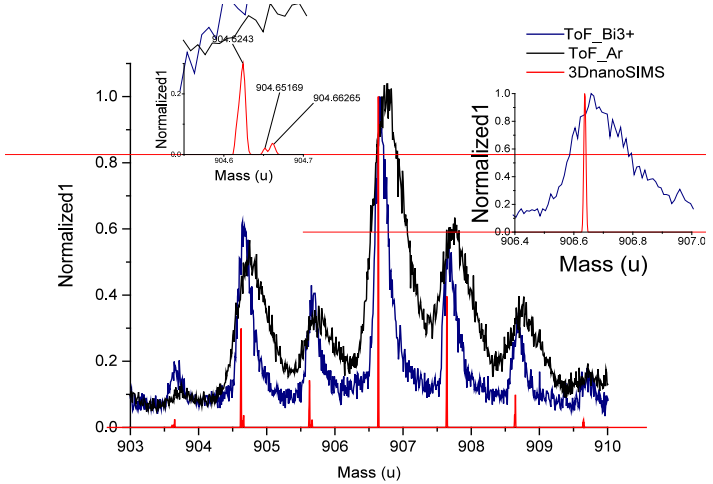
CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

MULTISCALE DATA

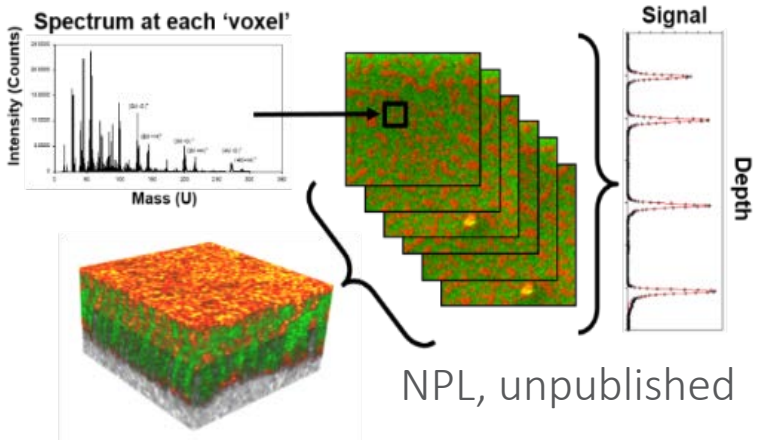


RESOLUTION, DIMENSION AND SCALE

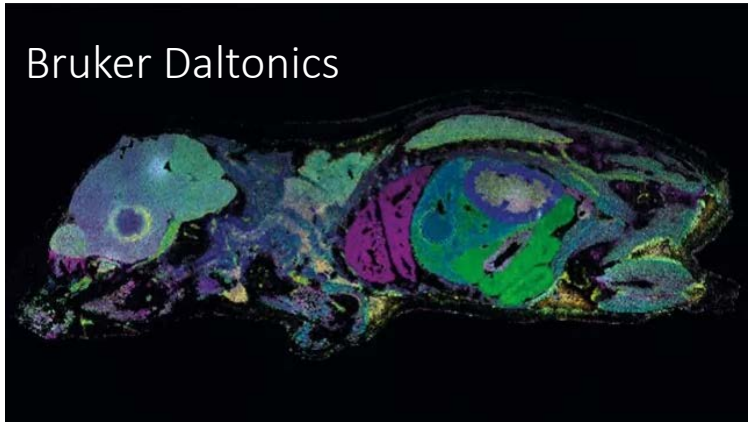




Passarelli et al 2017, Nature Methods



NPL, unpublished

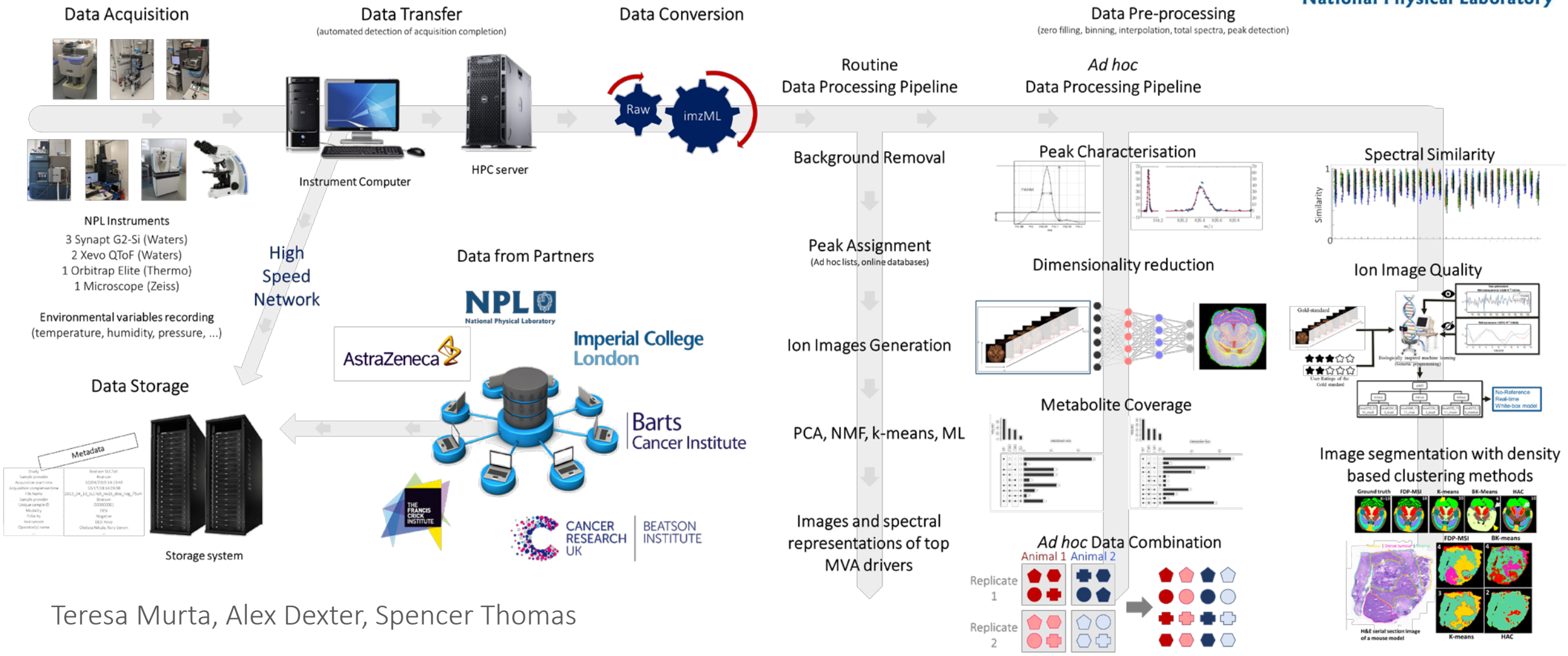


WHY MATLAB

NPL National Physical Laboratory

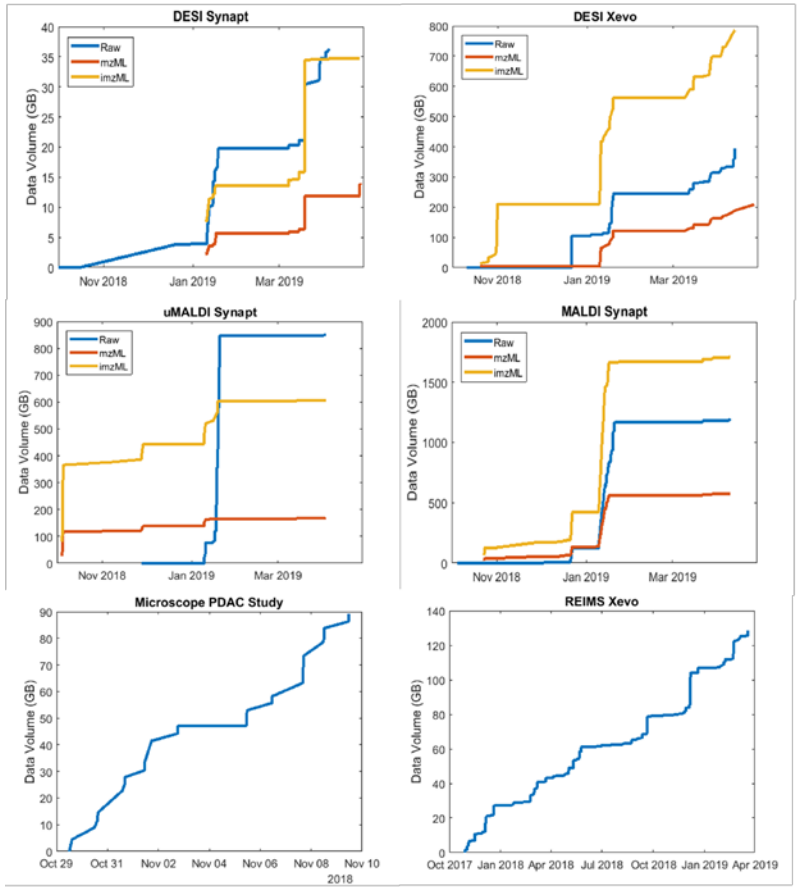
- Wide range of very well supported and documented packages
- Programming is easy, can quickly test ideas with little programming background
- Has attractive low level functionality (and interaction with other languages) if needed
  - Allows scientist for focus on what they are good at and want to explore
    - Experiments, chemistry and biology
  - the computer / data scientists can address any limitations within MATLAB
- Very aligned with MATLAB
  - Basic analysis is image and signal processing
  - MSI data treated as a matrix in almost all cases
- Large code base in MATLAB and no need to move away from it
  - Adding in to workflow rather than replacing it

HANDLING



Teresa Murta, Alex Dexter, Spencer Thomas

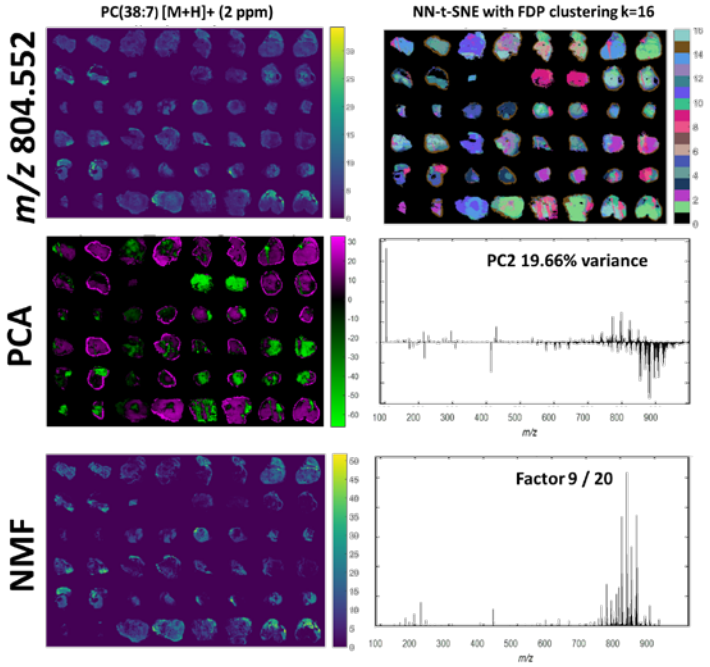
AUTOMATED PROCESSING



192 tissue sections
2 modalities, 2 polarities
217GB of raw data

Routine Data Processing Pipeline	Ad hoc Data Processing Pipeline	Manual Processing
1 day	2 weeks	120 days working 9am-5pm

DESI-MSI Positive Mode

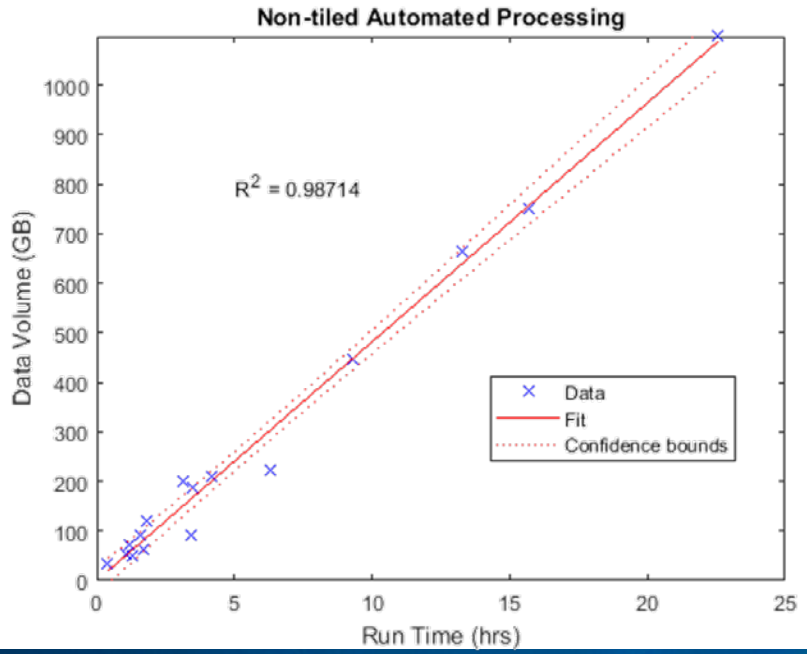


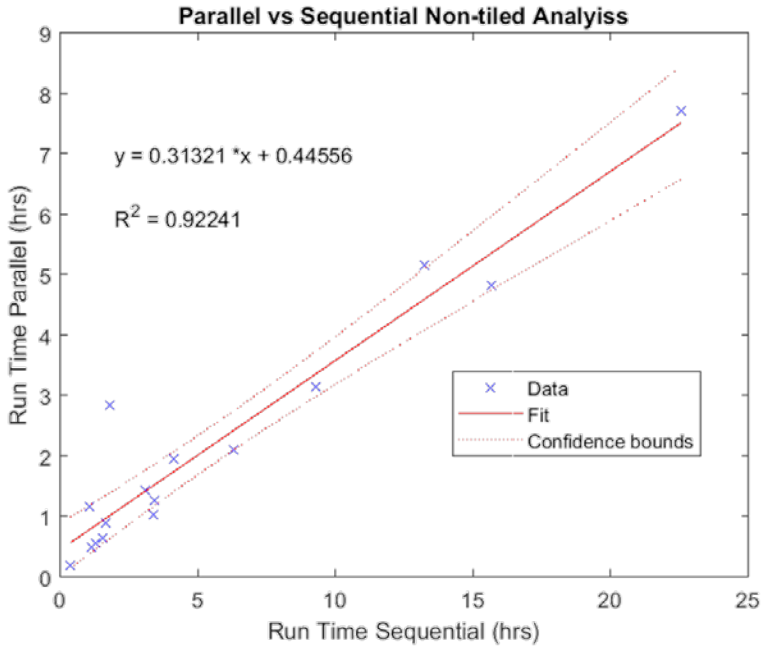
DISTRIBUTED AND PARALLEL PROCESSING

National Physical Laboratory

Large scale data processing

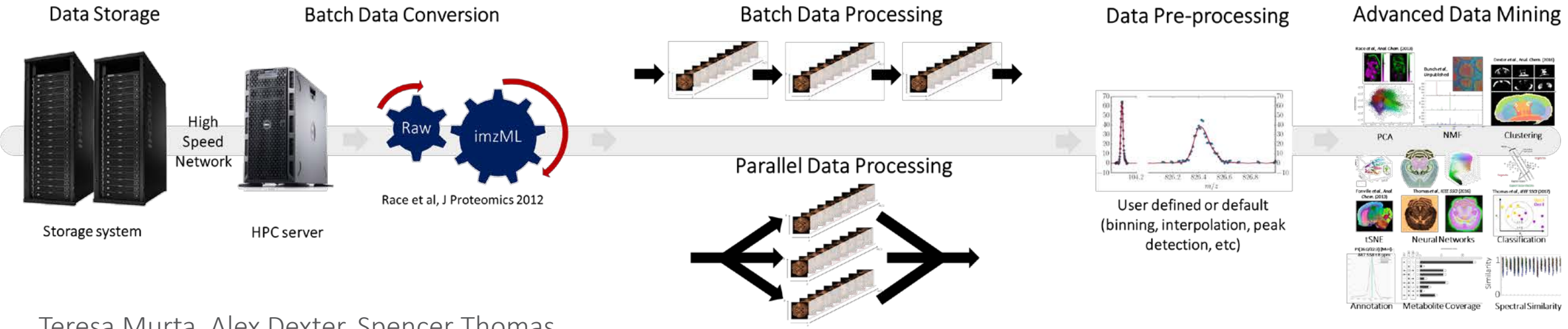
- **254 datasets**, containing **423 tissue sections**
- Comparable processing of large study MSI datasets defined by the user
- Run in batches for multiple studies
- Memory efficient processing of **4.3 TB** of imzML data
- Processed in **3.76 days** (sequence) or **1.5 days** (in parallel, 3 threads due to RAM)





CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

ADAPTIVE PROCESSING PIPELINES



Teresa Murta, Alex Dexter, Spencer Thomas

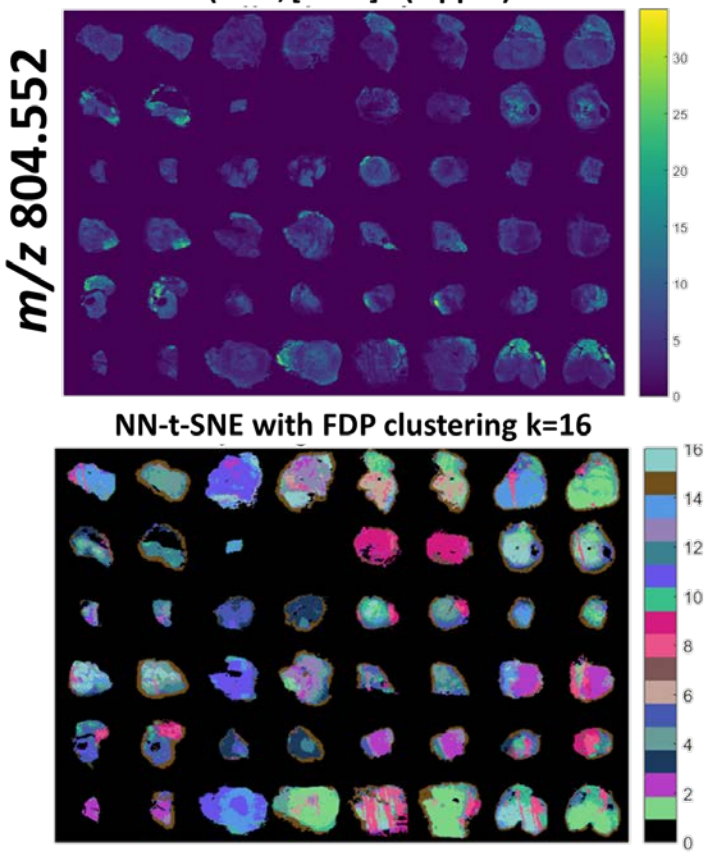
CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

LARGE SCALE PROCESSING

National Physical Laboratory

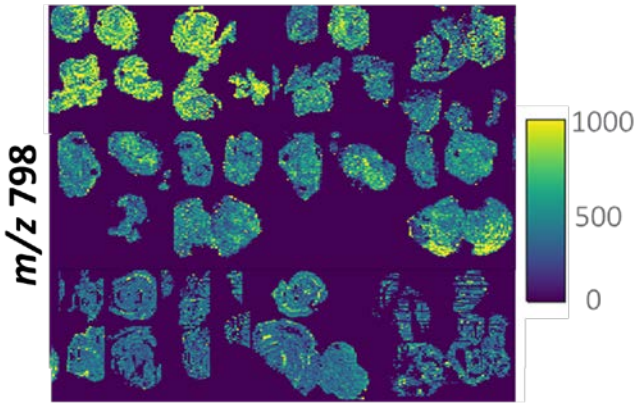
PDAC DESI-MSI (217 GB)

PC(38:7) [M+H]+ (2 ppm)

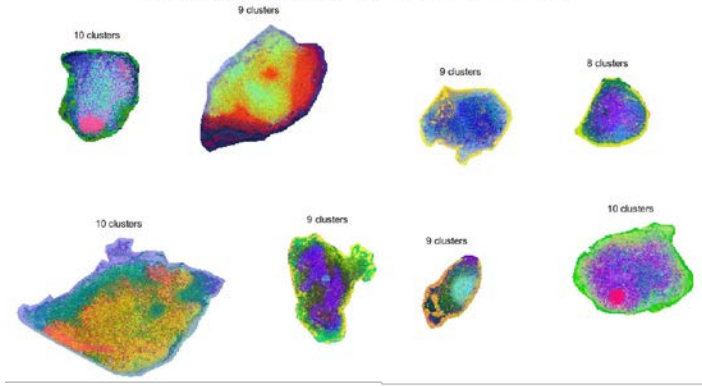


NN-t-SNE with FDP clustering k=16

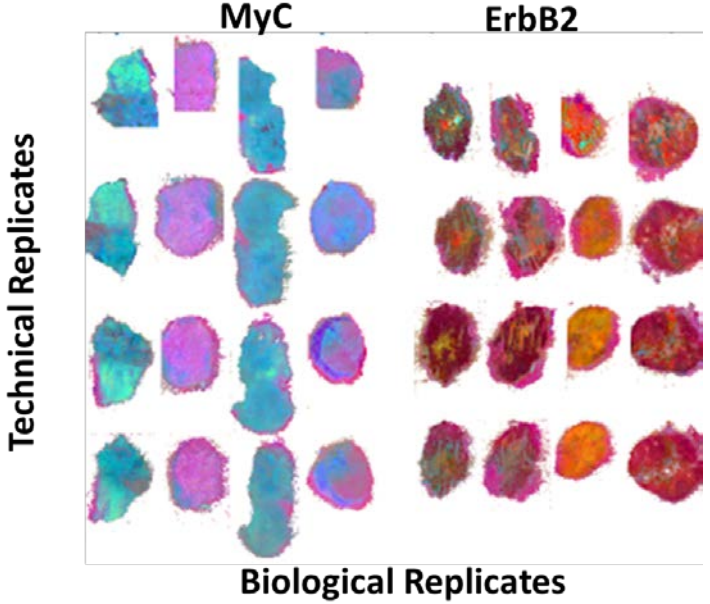
CRC MALDI-MSI (556 GB)



Xenograft MALDI-MSI (90 GB)



Breast DESI-MSI (197 GB)

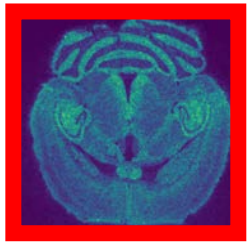


Spencer Thomas, Alex Dexter, Adam Taylor, Chelsea Nikula, Rory Steven, Josephine Bunch

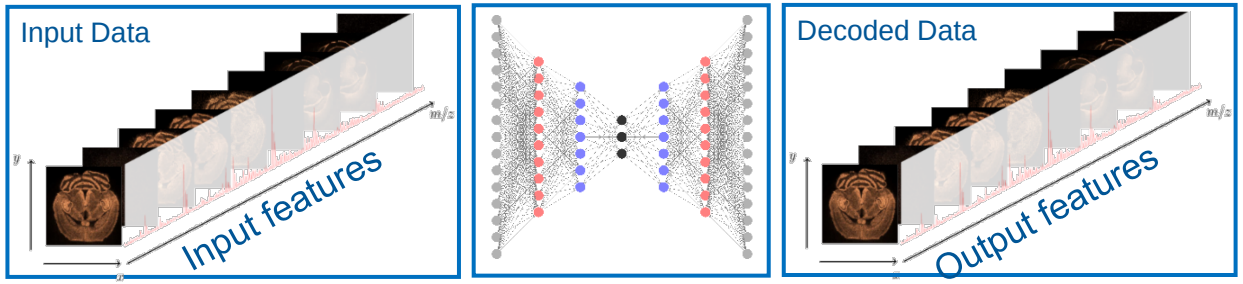
CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

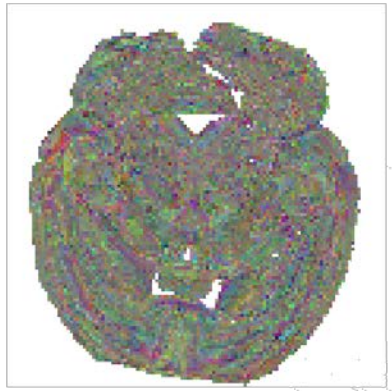
ADVANCED ANALYSIS

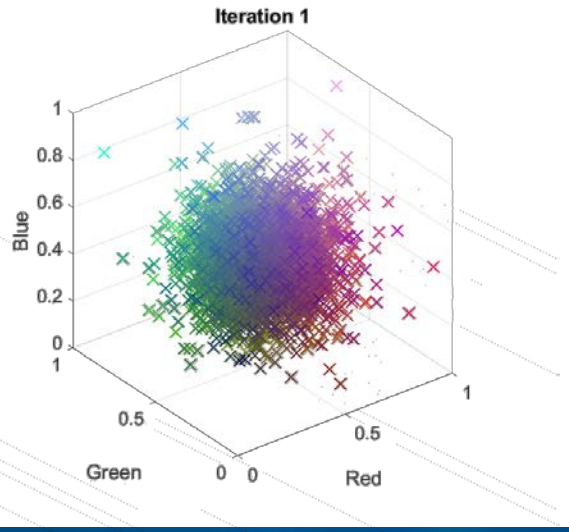
National Physical Laboratory

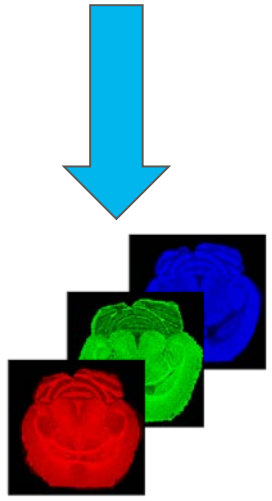


Unsupervised dimensionality reduction

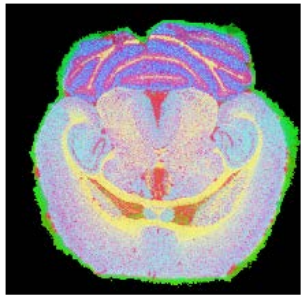




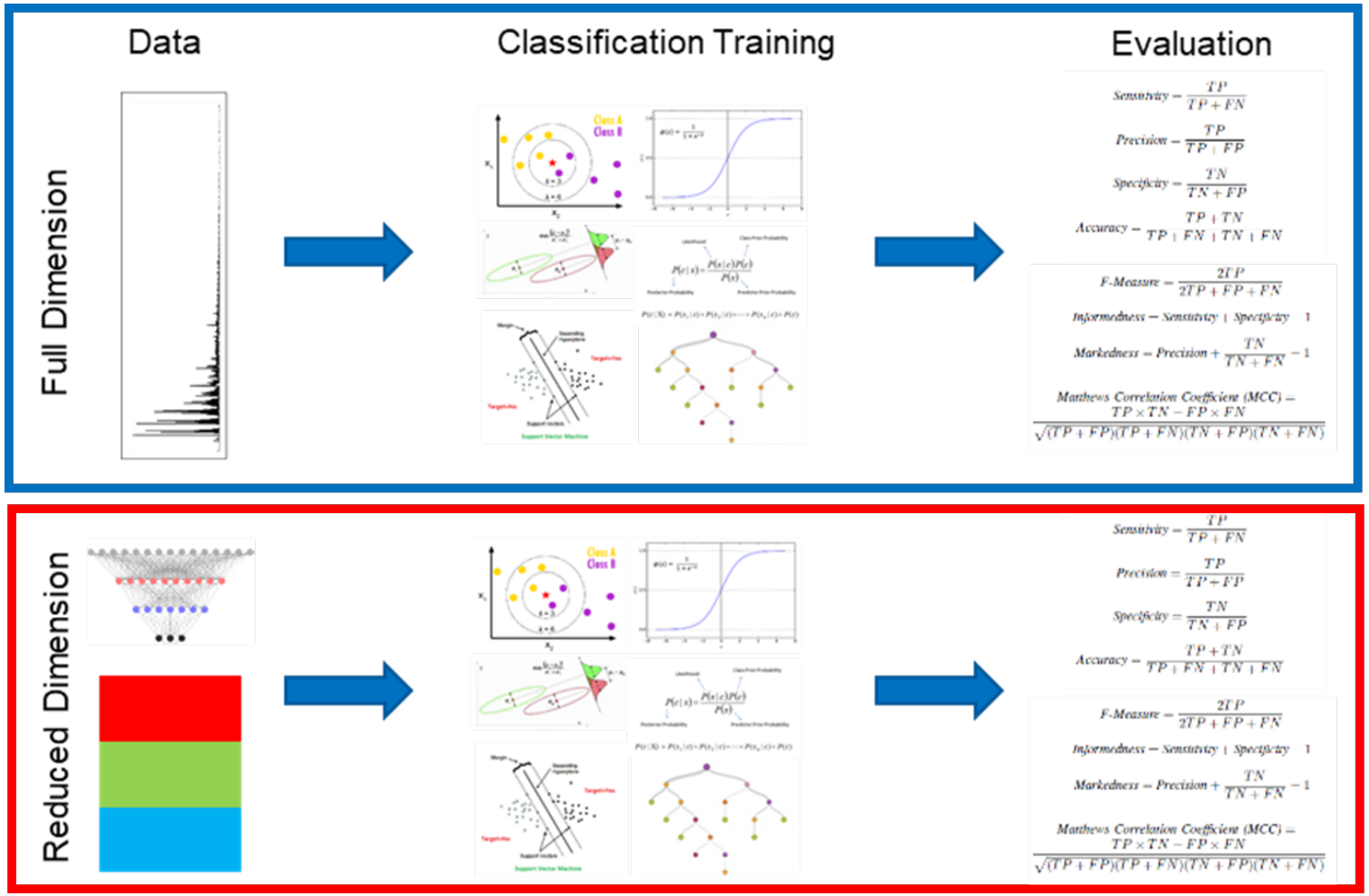




Visualisation of all features

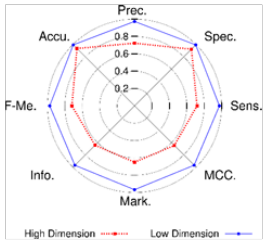


ADVANCED ANALYSIS

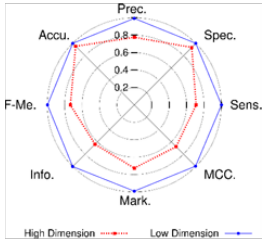


1 = ideal learning algorithm
1 > overfitting
0 = can not classify

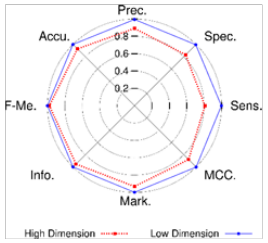
KNN



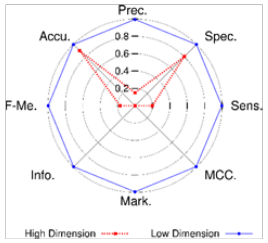
LDA



SVM



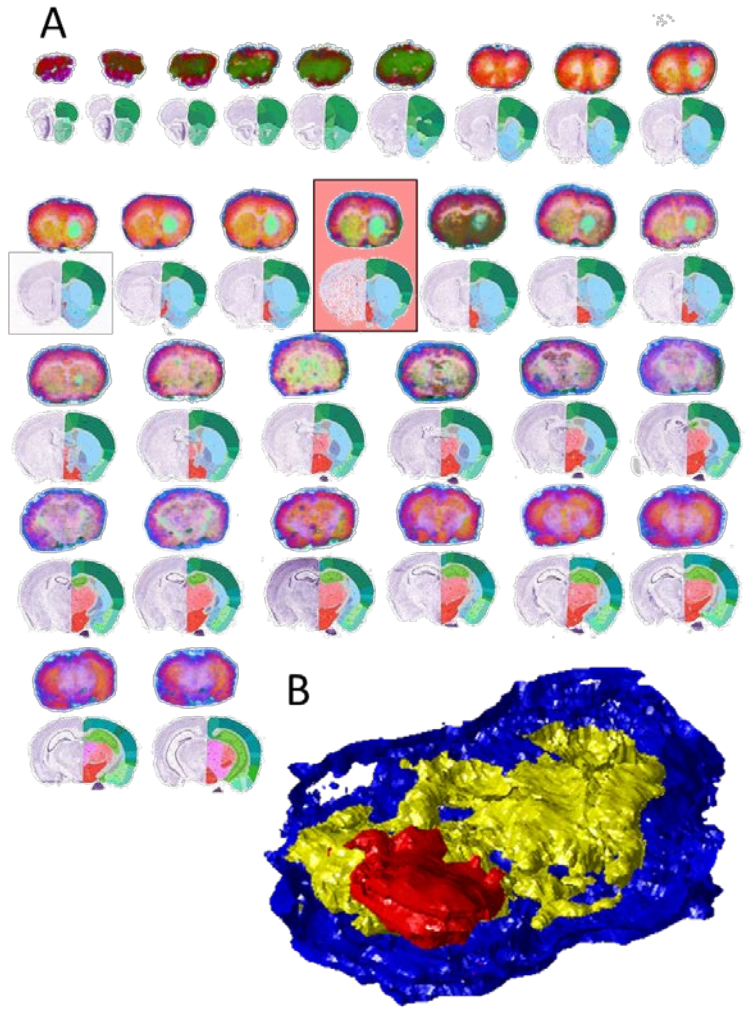
SVM-Gauss



CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

S. A. Thomas et al "Enhancing classification of mass spectrometry imaging data with deep neural networks," *2017 IEEE SSCI*

EXTENDING THE STATE-OF-THE-ART



A Dexter et al, Training deep learning on non-linear dimensionality reduction of subsampled data to analyse prohibitively large datasets, *BioArxiv*

IMPORTANCE OF DATA METROLOGY

"… the lack of reliability regarding the laboratory work at Harvard means that we are now retracting this paper."

Doi: 10.1016/S0140-6736(19)30542-2

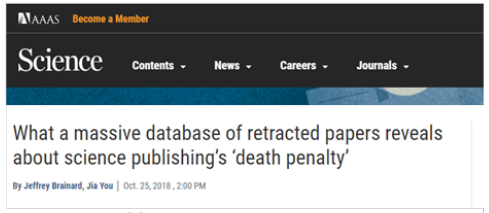
"… concerns that have been raised about the reliability of the database"

"… failed to adequately explain its data or methodology."

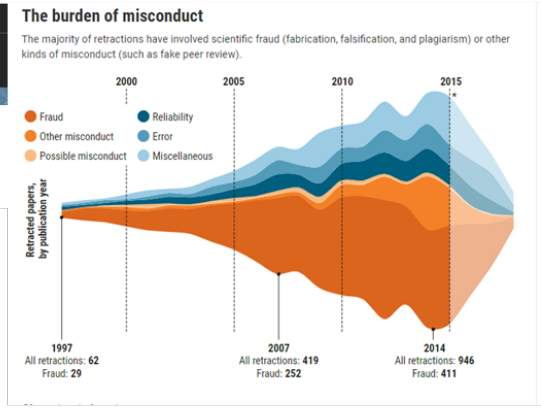
https://www.theguardian.com/world/2020/jun/04/

https://www.the-scientist.com/critic-at-large/

Lack of data transparency.
Misleading statistics.
Suspicious lab practices

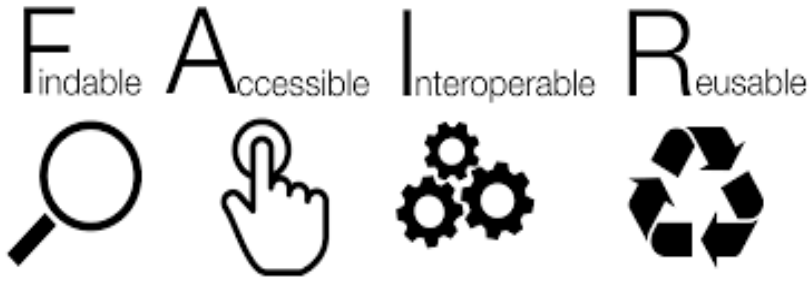


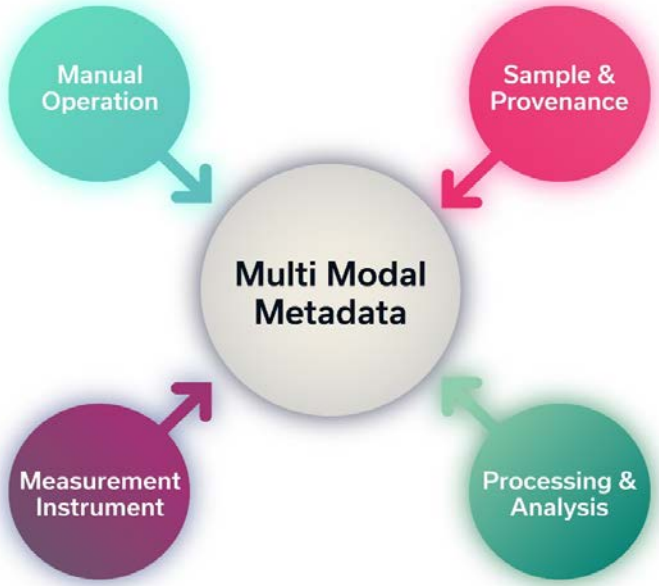
https://www.sciencemag.org/news/2018/10/

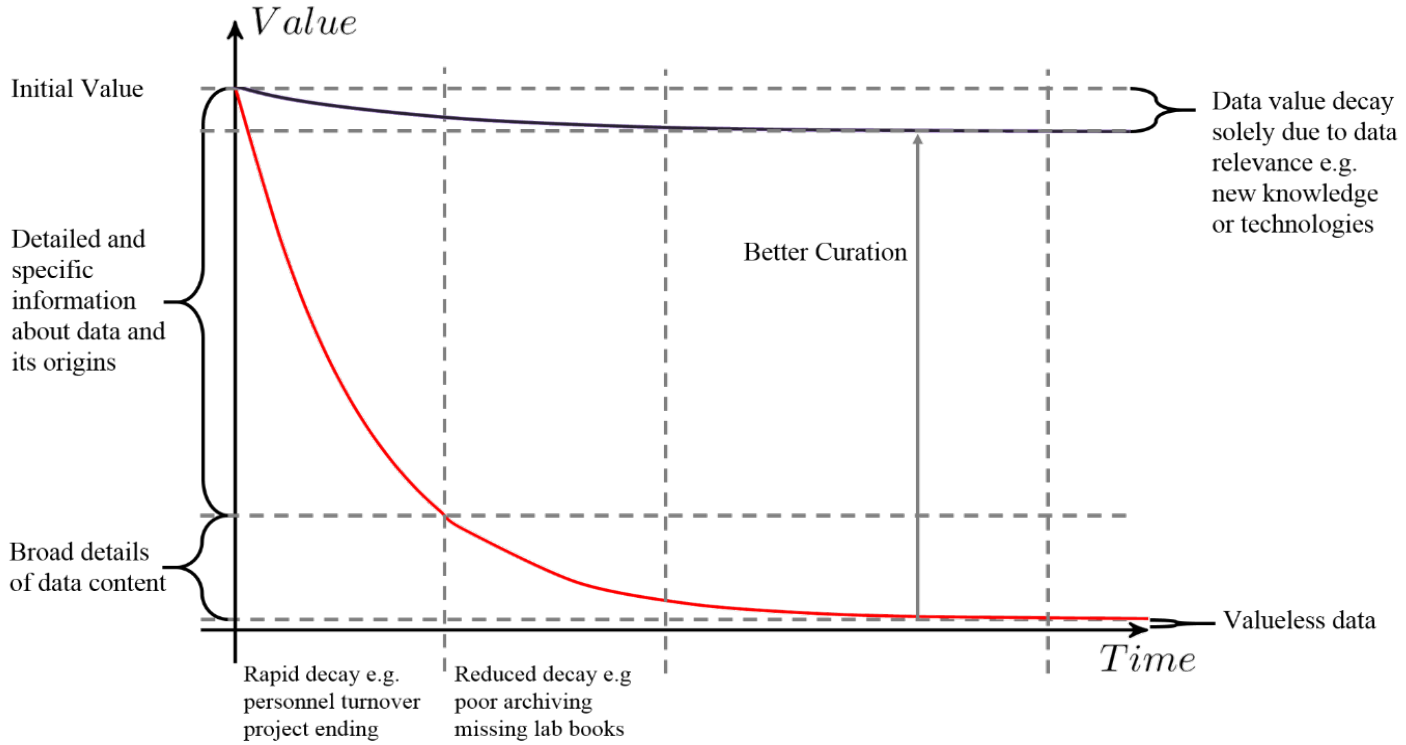


THE LONG-TERM VALUE IN DATA

NPL National Physical Laboratory



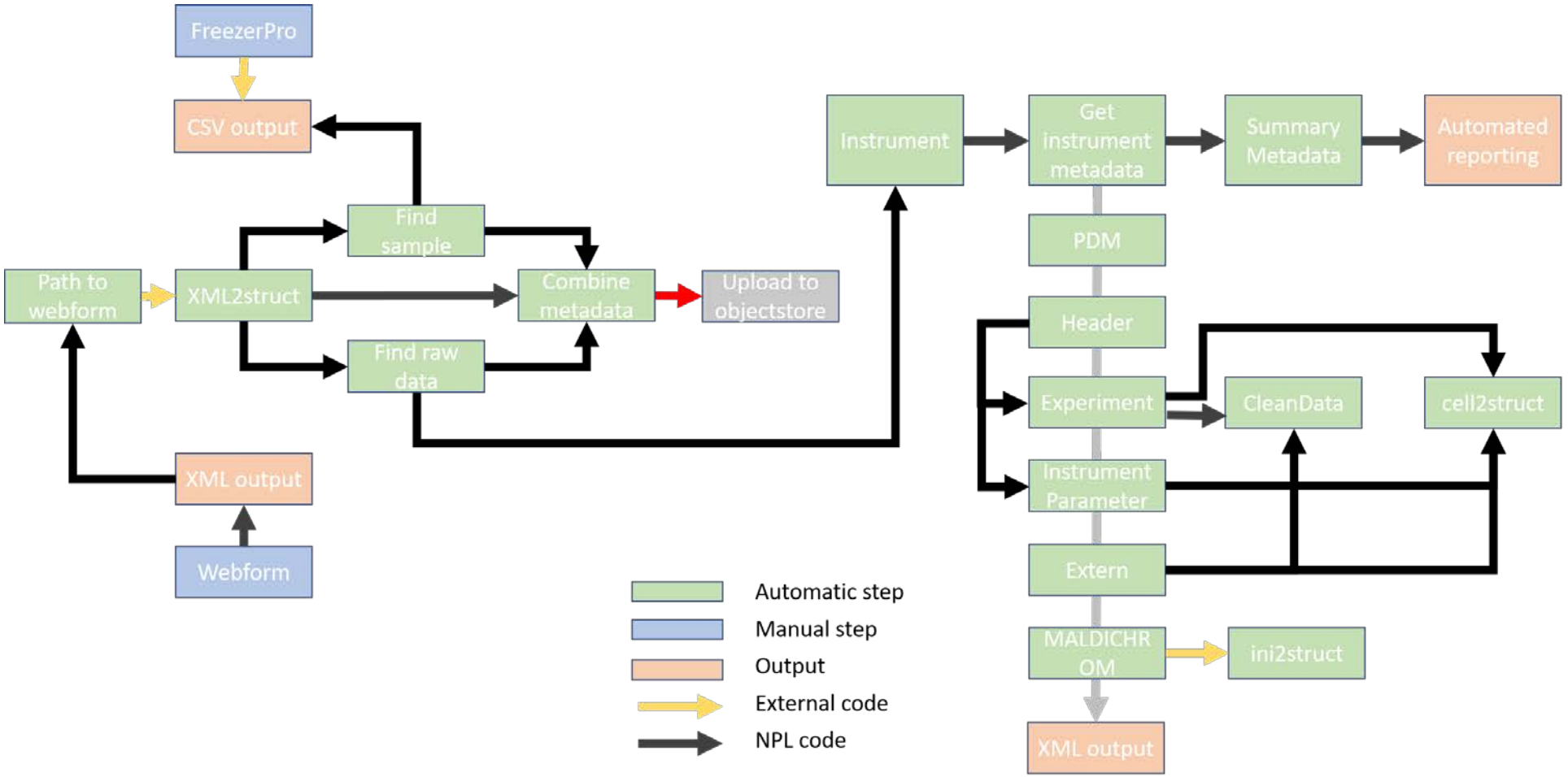




CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

N. A. S. Smith et al, Building confidence in digital health through metrology, *Br J Radiol*. 2020

AUTOMATING DATA CURATION



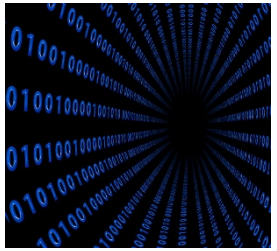
ROSETTA CASE STUDY

NPL National Physical Laboratory

Department for Business, Energy & Industrial Strategy

Traceable Science from data to reports

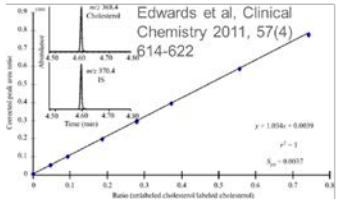
Raw Data



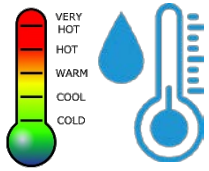
Processing and Analysis Codes

Science outputs

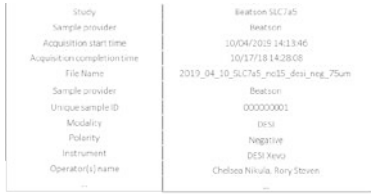
Calibration Data

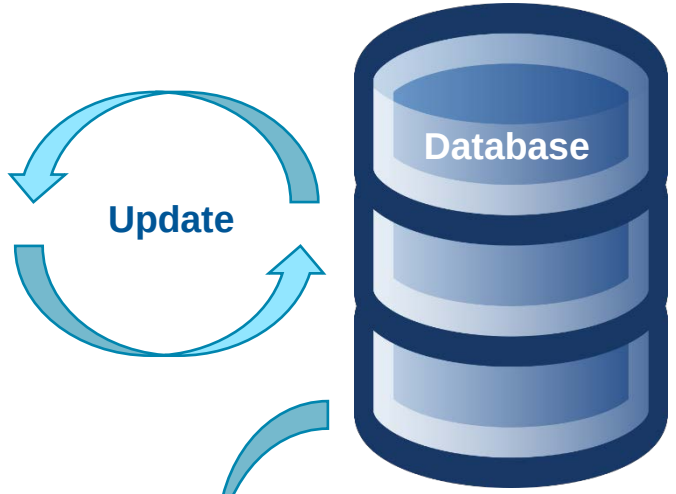


Environmental Conditions



Experimental Settings





Future-proof data analysis

Link to data from future technologies

CANCER RESEARCH UK | GRAND CHALLENGE ROSETTA

TOOLBOXES

NPL National Physical Laboratory

Statistics and Machine Learning Toolbox

Bioinformatics Toolbox

Deep Learning Toolbox

Image Processing Toolbox

Parallel Computing Toolbox

MATLAB Parallel Server

TAKEAWAYS

- Use ready made and well support elements to solve your challenges
- User friendly and extendible code facilitates innovative science from the experimentalist and computationalist
- Some solutions are simple, if you know what is possible
- Focusing on the biggest bottle neck has revolutionised our data processing capabilities

ACKNOWLEDGEMENTS

MATLAB EXPO organisers for the invitation

NPL colleagues in NiCE-MSI

Rosetta Grand Challenge Team

Thanks for listening

FUNDED BY BEIS

The National Physical Laboratory is operated by NPL Management Ltd, a wholly-owned company of the Department for Business, Energy and Industrial Strategy (BEIS).