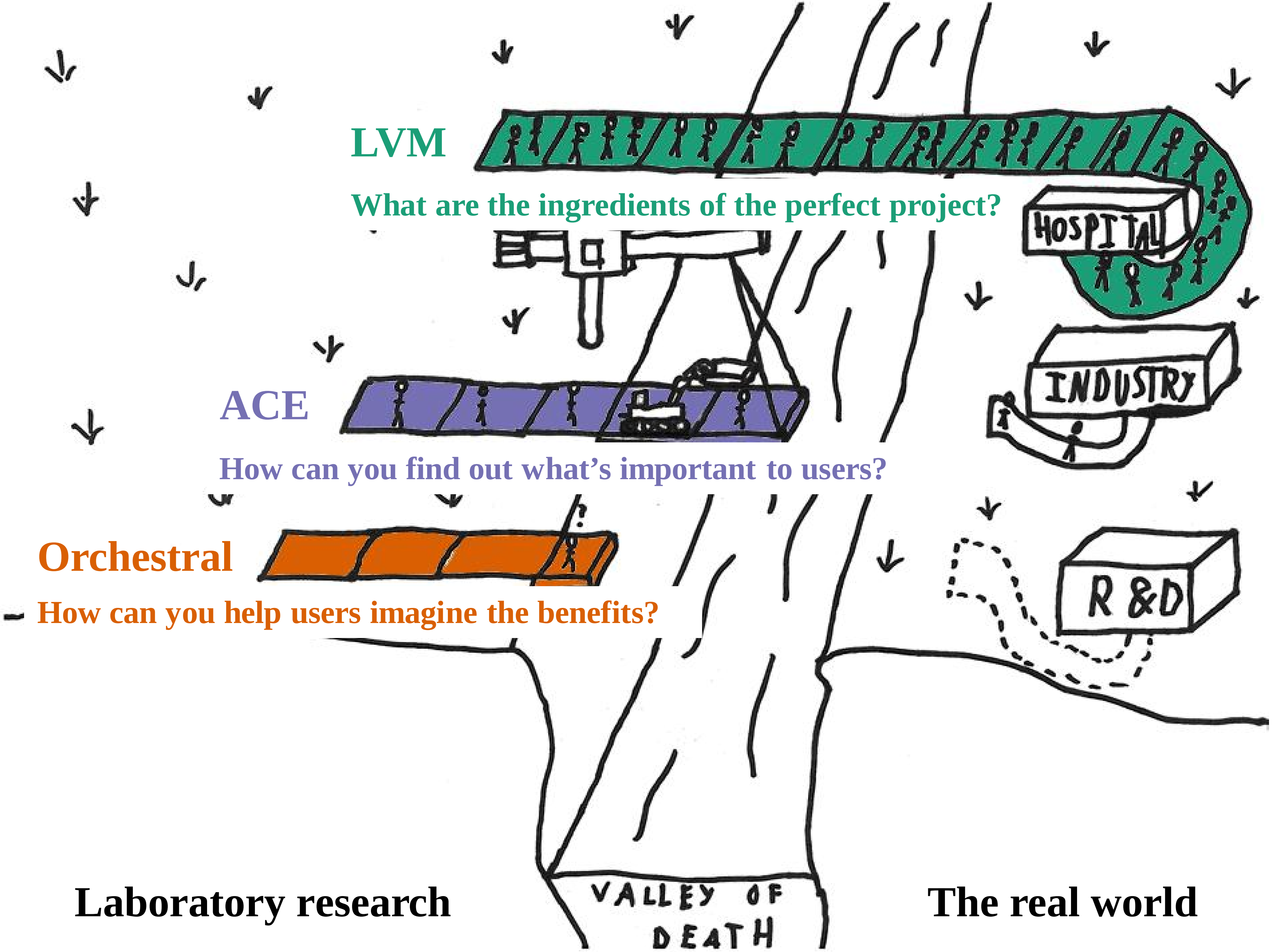
Visualizing health data – from fundamental research to successful applications

Professor Roy Ruddle

University of Leeds & Alan Turing Institute

https://raruddle.wordpress.com/



LVM

What are the ingredients of the perfect project?

ACE

How can you find out what's important to users?

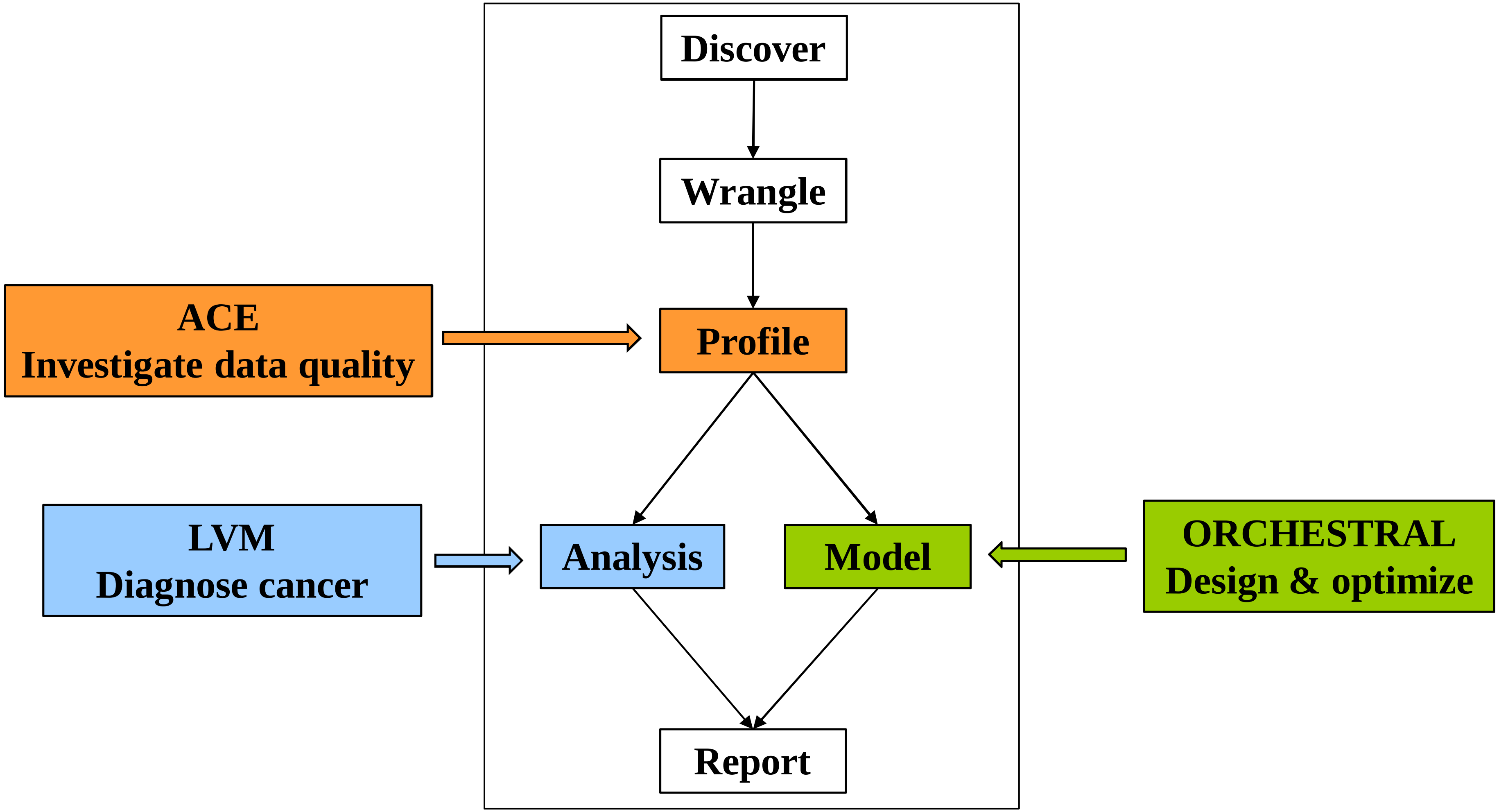
Orchestral

How can you help users imagine the benefits?

Laboratory research

The real world

Data science workflow[1]



[1] Alspaugh et al. (2018). *IEEE TVCG*.

ACE

A novel tool for investigating missing data

Big picture

- **Data wrangling & profiling take 50 – 80% of data scientists' time**
- **Many tools for investigating data quality**
  - **But they don't meet users' requirements**
- **Users lack of knowledge & rigour in data cleaning**[1]
- **Visualization methods for data quality**
  - **Limited research**[2]
  - **Unrealistic evaluation (toy datasets)**

[1] Visualizing the quality of data https://tinyurl.com/VizDataQuality

[2] Arbesser, et al. (2017). *IEEE TVCG*; Gotz, & Stavropoulos. (2014). *IEEE TVCG*; Gratzl, et al. (2013). *IEEE TVCG*; Gschwandtner, et al. (2014). *Proc. I-KNOW*; Kandel, et al. (2012)). *Proc. AVI*; Noselli, et al. (2017). *Proc. HEALTHINF*; Ruddle & Hall. (2019). *Proc. HEALTHINF*; Tennekes, et al. (2011). *Proc. NTTS*; Unwin, et al. 1996). *Comp. & Graph. Stat*; Xie, et al. (2006). *Proc. IEEE VAST*; Zhang, et al. (2014). Information Visualization

How can you find out users' requirements?

- **Tensions in applied research**
  - **Useful tool vs. novel research**
  - **Market research vs. requirements analysis**

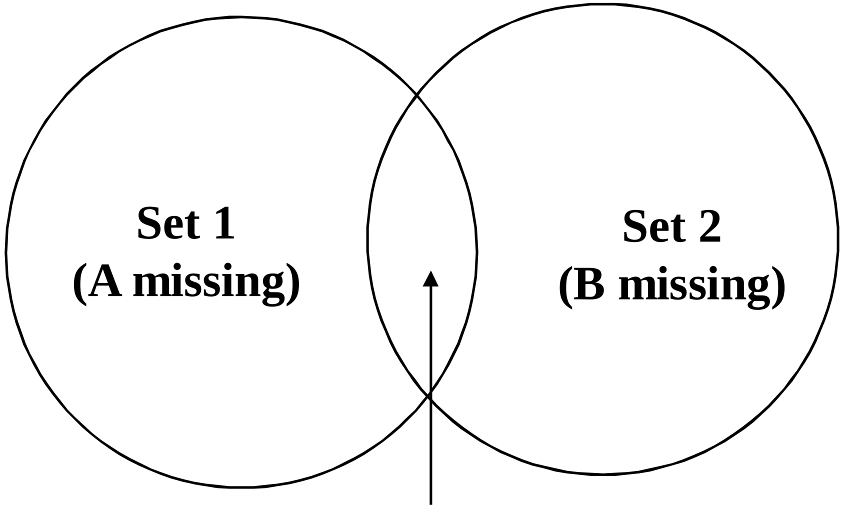
Three steps	Methods
Find out current situation	Questionnaire, interview, documentation, example data: • What analysis steps are involved? • What do you already do well? • What do you know?
Explore what's needed	Ask what would you like to do, but cannot do today? • What is hard or time-consuming (barriers & bottlenecks)? • What assumptions/simplifications are you forced to make? • Why don't current analysis tools solve these difficulties? • Let your self dream …
Check your understanding	Workshop • Encourage corrections & comments • Propose solutions (storyboard; throw-away prototype)

NHS Digital

- **Provides information, data and IT systems for National Health Service in England (£400 million)**
- **Current situation**
  - **Collect patient-level data from every NHS hospital**
    - **E.g., Admitted Patient Care (APC) data**
      - **500 fields and 20 million records/year**
  - **Mature data cleaning process, including**
    - **Business rules for data correction & validation**
    - **Threshold for missing values (only ≈8 fields)**
    - **Feedback to hospitals**
- **What's needed**
  - **Explore data quality patterns involving multiple fields**
  - **Exclude expected patterns, to reveal the unexpected**
  - **Develop new business rules**

Novel set visualization tool

Field A	Field B	Field C
101		M
102		F
		M
	99	F
	68	M



Set 1
(A missing)

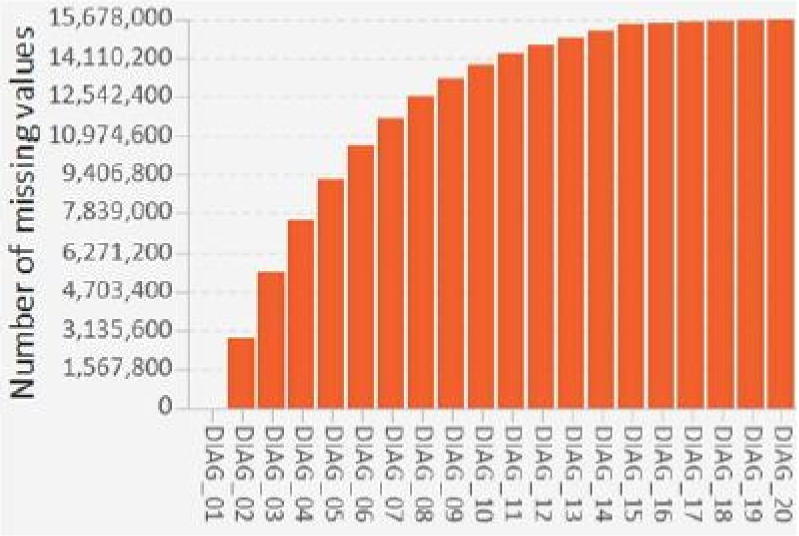
Set 2
(B missing)

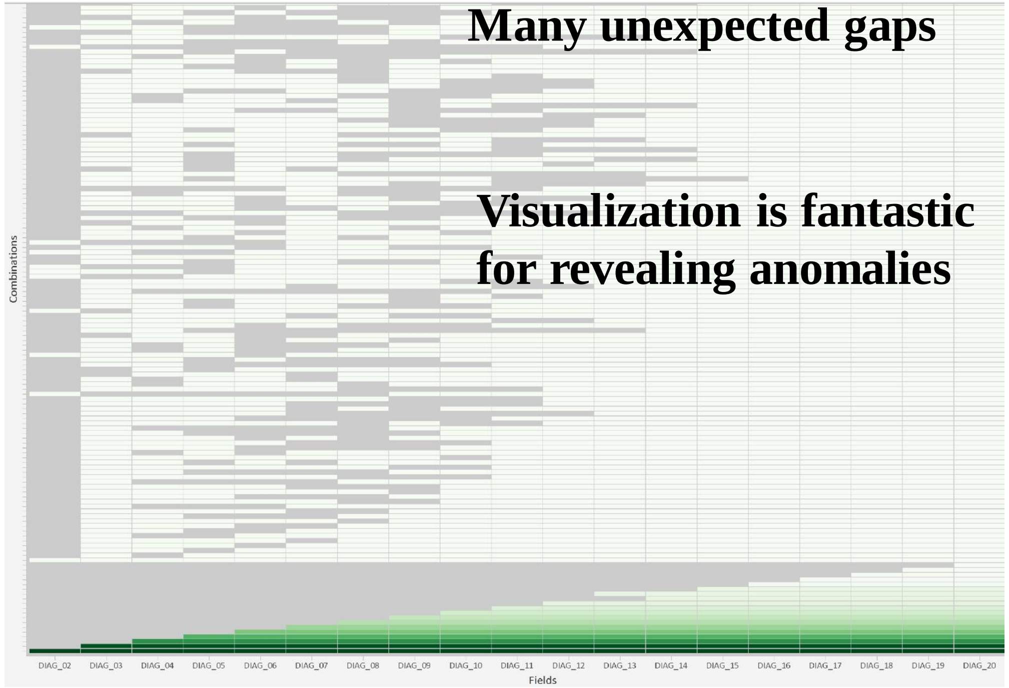
Set intersection
(A & B missing)

- **Scalable design**
  - **20 million records**
  - **500 fields**
  - **500,000 combinations of missing values**
- **Achieved using well-known techniques**
  - **Bar charts, heat maps and histograms**
  - **Reduce learning curve (avoid unnecessary novelty)**

Admitted Patient Care (APC) example

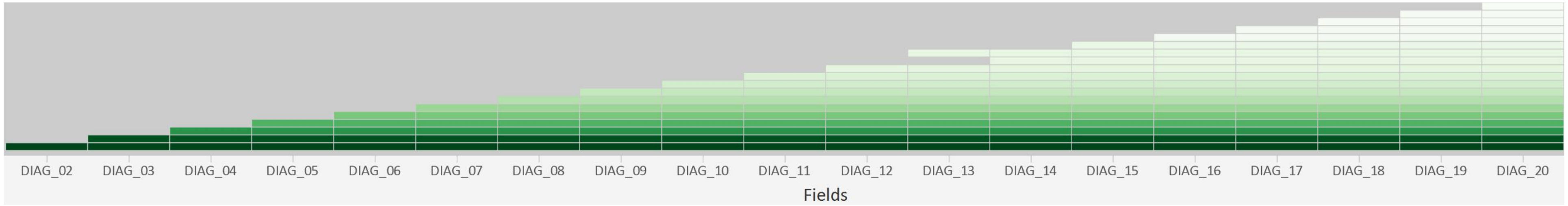
- **20 fields for diagnostic codes**
  - **Missing more often from DIAG_01 to DIAG_20**





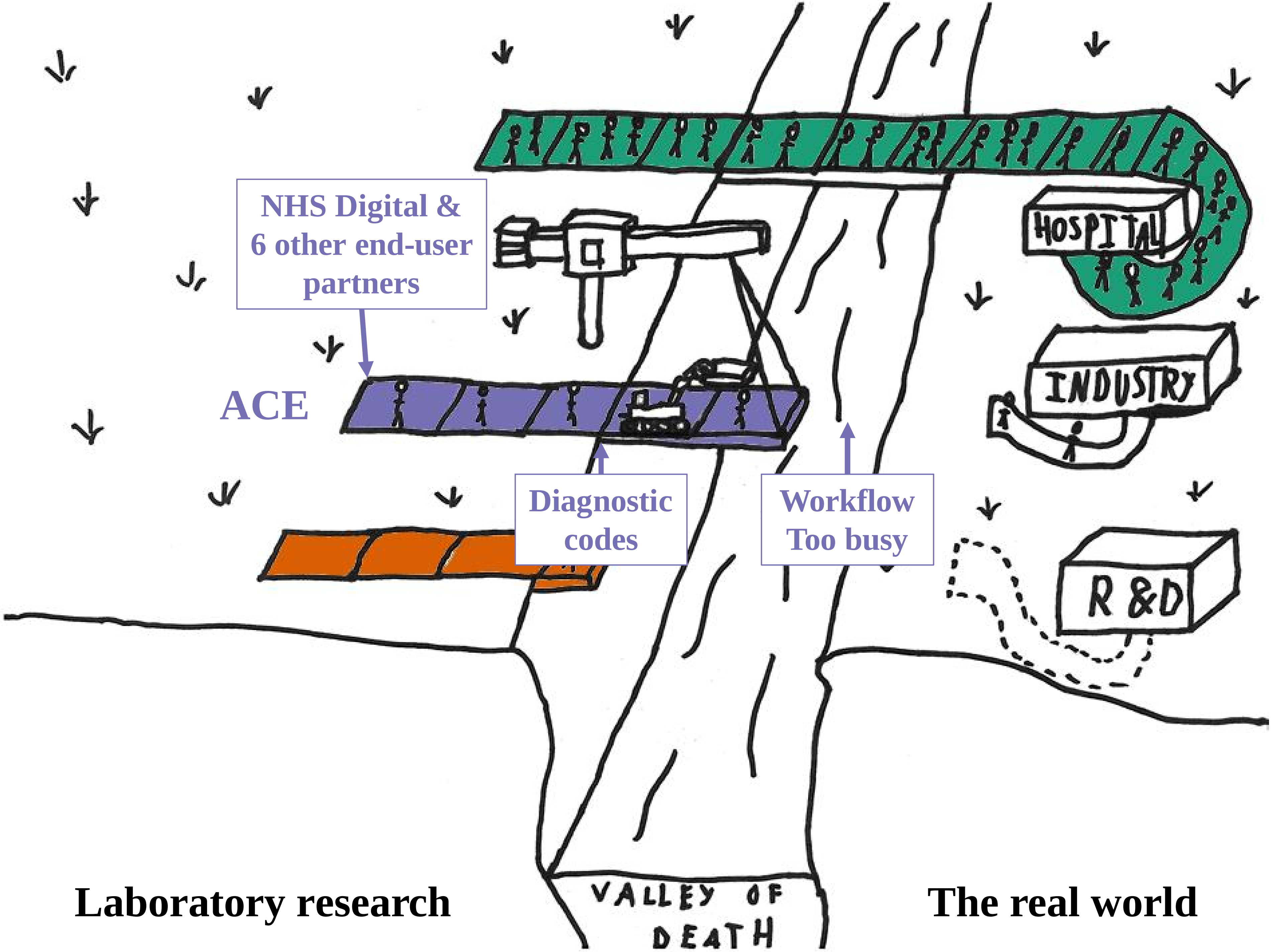
Many unexpected gaps

Visualization is fantastic for revealing anomalies



Actionable insights

- Widespread implications for data cleaning rules
- Gaps in diagnostic codes
  - Only 2000 records
  - 85+ % from one admission method in specific hospital
    - Improve data quality via established mechanism
- Gaps in operation codes
  - 2500 records
  - May affect NHS Payment by Results system for hospitals
- Millions of missing dates



NHS Digital & 6 other end-user partners

ACE

Diagnostic codes

Workflow Too busy

Laboratory research

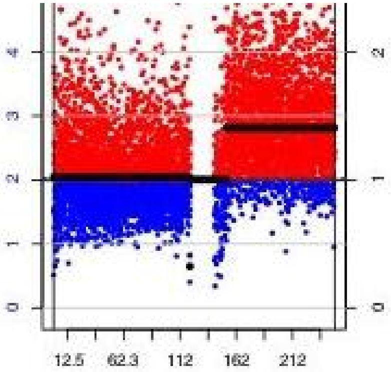
The real world

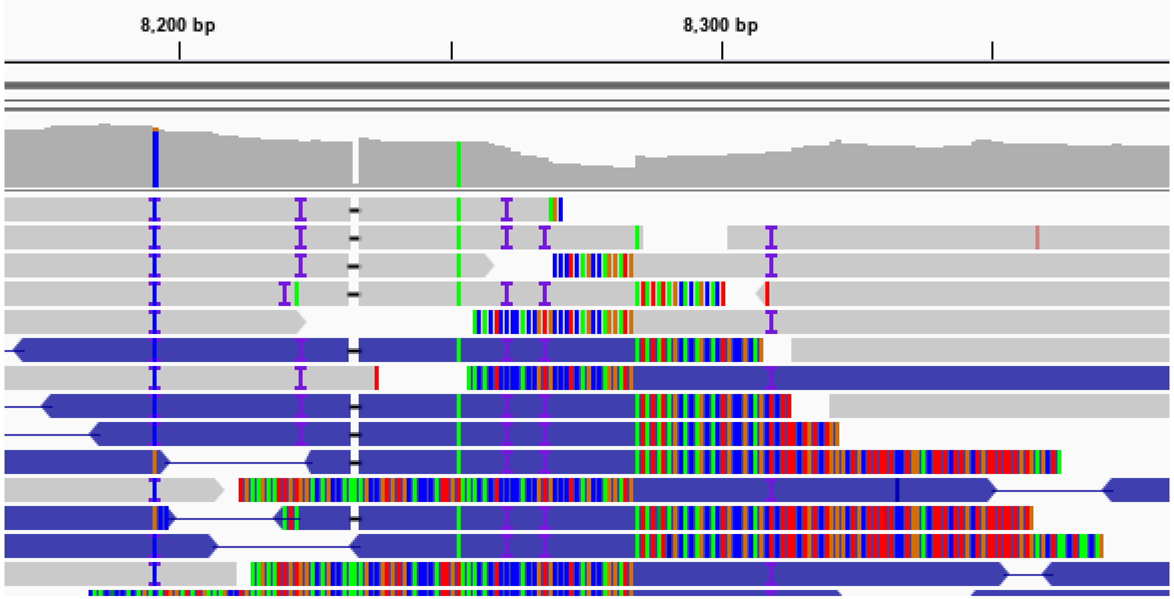
Orchestral

Visualizing genomics algorithms

Big picture - genomics

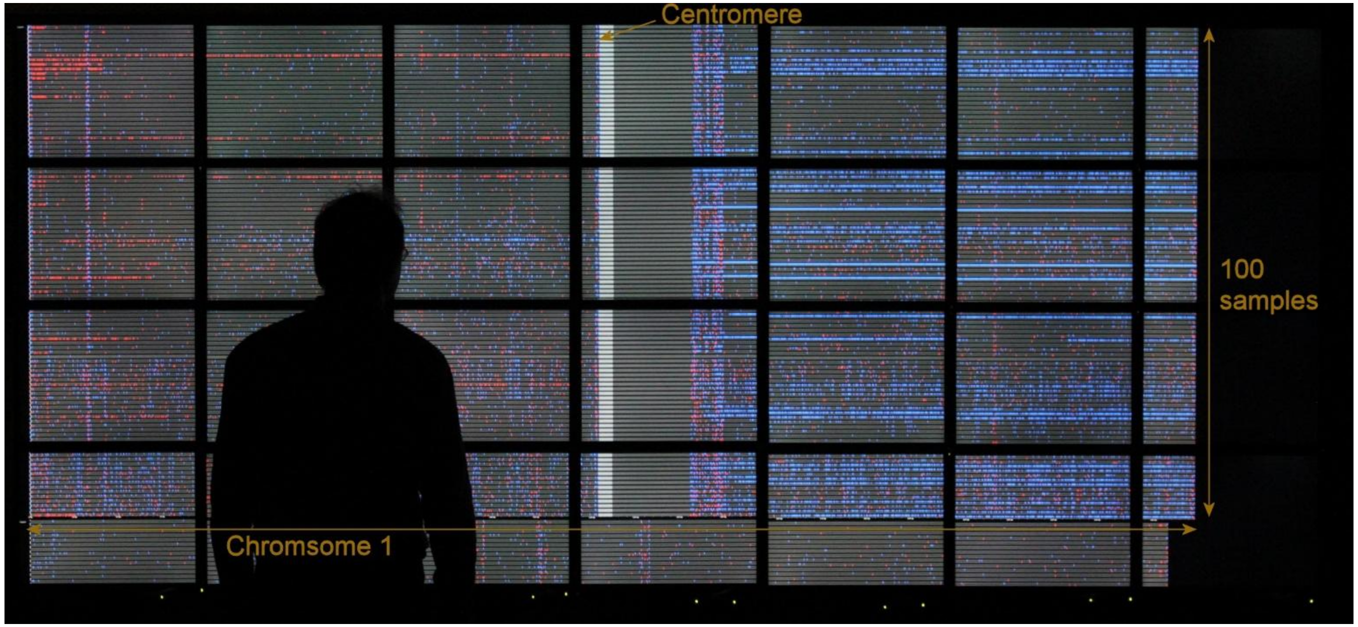
- Mature tools and pioneer in “big data”
- But unimaginative visualization
  - Massive over-plotting of data points
  - Have to pan thousands of times
- Example application: Breast cancer
  - Clear (subjective) differences between cohorts
  - Need to understand differences to design statistical test





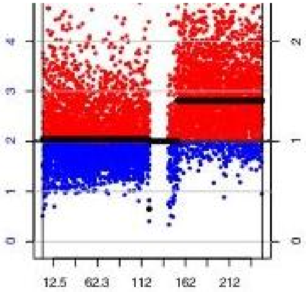
Hypothesis

- **Large high-resolution displays could transform scientists' ability to find patterns in genomic data**

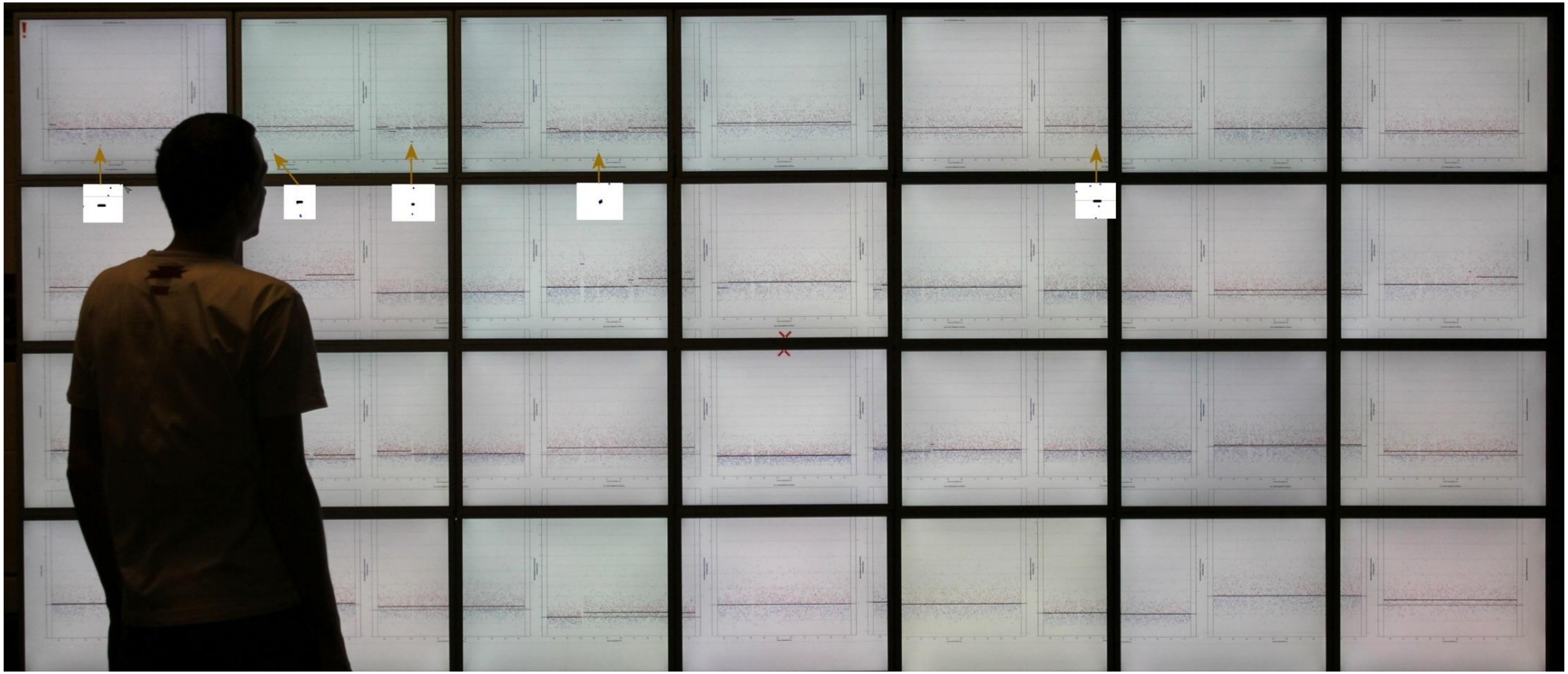


54 megapixel Powerwall (3 x 1.3 metres)

How can you help users imagine the benefits?

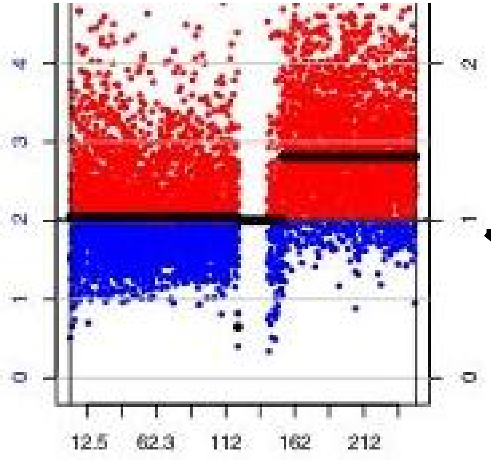


- Throw-away prototypes
  - Giant image (get user "pull")
  - Static visualization (spatial compression was too radical)
  - Interactive proof of concept

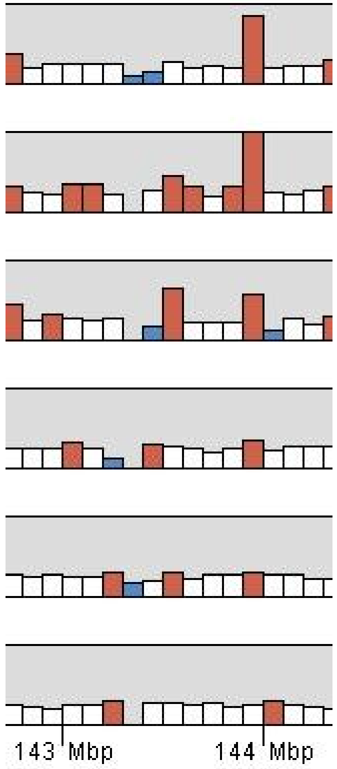


Current visualization vs. Orchestral

Current visualization

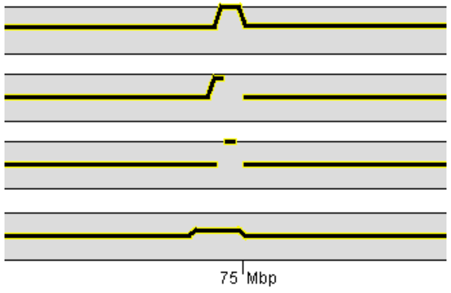


Copy number variation
(100 kilobase windows)



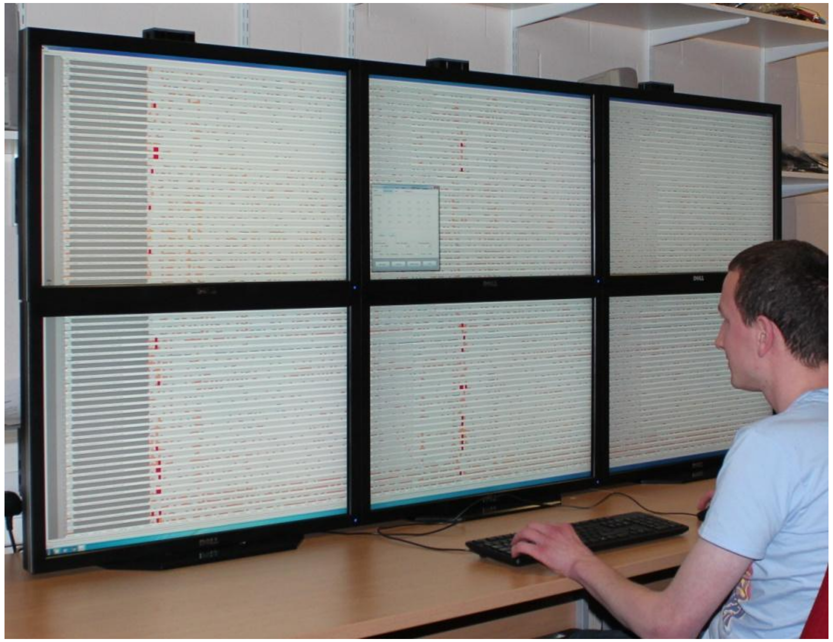
Segmented data
(noise removed for
statistical analysis)

Orchestral



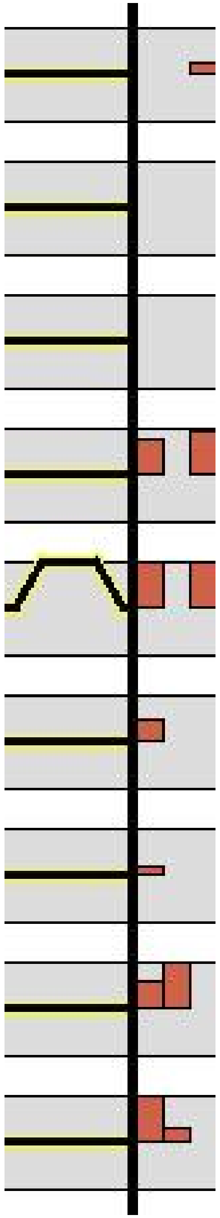
Open the black box by visualizing detail in context

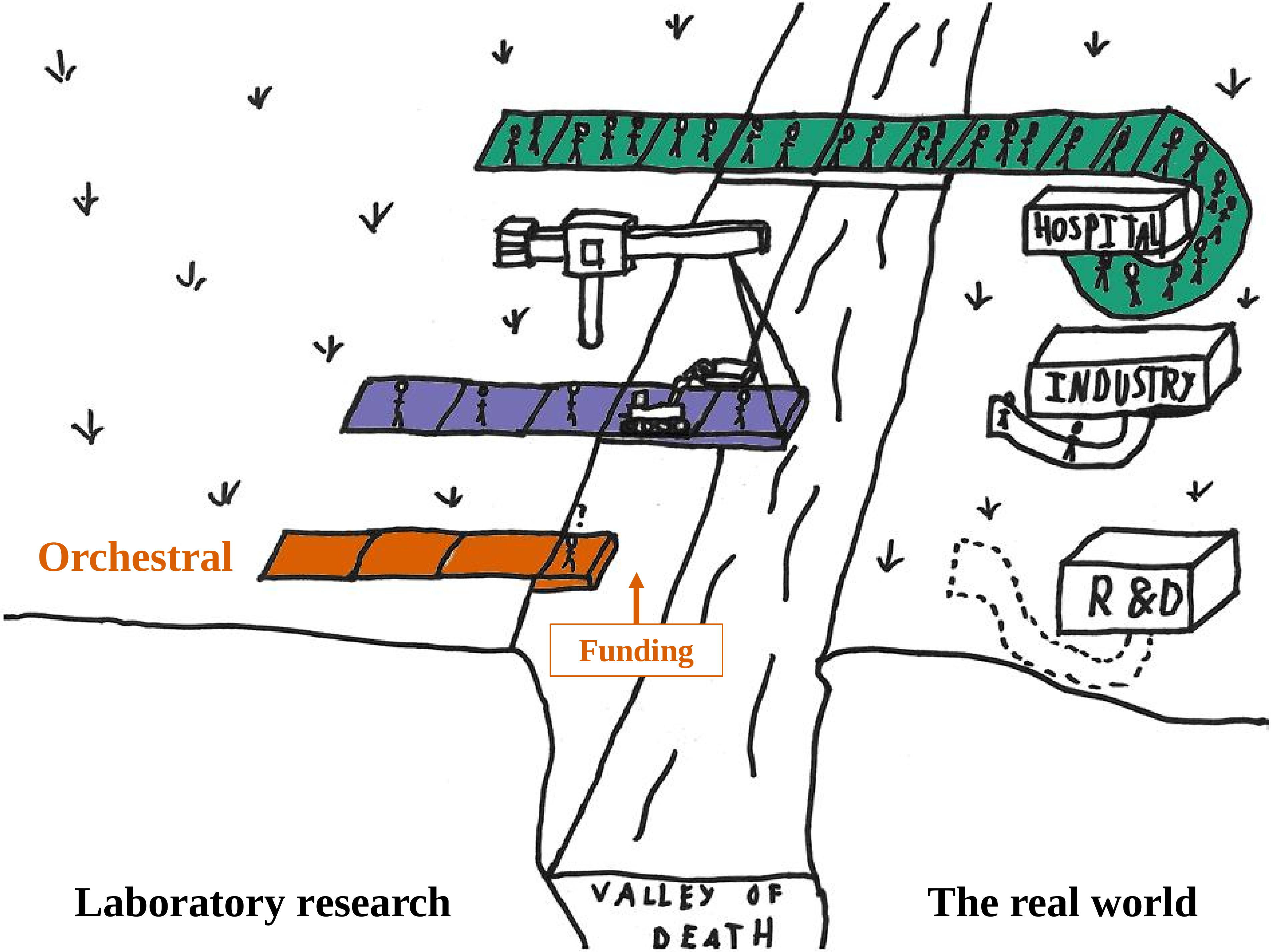
- **"Data looks abnormally similar, almost identical"**
  - **Processing error (incompatible steps)[1]**
- **Smoothing algorithm removes common feature**



24 megapixel workstation

[1] Ruddle et al. (2013). *Proc. Biovis.*





Orchestral

Funding

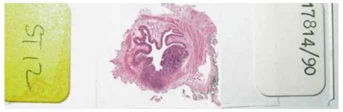
Laboratory research

The real world

Leeds Virtual Microscope

Diagnosing cancer from Amazon-sized images

What is pathology?



"Pathologists diagnose cancer by using a microscope to examine glass slides that contain thin sections of human tissue"

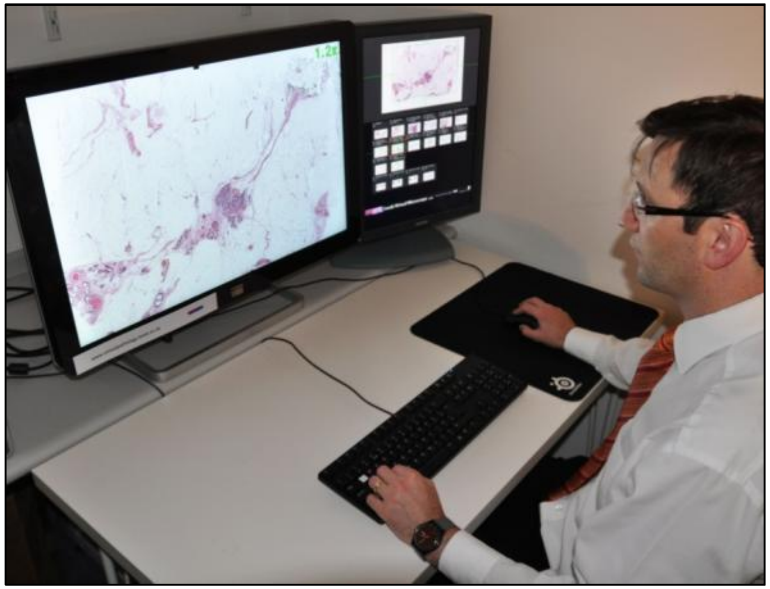
- **Large-scale operation**
  - **40 consultants (Leeds Teaching Hospitals)**
  - **150,000 slides/year, at 25 – 400× magnification**
- **The slides can be digitised for viewing on a computer**
  - **Advantageous for 2nd opinions, long-term survival and computer-assisted diagnosis**
  - **But it takes pathologist 60% longer to make a diagnosis[1]**
  - **Each slide is enormous**
    - **10 gigapixels (an "Amazon" of image data)**

- **http://www.youtube.com/watch?v=oZGkhKkDG5o**

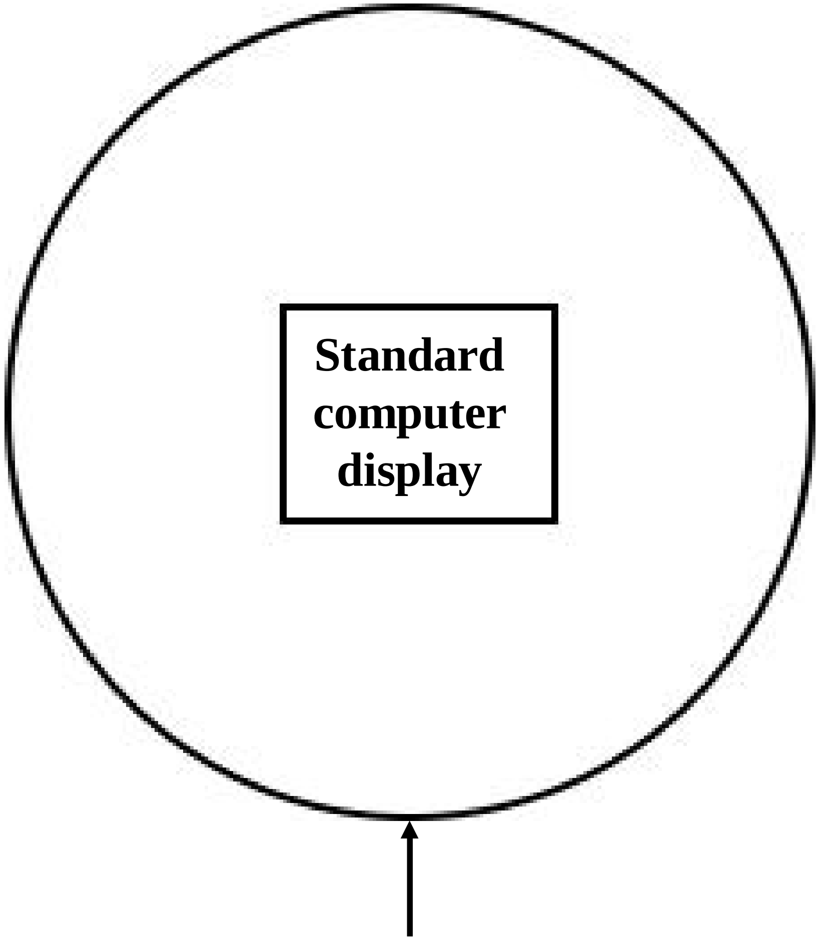
[1] Treanor & Quirke. *Pathological Society Glasgow, July 2007.*

Why is diagnosis 60% slower?

- Three reasons
  - Standard computer displays are too small
  - User interfaces of commercial products are inefficient
  - Doctors lack experience & training with digital slides



Standard displays – like looking through a keyhole

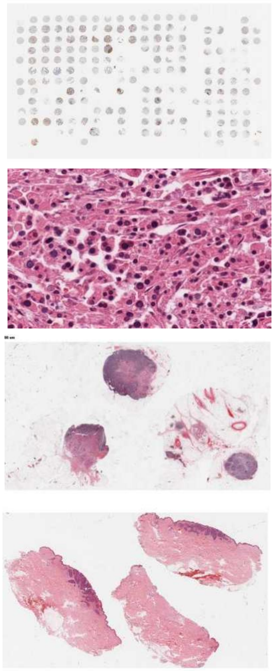


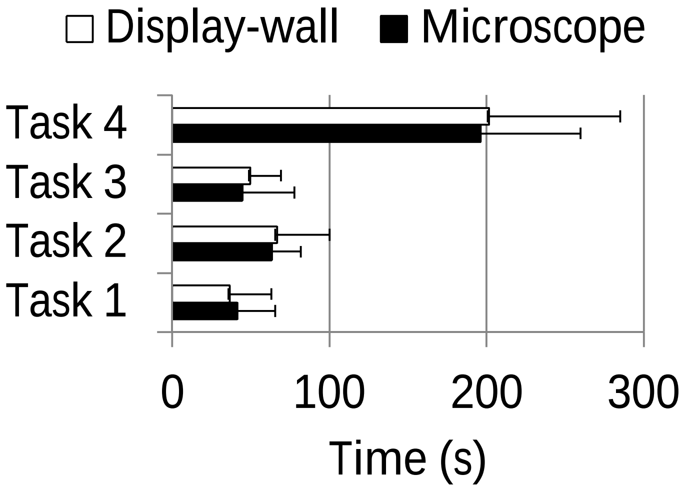
Microscope field of view (48°)

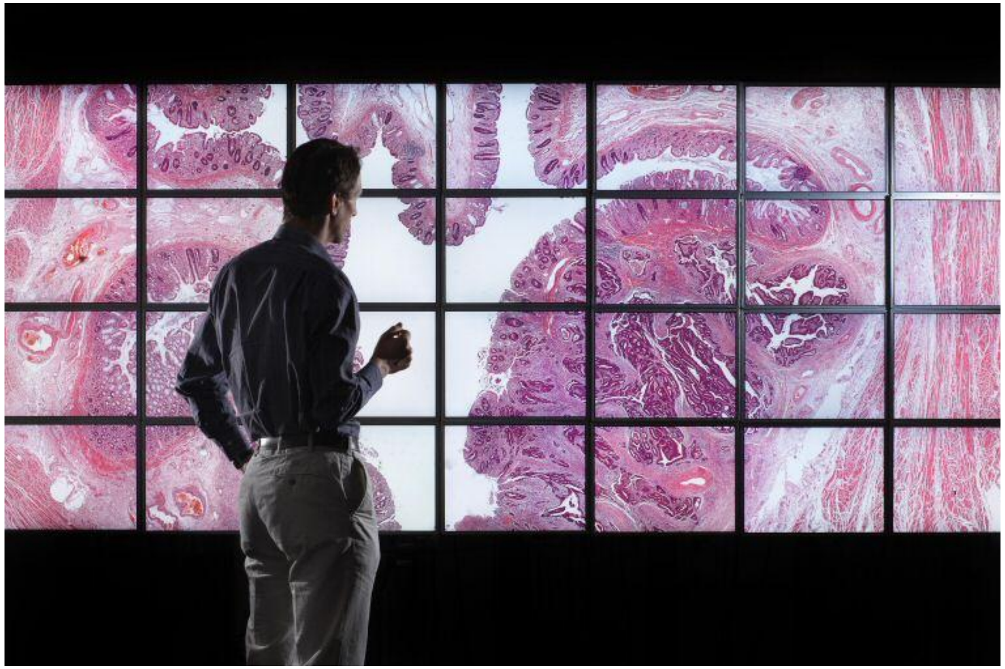
Solution: 54 megapixel Powerwall[1]

- 6× large field than a microscope, with
  - 3200 x 2400 pixel “thumbnail”
  - Gamepad user interface
- Microscope vs. Powerwall evaluation
  - 4 consultants & 4 trainees
  - Only a few minutes of training

Four tasks



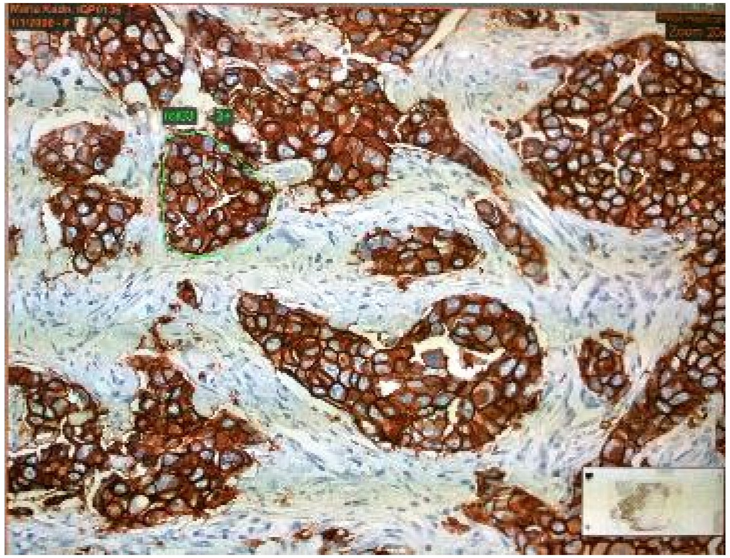




[1] Treanor et al. (2009). *Histopathology*.

Existing user interfaces – based on Photoshop/Google Maps

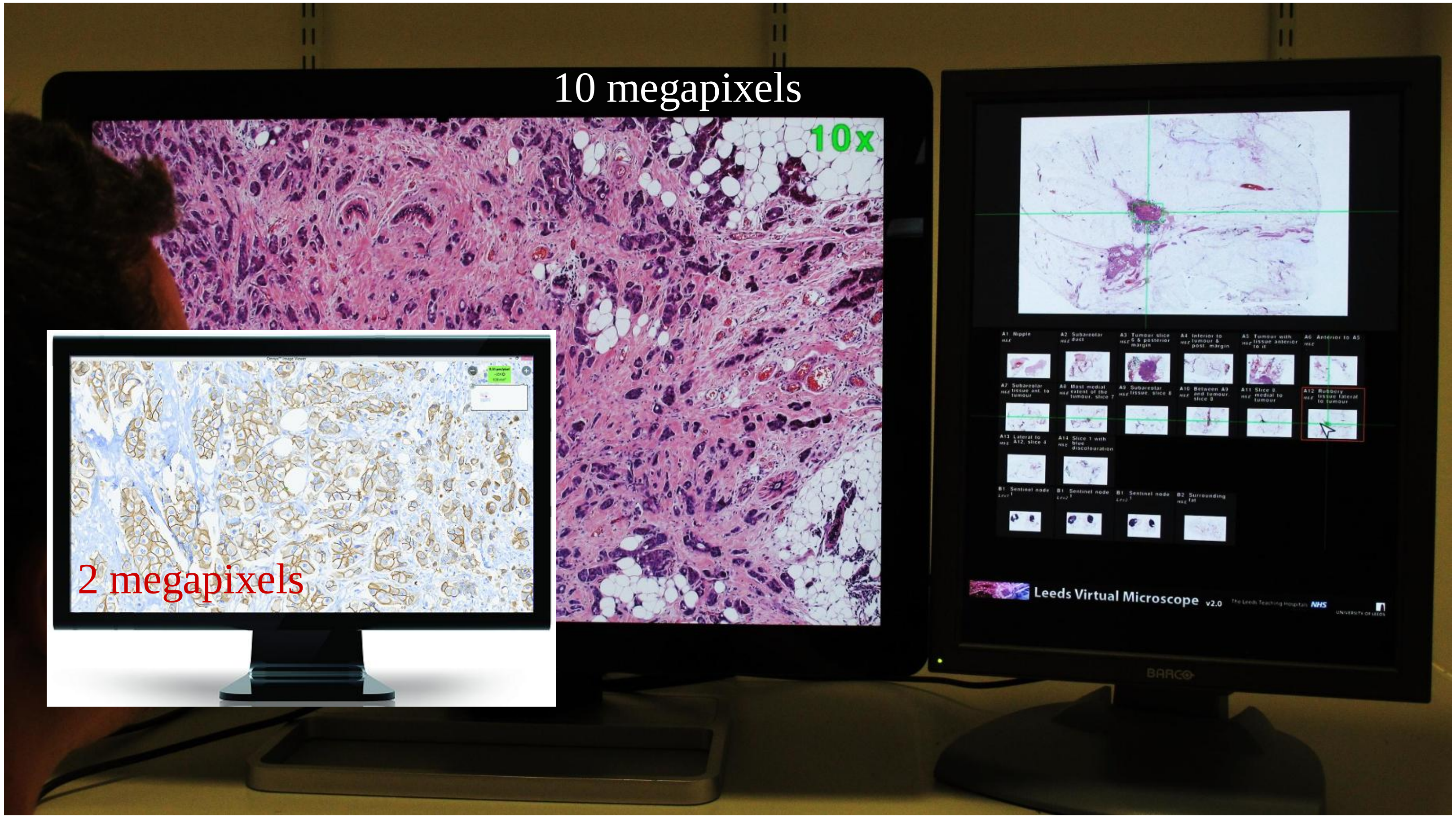
- Glass-sized thumbnail
- Real-time interaction
  - But thousands of panning movements
- Thmbnail scale difference
  - 1 : 1200 pathology
  - 1 : 30 Google Maps (established guidelines[1])



[1]Shneiderman (1998). *Designing the user interface.*

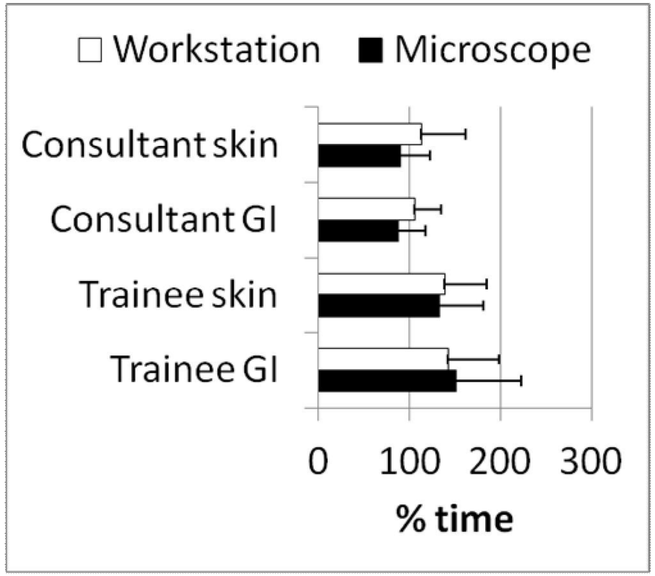
LVM solution

- **10 megapixel medical-grade display**
- **One third of space devoted to overviews**
- **Novel user interface**

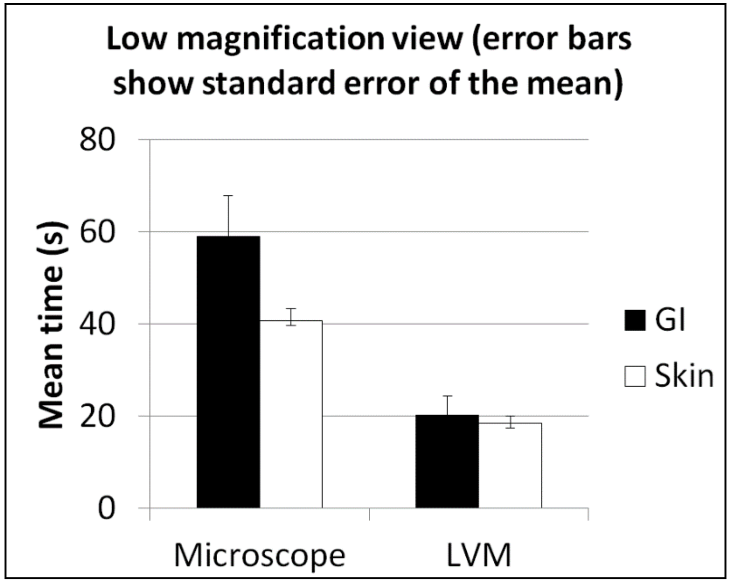

10 megapixels
10x
2 megapixels
A1 Nipple H&E
A2 Subareolar duct H&E
A3 Tumour slice 6 & posterior margin H&E
A4 Inferior to tumour & post. margin H&E
A5 Tumour with tissue anterior to it H&E
A6 Anterior to A5 H&E
A7 Subareolar tissue ant. to tumour H&E
A8 Most medial extent of the tumour, slice 7 H&E
A9 Subareolar tissue, slice 8 H&E
A10 Between A9 and tumour, slice 8 H&E
A11 Slice 8, medial to tumour H&E
A12 Rubbery tissue lateral to tumour H&E
A13 Lateral to A12, slice 4 H&E
A14 Slice 1 with blue discolouration H&E
B1 Sentinel node 1 Lev1
B1 Sentinel node 1 Lev2
B1 Sentinel node 1 Lev2
B2 Surrounding fat H&E
Leeds Virtual Microscope v2.0
BARCO

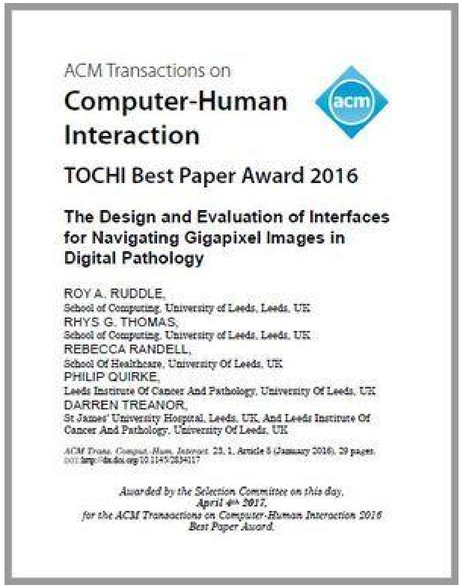
Evaluation: LVM vs. microscope

- **Controlled experiments**
  - **Real work (repeat diagnoses)**
  - **Participants were pathologists**
    - **0.5 – 28 years experience (microscope) vs. < 1 hour (LVM)**



Diagnosis time (Error bars = 1 SD)





ACM Transactions on

Computer-Human Interaction

TOCHI Best Paper Award 2016

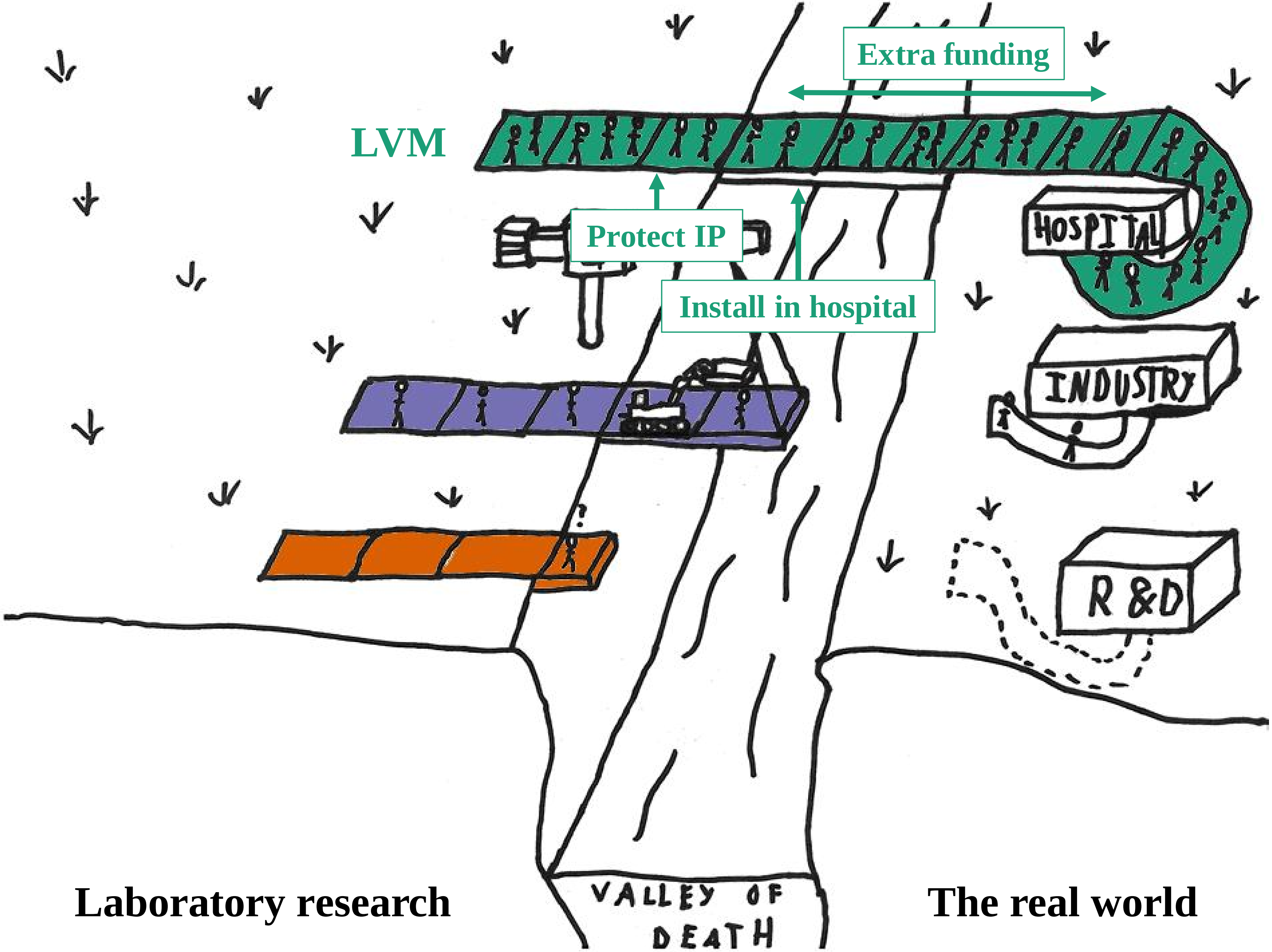
The Design and Evaluation of Interfaces for Navigating Gigapixel Images in Digital Pathology

ROY A. RUDDLE,
School of Computing, University of Leeds, Leeds, UK
RHYS G. THOMAS,
School of Computing, University of Leeds, Leeds, UK
REBECCA RANDELL,
School Of Healthcare, University Of Leeds, UK
PHILIP QUIRKE,
Leeds Institute Of Cancer And Pathology, University Of Leeds, UK
DARREN TREANOR,
St James' University Hospital, Leeds, UK, And Leeds Institute Of Cancer And Pathology, University Of Leeds, UK

ACM Trans. Comput.-Hum. Interact. 23, 1, Article 5 (January 2016), 29 pages.
DOI: http://dx.doi.org/10.1145/2834117

Awarded by the Selection Committee on this day,
April 4th 2017,
for the ACM Transactions on Computer-Human Interaction 2016
Best Paper Award.

Single-slide cases	Long (12-25 slide) cases	Meta-analysis
Randell et al. (2013). *Histopathology.*	Randell et al. (2014). *Human Pathology.*	Ruddle et al. (2016). *ACM ToCHI.*



Extra funding

LVM

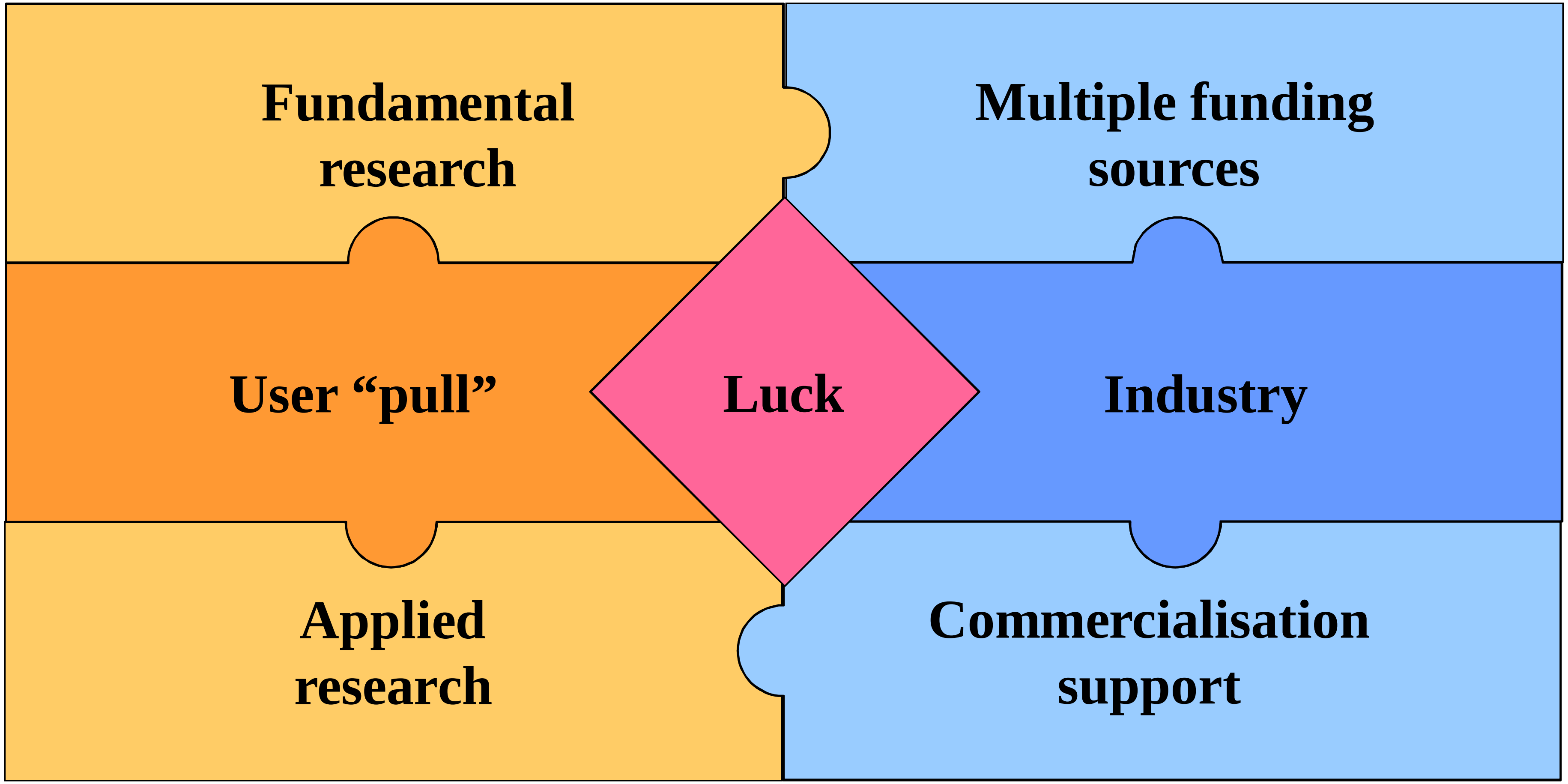
Protect IP

Install in hospital

Laboratory research

The real world

What are the ingredients of the perfect project?



Conclusions & future work

- **Generic**
  - **How can you find out what's important to users?**
  - **How can you help users imagine the benefits?**
  - **What are the ingredients of the perfect project?**
- **Visualization is fantastic for revealing anomalies**
  - **Unrealised Powerwall potential (4k is a commodity)**
- **Open the black box by visualizing detail in context**
  - **Visualization for pipeline design**[1]
  - **Visualization for machine learning (Vis4ML)**[2]
- **User interfaces**
  - **Minimise the cost ("… achieved something in minutes that would previously taken days"**[3]**)**

[1] von Landesberger et al. (2017). *IEEE TVCG*.
[2] Sacha et al. (2018). *IEEE TVCG*.
[3] Harrison et al. (2017). *IEEE TVCG*.

Acknowledgements

Project	Collaborators
LVM	Rhys Thomas, Rebecca Randell, Phil Quirke, Darren Treanor (LTHT)
Orchestral	Peter Sondergeld, Waleed Fateen, Phil Quirke Darren Treanor (LTHT)
QuantiCode	Georgios Aivaliotis, Mark Birkin, Justin Keen, Kevin Macnish, Alex Markham, Chris Megone, Muhammad Adnan, Richard Kavanagh, Anna Palczewska, Jan Palczewski. aql, Bradford Institute for Health Research, Consumerdata, Leeds City Council, Leeds Informatics Board, NHS Digital, Sainsbury's
Artwork	Sebastian Ruddle

EPSRC

NHS NIHR

The Alan Turing Institute

NHS Health Education Yorkshire and the Humber

Pathological Society Understanding Disease

yorkshire cancer research

Jisc

hefce

UNIVERSITY OF LEEDS

Questions?