# Final Results Report Template

**Visitor Name**:
Alan Dix

**Host Organisation**:
UNIPI

**Duration**:
5 weeks

**Title**: Humanising big data and computation

**Objectives**:

***A) Theoretical/Ethical*** – Theory and modelling relating to privacy, bias and explainability. "*The aim will be to document a semi-formal exploration of emerging human issues, in particular, using new algorithmic developments as case studies or inspirations to highlight or challenge more conceptual understanding.*"

***B) Technical*** – Developing interaction and infrastructure for working with large data in ways that can enhance user understanding, whilst protecting personal data sovereignty. "*The visit will allow further exploration of these issues and early steps to create better means for user interaction with big data that takes into account both the human capabilities of the data users and human rights of the data subjects*."

[*quotes* from application]

**Description of work carried out**:

The time flew by and where possible it seemed best to dedicate time to meeting people and discussions. This has led to a substantial number of working documents (described below) and planned future work more than things completed during the visit.

In addition to the core work at UniPisa and other Pisa institutions, the visit included wider interactions:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

- Delivering a virtual keynote at the FUSION 2024 conference in Malaysia on ‘patient interaction’
- A trip to Paris of the TANGO Horizon project all-hands meeting, which also included several colleagues from Pisa.
- Two days at the end of the trip visiting La Sapienza University

**Results and Conclusions**:

## Patient Users and Big Data

One of my first acts at Pisa was to give a virtual keynote at FUSION 2024, the annual conference of myHCI-UX, the ACM SIGCHI Kuala Lumber Chapter [Dxp24]. The abstract of this read:

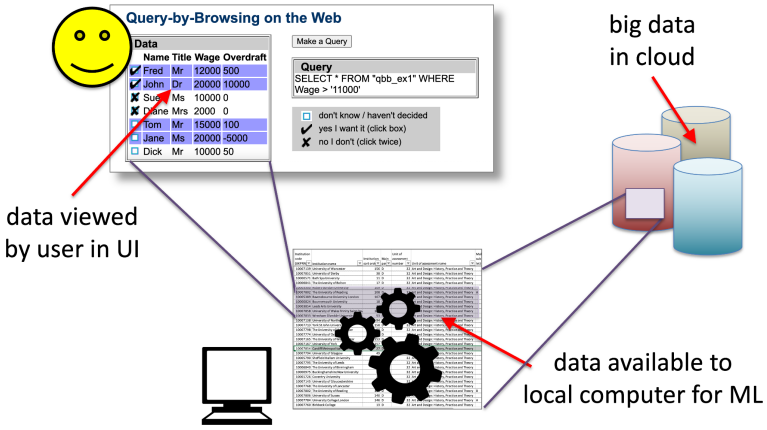
> “We live in a world of instant results and fleeting gratification. In HCI no less: the design principles for direct manipulation require immediate feedback and, in the case of graphical actions, sub-second responses. In addition, computers expect us to give them our undivided attention and continually seize it through notifications irrespective of what we are doing or how critical the interruption. This has a clear impact on well-being, and also on productivity as the myth of effective multi-tasking has been comprehensively dissolved. Furthermore, the need for instant and ever more complex computational response has major environmental impacts in terms both of the energy for computation itself and of the digital fast-fashion of discarded devices. Can we reimagine a world of patient human–computer interaction, where interfaces are designed to enable and encourage less feverish use of computational resources and more thoughtful engagement.”

At first this seemed to be unrelated to the work of the visit, except in the broadest terms of human interaction, and it was only when delivering it, that I realised that the central section on ‘patient users’ was addressing precisely the issues of “comprehensible human interaction that is example based, whilst dealing with datasets that are too large for real-time processing” which were raised under the ‘Technical’ strand of my application. However,in addition it had become clear that this was important for environmental sustainability as well as ease of interaction.

I repeated the talk in CNR Pisa whilst visiting Fabio Paterno, Carmen Santori and their group. In particular, some of the areas connect to their work on end user development.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871042

Detailed notes of the talks are under preparation and the slides can be found at:
https://alandix.com/academic/talks/FUSION2024/




*dealing with different subsets of data*
*for faster interaction and lower carbon footprint*

## Privacy

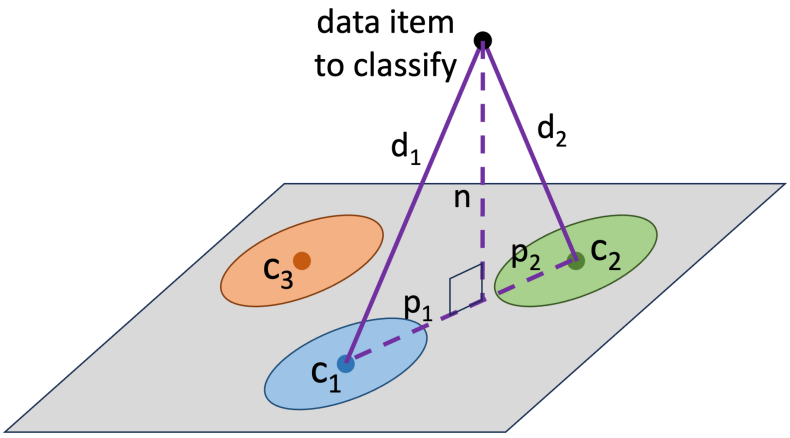
As noted in my application, I first wrote about issues of privacy from a HCI perspective in 1990 [Dx90]. This was a core topic in discussions with Anna Monreale and Francesca Naretto, not least because Francesca's PhD Thesis had identified conflicts between standard techniques used for XAI (eXplainable AI). One outcome of this is that I wrote up notes during my visit of talks I'd given some years previously on "three myths of privacy" [Dx19, WD1].

Rather like the way statistical queries can be used to compromise traditional databases [dJ83], Francesca's work showed that explanations with certainty factors could allow an adversary to home in on the training data [NM22]. This added to the examples of potential conflicts, in this case between explainability (important for bias) and privacy. Francesca ameliorated these problems by generating surrogate data to obfuscate the training data. This surrogate data has to be like the real data (so as to maintain the distribution of the model), but different enough that it does not 'average' back to the training data.

It became clear that the 'cluster-centroid latent space' (CCLS), might be a useful tool for this. While I'd posited the utility of this latent space in previous talks and personal communications, I had never written a coherent account, so a first step was

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

to do this [WD2]. We plan simulation techniques to create general validation, but have already the sketch of an asymptotic proof that, if the CCLS is not favoured, nearest neighbour approaches asymptotically tend towards random allocation of samples. It is hoped that adding some weighting towards creating surrogates close within the CCLS, but more widespread in orthogonal dimensions will help the obfuscation process.




*Cluster-centroid latent space*

## Modelling reliance and trust

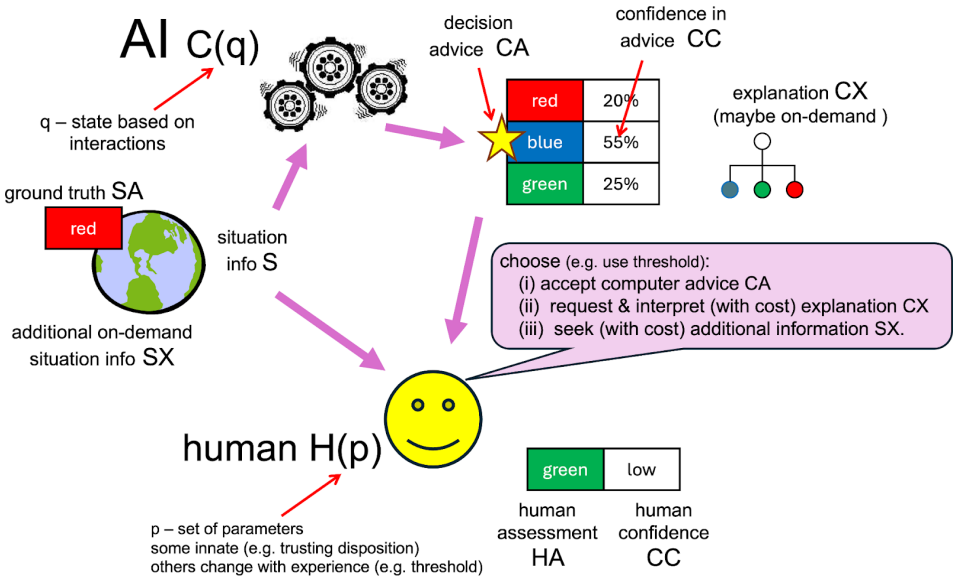
The visit picked up earlier online discussions about appropriate reliance and trust in automated systems, particularly working with Daria Mikhaylova, Tommaso Turchi and Alessio Malizia (all UniPisa), Andrea Beretta (CNR) and colleagues in the UK. We discussed models at different levels of formality:

1. *Conceptual models* that describe the vocabulary of a domain and the way that these interact with one another.
2. *Formal models* that are largely definitional – that is they are expressed formally, but the main aim is not to be able to mathematically reason with them, but to expose interesting issues and clarify meanings.
3. *Computational models* that are amenable to simulation.
4. *Mathematical models* that are suitable for reasoning with, either logically or through closed-form numerical analysis.

Discussions with Andrea more around (1) and those with Daria more in the spectrum (2–4), with a special focus on simulation (3).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

Workin this area is continuing as part of the TANGO Horizon project, and also initial ideas on computational simulation in [WD3].




*Modelling a single decision where the human uses AI assistance*

## System and user explanations and utterances

Interaction with machine learning systems can be viewed simplistically as of the form:

data == [ training ] ⇒ model

model == [ situation ] ⇒ advice/prediction

In this view the utterances of the user can be seen in terms of providing training data and the utterances of the system as concrete advice or predictions about a particular state of the world. However, the reality is far more rich and complex.

From the system side predictions are increasingly required to include an explanation (XAI), which may be phrased in terms of the *model*, for example, creating simplified local rule, or in terms of *data*, for example a counterfactual or close positive case. Communicating in terms of data has the advantage that it is a common language, whereas the computational space in, say, the layers of a deep neural network, may be very different from the concepts of the human. Of course, there may be privacy implications of data focused interactions.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

On the other hand the user often makes 'utterances' at the data level, including providing new examples (as united earlier), but also labelling training data. They may also be asked for relevance feedback on specific predictions/advice.

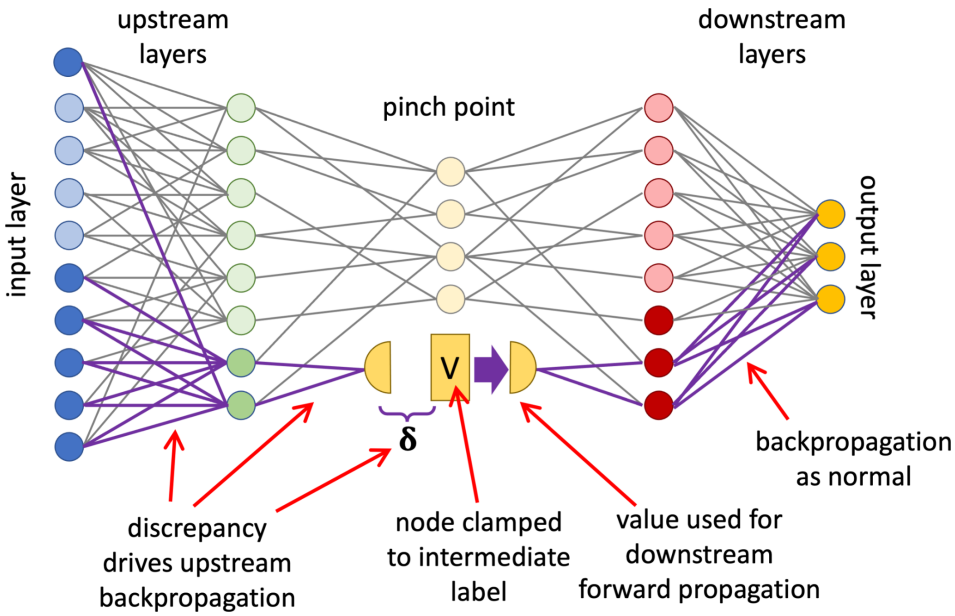
This much is not radically new, but we are working to clarify this rich space of interaction. In some cases this is using human–human and human–thing parallels while also being constantly aware that computation is neither human not a passive object. In addition, the fact that humans and AI have different roles, can act as an inspiration.

Roberto Pellungrini (Scuola Normale Superiore di Pisa) has been using real data counterfactuals to help explain decisions. We talked particularly about how, when two data items A and B are given different classifications, a third example C at different points within or outside the 'line' A–B can help either show why two apparently close items are given the same or different classifications. We also discussed whether the cluster-centroid latent space might help here also. Crucially also, good 'in between' examples are very powerful at helping to clarify decision boundaries (human or machine).

Discussions during the Paris TANGO meeting with Tomasso Turchi (UniPisa), Andrea Beretta (CNR Pisa) and others raised the issue of human feedback to AI systems. As noted this is often limited to explicitly or implicitly creating and/or labelling training examples, or giving some sort of relevance feedback. While the AI system is asked to explain decisions to the user, the reverse does not happen. In some forms of interactive learning studied at Pisa, the user may be challenged if a new example seems to contradict earlier training, but can only respond by withdrawing, relabelling or confirming the example. We imagine of the user could say (maybe in a formalised fashion) "this is an example of a strong researcher candidate because they have a PhD and lots of publications" the explanation, would then both offer a partial local rule (to constrain/guide ML) and a vocabulary for the system to use in its future explanations. It is easy to see how this could be used to create weakly constrained rules using genetic algorithms, and we are beginning to explore how this might be used in decision trees and even neural networks. We have started to draft an initial ideas paper for a workshop at IUI next March, and the work will also feed into a larger design-space analysis paper.

See working documents [WD4], [WD5] and [WD6] for further notes in this area.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

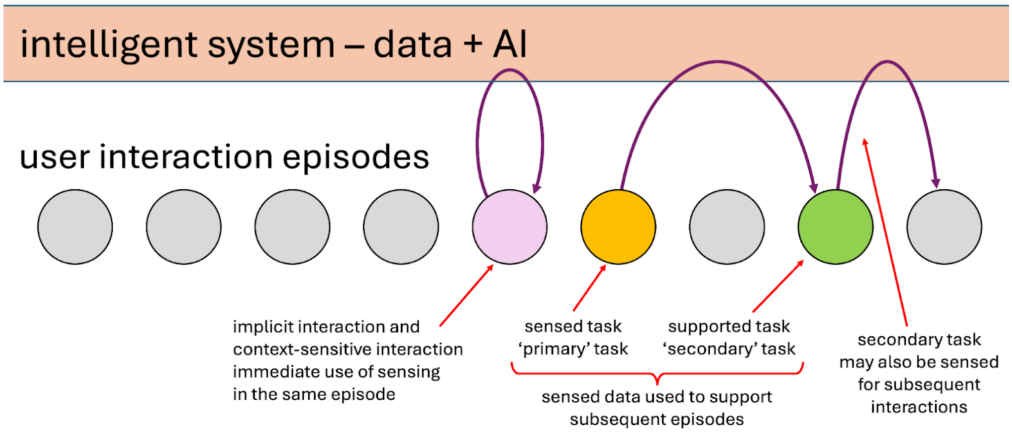



*Clamping concept nodes in a deep neural network*

## Evaluation of sensor-based and data mediated interactions

During my two day visit to Rome I picked up work with Emanuele Panizzi's group on the evaluation of sensor-driven interactions, where implicitly gathered sensor-data is used to update data or models that then improve subsequent interactions. Evaluation methods building on primary/secondary (sensed/supported) task distinction had already been submitted to IUI and will be included in my forthcoming book on "AI for Human–Computer Interaction". In this visit we focused on techniques to validate the evaluation methods.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042




*primary/sensed and secondary/supported tasks mediated by data*

**Future Work**:

Time during the visit was always dedicated to face-to-ace meetings where possible as this was the precious resource. This has created an open to-do list of research avenues opened! Indeed, most of the above sections have at least one open issue or part-written paper.

Much of this is collaborative, but I also have a few more personal targets, that I hope to complete over coming months.

- Keynote at HCI-E 2024 conference in Istanbul at the end of November, which will draw on some of the new work started during the visit.
- Complete the detailed notes for the Patient Interaction Fusion 2024 keynote and CNR talk, and then see how these might form the basis of a journal article.
- Investigate ways that Query-by-Browsing can be modified/extended to demonstrate some of the more conceptual ideas, including user 'explanations' [WD6].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042


**Link to Blog/video**:

– – see draft blog – –

**References**:

## Working documents

[WD1] Being Secure – some myths of privacy and personal security
[WD2] Interesting Latent Spaces
[WD3] Models of Trust and Reliance
[WD4] Two way communication and local human explanations

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871042

[WD5] A third way – XAI betwixt symbolic and sub-symbolic
[WD6] QbB as informal case study / worked example

## Other references

[Dx90] A. J. Dix (1990). Information processing, context and privacy. Human-Computer Interaction - INTERACT'90, Ed. D. G. D. Diaper G. Cockton & B. Shakel. North-Holland. pp. 15-20. https://alandix.com/academic/papers/int90/

[Dx19] A. Dix. (2019). Being secure: some myths of privacy and personal security. Talk at Cardiff University, 6th February 2019. https://alandix.com/academic/talks/being-secure-2019/

[Dxp24] Alan Dix (2024). Patient Interaction – for well-being, productivity and sustainability. FUSION 2024, Kuala Lumpur, Malaysia, 28 Sept. 2024. https://www.alandix.com/academic/talks/ FUSION2024/

[dJ83] Wiebren de Jonge, ‘‘Compromising statistical databases responding to queries about means’’, ACM Trans. On Database Systems 8(1), pp. 60-80, ACM (March 1983). https://doi.org/10.1145/319830.319834

[NM22] Naretto, F., Monreale, A., & Giannotti, F. (2022). Evaluating the privacy exposure of interpretable global explainers. In 2022 IEEE 4th International Conference on Cognitive Machine Intelligence (CogMI) (pp. 13-19). IEEE.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 871042