A quick note about the reading list

Available now on course materials page: http://www.cl.cam.ac.uk/teaching/1718/R230/IWML-reading-list.pdf

Don't try to read all of it!

"Starred" entries are particularly good for one or more of the following reasons:
- Influential
- Well-executed research
- Interesting/unique angle

Read at least the abstracts of all of the starred entries.

Use as a basis for your own research question/study design.

IWML 2018 Reading List

User research

Solomon, J. (2016). Heterogeneity in Customization of Recommender Systems By Users with Homogenous Preferences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 4166–4170). New York, New York, USA: ACM Press. http://doi.org/10.1145/2858036.2858513

Daee, P., Peltola, T., Vehtari, A., & Kaski, S. (2017). User Modelling for Avoiding Overfitting in Interactive Knowledge Elicitation for Prediction, 1–9. Retrieved from http://arxiv.org/abs/1710.04881

Fothergill, S., Mentis, H., Kohli, P., & Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12* (p. 1737). New York, New York, USA: ACM Press. http://doi.org/10.1145/2207676.2208303

★ Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (p. 31). New York, New York, USA: ACM Press. http://doi.org/10.1145/1240624.1240630

Christel, M. G. (2006). Evaluation and user studies with respect to video summarization and browsing. In E. Y. Chang, A. Hanjalic, & N. Sebe (Eds.), (p. 60730M–60730M–15). http://doi.org/10.1117/12.642841

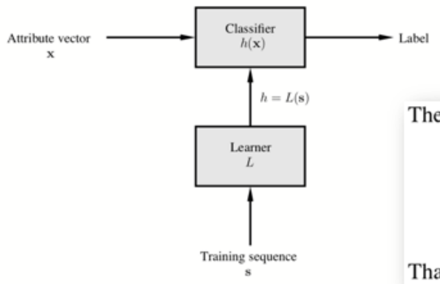
LABELLING

Interaction with Machine Learning
Cambridge MPhil ACS 2017-2018

What’s the big deal?

Supervised learning: a quick reminder

We don’t want to design h explicitly.

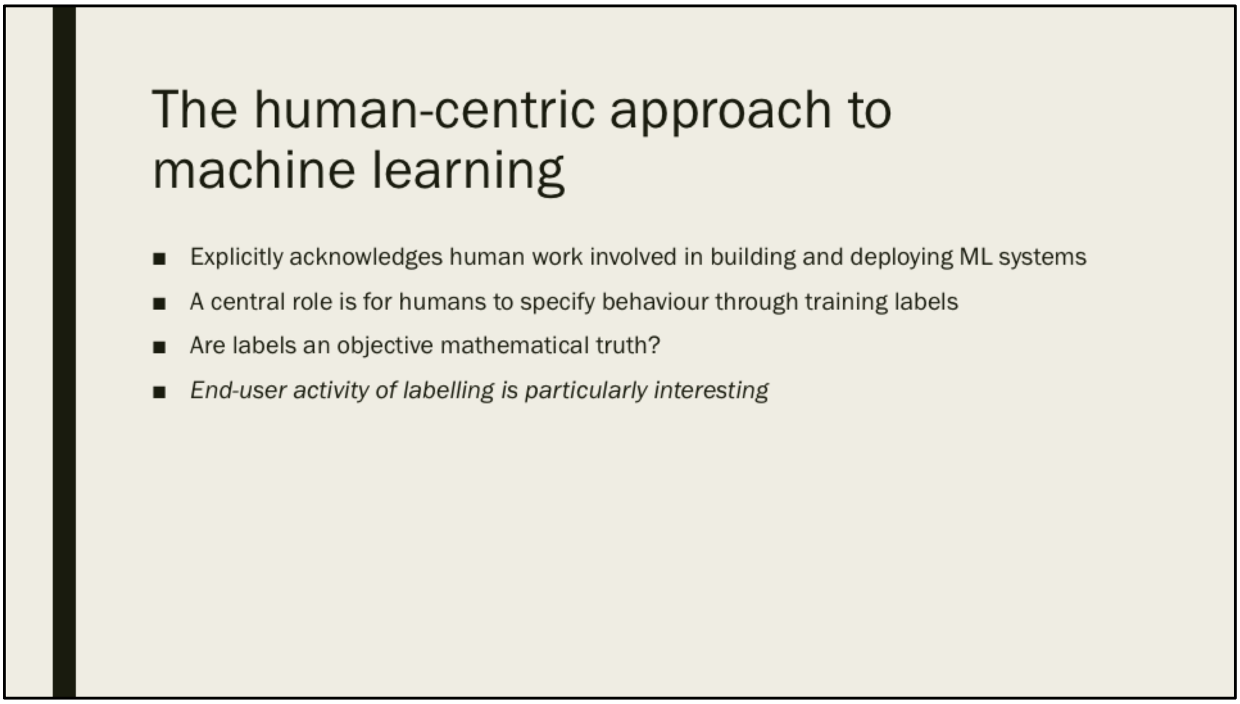


So we use a *learner* L to infer it on the basis of a sequence $\mathbf{s}$ of *training examples*.

The *training sequence* $\mathbf{s}$ is a sequence of m *labelled examples*.

$$\mathbf{s} = \begin{pmatrix} (\mathbf{x}_1, y_1) \\ (\mathbf{x}_2, y_2) \\ \vdots \\ (\mathbf{x}_m, y_m) \end{pmatrix}$$

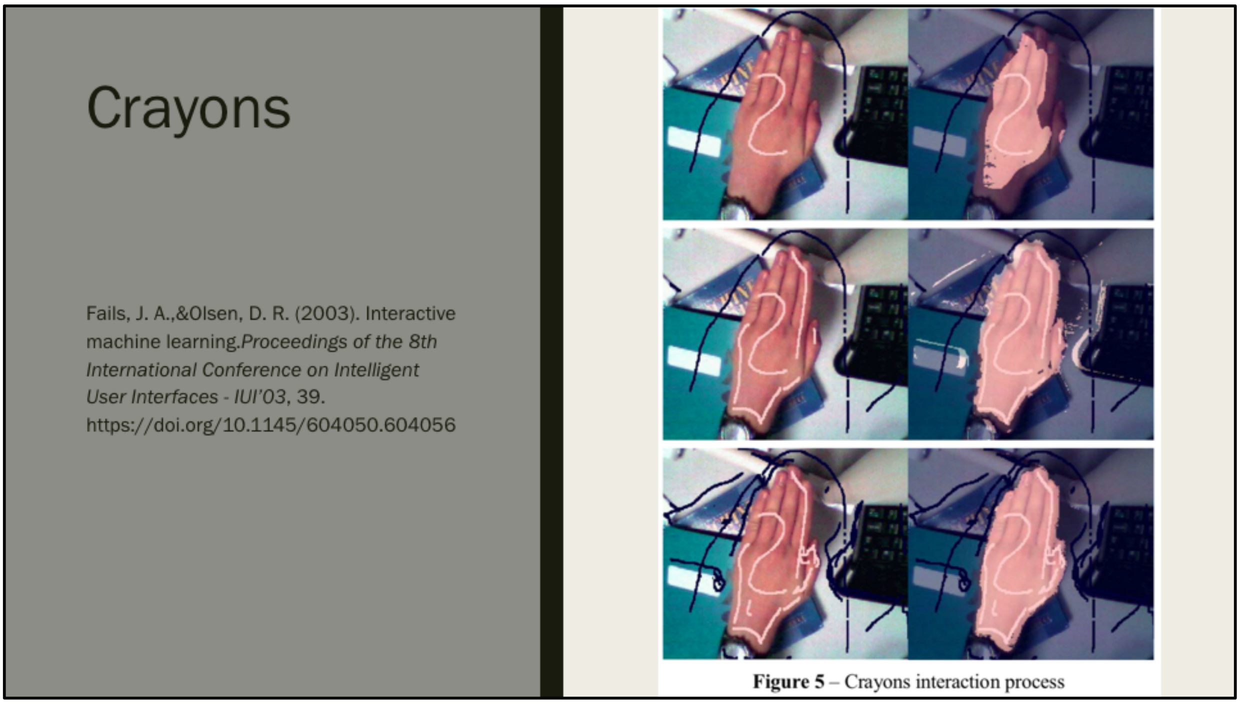
That is, examples of attribute vectors $\mathbf{x}$ with their correct label attached.



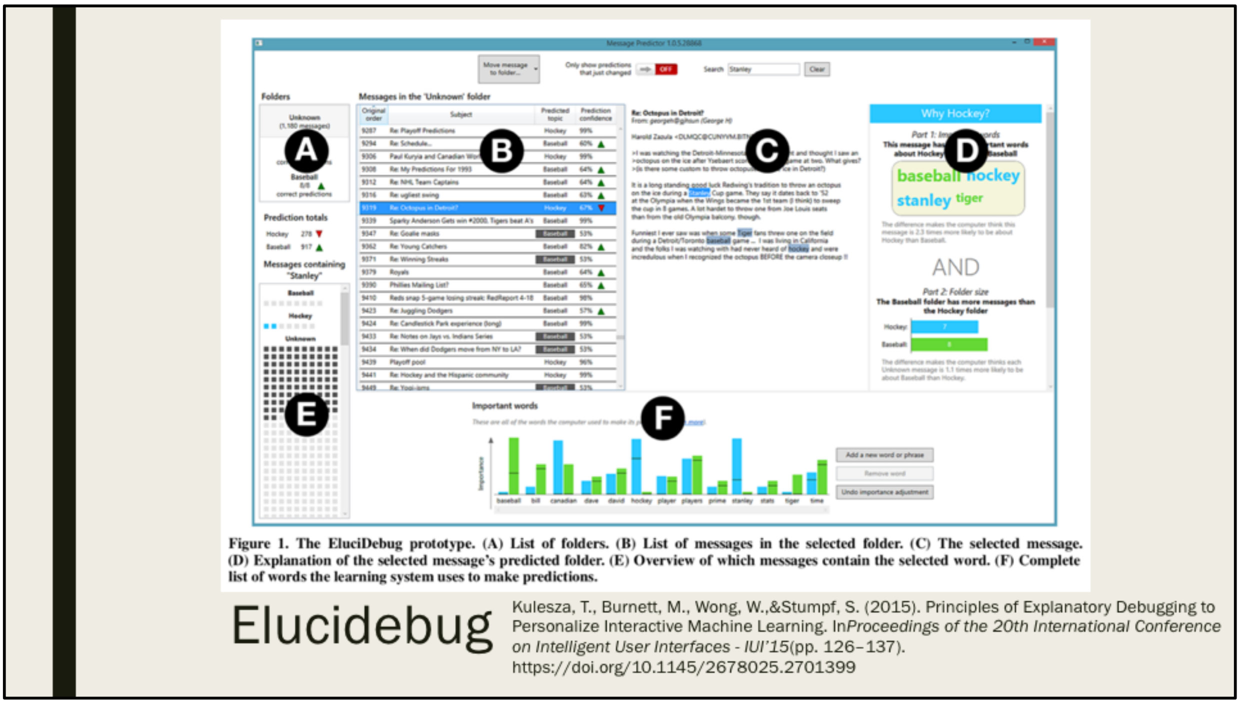
The*human-centric*approach to machine learning explicitly acknowledges the humanwork involved in building and deploying machine learning systems. A central role forhumans is to specify the desired behaviour of the system through the provision oftraining data with labels. When viewed through the lens of traditional statisticalphilosophy, these labels are intended to capture an objective mathematical property ofthe data. However, when faced with the irregular, noisy, and subjective applicationdomains of human-centric systems, this assumption unfortunately produces numerouschallenges which can result in both a poor user experience as well as poorer resultantmodels.

These challenges can be effectively addressed by addressing theinteraction design of the end-user activity of *labelling.*This is because not only islabelling the primary mechanism for non-expert interaction with machine learning,but also because it is where the end-user most clearly encounters the tension betweenthe statistical ideals of supervised learning and human-centricity.
Interactive machine learning (IML) systems enable users to train, customise, andapply machine learning models in a variety of domains. The end-users of these systemsare typically non-experts with no knowledge of machine learning or programming. Incontrast, the professional practice of machine learning, engineering or ‘data

science’typically requires expertise in both those areas. The key design strategy for reducingthe expertise requirements of applied IML systems is to abstract away usingautomation nearly all technical aspects of training and applying models,*except*theprovision of training data.



In the*Crayons*application (Fails&Olsen, 2003), userscan train a model to segment images into different parts. Crayons enables end-usersto build image segmentation classifiers, that is, pixel-level binary classifiers whichsegment portions of an image as falling into one of two classes. For example, a 'humandetector' classifier would take a 2D image of size$w \times h$as input, and as output, produce$w \cdot h$binary labels, one for each pixel, corresponding to whether or not the pixel is partof a human in the image. To build such a classifier in Crayons, users paint labels onan image as they would using a brush tool in a graphics application such as MicrosoftPaint or Adobe Photoshop, being able to toggle between two 'brushes' for the twoclasses. As the user paints, a model is trained, and the output of the model is renderedonto the same image, through a translucent overlay. This allows the user to focus further annotation on misclassified areas.



Another example of an end-user controlled IML system is *EluciDebug* (Kulesza,Burnett, Wong,&Stumpf, 2015). EluciDebug allows end-users to build multi-class classifiers for organising short to medium-length pieces of text, such as email. The userperforms manual annotation by moving emails to folders, where each folder representsa class. As the user organises their email, a model is trained, and the output of themodel is presented as suggestions for classification within the email client itself, whichthe user may accept or overrule. The key thing to note is that both systems involve atraining loop, where the user provides annotations either in the form of trainingexamples or potentially by manually adjusting model parameters (as can be done inEluciDebug). Next, a model is trained and the model output is somehow presented backto the user for further action in such a way as to directly suggest which furtherannotation or adjustment actions would be useful.

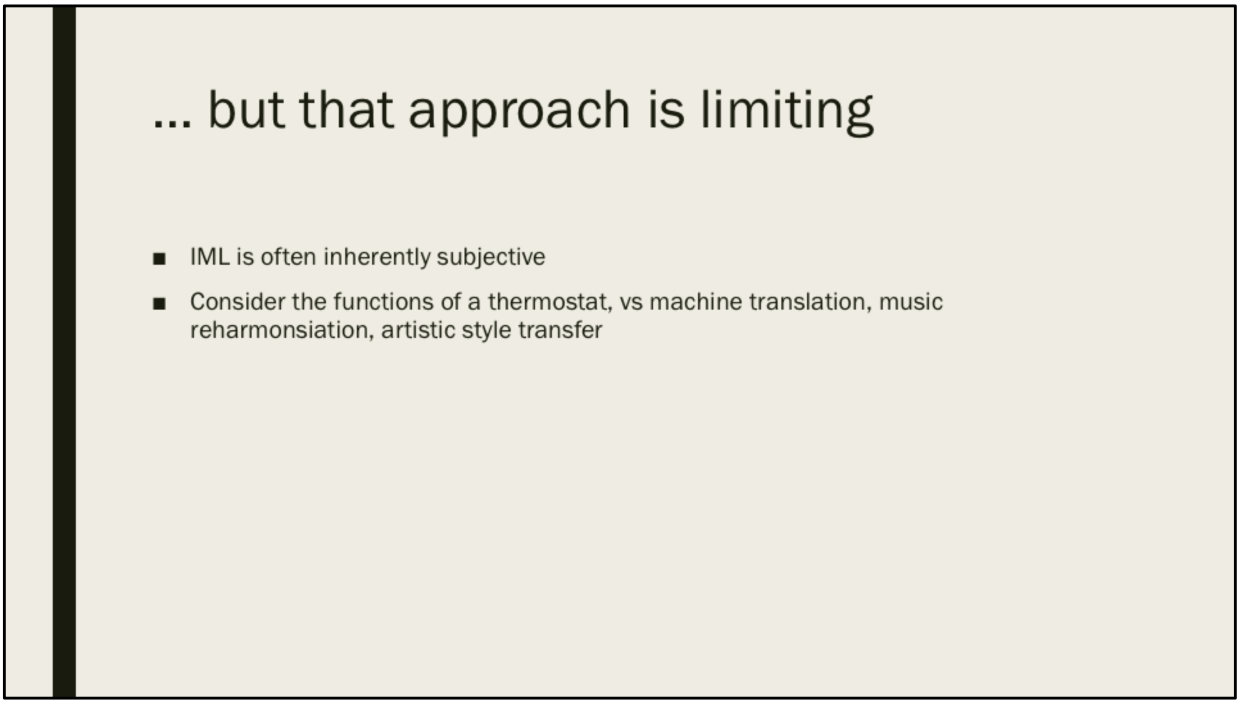
- Users interact with IML systems by providing labelled training instances that exemplify how the system ought to behave
- In labelling data in this way, users are forced to abide by statistical assumptions ofsupervised machine learning models that have been implicitly embedded in IMLsystems.

Labelling *could* be viewed as model construction…

- Fitting models to data
- Uncovering 'natural law' (Breiman, L. (2001). Statistical Modeling: The Two Cultures.*Statistical Science*, *16*(3), 199–215. https://doi.org/10.2307/2676681)
- A 'techno-pragmatist' view

These examples of interacting with a system in order to control its future behaviourcan be considered either as programing, or as model construction. The programmingperspective suggests that the user wants the system to behave in a certain way, and istraining it to do so. The model construction perspective suggests that the system istrying to discover what the user wants, and is building a model of the user's intentionsbased on observations of the user's behaviour. These two perspectives carry verydifferent philosophical assumptions.

The practice of fitting models to data has its roots in the statistical philosophy thatthere exists some natural law underlying observed data (Breiman, 2001). Due toimperfections in the data collection process, the observed data is subject to noise. Theobjective of data modelling, then, is to uncover the parameters of the underlying law.This philosophy has influenced the design of supervised learning algorithms, and inturn, the assumptions of supervised learning have, by default, driven the design ofIML systems. This design influence may be termed 'techno-pragmatism', where theinteraction is designed around satisfying the technical needs of statistical models. Thepurpose of the user, within the overall system design, is to satisfy the requirement foran 'objective' function, encoding the underlying 'law', in which the labels provided bythe user define the 'ground truth' of that law. The techno-pragmatist statistical viewof IML is therefore fundamentally concerned with notions of truth, law and objectivity.



In contrast to the techno-pragmatist view, in which the user is regarded as a source ofobjective ground truth for a statistical inference algorithm, we argue that the functionof an intelligent machine learning system is to be subjective, or more precisely, toreplay versions of subjective behaviour that has previously been captured fromhumans. This type of “intelligence” can be distinguished from mere objectiveautomation, of the kind exhibited by a heating thermostat or adaptive suspension,where behaviour is determined by direct measurement and physical laws. Thoseobjective systems do not require labelling (or at least, the labels are implicit in thedesign of the sensing channels). Examples of subjective judgements include givingnames to things, composing texts, making valuations, or expressing desires – all related to human needs and interpretations. None would be meaningful in the absenceof any human to interpret the result, meaning that they are inherently subjective.

A machine learning system is therefore expected to emulate subjective humanjudgments, and it does this by replicating judgments that humans have been seen tomake. Here are some extreme examples: machine translation systems are trainedusing texts that have been written by humans; music harmonisation systems aretrained using music that has been written by humans; and artistic style generatorsare trained using pictures painted by humans. In a sense, these “intelligent”algorithms offer a kind of institutionalised plagiarism, in which the statisticalalgorithm simply mashes up and disguises the original works until it is impossible tosort out who the rightful authors were.

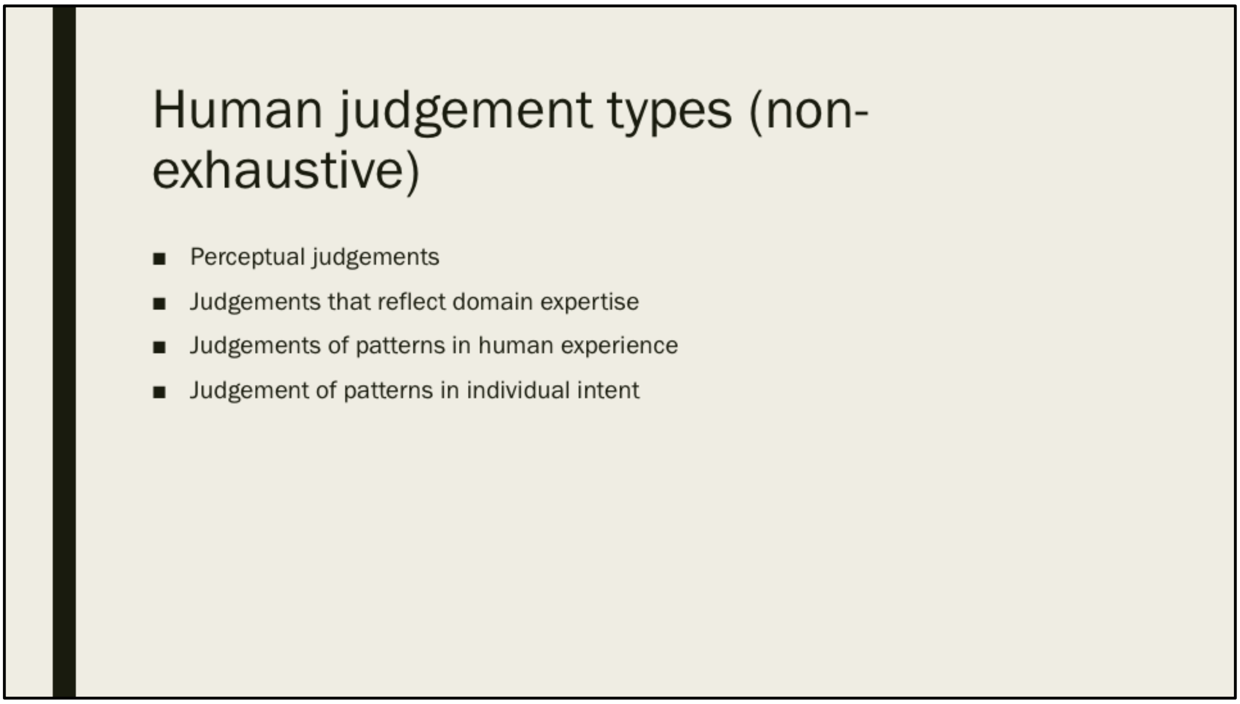
These kinds of creative “intelligence” offer an extreme case of machine behaviour thatis derived from subjective human decisions, but almost all supervised learning systemsdemonstrate similar dependencies. Data is acquired by observing humans (whetherresearchers, volunteers, anonymous Mechanical Turkers or Google searchers) makingdecisions and expressing themselves. The actions of those humans are then replayedby the system as appropriate, based on statistical likelihood that a human would dothe same thing in that situation.

Labelling is an act of programming

- A label is an instruction to the system
- Label providers are engaging in intentional creative acts, which are statistically encoded

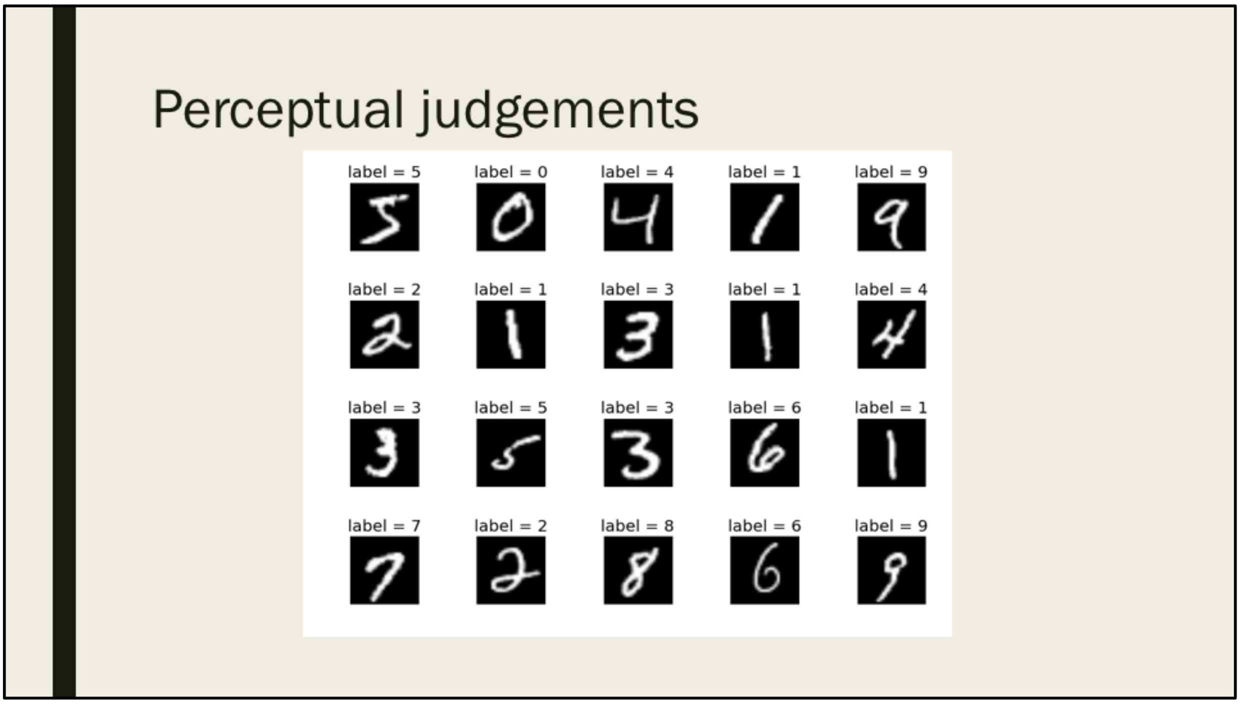
This human-centred perspective on machine learning systems focuses on the ways inwhich system behaviour depends on human actions rather than following physicallaws. When a machine appears to behaviour autonomously, we ask whether thisbehaviour has been derived by observing humans. The observation may either becovert, in which case the intelligence of the system has been achieved by appropriatingthe subjectively authored intentions of others, or else it is done with their awarenessand permission. In the overt case those users become programmers, determiningfuture system behaviour by authoring examples of what that behaviour should looklike.

Labelling is thus a kind of programming, albeit one that is often highly collaborative.A label is an instruction to the system, instructing it by example to behave in a certainway in a certain kind of situation. The system users who provide category labels forsupervised learning systems are engaging in (minor) intentional creative acts. Ofcourse, these intentional acts are statistically encoded and aggregated in ways thatmake it difficult or impossible to acknowledge who the original author was – but theoriginal authors are undeniably humans.

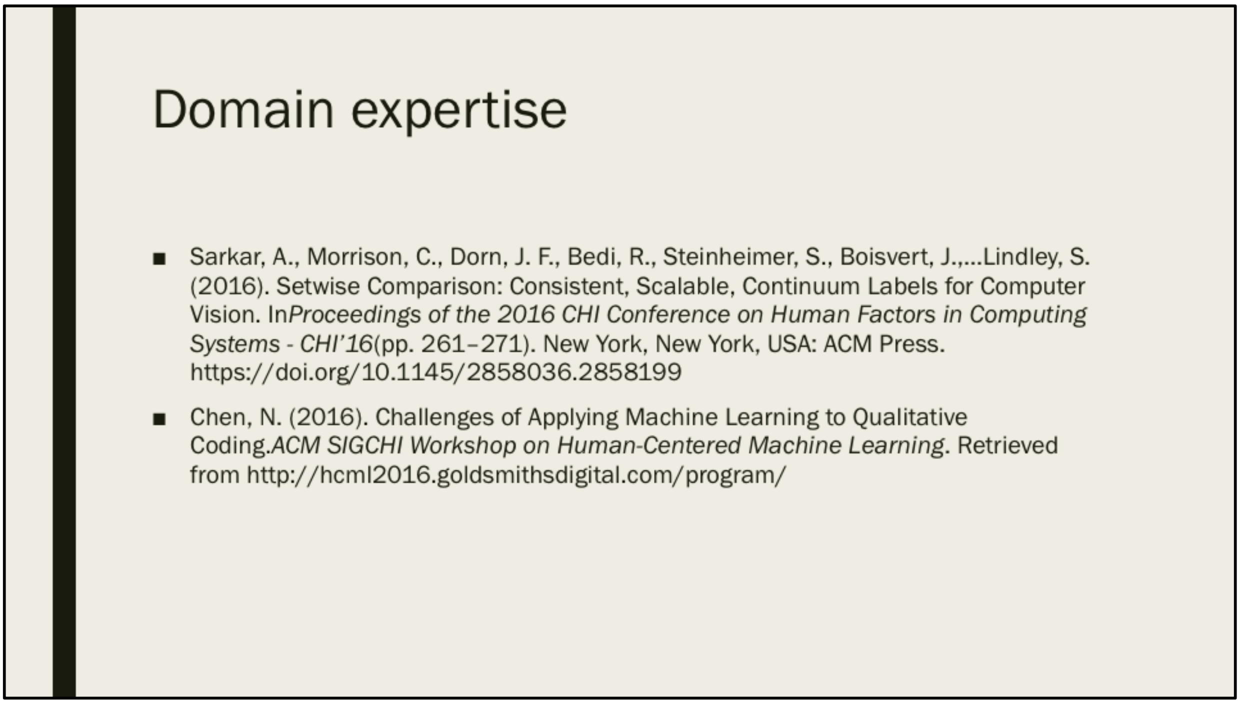


As discussed in the previous section, the purpose of the statistical model in an IMLsystem is not to capture a natural law. Rather, an IML system aims to reproducehuman judgment ability. In order to analyse the implications, we categorise humanjudgments into four (non-exhaustive) types.

perceptual judgements,
judgements that reflect domain expertise,
judgement of patterns in human experience, and
judgement of patterns in individual intent.

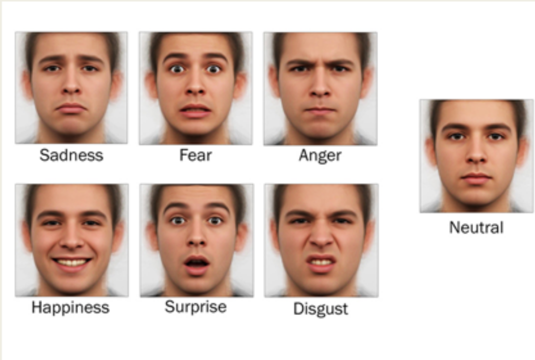


Perceptual judgments are those that rely principally on the human perceptual systemfor assignment of a stimulus to a perceptual category. An example is labelling digits inthe MNIST database (LeCun Yann, Cortes Corinna,&Burges Christopher, 1998).These are often presented as ‘objective’ judgments, although the assumption ofobjectivity is only possible because the training examples themselves have beenselected to reflect a consensus judgment that the labeller is assumed to share. TheMNIST database does not include invalid ‘digits’, non-digits, ambiguous shapes, orartistic subversions of the concept of a digit. Are labels representative of objective ‘facts’about the neuroscience of human vision, or the subjective assumptions shared by thelabellers and data set designers?

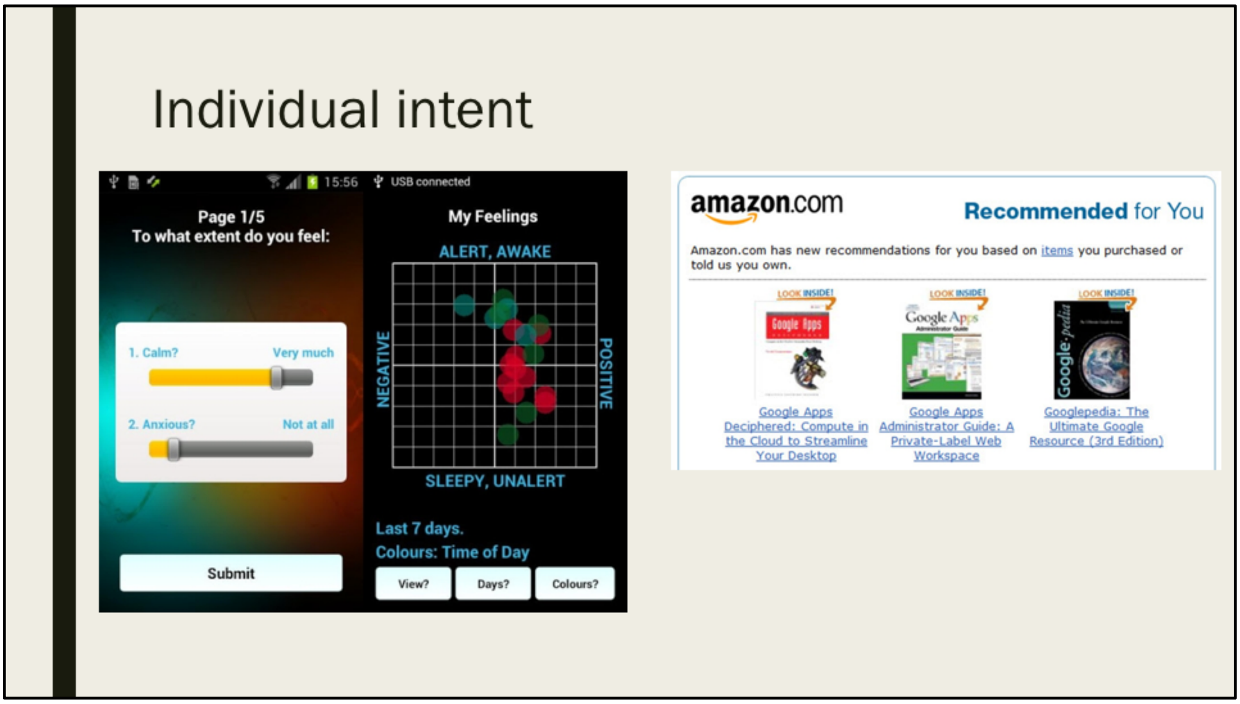


*Domain expertise*judgments rely on labellers' recognised expertise in a particular area. Two example are multiple sclerosis assessment through the analysis of patient videos(Sarkar et al., 2016), and assigning qualitative codes to social science research data (Chen, 2016). Despite these judgments being provided by experts, the concepts beinglabelled may have unclear definitions, impairing label quality. Moreover, manysources may contribute to inter-rater variability, such as variations in previousexperience, training, methods and heuristics used for labelling. Finally, for domainexpertise judgments, access to experts is clearly a prerequisite, which may poselogistical challenges if such expertise is rare.

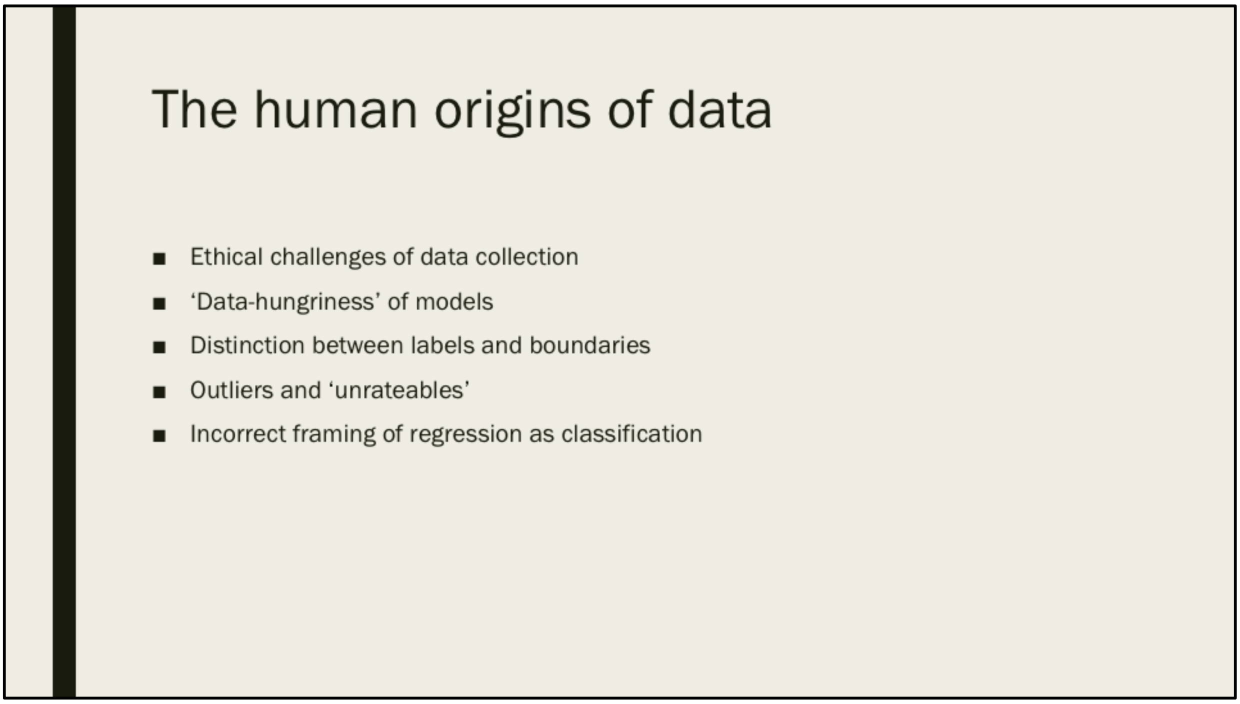
Human experience judgements



Human experience judgments are those that aim to capture some universal aspect ofthe human experience. This might be regarded as a special case of the domain expertise judgment where the domain is being human, as opposed to say, a dog or a monkey. An example is capturing labels for affect recognition (Picard, 1997). Here, there is atenuous assumption that any given person is acting as a representative judge on behalfof all humanity, in relation to universal human experience. In practice, people differ.Typical approaches to mitigate this variation include crowdsourcing and averagingacross labellers. Nonetheless, affect labelling is subject to variations across age, gender,culture, and other factors which are yet to be modelled. While such variation isrecognised as a primary challenge for affective computing (Picard, 2003), it is notexplicitly modelled or acknowledged in the labelling interface (for example, by askingthe labeller to assess the extent of their own individuality).

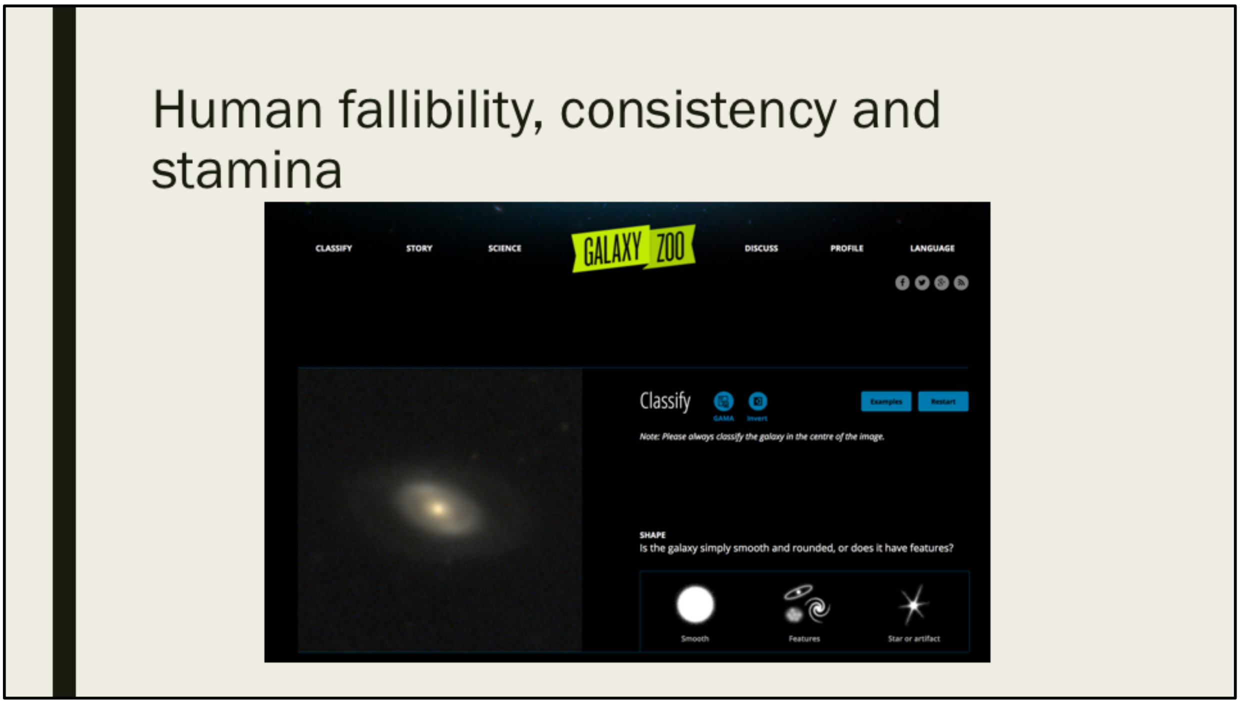


Individual intent judgments reflect personal feelings, desires, and attributes. Unlikethe previous three categories, which appeal to different standards of objectivity(perceptual reality, objective expertise, and universality) these judgements areacknowledged to be inherently subjective because they model an individual. Forexample, applications built with the EmotionSense platform (Lathia et al., 2013) aim to use emotional inference from mobile phone sensors to induce behavioural change, as a sort of personal therapist. However, the system relies at least partially on self-reporting affective states, which suffers from two issues: users may not be motivatedto provide this information repeatedly and consistently, and more importantly, theymay not be capable of consistently self-reporting their emotional state (Afzal&Robinson, 2014). Recommender systems such as Amazon's product recommendationscircumvent this issue by measuring judgments from concrete actions supposedlyreflecting revealed intent rather than expressed intent: products which were viewedor not viewed, bought or not bought. Such actions are unambiguous signals of intent(because the user interface paradigm enforces this), but are still not immune tomisdirection, for example when a user clicks on multiple irrelevant links in order todisguise their search history.



Even before it has been labelled, training data reflects human judgements andpriorities. Modern supervised learning techniques require large training sets to buildstable models, but the scale of data acquisition can raise ethical challenges, includingconsent to use data for new purposes, protected categories of data such as clinicalpatient data, and privacy and anonymity concerns which make it difficult to aggregatedata.

Moreover, some applications require fast convergence. For instance, the TrueSkillsystem (Herbrich, Minka,&Graepel, 2006) was developed for matching players inonline games. A gross mismatch in skill results in a less enjoyable experience for allplayers: the weaker player outclassed, and the stronger player unchallenged. A fastestimate of the player's skill, requiring only a few games, is also desirable, as repeatedmismatches may cause players to stop playing the game. Another example of atechnical approach dealing with fast convergence is one-shot learning (Fei-Fei, Fergus,&Perona, 2006).
Data itself carries epistemological assumptions that have been embedded in the wayit was collected. From the machine learning perspective, there may not be a formaldistinction between*examples*which cannot be placed exactly in the space of labels, andlabel*boundaries*which are not precise. However, they are very different from theperspective of a human labeller. Imprecise label boundaries may undermine labellerconfidence throughout the entire labelling activity. Training examples may also poseproblems because they are outliers, or simply unrateable. As noted by Chen (Chen,2016), outliers are typically discarded in quantitative analyses, but become the focusof attention in qualitative analyses. Examples that are unratable (perhaps because ofdata corruption or because they contain no meaningful information) may impair thelabelling process if the labelling tool has no provision to mark examples as unrateable,or the labeller is not equipped to identify such a situation should it arise.
In some cases, a regression problem is incorrectly framed as a classification problemfor the purpose of labelling – it is easier to ask labellers to provide one of a discrete setof labels than a real number on a continuous scale. However, this can result in theunnecessary conceptualisation of examples as belonging to a set of discrete categories,which causes issues for examples on the boundaries of different categories. This is theproblem faced by the Assess MS problem, detailed in the next section. Unclear conceptscause problems generally in precision, but less so for accuracy.



Humans are fallible. If there are large amounts of data to be labelled, the quality ofjudgements can be impaired as the labeller becomes tired. In the Assess MS projectdescribed in the next section, neurologists would spend an entire workday, sometimestwo, continuously labelling short video clips (Sarkar et al., 2016). Appropriate tools,such as the setwise comparison tool developed for Assess MS, can mitigate this problem.Explicit strategies to maintain interest and prevent boredom have been applied inexperiments such as the Galaxy Zoo (Lintott et al., 2008) which show compellingevidence for the benefit of ludic and engaging labelling tools.
Even in optimum conditions, people still make mistakes, misinterpret instructions ordisagree with each other. This is well understood in scientific studies where data mustbe categorised by an observer, such as coding of free-text questionnaire responses.Where one researcher might interpret an observed response in one way, another seesit differently. This difference might come from not stating or communicating criteriathat have been applied by one rater, or from terminological imprecision, for example,stemming from a different understanding of the criteria that two raters might have,or simply their wishful thinking in relation to a hypothesis.

Measuring label reliability

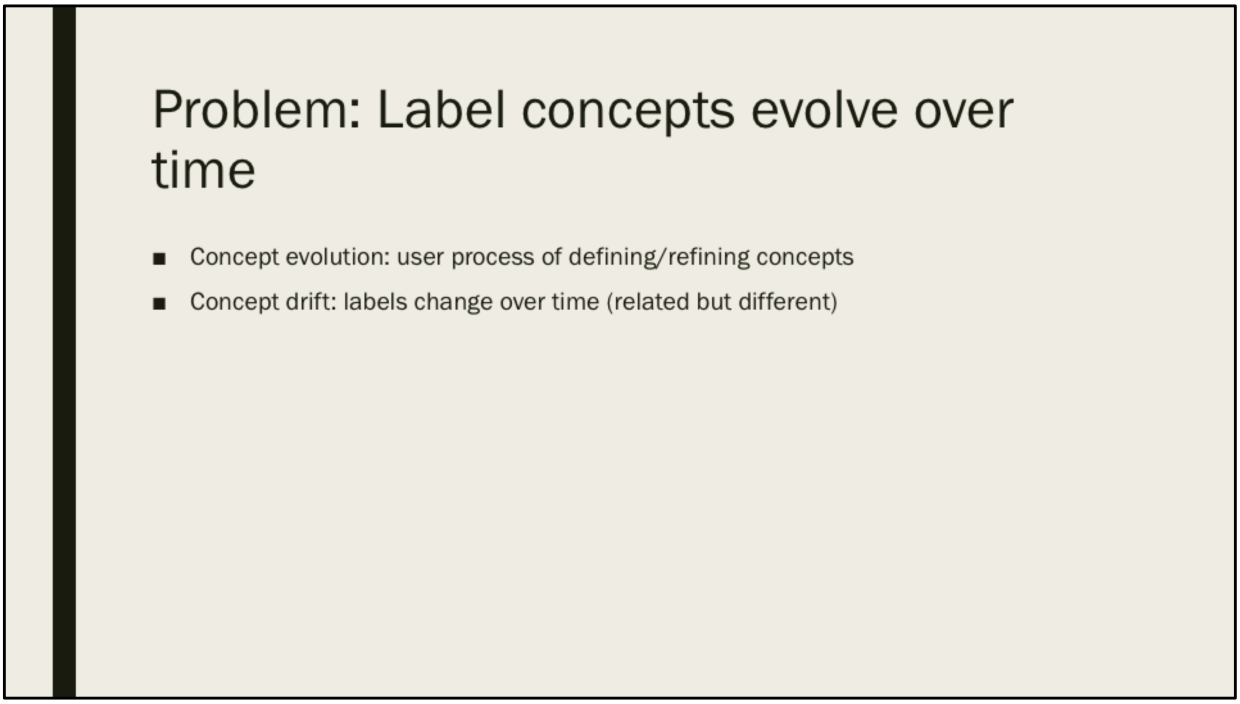
- Inter and intra-rater reliability measurements
  - *E.g., Cohen's Kappa, Krippendorff's Alpha*
- Error with respect to 'ground truth'

In response to this problem, qualitative social science researchers monitor thereliability of classification judgments. They want to know whether a judge consistentlymakes the same judgment in equivalent cases, and also whether two judges make thesame decision as each other. The second is more often discussed, because it happensso consistently. It is described as inter-rater reliability (IRR), and is often summarisedby a statistical measure such as Cohen's kappa (for the case of two raters), whichcompares the level of agreement to what might be expected from chance. IRR testingis intuitively appealing to computer scientists such as HCI researchers, because thefirst rating can be considered as a design decision, and the second rating as a test ofthat decision. Inter-rater reliability is never 100%, but pragmatic allowance for thelimits of human performance means that certain thresholds are considered acceptablewithin the range of observation error.

The question of whether a single person agrees with themselves (when repeating thesame judgment) is less often asked in computer science, but of more concern inmedicine, where it is quite likely that a clinician might assess the same patient morethan once, with a considerable interval between the assessments. Clinical researchsuggests that this test-retest reliability is also imperfect, with clinicians applyingdifferent criteria at different times, perhaps because of explicit training and correction,or perhaps because of changing tacit or contextual factors that the clinician may notbe consciously aware of. We discuss this issue further next.

STRUCTURED LABELLING FOR CONCEPT EVOLUTION

Case study I

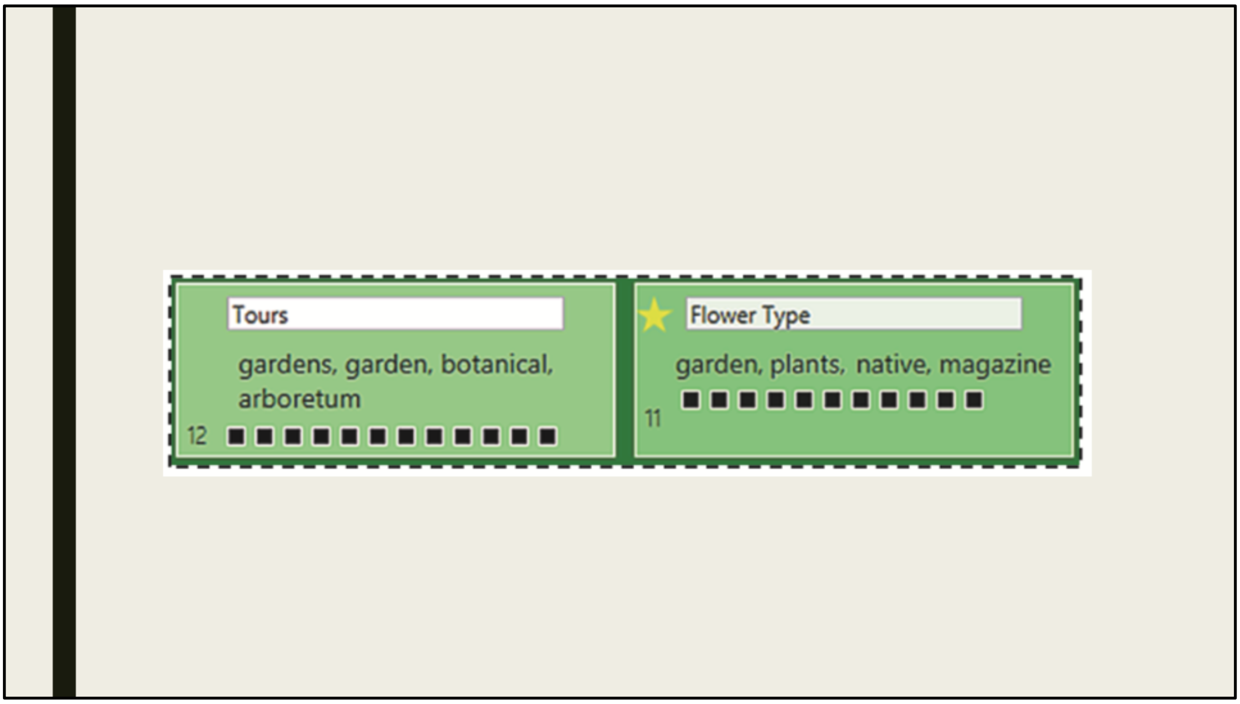


While labeling data is a seemingly simple task, it is actuallyfraught with problems (e.g., [9, 19, 26]). Labels reflect alabeler’s mapping between the data and their underlying*concept*(i.e., their abstract notion of the target class). Thus,label quality is affected by factors such as the labeler’sexpertise or familiarity with the concept or data, theirjudgment ability and attentiveness during labeling, and theambiguity and changing distribution of the data itself. This paper addresses a distinct problem in labeling data thatwe refer to as*concept evolution*. Concept evolution refers tothe labeler’s process of defining and refining a concept intheir minds, and can result in different labels being appliedto similar items due to changes in the labeler’s notion of theunderlying concept. In a formative study presented later inthis paper, we found that people labeling a set of web pagestwice with a four-week gap between labeling sessions were,on average, only 81% consistent with their initial labels. Thisinconsistency in labeling similar items can be harmful tomachine learning, which is fundamentally based on the ideathat similar inputs should have similar outputs

An even more insidious problem in data labeling is*conceptdrift*, where the underlying data is fundamentally changingover time [29]. An example of concept drift is a newsrecommender that attempts to recommend the mostinteresting recent news. Here, the concept of*interesting*mayremain the same over time, but the data (in this case thenews) is constantly drifting as a result of changing currentevents. Most solutions to concept drift model conceptstemporally, such as by discarding or weighting informationaccording to a moving window over the data (e.g., [27, 33)or by automatically identifying new types of data (e.g., [5,15]). Critically, none of these solutions are intended to helpa*user*refine their own idea of a concept, a problem whichmay be exacerbated in the presence of concept drift.

we introduce *structured labelling* (Figure 1), a novel interaction techniquefor helping people define and refine their concepts as they label data. Structured labeling allows people to organize theirconcept definition by grouping and tagging data (as much or as little as they choose) within a *traditional labelling* scheme (e.g., labeling into mutually exclusive categories such as‘yes’, ‘no’, and ‘could be’). This organization capability helps to increase label consistency by helping people explicitly surface and recall labeling decisions. Further, because the structure is malleable (users can create, delete,split, and merge groups), it is well-suited for situations where users are likely to frequently refine their concept definition as they observe new data.

Kulesza’s structured labeling approach allows people to group data in whatever way makes sense to them. By seeing theresulting structure, people can gain a deeper understanding of the concept they are modeling. Here, the user sees an uncategorizedpage (top left) and can drag it to an existing group (right), or create a new group for it. The thumbnails (bottom left) show similarpages in the dataset to help the user gauge whether creating a new group is warranted.



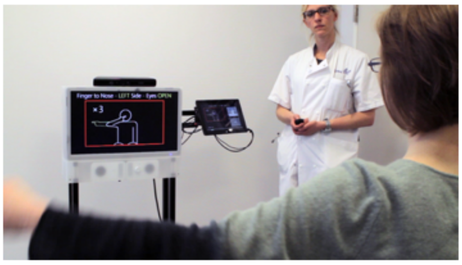
Our assisted structuring tool provides users with automatic summaries of each group's contents (below the user-supplied tag area) and recommends a group for thecurrent item via an animation and yellow star indicator. The black squares indicate how many items are in each group.

SORTABLE

Case study II

Assess MS

- Aim: a more consistent way of quantifying progression of motor illness in multiple sclerosis
- Input: Kinect RGB + depth videos of standard clinical movements
- Output: a standardised clinical disability score

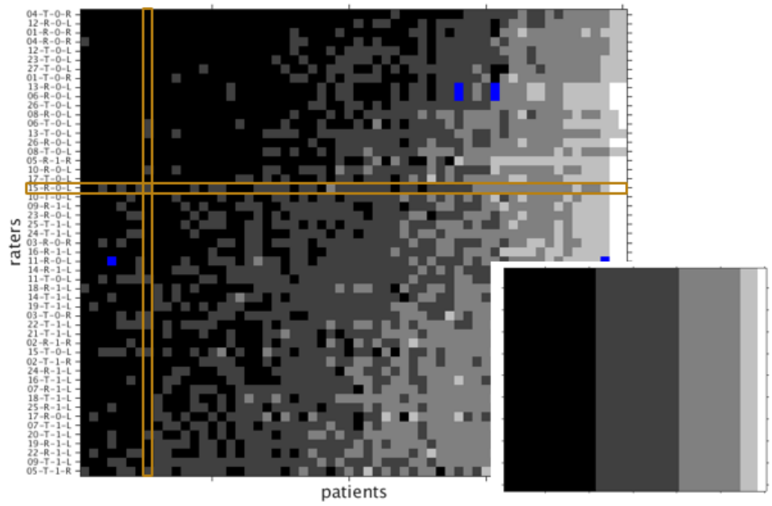


Problem: consistent labels

- Numeric scoring has poor labeller agreement
  - concept boundaries unclear even after iteration
- Crowdsource?
  ➡️ can't, need highly expert labellers
- Average across labellers?
  ➡️ can't, patient confidentiality
- Model individual labeller noise/bias?
  ➡️ can't, learning effects

25

Inter-rater consistency is limited



NOVARTIS

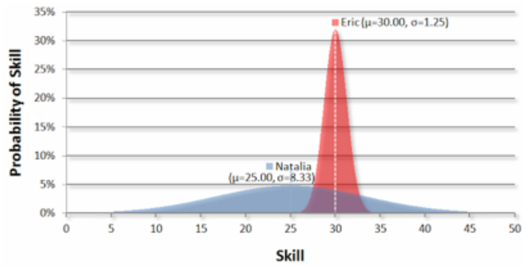
Partial solution

- Preference judgements
  - ‘this is **better / worse / equal** to that’ as opposed to ‘this is a **3**, that is a **4**’.
  - Not scalable :(

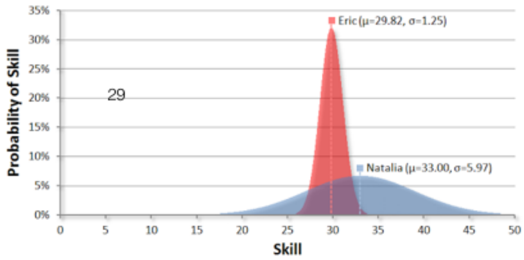
Full solution

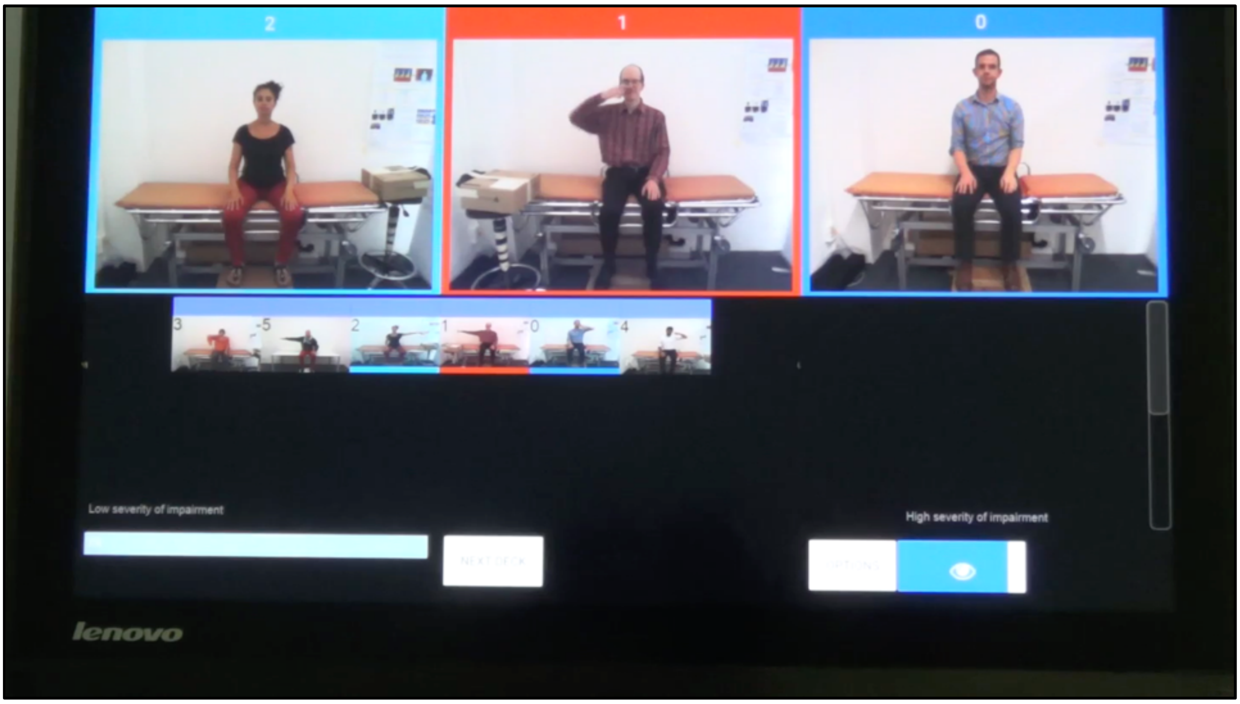
- Setwise comparison + TrueSkill inference
  - Order **sets** of videos with overlap
  - but don't need all pairwise comparisons
  - **Infer** remaining relationships

Prior



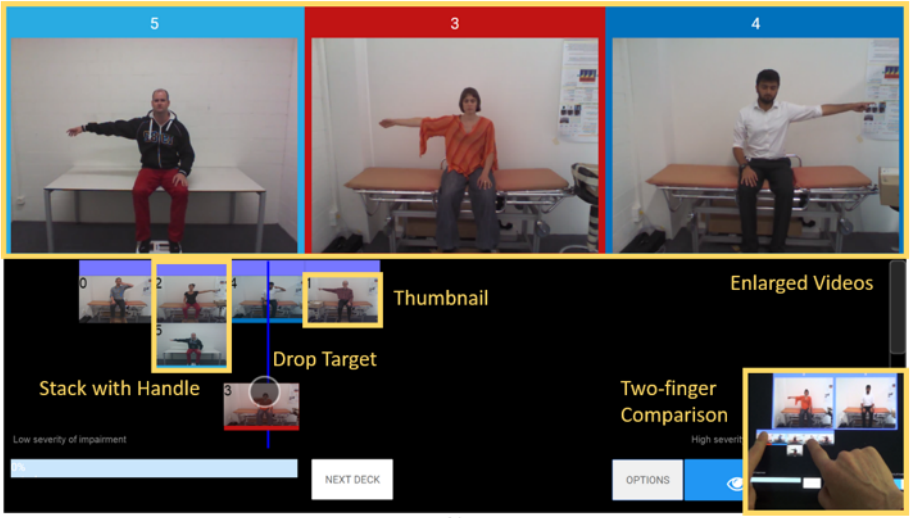
After Natalia wins




2
1
0
Low severity of impairment
High severity of impairment
lenovo

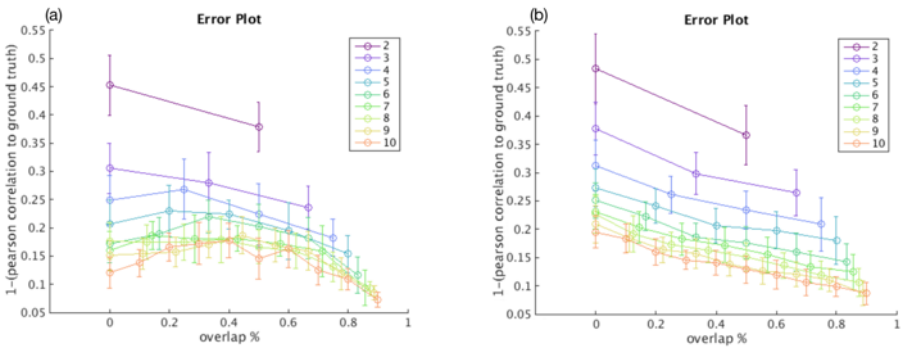
SorTable

an interface for setwise comparison

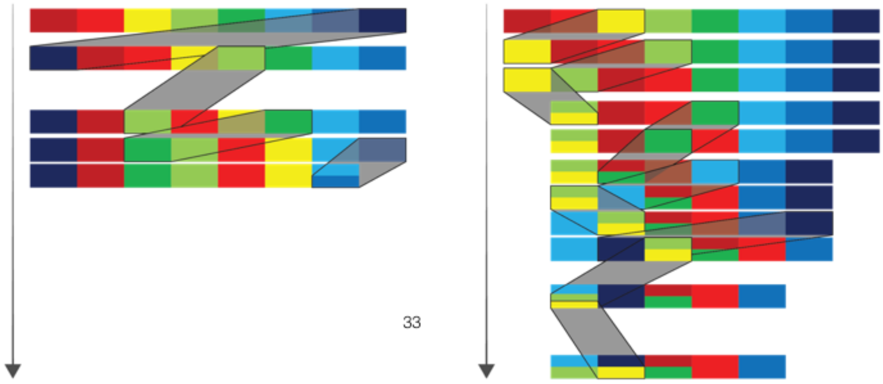


31

Choosing deck size and overlap



Sorting strategies



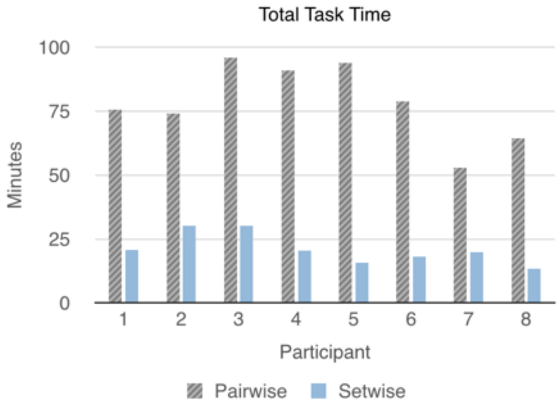
So, does it work?

- Already known: pairwise comparison achieves higher consistency than assigning numerical scores, but very slow
- **Question**: Does setwise comparison achieve a better efficiency-consistency tradeoff?
- Compared pairwise and setwise using 8 neurologists rating a set of 40 videos

34

Result 1:
Setwise comparison is more efficient

- Setwise task time was
  54 minutes less on average
  ($p = 4 \cdot 10^{-5}$)



35

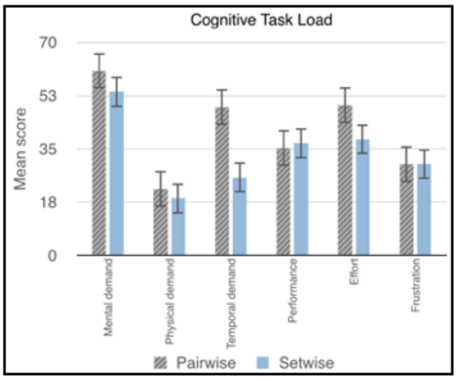
Result 2:
Setwise comparison is more consistent!

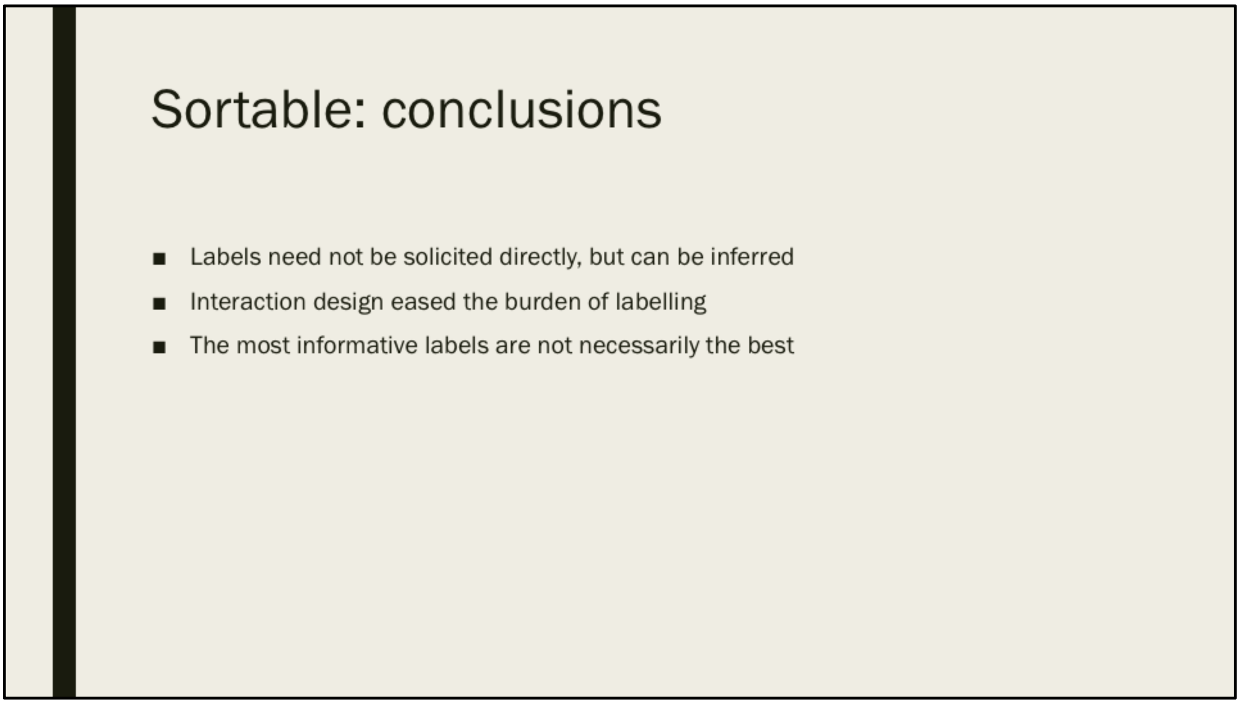
Agreement ***between*** labellers

	Global ICC	**Average ICC**
		mean±sd [min−max]
Pairwise	0.70	$0.77 \pm 0.1[0.64 - 0.94]$
Setwise	0.83	$0.85 \pm 0.07[0.72 - 0.95]$
t-test		$p = 5 \cdot 10^{-4}$

Why is it more consistent???

- *Inferring* missing comparisons was better than ***measuring*** all comparisons.
- Cognitive load assessment was inconclusive.
- Potential explanations:
  - Fatigue
  - TrueSkill's implicit noise[37] modelling
  - Increased reference points





We reframed the problem so that users were not providing labels directly, butproviding information from which labels could be reconstructed. In this way, we couldbuild upon strong human capability in relative judgement and still provide theclassification labels required by the Assess MS system. This overcame noisy labels,improving the accuracy of the algorithm by 10%.

A key insight was to by enabling setwise rather than pairwise comparison, achievingthree benefits for the users. First, the presentation of videos in sets builds upon humanshort-term memory to make multiple comparisons at once. Second, the ability to createstacks to indicate that videos are the same can substantially reduce the number ofcomparisons the labeller needs to make when sorting. Third, SorTable facilitatesmixed-strategy sorting, including the automatic display of the left and rightneighbours of the currently selected video, and the ability to compare any two videoswith a two-finger gesture. All interactions are touch based.

We found that choosing videos to label to maximise TrueSkill’s information gain andultimately decrease the number of required labels was not a good strategy for humanlabellers. It is less cognitively taxing for people to differentiate between very differentvideos rather than similar ones. Put differently, labels that satisfy a classifier’sinformation needs perfectly may also be the hardest for humans to give

(Lang&Baum,1992), and increase stress and fatigue.