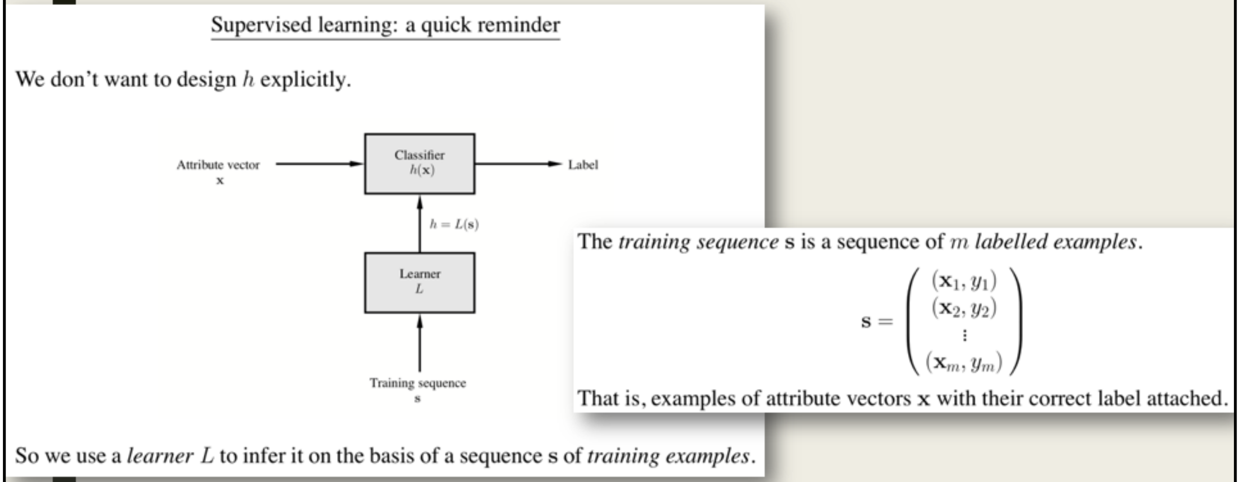
# LABELLING

Interaction with Machine Learning
Cambridge MPhil ACS 2018-2019

# What’s the big deal?



Supervised learning: a quick reminder

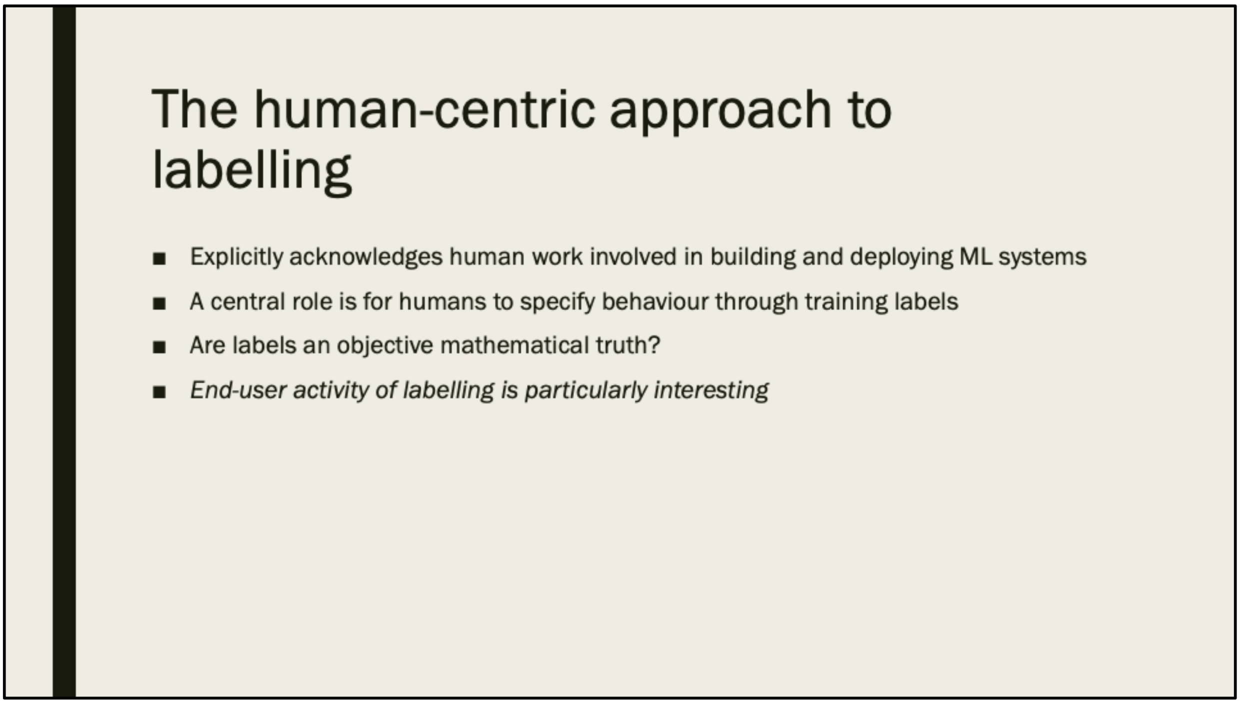
We don’t want to design $h$ explicitly.

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.



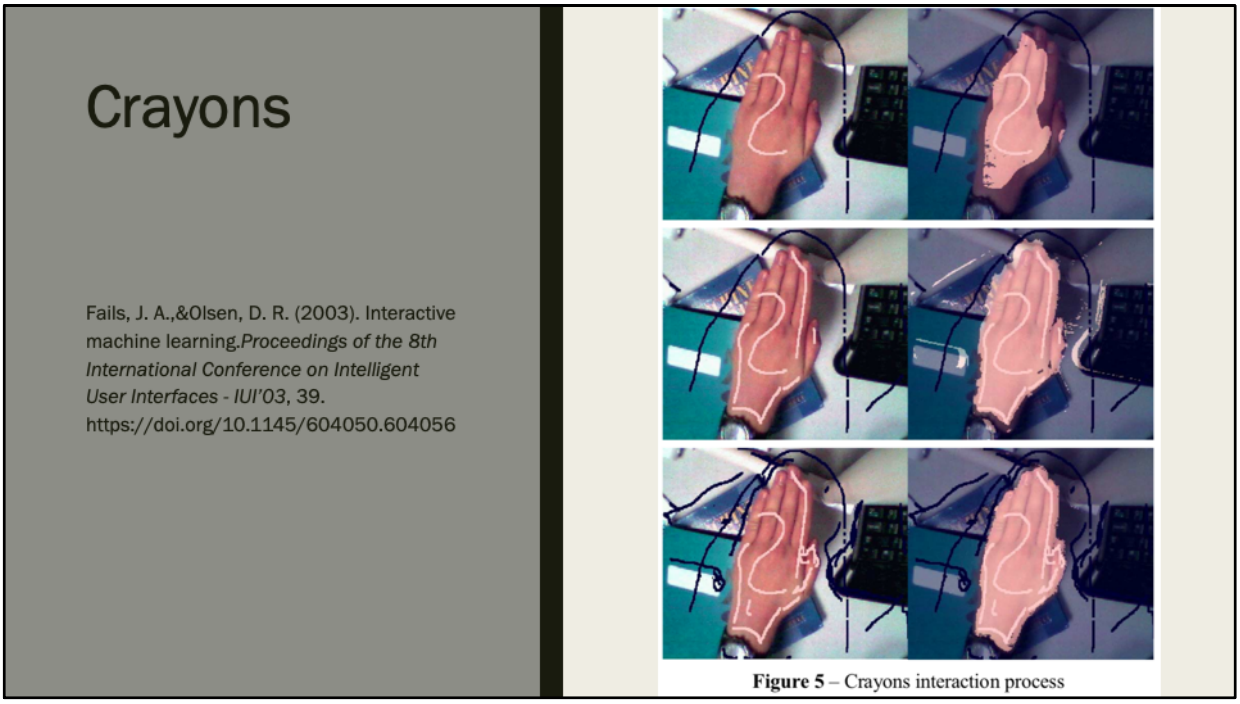
## The human-centric approach to labelling

- Explicitly acknowledges human work involved in building and deploying ML systems
- A central role is for humans to specify behaviour through training labels
- Are labels an objective mathematical truth?
- *End-user activity of labelling is particularly interesting*

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 the interaction design of the end-user activity of *labelling.* This is because not only is labelling the primary mechanism for non-expert interaction with machine learning, but also because it is where the end-user most clearly encounters the tension between the statistical ideals of supervised learning and human-centricity.
Interactive machine learning (IML) systems enable users to train, customise, and apply machine learning models in a variety of domains. The end-users of these systems are typically non-experts with no knowledge of machine learning or programming. In contrast, the professional practice of machine learning, engineering

or ‘data science’ typically requires expertise in both those areas. The key design strategy for reducing the expertise requirements of applied IML systems is to abstract away using automation nearly all technical aspects of training and applying models, *except* the provision of training data.



# Crayons

Fails, J. A.,&Olsen, D. R. (2003). Interactive machine learning.*Proceedings of the 8th International Conference on Intelligent User Interfaces - IUI'03*, 39. https://doi.org/10.1145/604050.604056

**Figure 5** – Crayons interaction process

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 'hand detector' classifier would take a 2D image of size*w×h*as input, and as output, produce*w·h*binary labels, one for each pixel, corresponding to whether or not the pixel is partof a hand 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.

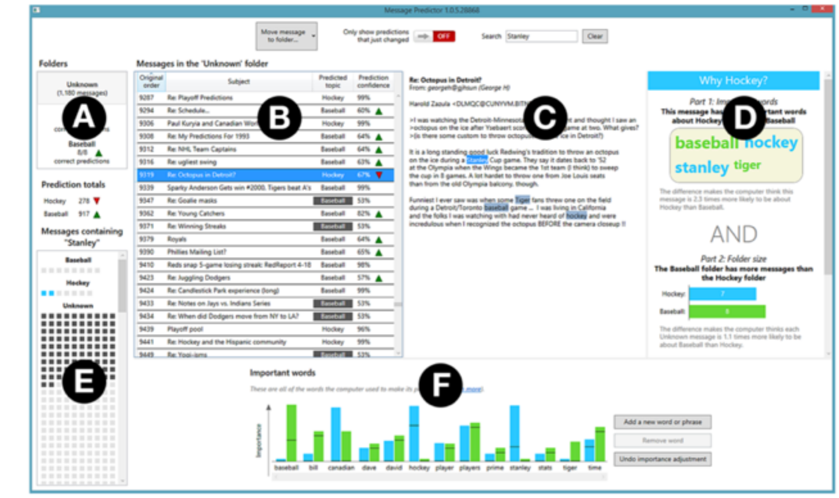
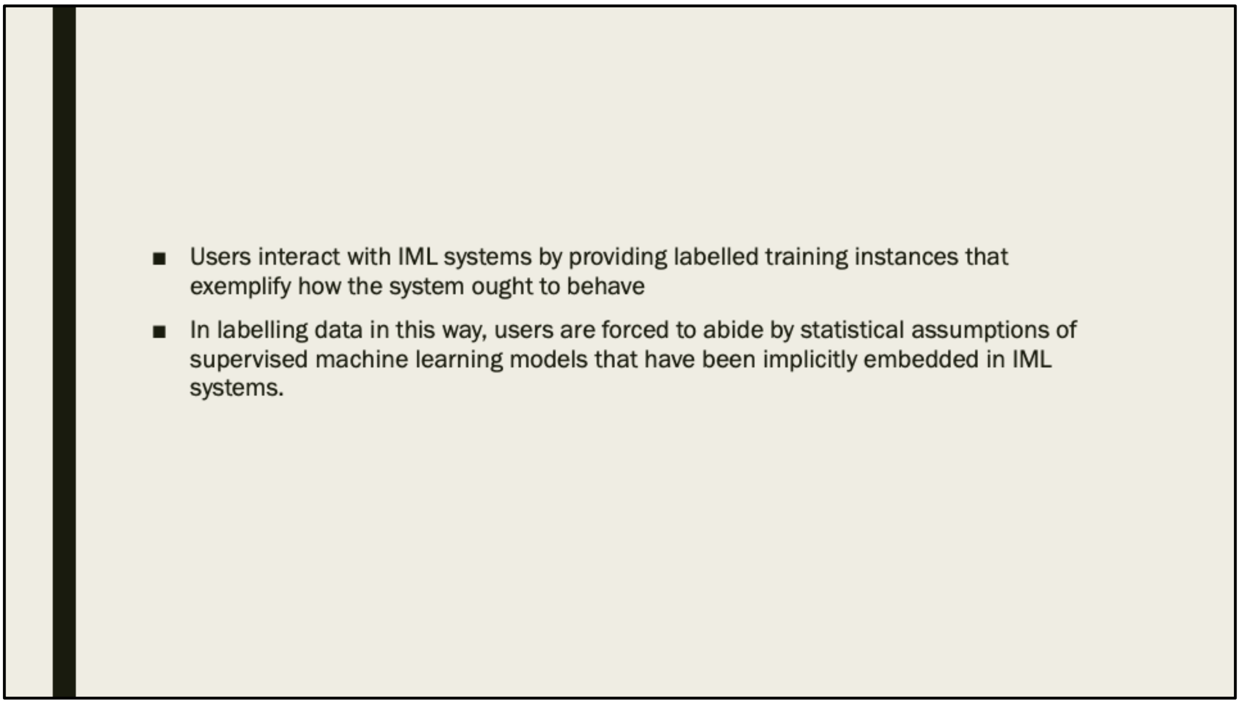


**Figure 1. The EluciDebug prototype. (A) List of folders. (B) List of messages in the selected folder. (C) The selected message. (D) Explanation of the selected message's predicted folder. (E) Overview of which messages contain the selected word. (F) Complete list of words the learning system uses to make predictions.**

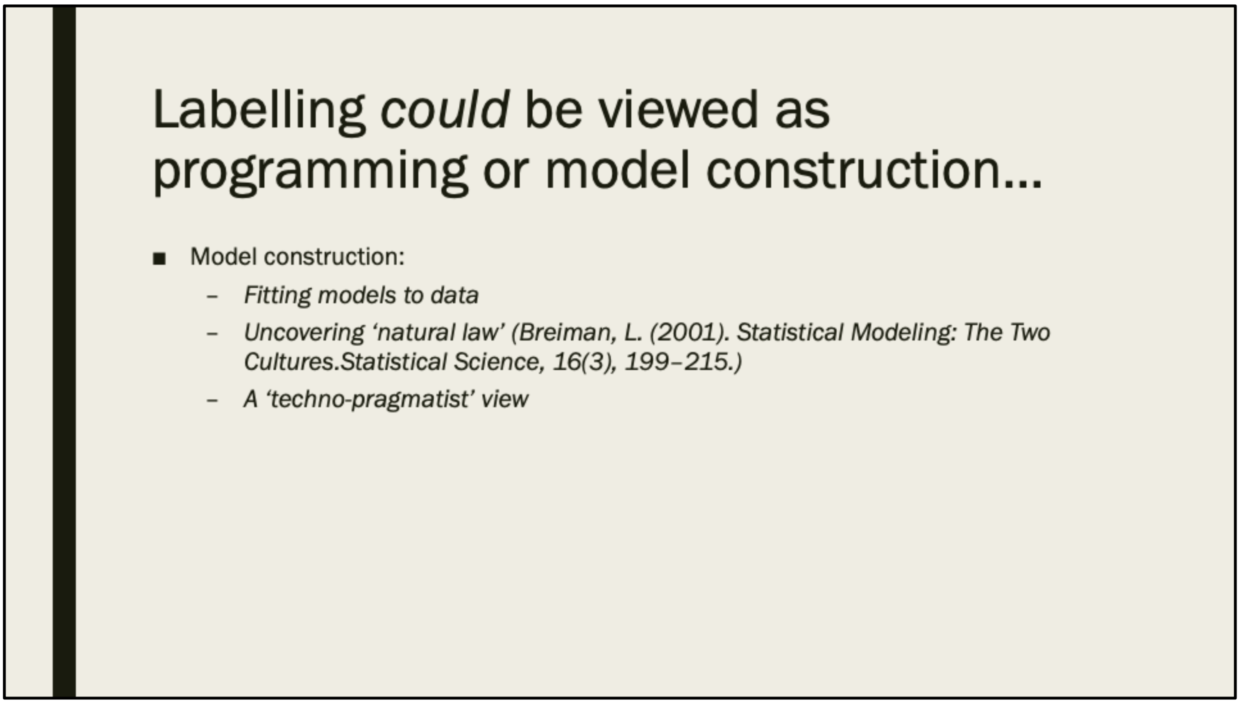
# Elucidebug

Kulesza, T., Burnett, M., Wong, W.,&Stumpf, S. (2015). Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In*Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI'15*(pp. 126–137). https://doi.org/10.1145/2678025.2701399

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 user performs manual annotation by moving emails to folders, where each folder represents a class. As the user organises their email, a model is trained, and the output of the model 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 a training 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 of supervised machine learning models that have been implicitly embedded in IML systems.



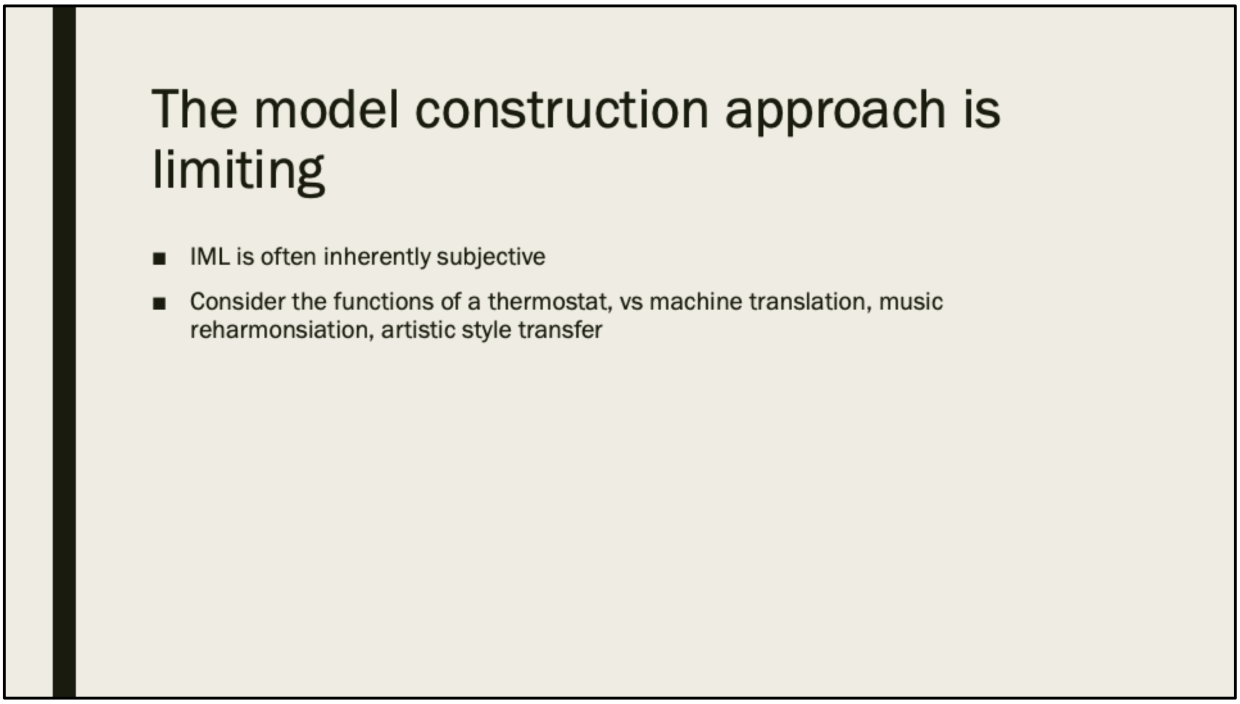
## Labelling *could* be viewed as programming or model construction…

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

These examples of interacting with a system in order to control its future behaviour can be considered either as programing, or as model construction. The programming perspective suggests that the user wants the system to behave in a certain way, and is training it to do so. The model construction perspective suggests that the system is trying to discover what the user wants, and is building a model of the user's intentions based on observations of the user's behaviour. These two perspectives carry very different philosophical assumptions.

Let's start with the model construction view:

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



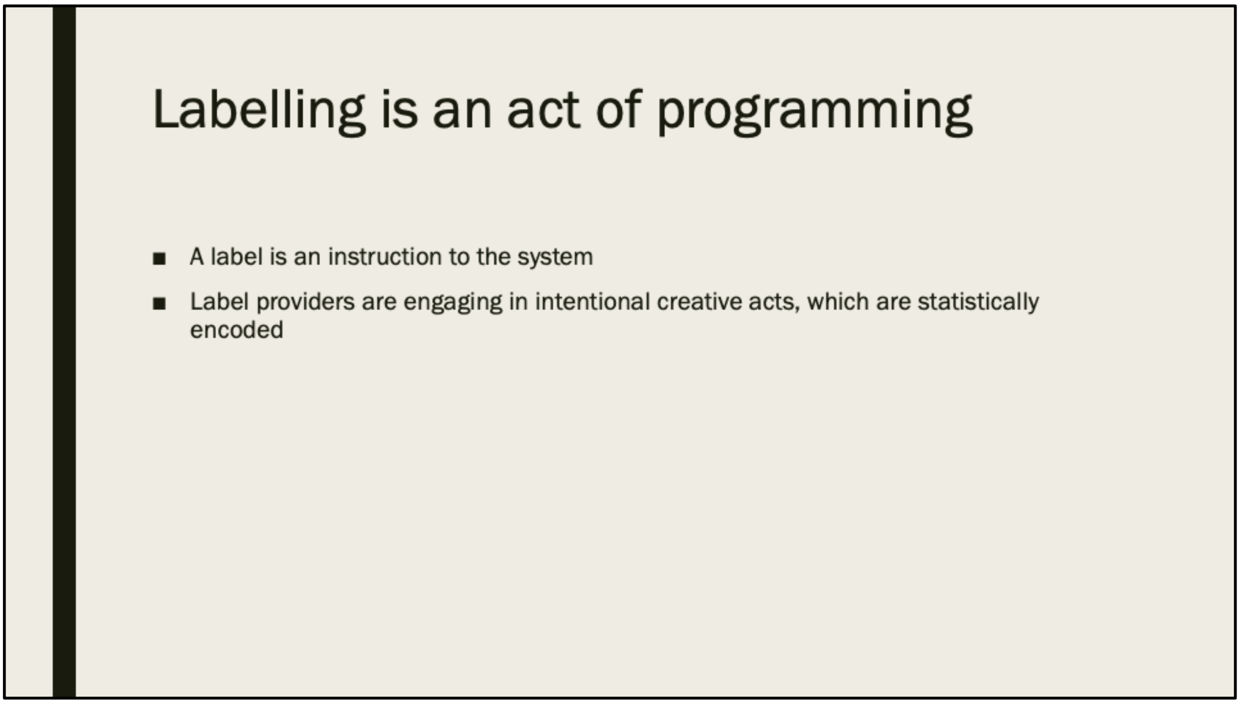
## The model construction approach is limiting

- IML is often inherently subjective
- Consider the functions of a thermostat, vs machine translation, music reharmonsiation, artistic style transfer

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

In many cases, a machine learning system is therefore expected to emulate subjective human judgments, and it does this by replicating judgments that humans have been seen to make. Here are some extreme examples: machine translation systems are trained using texts that have been written by humans; music harmonisation systems are trained using music that has been written by humans; and artistic style generators are trained using pictures painted by humans. In a sense, these “intelligent” algorithms offer a kind of mechanised plagiarism, in which the statistical algorithm simply mashes up and disguises the original works until it is impossible to sort out who the rightful authors were.

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

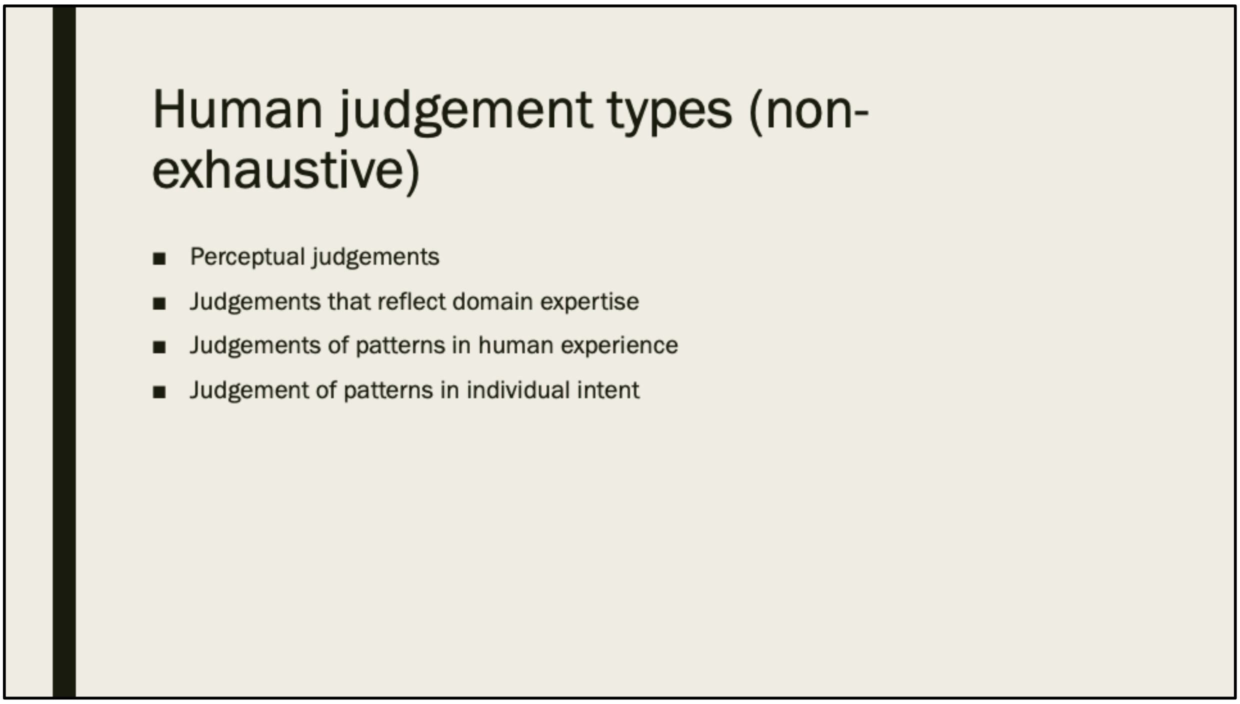


## Labelling is an act of programming

- 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 in which system behaviour depends on human actions rather than following physical laws. When a machine appears to behaviour autonomously, we ask whether this behaviour has been derived by observing humans. The observation may either be covert, in which case the intelligence of the system has been achieved by appropriating the subjectively authored intentions of others, or else it is done with their awareness and permission. In the latter (overt) case those users become programmers, determining future system behaviour by authoring examples of what that behaviour should look like.

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 certain way in a certain kind of situation. The system users who provide category labels for supervised learning systems are engaging in (minor) intentional creative acts. Of course, these intentional acts are statistically encoded and aggregated in ways that make it difficult or impossible to acknowledge who the original author was – but the original authors are undeniably humans.

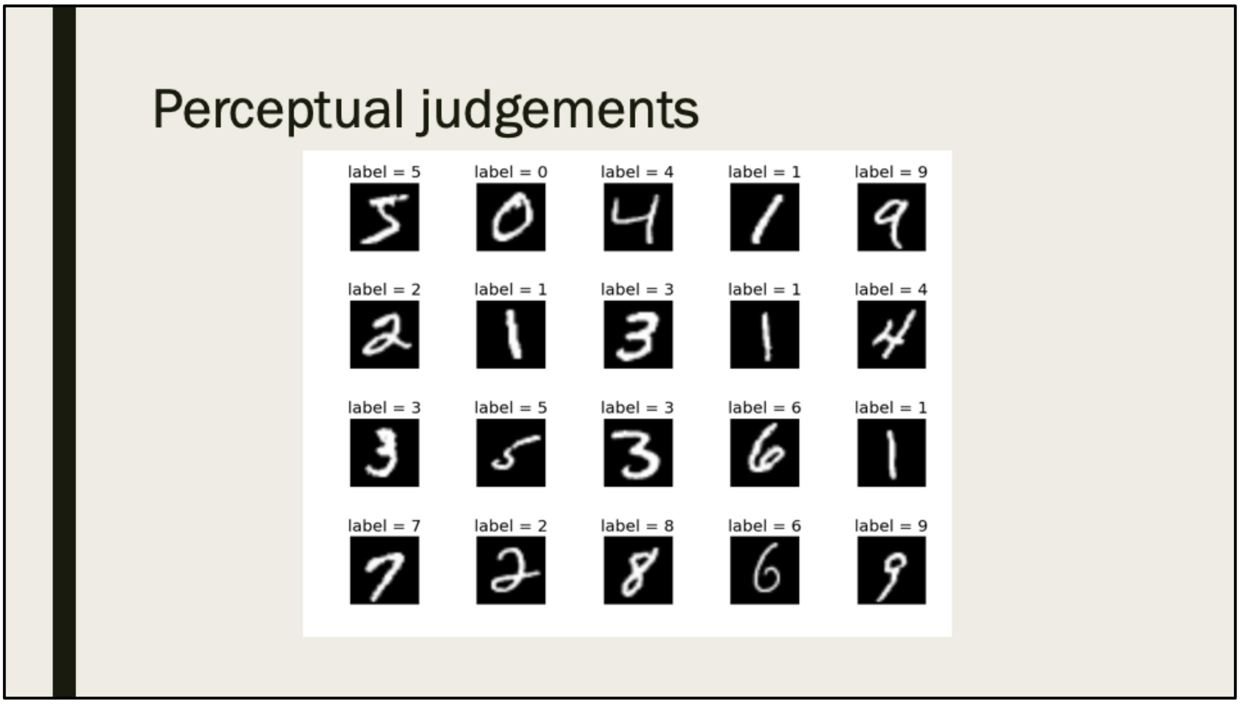


## Human judgement types (non-exhaustive)

- Perceptual judgements
- Judgements that reflect domain expertise
- Judgements of patterns in human experience
- Judgement of patterns in individual intent

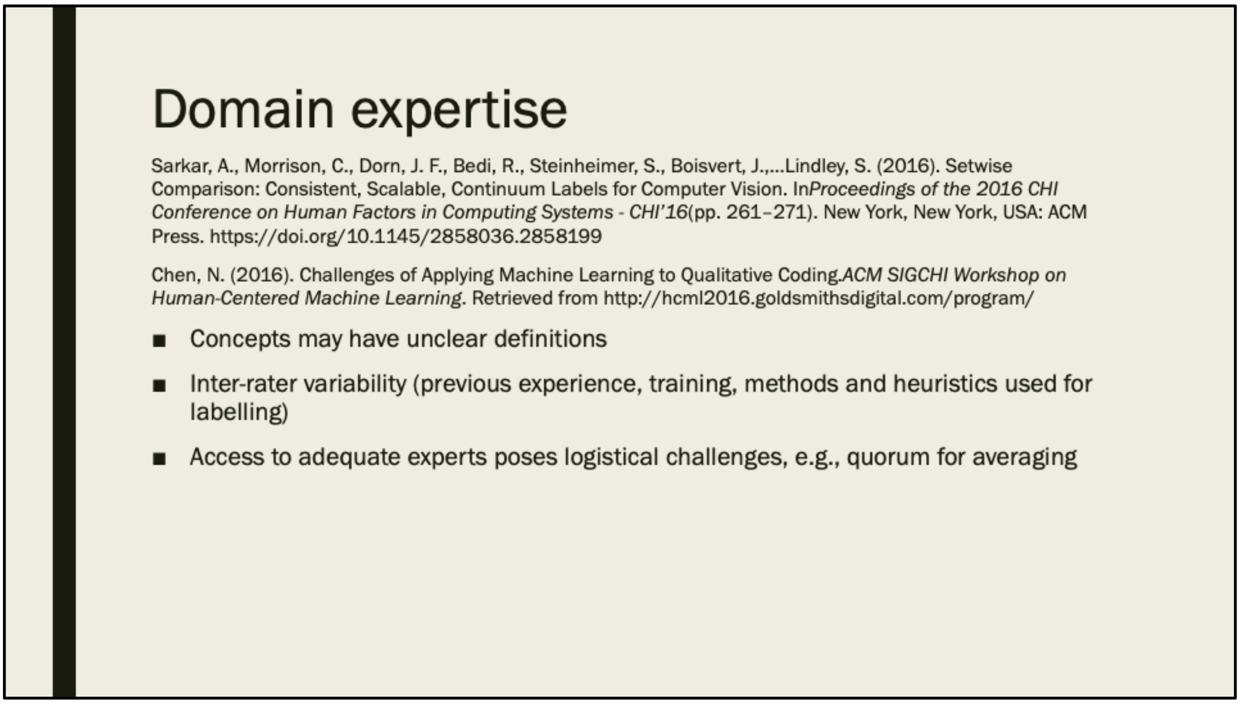
So, the purpose of the statistical model in an IML system is not to capture a natural law. Rather, an IML system aims to reproduce human judgment ability. In order to analyse the implications for design, we categorise human judgments into four (non-exhaustive) types.

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



## Perceptual judgements

*Perceptual* judgments are those that rely principally on the human perceptual system for assignment of a stimulus to a perceptual category. An example is labelling digits in the MNIST database (LeCun Yann, Cortes Corinna,&Burges Christopher, 1998).These are often presented as 'objective' judgments, although the assumption of objectivity is only possible because the training examples themselves have been selected to reflect a consensus judgment that the labeller is assumed to share. The MNIST database does not include invalid 'digits', non-digits, ambiguous shapes, or artistic subversions of the concept of a digit. Think about the following question: are labels representative of objective 'facts' about the neuroscience of human vision, or the subjective assumptions shared by the labellers and data set designers?



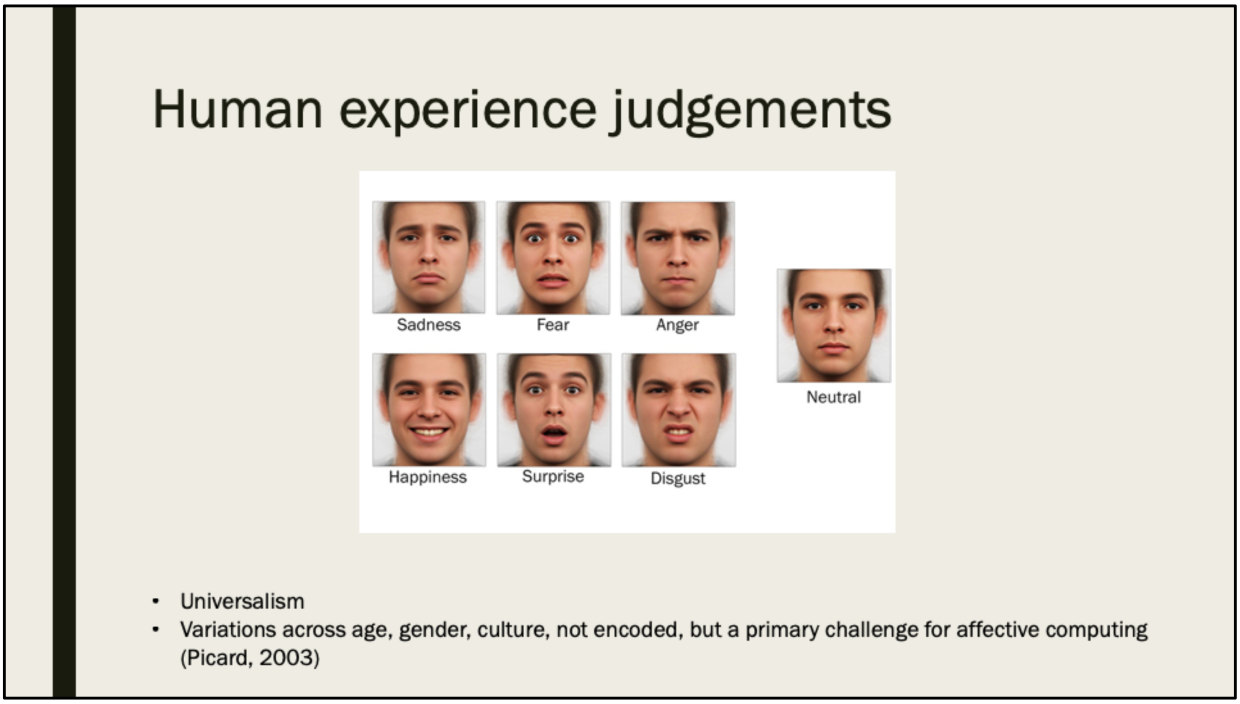
## Domain expertise

Sarkar, A., Morrison, C., Dorn, J. F., Bedi, R., Steinheimer, S., Boisvert, J.,...Lindley, S. (2016). Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision. In*Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI'16*(pp. 261–271). New York, New York, USA: ACM Press. https://doi.org/10.1145/2858036.2858199

Chen, N. (2016). Challenges of Applying Machine Learning to Qualitative Coding.*ACM SIGCHI Workshop on Human-Centered Machine Learning*. Retrieved from http://hcml2016.goldsmithsdigital.com/program/

- Concepts may have unclear definitions
- Inter-rater variability (previous experience, training, methods and heuristics used for labelling)
- Access to adequate experts poses logistical challenges, e.g., quorum for averaging

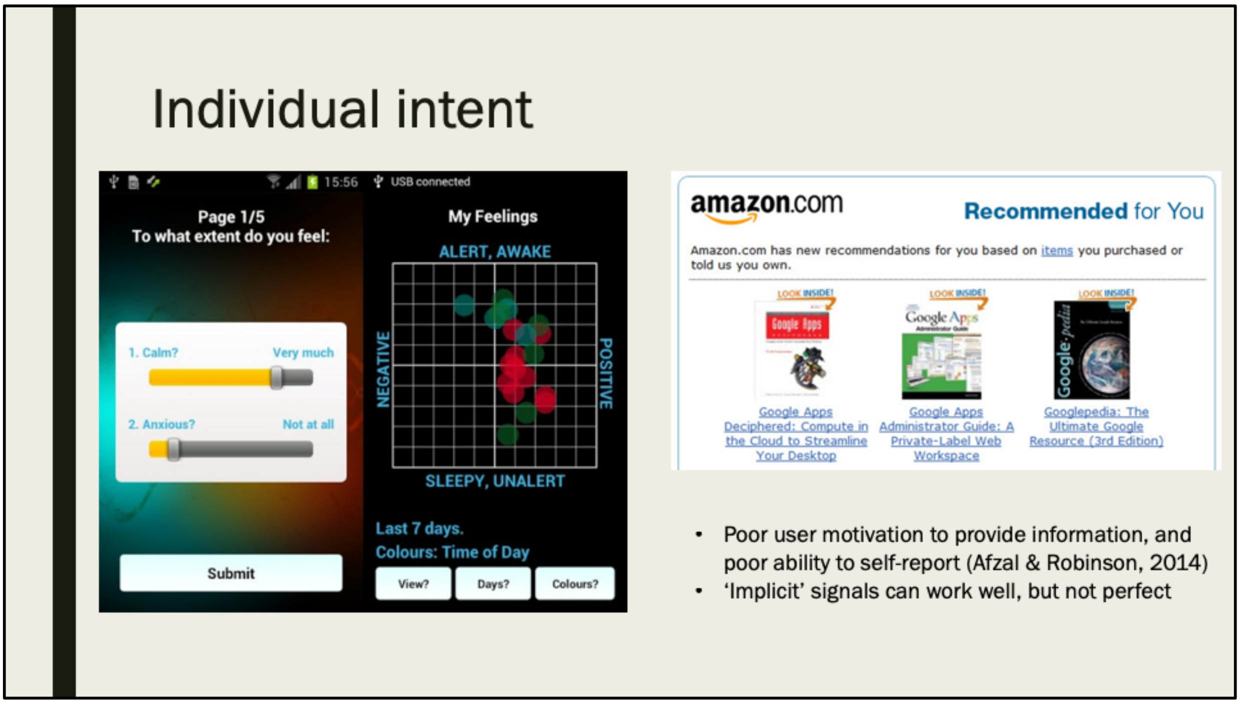
*Domain expertise* judgments rely on labellers' recognised expertise in a particular area. Two example are multiple sclerosis assessment through the analysis of patient videos (Sarkar et al., 2016), and assigning qualitative codes to social science research data (Chen, 2016). Despite these judgments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



# Human experience judgements

- Universalism
- Variations across age, gender, culture, not encoded, but a primary challenge for affective computing (Picard, 2003)

*Human experience* judgments are those that aim to capture some universal aspect of the 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 a tenuous assumption that any given person is acting as a representative judge on behalf of 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

Amazon.com has new recommendations for you based on items you purchased or told us you own.

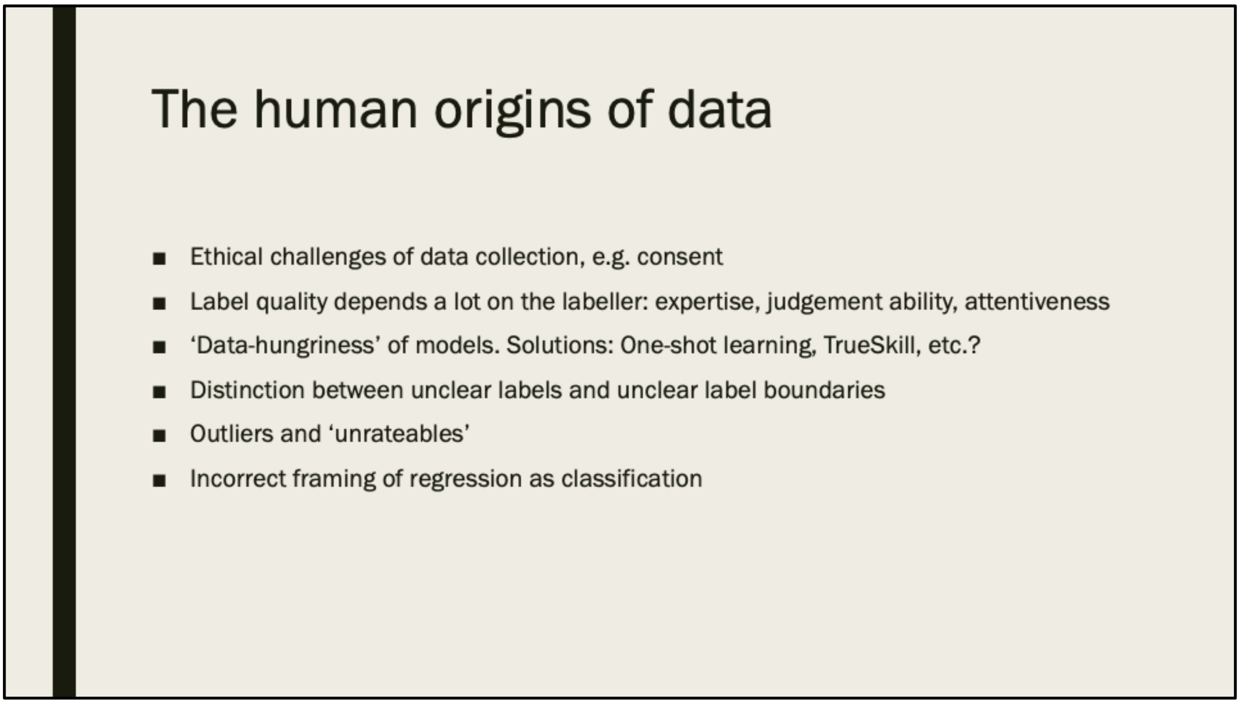
Google Apps Deciphered: Compute in the Cloud to Streamline Your Desktop

Google Apps Administrator Guide: A Private-Label Web Workspace

Googlepedia: The Ultimate Google Resource (3rd Edition)

- Poor user motivation to provide information, and poor ability to self-report (Afzal & Robinson, 2014)
- ‘Implicit’ signals can work well, but not perfect

*Individual intent* judgments reflect personal feelings, desires, and attributes. Unlike the previous three categories, which appeal to different standards of objectivity (perceptual reality, objective expertise, and universality) these judgements are acknowledged to be inherently subjective because they model an individual. For example, 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 motivated to 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.



# The human origins of data

- Ethical challenges of data collection, e.g. consent
- Label quality depends a lot on the labeller: expertise, judgement ability, attentiveness
- 'Data-hungriness' of models. Solutions: One-shot learning, TrueSkill, etc.?
- Distinction between unclear labels and unclear label boundaries
- Outliers and 'unrateables'
- Incorrect framing of regression as classification

Even before it has been labelled, training data reflects human judgements and priorities. Modern supervised learning techniques require large training sets to build stable models, but the scale of data acquisition can raise ethical challenges, including consent to use data for new purposes, protected categories of data such as clinical patient data, and privacy and anonymity concerns which make it difficult to aggregate data.

While labeling data is a seemingly simple task, it is actuallyfraught with problems (e.g., [9, 19, 26]). Labels reflect a labeler'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's expertise or familiarity with the concept or data, theirj udgment ability and attentiveness during labeling, and the ambiguity and changing distribution of the data itself.

Moreover, some applications require fast convergence. For instance, the TrueSkill system (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 way it was collected. From the machine learning perspective, there may not be a formal distinction between *examples* which cannot be placed exactly in the space of labels, and label *boundaries* which are not precise. However, they are very different from the perspective of a human labeller. Imprecise label boundaries may undermine labeller confidence throughout the entire labelling activity. Training examples may also pose problems because they are outliers, or simply unrateable. As noted by Chen (Chen,2016), outliers are typically discarded in quantitative analyses, but become the focus of attention in qualitative analyses. Examples that are unratable (perhaps because of data corruption or because they contain no meaningful information) may impair the labelling 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 problem for the purpose of labelling – it is easier to ask labellers to provide one of a discrete set of labels than a real number on a continuous scale. However, this can result in the unnecessary conceptualisation of examples as belonging to a set of discrete categories, which causes issues for examples on the boundaries of different categories. This is the problem faced by the Assess MS problem, detailed in the next section. Unclear concepts cause problems generally in precision, but less so for accuracy.

# Accommodating flexibility

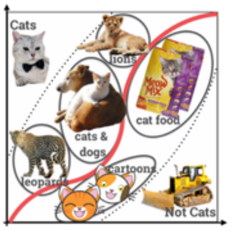
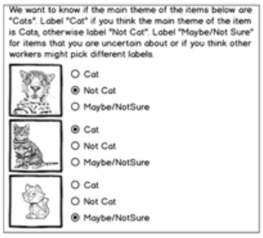


Figure 1. Revolt creates labels for unanimously labeled "certain" items (e.g., *cats* and *not cats*), and surfaces categories of "uncertain" items enriched with crowd feedback (e.g., *cats and dogs* and *cartoon cats* in the dotted middle region are annotated with crowd explanations). Rich structures allow label requesters to better understand concepts in the data and make post-hoc decisions on label boundaries (e.g., assigning *cats and dogs* to the *cats* label and *cartoon cats* to the *not cats* label) rather than providing crowd-workers with a priori label guidelines.

Revolt (Chee Chang et al., CHI 2017)



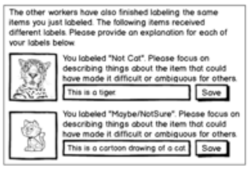
We want to know if the main theme of the items below are "Cats". Label "Cat" if you think the main theme of the item is Cats, otherwise label "Not Cat". Label "Maybe/Not Sure" for items that you are uncertain about or if you think other workers might pick different labels.

○ Cat
● Not Cat
○ Maybe/NotSure

● Cat
○ Not Cat
○ Maybe/NotSure

○ Cat
○ Not Cat
● Maybe/NotSure

Figure 3. Human Intelligence Task (HIT) interface for the Vote Stage. In addition to the predefined labels, crowdworkers can also select *Maybe/NotSure* when they were uncertain about the item.

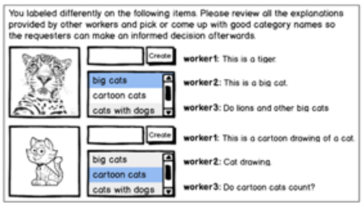


The other workers have also finished labeling the same items you just labeled. The following items received different labels. Please provide an explanation for each of your labels below.

You labeled "Not Cat". Please focus on describing things about the item that could have made it difficult or ambiguous for others.
This is a tiger

You labeled "Maybe/NotSure". Please focus on describing things about the item that could have made it difficult or ambiguous for others.
This is a cartoon drawing of a cat

Figure 4. Human Intelligence Task (HIT) interface for the Explain Stage. Crowdworkers enter a short description for each item that was labeled differently in the Vote Stage. They were informed that disagreement occurred, but not the distribution of different labels used.

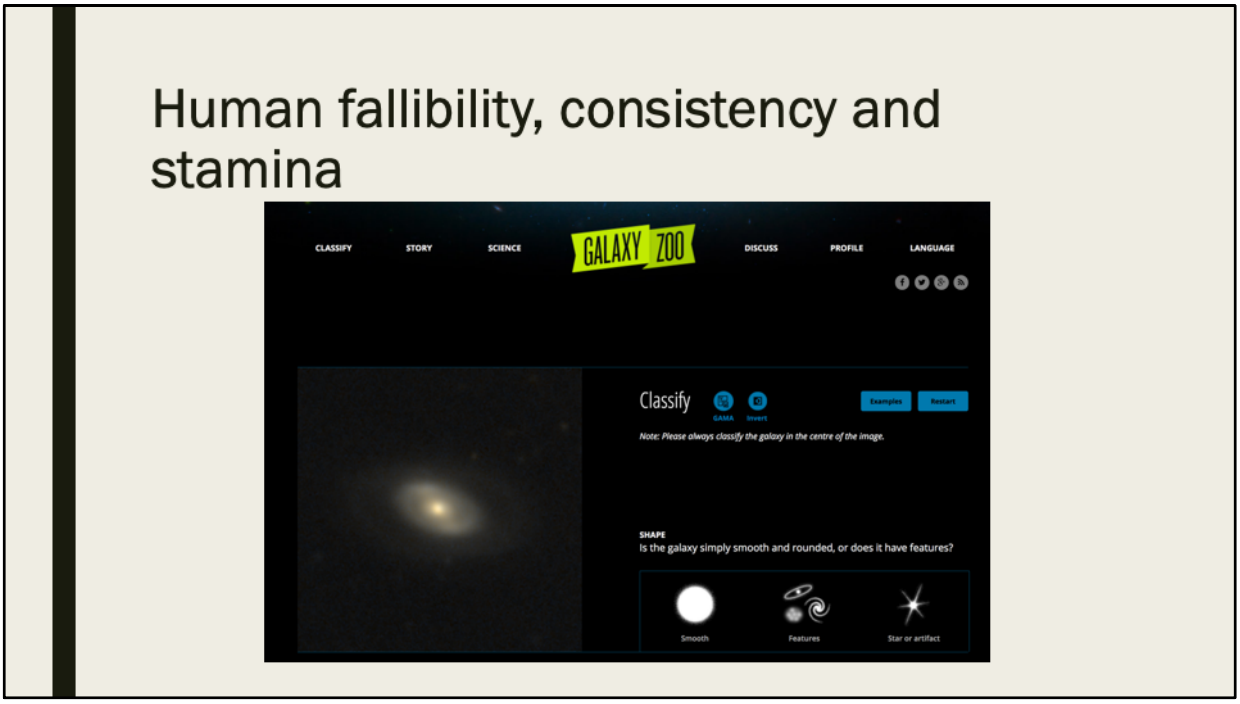


You labeled differently on the following items. Please review all the explanations provided by other workers and pick or come up with good category names so the requesters can make an informed decision afterwards.

**worker1**: This is a tiger.
**worker2**: This is a big cat.
**worker3**: Do lions and other big cats

**worker1**: This is a cartoon drawing of a cat.
**worker2**: Cat drawing.
**worker3**: Do cartoon cats count?

Figure 5. Human Intelligence Task (HIT) interface for the Categorize Stage. Crowdworkers select or create categories for items that were labeled differently in the Vote Stage, based on explanations from all three crowdworkers in the same group.



# Human fallibility, consistency and stamina

Note: Please always classify the galaxy in the centre of the image.

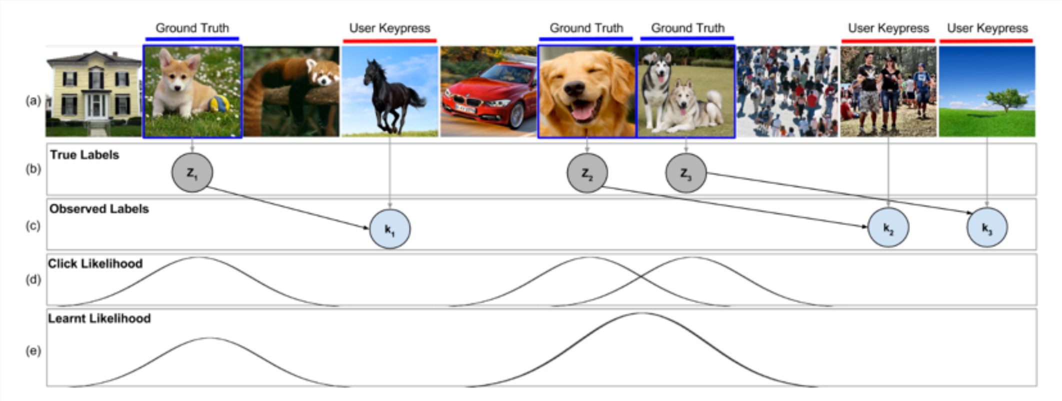
SHAPE

Is the galaxy simply smooth and rounded, or does it have features?

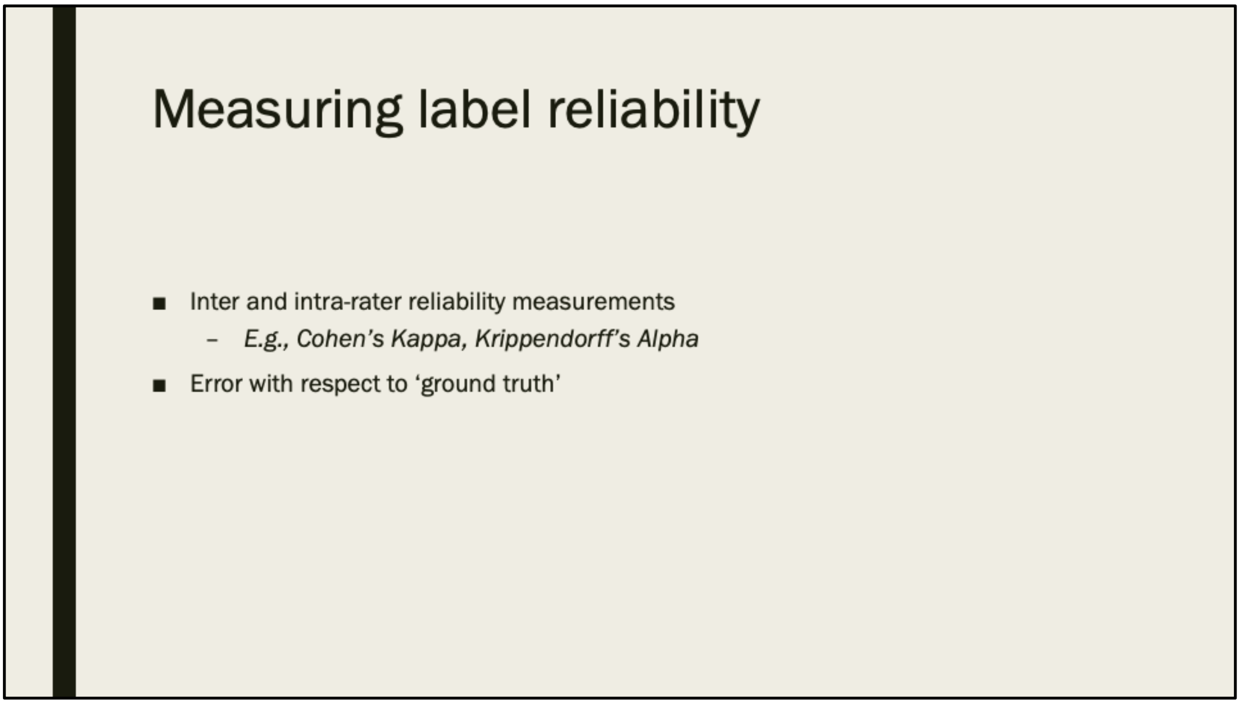
Humans are fallible. If there are large amounts of data to be labelled, the quality of judgements can be impaired as the labeller becomes tired. In the Assess MS projectdescribed in the next section, neurologists would spend an entire workday, sometimes two, continuously labelling short video clips (Sarkar 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.

# Embracing error to improve speed



Krishna et al., 2016 (Embracing Error to Enable Rapid Crowdsourcing. CHI 2016)



## Measuring label reliability

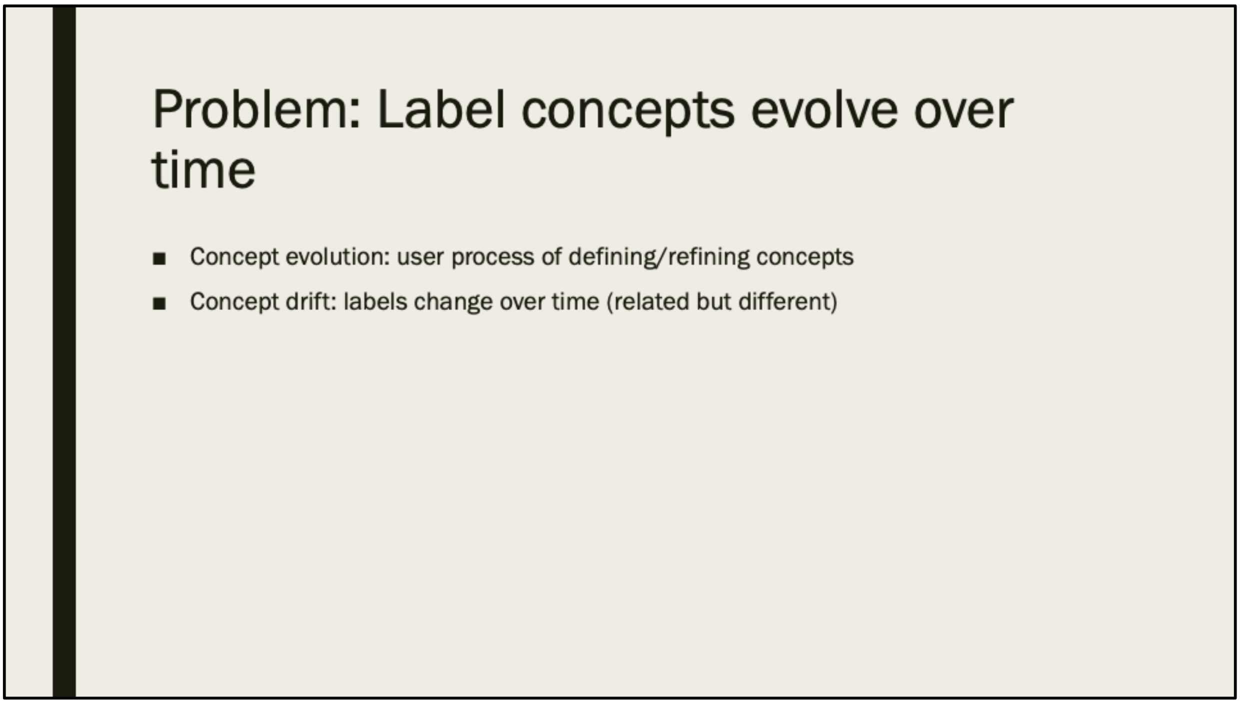
- Inter and intra-rater reliability measurements
  - *E.g., Cohen's Kappa, Krippendorff's Alpha*
- Error with respect to 'ground truth'

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



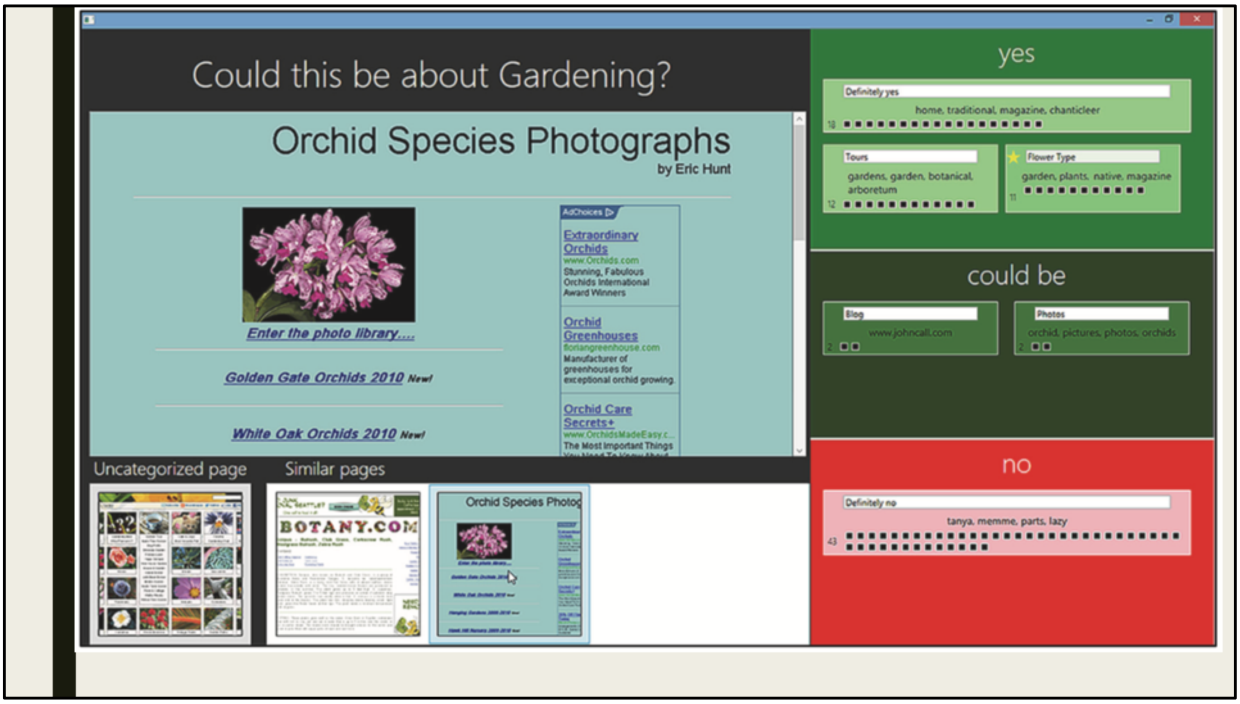
# Problem: Label concepts evolve over time

- Concept evolution: user process of defining/refining concepts
- Concept drift: labels change over time (related but different)

(Mostly from the paper)

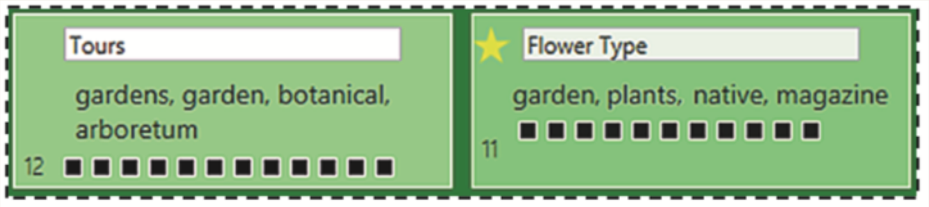
This paper addresses a distinct problem in labeling data that we refer to as *concept evolution*. Concept evolution refers to the labeler’s process of defining and refining a concept in their minds, and can result in different labels being applied to similar items due to changes in the labeler’s notion of theunderlying concept. The paper presents a formative study where the authors found that people labeling a set of web pages twice with a four-week gap between labeling sessions were,on average, only 81% consistent with their initial labels. This inconsistency in labeling similar items can be harmful to machine learning, which is fundamentally based on the ideathat similar inputs should have similar outputs

A separate problem in data labeling is *concept drift*, where the underlying data is fundamentally changingover time [29]. An example of concept drift is a news recommender that attempts to recommend the most interesting recent news. Here, the concept of *interesting* may remain the same over time, but the data (in this case the news) is constantly drifting as a result of changing current events. Most solutions to concept drift model concepts temporally, such as by discarding or weighting information according 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 help a *user* refine their own idea of a concept, a problem which may be exacerbated in the presence of concept drift.


Could this be about Gardening?
Orchid Species Photographs
by Eric Hunt
Enter the photo library....
Golden Gate Orchids 2010 New!
White Oak Orchids 2010 New!
AdChoices
Extraordinary Orchids
www.Orchids.com
Stunning, Fabulous Orchids International Award Winners
Orchid Greenhouses
floriangreenhouse.com
Manufacturer of greenhouses for exceptional orchid growing.
Orchid Care Secrets+
www.OrchidsMadeEasy.c...
The Most Important Things
Uncategorized page
Similar pages
BOTANY.COM
Orchid Species Photog
yes
Definitely yes
home, traditional, magazine, chanticleer
18
Tours
gardens, garden, botanical, arboretum
12
Flower Type
garden, plants, native, magazine
11
could be
Blog
www.johncall.com
2
Photos
orchid, pictures, photos, orchids
2
no
Definitely no
tanya, memme, parts, lazy
43

we introduce *structured labelling* (Figure 1), a novel interaction technique for helping people define and refine their concepts as they label data. Structured labeling allows people to organize their concept 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 uncategorized page (top left) and can drag it to an existing group (right), or create a new group for it. The thumbnails (bottom left) show similar pages in the dataset to help the user gauge whether creating a new group is warranted.**



Tours

gardens, garden, botanical, arboretum

12

Flower Type

garden, plants, native, magazine

11

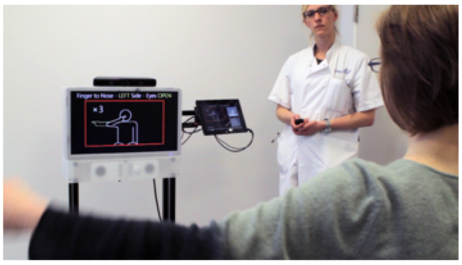
**Our assisted structuring tool provides users with automatic summaries of each group’s contents (below the user-supplied tag area) and recommends a group for the current item via an animation and yellow star indicator. The black squares indicate how many items are in each group.**

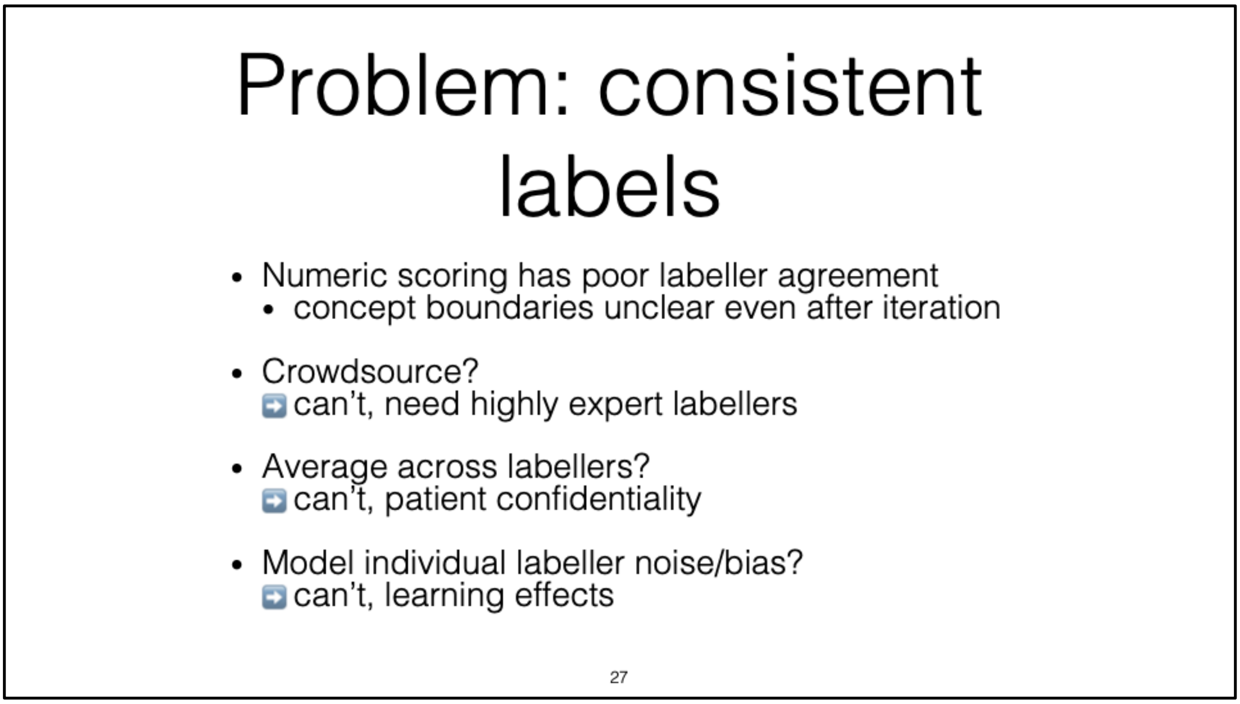
# SORTABLE

Case study II

# Assess MS

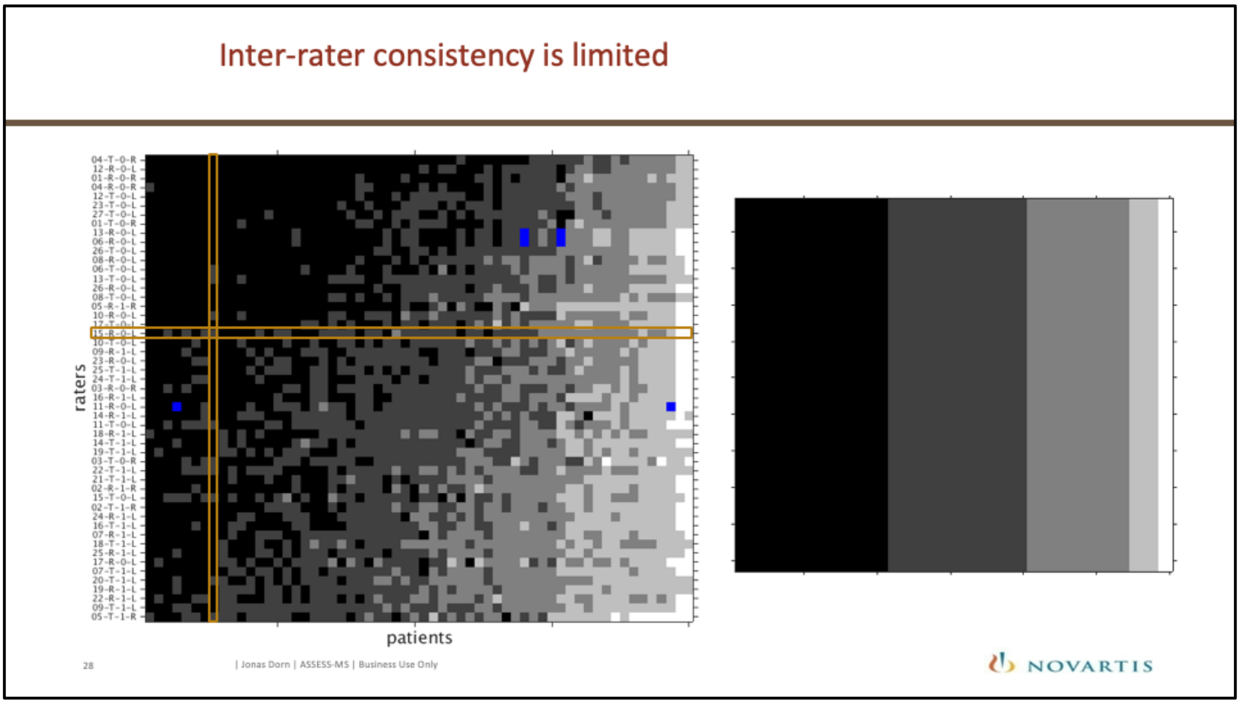
- Aim: a more consistent way of quantifying progression of motor illness in multiple sclerosis
- Input: Kinect RGB + depth videos of standard clinical movements
- Output: a standardised clinical disability score





# Problem: consistent labels

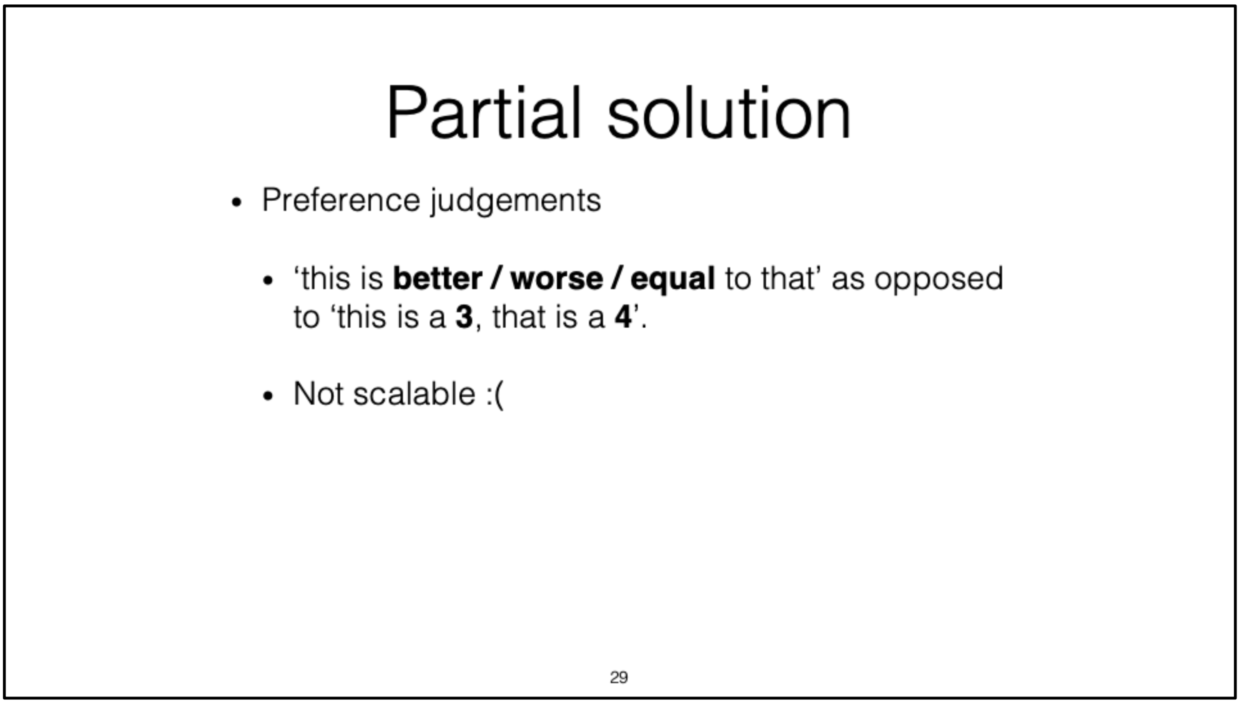
- Numeric scoring has poor labeller agreement
    - concept boundaries unclear even after iteration
- Crowdsource?
  ➡ can’t, need highly expert labellers
- Average across labellers?
  ➡ can’t, patient confidentiality
- Model individual labeller noise/bias?
  ➡ can’t, learning effects



## Inter-rater consistency is limited

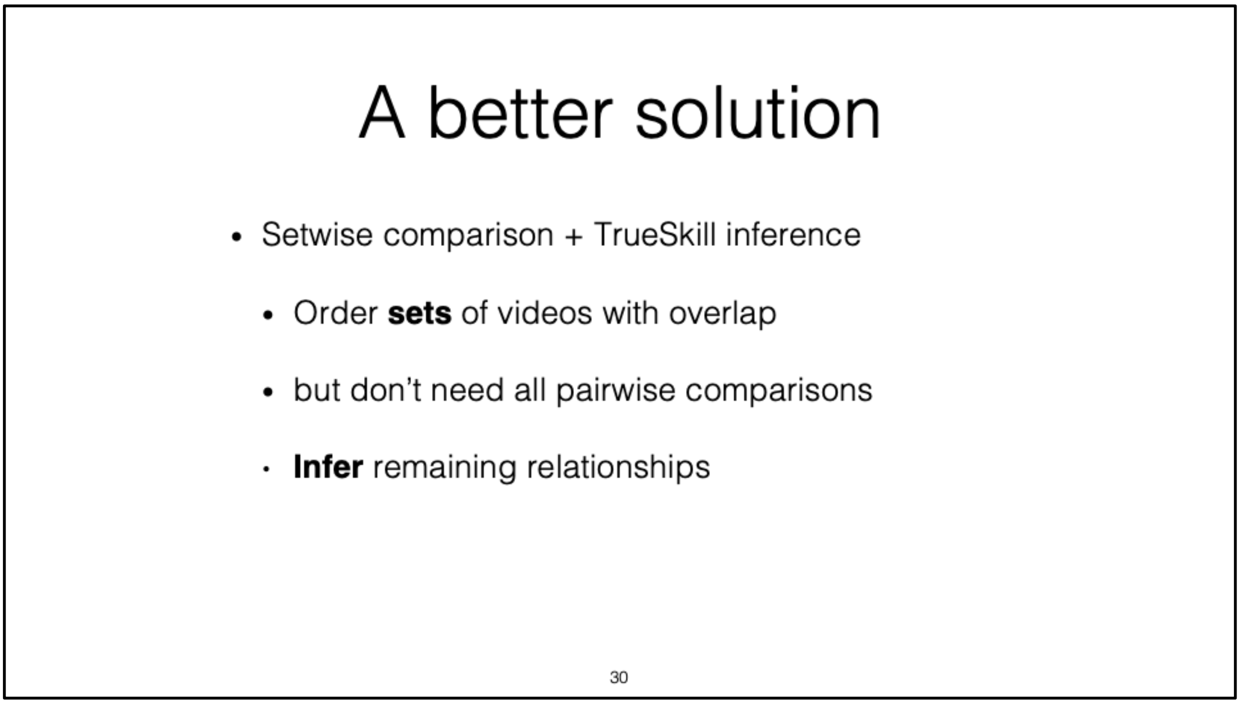
28 | Jonas Dorn | ASSESS-MS | Business Use Only

NOVARTIS



# Partial solution

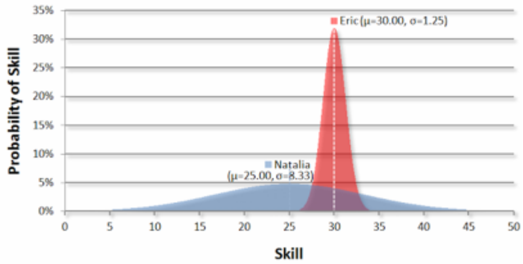
- Preference judgements
  - ‘this is **better / worse / equal** to that’ as opposed to ‘this is a **3**, that is a **4**’.
  - Not scalable :(



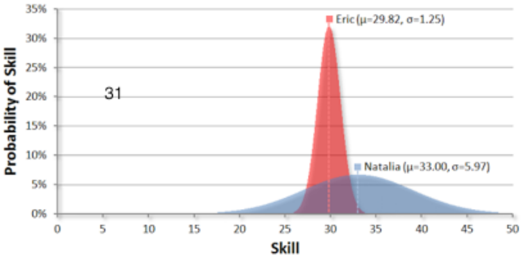
# A better solution

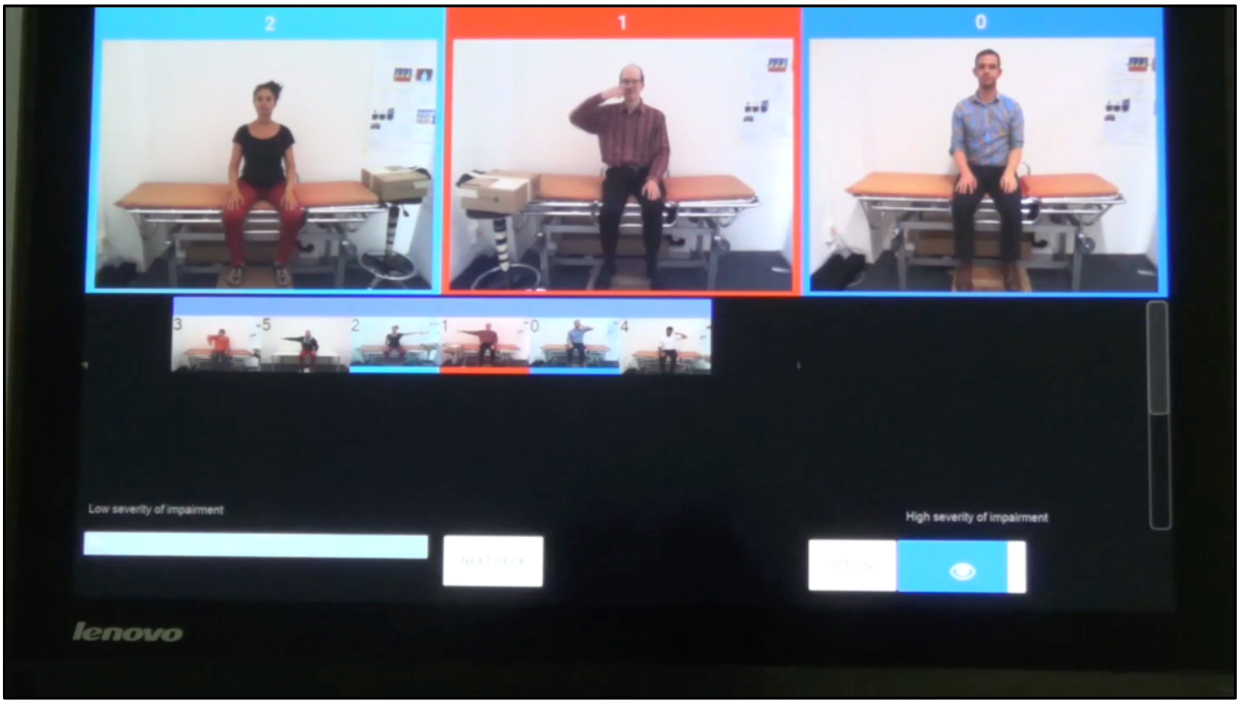
- Setwise comparison + TrueSkill inference
  - Order **sets** of videos with overlap
  - but don't need all pairwise comparisons
  - **Infer** remaining relationships

Prior

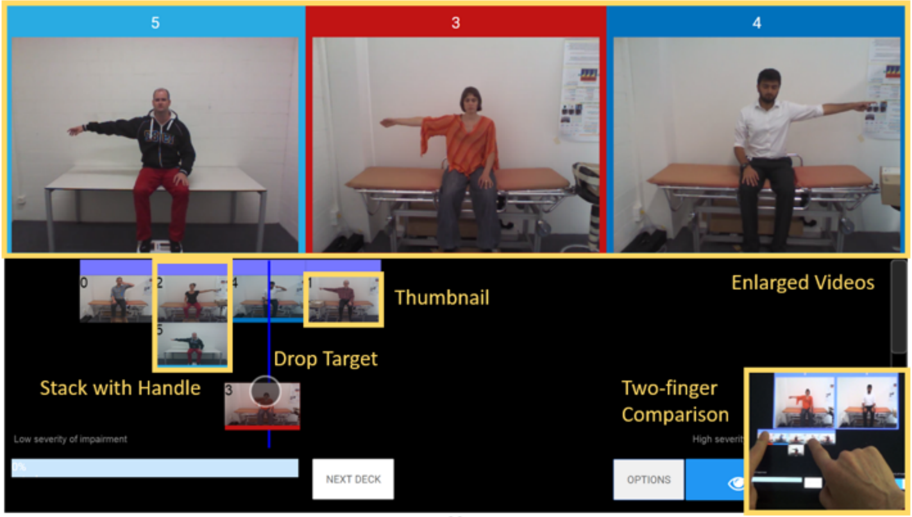


After Natalia wins

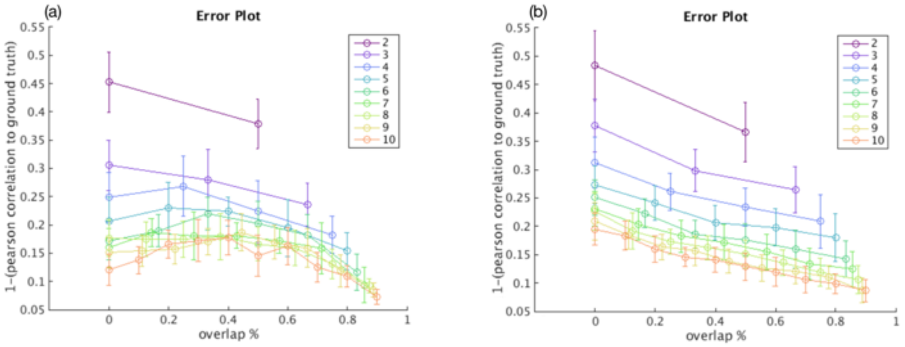


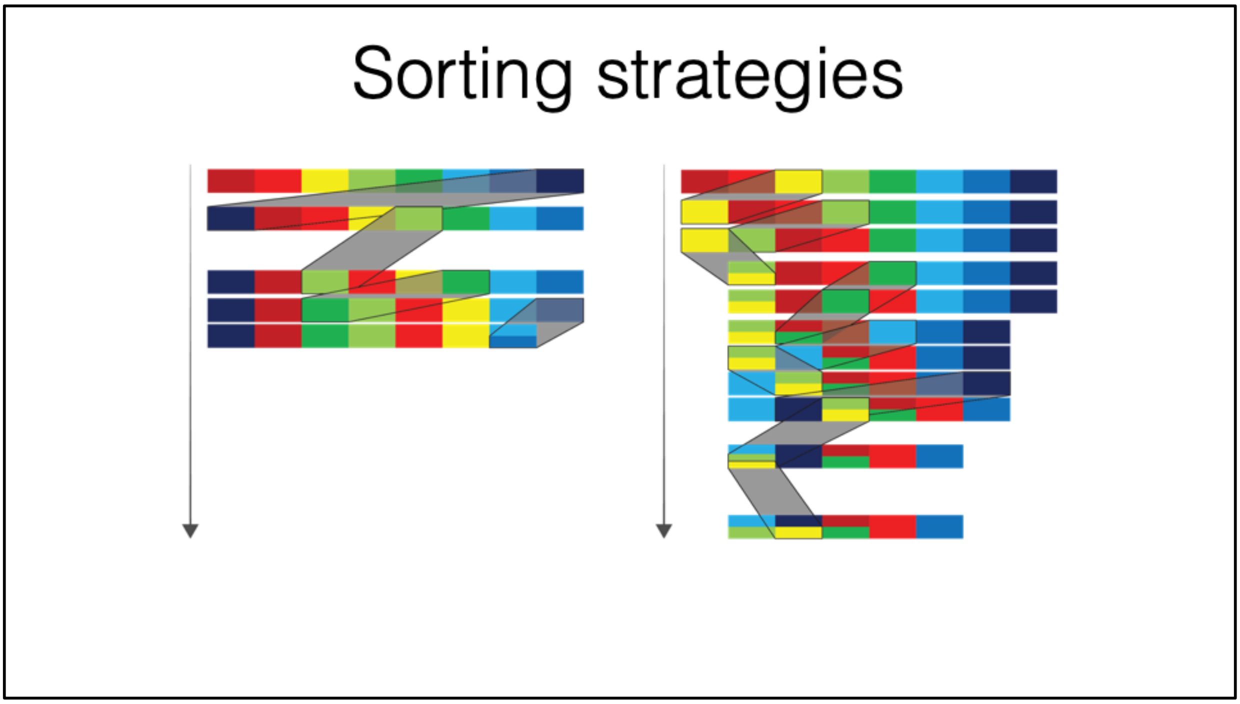

2
1
0
3
5
2
1
0
4
Low severity of impairment
NEXT DECK
High severity of impairment
OPTIONS
lenovo

# SorTable
## an interface for setwise comparison

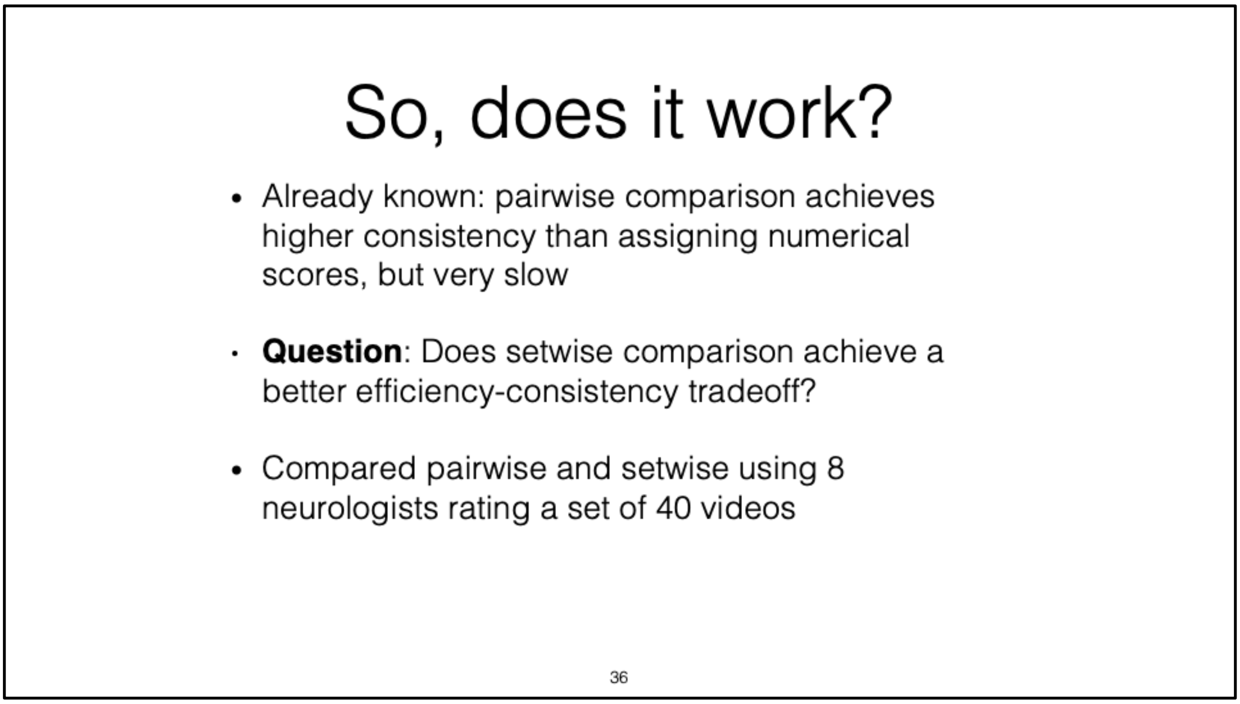


# Choosing deck size and overlap



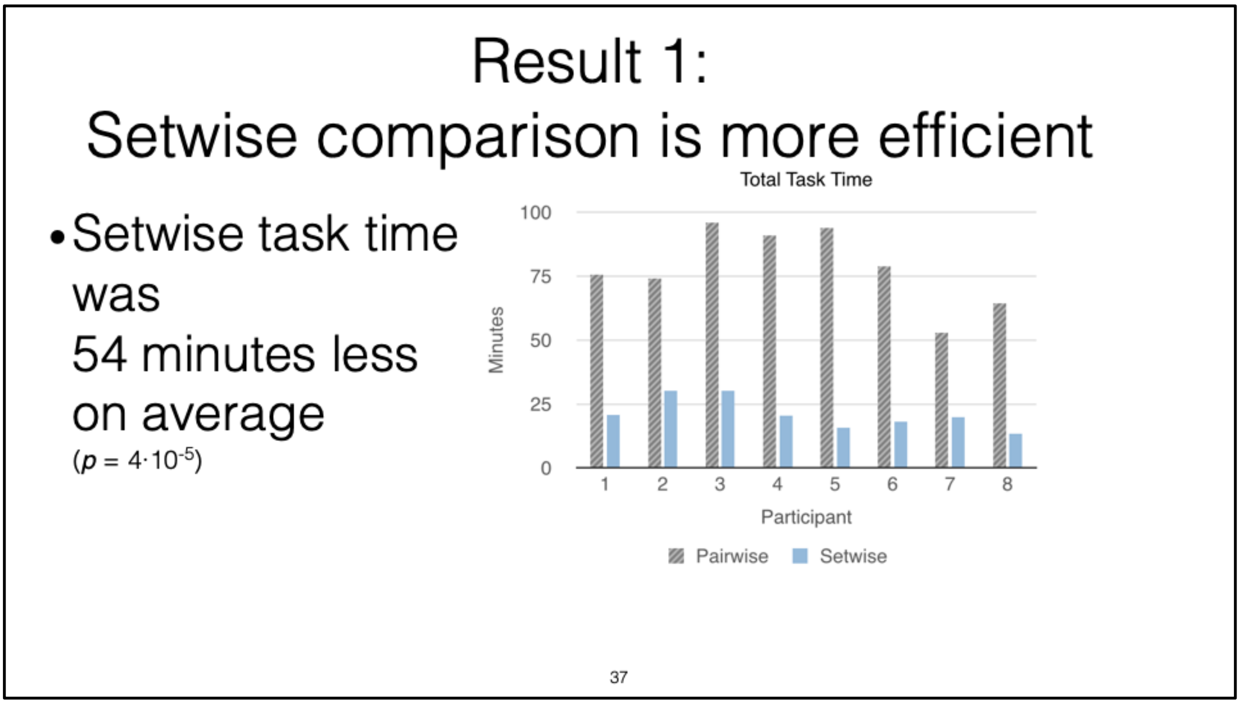


# Sorting strategies



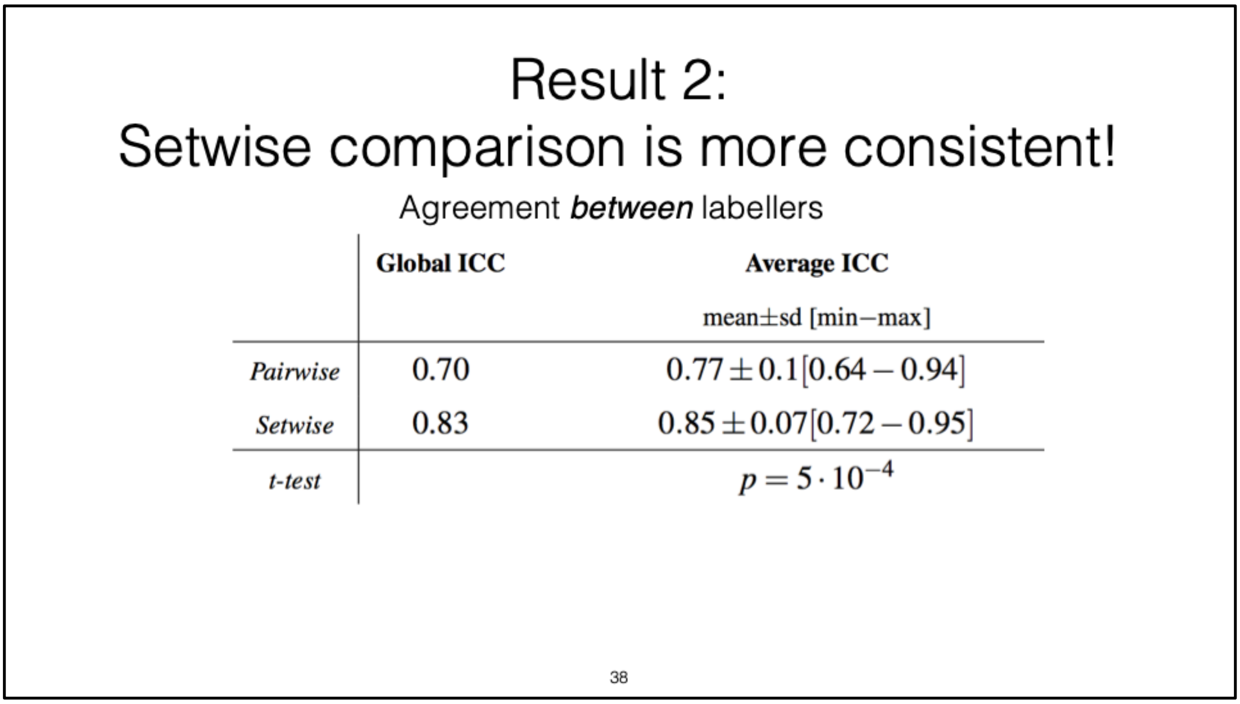
# So, does it work?

- 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



# Result 1:
# Setwise comparison is more efficient

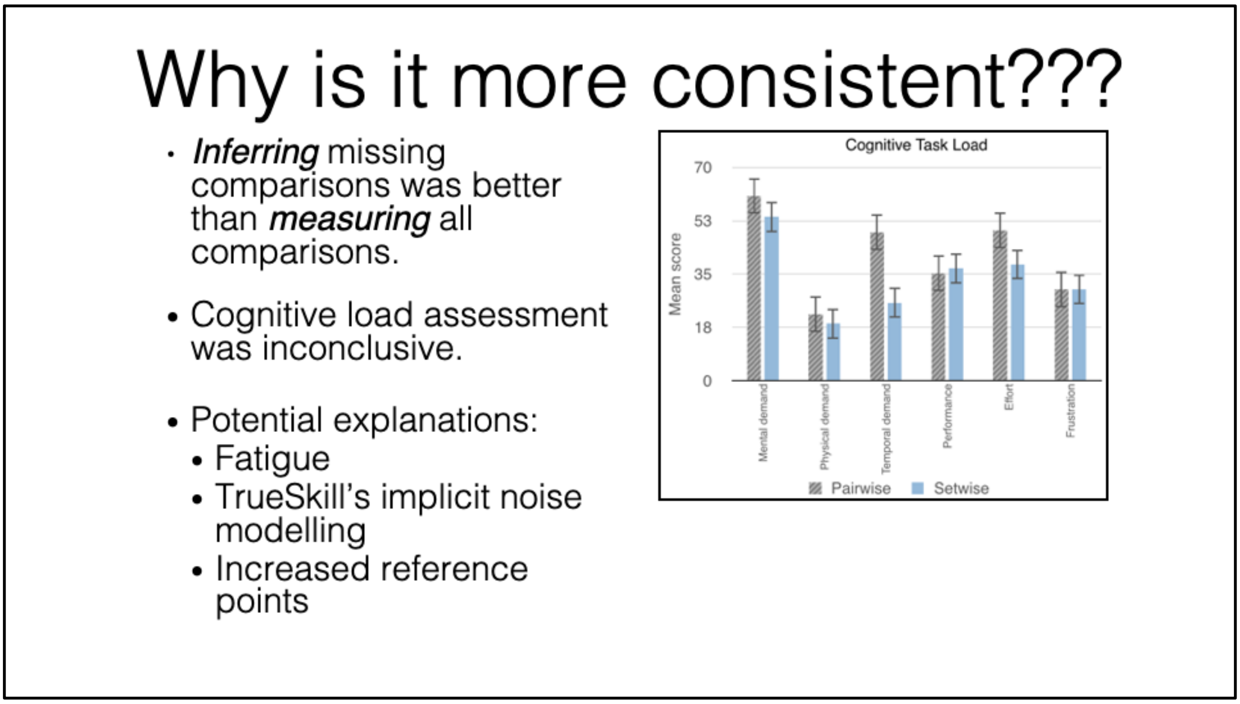
- Setwise task time was
  54 minutes less
  on average
  ($p = 4 \cdot 10^{-5}$)



# Result 2:
# Setwise comparison is more consistent!

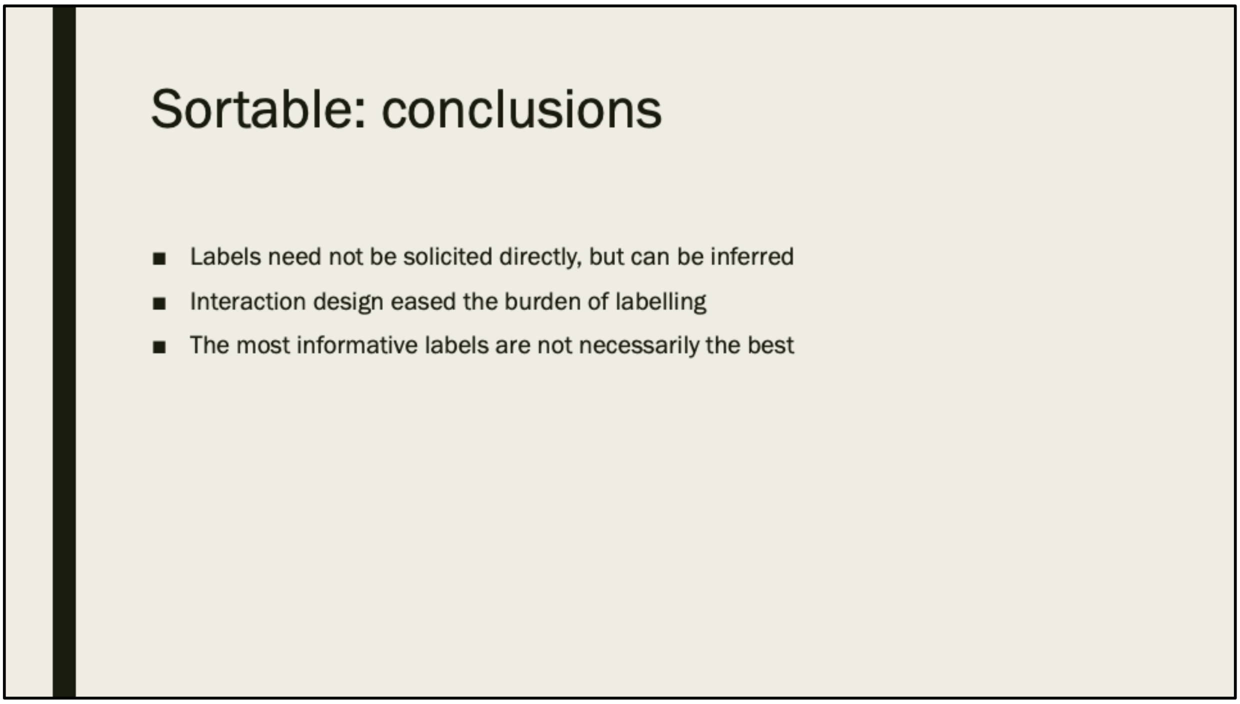
Agreement ***between*** labellers

| | **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 modelling
  - Increased reference points



## Sortable: conclusions

- Labels need not be solicited directly, but can be inferred
- Interaction design eased the burden of labelling
- The most informative labels are not necessarily the best

We reframed the problem so that users were not providing labels directly, but providing information from which labels could be reconstructed. In this way, we could build upon strong human capability in relative judgement and still provide the classification 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, achieving three benefits for the users. First, the presentation of videos in sets builds upon human short-term memory to make multiple comparisons at once. Second, the ability to create stacks to indicate that videos are the same can substantially reduce the number of comparisons the labeller needs to make when sorting. Third, SorTable facilitates mixed-strategy sorting, including the automatic display of the left and right neighbours of the currently selected video, and the ability to compare any two videos with a two-finger gesture. All interactions are touch based.

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

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