Automatically assigning UCDs using machine learning

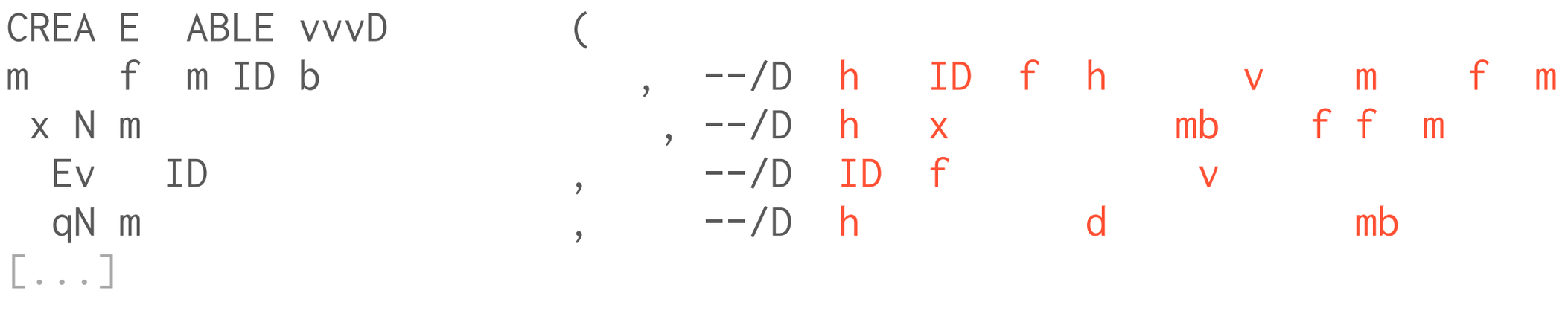
Norman Gray

University of Glasgow, UK

Co-SADIE Tech Forum, Trieste, 2014 March 12

Funded by

ROE



what are the UCD1+ for these columns?

--/D is description
--/U units
--/F FITS table TTYPE
--/K FITS keywords

schemas: what we have

- *UCD1* on many columns
- *HIERARCH tags* on some columns
- *units* on many columns
- *comments* on almost all columns

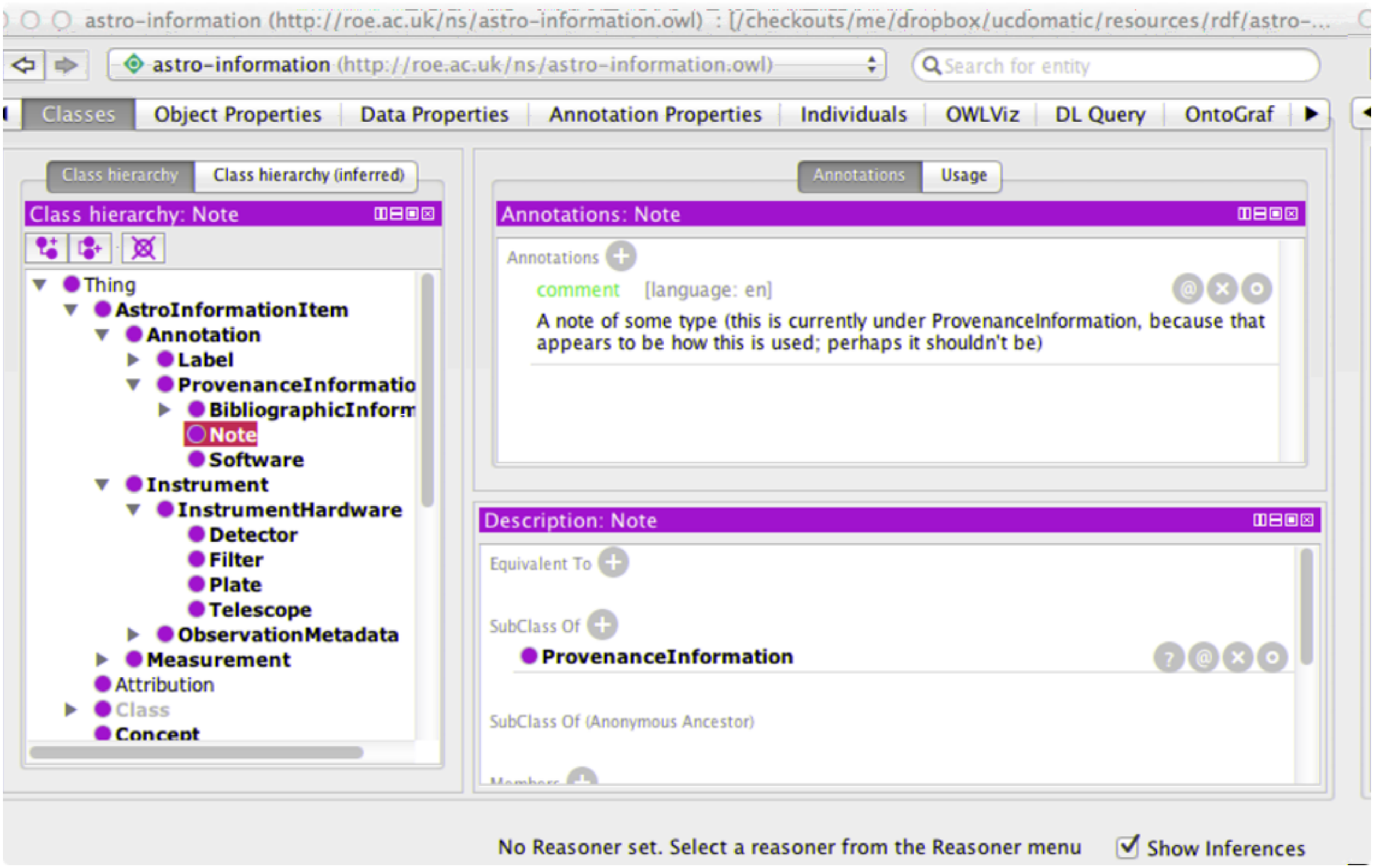
norman gray

‘comments’ are free text; hard to see it’s much use

semantics: ucd++

- 1. create an ontology of astronomical information
- 2. heuristically assign columns to elements of that
- 3. associate a UCD1+ with each class
- 4. read off the UCD1+
- 5. profit!

1. is subjective, but flexible
3 & 4 are straightforward (technically fiddly, that's all)
2. turns out to be hard



Hand-done ontology, inspired by the UCD1 & UCD1+ structures
...but not mechanically derived from either

```
:C          P          Eq          RA
   df :   bC     Of
    <h    ://    .   . k/   /   d1+ #    . q.   > ;
   w : q   v      C         d1:    _ q_    .

:B b          h   D    J
   w : q   v      C
    <h    ://    .   . k/   /   d1+ #m    .b b.          >,
         d1:   f   _         .

   <h    ://    .   . k/   /          - #IN¶.FIL >
      df :   bC     Of :F        .
```

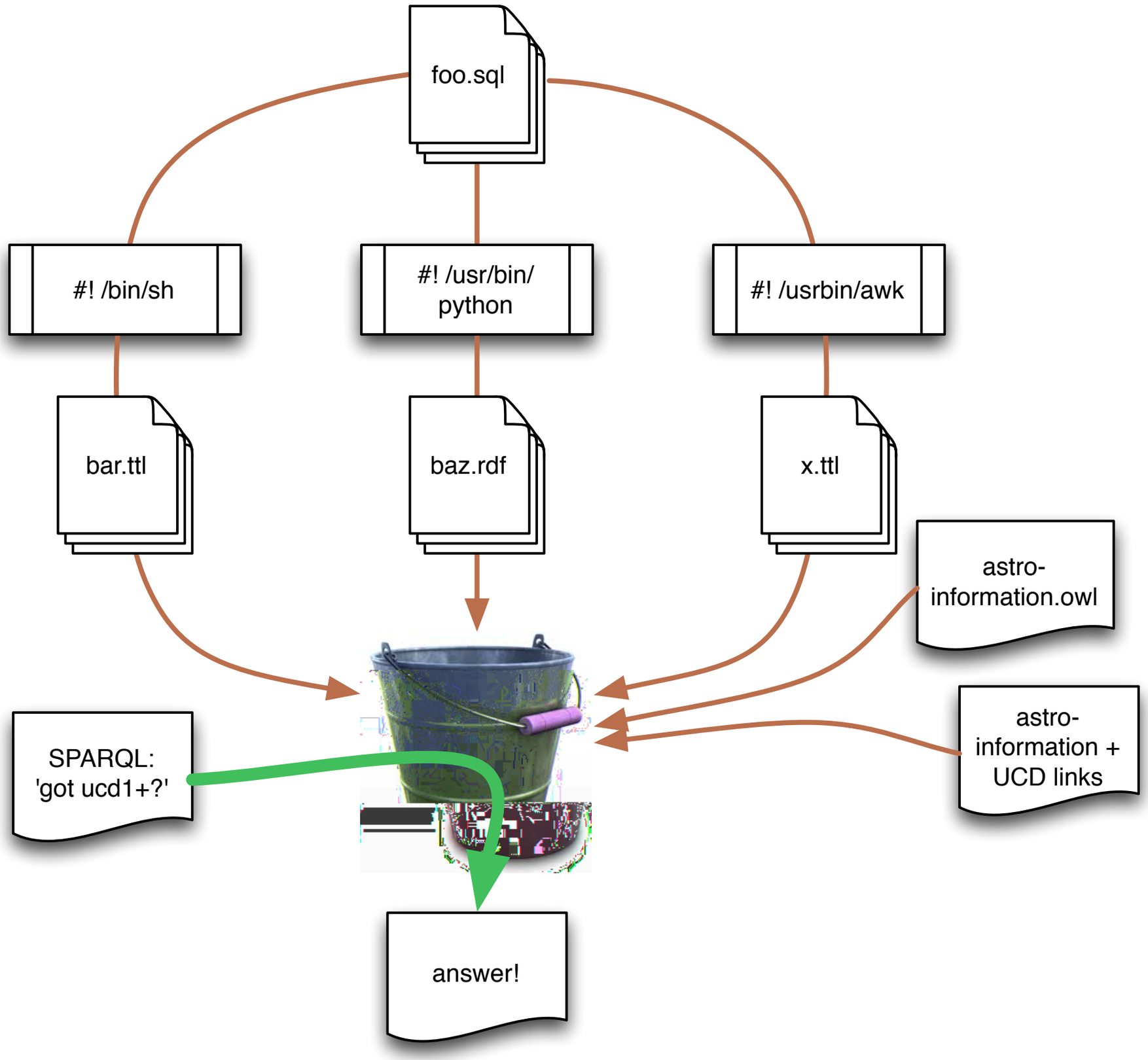
Looks messy, but it’s just a set of very simple ‘isA’ or ‘subClassOf’ relationships

anything else?

- There is some units information
- ...which *eventually* turns out not to be a lot of help

norman gray

Might still be useful for consistency checking



Extract information any way you can
Put it all in the bucket
Let the reasoner sort out the mess, then ask questions
Good idea, and best way to manage heterogeneity
...somewhat overwhelmed by fiddliness...
not quite enough information for what we want

let's have another look at those comments...

‘Intrinsic rms in H-band’

⇒ . ; m.IR.H

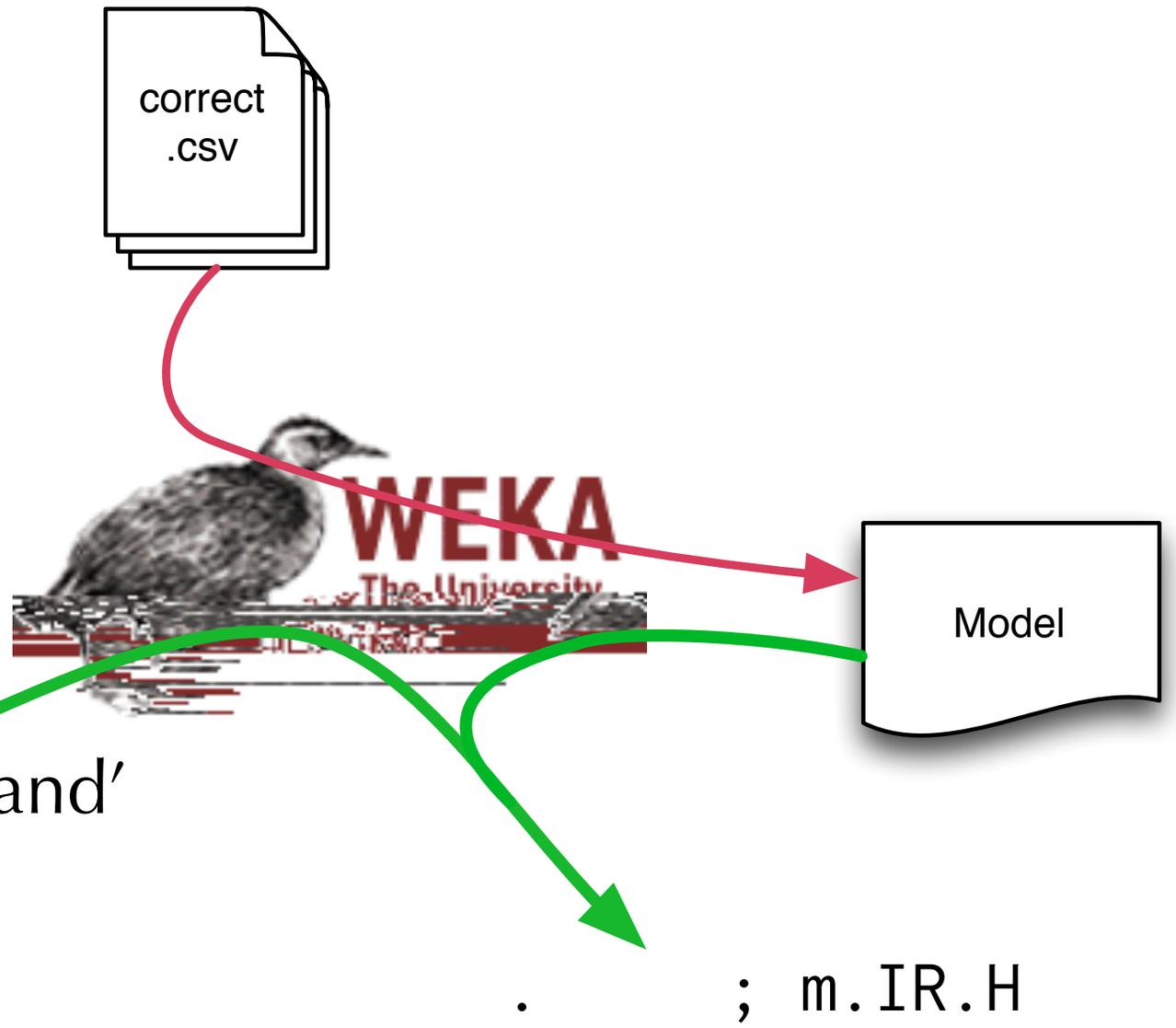
Classification of variability in this band’

⇒ m . d . ; .v

‘Angular separation between neighbours’

⇒ . D

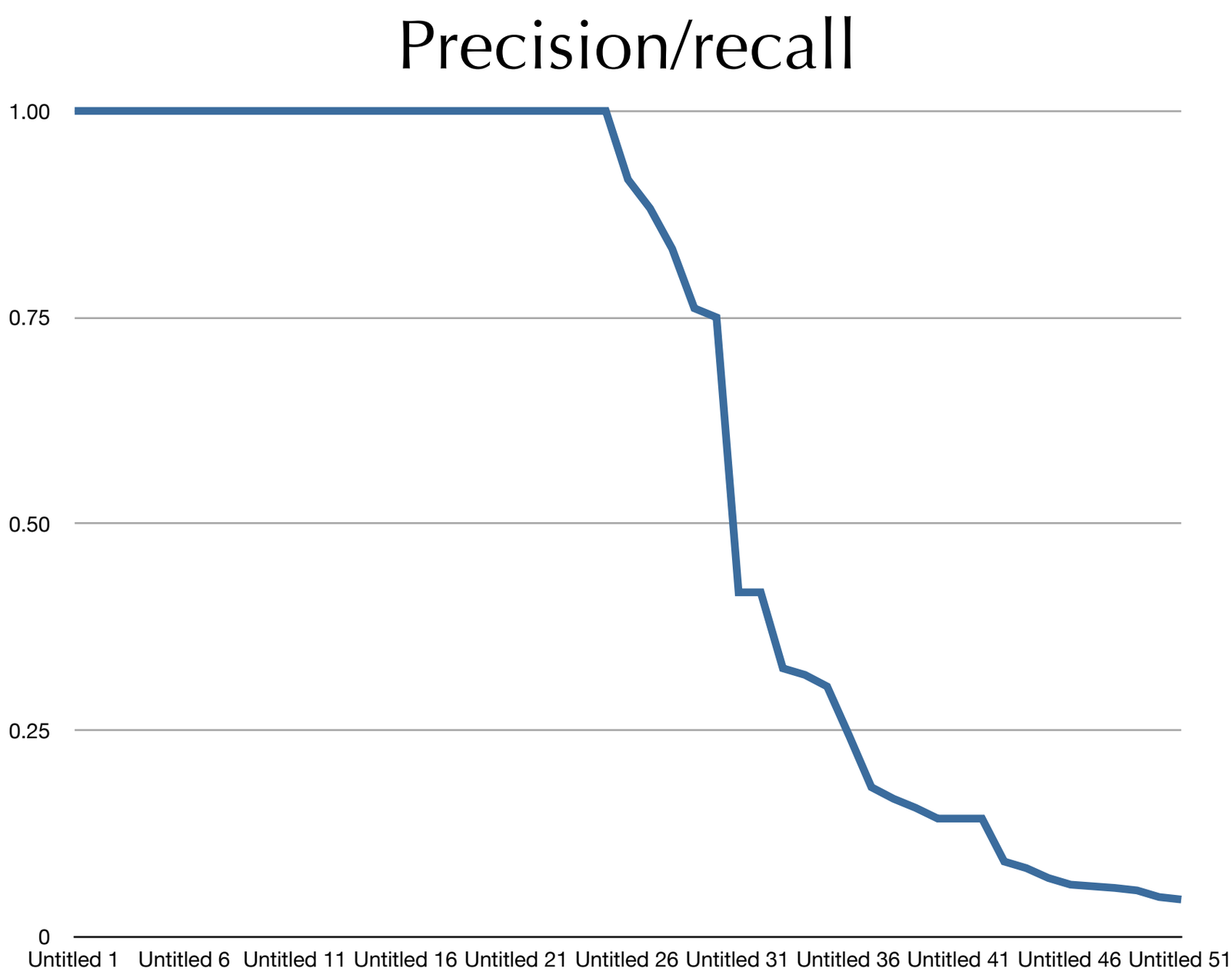
Weka is a Machine Learning toolkit



‘Intrinsic rms in H-band’

. ; m.IR.H

Start with a ‘training set’ of known-good assignments



Based on cross-validation
Precision/recall related to confidence of classification
Some clearly very good, but falls off rapidly
Sensitive to training set; haven't experimented with different algorithms and training sets

- use other features in input
- use other features (units/dimensions) to veto assignments
- enlarge training set (might be quite biased right now)
- package and release