

Automatically assigning UCDs using machine learning

Norman Gray

University of Glasgow, UK

Co-SADIE Tech Forum, Trieste, 2014 March 12



University
of Glasgow

Funded by



ROE

```

CREA E ABLE vvvD
m f m ID b
x N m
Ev ID
qN m
[...]
(
, --/D h ID f h v m f m
, --/D h x mb f f m
, --/D ID f v
, --/D h d mb
)

```

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

- 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!

norman gray

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

astro-information (http://roe.ac.uk/ns/astro-information.owl) : [/checkouts/me/dropbox/ucdomatic/resources/rdf/astro-...

astro-information (http://roe.ac.uk/ns/astro-information.owl) Search for entity

Classes | Object Properties | Data Properties | Annotation Properties | Individuals | OWLViz | DL Query | OntoGraf

Class hierarchy | Class hierarchy (inferred)

Class hierarchy: Note

- Thing
 - AstroInformationItem
 - Annotation
 - Label
 - ProvenanceInformation
 - BibliographicInform
 - Note**
 - Software
 - Instrument
 - InstrumentHardware
 - Detector
 - Filter
 - Plate
 - Telescope
 - ObservationMetadata
 - Measurement
 - Attribution
 - Class
 - Concent

Annotations | Usage

Annotations: Note

Annotations +

comment [language: en] @ x o

A note of some type (this is currently under ProvenanceInformation, because that appears to be how this is used; perhaps it shouldn't be)

Description: Note

Equivalent To +

SubClass Of +

- ProvenanceInformation** ? @ x o

SubClass Of (Anonymous Ancestor)

Members +

No Reasoner set. Select a reasoner from the Reasoner menu Show Inferences

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      .

```

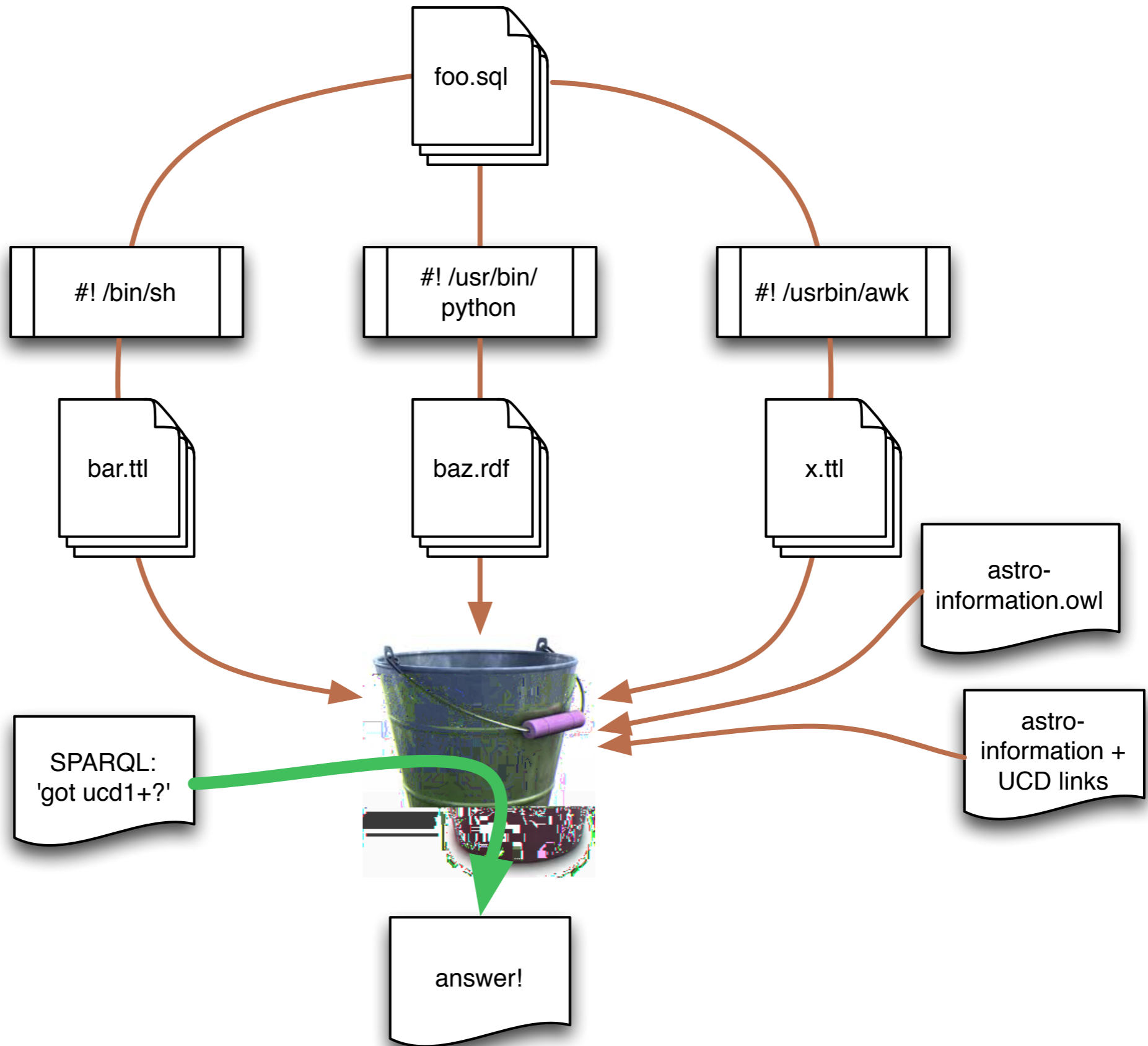
Looks messy, but it's just a set of very simple 'isA' or 'subClassOf' relationships

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...

want...

'Intrinsic rms in H-band'

\Rightarrow . ; m.IR.H

'Classification of variability in this band'

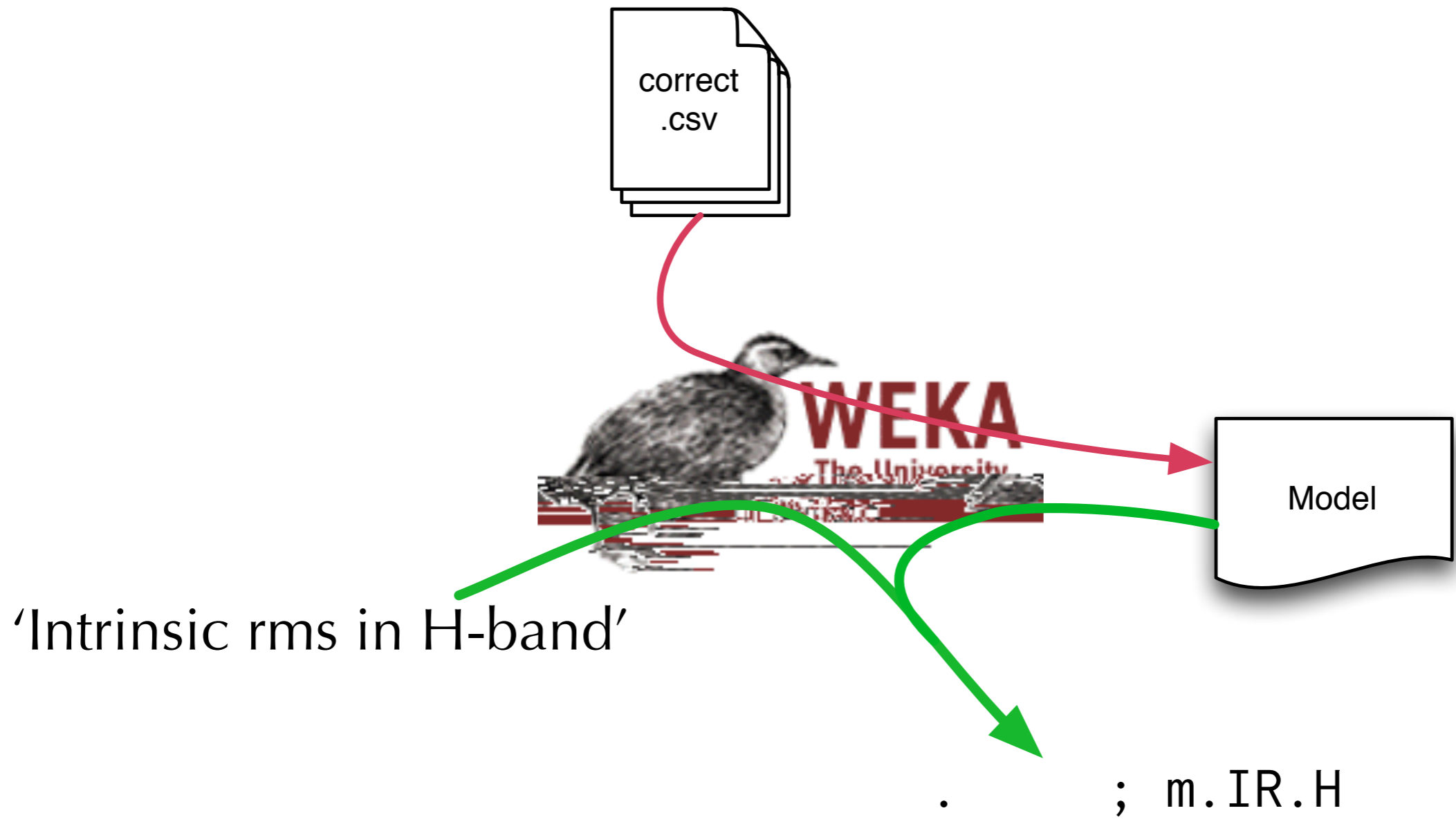
\Rightarrow m . d . ; .v

'Angular separation between neighbours'

\Rightarrow . D

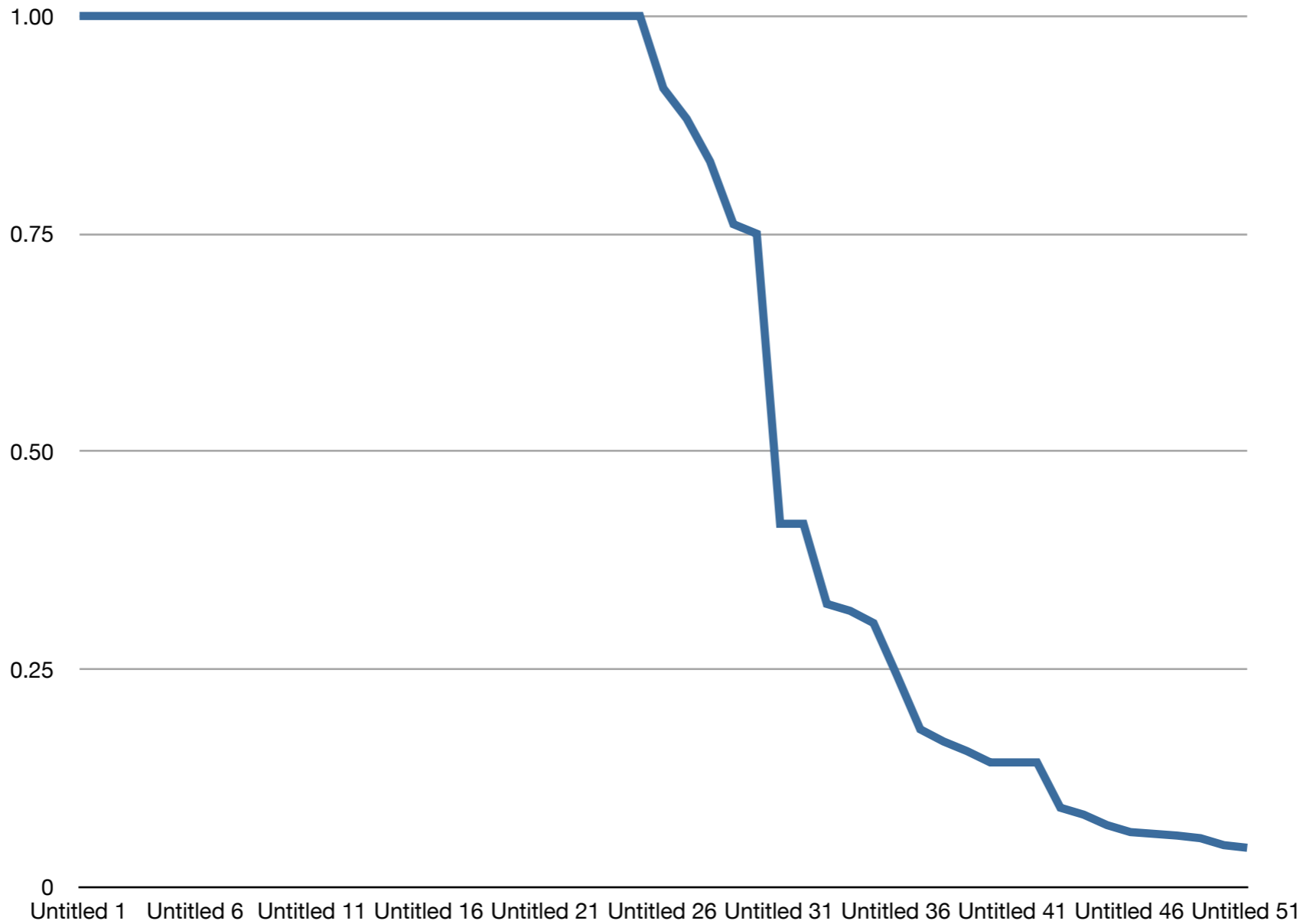


Weka is a Machine Learning toolkit



Start with a 'training set' of known-good assignments

Precision/recall



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