CONTENT MINE

SOFTWARE OVERVIEW

quickscrape & thresher
norma
AMI fact extraction

SHUTTLEWORTH FUNDED

CONTENT MINE

OUR MISSION

"make 100,000,000 facts from the scholarly literature open, accessible and reusable"

CONTENT MINE

THE SCALE OF THE TASK

- ~ 27,000 peer reviewed journals*
- > 5,000 publishers
- ~ 3,000 new papers per day

**Ulrich's database: http://ulrichsweb.serialssolutions.com/login*

CONTENT MINE

STRUCTURED INFORMATION

- chemical names and structures
- species
- metabolism
- phylogenetic trees

CONTENT MINE

COLLABORATIONS

- Mint phylogeny working group
- Phyloinformatic Literature Unlocking Tools (PLUTo)
- EBI - MetaboLights
- OpenFarm
- OpenOil / OpenCorporates

CONTENT MINE

SOFTWARE PIPELINE

PRODUCT: **journals (ISSNs)** → **fulltext URLs** → **metadata + content + files** → **facts**

PROCESS: **crawl** → **scrape** → **extract**

CONTENT MINE

CRAWLING

The latest journal tables of contents at **Journal TOCs**

http://www.journaltocs.hw.ac.uk/

CONTENT MINE

SCRAPERS

- all have the same plumbing
- scraping software (thresher) handles the plumbing
- scraperJSON is a config file
  - supports large collections of scrapers
  - no programming required
  - not limited to one piece of software

CONTENT MINE

BASIC SCRAPER JSON

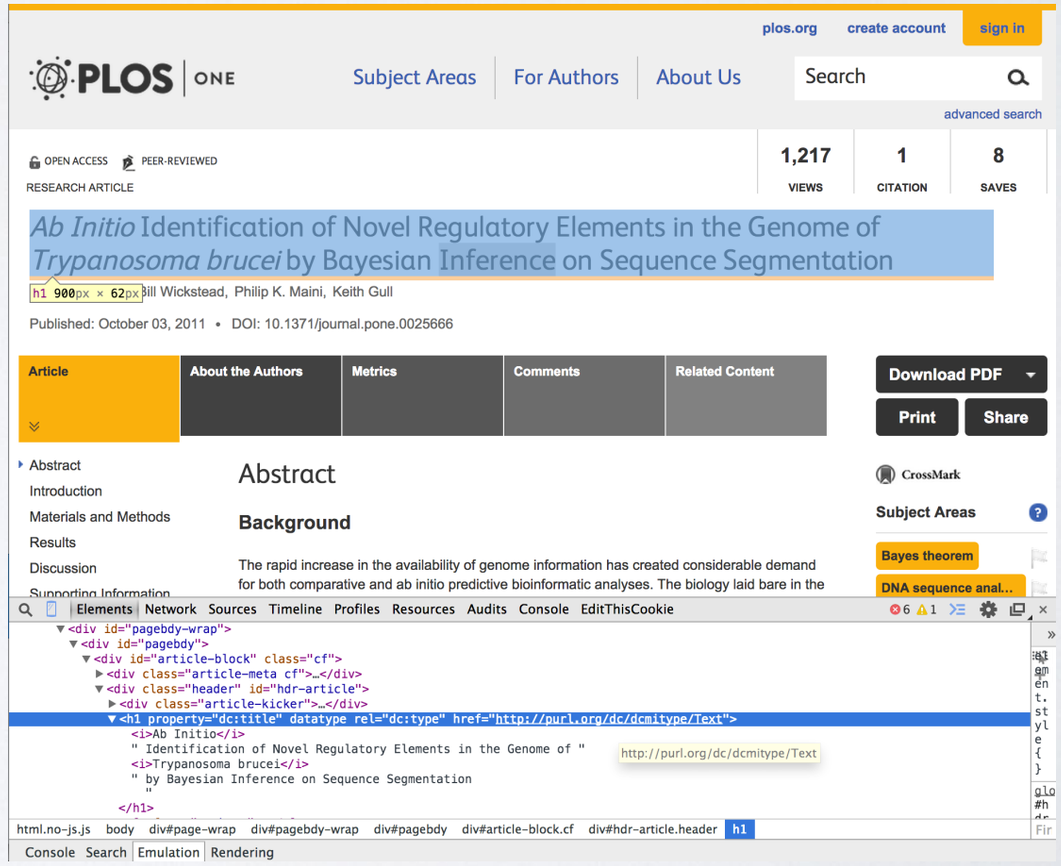
name of the scraper:
the URL(s) it applies to:
the elements to capture:
element name:
where to find it:

```
{

  "name": "PLOS",

  "url": "plos\\w*.org",

  "elements": {

    "title": {

      "selector": “//h1[@property='dc:title']”,

    }

  }

}
```

http://github.com/ContentMine/scraperJSON

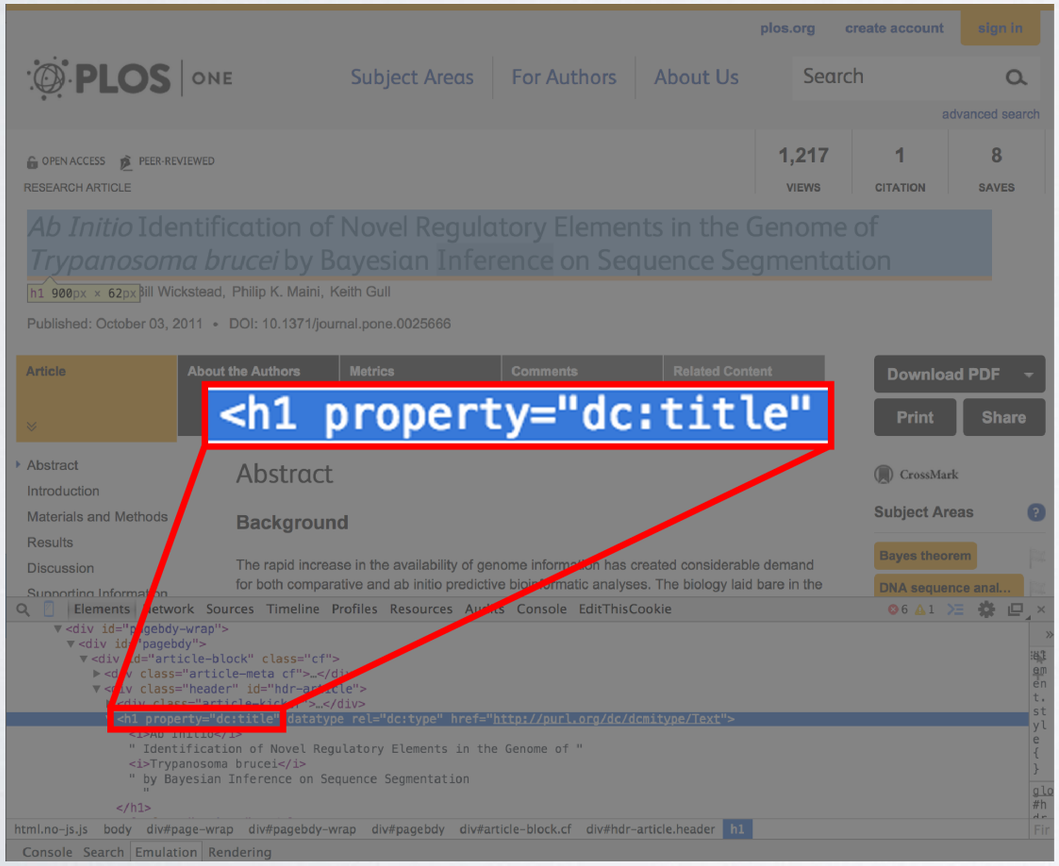
CONTENT MINE

SCRAPERS



CONTENT MINE

SCRAPERS




```
{
  "name": "PLoS",
  "url": "plos\\w*.org",
  "elements": {
    "title": {
      "selector": “//h1[@property='dc:title']”,
    }
  }
}
```

CONTENT MINE

SCRAPERS

bibJSON output


```
{

  "title": "Ab Initio Identification of Novel
Regulatory Elements in the Genome of Trypanosoma
brucei by Bayesian Inference on Sequence
Segmentation"

}
```

CONTENT MINE

THRESHER & QUICKSCRAPE

- reference implementation of scraperJSON
- thresher is the scraping library
  - *http://github.com/ContentMine/thresher*
- quickscrape is the command-line tool
  - *http://github.com/ContentMine/quickscrape*
- Node.js, MIT licensed

CONTENT MINE

JOURNAL SCRAPERS

http://github.com/ContentMine/journal-scrapers

a self-testing collection of scraperJSON scrapers for academic journals

PLOS

PeerJ

ScienceDirect

NPG, AAAS, RSC, ACS

MDPI

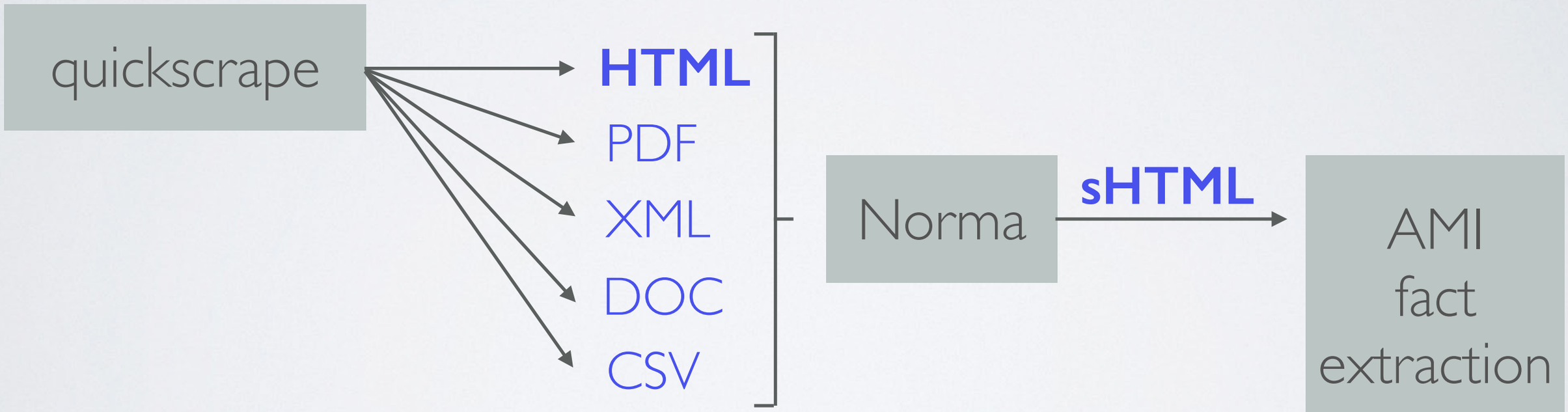
Wiley

Taylor & Francis

Springer

CONTENT MINE

NORMALISATION



CONTENT MINE

NORMALISATION

before

- un-navigable
- non-unicode
- pixel glyphs
- no structure

after

- processable
- sectioned
- tagged
- structured

CONTENT MINE

NORMALISATION

mending on a journal-by-journal basis

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml"
      xmlns:foaf="http://xmlns.com/foaf/0.1/"
      lang="en" xml:lang="en"
      itemscope itemtype="http://schema.org/Article"
      class="no-js">
```

invalid XHTML
from PLOS ONE

```
DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 
tml id="nojs" xmlns="http://www.w3.org/1999
  xmlns:og="http://ogp.me/ns#" xml:lang="en
  xmlns:wb="http://open.weibo.com/wb">
```

invalid XHTML
from BMC

CONTENT MINE

NORMALISATION

document structure

before: un-sectioned HTML from Hindawi

```
<h5 id="sec3.2">3.2. Lake Carl Blackwell, OK (2012)</h5>
<h6 id="sec3.2.1">3.2.1. Vegetative Growth</h6>
<p>

No significant differences were observed in either the
irrigated or rain-fed NDVI values for any of the growth stages
evaluated (Figure
```

after: sectioned and tagged HTML

```
<h:div class="xml-content" xmlns:h="http://www.w3.org/
    xmlns:svg="http://www.w3.org/2000/svg">
    <h:section title="Abstract" tag="abstract">
        <h4 class="header">Abstract</h4>
        <p>
            With the demand for maize increasing, produ
            into more water limited, semiarid regions.
            increasing nitrogen (N) fertilizer costs ar
```

CONTENT MINE

FACT EXTRACTION

we can't turn a

hamburger into a cow

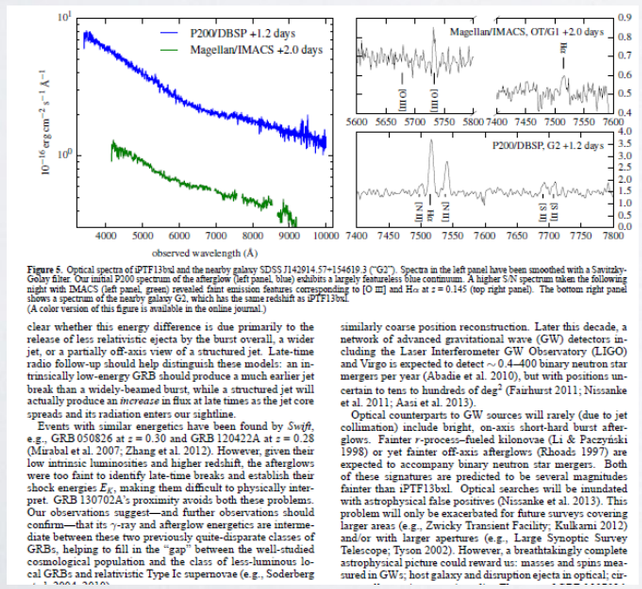


Figure 5. Optical spectra of iPTF13bxl and the nearby galaxy SDSS J142914.57+154619.3 ("G2"). Spectra in the left panel have been smoothed with a Savitzky-Golay filter. Our initial P200 spectrum of the afterglow (left panel, blue) exhibits a largely featureless blue continuum. A higher S/N spectrum taken the following night with IMACS (left panel, green) revealed faint emission features corresponding to [O III] and $H\alpha$ at $z = 0.145$ (top right panel). The bottom right panel shows a spectrum of the nearby galaxy G2, which has the same redshift as iPTF13bxl.
(A color version of this figure is available in the online journal.)

clear whether this energy difference is due primarily to the release of less relativistic ejecta by the burst overall, a wider jet, or a partially off-axis view of a structured jet. Late-time radio follow-up should help distinguish these models: an intrinsically low-energy GRB should produce a much earlier jet break than a widely-beamed burst, while a structured jet will actually produce an *increase* in flux at late times as the jet core spreads and its radiation enters our sightline.

Events with similar energetics have been found by *Swift*, e.g., GRB 050826 at $z = 0.30$ and GRB 120422A at $z = 0.28$ (Mirabal et al. 2007; Zhang et al. 2012). However, given their low intrinsic luminosities and higher redshift, the afterglows were too faint to identify late-time breaks and establish their shock energies E_K, making them difficult to physically interpret. GRB 130702A's proximity avoids both these problems. Our observations suggest—and further observations should confirm—that its γ-ray and afterglow energetics are intermediate between these two previously quite-disparate classes of GRBs, helping to fill in the "gap" between the well-studied cosmological population and the class of less-luminous local GRBs and relativistic Type Ic supernovae (e.g., Soderberg

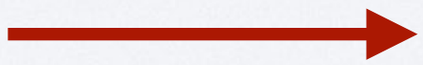
similarly coarse position reconstruction. Later this decade, a network of advanced gravitational wave (GW) detectors including the Laser Interferometer GW Observatory (LIGO) and Virgo is expected to detect ~ 0.4–400 binary neutron star mergers per year (Abadie et al. 2010), but with positions uncertain to tens to hundreds of deg^2 (Fairhurst 2011; Nissanke et al. 2011; Aasi et al. 2013).

Optical counterparts to GW sources will rarely (due to jet collimation) include bright, on-axis short-hard burst afterglows. Fainter *r*-process–fueled kilonovae (Li & Paczyński 1998) or yet fainter off-axis afterglows (Rhoads 1997) are expected to accompany binary neutron star mergers. Both of these signatures are predicted to be several magnitudes fainter than iPTF13bxl. Optical searches will be inundated with astrophysical false positives (Nissanke et al. 2013). This problem will only be exacerbated for future surveys covering larger areas (e.g., Zwicky Transient Facility; Kulkarni 2012) and/or with larger apertures (e.g., Large Synoptic Survey Telescope; Tyson 2002). However, a breathtakingly complete astrophysical picture could reward us: masses and spins measured in GWs; host galaxy and disruption ejecta in optical; cir-

but we can

turn PDFs

into science



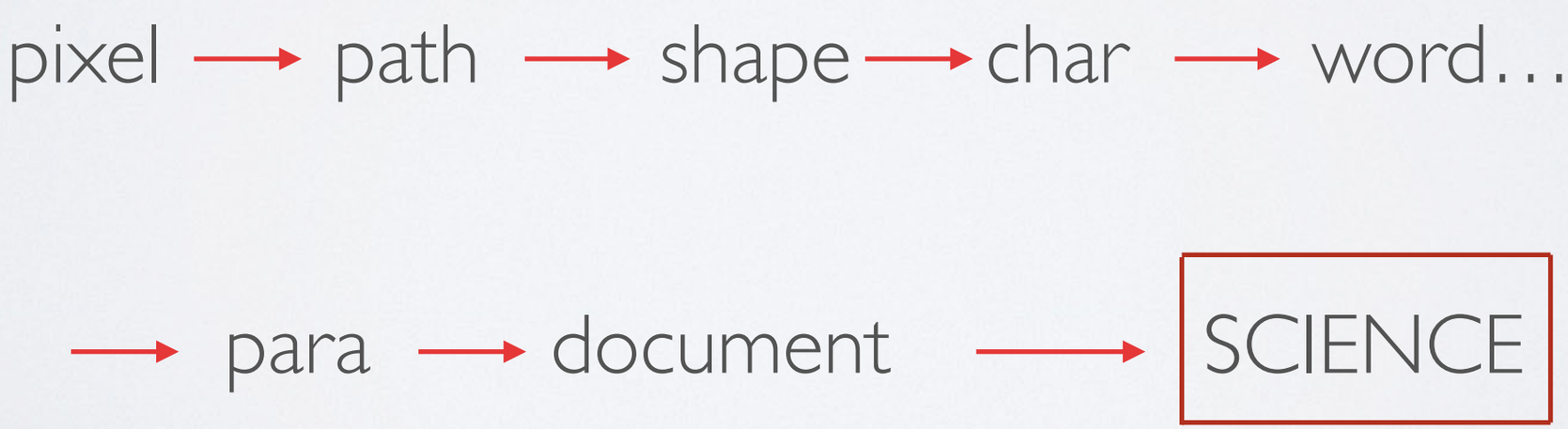
	A	B
1	99.056	98.563
2	99.257	85.249
3	99.324	84.918
4	99.358	85.129
5	99.592	87.833
6	99.626	87.65
7	99.726	85.133

1914	305.073	182.982
1915	305.107	182.142
1916	305.207	173.761
1917		

CONTENT MINE

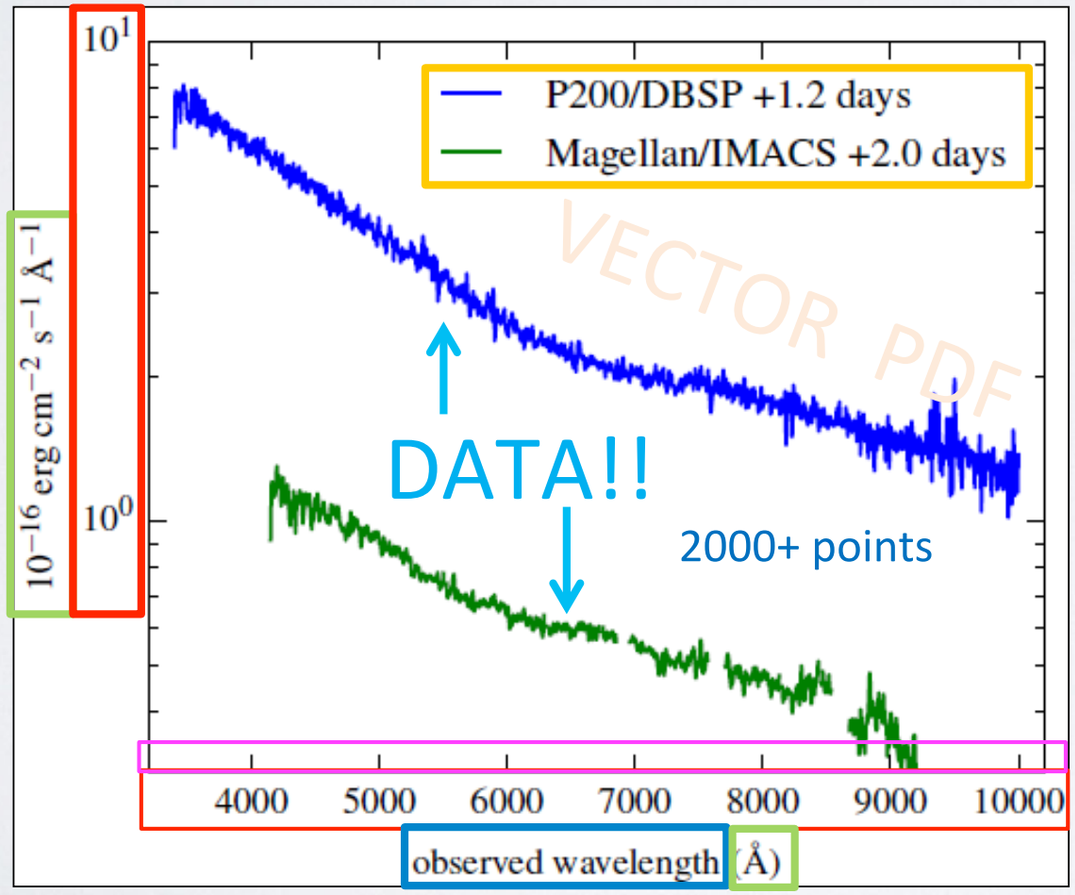
FACT EXTRACTION

AMI software: https://bitbucket.org/petermr/ami-core



CONTENT MINE

FACT EXTRACTION

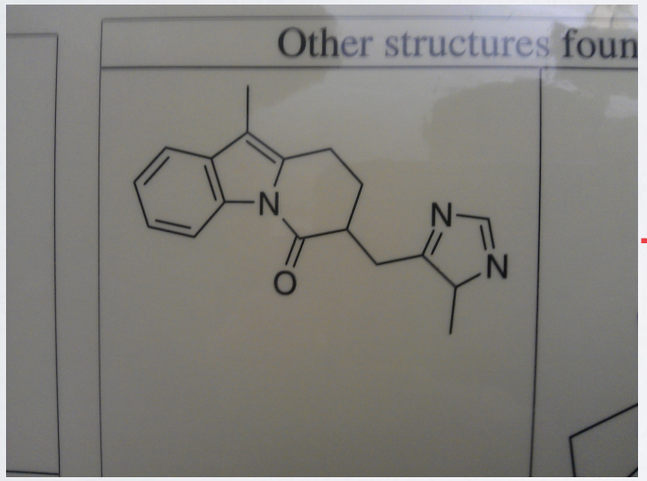


- titles
- scale
- units
- ticks
- quantity
- + data

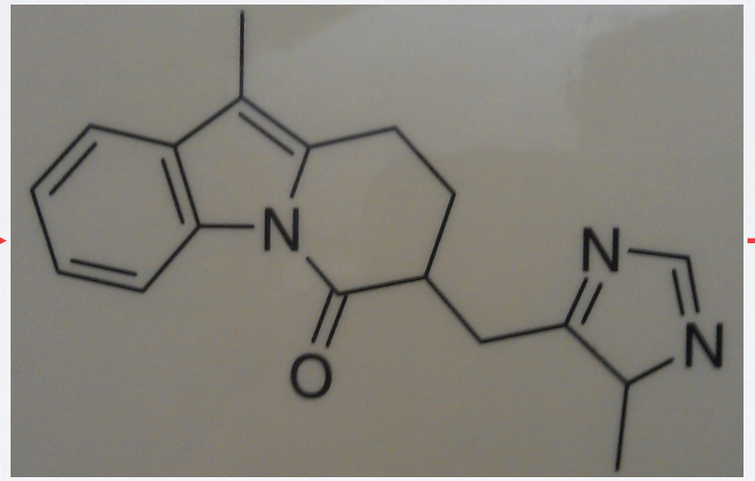
CONTENT MINE

FACT EXTRACTION

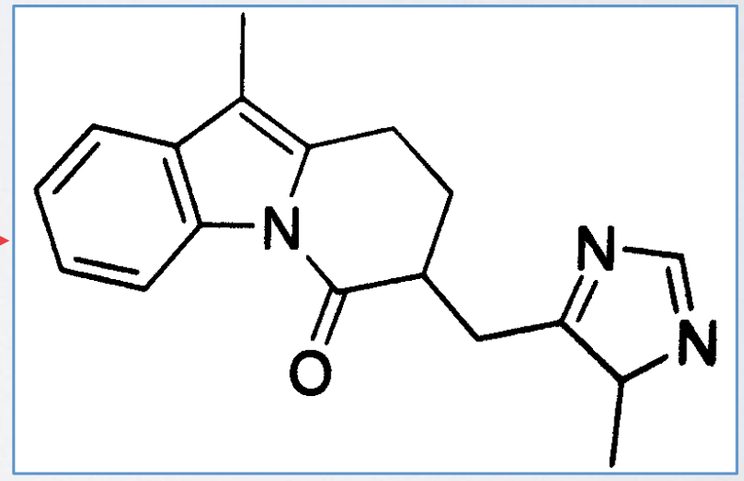
AMI-chem for extracting chemical formulae



raw mobile photo
shadows, contrast,
noise, skew



clipping

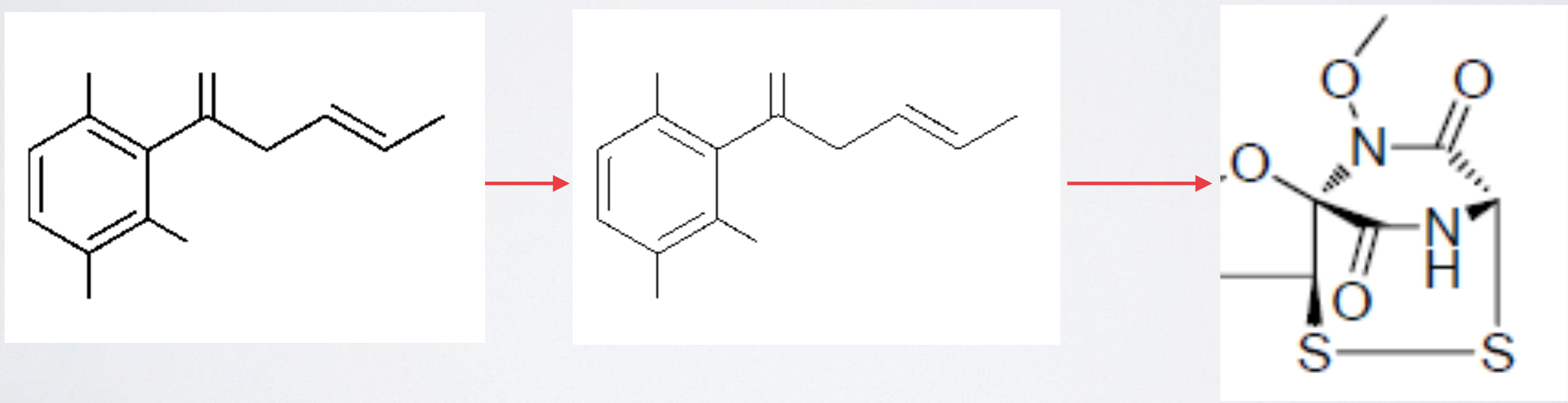


binarization:
pixels = 0, 1

CONTENT MINE

FACT EXTRACTION

AMI-chem for extracting chemical formulae



thinning

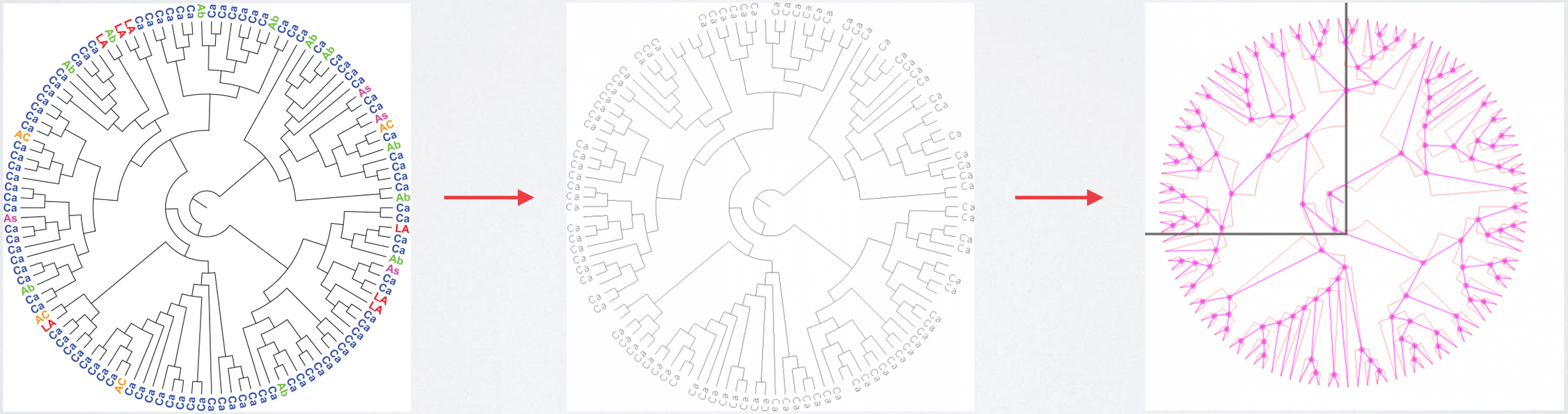
down to 1- pixel

chemical optical character recognition

CONTENT MINE

FACT EXTRACTION

AMI-phylo for extracting phylogenetic trees



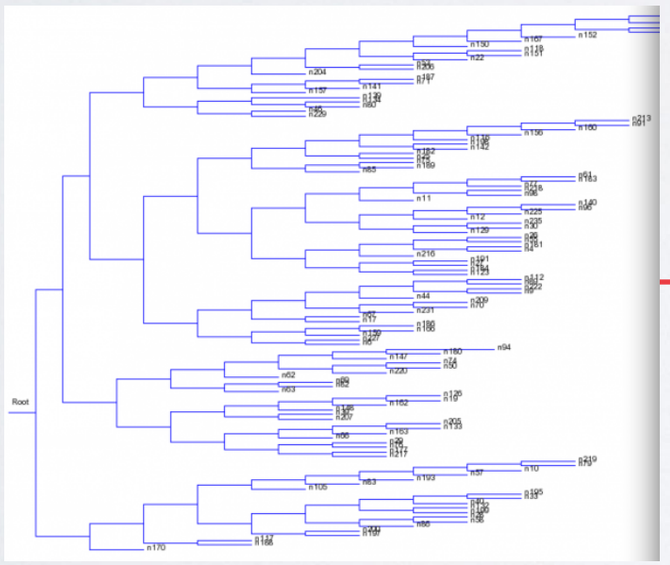
thinning

topology

CONTENT MINE

FACT EXTRACTION

AMI-phylo for extracting phylogenetic trees



serialization

```
((n122,((n121,n205),((n39,(n84,((((n35,n98),n191),n22),n17))),((n10,n182),
    ((((n232,n76),n68),(n109,n30)),(n73,(n106,n58)))))),((((((n103,n86),
    (n218,(n215,n157))),((n164,n143),((n190,((n108,n177),(n192,n220))),
    ((n233,n187),n41)))),(((n59,n184),((n134,n200),(n137,(n212,
    ((n92,n209),n29))))),(n88,(n102,n161))),(((n70,n140),(n18,n188)),(n49,
    ((n123,n132),(n219,n198)))),(((n37,(n65,n46)),(n135,(n11,
    (n113,n142)))),(n210,((n69,(n216,n36)),(n231,n160)))))))),(((n107,n43),
    ((n149,n199),n74)),(((n101,(n19,n54)),n96),(n7,((n139,n5),((n170,
    (n25,n75)),(n146,(n154,(n194,(((n14,n116),n112),(n126,n222))))))))))),
    (((((n165,(n168,n128)),n129),((n114,n181),(n48,n118))),((n158,(n91,
    (n33,n213))),(n87,n235))),((n197,(n175,n117)),(n196,((n171,
    (n163,n227)),((n53,n131),n159)))))));
```

Newick format can be viewed at:

http://www.unc.edu/~bdmorris/treelib-js/demo.html

CONTENT MINE

ACKNOWLEDGEMENTS

Richard Smith-Unna, Dept Plant Sci, Univ. Cambridge

Andy Howlett, Dept Chemistry, Univ. Cambridge

Mark Williamson, Dept Chemistry, Univ. Cambridge

Ross Mounce, Biology, Univ. Bath