

# Practical Astroinformatics

... or what I wish to knew when I was younger

Jaroslav Vážný

Masarykova univerzita

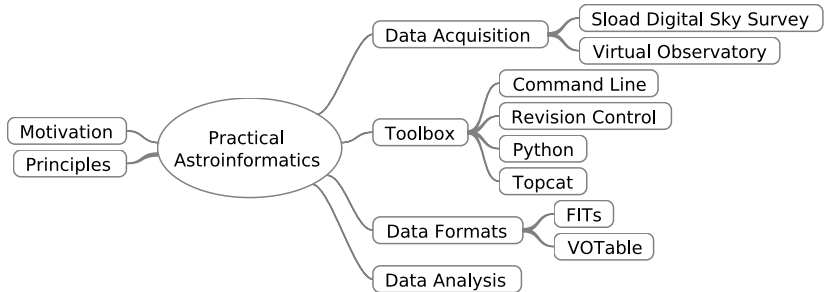
26. června 2012

# Prelude

motto: The only way to keep away from computers in science is to understand them ...

<https://www.coursera.org/>

# Concepts introduced in this talk



# Data Avalanche?

- Large Synoptic Survey Telescope
  - 20 TB per night
  - 60 PB for the raw data (after 10 years)
  - 15 PB for the catalog database

The total data volume after processing will be several hundred PB

- Where I can learn more?
  - <http://www.lsst.org/>

# Sloan Digital Sky Survey

- Why is it important?
  - Lots of data ( $>10^6$  objects)
  - Perfect documentation
  - Tools to access the data
- Where I can learn it?
  - <http://www.sdss3.org/>

# Virtual Observatory

- Why is it important?
  - Uniform access to astronomy data
  - Based on Web standards
  - Many tools with vo support (Topcat, Aladin, Tapsh)
- Where I can learn it?
  - [http://physics.muni.cz/~vazny/wiki/index.php/Diploma\\_work](http://physics.muni.cz/~vazny/wiki/index.php/Diploma_work)

# Example: Virtual Observatory Protocols

## Cone Search Protocol

1 `http://simbad.u-strasbg.fr/simbad-conesearch.pl?RA=24.5&  
DEC=-57.2&SR=0.1`

## Simple Image Access Protocol

1 `http://hubblesite.org/cgi-bin/sia/hst_pr_sia.pl?POS  
=83.6,22.0&SIZE=1.0`

## Simple Spectra Access Protocol

1 `http://archive.eso.org/apps/ssaserver/EsoProxySsap?  
REQUEST=queryData&POS=83.63,22&SIZE=1`

# Example: Virtual Observatory Protocols

## Table Access Protocol

```
1  -- Display all identifiers of a given object.  
2  SELECT id2.id  
3  FROM ident AS id1 JOIN ident AS id2 USING(oidref)  
4  WHERE id1.id = 'M1';
```

<http://simbad.u-strasbg.fr/simbad/sim-tap>



# Command Line

- Why is it important?
  - Efficient dialog computer  $\iff$  human
  - In all advanced tools (Programming, mathematica, CAD, ...)
  - Cooperation, re-usability, automatize
- Where I can learn it?
  - PEEPCODE: Meet the Command Line, Advanced Command Line

# Examples

- TAB, CTRL-A, CTRL-E (=Emacs)
- !! Repeat last command
- !\$ Repeat last argument
- history command history
- CTRL+R search in history

# Text tools

- Why is it important?
  - "Everything" is a text
  - head, tail, sed, awk, join, paste, vim, emacs ...
- Where I can learn it?
  - PEEPCODE: Meet Emacs, Smash Into Vim, Vim Emacs tutorials

# Revision Control Systems

- Why is it important?
  - Distributed systems (Git, Mercurial)
  - Almost everything is local
  - Branching
  - Natural (subjective?)
- Where I can learn it?
  - PEEPCODE: Git, Mercurial
  - <https://github.com>
  - <http://gitref.org/>
  - <http://www.youtube.com/watch?v=ZDR433b0HJY>

# Python

- Why is it important?
  - Language of science ?
  - Cooperation between scientist (Scipy conference)
  - Perfect for experiments (iPython)
  - Real free language (!= MATBLAB)
- Where I can learn it?
  - <http://pyvideo.org/>
  - <http://www.youtube.com/watch?v=B9MvjMFokLc>
  - <http://ipython.org/>

# Topcat

- Why is it important?
  - Perfect for big data (not only astro)
  - Example of cooperation between GUI applications
  - Learning Astrophysics
- Where I can learn it?
  - <http://www.star.bris.ac.uk/~mbt/topcat/>
  - <http://www.eurovo-ice.eu/twiki/bin/view/EuroVOICE/ICESchool>

# FITs

- Why is it important?
  - De-Facto standard in Astronomy
  - Flexible, Efficient, ASCII Meta-Data
- Where I can learn it?
  - <http://fits.gsfc.nasa.gov>

## Example: Reading FITS file

```

1 In [1]: import pyfits
2 In [2]: hdulist = pyfits.open('spSpec-53237-1886-248.fit')
3 In [3]: hdulist.info()
4 Filename: spSpec-53237-1886-248.fit
5
6 No.      Name          Type          Cards   Dimensions   Format
7 0        PRIMARY      PrimaryHDU    213      (3874, 5)    float32
8 1                   BinTableHDU   54      6R x 23C    [1E, 1E, ...
9 2                   BinTableHDU   54      44R x 23C    [1E, 1E, ...
10 3                   BinTableHDU   18      1R x 5C     [1E, 1E, ...
11 4                   BinTableHDU   32      53R x 12C    [1J, 1J, ...
12 5                   BinTableHDU   26      36R x 9C     [19A, 1E,
    ...
    6                   BinTableHDU   14      3874R x 3C    [1J, 1J, 1E]
  
```



# VOTable

- Why is it important?
  - Standard in Virtual Observatory
  - Flexible, Efficient, XML
- Where I can learn it?
  - <http://www.ivoa.org>

## Example: VOTable

```
1 <?xml version="1.0" encoding="utf-8"?>
2   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
3   xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/
   VOTable/v1.0"
4   xmlns="http://www.ivoa.net/xml/VOTable/v1.0">
5   <RESOURCE type="results" >
6     <TABLE >
7       <FIELD ID="col0" name="wave" datatype="float" unit=""
8         precision="F9"/>
9       <DATA>
10        <TABLEDATA>
11          <TR>
12            <TD>4012.50757</TD>
13          </TR>
14        </TABLEDATA>
15      </DATA>
```

## Example: Working with FITs in Python

```
1 In [1]: import atpy
2 In [2]: tbl = atpy.Table('spSpec-53401-2052-458.fit')
3 Auto-detected input type: fits
4 In [3]: tbl.write('votableExample.xml')
5 Auto-detected input type: vo
```

Updating FITS file.

```
1 In [1]: prihdr = hdulist[0].header
2 In [2]: prihdr.update('observer', 'Astar')
3 In [3]: prihdr.add_history('Updated 3/27/11')
```

# Data Mining

- Why is it important?
  - Astrology of data
  - Data preprocessing
- Where I can learn it?
  - Stanford(Andrew Ng)
  - `www.avc.cvut.cz`

## Example: Decison Tree

```
1 ug <= 0.663668
2 |   gr <= -0.191208: 1 (7.0)
3 |   gr > -0.191208: 3 (104.0/5.0)
4 ug > 0.663668
5 |   ri <= 0.285854: 1 (88.0/5.0)
6 |   ri > 0.285854
7 | |   ri <= 0.314657
8 | | |   gr <= 0.692108: 2 (6.0)
9 | | |   gr > 0.692108: 1 (3.0)
10 | |   ri > 0.314657: 2 (90.0/2.0)
```

# Discussion