Introduction to command-line tools and biomolecule drawing software

S. *Ravi*chandran
ravichandran@hood.edu

Discussion

- What is the key difference between Linus Pauling's DNA model and Watson and Crick's model?

- What was the key conclusion of Watson and Crick's paper?

Agenda

- Drawing/Viewing molecules: Demonstration using ChemSketch
- **Short Lecture**
- Back to hands-on
  - Login to Biolinux using Putty
  - **Linux Tutorial**
    - **Supporting webpage**
    - https://web.stanford.edu/group/farmshare/cgi-bin/wiki/index.php/How_to_learn_linux
    - PDF from the above link (available on BlackBoard)
  - **Shell tutorial (if time permits)**

Agenda (contd.)

- **Installation of NCBI e-utilities (Edirect) and will work out few examples**
  - **Exercise files will be made available in the class**

- **Introduction to R/R-Studio**
  - Introduction to RStudio
  - R-hands-on via RStudio

Unix/Linux

- **Warning**
- This is just a basic introduction to the Linux OS. For newcomers, it needs several weeks of exposure to become comfortable with this OS
- Today, we are just going to barely scratch the surface
- Please refer to **SR-Linuxtutorial.docx (available from Blackboard "Needed in Class" folder)**

Basic Intro to Shell

- Shell is a program that sits between the User and the OS.
- There are many flavors
  - C, K, Bourne, Bourne-again etc.
- C or tcsh is very user-friendly but rarely used by system administrators.
- I mostly use C-shell (tcsh)
- **I will demonstrate the capability in the class**

Entrez Direct (Edirect)

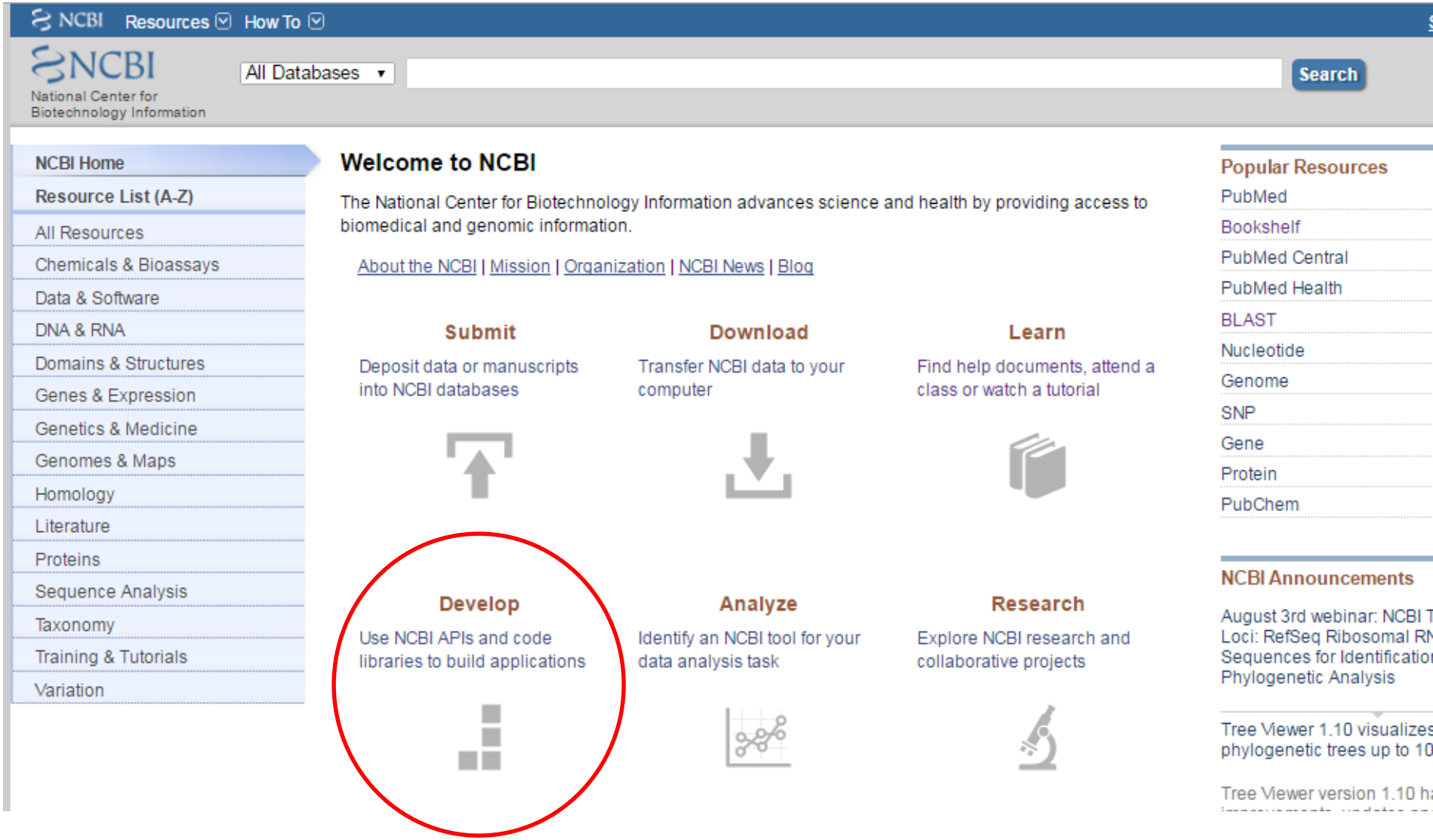
- Advanced method for accessing (not readily available via web-interface) the NCBI's databases
- Perl scripts accessible via Linux/Unix Shell scripts
- Will run on Unix/Linux/Mac systems that has Perl installed
  - If you are using Windows; use Cygwin (not complicated and I will not cover this in the class)

What is needed to run Edirect?

- Unix/Linux/Mac OSX
  - Windows
    - Unix/Linux emulators (SeqWin etc.)
- Perl with LWP::Simple
- Basic Idea
  - Each script outputs an XML
    - These XML outputs can in turn be piped using other Edirect commands
      - ESearch, Efetch, Elink, Epost xtract etc.
        - » Xtract (Powerful XML parser)

Why XML?
Easy to parse an XML
NCBI adopted XML based output

Where do we get this software?



FEEDBACK

Develop

NCBI provides a variety of resources that allow developers to access and manipulate NCBI data in their applications.

APIs

Programming interfaces including the E-utilities, BLAST URL API and PUG

Code Libraries

Software libraries including the C++ and SRA toolkits

Data Formats

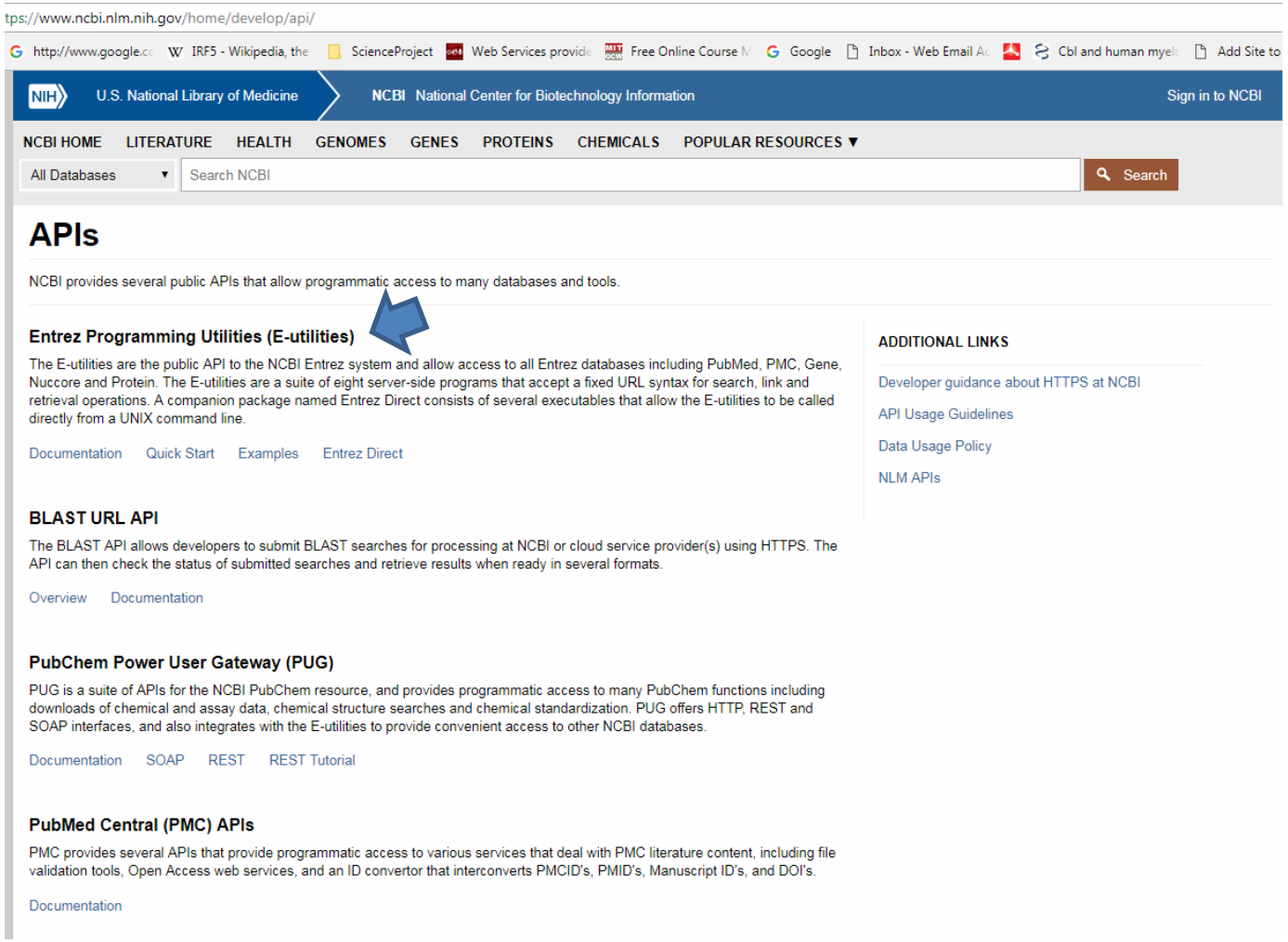
Schema, DTDs and other data specifications

GitHub

NCBI GitHub repository

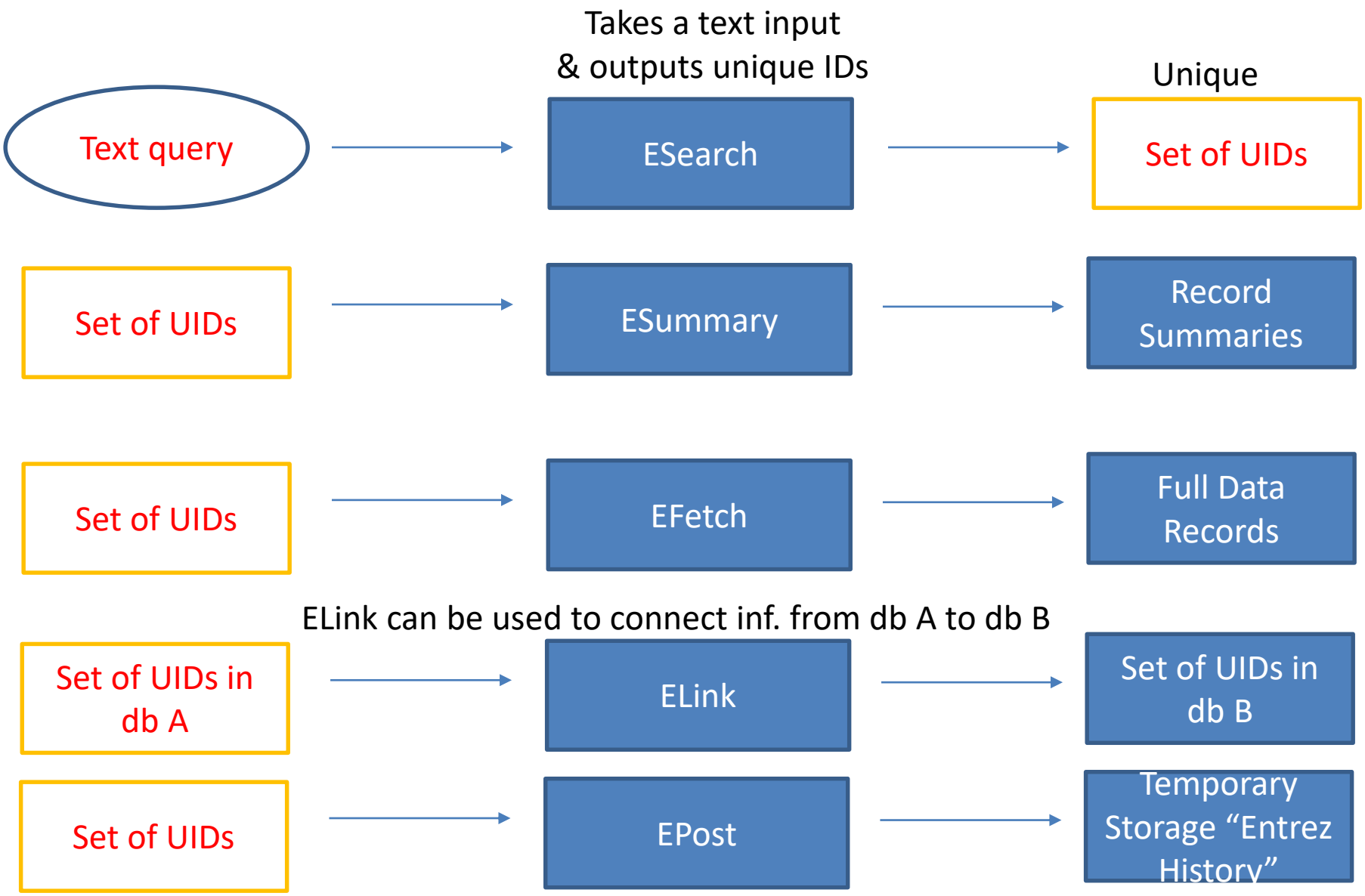
News & Blog

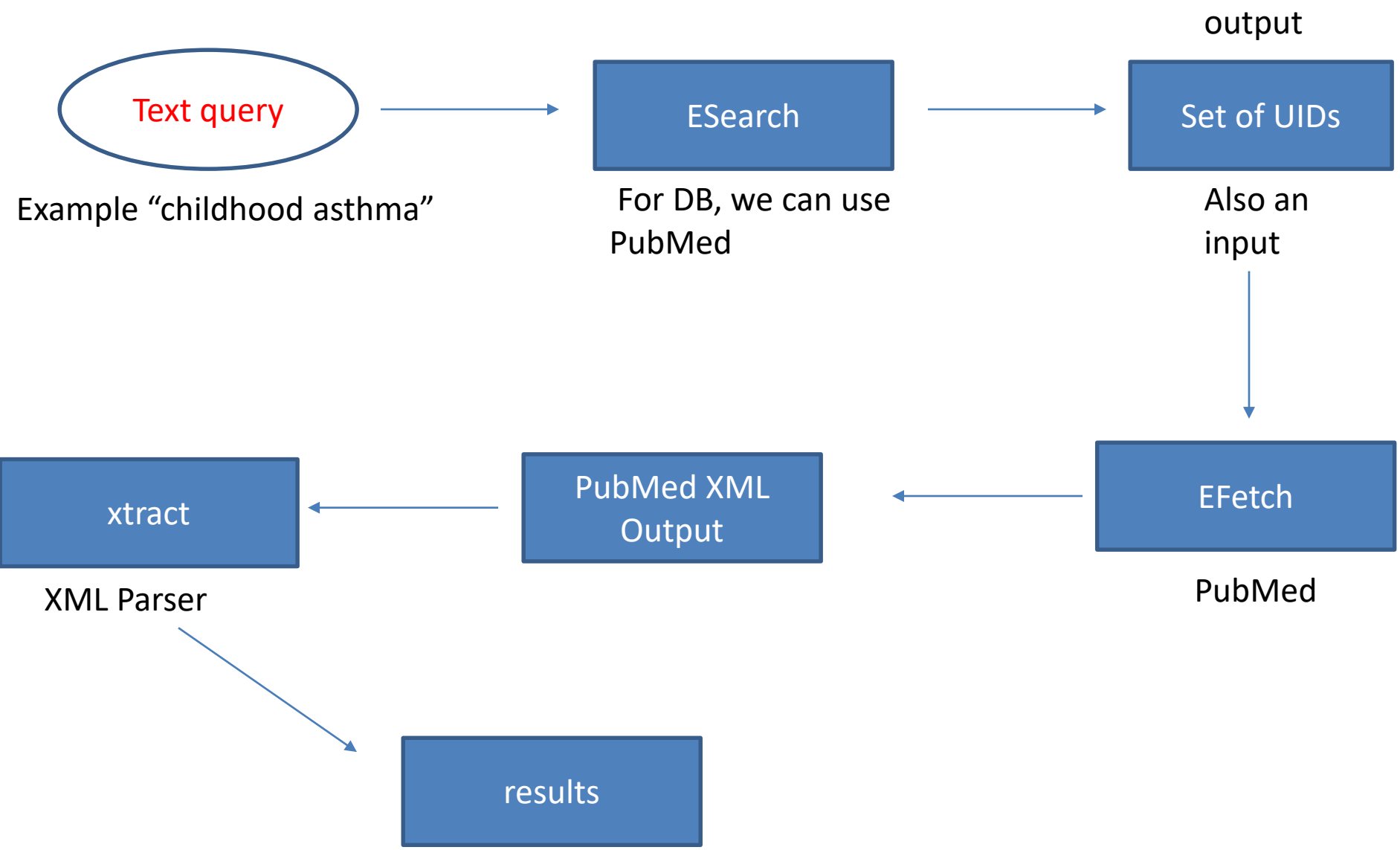
Entrez Utilities News
C++ toolkit News
API blog posts


tps://www.ncbi.nlm.nih.gov/home/develop/api/
http://www.google.cc
IRF5 - Wikipedia, the
ScienceProject
Web Services provide
Free Online Course M
Google
Inbox - Web Email Ac
Cbl and human myel
Add Site to
NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
NCBI HOME
LITERATURE
HEALTH
GENOMES
GENES
PROTEINS
CHEMICALS
POPULAR RESOURCES ▼
All Databases
Search NCBI
Search
APIs
NCBI provides several public APIs that allow programmatic access to many databases and tools.
Entrez Programming Utilities (E-utilities)
The E-utilities are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene, Nuccore and Protein. The E-utilities are a suite of eight server-side programs that accept a fixed URL syntax for search, link and retrieval operations. A companion package named Entrez Direct consists of several executables that allow the E-utilities to be called directly from a UNIX command line.
Documentation
Quick Start
Examples
Entrez Direct
BLAST URL API
The BLAST API allows developers to submit BLAST searches for processing at NCBI or cloud service provider(s) using HTTPS. The API can then check the status of submitted searches and retrieve results when ready in several formats.
Overview
Documentation
PubChem Power User Gateway (PUG)
PUG is a suite of APIs for the NCBI PubChem resource, and provides programmatic access to many PubChem functions including downloads of chemical and assay data, chemical structure searches and chemical standardization. PUG offers HTTP, REST and SOAP interfaces, and also integrates with the E-utilities to provide convenient access to other NCBI databases.
Documentation
SOAP
REST
REST Tutorial
PubMed Central (PMC) APIs
PMC provides several APIs that provide programmatic access to various services that deal with PMC literature content, including file validation tools, Open Access web services, and an ID convertor that interconverts PMCID's, PMID's, Manuscript ID's, and DOI's.
Documentation
ADDITIONAL LINKS
Developer guidance about HTTPS at NCBI
API Usage Guidelines
Data Usage Policy
NLM APIs

Based on NCBI Help Pages

E-utilities

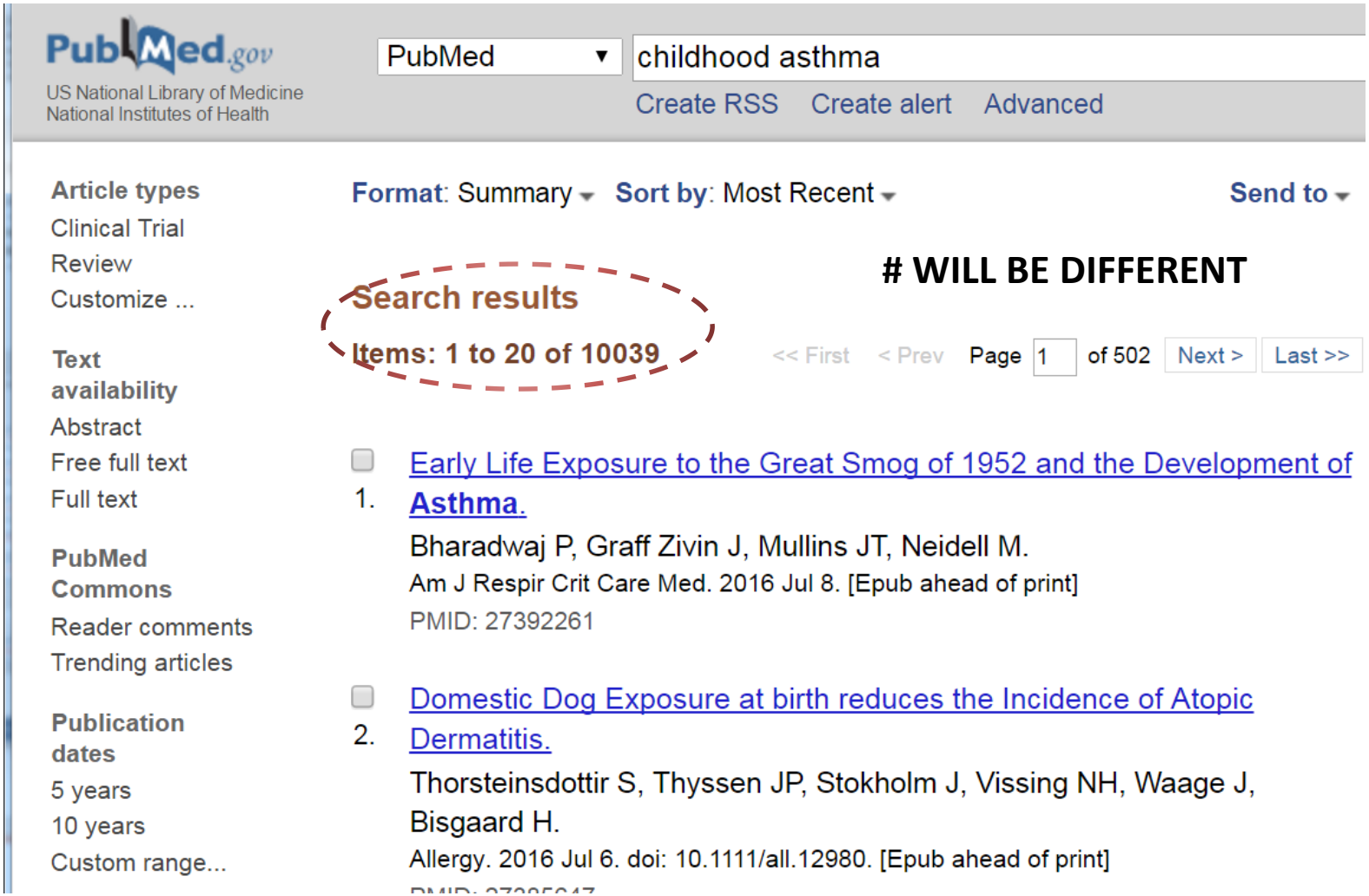



output
Text query
ESearch
Set of UIDs
Example “childhood asthma”
For DB, we can use PubMed
Also an input
EFetch
PubMed XML Output
xtract
PubMed
XML Parser
results

Edirect/E-utilies

- Why is it useful?
- Let us do a couple of examples

NOTE: DO NOT USE QUOTES FOR THE STRING THAT YOU WILL USE FOR THE SEARCH



Let us see how we can use E-Direct to get the same output using Command Prompt

esearch -db pubmed -query "childhood asthma"

```
<ENTREZ_DIRECT>
  <Db>pubmed</Db>
  <WebEnv>NCID_1_90103729_130.14.18.34_9001_1468093986_406367782_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>10039</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

Look at the count, it is same as the web based search.

Is it useful? It just gives you a total count. If you want individual records, you have to use esummary

esearch -db pubmed -query "childhood asthma“ | esummary

Note this will keep dumping all the 10,039 records in the XML format.

Edirect/E-utilies

- Second example
- For example, if you want to search PubMed for articles that contain the keyword, “lycopene cyclase” and download all the resulting article abstracts into a file.

Applications of E-Direct

- Go to PubMed, type “lycopene cyclase”
  - Hit “**Search**”
- From the hits page, change the Format of the page to **“Abstract”** and also ask it to show 200 hits per page to see them all
- Use **Send to** a file for download
- **E-Utils**
- **esearch -db pubmed -query "lycopene cyclase" | \ efetch -format abstract > HitsFile.txt**

- esearch –db protein –query “lycopene cyclase” | efetch –format fasta
- Please read more on the following NCBI link:
- http://www.ncbi.nlm.nih.gov/books/NBK179288/

Key flags that work well with xtract

- **-sep** and **–tab**

- **esearch –db pubmed –query “arthritis”**
  - **Will get a XML output**
  - **Which is a summary file # of hits etc,**
  - **Not very useful; to extract summaries, we pipe it to esummary**
- **esearch –db pubmed –query “arthritis” | esummary**

```
esearch –db pubmed –query “arthritis” \
  | efetch –format xml | xtract –pattern  \
Pubmed Article  -block Author –element \
LastName, Initials
```

-pattern Pubmed Article will loop over each article

Within each article, block will search

Output will be a bunch of names LastName, Initials separated by tab character (default).

To force the script to use space, use “-sep” flag

```
esearch –db pubmed –query “arthritis” \
  | efetch –format xml | xtract –pattern  \ Pubmed Article  -block \
Author –sep “ “ –element \ LastName, Initials
```

If we want each Author in a new line

```
esearch –db pubmed –query “arthritis” \
  | efetch –format xml | xtract –pattern  \ Pubmed Article  -block \
Author –sep “ “ –tab “\n” –element \ LastName, Initials
```

? How do we store the results in a file? (Linux shell command)

Take that file and sort them and see the frequency of each author

Hint we will use two shell command: sort and uniq

More complicated example

- Task
  - Search for the PNAS journal published papers of Garber ED
  - Collect all similar articles
  - Look for the keyword “mouse”
  - Extract the articles, first author name, PMID and the article title.
- This task is very difficult in some cases not possible to carry out via web interface

Websearch

- NCBI → PubMed → **Garber ED [AUTH] AND PNAS [JOUR]** → You get 4 or more hits → Collect articles from "Similar articles" → Look for keyword **mouse → Get the article PMID, article title, Author first name**
- **esearch** -db pubmed -query "Garber ED [AUTH] AND PNAS [JOUR]" | **elink** -related | **efilte**r -query "mouse" | **efetch** -format docsum | **xtract** -pattern DocumentSummary -element Id SortFirstAuthor Title

Links

- Linux
  - https://web.stanford.edu/group/farmshare/cgi-bin/wiki/index.php/How_to_learn_linux
- NCBI E-utilities
  - API to NCBI Entrez system
  - Need: Internet, web-browser, Perl etc.
  - http://www.ncbi.nlm.nih.gov/books/NBK25501/
- R
  - https://www.r-project.org/ (Windows/Mac/Linux)
- Rstudio
  - https://www.rstudio.com/ (Windows/Mac/Linux)

Statistics

- R/Rstudio
- R
  - Free Software for statistical computing and graphics.
  - It compiles and runs on a wide variety of OS
    - Unix, Windows, MacOS. UNIX platforms, Windows and MacOS.
- RStudio
  - GUI for R; Very helpful to run R via RStudio