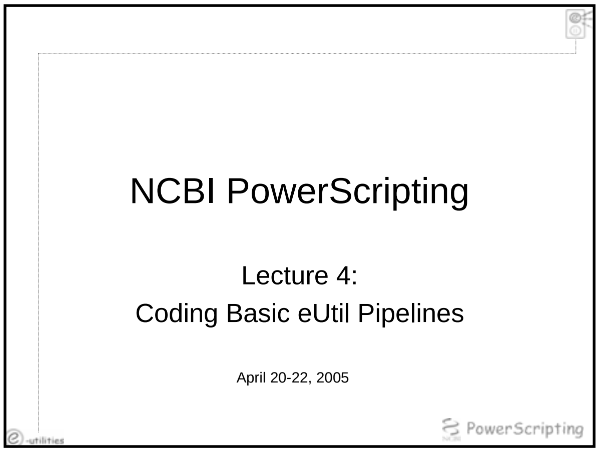
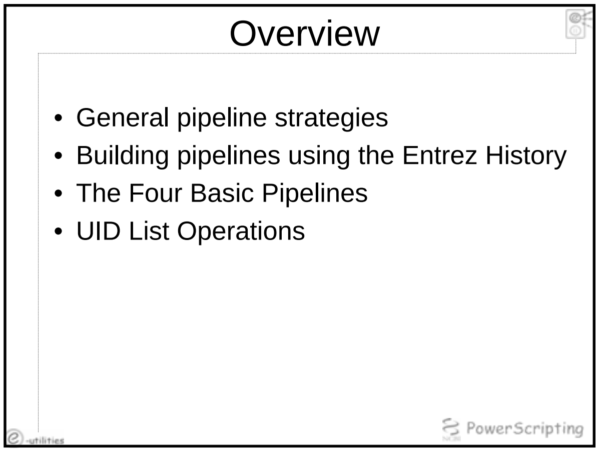


NCBI PowerScripting

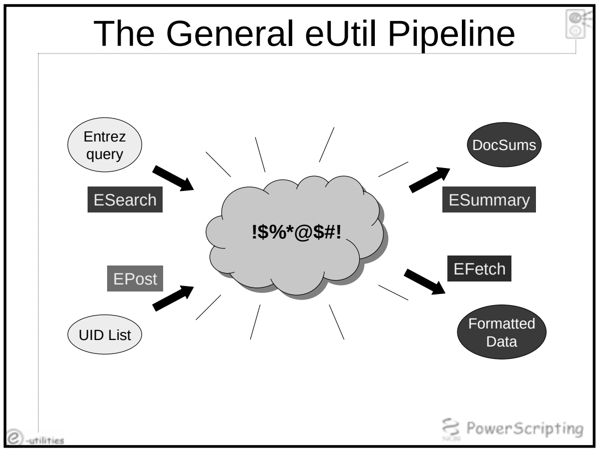
Lecture 4:
Coding Basic eUtil Pipelines

April 20-22, 2005



Overview

- General pipeline strategies
- Building pipelines using the Entrez History
- The Four Basic Pipelines
- UID List Operations



The General eUtil Pipeline

Entrez query

ESearch

EPost

UID List

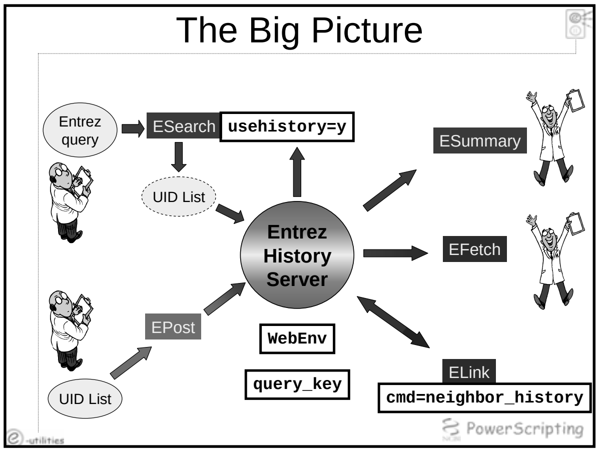
!$%*@$#!

DocSums

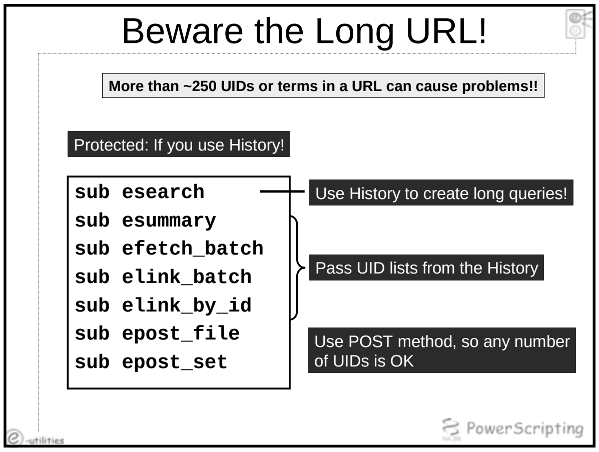
ESummary

EFetch

Formatted Data



The Big Picture



Beware the Long URL!

More than ~250 UIDs or terms in a URL can cause problems!!

Protected: If you use History!

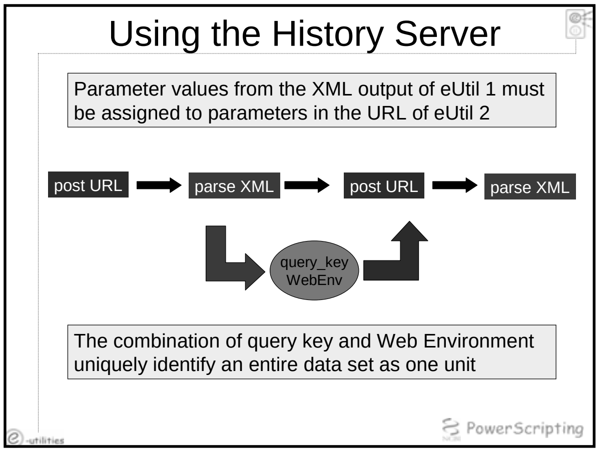
```
sub esearch
sub esummary
sub efetch_batch
sub elink_batch
sub elink_by_id
sub epost_file
sub epost_set
```

Use History to create long queries!

Pass UID lists from the History

Use POST method, so any number of UIDs is OK

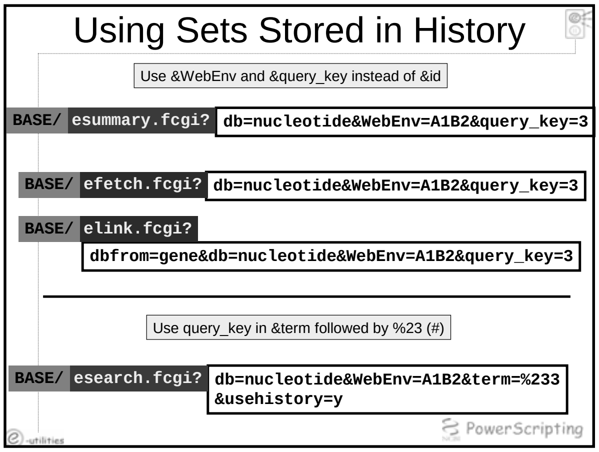
PowerScripting



Using the History Server

Parameter values from the XML output of eUtil 1 must be assigned to parameters in the URL of eUtil 2

The combination of query key and Web Environment uniquely identify an entire data set as one unit



Using Sets Stored in History

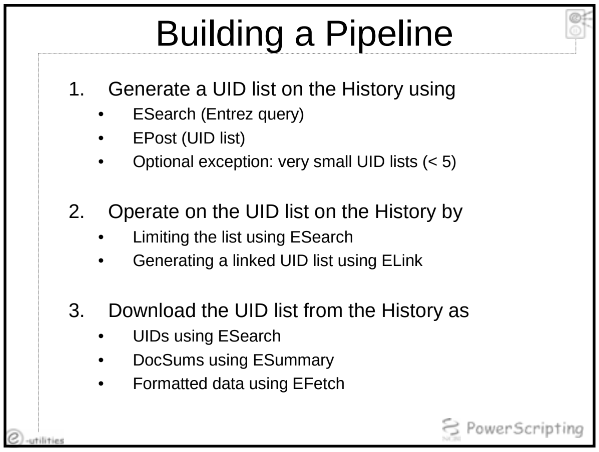
Use &WebEnv and &query_key instead of &id

```
BASE/ esummary.fcgi? db=nucleotide&WebEnv=A1B2&query_key=3

BASE/ efetch.fcgi? db=nucleotide&WebEnv=A1B2&query_key=3

BASE/ elink.fcgi?
      dbfrom=gene&db=nucleotide&WebEnv=A1B2&query_key=3
```

Use query_key in &term followed by %23 (#)

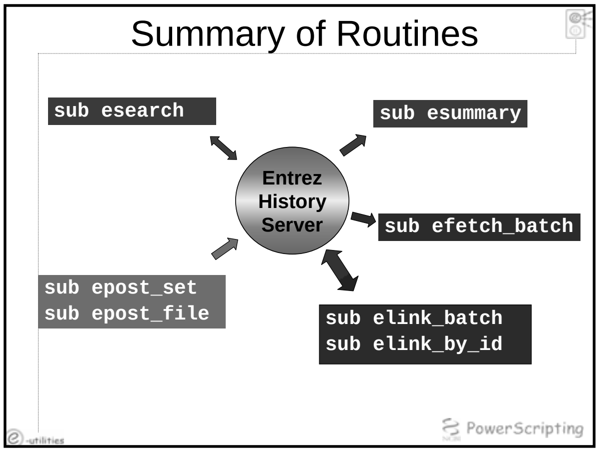
```
BASE/ esearch.fcgi? db=nucleotide&WebEnv=A1B2&term=%233
                    &usehistory=y
```

PowerScripting
e-utilities

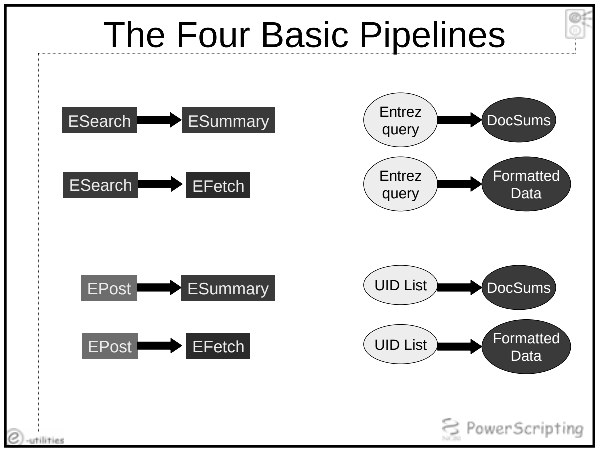


Building a Pipeline

1. Generate a UID list on the History using
   - ESearch (Entrez query)
   - EPost (UID list)
   - Optional exception: very small UID lists (< 5)
2. Operate on the UID list on the History by
   - Limiting the list using ESearch
   - Generating a linked UID list using ELink
3. Download the UID list from the History as
   - UIDs using ESearch
   - DocSums using ESummary
   - Formatted data using EFetch



Summary of Routines



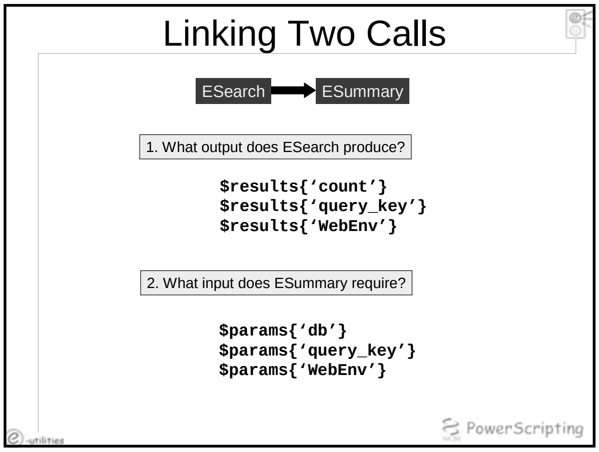
The Four Basic Pipelines

ESearch → ESummary — Entrez query → DocSums

ESearch → EFetch — Entrez query → Formatted Data

EPost → ESummary — UID List → DocSums

EPost → EFetch — UID List → Formatted Data



Linking Two Calls

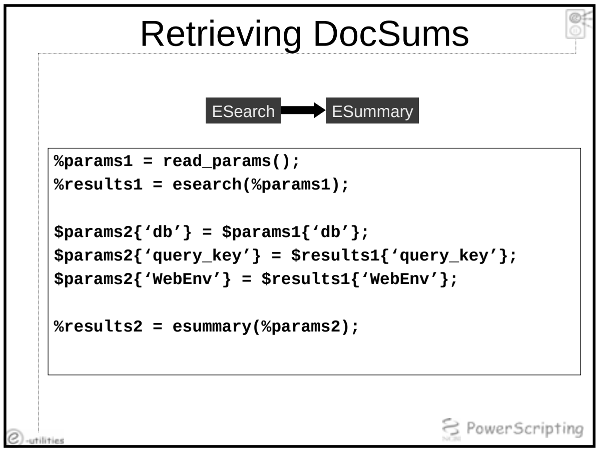
ESearch → ESummary

1. What output does ESearch produce?

```
$results{‘count’}
$results{‘query_key’}
$results{‘WebEnv’}
```

2. What input does ESummary require?

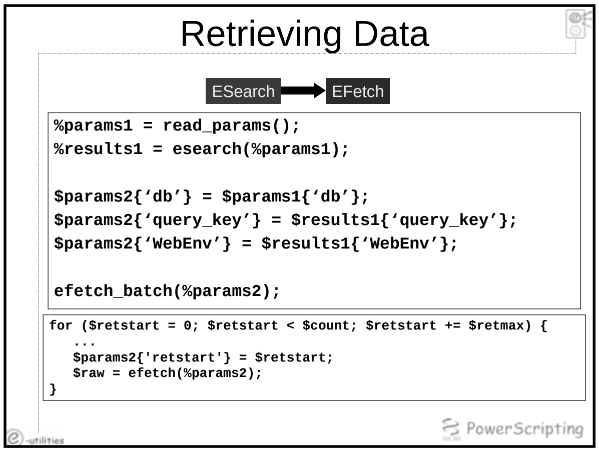
```
$params{‘db’}
$params{‘query_key’}
$params{‘WebEnv’}
```



Retrieving DocSums

ESearch → ESummary

```
%params1 = read_params();
%results1 = esearch(%params1);

$params2{‘db’} = $params1{‘db’};
$params2{‘query_key’} = $results1{‘query_key’};
$params2{‘WebEnv’} = $results1{‘WebEnv’};

%results2 = esummary(%params2);
```

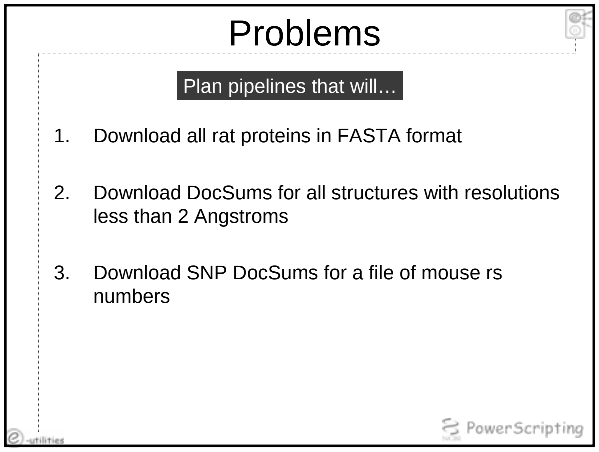


Retrieving Data

ESearch → EFetch

```
%params1 = read_params();
%results1 = esearch(%params1);

$params2{‘db’} = $params1{‘db’};
$params2{‘query_key’} = $results1{‘query_key’};
$params2{‘WebEnv’} = $results1{‘WebEnv’};

efetch_batch(%params2);
```

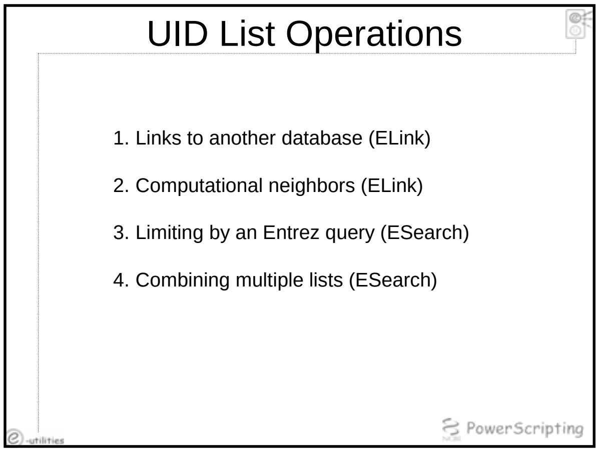
```
for ($retstart = 0; $retstart < $count; $retstart += $retmax) {
   ...
   $params2{'retstart'} = $retstart;
   $raw = efetch(%params2);
}
```



Problems

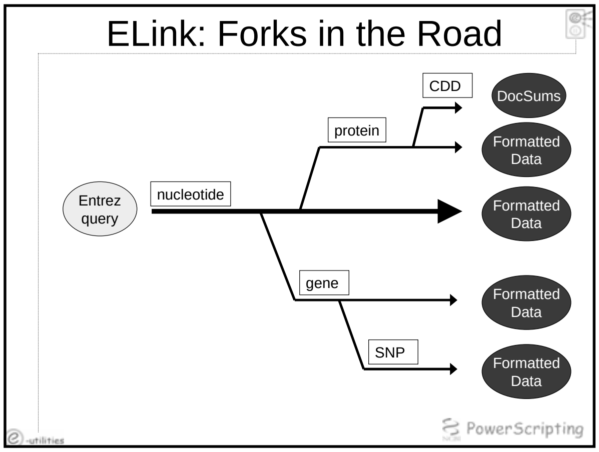
Plan pipelines that will…

1. Download all rat proteins in FASTA format
2. Download DocSums for all structures with resolutions less than 2 Angstroms
3. Download SNP DocSums for a file of mouse rs numbers



UID List Operations

1. Links to another database (ELink)
2. Computational neighbors (ELink)
3. Limiting by an Entrez query (ESearch)
4. Combining multiple lists (ESearch)



ELink: Forks in the Road

Entrez query

nucleotide

protein

CDD

gene

SNP

DocSums

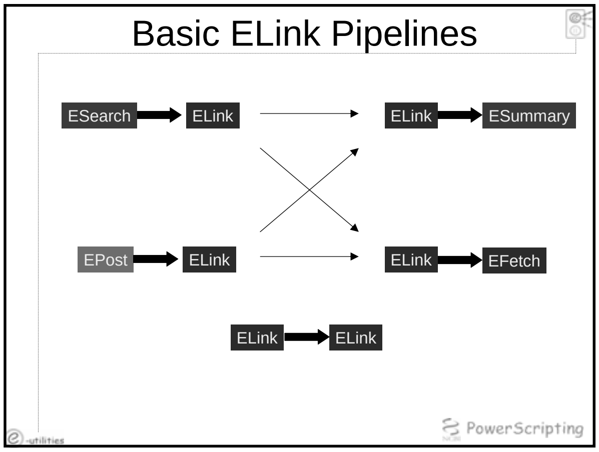
Formatted Data

Formatted Data

Formatted Data

Formatted Data

PowerScripting



Basic ELink Pipelines

ESearch → ELink

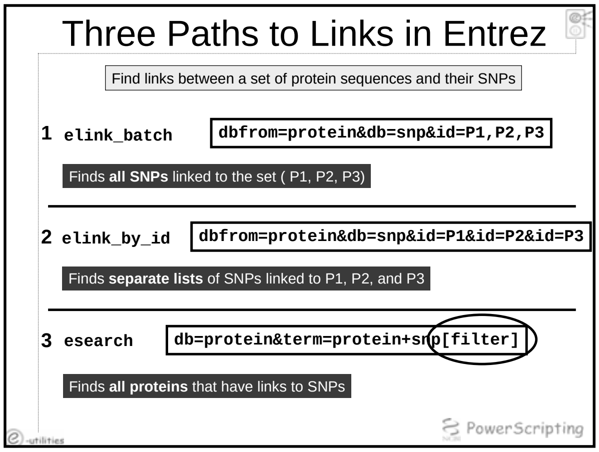
EPost → ELink

ELink → ESummary

ELink → EFetch

ELink → ELink

PowerScripting



Three Paths to Links in Entrez

Find links between a set of protein sequences and their SNPs

1 `elink_batch` `dbfrom=protein&db=snp&id=P1,P2,P3`

Finds **all SNPs** linked to the set (P1, P2, P3)

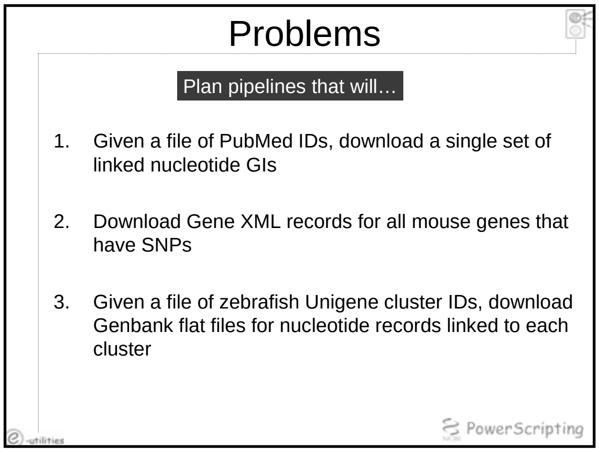
2 `elink_by_id` `dbfrom=protein&db=snp&id=P1&id=P2&id=P3`

Finds **separate lists** of SNPs linked to P1, P2, and P3

3 `esearch` `db=protein&term=protein+snp[filter]`

Finds **all proteins** that have links to SNPs

PowerScripting

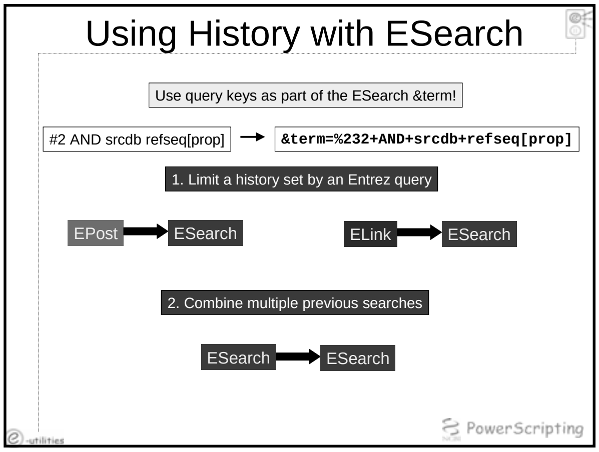


Problems

Plan pipelines that will…

1. Given a file of PubMed IDs, download a single set of linked nucleotide GIs
2. Download Gene XML records for all mouse genes that have SNPs
3. Given a file of zebrafish Unigene cluster IDs, download Genbank flat files for nucleotide records linked to each cluster

PowerScripting



Using History with ESearch

Use query keys as part of the ESearch &term!

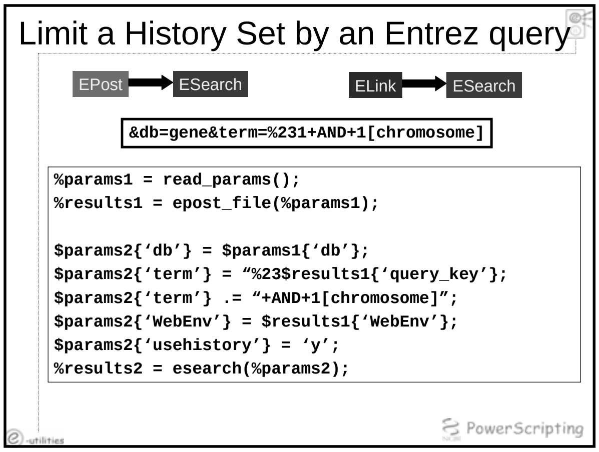
#2 AND srcdb refseq[prop] → `&term=%232+AND+srcdb+refseq[prop]`

1. Limit a history set by an Entrez query

EPost → ESearch   ELink → ESearch

2. Combine multiple previous searches

ESearch → ESearch



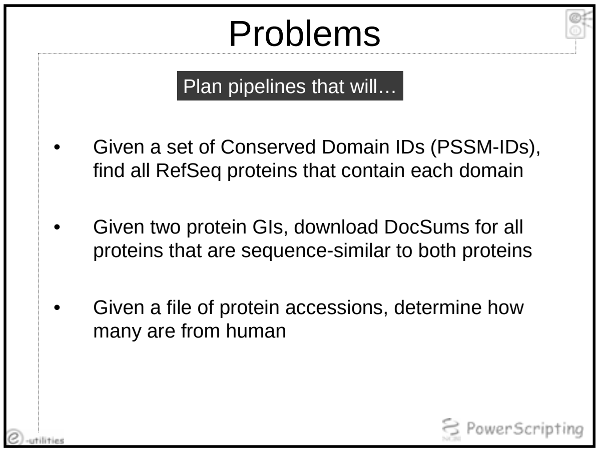
Limit a History Set by an Entrez query

EPost → ESearch   ELink → ESearch

`&db=gene&term=%231+AND+1[chromosome]`

```
%params1 = read_params();
%results1 = epost_file(%params1);

$params2{‘db’} = $params1{‘db’};
$params2{‘term’} = “%23$results1{‘query_key’};
$params2{‘term’} .= “+AND+1[chromosome]”;
$params2{‘WebEnv’} = $results1{‘WebEnv’};
$params2{‘usehistory’} = ‘y’;
%results2 = esearch(%params2);
```

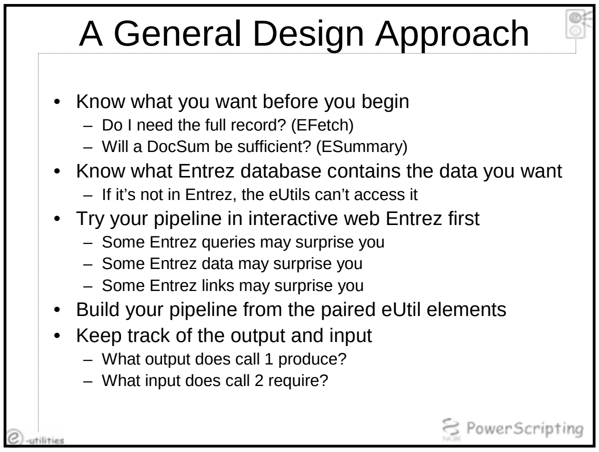
PowerScripting



Problems

Plan pipelines that will…

- Given a set of Conserved Domain IDs (PSSM-IDs), find all RefSeq proteins that contain each domain
- Given two protein GIs, download DocSums for all proteins that are sequence-similar to both proteins
- Given a file of protein accessions, determine how many are from human



A General Design Approach

- Know what you want before you begin
  - Do I need the full record? (EFetch)
  - Will a DocSum be sufficient? (ESummary)
- Know what Entrez database contains the data you want
  - If it’s not in Entrez, the eUtils can’t access it
- Try your pipeline in interactive web Entrez first
  - Some Entrez queries may surprise you
  - Some Entrez data may surprise you
  - Some Entrez links may surprise you
- Build your pipeline from the paired eUtil elements
- Keep track of the output and input
  - What output does call 1 produce?
  - What input does call 2 require?