HATCH

USGS-CRRP

?

Why?

Introduction

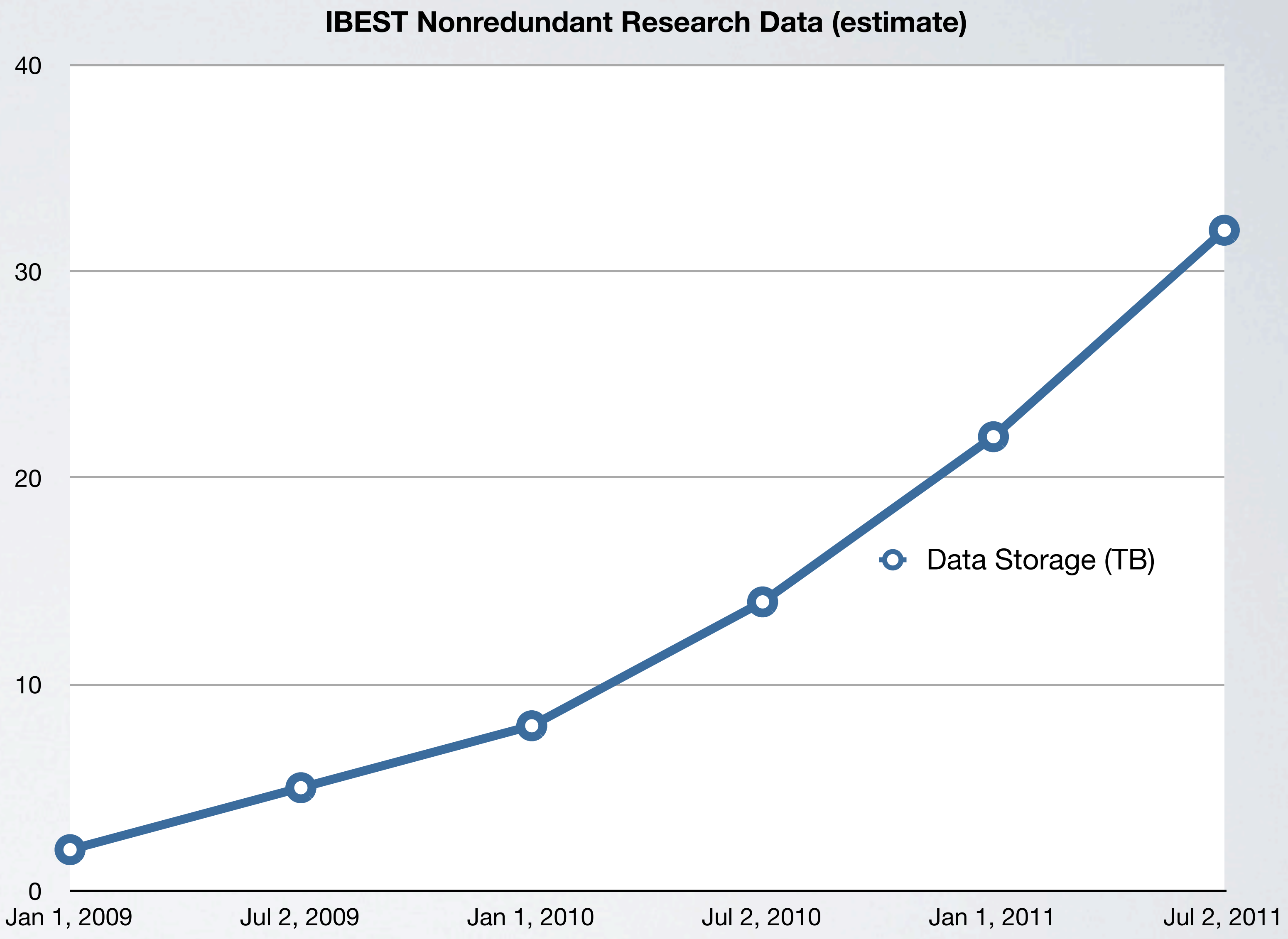
- Exponential scientific data growth
- CPU computer power cannot keep up
- Huge emerging gap between researchers and results
- Big problems will remain unanswered
- Less investment in computing equipment, more in staff

[1] Szalay, Alex; Gray, Jim. **“2020 Computing: Science in an exponential world”**. Nature 440, 413-414 (23 March 2006)

[2] Foster, James. **Visualizing Human Microbiome Ecosystems**. University of Idaho: Computer Science Colloquium, December 7th 2010. Seminar.

Exponential data growth

- Doubling every 2 years [1]
- U of I IBEST data is ~33 TB, growing
- <10% data used in final research [2].



[1] Szalay, Alex; Gray, Jim. **"2020 Computing: Science in an exponential world"**. Nature 440, 413-414 (23 March 2006)

[2] Foster, James. **Visualizing Human Microbiome Ecosystems**. University of Idaho: Computer Science Colloquium, December 7th 2010. Seminar.

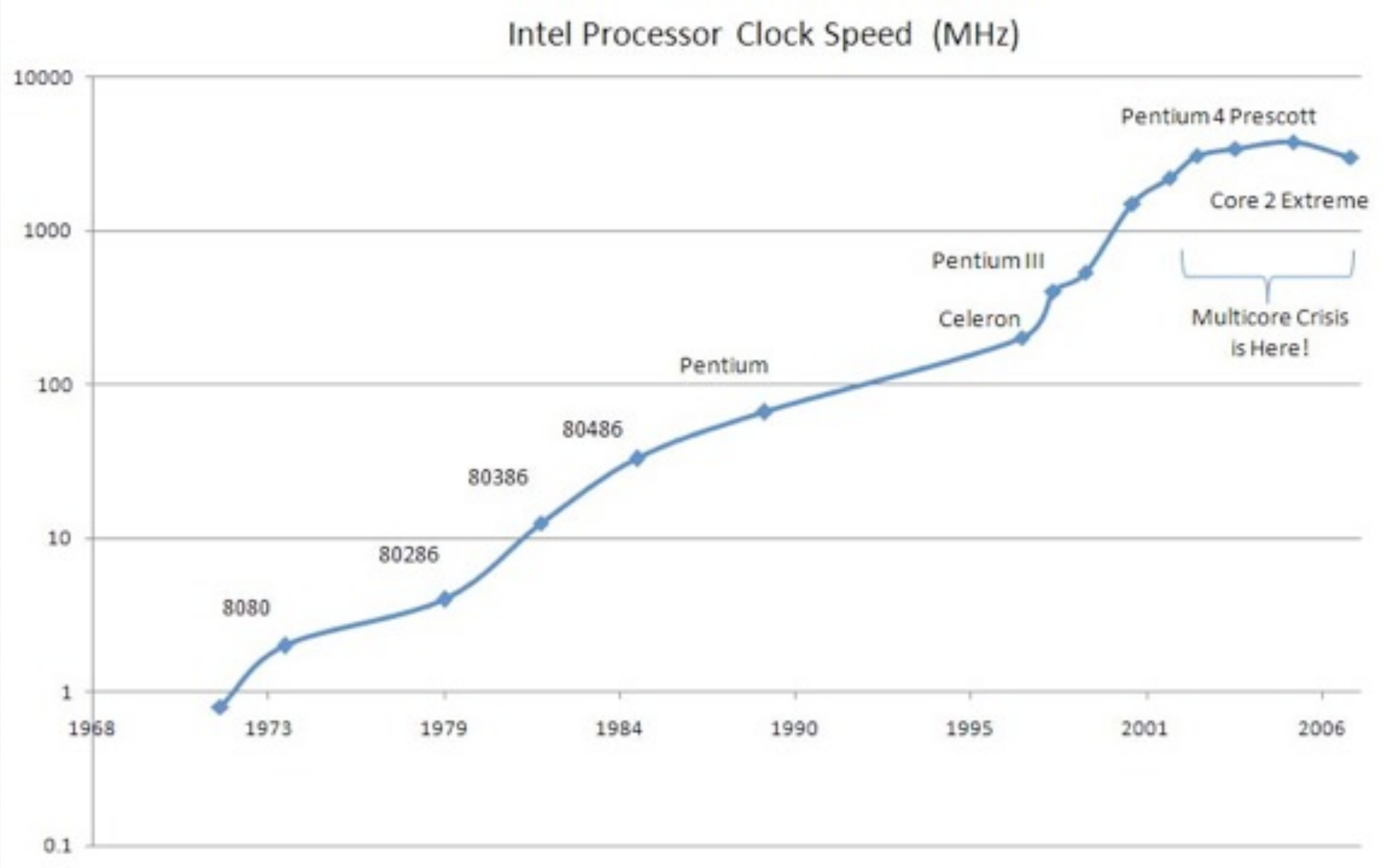
[1] Szalay, Alex; Gray, Jim. **“2020 Computing: Science in an exponential world”**. Nature 440, 413-414 (23 March 2006)

[2] Foster, James. **Visualizing Human Microbiome Ecosystems**. University of Idaho: Computer Science Colloquium, December 7th 2010. Seminar.

[3] Manek Dubash (2005-04-13). "Moore's Law is dead, says Gordon Moore". Techworld. Retrieved 2006-06-24

The death of Moore's Law

- Back in 1965, Moore's Law declared [3]
- Transistors density doubled every 2 years
- CPU power roughly the same
- Growth continued for 40 years
- Transistors approach to size of atom
- Who boldly declared the law dead?

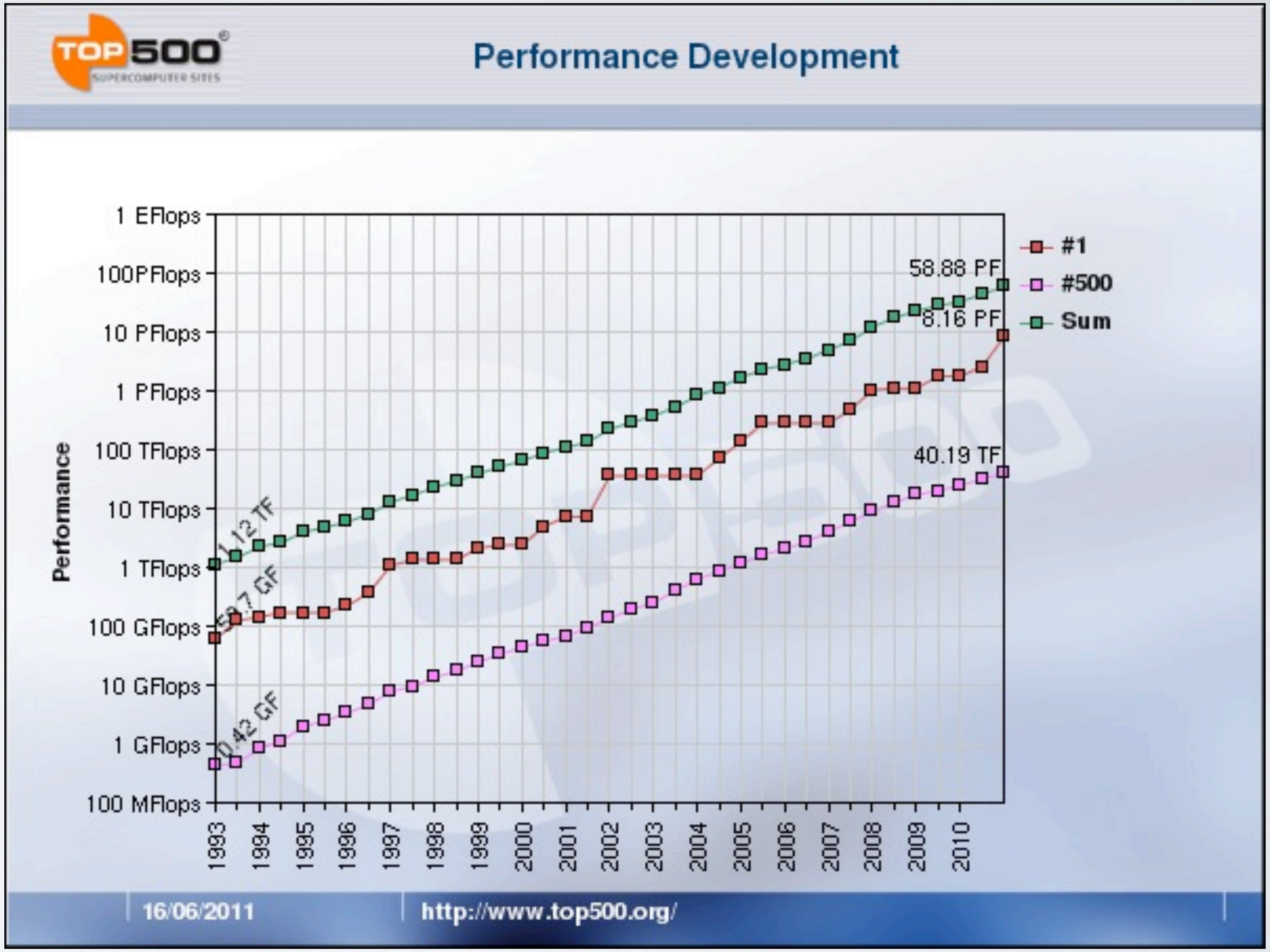


[3] Manek Dubash (2005-04-13). "Moore's Law is dead, says Gordon Moore". Techworld. Retrieved 2006-06-24

[3] Manek Dubash (2005-04-13). "Moore's Law is dead, says Gordon Moore". Techworld. Retrieved 2006-06-24

CPU Chips cannot keep up

- High Performance Compute Clusters
- Groups of computers, one huge CPU
- Amazingly powerful and complicated



Today's scientific computation is too complicated

- the gap between data and analysis is broad
- data assets become management liabilities
- unnecessary computation and data redundancy is common
- the research workspace is more fractured than it is unified

An interactive solution is needed

- requires no new training
- maximizes use of existing standards and libraries
- takes advantage of the existing use cases
- answers the current administrative questions
- doesn't create new ones

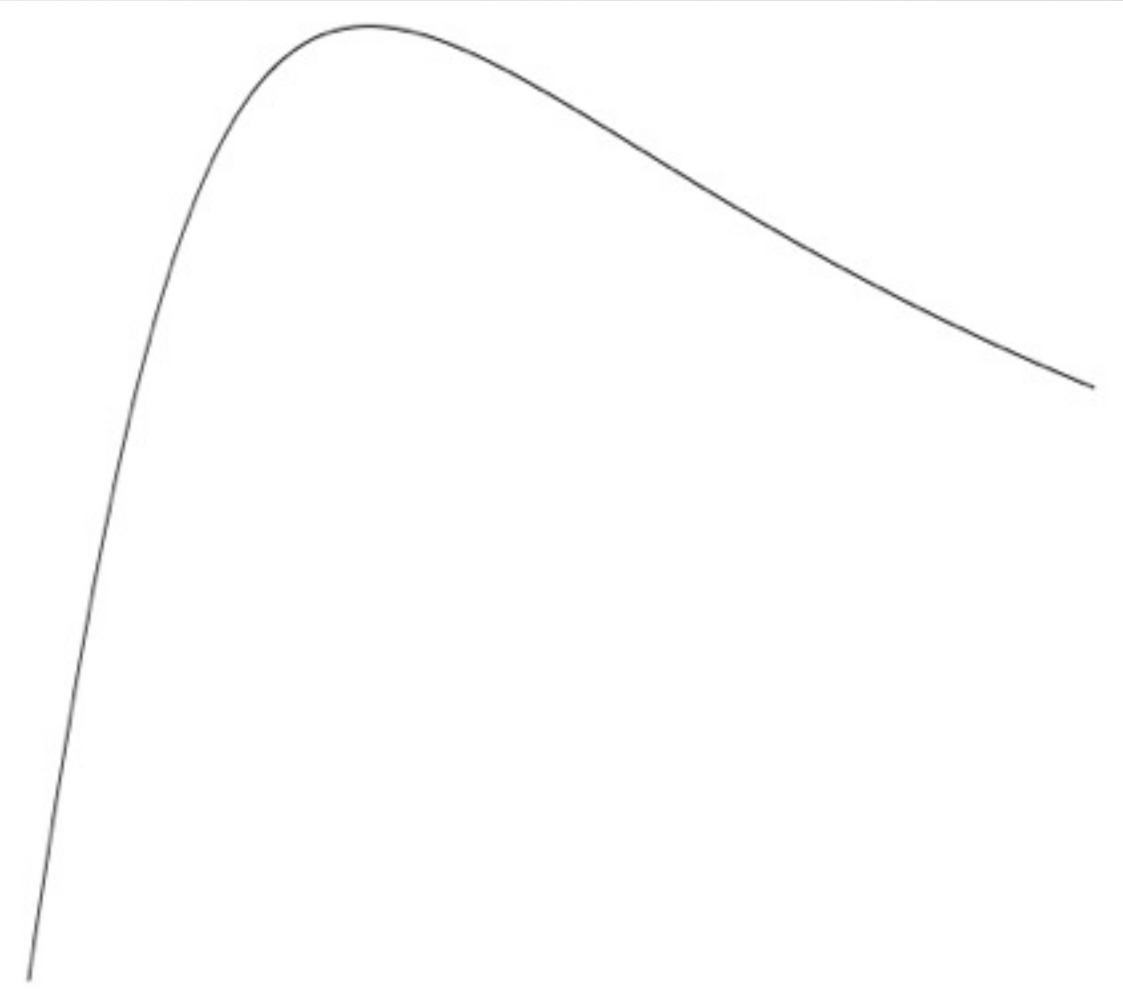
Hatch

?

What is Hatch?

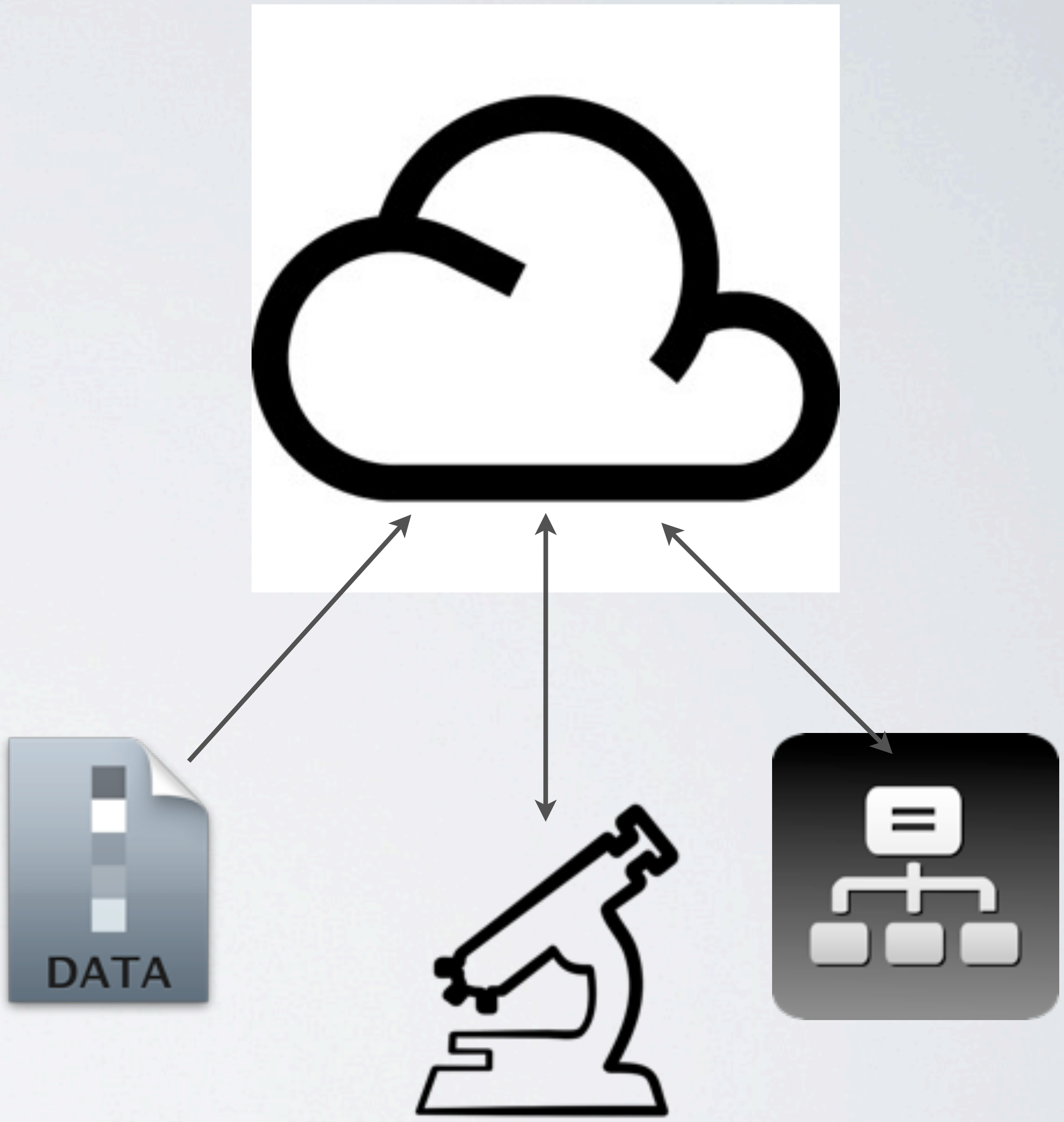
Hatch is a simplified:

- data visualization and filtration system
- entry for local data to a research cloud
- minimization of redundancy in data sharing
- linkage between data acquisition and analysis

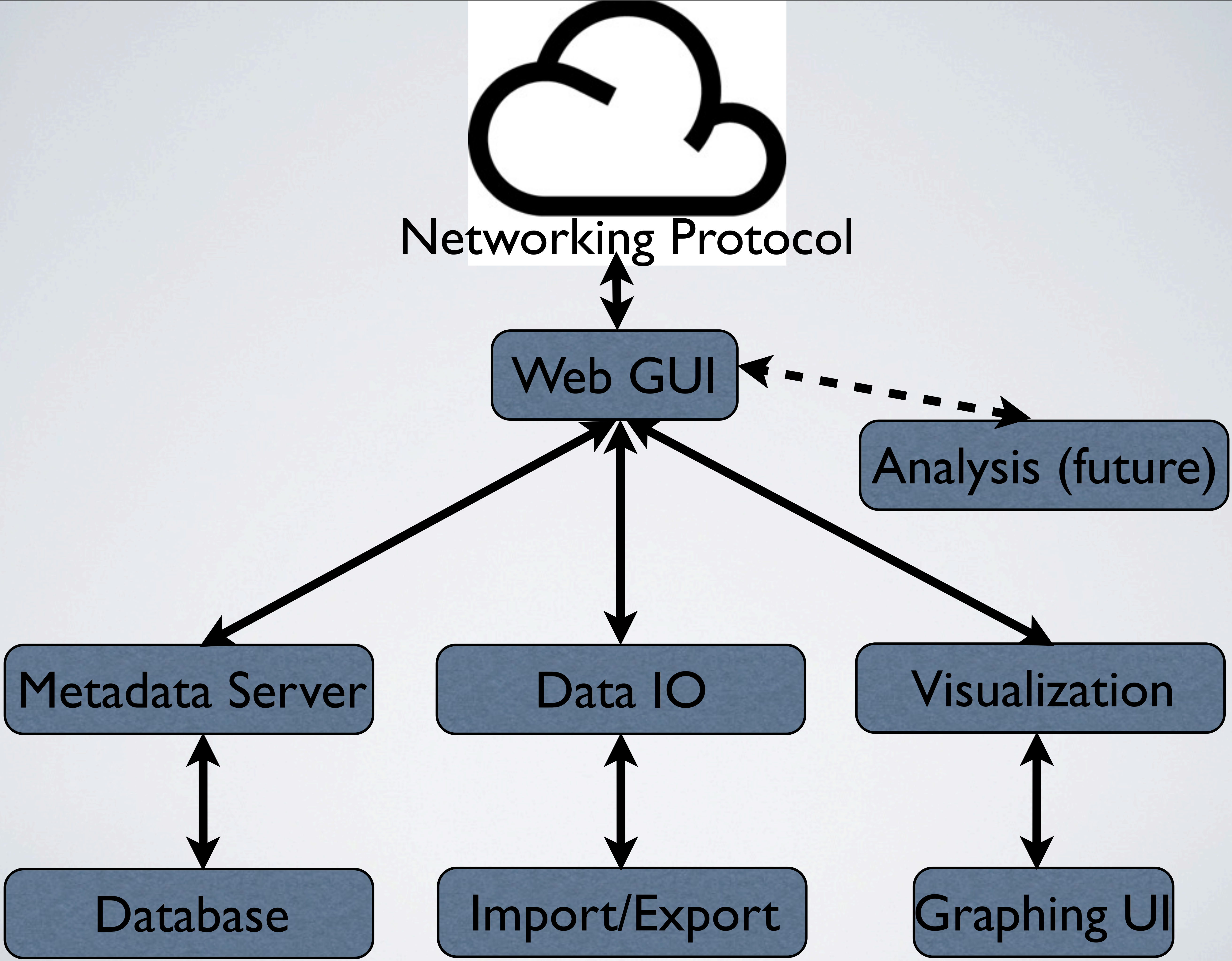


A graphical way to manage:

- data sets
- computational resources
- cloud network topologies
- processes



What is Hatch?



Networking Protocol

Web GUI

Analysis (future)

Metadata Server

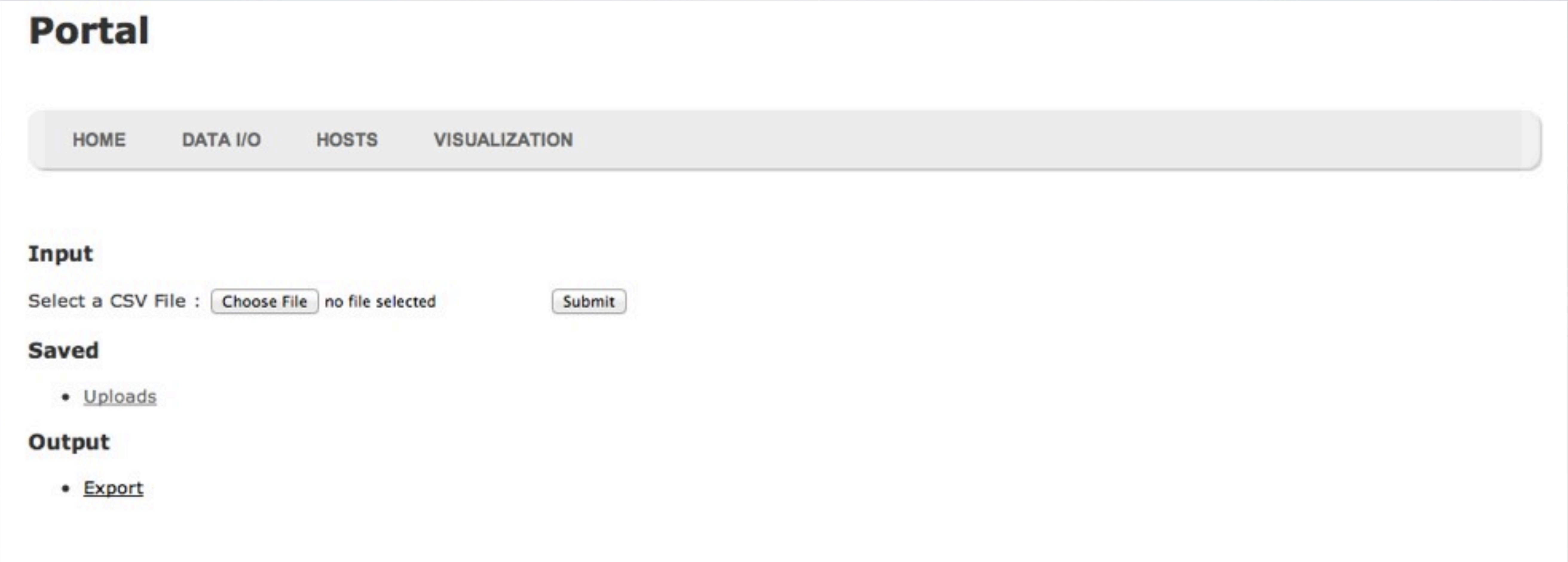
Data IO

Visualization

Database

Import/Export

Graphing UI



Portal

Input

Select a CSV File : no file selected

Saved

- Uploads

Output

- Export

- decentralized cloud
- seamless data management
- communication encouragement
- preservation of group autonomy

Completed so far

Import Data

- Currently, EcoData supports data that is formatted in well-structured Comma-Separated Values (CSV) format.
- A CSV data file is uploaded via the web interface (pictured above) and stored in the database.

Manipulate Data

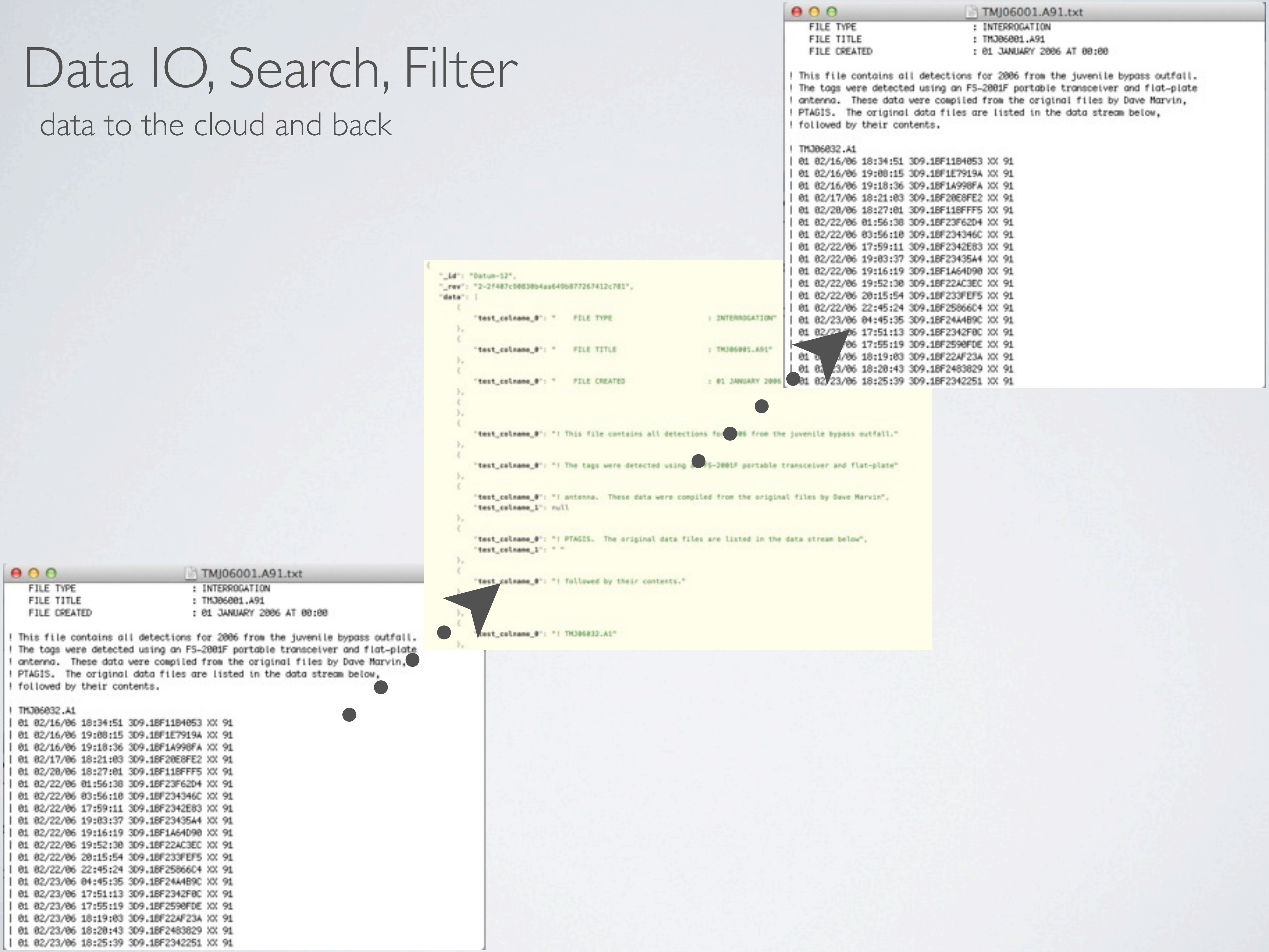
- Currently all data must be manipulated before being imported into EcoData. This limitation will be overcome as the database backend is finalized.

Visualize Data

- A visualization framework has been put in place and is nearly ready to accept user data. Once the database structure is finalized, data will be graphed as seen in "Visualizations in the Web Browser."

Data IO, Search, Filter

data to the cloud and back

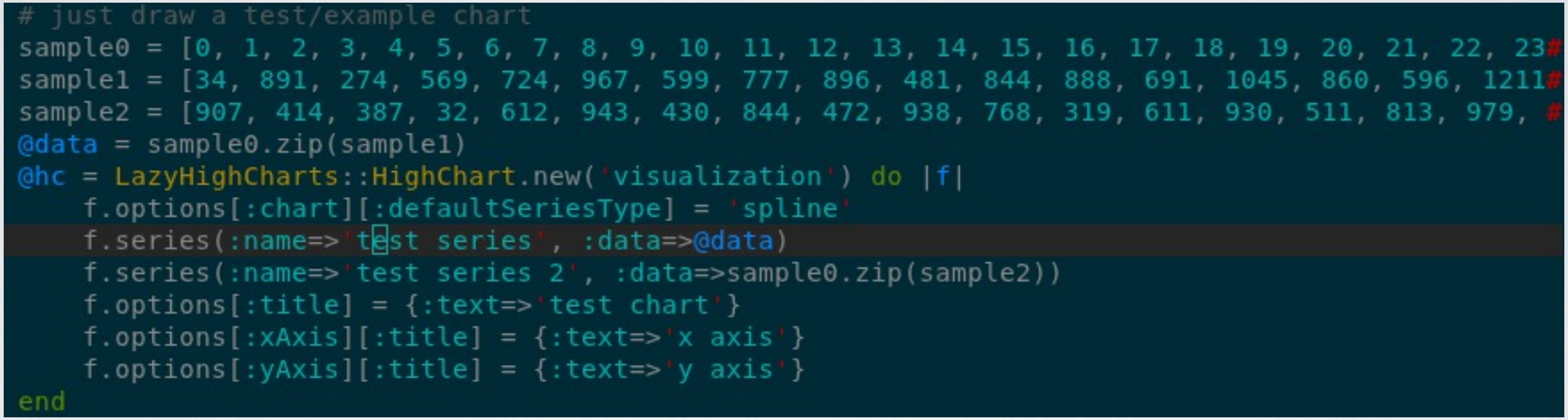


```
   FILE TYPE                    : INTERROGATION
   FILE TITLE                   : TMJ06001.A91
   FILE CREATED                 : 01 JANUARY 2006 AT 00:00

! This file contains all detections for 2006 from the juvenile bypass outfall.
! The tags were detected using an FS-2001F portable transceiver and flat-plate
! antenna.  These data were compiled from the original files by Dave Marvin,
! PTAGIS.  The original data files are listed in the data stream below,
! followed by their contents.

! TMJ06032.A1
| 01 02/16/06 18:34:51 3D9.1BF11B4053 XX 91
| 01 02/16/06 19:00:15 3D9.1BF1E7919A XX 91
| 01 02/16/06 19:18:36 3D9.1BF1A998FA XX 91
| 01 02/17/06 18:21:03 3D9.1BF20E8FE2 XX 91
| 01 02/20/06 18:27:01 3D9.1BF11BFFF5 XX 91
| 01 02/22/06 01:56:38 3D9.1BF23F62D4 XX 91
| 01 02/22/06 03:56:10 3D9.1BF234346C XX 91
| 01 02/22/06 17:59:11 3D9.1BF2342E83 XX 91
| 01 02/22/06 19:03:37 3D9.1BF23435A4 XX 91
| 01 02/22/06 19:16:19 3D9.1BF1A64D90 XX 91
| 01 02/22/06 19:52:30 3D9.1BF22AC3EC XX 91
| 01 02/22/06 20:15:54 3D9.1BF233FEF5 XX 91
| 01 02/22/06 22:45:24 3D9.1BF25866C4 XX 91
| 01 02/23/06 04:45:35 3D9.1BF24A4B9C XX 91
| 01 02/23/06 17:51:13 3D9.1BF2342F0C XX 91
| 01 02/23/06 17:55:19 3D9.1BF2590FDE XX 91
| 01 02/23/06 18:19:03 3D9.1BF22AF23A XX 91
| 01 02/23/06 18:20:43 3D9.1BF2483829 XX 91
| 01 02/23/06 18:25:39 3D9.1BF2342251 XX 91
```

```
{
   "_id": "Datum-12",
   "_rev": "2-2f407c90830b4aa649b877267412c781",
   "data": [
       {
           "test_colname_0": "    FILE TYPE                    : INTERROGATION"
       },
       {
           "test_colname_0": "    FILE TITLE                   : TMJ06001.A91"
       },
       {
           "test_colname_0": "    FILE CREATED                 : 01 JANUARY 2006
       },
       {
       },
       {
           "test_colname_0": "! This file contains all detections for 2006 from the juvenile bypass outfall."
       },
       {
           "test_colname_0": "! The tags were detected using an FS-2001F portable transceiver and flat-plate"
       },
       {
           "test_colname_0": "! antenna.  These data were compiled from the original files by Dave Marvin",
           "test_colname_1": null
       },
       {
           "test_colname_0": "! PTAGIS.  The original data files are listed in the data stream below",
           "test_colname_1": " "
       },
       {
           "test_colname_0": "! followed by their contents."
       },
       {
       },
       {
           "test_colname_0": "! TMJ06032.A1"
       },
```

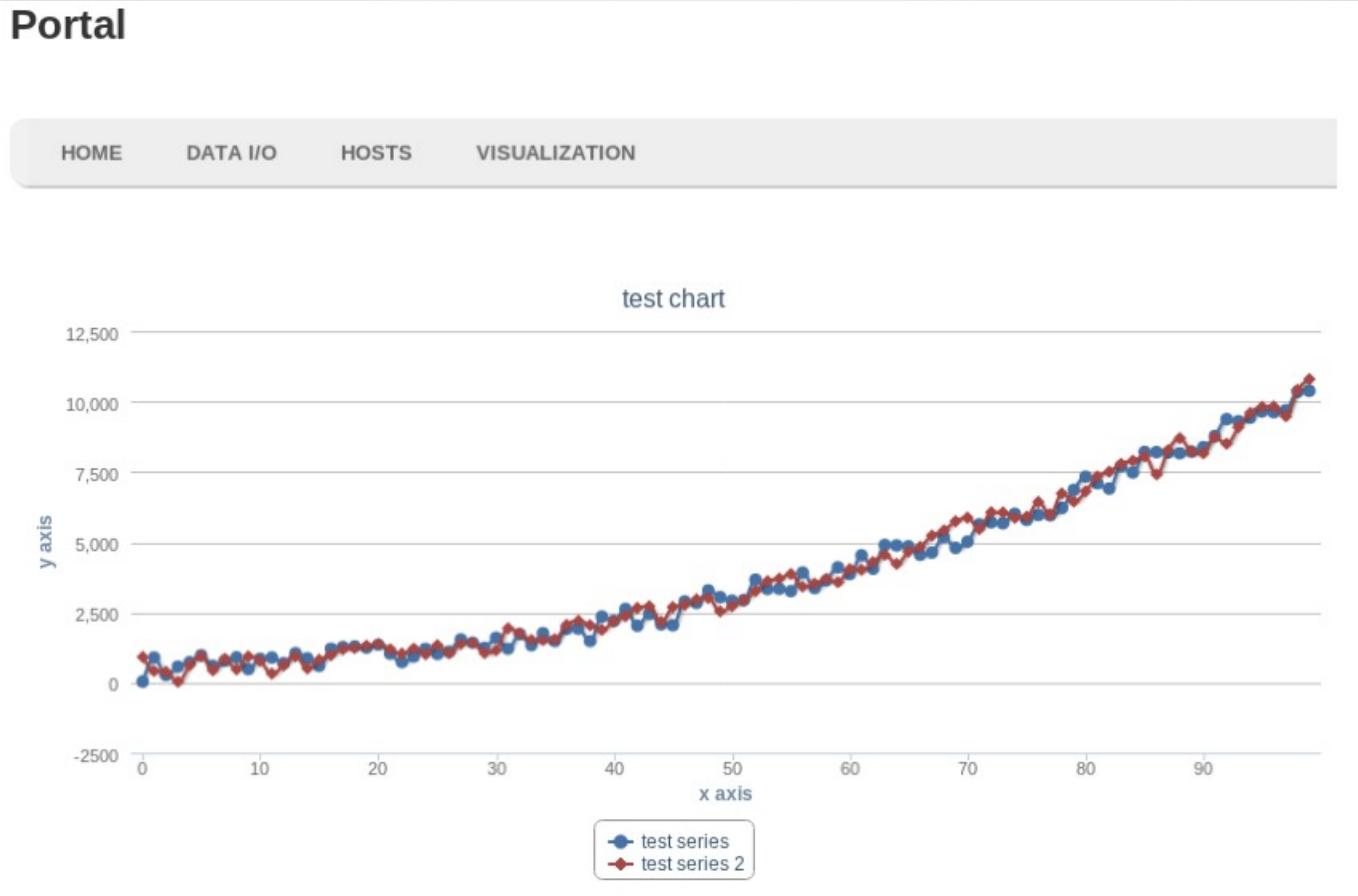
```
   FILE TYPE                    : INTERROGATION
   FILE TITLE                   : TMJ06001.A91
   FILE CREATED                 : 01 JANUARY 2006 AT 00:00

! This file contains all detections for 2006 from the juvenile bypass outfall.
! The tags were detected using an FS-2001F portable transceiver and flat-plate
! antenna.  These data were compiled from the original files by Dave Marvin,
! PTAGIS.  The original data files are listed in the data stream below,
! followed by their contents.

! TMJ06032.A1
| 01 02/16/06 18:34:51 3D9.1BF11B4053 XX 91
| 01 02/16/06 19:00:15 3D9.1BF1E7919A XX 91
| 01 02/16/06 19:18:36 3D9.1BF1A998FA XX 91
| 01 02/17/06 18:21:03 3D9.1BF20E8FE2 XX 91
| 01 02/20/06 18:27:01 3D9.1BF11BFFF5 XX 91
| 01 02/22/06 01:56:38 3D9.1BF23F62D4 XX 91
| 01 02/22/06 03:56:10 3D9.1BF234346C XX 91
| 01 02/22/06 17:59:11 3D9.1BF2342E83 XX 91
| 01 02/22/06 19:03:37 3D9.1BF23435A4 XX 91
| 01 02/22/06 19:16:19 3D9.1BF1A64D90 XX 91
| 01 02/22/06 19:52:30 3D9.1BF22AC3EC XX 91
| 01 02/22/06 20:15:54 3D9.1BF233FEF5 XX 91
| 01 02/22/06 22:45:24 3D9.1BF25866C4 XX 91
| 01 02/23/06 04:45:35 3D9.1BF24A4B9C XX 91
| 01 02/23/06 17:51:13 3D9.1BF2342F0C XX 91
| 01 02/23/06 17:55:19 3D9.1BF2590FDE XX 91
| 01 02/23/06 18:19:03 3D9.1BF22AF23A XX 91
| 01 02/23/06 18:20:43 3D9.1BF2483829 XX 91
| 01 02/23/06 18:25:39 3D9.1BF2342251 XX 91
```

Sample data

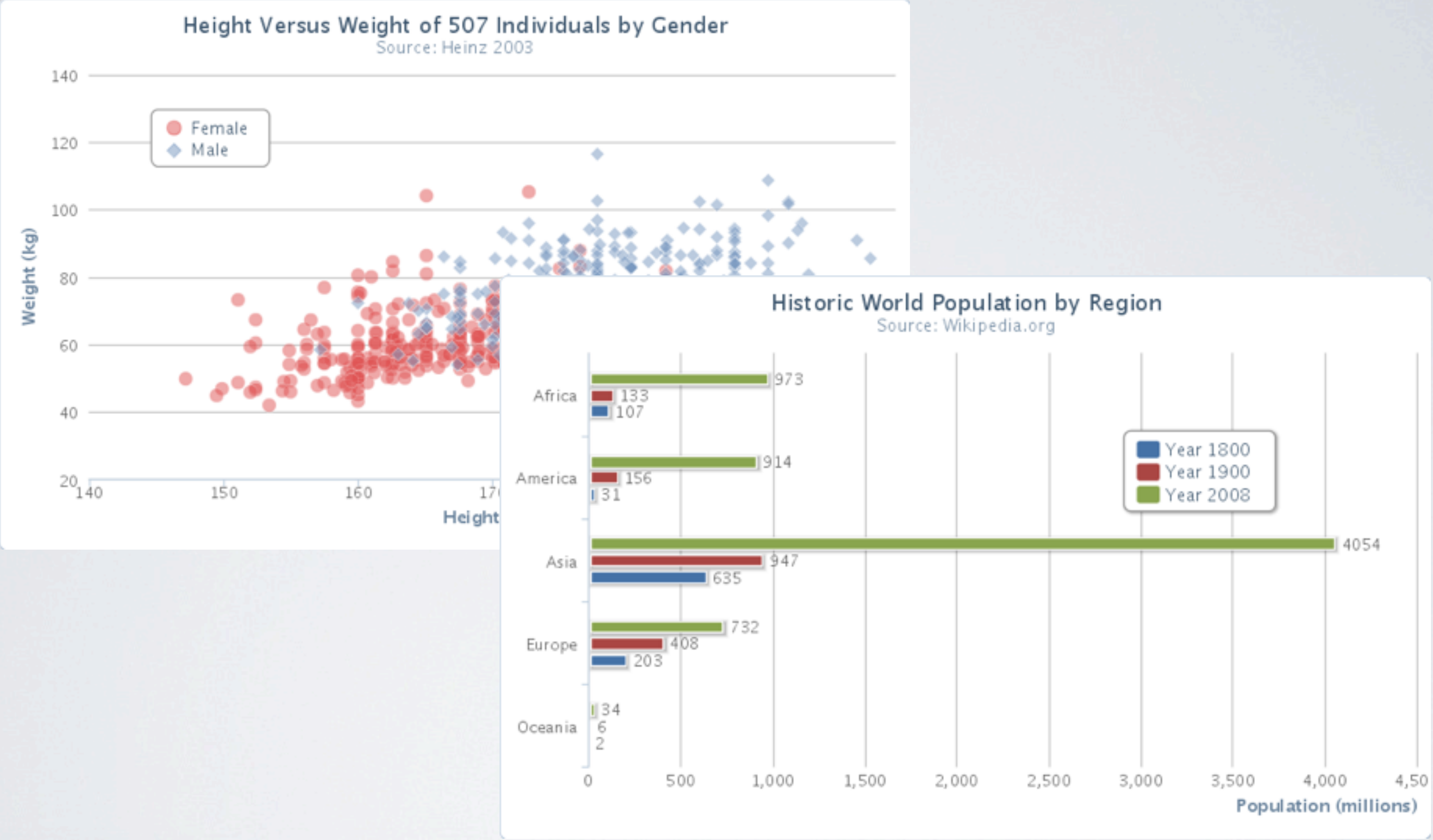


```
# just draw a test/example chart
sample0 = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23#
sample1 = [34, 891, 274, 569, 724, 967, 599, 777, 896, 481, 844, 888, 691, 1045, 860, 596, 1211#
sample2 = [907, 414, 387, 32, 612, 943, 430, 844, 472, 938, 768, 319, 611, 930, 511, 813, 979, #
@data = sample0.zip(sample1)
@hc = LazyHighCharts::HighChart.new('visualization') do |f|
    f.options[:chart][:defaultSeriesType] = 'spline'
    f.series(:name=>'test series', :data=>@data)
    f.series(:name=>'test series 2', :data=>sample0.zip(sample2))
    f.options[:title] = {:text=>'test chart'}
    f.options[:xAxis][:title] = {:text=>'x axis'}
    f.options[:yAxis][:title] = {:text=>'y axis'}
end
```

Sample graph



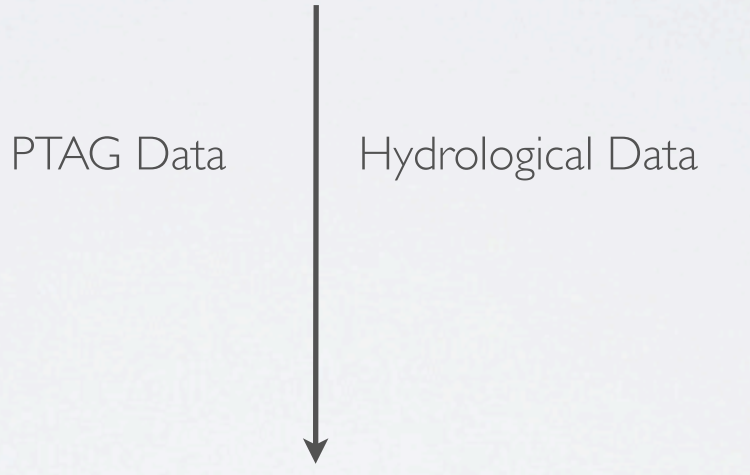
Sample graphs



College of
Natural Resources

University of Idaho

Alex Fremier, Associate Professor at the University of Idaho College of Natural Resources



US Fish and Game, Methow River

Customers

UAF™
UNIVERSITY OF ALASKA
FAIRBANKS

PTAGIS

NSF

Columbia River DART
Data Access in Real Time

Who else?

Trial and error

- coupling design too closely to the data
- tool dependencies
- coupling data too closely to the tools
- tool module coupling

Long term goals

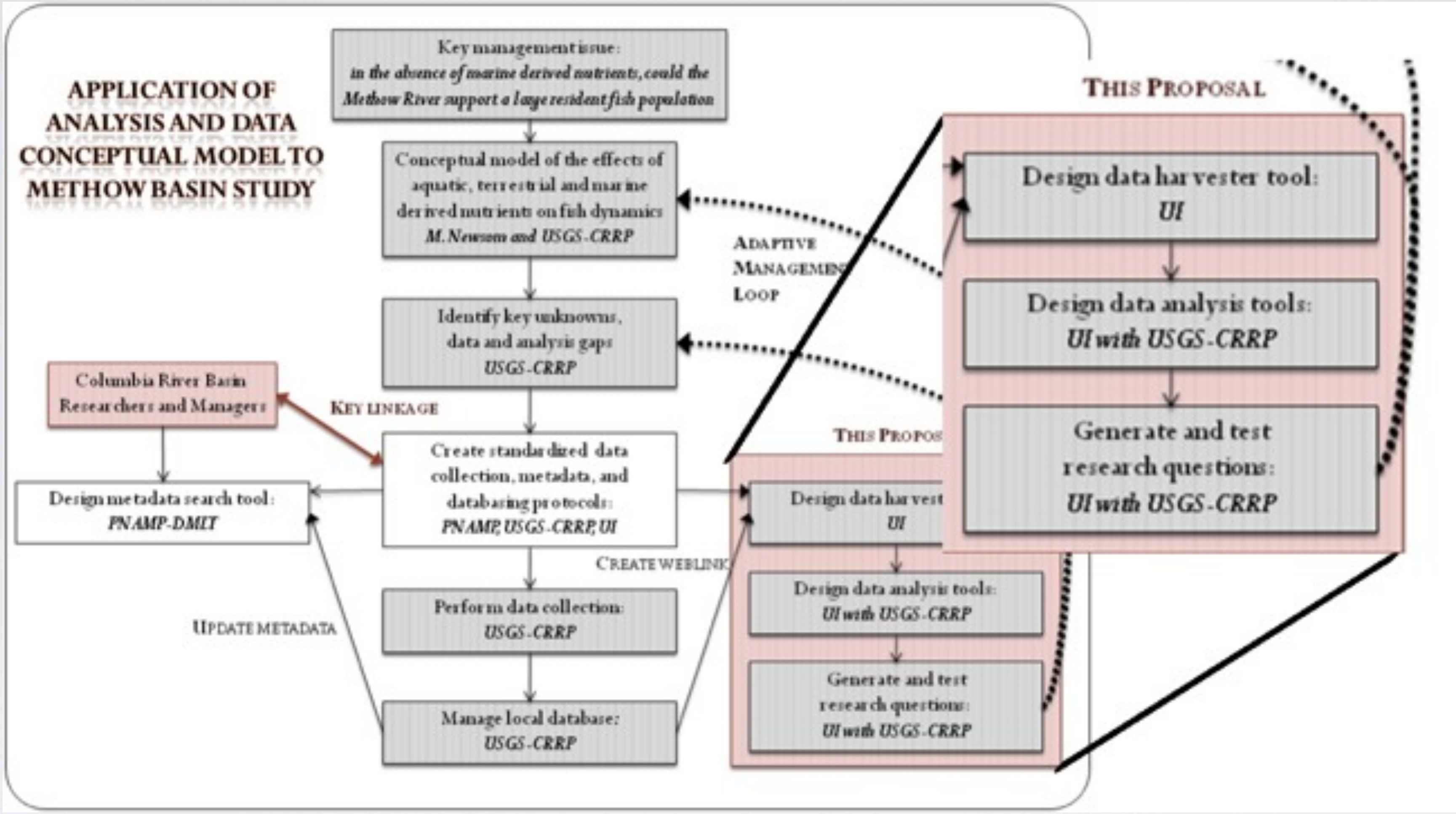
While the simplest use of EcoData is a simple process involving importing and storage of data followed by manipulations and visualizations, each stage of the process can and will be elaborated upon.

- Import data
  - Support other data types
- Manipulate Data
  - Filtering
  - Joins between tables of data
- Visualize data
  - Add more types of visualizations

Other objectives include:

- Authentication and security for data
- Availability of data between servers
- Cross-server data availability
  - Fast transfers of data between remote servers

Future of Portal


APPLICATION OF ANALYSIS AND DATA CONCEPTUAL MODEL TO METHOW BASIN STUDY
Key management issue: in the absence of marine derived nutrients, could the Methow River support a large resident fish population
Conceptual model of the effects of aquatic, terrestrial and marine derived nutrients on fish dynamics M. Newsom and USGS-CRRP
Identify key unknowns, data and analysis gaps USGS-CRRP
Columbia River Basin Researchers and Managers
KEY LINKAGE
Create standardized data collection, metadata, and databasing protocols: PNAMP, USGS-CRRP, UI
Design metadata search tool: PNAMP-DMIT
UPDATE METADATA
CREATE WEBLINK
Perform data collection: USGS-CRRP
Manage local database: USGS-CRRP
ADAPTIVE MANAGEMENT LOOP
THIS PROPOSAL
Design data harvester tool: UI
Design data analysis tools: UI with USGS-CRRP
Generate and test research questions: UI with USGS-CRRP
THIS PROPOSAL
Design data harvester tool: UI
Design data analysis tools: UI with USGS-CRRP
Generate and test research questions: UI with USGS-CRRP

- Portal is the first piece in a larger data management and analysis system.
- It provides the tools necessary to convert and store data in a structured format and to produce basic visualizations of that data.
- Eventually, it will be used as one piece in a more complex data analysis pipeline, allowing for more sophisticated data manipulations and visualizations.

Conclusion

- Dire need of more processing power
- Data collection eclipsed by the inability to analyze it
- Big answers lie in the balance
- Data deluge will hold back advances
- People disciplined in computers are needed
- NSF and NIH grants favor funding data collection
- Actually analyzing the data
- Research will become less and less meaningful