Classifying congestion in Ark measurements

Steven Bauer

MIT

March 31, 2015

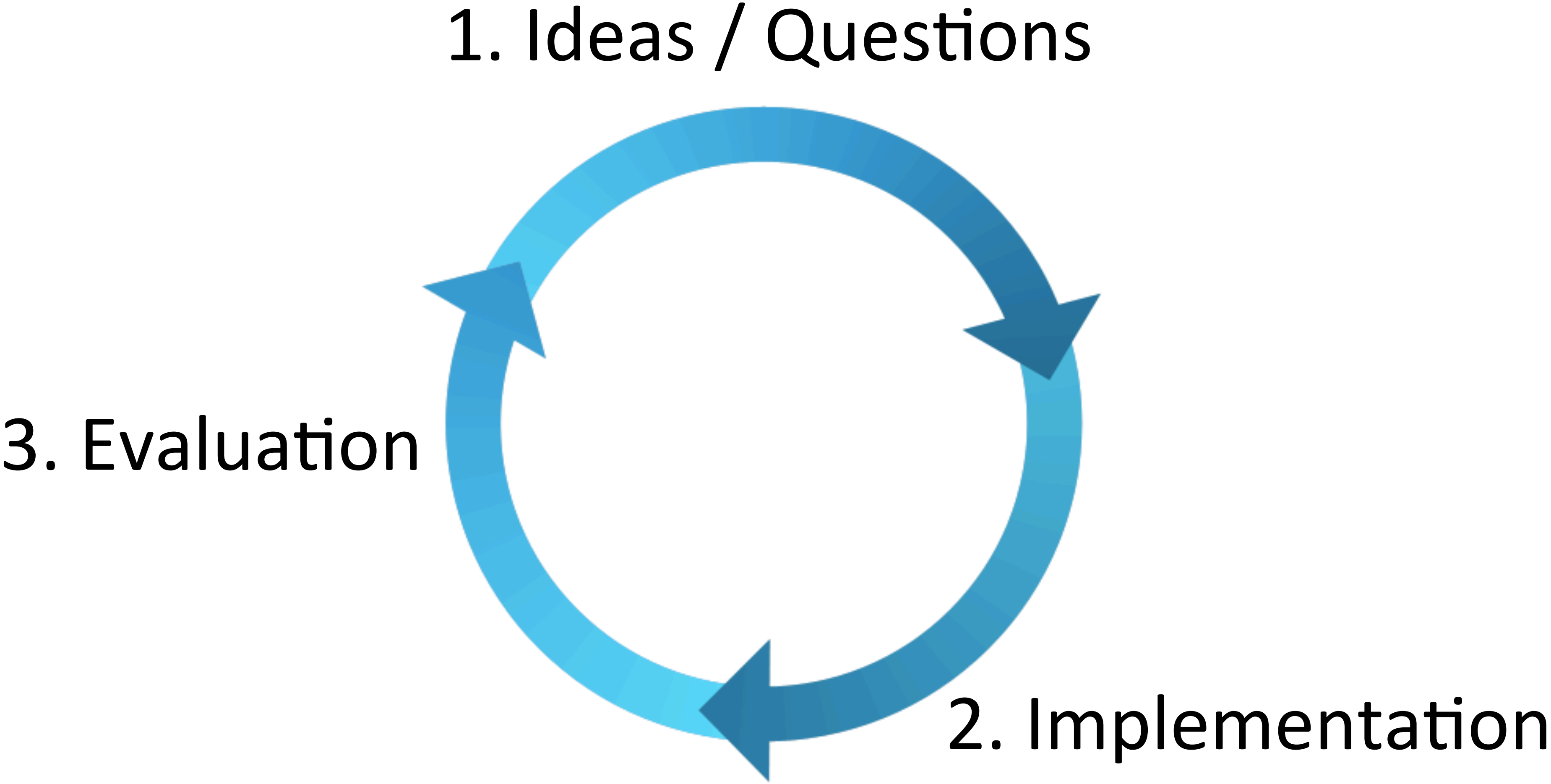
How Ark (and the Ark community) could further my research

We should do a better job sharing best common practices and lessons learned from working with large networking data sets

I am particularly interested in the:

- Ark data
- FCC / Samknows data
- Measurement Lab data

Accelerate this loop



1. Ideas / Questions

2. Implementation

3. Evaluation

Lots of possible solutions (and problems)

Compute

Spark

STORM

hadoop

Infrastructure

docker

openstack™

MESOS

Storage

SciDB

Neo4j

PostgreSQL

(Just some examples)

Some observations on how other communities facilitate replicating results

- IPython Notebooks: all the rage in some communities
  - Facilitate easy exploration and initial experimentation of code and data
  - Entire books with text, code, data, and visual results bound together
- Other scientific communities have extensive experience and lessons learned from data sharing

Lots of work has been funded to make sharing large amounts of complicated scientific data easier

OPeNDAP

NOAA

ERDDAP

Easier access to scientific data

HDF

The HDF Group

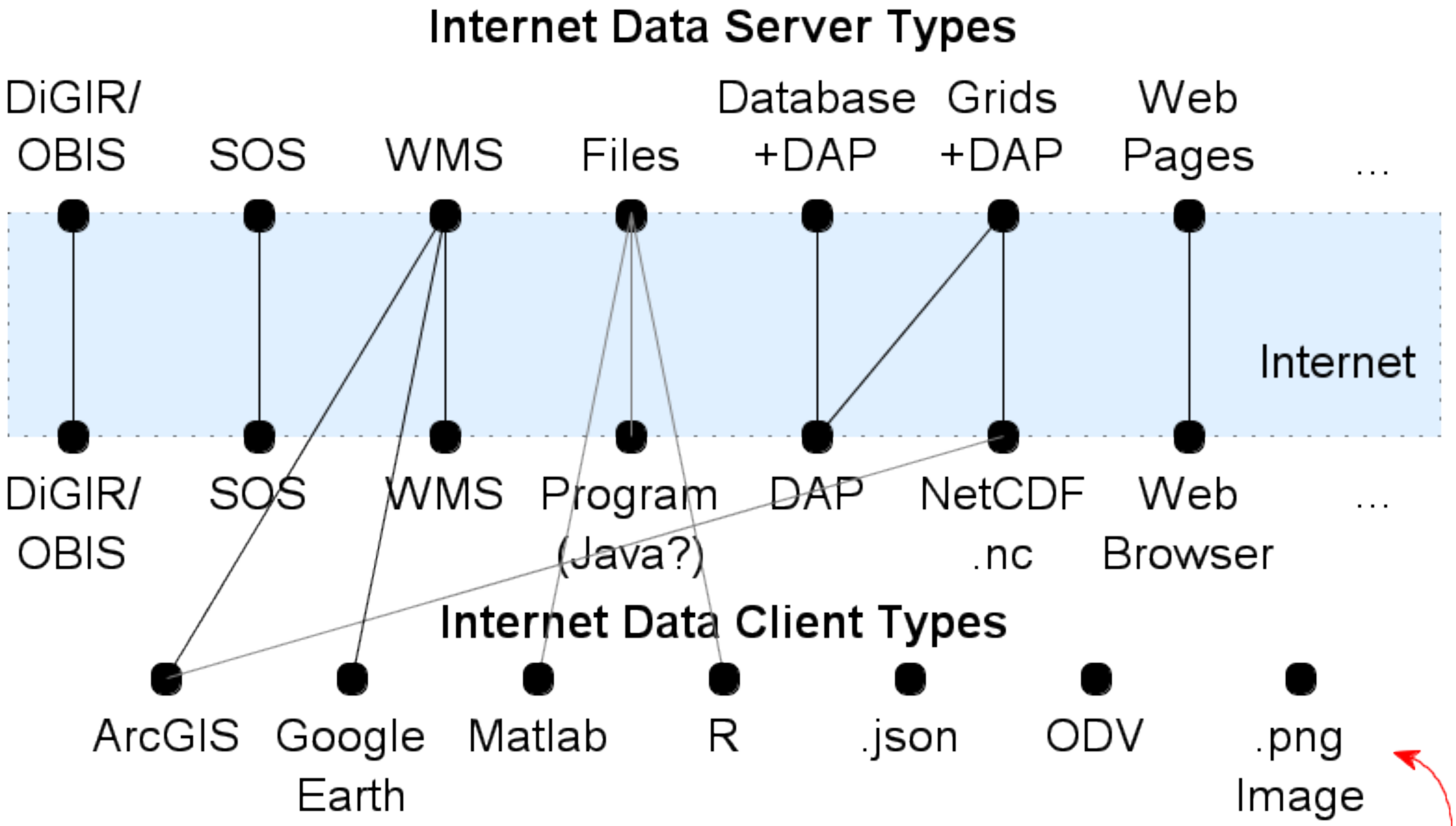
netCDF

CBOR

RFC 7049

(Just some examples)

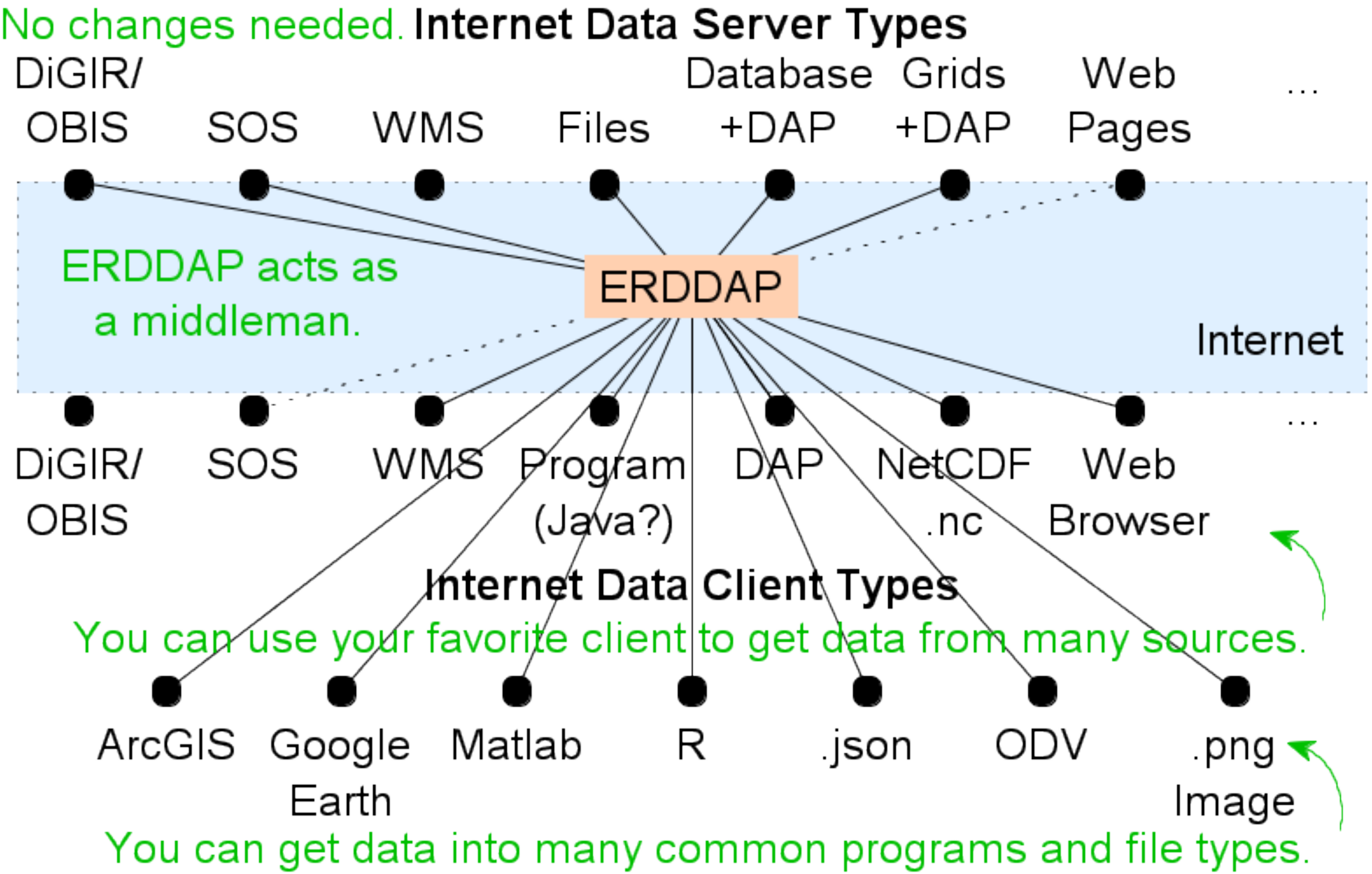
Different communities use different data servers. Each is fine by itself. But they all work differently!



So there's no easy way to get data into other programs or file types.

Slide from http://www.opendap.org/sites/default/files/TabularData.pptx

ERDDAP solves those problems by acting as a middleman.



No changes needed. **Internet Data Server Types**

DiGIR/ OBIS

SOS

WMS

Files

Database +DAP

Grids +DAP

Web Pages

...

ERDDAP acts as a middleman.

ERDDAP

Internet

DiGIR/ OBIS

SOS

WMS

Program (Java?)

DAP

NetCDF .nc

Web Browser

...

Internet Data Client Types

You can use your favorite client to get data from many sources.

ArcGIS

Google Earth

Matlab

R

.json

ODV

.png Image

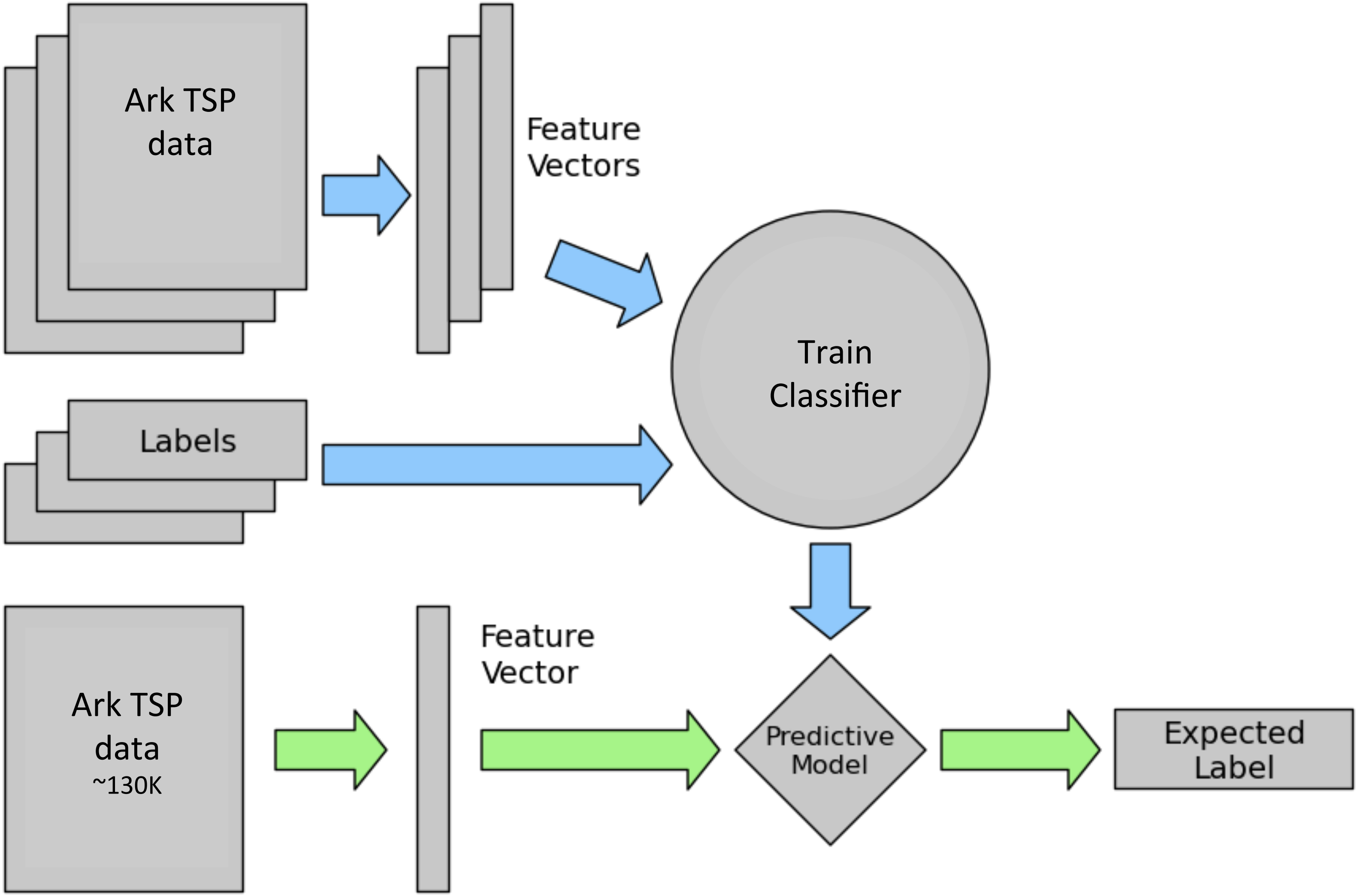
You can get data into many common programs and file types.

Slide from http://www.opendap.org/sites/default/files/TabularData.pptx

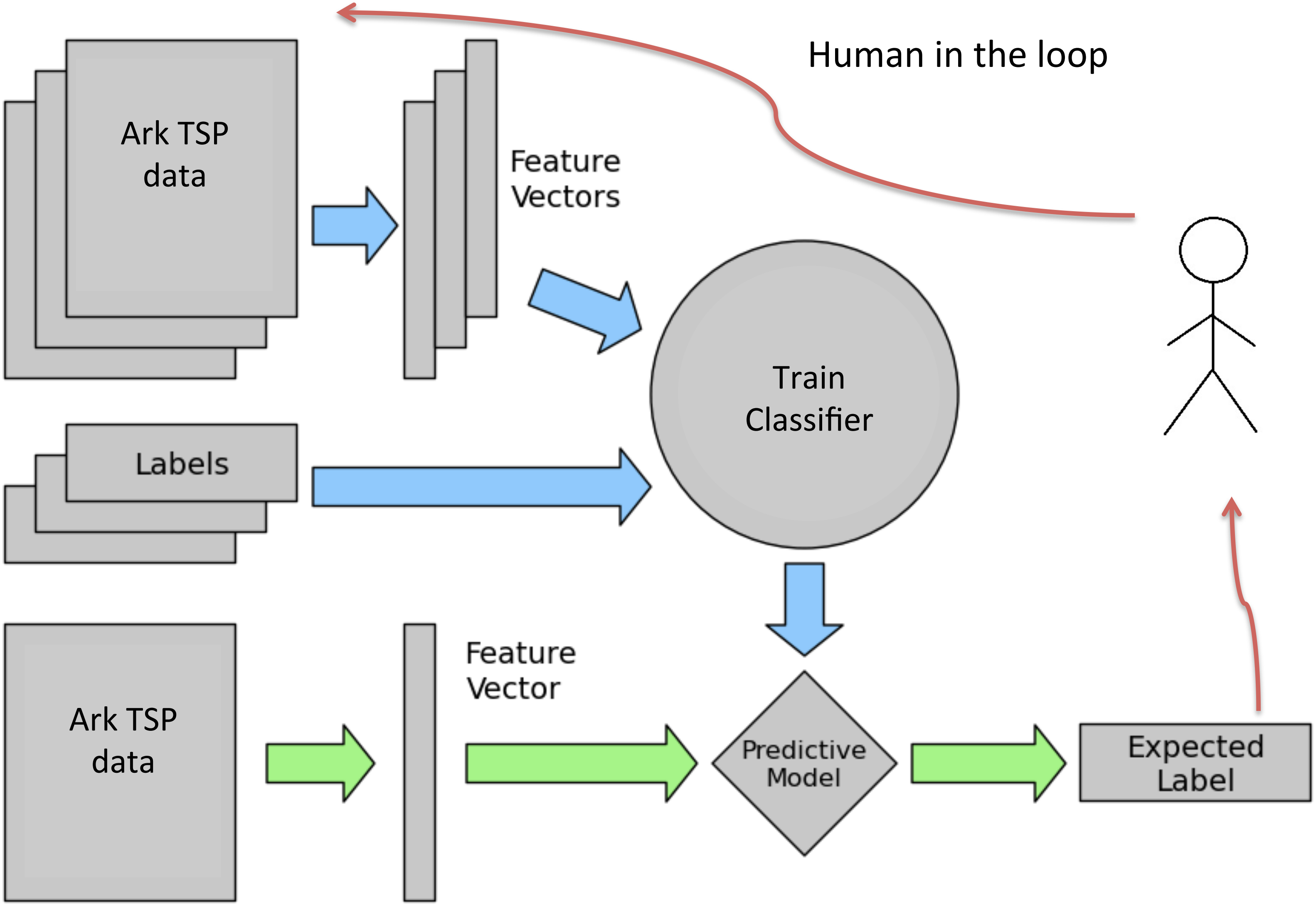
Classifying congestion in Ark measurements

(See earlier talks by
Among and Matthew
for background)

Classification system



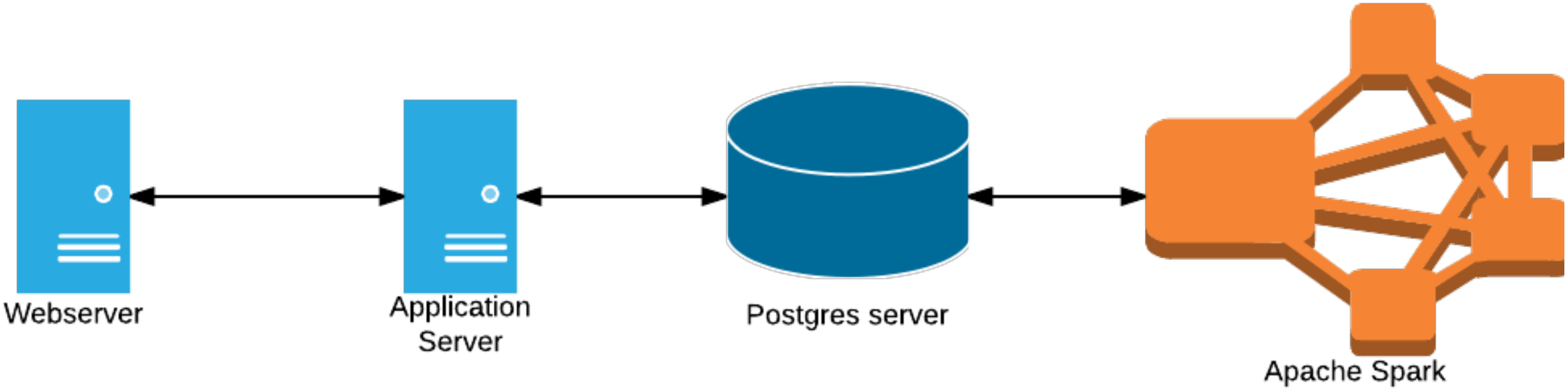
Classification system



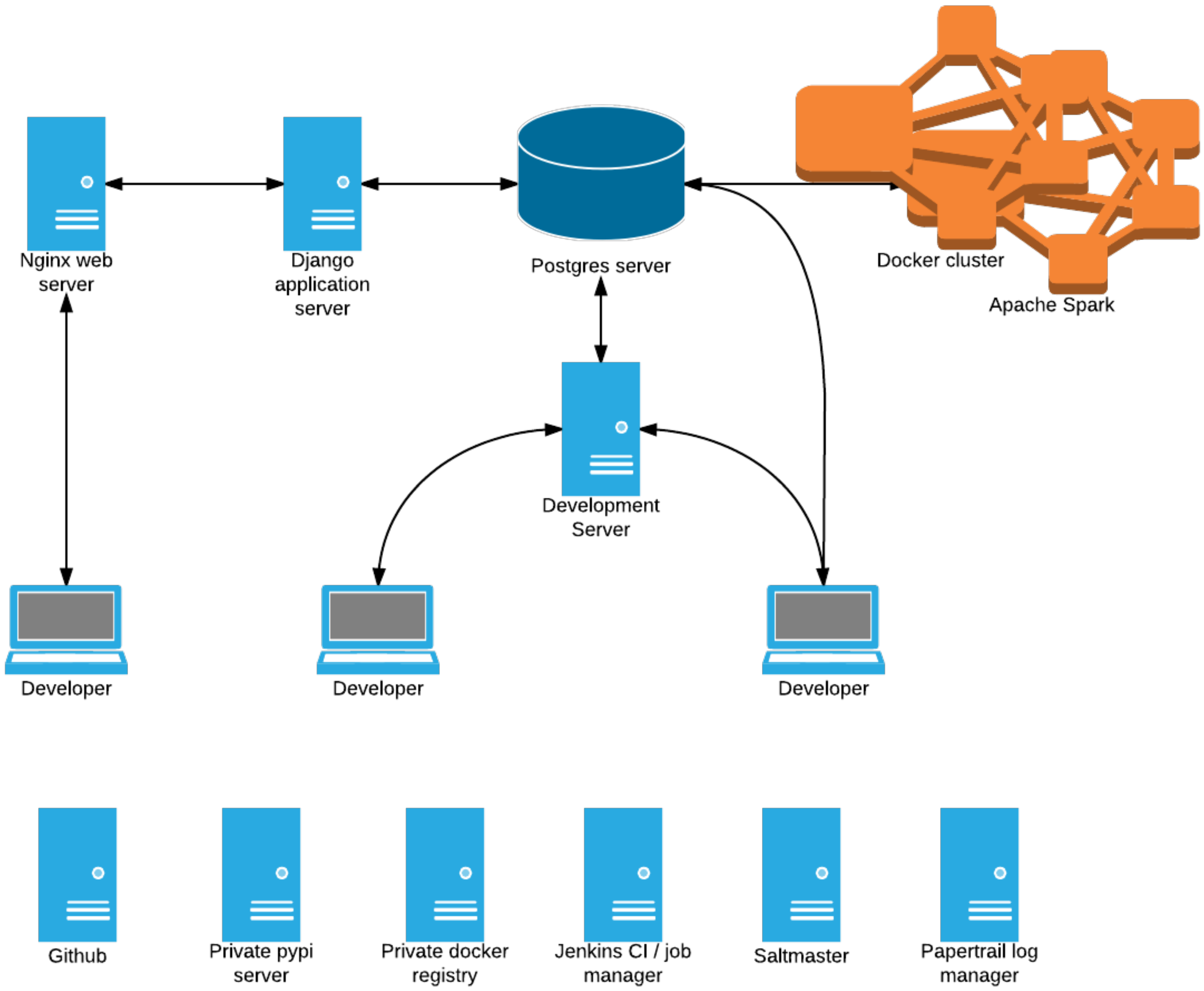
Objectives of human reviewers

1. Review links with “new” or “interesting” congestion changes
   - Congestion on formerly uncongested link
   - No congestion on formerly congested link
   - Change in the nature of congestion on a link
2. Generate annotation data that will be useful in improving congestion classifier
3. Development of good features

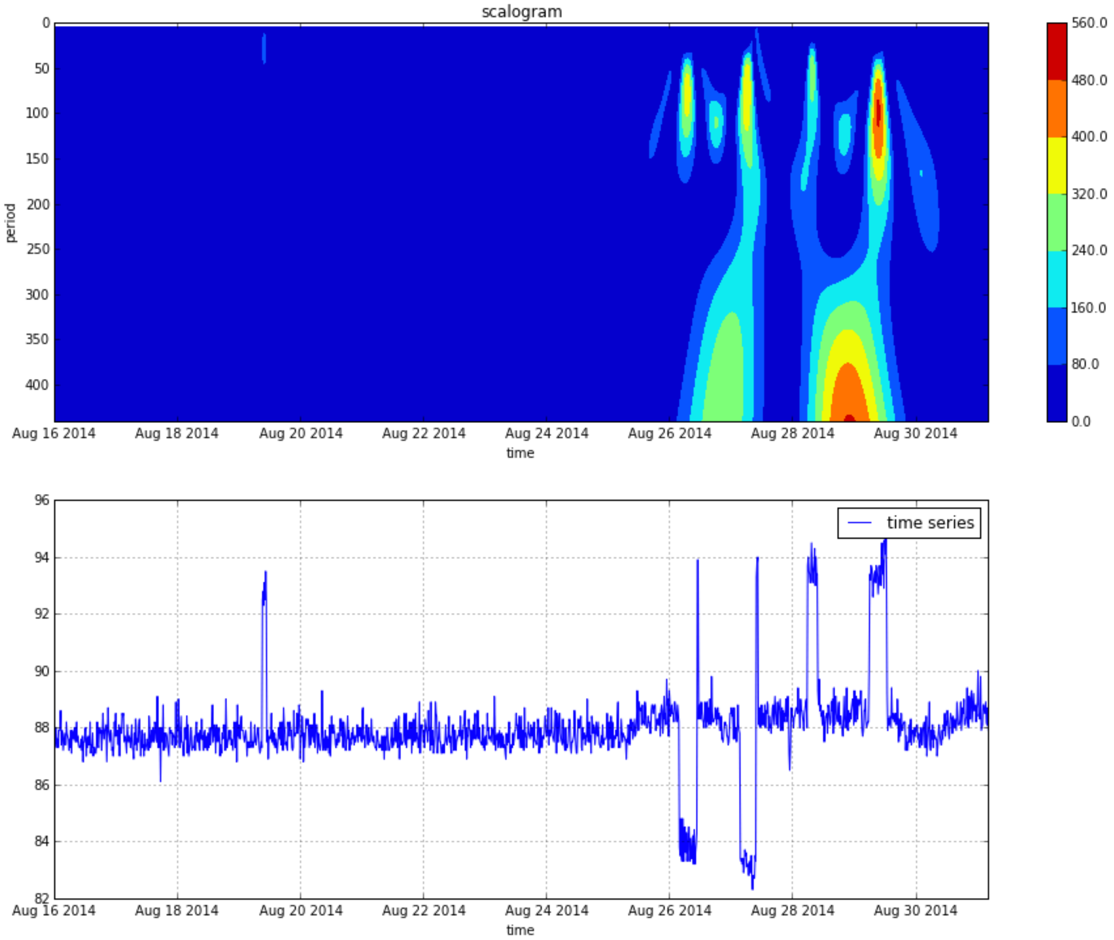
Simple system architecture



Detailed system architecture



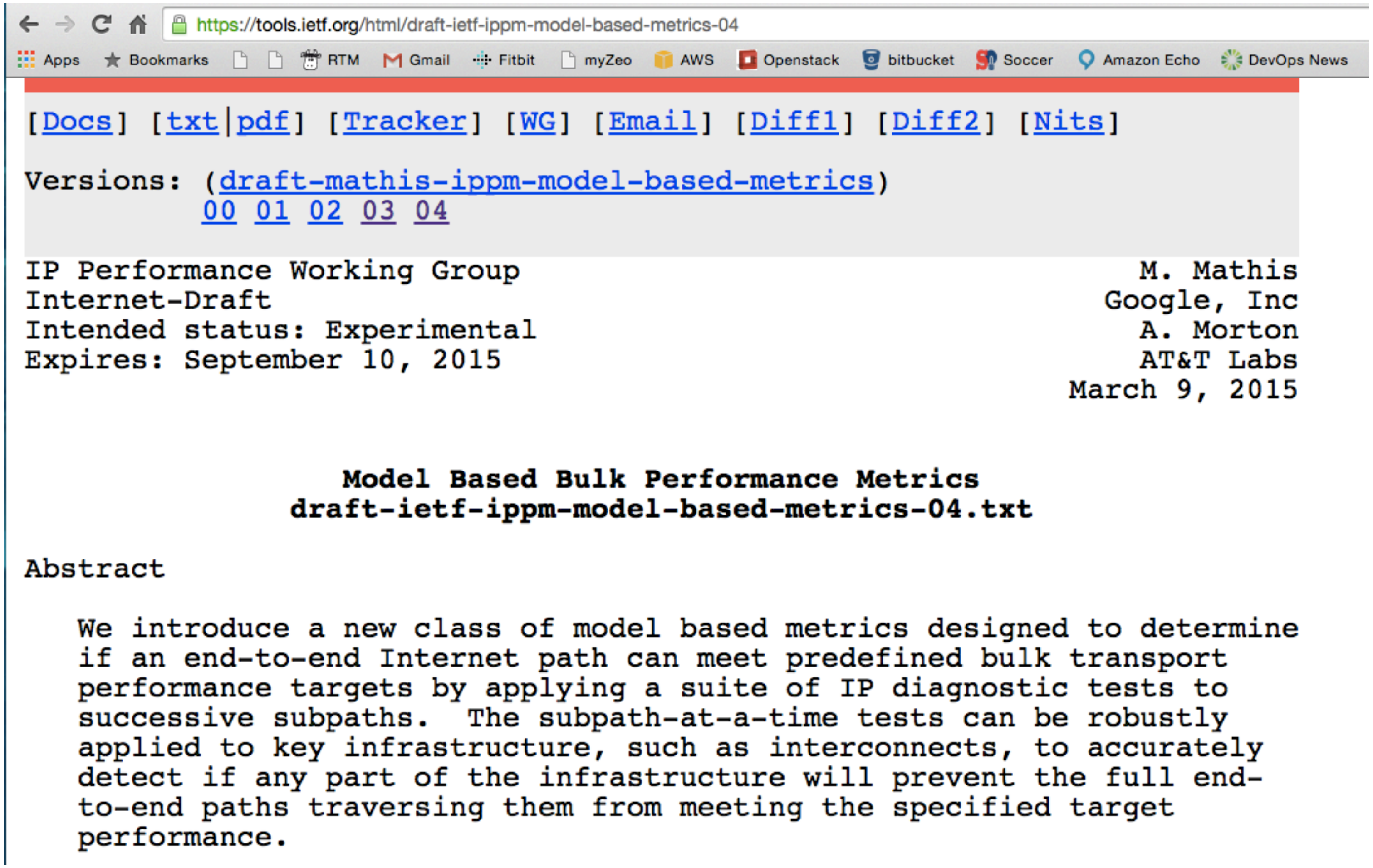
Wavelet features



Research agenda

- Improve features being extracted
- Classification algorithms
  - Random forest
  - K-NN
- Trigger additional tests:
  - Model Based Metrics tests on Ark nodes

Model Based Metrics (MBM)



[Docs] [txt|pdf] [Tracker] [WG] [Email] [Diff1] [Diff2] [Nits]

Versions: (draft-mathis-ippm-model-based-metrics)
00 01 02 03 04

```
IP Performance Working Group                                   M. Mathis
Internet-Draft                                               Google, Inc
Intended status: Experimental                                  A. Morton
Expires: September 10, 2015                                    AT&T Labs
                                                           March 9, 2015
```

Model Based Bulk Performance Metrics
draft-ietf-ippm-model-based-metrics-04.txt

Abstract

We introduce a new class of model based metrics designed to determine if an end-to-end Internet path can meet predefined bulk transport performance targets by applying a suite of IP diagnostic tests to successive subpaths. The subpath-at-a-time tests can be robustly applied to key infrastructure, such as interconnects, to accurately detect if any part of the infrastructure will prevent the full end-to-end paths traversing them from meeting the specified target performance.

Model Based Metrics

Suppress equilibrium behavior by open looping TCP

- IP test traffic mimics TCP independent of the network details
- IP success criteria is based on TCP models
- Eliminate circular interactions between RTT, packet loss and data rate

Slide details from:
http://www.ietf.org/proceedings/92/slides/slides-92-ippm-7.pdf