

# Repeatable research, measurement, and cybersecurity – opportunity and necessity –



Andrew W. Moore  
Computer Laboratory  
Dept of Computer Science and Technology

ACM SIGCOMM 2018 Workshop on Traffic Measurements for Cybersecurity (WTMC 2018)

<http://www.cl.cam.ac.uk/~awm22/slides/2018-sigcomm-wtmc-moore.pdf>

# Reproducibility in Science

- **Validate Correct Results**

supporting the conclusions  
and  
compare with new ideas

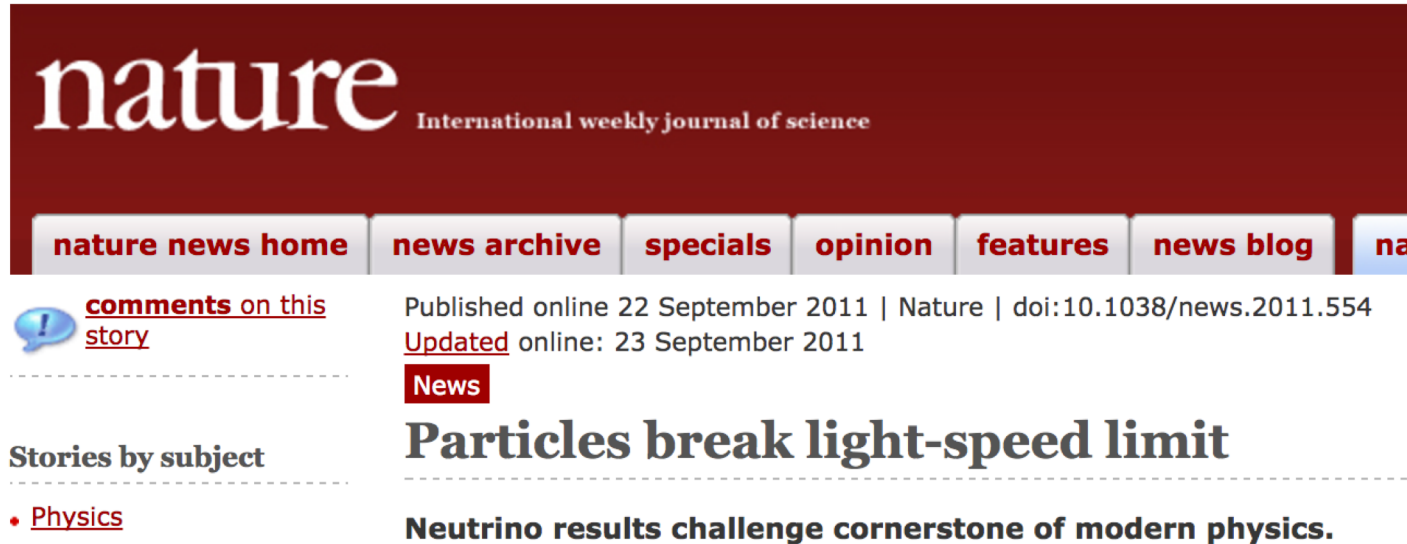
- **Invalidate Incorrect results**

refuting the conclusions  
and  
improve and refine

# Reproducibility as validation



# Reproducibility as invalidation




The screenshot shows the top of the Nature website. The header is dark red with the word "nature" in white serif font, followed by "International weekly journal of science" in a smaller white sans-serif font. Below the header is a navigation bar with white buttons containing red text: "nature news home", "news archive", "specials", "opinion", "features", "news blog", and "na".

On the left side, there is a section titled "Stories by subject" with a list of subjects, including "Physics". Above this, there is a link "comments on this story" with a blue speech bubble icon.

The main content area on the right features a red "News" tag, the headline "Particles break light-speed limit", and a sub-headline "Neutrino results challenge cornerstone of modern physics." Above the headline, it says "Published online 22 September 2011 | Nature | doi:10.1038/news.2011.554" and "Updated online: 23 September 2011".

**nature** International weekly journal of science

[nature news home](#) [news archive](#) [specials](#) [opinion](#) [features](#) [news blog](#) [na](#)

 [comments on this story](#)

Published online 22 September 2011 | Nature | doi:10.1038/news.2011.554  
[Updated](#) online: 23 September 2011

**News**

## Particles break light-speed limit

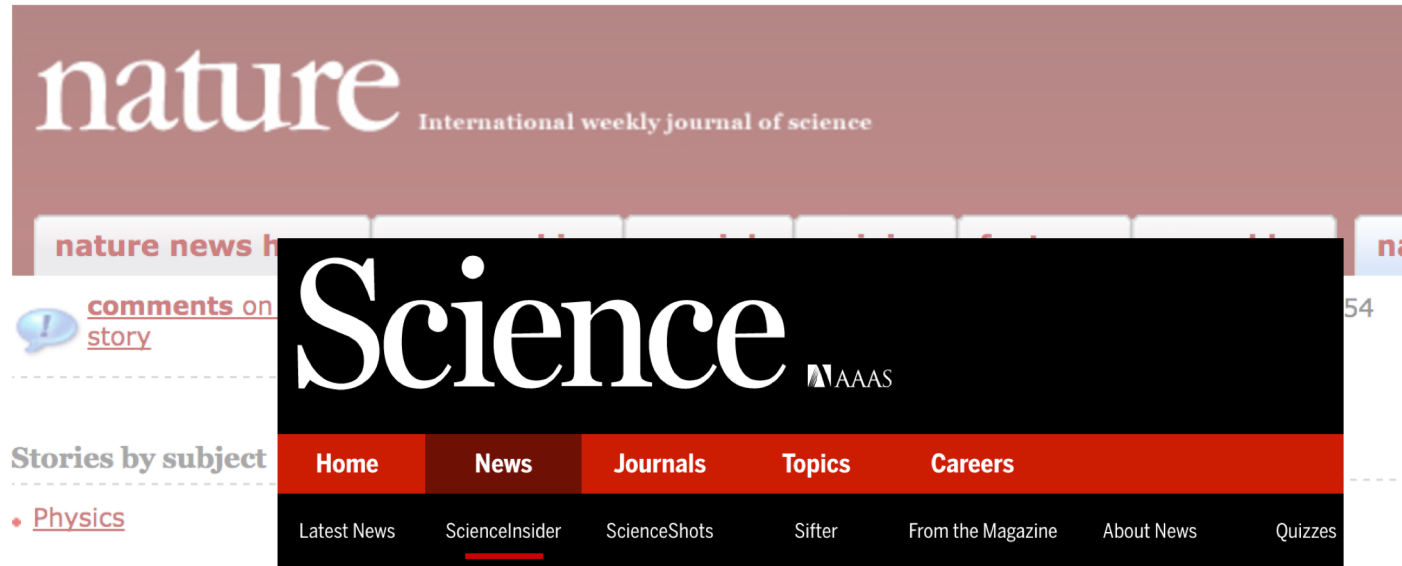
Neutrino results challenge cornerstone of modern physics.

Stories by subject

- [Physics](#)



# Reproducibility as invalidation



**SHARE**

## Once Again, Physicists Debunk Faster-Than-Light Neutrinos

By [Adrian Cho](#) | Jun. 8, 2012, 3:39 PM

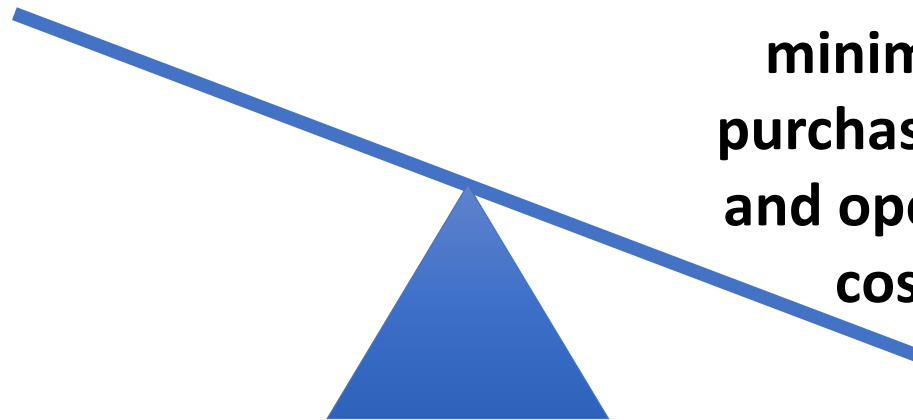
Researchers: Novel &  
Disruptive ideas

# Flexibility: the network architect (& operator) dilemma...

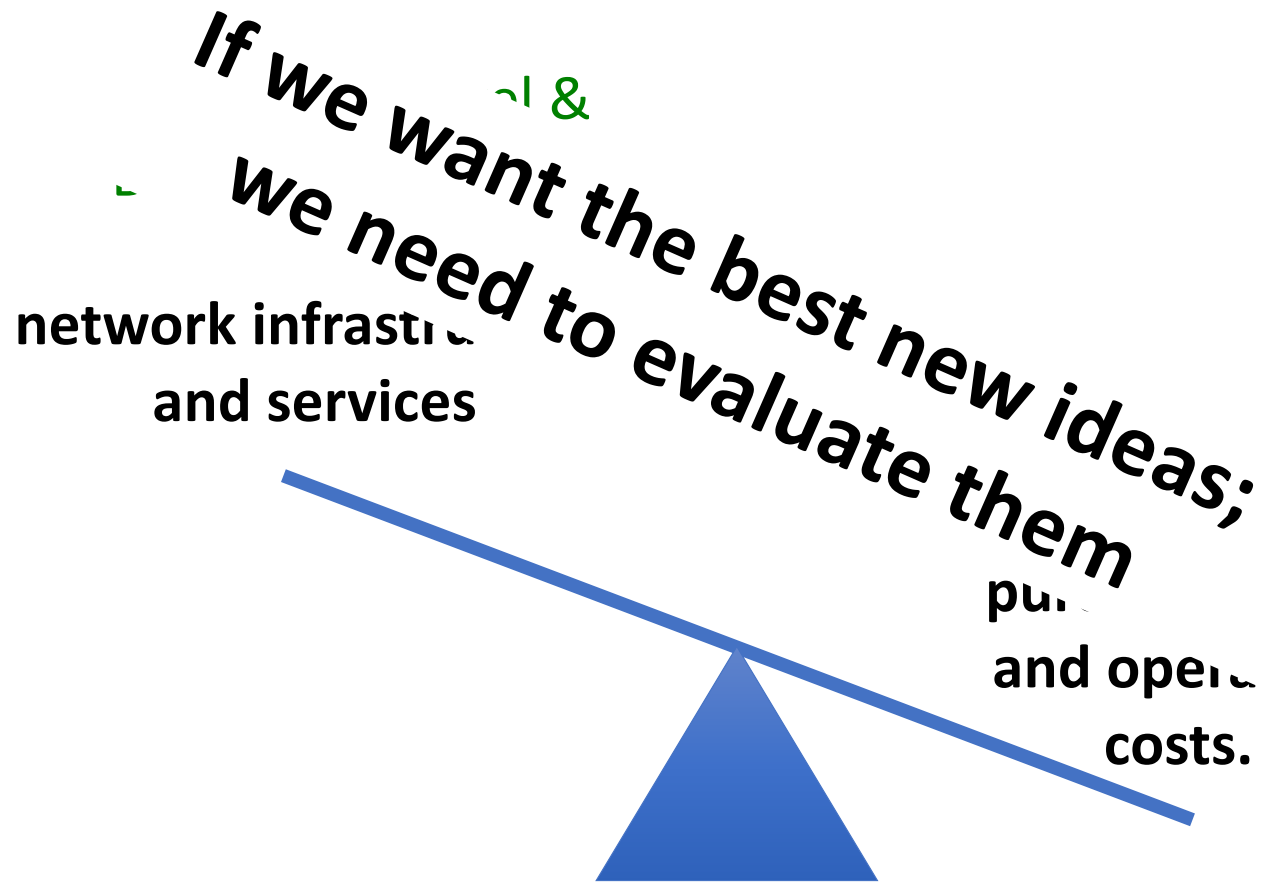
Researchers: Novel &  
Disruptive ideas

**network infrastructure  
and services**

**minimizing  
purchase costs  
and operation  
costs.**



Flexibility: the network architect  
(& operator) dilemma...



What do Linux apache MySQL Firefox BSD BIND and Bro



have in common with the resurgence in  
Software Defined Networking (SDN)?

What do Linux apache MySQL Firefox BSD BIND and Bro



have in common with the resurgence in SDN?

Openly available source and open standards



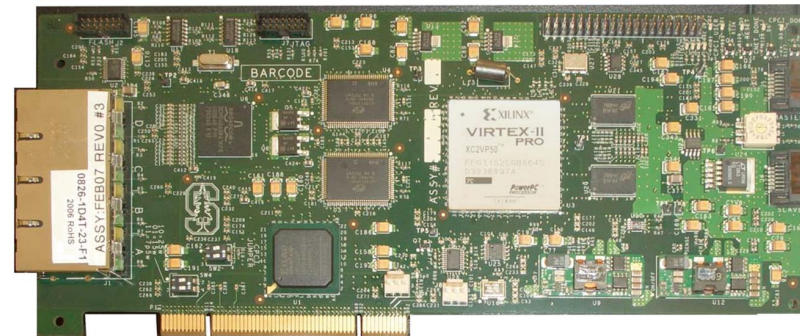
# An approach (perhaps not the solution)

- SDN, not a new idea but one that has definitely had rockets in recent years.



# An approach (perhaps not the solution)

- SDN, not a new idea but one that has definitely had rockets in recent years.





# And not just an implementation...

- OFLOPS – the OpenFlow performance tester
- OFtest – OpenFlow compliance tester
- OvS – software only implementation
- Numerous OpenFlow controllers:



From Ryu to OpenDaylight





Each a stable platform

- enabling extension, and
- a process for adopting contributions and improvements



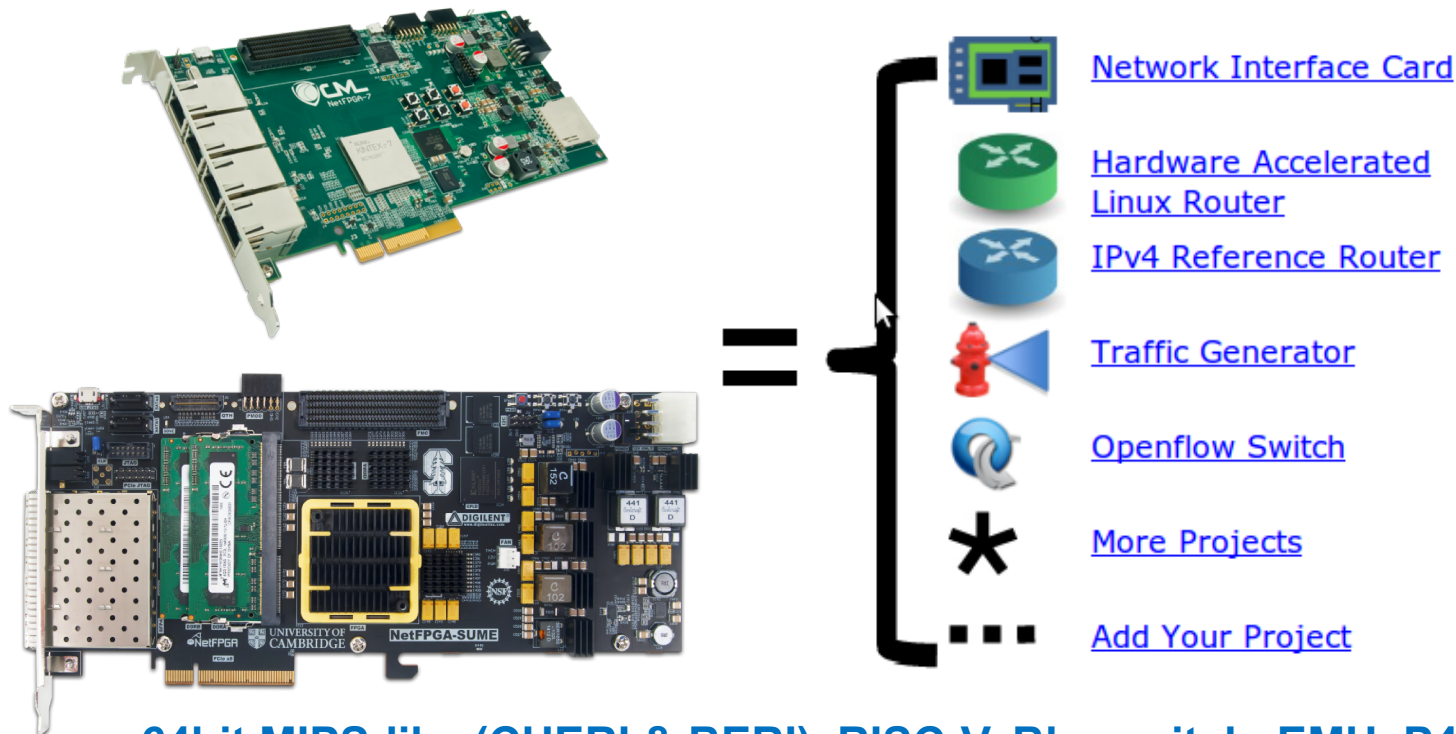
Because **network research**, and **education**  
needs a good **platform**

**[www.netfpga.org](http://www.netfpga.org)**

# So what is NetFPGA?

## NetFPGA = Networked FPGA

A line-rate, flexible, open networking platform for teaching and research



**64bit MIPS-like (CHERI & BERI), RISC-V, Blueswitch, EMU, P4 FPGA...**

# **Some thoughts on Application Identification and Classification**

Ten-year old

## Some thoughts on Application Identification and Classification



"**Those** who cannot remember the past are **condemned to repeat** it."

But this time with different buzzwords?

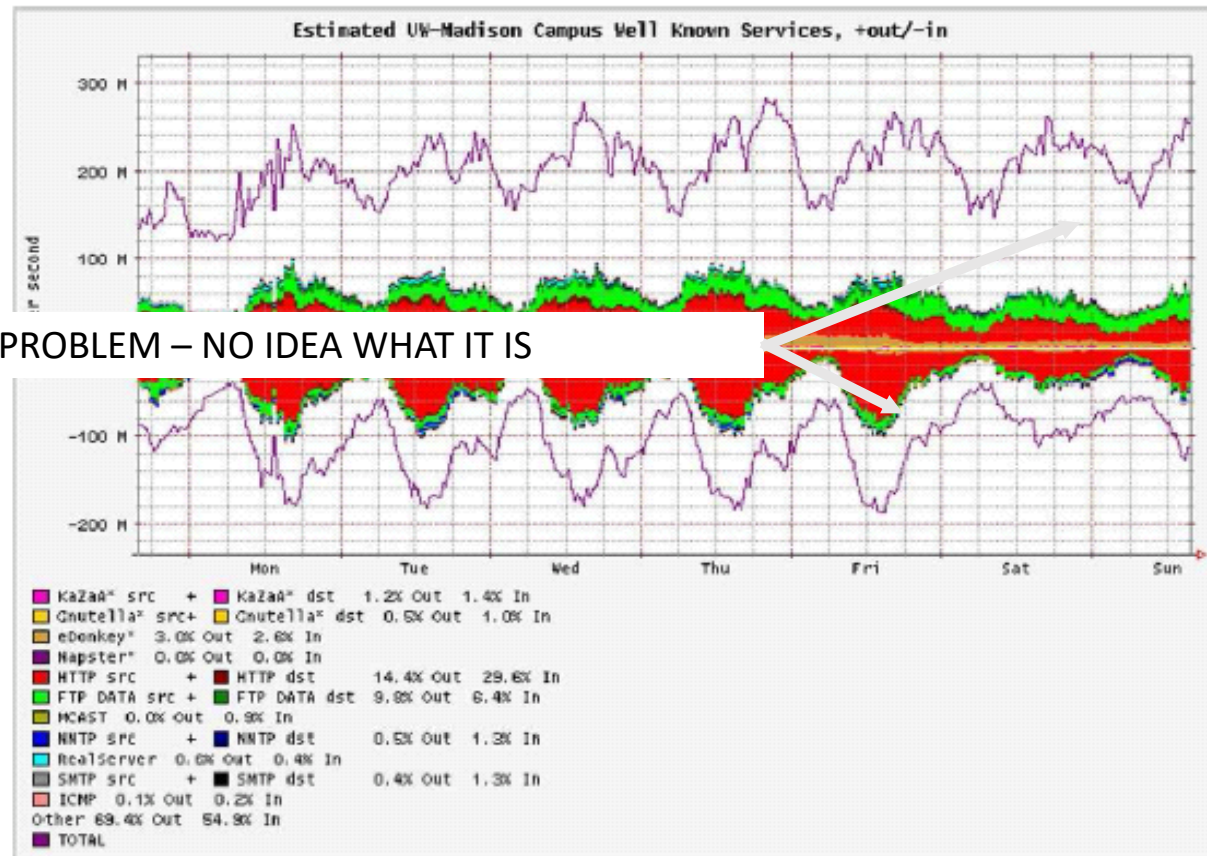
# Why Characterize?

- **Identify:** *“Hmmm, So this is what an attack looks like”*
- **Understanding:** *“So what is my network doing anyway?”*
- **Accountability:** *“What has caused this enormous bill?”*
- **Application Enabler:** Dynamic (application-specific) handling (e.g. routing) by end systems
- **Performance Tracking:** *“What is causing my application to go so very slow?”*
- **Application identification:** *“...telling helpdesk what the users won’t or can’t find out”*
- **Better Models:** Leading to better/more-realistic test traffic

# Understanding

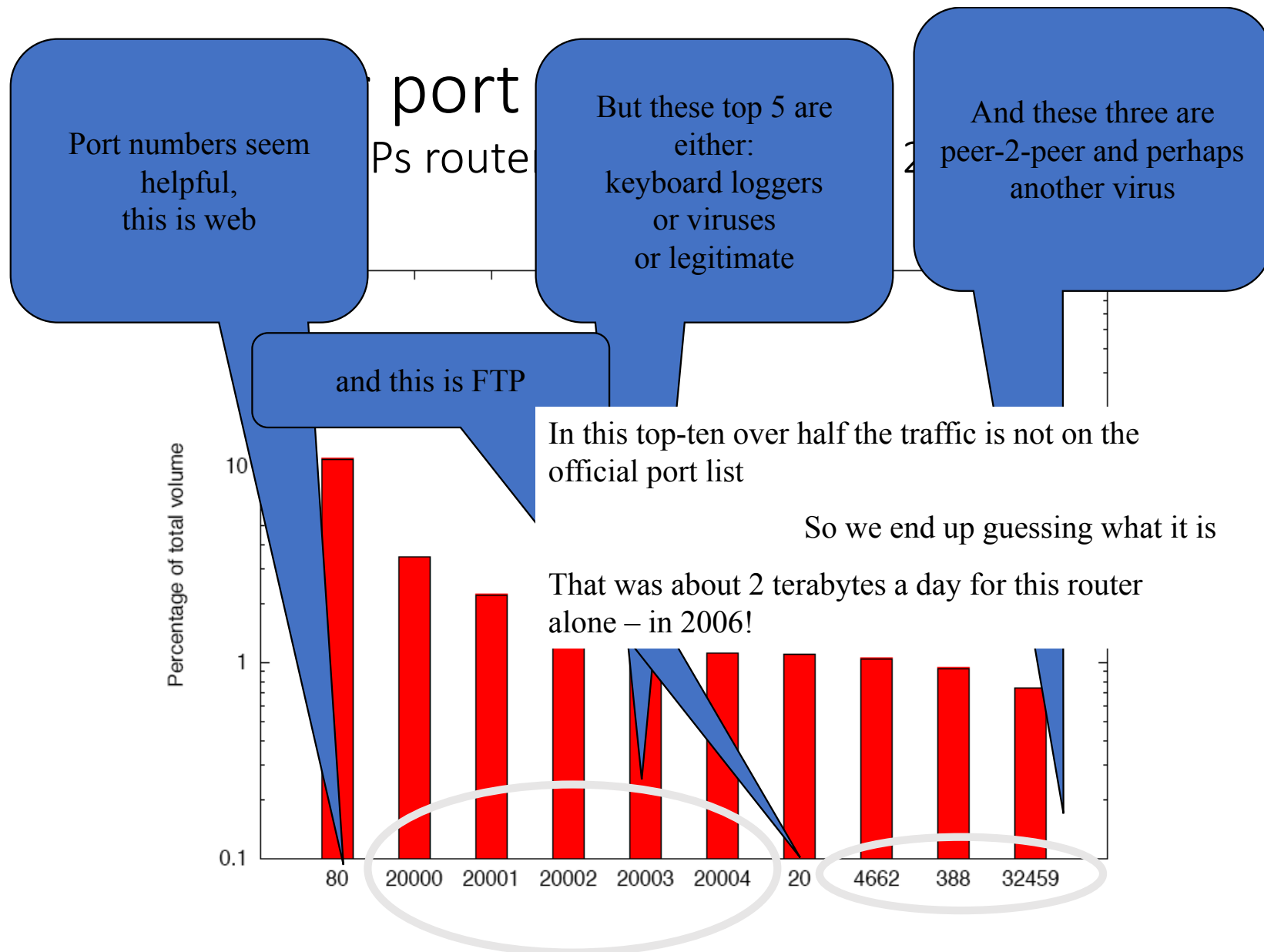
traffic for a large university - not Cambridge

THIS IS THE PROBLEM – NO IDEA WHAT IT IS



Traffic Distribution of the network of the University of Wisconsin for the week 7-13 Sept. 2003. Courtesy of wwstats.net.wisc.edu





# Why is this a problem?

For one particular traffic sample...

- Using a port-based method we could not identify 30% of the traffic **at all**

*Why?* Many ports are not “designated”, have unofficial uses or an ambiguous designation

32343: Err no-idea

4662: that would be eMule, but it isn't in any “official” list

- Of the 70% we could identify with port-based schemes a further 29% was **incorrectly** identified

*Why?* Official port lists don't tell the whole tale

“If I wrap my new application up to look like HTTP it will get through the firewall”

80: HTTP is that a server or a proxy or a VPN or a ...?



## Ports as poor practice

- Ports are still used as some sort of definitive classifier
- Commonly by studies examining the effectiveness of new methods  
(using traffic without “ground-truth”)
- BUT  
ground-truth error >> evaluation accuracy

# What is an application anyway?

- port 80?
- http on port 80?
- html on http on port 80?
- web page on html on http on port 80?
- So what about gmail?
  - email or web (browser) traffic?
  - What about when my MUA gets the email via the webmail interface?

# Email

- MTA vs MUA
- Spam vs Ham
- Commercial vs Domestic
- Decent vs Wicked



A modern equivalent...



*Bad guys or bad content is coming from some IP addresses*

*Lets block them!*

*But those are the IP addresses of*

- *Amazon AWS*
- *Akamai CDN*
- *...*

# Domain Knowledge

- Each of the motivations for “Why?” is a different domain of knowledge:
  - Hard to compare methods applied to different domains  
(Helping helpdesk may require significant site knowledge & historical knowledge)
  - Hard to compare data used in/by/for different methods
- ML “headline”: These approaches encode domain knowledge

Anyone remember expert-systems?

# *Elephants in the*

*Hallway/Driveway/Kitchen/Lounge(room)/Bathroom/Bedroom*

- Limited engagement of/with the M-L community
  - Mea Cupla - I don't read KDD output either  
(in fact SIGKDD is happening right now – in London)
- Difficult-to-compare methodologies
- Difficult-to-compare datasets
- Lack of (annotated) Data
  - We don't/**can't** play nicely together
  - Privacy/Law





# Classes as confusion

Network traffic Paper 1	Network traffic Paper 2	Network traffic Paper 3	Typical IDS paper
7 meta-classes (? classes)	11 meta-classes (40-50 classes)	11 meta-classes (40-50 classes)	2/3 meta-classes
domain, ftp-data, https, kazaa, realmedia, telnet, www	bulk(ftp), database, interactive, mail, services, www, p2p, attack, games, multimedia, unknown	web, p2p, data(ftp), network management, mail, news, chat/irc, streaming, gaming, nonpayload, unknown	Good, Bad, Ugly



How can I compare these methods?  
I certainly can't compare the output

Upshot - one persons great performance  
is another persons rubbish performance

**DON'T  
PANIC**

# What can we do?

- Raise the bar on acceptable research
  - Insist on the artefacts being published
  - Insist on the results being available
  - Insist on the experiments being repeatable
  - Accept reproduction studies (well done CCR!)
- Enable Research Repeatability

Lead by example

# Lead by example

Moore & Zuev 2005, we published the dataset and code to create this dataset

<https://www.cl.cam.ac.uk/research/srg/netos/projects/archive/nprobe/data/papers/sigmetrics/index.html>

In a form that was preserved anonymity and

1. Enabled reproducibility,
2. Enabled comparison against new algorithms, and
3. Actively encouraged others to create their own datasets

# Lead by example

## **Queues don't matter when you can JUMP them!**

Matthew P. Grosvenor    Malte Schwarzkopf    Ionel Gog    Robert N. M. Watson  
Andrew W. Moore    Steven Hand    Jon Crowcroft

*University of Cambridge Computer Laboratory*

*Simplicity is the shortest path to a solution.*  
– Ward Cunningham

### **Abstract**

QJUMP is a simple and immediately deployable approach to controlling network interference in datacenter networks. Network interference occurs when congestion from throughput-intensive applications causes queueing that delays traffic from latency-sensitive applications. To mitigate network interference, QJUMP applies Internet

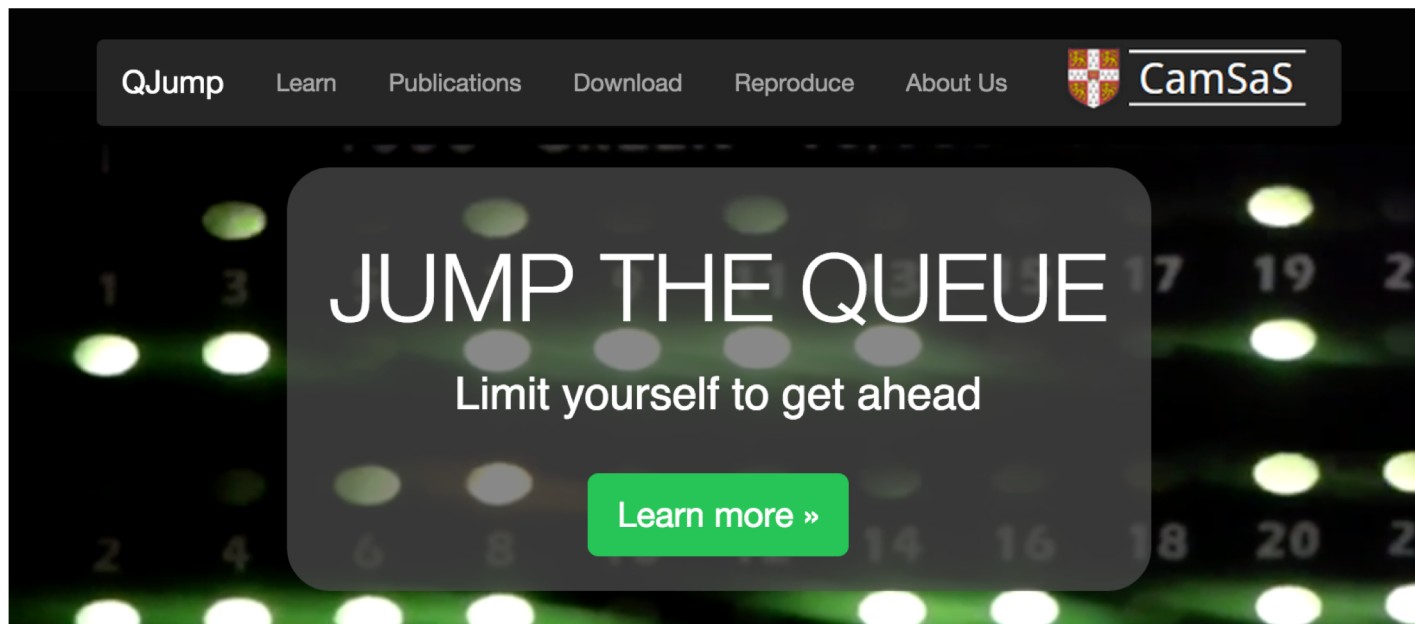
If memcached packets can somehow be prioritized to “jump-the-queue” over Hadoop’s packets, memcached will no longer experience latency tails due to Hadoop. Of course, multiple instances of memcached may still interfere with *each other*, causing long queues or incast collapse [10]. However, if each memcached instance can be appropriately rate-limited at the origin, this too can be mitigated.

These observations are not new: QoS technologies like DiffServ [7] demonstrated that coarse-grained classifica-

# Lead by example

**Queues don't matter when you can JUMP them!**

<http://www.camsas.org/qjump>



# Lead by example

**Queues don't matter when you can JUMP them!**

<http://www.camsas.org/qjump>

QJump

Learn

Publications

Download

Reproduce

About Us



CamSaS

## Reproducing the QJump Experiments

As scientists and researchers we take the reproducibility of our work very seriously. We don't expect you to trust our results, in fact, we hope that you don't! On the following page we provide links to detailed descriptions the experiments behind each of the figures in our research publications. The full experimental descriptions include the precise configuration of our test equipment, links to the source code for our tools, patches that we made to other people's tools and original preprocessed data-sets that we gathered. Our hope is that anyone can use these descriptions re-run any of our experiments. We believe that this kind of openness is the way that all good, scholarly scientific research should be conducted.








For details of our publications including links to the manuscripts please see the [publications](#) page.

## NSDI 2015 - Queues don't matter when you can














# Enable others

Reproducible research needs,  
widely available test-equipment

	Cost	Flexibility	Resolution	Line Rate
	\$\$\$ \$\$\$			
DPDK, SW tools	(\$)			

# Enable others

Reproducible research needs,  
widely available test-equipment

	Cost	Flexibility	Resolution	Line Rate
	\$\$\$ \$\$\$			
DPDK, SW tools	(\$)			
	(\$)			

# Open-Source Network Tester



A platform for testing powered by



- Open source hardware and software platform for network test, publicly available

<https://osnt.org/>

<https://github.com/NetFPGA/OSNT-Public/wiki>

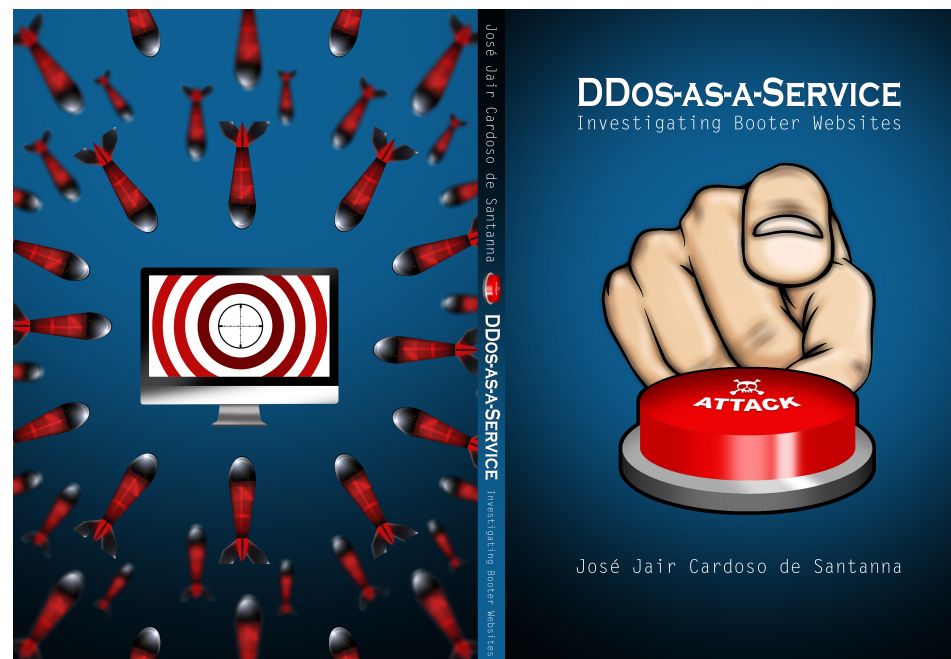
- Low cost, low jitter, flexible to update, scale-out,  
no CPU usage, nano-second resolution measurement

# What can we do?

- Raise the bar on acceptable research
  - Insist on the artefacts being published
  - Insist on the results being available
  - Insist on the experiments being repeatable
  - Accept reproduction studies (well done CCR!)
- Enable Research Repeatability  
and then *Just Do It!* Every time
- Open-source (research) platforms **work!**  
Build on platforms, and support platforms

# What can we do?

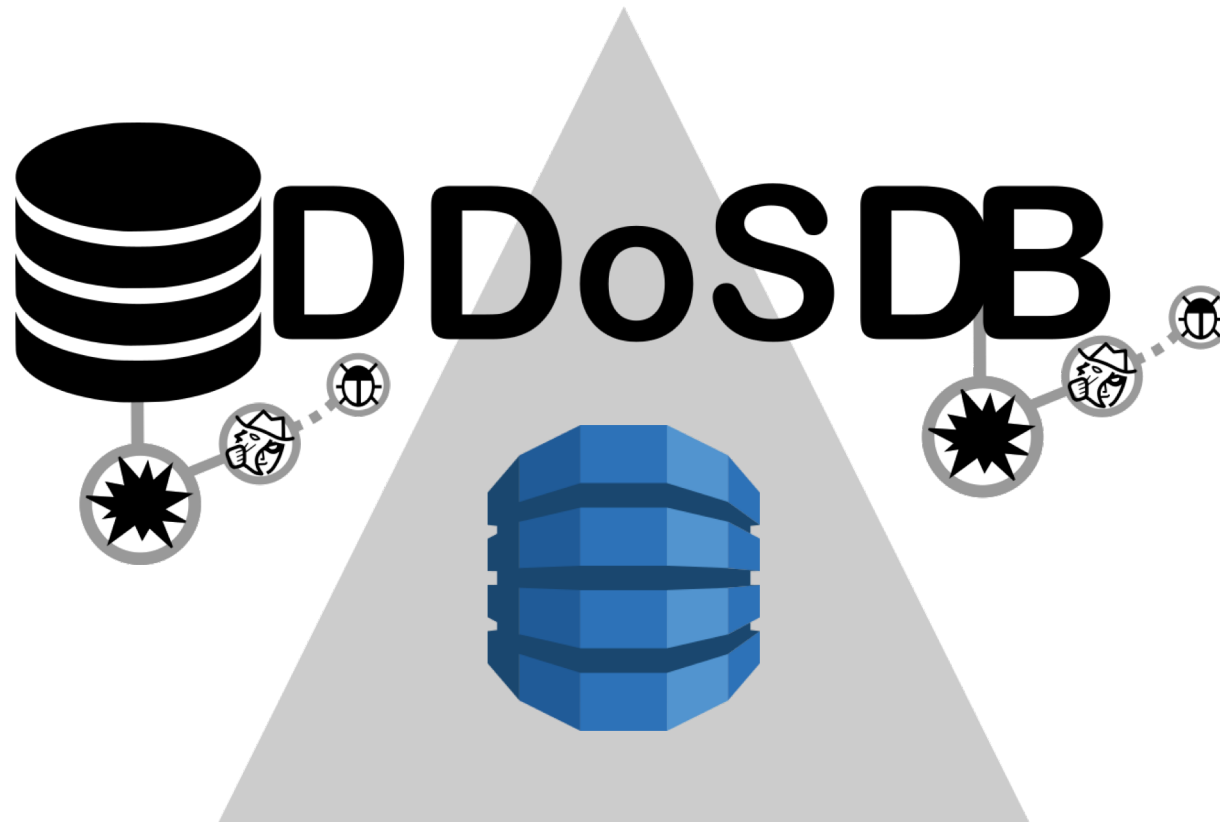
“Ok, that’s nice... still waiting for the cybersecurity connection?”



***Jair Santanna***

[jairsantanna.com](http://jairsantanna.com)

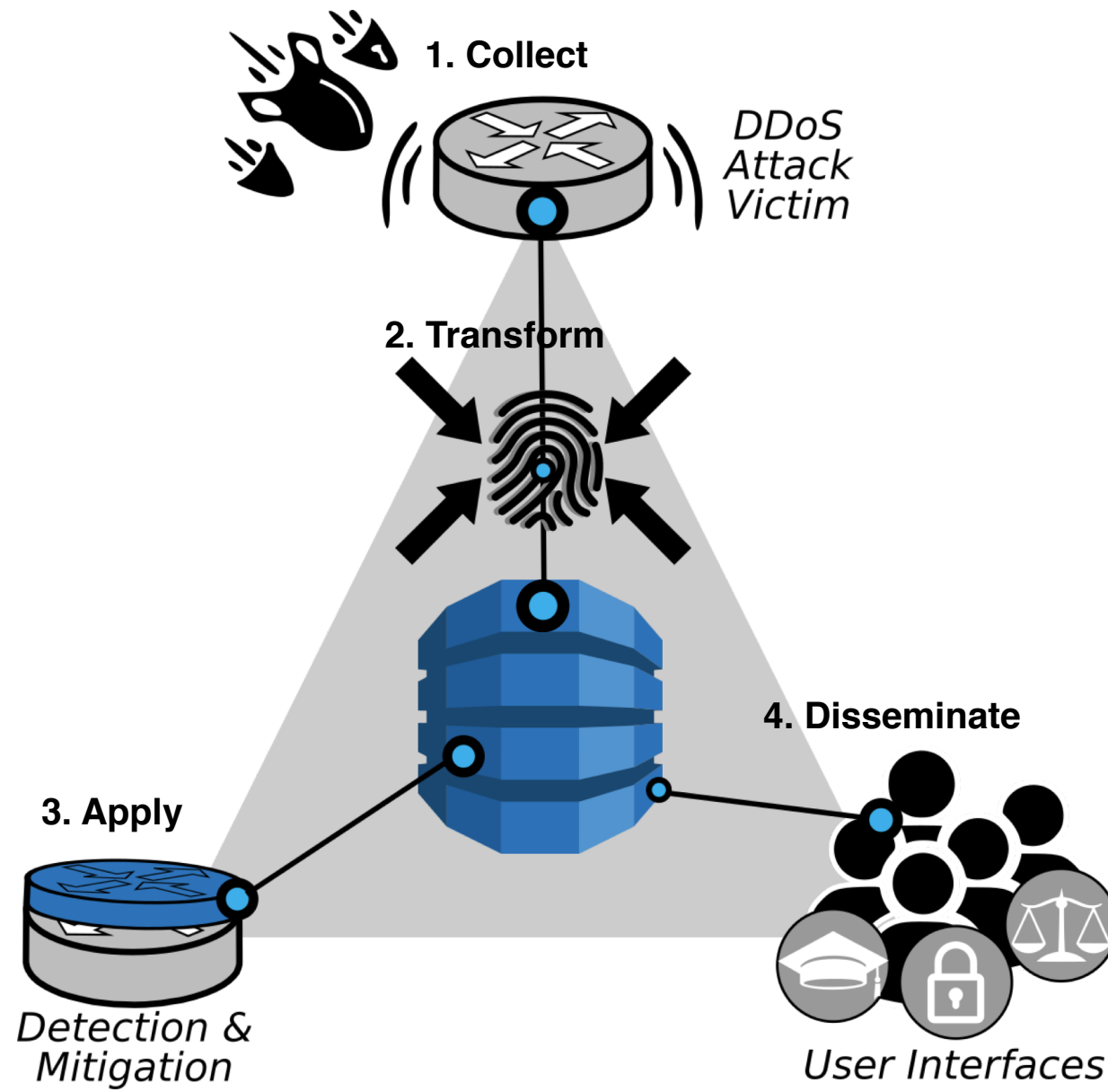
[j.j.santanna@utwente.nl](mailto:j.j.santanna@utwente.nl)



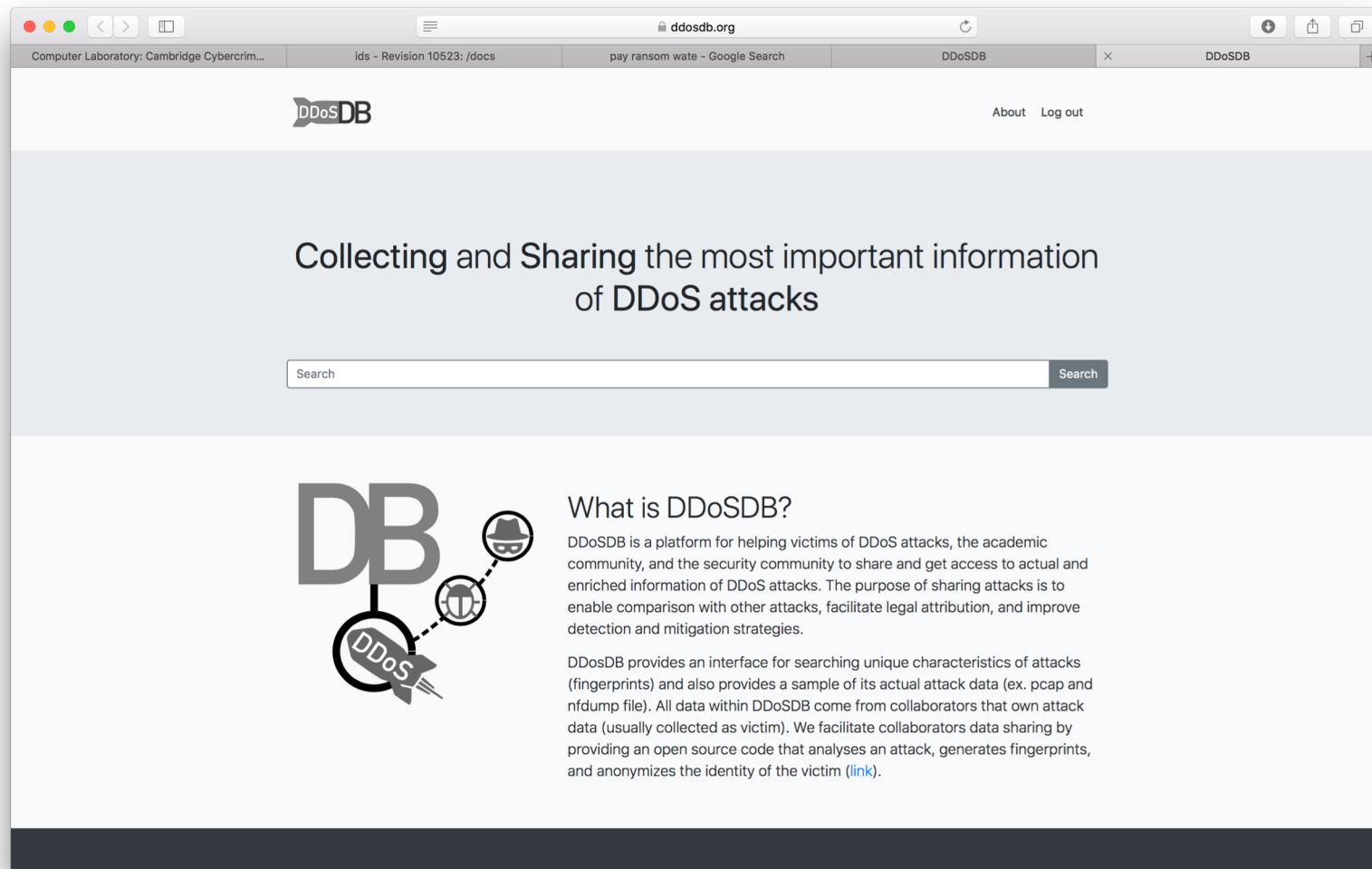
***Jair Santanna***

[jairsantanna.com](http://jairsantanna.com)

[j.j.santanna@utwente.nl](mailto:j.j.santanna@utwente.nl)







[About](#) [Log out](#)

## Collecting and Sharing the most important information of DDoS attacks

Search



### What is DDoSDB?

DDoSDB is a platform for helping victims of DDoS attacks, the academic community, and the security community to share and get access to actual and enriched information of DDoS attacks. The purpose of sharing attacks is to enable comparison with other attacks, facilitate legal attribution, and improve detection and mitigation strategies.

DDoSDB provides an interface for searching unique characteristics of attacks (fingerprints) and also provides a sample of its actual attack data (ex. pcap and nfdump file). All data within DDoSDB come from collaborators that own attack data (usually collected as victim). We facilitate collaborators data sharing by providing an open source code that analyses an attack, generates fingerprints, and anonymizes the identity of the victim ([link](#)).

# Cambridge Cybercrime Centre

(another unpaid advertisement)

- Rich expertise: University of Cambridge's [Department of Computer Science and Technology](#), [Institute of Criminology](#) and [Faculty of Law](#).
- Data-driven approach: the need for real-data hampers everyone, so..... Others can use our data too
- Importantly: This is not a competition –

*“... I want to be judged not on how many papers we wrote in Cambridge, but ...(the work we enabled)... new ways to prevent crime, to detect and deter criminals ... that’s why society funds our work...”*

Richard Clayton (Director)

**<https://www.cambridgecybercrime.uk/process.html>**

## Computer Laboratory

### Cambridge Cybercrime Centre: Process for working with our data

This page sets out the steps in the process for obtaining data from the Cybercrime Centre.

#### **Assess whether you will be allowed to use our data**

Our datasets are intended for research and analysis into methods to find, understand, investigate and counter cybercrime so your project must clearly fall into this space. Although we do not require researchers to be academics, there are significant restrictions on using our data for commercial purposes.

Although some of our data was generated internally and so we can make it available for other types of project and for commercial purposes, much of our data has come from third parties and they have only provided us with the data because of the framework under which it will be shared.

#### **Identify the data you wish to use**

We describe our various datasets on this page [ [LINK](#) ]. The descriptions are public and necessarily fairly high level. We do however try to indicate the size of the datasets, the period over which they was collected, along with any known biases.

We strongly encourage the use of prepacked datasets rather than "live feeds". Although a live feed may be superficially attractive it makes it harder to arrange that other researchers can receive the same data that you did -- a key aim of the Cybercrime Centre is to enable reproducible research. If the issue is that you need to collect a further "field" over and above what we supply then talk with us and we may well be able to do this for you.

#### **Read about our legal framework**

It is important that you understand the basis on which we share data and the paperwork that will need to be signed.

There's several pages of explanations and FAQs about our agreements, starting here at <https://www.cambridgecybercrime.uk/data.html>, which you should read before contacting us.

#### **Make an application**

You will need to make a formal application to use our data. In the first instance you should send an email to the Director of the Cybercrime Centre,

# Datasets at Cambridge Cybercrime Centre

- Underground Forums (>> 40m posts)
- Blog spam (>300K posts)
- Reflected DDoS victims (4+ years data)
- Mirai scanning data (of Cambridge and elsewhere)
- Mirai (etc) malware (since Dec 2016, 20K samples!)
- SSH honeypot datasets (> 2 years)
- Email spam (back to 2004, and some from the 1990s!)
- 419 scam emails (> 60K, dating back to 2006)
- Phishing emails (50K plus, over 10 years)
- Phishing URLs and pages

New field – new methods, new research, new science.....

**Exploring the provision of online booter services**

Alice Hutchings<sup>1</sup> and Richard Clayton<sup>2</sup>

1000 days of UDP amplification DDoS attacks

Daniel R. Thomas

Richard Clayton

Alastair R. Beresford

**Ethical issues in research using datasets of illicit origin**

Daniel R. Thomas

Sergio Pastrana

Alice Hutchings

Richard Clayton

Alastair R. Beresford

Cambridge Cybercrime Centre, Computer Laboratory

University of Cambridge

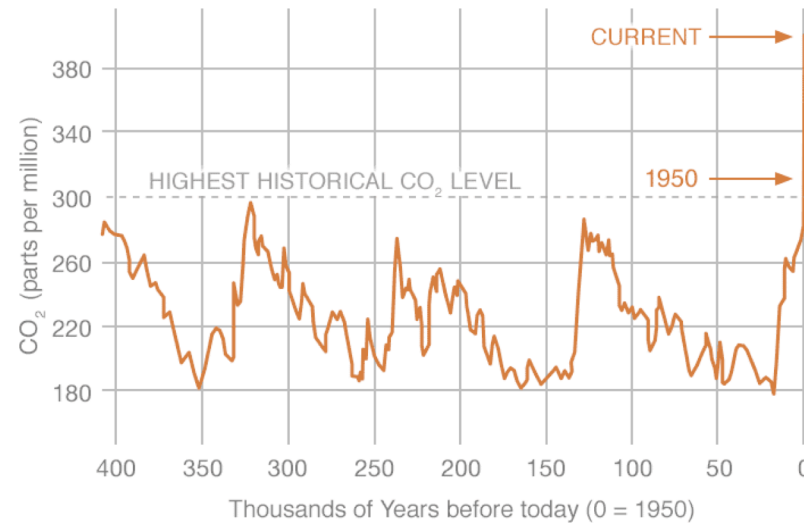
United Kingdom

Firstname.Lastname@cl.cam.ac.uk

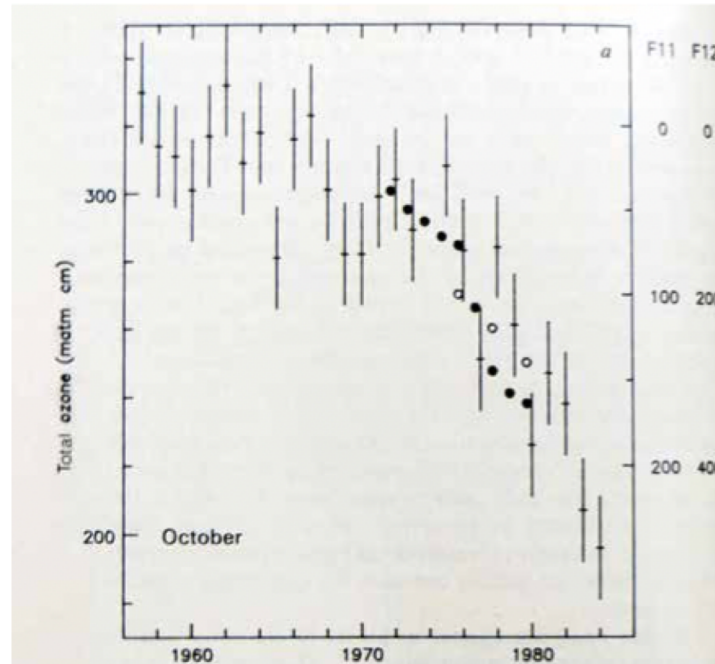
# What can we do?

- Raise the bar on acceptable research
- Enable Research Repeatability
- Open-source (research) platforms
- Share our research, our tools, our interpretation, our insights

# Good data outlives bad theory.....



# Good data outlives bad theory.....



\*JC Farman, BG Gardiner, JD Shanklin 1985 'Large losses of total ozone in Antarctica reveal seasonal  $\text{ClO}_x/\text{NO}_x$  interaction' *Nature* **315** 207-10.

## Before global warming there was ozone depletion



# What can we do?

- Raise the bar on acceptable research
- Enable Research Repeatability
- Open-source (research) platforms
- Share our research, our tools, our interpretation, our insights

***Sharing is hard*** (ask any 5 year old)  
but still worth doing!

# Acknowledgements



UNIVERSITY OF  
CAMBRIDGE

EPSRC

Pioneering research  
and skills



XILINX

ALL PROGRAMMABLE™



NEC



Microsoft



Google

Atomic  
Rules

H<sup>G</sup>GLOBAL  
ITECH



The Leverhulme Trust



HUAWEI

ALGO-LOGIC



imc