

# SCC.400 Research Methods

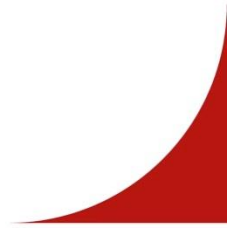
## SCC.460/SCC.460-HDS Data Science Fundamentals

Module Introduction



# Module Overview

---

- Teaching teams
  - Motivation
  - Aims
  - Structure
  - Assessment
  - Expectations
- 

# Teaching Teams SCC.400 Research Methods

---



Professor Hans Gellersen  
Convenor



Dr Elisa Rubegni

# Teaching Teams SCC.460/SCC.460-HDS Data Science Fundamentals

---



Professor Cheverst Keith  
Convenor



Dr Richard Jiang



Dr Ignatius Ezeani



*"They're harmless when they're alone, but get a bunch of them together with a research grant and watch out."*

# CORE PRINCIPLES IN RESEARCH



## OCCAM'S RAZOR

"WHEN FACED WITH TWO POSSIBLE EXPLANATIONS, THE SIMPLER OF THE TWO IS THE ONE MOST LIKELY TO BE TRUE."



## OCCAM'S PROFESSOR

"WHEN FACED WITH TWO POSSIBLE WAYS OF DOING SOMETHING, THE MORE COMPLICATED ONE IS THE ONE YOUR PROFESSOR WILL MOST LIKELY ASK YOU TO DO."

# Module Motivation (SCC.400/460/460-HDS)

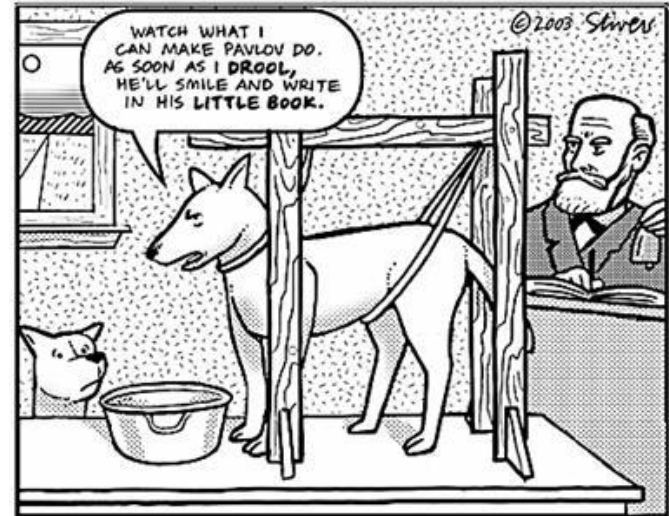
- Regardless of your career choice, **you will do research.**
  - Facing problems and questions to which there isn't a ready answer.
  - Research methods is about finding answers systematically.
- A foundation for your course.
  - Research understanding and skills you need for other modules, placements and projects you undertake over the next year.
  - Learn about skills related to data-driven research.



"I have an in-depth background in research and analysis, which means I like Googling."

# Learning aims: Research Methods (SCC.400)

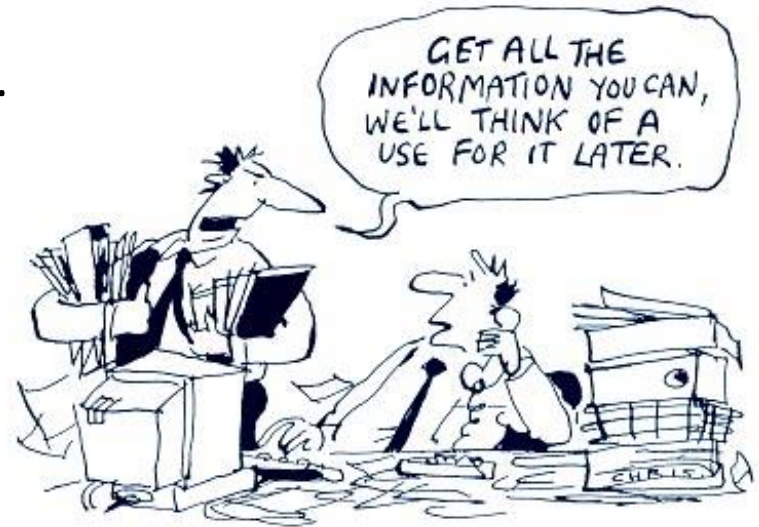
- Principled understanding of “how research works”.
- Awareness of different methods and different ways of thinking about research.
- Ability to critically reflect on research approaches and practices.





# Learning aims: Data Science skills (SCC.460/460-HDS)

- Choose appropriate data sources, methods and technologies.
- Design experiments and analyse findings.
- Awareness of professional practice in industry.



# Learning aims: Complementary skills

---

- SCC.400
  - Awareness of expectations in preparation of placements
  - Understanding of how to design and plan a research project
  - Developing skills for engagement with industry
  - Awareness of research areas in Computing & Communications
- SCC.460/460-HDS
  - Developing well-argued research plans
  - Design and plan group work tasks
  - Apply data skills on real data
  - Communicate results to a non-technical audience

# Two modules with a shared five-week core

<p>Weeks 1-5 (SCC. 400/460/460-HDS)</p>	<p><b>A crash course in research and data science methods</b></p> <ul style="list-style-type: none"> <li>• Lectures</li> <li>• In-class exercises</li> <li>• Industry guest talks</li> <li>• A bit of reading</li> <li>• Week-on-week quizzes (5×10%)</li> </ul>
<p>Weeks 6-10 (SCC.400) Weeks 6-13 (SCC. 460/460-HDS)</p>	<p><b>Group Work (50%)</b></p> <ul style="list-style-type: none"> <li>• Draft report</li> <li>• Final report</li> </ul>

# Module Structure: SCC.400 students

---

## MSci Students

### Michaelmas

SCC.400

*Research skills  
Skills for placement*

### Lent

SCC.419

*Placement in industry  
or research group*

### Summer

SCC.421

*Carry out your  
final project*

# Module Structure: SCC.460/460-HDS students

---

Michaelmas

Lent

Summer

SCC. 460/460-HDS

Placements

*Learn skills for  
carrying out research*

*Develop a proposal,  
exercise analysis skills*

*Carry out your  
industry placement*

# Assessment

---

- 100% coursework assessed. No exam!
- The five quizzes are shared by both SCC.400 and SCC.460 /460-HDS.
  - Weekly Moodle quiz based on lecture material and/or reading assignment.
  - Six questions in ten minutes, each worth 10% of course mark.
  - Each quiz opens on Friday afternoon (at 16:00) and closes one week later.
- The group-work **is not the same** for both modules!
  - SCC.400: Work in groups in weeks 6-10.
  - SCC.460: Works in groups in weeks 6-13.
  - [Draft Report (SCC.400) or Presentation (SCC.460 /460-HDS)] and Report: worth 50%

# Coursework submission / Moodle logistics

---

- Separate learning spaces for SCC.400 and SCC.460 /460-HDS
  - Contain separate overviews
  - Contain all coursework descriptions
  - Complete and submit quizzes in weeks 1 to 5
  - Complete and submit group-work in the second part of the module
  - Forum for discussion
  - Material used in weeks 1-5 will be uploaded on both Moodle spaces
  - Additional reading and learning resources may become available during the course of each module

# Feedback

- Automated feedback on quizzes
- Overall feedback on quizzes and class performance in lectures
- Feedback on coursework
  - Formative feedback: Detailed feedback on progress (oral / written)
  - Summative feedback



# What is Plagiarism?

---

- Passing off someone else's work as your own, including:
  - Submitting work that someone else wrote
  - Paying for someone else to do it for you
  - Working on a piece of non-group work together as a group, and submitting it as individual work
  - Sharing of code that you then possibly adapt
  - Describing work of others by copying their description
- Coursework is submitted online and automatically checked for plagiarism.

# Syllabus: Research Methods (SCC.400)

---

- Week 1: What is Research?
- Week 2: Language of Research and Framing Research Projects
- Week 3: Empirical Research Methods and Experimental Research
- Week 4: Study Design and Validity
- Week 5: Ethics

# Syllabus: Data Science Fundamentals (SCC.460/460-HDS)

---

- Week 1: Data Collection and Cleaning
- Week 2: Data Biases, Integration, and Transformation
- Week 3: Big Data Technologies
- Week 4: Modelling and Experimental Design
- Week 5: Testing Hypotheses, Analysis and Visualisation

# What we expect from you: Reading

---

- Literature on research methods
- Practice articles that you will analyze
  - Focus analysis on the research argument and method
  - The subject-specific content is secondary
- Deep reading
  - Read and read again; take notes
- Critical reading:
  - Don't just absorb
  - Question what you read, form your own opinion

# What we expect from you: Interaction

---

- This is your course
  - Bring questions, flag material for discussion
  - Read around the topic
  - Use resources (web, library, etc.)
- Discussion
  - Large class, but the intention is to make it as interactive as possible
  - Don't hide
- Plan your time carefully!

# What to expect from us

---

- We will do our best to:
  - Make all our lecture material available on Moodle, before class
  - Give you follow up references and additional reading material
  - Offer prompt feedback on coursework, answer questions
- Respond to queries (sent by email or posted on the Discussion Forum)
  - Hans and Elisa (SCC.400); Keith, Richard and Ignatius (SCC.460/460-HDS)
  - Use your University email addresses
  - Please don't send email out of hours

# Quiz 1: Quiz on Research Methods Foundations

---

- Trochim's Research Methods Knowledge Base
  - Language of research; Fundamentals of sampling, measurement, design, analysis; Different types of validity.
  - Quite a lot of material: **start early and make sure to allocate quality time to this.**
- Presented from a social science background
  - Wherever they use examples from social sciences, practice making up examples from computing, engineering or data science
- Quiz will go live at 16:00 today
  - 6 multiple choice questions selected at random.
  - 10 minutes to complete once you have started the quiz
  - Deadline: 16:00 on Friday Week 2. Automated scoring.

# SCC.400 Research Methods

## SCC.460/460-HDS Data Science Fundamentals

What is Research - Motivations, Types and Outcomes of Research





# Today's Agenda

---

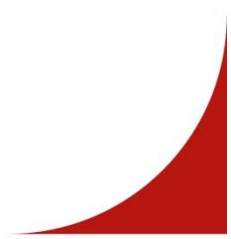
- Introduction to research
  - What is research
  - Types
  - Motivations
  - Outcomes
  - Reading exercise
  - Quiz 1

# What is research?

---

research, *noun*.

Systematic investigation into some domain or problem to reach new conclusions and derive new insights.

- Planned and managed to attain credible outcomes.
  - Expected to lead to a reliable, well-grounded contribution i.e. one that establishes new and relevant understanding.
  - Cyclical: always leads to further questions.
  - Creative: requires imagination, initiative, intuition, and curiosity.
- 

# What is *not* research?

---

- Accidental discovery or anecdotal reporting.
- Data collection.
- Reviewing a lot of literature.
- Talking to people about your results.

# Types of research studies

---

- Quantitative vs. Qualitative
  - Quantitative: measures variables using numerical methods
  - Qualitative: limited variables for in-depth study, e.g. interviews and ethnography
- Theoretical vs. Applied
  - Theoretical: deals with fundamental facts, concepts and relationships
  - Applied: targets a specific problem in the world
- Descriptive vs. Explanatory
  - Descriptive: answers basic questions (who? what? when?) in detail
  - Explanatory: explains phenomena and relationships (why? how?)
- Cross-sectional vs. Longitudinal
  - Cross-sectional: making observations at a distinct point in time
  - Longitudinal: measurement of a variable over a finite number of discrete time points

# Types: Quantitative or Qualitative?

## Probing Communities: Study of a Village Photo Display

Nick Taylor, Keith Cheverst, Dan Fitton, Nicholas J. P. Race, Mark Rouncefield and Connor Graham

Lancaster University  
Computing Department  
Lancaster, LA1 4WA, UK  
+44(0)1524 510311

{n.taylor,kc.df,race}@comp.lancs.ac.uk, {m.rouncefield,c.graham}@lancaster.ac.uk

### ABSTRACT

In this paper we describe a technology probe aiming to aid understanding of how digital displays can help support communities. Using a simple photo gallery application, deployed in a central social point in a small village and displaying user-generated photos and videos, we have been able to gain an understanding of this setting, field test our device and inspire new ideas directly from members of the community. We explore the process of deploying this display, the response from residents and how the display has taken a place within the community.

### Categories and Subject Descriptors

H.5.m [Miscellaneous].

### General Terms

Design, Human Factors.

### Keywords

Technology probe, situated display, community, interaction, content sharing, deployment, field study, digital photos.

## 1. INTRODUCTION

Our work aims to understand the ways in which the physical

- **Relationships.** Communities are based on meaningful relationships within the group, which define expectations and acceptable practices.
- **Change.** Communities are dynamic and develop a sense of history as they change and evolve.

We have been able to see the impact of these features on digital displays ourselves with the assistance of residents in Wray (Figure 1), a village situated 16km from Lancaster in the North West of England covering a geographical area of approximately 2km<sup>2</sup>. It is a vibrant community with a mix of attractive historic stone cottages and some newer developments on the edge of the village where farms and other rural industries used to reside. There are around 120 houses with a total population of under 500 representing all age groups, but with a slight bias towards the older generation.



Figure 1. Wray Village, Lancashire.

## Watching Television Over an IP Network

Meeyoung Cha  
MPI-SWS  
Saarbrücken, Germany

Pablo Rodriguez  
Telefonica Research  
Barcelona, Spain

Jon Crowcroft  
University of Cambridge  
Cambridge, UK

Sue Moon  
KAIST  
Daejeon, Korea

Xavier Amatralain  
Telefonica Research  
Barcelona, Spain

### ABSTRACT

For half a century, television has been a dominant and pervasive mass media, driving many technological advances. Despite its widespread usage and importance to emerging applications, the ingrained TV viewing habits are not completely understood. This was primarily due to the difficulty of instrumenting monitoring devices at individual homes at a large scale. The recent boom of Internet TV (IPTV) has enabled us to monitor the user behavior and network usage of an entire network. Such analysis can provide a clearer picture of how people watch TV and how the underlying networks and systems can better adapt to future challenges. In this paper, we present the first analysis of IPTV workloads based on network traces from one of the world's largest IPTV systems. Our dataset captures the channel change activities of 250,000 households over a six-month period. We characterize the properties of viewing sessions, channel popularity dynamics, geographical locality, and channel switching behaviors. We discuss implications of our findings on networks and systems, including the support needed for fast channel changes. Our data analysis of an operational IPTV system has important implications on not only existing and future IPTV systems, but also the design of the open Internet TV distribution systems such as Joost and BBC's iPlayer that distribute television on the wider Internet.

## 1. INTRODUCTION

Since the 1950's, television has been a dominant and pervasive mass media; it is watched across all age groups and by almost all countries in the world. Over the years, television has transformed itself into a new media. The number of channels has increased from a few free-to-air broadcasts to several hundreds for cable, satellite, and Internet TV networks, that transmit more channels to each user. The video signal itself has changed from black & white and color analog to high-quality digital stream.

Many technological advances were produced by trying to meet user needs and expectations in such a widespread media. For example, the large number of users that concurrently watch TV sparked the use of IP multicast by network operators to provide Internet TV (IPTV) services with low transmission costs. And now traditional media is converging with newer Internet-based services (e.g., Joost [1], Zattoo [2], Livestation [3], and BBC's iPlayer [4]). In such a context, IPTV is a promising starting point for research because it opens up the door for many innovations while still keeping its roots in the traditional TV watching paradigm.

Despite the widespread usage of television and its importance to emerging applications, the ingrained TV viewing habits are not completely understood. Nielsen Media Research [5] spearheads a long-standing research effort to en-

# Types: Theoretical or Applied?

## Facebook Linked Data via the Graph API

Editor(s): Pascal Hitzler, Wright State University, U.S.A.

Solicited review(s): Michael Hausenblas, DERI Galway, Ireland; Ivan Herman, W3C; Amit Joshi, Wright State University, U.S.A.

Jesse Weaver <sup>a,\*</sup> and Paul Tarjan <sup>b</sup>

<sup>a</sup> *Tetherless World Constellation, Rensselaer Polytechnic Institute, 110 8th St., Troy, NY, USA*

*E-mail: weavej3@cs.rpi.edu*

<sup>b</sup> *Facebook Inc., 1601 Willow Road, Menlo Park, CA, USA*

*E-mail: pt@fb.com*

**Abstract.** Facebook's Graph API is an API for accessing objects and connections in Facebook's social graph. To give some idea of the enormity of the social graph underlying Facebook, it was recently announced that Facebook has 901 million users, and the social graph consists of many types beyond just users. Until recently, the Graph API provided data to applications in only a JSON format. In 2011, an effort was undertaken to provide the same data in a semantically-enriched, RDF format containing Linked Data URIs. This was achieved by implementing a flexible and robust translation of the JSON output to a Turtle output. This paper describes the associated design decisions, the resulting Linked Data for objects in the social graph, and known issues.

**Keywords:** Linked Data, Facebook, Graph API, Turtle, JSON

### 1. Introduction

Facebook's Graph API<sup>1</sup> "presents a simple, consistent view of the Facebook social graph, uniformly rep-

resenting (HTTP(S)) URIs that dereference in accordance with httpRange-14 [2]. The effort had three primary restrictions: (1) to make only minimal changes to existing code, (2) to make the solution robust enough to

## Assessing Linked Data Mappings Using Network Measures

Christophe Guéret<sup>1</sup>, Paul Groth<sup>1</sup>, Claus Stadler<sup>2</sup>, and Jens Lehmann<sup>2</sup>

<sup>1</sup> Free University Amsterdam, De Boelelaan 1105, 1081HV Amsterdam  
{c.d.m.gueret,p.t.groth}@vu.nl

<sup>2</sup> University of Leipzig, Johannisgasse 26, 04103 Leipzig  
{cstadler,lehmann}@informatik.uni-leipzig.de

**Abstract.** Linked Data is at its core about the setting of links between resources. Links provide enriched semantics, pointers to extra information and enable the merging of data sets. However, as the amount of Linked Data has grown, there has been the need to automate the creation of links and such automated approaches can create low-quality links or unsuitable network structures. In particular, it is difficult to know whether the links introduced improve or diminish the quality of Linked Data. In this paper, we present LINK-QA, an extensible framework that allows for the assessment of Linked Data mappings using network metrics. We test five metrics using this framework on a set of known good and bad links generated by a common mapping system, and show the behaviour of those metrics.

**Keywords:** linked data, quality assurance, network analysis.

# Types: Descriptive or Explanatory?

## Why We Twitter: Understanding Microblogging Usage and Communities

Akshay Java  
University of Maryland Baltimore County  
1000 Hilltop Circle  
Baltimore, MD 21250, USA  
aks1@cs.umbc.edu

Tim Finin  
University of Maryland Baltimore County  
1000 Hilltop Circle  
Baltimore, MD 21250, USA  
finin@cs.umbc.edu

Xiaodan Song  
NEC Laboratories America  
10080 N. Wolfe Road, SW3-350  
Cupertino, CA 95014, USA  
xiaodan@sv.nec-labs.com

Belle Tseng  
NEC Laboratories America  
10080 N. Wolfe Road, SW3-350  
Cupertino, CA 95014, USA  
belle@sv.nec-labs.com

### ABSTRACT

Microblogging is a new form of communication in which users can describe their current status in short posts distributed by instant messages, mobile phones, email or the Web. Twitter, a popular microblogging tool has seen a lot of growth since it launched in October, 2006. In this paper, we present our observations of the microblogging phenomena by studying the topological and geographical properties of Twitter's social network. We find that people use microblogging to talk about their daily activities and to seek or share information. Finally, we analyze the user intentions associated at a community level and show how users with similar intentions connect with each other.

### Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information

provided by several services including Twitter<sup>2</sup>, Jaiku<sup>3</sup> and more recently Pownce<sup>4</sup>. These tools provide a light-weight, easy form of communication that enables users to broadcast and share information about their activities, opinions and status. One of the popular microblogging platforms is Twitter [29]. According to ComScore, within eight months of its launch, Twitter had about 94,000 users as of April, 2007 [9]. Figure 1 shows a snapshot of the first author's Twitter homepage. Updates or posts are made by succinctly describing one's current status within a limit of 140 characters. Topics range from daily life to current events, news stories, and other interests. IM tools including Gtalk, Yahoo and MSN have features that allow users to share their current status with friends on their buddy lists. Microblogging tools facilitate easily sharing status messages either publicly or within a social network.

## Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi

Palo Alto Research Center, Inc.  
Palo Alto, CA, U.S.A.  
{suh, hong, pirolli, ech}@parc.com

**Abstract**— Retweeting is the key mechanism for information diffusion in Twitter. It emerged as a simple yet powerful way of disseminating information in the Twitter social network. Even though a lot of information is shared in Twitter, little is known yet about how and why certain information spreads more widely than others. In this paper, we examine a number of features that might affect retweetability of tweets. We gathered content and contextual features from 74M tweets and used this data set to identify factors that are significantly associated with retweet rate. We also built a predictive retweet model. We found that, amongst content features, URLs and hashtags have strong relationships with retweetability. Amongst contextual features, the number of followers and followers as well as the age of the account seem to affect retweetability, while, interestingly, the number of past tweets does not predict retweetability of a user's tweet. We believe that this research would inform the design of sensemaking and analytics tools for social media streams.

**Keywords**—Twitter; retweet; tweet; follower; social network; social media; factor analysis

retweeting actions are often associated with certain values of the original information items [2]. Retweeting may be to entertain a specific audience, to comment on someone's tweet, to publicly agree with someone, or to save tweets for future personal access. These actions suggest that the original tweet contains valuable information [2].

Another interesting investigation on retweeting is Zarrella's series of blog posts [12]. Zarrella showed that retweets have quite different content characteristics from normal tweets. For example, he reported that 56.7% of retweets have URLs in them while only 19.0% of regular tweets have URLs. This suggests that retweets are used to spread interesting web pages, videos, and other web content to other users. Zarrella's work focused mainly on direct content analysis of the retweets and the original tweets themselves, such as the most likely words to be retweeted, types of URL shortening services used, and reading grade level of the retweets. His recent posts have also examined the depth that tweets reach.

*In our work, we are interested in extending his findings to*



# Types: Cross-sectional or Longitudinal?

## It Bends but Would it Break? Topological Analysis of BGP Infrastructures in Europe

Sylvain Frey, Yehia Elkhatib, Awais Rashid, Karolina Follis, John Vidler, Nicholas Race, Christopher Edwards  
Security Lancaster Research Centre, Lancaster University, United Kingdom  
{s.frey, y.elkhatib, a.rashid, k.follis, j.vidler, n.race, c.edwards}@lancaster.ac.uk

**Abstract**—The Internet is often thought to be a model of resilience, due to a decentralised, organically-grown architecture. This paper puts this perception into perspective through the results of a security analysis of the Border Gateway Protocol (BGP) routing infrastructure. BGP is a fundamental Internet protocol and its intrinsic fragilities have been highlighted extensively in the literature. A seldom studied aspect is how robust the BGP infrastructure actually is as a result of nearly three decades of perpetual growth. Although global black-outs seem unlikely, local security events raise growing concerns on the robustness of the backbone. In order to better protect this critical infrastructure, it is crucial to understand its topology in the context of the weaknesses of BGP and to identify possible security scenarios. Firstly, we establish a comprehensive threat model that classifies main attack vectors, including but not limited to BGP vulnerabilities. **We then construct maps of the European BGP backbone based on publicly available routing data.** We analyse the topology of the backbone and establish several disruption scenarios that highlight the possible consequences of different types of attacks, for different attack capabilities. We also discuss existing mitigation and recovery strategies, and we propose improvements to enhance the robustness and resilience of the backbone. To our knowledge, this study is the first to combine a comprehensive threat analysis of BGP infrastructures with advanced network topology considerations. We find that the BGP infrastructure is at higher risk than already understood, due to topologies that remain vulnerable to certain targeted attacks as a result of organic deployment over the years. Significant parts of the system are still uncharted territory, which warrants further investigation in this direction.

are necessary for the Internet to be considered truly resilient against targeted attacks.

The Border Gateway Protocol (BGP) is the de facto standard for routing between large IP networks, called Autonomous Systems (AS). ASs advertise their existence and the range of addresses they own to the rest of the Internet. BGP ensures all other ASs know about the various subnets and how to reach them. Without BGP, each sub-network would be isolated and unreachable.

Since its introduction in the 1980's, BGP has grown to become an important part of running the Internet's core. The infrastructure, both logical – i.e. AS routing tables – and physical – i.e. BGP routers and the traffic highways connecting them – is thus quite a critical one. However, like many other key internet technologies, BGP as a protocol was not designed with security requirements. There is plenty of evidence to show that it is intrinsically fragile (e.g. [1]–[3]) and significant disruptions to parts of the backbone raise the question of its robustness as a whole. The BGP infrastructure has grown over the years into what is now a particularly complex system. Such an organic growth results notably in a scale-free topology that is known to be particularly resilient against random failures but remains vulnerable to targeted attacks [4].

This paper aims at shedding some light onto the security of BGP infrastructures in Europe. Our methodology is reproducible for other portions of the network, and the findings of the paper are of high importance to network operators and regulators both within and outside European countries. Understanding the Internet as a critical infrastructure from a security perspective is essential. Such a relatively young system that has grown mostly unsupervised is a tempting

## Studying Interdomain Routing over Long Timescales

Giovanni Comarela  
Boston University  
Boston, USA  
gcom@bu.edu

Gonca Gürsun  
Boston University  
Boston, USA  
goncag@bu.edu

Mark Crovella  
Boston University  
Boston, USA  
crovella@bu.edu

### ABSTRACT

The dynamics of interdomain routing have traditionally been studied through the analysis of BGP update traffic. However, such studies tend to focus on the volume of BGP updates rather than their effects, and tend to be local rather than global in scope. Studying the global state of the Internet routing system over time requires the development of new methods, which we do in this paper. We define a new metric, MRSD, that allows us to measure the similarity between two prefixes with respect to the state of the global routing system. Applying this metric over time yields a measure of how the set of total paths to each prefix varies at a given timescale. We implement this analysis method in a MapReduce framework and apply it to a dataset of more than 1TB, collected daily over 3 distinct years and monthly over 8 years. We show that this analysis method can uncover interesting aspects of how Internet routing has changed over time. We show that on any given day, approximately 1% of the next-hop decisions made in the Internet change, and this property has been remarkably constant over time; the corresponding amount of change in one month is 10% and in two years is 50%. Digging deeper, we can decompose next-hop decision changes into two classes: churn, and structural (persistent) change. We show that structural change shows a strong 7-day periodicity and that it represents approximately 2/3 of the total amount of changes.

### 1. INTRODUCTION

There are many aspects of the interdomain routing system that are important to understand, including its stability, scalability, and security. However, a particularly difficult problem is understanding the overall structure of interdomain routing and how it evolves over time. The immense size, complexity, and continuous growth of the system make it challenging to gain a useful understanding of the nature of routing changes over time.

This problem is important because currently there are no metrics able to provide useful information about the rate of routing changes of the interdomain routing system. Such metrics could contribute at the understanding the impact of de-peering disputes, link failures, merging of autonomous systems and any other event that may affect globally the routing structure of the Internet.

To address this challenge, we focus in this paper on answering basic questions about how Internet routing has changed over time. As a first step, we are interested in characterizing the rate of change of the routing system and the dynamics of its change. We seek to answer these questions for the system as a whole (i.e., globally) and over long timescales. In this respect we differ from most prior work which has asked more specific questions about the evolution of Internet routing.

We believe that one reason that temporal change has not been extensively studied to date is that good methods and metrics to study it have not existed. Accordingly, the first question we make is to



# Research philosophies

---

- Positivist:
  - Knowledge is absolute and is to be discovered
  - Knowledge is part of the truth of the world
- Relativist:
  - Knowledge is relative to context and is produced
  - Knowledge is a way of looking at the world
- Many implications about whether science is objective or subjective. The argument is beyond the scope of this lecture.
- More reading: 'Research Methods' by McNeill & Chapman, 3<sup>rd</sup> edition and Trochim (see this week's quiz)

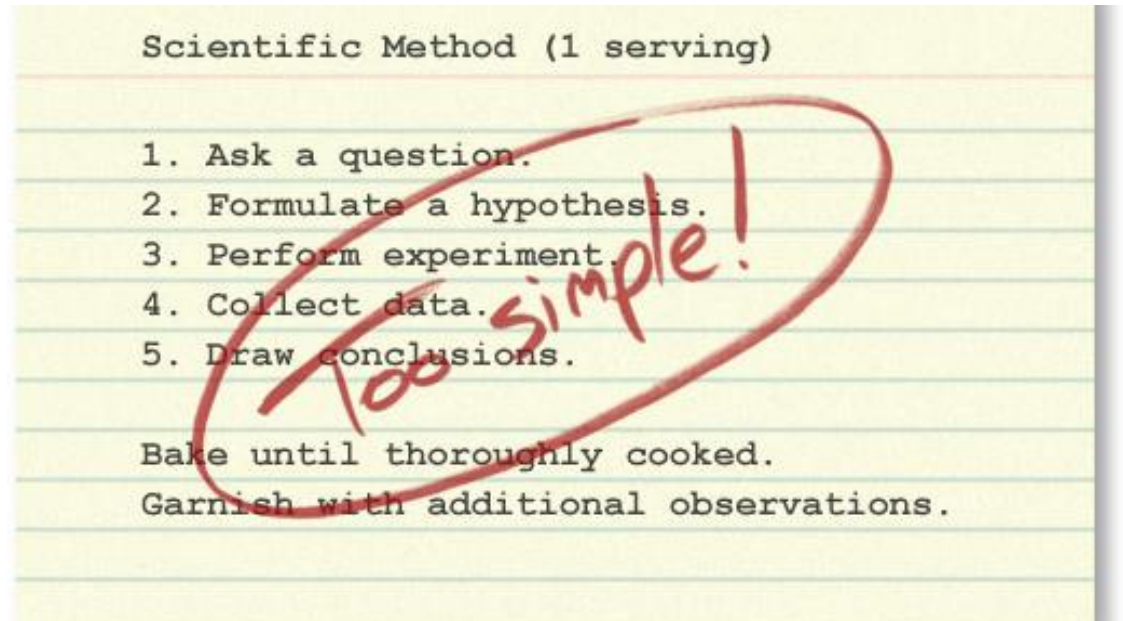
# The research process (1/2)

---

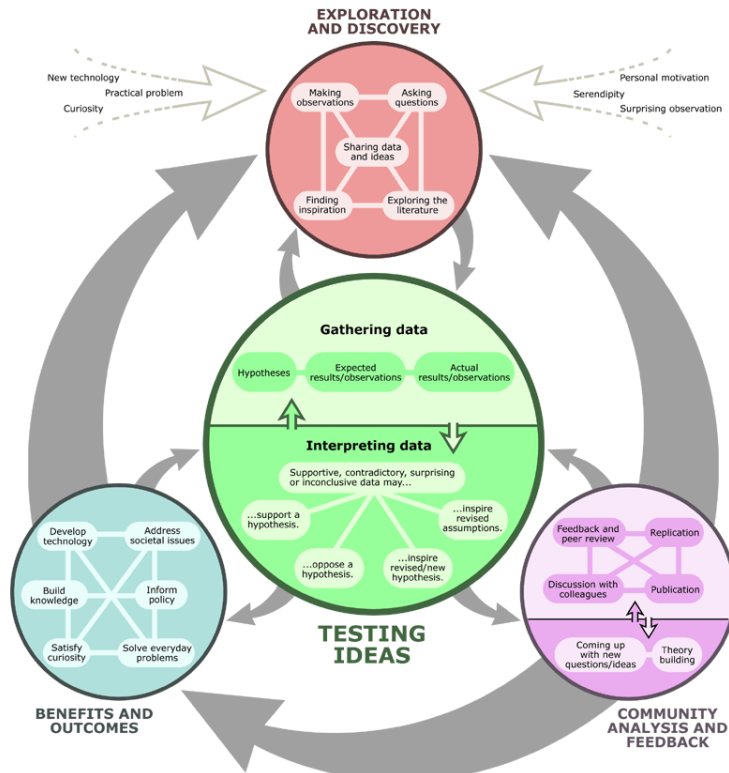
- A sequential model:
  1. Identify the broad area of study
  2. Select a research topic
  3. Review the literature
  4. Form hypotheses
  5. Decide on an approach
  6. Plan how you will perform the research
  7. Gather data and information
  8. Analyse and interpret the data
  9. Document and present the results and findings

## The research process (2/2)

In practice, it is never so simple ...



# 'Real' research process



- The process of research / science is iterative.
- The process of research / science is not predetermined.

[Visit "Understanding Science" by the University of Berkeley](#)

# Motivations for research

---

## What drives research?

- Add to the body of knowledge
- Solve a problem
- Come up with a better way
- Find out what happens
- Predict, plan, control
- Find evidence informing practice
- Contribute to societal needs
- Make life better for other people
- To gain academic degrees
- To gain research employment
  - Both academia & industry

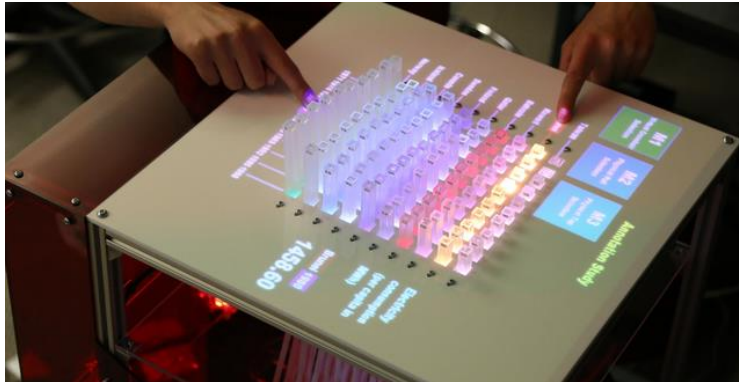
# Motivation examples: Solving a problem

- Quality of experience (QoE) fairness for competing video streams



# Motivation examples: Find out what happens

- Novel devices for interaction to find out what new capabilities they offer for human-computer interaction.
- Jason Alexander's work
  - shape-changing displays
  - 3D displays



# Motivation examples: Predict, plan, control

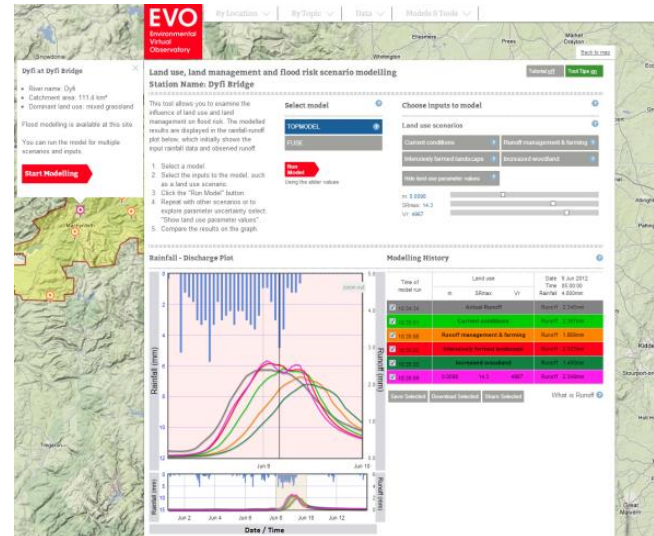
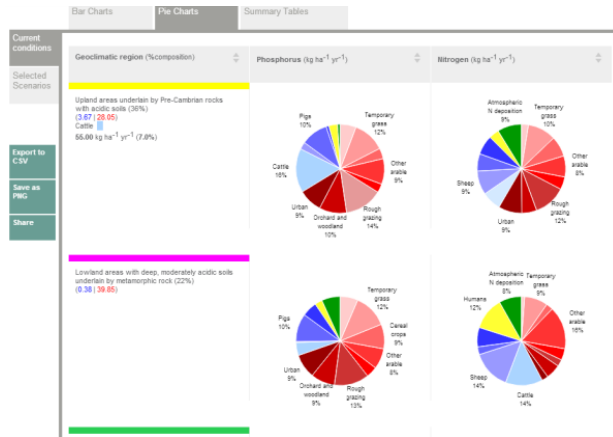
- Evolving Systems Toolbox by Professor Plamen Angelov
- Used by Ford to monitor:
  - machines that produce vehicles
  - the “health” of vehicles
  - reactions of drivers, e.g. how they press pedals
- Other industrial applications, e.g. Dow Chemical





# Motivation examples: Evidence informing practice

- EVOp: Helping policy makers answer ‘what if’ questions relating to land management and environmental impact.



# Motivation examples: Societal challenge

- Alistair Baron, Paul Rayson, Awais Rashid, *et al.*: Technology to protect children on social networks by automatically detecting and signaling grooming in chat rooms.



# Motivation examples: Make life better

- Research with specific user communities.
- Catalyst Access ASD (Jon Whittle, Maria Angela Ferrario, Debbie Stubbs, *et al.*): digital technology to reduce social barriers amongst people on the Autism Spectrum.



# Computer Science research is different

“Computer science research is different from these more traditional disciplines. Philosophically it **differs from the physical sciences** because it seeks not to discover, explain or exploit the natural world, but instead to study the properties of machines of human creation. In this, **it is analogous to mathematics**, and indeed the ‘science’ part of computer science is, for the most part mathematical in spirit.”

– **Dennis Ritchie (1941-2011)**



# Outcomes of research (1/2)

---

- What does research produce?
- What kinds of contribution?
- Is science the knowledge, or the methods to arrive at it?

## Outcomes of research (2/2)

---

- New knowledge / understanding
  - Propositional: know-that
  - Procedural: know-how
- “Products”
  - New or improved system / tool / technique
  - New or improved methodology
  - New or improved concepts / models / theories
  - New or improved evidence
  - New or improved analysis

# Reminder of Quiz 1: Reading and Quiz on Research Methods Foundations

---

- Trochim's Research Methods Knowledge Base
  - Quite a lot of material
  - **start early, and make sure to allocate quality time to this**
- Quiz will go live at 16:00 today
  - 6 multiple choice questions selected at random
  - 10 minutes to complete once you have started the quiz
  - Automated scoring
  - Deadline: 16:00 on Friday Week 2.

# What's Next

---

- **This afternoon (15:00-17:00, by Keith):**
  - Introduction to fundamentals of data science
  - Data collection and cleaning
- **Next week's Research Methods class (by Elisa):**
  - The Language of Research
  - Framing research and asking research questions