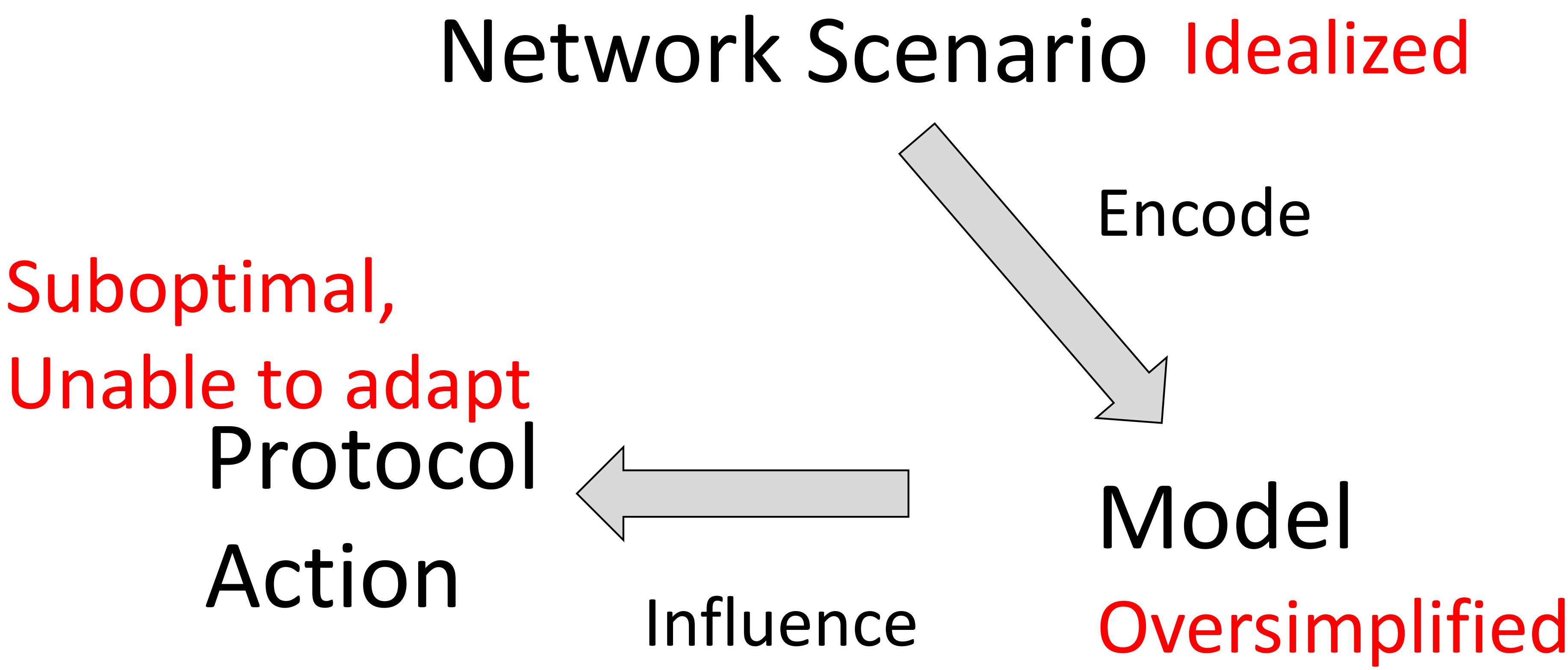
Smart, Thy Name is Data!

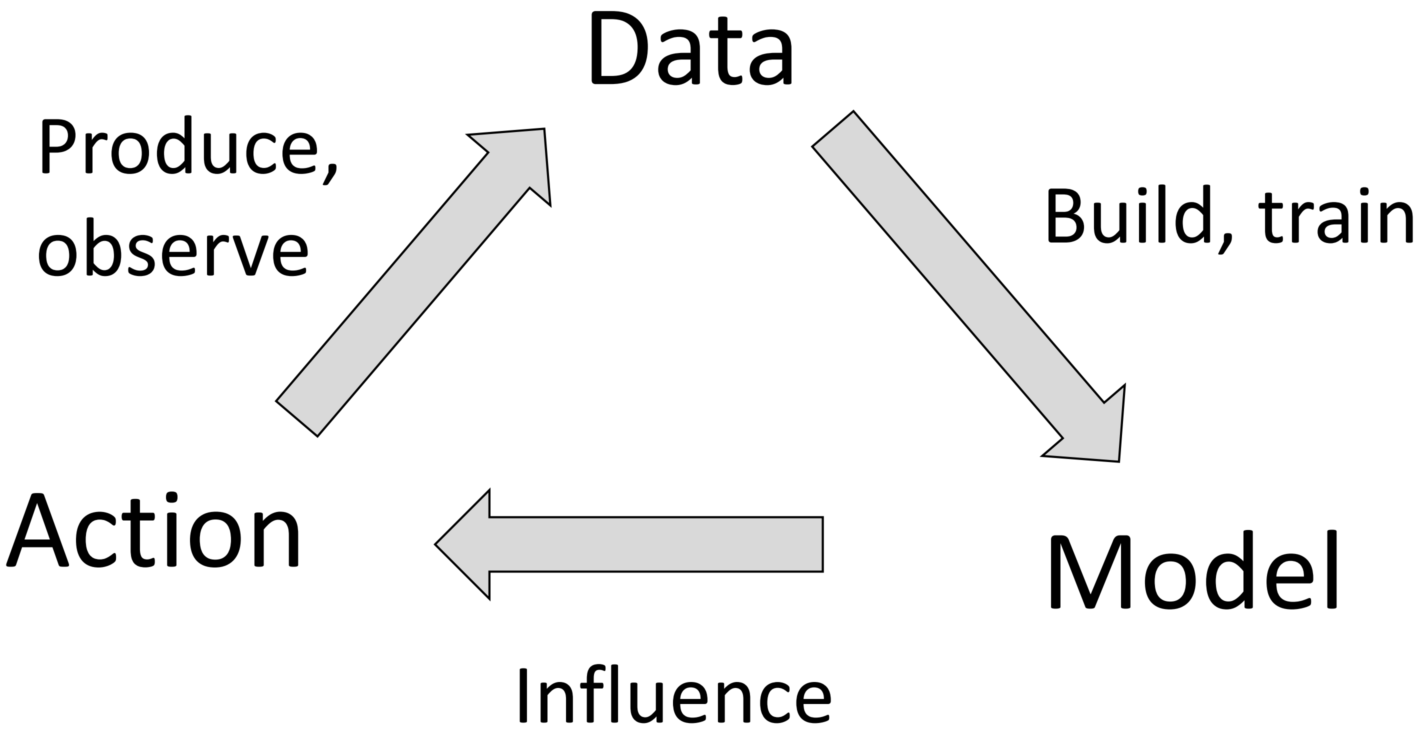
Data-Driven Approaches to Protocol Design

Samir Das
Stony Brook University

Network Protocol Design



Harness Power of Data



The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, *Google*

Eugene Wigner's article "The Unreasonable Effectiveness of Mathematics in the Natural Sciences"[1] examines why so much of physics can be neatly explained with simple mathematical formulas

such as $f = ma$ or $e = mc^2$. Meanwhile, sciences that involve human beings rather than elementary particles have proven more resistant to elegant mathematics. Economists suffer from physics envy over their inability to neatly model human behavior. An informal, incomplete grammar of the English

behavior. So, this corpus could serve as the basis of a complete model for certain tasks—if only we knew how to extract the model from the data.

Learning from Text at Web Scale

The biggest successes in natural-language-related machine learning have been statistical speech recognition and statistical machine translation. The reason for these successes is not that these tasks are easier than other tasks; they are in fact much harder than tasks such as document classification that extract just a few bits of information from each doc-

[IEEE Intelligent Systems 2009]

[SIGCOMM 2014]

An Experimental Study of the Learnability of Congestion Control

Anirudh Sivaraman, Keith Winstein, Pratiksha Thaker, and Hari Balakrishnan
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology, Cambridge, Mass.
{anirudh, keithw, prthaker, hari}@mit.edu

ABSTRACT

When designing a distributed network protocol, typically it is infeasible to fully define the target network where the protocol is intended to be used. It is therefore natural to ask: How faithfully do protocol designers really need to understand the networks they design for? What are the important signals that endpoints should listen to? How can researchers gain confidence that systems that work well on well-characterized test networks during development will also perform adequately on real networks that are inevitably more complex, or future networks yet to be developed? Is there a tradeoff between the performance of a protocol and the breadth of its intended operating range of networks? What is the cost of playing fairly with cross-traffic that is governed by another protocol?

We examine these questions quantitatively in the context of congestion control, by using an automated protocol-design tool to ap-

line of work has explored the use of in-network algorithms ning at bottleneck routers to help perform this function more ciently [11, 10, 9, 18, 23, 21, 16].

As the Internet grows and evolves, it appears likely that new work protocols will continue to be developed to accommodate subnetwork behaviors and shifting application workloads and g Some of these may be intended for specialized environmen e.g., inside a centrally-managed datacenter—while some will b broad use across the wide-area Internet, or over cellular net paths.

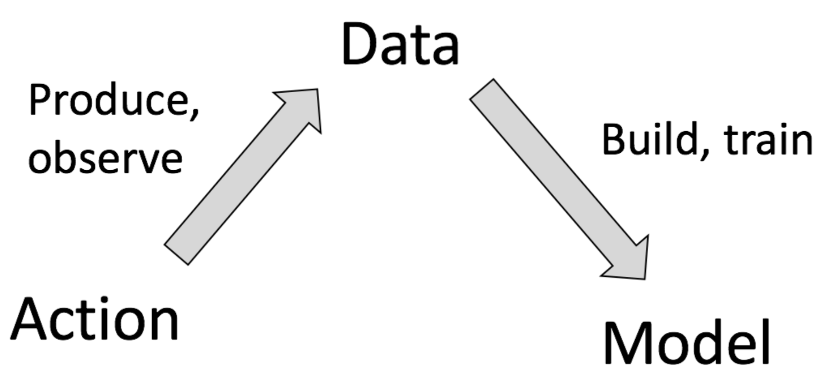
In practice, however, it is challenging to guarantee that a tributed system's performance on well-characterized test netw will extend to real networks, which inevitably differ from thos visioned in development and will continue to evolve over time. uncertain generalizability presents an obstacle to any new p col's deployment.

Examples from Home

- ExBox: QoE management middlebox for enterprise wireless networks [CONEXT 2016]
  - Need a capacity model.
  - Too many parameters for the network and applications. Many of them unknown.
- Managing mobile, distributed spectrum sensors [Ongoing]
  - Need a sensor performance model.
  - Too many device level characteristics influence model. Many of them unknown.

Approaches

- Classical ‘whitebox’ modeling approaches deliver very poor performance.
- Use a ‘blackbox’ approach. Probe the system’, observe response, learn a model.



Research Agenda

- How much data? Supervised vs Unsupervised?
- Scalability? Tools & systems?
- Time-scale? Online vs Offline?
- Are existing learning tools sufficient?
- Build bridges with ML community. Networks provide a rich experimental platform.