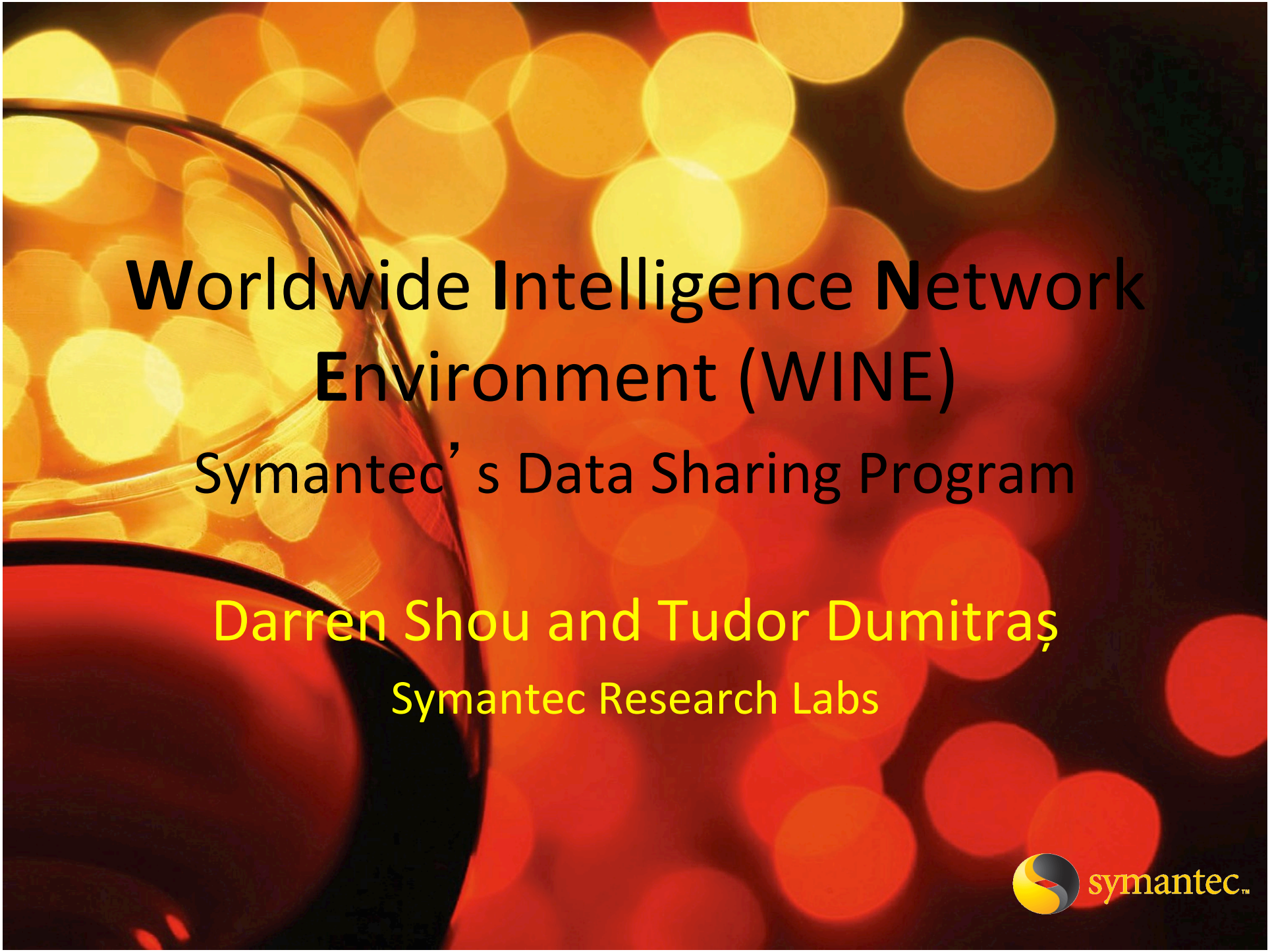


Worldwide **I**ntelligence **N**etwork **E**nvironment (WINE)

Symantec's Data Sharing Program

Darren Shou and Tudor Dumitraș

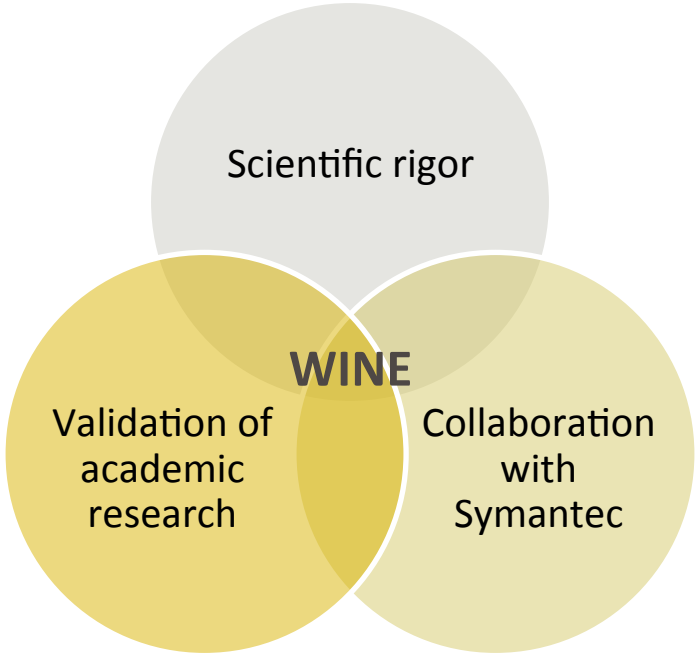
Symantec Research Labs

symantec.

Goals of WINE Project

- Enable sound experimentation for computer security
  - Create platform for repeatable research, comparable results
  - Allow re-running experiments against reference data sets

- Promote good science
  - Enable independent verification
  - Ensure statistical relevance
  - Reflect field data



http://www.symantec.com/about/profile/universityresearch/sharing.jsp

Why Symantec?

Have real-world data

- Norton Antivirus
  - Information on malware and attacks
- Brightmail spam appliance
  - Used by Fortune 500 companies
- Insight reputation security
  - Executables downloaded
- Honeypots around the world
  - Information on botnets
- Norton DNS
  - 17B+ DNS queries / day
- Message Labs
- Norton Online backup
- Shasta URL reputation
- Symantec Management Platform (Altiris)
- Endpoint Virtualization
- Data Loss Prevention
- Backup Exec
- Veritas Storage Foundation
- …

Why Symantec?

Want transformative, new techniques

- New approaches for fighting cybercrime
  - Malware
  - Spam
  - Botnets

- Can we tip the balance of the security arms race?

Malware Data Set

Malware samples collected by Symantec over years

- What
  - Binaries, statistical history
- How much
  - 5.5 million samples
- Growth rate
  - 10+ thousand samples / day

Reputation-Based Security Data Set

Reputation-based whitelisting of executables downloaded

- What
  - Machine hygiene rating history, file hashes, computed file rating
- How much
  - 30 TB
- Growth rate
  - 2 TB / month

Spam Data Set

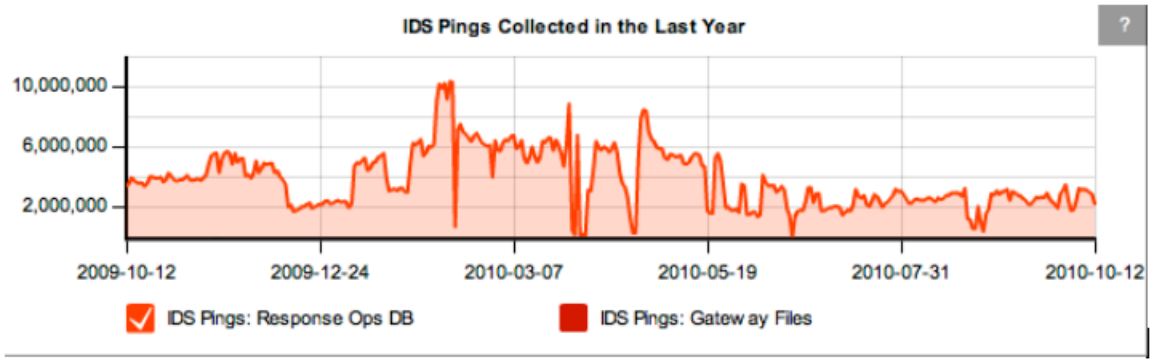
Logs of spam-filtering appliances

- What
  - Samples, statistical history
- How much
  - 100,000 samples
- Growth rate
  - Variable

Telemetry Data Sets

Notifications of threats detected by Norton products

- What
  - Attacking addresses, OS version, process name, geographic location
  - 18 different data feeds
- How much
  - 75 million machines
- Growth rate
  - Variable



Example: statistics for intrusion-detection telemetry

Operational Model

- Project proposals
  - Researchers in academia request access to data sets
  - NSF support: Trustworthy Computing program
    `http://www.gtisc.gatech.edu/nsf_workshop10_data.html`

- Internal operations
  - Collect data continuously
  - Each proposal's requested data is frozen as reference
  - Experimental environment is hosted **on SRL site**

- Selection of projects
  - Advisory board: senior researchers (external and internal)

Intellectual Property and Usage

- NDA to protect confidentiality of data
  - Provision for publication
- Symantec receives internal use copies of all research products
- Researchers assume all risks and liabilities from use of data
- All right, title and interest belong to the researchers
  - Unless licensing exception is negotiated beforehand
  - Data set should be acknowledged in publications

Many Ways to Use the Data

- Security
  - How many zero-day attacks are there?
  - Malware detection: can we do better than signatures and heuristics?
  - How do botnets spread? (and how can we stop them?)
- Machine learning
  - Belief propagation
  - Structure of large graphs (> 1B nodes)
- Software engineering and programming languages
  - Validate exploit-protection approaches

Challenges for the WINE System

- Data-intensive system
  - Store 100+ TB data
  - Ingest 10-20 TB/day, from multiple sources
  - Snapshots and clones
  - Analytics on all the data

- Platform for repeatable experimentation
  - Preserve reference data sets used in past experiments
  - Record minutiae of experimental procedures (lab book)
  - Produce comparable results in the future

What would you do with this data?

Thank you!