Applications of Principal Component Analysis

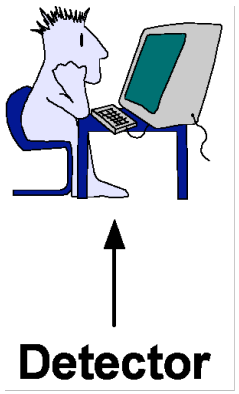
Todd Heberlein
Net Squared, Inc.

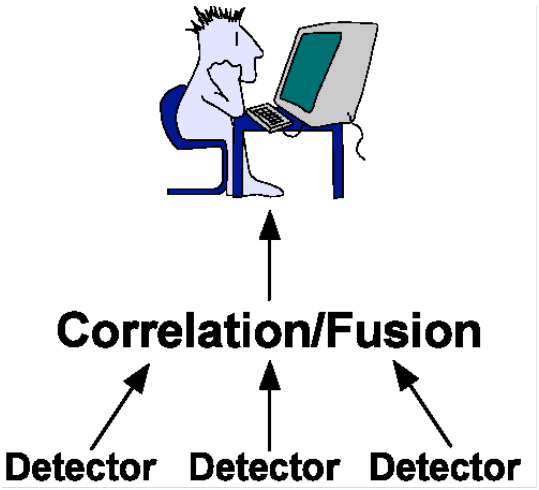
todd@NetSQ.com *heberlei@jps.net*

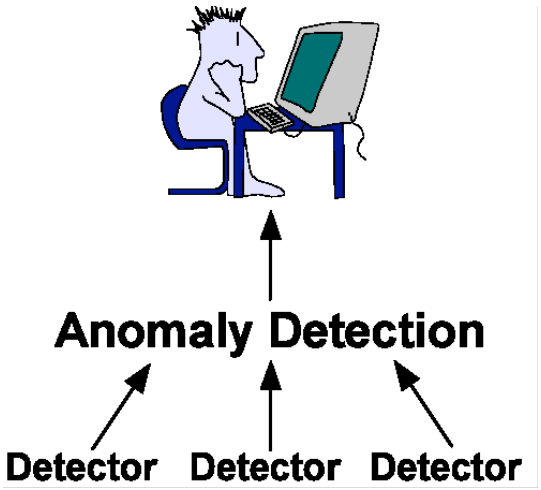
Outline

- Context for the PCA work
- Related PCA work
- PCA fundamentals
- PCA applications
  - PcaStream
  - ThumbprintStream
- Linux Audit System

History of Intrusion Detection



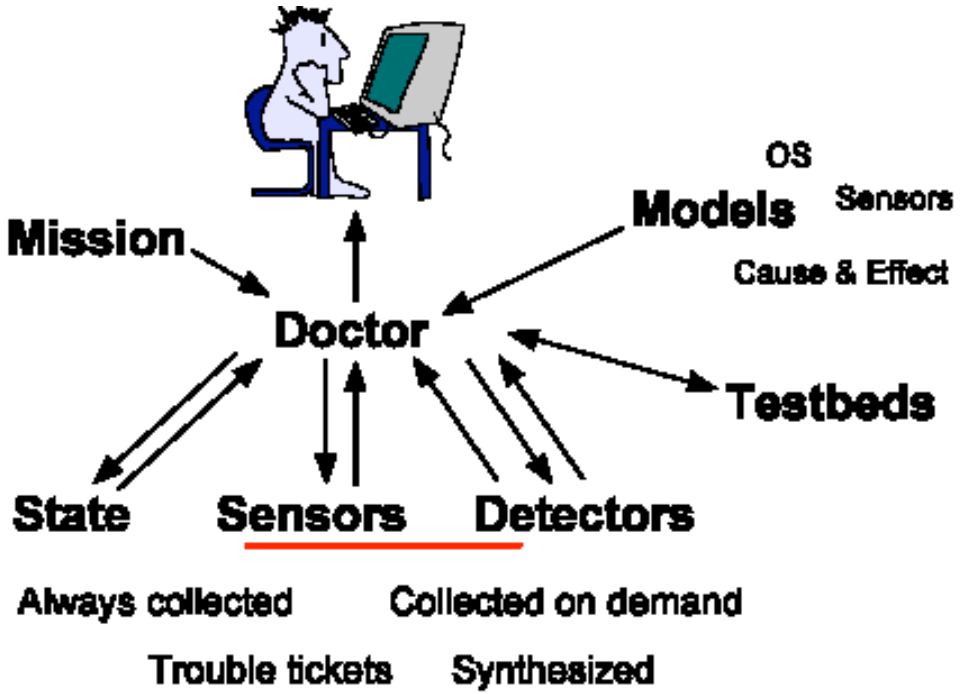




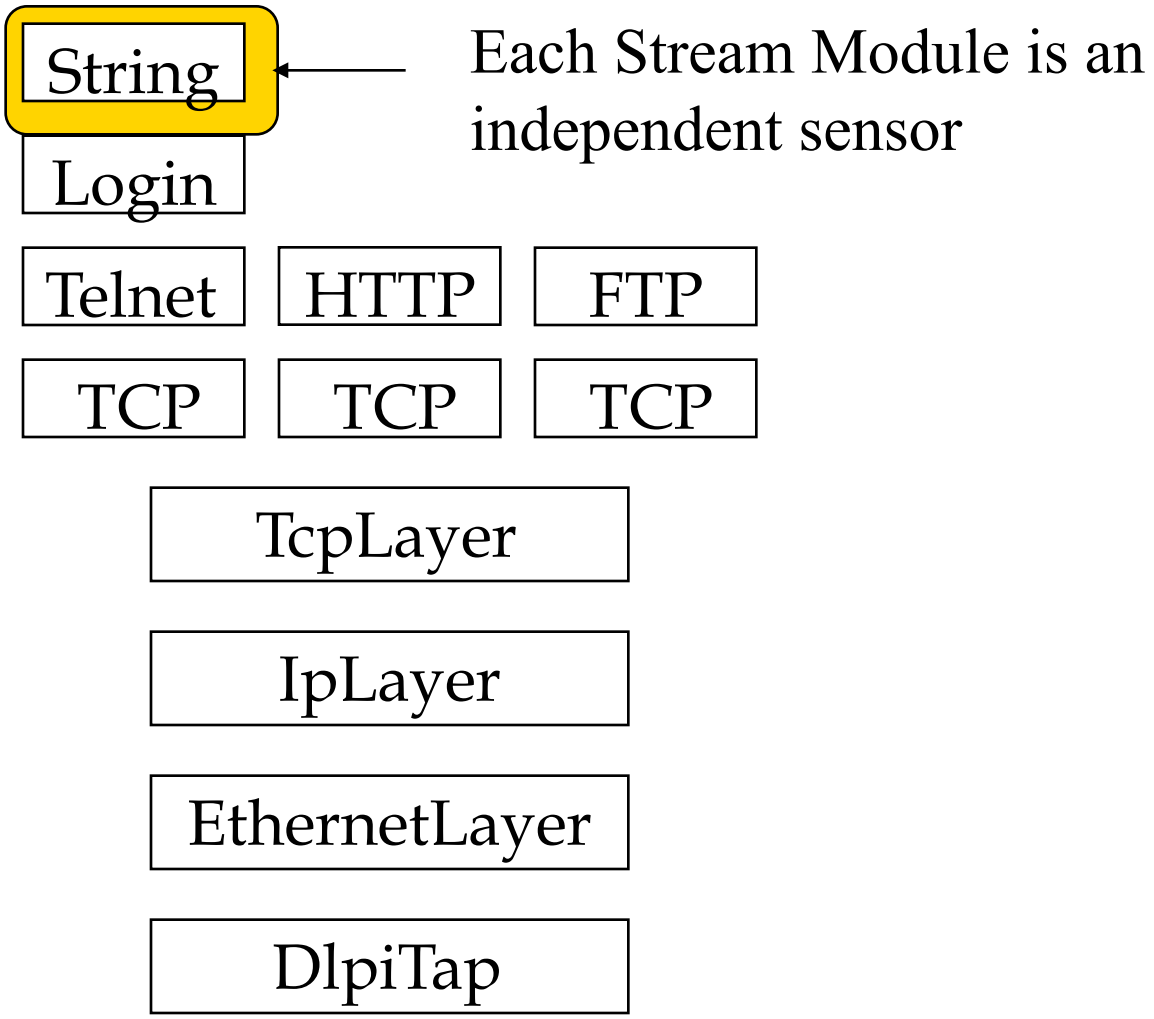
Understanding Original Events

- Invisible traffic
  - TCP only seeing 10% ok pkts seen by IP
  - Use snoop to identify individual fields
- Giant Packet (13,000+ bytes)
  - Only at Rome
  - Crashes our system, crashes snoop
  - Hex dumps, work arounds, still searching
- Lincoln's new protocol attacks

Future Intrusion Detection



Network Radar at a Glance



Each Stream Module is an independent sensor

Finding Related Content

- Repeated original attacks
  - worms
- Retroactive signatures
- Tracking users
- Similar documents
  - Downgrading?
  - Espionage (FTP)

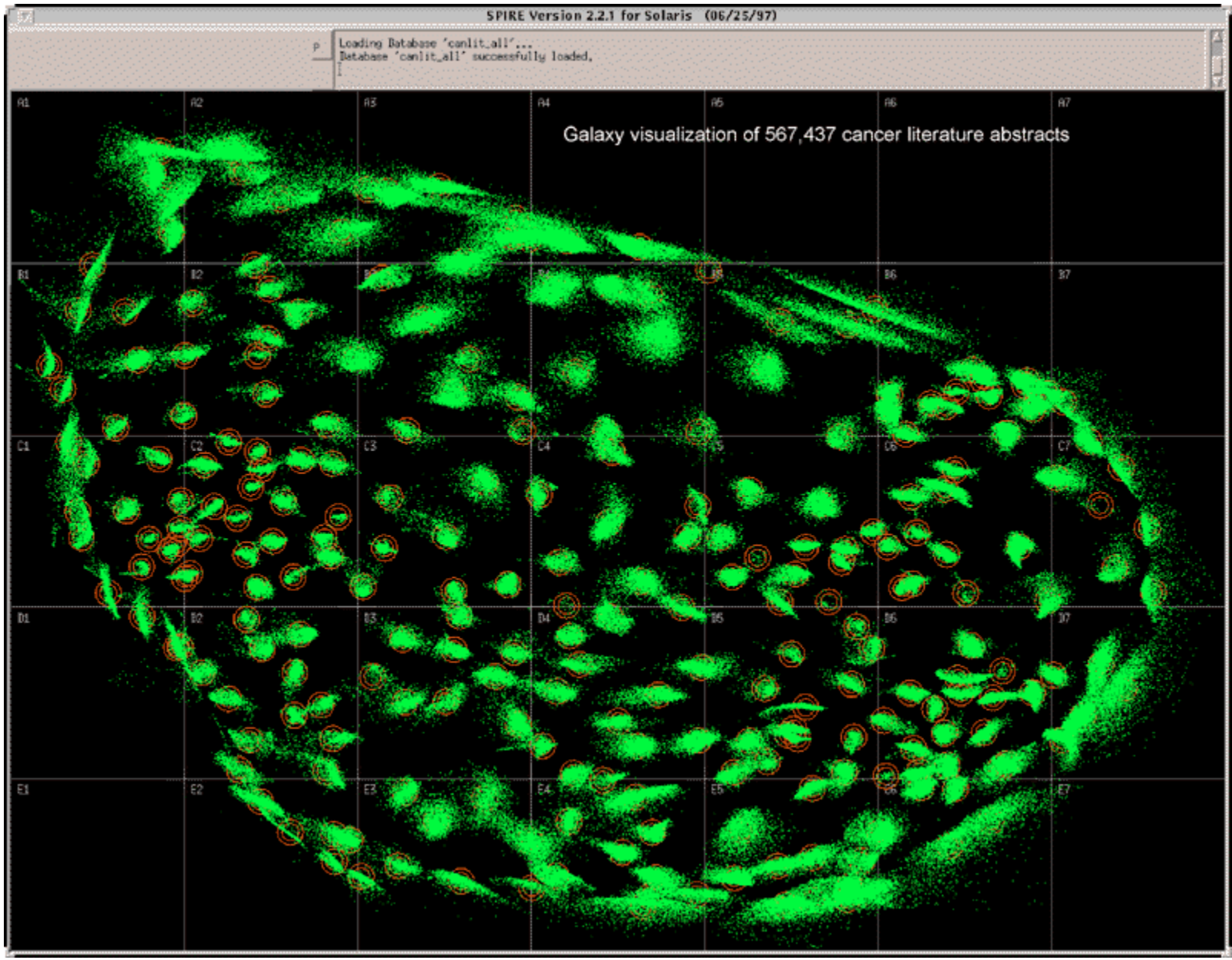
Related Work

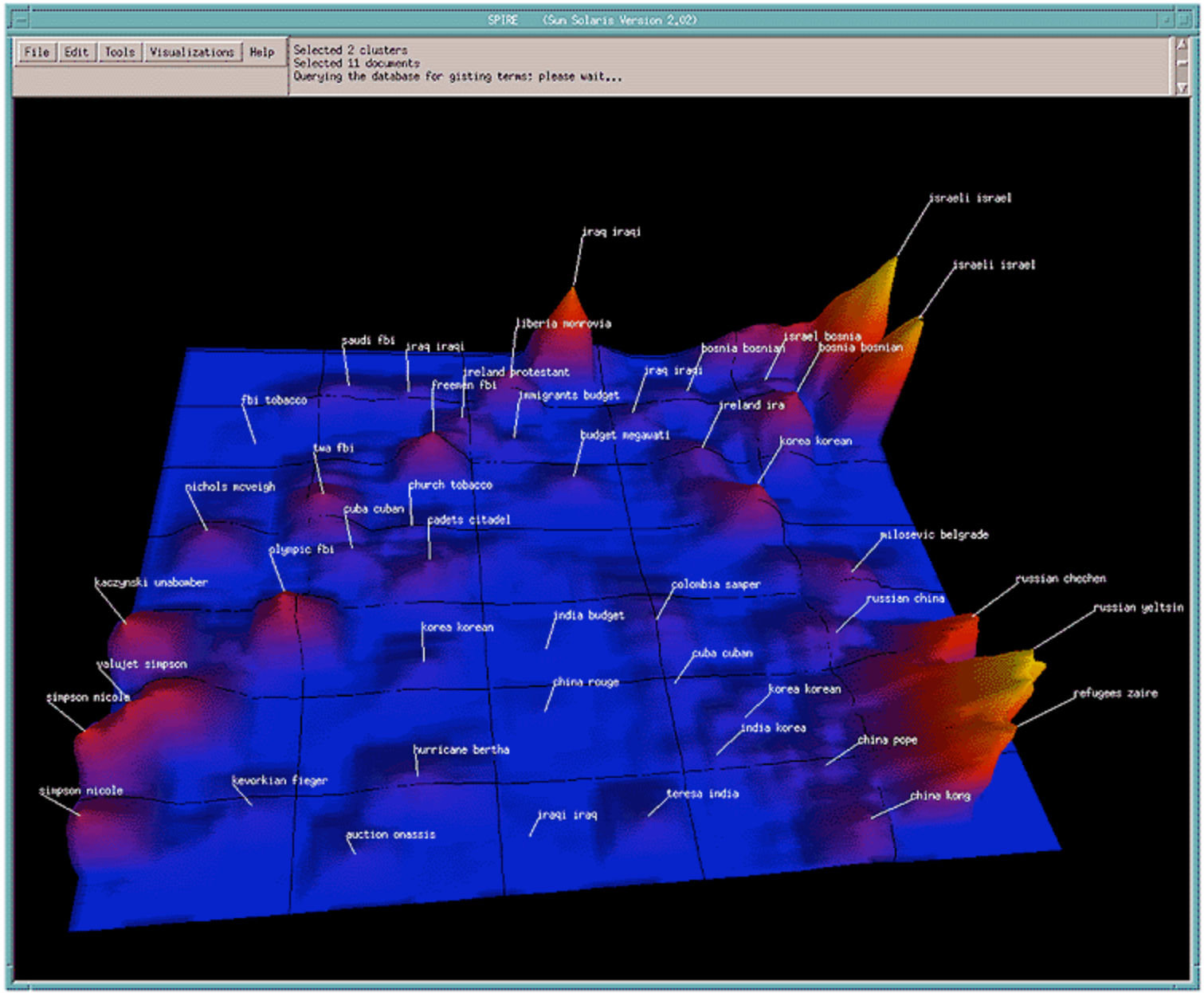
- SPIRE
- Themescape
- Previous thumbprinting work
- Tons of uses for PCA in lossy compression

SPIRE

- **S**patial **P**aradigm for **I**nformation **R**etrieval and **E**xploration
- Family of tools to perform interactive data exploration
- Focus is on text documents
- Performing the same functions we are

http://multimedia.pnl.gov:2080/infoviz/index.html


SPIRE Version 2.2.1 for Solaris (06/25/97)
Loading Database 'canlit_all'...
Database 'canlit_all' successfully loaded.
Galaxy visualization of 567,437 cancer literature abstracts


SPIRE (Sun Solaris Version 2.02)
File
Edit
Tools
Visualizations
Help
Selected 2 clusters
Selected 11 documents
Querying the database for gisting terms: please wait...
israeli israel
iraq iraqi
israeli israel
liberia monrovia
saudi fbi
iraq iraqi
israel bosnia
bosnia bosnian
bosnia bosnian
ireland protestant
iraq iraqi
freemen fbi
immigrants budget
fbi tobacco
ireland ira
budget megawati
korea korean
twa fbi
nichols mcveigh
church tobacco
cuba cuban
cadets citadel
milosevic belgrade
olympic fbi
kaczynski unabomber
colombia sanger
russian chechen
russian china
russian yeltsin
india budget
korea korean
valujet simpson
cuba cuban
china rouge
korea korean
simpson nicole
refugees zaire
india korea
china pope
hurricane bertha
kevorkian fieger
simpson nicole
teresa india
china kong
iraqi iraq
auction onassis

Themescape

- Themescape discovers clusters of related data sources
- NSA applied similar tools (parentage/acquaintance) to session data from a large DOD break-in
- Pacific Northwest National Laboratory is developing visualization tools

http://demo.cartia.com

Information as Document Vectors

- Identify document (information) of interest
- Represent document as an n-dimensional vector, n may be very large (128-500)
- **Goal:** document vectors representing similar information should be close in the n-dimensional space; document vectors representing dissimilar information should **not** be close in this n-dimensional space

Problems with Large Dimensions

- Impossible to visualize
- Memory intensive
  - calculating document vectors for 50,000+ simultaneous sessions
  - storing document vectors for millions of sessions per day
- Expensive to compare “closeness”
- Goal: Map n-dimensional vectors down to k-dimensional vectors (k << n), while preserving the information

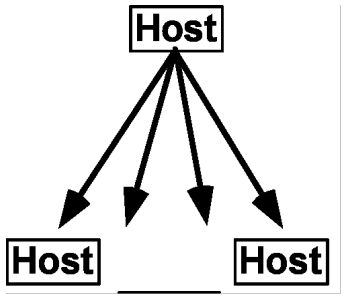
From n to k Dimensions

- Calculate covariance matrix for your set of document vectors
- Calculate eigen vectors and values for this matrix
- Select k largest eigen values - their eigen vectors become your new basis
- Project n-dimensional document vectors into k dimensional space
- Don’t need to calculate original n-dimensional vector

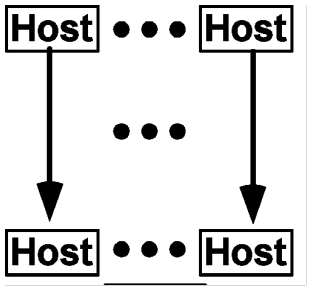
PcaStream

- Document model
  - transactional sessions
  - two documents per session: client and server
- Parameterized on each service
  - set of eigenvectors for client and server for www, sendmail, finger, etc.
- Two dimensions for each direction

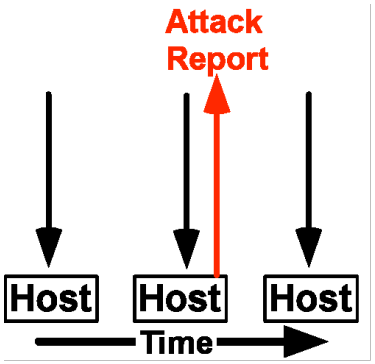
Uses of Session-based Thumbprints



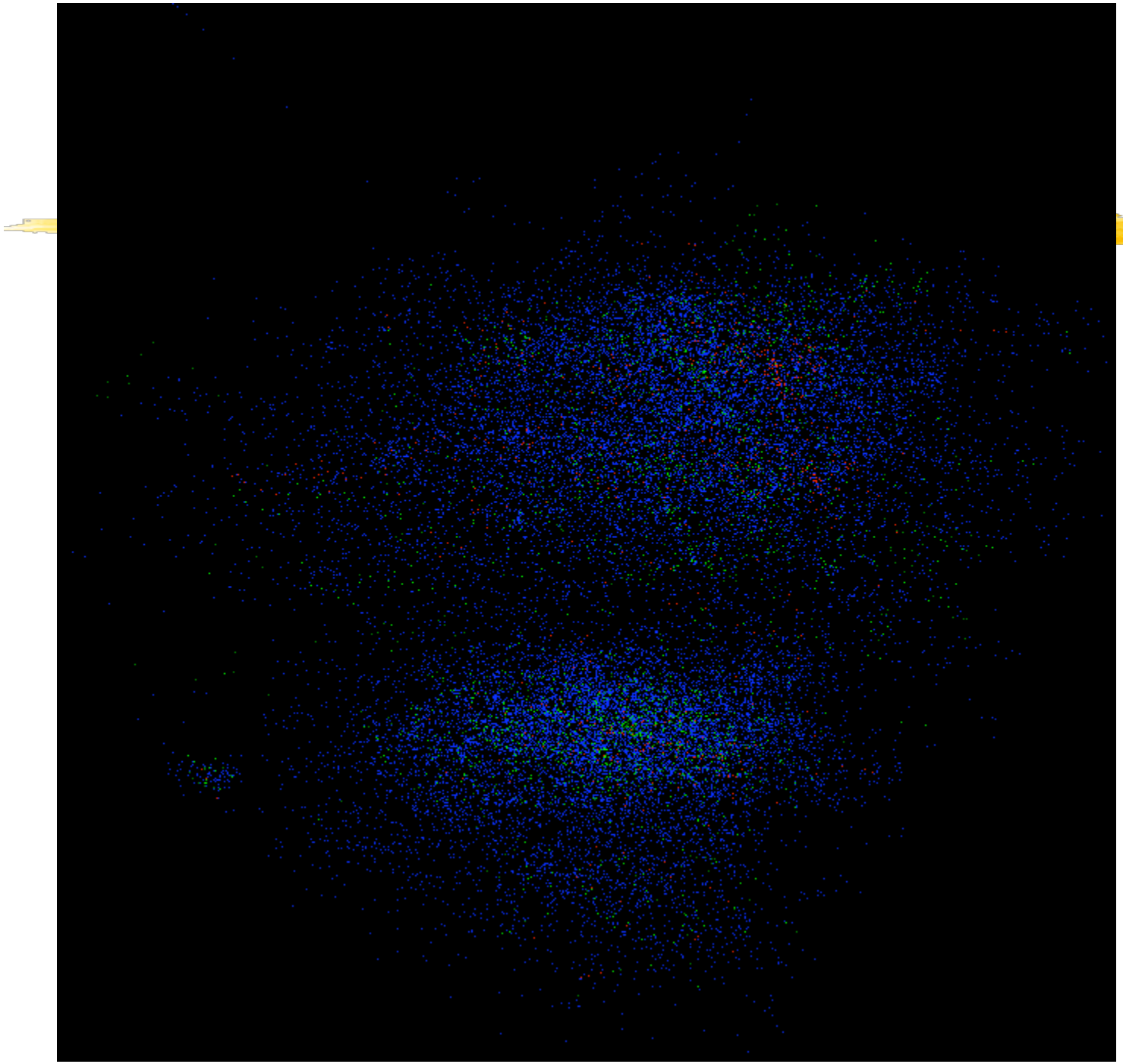
Detection of canned attack launched by a script - single mountain

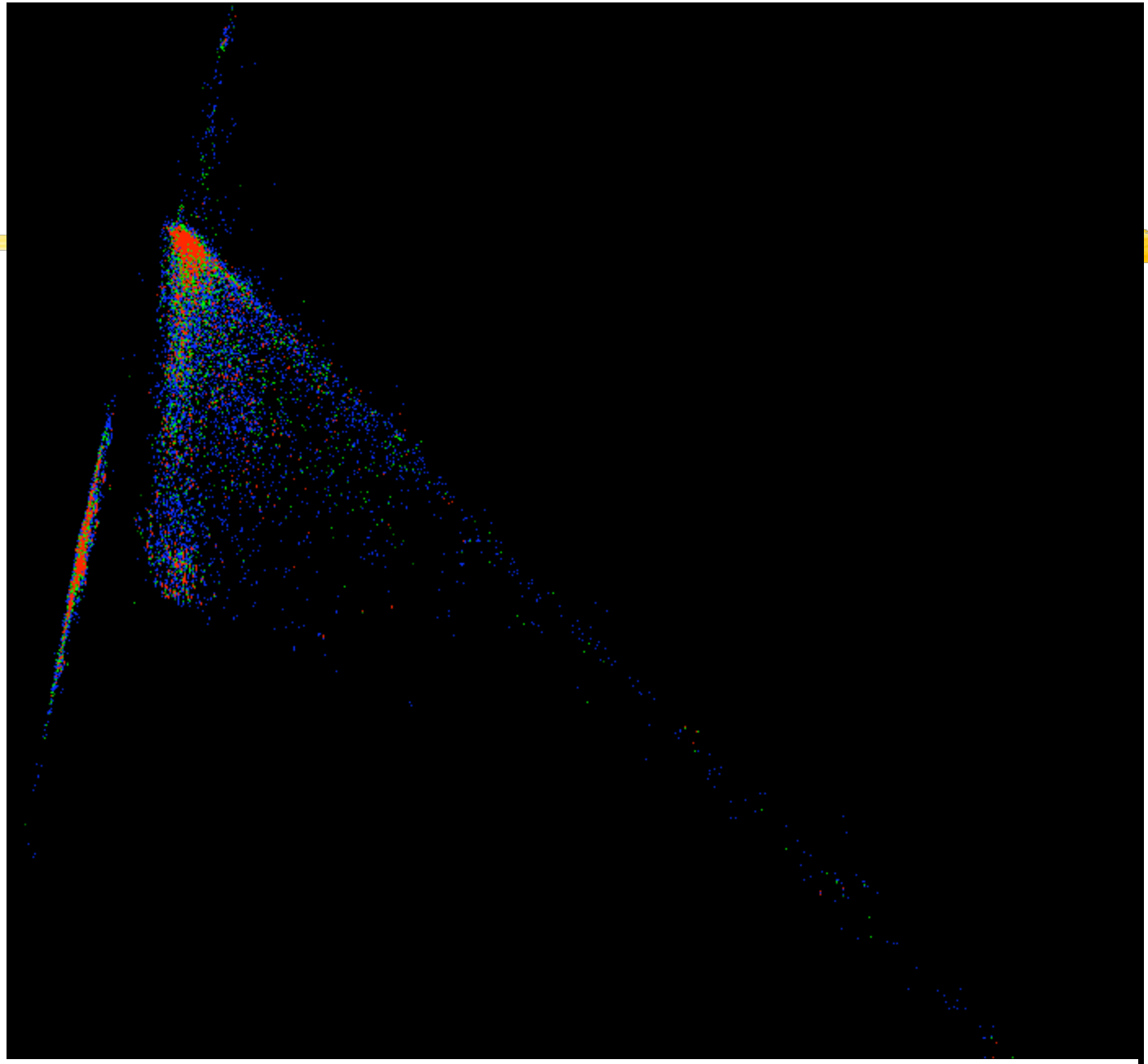


Distributed attack from many sources, same detection - single mountain



Used as instant signature - looking forward and **backward** in time

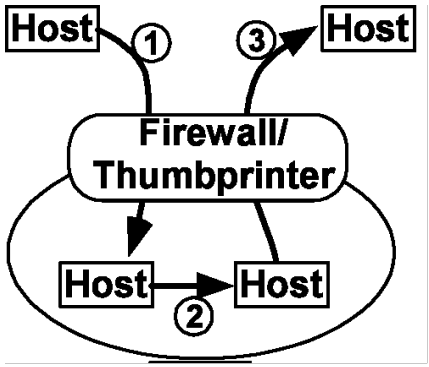




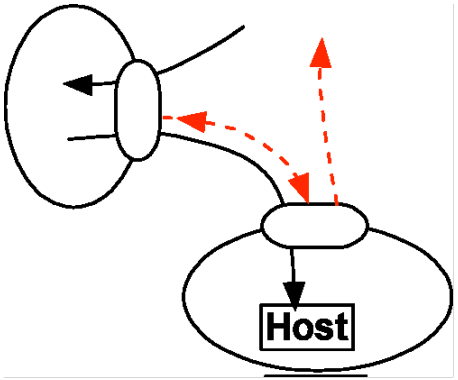
http://www.netsq.com/Projects/NetworkLandscapes/

ThumbprintStream

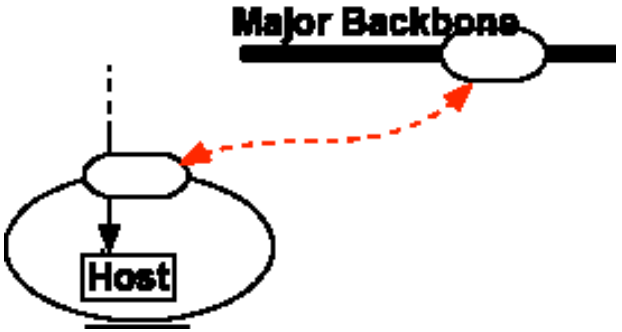
- Seven years in the making



Identifying connection laundering at your own site

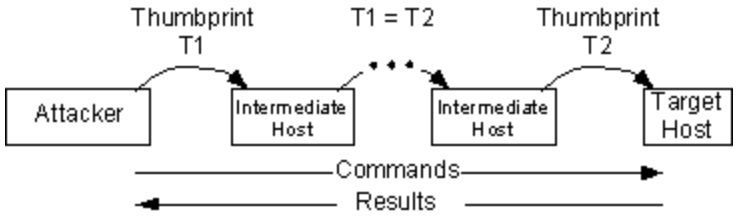


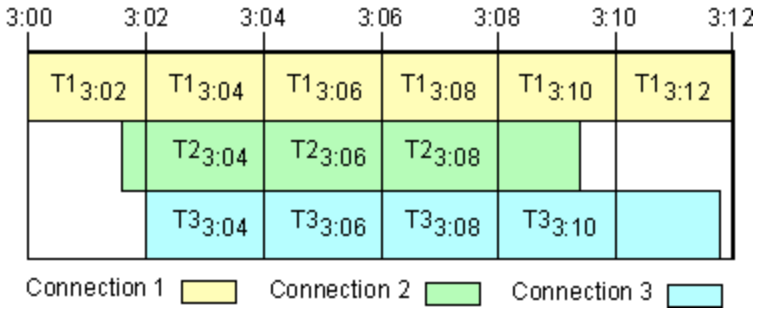
Automated support from intermediate site



Bypassing many intermediate sites with support of thumbprinting on a major backbone

Thumbprint Document

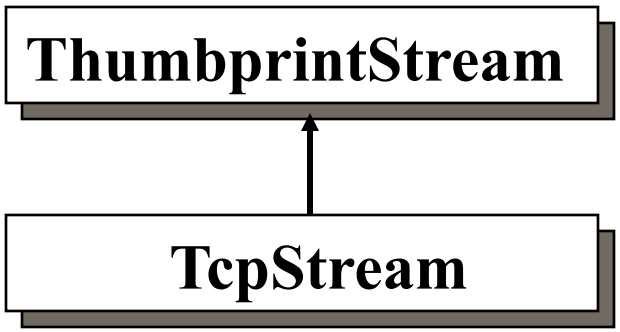


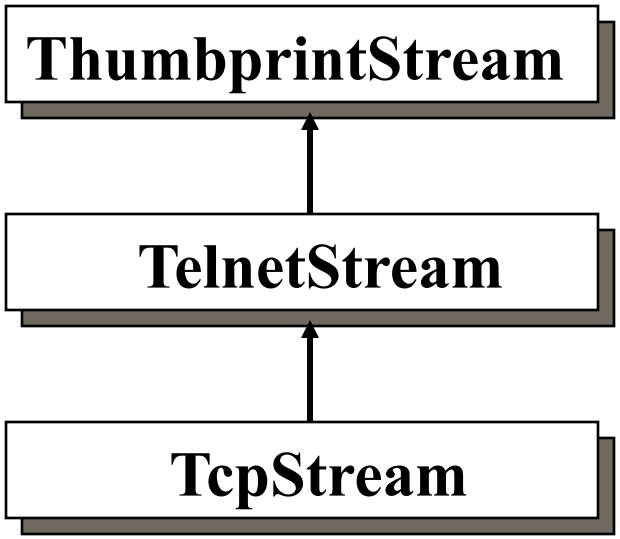


Art of the Document

- Elizabethan plays
  - Similar stories
  - Similar writing structure
- Web traffic
  - client data, server data, mixture
- Login traffic
  - Same extended session
  - Same style, looking for the same person

Problems with Telnet

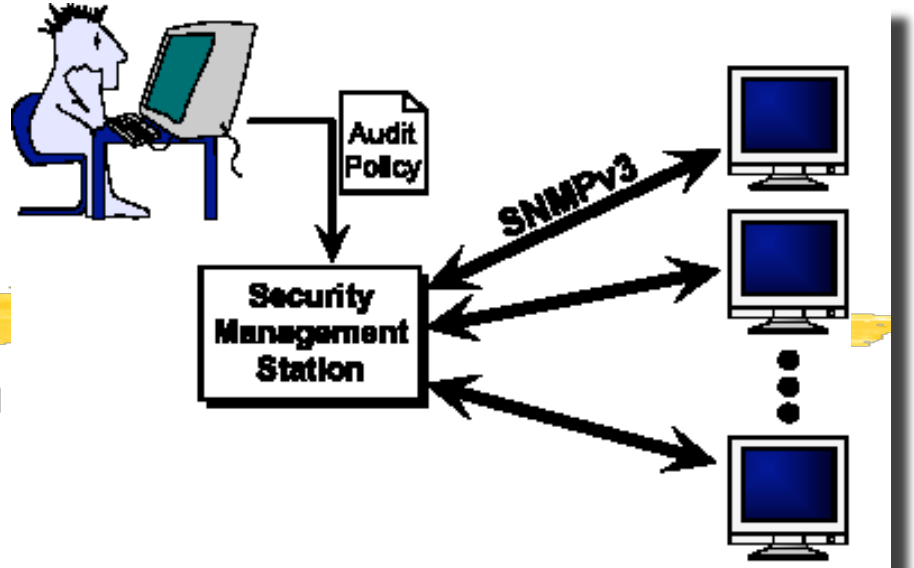




18	17	20	20
37	23	26	26
35	26	15	15

WWW Example

Linux Audit



- Part of the Audit Workbench
- Motivations
  - What happened to C2 by '92?
  - Need audit trail on Linux
  - Linux source of many penetrations
  - Research platform
- Goals
  - Preserve operations experience and development work
  - Almost configuration compatible with Solaris BSM
  - Simplify Audit Workbench project

http://www.netsq.com/Projects/LinuxAudit/

http://soledad.cs.ucdavis.edu/