Issues of Big Data Sharing in a Global Science Collaboration

Is it networking issue?
Or is it a security issue?

Jerry Sobieski
Chief Research Officer
NORDUnet

Presented to the Internet2 Global Summit 2017

Sharing large data assets

- Redistributing and correlating large data has two major challenges:
  - Moving large data sets across large physical distances -> The classic network capacity/performance issue (This assumes the two locations are trusted)
  - Secured access to information – once outside a secure perimeter, there is no longer effective control of access to that info. (i.e. how do we “trust” remote locations?)

- Moving the algorithm to the data:
  - Useful where the distributed data sets are already integrated in a single “location”
  - Does not solve the problem of gathering distributed data sets for correlation or other integrated analysis algorithms,

- Exposes the algorithm to potential security breaches
  - Proprietary algorithms may be compromised

Security Issues in data sharing

- Jurisdictional restrictions
  - (E.g. national borders)
- Proprietary restrictions
  - e.g. business policy, IP algorithms
- Privacy restrictions
  - E.g. personal financial info, medical data, etc.
- Trust – but verify
  - Verifiably compliance – can we authorize each access of information? Or limit the use to a single trusted agent?
- Provinence – how do we handle provinence / reproducibility where data access is secured or constrained?

The virtualized world

- “Virtualization” poses important challenges
  - The physical location of information is no longer determined
  - What constitutes a secure (trusted) perimeter in virtual service environments?
- “Cloud” services have not solved the security problem:
  - We can store encrypted data
  - We can transport encrypted data
  - We cannot [yet?] compute on encrypted data (homomorphic computing)
  - This exposes data in the clear

Key open questions:

- Can we *verifiably* secure computational processes short of physical secure perimeters?
  - Security thru obscurity? Distributed computation, interchangable algorithmic components,
  - Who verifies and signs “trusted” code – can we trust them? -> trusted security services who’s business value proposition is their reliability in terms of security analysis of components.
  - Homomorphic (encrypted) computing?

- We can authorize access to information, but having authorized access to some agent, we lose control over the information because that info is now in the clear...
  - Can we encrypt and “sign” data in such a way that only authorized agent(s) can interpret the data and make use of it?