Provenance Data Infrastructure Building Blocks

Zachary G. Ives
University of Pennsylvania

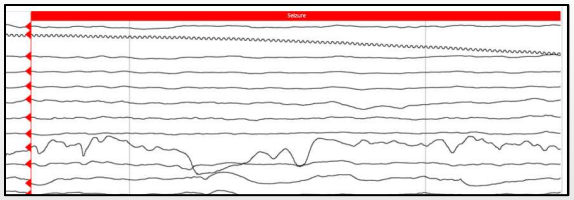
Susan B. Davidson, Sampath Kannan, and Val Tannen (CS)
With Brian Litt and Junhyong Kim (Neuroscience, Biology)

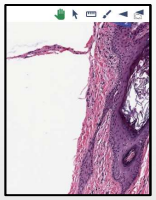
Data Provenance: Data → **Trusted, Reusable Data**

How / when we can **reuse data** across studies?

- Different fields, formats, population, initial conditions

How to scale-up human-annotated **"gold standard" data**?





- Combine **annotations** across expertise + skill levels?

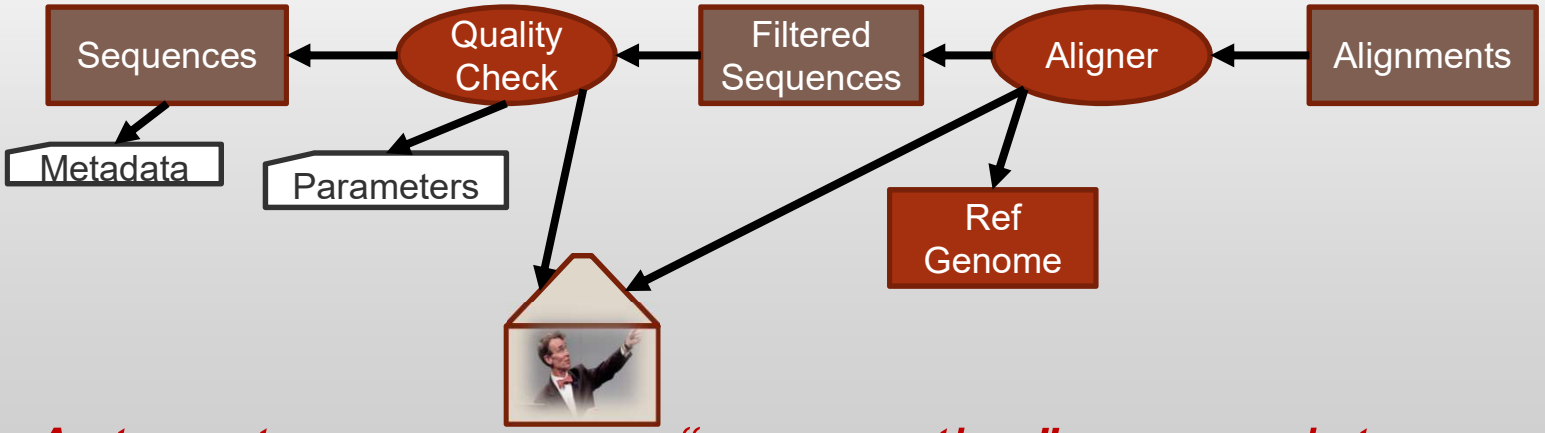
Connect **provenance of** data / annotations, **consensus**, and **how much to trust** contributors / contributions

- Across data modalities (strings, arrays, images, timeseries…)

Provenance is More than Metadata Fields!

Let's consider a computation:

- **Entities** - data objects, typically files
- **Activities** - program executions, processing steps
- **Agents** - users, tools invoked on behalf of users



Automate provenance "propagation" across data modalities, look at provenance "network"

Our Approach

- Make provenance **logging "invisible"** – it should "naturally happen" as part of running a script, a piece of Matlab code, …

- Enable **reasoning about common operations**
  - Filters, zoom, project, … and how they affect data
- Support **reasoning about annotations and agreement**
- Build **trust policy language** based on the above

Our Project: A Staged Effort

- Stage 1: provenance collection, modeling for many modalities and applications. "Gather use cases."
  - pennprovenance.net: storage and logging of computation!
  - "Gray box provenance" model – we can "see into" data components, semantics of operations (matrix, string, …)
- Stage 2: measures of overlap → distributions → consensus. "Formalize measures for use cases."
- Stage 3: trustworthiness and interplay with consensus. "Incorporate trust into application."

Early Artifacts and Results

- At pennprovenance.net: seamless provenance infrastructure
  - **Provenance Tracker** – Mac/Linux/Windows event logging to capture provenance from scripts, program executions
  - **Provenance Storage** – collect data provenance in a central repository (PROV-DM data model)
- Papers: provenance for linear algebra, provenance summarization

- Early scientific impact: Supporting high-throughput gene seq in Junhong Kim's lab @ Penn
  - Provenance for byte sequences and strings with similarity
  - Experiments with TrustRank on data from the lab – what are the most impactful sequences, tools, …?