Confession #1

Provenance and Causality

Provenance-based Belief

Adriane Chapman, Barbara Blaustein
and Chris Elsaesser

Confession #2

Causality Arguments

MITRE

Approved for Public Release; Distribution Unlimited - 09-5315

Motivation

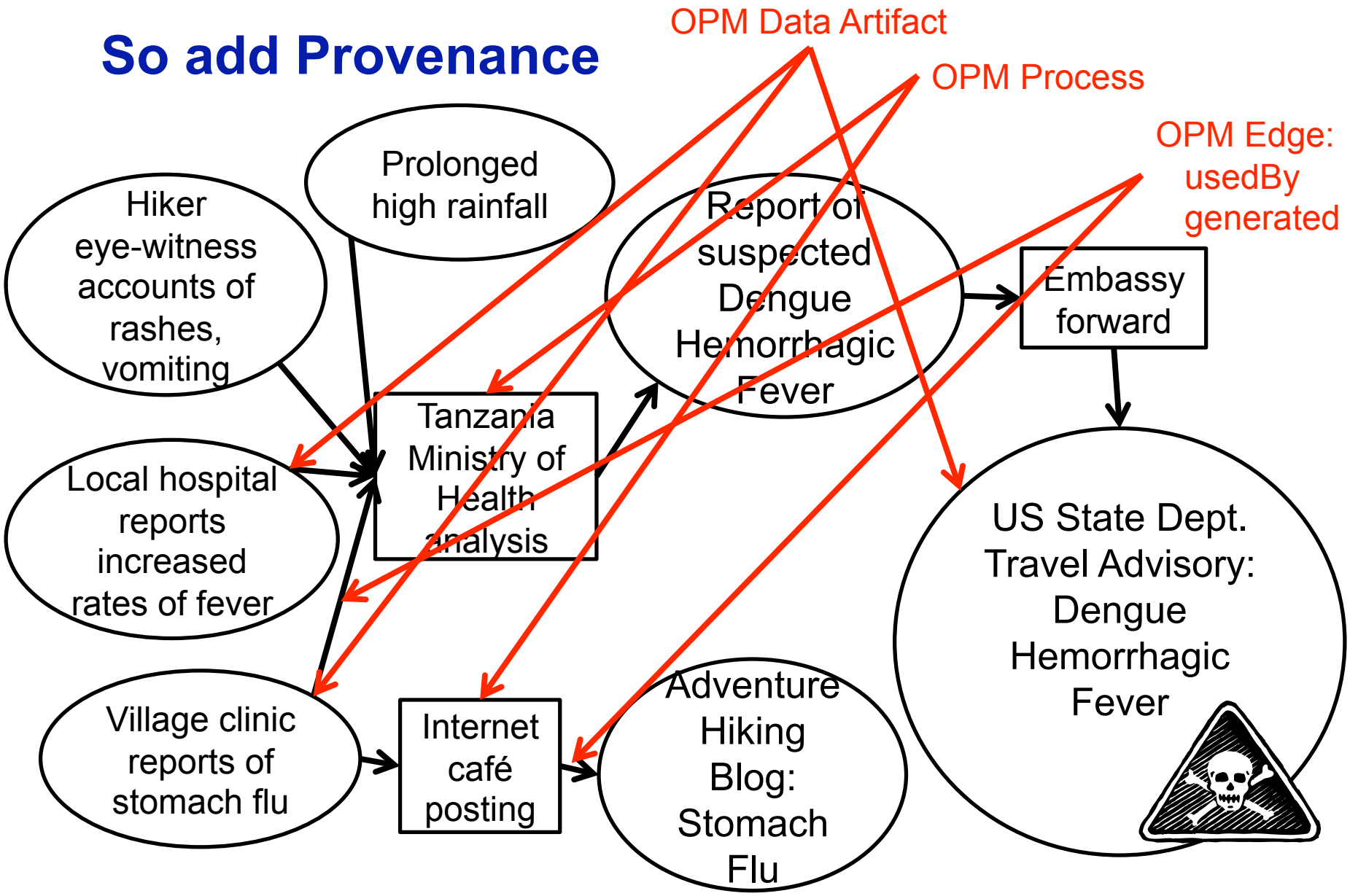
US State Dept. Travel Advisory: Dengue Hemorrhagic Fever

Adventure Hiking Blog: Stomach Flu

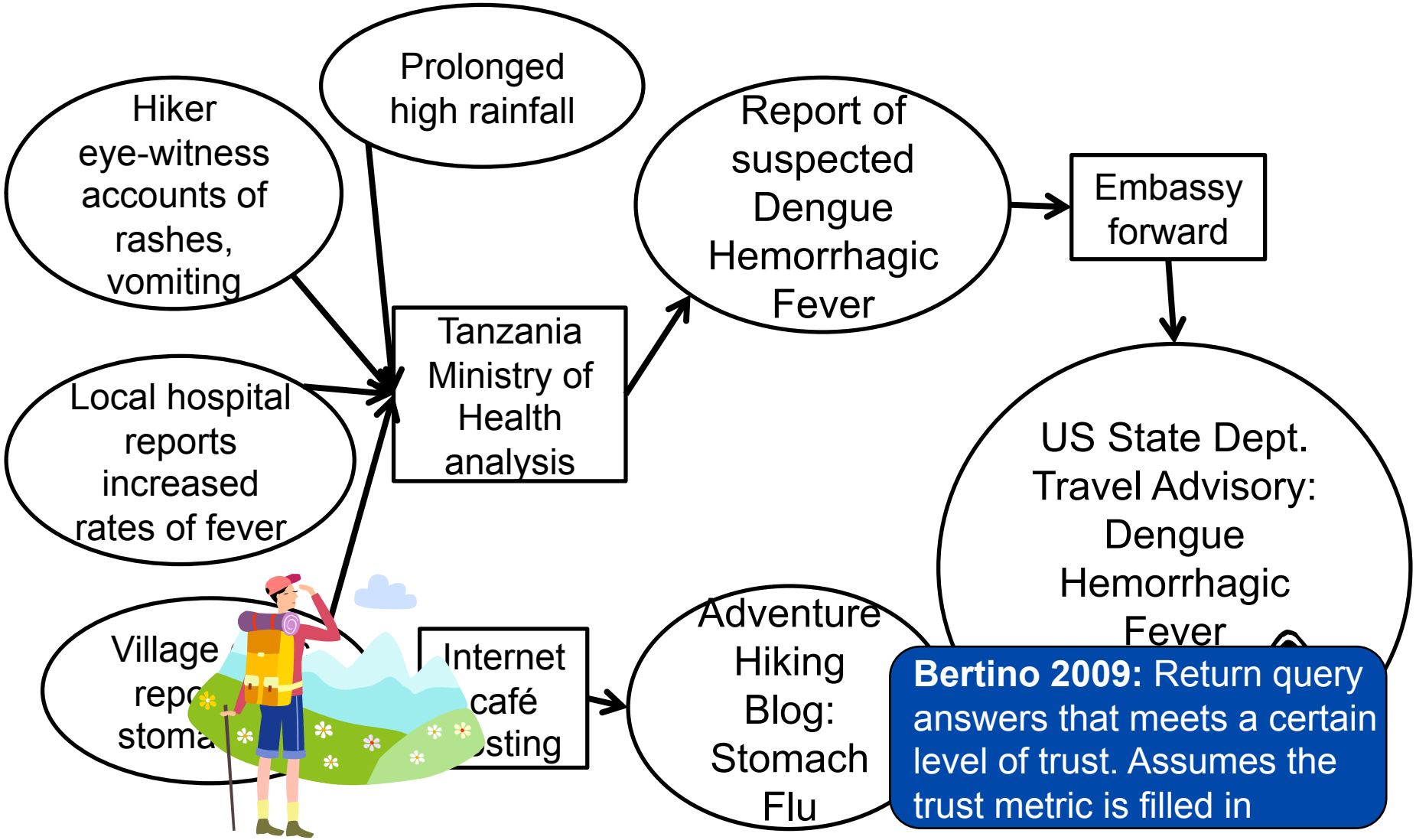
"Provenance can be used to determine how much to trust the data"

Provenance metadata is essential for data consumers to assess the authoritativeness and trustworthiness of the data asset – IA Metadata COI

So add Provenance



Now what do we do?



MITRE

Computing a Trust Value – Current Work

- **Prat and Madnick, 2008**
  - Requires "reasonableness of data" evaluation
- **Gil and Artz, 2007**
  - Use data quality metrics
- **de Keijzer and van Keulen, 2007**
  - Looks at the uncertainty of the data
- **Hartig and Zhao, 2009**
  - Timeliness based on data expiry date
- **Becker et. al., 2008**
  - Measures accuracy of data

Rely on information in the data, not the provenance

Calculating Uncertainty in Probabilistic DBs – Current Work

- **Widom et al, 2006**
  - Data values assigned a base probability. Values propagated based on relational manipulations.
- **Gatterbauer and Sucio et. al., 2009**
  - Create all possible worlds based on belief annotations
- **Ives et. al., 2008**
  - Use semi-rings to propagate stated trust annotations

Report probabilistic combinations based on relational algebra manipulations

What we want

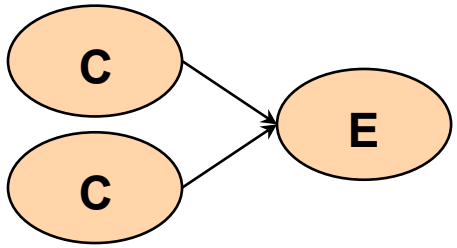
- Provide a mechanism for calculating a trust value that does not require access to the data itself
  - Some data are not accessible from the provenance store
    - E.g. The provenance store cites a HUMINT report that you aren't cleared to read
  - Some information cannot be determined until well after an event
    - E.g. The accuracy of the estimation for "2010 corn eaten" cannot be assessed until 2011.
- Provide a mechanism for calculating a trust value for processes other than the relational algebra
  - Some processes are worse than others
    - E.g. a broken AVERAGE function that forgets to divide by the number of items summed.

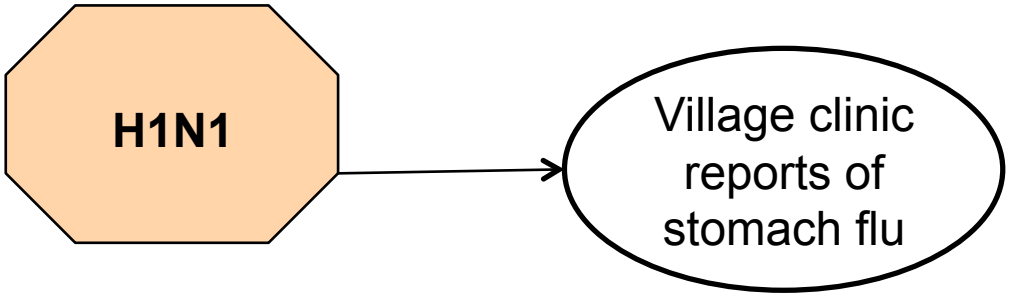
A Brief Introduction to Bayesian Models

- ***Proposition***: a sentence expressing something is T or F
  - E.g. C = there are symptoms of DHF at Kilimanjaro
- ***Belief***: subjective probability that the proposition is true
  - For proposition C, belief = p(C)
- ***Evidence***: a proposition related to another proposition
  - E.g. p(C|E) = there are symptoms of DHF at Kilimanjaro given the State Department Report
- ***Bayes' Rule***: allows calculation of p(C|E)
  - p(C|E) = p(E|C)p(C) / p(E)

Causal Reasoning

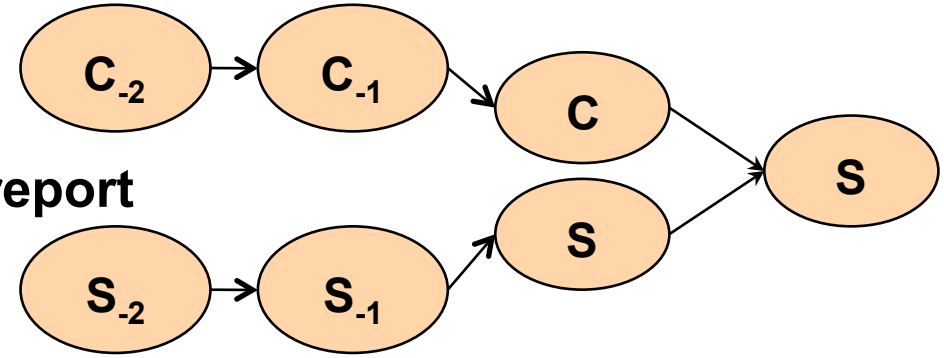
- **Evidence such as a report is *caused*, in the Bayesian sense, by**
  - **The event it reports**
  - **The source that produced the reporting**

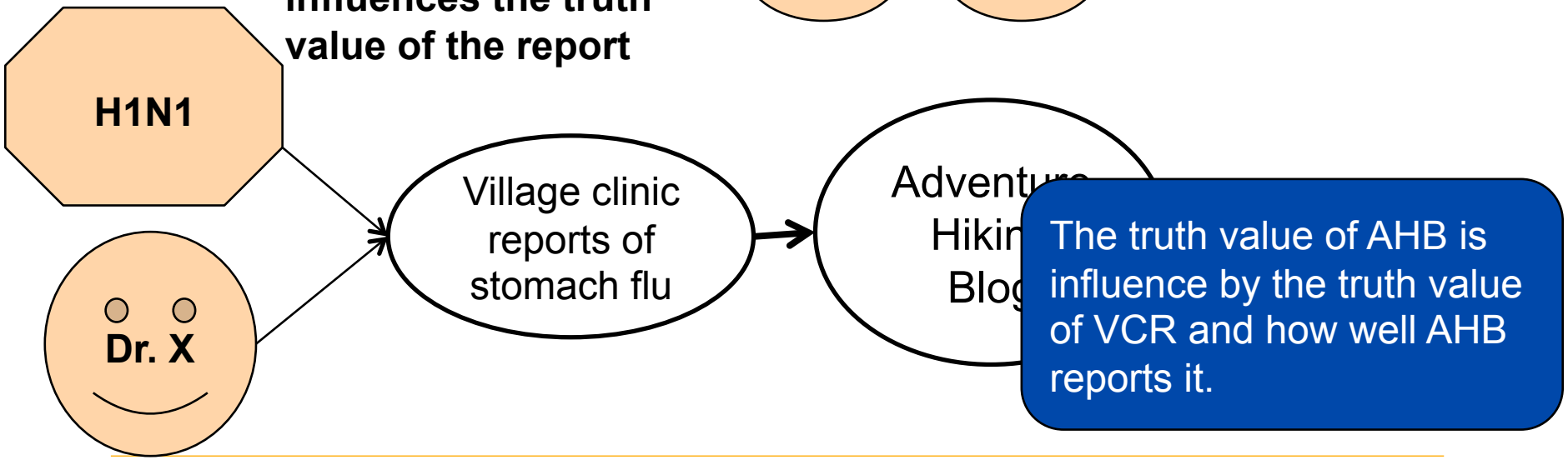




Integrating causal reasoning and provenance

- **Using this formulation, provenance can**
  - **be integrated with causal models for inference:**
  - **account for preceding data manipulations**
  - **Dr. X doesn't cause flu report**
    - **His provenance influences the truth value of the report**

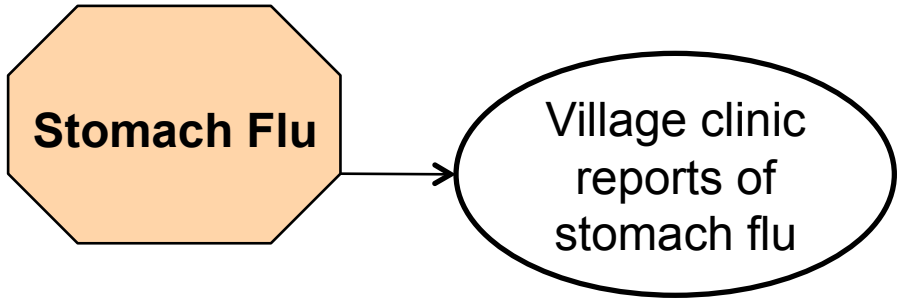




The truth value of AHB is influence by the truth value of VCR and how well AHB reports it.

MITRE

Generating Conditional Probability Tables



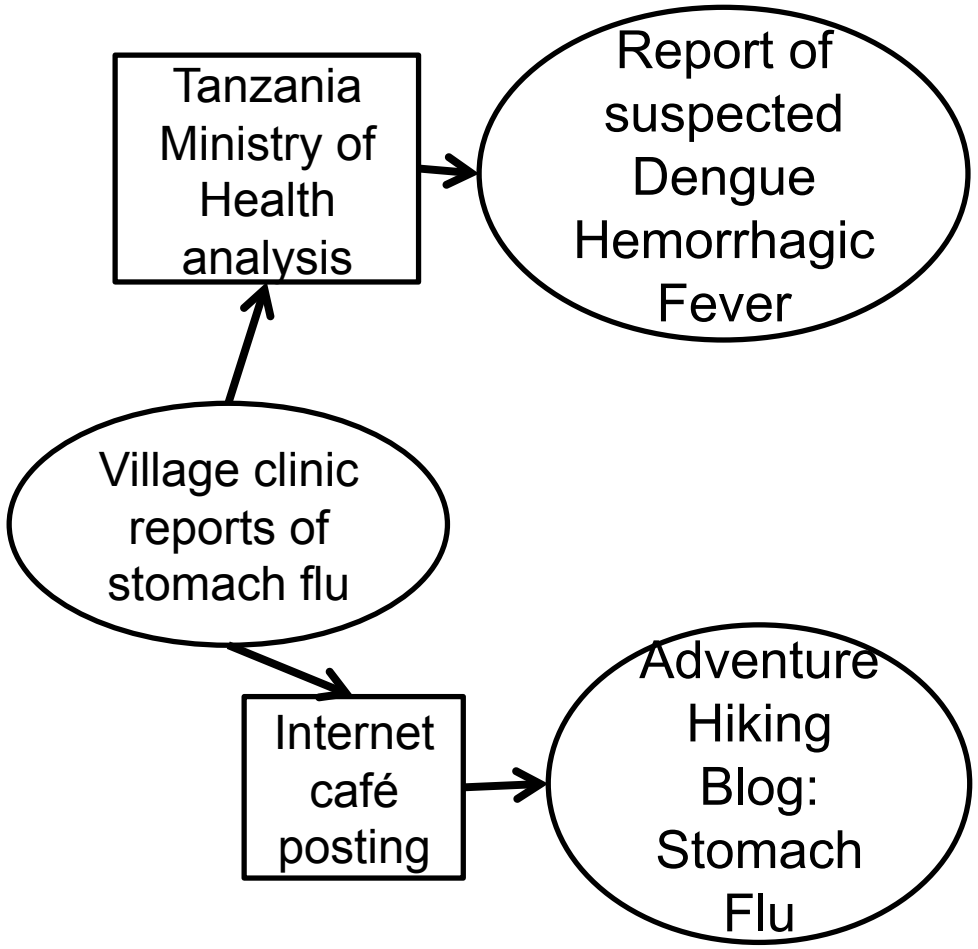
p(Village Report | Stomach Flu)

SF \ VR	R says flu	R no flu
true	.8	.2
false	.4	.6

- Ask an expert on the accuracy for each source
- Learn the accuracy values over time
  - How often does the WHO report an outbreak (when there is an outbreak and when there isn't an outbreak)?
  - Requires knowledge of results
- Use the provenance store to determine conditional probabilities of shared sources

Independence and Single Sources

- **Provenance will show when there are shared sources**
- **Modelling provenance with a causal model will allow us to propogate beliefs based on shared and independent sources**



Processes

- **Processes have their own conditional probability tables that reflect how accurately they manipulate the information.**

Default Process

E1 \E2	T	F
T	.9	.1
F	.1	.9

Computer Copy

SF \ VR	R says flu	R no flu
true	1	0
false	0	1

Bad Intern Copy

SF \ VR	R says flu	R says no
true	.6	.4
false	.4	.6

Conclusions and Future Work

- Taking a causal view of provenance (in some cases) allows computation of trustworthiness of data
- Trust can be computed using only provenance (no data)
- Implementing this into a real system for further evaluation

Bertino 2009: Return query answers that meets a certain level of trust. Assumes the trust metric is filled in