Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Integration of Experimental and Textual Data for Biosurveillance

BOBBIE-JO WEBB-ROBERTSON
BJ@PNNL.GOV

August 28, 2012

Biosurveillance Conference, Washington, DC

Motivation

Investigators/analysts need “confidence” metrics to enable justified and rapid decision making.

Motivation

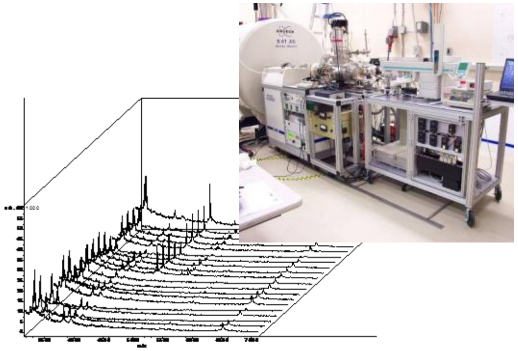
Investigators/analysts need "confidence" metrics to enable justified and rapid decision making.

Sample

Motivation

Investigators/analysts need "confidence" metrics to enable justified and rapid decision making.

Sample

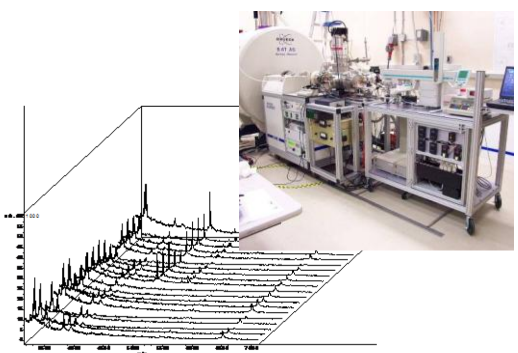


Experimental Signatures

Motivation

Investigators/analysts need “confidence” metrics to enable justified and rapid decision making.

Sample

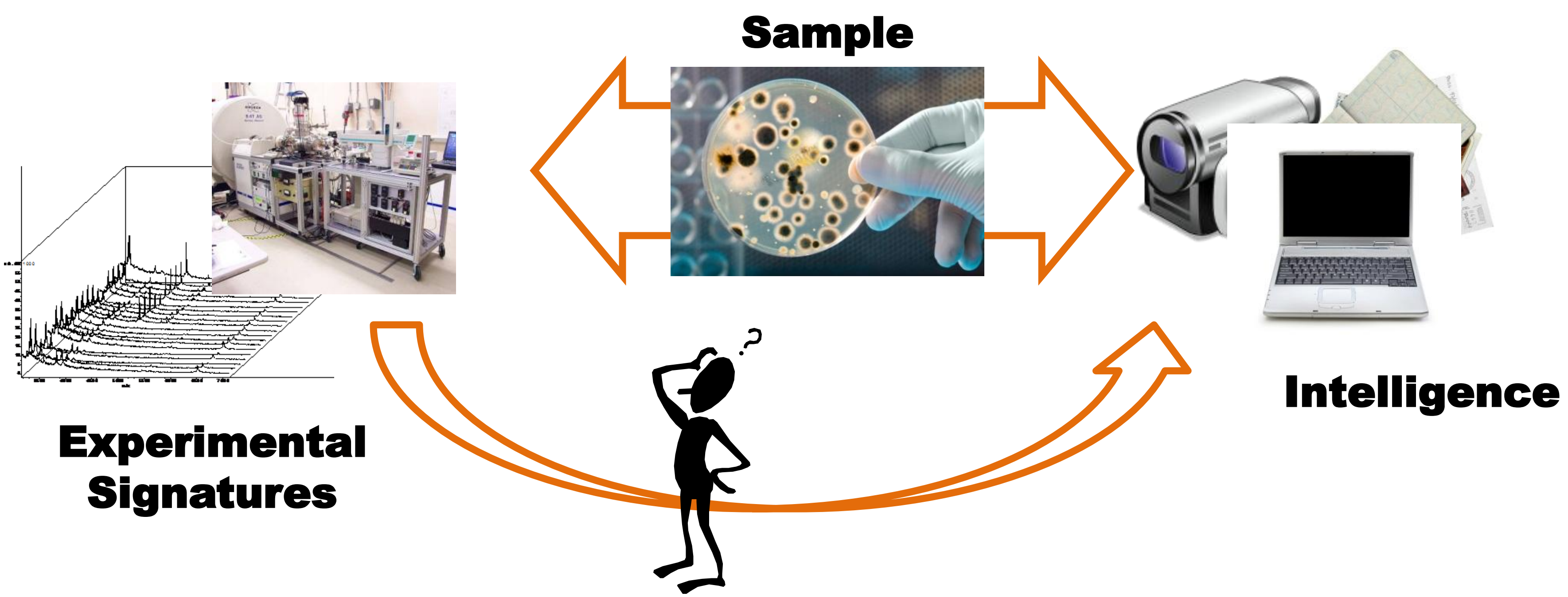


Experimental Signatures

Intelligence

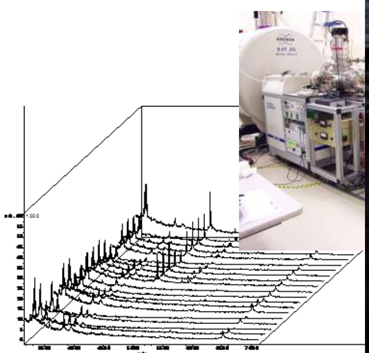
Motivation

Investigators/analysts need "confidence" metrics to enable justified and rapid decision making.



Motivation

Investigators/analysts need "confidence" metrics to enable justified and rapid decision making.



Experiment
Signatures

Intelligence

Integration Problem

Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

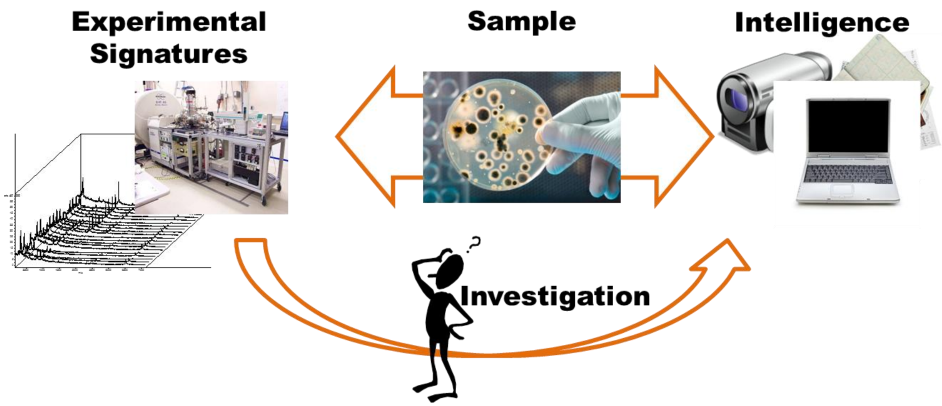
How do we tie together the “experimental” and “intelligence” signatures to help the analyst/investigator?

Integration Problem

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

How do we tie together the "experimental" and "intelligence" signatures to help the analyst/investigator?

- Challenge
  - Research is compartmentalized into domains
  - Statistical confidence metrics from multiple sources of evidence have not been well defined for bioforensics/ biosurveillance



Approach – Bayesian networks

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Bayesian Statistics Naturally fits forensic and surveillance type problems

Outcome is conditionally related to the sources of evidence

Approach – Bayesian networks

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle *Since 1965*

Bayesian Statistics Naturally fits forensic and surveillance type problems

Outcome is conditionally related to the sources of evidence

- **Bayes theorem**

Posterior ***Likelihood*** ***Prior***

$$\underbrace{P(O|E)}_{\text{Posterior}} = \frac{\overbrace{P(E|O)}^{\text{Likelihood}}\overbrace{P(O)}^{\text{Prior}}}{P(E)}$$

Approach – Bayesian networks

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

Bayesian Statistics Naturally fits forensic and surveillance type problems

Outcome is conditionally related to the sources of evidence

- **Bayes theorem**

Posterior ***Likelihood*** ***Prior***

$$\underbrace{P(O\,|\,E)}_{\text{Posterior}} = \frac{\overbrace{P(E\,|\,O)}^{\text{Likelihood}}\overbrace{P(O)}^{\text{Prior}}}{P(E)}$$

Probability that a person become sick with the flu given (*O*) their age (*E*)

Approach – Bayesian networks

Bayesian Statistics Naturally fits forensic and surveillance type problems

Outcome is conditionally related to the sources of evidence

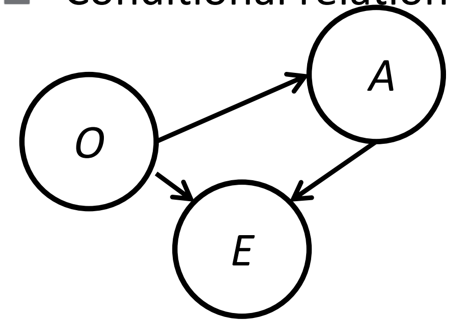
- **Bayes theorem**

Posterior *Likelihood* *Prior*

$$P(O|E) = \frac{P(E|O)P(O)}{P(E)}$$

Probability that a person become sick with the flu given (*O*) their age (*E*)

- **Bayes network**
  - Conditional relationships

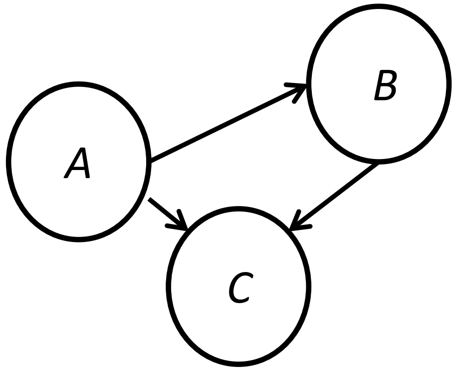


$$P(O|E,G) \prec P(E|G,O)P(G|O)P(O)$$

Probability that a person become sick with the flu given (*O*) their age (*E*) and gender (G)

Approach – Bayesian networks

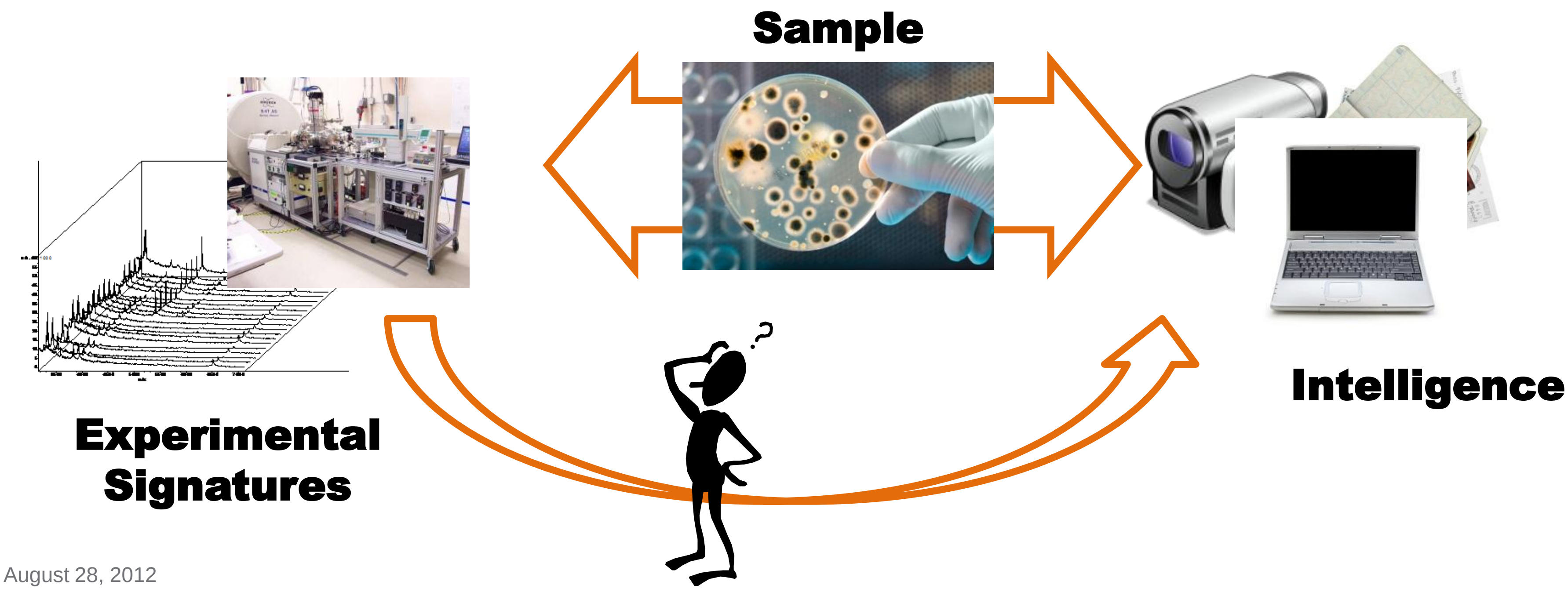
- Allows
  - Integration of heterogeneous data types
  - Multiple complex relationships
  - Incomplete information
- Yields
  - Probabilistic measure of the outcome
  - Probabilistic Interrogation of intermediate nodes



$$P(C \mid A, B) P(B \mid A) P(A)$$

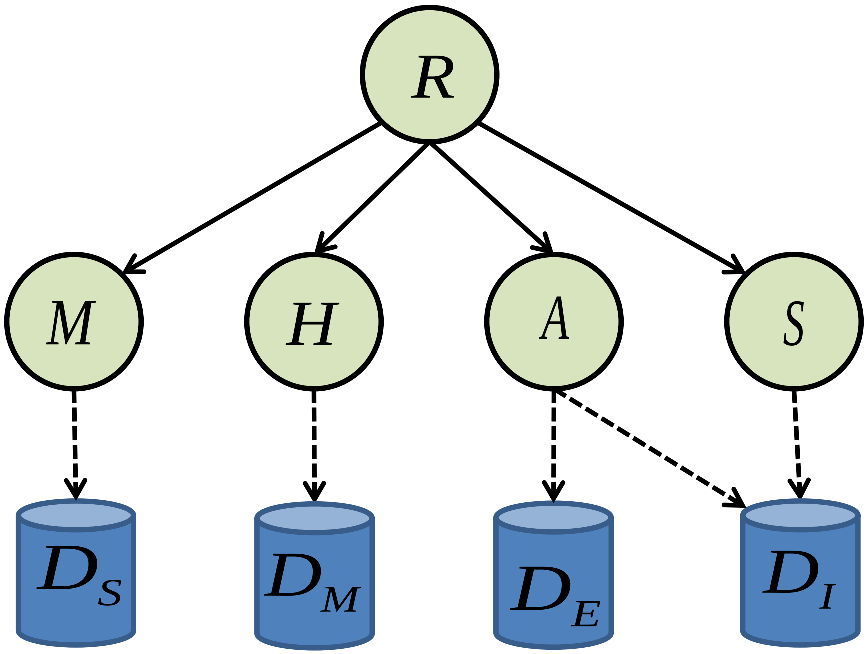
Microbial Forensics

Microorganism-based forensics do not offer investigators "confidence" metrics associated with the sample to gain insight into individuals or places with information pertinent to the investigation.



Approach – Existing Experimentally deriving network (culture media recipe)

Prior work (Jarman et al., 2008) demonstrated that using disparate analytical measurements (*D_S, D_M, D_E, D_I*) of Bacillus spores could yield a predictive model of production environment (*R*)

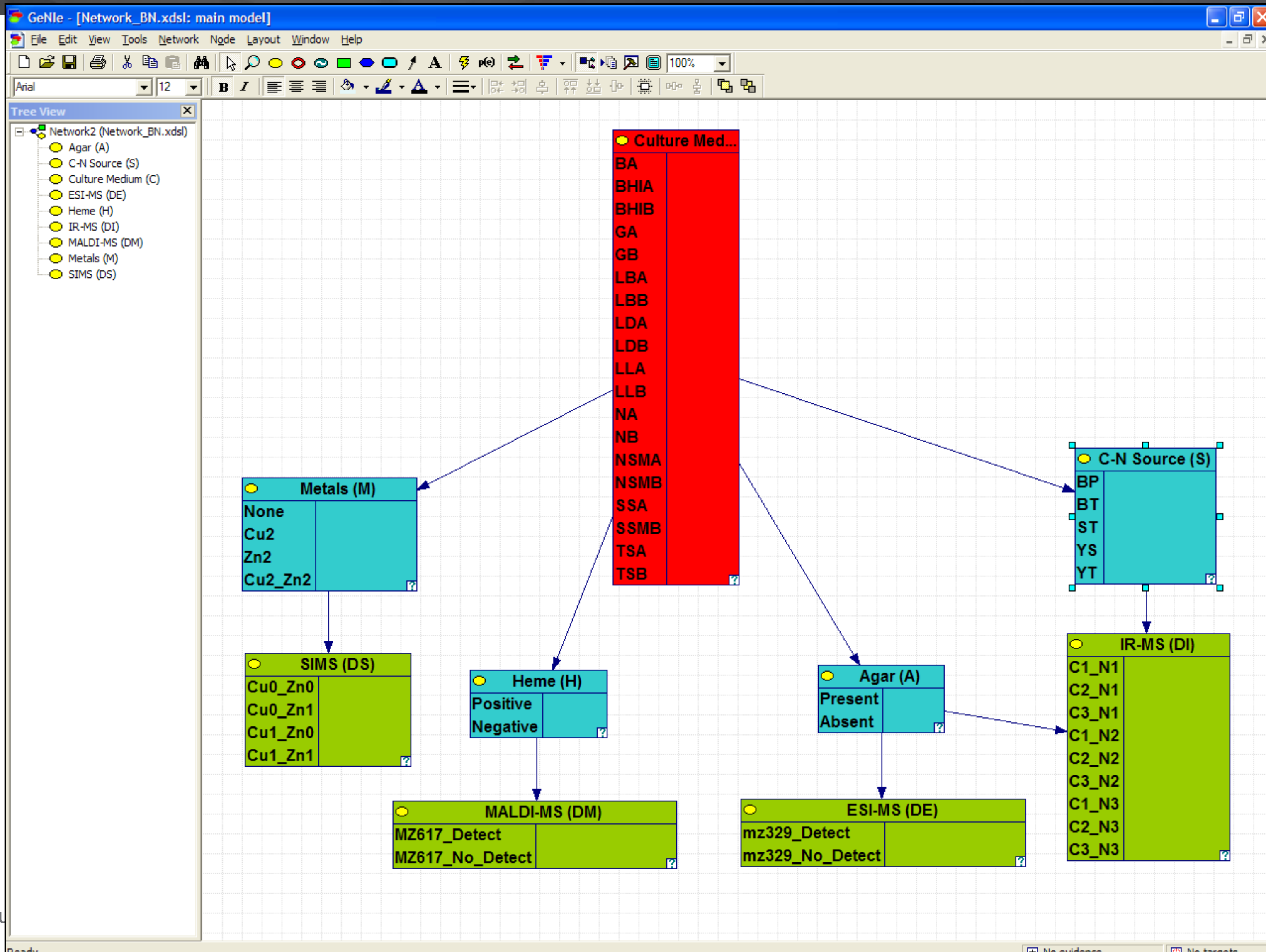


$$P(R \mid D_S, D_M, D_E, D_I)$$

Computed using GeNIe tool for visualization

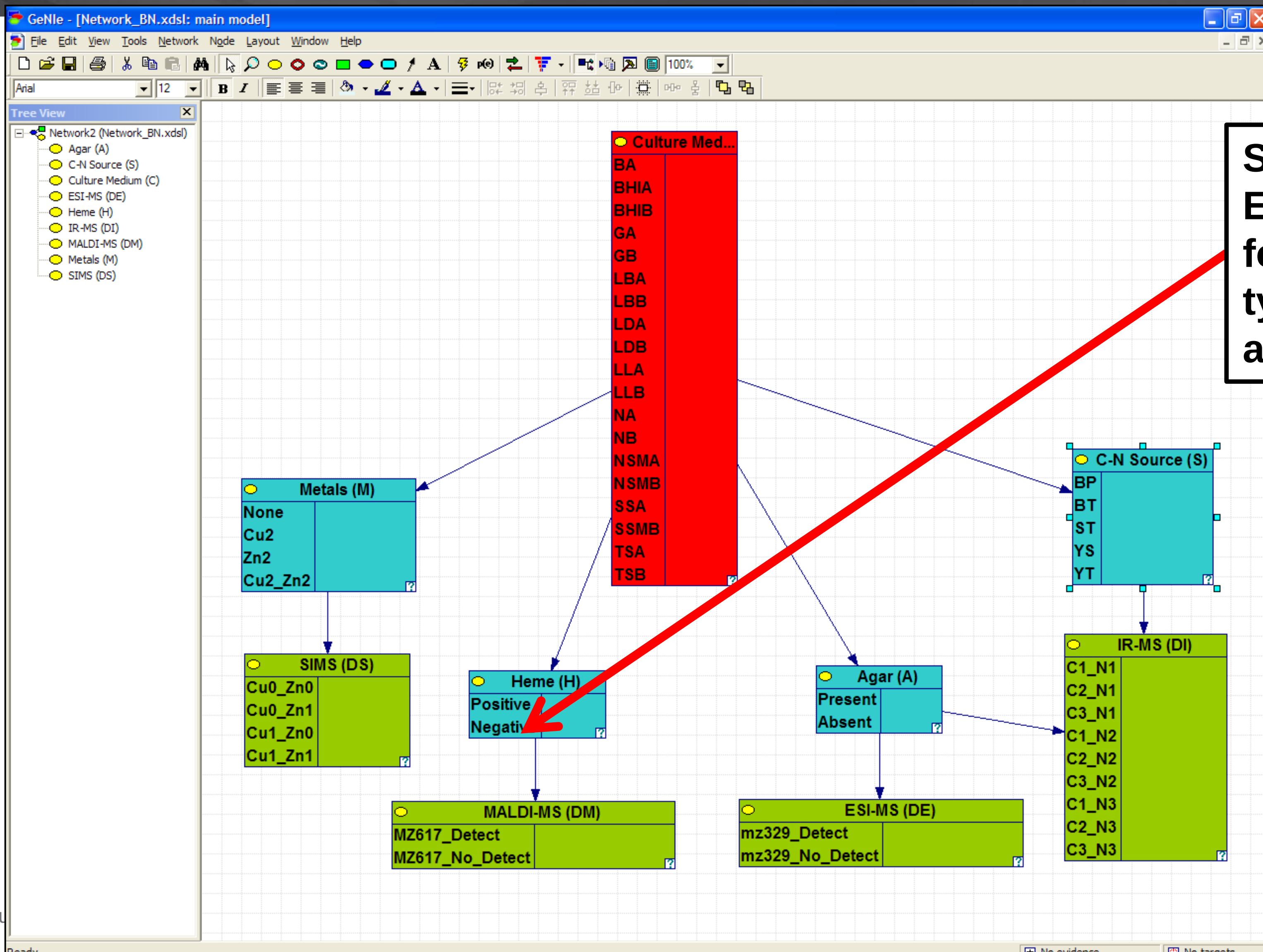
Jarman et al., (2008) AEM

Approach – Existing Experimentally deriving network (culture media recipe)



Approach – Existing Experimentally deriving network (culture media recipe)

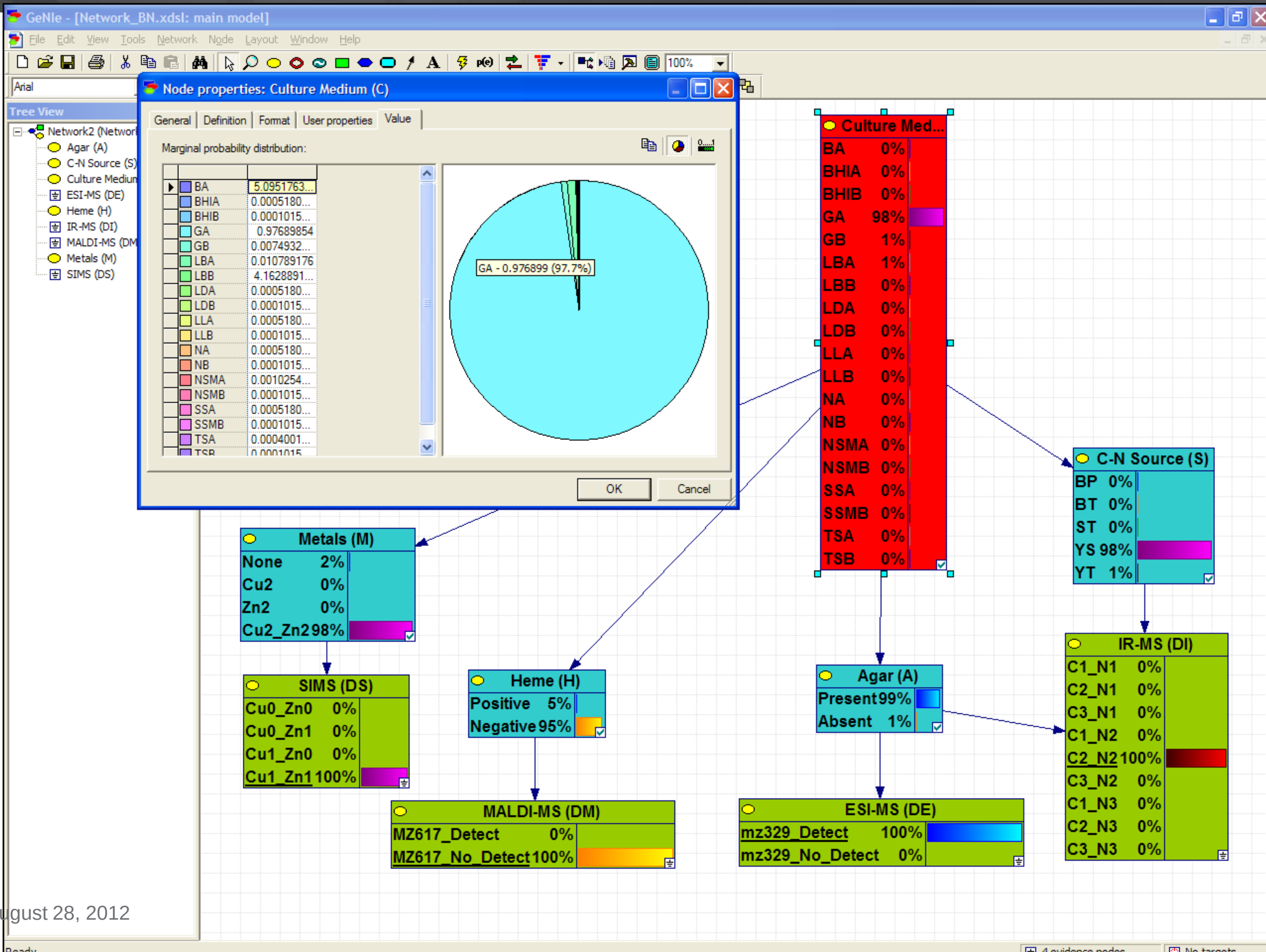
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



Set Evidence for data types available

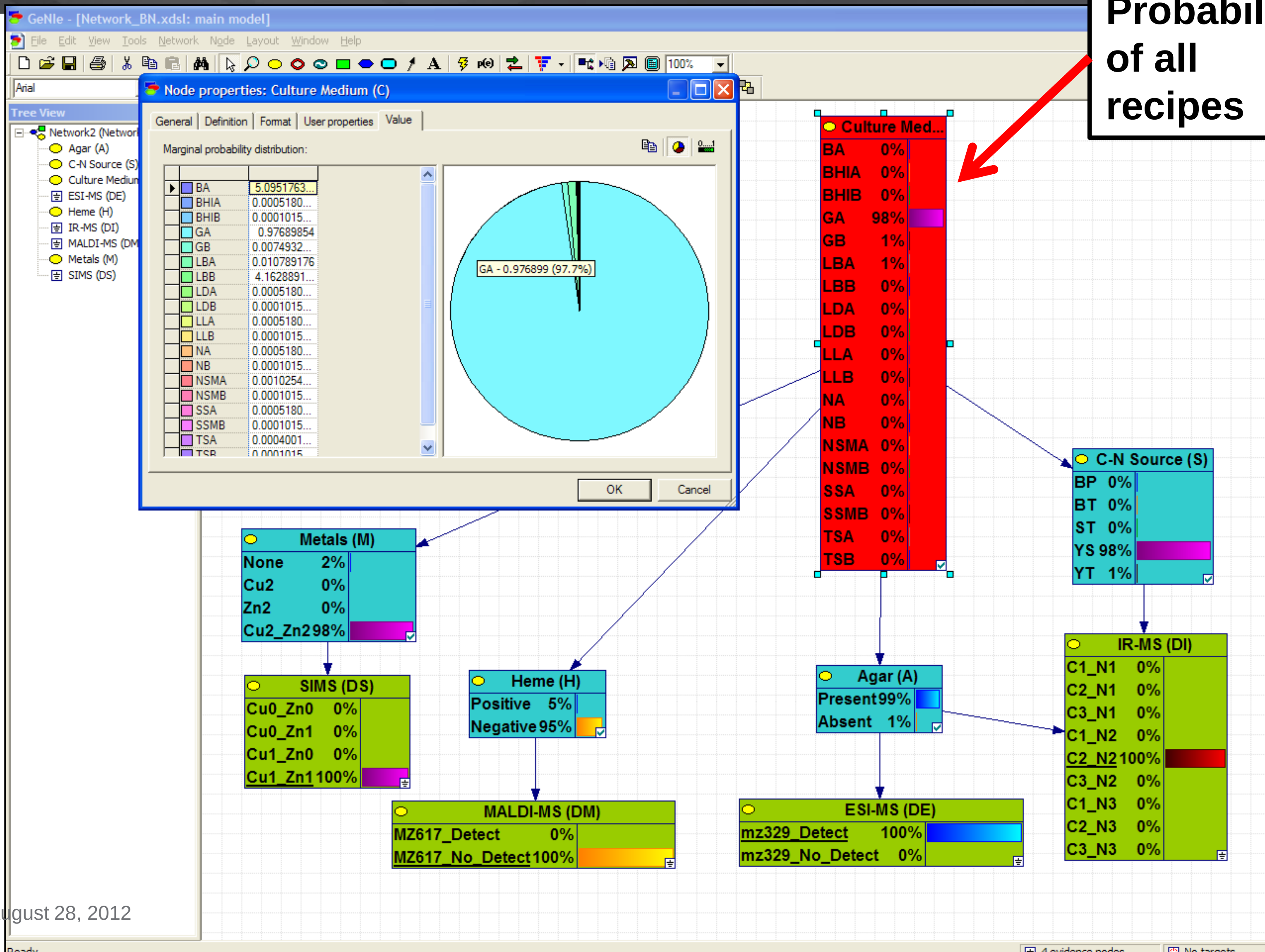
Approach – Existing Experimentally deriving network (culture media recipe)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*



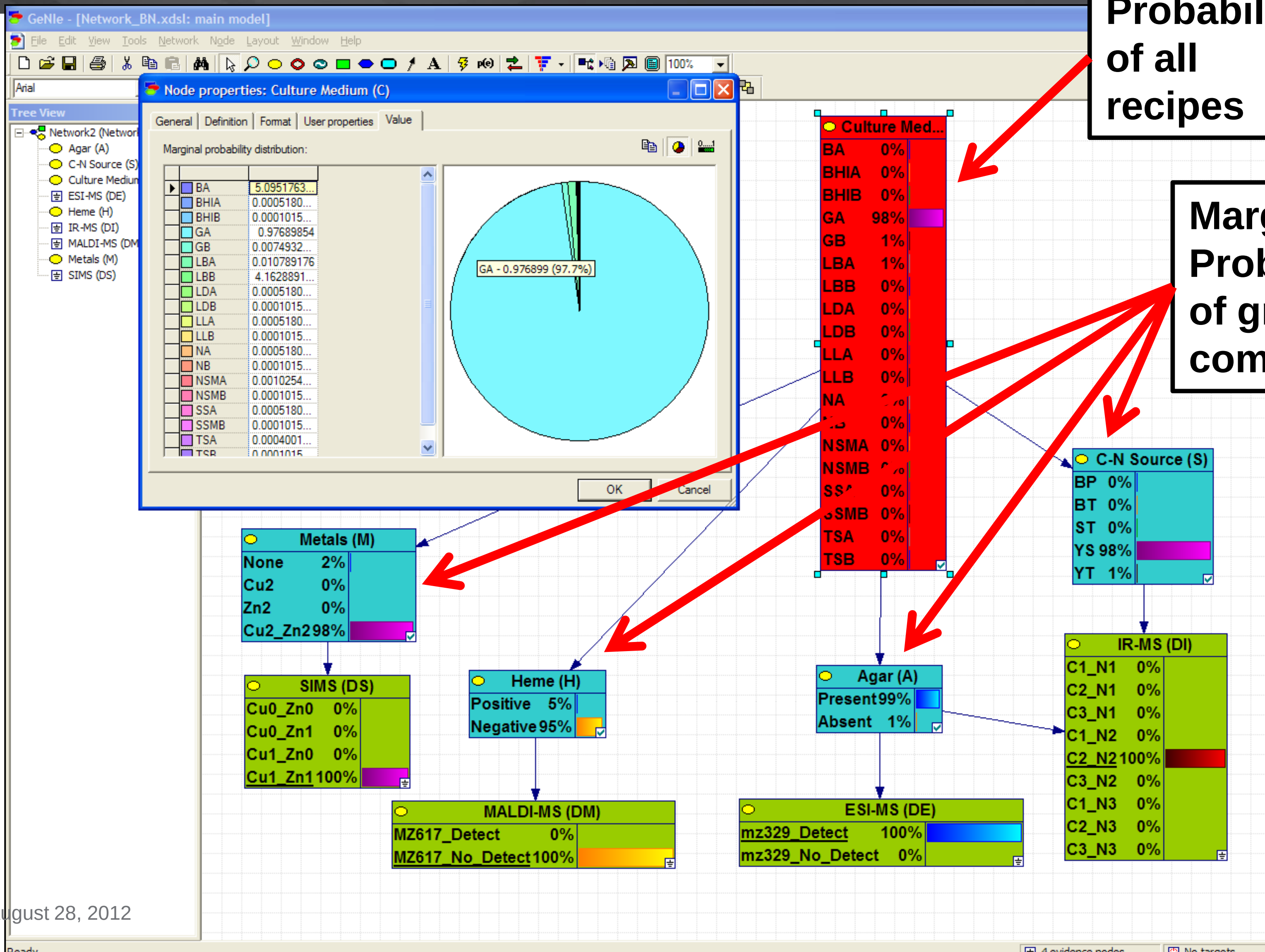
Approach – Existing Experimentally deriving network (culture media recipe)

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*



Probability of all recipes

Approach – Existing Experimentally deriving network (culture media recipe)

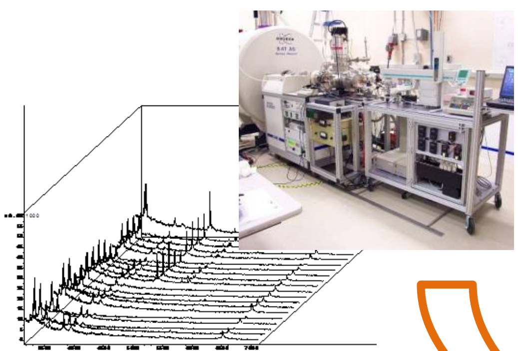


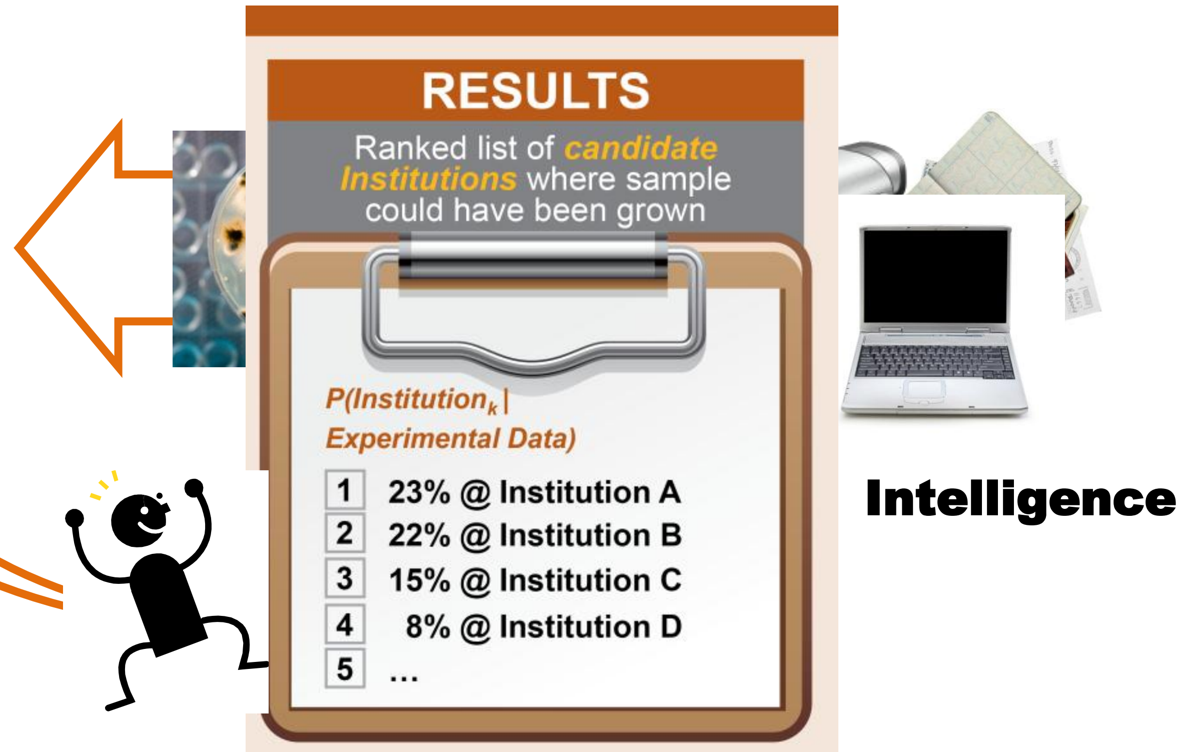
Probability of all recipes

Marginal Probabilities of growth components

Integration Problem – Bayesian network

How can you identify institutions that have experience with the kind of culturing practice pointed to by the experimental evidence?

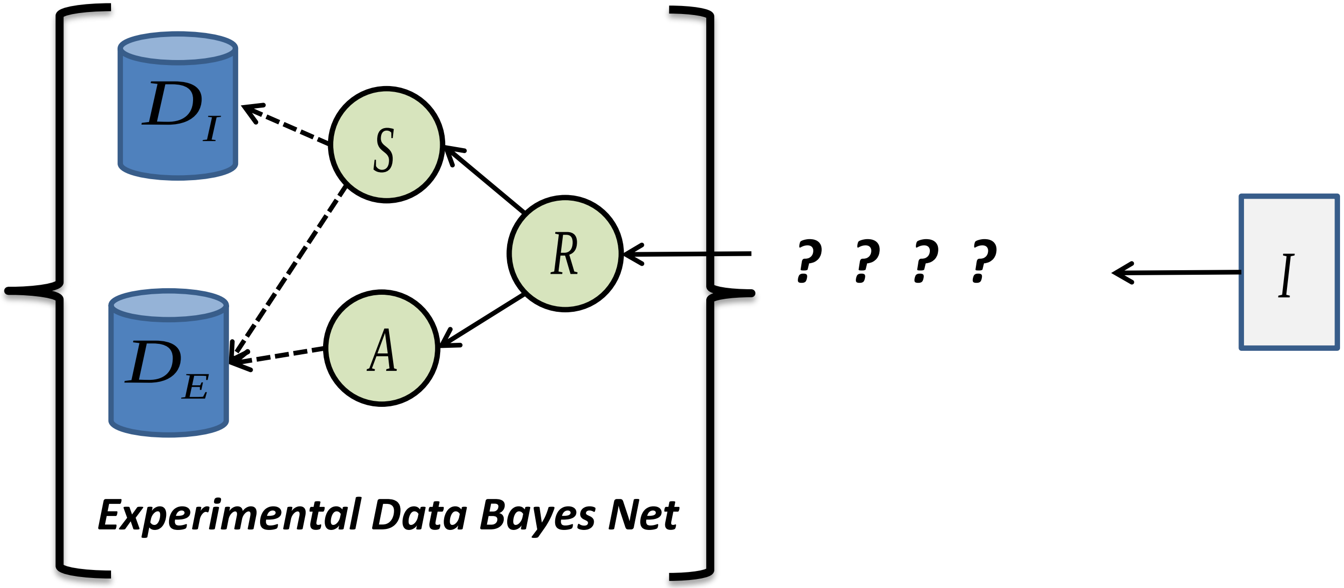




Integration Problem – Building the Bayesian network

How can you identify institutions that have experience with the kind of culturing practice pointed to by the experimental evidence?

$$\boxed{P(I_j \mid D_E, D_I)}$$

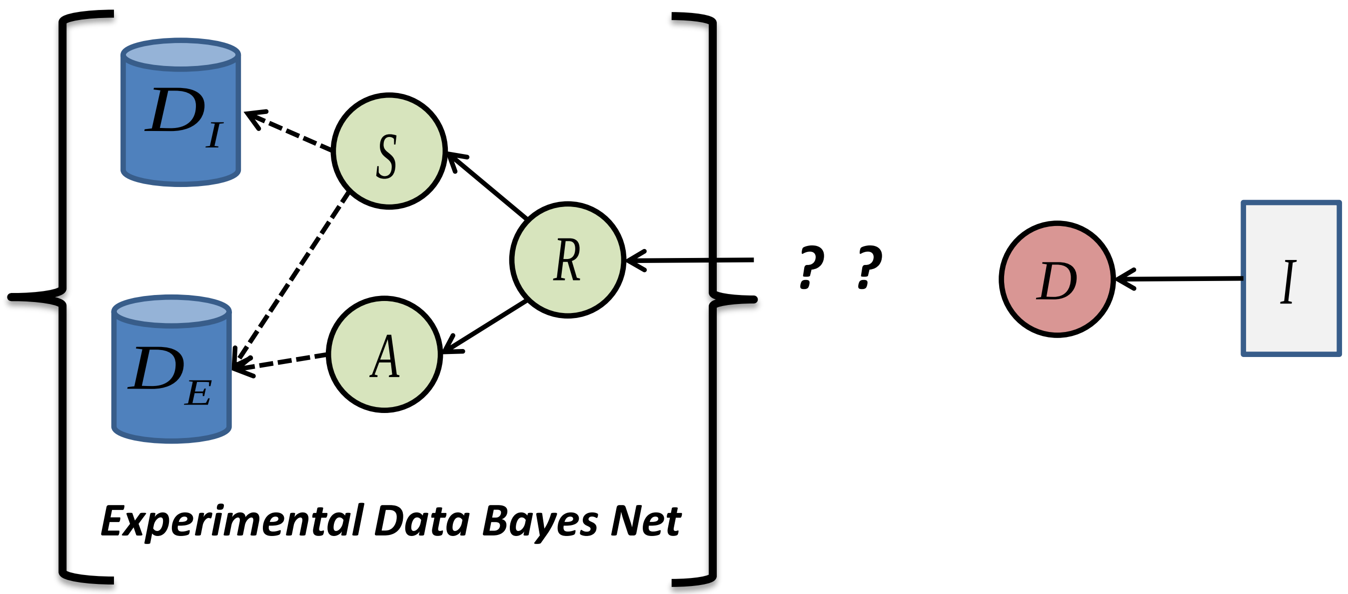


Prediction of culturing recipe from institution is not feasible.

Integration Problem – Building the Bayesian network

How can you identify institutions that have experience with the kind of culturing practice pointed to by the experimental evidence?

$$\boxed{P(I_j \mid D_E, D_I)}$$



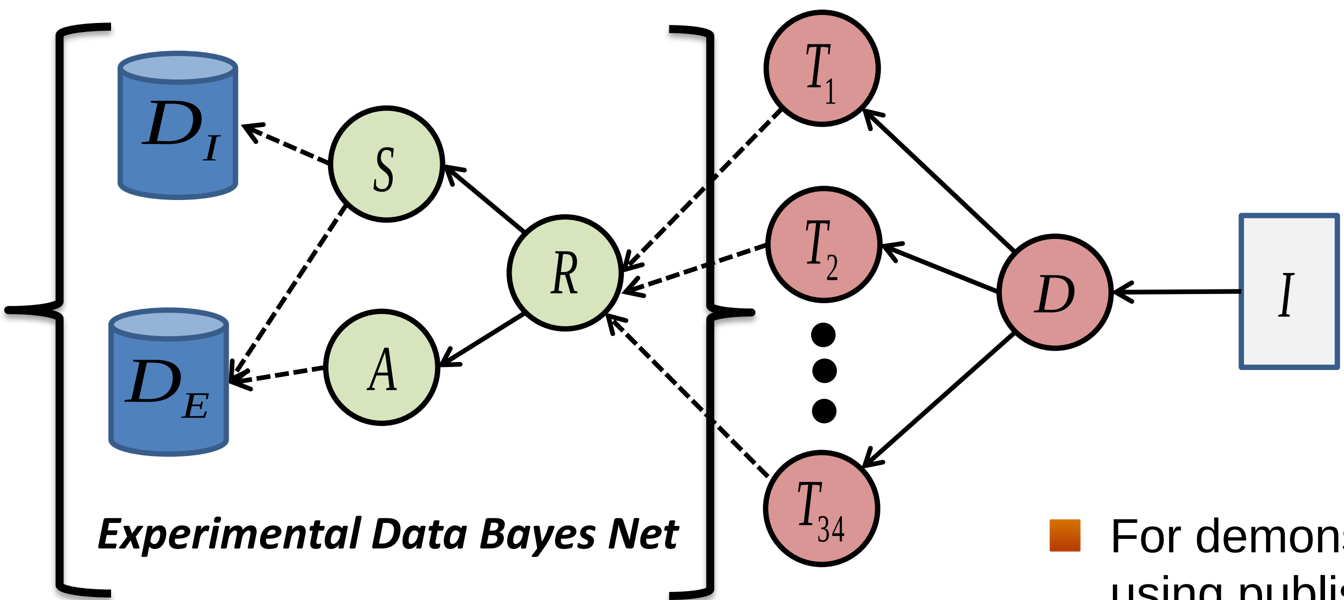
Institutions tie to documents

Challenge to predict recipes directly from document

Integration Problem – Building the Bayesian network

How can you identify institutions that have experience with the kind of culturing practice pointed to by the experimental evidence?

$$\boxed{P(I_j \mid D_E, D_I)}$$



Use automated text scanning (key words)

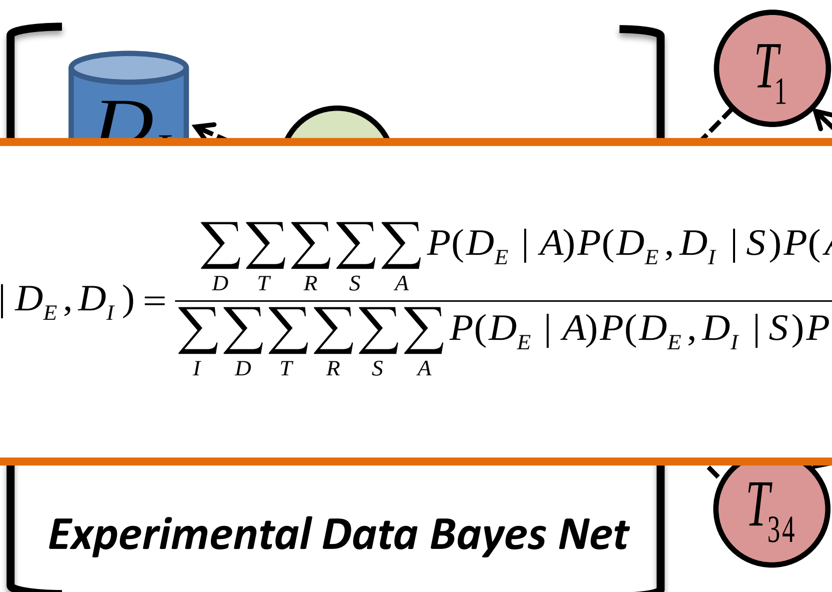
- For demonstration we focus on using published journal articles in the public domain.

Integration Problem – Building the Bayesian network

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

How can you identify institutions that have experience with the kind of culturing practice pointed to by the experimental evidence?

$$\boxed{P(I_j \mid D_E, D_I)}$$



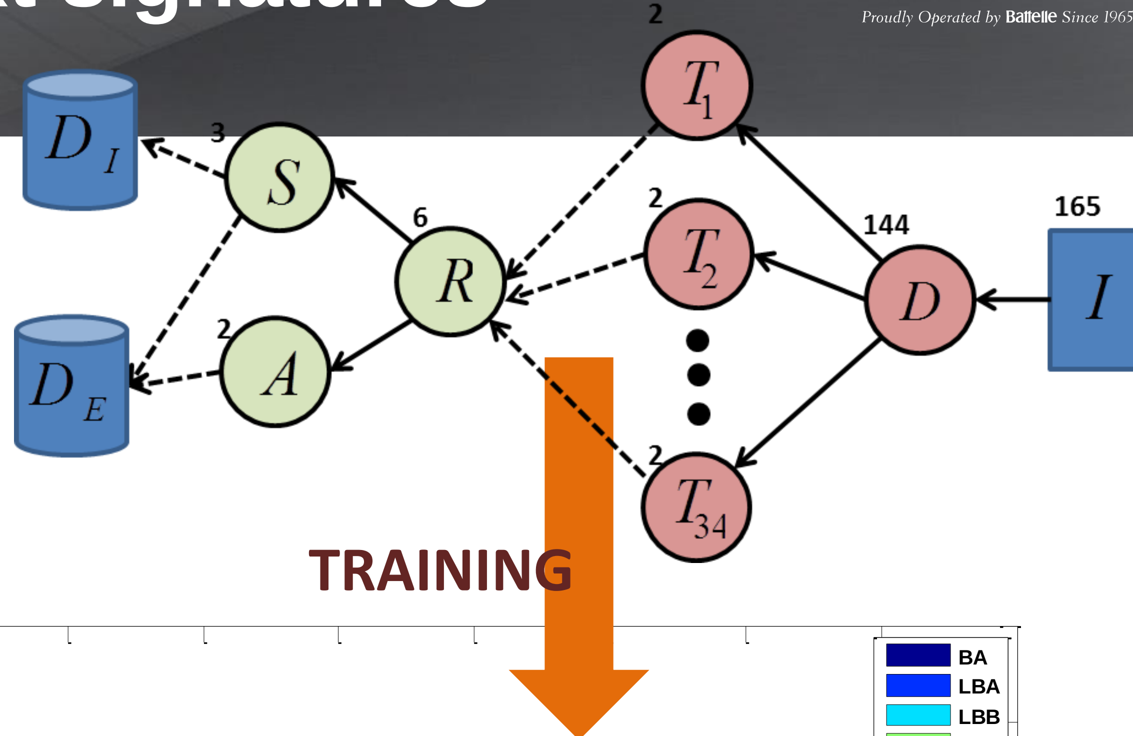
Use automated

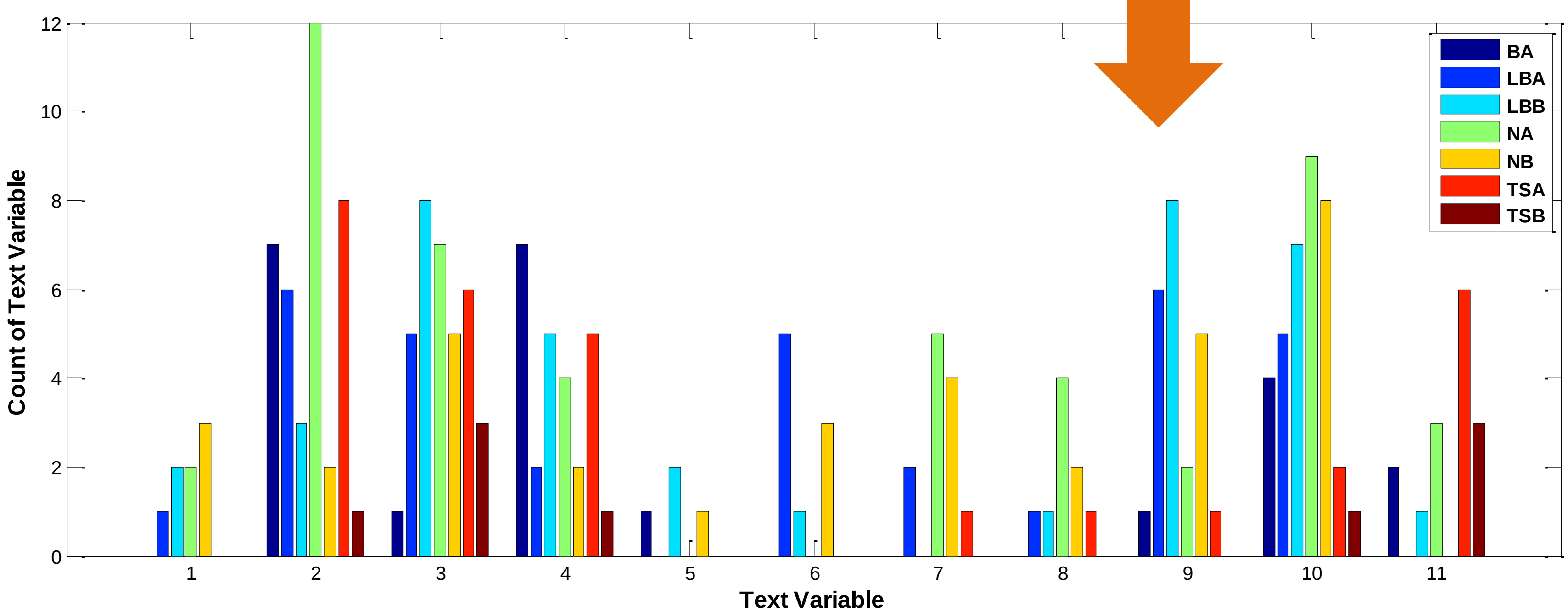
$$P(I_j \mid D_E, D_I) = \frac{\sum_D \sum_T \sum_R \sum_S \sum_A P(D_E \mid A) P(D_E, D_I \mid S) P(A \mid R) P(S \mid R) \prod_q \left[P(R \mid T^{(q)}) P(T^{(q)} \mid D) \right] P(D \mid I) P(I)}{\sum_I \sum_D \sum_T \sum_R \sum_S \sum_A P(D_E \mid A) P(D_E, D_I \mid S) P(A \mid R) P(S \mid R) \prod_q \left[P(R \mid T^{(q)}) P(T^{(q)} \mid D) \right] P(D \mid I) P(I)}$$

- For demonstration we focus on using published journal articles in the public domain.

Open-source text signatures

Hand curated documents show a discriminatory pattern between culture medium recipes





Validation

INFORMATION

- 144 total documents
  - 52 documents hand curated
  - 92 additional documents
- 165 institutions

EVALUATION

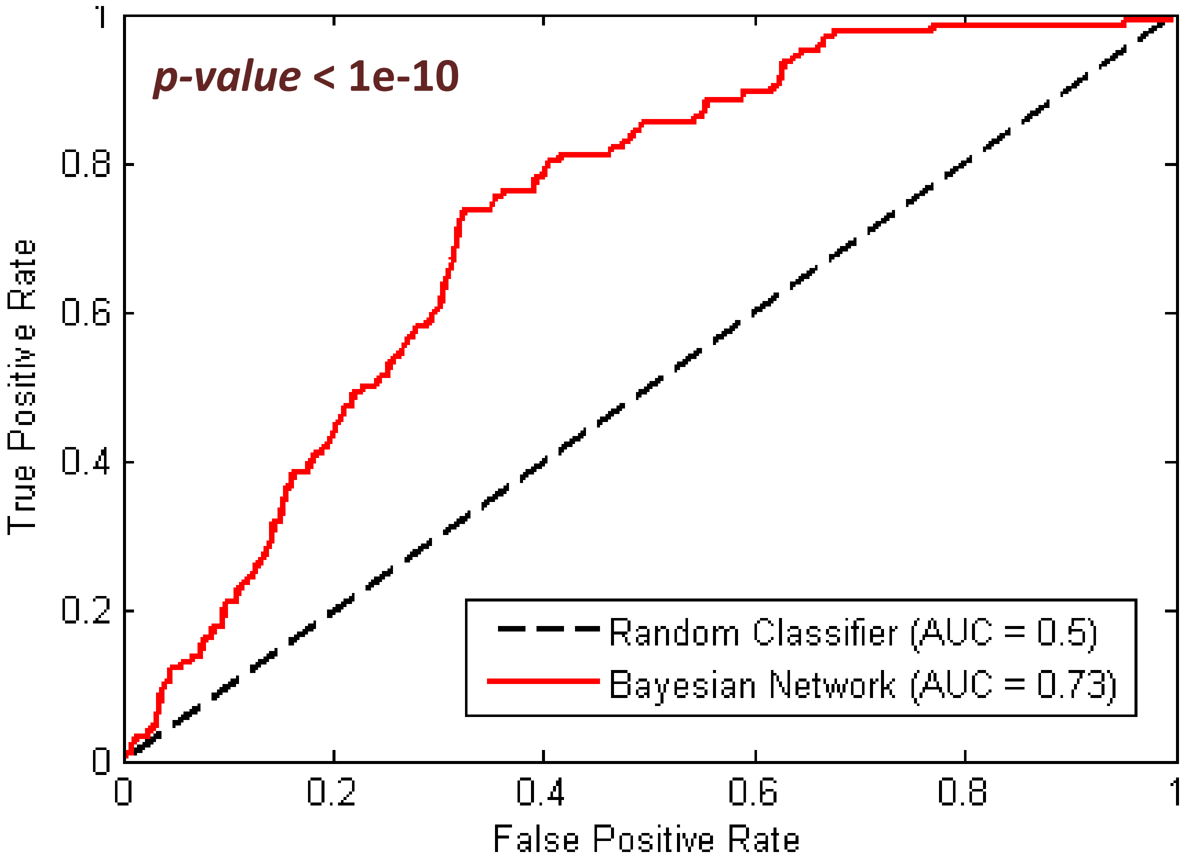
- Cross-validation (bootstrapping): 52 documents
- Area under Receiver Operating Characteristic curve (AUC)

Random Classifier will given an AUC of 0.5

Perfect Classifier will give an AUC of 1.0

AUC Statistically Higher than Random

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965



Bayesian	Random
0.71±0.17	0.48±0.124

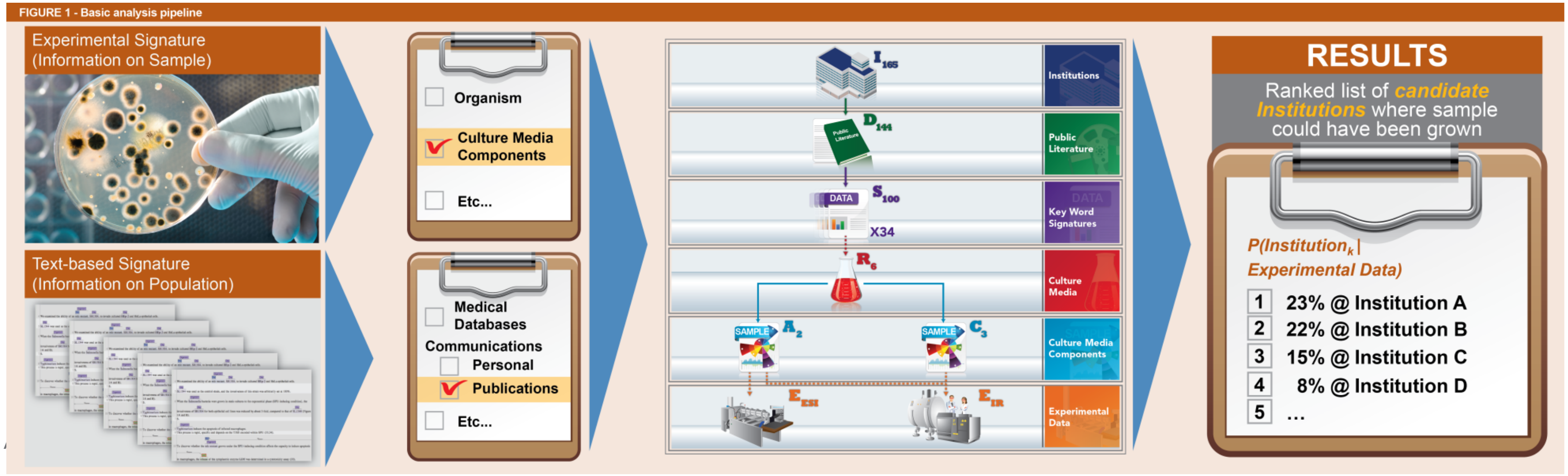
- Issues with Validation
  - Presumably many “false” are “true”
  - Limited to the culture medias of the hand curation

Advantages of the Bayesian Network Approach

- More experimental and/or soft data streams can be added
- Modify the final probability (e.g., foreign vs. domestic, individual researchers)
- Automated approach, any number of documents (institutions, people) can be evaluated

- ***Yields a easy to interpret confidence metric***

FIGURE 1 - Basic analysis pipeline

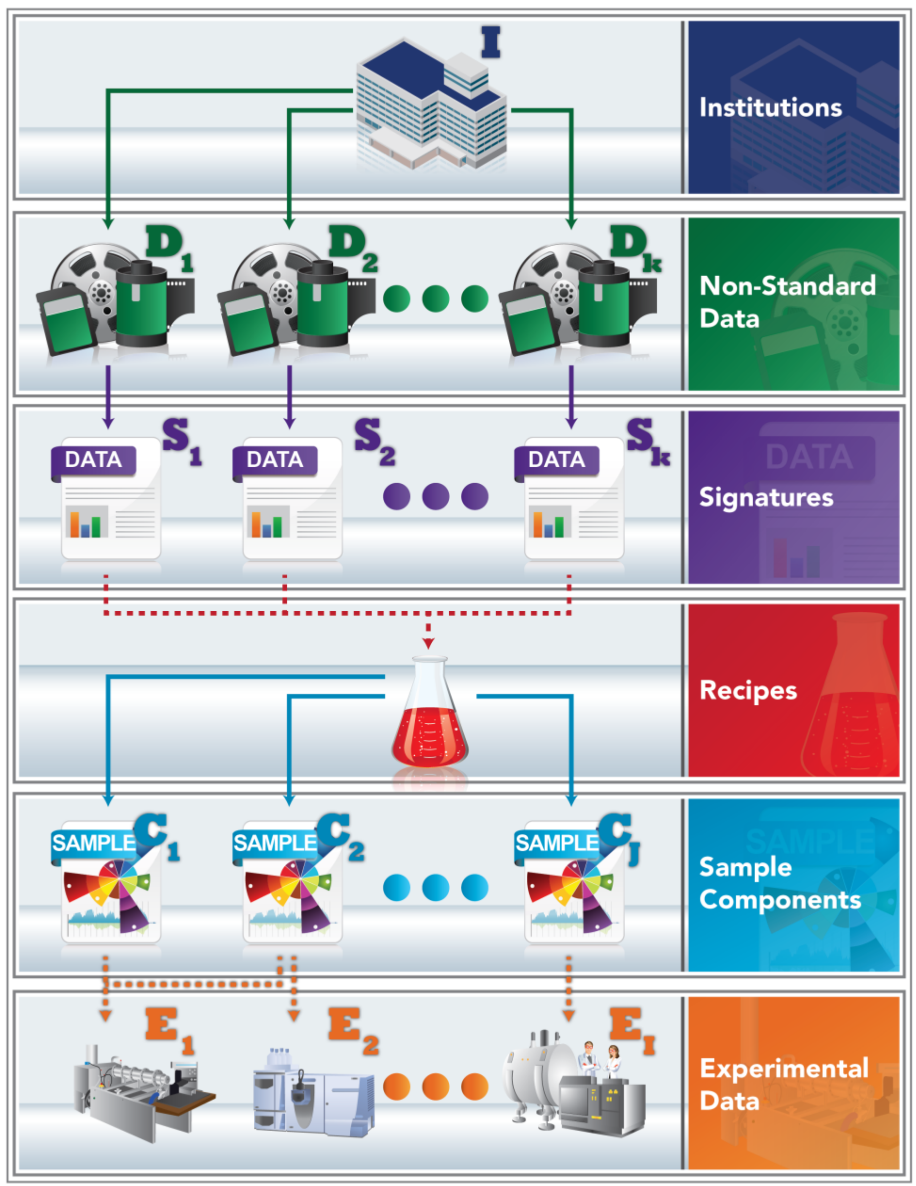


Looking Forward: Bioforensics and Biosurveillance

Pacific Northwest NATIONAL LABORATORY

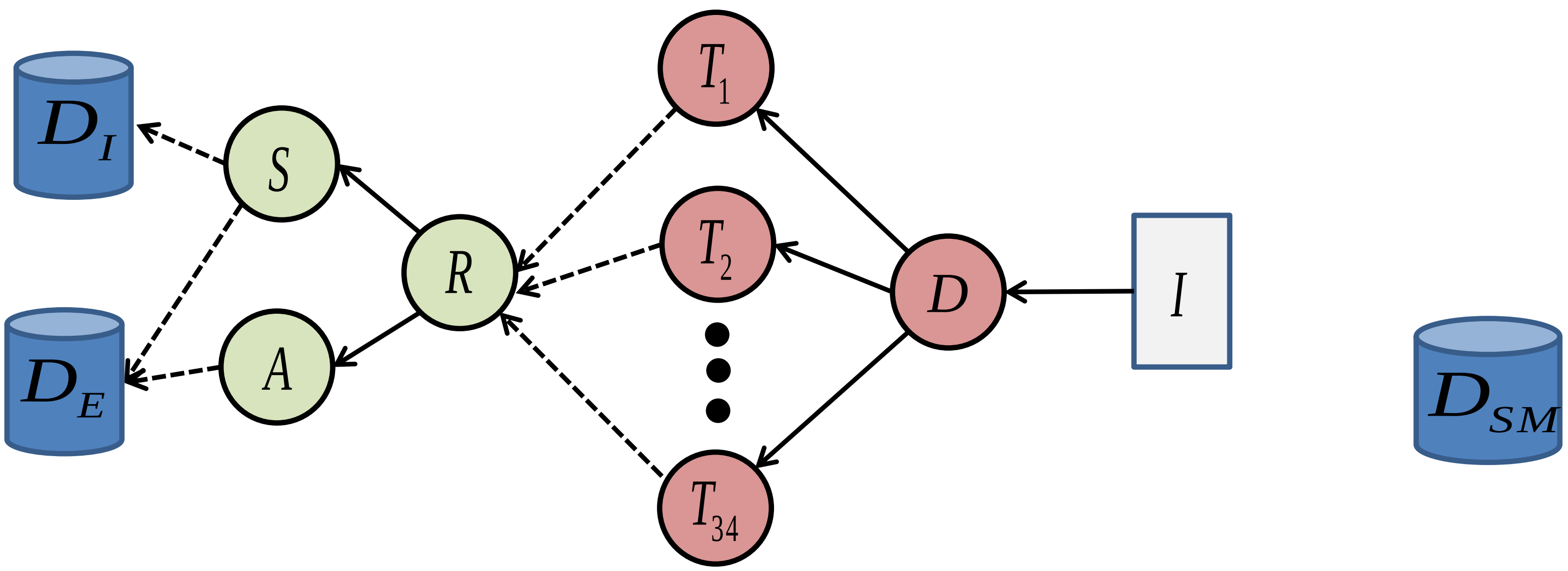
Proudly Operated by Battelle Since 1965

- Expand to include more “who” and “where”
  - Means more nodes, types of information (e.g., social media)
- Dynamic Bayesian networks
  - Evaluate a “threat” over time



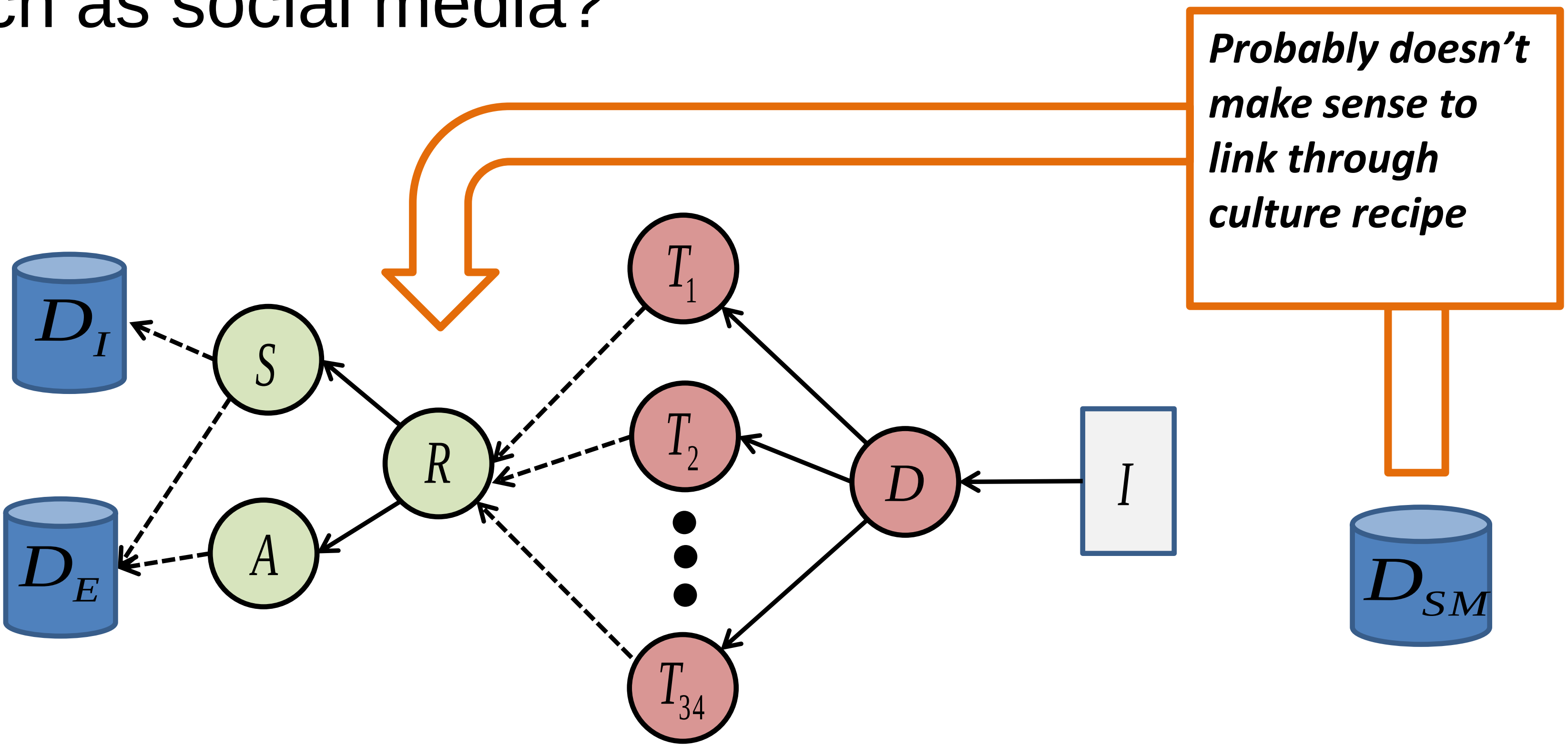
Adding non-traditional "soft" data to the existing network

How can we link in some new source of soft data, such as social media?



Adding non-traditional "soft" data to the existing network

How can we link in some new source of soft data, such as social media?

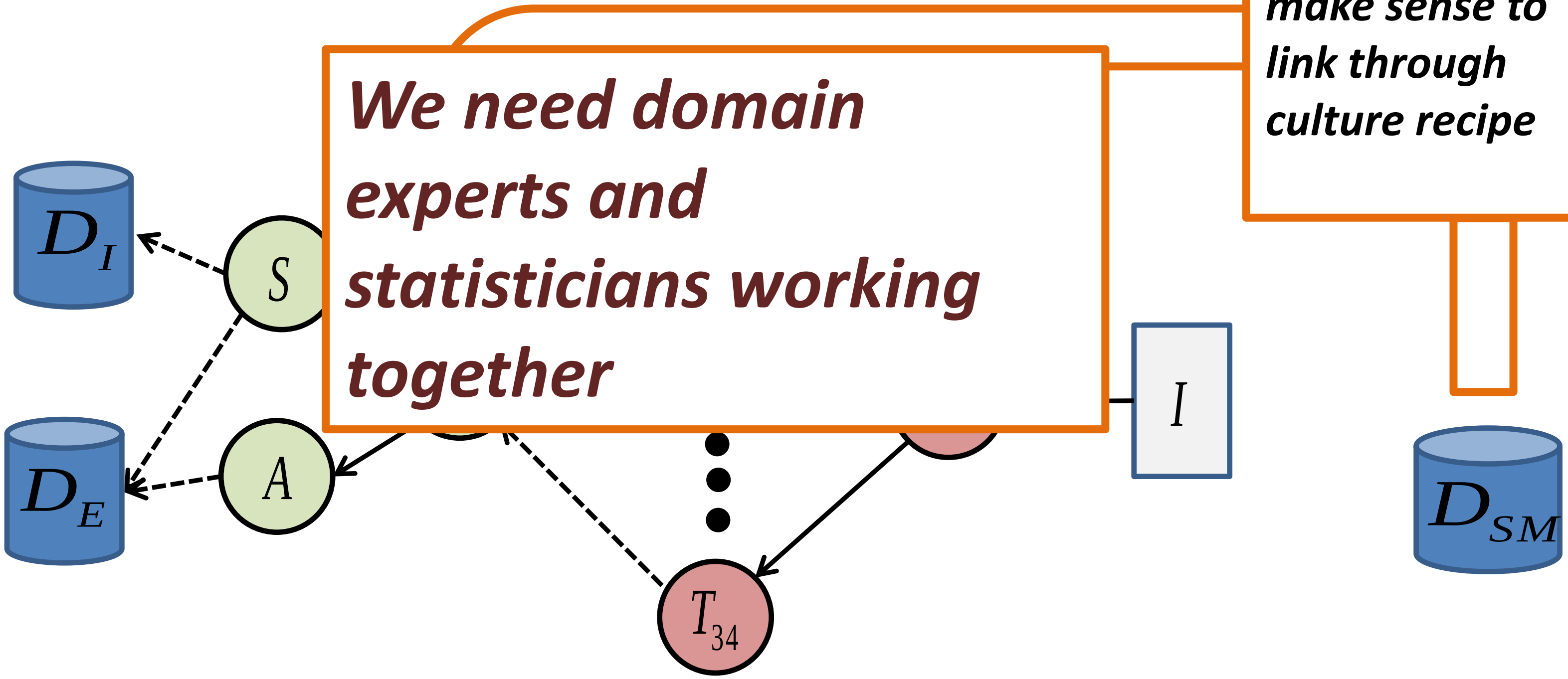


Adding non-traditional “soft” data to the existing network

How can we link in some new source of soft data, such as social media?

Probably doesn’t make sense to link through culture recipe

We need domain experts and statisticians working together



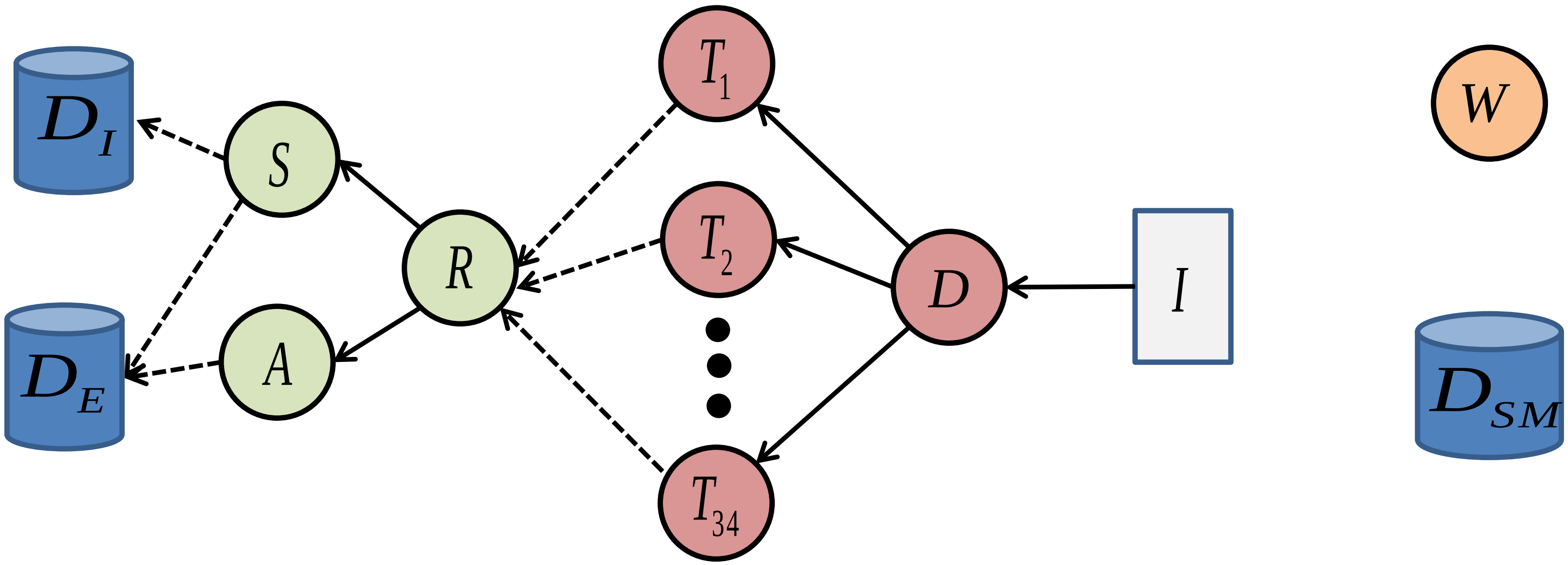
Adding non-traditional "soft" data to the existing network

Pacific Northwest NATIONAL LABORATORY

Proudly Operated by Battelle Since 1965

One approach would be to add a "warning" node

- Compute the probability that there is a threat (*W*) given the "individual" and data source (D_{SM})

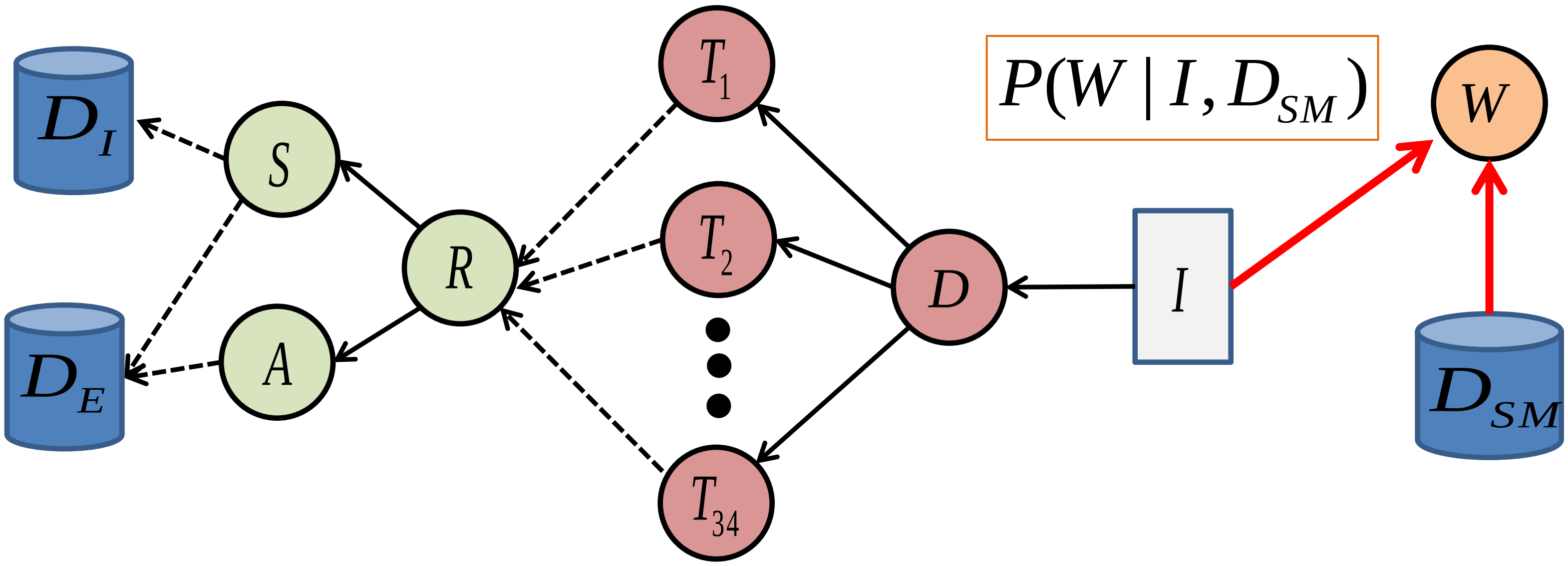


Adding non-traditional "soft" data to the existing network

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

One approach would be to add a "warning" node

- Compute the probability that there is a threat (*W*) given the "individual" and data source (D_{SM})

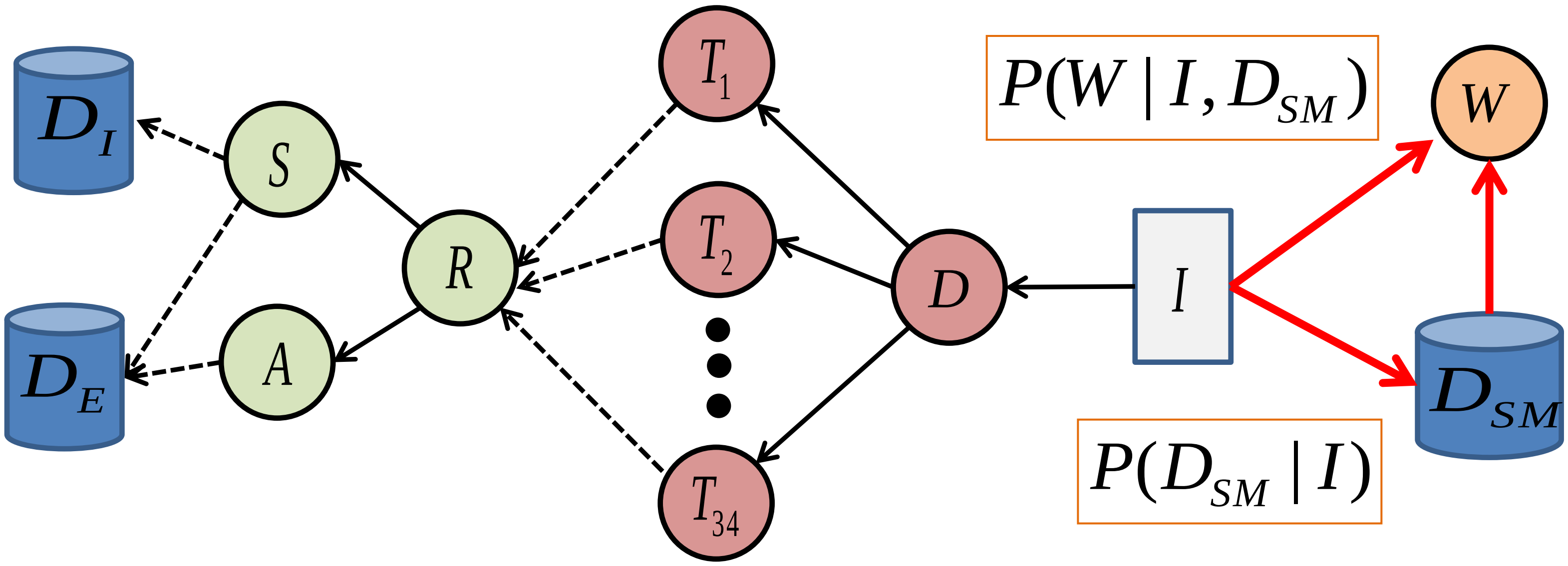


Adding non-traditional "soft" data to the existing network

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

One approach would be to add a "warning" node

- Compute the probability that there is a threat (*W*) given the "individual" and data source (D_{SM})
- Link individuals/institutions to social media

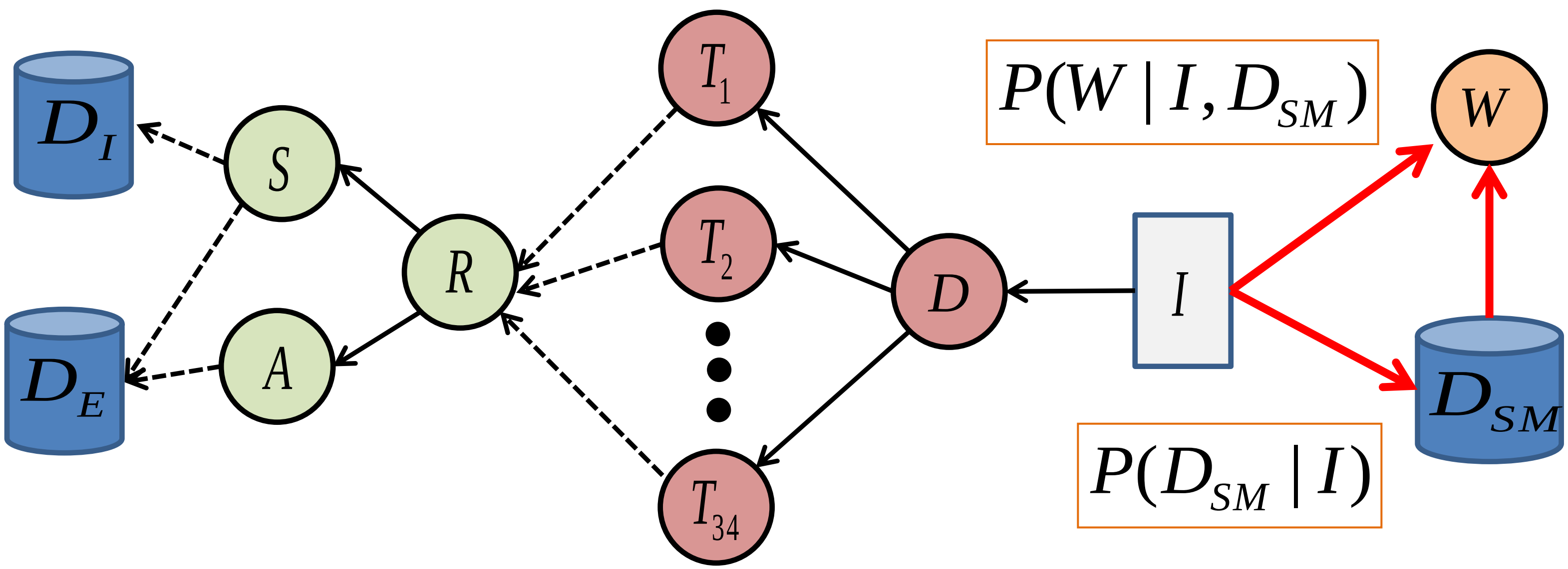


Adding non-traditional "soft" data to the existing network

Pacific Northwest NATIONAL LABORATORY

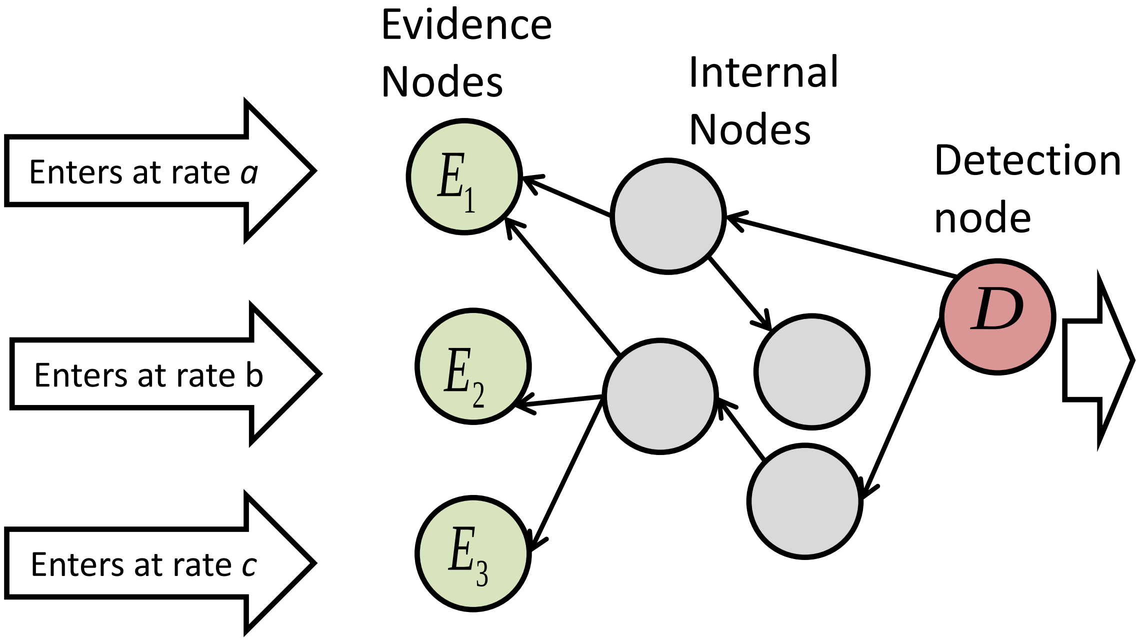
Proudly Operated by Battelle Since 1965

$$P(I_j \mid D_E, D_I, D_{SM}) = \frac{\sum_{\ldots\ldots}\sum_{W} P(D_E, D_I \mid I) \quad P(W \mid I, D_{SM}) P(D_{SM} \mid I) P(I)}{\sum_{I}\sum_{\ldots\ldots}\sum_{W} P(D_E, D_I \mid I) \quad P(W \mid I, D_{SM}) P(D_{SM} \mid I) P(I)}$$



Adding a dynamic component

Generally, integration of multiple 'orthogonal' streams of data improves predictive capability



$$P(Alert \mid E_1, E_2, E_3)$$

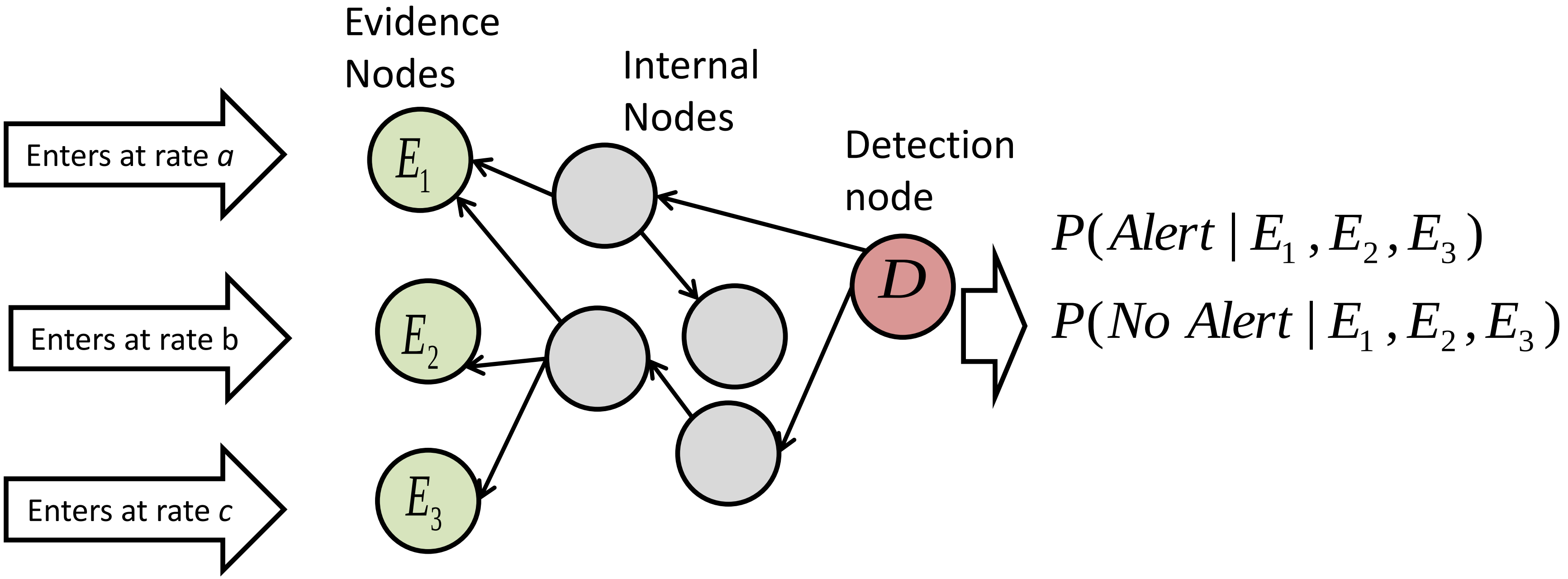
$$P(No\ Alert \mid E_1, E_2, E_3)$$

Webb-Robertson et al., (2009) PSB

Adding a dynamic component

Generally, integration of multiple 'orthogonal' streams of data improves predictive capability

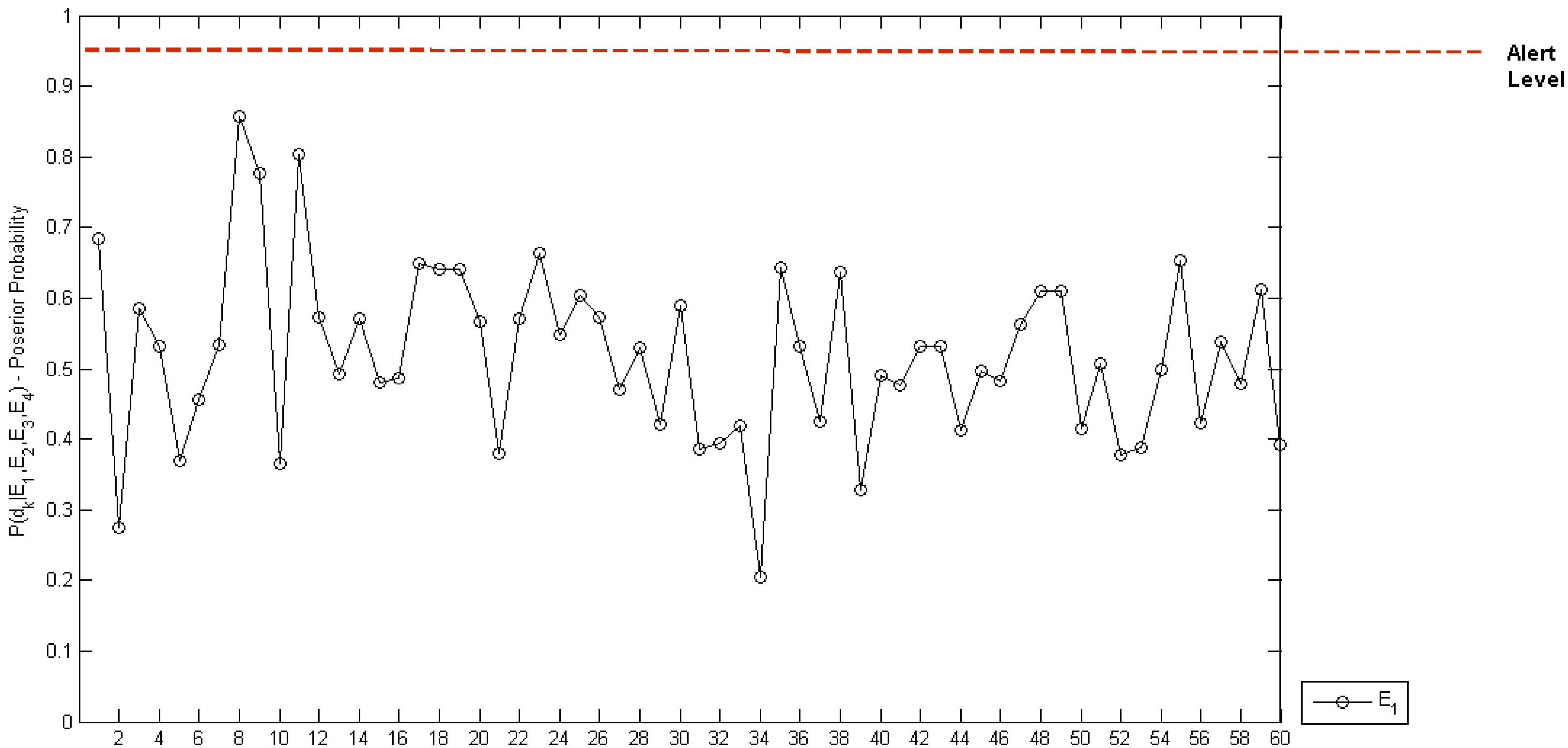
- Automated nature of the network allows continual update of the probability at rate of the fastest source of data.



Webb-Robertson et al., (2009) PSB

Adding a dynamic component

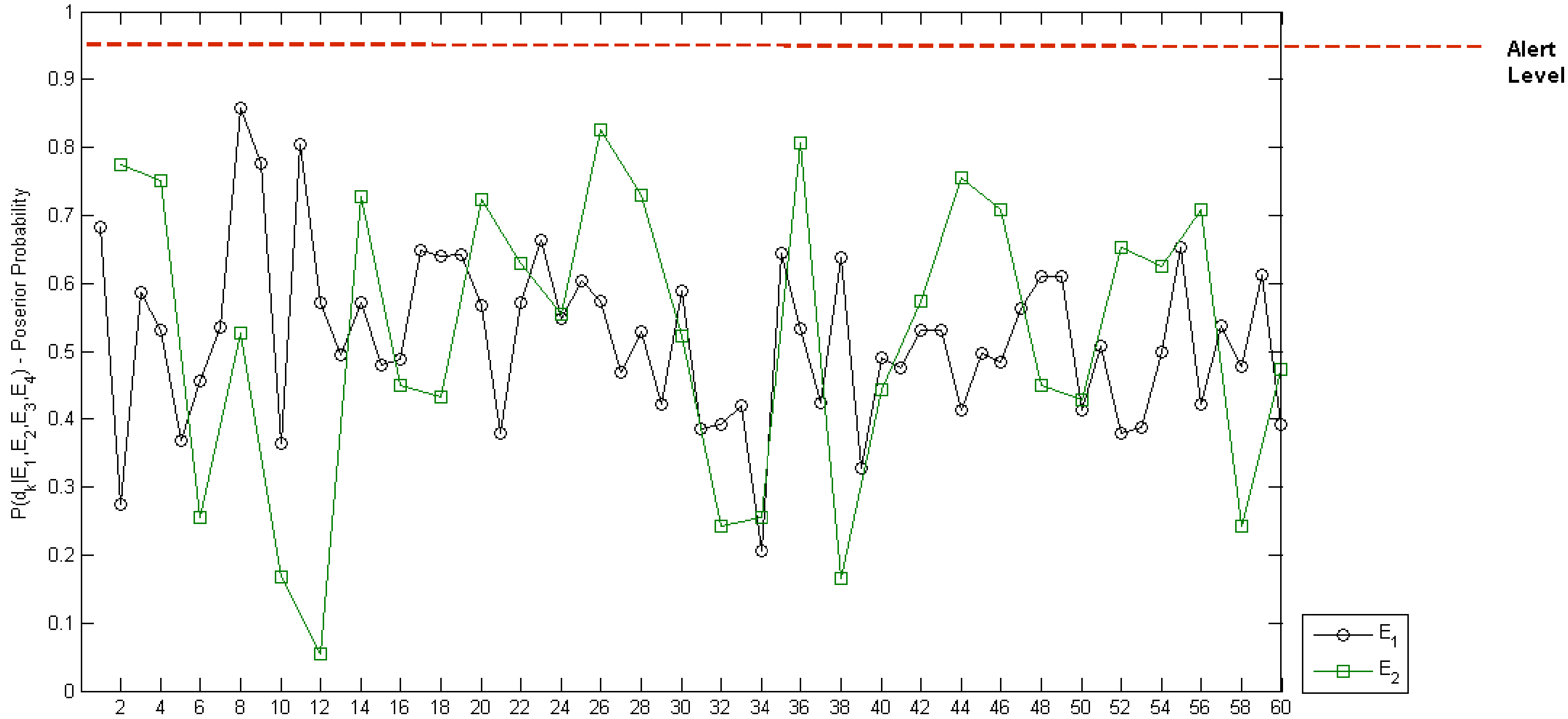
Integration can identify an “alert” where individual data streams may not



Adding a dynamic component

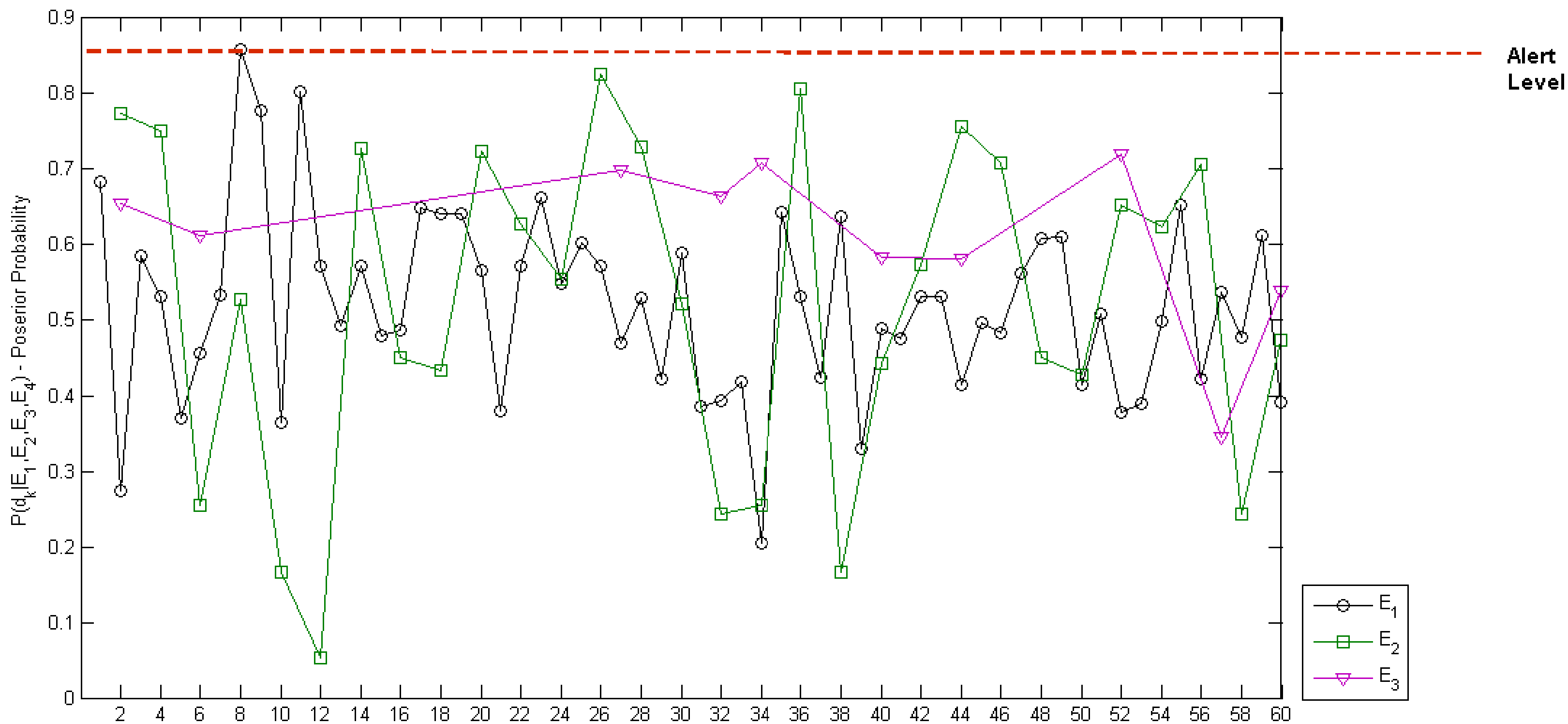
Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Integration can identify an “alert” where individual data streams may not



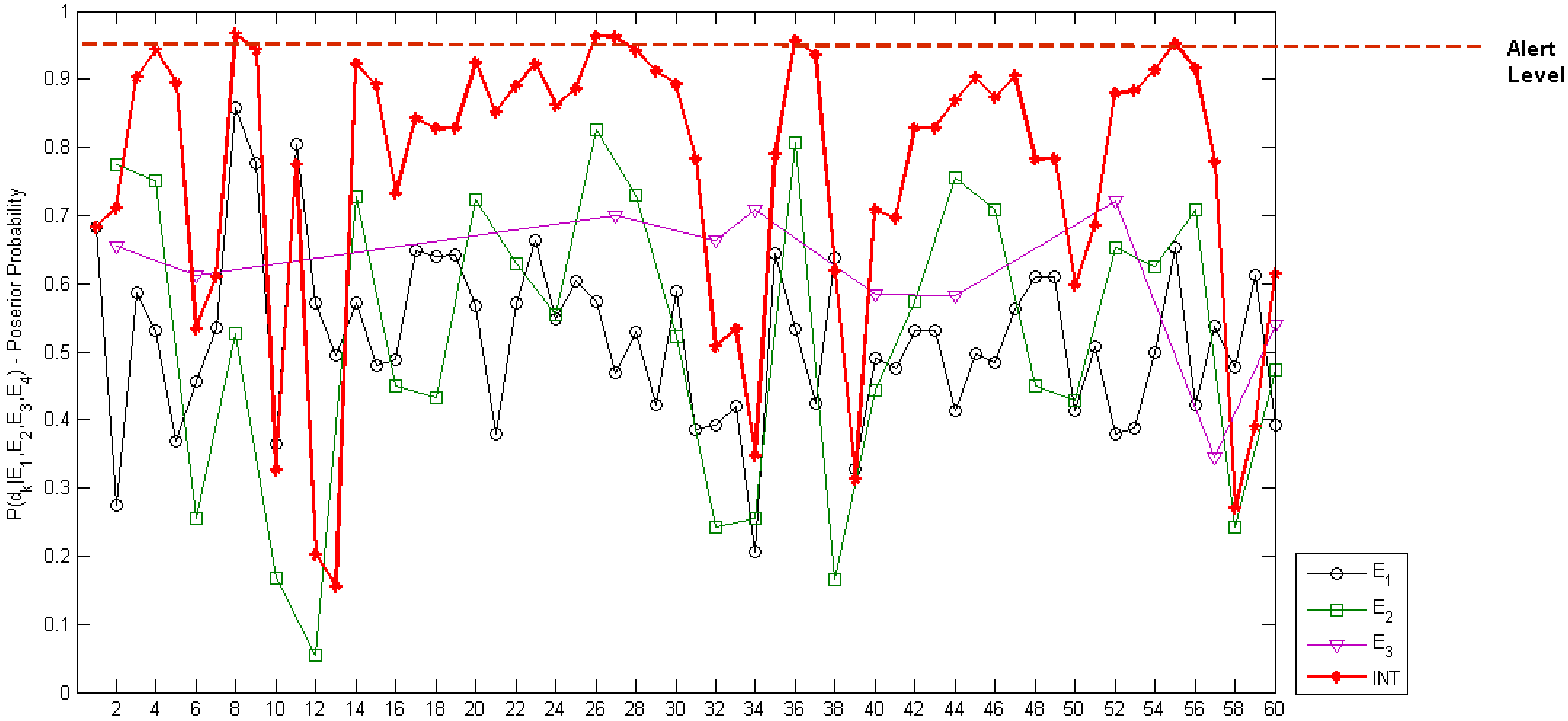
Adding a dynamic component

Integration can identify an “alert” where individual data streams may not



Adding a dynamic component

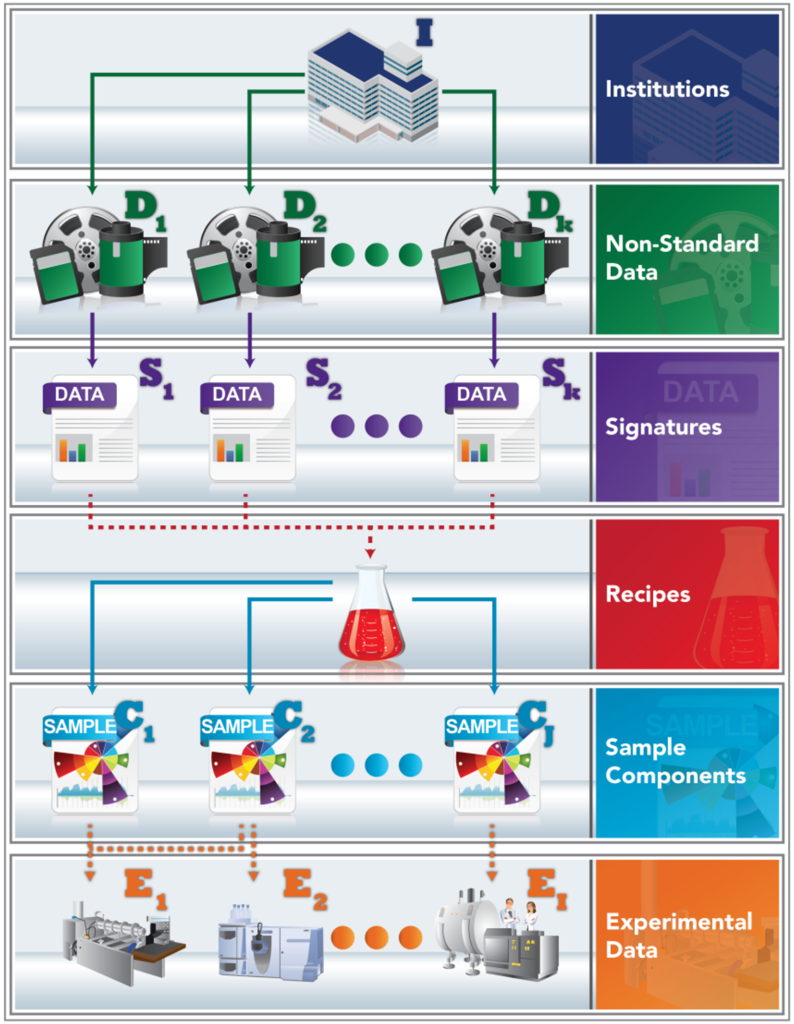
Integration can identify an “alert” where individual data streams may not



Acknowledgments

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

- Funding
  - Department of Homeland Security
  - LDRD (Signature Discovery Initiative)
- Staff
  - B Webb-Robertson (statistics)
  - Courtney Corley (informatics/text analytics)
  - Helen Kreuzer (bioforensics/ experimentation)
  - Lee Ann McCue (microbiology/ Computational Biology)
  - Karen Wahl (bioforensics/ experimentation)



Contact Information

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965

Bobbie-Jo Webb-Robertson

Senior Research Scientist

Computational Biology & Bioinformatics

Pacific Northwest National Laboratory

902 Battelle Blvd / J4-33

Richland, WA 99352

Tel: (509) 375-2292

bj@pnnl.gov