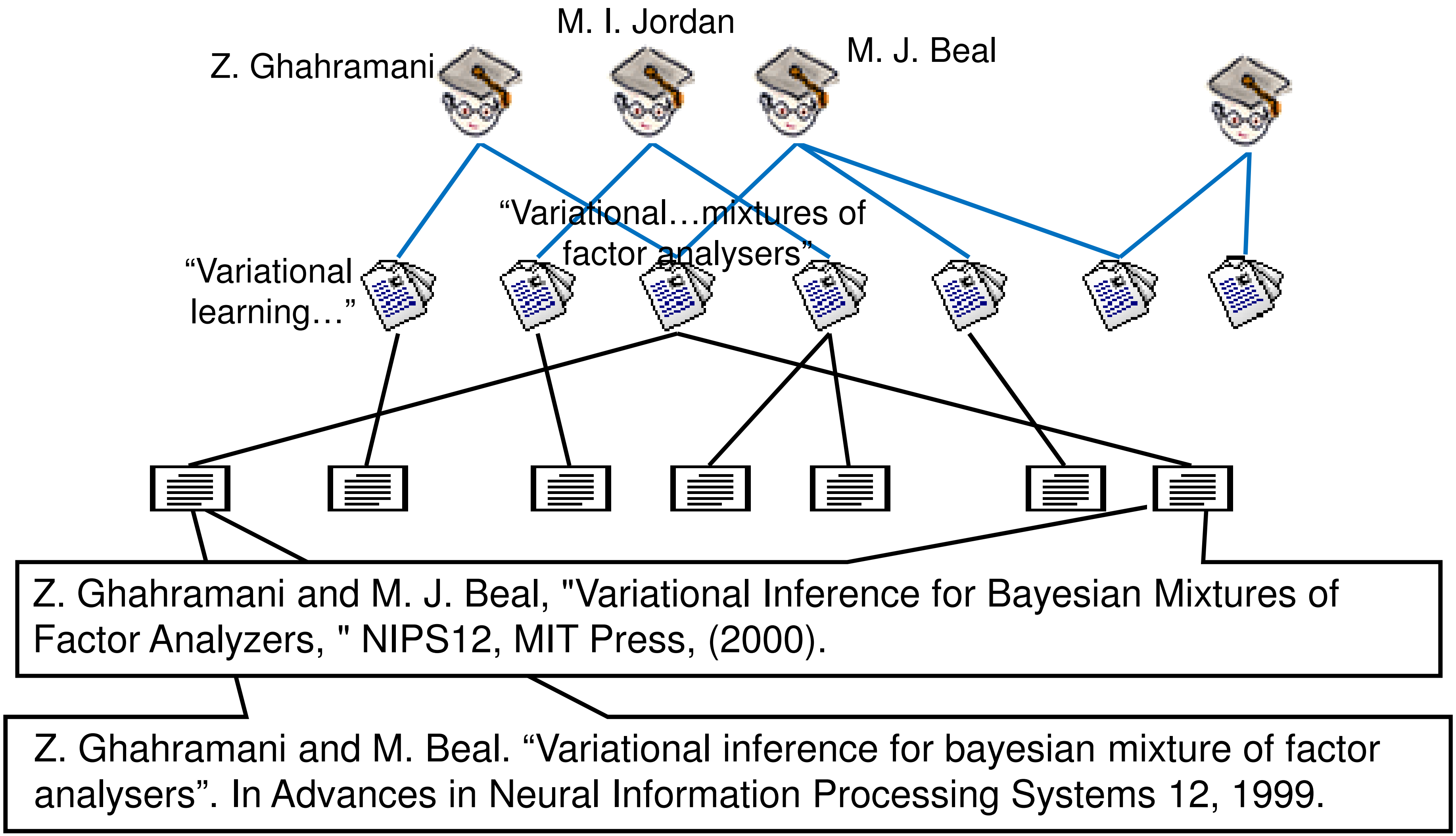
Probabilistic Reasoning about Possible Worlds

Brian Milch

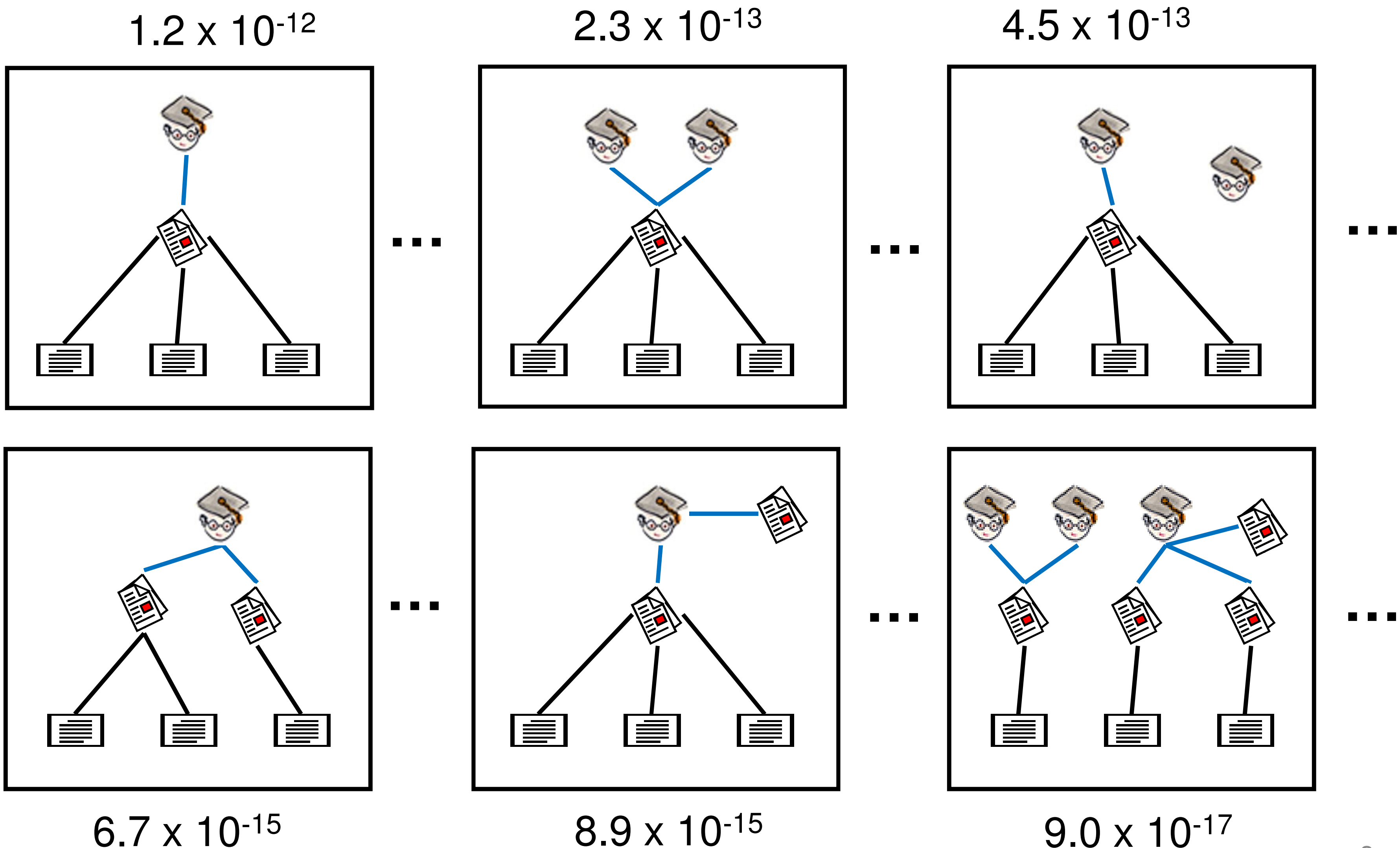
Google
(based on work at UC Berkeley with Stuart Russell and his group)

IPAM Summer School
July 8, 2011

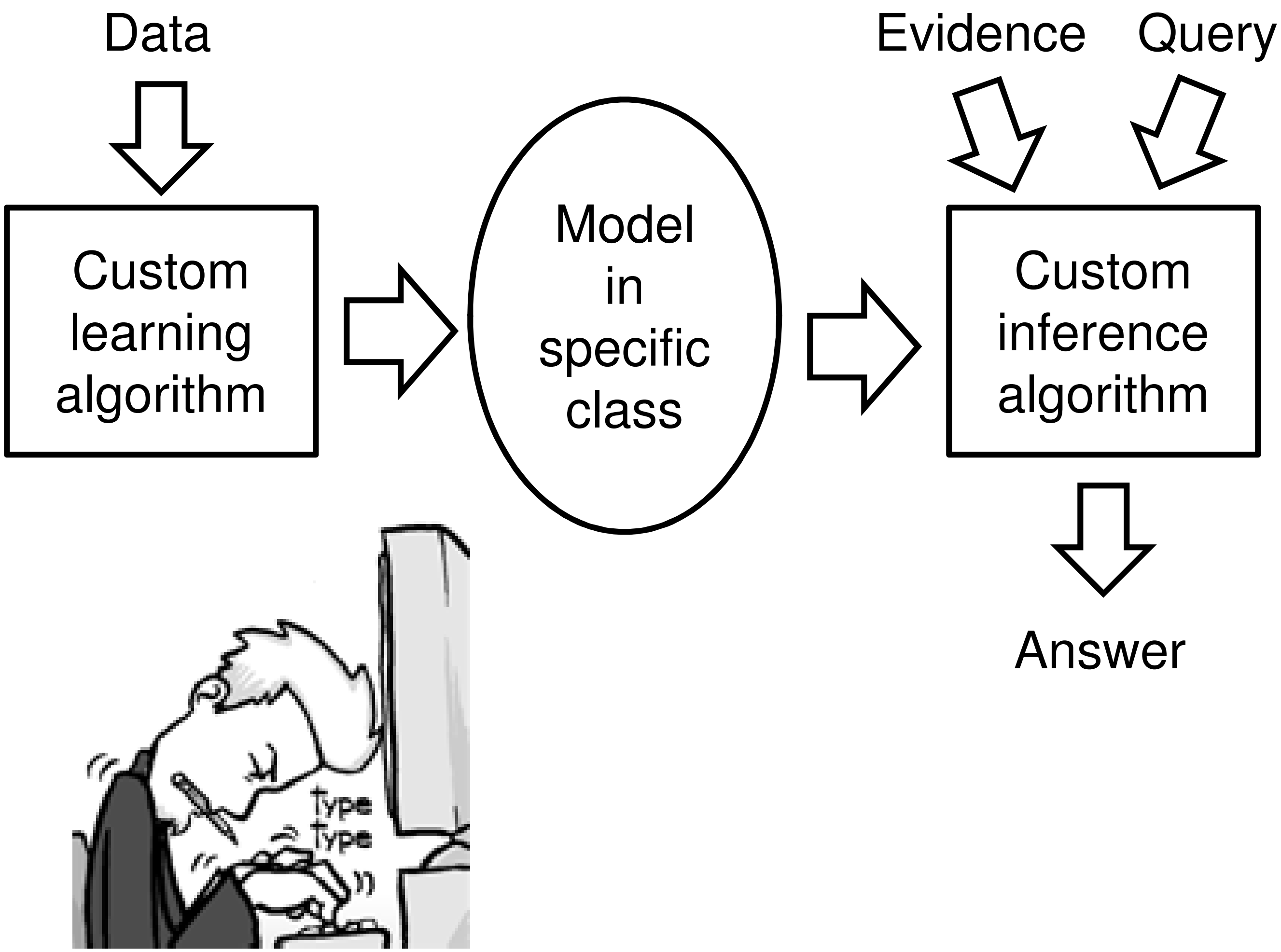
Example: Bibliographies



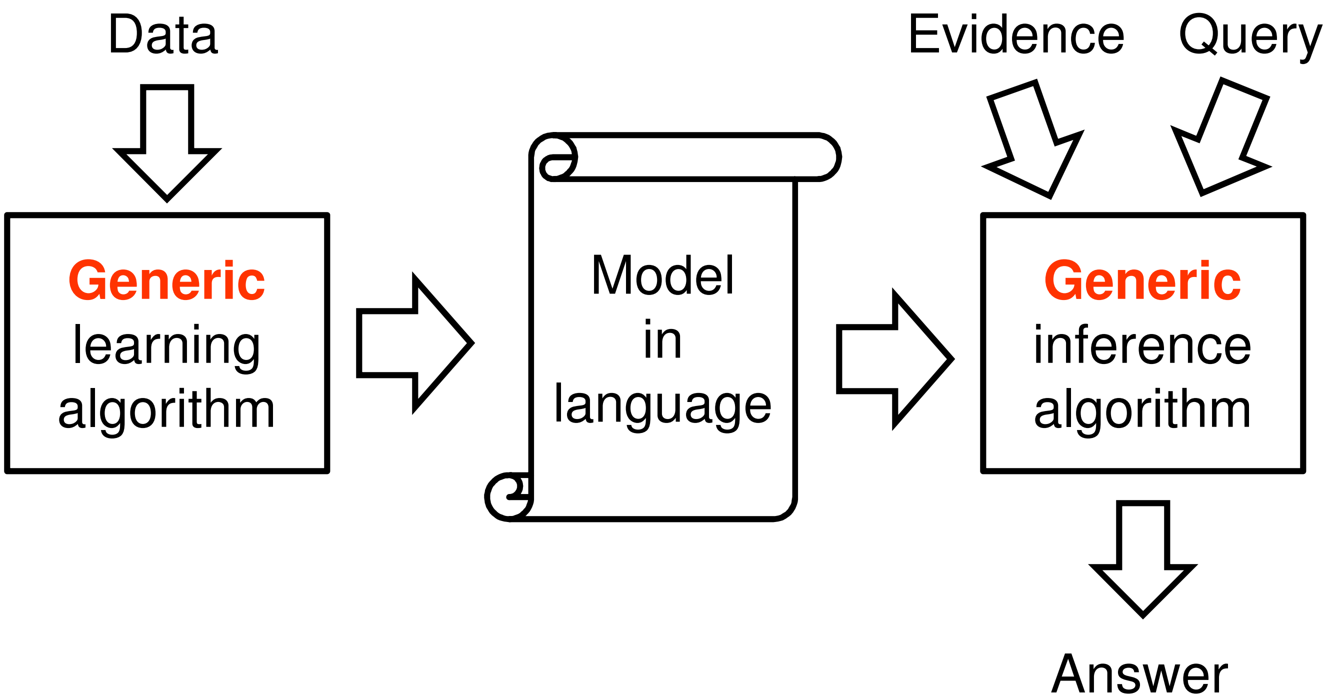
Probabilities on Possible Worlds



Traditional Approach



Role of Representation Language



Outline

- Approaches to probability over possible worlds
- Bayesian Logic (BLOG)
  - Generalization of Bayesian networks
  - Unknown objects
- Inference: MCMC over possible worlds
- Application: Seismic monitoring

Levels of Expressivity

- Atomic: Enumerate all worlds (states)
  - Examples: histogram, Markov model
- Propositional: Describe worlds using propositions or random variables (RVs)
  - Examples: propositional logic, graphical models
- First-order: Parameterize propositions or RVs with logical variables
  - Examples: first-order logic, programming languages, plate models, BLOG, Markov Logic, …

Example: Rules of Chess

- 1 page in first-order logic
  On(*color,piece,x,y,t*)

- ~100000 pages in propositional logic
  WhiteKingOnC4Move12

- ~10000000000000000000000000000000000000000 pages as atomic-state model

  R.B.KB.RPPP..PPP..N..N…..PP….q.pp..Q..n..n..ppp..pppr.b.kb.r

(Note: Chess is still tiny compared to the real world)

First-Order Logical Languages

- For a given application, declare:
  - Constant symbols: Citation1, Ghahramani, …
  - Predicate symbols: IsProf(*person*), AuthorOf(*person*, *publication*), …
  - Function symbols: PubCited(*citation*), Name(*person*), Affiliation(*person*, *date*)…

First-Order Structures

- Formalization of possible worlds
- Each structure specifies domain of objects, and maps:
  - Constant symbols → objects
  - Predicate symbols → relations on objects
  - Function symbols → functions on objects
- Each sentence of logical language is true or false in each structure

Approach 1: Probabilities for First-Order Formulas

- Specify probabilities of sentences or formulas, as in:
  - **0.2: ∀x IsProf(x)** [Gaifman 1964]
  - **∀x(μ(IsProf(x)) = 0.7)** [Halpern 1990]

- Drawback: hard to tell if there's exactly one distribution over structures satisfying these constraints

Approach 2: Weights for Formulas

2.5: IsProf(x) → Busy(x)

		Busy(Smith) T	Busy(Smith) F
IsProf(Smith)	T	$e^{2.5}$	1
	F	$e^{2.5}$	$e^{2.5}$

- Weight of world x is $\exp(\sum_{\varphi} w_{\varphi}\ \#sat(\varphi\ ,x))$
- Get probability distribution by normalizing
- Weights need to be learned jointly

(Markov logic [Richardson & Domingos 2006], relational Markov networks [Taskar et al. 2002])

Approach 3: Logic Plus Coin Flips

0.8: BeingProfCausesBusy(x)

IsProf(x) $\wedge$ BeingProfCausesBusy(x) $\rightarrow$ Busy(x)

- Coin flips are mutually independent
- Coin flips yield a structure by logical implication; anything not implied is false

(probabilistic Horn abduction [Poole 1993],
independent choice logic [Poole 1997],
PRISM [Sato & Kameya 1997])

Approach 4: Conditional Probability Statements

```
Busy(x) {
  if IsProf(x) then ~ Bernoulli[0.8]
  else ~ Bernoulli[0.3]
}
```

- Specify conditional probability distribution (CPD) for each predicate and function
  - Defines directed graphical model
  - Model must be acyclic

(BUGS/plates [Thomas et al. 1992], relational Bayes nets [Jaeger 1997], probabilistic relational models [Friedman et al. 1999], Bayesian logic programs [Kersting & De Raedt 2001], Bayesian logic [Milch et al. 2005], multi-entity Bayes nets [Laskey 2008], many others)

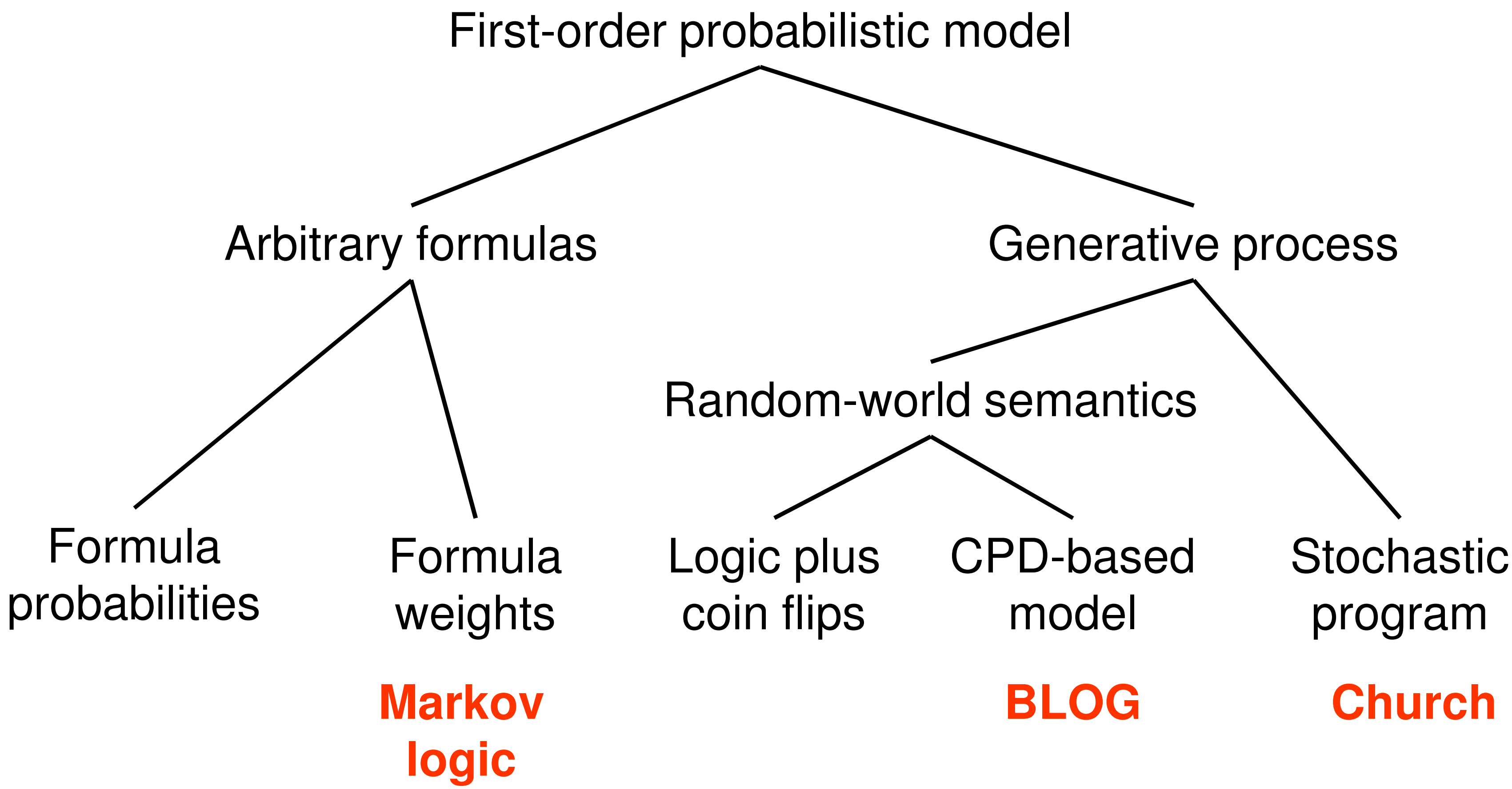
Approach 5: Stochastic Programs

```
(define (Busy is_prof) (cond is_prof
                             (flip 0.8)
                             (flip 0.2)))
```

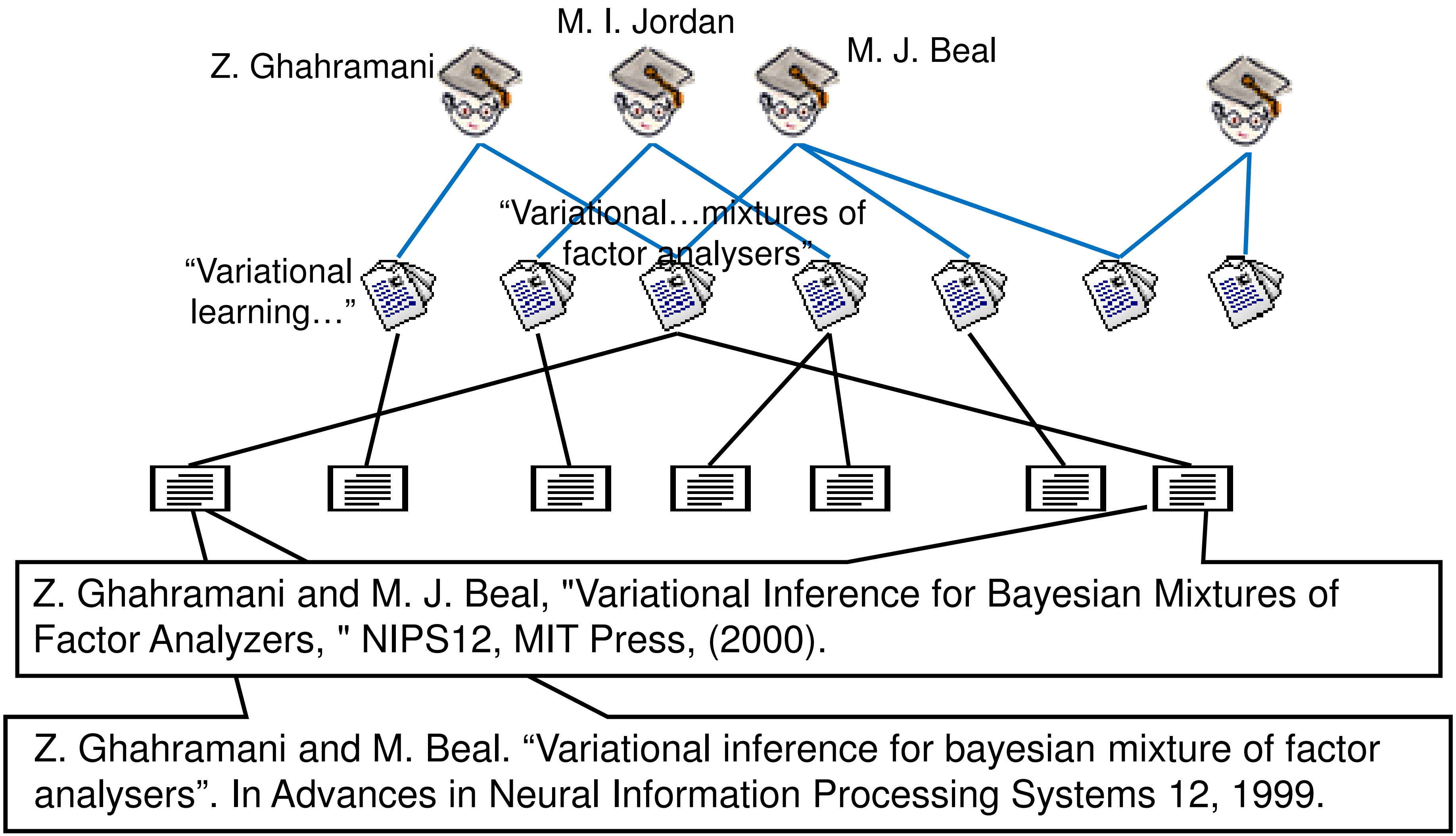
- Functions are stochastic: each call may return a different value (but see "mem" in Church)
  - Contrast with **IsProf(x)** in other approaches
  - Store results in program variables for re-use, implicitly define possible world

(Stochastic logic programs [Muggleton 1996], stochastic programs [Koller et al. 1997], IBAL [Pfeffer 2001], λ_O [Frank et al. 2005], Church [Goodman et al. 2008])

Summary of Approaches



Bibliographies Again



BLOG Model: Object Existence

```
type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();


#Pub ~ NumPubsPrior();
```

BLOG Model: Relational Structure

```
type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();


#Pub ~ NumPubsPrior();
random NaturalNum NumAuthors(Pub p) ~ NumAuthorsPrior();
random Researcher NthAuthor(Pub p, NaturalNum n)
    if (n < NumAuthors(p)) then ~ Uniform({Res r});


random Pub PubCited(Citation c) ~ Uniform({Pub p});
```

Full BLOG Model for Citations

```
type Researcher; type Pub; type Citation;
guaranteed Citation Cit1, Cit2, Cit3, Cit4;

#Researcher ~ NumResearchersPrior();
random String Name(Researcher r) ~ NamePrior();

#Pub ~ NumPubsPrior();
random NaturalNum NumAuthors(Pub p) ~ NumAuthorsPrior();
random Researcher NthAuthor(Pub p, NaturalNum n)
    if (n < NumAuthors(p)) then ~ Uniform({Res r});
random String Title(Pub p) ~ TitlePrior();

random Pub PubCited(Citation c) ~ Uniform({Pub p});
random String Text(Citation c)
    ~ FormatCPD(Title(PubCited(c)),
          {n, Name(NthAuthor(PubCited(c), n)) for
           NaturalNum n : n < NumAuthors(PubCited(c))});
```

Flexibility of Number Variables

- Can have whole family of number variables for a given type:

```
#Pub(FirstAuthor = r) ~ PubsPerAuthorPrior();
```

For each researcher ***r***, there is random variable indicating $|\{\boldsymbol{p} : \text{FirstAuthor}(\boldsymbol{p}) = \boldsymbol{r}\}|$

- Number can depend on other variables:

```
#Pub(FirstAuthor = r)
    if IsProf(r) then ~ ProfPubsPrior()
    else ~ IndividualPubsCPD(Diligence(r));
```

What Exactly Are the Objects?

```
#Researcher ~ NumResearchersPrior();
```

Objects are (Res, 1), (Res, 2), …

```
#Pub(FirstAuthor = r) ~ PubsPerAuthorPrior();
```

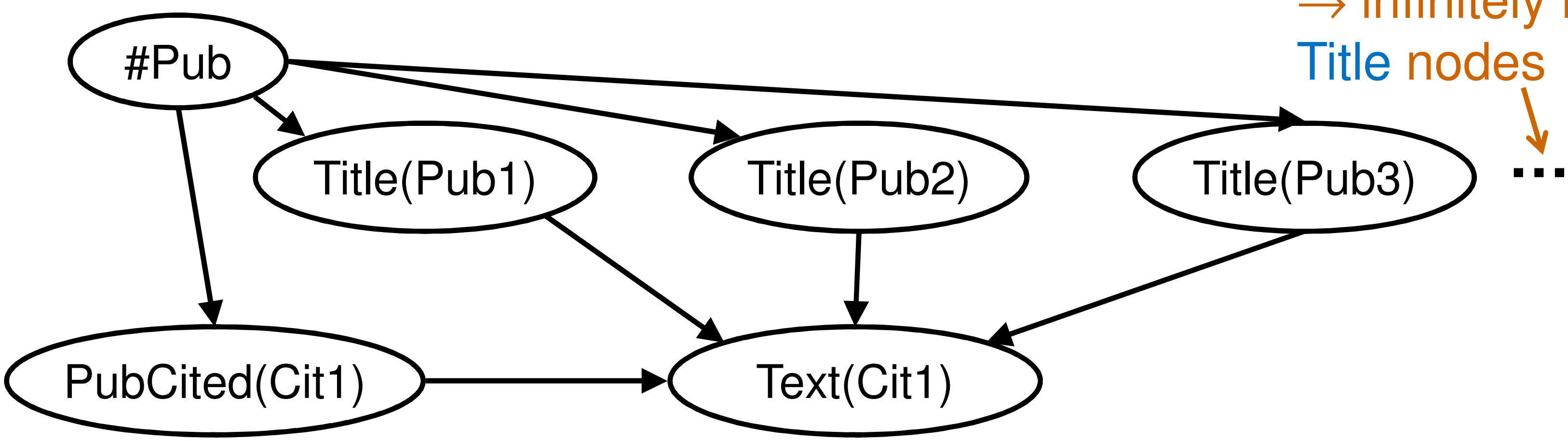
Publications with FirstAuthor = (Res, 2) are:
(Pub, (FirstAuthor, (Res, 2)), 1)
(Pub, (FirstAuthor, (Res, 2)), 2)
…

Result: Given values of number variables, there's no ambiguity about *which* publications have which first authors

BN Defined by BLOG Model

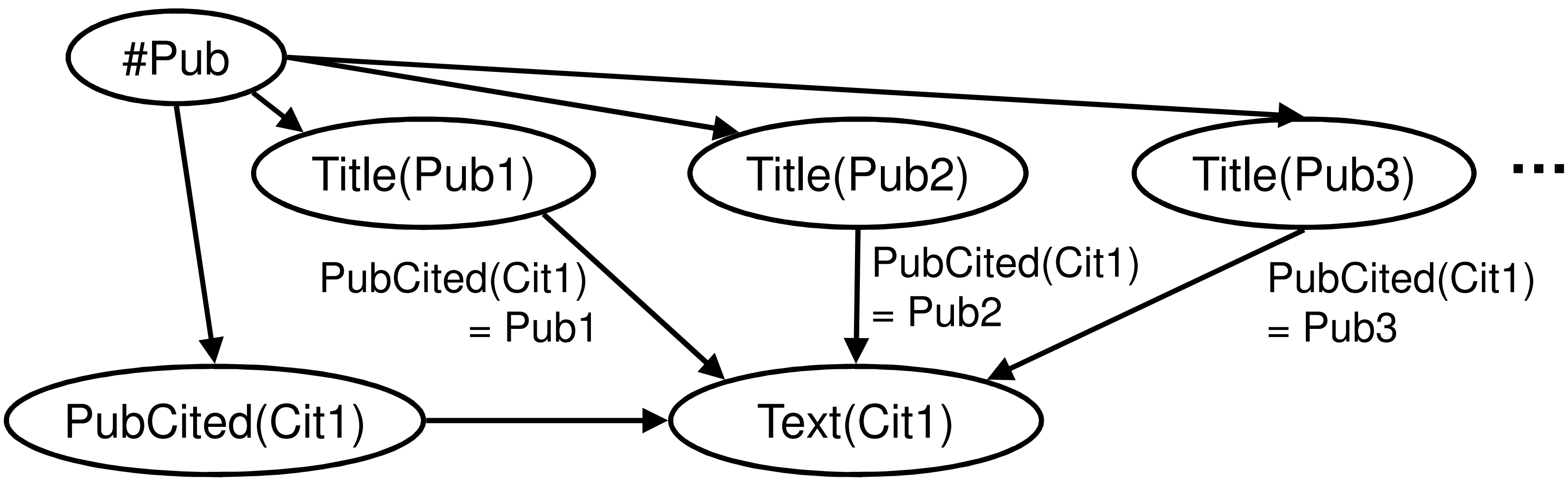
- BN includes
  - All variables that have values in any world
  - All potential dependencies

#Pub unbounded
→ infinitely many
Title nodes



BN has no topological numbering ⇨ Previous results don't guarantee that CPDs define joint distribution

Contingent Bayes Nets



- Contingent Bayes net (CBN)
  – Each edge labeled with condition
  – Each world yields subgraph of active edges

[Milch *et al.*, AI/Stats 2005]

BLOG Semantics Using CBNs

- BLOG model ***M*** defines set of possible worlds $\Omega_{\boldsymbol{M}}$, contingent Bayes net $\boldsymbol{B}_{\boldsymbol{M}}$
- *Theorem*: If for each world in $\Omega_{\boldsymbol{M}}$, the active subgraph of $\boldsymbol{B}_{\boldsymbol{M}}$ has a topological numbering, then the model ***M*** fully defines a probability distribution on $\Omega_{\boldsymbol{M}}$

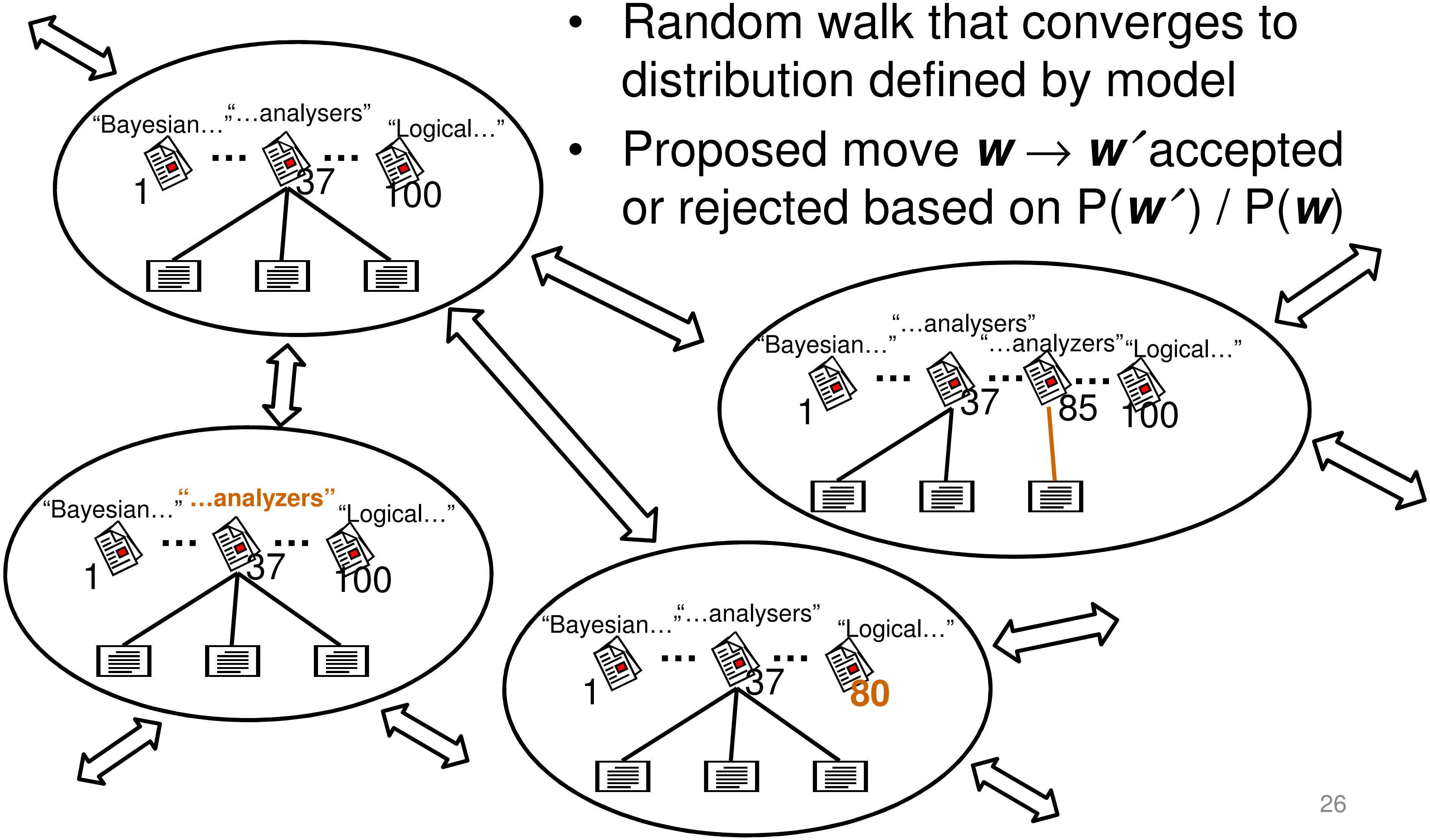
Topological numbering can vary from world to world

⇩

Flexibility in modeling unknown objects, relations

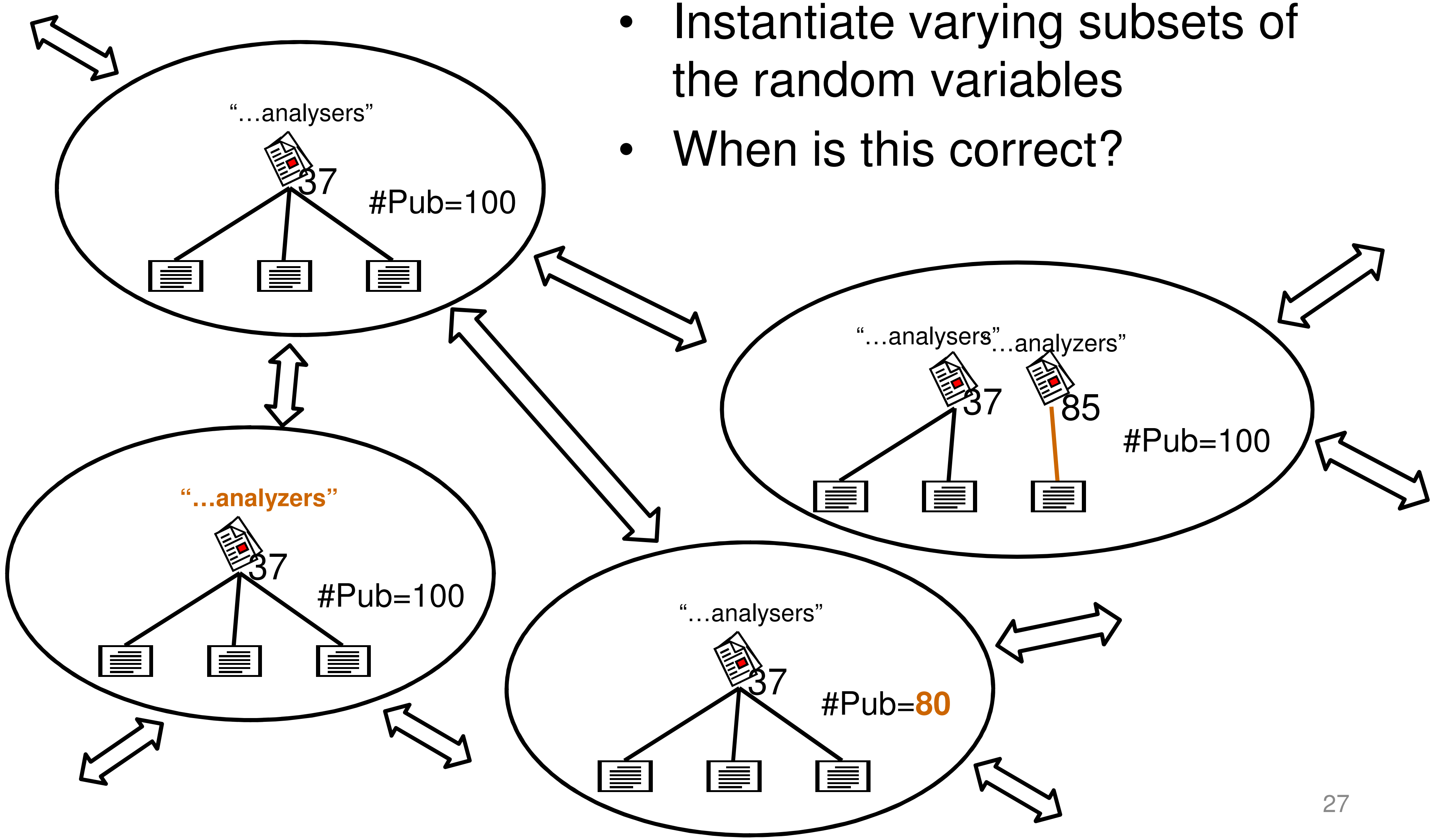
Markov Chain Monte Carlo (MCMC) over Possible Worlds

- Random walk that converges to distribution defined by model
- Proposed move $\boldsymbol{w} \rightarrow \boldsymbol{w}'$ accepted or rejected based on $P(\boldsymbol{w}') / P(\boldsymbol{w})$

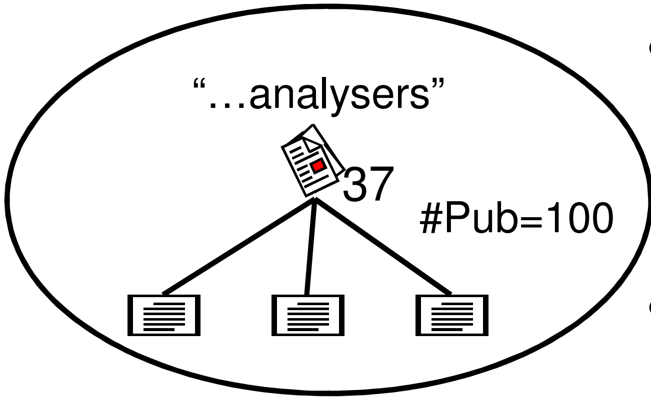


Partial Worlds

- Instantiate varying subsets of the random variables
- When is this correct?



MCMC over Sets of Worlds

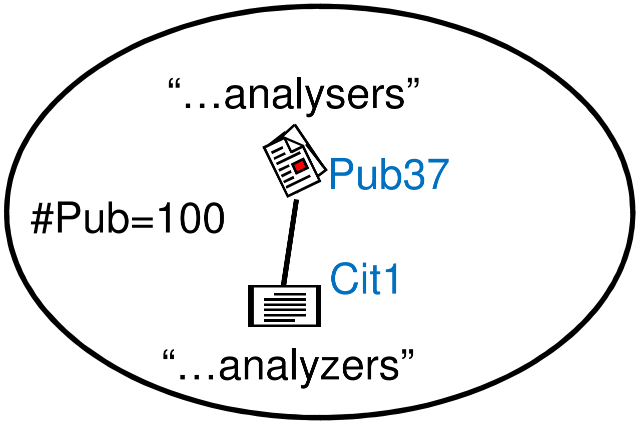


- Think of partial world as description satisfied by whole set of worlds
- Probability of partial world is probability of this whole set

- *Theorem*: MCMC over partial worlds is correct if:
  - Each partial world is specific enough to affirm the given evidence and answer the query
  - Partial worlds define non-overlapping sets of possible worlds

[Milch & Russell, UAI 2006]

Probabilities of Partial Worlds



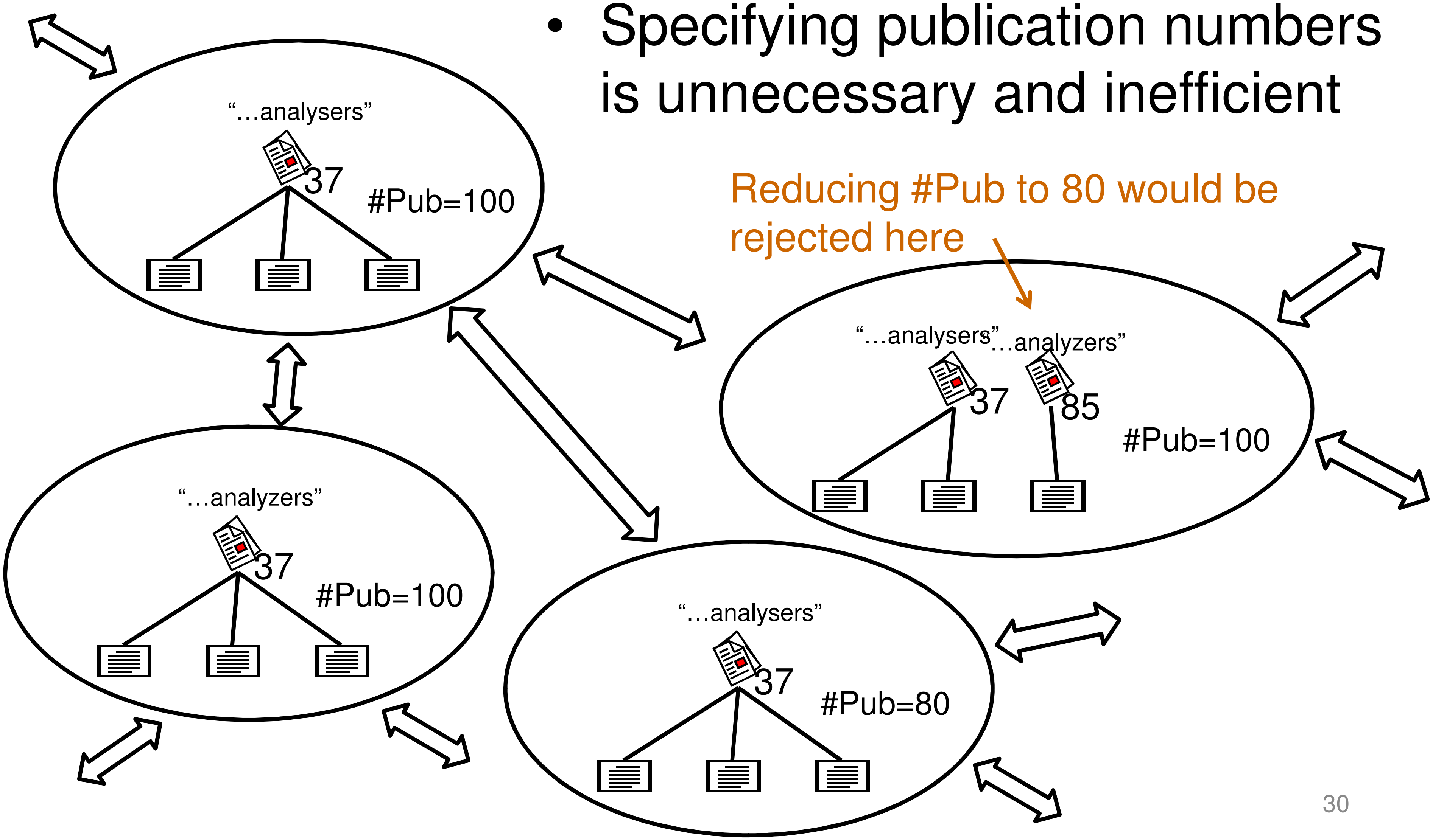
- What's the probability of the set of worlds satisfying this description?
- As in BN, it's product of CPDs:

$$\begin{aligned}
&p(\#Pub = 100) \\
&\times p(\text{Title}(\text{Pub37}) = \text{“…analysers”} \mid \#Pub = 100) \\
&\times p(\text{PubCited}(\text{Cit1}) = \text{Pub37} \mid \#Pub = 100) \\
&\times p(\text{Text}(\text{Cit1}) = \text{“…analyzers”} \mid \text{PubCited}(\text{Cit1}) = \text{Pub37}, \\
&\qquad\qquad\qquad\qquad \text{Title}(\text{Pub37}) = \text{“…analysers”})
\end{aligned}$$

Probability of partial world is product of CPDs
if the world includes all the active parents
of the variables it instantiates

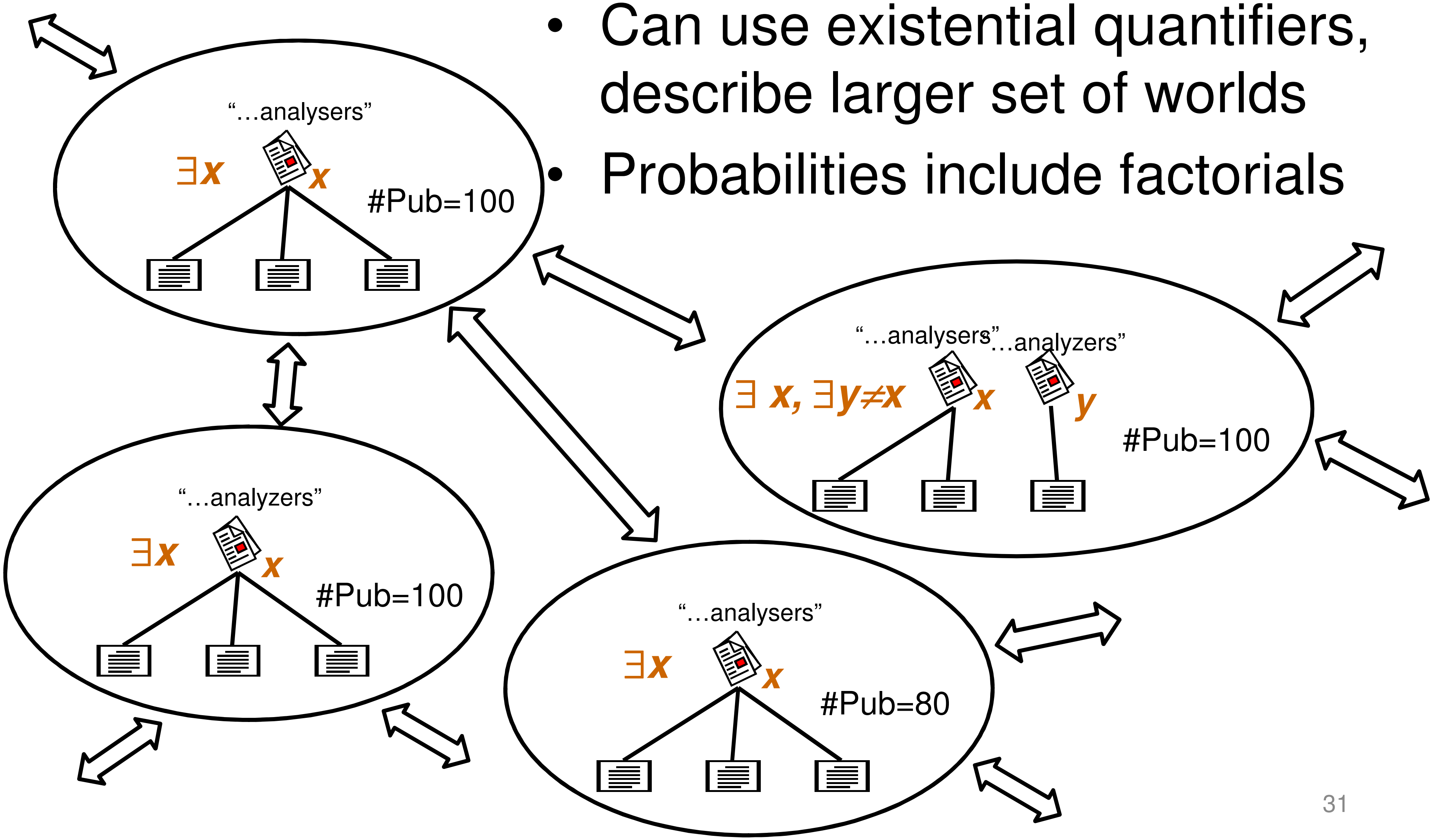
Numbering of Objects?

- Specifying publication numbers is unnecessary and inefficient

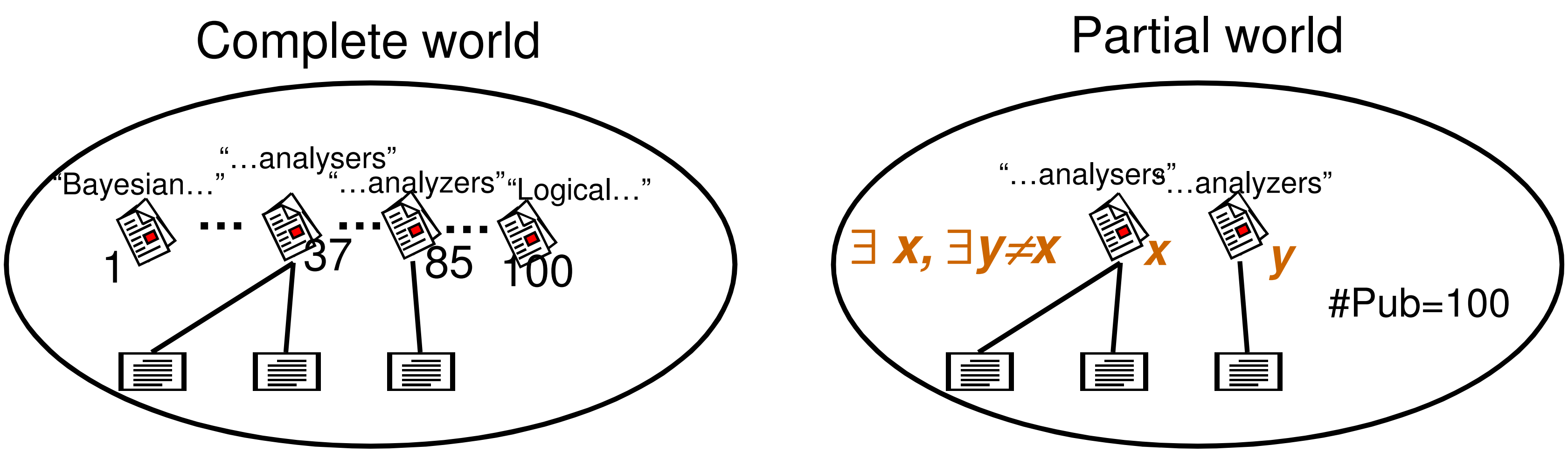


Abstraction over Objects

- Can use existential quantifiers, describe larger set of worlds
- Probabilities include factorials



MCMC over Partial Worlds



- Derived conditions under which generic MCMC engine can use partial worlds
- Engine avoids spending time changing irrelevant aspects of world

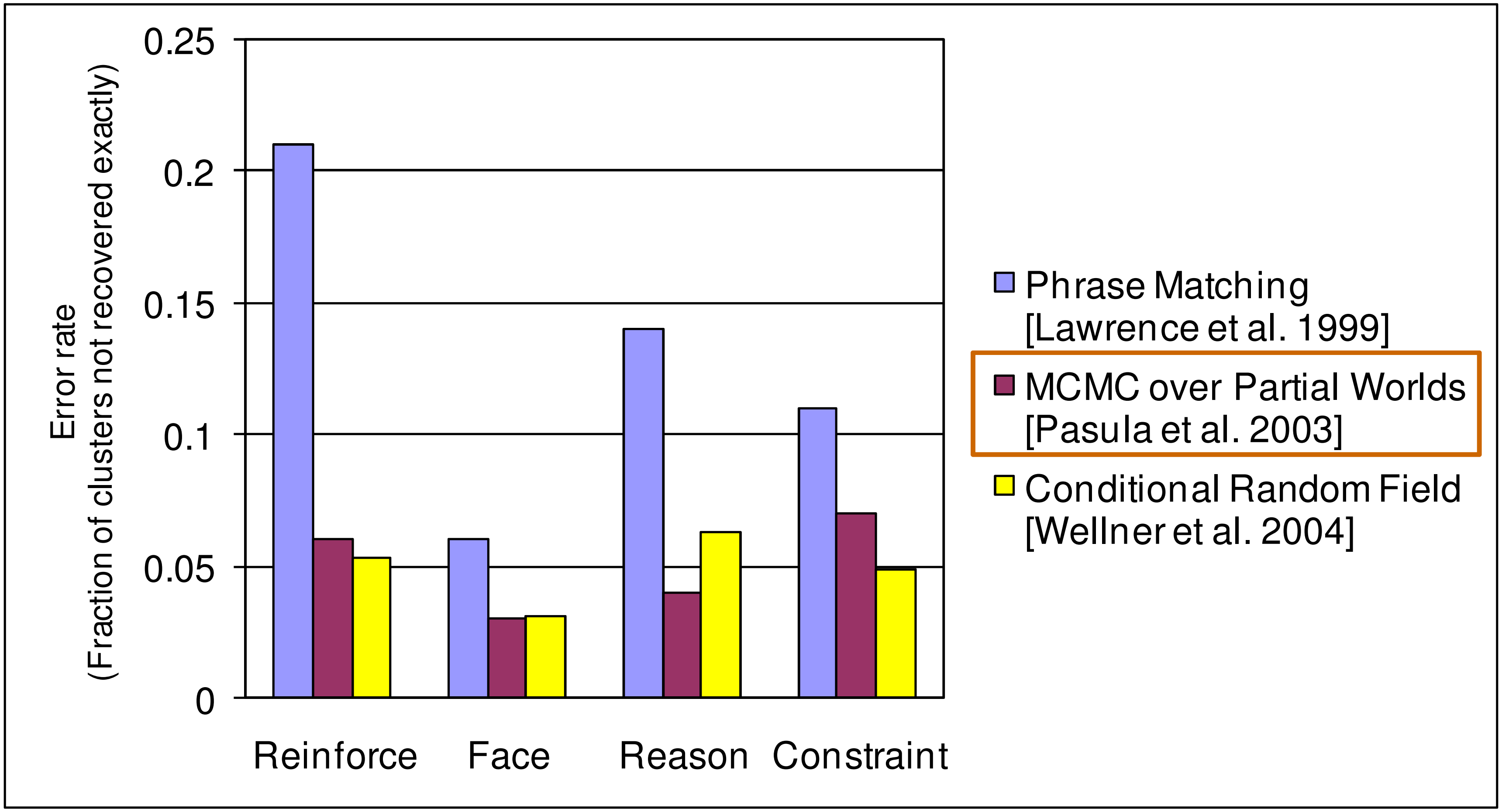
Experiment: Citation Data Set

- Citeseer created four sets of 300-500 citations by searching for certain strings ("face", "reinforcement", "reasoning", "constraint")
- Lots of noise in citation strings

```
[Lashkari et al 94] Collaborative Interface Agents, Yezdi
Lashkari, Max Metral, and Pattie Maes, Proceedings of the
Twelfth National Conference on Articial Intelligence, MIT
Press, Cambridge, MA, 1994.
```

```
Metral M. Lashkari, Y. and P. Maes. Collaborative interface
agents. In Conference of the American Association for
Artificial Intelligence, Seattle, WA, August 1994.
```

Experiment: Accuracy Results

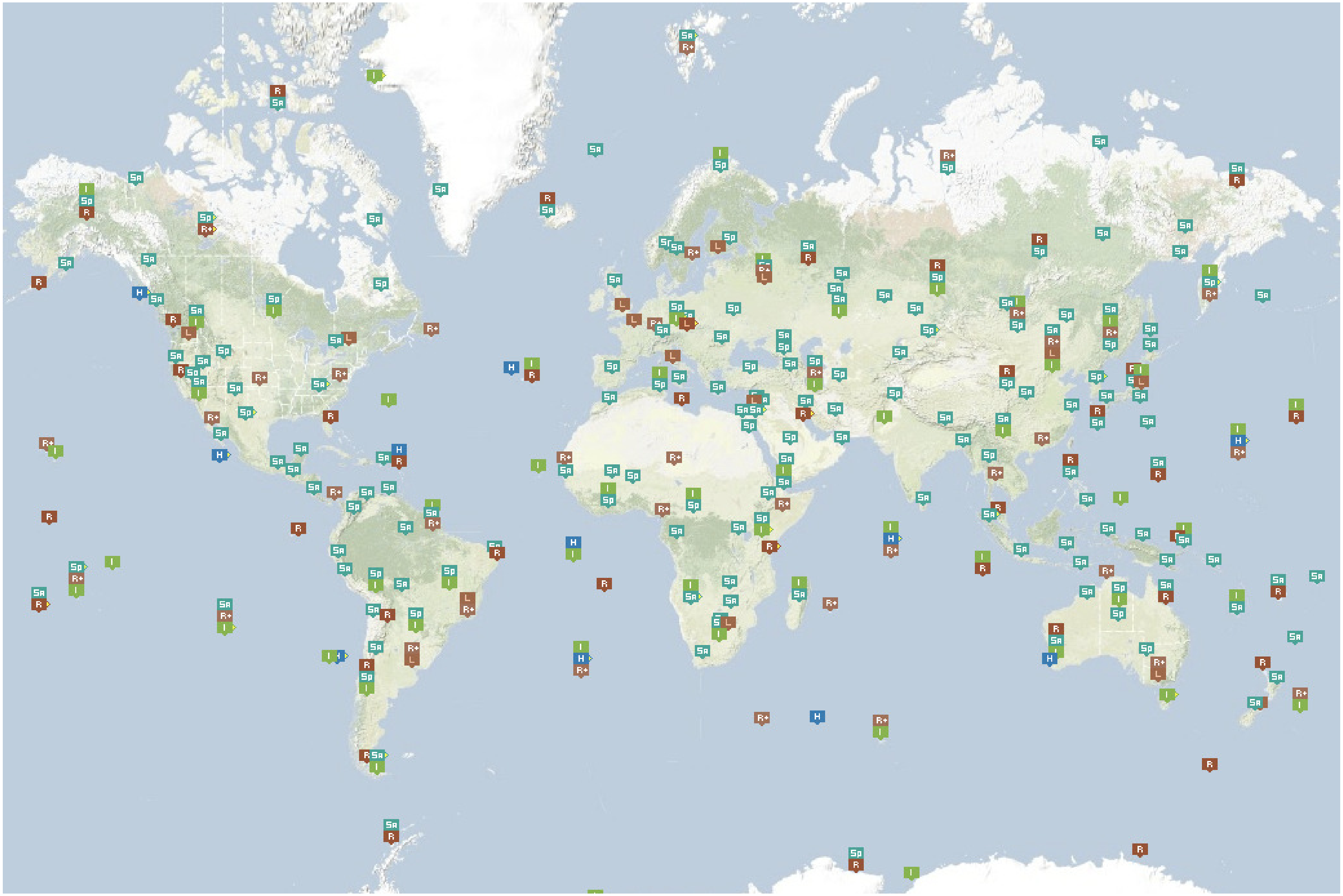


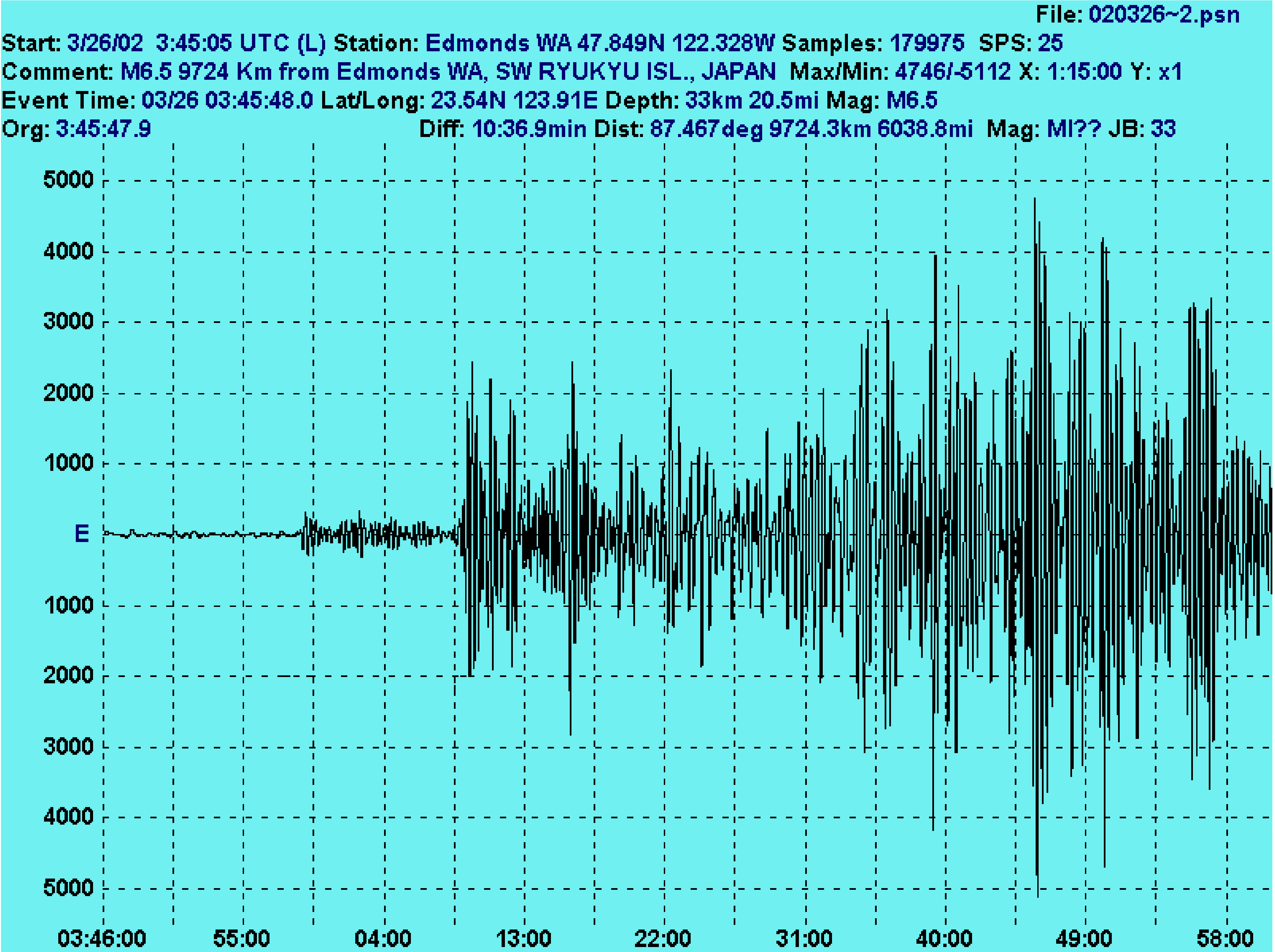
Four data sets of ~300-500 citations, referring to ~150-300 papers

Seismological Monitoring

- Comprehensive Nuclear Test Ban Treaty (CTBT) organization responsible for detecting nuclear explosions
- Detect seismological events, then figure out if earthquakes or nuclear tests
- Vertically Integrated Seismic Analysis (VISA) project using BLOG-style model [Arora et al., NIPS 2010]

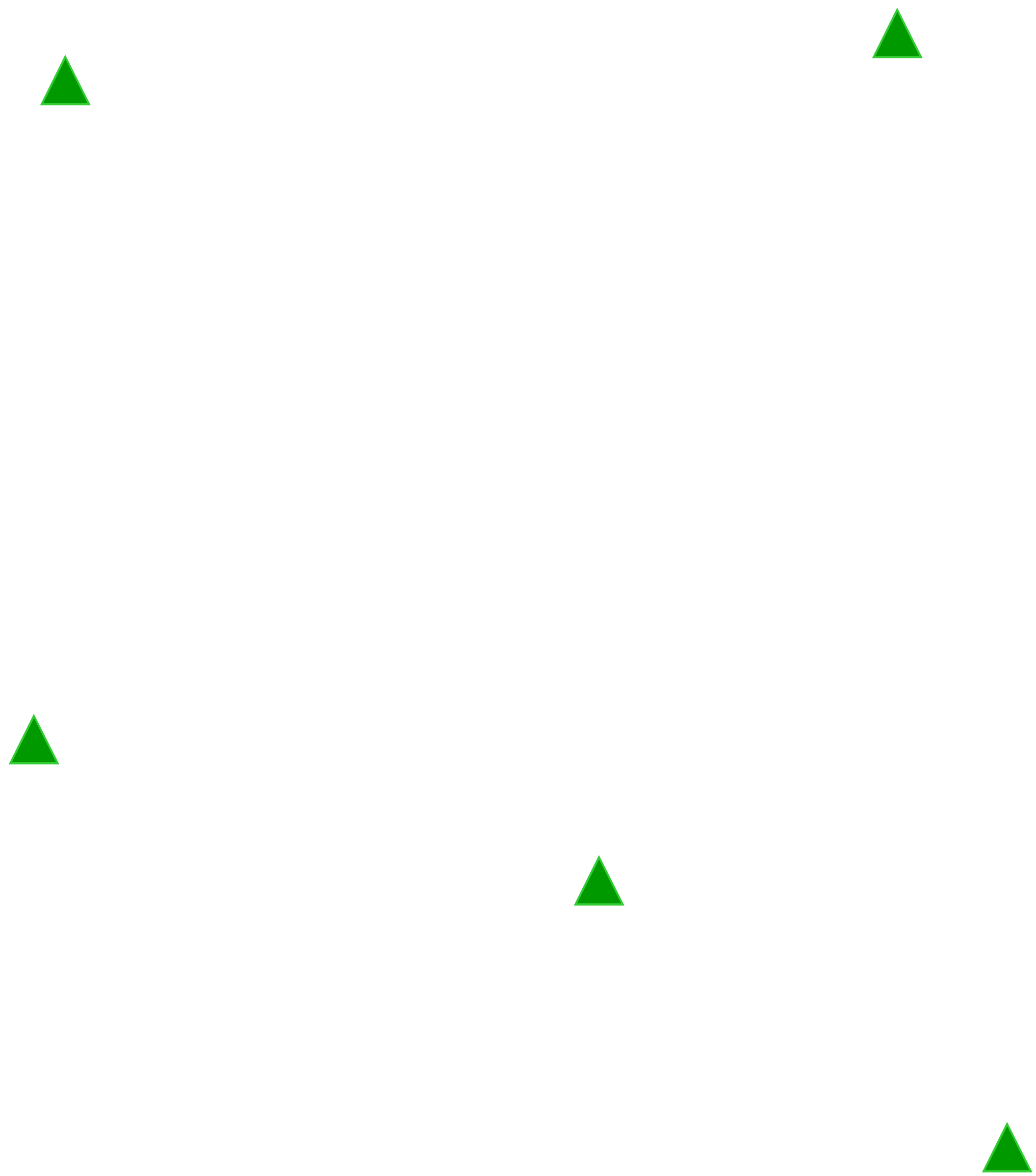
254 monitoring stations

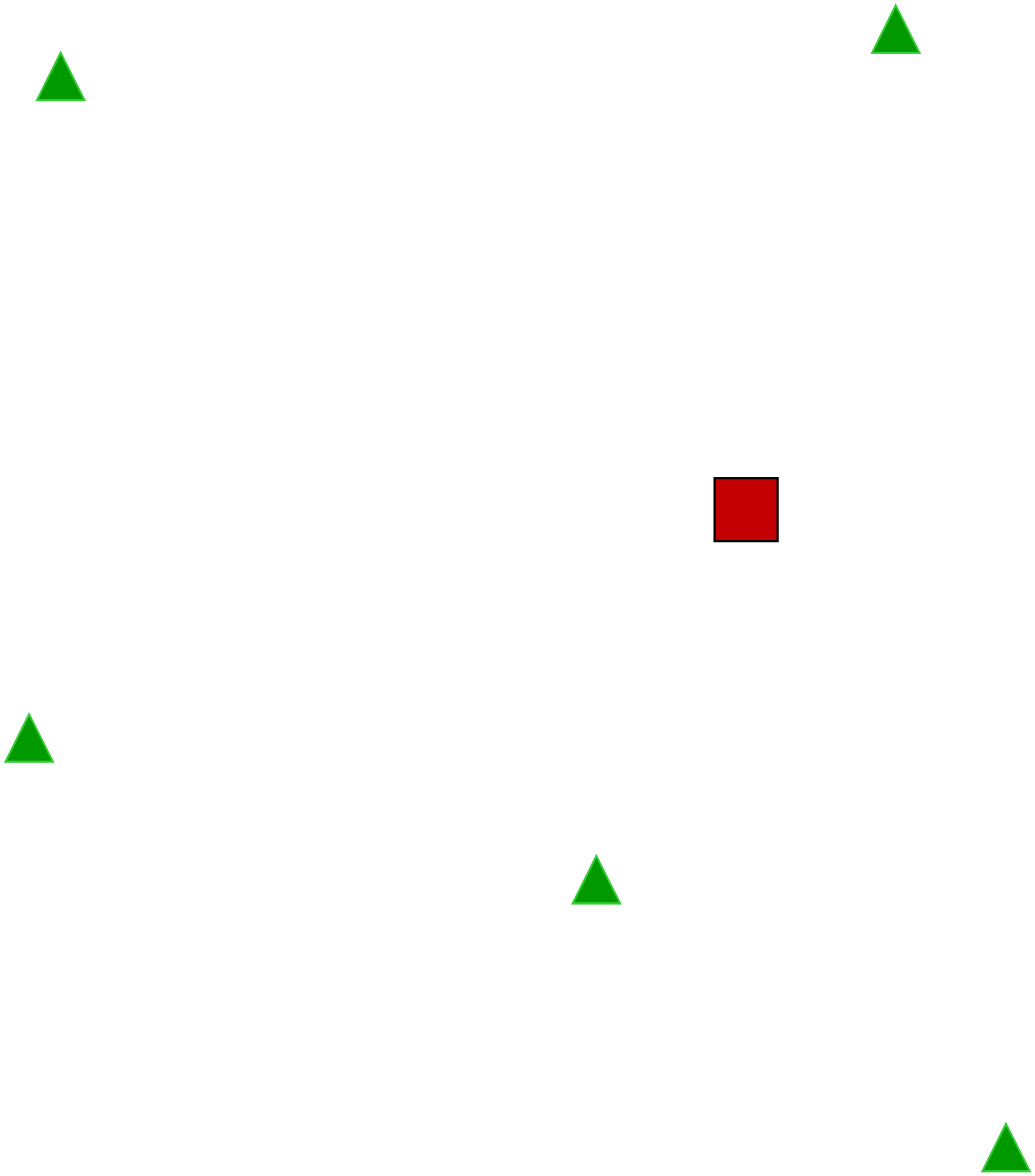


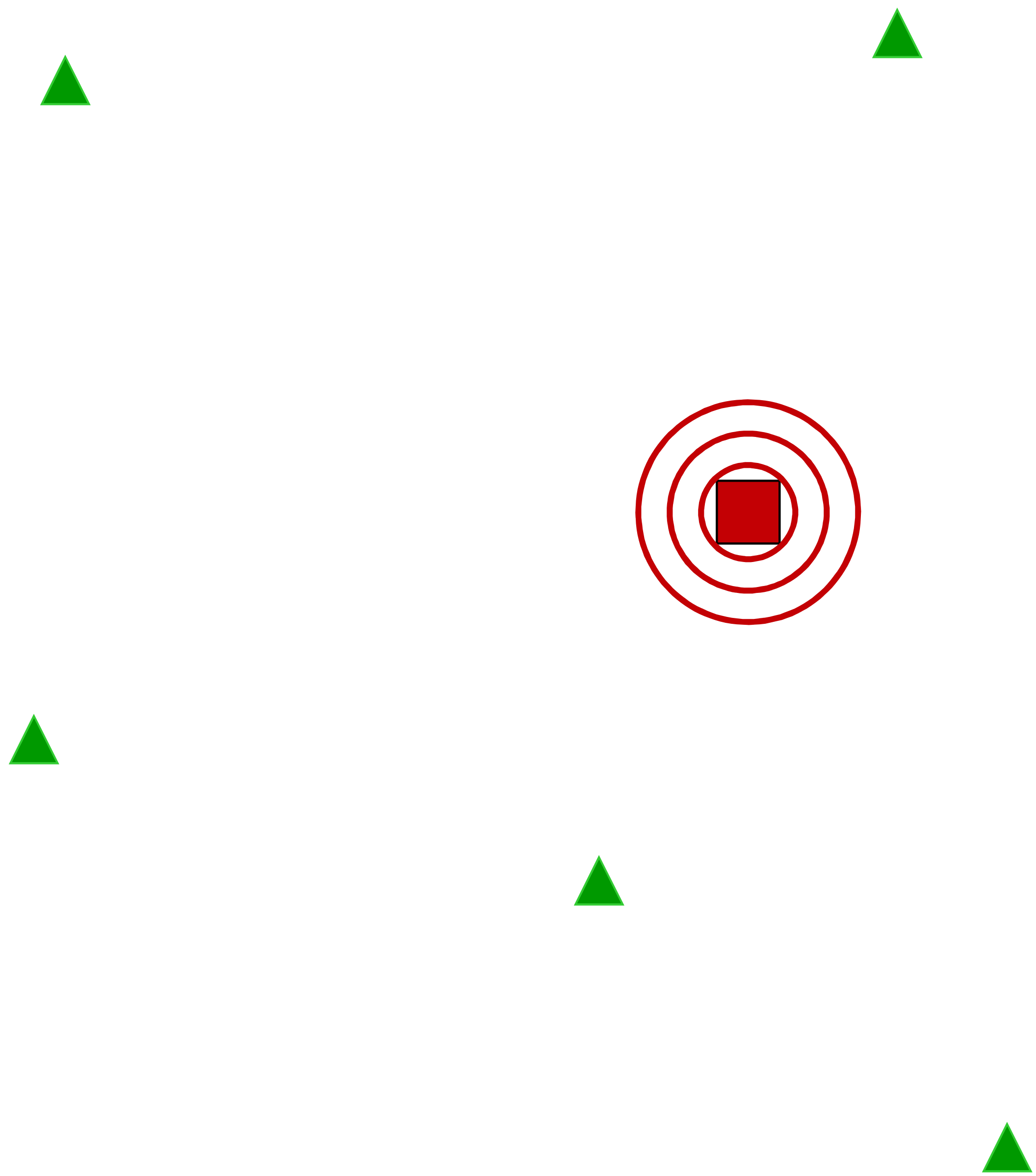

File: 020326~2.psn
Start: 3/26/02 3:45:05 UTC (L) Station: Edmonds WA 47.849N 122.328W Samples: 179975 SPS: 25
Comment: M6.5 9724 Km from Edmonds WA, SW RYUKYU ISL., JAPAN Max/Min: 4746/-5112 X: 1:15:00 Y: x1
Event Time: 03/26 03:45:48.0 Lat/Long: 23.54N 123.91E Depth: 33km 20.5mi Mag: M6.5
Org: 3:45:47.9 Diff: 10:36.9min Dist: 87.467deg 9724.3km 6038.8mi Mag: Ml?? JB: 33
5000
4000
3000
2000
1000
E
1000
2000
3000
4000
5000
03:46:00
55:00
04:00
13:00
22:00
31:00
40:00
49:00
58:00

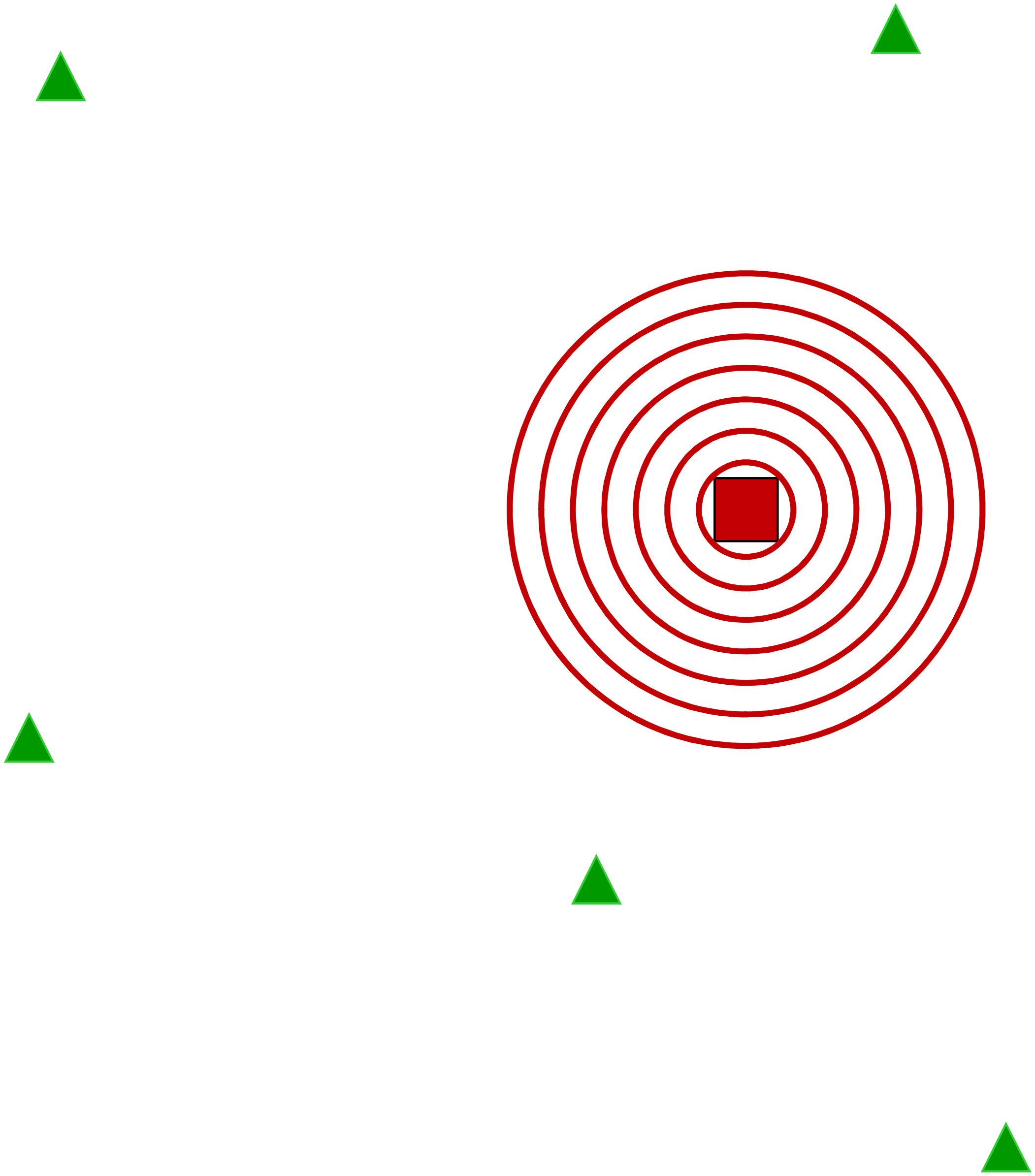
Vertically Integrated Seismic Analysis

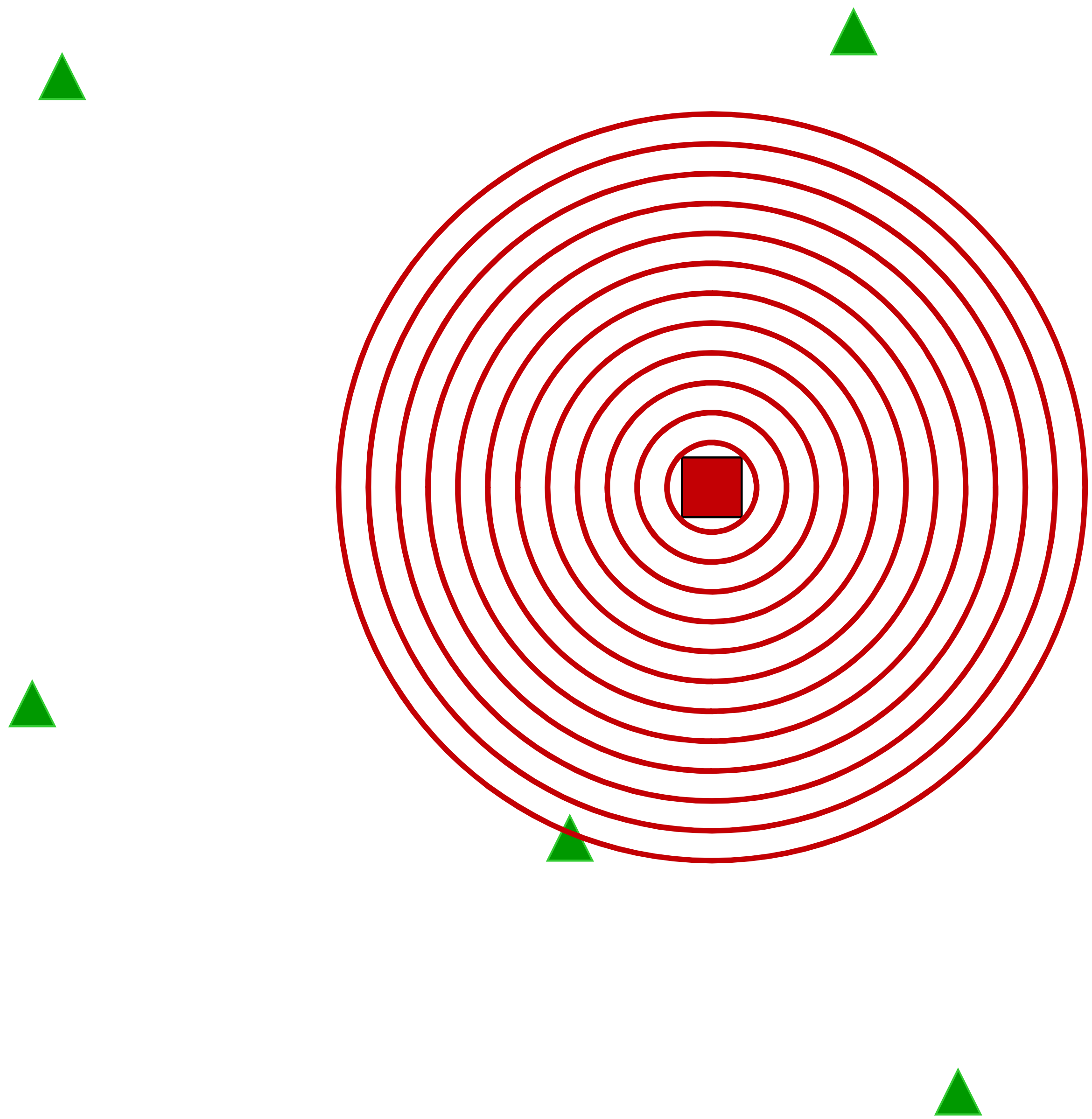
- The problem is hard:
  - ~10000 “detections” per day, 90% false
  - CTBT system (SEL3) finds 69% of significant events plus about twice as many spurious (nonexistent) events
  - 16 human analysts find more events, correct existing ones, throw out spurious events, generate LEB (“ground truth”)
  - Unreliable below magnitude 4 (1kT)
- Solve it by global probabilistic inference
  - NET-VISA finds around 88% of significant events

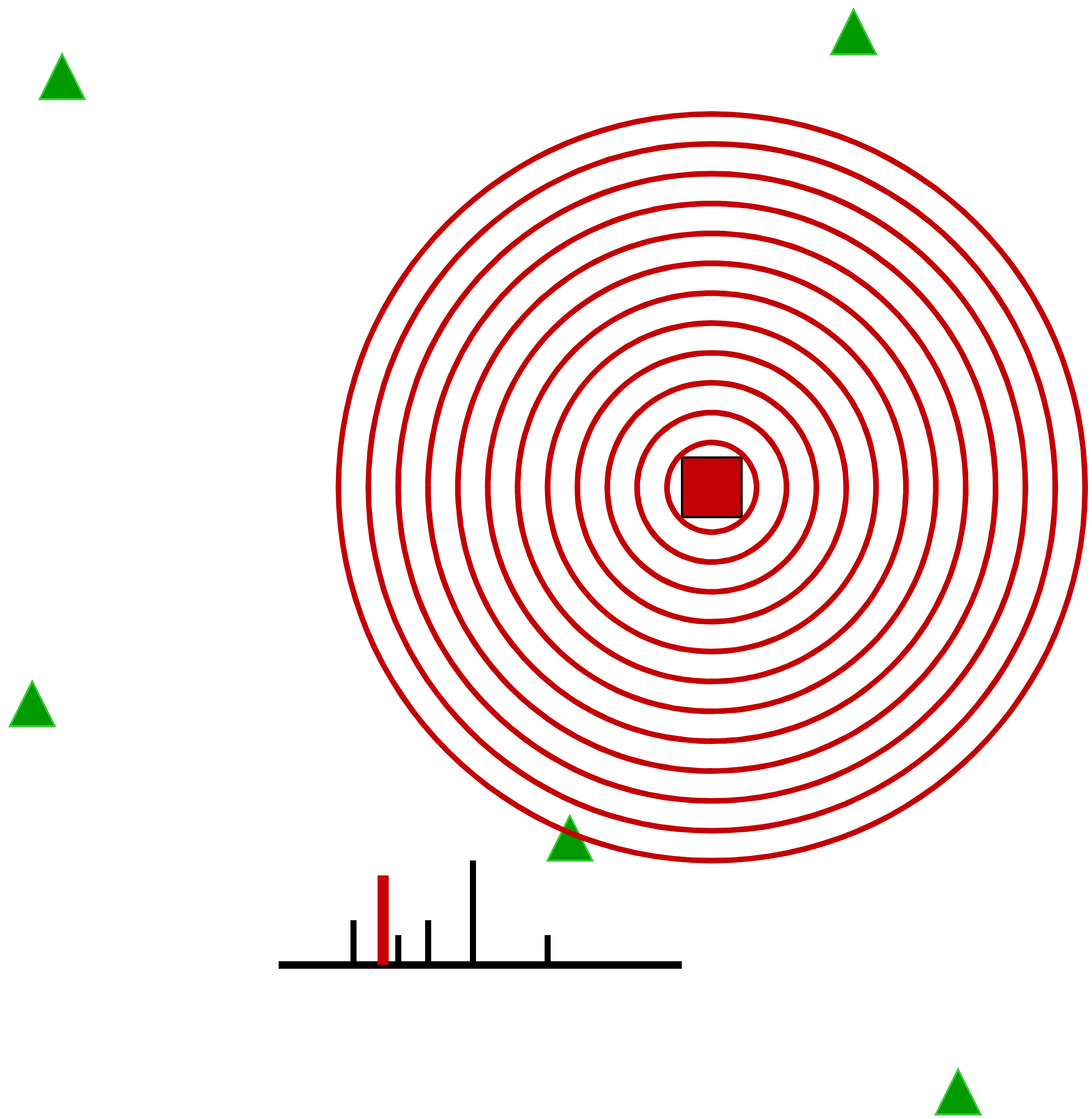


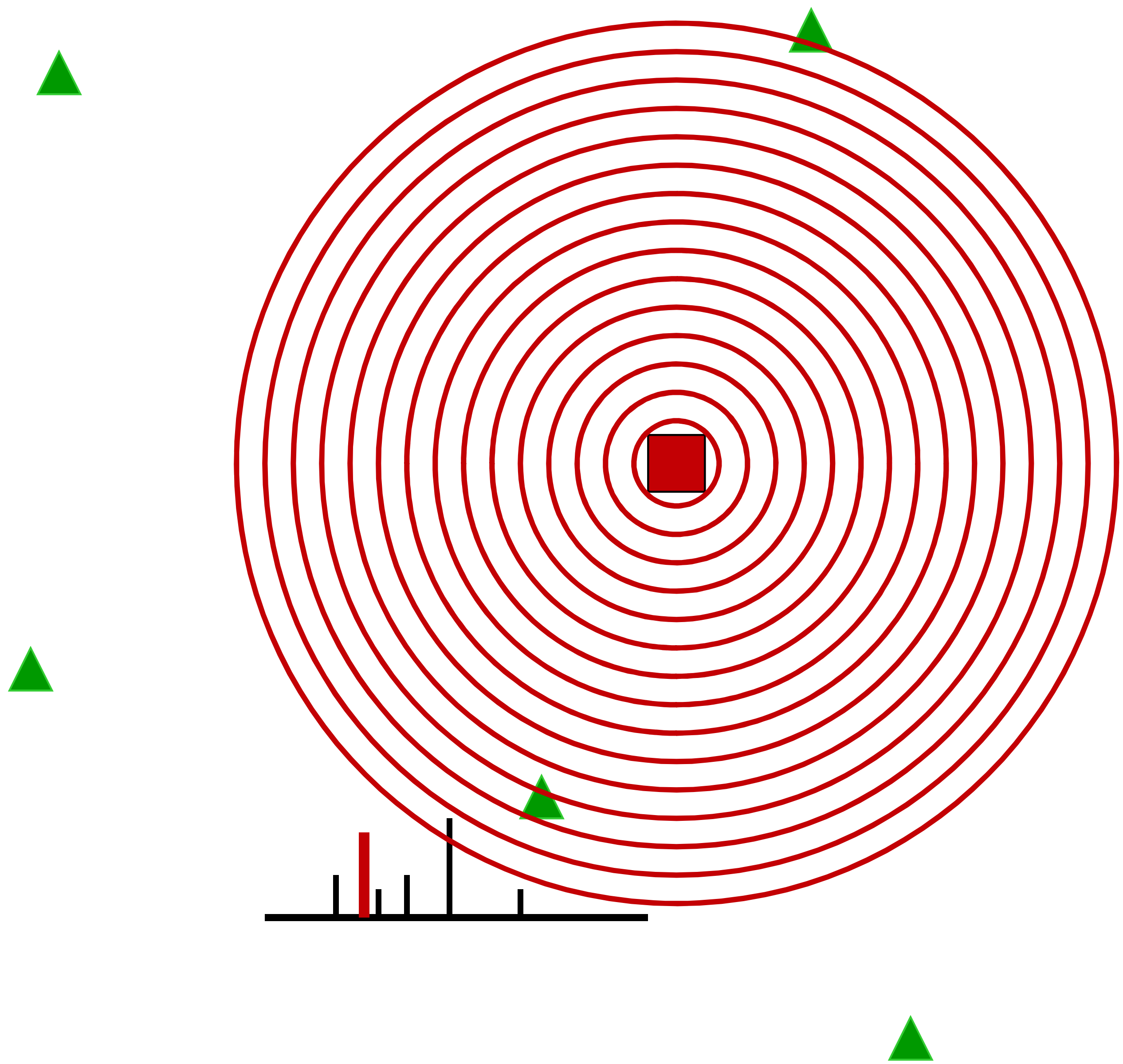


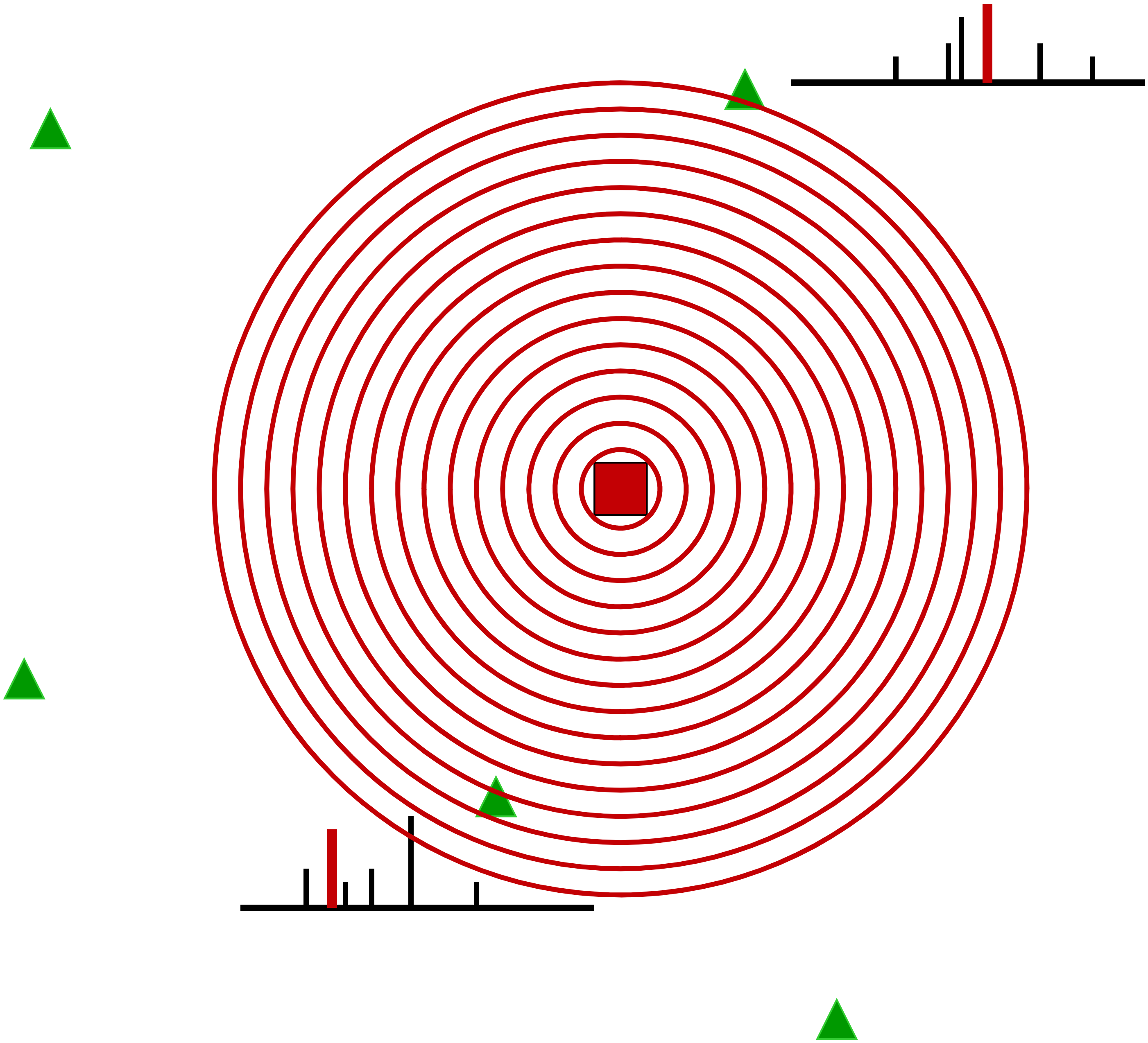


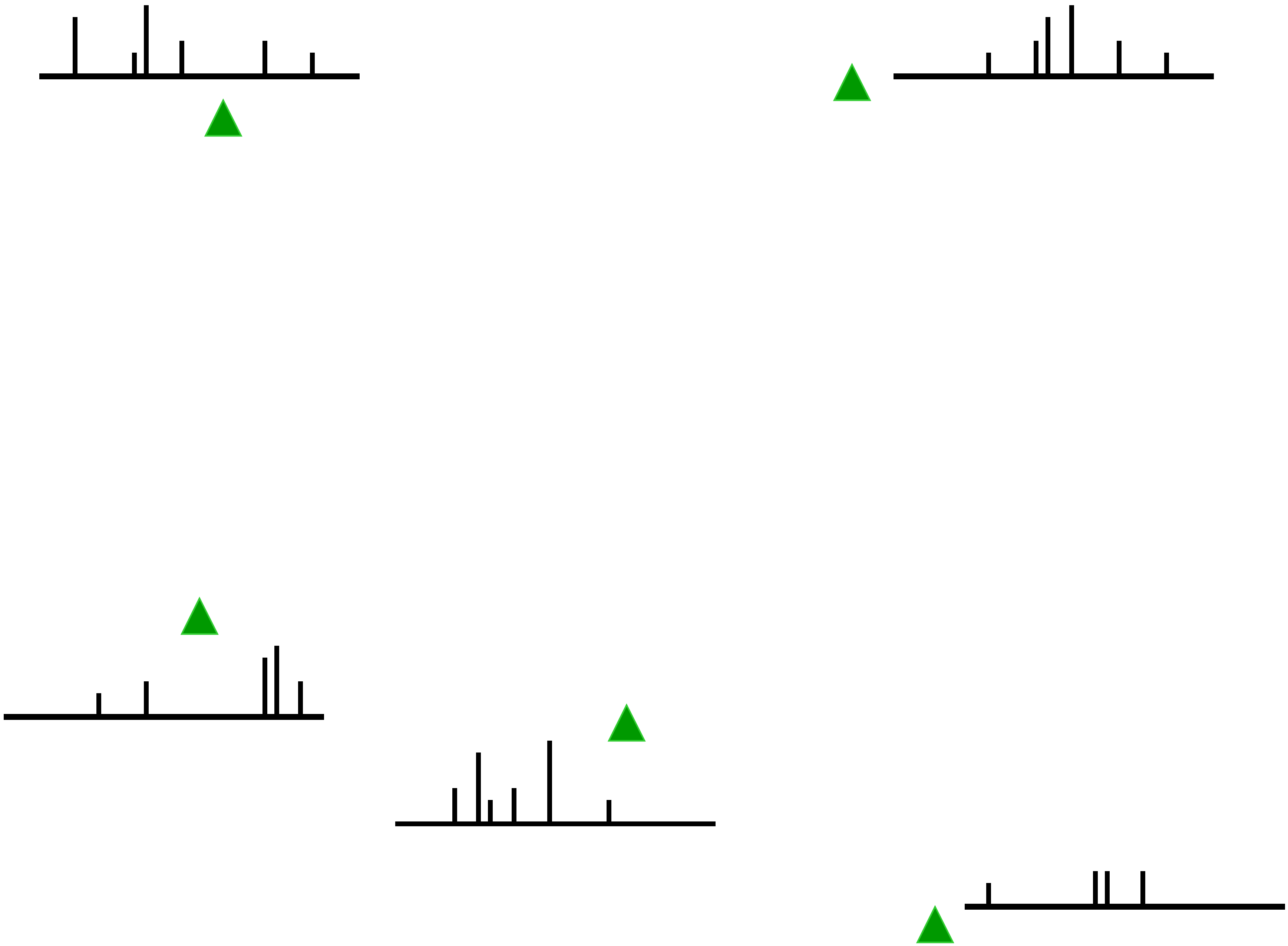










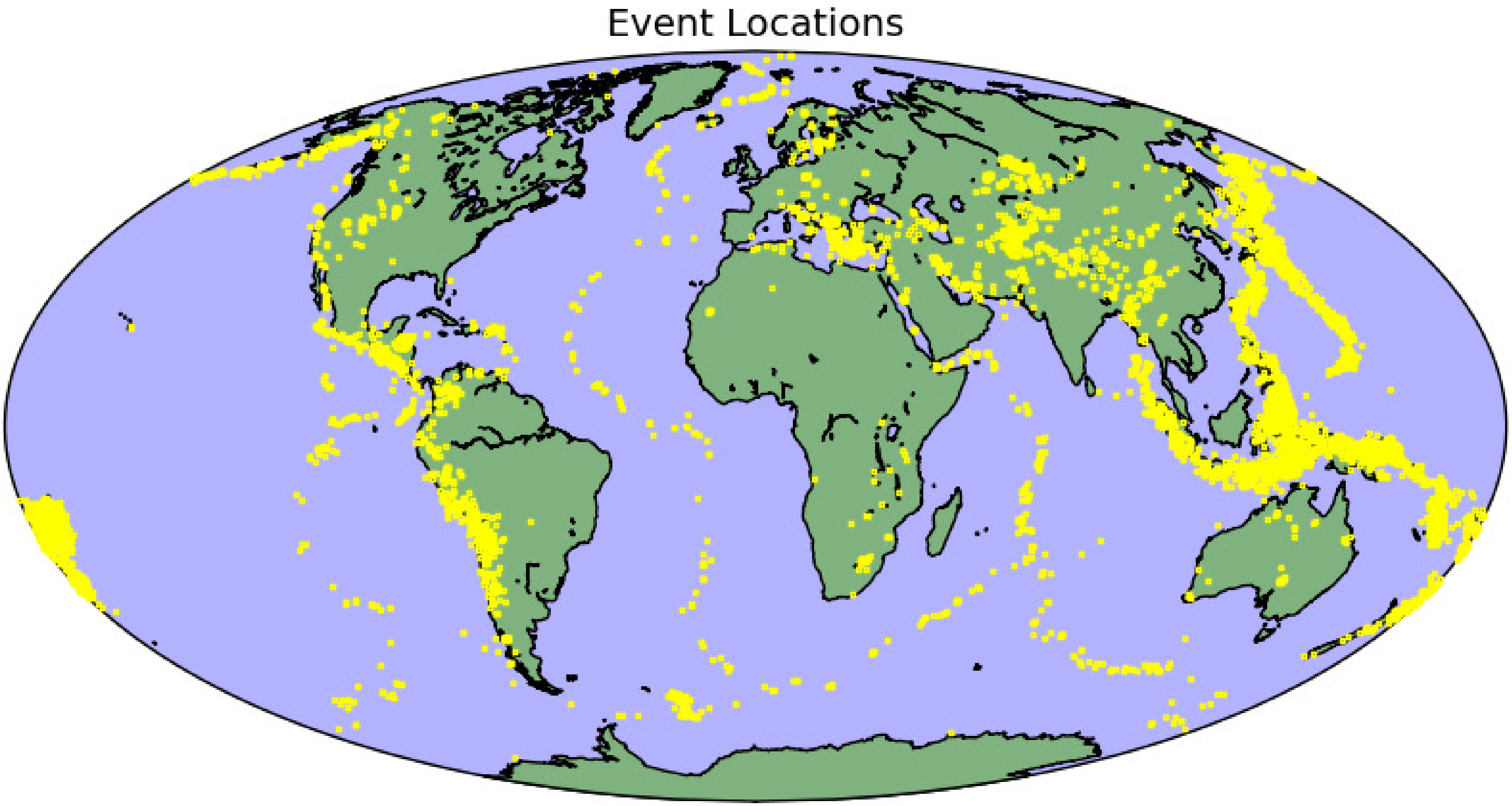


Generative model for detections

- Events occur in time and space with magnitude
  - Natural spatial distribution a mixture of Fisher-Binghams
  - Man-made spatial distribution uniform
  - Time distribution Poisson with given spatial intensity
  - Magnitude distribution Gutenberg-Richter (exponential)
  - Aftershock distribution (not yet implemented)
- Travel time according to IASPEI91 model plus Laplacian error distribution for each of 14 *phases*
- Detection depends on magnitude, distance, station
- Detected azimuth, slowness have Laplacian error
- False detections with station-dependent distribution

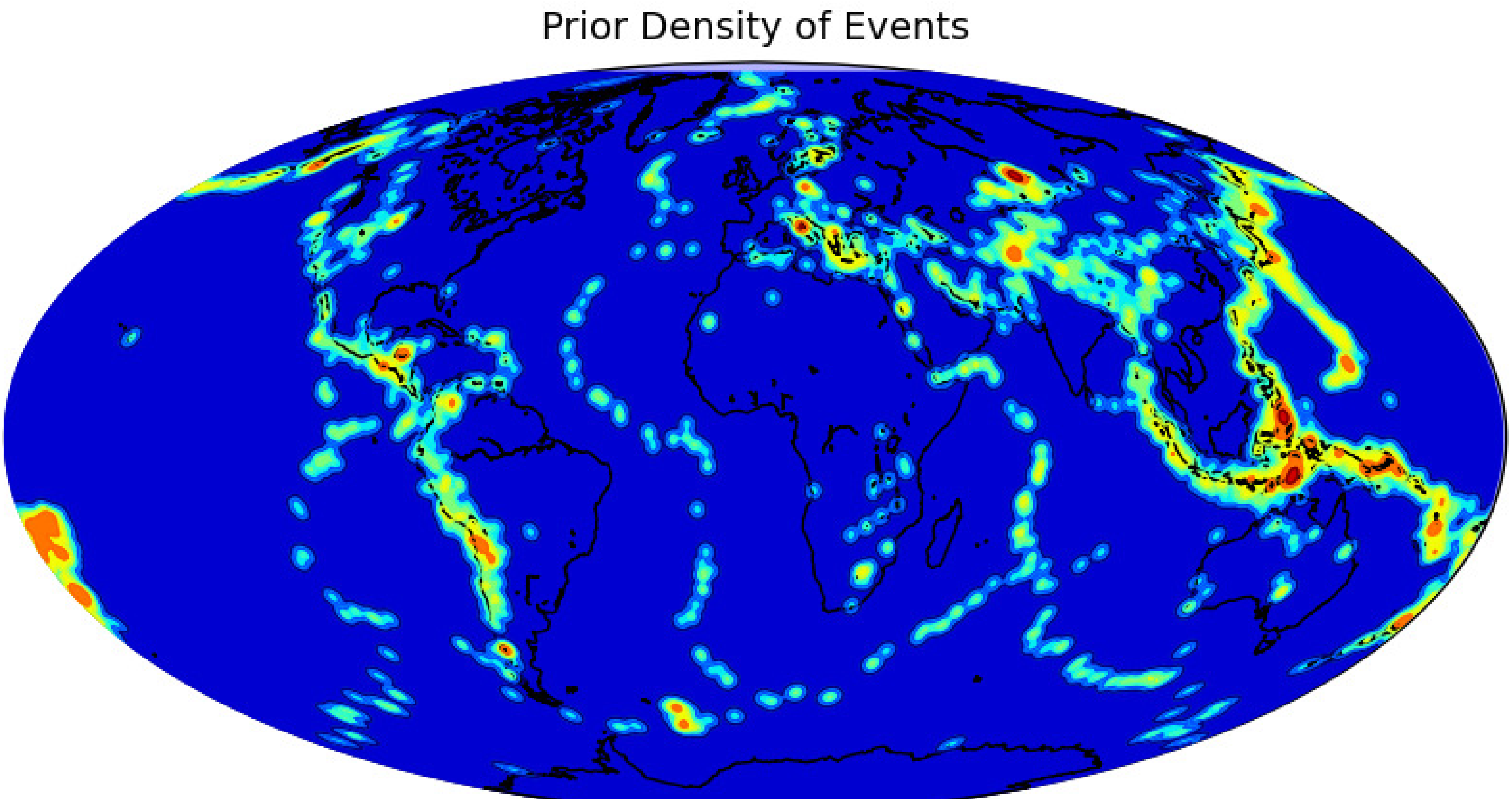
Model for Seismic Monitoring

 # SeismicEvents ~ Poisson[TIME_DURATION*EVENT_RATE];
IsEarthQuake(e) ~ Bernoulli(.999);
EventLocation(e) ~ If IsEarthQuake(e) then EarthQuakeDistribution()
Else UniformEarthDistribution();
Magnitude(e) ~ Exponential(log(10)) + MIN_MAG;
Distance(e,s) = GeographicalDistance(EventLocation(e), SiteLocation(s));
***IsDetected(e,p,s)* ~** Logistic[SITE_COEFFS(s,p)](Magnitude(e), Distance(e,s)**;**
#Arrivals(site = s) ~ Poisson[TIME_DURATION*FALSE_RATE(s)];
#Arrivals(event=e, phase = p, site = s) = If IsDetected(e,p,s) then 1 else 0;
***Time(a)* ~** If (event(a) = null) then Uniform(0,TIME_DURATION)
else IASPEI(EventLocation(event(a)),SiteLocation(site(a)),Phase(a)) + TimeRes(a);
TimeRes(a) ~ Laplace(TIMLOC(site(a)), TIMSCALE(site(a)));
***Azimuth(a)* ~** If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(EventLocation(event(a)),SiteLocation(site(a)) + AzRes(a);
AzRes(a) ~ Laplace(0, AZSCALE(site(a)));
***Slow(a)* ~** If (event(a) = null) then Uniform(0,20)
else IASPEI-SLOW(EventLocation(event(a)),SiteLocation(site(a)) + SlowRes(site(a));


Event Locations

Prior on Event Locations

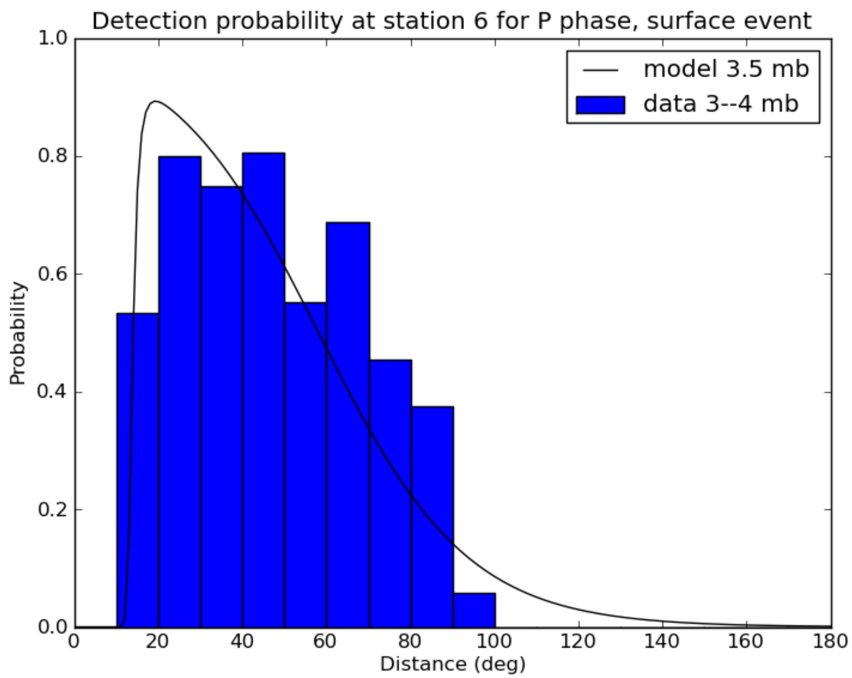
EventLocation(e) **~** If IsEarthQuake(e) then EarthQuakeDistribution()



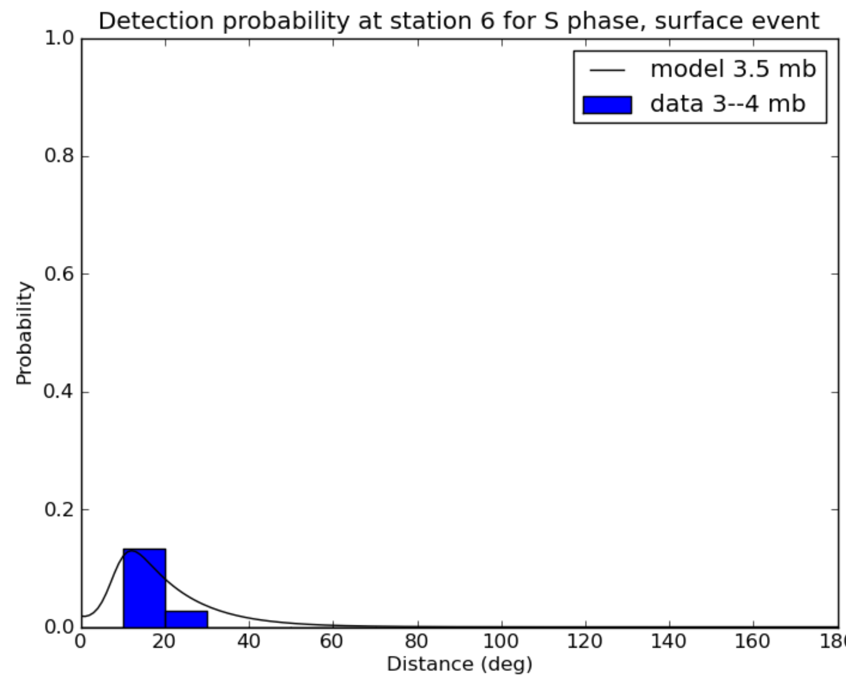
Detection Probability

IsDetected(e,p,s) **~** Logistic[SITE_COEFFS(s,p)](Magnitude(e), Distance(e,s);

P phase

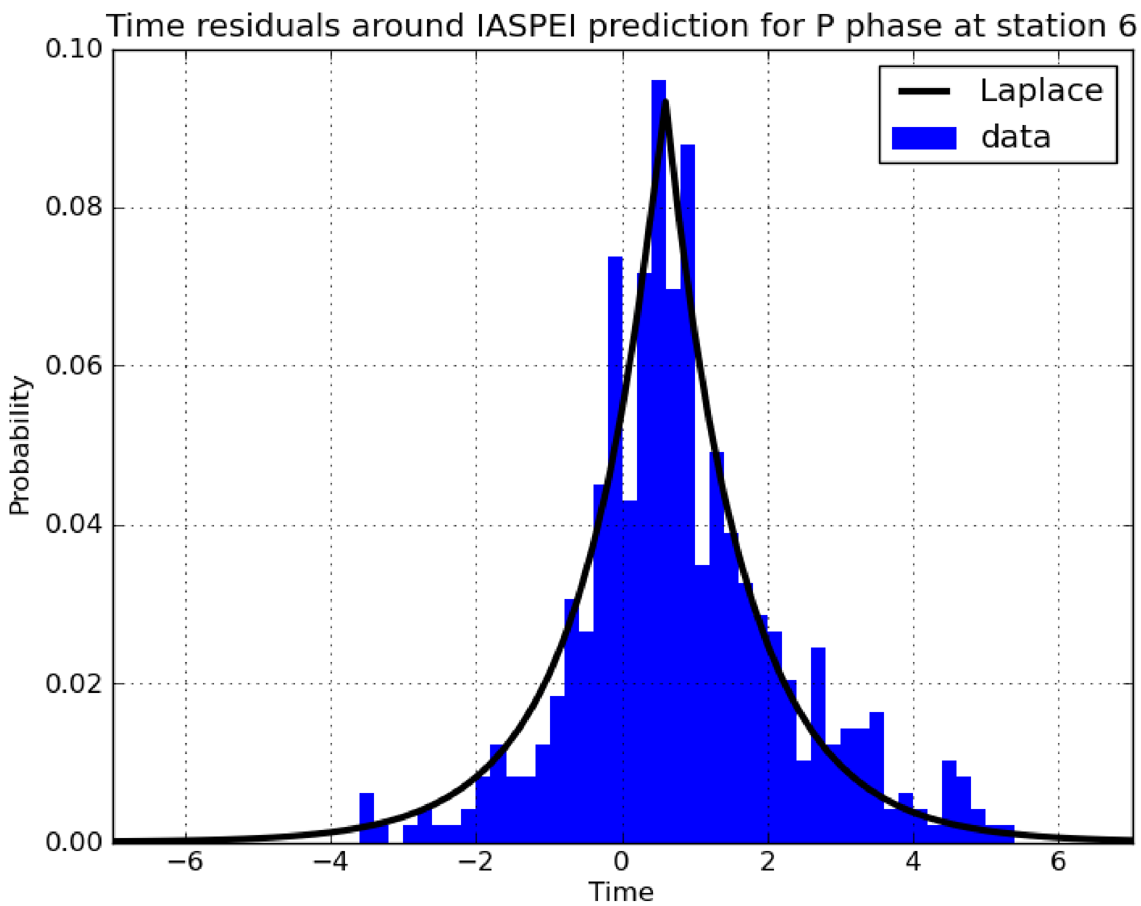


S phase



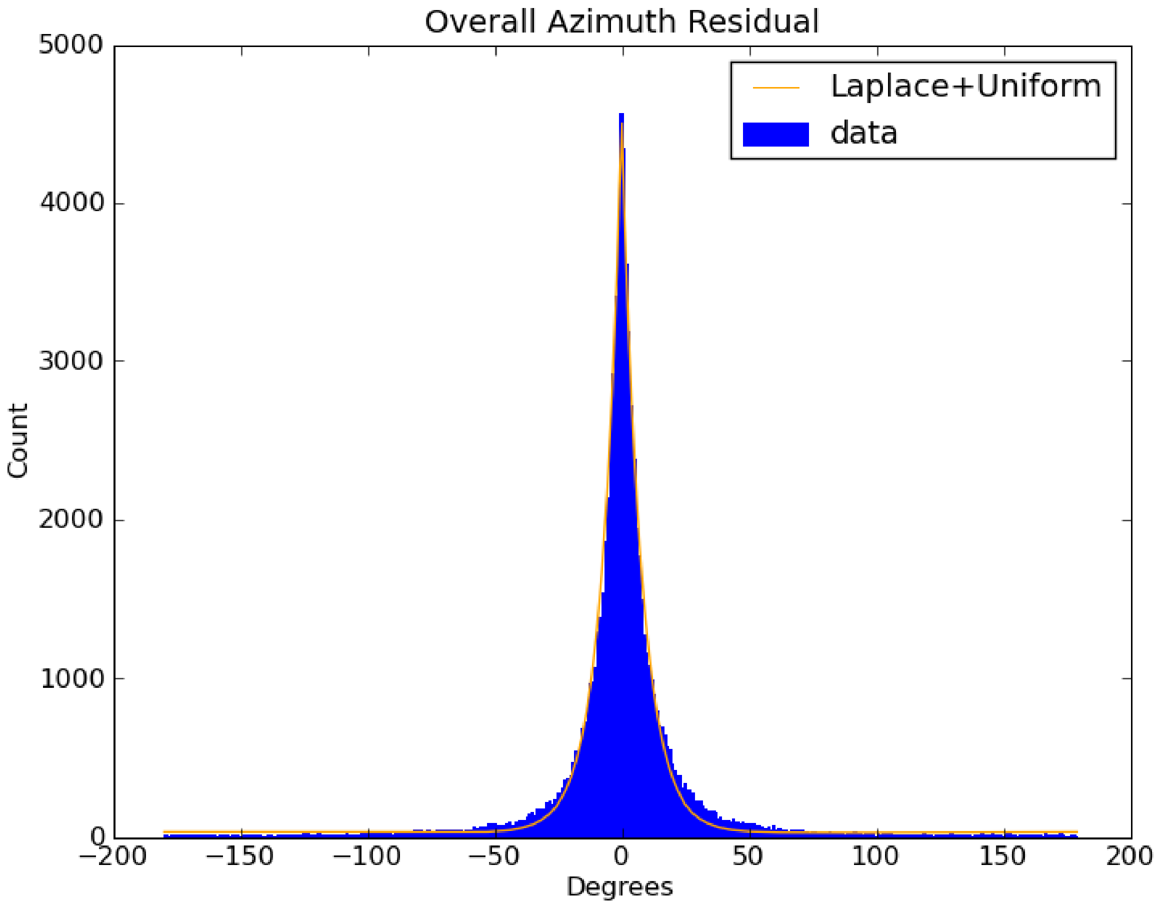
Travel time

Time(a) ~ IASPEI(EventLocation(event(a)),SiteLocation(site(a)),Phase(a))
+ TimeRes(a);

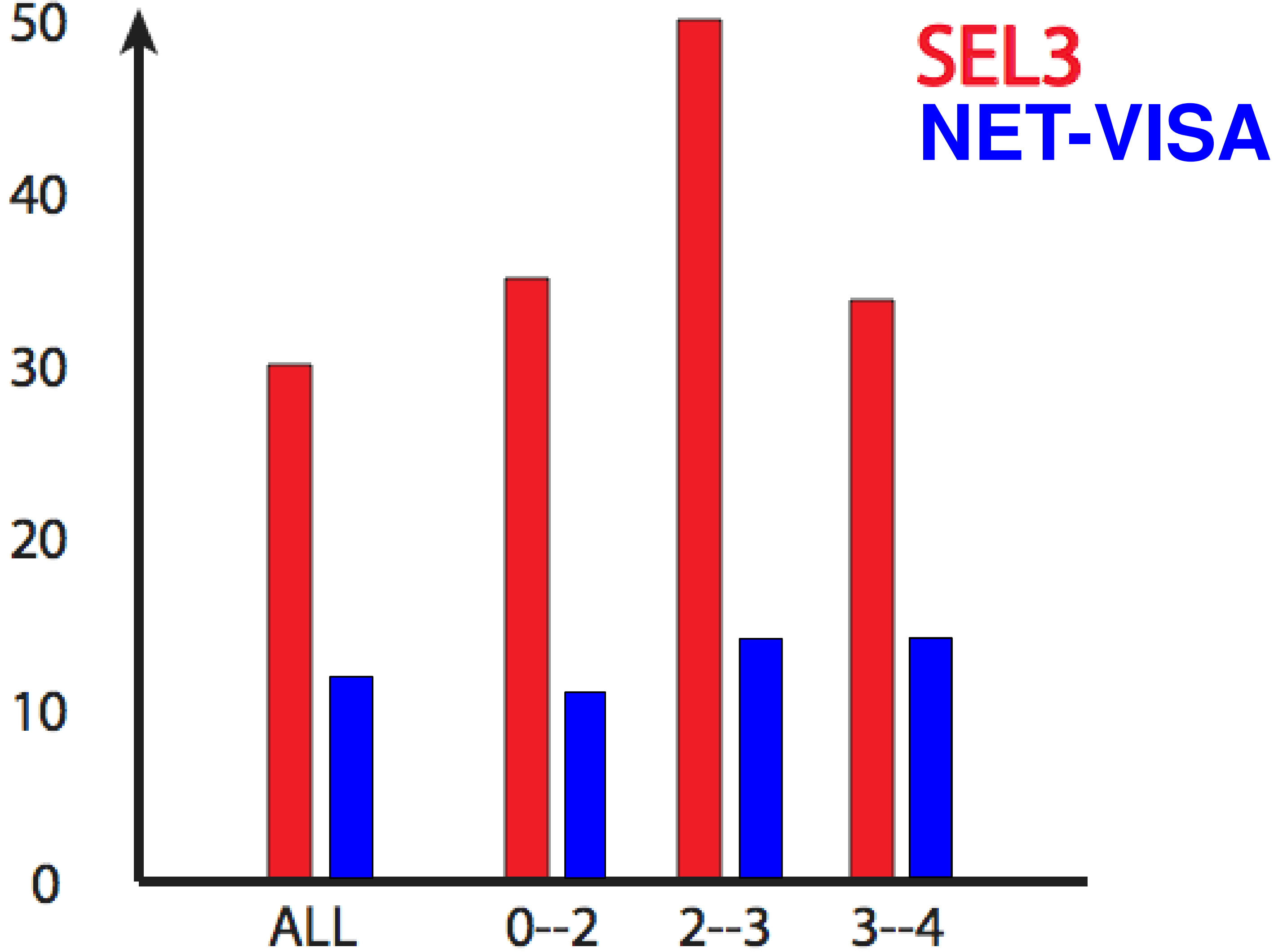


Detected Azimuth

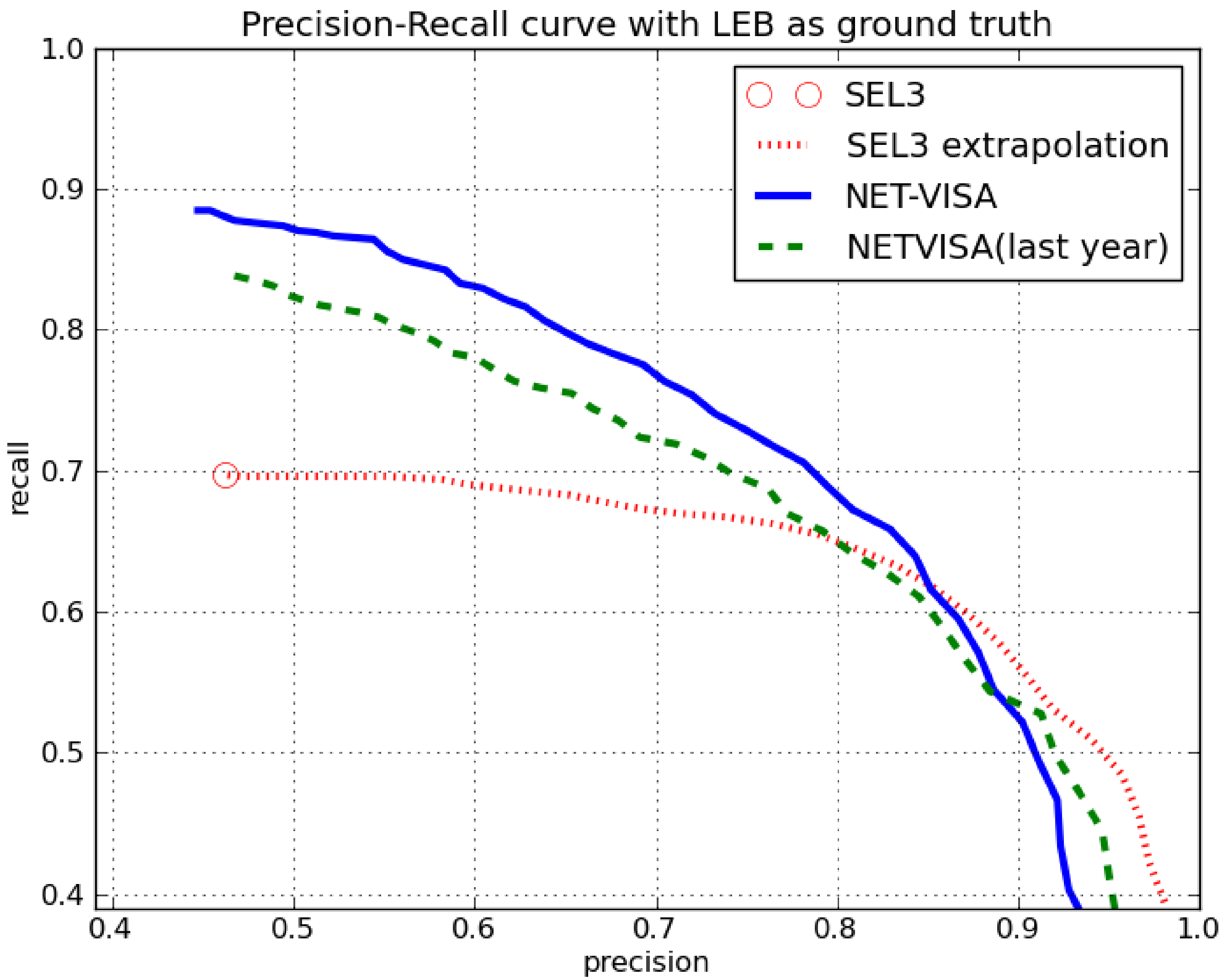
Azimuth(a) ~ GeoAzimuth(EventLocation(event(a)),SiteLocation(site(a))
+ AzRes(a);



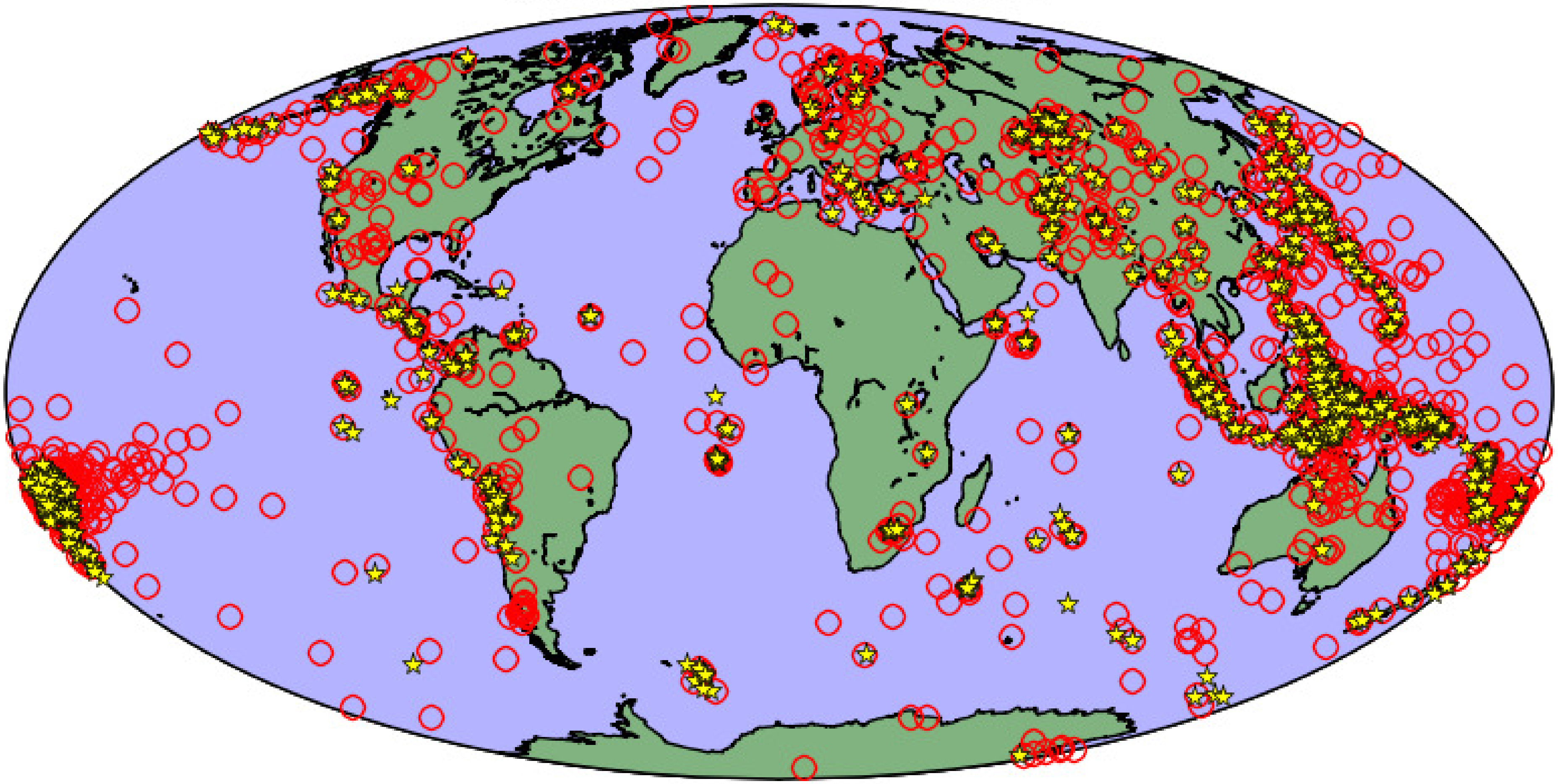
Fraction of LEB events missed



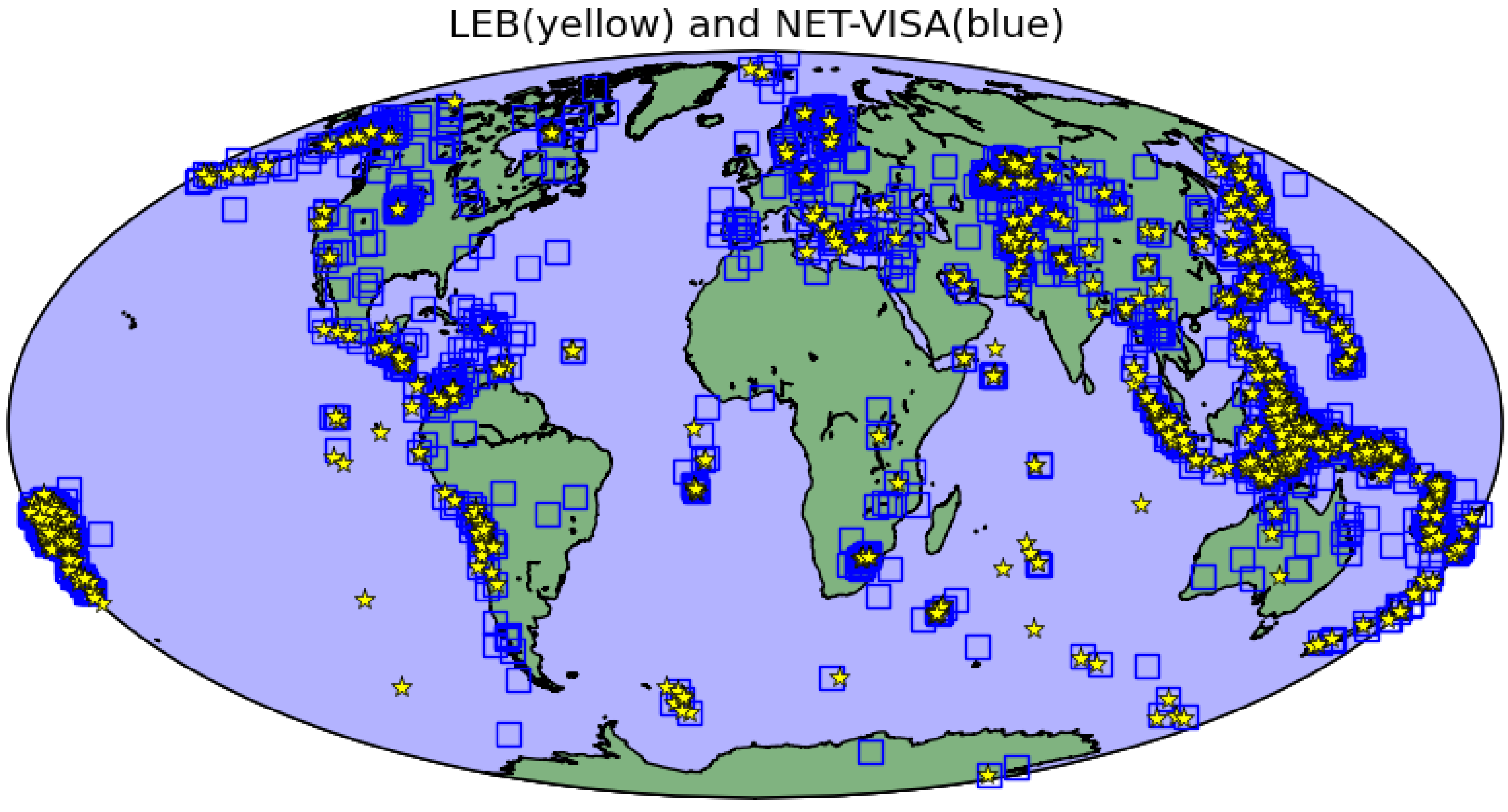
Precision-Recall Curve



Event distribution: LEB vs SEL3



Event distribution: LEB vs NET-VISA



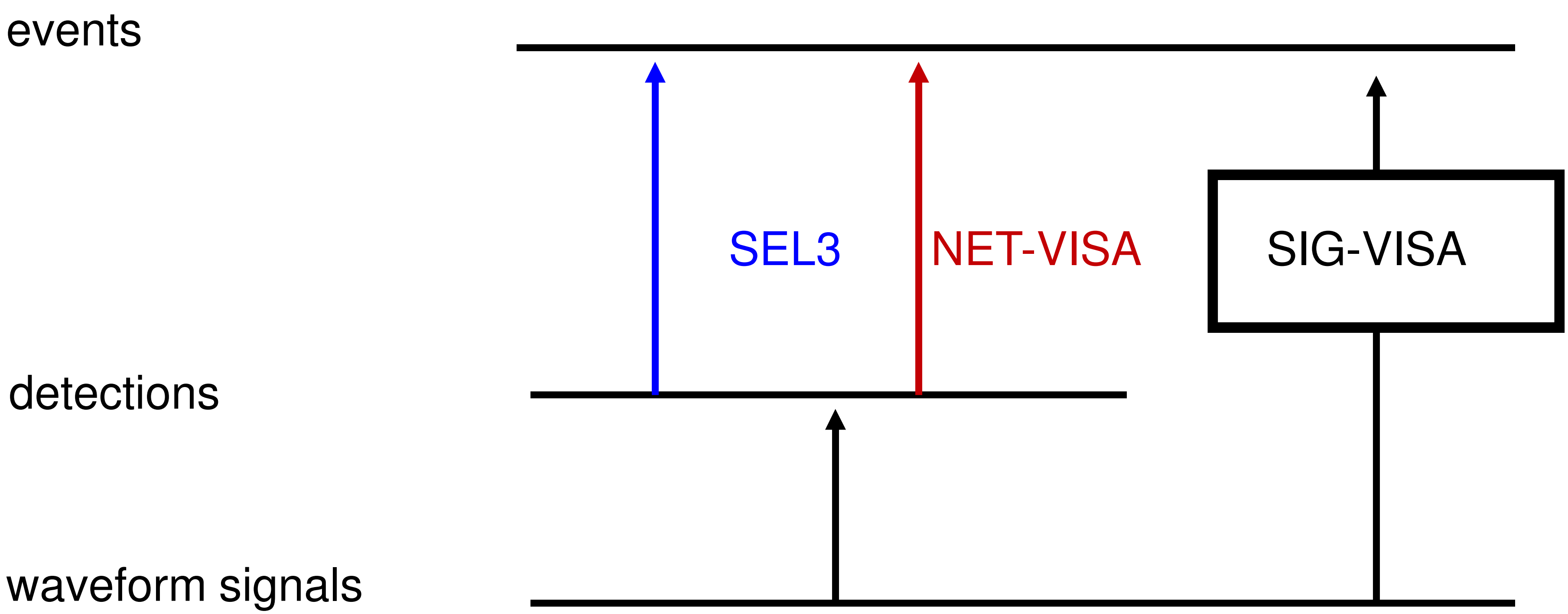
Why does NET-VISA work?

- Empirically calibrated *seismological* models
  - Improving model structure and quality improves the results
- Sound Bayesian combination of evidence
- Measured arrival times, phase labels, azimuths, etc., NOT taken as truth
- *Absence* of detections provides negative evidence
- More detections per event than SEL3 or LEB

Why does NET-VISA not work (perfectly)?

- Needs hydroacoustics for mid-ocean events
- Weaknesses in model:
  - Travel time residuals for *all phases* along a single path are assumed to be *uncorrelated*
  - *Each* phase arrival is assumed to generate *at most one* detection; in fact, *multiple detections* occur
- *Arrival detectors use high SNR thresholds, look only at local signal to make hard decisions*

Detection-based and signal-based monitoring



Summary

- CPD-based first-order probabilistic models
  - Describe generative processes for worlds
  - Generalize Bayesian networks
- BLOG adds support for unknown objects
- Can run MCMC over partial worlds
- Strong results when CPDs come from science, as in seismological monitoring