Probabilistic Programming and AI

Stuart Russell
Computer Science Division, UC Berkeley

Deep learning ad infinitum?

François Chollet: “Many more applications are completely out of reach for current deep learning techniques – even given vast amounts of human-annotated data.

…

The main directions in which I see promise are models closer to general-purpose computer programs.”

AI: intelligent systems in the real world

The world has things in it!!

The world is uncertain!!

GOFAI: first-order logic

Modern AI: Bayes nets

The world is uncertain!!

The world has things in it!!

first-order probabilistic languages

The world has unknown things in it!!

first-order open-universe probabilistic languages

AI: intelligent systems in the real world

The world is uncertain!!

Modern AI: graphical models

Bayes nets are machines for generating samples

Stochastic programs are bigger and better machines for generating samples

probabilistic programming languages

Expressive power

- Inexpressive language (e.g., feedforward circuits) =>
  - Very large function representations
  - Very little generalization, very slow learning
- Ditto for time-efficient languages (e.g., linear-time)
- DL community may be repeating history
  - 1960s BOFAI ignored computational intractability
  - 2010s DL seems to be ignoring sample intractability

BLOG

- Elements of probability space are the "model structures" (possible worlds) of first-order logic
  - Objects/functions/relations, not traces with state
  - Different worlds can have different objects, names, relational structures, etc.: ***open-universe models***
- Every well-formed BLOG model defines a proper distribution over possible worlds
- Sound and complete* Monte Carlo inference
  - Constant time per sample (locality, incrementalization)
  - Sequential MC for DBLOG inference

Citation information extraction

- Given: a set of text strings from reference lists:
  - [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
- Decide:
  - What papers and researchers exist
  - For each paper
    - The real title
    - The real authors
    - The papers it cites

BLOG model (single-author)

```
#Researcher ~ OM(5,1);

Name(r) ~ CensusDB_NamePrior();

#Paper(Author=r) ~
    if Prof(r) then OM(2,0.5)
                else OM(1.0.5)

Title(p) ~ CSPaperDB_TitlePrior();

PubCited(c) ~ Uniform({Paper p});

Text(c) ~ NoisyCitationGrammar
       (Name(Author(PubCited(c))),
        Title(PubCited(c)));
```

BLOG model (single-author)

```
#Researcher ~ OM(5,1);

Name(r) ~ CensusDB_NamePrior();

#Paper(Author=r) ~
    if Prof(r) then OM(2,0.5)
               else OM(1,0.5)

Title(p) ~ CSPaperDB_TitlePrior();

PubCited(c) ~ Uniform({Paper p});

Text(c) ~ NoisyCitationGrammar
      (Name(Author(PubCited(c))),
       Title(PubCited(c)));
```

BLOG model (single-author)

```
#Researcher ~ OM(5,1);

Name(r) ~ CensusDB_NamePrior();

#Paper(Author=r) ~
    if Prof(r) then OM(2,0.5)
               else OM(1,0.5)

Title(p) ~ CSPaperDB_TitlePrior();

PubCited(c) ~ Uniform({Paper p});

Text(c) ~ NoisyCitationGrammar
      (Name(Author(PubCited(c))),
       Title(PubCited(c)));
```

BLOG model (single-author)

```
#Researcher ~ OM(5,1);

Name(r) ~ CensusDB_NamePrior();

#Paper(Author=r) ~
    if Prof(r) then OM(2,0.5)
               else OM(1,0.5)

Title(p) ~ CSPaperDB_TitlePrior();

PubCited(c) ~ Uniform({Paper p});

Text(c) ~ NoisyCitationGrammar
      (Name(Author(PubCited(c))),
       Title(PubCited(c)));
```

Fraction of citation clusters imperfectly recovered

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

Multi-target tracking + data association

```
#Aircraft(EntryTime = t) ~ Poisson(λa);

Exits(a,t) if InFlight(a,t) then ~ Boolean(αe);

InFlight(a,t) = (t == EntryTime(a))
                |(InFlight(a,t-1) & !Exits(a,t-1));

X(a,t) if t = EntryTime(a) then ~ InitState()
       elseif InFlight(a,t) then
              ~ Normal(F*X(a,t-1),Σx);

#Blip(Source=a, Time=t)
   if InFlight(a,t) then
      ~ Bernoulli(DetectionProbability(X(a,t)));

#Blip(Time=t) ~ Poisson(λf);

Z(b) if Source(b)=null then ~ Uniform(R)
     else ~ Normal(H*X(Source(b),Time(b)),Σz);
```

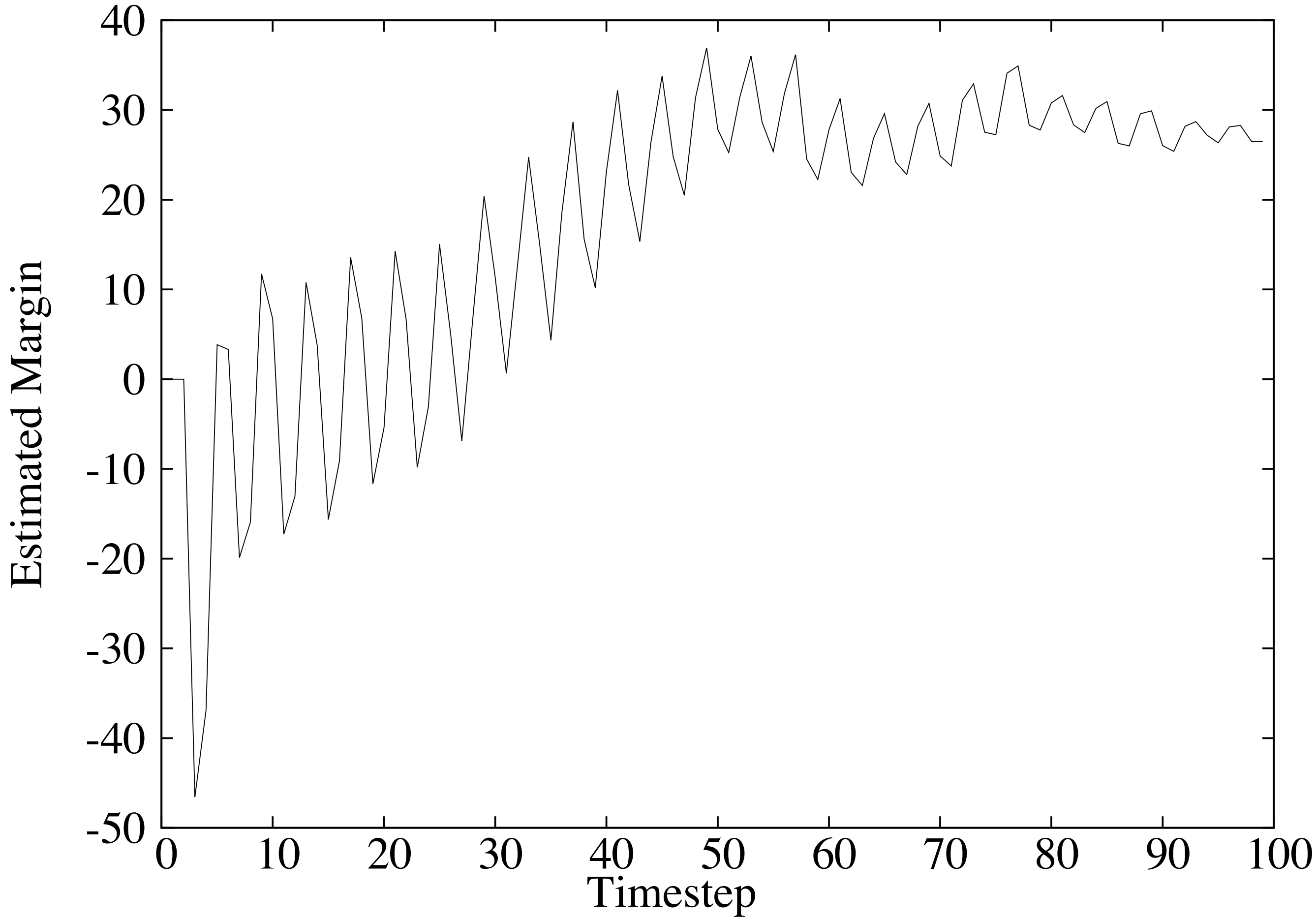
Multi-target tracking + data association

```
#Aircraft(EntryTime = t) ~ Poisson(λa);

Exits(a,t) if InFlight(a,t) then ~ Boolean(αe);

InFlight(a,t) = (t == EntryTime(a))
                |(InFlight(a,t-1) & !Exits(a,t-1));

X(a,t) if t = EntryTime(a) then ~ InitState()
       elseif InFlight(a,t) then
            ~ Normal(F*X(a,t-1),Σx);

#Blip(Source=a, Time=t)
  if InFlight(a,t) then
     ~ Bernoulli(DetectionProbability(X(a,t)));

#Blip(Time=t) ~ Poisson(λf);

Z(b) if Source(b)=null then ~ Uniform(R)
     else ~ Normal(H*X(Source(b),Time(b)),Σz);
```

OUPOMDPs: Blind Monopoly

- Cannot observe opponent's position & holdings
- Get observations for rent payments and receipts
- Observation-based policy hopeless, BSQ policy is very natural:
    - If agent owns one of c, and Pr(opponent doesn't own c) > $\theta 1$, buy
    - If agent doesn't own one of c, and Pr(opponent owns c) > $\theta 2$, buy
    - etc
- OUPOMDPs raise issues of referential transparency (cf modal logic)

Results: Blind Monopoly

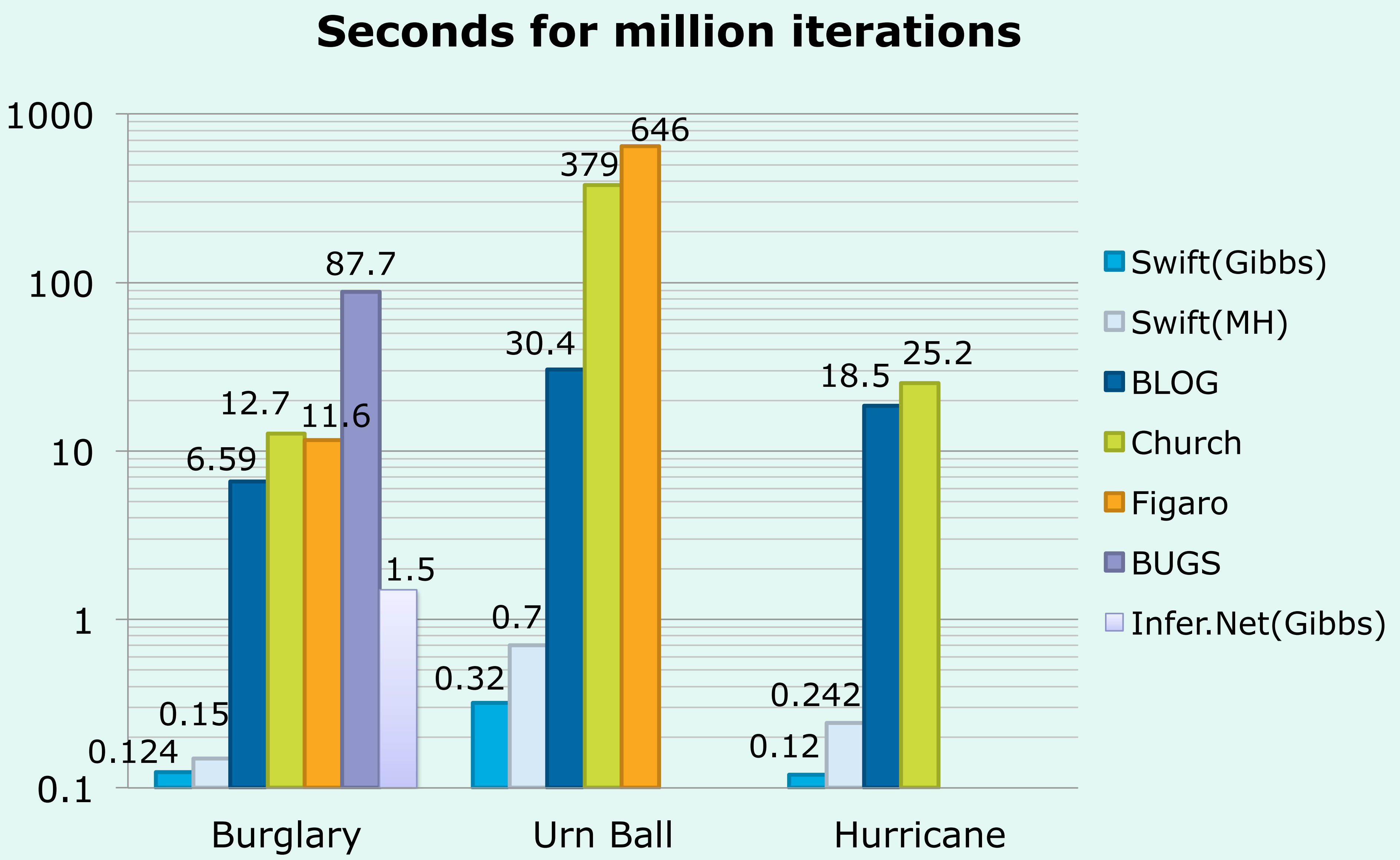


θ1 = 0.75
θ2 = 0.5

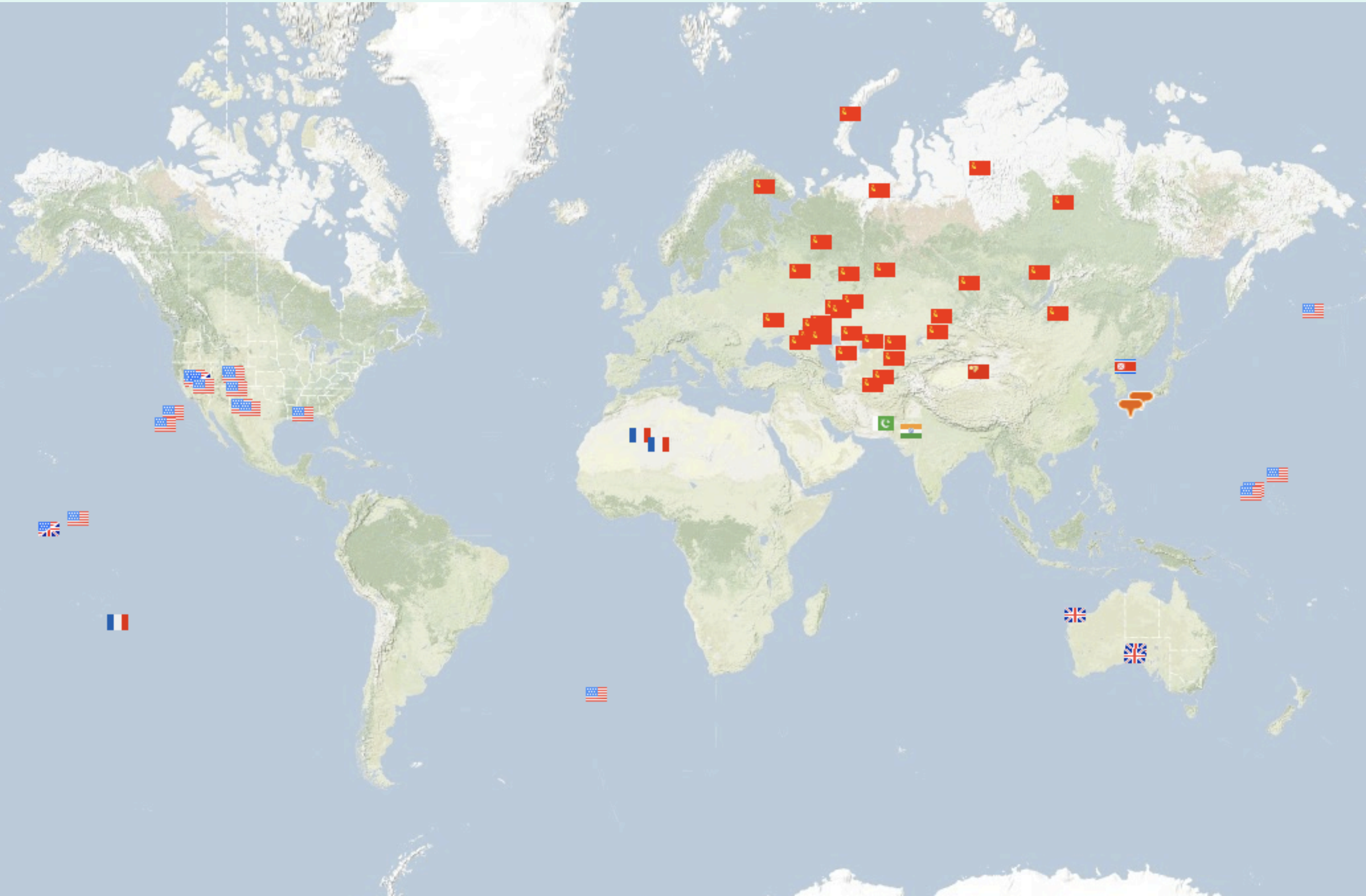
Efficient inference

- Application-specific proposals (make it easy to do!)
- Adaptive proposal distributions
  - IS: see Frank Wood's talk at 10.00am
  - MCMC: neural block sampling (arXiv 2017)
- Automatic model-specific code generation
  - BLOG compiler gives 100x-300x speedup over original engine
    - 700x compared to widely used Bayes net packages
- Modular design with "plug-in" expert samplers
  - Sample $X_1,\ldots,X_k$ given sum; sample parse tree given sentence+PCFG
  - cf JAGS, Edward, MRS
- Data and process parallelism, GPUs, TPUs, etc.
- Lifted inference

Markov Chain Monte Carlo

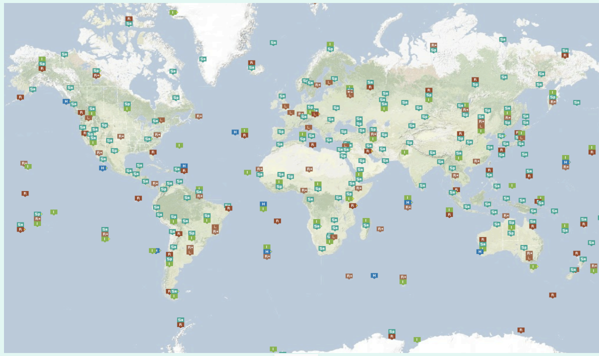


2055 nuclear explosions, 300K deaths

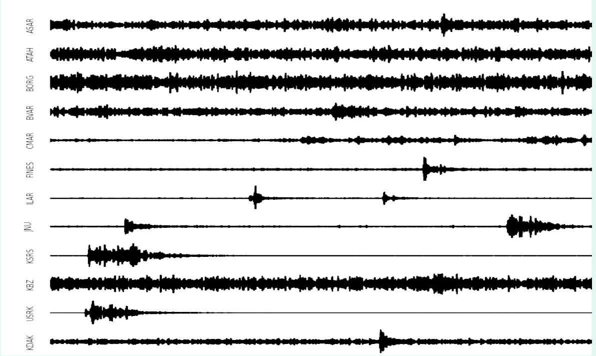


Global seismic monitoring for the Comprehensive Nuclear Test-Ban Treaty

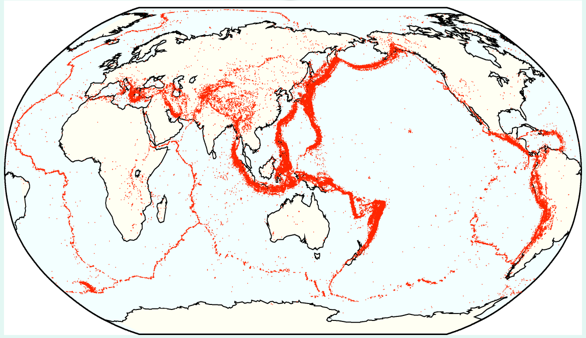
- **Evidence**: waveforms from 150 seismic stations
- **Query**: what happened?
- **Model**: geophysics of event occurrence, signal transmission, detection, noise



IMS



waveforms



bulletin

#SeismicEvents ~ Poisson[T*λ_e];

Time(e) ~ Uniform(0,T)

IsEarthQuake(e) ~ Bernoulli(.999);

Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();

Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;

Magnitude(e) ~ Exponential(log(10));

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

#Detections(site = s) ~ Poisson[T*λ_f(s)];

#Detections(event=e, phase=p, station=s) **=** if IsDetected(e,p,s) then 1 else 0;

OnsetTime(a,s) ~ if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
+ Laplace(μ_t(s), σ_t(s))

Amplitude(a,s) ~ If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

Azimuth(a,s) ~ If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

Slowness(a,s) ~ If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

ObservedPhase(a,s) ~ CategoricalPhaseModel(phase(a))

#SeismicEvents ~ Poisson[T*λ_e];

Time(e) **~** Uniform(0,T)

IsEarthQuake(e) ~ Bernoulli(.999);

Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();

Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;

Magnitude(e) ~ Exponential(log(10));

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

#Detections(site = s) ~ Poisson[T*λ_f(s)];

#Detections(event=e, phase=p, station=s) **=** if IsDetected(e,p,s) then 1 else 0;

OnsetTime(a,s) **~** if (event(a) = null) then Uniform[0,T] else

Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))

+ Laplace(μ_t(s), σ_t(s))

Amplitude(a,s) **~** If (event(a) = null) then NoiseAmplitudeDistribution(s)

else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

Azimuth(a,s) **~** If (event(a) = null) then Uniform(0, 360)

else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

Slowness(a,s) **~** If (event(a) = null) then Uniform(0,20)

else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

ObservedPhase(a,s) **~** CategoricalPhaseModel(phase(a))

#SeismicEvents ~ Poisson[T*λ_e];

Time(e) **~** Uniform(0,T)

IsEarthQuake(e) ~ Bernoulli(.999);

Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();

Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;

Magnitude(e) ~ Exponential(log(10));

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

#Detections(site = s) ~ Poisson[T*λ_f(s)];

#Detections(event=e, phase=p, station=s) **=** if IsDetected(e,p,s) then 1 else 0;

OnsetTime(a,s) **~** if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
+ Laplace(μ_t(s), σ_t(s))

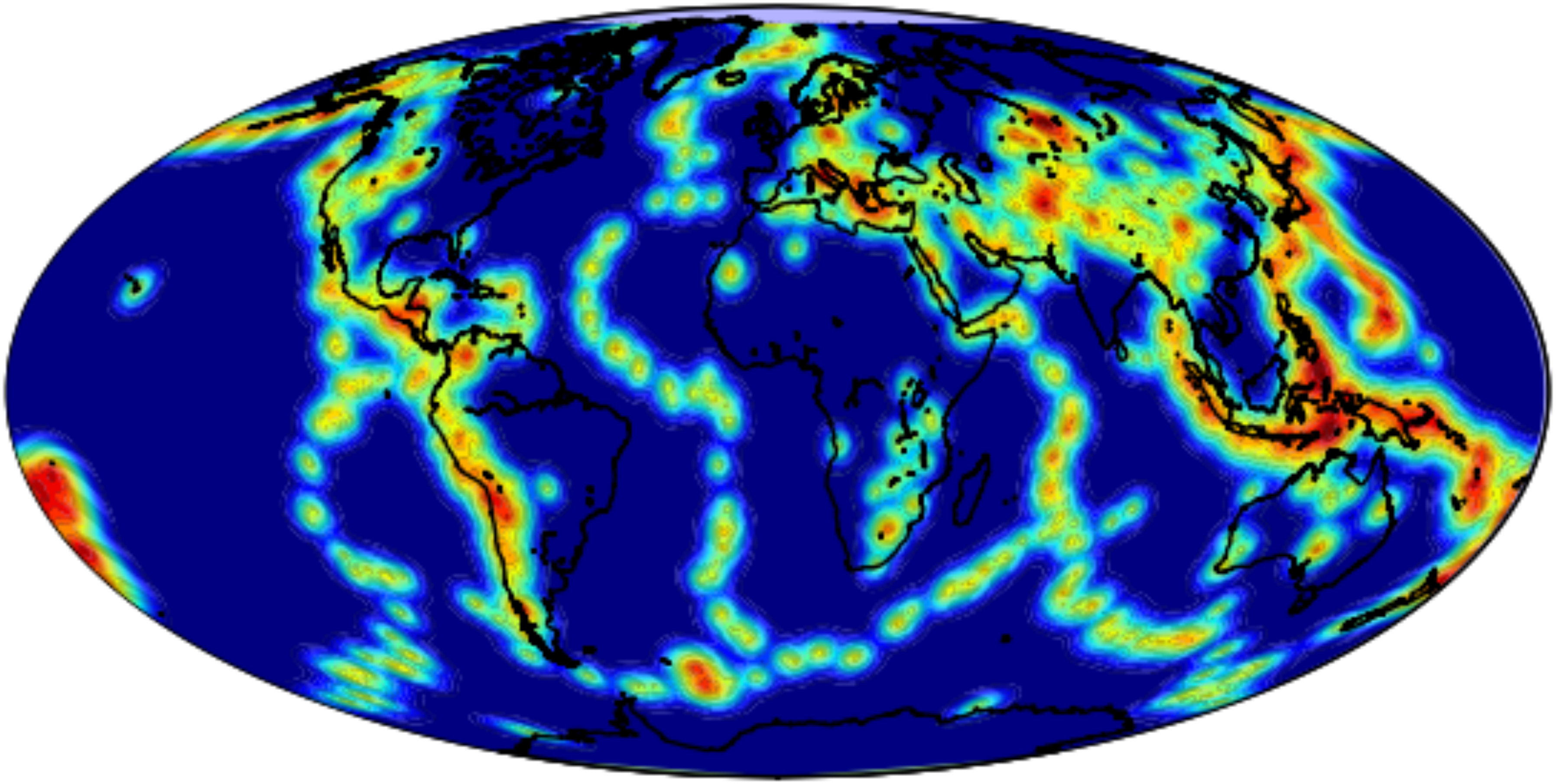
Amplitude(a,s) **~** If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

Azimuth(a,s) **~** If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

Slowness(a,s) **~** If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

ObservedPhase(a,s) **~** CategoricalPhaseModel(phase(a))

Event location prior



#SeismicEvents ~ Poisson[T*λ_e];
***Time(e)* ~** Uniform(0,T)
IsEarthQuake(e) ~ Bernoulli(.999);
Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();
Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;
Magnitude(e) ~ Exponential(log(10));

IsDetected(e,p,s) ~ Logistic[weights(s,p)](Magnitude(e), Depth(e), Distance(e,s))**;**
#Detections(site = s) ~ Poisson[T*λ_f(s)];
***#Detections(event=e, phase=p, station=s)* =** if IsDetected(e,p,s) then 1 else 0;
***OnsetTime(a,s)* ~** if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
+ Laplace(μ_t(s), σ_t(s))

***Amplitude(a,s)* ~** If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

***Azimuth(a,s)* ~** If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

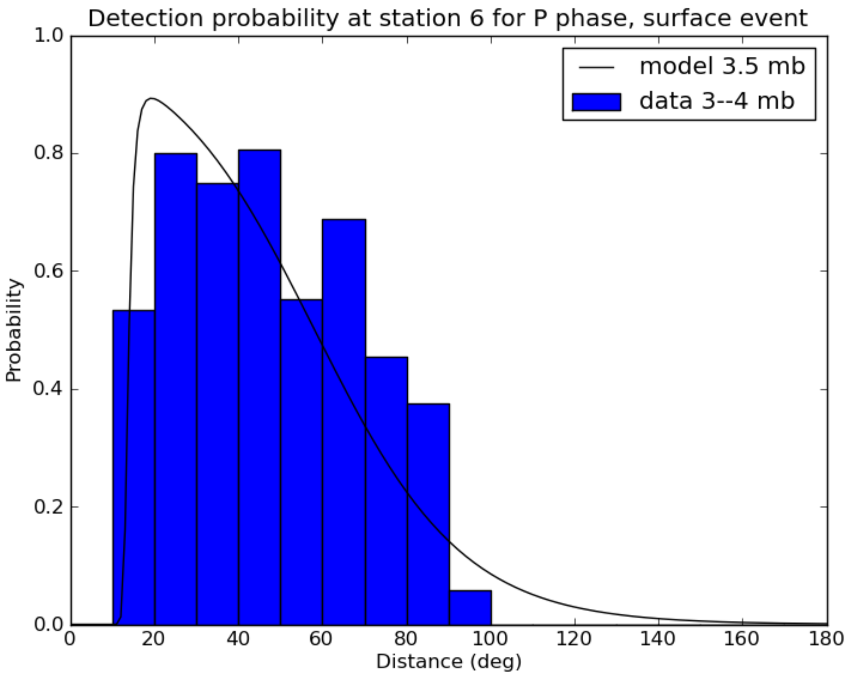
***Slowness(a,s)* ~** If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

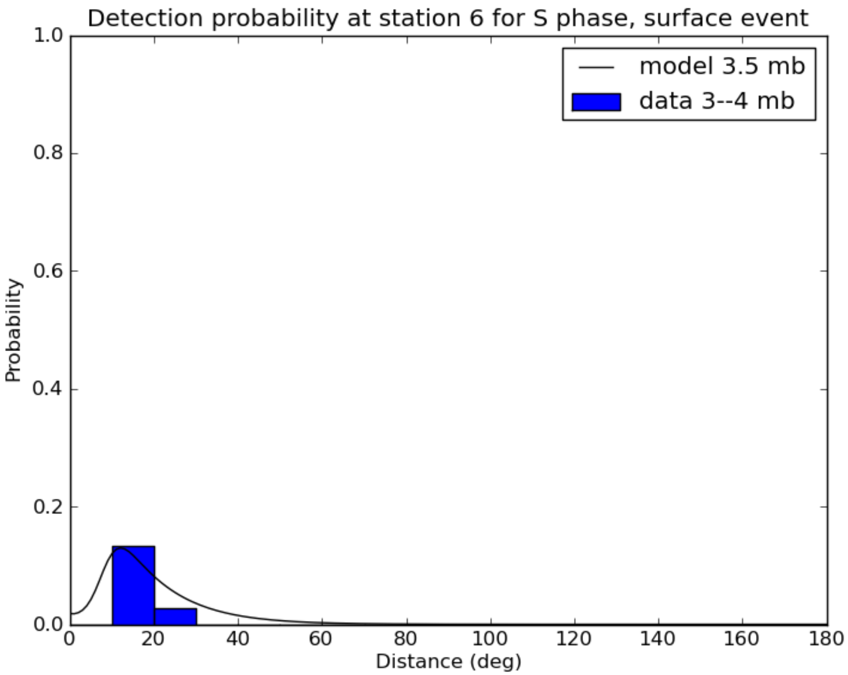
***ObservedPhase(a,s)* ~** CategoricalPhaseModel(phase(a))

Detection probability as a function of distance (Alice Springs station, m_b 3.5)

P phase

S phase





#SeismicEvents ~ Poisson[T*λ_e];
***Time(e)* ~** Uniform(0,T)
IsEarthQuake(e) ~ Bernoulli(.999);
Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();
Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;
Magnitude(e) ~ Exponential(log(10));

***IsDetected(e,p,s)* ~** Logistic[weights(s,p)](Magnitude(e), Depth(e), Distance(e,s))**;**
#Detections(site = s) ~ Poisson[T*λ_f(s)];
***#Detections(event=e, phase=p, station=s)* =** if IsDetected(e,p,s) then 1 else 0;
***OnsetTime(a,s)* ~** if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
+ Laplace(μ_t(s), σ_t(s))

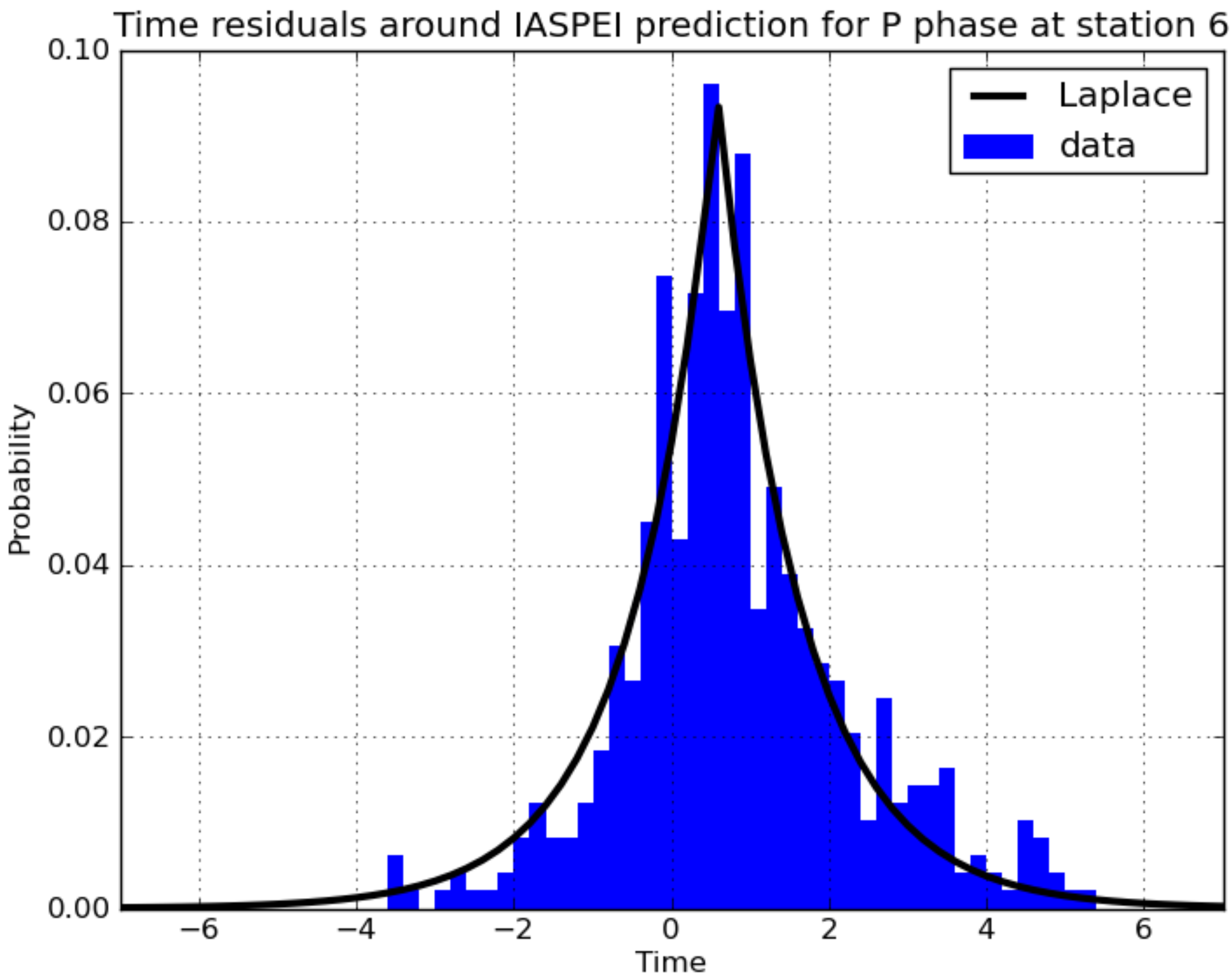
***Amplitude(a,s)* ~** If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

***Azimuth(a,s)* ~** If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

***Slowness(a,s)* ~** If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

***ObservedPhase(a,s)* ~** CategoricalPhaseModel(phase(a))

Travel-time residual (Alice Springs)



#SeismicEvents ~ Poisson[T*λ_e];

Time(e) ~ Uniform(0,T)

IsEarthQuake(e) ~ Bernoulli(.999);

Location(e) ~ if IsEarthQuake(e) then SpatialPrior() else UniformEarthDistribution();

Depth(e) ~ if IsEarthQuake(e) then Uniform[0,700] else 0;

Magnitude(e) ~ Exponential(log(10));

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

#Detections(site = s) ~ Poisson[T*λ_f(s)];

#Detections(event=e, phase=p, station=s) = if IsDetected(e,p,s) then 1 else 0;

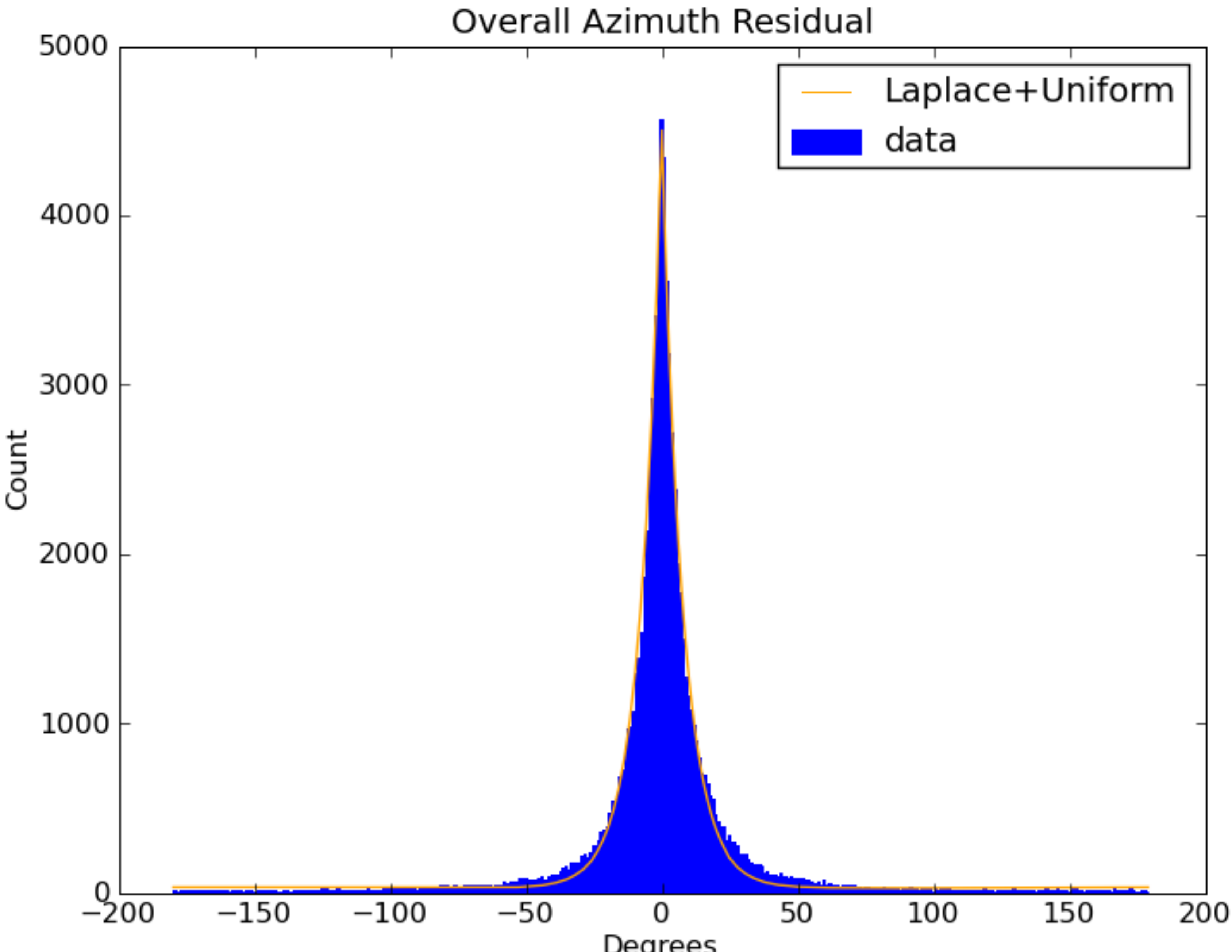
OnsetTime(a,s) ~ if (event(a) = null) then Uniform[0,T] else
Time(event(a)) + GeoTravelTime(Distance(event(a),s),Depth(event(a)),phase(a))
+ Laplace(μ_t(s), σ_t(s))

Amplitude(a,s) ~ If (event(a) = null) then NoiseAmplitudeDistribution(s)
else AmplitudeModel(Magnitude(event(a)), Distance(event(a),s),Depth(event(a)),phase(a))

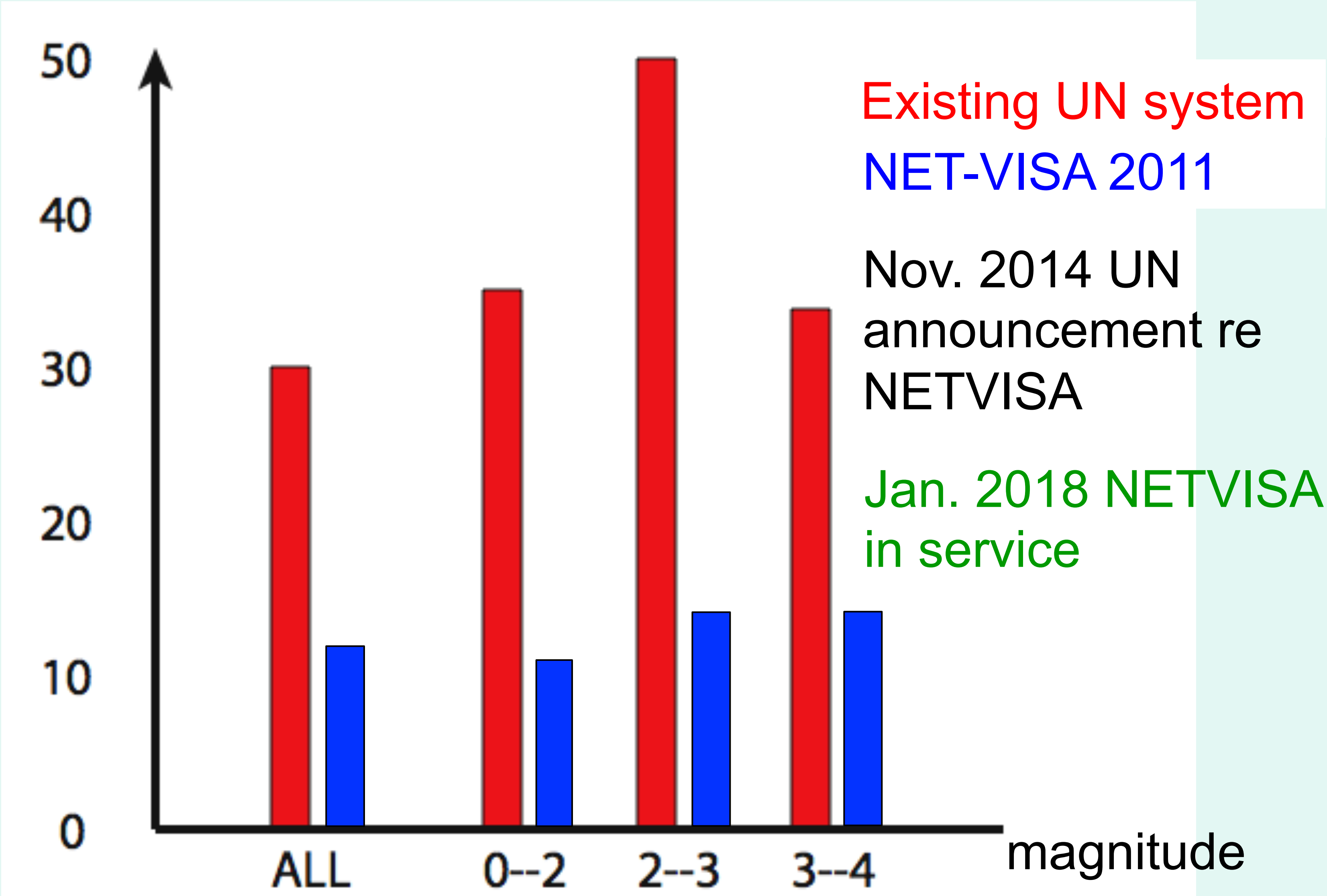
Azimuth(a,s) ~ If (event(a) = null) then Uniform(0, 360)
else GeoAzimuth(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

Slowness(a,s) ~ If (event(a) = null) then Uniform(0,20)
else GeoSlowness(Location(event(a)),Depth(event(a)),phase(a),Site(s)) + Laplace(0,σ_a(s))

ObservedPhase(a,s) ~ CategoricalPhaseModel(phase(a))


Overall Azimuth Residual
Laplace+Uniform
data
Count
Degrees
5000
4000
3000
2000
1000
0
−200
−150
−100
−50
0
50
100
150
200

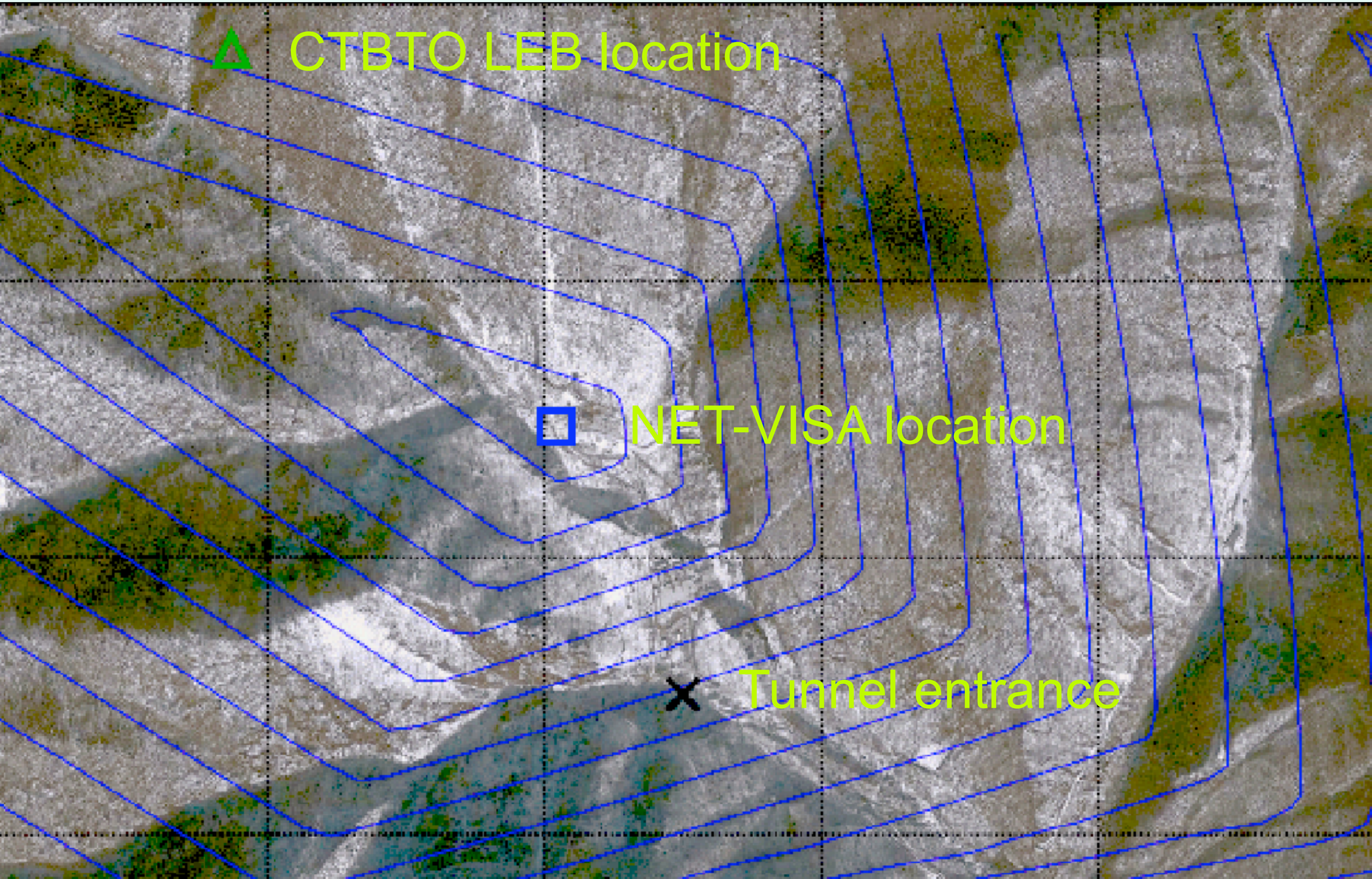
Fraction of events missed



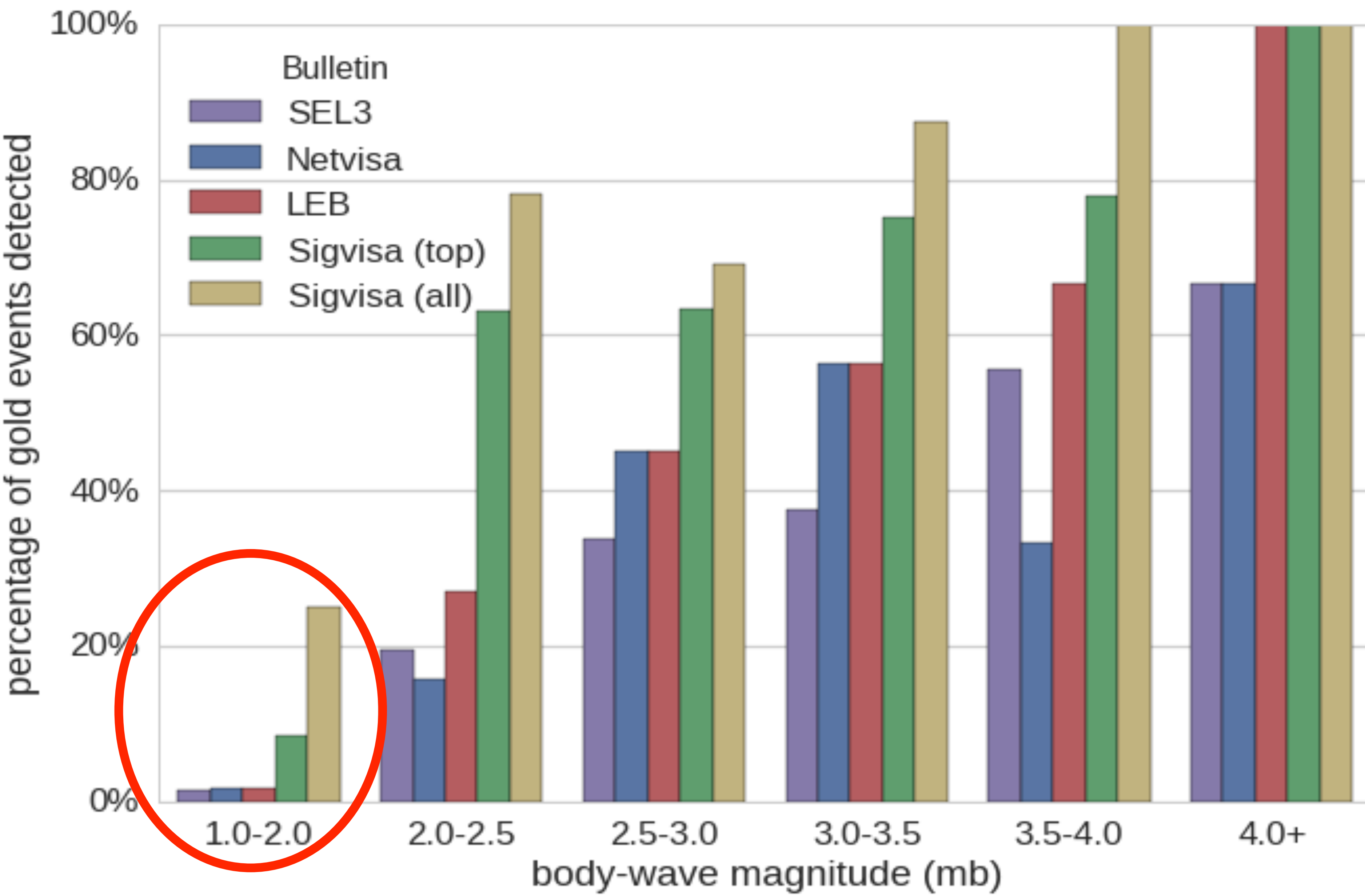
Nov. 2014 UN announcement re NETVISA

Jan. 2018 NETVISA in service

February 12, 2013 DPRK test



SIGVISA on Western US dataset



Tracking moving objects in video

- Standard model (Friedman & Russell, 1997; Stauffer & Grimson, 1998):
  - Each pixel sampled from 3-component mixture for background, foreground, shadow
  - Mixture parameters estimated online
- Improvement 1: temporal persistence for each pixel state: per-pixel HMM
- Improvement 2: spatial contiguity of pixel states within each frame: Ising potential

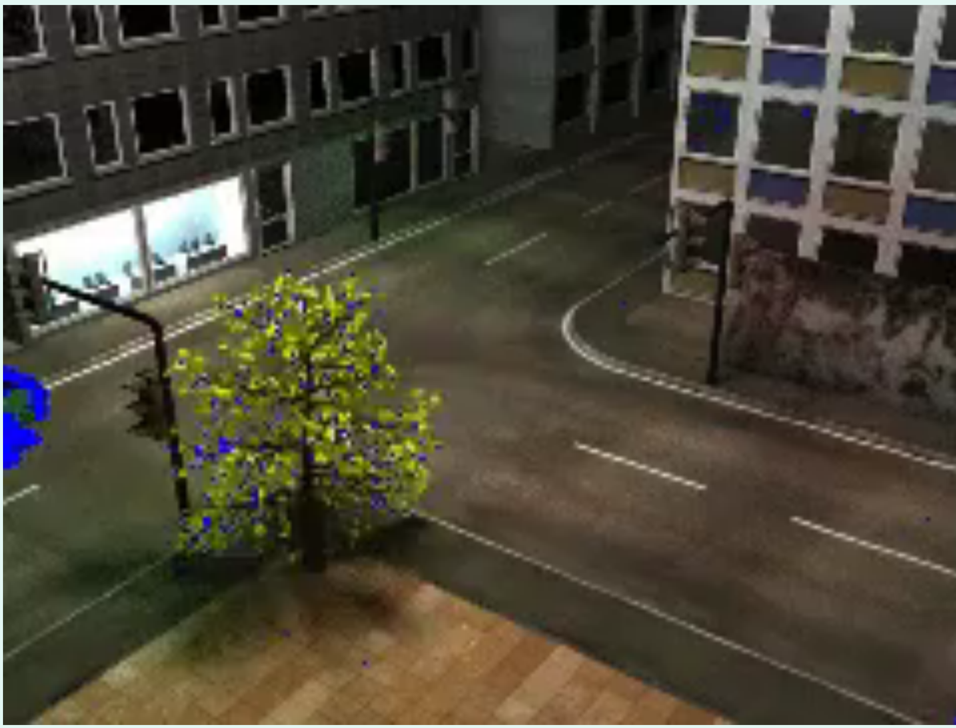
BLOG model (cleaned up a bit)

```
…
Intensity(x,y,t) ~
    MultivarGaussian(Mean(PixelState(x,y,t),x,y),
                     Variance(PixelState(x,y,t),x,y));
PixelState(x,y,t) ~
    Categorical(MixtureWeights(x,y,t));
…
```

BLOG model (cleaned up a bit)

```
…
Intensity(x,y,t) ~
    MultivarGaussian(Mean(PixelState(x,y,t),x,y),
                     Variance(PixelState(x,y,t),x,y));
PixelState(x,y,t) ~
    if t==0 then Categorical(MixtureWeights(x,y,0))
    else TransitionModel(x,y,PixelState(x,y,t-1));
…
```

OpenCV



BLOG model

Completely unsupervised text understanding

#Object ~ OM(3,1);
#Relation ~ OM(2,1)
Dictionary(r) ~ Dirichlet(α,StringList);
Sparsity(r) ~ Beta(10,1000);
Holds(r,x,y) ~ Boolean(Sparsity(r));
ChosenFact(s) ~ Uniform({f : Holds(f)})
Subject(s) = Arg1(ChosenFact(s))
Object(s) = Arg2(ChosenFact(s))
Verb(s) ~
Categ(Dictionary(Rel(ChosenFact(s))))

There are many objects in the world

There are quite a few relations

Relations are expressed by strings

Some (very few) objects are related to each other by any given relation

People somehow choose facts to say

Subject of sentence is 1^{st} arg of fact

Object of sentence is 2^{nd} arg of fact

Verb of sentence is the relation string

Evidence: unsupervised sentence data (NYT subset, from McCallum group)

"… J. Edgar Hoover, who was Director of the FBI, …"

Query: what is true in the world?

Relation [rel_46] : text patterns

appos->**unit**->prep->**of**->pobj

appos->**part**->prep->**of**->pobj

nn<-**unit**->prep->**of**->pobj

partmod->**own**->prep->**by**->pobj

rcmod->**own**->prep->**by**->pobj

appos->**subsidiary**->prep->**of**->pobj

rcmod->**part**->prep->**of**->pobj

rcmod->**unit**->prep->**of**->pobj

poss<-**parent**->appos

appos->**division**->prep->**of**->pobj

pobj<-**of**<-prep<-**office**->appos->**part**->prep->**of**->pobj

pobj<-**of**<-prep<-**unit**->appos->**part**->prep->**of**->pobj

nn<-**division**->prep->**of**->pobj

appos->**unit**->nn

nsubjpass<-**own**->prep->**by**->pobj

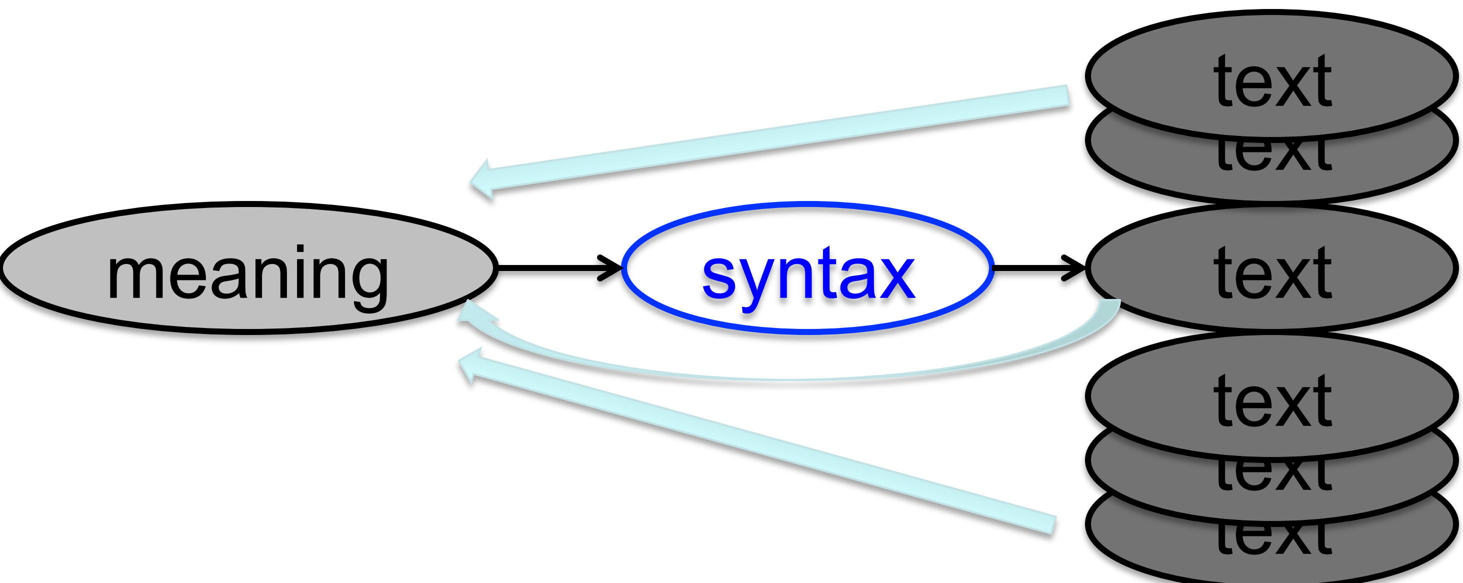
nn<-**office**->prep->**of**->pobj

Relation [rel_46] : extracted facts

rel_46(ABC, Walt Disney Company)
rel_46(American Airlines, AMR Corporation)
rel_46(American, AMR Corporation)
rel_46(Arnold Worldwide, Arnold Worldwide Partners)
rel_46(BBDO Worldwide, Omnicom Group)
rel_46(Bozell Worldwide, Bozell)
rel_46(Chicago, DDB Worldwide)
rel_46(Conde Nast, Advance Publications)
rel_46(DDB Needham Worldwide, Omnicom Group)
rel_46(DDB Worldwide, Omnicom Group)
rel_46(Eastern, Texas Air Corporation)
rel_46(Electronic Data Systems, General Motors)
rel_46(Euro RSCG Worldwide, Havas Advertising)
rel_46(Euro RSCG Worldwide, Havas)
rel_46(Fallon Worldwide, Publicis Groupe)
rel_46(Foote, True North Communications)
rel_46(Fox, News Corporation)
rel_46(Goodby, Omnicom Group)
rel_46(Grey Worldwide, Grey Global Group)
rel_46(Hughes, General Motors Corporation)
rel_46(J. Walter Thompson, WPP Group)
rel_46(Kellogg Brown & Root, Halliburton)
rel_46(Kellogg, Halliburton)
rel_46(Kraft General Foods, Philip Morris Cos.)
rel_46(Lorillard Tobacco, Loews Corporation)
rel_46(Lowe Group, Interpublic Group)
rel_46(McCann-Erickson, Interpublic Group)
rel_46(NBC, General Electric Company)
rel_46(New York, BBDO Worldwide)
rel_46(New York, Hill)
rel_46(Ogilvy & Mather Worldwide, WPP Group)
rel_46(Saatchi & Saatchi, Publicis Groupe)
rel_46(Salomon Smith Barney, Citigroup)
rel_46(San Francisco, Foote)
rel_46(Sears Receivables Financing, Sears)
rel_46(TBWA Worldwide, Omnicom Group)
rel_46(United, UAL Corporation)
rel_46(United, UAL)
rel_46(Young & Rubicam, WPP Group)

FURTHER ELABORATIONS

- Entity resolution: generative models of entity mentions
- Ontology: types, time, events, vector-space meaning
- Pragmatics: choice of facts, effect of context
- Grammar: learning the missing link



Summary

- Unifying logic and probability leads to a declarative PPL for *open-universe* probability models
- Let many flowers bloom! We need to try many approaches to modeling and inference
  - But not too many!
- Future topics:
  - QA from a large body of evidence: intermediate structures?
  - Cumulative structural learning: predicate and function invention
  - Real understanding of language, images: coarse-to-fine models
  - Continued integration with data-driven learning methods

Thank you!

Research funded by DARPA, CTBTO, DTRA, ANR, and Région d’Île de France