Basics of Bayesian Formalism

Ashish Mahabal
(PQ, CSS, JPL collabs)
Ay/Bi 199
Caltech, 12 May 2011

Advantages of Bayesian Networks

- Handling of incomplete data
  - Real-world cases
- Learning causal connections
  - What variable caused what
- Incorporating domain knowledge
  - Experts can weight in at different points
- Memorizing (aka overfitting) avoided
  - No holdout necessary

- Tue May 17 Moghaddam

Bayesian Methods

Nonparametric Bayes and Gaussian Processes

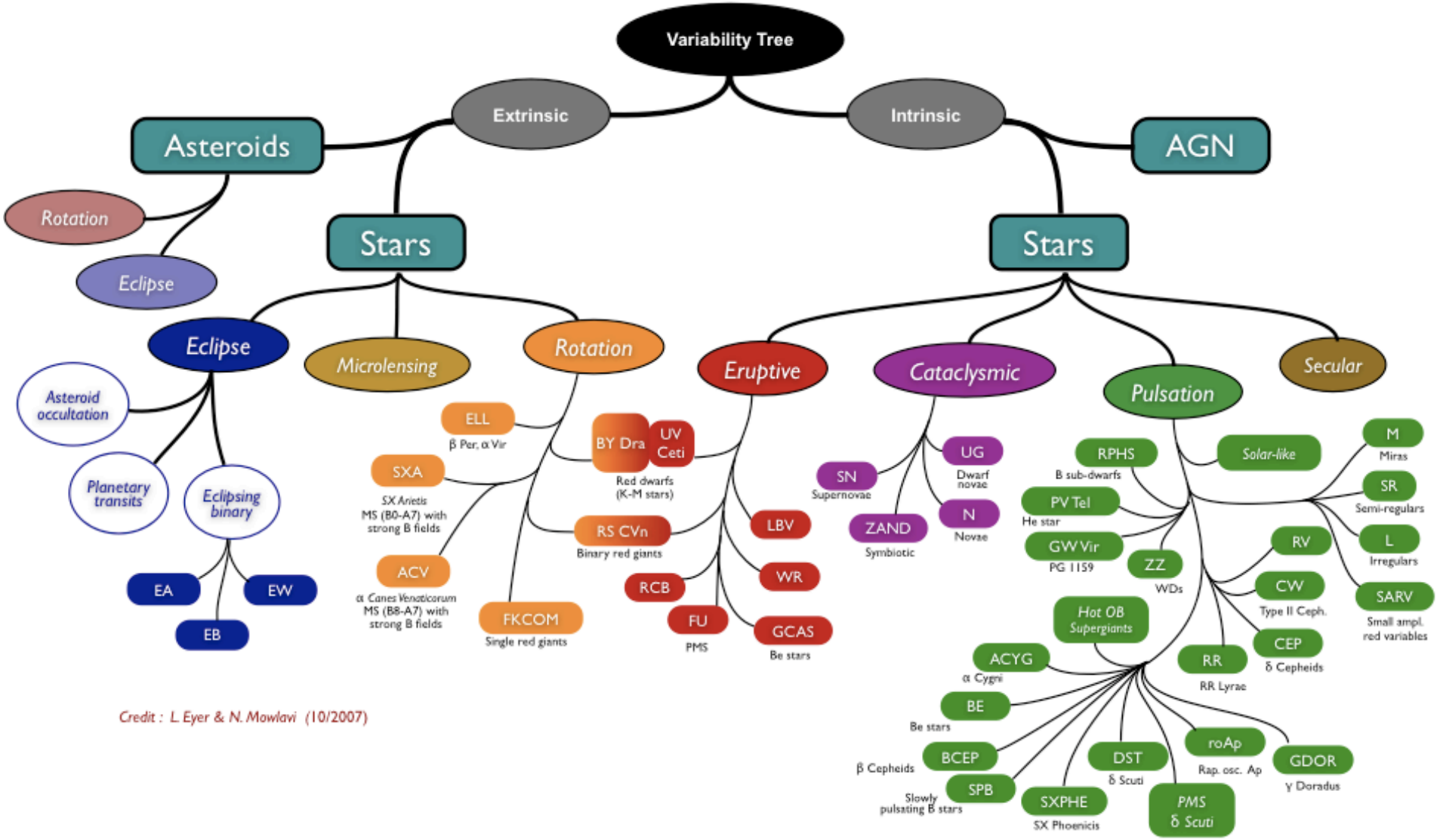
Ambitious Outline

- Basic astronomy classification trivia
- Time and place for Bayesian techniques
- Basic concepts related to belief
- Logic and probability theory
- The inversion formula
- Application to astronomy

Astronomical Classification and the time domain

- Moving objects (asteroids, TNOs, KBOs)
- SNe (cosmological standard candles, endpoints of stellar evolution)
- GRB orphan afterglows (constraining beaming models)
- Variable stars (stellar astrophysics, galactic structure)
- AGN (QSOs, fuelling mechanisms, lifetimes)
- Blazars, Cosmic Rays, …

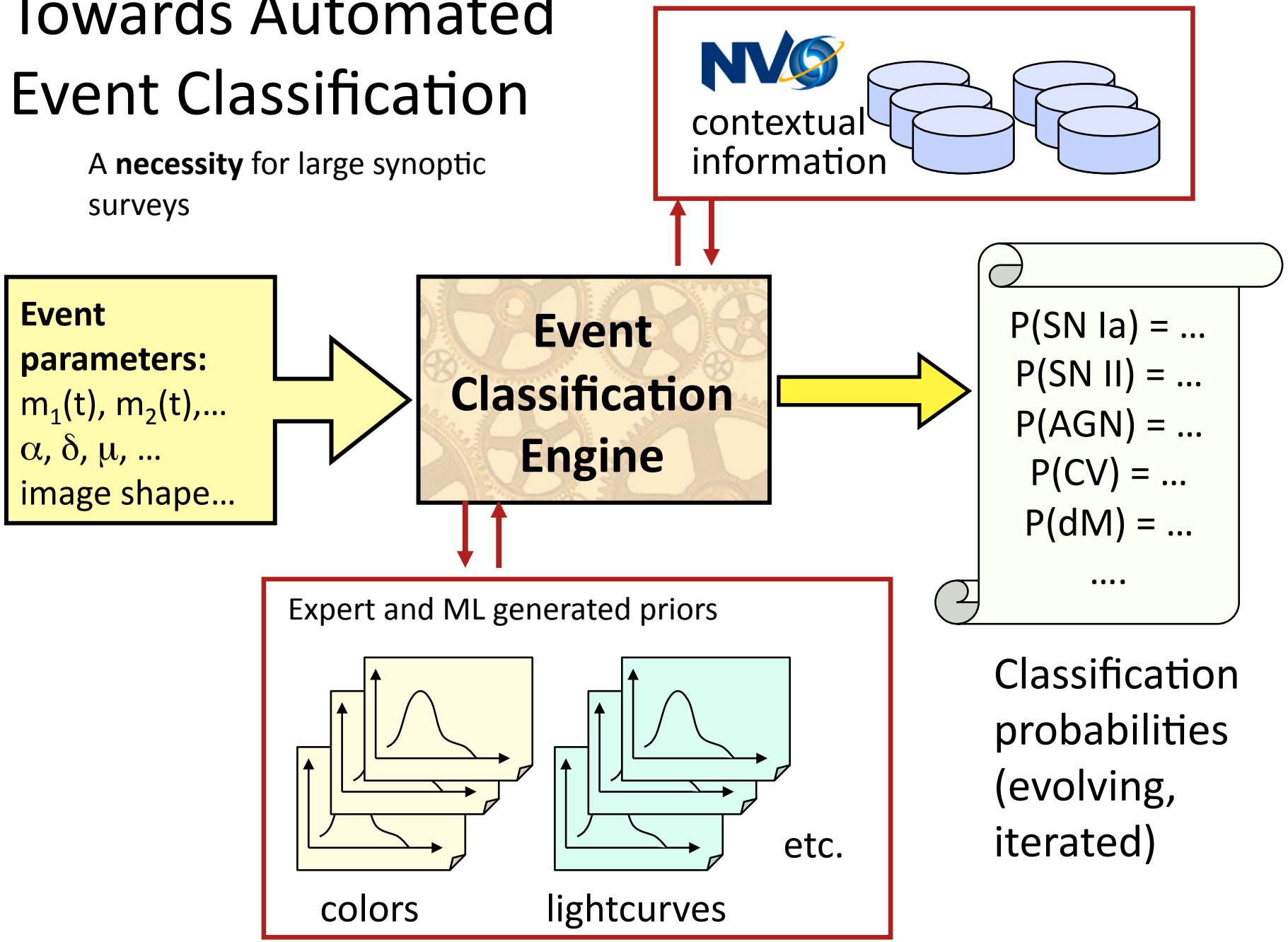
Rapid follow-up ➔ keys to understanding



Credit : L. Eyer & N. Mowlavi (10/2007)

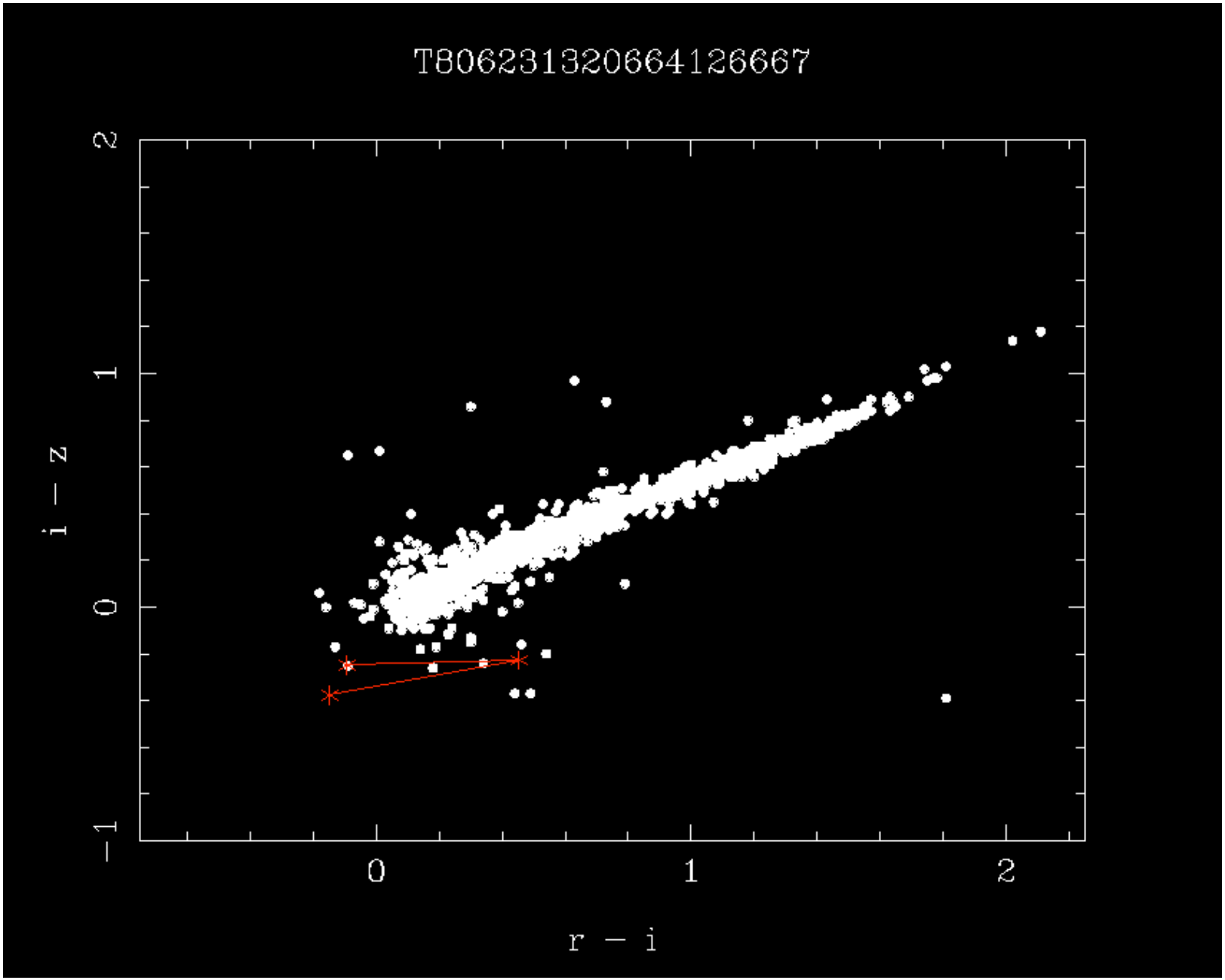
Towards Automated Event Classification

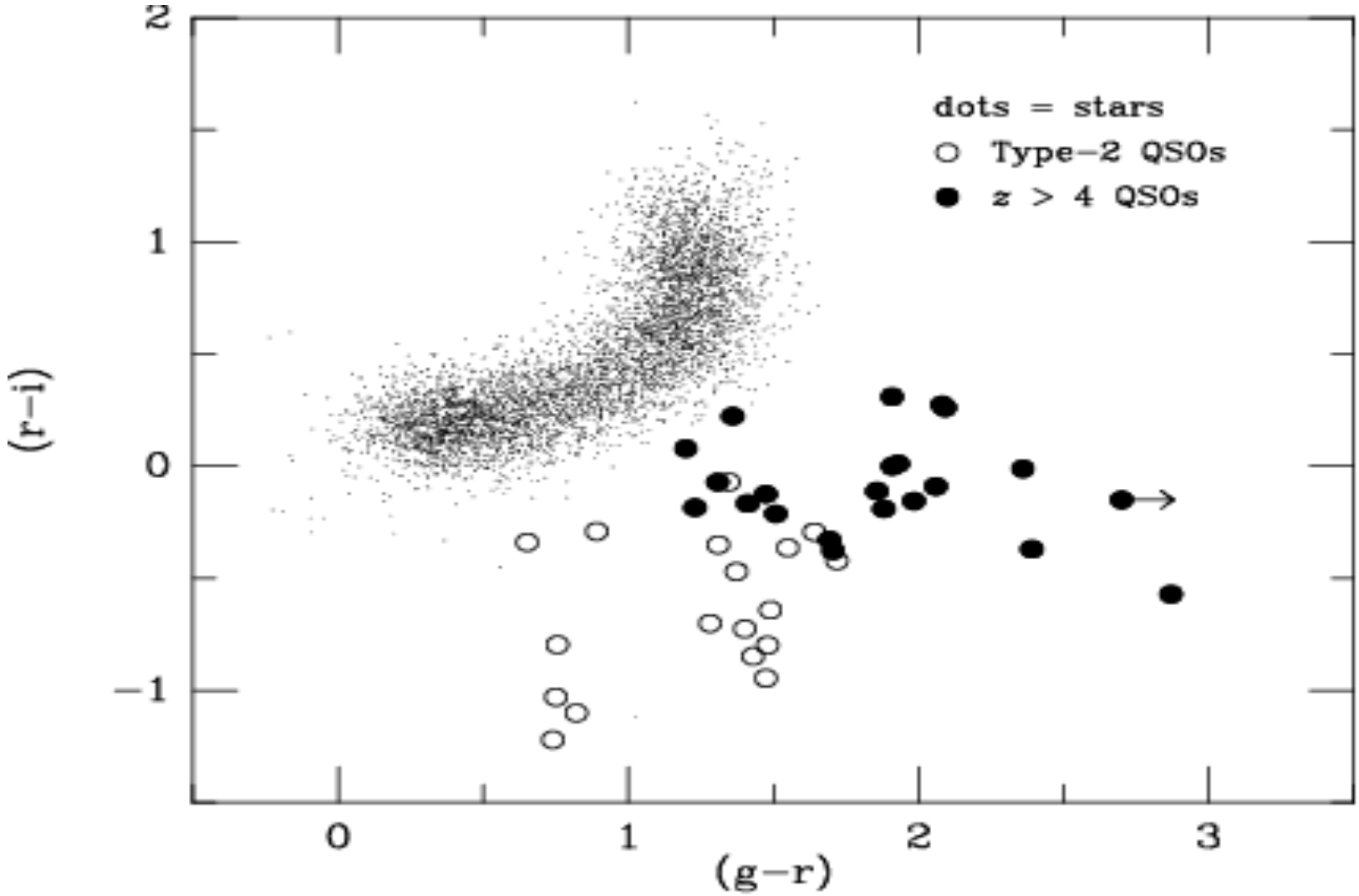
A **necessity** for large synoptic surveys



Basic astronomy classification trivia
Colors

- Magnitude as basic observation (flux)
- Color as flux ratio
- Color-color diagram as a diagnostic
- Ambiguity


TB06231320664126667
2
1
0
−1
i − z
0
1
2
r − i

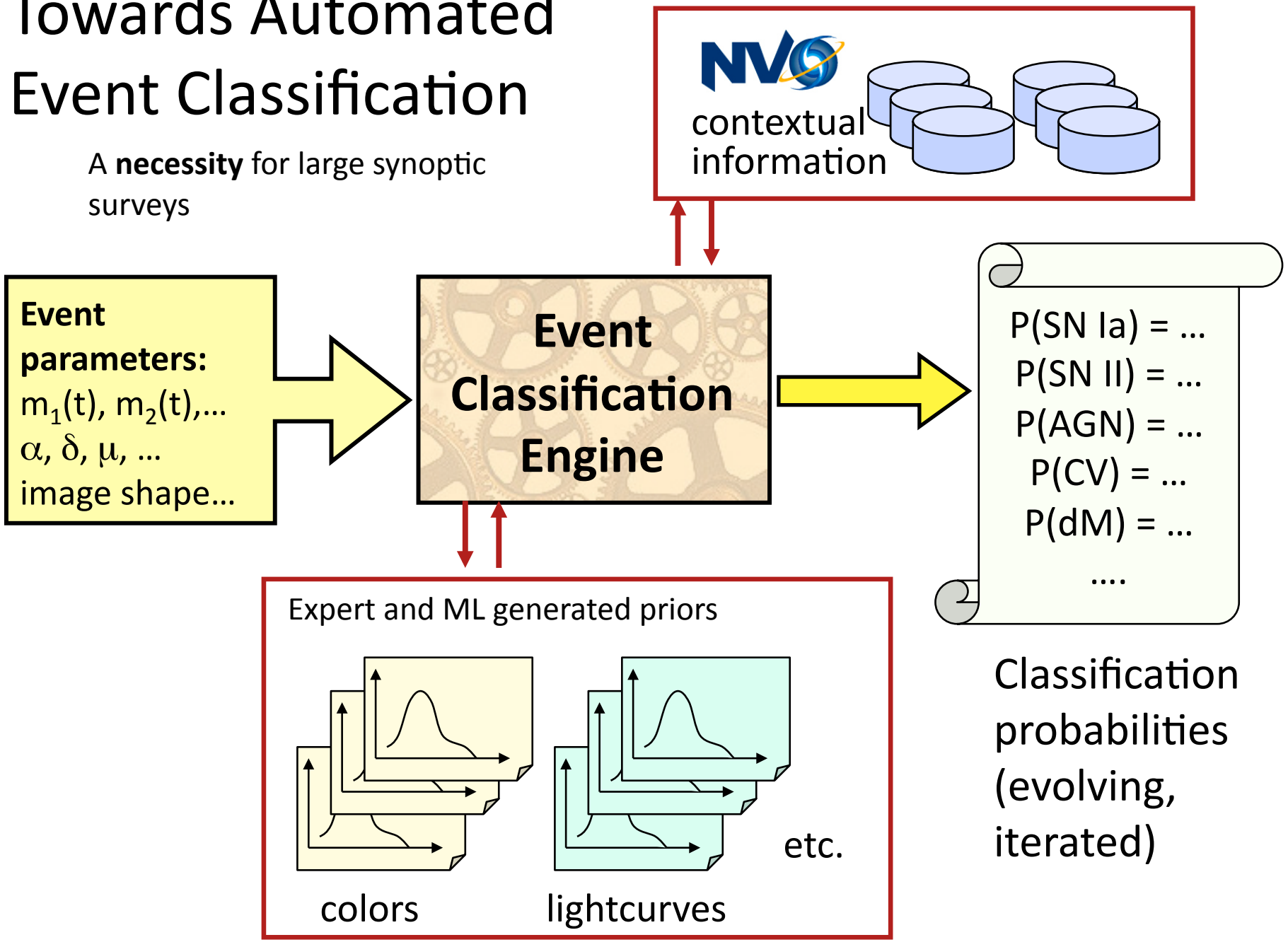

dots = stars
Type−2 QSOs
z > 4 QSOs
(r−i)
(g−r)
2
1
0
−1
0
1
2
3

Basic astronomy classification trivia

- Attach probabilities through priors to various classes and determine what class a newly looked at object belongs to
- Bayesian techniques allow us to do this in a rational manner even when some of the data is uncertain or missing

Towards Automated Event Classification

A **necessity** for large synoptic surveys



Bayesian techniques

- Bayesian methods provide a formalism for reasoning about **partial beliefs** under **conditions of uncertainty**

Belief is going to be a crucial word

A: World will end in 2012

p(A|K) → belief about A given a body of knowledge K. Often written simply as p(A).

- p(NOT A) → belief that A will not happen
- When K changes, p(A) and p(NOT A) change accordingly
- In general:
  - $0 \leq p(A) \leq 1$
  - p(sure proposition) = 1
  - p(A or B) = p(A) + p(B) when A and B are mutually exclusive

- $p(A) + p(\text{NOT } A) = 1$
- $p(A) = p(A,B) + p(A, \text{NOT } B)$
  - $p(A,B) == p(A \text{ and } B)$

  In general:

  - For Bi = B1, B2, B3, …Bn mutually exclusive

  $p(A) = \Sigma\, p(A,Bi) = p(A,B1) + p(A,B2) + .. + p(A,Bn)$

  We will revisit these later. Catchword: **Belief**

An example

- A: outcome of 2 dice is equal
- B1: 2nd die is 1
- B2 .. B6: 2nd die is 2..6
- Bi = B1 .. B6 form a partition (are exhaustive)
- $p(A) = \Sigma p(A,Bi) = p(A,B1) + .. + p(A,B6)$

  $= 1/36 + .. + 1/36$

  $= 1/6$

About belief

- Rules have exceptions
- Ignoring exceptions leads to uncertainty
- If we consider all exceptions, we may not be able to proceed
- Middle way is to summarize exceptions

That's where Bayes formalism leads us to

An example

Suppose I have a bird

What can be said about its ability to fly?

Birds fly. So the bird under question should be able to fly

An example

Suppose I have a bird

What can be said about its ability to fly?

Birds fly. So the bird under question should be able to fly

- What if it's a penguin?
- Dead?
- With wings cut?
- Made of paper?

Summary: Most birds can fly. Need to make it quantitative

Logical approach?

- Assign numerical values to uncertainty and combine them like truth values

- But sources of uncertainty are not independent and it is not easy to evaluate effect of additional evidence e.g. in the last example, if we are told that the said bird has existed for 1 year, how do we take that into consideration?

What is p(B|A)?

- ***Not*** "Given A, the probability that B is true"

  That is true only if B does not (also) depend on anything else. Because if we knew other things, may be the probability will be different.

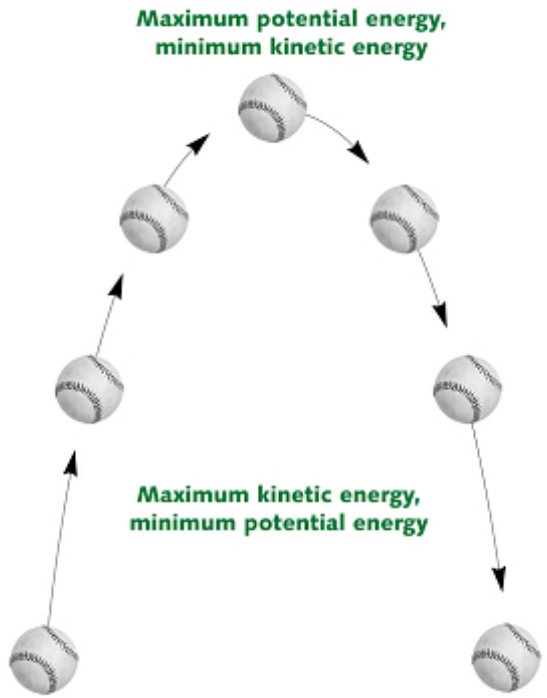
  If of 10000 species 10 can't fly, then:

  p(cant fly | bird) = 1/1000 (blanket statement)

  That does not indicate if we know anything else (here, e.g. that the bird has existed for 1 year).

Verification of irrelevancy is crucial

- Rule: What goes up must come down
  - A: foo comes down
  - B: foo goes up
  - P(A|B) = 1



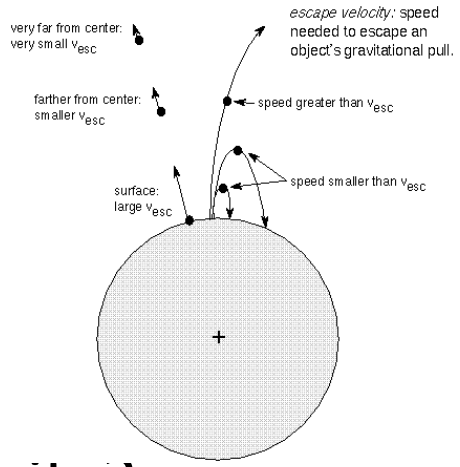
- Rule: What goes up must come down
  - A: foo comes down
  - B: foo goes up
  - P(A|B) = 1

  Is that really true?

  What if upward velocity > escape velocity?

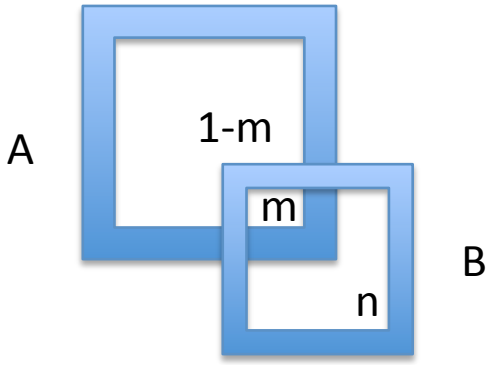
  Where did escape velocity come from?

  Escape velocity was always there



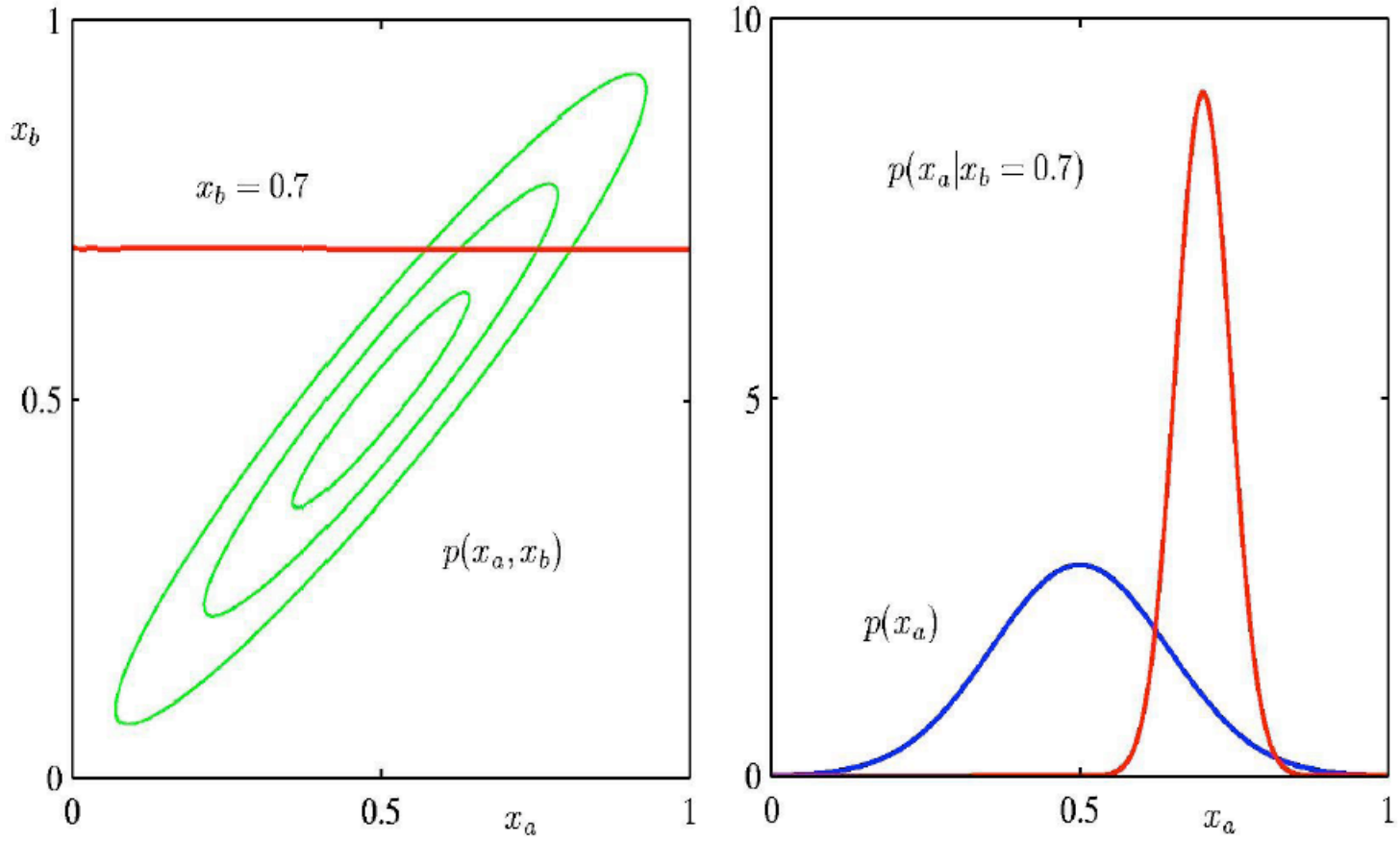
Intentional v. extensional

- These rule based systems are computationally convenient, but semantically inconvenient.
- The opposite is true of Bayes formalism: its declarative and model based



P(B|A) = m =>
In all worlds that satisfy A, those also satisfying B are a fraction m

Partitioned Conditionals and Marginals



A bit more about logical systems

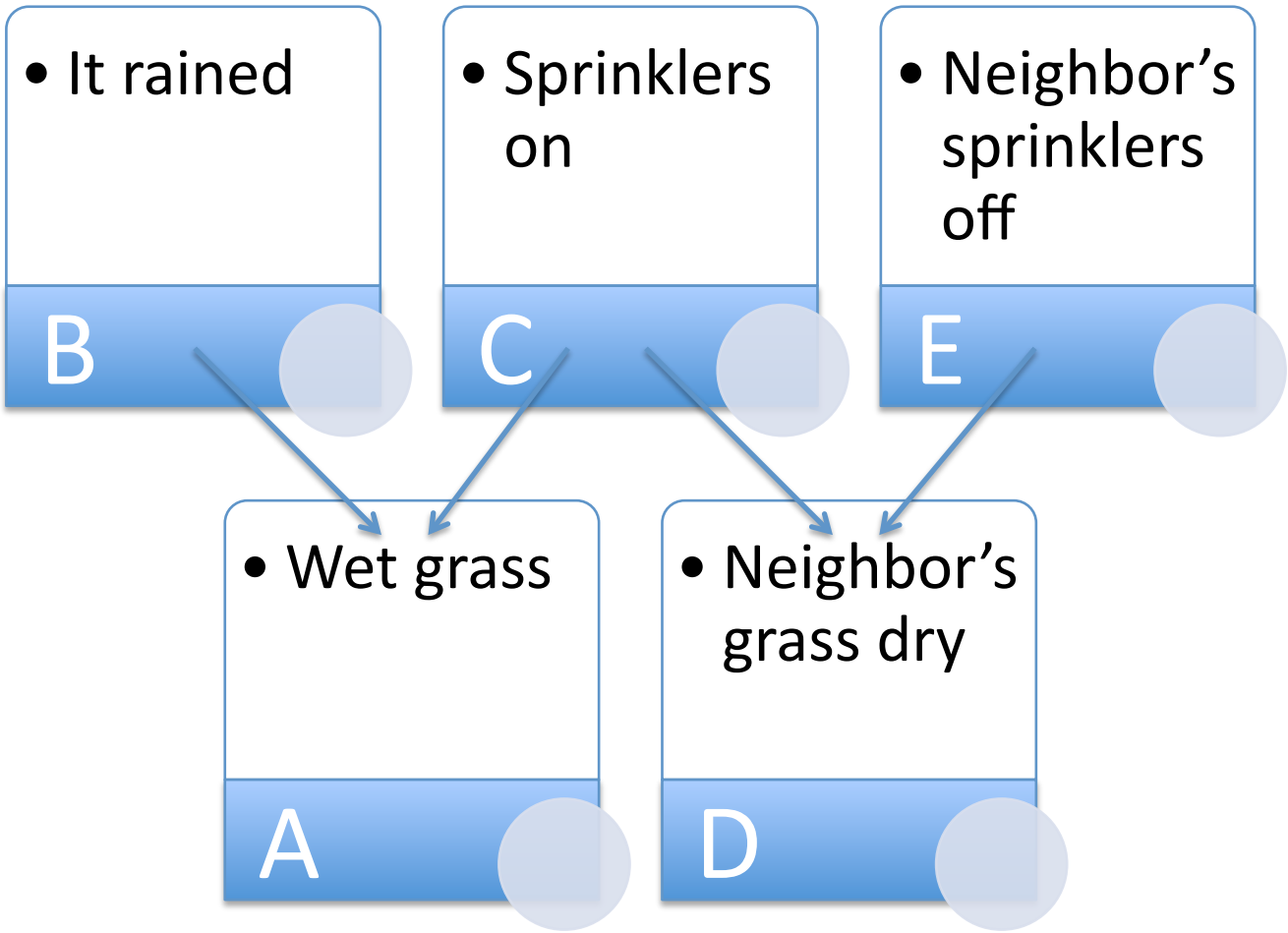
- **A: The grass is wet**
- **B: It rained**

A gives credibility to B. In a rule based system, that weight increases irrevocably.

If, later, **C: The sprinkler was on**,

And **D: The neighbor's grass is dry**

It is difficult to connect **A** and **C**



If anything, **C** becomes less credible once we know **A** to be true

Chaining?

- Logic: If A then B and if B then C => if A then C
- In plausible reasoning, it can lead to problems:
  - If the ground is wet, then it rained
  - If the sprinklers are on, the grass is wet

  Does that mean: if sprinklers are on, it rained?

  If anything, it takes away support for such a suggestion.

A system of rules produces coherent updates if and only if rules form a directed tree i.e. no two rules may stem from the same premise. Wet grass here points to two possible explanations.

Why networks?

To make p(B|A) meaningful, we have to show:

- Other items in knowledge base are irrelevant to B
- Better still, make unignorable quickly identifiable and accessible

Neighboring nodes in a graph allow that. What is not local does not matter!

Networks are also used in AI etc. but BN have clear semantics. Most features can be derived from the knowledge base.

These play central role in uncertainty formalisms

- BN
- causal nets
- influence diagrams
- Constraint networks

More important terms

- Likelihood: in blackjack the likelihood of getting 10 is higher (because it can be 10, J, Q or K)
- Conditioning:

$$P(A|C) = p(A,C)/p(C)$$

Belief

Conditioning bar

Knowledge/ context

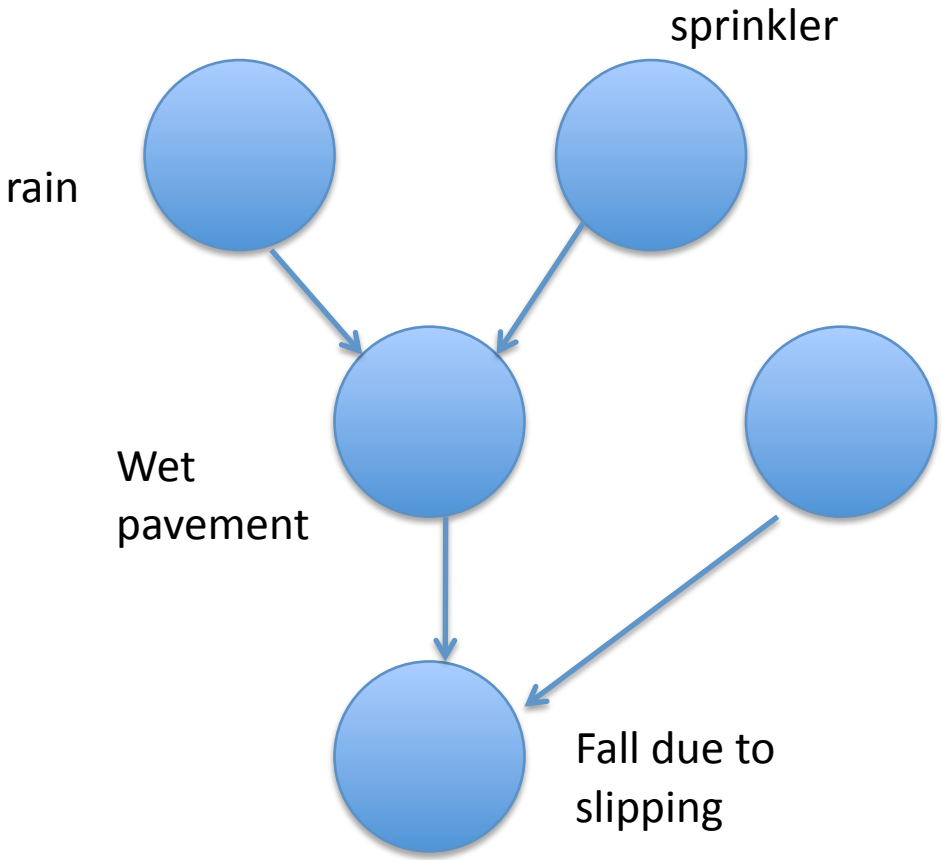
P(fly(a)|bird(a)) = HIGH
P(fly(a)|bird(a),sick(a)) = LOW → (non-zero!)

(retraction possible)

$P(fl|bird) = \alpha * p(fl|bird,sick) + (1-\alpha) * p(fly(bird, NOT\ sick)$ $0<\alpha<1$

- Relevance: potential change in belief due to change in knowledge

- Causation:



Once falling is observed its irrelevant why the pavement was wet

Case study: college plans (Heckerman, 1995, MSR-TR-95-06)

- Sex (SEX: M, F)
- Socioeconomic Status (SES: L, M, U, H)
- IQ (IQ: L, M, U, H)
- Parental encouragement (PE: L, H)
- College plans: (CP: Y, N)

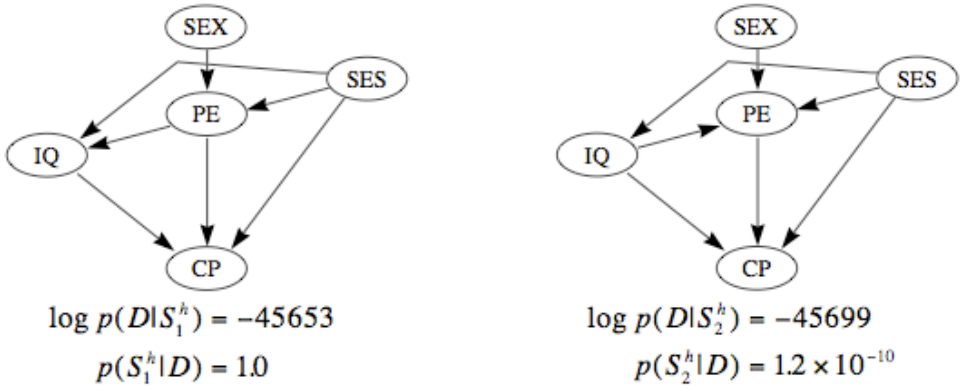
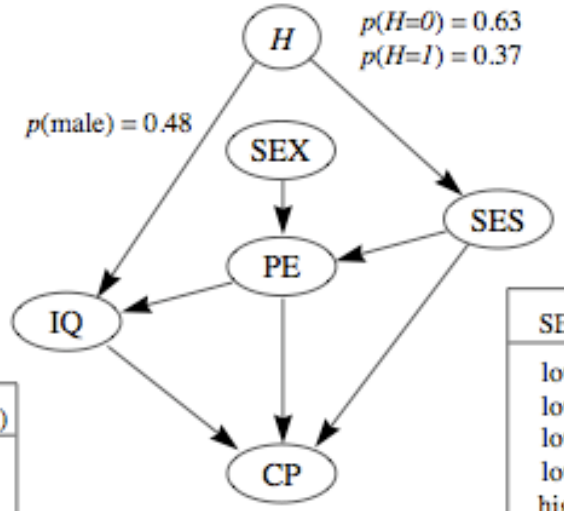


Figure 10: The a posteriori most likely network structures without hidden variables.

Hidden variable



PE	H	p(IQ=high|PE,H)
low	0	0.098
low	1	0.22
high	0	0.21
high	1	0.49

H	p(SES=high|H)
low	0.088
high	0.51

SES	SEX	p(PE=high|SES,SEX)
low	male	0.32
low	female	0.166
high	male	0.86
high	female	0.81

SES	IQ	PE	p(CP=yes|SES,IQ,PE)
low	low	low	0.011
low	low	high	0.170
low	high	low	0.124
low	high	high	0.53
high	low	low	0.093
high	low	high	0.39
high	high	low	0.24
high	high	high	0.84

$$\log p(S^h|D) \cong -45629$$

Using the terms just explained we can use probability for describing qualitative phenomena.

One can see if refinements, extensions are possible. If it is based on theory, we can understand exactly what adjustments need to be made.

The inversion formula

$p(H|e) = p(e|H)*p(H)/p(e)$

Posterior=likelihood*prior/normalizing constant

$p(e) = p(e|H)*p(H) + p(e|\ \text{NOT H})*p(\text{NOT H})$

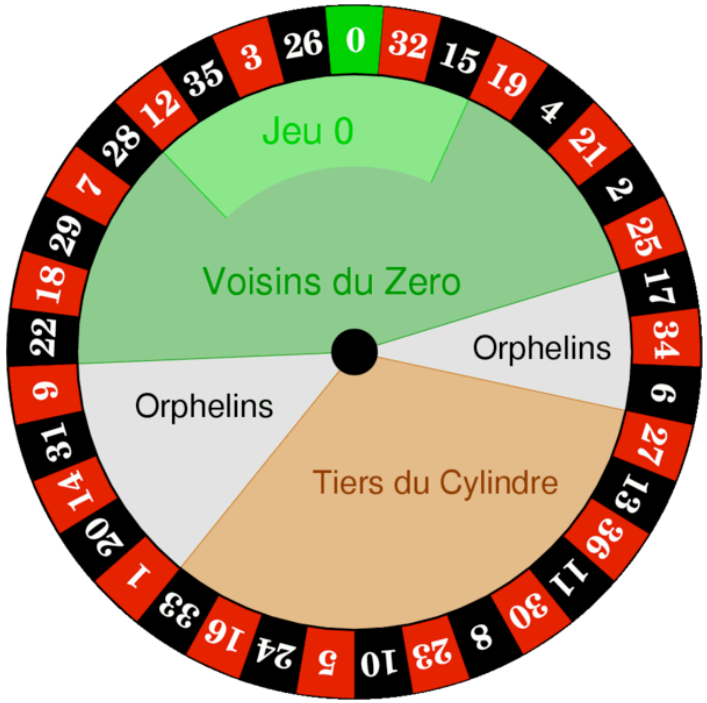
The formula seems to come from:

$p(A|B) = p(A,B)/p(B)$ and $p(B|A) = p(A,B)/p(A)$

Assesing p(H|e)

In a gambling room someone calls 12

Is it from a pair of dice, or from a roulette wheel?



- For dice: p(e|H) = p(12|dice) = 1/36
- For roulette: p(e|H) = p(12|roulette) = 1/38

Thus if there are more than 38/36 roulette wheels in the room, p(roulette) is more likely

Talking about p(roulette|12) would have been much more difficult.

3 prisoner problem



1 prisoner. 3 doors. 2 lead to death, 1 to escape. S/He is asked to choose one. Once he has indicated his choice, one of the other doors is indicated to be leading to death. He is given a chance to switch to the third door. Should he?

What if there are 1000 doors (999 leading to death)? Should he switch?

The Bayesian twist

Of three prisoners A, B, C only one is going to be hanged and the other two pardoned.

A says to guard: Give this letter to one of the pardoned ones.

An hour later, A asks the guard: tell me who did you give it to?

The guard answers “B”.

A reasons: So either C will be hanged, or I will be. Prob. That I will be is 50%. Is he right?

$p(G\_A|I\_B) = p(I\_B|G\_A)*p(G\_A)/p(I\_B)$

$= 1*(1/3)/(2/3) = ½$

So, applying Bayesian logic incorrectly leads to false results. The error here is misinterpreting the context.

Rather than: I_B = B will be released,

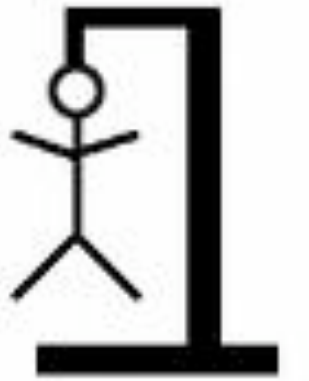
I'_B = Guard said B will be released.

$p(G\_A|I'\_B) = p(I'\_B|G\_A)*p(G\_A)/p(I'\_B)$

$=(1/2)*(1/3)/(1/2) = 1/3$

Case of 1000?

- 1000 prisoners of which 1 is to be put to death
- A finds a list of 998 to be released without his name on it
- What should our belief be that he will be the one being put to death?



Case of 1000?

- List of 998 to be released
- Query associated with the printout: 998 right handers
- A is left-handed
- ??

Difficult to treat such ignorance in general.

- Multivalued hypothesis
- Uncertain evidence
- Virtual (intangible) evidence
- Predicting future events
- Multiple causes, explaining away
- Patterns of plausible reasoning
- …

Naïve Bayes

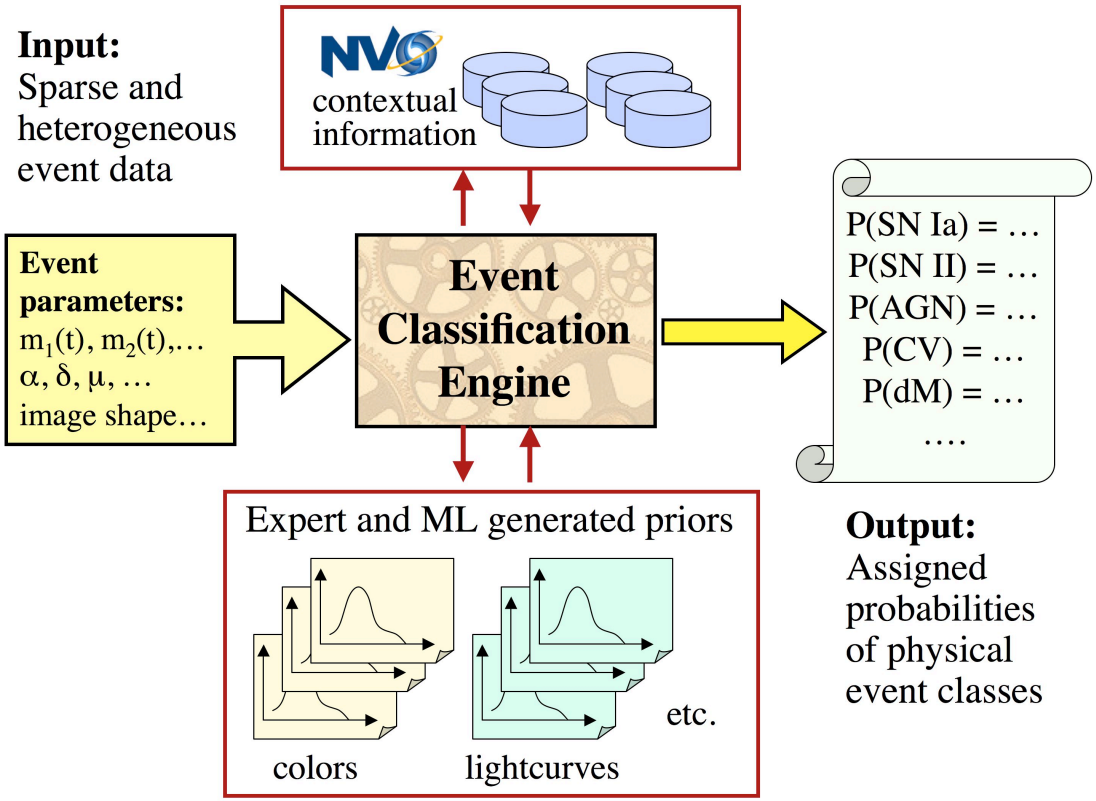
$$P(y = k \mid x) = P(x \mid y = k)P(k)/P(x) \propto P(k)P(x \mid y = k) \approx P(k)\prod_{b=1}^{B} P(x_b \mid y = k)$$

- x: feature vector of event parameters
- y: object class that gives rise to x (1<y<k)
- Certain features of x known:
  - Position
  - Flux at observed wavelength
- Others will be unknown
  - Color
  - Change in mag/flux over time baselines

Naïve Bayes (contd.)

$$P(y = k \mid x) = P(x \mid y = k)P(k)/P(x) \propto P(k)P(x \mid y = k) \approx P(k)\prod_{b=1}^{B} P(x_b \mid y = k)$$

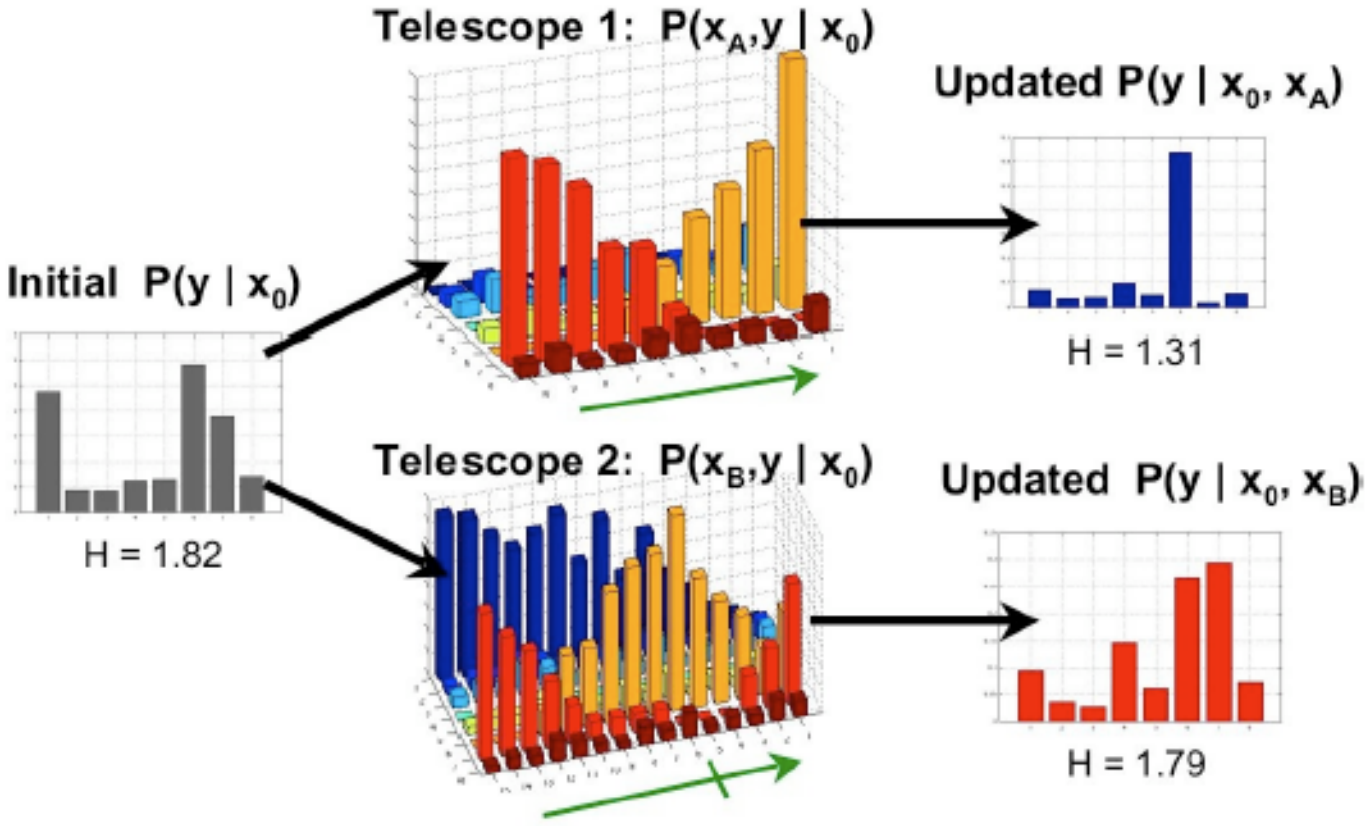
- Assumption: based on y, x is decomposable into B distinct independent classes (labeled x_b)
- This helps with the curse of dimensionality
- Also allows us to deal with missing values
- Alternate parallel supplemental supervised classification
  - Automated Neural Networks (ANN)
  - Support Vector Machines (SVM)


Input:
Sparse and heterogeneous event data
NVO
contextual information
Event parameters:
$m_1(t)$, $m_2(t)$,…
α, δ, μ, …
image shape…
Event Classification Engine
P(SN Ia) = …
P(SN II) = …
P(AGN) = …
P(CV) = …
P(dM) = …
….
Output:
Assigned probabilities of physical event classes
Expert and ML generated priors
etc.
colors
lightcurves

Follow-up (for missing values)

- Such that it will help discriminate better
- Serve probabilities so that consumers can choose their types of transients
- Widest possible models

Choosing follow-up configs



r-i color, hi-z quasar, blue star

Transient classification mantra

- Obtain a couple of epochs in one or more filters
- Assigns probabilities for different classes
- Choose observations (filters, wavelengths) for best discrimination
- Feed the new observations back in
- Revise probabilities, choose observations, …
- Based on confirmed class (how?) revise priors

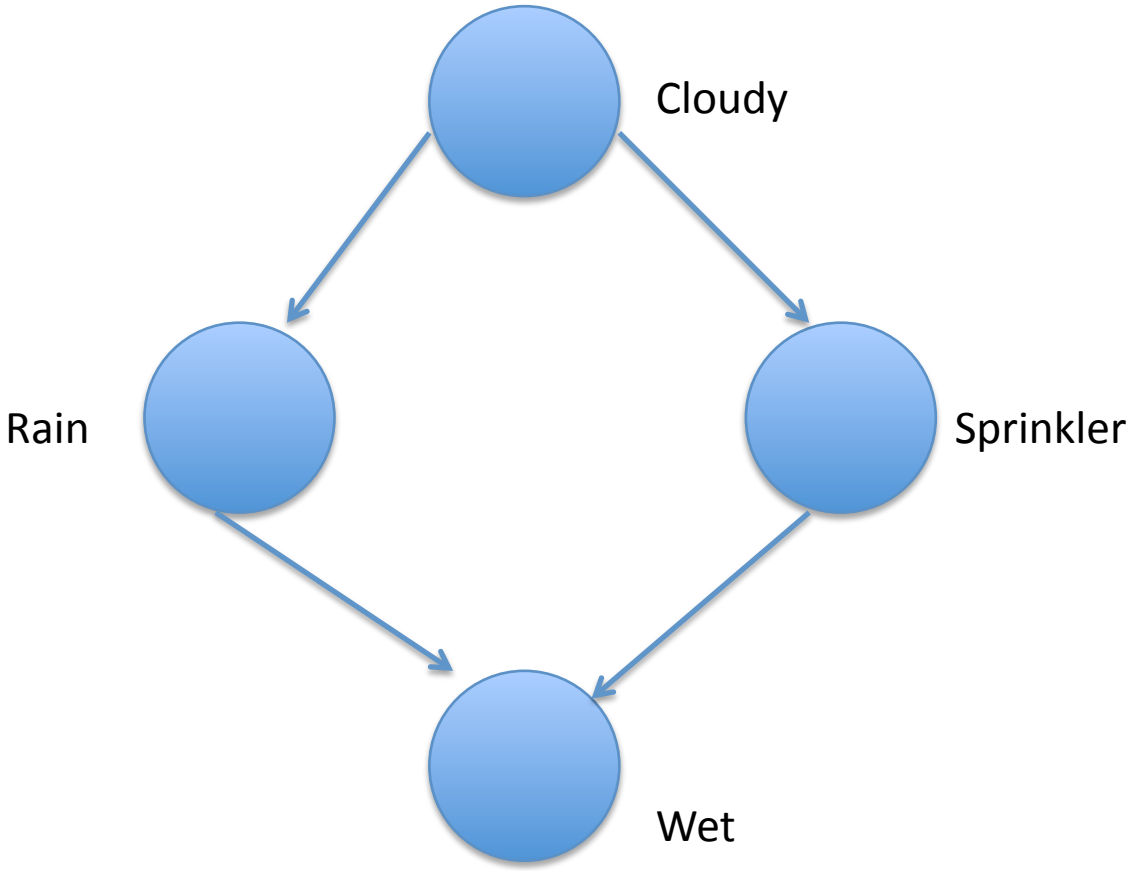
Summary

- Modeling is all important (to predict/explore/ explain)
- Local dependencies, irrelevancies to be evaluated
- Priors, likelihoods to be obtained
- Directed Acyclic Graph to be constructed
- Data define network
- No “training” necessary

Bayesian Network Toolbox

http://bnt.googlecode.com

A Bayesian Network



Creating a DAG

To specify this directed acyclic graph (dag), we create an adjacency matrix:

```
N = 4;
dag = zeros(N,N);
C = 1; S = 2; R = 3; W = 4;
dag(C,[R S]) = 1;
dag(R,W) = 1;
dag(S,W)=1;
```

Multivalued nodes

```
discrete_nodes = 1:N;
node_sizes = 2*ones(1,N);
```

If the nodes were not binary, you could type e.g.,

```
node_sizes = [4 2 3 5];
```

Making the bnet

```
bnet = mk_bnet(dag, node_sizes, 'discrete', discrete_nodes);
```

By default, all nodes are assumed to be discrete, so we can also just write

```
bnet = mk_bnet(dag, node_sizes);
```

Naming parameters

It is possible to associate names with nodes, as follows:

```
bnet = mk_bnet(dag, node_sizes, 'names', {'cloudy','S','R','W'}, 'discrete', 1:4);
```

You can then refer to a node by its name:

```
C = bnet.names('cloudy'); % bnet.names is an associative array
bnet.CPD{C} = tabular_CPD(bnet, C, [0.5 0.5]);
```

Conditional probability distribution

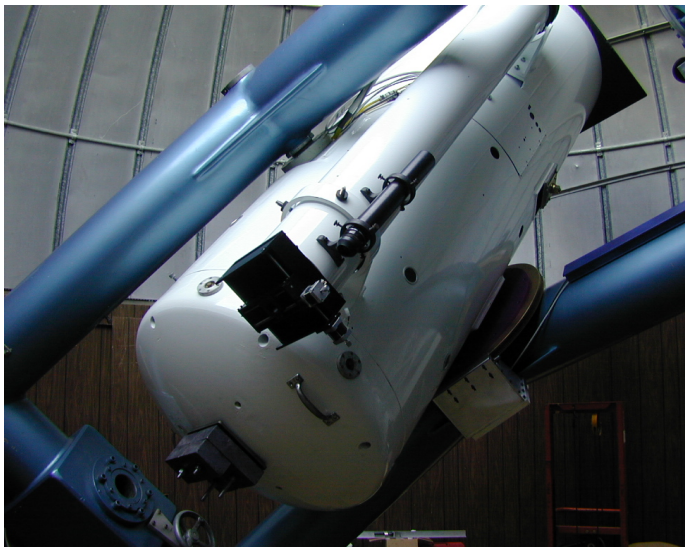
```
bnet.CPD{C} = tabular_CPD(bnet, C, [0.5 0.5]);
bnet.CPD{R} = tabular_CPD(bnet, R, [0.8 0.2 0.2 0.8]);
bnet.CPD{S} = tabular_CPD(bnet, S, [0.5 0.9 0.5 0.1]);
bnet.CPD{W} = tabular_CPD(bnet, W, [1 0.1 0.1 0.01 0 0.9 0.9 0.99]);
```

Entering evidence

```
evidence = cell(1,N);
evidence{W} = 2;
engine = enter_evidence(engine, evidence);
m = marginal_nodes(engine, W);
m.T
ans =
     1
```

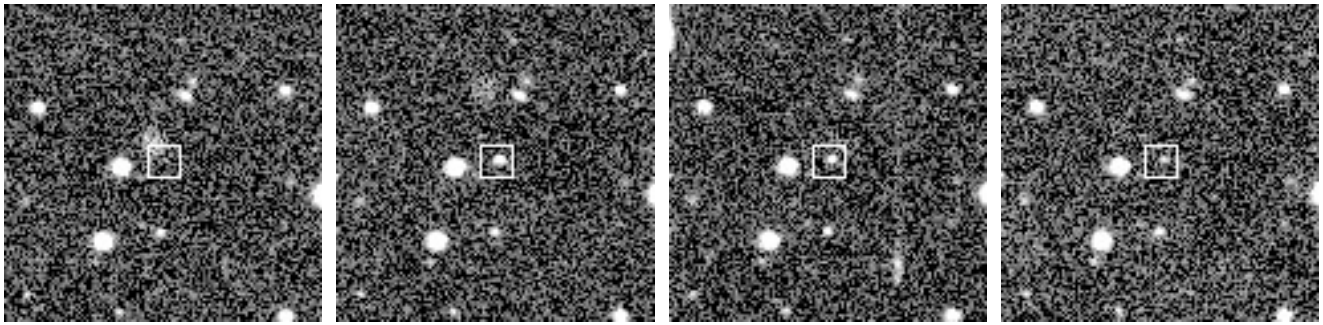
The Catalina Realtime Transient Survey

CRTS is a search for transients being done at Caltech piggybacking on the data from the search for near-Earth, potentially hazardous asteroids (this later is led by S. Larson, E. Beshore, et al. at UAz LPL). The survey uses the 24-inch Schmidt on Mt. Bigellow, and a single, unfiltered 4kx4k CCD (and also telescopes at Mt. Lemmon and Siding Spring). Coverage of well over 1000 deg^2/night

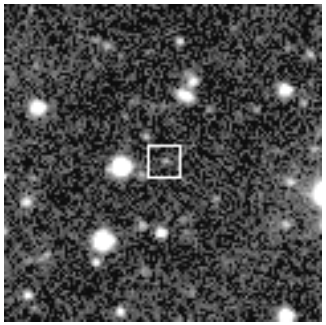


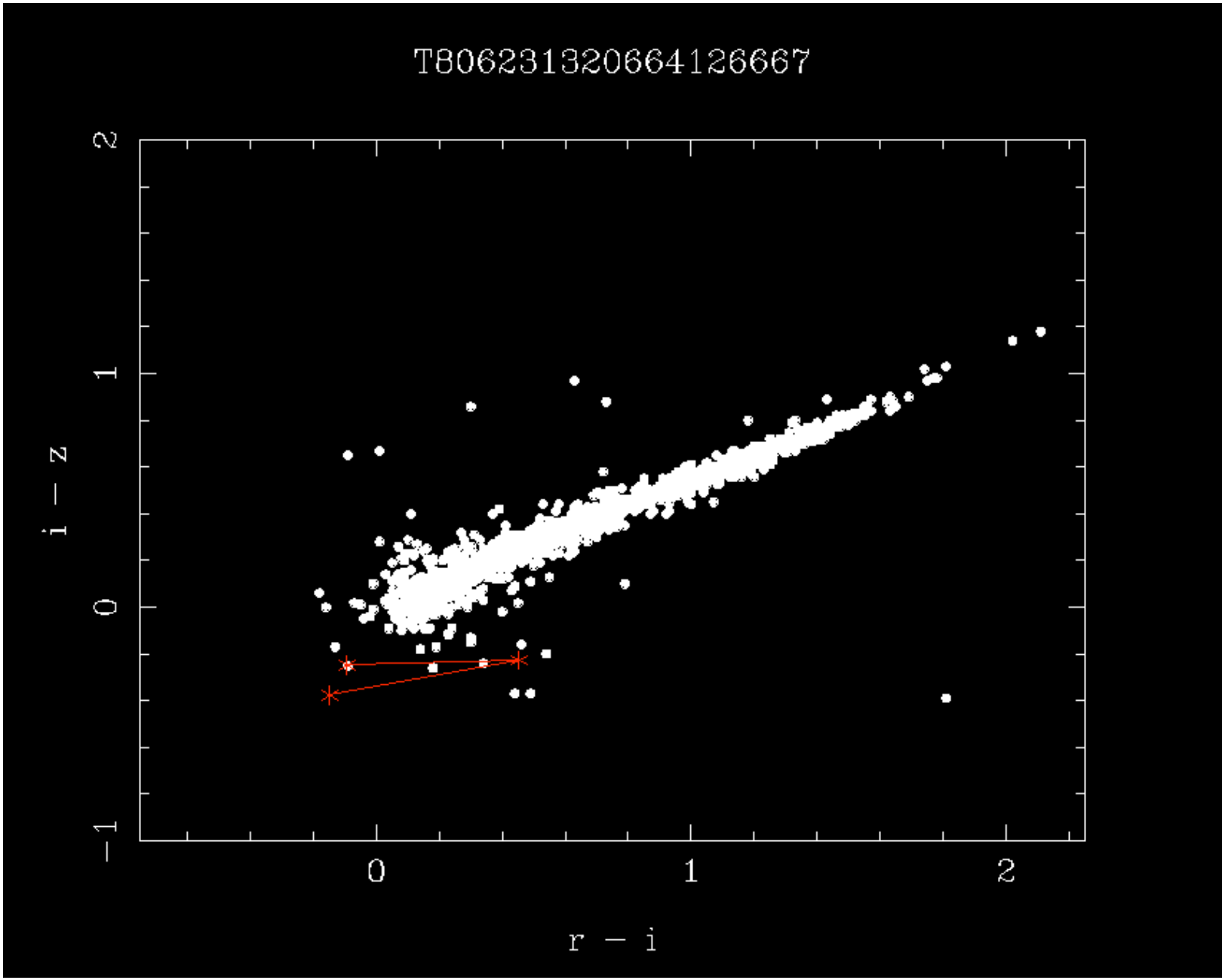
Catalina Survey Fast Transient (a flare star), 02 Nov 2007 UT:

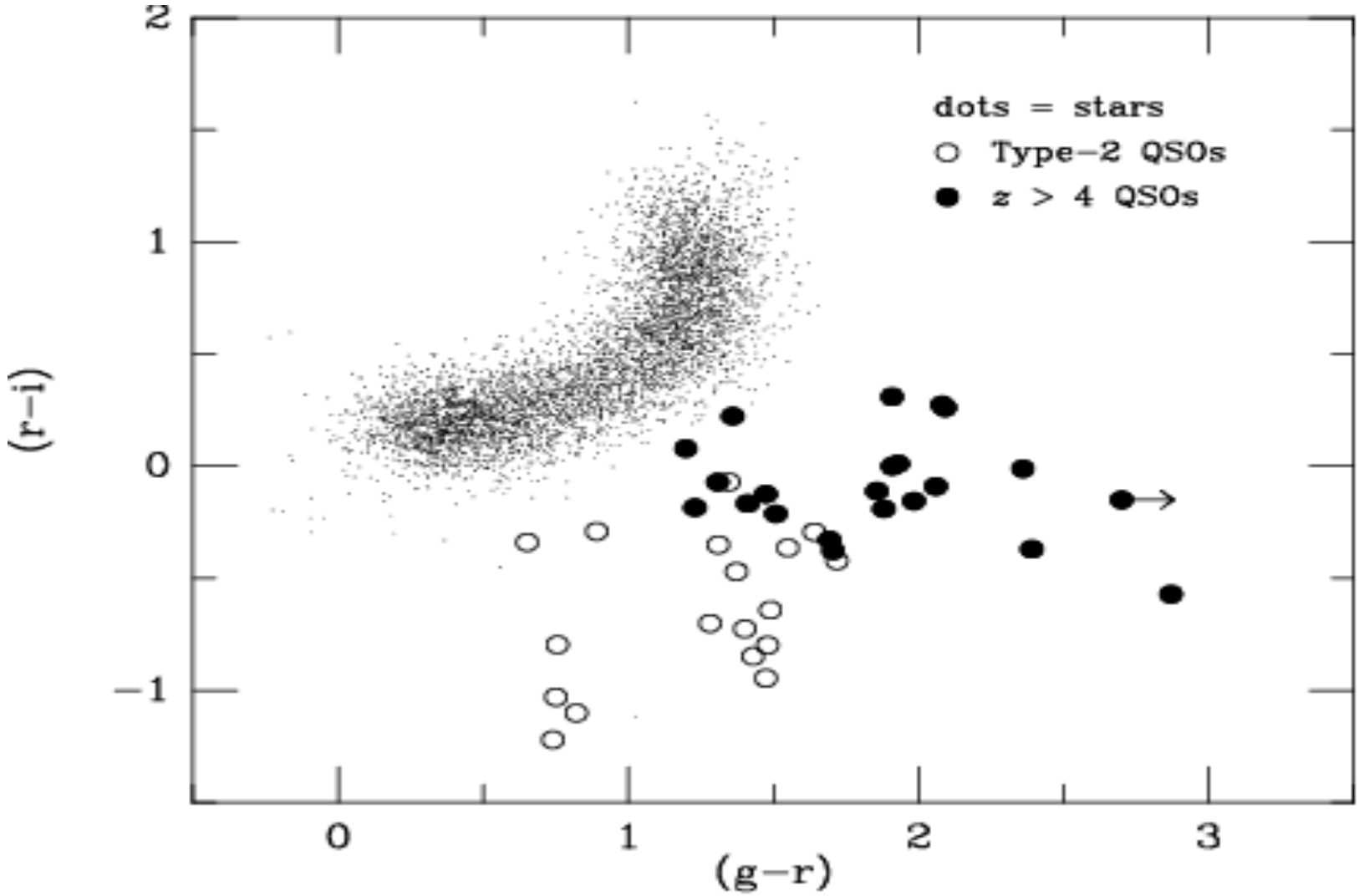
4 individual exposures, separated by 10 min



Baseline coadd:

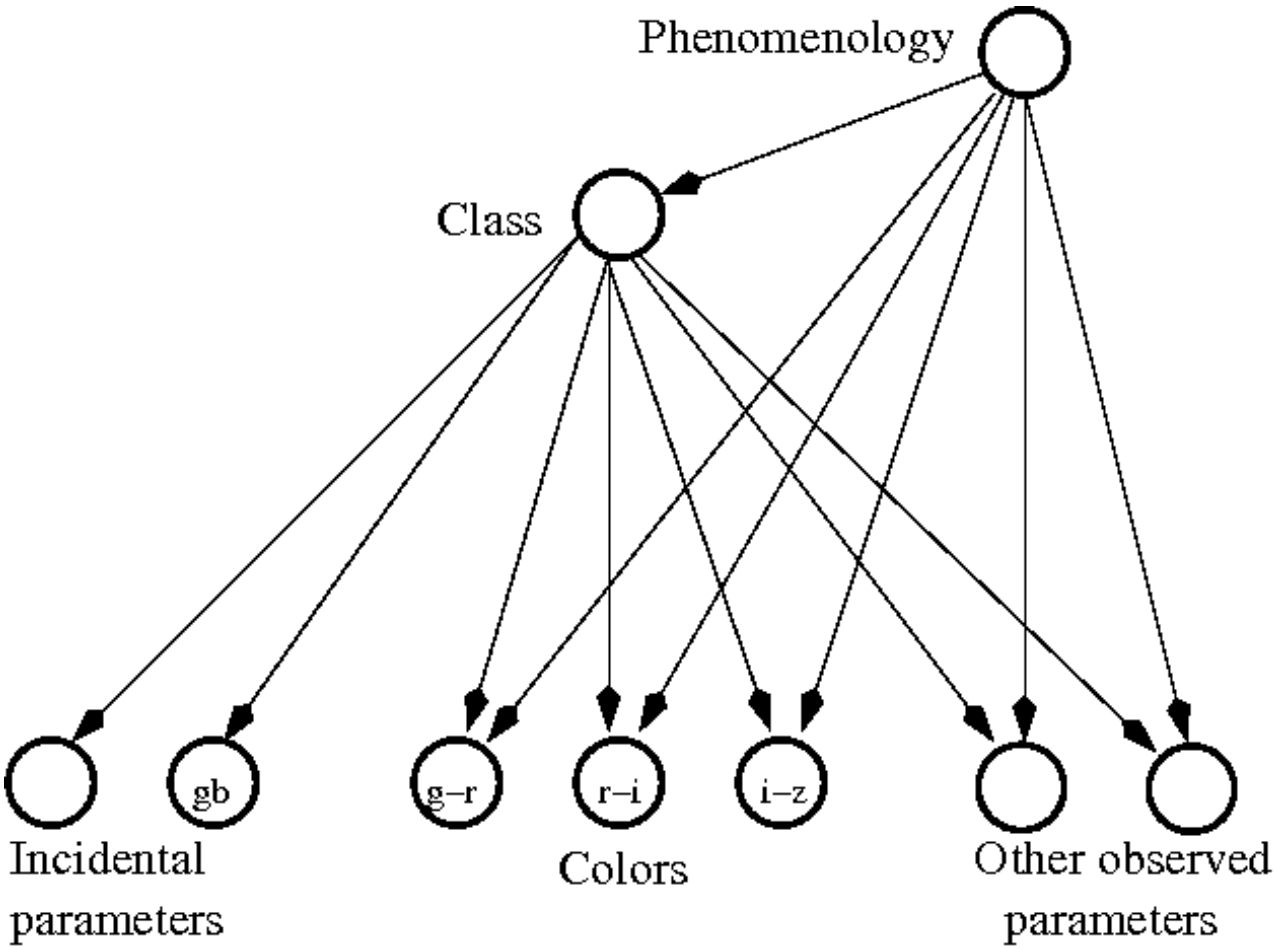


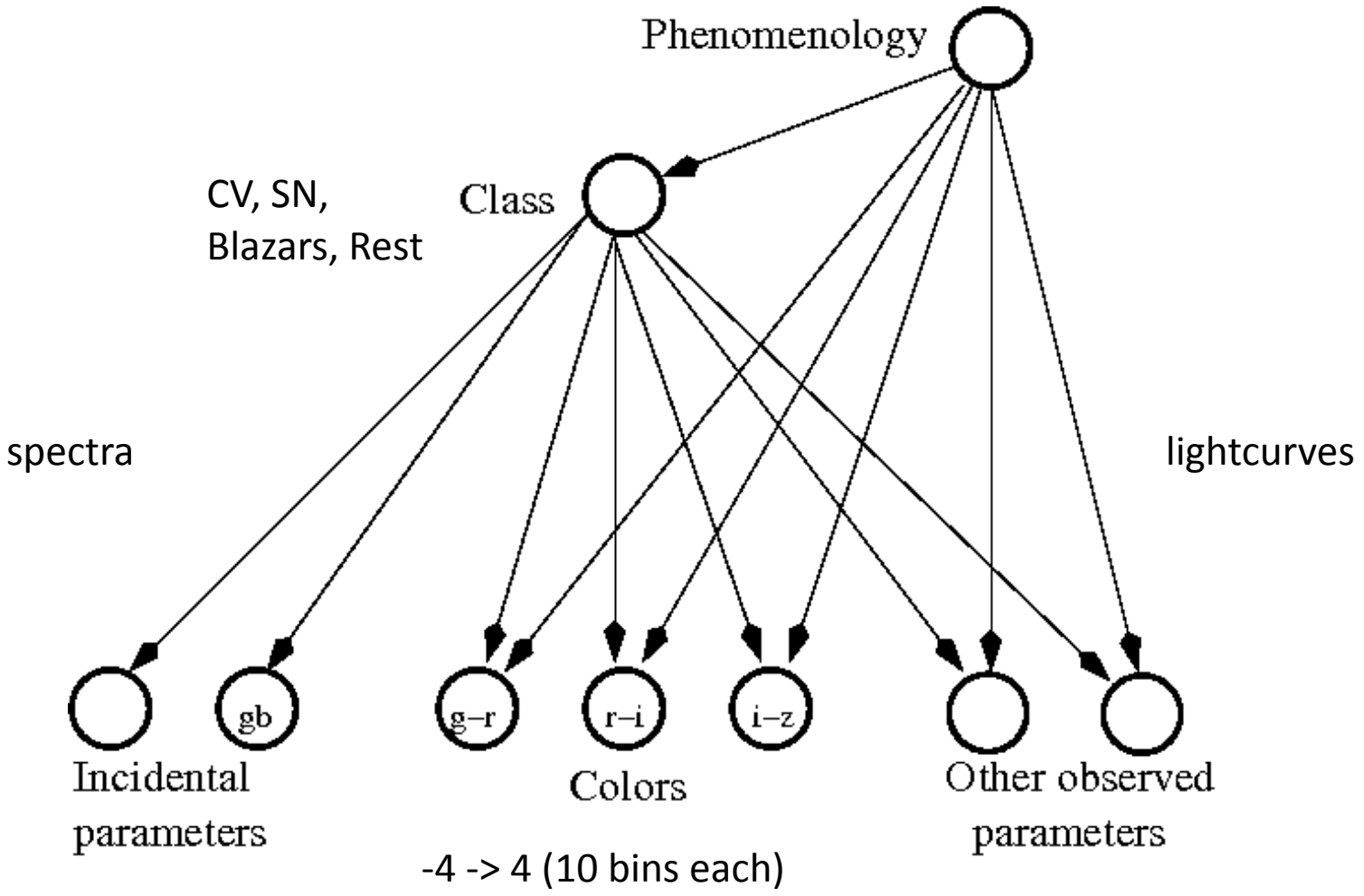

TB06231320664126667
2
1
0
−1
i − z
0
1
2
r − i


dots = stars
Type−2 QSOs
z > 4 QSOs
(r−i)
(g−r)
2
1
0
−1
0
1
2
3

Basic astronomy classification trivia context based information

- Galactic latitude – Galacticness
- Proximity to a galaxy – SN


Phenomenology
Class
gb
g−r
r−i
i−z
Incidental parameters
Colors
Other observed parameters



-4 -> 4 (10 bins each)

- N=5
- Dag = zeroes(N,N)
- C=1;g=2;c1=3;c2=4;c3=5;
- Dag(c,[g,c1,c2,c3])=1
- Discrete_nodes(1:N)
- Node_sizes=[4,10,10,10,10]
- Bnet=mk_bnet(dag,node_sizes,names, {'class','galactic_latitude','g-r','r-i','i-z'},'discrete', 1:5)

Advantages of Bayesian Networks

- Handling of incomplete data
  - Real-world cases
- Learning causal connections
  - What variable caused what
- Incorporating domain knowledge
  - Experts can weight in at different points
- Memorizing (aka overfitting) avoided
  - No holdout necessary