Constructing Hypothesis Lattices for Virtual Experiments in Data Intensive Research

Dmitry Kovalev, Sergey Stupnikov

Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences

Motivation

- Science is increasingly dependent on data as the core source for discovery
- Data deluge has affected the way scientific experiment is done
  - X-informatics were the first to deal with it
  - Providing valuable insight into how modern data intensive research is done (scheme, methods, algorithms, etc.)
- Hypothesis remains the central research unit in DIR

Research hypotheses in Data Intensive Domains

From classical hypothesis to the DIS hypothesis

- Mathematical equation
  - $a(t) = -g$
  - $v(t) = -gt + v_0$
  - $s(t) = -(g/2)t^2 + v_0 t + s_0$
- Existentional formula
  - $\forall x \in X \, \forall y \in Y, \quad p(x) \rightarrow q(y)$

Database relation

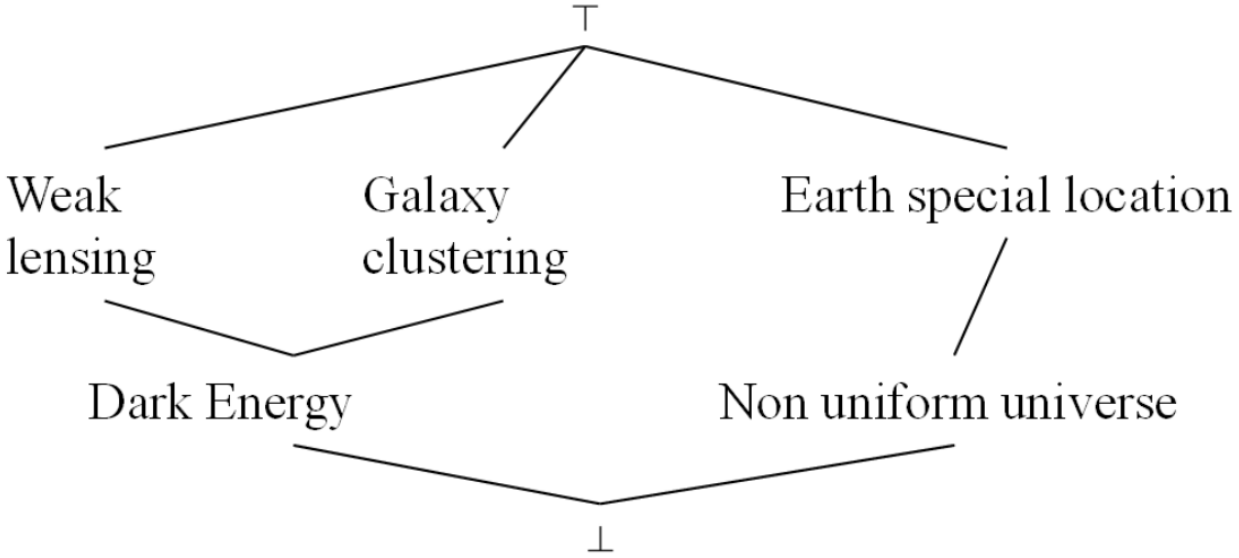
t	v	s
0	0	5000
1	-32	4984
2	-64	4936
3	-96	4856
4	-128	4744

Algorithm

```
for k = 0:n;
    t = k * dt;
    v = -g*t + v_0;
    s = -(g/2)*t^2 +
        v_0*t + s_0;
    t_plot(k) = t;
    v_plot(k) = v;
    s_plot(k) = s;
end
```

- Multiple ways to represent research hypotheses
- Focus is on formula representations

Hypotheses interaction



Definition

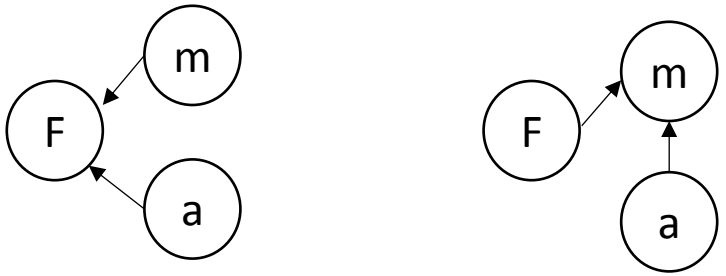
A hypothesis lattice is formed by considering a set of hypotheses equipped with wasDerivedFrom as a strict order $<$ (from the bottom to the top). Hypotheses directly derived from exactly one hypothesis are *atomic*, while those directly derived from at least two hypotheses are *complex*.

Why interaction is important?

… However, we want to emphasize that the *γ* parameter is correlated with the values of other parameters used in the model and especially with the slopes of *IMF* and the age of the disc. …
This multi-dependency and interplay between different model's ingredients oblige us to always look for the best global fit.

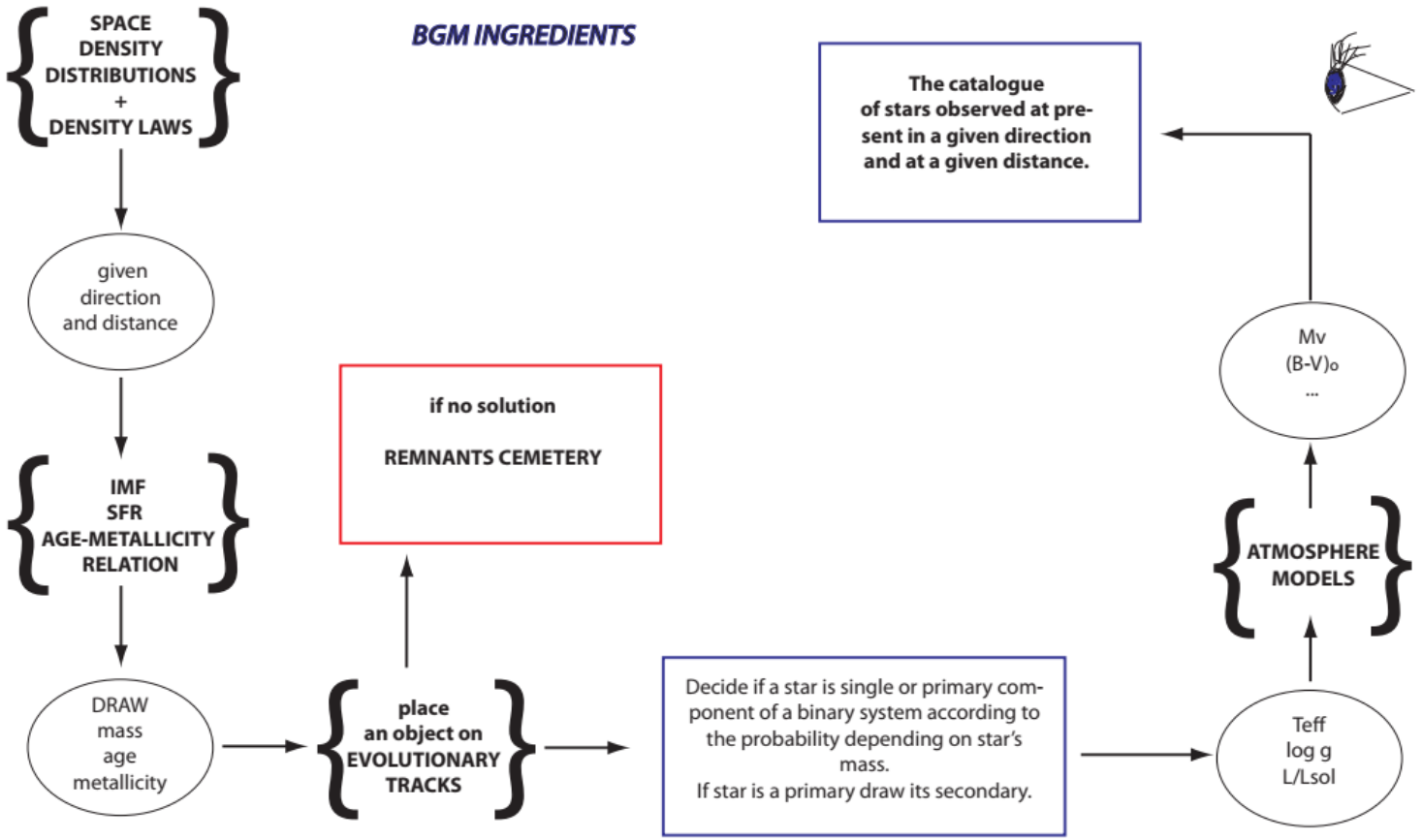
Building Hypotheses Lattices

- System of linear equation are parsed into graphs of variables (Casual Ordering Algorithm and Problem)



- Not always understand causality order between several systems of equations
- Workflow defines the order of hypotheses invocation

Workflow example

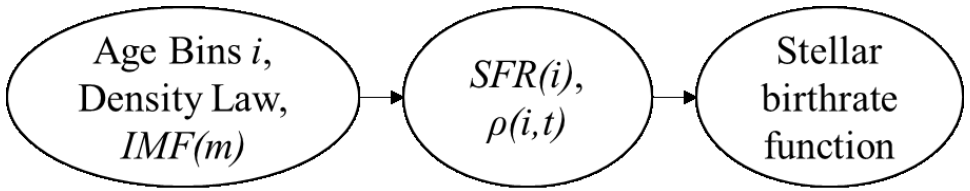


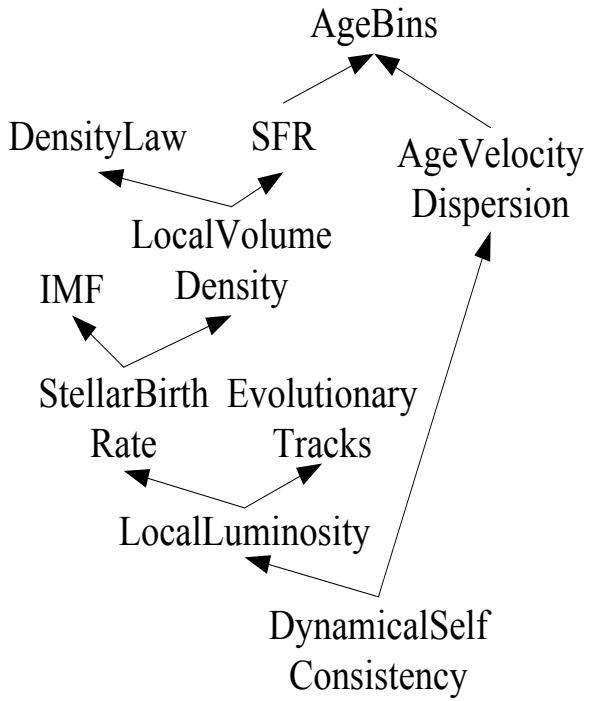
** From M. Czekaj PhD Thesis*

Algorithm:
Construction of Hypotheses Lattice

- As the input, the algorithm takes a workflow *W* and the set of hypotheses *H*
- System of equations -> direct causality graphs of variables -> workflow provides who comes first
- Hypothesis lattice *L* is returned
- Complexity of the algorithm is bounded by $O(|W|^2*|S|*|V|*|H|)$

Besancon Galaxy Model as example



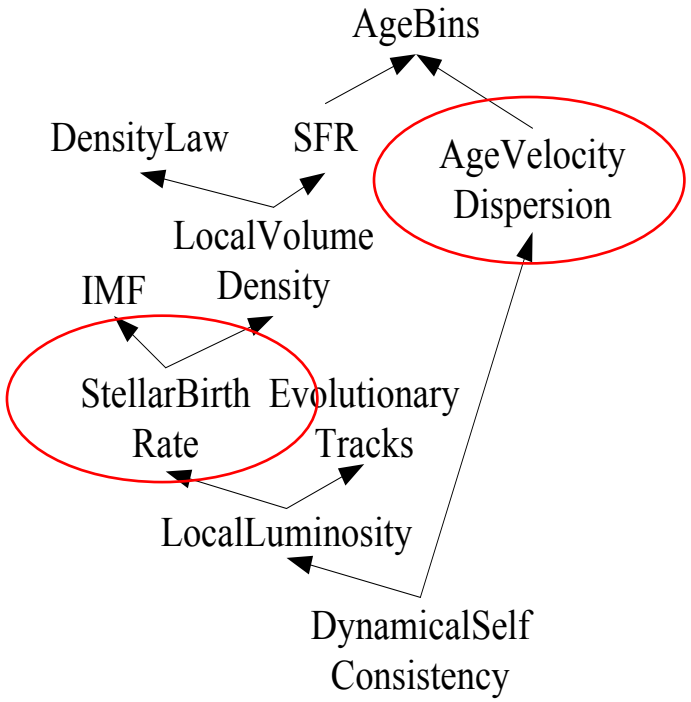


$$SFR(i) = exp(\gamma \times x_c(i)) \times d$$

$$IMF(m) = m^{-\alpha}$$

$$b(m, t)dmdt \approx f(SFR(i), \rho(i, SFR(i)) \times IMF(m)\Delta m \Delta i$$

Besancon Galaxy Model as example



- Now can change some parameters independently
- Still no tracking of meaningful correlations between variables