HEALTHY BIRTH,
GROWTH & DEVELOPMENT

New Information-Theory-Based Methods in the Analysis of Childhood Development Data

Nikita A. Sakhanenko,[1] David J. Galas,[1] Representing the Healthy Birth, Growth, and Development knowledge integration (HBGD*ki*) Community[2]

[1]Pacific Northwest Diabetes Research Institute, Seattle, WA, USA; [2]Bill & Melinda Gates Foundation, Seattle, WA, USA

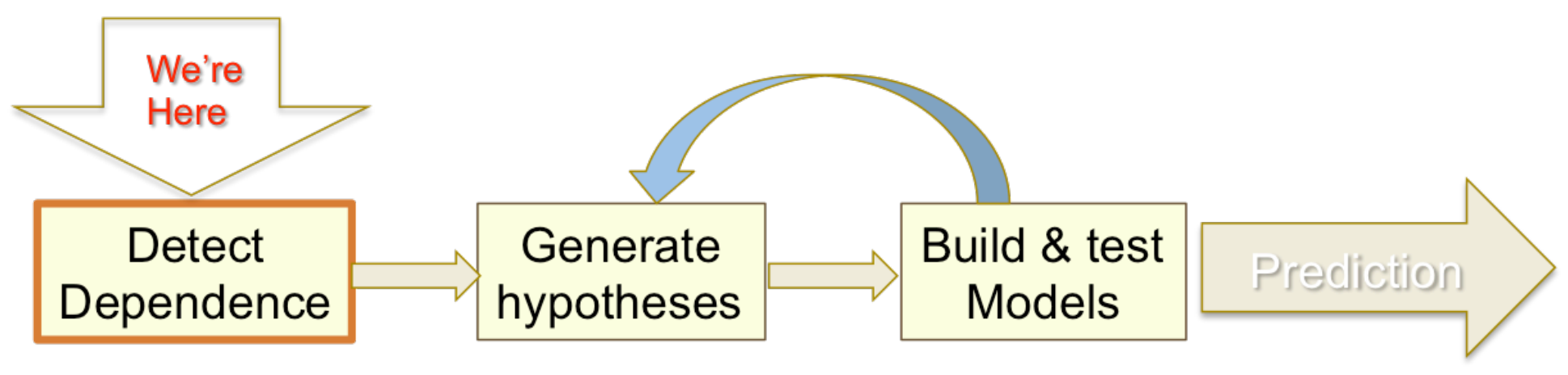
Introduction

The complexity of infant growth and development processes and resultant data reflect the deep complexity of biological systems.

The Problem: Can we detect complex dependencies of biological variables?

Given data sets with many instances of many variables:

- Can we find a measure for a subset of variables that is significantly nonzero if and only if the variables are interdependent?
- Can we separate detection of the existence of dependence from any models of the nature of dependence?


We're Here
Detect Dependence
Generate hypotheses
Build & test Models
Prediction

Dependency Measures

Information about one variable is contained in information about another.

Mutual information
$I(X_i, X_j) = H_i + H_j - H_{ij}$

"Interaction information"

$$I(X_i, X_j, X_k) = I(X_i, X_j) - I(X_i, X_j | X_k)$$

$$I(X_i,X_j,X_k) = H(X_i)+H(X_j)+H(X_k)-H(X_i,X_j)-H(X_i, X_k)-H(X_j,X_k)+H(X_i,X_j,X_k)$$

"Conditional mutual information, or differential interaction information"

$$-I(X_i, X_j | X_k) = I(X_i, X_j, X_k) - I(X_i, X_j)$$

$$\Delta_{ijk} = -I(X_i, X_j | X_k) = H_k - H_{ki} - H_{kj} + H_{ijk}$$

General dependency measures:

based on interaction information

Symmetric deltas

If $\{\tau_m\}$ are all subsets of $\nu_m \subseteq V_m$ that contain X_m

$$\Delta(V_{m-1}; X_m) = I(V_m) - I(V_{m-1}) = \sum_{\{\tau_m\}} (-1)^{|\tau_n|+1} H(\tau_m)$$

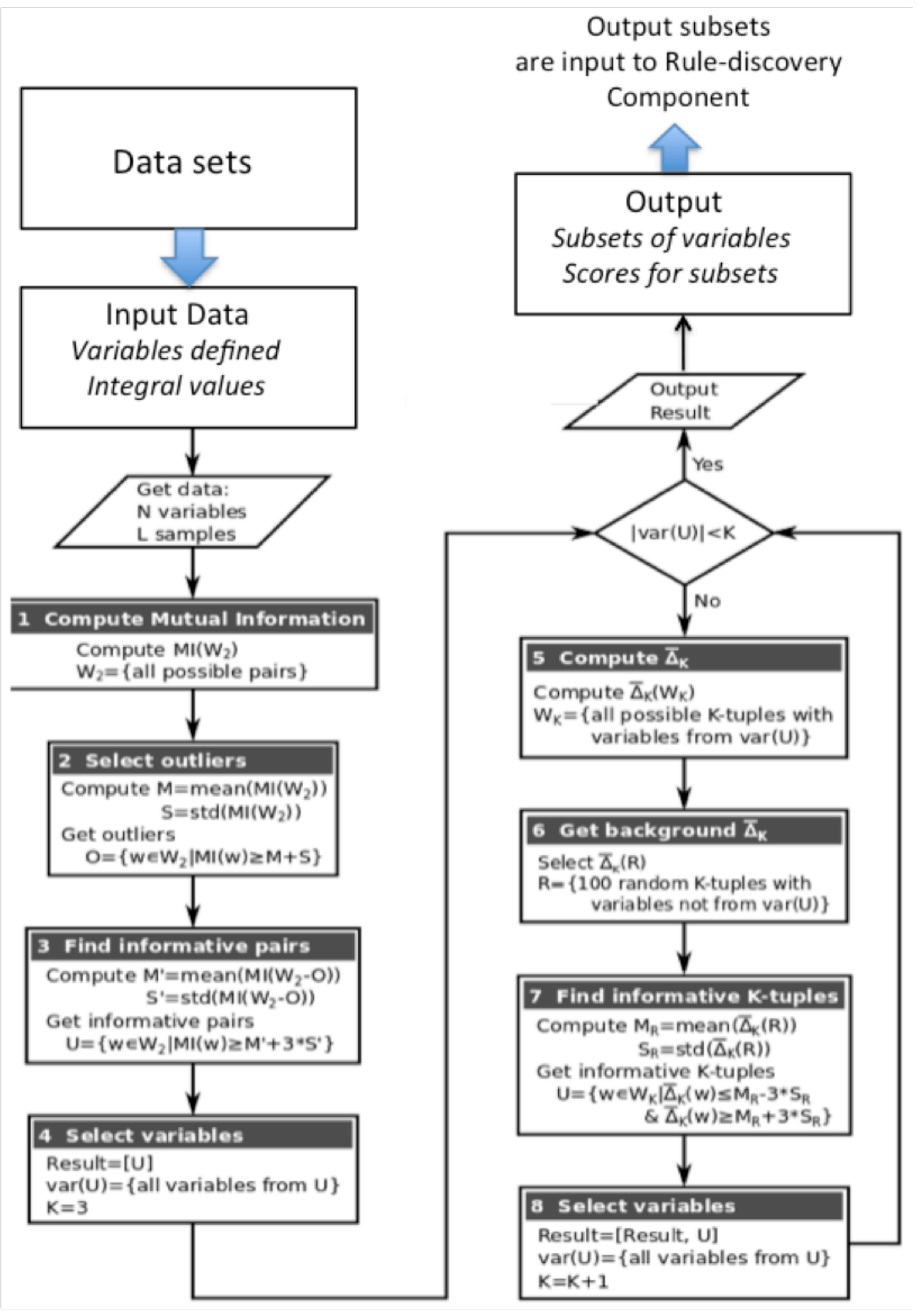
The general, symmetric measure is then

$$\bar{\Delta}(V_m) \equiv \prod_{permutations} \Delta(V_{m-1}; X_m)$$

Galas and Sakhanenko, J. Comp. Biol. (2014)

Avoiding the Combinatorial Explosion: the "Shadows Algorithm"

- For *N* variables, and subsets of size *m*,
- Number of combinations increases like $\sim N^m$
- For ~100,000 variables *m* = 3 & 4 gives 10^{15} & 10^{20}
- "Shadows algorithm" tracks *shadows* of the high degree dependencies in lower degree calculations.


Output subsets are input to Rule-discovery Component

Data sets

Input Data
Variables defined
Integral values

Get data:
N variables
L samples

1 Compute Mutual Information
Compute MI(W_2)
W_2={all possible pairs}

2 Select outliers
Compute M=mean(MI(W_2))
S=std(MI(W_2))
Get outliers
O={w∈W_2|MI(w)≥M+S}

3 Find informative pairs
Compute M'=mean(MI(W_2-O))
S'=std(MI(W_2-O))
Get informative pairs
U={w∈W_2|MI(w)≥M'+3*S'}

4 Select variables
Result=[U]
var(U)={all variables from U}
K=3

|var(U)|<K
Yes
No

5 Compute $\bar{\Delta}_K$
Compute $\bar{\Delta}_K(W_K)$
W_K={all possible K-tuples with variables from var(U)}

6 Get background $\bar{\Delta}_K$
Select $\bar{\Delta}_K$(R)
R={100 random K-tuples with variables not from var(U)}

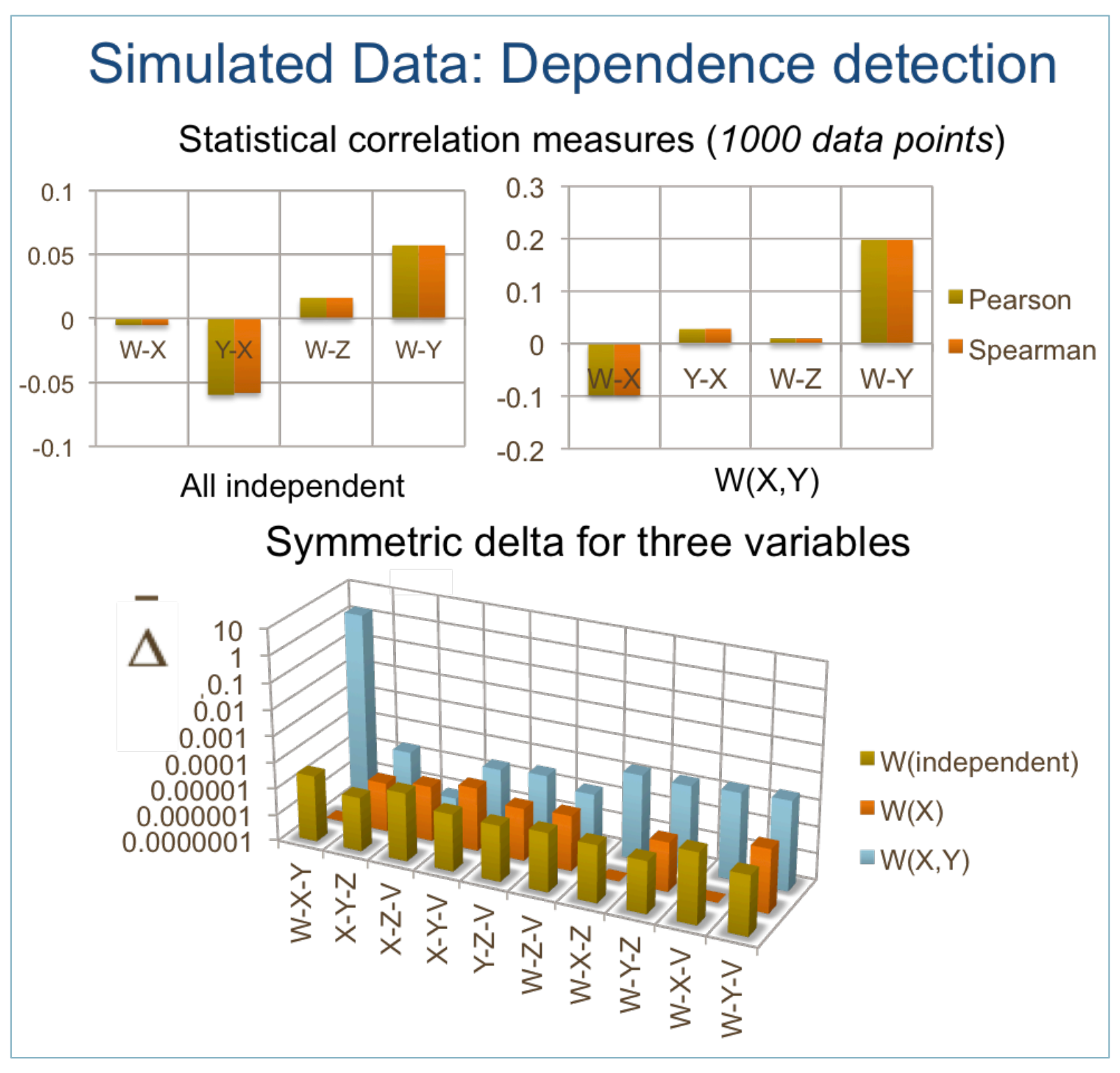
7 Find informative K-tuples
Compute M_R=mean($\bar{\Delta}_K$(R))
S_R=std($\bar{\Delta}_K$(R))
Get informative K-tuples
U={w∈W_K|$\bar{\Delta}_K$(w)≤M_R-3*S_R & $\bar{\Delta}_K$(w)≥M_R+3*S_R}

8 Select variables
Result=[Result, U]
var(U)={all variables from U}
K=K+1

Output Result

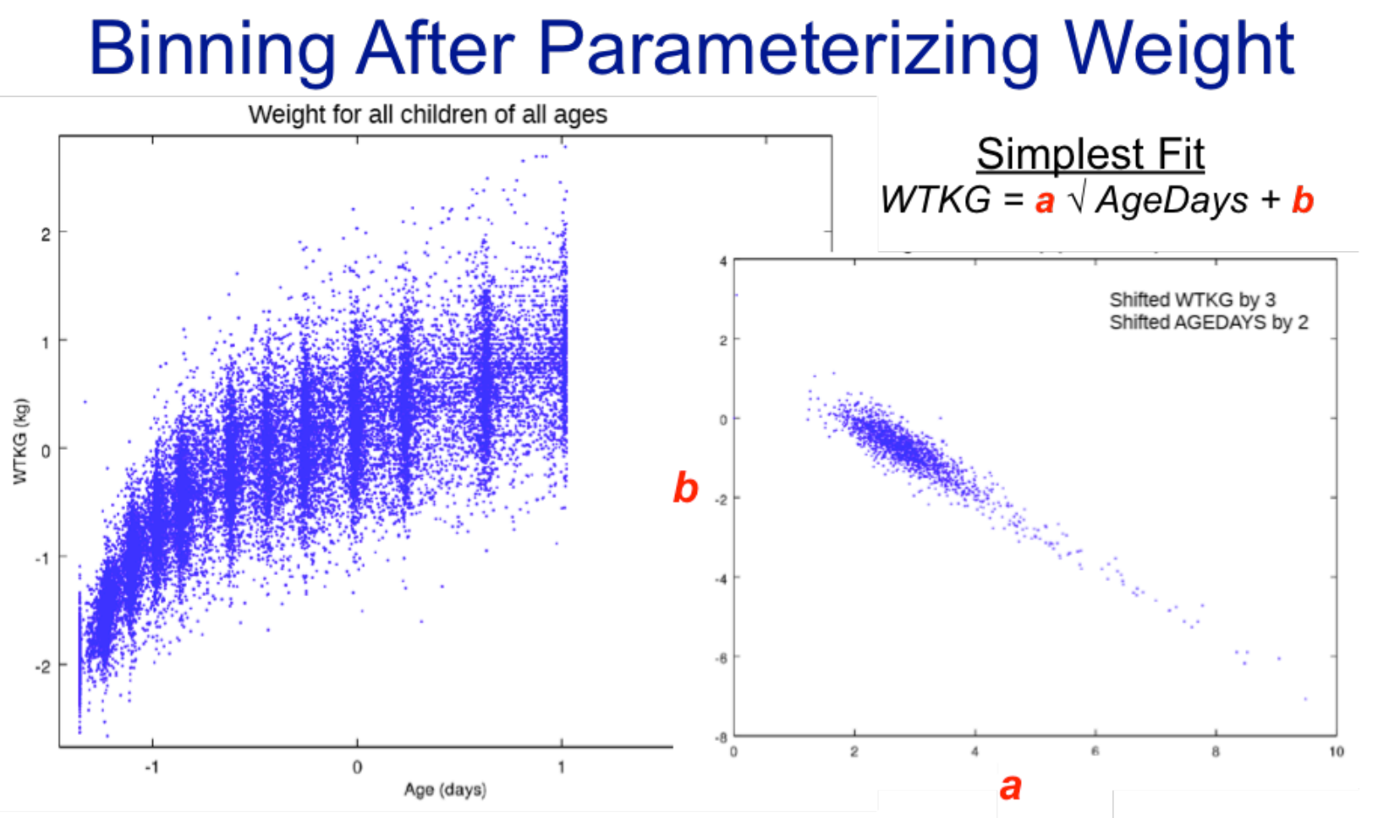
Output
Subsets of variables
Scores for subsets

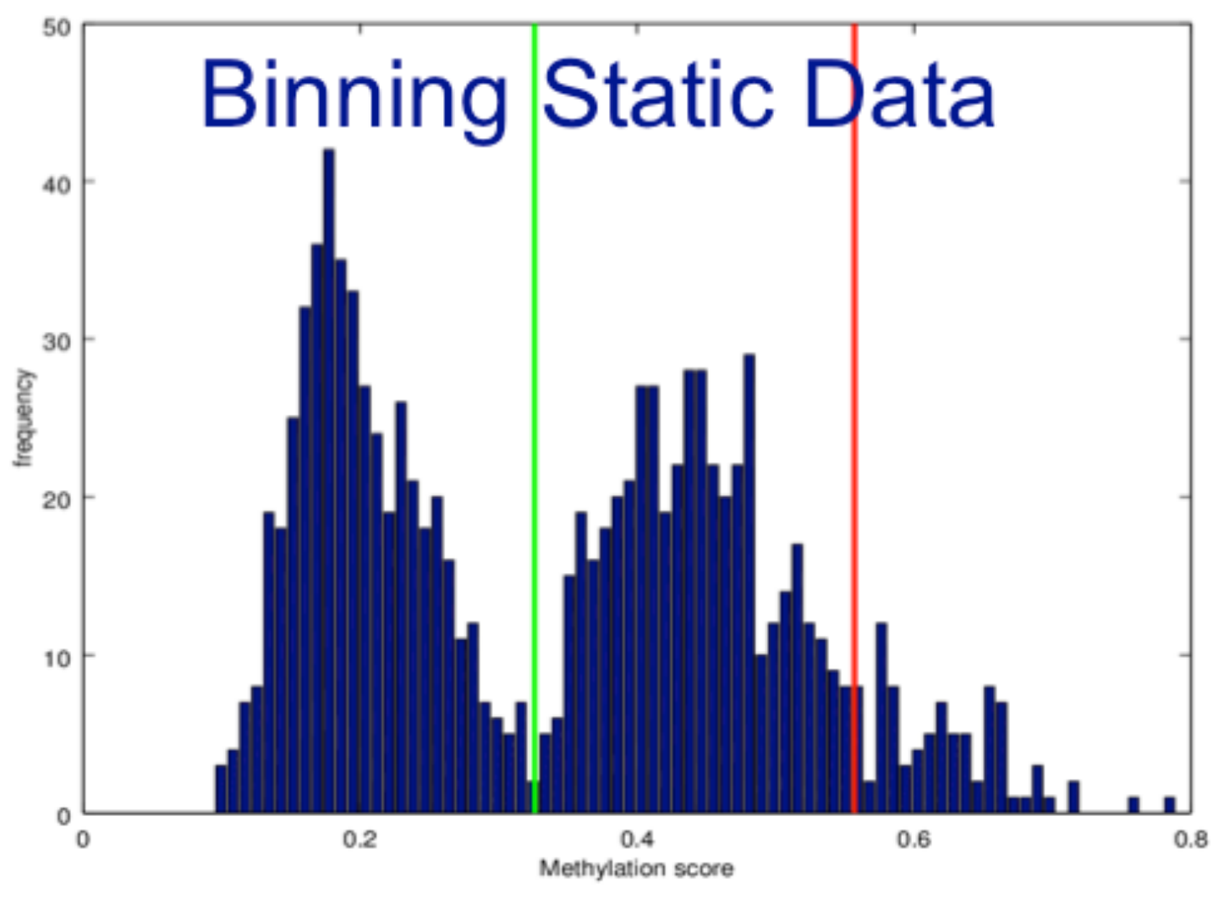
Sakhanenko and Galas , *J. Comp. Biol.* (2015)

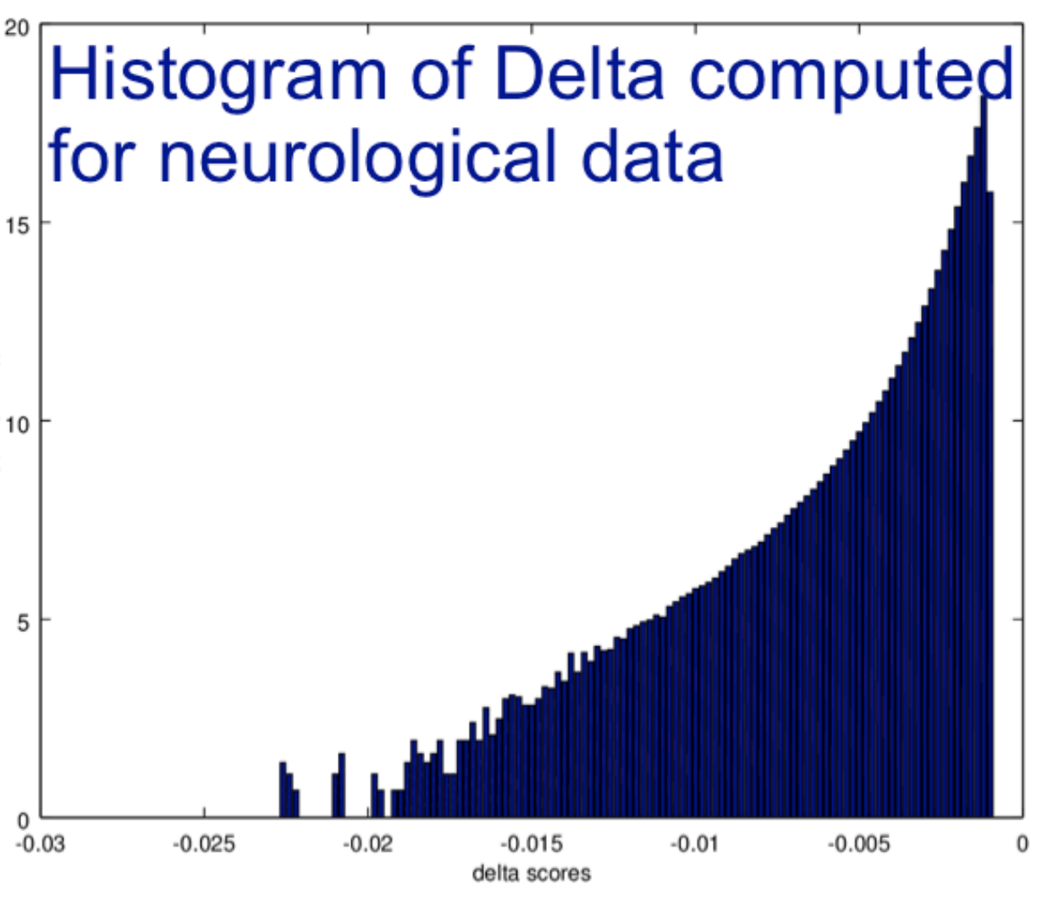

Simulated Data: Dependence detection

Statistical correlation measures (*1000 data points*)

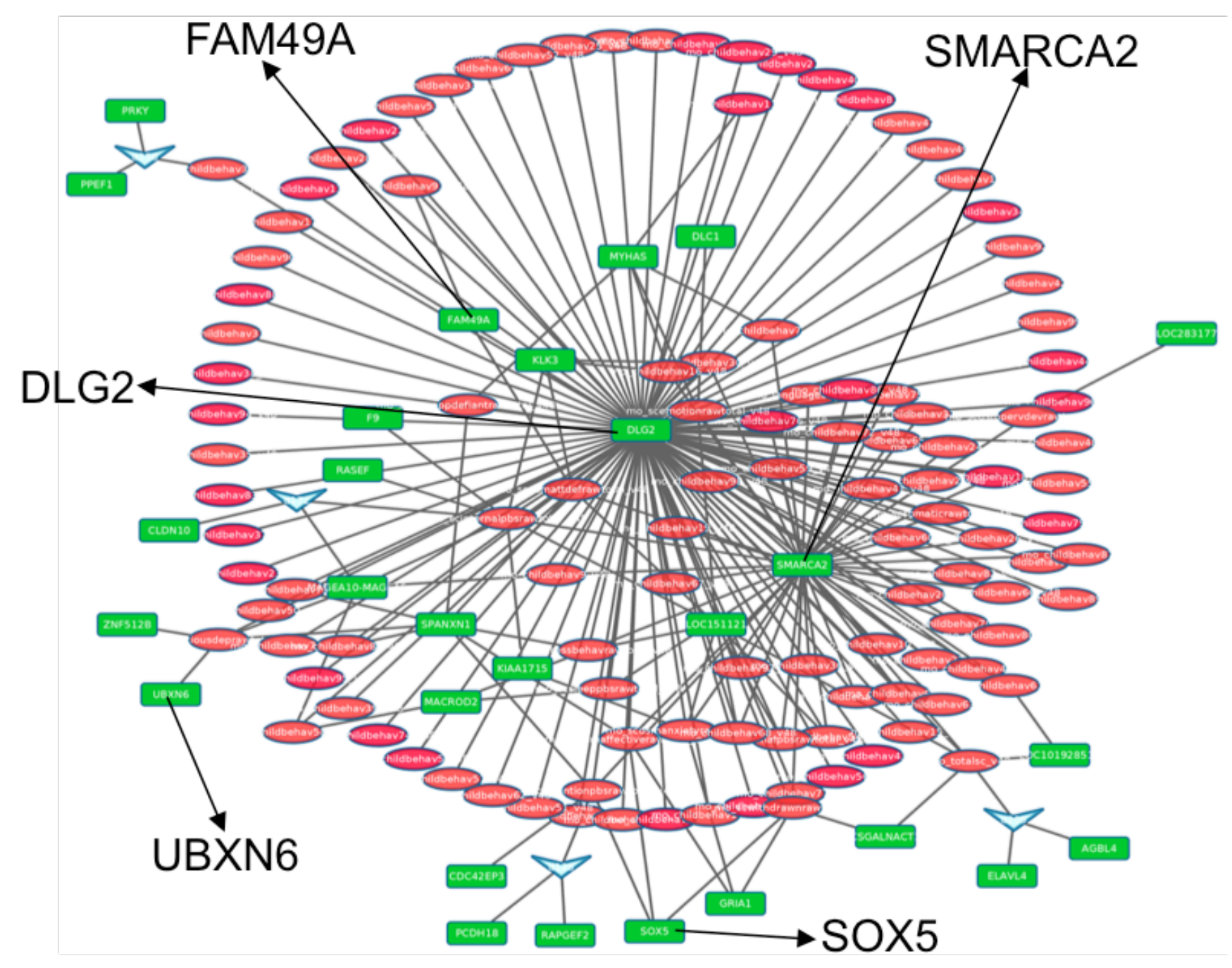
Symmetric delta for three variables


Binning After Parameterizing Weight


Binning Static Data


Histogram of Delta computed for neurological data

Example: 3-variable dependence (2 SNPs and 1 phenotype), illustrated by a hypergraph



Neuro data: top CBCL phenotypes & SNP connections
All point to genes with known neurological effects.

SNP1 (Gene1)	SNP2 (Gene2)	Phenotype	Description of the phenotype
Rs896996 **(DLG2)**	Rs1022308 **(DLG2)**	mo_totalsc_v48	Total score summary
Rs6475635 **(SMARCA2)**	Rs1555646 **(SMARCA2)**	mo_childbehav18_v48	Destroys things belonging to family
Rs751192 **(FAM49A)**	Rs888575 **(FAM49A)**	mo_scdsmattdefrawtotal_v48	Attention Deficit/ Hyperactivity total score
Rs741923 **(UBXN6)**	Rs4807584 (MPND)	mo_scanxiousdeprawtotal_v48	Anxious/Depressed: raw total score

Methods

- This method was model-free and insensitive to under sampling.
- The method was used to detect multivariable dependencies among variables.
- The measures were significantly nonzero only when the subset of variables had an essential, collective dependency [1].
- We used our approach to detect multivariable dependencies in childhood data in a large cohort of children with known genotype, environmental, and phenotype information (GUSTO data from Singapore).
- The approach was used to develop new hypotheses about causal relations.
- We calculated dependency values for variable sets of large degree; this enabled us to identify dependent subsets, but was limited by the combinatorial explosion:

Calculations ~ (Number of variables)^{degree}

- We used properties of the measure to avoid the combinatorial explosion by following the "shadows" that the multivariable dependency cast onto smaller subsets [2].
- The Shadows Algorithm enabled us to calculate measures of any degree of dependency.

Results

- We analyzed a large, high-dimensional data set about the development of Singapore children (GUSTO).
- The GUSTO study collected a diverse range of information about children to capture a full view of child development.
- We considered 3 categories of phenotypes – anthropometric, neurological, and asthma/eczema – and their dependence on genetic variation.
- We identified a small set of strong 2- and 3-variable collective dependencies among phenotypes and SNPs.
- These dependencies formed interconnected networks of variables and enabled us to seek biological relations in these dependencies and form new hypotheses.
- Our method returned a set of candidate multivariable dependencies, which were input to functional analysis.
- The SNP-phenotype dependencies and their networks suggested several involved biological pathways – essential for precise models.
- Genetics can help stratify populations to better detect environmental effect signals.

Conclusions

A. The method works.

- **The application of our method to the Singapore data (GUSTO) showed promising initial results.**
- **We identified complex dependencies in very large and heterogeneous data sets.**
- **We are adding other types of data to the analysis and integrating them into a more unified network.**

B. Preprocessing of data is extremely important.

- **Missing data and other noise can strongly affect our ability to detect dependencies.**
- **Binning variable quantities also is key.**
- **These issues are being actively investigated.**

References

1. Galas DJ, Sakhanenko NA, Skupin A, Ignac T. Describing the complexity of systems: multivariable "set complexity" and the information basis of systems biology. *J Comput Biol*. 2014;21(2):118-140.
2. Sakhanenko NA, Galas DJ. Biological data analysis as an information theory problem: multivariable dependence measures and the shadows algorithm. *J Comput Biol*. 2015;22(11):1005-1024.

Sponsored by the Healthy Birth, Growth, and Development (HBGD) initiative representing Discovery & Translational Sciences, Integrated Development, Integrated Delivery, Nutrition, Agriculture, Water, Sanitation & Hygiene, Maternal, Newborn & Child Health, Enteric & Diarrheal Diseases, and Pneumonia Program Strategy Teams. HBGD*ki* was conceived of and is directed by shasha.jumbe@gatesfoundation.org.

Acknowledgments

We thank the GUSTO project staff and Sir Peter Gluckman for access to the data and their collaboration, and members of the Galas Laboratory for their contributions. This work was supported by the Bill & Melinda Gates Foundation HBGD initiative.