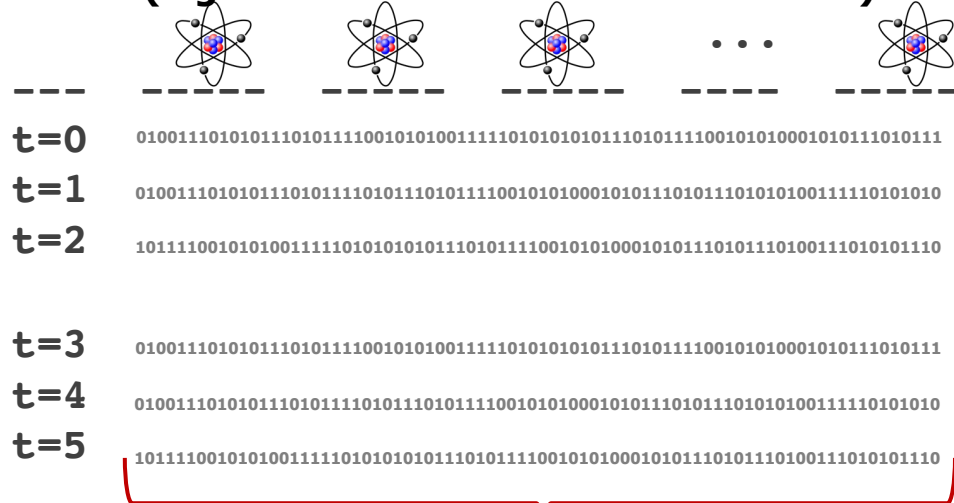# Discovering Informative Factors in Gene Expression

Greg Ver Steeg

AI4Health

Dec. 1, 2022
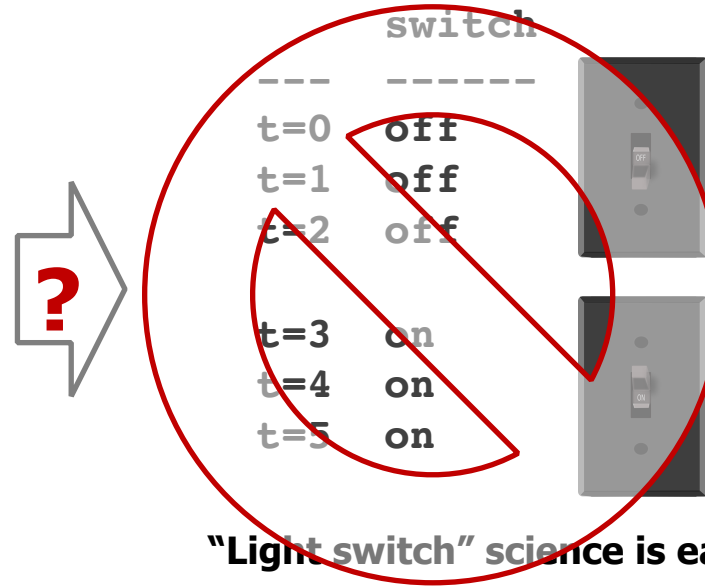
The Light Switch Problem

Micro variable data
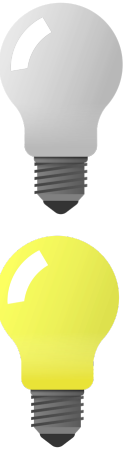**(e.g. about all the atoms in the switch)**

Macro variable description

Outcome we care about

"curse of dimensionality" with more variables

"Light switch" science is easy!

Problem: Hard to identify "macro" switch states from "micro" data

Micro variable data

Macro variable description

Outcome we care about

Human behavior

Gene expression

Macro state of individual

Phenotypes

Response to treatment

Factor discovery applications and impact

Human Behavior

(Extroversion)
Z_1

(Neuroticism)
Z_2

Z_1 ← Life of the party?
Z_1 ← Talk a lot?

Z_2 ← Easily stressed?
Z_2 ← Mood swings?

Gene expression

(Immune activation)

(Metastasis)

Z_1

Z_2

Brain imaging

$Z_3 \ldots$

Z_1

Z_2

Superior

Right

Anterior

Left

Posterior
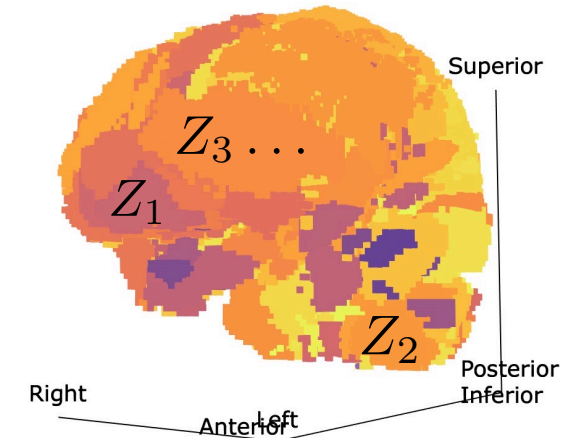Inferior
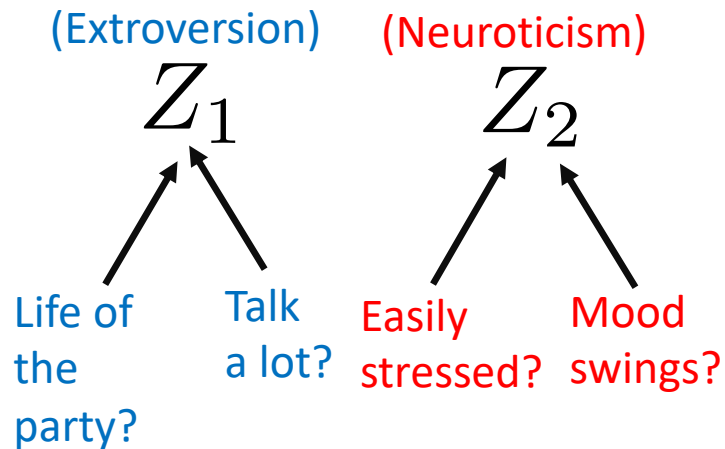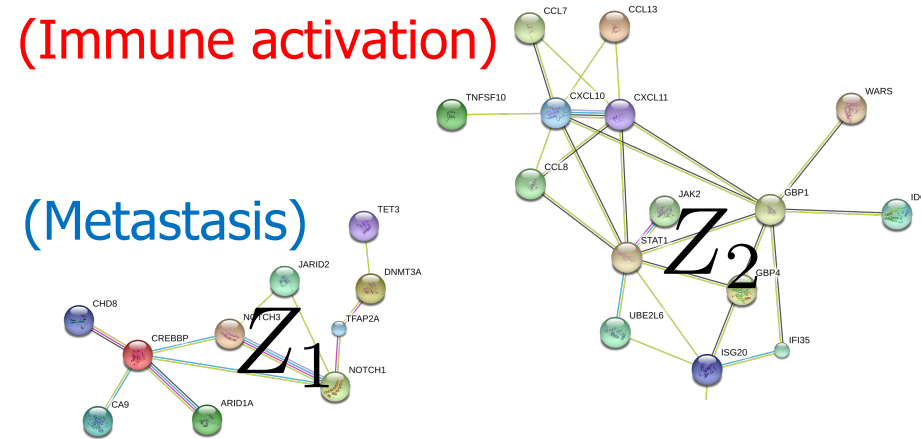
Used by *Pew Research* in "**Where Americans Find Meaning**" study to discover latent factors from open-ended responses.
https://bit.ly/2UKWcad

Discovered factors with **4x gene ontology enrichment** leading to **new, targeted treatment options** for ovarian cancer
http://corex.isi.edu
http://wapo.st/2sabpAg

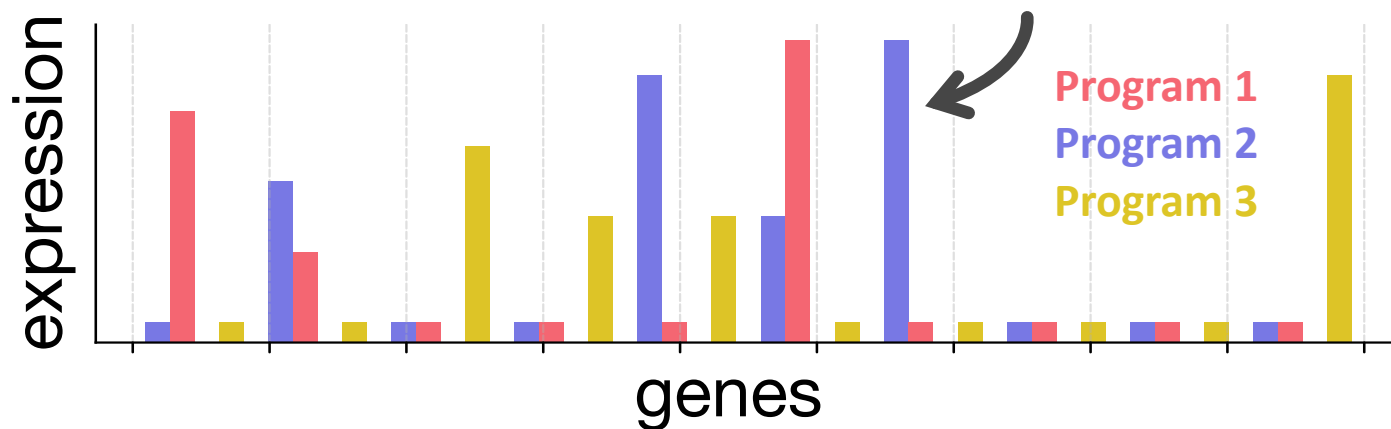**Blessing of dimensionality**
New biomarkers for disease progression
(Ver Steeg et al, NeurIPS 2019)

"Gene expression programs" via "q-diffusion"

A "program" or "meta-gene" is a large set of genes that change in coordinated ways, as part of some biological process

Program 1
Program 2
Program 3

expression
genes

"q-diffusion" , a similarity measure for high-dimensional data:
- **non-Euclidean**
- **Multi-level interactions**
- **Power law decay**

pairwise interactions

gene
activities
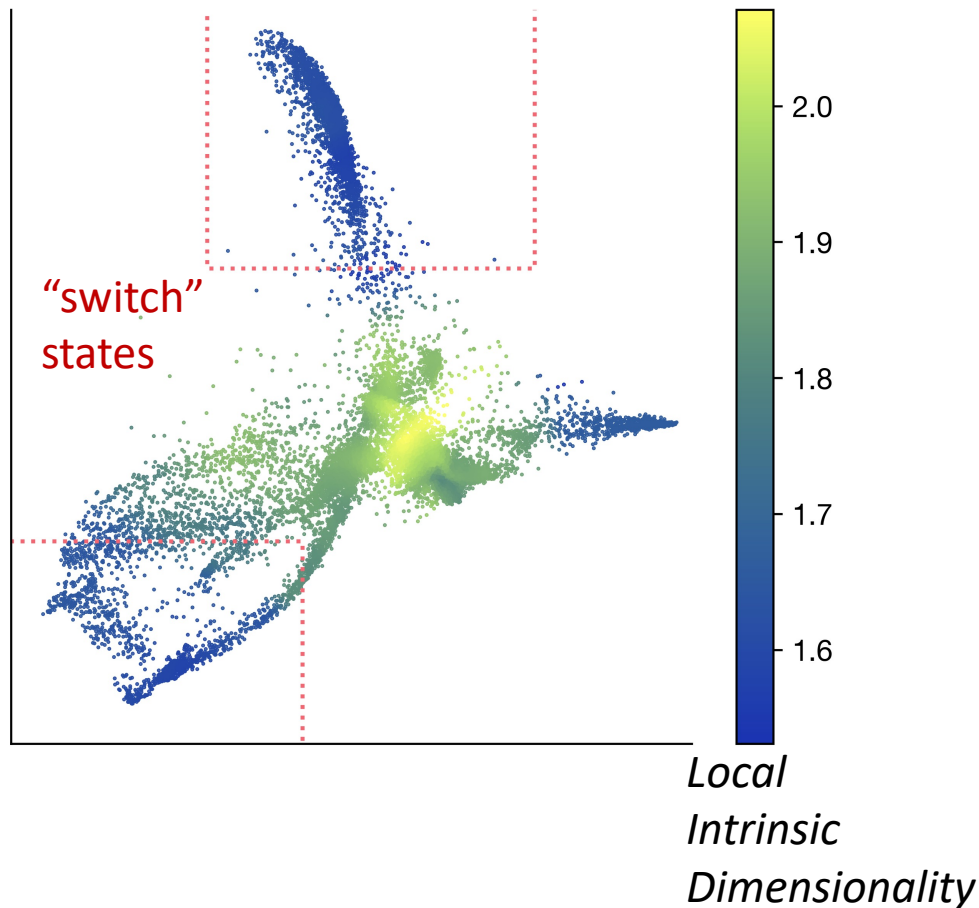
higher-order interactions
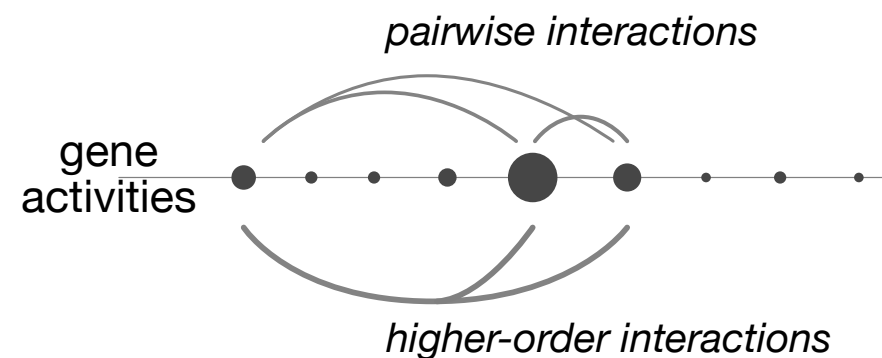
How should we identify gene expression programs?
Don't say single gene association studies

q-diffusive embedding reveals "light switches" in colorectal cancer study

Patient embeddings

"switch" states

Local Intrinsic Dimensionality

- Single cell RNAseq study for colorectal cancer
  - Few samples, high-d
  - Single cell data more targeted, but inherently noisy

- Unlike embeddings from PCA, ICA, NMF..., q-diffusive embeddings reveal characteristic branching structures
- "Programs" exhibit **2-10x gene ontology enrichment** compared to other methods
- **Branches reveal differential outcomes for treatments:** *bevacizumab, cetuximab*

Conclusion: discover macro "switch" states from microscopic data

Thanks to collaborators:

Heinz-Josef Lenz (Keck), Joshua Millstein (Keck), and group, working on colorectal cancer

Paul Thompson (Keck) group, neuroimaging

Jose-Luis Ambite's (ISI) group: bioinformatics, federated learning

Myrl Marmarelis (ISI), PhD student in my group

Thanks!
Contact: gversteeg@gmail.com