AYASDI

Discover What You Don’t Know

Pek Lum, Ph.D.
Chief Data Scientist & VP of Solutions

Complex
Big Data is Nothing New to Pharma
Pharma R&D live and die by data

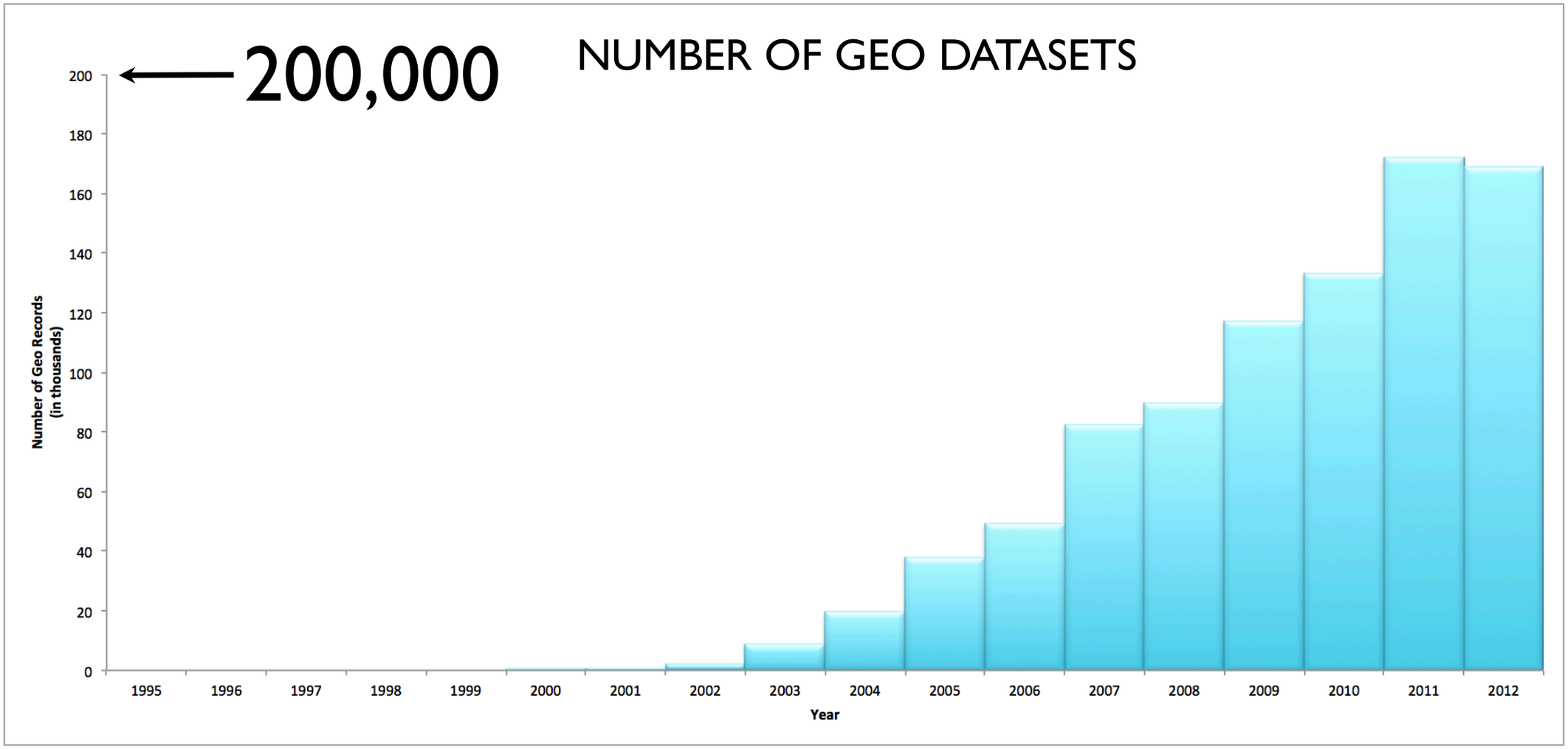
~~Big~~ Complex Data is Nothing New to Pharma

Pharma R&D live and die by data

One of the greatest data collection to date: Gene Expression Omnibus

Start Date: 1995

Amount of Data: LOTS



AYASDI

Discover what you don't know.

The Cancer Genome Atlas

Start Date: 2006

Amount of Data: 150TB

Cost to date: over $500M

"Forty Years' War: Advances Elusive in the Drive to Cure Cancer."

The New York Times

Finding new ways of unlocking big data is KEY

PHARMA IS GOOD AT:

data storage

data management

even more data generation

PHARMA MUST START THINKING ABOUT:

radical new ways to extract insight

data exploration

automation

computer-augmented analysis

BIG DATA is all about:

Extracting insights

Exploration

Letting the data show you the questions you never thought of asking

Automation

Introducing "The Shape of Data"

A mathematical concept that began in the 1700's.

Uses the shape of data to find unknown phenomena.

Topological Data Analysis (TDA)

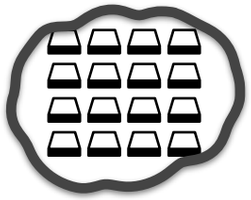
Math

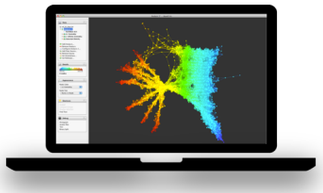
+ Computer Science

+ User Experience

$$w(x,y) = \exp\left(\frac{-d^2(x,y)}{\epsilon}\right)$$

Automated discovery of shapes





"TDA methods will transform the way that doctors triage patients, through construction of non-linear, non-invasive medical statistics to assess patients in intensive and critical care situations."

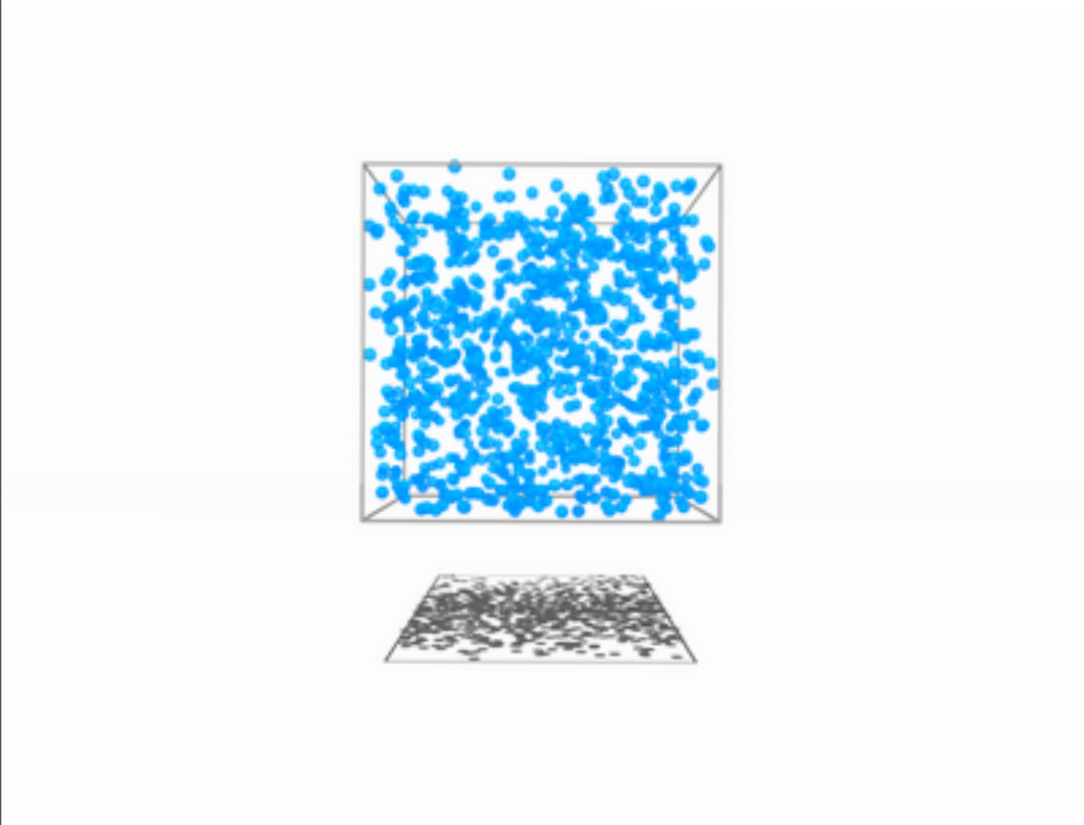
2009 DARPA Strategic Plan

AYASDI

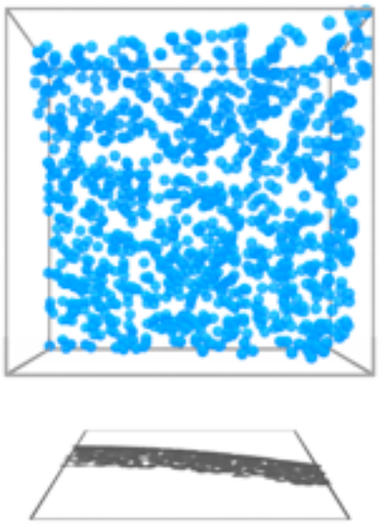
Discover what you don't know.

Data has shape and shape **has meaning.**

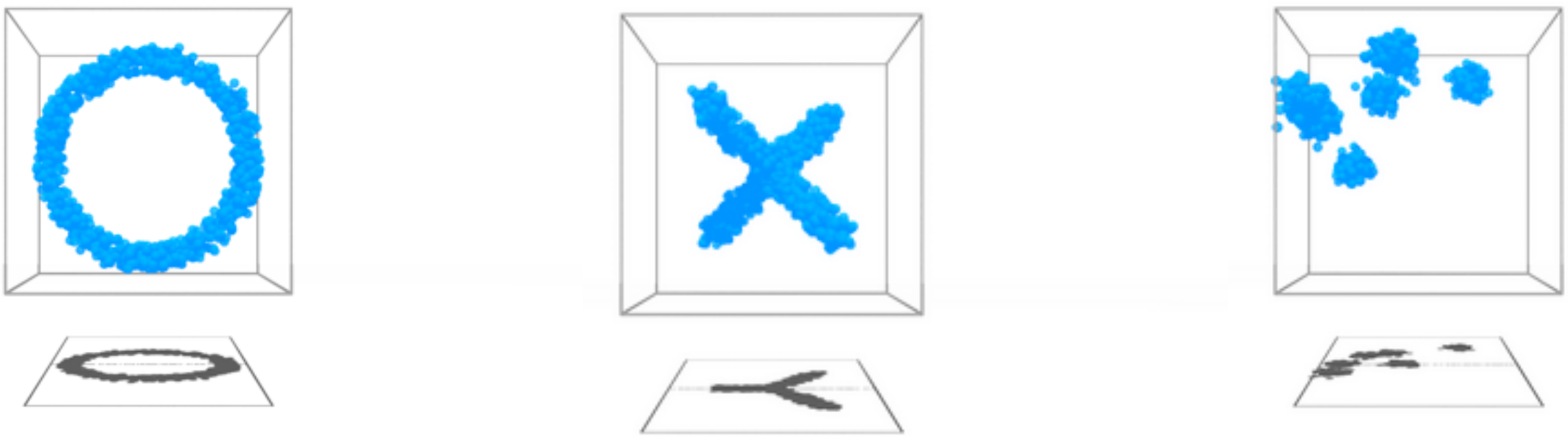
What do I mean by data having shape ?



Age, Weight and Height sampled uniformly at ***random***



In reality, age, weight and height are correlated and that ***data has a shape***

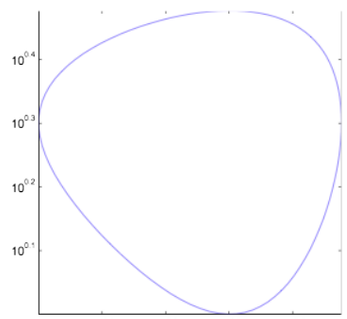


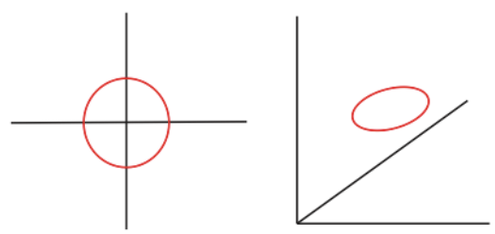
AYASDI

Discover what you don't know.

3 Key Properties of TDA

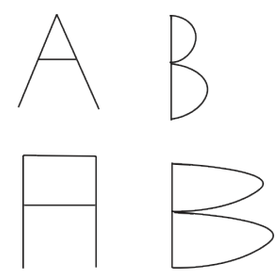
Coordinate Freeness





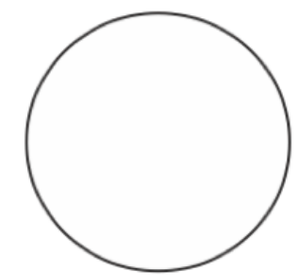
COMBINE DISPARATE DATA SOURCES

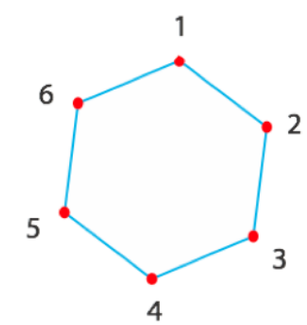
Deformation Invariance



NOISE & NULL TOLERANT

Compressed Representation





EXPOSES SIGNAL

Ayasdi's approach using Topological Data Analysis (TDA) is one of the top 10 innovations developed at DARPA in the last decade.

Tony Tether, Director
Defense Advanced Research Projects Agency (2001-2009)

Ayasdi Iris Basics

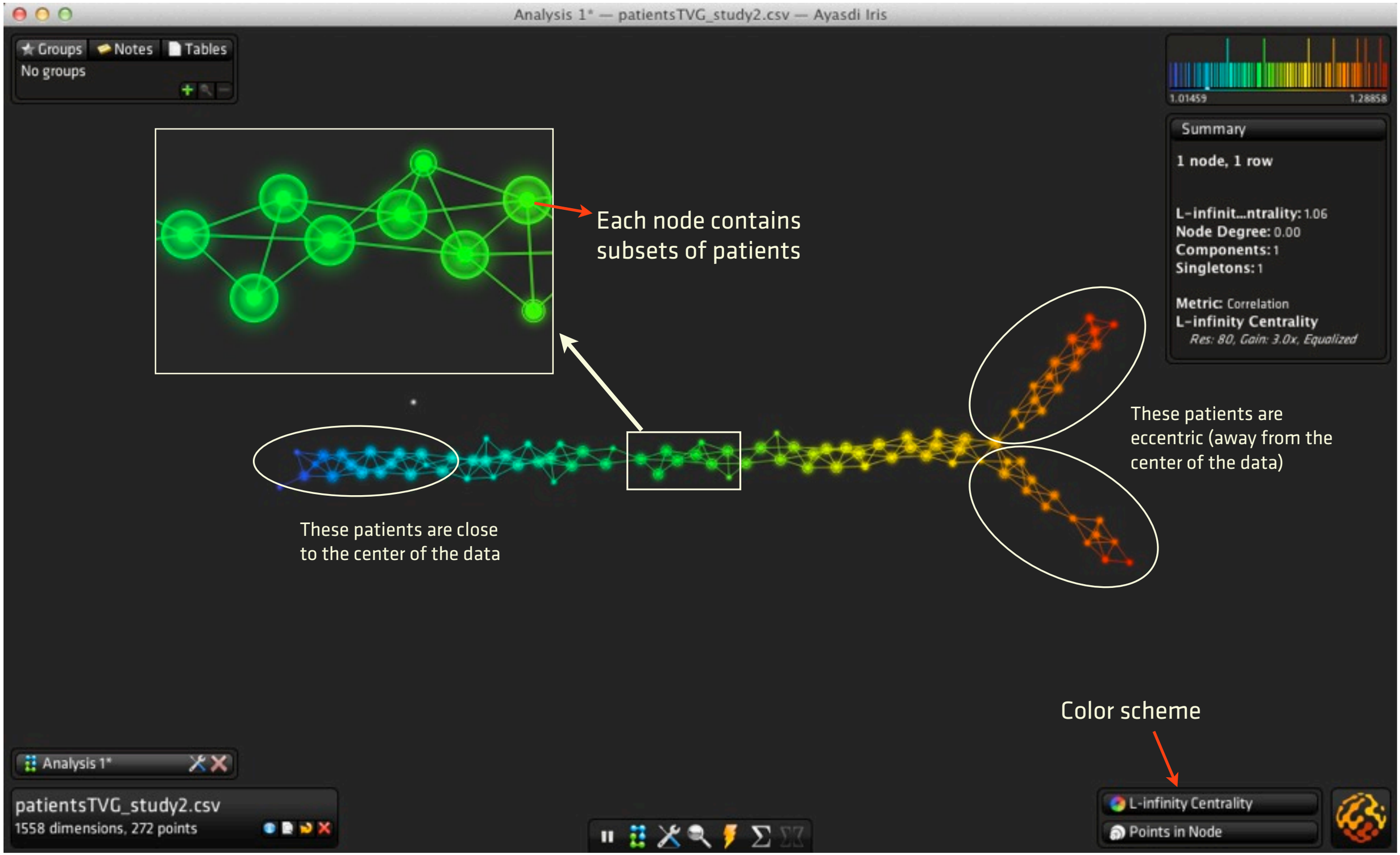
1. Nodes are groups of similar data points

   Size scales with number of data points

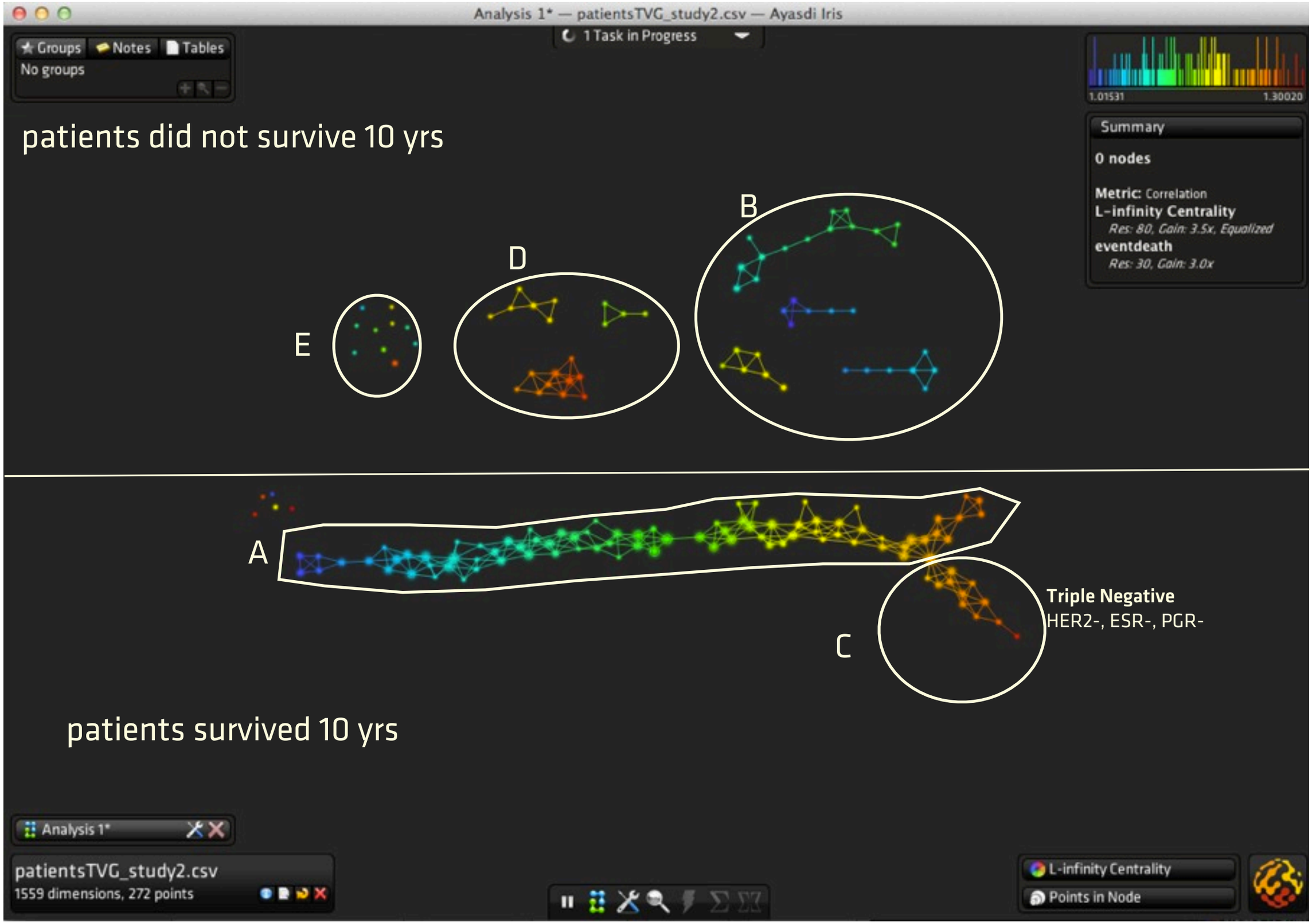
   Color schemes highlight properties

2. Nodes are connected when they have data points in common

3. Network shapes automatically reveal patterns within your data

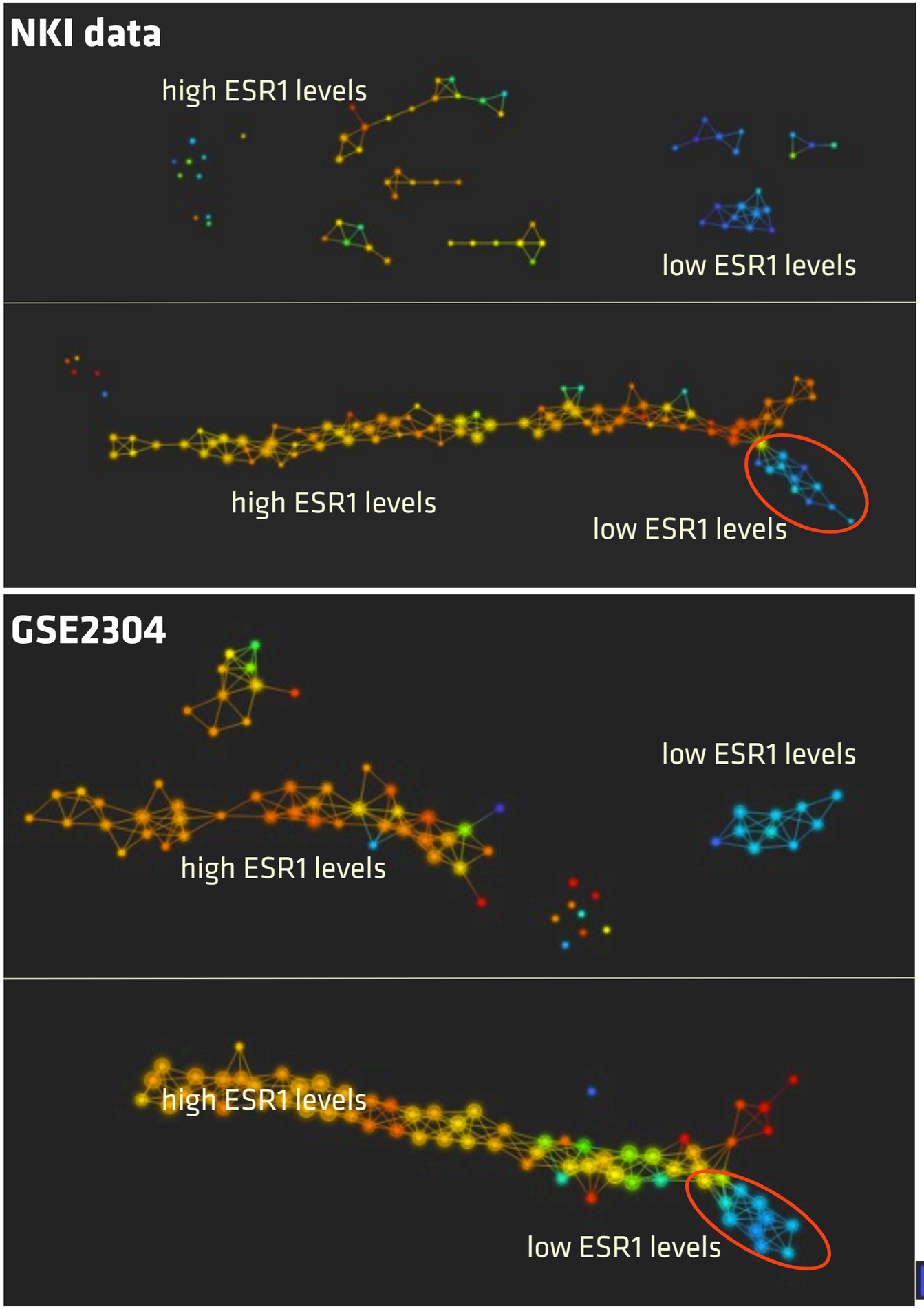


Topological Map of Patient-Patient Relationships according to their tumors' molecular characteristics (in this case, gene expression)



AYASDI

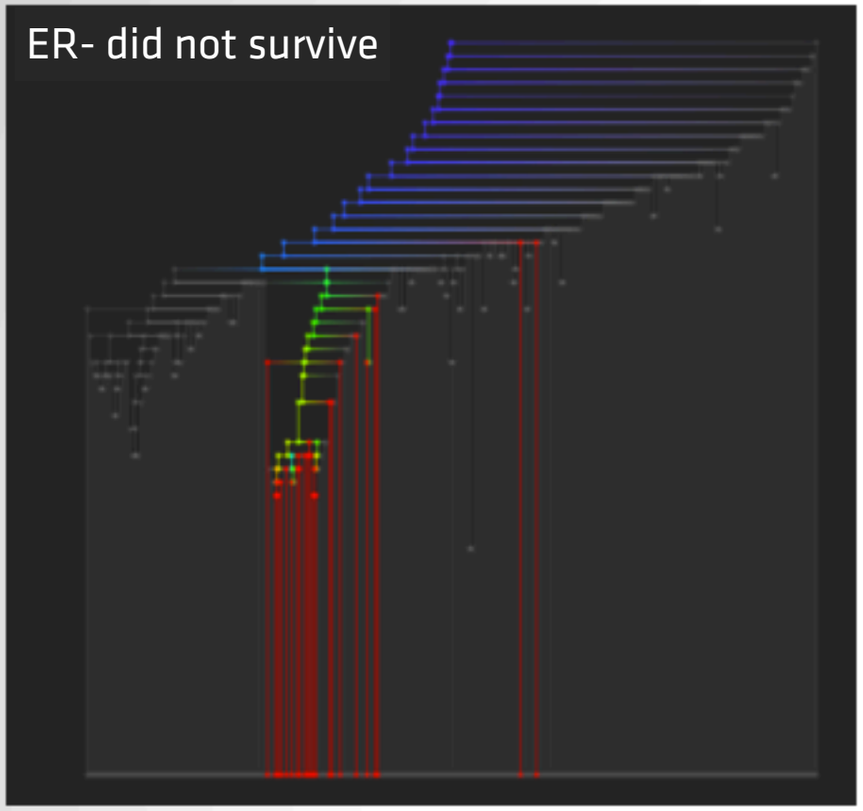
Discover what you don't know.

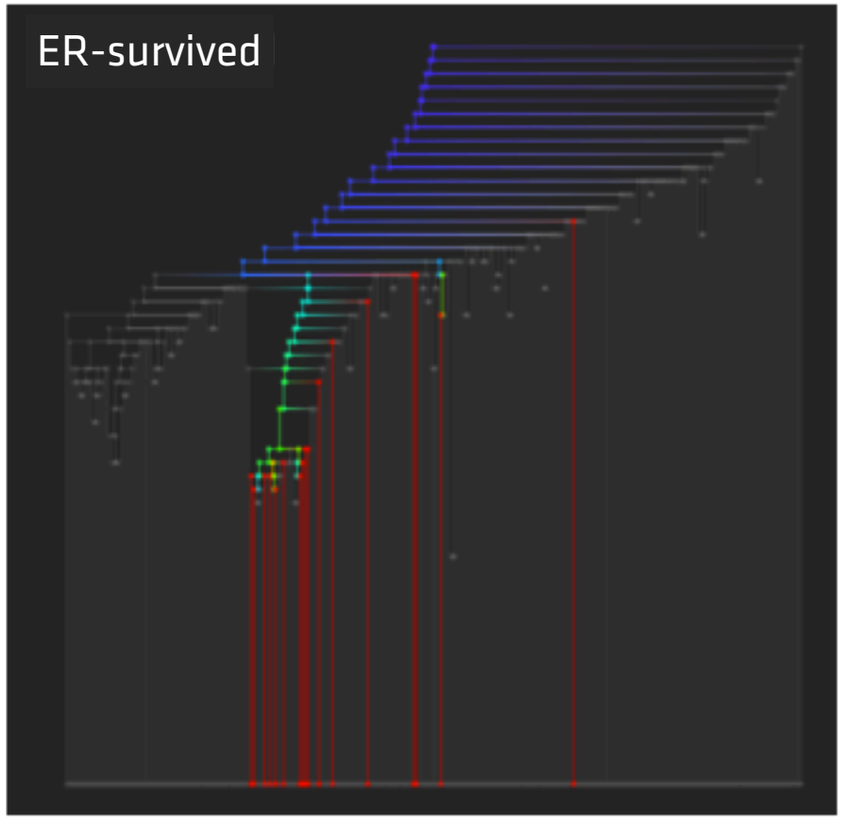


Topological maps from two independent cancer data sets are visually and molecularly very similar

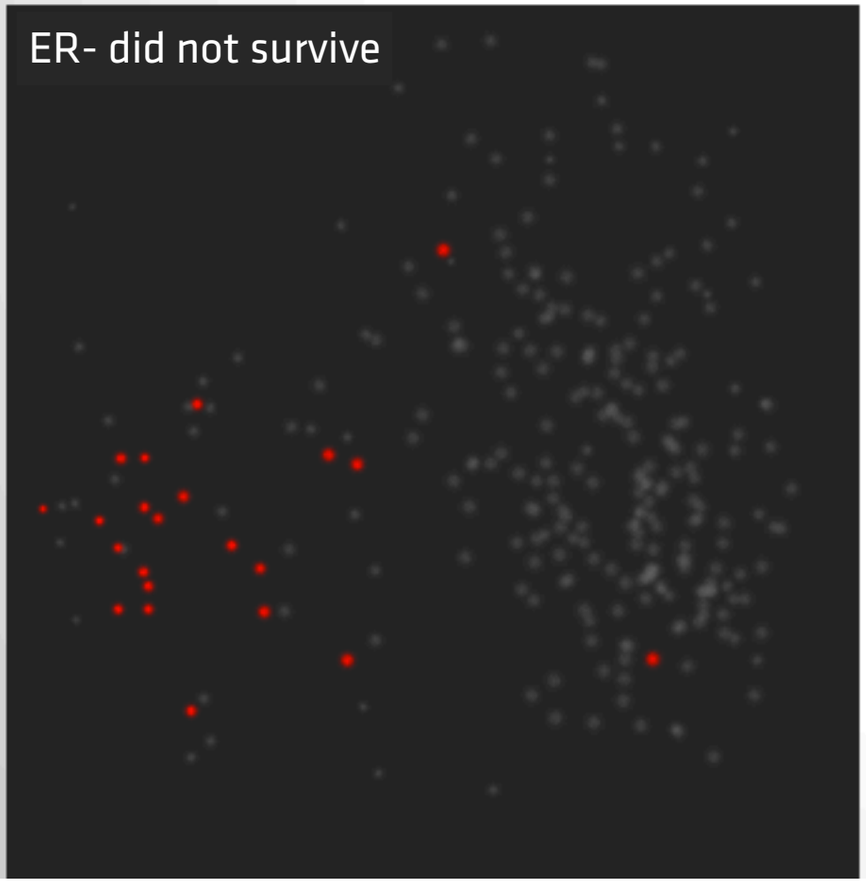
Difficult to identify the subtypes by conventional methods

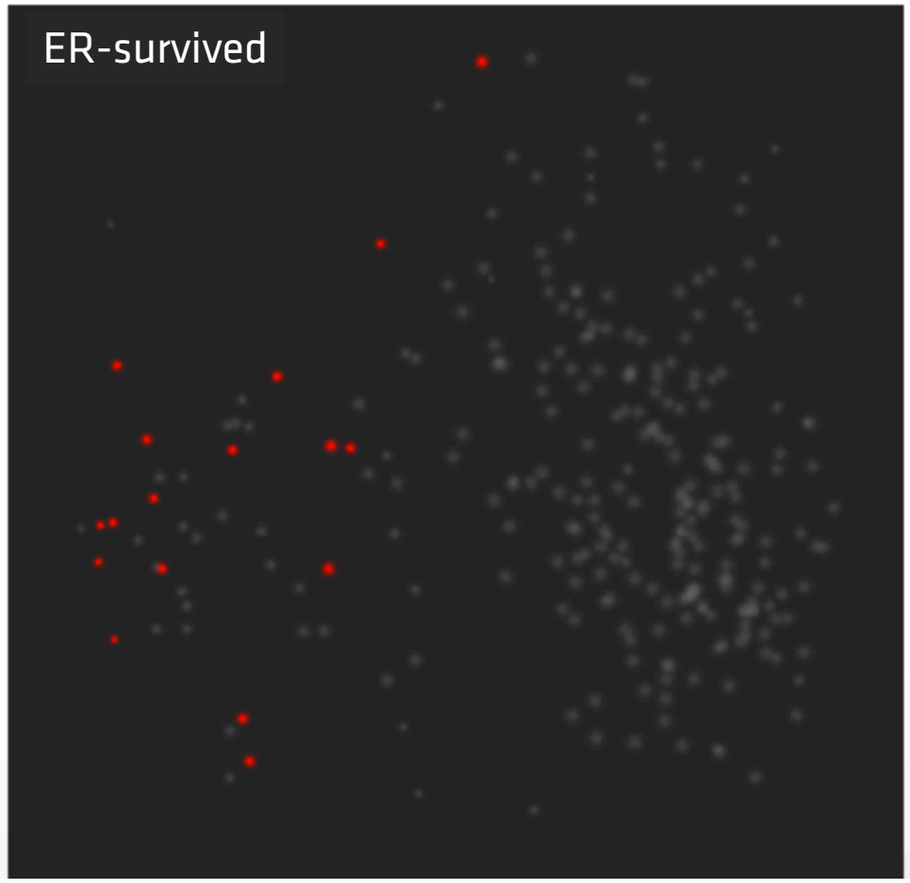
Clustering





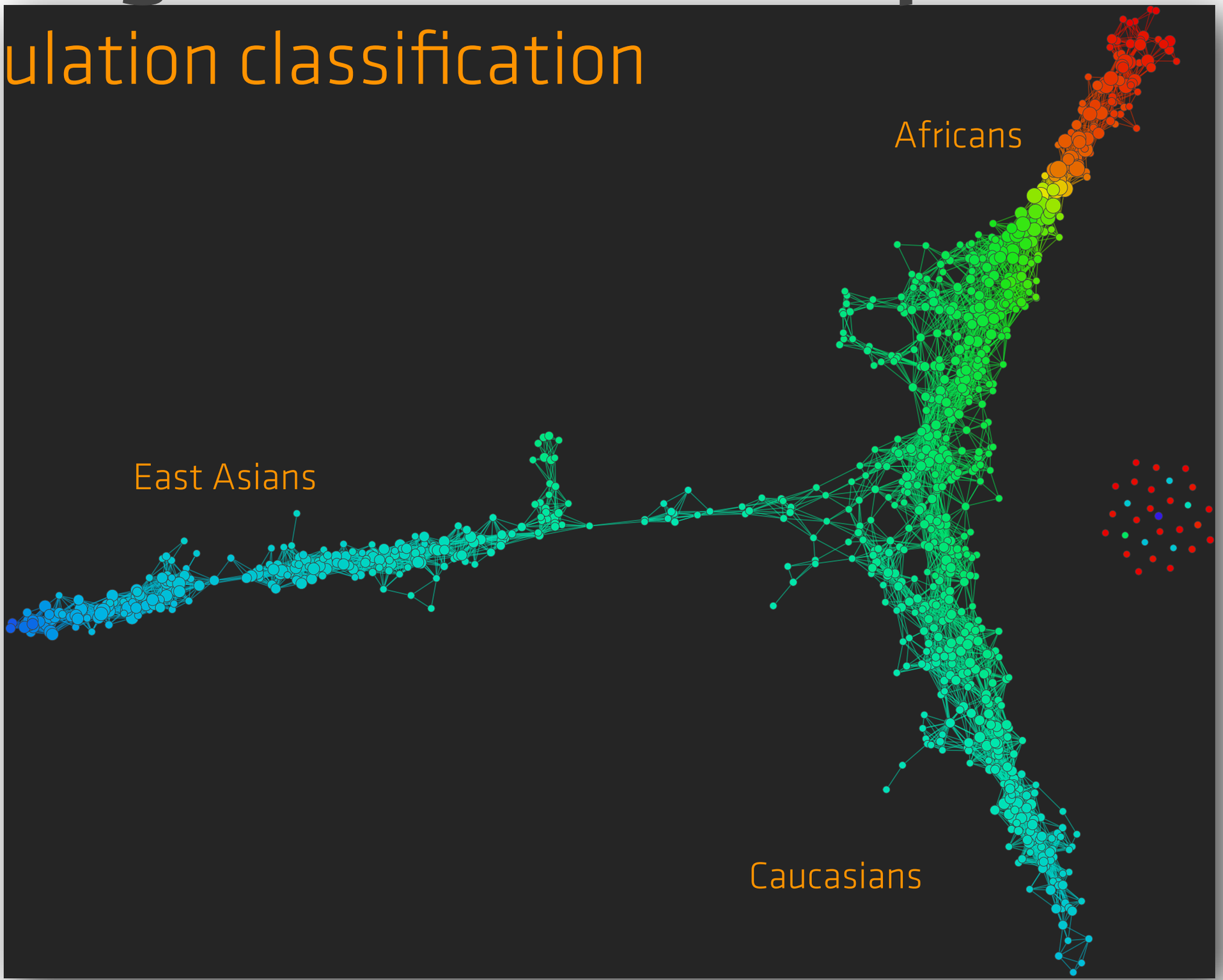
PCA





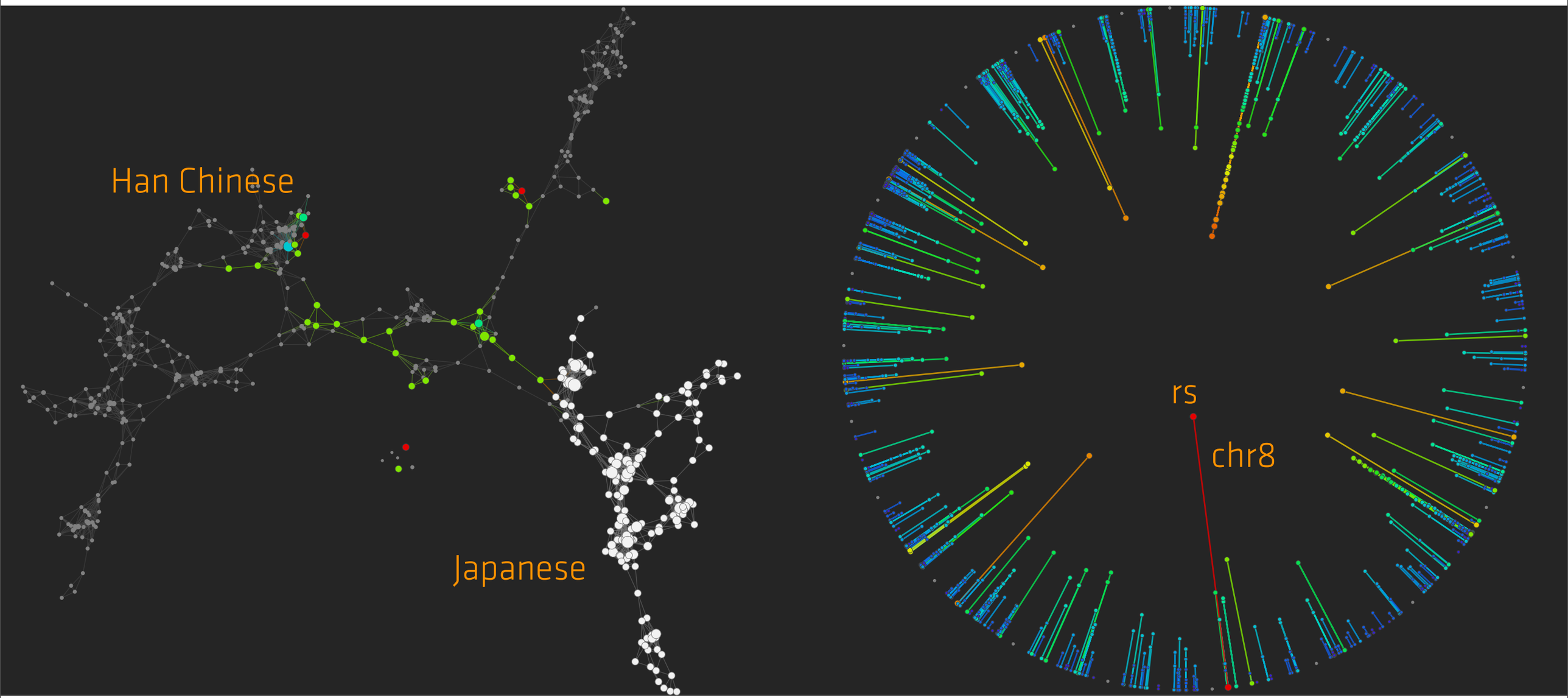
1000 genomes DNA-seq data

population classification

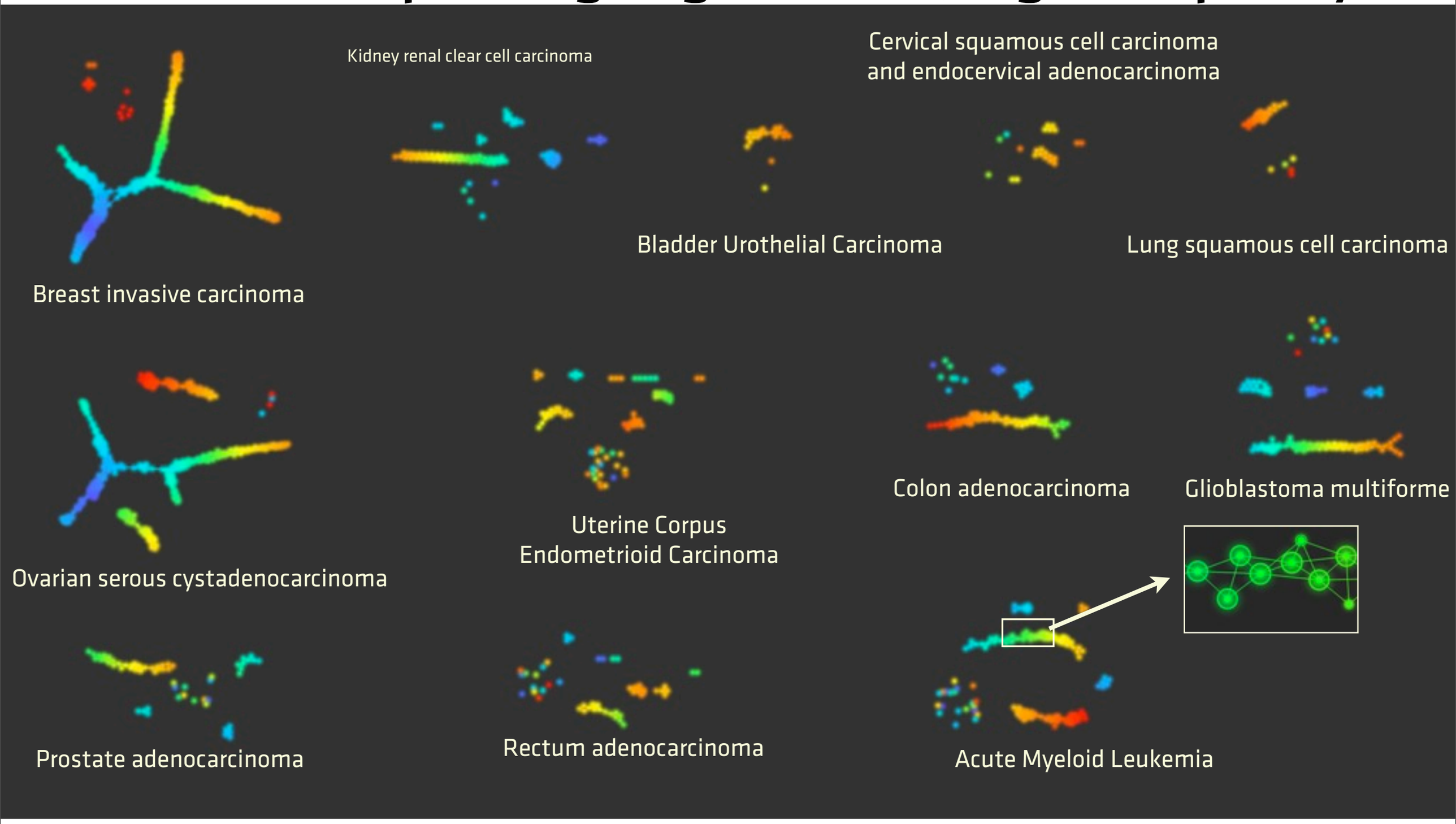


Identification and visualization of significant variants

Han Chinese and Japanese can be easily distinguished and visualized with TDA and Iris

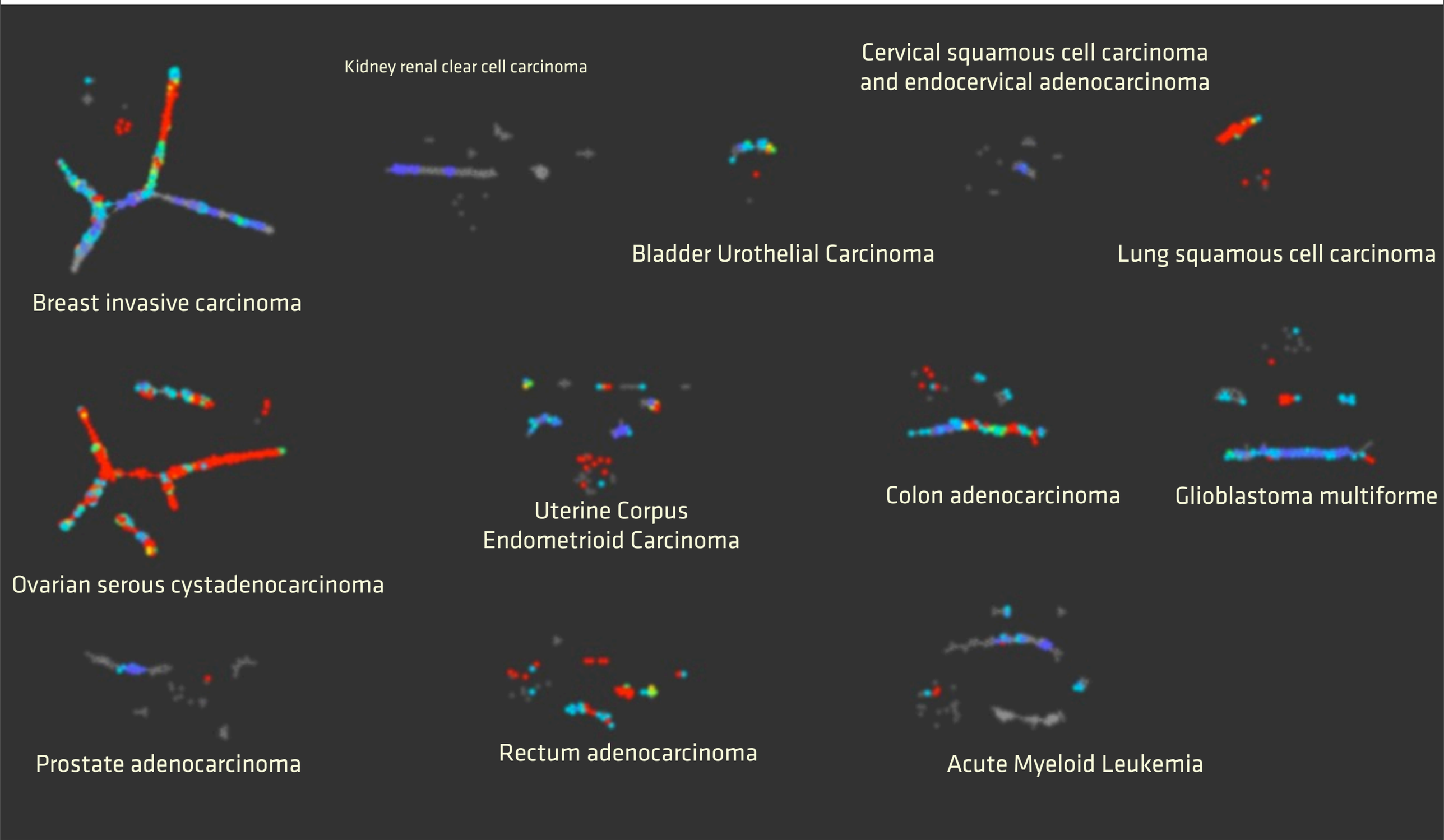


12 cancer types from The Cancer Genome Atlas
DNA exome sequencing: High Volume, High Complexity



Over 2400 tumors, 12 cancer types, over half a million unique variants analyzed simultaneously

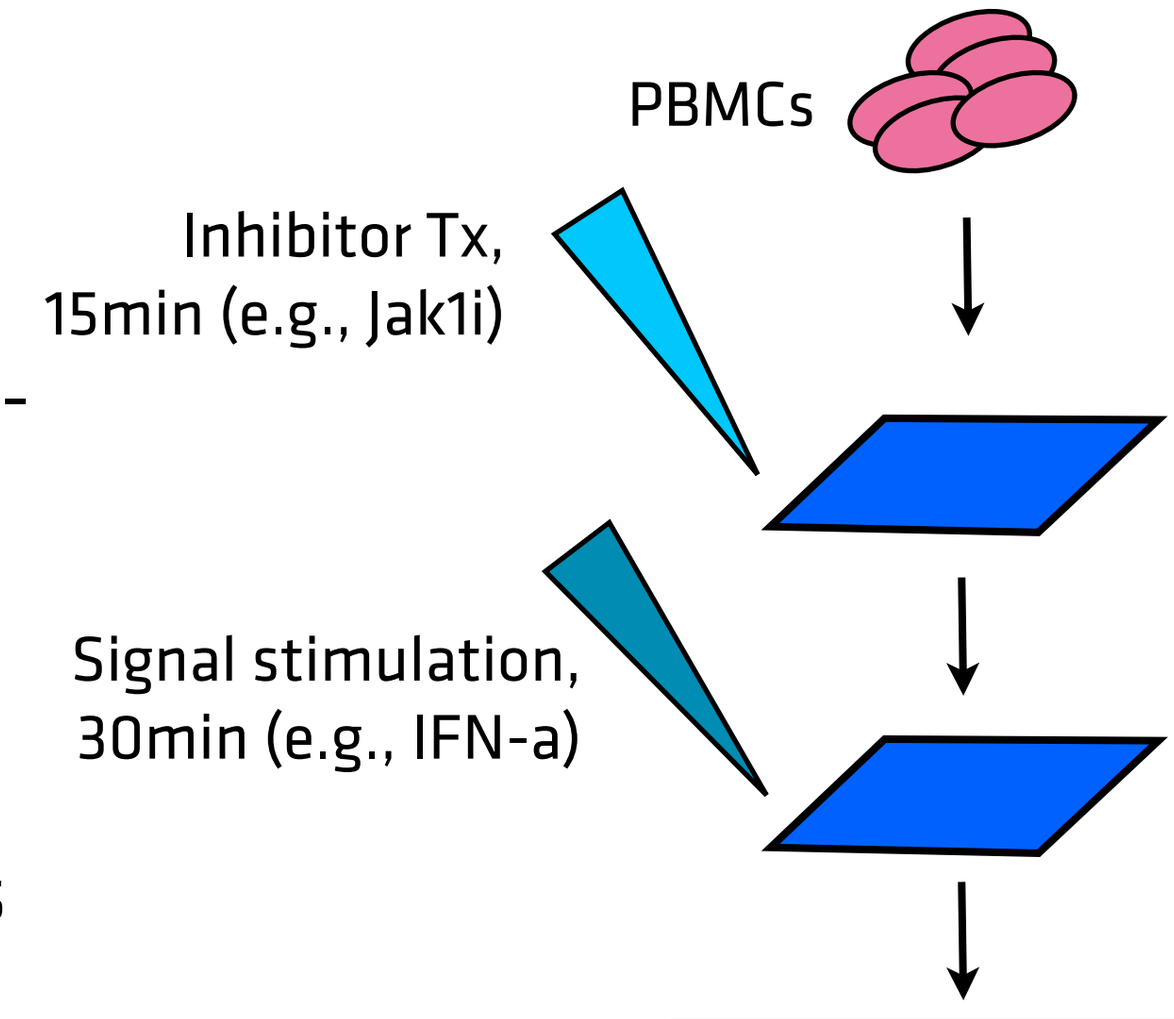
Landscape of p53 mutations across all 12 cancers



red= enriched for p53 mutations; blue=not enriched

Mass cytometry data under analysis

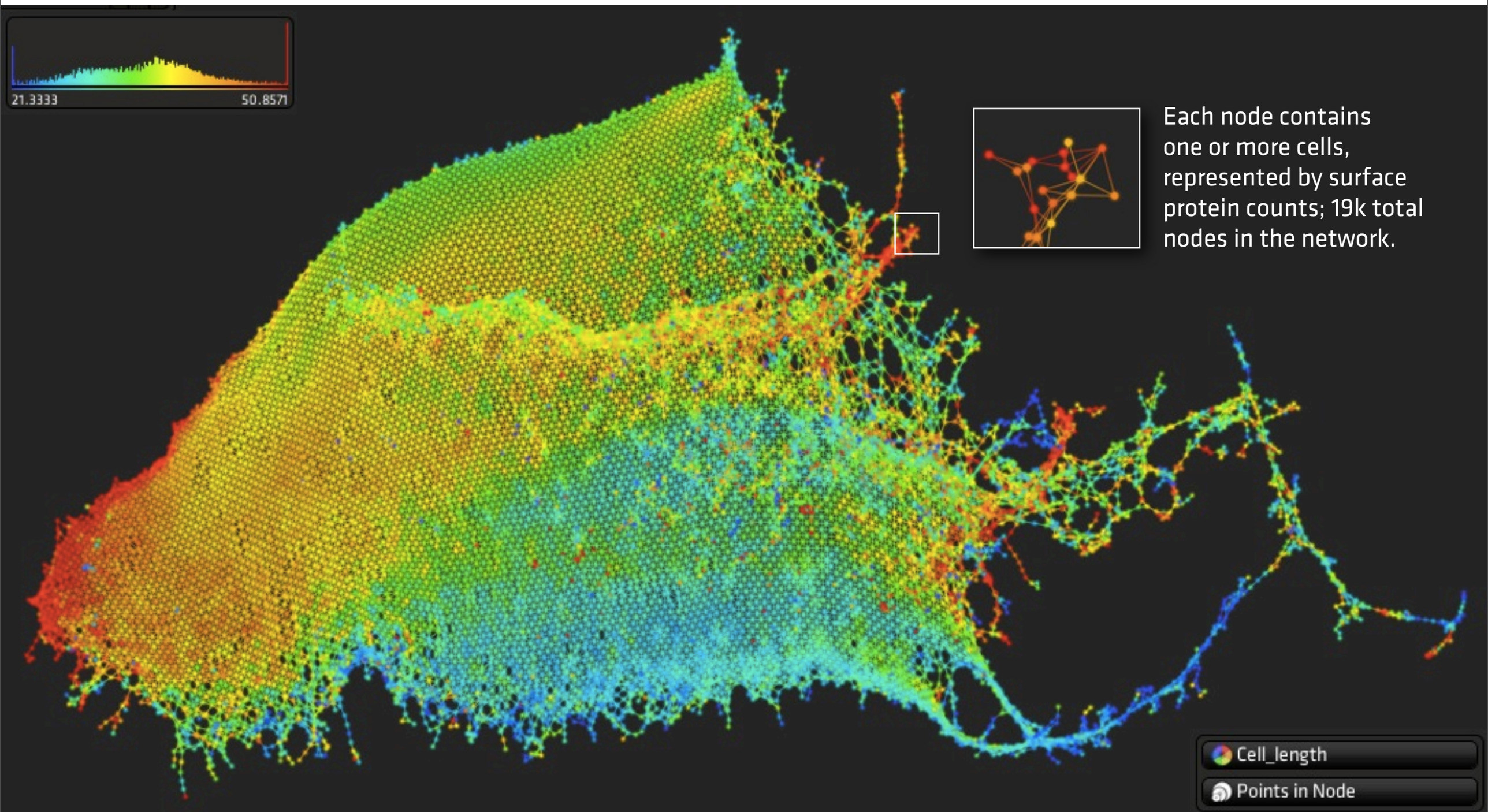
- Data powered by Cytobank (Nolan Lab at Stanford for mass-tag cellular barcoding (Nat. Biotech. 2012))
- 27 inhibitors at 8 concentrations over 14 stimulations
- 14 cell types identified via gating procedures



Processing pipe from FCS to comma-delimited format for topological analysis

pAkt(Sm152)	pBtk(Er166)(	pErk(Er168)(	pLat(Er170)(	pNFkB(Nd14	pPlcg2(Er167	p56(Yb172)D
-0.2471466	1.44065857	0.0868454	3.41189575	-0.6493683	1.84163666	0.09715271
2.16613007	-0.2298965	-0.2900467	-0.2257767	-0.8921585	0.26354218	-0.1501007
1.19752502	-0.7157669	-0.3414612	1.14325714	14.1423874	14.3337173	1.96299744
3.42930603	1.92115021	-0.2364349	-0.8923416	3.87947845	3.65786743	-0.5170441
-0.6203766	-0.7472	0.25080872	2.81351471	-0.1446533	-0.6800995	0.24423981
2.48509216	898.824463	19.489296	-0.7567444	24.6196442	16565.1973	-0.0169678
-0.3382874	-0.6339111	3.02802277	1.56978607	1.21856689	14.9641418	-0.6439972
0.23661041	0.14953613	4.96395111	-0.9538727	2.07623291	43.3572083	1.96287537
-0.8797989	0.43587494	-0.1377487	-0.0057983	-0.1749802	1.53496552	-0.9740372
1.68782043	-0.3572464	1.75392914	-0.3883667	-0.156517	-0.3858109	-0.5853806
-0.8083954	-0.4314651	-0.2190857	0.8339386	-0.9462814	-0.7055206	-0.0355453
6.83000183	1.55770874	6.79499054	-0.5504532	2.08599854	-0.9033508	0.29000092
-0.913353	-0.1796722	0.15005493	-0.7339935	10.0560684	-0.6094437	-0.7047195

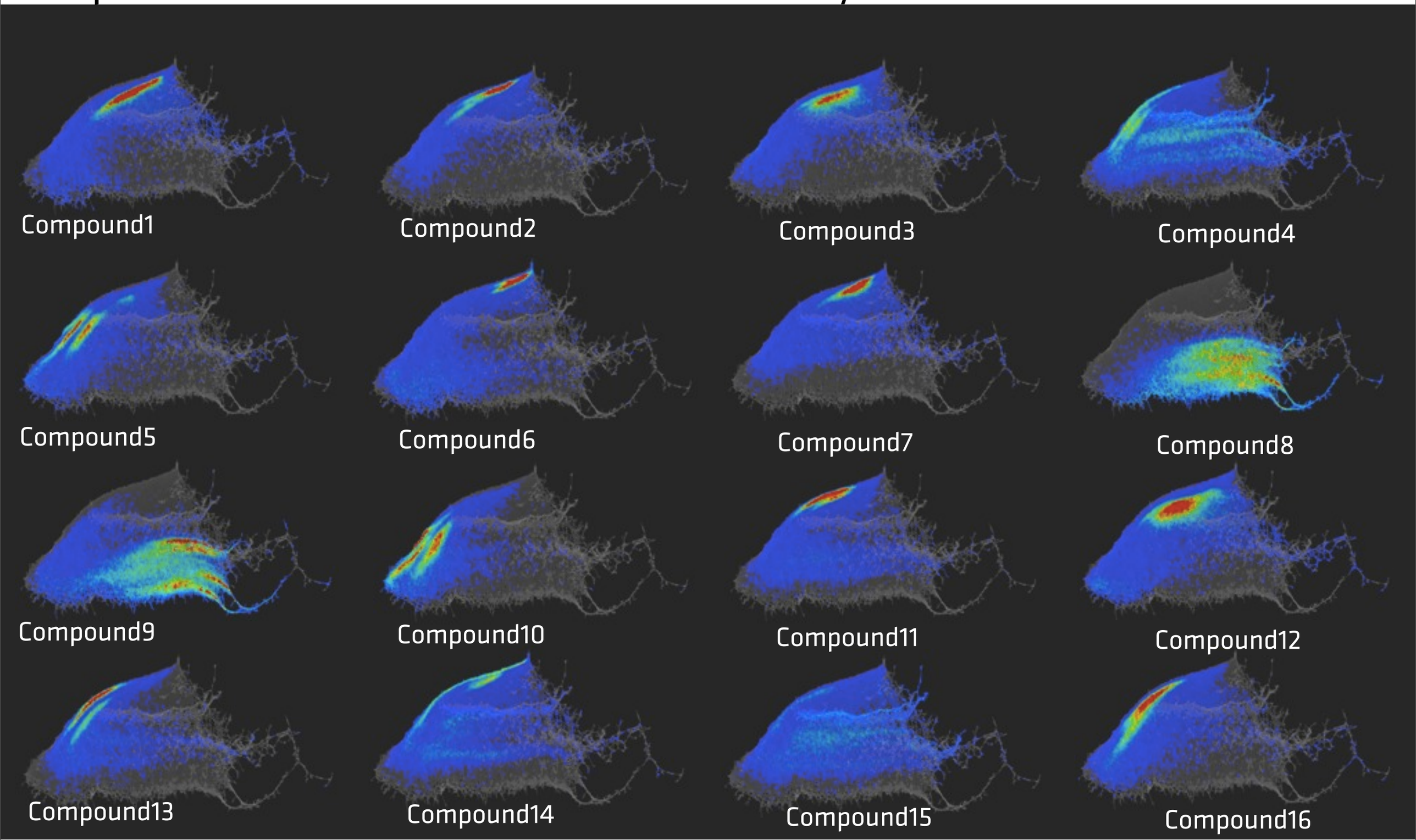
A topological network of 27 inhibitors comprised of over 2 million single cell Cytof outputs



AYASDI

Discover what you don't know.

Regions of the network are strongly enriched for different inhibitors provide a broad view of inhibitor similarity in reference stimulation



AYASDI

Discover what you don't know.

Questions? pek@ayasdi.com

AYASDI

Discover What You Don't Know

www.ayasdi.com

"Ayasdi's approach using Topological Data Analysis (TDA) is one of the **top 10 innovations** developed at DARPA in the last decade."

DARPA

Tony Tether, Director
Defense Advanced Research Projects Agency (2001–2009)