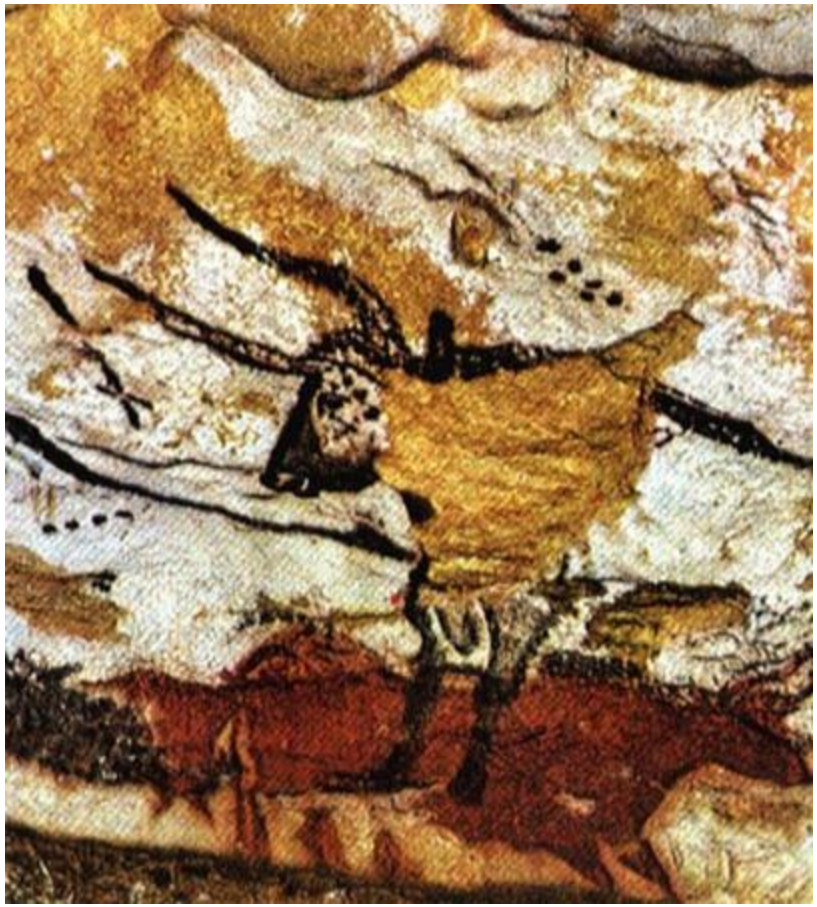
From latent spaces to living systems: Lecture 3 & 4

Dynamic tissue cartography

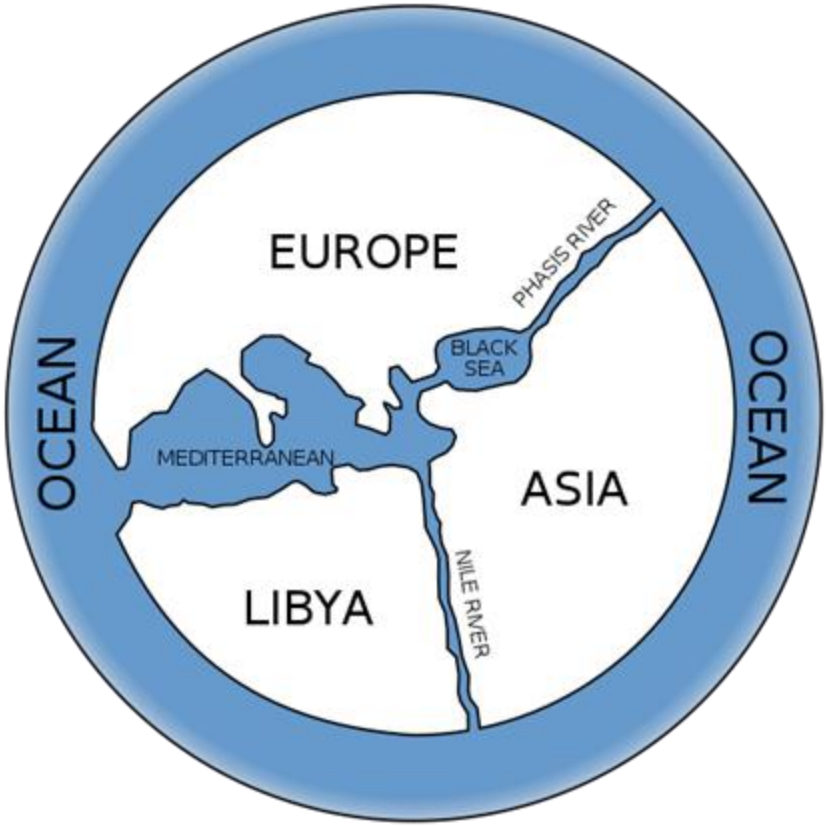
Tiled Illustration: Creation of the Colors, Máximo Laura

Maps and Census as ways to record data and tell stories

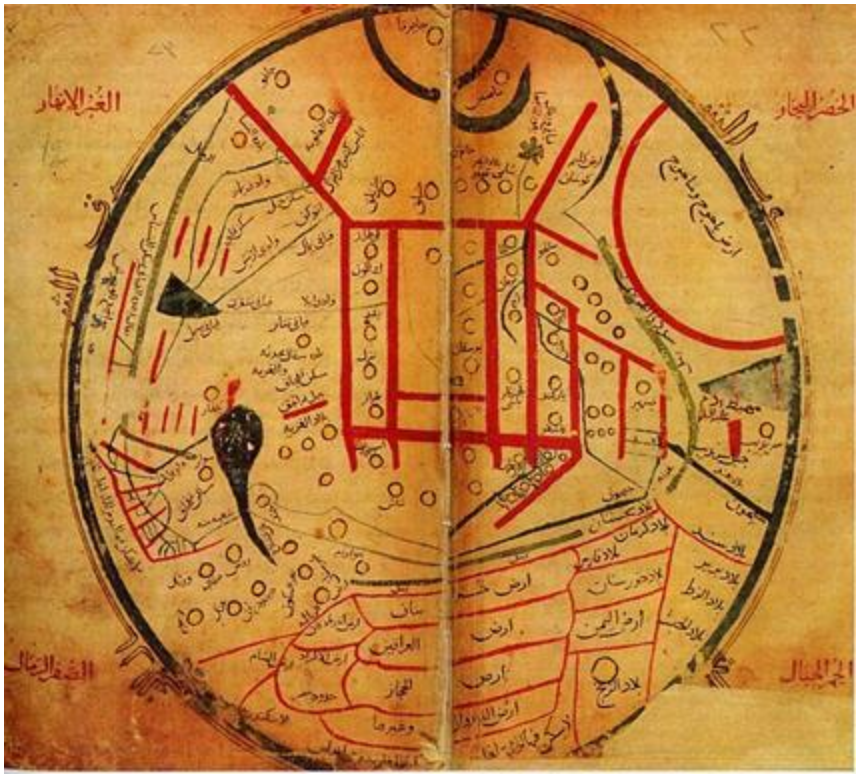


Year Created: c.17,000 BCE
Country of Origin: near Montignac, France, **Creator:** Unknown
Area Depicted: Area around Abauntz Lamizulo cave and animals such as red deer and ibex

Maps can tell many different stories



Year Created: c.610 – 546 BCE
Country of Origin: Ancient Greek city of Miletus
Creator: Anaximander



Kara-Khanid scholar Mahmud al-Kashgari compiled a *Dīwān Lughāt al-Turk* (*Compendium of the languages of the Turks*) in the 11th century. Source: Wikipedia

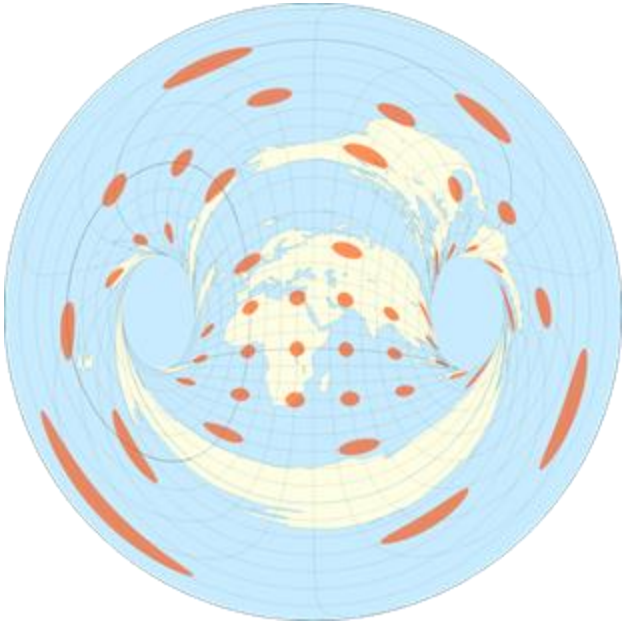
Census records tell complementary stories

Quipus from Highlands of southern Peru
Imperial Epoch (1300 d.C. – 1532 d.C.) Museo Larco

Maps and Censuses

- Usually depict elements that are **known** (identifiable) with the appropriate scales and legends (rivers, factual information about number of cattle)
- Information can be spatial (map), temporal (census) or both (climate data)

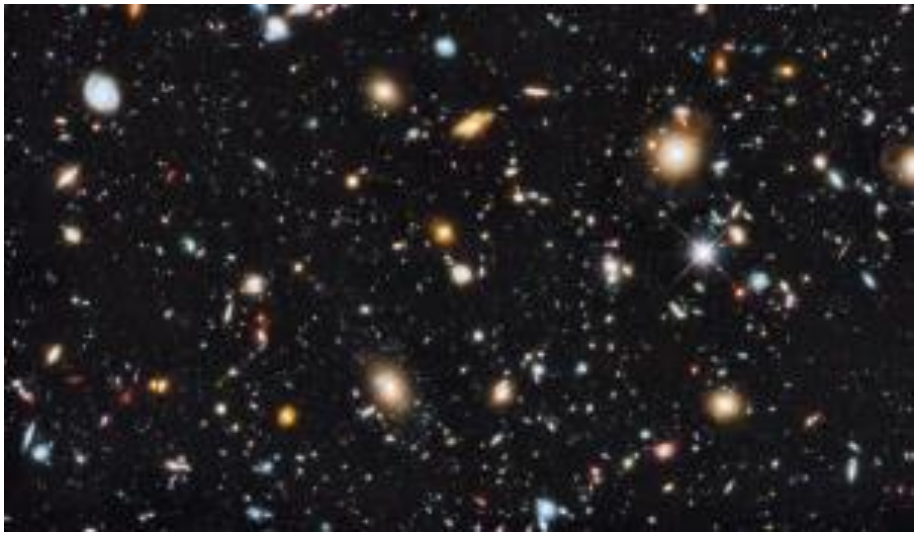
Maps and Censuses



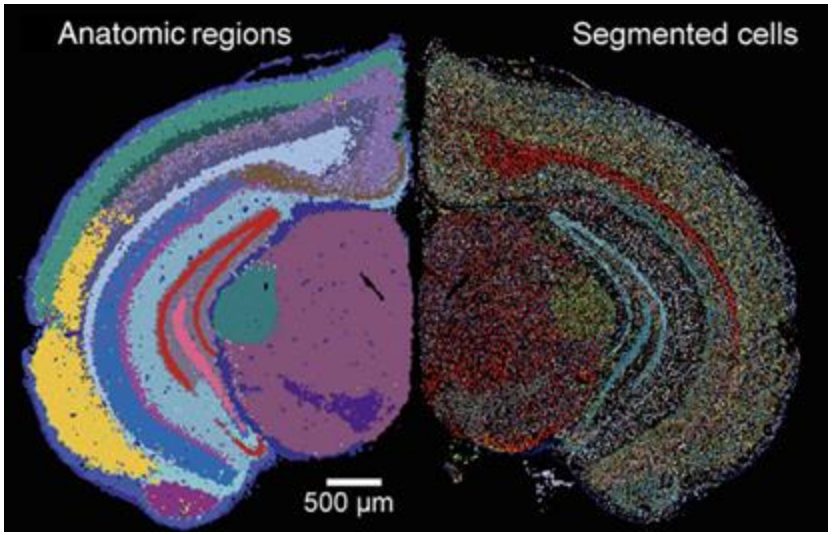
- Reflect a particular point of view and mislead (e.g.polling numbers)
- Subject to bias and error
- Require clear knowledge of the elements presented
- Need to be continuously updated to reflect changes

Figure: Hammer retroazimuthal projection.
Source:Wikipedia

Astronomy and Biology give us new Maps



Hubble NASA



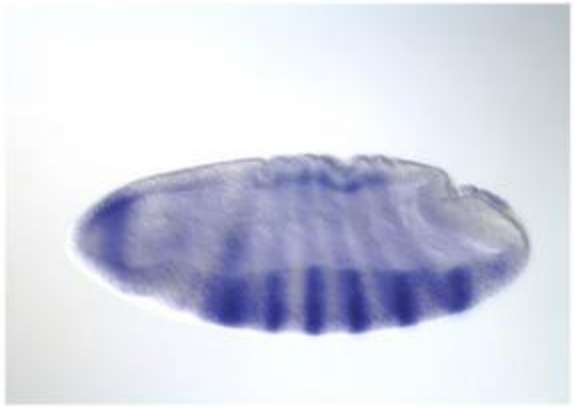
Mouse organogenesis spatiotemporal transcriptomic atlas. (Chen et al., 2022)

To make Maps that represent Living Systems we need two answer at least two Questions

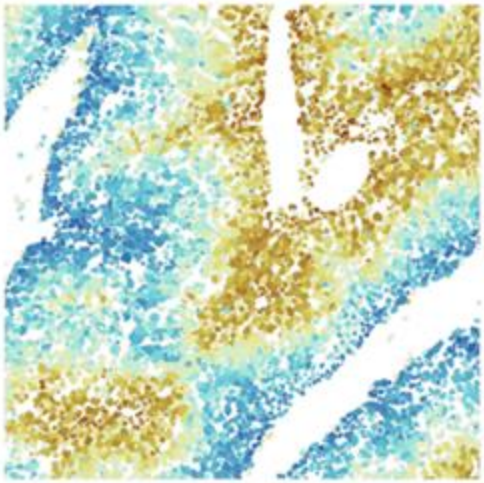
- **What are the units?**
  If we do not yet know the key components or players, how can we discover and define them?

- **How are they organized and how do they interact?**
  Given the players, what are the rules and patterns that govern their interactions **across spatial and temporal scales**?

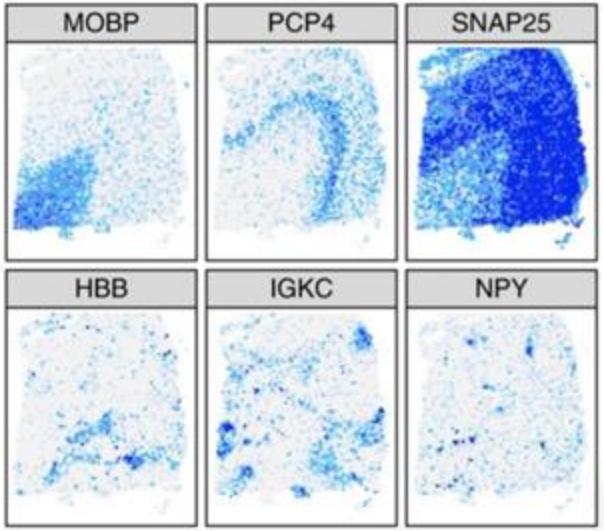
Spatial transcriptomics: mapping living systems



BDGP



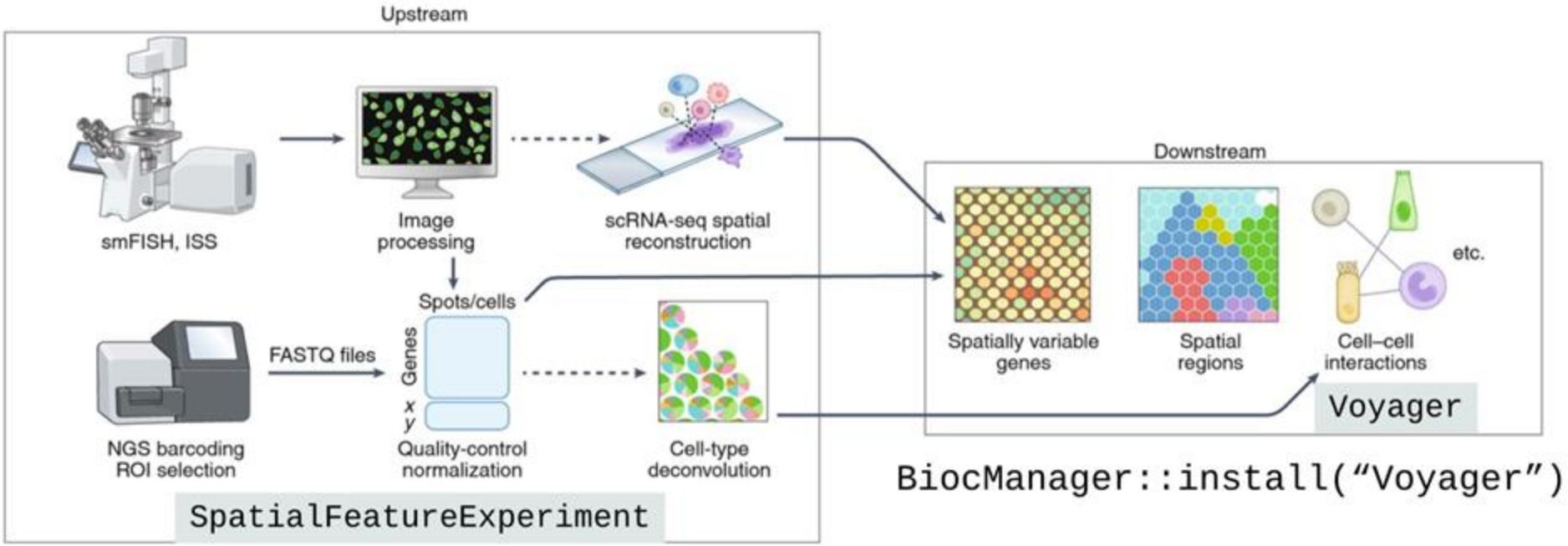
Voyager website



Maynard 2021 DLPFC

Figure: Lambda Moses

Spatial transcriptomics: mapping living systems

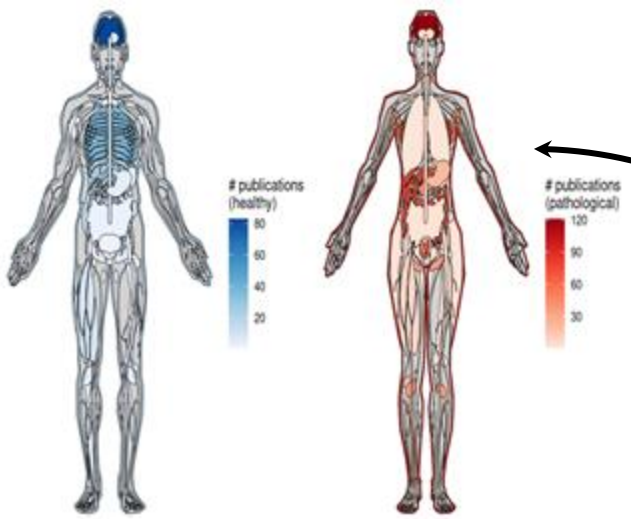


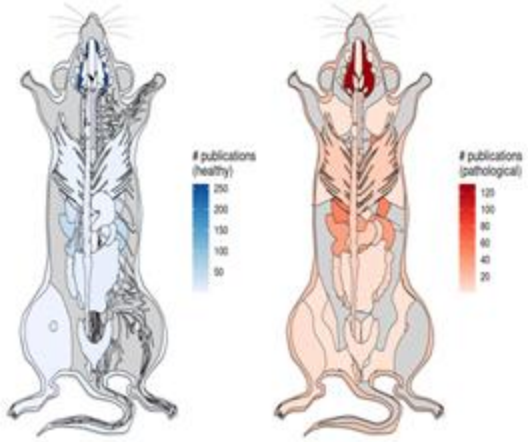

```
BiocManager::install("Voyager")
```

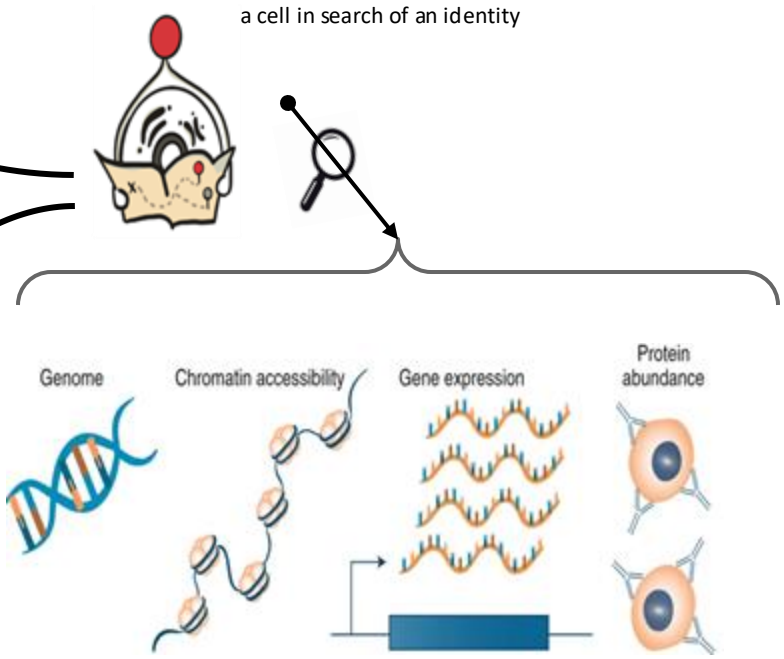
```
BiocManager::install("SpatialFeatureExperiment")
```

https://pachterlab.github.io/voyager/

Cellular heterogeneity across scales underlies healthy development and disease



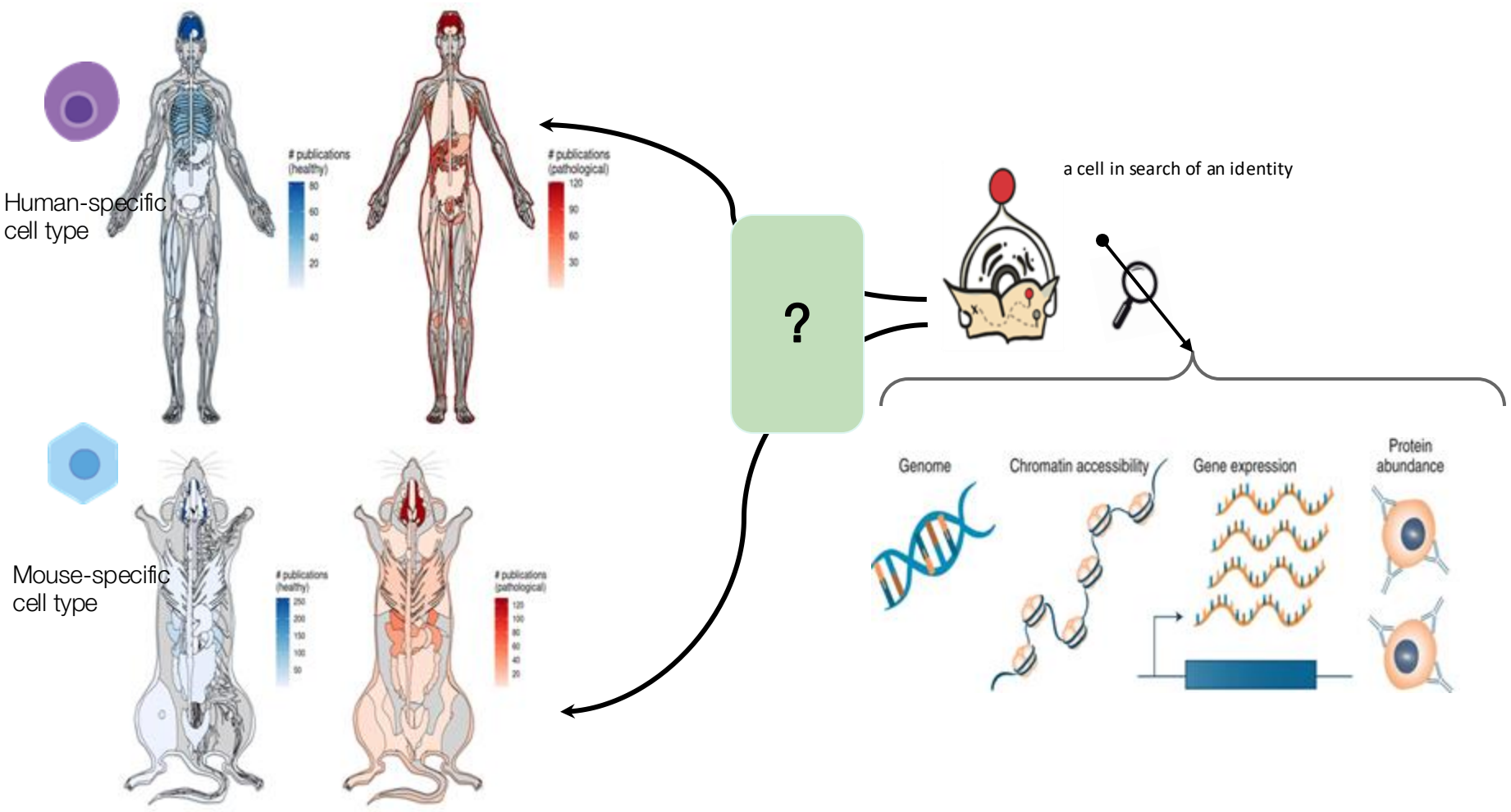




Adapted from Moses, Lambda. *Museum of spatial transcriptomics*. Nat Methods. 2022.

Adapted from Efremova. *Computational methods for single-cell omics across modalities*. Nat Methods. 2020.

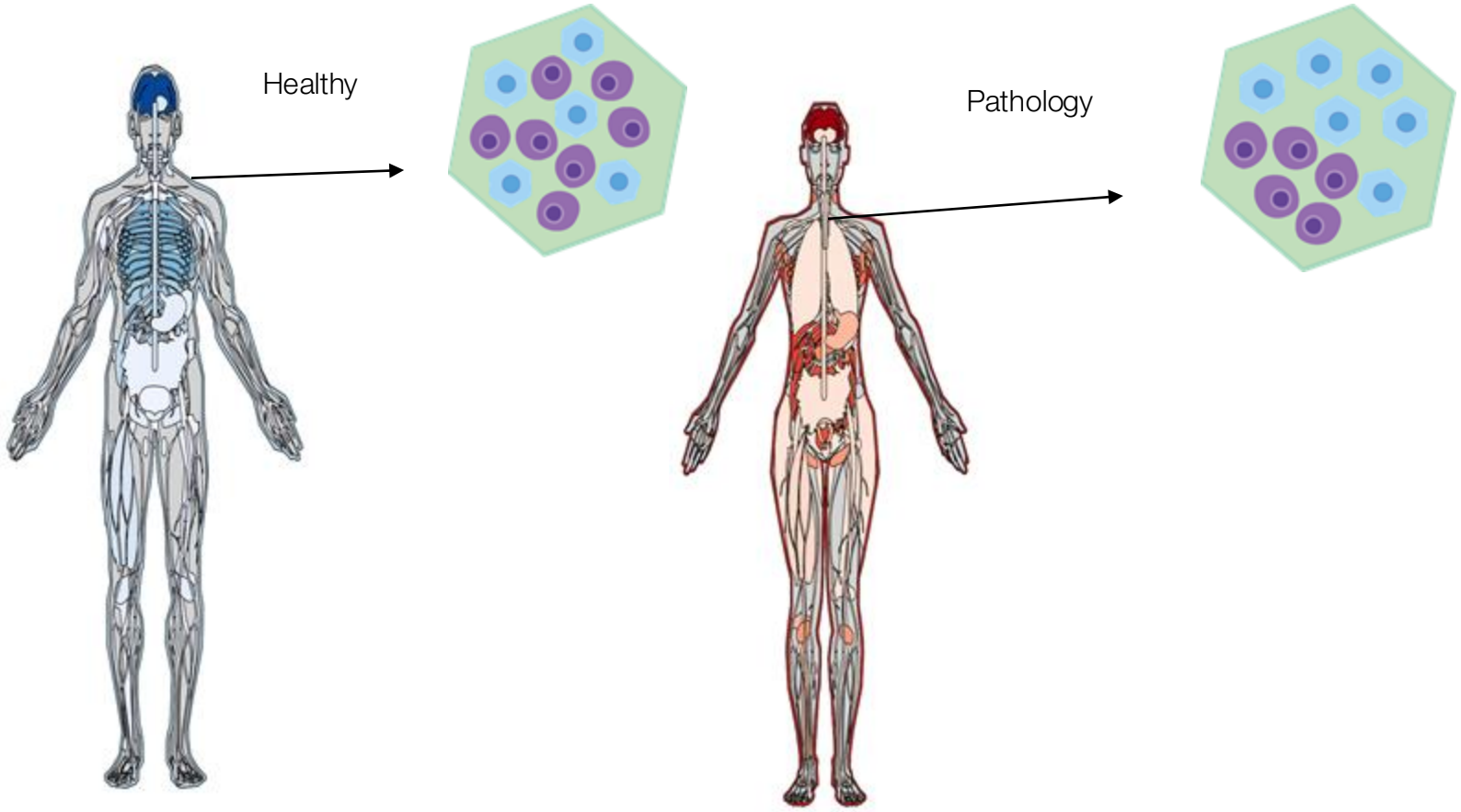
Cellular heterogeneity across scales underlies healthy development and disease



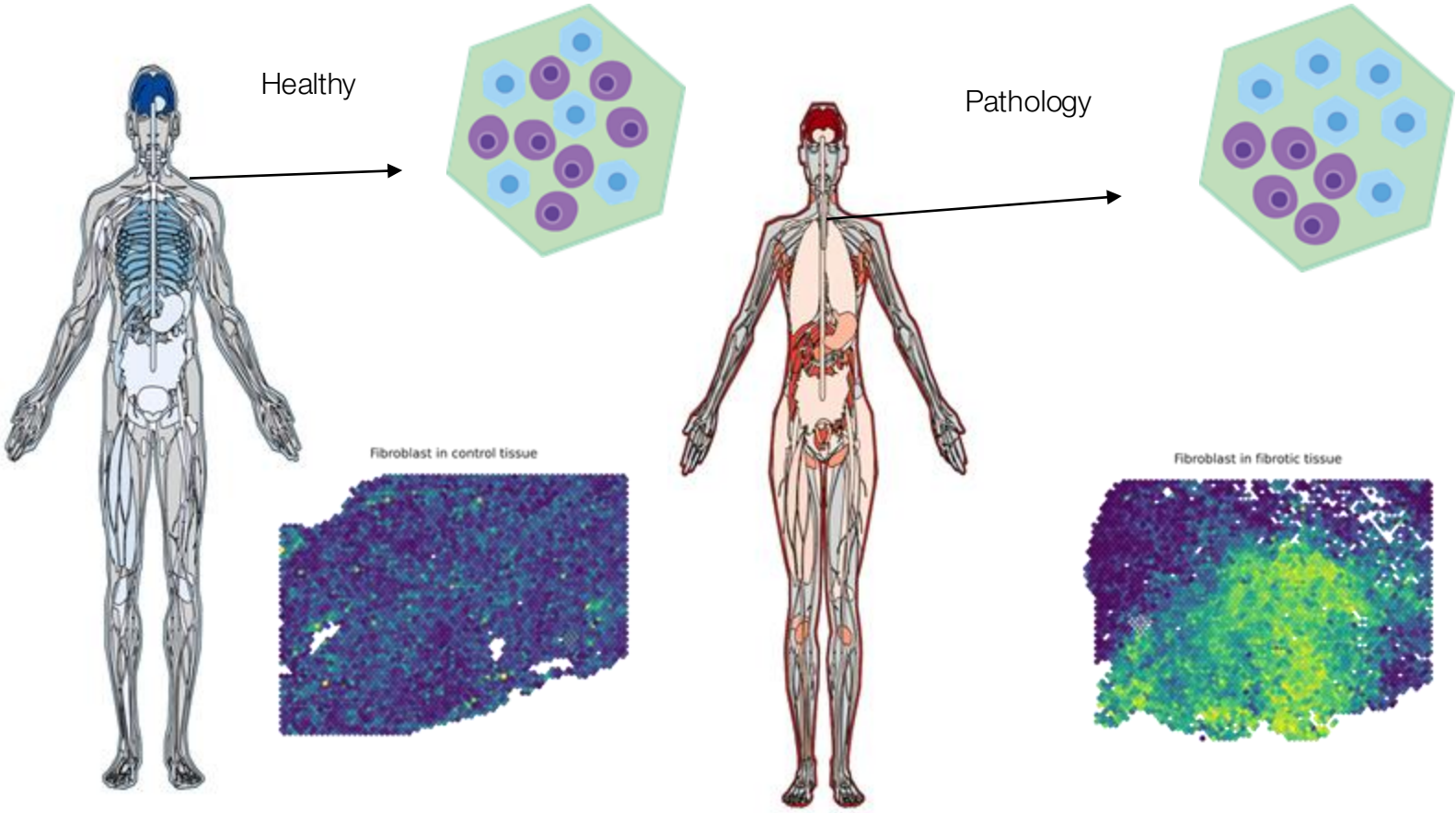
Adapted from Moses, Lambda. *Museum of spatial transcriptomics*. Nat Methods. 2022.

Adapted from Efremova. *Computational methods for single-cell omics across modalities*. Nat Methods. 2020.

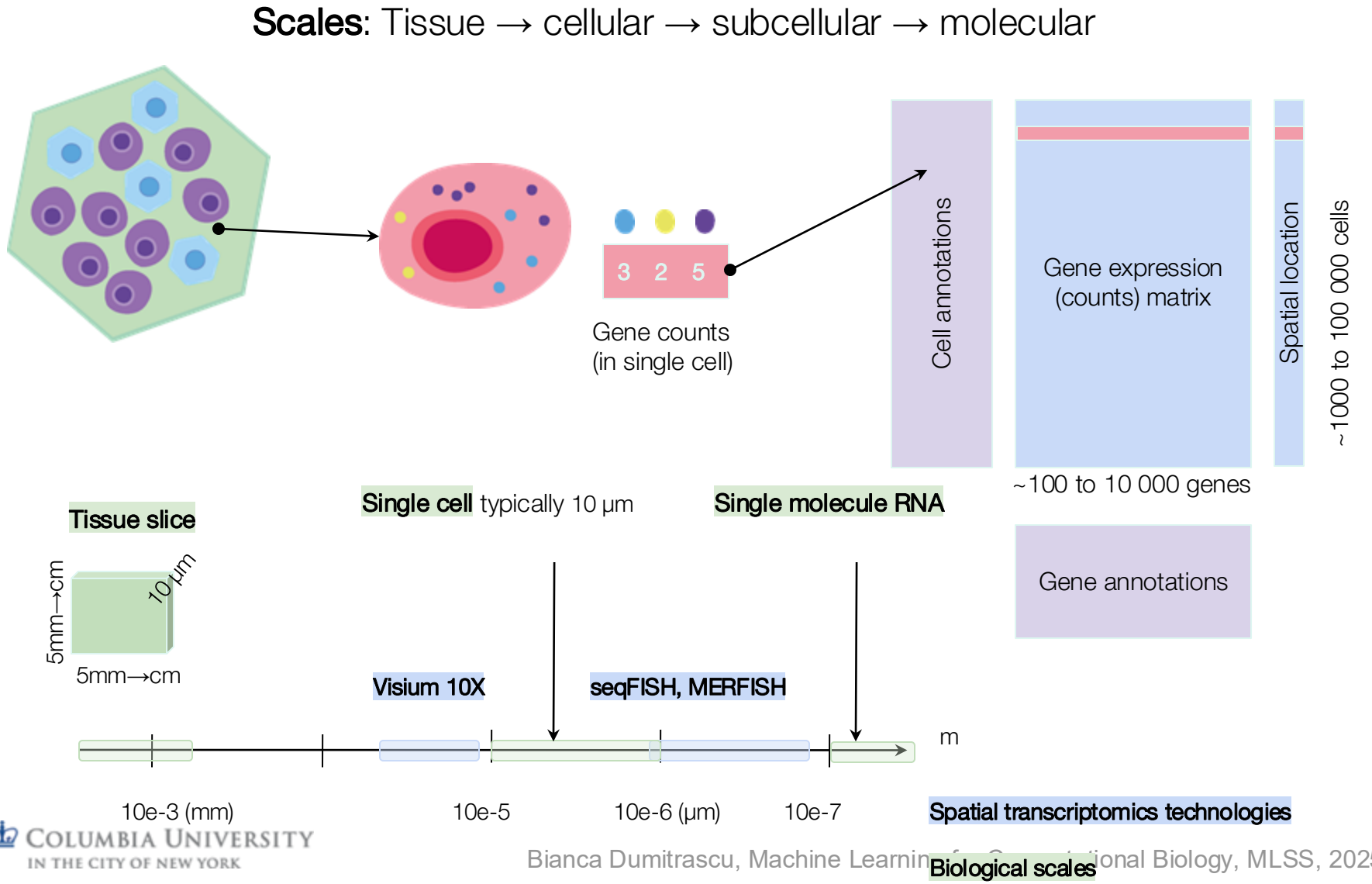
Spatial contexts can drive differences in health and disease contexts



Cellular heterogeneity across scales underlies healthy development and disease



Cellular heterogeneity: a multiscale problem

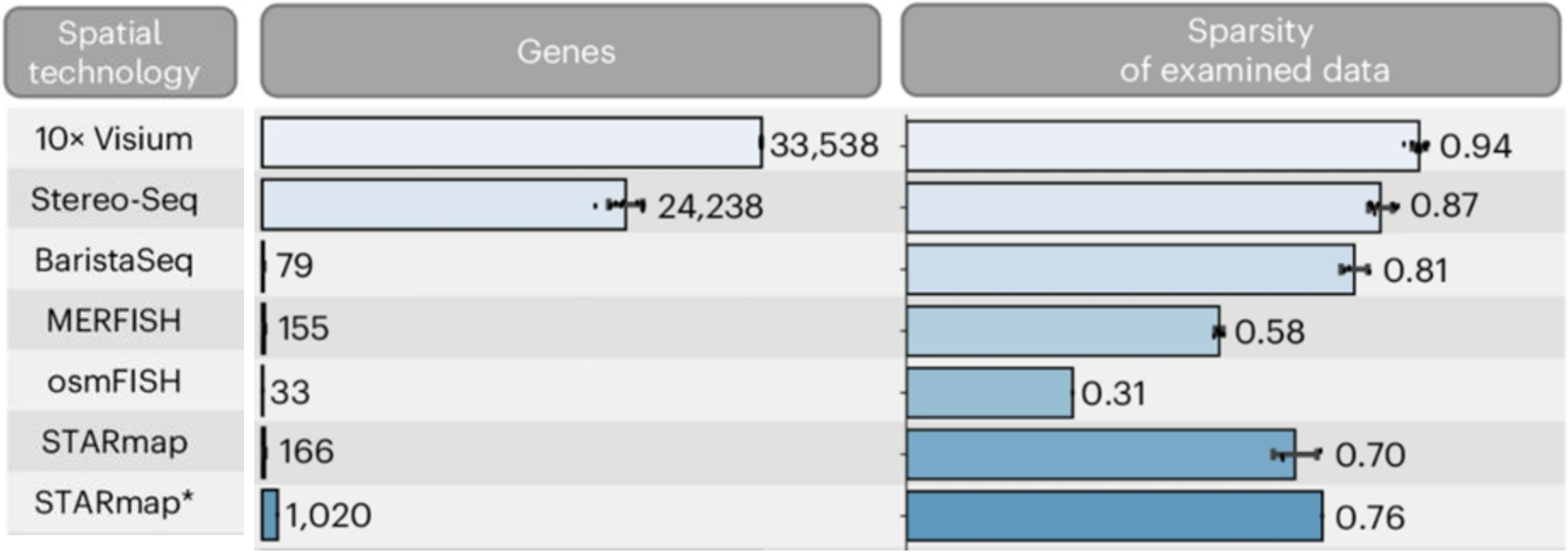


Cellular heterogeneity: a multiscale problem with many tools to query it

Spatial technology	Data number	Size/ radius	Spots
10× Visium	12	55 μm	3,844
Stereo-Seq	9	25 μm	5,913
BaristaSeq	3	≤1 cell	1,690
MERFISH	5	≤1 cell	5,557
osmFISH	1	≤1 cell	4,839
STARmap	3	≤1 cell	1,053
STARmap*	1	≤1 cell	1,207

From: Yuan, Z., Zhao, F., Lin, S. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data.2024

Cellular heterogeneity: a multiscale problem with many tools to query it



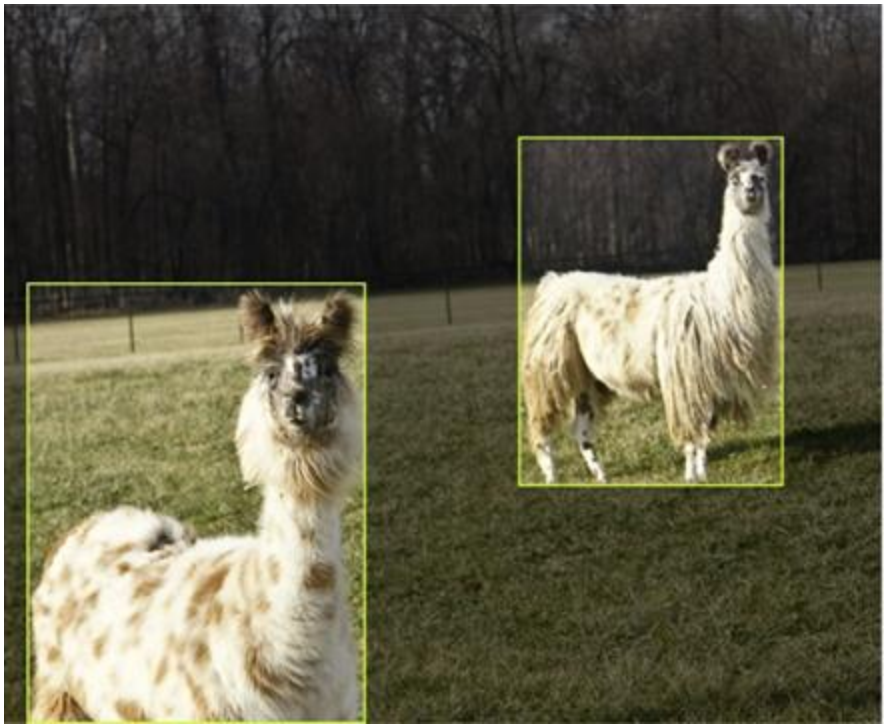
From: Yuan, Z., Zhao, F., Lin, S. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data.2024

To make Maps that represent Living Systems we need two answer at least two Questions

- **What are the units? (This lecture: Spatial Patterns)**
  If we do not yet know the key components or players, how can we discover and define them?

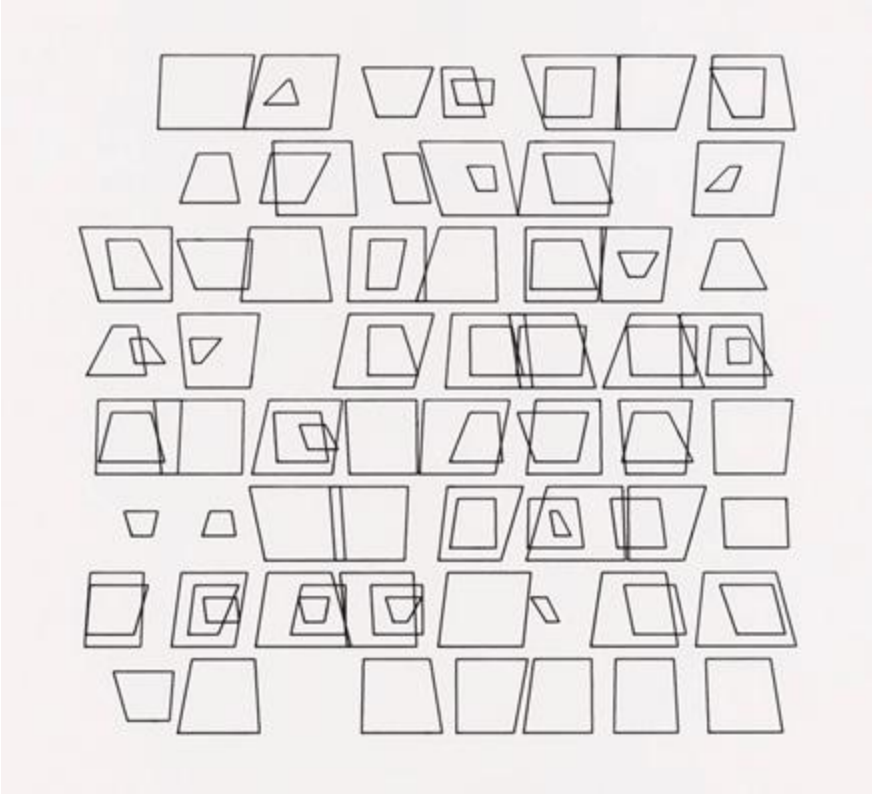
- **How are they organized and how do they interact?**
  Given the players, what are the rules and patterns that govern their interactions **across spatial and temporal scales**?

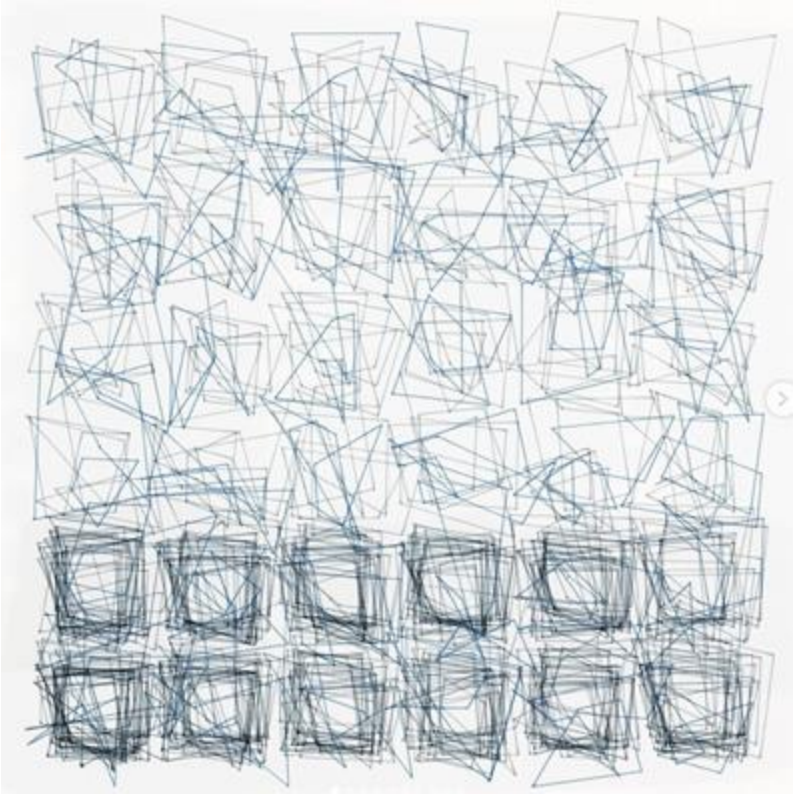
Broader goal: quantify spatial patterns!



Con: Spatial Patterns can be difficult to define

Patterns can vary in space

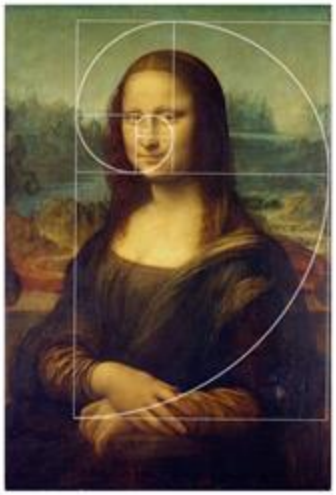


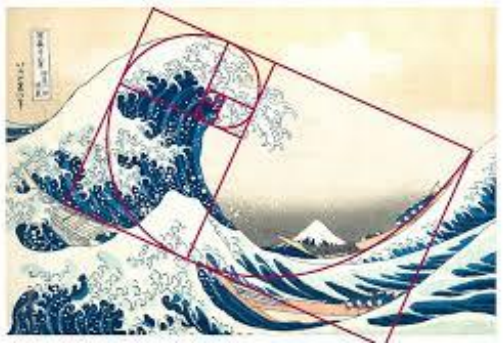


Vera Molnar, Génèse du Trapèze, 1974: art between the three cons: Conceptualists, Constructivists, and computers

Patterns can segment space

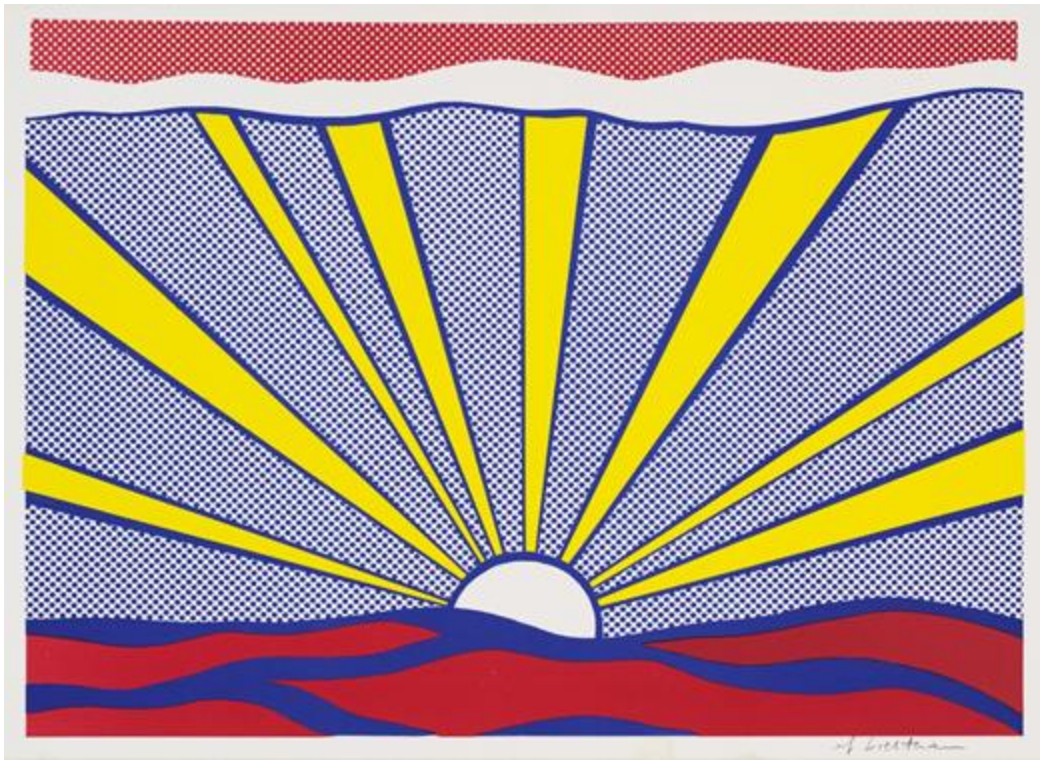
Patterns can be latent





Golden ratios

Patterns can make sense semantically



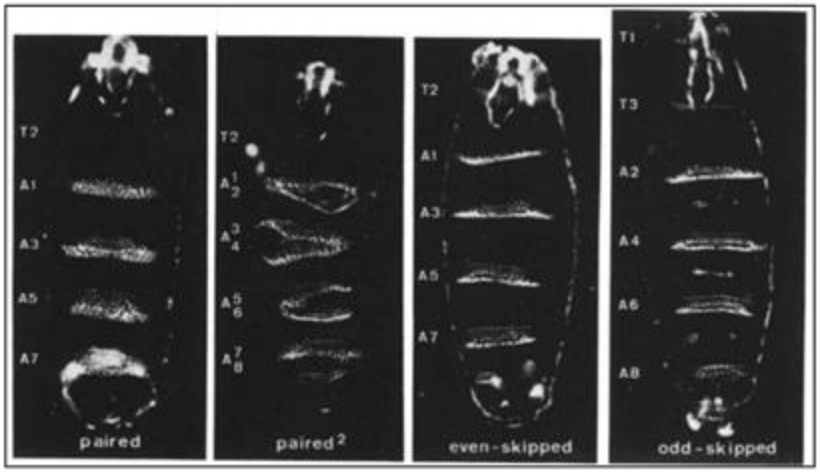
Sunrise (1965). Roy Lichtenstein.
Whitney Museum of American Art, New York;

George Seurat cc 1884. Top: Farm people,
Guggenheim Museum

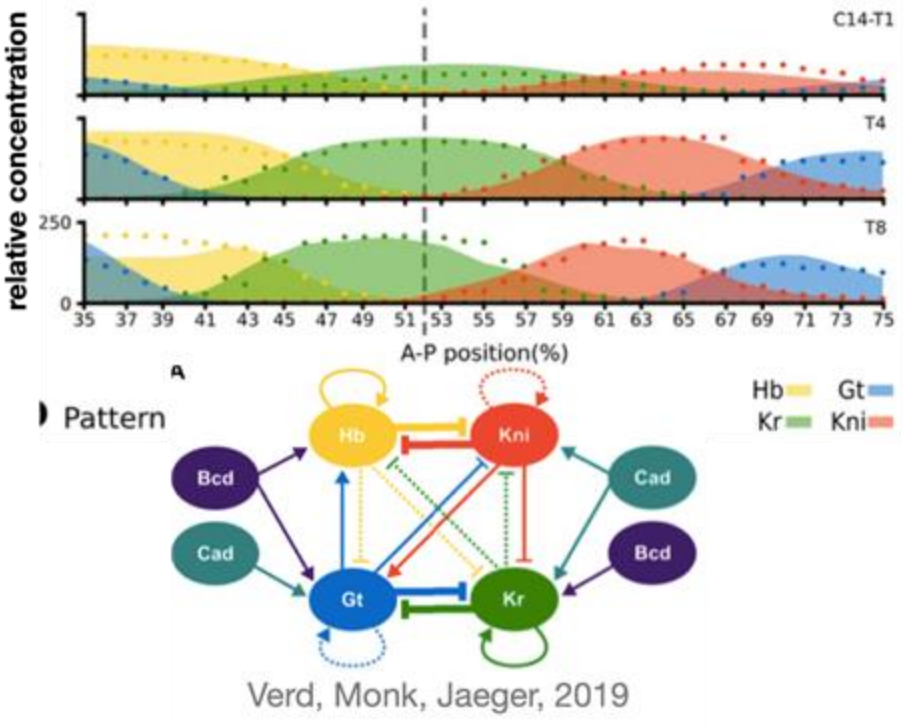
Patterns can be present at different scales

Paracas mantles. Museo Larco

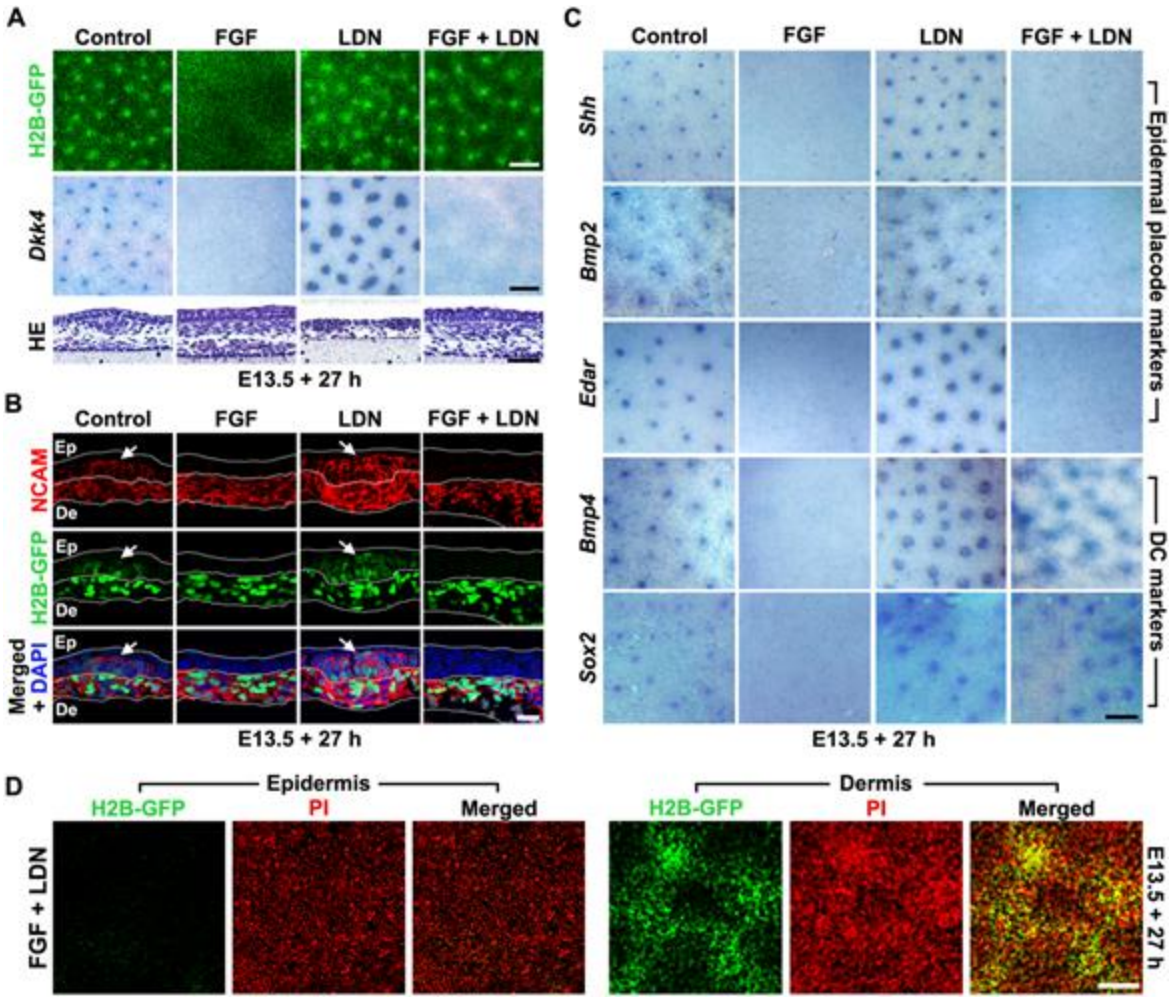
Biological Patterns: Fruit Fly Stripes



"Mutations affecting segment number and polarity in *Drosophila*"
by C. Nüsslein-Volhard and E. Wieschaus, 1980

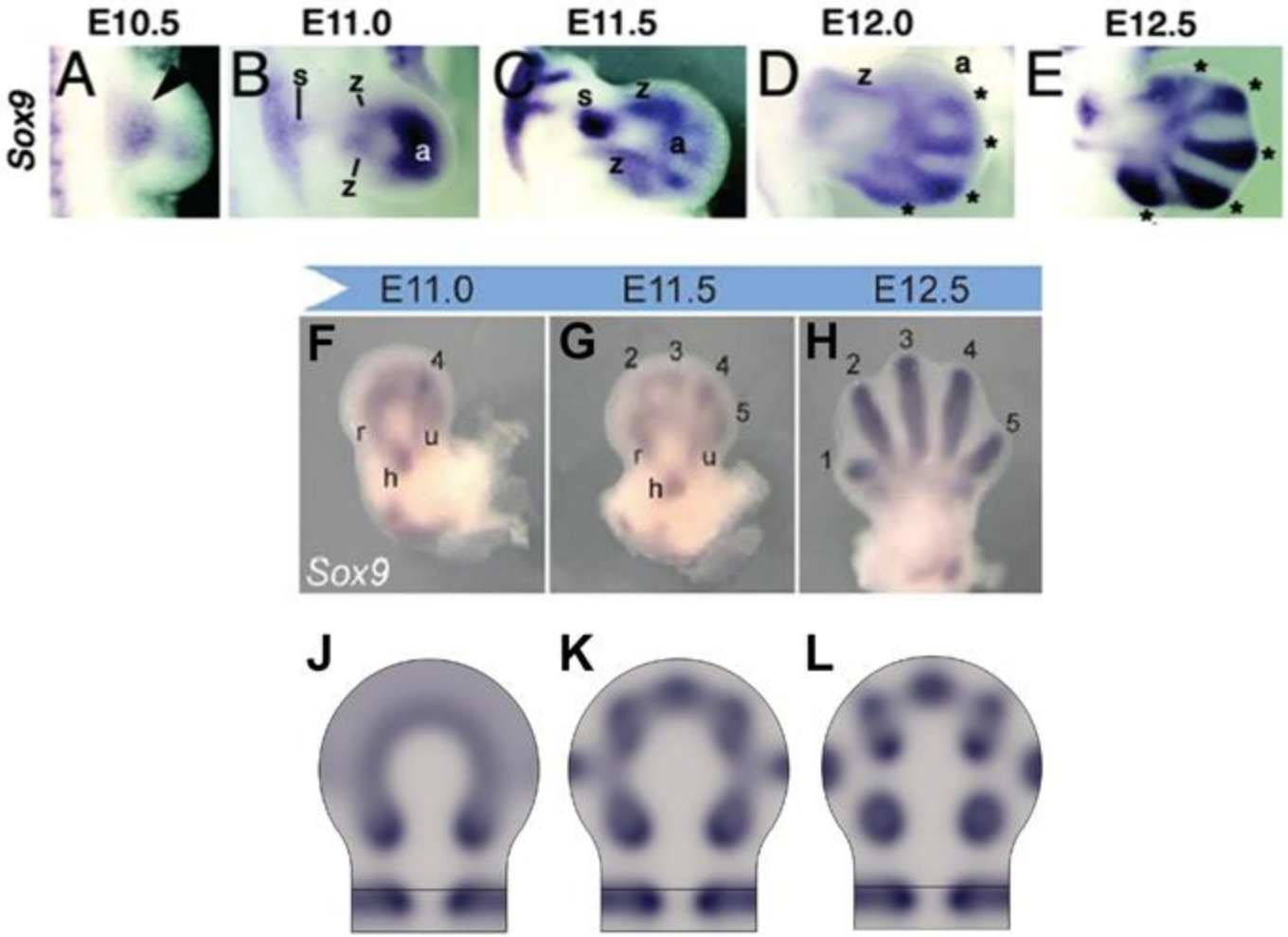


Biological Patterns: Hair Follicles



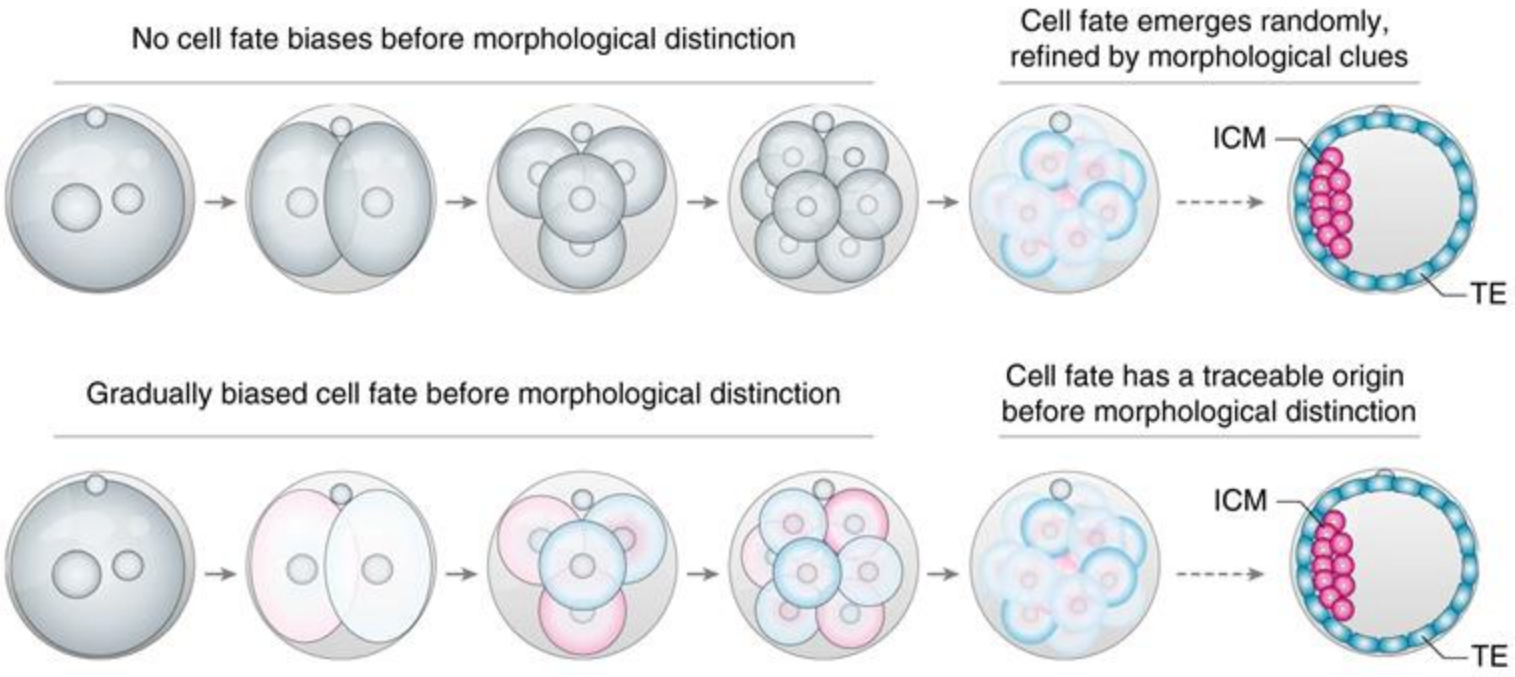
From: Hierarchical patterning modes orchestrate hair follicle morphogenesis. Glover et al. (2017)

Biological Patterns: Digit Formation



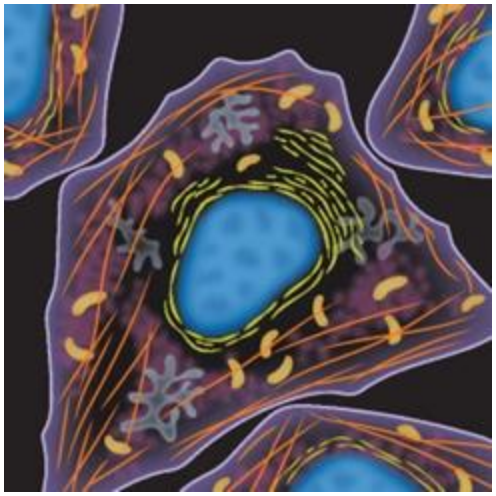
From: Digit patterning during limb development as a result of the BMP-receptor interaction. Badugu et al (2012)

Biological Patterns: Early Embryo Symmetry Breaking



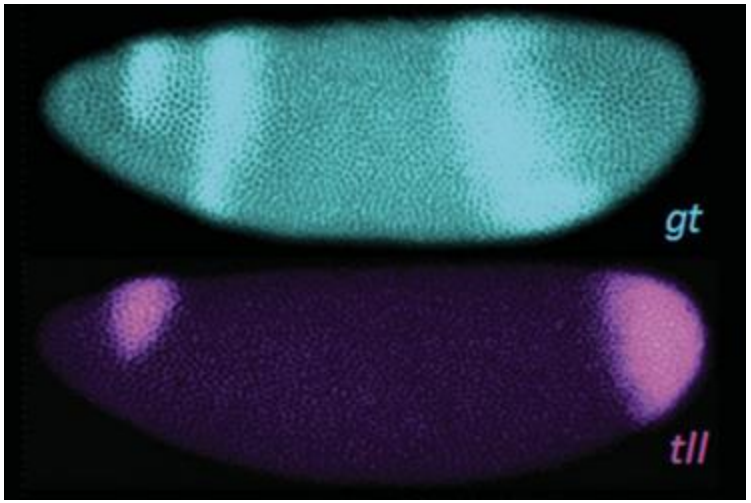
Adapted from: Tracing the origin of heterogeneity and symmetry breaking in the early mammalian embryo. Chen et al. (2018)

Examples of Spatial Patterns at the Single Cell Level



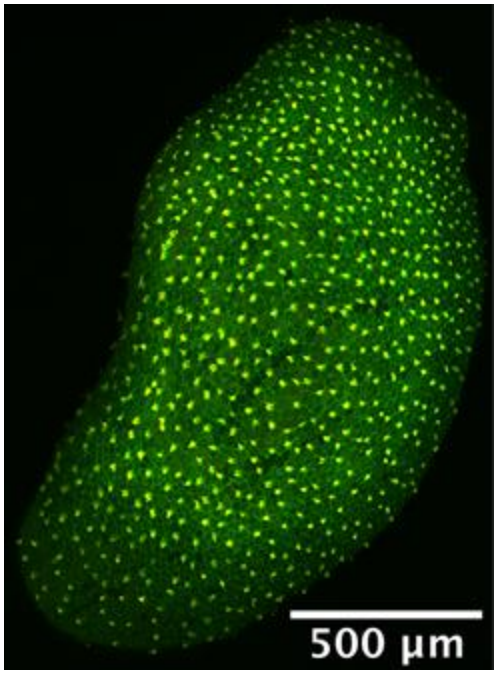
An artist's rendition of subcellular imaging. Credit: Debbie Maizels

proteins are spatially distributed at subcellular level



Credit: S. Shvartzman & R. Baker

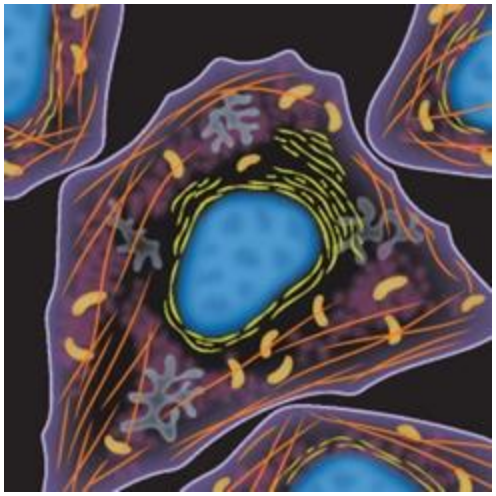
genes or gene programs can vary spatially, in a cell type dependent or cell type independent manner



Credit: Frog embryos from Jakub Sedzinski

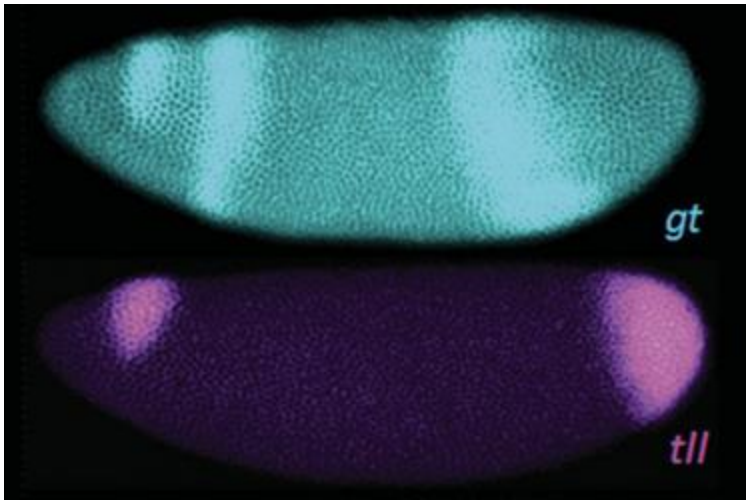
cell types themselves can form spatial units

Examples of Spatial Patterns at the Single Cell Level



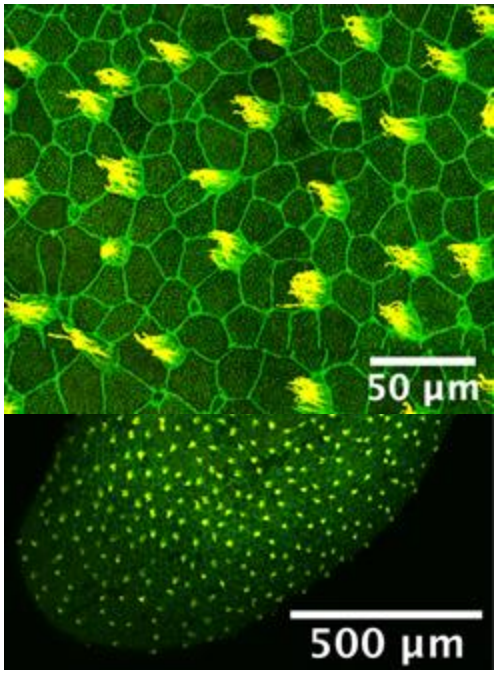
An artist's rendition of subcellular imaging. Credit: Debbie Maizels

proteins are spatially distributed at subcellular level



Credit: S. Shvartzman & R. Baker

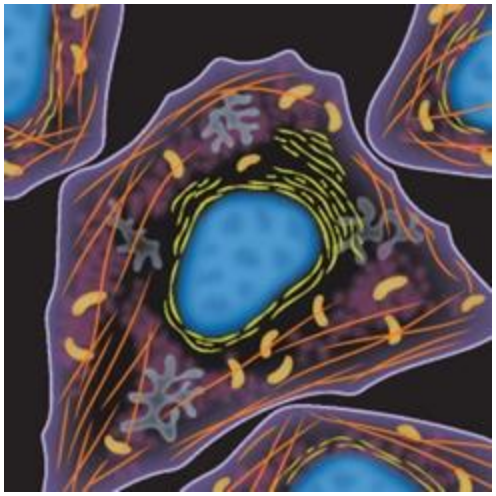
genes or gene programs can vary spatially, in a cell type dependent or cell type independent manner



Credit: Frog embryos from Jakub Sedzinski

cell types themselves can form spatial units

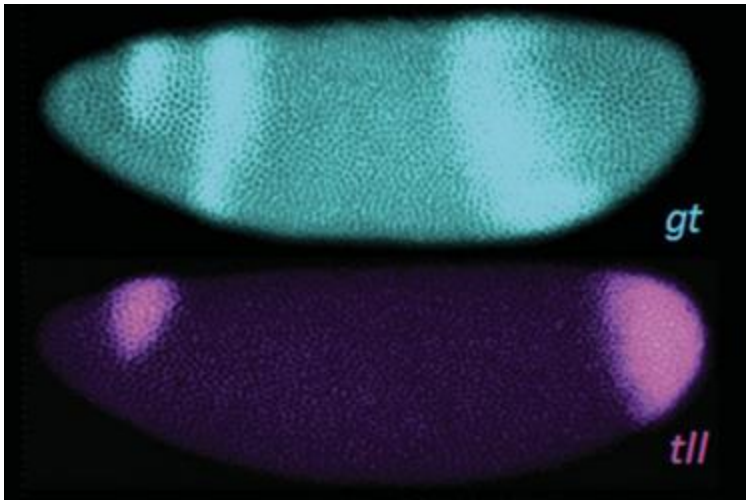
Examples of Spatial Patterns at the Single Cell Level



An artist's rendition of subcellular imaging. Credit: Debbie Maizels

Testing for randomness of spatial events

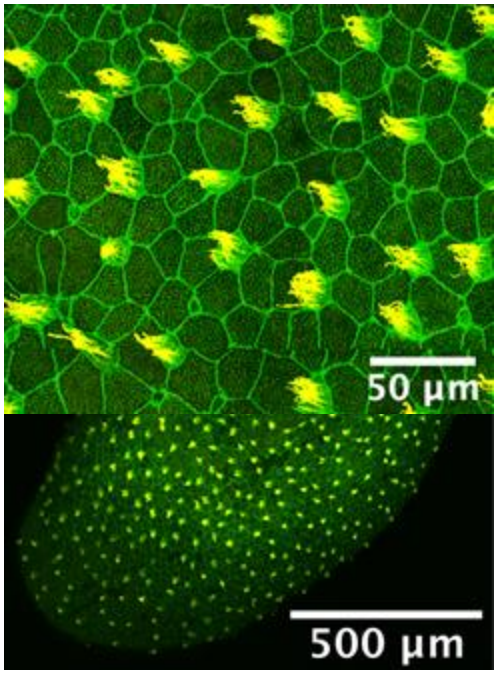
proteins are spatially distributed at subcellular level



Credit: S. Shvartzman & R. Baker

Spatial Variable Gene Detection

genes or gene programs can vary spatially, in a cell type dependent or cell type independent manner



Credit: Frog embryos from Jakub Sedzinski

Spatial Clustering

cell types themselves can form spatial units

Computational tools for spatial data analysis

Areal models:
- Moran's I
- Conditional Autoregressive Models

In context
- (gene) SpatialDE, nnSVG
- (gene programs)Spatial NMF

GNNS
- GraphST

Testing for randomness of spatial events

proteins are spatially distributed at subcellular level

Spatial Variable Gene Detection

genes or gene programs can vary spatially, in a cell type dependent or cell type independent manner

Spatial Clustering

cell types themselves can form spatial units

Modelling Spatial Correlation with Areal Models

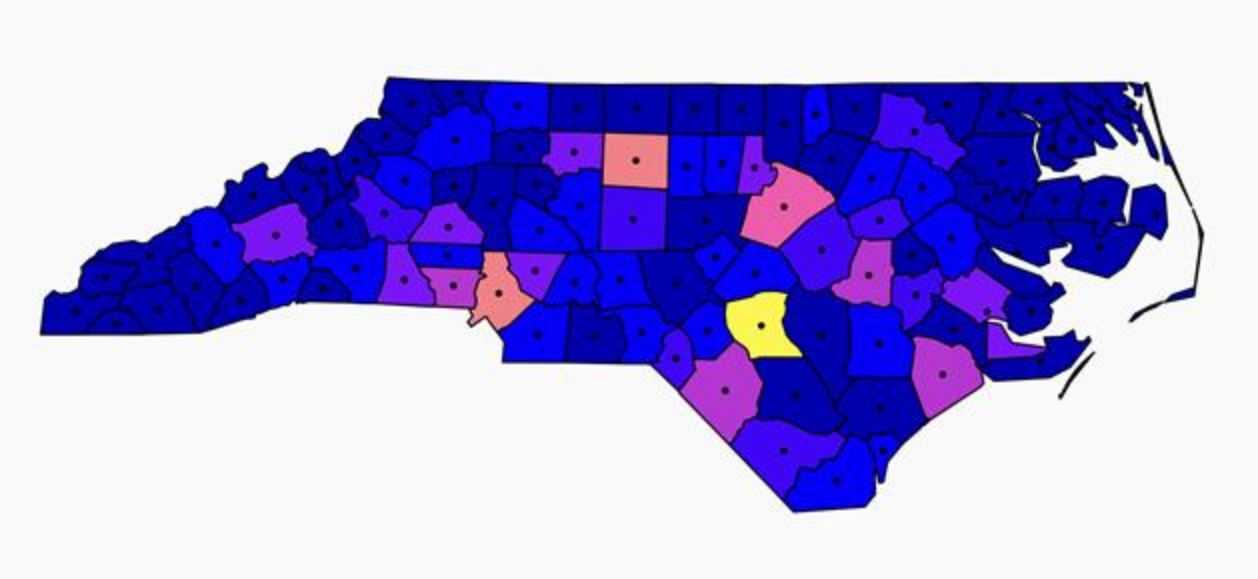
Areal models:

- Moran's I
- Conditional Autoregressive Models

Modelling Spatial Correlation with Areal Models

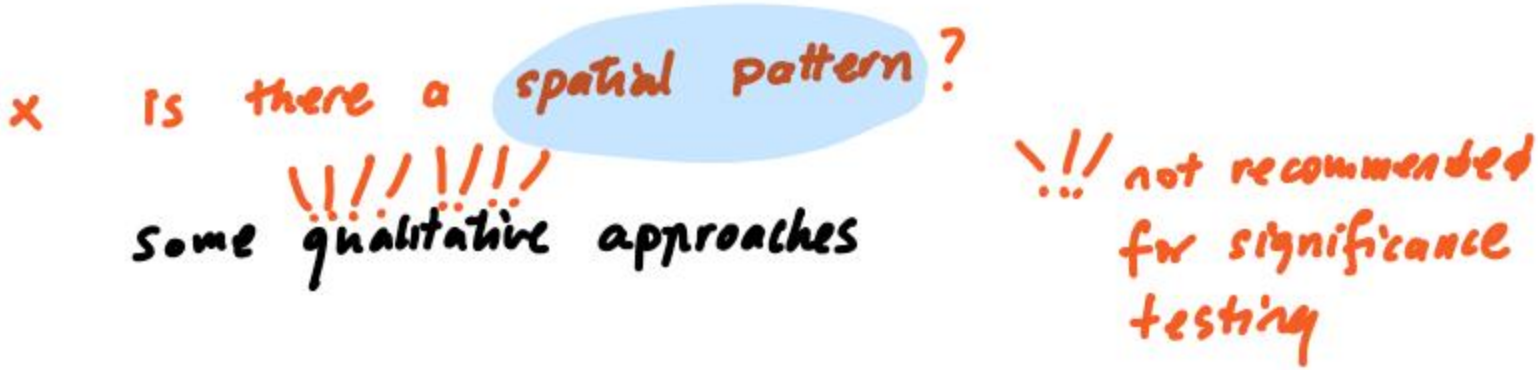
Discrete Spatial Modeling
(Joint model for a finite set of variables)

Cases
- irregular geographic units
- regular grid cells (pixels)



Areal Data Models : statistical questions

- is there a spatial pattern? (random vs units close by are more similar to one another that units far away)
- do we need to smooth the data?
- for new areal units can we make new predictions?


x Is there a spatial pattern?
!!!!!!!!
some qualitative approaches
!! not recommended for significance testing

x Is there a spatial pattern ?

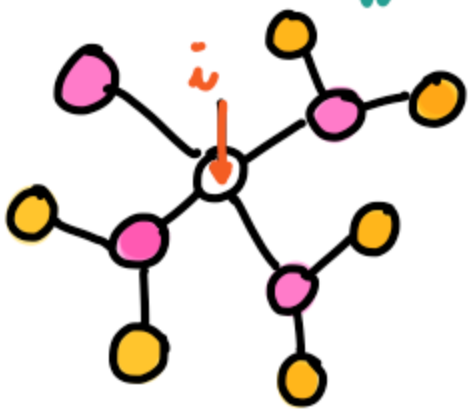
!!!!!!!!

some qualitative approaches

!!! not recommended for significance testing

setup: $Y_1, Y_2, \ldots Y_n$

W: proximity matrix $w_{ij} = \begin{cases} 1 \\ 0 \end{cases}$ ← if a boundary is shared



$$w_{i+} = \sum_j w_{ij}$$

$$\tilde{w}_{ij} = \frac{w_{ij}}{w_{i+}}$$

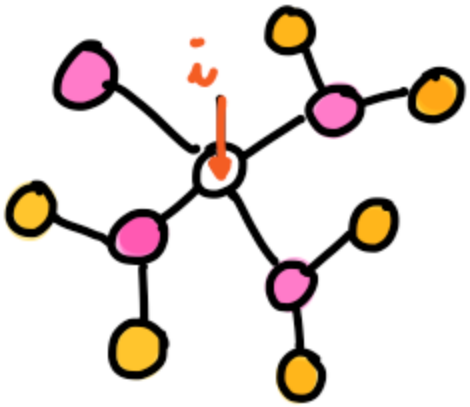
$w_{ij}^{1} = 1$ first order neighbours

$w_{ij}^{2} = 1$ 2nd

x Is there a spatial pattern ?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations

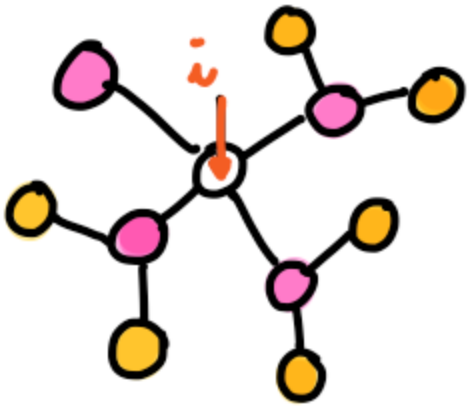


Moran's I : $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

x Is there a spatial pattern ?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations



Moran's I : $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

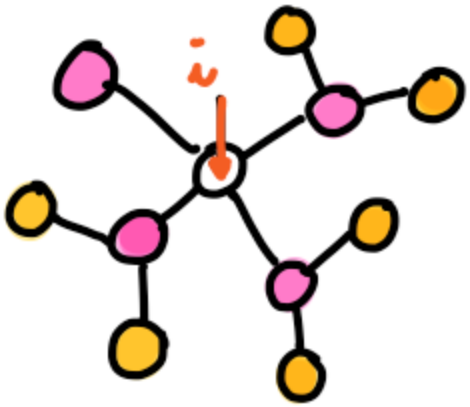
Statistically we care, for example, about the asymptotic behavior of I

x delta method

x Is there a spatial pattern ?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations



Moran's I : $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

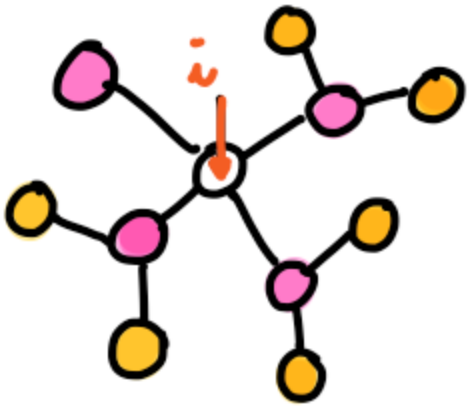
Statistically we care, for example, about the asymptotic behavior of I $\sqrt{n}[x_n - \theta] \xrightarrow{D} N(0, \sigma^2)$

x delta method $\sqrt{n}[g(x_n) - g(\theta)] \xrightarrow{D} N(0, \sigma^2 [g'(\theta)]^2)$

x Is there a spatial pattern ?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations



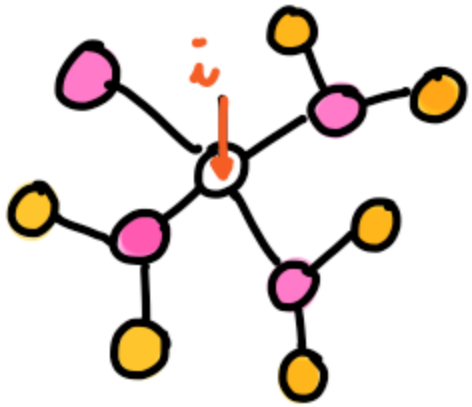
Moran's I : $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

Q :
- multivariate Moran's I ?
- use Moran's I as a selection criteria ?

× Is there a spatial pattern?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations



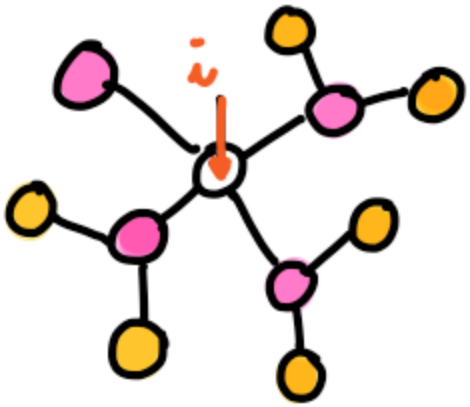
Moran's I: $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

Geary's C: $$C = \frac{(n-1) \sum_i \sum_j w_{ij} (y_i - y_j)^2}{2 \left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

x Is there a spatial pattern ?

!!!!!!!!

Some qualitative ~~approaches~~ measures of spatial associations



Moran's I : $$I = \frac{n \sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (y_i - \bar{y})^2}$$

$I^1 : w^1$

$I^2 : w^2$

$w_{ij}^1 = 1$ first order neighbours

$w_{ij}^2 = 1$ 2nd

x do we need to smooth the data?

why would we want to do this in the first place?

x do we need to smooth the data?

$Y_1, \ldots Y_N$; W

$$Y_i \rightarrow \hat{Y}_i = \sum_j \frac{w_{ij}}{w_{i+}} Y_j$$

what's strange about this?

x do we need to smooth the data?

$Y_1, \ldots Y_N$; W

$$Y_i \rightarrow \hat{Y}_i = \sum_j \frac{w_{ij}}{w_{i+}} Y_j$$

what's strange about this?

x do we need to smooth the data?

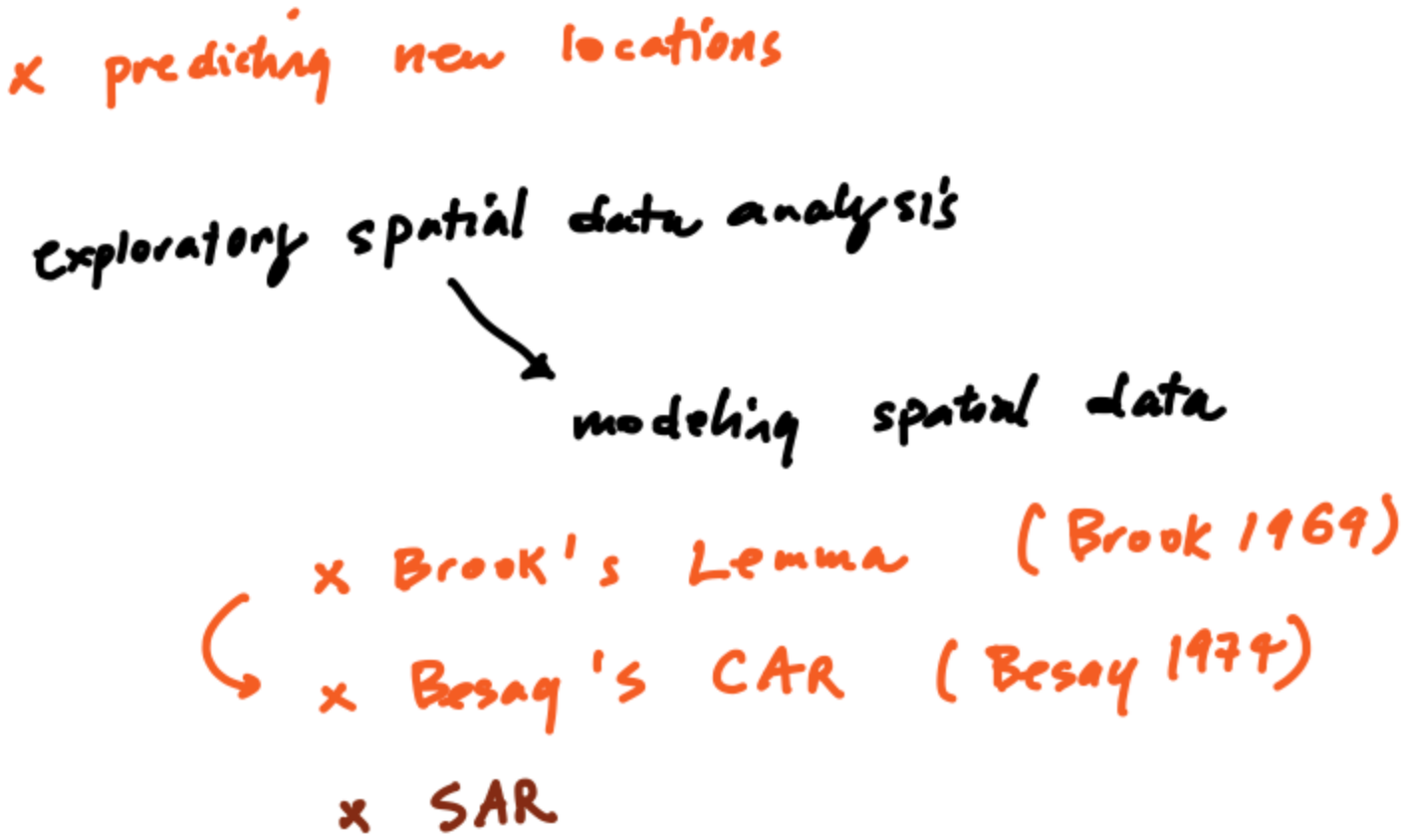
$Y_1, \dots Y_N$; W

$$Y_i \rightarrow \hat{Y}_i = \sum_j \frac{w_{ij}}{w_{i+}} Y_j$$

alternative: $$\hat{Y}_i^* = (1-\alpha) Y_i + \alpha \hat{Y}_i$$

Q: • GNN as nonlinear smoothers?

• what's the optimal α?


x predicting new locations
exploratory spatial data analysis
modeling spatial data
x Brook's Lemma (Brook 1969)
x Besag's CAR (Besag 1974)
x SAR

Brook's Lemma

Given $p(y_1, y_2, \dots y_n)$ the full conditionals

$p(y_i | y_j, j \neq i)$, $i = 1, n$

are uniquely determined

= WHY ? =

Brook's Lemma

"$\Rightarrow$"

Given $p(y_1, y_2, \dots y_n)$ the full conditionals

$p(y_i | y_j, j \neq i)$, $i = 1, n$

are uniquely determined

Is the "$\Leftarrow$" statement also true?

Brook's Lemma

$$Y_1 | Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$$
$$Y_2 | Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2)$$

$$E[Y_1] = E[E[Y_1 | Y_2]] = E[\alpha_0 + \alpha_1 Y_2]$$
$$= \alpha_0 + \alpha_1 E[Y_2]$$

$$E[Y_2] = E[E[Y_2 | Y_1]] = E[\beta_0 + \beta_1 Y_1^3] =$$
$$= \beta_0 + \beta_1 E[Y_1^3]$$

Brook's Lemma

$$Y_1 | Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$$

$$Y_2 | Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2)$$

$$E[Y_1] = E[E[Y_1 | Y_2]] = E[\alpha_0 + \alpha_1 Y_2]$$
$$= \alpha_0 + \alpha_1 E[Y_2]$$

$$E[Y_2] = E[E[Y_2 | Y_1]] = E[\beta_0 + \beta_1 Y_1^3] =$$
$$= \beta_0 + \beta_1 E[Y_1^3]$$

$f(y_1 | y_2)$ and $f(y_2 | y_1)$ are incompatible with regard to determining $p(y_1, y_2)$

Brook's Lemma

$$p(y_1, y_2) \propto \exp\left[-\frac{1}{2}(y_1 - y_2)^2\right] \quad (*)$$

$p(y_1 | y_2)$ is $N(y_2, 1)$

$p(y_2 | y_1)$ is $N(y_2, 1)$

why is $(*)$ improper ?

Brook's lemma

$$p(y_1, \dots y_n) = \frac{p(y_1 | y_2 \dots y_n)}{p(y_{10} | y_2 \dots y_n)} \cdot \frac{p(y_2 | y_{10}, y_3, \dots y_n)}{p(y_{20} | y_{10}, y_3, \dots y_n)} \dots \frac{p(y_n | y_{10}, \dots y_{n-1,0})}{p(y_{n0} | y_{10}, y_{20} \dots)} \cdot$$

$$p(y_{10}, \dots y_{n0})$$

where $y_0 = (y_{10}, \dots y_{n0})'$ is any fixed point in the support of $p(y_1 \dots y_n)$

Brook's lemma

conditional

constant

$$p(y_1, \dots y_n) = \frac{p(y_1 | y_2 \dots y_n)}{p(y_{10} | y_2 \dots y_n)} \cdot \frac{p(y_2 | y_{10}, y_3, \dots y_n)}{p(y_{20} | y_{10}, y_3, \dots y_n)} \dots \frac{p(y_n | y_{10}, \dots y_{n-1,0})}{p(y_{n0} | y_{10}, y_{20} \dots)} \cdot p(y_{10}, \dots y_{n0})$$

where $y_0 = (y_{10}, \dots y_{n0})'$ is any fixed point in the support of $p(y_1 \dots y_n)$

it has to integrate to 1 so we can find the constant.

Spatial considerations

- when the areal units span a large area one prefers to model exclusively the n corresponding conditional distributions

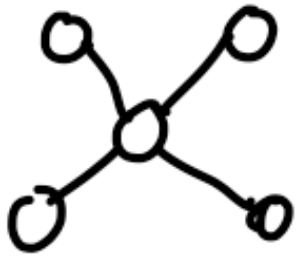
- we'd like to save on parameters, share info, make conditionals 'local'

Spatial considerations

In plain English:
Encoding all the spatial marginals is expensive, we want to define a model that uses **local information!**

Spatial considerations

$$p(y_i | y_j, j \neq i) = p(y_i | y_j, j \in \partial i)$$

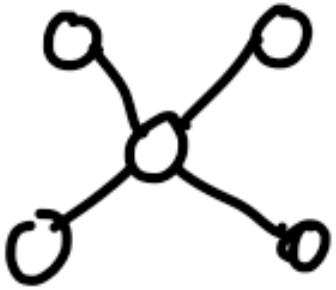


MRF

Clique

Spatial considerations

$$p(y_i | y_j, j \neq i) = p(y_i | y_j, j \in \partial i)$$

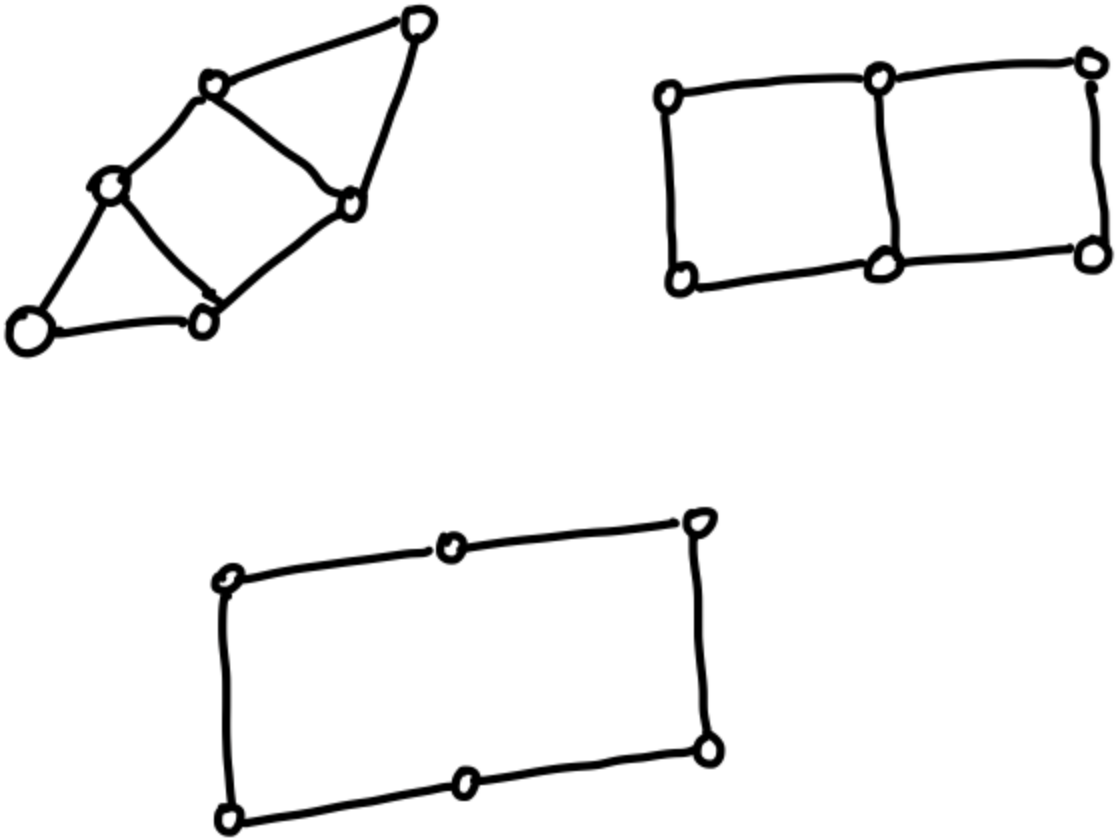


MRF

Clique

a set of cells
every element is
a neighbor of
any other element

what are the cliques here?



Spatial considerations

- MRF
- CLIQUE
- POTENTIAL FUNCTIONS – function of k arguments that is exchangeable in those arguments

Spatial considerations

- MRF
- CLIQUE
- POTENTIAL FUNCTIONS – function of k arguments that is exchangeable in those arguments

$$p(y_1, y_2, \ldots y_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c)$$

$$p(x,y,z) = p(y)\,p(x|y)\,p(z|y)$$

$$p(x,y)\;p(z|y)$$

$$p(x|y)\;p(z|y)$$

$$\phi_1(x,y) \qquad \phi_2(y,z)$$

$$\frac{f_1(x,y)}{z_1} \cdot \frac{f_2(y,z)}{z_2}$$

$$\phi(y_i, y_j) = (y_i - y_j)^2 \quad , i \sim j$$

Binary example

$$I(y_i = y_j) = y_i y_j + (1 - y_i)(1 - y_j)$$

Gibbs distribution

$p(y_1, \dots y_n)$ is a Gibbs distrib if it is a function of y_i only through potentials on cliques

$$p(y_1, \dots y_n) \propto \exp\left\{ \gamma \sum_K \sum_{\alpha \in M_K} \phi^{(k)}(y_{\alpha 1} \dots y_{\alpha k}) \right\}$$

$\phi^{(k)}$ is a potential of order K

$\alpha \in M_K$: subsets of order K

Recall : AR(1) process (Yarin's talk)

$$y_t = \delta + \phi y_{t-1} + w_t$$

$$w_t \sim N(0, \sigma^2) \quad |\phi| < 1$$

Then $$E(y_t) = \frac{\delta}{1-\phi}$$

$$Var(y_t) = \frac{\sigma^2}{1-\phi}$$

Recall : AR(1) process (Yarin's talk)

$$y_t = \delta + \phi y_{t-1} + w_t$$

We can show that

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \sim N\left(\frac{\delta}{1-\phi} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \frac{\sigma^2}{1-\phi} \begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & & \phi^{n-2} \\ \vdots & \vdots & & \vdots \\ \phi^{n-1} & \phi^{n-2} & & 1 \end{pmatrix} \right)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \sim N\left(\frac{\delta}{1-\phi} \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \frac{\sigma^2}{1-\phi} \begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & & \phi^{n-2} \\ \vdots & \vdots & & \vdots \\ \phi^{n-1} & \phi^{n-2} & & 1 \end{pmatrix} \right)$$

AR(1) is 1st order Markovian

$$f(y_1, y_2, \cdots y_n) = f(y_1) f(y_2|y_1) f(y_3|y_2, y_1) \cdots$$
$$= f(y_1) f(y_2|y_1) \cdots f(y_n|y_{n-1})$$

So an AR(1) process can be written as

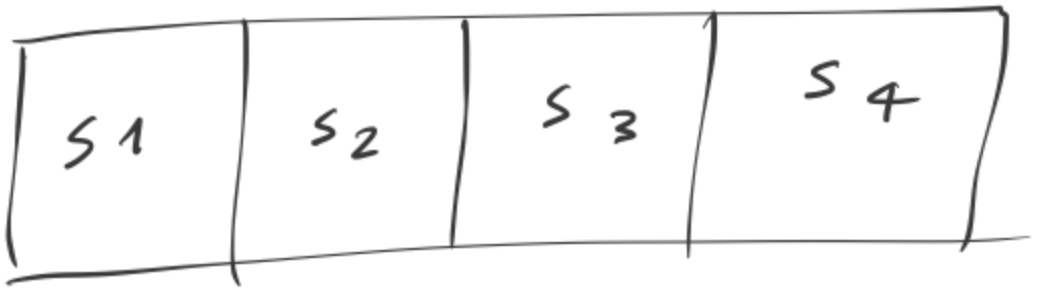
$$y_t = \delta + \phi y_{t-1} + w_t$$

or

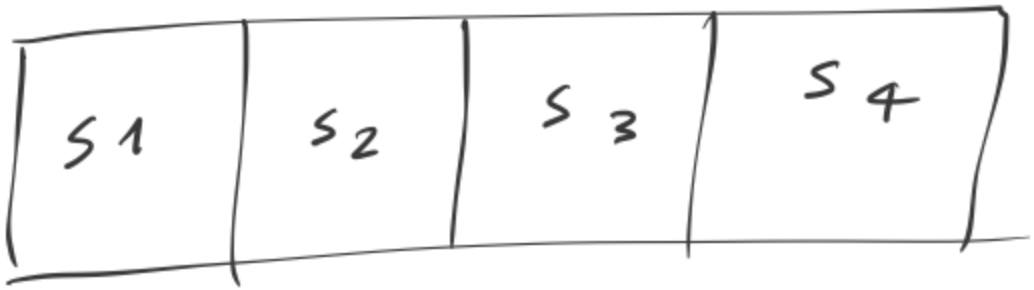
$$y_t | y_{t-1} \sim \mathcal{N}(\delta + \phi y_{t-1}, \sigma^2)$$

! They both give the same multivariate distrib for y

Let's do AR(1) in space!



Let's do AR(1) in space!



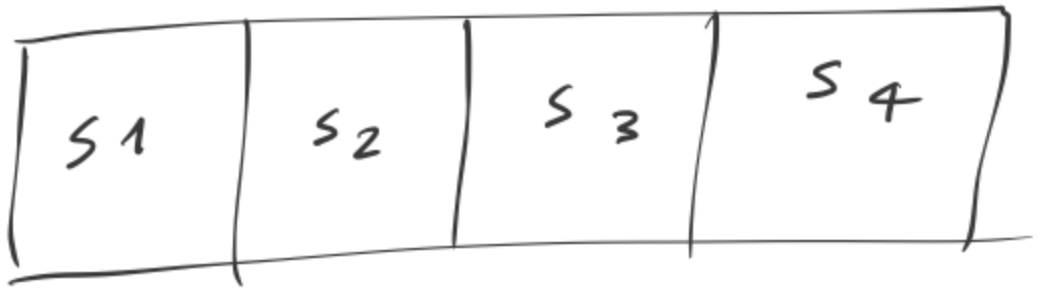
No unique ordering

$$f(y(s_1), y(s_2), y(s_3), y(s_4)) =$$

$$= f(y(s_1)) f(y(s_2) \mid y(s_1)) \cdot f(y(s_3) \mid y(s_1), y(s_2)) \ldots$$

$$= f(y(s_3)) f(y(s_4) \mid y(s_3)) \cdot f(y(s_2) \mid y(s_3)) \ldots$$

Let's do AR(1) in space!



Instead: we need to think about neighbors & cliques

- Simultaneous Autoregressive

$$y(s) = \delta + \phi \frac{1}{N(s)} \sum_{s' \in N(s)} y(s') + N(0, \sigma^2)$$

- Conditional Autoregressive (CAR)

$$y(s) | y_{-s} \sim N\left(\delta + \phi \frac{1}{N(s)} \sum_{s' \in N(s)} y(s'), \sigma^2\right)$$

- **Conditional Autoregressive (CAR)**

$$y(s) \mid y_{-s} \sim N\left(\delta + \phi \frac{1}{N(s)} \sum_{s' \in N(s)} y(s'), \sigma^2\right)$$

What's the joint distribution of the CAR model?

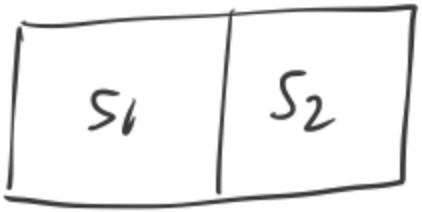
- **Conditional Autoregressive (CAR)**

$$y(s) \mid y_{-s} \sim N\left(\delta + \phi \frac{1}{N(s)} \sum_{s' \in N(s)} y(s'), \sigma^2\right)$$

What's the joint distribution of the CAR model?

Solution: Brook's Lemma

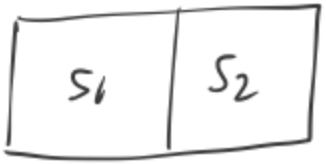
Brook's Lemma Applied to the CAR Model



$$y(s_1) \mid y(s_2) \sim N(\phi w_{12} y(s_2), \sigma^2)$$

$$y(s_2) \mid y(s_1) \sim N(\phi w_{21} y(s_1), \sigma^2)$$

Brook's Lemma Applied to the CAR model



$$y(s_1) \mid y(s_2) \sim N(\phi w_{12} y(s_2), \sigma^2)$$
$$y(s_2) \mid y(s_1) \sim N(\phi w_{21} y(s_1), \sigma^2)$$

Show that

$$p(y(s_1), y(s_2)) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\mu)\begin{pmatrix} 1 & -\phi w_{12} \\ -\phi w_{12} & 1 \end{pmatrix}(y-\mu)^t\right)$$

Brook's Lemma Applied to the CAR Model

This yields that for $y = (y(s_1), y(s_2))^t$

$$y \sim N\left(0, \sigma^2 (I - \phi W)^{-1}\right)$$

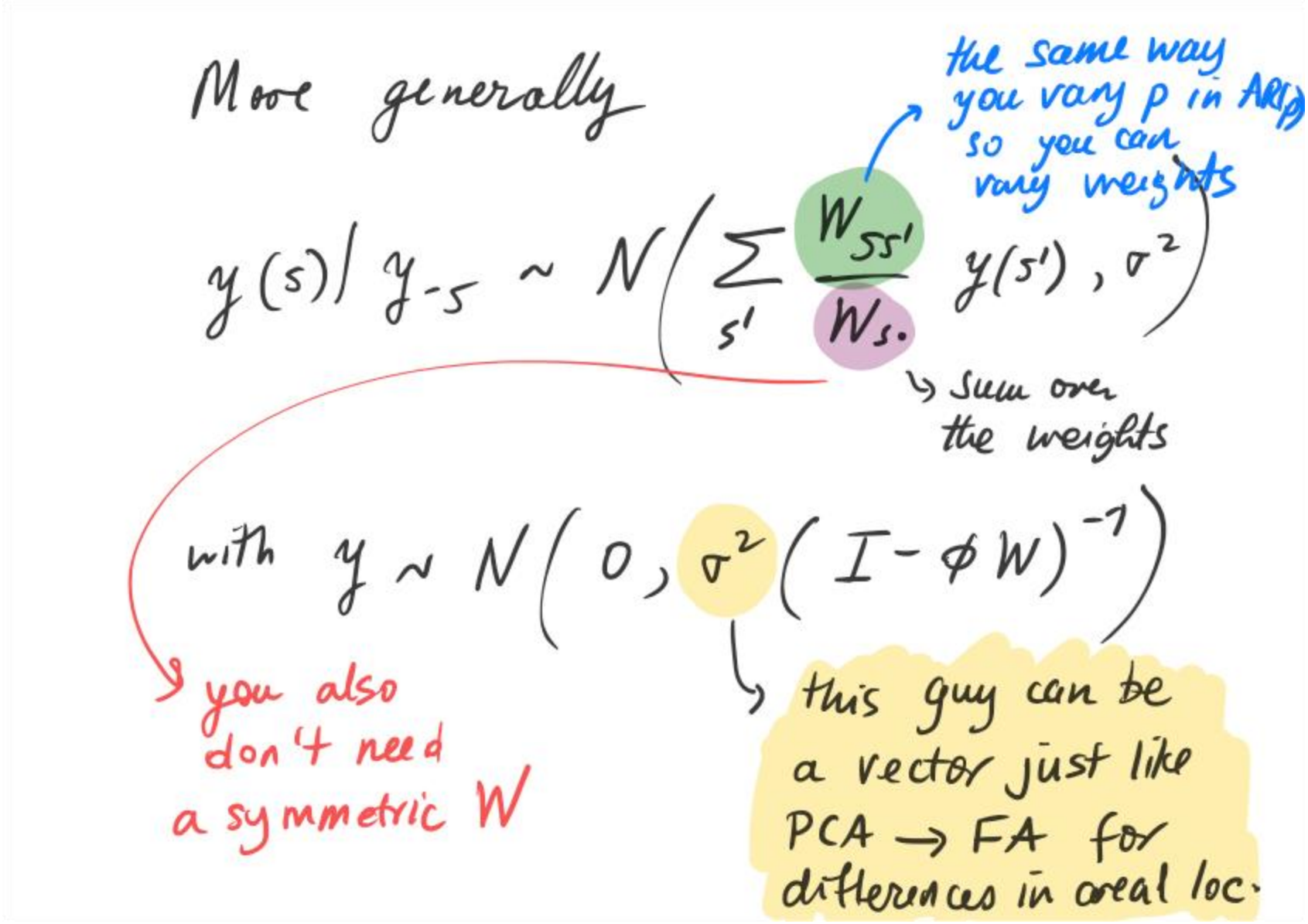
[with $$y(s) | y_{-s} \sim N\left(\phi \frac{1}{N(s)} \sum_{s' \in N(s)} y(s'), \sigma^2\right)$$]

More generally

$$y(s) \mid y_{-s} \sim N\left(\sum_{s'} \frac{W_{ss'}}{W_{s\cdot}} y(s'), \sigma^2 \right)$$

↳ sum over the weights

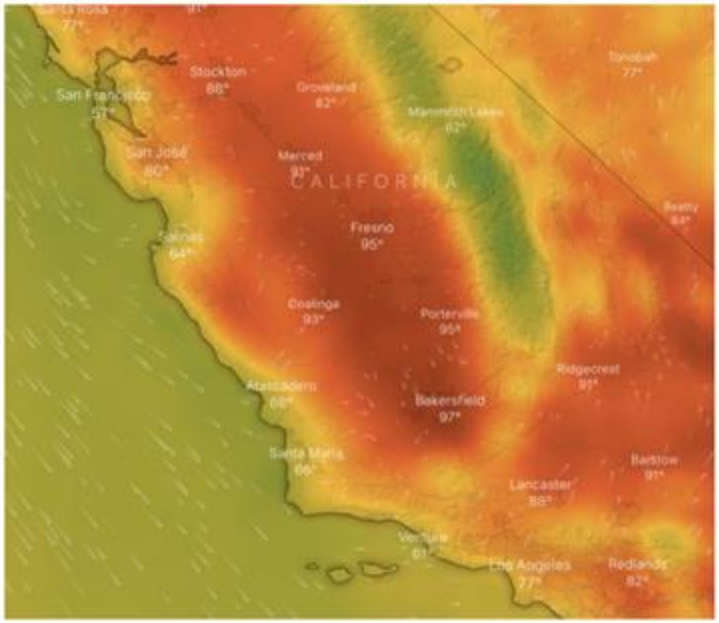
with $y \sim N\left(0, \sigma^2 (I - \phi W)^{-1}\right)$


More generally
$y(s) \mid y_{-s} \sim N\left(\sum_{s'} \frac{W_{ss'}}{W_{s\cdot}} y(s'), \sigma^2\right)$
the same way you vary ρ in AR(p) so you can vary weights
↳ sum over the weights
with $y \sim N\left(0, \sigma^2 (I - \phi W)^{-1}\right)$
you also don't need a symmetric W
this guy can be a vector just like PCA → FA for differences in areal loc.

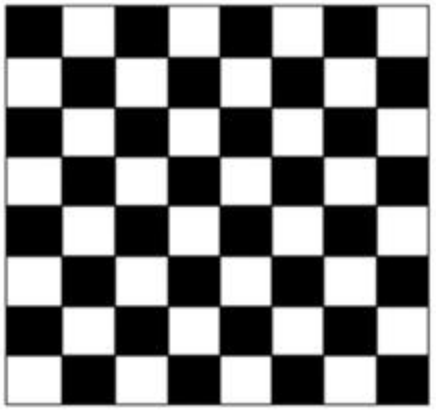
Exploratory Data Analysis

Tobler's first law of geography: Everything is related to everything else.
But near things are more related than distant things.

Positive

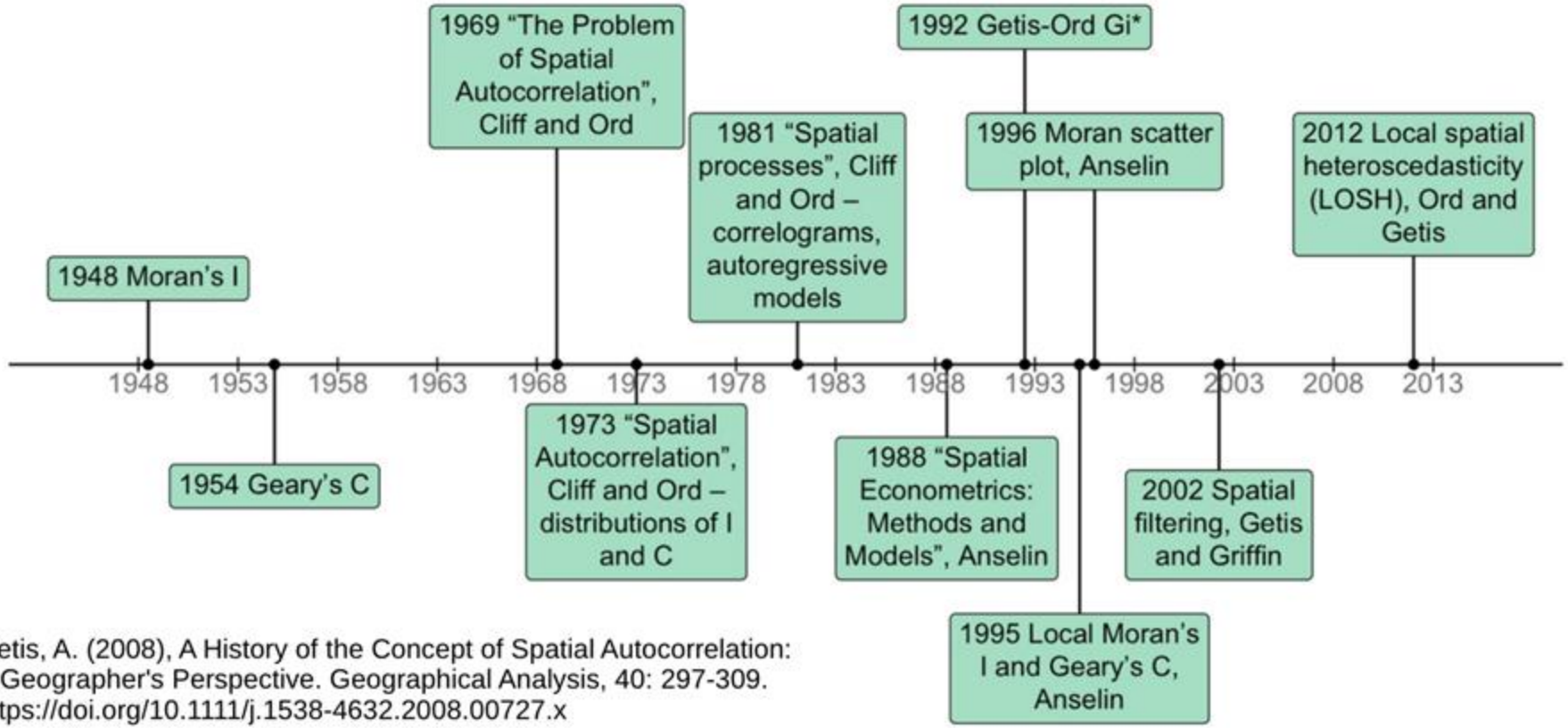


Negative



Slide credit: Lambda Moses

Exploratory Data Analysis



Getis, A. (2008), A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective. Geographical Analysis, 40: 297-309. https://doi.org/10.1111/j.1538-4632.2008.00727.x

Slide credit: Lambda Moses

Exploratory Data Analysis: Moran's I

Pearson correlation: coexpression of 2 genes

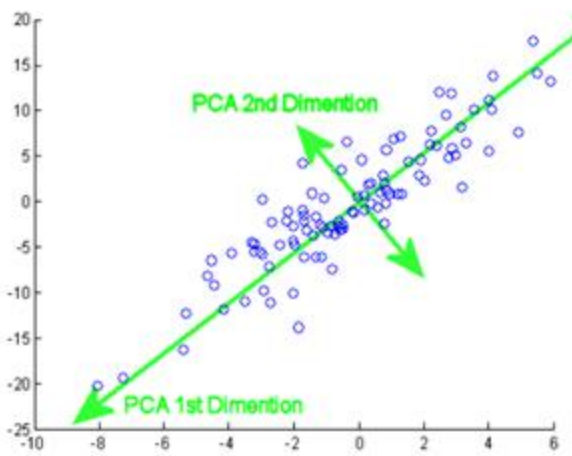
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

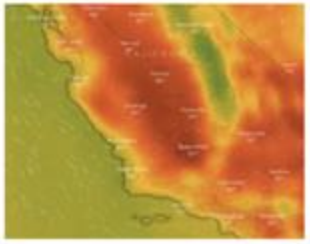
Moran's I: spatial autocorrelation of 1 gene

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

w_{ij}: spatial weights indicating neighborhood

PCA: correlation among numerous genes





Slide credit: Lambda Moses

Modeling Spatial Variability in Gene Space

In Context

- (gene) SpatialDE, nnSVG
- (gene programs)Spatial NMF

Spatial Variable Gene Selection

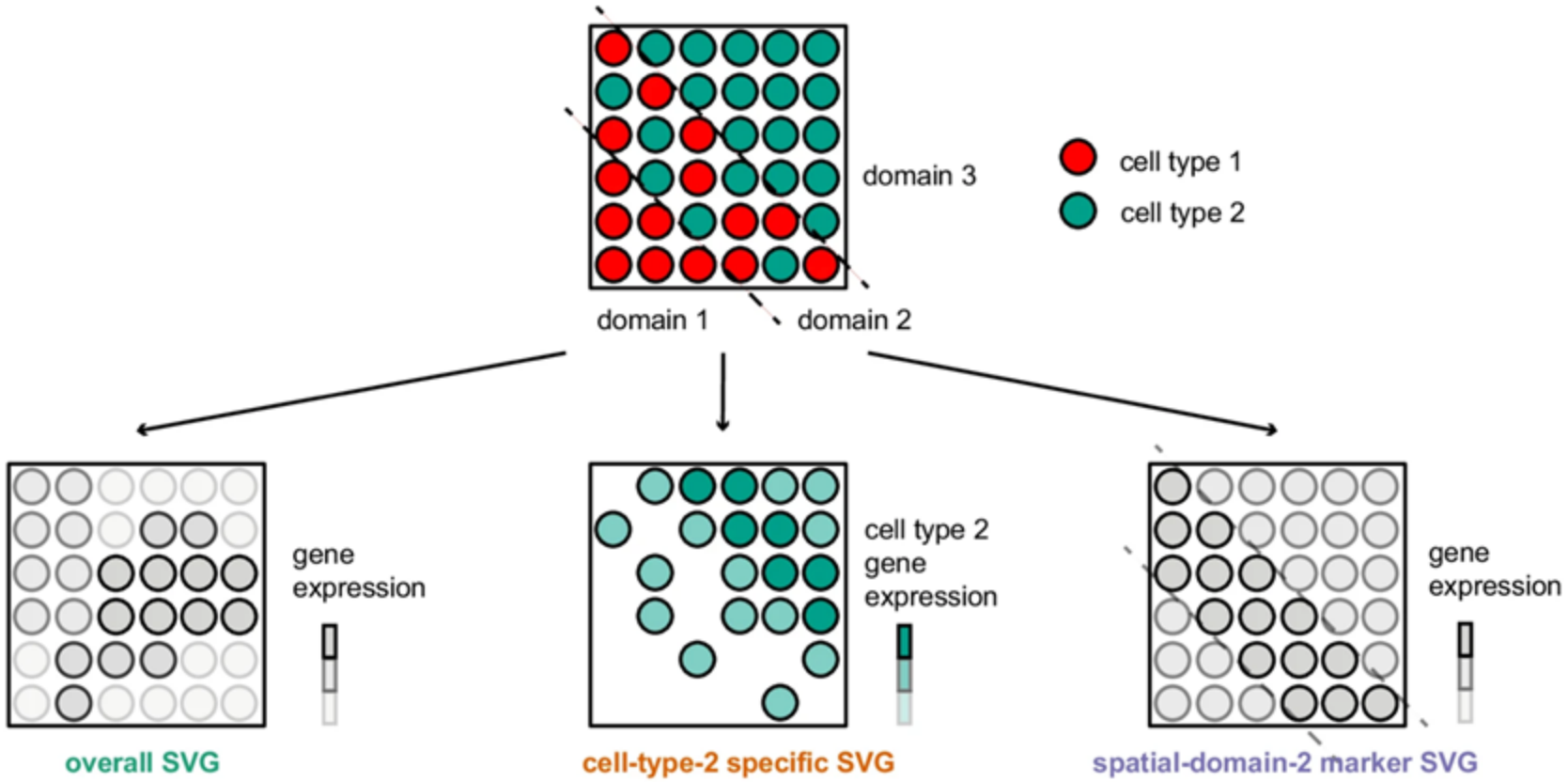
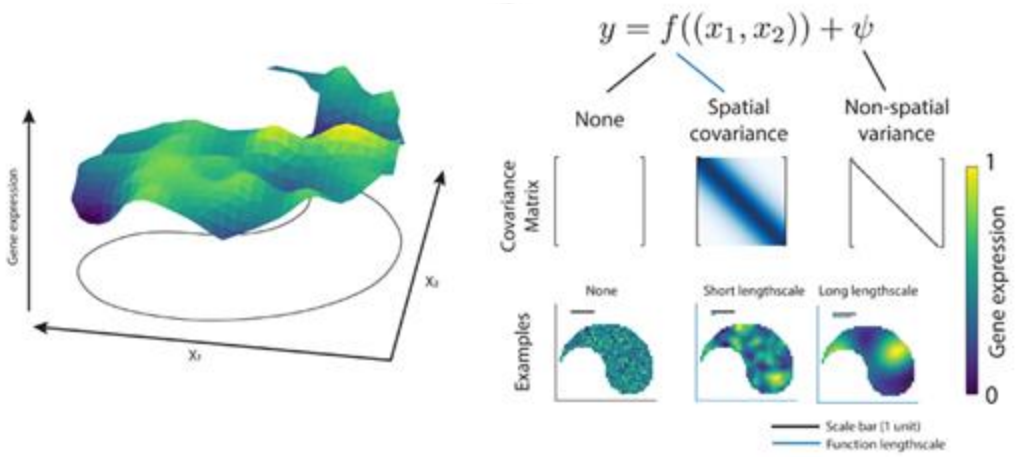
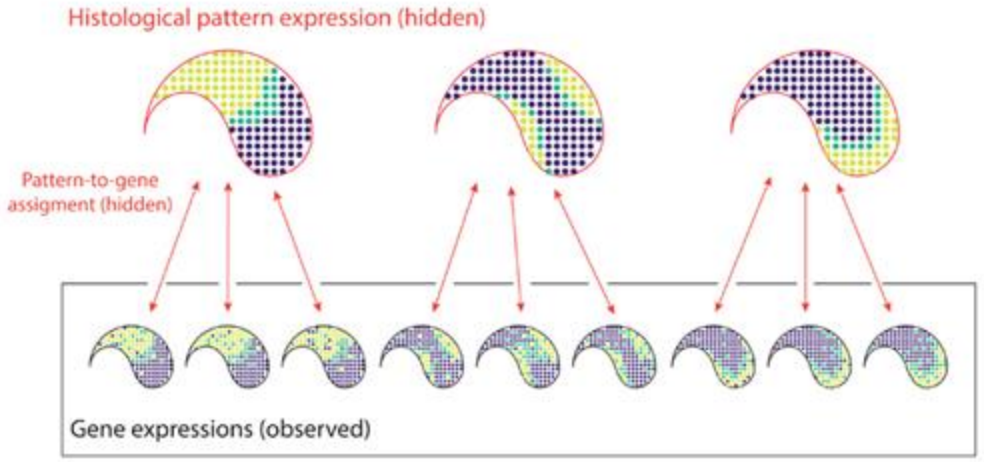


Figure from: Yan, G., Hua, S.H. & Li, J.J. Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data. (2025)

Spatial Variable Gene Selection: SpatialDE





Adapted from Svensson. *et al.* SpatialDE: identification of spatially variable genes (2018).

Spatial Variable Gene Selection: SpatialDE

x_1^n, x_2^n coordinates

$$y_n = f((x_1^n, x_2^n)) + \psi_n$$

Spatial covariance

non-spatial covariance

Null model

$$\varepsilon \sim N(0, \sigma^2 \cdot I)$$

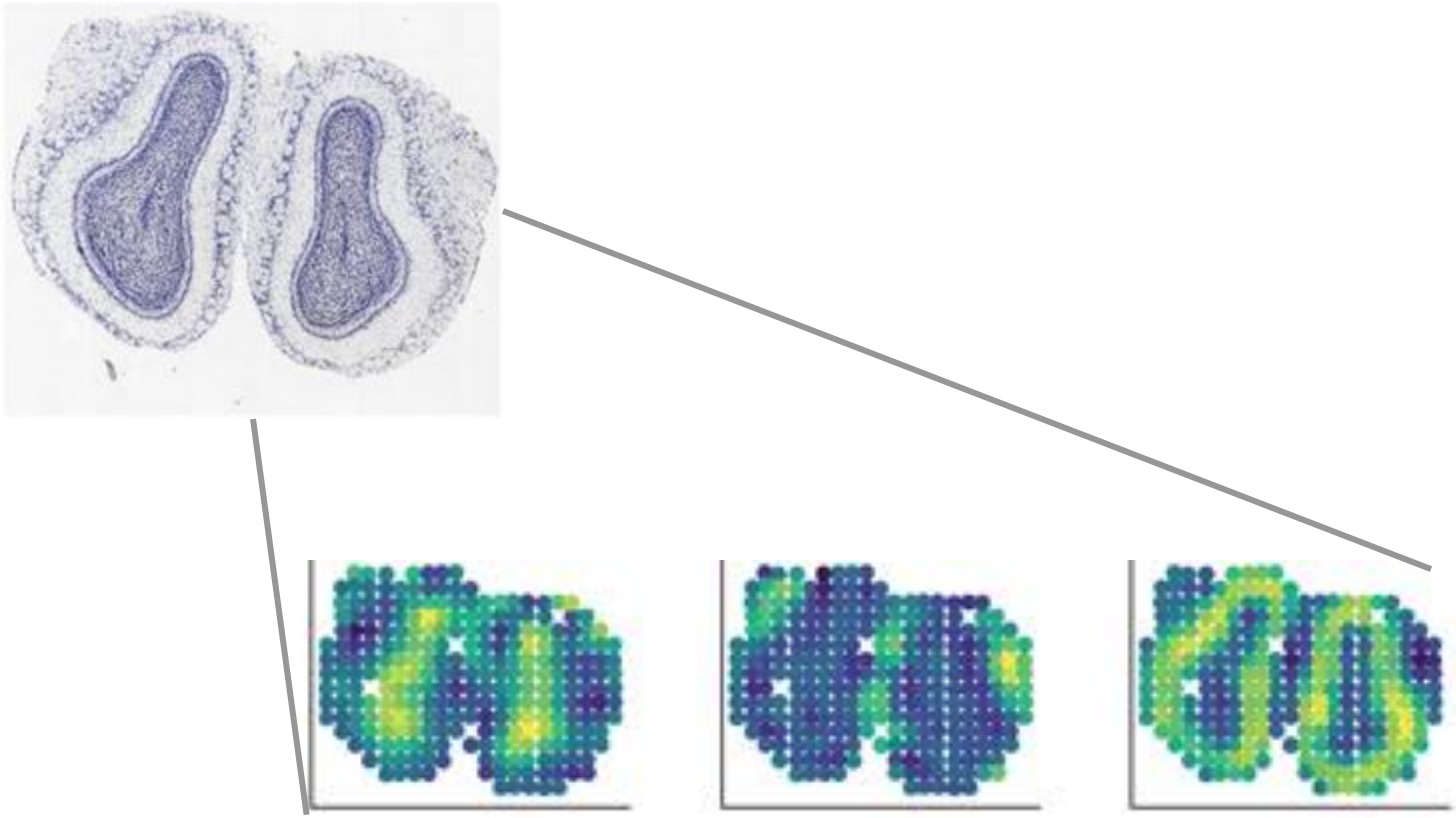
$$y_n = \mu_n + \varepsilon_n$$

$$\varepsilon_n \sim N(0, \sigma_s^2 \Sigma + \sigma_s^2 \cdot \delta \cdot I)$$

$$\Sigma_{ij} = \exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2}\right)$$

i, j, n indices of cells

Spatial Variable Gene Selection: SpatialDE



Adapted from Svensson. *et al.* SpatialDE: identification of spatially variable genes (2018).

Spatial Variable Gene Selection: nnSVG

Motivation: why should all the genes share the same scale?

$$y \sim N\left(X\beta, \tilde{\Sigma}(\theta, \tau^2)\right)$$

$y = (y_1, y_2, \cdots y_N)$ gene expression value for a gene g

at spatial locations $s = (s_1, s_2, \cdots s_N)$

$$\tilde{\Sigma}(\theta, \tau^2) = \Sigma(\theta, \tau^2) = C(\theta) + \tau^2 I$$

↳ spatially correlated info

$$C_{ij}(\theta) = \sigma^2 \exp\left(\frac{-\|s_i - s_j\|}{\ell}\right)$$

Based on Weber, L.M., Saha, A., Datta, A. *et al.* nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. (2023).

Spatial Variable Gene Selection: nnSVG

$$y^g \sim N\left(X\beta, \tilde{\Sigma}(\theta^g, \tau_g^2)\right)$$

$$y^g = (y_1^g, y_2^g, \cdots y_N^g)$$

gene expression value for a gene g

at spatial locations $S = (s_1, s_2, \cdots s_N)$

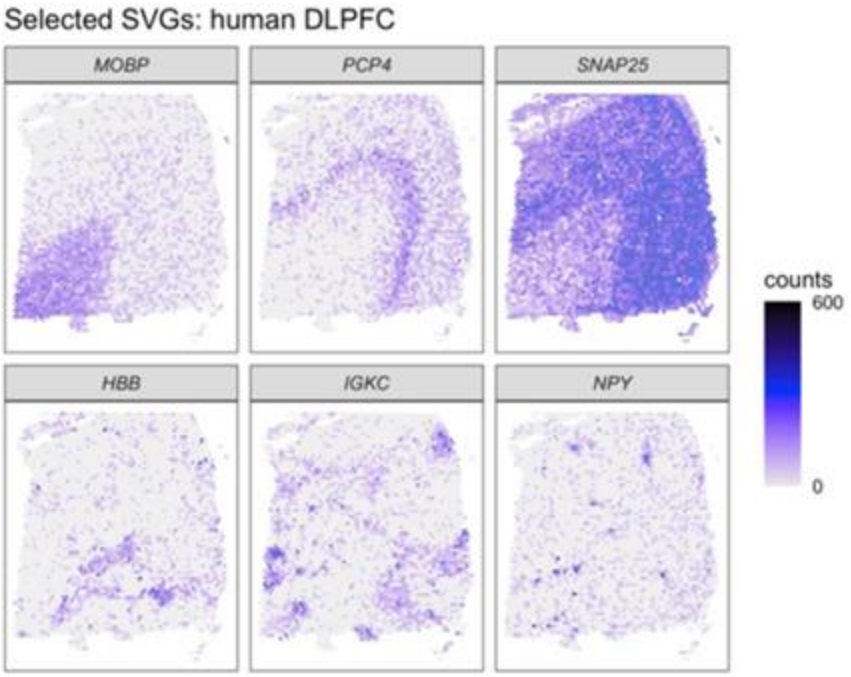
$$\tilde{\Sigma}(\theta^g, \tau_g^2) = \Sigma(\theta^g, \tau_g^2) = C(\theta_g) + \tau_g^2 I$$

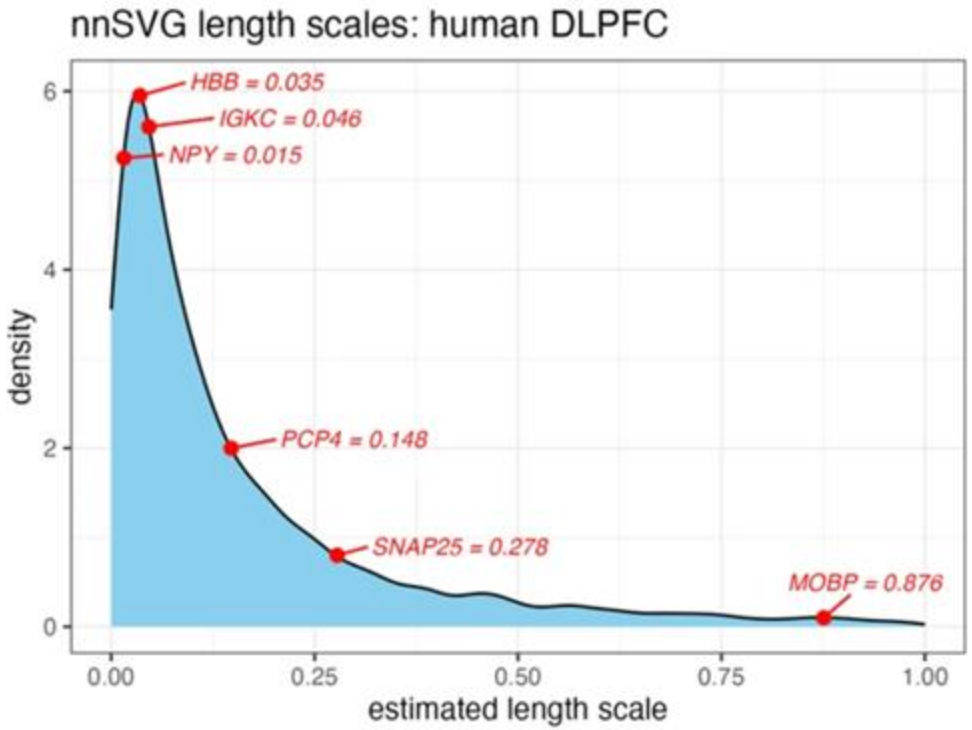
↳ spatially correlated info

$$C_{ij}(\theta_g) = \sigma_g^2 \exp\left(\frac{-\|s_i - s_j\|}{\ell_g}\right)$$

each gene has its scale!

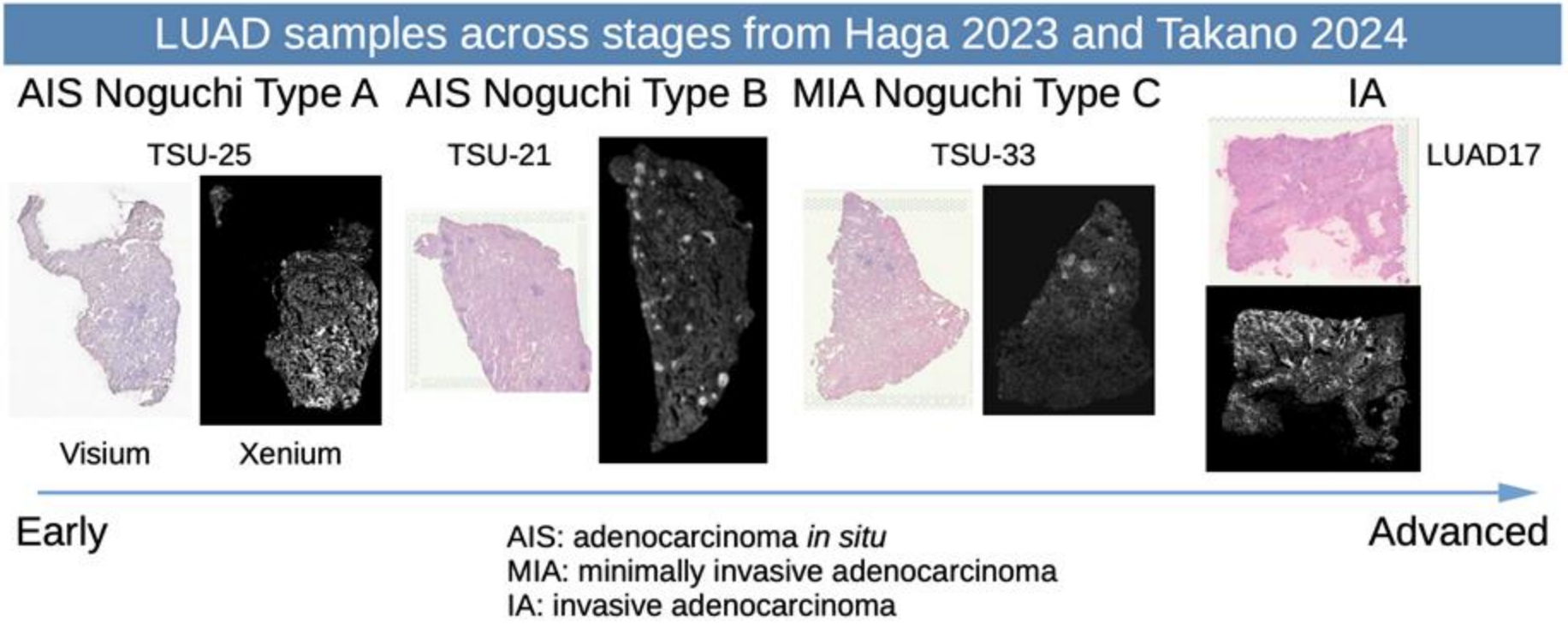
Spatial Variable Gene Selection: nnSVG



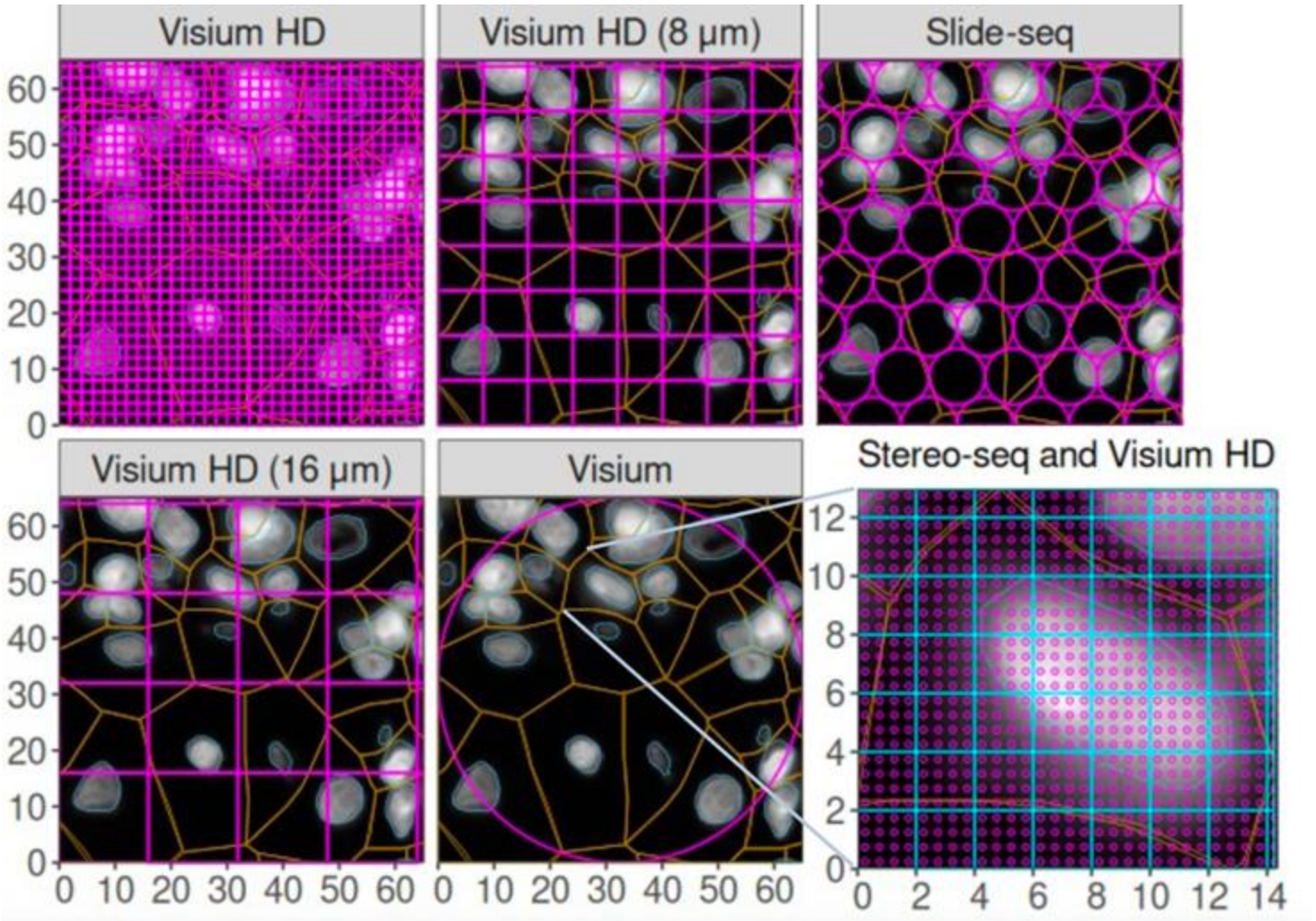


Adapted from Weber, L.M., Saha, A., Datta, A. *et al.* nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes. (2023).

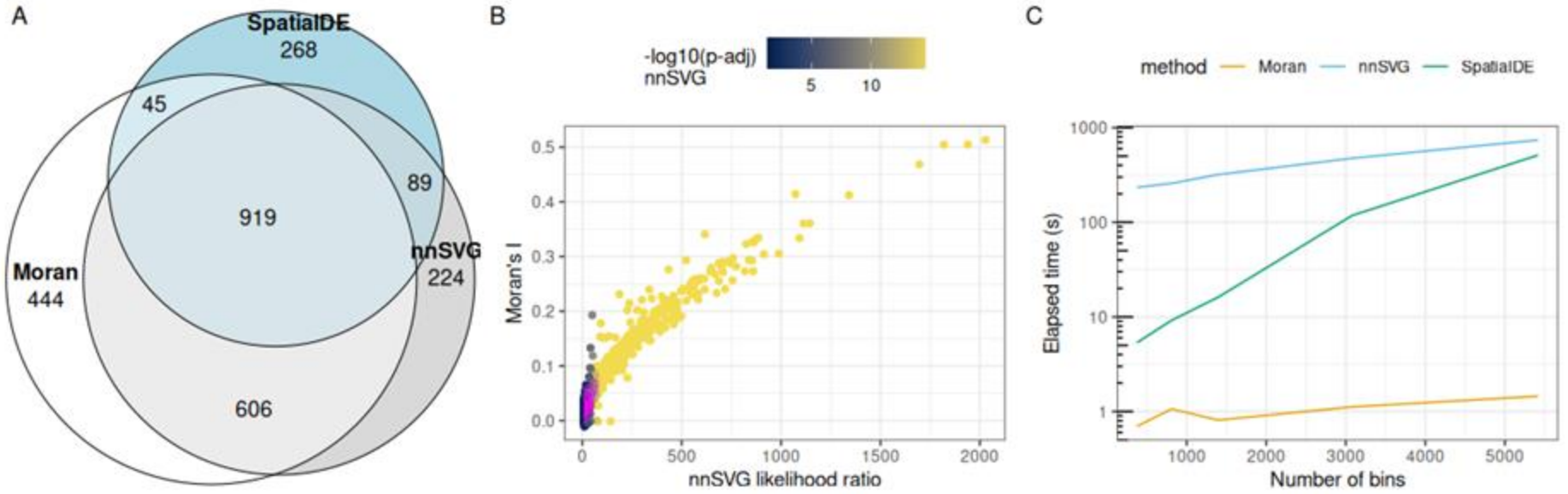
SVG Case study: a problem of scale



SVG Case study: a problem of scale



SVG Case study: a problem of scale



Spatially Variable Gene Expression Programs

Spatial Process Factorization

outcome (gene expression) $Y \in \mathbb{R}^{N \times G}$

Spatial coordinates $X \in \mathbb{R}^{N \times D}$

$$y_{ig} \sim N(\mu_{ig}, \sigma_g^2)$$

$$\mu_{ig} = \sum_{l=1}^{L} w_{gl} f_{il}$$

$$f_{il} = f_l(x_i) \sim GP(\mu_l(x_i), k_l(x_i, X))$$

based on Spatial Nonnegative Matrix Factorization, Townes & Engelhardt (2023)

Spatially Variable Gene Expression Programs

outcome (gene expression) $Y \in \mathbb{R}^{N \times G}$

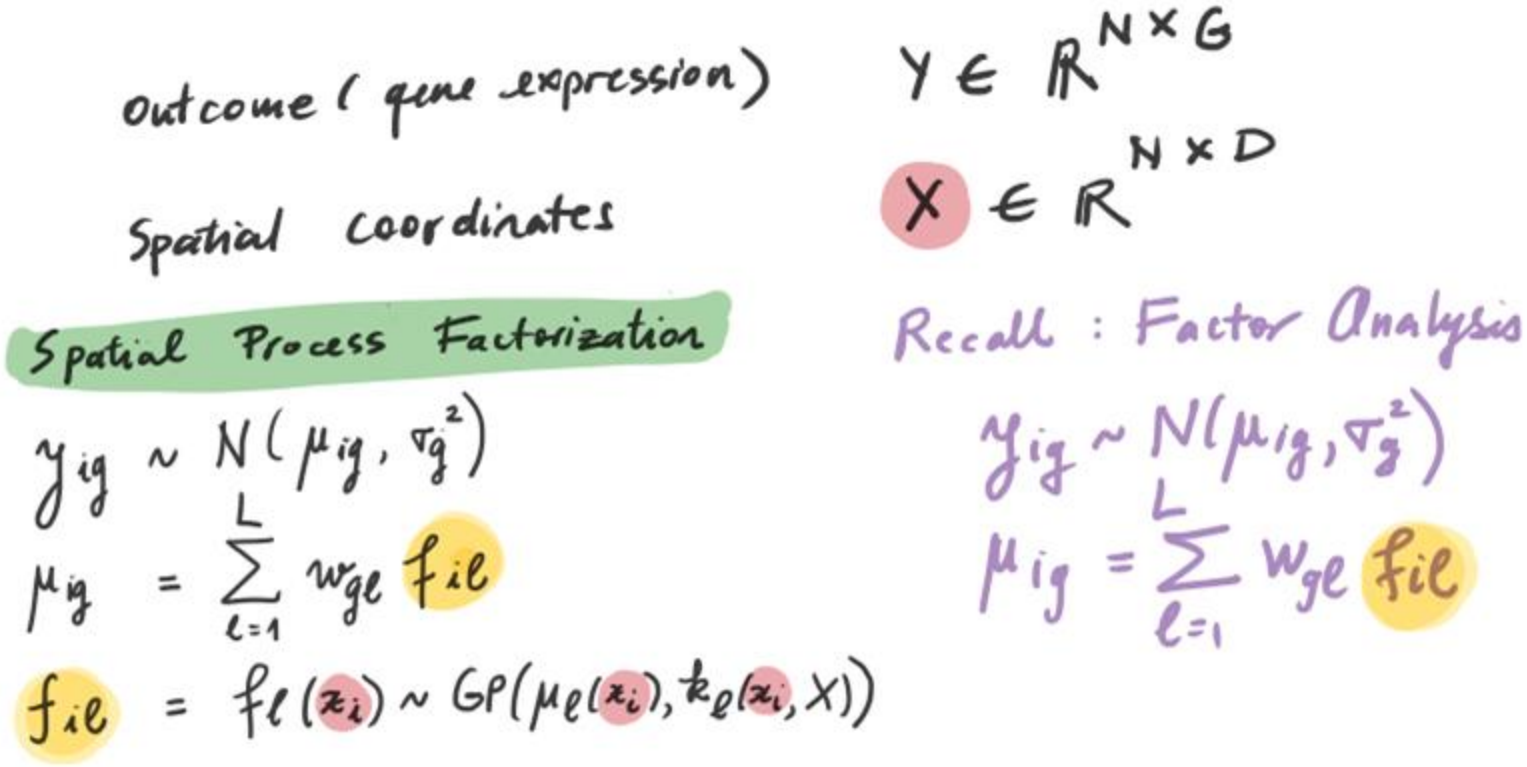
Spatial coordinates $X \in \mathbb{R}^{N \times D}$

$$y_{ig} \sim N(\mu_{ig}, \sigma_g^2)$$

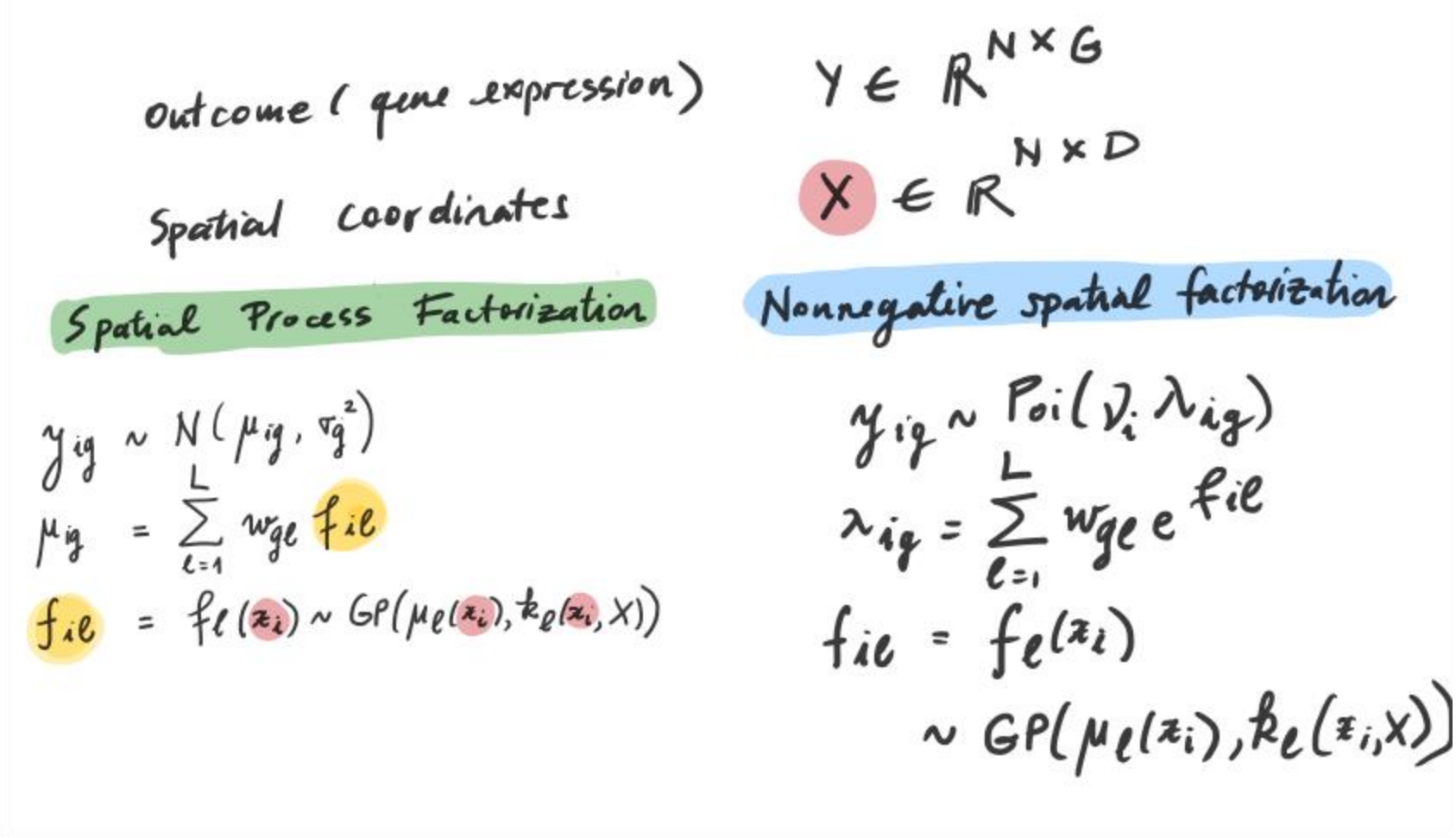
$$\mu_{ig} = \sum_{l=1}^{L} w_{gl} f_{il}$$

$$f_{il} = f_l(x_i) \sim GP(\mu_l(x_i), k_l(x_i, X))$$

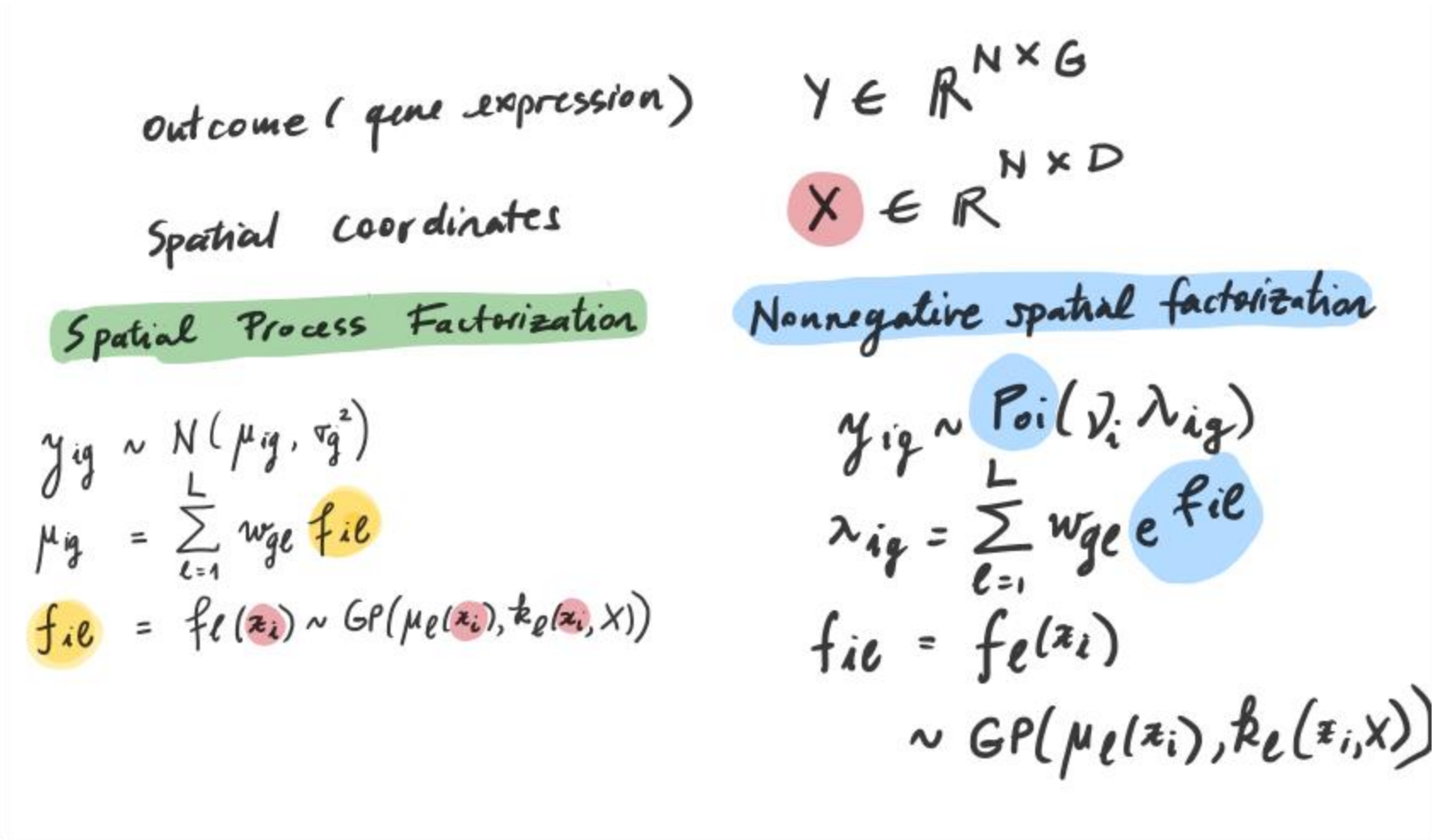
Spatially Variable Gene Expression Programs



Spatially Variable Gene Expression Programs



Spatially Variable Gene Expression Programs



Spatially Variable Gene Expression Programs

outcome (gene expression) $Y \in \mathbb{R}^{N \times G}$

Spatial coordinates $X \in \mathbb{R}^{N \times D}$

Nonnegative spatial factorization hybrid

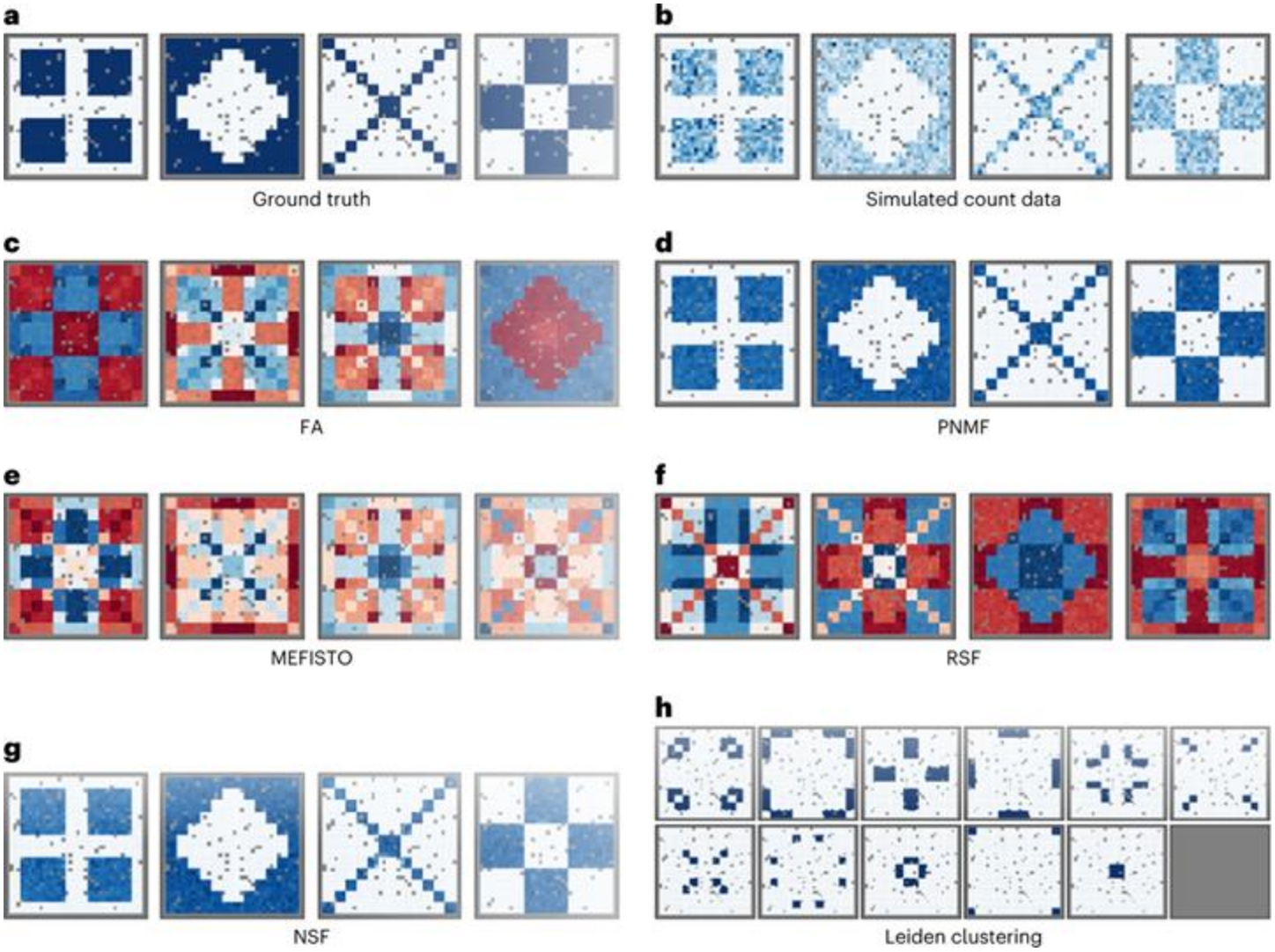
$$y_{ig} \sim \text{Poi}(v_i \lambda_{ig})$$

$$\lambda_{ig} = \sum_{\ell=1}^{L} w_{g\ell} e^{f_{i\ell}} + \sum_{\ell=L+1}^{L+L_{ns}} v_{g\ell} e^{h_{i\ell}}$$

$$f_{i\ell} = f_\ell(x_i) \sim GP(\mu_\ell(x_i), k_\ell(x_i, X))$$

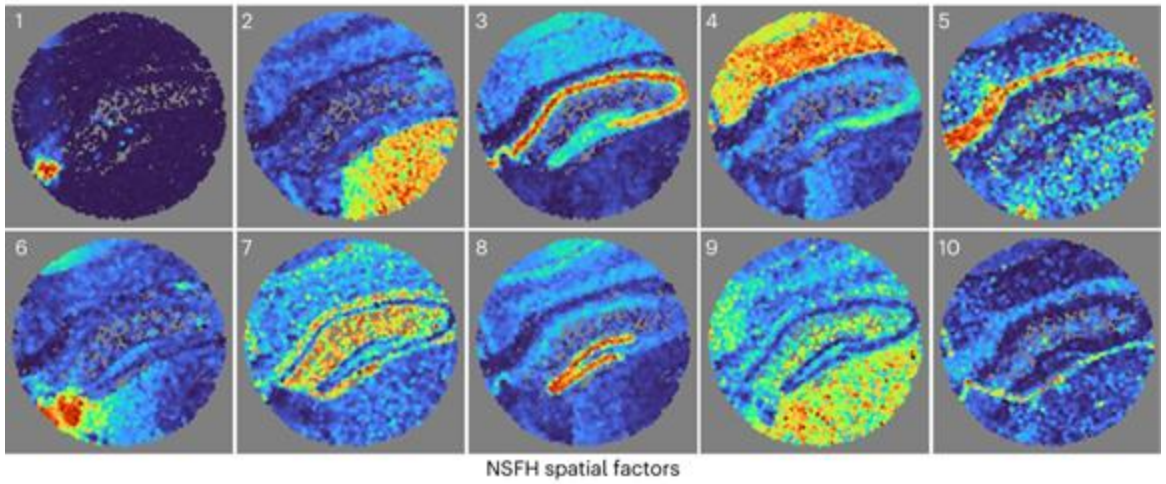
$$h_{i\ell} \sim N(m_\ell, s_\ell^2)$$

Spatially Variable Gene Expression Programs

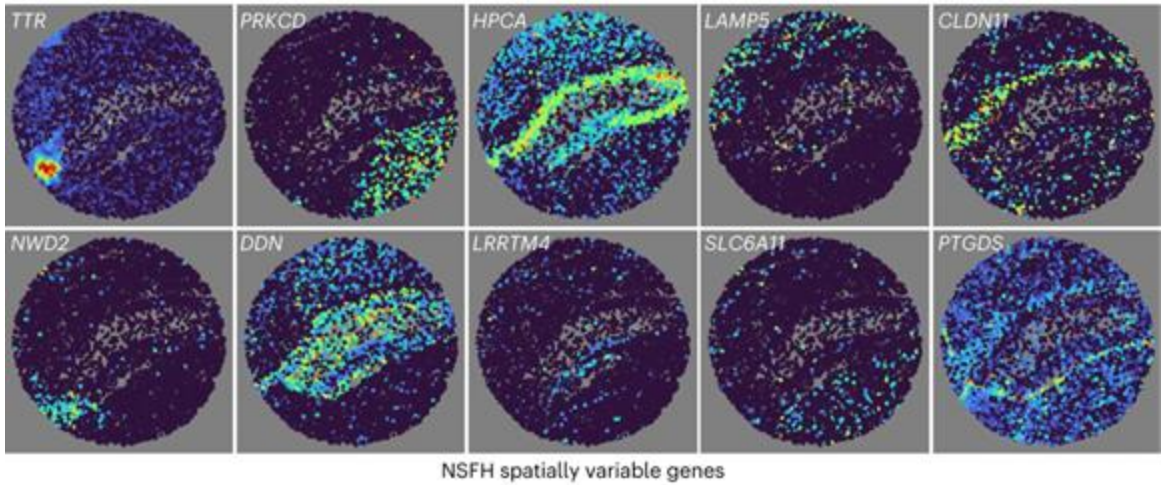


Spatial Nonnegative Matrix Factorization, Townes & Engelhardt (2023)

Spatially Variable Gene Expression Programs



NSFH spatial factors



NSFH spatially variable genes

Slide-seqV2 mouse hippocampus gene expression data

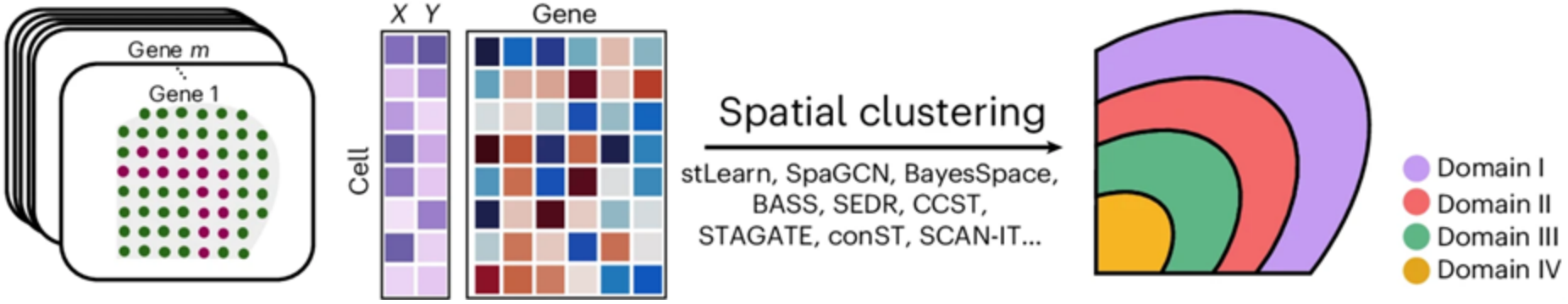
Spatial Nonnegative Matrix Factorization, Townes & Engelhardt (2023)

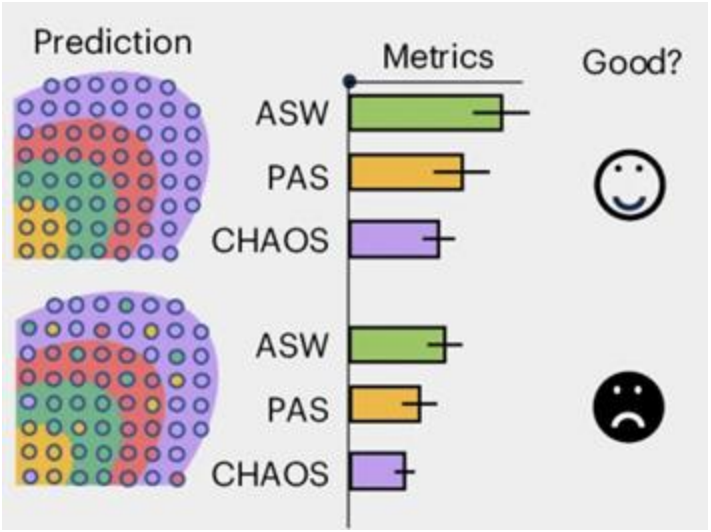
Modelling spatial neighbourhoods

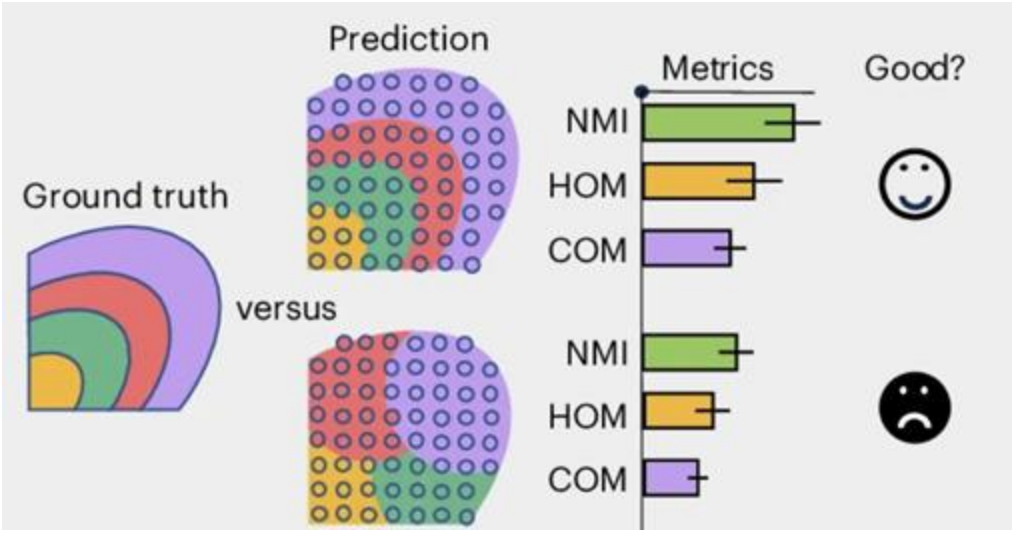
Spatial clustering via Graph Neural Networks (GNNs)

- **GraphST**

Spatial Clustering

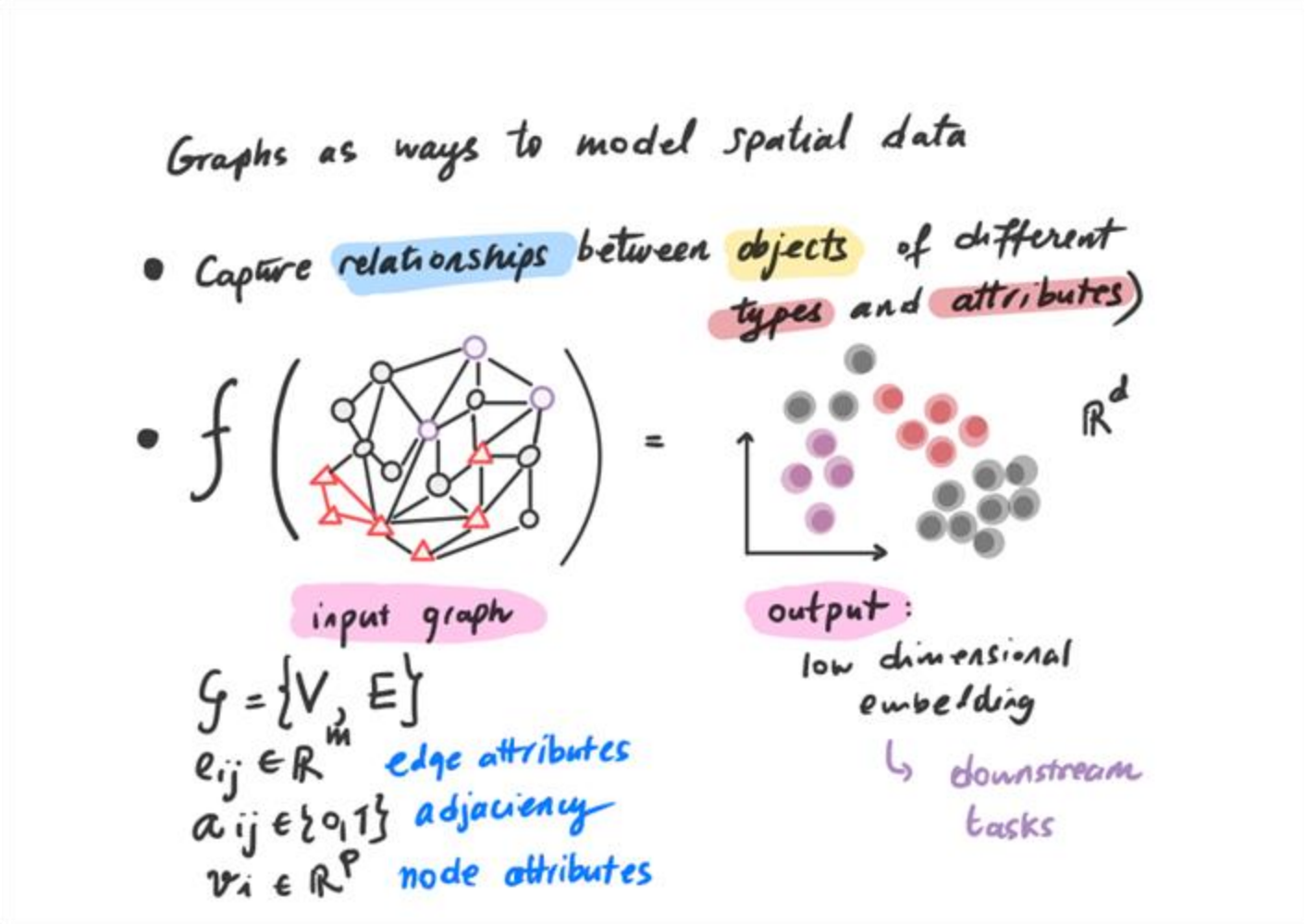






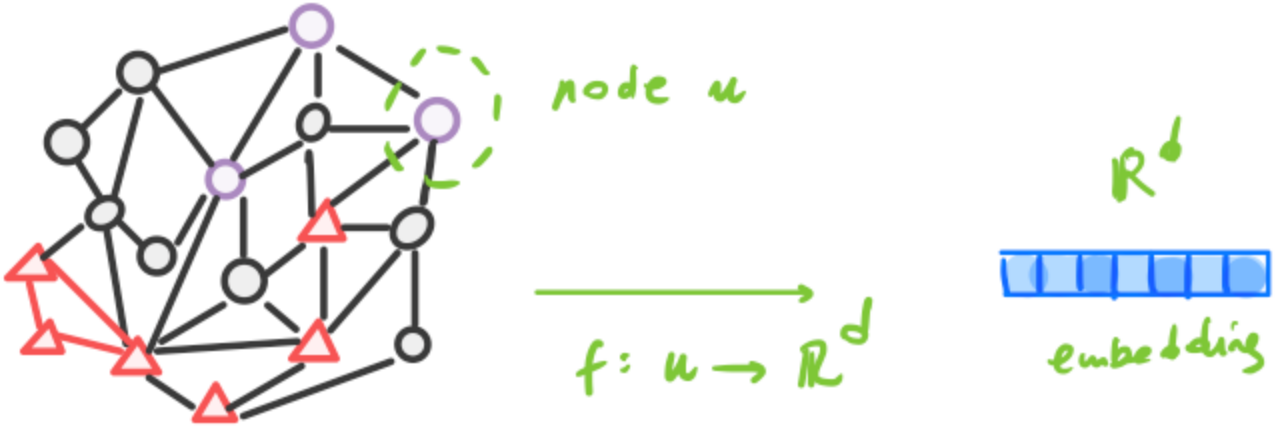
From: Yuan, Z., Zhao, F., Lin, S. *et al.* Benchmarking spatial clustering methods with spatially resolved transcriptomics data.2024

GraphST and GNNs for Spatial Transcriptomics

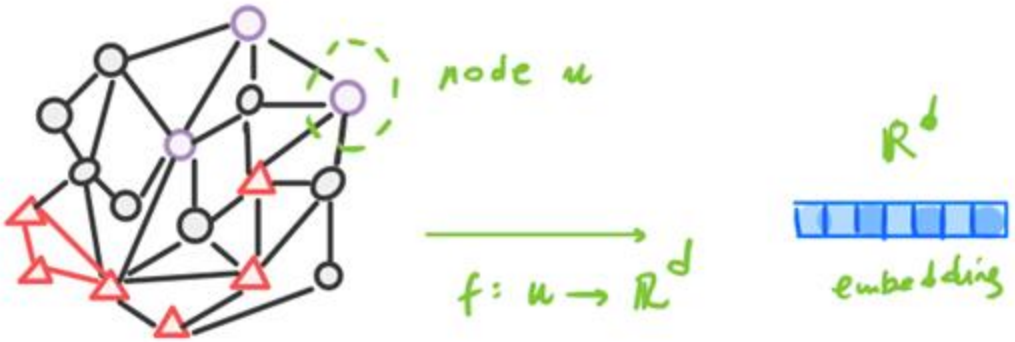


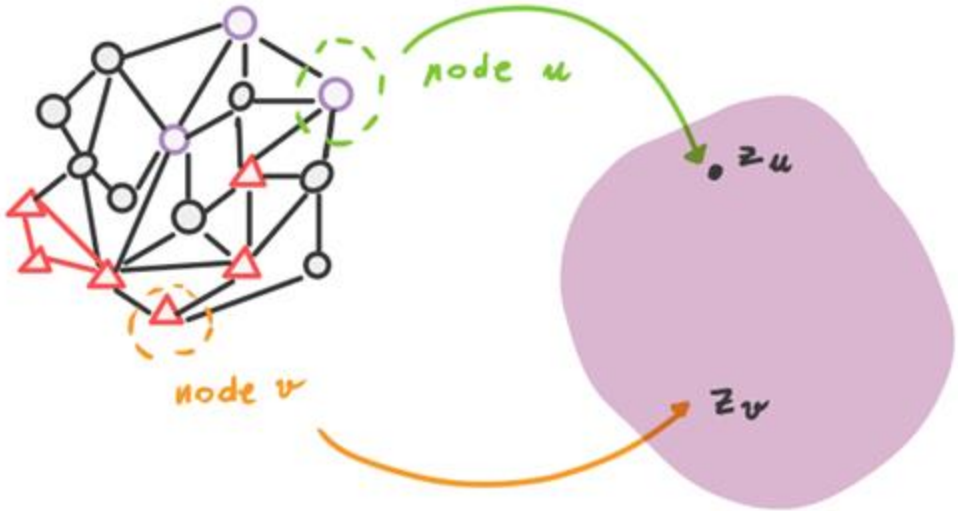
based on: Long, Y., Ang, K.S., Li, M. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. (2023).

GraphST and GNNs for Spatial Transcriptomics

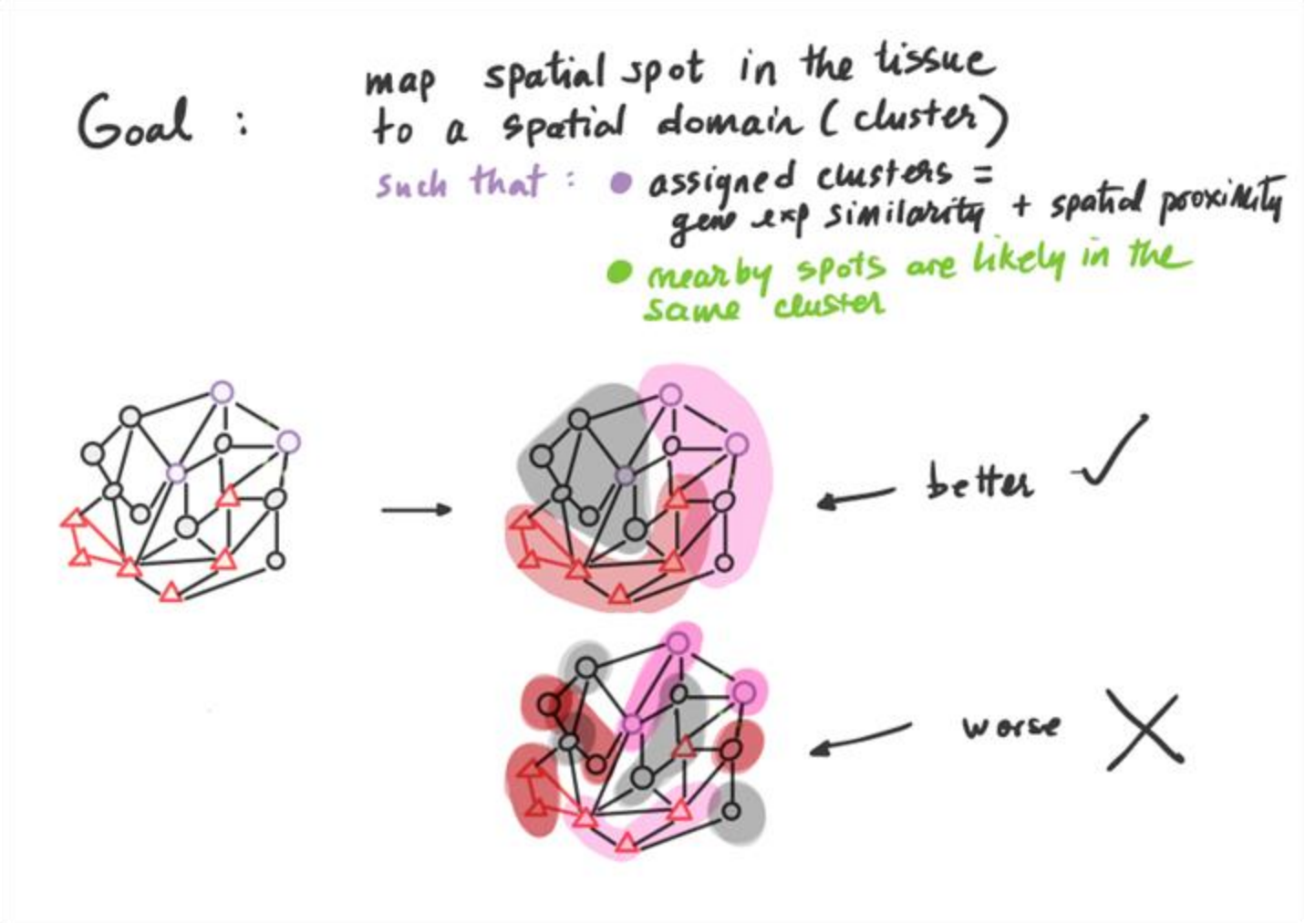


GraphST and GNNs for Spatial Transcriptomics





GraphST and GNNs for Spatial Transcriptomics



GraphST and GNNs for Spatial Transcriptomics

Input

$X \in \mathbb{R}^{N \times G}$

N spots, G genes

$S \in \mathbb{R}^{N \times 2}$

spatial coordinates

Step 1

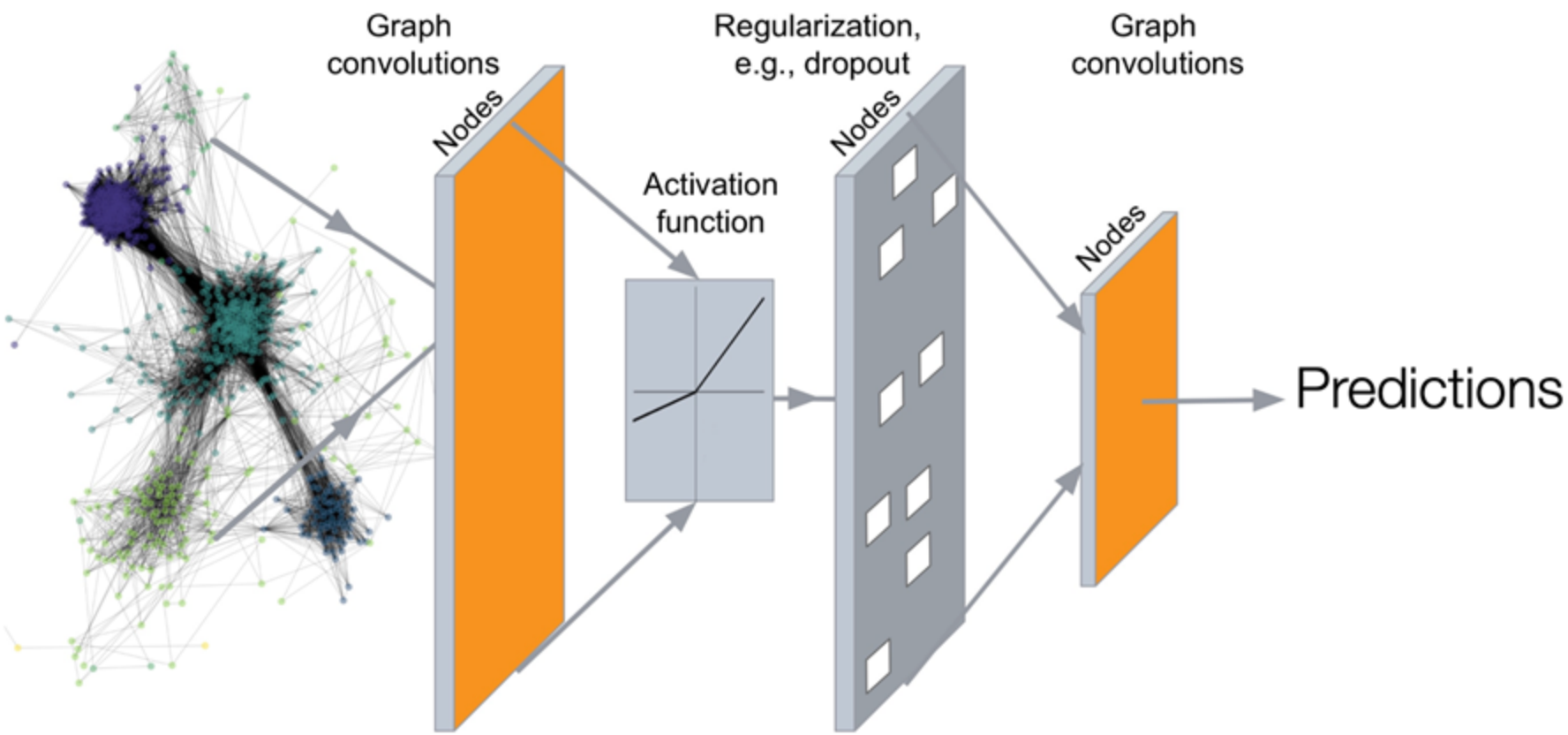
construct a spatial graph (k nearest neigh)

$\mathcal{G} = (V, E)$

$A \in \{0,1\}^{N \times N}$

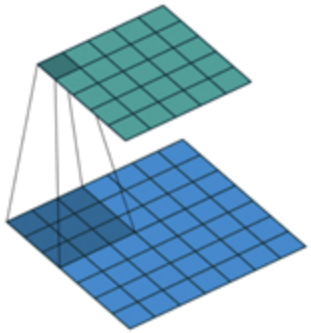
$d_{ij} = \|x_i - x_j\|_2^2$

GraphST and GNNs for Spatial Transcriptomics

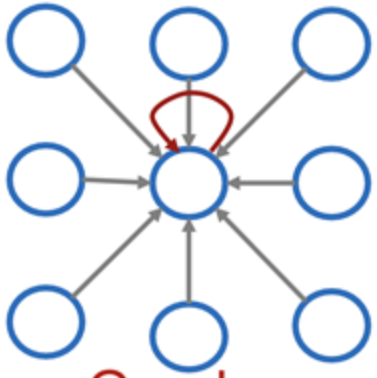


based on https://web.stanford.edu/class/cs224w/

Image Convolutions

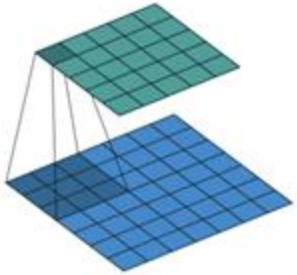


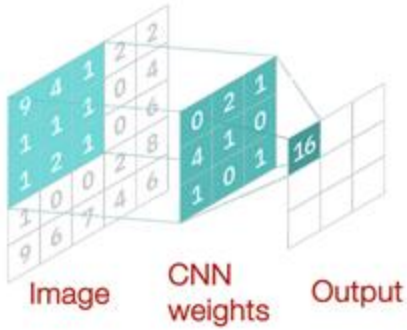
Image



Graph

Convolutional neural network (CNN) layer with 3x3 filter:

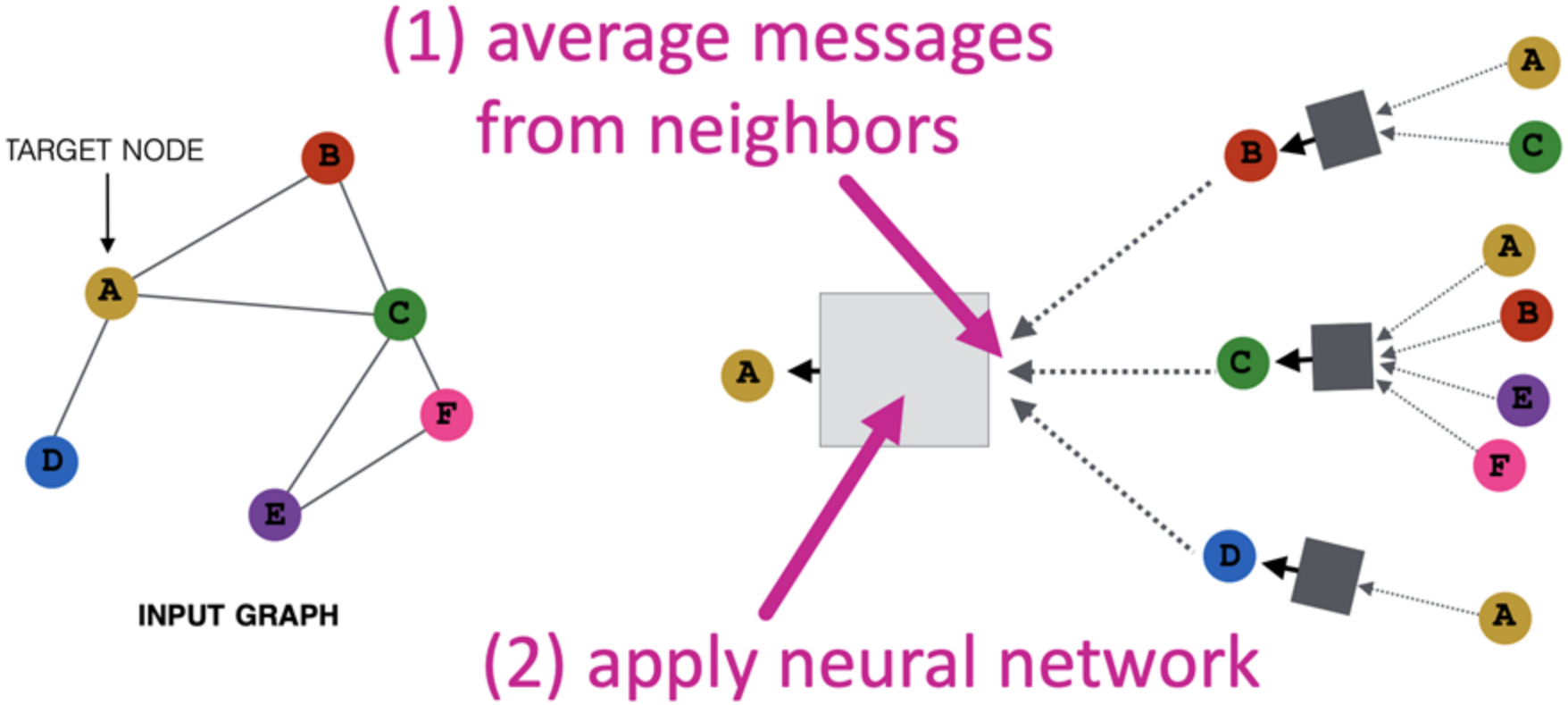




CNN formulation: $h_v^{(l+1)} = \sigma(\sum_{u \in N(v) \cup \{v\}} W_l^u h_u^{(l)}), \quad \forall l \in \{0, \ldots, L-1\}$

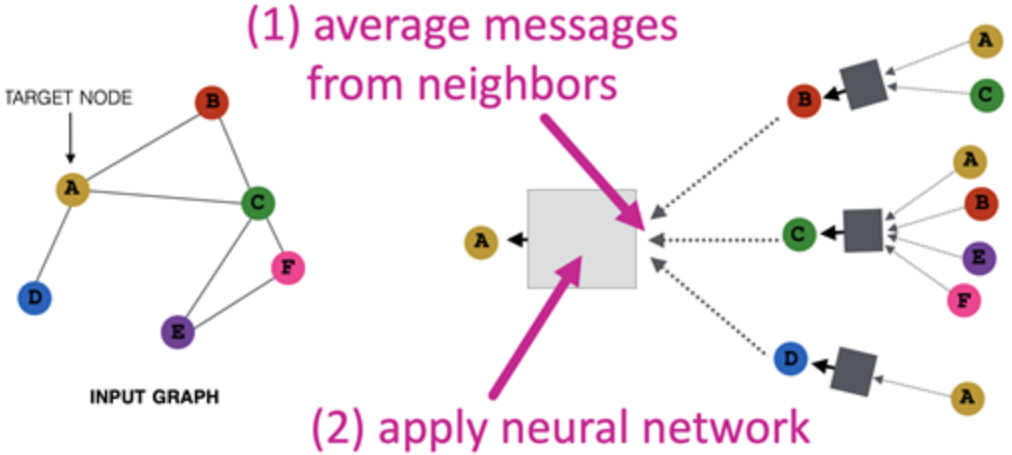
based on https://web.stanford.edu/class/cs224w/

GraphST and GNNs for Spatial Transcriptomics



based on https://web.stanford.edu/class/cs224w/

GraphST and GNNs for Spatial Transcriptomics



Trainable weight matrices
(i.e., what we learn)

$$\mathrm{h}_v^{(0)} = \mathrm{x}_v$$

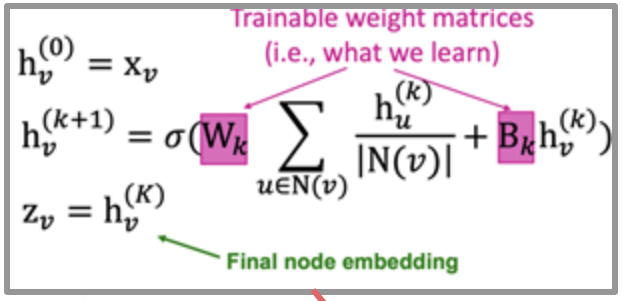
$$\mathrm{h}_v^{(k+1)} = \sigma\left(\mathrm{W}_k \sum_{u \in \mathrm{N}(v)} \frac{\mathrm{h}_u^{(k)}}{|\mathrm{N}(v)|} + \mathrm{B}_k \mathrm{h}_v^{(k)}\right)$$

$$\mathrm{z}_v = \mathrm{h}_v^{(K)}$$

Final node embedding

based on https://web.stanford.edu/class/cs224w/

GraphST



Step 2

Learn spot embedding using GCN (Graph Convolution Network)

$$H^{(l+1)} = \sigma(\hat{A} H^{(l)} W^{(l)} + b^{(l)})$$

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

$$H^{(0)} = X$$

Final output $Z = H^{(L)}$

Step 3

Self supervised contrastive learning

- positive pairs
  - spot embedding z_i
  - local context vector

$$s_i = \sigma\left(\frac{1}{|N(i)|} \sum_{j \in N(i)} z_j\right)$$

- negative pairs $\tilde{z}_i$
  - (preserve the spatial structure)
  - shuffle the rows of X

- contrastive loss

$$L_{CL} = -\log D(z_i, s_i) - \log(1 - D(\tilde{z}_i, s_i))$$

$$D(z, s) = \text{sigmoid}(z^T W s)$$

GraphST

Step 3

Self supervised contrastive learning

- positive pairs
  - spot embedding z_i
  - local context vector

$$s_i = \sigma\left(\frac{1}{|N(i)|}\sum_{j \in N(i)} z_j\right)$$

- negative pairs $\tilde{z}_i$
  - (preserve the spatial structure)
  - shuffle the rows of X

- contrastive loss

$$L_{CL} = -\log D(z_i, s_i) - \log\left(1 - D(\tilde{z}_i, s_i)\right)$$

$$D(z, s) = \text{sigmoid}(z^T W s)$$

Step 4

Add a reconstruction loss

$$\mathcal{L}_{recon} = \sum_{i=1}^{N} \|x_i - Dec(z_i)\|_2^2$$

Final loss:

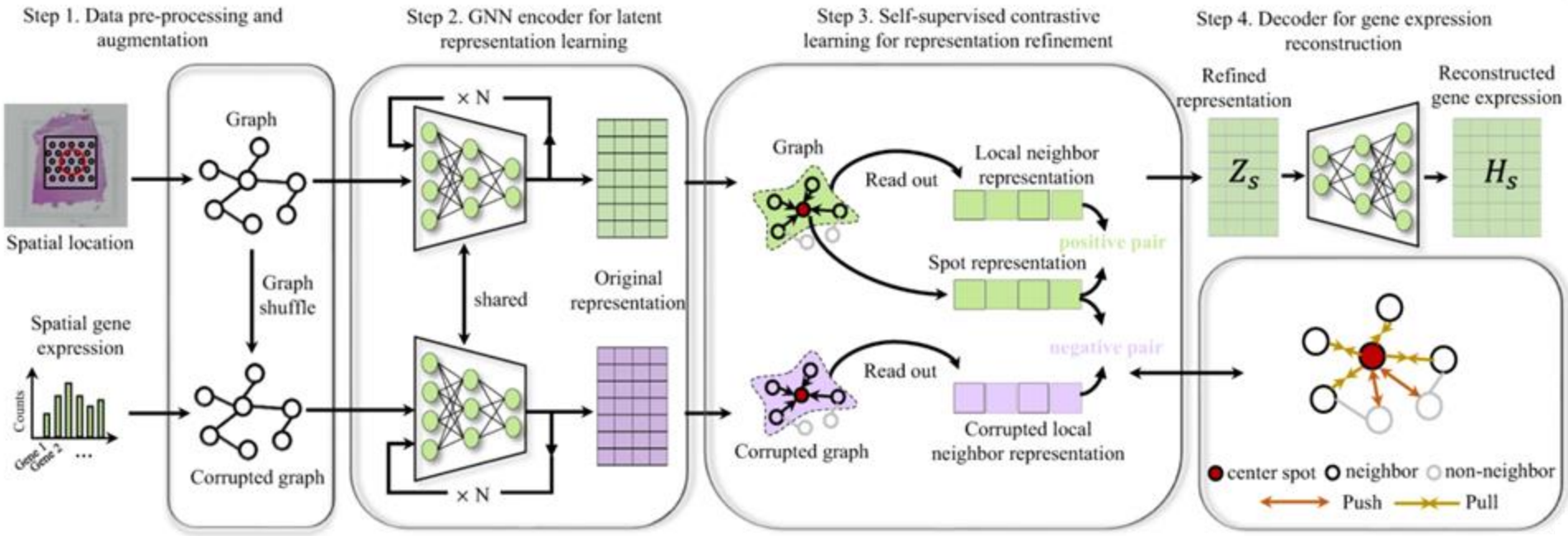
$$\lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{CL}$$

Step 5

mclust (Gaussian Mixture Model)

on $\hat{X} = Dec(Z)$

GraphST



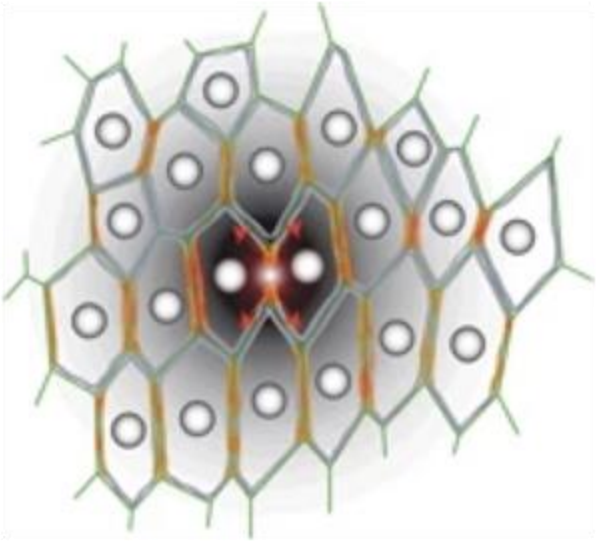
adapted from: Long, Y., Ang, K.S., Li, M. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. (2023).

To make Maps that represent Living Systems we need two answer at least two Questions

- **What are the units?**
  If we do not yet know the key components or players, how can we discover and define them?

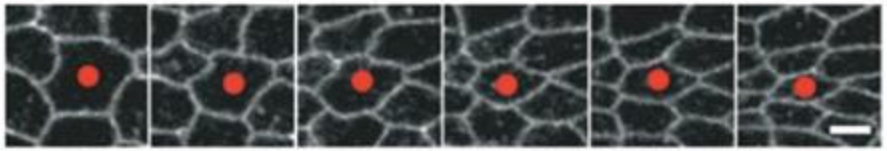
- **How are they organized and how do they interact?**
  Given the players, what are the rules and patterns that govern their interactions **across scales**?

Case study: material science meets biology meets machine learning



https://www.nature.com/articles/nrm2606

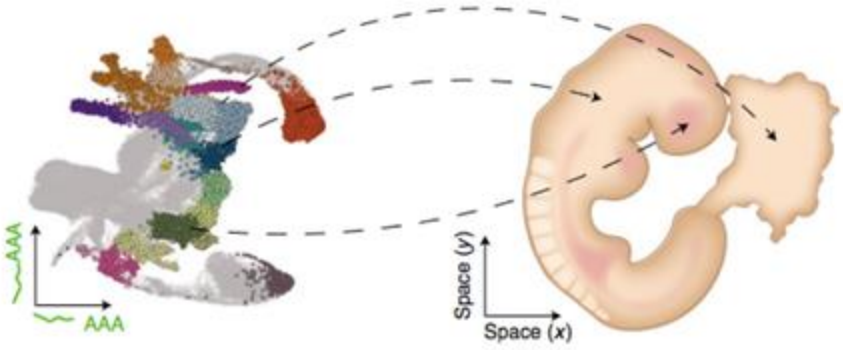
Tissues are shaped by mechanical forces,
* What are the origins and the nature of these forces?
* How do tissues fold to create 3D shapes?



Martin, Kaschube, & Wieschaus.
Pulsed contractions of an actin–myosin network drive apical constriction.
Nature 457, 495–499 (2009)

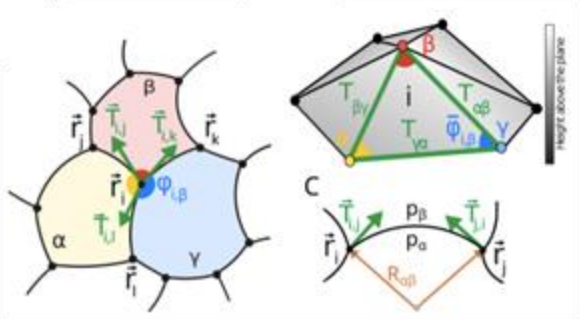
Case study: material science meets biology meets machine learning

Quantifying how multiple modalities interact:
How do mechanics and genomics interact in the context of early boundary formation?

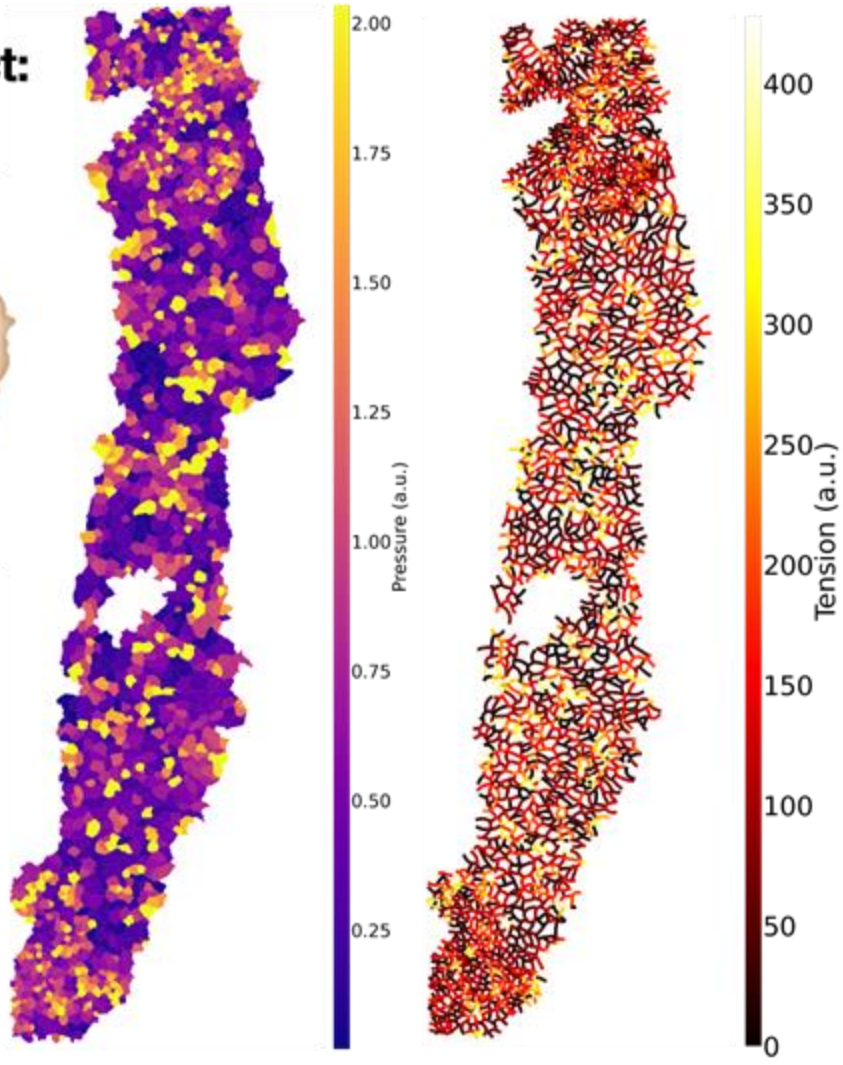


Arguelaget, Cuomo, Stegle ,Marioni (2021)

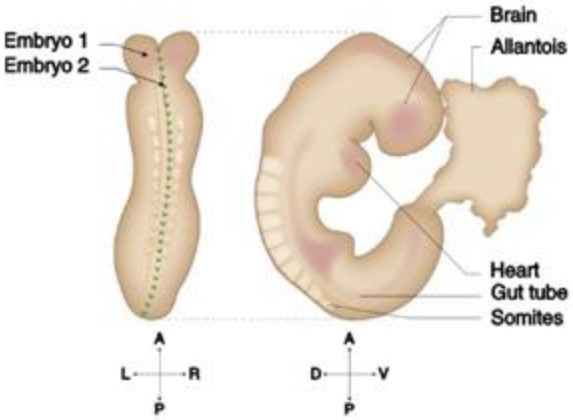
Spatial Mouse, Lohoff et al, Nat. Biotech 2022



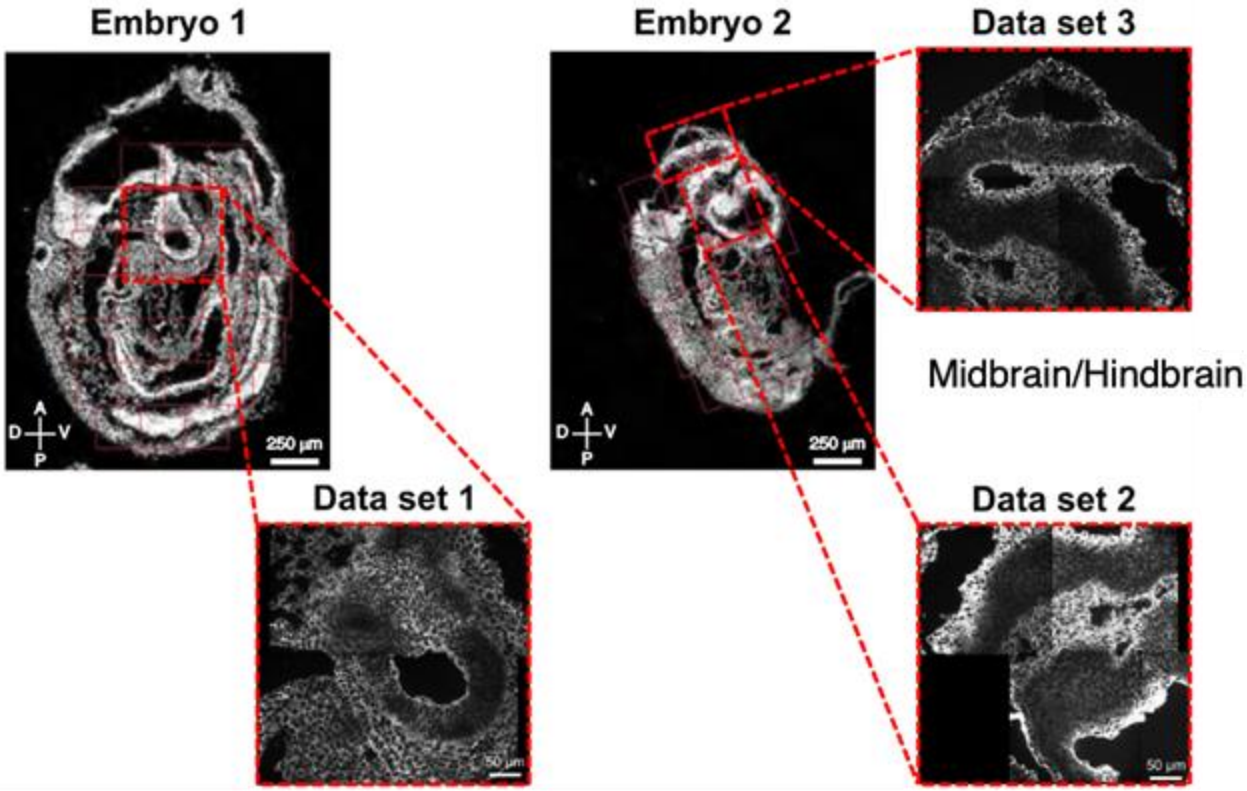
Noll, Streichan, Schraiman, Phys Rev X. 2020



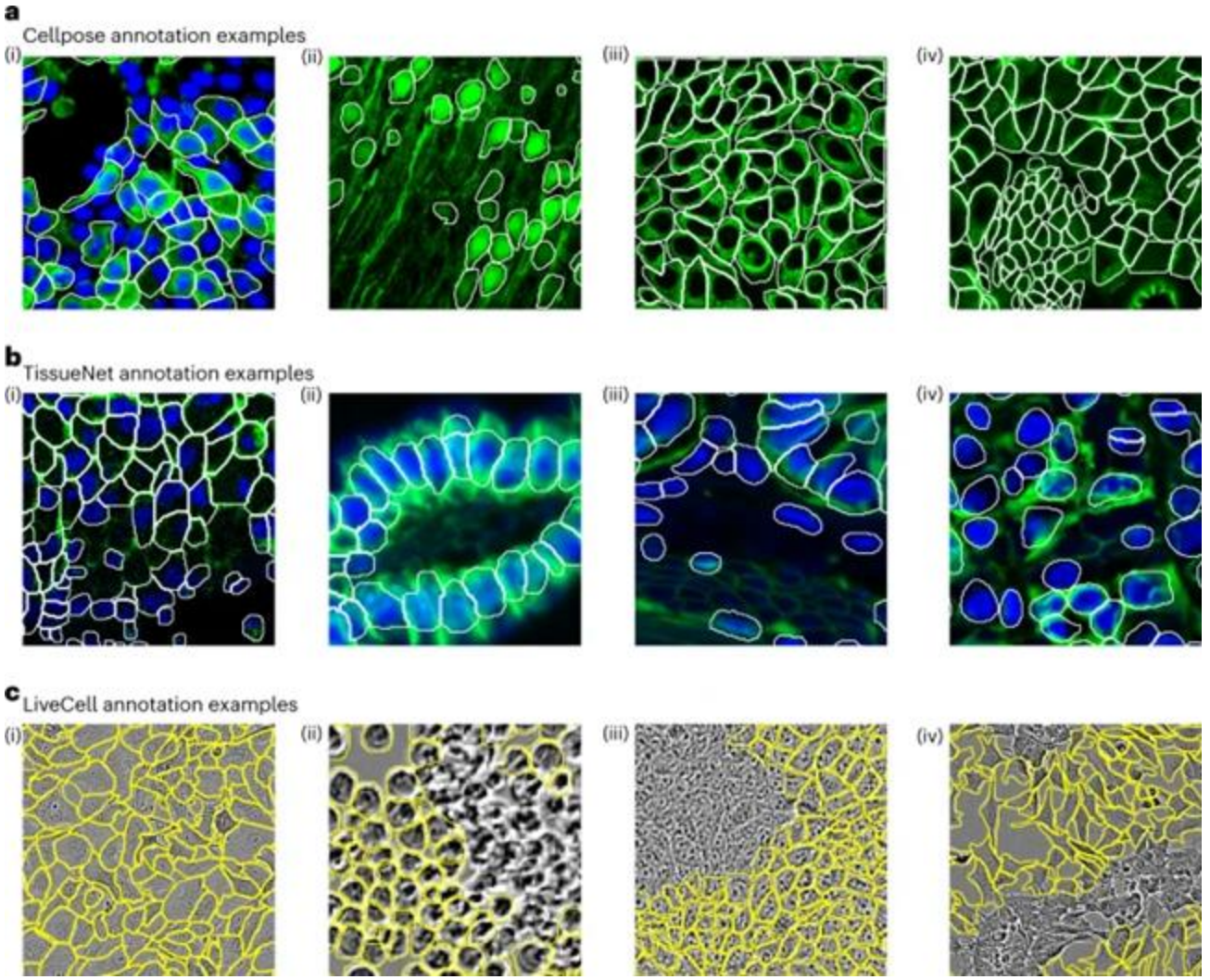
Case study: material science meets biology meets machine learning



Lohoff et al, Nat. Biotech 2022



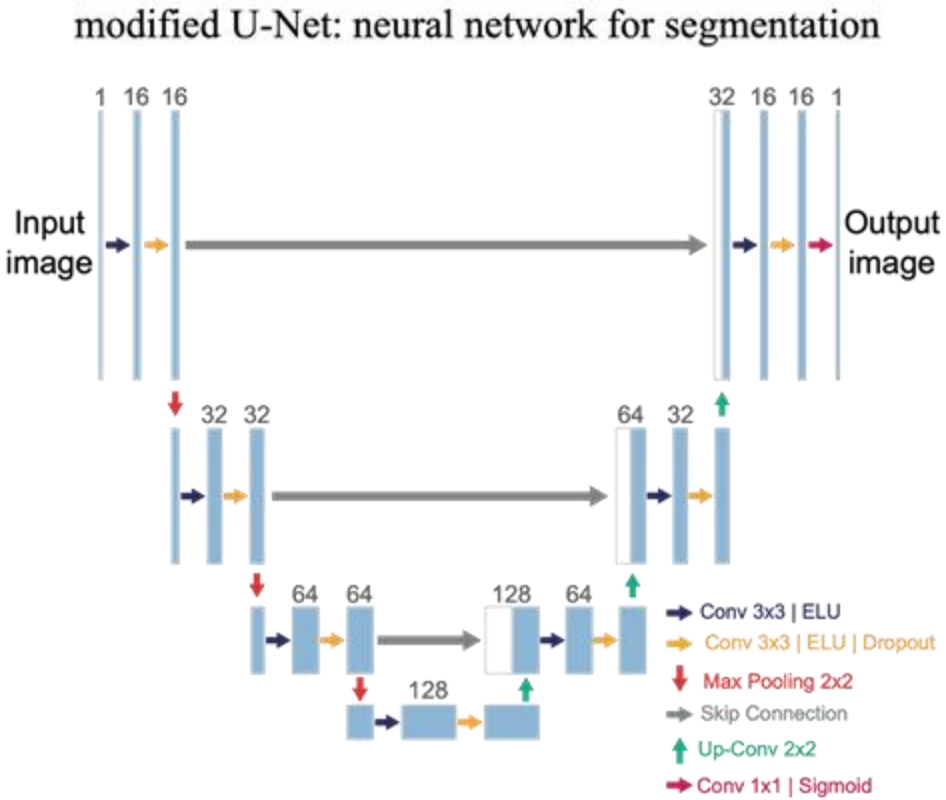
Ingredient: segmentation



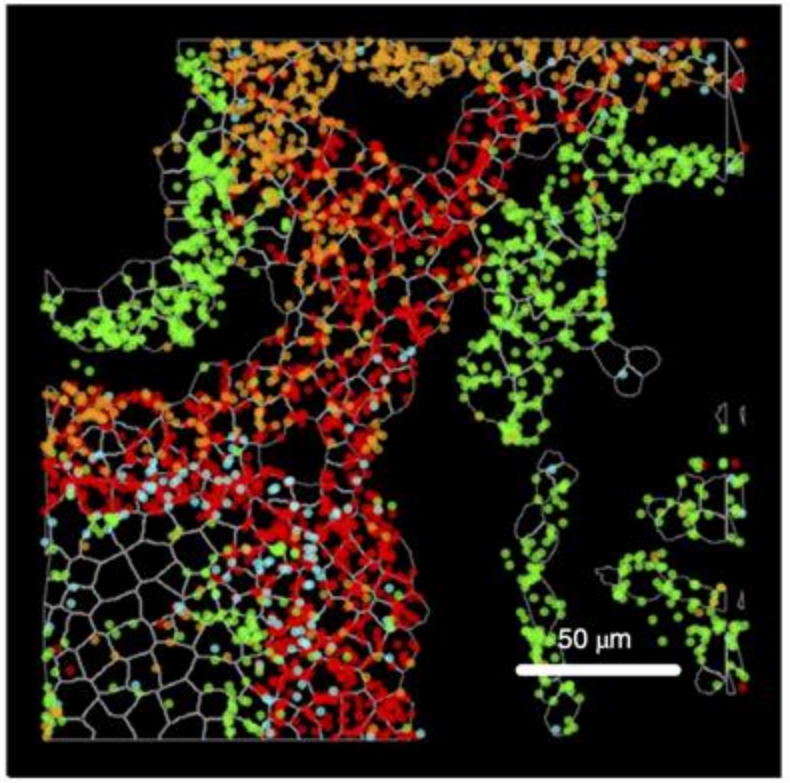
Cellpose 2.0: how to train your own model. Pachitariu & Stringer, 2022

Scale: cellular organization and data processing

Ingredient: segmentation

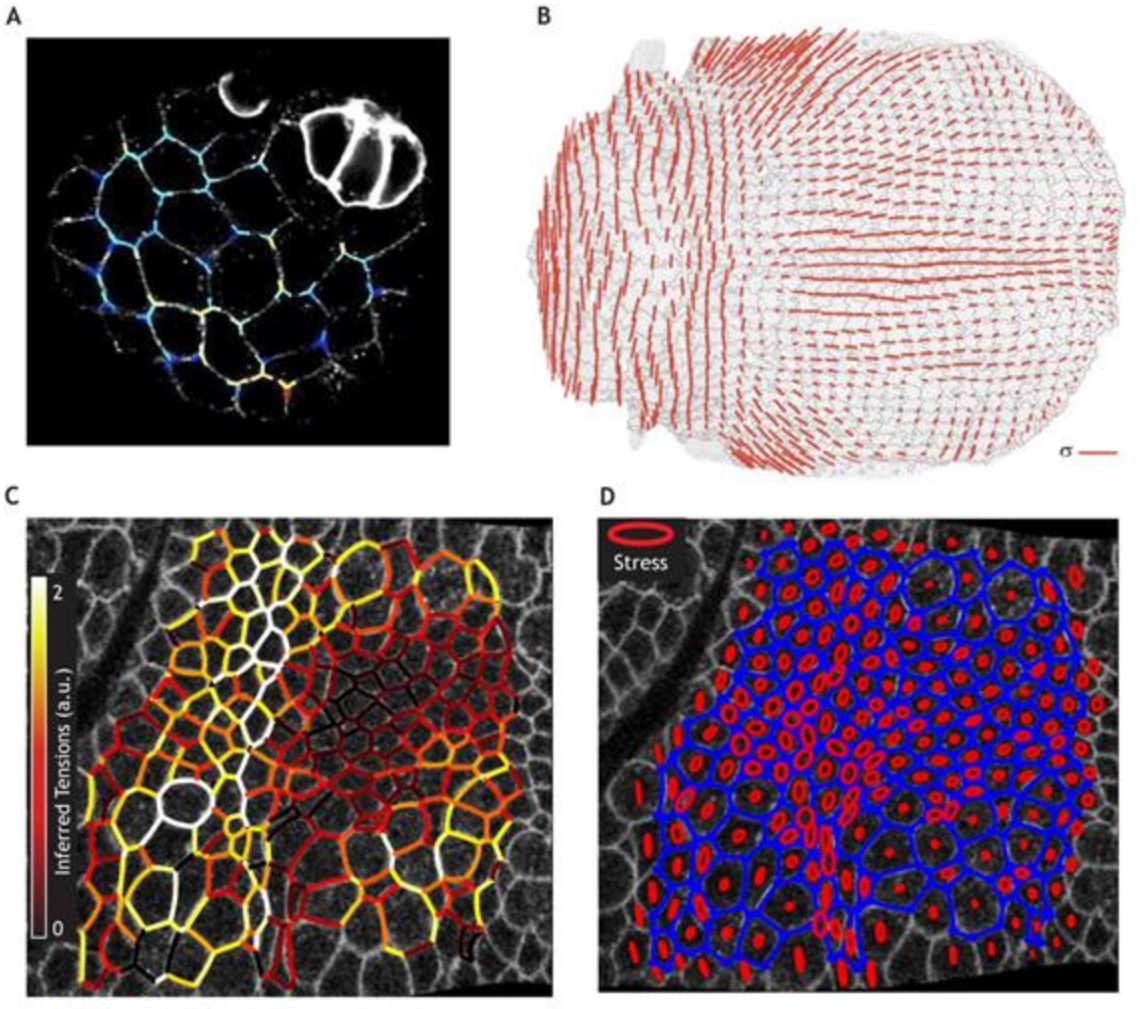


Ronneberger, Fischer, Brox, MICCAI, 2015



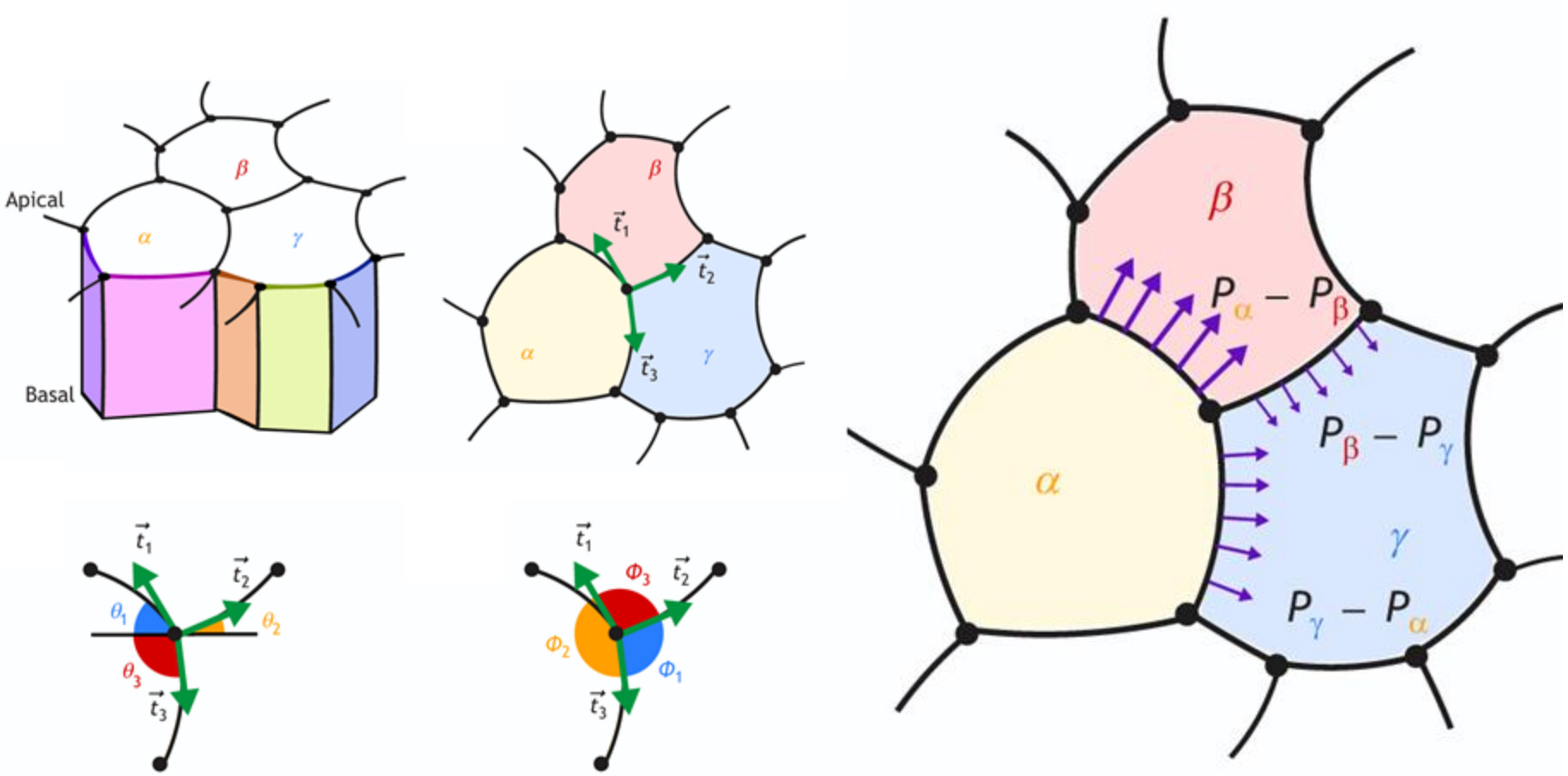
Lohoff et al., Nature Biotech., 2022 (seqFISH)

Ingredient: inferring tension without measuring it



Examples of stress inference in development. Adapted from: Inferring cell junction tension and pressure from cell geometry. Roffay et al, 2021

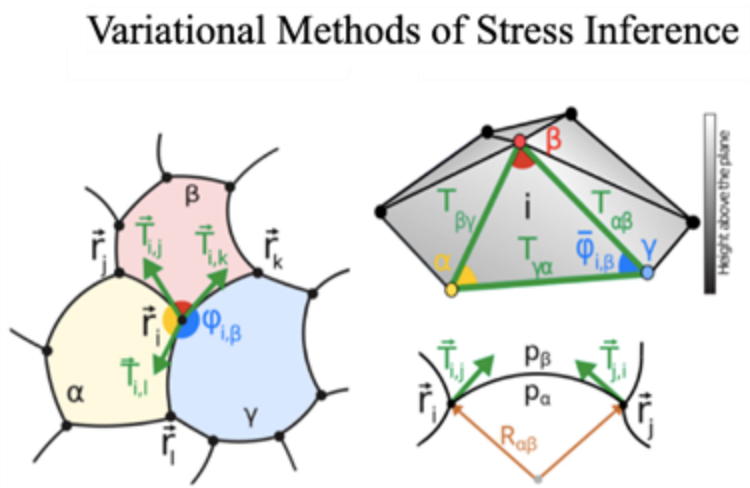
Ingredient: inferring tension without measuring it



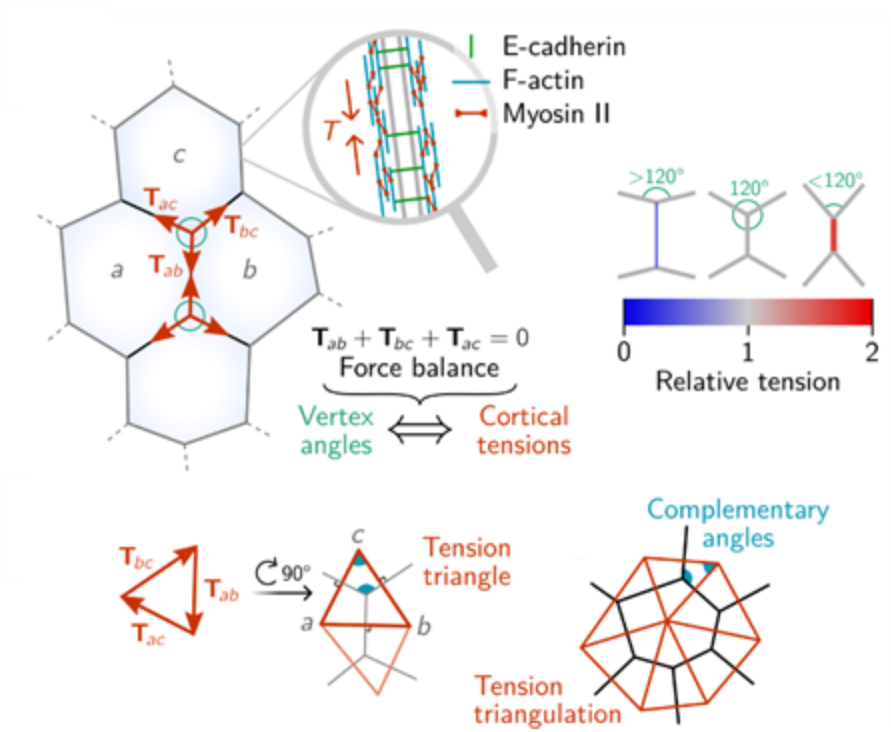
Adapted from: Inferring cell junction tension and pressure from cell geometry. Roffay et al, 2021

Scale: morphological and mechanical phenotypes

Ingredient: estimating tension without measuring it

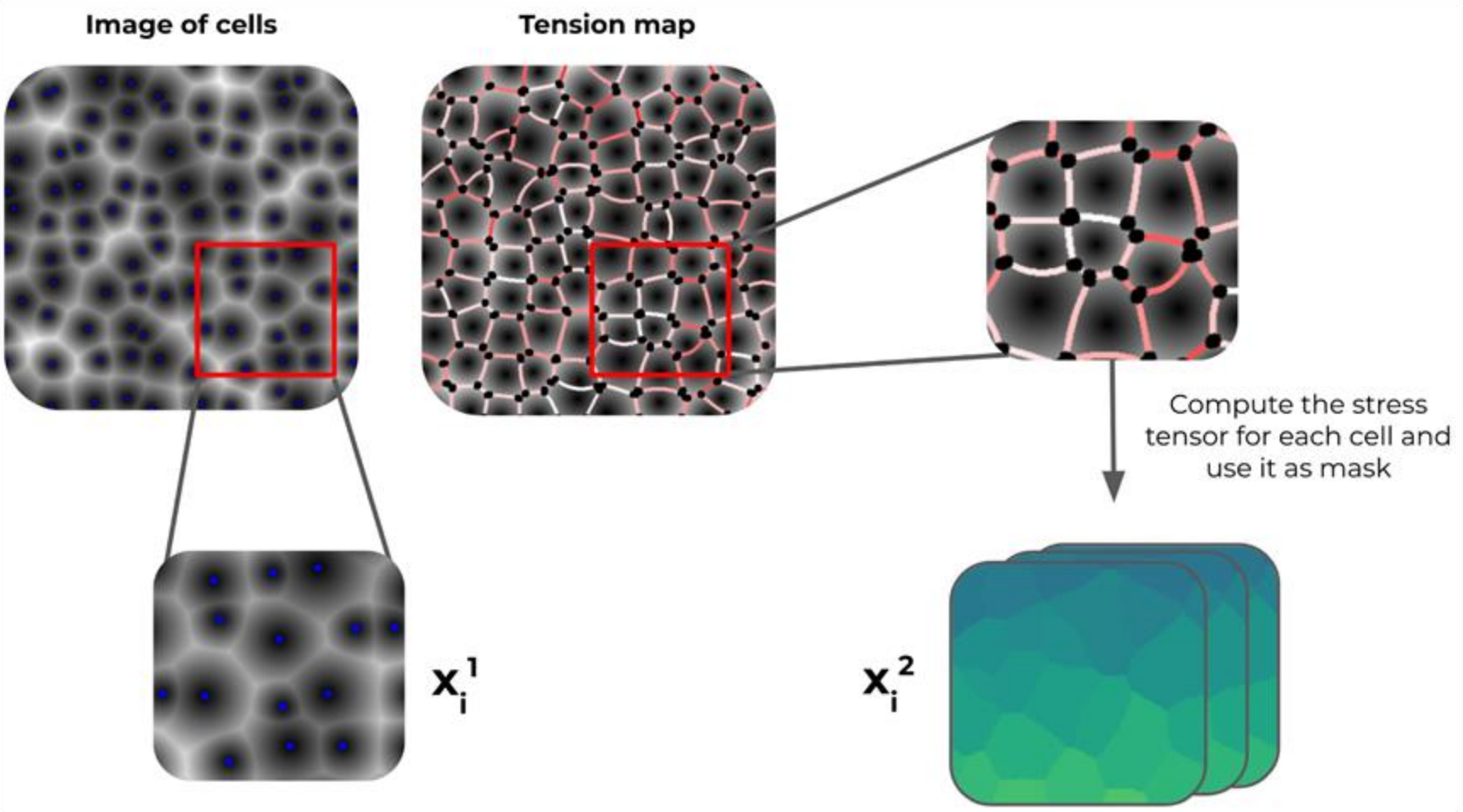


Noll, Streichan, Schraiman, Phys Rev X. 2020

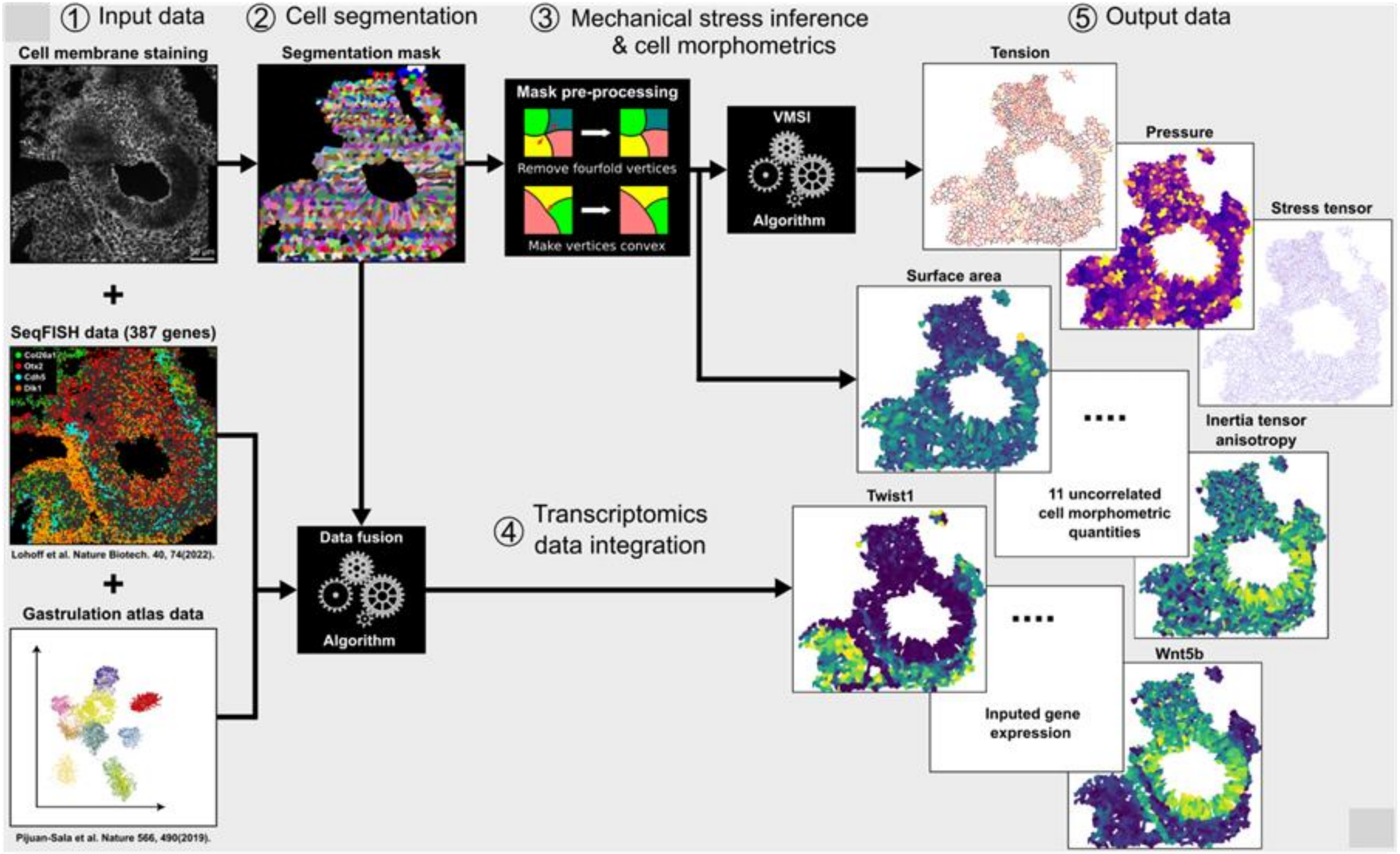


Brauns, Claussen, Wieschaus, Schraiman, arxiv 2023

Inferring tension from images: a pseudo-gene toy case

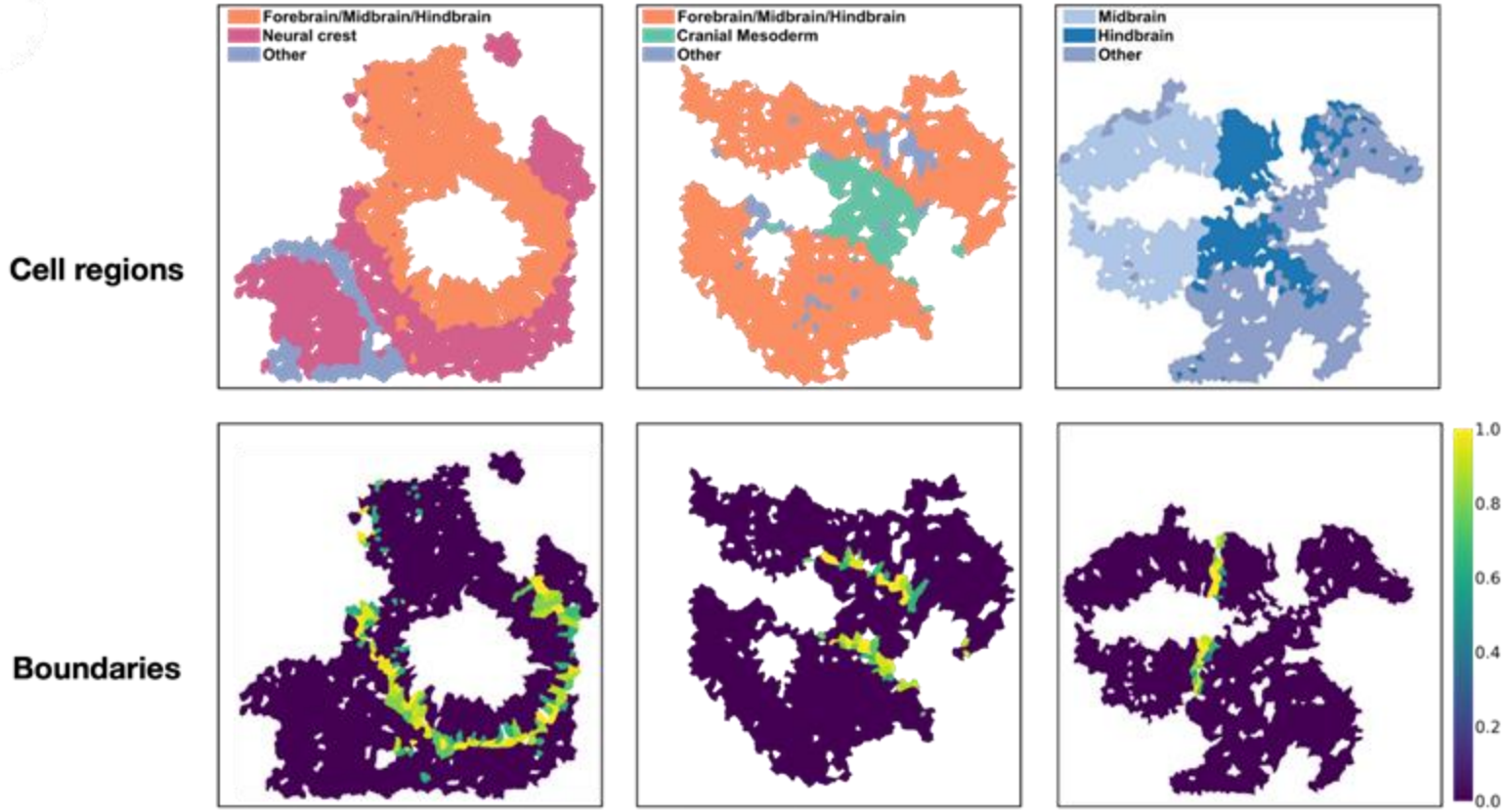


A machine learning pipeline for multi-scale integration

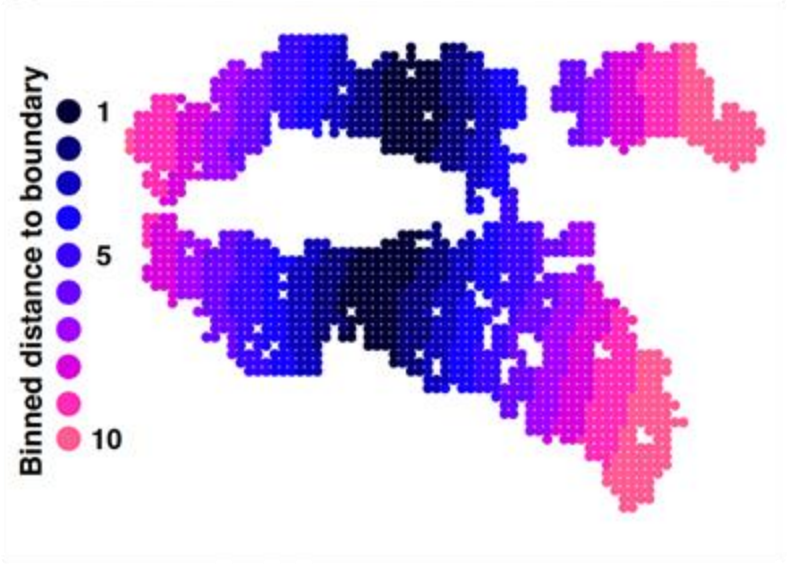


Cell type boundaries recapitulate mechanical properties

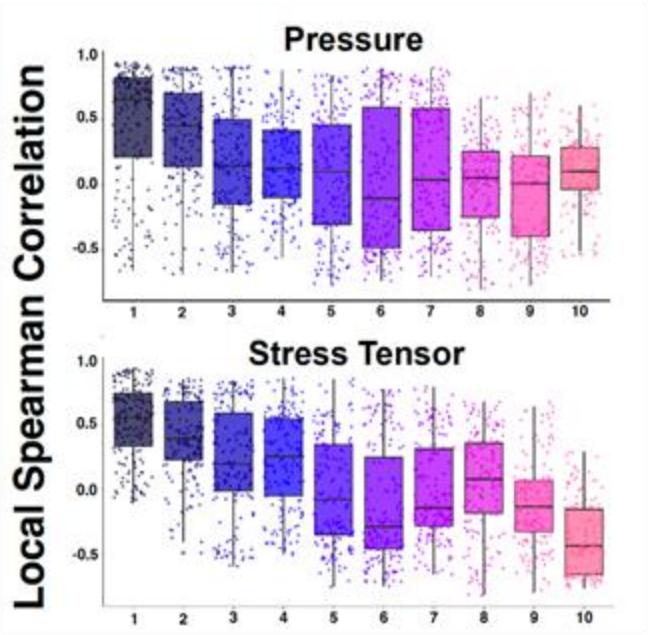
Gene expression identified boundaries recapitulate tension driven tissue compartments



Gene expression boundaries recapitulate tension defined compartments



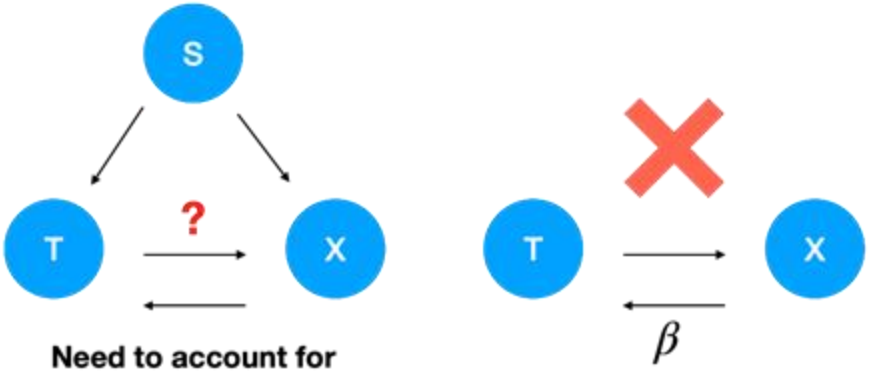
Midbrain-hindbrain region



Binned distance to boundary

Spatial regression testing framework

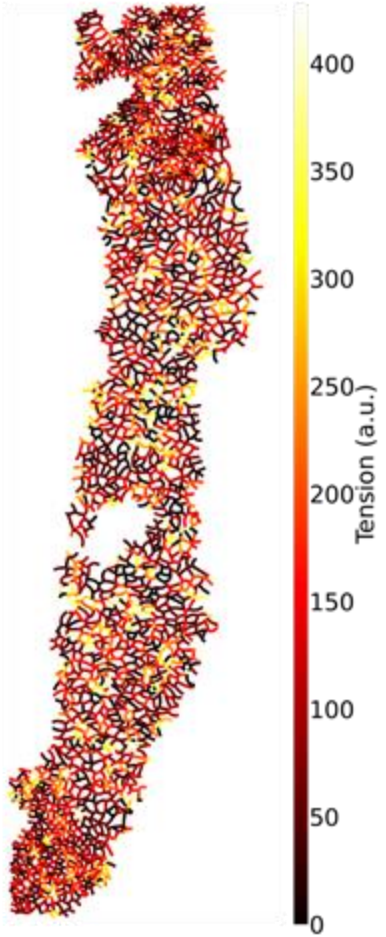
Geoadditive structural equation (gSEM) models identify groups of genes associated with tension



Need to account for spatial confounders!

$$x_i^g = f^x(s_i) + \epsilon_i^x, \epsilon_i^x \sim N(0,\sigma^2 I) \qquad r_i^{x,g} = x_i^g - \hat{f}^x(s_i)$$
$$t_i = f^t(s_i) + \epsilon_i^t, \epsilon_i^t \sim N(0,\sigma^2 I) \qquad r_i^t = t_i - \hat{f}^t(s_i)$$

$$r_i^{x,g} = \beta_g r_i^t + \epsilon_i,$$

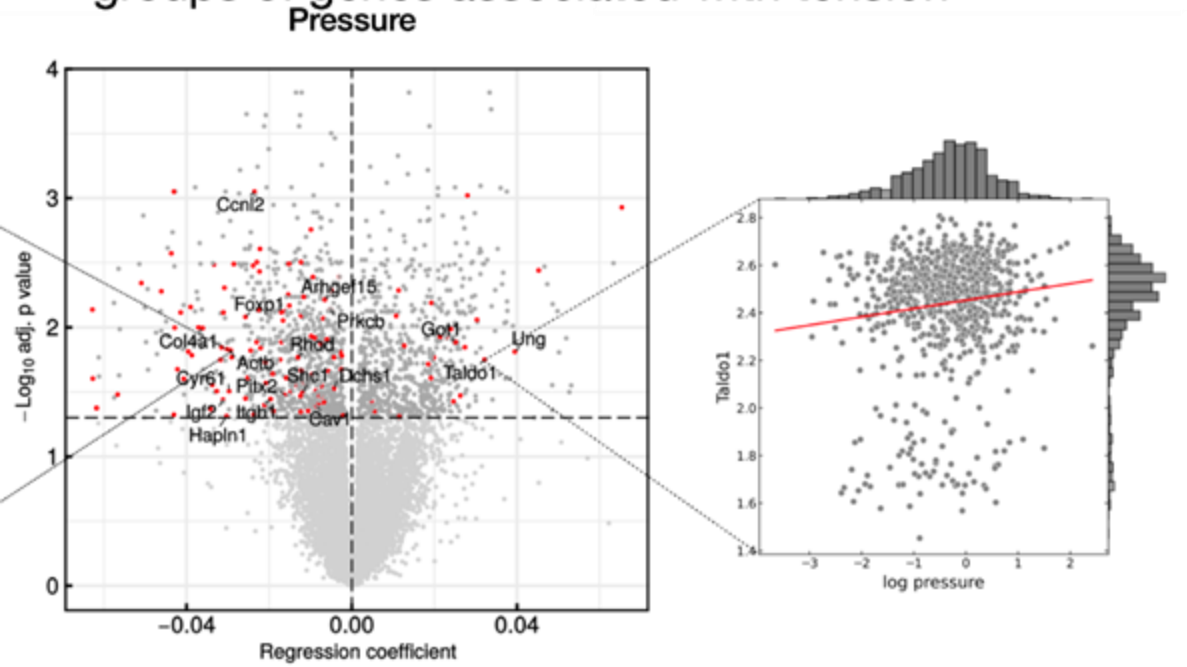


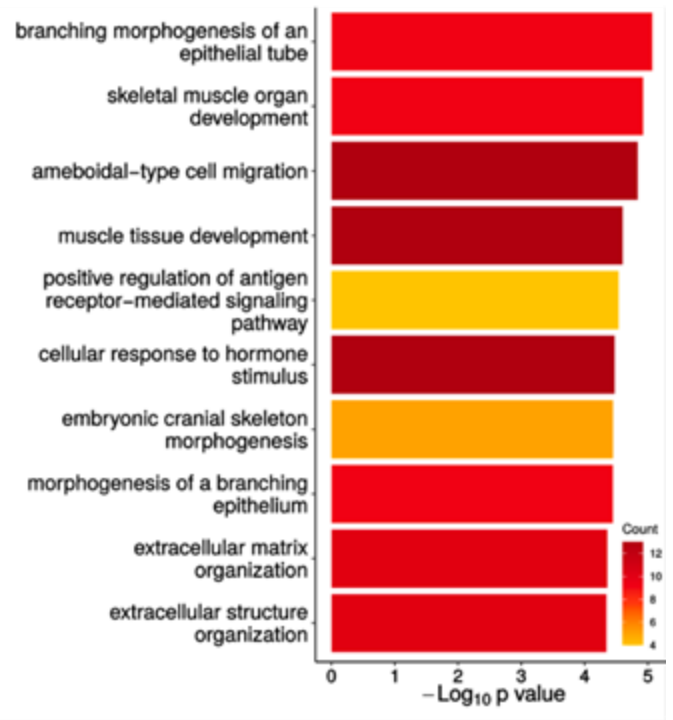
gSEM:Thaden and Kneib (2018)

Structural Equation Models for Dealing With Spatial Confounding

Spatial regression testing framework

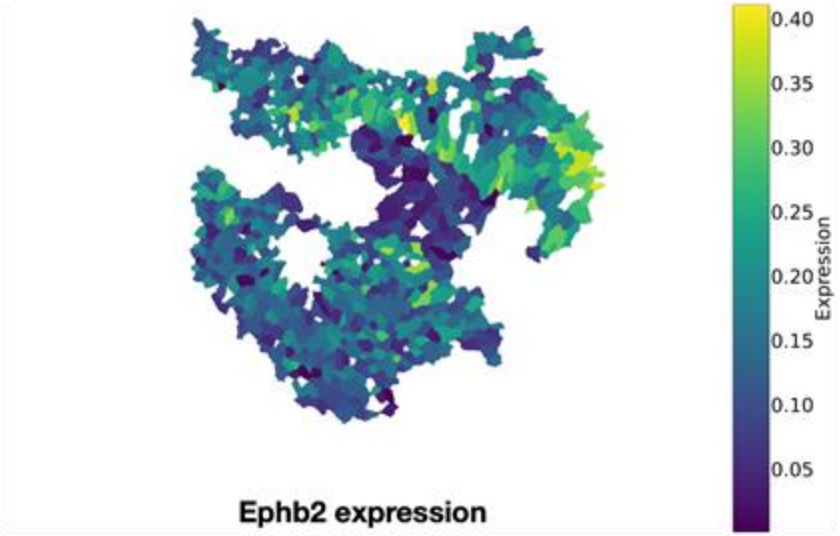
Geoadditive structural equation (gSEM) models identify groups of genes associated with tension



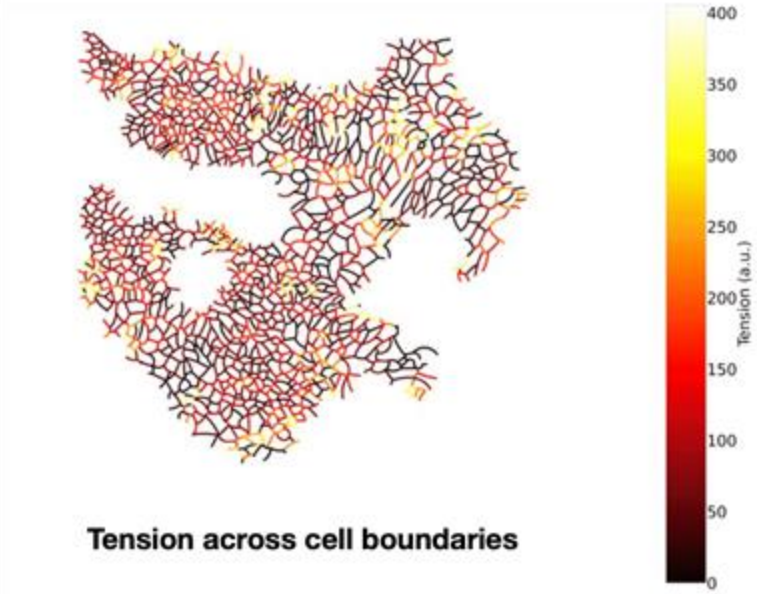


Spatial gene variability informs tissue mechanical stability

Geoadditive structural equation (gSEM) models identify groups of genes associated with tensional differences

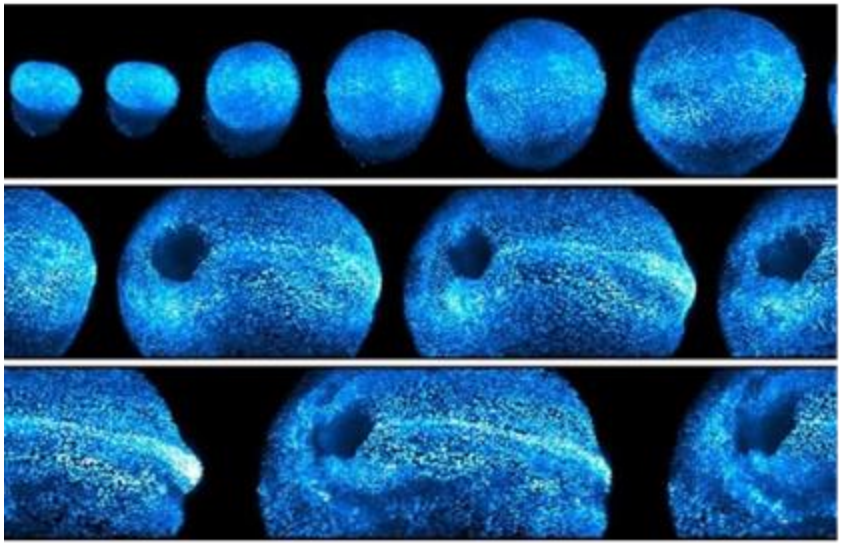


Ephb2 expression

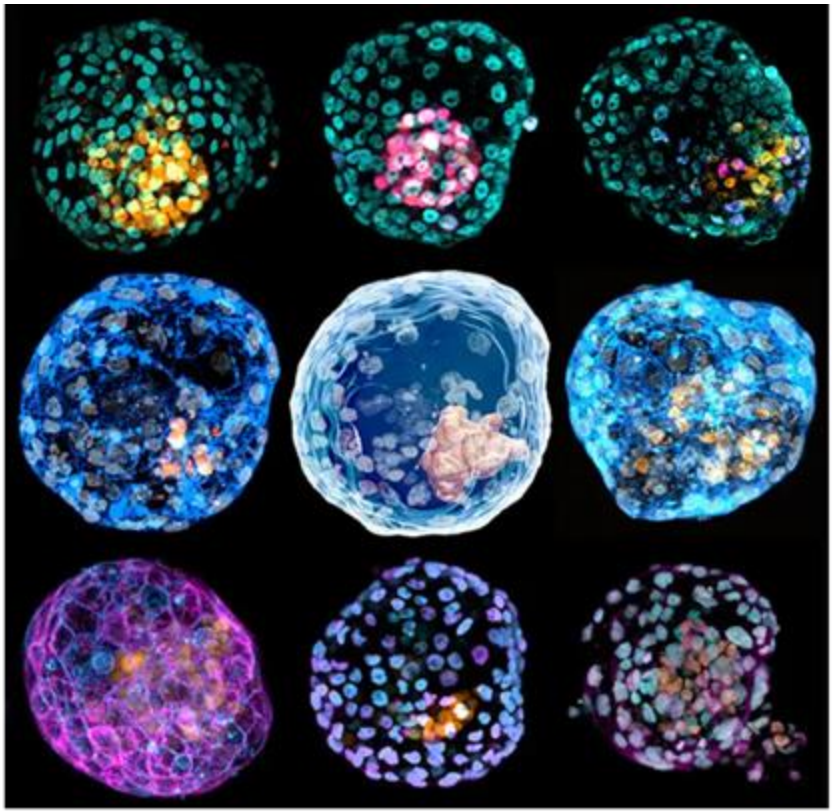


Tension across cell boundaries

Integrating tension & genomics in a spatial context: Going forward

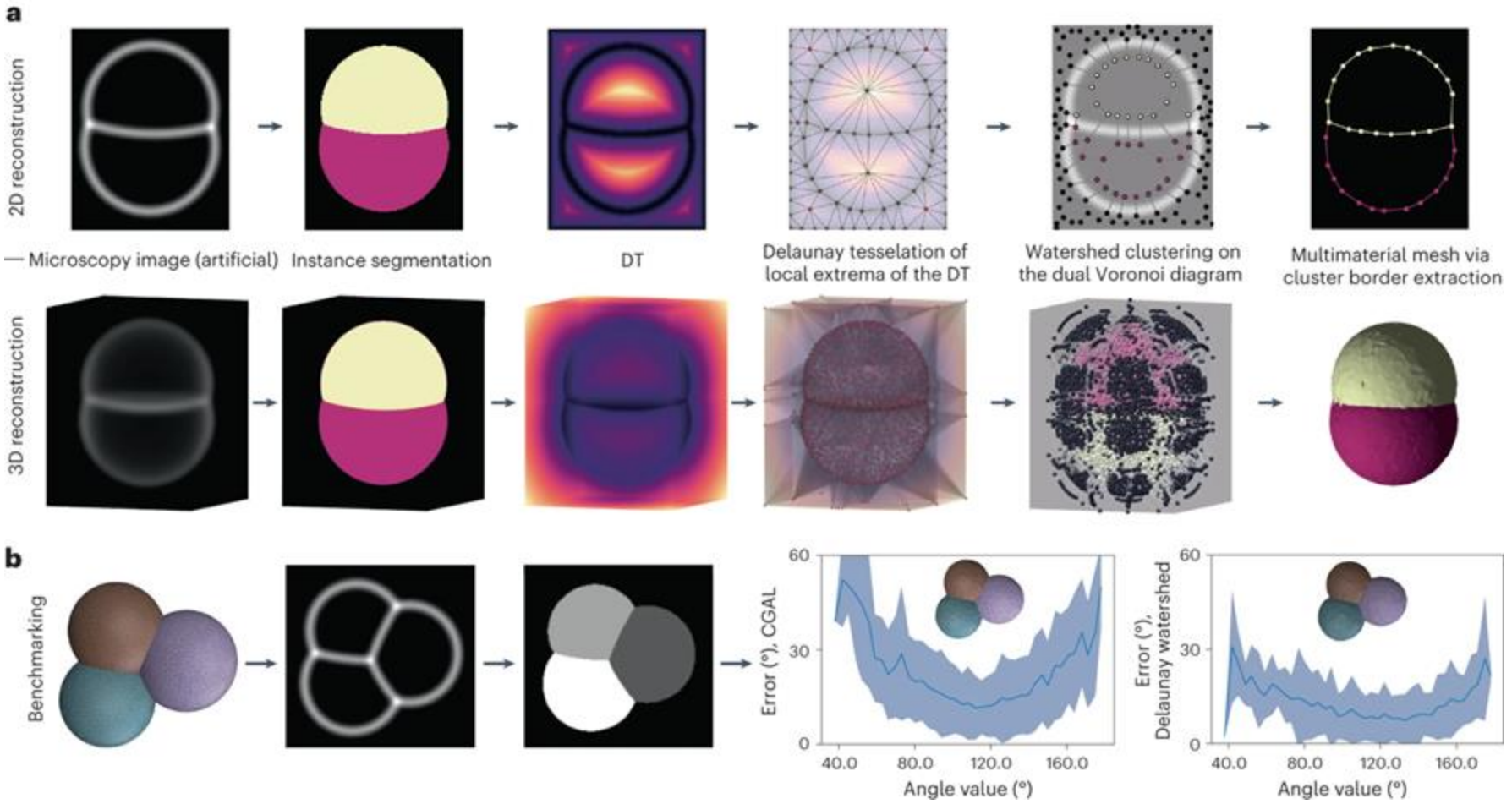


K McDole et al., Cell 2018



X. Liu et al., Nature 2021 (iBlastoids)

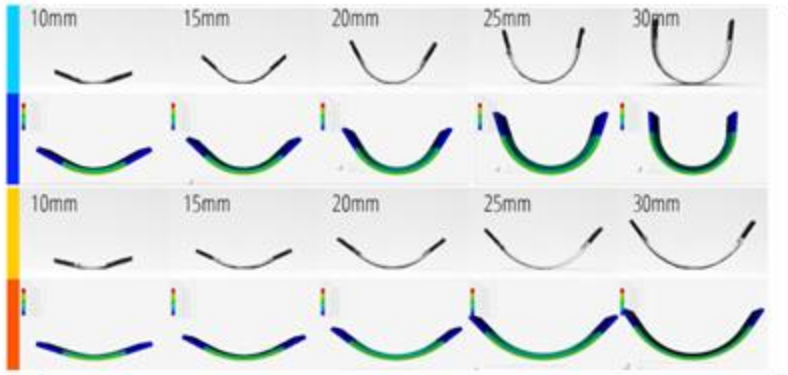
Integrating tension & genomics in a spatial context: Going forward

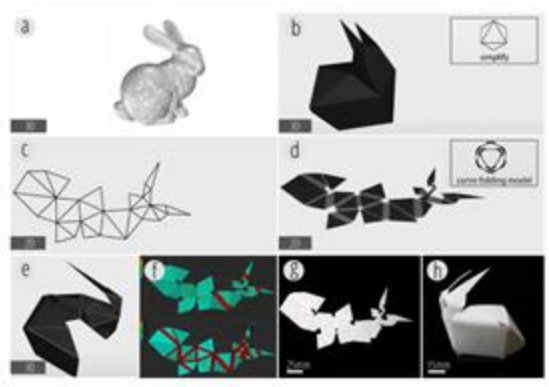


Multimaterial mesh generation algorithm. From: Embryo mechanics cartography: inference of 3D force atlases from fluorescence microscopy. Ichbiah et al, 2023

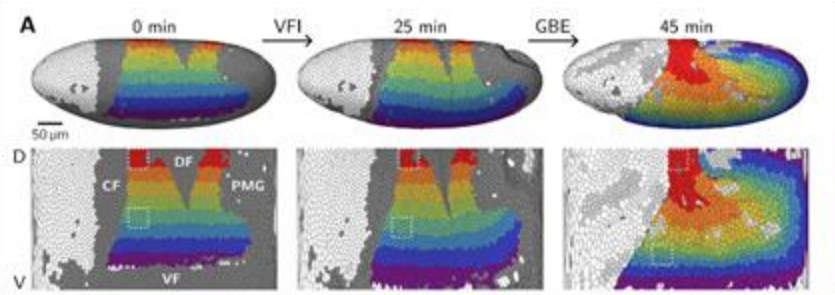
Beyond representation learning: digital twins + modeling and inverting the dynamics of emergent systems

Source: Jo Nakashima





Thermorph: Democratizing 4D Printing of Self-Folding Materials and Interfaces; An et. al, 2018



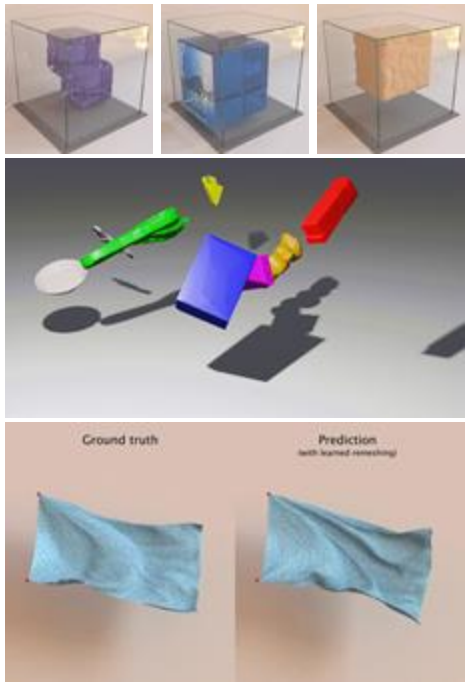
Brauns, Claussen, Wieschaus, Schraiman, arxiv 2023

Next:

* Embryos as **materials**: where do the forces that fold embryos come from? Can we distinguish between active and passive forces?
* Embryos as **materials**: inverse design - which tensions are required to generate a final shape/ final tensional distribution?

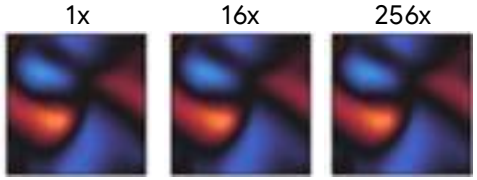
Beyond representation learning: digital twins + modeling and inverting the dynamics of emergent systems

AI for simulation

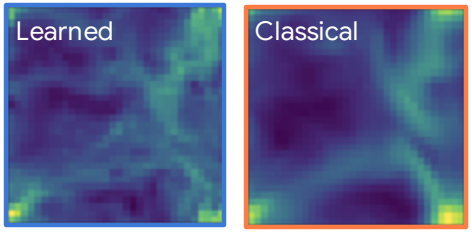


Mrowca et al (2018); Sanchez-Gonzalez et al. (2020); Pfaff et al (2021); Stachenfeld et al. (2022); Allen, Rubanova* et al. (2023); Wu et al (2023)*

Low temporal resolution

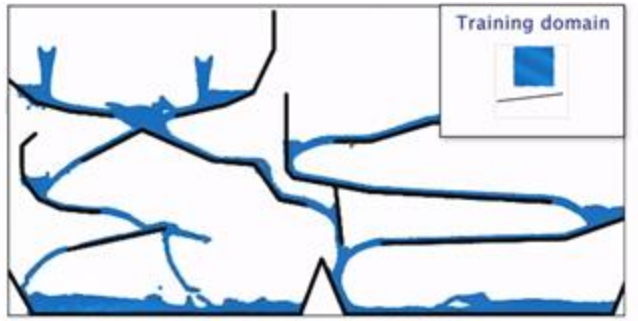


Low spatial resolution



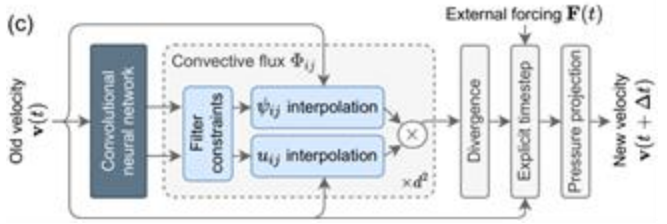
Stachenfeld et al 2021

Locality, Translation Equivariance



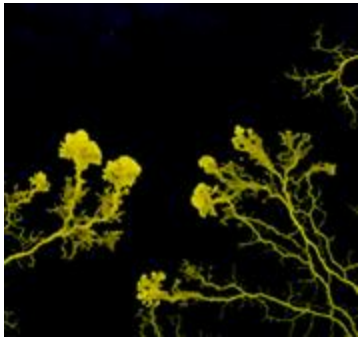
Pfaff et al 2021

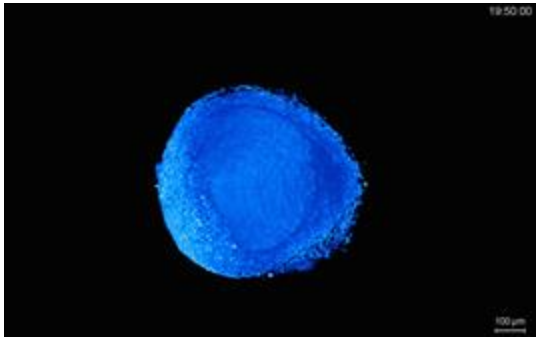
Physics-informed

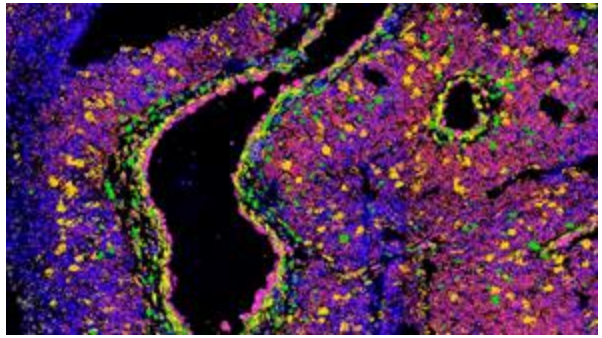


Kochkov et al 2021

Beyond representation learning: digital twins + modeling and inverting the dynamics of emergent systems







Thank you!

More @ https://www.morpho-lab.com/; illustration by Elena Bansh

**** Accepting postdoc and grad student applications ! ****
Get in touch: bmd2151@columbia.edu

NIH NIGMS

CIFAR

Columbia University
IN THE CITY OF NEW YORK