

指定討論：データサイエンスの世界的潮流と その展望

熊坂 夏彦

Wellcome Trust Sanger Institute

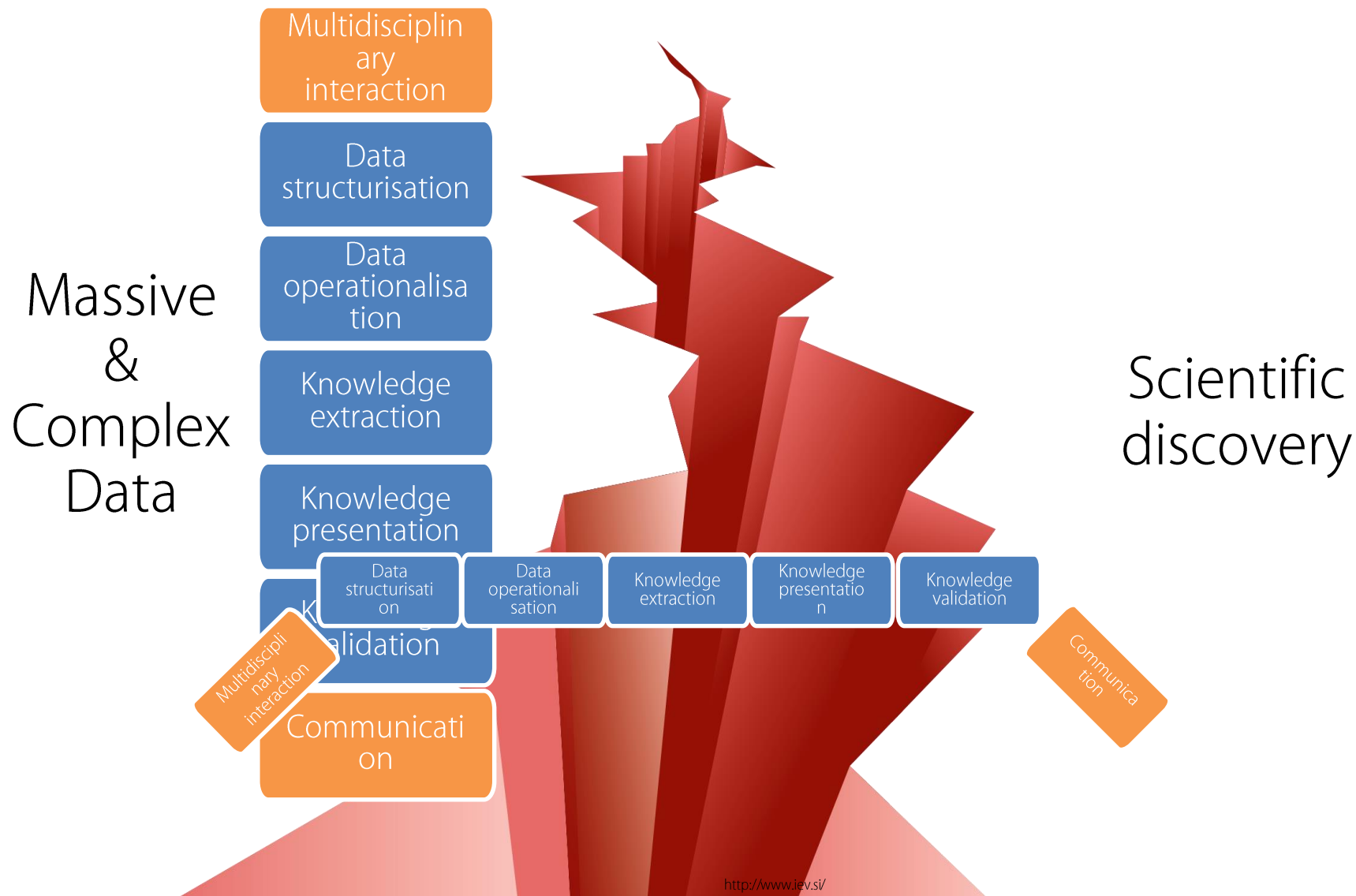
Role of statistics

Data



Scientific
discovery

Statistics is not enough to cover things



We need Data Science!

- Is *data* really given?
- How can we transfer **data science technologies** in one discipline to another?
- Can **TextilePlot** be an ubiquitous technique in data science?

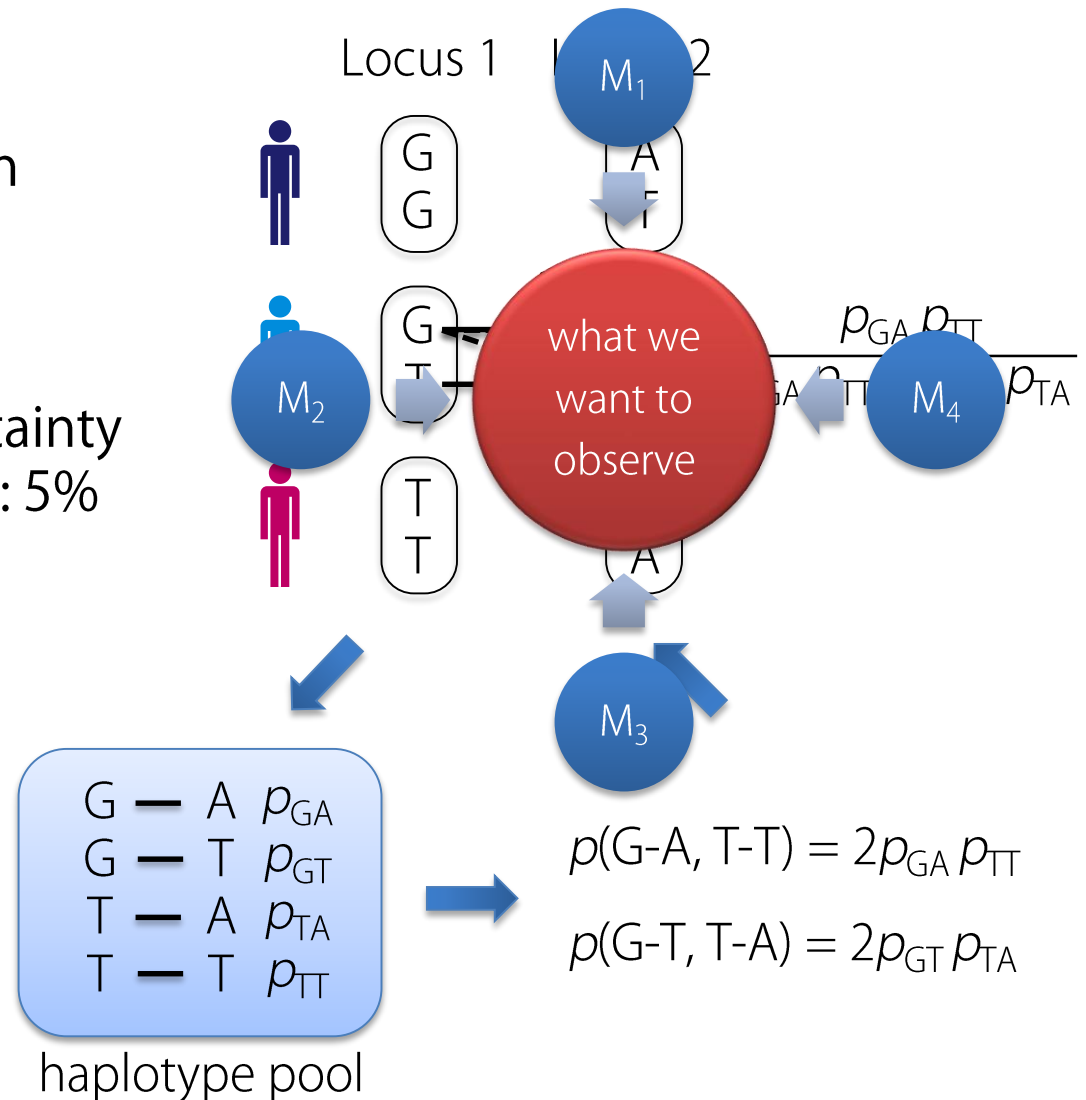
Is X really given?

$$L(\theta|X)$$

- Data is generated by someone somewhere
- Data is not easy to access → InterDatabase
- Data is not quite ready

Data is not quite ready

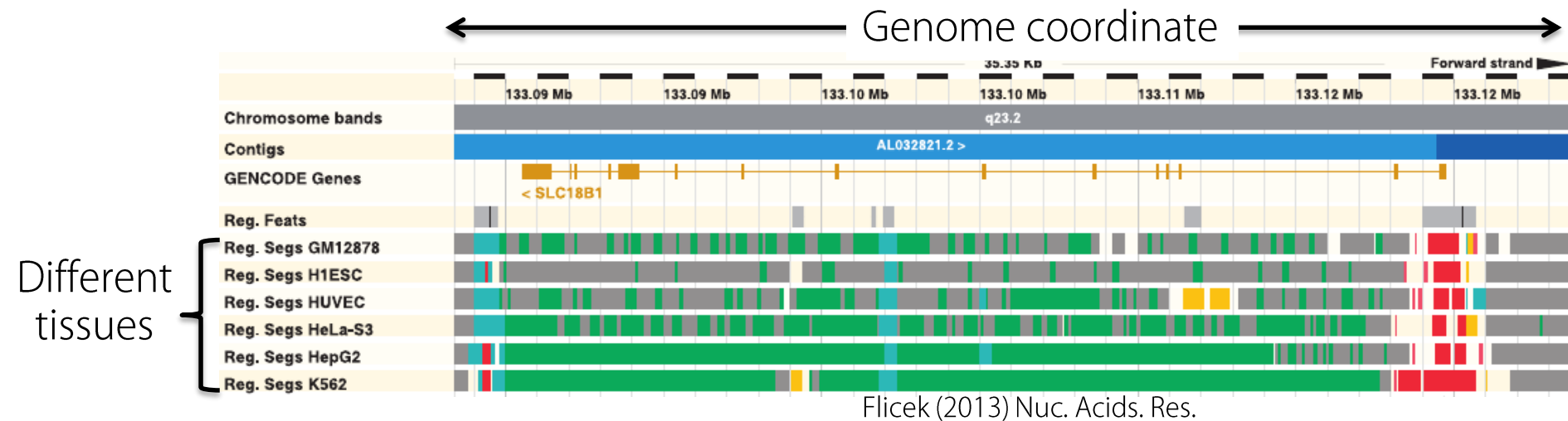
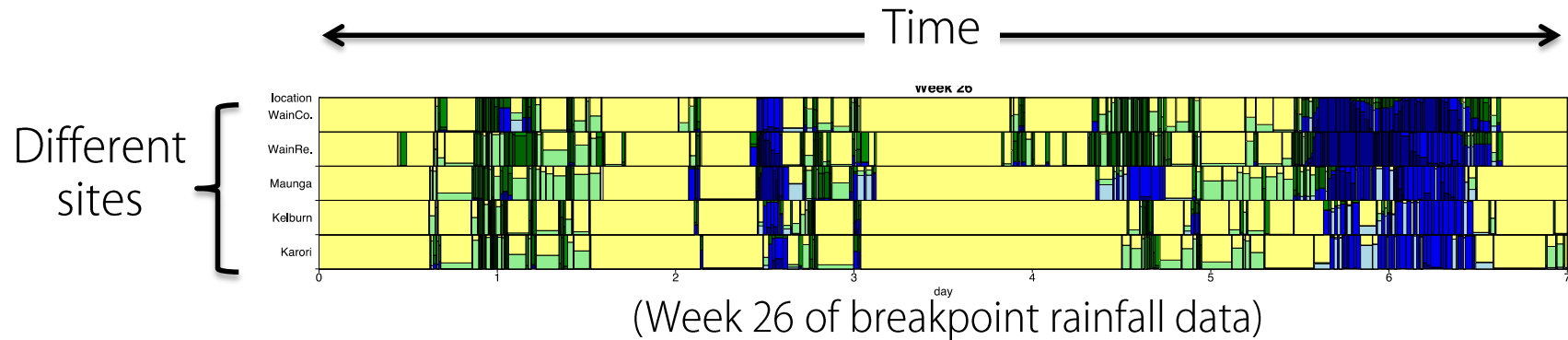
- **Indirect**
 - e.g., DNA methylation
- **Uncertain**
 - e.g., genotype uncertainty
AA: 80%; AB: 15%; BB: 5%
- **Incomplete**
 - haplotype
- **etc...**



We need Data Science!

- Is *data* really given?
- How can we transfer **data science technologies** in one discipline to another?
- Can TextilePlot be an ubiquitous technique in data science?

Similar technologies seen in different fields

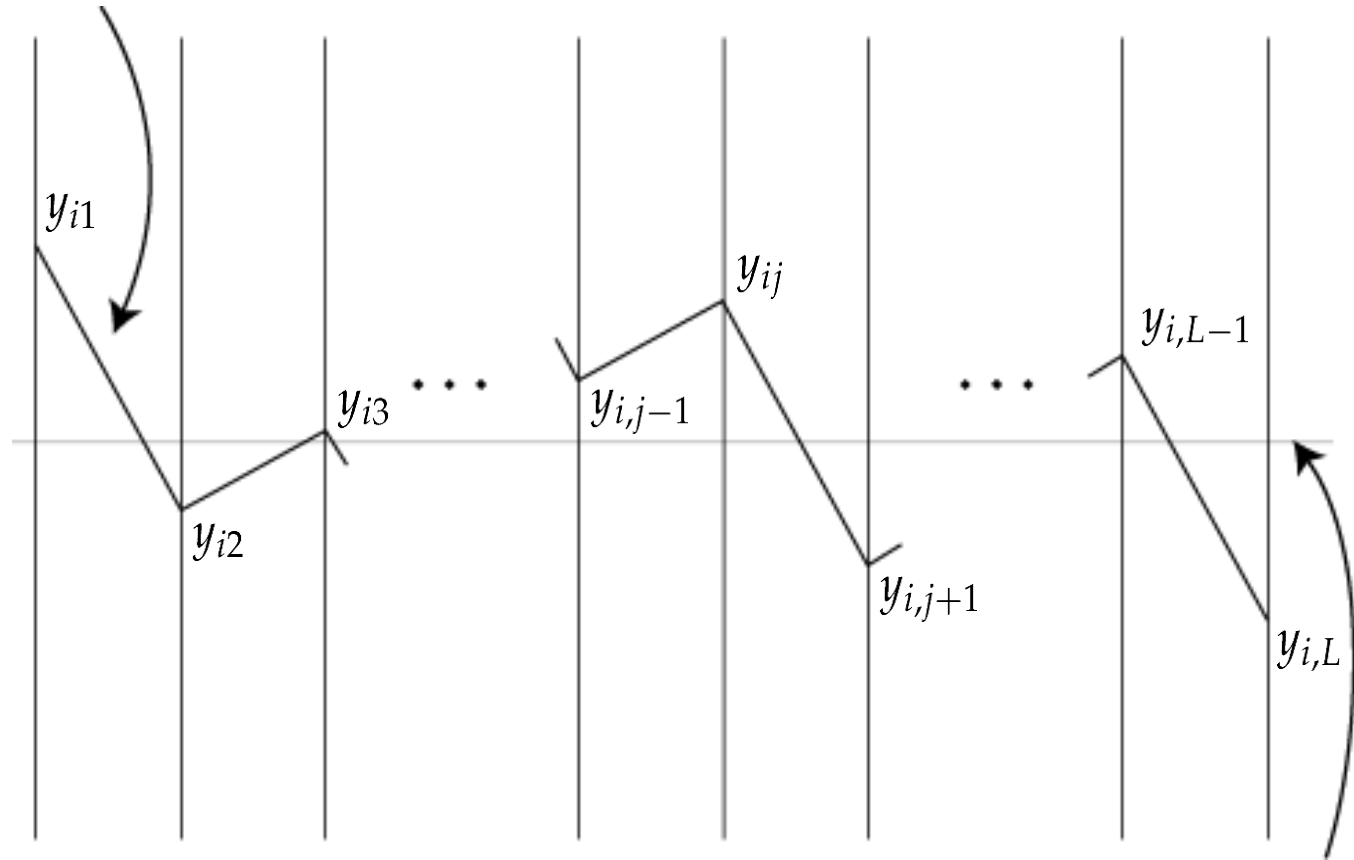


We need Data Science!

- Is *data* really given?
- How can we transfer data science technologies in one discipline to another?
- Can **TextilePlot** be an ubiquitous technique in data science?

Textile Plot

i th connecting line



i th horizontal line

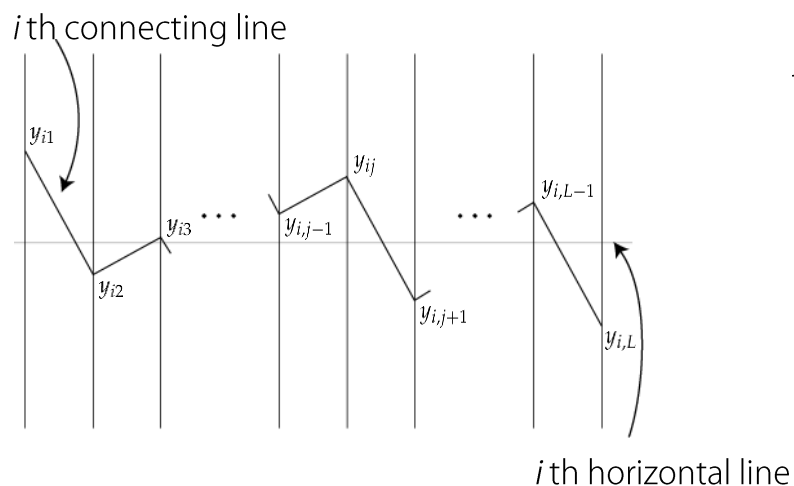
Horizontalization criterion

Optimization Problem

$$\begin{cases} \sum_{ij} (y_{ij} - \xi_i)^2 \xrightarrow{\{\xi_i\}, \{\alpha_j, \beta_j\}} \min \\ \text{s.t. } \sum_{ij} (y_{ij} - \bar{y}_i)^2 = \text{const} \end{cases}$$

$\Leftrightarrow$
Kumasaka & Shibata 2008

$$A\hat{\beta} = \lambda B\hat{\beta}$$



$$A = \begin{pmatrix} C_{11} & \cdots & C_{p1} \\ \vdots & \ddots & \vdots \\ C_{1p} & \cdots & C_{pp} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$B = \begin{pmatrix} C_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & C_{pp} \end{pmatrix}$$

TextilePlot in genetics (disentangler)

LETTERS

nature
genetics

Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese

Yukinori Okada^{1,2}, Yukihide Momozawa³, Kyota Ashikawa³, Masahiro Kanai¹, Koichi Matsuda⁴, Yoichiro Kamatani², Atsushi Takahashi² & Michiaki Kubo³

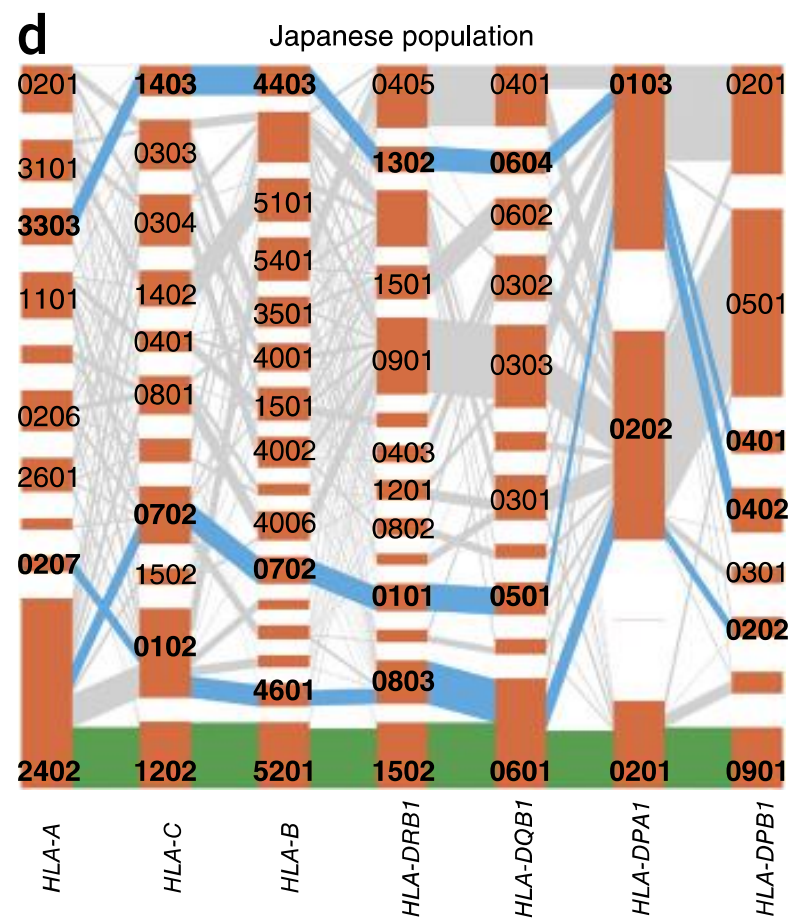
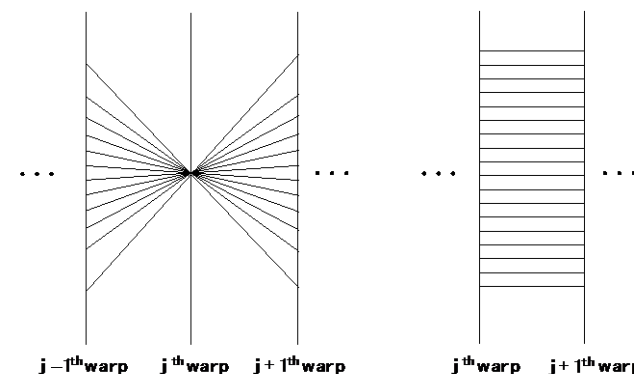


Figure 1d (Okada *et al.* 2015 *Nat. Genet.*)

Mathematical/Statistical interpretation

- Multiple Correspondence Analysis / PCA / ANOVA (Kumasaka & Shibata 2008, 2009)

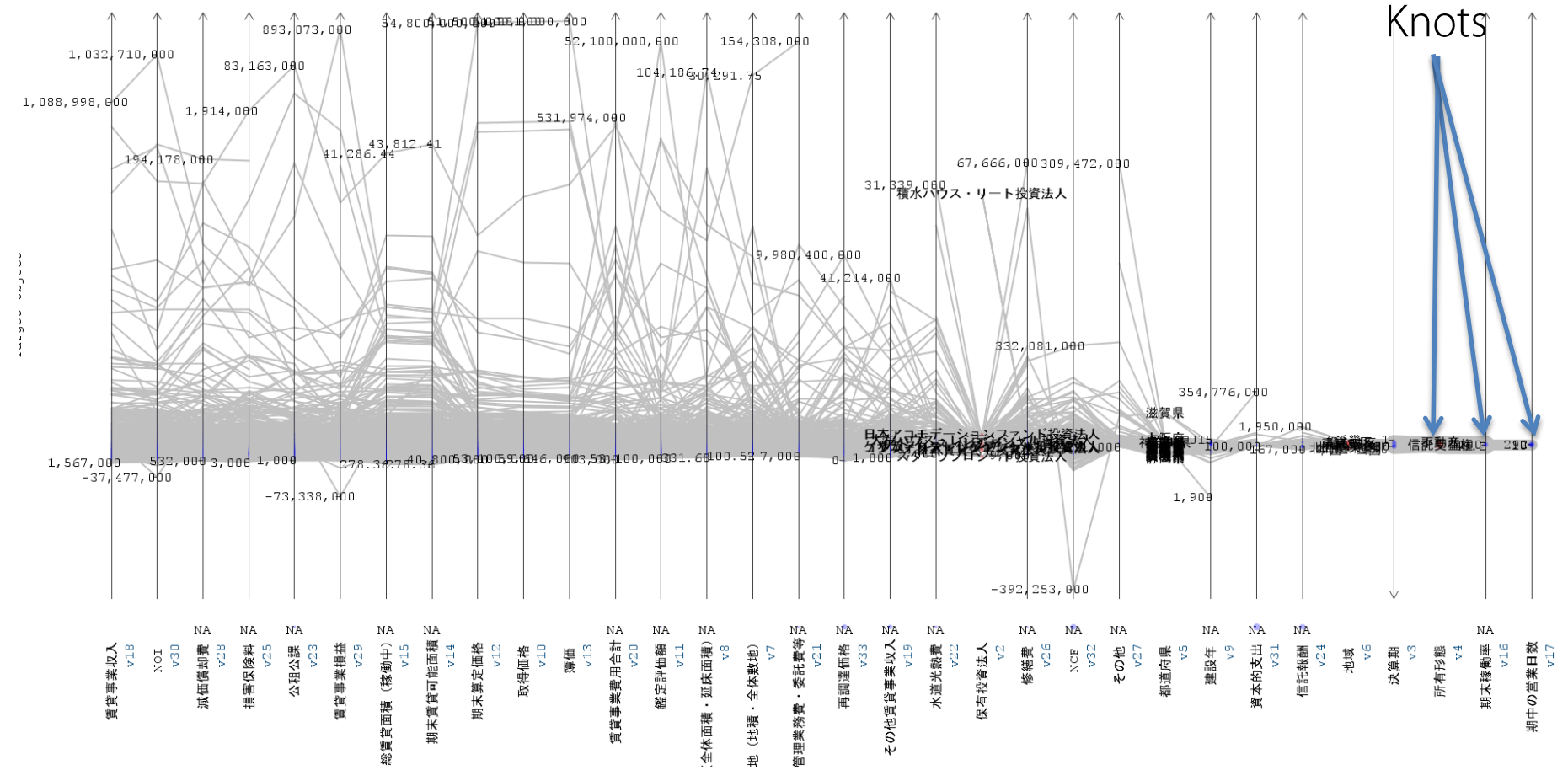
- Knot / parallel wefts



- Differential geometry (Sei & Tanaka 2015)

$$A\hat{\beta} = \lambda B\hat{\beta} ?$$

Knot on TextilePlot



Combination of TextilePlot and OGI plot can improve interpretation

Rejoinder to discussions welcome

- Is *data* really given?
- How can we transfer technologies in one discipline to another?
- Can TextilePlot be an ubiquitous technique in data science?