Reducing, selecting and leveraging structure in contemporary data – some reflections, looking ahead

Francesca Chiaromonte

PennState
ECoS Statistics
THE HUCK INSTITUTES OF THE LIFE SCIENCES

SSSA
IdE
Department of Excellence 2018 - 2022
EMbeDS
Economics and Management in the era of Data Science

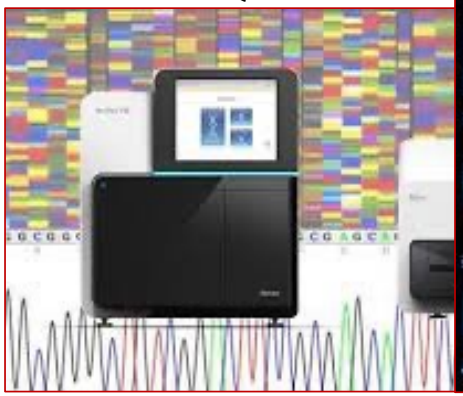
the bright side of contemporary data

Science, the economy, trade, public administration, policy & daily life – transformed by the availability of

massive, complex (and dirty) data

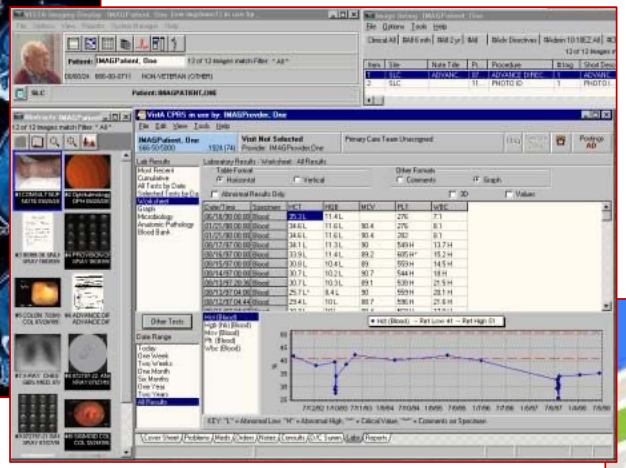
Fast-evolving **high-throughput technologies**, **online activities**, our own expanding **digital traces**:

NEXT GEN SEQUENCING

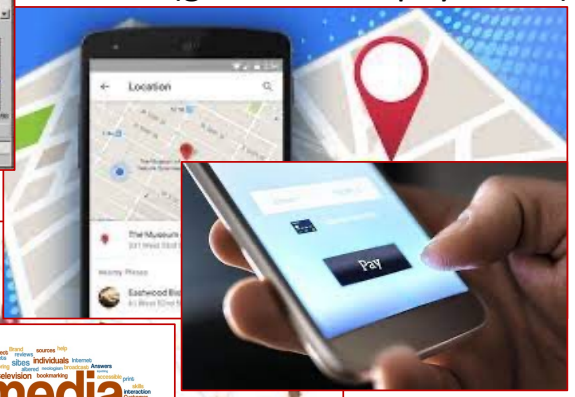


MEDICAL IMAGING

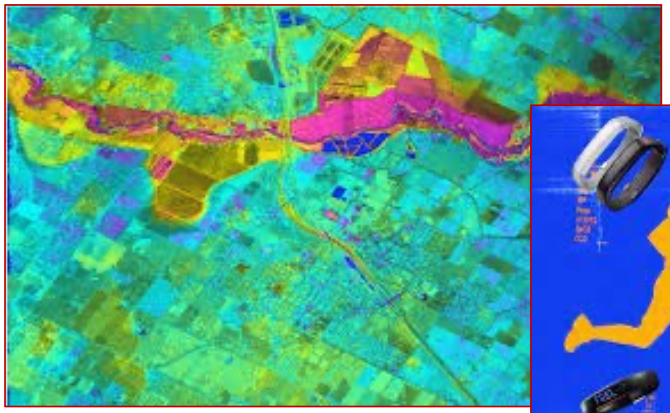
ELECTRONIC HEALTH RECORDS



APPS (geolocation, payments)



REMOTE SENSING



WEARABLE DEVICES

SOCIAL MEDIA

Sieving through & making sense of massive, complex data: my cross-disciplinary journey

ABMs
- DESIGN, STATISTICS
- USE IN ECONOMICS

METEO
- TROPICAL STORMS CHARACTERIZATION
- FORECAST, ENSAMBLES

ECON
- SOCIO-ECONOMIC IMPACTS OF CLIMATE CHANGE & DISASTERS
- SCIENCE, TECH & INNOVATION

HUMAN DISEASE
- mtDNA AND MATERNAL AGE
- MULTIPLE "OMICS" OF CHILDHOOD OBESITY, AND MORE…

"OMICS"
- HOW GENOMES CHANGE
- HOW GENOMES FUNCTION

PEERS

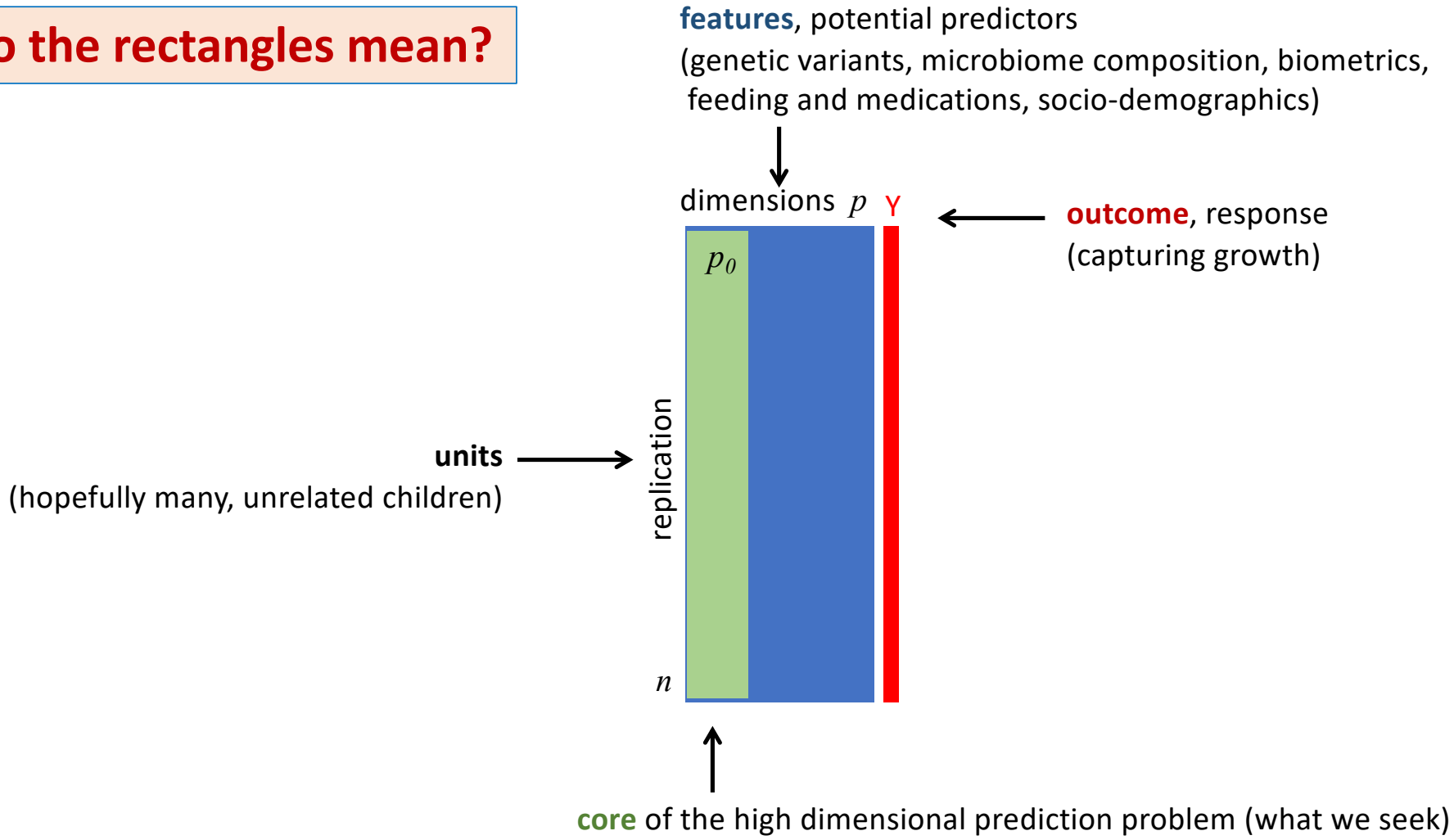
(virtuous) loop:

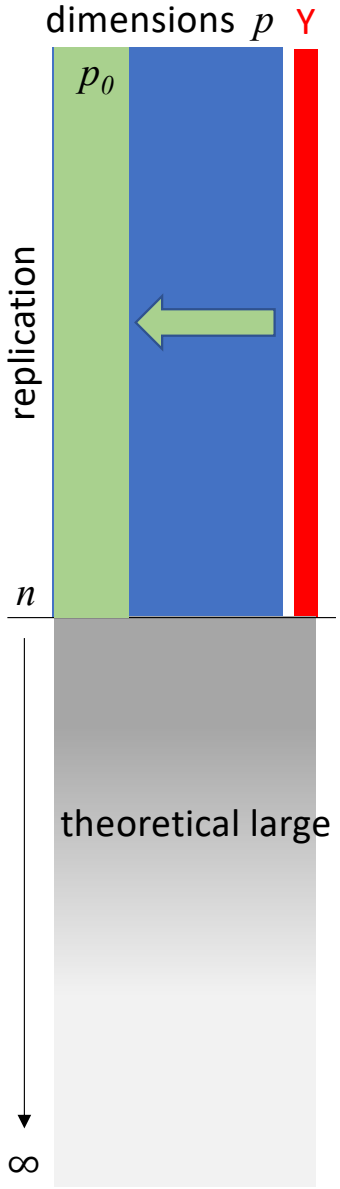
METHODS DEVELOPMENT
- REDUCE LARGE, COMPLEX DATA
- INTEGRATE DIVERSE DATA, LEVERAGE STRUCTURE
- REPRODUCIBILITY OF ANALYSIS PIPELINES, STABILITY OF RESULTS, SCALABILITY OF TECHNIQUES

time

reducing, selecting, leveraging structure

What do the rectangles mean?





SUFFICIENT DIMENSION REDUCTION
(low-dimensional core)

$$Y \perp X \mid P_S X \quad i.e. \quad Y|X \sim Y|P_S X$$

$S_{Y|X} = \cap S$ **Central Subspace (CS)**

R.D. Cook *et al*.
K.C. Li *et al*.
Bing Li *et al*. … and many more

FEATURE SELECTION
(sparse core)

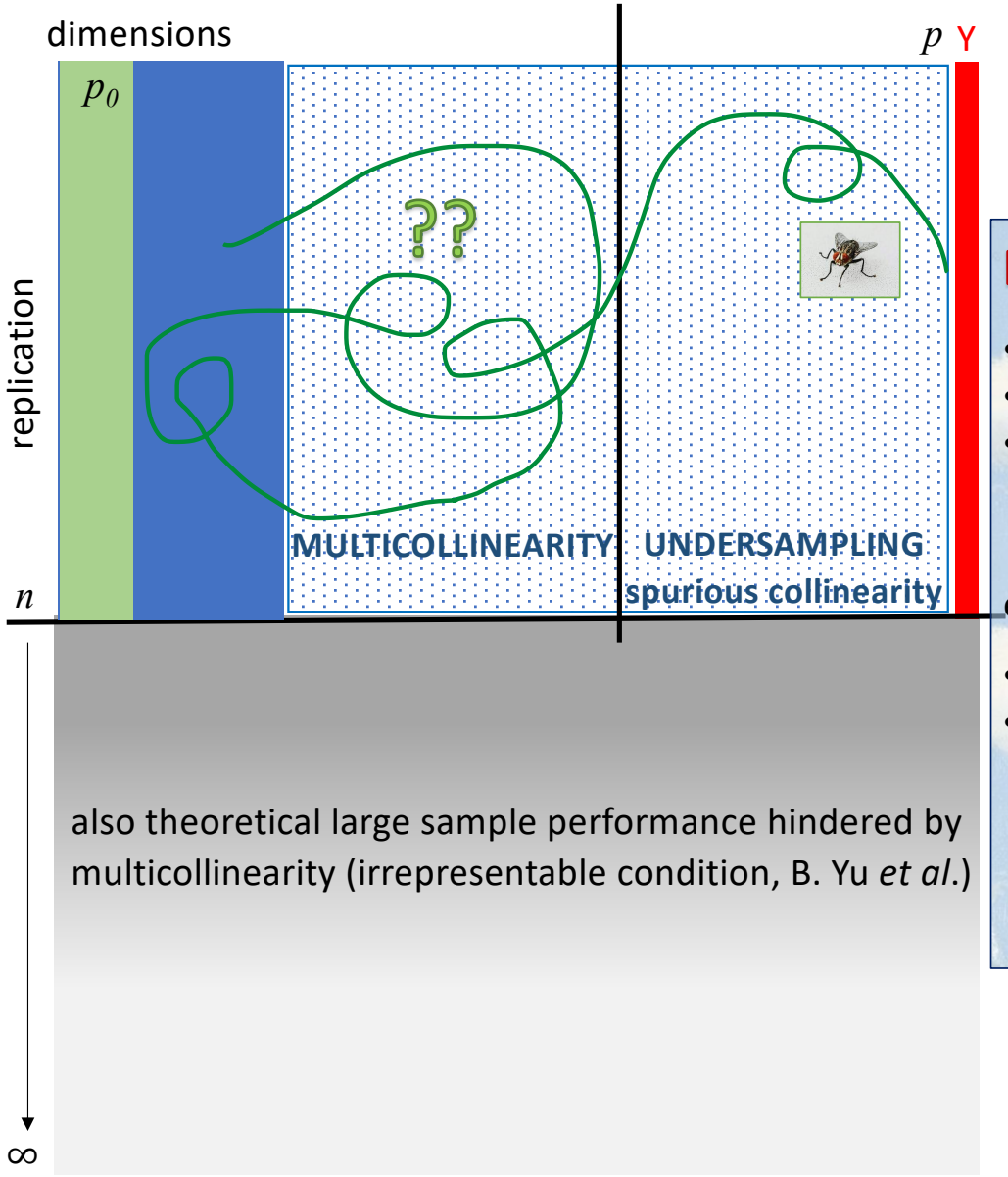
$$\hat{\beta}_{LASSO} = \text{argmin}\left\{ \| \underline{Y} - \underline{X}\beta \|^2 + \lambda \| \beta \|_{(1)} \right\}$$

T. Hastie, R. Tibshirani *et al*. … and many more

non-convex penalties
J. Fan *et al*.
Runze Li *et al*. … and many more

L_0 penalty (Best Subset) Mixed Integer Programs
D. Bertzimas *et al*. … and many more

theoretical large sample performance



also theoretical large sample performance hindered by multicollinearity (irrepresentable condition, B. Yu *et al.*)

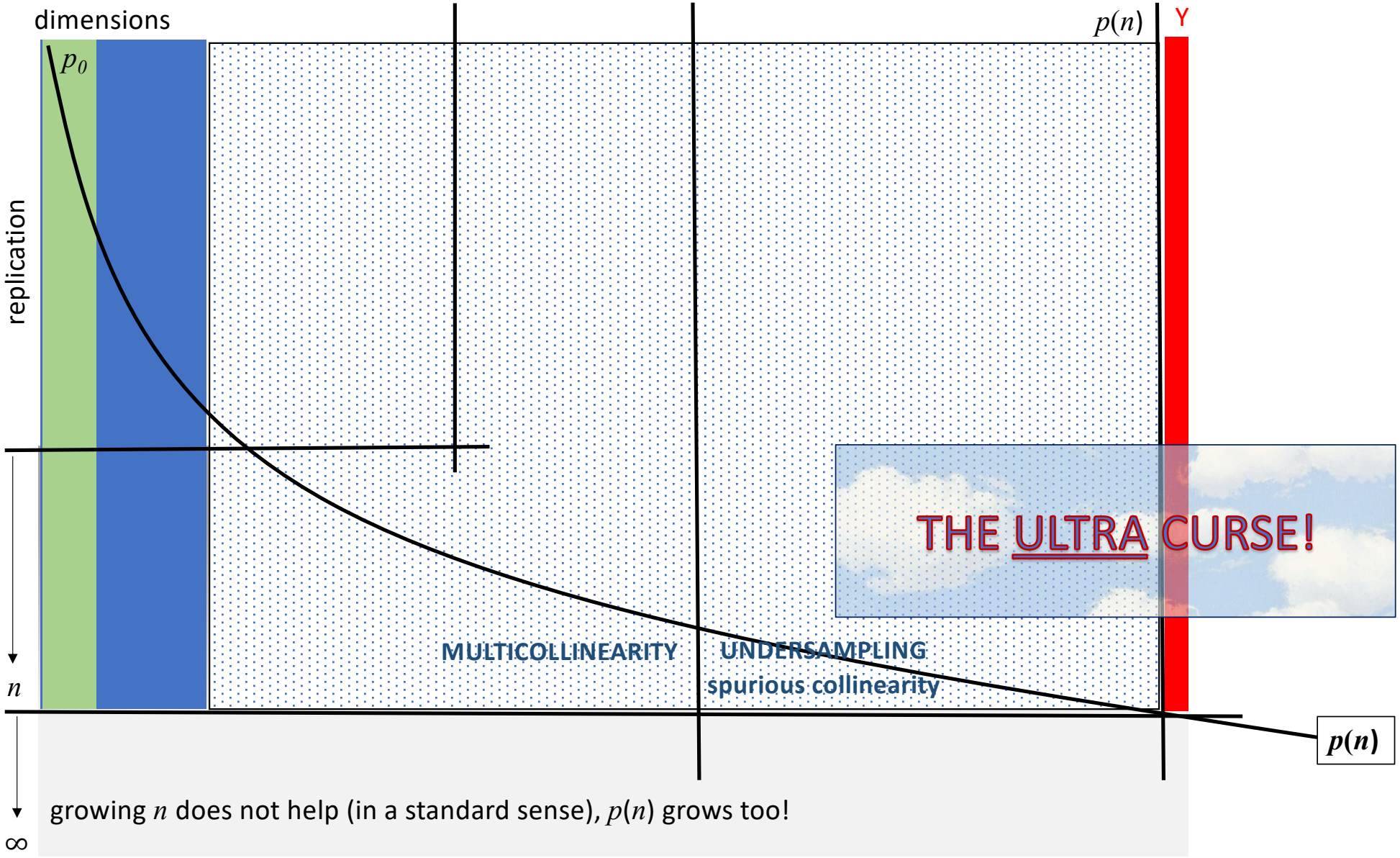
If p grows very large

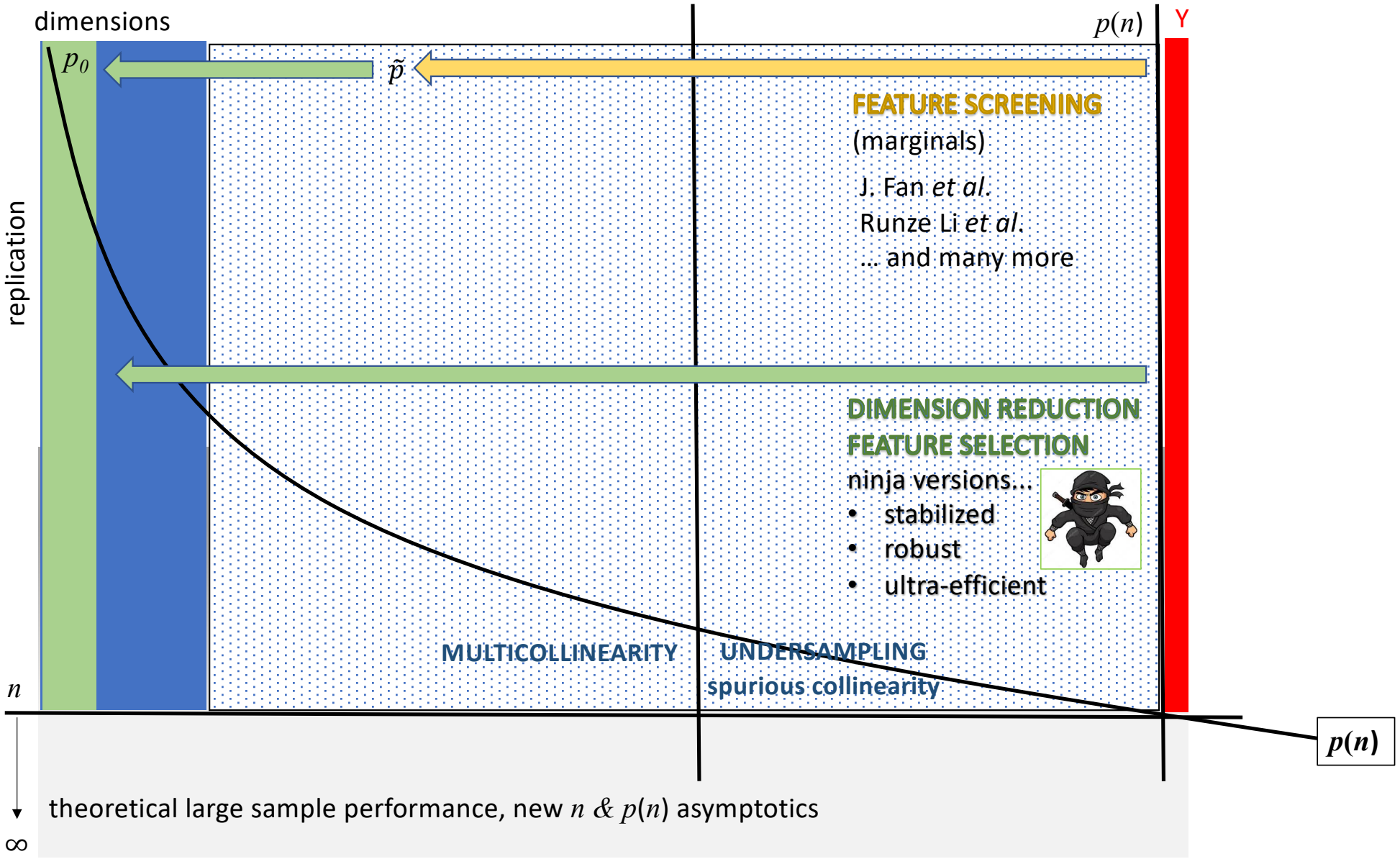
- multicollinearity becomes more likely
- $p > n$, undersampling creates spurious collinearity
- true and spurious collinearities create **instability** in selection, estimation and prediction (… variance inflation, overfitting)

even procedures to curb p can fail to gain traction

- **contaminations** become more likely
- for some procedures, **computational burden** becomes prohibitive

Rene Magritte, THE CURSE (c. 1963)


dimensions
$p(n)$
Y
p_0
replication
n
∞
MULTICOLLINEARITY
UNDERSAMPLING
spurious collinearity
THE ULTRA CURSE!
$p(n)$
growing n does not help (in a standard sense), $p(n)$ grows too!


dimensions
$p(n)$
Y
p_0
$\tilde{p}$
FEATURE SCREENING
(marginals)
J. Fan *et al.*
Runze Li *et al.*
… and many more
replication
DIMENSION REDUCTION
FEATURE SELECTION
ninja versions…
• stabilized
• robust
• ultra-efficient
MULTICOLLINEARITY
UNDERSAMPLING
spurious collinearity
n
$p(n)$
theoretical large sample performance, new n & $p(n)$ asymptotics
∞

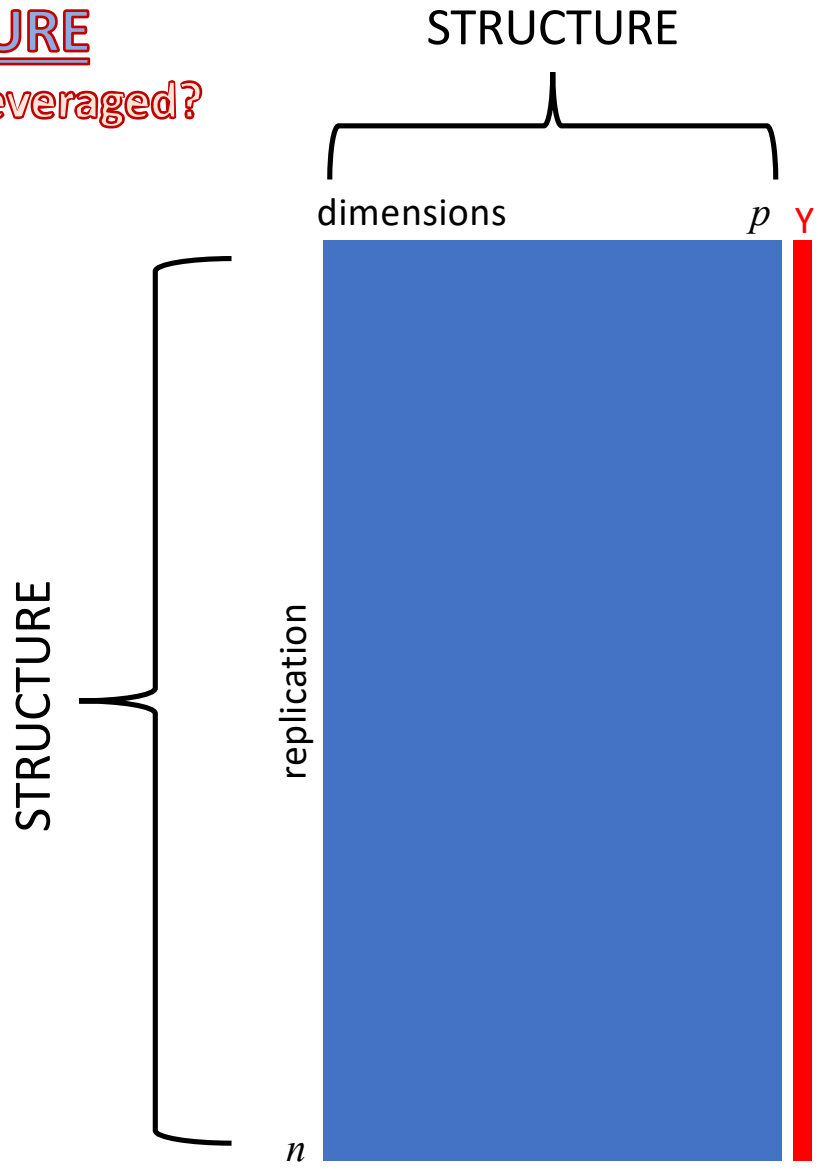
SDR and screens with (pseudo) Fisher Information

W. Yao, **D. Nandy**, B.Lindsay, F. Chiaromonte (2019) Covariate Information Matrix for Sufficient Dimension Reduction. JASA, 114(528) 1752-1764.
D. Nandy, F. Chiaromonte, R. Li (2022) Covariate Information Number for Feature Screening in Ultrahigh-Dimensional Supervised Problems. JASA, 117(539) 1516-1529.

L_0 selection with Mixed Integer Programs

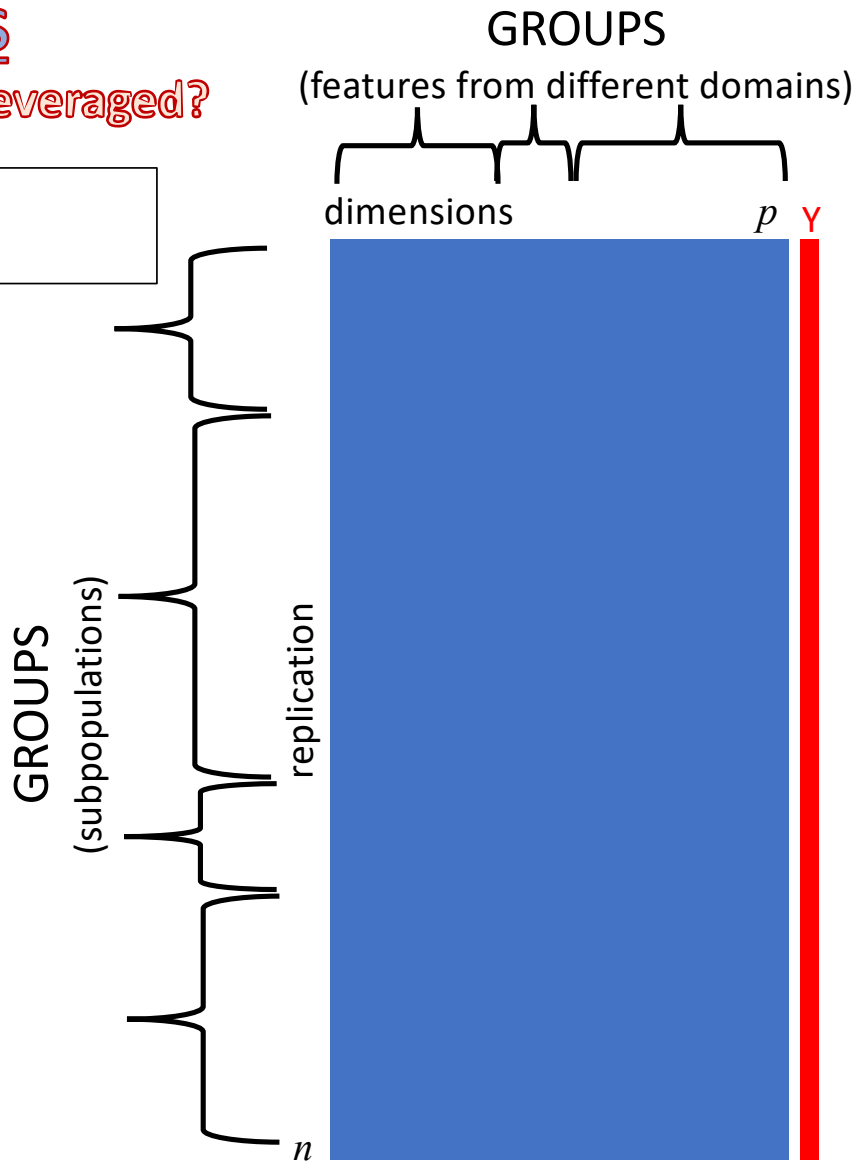
A. Kenney, F. Chiaromonte, G. Felici (2021) MIP-BOOST: efficient an effective L_0 feature selection for linear regression. JCGS, 30(3) 566-577.
L. Insolia, **A. Kenney**, F. Chiaromonte, G. Felici (2021) Simultaneous feature selection and outlier detection with optimality guarantees. Biometrics.

How can STRUCTURE be accounted for, leveraged?

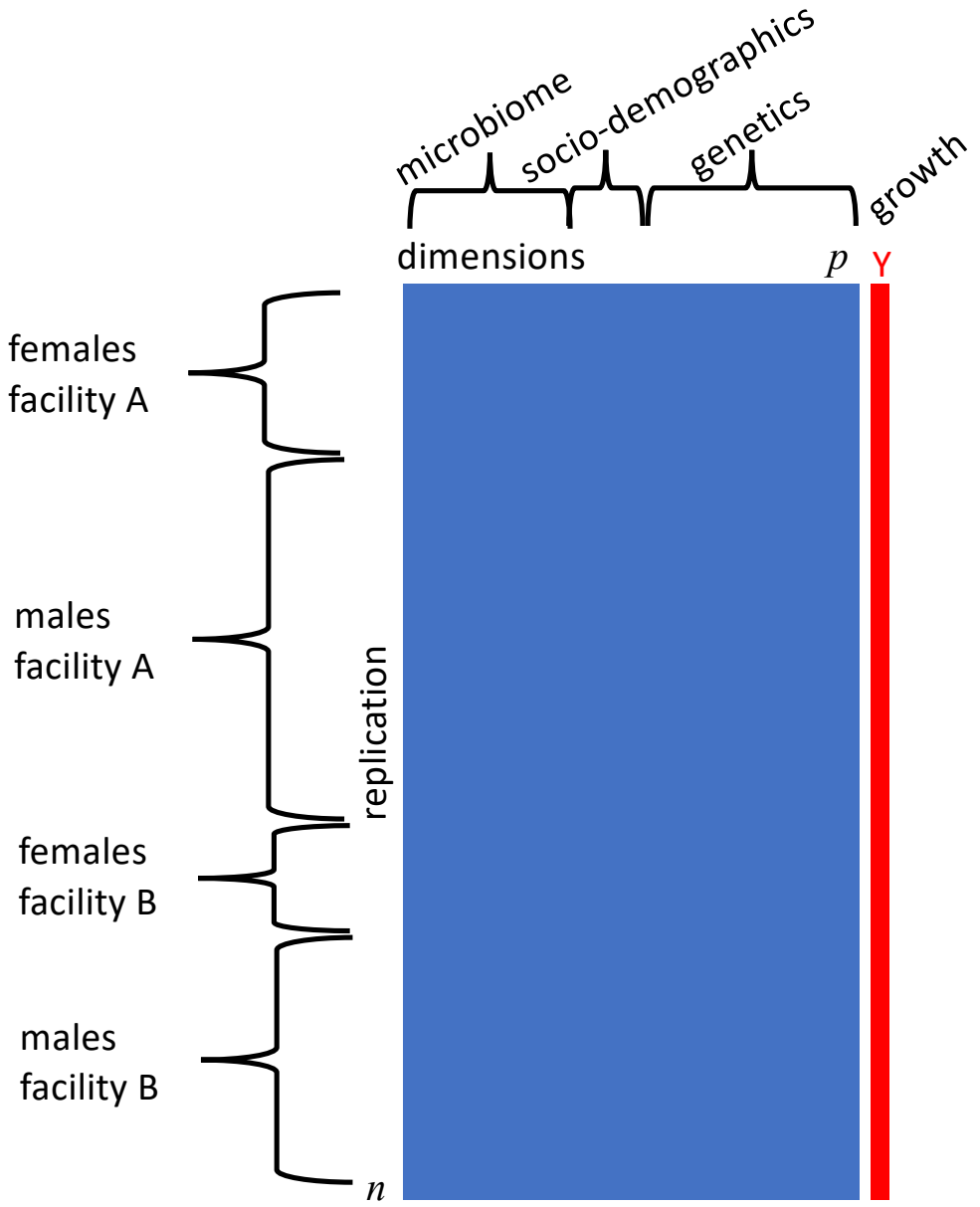


How can GROUPS be accounted for, leveraged?

DIMENSION REDUCTION
(low-dimensional core)



R.D. Cook *et al*.
Lexin Li *et al.*
Bing Li *et al*. … and many more

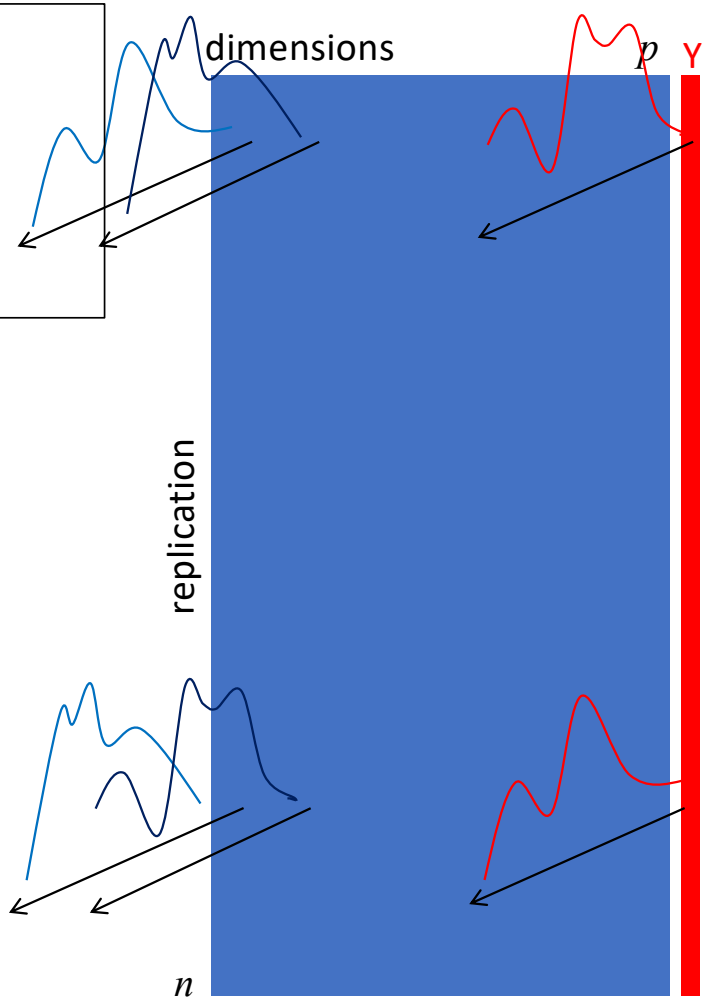


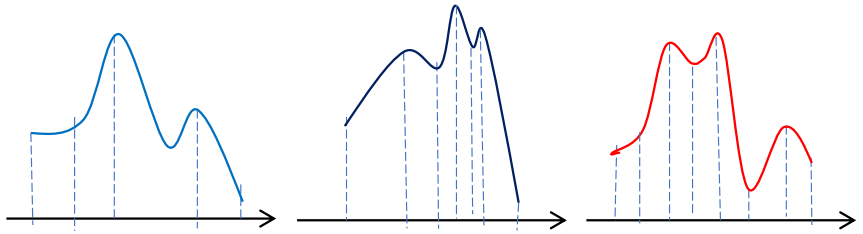
INTEGRATING DATA
from different sources

How can FUNCTIONAL DATA be accounted for, leveraged?

DIMENSION REDUCTION
(low-dimensional core)
FEATURE SELECTION
(sparse core)
FEATURE SCREENING
(marginals)

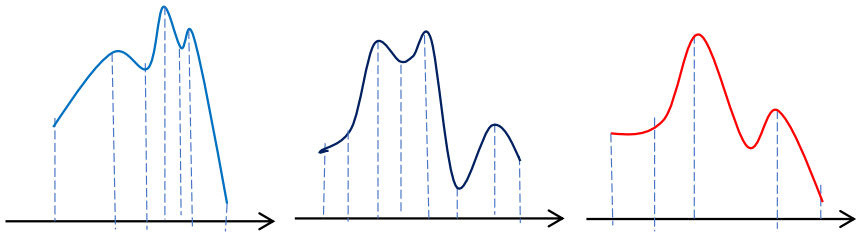
P. Secchi *et al.*
M. Reimherr, R. Li *et al.*
Bing Li *et al.* … and many more

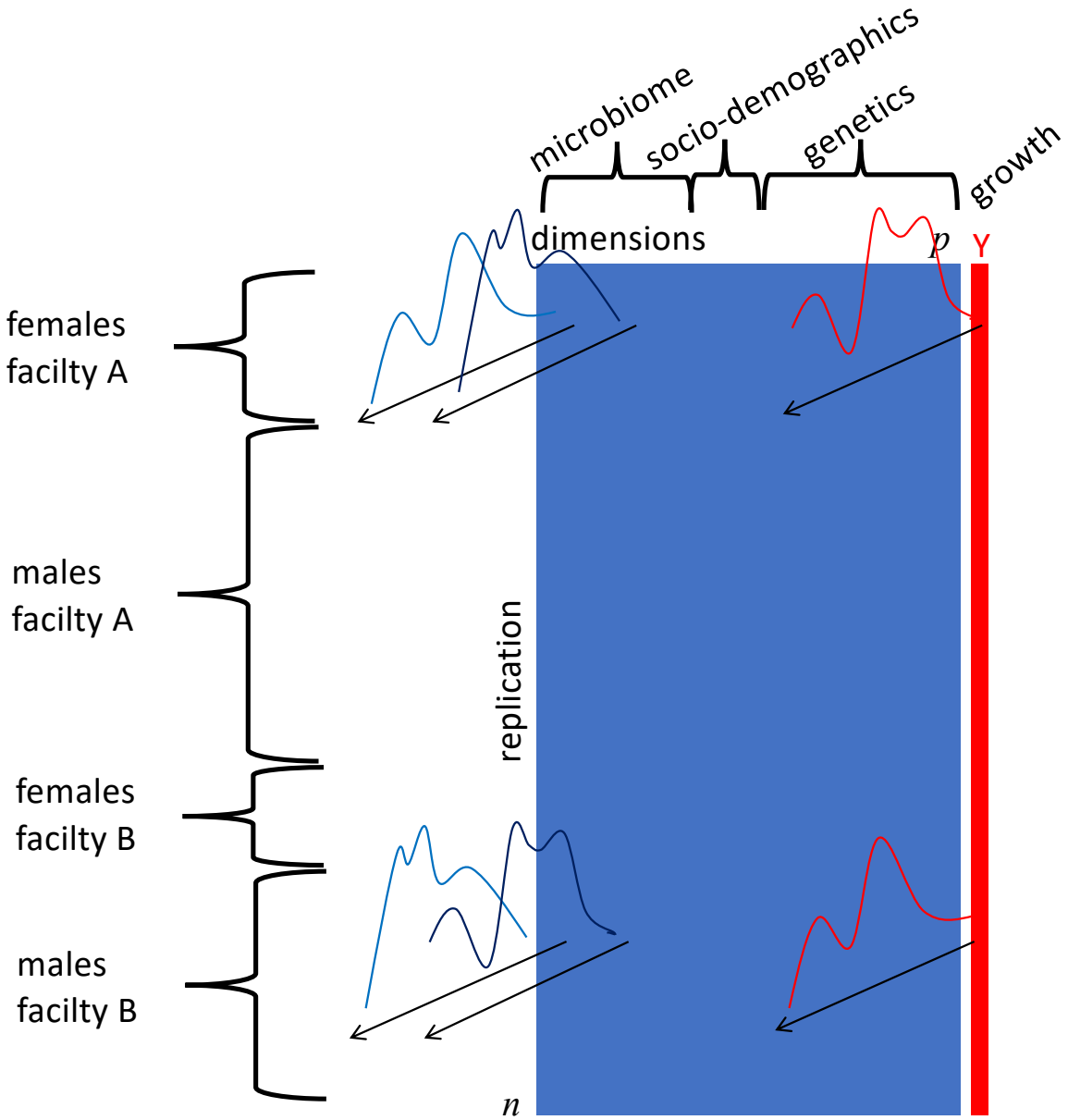




domains, resolution, location of measurements can vary across variables and units

SMOOTHING (denoise, create curves)
ALIGNMENT (parse vertical & horizontal variation)




microbiome
socio-demographics
genetics
growth
dimensions
p
Y
females
facilty A
males
facilty A
females
facilty B
males
facilty B
replication
n

SDR with groups

F. Chiaromonte, R.D. Cook, B. Li (2002) Sufficient dimension reduction in regressions with categorical predictors. AoS, 30(2) 475-497.
Y. Liu, F. Chiaromonte, B. Li (2017) Structured Ordinary Least Squares: A Sufficient Dimension Reduction approach for regressions with partitioned predictors and heterogeneous units. Biometrics, 73(2) 529-539.

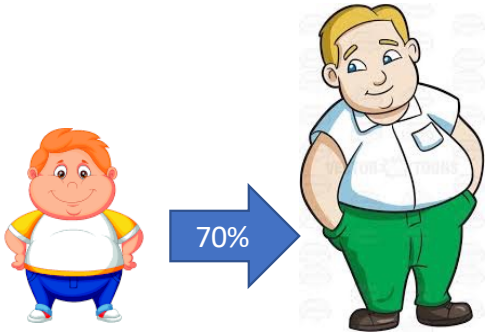
'Elastic Net' selection with Augmented Lagrangians, FDA

T. Boschi, M. Reimherr, F. Chiaromonte (2021) A Highly-Efficient Group Elastic Net Algorithm with an Application to Function-On-Scalar Regression. NeurIPS, 34 9264-9277.
T. Boschi, L. Testa, F. Chiaromonte, M. Reimherr (2022+) FAStEN: an efficient adaptive method for feature selection and estimation in high-dimensional functional regressions. (submitted manuscript)

does it work?
(applications)

the Childhood Obesity epidemic

- by 2030, 44% of Americans may be overweight or obese
- increasing prevalence in children
- more likely to become obese adults



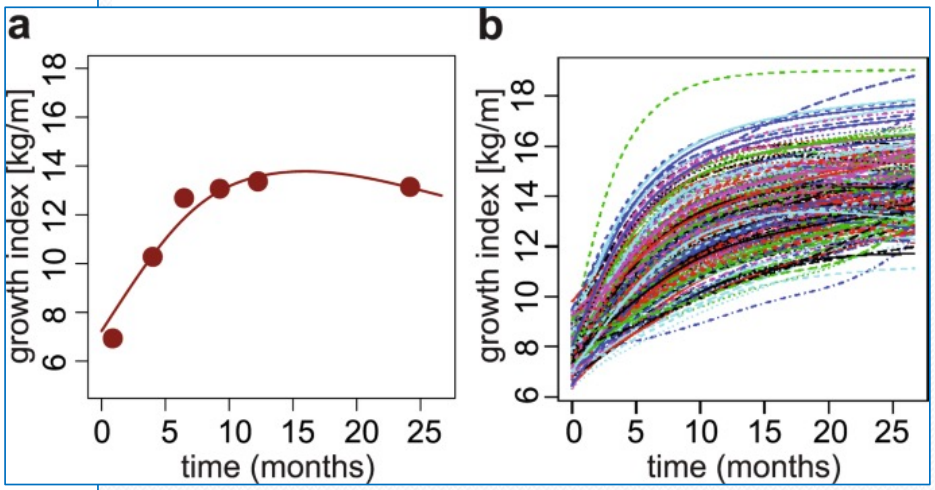
Using curves to analyze multifaceted data may hold the key to understanding childhood obesity

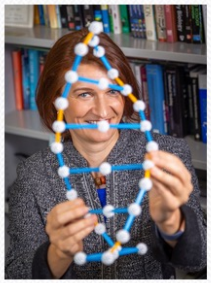
Gail McCormick

9 April 2020

Penn State News **Connecting the dots**

GROWTH CURVES



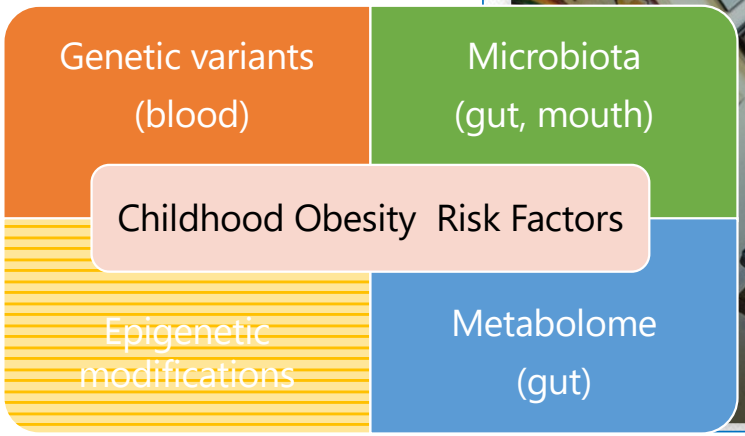


K. Makova

multiple 'Omics'

on INSIGHT cohort (I. Paul *et al.*)

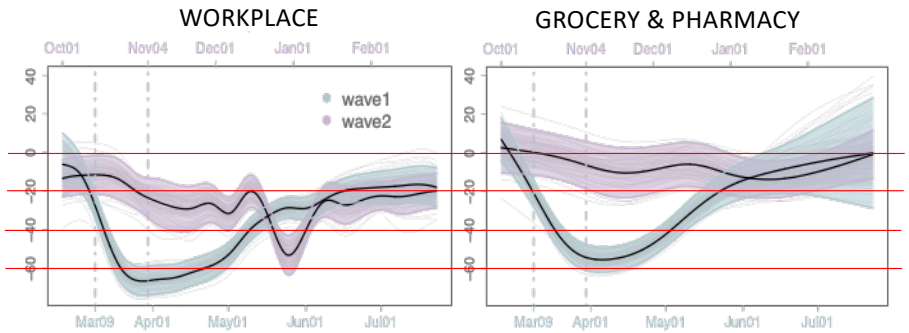
- p = 100,000s features (interdependent)
- n = 100s children
- covariates (feeding, meds, socio-demographics)



M. Reimherr, A. Kenney, S. Craig

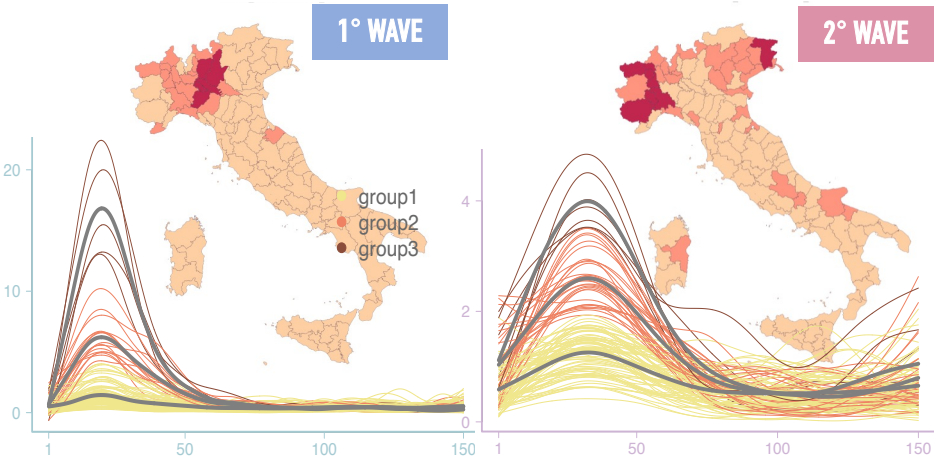
COVID-19: what explains the differences?

mobility



BANDS: pointwise average +/- 1.96*pointwise standard deviation

mortality



Penn State News **COVID-19 in Italy**

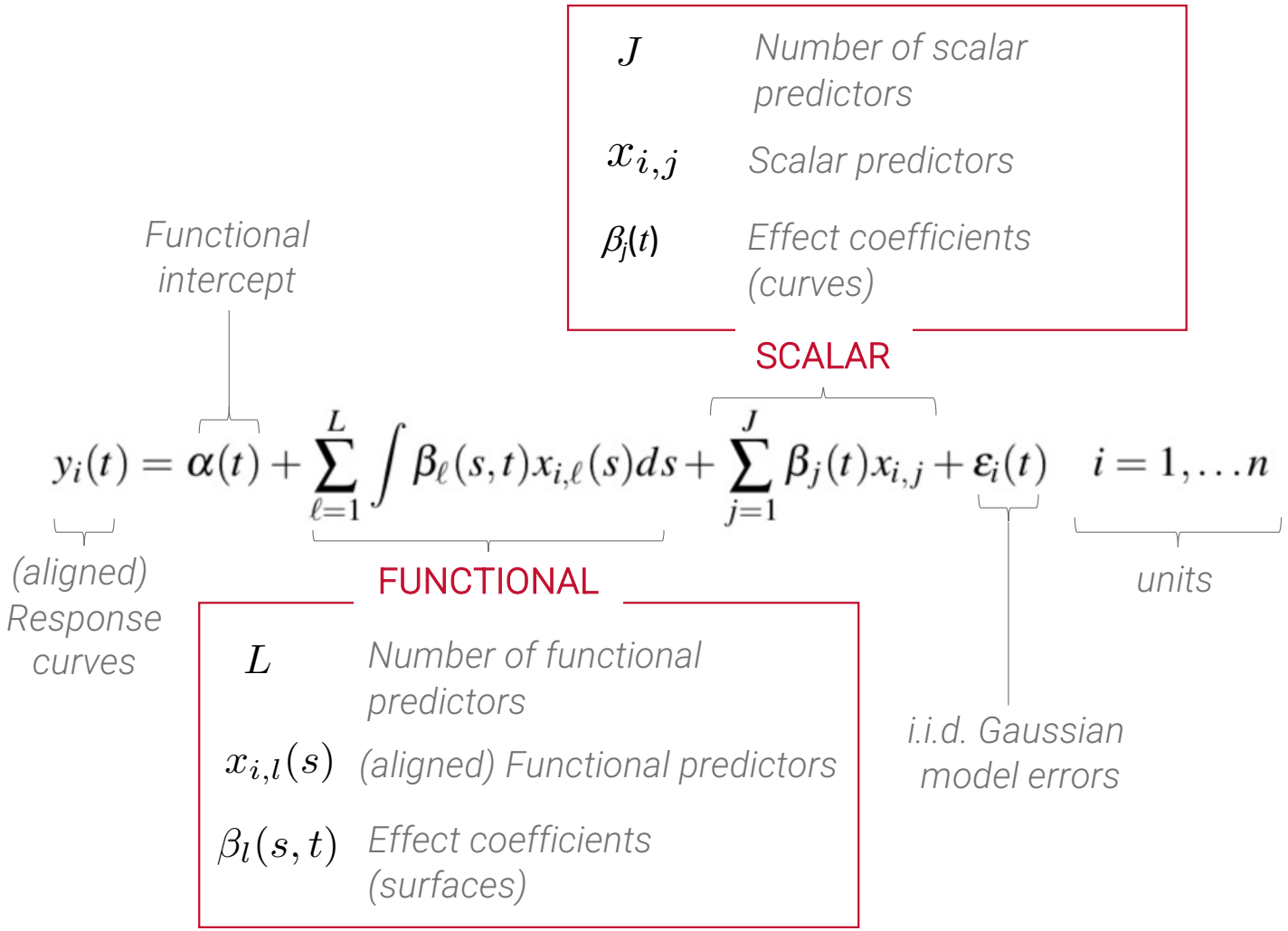
Staying home, primary care, and limiting contagion hubs may curb COVID-19 deaths

New study uses novel statistics to understand why some regions in Italy were hit harder than others during the first wave of the pandemic

New research reveals that staying home and limiting travel, supporting access to primary health care, and limiting contacts in contagion hubs like hospitals and schools helped reduce mortality due to COVID-19 during the first wave of the epidemic in Italy. **Credit: Gabriella Clare Marino, Unsplash. All Rights Reserved.**

COLLAPSE –

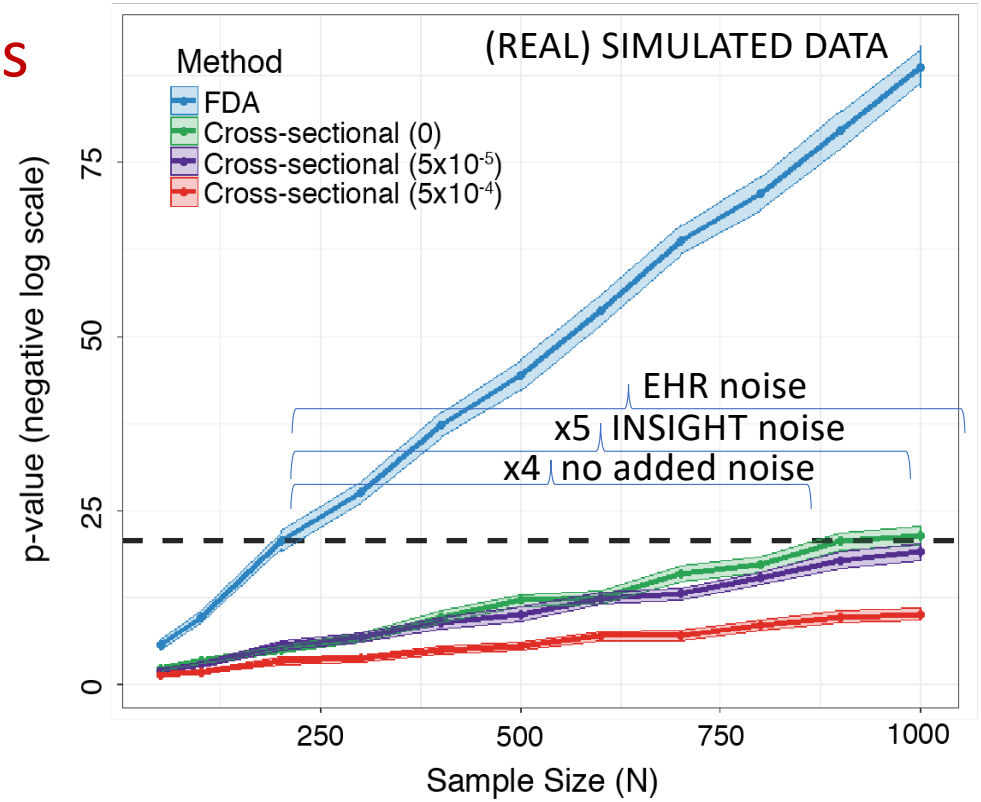
REGRESSION: function on function and/or scalar



Childhood Obesity and genetic variants

S. Craig, A. Kenney *et al*. (2021) Constructing a polygenic risk score for childhood obesity using functional data analysis. Econ & Stat.

new approach for Genome-Wide Association Studies on small cohorts… the power of rich, longitudinal data



parsimonious FDA PRSs (Polygenic Risk Scores; 20 SNPs, 5 SNPs); interpretable, and they can be used in practice!

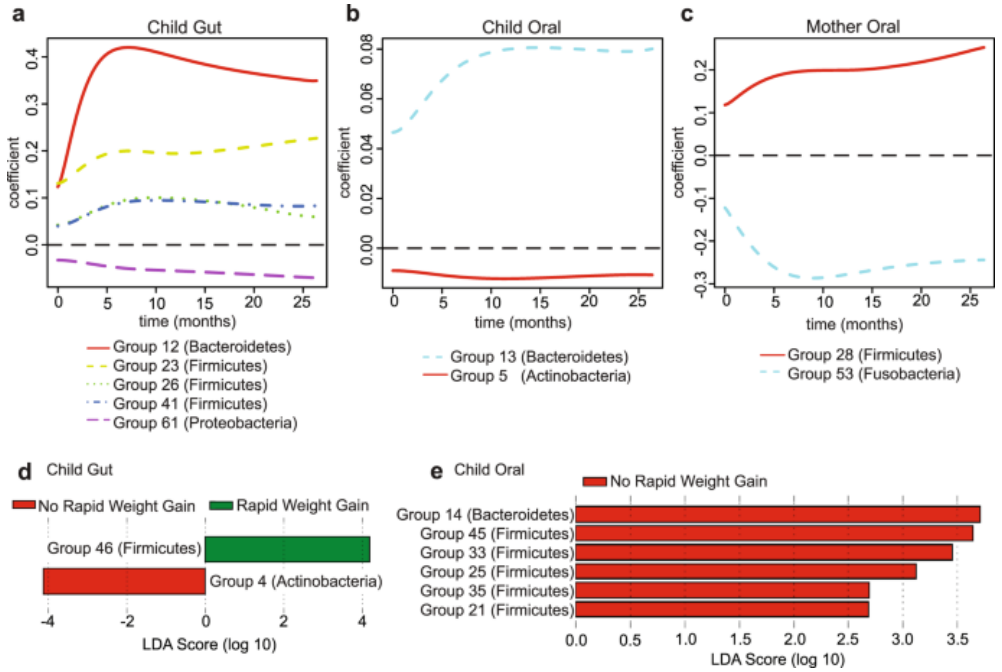
different signatures: adult PRSs do not validate in children, our FDA PRSs validate in older children and adults

- genetics of early life weight gain, which is reflected in adult obesity
- separate genetics of adult obesity, which is the dominant signal in adults, but is not reflected in early life

FDA PRSs remain dominant predictors of early life weight gain even controlling for environmental and behavioral factors

Childhood Obesity and microbiota

S. Craig *et al*. (2018) Child weight gain trajectories linked to oral microbiota composition. Scientific Reports, 8 1–14.



low diversity and high Firmicutes-to-Bacteroidetes ratios in a child's mouth (not gut) are signatures of early life weight gain

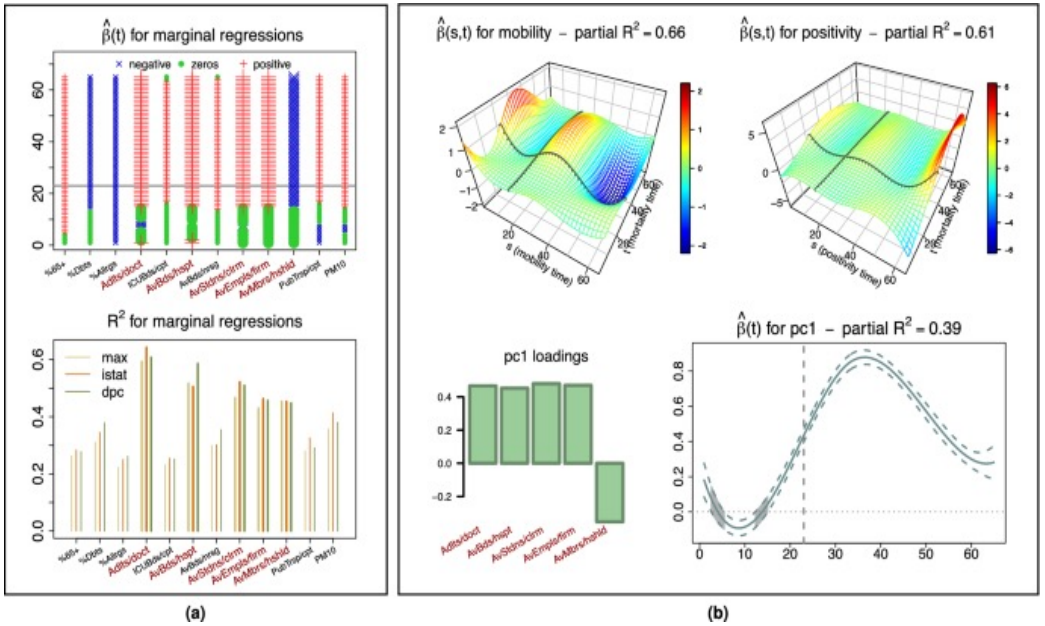
mirrored in the mother's mouth; oral microbiota of child and mother are similar (correlated diversity and F:B ratios)

typically found in the gut microbiome of obese adults; established in the mouth but not yet in the gut at 2y (a child's gut microbiome is not significantly predictive)

using Conditional Weight Gain (scalar), MIP-based simultaneous feature selection and outlier detection identifies ~10-15 bacterial group abundances in oral microbiota of child, incl. Bacteroidetes, and mother, incl. Fusobacteria (strongest outliers: children of mothers who smoked during pregnancy)

COVID-19 in Italy

T. Boschi, J. Di Iorio, L. Testa, M. Cremona, F. Chiaromonte (2021) Functional data analysis characterizes the shapes of the first COVID-19 epidemic wave in Italy. Scientific Reports, 11 1–15.



flowed epidemiological data (cases, hospitalizations, deaths)

many problems with data on potential predictors

2-3 different epidemic patterns unfolding in different areas of the country (exponential, flat(tened), intermediate)

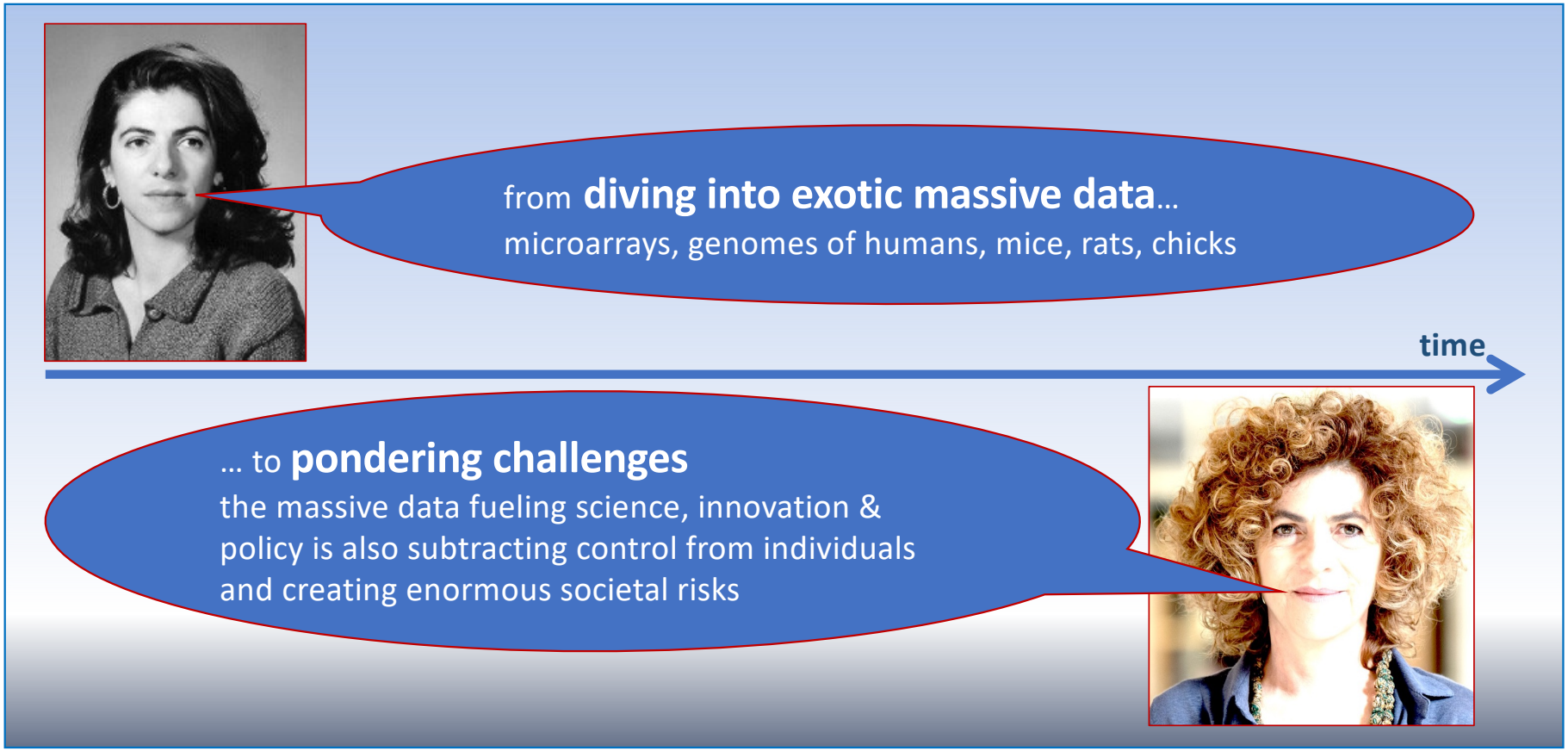
strong lagged positive association of mortality with mobility (early and mid-mobility on mortality at its peak) and positivity

distributed primary healthcare may mitigate mortality

hospitals, schools, and workplaces may work as contagion hubs

the darker side of contemporary data

The decades have delineated daunting methodological and social challenges



Are we (as scientists of the data) wise enough, and willing to take on our responsibilities?

Challenges ("keyword" dropping)

- scalability
- stability (reproducibility)
- power
- bias
- ~~interpretability~~
- privacy…

STATISTICS (when I went to school)

the science of characterizing (often **strong**) signals parsing through (**well understood**) errors

time

STATISTICS (now)

the science of characterizing (often **weak**) signals parsing through (**poorly understood**) spurious variability

I shall offer some thoughts…
if they sound familiar (obvious)

… I will be happy!

veridical data science

Bin Yu (UC Berkeley)

How good is your favorite statistical procedure?

traditionally **accuracy in predicting a response and in assessing effects**

recently **scalability**

- Sophisticated optimization, numerics; how does computational burden scale with the size of the data?
- Is this even an issue? computational power increases too

… more shortly (*it ain't the the cost of a single run!*)

more recently **stability**

- sensitivity to input data, general (some FDA tools) or under conditions (highly collinear regressions)
- sensitivity to tuning parameters
- but also, choices in data preprocessing

can generate instability in outcomes (cousin: **reproducibility**; aunties: **statistical robustness, influence**)

We evaluate accuracy and computational burden, but do we check if a procedure (or a pipeline, preprocessing + procedure + tuning) is stable?

… again, more shortly

surveillance capitalism

Shoshana Zuboff
(Harvard)

hackable humans

Yuval Noah Harari
(Hebrew U of Jerusalem)

power grabbing onto weak signals amid poorly understood spurious variability, more data or more sensitive procedures may not lead to more reliable (stable, reproducible) outcomes – in both prediction and effects assessment.

bias (shorthand) collection skews data wrt an intended analysis, biasing outcomes and hindering fairness – for instance, imbalanced classification/learning problems:

- fewer cases than controls available (a rare disease)
- over-representation of a subpopulation (male vs female, white vs non-white; clinical trials, credit worthiness)

privacy (shorthand) ethical and regulatory dimensions of data collection, governance, usage:

- who does the data belong to? Who can access them? Who benefits from them?
- who regulates their use? for research (academia), innovation and marketing (private sector), policy (public sector)

some thoughts looking ahead

1. Stress-test your procedure (really!)

importance of simulation experiments, even when you have wonderful theorems

SIMULATING ADVERSE CONDITIONS

generative models can be too benign, or legacy-driven

- explore systematically when and how fast performance deteriorates, when it breaks down
- what good is it if you are (somewhat) better than competitors and computationally viable with SNR=1000?
  … always know your SNR! (in close form, or numerically estimated)

SIMULATING REALISTIC GROUND TRUTHS

generative models can be too simplistic

- **real simulated data**, the new trend to render realistic complexity (interdependencies, structure, signal strengths)
- tweak a real dataset in a controlled fashion, for instance
  - spiking in (known artificial) signals
  - adding (known artificial) noise
  - removing (known) portions of information – features and/or units

  and try out your procedure under different tweaking scenarios

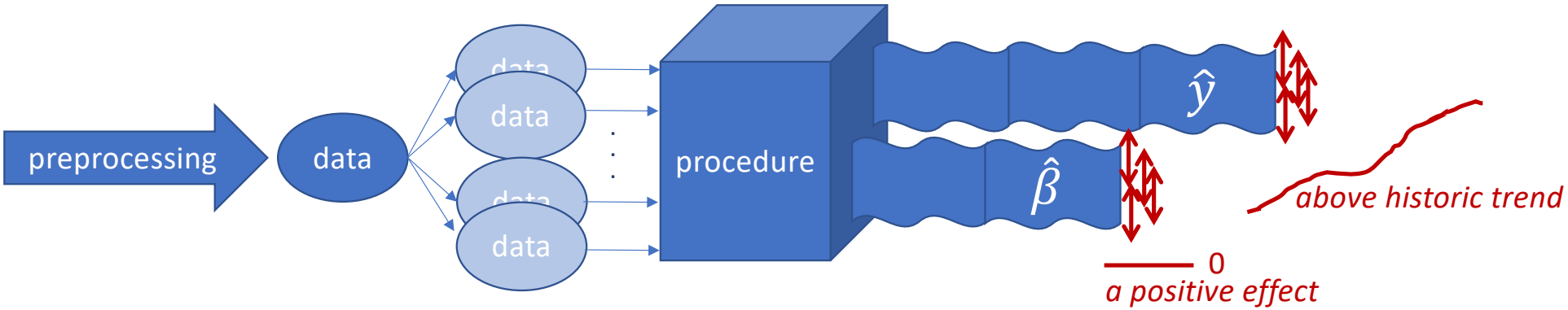
Controlling computational burden is of the essence:

… you will not properly stress-test if it takes too long to run the procedure

2. Shake it, shake it, shake it, shake it baby! … and repeat

- reducing dimension
- screening and selecting features
- downweighing/eliminating contaminations
- leveraging structure
- tuning

make for a good, hopefully stable box!



feed in repeated **perturbations*** of the data… and build ranges, see what survives

*resampling, subsampling, random deletions, superposition of noise, transformations

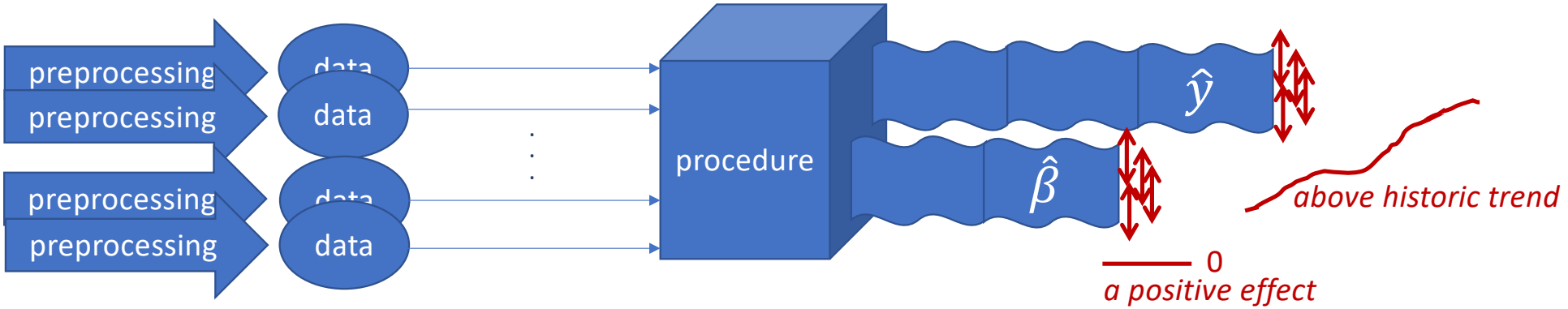
Controlling computational burden is of the essence:

… you will not properly explore perturbations if it takes too long to run the procedure

3. It's the ~~economy~~ preprocessing stupid! … so, repeat again

preprocessing (from the raw measurements to the data we run the procedures on) may matter more than the procedure itself!

critical issue in massive data driven **biomedical sciences** (omics, EHR, precision medicine) and **social sciences** (social media, mobility, consumption data)



feed in data produced by different preprocessing specifications … and build ranges, see what survives

Controlling computational burden is of the essence:

… you will not properly explore preprocessing if it takes too long to run the procedure

B. Yu, K. Kumbier (2020) Veridical Data Science. PNAS, 117 (8) 3920-3929.

4. Let's stop using (real) data?

built-in mitigation with SYNTHETIC DATA generated from real data

birds do it (computer scientists)…

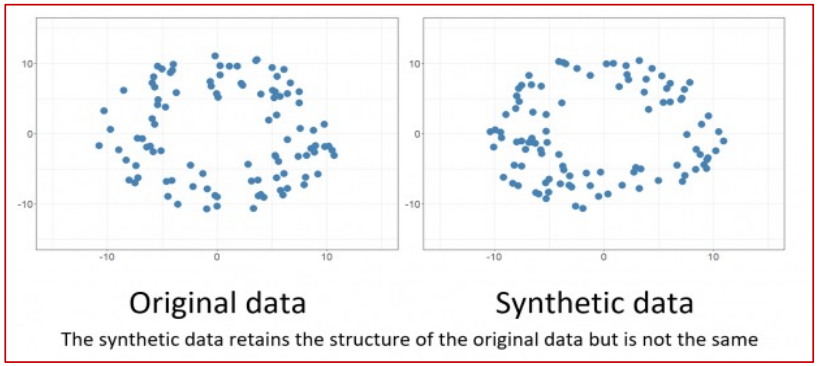
- **data augmentation for (deep) learning**: augment (w transformations) data to guarantee invariances, enrich learning
- **data re-balancing**: augment (w noise, transformations) or thin (w removals) data to balance classes/mitigate biases

… bees do it (statisticians)

- **bootstrap**: data perturbations to simulate/gauge variability due to chance in sampling
- **outliers/influence assessment**: techniques based on thinning data by removing points
- **regularization and shrinkage**: can be mimicked by data augmentation (w noise)
- **extended notions of stability**: data perturbations to simulate/gauge variability due to multiple mechanisms (sampling, also spurious sources – contaminations, batch effects, preprocessing specifications)

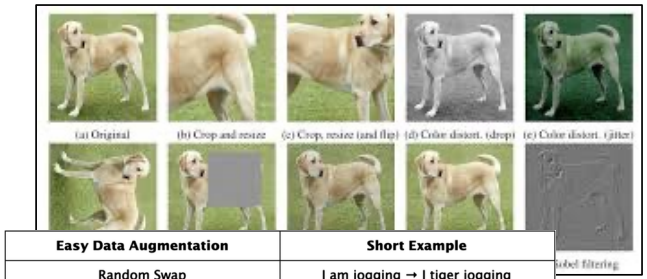
… even educated flees do it (governments, statistical offices)

- **privacy preserving data releases**:
  - noise data prior to release, or
  - fit models on data and release model-generated values

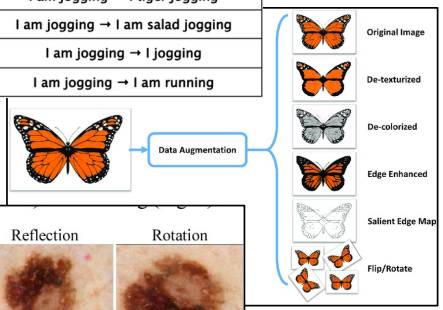


dataingovernment.blog.gov.uk

Data augmentation (Deep Learning)



Easy Data Augmentation	Short Example
Random Swap	I am jogging → I tiger jogging
Random Insertion	I am jogging → I am salad jogging
Random Deletion	I am jogging → I jogging
Random Synonym Replacement	I am jogging → I am running



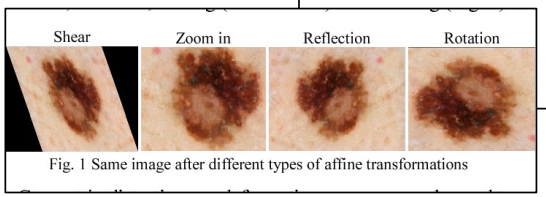
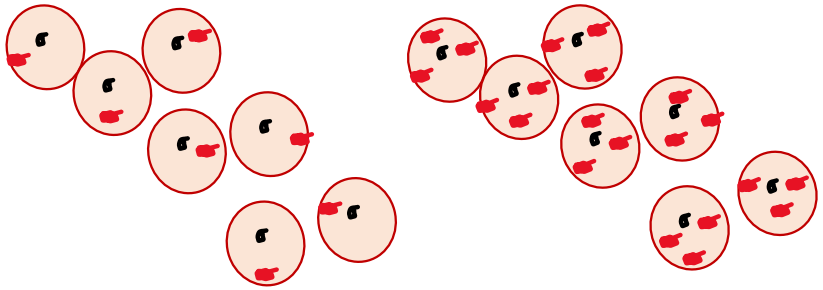


Fig. 1 Same image after different types of affine transformations

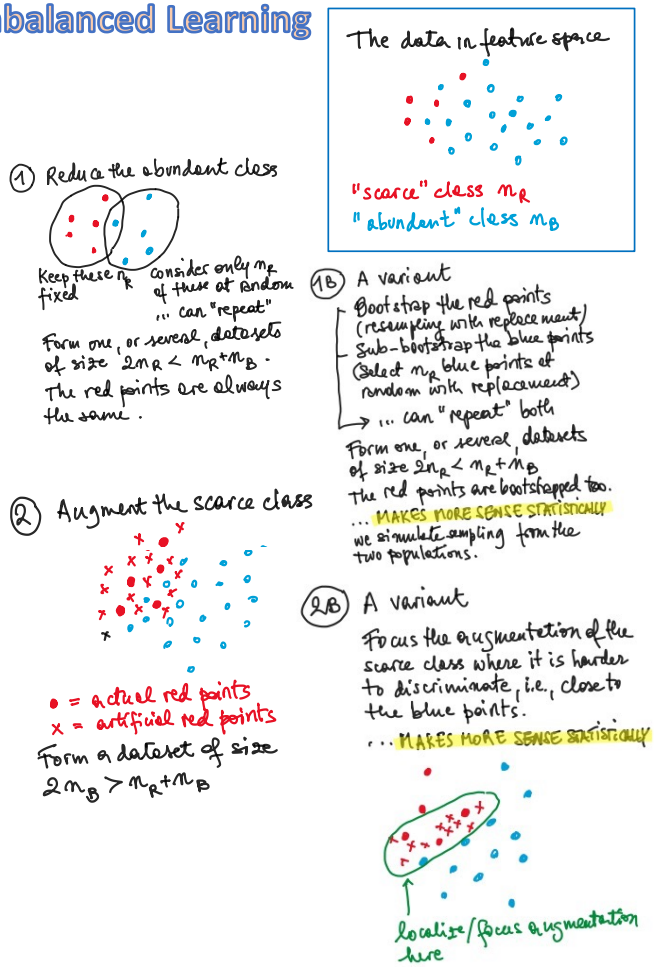
BOOTSTRAP

'Fudged' ($n \rightarrow n$) 'Augmented' ($n < p \rightarrow (k+1)n > p$)



S. Tyekucheva, F. Chiaromonte (2008) Augmenting the bootstrap to analyze high dimensional genomic data. Test, 17 1-18.

Imbalanced Learning



MIT tools (including SMOTE)
https://imbalanced-learn.org/stable/index.html

W. Fithian, T. Hastie (2016) Local case-control sampling: efficient subsampling in imbalanced data sets. AoS 42(5) 1693-1724

… let's do it, let's fall in love!

use SYNTHETIC DATA generated starting from real data as to

- preserve key signals, with some acceptable loss of information

but at the same time

- reduce the risks connected with privacy breaches
- control/mitigate biases
- regularize/robustify outcomes wrt overfitting and contaminations
- increase stability and reproducibility of outcomes wrt spurious variability and preprocessing specifications

switch to (quasi-equivalent) synthetic data

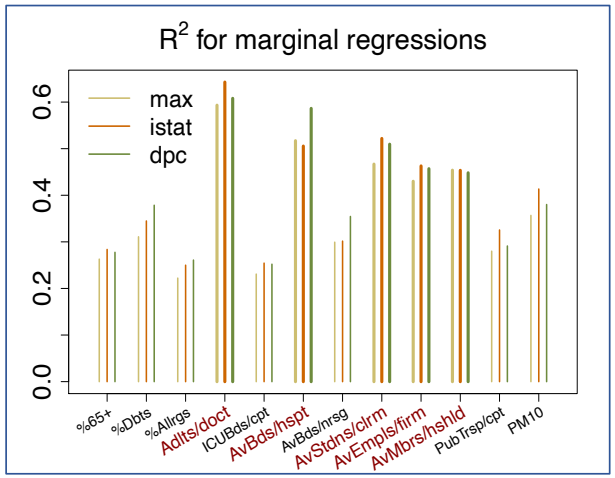
- prior to releasing data (the data owner has the real data)
- even during data collection/harvesting (the data owner never has the real data)

– CARBON FOOTPRINT ☹ !

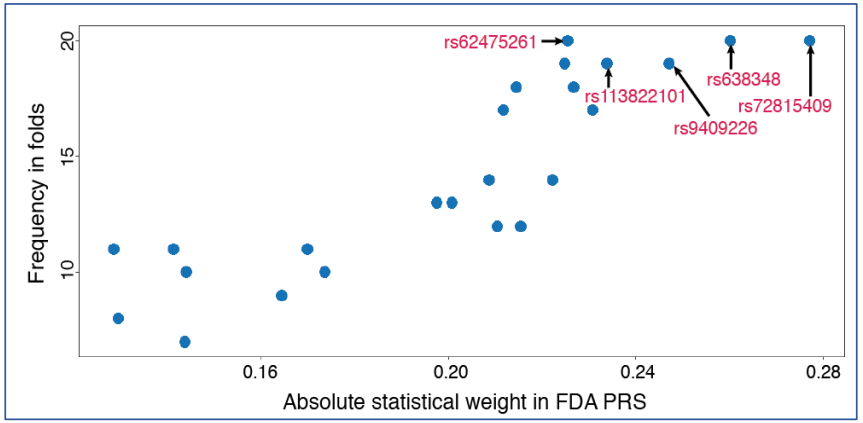
… and let's do it right!

unified theoretical framework, efficient operational approaches for synthetic data generation & use

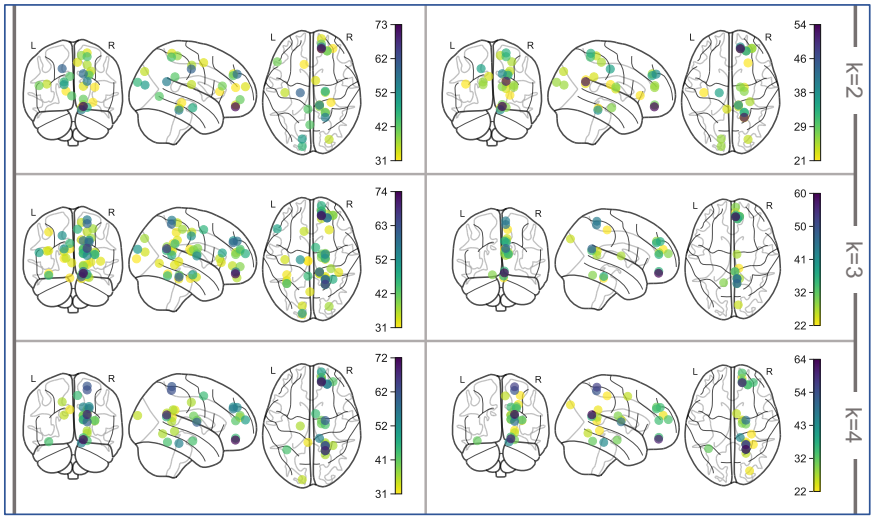
5. What are we going to tell the children?!?



CHILDHOOD OBESITY: repeatedly selected SNPs, growth prediction, fct-on-scal regr. (subsampling)

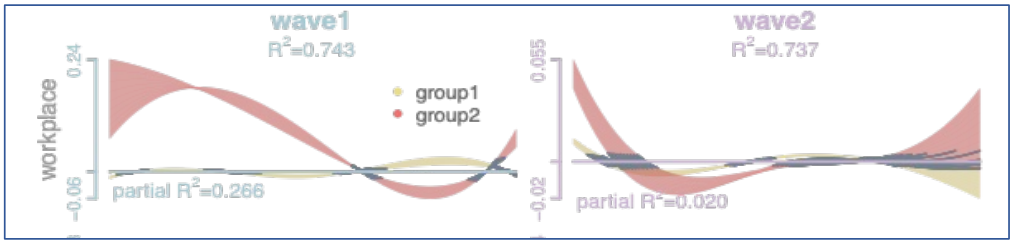


BRAIN fMRI: repeatedly selected voxels, heart rate prediction, fct-on-fct regr. (bootstrap; 2 mod. sel. criteria, 3 levels of curve detail – FPCA)



COVID-19 IN ITALY: 3 ways of generating mortality curves, fct-on-scal regr.

COVID-19 IN ITALY: 20 lags for mortality vs mobility, fct-on-fct concurrent regr., estimated effect curves



Acknowledgments (… they are way ahead of us!)

Giorgio Tripodi
SSSA/SNS, Northwestern

Prabhani K. Don
PSU, Harvard, URI, PSU

Luca Insolia
SSSA/SNS (PSU), U Geneva

Ana Kenney
PSU, UC Berkeley

Lorenzo Testa
SSSA (PSU), Bocconi

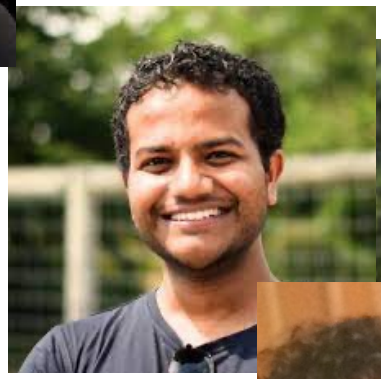
Simone Tonini
SSSA

Jacopo Di Iorio
PoliMi, SSSA, PSU

Tobia Boschi
PoliMi, PSU, IBM Research

Svitlana Tyekucheva
PSU, JHU, Harvard

Marzia Cremona
PoliMi, PSU, U Lavàl



Debmalya Nandy
PSU, U Colorado

Huy Dang
PSU, Moderna

James Taylor PSU, Emory, JHU – in memoriam

More Acknowledgments

all my extraordinary (young at heart) colleagues & collaborators!

funding from:

PennState
THE HUCK INSTITUTES
OF THE LIFE SCIENCES

NSF
National Science
Foundation

NIH
National Institutes
of Health

MIUR
MINISTERO DELL'ISTRUZIONE,
DELL'UNIVERSITÀ E DELLA RICERCA

Thank you!