Synthetic generation and data augmentation

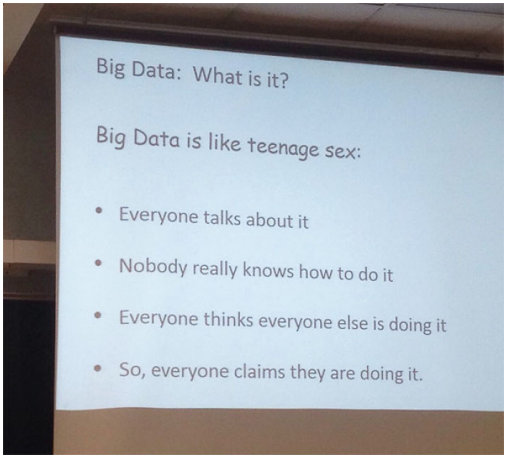
Enrico Giampieri

University of Bologna - enrico.giampieri@unibo.it

nextAIM - 2022-02-17

Big Data Science

Big Data Science



Back to the Future: Big Data Edition

Have we gone back to the Baconian Method?

1. collect a lot of data ... **ALL the data!**

Back to the Future: Big Data Edition

Have we gone back to the Baconian Method?

1. collect a lot of data ... **ALL the data!**
2. remove unrelated observations

Back to the Future: Big Data Edition

Have we gone back to the Baconian Method?

1. collect a lot of data ... **ALL the data!**
2. remove unrelated observations
3. stare at them **reeaally hard**

Back to the Future: Big Data Edition

Have we gone back to the Baconian Method?

1. collect a lot of data ... **ALL the data!**
2. remove unrelated observations
3. stare at them **reeaally hard**
4. ...

Back to the Future: Big Data Edition

Have we gone back to the Baconian Method?

1. collect a lot of data ... **ALL the data!**
2. remove unrelated observations
3. stare at them **reeaally hard**
4. ...
5. science?

Big Data ⇒ better results?

more data make you more certain, not more right

Big Data ⇒ better results?

more data make you more certain, not more right

if the data is biased . . .

Big Data ⇒ better results?

more data make you more certain, not more right

if the data is biased . . .

you get more certain of the wrong thing!

50 Shapes of data?

- wide data (many variables)
- long data (many subjects)
- deep data (time series)
- connected data (networks and relational databases)
- complex data:
  - unstructured data
  - context/domain dependent
  - interval data
  - missing data

Data entropy

let's not measure data by size, but information richness

Bayesian inception

no model has no assumption, let's not try to pretend otherwise

Bayesian inception

no model has no assumption, let's not try to pretend otherwise

make your assumptions explicit

Bayesian inception

no model has no assumption, let's not try to pretend otherwise

make your assumptions explicit

make your **knowledge** explicit

data augmentation

data augmentation

data augmentation

do our models respects our data?

A Noether Theorem for models?

all the problems have intrinsic invariants to them

A Noether Theorem for models?

all the problems have intrinsic invariants to them

they are often implicit, or unspoken

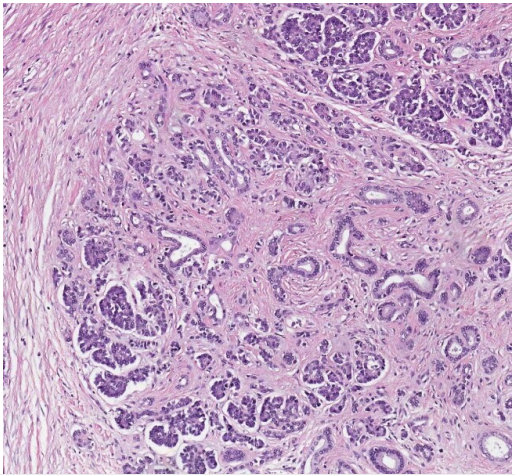
A Noether Theorem for models?

all the problems have intrinsic invariants to them

they are often implicit, or unspoken

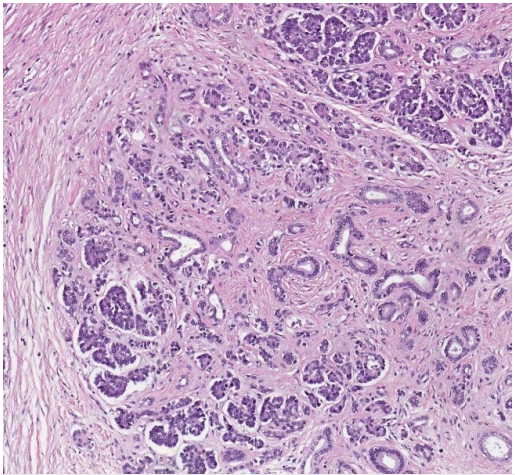
. . . or mispoken

the pathomic case



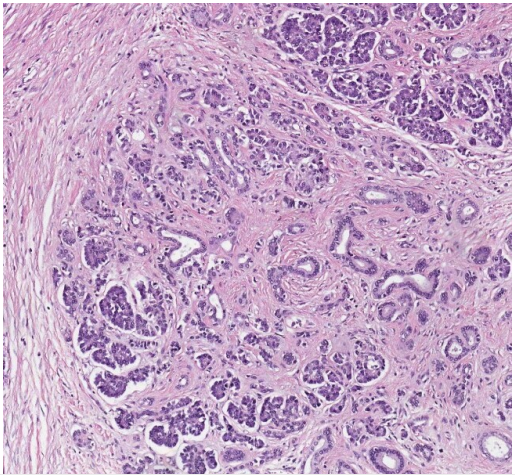
- invariance for rotation

the pathomic case



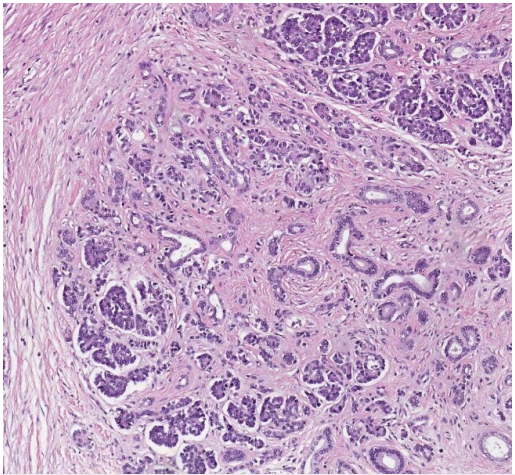
- invariance for rotation
- invariance for saturation/colorization

the pathomic case



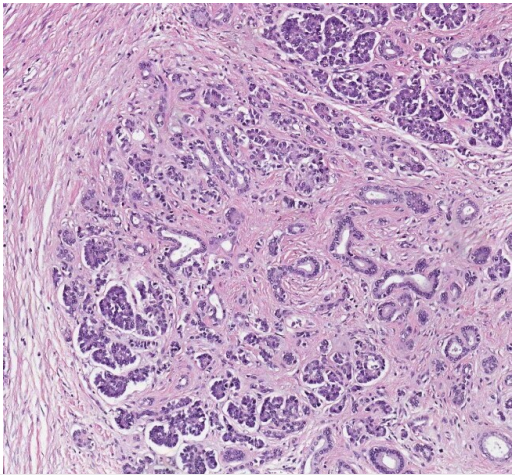
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity

the pathomic case



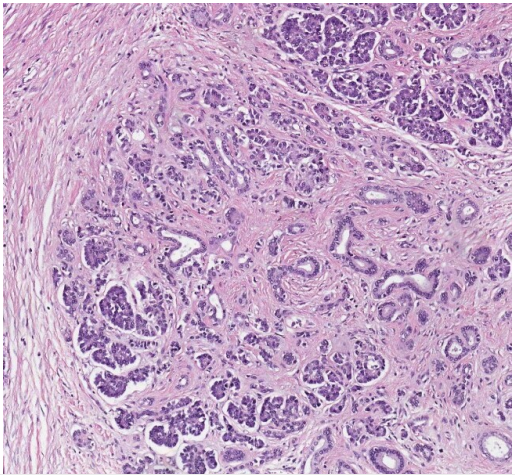
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast

the pathomic case



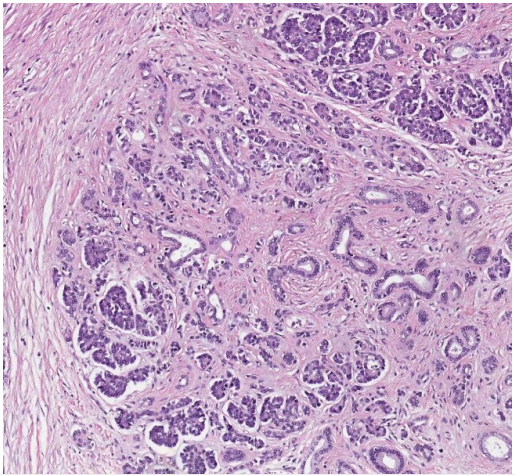
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection

the pathomic case



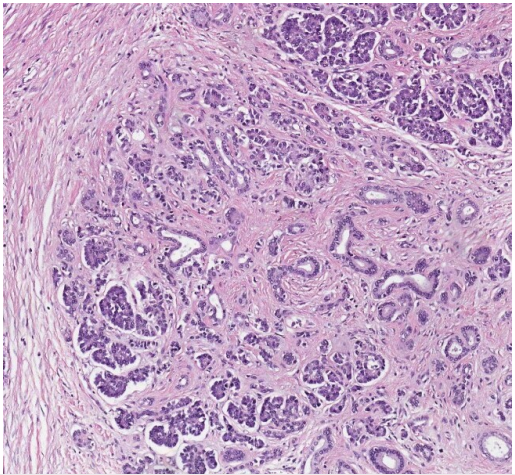
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection
- invariance (almost) for scaling

the pathomic case



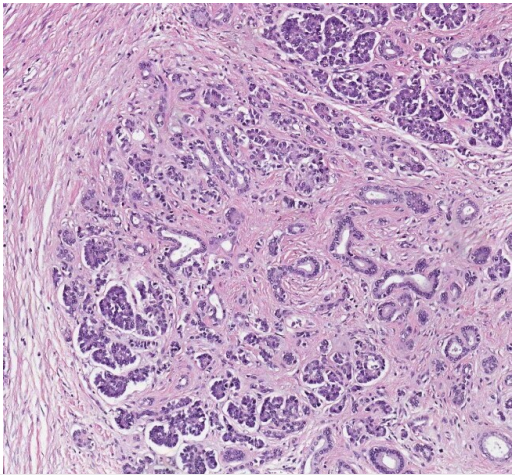
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection
- invariance (almost) for scaling
- invariance (almost) for deformation

the pathomic case



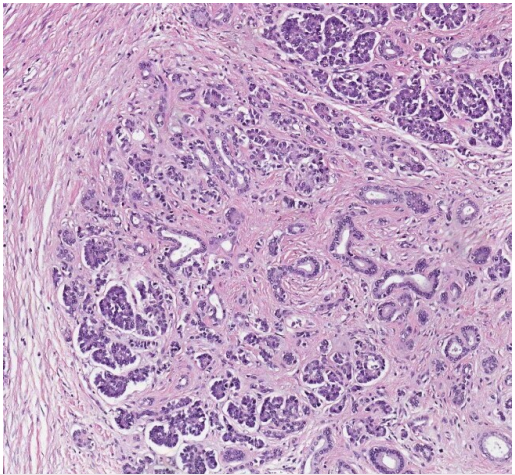
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection
- invariance (almost) for scaling
- invariance (almost) for deformation
- invariance (almost) for blurring

the pathomic case



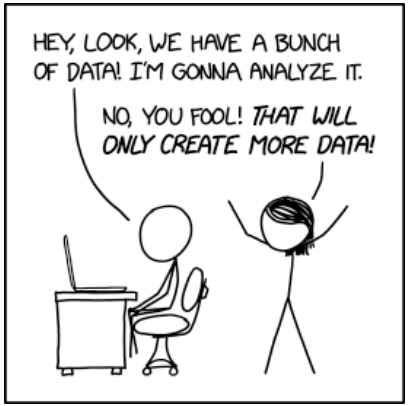
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection
- invariance (almost) for scaling
- invariance (almost) for deformation
- invariance (almost) for blurring
- invariance (almost) photon shot noise

the pathomic case



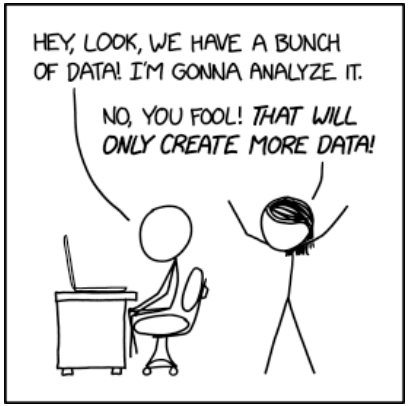
- invariance for rotation
- invariance for saturation/colorization
- invariance for luminosity
- invariance for contrast
- invariance for reflection
- invariance (almost) for scaling
- invariance (almost) for deformation
- invariance (almost) for blurring
- invariance (almost) photon shot noise
- invariance (almost) blue noise (salt and pepper)

if only it was so easy!



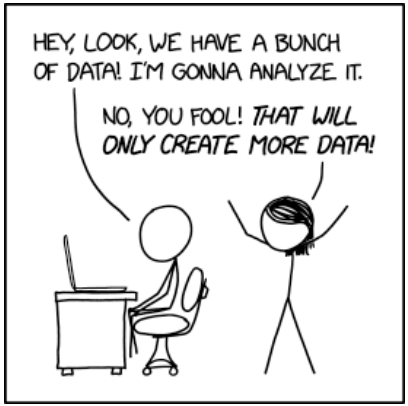
- how does one manage borders?

if only it was so easy!



- how does one manage borders?
- how many augmentation is too many?

if only it was so easy!



- how does one manage borders?
- how many augmentation is too many?
- how to describe: "basically the same but not the same"?

model augmentation

with current neural networks there are very few methods to incorporate this knowledge **directly**

model augmentation

with current neural networks there are very few methods to incorporate this knowledge **directly**

we have to rely on data augmentation, *i.e.* repeating data with variations... the intention is good!

model augmentation

with current neural networks there are very few methods to incorporate this knowledge **directly**

we have to rely on data augmentation, *i.e.* repeating data with variations... the intention is good!

but the road to hell is paved with good intentions!

model augmentation

with current neural networks there are very few methods to incorporate this knowledge **directly**

we have to rely on data augmentation, *i.e.* repeating data with variations... the intention is good!

but the road to hell is paved with good intentions!

we need better methods to incorporate invariants in our models from the ground up!

data syntesis

data syntesis

data syntesis

if we have few controls, but we know how they should look like...

data syntesis

if we have few controls, but we know how they should look like...

could we frankestein them?

a problematic organ

pancreas is a hard organ to work with

a problematic organ

pancreas is a hard organ to work with

it is:

- autolytic
- small
- uncomfortable to reach

a problematic organ

pancreas is a hard organ to work with

it is:

- autolytic
- small
- uncomfortable to reach

can we create our own?

This person does not exists

- GANs can generate interesting samples

This person does not exists

- GANs can generate interesting samples
- we fall back to the problems of model properties

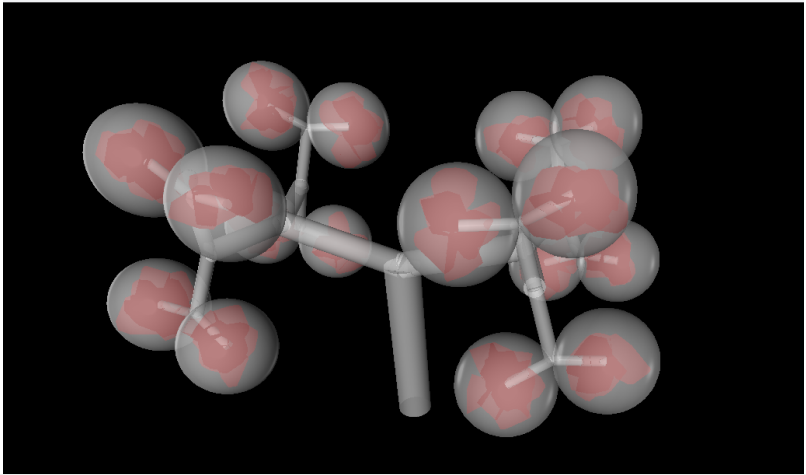
This person does not exists

- GANs can generate interesting samples
- we fall back to the problems of model properties
- still we don't have explicit knowledge of the structure

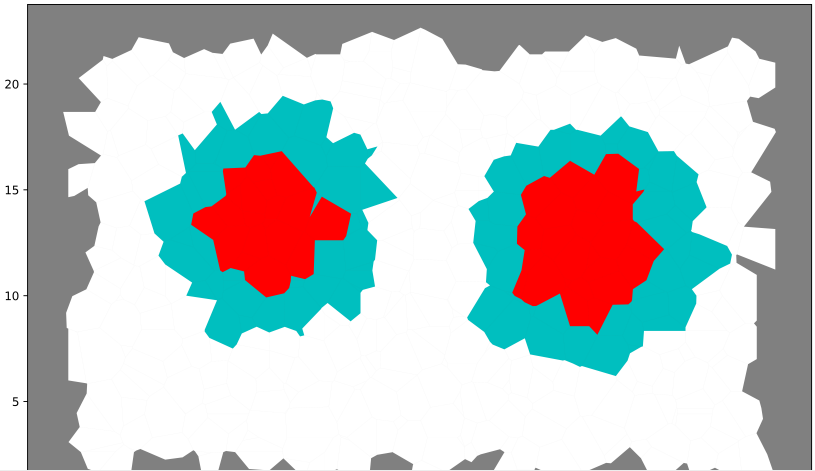
This person does not exists

- GANs can generate interesting samples
- we fall back to the problems of model properties
- still we don't have explicit knowledge of the structure
- and can still create monsters

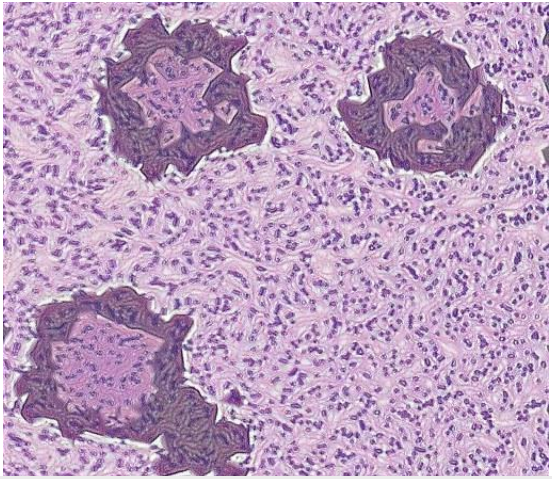
duct network with L-systems



virtual tomography



style transfer



In conclusion

In conclusion

- let's celebrate "not so big data"
  - necessity is the mother of invention
- let's create models that better encode:
  - our assumptions
  - our knowledge
  - the system's invariants
- create the data you want but don't have

Thank you for you attention