Navigating Privacy and Utility with Multiple Imputation, Satellite Imaging and Deep Learning

Mohammad Kakooei[1], James Bailie[2], Xiao-Li Meng[2], Adel Daoud[1,3]

[1] Chalmers University of Technology, Sweden.
[2] Harvard University, USA.
[3] Linköping University, Sweden

JSM 2024 — Portland, Oregon

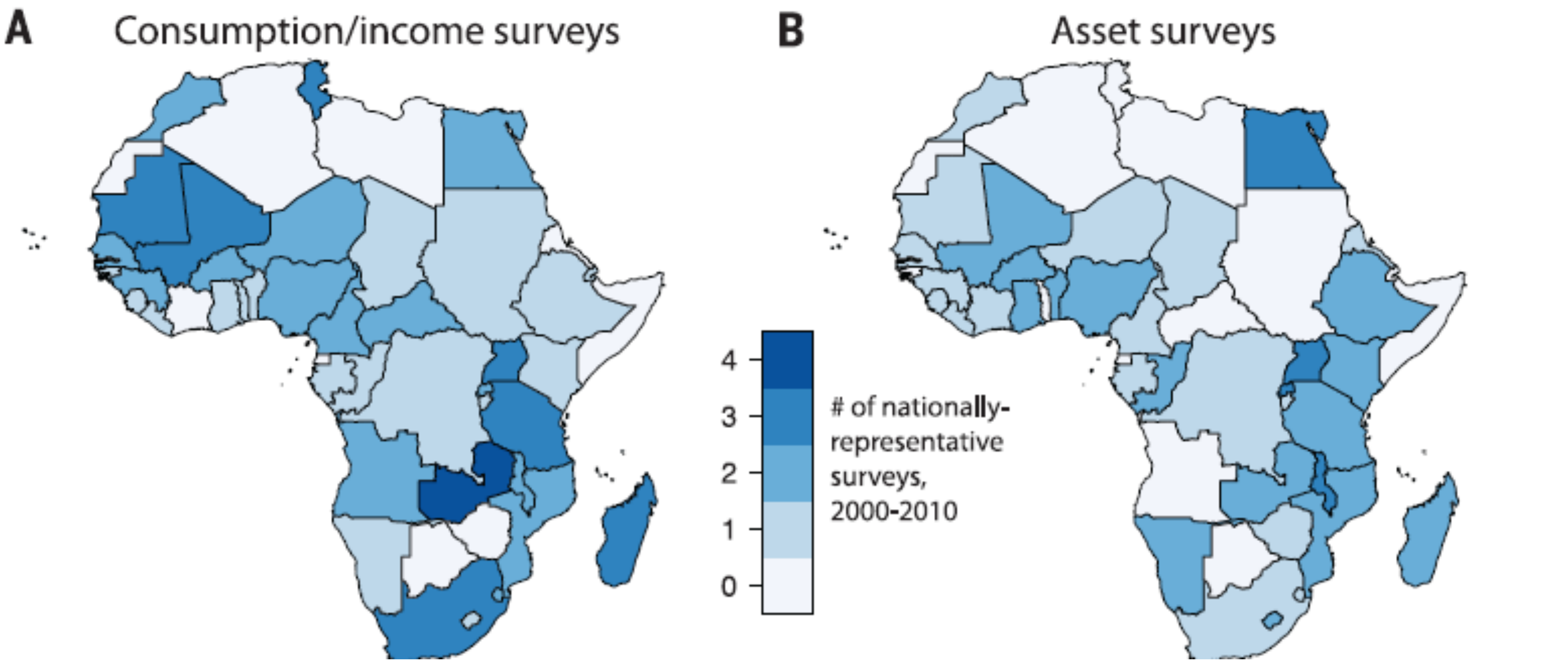
LINKÖPING UNIVERSITY

HARVARD UNIVERSITY

The AI and Global Development Lab

Funded mainly by the Swedish Research Council (SRC)

Because of a lack of high-frequency human-development data across time and space, scholarship on poverty is limited.



Source: Jean et al 2016

Source: Google Earth Timelapse (Google, Landsat, Copernicus)

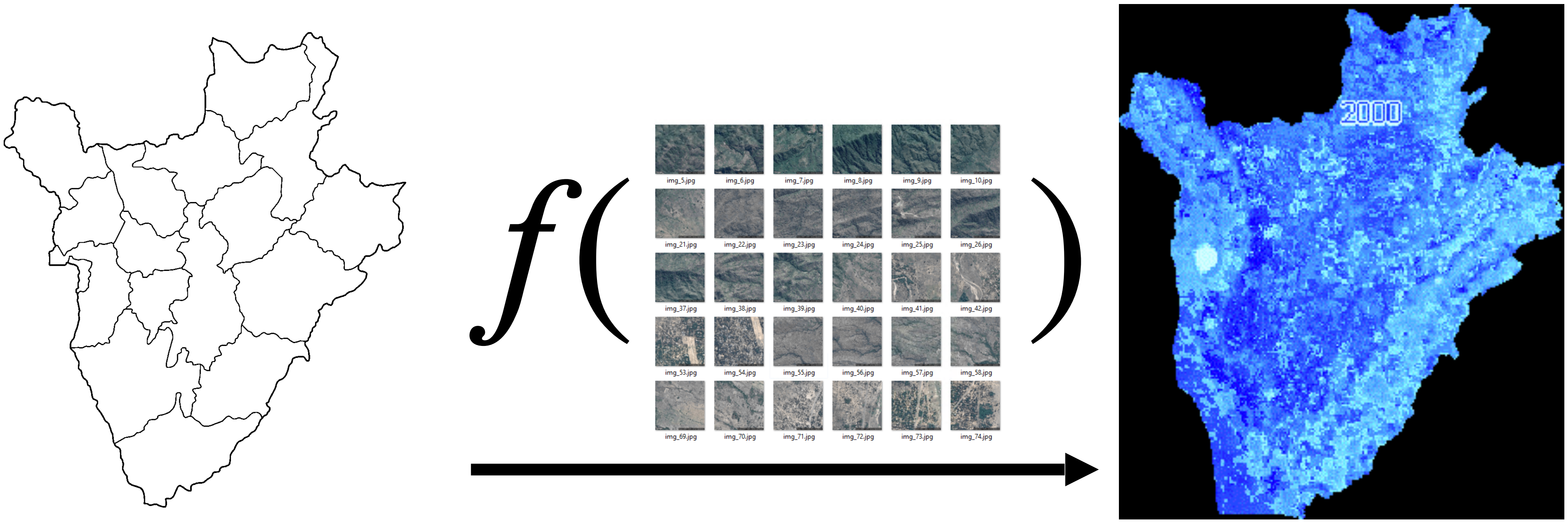
Source: Google Earth Timelapse (Google, Landsat, Copernicus)

Source: Google Earth Timelapse (Google, Landsat, Copernicus)

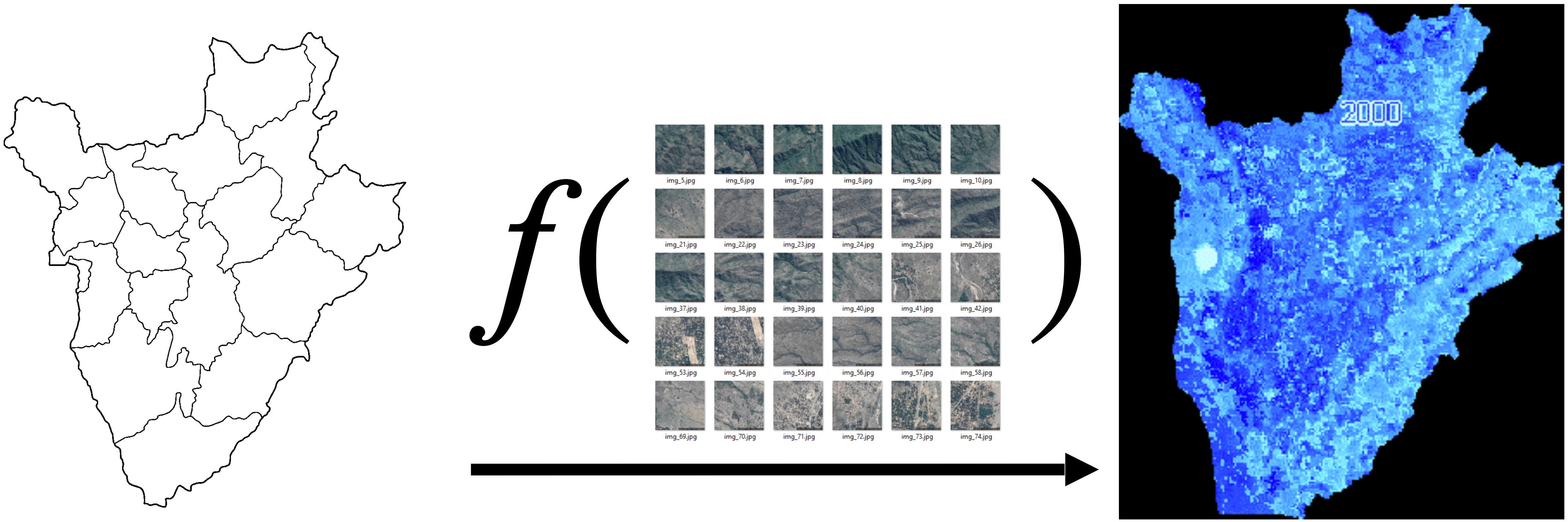
Source: Google Earth Timelapse (Google, Landsat, Copernicus)

Source: Google Earth Timelapse (Google, Landsat, Copernicus)

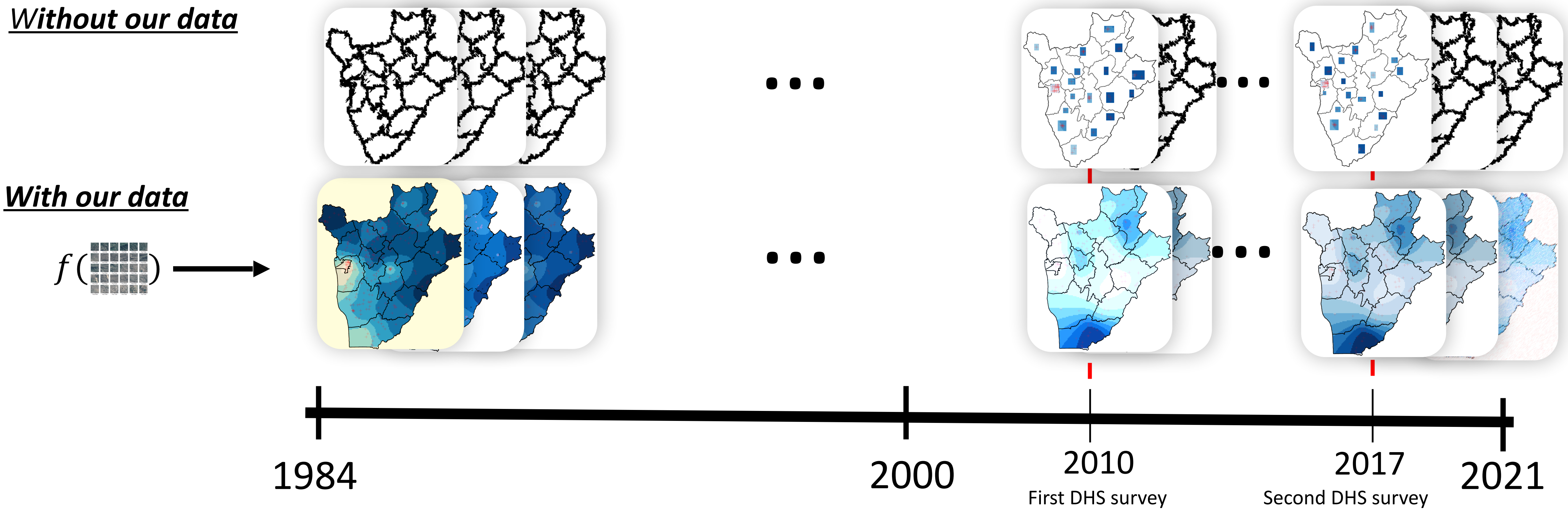
Constructing an Algorithm for Poverty Measurement



Constructing an Algorithm for Poverty Measurement

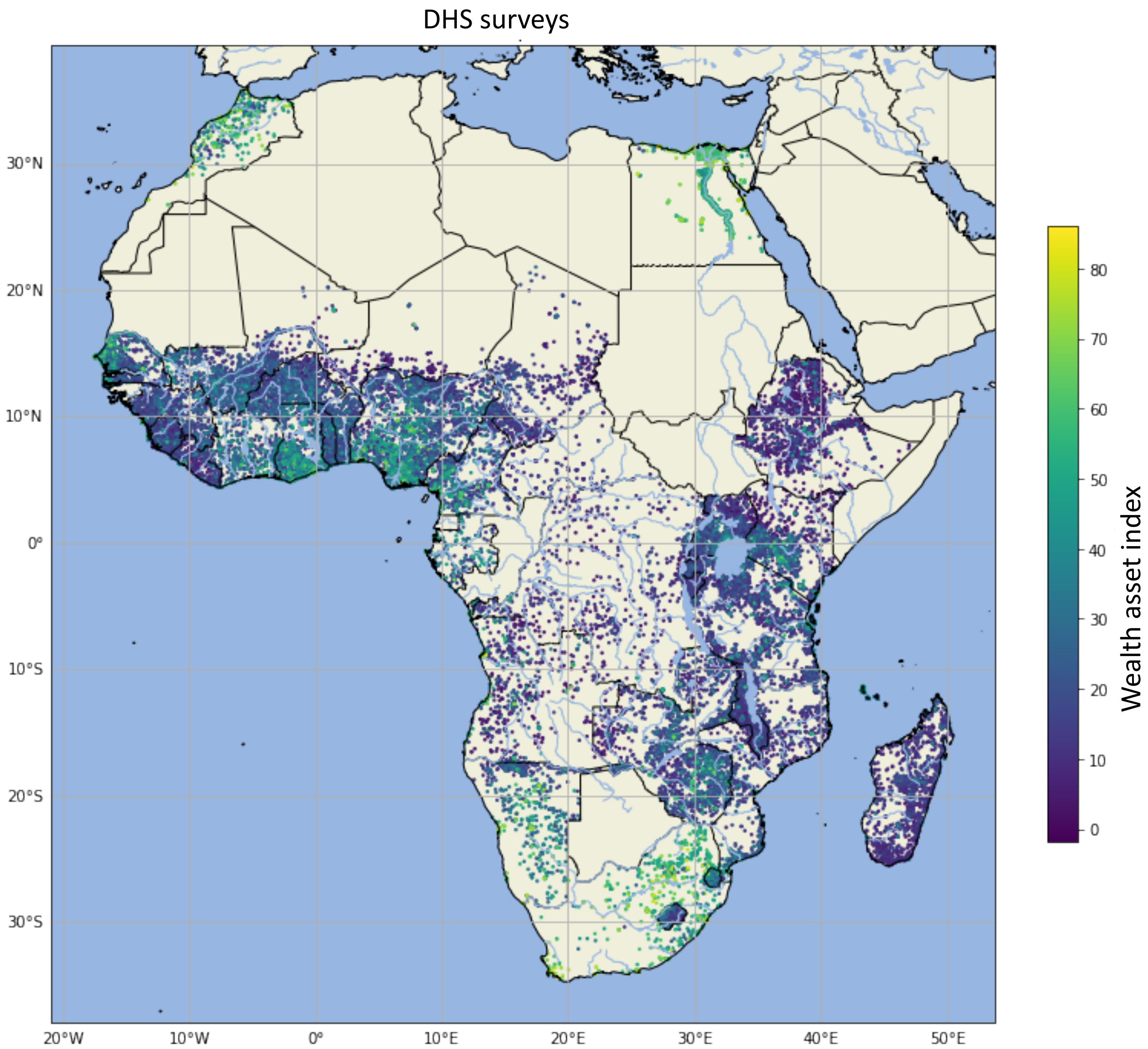


Our Data Product



Ground "truth"

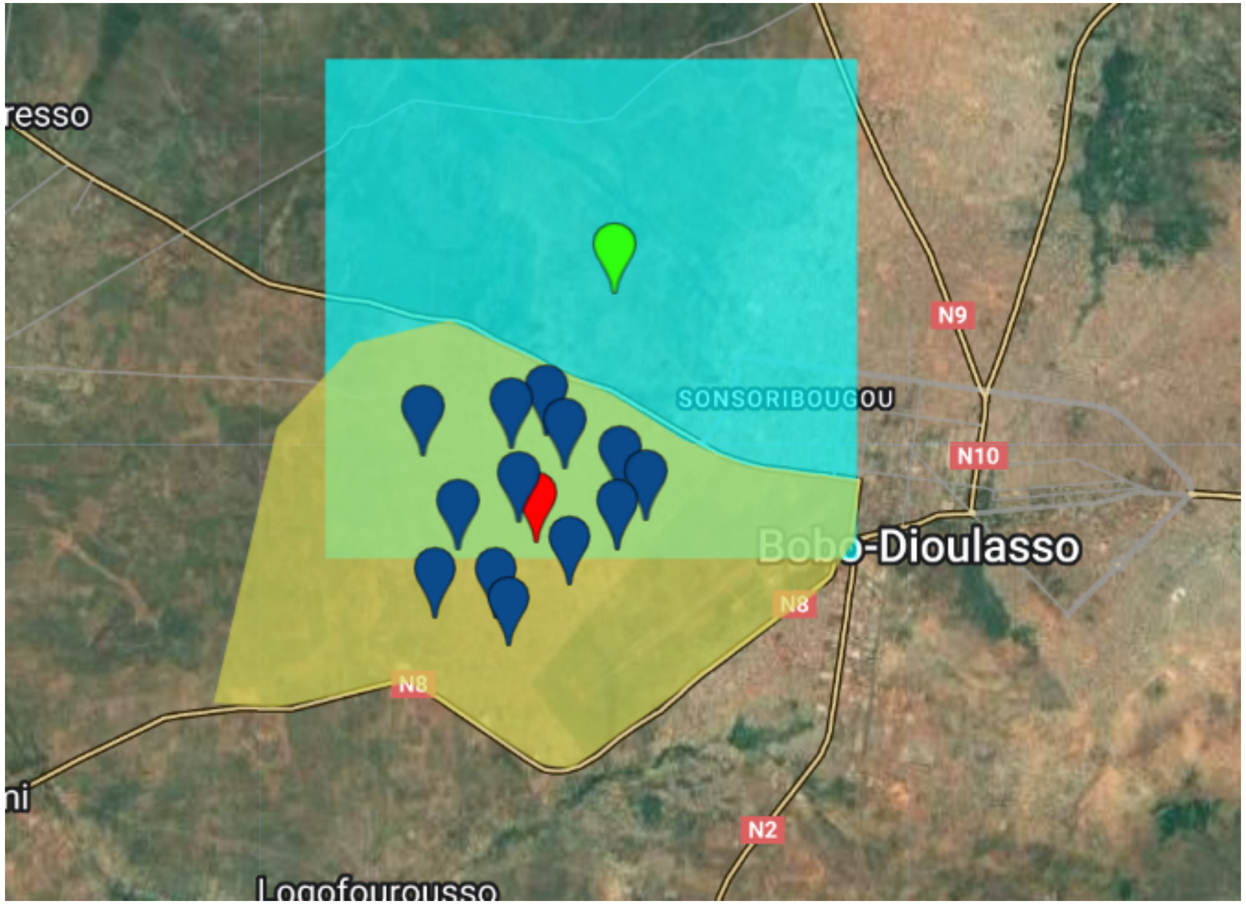
- International wealth index (material assets)
- ≈57 000 DHS survey units ("clusters")
- From 36 countries
- 1984 – 2019
- Unit of analysis: clusters consisting of about 20-30 households



But…

But… Noise Is Added For Privacy

But… Noise Is Added For Privacy



Households

Cluster center

Displaced location (released coordinates)

Correcting For Privacy Using Multiple Imputation?

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
  - True location L of each cluster i

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
    - True location L of each cluster i

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
  - True location L of each cluster i

- Known: Perturbed location D_i and perturbation distribution $\Pr(D_i \mid L_i)$

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
  - True location L of each cluster i

- Known: Perturbed location D_i and perturbation distribution $\Pr(D_i \mid L_i)$

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
  - True location L of each cluster i

- Known: Perturbed location D_i and perturbation distribution $\Pr(D_i \mid L_i)$

- Imputation: Given a prior $\pi(L_i)$, sample from posterior $\pi(L_i \mid D_i) \propto \pi(D_i)\Pr(D_i \mid L_i)$

Correcting For Privacy Using Multiple Imputation?

- What is being imputed?
    - True location L of each cluster i

- Known: Perturbed location D_i and perturbation distribution $\Pr(D_i \,|\, L_i)$

- Imputation: Given a prior $\pi(L_i)$, sample from posterior $\pi(L_i \,|\, D_i) \propto \pi(D_i) \Pr(D_i \,|\, L_i)$

Correcting For Privacy Using Multiple Imputation?

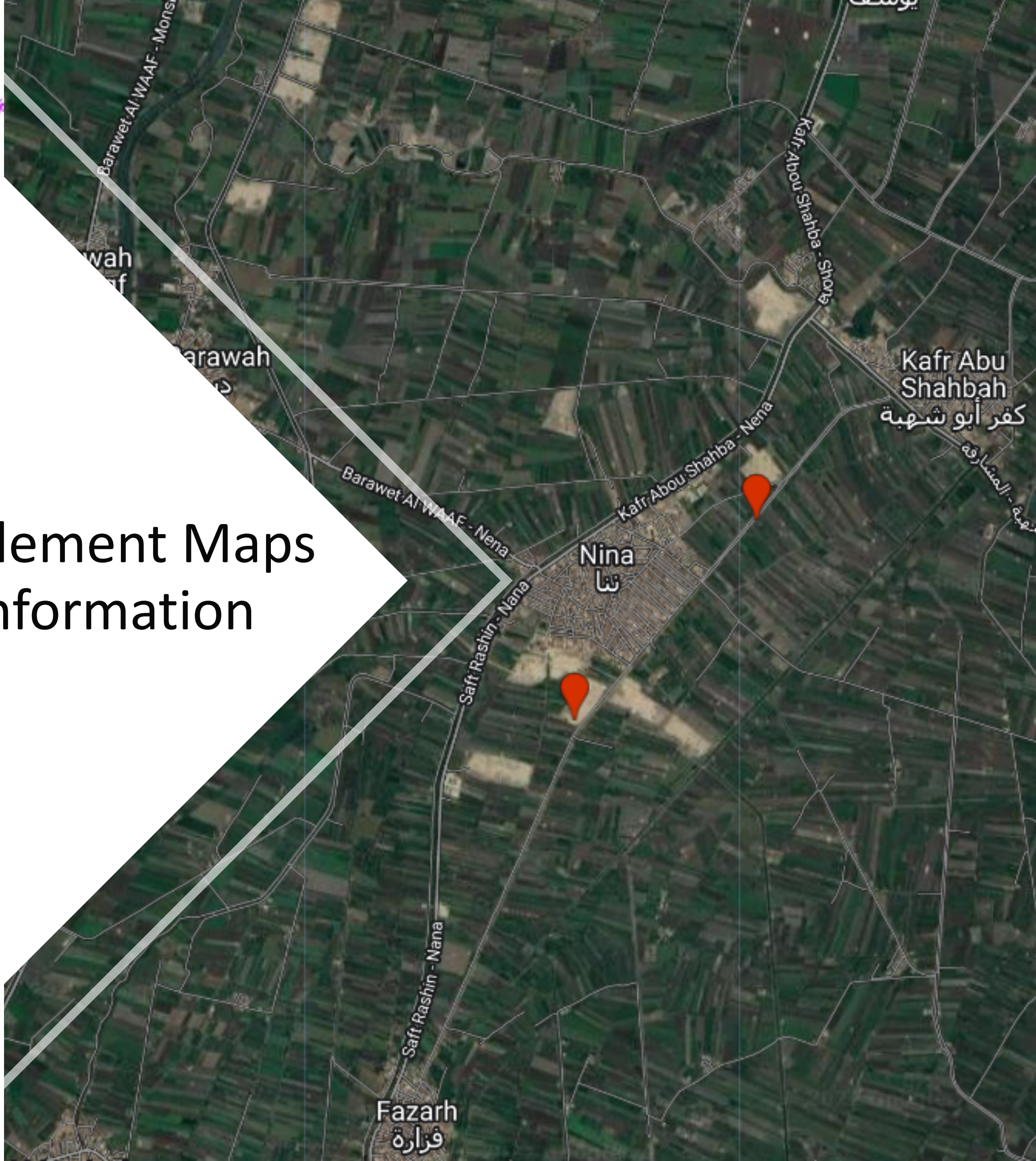
- What is being imputed?
  - True location L of each cluster i

- Known: Perturbed location D_i and perturbation distribution $\Pr(D_i \mid L_i)$

- Imputation: Given a prior $\pi(L_i)$, sample from posterior $\pi(L_i \mid D_i) \propto \pi(D_i)\Pr(D_i \mid L_i)$

- Train and test model using the satellite images at the imputed locations $\hat{L}_i$.

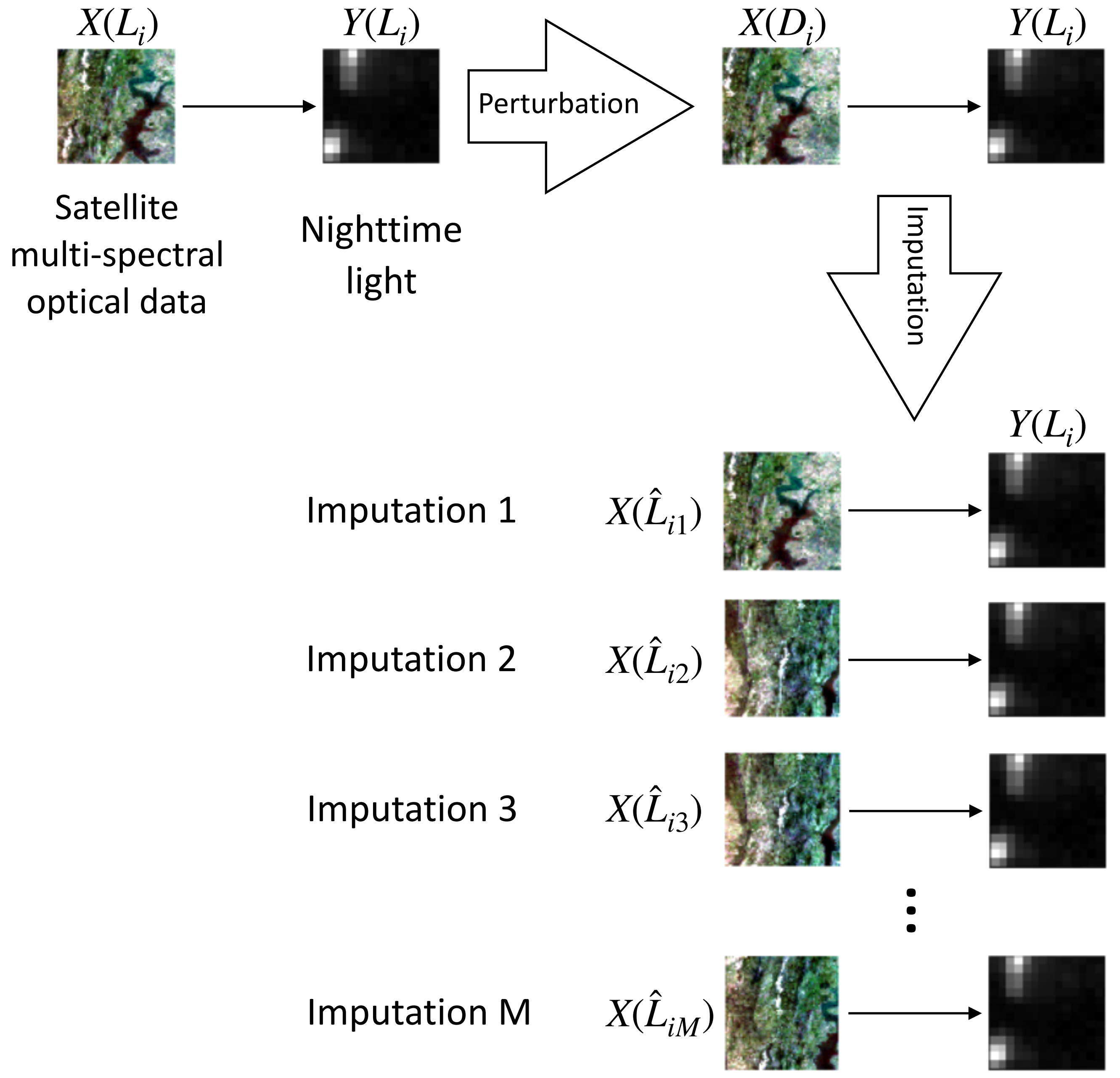
Human Settlement Maps as Prior Information



A simulation study predicting nighttime light intensity

A simulation study predicting nighttime light intensity

A simulation study predicting nighttime light intensity



Can We Trust the Imputed Data?

Can We Trust the Imputed Data?

- Ideal (A): Evaluate a fitted model $\mathscr{A}$ on the confidential dataset $\mathscr{D}$.

Can We Trust the Imputed Data?

- Ideal (A): Evaluate a fitted model $\mathscr{A}$ on the confidential dataset $\mathscr{D}$.
- Pragmatic (B): Evaluate $\mathscr{A}$ on a 'synthetic' dataset $\mathscr{D}_{\mathrm{Syn}}$.

Can We Trust the Imputed Data?

- Ideal (A): Evaluate a fitted model $\mathscr{A}$ on the confidential dataset $\mathscr{D}$.

- Pragmatic (B): Evaluate $\mathscr{A}$ on a 'synthetic' dataset $\mathscr{D}_{\text{Syn}}$.

- What can (B) tell us about (A), specifically with respect to R-squared: $R^2 = 1 - RSS/TSS$?

Can We Trust the Imputed Data?

- Ideal (A): Evaluate a fitted model $\mathscr{A}$ on the confidential dataset $\mathscr{D}$.

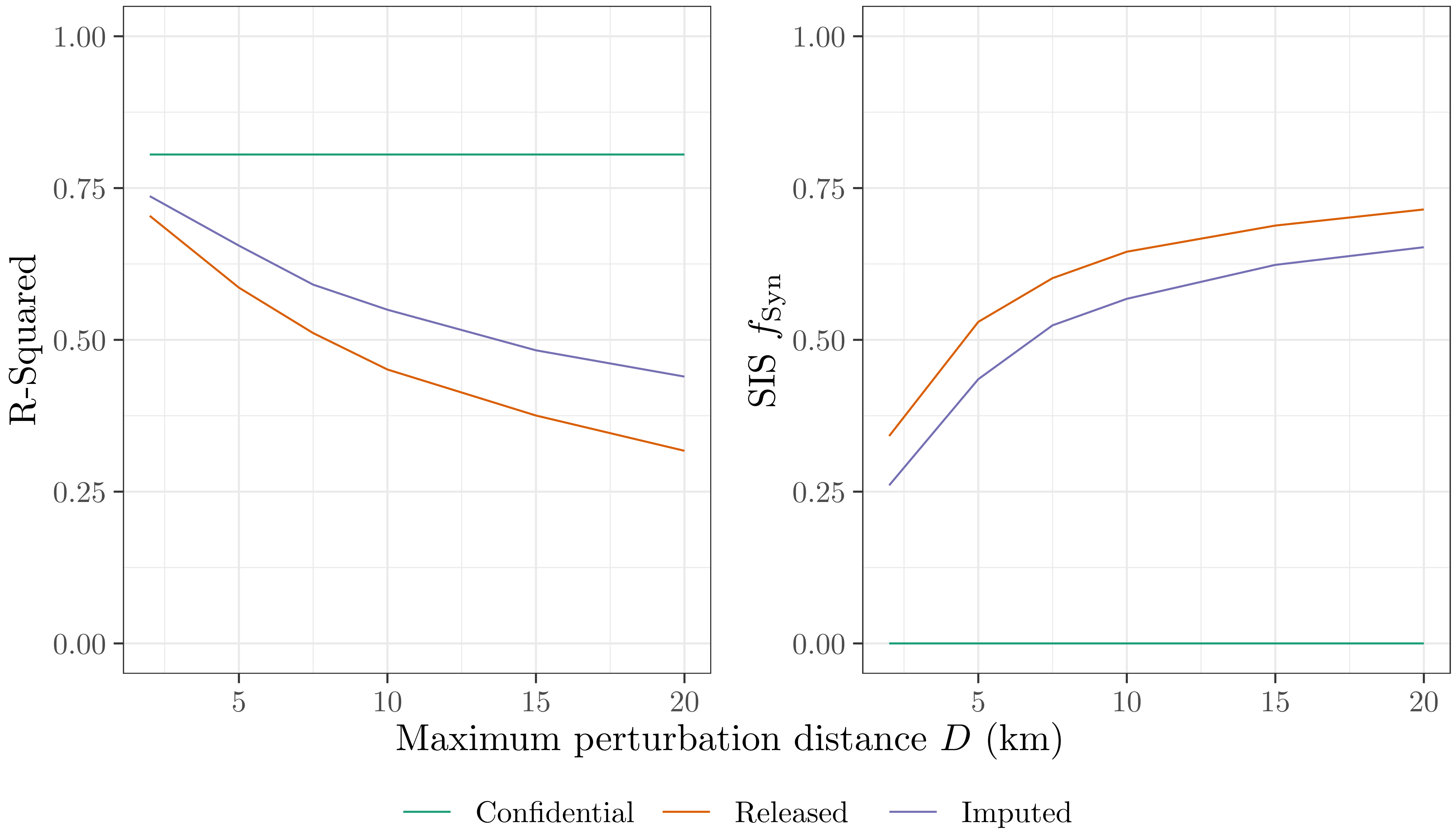
- Pragmatic (B): Evaluate $\mathscr{A}$ on a 'synthetic' dataset $\mathscr{D}_{\mathrm{Syn}}$.

- What can (B) tell us about (A), specifically with respect to R-squared: $R^2 = 1 - RSS/TSS$?

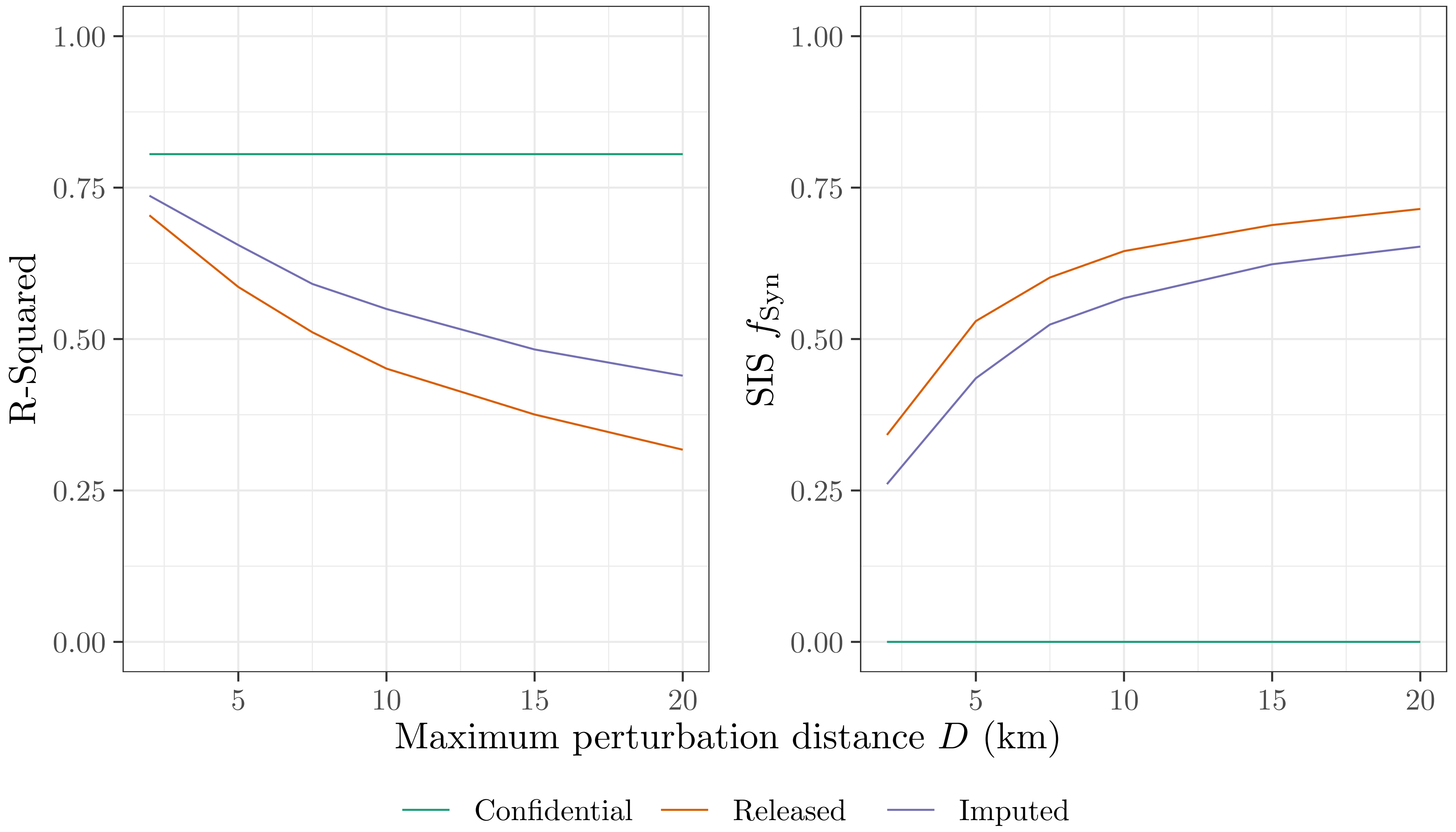
- With some simple algebra, $R^2 = R^2_{\mathrm{Syn}} + (1 - R^2_{\mathrm{Syn}})f_{\mathrm{Syn}}$, where

$$f_{\mathrm{Syn}} = \frac{RSS_{\mathrm{Syn}}/RSS - TSS_{\mathrm{Syn}}/TSS}{RSS_{\mathrm{Syn}}/RSS}$$

Can We Trust the Imputed Data?



Can We Trust the Imputed Data?



Can We Trust the Imputed Data?

Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance

Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance

- We have $RSS_{\text{Syn}} = RSS + [1 - 2\hat{\beta}_{r,\delta}] \sum_i \delta_i^2$

Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance

- We have $RSS_{\text{Syn}} = RSS + [1 - 2\hat{\beta}_{r,\delta}] \sum_i \delta_i^2$

where $\hat{\beta}_{r,\delta}$ is the regression coefficient when regressing the benchmark residuals r_i on the difference of residuals $\delta_i = r_i - r_i^{\text{Syn}}$.

Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance

- We have $RSS_{\mathrm{Syn}} = RSS + [1 - 2\hat{\beta}_{r,\delta}] \sum_i \delta_i^2$

where $\hat{\beta}_{r,\delta}$ is the regression coefficient when regressing the benchmark residuals r_i on the difference of residuals $\delta_i = r_i - r_i^{\mathrm{Syn}}$.

- Then $R^2 \geq R^2_{\mathrm{Syn}}$ if and only if $\hat{\beta}_{r,\delta} \leq 0.5$ (assuming $TSS = TSS_{\mathrm{Syn}}$).

Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance

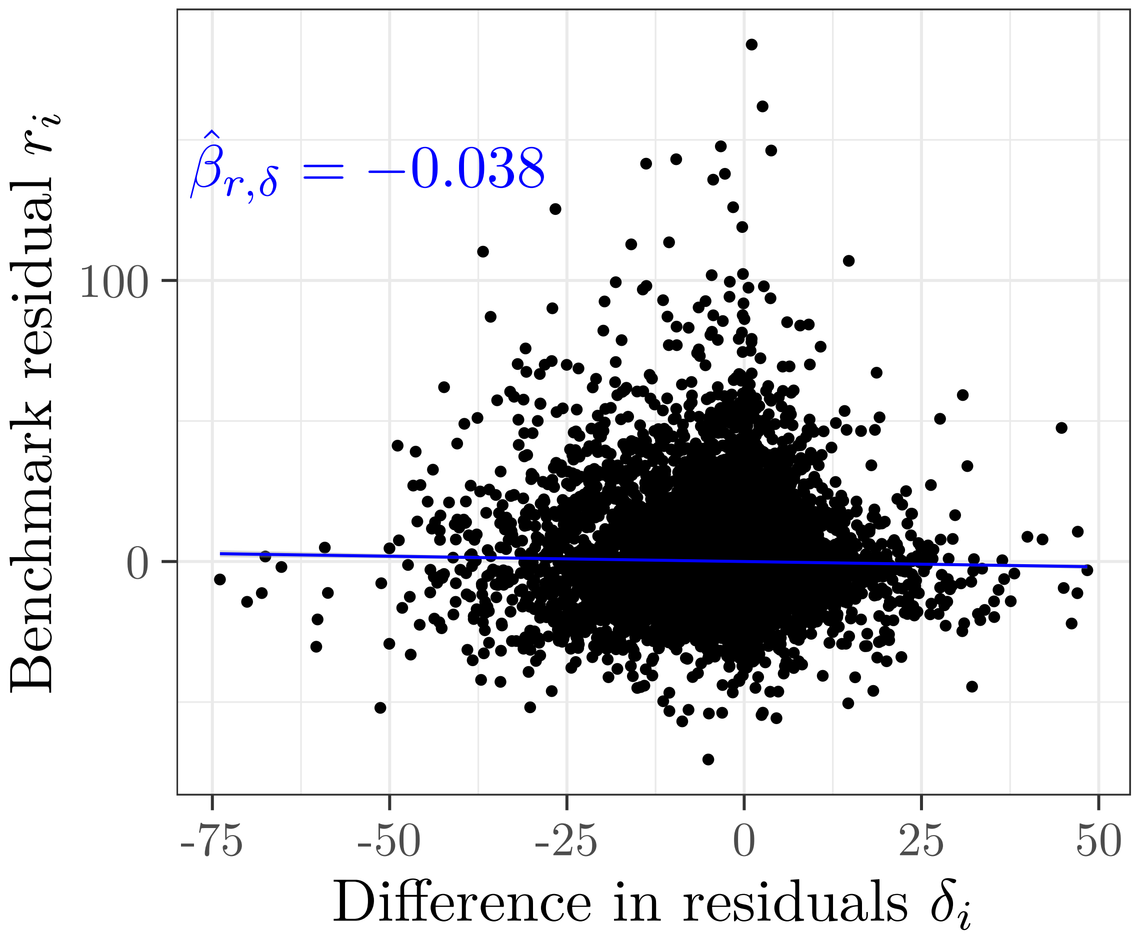
- We have $RSS_{\text{Syn}} = RSS + [1 - 2\hat{\beta}_{r,\delta}] \sum_i \delta_i^2$

where $\hat{\beta}_{r,\delta}$ is the regression coefficient when regressing the benchmark residuals r_i on the difference of residuals $\delta_i = r_i - r_i^{\text{Syn}}$.

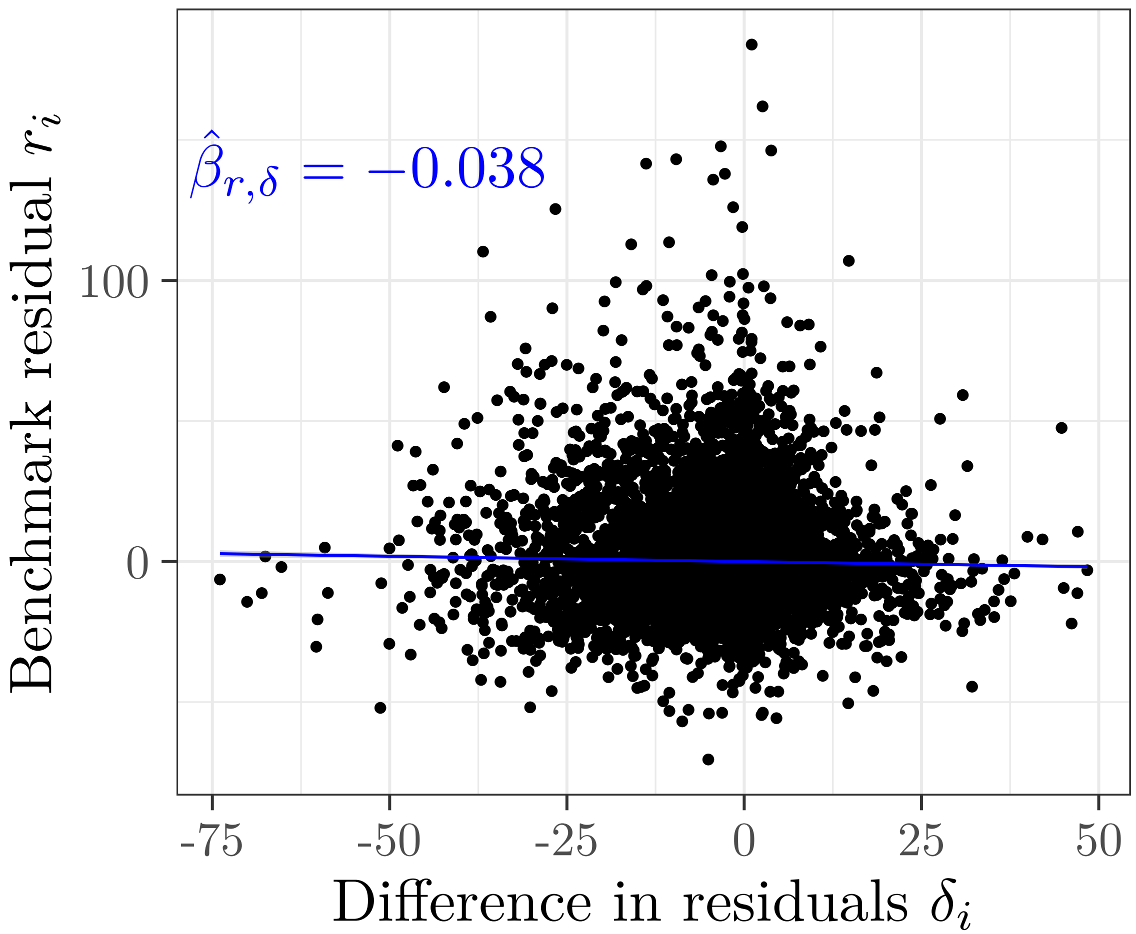
- Then $R^2 \geq R^2_{\text{Syn}}$ if and only if $\hat{\beta}_{r,\delta} \leq 0.5$ (assuming $TSS = TSS_{\text{Syn}}$).

- I.e. R^2_{Syn} is a lower bound as long as δ_i is not informative of r_i.

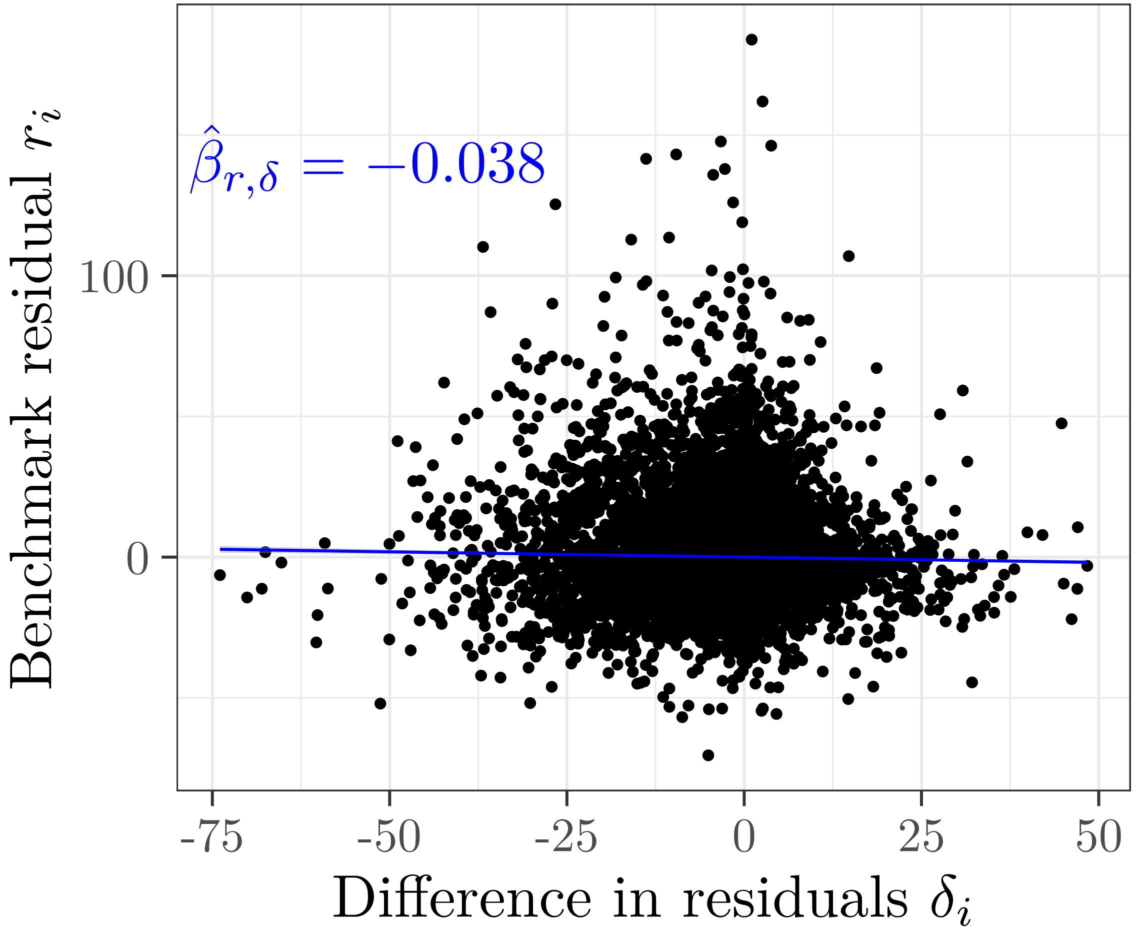
Can We Trust the Imputed Data?

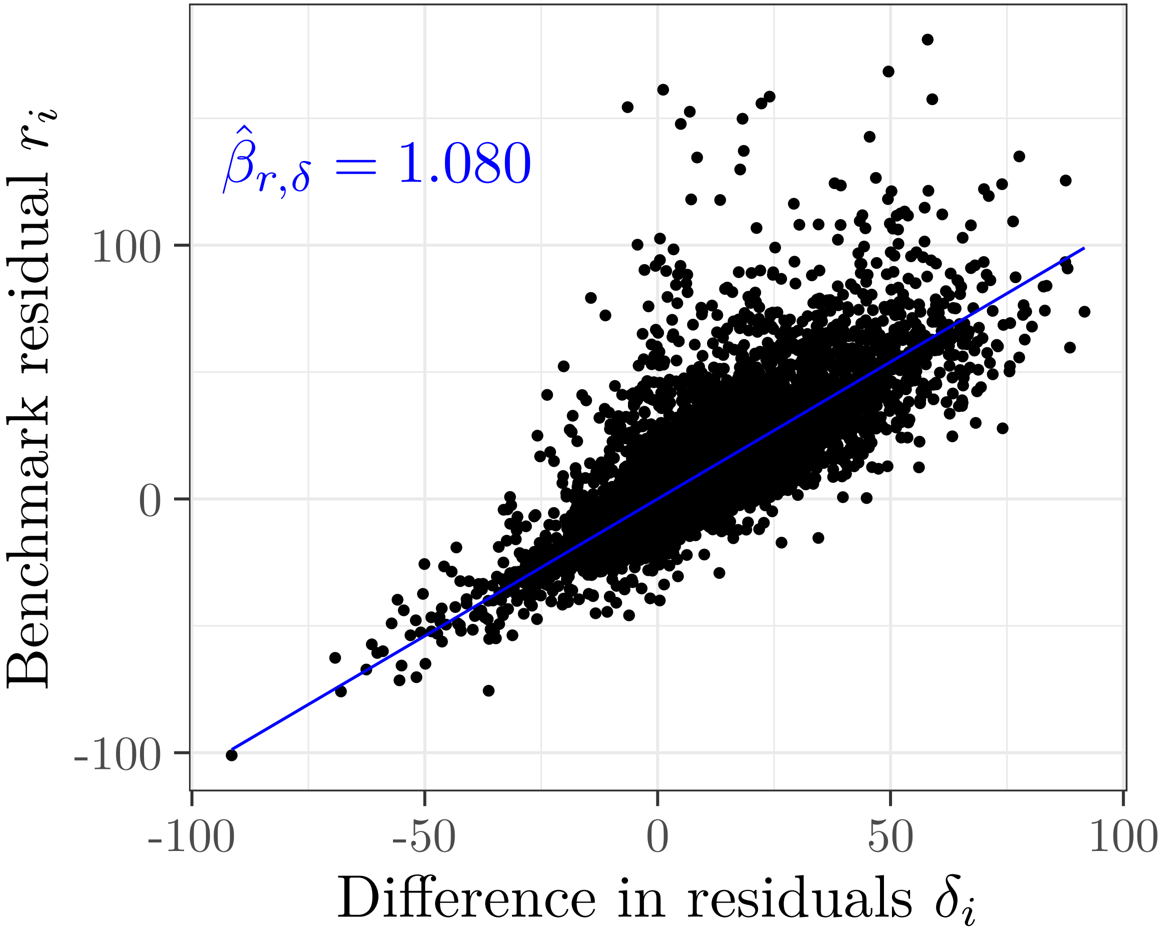


Can We Trust the Imputed Data?

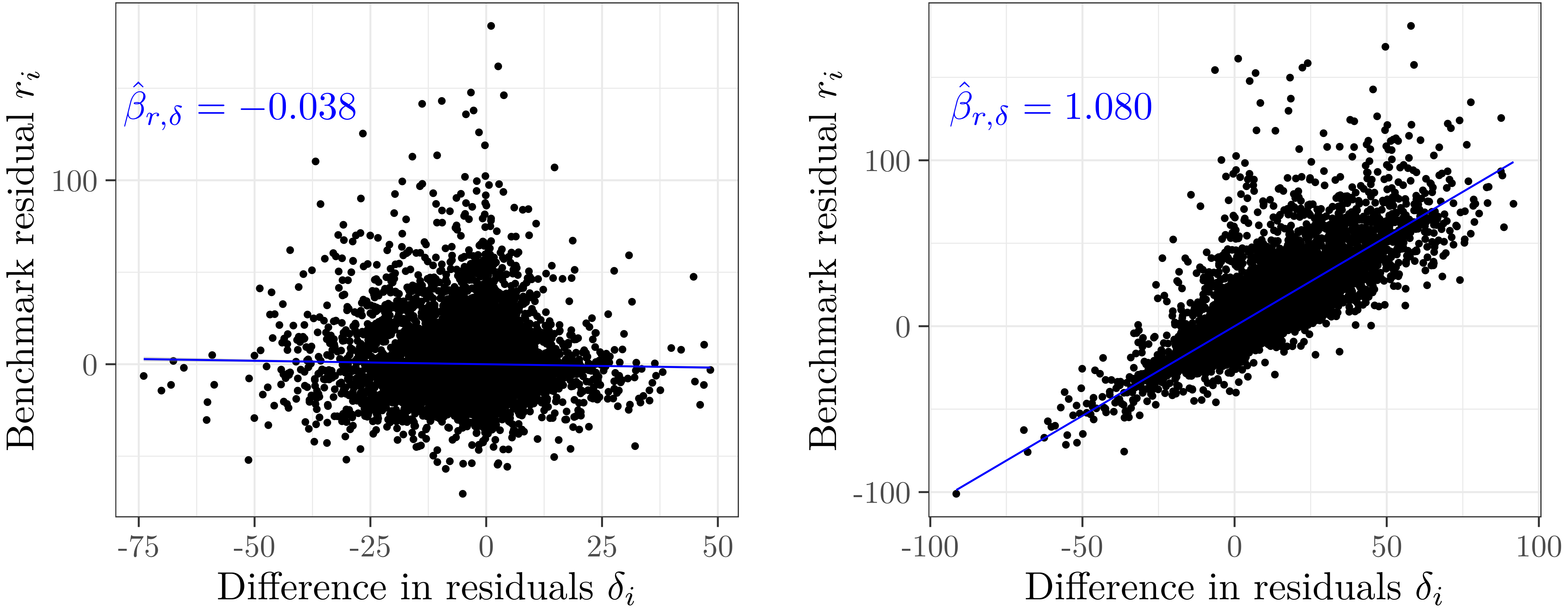


Can We Trust the Imputed Data?



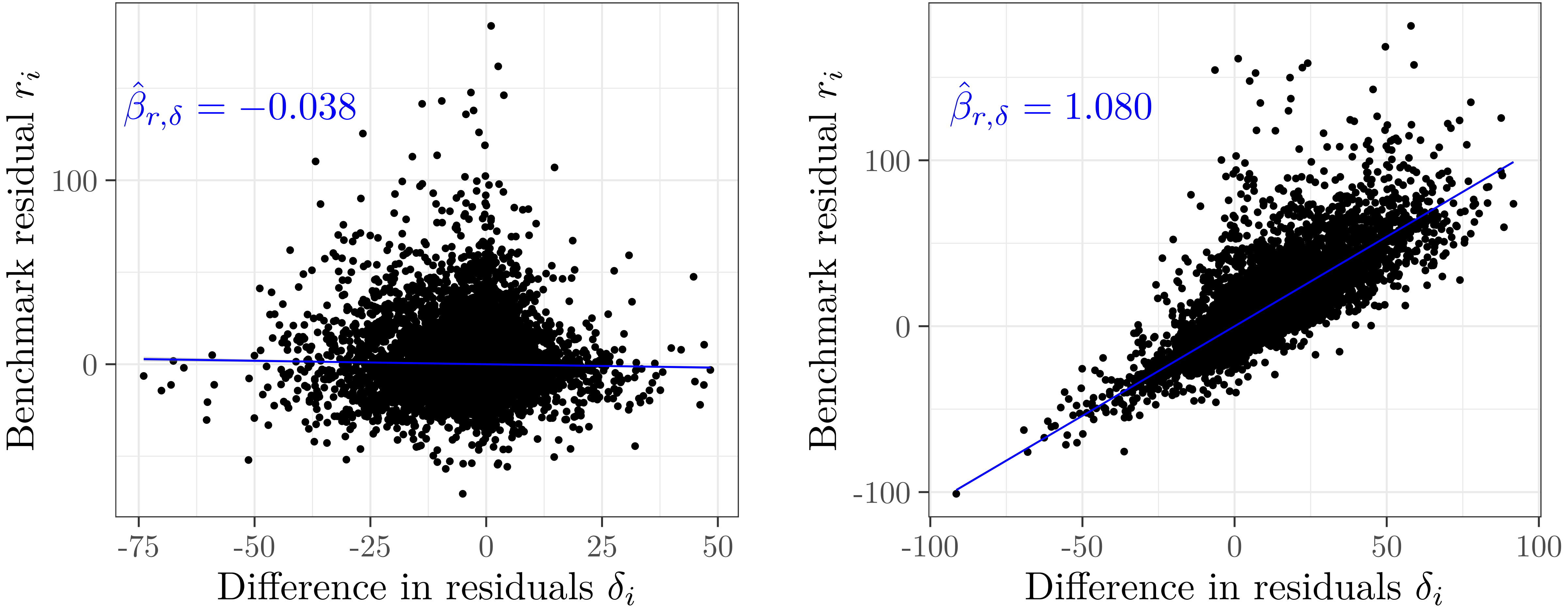


Can We Trust the Imputed Data?

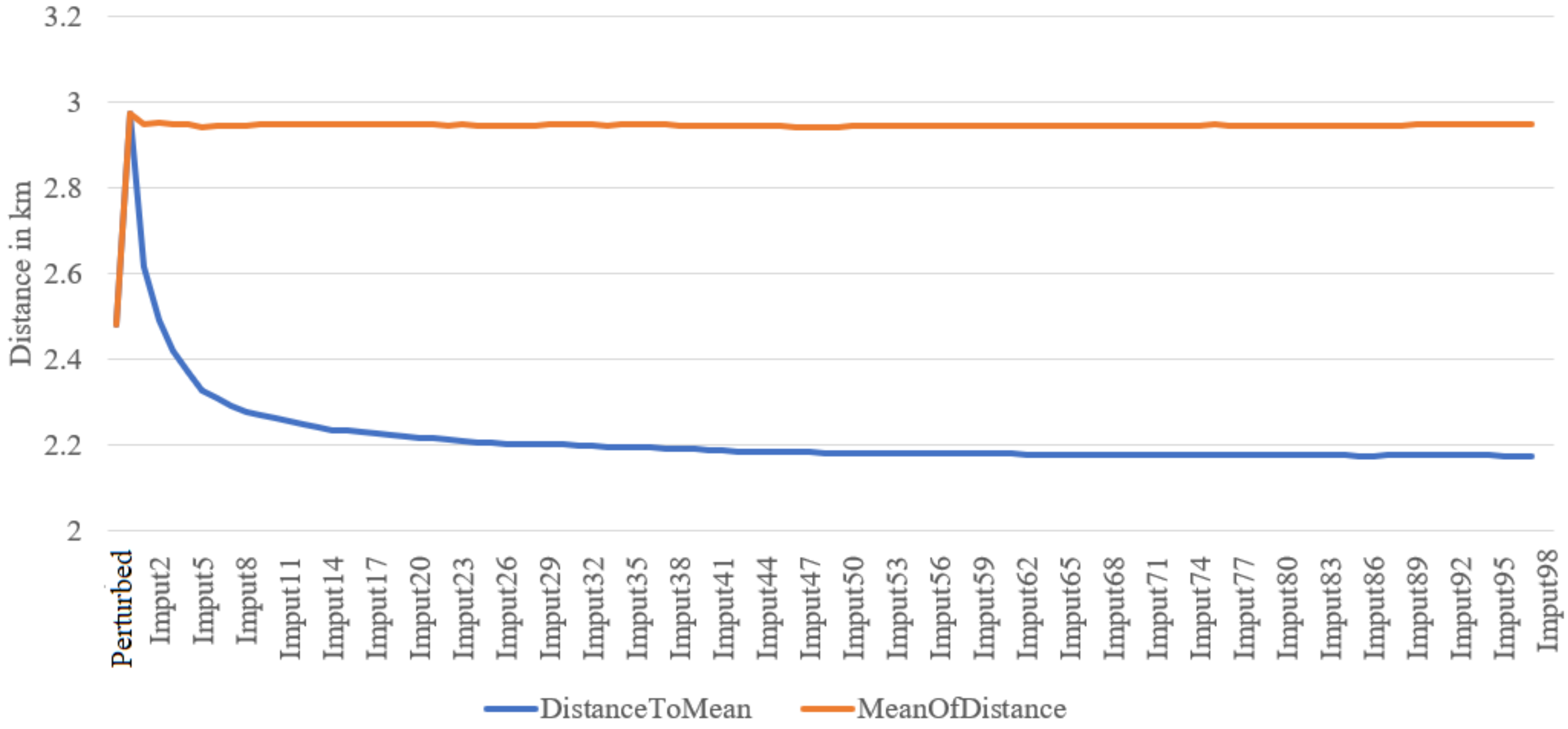


Can We Trust the Imputed Data?

Yes, at least for a lower bound on the true performance



Compare MI average distance with distance to average of MI



Comparing 5 DL models

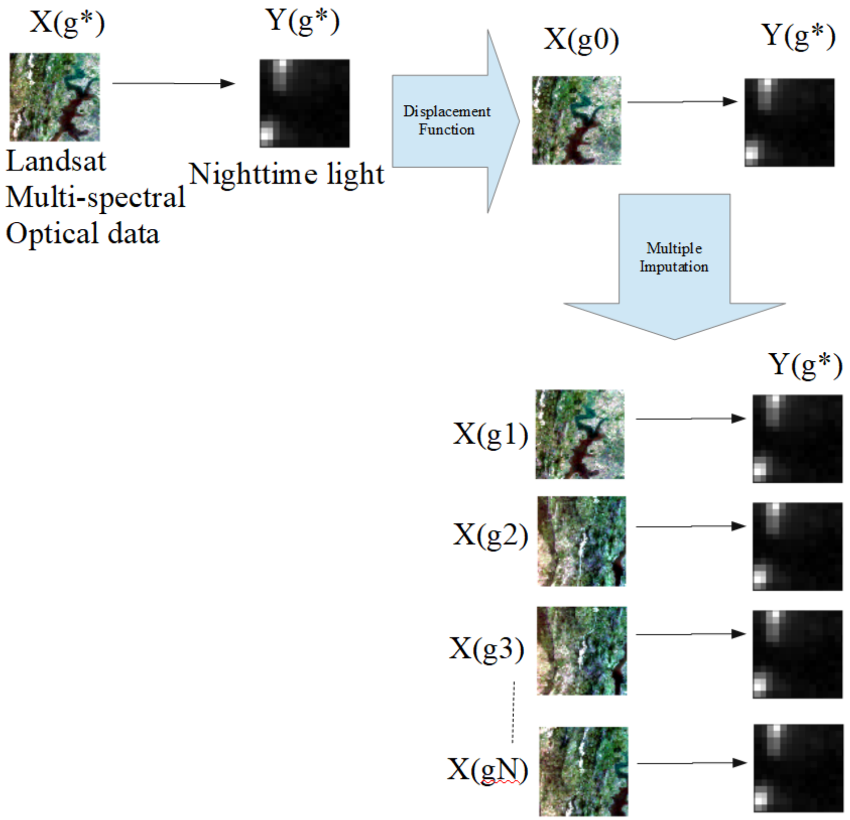
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

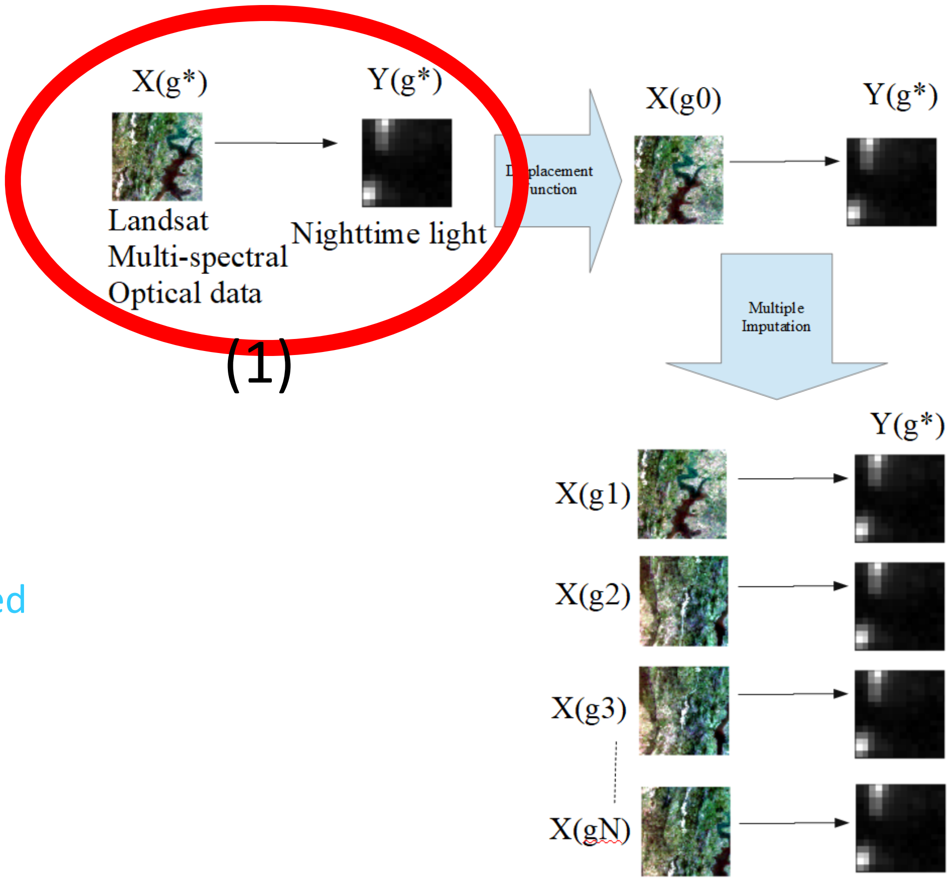
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

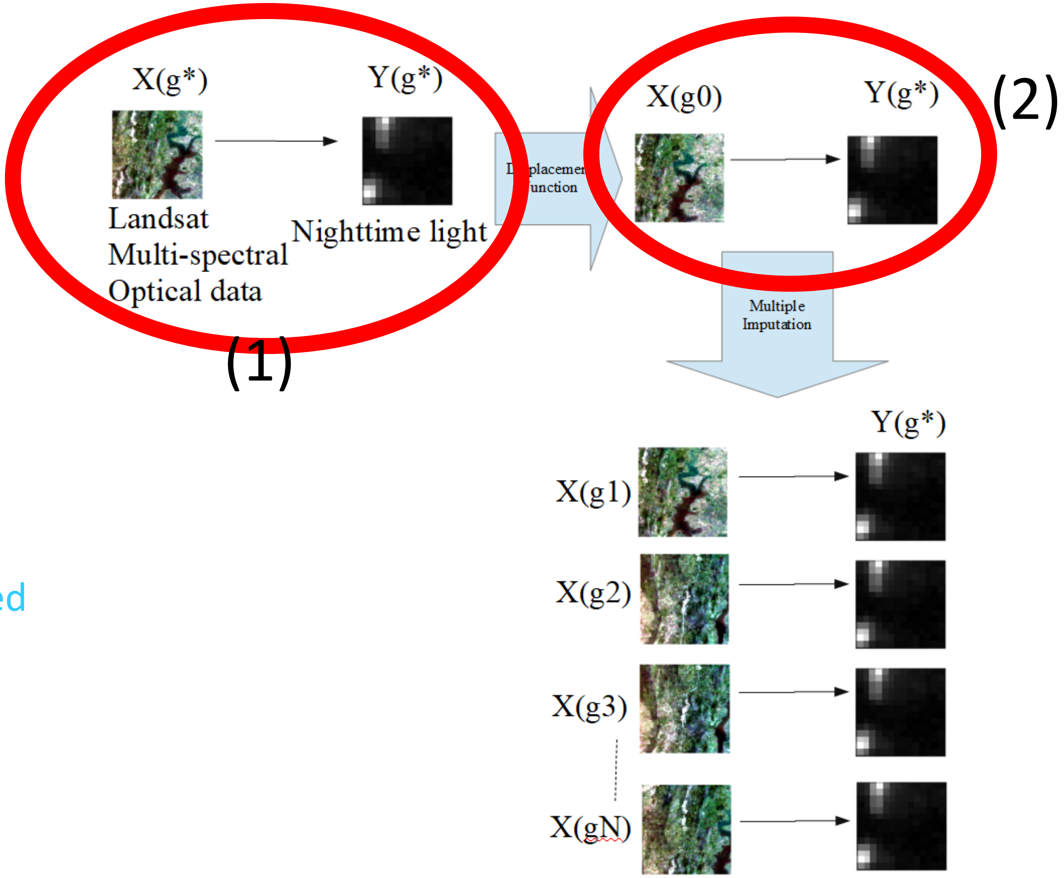
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

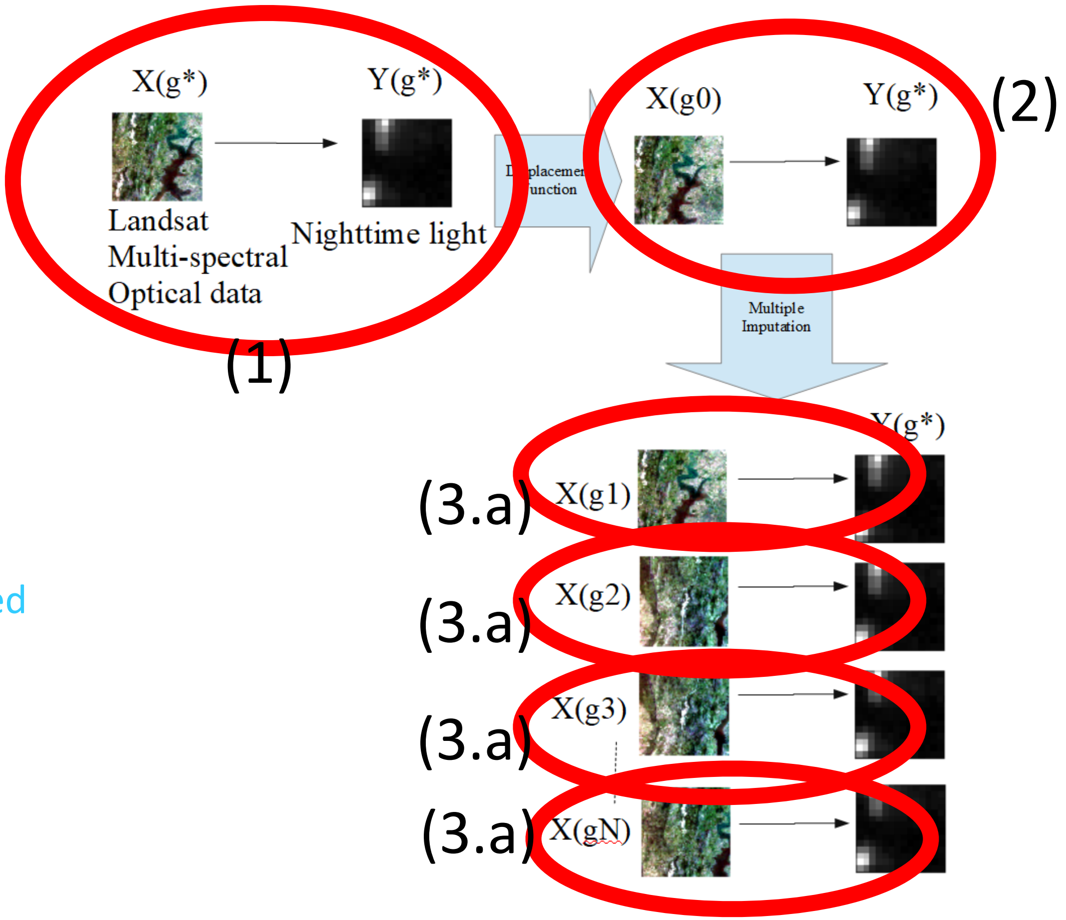
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

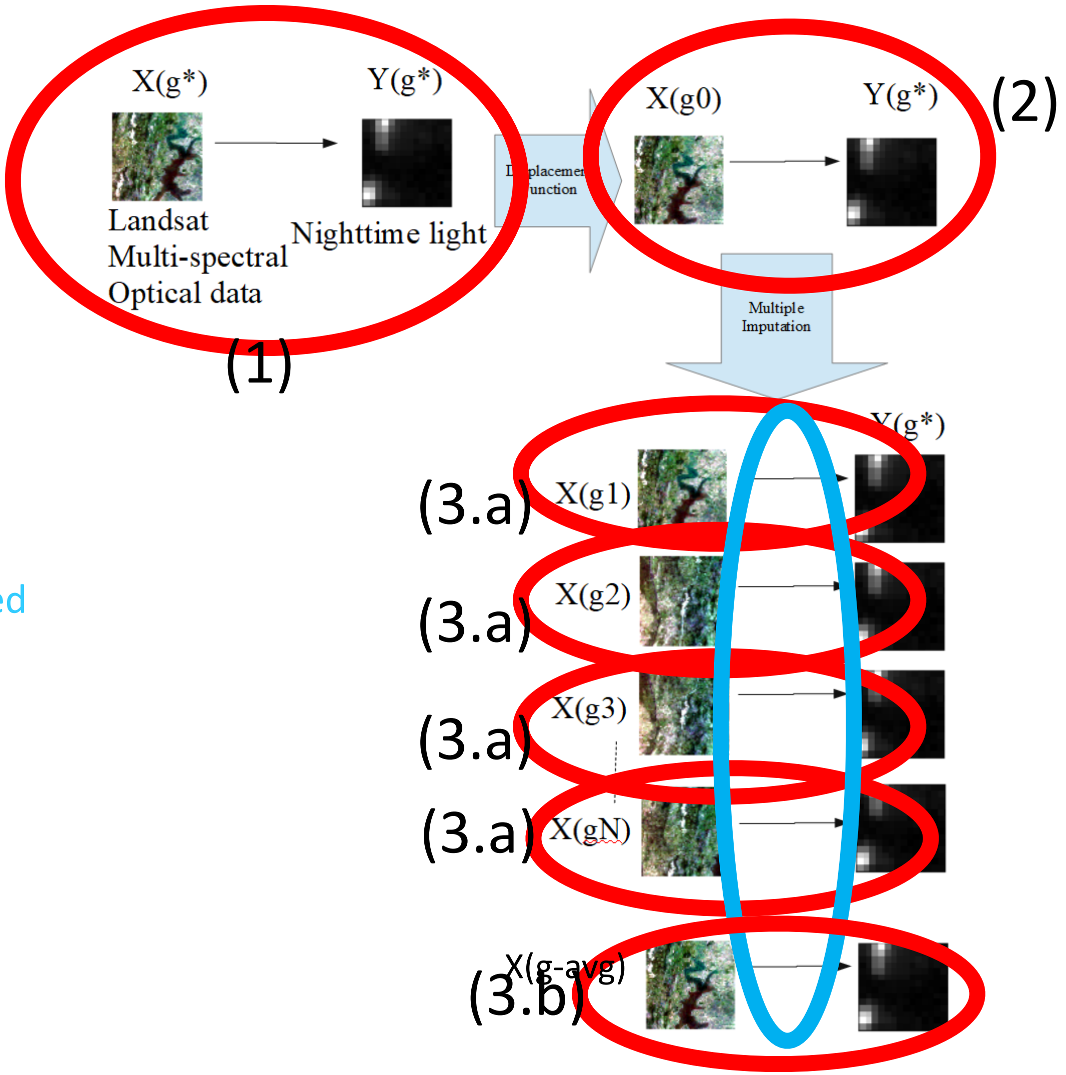
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

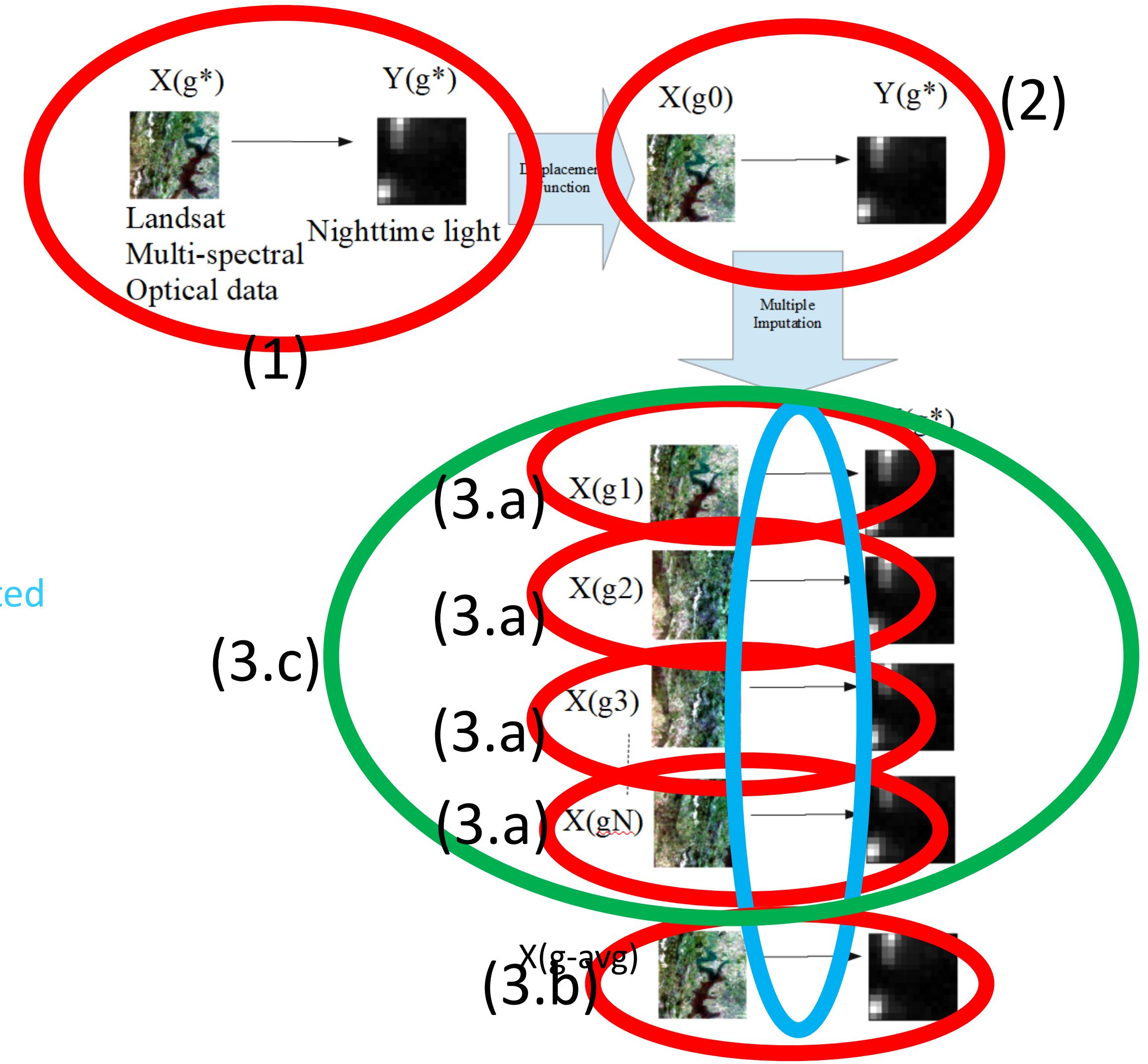
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Comparing 5 DL models

(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively

Which one predicts most accurately, and which one least?

Comparing 5 DL models

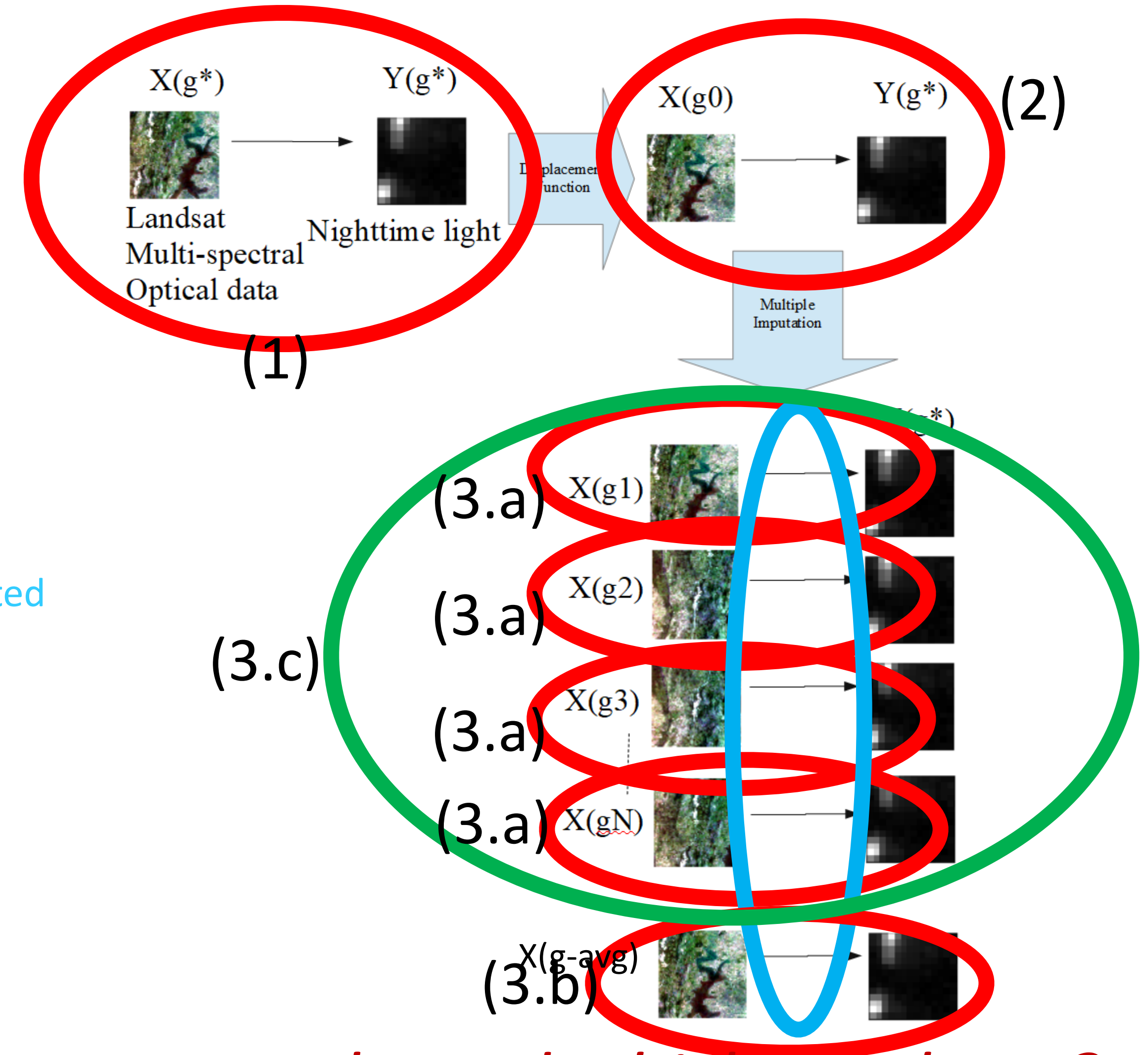
(1) DL trained on confidential data

(2) DL on released data

(3.a) DL on each imputation and than taking average

(3.b) DL on the average location of the imputed data

(3.c) DL on all imputed data collectively



Which one predicts most accurately, and which one least?
When measuring accuracy against what benchmark?

Evaluating the 5 DL models on five different test datasets

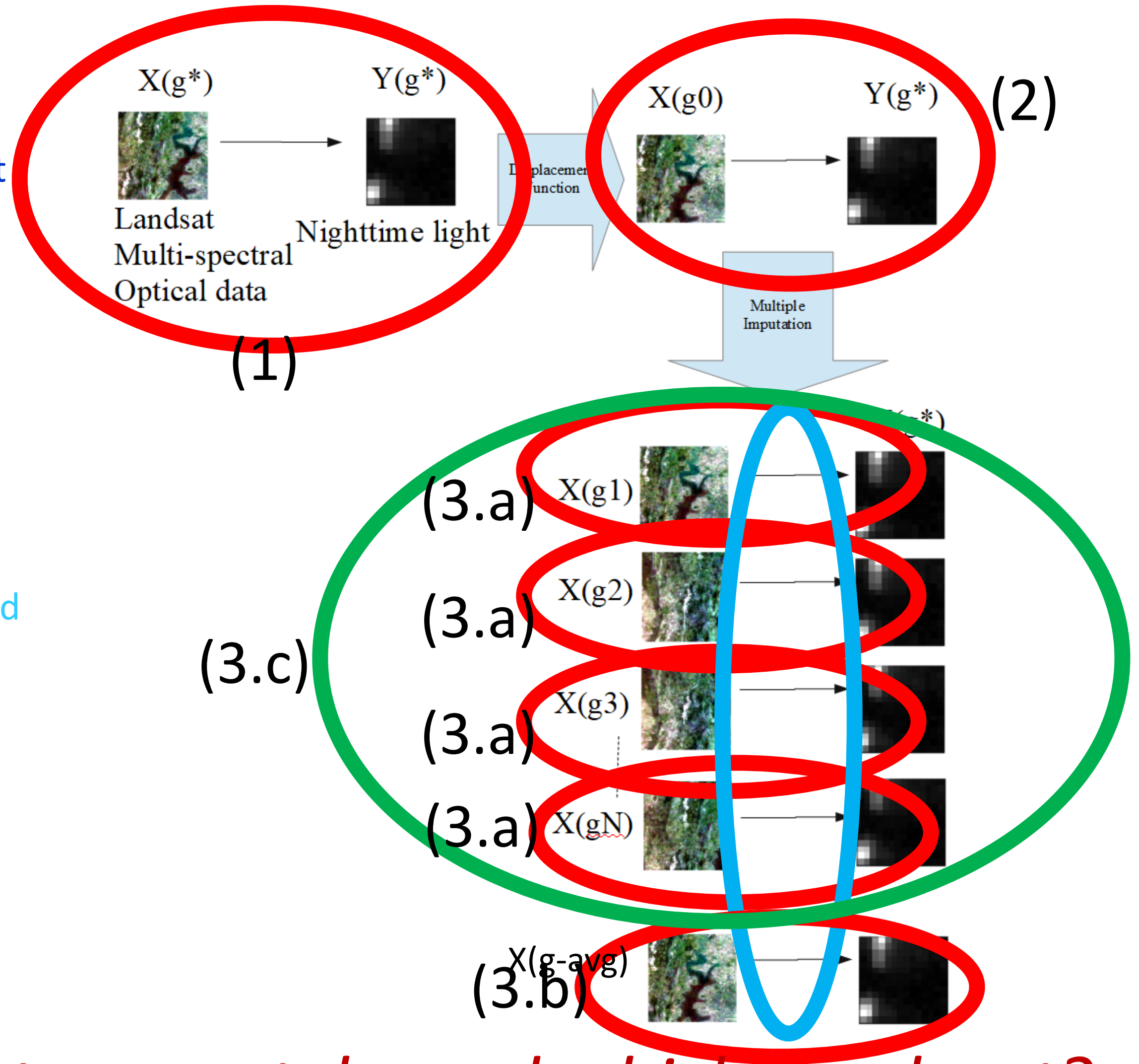
(1) Test on confidential data

(2) Test on released data

(3.a) Test on each imputation and than taking average

(3.b) Test on the average location of the imputed data

(3.c) Test on all imputed data collectively



Which one predicts most accurately, and which one least? When measuring accuracy against what benchmark?

Evaluating the 5 DL models on five different te
datasets

(1) Test on confidential data

(2) Test on released data

(3.a) Test on each imputation and than taking average

(3.b) Test on the average location of the imput data

(3.c) Test on all imputed data collectively

(4) Test on a single imputed data

Training dataset(s) $\mathscr{D}^{\mathrm{Tr}}$			Test dataset(s) $\mathscr{D}^{\mathrm{Te}}$: Single (1)	(2)	(4)	(3b)	(3c)	Multiple (3a)
Single		(1)	0.77	0.56	0.58	0.62	0.58	0.69
		(2)	0.69	0.64	0.62	0.64	0.62	0.66
		(4)	0.70	0.64	0.64	0.66	0.63	0.68
		(3b)	0.72	0.63	0.62	0.67	0.63	0.68
		(3c)						
Multiple		(3a)	0.73	0.67	0.69	0.69	0.63	0.69
	w/ diff. seeds	(1)	0.81	0.59	0.61	0.66	0.57	0.70
		(2)	0.70	0.65	0.63	0.66	0.59	0.66
		(4)	0.72	0.66	0.66	0.68	0.62	0.68
		(3b)	0.74	0.65	0.65	0.69	0.62	0.69
		(3c)						

Which one predicts most accurately, and which one least?
When measuring accuracy against what benchmark?

International Wealth Index (IWI)

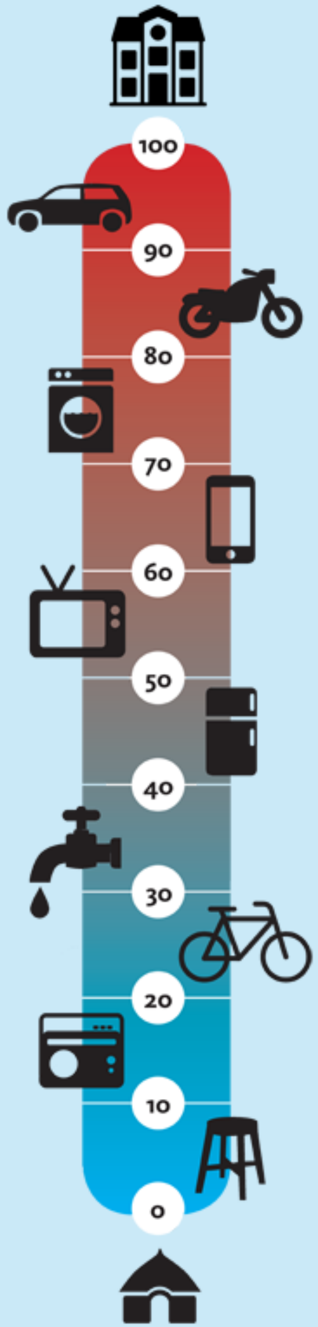
Does the household own or have a:

TV:	Yes	No	Unknown
Refrigerator:	Yes	No	Unknown
Phone:	Yes	No	Unknown
Bike:	Yes	No	Unknown
Car:	Yes	No	Unknown
Cheap utensils (<$50):	Yes	No	Unknown
Expensive utensil (>$300):	Yes	No	Unknown
Electricity:	Yes	No	Unknown

What is the quality of the...

Main source drinking water?:	Low	Middle	High	Unknown
Toilet facility usually used?:	Low	Middle	High	Unknown
Main floor material?:	Low	Middle	High	Unknown
Nr. of rooms used for sleeping:	One	Two	Three+	Unknown

The household's IWI score is: **12.73**



https://globaldatalab.org/iwi/form/

International Wealth Index (IWI)

Does the household own or have a:

TV:	Yes	No	Unknown	
Refrigerator:	Yes	No	Unknown	
Phone:	Yes	No	Unknown	
Bike:	Yes	No	Unknown	
Car:	Yes	No	Unknown	
Cheap utensils (<$50):	Yes	No	Unknown	
Expensive utensil (>$300):	Yes	No	Unknown	
Electricity:	Yes	No	Unknown	

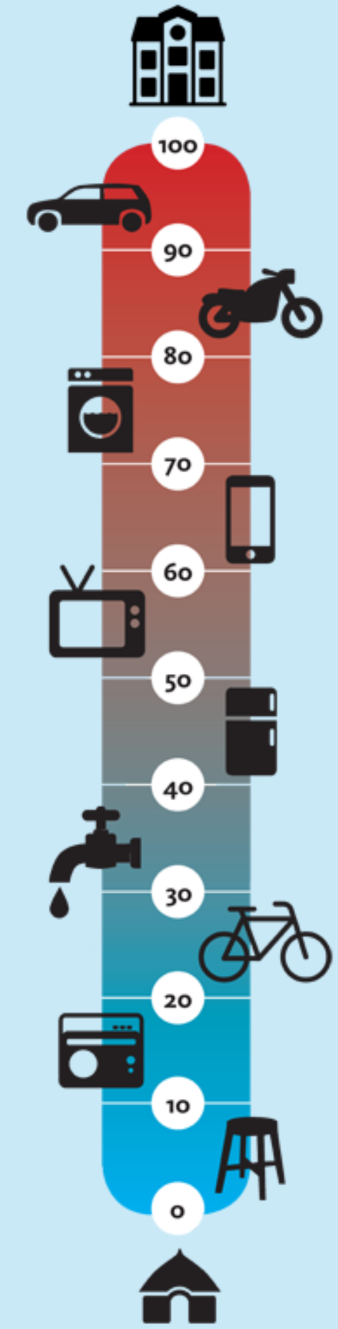
What is the quality of the...

Main source drinking water?:	Low	Middle	High	Unknown
Toilet facility usually used?:	Low	Middle	High	Unknown
Main floor material?:	Low	Middle	High	Unknown
Nr. of rooms used for sleeping:	One	Two	Three+	Unknown

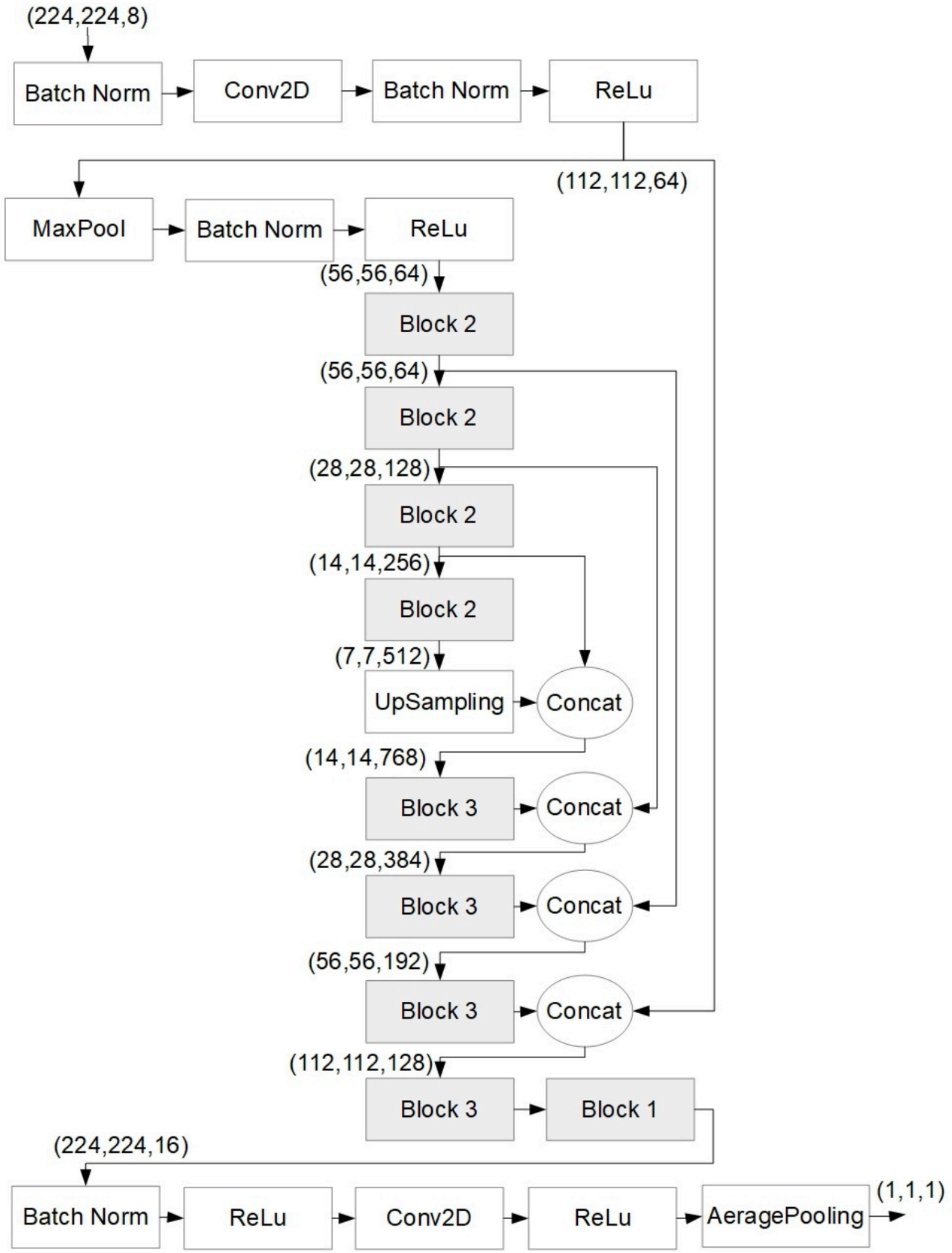
The household's IWI score is: **12.73**

<u>With TV</u> =12.73

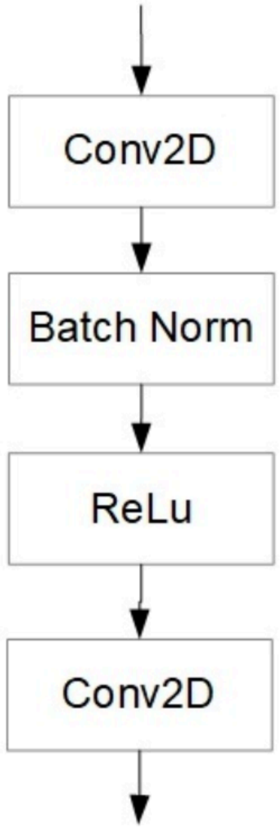
<u>Without TV</u> = 4.12



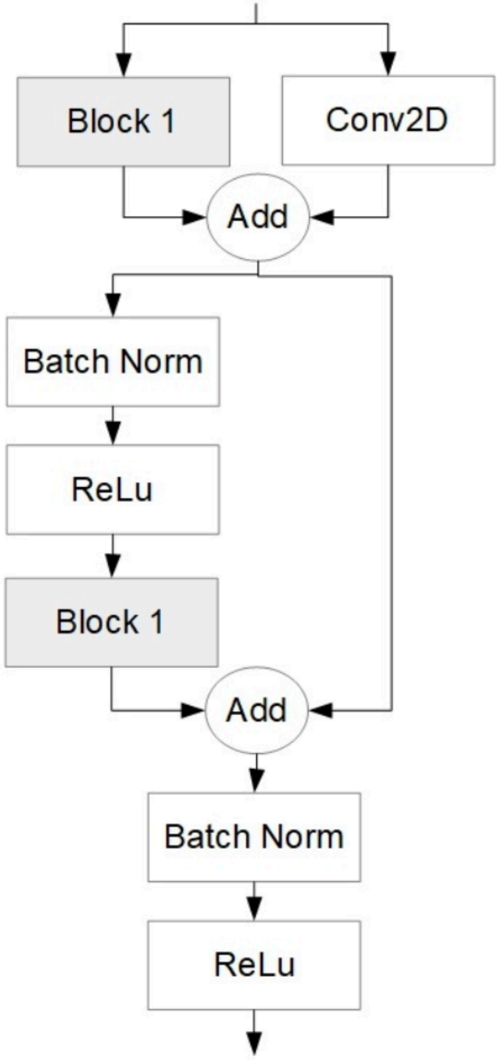
https://globaldatalab.org/iwi/form/



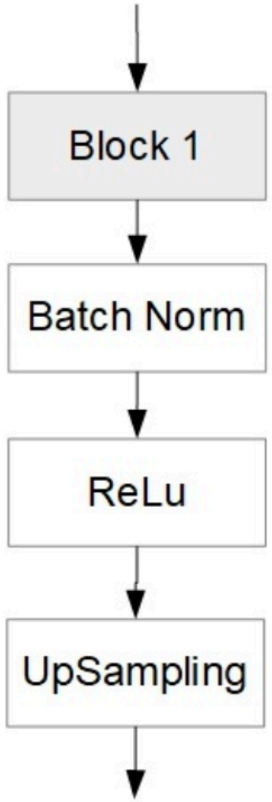
(224,224,8)
Batch Norm
Conv2D
Batch Norm
ReLu
(112,112,64)
MaxPool
Batch Norm
ReLu
(56,56,64)
Block 2
(56,56,64)
Block 2
(28,28,128)
Block 2
(14,14,256)
Block 2
(7,7,512)
UpSampling
Concat
(14,14,768)
Block 3
Concat
(28,28,384)
Block 3
Concat
(56,56,192)
Block 3
Concat
(112,112,128)
Block 3
Block 1
(224,224,16)
Batch Norm
ReLu
Conv2D
ReLu
AeragePooling
(1,1,1)



Block 1



Block 2



Block 3