

Machine Learning for Official Statistics

InKyung Choi

UN Economic Commission for Europe
Statistics Division

EXPO2020

Mobilizing Big Data and Data Science for the Sustainable Development Goals Event

(January 26, 2022)



Content

1. **What** is machine learning?
2. **Why** machine learning?
3. **How** can machine learning help official statistics?

1. What is Machine Learning?

Machine Learning

- Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed

Machine Learning

Description of work-related injury of someone

“I cut my fingers while chopping vegetables”

Is this person a cook or a statistician?



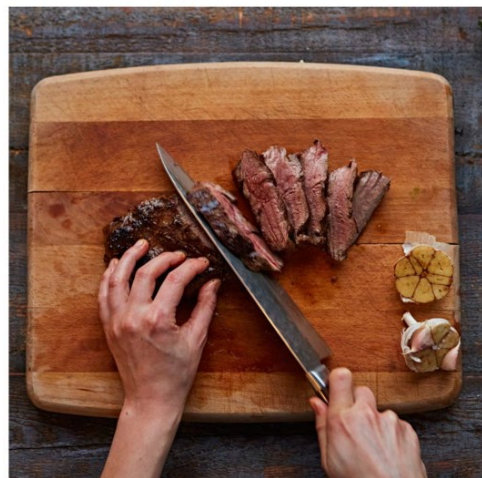
Machine Learning

"I cut my fingers while chopping vegetables"

"burnt hands from oven"

"strained arms while carrying frozen meat"

"I got hill pain from 'Chef's foot' "



"Files fell on me"

"Backpain, need better workstation ergonomics"



"I looked at monitor too long, I got dry eyes"

"Office gate shut too hard and bruised my arm"

Machine Learning

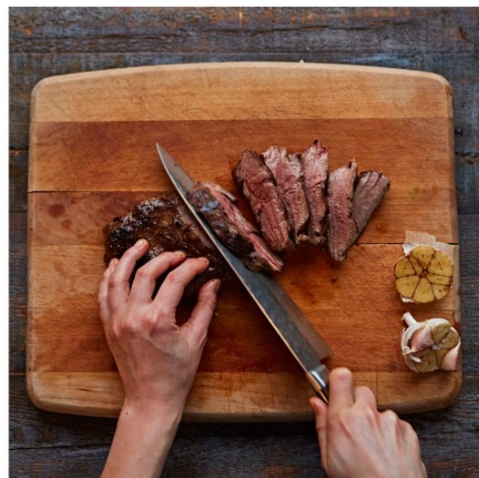
What humans know from **experiences**, machines can learn from **data**

*"I cut my fingers while chopping **vegetables**"*

*"**burnt** hands from **oven**"*

*"strained arms while carrying **frozen meat**"*

*"I got hill pain from '**Chef's foot**' "*



*"**Files** fell on me"*

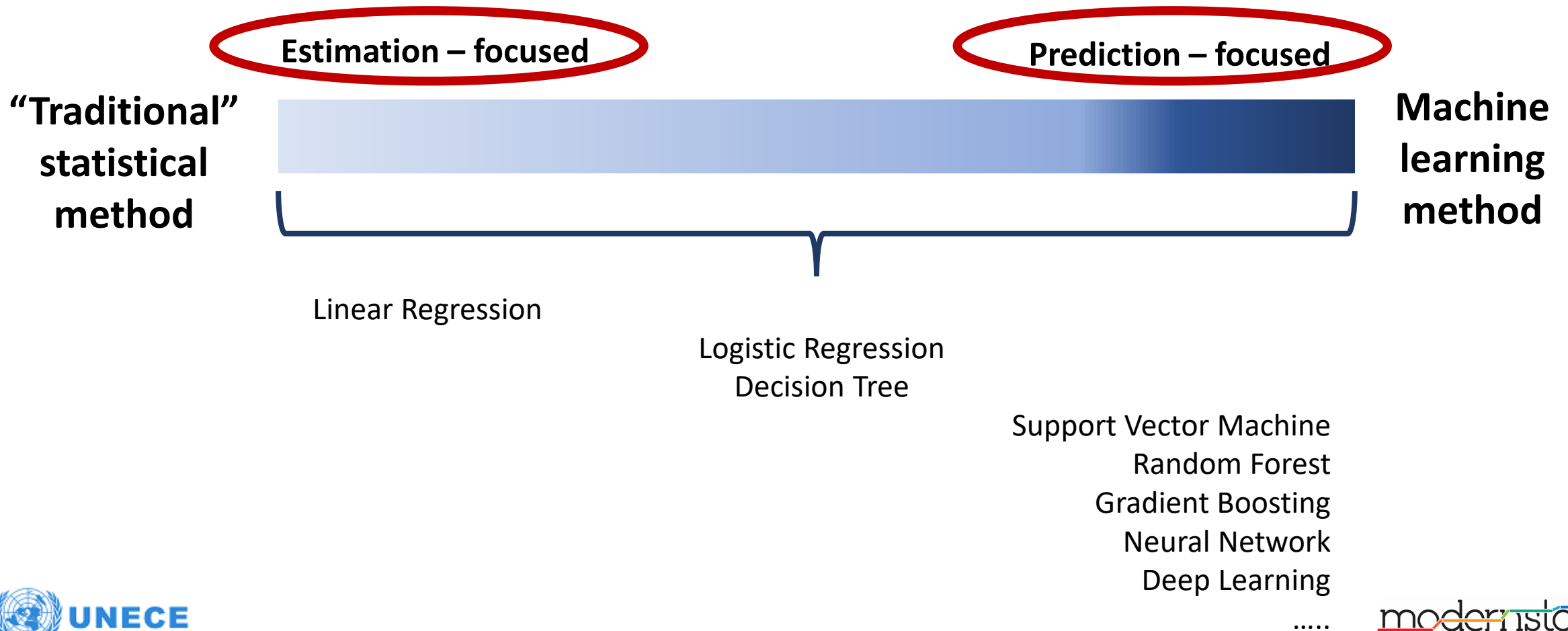
*"Backpain, need better **workstation ergonomics**"*



*"I looked at **monitor** too long, I got dry eyes"*

*"**Office** gate shut too hard and bruised my arm"*

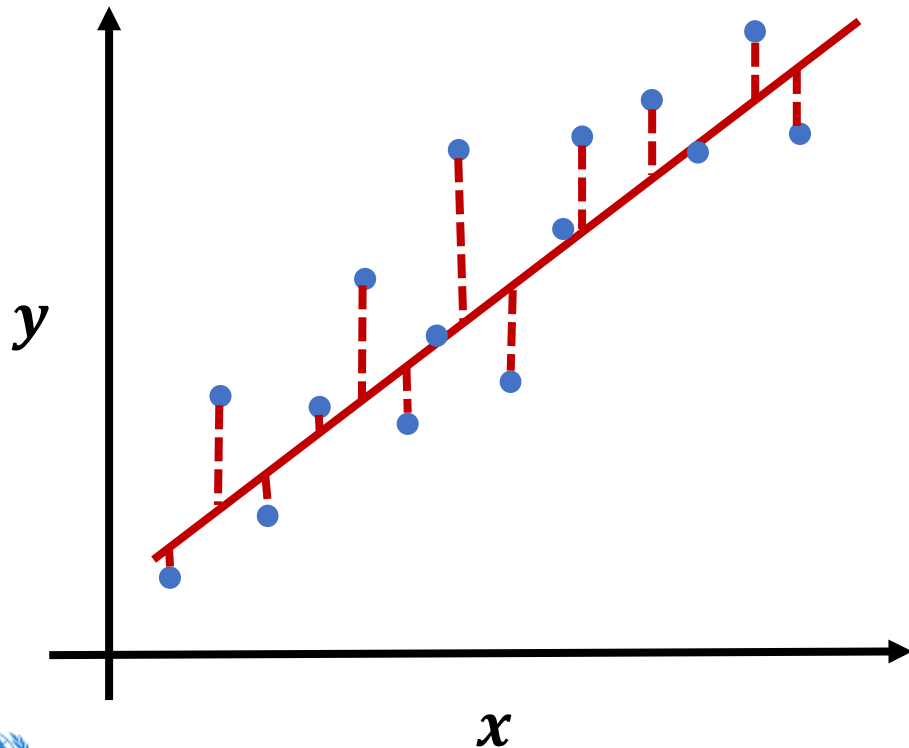
What is difference from “traditional” statistical method?



What is difference from “traditional” statistical method?

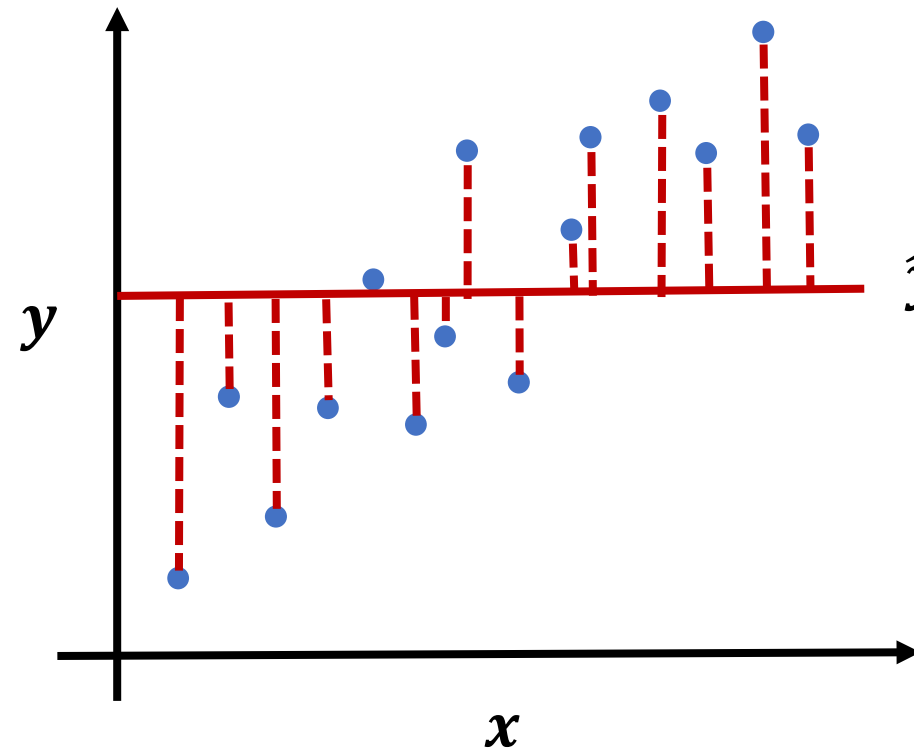
Linear regression

$$\hat{y} = \beta_0 + \beta_1 x$$



Gradient boosting

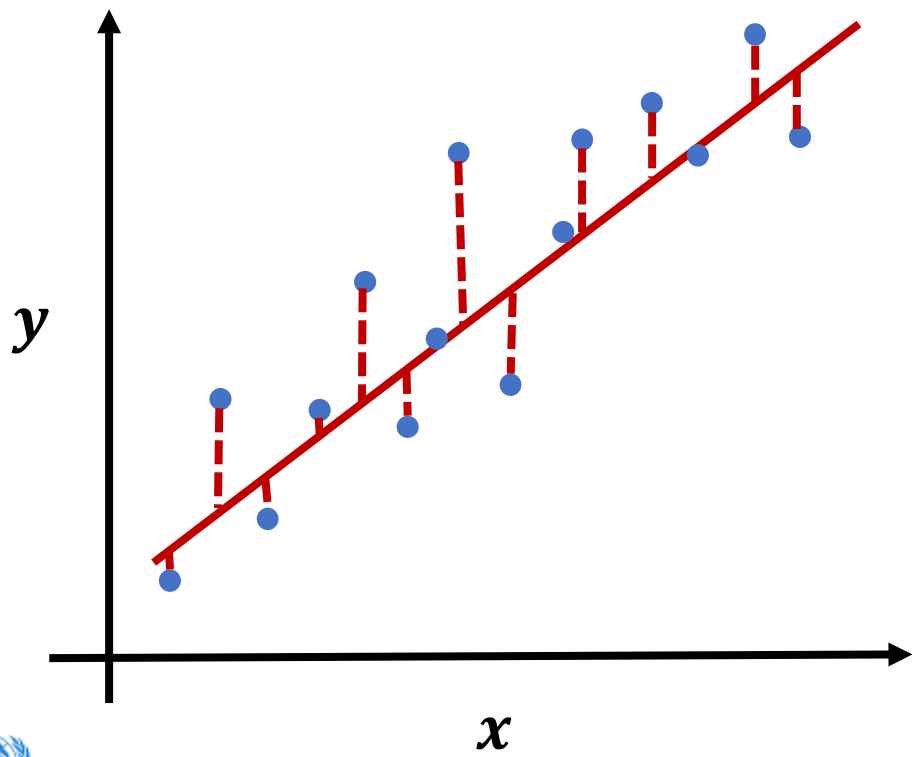
$$\hat{y} = f_0(x)$$



What is difference from “traditional” statistical method?

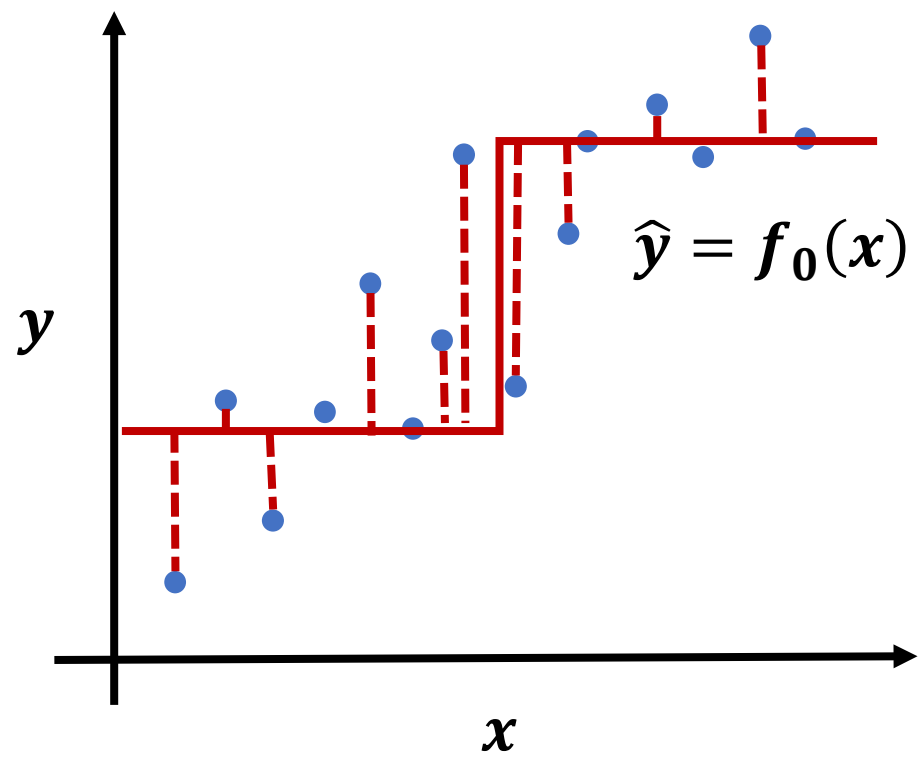
Linear regression

$$\hat{y} = \beta_0 + \beta_1 x$$



Gradient boosting

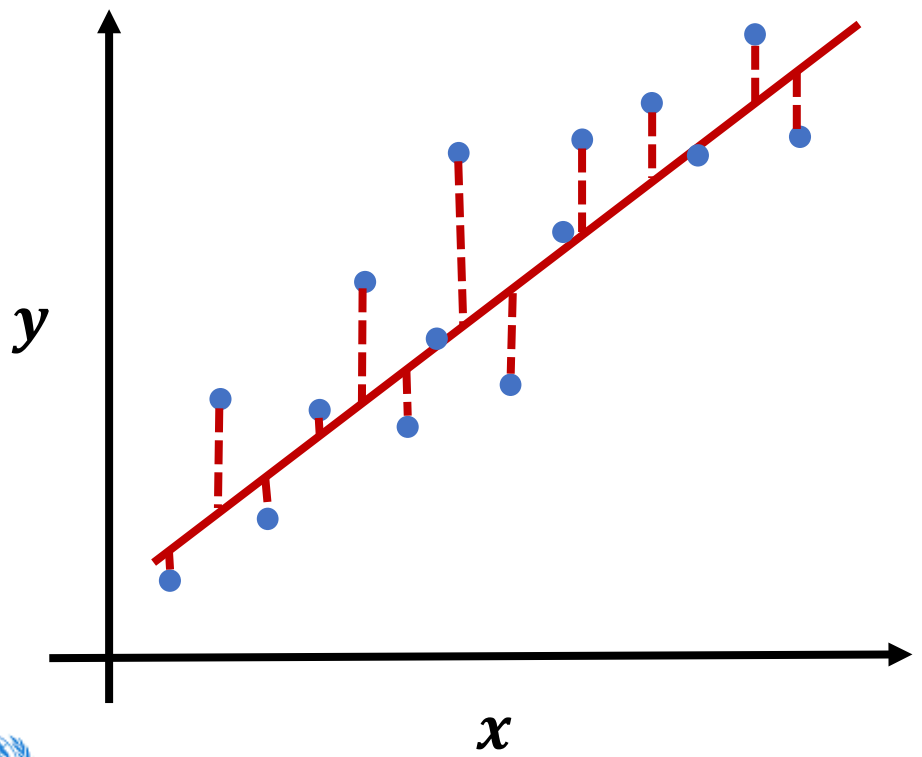
$$\hat{y} = f_0(x) + \Delta_1(x)$$



What is difference from “traditional” statistical method?

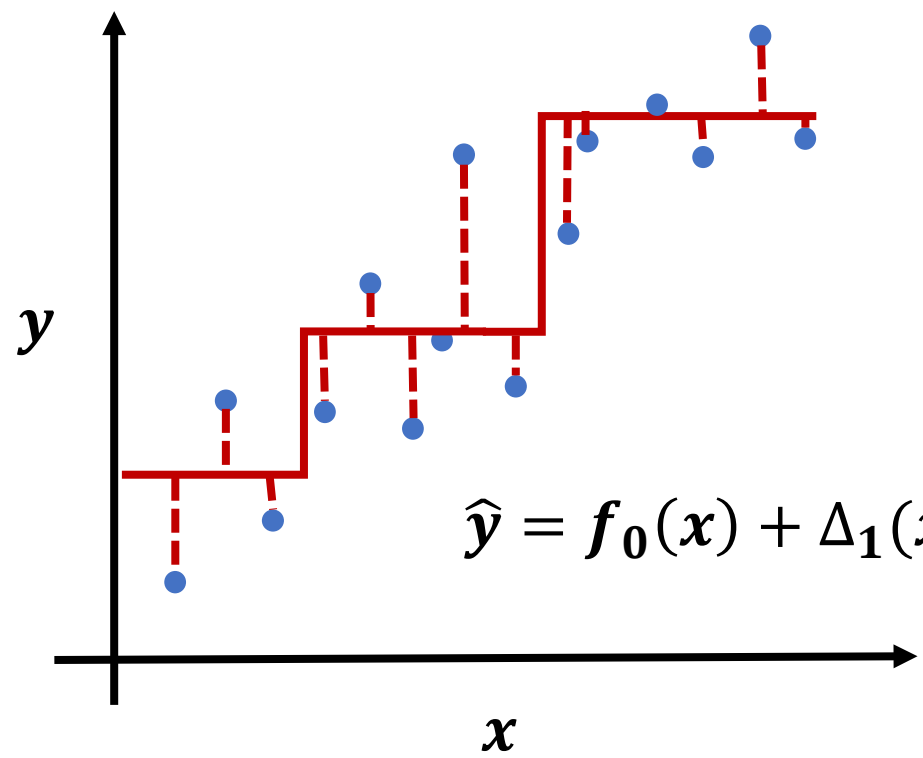
Linear regression

$$\hat{y} = \beta_0 + \beta_1 x$$



Gradient boosting

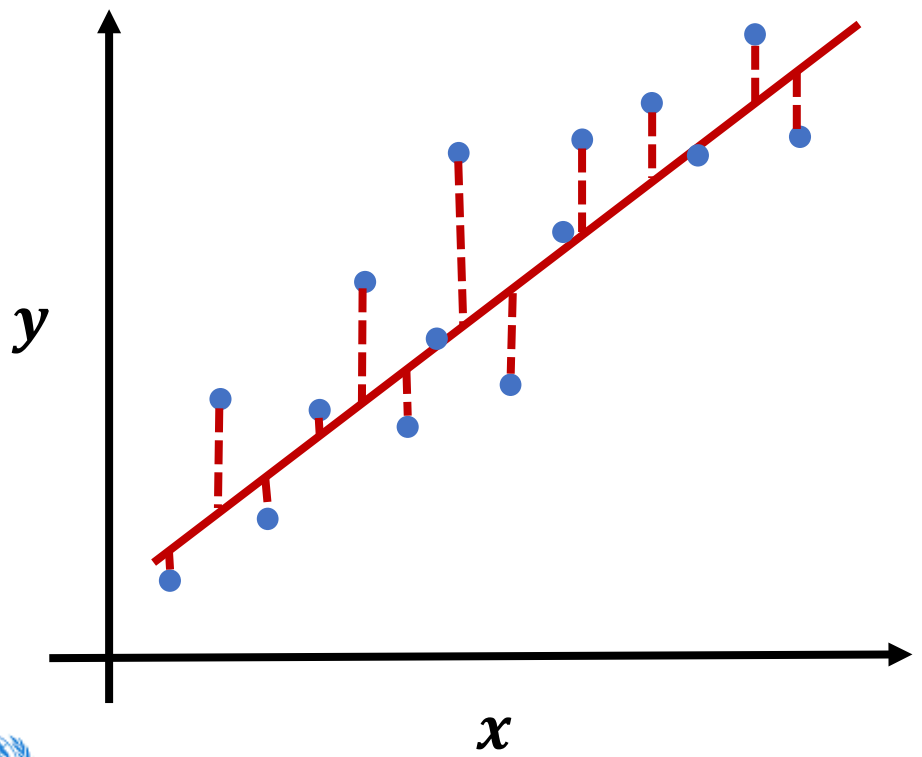
$$\hat{y} = f_0(x) + \Delta_1(x) + \Delta_2(x)$$



What is difference from “traditional” statistical method?

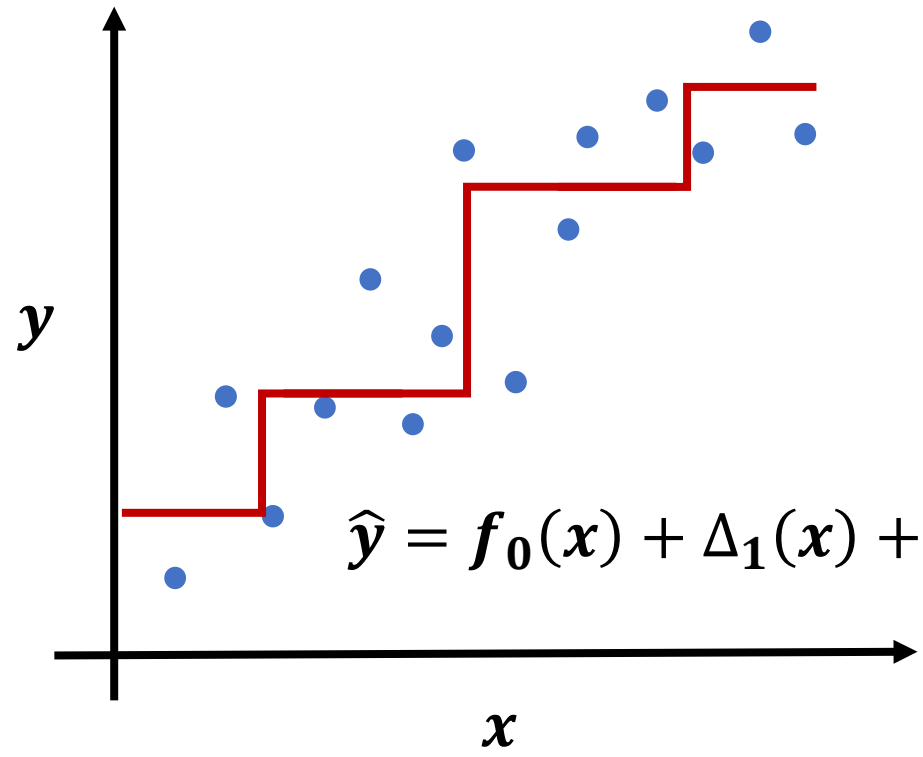
Linear regression

$$\hat{y} = \beta_0 + \beta_1 x$$



Gradient boosting

$$\hat{y} = f_0(x) + \Delta_1(x) + \dots + \Delta_m(x)$$



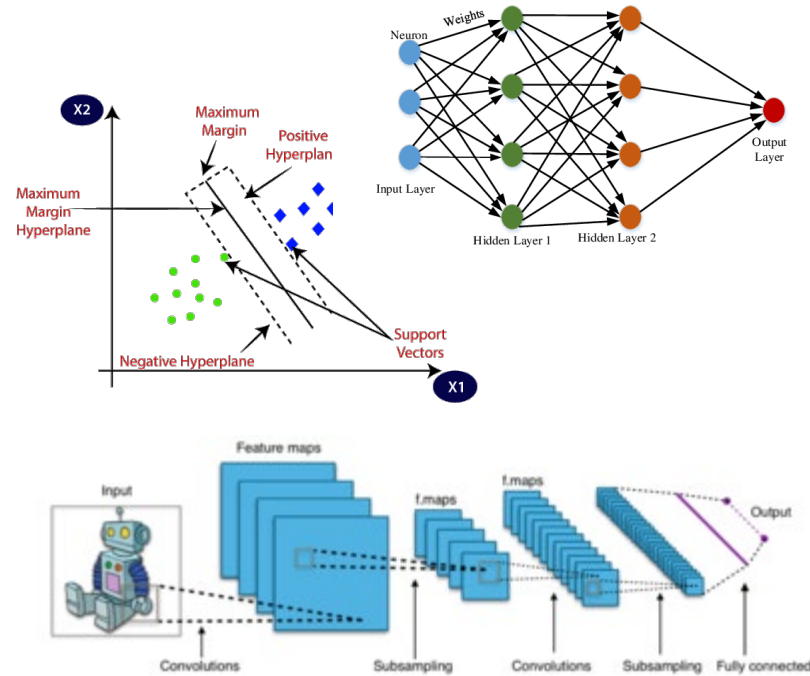
2. Why Machine Learning?

Machine Learning , why now?

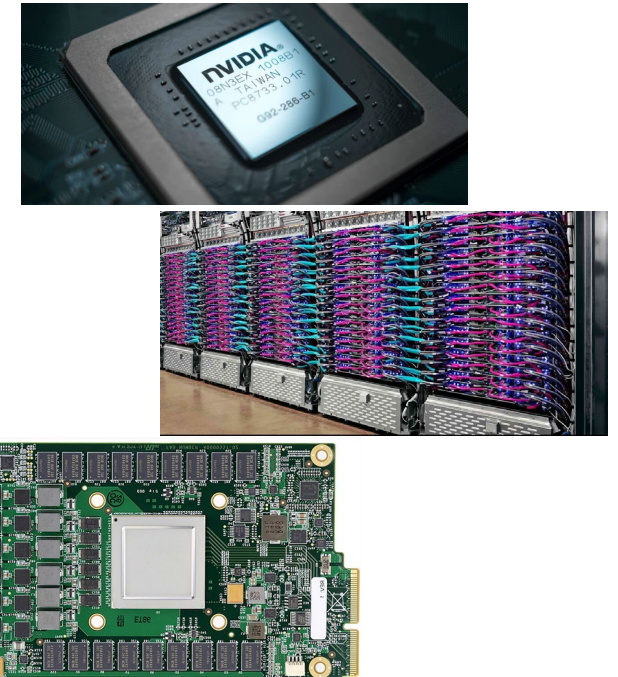
Data



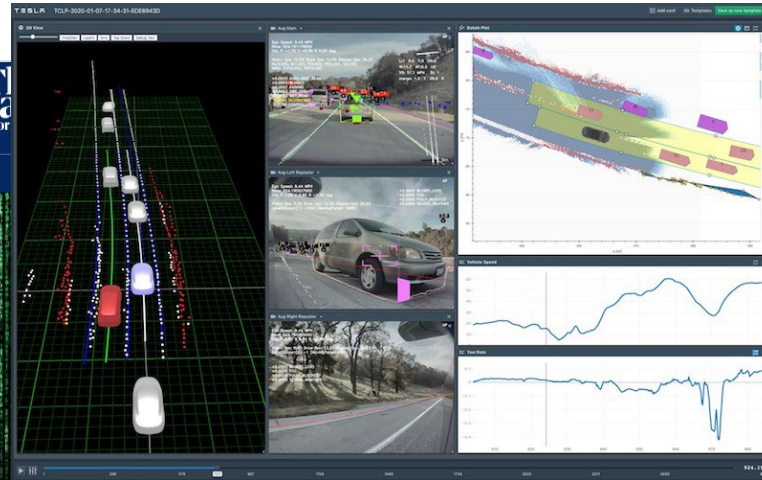
Methodology



Technology



Machine Learning , why now?



Opinion
A robot wrote this en
you scared yet, huma
GPT-3

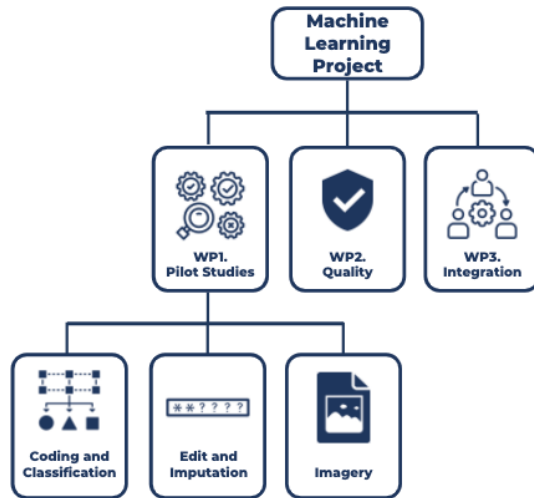


Next Rembrandt project

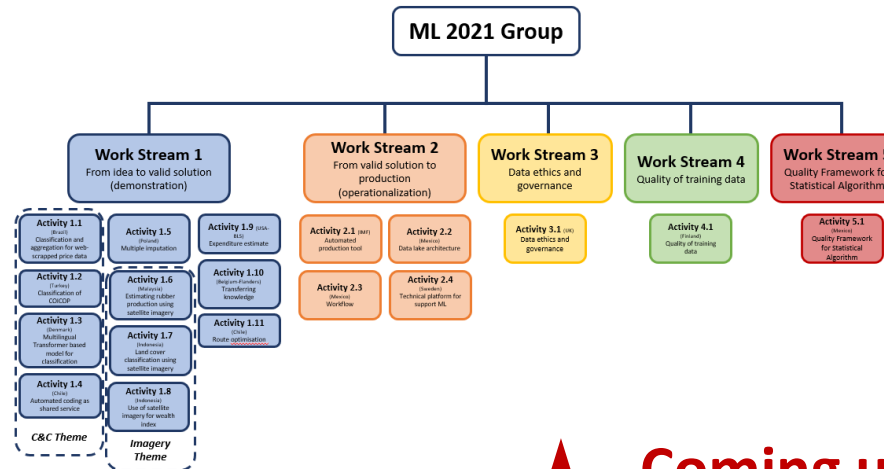


Is it for official statistics....?

UNECE HLG-MOS Machine Learning Project
(2019-20)



UK ONS-UNECE Machine Learning Group
(2021)

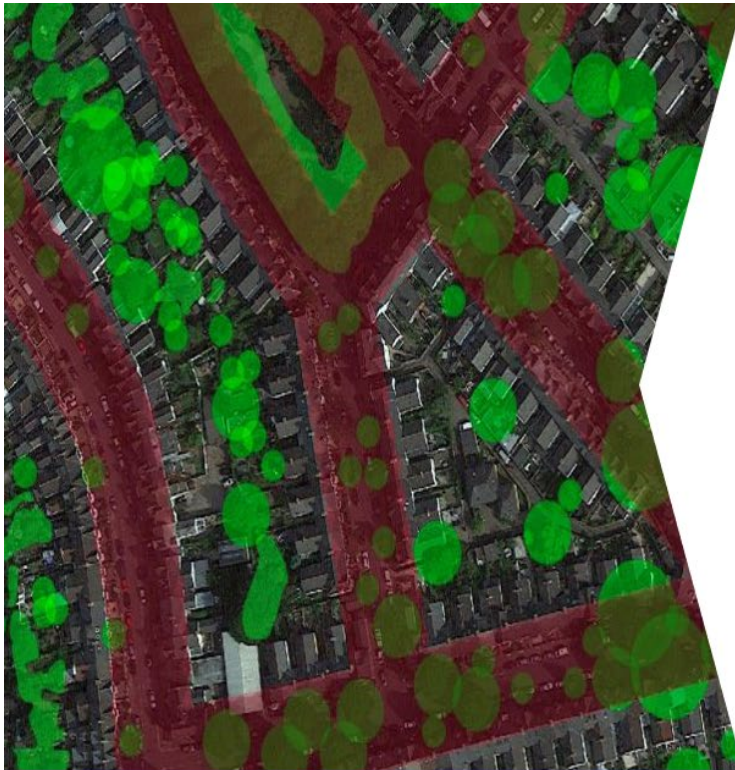


**250+ members
around 33 countries
38 pilot studies &
researches**

★ Coming up
UK ONS-UNECE
Machine Learning Group **2022**

3. How can machine learning help official statistics?

Machine learning for official statistics



Why?

"Any industry with very large amounts of data — so much that humans can't possibly analyz[s]e or understand it on their own — can utilize AI"

[Gartner 2017](#)

© Satellite imagery copyright Google

datasciencecampus.ons.gov.uk

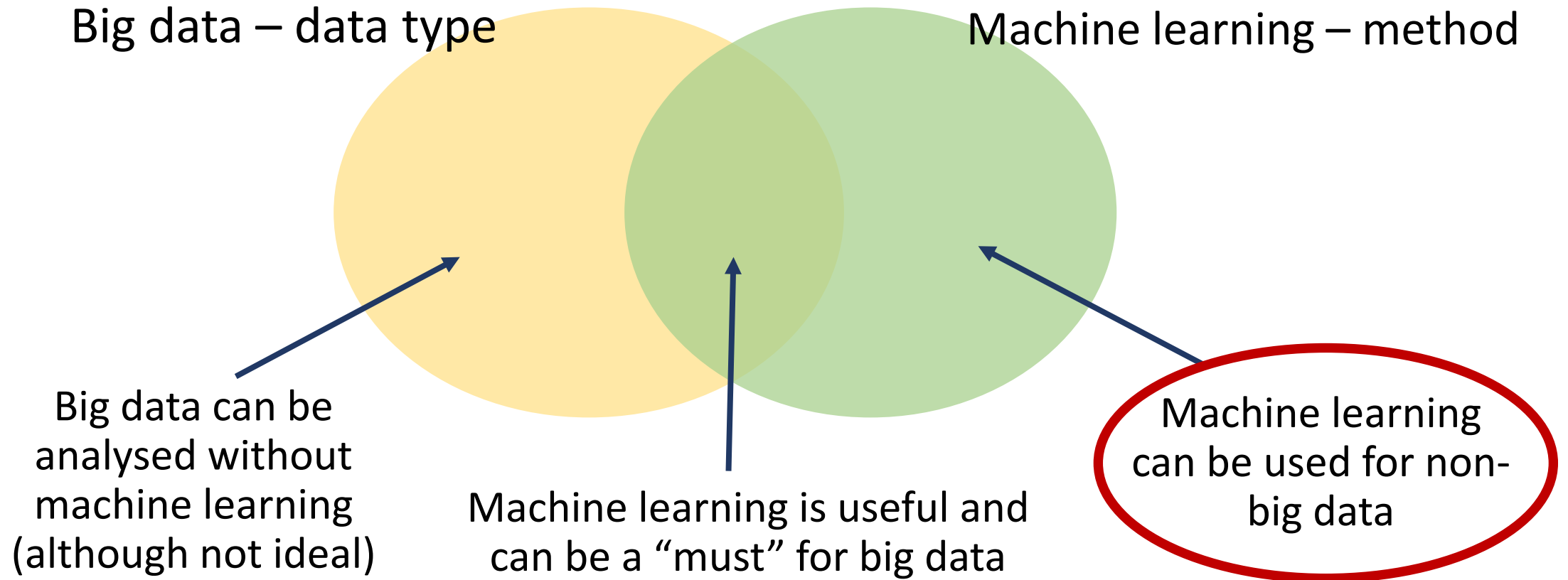
datasciencecampus@ons.gov.uk

[@DataSciCampus](https://twitter.com/DataSciCampus)

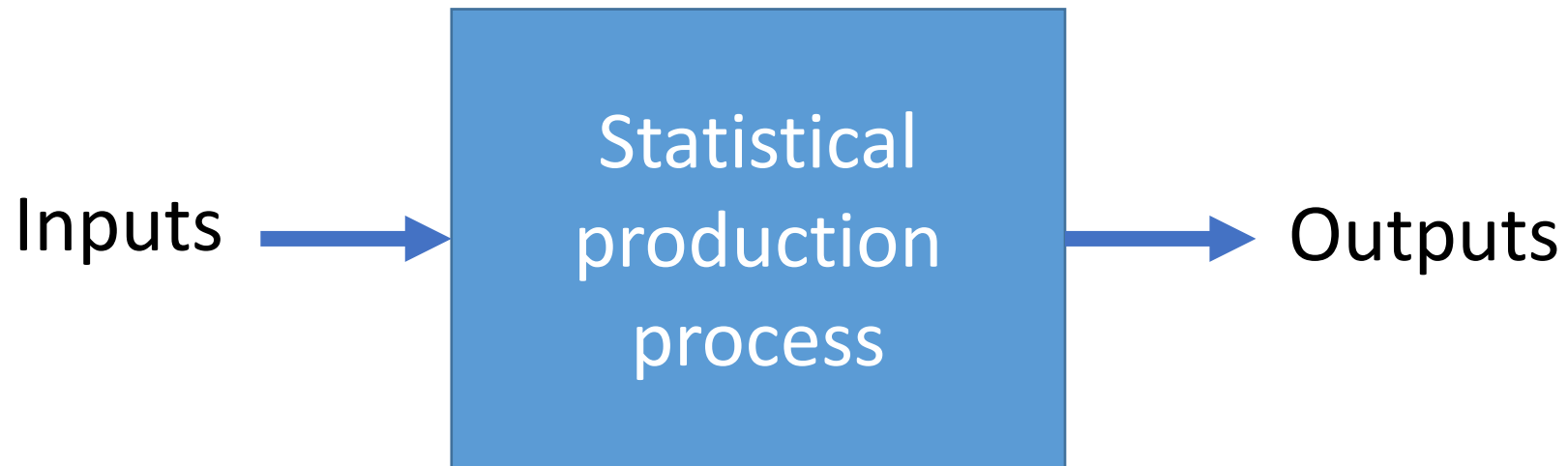


Office for
National Statistics

Machine learning and big data

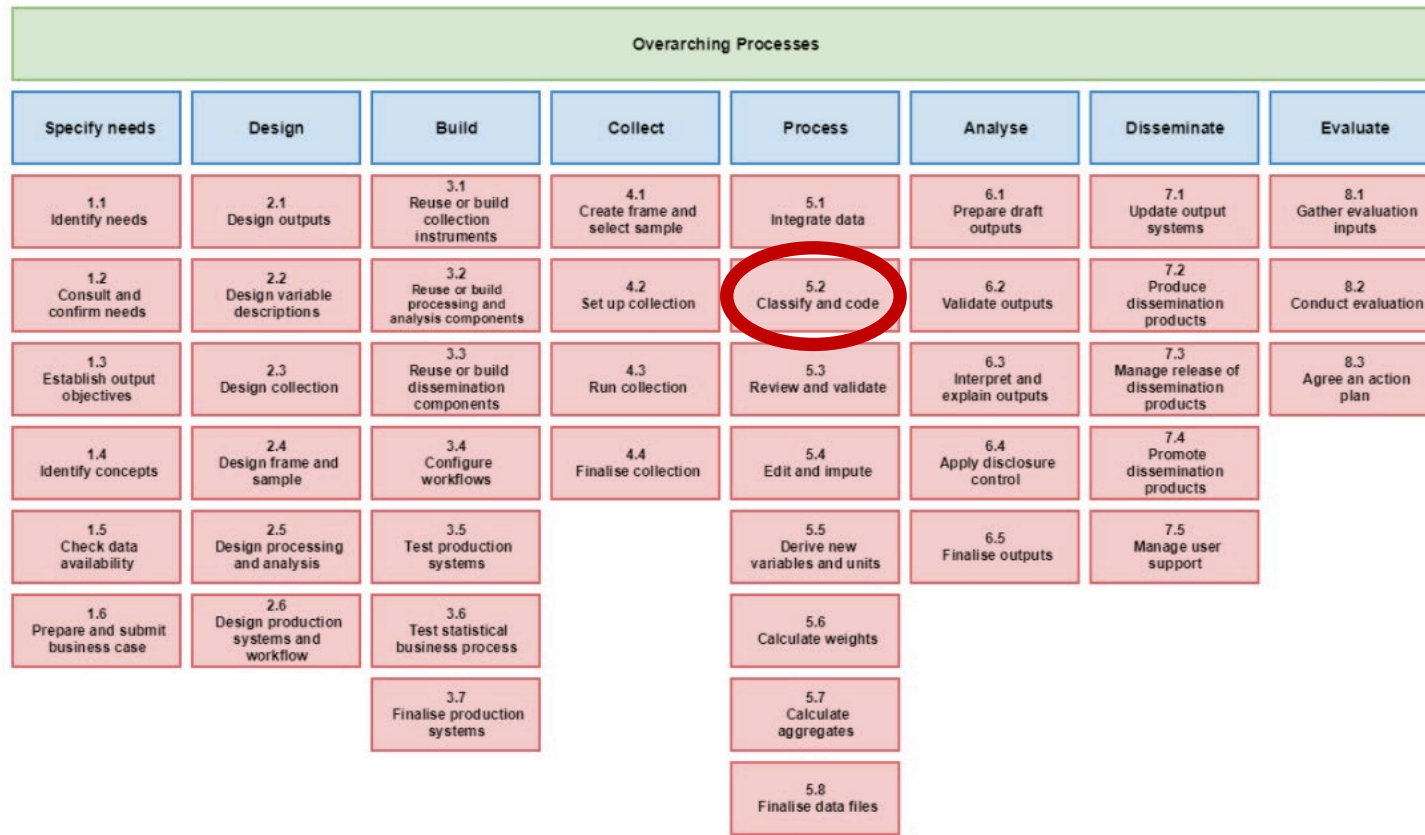


Statistical production process



Statistical production process

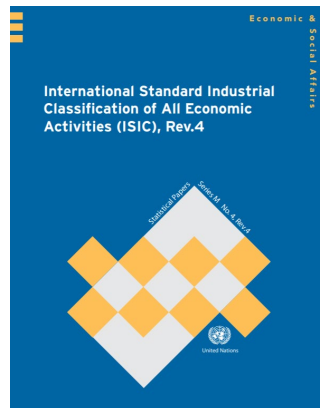
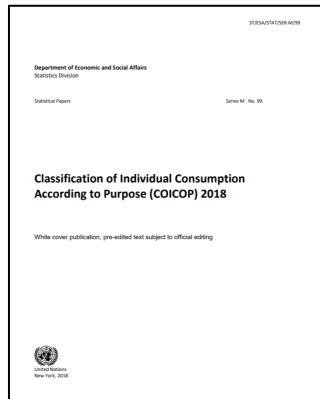
Inputs



Outputs

Machine learning – application areas

5.2 Classify and code



Survey of Occupational Injuries and Illnesses

Example Narrative

Job title: sanitation worker

What was the employee doing just before the incident?
mopping floor in gym

What happened?
slipped on water on floor and fell

What part of the body was affected?
fractured right arm

What object directly harmed the employee?
wet floor



Codes Assigned

Occup: 37-2011 (Janitor)

Nature: 111 (Fracture)

Part: 420 (Arm)

Event: 422 (Fall, slipping)

Source: 6620 (Floor)

Secondary: 9521(Water)

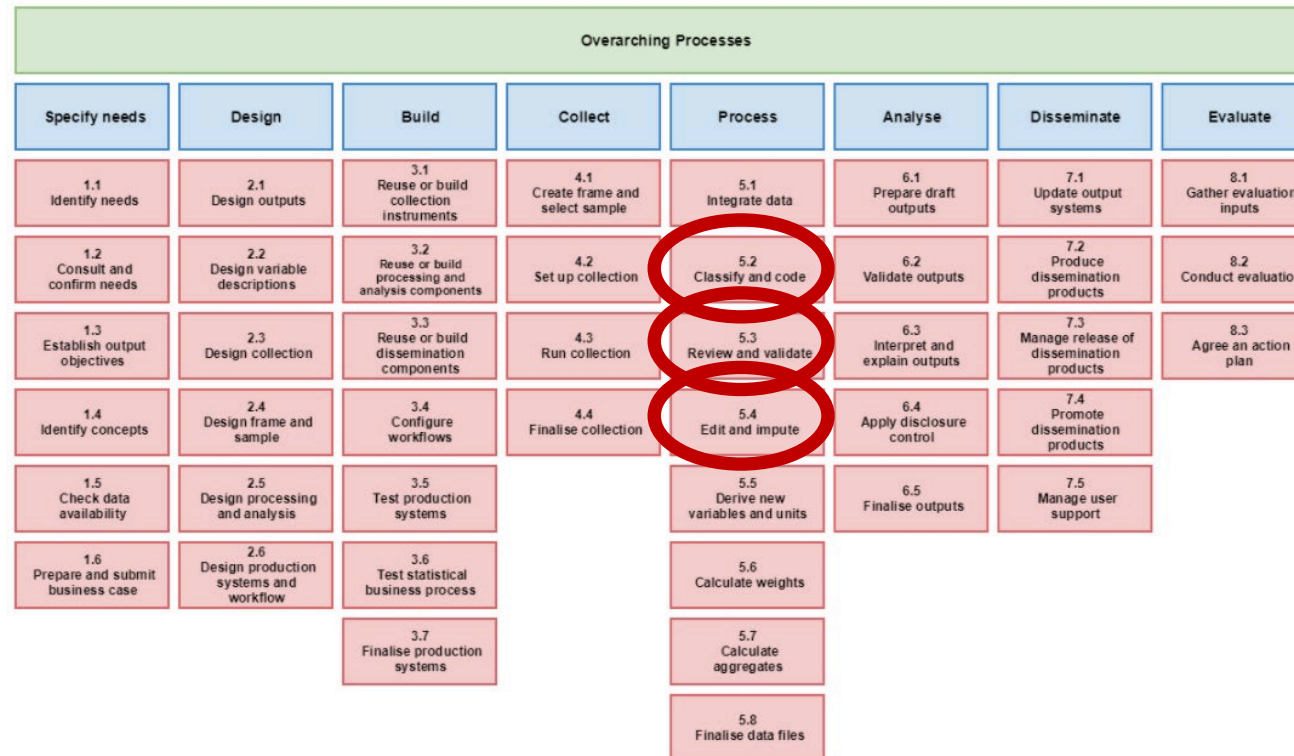
2 — U.S. BUREAU OF LABOR STATISTICS • bls.gov

2



ML Project Coding Pilot Study from U.S.

Machine learning – application areas

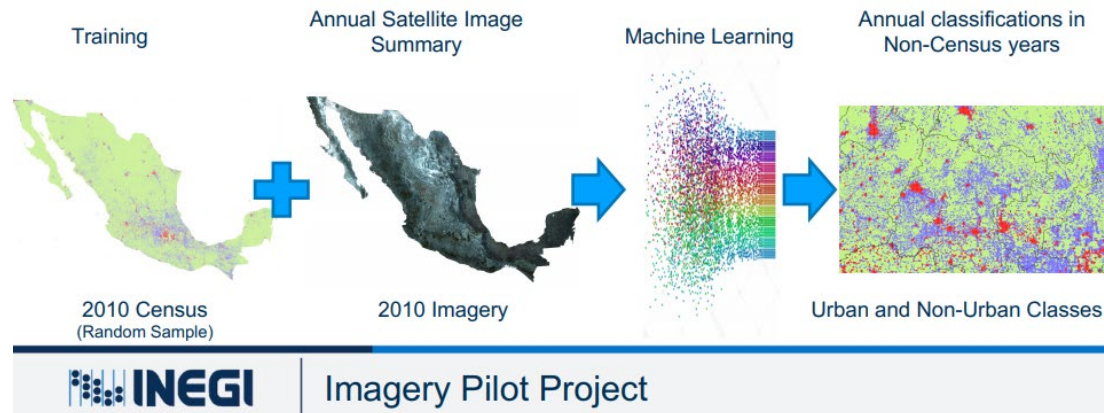


Areas with manual, repetitive works can be automated with help of machine learning

Machine learning – application areas

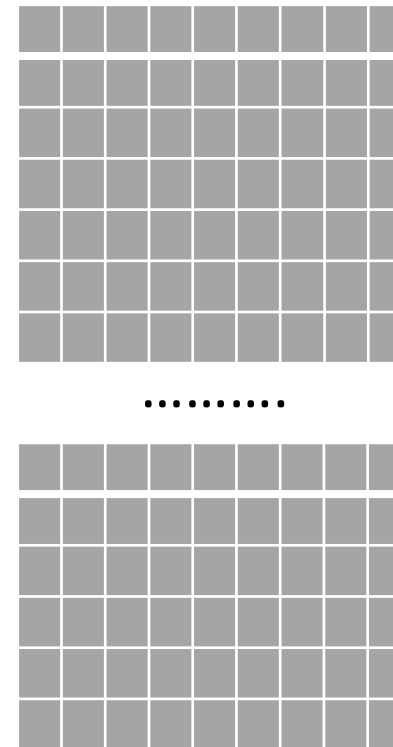
Objective of this Imagery Pilot Project (Practical Application)

Expand the use of imagery data in the production of official statistics through the further development of knowledge and sharing of ML solutions and practices.



ML Project Imagery Pilot Study from Mexico

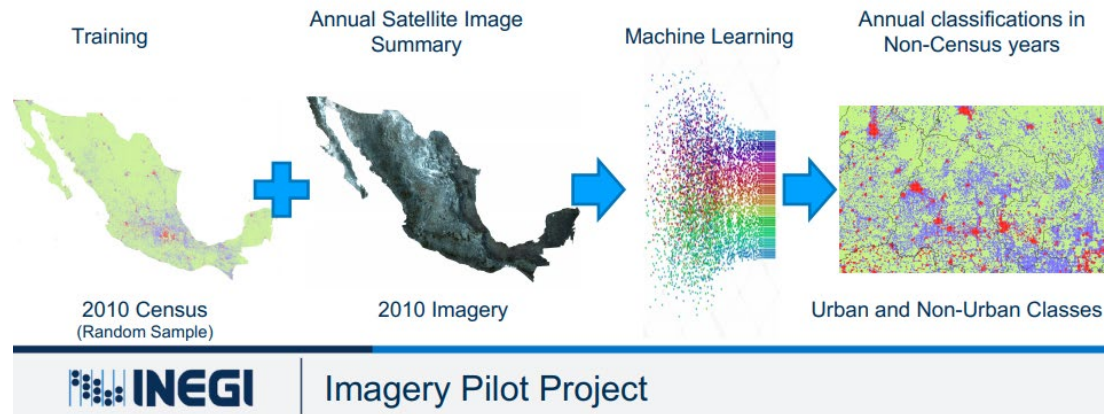
1,975,719 (1km x 1km) grid cells



Machine learning – application areas

Objective of this Imagery Pilot Project (Practical Application)

Expand the use of imagery data in the production of official statistics through the further development of knowledge and sharing of ML solutions and practices.

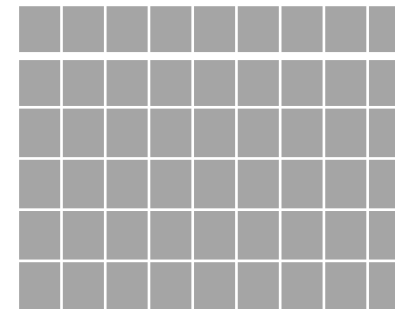
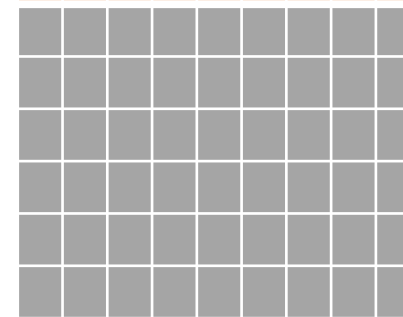


ML Project Imagery Pilot Study from Mexico

1,975,719 (1km x 1km) grid cells



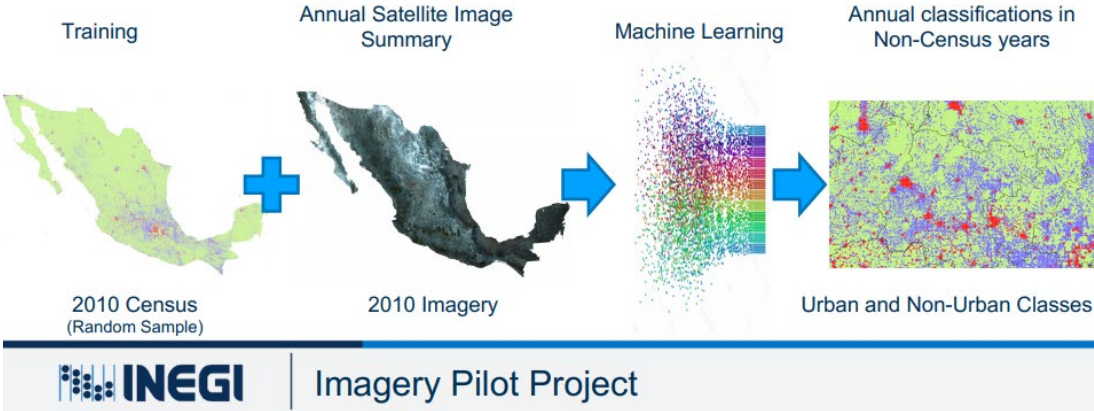
40,000 done by human



Machine learning – application areas

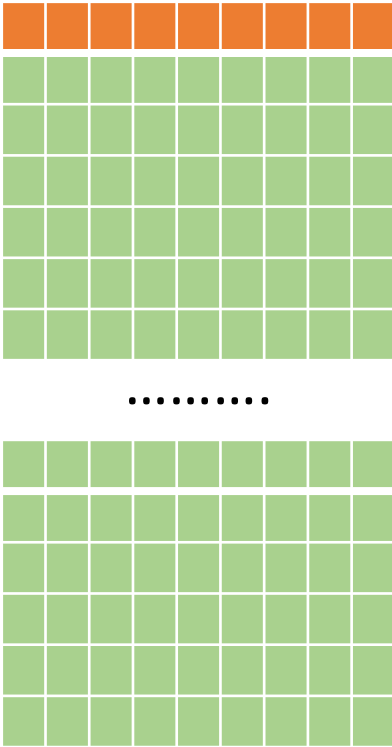
Objective of this Imagery Pilot Project (Practical Application)

Expand the use of imagery data in the production of official statistics through the further development of knowledge and sharing of ML solutions and practices.



ML Project Imagery Pilot Study from Mexico

1,975,719 (1km x 1km) grid cells



40,000 done by human

based on this data
1,935,719 done by ML

Machine learning - quality implications

United Nations National Quality Assurance Framework quality principles and supporting Fundamental Principles of Official Statistics

Quality principles	Fundamental Principles of Official Statistics									
	1	2	3	4	5	6	7	8	9	10
Level A: Managing the statistical system										
1: Coordinating the national statistical system								*		
2: Managing relationships with data users, data providers and other stakeholders	*				*			○		○
3: Managing statistical standards									*	
Level B: Managing the institutional environment										
4: Assuring professional independence	○	*						○		
5: Assuring impartiality and objectivity	*	○	○	○	○			○		
6: Assuring transparency			*					○		
7: Assuring statistical confidentiality and data security						*				
8: Assuring commitment to quality		*								
9: Assuring adequacy of resources	○									
Level C: Managing statistical processes										
10: Assuring methodological soundness		*			○				○	○
11: Assuring cost-effectiveness					*				○	
12: Assuring appropriate statistical procedures		*			○					
13: Managing the respondent burden					*					
Level D: Managing statistical outputs										
14: Assuring relevance	*		○		○					
15: Assuring accuracy and reliability	*				○					
16: Assuring timeliness and punctuality	*				○					
17: Assuring accessibility and clarity	*		○							
18: Assuring coherence and comparability	*		○						○	
19: Managing metadata			*						○	

Quality Framework for Statistical Algorithms

- Timeliness
- Accuracy
- Cost-effectiveness
- Explainability
- Reproducibility

UN National Quality Assurance Framework

Machine learning - quality implications

United Nations National Quality Assurance Framework quality principles and supporting Fundamental Principles of Official Statistics

Quality principles	Fundamental Principles of Official Statistics									
	1	2	3	4	5	6	7	8	9	10
Level A: Managing the statistical system										
1: Coordinating the national statistical system								*		
2: Managing relationships with data users, data providers and other stakeholders	*				*			○		○
3: Managing statistical standards									*	
Level B: Managing the institutional environment										
4: Assuring professional independence	○	*						○		
5: Assuring impartiality and objectivity	*	○	○	○	○			○		
6: Assuring transparency			*					○		
7: Assuring statistical confidentiality and data security						*				
8: Assuring commitment to quality		*								
9: Assuring adequacy of resources	○									
Level C: Managing statistical processes										
10: Assuring methodological soundness		*			○				○	○
11: Assuring cost-effectiveness					*				○	
12: Assuring appropriate statistical procedures		*			○					
13: Managing the respondent burden					*					
Level D: Managing statistical outputs										
14: Assuring relevance	*		○		○					
15: Assuring accuracy and reliability	*				○					
16: Assuring timeliness and punctuality	*				○					
17: Assuring accessibility and clarity	*		○							
18: Assuring coherence and comparability	*		○						○	
19: Managing metadata			*						○	

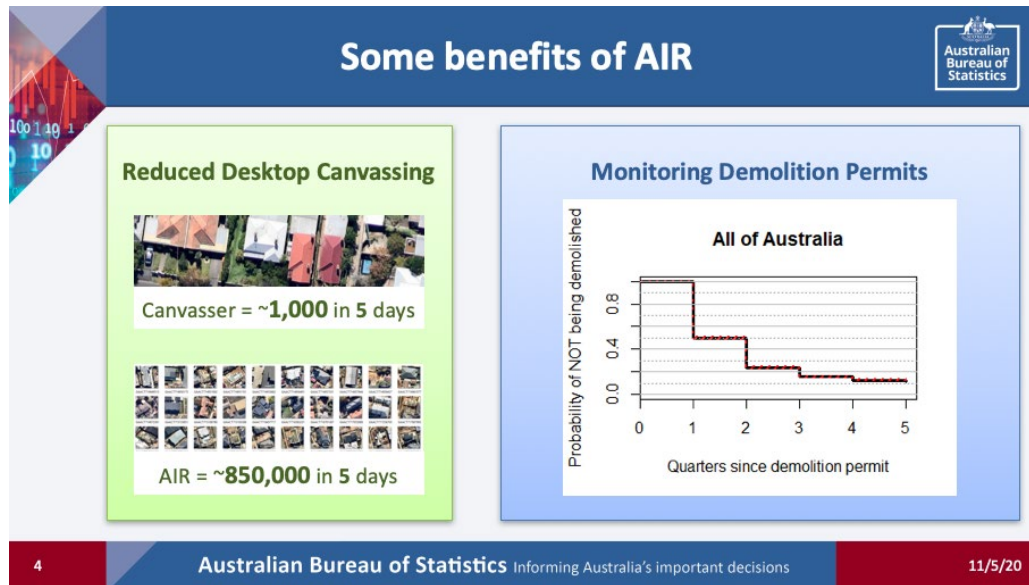
Quality Framework for Statistical Algorithms

- **Timeliness**
- **Accuracy**
- Cost-effectiveness
- **Explainability**
- Reproducibility

UN National Quality Assurance Framework

Machine learning - quality implications

Timeliness



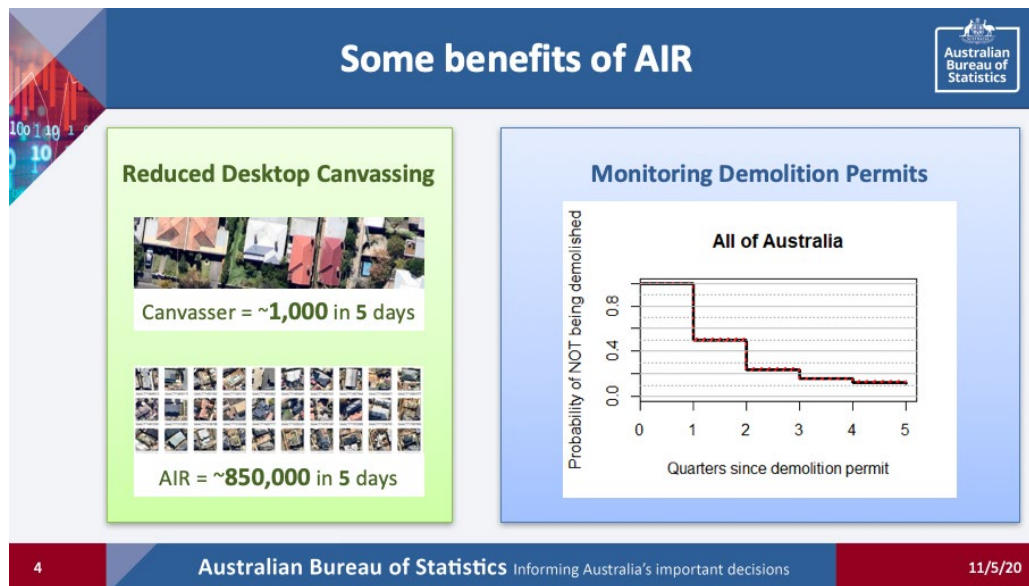
ML Project Imagery Pilot Study from Australia

Accuracy

$$\text{Prediction error} = \frac{\# \text{ correctly predicted}}{\# \text{ total}}$$

Machine learning - quality implications

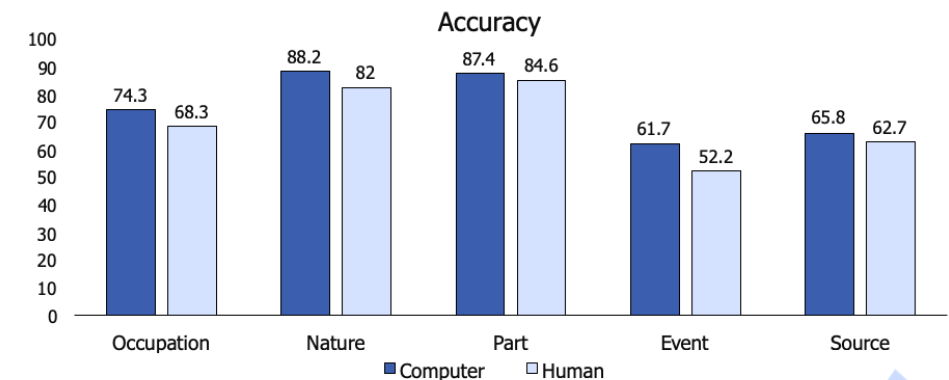
Timeliness



ML Project Imagery Pilot Study from Australia

Accuracy

Machine Learning vs. Manual Process



4 — U.S. BUREAU OF LABOR STATISTICS • bls.gov



ML Project Coding Pilot Study from U.S.

Machine learning - quality implications

Explainability



Introduction to Quality Framework for Statistical Algorithms (QF4SA) in Session 1.2

Machine learning for official statistics

Some final remarks

- ML can be used not only for big data but also for non-big data
- There are advanced methods, but simple methods work well too
- Depending on use case, different emphasis on different quality dimensions
- Sharing and collaboration is key to facilitating ML

Thank you for your attention

Resources

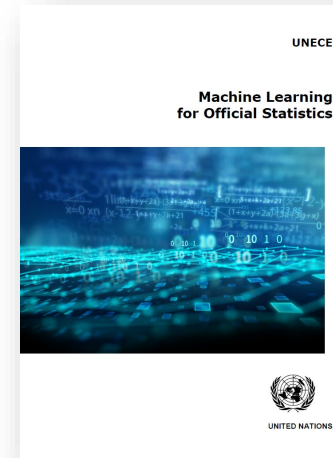
Machine Learning for Official Statistics Home

Created by Inkyung Choi, last modified on 07 Jan, 2022

Machine Learning for Official Statistics

- HLG-MOS Machine Learning Project (2019-20)**
The project launched in 2019 attracted more than 100 experts around the world. All the project outputs are made publicly available.
[Click here for more information](#)
- Learning and training**
Machine learning is widely used in many areas and there is certainly not lack of resources. Check out learning materials produced or recommended by HLG-MOS ML project team.
[Click here for more information](#)
- Studies and codes**
You can search ML studies using filter such as method (e.g. neural network, fasttext), programme language (e.g. python, R) or availability of the codes. Do you have a study that you want to share with community? Feel free to send it to us!
[Click here for more information](#)
- ONS-UNECE Machine Learning Group 2021**
Building on the work of the ML Project (2019-2020), the UK ONS, in partnership with the UNECE, launched Machine Learning Group 2021. It consisted of 5 Work Streams and conducted various knowledge sharing activities.
[Click here for more information](#)
- New ONS-UNECE Machine Learning Group 2022**
The international efforts for advancing the use of ML for official statistics continue in 2022.
[Click here for more information](#)

[Machine Learning for Official Statistics Wiki](#): all reports, pilot studies, codes, learning resources, etc.



*UNECE publication on
Machine Learning for
Official Statistics*