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)

UNECE

modernstats by HLG - MOS

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

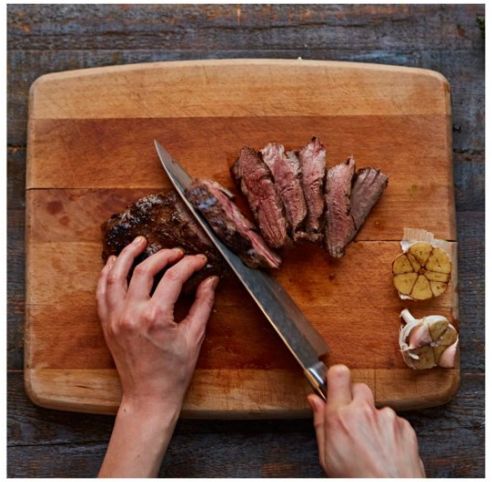
- Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed

Machine Learning

Description of work-related injury of someone

“I cut my fingers while chopping vegetables”

Is this person a cook or a statistician?

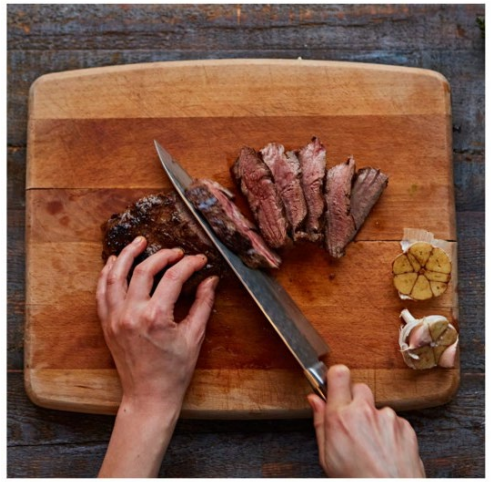


Machine Learning

"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"

UNECE

modernstats by HLG - MOS

Machine Learning

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

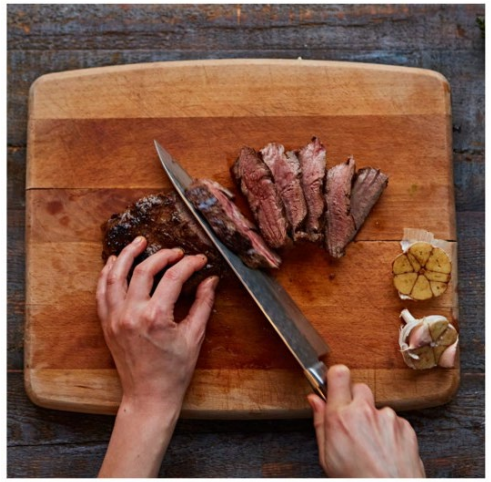
*"I cut my fingers while chopping **vegetables**"*

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

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

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

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



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

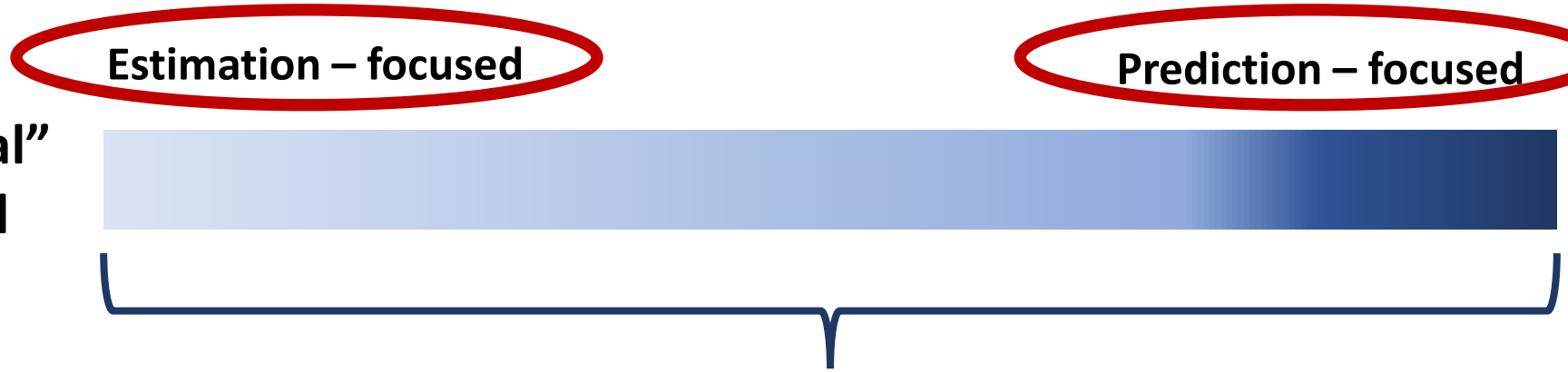
*"I got hill pain from '**Chef**'s foot' "*

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

UNECE

modernstats by HLG - MOS

What is difference from “traditional” statistical method?



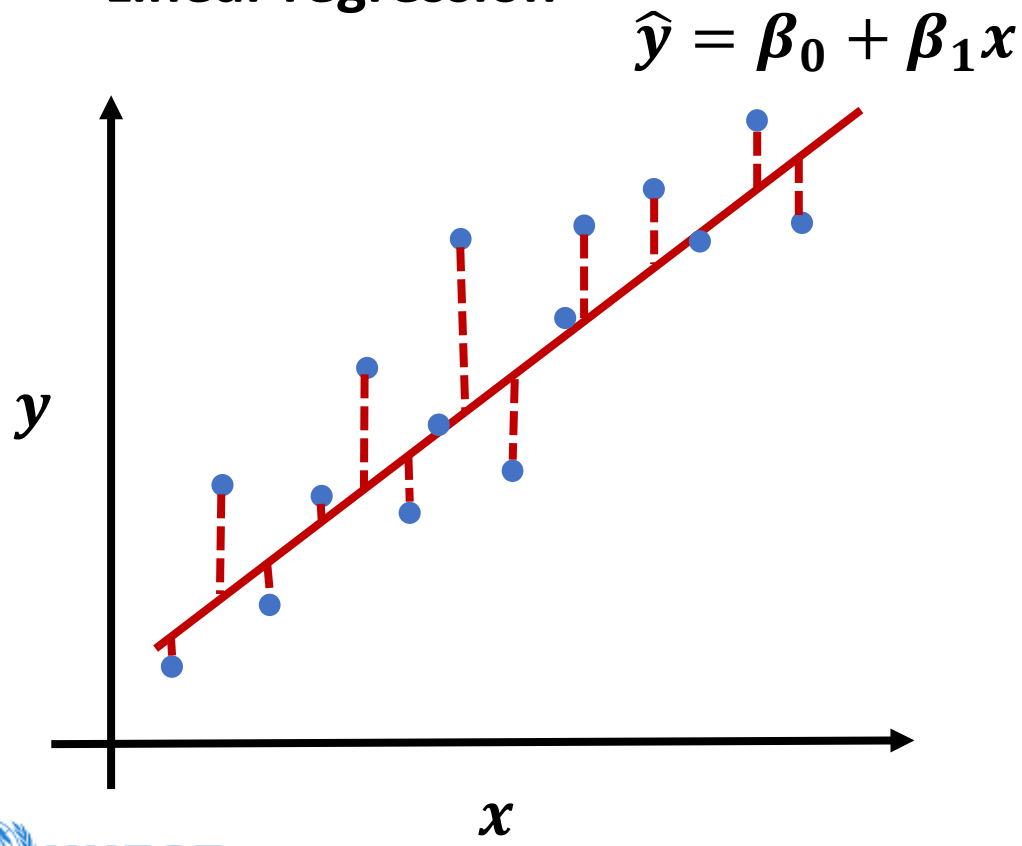
Linear Regression

Logistic Regression
Decision Tree

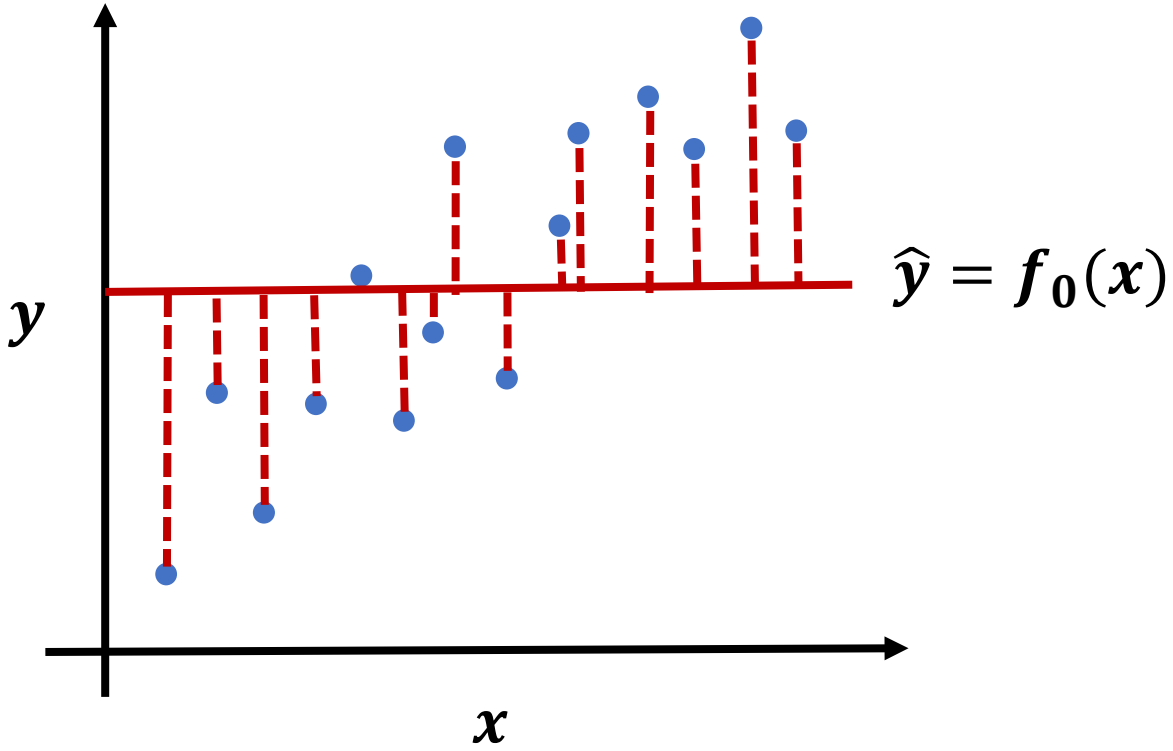
Support Vector Machine
Random Forest
Gradient Boosting
Neural Network
Deep Learning
…..

What is difference from “traditional” statistical method?

Linear regression

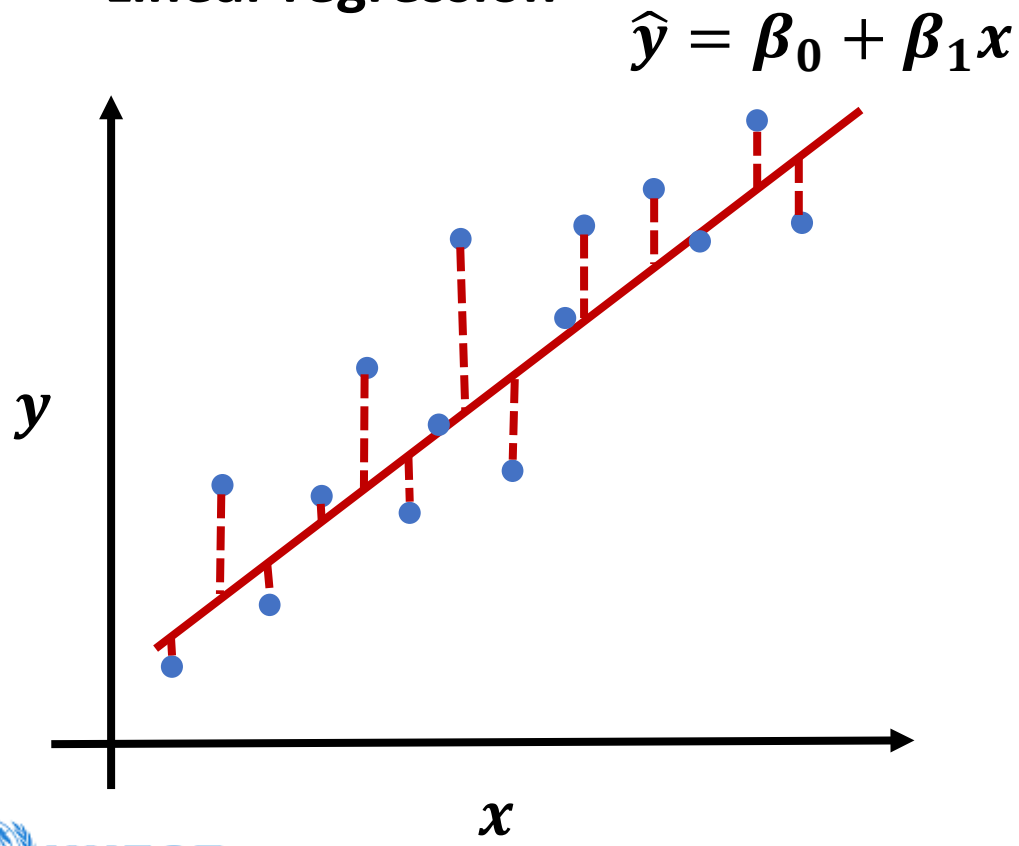


Gradient boosting

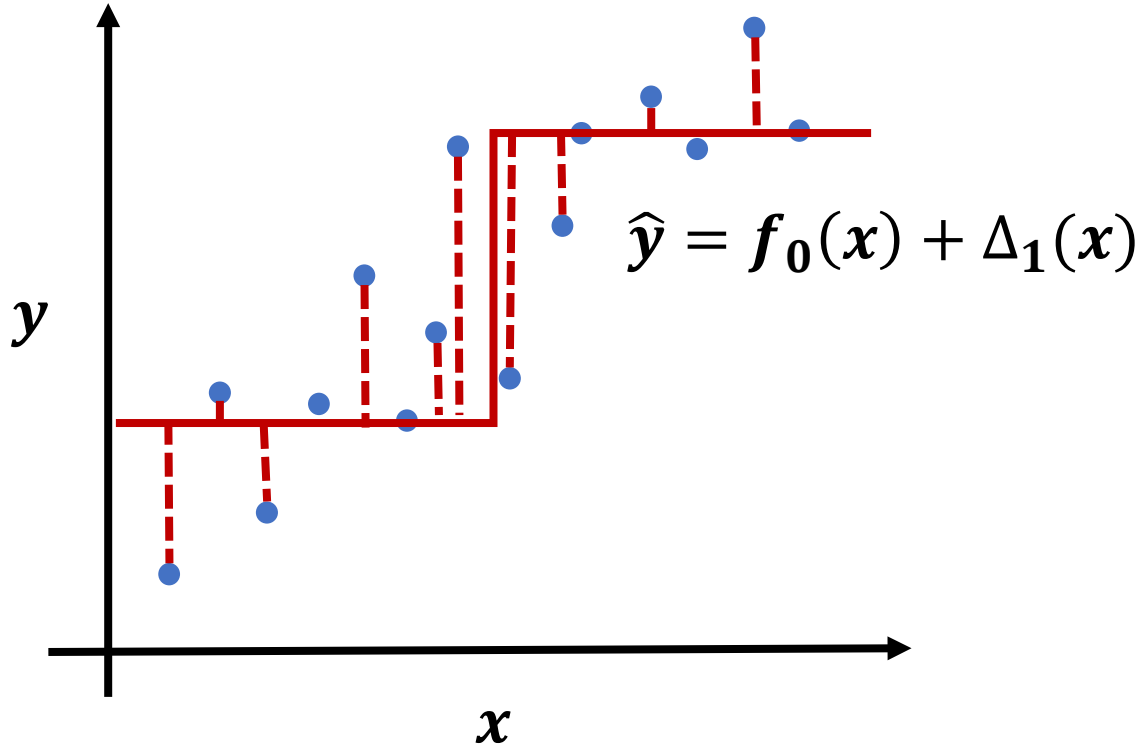


What is difference from “traditional” statistical method?

Linear regression



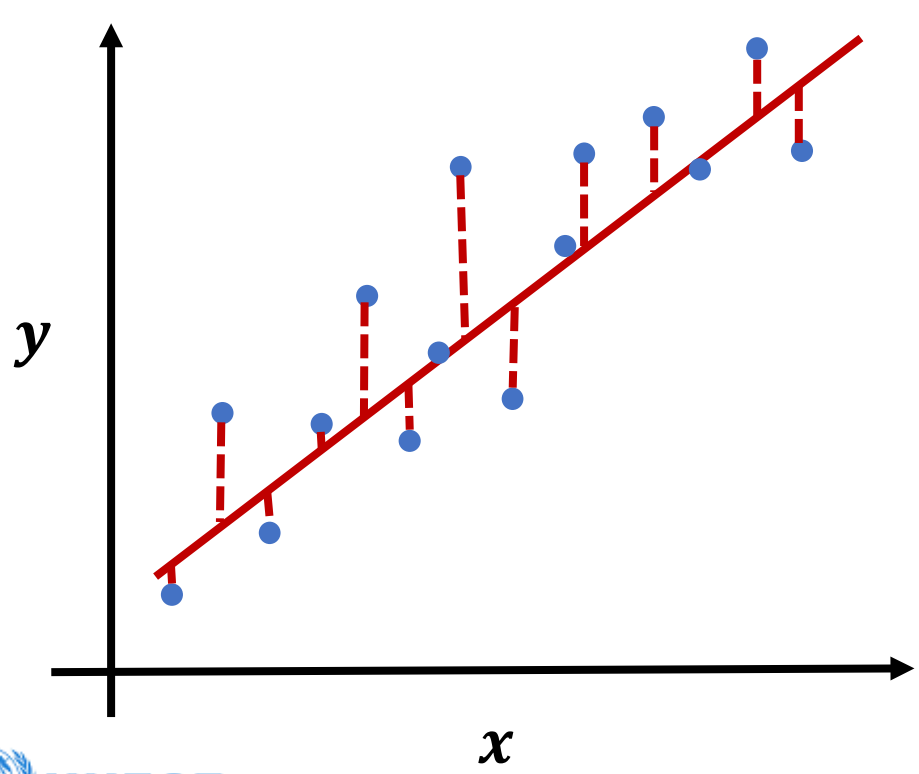
Gradient boosting



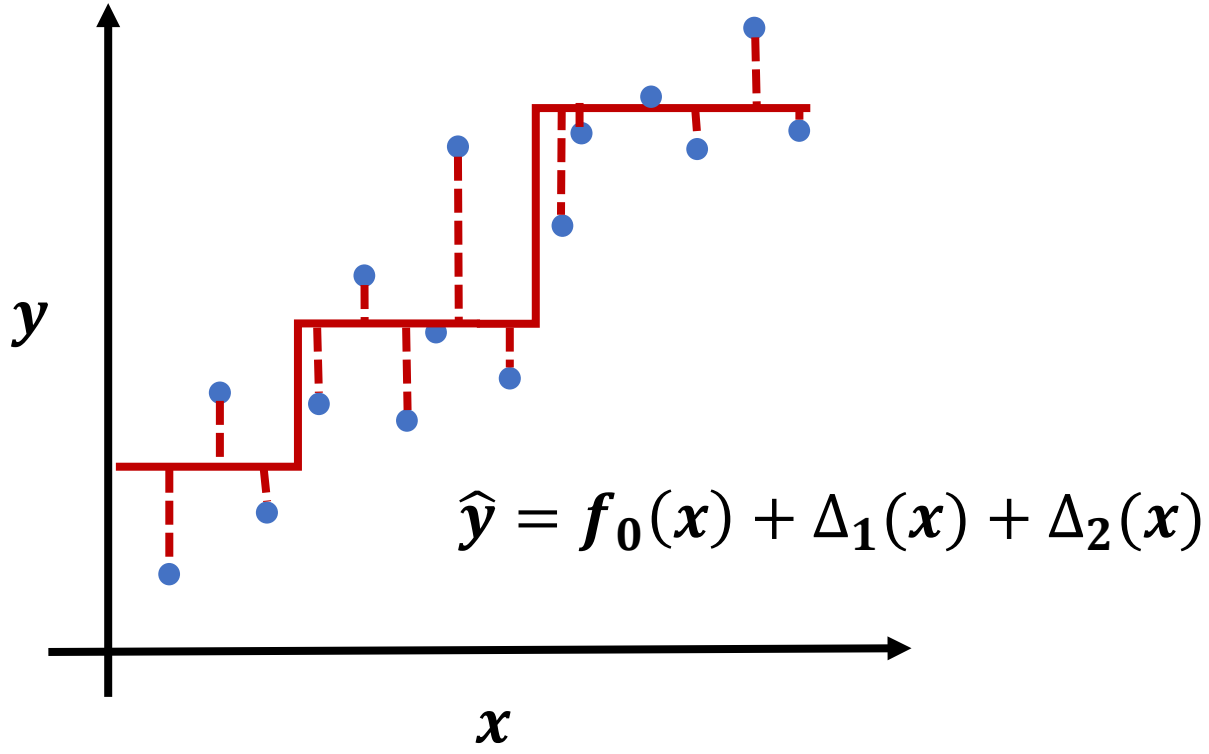
What is difference from “traditional” statistical method?

Linear regression

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



Gradient boosting

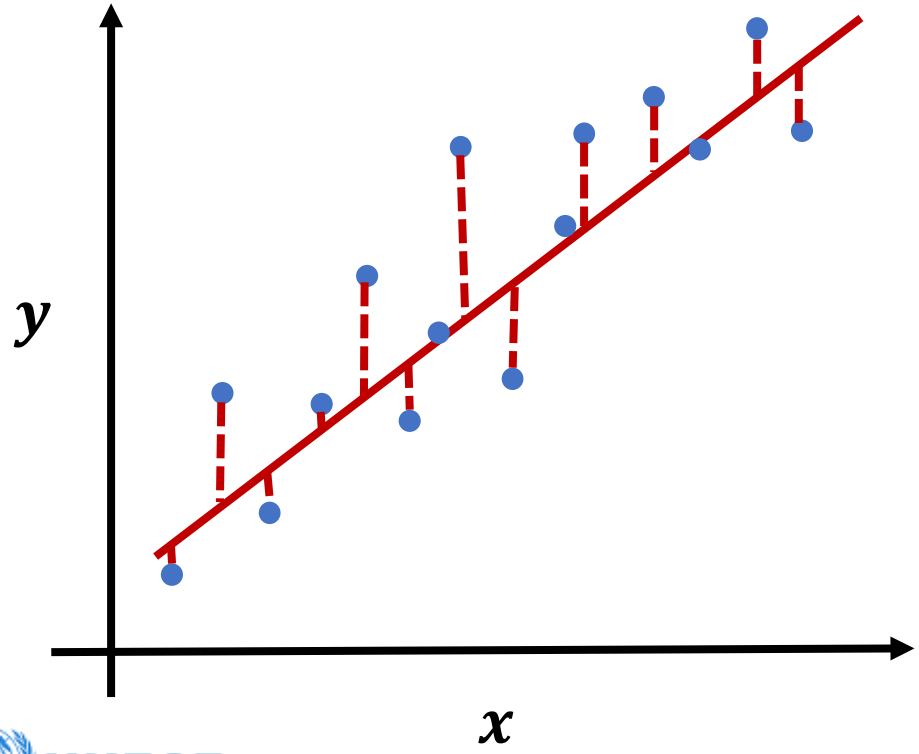


$$\hat{y} = f_0(x) + \Delta_1(x) + \Delta_2(x)$$

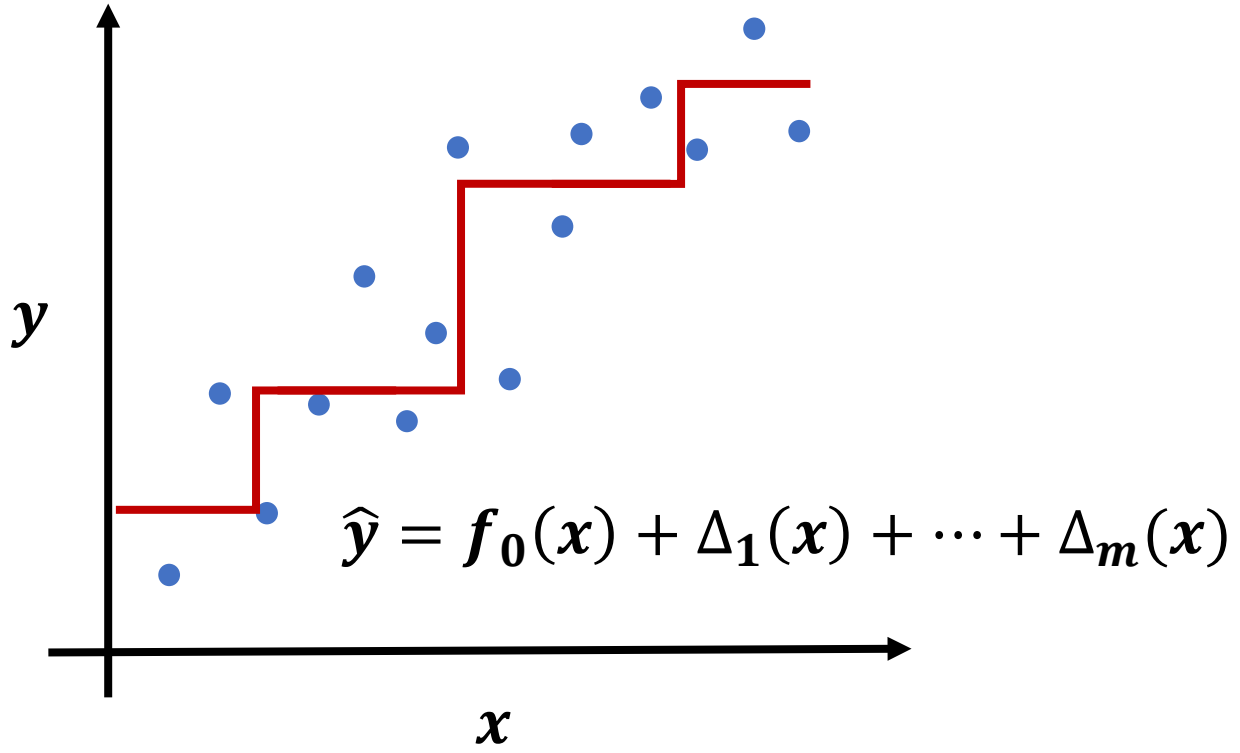
What is difference from “traditional” statistical method?

Linear regression

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



Gradient boosting



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

UNECE

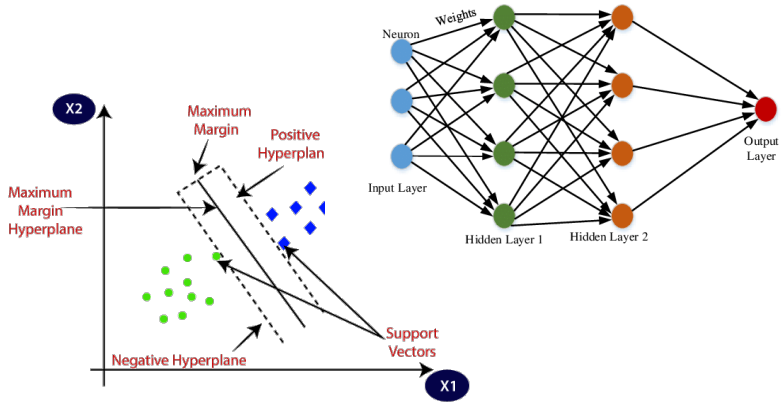
modernstats

2. Why Machine Learning?

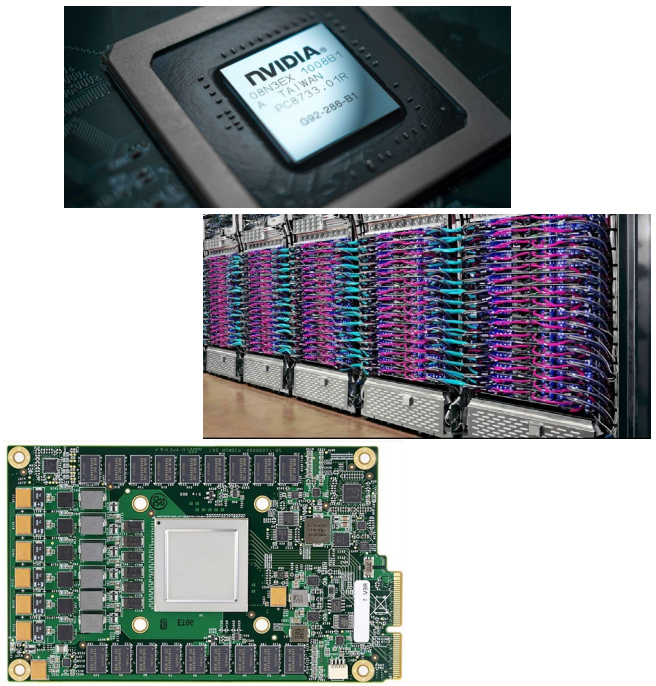
Machine Learning , why now?

Data

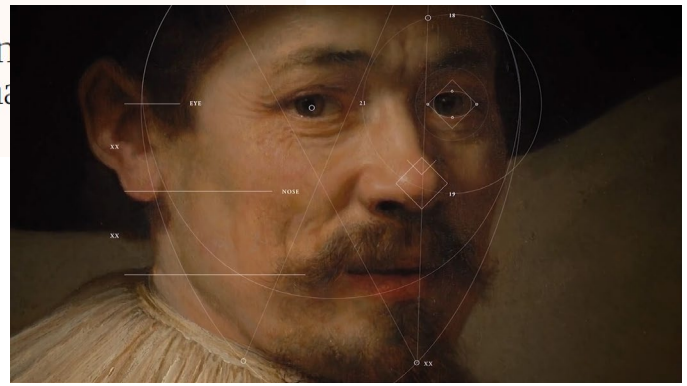
Methodology



Technology



Machine Learning , why now?



Next Rembrandt project

Google

amazon

facebook

twitter

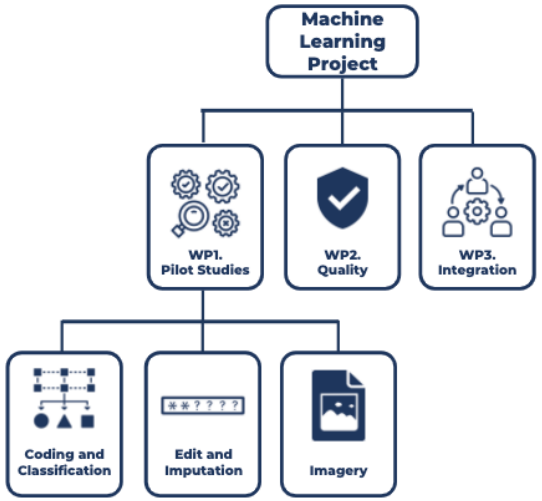
NETFLIX

UNECE

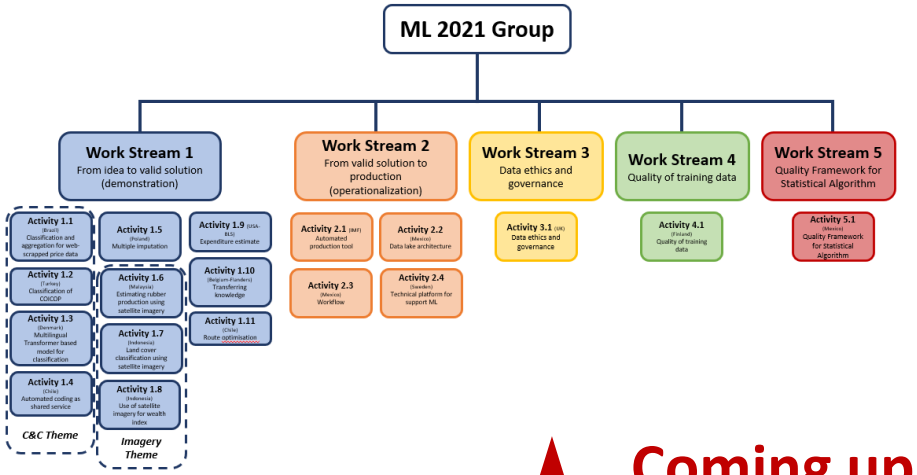
modernstats
by HLG - MOS

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

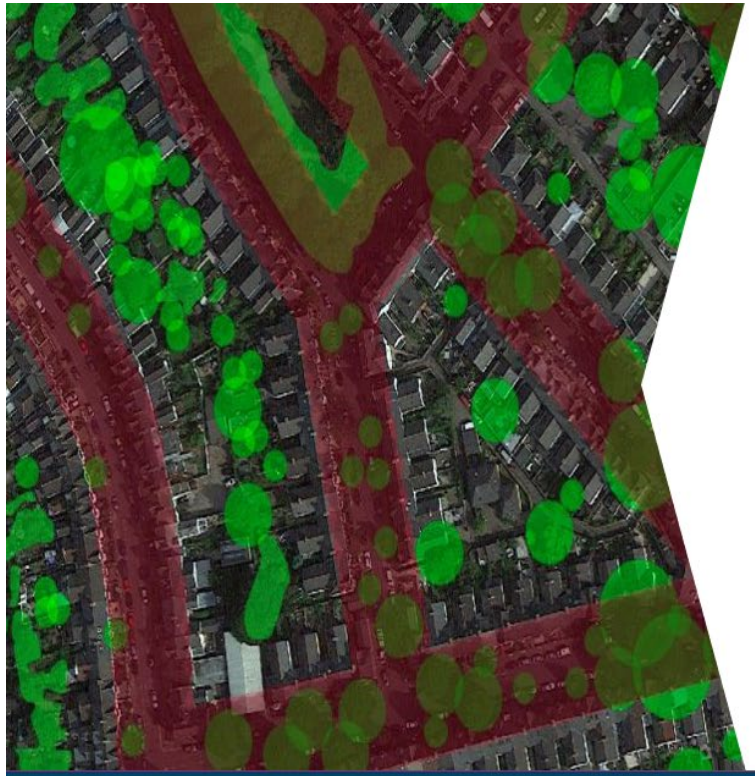
UK ONS-UNECE Machine Learning Group **2022**

UNECE

modernstats by HLG-MOS

3. How can machine learning help official statistics?

Machine learning for official statistics

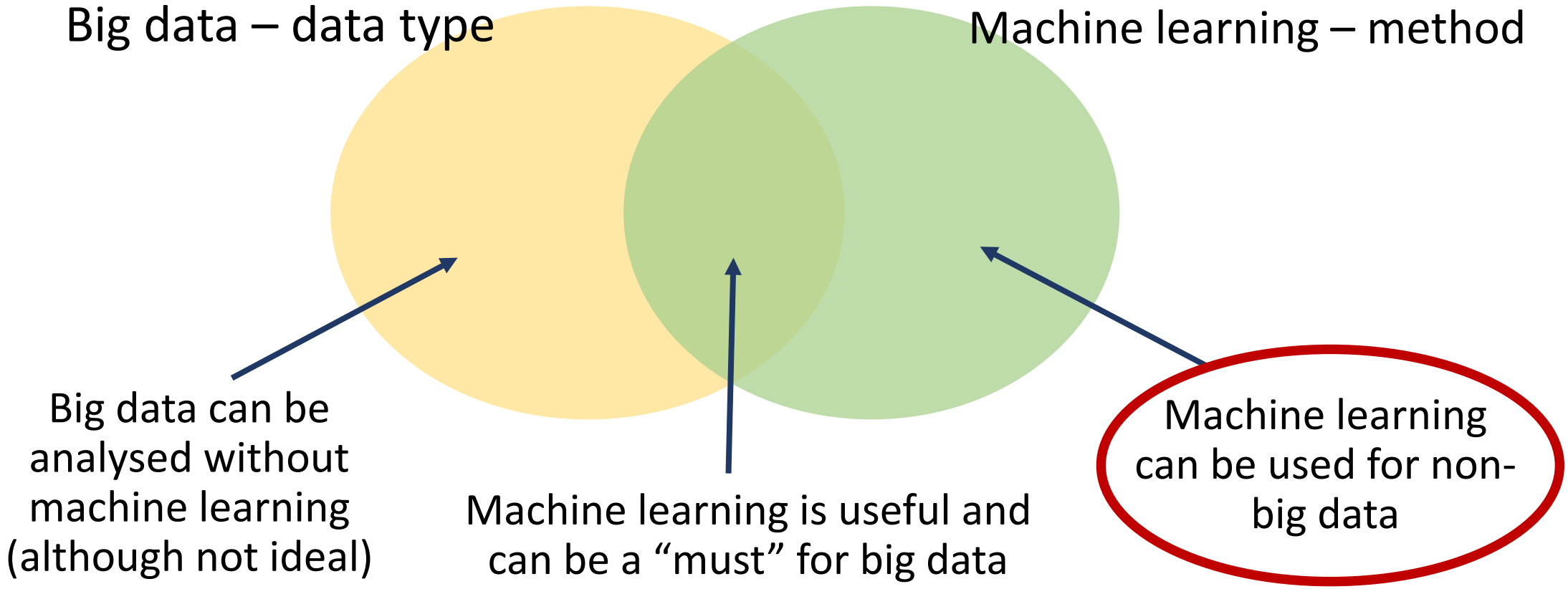


Why?

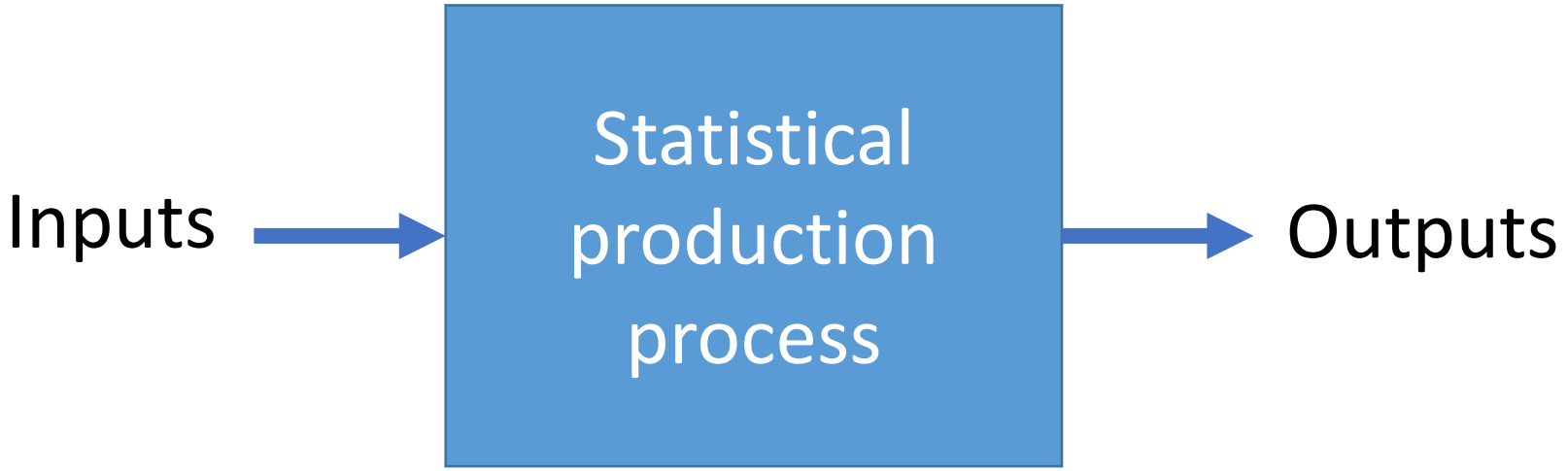
"Any industry with very large amounts of data — so much that humans can't possibly analyz[s]e or understand it on their own — can utilize AI"

Gartner 2017

Machine learning and big data

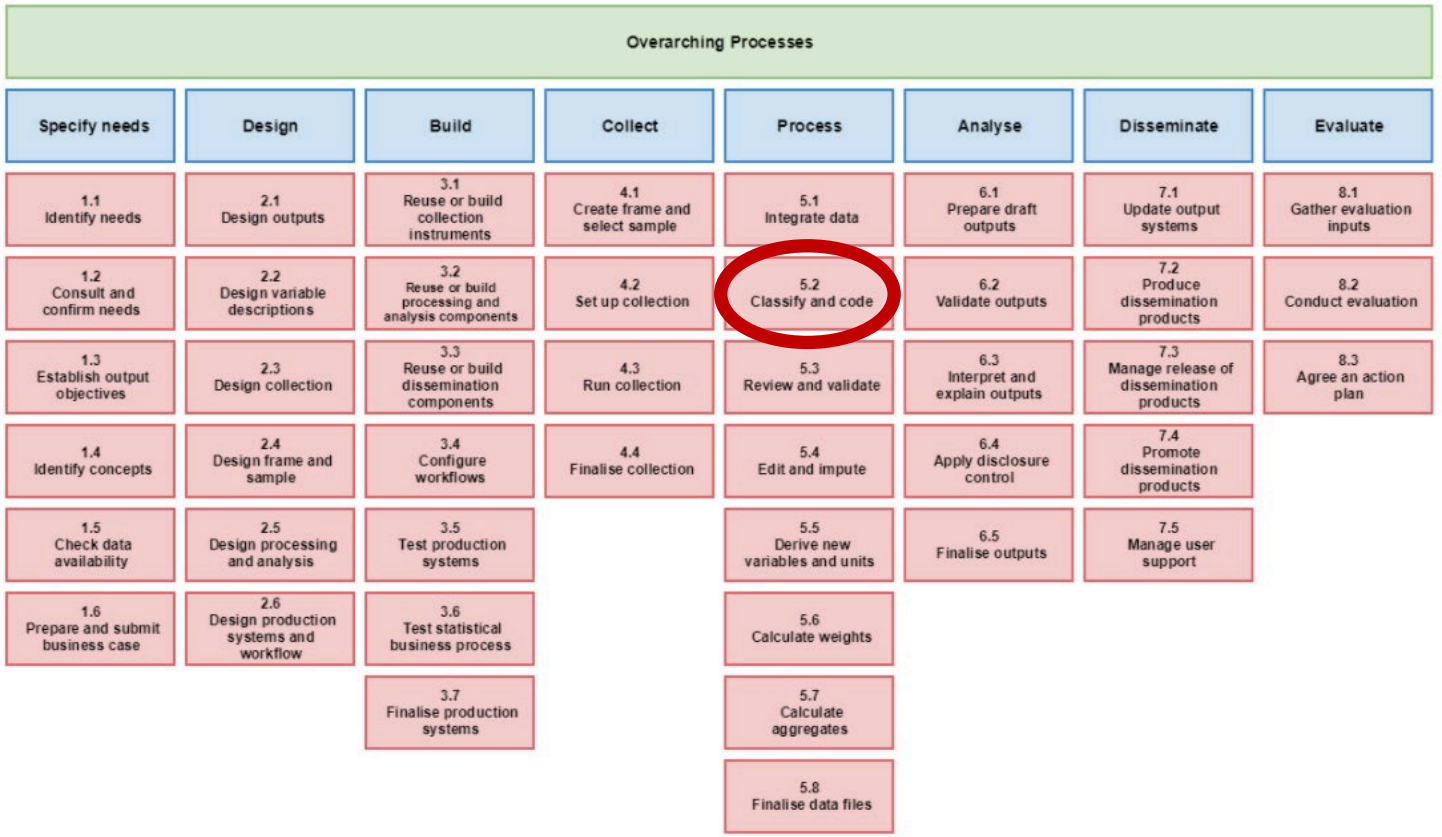


Statistical production process



Statistical production process

Inputs →

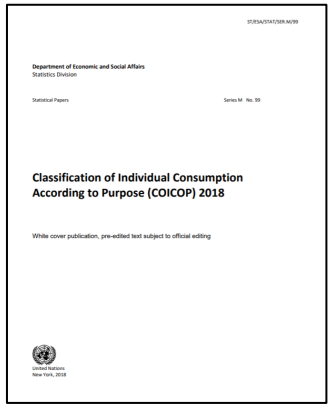


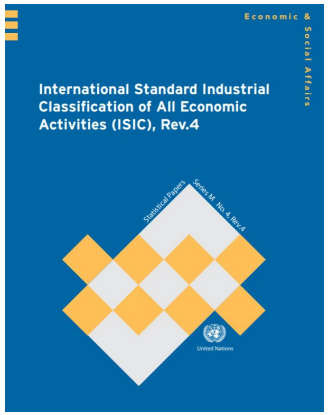
→ Outputs

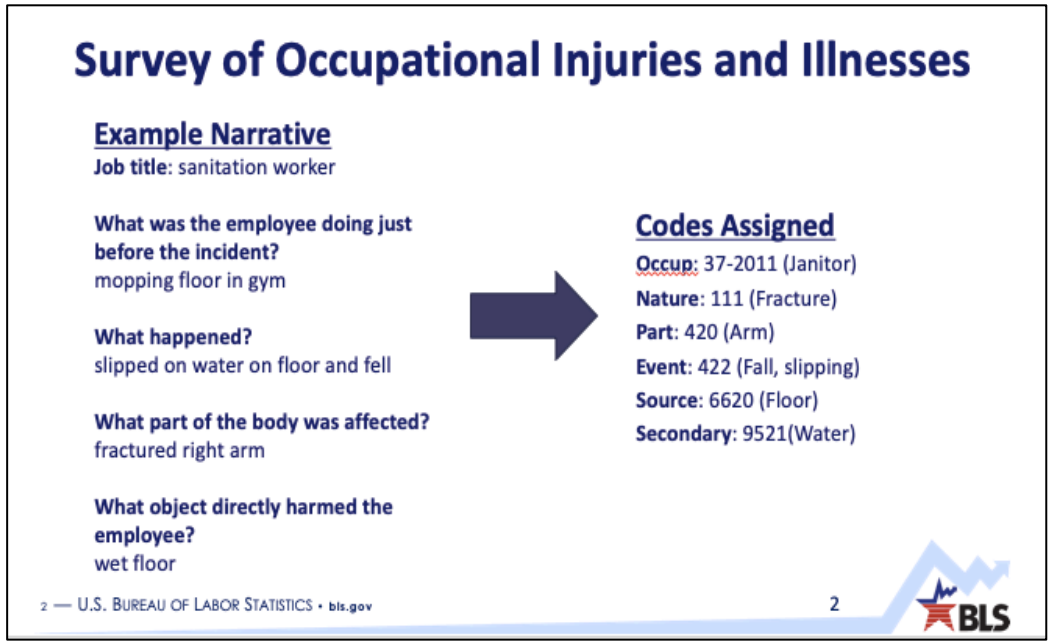
Generic Statistical Business Process Model (GSBPM)

Machine learning – application areas

5.2
Classify and code

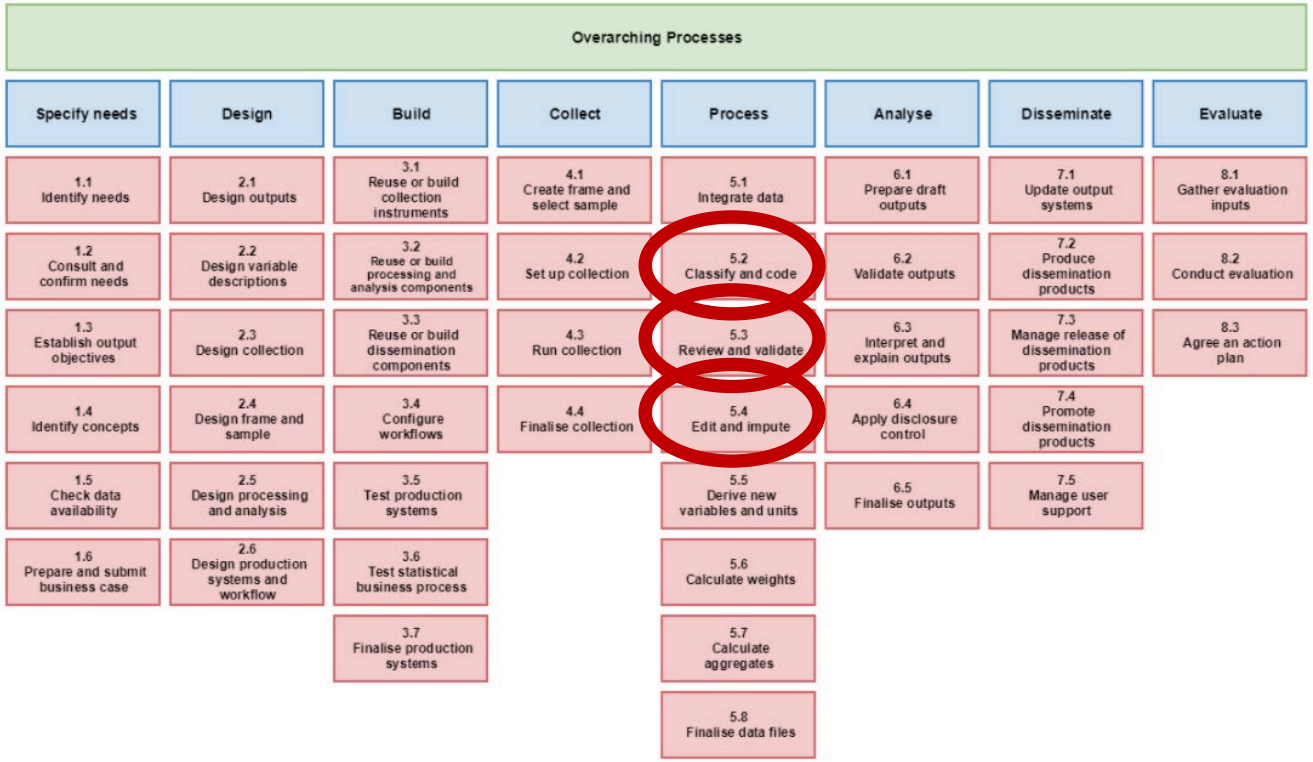






ML Project Coding Pilot Study from U.S.

Machine learning – application areas

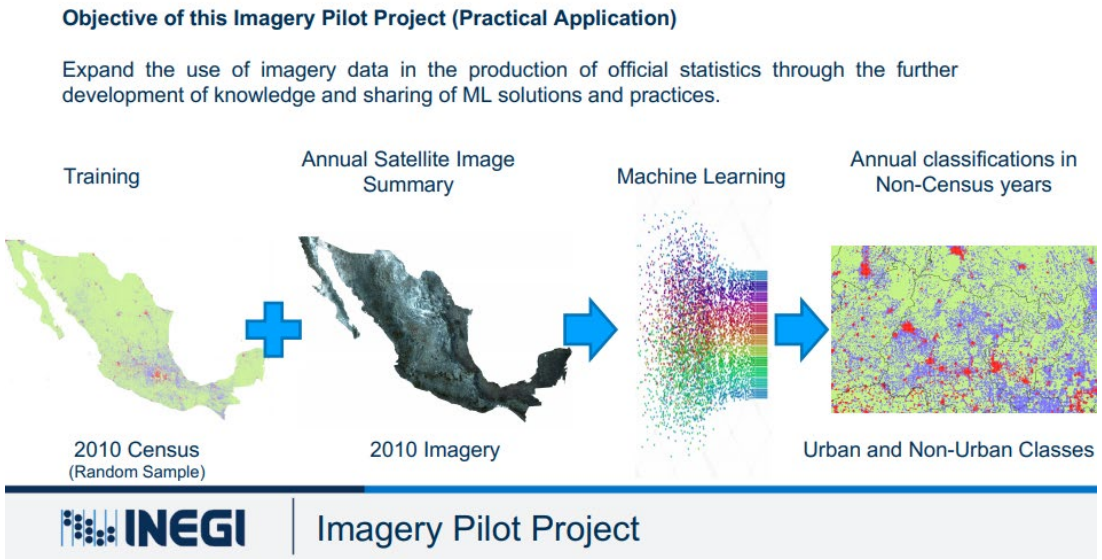


Areas with manual, repetitive works can be automated with help of machine learning

UNECE

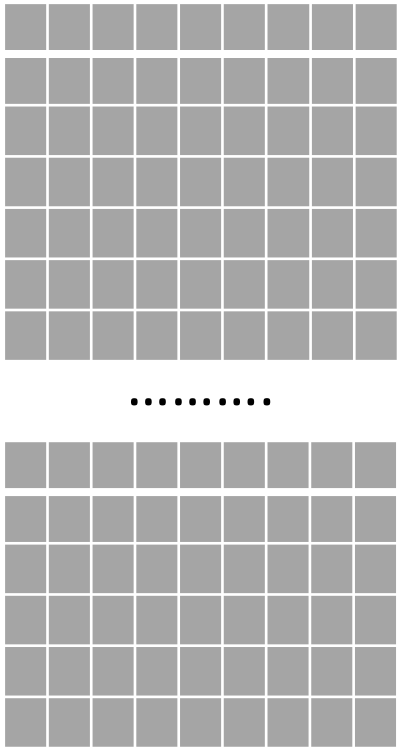
modernstats by HLG - MOS

Machine learning – application areas

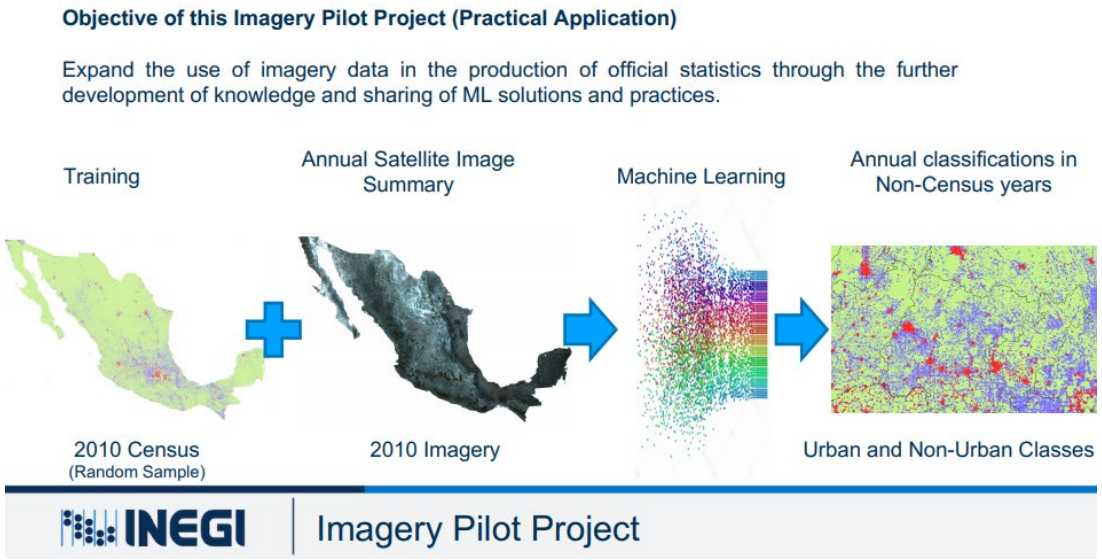


ML Project Imagery Pilot Study from Mexico

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

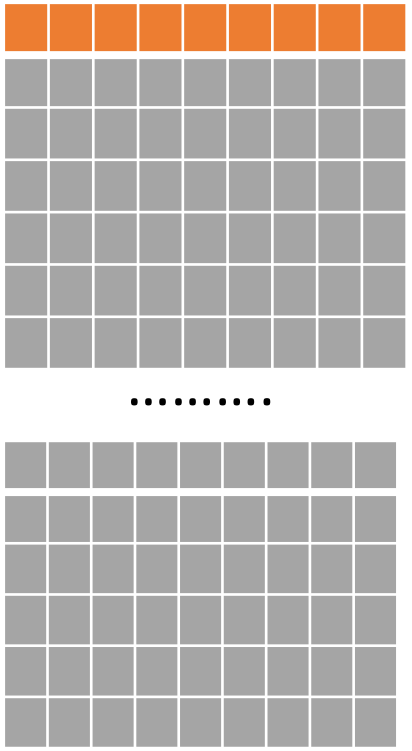


Machine learning – application areas



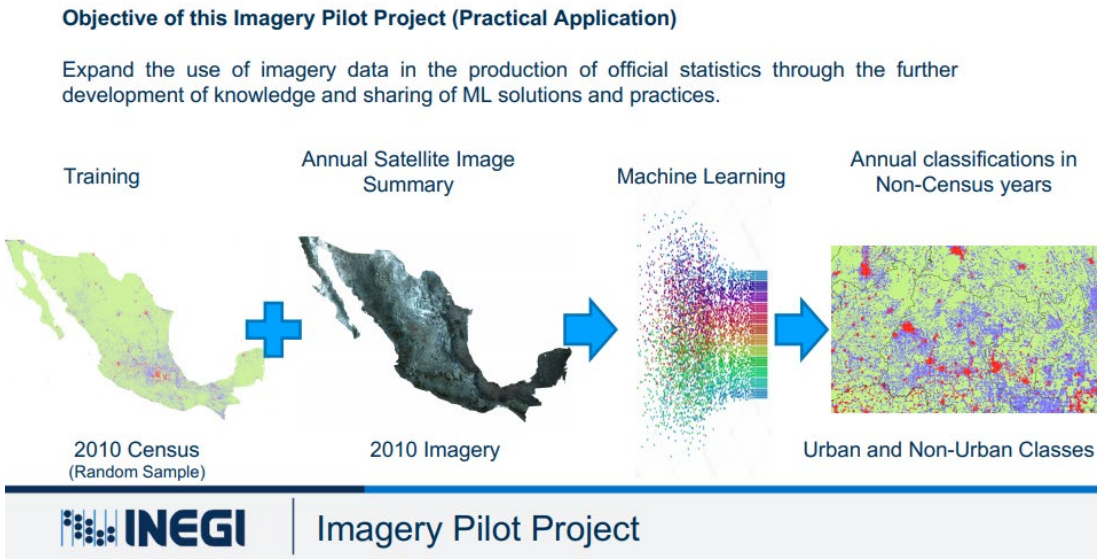
ML Project Imagery Pilot Study from Mexico

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



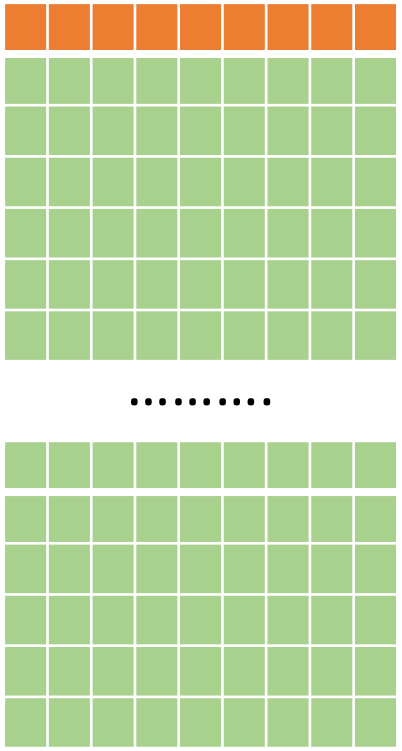
40,000 done by human

Machine learning – application areas



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			✸						○	

UN National Quality Assurance Framework

Quality Framework for Statistical Algorithms

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

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			✸						○	

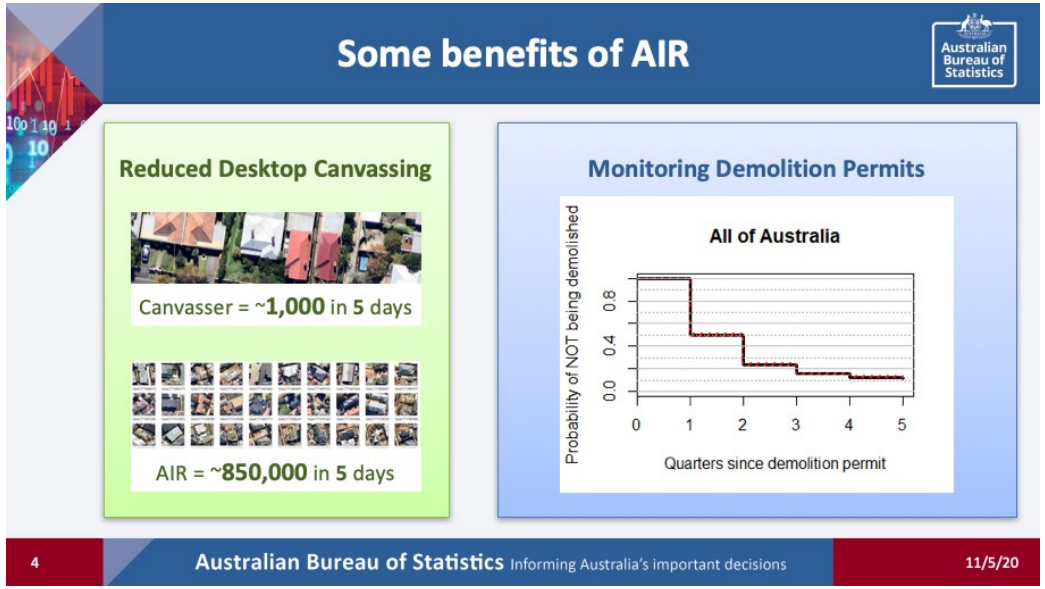
UN National Quality Assurance Framework

Quality Framework for Statistical Algorithms

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

Machine learning - quality implications

Timeliness



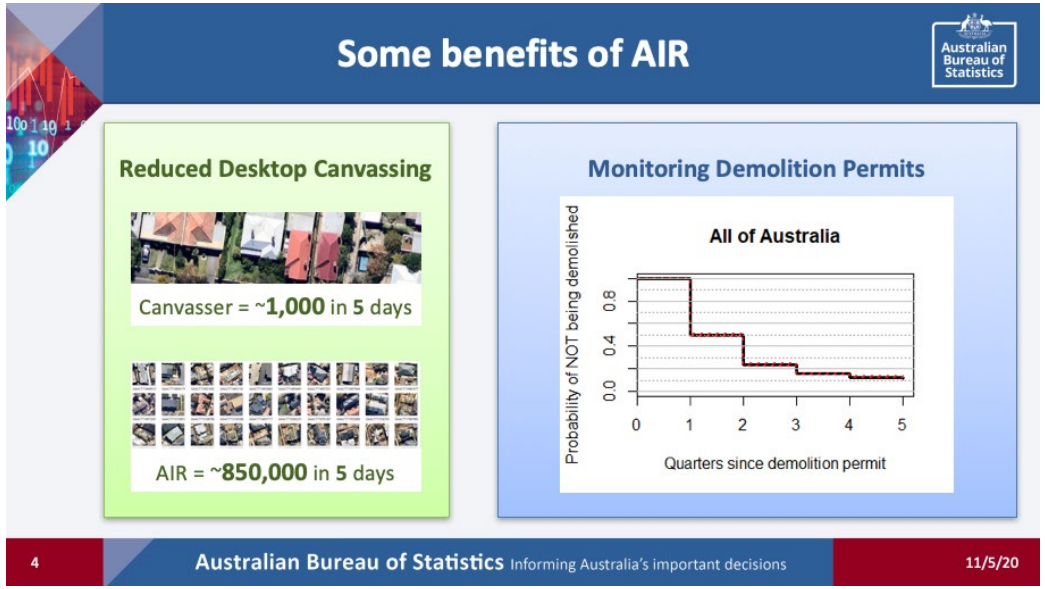
ML Project Imagery Pilot Study from Australia

Accuracy

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

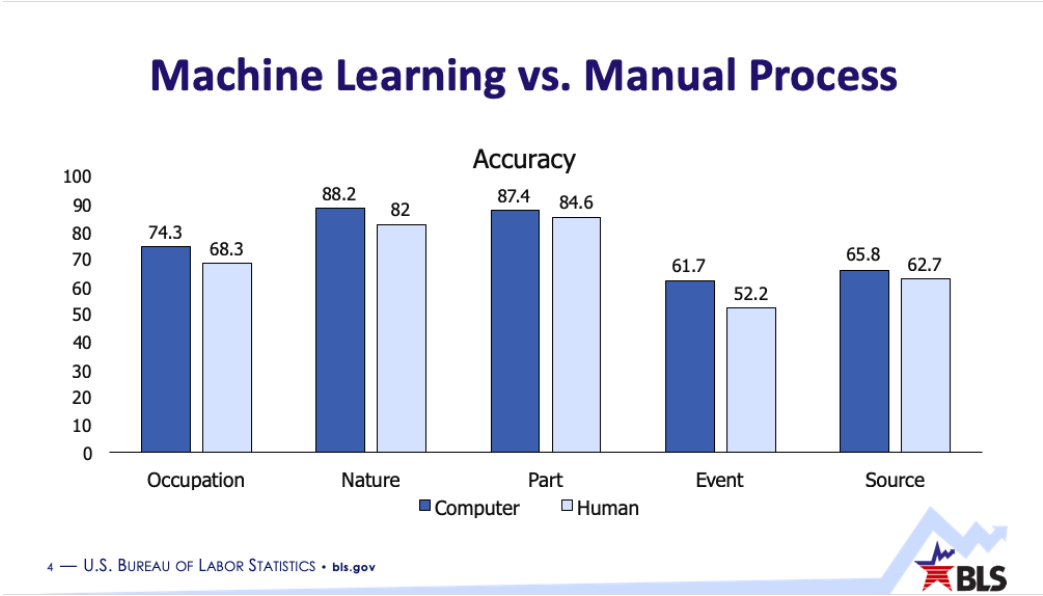
Machine learning - quality implications

Timeliness



ML Project Imagery Pilot Study from Australia

Accuracy



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

UNECE

modernstats

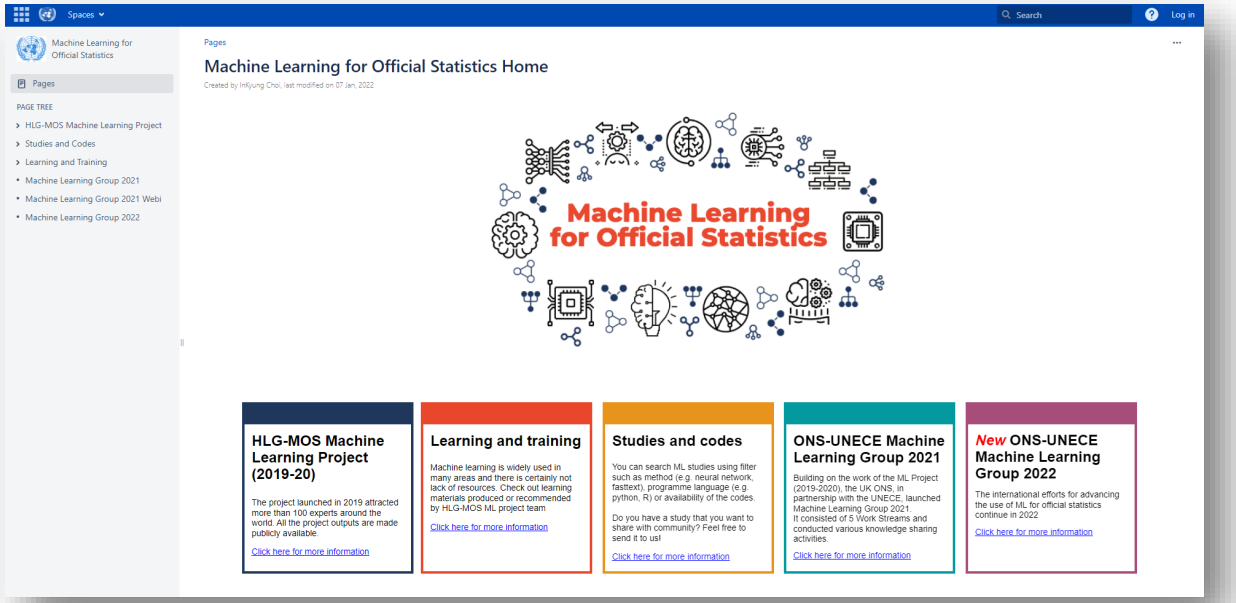
Machine learning for official statistics

Some final remarks

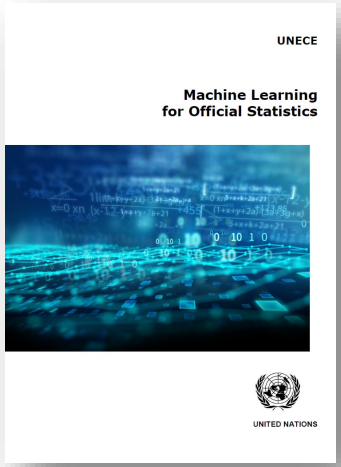
- ML can be used not only for big data but also for non-big data
- There are advanced methods, but simple methods work well too
- Depending on use case, different emphasis on different quality dimensions
- Sharing and collaboration is key to facilitating ML

Thank you for your attention

Resources



Machine Learning for Official Statistics Wiki: all reports, pilot studies, codes, learning resources, etc.



UNECE publication on Machine Learning for Official Statistics

UNECE

modernstats by HLG-MOS