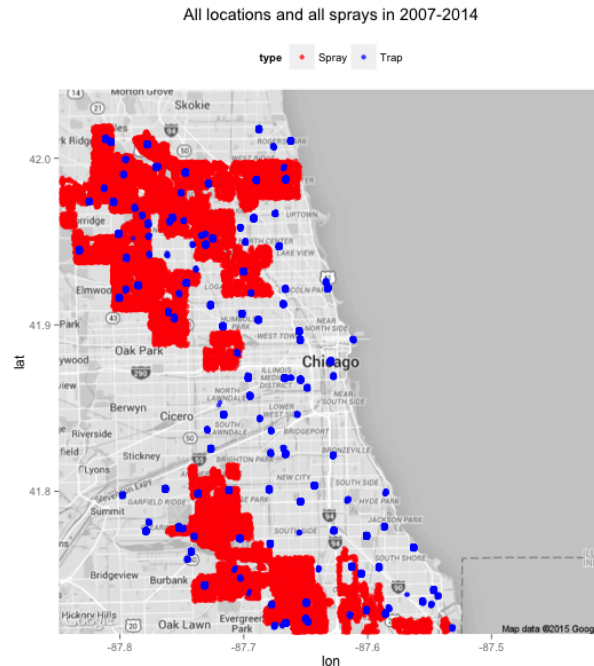


# Machine learning in outbreaks predictions (Łukasz Czekaj)

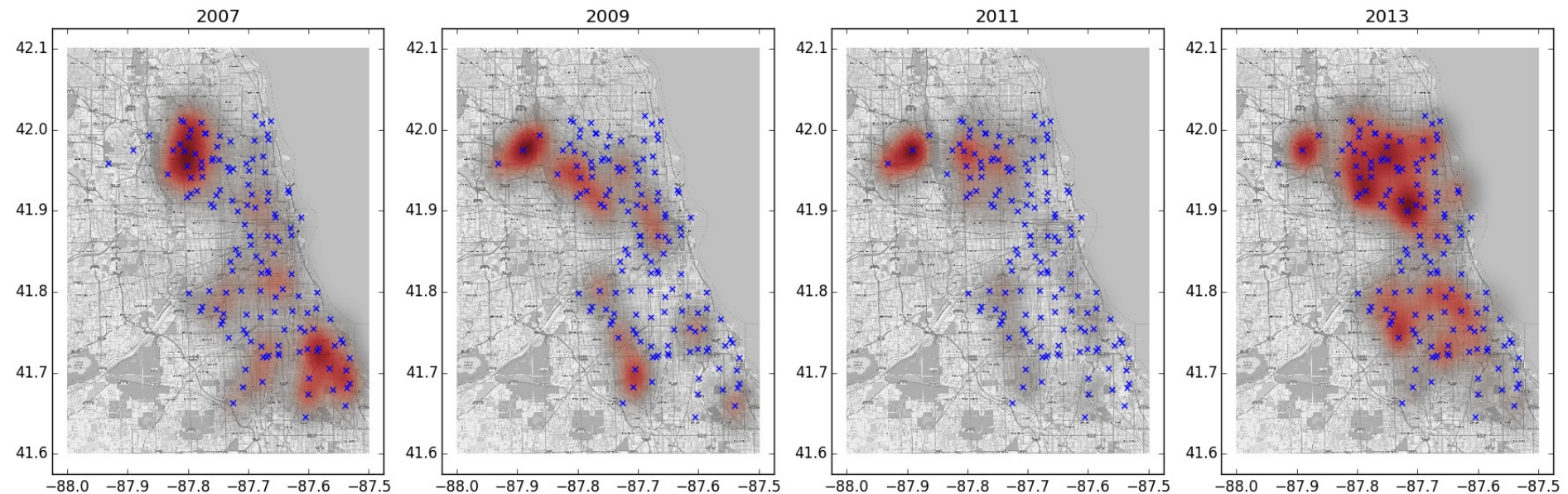
# West Nile Virus

- „Every week from late spring through the fall, mosquitoes in traps across the city are tested for the virus.”
- „Given weather, location, testing, and spraying data, this competition asks you to predict when and where different species of mosquitos will test positive for West Nile virus.”
- <https://www.kaggle.com/c/predict-west-nile-virus>



# West Nile Virus

- #mosquitoes, WNV test, species for trap, date
- Spray location and date
- Detailed weather information from 3 stations
- Predict WNV test result for given location nad date



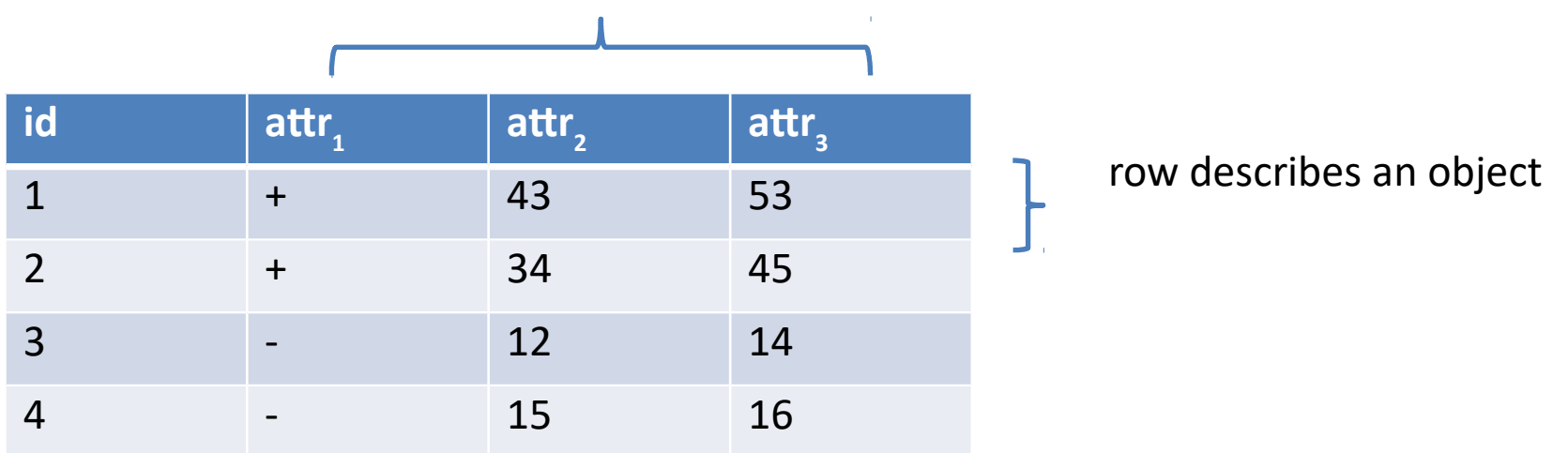
# Types of machine learning

- Unsupervised learning
- Supervised learning

# Unsupervised learning

- Data summarization, data reduction
- Discover latent structure (clustering)

descriptive attributes



id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>
1	+	43	53
2	+	34	45
3	-	12	14
4	-	15	16

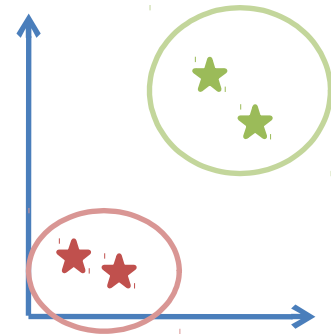
row describes an object

# Unsupervised learning (clustering)

id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>
1	+	43	53
2	+	34	45
3	-	12	14
4	-	15	16



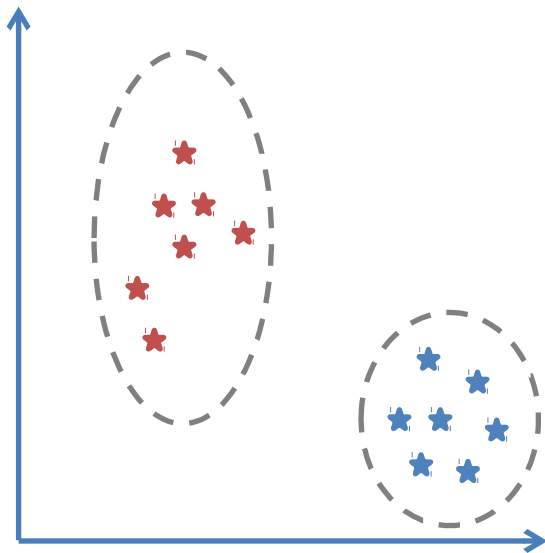
model



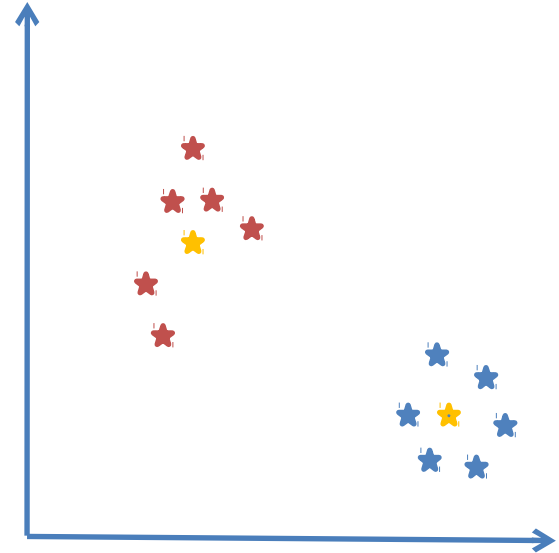
id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>	class
1	+	43	53	A
2	+	34	45	A
3	-	12	14	B
4	-	15	16	B

# Unsupervised learning (clustering)

- Automatic clustering



Mixture model



Centroids

# Supervised learning

- Predict unknown attribute on the base of other attributes

descriptive attributes

predicted attribute

id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>	attr <sub>4</sub>
1	+	43	53	A
2	+	34	45	A
3	-	12	14	B
4	-	15	16	?

training examples

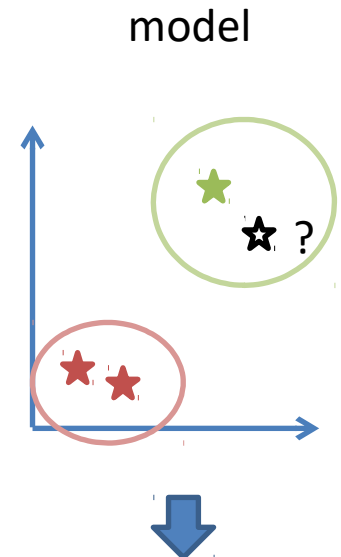
predicted example

$\text{attr}_1, \text{attr}_2, \text{attr}_3 \Rightarrow \text{attr}_4$



# Supervised learning

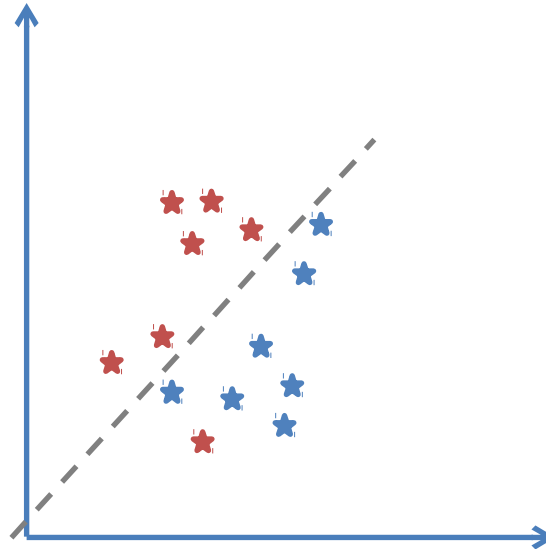
id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>	attr <sub>4</sub>
1	+	43	53	A
2	+	34	45	A
3	-	12	14	B
4	-	15	16	?



id	attr <sub>1</sub>	attr <sub>2</sub>	attr <sub>3</sub>	attr <sub>4</sub>
1	+	43	53	A
2	+	34	45	A
3	-	12	14	B
4	-	15	16	B

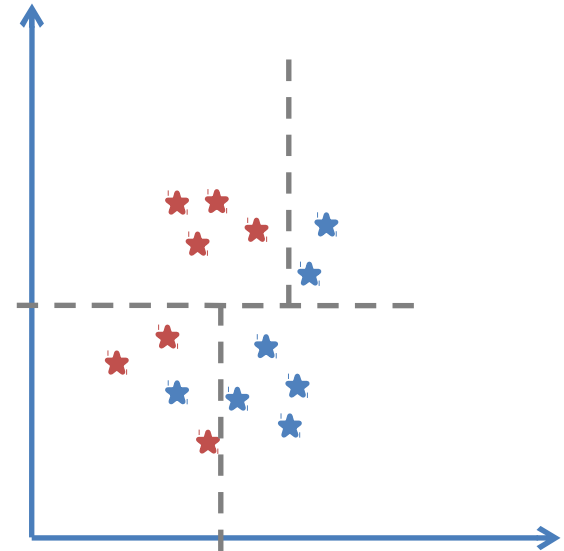
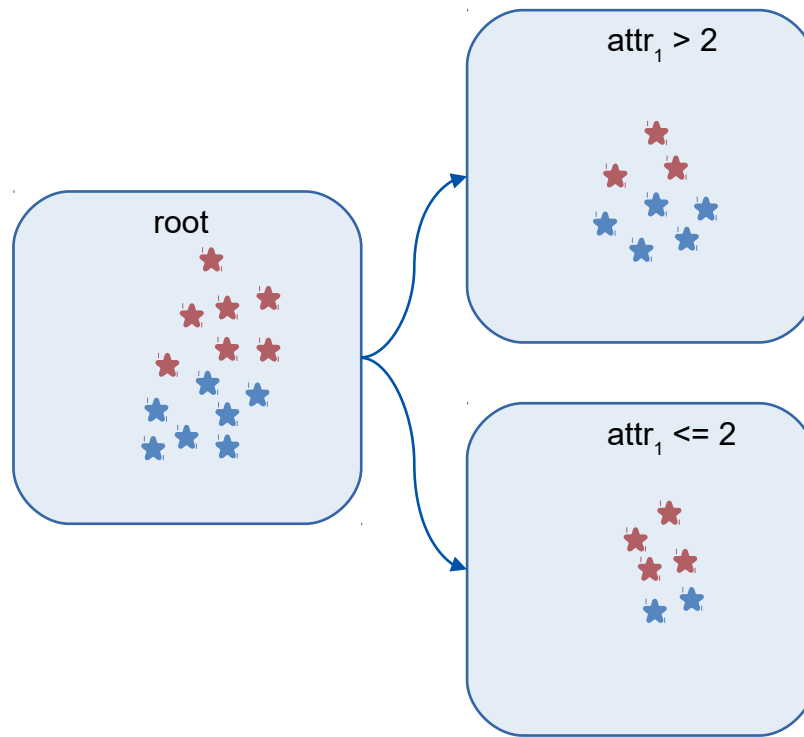
# Supervised learning

- Logistic regression (glm)



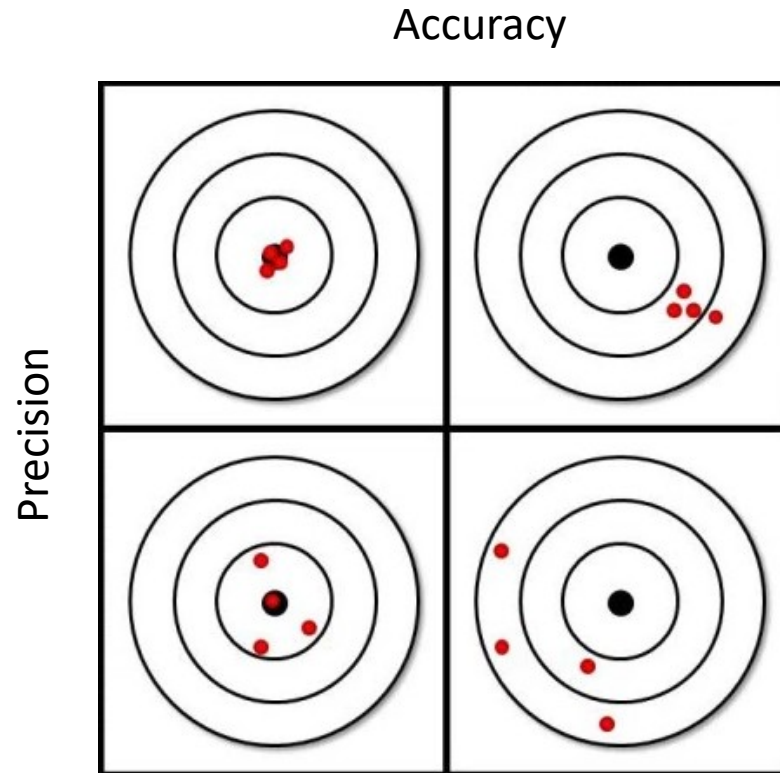
# Supervised learning

- Classification tree



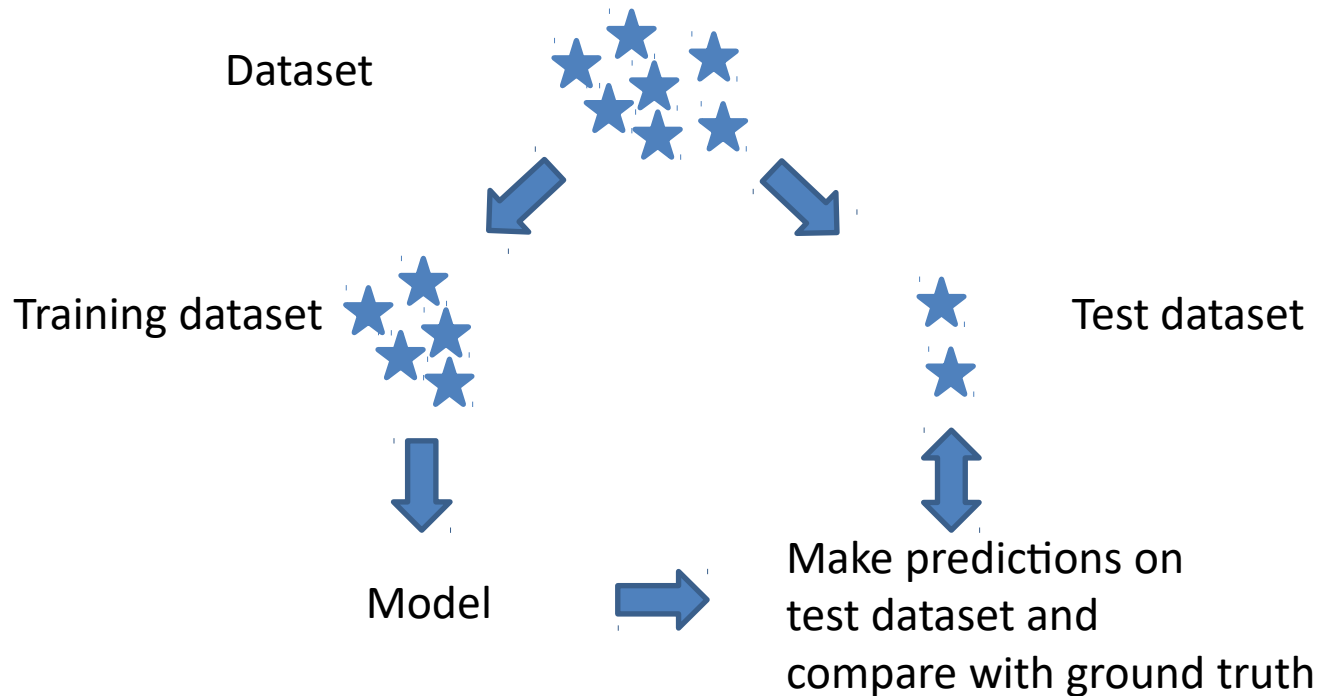
# Model quality

- Precision and accuracy



# Model quality

- Validation



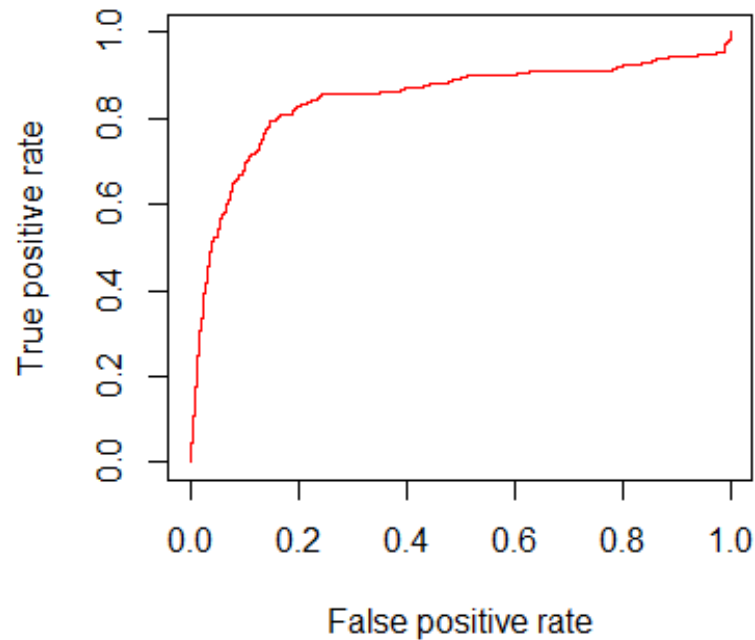
# Model quality

- Confusion matrix

Data\Predictions	Positive	Negative
Positive	True positive (sensitivity)	False negative
Negative	False positive (false alarm)	True negative (specificity)

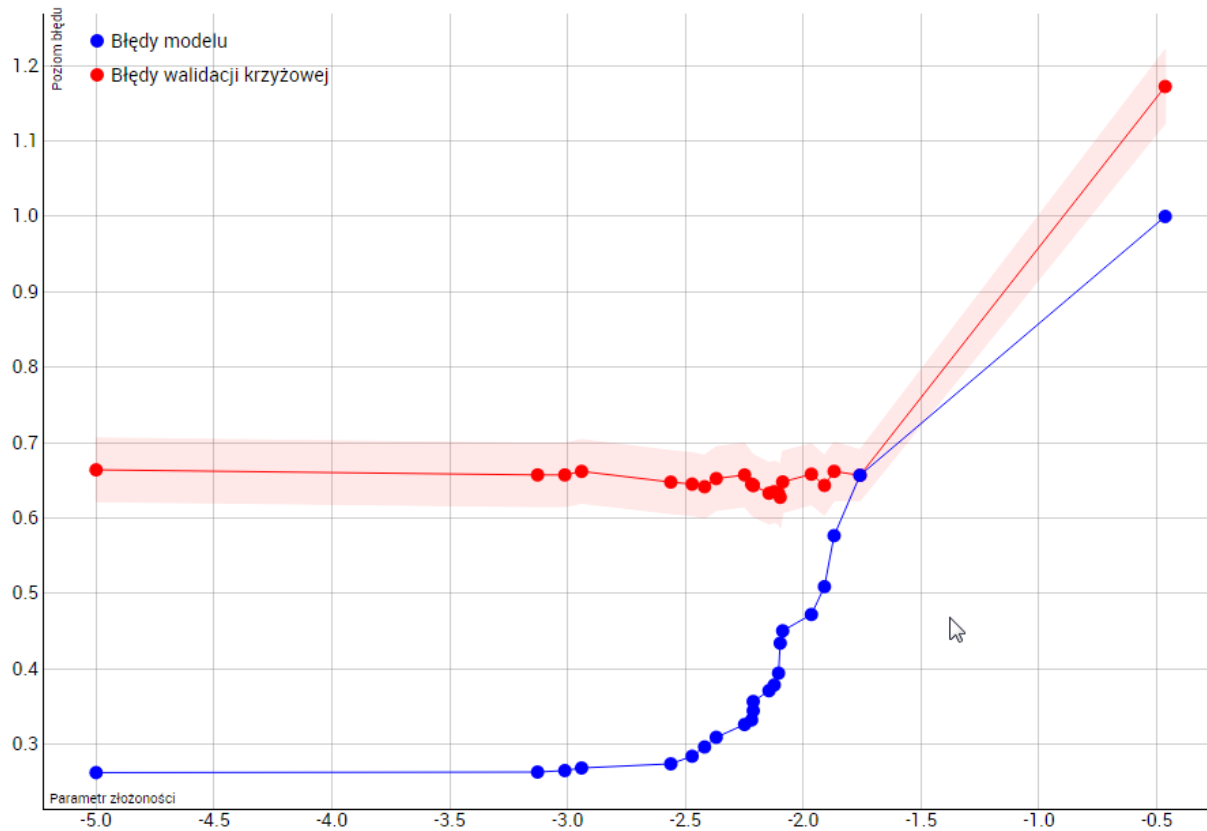
# Model quality

- Receiver Operating Characteristic (ROC) curve



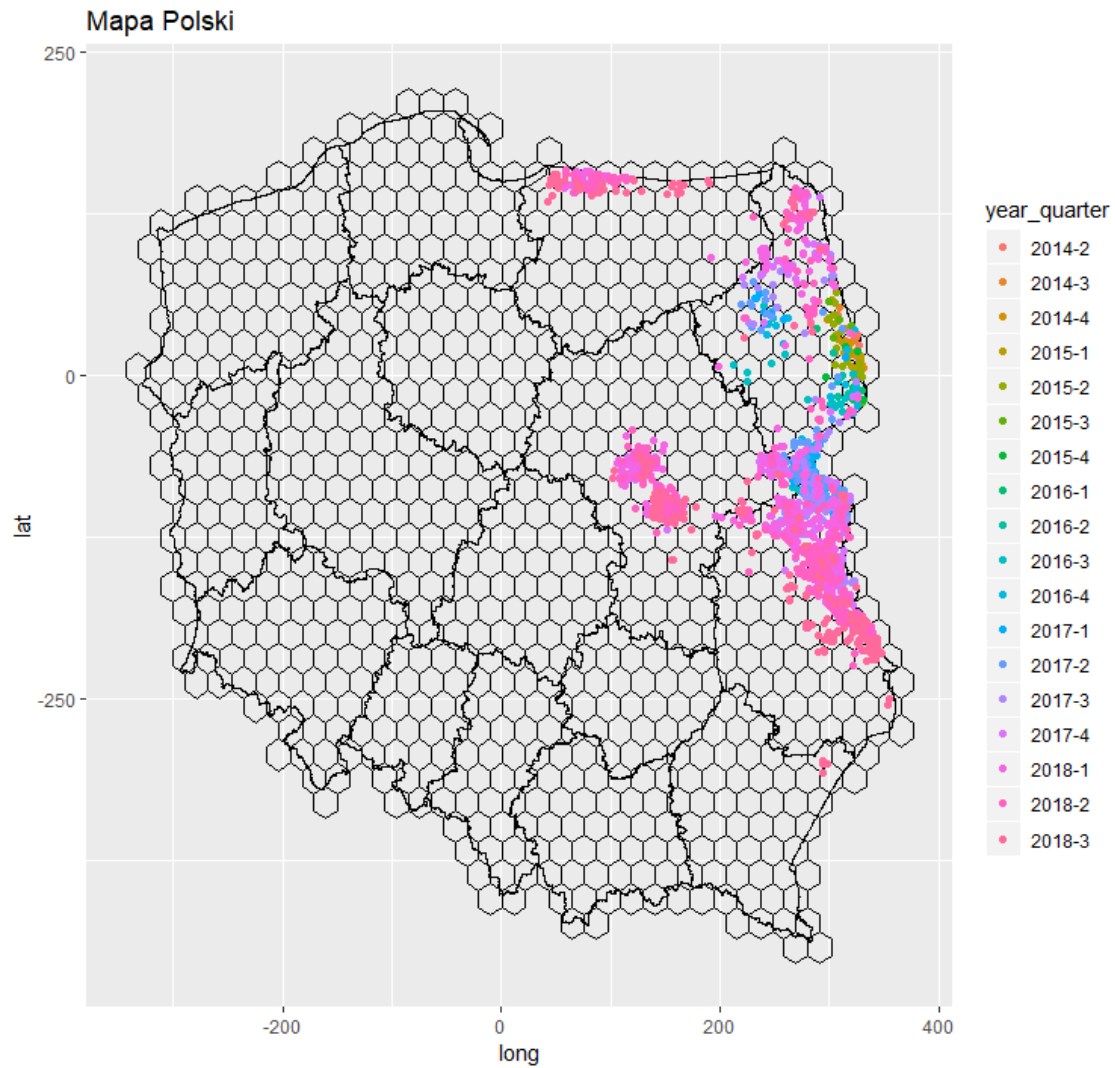
# Model quality

- Overfit



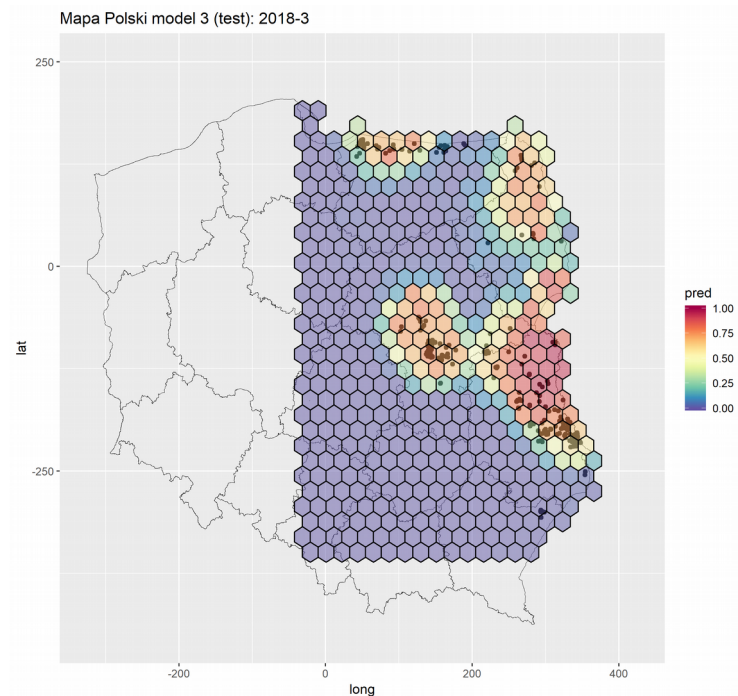
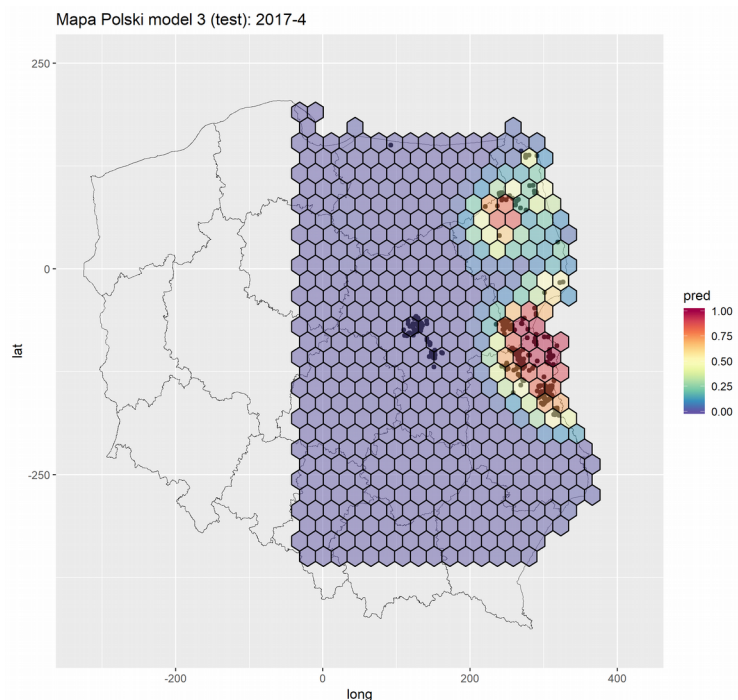


# ASF in Poland

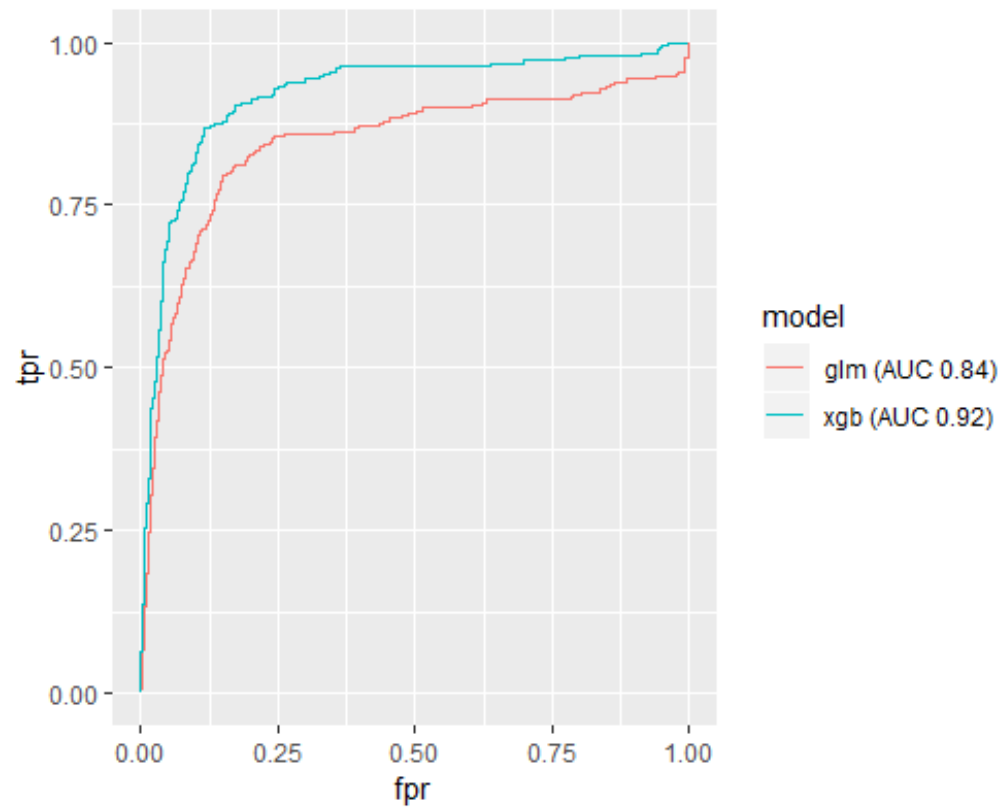


# ASF in Poland

- Training Data: 2015 Q2 – 2017 Q3 (10 quarters)
- Training Data: 2017 Q4 – 2018 Q3 (4 quarters)
- Predicted event: ASF positive dead pig in hex
- Descriptive attributes: hex history, hex neighbourhood history, human, forest and pigs density, season



# ASF in Poland



# Modeling approach comparison

	Machine Learning	Phenomenological
Model	Blackbox, nonparametric, easy to add new attributes with complicated interactions.	Fixed parametric equation inspired by „mechanics” of the process.
Goal	Provide accurate predictions	Understanding „mechanics”, select from competitive models
Quality	Validation	AIC, BIC, DIC, Bayes Factor

