Introduction to machine learning in Hydrology

Lazaro J. Perez & Marc Berghouse

contact: lazaro.perez@dri.edu

Outline

- Global picture
- Statistics
- Machine learning in Hydrology

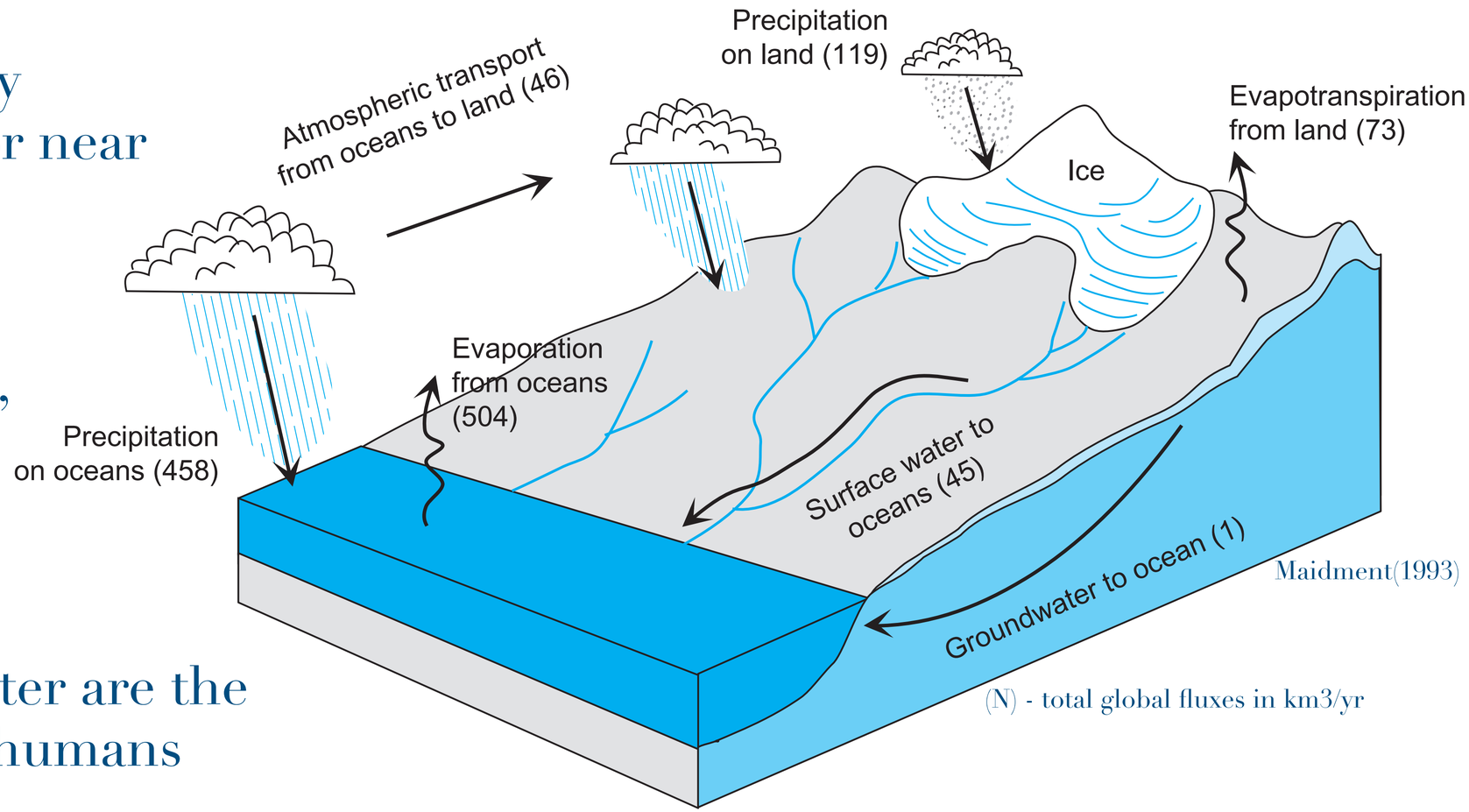
The big picture

Global hydrologic cycle

Water exists in virtually every accessible environment on or near the earth's surface

Of the fresh water reservoirs, glacial ice and groundwater are by far the largest

Groundwater and surface water are the two reservoirs most used by humans because of their accessibility



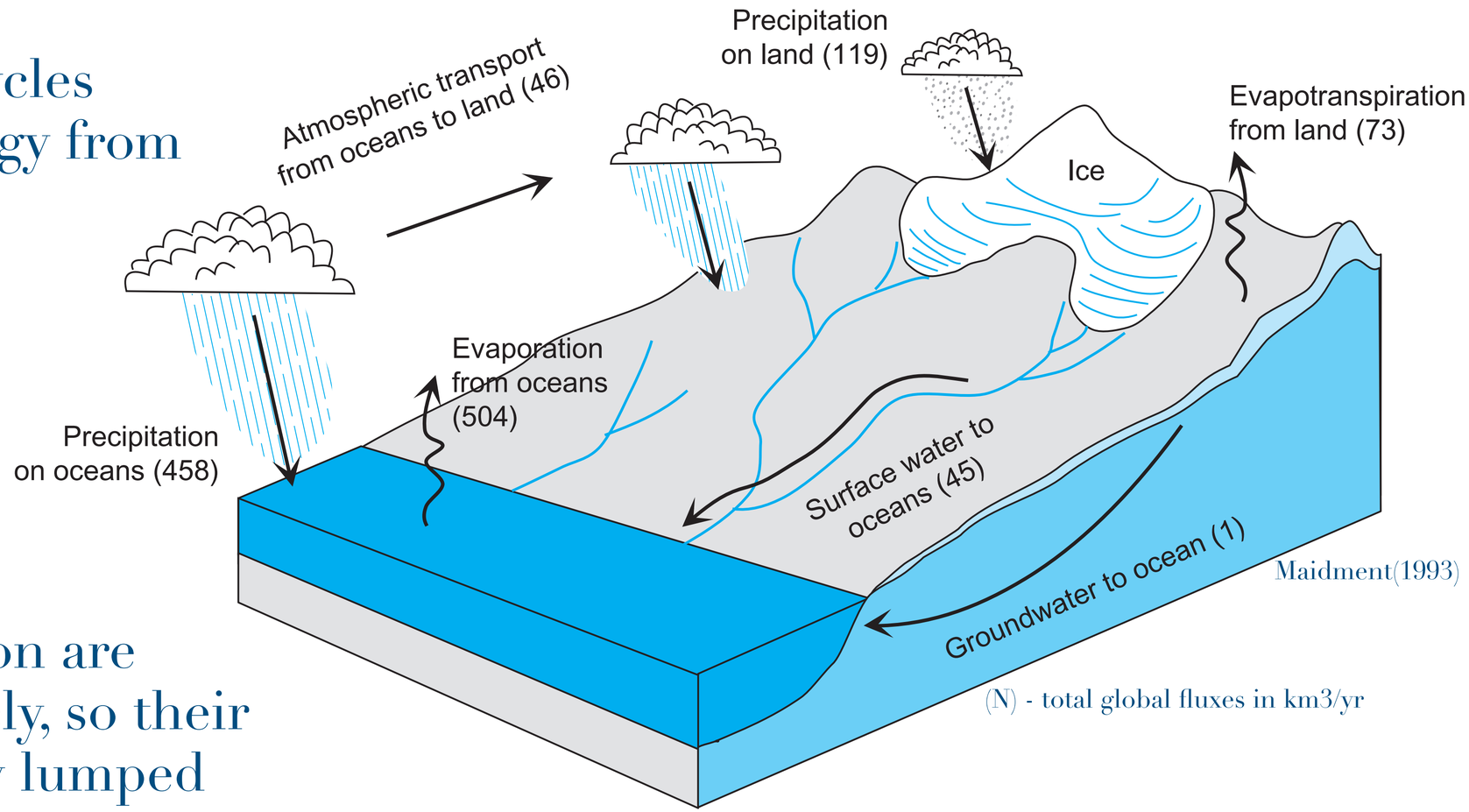
The big picture

Global hydrologic cycle

Water changes phase and cycles continuously fueled by energy from solar radiation

Solar energy drives evaporation, transpiration, atmospheric circulation, and precipitation

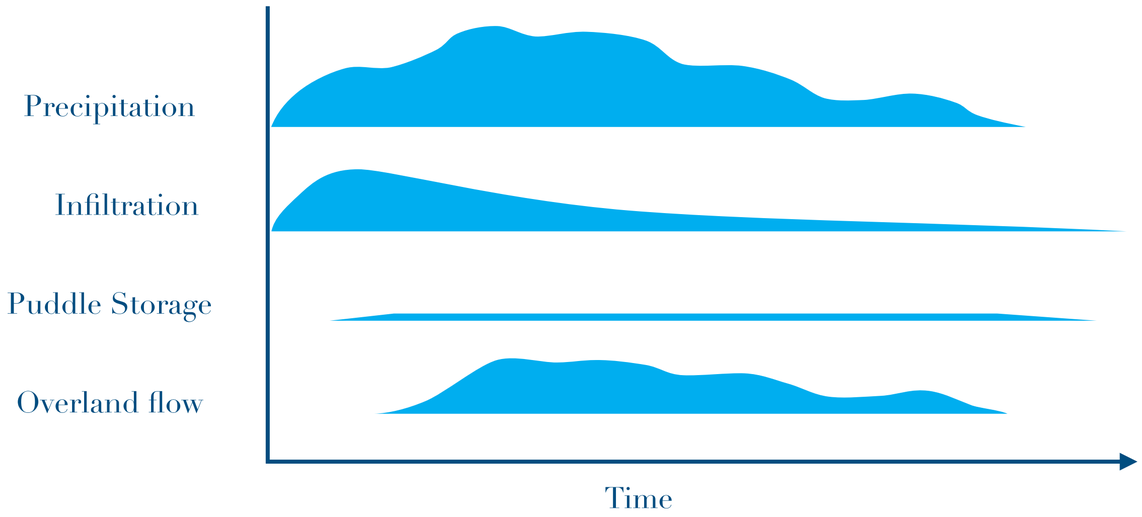
Evaporation and transpiration are difficult to measure separately, so their combined effects are usually lumped together and called **evapotranspiration**

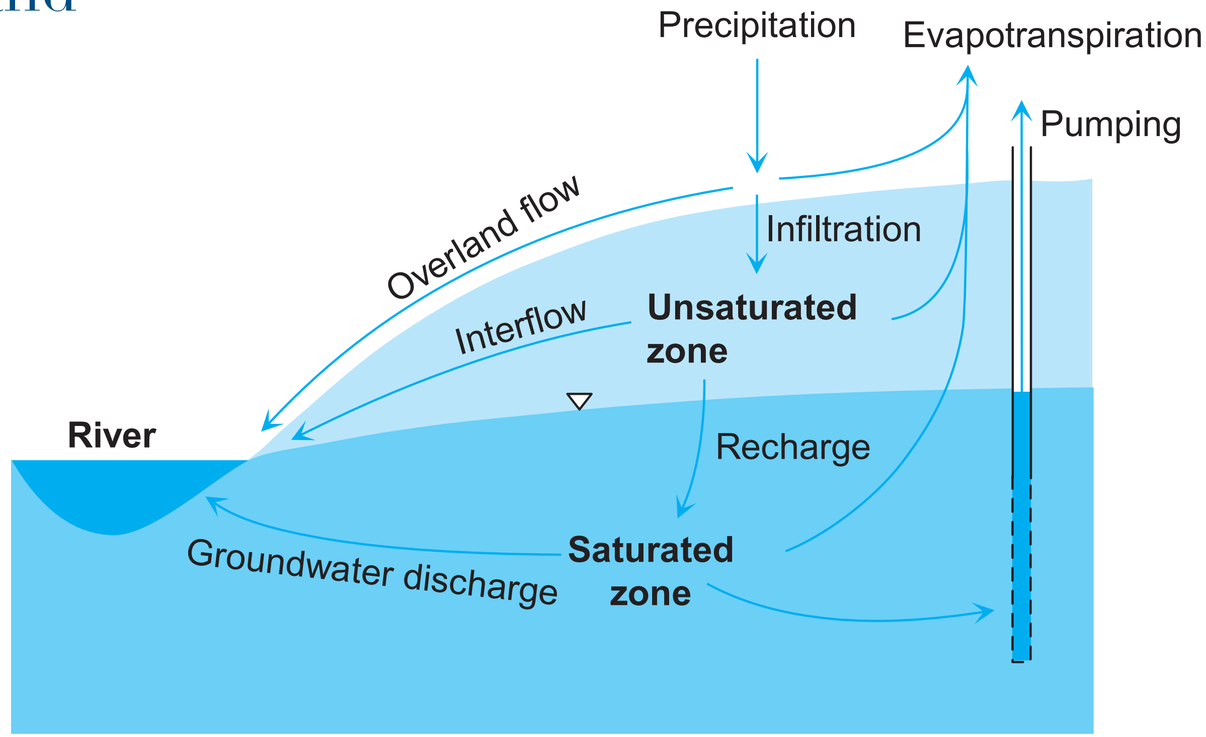


The big picture

Infiltration and recharge

Infiltration is favored where there is porous and permeable soil or rock, flat topography, and a history of dry conditions





Fitts (2002)

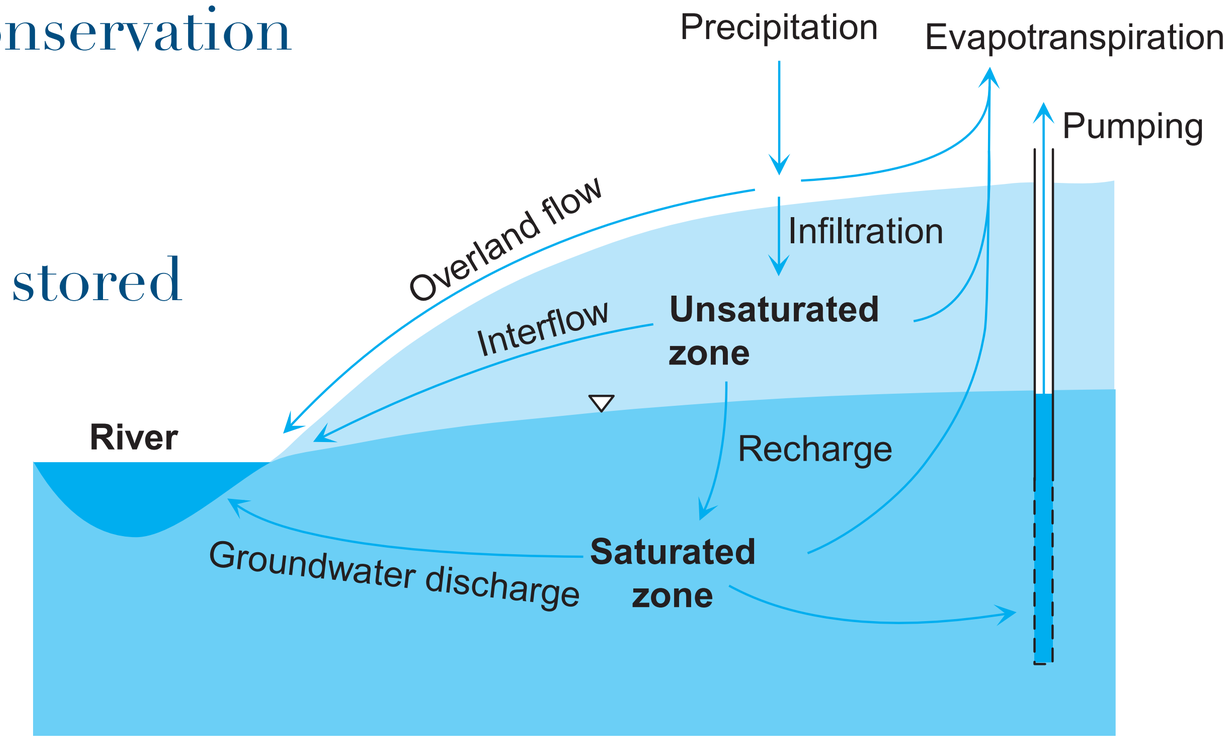
The big picture

Hydrologic balance

Hydrologic balance is the basic concept of conservation of mass concerning water fluxes

flux in – flux out = rate of change in water stored

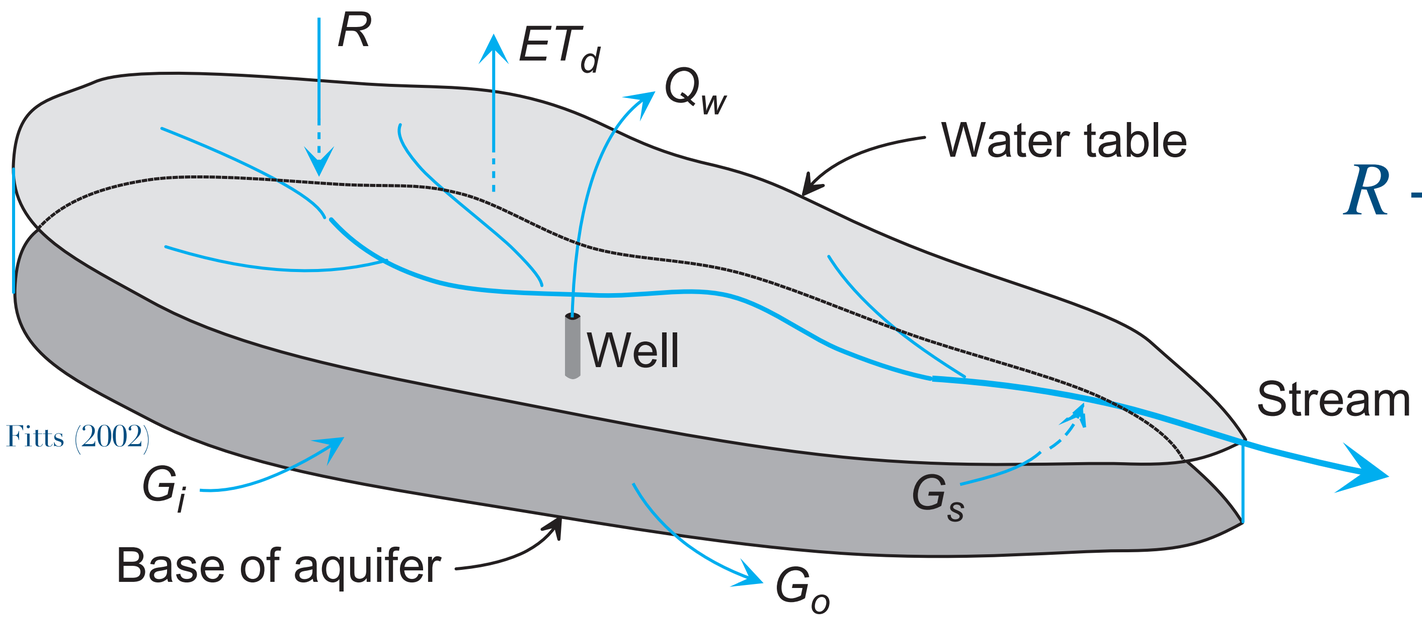
Hydrologic balance is useful for estimating unknown fluxes in many different hydrologic systems



Fitts (2002)

The big picture

Hydrologic balance



Fitts (2002)

Transient conditions

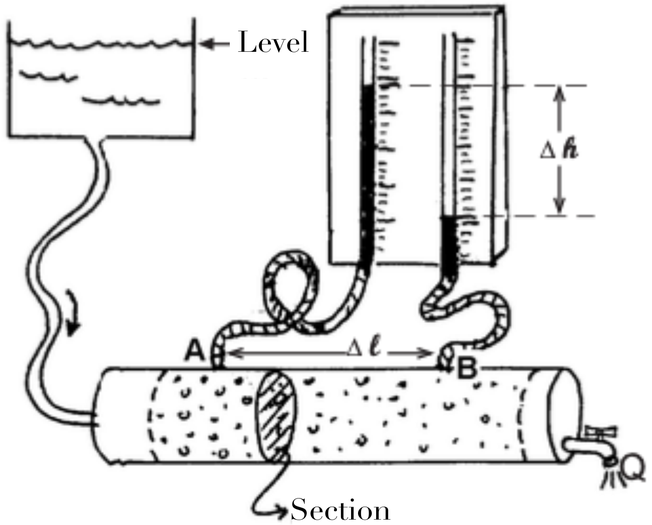
$$R + G_i - G_0 - G_s - ET_d - Q_w = \frac{dV}{dt}$$

Steady-State conditions

$$R + G_i - G_0 - G_s - ET_d - Q_w = 0$$

The big picture

Hydrology and geology



Darcy: Principles of flow

$$Q \ \alpha \ \Delta h \quad \& \quad Q \ \alpha \ \frac{1}{\Delta l}$$

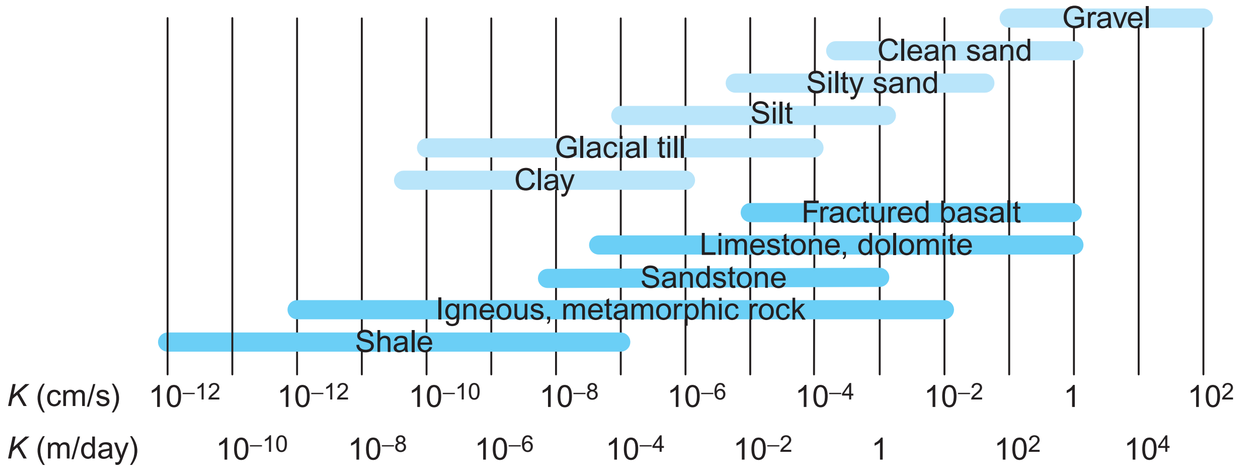
$$Q = -k\frac{dh}{dl}A$$

Henry Darcy was in charge of studying the water supply network of Dijon (1856)

He was interested in the factors that influence water flow through sand materials

The big picture

Hydrology and geology



Darcy: Principles of flow

$$Q \ \alpha \ \Delta h \quad \& \quad Q \ \alpha \ \frac{1}{\Delta l}$$

$$Q = -k\frac{dh}{dl}A$$

Henry Darcy was in charge of studying the water supply network of Dijon (1856)

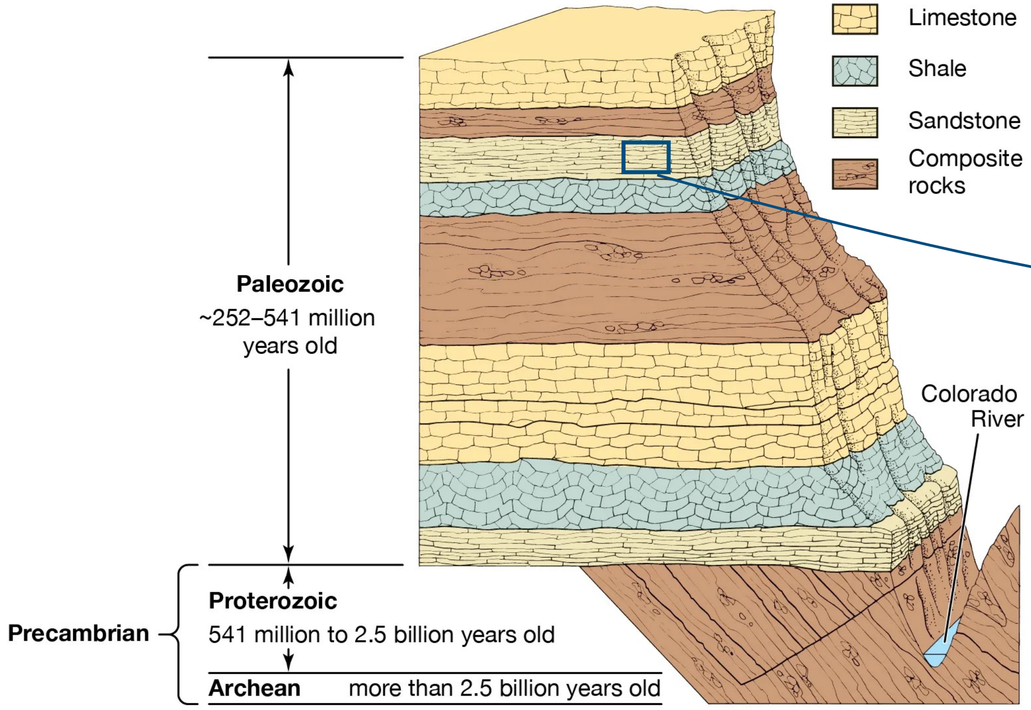
He was interested in the factors that influence water flow through sand materials

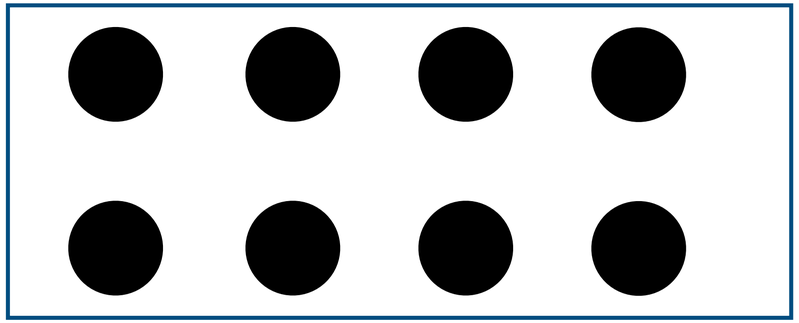
The big picture

Hydrology and geology

Darcy: Limitations

Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great



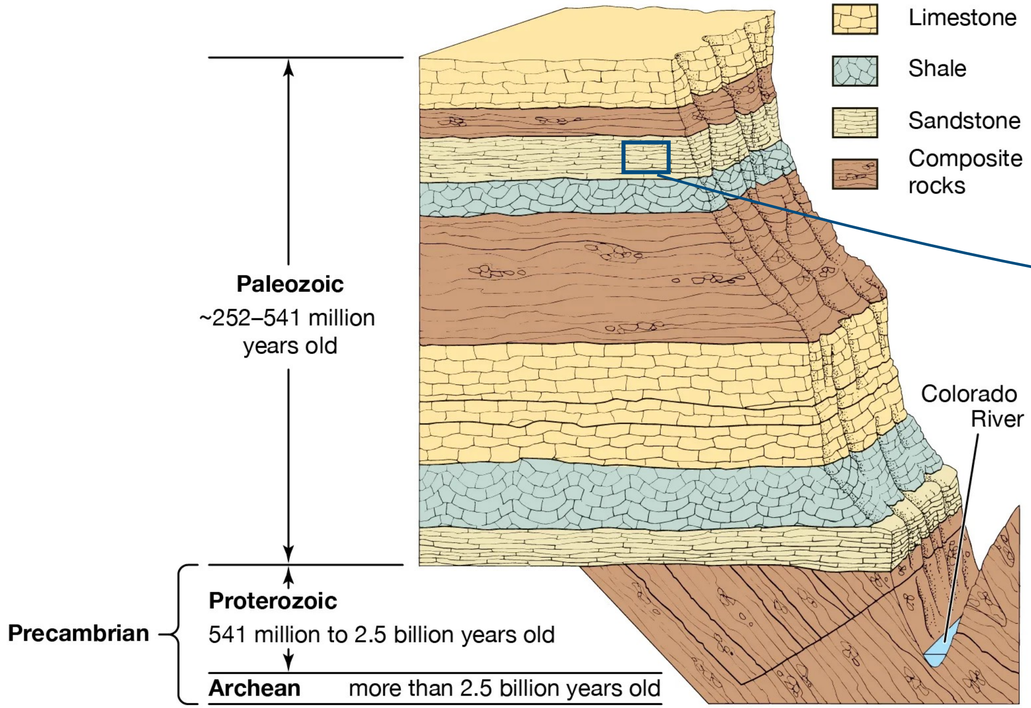


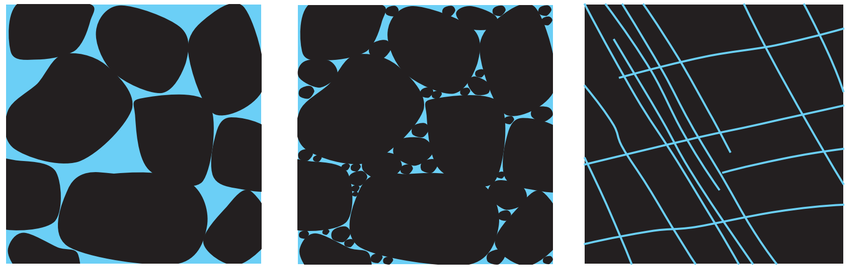
The big picture

Hydrology and geology

Darcy: Limitations

Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great



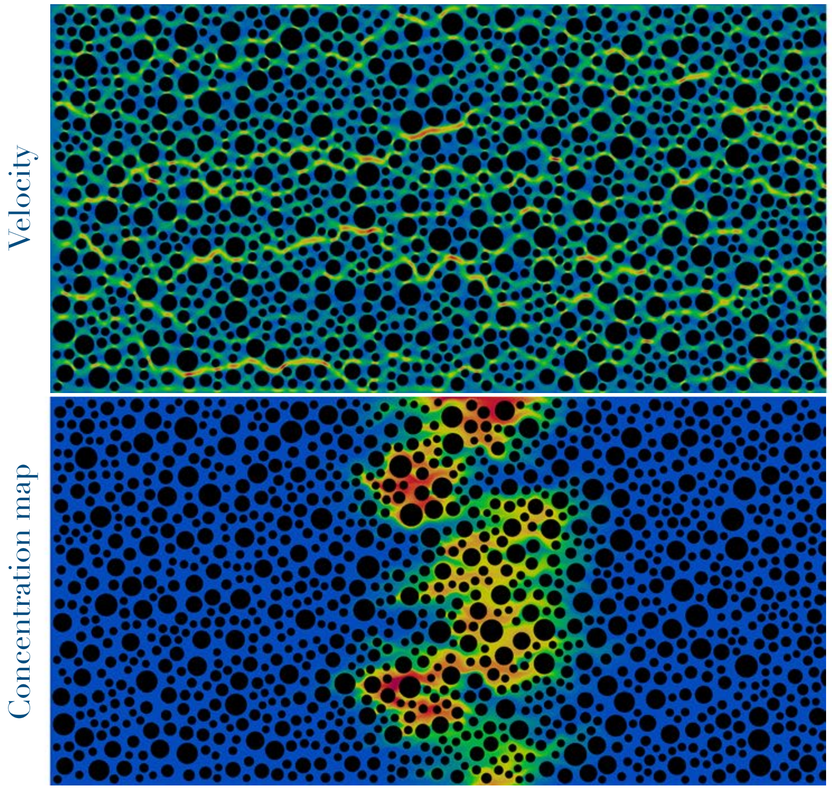


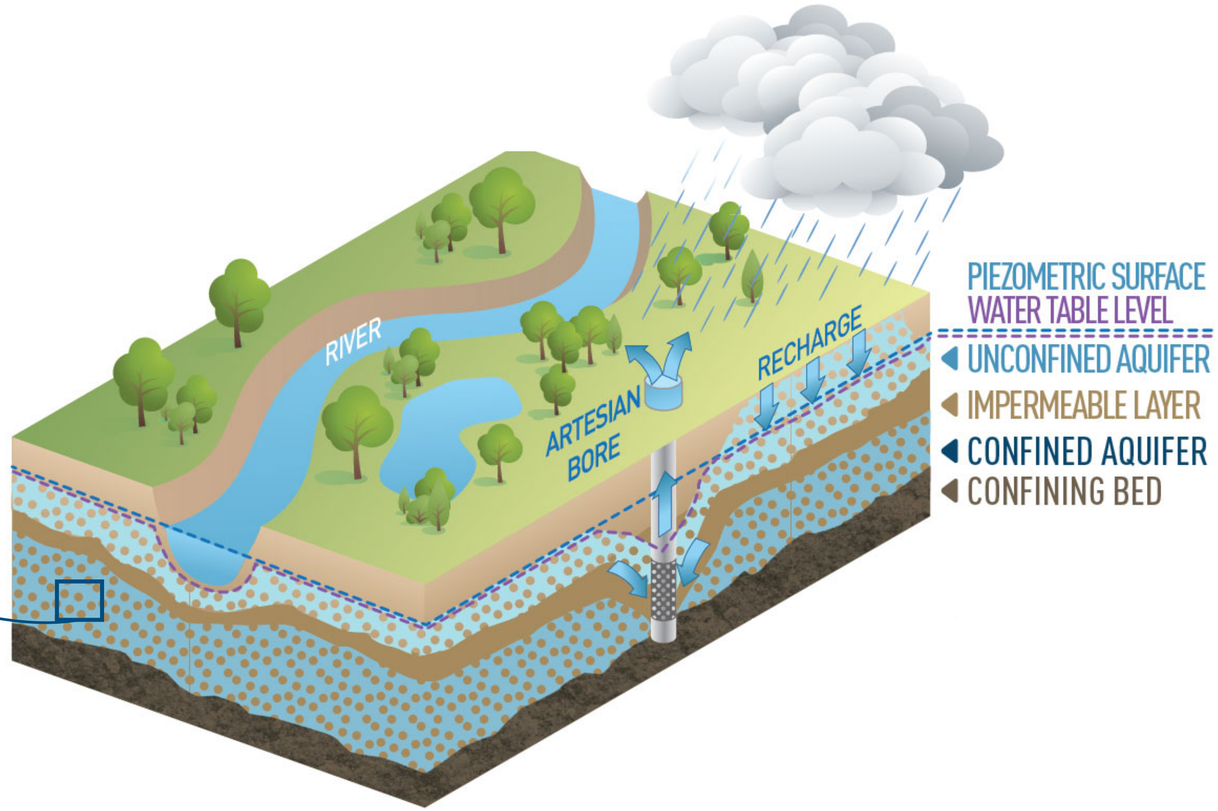
The big picture

Hydrology and geology

Darcy: Limitations

Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great





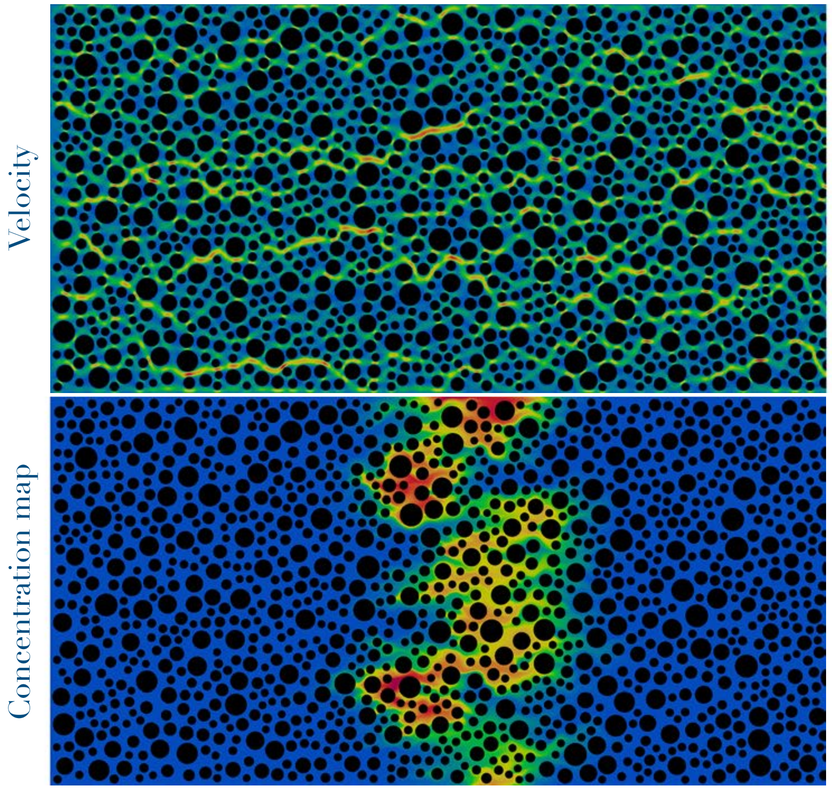
Dentz et al., 2022

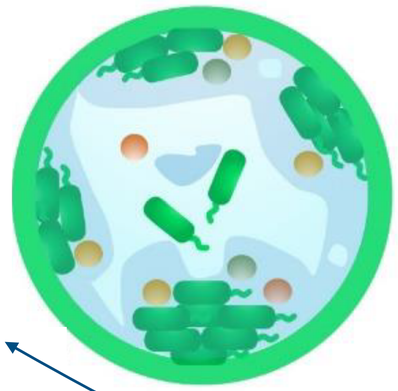
The big picture

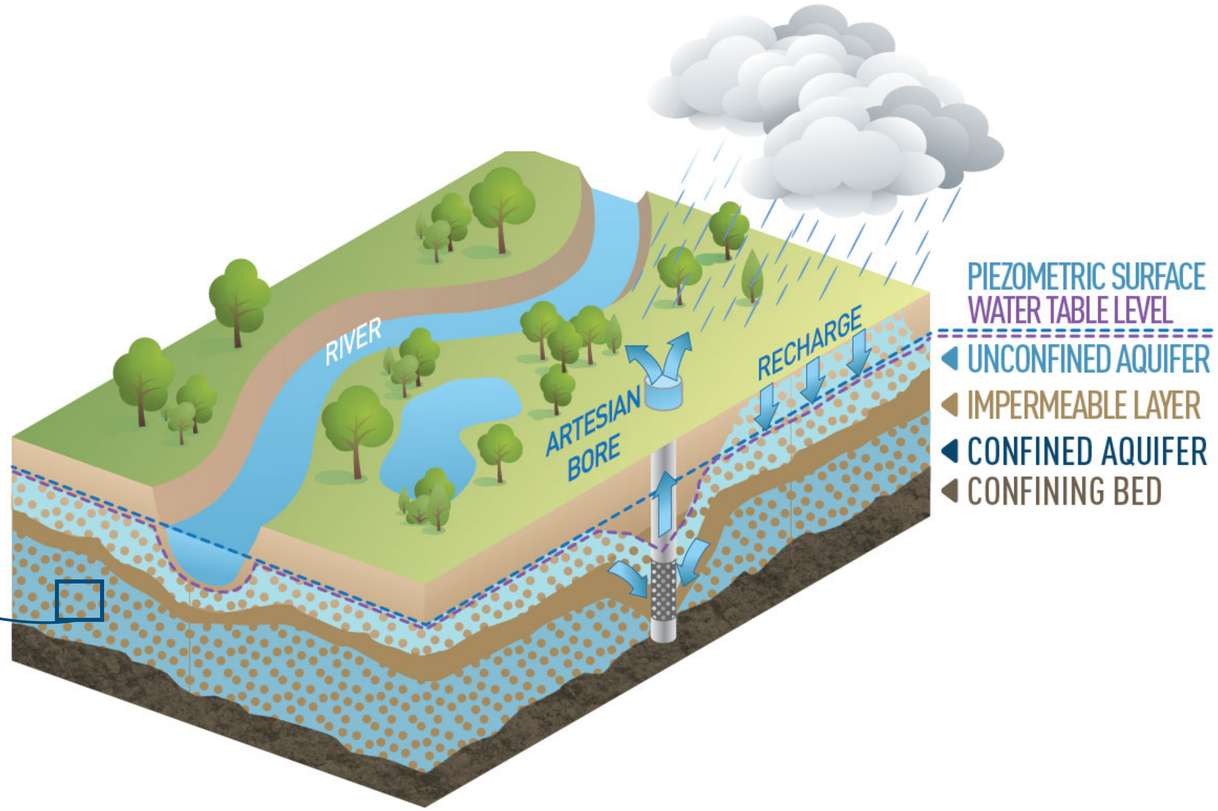
Hydrology and geology

Darcy: Limitations

Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great







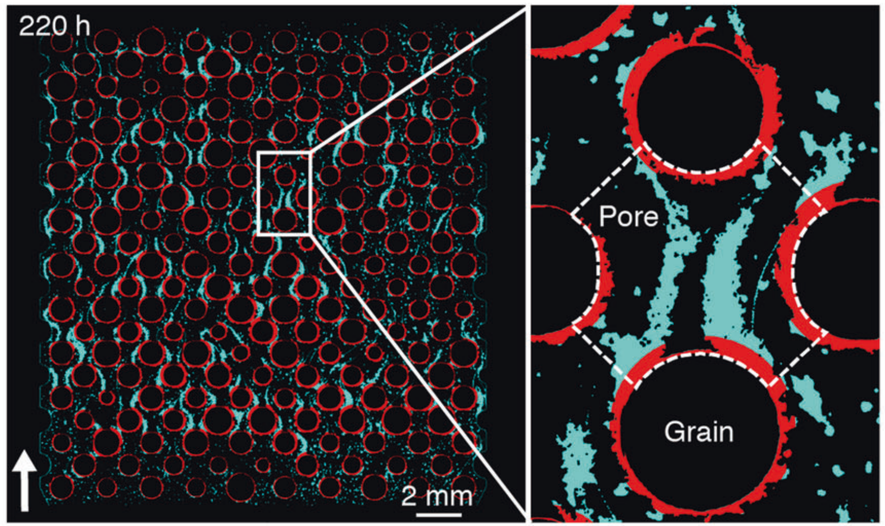
Dentz et al., 2022

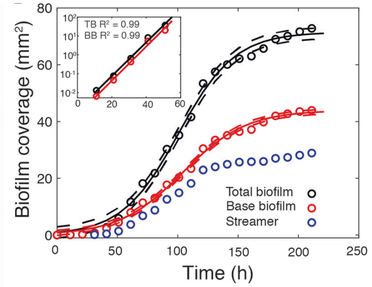
The big picture

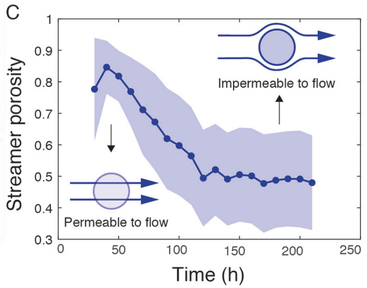
Hydrology and geology

Darcy: Limitations

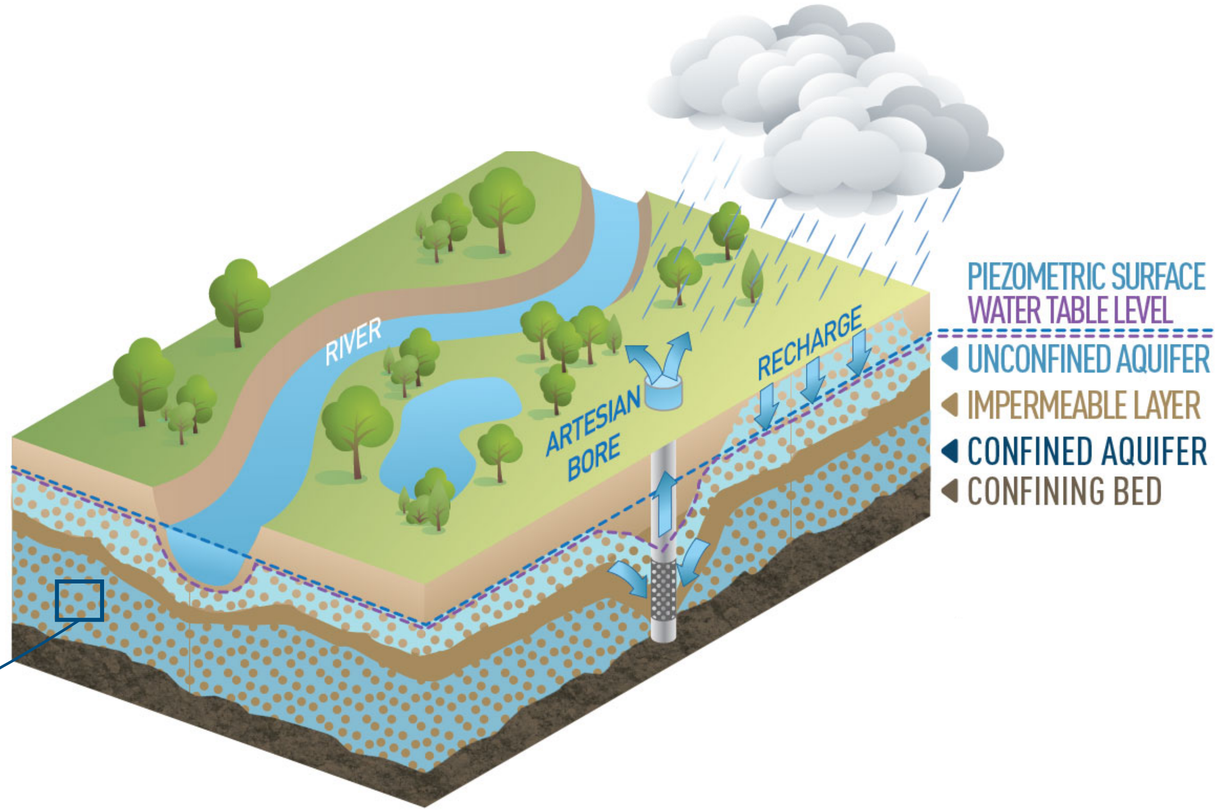
Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great







Schweidler et al., 2021

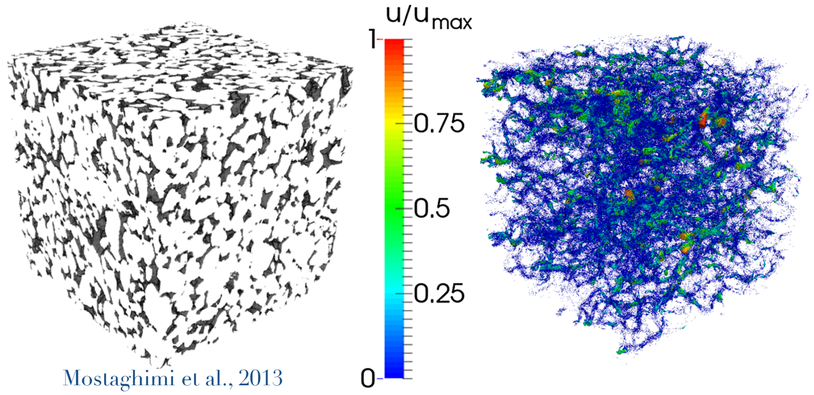


The big picture

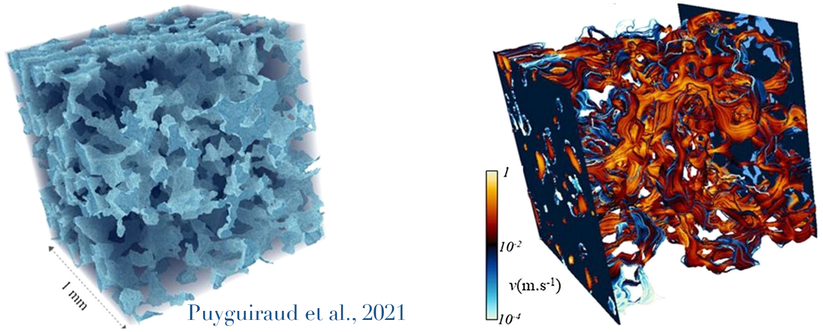
Hydrology and geology

Darcy: Limitations

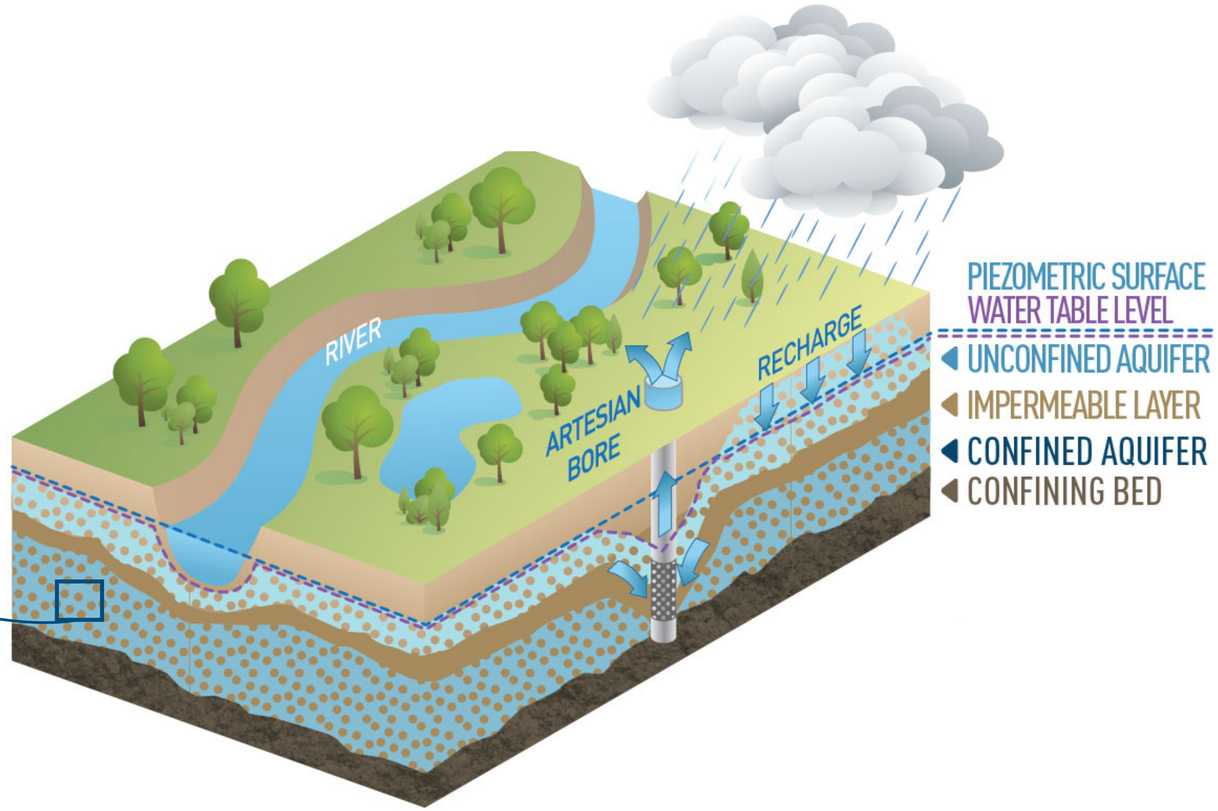
Darcy's law can be inappropriate if the medium is too irregular or if the flow velocity is too great



Mostaghimi et al., 2013

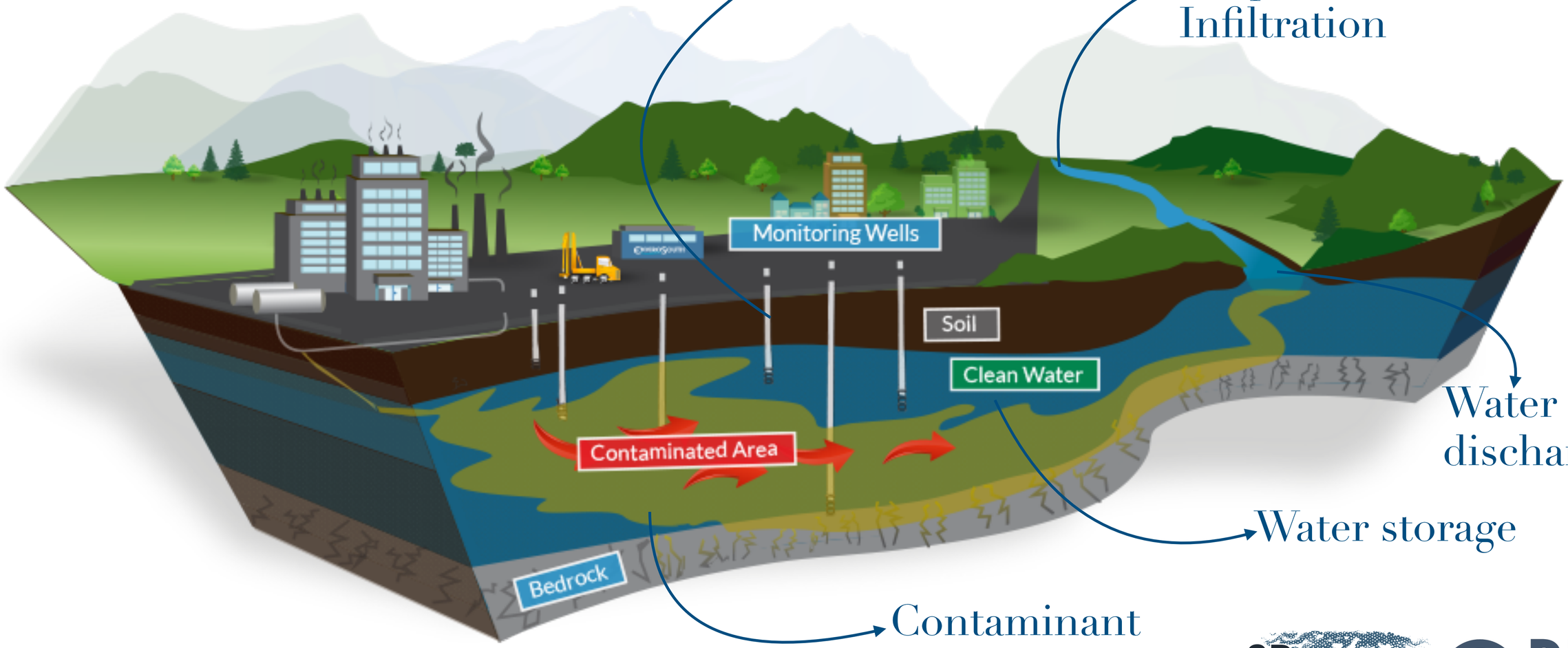


Puyguiraud et al., 2021

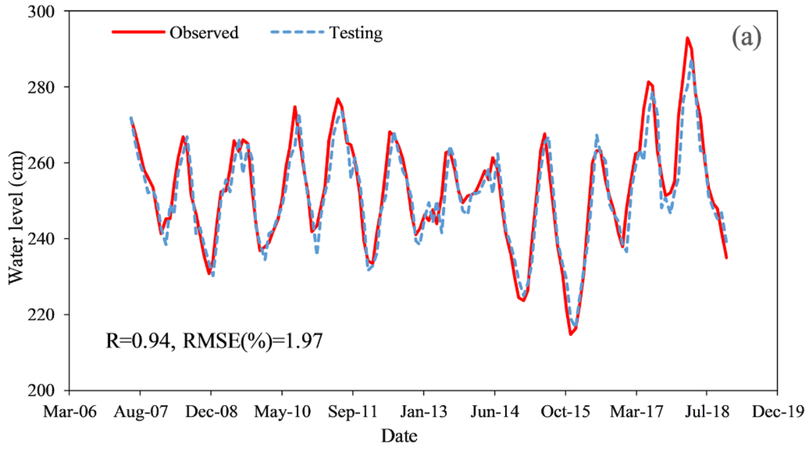


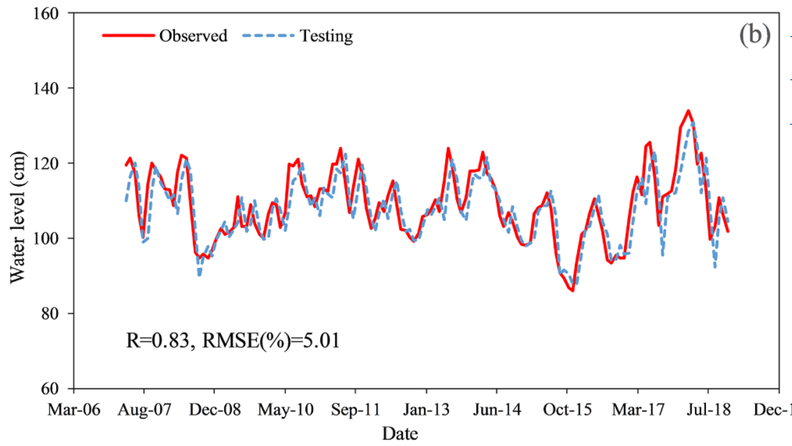
The big picture

Hydrology and geology



Machine learning in hydrology

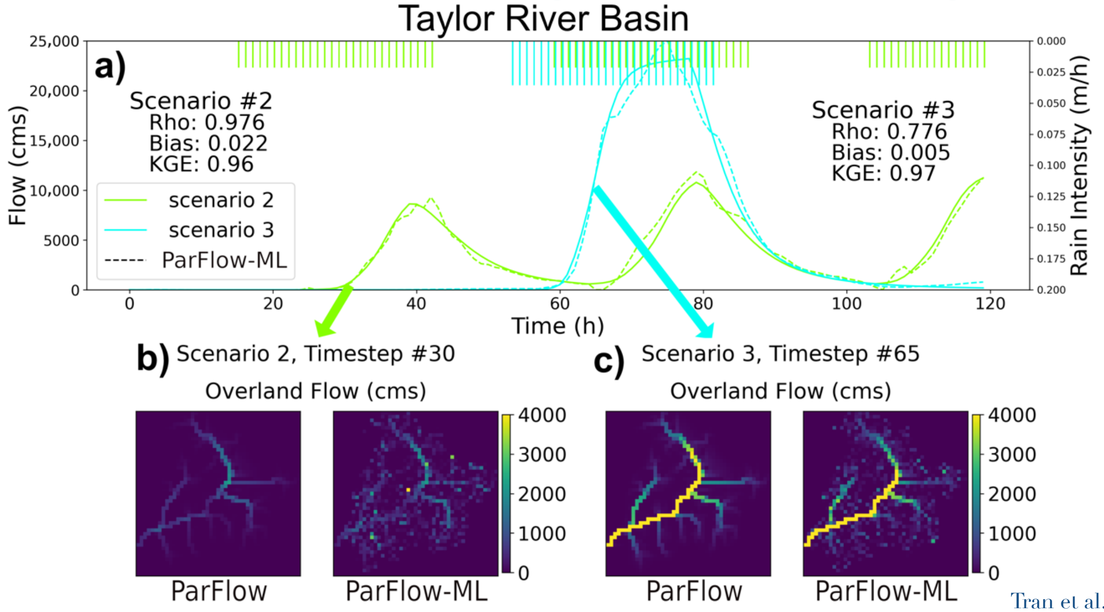




Predicting water levels using decision trees

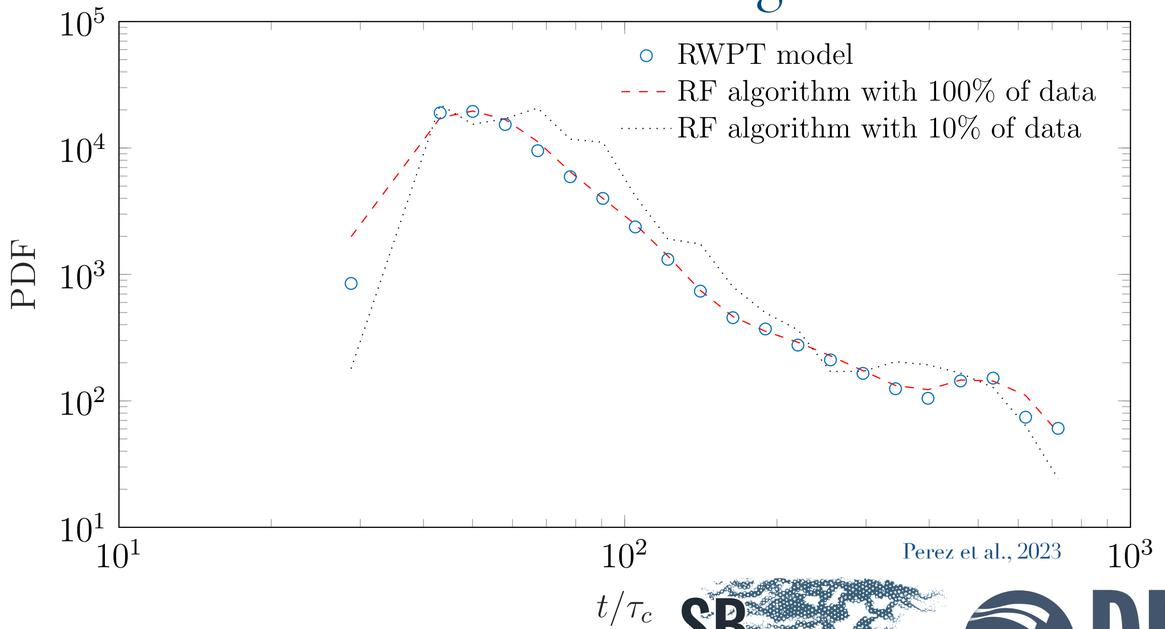
Zhu et al., 2020

Overland flow using Gaussian Regression



Tran et al., 2021

Contaminant arrival using random forests



Perez et al., 2023

Machine learning in hydrology

Date	Topic	Assignment
01/24	Intro & ML Basics (Laz)	
01/31	Applications of ML in Hydrology (Laz)	
02/02	Matlab & Random Forests (Laz)	H1
02/09	Python (Marc)	H2
02/14	K-Nearest Neighbor (Marc)	
02/21	Stochastic Gradient Descent (Marc)	H3
02/28	Support Vector Machine (Laz)	

Statistics

Regression statistics

Where is the model performing well?

Where is the model performing poorly?

What practical implications do the metrics have?

Is the performance reasonable given the model's application?

Statistics

Regression statistics

- Residuals

$$r_i = y_i - \hat{y}_i$$

This is the model's error for each data point.
All the regression metrics are summary statistics for these values

- Mean absolute error

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i |$$

The average magnitude of the residuals. This is an easy-to-interpret metric that has the same units as the response

Statistics

Regression statistics

- Mean square error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The average of the squared residuals. Most types of regression will minimize this term to train the model. Because of the squaring term, it's more sensitive to significant errors and outliers than the MAE

- Root mean square error

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Same units as MAE but also emphasizes large errors

Statistics

Regression statistics

- Sum of squared errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Sum of squares total

$$SST = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$$

Statistics

Regression statistics

$$R^2$$

$$R^2 = \frac{SST - SSE}{SST}$$

The relative difference in the total error obtained by fitting a model, so a value between 0 and 1.
If a model fits the data well, the model error is small and will be close to 1. If the model fits the data poorly, then the model error is large and will be close to 0.
This metric is also called the Coefficient of Determination.

Statistics

Regression statistics

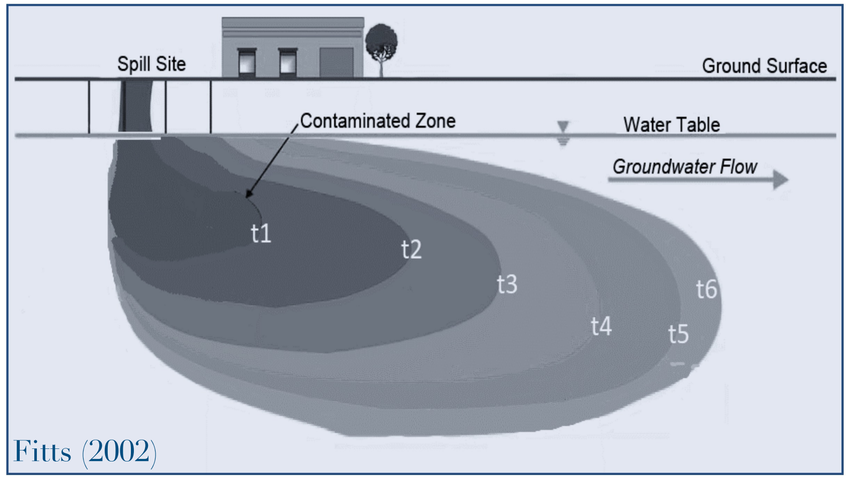
Where is the model performing well?

Where is the model performing poorly?

What practical implications do the metrics have?

Is the performance reasonable given the model's application?

Statistics



Fitts (2002)

Where is the model performing well?

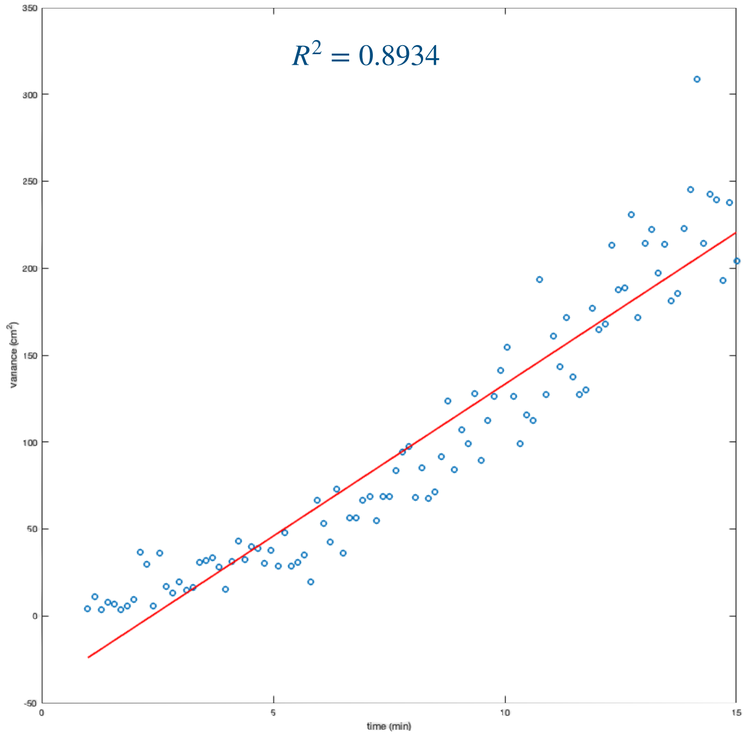
What practical implications do the metrics have?

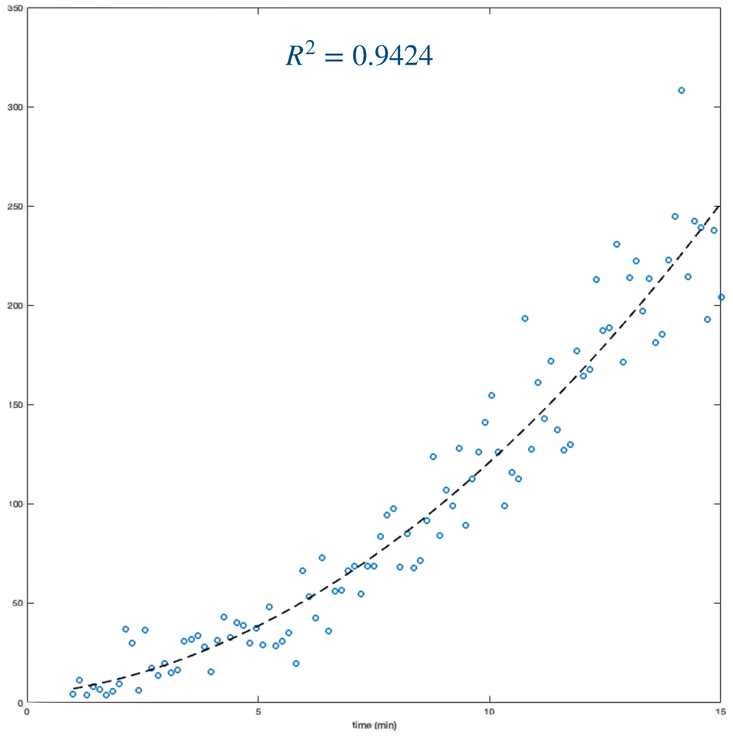
Where is the model performing poorly?

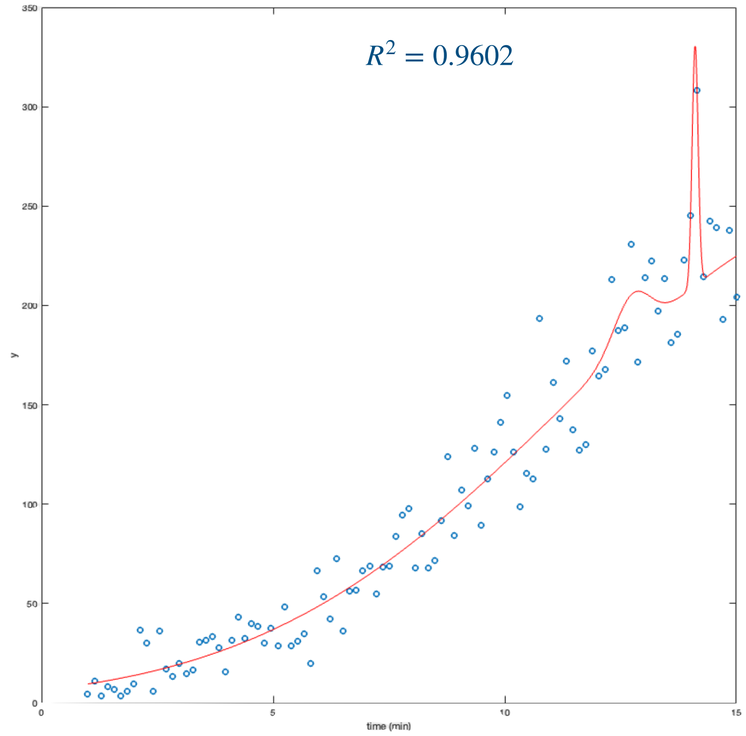
Is the performance reasonable given the model's application?

Error

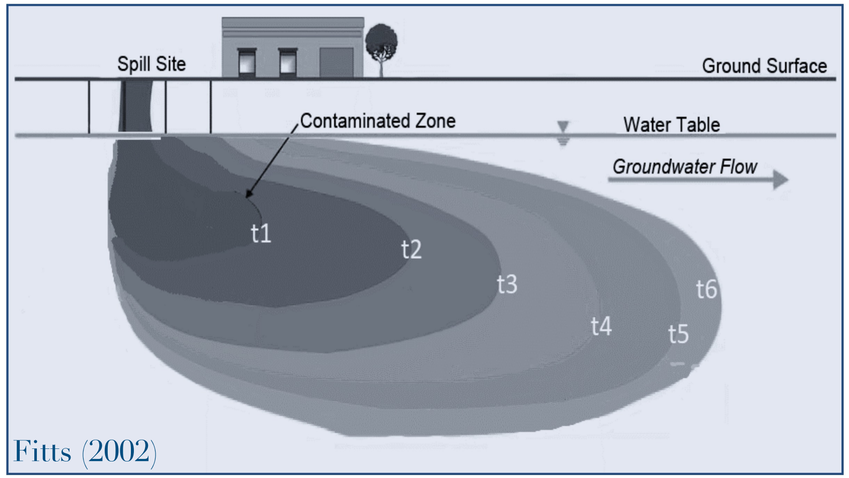
Complexity







Statistics



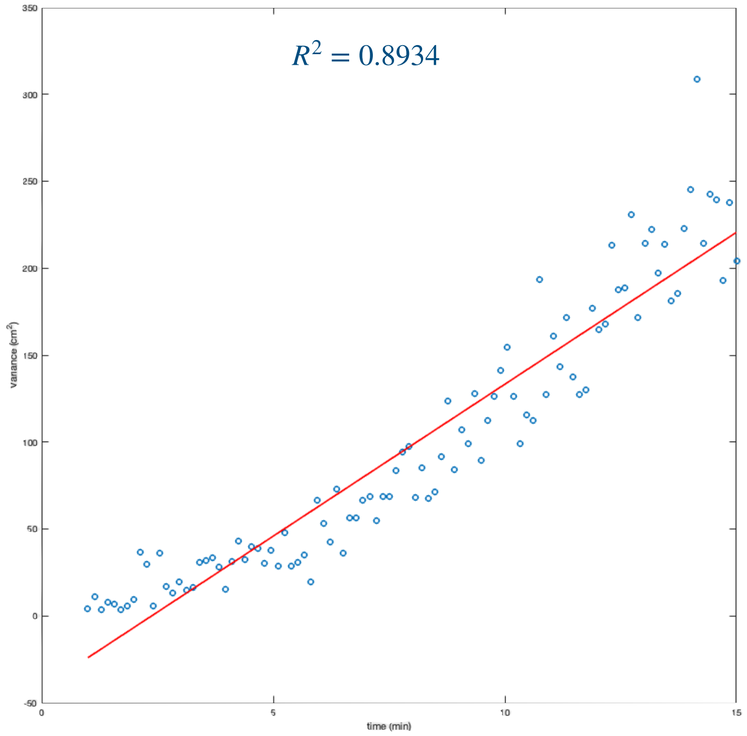
Fitts (2002)

Where is the model performing well?

What practical implications do the metrics have?

Where is the model performing poorly?

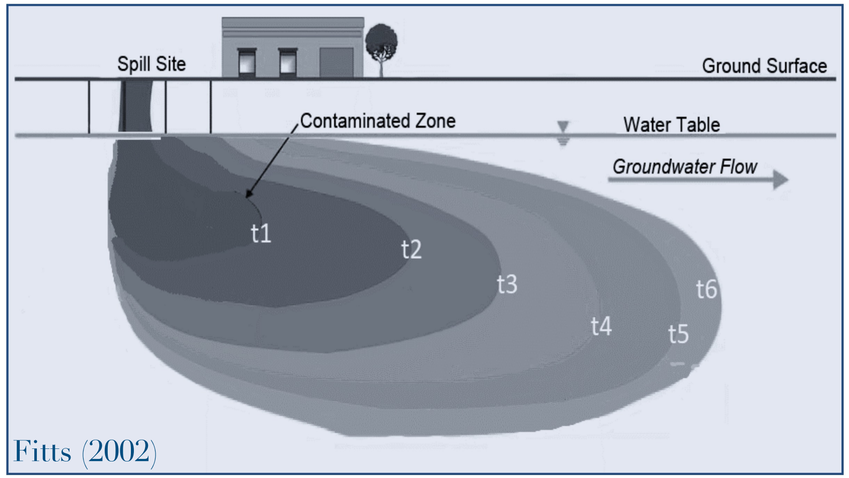
Is the performance reasonable given the model's application?



Linear models

- ✅ Easy to interpret
- ❌ Do not capture complex trends
- ❌ Large errors

Statistics



Fitts (2002)

Where is the model performing well?

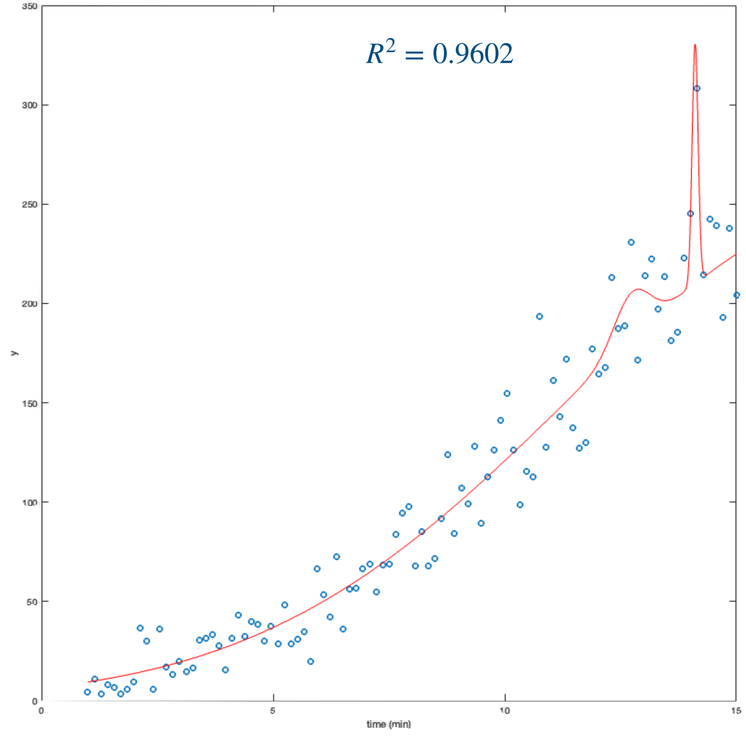
What practical implications do the metrics have?

Where is the model performing poorly?

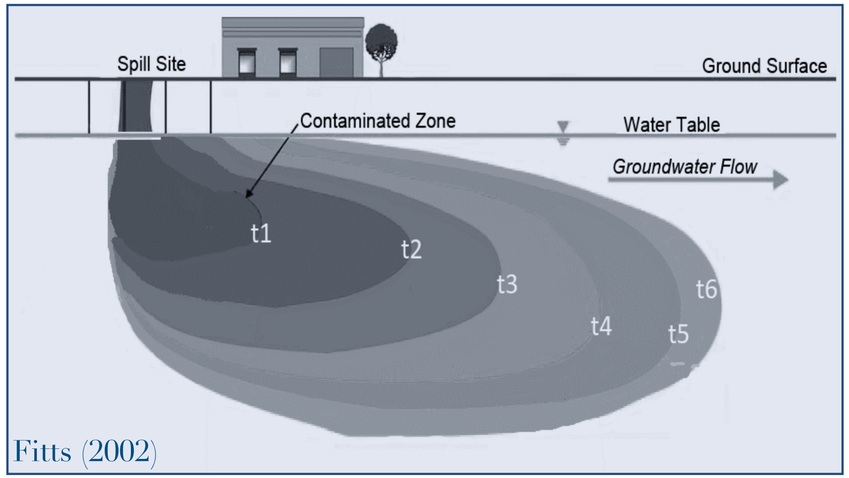
Is the performance reasonable given the model's application?

Complex models

- ✅ Capture complex trends
- ❌ Difficult to interpret
- ❌ May model noisy data



Statistics



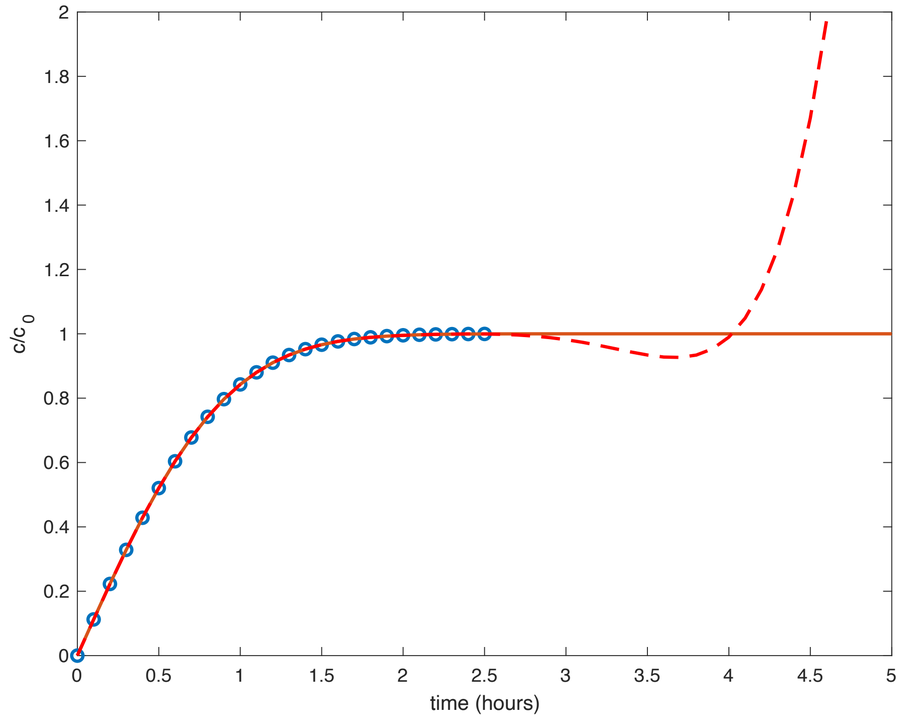
Fitts (2002)

Where is the model performing well?

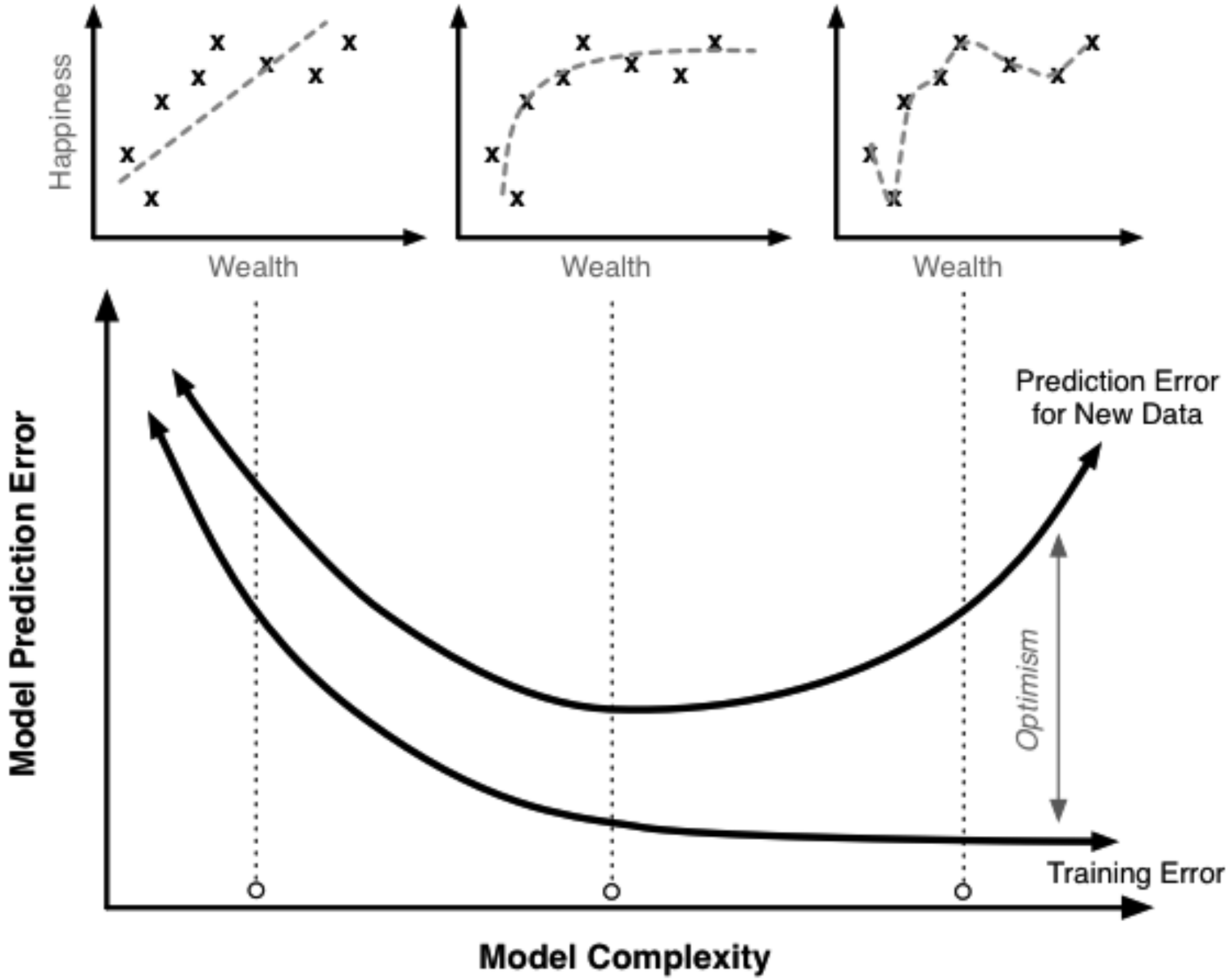
What practical implications do the metrics have?

Where is the model performing poorly?

Is the performance reasonable given the model's application?



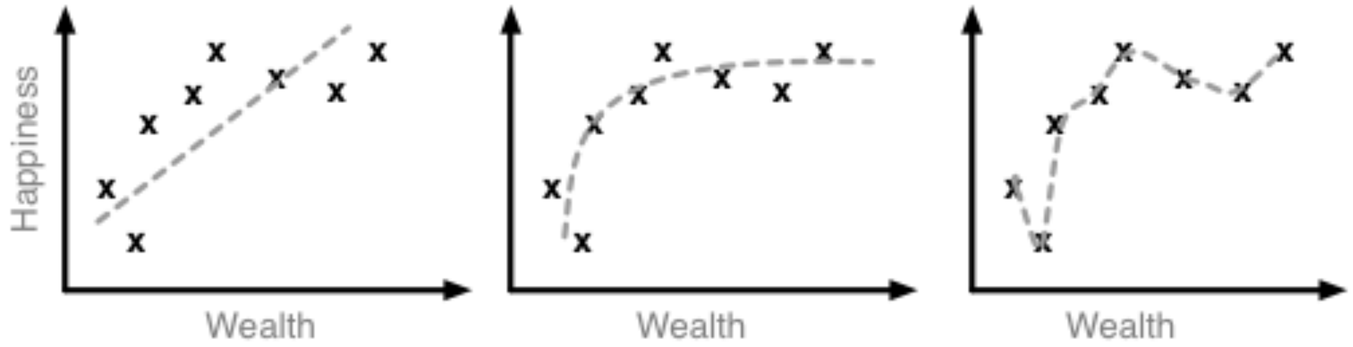
Statistics

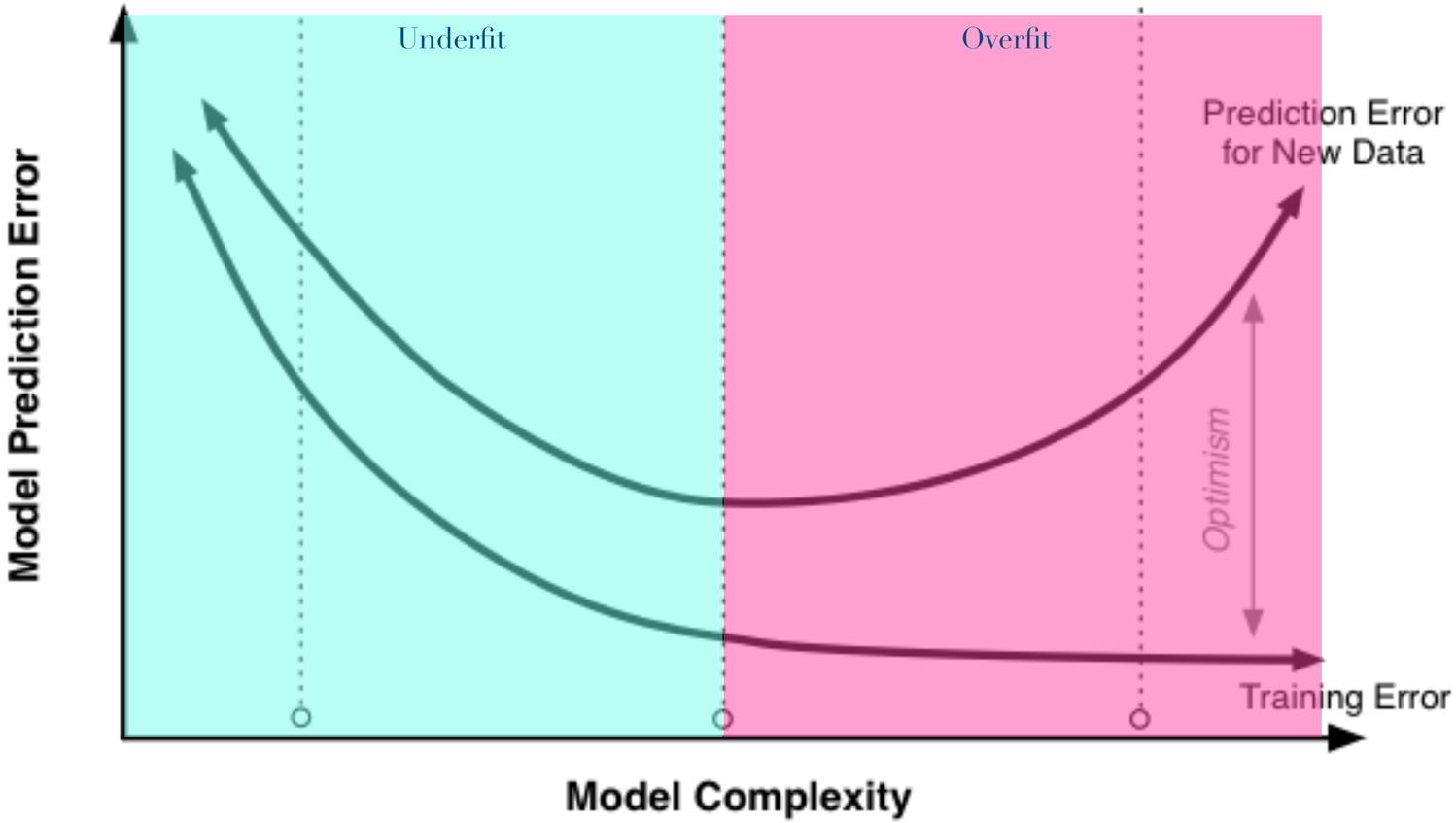


Decreasing error with training data by increasing model complexity produces higher error with new data

Find optimal point where model captures trends in training data but is flexible enough to predict new data

Statistics





What can you do?

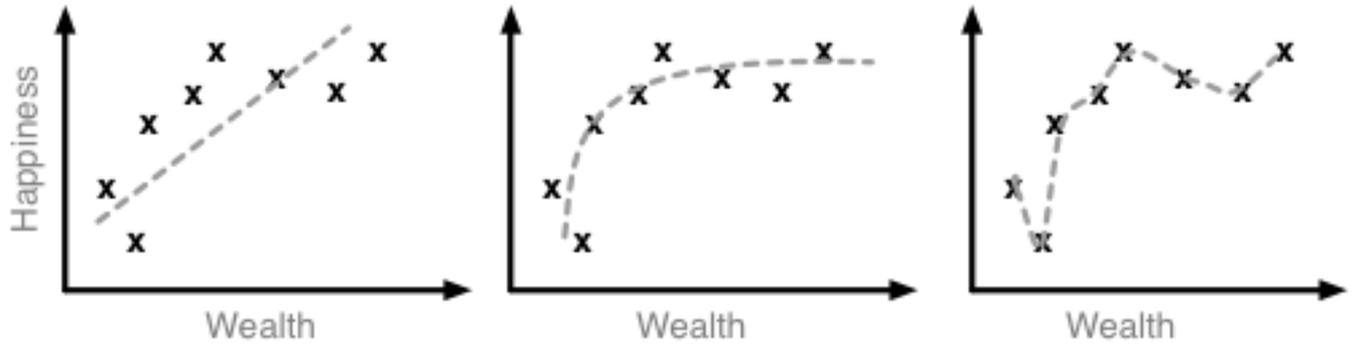
Decrease model complexity

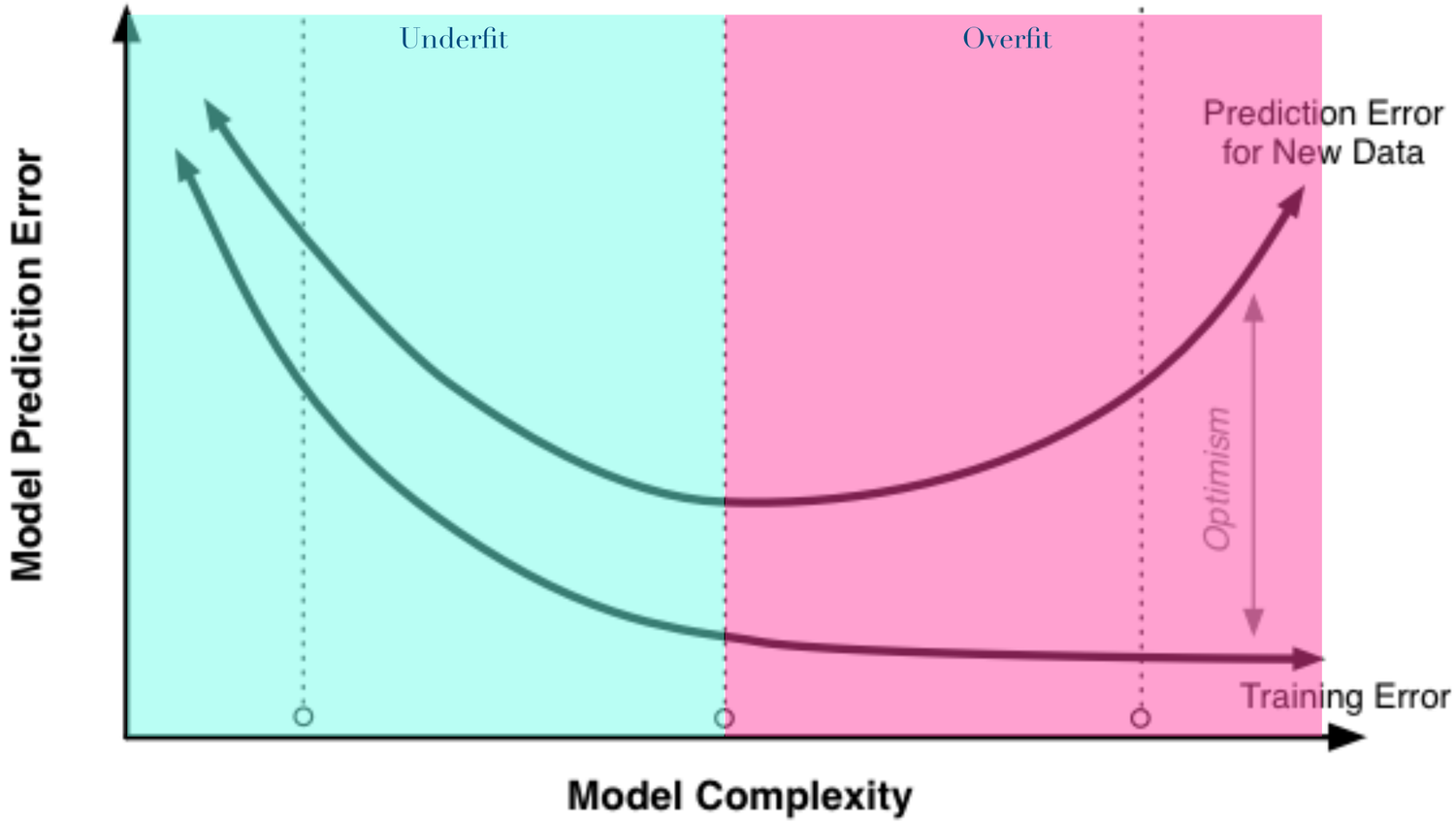
Add more data

Reduce number of features

Regularize model

Statistics





What can you do?

Decrease model complexity

Add more data

Reduce number of features

Regularize model

Will be covered in decision trees