# Systemy uczące się
# **Drzewa regresji**
# wykład 6

Jerzy Stefanowski
Instytut Informatyki PP
2021

Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI-TECH)
projekt finansowany z środków Programu Operacyjnego Polska Cyfrowa
POPC.03.02.00-00-0001/20

Fundusze Europejskie
Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

# Plan wykładu

1. Zadanie regresji w uczeniu maszynowym
2. Ograniczenia klasycznych modeli liniowych, podejścia nieparametryczne i wykorzystanie podziałów dziedziny zmiennych niezależnych
3. Drzewa - rekurencyjny podział przestrzeni cech oraz estymacja predykcji zmiennej wyjściowej – ilustracja oraz przykład
4. Drzewa regresji vs. drzewa klasyfikacyjne
5. Kryterium podziału w węźle
6. Zatrzymanie budowy drzewa vs. tzw. post-pruning
7. Inne rodzaje drzew, tzw. model trees

# Przypomnienie regresji

**Zadanie regresji (predykcja zmiennej liczbowej)**

- Metoda oszacowania wartości liczbowej zmiennej zależnej (objaśnianej) *y* na podstawie wartości zmiennych niezależnych **x** [klasyczne w statystyce]
- Poszukujemy modelu $\hat{y} = f(\mathbf{x},\boldsymbol{\beta})$ − wybór postaci funkcji *f* oraz estymacja parametrów

Popularne modele liniowe − regresja wieloraka / wielowymiarowa

$$y=x_1w_1+x_2w_2+\ldots+x_mw_m+w_0$$

Na ogół minimalizacja funkcji straty w postaci RSME (dot reszty $y - \hat{y}$)

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

Różne metody szacowania (MNK, Est.Najw. Wiaryg., Spadek Gradientu)

Dostępna w wielu programach, np. SAS, SPSS, R lub Statistica,…

Liczne zastosowania praktyczne

# Liczne zastosowania

Predykcja:

- Wyceny produktów finansowych, akcji giełdowych, portfolio analiza
- Cen sprzedaży, wynajmu mieszkań
- Sektor sprzedaży różnych produktów
- Poziomu satysfakcji klientów oraz czasu współpracy w CRM, rynku ubezpieczeniowym, itd.

Ocena pracochłonności projektów (COCOMO)

Model oceny efektywności systemów (np. komputerowych)

Analiza ryzyka przedsięwzięć

I wiele innych, ….

# Przykład predykcji cen mieszkań

- Harrison i Rubinfeld – badanie związku między róznymi wskaźnikami jakości życia a cenami nieruchomości w okolicach Bostonu tzw. **Boston Housing** patrz lib.stat.cmu.edu/datasets/boston
- 506 domów opisanych przez 14 cech
- Zadanie – predykcja ceny nieruchomości,  pośrednio poziom zanieczyszczenia (koncentracja tlenku azotu)
  1. CRIM - per capita crime rate by town
  2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS - proportion of non-retail business acres per town.
  4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
  5. NOX - nitric oxides concentration (parts per 10 million)
  6. RM - average number of rooms per dwelling
  7. AGE - proportion of owner-occupied units built prior to 1940
  8. DIS - weighted distances to five Boston employment centres
  9. RAD - index of accessibility to radial highways
  10. TAX - full-value property-tax rate per $10,000
  11. PTRATIO - pupil-teacher ratio by town
  12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
  13. LSTAT - % lower status of the population
  14. MEDV - Median value of owner-occupied homes in $1000's

# Inne repozytoria

- **Kaggle** – kilka konkursów predykcji cen nieruchomości (np. Ames data lub new york)
- **UCI ML repository** – specjalna kolekcja benchmarkowych danych dla regresji (**134 zbiory**) https://archive.ics.uci.edu/ml/index.php
- Spojrzeć do artykułu: PMLB: a large benchmark suite for machine learning evaluation and comparison (2017) – repozytorium https://github.com/EpistasisLab/pmlb

# Poprzednie wykłady i laboratorium

**Modele liniowe** − też regresja

Różne formy funkcji straty (nie tylko błąd resztowy $y - \hat{y}$)

Zasady estymacji metodą największej wiarygodności

Przeuczenie − na przykładzie różnych rodzajów regresji

W systemach uczących − modele predykcji zmiennych liczbowych

Nie tylko o podłożu liniowych modeli statystycznych

Także nauczone sieci neuronowe − predykcja wielu zmiennych liczbowych $y_1 y_2,\ldots,y_k$ − neurony wyjściowe

# Regularyzacja w regresji

- Ridge (regresja grzbietowa) - Czynnik dodatkowy

$$\sum_{j=1}^{m} (w_j)^2 \leq t$$

gdzie *t* jest ograniczeniem, a całość sformułowania

$$w = \text{argmin}(\sum_{i=1}^{n} (y_i - \sum_{j=1}^{m} x_{ij} w_j - w_0)^2 + \lambda \sum_{j=1}^{m} w_j^2)$$

λ − mnożniki Lagrange'a

Lasso − inny czynnik dodatkowy

$$\sum_{j=1}^{m} |w_j| \leq t$$

# Dyskusja klasycznej regresji

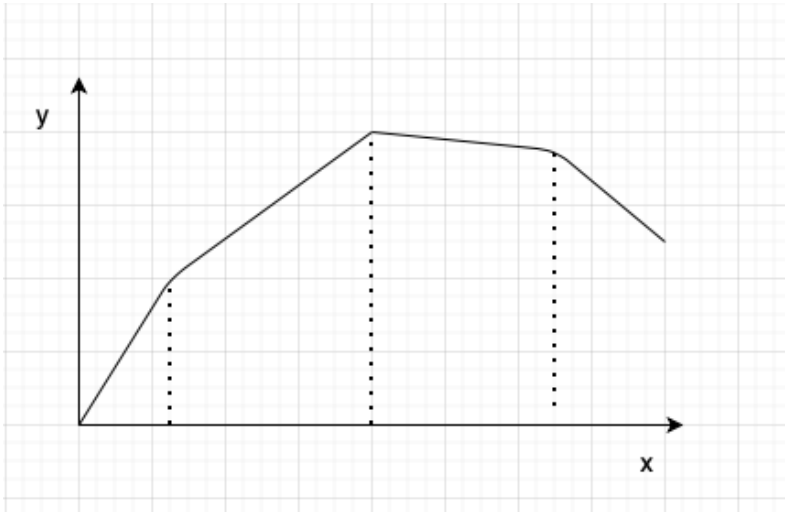
- **Liniowa regresja**
  - Model globalny − zmienne x obejmują całą przestrzeń cech,
  - Założenie liniowości i nieskorelowanie zmiennych – lecz rzeczywiste dane / przykłady uczące mają rozkłady, gdzie cechy mogą być współzależne nieliniowe („świat dla sztucznych sieci neuronowych”)
  - W problemach uczenia maszynowego na ogół „mieszaniny” różnych typów cech / atrybutów
- **Nieliniowa regresja**
  - Metody nieparametryczne estymacji – patrz literatura, np. książka J.Koronackiego Statystyczne systemu uczące
  - Także podział funkcji na segmenty / części

# Aproksymacje lokalne – regresja nieparametryczna

- Więcej w książka J.Koronacki, J.Ćwik: Statystyczne systemy uczące się.
- Estymując funkcję regresji staramy się uwzględnić w modelu własności lokalne
- Składanie kilku „funkcji podstawowych” zdolnych lokalnie przybliżyć własności pewnych podobszarów dziedziny
- Regresyjne funkcje sklejane z „węzłami”
- Tzw. regresja lokalnie ważona

Locally weighted regression

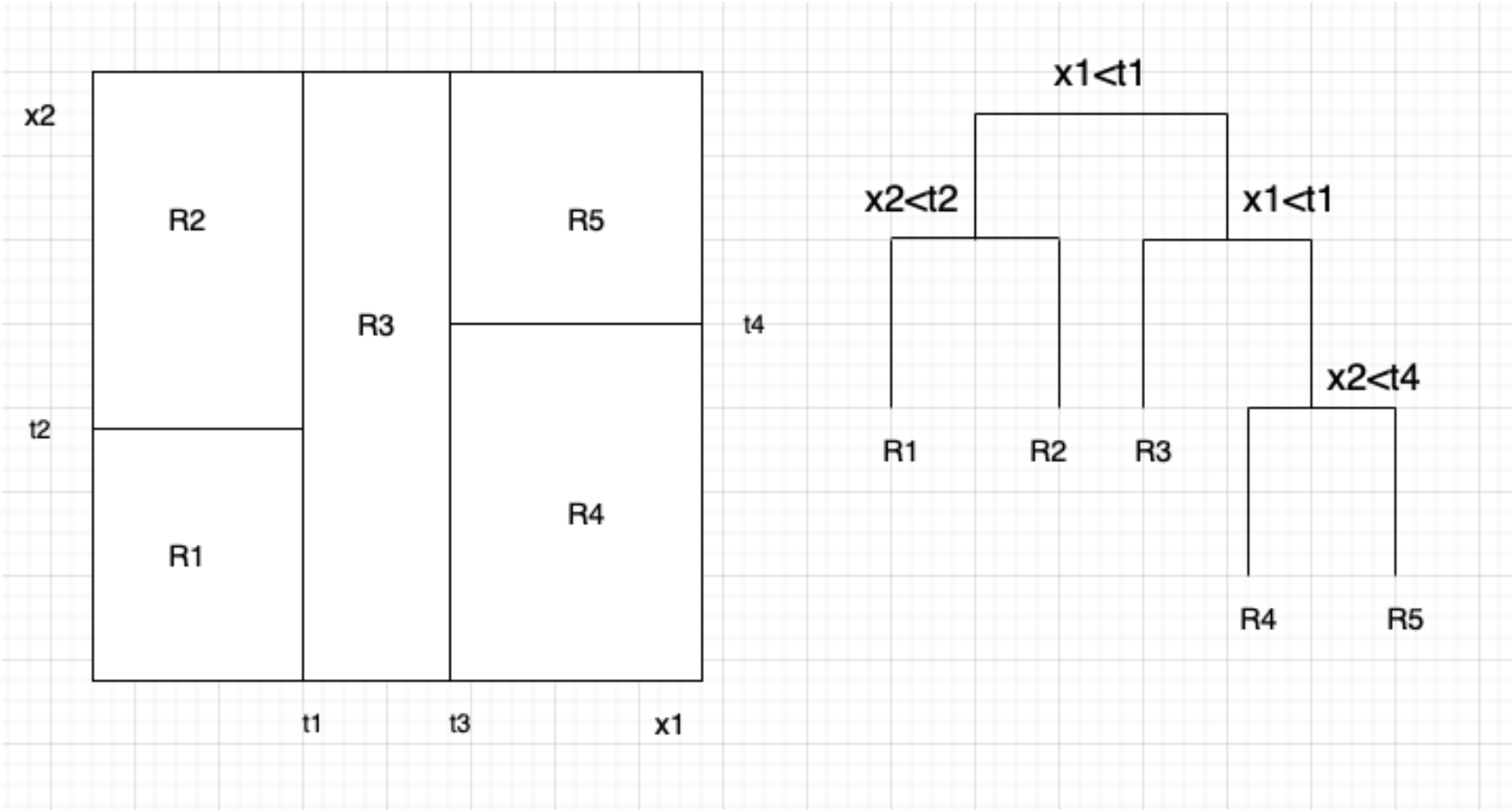
$$y = \alpha + \sum_{j=1}^{p} f_j(\mathbf{x}, \beta)$$



# W stronę drzew regresji

- **Inne podejście do podziałów wielowymiarowej przestrzeni cech /atrybutów**
  - Stopniowo podziel przestrzeń na obszary (CART – hiperkostki - prostopadłościany),
  - Procedura rekurencyjna podziału (top-down) jak w drzewach klasyfikacyjnych)
  - Uwzględnianie różnych typów cech / atrybutów
- **Predykcja**
  - W końcowym obszarze można zbudować prostszy model predykcji *y*
  - Drzewa regresji – estymacja pojedynczej wartości y na podstawie rozkładu przykładów należących do obszaru
  - ang. Model trees – zbuduj model regresji liniowej, jeśli jest wystarczająca liczba przykładów  w obszarze

# Ilustracja drzewa regresji i zasad podziału przestrzeni cech



# Drzewa regresji

- **Węzły drzewa** − sekwencja pytań o testy na wartościach atrybutów (np. is horsepower > 50 and is gradutestudent)
- **Predykcja w obszarze** (hiper-kostce)
  - CART – przykłady należące do hiperkostki $R_j$− w miarę jednorodne (ze względu na charakterystykę $x$ + możliwe wyjście $y$ − czyli posiadają dość podobne wartości $y$ dla $x_i$ z $R_j$
  - Oszacowujemy wartość przeciętną wśród dla $x_i$ z $R_j$ -> średnia arytmetyczna ŷ w ($R_j$ )

$$\hat{y} = \frac{1}{\left|R_j\right|} \sum_{x_i \in R_j} y_i$$

  - Podziały w liściach muszą być dość homogeniczne i reprezentowane są przez stała wartość (średnią)! - lepszy estymator niż mediana z uwagi na kryterium oceny drzewa

# Przykład danych cars

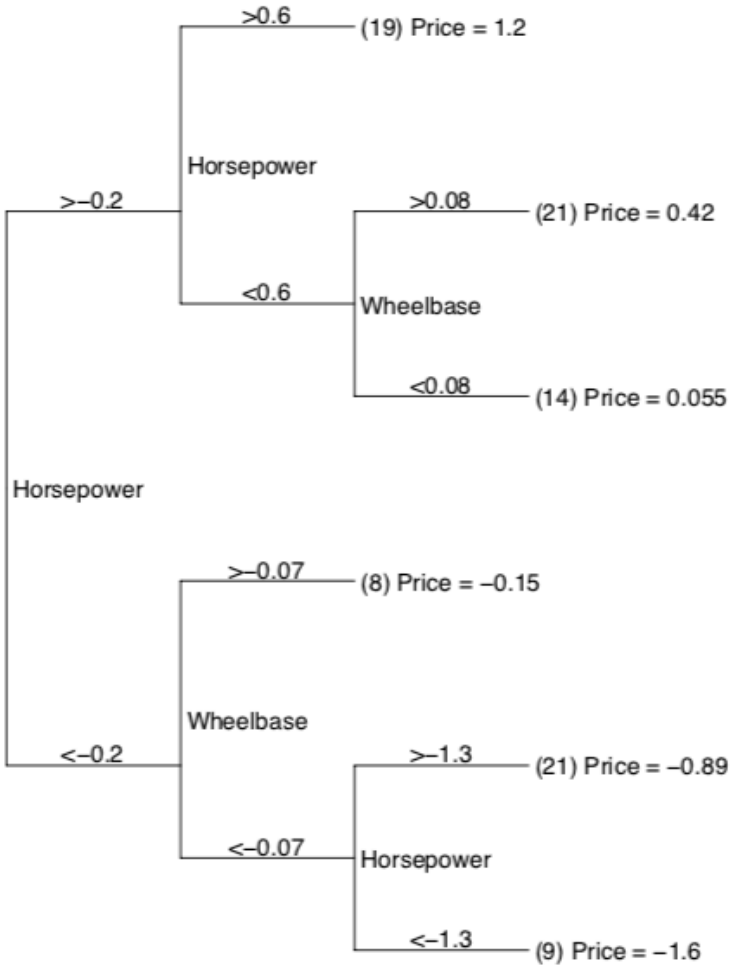



Figure 1: Regression tree for predicting price of 1993-model cars. All features

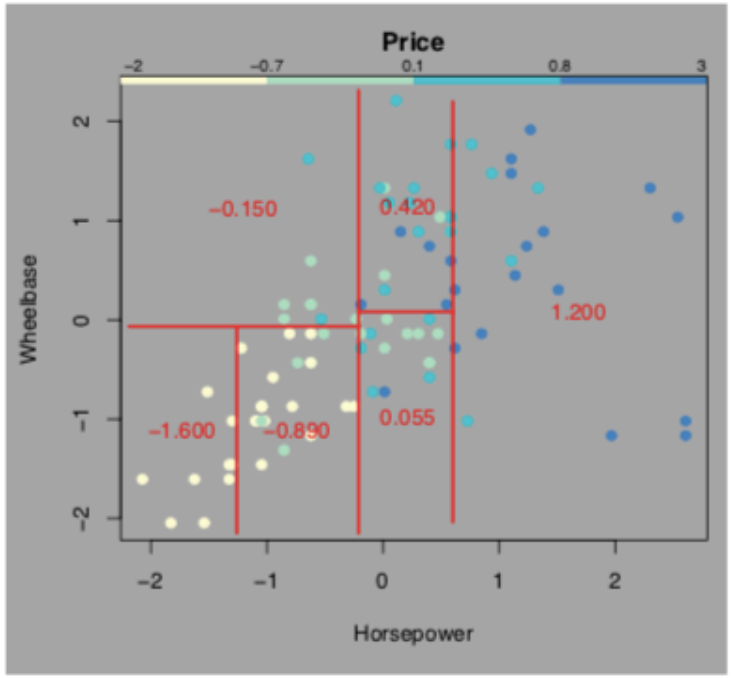



Figure 2: The partition of the data implied by the regression tree from Figure 1

# Drzewa regresyjna a klasyfikacyjne

- Podobne zasady rekurencyjnego podziału zbioru przykładów uczących
- Podobna struktura drzewa
- Inne kryteria (podziału, stopu, ..)
- Inne miara oceny predykcji − funkcja straty ciągła, błąd średniokwadratowy $(y - \hat{y})^2$
- Ponadto możliwość upraszczania, redukcji wielkości drzewa
- Spójrz do pracy przeglądowej Wei-Yin Loh: Fifty Years of Classification and Regression Trees, International Statistical Review (2014)

# Problemy w budowie drzew regresji

- Kryteria oszacowania jakości drzewa
- Zasady wykonania podziału w węźle
- Określenie kryterium stopu (kiedy węzeł drzewa stanie się liściem)
- Alternatywne upraszczanie drzewa (tzw. ang. pruning)

# Kryterium oceny

- Ogólne kryterium oceny predykcji drzewa *T*, wybierz drzewo min. błąd predykcji

$$\min_{T} \quad \frac{1}{n} \sum_{i=1}^{n} (y_i - T(x_i))^2$$

- W przypadku drzewa regresji dzielącego przestrzeń na *J* obszarów $R_j$ − można minimalizować

$$\sum_{j=1}^{J} \sum_{i \in Rj} (y_i - \hat{y}_{Rj})$$

- Dalsze wersje drzew regresji − można dodać czynnik regularyzacji, zwłaszcza dla upraszczania (patrz książka E. Gatnara)

# Kryterium podziału w węźle

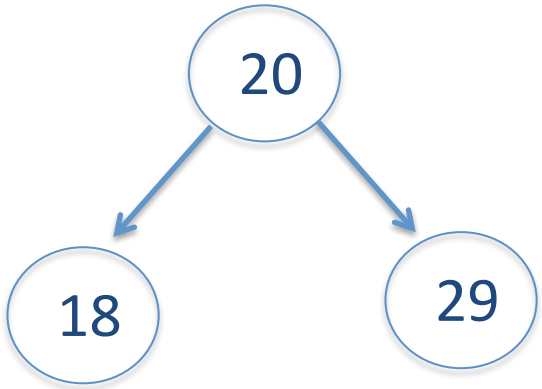
- Zbiór przykładów S zostanie podzielony na dwie części S1 i S2 (np., wg. progu $\tau$ na atrybucie $A$)
- Ocena podziału wg. kryterium błędu

$$\sum_{i \in S1} (y_i - \hat{y}_{S1})^2 + \sum_{i \in S2} (y_i - \hat{y}_{S2})^2$$

- Dla atrybutu $A$ sprawdza się możliwe progi $\tau$ i wybiera ten, który minimalizuje błąd po podziale
- Dla atrybutów nominalnych – możliwe więcej podziałów niż dwa S = S1,S2,..,Sv − wtedy suma v elementów
- Niektóre źródła − ważenie liczbą przykładów w $S_j$ lub ich prawdopodobieństwem $p_j$

# Intuicja podziałów

- Niezależnie od minimalizacji kryterium błędu
- Dąży się do tego, aby w obszarach (kostkach) przykłady miały zbliżone do siebie wartości
- Ponadto stara się rozdzielić wartości y niższe od wyższych i przydzielać je wydzielonych węzłów




Średnie warrtości y
w kostce / obszarze

# Ogólny schemat

- Rozpocznij od pojedynczego węzła z S przykładami – oblicz ŷ na ich podstawie
- Jeśli wszystkie przykłady w S mają tą samą wartość − stop;
  W przeciwnym razie - po wszystkim atrybutach poszukaj najlepszego podziału S, który minimalizuje błąd; jeśli spadek błędu jest zbyt mały lub zbyt mało przykładów (najczęściej wymaga się się aby $n_j >= 5$) to także zatrzymaj
- Gdy stop, to utwórz liść odpowiadający $R_j$ (obl. $\hat{Y}_{Rj}$), w przeciwnym razie powróć do pierwszego punktu

Na ogół w liściu/węźle wymaga się minimalnej liczby przykładów dla obliczeń (CART do 5 przykładów)

# Upraszczanie drzew

Podobnie jak w drzewach klasyfikacyjnych – zbuduj pełne drzewo i i zastosuj tzw. post-pruning

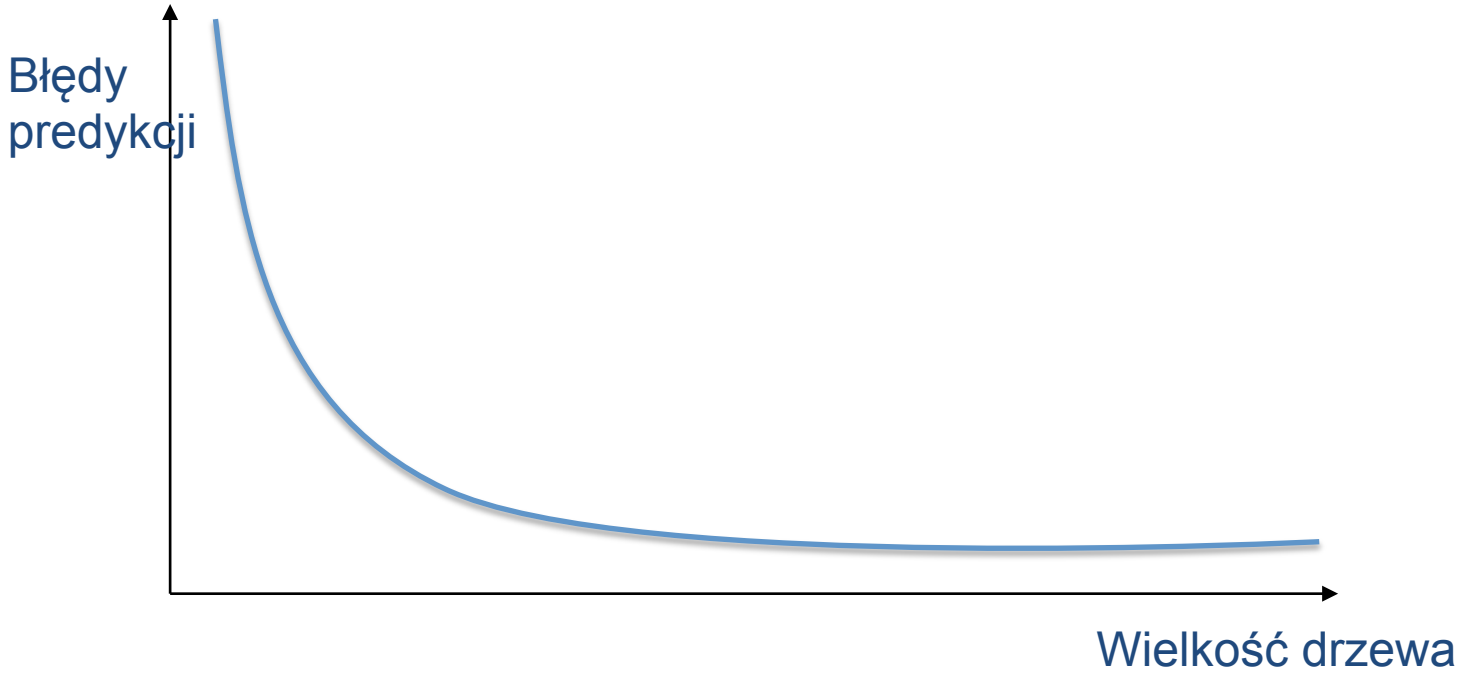
- CART – rozwiązanie cost-complexity
- T – (pod)drzewo do ew. uproszczenia z |T| węzłami ($R_m$)
- $N_m$ = #{$x_i$ in $R_m$} oraz $\hat{y}_{Rm} = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{y}_{Rm})^2$$

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|$$

- α - współczynnik przetargu pomiędzy dobrym dopasowaniem a preferencją dla mniejszych drzew (odpowiednik regularyzacji)
- Ocena na zbiorze walidacyjnym (CART wewn.ocena krzyżowa) tworzy się sekwencje coraz mniejszych drzew i poszukuje najm,

# Upraszczanie drzew regresyjnych

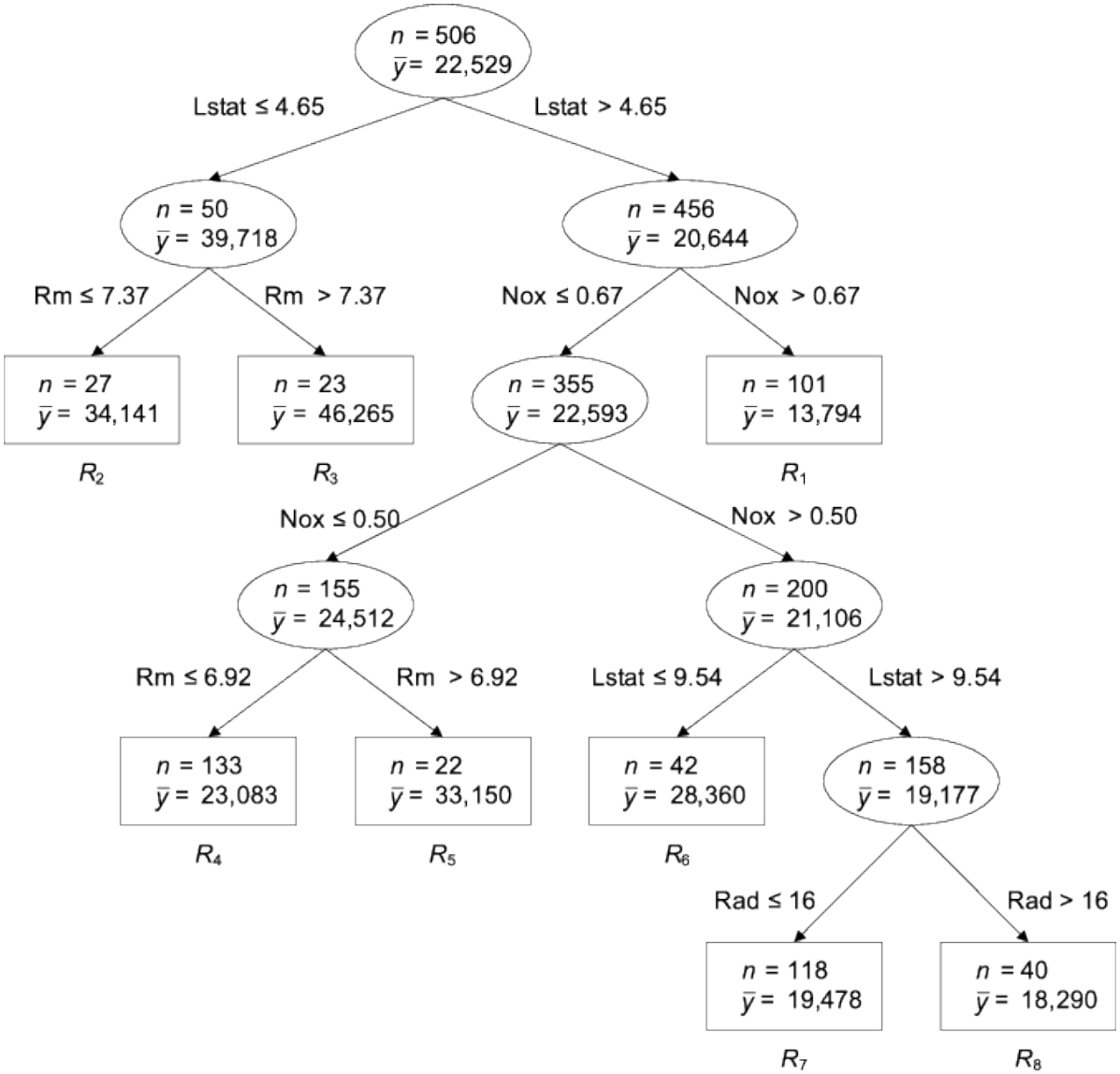



Wykres zmian błędu, w odróżnieniu od drzew klasyfikacyjnych bardziej płaski wykres;

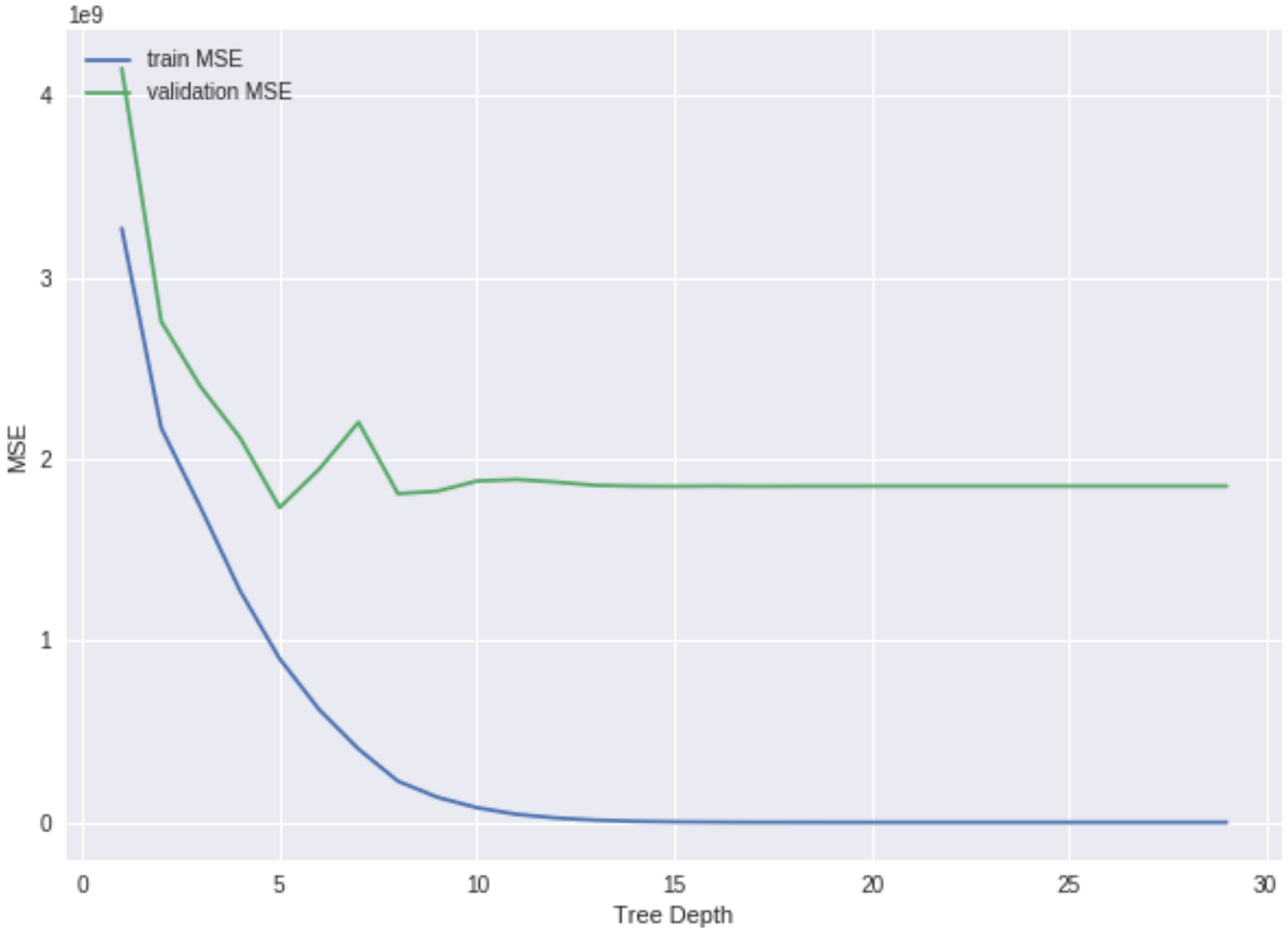
Przycinając drzewo − redukujemy liczbę liści przy jak najmniejszym przyroście błędu − błąd $Q_m$ mniejszy od najmniejszego, powiększonego o jedno odchylenie standardowe w sekwencji drzew

Liczne inne propozycje, także z wykorzystaniem metody LASSO

# Przykład drzewa CART nauczonego z danych Boston housing (pruned)



# Boston housing – upraszczanie drzew



# Cechy drzew regresji

Różnice wobec klasycznych metod regresji:

- Możliwość bezpośredniego użycia różnorodnych zmiennych, w tym wielowartościowych jakościowych (bez specjalnego kodowania zerojedynkowego)
- Nie ma potrzeby standaryzacji, normalizacji zmiennych
- Rozkłady zmiennych nie muszą być normalne (typowa regresja liniowe)
- Modelowanie złożonych nieliniowych współzależności zmiennych (kiedy powierzchnia wielowymiarowej regresji jest bardzo złożona i nieregularna0
- Szybka predykcja
- Wspierania interpretacji struktury danych oraz oceny ważności zmiennych

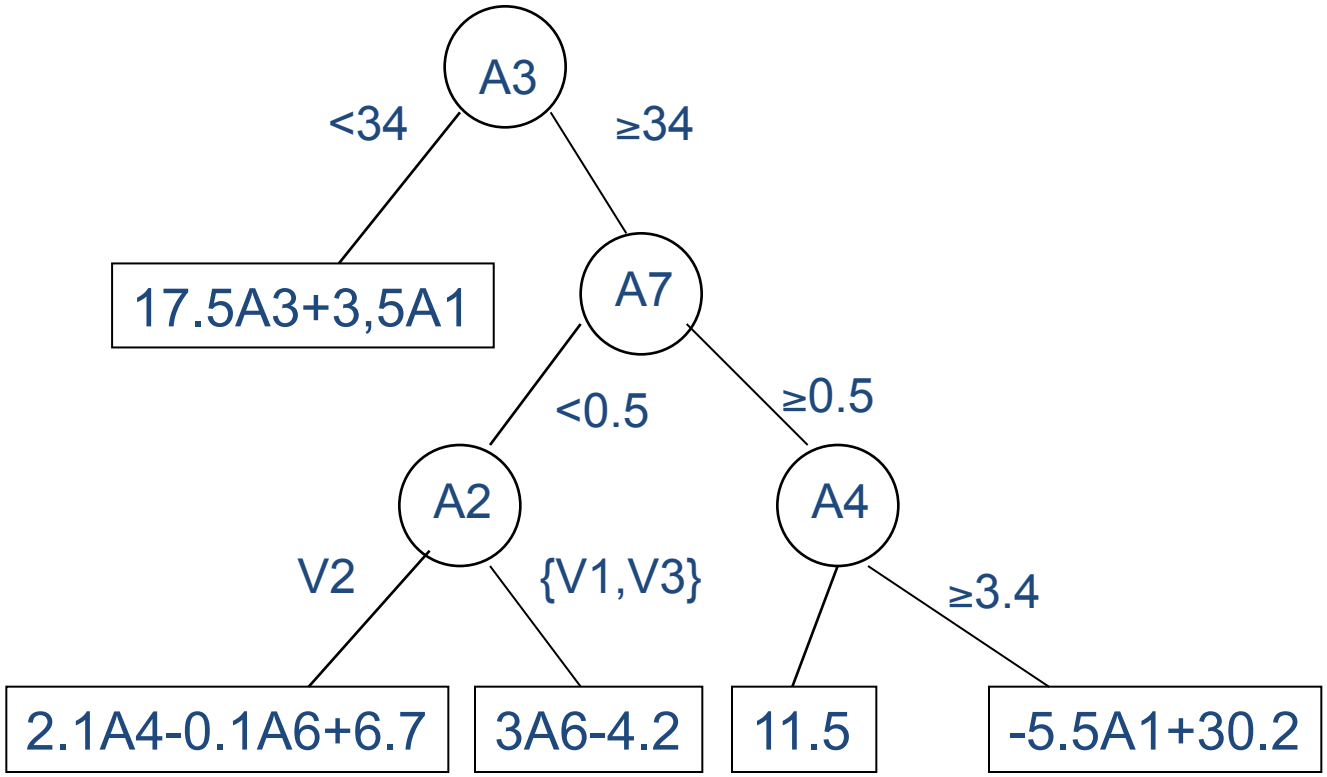
# Rozwój drzew

- Drzewa regresji wprowadzono w CART - (Breiman et al., 1984)
- Później model trees, od M5 (Quinlan, 1992)
- Wykorzystanie w zespołach klasyfikatorów (bagging, Random Forest) Breiman 1994, 2001
- Gradient boosted trees (Friedman 1999)
- Option trees (Buntine)
- Wersje przyrostowe dla strumieni danych (Ekomorovska)
- Multi-targer trees – wiele wyjść **y** (S.Dzeroski et al.)
- Oraz wiele innych – patrz artykuł przeglądowy Wei-Yin Loh: Fifty Years of Classification and Regression Trees, International Statistical Review (2014)

# Rozwinięcie do model trees

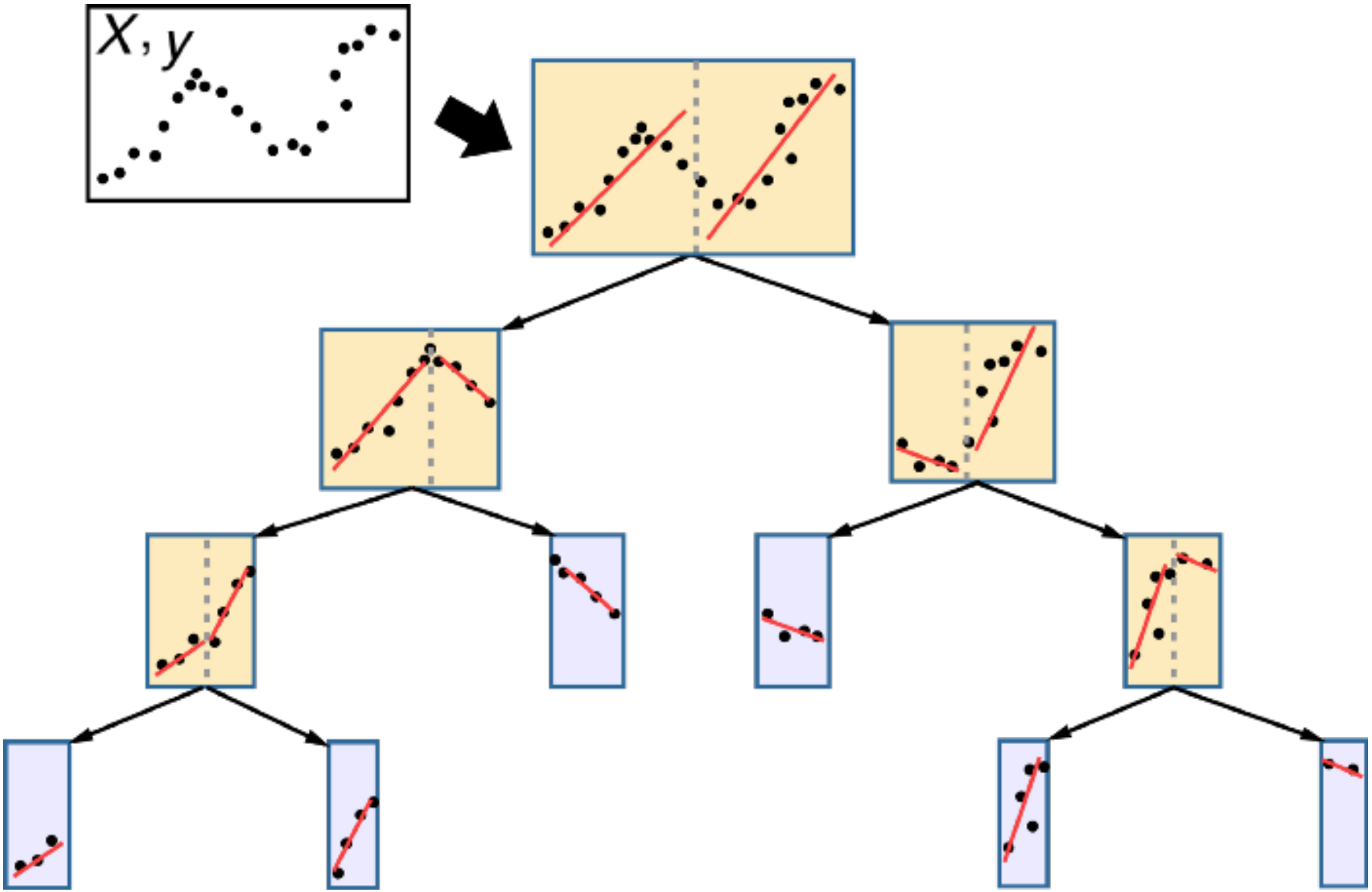
- W liściach wprowadzono odcinkami linowe funkcje regresji
- Efektywnie rozwiązane przez Quinlana w M5, rozwijane później, np. a stepwise linear regression model w węzłach (Kardic, Malerba,…)
- Torgo 1997 – zaproponował użycie regresji z funkcjami jądrowymi
- Oraz wiele innych − spójrz do literatury oraz wypróbuj oprogramowanie

# Model trees

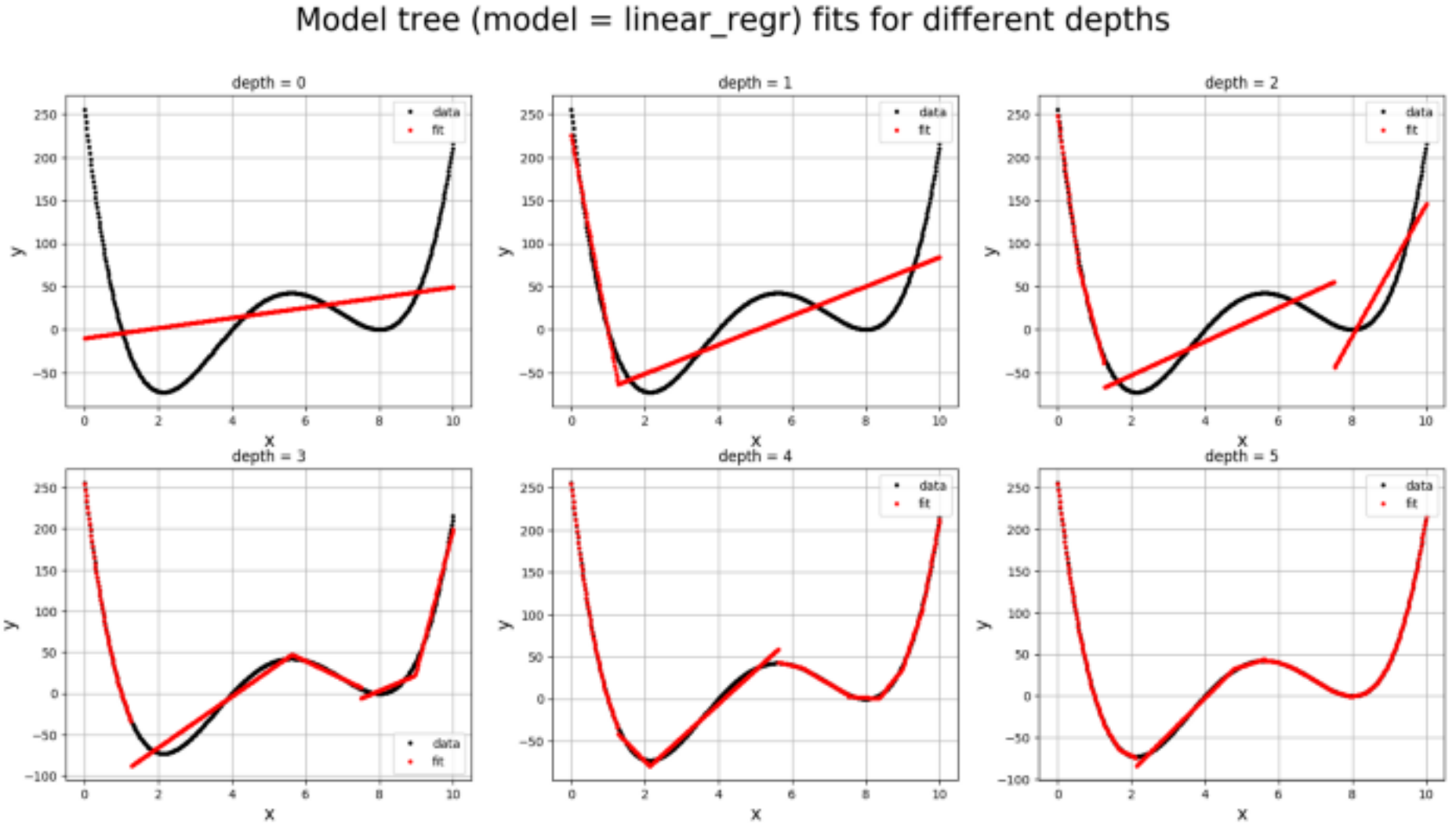



Jeśli w kostce obszaru związanej z liściem jest wystarczająca liczba przykładów uczących, to buduj lokalny model regresji liniowej

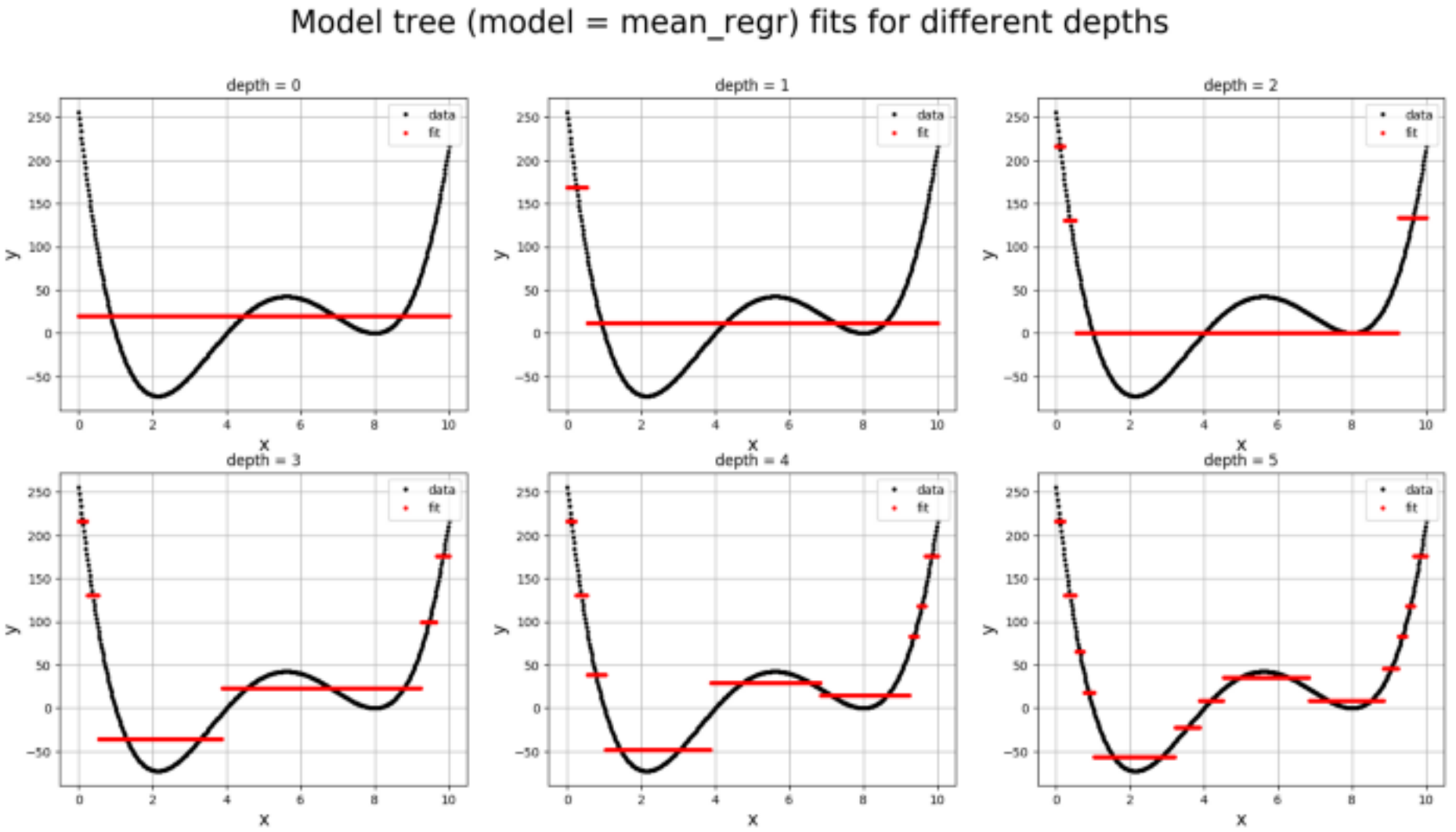
# Ilustracja działania model tree



# Przybliżenie modelem linowym funkcji nieliniowej na różnych poziomach drzewa (model tree)



# Przybliżenie modelem linowym funkcji nieliniowej na różnych poziomach drzewa regresji (średnia w kostce)

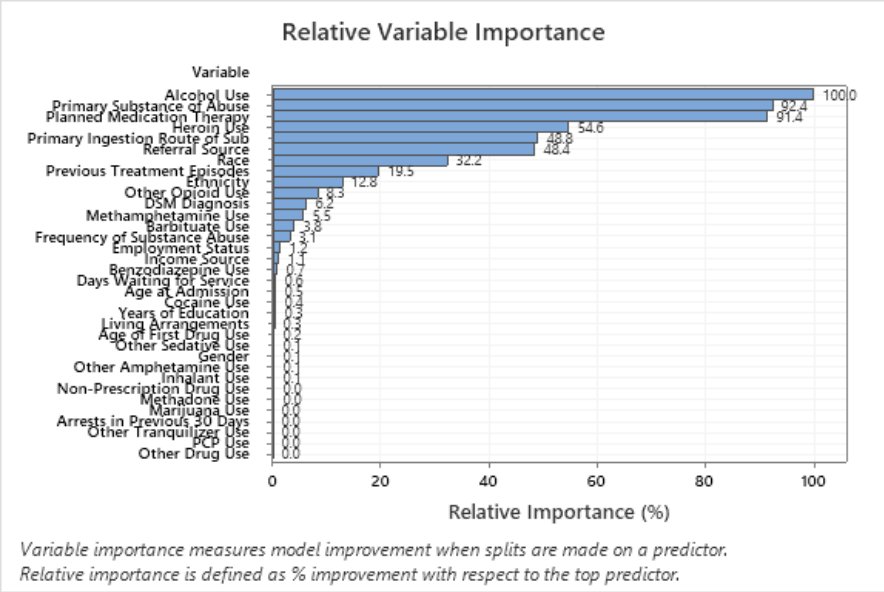


# Ilustracja przybliżenia funkcji

- Rysunki oraz szerszy opis intuicji budowy model trees z odcinkami liniowymi funkcjami regresji pochodzą z blogu pt. introduction to Model Trees from scratch | Anson Wong | na Towards Data Science − znajdź samodzielnie w internecie

# Interpretowalność drzewa i predykcji

- Możliwość interpretacji symbolicznej struktury drzewa – jeśli nie jest bardzo złożone
  - Podobnie jak dla drzew klasyfikacyjnych
- Ocena znaczenia najważniejszych cech dla predykcji – wykorzystanie propozycji tzw. feature importance zaproponowanej przez Briemana



# Dalsze pytania

- Złożone pytania w węzłach drzewa − lepsze przybliżenie skomplikowanych funkcji
- Lecz koszty obliczeniowe
- Otwarte pytania na dalszy wykład:
  - Jak wykorzystać drzewa regresji i model trees w zespołach klasyfikatorów
  - Inne funkcje straty *L* do optymalizacji i lepsze dopasowanie drzewa do danych, zwłaszcza w zespołach drzew regresyjnych (patrz np. gradient boosted trees)

# Odnośniki do literatury

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. Classification and Regression Tree. (1984).
- Rozdziały w książce Hastie, Tibushirani, Friedman: Elements of statistical learning (dostępna online pdf)
- Wei-Yin Loh: Fifty Years of Classification and Regression Trees, International Statistical Review (2014) − dobra lista cytowań do podstawowych prac oraz wielu rozszerzeń drzew
- Po polsku: M.Krzyśko, T.Górecki i inni, książka pt. Systemy uczące się
- E.Gatnar: Nieparametryczna metoda dyskryminacji i regresji
- J.Koronacki: Statystyczne systemu uczące się

Przykład analizy Boston house w Python

https://quantdev.ssri.psu.edu/sites/qdev/files/07_Trees_2017_1125.html

# Pytanie i komentarze?

Dalszy kontakt:

jerzy.stefanowski@cs.put.poznan.pl

http://www.cs.put.poznan.pl/jstefanowski/

Fundusze Europejskie Polska Cyfrowa

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego