České vysoké učení technické v Praze

Fakulta elektrotechnická

Katedra kybernetiky
Katedra počítačů

Vytěžování dat – přednáška 11

Kombinování modelů (ensembling)

Osnova přednášky

- Netflix prize
- Teoretické aspekty kombinování modelů
  - Skupina modelů
  - Různorodost modelů
  - Dekompozice chyb (rozptyl dat/systematická chyba modelu)
  - Jaké modely kombinovat?
- Populární ensemble metody
  - Bagging
  - Boosting
  - Stacking
- Algoritmus, co vyhrál milión $

Netflix Prize

- Vyrobte mi klasifikátor, který je o 10% lepší než to co mám, a dostanete milion dolarů!
- http://www.netflixprize.com/

- Učení s učitelem
  - Trénovací data jsou množiny respondentů a jejich hodnocení (1,2,3,4,5 hvězdiček), které udělili jednotlivým filmům.
  - Vytvoř klasifikátor, který dostane na vstup identifikaci uživatele a jím ještě nehodnocený film a vyprodukuje počet hvězdiček.

A kdo tedy vyhrál?

Leaderboard

Display top 20 leaders.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelay Industries !	0.8579	9.83	2009-07-26 02:49:53
6	PragmaticTheory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dace	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8633	9.26	2009-07-21 02:04:40
14	Gravity	0.8634	9.25	2009-07-26 15:58:34
15	Ces	0.8642	9.17	2009-07-25 17:42:38

COMPUTATION INTELLIGENCE GROUP

IDA Intelligent Data Analysis RESEARCH GROUP

The Ensemble

- And lo, as if powered by gravity, Grand Prize Team and Vandelay Industries ! began to draw in more and more members. And Vandelay went on to join forces with Opera Solutions, and then Vandelay and Opera united with Grand Prize Team, and then ... and then ... well, things got so complex we decided just to call ourselves **The Ensemble**.
- http://www.the-ensemble.com/content/meet-team

- Ale i druhý vítěz BellKor's Pragmatic Chaos je vlastně ensemble!

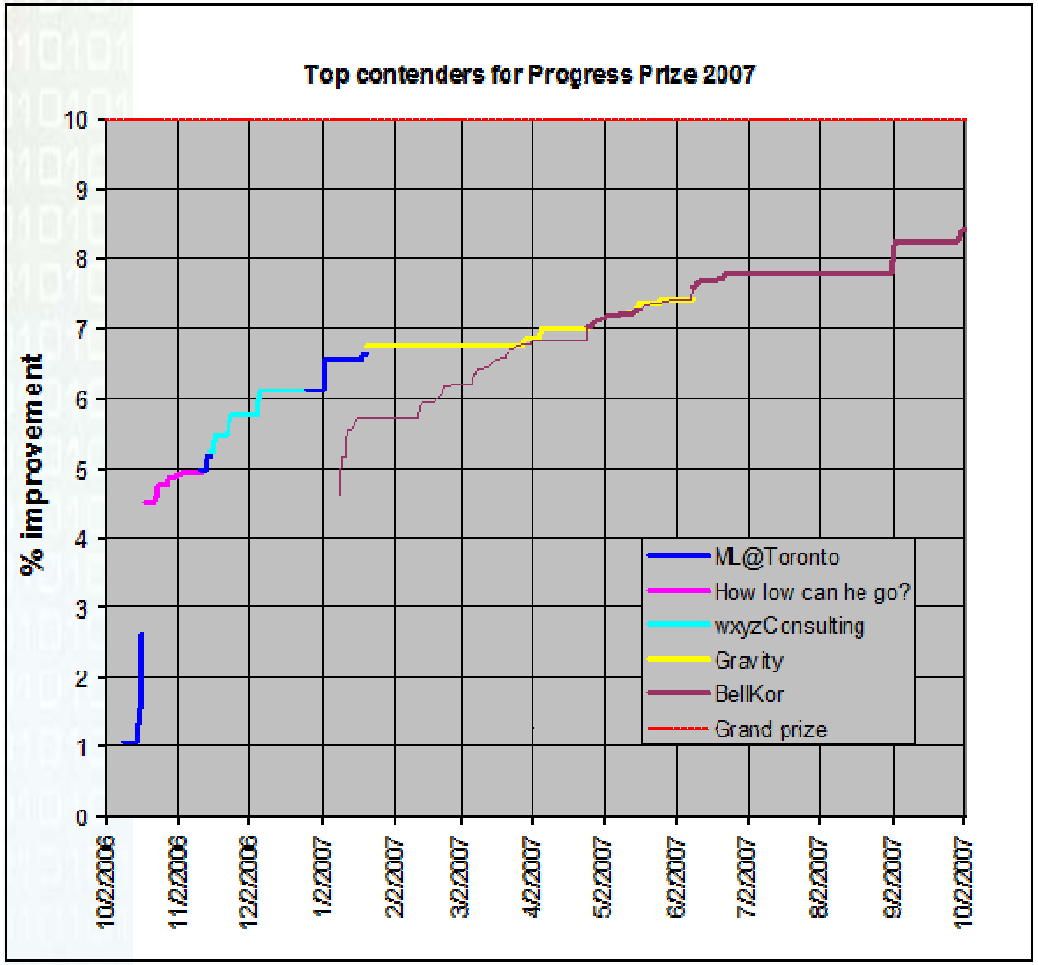
Takhle to vypadalo v květnu 2009

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
1	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
2	PragmaticTheory	0.8595	9.66	2009-05-19 13:23:12
3	Grand Prize Team	0.8597	9.64	2009-05-08 14:10:33
4	Dace	0.8604	9.56	2009-04-22 05:57:03
5	BigChaos	0.8613	9.47	2009-05-09 21:06:09
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
6	BellKor	0.8621	9.39	2009-05-10 09:39:48
7	Gravity	0.8634	9.25	2009-04-22 18:31:32
8	Opera Solutions	0.8642	9.17	2009-05-18 05:53:59
9	Ces	0.8642	9.17	2009-05-14 02:13:41
10	majia2	0.8642	9.17	2009-05-18 09:29:55
11	BruceDengDaoCiYiYou	0.8642	9.17	2009-05-18 09:59:21
12	Feeds2	0.8649	9.09	2009-05-17 10:47:21
13	Just a guy in a garage	0.8651	9.07	2009-05-19 18:13:50
14	pengpengzhou	0.8654	9.04	2009-05-05 18:18:03
15	NewNetflixTeam	0.8657	9.01	2009-05-13 01:01:36
16	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
17	Vandelay Industries !	0.8658	9.00	2009-05-11 00:43:14
18	acmehill	0.8659	8.99	2009-04-16 06:29:35
19	MonteCarlo	0.8661	8.97	2009-03-25 15:00:05
20	IDEA2	0.8661	8.97	2009-03-25 15:37:59
21	Newman, George, and Peterman !	0.8665	8.92	2009-05-07 05:04:09
22	xiangliang	0.8665	8.92	2009-05-19 14:19:10
23	Team ESP	0.8666	8.91	2009-05-19 14:33:15
24	My Brain and His Chain	0.8668	8.89	2008-09-30 02:19:47
25	xlvector	0.8668	8.89	2009-05-19 07:33:38

COMPUTATIONAL INTELLIGENCE GROUP

IDA Intelligent Data Analysis RESEARCH GROUP

Vývoj soutěže



--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

Rookies

"Thanks to Paul Harrison's collaboration, a simple mix of our solutions improved our result from 6.31 to 6.75"

--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

Arek Paterek

"My approach is to **combine the results of many methods** (also two-way interactions between them) using linear regression on the test set. The best method in my ensemble is regularized SVD with biases, post processed with kernel ridge regression"

http://rainbow.mimuw.edu.pl/~ap/ap_kdd.pdf

--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

U of Toronto

"When the predictions of **multiple** RBM models and **multiple** SVD models are linearly combined, we achieve an error rate that is well over 6% better than the score of Netflix's own system."

http://www.cs.toronto.edu/~rsalakhu/papers/rbmcf.pdf

--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

A co „tahoun“ celé soutěže?

BellKor / KorBell

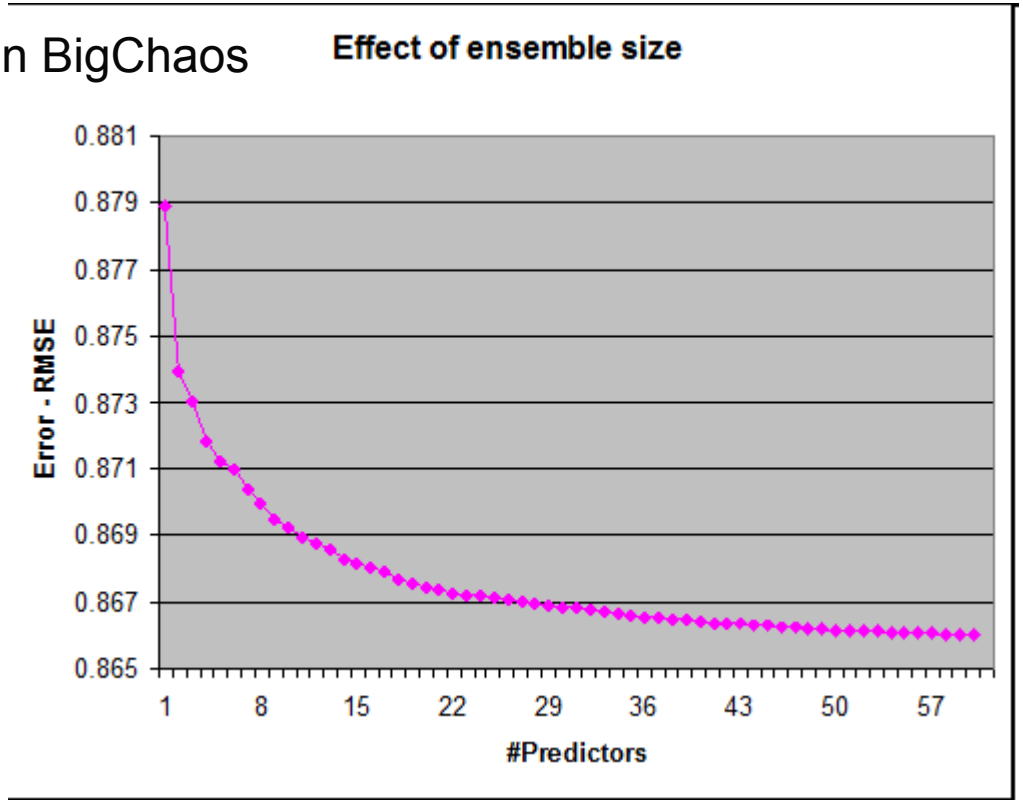
“Our final solution (RMSE=0.8712) consists of blending 107 individual results. “

„blendingem“ rozuměj kombinování modelů

--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

Chyba klesá s počtem modelů

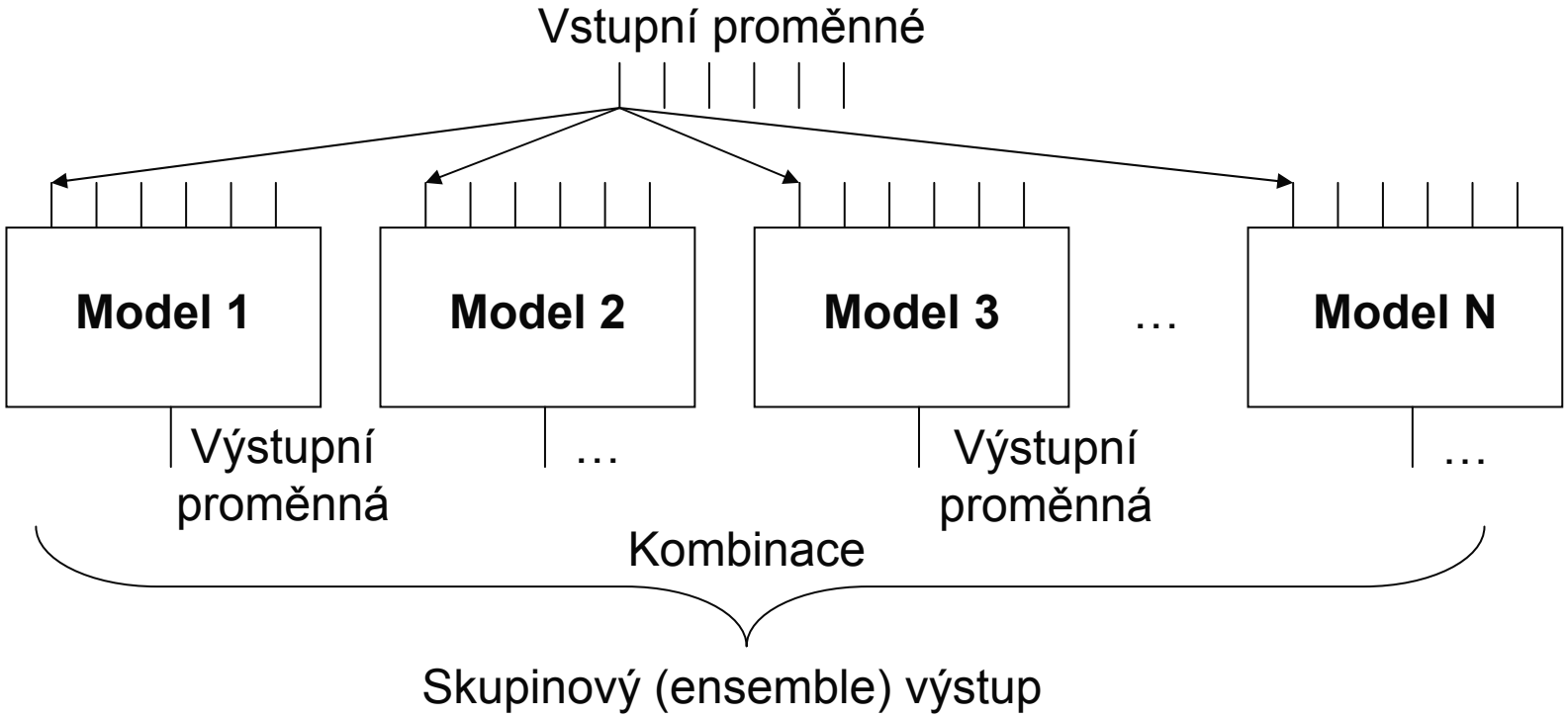
- BellKor in BigChaos



- Jak a jaké modely kombinovat?

Princip kombinování modelů

- Skupina modelů (např. rozhodovacích stromů) se naučí na stejný (podobný) úkol.
- Výstupy naučených modelů se kombinují.



Různorodost ensemble modelů

- Co se stane, když budou všechny modely totožné?
  => Degradace na jeden model.
- Jak zajistíme, aby byly modely různorodé?
  - Různé množiny trénovacích dat (počáteční podmínky)
  - Různé metody konstrukce modelů
- Jak se dá měřit různorodost modelů?
  - Odchylky výstupů na jednotlivých testovacích datech.
  - Strukturální odlišnosti

Funguje to vůbec?

- Chceme zjistit, za jakých předpokladů se vyplatí modely kombinovat.
- Zajímá nás proč ensembling funguje.

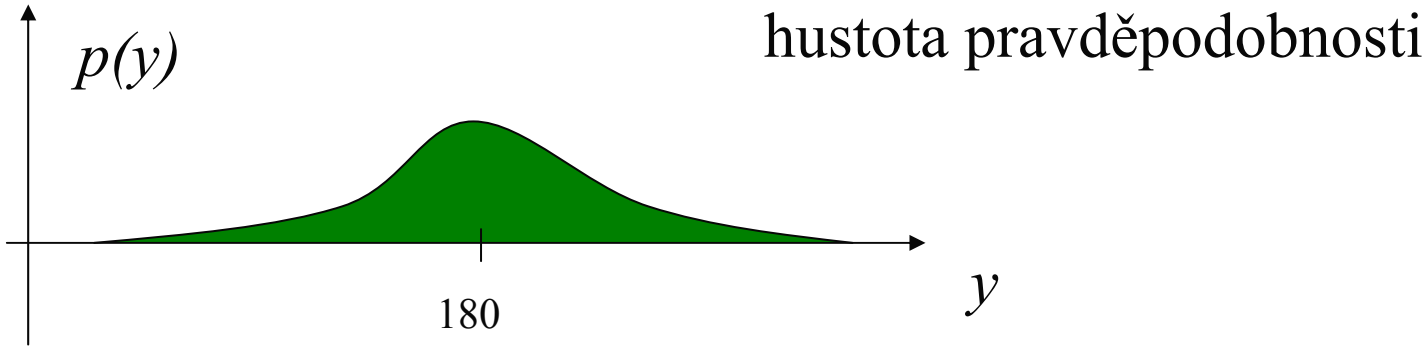
- Potřebujeme k tomu analyzovat, čím je chyba modelů způsobena

Příklad – výška lidí

Cíl: odhadnout výšku dospělého muže.

Přesněji:

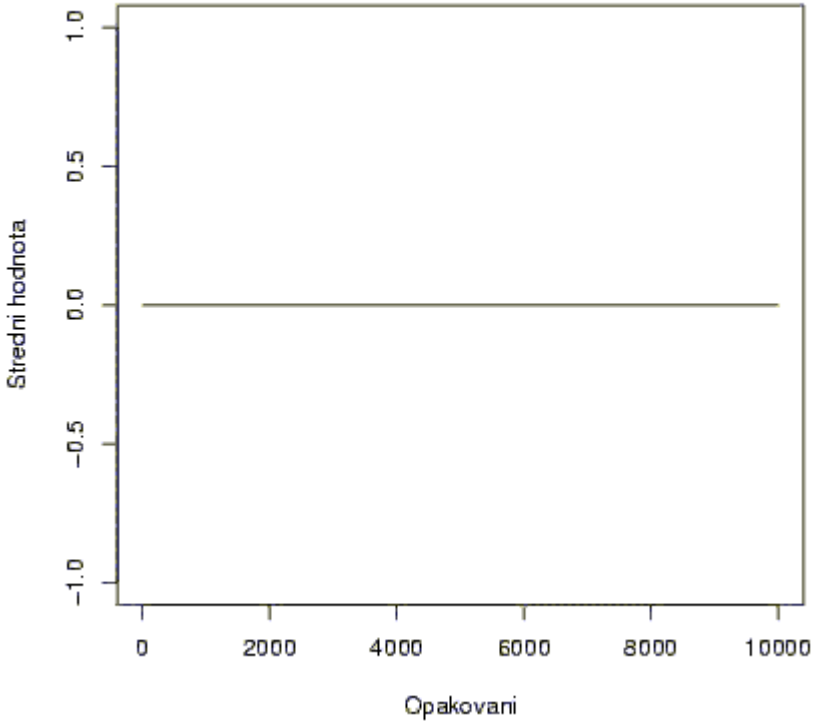
- Najít odhad $\hat{y}$ který minimalizuje střední hodnotu chyby $E_y\{(y-\hat{y})^2\}$ přes celou populaci.



Co to je ta střední hodnota?

Jednoduché metody výpočtu:
Průměr: je aritmetický průměr: 2, 3, 3, 5, 7 a 10 je 30 děleno 6, což je 5.
Medián: je prostřední číslo ve skupině čísel, kdy má polovina čísel hodnotu vyšší než medián a polovina čísel hodnotu nižší než medián.
Medián čísel 2, 3, 3, 5, 7 a 10 je 4.

Animace:
generujeme náhodná
čísla <-0.5,0.5>
a počítáme průměr – ten se
limitně blíží k nule



Jak by na to šel pan Bayes?

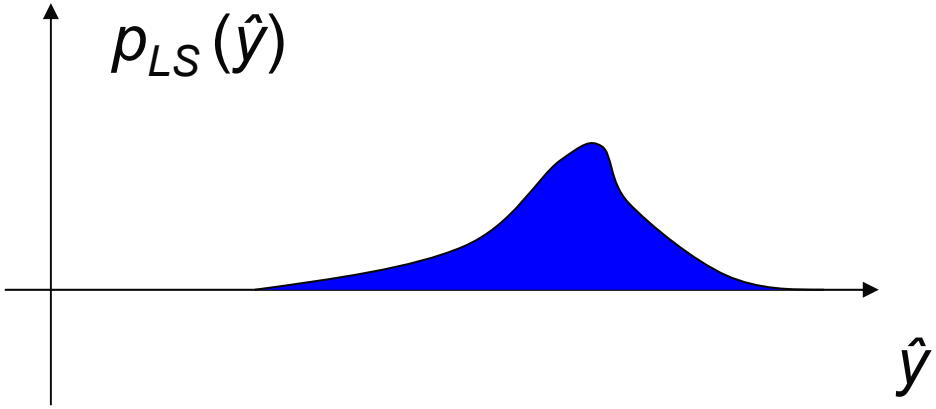
- Změřil by všechny dospělé muže v celé populaci, vypočetl střední hodnotu $\mathrm{E}_y\{y\}$ a prohlásil ji za optimální odhad.
- Potlesk, úloha vyřešena, výsledek je:

$$\mathrm{E}_y\{y\} = \sum_{y=-\infty}^{\infty} p(y)y$$

- Ale v reálu je $p(y)$ neznámá, její odhad je třeba zkonstruovat pomocí skupiny mužů $LS=\{y_1,y_2,\ldots,y_N\}$, jejichž výšku jsme změřili.

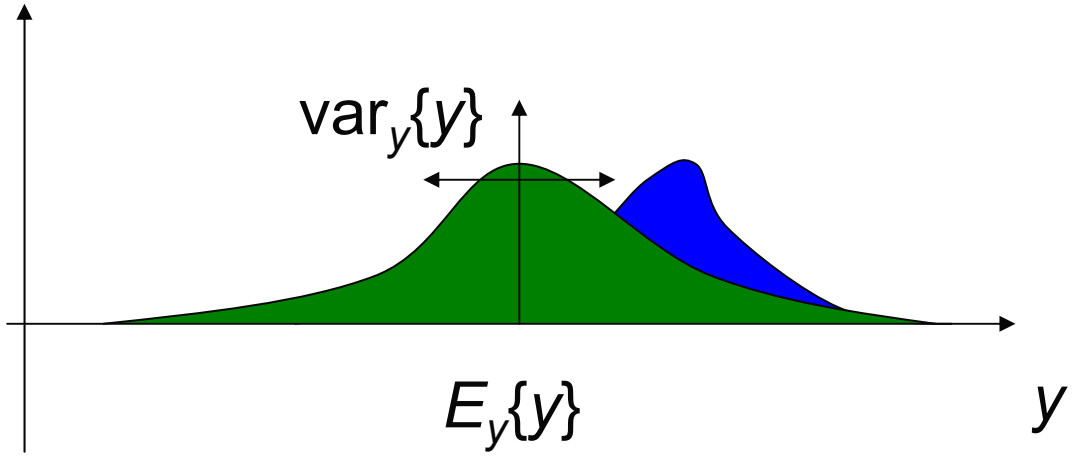
Výstup nějakého našeho modelu

- Protože skupina *LS* je náhodně zvolena, odhad $\hat{y}$ bude také náhodná veličina.



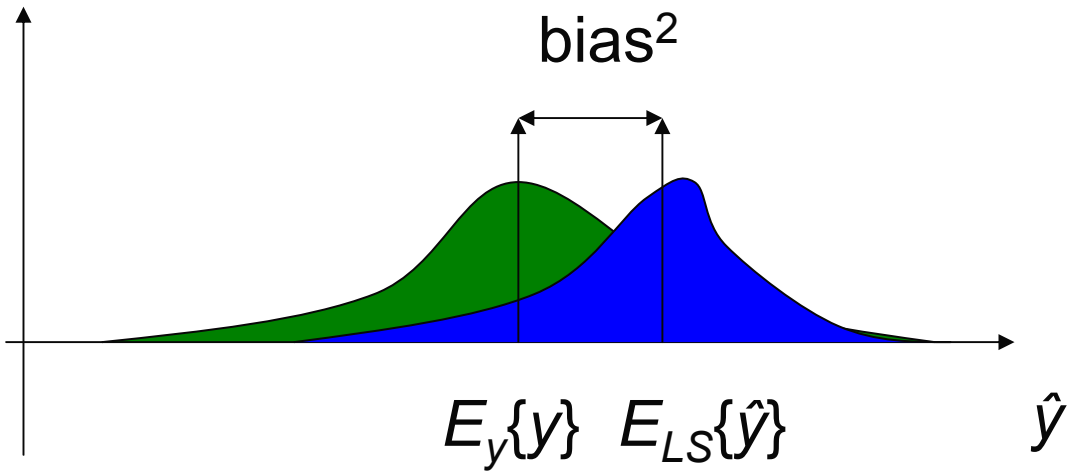
- Dobrý učicí algoritmus by měl fungovat dobře na libovolně zvolené skupině *LS*, tedy poskytnout minimální chybu v průměru pro různé *LS*: $E=E_{LS}\{E_y\{(y-\hat{y})^2\}\}$

Dekompozice bias/variance (1)



$$E = \underbrace{E_y\{(y - E_y\{y\})^2\}}_{\text{= residuální chyba = minimální dosažitelná chyba} = \mathrm{var}_y\{y\}} + E_{LS}\{(E_y\{y\} - \hat{y})^2\}$$

Dekompozice bias/variance (2)



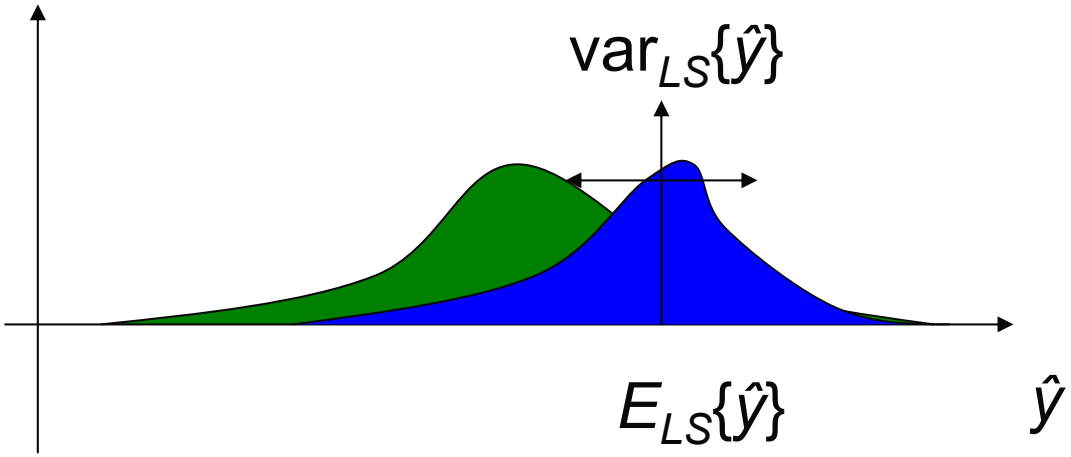
$$E = \mathrm{var}_y\{y\} + (E_y\{y\} - E_{LS}\{\hat{y}\})^2 + \ldots$$

$E_{LS}\{\hat{y}\}$ = průměrný model (přes všechny *LS*)

bias² = chyba průměrného modelu oproti modelu pana Bayese (optimálnímu)

Náš algoritmus učení

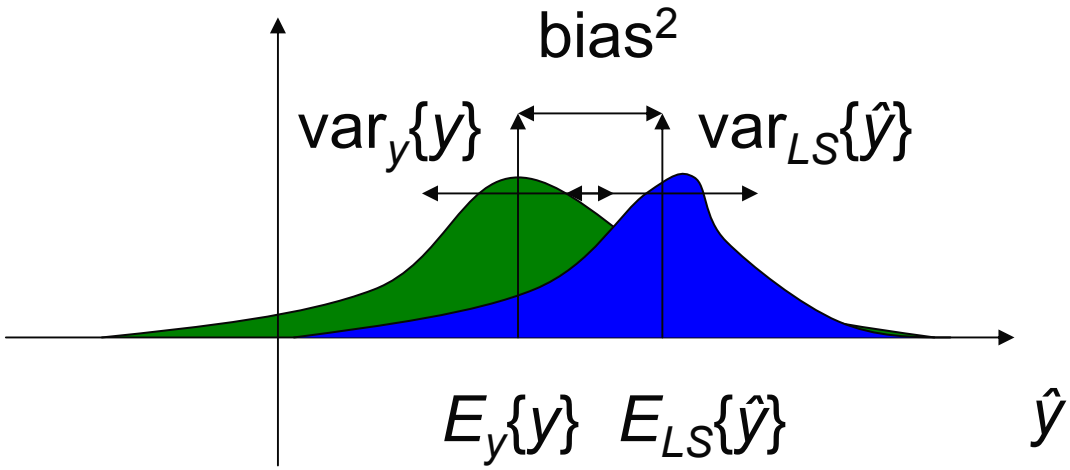
Dekompozice bias/variance (3)



$$E = \mathrm{var}_y\{y\} + \mathrm{bias}^2 + \boldsymbol{E_{LS}\{(\hat{y} - E_{LS}\{\hat{y}\})^2\}}$$

$\mathrm{var}_{LS}\{\hat{y}\}$ = variance našich modelů pro různé LS = důsledek přeučení

Dekompozice bias/variance (4)



$$E = \text{var}_y\{y\} + \text{bias}^2 + \text{var}_{LS}\{\hat{y}\}$$

Dekompozice bias/variance (5)

$E_{LS}\{E_{y/\underline{x}}\{(y-\hat{y}(\underline{x}))^2\}\}$=Noise($\underline{x}$)+Bias^2($\underline{x}$)+Variance($\underline{x}$)

- **Šum(x)** = $E_{y/\underline{x}}\{(y-h_B(\underline{x}))^2\}$

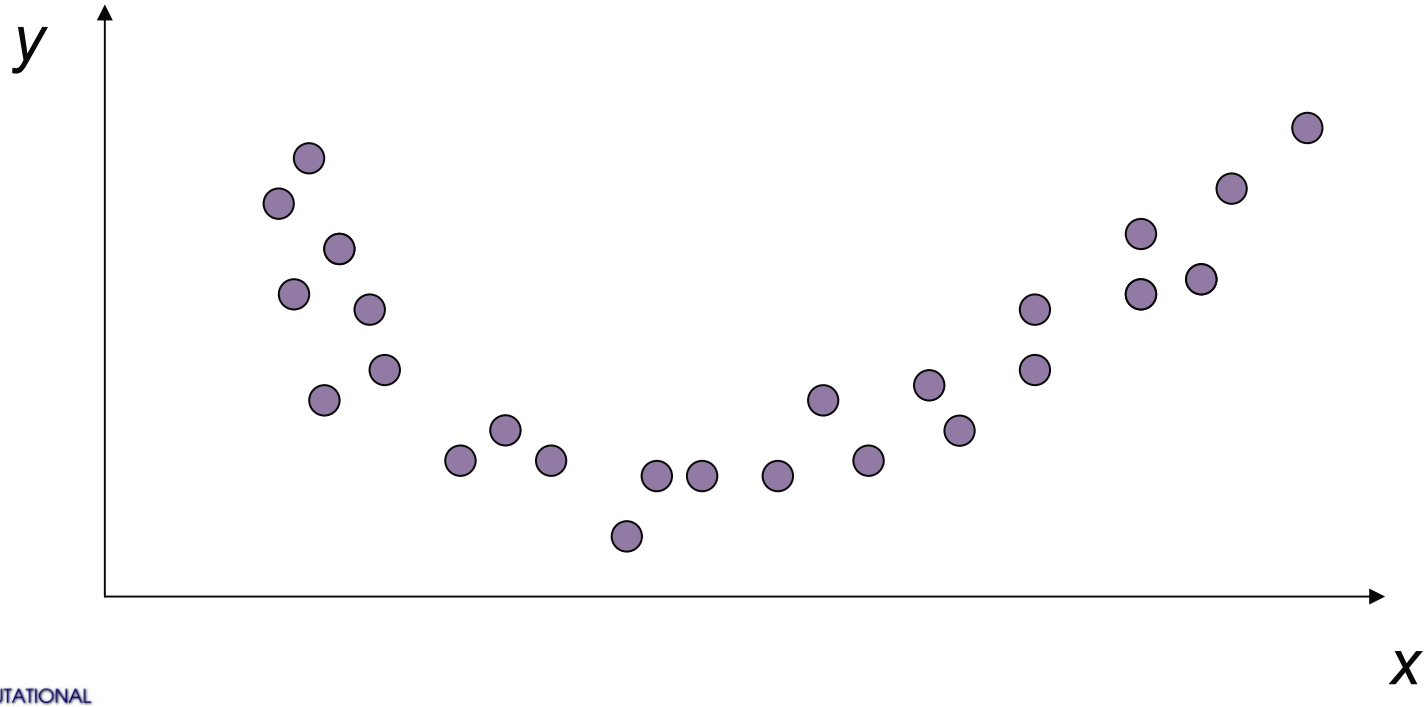
  Kvantifikuje odchylku výstupu y od optimálního modelu $h_B(\underline{x}) = E_{y/\underline{x}}\{y\}$.
- **Bias^2(x)** = $(h_B(\underline{x})-E_{LS}\{\hat{y}(\underline{x})\})^2$:

  Chyba průměrného modelu vzhledem k optimálnímu.
- **Variance(x)** = $E_{LS}\{(\hat{y}(\underline{x})-E_{LS}\{\hat{y}(\underline{x})\})^2\}$:

  Jak moc se predikce $\hat{y}(\underline{x})$ liší pro různé učicí množiny *LS*.

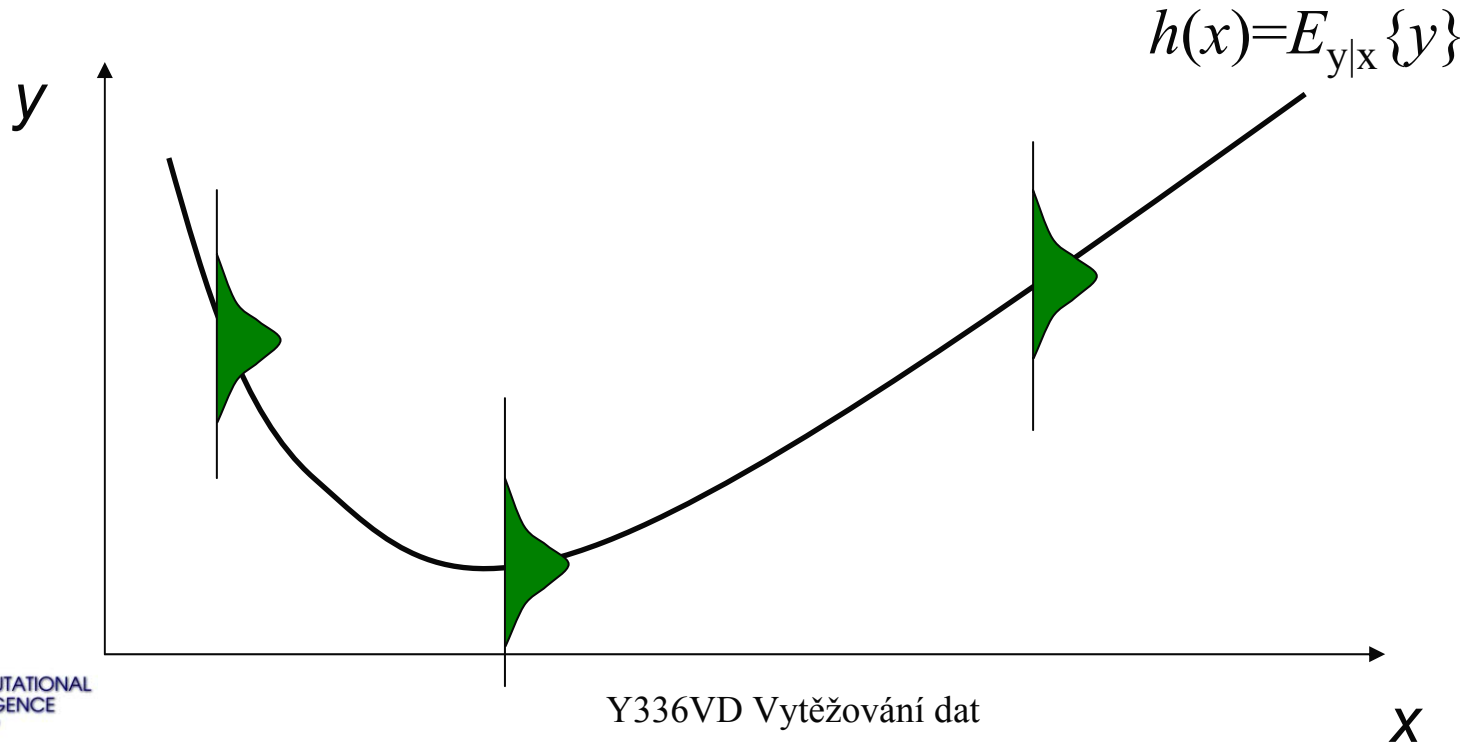
Příklad (1)

- Najděte algoritmus produkující co nejlepší modely pro následující data:



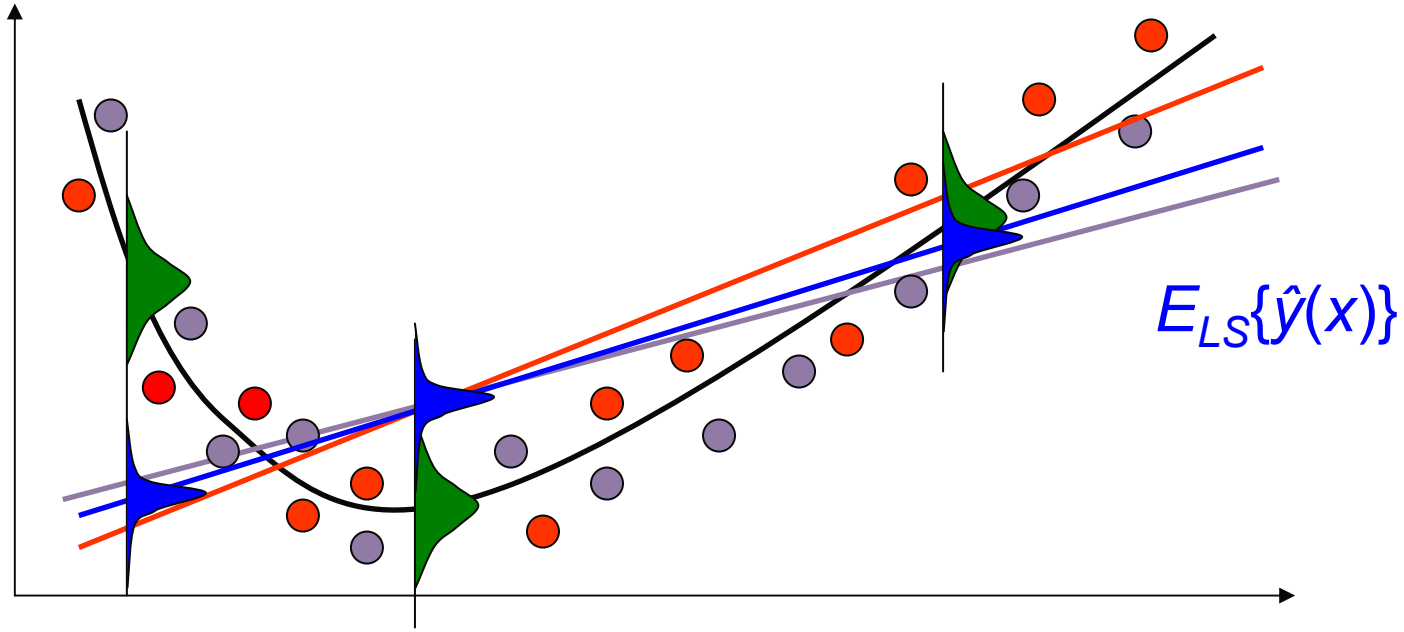
Příklad (2)

- Optimální model:
  - Vstup *x*, náhodná proměnná rovnoměrně rozložená v intervalu [0,1]
  - $y=h(x)+\varepsilon$, kde $\varepsilon \sim N(0,1)$ je Bayesovský model y a šum



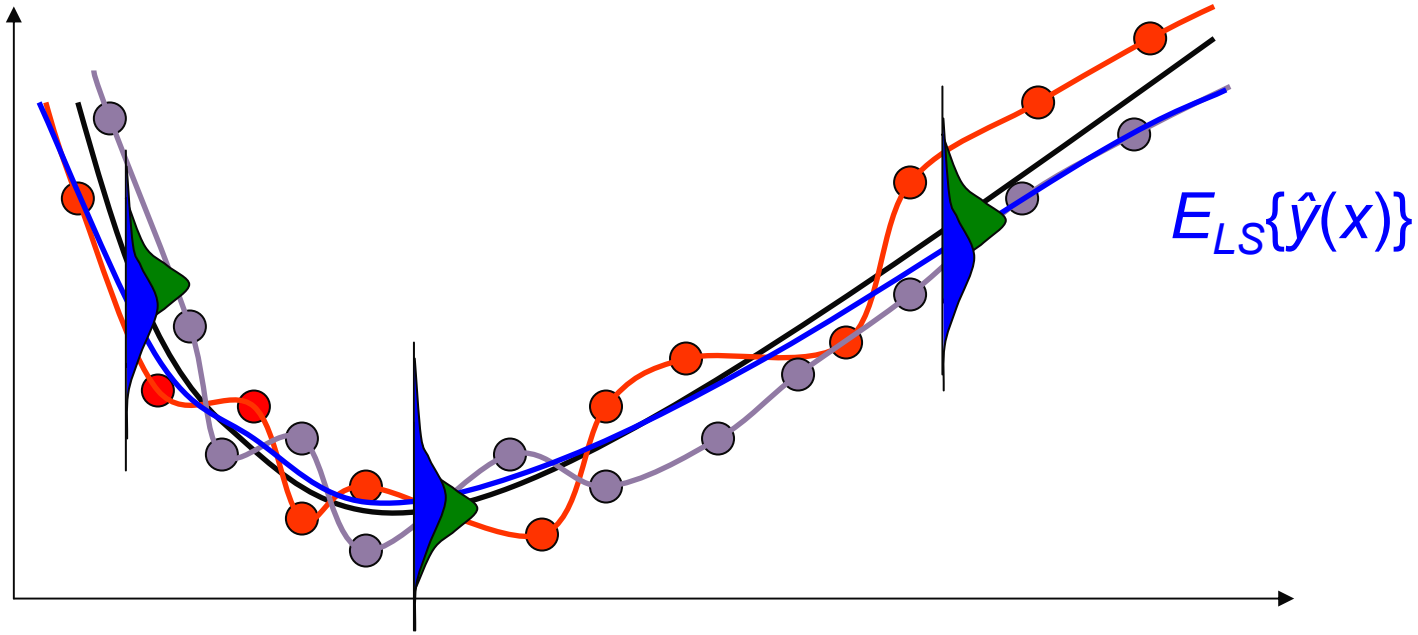
Příklad – algoritmus lineární regrese

- Modely mají malý rozptyl (variance), ale velké zaujetí (bias) $\Rightarrow$ nedoučení



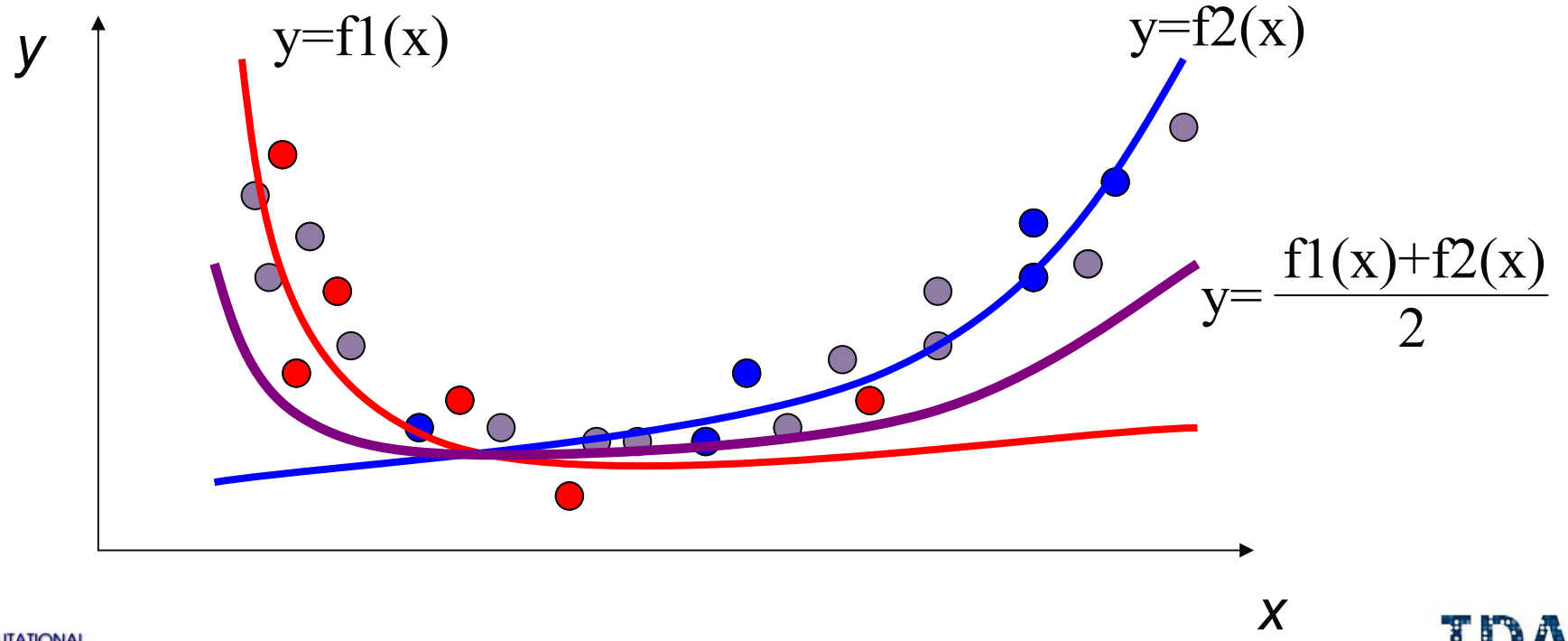
Příklad – algoritmus RBFN s počtem neuronů = mohutnost LS

- Nízké zaujetí (bias), velký rozptyl (variance) modelů $\Rightarrow$ přeučení

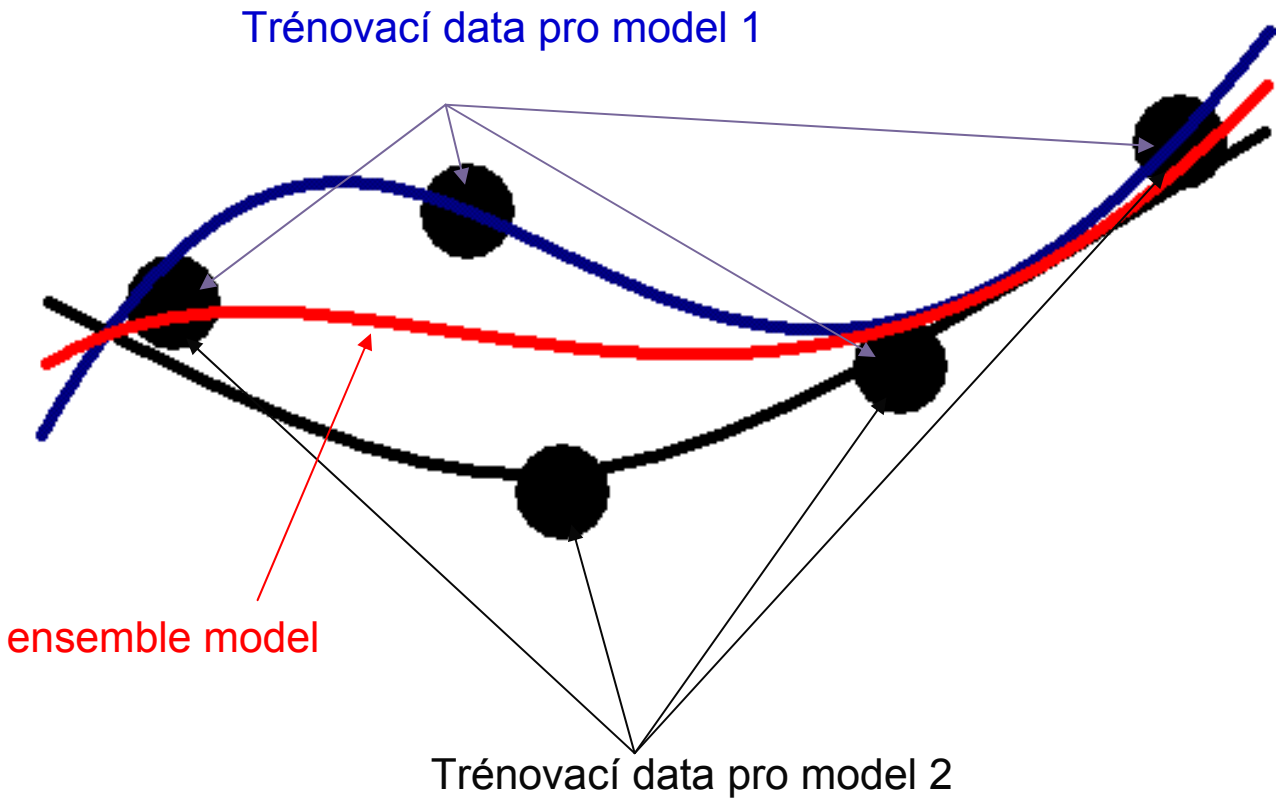


Zpět ke kombinování modelů

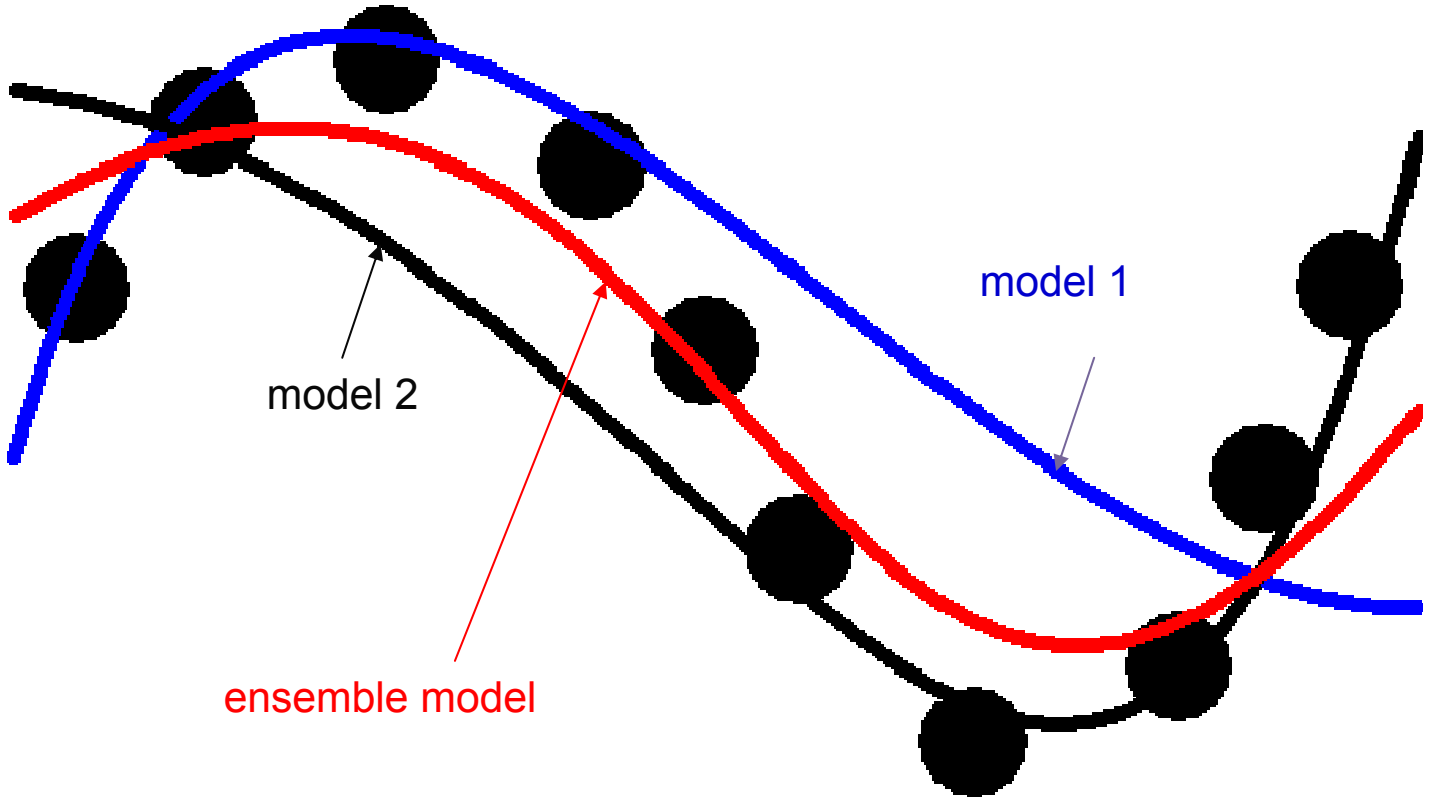
- Co se stane, když naučím 2 jednoduché modely na různých podmnožinách LS?



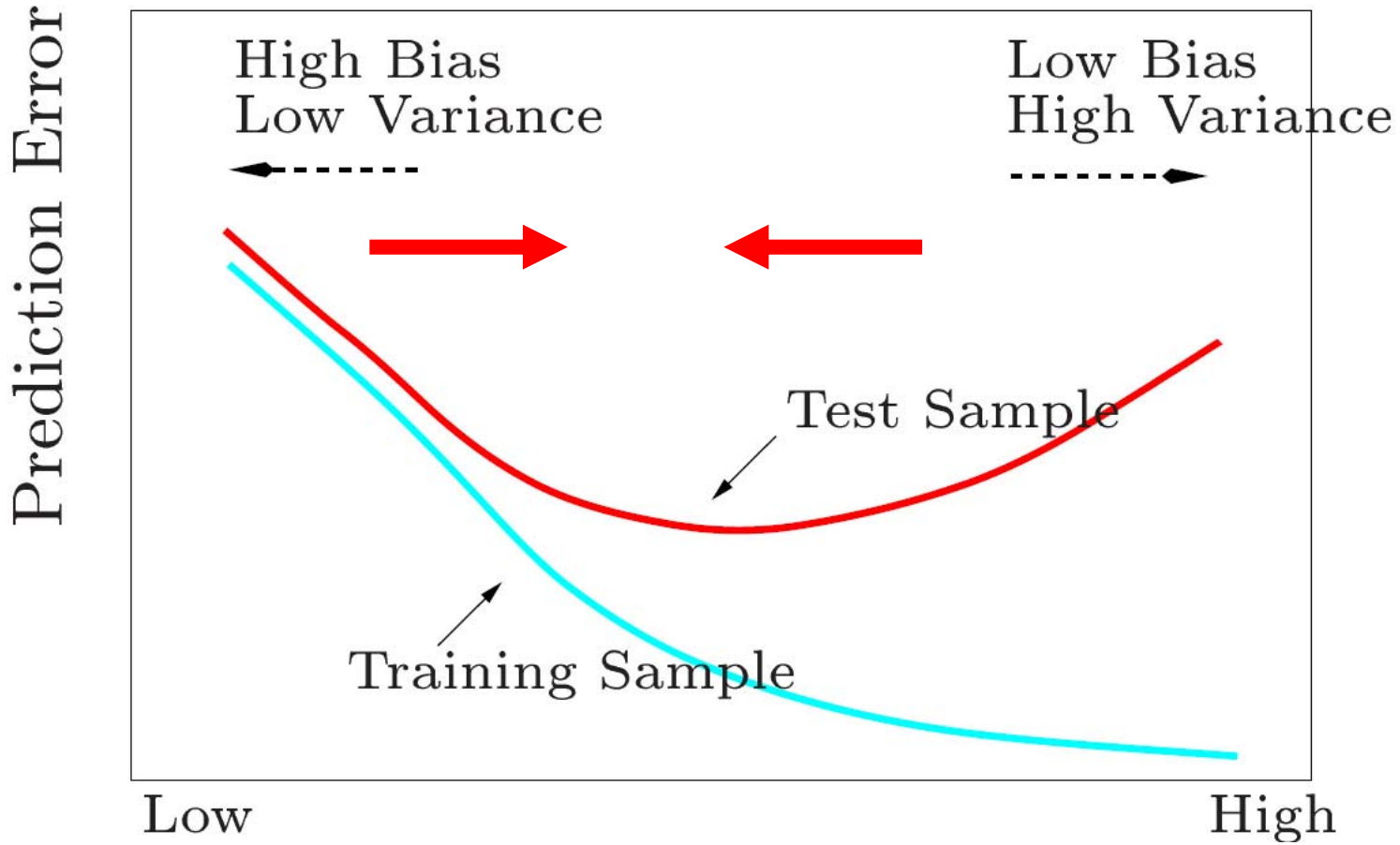
Ensembling snižuje **varianci**



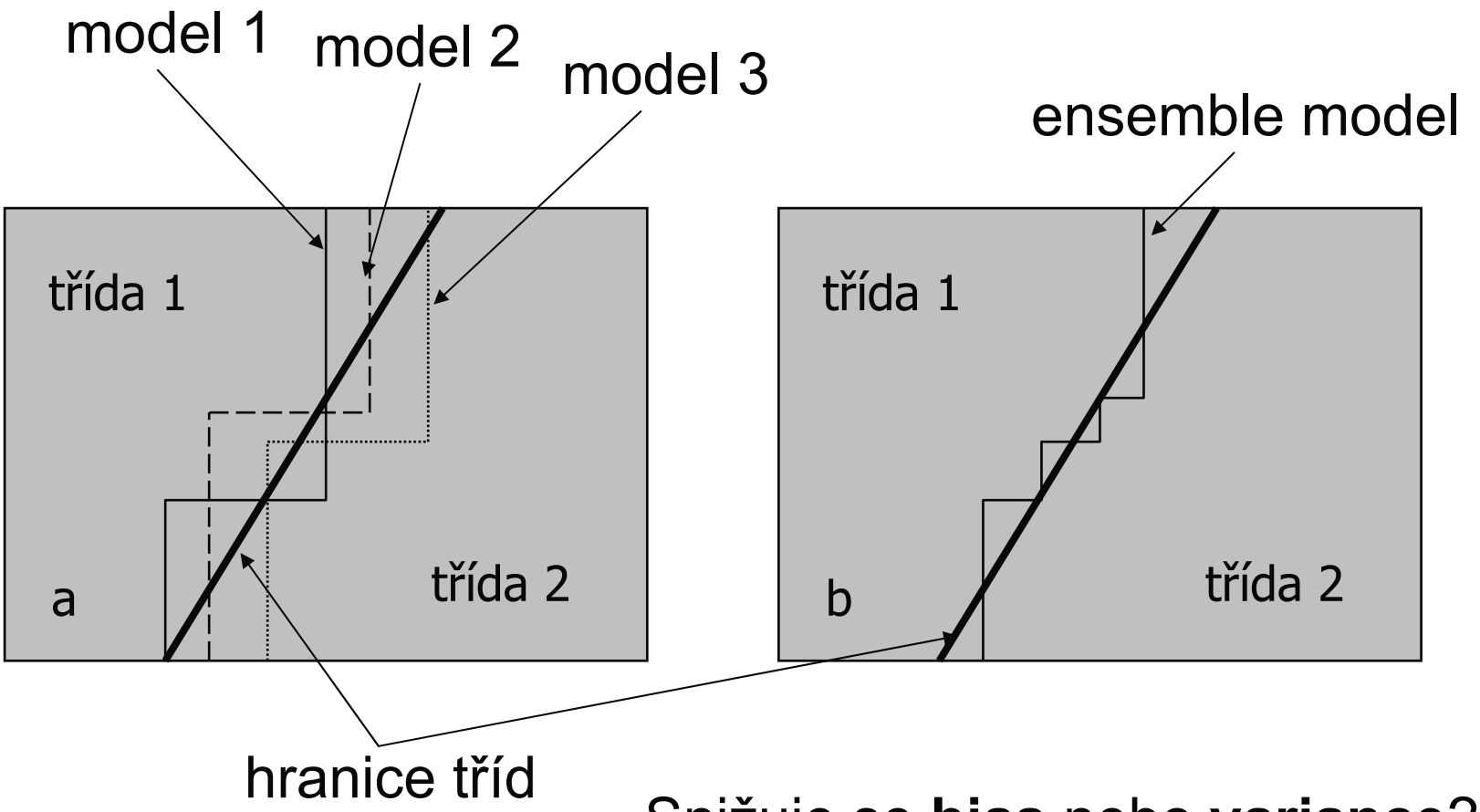
Ensembling snižuje **bias**



Tedy:


Prediction Error
High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Low
High

Podobně pro klasifikaci?

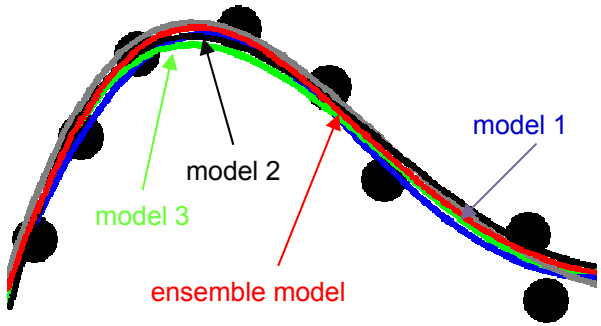


Snižuje se **bias** nebo **variance**?

Obrázky z TCD AI Course, 2005

Jaké modely kombinovat?

- Co se stane, když zkombinují optimálně naučené modely?



- Výhodně se dají kombinovat jednoduché modely (tzv. weak learners).
- Modely musí být různorodé! Musejí vykazovat různé chyby na jednotlivých trénovacích vzorech.

redukujeme **bias**

redukujeme **varianci**

Populární ensemble metody

- Bagging (Bootstrap Aggregating)
  - Modely naučím nezávisle a jednoduše zkombinuji jejich výstup
- Boosting
  - Modely se učí sekvenčně, trénovací data jsou závislá na chybách předchozích modelů
- Stacking
  - Modely se učí nezávisle, kombinují naučením speciálního modelu

Bagging (1)

$$E_{LS}\{Err(\underline{x})\} = E_{y/\underline{x}}\{(y - h_B(\underline{x}))^2\} + (h_B(\underline{x}) - E_{LS}\{\hat{y}(\underline{x})\})^2 + E_{LS}\{(\hat{y}(\underline{x}) - E_{LS}\{\hat{y}(\underline{x})\})^2\}$$

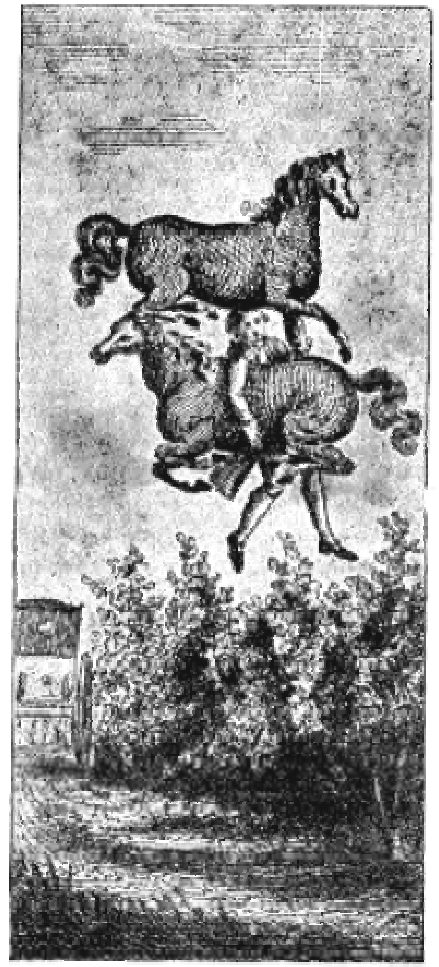
- **Myšlenka:** průměrný model $E_{LS}\{\hat{y}(\underline{x})\}$ má stejný bias jako původní metoda, ale nulovou varianci
- **Bagging** (**B**ootstrap **AGG**regat**ING**) :
  - K vypočtení $E_{LS}\{\hat{y}\ (\underline{x})\}$, potřebujeme nekonečně mnoho skupin *LS* (velikosti N)
  - Máme však jen jednu skupinu *LS,* musíme si sami nějak pomoc …

Bootstrapping: trocha historie

• Rudolf Raspe, *Baron Munchausen's Narrative of his Marvellous Travels and Campaings in Russia*, 1785
He hauls himself and his horse out of the mud by lifting himself by his own hair.

•This term was also used to refer to doing something on your own, without the use of external help since 1860s

• Since 1950s it refers to the procedure of getting a computer to start (to boot, to reboot)



The BARON helps his horses over the hedge.

Co je Bootstrap?

X=(3.12, 0, 1.57, 19.67, 0.22, 2.20)

Mean=4.46

X1=(1.57,0.22,19.67, 0,0,2.2,3.12)

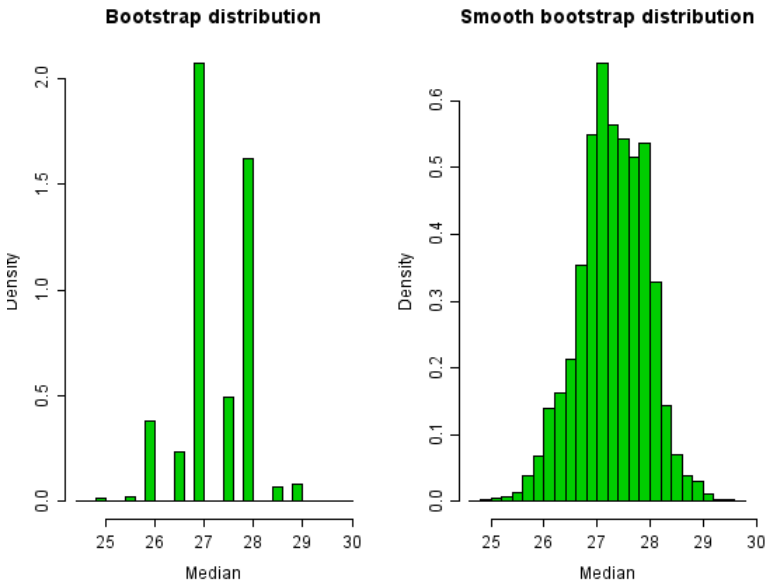
Mean=4.13

X2=(0, 2.20, 2.20, 2.20, 19.67, 1.57)

Mean=4.64

X3=(0.22, 3.12,1.57, 3.12, 2.20, 0.22)

Mean=1.74



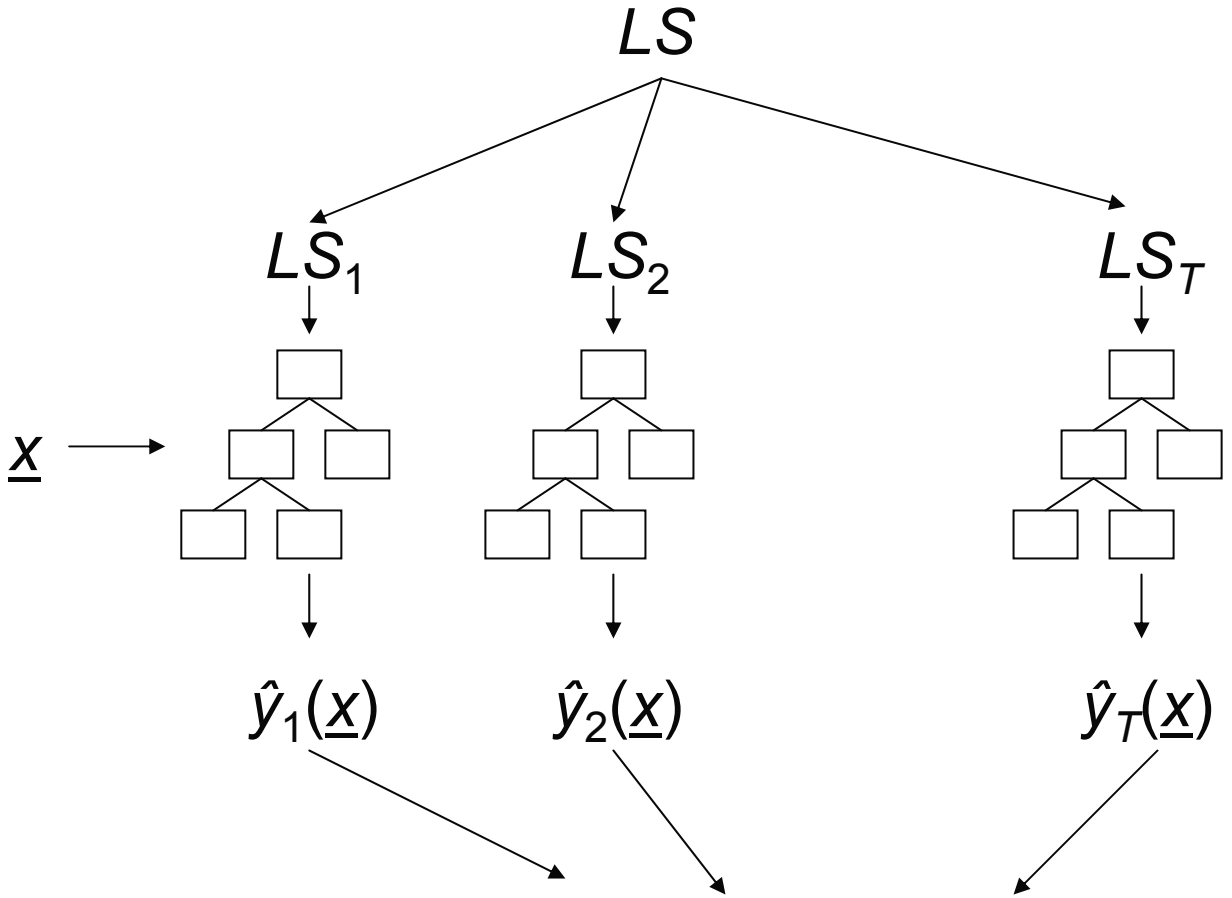
statistika:

– odhad intervalu spolehlivosti

Příklad bootstrap (Opitz, 1999)

Trénovací vzory	1	2	3	4	5	6	7	8
vzorek 1	2	7	8	3	7	6	3	1
vzorek 2	7	8	5	6	4	2	7	1
…	…	…	…	…	…	…	…	…
vzorek M	4	5	1	4	6	4	3	8

Bagging (2)

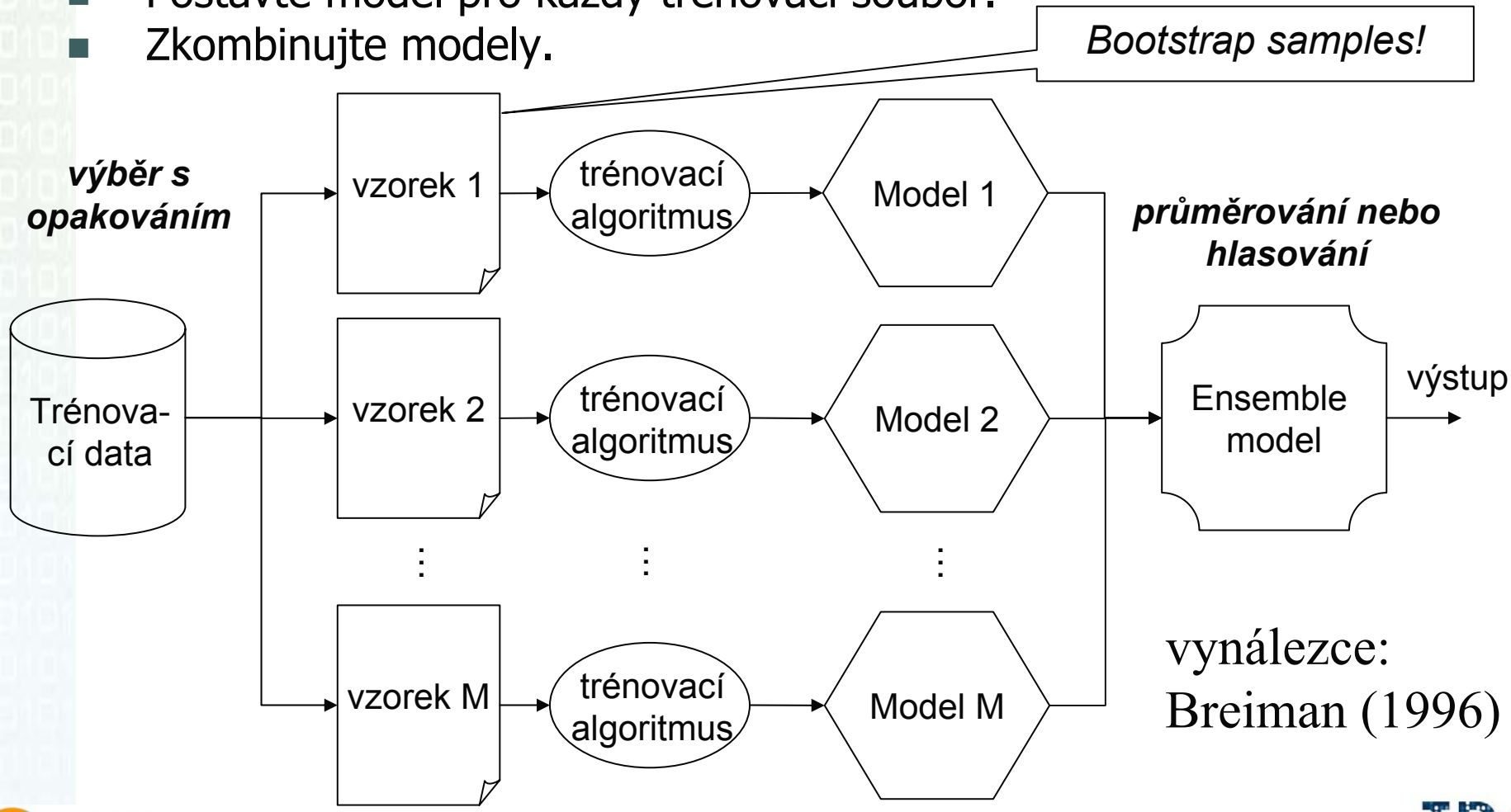


Pro regresi: $\hat{y}(\underline{x}) = 1/k * (\hat{y}_1(\underline{x})+\hat{y}_2(\underline{x})+\ldots+\hat{y}_T(\underline{x}))$

Pro klasifikaci: $\hat{y}(\underline{x})$ = majoritní třída z $\{\hat{y}_1(\underline{x}),\ldots,\hat{y}_T(\underline{x})\}$

Bagging (Bootstrap agregating)

- Výběrem s opakováním utvořte M trénovacích souborů o ***n*** vzorech (místo jednoho původního souboru o ***n*** vzorech).
- Postavte model pro každý trénovací soubor.
- Zkombinujte modely.



vynálezce: Breiman (1996)

Bagging – př. rozhodovací stromy

- Obvykle bagging podstatně sníží varianci a zachová zaujetí modelů.
- Podívejme se na příklad s rozhodovacími stromy:

Metoda	E	Bias	Variance
3 Test regr. Tree	14.8	11.1	3.7
Bagged (T=25)	11.7	10.7	1.0
Full regr. Tree	10.2	3.5	6.7
Bagged (T=25)	5.3	3.8	1.5

- V tomto případě jsou pro učení rozhodovacích stromů použity všechny vstupní atributy a diverzita je způsobena proměnnou učicí množinou

Náhodné (rozhodovací) lesy – random forests

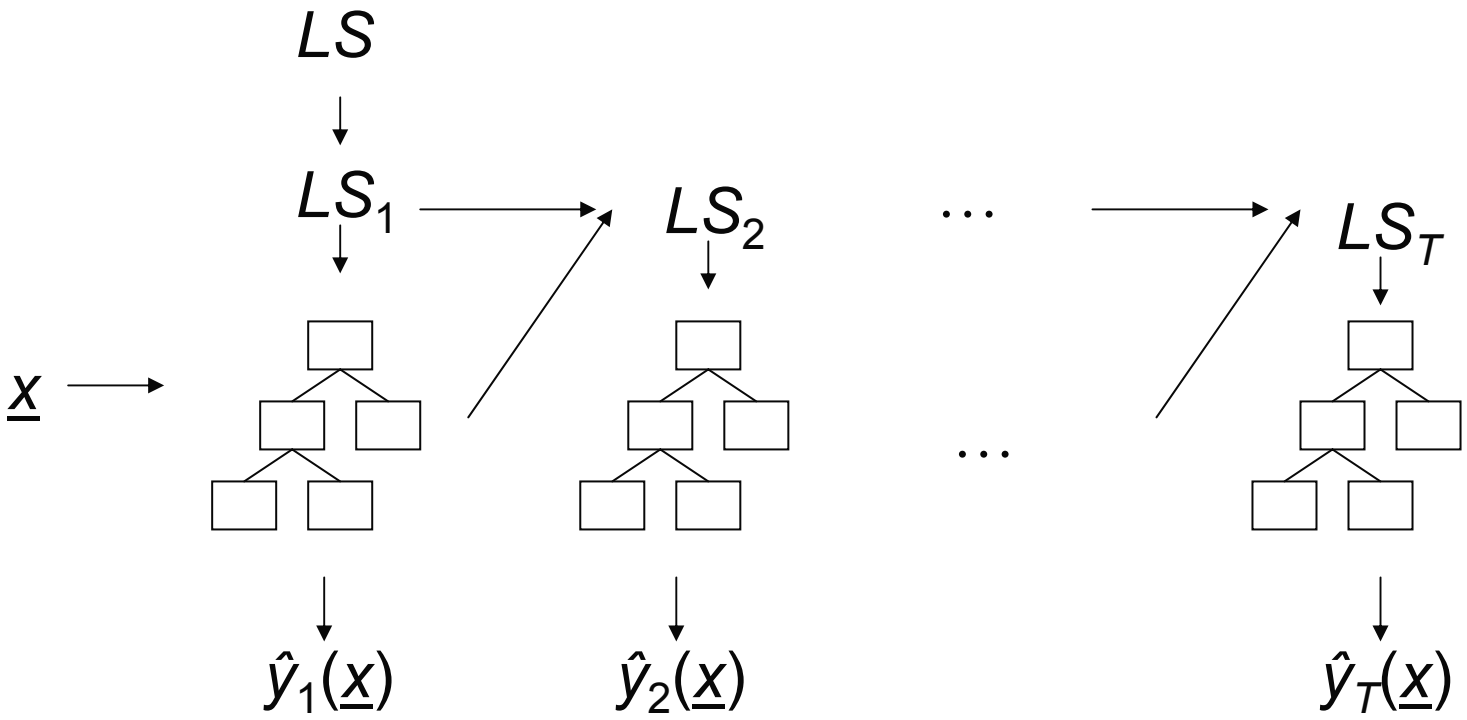
- K baggingu se ještě přidá to, že náhodně vybíráme podmnožinu vstupních atributů
- Tedy:
  - Postav rozhodovací strom z bootstrap vzorku
  - Najdi „best split" mezi náhodnou podmnožinou **k** atributů, ne mezi všemi jako normálně

  (= bagging, když **k** je rovno počtu attributů)
- Odhadnete vliv **k**?
  - Menší **k** redukuje varianci a zvyšuje bias

Boosting

- Iterativní procedura adaptivně měnící rozložení učicích dat zvýrazňujíce špatně klasifikované vzory
- Používá se zejména ke kombinaci slabých modelů (weak learners), které mají velké zaujetí
- Výrazně redukuje bias – náchylnost k přeučení

Boosting (2)



Pro regresi: $\hat{y}(\underline{x}) = \beta_1 * \hat{y}_1(\underline{x}) + \beta_2 * \hat{y}_2(\underline{x}) + \ldots + \beta_T * \hat{y}_T(\underline{x}))$

Pro klasifikaci: $\hat{y}(\underline{x})$ = majorita tříd z $\{\hat{y}_1(\underline{x}),\ldots,\hat{y}_T(\underline{x})\}$
s použitím vah $\{\beta_1,\beta_2,\ldots,\beta_T\}$

Boosting (3)

- Na počátku mají všechny vzory stejnou váhu
- Po naučení modelu zvýšíme váhu špatně klasifikovaným vzorům a snížíme ji dobře klasifikovaným
- **Příklad:** vzor 4 je špatně klasifikovatelný
- Postupně se zvyšuje jeho váha a tedy i pravděpodobnost zařazení do učicí množiny

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

Algoritmus AdaBoost (1)

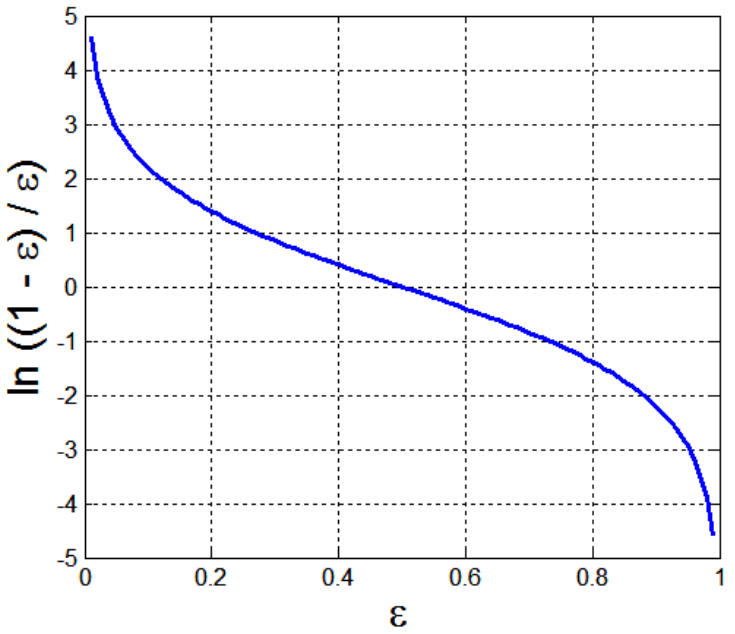
- Klasifikátory: C_1, C_2, …, C_T
- Jejich chyby:

váha vzoru (chyby na zmnožených vzorech bolí více!)

$$\varepsilon_i = \frac{1}{N}\sum_{j=1}^{N} w_j \delta\left(C_i(x_j) \neq y_j\right)$$

- Důležitost klasifikátoru:

$$\beta_i = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$



Algoritmus AdaBoost (2)

- Update vah:

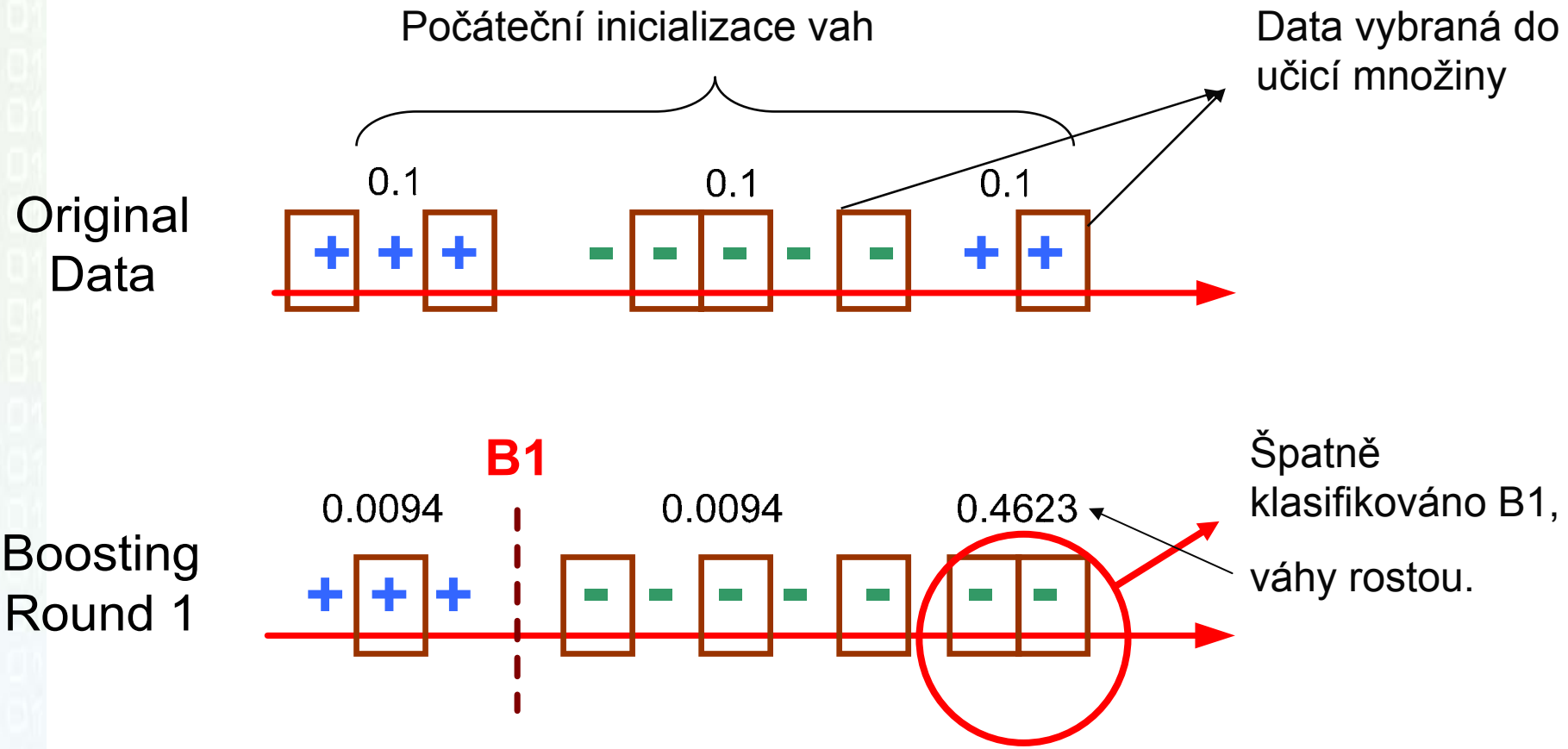
$$w_i^{(j+1)} = \frac{w_i^{(j)}}{Z_j} \begin{cases} \exp^{-\beta_j} & \text{if } C_j(x_i) = y_i \\ \exp^{\beta_j} & \text{if } C_j(x_i) \neq y_i \end{cases}$$

kde Z_j je normalizační konstanta

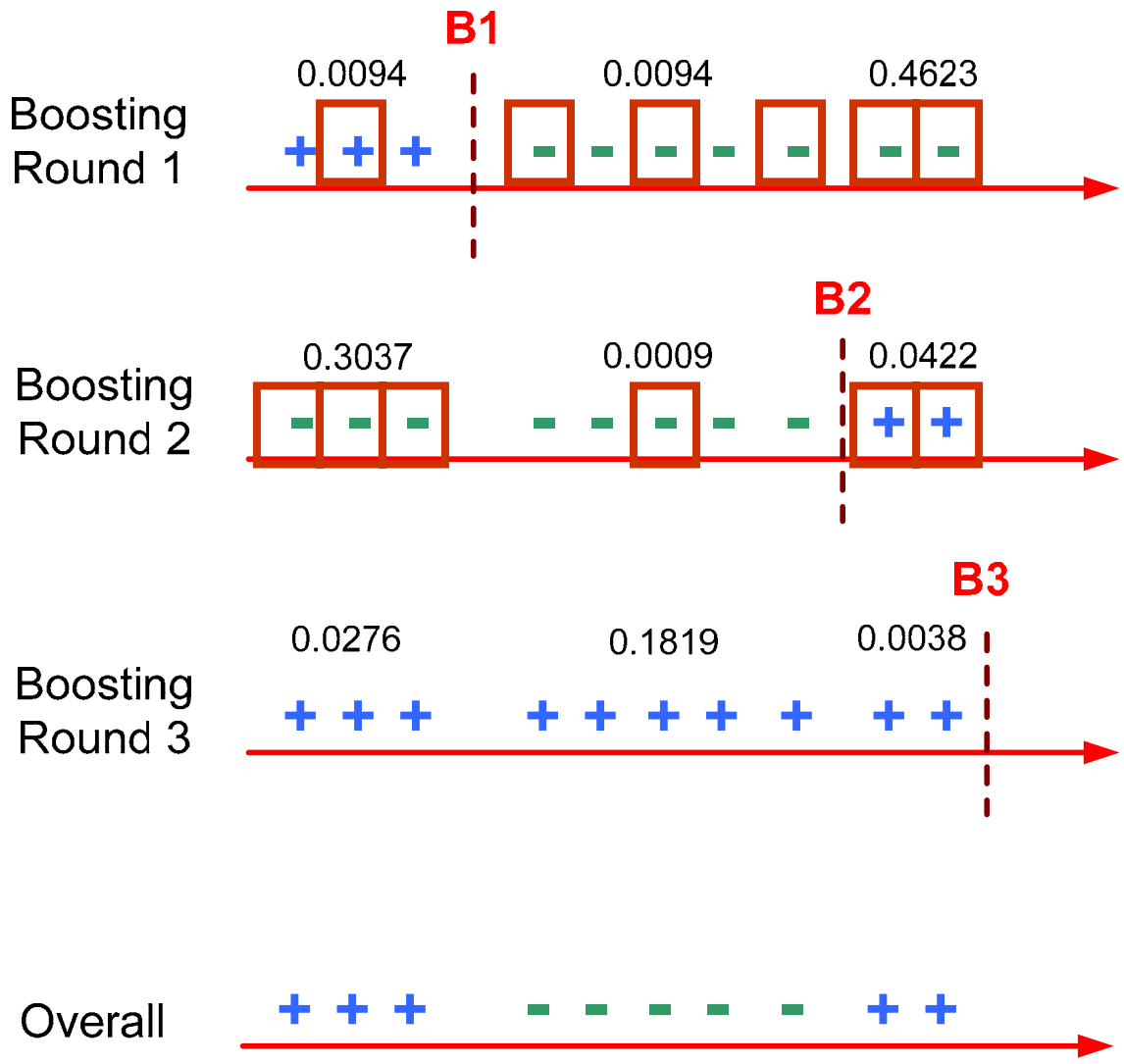
- Finální klasifikace:

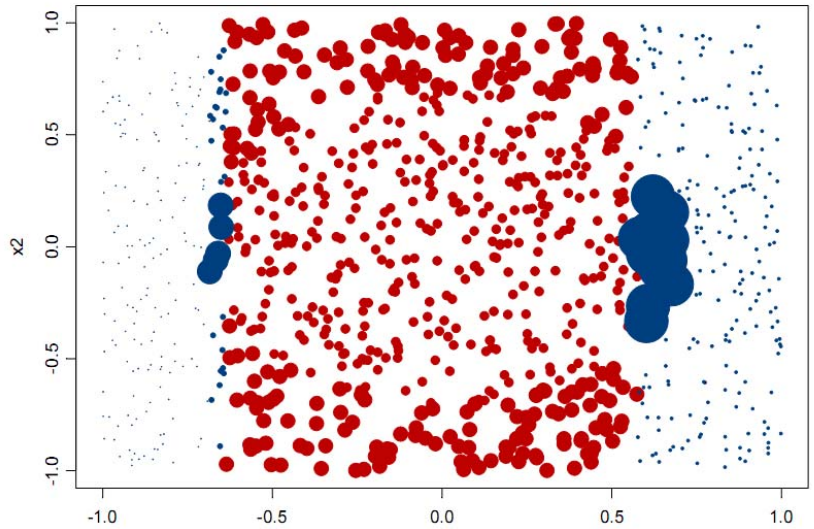
$$C^*(x) = \arg\max_y \sum_{j=1}^{T} \beta_j \delta\left(C_j(x) = y\right)$$

Příklad – AdaBoost



Příklad – AdaBoost

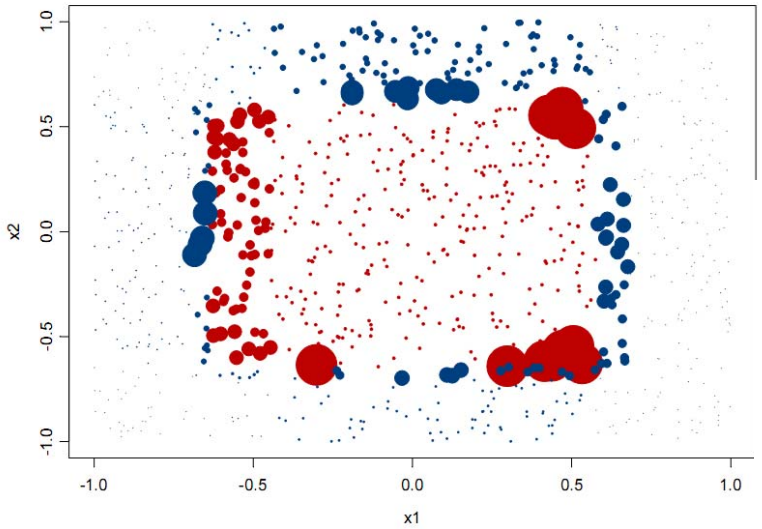




Při inicializaci jsou váhy vzorů stejné, puntíky mají stejnou velikost. Vzory se mají klasifikovat do modré a červené třídy.

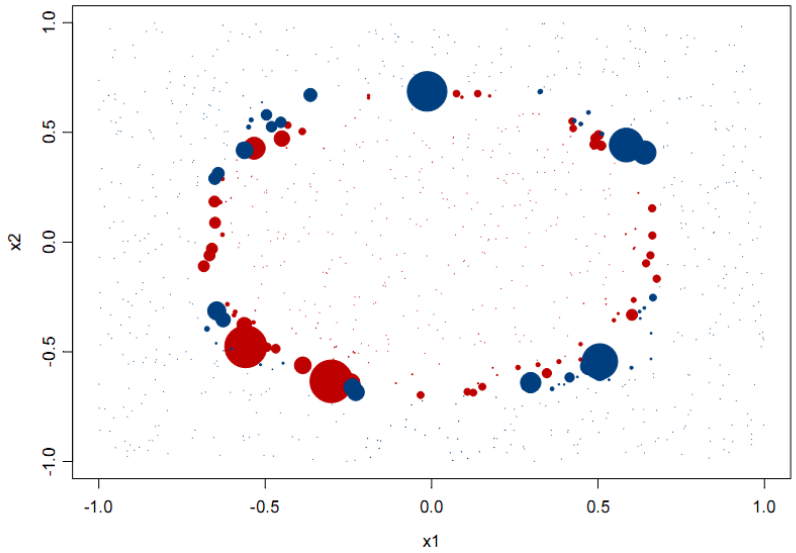
Po 1. iteraci AdaBoostu rostou váhy špatně klasifikovaných vzorů na pomezí tříd

3. iterace



Po naučení 20 klasifikátorů jsou váhy vzorů mimo hranici tříd téměř nulové (nejsou vybírány do trénovací množiny)

20 iterace

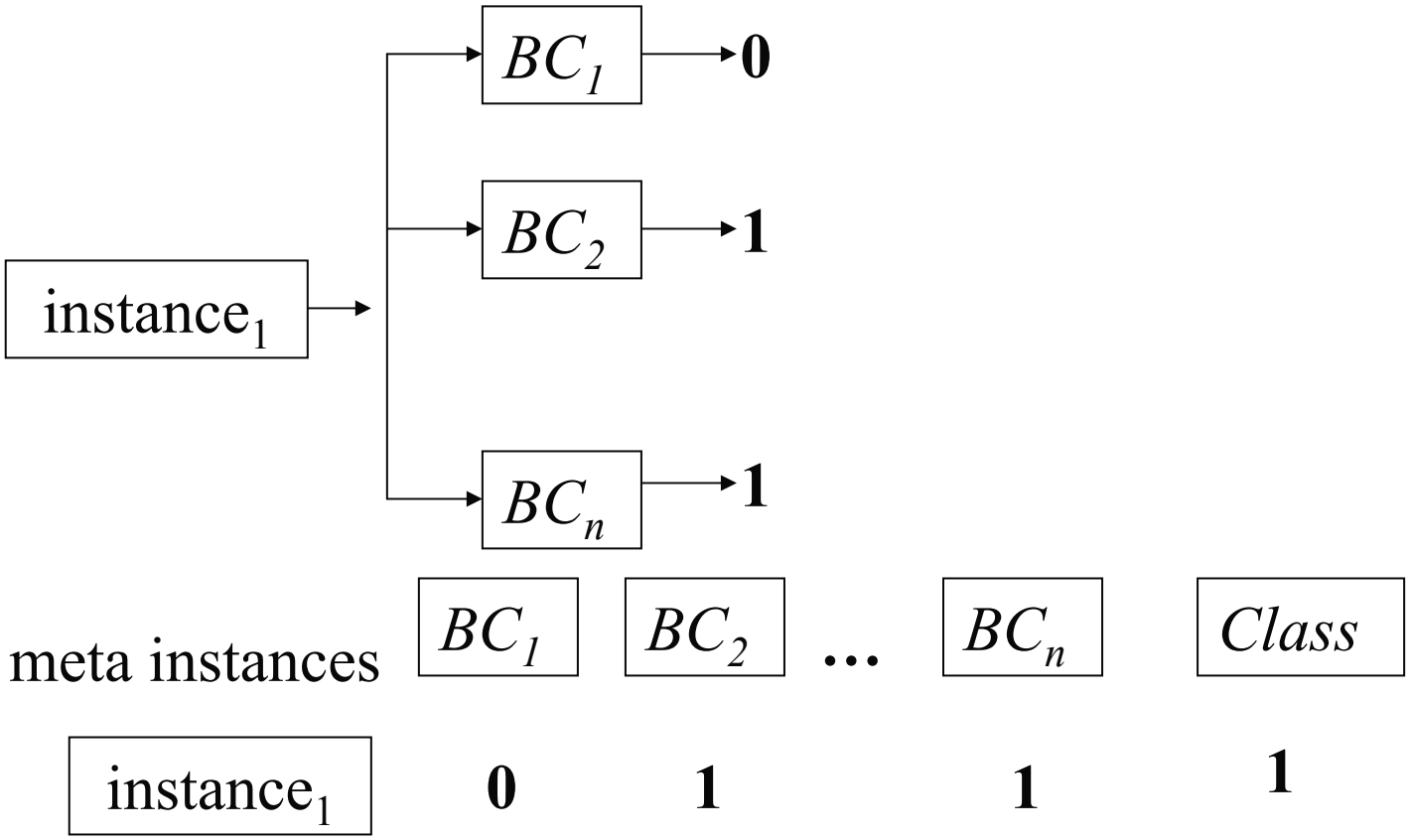


from Elder, John. From Trees to Forests and Rule Sets - A Unified Overview of Ensemble Methods. 2007.

Stacking

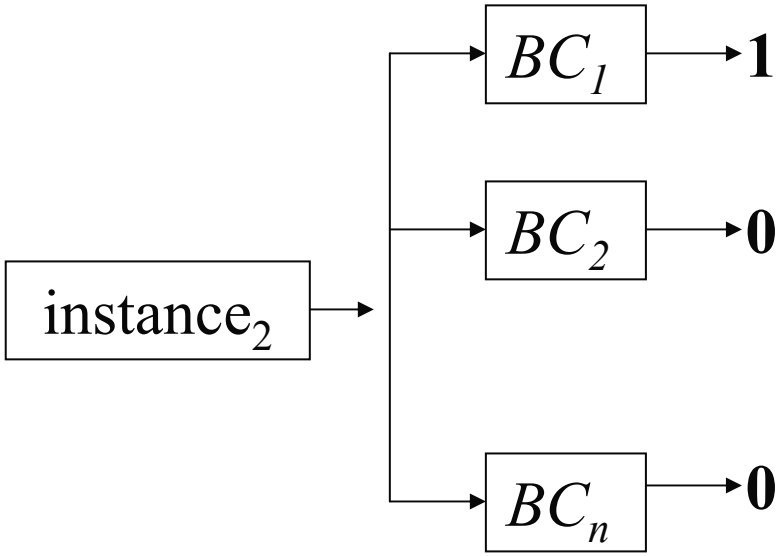
- Používá *meta model* pro kombinaci výstupů ensemble modelů (oproti jednoduchému průměrování, nebo hlasování)
  - Výstupy ensemble modelů jsou použity jako trénovací data pro meta model
- Ensemble modely jsou většinou naučeny různými algoritmy
- Teoretická analýza stackingu je obtížná, jedná se o "black magic"

Stacking



Slide from Ensembles of Classifiers by *Evgueni Smirnov*

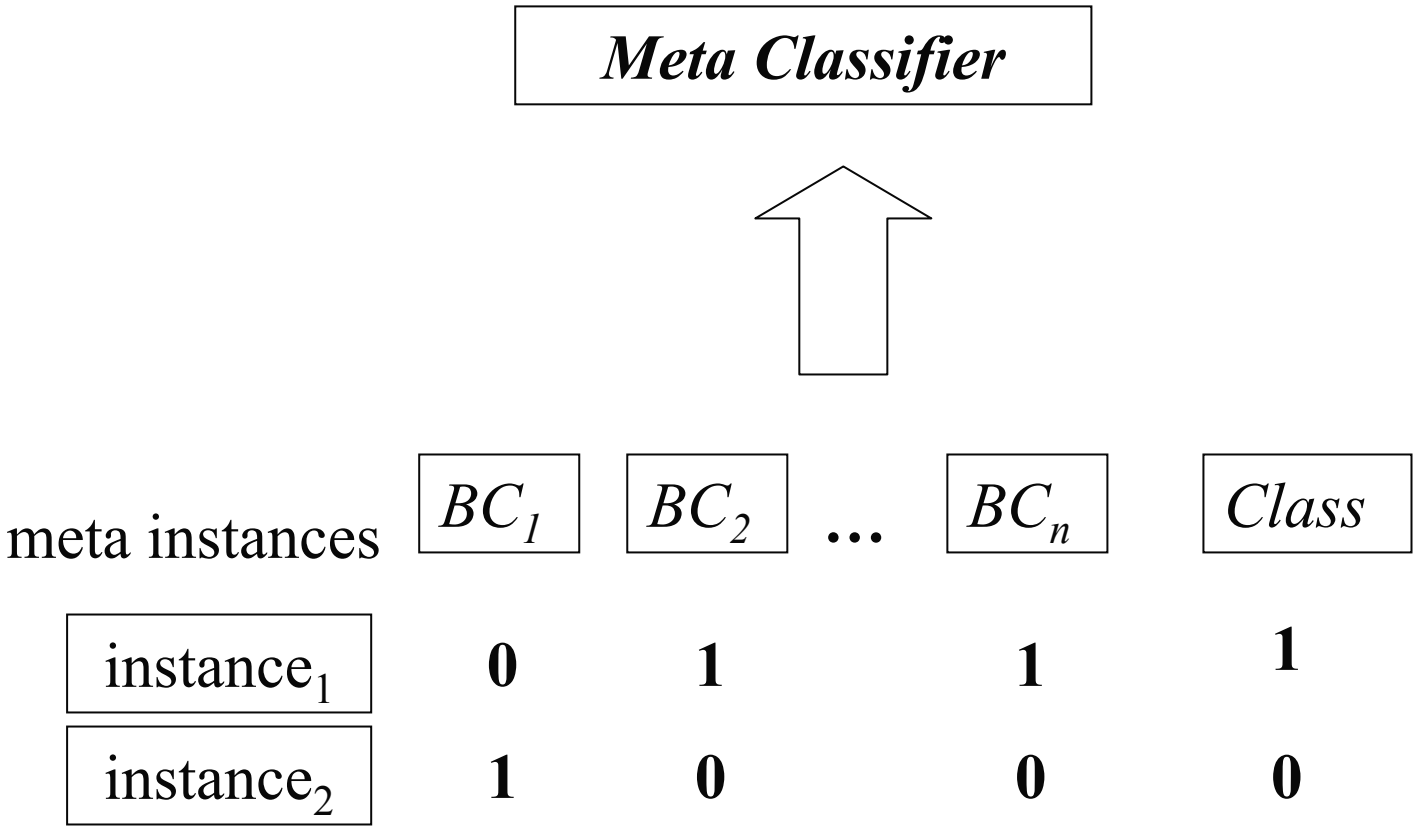
Stacking



meta instances	BC_1	BC_2	…	BC_n	$Class$
instance$_1$	**0**	**1**		**1**	**1**
instance$_2$	**1**	**0**		**0**	**0**

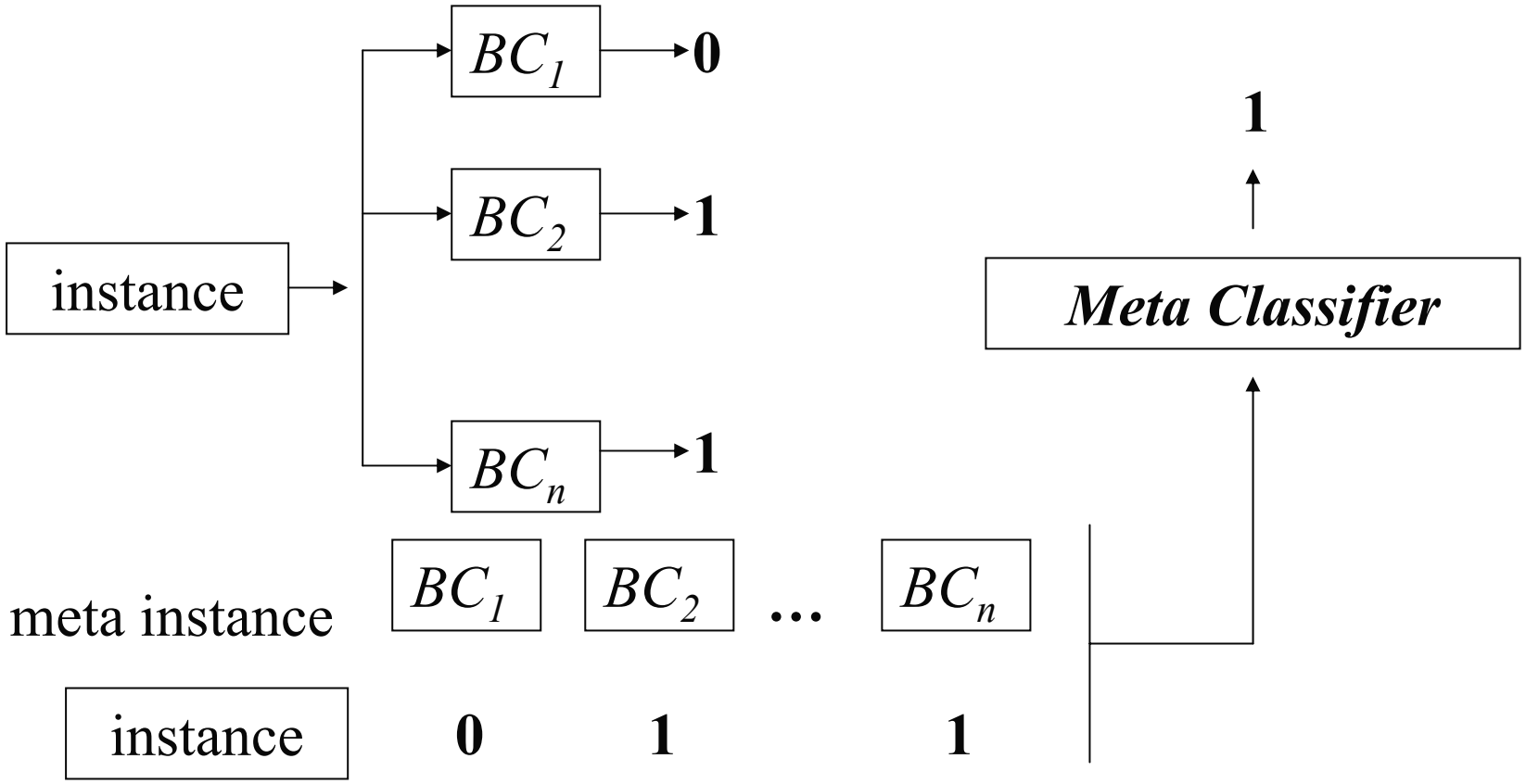
Slide from Ensembles of Classifiers by *Evgueni Smirnov*

Stacking



Slide from Ensembles of Classifiers by *Evgueni Smirnov*

Stacking



Slide from Ensembles of Classifiers by *Evgueni Smirnov*

Argumenty proti kombinování modelů?

- Okamova břitva – v jednoduchosti je síla
  - Je lepší mít jednoduchý optimální model, než kombinaci mnoha modelů
  - … ale jak najít **optimální** model?
  - Domingos, P. Occam's two razors: the sharp and the blunt. KDD 1998.
- Kombinováním modelů se často kamufluje nedokonalost metod produkujících nedoučené nebo přeučené modely
- Kombinováním dostanu model s horšími výsledky na testovacích datech, než mají kombinované modely

Argumenty pro kombinování

- Většinou zlepším výsledky na testovacích datech
  - Algoritmy jsou implicitně nastaveny, je třeba experimentovat s jejich konfigurací, aby produkované modely byly optimální konkrétních datech
- Dostanu povědomí o jistotě modelu
  - když se pro jeden vstupní vektor jednotlivé modely hodně liší, zřejmě jsme mimo oblast trénovacích dat
- Netflix prize

Co přesně použil vítěz?

- Víceúrovňový Stacking pomocí MLP
- Gradient Boosting rozhodovacích stromů
- A samozřejmě výborné base-klasifikátory:
  - KNN
  - SVD
  - RBM
  - …

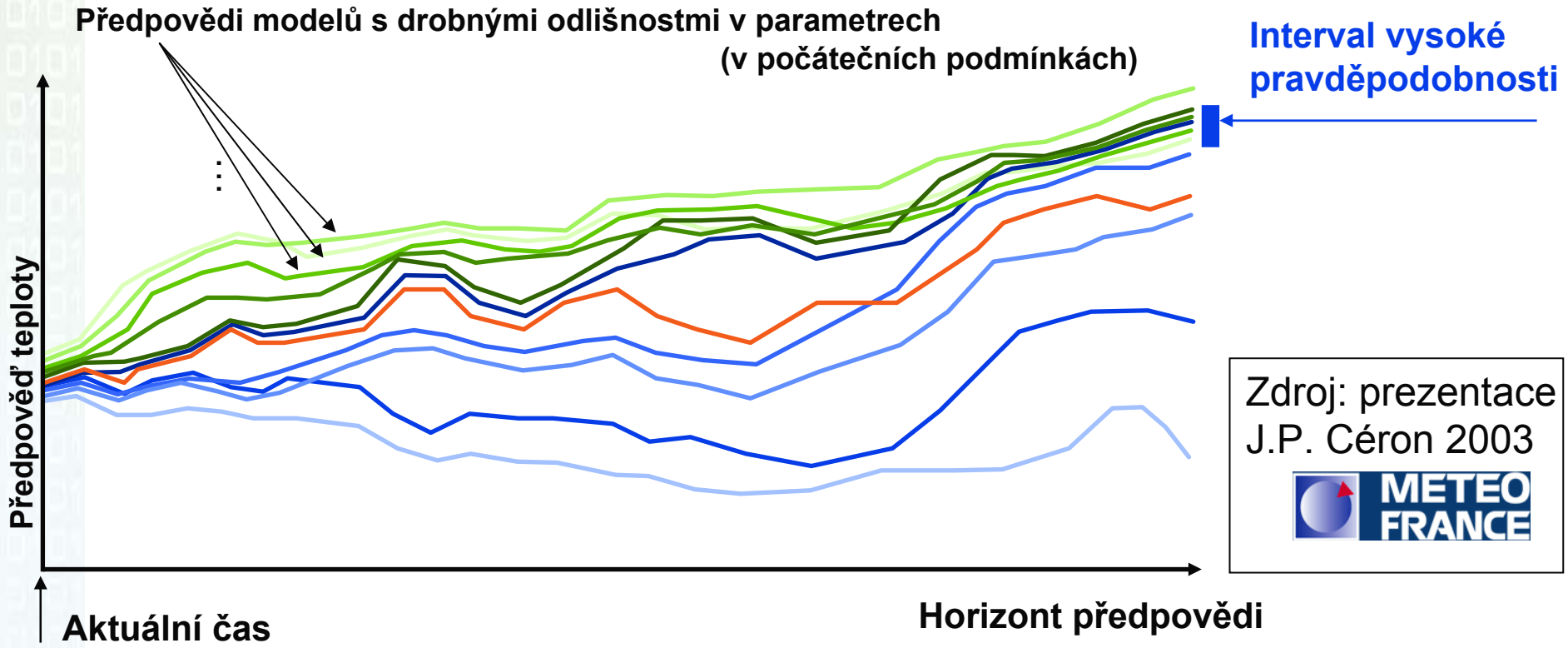
Netflix prize a další aplikace

- Podrobný popis vítězných algoritmů
  - http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf
  - http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf
  - http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf
- Ještě něco zajímavého na ensemblech?

Příklady použití ensemble metod (klimatické modely)

Skupinová předpověď (Ensemble forecast)

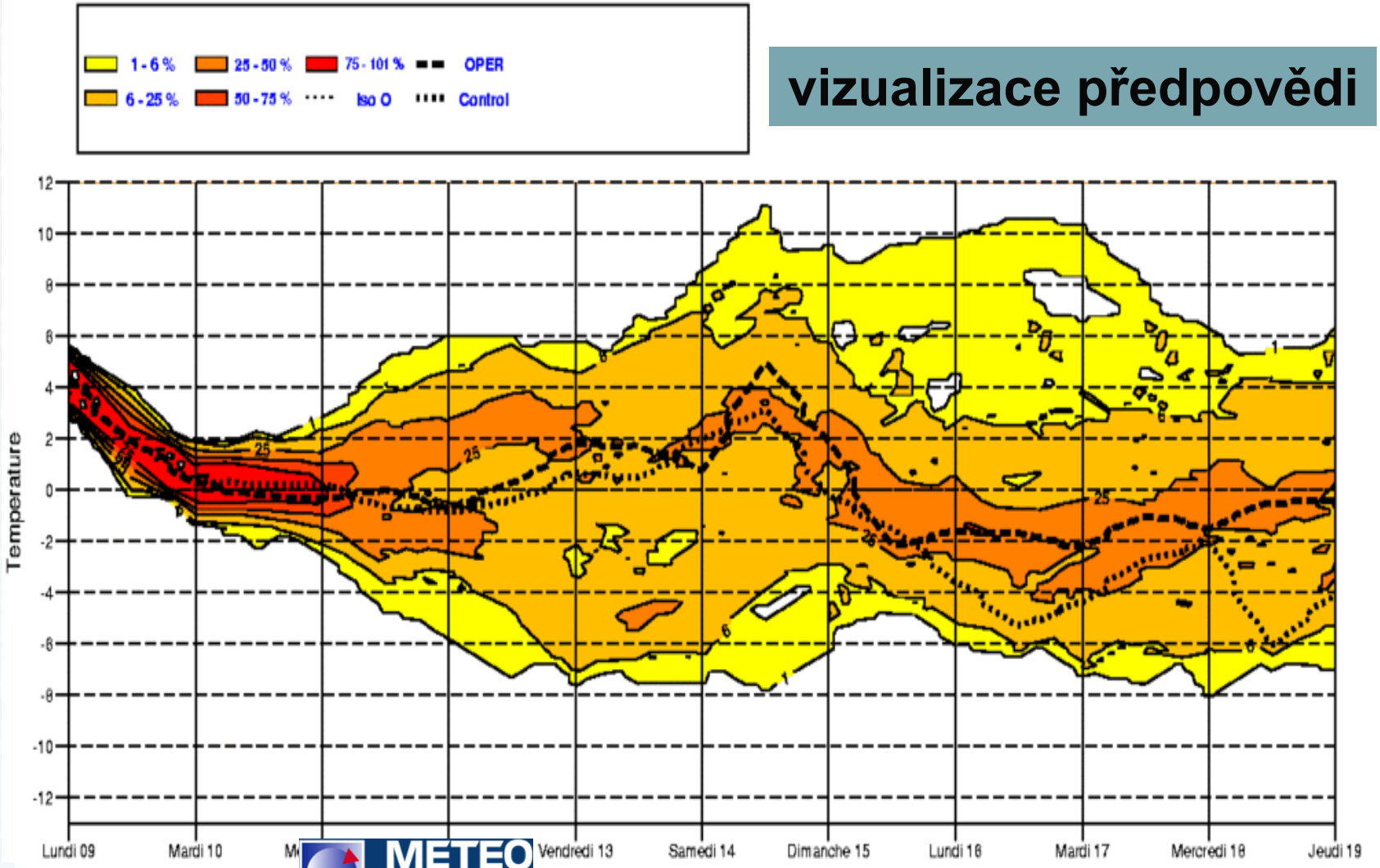
Ensemble je soubor předpovědí, které kolektivně mapují pravděpodobné budoucí stavy, s ohledem na nejistotu provázející proces předpovědi.



Zdroj: prezentace J.P. Céron 2003

METEO FRANCE

Ensemble modely pro předpověď počasí

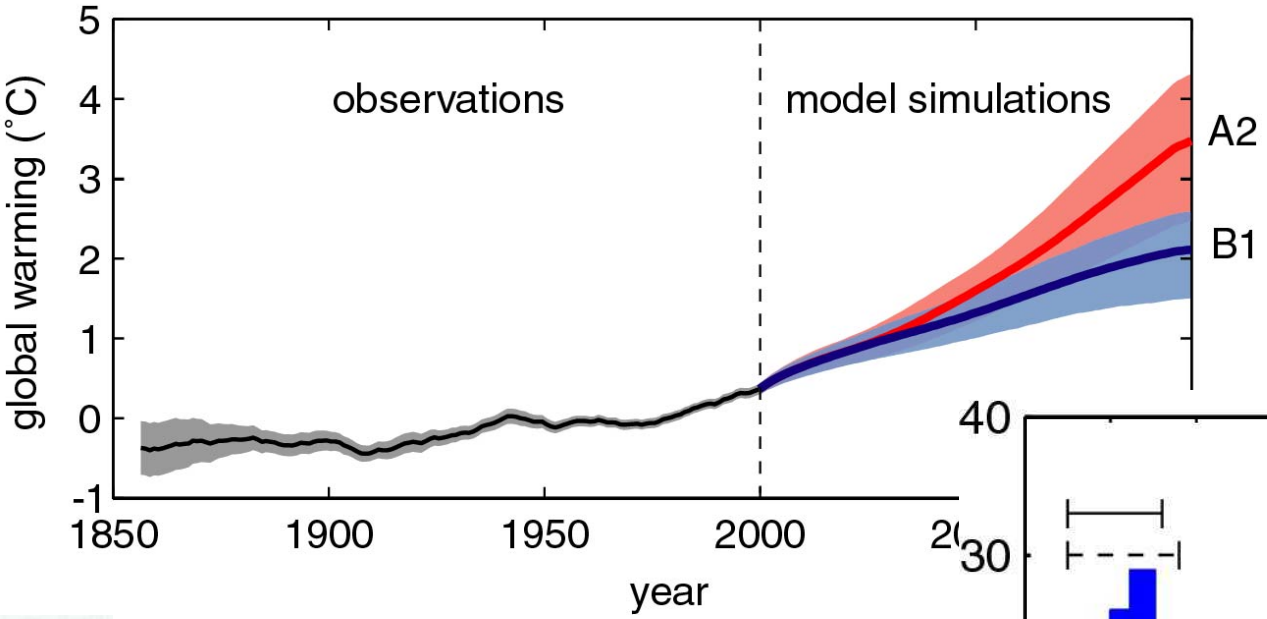


vizualizace předpovědi

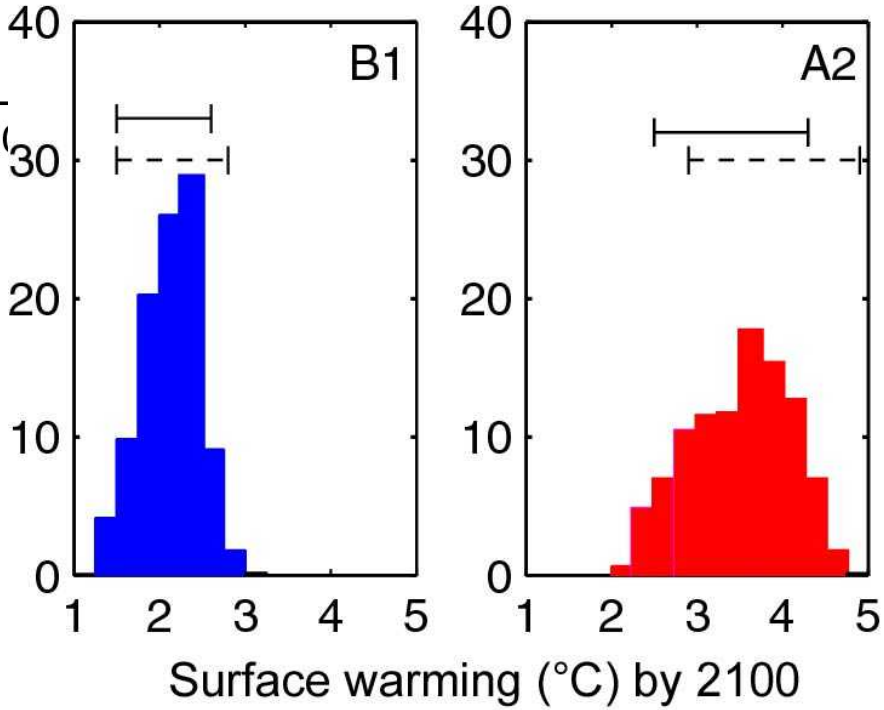
METEO FRANCE

zdroj: J.P. Céron

Ensemble modely a globální oteplování



Knutti et al. (2002)



Pouze jeden model ignoruje neodmyslitelné mezery v našich znalostech a limity předvídatelnosti vývoje Země.

Potřebujeme celou skupinu (ensemble) modelů, které se liší složitostí, počátečními podmínkami a parametry.

Dr. Thomas Stocker, Climate and Environmental Physics
University of Bern, Switzerland

www.climate.unibe.ch, 2003

Otázky

- Jakou novou informaci získáme použitím skupiny modelů oproti použití pouze jednoho modelu?
- Může ensemble zpřesnit předpověď, pro jaké modely?
- Jaké jsou nevýhody ensemble předpovědi?