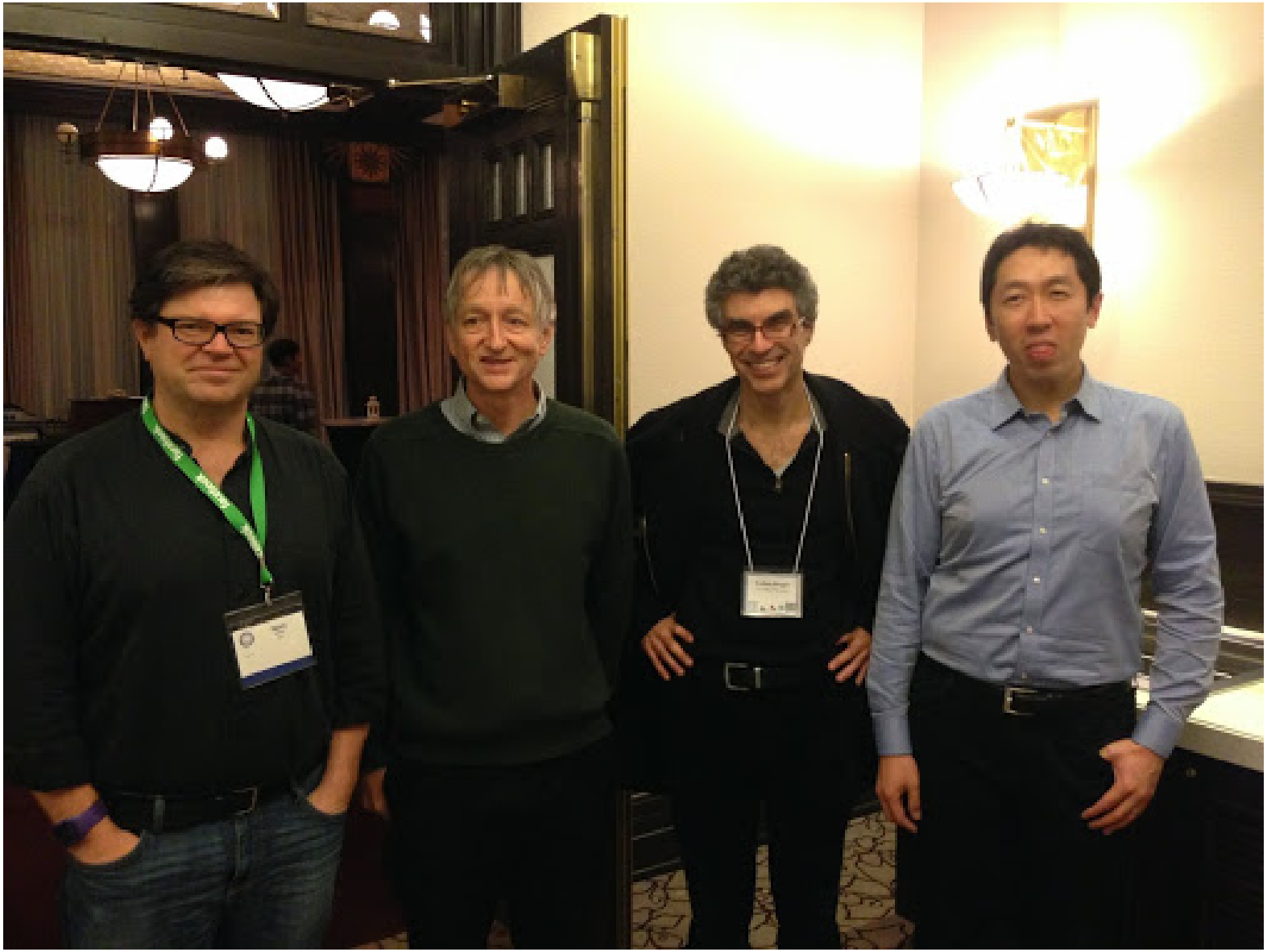
Klasyfikacja z milionami etykiet

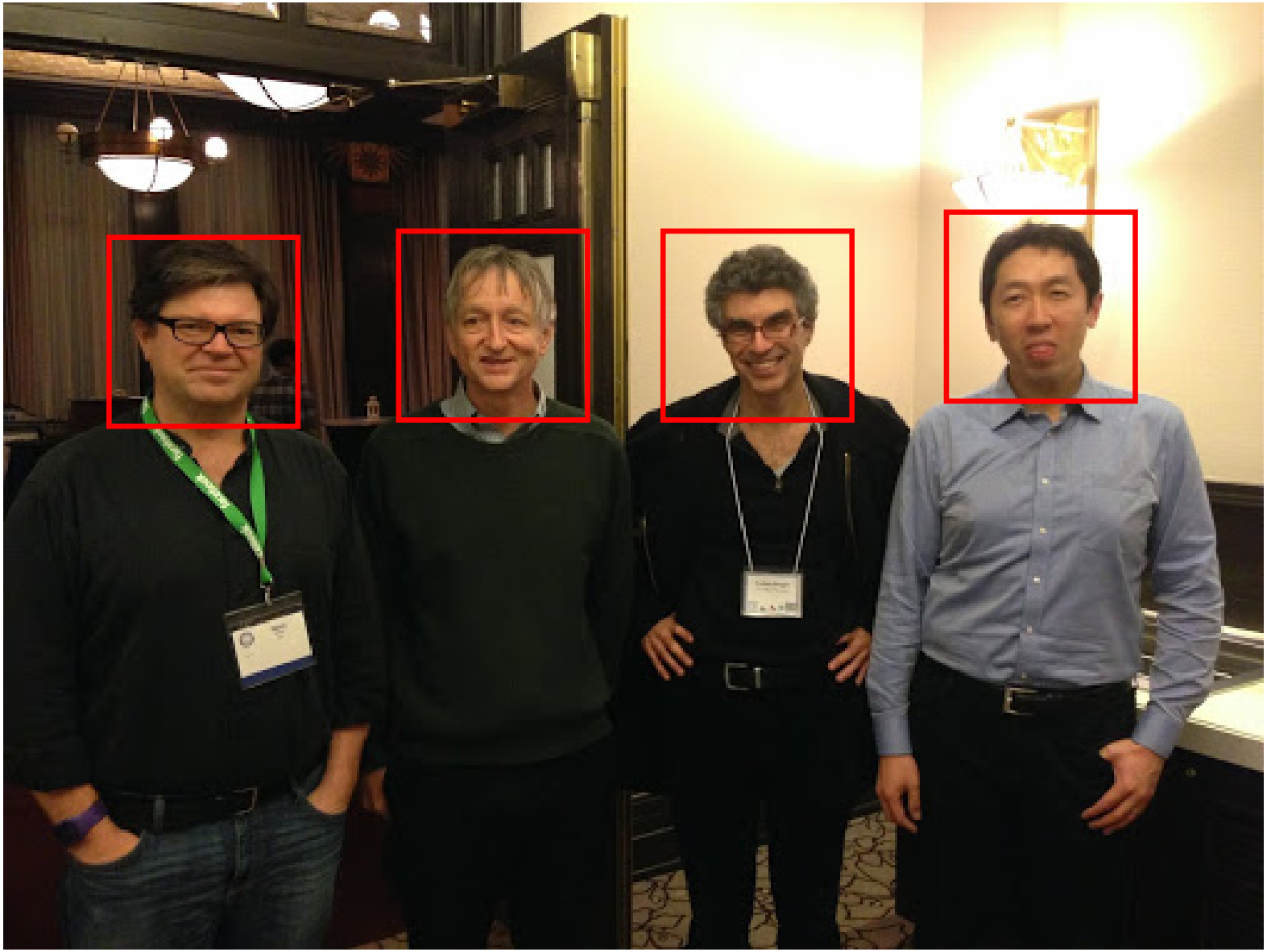
Krzysztof Dembczyński

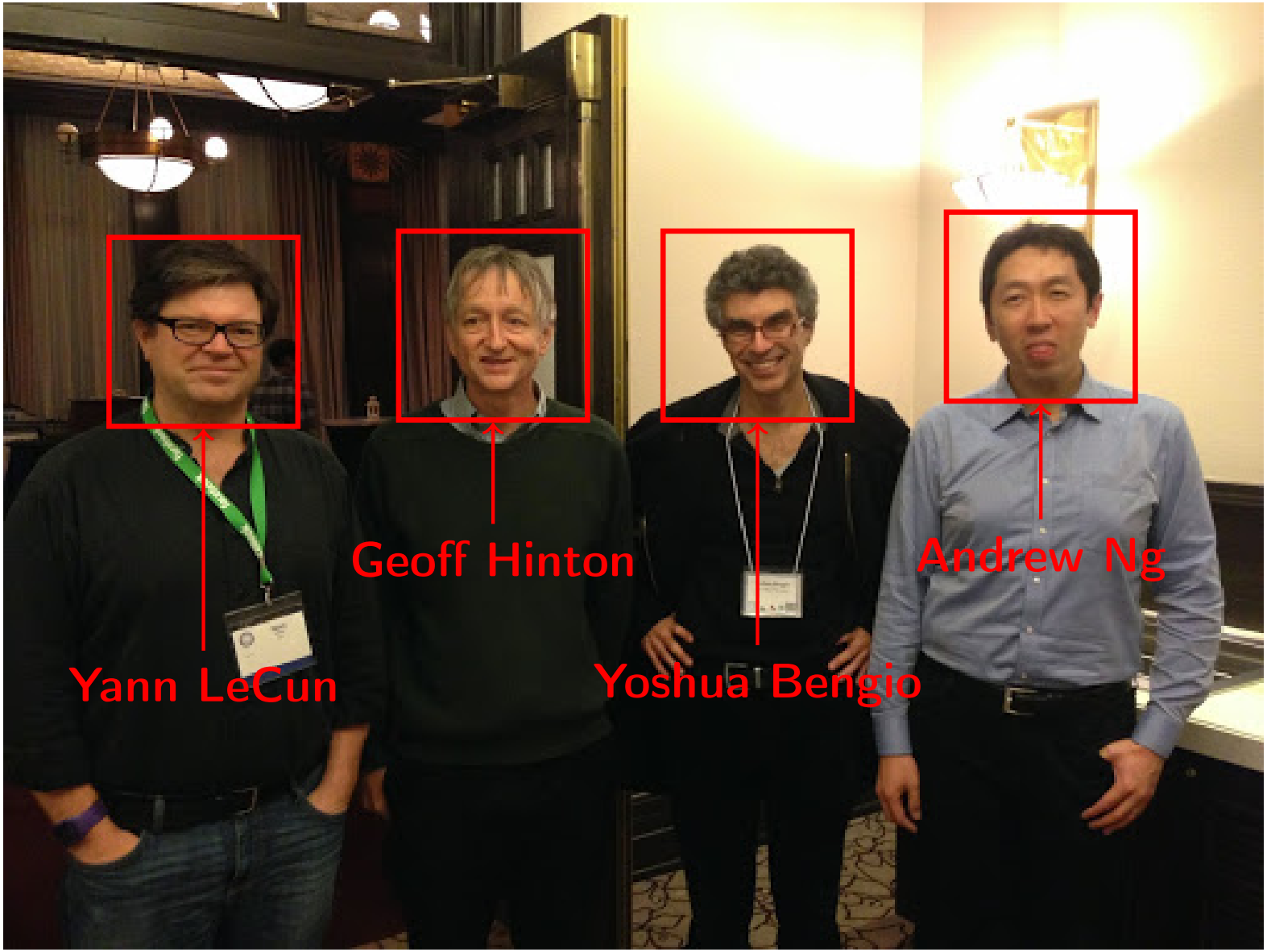
Zakład Inteligentnych Systemów Wspomagania Decyzji
Politechnika Poznańska

Big Data: Przetwarzanie i eksploracja

Poznań, 22 kwietnia 2016 r.






Yann LeCun
Geoff Hinton
Yoshua Bengio
Andrew Ng

Dwa rodzaje problemów

- **Klasyfikacja wieloklasowa**: jedna etykieta przypisana do obiektu.
- **Klasyfikacja wieloetykietowa**: żadna, jedna lub więcej etykiet przypisanych do obiektu.

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p \xrightarrow{\boldsymbol{h}(\boldsymbol{x})} \boldsymbol{y} = (y_1, y_2, \ldots, y_m) \in \{0, 1\}^m$$

	x_1	x_2	$\ldots$	x_p	y_1	y_2	$\ldots$	y_m
$\boldsymbol{x}$	4.0	2.5		-1.5	?	?		?

Dwa rodzaje problemów

- **Klasyfikacja wieloklasowa**: jedna etykieta przypisana do obiektu.
- **Klasyfikacja wieloetykietowa**: żadna, jedna lub więcej etykiet przypisanych do obiektu.

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p \xrightarrow{\boldsymbol{h}(\boldsymbol{x})} \boldsymbol{y} = (y_1, y_2, \ldots, y_m) \in \{0, 1\}^m$$

	x_1	x_2	$\ldots$	x_p	y_1	y_2	$\ldots$	y_m
$\boldsymbol{x}$	4.0	2.5		-1.5	1	1		0

Dwa rodzaje problemów

- **Klasyfikacja wieloklasowa**: jedna etykieta przypisana do obiektu.
- **Klasyfikacja wieloetykietowa**: żadna, jedna lub więcej etykiet przypisanych do obiektu.

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p) \in \mathbb{R}^p \xrightarrow{\boldsymbol{h}(\boldsymbol{x})} \boldsymbol{y} = (y_1, y_2, \ldots, y_m) \in \{0, 1\}^m$$

	x_1	x_2	$\ldots$	x_p	y_1	y_2	$\ldots$	y_m
$\boldsymbol{x}$	4.0	2.5		-1.5	1	1		0

- $m \to \infty \Rightarrow$ **klasyfikacja ekstremalna** (ang.*extreme classification*)

Zbiory danych

	#etykiet	#cech	#test.	#trening.	przk./etyk.	etyk./przk.
RCV1	2456	47236	155962	623847	1218.56	4.79
AmazonCat	13330	203882	306782	1186239	448.57	5.04
Wiki10	30938	101938	6616	14146	8.52	18.64
Delicious	205443	782585	100095	196606	72.29	75.54
WikiLSHTC	**325056**	**1617899**	**587084**	**1778351**	**17.46**	**3.19**
Amazon	670091	135909	153025	490449	3.99	5.45

Tabela: Zbiory danych z repozytorium `XMLRepository`.[1]

[1] `http://research.microsoft.com/en-us/um/people/manik/downloads/XC/XMLRepository.html`

Podejście tradycyjne/naiwne

Podejście tradycyjne/naiwne

- **Gęsty model liniowy** dla każdej etykiety:

$$\boldsymbol{X}^T \boldsymbol{W} = \hat{\boldsymbol{Y}}$$

Podejście tradycyjne/naiwne

- **Gęsty model liniowy** dla każdej etykiety:

$$\boldsymbol{X}^T \boldsymbol{W} = \hat{\boldsymbol{Y}}$$

- Rozmiar problemu **WikiLSHTC**:
  - # przykładów treningowych: $n = 1\,778\,351$
  - # przykładów testowych: $n' = 587\,084$
  - # cech: $d = 1\,617\,899$
  - # etykiet: $m = 325\,056$

Podejście tradycyjne/naiwne

- **Gęsty model liniowy** dla każdej etykiety:

$$\boldsymbol{X}^T \boldsymbol{W} = \hat{\boldsymbol{Y}}$$

- Rozmiar problemu **WikiLSHTC**:
  - # przykładów treningowych: $n = 1\,778\,351$
  - # przykładów testowych: $n' = 587\,084$
  - # cech: $d = 1\,617\,899$
  - # etykiet: $m = 325\,056$
- Złożoność pamięciowa:

$$325\,056 \times 1\,617\,899 = 5.26 \times 10^{12}$$

- Złożoność obliczeniowa uczenia:

$$1\,778\,351 \times 325\,056 \times 1\,617\,899 = 9.35 \times 10^{17}$$

- Złożoność obliczeniowa testowania:

$$587\,084 \times 325\,056 \times 1\,617\,899 = 3.09 \times 10^{17}$$

Podejście tradycyjne

- **Nie musi być aż tak źle**:

Podejście tradycyjne

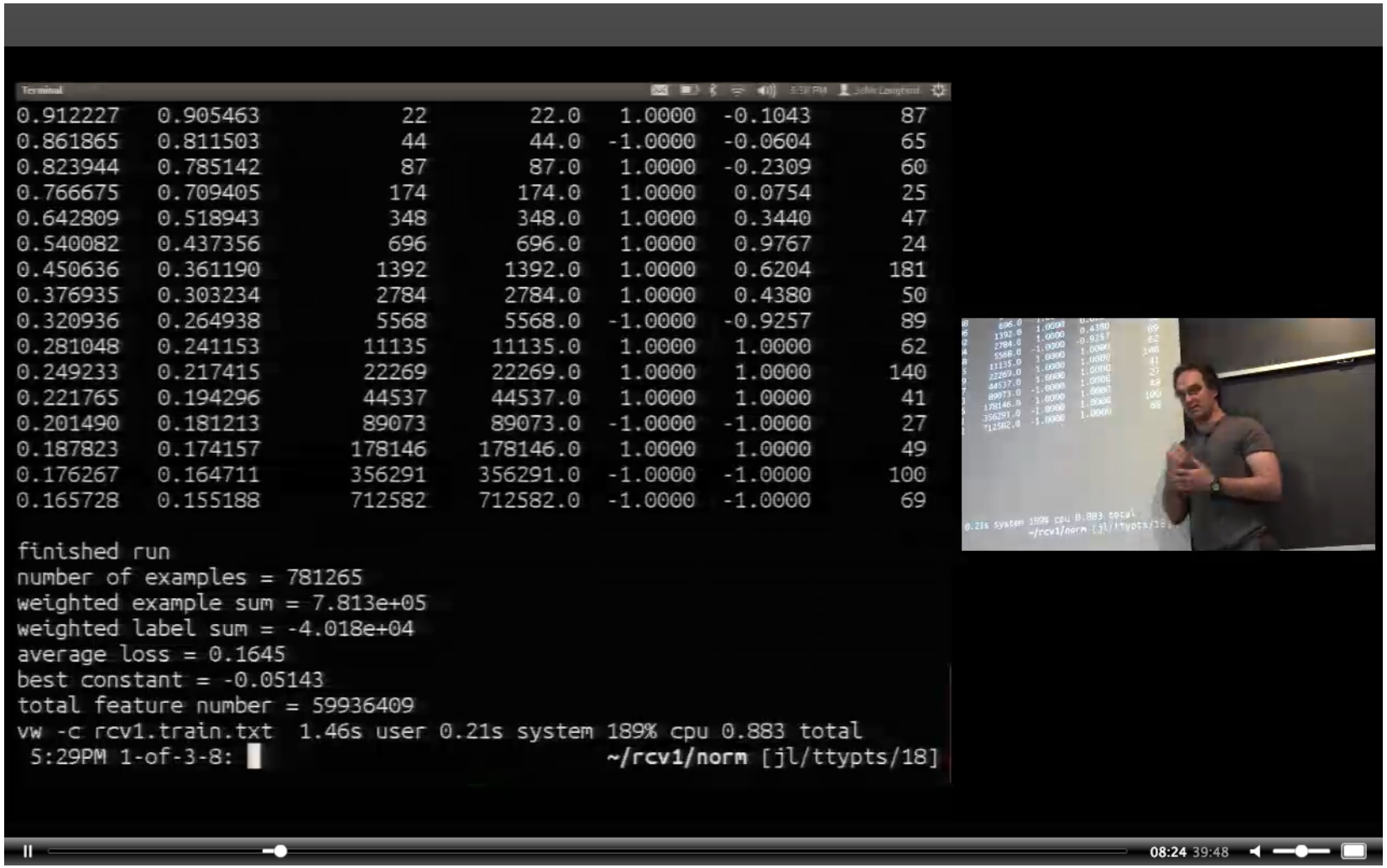
- **Nie musi być aż tak źle**:
  - Duże dane $\longrightarrow$ rzadkie dane (rzadkie cechy i rzadkie etykiety)

Podejście tradycyjne

- **Nie musi być aż tak źle**:
  - Duże dane $\longrightarrow$ rzadkie dane (rzadkie cechy i rzadkie etykiety)
  - Szybkie algorytmy dla tradycyjnych problemów uczenia maszynowego

Podejście tradycyjne

- **Nie musi być aż tak źle**:
  - Duże dane $\longrightarrow$ rzadkie dane (rzadkie cechy i rzadkie etykiety)
  - Szybkie algorytmy dla tradycyjnych problemów uczenia maszynowego
  - Duże moce obliczeniowe



Rysunek: Vowpal Wabbit[2] na wykładzie Johna Langforda[3]

[2] Vowpal Wabbit, `http://hunch.net/~vw`

[3] `http://cilvr.cs.nyu.edu/doku.php?id=courses:bigdata:slides:start`

Szybka klasyfikacja binarna[4]

- Zbiór danych: **RCV1**
- Predykcja kategorii: CCAT
- # przykładów uczących: 781 265
- # cech: 60M
- Rozmiar: 1.1 GB
- Linia komend: `time vw -sgd rcv1.train.txt -c`
- Czas uczenia: 1-3 sekundy na laptopie.

[4] `http://cilvr.cs.nyu.edu/doku.php?id=courses:bigdata:slides:start`

Przyśpieszanie modeli liniowych

- Uczenie modeli:
  - Stochastyczny spadek wzdłuż gradientu.[5]
  - Obsługa rzadkich cech.[6]
  - Negatywne samplowanie.[7]
- Rozmiar modelu:
  - Regularyzacja: L_1 vs L_2.
  - Haszowanie cech.[8]
- Dwa kryteria na raz:
  - Projekcja do nisko-wymiarowej przestrzeni $\boldsymbol{X}$, $\boldsymbol{W}$, $\boldsymbol{Y}$, etc.[9]

---

[5] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*. Springer

[6] Duchi, J. i Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *JMLR*, 10:2899–2934

[7] Collobert, R. i Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*

[8] Weinberger, K., Dasgupta, A., Langford, J., Smola, A., i Attenberg, J. (2009). Feature hashing for large scale multitask learning. In *ICML*

[9] Hsu, D., Kakade, S., Langford, J., i Zhang, T. (2009). Multi-label prediction via compressed sensing. In *NIPS*

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!
- **Redukcja** złożoności poprzez wykorzystanie odpowiednich **struktur danych**:

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!
- **Redukcja** złożoności poprzez wykorzystanie odpowiednich **struktur danych**:
  - Funkcje mieszające

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!
- **Redukcja** złożoności poprzez wykorzystanie odpowiednich **struktur danych**:
  - Funkcje mieszające ($\longrightarrow$ grupowanie).

Klasyfikacja ekstremalna

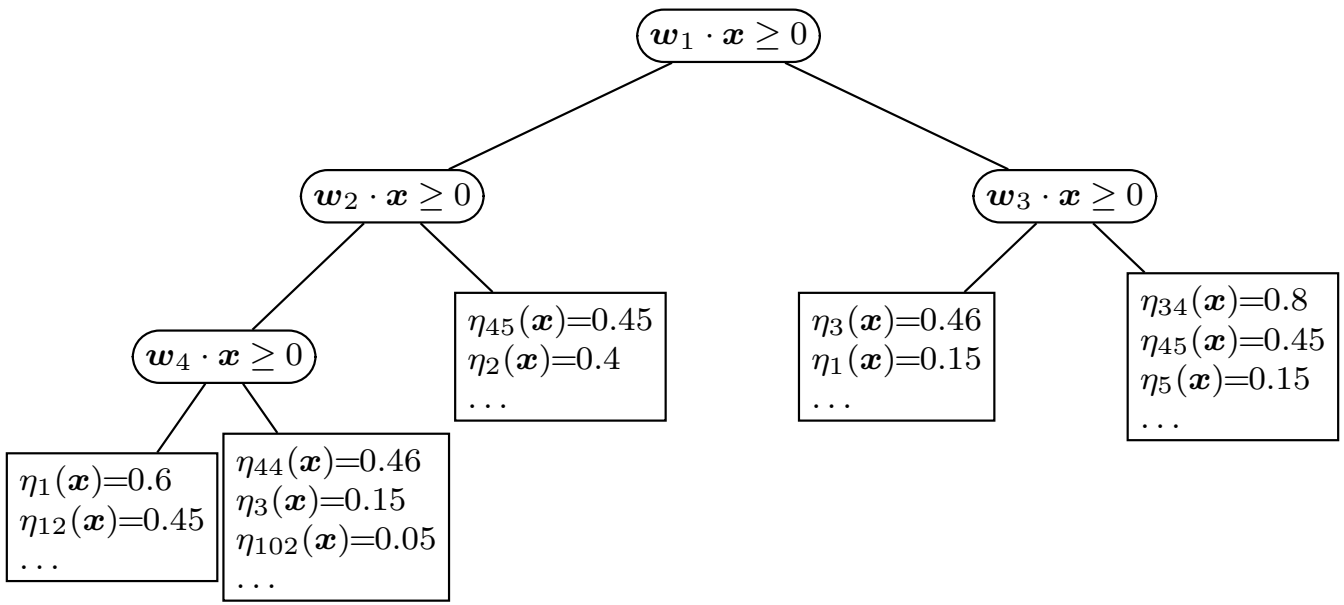
- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!
- **Redukcja** złożoności poprzez wykorzystanie odpowiednich **struktur danych**:
  - Funkcje mieszające ($\longrightarrow$ grupowanie).
  - Sortowanie $\longrightarrow$ **drzewa**

Klasyfikacja ekstremalna

- Wszystkie powyższe techniki poprawiają wydajność, ale ...
- Czas predykcji ciągle rośnie **linowo** z liczbą etykiet!
- **Redukcja** złożoności poprzez wykorzystanie odpowiednich **struktur danych**:
  - Funkcje mieszające ($\longrightarrow$ grupowanie).
  - Sortowanie $\longrightarrow$ **drzewa**
  - $\longrightarrow$ **drzewa decyzyjne**.
  - $\longrightarrow$ **drzewa etykiet**.

Drzewa decyzyjne dla klasyfikacji ekstremalnej

- Dwa przykładowe podejścia: **FastXML**[10] i **LomTrees**[11]
- Klasyfikatory liniowe w wierzchołkach drzewa.
- Szybki podział obiektów na **prawe**/**lewe** w wierzchołkach.
- Jedna ścieżka wzdłuż drzewa podczas predykcji.
- Predykcja **logarytmiczna** w liczbie **przykładów uczących**.

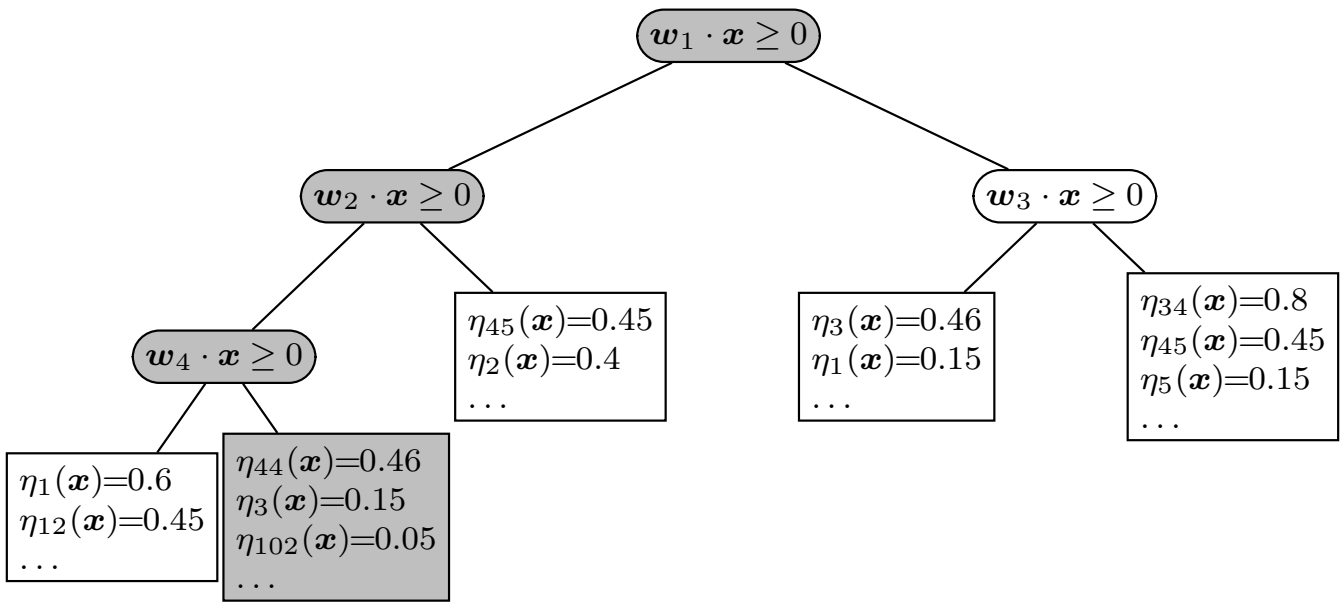


[10] Prabhu, Y. i Varma, M. (2014). FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*

[11] Choromanska, A. i Langford, J. (2015). Logarithmic time online multiclass prediction. In *NIPS 29*

Drzewa decyzyjne dla klasyfikacji ekstremalnej

- Dwa przykładowe podejścia: **FastXML**[10] i **LomTrees**[11]
- Klasyfikatory liniowe w wierzchołkach drzewa.
- Szybki podział obiektów na **prawe**/**lewe** w wierzchołkach.
- Jedna ścieżka wzdłuż drzewa podczas predykcji.
- Predykcja **logarytmiczna** w liczbie **przykładów uczących**.

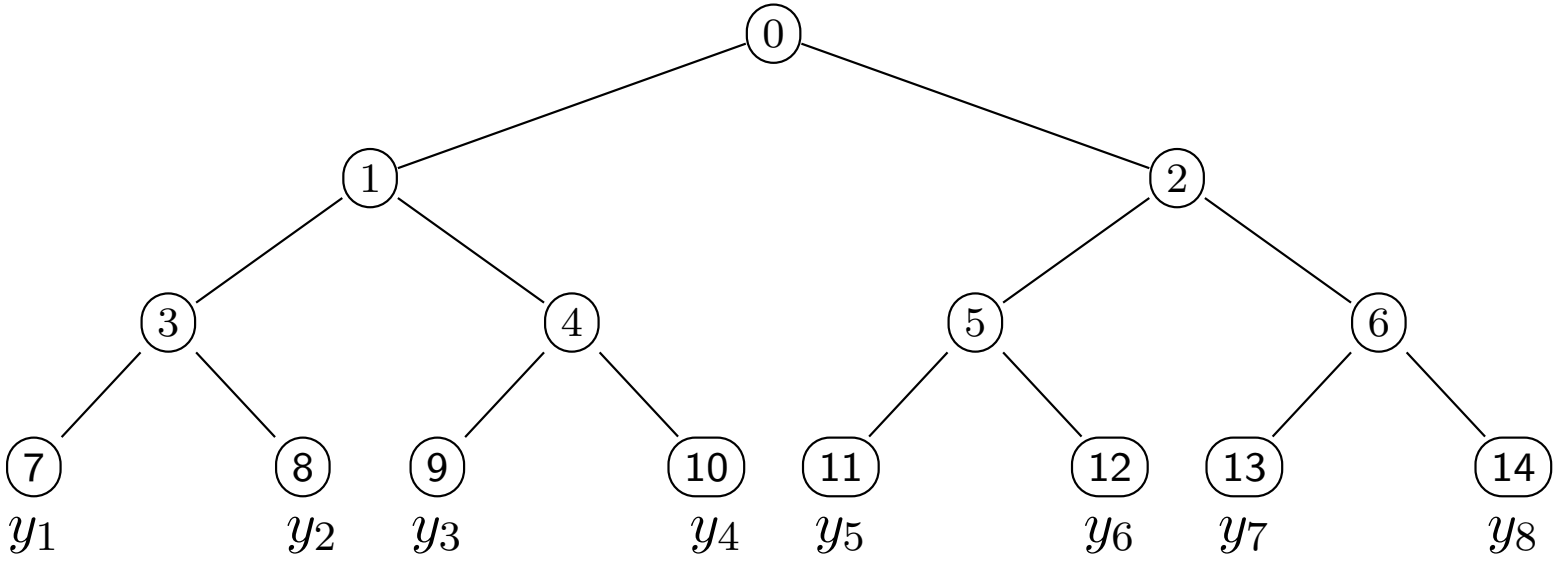


[10] Prabhu, Y. i Varma, M. (2014). FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*

[11] Choromanska, A. i Langford, J. (2015). Logarithmic time online multiclass prediction. In *NIPS 29*

Drzewa etykiet

- Bazują na b-arnych drzewach.[12]



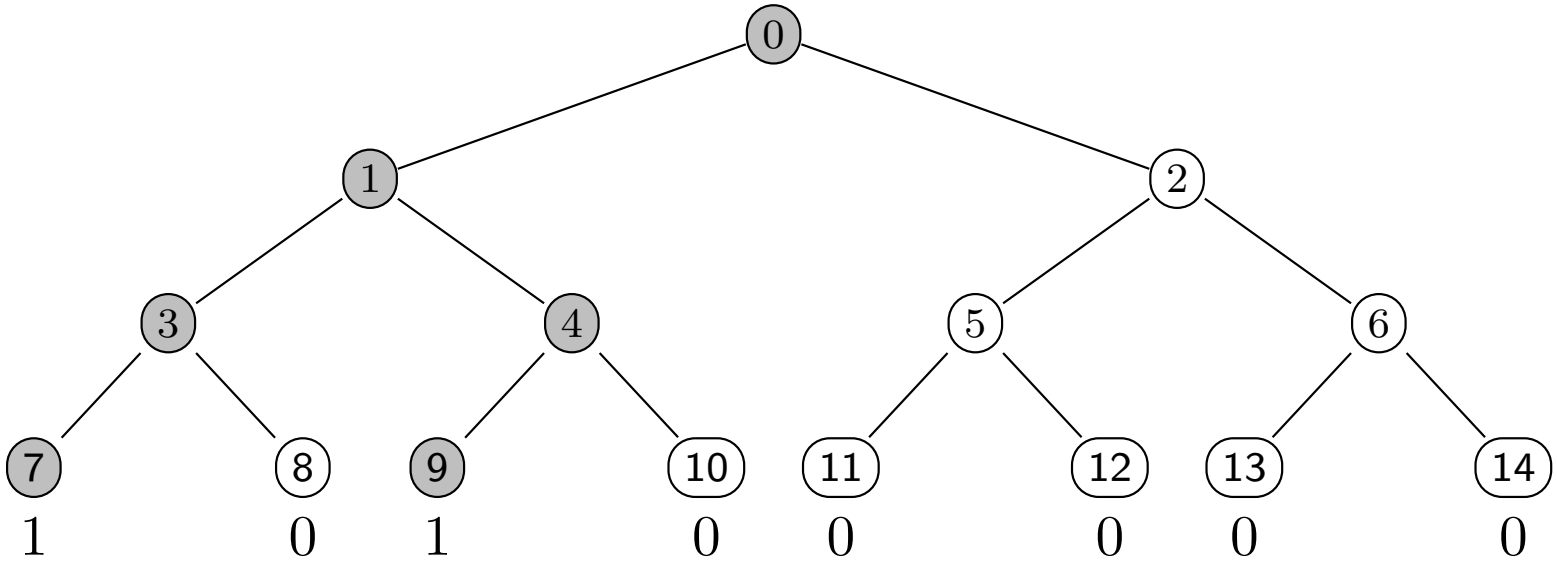
- Każdy **liść** odpowiada jednej etykiecie.
- **Wewnętrzne** klasyfikatory decydują, czy przejść do potomków.
- **Liść** może zawierać ostateczny klasyfikator dla danej etykiety.
- Przykład testowy może odwiedzić **wiele ścieżek** od korzenia do liści.

[12] Bengio, S., Weston, J., i Grangier, D. (2010). Label embedding trees for large multi-class tasks. In *NIPS*, pages 163-171. Curran Associates, Inc

Jasinka, K., Dembczynski, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., i Hüllermeier, E. (2016). Extreme F-measure maximization using sparse probability estimates

Drzewa etykiet

- Bazują na b-arnych drzewach.[12]



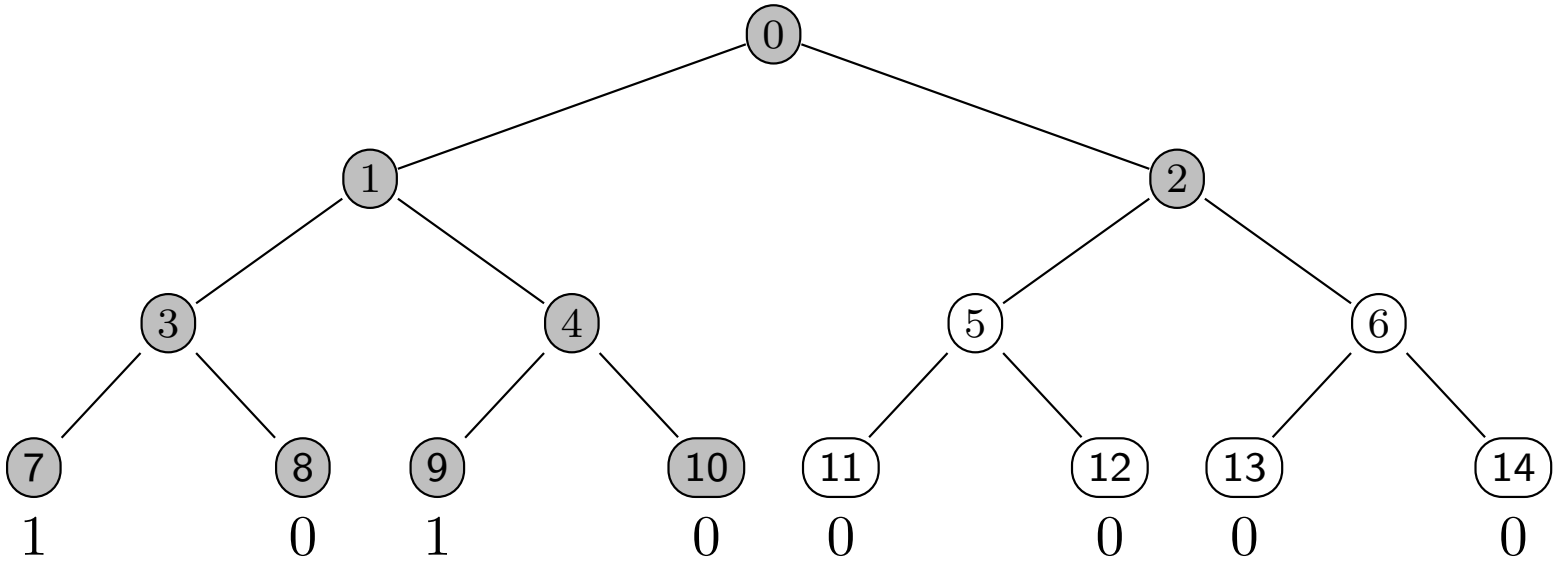
- Każdy **liść** odpowiada jednej etykiecie.
- **Wewnętrzne** klasyfikatory decydują, czy przejść do potomków.
- **Liść** może zawierać ostateczny klasyfikator dla danej etykiety.
- Przykład testowy może odwiedzić **wiele ścieżek** od korzenia do liści.

[12] Bengio, S., Weston, J., i Grangier, D. (2010). Label embedding trees for large multi-class tasks. In *NIPS*, pages 163-171. Curran Associates, Inc

Jasinka, K., Dembczynski, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., i Hüllermeier, E. (2016). Extreme F-measure maximization using sparse probability estimates

Drzewa etykiet

- Bazują na b-arnych drzewach.[12]



- Każdy **liść** odpowiada jednej etykiecie.
- **Wewnętrzne** klasyfikatory decydują, czy przejść do potomków.
- **Liść** może zawierać ostateczny klasyfikator dla danej etykiety.
- Przykład testowy może odwiedzić **wiele ścieżek** od korzenia do liści.

[12] Bengio, S., Weston, J., i Grangier, D. (2010). Label embedding trees for large multi-class tasks. In *NIPS*, pages 163-171. Curran Associates, Inc

Jasinka, K., Dembczynski, K., Busa-Fekete, R., Pfannschmidt, K., Klerx, T., i Hüllermeier, E. (2016). Extreme F-measure maximization using sparse probability estimates

Wyniki eksperymentalne

	#etykiet	#cech	#test.	#trening.	przk./etyk.	etyk./przk.
RCV1	2456	47236	155962	623847	1218.56	4.79
AmazonCat	13330	203882	306782	1186239	448.57	5.04
Wiki10	30938	101938	6616	14146	8.52	18.64
Delicious	205443	782585	100095	196606	72.29	75.54
WikiLSHTC	325056	1617899	587084	1778351	17.46	3.19
Amazon	670091	135909	153025	490449	3.99	5.45

Tabela: Zbiory danych z repozytorium `XMLRepository`.[13]

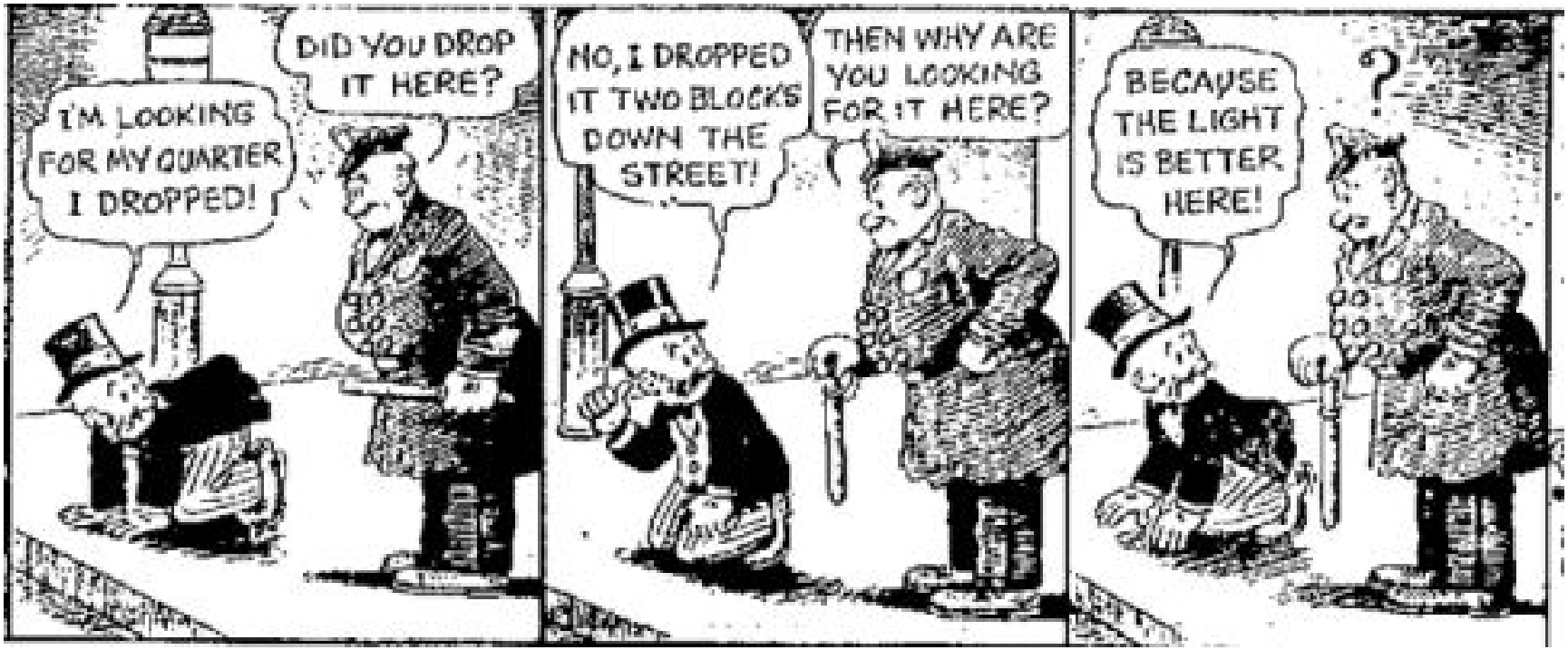
[13] `http://research.microsoft.com/en-us/um/people/manik/downloads/XC/XMLRepository.html`

Wyniki eksperymentalne

	PLT			FastXML		
	P@1	P@3	P@5	P@1	P@3	P@5
RCV1	90.46	72.4	51.86	**91.13**	**73.35**	**52.67**
AmazonCat	91.47	75.84	61.02	**92.95**	**77.5**	**62.51**
Wiki10	**84.34**	**72.34**	**62.72**	81.71	66.67	56.70
Delicious	**45.37**	**38.94**	35.88	42.81	38.76	**36.34**
WikiLSHTC	45.67	29.13	21.95	**49.35**	**32.69**	**24.03**
Amazon	**36.65**	**32.12**	**28.85**	34.24	29.3	26.12

	PLT					FastXML			
	uczen. [min]	test. [ms]	b	wysok. drzewa	#ilocz. skalar.	uczen. . [min]	test. [ms]	wysok. drzewa	#ilocz. skalar.
RCV1	**64**	0.22	32	2,25	**43,46**	78	0.96	14.95	747
AmazonCat	**96**	0.17	16	3,43	**54,39**	561	1.14	17.44	871
Wiki10	290	2.66	32	2,98	**121,98**	**16**	3.00	10.83	541
Delicious	1327	32.97	2	17,69	11779,65	**458**	4.01	14.79	**739**
WikiLSHTC	**653**	3.00	32	3,66	**622,27**	724	1.17	18.01	900
Amazon	**54**	0.99	8	6,45	**374,30**	422	1.39	15.92	796

Czy szukamy rozwiązania w dobrym miejscu?



Rysunek: [14] Podobny rysunek użyty wcześniej przez Aselę Gunawardanę.[15]

[14] Oryginał: Florence Morning News, Mutt and Jeff Comic Strip, Page 7, Florence, South Carolina,1942

[15] Asela Gunawardana, *Evaluating Machine Learned User Experiences*. Extreme Classification Workshop. NIPS 2015

Wyzwania

- **Redukcja złożoności obliczeniowej**:
  - czas vs. pamięć
  - #przykładów vs. #cech vs. #etykiet
  - uczenie vs. walidacja vs. testowanie
- **Poprawa wydajności predykcyjnej**:
  - Miary wydajności: błąd Hamminga, prec@k, miara F, ...
  - Statystyczna teoria uczenia dla dużego m.
  - Uczenie rzadkich etykiet.

Koniec

- Więcej na stronie:

  `http://www.cs.put.poznan.pl/kdembczynski`