# Jak sobie radzić z etykietami, gdy eksperci są mało wiarygodni?

Łukasz Czekaj

# Plan prezentacji

- Etykiety dostarczone przez „tłum” (crowdsourcing)
- Modelowanie jakości „ekspertów”
- Uczenie z uwzględnieniem modelu eksperta

# Etykiety dostarczone przez „tłum”

## Budowanie zbioru danych

- Coś trudniejszego niż klasyfikacja kot vs pies,
  np. analiza danych medycznych
- Problem ze znalezieniem eksperta:
  - lekarze są za drodzy; nie mają czasu; nie chcą wykonywać żmudnej i rutynowej pracy;
  - technicy – mniej wykwalifikowani ale tańsi i łatwiej dostępni

# Etykiety dostarczone przez „tłum”

## Budowanie zbioru danych

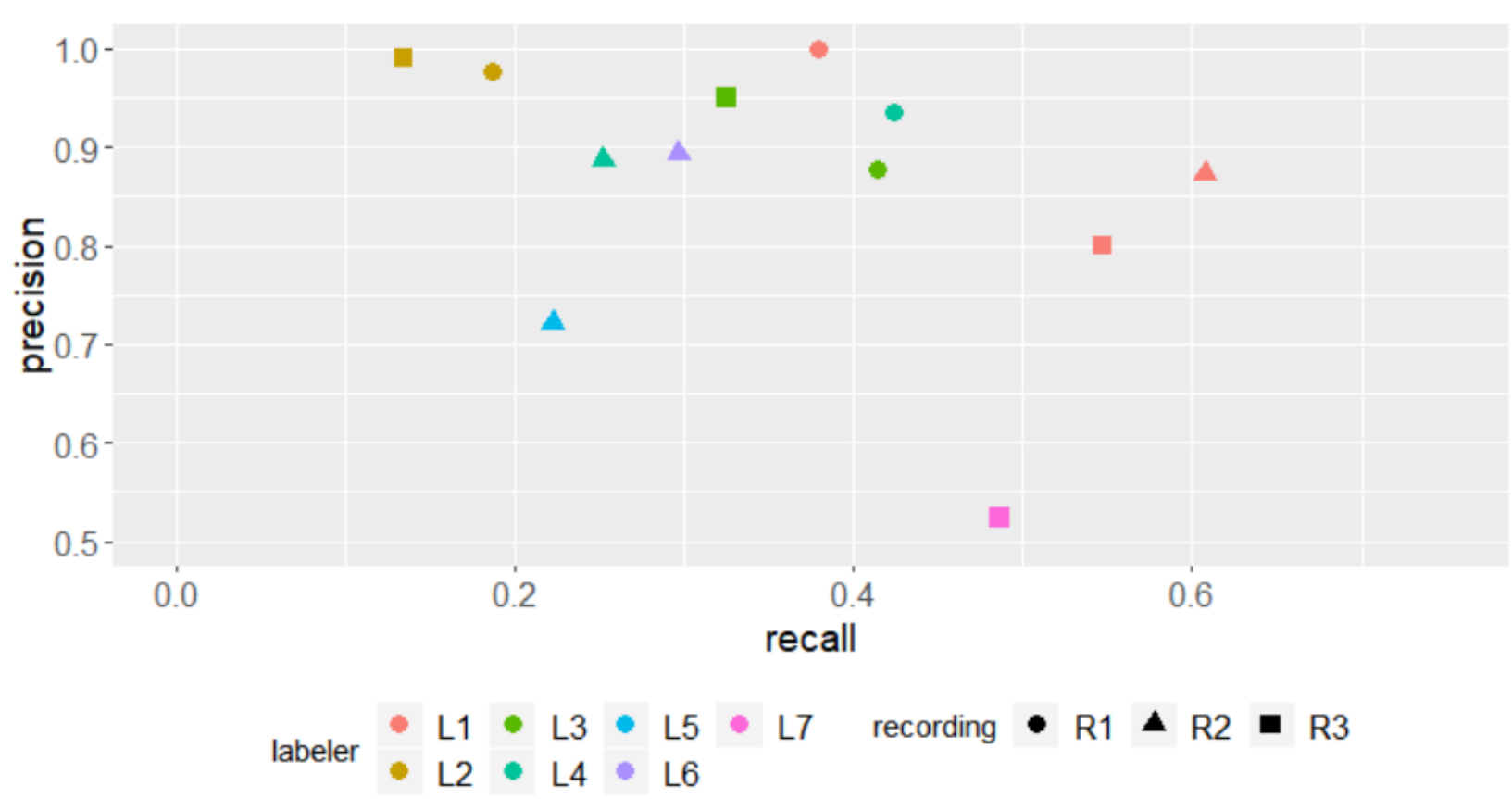
- Eksperci nie są zgodni:
  - różne szkoły, różne progi (czułość, precyzja)
- Eksperci też się mylą:
  - zmęczenie, motywacja
- Nie do końca możemy ufać etykietom

# Etykiety dostarczone przez „tłum”

## Budowanie zbioru danych

- Masowe etykietowanie danych:
  - Amazon Mechanical Turk
  - Figure Eight (CrowedFlower)
  - Zooniverse

# Etykiety dostarczone przez „tłum”
# Zgodność między ekspertami



# Etykiety dostarczone przez „tłum”

## Budowanie zbioru danych

- Model danych:
  - $\{(X_i,\{y_{i,j}\}_j,z_i=?)\}_i$
  - i – indeks przykładu uczącego
  - j – indeks eksperta
  - $X_i$ – wektor cech
  - $y_{i,j}$ – etykieta którą ekpert j przydzielił dla przykładu i, $y_{i,j}=\varnothing$ jeśli ekspert nie j nie etykietował przykładu i
- Jeden przykład może mieć przypisane wiele etykiet
- Etykiety nie muszą być zgodne
- Nie wiemy jaka jest „prawdziwa” etykieta $z_i$

# Etykiety dostarczone przez „tłum”

## Alokacja pracy

- Zbiór treningowy:
  - Wielokrotne adnotacje każdego przykładu – mniej przykładów treningowych
  - Pojedyncze adnotacje każdego przykładu – mniejsza jakość etykiet

    [Sheng, Victor et all. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 614-622. 10.1145/1401890.1401965. ]
  - Uczenie aktywne (ale to temat na zupełnie inną opowieść)

# Etykiety dostarczone przez „tłum”

## Alokacja pracy

- Zbiór testowy (jak ustalamy „prawdziwe” etykiety):
  - Wielokrotne adnotacje każdego przykładu – liczba osób w zależności od jakości adnotujących (3-7), mniejsza liczba przykładów niż w zbiorze treningowym
  - Każdy przykład adnotują 2 osoby, w razie niezgodności konflikt rozwiązuje „super ekspert” - mniejsze koszty
  - Konsensus – każdy przykład testowy jest omawiany na panelu ekspertów, etykieta na podstawie wypracowanego konsensusu

# Etykiety dostarczone przez „tłum”

## Alokacja pracy

Jak to robi Adnrew Ng?

[https://arxiv.org/pdf/1707.01836.pdf]

- Trening: 64,121 ECG zapisów; 29,163 pacjentów; 30s każdy; etykiety dostarczane przez pojedynczego eksperta;
- Testy: 336 zapisów; 328 pacjentów; etykiety na podstawie konsensusu (dyskusja) 3 ekspertów;

# Modelowanie jakości eksperta (tylko etykiety, brak X)

Estymacja macierz konfuzji $p_j(Y|Z)$

[A. P. Dawid and A. M. Skene, Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm (1997)]

- Prawdziwe etykiety ($Z_i$) są znane:

  $p(\{dane\}) = p(\{\{Y_{i,j}\},Z_i\}) = \Pi_{i,j} p_j(Y_{i,j}|Z_i)$

  $p_j(\{Y_{i,j}\},Z_i)$ – rozkład Bernoulliego

  $\alpha_j = p_j(Y=1|Z=1)$

  $\beta_j = p_j(Y=0|Z=0)$

  MLE: $\text{argmax}\{\alpha_j,\beta_j\}\ p(\{\{Y_{i,j}\},Z_i\})$

# Modelowanie jakości eksperta (tylko etykiety, brak X)

Estymacja macierz konfuzji $p_j(Y|Z)$

[A. P. Dawid and A. M. Skene, Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm (1997)]

- Prawdziwe etykiety ($Z_i$) nie są znane:

  $p(\{dane\})= p(\{\{Y_{i,j}\}\})=\Pi_{i,j}\sum_z p_j(Y_{i,j}|Z_i)p_i(Z_i)$
- Wiele adnotacji dla jednego przykładu
- $p_i \in [0,1]$ => Bayes (montecarlo)
- $p_i \in \{0,1\}$ => algorytm EM:
  - Inicjalizacja: losowo przypisujemy {0,1} do $p_i$
  - M: argmax$\{\alpha_j,\beta_j\}$ $p(\{\{Y_{i,j}\},Z_i\})$; ustalone $p_i$
  - E: argmax$\{p_i\}$ $p(\{\{Y_{i,j}\},Z_i\})$; ustalone $\alpha_j,\beta_j$

# Modelowanie jakości eksperta (dane etykiety oraz X)

Estymacja macierz konfuzji $p_j(Y|Z)$ lub $p_j(Y|Z,X)$

Estymacja klasyfikatora $p(Z|X)$

[V. C. Raykar, etal., Learning From Crowds (2010);
Yan Yan, etal., Modeling annotator expertise: Learning when everybody knows a bit of something (2010);
Ashish Khetan, etal., Learning from noisy singly-labeled data (2018);]

- Prawdziwe etykiety ($Z_i$) nie są znane
- Estymujemy etykiety ($Z_i$) na podstawie modelu $p_i(Z_i|X_i)$
- Wystarczą pojedyncze adnotacje

# Modelowanie jakości eksperta (dane etykiety oraz X)

Estymacja macierz konfuzji $p_j(Y|Z)$ lub $p_j(Y|Z,X)$

Estymacja klasyfikatora $p(Z|X)$

[V. C. Raykar, etal., Learning From Crowds (2010);
Yan Yan, etal., Modeling annotator expertise: Learning when everybody knows a bit of something (2010);
Ashish Khetan, etal., Learning from noisy singly-labeled data (2018);]

- Macierz konfuzji zależy tylko od klasy przykładu ($Z_i$)

  $p(\{dane\}) = p(\{X_i,\{Y_{i,j}\}\}) = \Pi_{i,j} p_j(Y_{i,j}|Z_i) p_i(Z_i|X_i)$

- Dla każdej klasy są przykłady trudniejsze i łatwiejsze

  $p(\{dane\}) = p(\{X_i,\{Y_{i,j}\}\}) = \Pi_{i,j} p_j(Y_{i,j}|Z_i,X_i) p_i(Z_i|X_i)$

# Modelowanie jakości eksperta (dane etykiety oraz X)

Estymacja macierz konfuzji $p_j(Y|Z)$ lub $p_j(Y|Z,X)$

Estymacja klasyfikatora $p(Z|X)$

[V. C. Raykar, etal., Learning From Crowds (2010);
Yan Yan, etal., Modeling annotator expertise: Learning when everybody knows a bit of something (2010);
Ashish Khetan, etal., Learning from noisy singly-labeled data (2018);]

- Algorytm EM:
  - M: MLE dla $p_j$, $p(Z|X)$;

    proste modele – analitycznie
    bardziej złożone (xgboost) – numerycznie
  - E: $p_i(Z_i)$,$p_{i,j}(Z_{i,j})$ – na podstawie tw. Bayesa

# Modelowanie jakości eksperta (dane etykiety oraz X)

Estymacja macierz konfuzji $p_j(Y|Z)$ lub $p_j(Y|Z,X)$

Estymacja klasyfikatora $p(Z|X)$

[V. C. Raykar, etal., Learning From Crowds (2010);
Yan Yan, etal., Modeling annotator expertise: Learning when everybody knows a bit of something (2010);
Ashish Khetan, etal., Learning from noisy singly-labeled data (2018);]

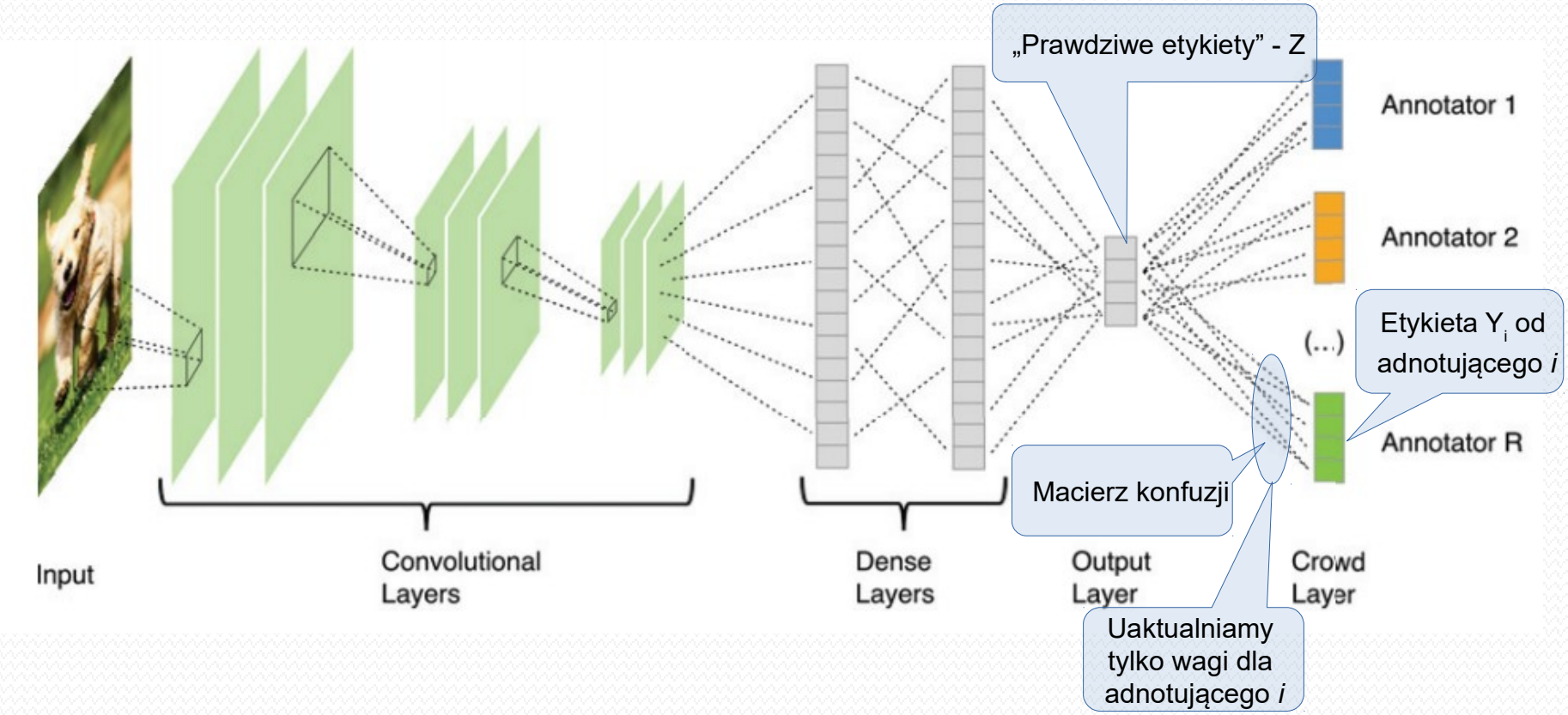
- Algorytm EM:
  $p_i(Z_i)$, $p_{i,j}(Z_{i,j})$ => modyfikacja funkcji straty $p(Z|X)$
  - $\{ X_i, Y_{i,j}, w = \max\{Y_{i,j}(p_{i,j}(1)-p_{i,j}(0)),0\} \}$ - dla każdego przykładu wiele adnotacji
  - $\{ X_i, Z^{est}_i, w = \max\{Z^{est}_i(p_i(1)-p_i(0)),0\} \}$ - pojedyncze adnotacje dla przykładu
  - $\{ X_i, 0, w = p_i(0)\}$, $\{X_i, 1, w = p_i(1)\}$ – p. a posteriori

# Modelowanie jakości eksperta (dane etykiety oraz X)

Estymacja macierzy konfuzji p(Y|Z) w epoce DeepLearning

CrowdLayer (uczenie w jednym etapie, bez EM)

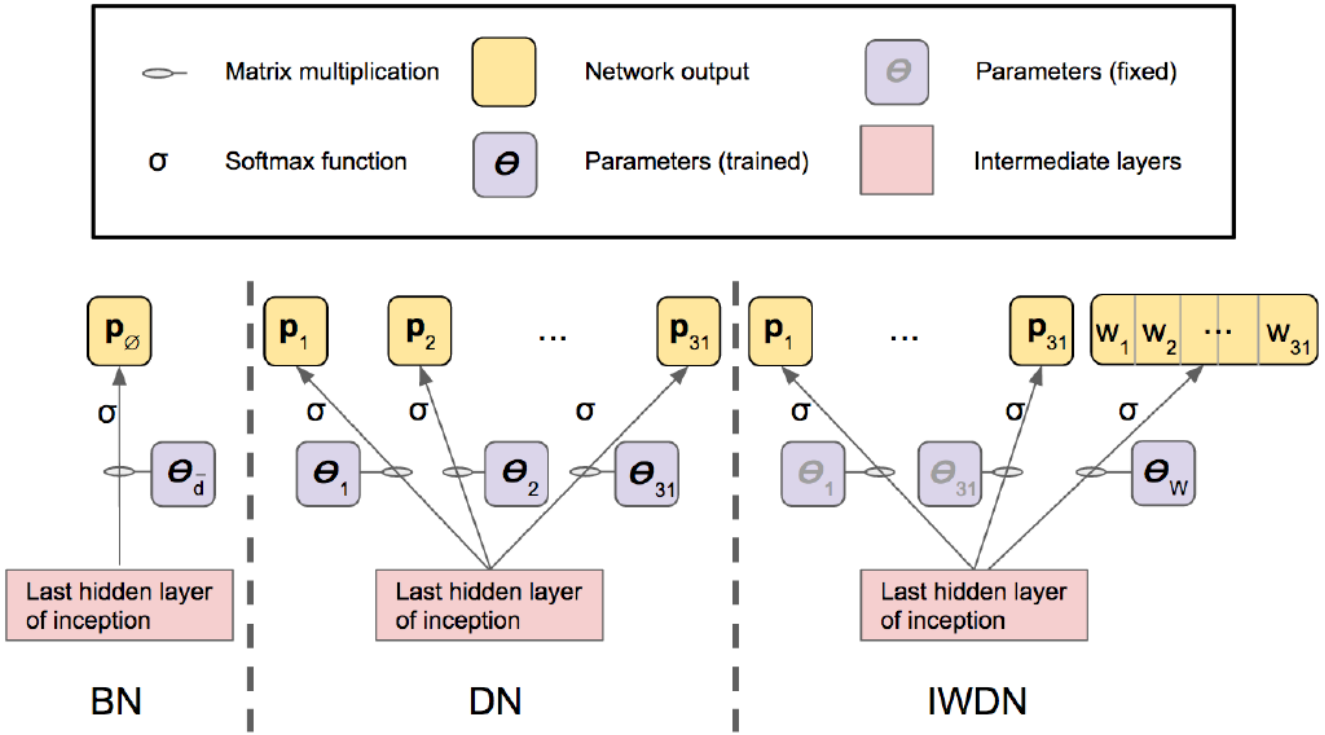
[F. Rodrigues, F. Pereira, Deep learning from crowds, arXiv:1709.01779 (2017)]



# Uczenie z uwzględnieniem modelu Eksperta

Sieci Głębokie – p(Y|X) + głosowanie
(uczenie 2 etapowe oparte na „reprezentacji ukrytej”)

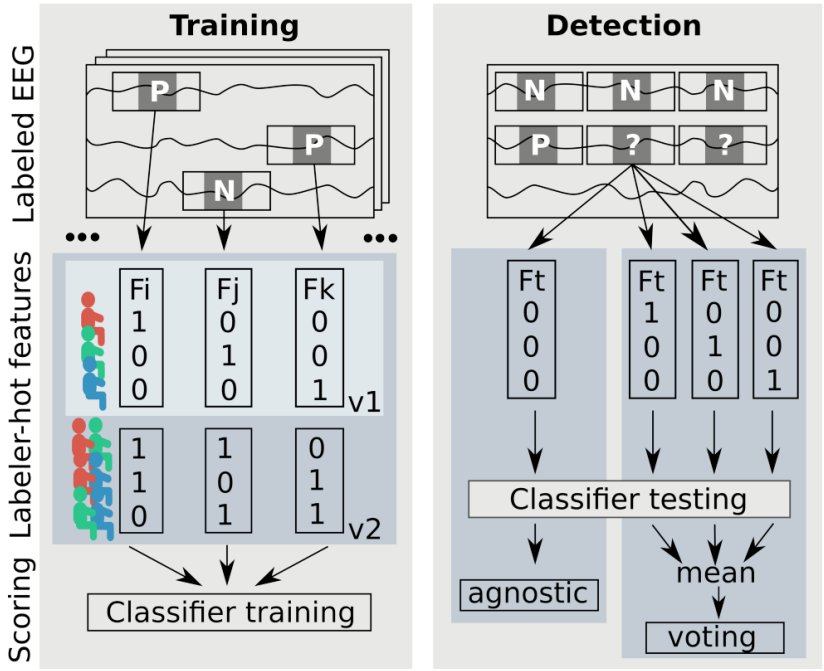
[M. Y. Guan, V. Gulshan, A. M. Dai, G. E. Hinton, Who Said What: Modeling Individual Labelers Improves Classification, arXiv:1703.08774v2 (2018)]

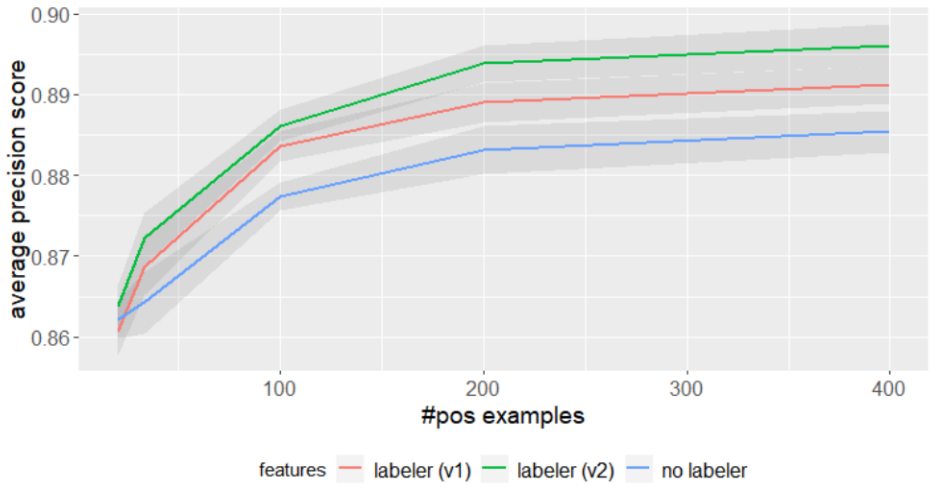


# Uczenie z uwzględnieniem modelu Eksperta

XGBoost – p(Y|X, i) + głosowanie
(identyfikator eksperta jako parametr modelu)

[L. Czekaj, W. Ziembla, P. Jezierski, P. Swiniarski, A. Kolodziejak, P. Ogniewski, P. Niedbalski, A. Jezierska, D. Wesierski;
Labeler-hot Detection of EEG Epileptic Transients, arXiv:1903.04337 (2019)]

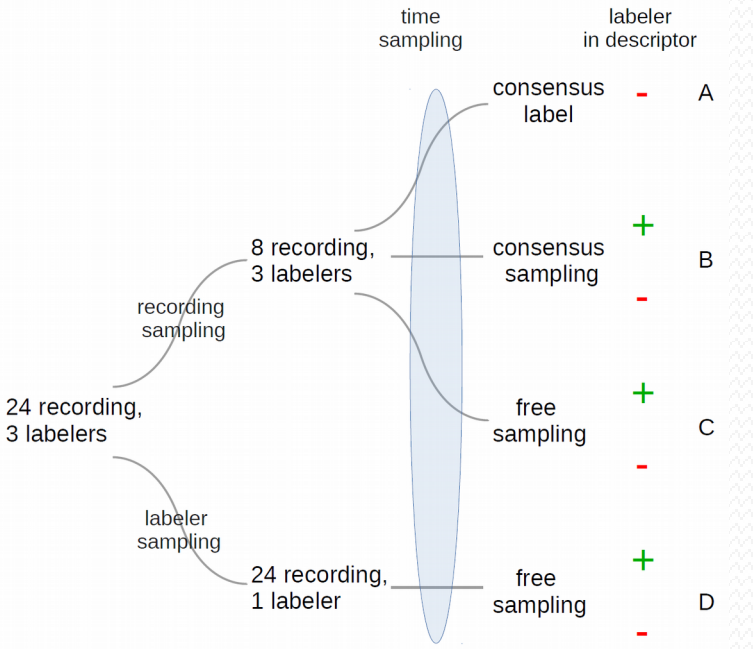


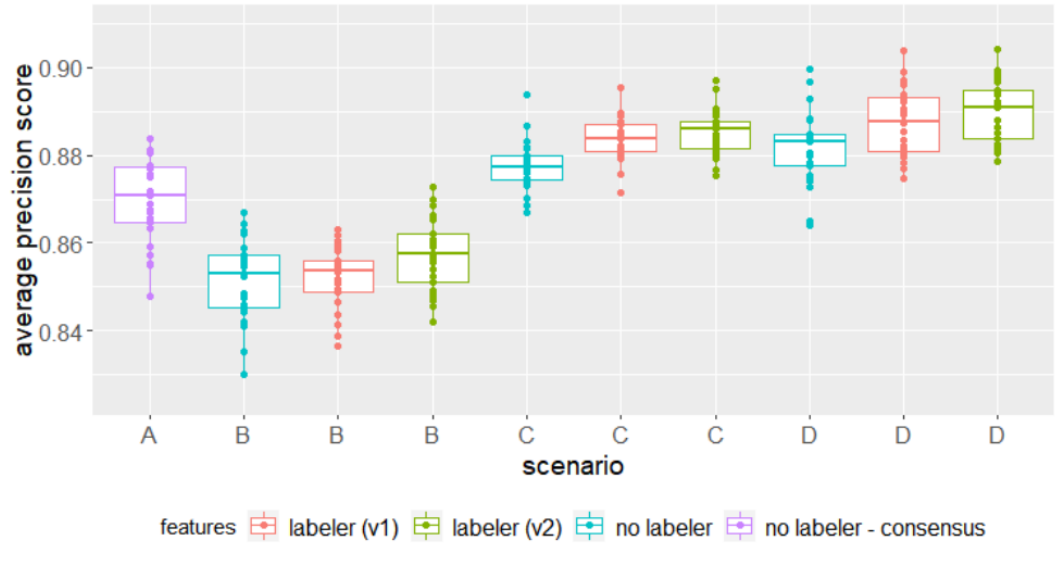



# Uczenie z uwzględnieniem modelu Eksperta

## Strategie zbierania danych, modelowanie eksperta vs konsensus

[L. Czekaj, W. Ziembla, P. Jezierski, P. Swiniarski, A. Kolodziejak, P. Ogniewski, P. Niedbalski, A. Jezierska, D. Wesierski; Labeler-hot Detection of EEG Epileptic Transients, arXiv:1903.04337 (2019)]



# Podsumowanie

Integracja etykiet od wielu ekspertów

- Niska zgodność ekspertów;
- Eksperci etykietują różne przykłady, nie możemy ich bezpośrednio porównać;
- Za dużo ekspertów, żeby „ręcznie” oceniać ich jakość;
- Duża rozbieżność między jakością ekspertów – największy zysk z modelu – model automatycznie wyłapuje słabych;
- Równoległe etykietowanie – proces nie pozwala na wyłączenie słabych ekspertów w trakcie zbierania danych – więc warto czegoś się od nich nauczyć;