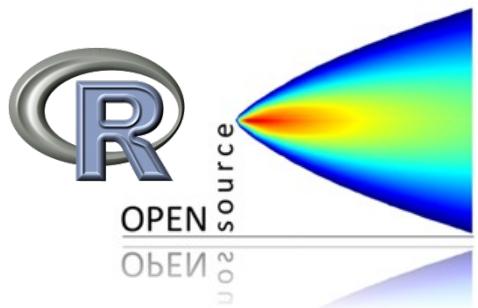


ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Department of Climatology | Faculty of Geosciences



FasteR - kiedy warto przenieść obliczenia na superkomputery?

- przykładowe rozwiązania z perspektywy użytkownika desktopa i PCSS/WCSS

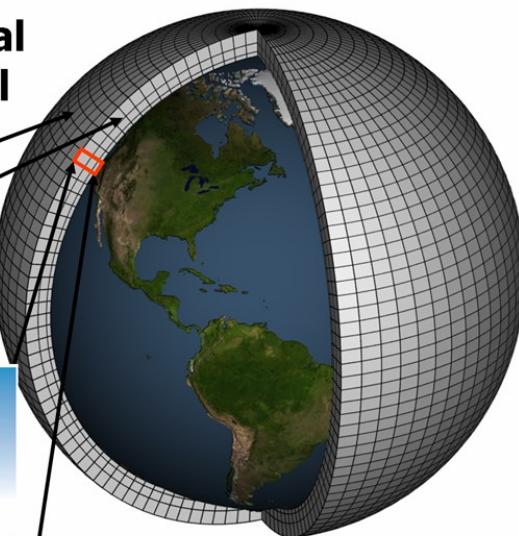
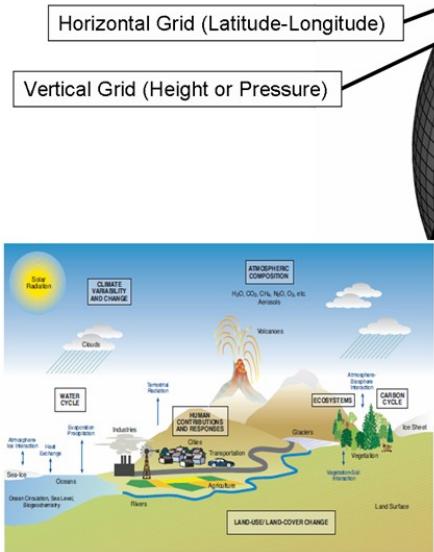
Bartosz Czernecki



Wrocławskie Centrum Sieciowo-Superkomputerowe

Dane modelowe (prognoza/reanalizy)

Schematic for Global Atmospheric Model



Global Forecasting System (GFS):

Siatka: 0.25x0.25 stopnia (1440 x 1440)

Liczba warstw: 64 warstwy w pionie

Liczba zmiennych (elementów meteo): 136

Horyzont prognozy: 0-120 h → co 1h,
123-384h → co 3h

Liczba prognoz na dobę: 4
(00, 06, 12, 18 UTC)

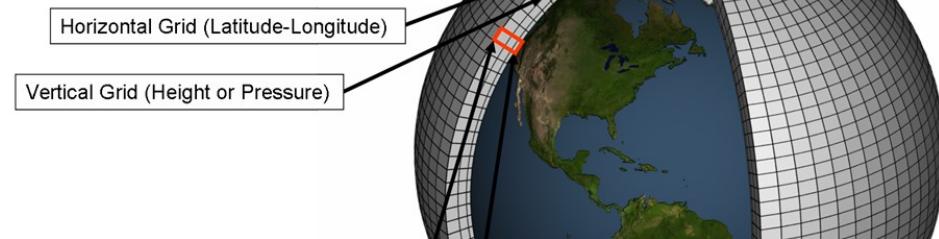
Potencjalna liczba wartości w prognozie na:

$$1\text{h: } = 1440 * 1440 * 64 * 136 = 18,048,614,400$$

$$3 \text{ dni: } \sim 1,300,000,000,000$$

Dane modelowe (prognoza/reanalizy)

Schematic for Global Atmospheric Model



Ciekawostka: ile GB zajmuje (w CSV) prognoza GFS na 90h?

- dla Polski (~3.2 GB, 41,8 mln rzędów)
- dla Europy (~117 GB)

Global Forecasting System (GFS):

Siatka: 0.25x0.25 stopnia (1440 x 1440)

Liczba warstw: 64 warstwy w pionie

Liczba zmiennych (elementów meteo): 136

Horyzont prognozy: 0-120 h → co 1h,
123-384h → co 3h

Liczba prognoz na dobę: 4
(00, 06, 12, 18 UTC)

Potencjalna liczba wartości w prognozie na:

$$1\text{h: } = 1440 * 1440 * 64 * 136 = 18,048,614,400$$

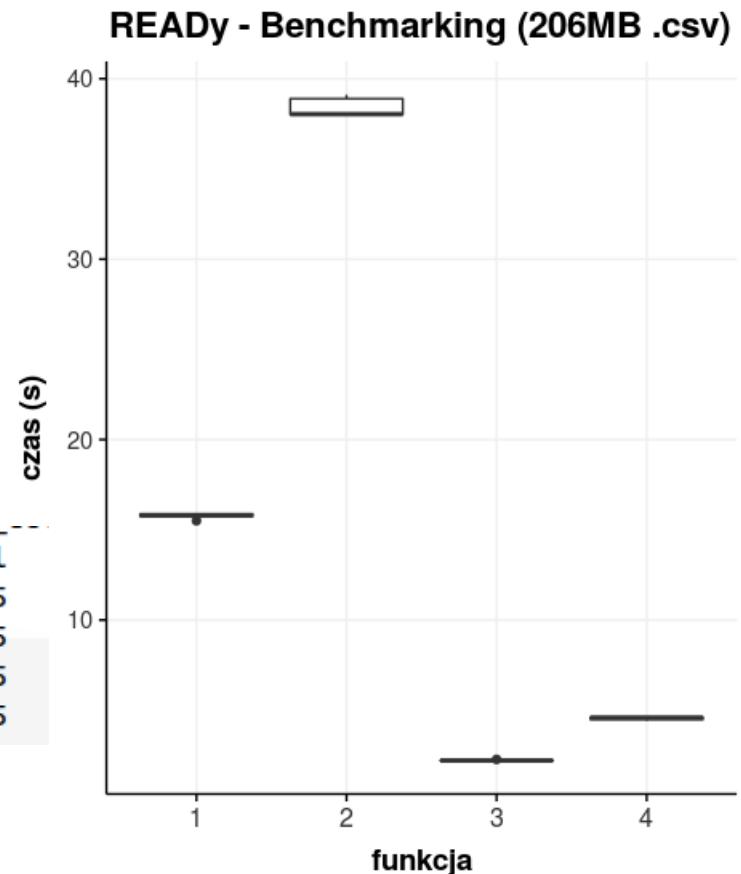
$$3 \text{ dni: } \sim 1,300,000,000,000$$

Problem: ::wczyt_danych read_XXX::benchmark

Dane: domena PL (1125 punktów * 90h) →
206MB

- 1) `read.csv()`
- 2) `read.csv(colClasses = zdefiniowane)`
- 3) `data.table::fread()`
- 4) `readr::read_delim()`

	<code>min</code>	<code>lq</code>	<code>mean</code>	<code>median</code>	<code>uq</code>	<code>max</code>	<code>neval</code>
	15.501717	15.741768	15.765353	15.806758	15.878219	15.898306	5
	37.986519	38.021926	38.410795	38.036555	38.894263	39.114712	5
	2.203171	2.208461	2.224006	2.215797	2.218906	2.273693	5
	4.427898	4.488693	4.553214	4.588538	4.612542	4.648399	5



Możliwe przyspieszenie:
 $38.4/2.2 = \sim 17.4x$



Diagram termodynamiczny Skew-T CAPE/CIN

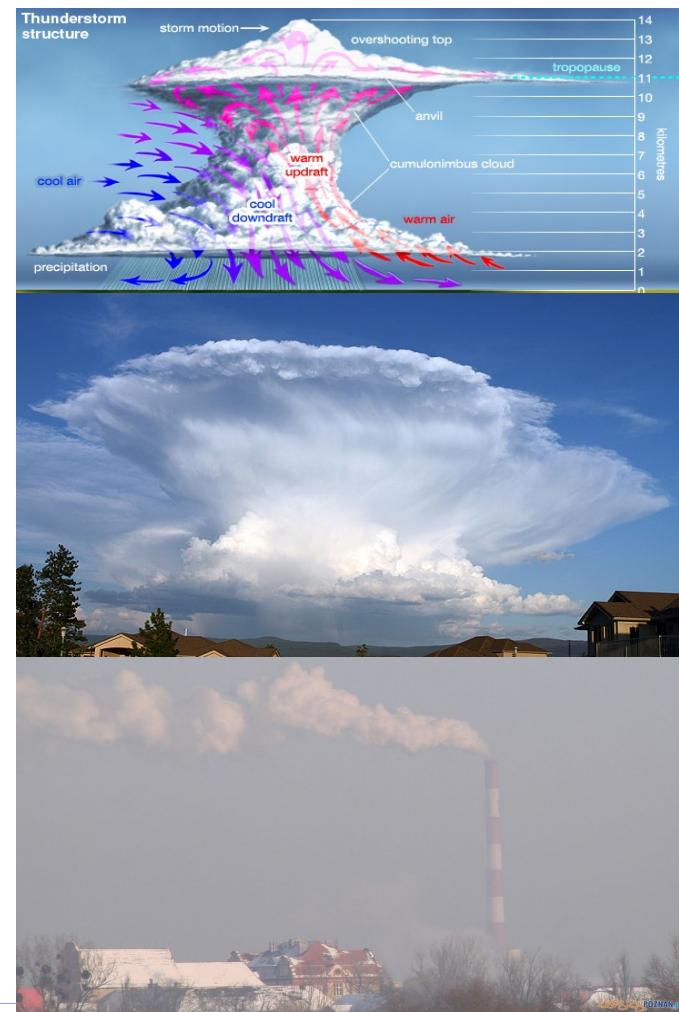
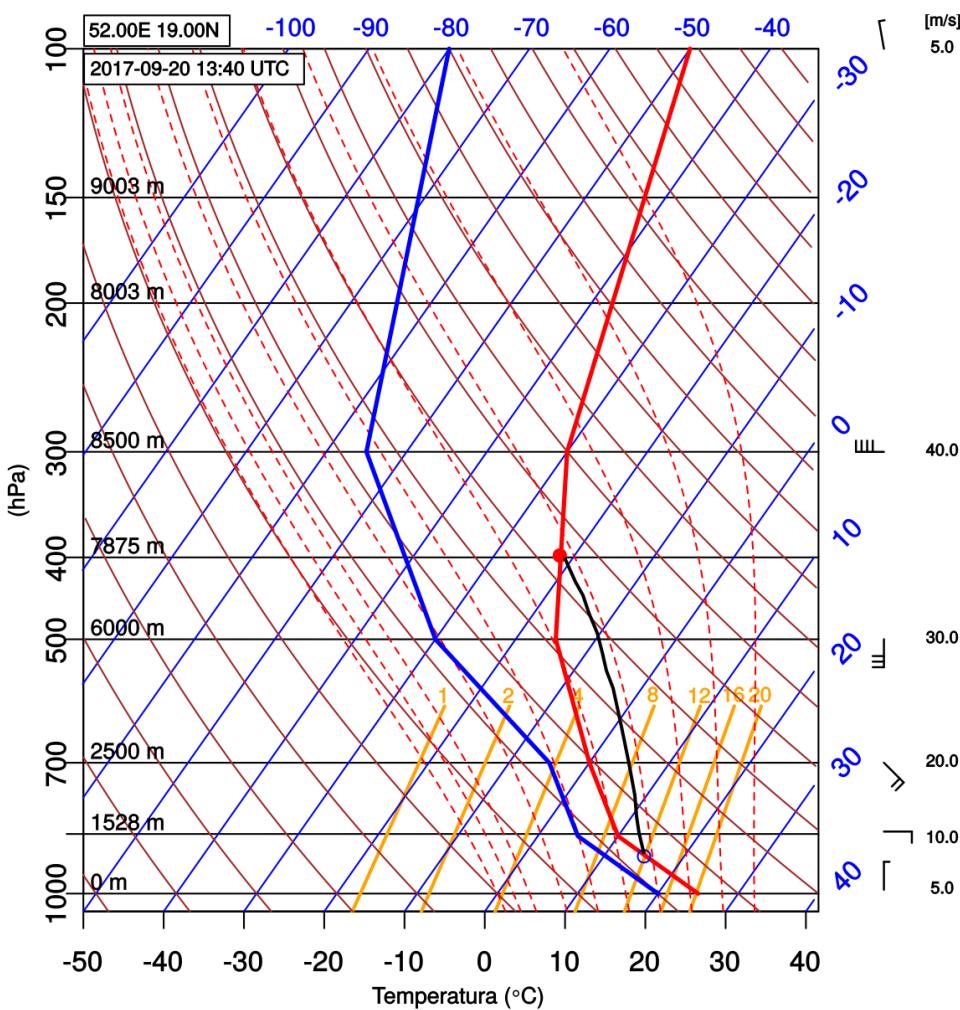
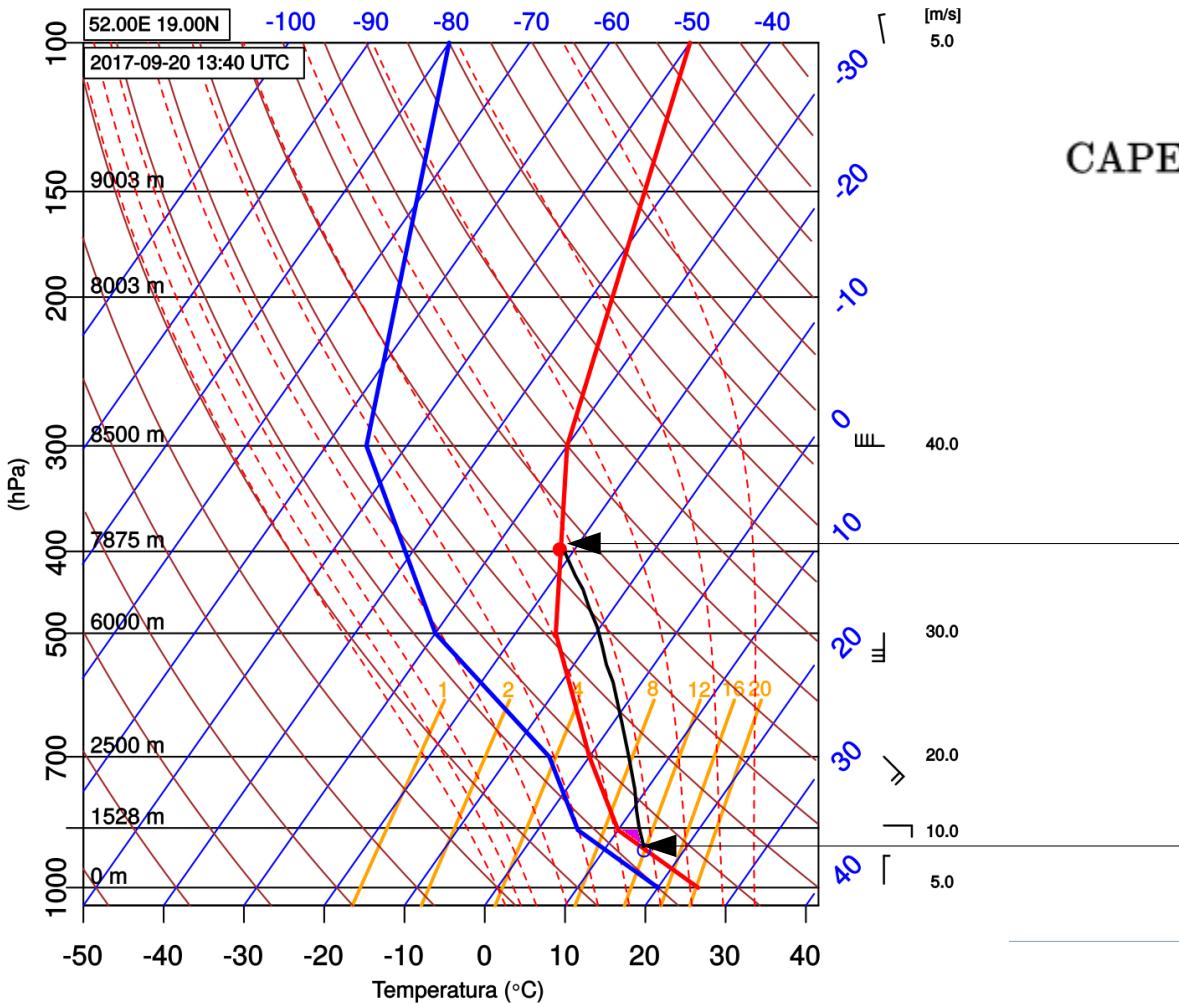


Diagram termodynamiczny Skew-T CAPE/CIN

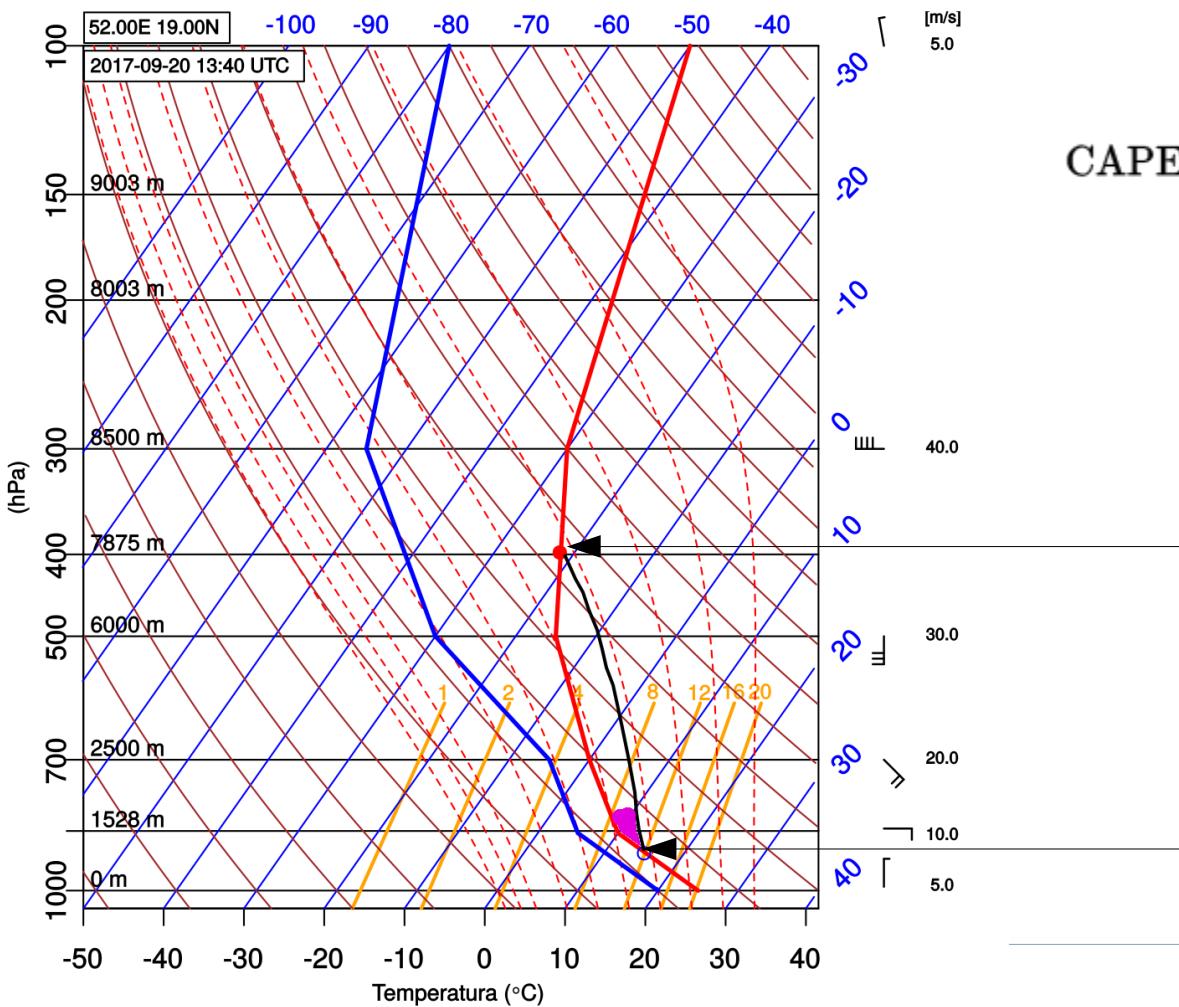


$$\text{CAPE} = \int_{z_f}^{z_n} g \left(\frac{T_{v,\text{parcel}} - T_{v,\text{env}}}{T_{v,\text{env}}} \right) dz$$

Equilibrium level (EL)

**Level of free convection
(LFC)**

Diagram termodynamiczny Skew-T CAPE/CIN



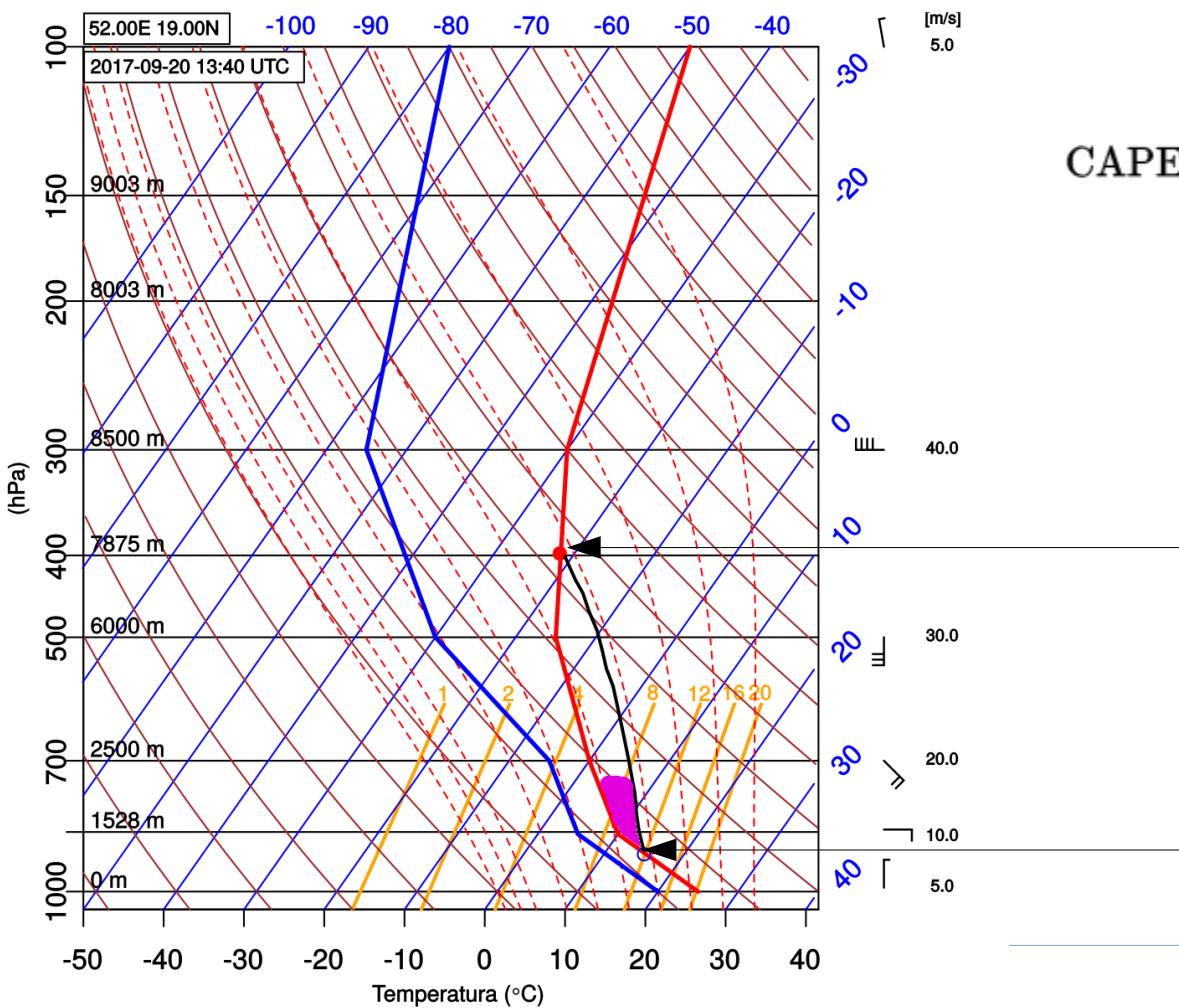
$$\text{CAPE} = \int_{z_f}^{z_n} g \left(\frac{T_{v,\text{parcel}} - T_{v,\text{env}}}{T_{v,\text{env}}} \right) dz$$

Equilibrium level (EL)

**Level of free convection
(LFC)**



Diagram termodynamiczny Skew-T CAPE/CIN

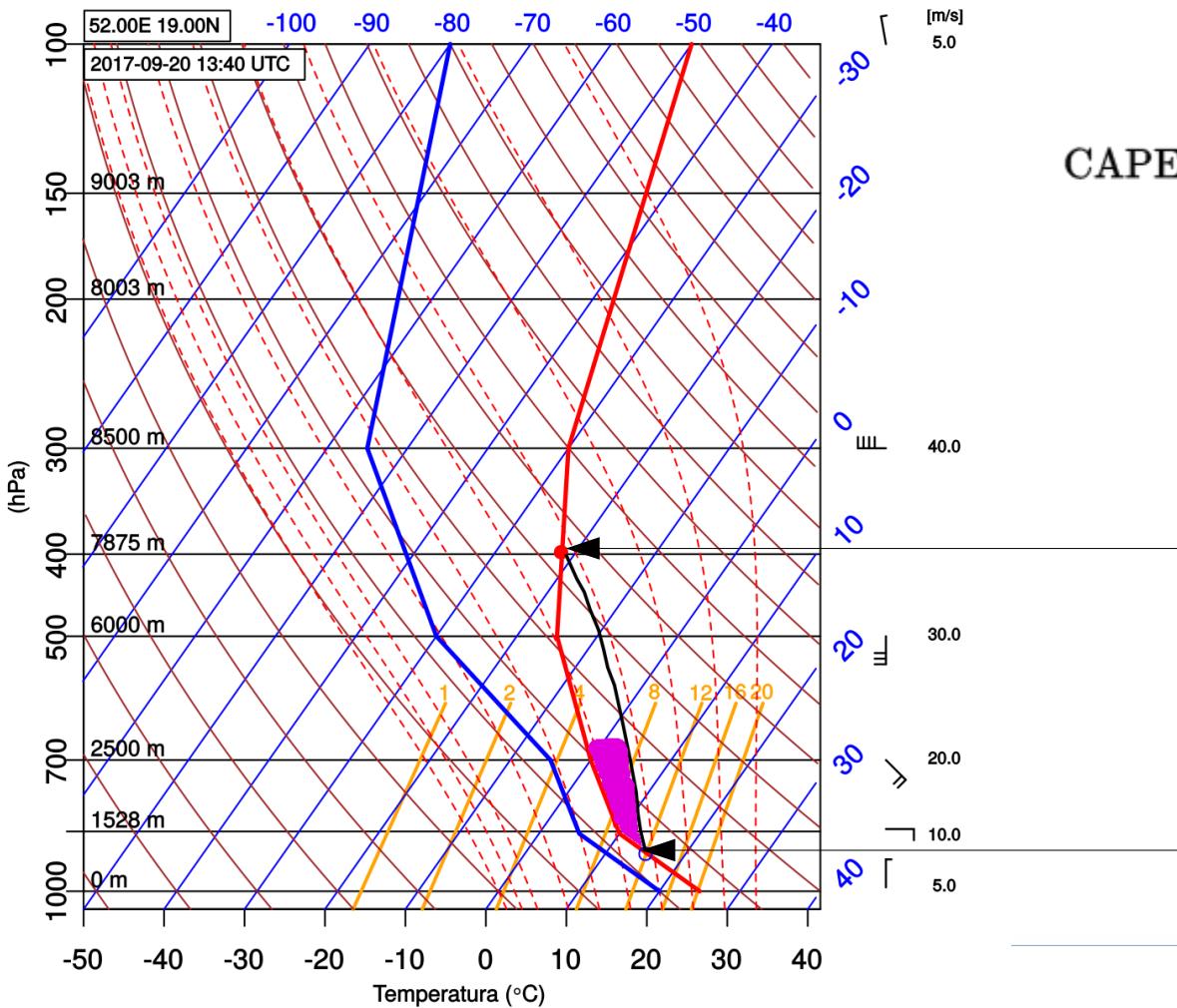


$$\text{CAPE} = \int_{z_f}^{z_n} g \left(\frac{T_{v,\text{parcel}} - T_{v,\text{env}}}{T_{v,\text{env}}} \right) dz$$

Equilibrium level (EL)

Level of free convection (LFC)

Diagram termodynamiczny Skew-T CAPE/CIN

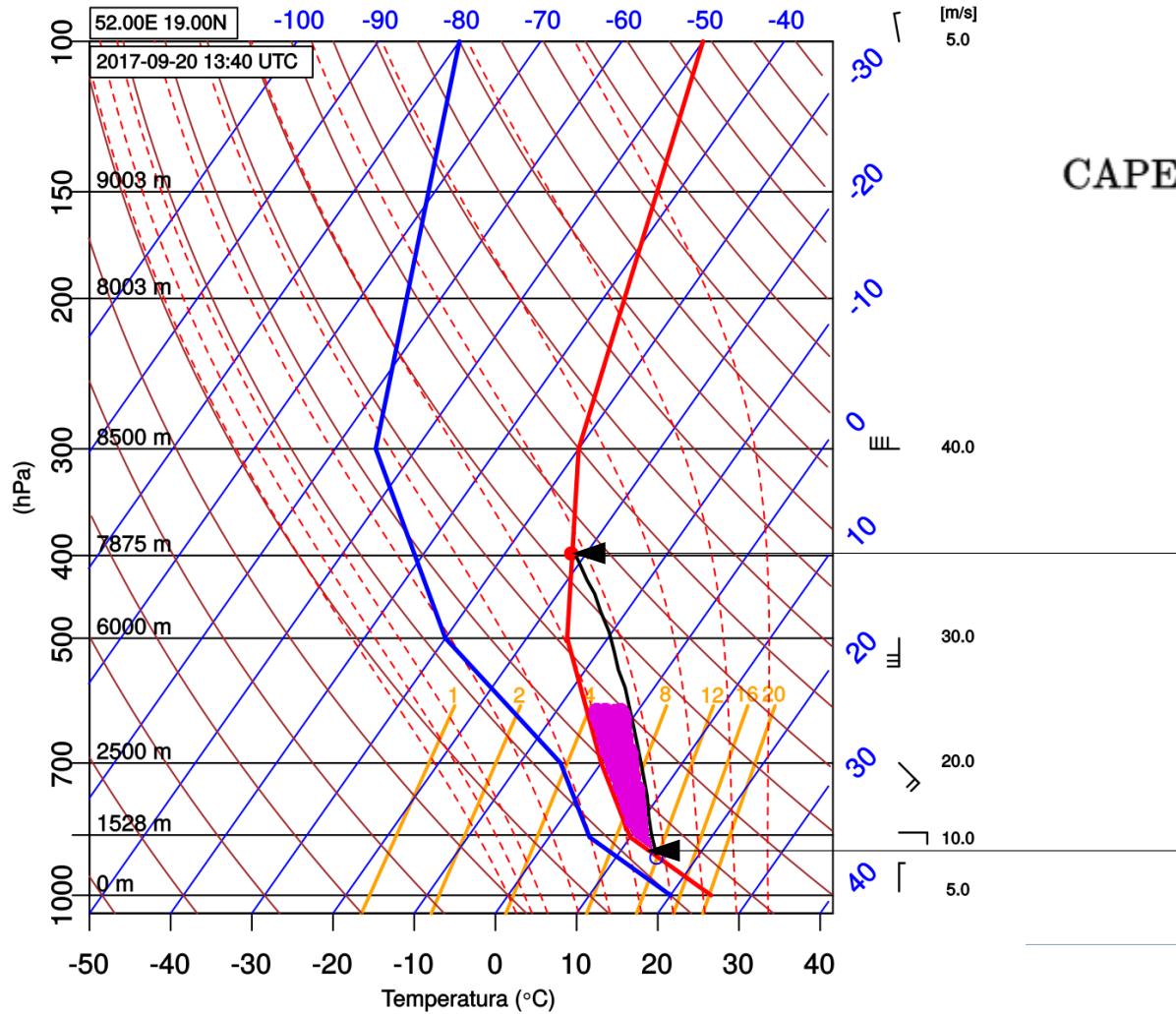


$$\text{CAPE} = \int_{z_f}^{z_n} g \left(\frac{T_{v,\text{parcel}} - T_{v,\text{env}}}{T_{v,\text{env}}} \right) dz$$

Equilibrium level (EL)

**Level of free convection
(LFC)**

Diagram termodynamiczny Skew-T CAPE/CIN

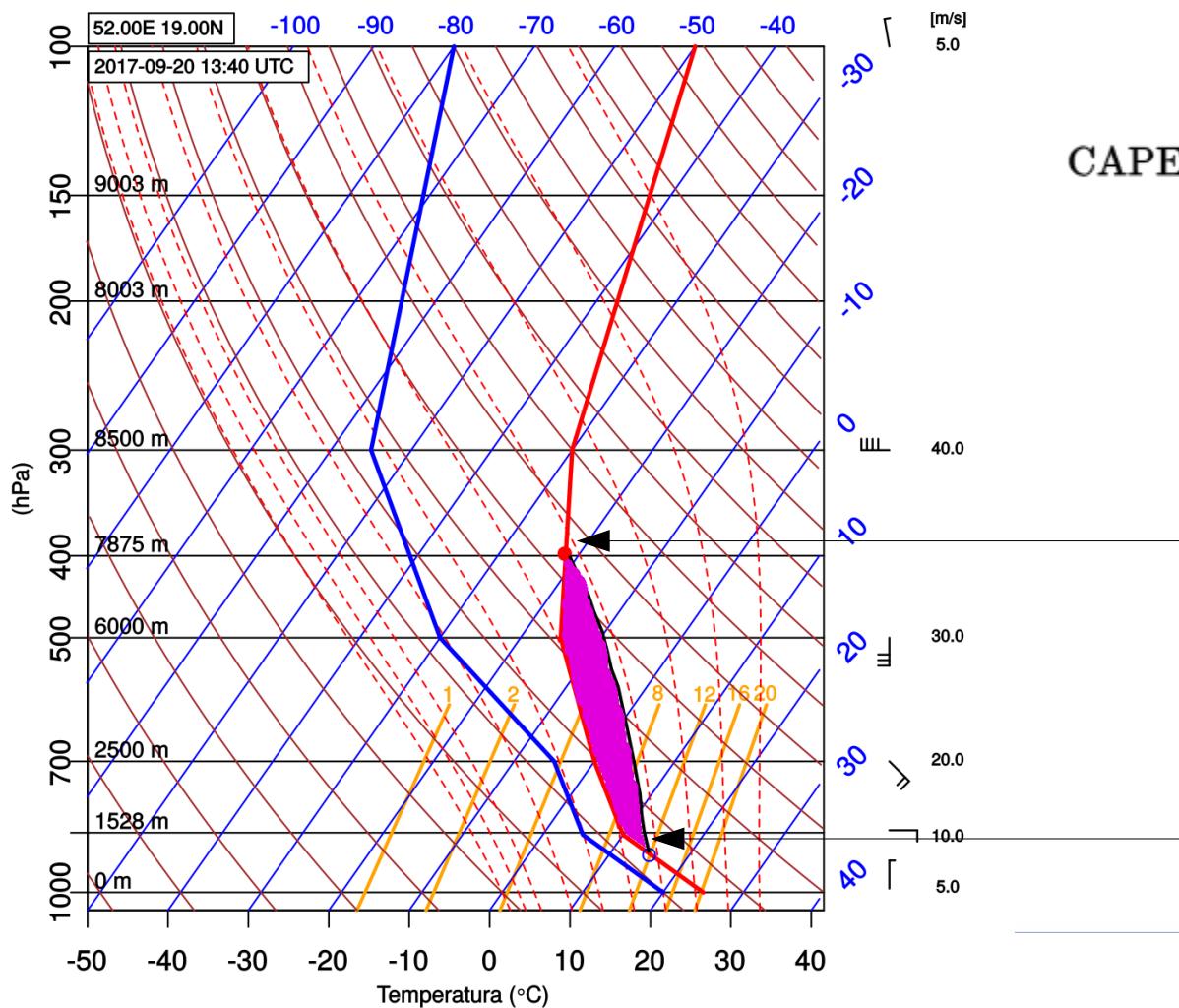


$$\text{CAPE} = \int_{z_f}^{z_n} g \left(\frac{T_{v,\text{parcel}} - T_{v,\text{env}}}{T_{v,\text{env}}} \right) dz$$

Equilibrium level (EL)

**Level of free convection
(LFC)**

Diagram termodynamiczny Skew-T CAPE/CIN



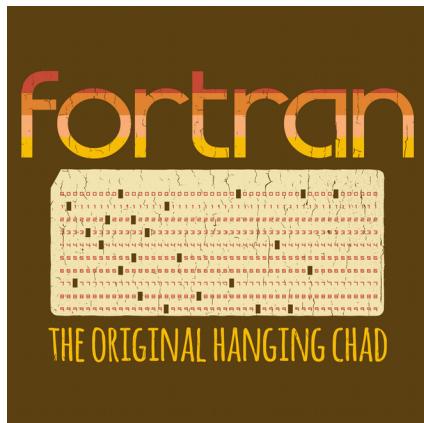
FasteR:: FORTRAN (90) + R

1) Pierwotna wersja kodu na obliczanie CAPE/CIN dla pojedynczego profilu pionowego (~2010) (bez rysowania)



~450 linii z pętlami → 0,78 s / sondaż
po wektoryzacji → 0,49 s / sondaż

2) R + Fortran



```
bash#> R CMD SHLIB sounding.f  
R > dyn.load(„sounding.so”)  
... 0,03 s / sondaż
```

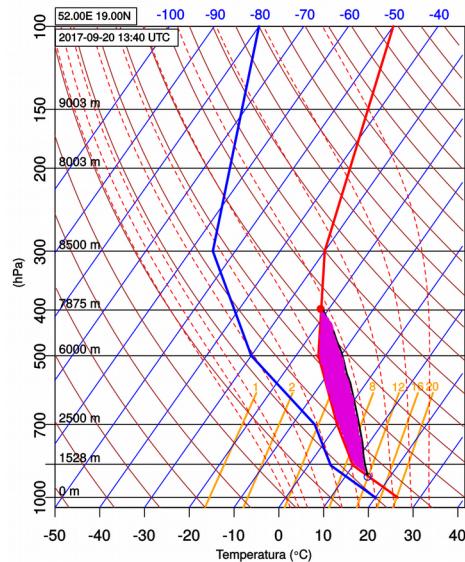
Możliwe przyspieszenie (bez zrównoleglania):
0,49/0,03 = ~ 16,3x

24h * 1125 punktów (dla PL) indeksu CAPE:
~3h40 min vs. 13 min

3) R + Rcpp (+cache)

Możliwe przyspieszenie (bez zrównoleglania):
~5-10x szybciej niż Fortran

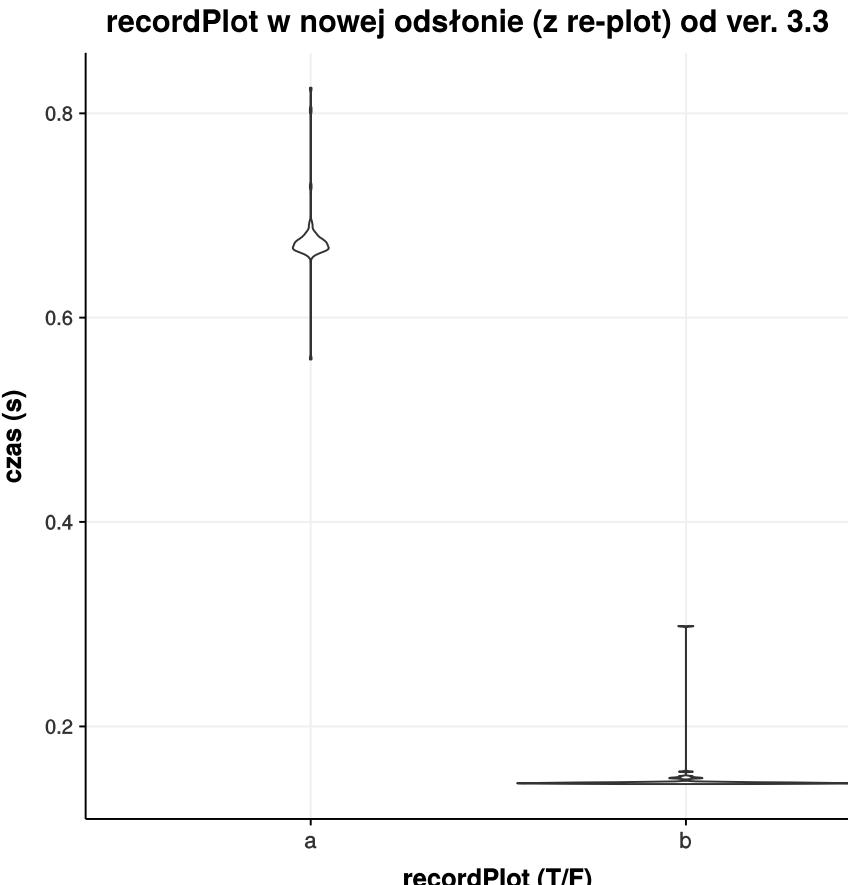
Faster::plotowanie z szablonu



Jeśli mamy wiele takich wykresów do wygenerowania na jednym szablonie warto pomyśleć o funkcji recordPlot():

Rysowanie diagramu Skew-T bez i z użyciem szablonu →

Możliwe przyspieszenie:
676ms/148ms = ~ 4.57x



expr	min	lq	mean	median	uq	max	neval	
1	a	559.1534	666.7048	676.9848	672.7972	676.7871	825.0011	50
2	b	143.5537	144.1003	148.3704	144.6700	145.1854	298.0670	50

FasteR:: wielowątkowość

Przepis na najprostsze zrównoleglenie kodu (np. z prognozami pogody) w **R**:

- 1) Rozbij (np. `split()`) bazę danych (`data.frame`) na listę
- 2) Napisz funkcję „czyszczącą” (jeśli potrzebna) i robiącą wykresy/obliczenia
- 3) `library(parallel)`
- 4) skorzystaj ze zrównoleglonego `lapply()` → `mclapply()`

```
> head(dt, 15)
   date2    lev lon   lat DPT     HGT PRES TMP WD WS
1: 2017-09-05 06:00:00 1000 14 49.00 10.45 176.632 NA 13.65 243 2.443866
2: 2017-09-05 06:00:00 1000 14 49.25  9.05 172.184 NA 13.85 257 2.977459
3: 2017-09-05 06:00:00 1000 14 49.50  9.11 171.848 NA 12.95 273 3.134176
4: 2017-09-05 06:00:00 1000 14 49.75  9.06 171.272 NA 12.55 258 2.821284
5: 2017-09-05 06:00:00 1000 14 50.00  8.16 166.856 NA 12.75 246 3.338212
6: 2017-09-05 06:00:00 1000 14 50.25  7.50 165.192 NA 12.55 241 2.856708
7: 2017-09-05 06:00:00 1000 14 50.50  7.07 161.928 NA 13.15 264 1.537956
8: 2017-09-05 06:00:00 1000 14 50.75  7.91 161.784 NA 13.15 282 1.776396
9: 2017-09-05 06:00:00 1000 14 51.00  7.28 158.984 NA 13.05 258 3.180268
10: 2017-09-05 06:00:00 1000 14 51.25  7.31 158.152 NA 12.35 254 4.005302
11: 2017-09-05 06:00:00 1000 14 51.50  6.97 154.968 NA 12.15 259 4.353213
12: 2017-09-05 06:00:00 1000 14 51.75  7.35 152.392 NA 12.35 271 4.920718
13: 2017-09-05 06:00:00 1000 14 52.00  7.69 149.896 NA 12.75 288 5.495212
14: 2017-09-05 06:00:00 1000 14 52.25  7.67 148.232 NA 12.75 298 5.936323
15: 2017-09-05 06:00:00 1000 14 52.50  7.51 146.456 NA 12.35 301 6.479519
```

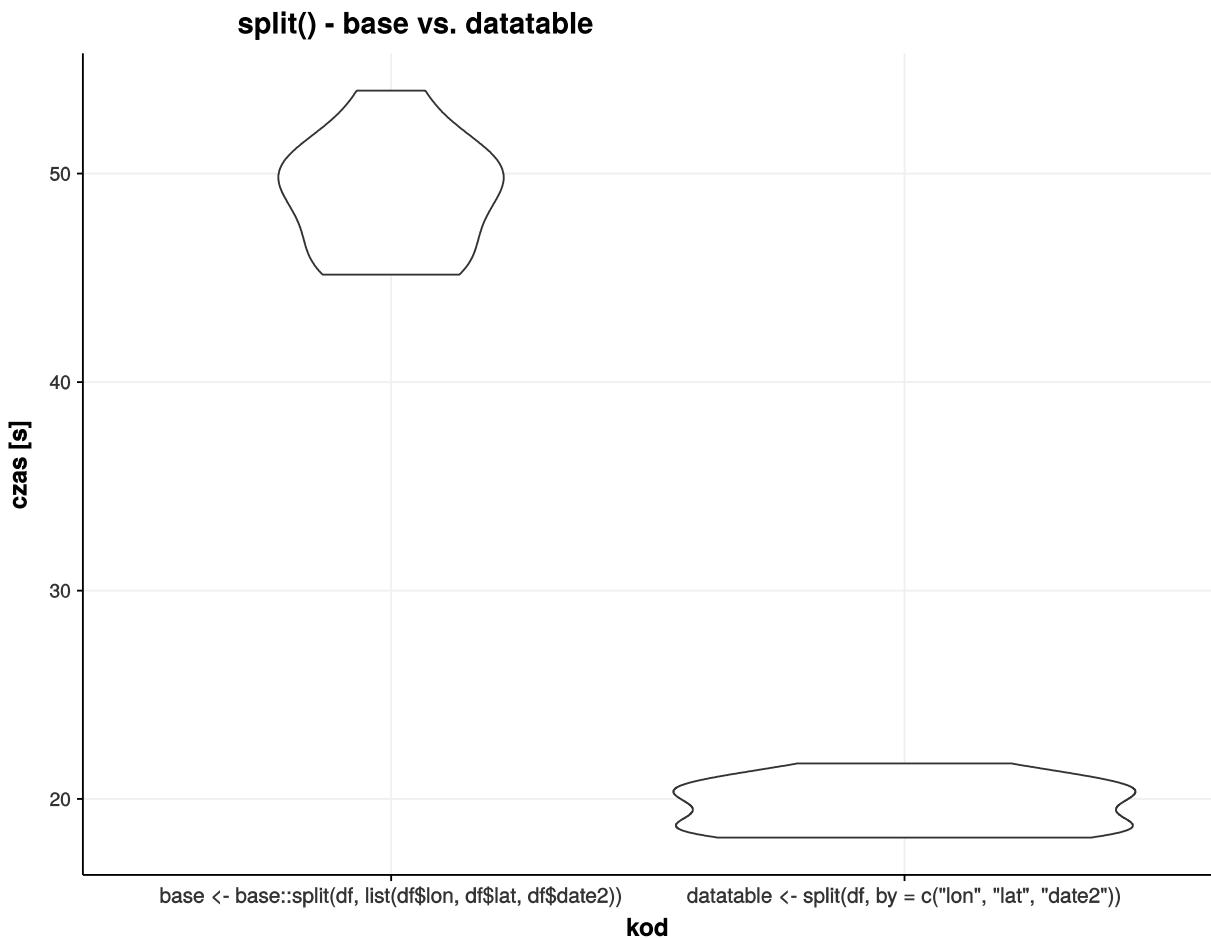
Faster:: rozdzielimy do listy → split()

base vs. data.table:

```
base::split()  
data.table::split()
```

**Możliwe przyspieszenie
(praktycznie bez zmiany kodu):**

$49,5/19,8 = \sim 2,5x$

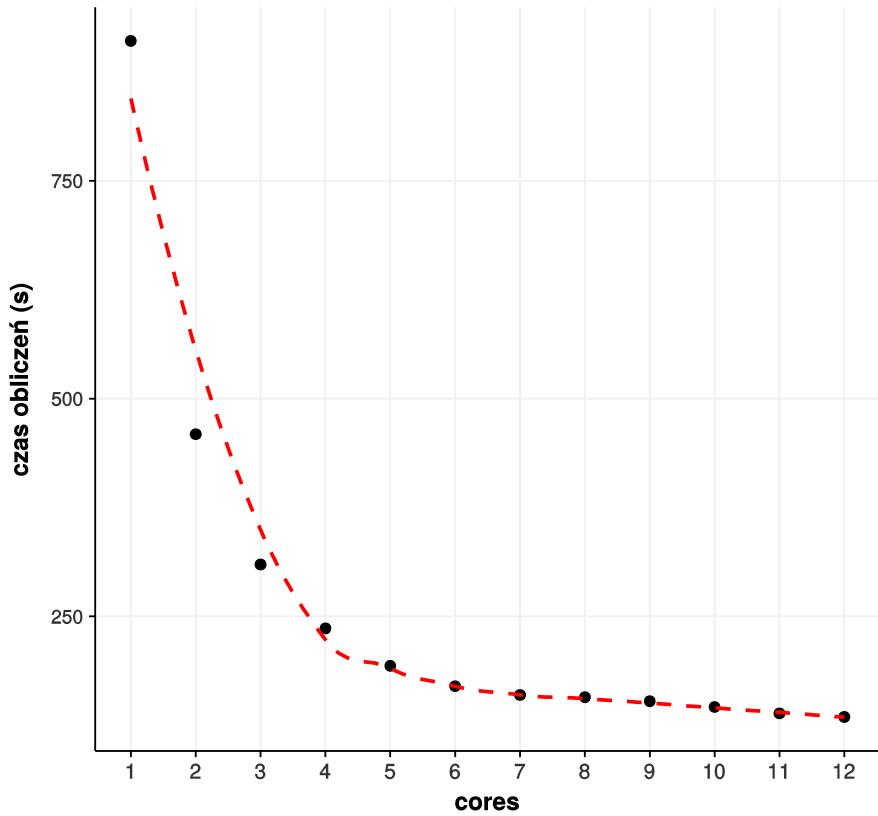


FasteR::mclapply

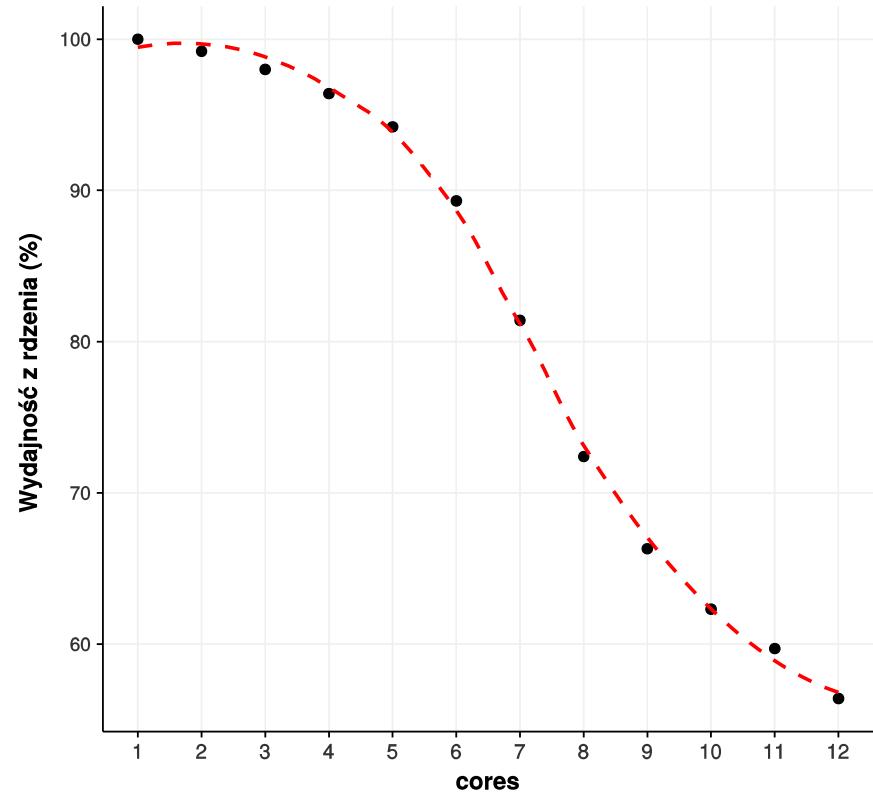
Czy warto zrównolegać za wszelką cenę? → skalowalność

Intel(R) Core(TM) i7-5930K CPU @ 3.50GHz, 24 GB RAM

Czas obliczeń vs. liczba rdzeni



Średnia wydajność z rdzenia vs. single core lapply (%)



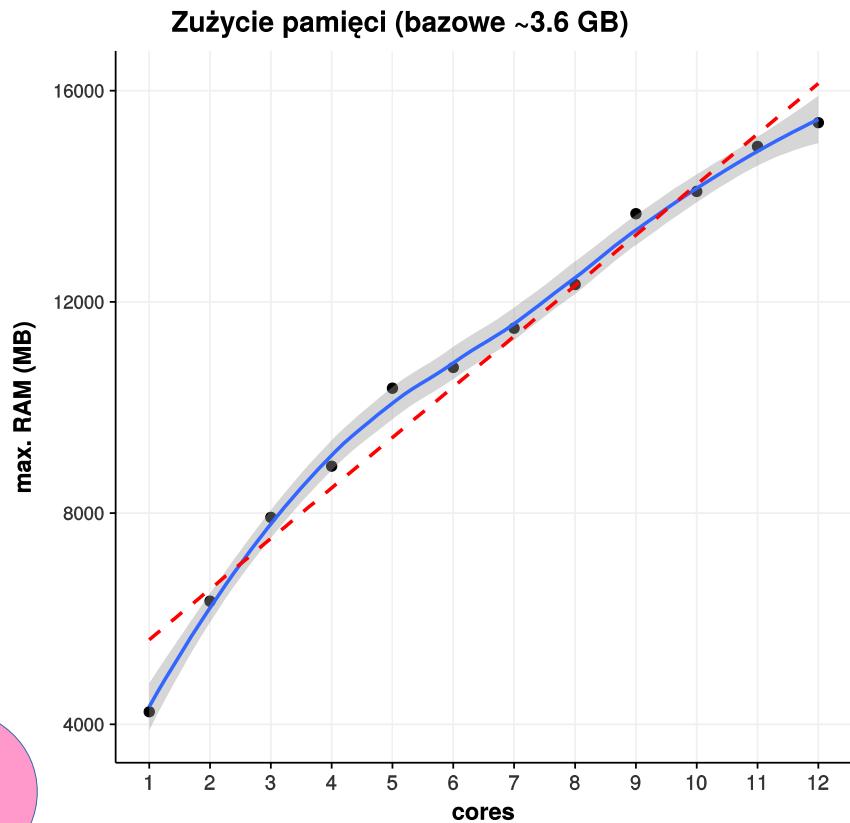
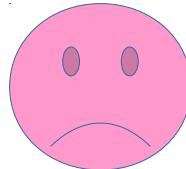
FasteR::mclapply

Maksymalne zużycie pamięci związane z użyciem mclapply()

Max. zużycie chwilowe przy
mclapply()
= ~liczba rdzeni * object.size()

- Zagrożenie na klastrach
PCSS/WCSS (brak SWAPa!)

- W praktyce max. no_cores dla
najmocniejszej konfiguracji R w PCSS
(28 rdzeni, 256 GB RAM) → źródłowa
baza danych <4.5GB (bo musimy
jeszcze obiekt zapisać)





FasteR:: PCSS/WCSS → wnioski

(Jeśli to możliwe) **dziel bazę na mniejsze fragmenty i uruchamiaj zadania jako oddzielne procesy bez zrównoleglania (sic!)**

- **krótsza kolejka**
- **mniejsza liczba zużytych godzin w grancie**
- **... pętla w Bashu do systemu kolejkowego**
- **przygotowanie skryptu R do wywoływania z argumentami:**

```
#!/bin/bash
#SBATCH -N1
#SBATCH -n 4
#SBATCH -t 900
#SBATCH --mem=125G
```

```
module load r
Rscript /home/users/barcze/era/prepare.R 1979 1980
```

FasteR:: PCSS/WCSS → wnioski

„Ile da się” policzyć w ten sposób policzyć?

Projekt: „stwórz bazę danych warunków termodynamicznych do *machine learningowego* prognozowania burz i zjawisk ekstremalnych nad Europą”

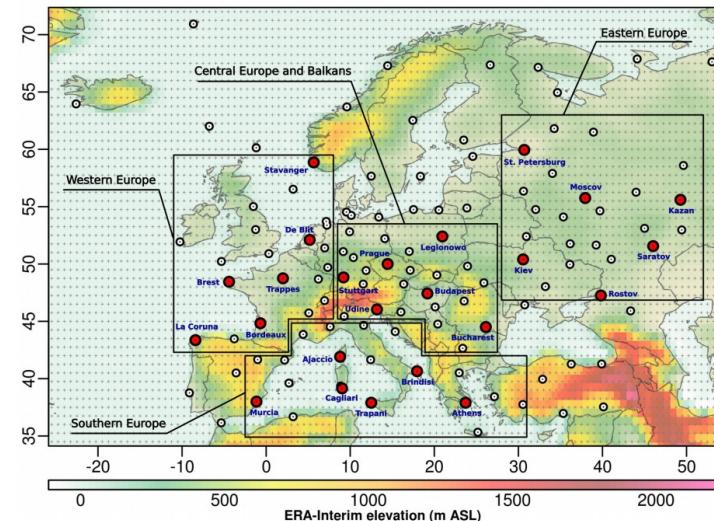
1) Funkcja Rcpp obliczająca 120 indeksów termodynamicznych

**2) Baza danych dla Europy 1979-2016 co 6h
≈1.1 mld profilów atmosfery, >100 mld liczb...**

3) Wczytanie tylko 1 miesiąca/zadanie = 444 zadania

4) ... Wyniki po 1.5 dniu ... zamiast po 2 miesiącach

5) Łącznie zużyto ok. 6000 z przyznanych 50000h





- 1 Climatological aspects of convective parameters over Europe: a
- 2 comparison of ERA-Interim and sounding data.
- 3 Mateusz Taszarek^{*1,5}, Harold E. Brooks², Bartosz Czernecki¹, Piotr Szuster³, Krzysztof Fortuniak⁴
- 4
- 5 ¹- Department of Climatology, Institute of Physical Geography and Environmental Planning, Adam Mickiewicz University, Poland
- 6 ² - NOAA/National Severe Storms Laboratory, Norman, Oklahoma, United States.
- 7 ³ - Cracow University of Technology, Poland.
- 8 ⁴ - Department of Meteorology and Climatology, Faculty of Geographical Sciences, University of Łódź, Poland.
- 9 ⁵ - Skywarn Poland

Dziękuję za
uwagę

<http://www.enwo.pl>



Part of computations were carried out in
the Poznań Centre for Networking and
Supercomputing
(<http://hpc.man.poznan.pl>), Grant No. 331.