

Proces kategoryzacji semantycznej pojęć w
przestrzeni barw - analiza danych światowych
i modelowanie populacyjne

dr inż. Franciszek Rakowski
ICM UW
19 III 2015

Jak to się dzieje że gdy mówimy “zielony” rozumiemy “zielony”, i czy tak jest do końca...

- **Natywizm** Chomsky & Fodor

Struktura - ogólny konstrukt - języka jest wrodzona. Zbiór kategorii językowych - zarówno pojęć jak i gramatyk jest współdzielony przez wszystkich ludzi od urodzenia. Uczenie się języka polega na wypełnianiu aktualnymi formami językowymi struktur uprzednio istniejących

- **Empirycyzm**

Mechanizm nabywania języka jest współdzielony przez społeczności, jednak samo kształtowanie się jego struktur jest odbiciem otaczającej nas rzeczywistości. Funkcjonalizm.

- **Kulturalizm**

Poza językotwórczym oddziaływaniem środowiska silny wpływ na strukturę języka i zbiór pojęć ma konsensus kulturowy. Przejmujemy od otaczających nas ludzi strukturę języka.

Pomiary WCS

World Color Survey, Berlin & Kay 1969

Berlin, Brent and Paul Kay. Basic Color Terms: Their Universality and Evolution.
Berkeley and Los Angeles. University of California Press, 1969.

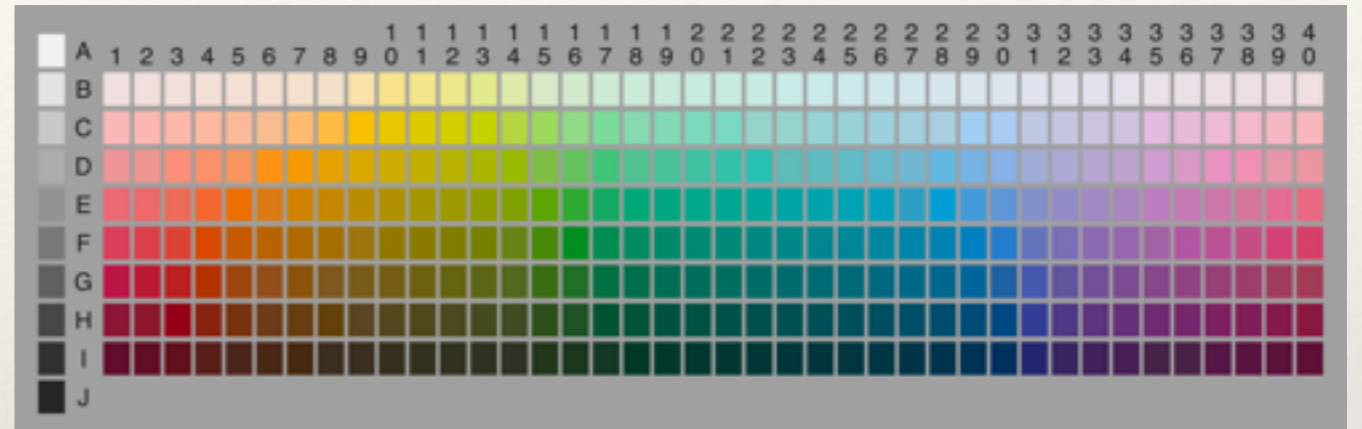
- Badano przyporządkowanie nazwy barwy do pól palety Munsella
- wykorzystano przestrzeń barw CIE L*a*b



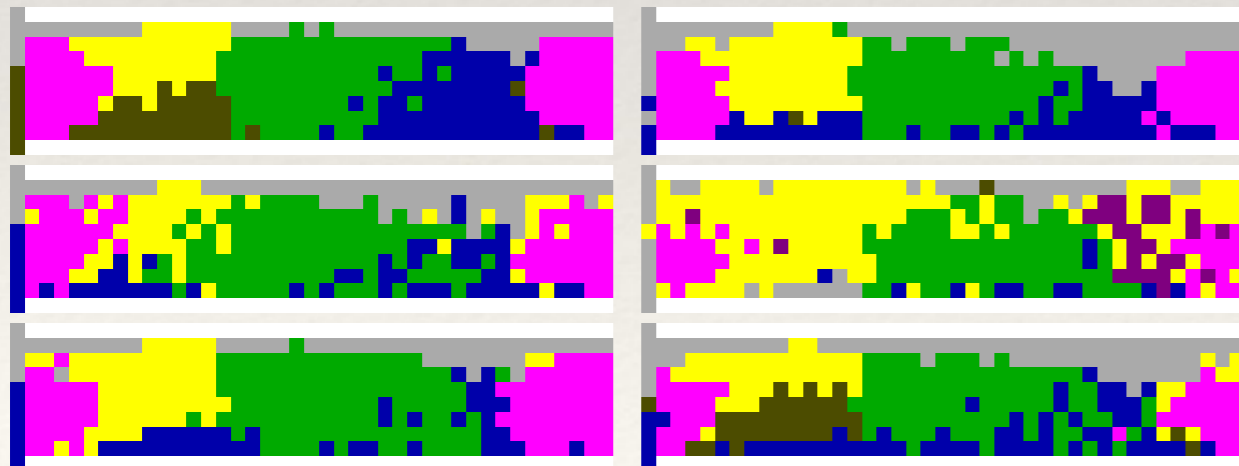
- zbadano 110 niepiśmiennych kultur z całego świata.

Jak odpowiadali poszczególni badani?

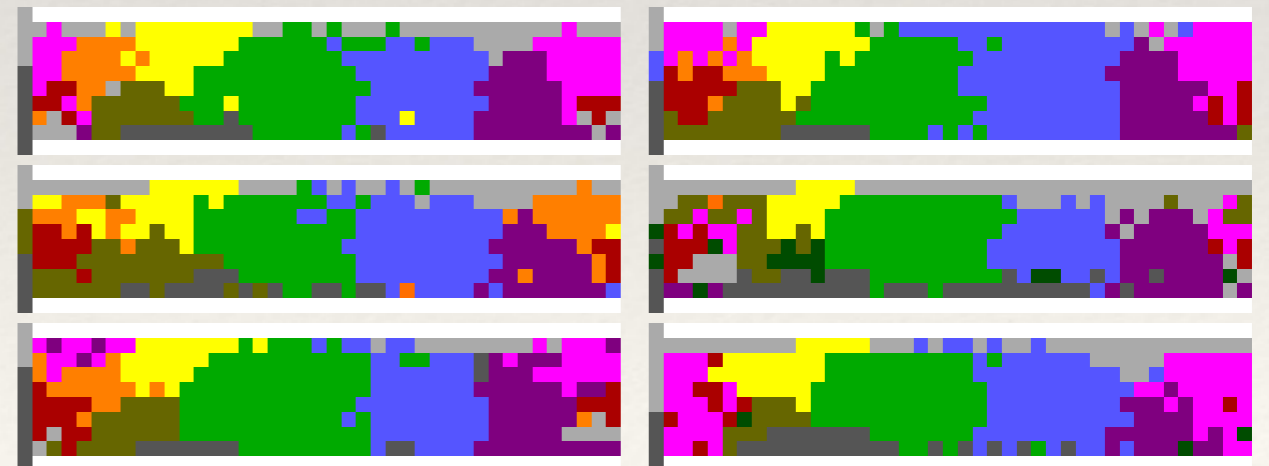
- ❖ Nie każdy z badanych tak samo dzieli przestrzeń kolorów
- ❖ Istnieje jednak duży stopień uwspólnienia nazw...



Chavacano Philippines (WCS 24)

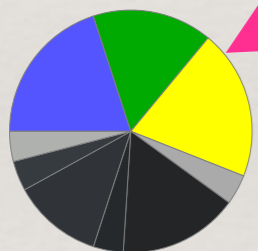
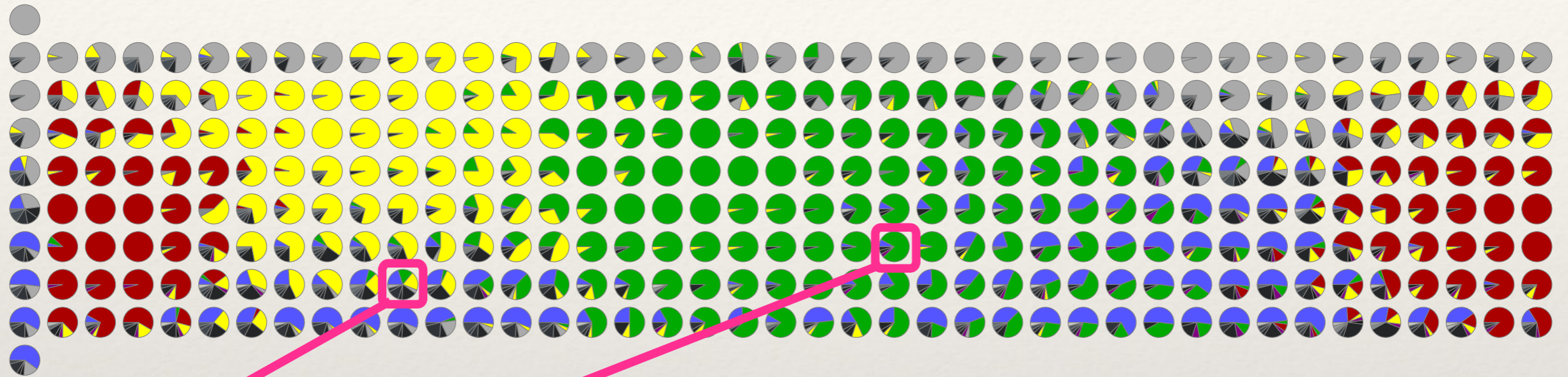


Iduna Papua New Guinea (WCS 47)

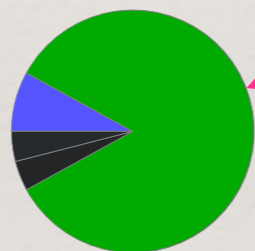


Iduna Papua New Guinea (WCS 47)

kategoryzacja uwspólniona

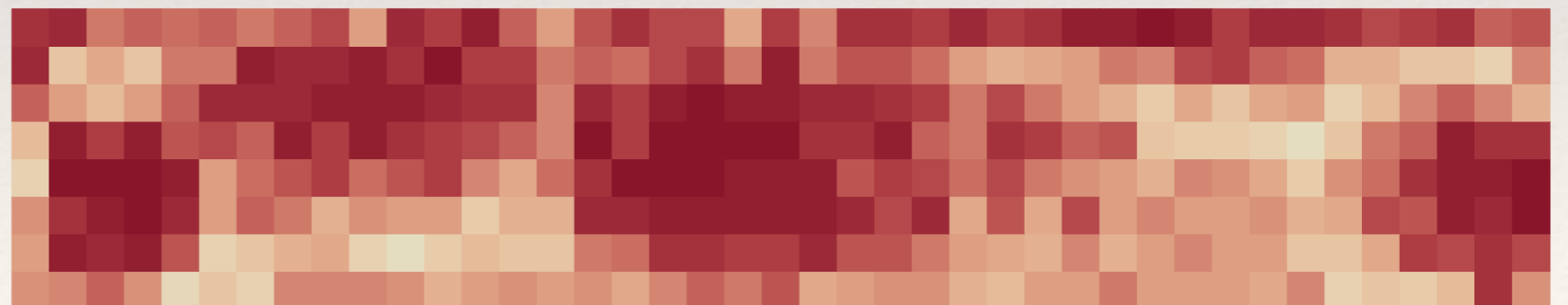


$Mc = 0.2$



$Mc = 0.84$

Na podstawie udziału mody, można utworzyć tzw. mapę kongruencji. W ten sposób otrzymujemy zredukowaną reprezentację danych.



Jakie były całościowe wnioski z badań?

World Color Survey, Berlin & Kay 1969

Berlin, Brent and Paul Kay. Basic Color Terms: Their Universality and Evolution. Berkeley and Los Angeles. University of California Press, 1969.

- pokazano hierarchiczność w słownikach pomiędzy poszczególnymi językami,
- argumentacja na rzecz koncepcji uniwersalistycznej

All languages contain terms for black and white.

If a language contains three terms, then it also contains a term for red.

If a language contains four terms, then it also contains a term for either green or yellow (but not both).

If a language contains five terms, then it contains terms for both green and yellow.

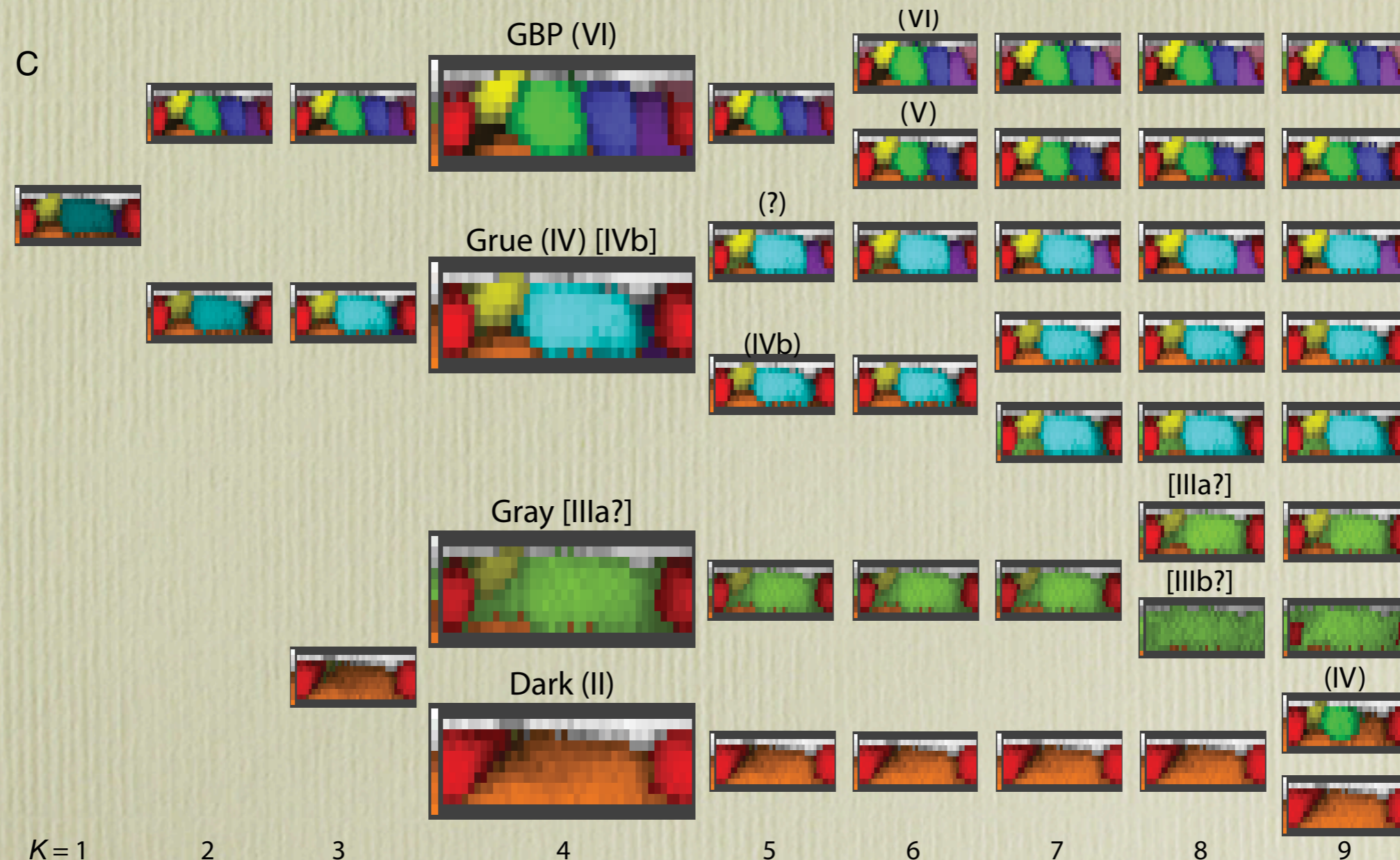
If a language contains six terms, then it also contains a term for blue.

If a language contains seven terms, then it also contains a term for brown.

If a language contains eight or more terms, then it contains a term for purple, pink, orange, and/or grey.

Argumenty na rzecz Uniwersalizmu

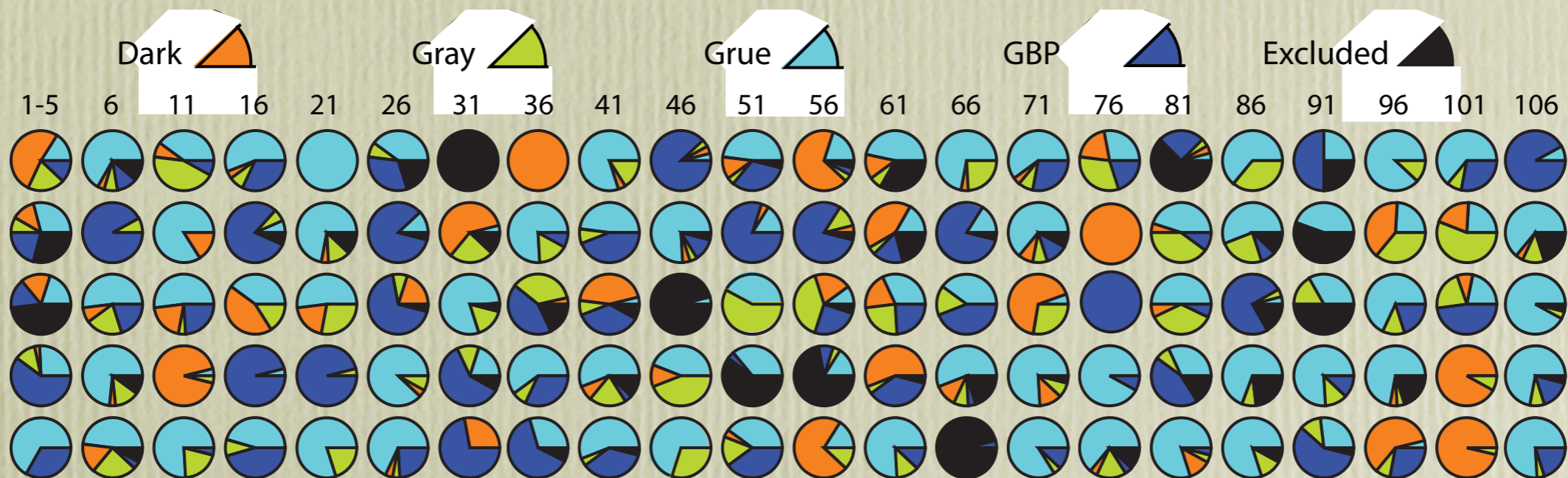
- wykonano różne analizy tych danych. Np. klastrowanie badanych osób w całym zespole badanych.



1. Lindsey DT, Brown AM. World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(47):19785-90.

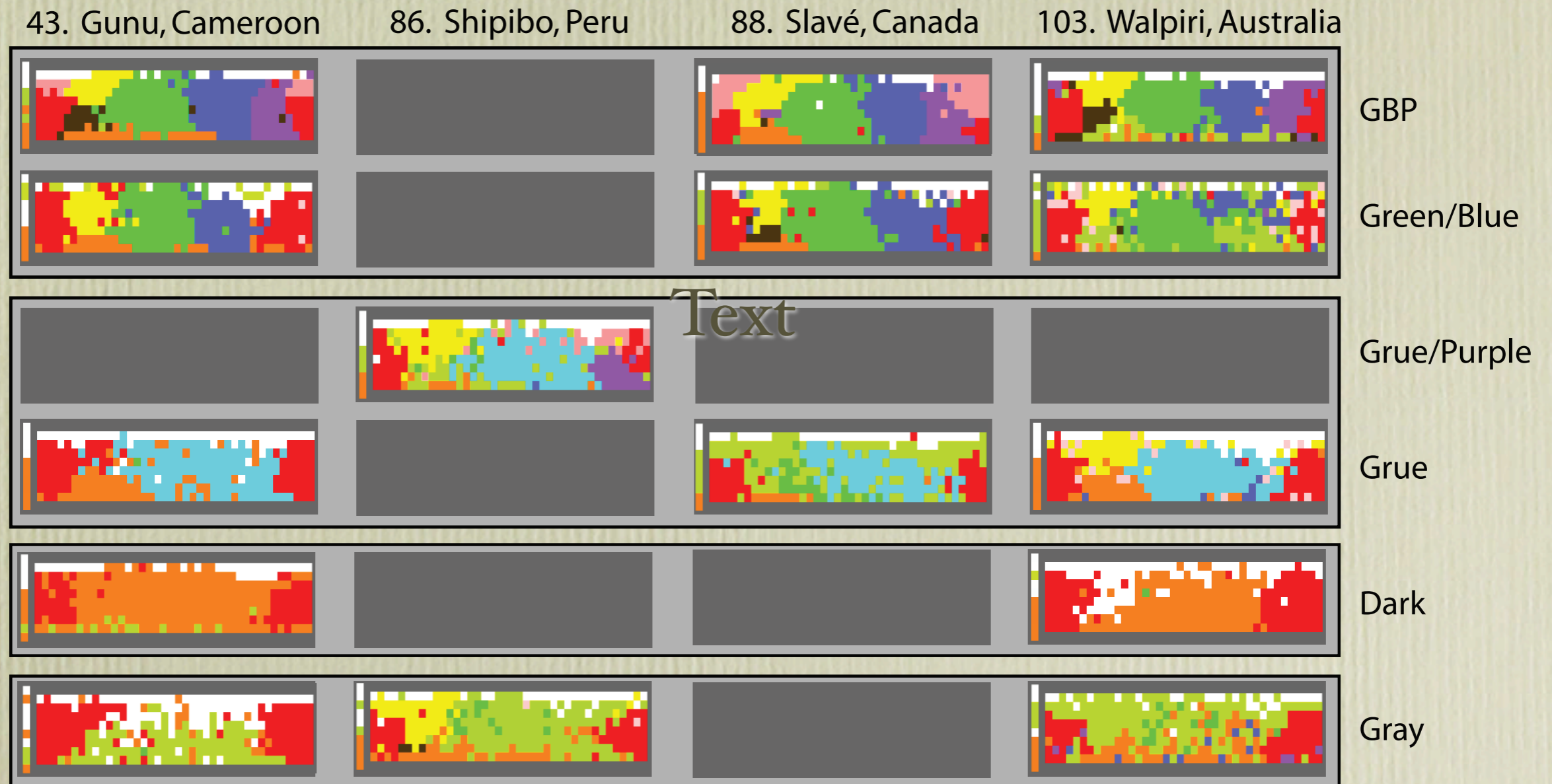
Argumenty na rzecz Uniwersalizmu

- motywy w nazewnictwie kolorów grupują osoby z różnych, niespokrewnionych grup językowych.
- w większości języków istnieją podstawowe motywy nazewnictwa kolorów GBP, Grue, Grey, Dark.



1. Lindsey DT, Brown AM. World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(47):19785-90.

Wygląd motywów w wybranych językach



Text

1. Lindsey DT, Brown AM. World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences of the United States of America*. 2009;106(47):19785-90.

Modelowanie

powstawania kategorii w przestrzeni barw i przypisania nazwy do danej kategorii.

- Spektrum oraz intensywność barw prezentowana jest poprzez 1269 (lub 330) kostek Munsella
- transformacja bodźca $S(\lambda) \Rightarrow \{L,a,b\}$
- jednostki centrujące (reactive units)

$$z_j(x) = e^{-\frac{1}{2} \sum_{i=1}^N \left(\frac{x_i - m_{ij}}{\sigma} \right)^2}$$

1. Puglisi A, Baronchelli A, Loreto V. Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(23):7936-40.

2. Luc Steels, Tony Belpaeme Coordinating perceptually grounded categories through language: A case study for colour *Behavioral and Brain Sciences*, Vol. 28, No. 4. (2005), pp. 469-489.

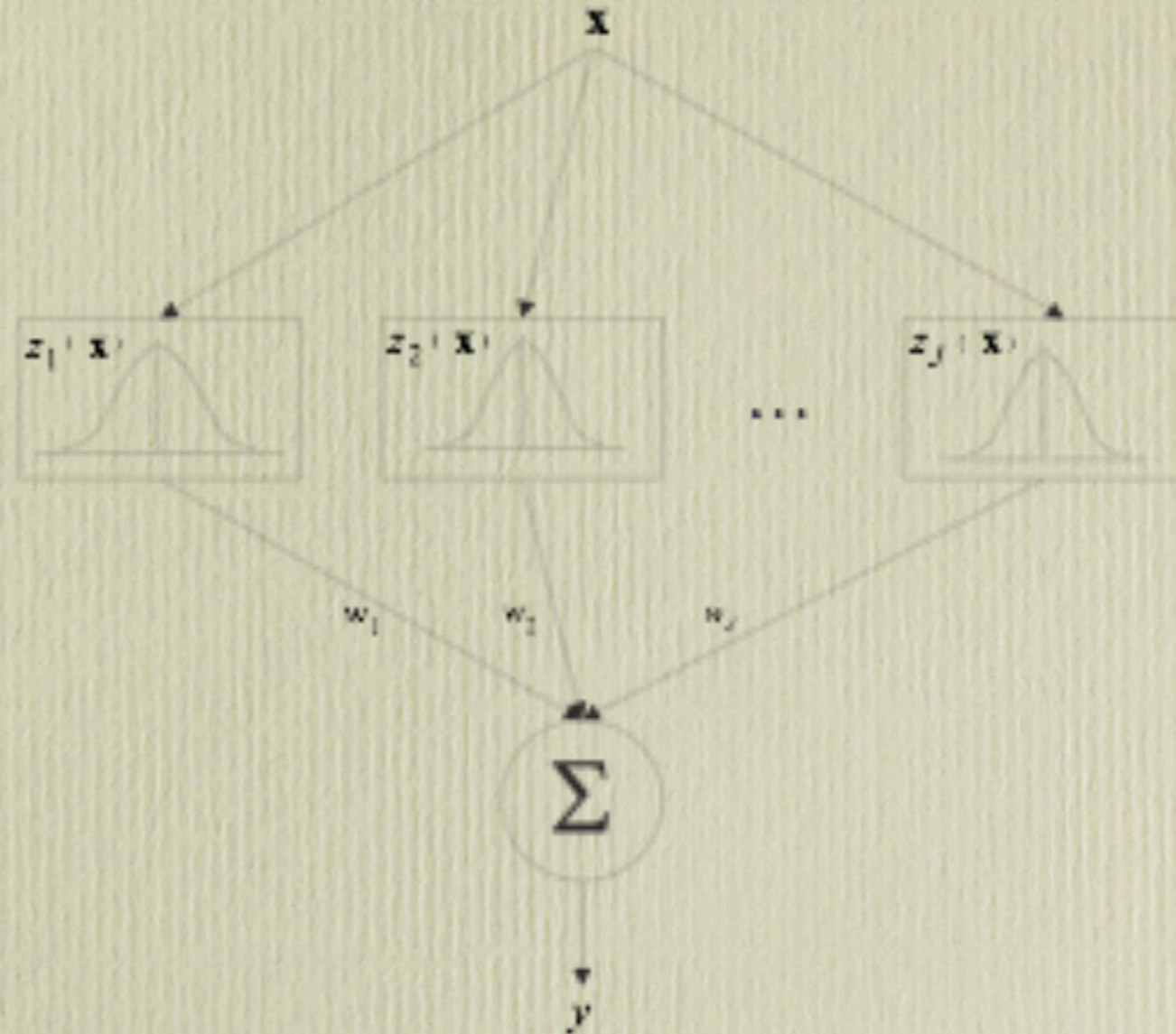
Kategoryzacja bodźca

- kategoria koloru jest zdefiniowana przez sieć adaptacyjną

$$y_k(x) = \sum_{j=1}^J w_j z_j(x)$$

- każdy agent posiada zbiór kategorii i wybiera kategorię najmocniej odpowiadającą na bodziec - $\text{argmax}(y(x))$ w trakcie gry w dyskryminację.

perceptrony poznawcze:



1. Puglisi A, Baronchelli A, Loreto V. Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences of the United States of America*. 2008;105(23):7936-40.

2. Luc Steels, Tony Belpaeme Coordinating perceptually grounded categories through language: A case study for colour *Behavioral and Brain Sciences*, Vol. 28, No. 4. (2005), pp. 469-489.

Gra w dyskryminację

- opisuje funkcjonalizm środowiskowy
- algorytm:
 - prezentacja kontekstu $O = \{o_1, \dots, o_N\}$ wraz z próbką wyróżnioną.
 - wybór najsilniej odpowiadającej kategorii $c_{s_o} = \operatorname{argmax}(y_c)$
 - sukces dyskryminacyjny jeśli:
 - istnieje kategoria zwracająca najwyższą wartość dla próbki wyróżnionej (jeśli nie - modyfikacja kategorii)
 - żadna inna próbka w kontekście nie ma równie wysokiego wyniku

Gra w zgadywanie I

proces uzgadniania przestrzeni nazw w populacji

- opisuje wpływ kulturowy i środowiskowy, występują: mówca i słuchacz
- algorytm:
 - prezentacja kontekstu $O = \{o_1, \dots, o_N\}$ mówcy i słuchaczowi
 - prezentacja próbki wyróżnionej mówcy
 - mówca gra w grę dyskryminacyjną - jeśli gra odnosi sukces - kontynuacja, w innym przypadku gra w zgadywanie jest przerywana
 - w przypadku sukcesu wybrana zostaje kategoria zwycięska c^s

Gra w zgadywanie II

- mówca próbuje przypisać kategorię zwycięską c^s do słowa z leksykonu
 - jeśli istnieje słowo (najmocniej) związane z kategorią c^s jest ono wybierane
 - jeśli takiego słowa nie ma - tworzone jest ze zbioru predefiniowanych sylab.
- słowo przekazywane jest słuchaczowi
- jeśli słuchacz nie znajduje przekazanego słowa w swoim leksykonie to:
 - jest informowany przez mówcę która próbka jest próbką wyróżnioną
 - gra w grę dyskryminacyjną (z tym samym kontekstem i próbką wyróżnioną co mówca na początku gry w zgadywanie)
 - jeśli gra w dyskryminację kończy się sukcesem - przypisuje słowo usłyszane od mówcy do kategorii wygranej
 - jeśli nie: nowa kategoria jest tworzona i przypisywane jest do niej słowo – w obrębie informacji prywatnych słuchacza

Gra w zgadywanie III

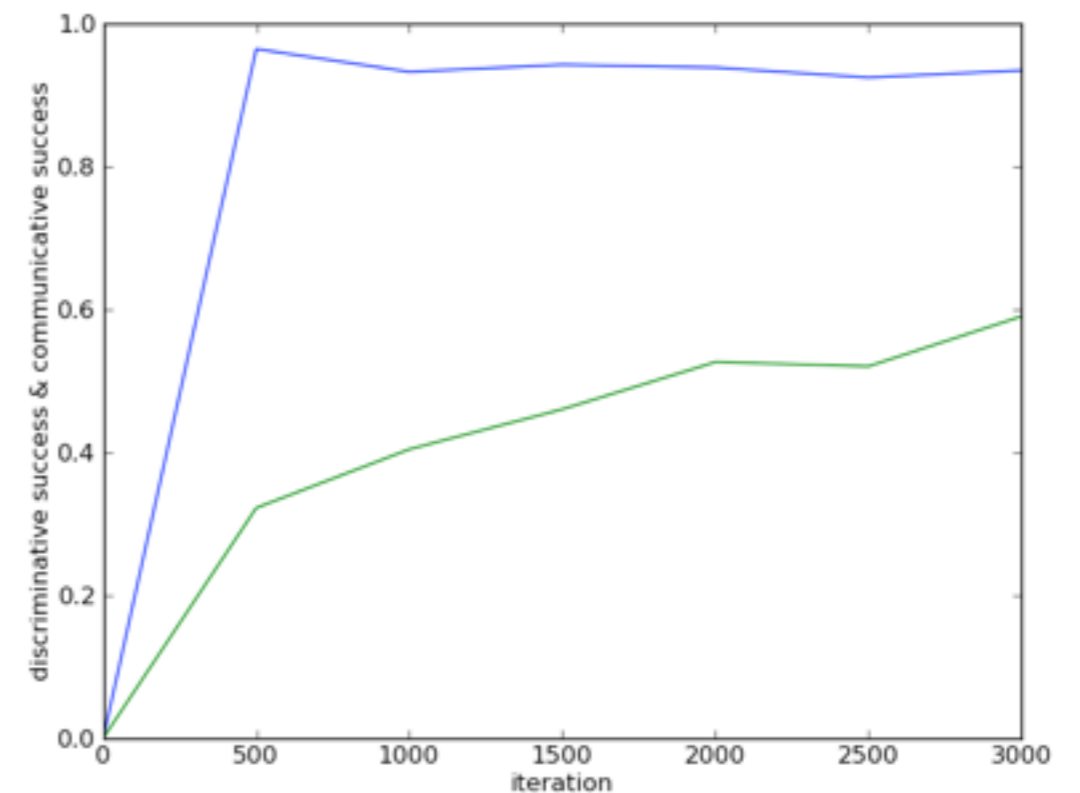
- jeśli słuchacz posiada słowo usłyszane od mówcy w swoim leksykonie to:
 - szuka kategorii w swoim leksykonie związanej z usłyszonym słowem i identyfikuje główną próbkę odpowiadającą tej kategorii (argmax)
 - wskazuje mówcy tą próbkę
- mówca porównuje próbki: wybraną przez siebie na początku gry i tą prezentowaną w tej chwili przez słuchacza
 - jeśli się zgadzają gra kończy się sukcesem
 - jeśli nie: mówca prezentuje swoją próbkę wyróżnioną słuchaczowi, a słuchacz wykonuje operacje zaznaczone na **zielono**

Miary monitorujące eksperyment numeryczny:

- sukces dyskryminacyjny dla poszczególnego agenta:
- sukces dyskryminacyjny na poziomie populacji.

$$DS_j^A = \frac{1}{n} \sum_{i=j-n}^j ds_i^A$$

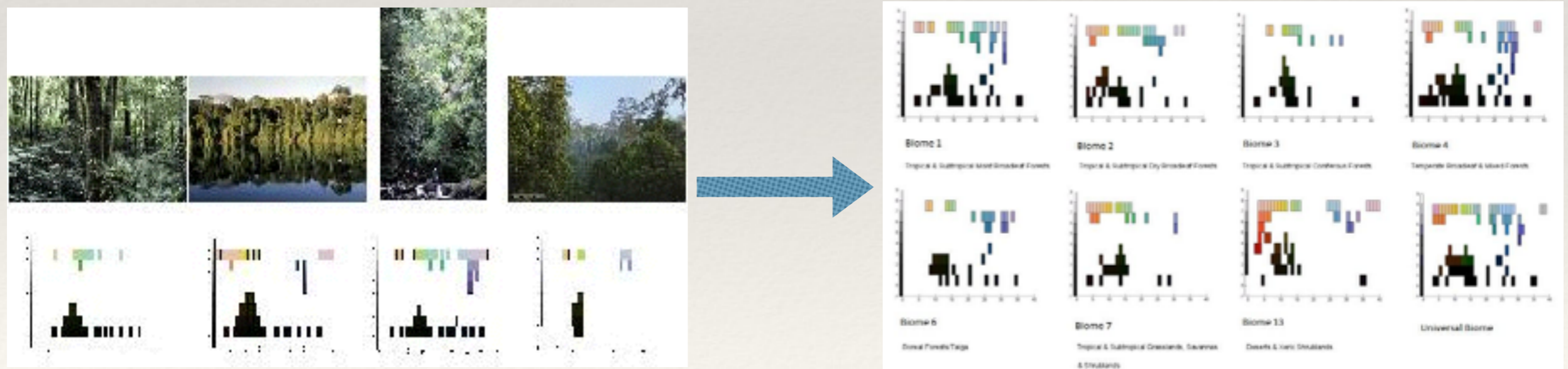
$$\mathcal{D}S_j = \frac{1}{m} \sum_A DS_j^A$$



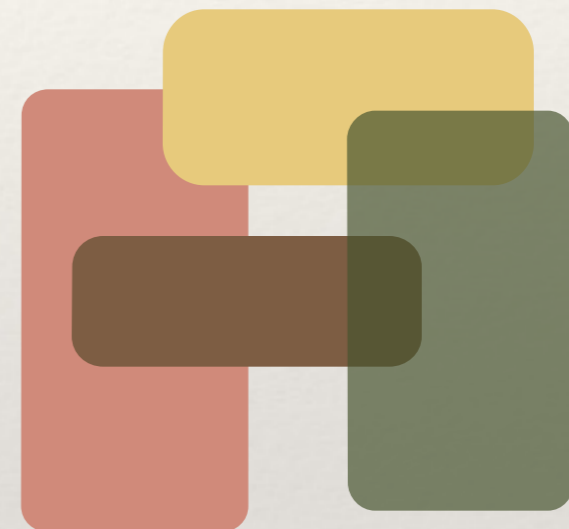
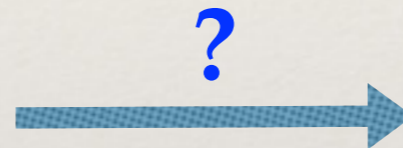
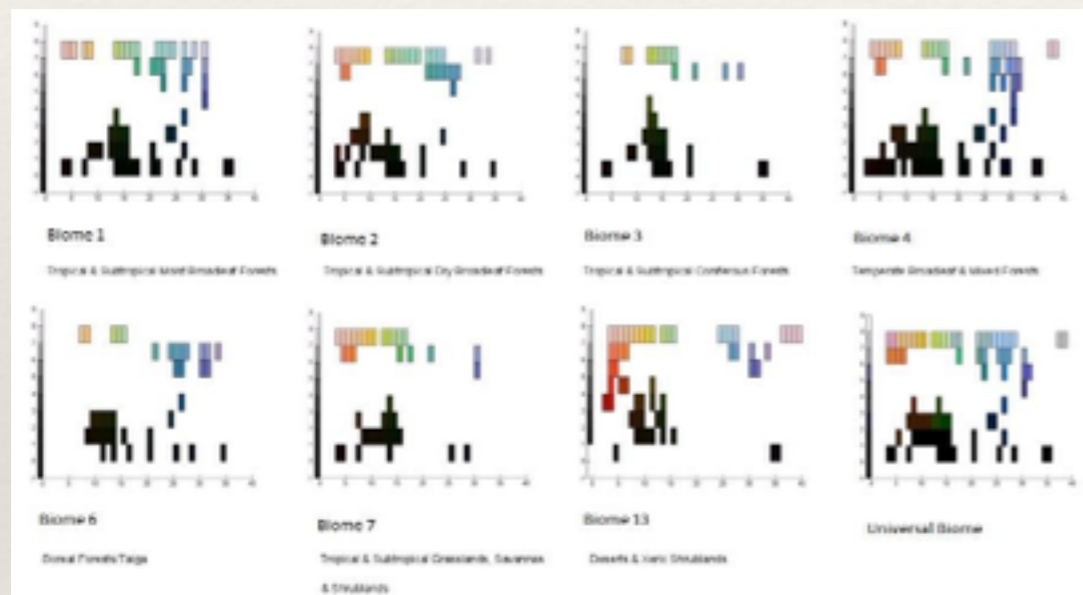
Czy można zbadać wpływ środowiska?

- ❖ Following the World Wildlife Fund (Olson et al. 2001) we have applied the environment type categorisation based on biomes.
- ❖ The WWF has determined 14 biomes regarding climate, fauna, flora, landscape, habitat, biodiversity and human activity, 110 cultures fall into subset of 7 biomes (1,2,3,4,6,7,13).
- ❖ Collection of the pictures with unambiguous assignment to a biome was used as the data set, based on which the so called **biome regions** - in the colour space - were determined.

eg. Boreal Forest



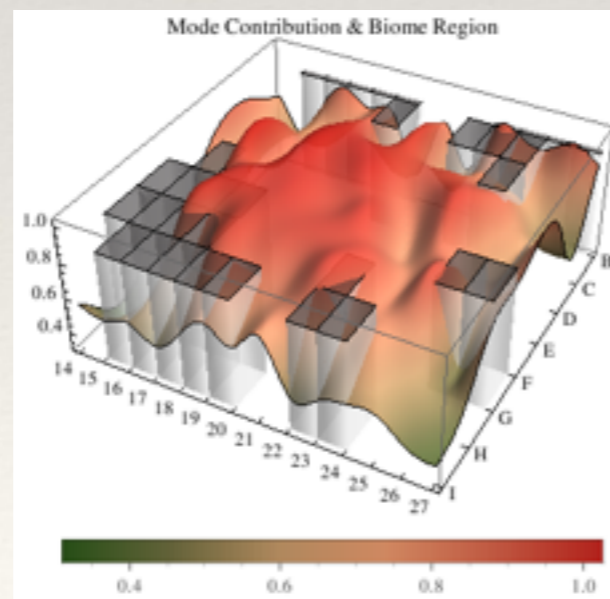
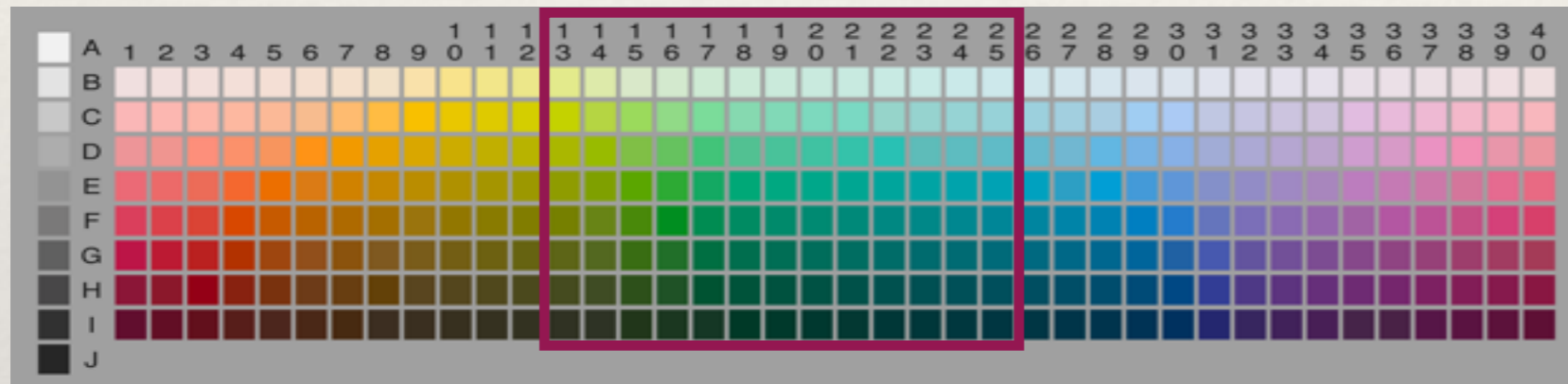
Jak częstość kolorów występujących w środowisku determinuje system kategoryzacji barw?



co mówią nam mapy kongruencji z nałożonym na nie wzorcem częstotliwości bodźców?

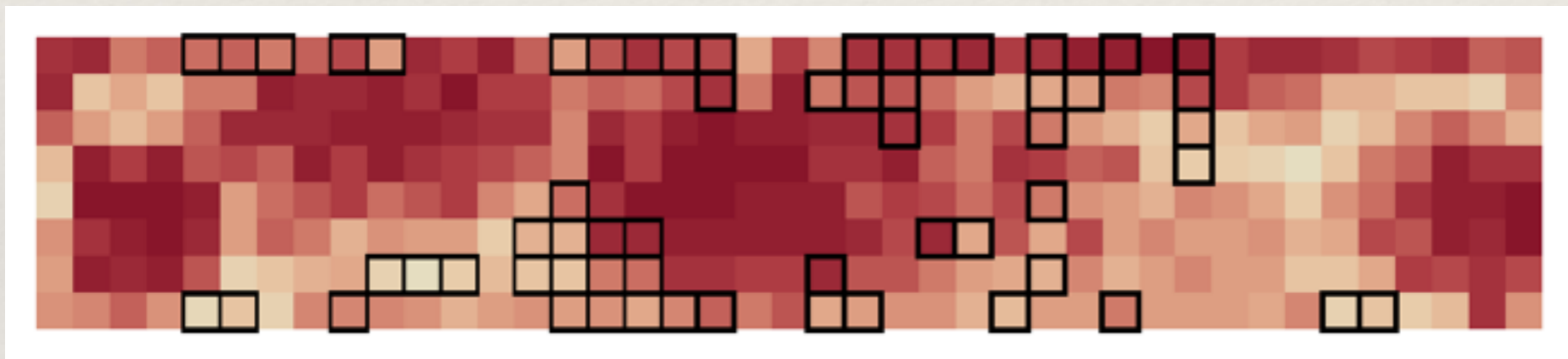
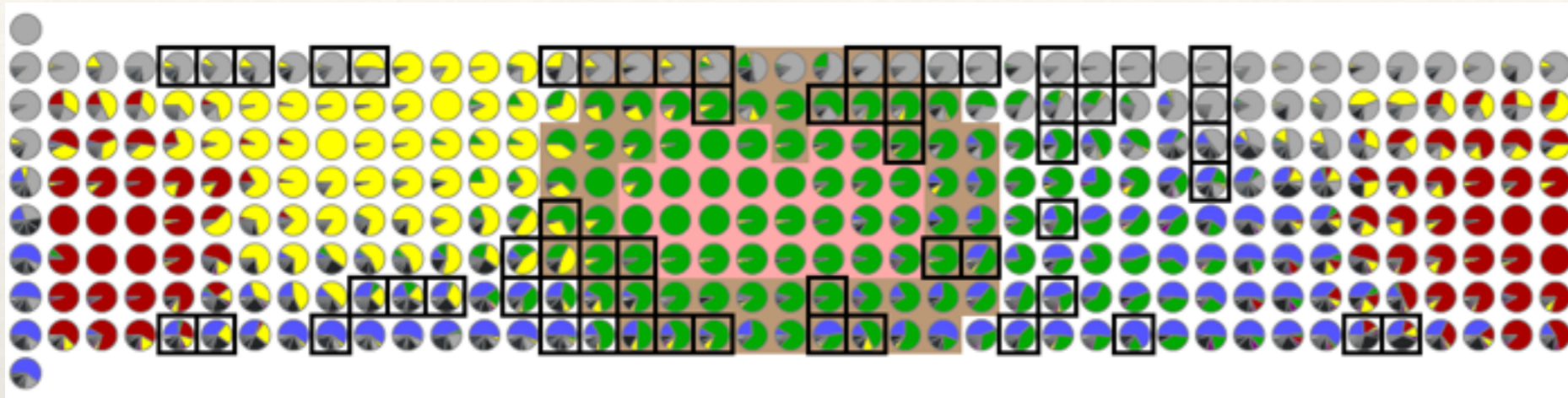


Iduna Papua New Guinea



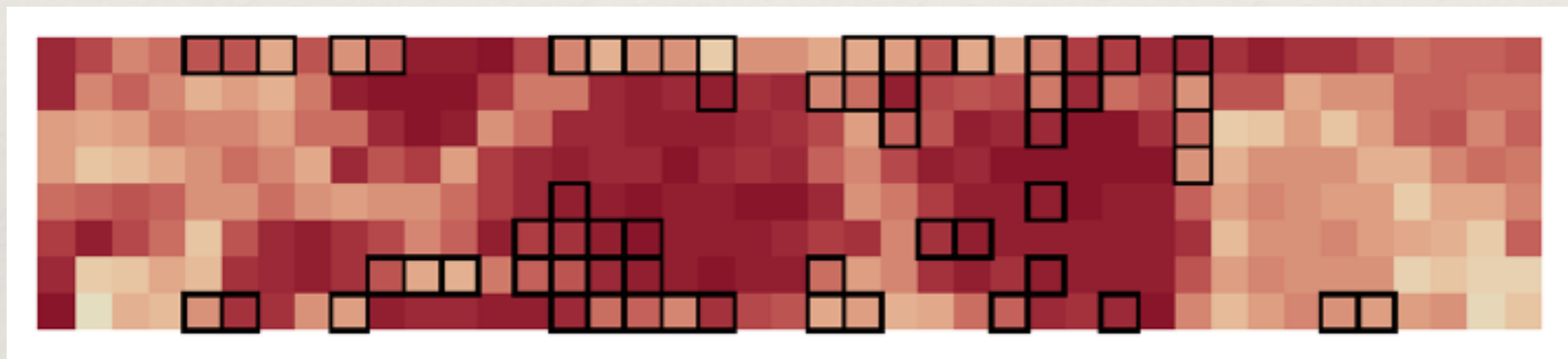
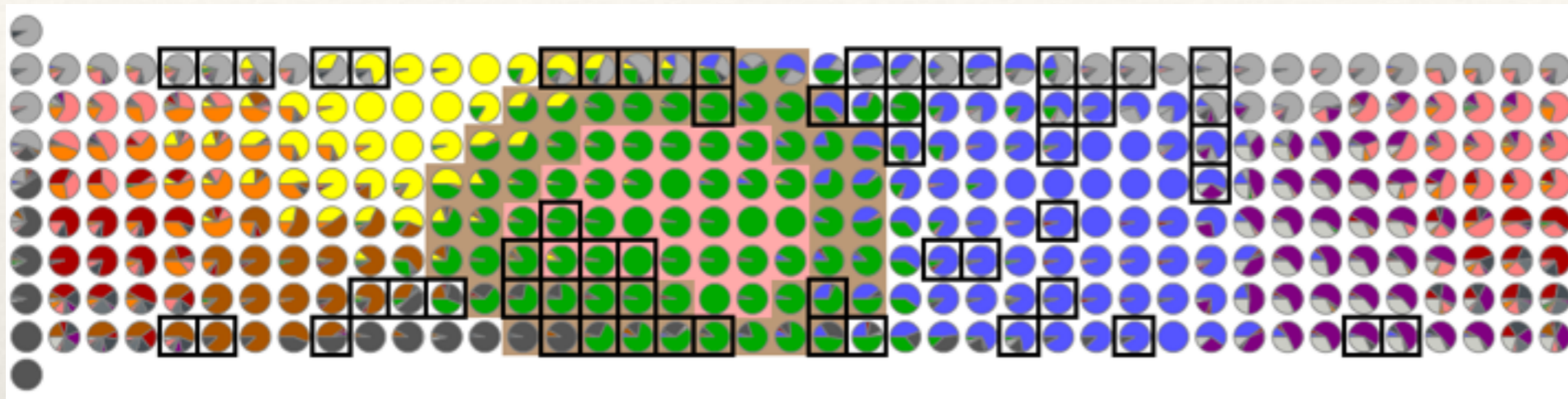
We see that the region of high congruence in naming is free from frequent environmental stimuli, which are rather surrounding the plateau.

Iduna Papua New Guinea (WCS 47)



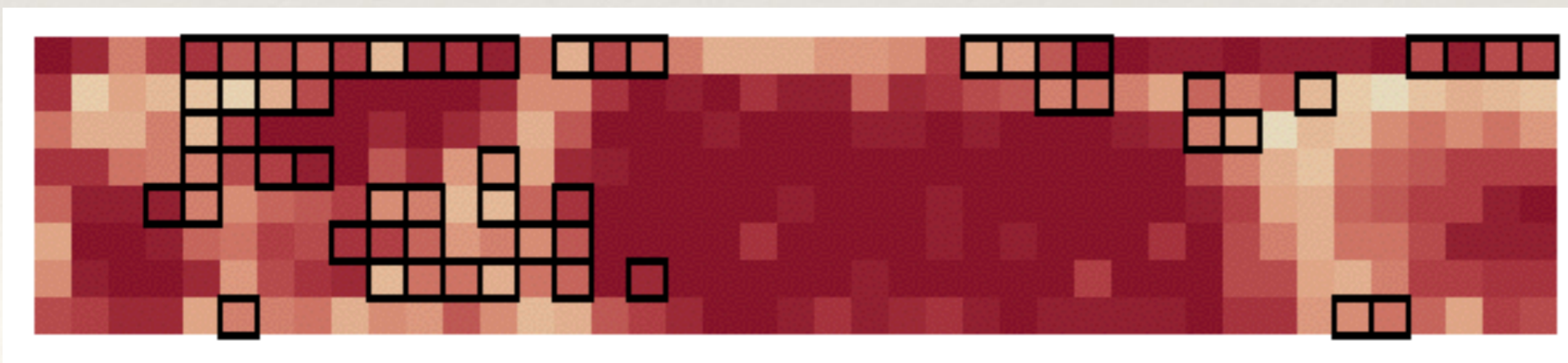
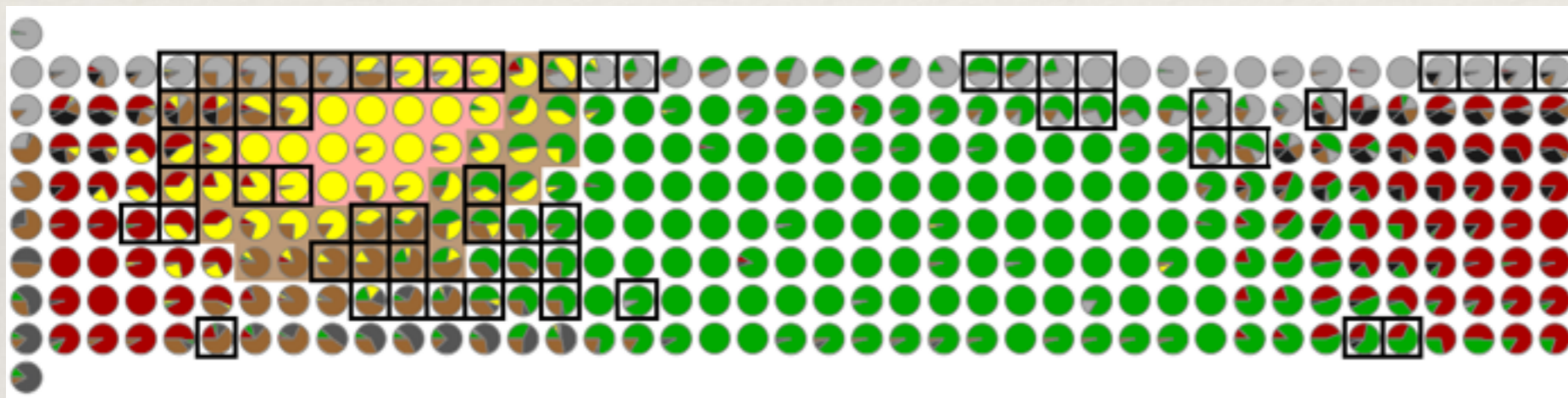
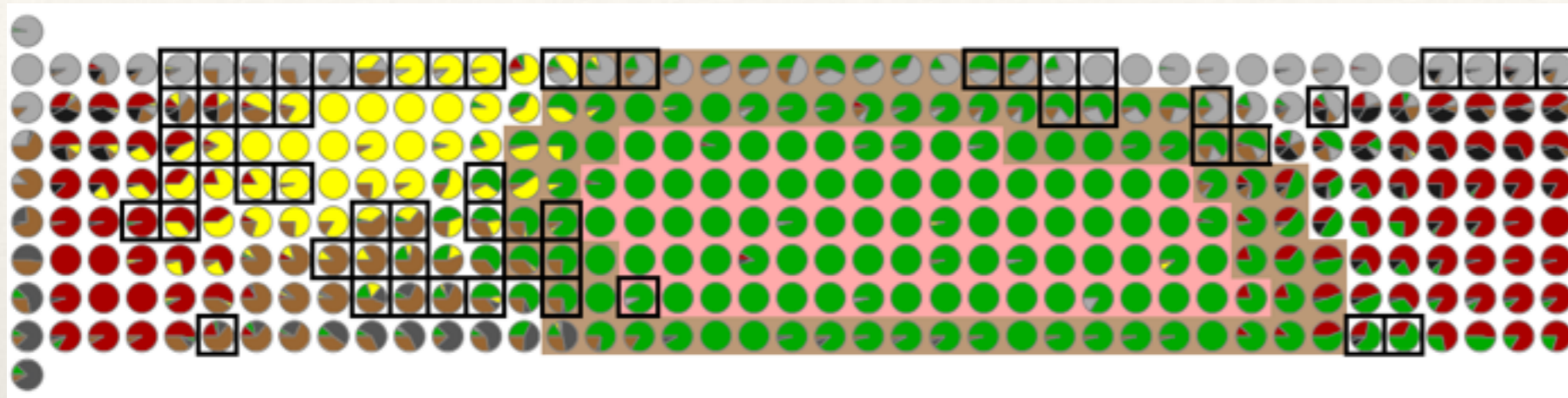
- ❖ There is a good agreement amongst individuals (often just one colour name) in the central region of the patch (the region of high congruence in naming)
- ❖ Regions of high stimuli frequency are grouped at the border regions of the patches.

Chavacano Philippines (WCS 24)



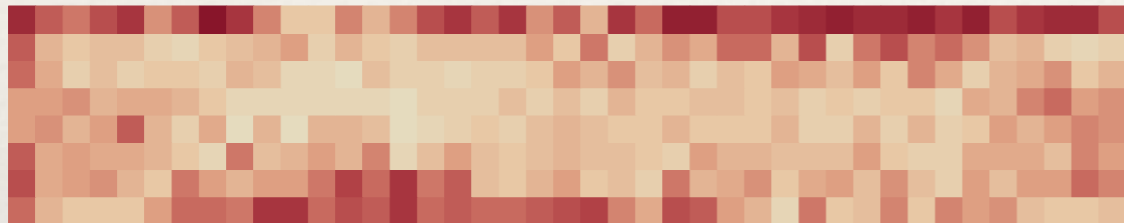
- ❖ Population uses the diversity of words for frequently present stimuli. But - they do not add new words to the vocabulary, they rather adapt the collective categorisation system to the frequency map of the environmental stimuli.

Seri Mexico (WCS 85)

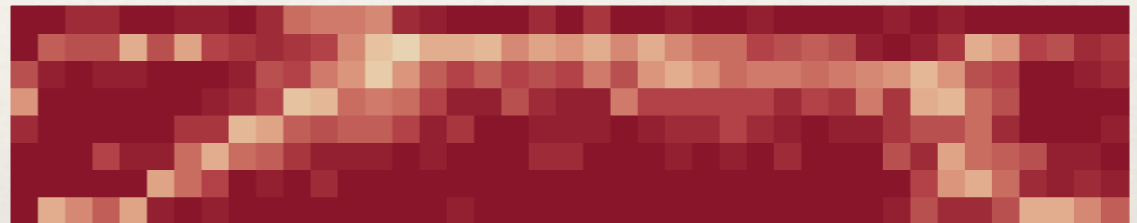


Methodological remarks

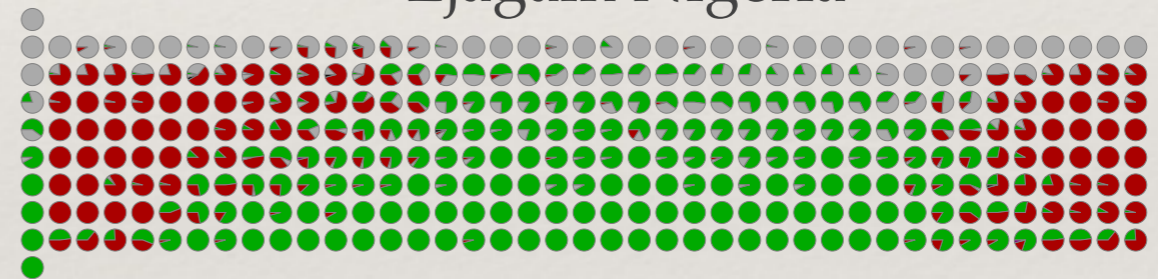
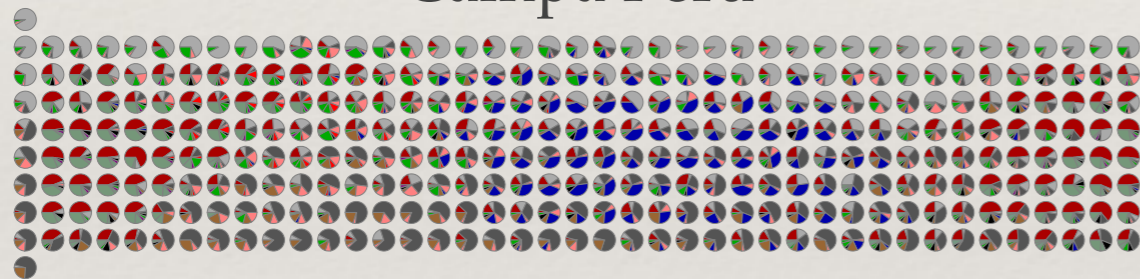
- ❖ not all cultures have patches: 1 (yellow) 2 (green) 3 (blue) or 4 (grue).
E.g. Campa Peru, Ejagam Nigeria.



Campa Peru



Ejagam Nigeria



- ❖ 64 cultures having at least one of them were chosen.