Searching for patterns in the World Color Survey

Gerhard Jäger
gerhard.jaeger@uni-tuebingen.de

July 2, 2009

University of Frankfurt

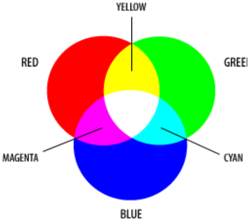
Overview

Structure of the talk

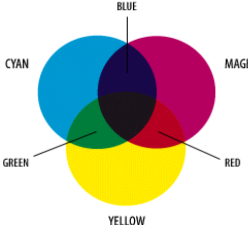
- the psychological color space
- Berlin and Kay's 1969 study
- the World Color Survey
- the distribution of focal colors
- categorization
- Principal Component Analysis
- clustering
- color categories are (more or less) convex

The psychological color space

- physical color space has infinite dimensionality — every wavelength within the visible spectrum is one dimension
- psychological color space is only 3-dimensional
- this fact is employed in technical devices like computer screens (additive color space) or color printers (subtractive color space)



additive color space

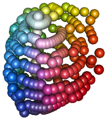
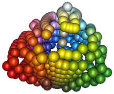


subtractive color space

The psychological color space

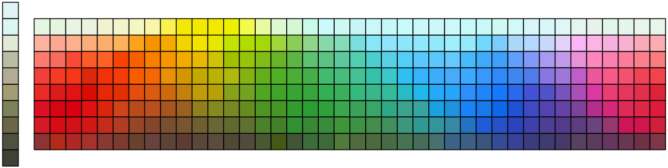
- psychologically correct color space should not only correctly represent the topology of, but also the distances between colors
- distance is inverse function of perceived similarity
- L*a*b* color space has this property
- three axes:
  - black — white
  - red — green
  - blue — yellow
- irregularly shaped 3d **color solid**

The color solid

The Munsell chart

- for psychological investigations, the *Munsell chart* is being used
- 2d-rendering of the surface of the color solid
  - 8 levels of lightness
  - 40 hues
- plus: black–white axis with 8 shaded of grey in between
- neighboring chips differ in the minimally perceivable way

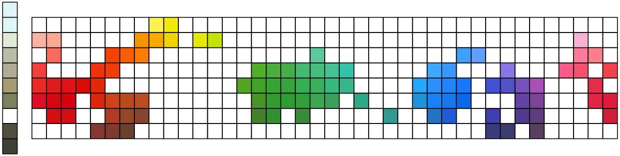


Berlin and Kay 1969

- pilot study how different languages carve up the color space into categories
- informants: speakers of 20 typologically distant languages (who happened to be around the Bay area at the time)
- questions (using the Munsell chart):
  - What are the basic color terms of your native language?
  - What is the extension of these terms?
  - What are the prototypical instances of these terms?
- results are not random
- indicate that there are universal tendencies in color naming systems

Berlin and Kay 1969

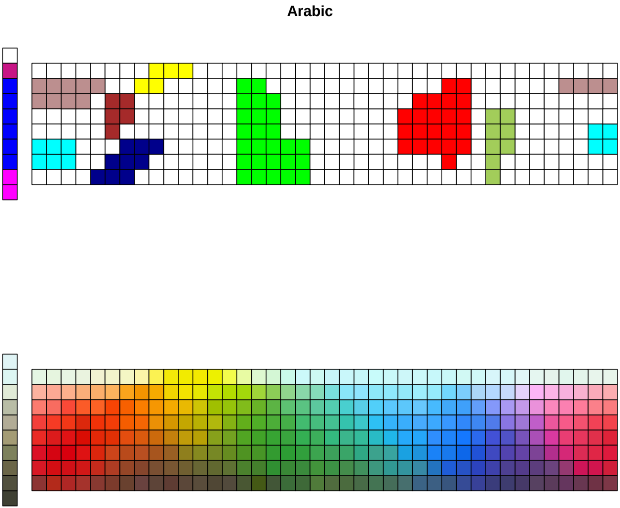
- distribution of focal colors:



- essentially correspond to the centers of the English categories *black, white, red, green, yellow, blue, purple, orange, brown, grey, pink*

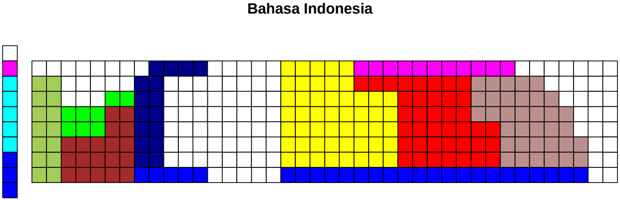
Berlin and Kay 1969

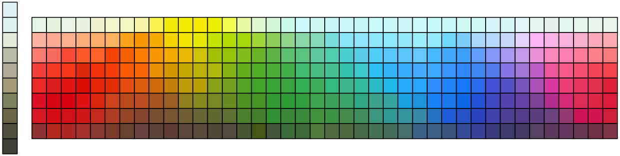
- extensions



Berlin and Kay 1969

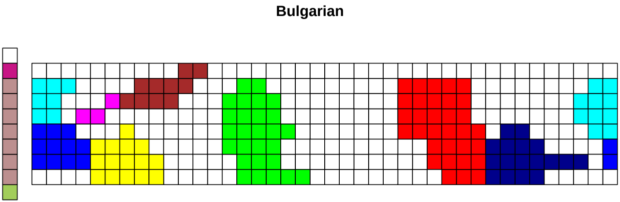
- extensions

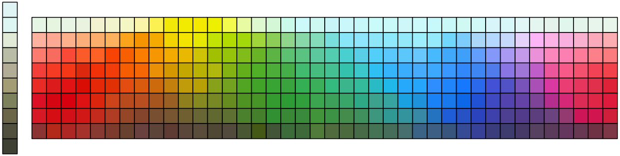




Berlin and Kay 1969

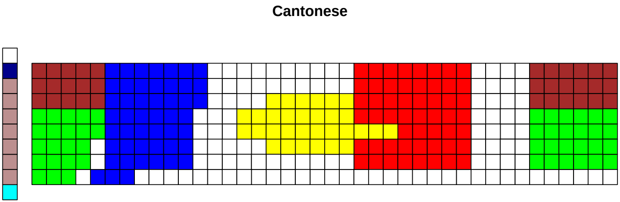
- extensions

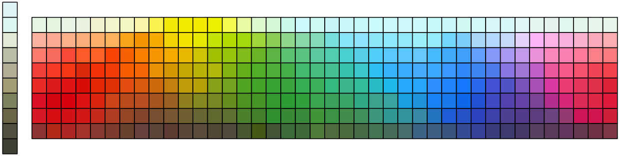




Berlin and Kay 1969

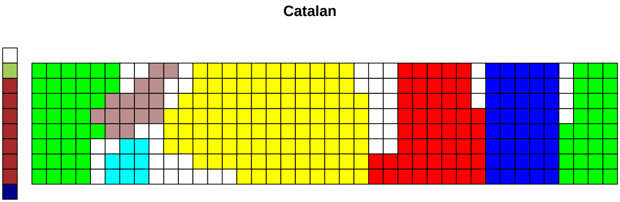
- extensions

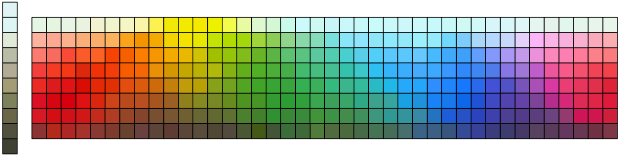




Berlin and Kay 1969

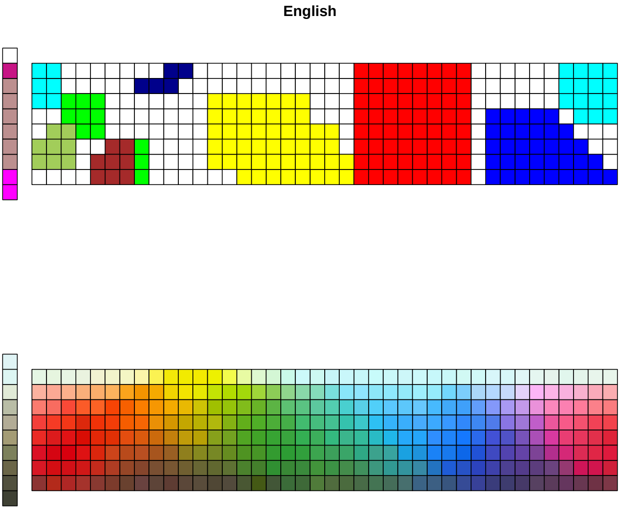
- extensions





Berlin and Kay 1969

- extensions

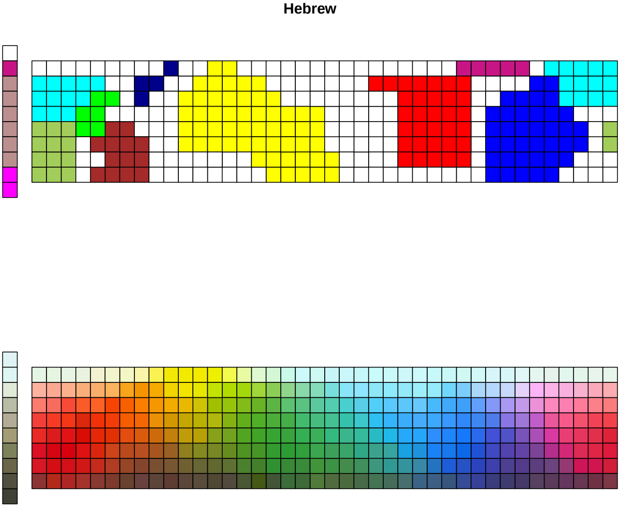


Eberhard Karls Universität Tübingen

sfs seminar für sprachwissenschaft

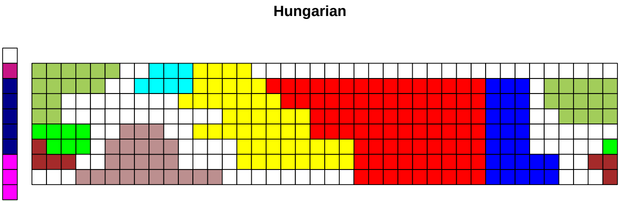
Berlin and Kay 1969

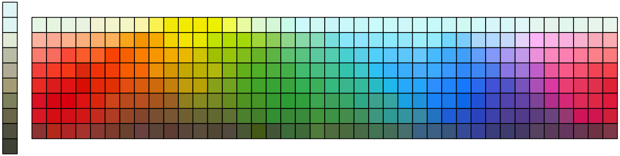
- extensions



Berlin and Kay 1969

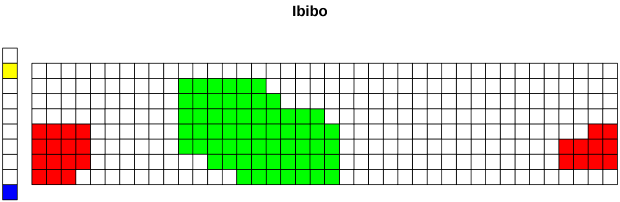
- extensions

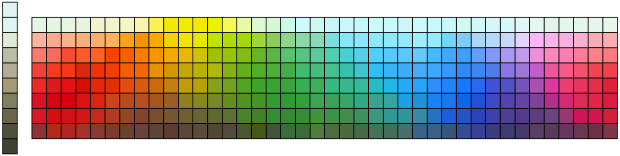




Berlin and Kay 1969

- extensions

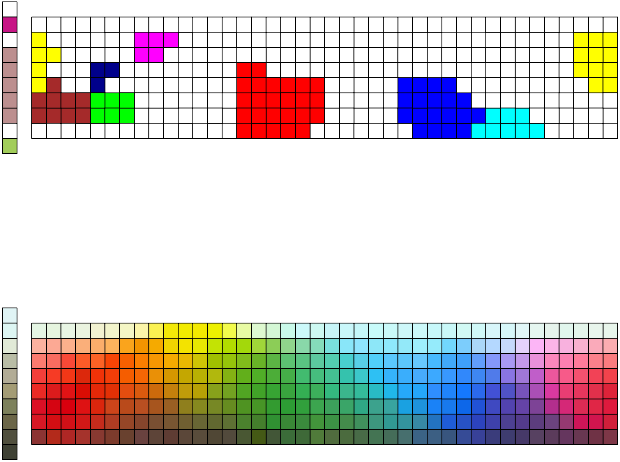




Berlin and Kay 1969

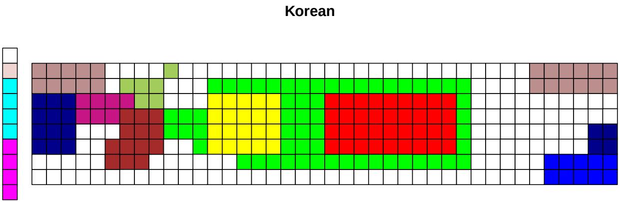
- extensions

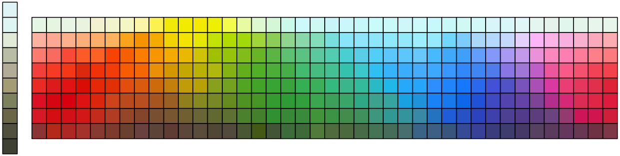
Japanese



Berlin and Kay 1969

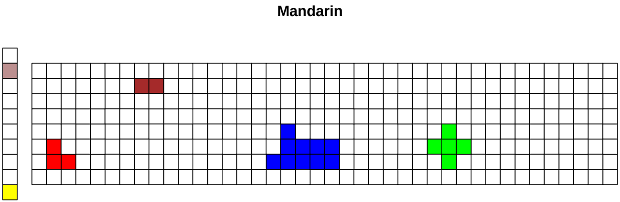
- extensions

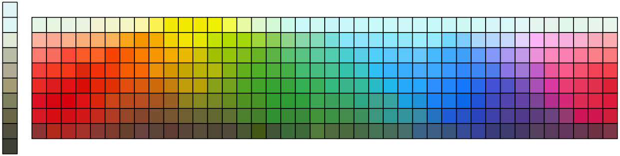




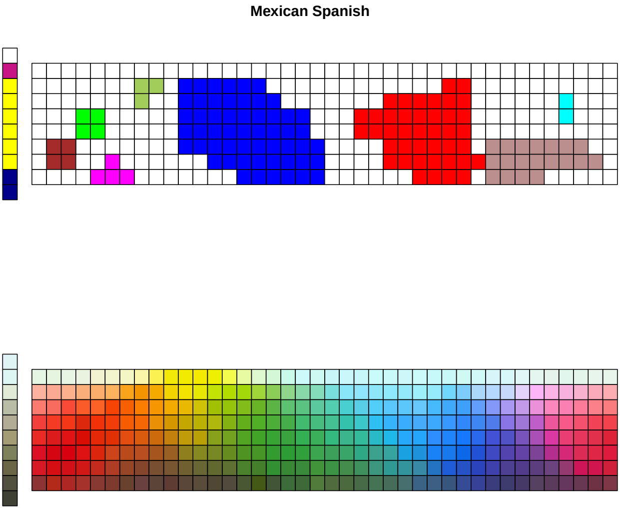
Berlin and Kay 1969

- extensions





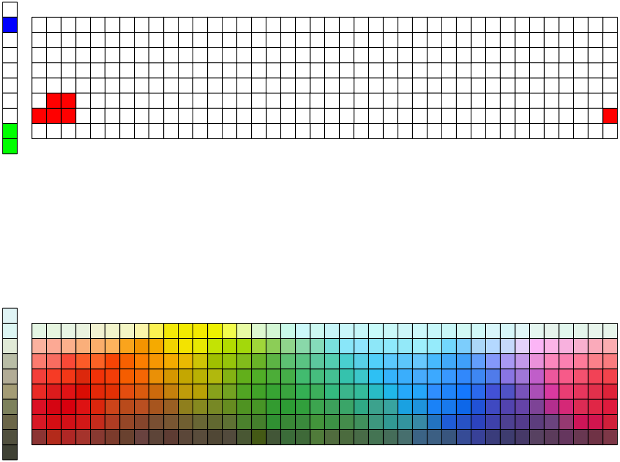
- extensions



Berlin and Kay 1969

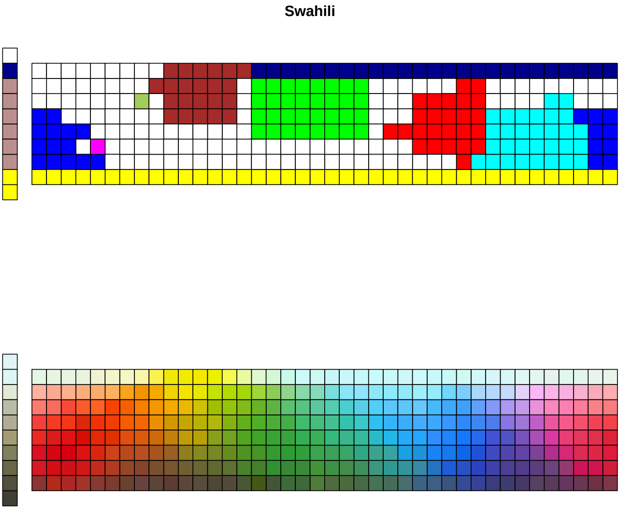
- extensions

Pomo



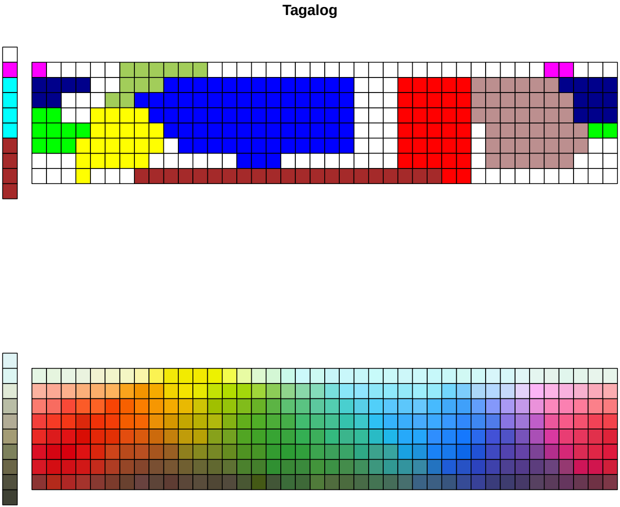
Berlin and Kay 1969

- extensions



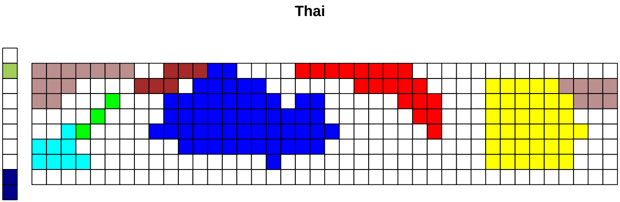
Berlin and Kay 1969

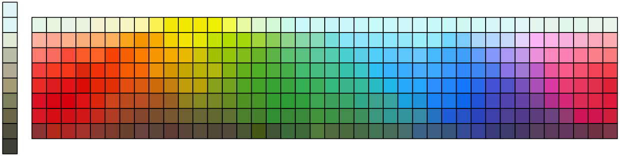
- extensions



Berlin and Kay 1969

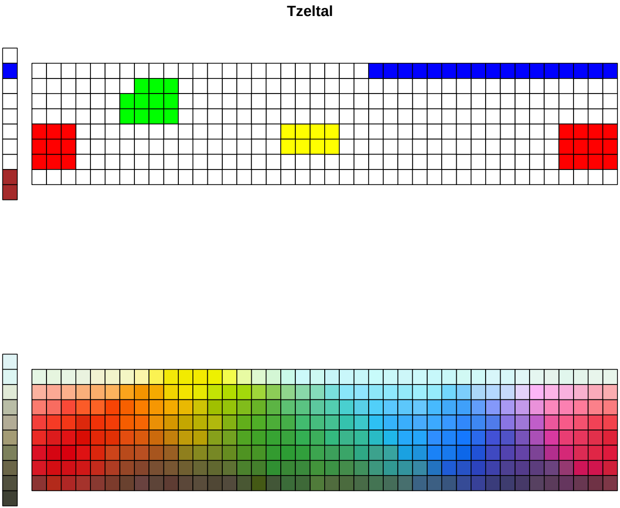
- extensions





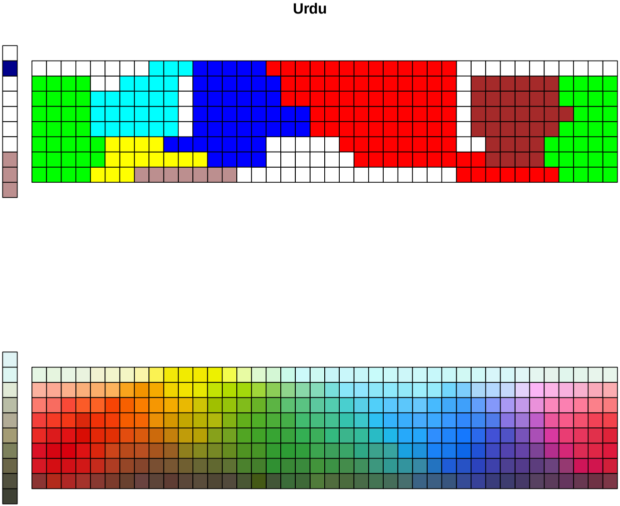
Berlin and Kay 1969

- extensions



Berlin and Kay 1969

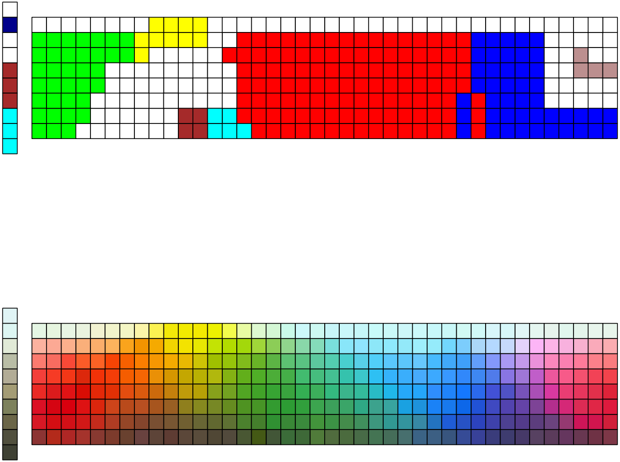
- extensions



Berlin and Kay 1969

- extensions

Vietnamese



Berlin and Kay 1969

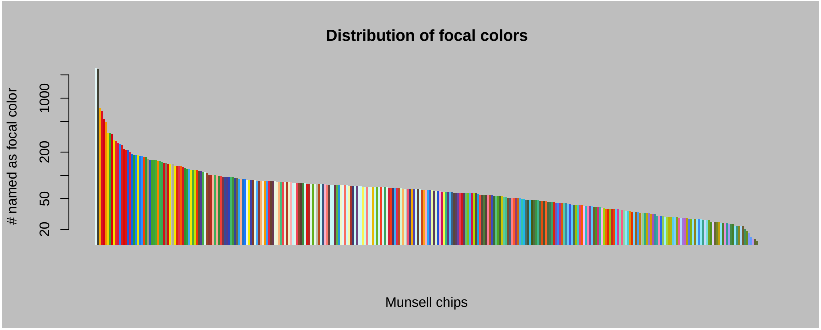
- identification of absolute and implicational universals, like
  - all languages have words for *black* and *white*
  - if a language has a word for *yellow*, it has a word for *red*
  - if a language has a word for *pink*, it has a word for *blue*
  - ...

The World Color Survey

- B&K was criticized for methodological reasons
- in response, in 1976 Kay and co-workers launched the *world color survey*
- investigation of 110 non-written languages from around the world
- around 25 informants per language
- two tasks:
  - the 330 Munsell chips were presented to each test person one after the other in random order; they had to assign each chip to basic some color term from their native language
  - for each native basic color term, each informant identified the prototypical instance(s)
- data are publicly available under http://www.icsi.berkeley.edu/wcs/

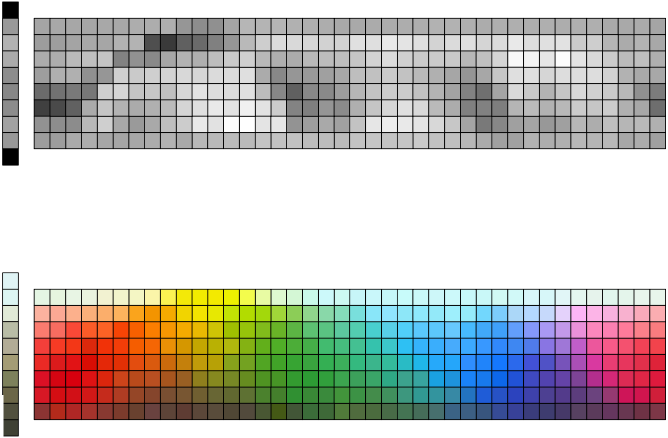
Data digging in the WCS

- distribution of focal colors across all informants:



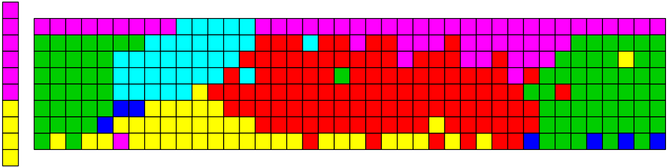
Data digging in the WCS

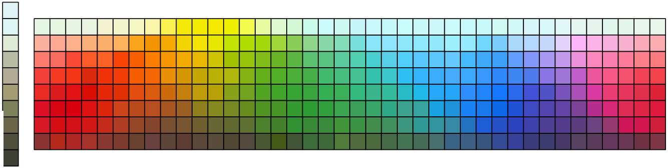
- distribution of focal colors across all informants:



Data digging in the WCS

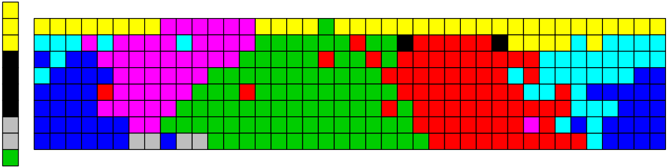
- partition of a randomly chosen informant from a randomly chosen language

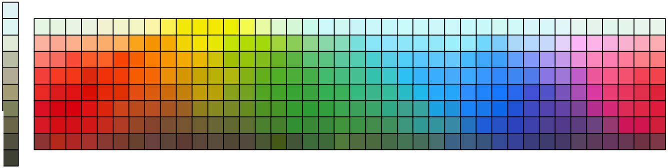




Data digging in the WCS

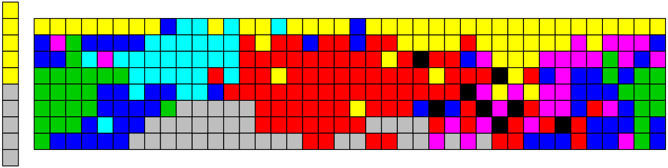
- partition of a randomly chosen informant from a randomly chosen language

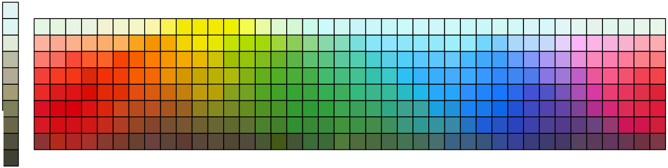




Data digging in the WCS

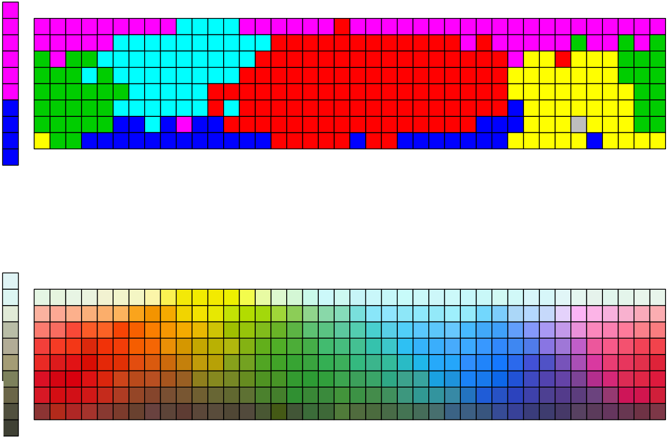
- partition of a randomly chosen informant from a randomly chosen language





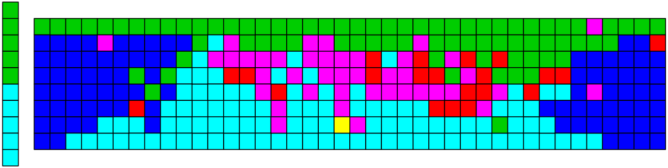
Data digging in the WCS

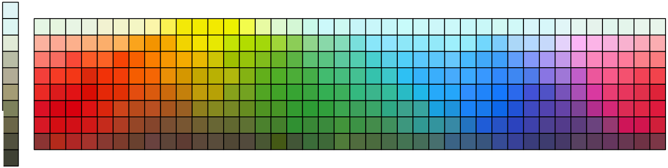
- partition of a randomly chosen informant from a randomly chosen language



Data digging in the WCS

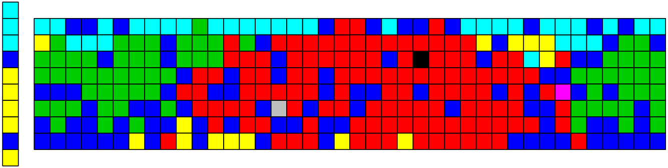
- partition of a randomly chosen informant from a randomly chosen language

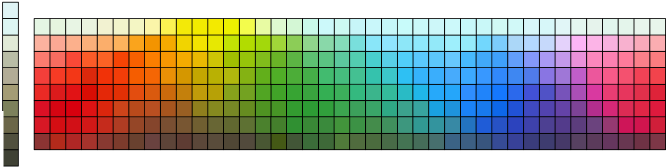




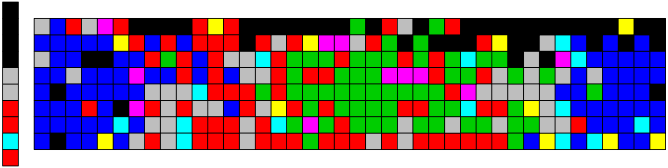
Data digging in the WCS

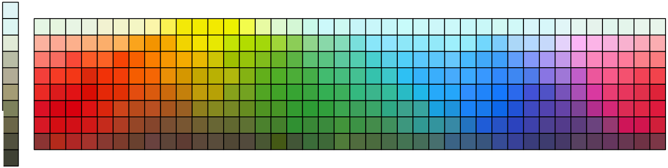
- partition of a randomly chosen informant from a randomly chosen language





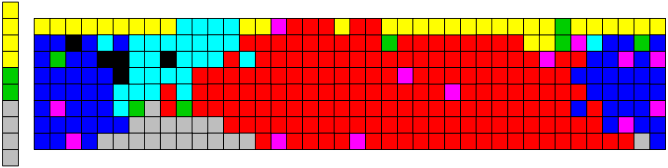
- partition of a randomly chosen informant from a randomly chosen language

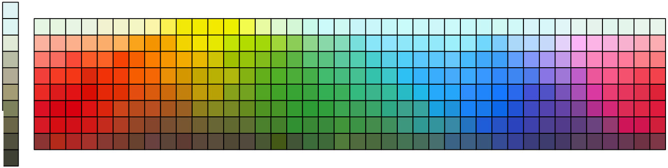




Data digging in the WCS

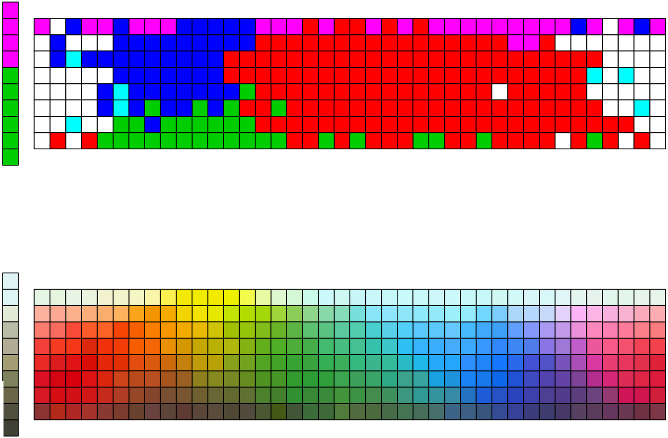
- partition of a randomly chosen informant from a randomly chosen language





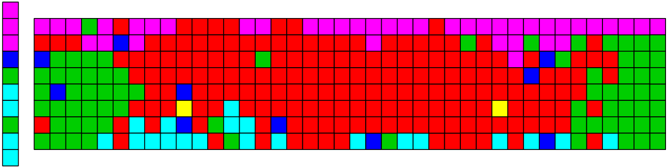
Data digging in the WCS

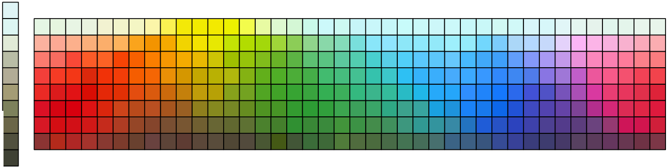
- partition of a randomly chosen informant from a randomly chosen language



Data digging in the WCS

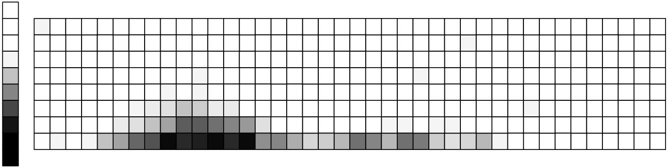
- partition of a randomly chosen informant from a randomly chosen language

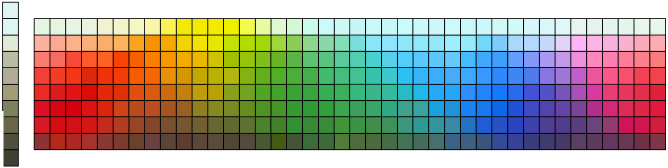




Data digging in the WCS

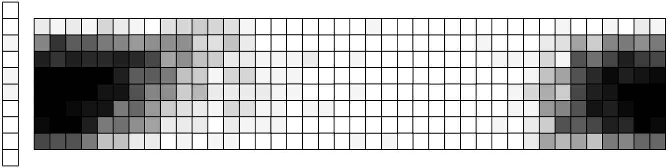
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

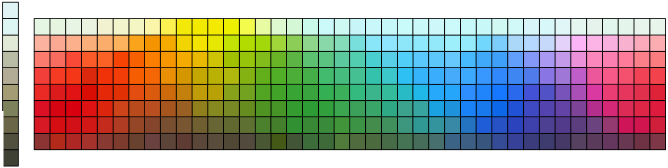




Data digging in the WCS

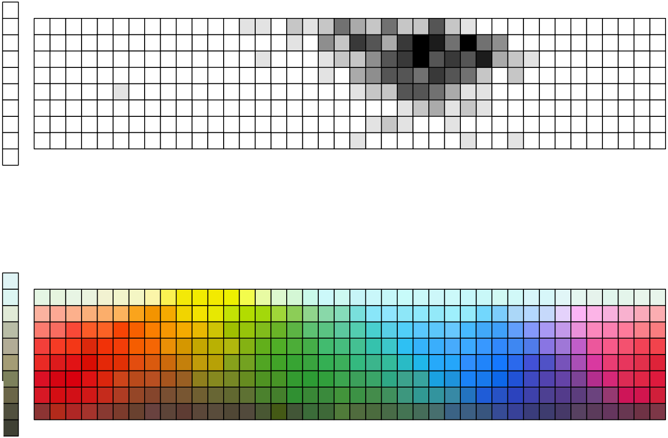
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language





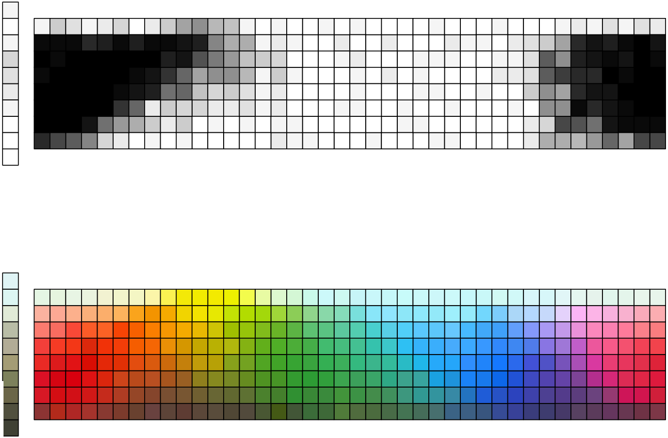
Data digging in the WCS

- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language



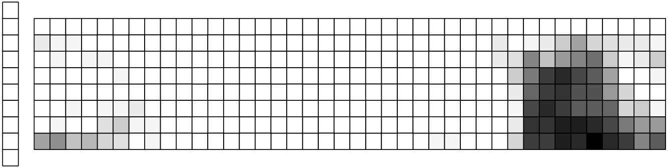
Data digging in the WCS

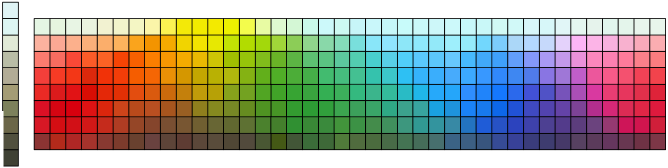
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language



Data digging in the WCS

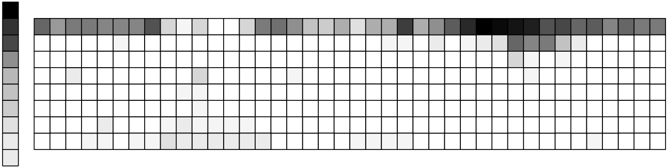
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

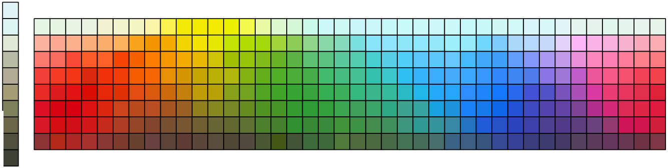




Data digging in the WCS

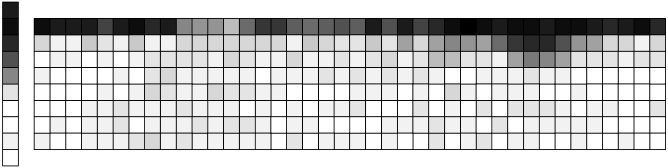
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

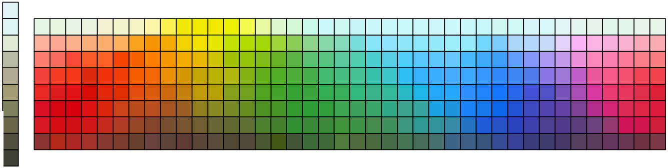




Data digging in the WCS

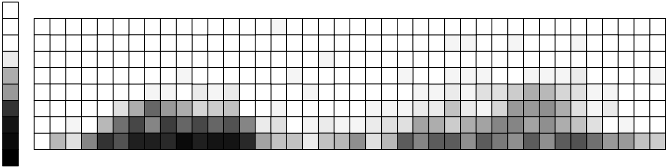
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

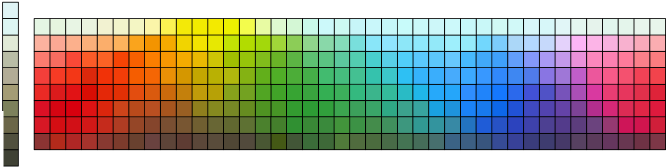




Data digging in the WCS

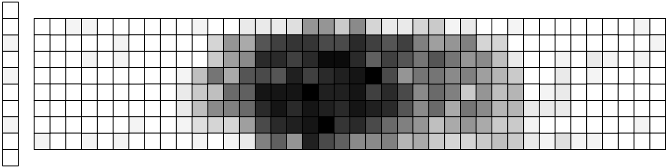
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

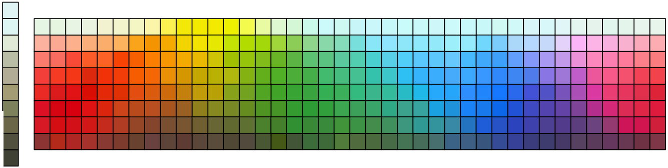




Data digging in the WCS

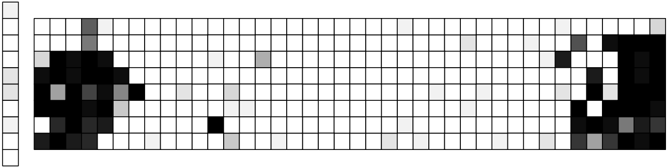
- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language

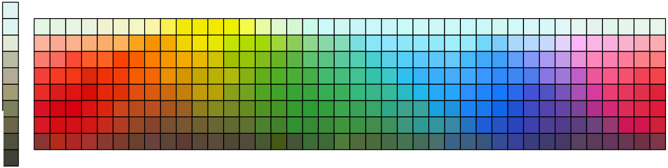




Data digging in the WCS

- extension of a randomly chosen term from a randomly chosen language, averaged over all informants from that language





What is the extension of categories?

- data from individual informants are extremely noisy
- averaging over all informants from a language helps, but there is still noise, plus dialectal variation
- desirable: distinction between “genuine” variation and noise

Principal Component Analysis

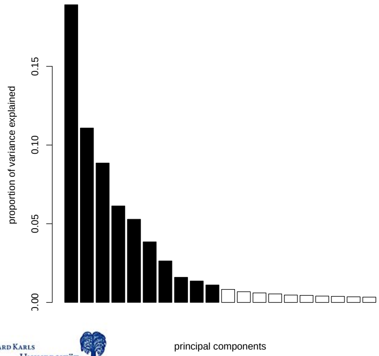
- technique to reduce dimensionality of data
- input: set of vectors in an n-dimensional space
- first step: rotate the coordinate system, such that
  - the new n coordinates are orthogonal to each other
  - the variations of the data along the new coordinates are stochastically independent
- second step:
  - choose a suitable $m < n$
  - project the data on those m new coordinates where the data have the highest variance

Principal Component Analysis

- alternative formulation:
  - choose an m-dimensional linear sub-manifold of your n-dimensional space
  - project your data onto this manifold
  - when doing so, pick your sub-manifold such that the average squared distance of the data points from the sub-manifold is minimized
- intuition behind this formulation:
  - data are “actually” generated in an m-dimensional space
  - observations are disturbed by n-dimensional noise
  - PCA is a way to reconstruct the underlying data distribution
- applications: picture recognition, latent semantic analysis, statistical data analysis in general, data visualization, ...

Applying PCA to WCS-categories

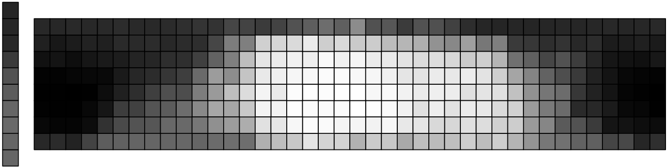
- data: informant-category pairs
- 330 dimensions (each Munsell color is one dimension)
- each informant-category pair assigns 1 to the colors that belong to that category, and 0 else

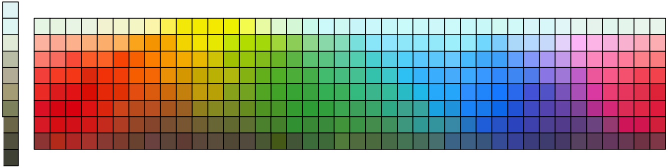


- first seven principal components jointly explain 60% of the variance in the data
- each PC after PC10 only marginally increases proportion of variance explained
- so let's say $m = 10$

PC1

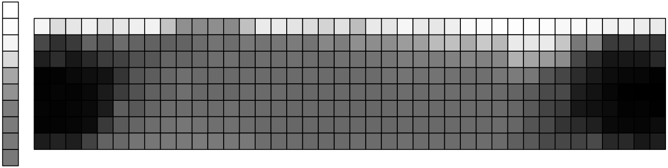
- green/blue vs. white/red/yellow

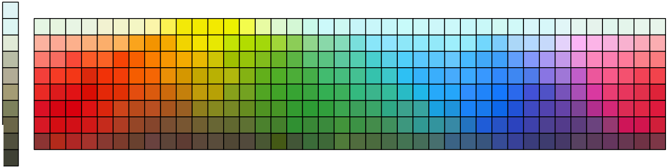




PC2

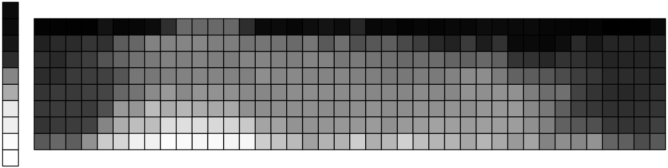
- white vs. red

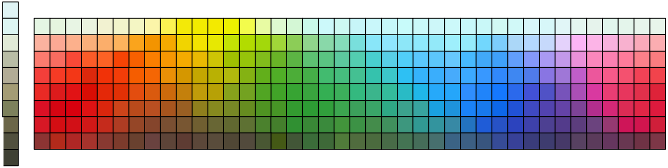




PC3

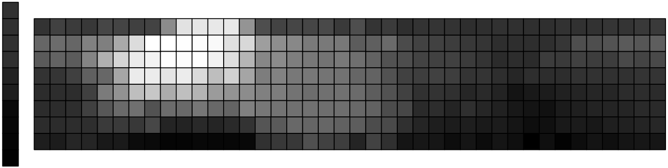
- black vs. red/white

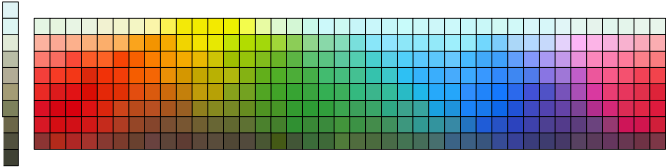




PC4

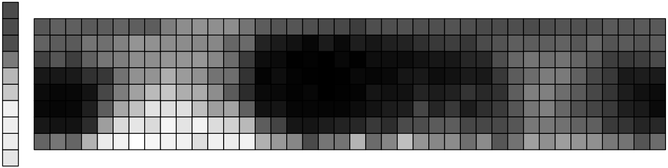
- yellow vs. black/white/blue/red

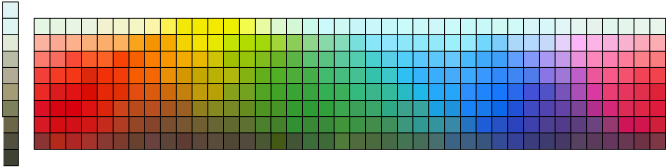




PC5

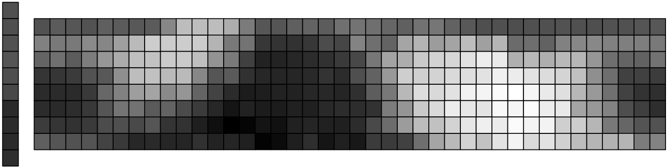
- black vs. red/green/blue

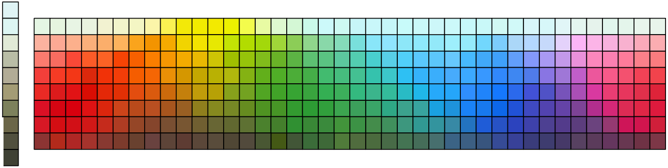




PC6

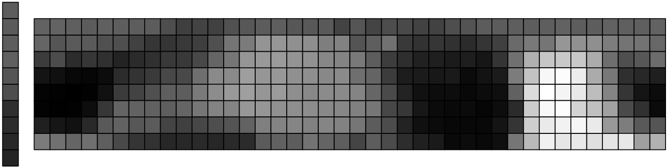
- blue/yellow vs. red/green

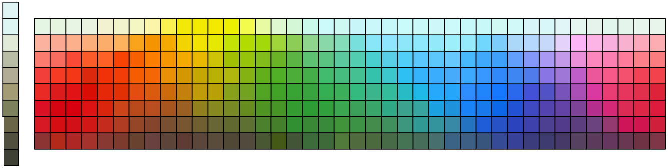




PC7

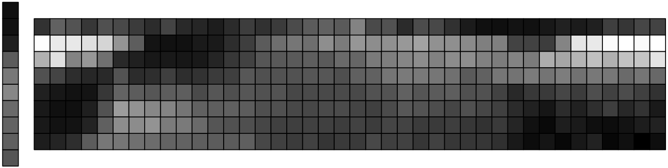
- purple vs. red/blue/black

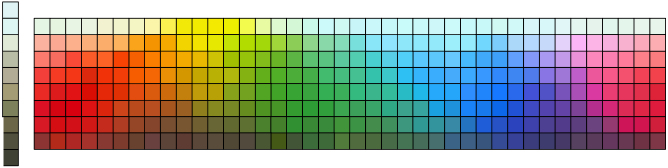




PC8

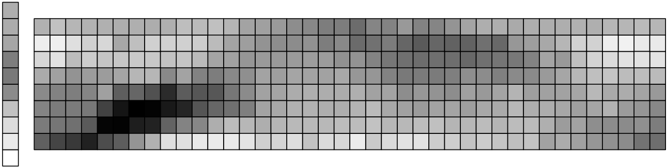
- pink vs. red/yellow/white

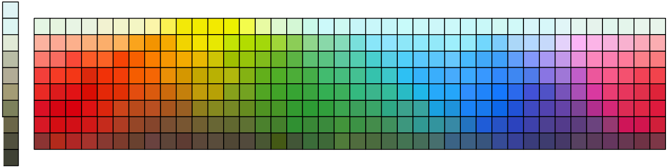




PC9

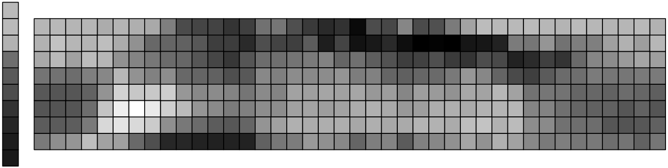
- brown vs. black/pink

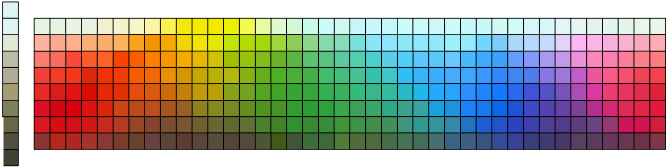




PC10

- brown vs. light blue/yellow/black

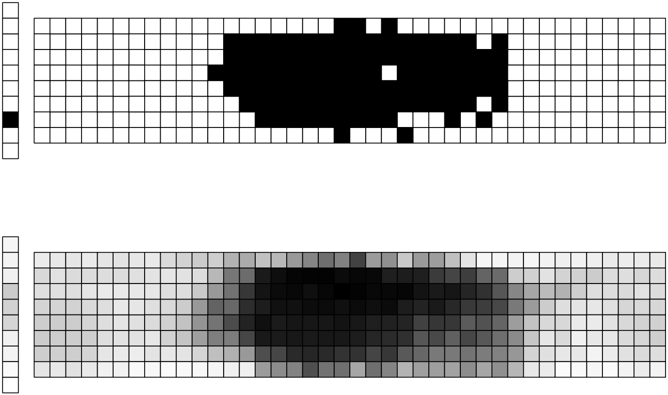




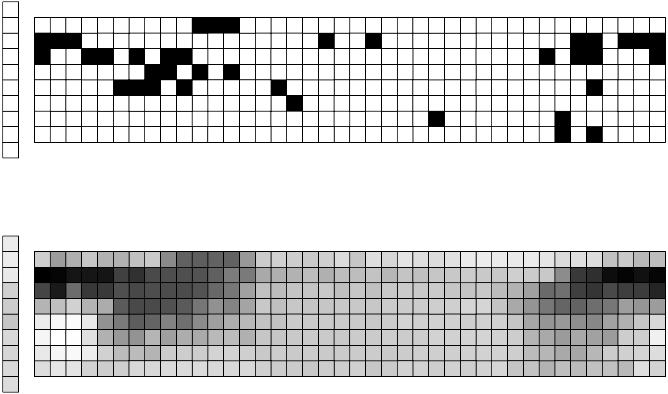
Projecting observed data on 10d-manifold

- noise removal: project observed data onto the lower-dimensional submanifold that was obtained via PCA
- in our case: noisy binary categories are mapped to smoothed fuzzy categories (= probability distributions over Munsell chips)
- some examples:

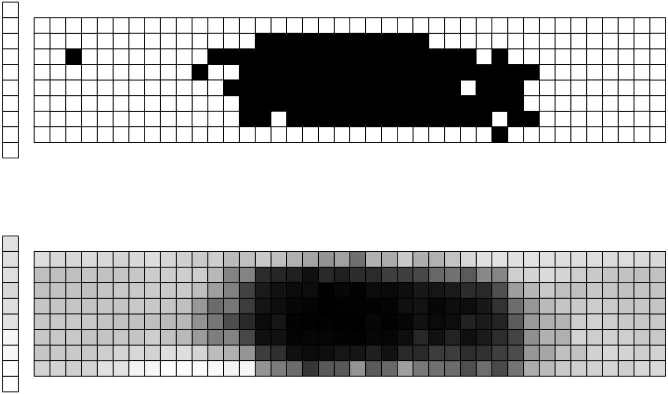
Projecting observed data on 10d-manifold



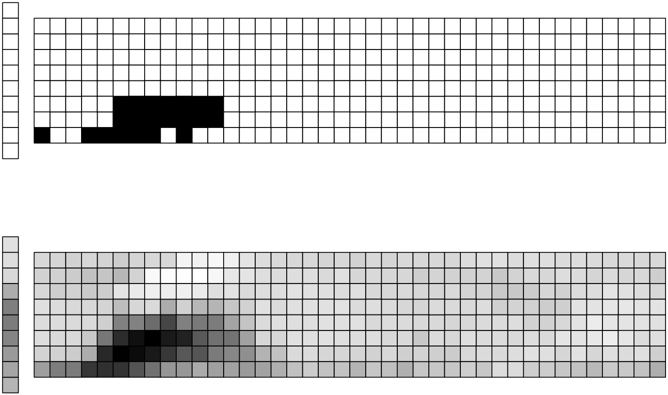
Projecting observed data on 10d-manifold



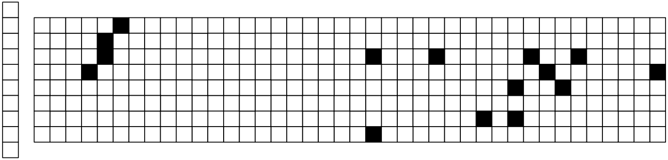
Projecting observed data on 10d-manifold

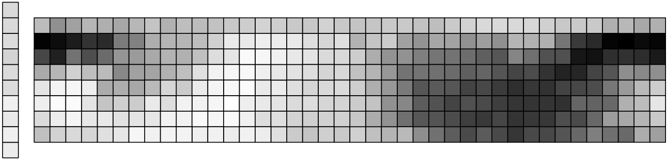


Projecting observed data on 10d-manifold

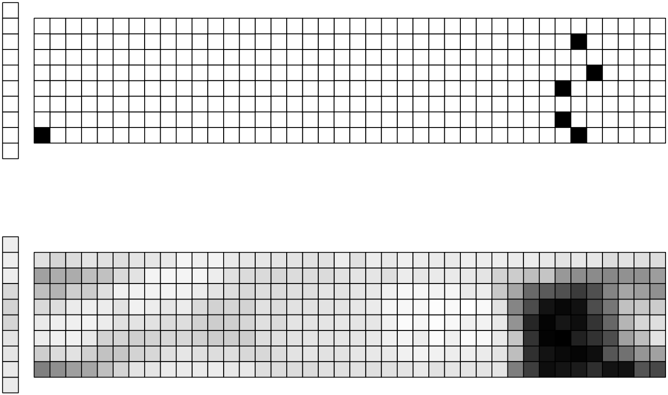


Projecting observed data on 10d-manifold

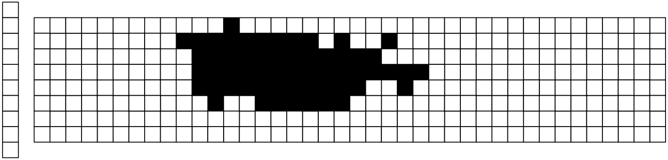


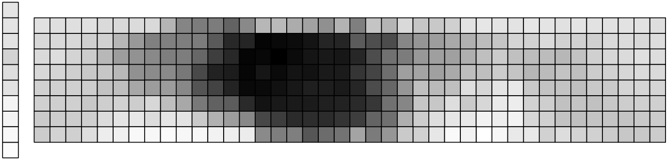


Projecting observed data on 10d-manifold

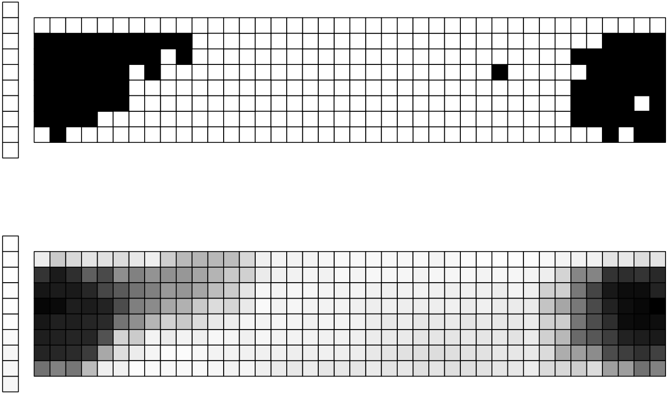


Projecting observed data on 10d-manifold

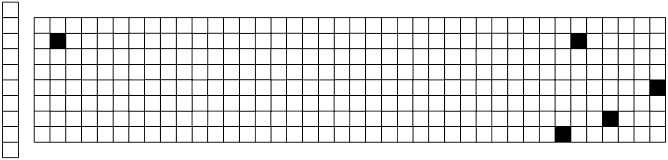


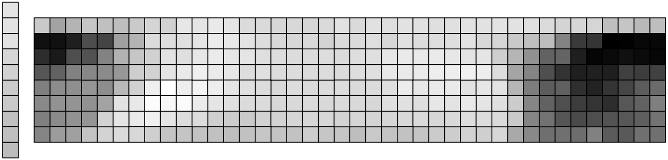


Projecting observed data on 10d-manifold

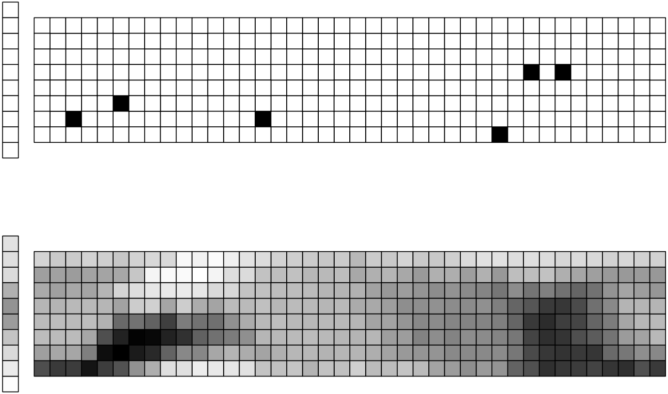


Projecting observed data on 10d-manifold

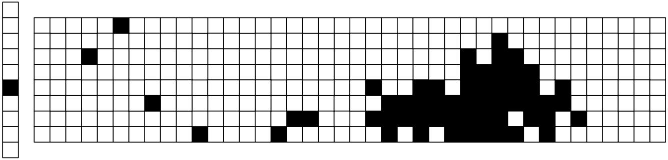


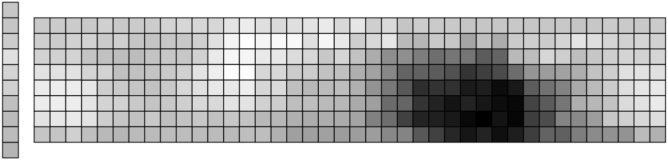


Projecting observed data on 10d-manifold

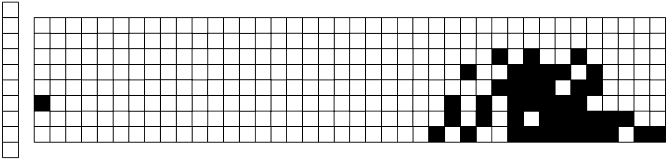


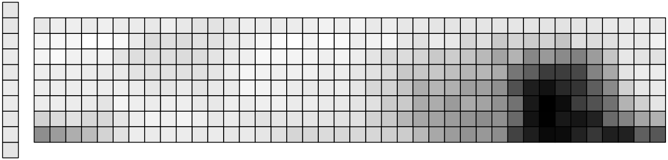
Projecting observed data on 10d-manifold



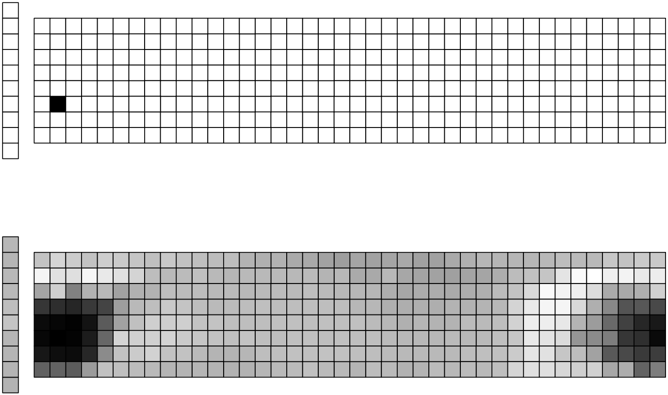


Projecting observed data on 10d-manifold

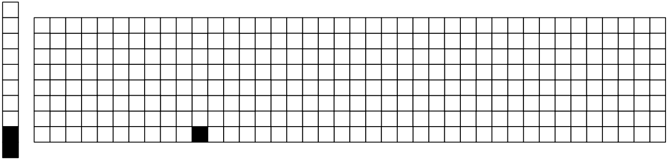


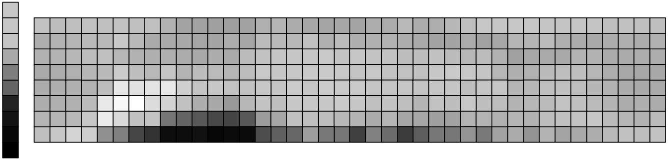


Projecting observed data on 10d-manifold

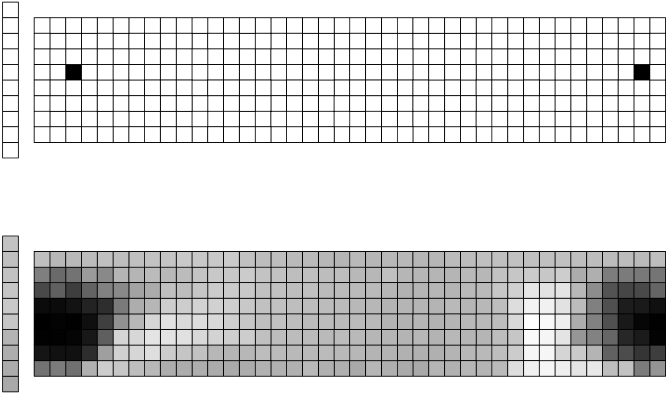


Projecting observed data on 10d-manifold

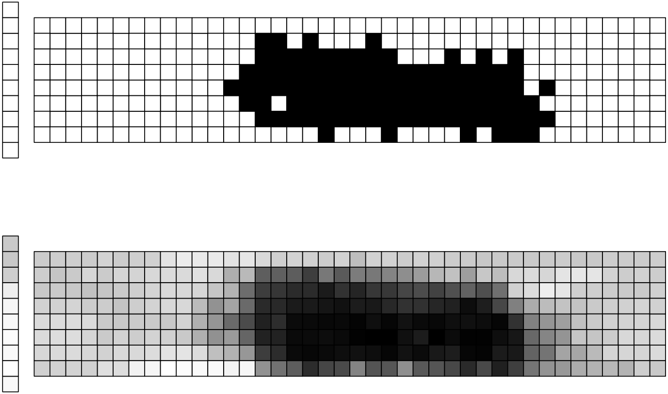




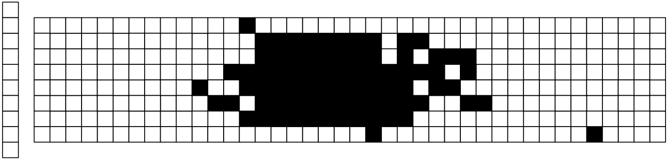
Projecting observed data on 10d-manifold

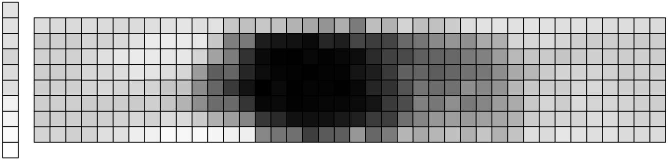


Projecting observed data on 10d-manifold

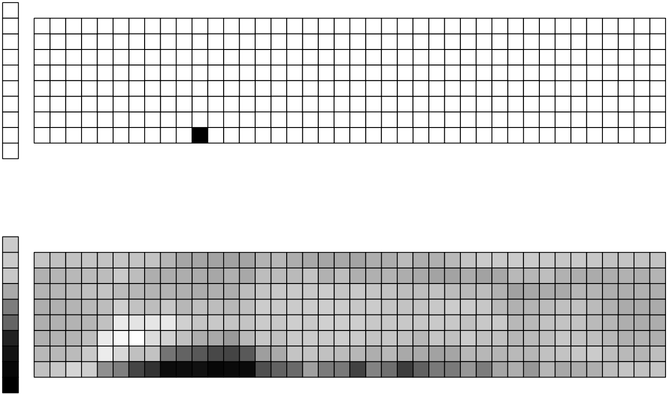


Projecting observed data on 10d-manifold

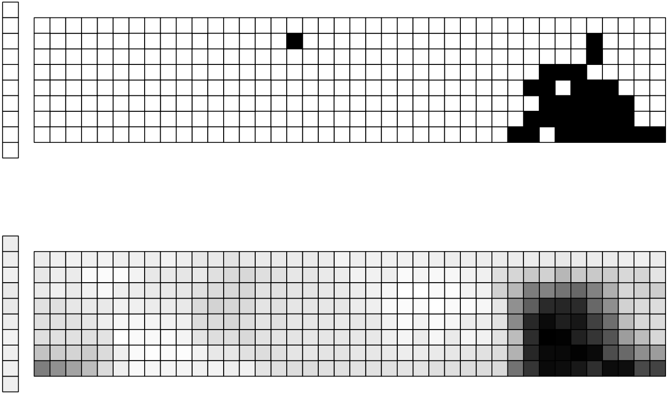




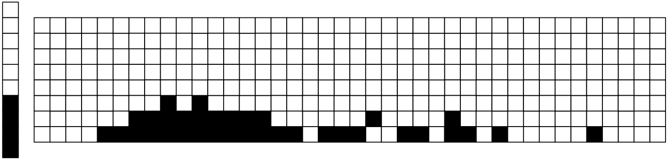
Projecting observed data on 10d-manifold

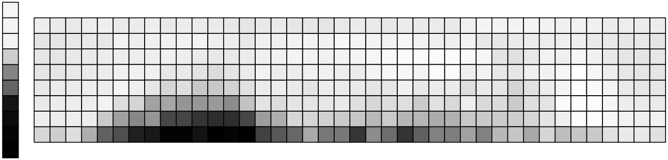


Projecting observed data on 10d-manifold



Projecting observed data on 10d-manifold

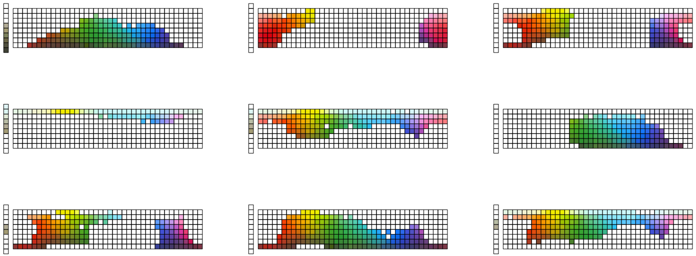




Smoothed partitions of the color space

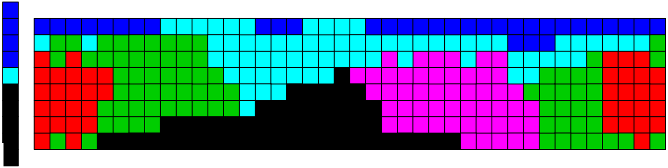
- vocabulary of a given language does not always form a partition
- many cases of (near) synonymy, hyponymy, and overlap
- for instance language 1 (Abidjy, Ivory Coast):

Smoothed partitions of the color space



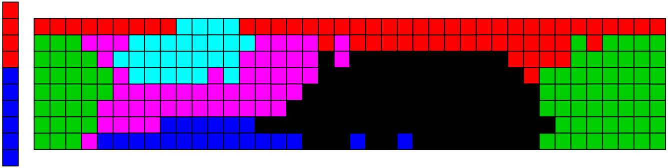
Smoothed partitions of the color space

- if two categories of one language have a correlation of at least $.5$, they are treated as synonyms
- process is repeated if remaining categories are independent or negatively correlated
- after this process, each Munsell chip c is assigned to the category that assigns the highest probability to c
- for Abidji, we get



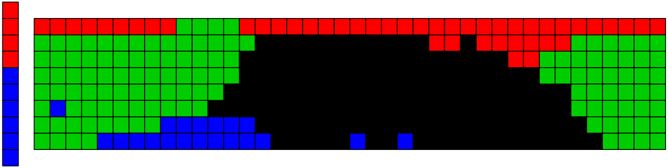
Smoothed partitions of the color space

- some more examples: Waorani (Ecuador)



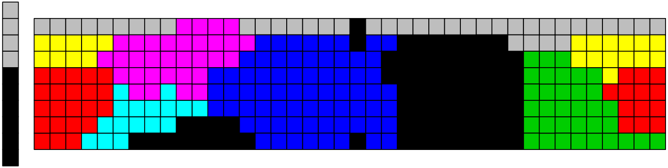
Smoothed partitions of the color space

- some more examples: Arabela (Peru)



Smoothed partitions of the color space

- some more examples: Camsa (Colombia)

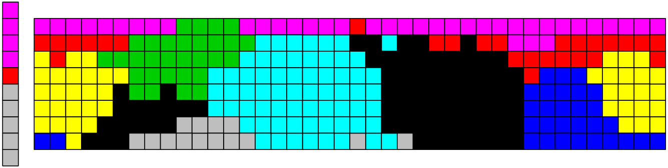


EBERHARD KARLS UNIVERSITÄT TÜBINGEN

sfs seminar für sprachwissenschaft

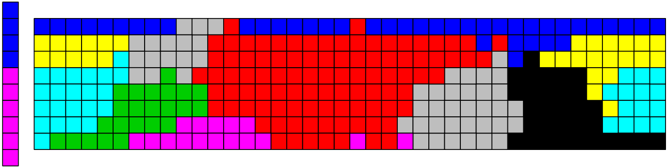
Smoothed partitions of the color space

- some more examples: Candoshi (Peru)



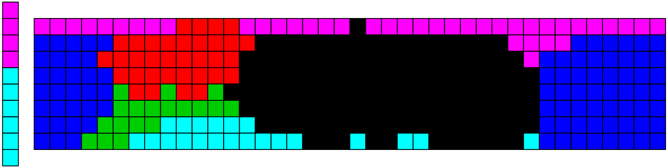
Smoothed partitions of the color space

- some more examples: Chinanteco (Mexico)



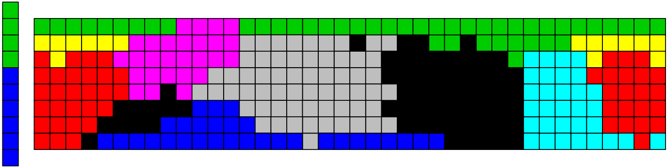
Smoothed partitions of the color space

- some more examples: Guarijio (Mexico)



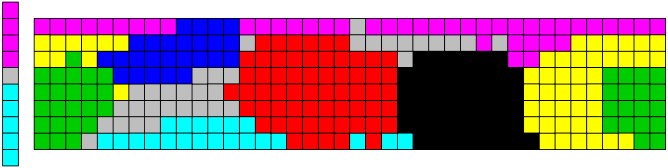
Smoothed partitions of the color space

- some more examples: Gunu (Cameroon)



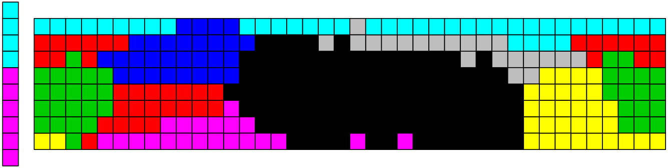
Smoothed partitions of the color space

- some more examples: Kalam (Papua New Guinea)



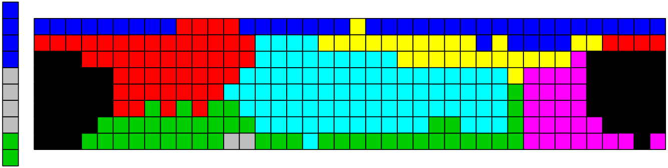
Smoothed partitions of the color space

- some more examples: Menye (Papua New Guinea)



Smoothed partitions of the color space

- some more examples: Tifal (Papua New Guinea)

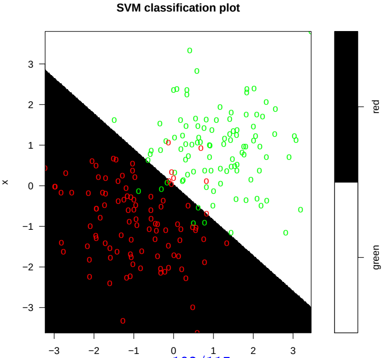


Convexity

- note: so far, we only used information from the WCS
- the location of the 330 Munsell chips in L*a*b* space played no role so far
- still, apparently partition cells always form continuous clusters in L*a*b* space
- Hypothesis (Gärdenfors): extension of color terms always form **convex** regions of L*a*b* space

Support Vector Machines

- supervised learning technique
- smart algorithm to classify data in a high-dimensional space by a (for instance) linear boundary
- minimizes number of mis-classifications if the training data are not linearly separable

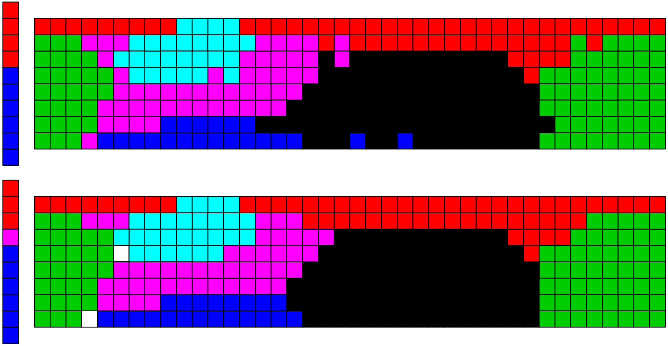


Convex partitions

- a binary linear classifier divides an n-dimensional space into two **convex** half-spaces
- intersection of two convex set is itself convex
- hence: intersection of k binary classifications leads to convex sets
- procedure: if a language partitions the Munsell space into m categories, train $\frac{m(m-1)}{2}$ many binary SVMs, one for each pair of categories **in L*a*b* space**
- leads to m convex sets (which need not split the L*a*b* space exhaustively)

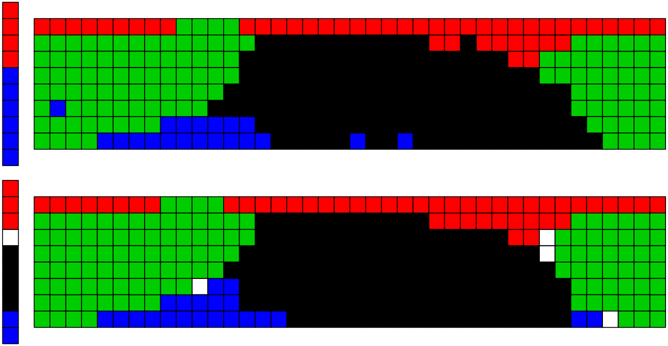
Convex approximation

- Waorani (Ecuador)



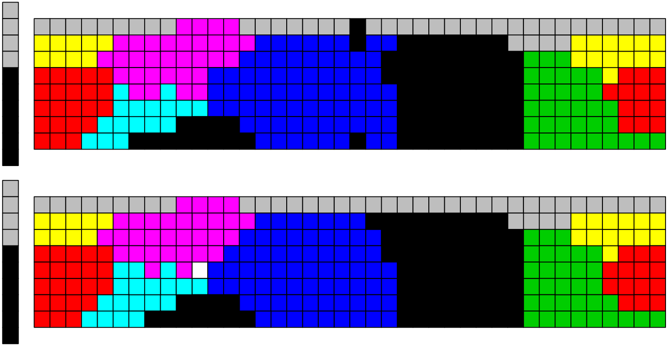
Convex approximation

- Arabela (Peru)



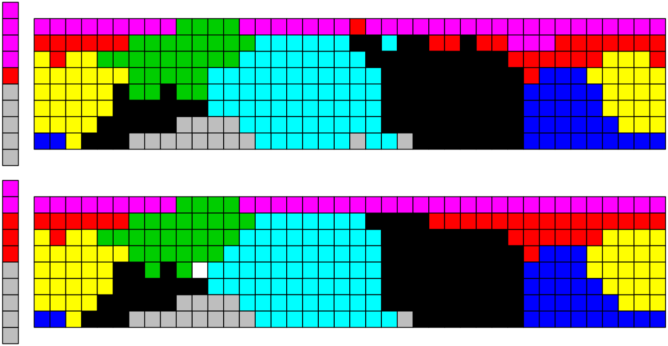
Convex approximation

- Camsa (Colombia)



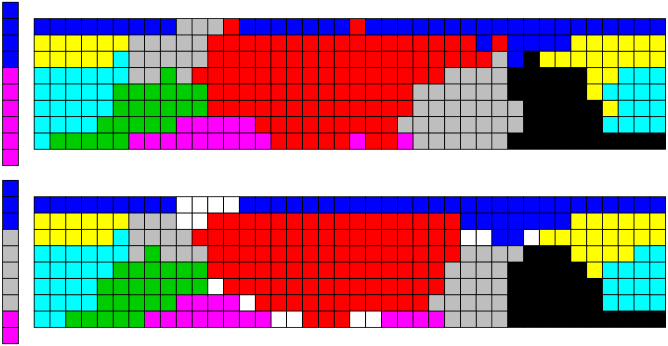
Convex approximation

- Candoshi (Peru)



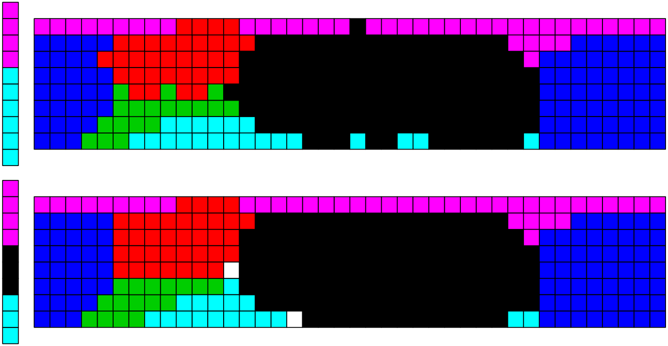
Convex approximation

- Chinanteco (Mexico)



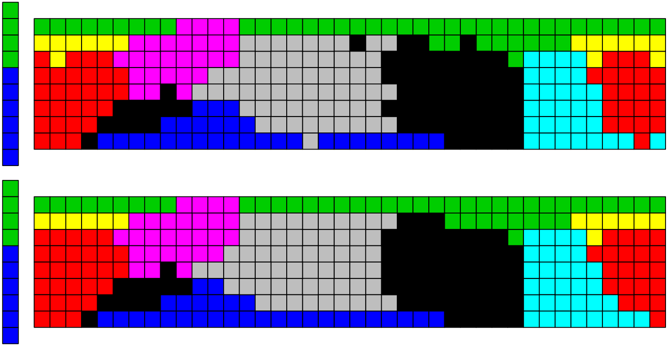
Convex approximation

- Guarijio (Mexico)



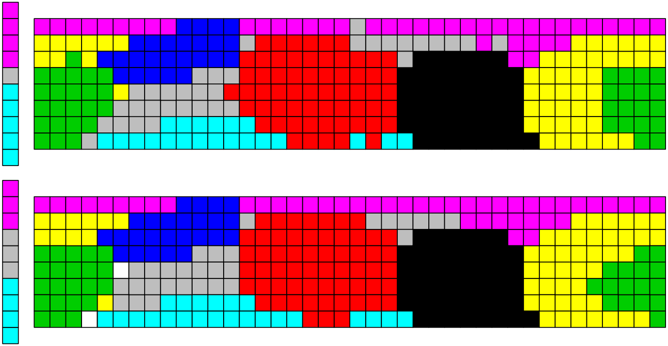
Convex approximation

- Gunu (Cameroon)



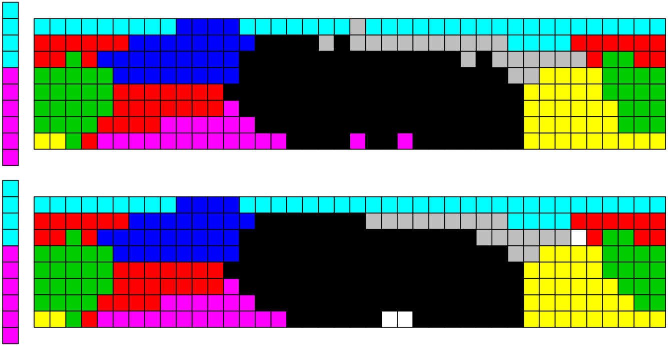
Convex approximation

- Kalam (Papua New Guinea)



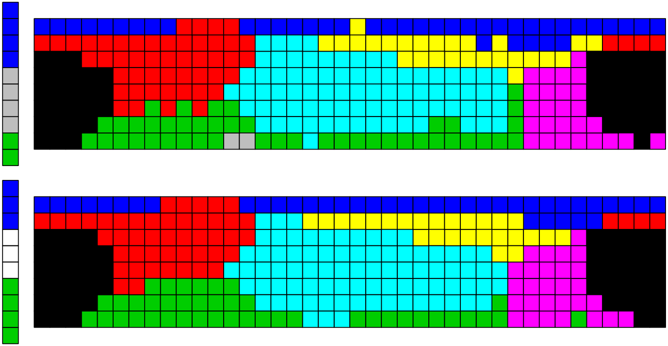
Convex approximation

- Menye (Papua New Guinea)



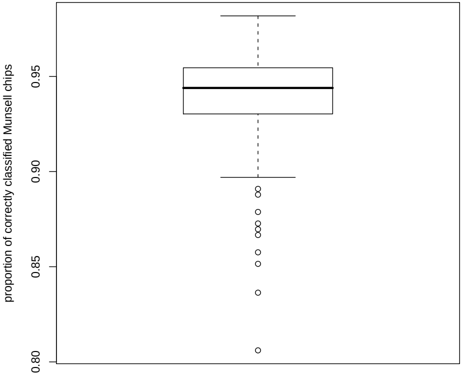
Convex approximation

- Tifal (Papua New Guinea)



Convex approximation

- on average, 93.7% of all Munsell chips are correctly classified by convex approximation

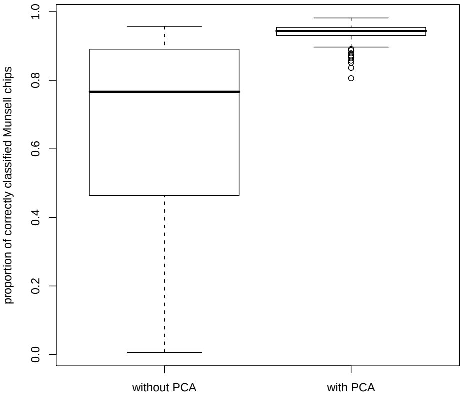


Eberhard Karls Universität Tübingen

sfs seminar für sprachwissenschaft

Convex approximation

- compare to the outcome of the same procedure without PCA:



Conclusion

- empirical support for Gärdenfors' thesis that natural properties are convex sets
- quantitative data analysis reveals robust universal tendencies
- techniques from statistical pattern recognition are useful for typological studies
- R is a great tool