Projections

As you watch the projections dance across the screen, we are scanning for directions that are “interesting”, providing us with a view into the clustering or grouping of data that might not be immediately evident otherwise

It turns out (a consequence of the Central Limit Theorem) that these projected views of the data will be “uninteresting” in that they will look like a bivariate normal distribution

This, then, becomes one possible definition of “uninteresting” and we can score views by how dissimilar they are from this distribution -- In the late 1970s and early 1980s, this led to a statistical technique known as projection pursuit

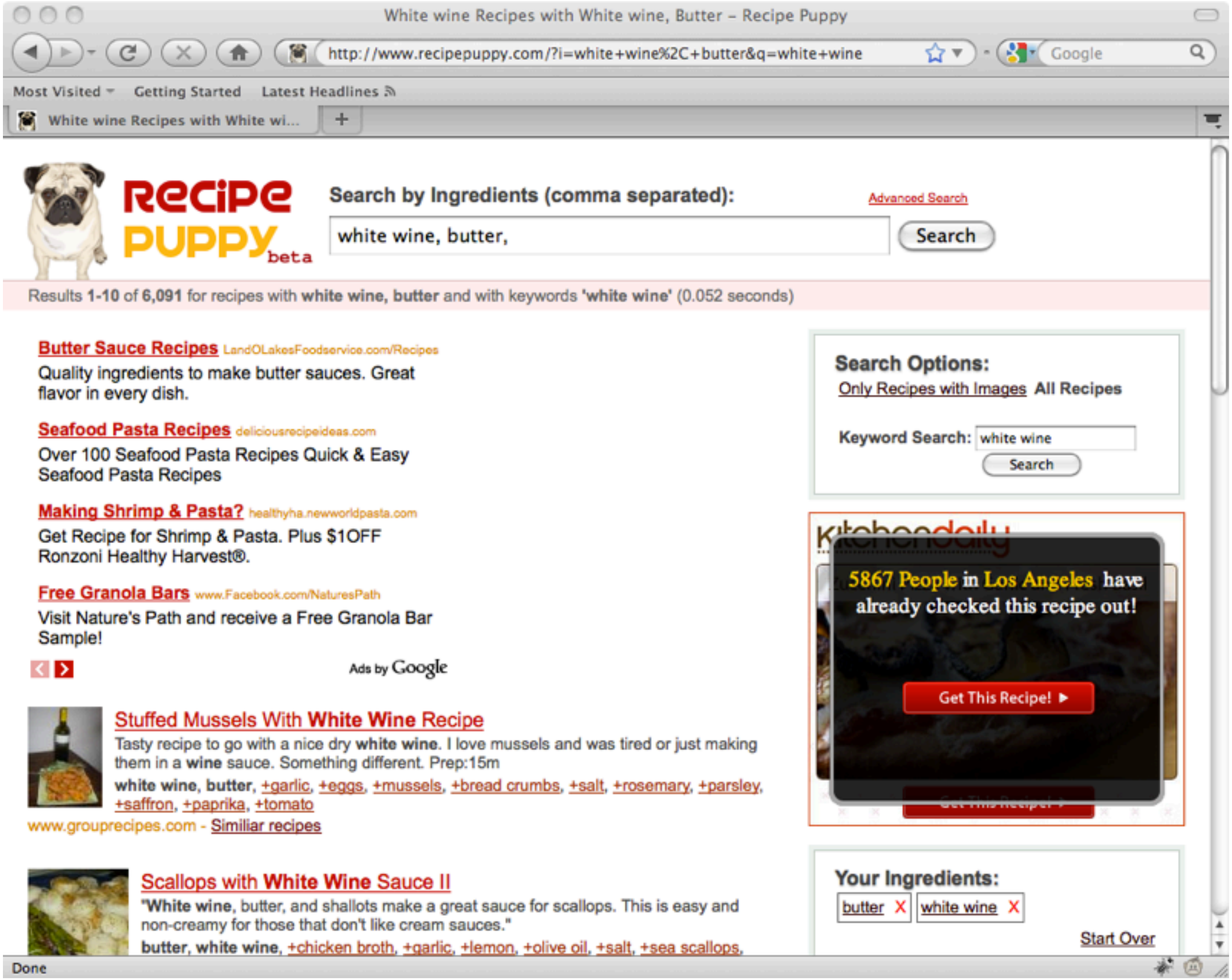
Projection indices were designed to respond to various features in a scatter (say, the presence of holes) -- The Grand Tour then becomes a kind of stochastic search for these “interesting” aspects of the data

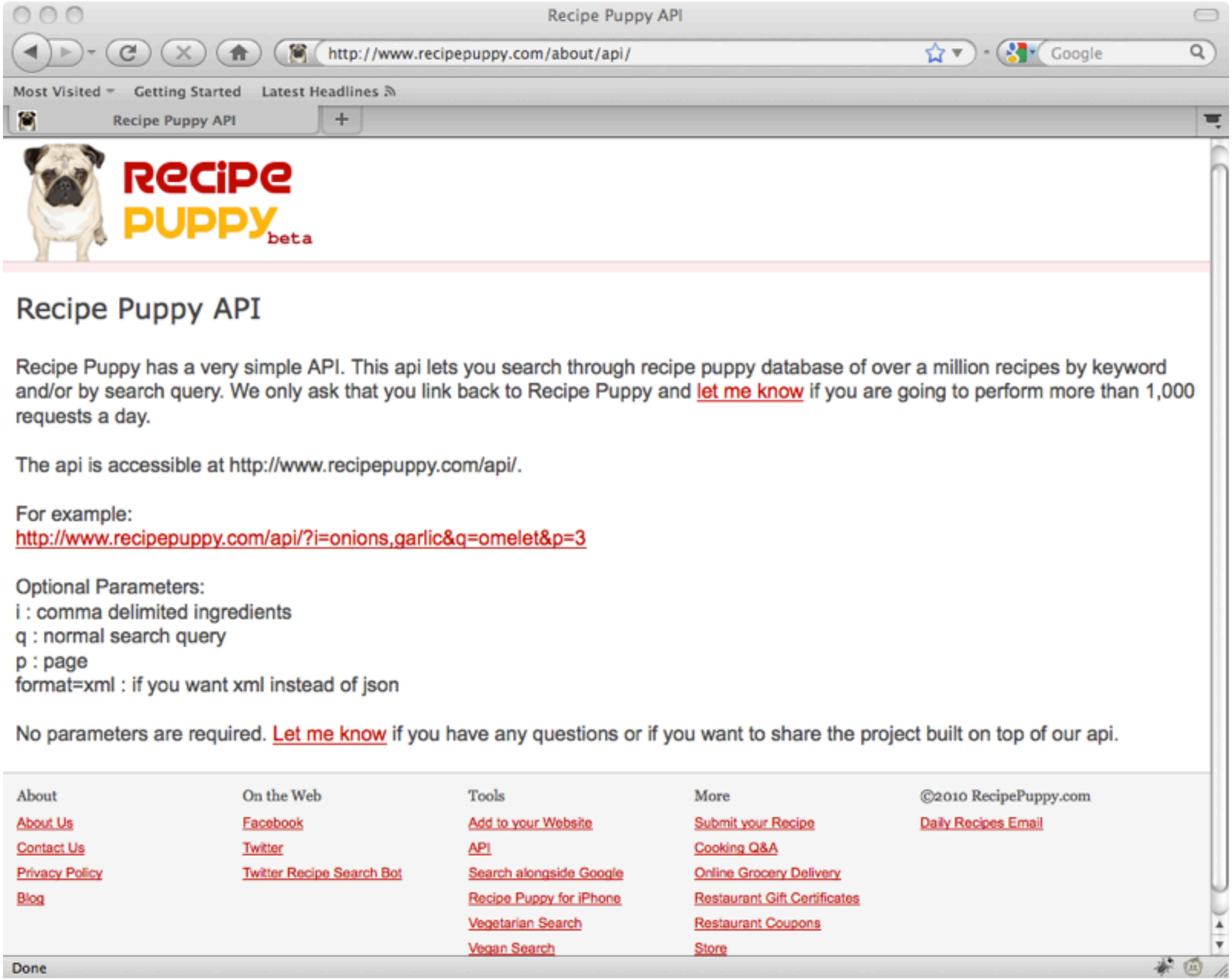
Documents as data

We are going to start with something simple and build up some of the statistical issues associated with handling and modeling text -- **We'll take as an initial case study a collection of recipes**

Recipes are interesting because they are not as unstructured as a tweet or an email message or a news article or a novel, but instead have a form -- You can usually point to **an ingredient list and then a list of instructions**, for example

Our recipes come from a site that distributes its data in an open way...


White wine Recipes with White wine, Butter – Recipe Puppy
http://www.recipepuppy.com/?i=white+wine%2C+butter&q=white+wine
Google
Most Visited
Getting Started
Latest Headlines
White wine Recipes with White wi...
Recipe Puppy beta
Search by Ingredients (comma separated):
Advanced Search
white wine, butter,
Search
Results 1-10 of 6,091 for recipes with white wine, butter and with keywords 'white wine' (0.052 seconds)
Butter Sauce Recipes LandOLakesFoodservice.com/Recipes
Quality ingredients to make butter sauces. Great flavor in every dish.
Seafood Pasta Recipes deliciousrecipeideas.com
Over 100 Seafood Pasta Recipes Quick & Easy Seafood Pasta Recipes
Making Shrimp & Pasta? healthyha.newworldpasta.com
Get Recipe for Shrimp & Pasta. Plus $1OFF Ronzoni Healthy Harvest®.
Free Granola Bars www.Facebook.com/NaturesPath
Visit Nature's Path and receive a Free Granola Bar Sample!
Ads by Google
Stuffed Mussels With White Wine Recipe
Tasty recipe to go with a nice dry white wine. I love mussels and was tired or just making them in a wine sauce. Something different. Prep:15m
white wine, butter, +garlic, +eggs, +mussels, +bread crumbs, +salt, +rosemary, +parsley, +saffron, +paprika, +tomato
www.grouprecipes.com - Similiar recipes
Scallops with White Wine Sauce II
"White wine, butter, and shallots make a great sauce for scallops. This is easy and non-creamy for those that don't like cream sauces."
butter, white wine, +chicken broth, +garlic, +lemon, +olive oil, +salt, +sea scallops,
Search Options:
Only Recipes with Images All Recipes
Keyword Search: white wine
Search
kitchendaily
5867 People in Los Angeles have already checked this recipe out!
Get This Recipe!
Your Ingredients:
butter X white wine X
Start Over
Done



Recipe Puppy API

http://www.recipepuppy.com/about/api/

Most Visited Getting Started Latest Headlines

Recipe Puppy API

Recipe Puppy beta

Recipe Puppy API

Recipe Puppy has a very simple API. This api lets you search through recipe puppy database of over a million recipes by keyword and/or by search query. We only ask that you link back to Recipe Puppy and let me know if you are going to perform more than 1,000 requests a day.

The api is accessible at http://www.recipepuppy.com/api/.

For example:
http://www.recipepuppy.com/api/?i=onions,garlic&q=omelet&p=3

Optional Parameters:
i : comma delimited ingredients
q : normal search query
p : page
format=xml : if you want xml instead of json

No parameters are required. Let me know if you have any questions or if you want to share the project built on top of our api.

About	On the Web	Tools	More	©2010 RecipePuppy.com
About Us	Facebook	Add to your Website	Submit your Recipe	Daily Recipes Email
Contact Us	Twitter	API	Cooking Q&A	
Privacy Policy	Twitter Recipe Search Bot	Search alongside Google	Online Grocery Delivery	
Blog		Recipe Puppy for iPhone	Restaurant Gift Certificates	
		Vegetarian Search	Restaurant Coupons	
		Vegan Search	Store	

Done

APIs

APIs (application programming interfaces) make data available as a kind of web service -- The Recipe Puppy API is just one of many you will find, but it was somehow the simplest and hence a reasonable choice to play with on pedagogical grounds

Aside from complications with authentication (often organizations want to know who is accessing their data) the broad principle is largely the same -- A specialized URL represents data, the results of a search, say, and not necessarily an HTML page

For example, enter these two into a Chrome browser -- What do you get?

```
http://www.recipepuppy.com/api/?q=cake
```

```
http://www.recipepuppy.com/api/?q=cake&format=xml
```

www.recipepuppy.com/api/?q=cake&p=100

{"title":"Recipe Puppy","version":0.1,"href":"http:\/\/www.recipepuppy.com\/","results":[{"title":"Best Lemon Blueberry Bundt Cake","href":"http:\/\/www.recipezaar.com\/Best-Lemon-Blueberry-Bundt-Cake-176927","ingredients":"baking powder, baking soda, blueberries, butter, buttermilk, eggs, flour, flour, lemon juice, lemon zest, salt, sugar, vanilla extract","thumbnail":"http:\/\/img.recipepuppy.com\/166414.jpg"},{"title":"Best Southern Pound Cake","href":"http:\/\/www.recipezaar.com\/Best-Southern-Pound-Cake-79972","ingredients":"baking powder, butter, shortening, eggs, milk, flour, salt, sugar, vanilla extract","thumbnail":""},{"title":"Best Wacky Cake","href":"http:\/\/www.recipezaar.com\/Best-Wacky-Cake-125923","ingredients":"flour, baking soda, butter, water, salt, semisweet chocolate, sugar, cocoa powder, vanilla extract, white vinegar","thumbnail":"http:\/\/img.recipepuppy.com\/166944.jpg"},{"title":"Better Than Grandma's Pound Cake","href":"http:\/\/www.recipezaar.com\/Better-Than-Grandmas-Pound-Cake-104343","ingredients":"crisco, eggs, margarine, milk, flour, flour, sugar, vanilla extract","thumbnail":""},{"title":"Better Than Sex Cake (With Bananas, Coconut, and Pineapple)","href":"http:\/\/www.recipezaar.com\/Better-Than-Sex-Cake-With-Bananas-Coconut-and-Pineapple-163122","ingredients":"banana, coconut, cool whip, pineapple, pudding, eggs, cream cheese, vegetable oil, sugar, water","thumbnail":""},{"title":"Better-For-You Pound Cake","href":"http:\/\/www.recipezaar.com\/Better-For-You-Pound-Cake-126395","ingredients":"buttermilk, egg whites, condensed milk, flour, almonds, applesauce, cake mix","thumbnail":""},{"title":"Betty Crocker Coffee Cake Circa 1973","href":"http:\/\/www.recipezaar.com\/Betty-Crocker-Coffee-Cake-Circa-1973-224163","ingredients":"baking powder, eggs, flour, milk, salt, sugar, vegetable oil","thumbnail":""},{"title":"Better Than Sex Cake (Toffee Bar Cake)","href":"http:\/\/www.recipezaar.com\/Better-Than-Sex-Cake-Toffee-Bar-Cake-38550","ingredients":"chocolate cake, chocolate syrup, cool whip, candy bars","thumbnail":""},{"title":"Bijan's Pina Colada Birthday Cake","href":"http:\/\/www.recipezaar.com\/Bijans-Pina-Colada-Birthday-Cake-125679","ingredients":"cool whip, cream cheese, pineapple, cream of coconut, vanilla pudding, milk, cake mix","thumbnail":"http:\/\/img.recipepuppy.com\/169120.jpg"},{"title":"Birdseed Cake","href":"http:\/\/www.recipezaar.com\/Birdseed-Cake-57924","ingredients":"flour, baking soda, banana, cinnamon, pineapple, eggs, vegetable oil, pecan, salt, sugar, vanilla extract","thumbnail":""}]}

www.recipepuppy.com/api/?q=cake&p=100&format=xml


```xml
<recipes>
  <recipe>
    <title>Best Lemon Blueberry Bundt Cake</title>
    <href>
      http://www.recipezaar.com/Best-Lemon-Blueberry-Bundt-Cake-176927
    </href>
    <ingredients>
      baking powder, baking soda, blueberries, butter, buttermilk, eggs, flour, flour, lemon juice, lemon zest, salt,
      sugar, vanilla extract
    </ingredients>
  </recipe>
  <recipe>
    <title>Best Southern Pound Cake</title>
    <href>
      http://www.recipezaar.com/Best-Southern-Pound-Cake-79972
    </href>
    <ingredients>
      baking powder, butter, shortening, eggs, milk, flour, salt, sugar, vanilla extract
    </ingredients>
  </recipe>
  <recipe>
    <title>Best Wacky Cake</title>
    <href>http://www.recipezaar.com/Best-Wacky-Cake-125923</href>
    <ingredients>
      flour, baking soda, butter, water, salt, semisweet chocolate, sugar, cocoa powder, vanilla extract, white
      vinegar
    </ingredients>
  </recipe>
  <recipe>
    <title>Better Than Grandma's Pound Cake</title>
    <href>
      http://www.recipezaar.com/Better-Than-Grandmas-Pound-Cake-104343
    </href>
    <ingredients>
      crisco, eggs, margarine, milk, flour, flour, sugar, vanilla extract
    </ingredients>
  </recipe>
  <recipe>
    <title>
      Better Than Sex Cake (With Bananas, Coconut, and Pineapple)
    </title>
    <href>
      http://www.recipezaar.com/Better-Than-Sex-Cake-With-Bananas-Coconut-and-Pineapple-163122
    </href>
    <ingredients>
```

Data formats

With the advent of Web 2.0, we have seen the rise of data services or web services like our recipe server -- You can think of these as URL's that return, well, data rather than HTML documents

While this sounds small, let's consider one of the cake recipes from Recipe Puppy -- Here is what the recipe looks like on the screen and then how we'd have to "read" it if we were processing the text

Gluten-Free Coconut Layer Cake

Epicurious | March 2008
by Zoe Singer

save to recipe box

Join now to set up a FREE recipe box to save all your favorite recipes!

recipe | reviews (40) | photo | video | my notes | find out more

enlarge image

yield: Makes one three-layer (9-inch) cake; 12 servings

This all-purpose sponge cake has a wonderfully rich flavor and golden color, thanks to the almond flour. For a variation, instead of white-chocolate...

more ›

user rating

92% would make it again

user rating: 4 forks

rate this recipe

review this recipe

at a glance

main ingredients
Coconut

type
Kid-Friendly, Cake

dietary considerations
Wheat/Gluten-Free

cooks' tools

conversion chart

technique videos

ingredients

subscribe to Bon Appétit

For fluffy white-chocolate whipped cream frosting

3 cups heavy cream, chilled
9 ounces white chocolate, coarsely chopped
2 teaspoons coconut or golden rum
1/8 teaspoon fine sea salt

For coconut layer cake

1 3/4 cup almond flour
2 tablespoons coconut flour
10 large eggs, at room temperature, separated
1 tablespoon coconut or golden rum
2 cups confectioner's sugar, sifted
1/4 teaspoon cream of tartar
1/4 teaspoon fine sea salt

To assemble

2 cups (3 ounces) unsweetened coconut flakes, for coating

view-source:www.epicurious ×

view-source:www.epicurious.com/recipes/food/views/Gluten-Free-Coconut-Layer-Cake-241926

```html

<div id="preparation" class="instructions">
    <h2>Preparation</h2>



            <p class="instruction">
                <strong>Make frosting</strong><br />
                Chill bowl of stand mixer and whisk attachment or large metal bowl and beaters for at least 15 minutes.
            </p>


            <p class="instruction">

                In small saucepan over moderate heat, bring 1 cup cream to simmer. Transfer white chocolate to medium heatproof bowl, pour hot cream over, and whisk until smooth. Whisk in rum and salt. Let cool at room temperature until thickened slightly, about 1 hour.
            </p>


            <p class="instruction">

                In chilled bowl of electric mixer fitted with whisk attachment, beat remaining 2 cups cream at moderately high speed until whisk leaves marks but cream does not quite hold soft peaks, 6 to 8 minutes. Turn mixer off, then add white chocolate mixture and beat just until stiff peaks begin to form, about 5 minutes. (Do not overbeat, or cream will curdle.) Refrigerate until firm, about 3 hours. (Frosting can be made ahead and refrigerated, covered, up to 8 hours.)
            </p>




            <p class="instruction">
                <strong>While frosting is chilling, make cake</strong><br />
                Preheat oven to 350F. Line bottoms of cake pans with parchment paper.
            </p>


            <p class="instruction">

                In large bowl, whisk together almond and coconut flours.
            </p>


            <p class="instruction">

                In bowl of electric mixer fitted with whisk attachment, beat egg yolks at high speed until pale yellow and fluffy, 2 to 3 minutes. Reduce speed to moderately low and beat in rum and all but 1 tablespoon confectioner's sugar. Scrape down bowl, then increase speed to high and beat until pale and thick, about 1 minute. Reduce speed to low and gradually add almond and coconut flour mixture, scraping down bowl and folding in last of flour by hand. Set aside.
            </p>

```

Data formats

HTML (the HyperText Markup Language) is the (primary) language used to represent documents on the web -- It describes the components of a document (headings, paragraphs, lists, tables, images) via special text “tags” known as markup

In the snippet of “code” on the previous page, you see references to paragraphs (denoted `<p>`) and headings (here `<h2>` for a “second level” heading) -- These notations tell your browser how to render the text

In addition you will notice `<div>` and `<span>` tags that contain attributes that also affect the way text is displayed -- Through so-called cascading style sheets (CSS), a web designer can have finer control over the visual properties of their rendered document

Data formats

The markup underlying HTML is designed to both describe the structure of the document and to control its visual layout -- The naming references in the recipe HTML file are a funny mix of formatting and (I’m guessing) database logic as these recipes are all served from a database

If we look at a recipe from a different site, you’ll notice a very different naming convention for the structures used in the document...

Survivor Birthday Party Poke Cake

(29) | rate it now!

photo by: kraft

add your photo 1 | 1

add to shopping list | add to recipe box | print | send/share

+1 Pin it Like 58

Fun, easy-to-make, & delicious! The kids and adults all loved the look and taste. Used this for a retirement party as an "I SURVIVED" cake. Also, used strawberry JELL-O ins...read more

posted by jessicasavory | on 2/6/2010

time

prep: 1 hr

total: 5 hr 40 min

servings

total: 16 servings

recipe | reviews (29) | nutrition

what you need

2	baked 9-inch round white cake layers, cooled
2	cups boiling water
2	pkg. (3 oz. each) JELL-O Cherry Flavor Gelatin
1	tub (8 oz.) COOL WHIP Whipped Topping, thawed
4	OREO Cookies, finely crushed (about 1/3 cup)
1-1/2	cups assorted bug- or worm-shaped chewy fruit snacks

make it

PLACE cake layers, top-sides up, in 2 clean 9-inch round pans. Pierce cakes with large fork at 1/2-inch intervals.

ADD boiling water to gelatin mixes in small bowl; stir 2 min. until completely dissolved. Pour over cake layers in pans. Refrigerate 3 hours.

you may also enjoy

- Low-Fat Strawberry Shortcut (26)
- BAKER'S ONE BOWL Chocolate Frosting (16)
- Frozen Yogurt Pie (31)
- Luscious Lemon Poke Cake (152)
- Patriotic Poke Cake (69)
- JELL-O Poke Cake (179)
- Gelatin Poke Cake (167)
- Easy Pudding Poke Cake (7)
- Pudding Poke Cake (24)
- Lemon Pudding Poke Cake (50)

view-source:www.kraftrecip

view-source:www.kraftrecipes.com/recipes/survivor-birthday-party-poke-64178.aspx

```
                                        <div class="table-row">
                                                <div class="column1">
                                                        <div class="textarea">
                                                <span rel="v:ingredient">
                                                         <span typeof="v:ingredient">
                                                                <div class="amount"><span property="v:amount">1-1/2</span></div>
                                                                <div class="desc">cups assorted bug- or <span property="v:name">worm-shaped chewy fruit snacks</span></div>
                                                        </span>
                                                </span>
                                                        </div>
                                                </div>
                                        </div>



        </div>
        <!--concordance-end-->








<div id="recipeGradHeading" class="recipeGradHeading">
        <div class="head">
                <h2>Make It</h2>
        </div><!--[if !IE]>  END head <![endif]-->
</div><!--[if !IE]>  END recipeGradHeading <![endif]-->
<div class="recipeMakeItText">
        <div class="stdContBlock">
                <div class="textarea">
                 <span rel="v:instructions">
            <span typeof="v:Instructions">
                <p><STRONG></STRONG></p>


                                        <p><span property="v:instruction">
                                                <STRONG>PLACE </STRONG>cake layers, top-sides up, in 2 clean 9-inch round pans. Pierce cakes with large fork at 1/2-inch intervals.
                                        </span>
                                        </p>


                                        <p><span property="v:instruction">
                                                <STRONG>ADD </STRONG>boiling water to gelatin mixes in small bowl; stir 2 min. until completely dissolved.  Pour over cake layers i
                                        </span>
```

Data formats

What you see in HTML is a tension between structure and layout with a wiff of content clues, if somewhat irregular and site-specific -- With the advent of Web 2.0, we have seen the rise of data services or web services like our recipe server

You can think of these as URL's that return, well, **data rather than HTML documents** -- That is, instead of hiding data behind markup designed for describing the structure and display of a document, introduce **a format designed for data**

Today, the format tends to be either XML (the eXtensible Markup Language) or JSON (JavaScript Object Notation) -- Both provide so-called humanly-readable files that are "self-describing" in some sense

Let's look at the Recipe Puppy output again...

```
<recipes>
   <recipe>
      <title>Best Lemon Blueberry Bundt Cake</title>
      <href>
         http://www.recipezaar.com/Best-Lemon-Blueberry-Bundt-Cake-176927
      </href>
      <ingredients>
         baking powder, baking soda, blueberries, butter, buttermilk, eggs, flour, flour, lemo
      </ingredients>
   </recipe>
   <recipe>
      <title>Best Southern Pound Cake</title>
      <href>
         http://www.recipezaar.com/Best-Southern-Pound-Cake-79972
      </href>
      <ingredients>
         baking powder, butter, shortening, eggs, milk, flour, salt, sugar, vanilla extract
      </ingredients>
   </recipe>

   ...

   <recipe>
      <title>Best Wacky Cake</title>
      <href>http://www.recipezaar.com/Best-Wacky-Cake-125923</href>
      <ingredients>
         flour, baking soda, butter, water, salt, semisweet chocolate, sugar, cocoa powder, va
      </ingredients>
   </recipe>
</recipes>
```

```
{"title":"Recipe Puppy",
 "version":0.1,
 "href":"http://www.recipepuppy.com/",
 "results":[
            {"title":"Best Lemon Blueberry Bundt Cake",
               "href":"http://www.recipezaar.com\/Best-Lemon-Blueberry-Bundt-Cake-176927",
               "ingredients":"baking powder, baking soda, blueberries, butter, buttermilk, blu

               "thumbnail":"http://img.recipepuppy.com/166414.jpg"
            },
            {"title":"Best Southern Pound Cake",
               "href":"http:\/\/www.recipezaar.com\/Best-Southern-Pound-Cake-79972",
               "ingredients":"baking powder, butter, shortening, eggs, milk, flour, salt, suga
               "thumbnail":""
            },
            {"title":"Best Wacky Cake","href":"http://www.recipezaar.com/Best-Wacky-Cake-12592
               "ingredients":"flour, baking soda, butter, water, salt, semisweet chocolate, su
               "thumbnail":"http://img.recipepuppy.com/166944.jpg"
            },
            {"title":"Better Than Grandma's Pound Cake",
               "href":"http://www.recipezaar.com\/Better-Than-Grandmas-Pound-Cake-104343",
               "ingredients":"crisco, eggs, margarine, milk, flour, flour, sugar, vanilla extr
               "thumbnail":""
            },
     ...
            {"title":"Better Than Sex Cake (With Bananas, Coconut, and Pineapple)",
               "href":"http://www.recipezaar.com\/Better-Than-Sex-Cake-With-Bananas-Coconut-ar
               "ingredients":"banana, coconut, cool whip, pineapple, pudding, eggs, cream chee
               "thumbnail":""},
           ]
}
```

Ingredients

We’ll start by considering just the ingredient lists from 1,000 recipes offered by Recipe Puppy -- In all, there are 381 ingredients, of which these are the most frequent

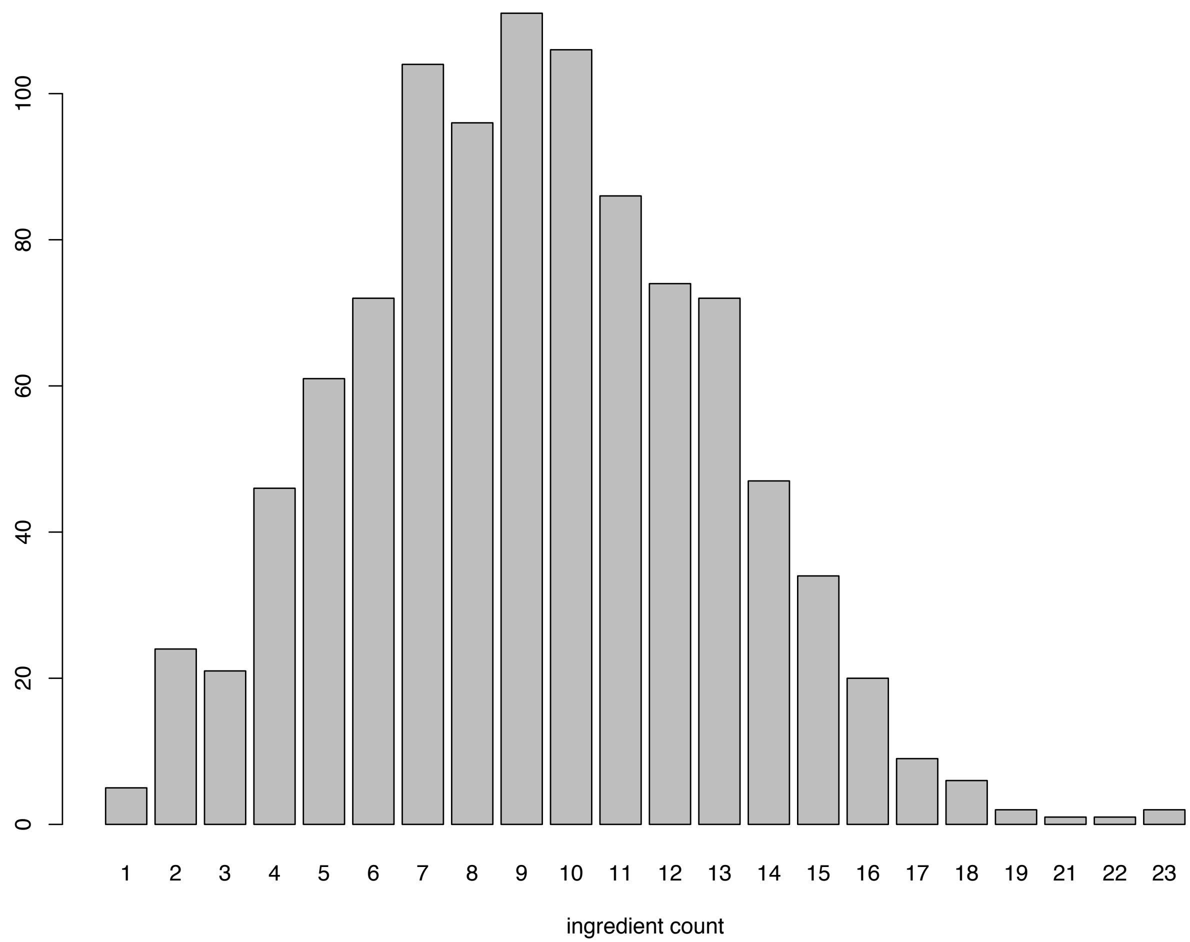
```
  eggs           sugar           flour   vanilla.extract       salt
   768             653             653               508        494

butter     baking.soda   vegetable.oil     baking.powder      water
   476             359             342               33         286
```

On the other end of the spectrum, there are 126 ingredients that appear only in one recipe -- Things like Nutella, various liqueurs, garlic and oyster sauce

Here’s a breakdown of the number of ingredients per recipe (There are some exceedingly simple and some crazy difficult recipes out there!)

Ingredients

We'll start by considering just the ingredient lists from 1,000 recipes offered by Recipe Puppy -- Here's a breakdown of the number of ingredients per recipe (There are some exceedingly simple and some crazy difficult recipes out there!)



Ingredients

Triple-Chocolate Celebration Cake (23 ingredients): baking powder, baking soda, semisweet chocolate, semisweet chocolate, semisweet chocolate, flour, cherries, cocoa powder, egg yolks, eggs, heavy cream, chocolate, corn syrup, semisweet chocolate chips, strawberries, blackberries, blueberries, raspberries, raspberry jam, salt, sour cream, sugar, cake, vanilla extract, vegetable oil, heavy cream

Spiced Pumpkin Cake with Caramel Icing (22 ingredients): allspice, baking powder, baking soda, flour, vegetable oil, cinnamon, cloves, cream cheese, rum, cranberries, eggs, ginger, heavy cream, orange zest, pumpkin puree, raisins, salt, sugar, sugar, orange zest, vanilla extract, vanilla ice cream, walnut, water"

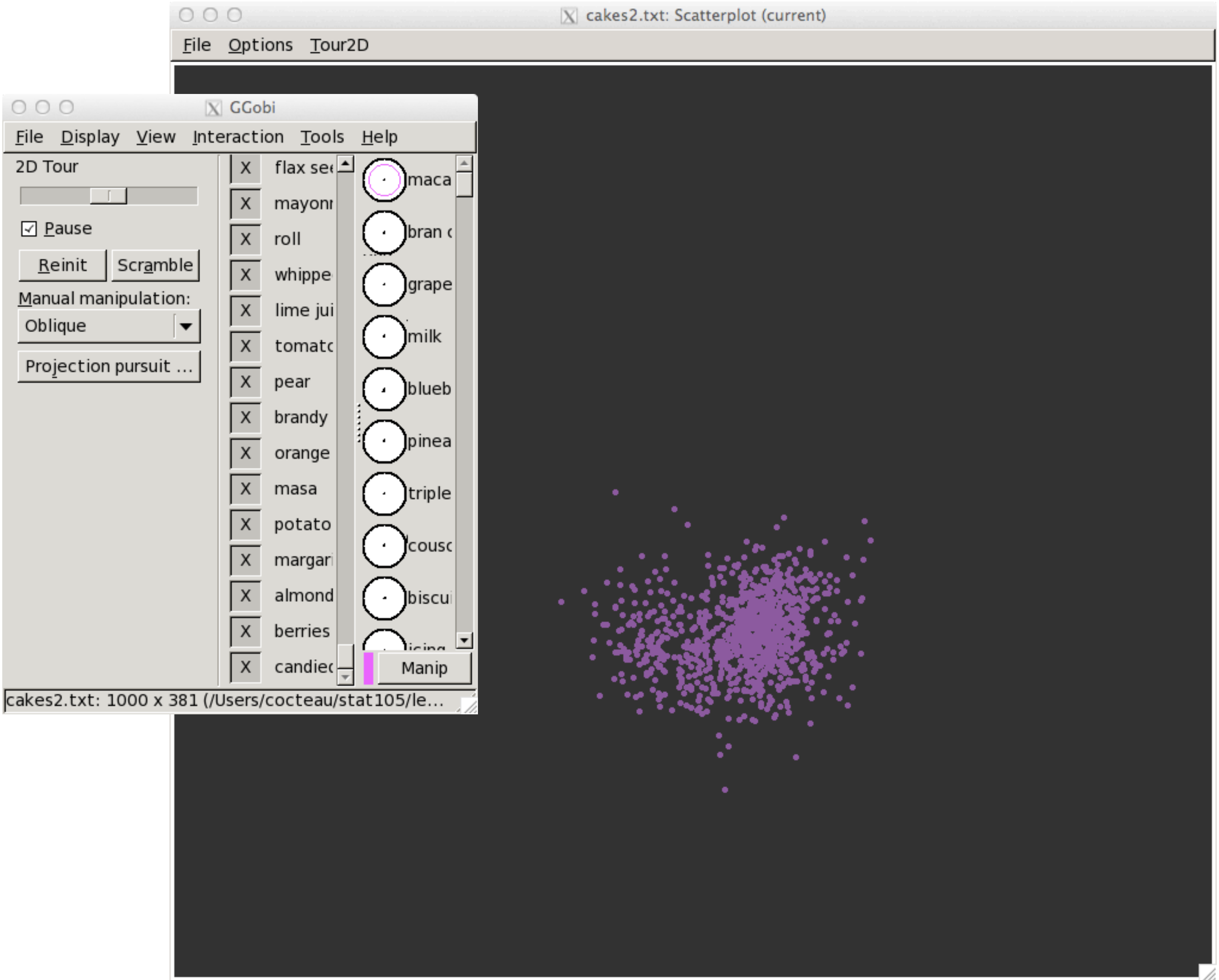
Funny Cake: pie shell

Examining the data

Our data set can be thought of as a 1000 x 381 binary matrix where each row is a recipe and each column represents and ingredient -- This is a “high dimensional” data set in the sense that the number of variables (the vector of unique ingredients) is large

This situation is common when dealing with text data -- The first step is often to reduce the text to a “bag of words” with indicators (or counts or weights, as we’ll see) for each

We can apply the projective methods from earlier in the quarter to have a look at these data -- Let’s fire up GGobi again!


cakes2.txt: Scatterplot (current)
File
Options
Tour2D
GGobi
File
Display
View
Interaction
Tools
Help
2D Tour
Pause
Reinit
Scramble
Manual manipulation:
Oblique
Projection pursuit ...
X
flax see
X
mayonn
X
roll
X
whippe
X
lime jui
X
tomato
X
pear
X
brandy
X
orange
X
masa
X
potato
X
margari
X
almond
X
berries
X
candied
maca
bran c
grape
milk
blueb
pinea
triple
cousc
biscui
Manip
cakes2.txt: 1000 x 381 (/Users/cocteau/stat105/le...

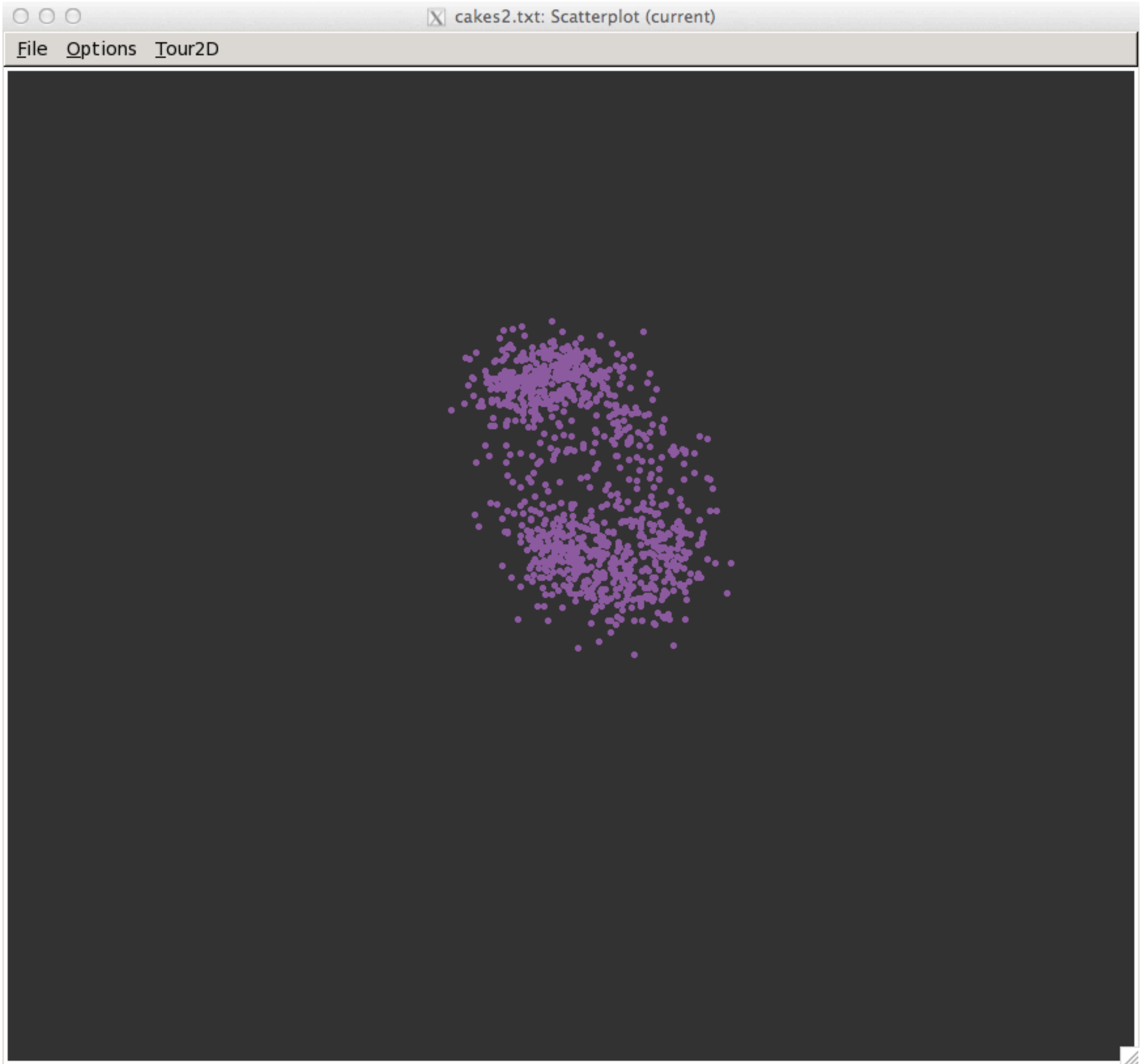
Projections

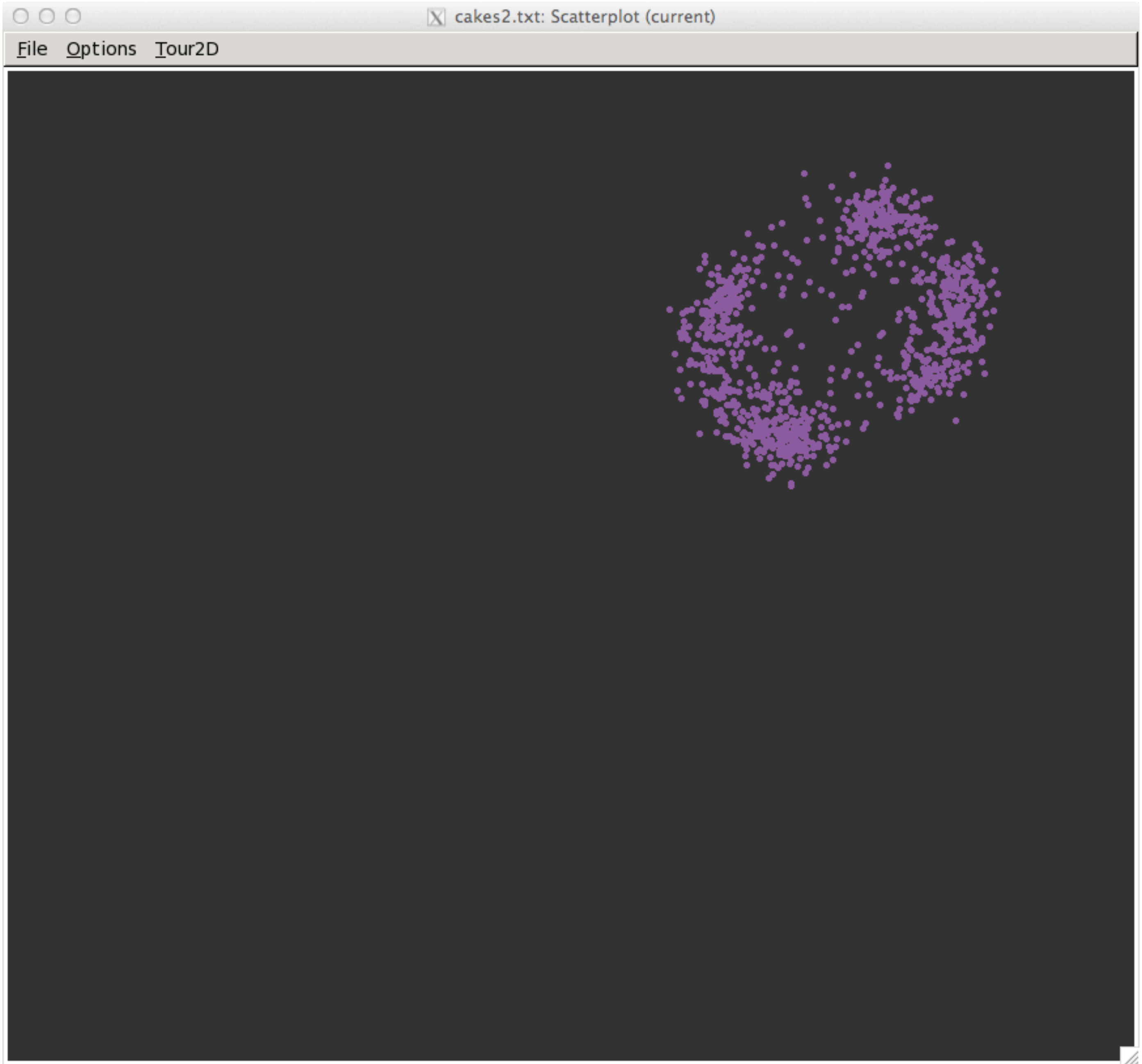
GGobi implements the “Grand Tour,” as smoothly interpolated set of projections of the data -- The Grand Tour selects the directions essentially at random and then moves from one to the next

The Grand Tour is a kind of fishing expedition -- We scan these plots for structure, for **something other than a formless, bivariate “splatter”** (or maybe in technical parlance, something that doesn’t look like observations from a bivariate normal distribution)

In the late 1970s and early 1980s, there was interest in driving the tour to interesting directions, those that don’t appear normal, say -- There were a variety of projection metrics, that scored the projected, bivariate data cloud

For example, one of these responds to “holes” in the data set...

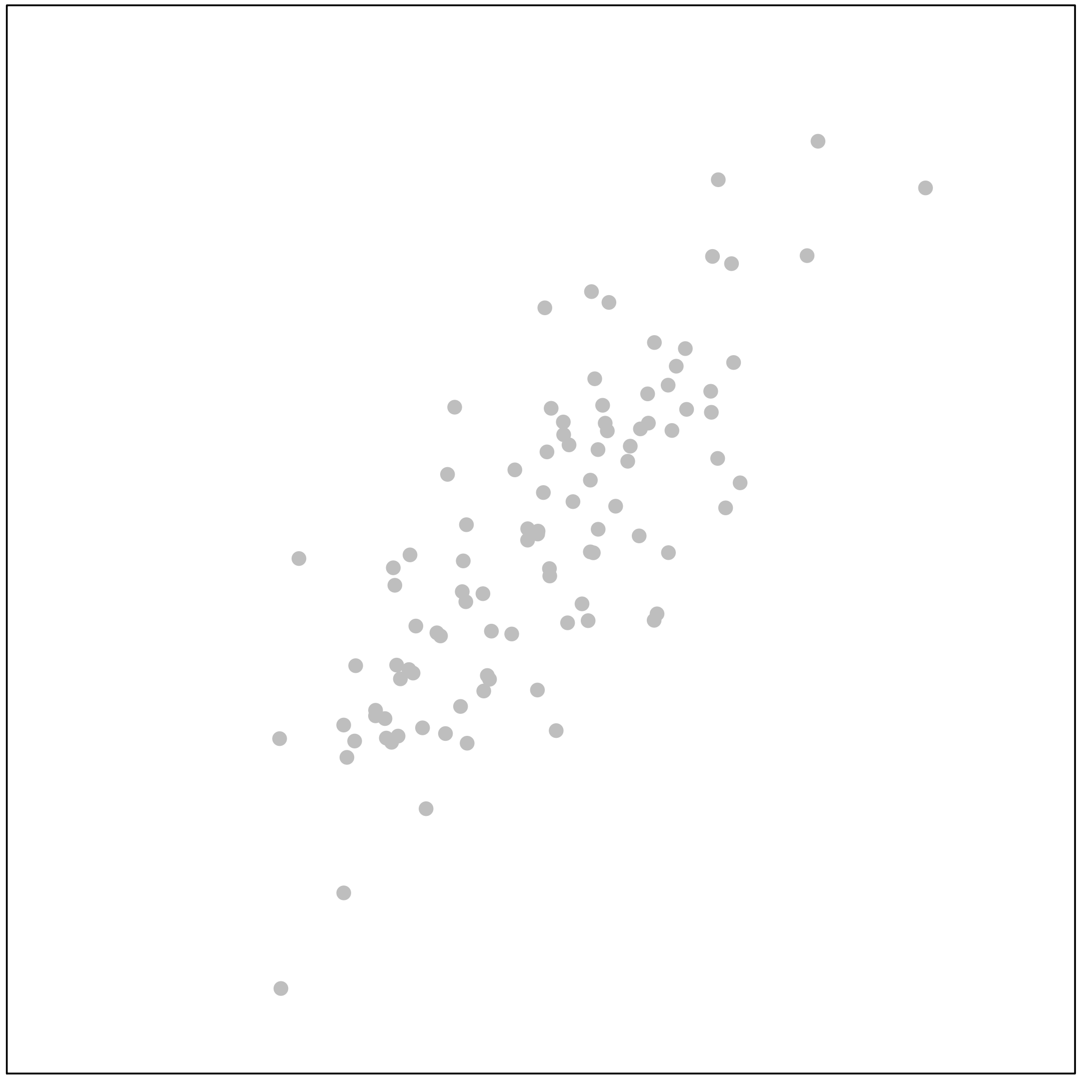

cakes2.txt: Scatterplot (current)
File
Options
Tour2D


cakes2.txt: Scatterplot (current)
File
Options
Tour2D

A distinguished (or at least storied) projection

One particular projection measures “interestingness” in terms of its ability to **represent as much of the variability in the data as possible in a two dimensional display** -- Because of this, the projection is said to be useful for “dimension reduction”

In principal components analysis (PCA) we derive **a new coordinate system** for the data, one that is “aligned” to the shape of the data cloud -- On the next few slides we illustrate the idea with a bivariate data set (we’ll return to our 381-dimensional data in a moment)

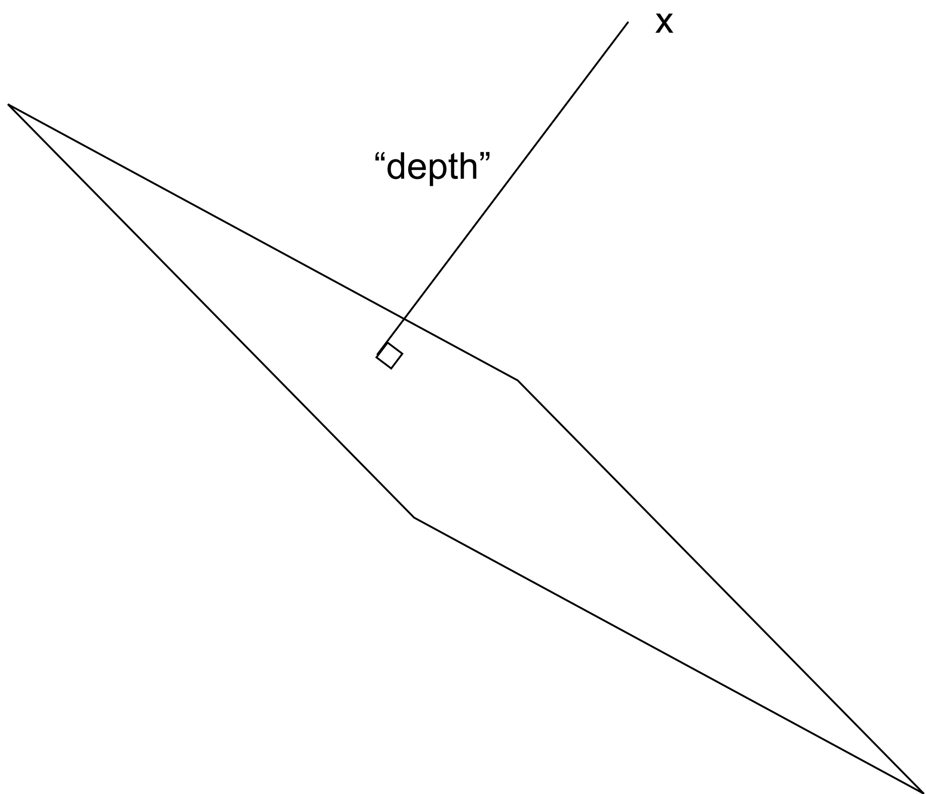


PCA

Now, consider creating **a new coordinate system for the data** (or, rather, come up with a new pair of orthogonal axes) that is “aligned” to the data’s shape -- You can formalize this in two (equivalent ways)

First, given a projection onto a plane, you can think about the distance from that point to the plane as being the “depth” of the point -- We then find the projection that gives us the smallest overall (over the whole data set) depth

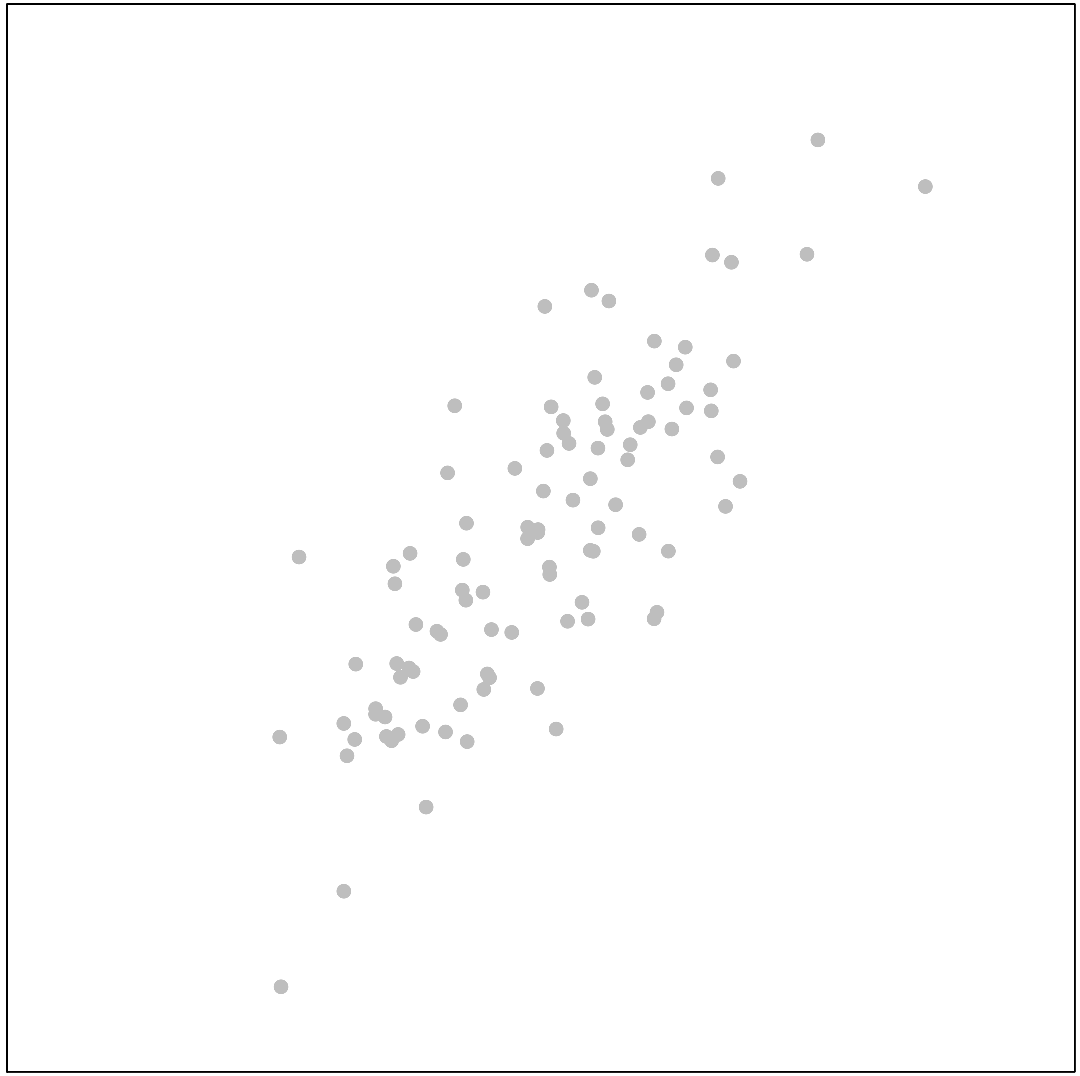
The depth of x when projected on the plane is just its orthogonal distance

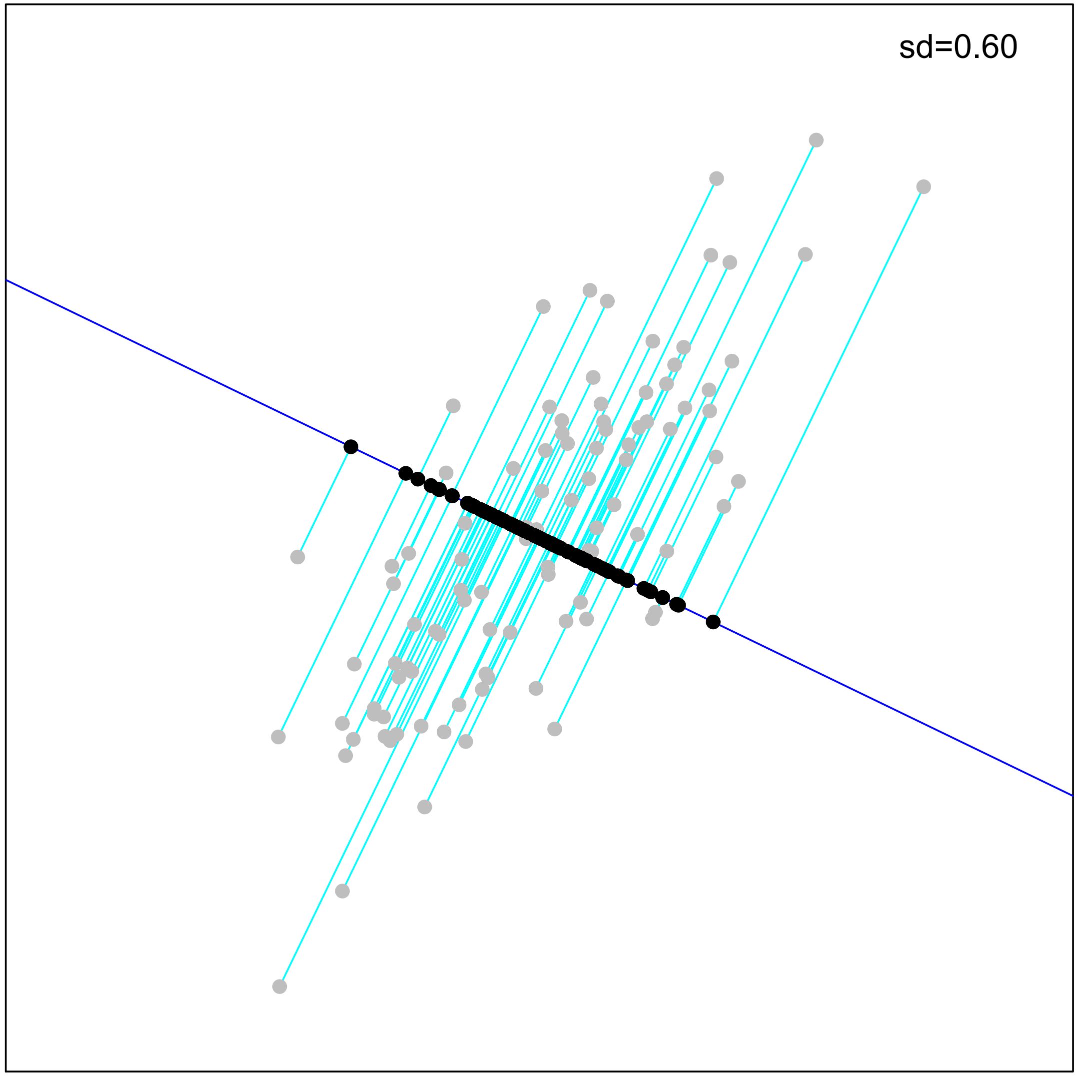


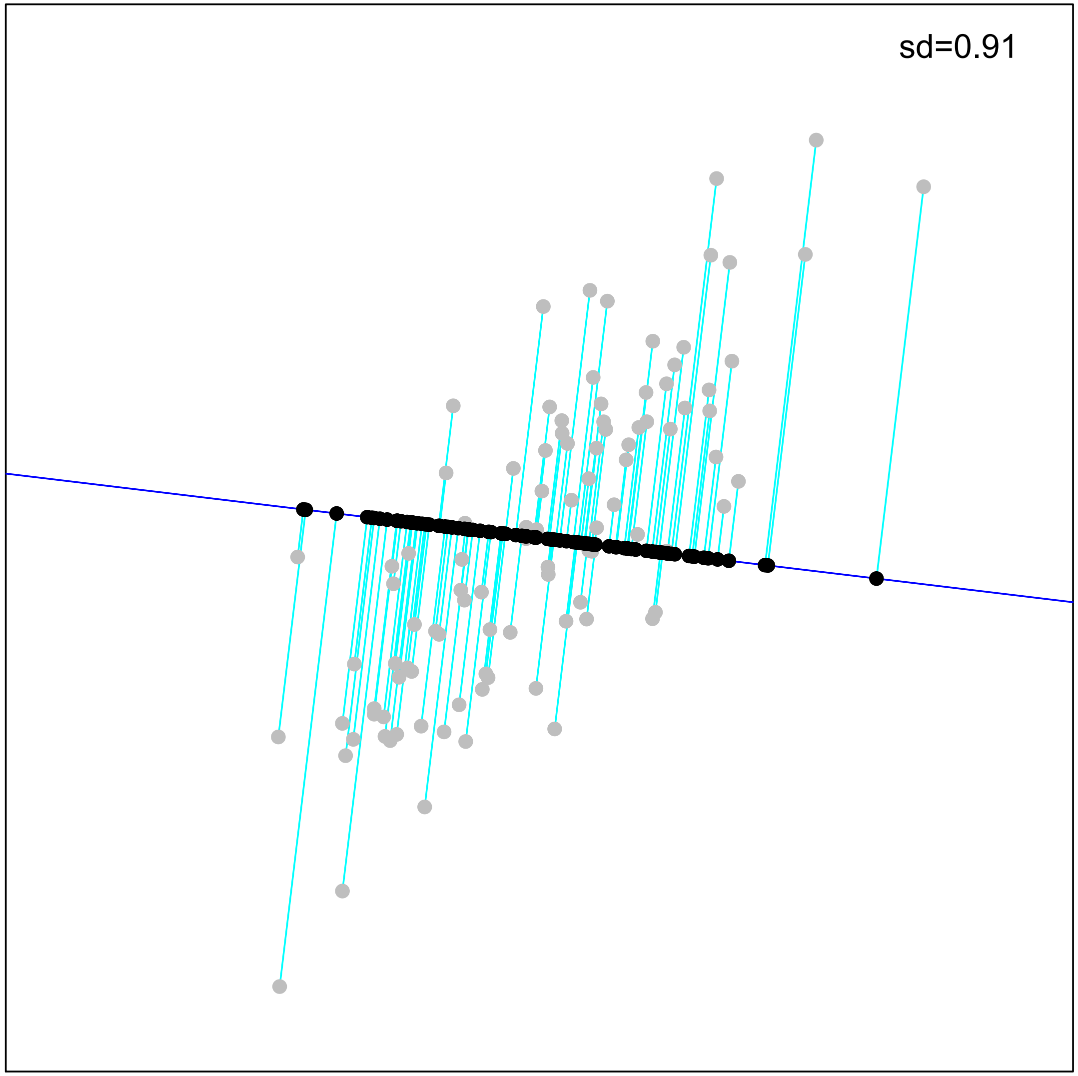
PCA

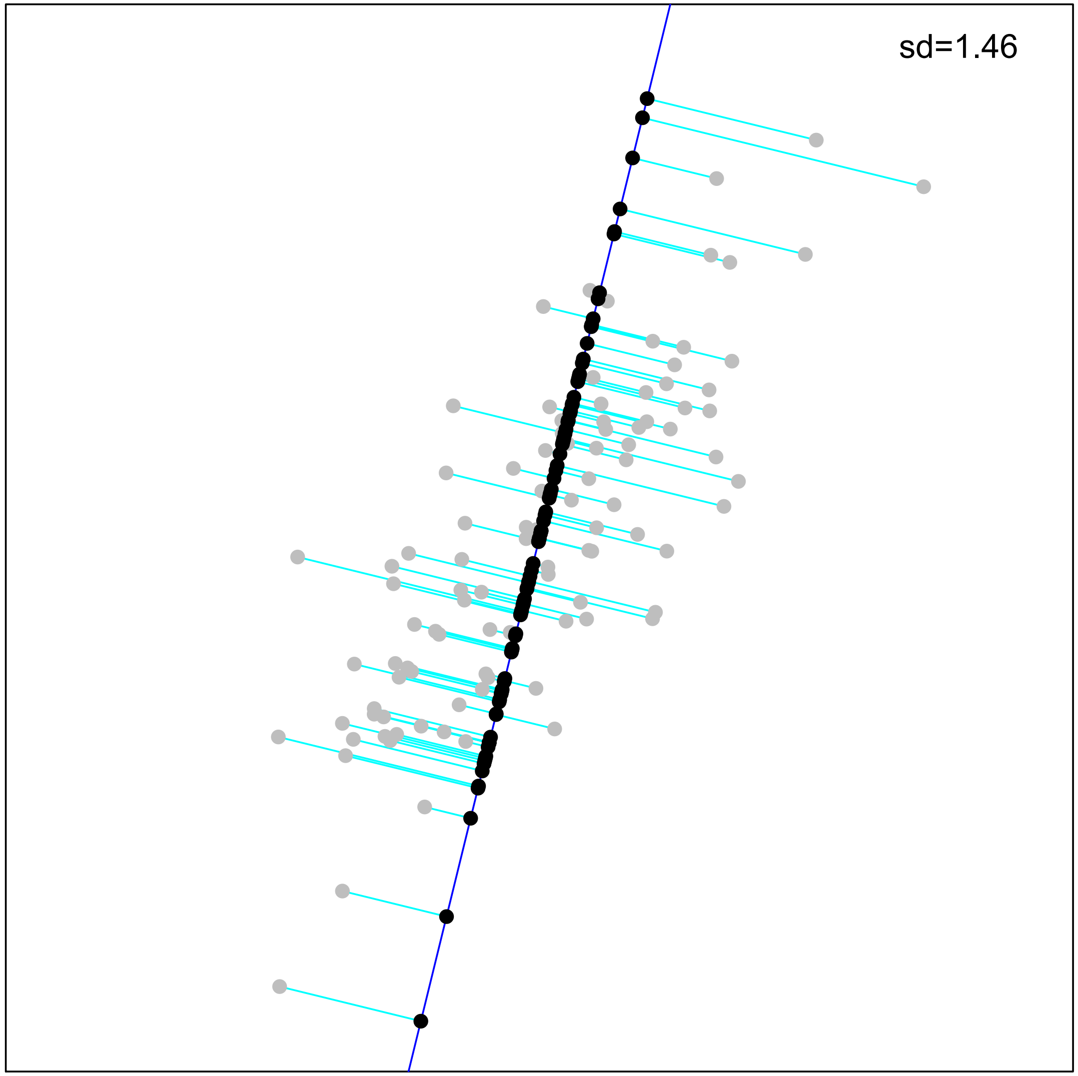
You can also formalize the notion of “aligned” by taking a sequential approach to PCA -- We define a new coordinate system one variable at at time, taking the first coordinate direction to be the one that gives us **the biggest spread** when we project our data along it

On the next three slides we just choose directions at random, project the data and examine its (univariate) spread -- The blue lines indicate the direction, the black dots are the projection of the data, and in the upper righthand corner we present the sample standard deviation of the projected data




sd=0.60


sd=0.91

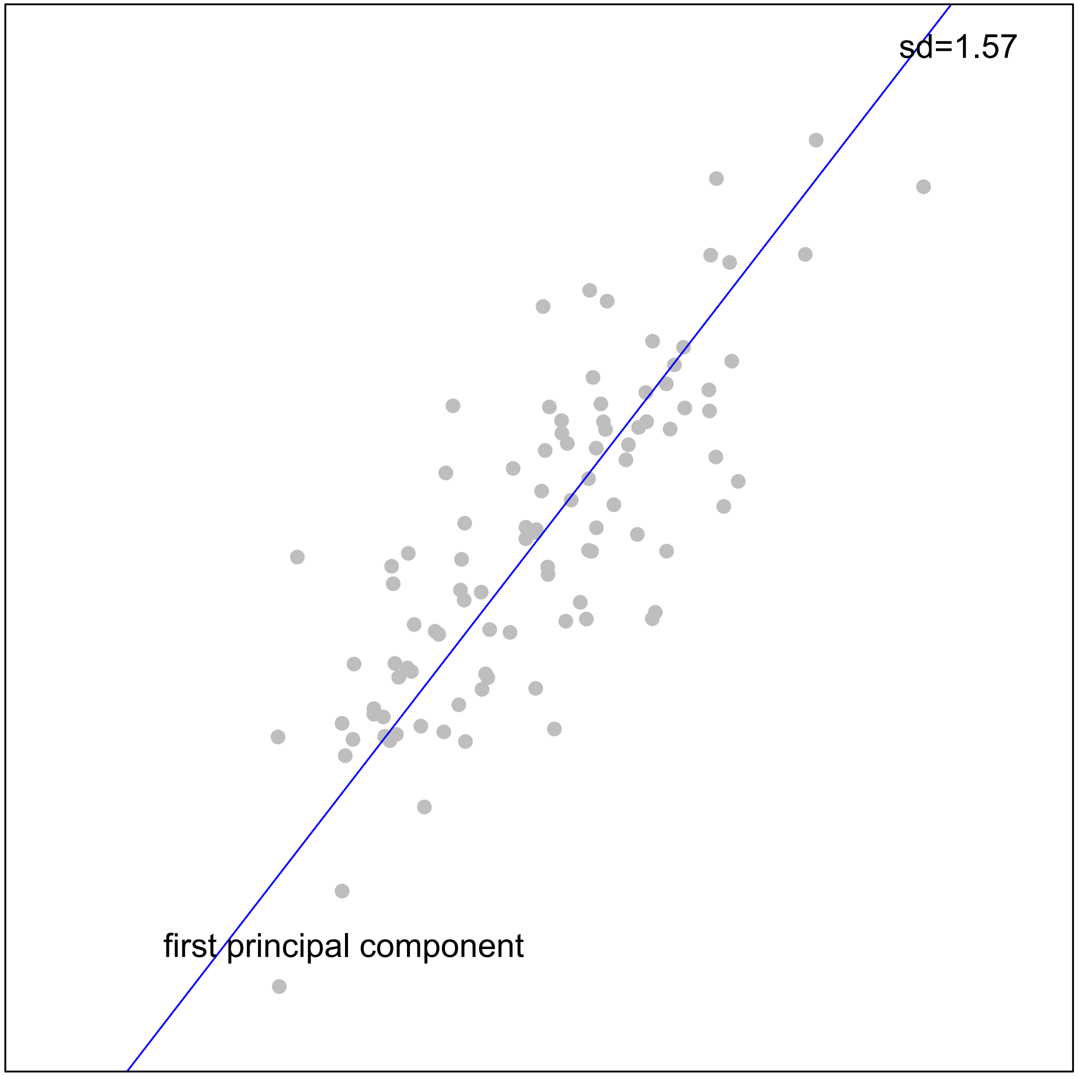

sd=1.46

PCA

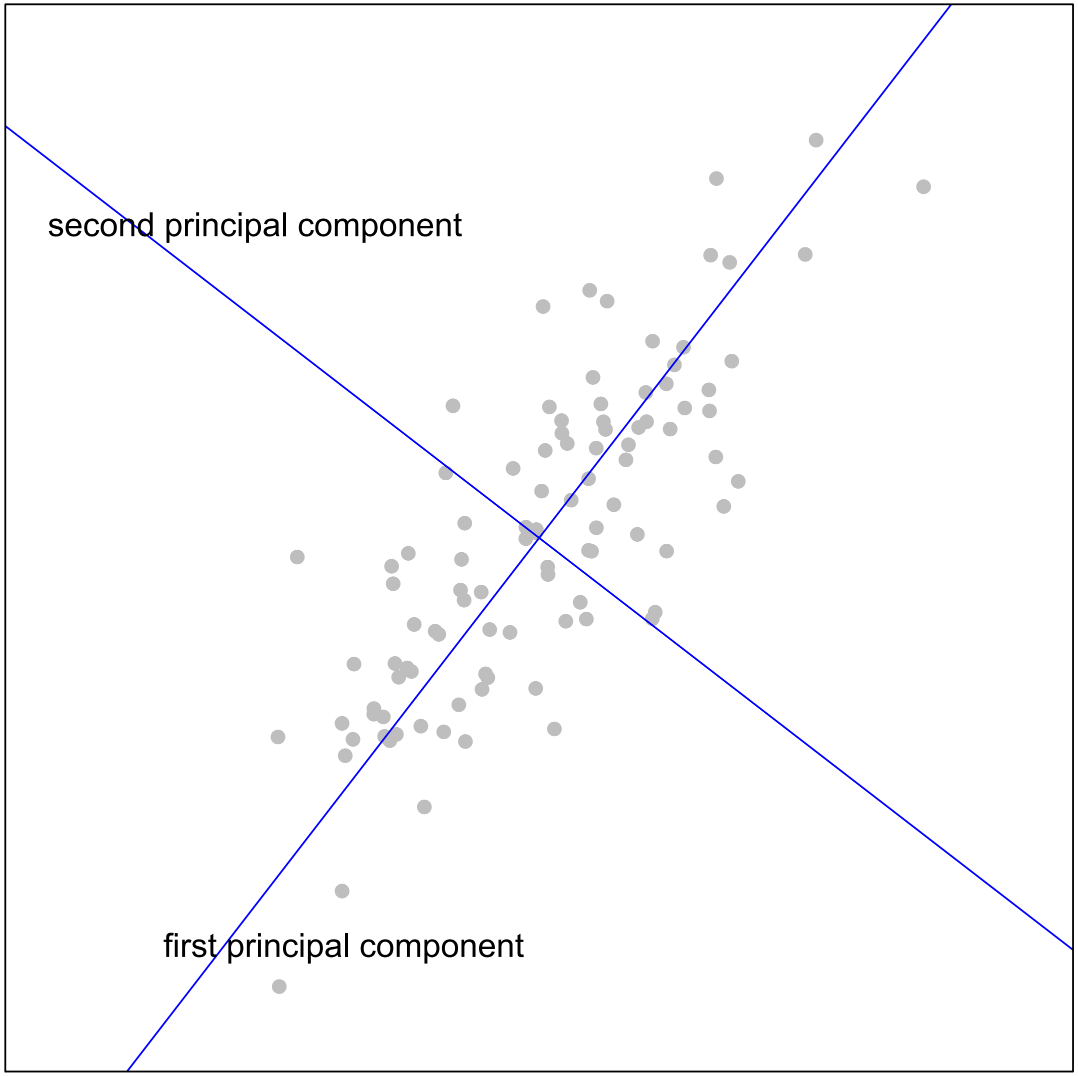
Clearly, when we select a direction that is “aligned” with the data, the projected values are more spread out -- The first principal component is just **the direction which yields the largest spread**

The second principal component is the direction, **orthogonal to the first, giving the next largest spread** -- For our little two dimensional data set, once we picked one direction, the second is set by orthogonality

For data that live in higher dimensions (like our 381-dimensional space of ingredients), we can continue adding directions, each time making sure **the new addition is orthogonal to the previous ones** and that, subject to this constraint, **the spread of the projected data is as large as possible**



sd=1.57
first principal component



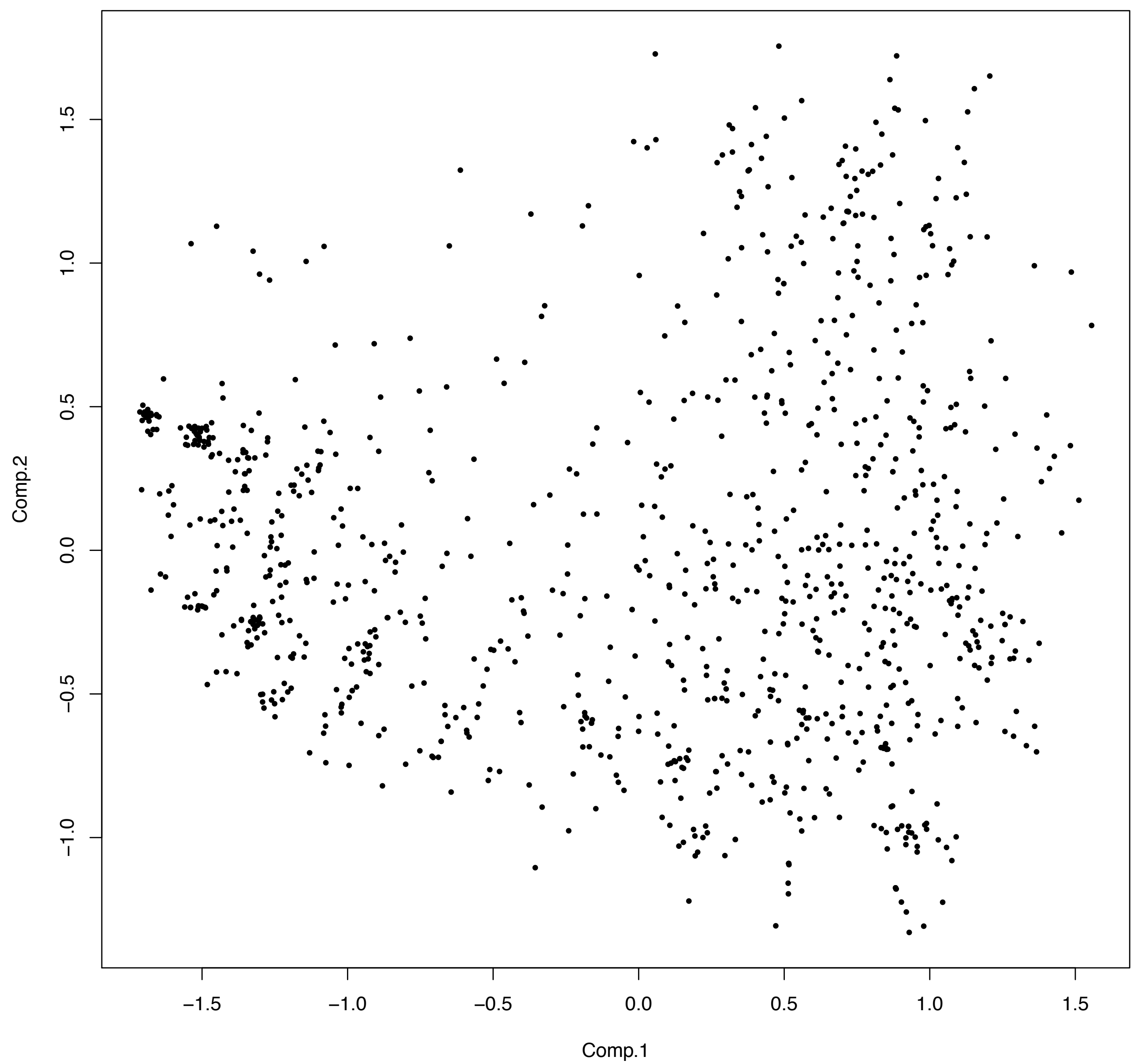
second principal component
first principal component

PCA

Principal components provide a new coordinate system that is aligned to the features of the data -- It is often used as **a tool for dimension reduction**, focusing our attention on the first few components as they capture as much of the variation in the data as possible

In this sense, then, we can **plot the data projected onto the first two principal components**, say, to generate the view from the Grand Tour that will account for the greatest variation in the data

Let's now apply the projection to our 381-dimensional recipe data and plot the "scores" for the first two components, in effect plotting our data in this new coordinate system...


Comp.2
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Naming

The columns in our original data set are meaningful, they represent the presence or absence of an ingredient -- When faced with the new coordinates of PCA, we should try to understand **what the coordinates represent**

Each principal component is just **a linear combination of our original variables** -- We have taken our 381 ingredient variables and have replaced them with 381 principal components, new variables that represent decreasing variation

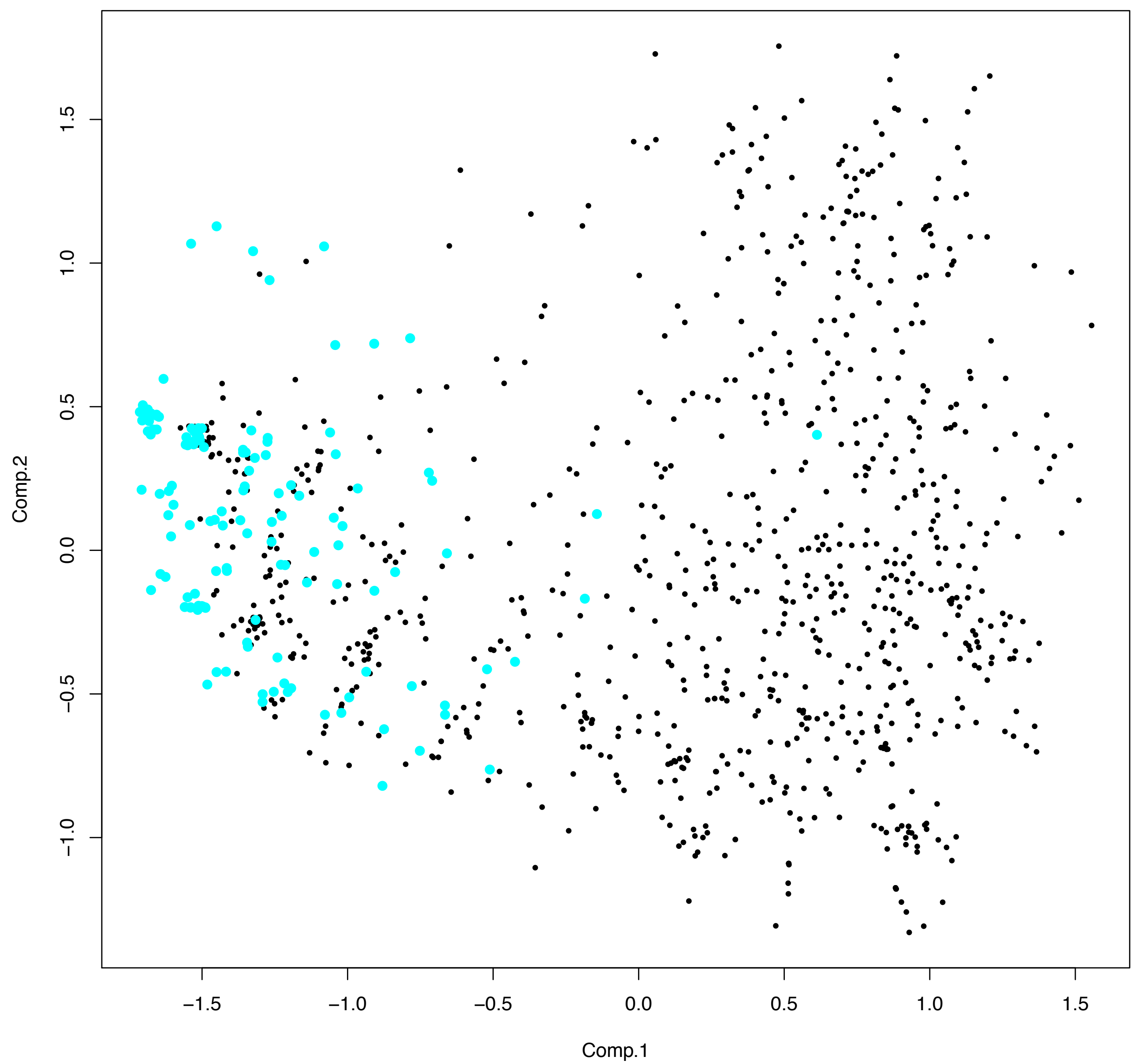
Naming

The factor loadings associated with each principal component are just the coefficients assigned to each of the original ingredients to make the new variables -- Here is what we get for the first, looking at the largest in absolute value

PC1 = -0.19 **cake.mix** - 0.13 **water** - 0.11 **vegetable.oil** - ...
+ 0.23 **butter** + 0.29 **baking.powder** + 0.31 **vanilla.extract** + 0.37 **sugar** + 0.42 **salt** + 0.44 **flour**

Those recipes scoring low in this first principal component direction involve ingredients like **cake mix, water and vegetable oil** -- Those scoring high involve **flour, sugar and baking powder**, ingredients that you need when you make a cake from "scratch"

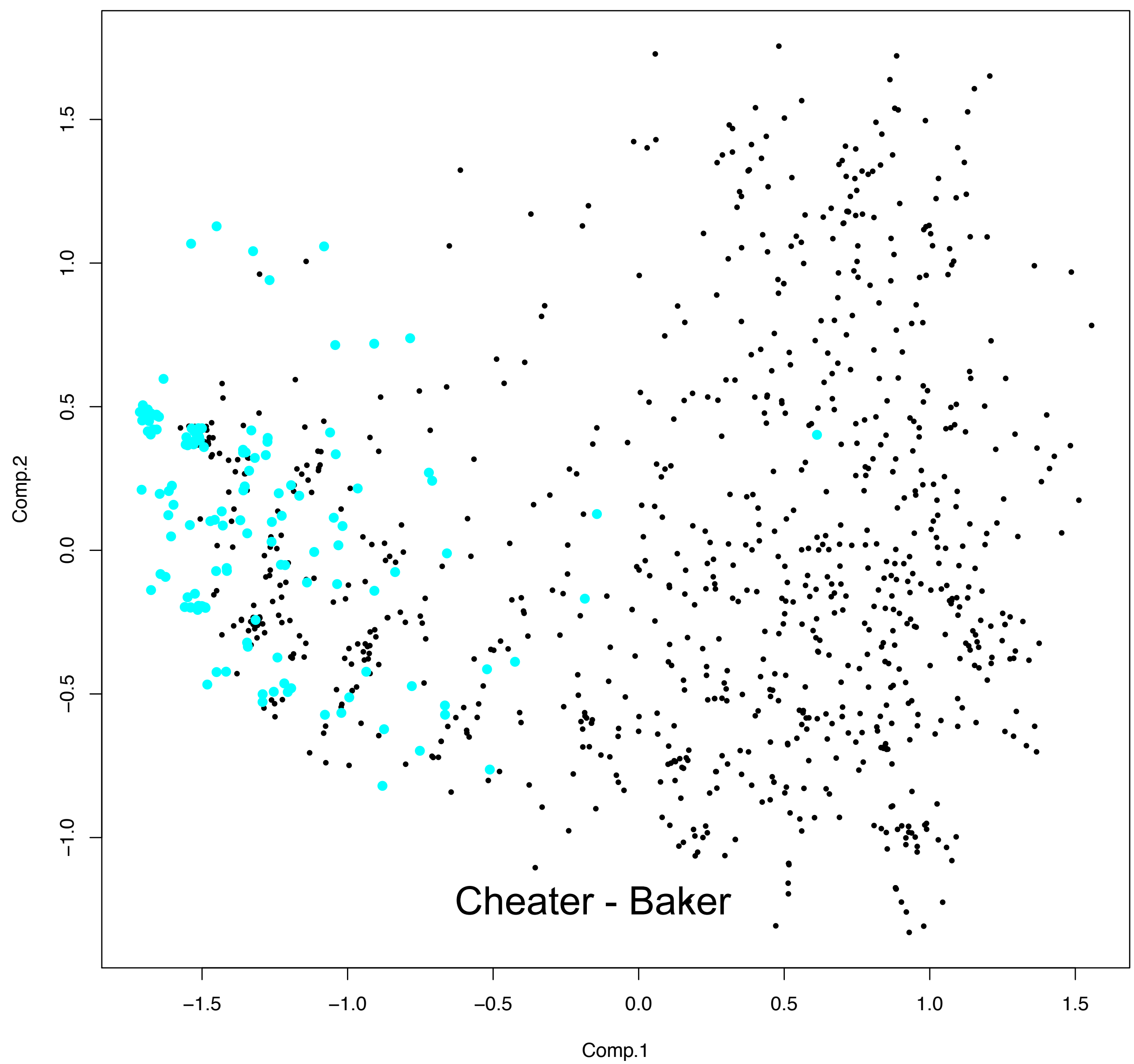
In the next slide we color recipes cyan that involve cake mix as an ingredient...


Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Naming

With this in mind, the first principal component tends to divide recipes based on whether someone needs to **bake from scratch or bootstrap with a cake mix** (or actual cake in some cases!)

We’ll call this the Cheater-Baker axis...

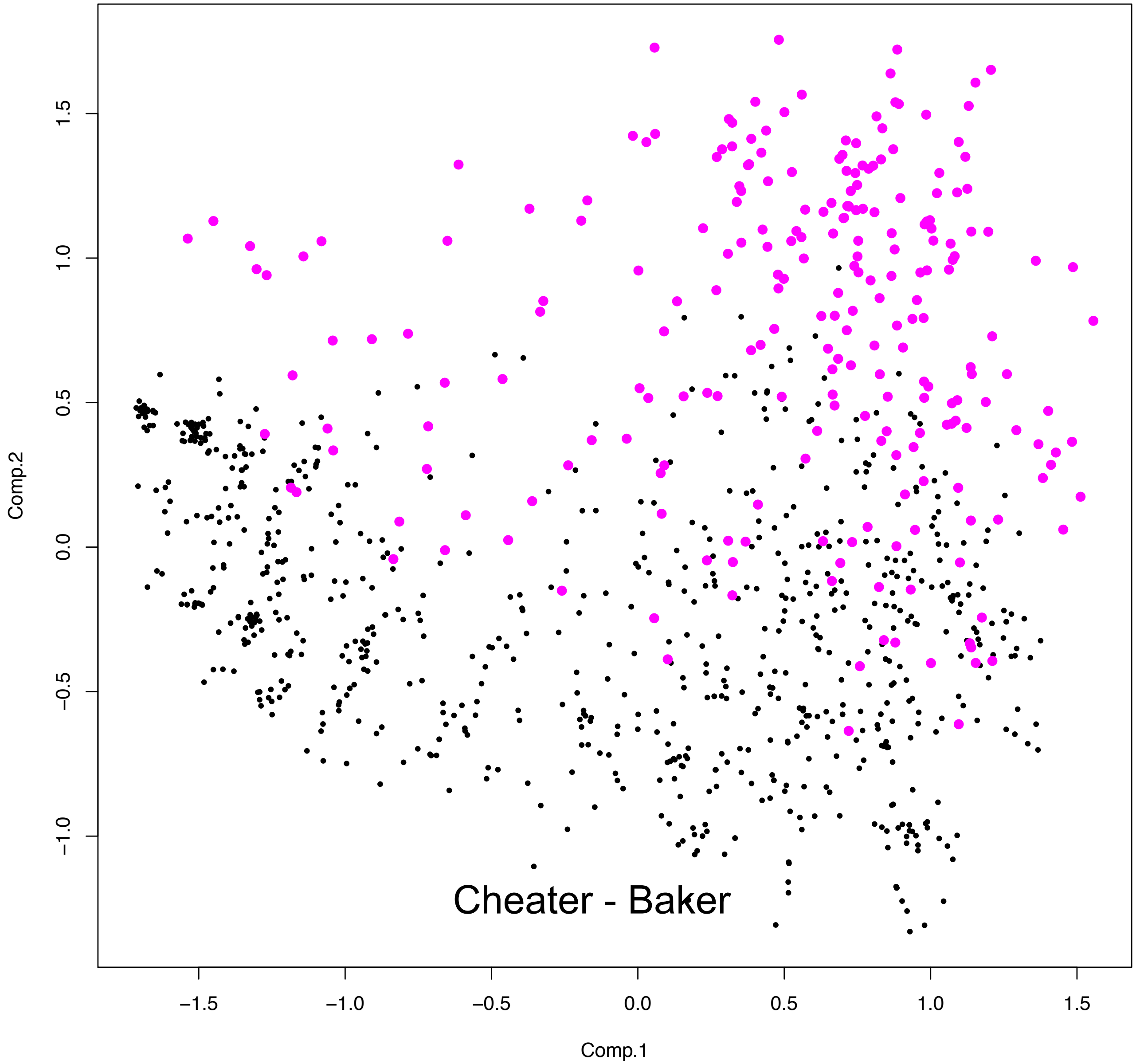


Cheater - Baker
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

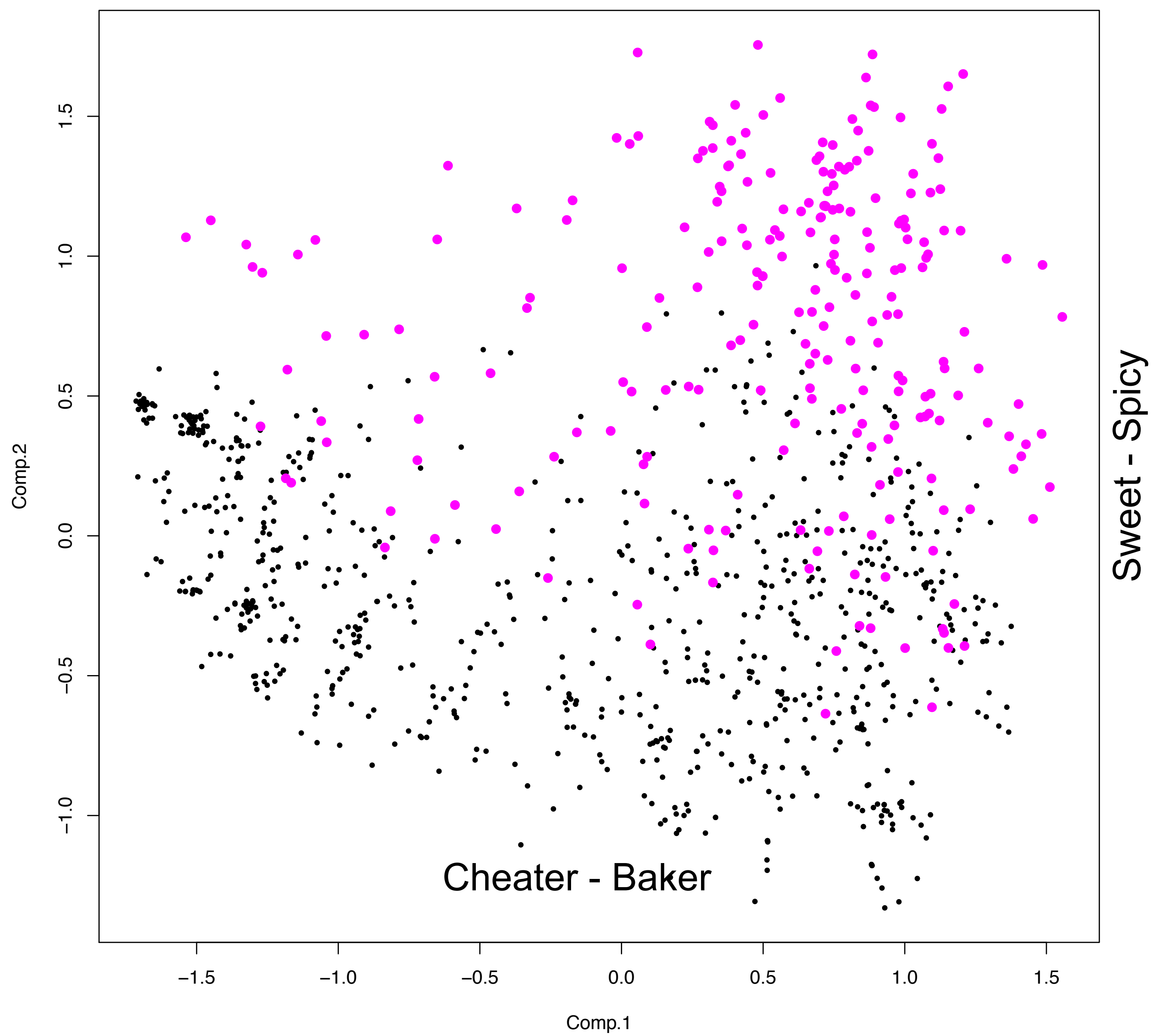
Naming

Looking at the loadings for the second principal component we see high scores for **cinnamon, nutmeg, raisins and carrots** -- On the low end we see **butter, powdered sugar and lemon zest**

We'll call the second principal component the Sweet-Spicy axis -- On the next slide, recipes containing cinnamon are highlighted in magenta



Cheater - Baker
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Sweet - Spicy
Cheater - Baker
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Under the hood

Principal components, with their focus on variability, operate on the sample variance-covariance matrix of our data -- Specifically a 381-by-381 matrix of covariances between the ingredient variables

After a little (linear) algebra, we find that the principal components are really just the eigenvectors of this matrix, and the variation they capture the eigenvalues -- So computationally, this technique is pretty straightforward

A general methodology

Using understanding or “reading” large quantities of text as an application, we introduced a number of methodological tools for viewing and structuring high-dimensional data

While pairwise scatterplots can gives us a sense of the (marginal) distribution of our data, **projection and the visualization of a series of (random or guided) projections** is an effective approach for uncovering more complex structure

Principal components analysis focuses on one particular projection of the data -- It is a technique for **reducing the dimension of a data set** by compacting its major variation into (typically) a few derived variables

It has a huge number of applications...

Behav Ecol Sociobiol (2009) 63:1057–1066
DOI 10.1007/s00265-009-0739-0

ORIGINAL PAPER

Eigenbehaviors: identifying structure in routine

Nathan Eagle · Alex Sandy Pentland

Received: 12 September 2007 / Revised: 24 February 2009 / Accepted: 24 February 2009 / Published online: 7 April 2009

Abstract Longitudinal behavioral data generally contains a significant amount of structure. In this work, we identify the structure inherent in daily behavior with models that can accurately analyze, predict, and cluster multimodal data from individuals and communities within the social network of a population. We represent this behavioral structure by the principal components of the complete behavioral dataset, a set of characteristic vectors we have termed eigenbehaviors. In our model, an individual's behavior over a specific day can be approximated by a weighted sum of his or her primary eigenbehaviors. When these weights are calculated halfway through a day, they can be used to predict the day's remaining behaviors with 79% accuracy for our test subjects. Additionally, we demonstrate the potential for this dimensionality reduction technique to infer community affiliations within the subjects' social network by clustering individuals into a "behavior space" spanned by a set of their aggregate eigenbehaviors. These behavior spaces make it possible to determine the behavioral similarity between both individuals and groups, enabling 96% classification accuracy of community affiliations within the population-level social network. Additionally, the distance between individuals in the behavior space can be used as an estimate for relational ties such as friendship, suggesting strong behavioral homophily amongst the subjects. This approach capitalizes on the large amount of rich data previously captured during the Reality Mining study from mobile phones continuously logging location, proximate phones, and communication of 100 subjects at MIT over the course of 9 months. As wearable sensors continue to generate these types of rich, longitudinal datasets, dimensionality reduction techniques such as eigenbehaviors will play an increasingly important role in behavioral research.

Keywords Behavioral modeling · Machine learning · Eigendecomposition

Introduction

While discrete observations of an individual's idiosyncratic behavior can appear almost random, typically there are repeating and easily identifiable routines in every person's life. These patterns become more apparent when the behavior is temporally, spatially, and socially contextualized. However, building models of long-term behavior has been hampered due to the lack of contextualized behavioral data. Additionally, traditional Markov models work well for specific set of behaviors, but have difficulty incorporating temporal patterns across different timescales (Clarkson 2002). We present a new methodology for identifying the repeating structures underlying behavior. These structures are represented by *eigenbehaviors*, the principal components of an individual's behavioral dataset.

To capture these characteristic behaviors, we compute the principal components of an individual's behavioral data.

Communicated by Guest Editor D. Lusseau

This contribution is part of the special issue "Social Networks: new perspectives" (Guest Editors: J. Krause, D. Lusseau and R. James).

N. Eagle · A. S. Pentland
MIT Media Laboratory, Massachusetts Institute of Technology,
E15-383, 20 Ames St.,
Cambridge, MA 02139, USA

N. Eagle (✉)
The Santa Fe Institute,
1399 Hyde Park Rd,
Santa Fe, NM 87501, USA
e-mail: nathan@mit.edu

Behav Ecol Sociobiol (2009) 63:1057–1066
DOI 10.1007/s00265-009-0739-0

ORIGINAL PAPER

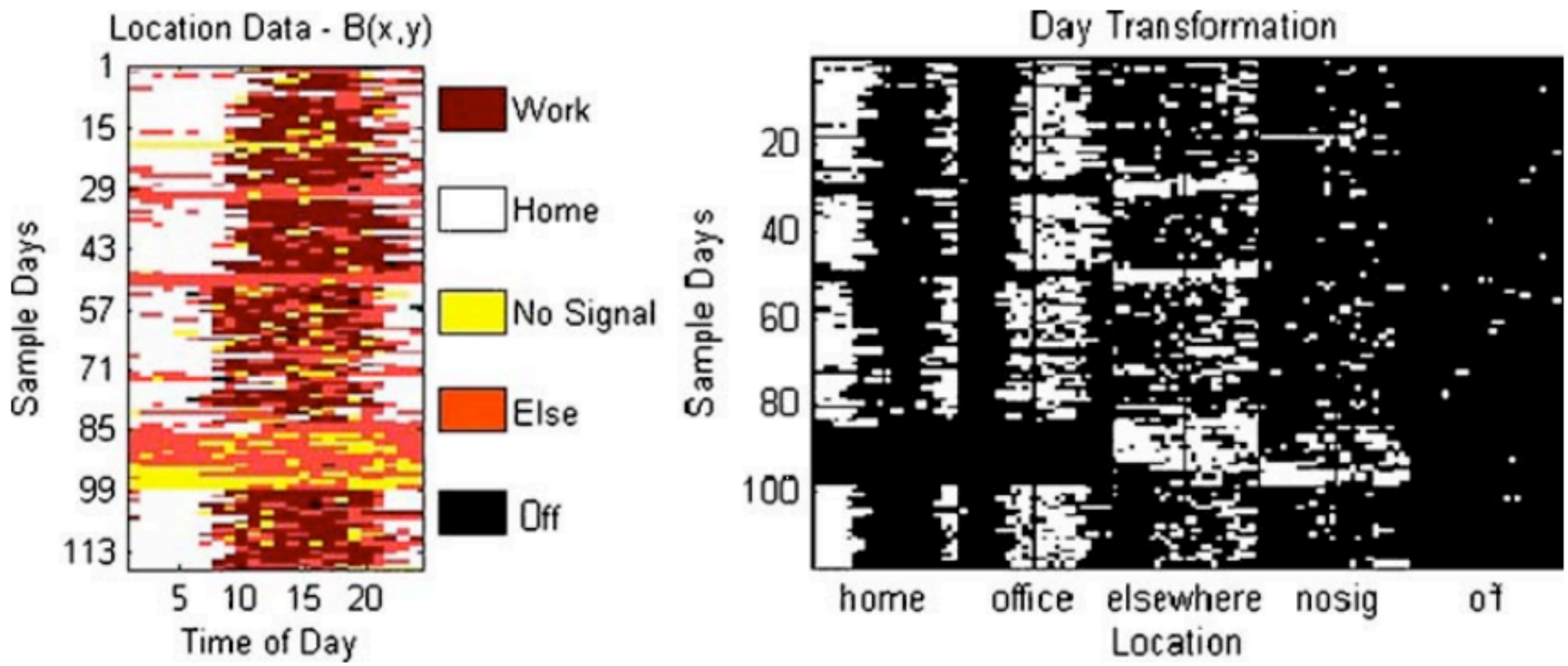
Eigenbehaviors: identifying structure in routine

Nathan Eagle · **Alex Sandy Pentland**

Received: 12 September 2007 / Revised: 24 February 2009 / Accepted: 24 February 2009 / Published online: 7 April 2009

Abstract Longitudinal behavioral data generally contains a significant amount of structure. In this work, we identify the structure inherent in daily behavior with models that can accurately analyze, predict, and cluster multimodal data from individuals and communities within the social network of a population. We represent this behavioral structure by the principal components of the complete behavioral dataset, a set of characteristic vectors we have termed eigenbehaviors. In our model, an individual's behavior over a specific day can be approximated by a weighted sum of his or her primary eigenbehaviors. When these weights are calculated halfway through a day, they can be used to predict the day's remaining behaviors with 79% accuracy for our test subjects. Additionally, we demonstrate the potential for this dimensionality reduction network. Additionally, the distance between individuals in the behavior space can be used as an estimate for relational ties such as friendship, suggesting strong behavioral homophily amongst the subjects. This approach capitalizes on the large amount of rich data previously captured during the Reality Mining study from mobile phones continuously logging location, proximate phones, and communication of 100 subjects at MIT over the course of 9 months. As wearable sensors continue to generate these types of rich, longitudinal datasets, dimensionality reduction techniques such as eigenbehaviors will play an increasingly important role in behavioral research.

Keywords Behavioral modeling · Machine learning · Eigendecomposition



Location Data - B(x,y)
Work
Home
No Signal
Else
Off
Sample Days
Time of Day
Day Transformation
Sample Days
home
office
elsewhere
nosig
off
Location

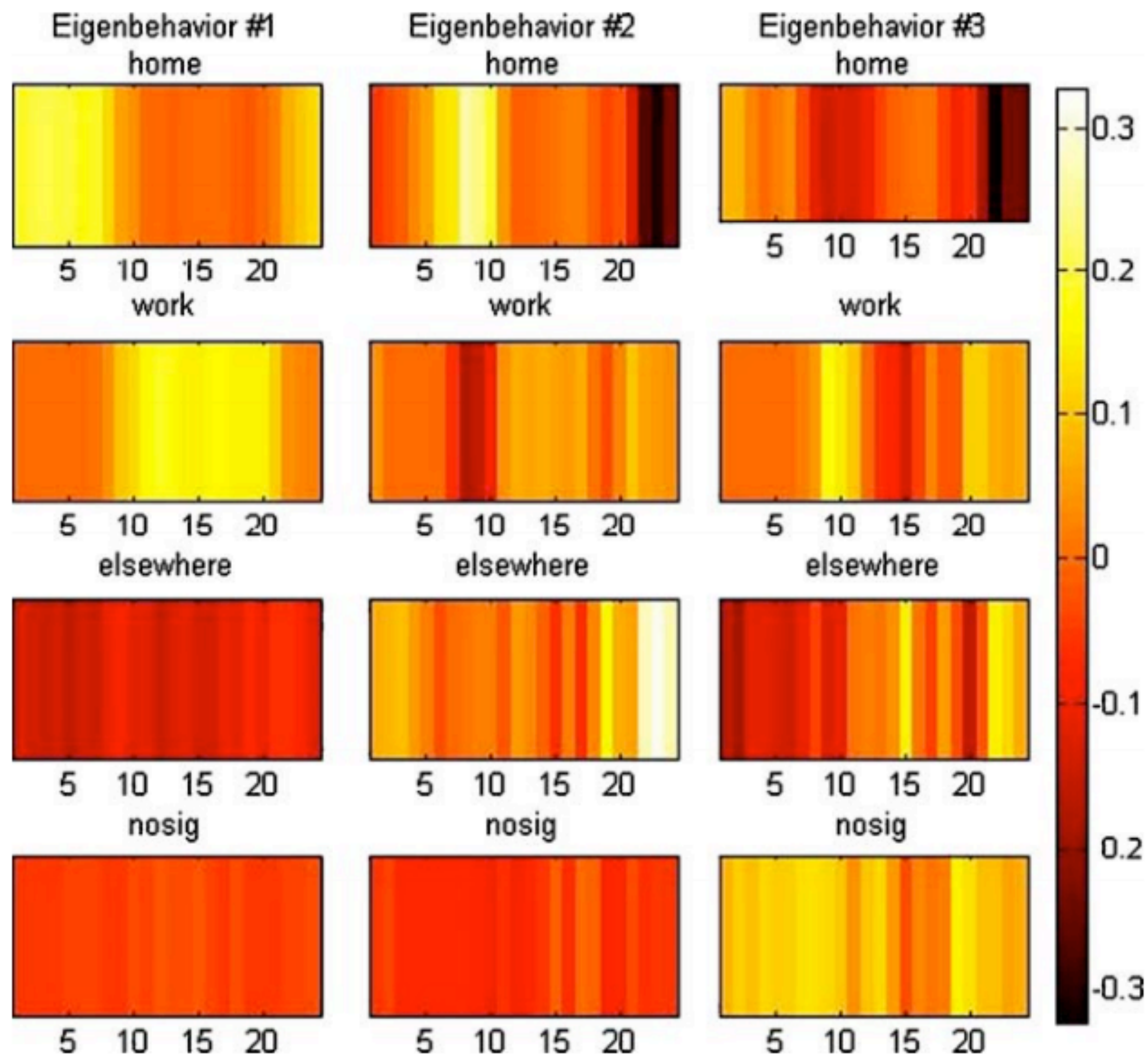


Fig. 2 The top three eigenbehaviors, $[u_1, u_2, u_3]$, for Subject 4. The first eigenbehavior (represented with the *first column of three figures*) corresponds to whether it is a normal day or whether the individual is traveling. If the first weight is positive, then this eigenbehavior shows that the subject's typical pattern of behavior consists of midnight to 9:00 at home, 10:00 to 20:00 at work, and then the subject returns home at approximately 21:00. The second eigenbehavior (and similarly the *middle column of three figures*) corresponds to typical weekend behavior. It is highly likely the subject will remain at home past 10:00 in the morning and will be out on the town ('elsewhere') later that evening. The third eigenbehavior is most active when the individual is in locations where the phone has no signal

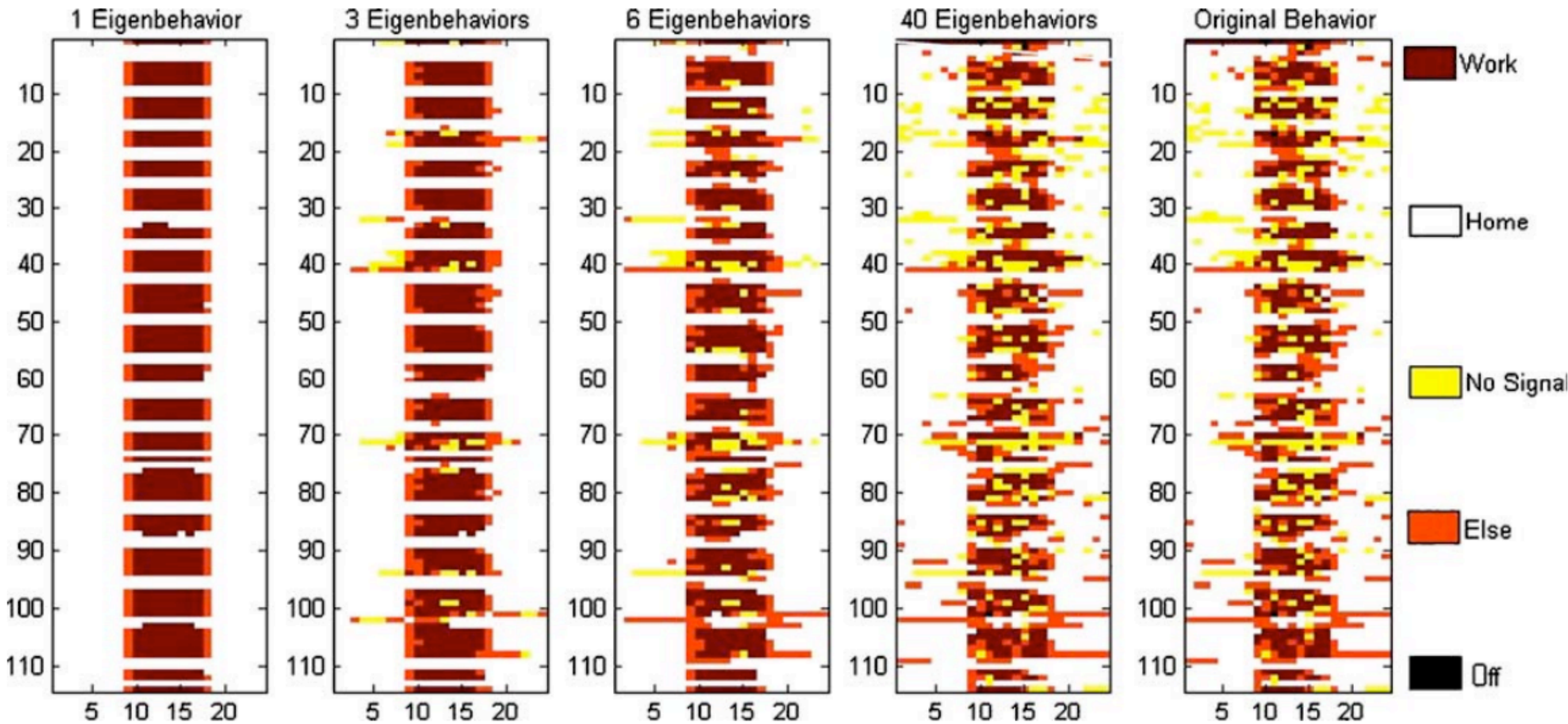


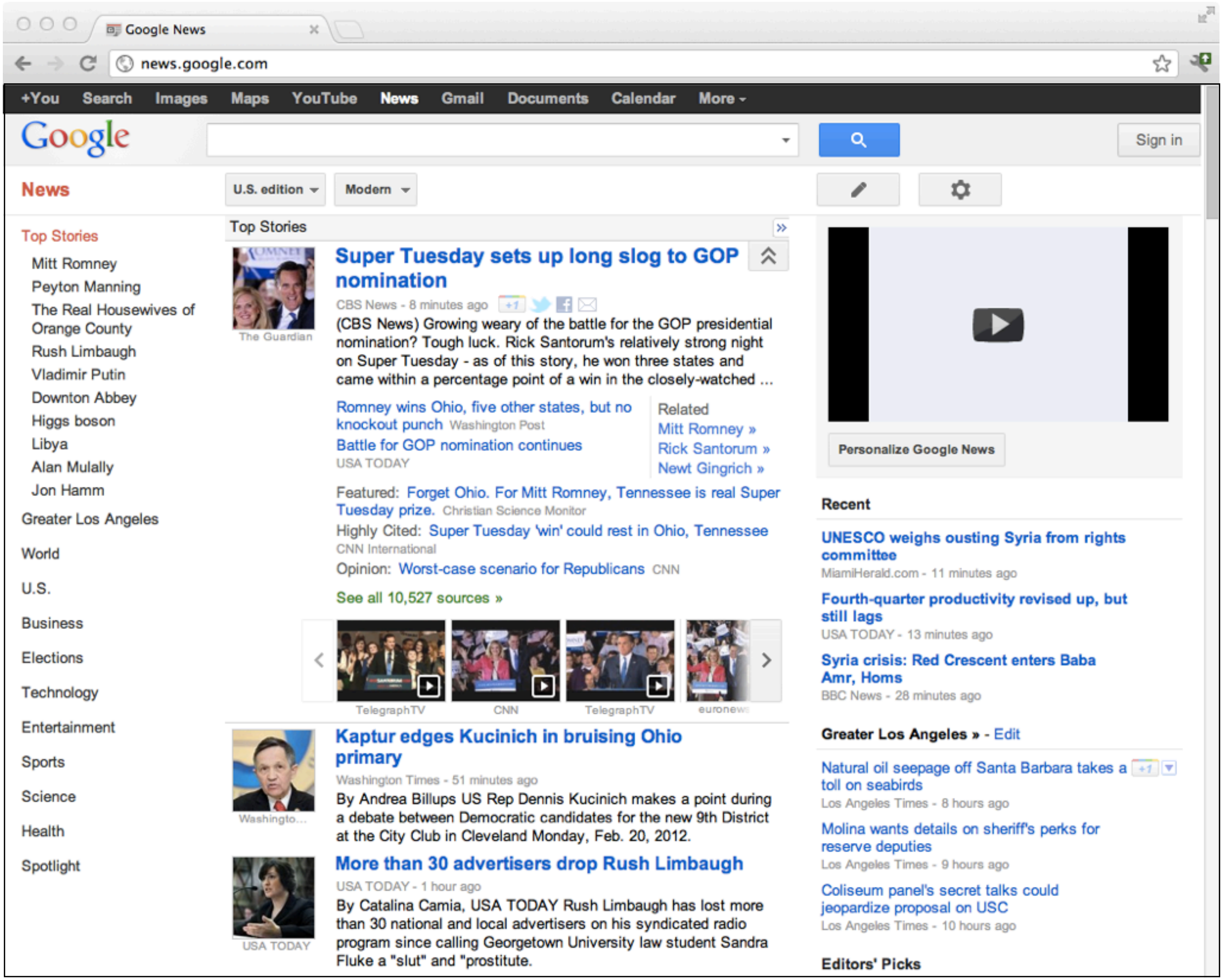
Fig. 3 Behavior approximation of 115 days using a varying number of eigenbehaviors. The *left-most* figure corresponds to behavioral approximation using only one eigenbehavior. The approximation accuracy increases with the number of eigenbehaviors

Document clustering

This simple investigation suggests that recipes fall into **natural groups depending on their ingredient lists** -- This is probably obvious, but to see it in the data is always a satisfying moment

We are interested in these groups or “clusters” not only because they seem to agree with our intuition, but also for a variety of practical reasons -- In web searching, for example, it might be useful to not just return a long list of pages containing a key word, but instead **organize things by “topic” or cluster**

Google News, for example, does this for news stories...



Google News
news.google.com
+You Search Images Maps YouTube News Gmail Documents Calendar More
Google
Sign in
News
U.S. edition
Modern
Top Stories
Mitt Romney
Peyton Manning
The Real Housewives of Orange County
Rush Limbaugh
Vladimir Putin
Downton Abbey
Higgs boson
Libya
Alan Mulally
Jon Hamm
Greater Los Angeles
World
U.S.
Business
Elections
Technology
Entertainment
Sports
Science
Health
Spotlight
Top Stories
Super Tuesday sets up long slog to GOP nomination
The Guardian
CBS News - 8 minutes ago
(CBS News) Growing weary of the battle for the GOP presidential nomination? Tough luck. Rick Santorum's relatively strong night on Super Tuesday - as of this story, he won three states and came within a percentage point of a win in the closely-watched ...
Romney wins Ohio, five other states, but no knockout punch Washington Post
Battle for GOP nomination continues USA TODAY
Related
Mitt Romney »
Rick Santorum »
Newt Gingrich »
Featured: Forget Ohio. For Mitt Romney, Tennessee is real Super Tuesday prize. Christian Science Monitor
Highly Cited: Super Tuesday 'win' could rest in Ohio, Tennessee CNN International
Opinion: Worst-case scenario for Republicans CNN
See all 10,527 sources »
TelegraphTV
CNN
TelegraphTV
Kaptur edges Kucinich in bruising Ohio primary
Washingto...
Washington Times - 51 minutes ago
By Andrea Billups US Rep Dennis Kucinich makes a point during a debate between Democratic candidates for the new 9th District at the City Club in Cleveland Monday, Feb. 20, 2012.
More than 30 advertisers drop Rush Limbaugh
USA TODAY
USA TODAY - 1 hour ago
By Catalina Camia, USA TODAY Rush Limbaugh has lost more than 30 national and local advertisers on his syndicated radio program since calling Georgetown University law student Sandra Fluke a "slut" and "prostitute.
Personalize Google News
Recent
UNESCO weighs ousting Syria from rights committee
MiamiHerald.com - 11 minutes ago
Fourth-quarter productivity revised up, but still lags
USA TODAY - 13 minutes ago
Syria crisis: Red Crescent enters Baba Amr, Homs
BBC News - 28 minutes ago
Greater Los Angeles » - Edit
Natural oil seepage off Santa Barbara takes a toll on seabirds
Los Angeles Times - 8 hours ago
Molina wants details on sheriff's perks for reserve deputies
Los Angeles Times - 9 hours ago
Coliseum panel's secret talks could jeopardize proposal on USC
Los Angeles Times - 10 hours ago
Editors' Picks

Clustering

Clustering or, rather, dividing the data into natural groups, is also known as "unsupervised learning" -- It is unsupervised in the sense that we don't have any tags telling us that a document begins to one class or another (these problems are referred to as "supervised learning")

Clustering, like regression, is a statistical idea that has been around for **a long, long time** -- Rather than review the history, we'll simply note that there are hundreds of different approaches to identifying groups in data

We will audition two here, in part because they have a lot of traction in the natural language processing community

K-means clustering

K-means clustering divides our data $X_1, \ldots, X_n$ into K groups $G_1, \ldots, G_K$ (you specify K), where the groups are associated with cluster centers $\mu_1, \ldots, \mu_K$ (points in the d-dimensional space along with the data)

A point X_i is associated with group G_j if its nearest center (distance again!) is μ_j -- So the cluster centers are attractors, and each group is defined as the data points closest to the given center

K-means clustering

Given data $X_1, \ldots, X_n$ and a pre-specified value of K, we "learn" this model by finding values for $\mu_1, \ldots, \mu_K$ so as to "minimize" the overall loss

$$V = \sum_{k=1}^{K} \sum_{X_i \in G_k} \|X_i - \mu_k\|^2$$

We solve this problem iteratively...

K-means clustering

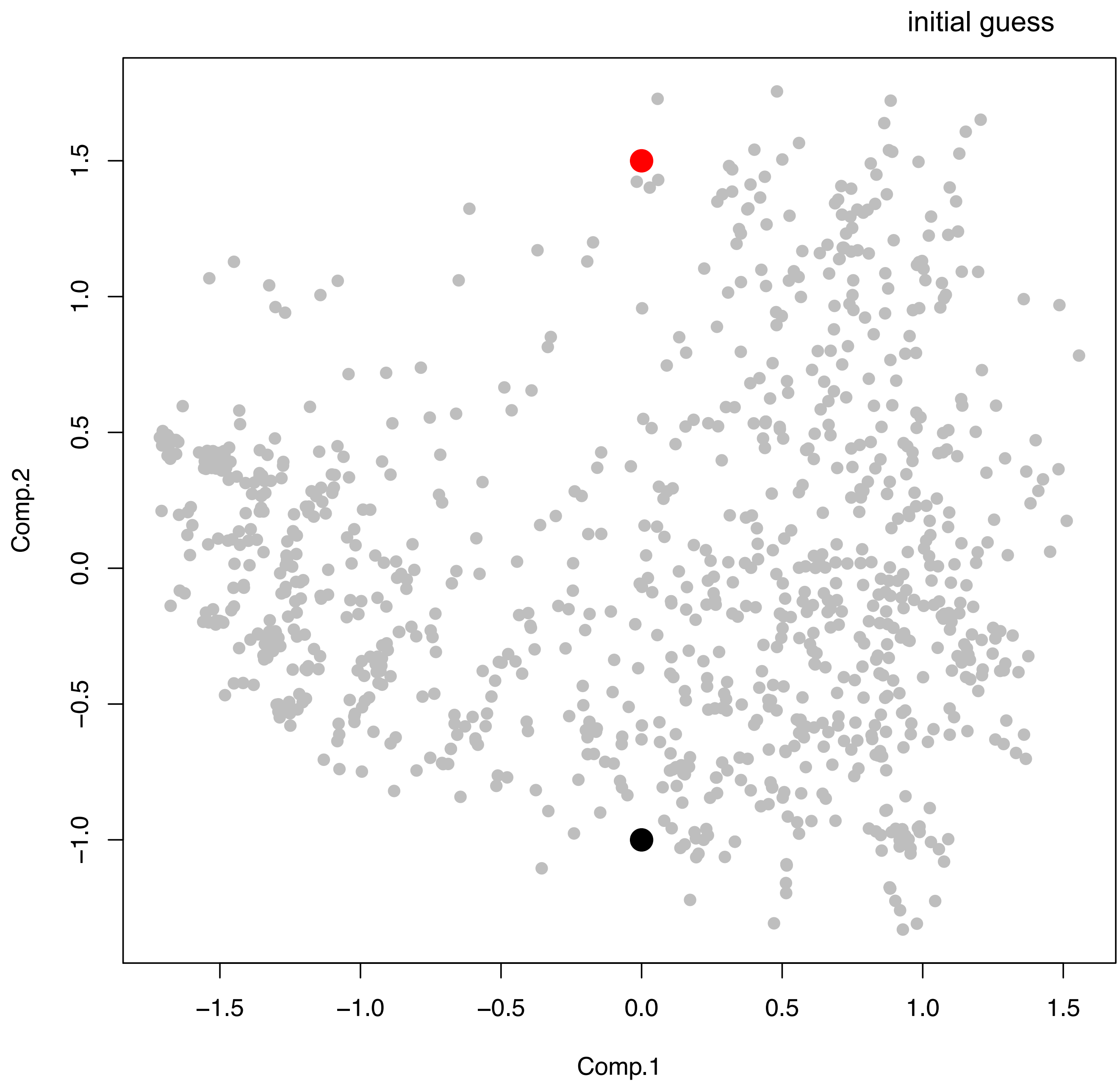
With K-means, we want to divide our data $X_1, \ldots, X_n$ into, well, K groups or clusters
-- The algorithm is pretty simple

Make an initial guess for the centers $\mu_1^0, \ldots, \mu_K^0$

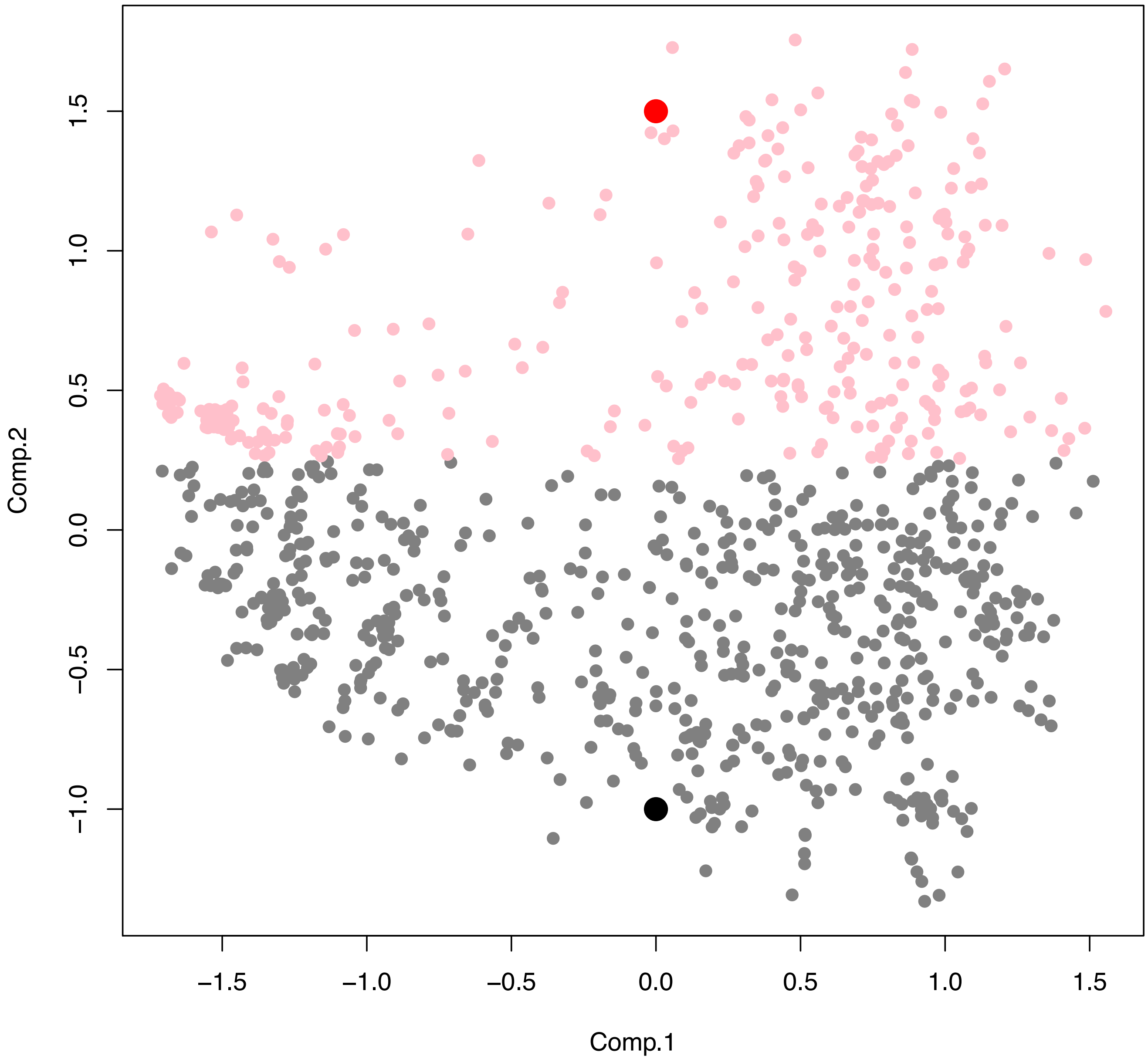
Until there's no change in these values do:

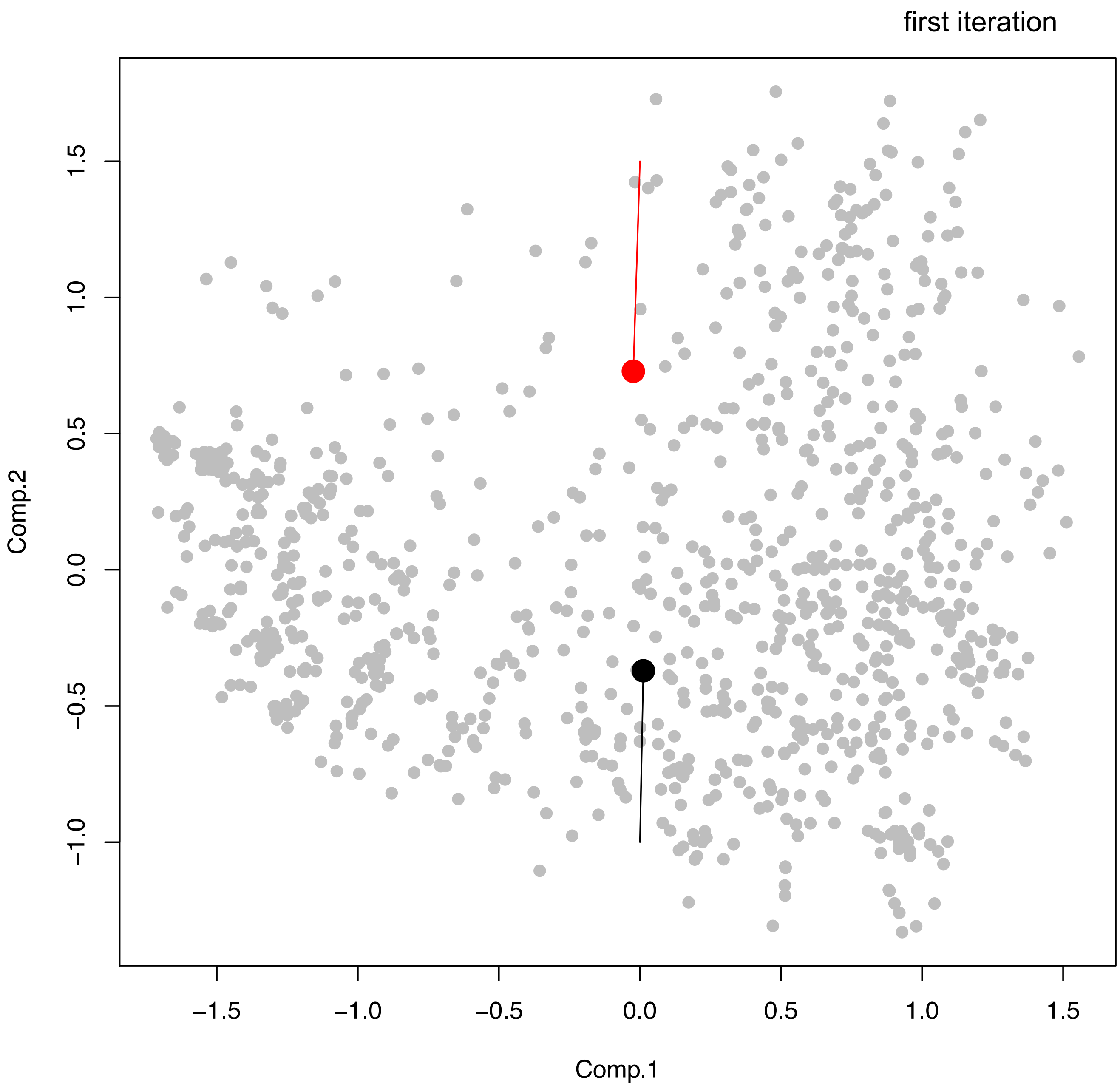
1. Assign each data point X_i to the nearest cluster center using simple Euclidean distance

2. For each cluster k, update μ_k^1 to be the mean of all the points associated with the group

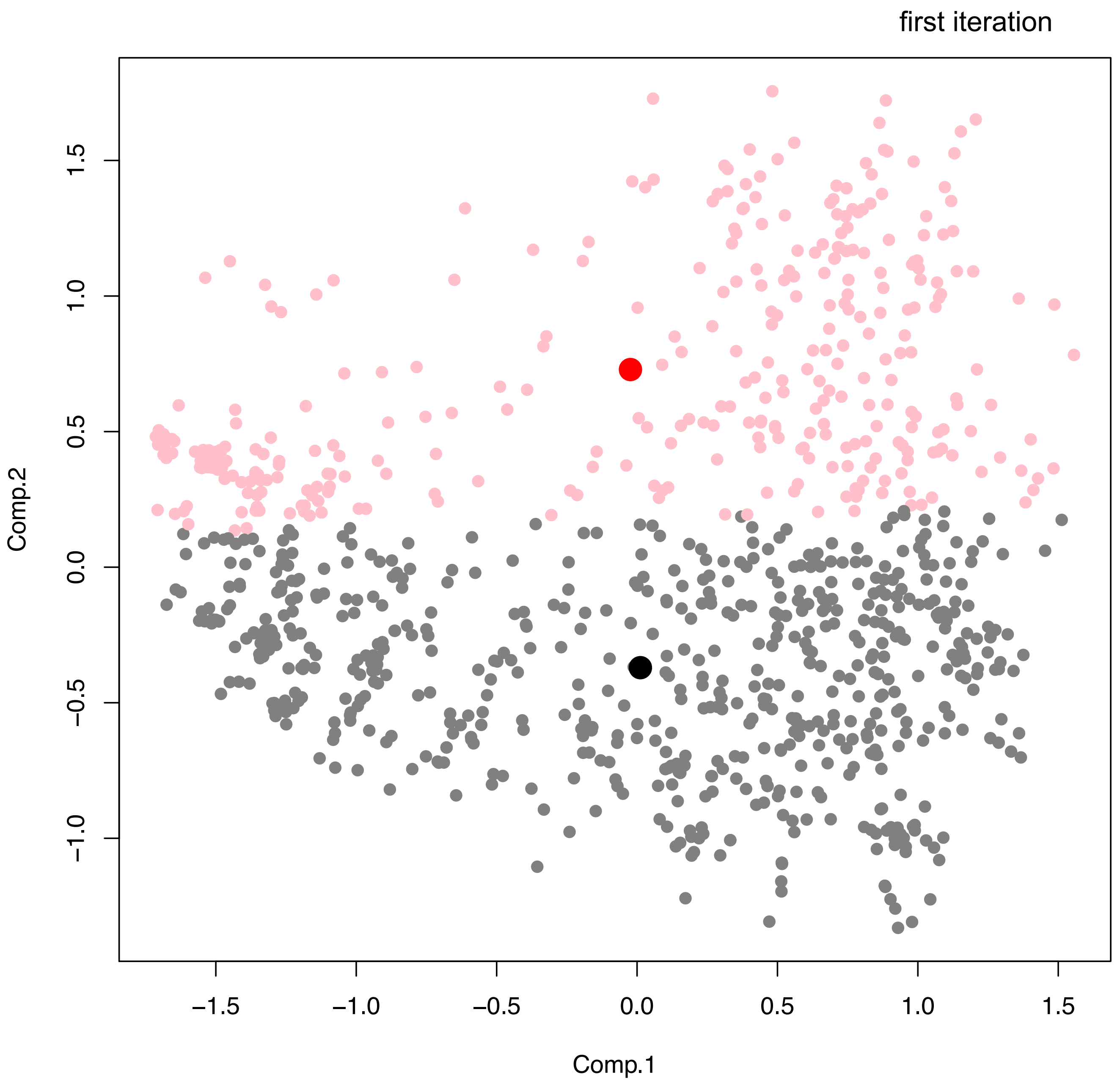

initial guess
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

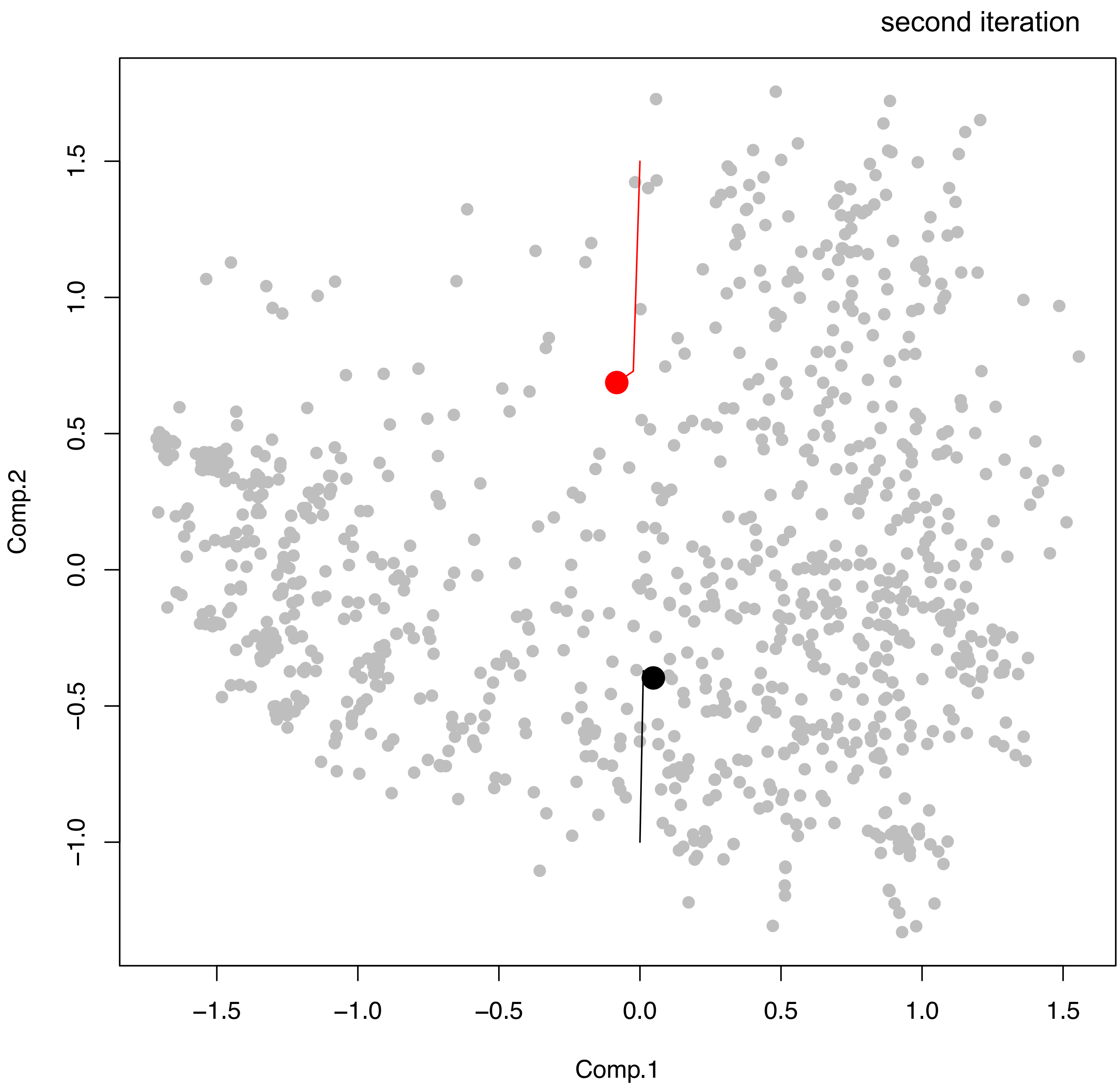
initial guess



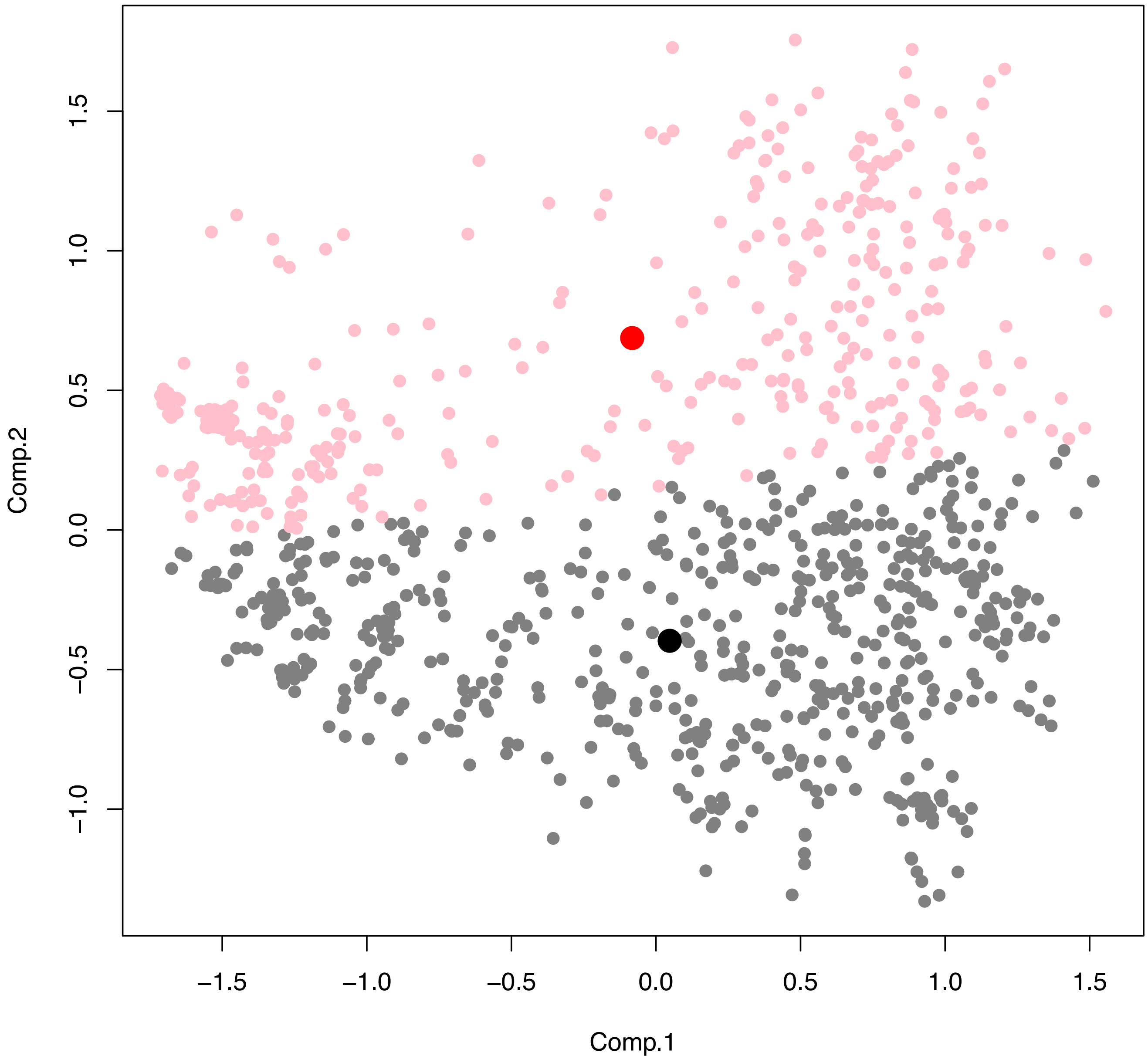


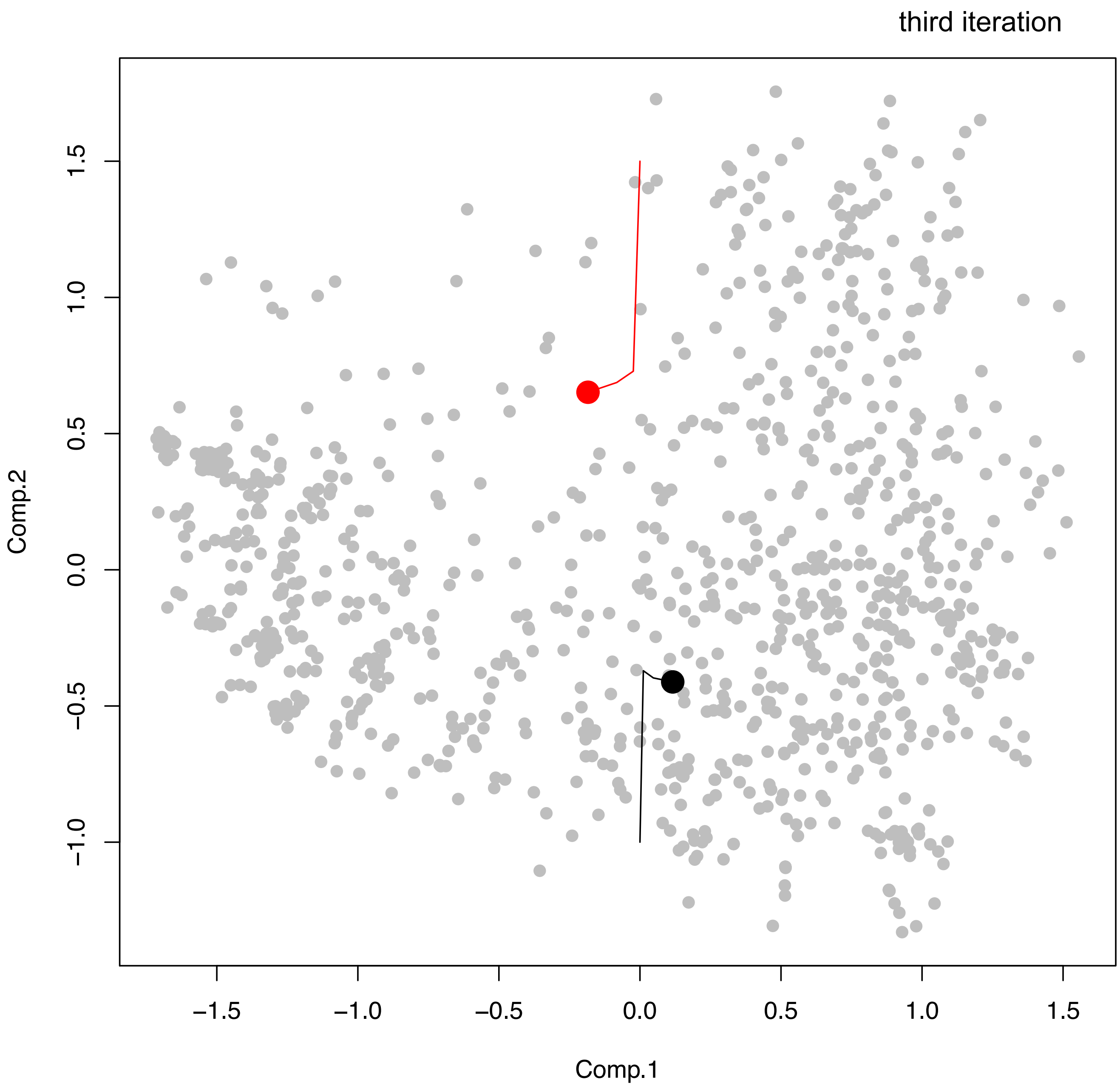
first iteration
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


first iteration
Comp.2
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

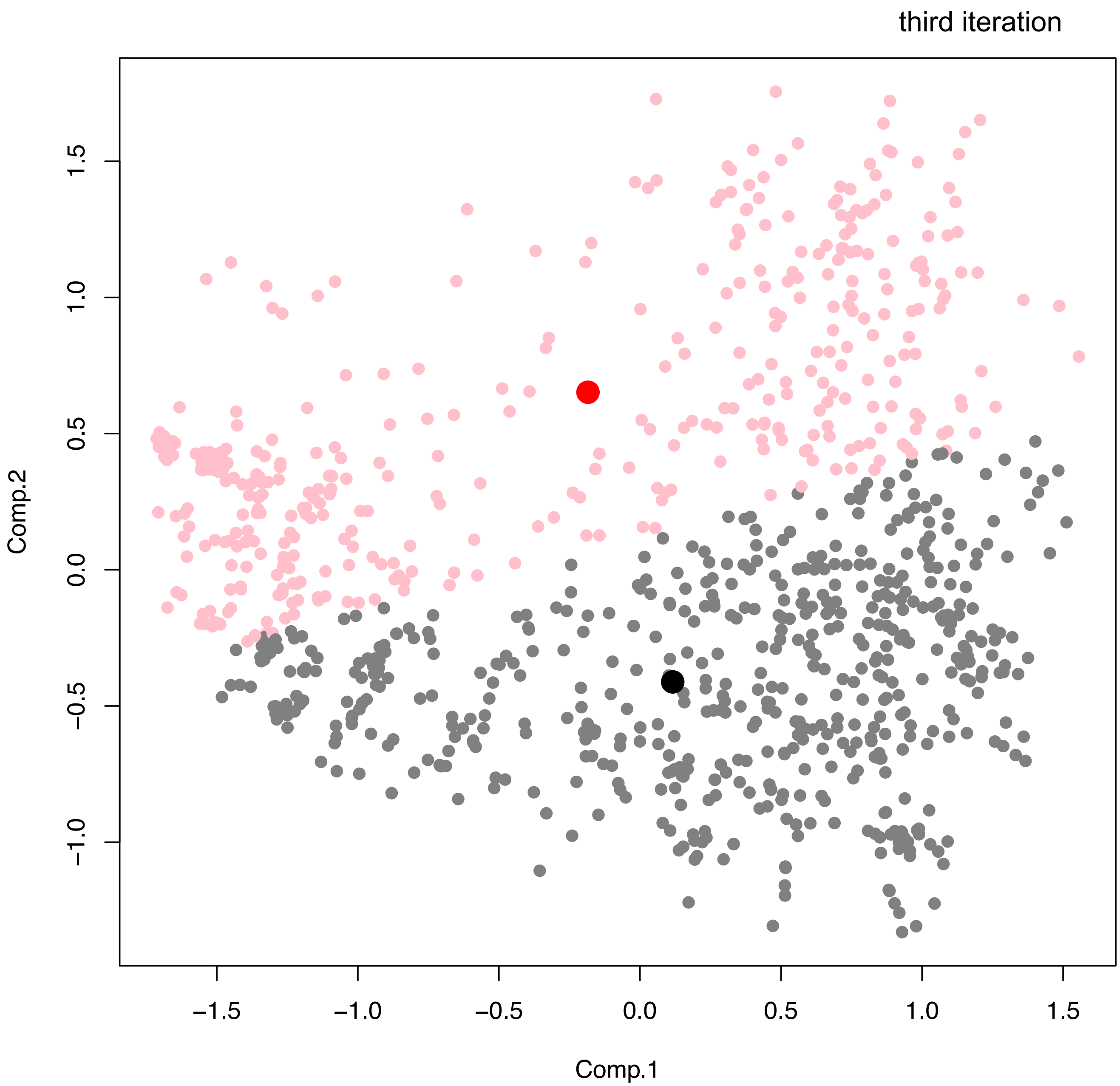

second iteration
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

second iteration

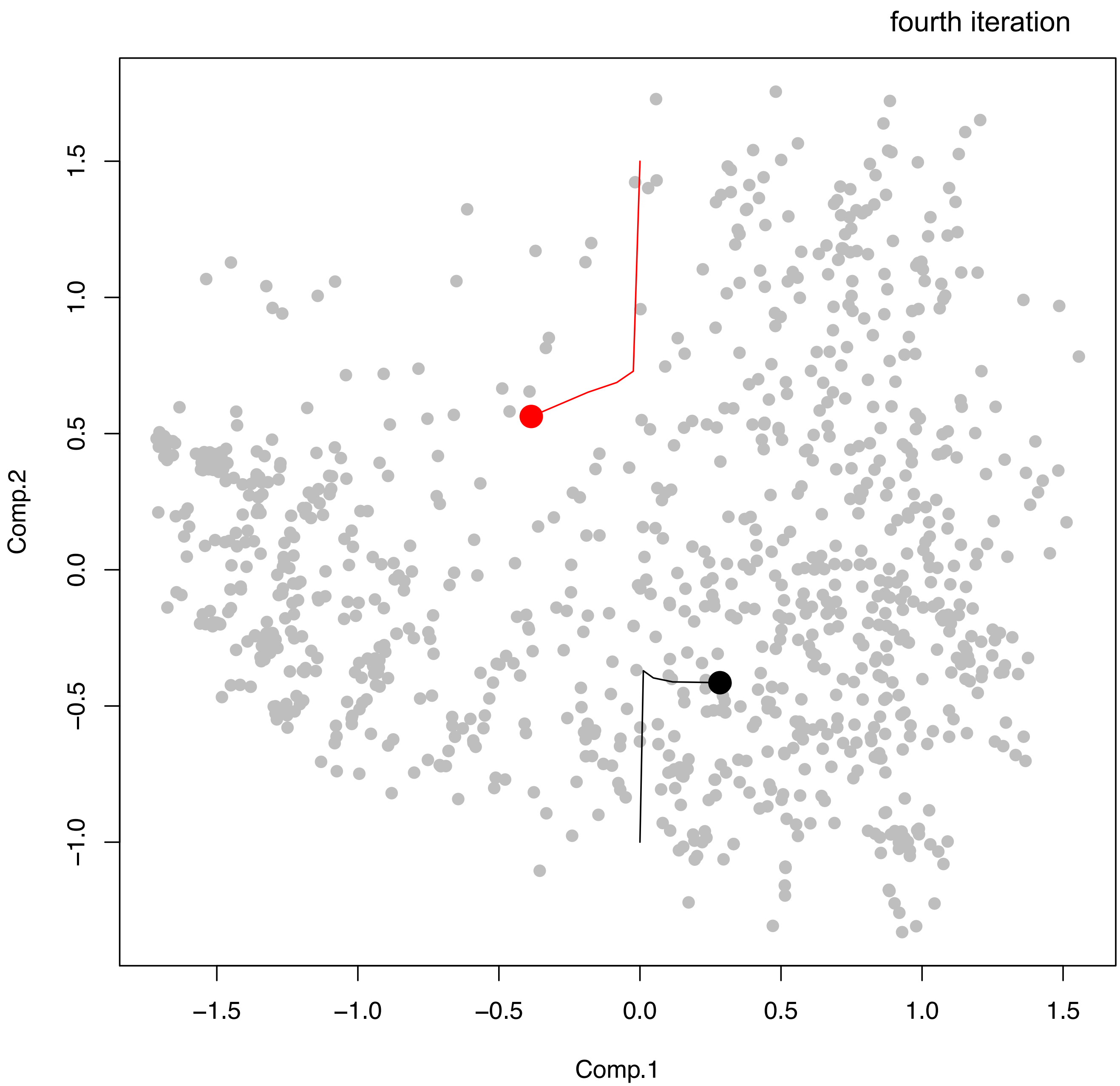


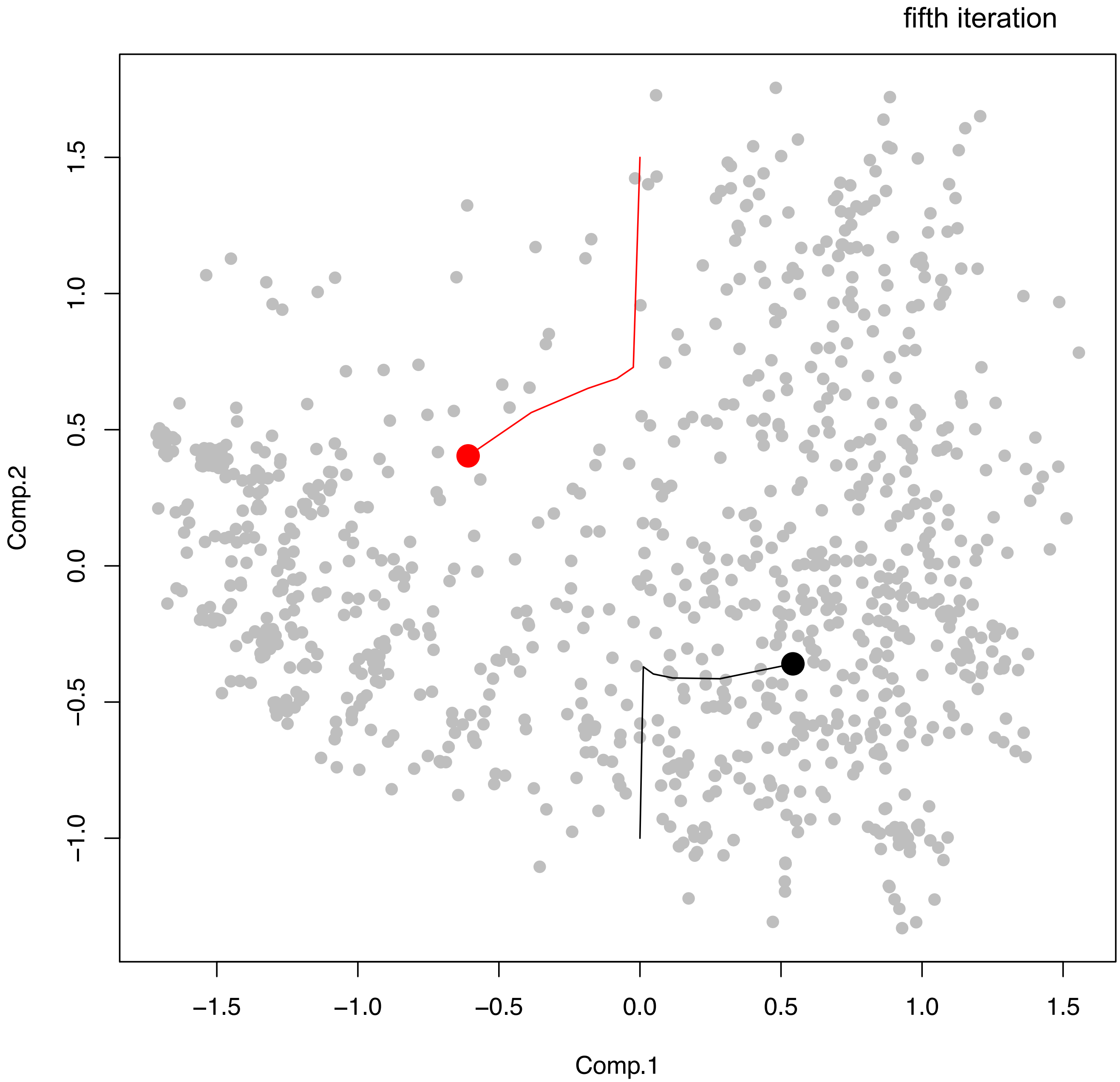


third iteration
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

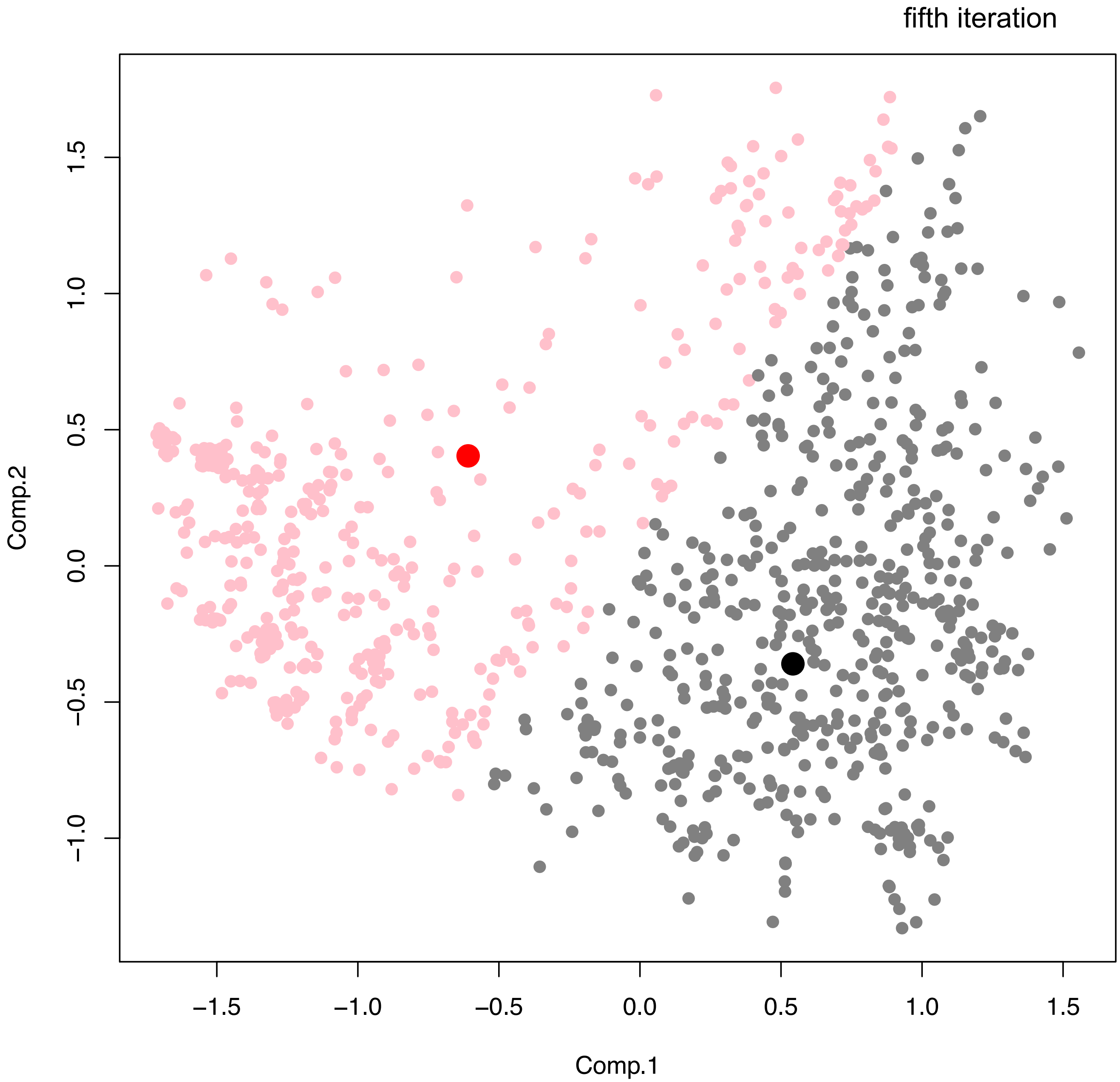


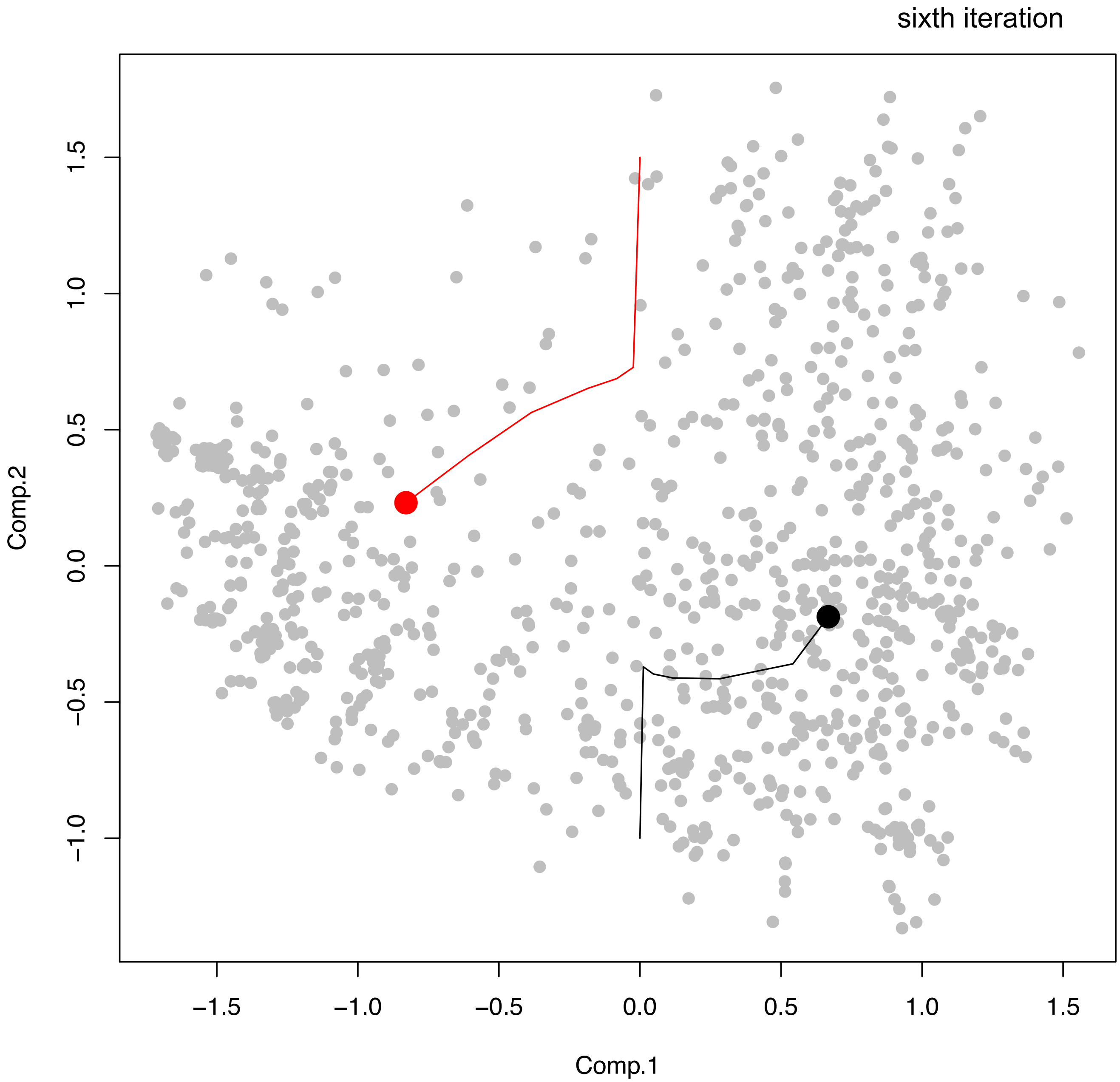
third iteration
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


fourth iteration
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

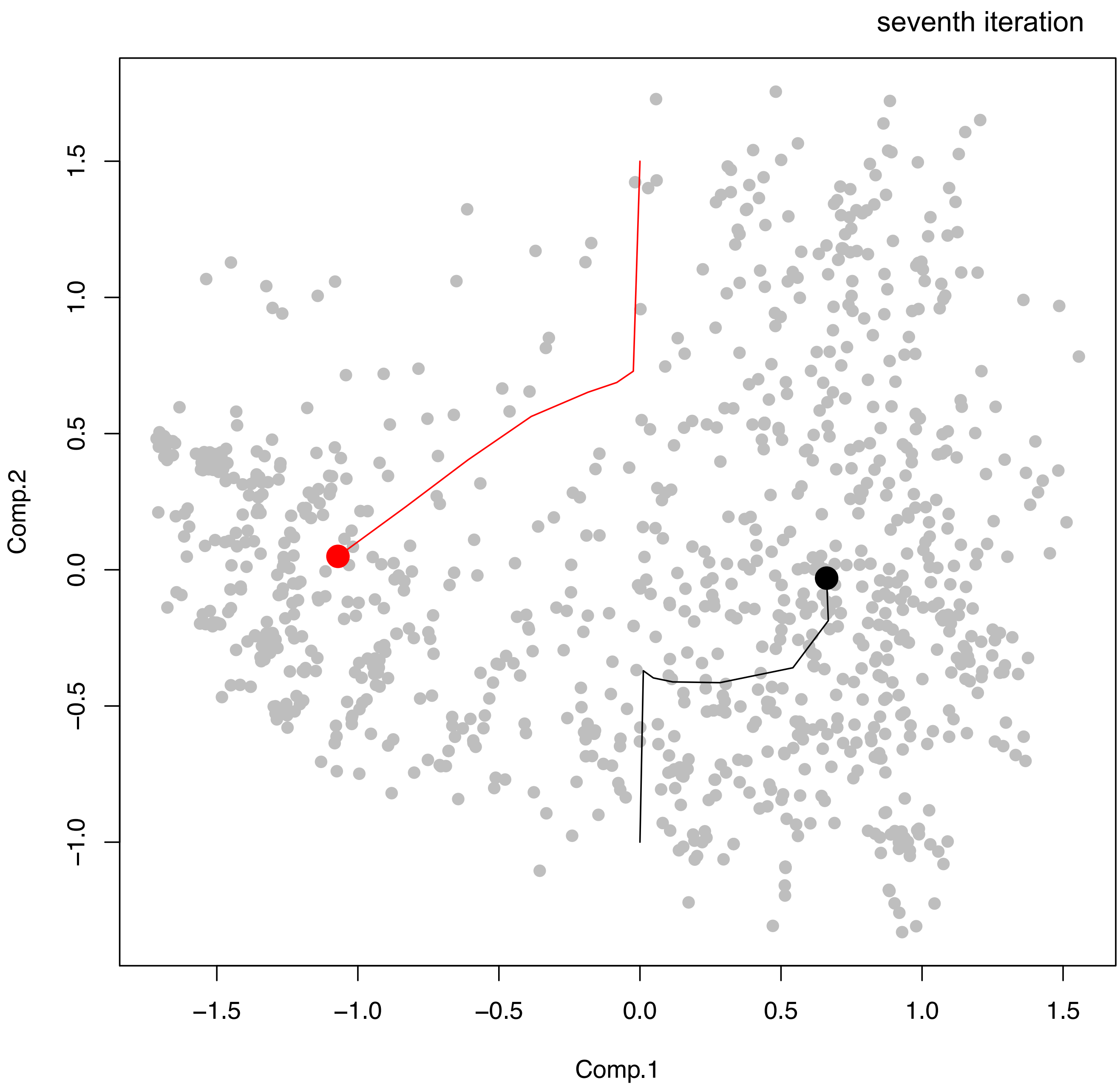


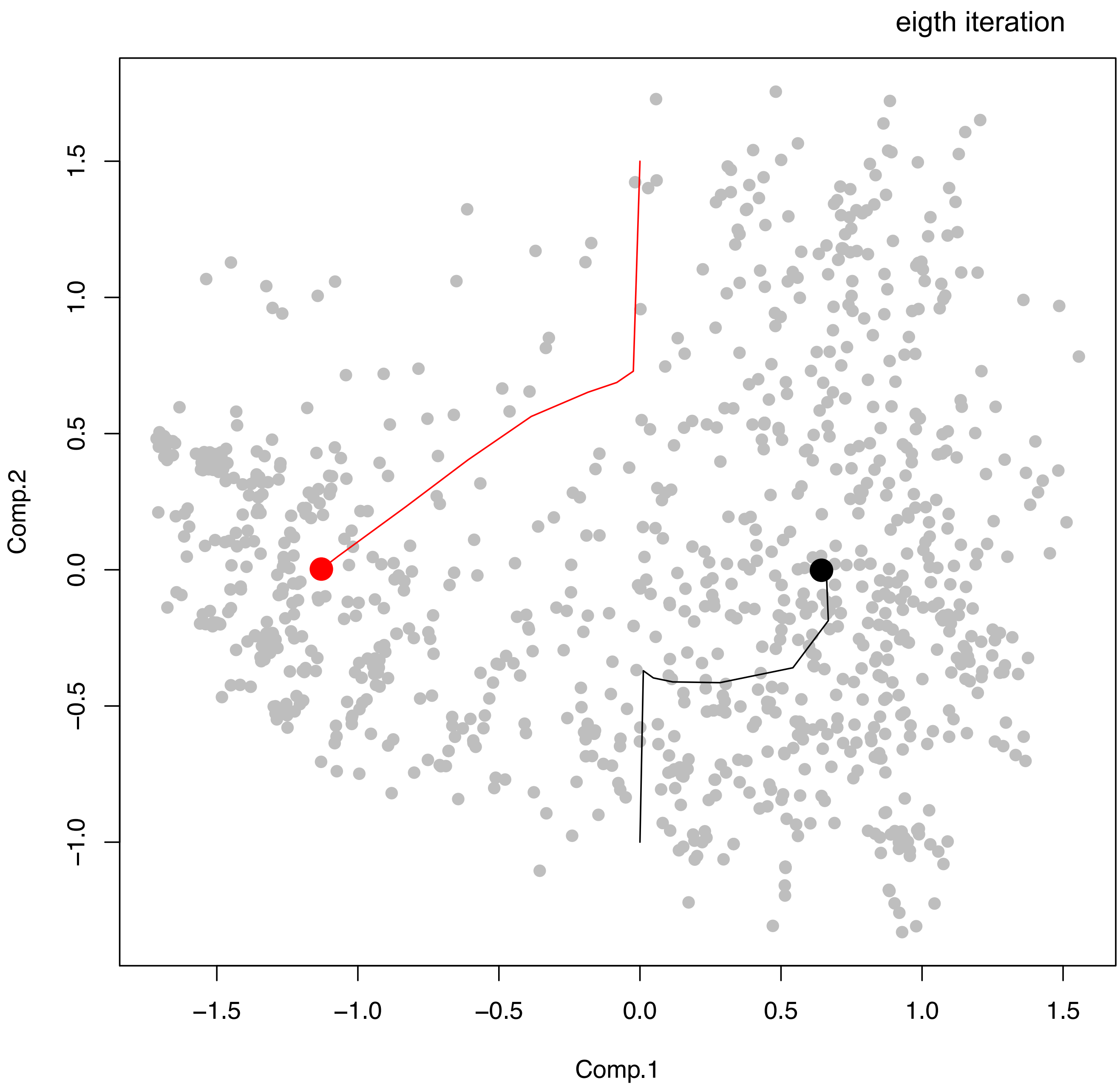
fifth iteration
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

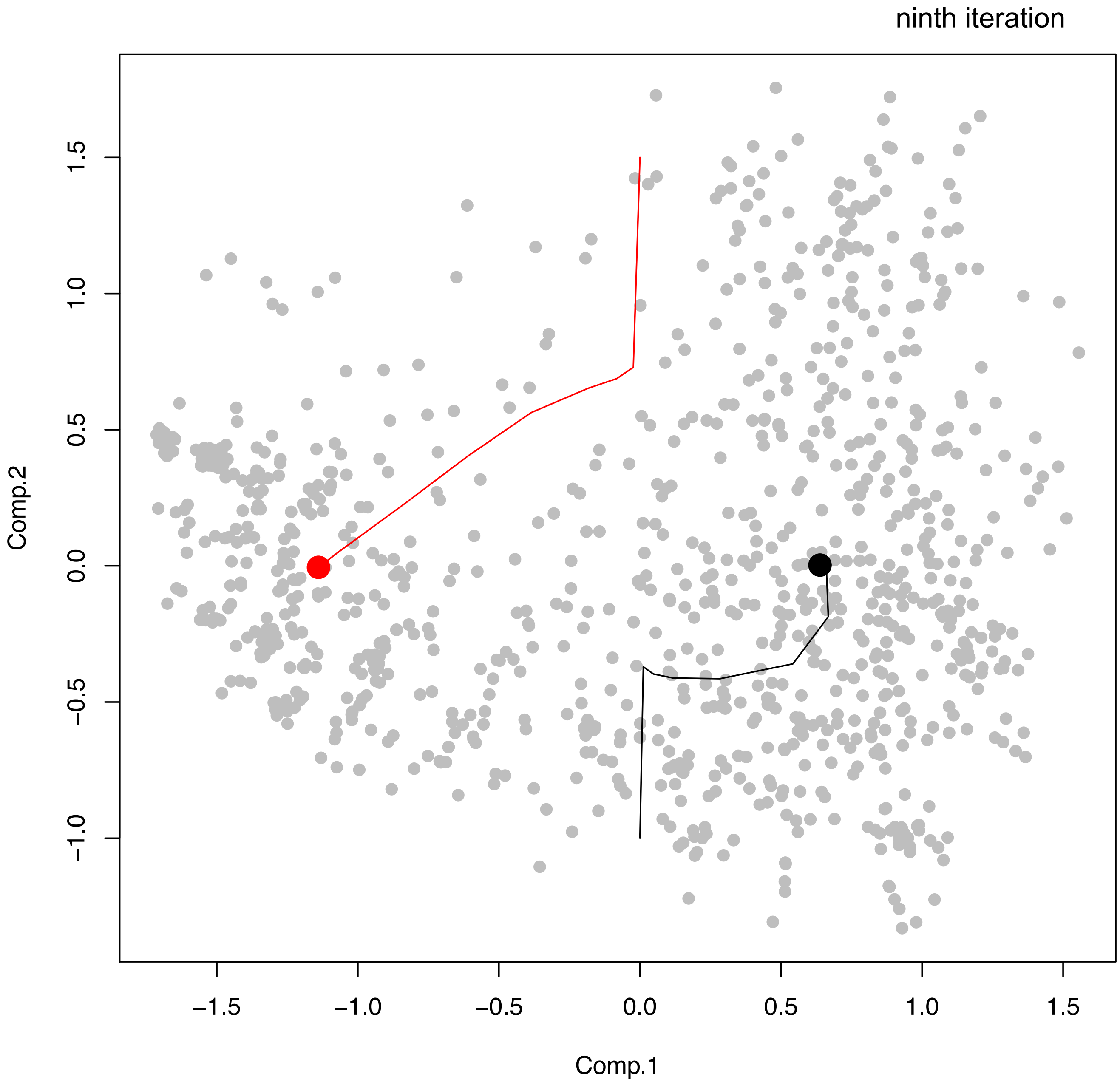

fifth iteration
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

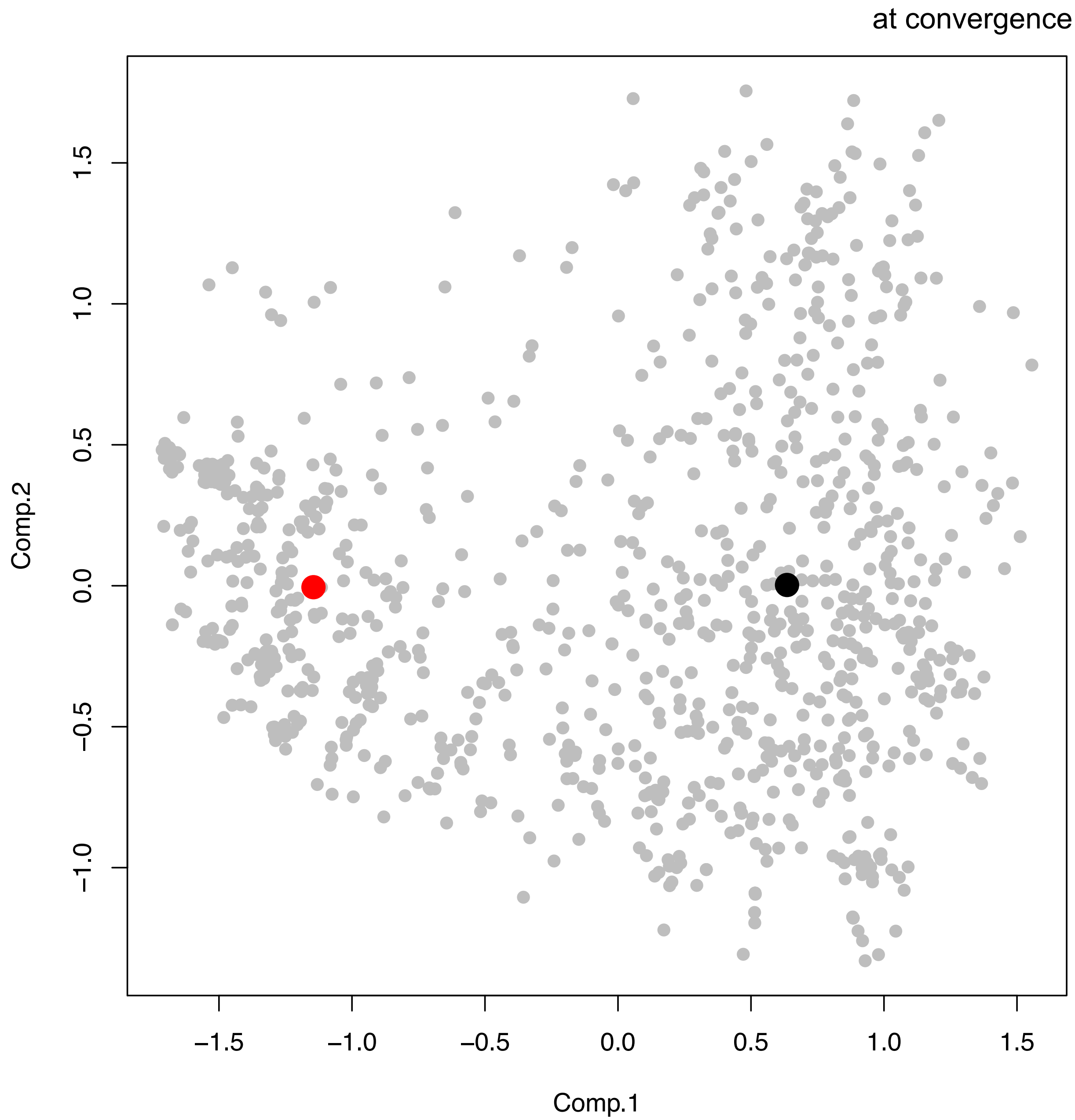


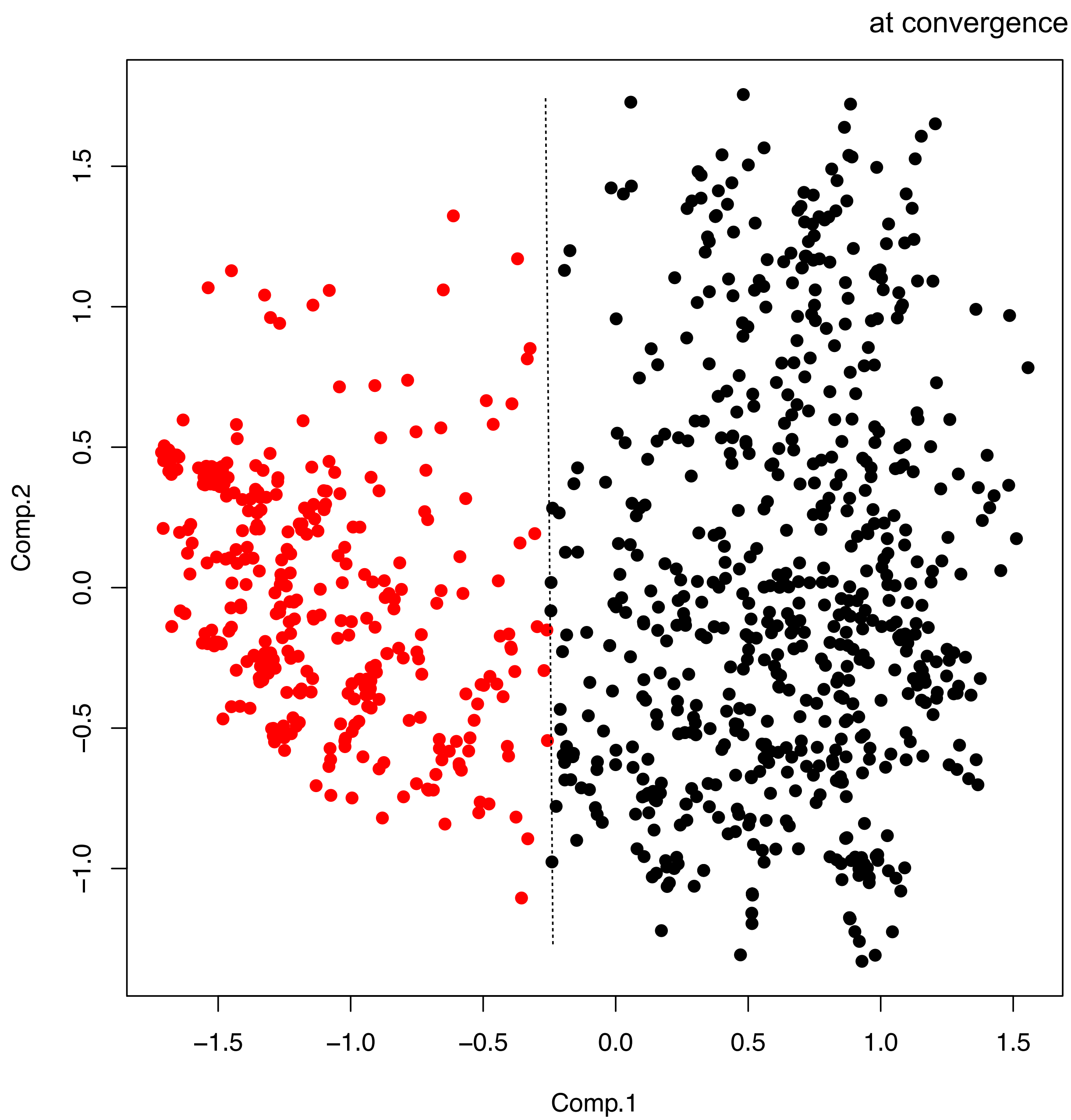
sixth iteration
Comp.2
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


seventh iteration
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


eigth iteration
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


ninth iteration
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

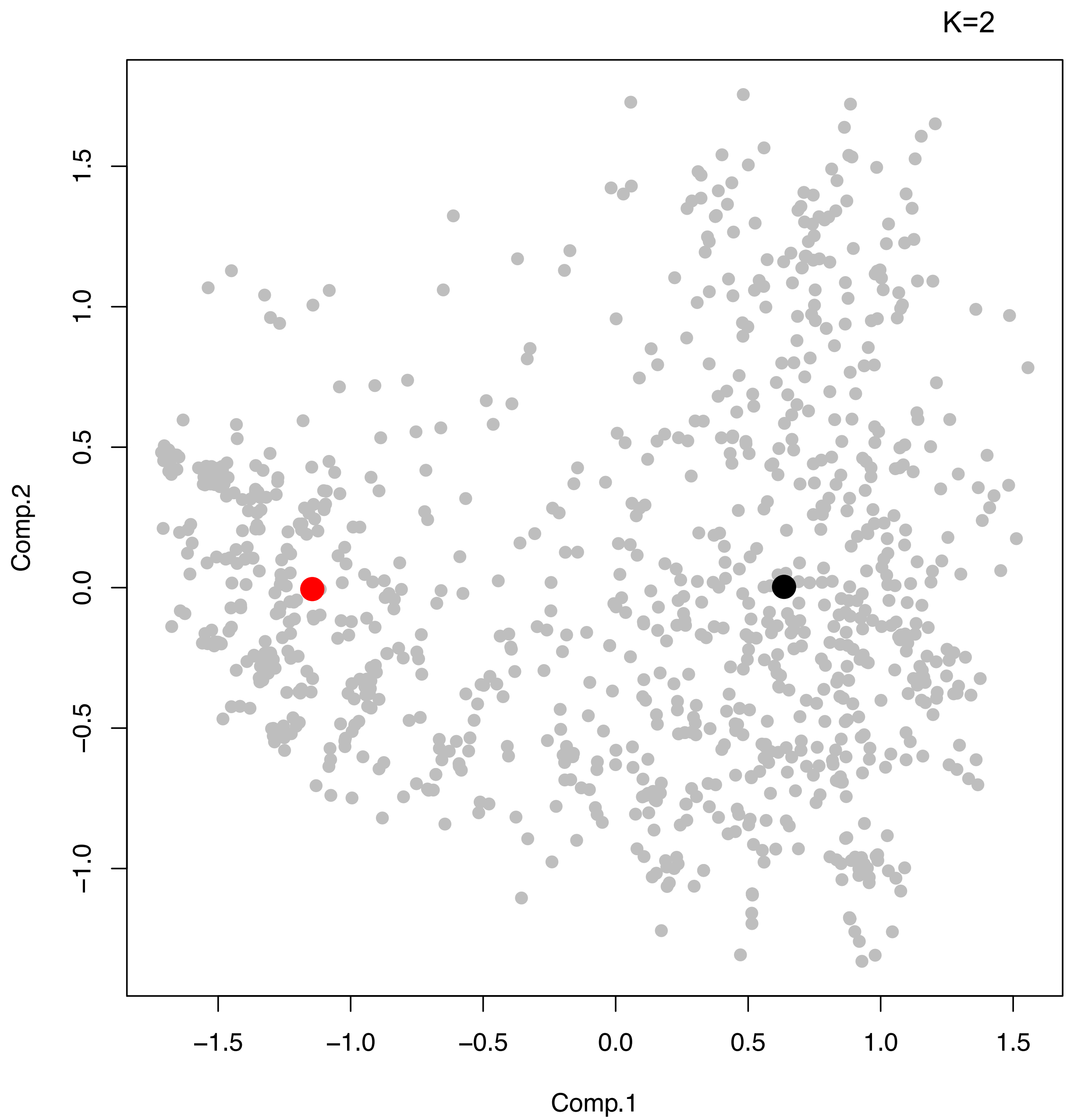

at convergence
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

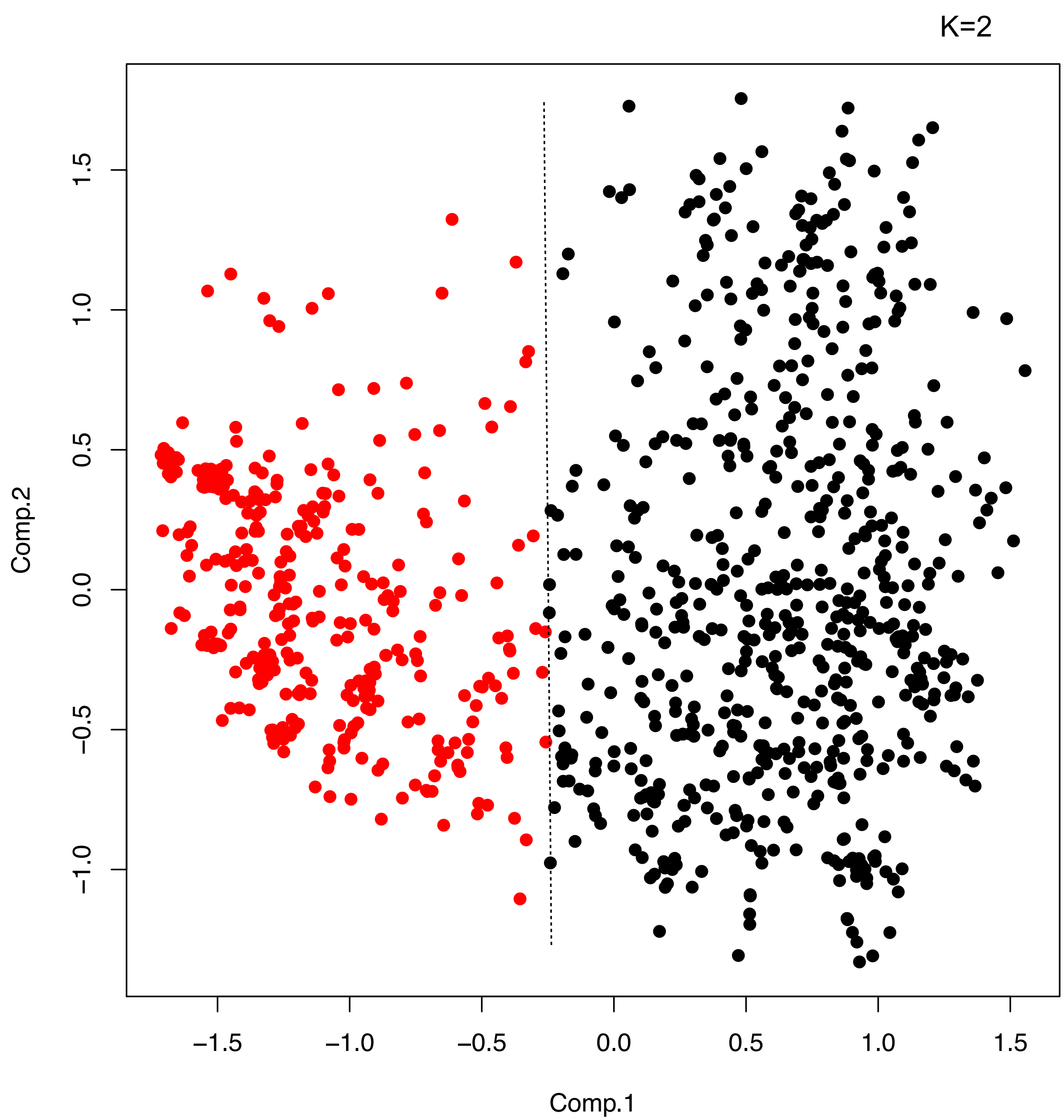

at convergence
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

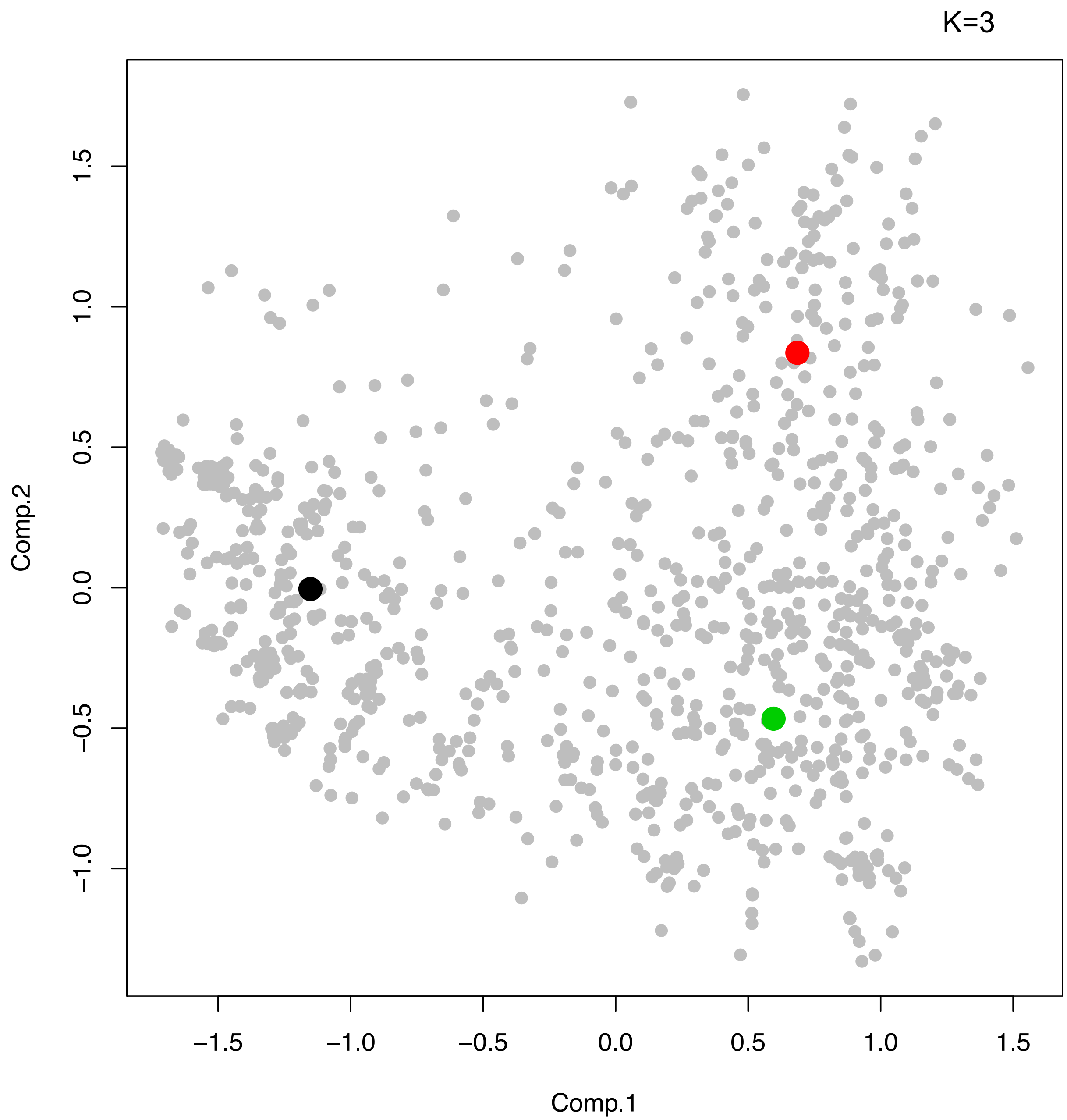
K-means

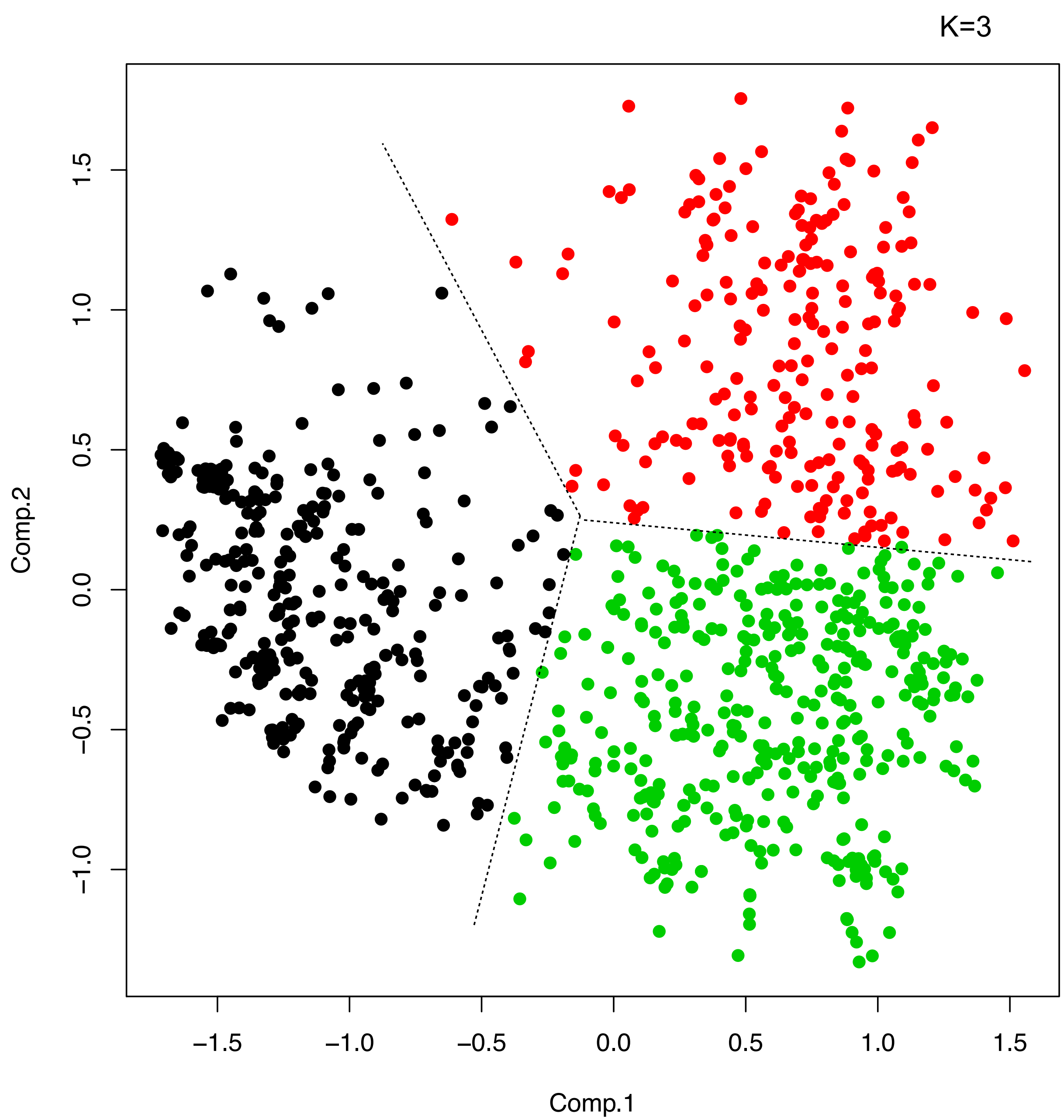
Through this simple iterative scheme, we construct a so-called **“hard” clustering**, in that each data point is associated with a single cluster (here K=2) -- There are **“soft” clustering schemes** that assign weights or probabilities that each point belongs to the different clusters

Here is how the algorithm behaves as we increase from K=2 to K=3, 4, 5...


K=2
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


K=2
Comp.2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Comp.1

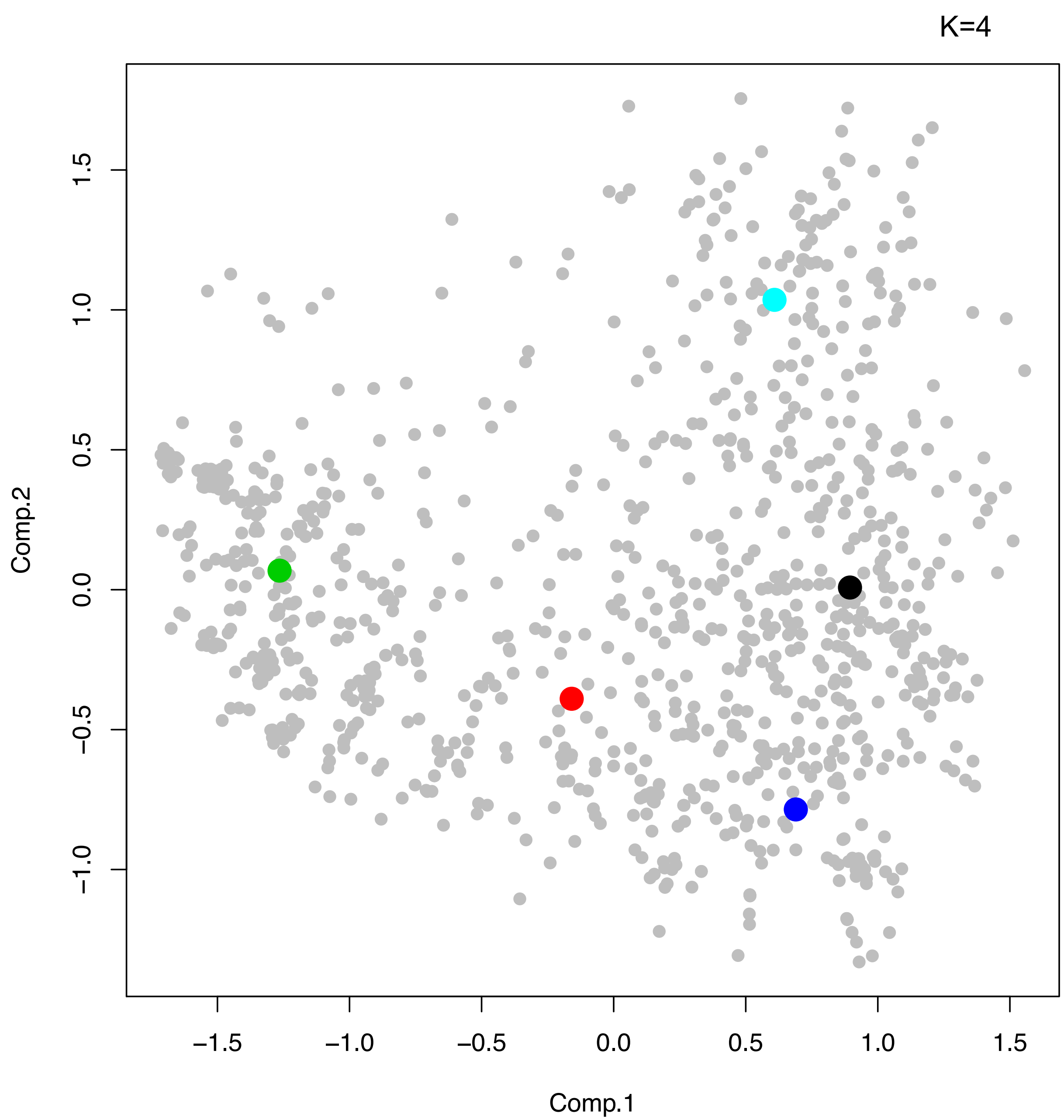

K=3
Comp.2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Comp.1

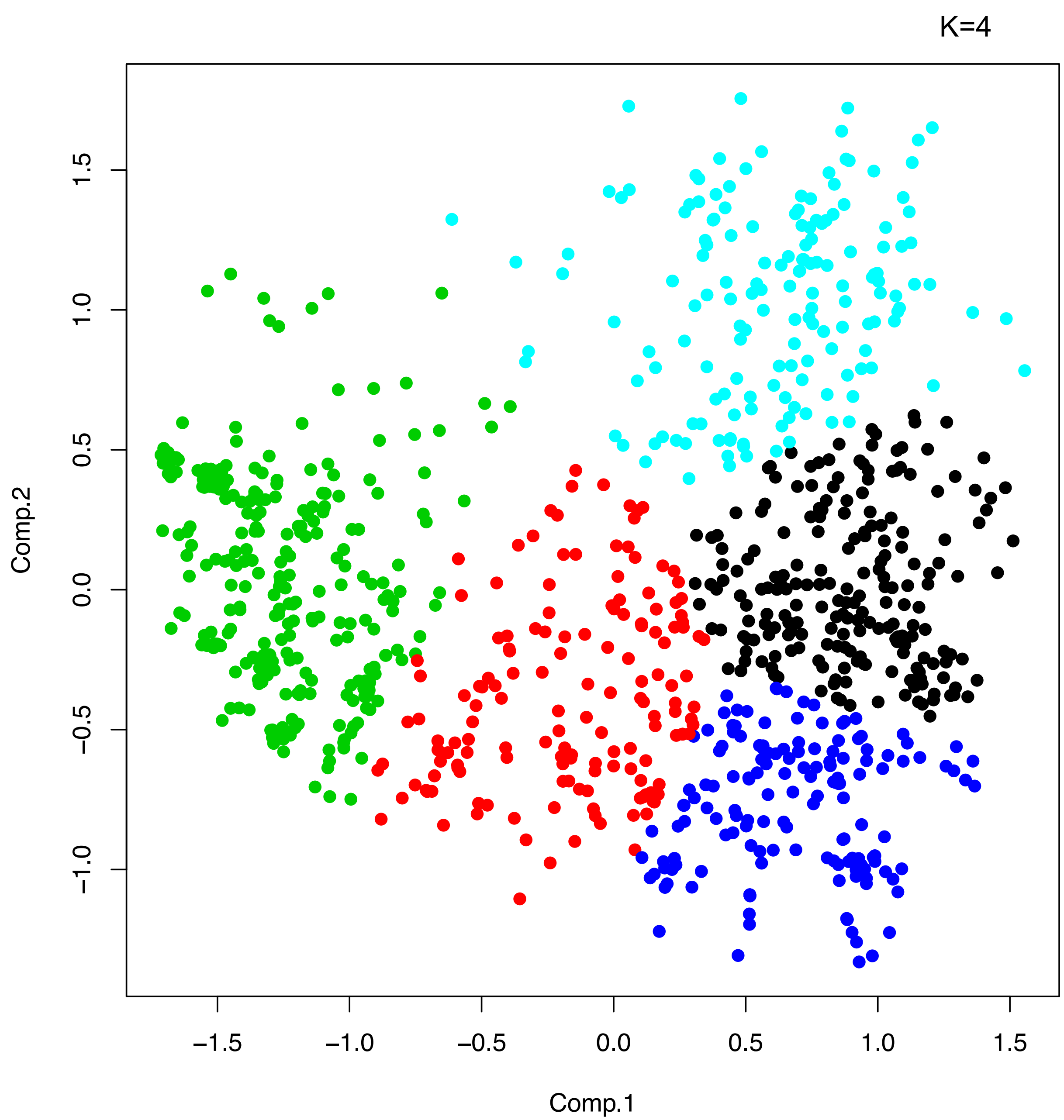

K=3
Comp.2
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Clustering

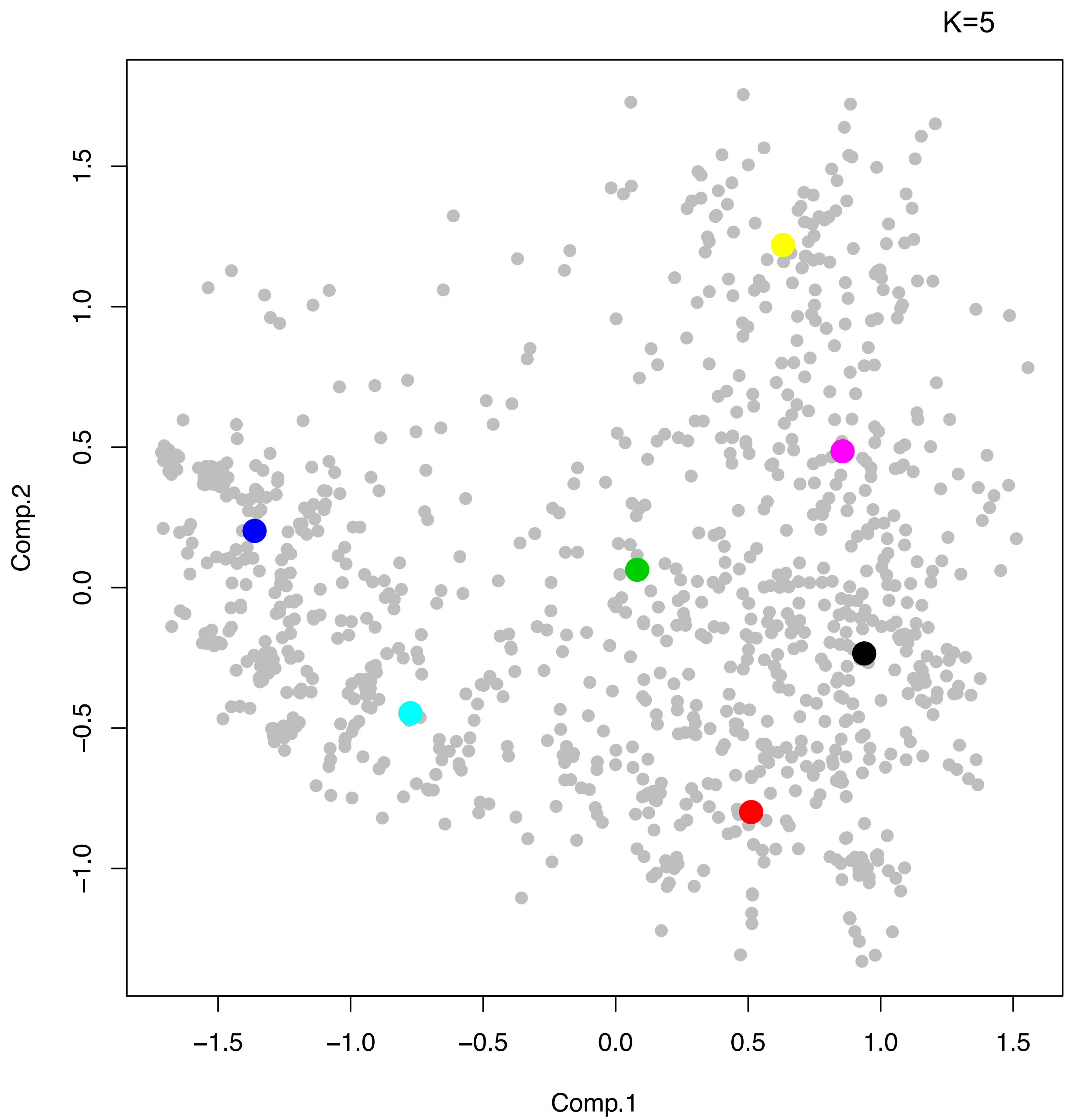
On the previous two slides, we asked for three clusters and displayed the three cluster centers (the three means) and indicated cluster membership

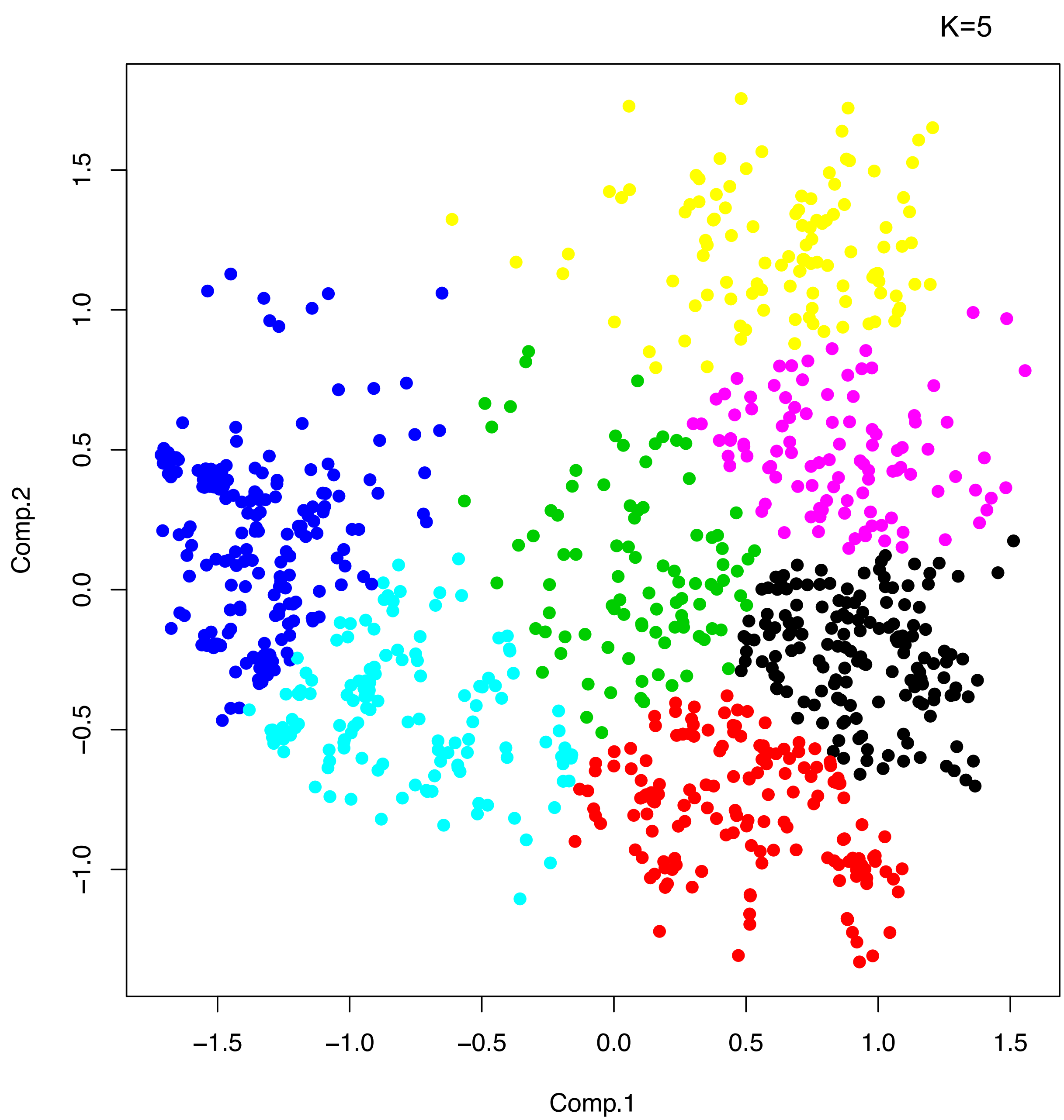
Note that the algorithm assigns points according to the nearest group mean and so in the end we have divisions based on the **Voronoi tessellation of these center points**


K=4
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5



K=4
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


K=5
Comp.2
Comp.1
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5



K=5
Comp.2
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

K-means and vector quantization

Vector quantization (VQ) is a lossy data compression method that builds a “block code” for a source -- Each point in our (in this case 2-d) space is represented by the nearest codeword

Historically this was a hard problem because it involved a lot of multi-dimensional integrals -- In the 1980s a VQ algorithm was proposed based on a training set of source vectors $X_1, \ldots, X_n$

In short, the goal was to design a codebook $\mu_1, \ldots, \mu_K$ and a partition $G_1, \ldots, G_K$ to represent the training set so that the overall distorition measure

$$V = \sum_{k=1}^{K} \sum_{X_i \in G_k} \|X_i - \mu_k\|^2$$

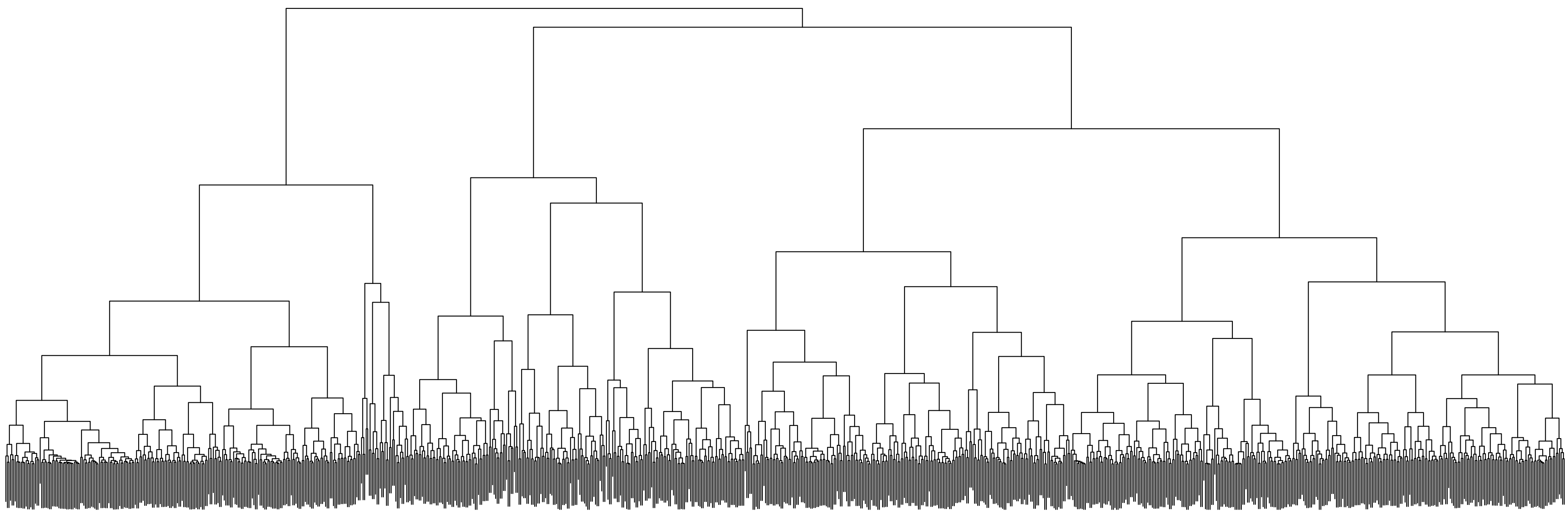
is a s small as possible

A second approach

A competitor to K-means is also best described algorithmically -- It builds groups **from the bottom up**, forming clusters of data points that are nearby in some notion of distance

The aptly named **hierarchical clustering methodology** organizes data into **a tree structure** in which close points appear in nearby leaves or branches -- Rather than belabor the algorithm, let’s have a look at the tree associated with our recipe data

Cluster Dendrogram

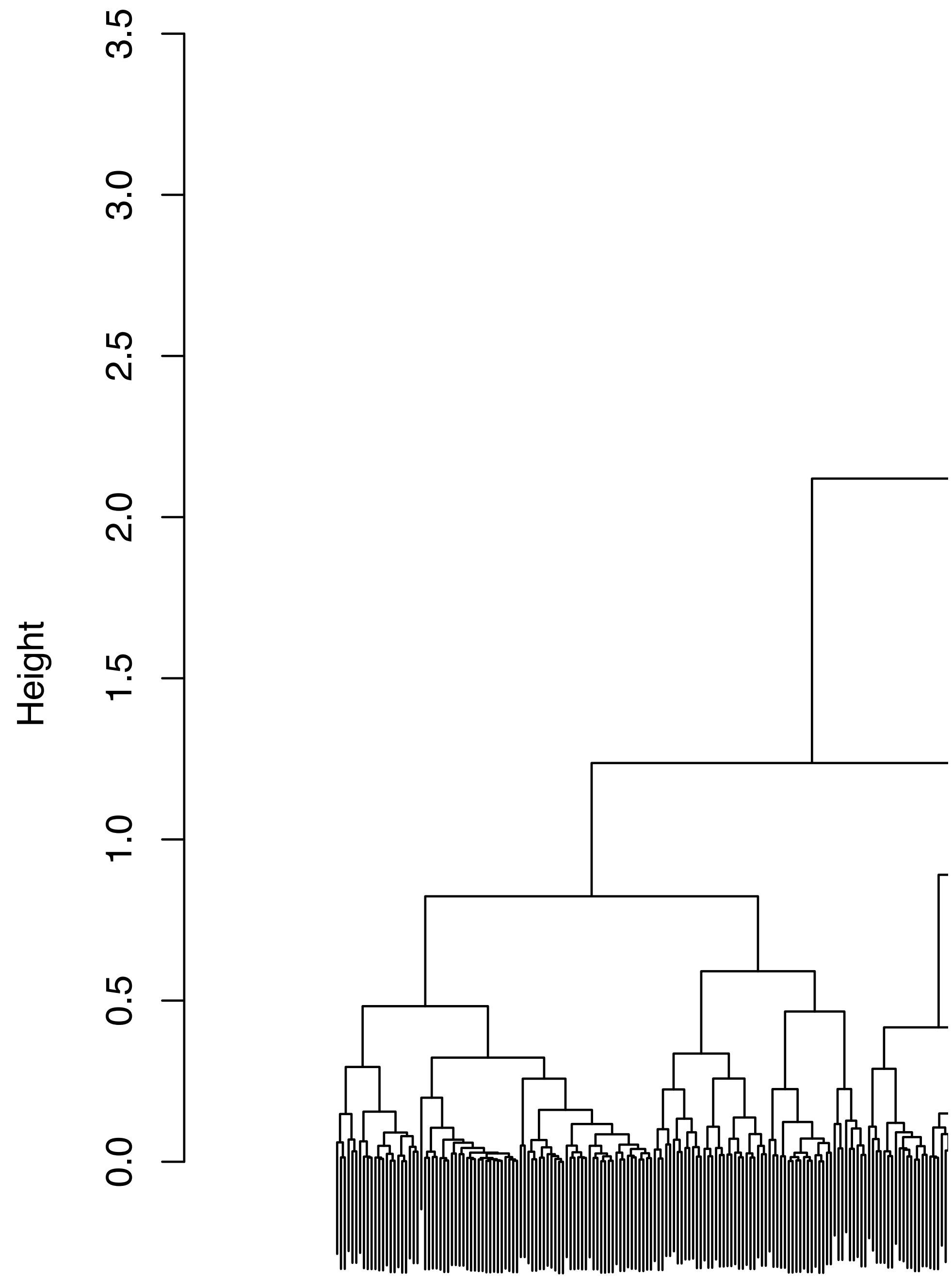


Hierarchical clustering

At the bottom of the tree (the leaves, assuming we grow trees upside down) are the 1,000 individual recipes

The algorithm starts by finding **the pair of recipes that are closest** (here using standard Euclidean distance) in the plot from a few slides back

These two points are **joined to form a branch** and then the pair is added to the other 998 recipes giving us 999 items to compare

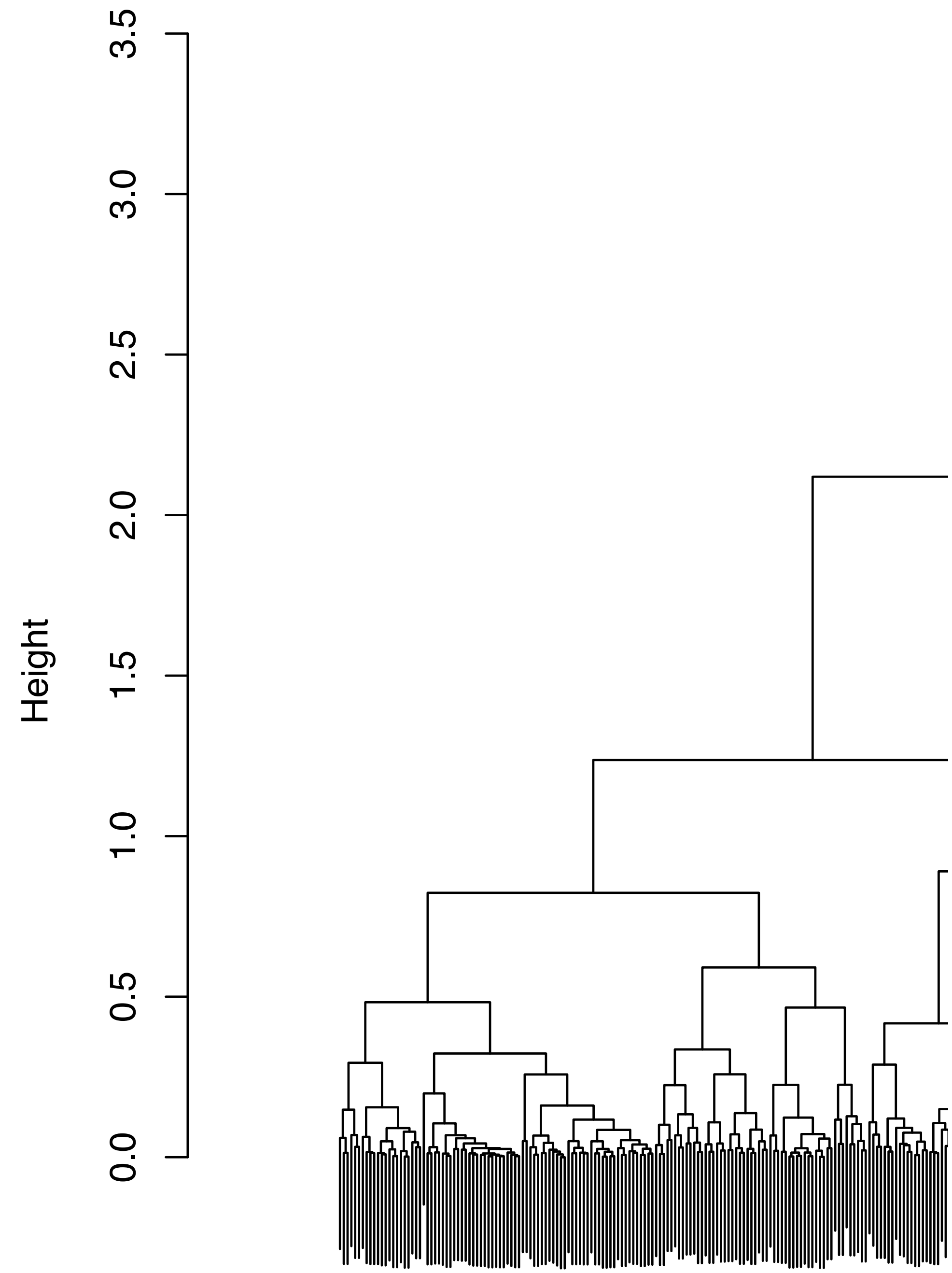


Hierarchical clustering

These two points are **joined to form a branch** and then the pair is added to the other 998 recipes giving us 999 items to compare

The algorithm then looks for the pair (either two points or a point and our branch) that has the smallest distance and joins them

To compute the distance between two branches or a branch and a point, we have choices -- One strategy is to assign their distance to be the **maximum pairwise distance** between points from the two branches

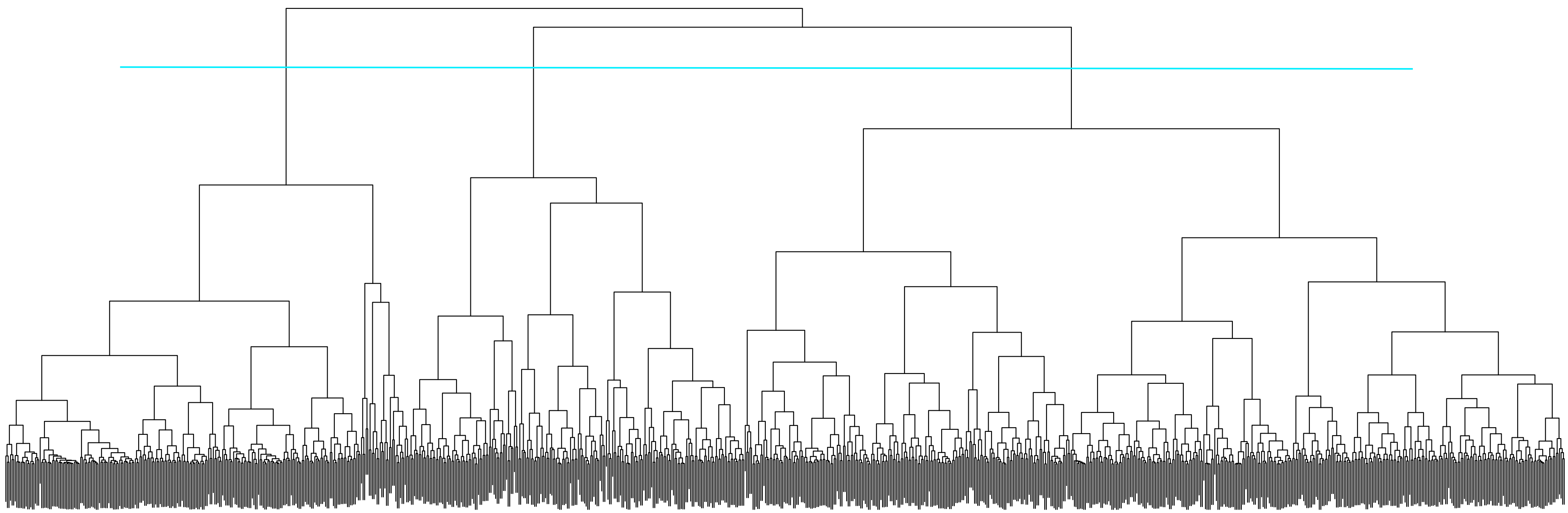


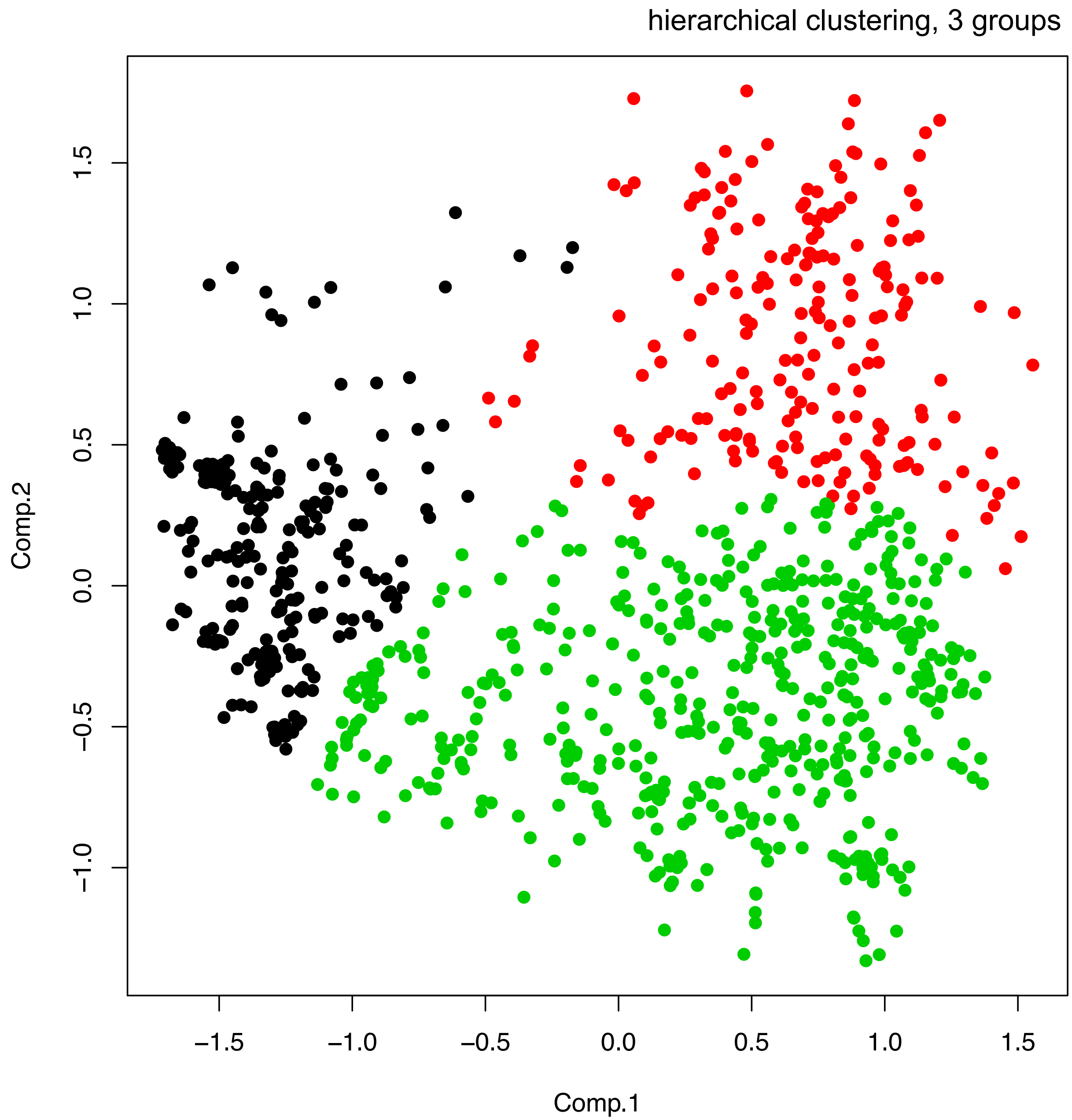
Hierarchical clustering

The height in the display refers to **the distance computed at the particular join** -- We can translate this tree structure into a clustering by “cutting” the tree either at a particular height or by asking for K groups

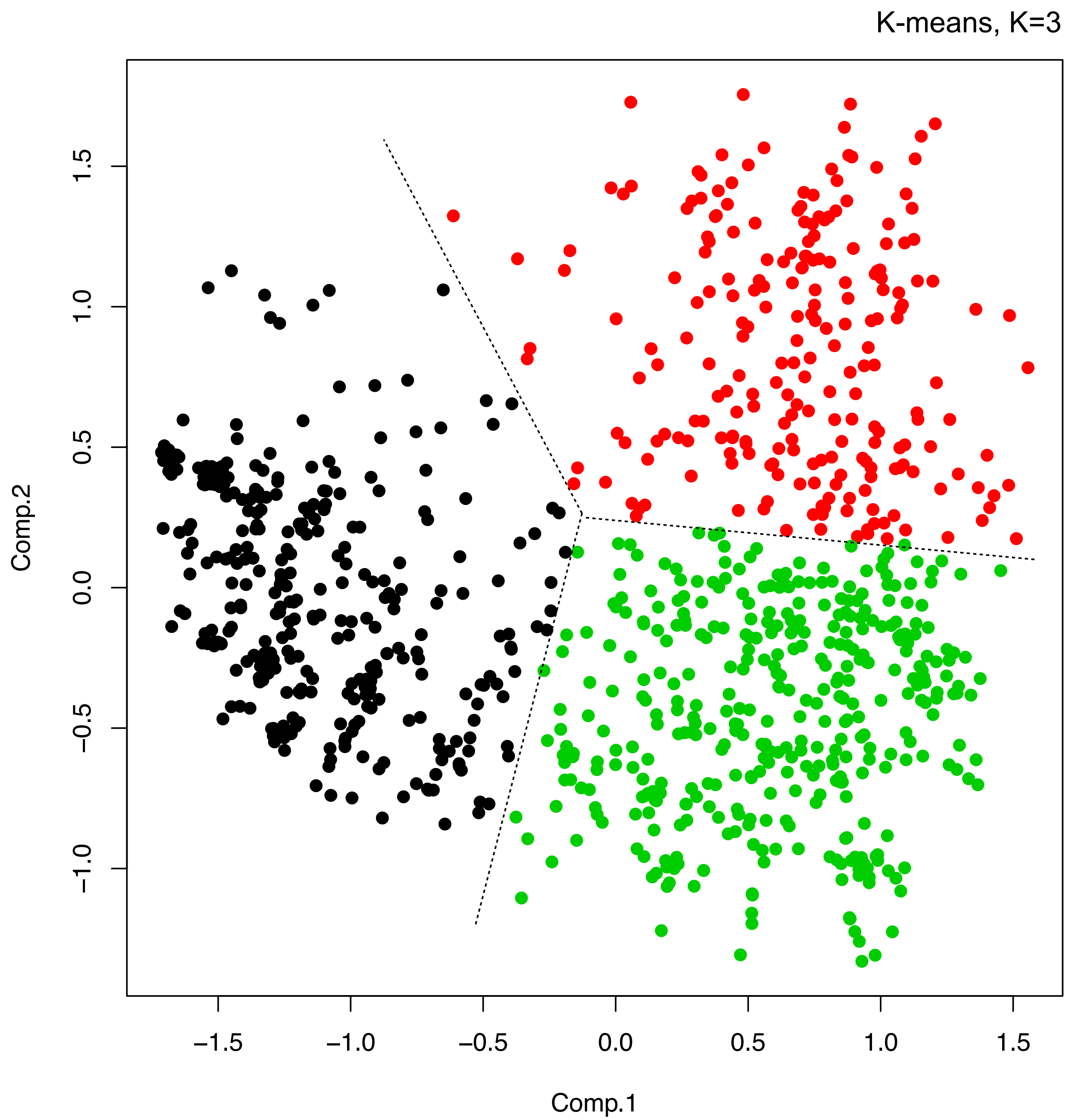
Here is what a 3-group clustering looks like on our data...

Cluster Dendrogram





hierarchical clustering, 3 groups
Comp.2
1.5
1.0
0.5
0.0
−0.5
−1.0
Comp.1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5



K-means, K=3
Comp.1
Comp.2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Another look: Hierarchical clustering

This approach starts with a distance matrix, an n-by-n matrix having as it's i,j element the distance between X_i and X_j -- Any distance measure can be used, and certain choices make more sense for certain data types

Note that in addition to specifying a grouping, the tree creates a linear ordering for the data that people have used to great effect...