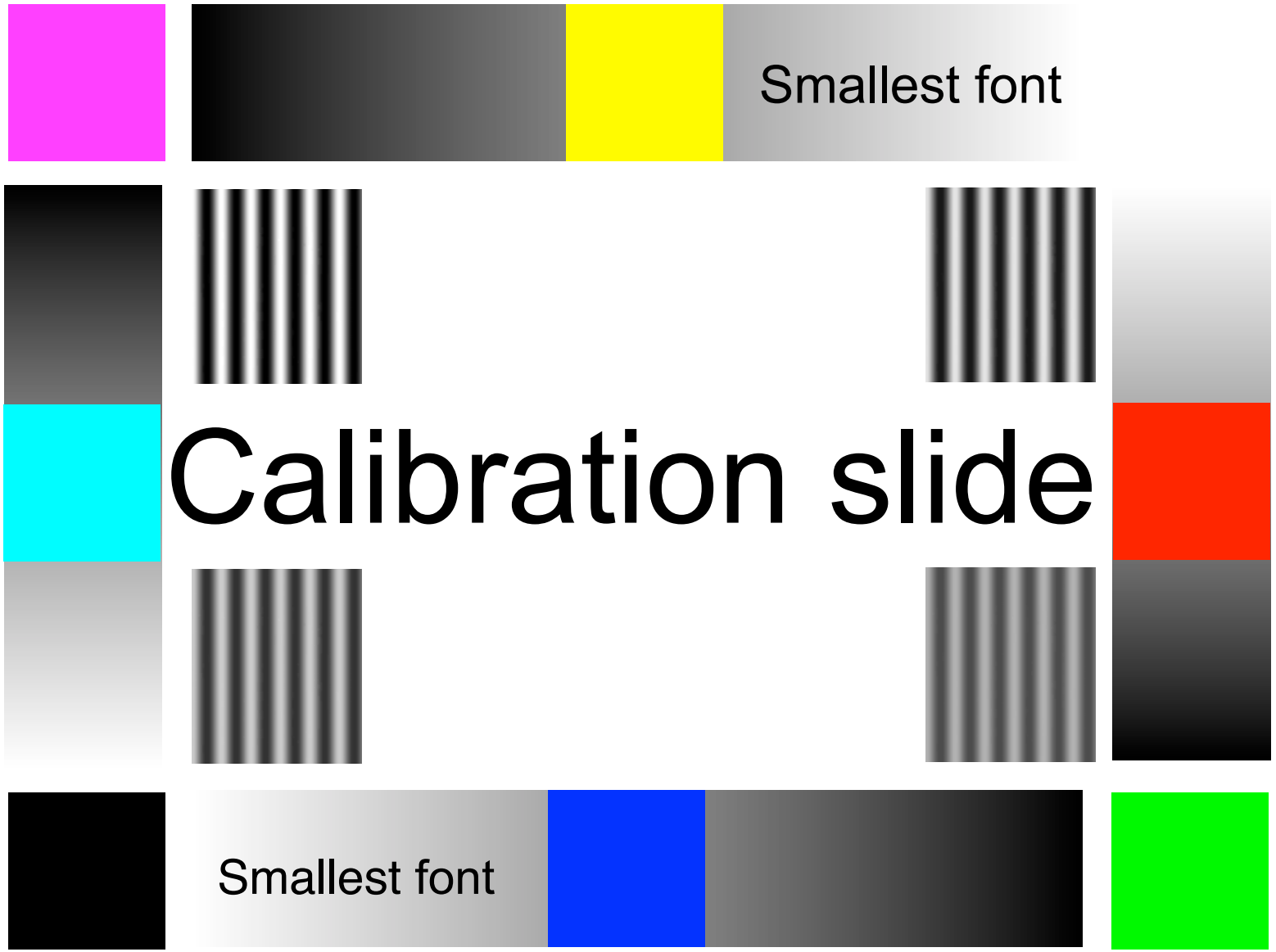

Smallest font
Calibration slide
Smallest font

Scientific Programming and Computing for the Behavioral Sciences

Data in low dimensional space: Dimension reduction
Factor analysis
Principal components analysis

Previously: Data analysis programs modeled after information encoding (perceptual areas)

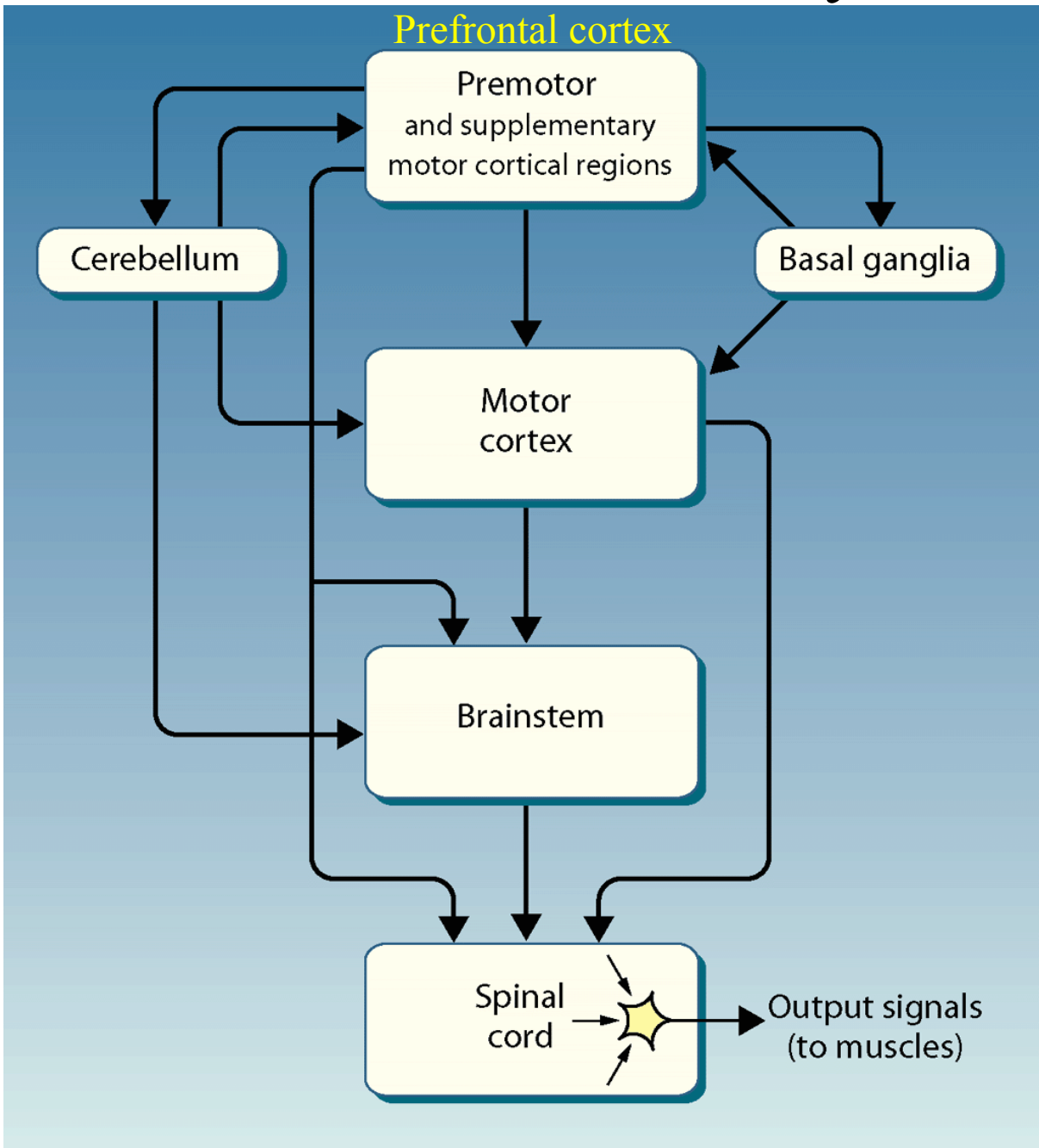
- Now: Introduce the concept of modeling programming in general as an implementation of the design principles of the primate motor system.
- After all, you are telling a computer what to do.

Neural basis of motor control hierarchy

Goal (movement selection & plan)

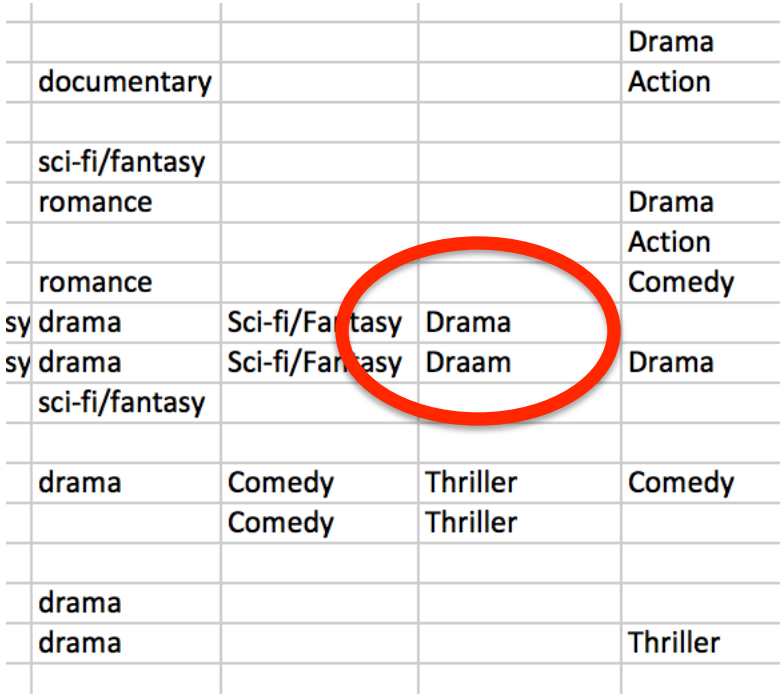
Tactics (spatio-temporal pattern of joint angles and muscle activations)

Execution



An example from scientific programming

- 1) Goal: Remove misspellings from participant input



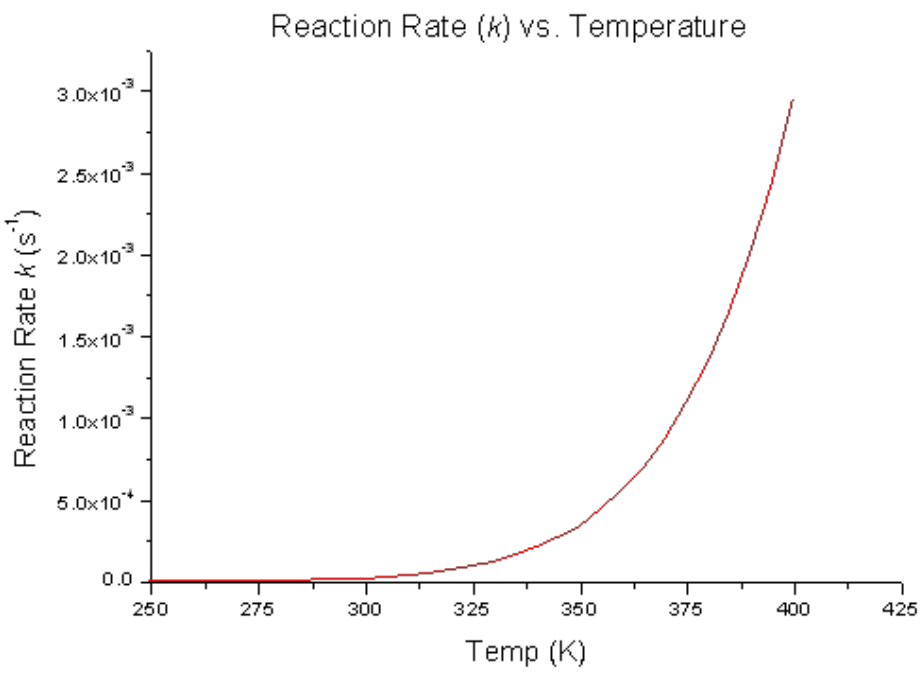
				Drama
	documentary			Action
	sci-fi/fantasy			
	romance			Drama
				Action
	romance			Comedy
sy	drama	Sci-fi/Fantasy	Drama	
sy	drama	Sci-fi/Fantasy	Draam	Drama
	sci-fi/fantasy			
	drama	Comedy	Thriller	Comedy
		Comedy	Thriller	
	drama			
	drama			Thriller

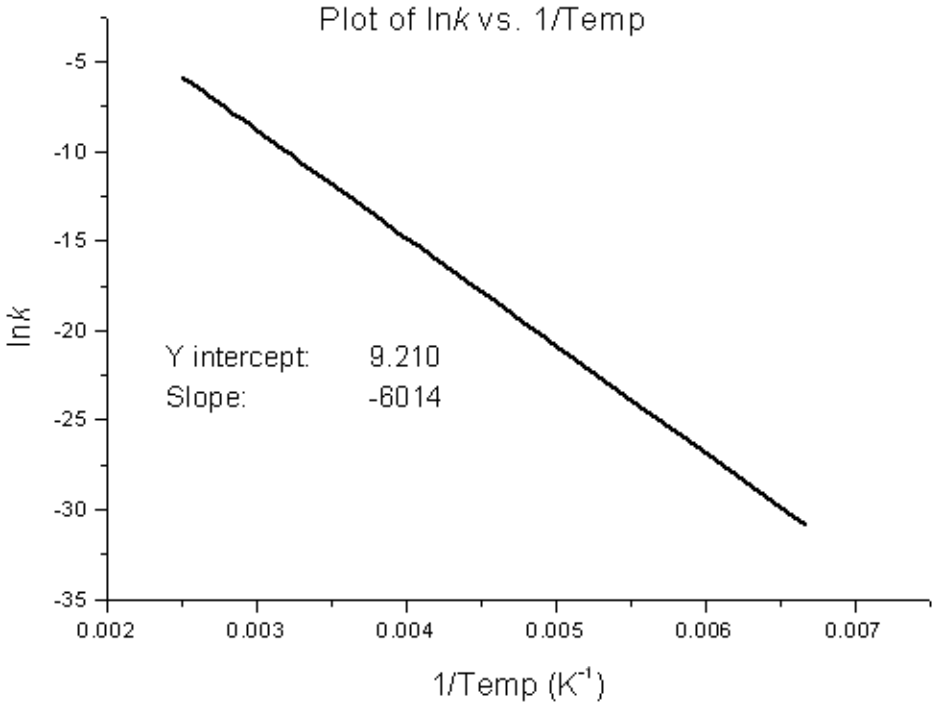
- 2) Tactics: The algorithm of how to do this in principle (there are many ways or degrees of freedom)
- 3) Implementation: Actually writing the Matlab code that executes the algorithm.

Here

- Goal: Dimension reduction because the data is too highly dimensional.
- Tactics: Factor Analysis, Principal Component Analysis (PCA), Independent Component Analysis (ICA)…
- Implementation: Writing the actual Matlab code.
- Suggestion: Use this for comments hierarchy (big goal, pseudocode, explaining in line variables)
- Why do we need to reduce the dimensionality of the data in the first place?

“Simple” science





Psychology is not like that

What sets psychology and neuroscience apart

- It is trying to study the casual relationships of complex phenomena.
- So far, we have focused on the **factors** that determine the relationship, the independent variables.
- But this is also true for the phenomena itself. They probably can't be measured by a single **dependent** variable.
- Designs and methods involving multiple dependent variables are called **multivariate**.

The problem

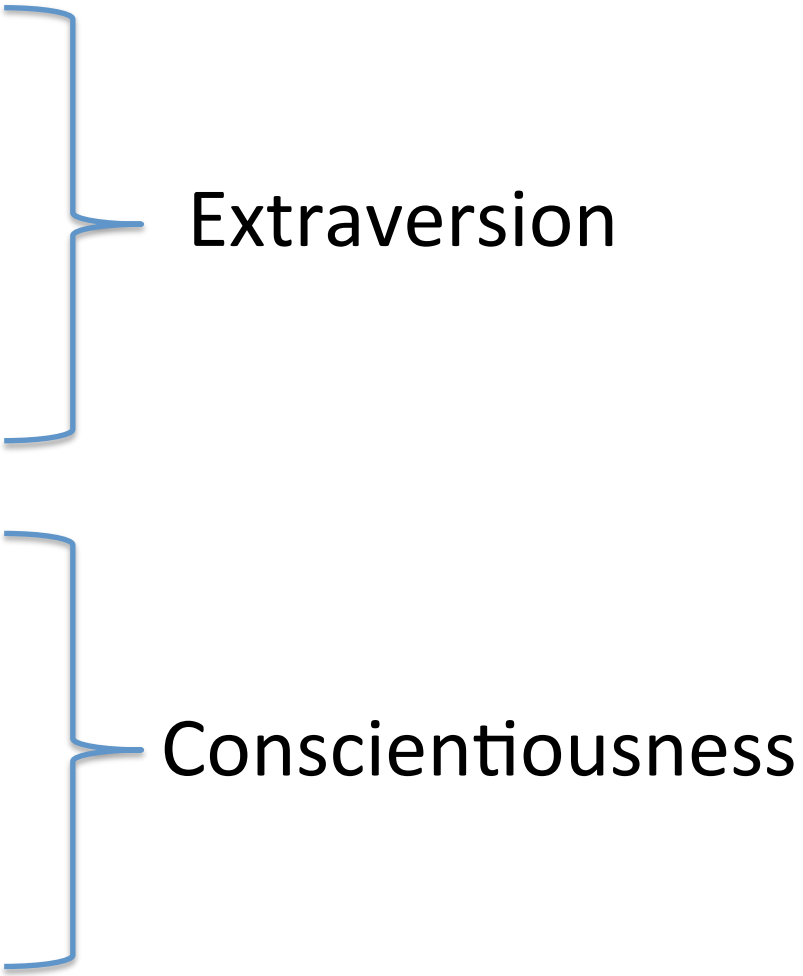
	1	2	3	4	5	6	7	8	9	10	11	12	13
1	8	6	8	7	8	2	1	8	4	1	3	5	7
2	5	1	7	6	1	6	4	4	1	2	8	4	4
3	9	1	8	9	7	6	2	4	3	5	8	9	2
4	3	8	2	3	9	5	8	4	7	8	1	3	8
5	7	1	8	5	3	1	9	2	4	6	2	4	2
6	6	9	5	2	9	2	4	5	5	2	3	4	1
7	3	2	1	8	8	3	4	5	1	7	8	7	9
8	1	4	5	7	9	6	3	5	6	7	9	5	9
9	8	3	4	8	5	3	9	4	7	4	2	1	3
10	1	5	7	8	5	6	1	2	2	9	2	6	3
11	4	4	6	7	8	8	4	9	8	1	1	4	6
12	8	8	4	6	1	3	1	8	6	3	5	2	5
13	8	1	2	4	5	5	1	2	3	6	6	3	7
14	1	8	9	6	2	3	1	2	6	8	6	2	7
15	4	9	7	1	9	6	3	3	2	4	6	6	6

The name of the game

- **Multivariate** data is usually too complex to understand by just looking at it.
- Dimension reduction methods like principal component analysis allow to extract underlying **factors** that account for the data.
- Then visualize them so you can look at it.
- This is called "**factor analysis**".
- Very popular in psychology and everywhere where one has multivariate data.

Example: Personality psychology

- I make friends easily
- I avoid crowds
- I prefer to be alone
- I love large parties
- I try to lead others
- I complete tasks successfully
- I like order
- I try to follow the rules
- I go straight for the goal
- I get chores done right away



Example: Personnel selection

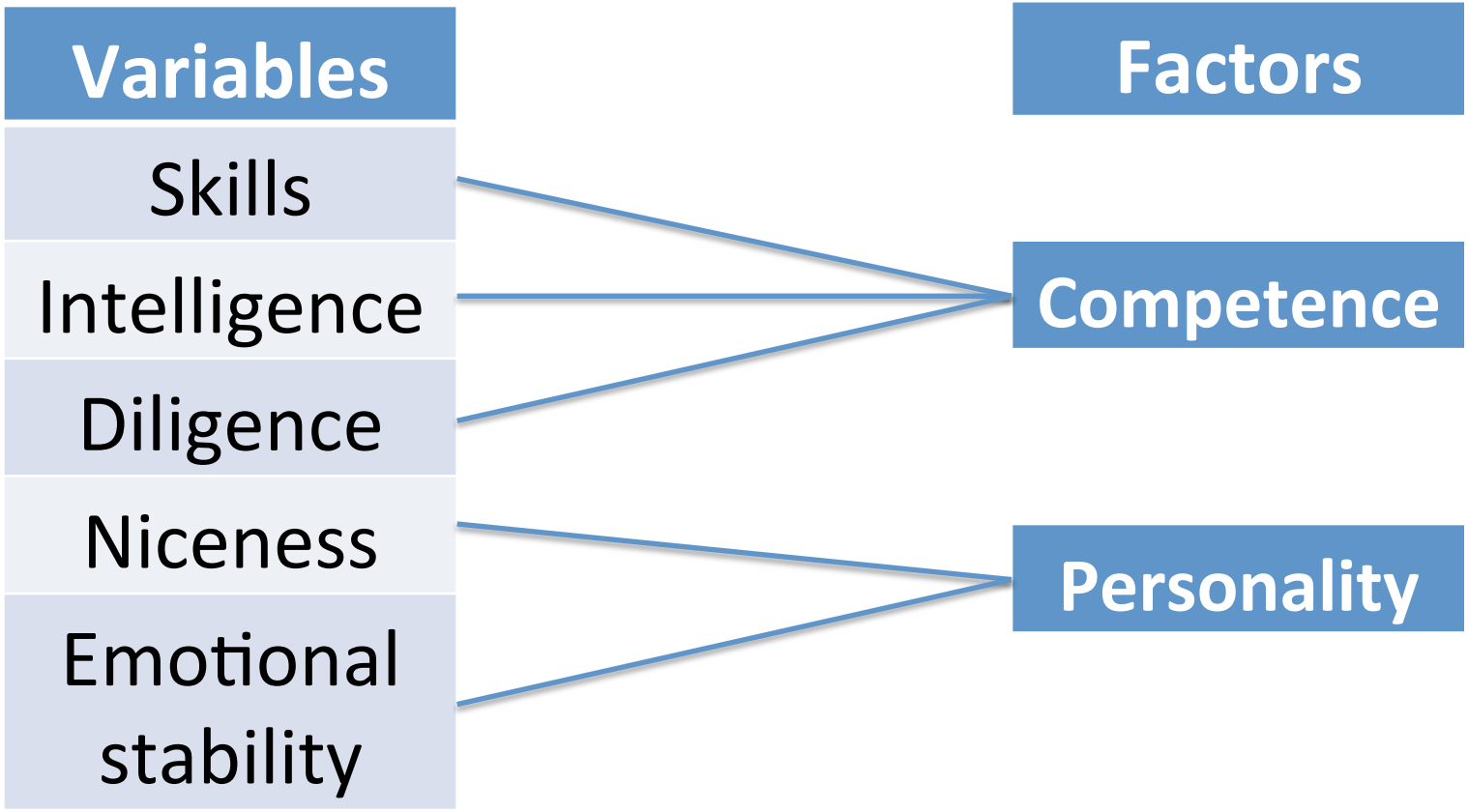
- Your job in the HR department of a major corporation is to conduct candidate interviews for personnel selection.
- After the interview, you rate each candidate in terms of these variables: Skills, Intelligence, Diligence, Niceness, Emotional stability.
- You then do a factor analysis to determine whom to hire.

Ratings of a typical day

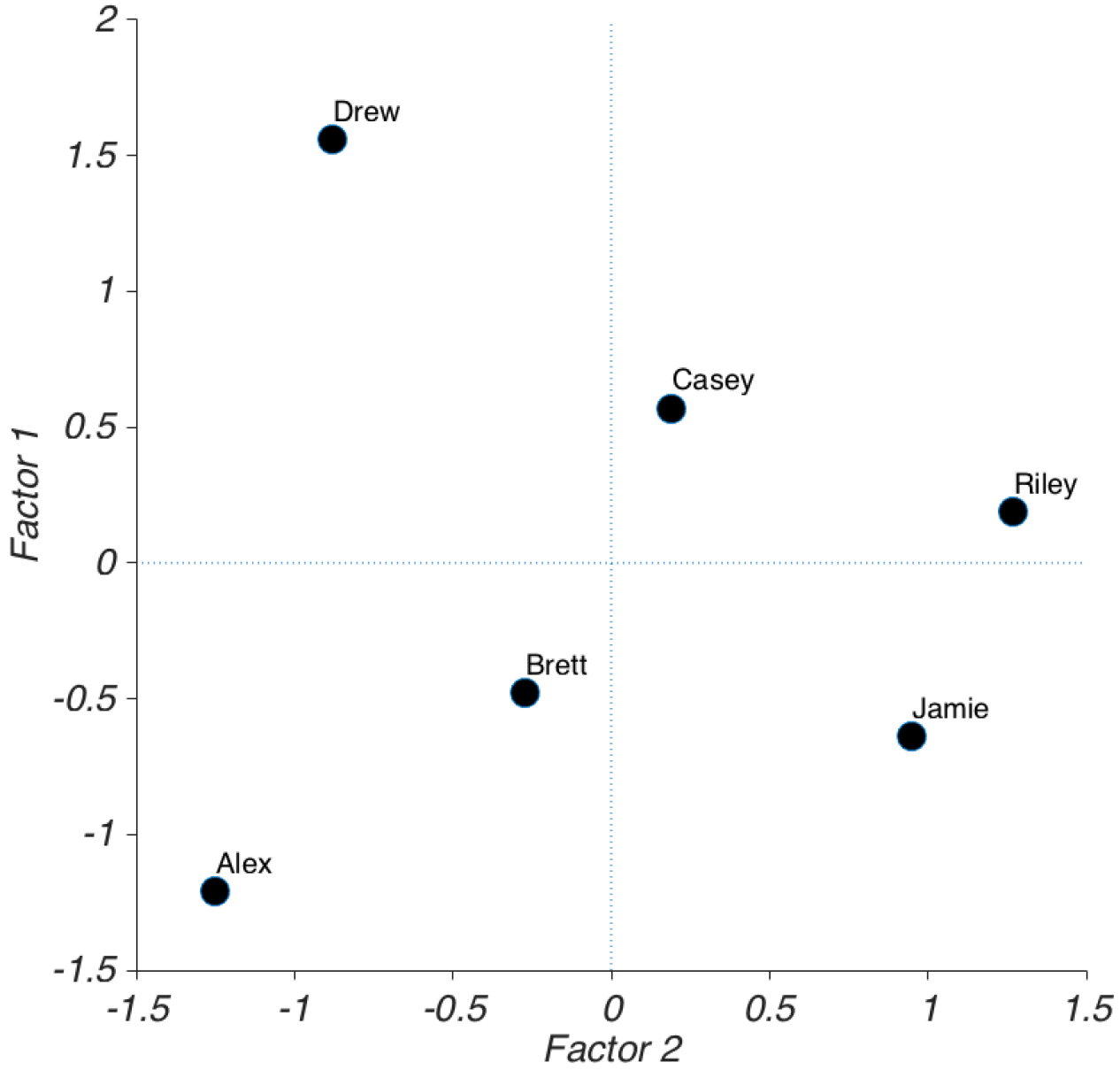
Candidate	Skills	Intelligence	Diligence	Niceness	Emotional stability
Alex	1	1	2	1	2
Brett	2	6	3	3	4
Casey	5	6	5	5	5
Drew	5	6	6	2	3
Jamie	2	3	3	5	7
Riley	3	4	4	6	7

We need to visualize this to make a decision

- We can't visualize 5 variables at once.
- We can visualize 2-3.
- So dimension reduction is critical:



The result


Drew
Casey
Riley
Brett
Jamie
Alex
2
1.5
1
0.5
0
-0.5
-1
-1.5
Factor 1
-1.5
-1
-0.5
0
0.5
1
1.5
Factor 2

How do we get there?

1) Picking variables and calculating correlations

2) Extracting factors

3) Determine the number of factors

4) Interpret their meaning

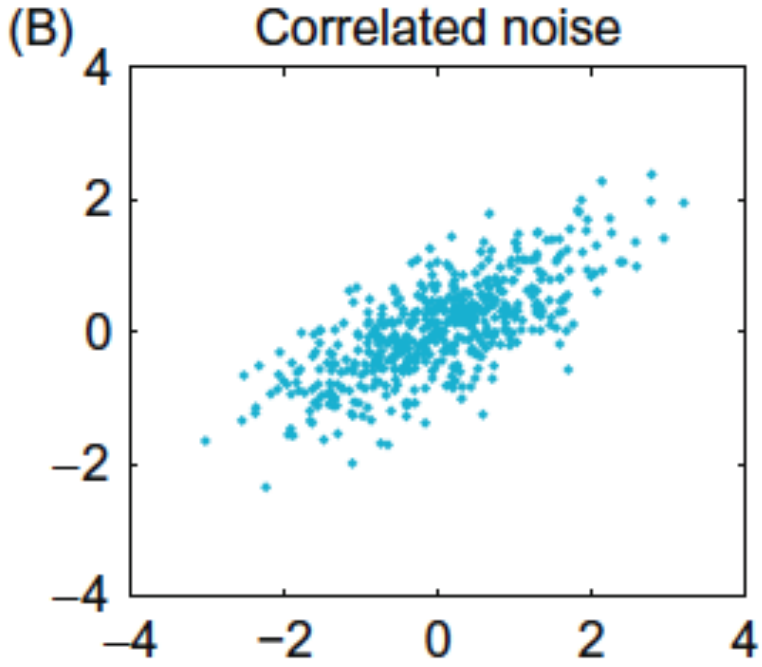
5) Determine factor values

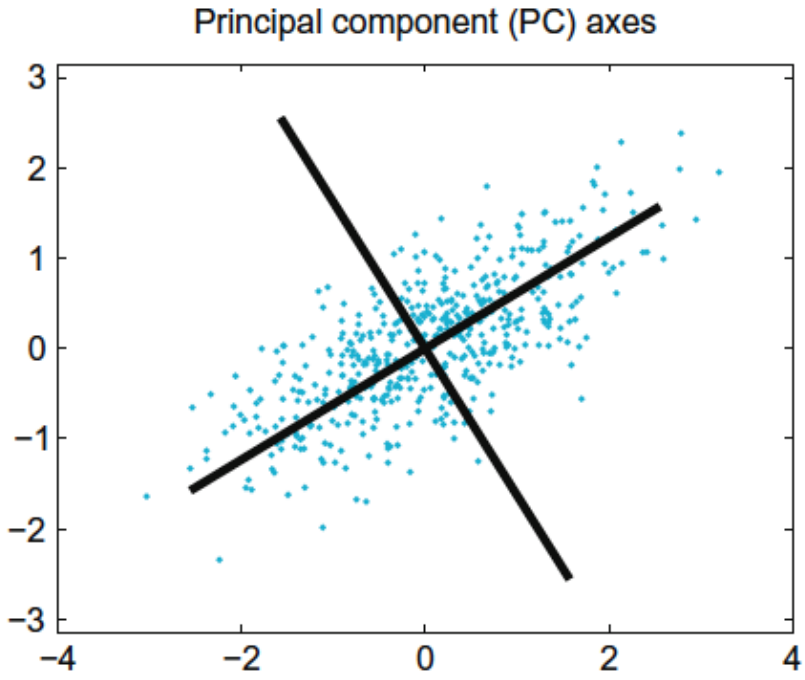
1) The correlation matrix

	Skills	Intelligence	Diligence	Niceness	Emotional Stability
Skills	1				
Intelligence	0.71	1			
Diligence	0.96	0.70	1		
Niceness	0.11	0.14	0.08	1	
Emotional stability	0.04	0.06	0.02	0.98	1

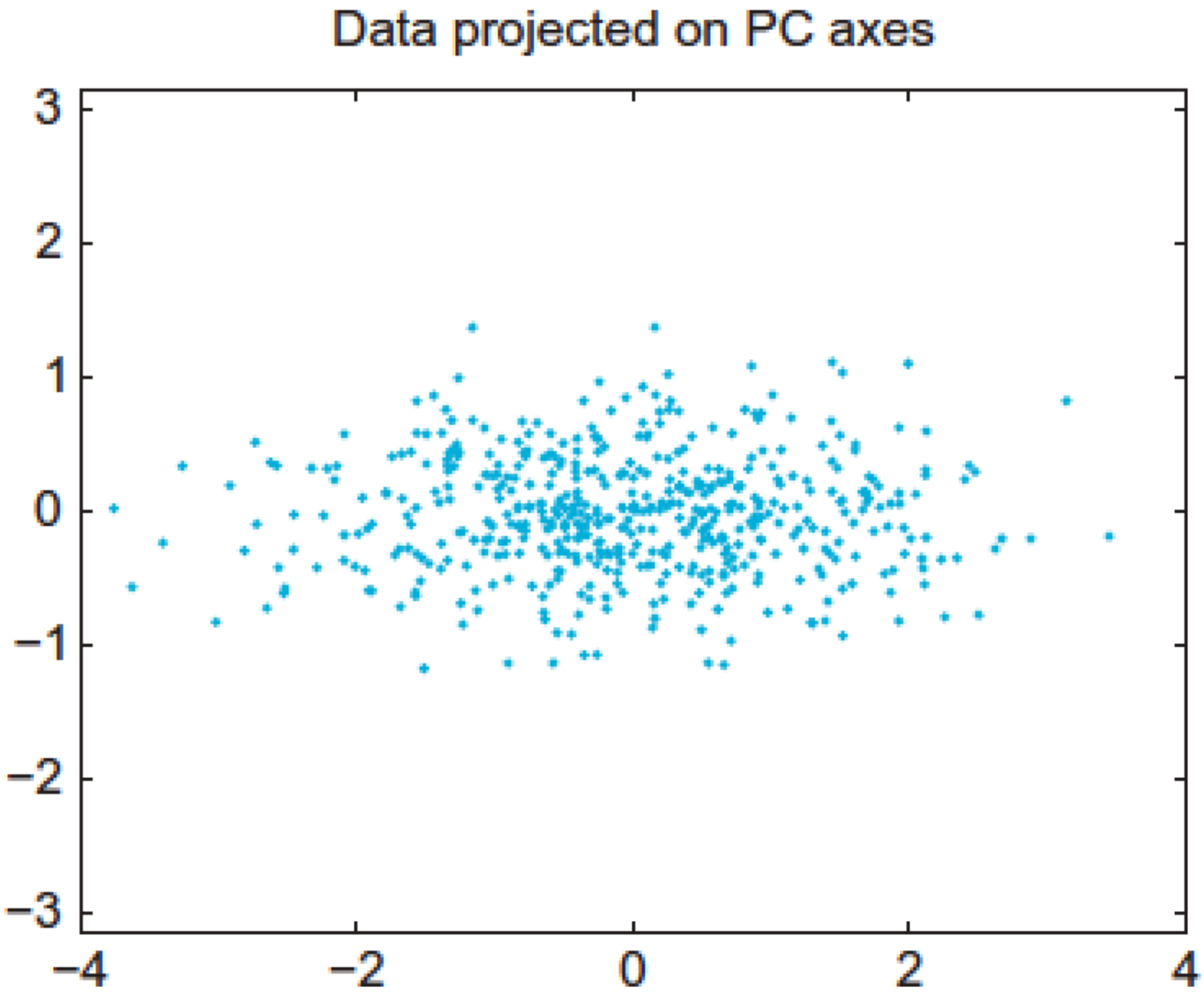
2) Factor extraction as an axis rotation

- The data live in some coordinate frame.
- Idea: Introduce a new coordinate system in the directions in which the data vary the most.
- Then recalculate the data values in terms of *those* axes.





The result?



Numerical factor extraction: The fundamental theorem

- The idea: The original data in terms of their variables can be expressed as a linear combination of (hypothetical) factors.

$$x_{kj} = a_{j1} \bullet p_{k1} + a_{j2} \bullet p_{k2} + a_{jQ} \bullet p_{kQ}$$

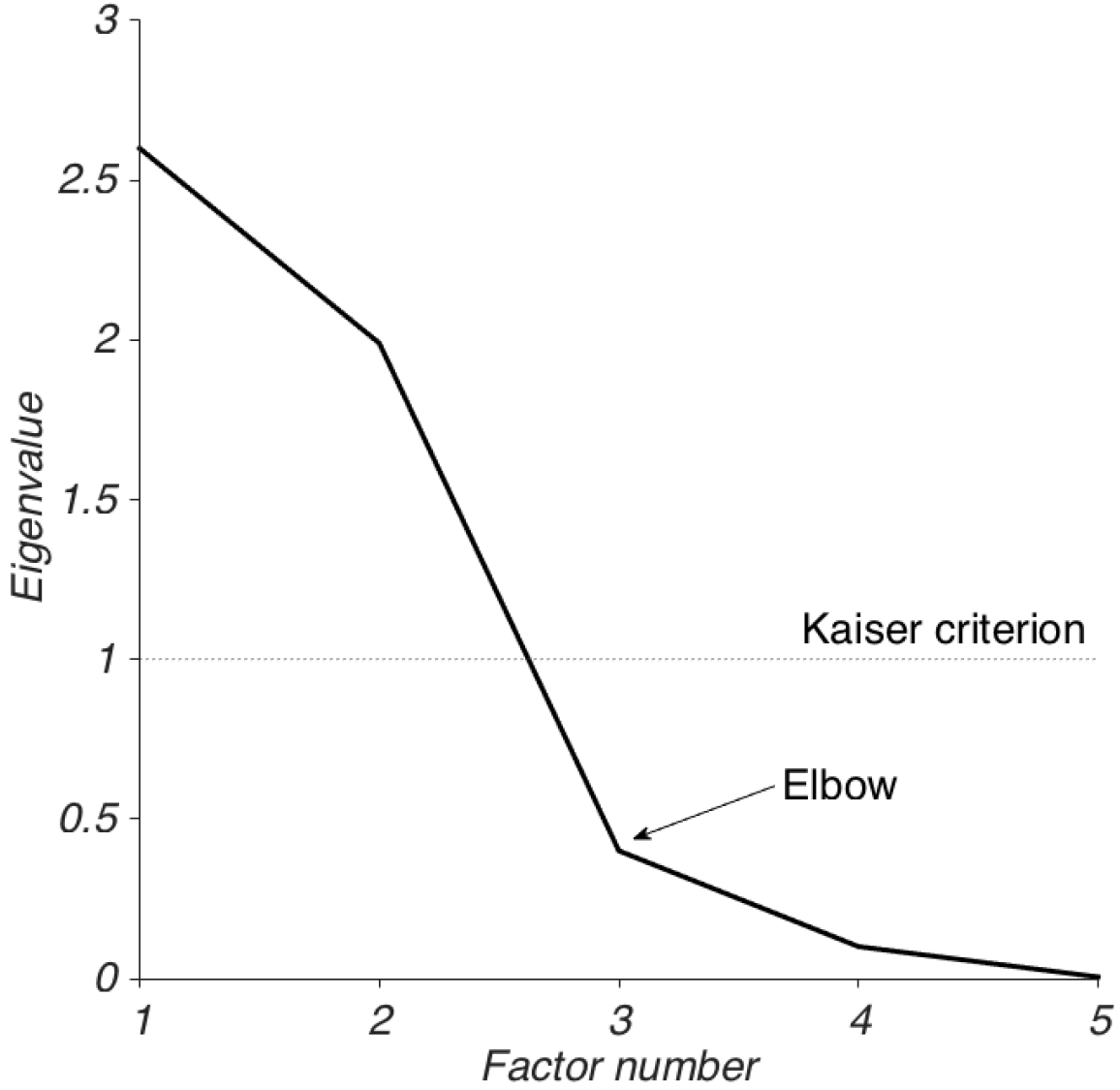
$$X = P \bullet A'$$

Determining “eigenvalues” from the factor matrix

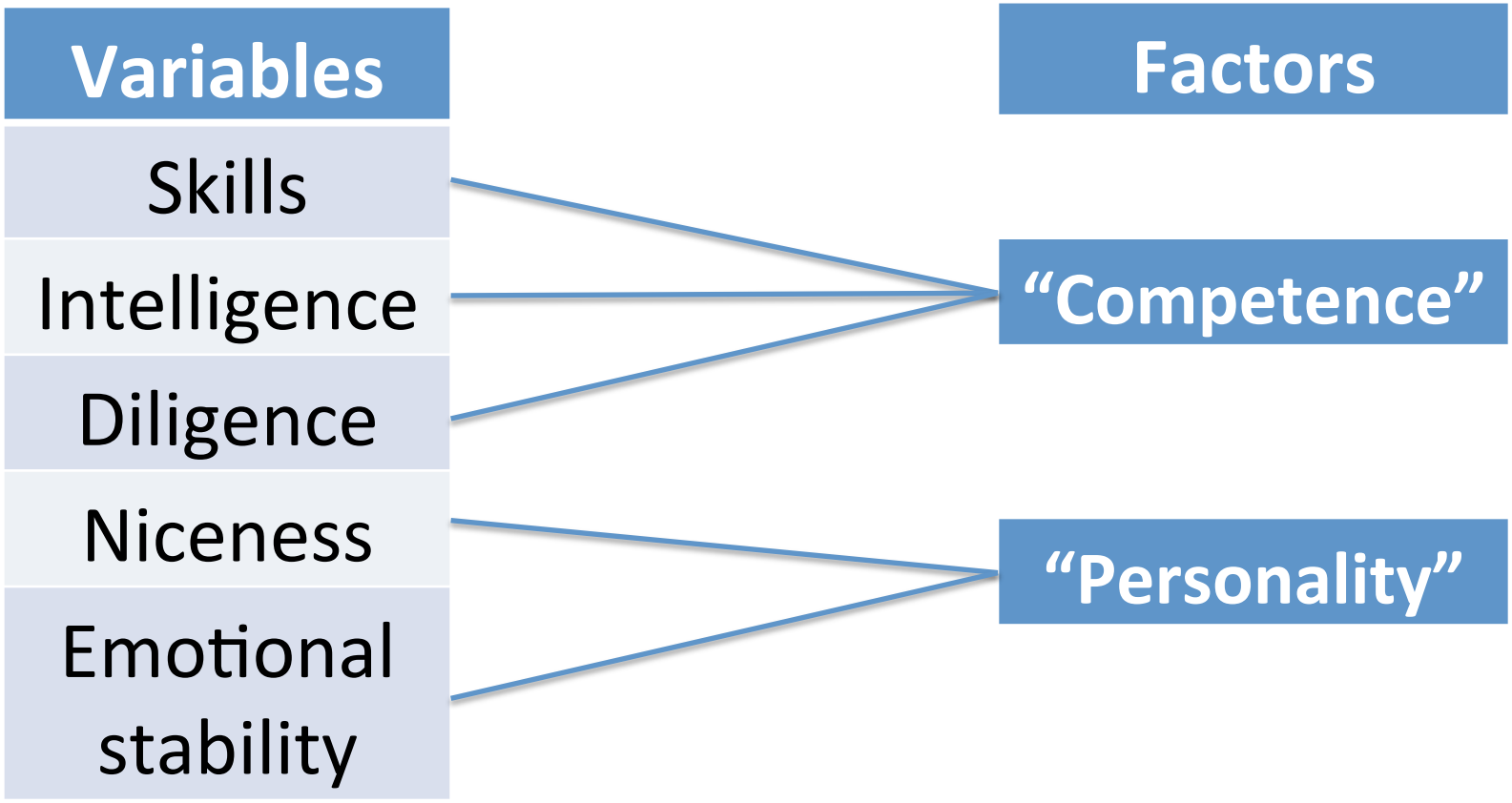
Factor matrix (loadings)

	Factor 1	Factor 2
Skills	0.89	0.08
Intelligence	0.50	0.03
Diligence	0.86	0.09
Niceness	0.15	0.84
Emotional stability	0.10	0.88
Eigenvalue	**2.51**	**1.91**

3) Number of factors: The Screeplot



4) Interpretation of meaning



5) Determine factor values

- The original data, expressed in terms of the factors (not the original variables).

$$X = P \bullet A'$$

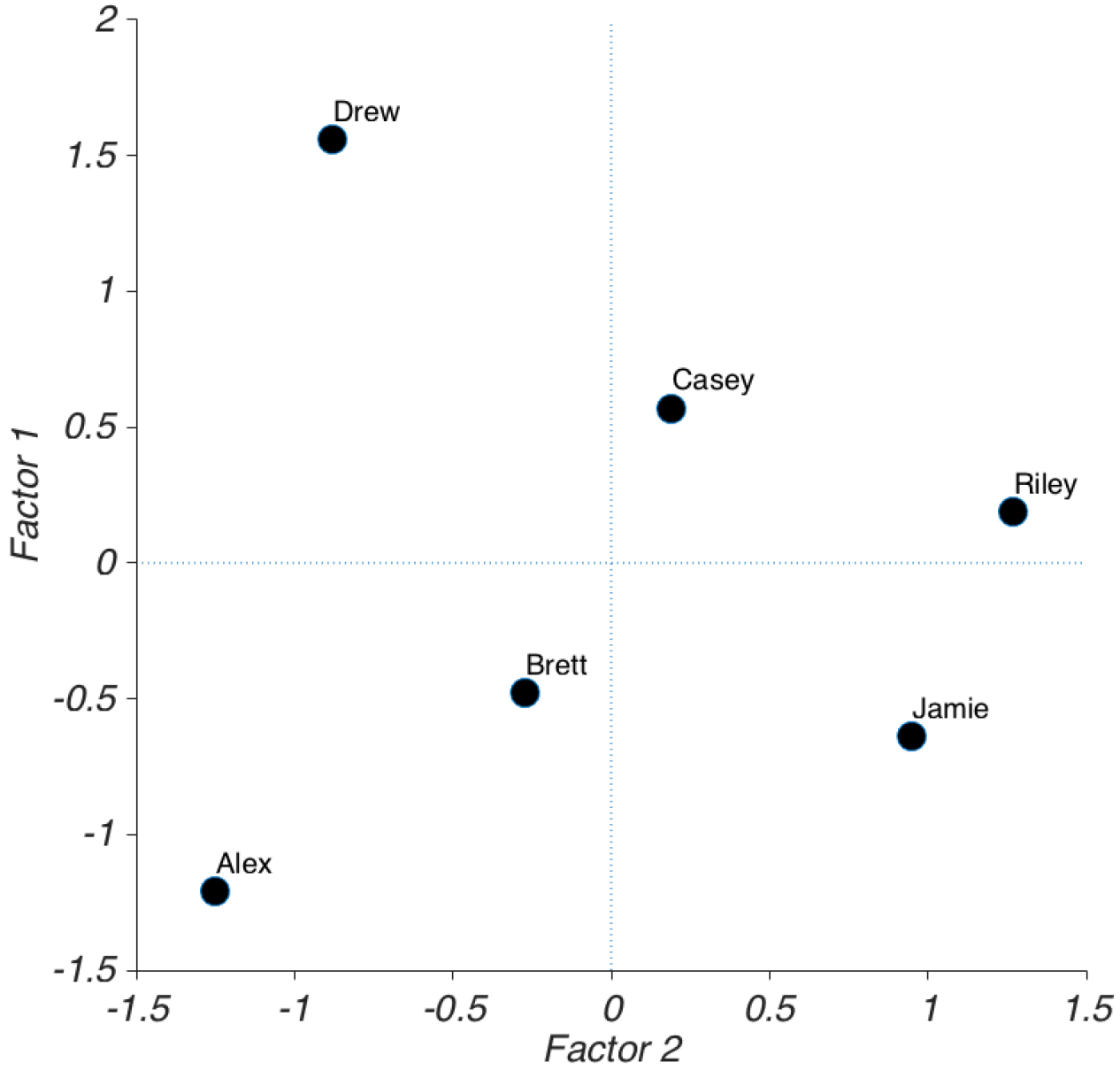
- Solve for P

$$P = X \bullet (A')^{-1}$$

Numerical solution in this case

Candidate	Factor 1	Factor 2
Alex	-1.21	-1.25
Brett	-.48	-.27
Casey	.57	.19
Drew	1.56	-.89
Jamie	-.64	.95
Riley	.20	1.27

Plot it


Drew
Casey
Riley
Brett
Jamie
Alex
2
1.5
1
0.5
0
-0.5
-1
-1.5
Factor 1
-1.5
-1
-0.5
0
0.5
1
1.5
Factor 2

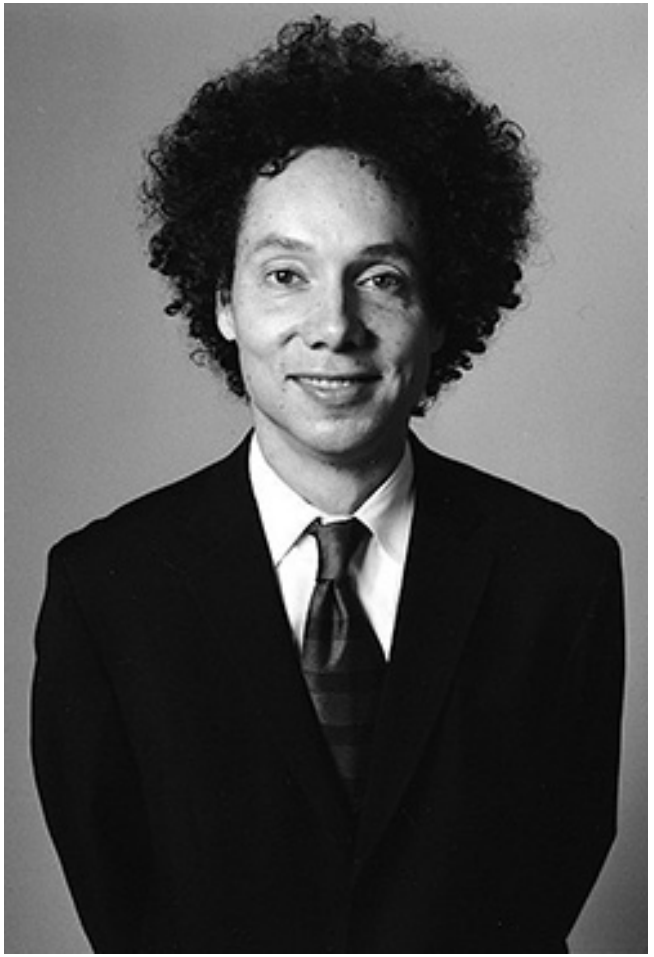
Principal Components Analysis (PCA)

- A linear transformation of data used to reduce multidimensional data to fewer dimensions.
- Idea: If one has many dimensions, they are likely not uncorrelated.
- Use the correlations between the dimensions to extract (fewer) underlying dimensions that *are* uncorrelated.
- It extracts independent “factors” that likely generated the observed (correlated, highly-dimensional) signal.
- In this sense, PCA is a coordinate transform.

It can be hard: The Igon value problem

"We say we have a Gaussian distribution, and you have the market switching from a low-volume regime to a high volume… you have your Igon value."

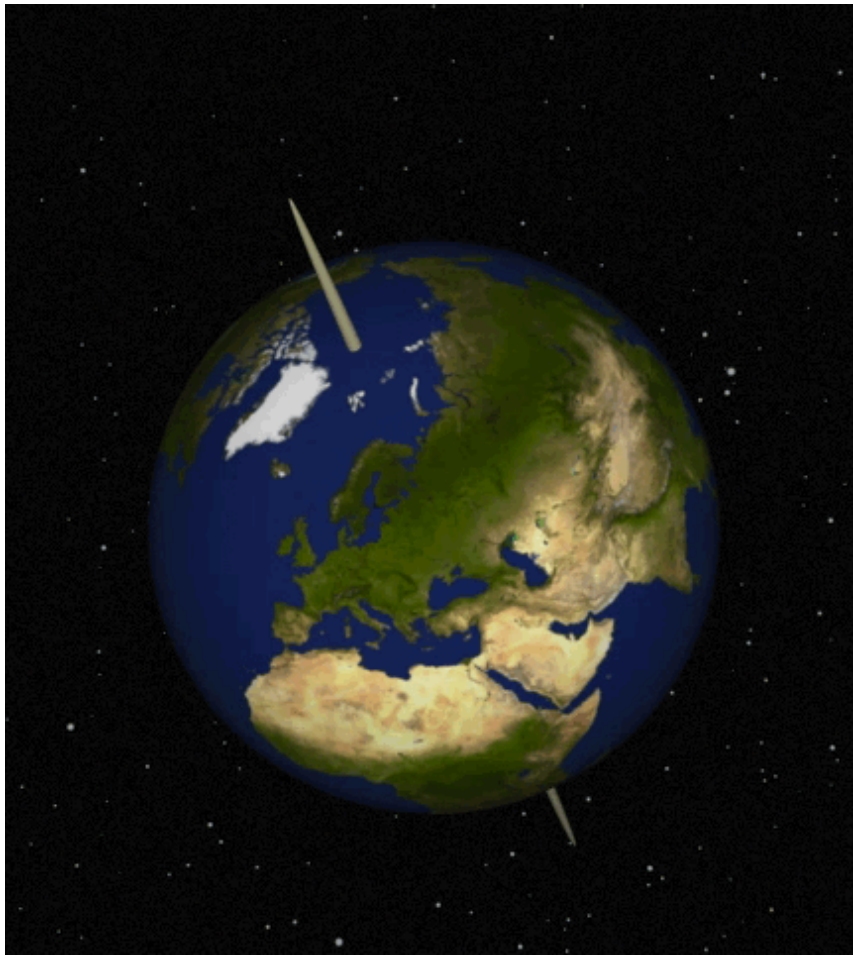
-Malcolm Gladwell, "Blowing up" (2002), relating a discussion with Nassim Taleb.



Let's do it!

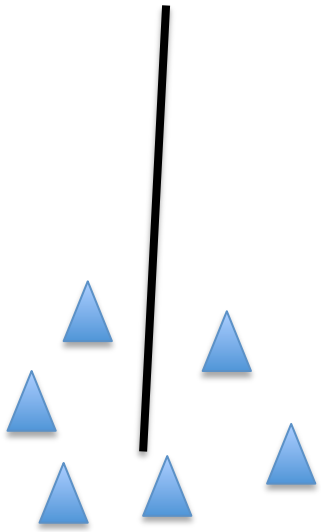
Terminology: Eigenvectors and Eigenvalues

- An Eigenvector is that (nonzero!) vector that doesn't change direction when a linear transformation is applied to it:
- $\mathbf{L(v) = \lambda v}$
- $\mathbf{Av = \lambda v}$
- Matlab will do it for you, so it is critical to understand what you are doing and why.



Our use case: Spike sorting

The fundamental problem



Spike sorting