Vorlesung 7b

Korrelationskoeffizient und Regressionsgerade

Teil 2

Der Korrelationskoeffizient

(Buch S. 62)

Definition.

Für zwei Zufallsvariable X, Y

mit positiven, endlichen Varianzen ist

$$\kappa_{XY} := \frac{\mathbf{Cov}[X,Y]}{\sqrt{\mathbf{Var}X}\sqrt{\mathbf{Var}Y}} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

der **Korrelationskoeffizient** von X und Y

(kurz auch: *die Korrelation* von X und Y).

Definition.

Für zwei Zufallsvariable X, Y

mit positiven, endlichen Varianzen ist

$$\kappa_{XY} := \frac{\mathbf{Cov}[X,Y]}{\sqrt{\mathbf{Var}X}\sqrt{\mathbf{Var}Y}}$$

der **Korrelationskoeffizient** von X und Y

(kurz auch: *die Korrelation* von X und Y).

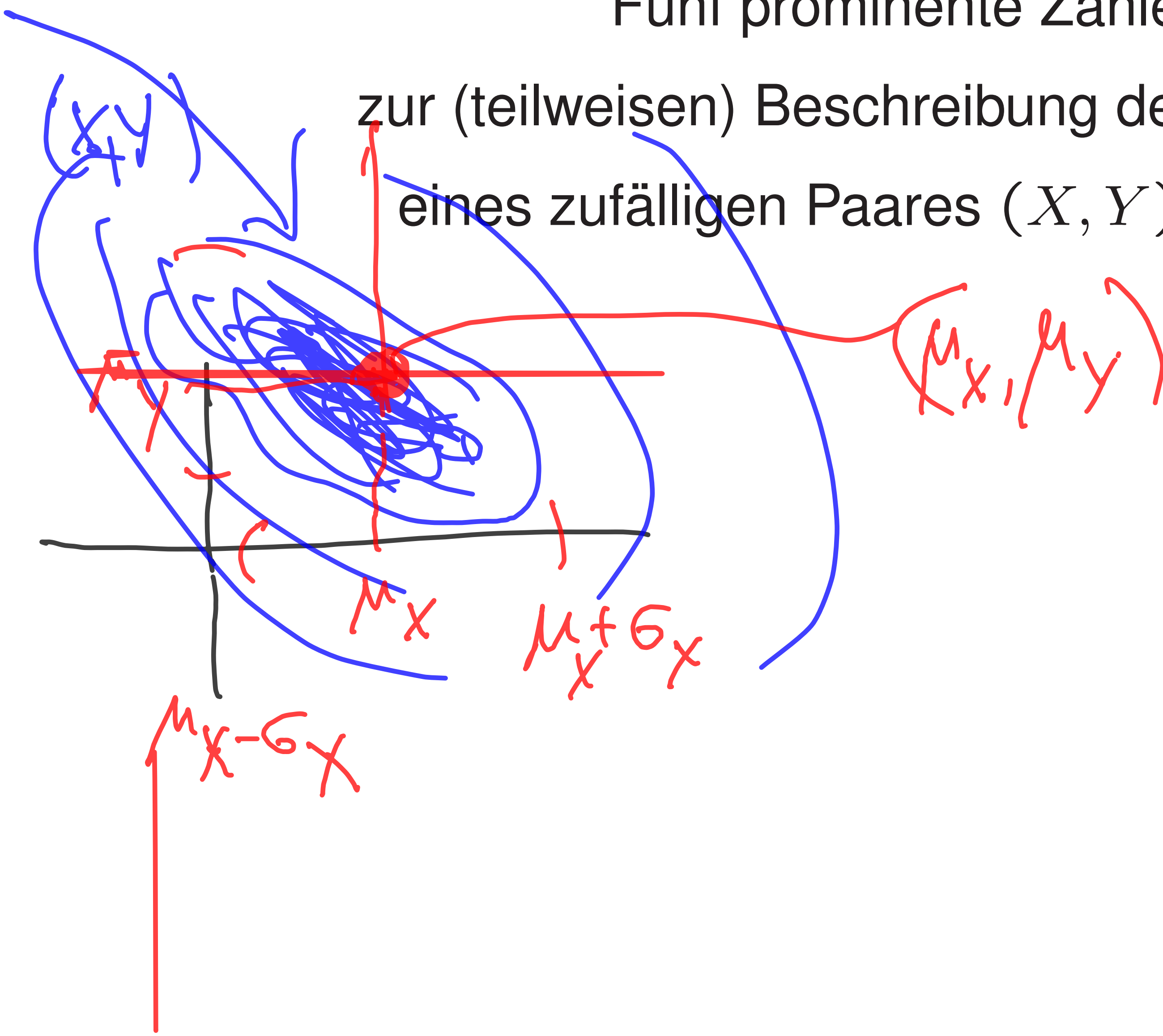
Aus der Kovarianz-Varianz-Ungleichung folgt sofort:

$$-1 \le \kappa_{XY} \le 1.$$

Fünf prominente Zahlen

zur (teilweisen) Beschreibung der Verteilung

eines zufälligen Paares (X, Y) in $\mathbb{R} \times \mathbb{R}$:



Fünf prominente Zahlen

zur (teilweisen) Beschreibung der Verteilung

eines zufälligen Paares (X, Y) in $\mathbb{R} \times \mathbb{R}$:

μ_X und μ_Y: die Erwartungswerte von X und Y

Fünf prominente Zahlen

zur (teilweisen) Beschreibung der Verteilung

eines zufälligen Paares (X, Y) in $\mathbb{R} \times \mathbb{R}$:

μ_X und μ_Y: die Erwartungswerte von X und Y

σ_X und σ_Y: die Standardabweichungen von X und Y

Fünf prominente Zahlen

zur (teilweisen) Beschreibung der Verteilung

eines zufälligen Paares (X, Y) in $\mathbb{R} \times \mathbb{R}$:

μ_X und μ_Y: die Erwartungswerte von X und Y

σ_X und σ_Y: die Standardabweichungen von X und Y

κ_{XY}: der Korrelationskoeffizient von X und Y

Die folgenden 11 Bilder zeigen jeweils die

Realisierungen von 1000 unabhängige Kopien (X_i, Y_i)

eines zufälligen Paares (X, Y), mit

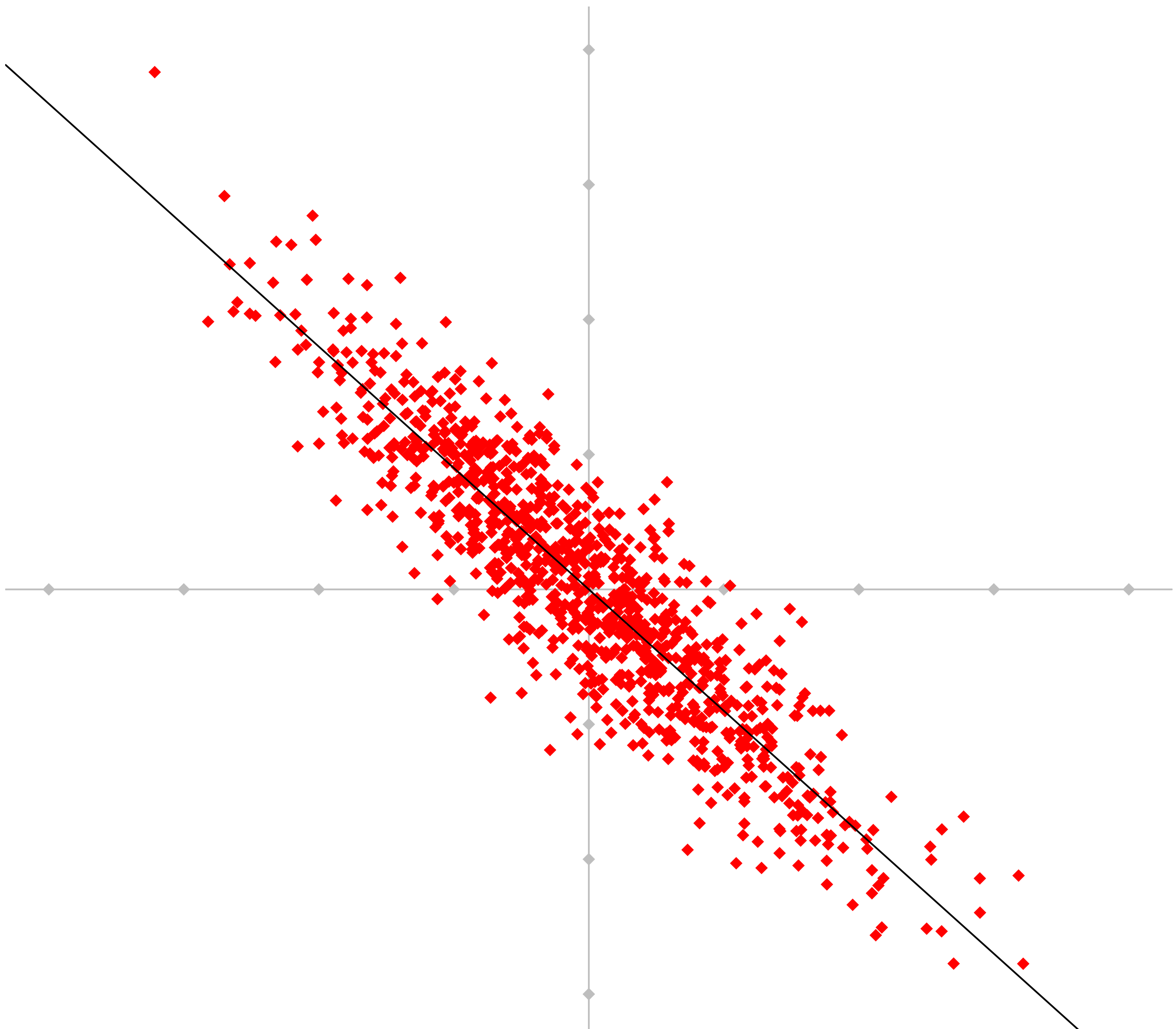
X N$(0, 1)$-verteilt , Y N$(0, 1)$-verteilt, und

$\kappa_{XY} = -0.9, -0.7, \ldots, -0.1,\ 0,\ 0.1, \ldots, 0.7, 0.9$,

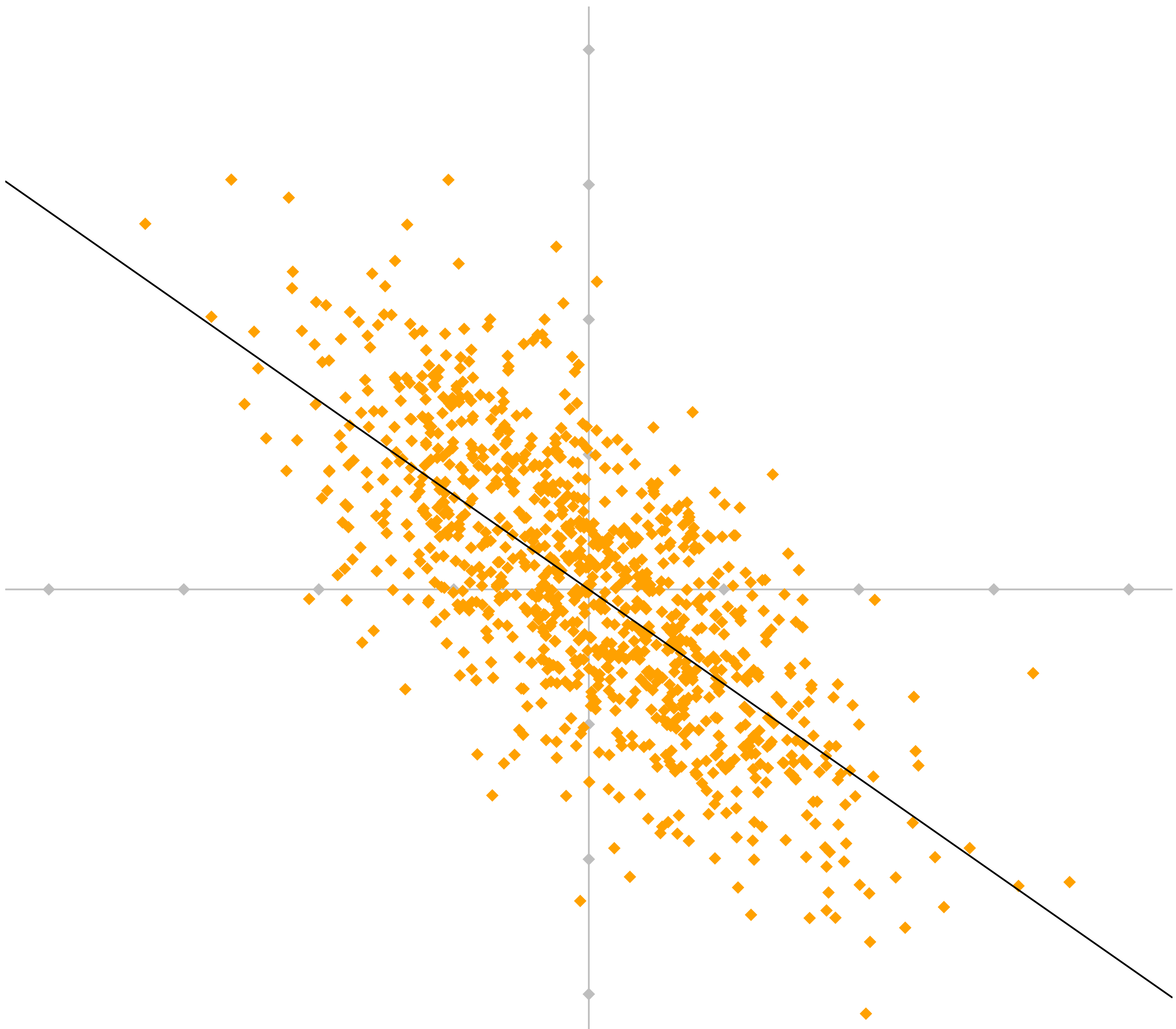
zusammen mit der

Geraden durch den Ursprung mit Anstieg κ_{XY}.

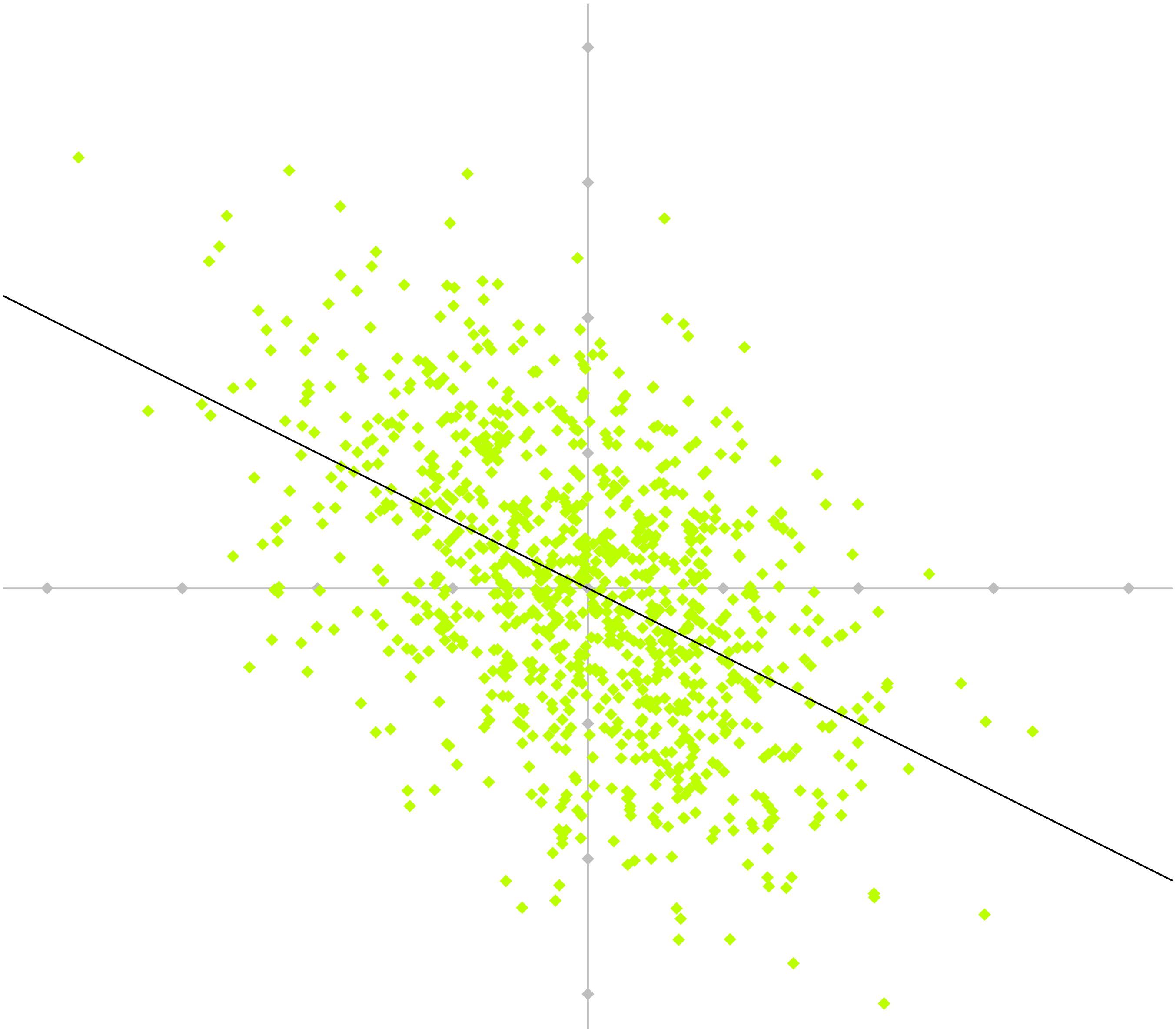
Korrelation = - 0.9



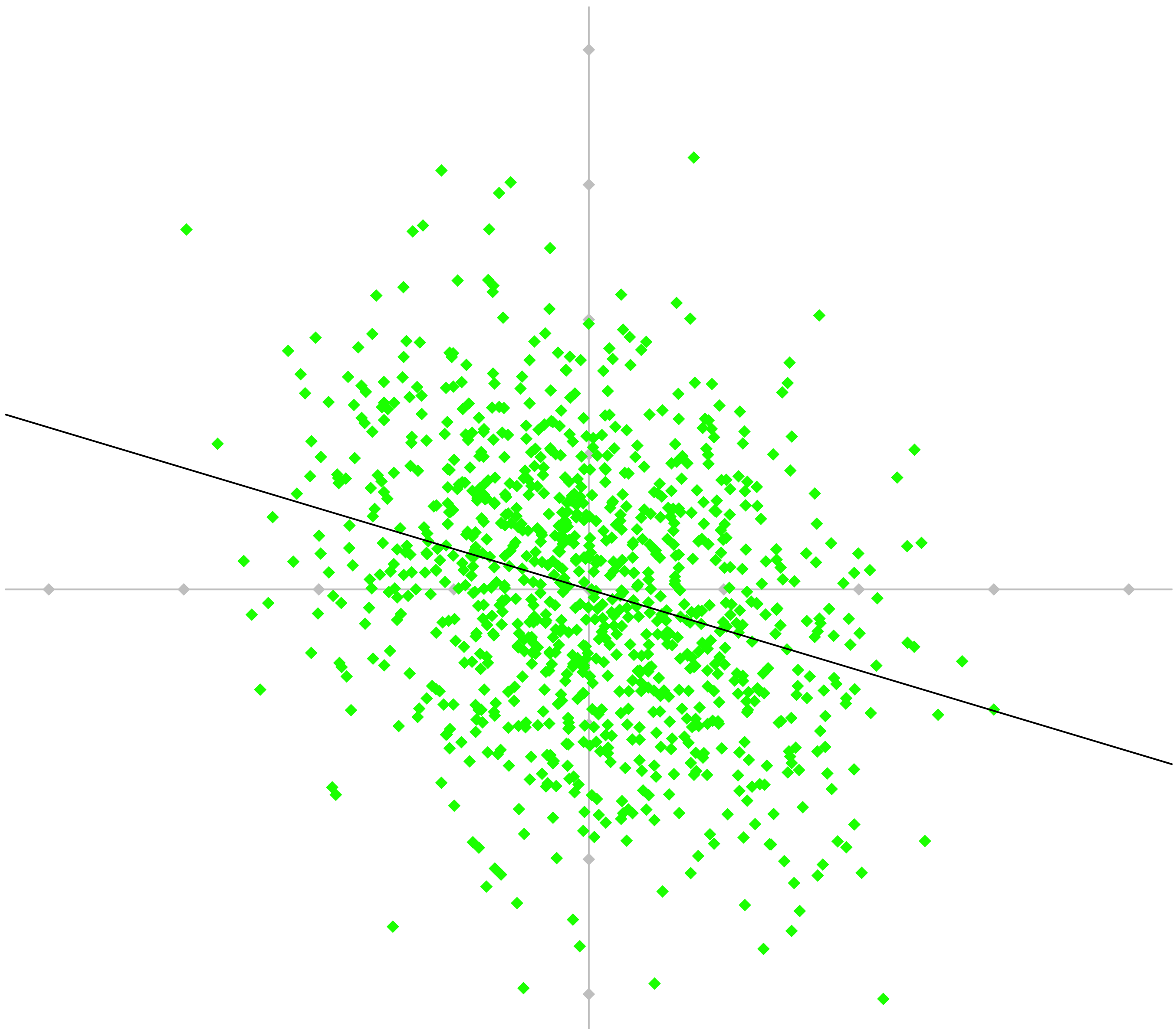
Korrelation = - 0.7



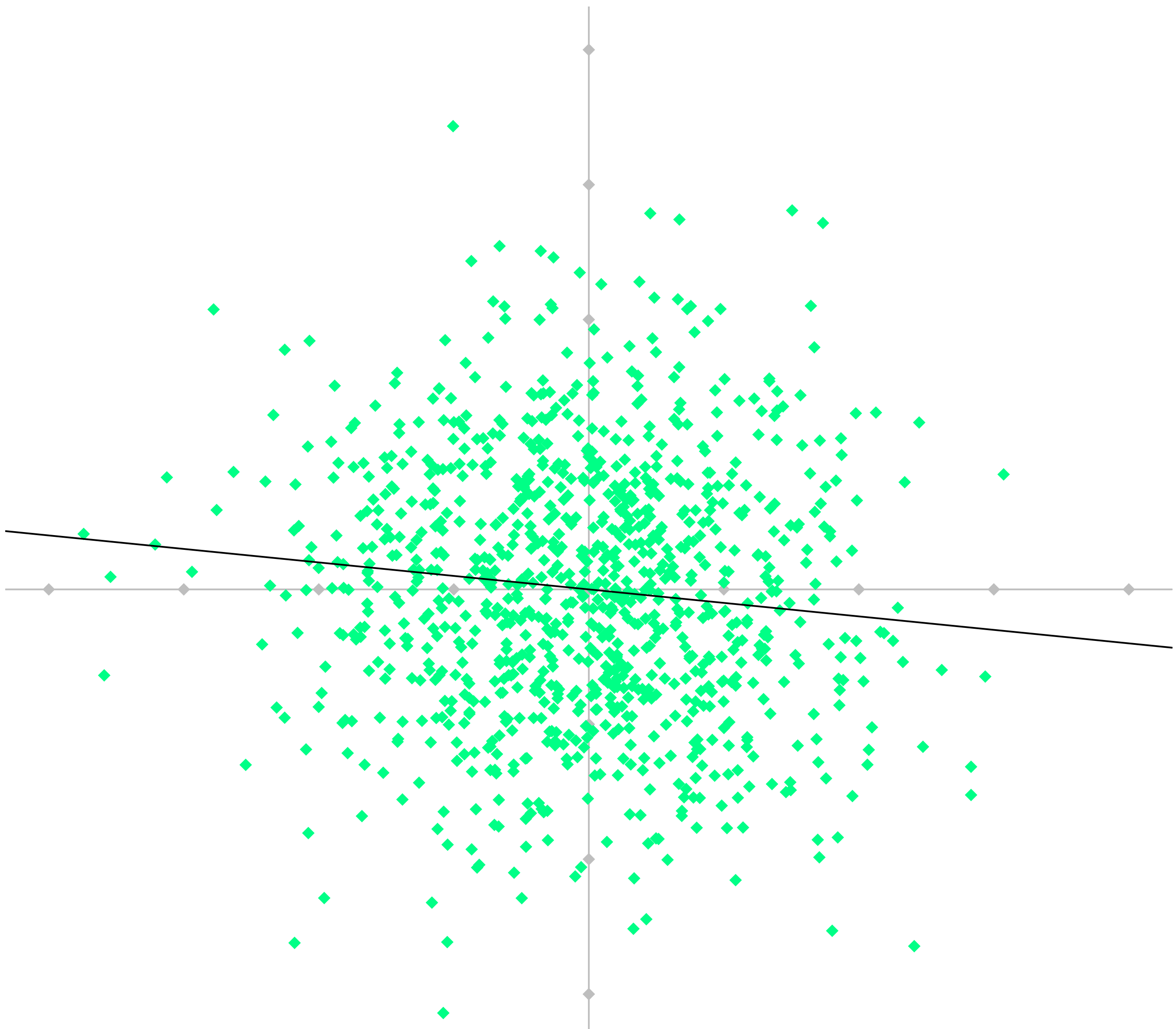
Korrelation =   - 0.5



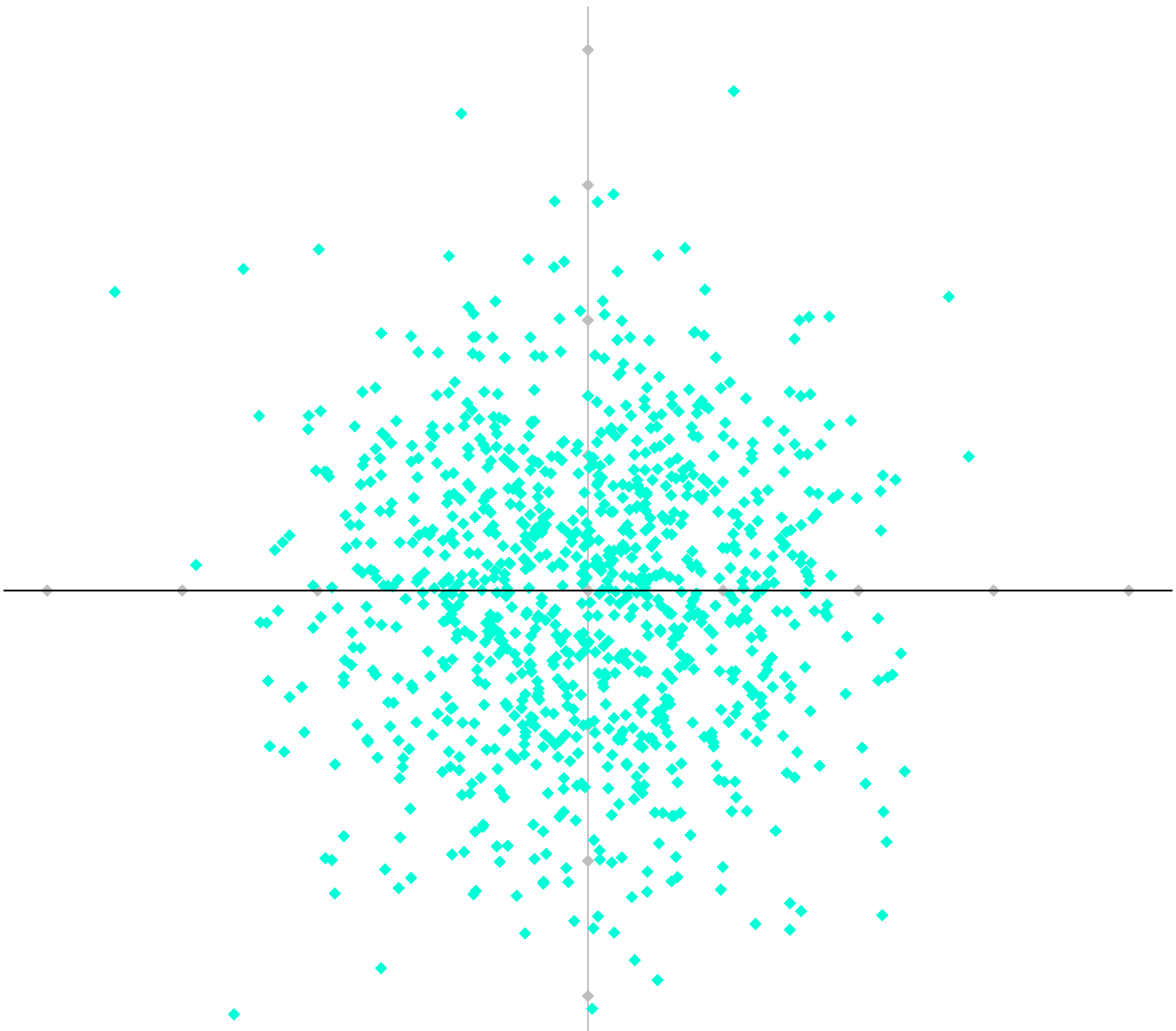
Korrelation = - 0.3



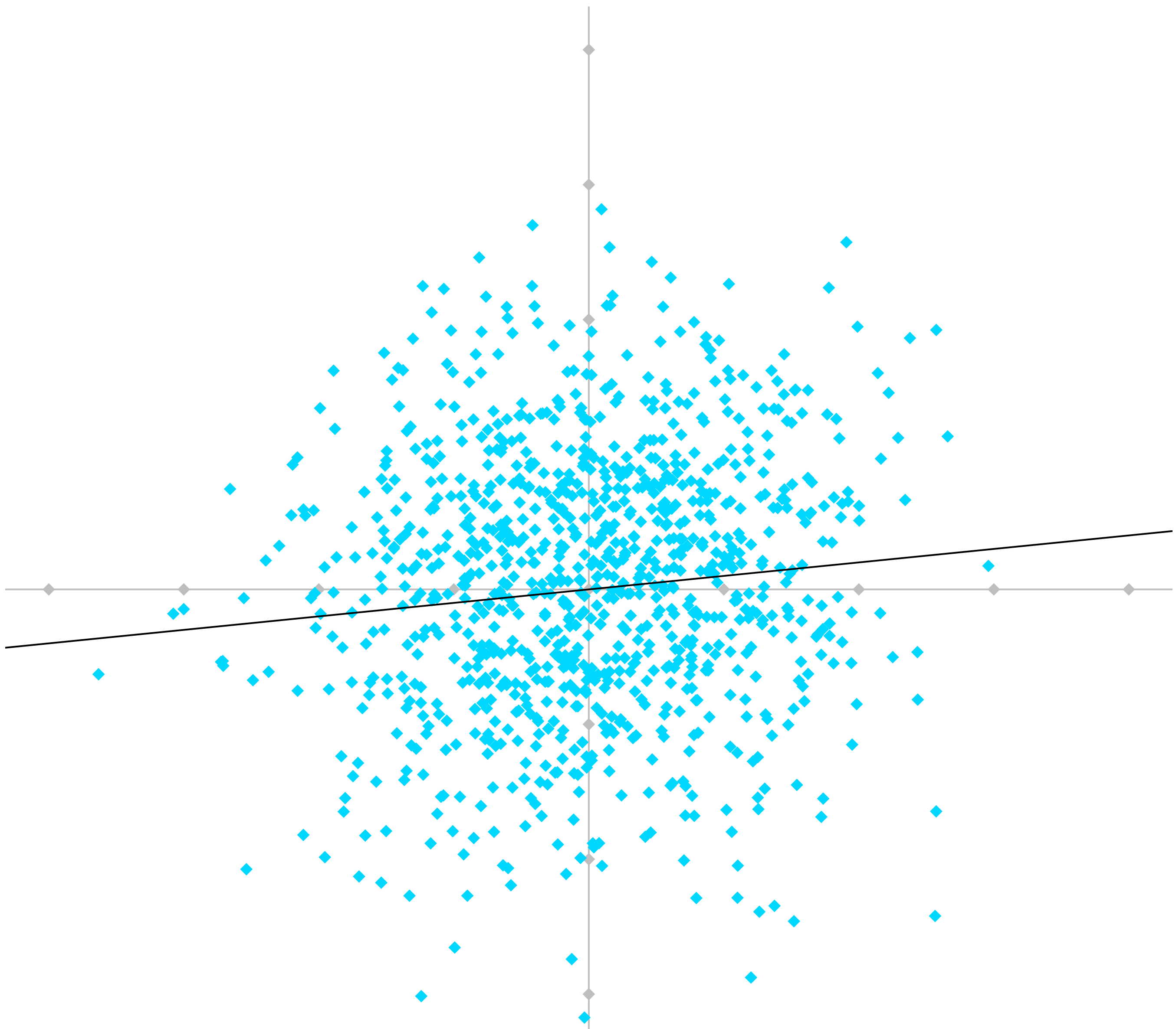
Korrelation = - 0.1



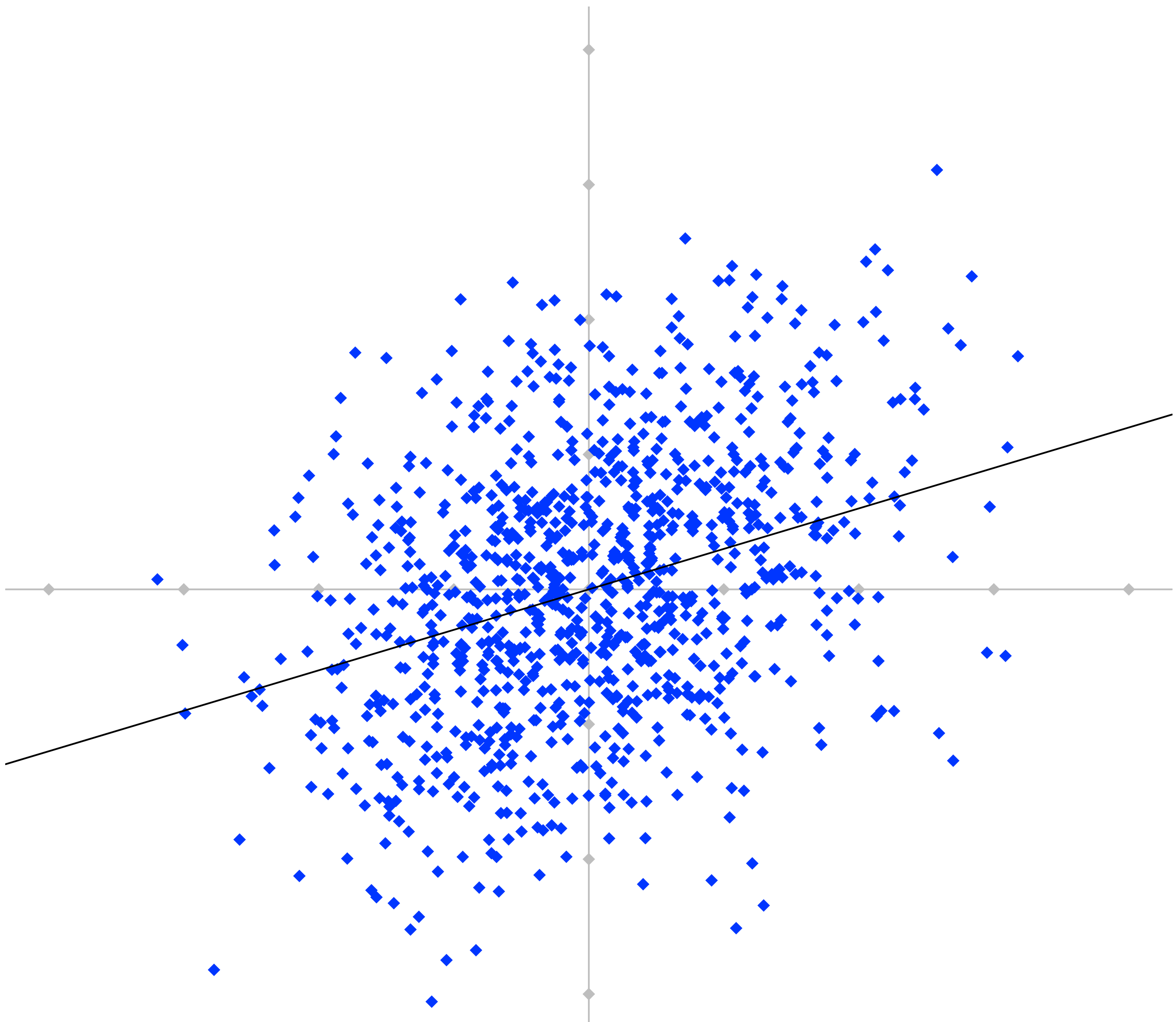
Korrelation = 0



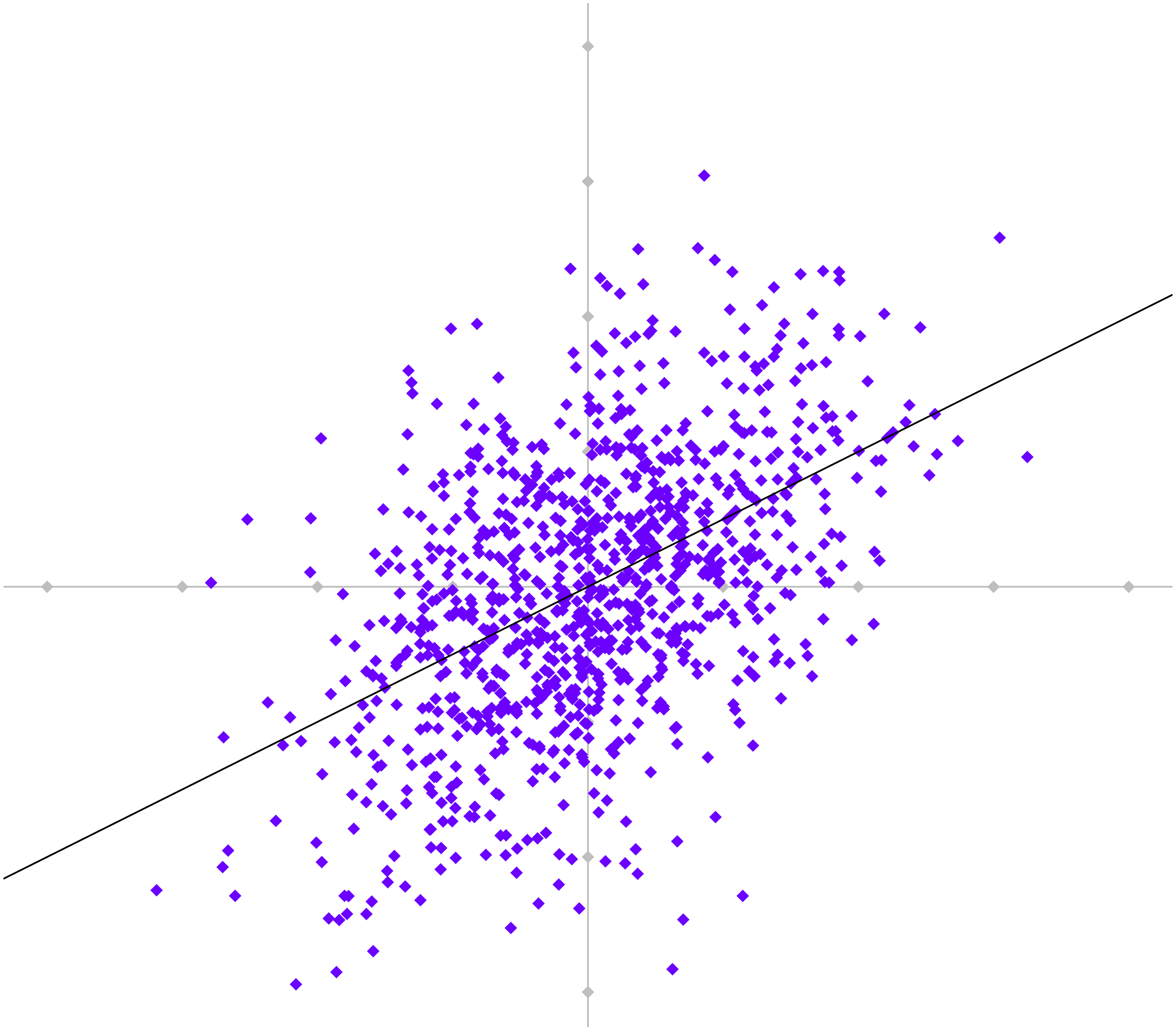
Korrelation = 0.1



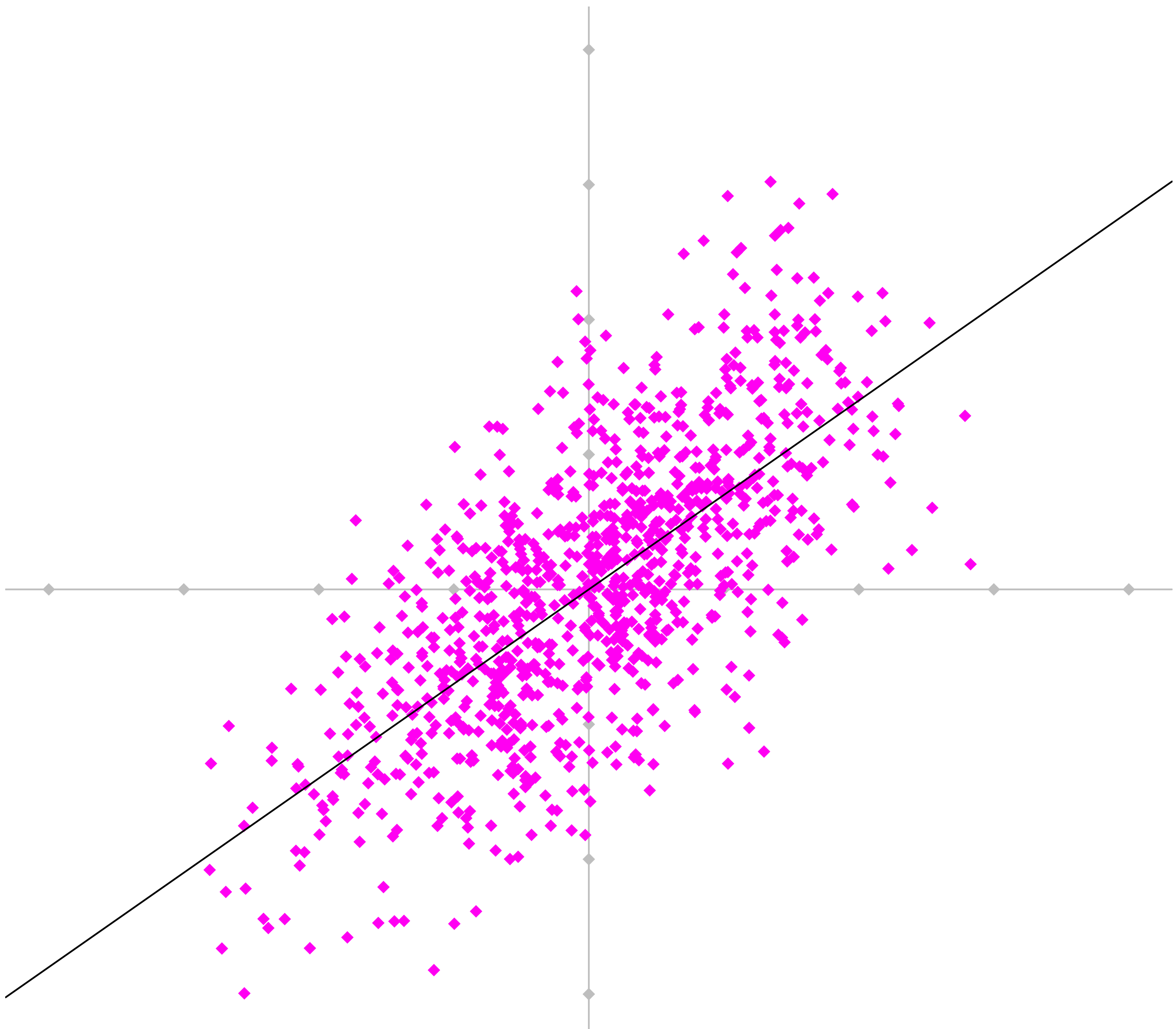
Korrelation = 0.3



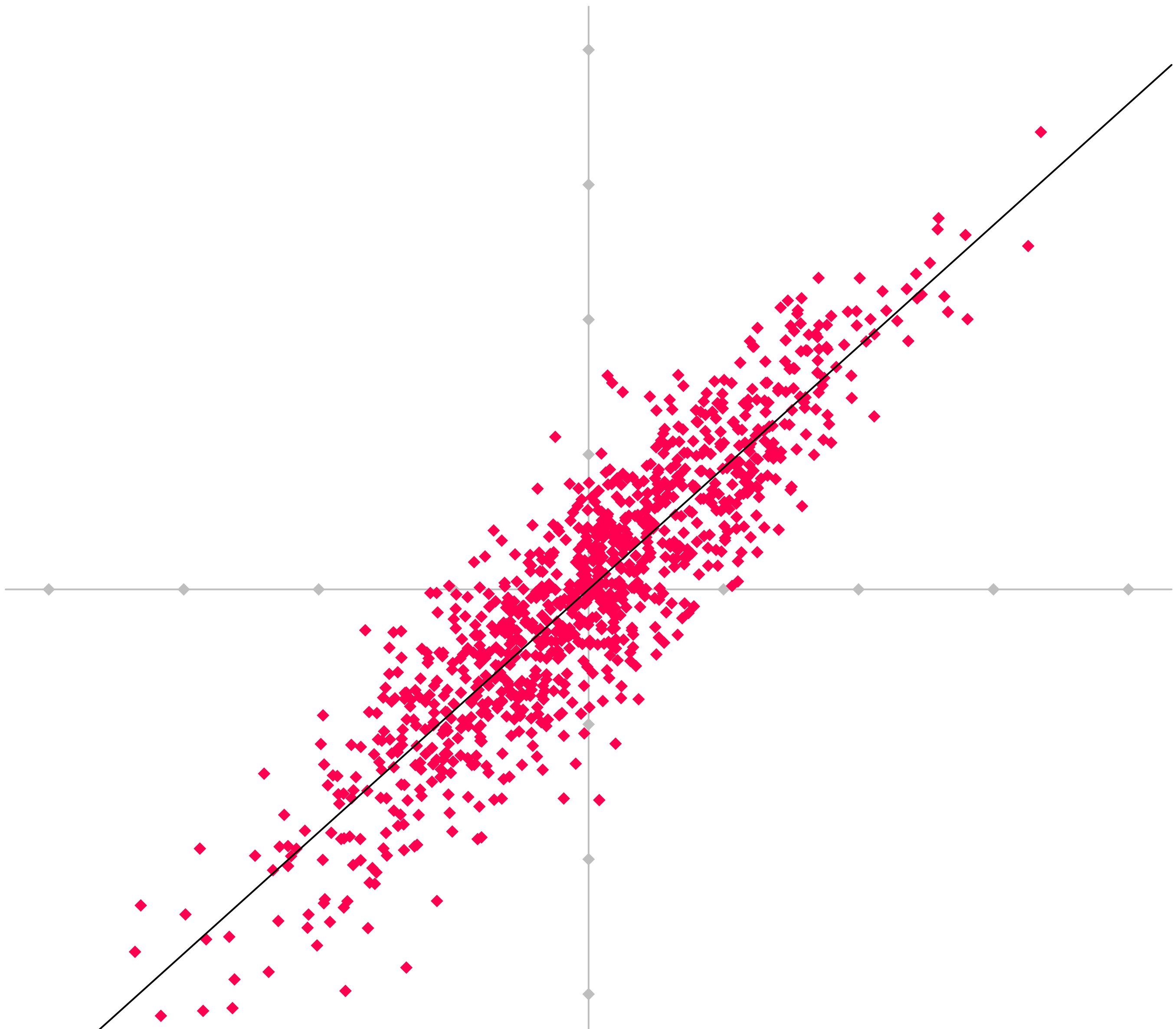
Korrelation = 0.5



Korrelation = 0.7



Korrelation = 0.9



$k^2 + ? = 1$

$k^2 + (1 - k^2) = 1$

Hier kommt das Rezept,

wie in den 9 vorangegangenen Bildern

jeweils die 1000 Punkte (X_i, Y_i) erzeugt wurden:

Z_1, Z_2 unabh. und $N(0,1)$-verteilt

$k = 1$

k_{xy}

$X = Y =: Z_1$

$\mathrm{cov}[Z_1, Z_1] = \mathrm{var}[Z_1] = 1$

Allgemein: k sei gegeben

$X := Z_1 \qquad Y =: k Z_1 + \sqrt{1-k^2}\, Z_2$

Beispiel:

Gemeinsam normalverteilte Zufallsvariable X und Y

mit $\sigma_X^2 = \sigma_Y^2 = 1$ und Korrelationskoeffizient κ:

Z_1, Z_2 seien unabhängig und standard-normalverteilt.

Beispiel:

Gemeinsam normalverteilte Zufallsvariable X und Y

mit $\sigma_X^2 = \sigma_Y^2 = 1$ und Korrelationskoeffizient κ:

Z_1, Z_2 seien unabhängig und standard-normalverteilt.

$$\boxed{X := Z_1, \quad Y := \kappa Z_1 + \sqrt{1-\kappa^2} Z_2.}$$

Aus der Bilinearität der Kovarianz folgt:

$$\mathbf{Cov}[X, Y] = \mathbf{Cov}[Z_1, \kappa Z_1] + 0 = \kappa.$$

$$\sigma_X^2 = 1, \quad \sigma_Y^2 = \kappa^2 + (1 - \kappa^2) = 1.$$

Also: $\kappa_{XY} = \kappa$.

Auch Y ist (standard-) normalverteilt (siehe V7a1 Folie 12).

Eine Interpretation von κ^2:

Wir werden sehen:

Für jedes zufällige Paar (X, Y) reellwertiger Zufallsvariabler mit endlichen Varianzen gilt:

κ^2 ist ein Maß dafür, um wieviel besser man Y durch eine affin lineare Funktion von X vorhersagen kann:

$Y = \beta_1 X + \beta_0 +$ "Fehler",

Eine Interpretation von κ^2:

Wir werden sehen:

Für jedes zufällige Paar (X, Y) reellwertiger Zufallsvariabler

mit endlichen Varianzen gilt:

κ^2 ist ein Maß dafür, um wieviel besser man Y

durch eine affin lineare Funktion von X vorhersagen kann:

$Y = \beta_1 X + \beta_0 +$ “Fehler”,

als durch eine Konstante:

$Y = c +$ “Fehler”.

(Die “Güte der Vorhersage” bezieht sich auf die Kleinheit des **erwarteten quadrierten “Fehlers” (mean sqare error)**.)