Wie indexieren Google & Co 13 Millionen Seiten?

Inetbib-Tagung Bonn, 05.11.2004

Florian Seiffert, HBZ

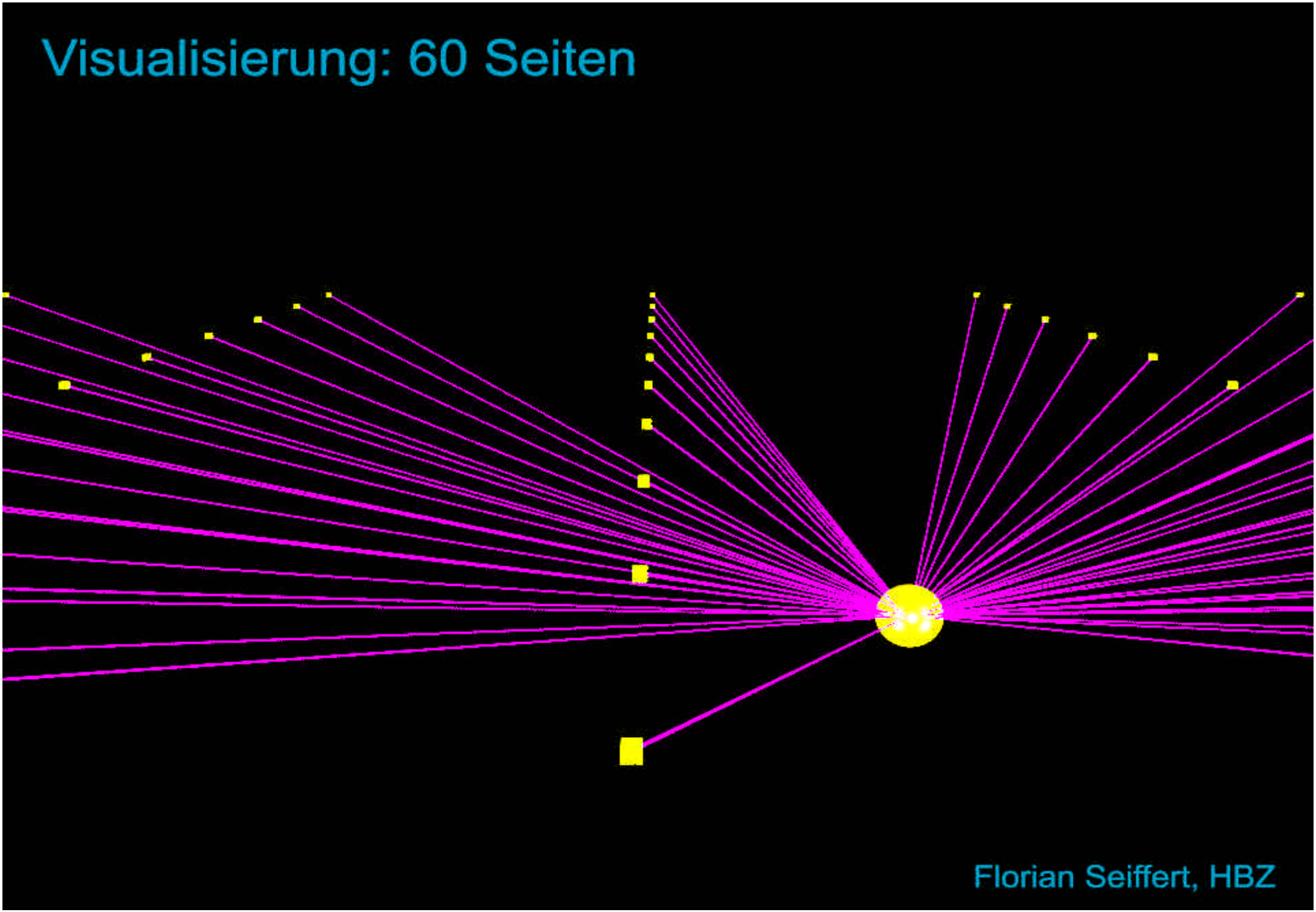
Was haben wir denn heute so vor ...

- Visualisierung
- Google
  - Indexierungsstrategie
  - Indexierungsleistung
  - Seitenzahlbegrenzung pro Server
  - Dauer bis zur Findbarkeit in der Suche
  - "Tiefe" der Indexierung
  - Dateigroessen
  - Einfluss auf die Besuchshaeufigkeit der Bots
- fast, msn, neofonie
- Was lernen wir daraus

Visualisierung: root

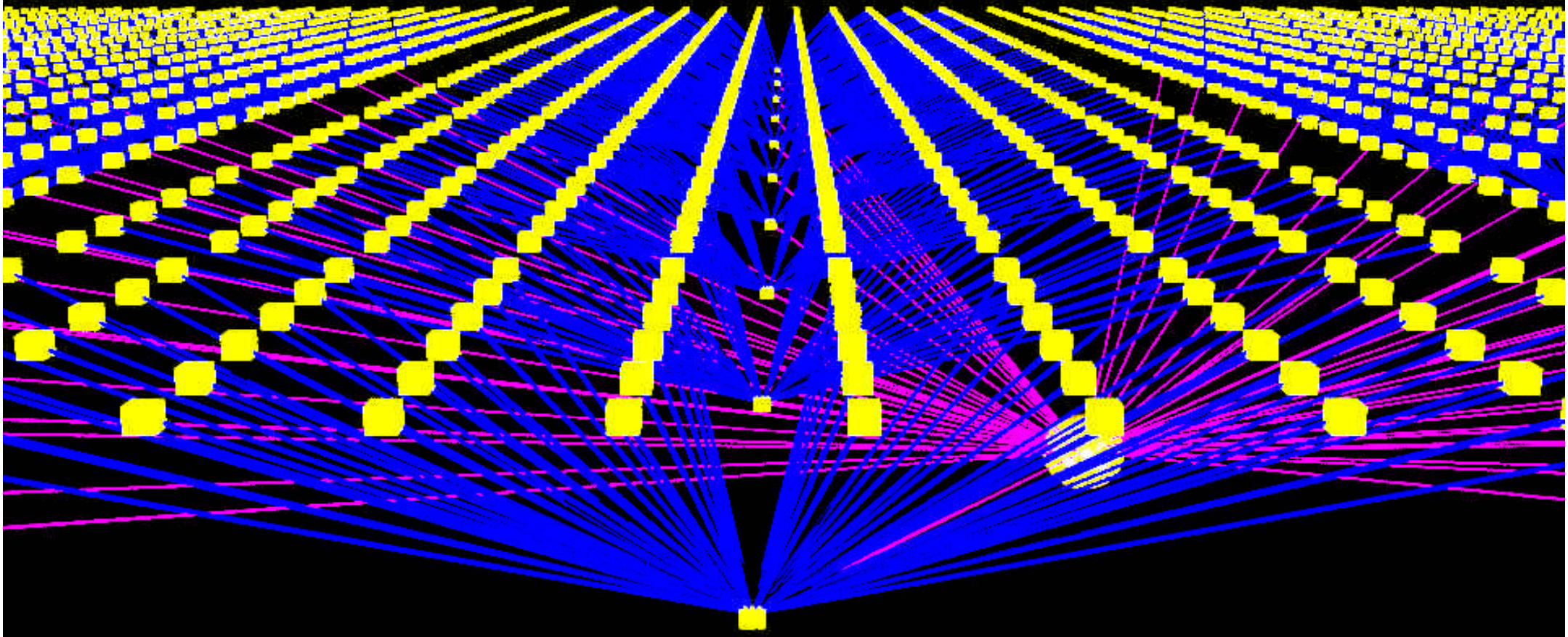
Florian Seiffert, HBZ

Visualisierung: 60 Seiten



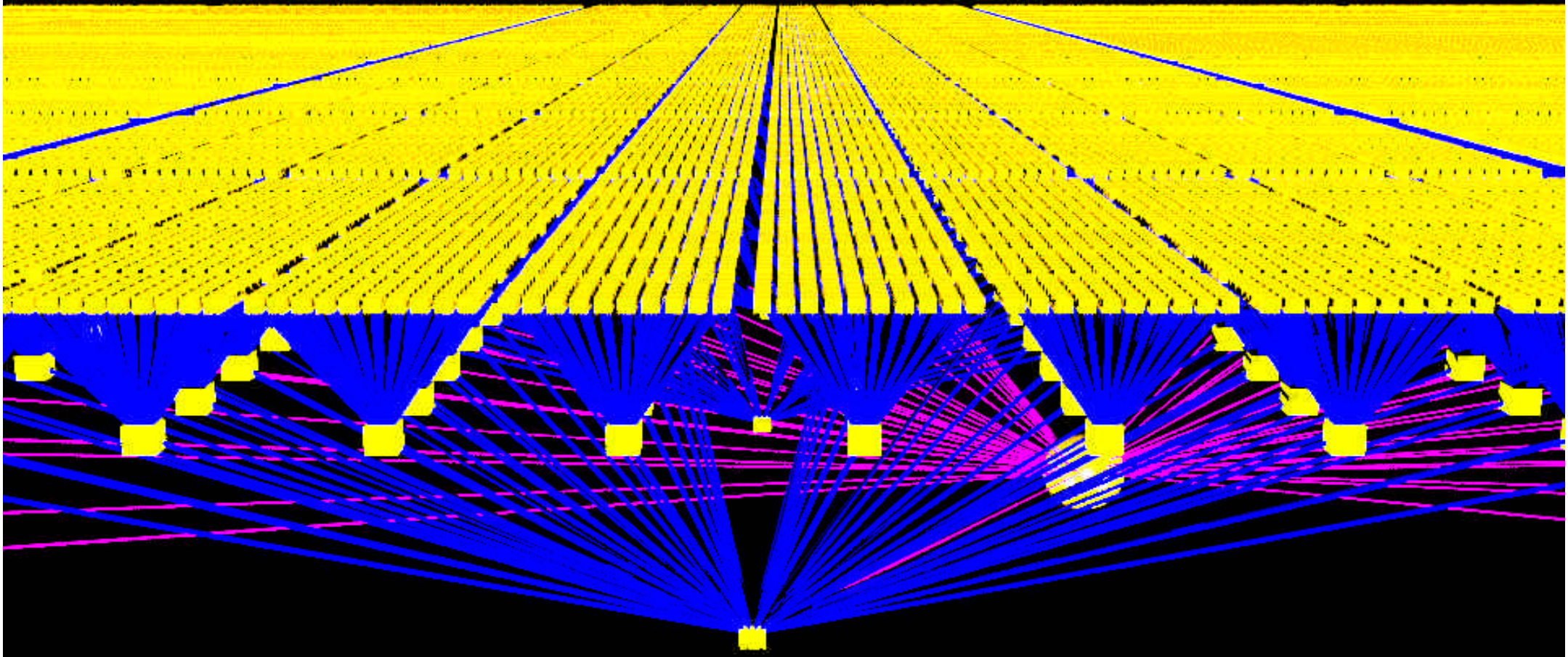
Florian Seiffert, HBZ

Visualisierung: 60x60 = 3600 Seiten



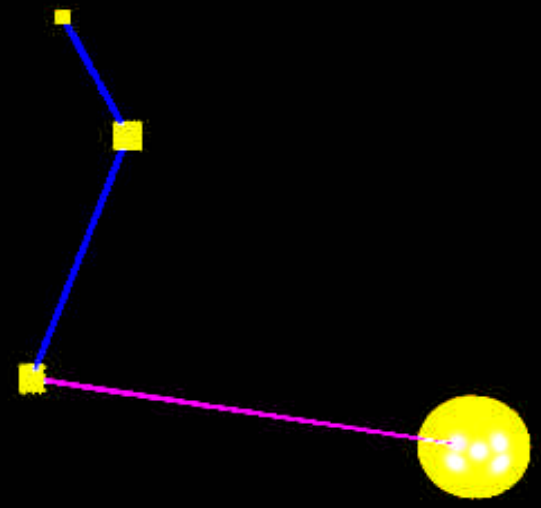
Florian Seiffert, HBZ

Visualisierung: 60x60x60 = 216.000 Zellen



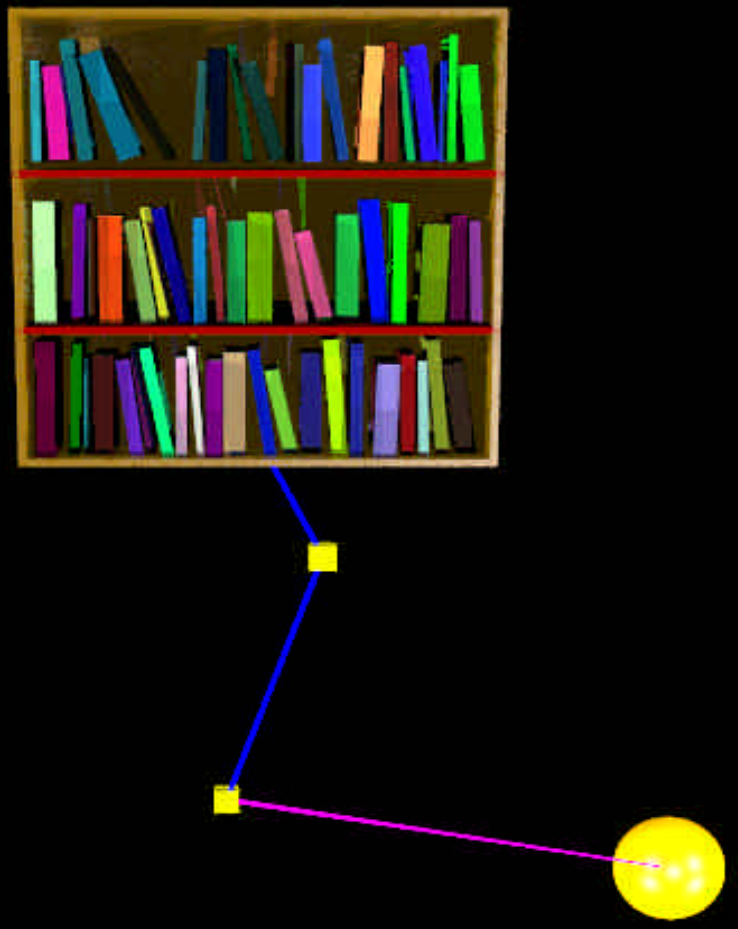
Florian Seiffert, HBZ

Visualisierung: Eine Zelle



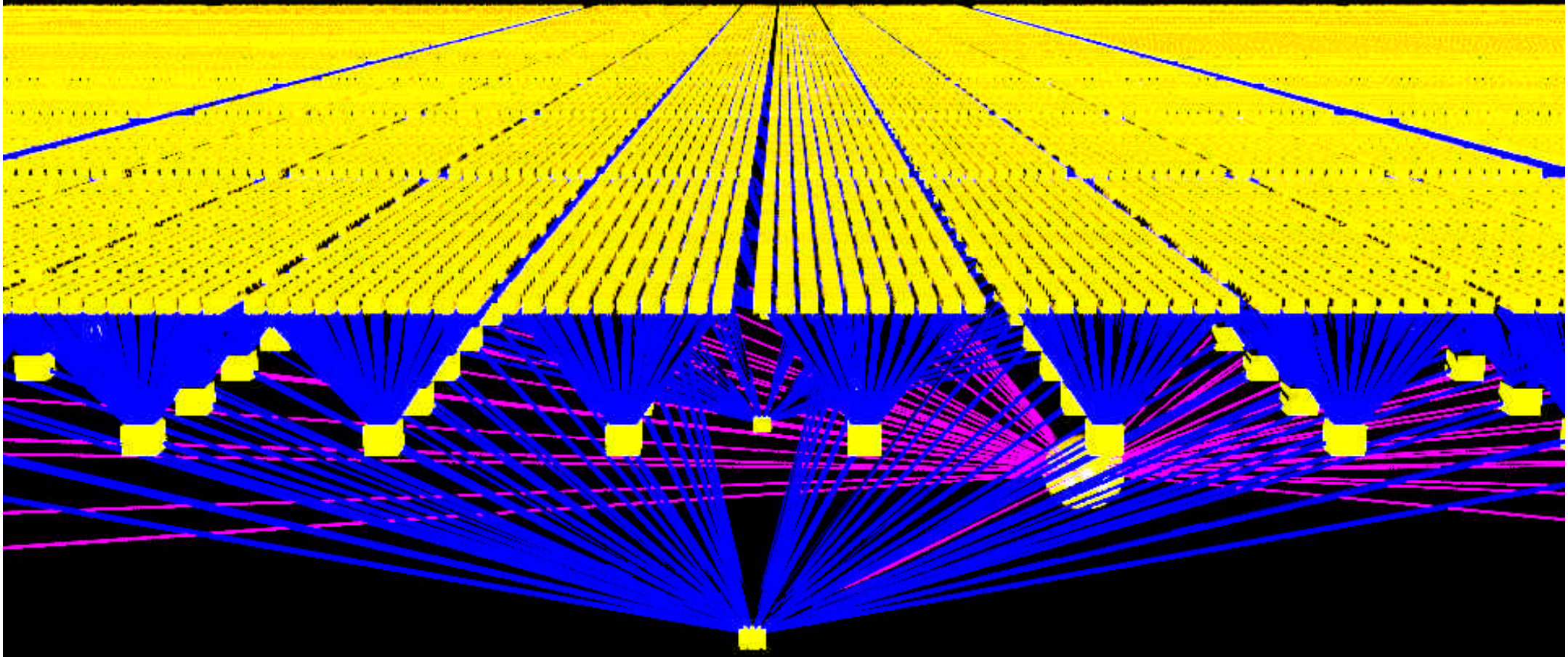
Florian Seiffert, HBZ

Visualisierung: 60 Buecher pro Zelle -> 12.96 Mio



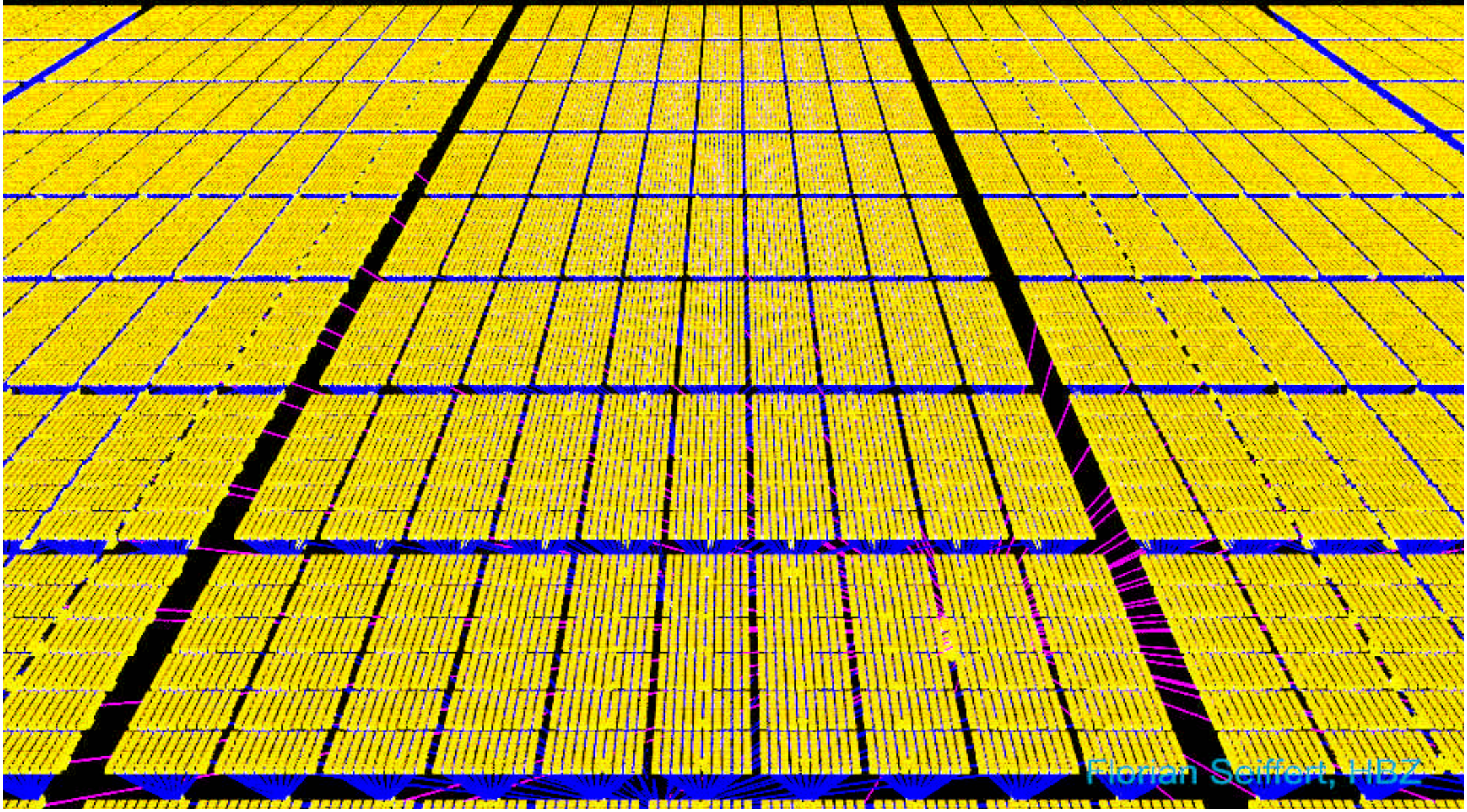
Florian Seiffert, HBZ

Visualisierung: 60x60x60 = 216.000 Zellen

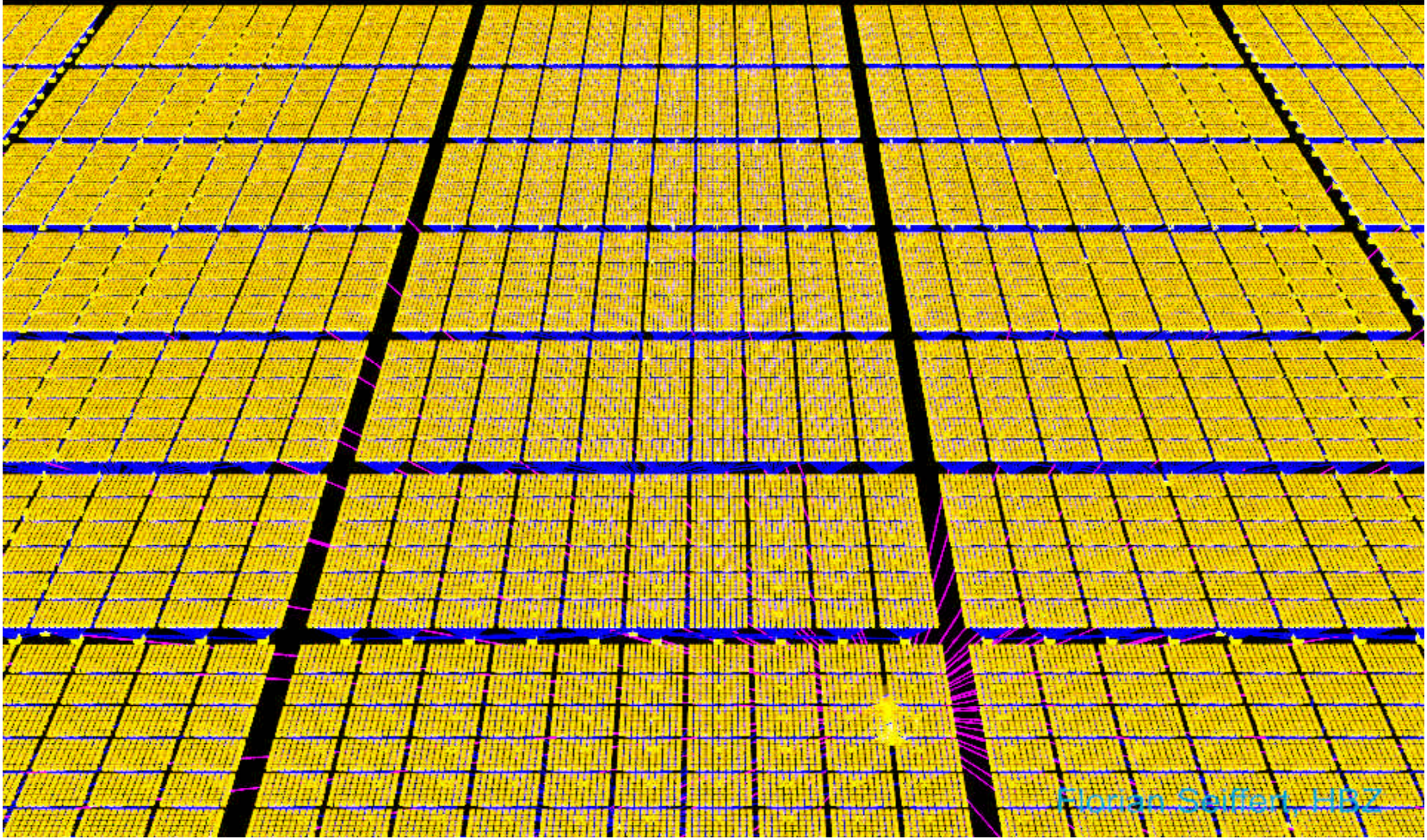


Florian Seiffert, HBZ

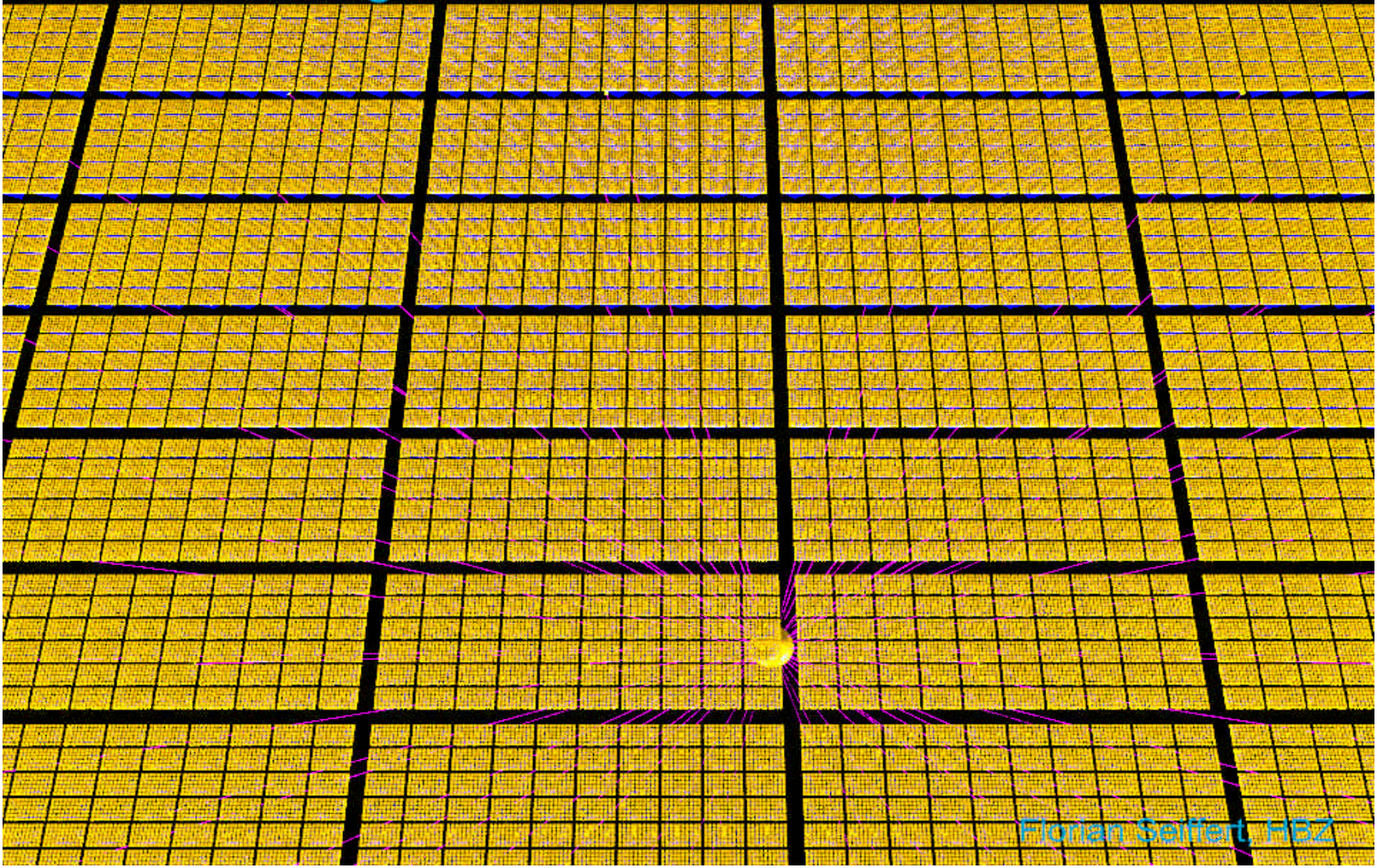
Visualisierung: Wir drehen die Konstruktion



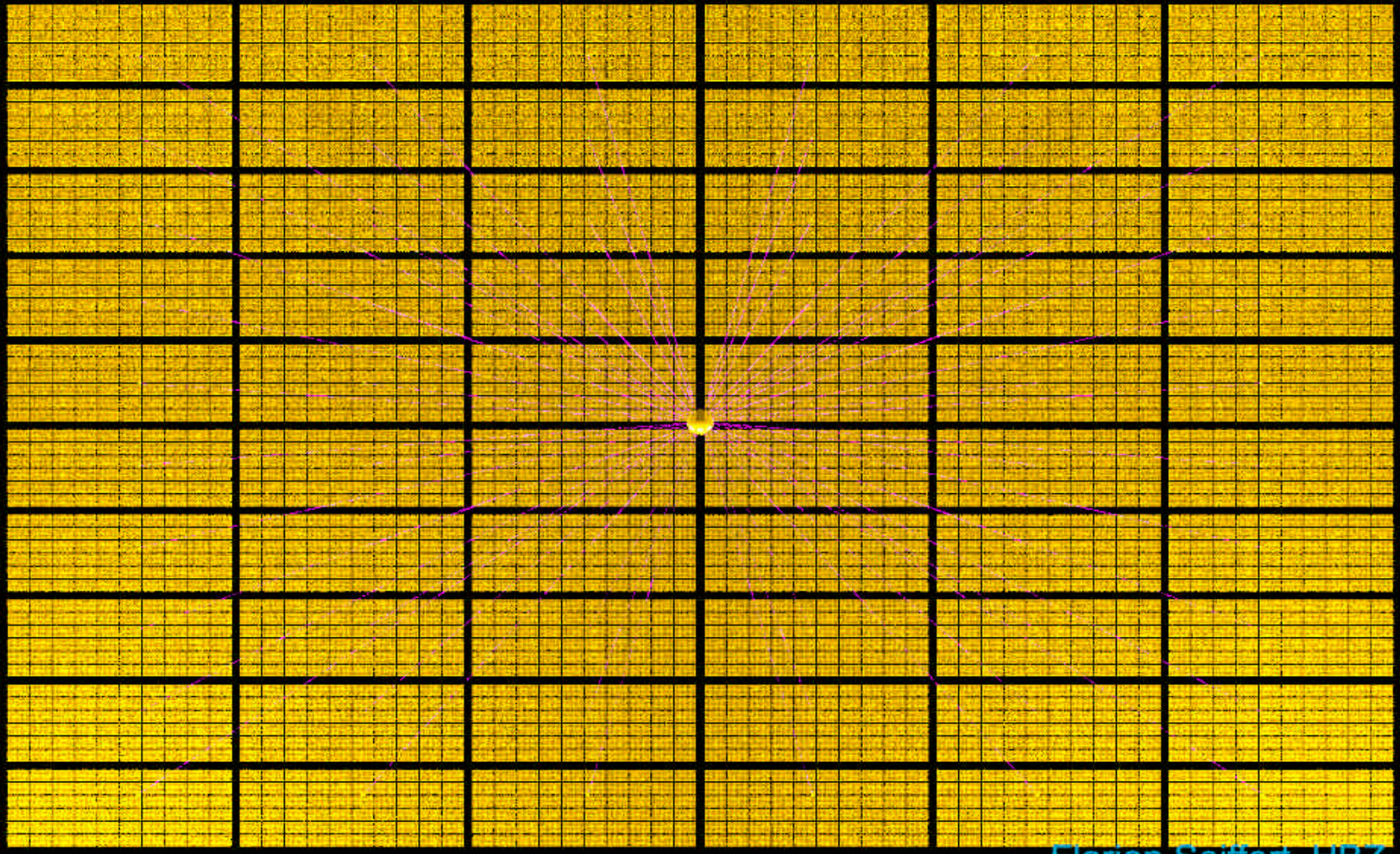
Visualisierung: Wir drehen die Konstruktion



Visualisierung: Wir drehen die Konstruktion



Visualisierung: Fertig

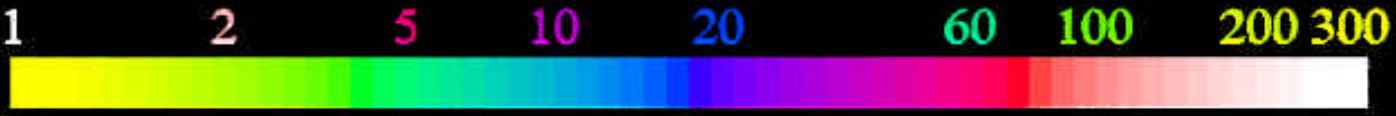


Florian Seiffert, HBZ

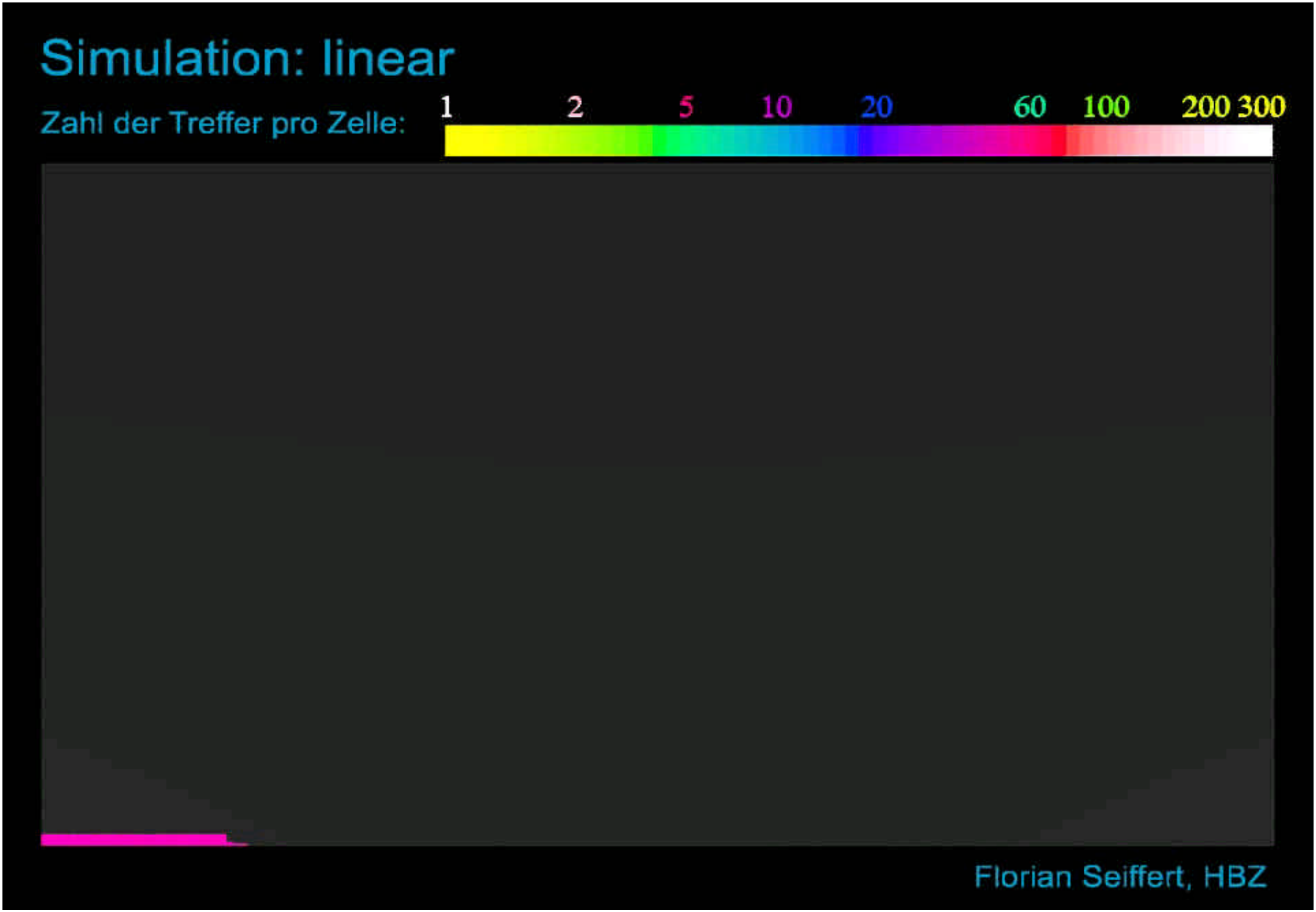
Visualisierung: Abstaende entfernt

Visualisierung: Monitor auf vbR

Zahl der Treffer pro Zelle:

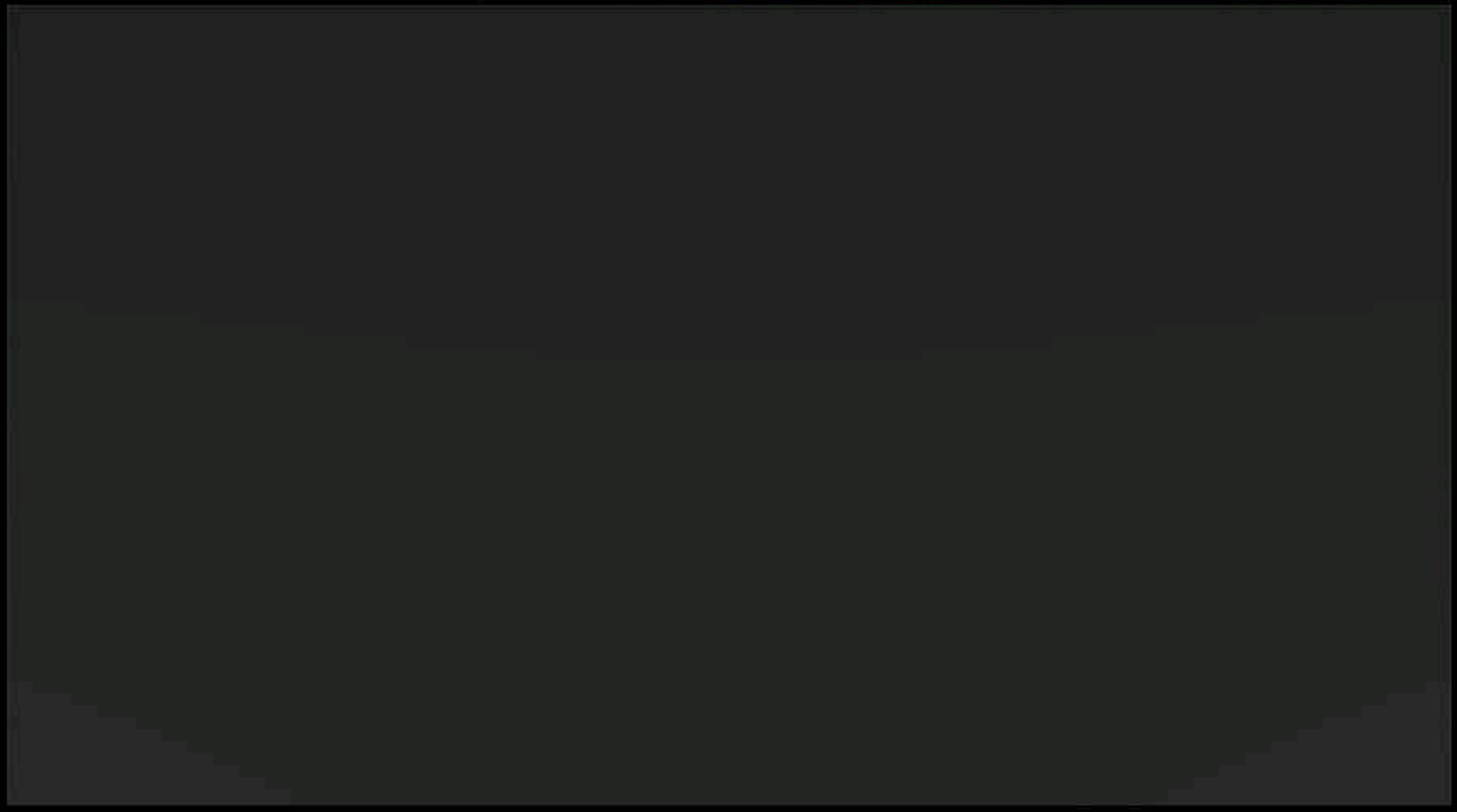


Florian Seiffert, HBZ


Simulation: linear
Zahl der Treffer pro Zelle:
1
2
5
10
20
60
100
200 300
Florian Seiffert, HBZ

Googlebot vom 31/Jul/2002

Zahl der Treffer pro Zelle: 1 2 5 10 20 60 100 200 300

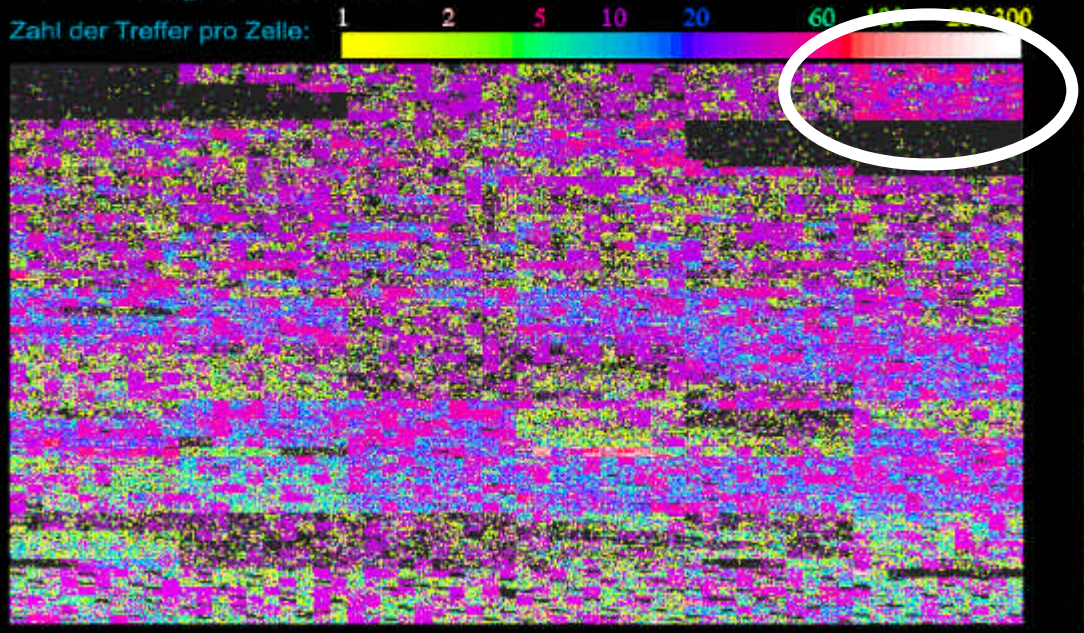


Florian Seiffert, HBZ

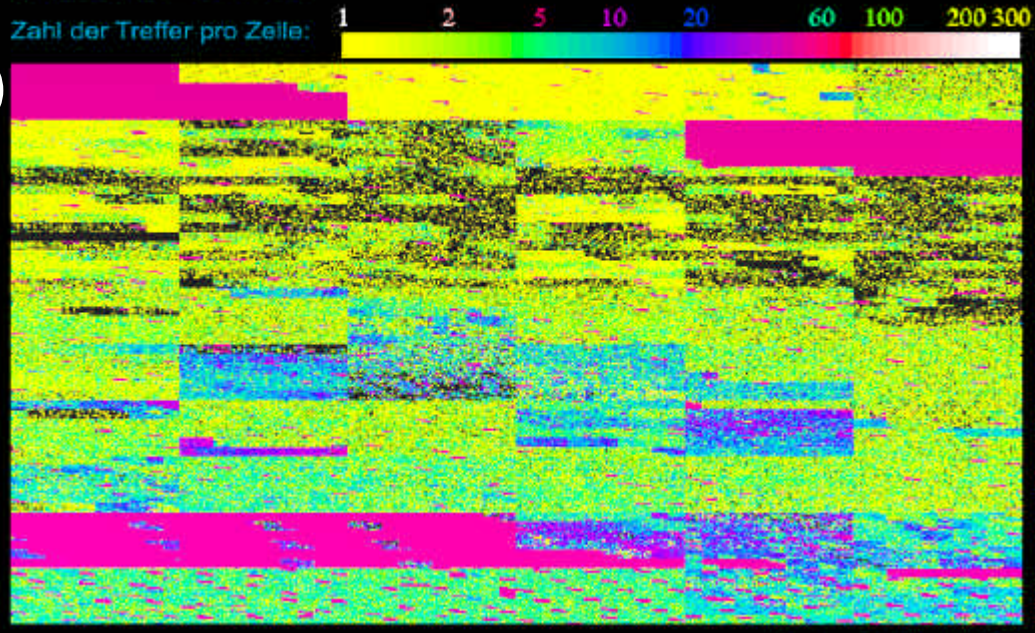
Was haben wir gesehen?

- Keine systematische Indexierung
  - nix: Seite 1, Seite 2, ..., Seite 13 Mio
  - statistische Verfolgung der Links
- Erklaerungsversuche
  - unbenutzte Ids
  - Schlagwoerter

von Google indexiert

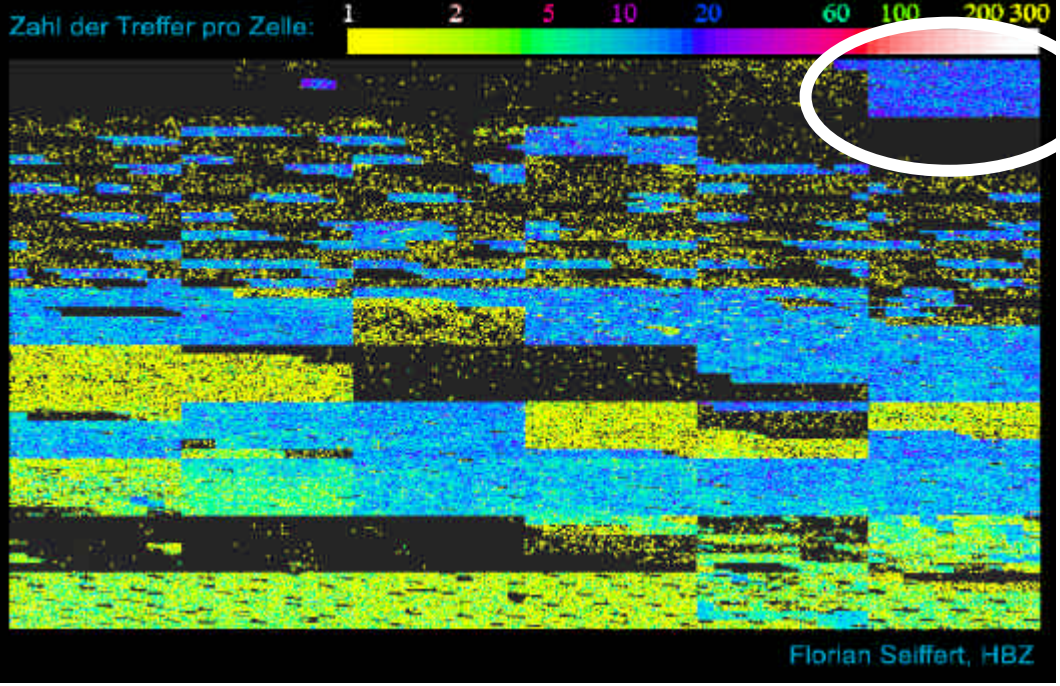


unbenutzte IDs



- unbenutzte Ids
- Schlagwoerter

Titel mit Schlagworten



Schlagworte ein Vorteil?

- Ja!
  - Google-Bots indexieren Seiten
  - mit Schlagworten: 686.561 (18.2%)
  - ohne Schlagworte: 3.090.095 (81.8%)
  - Suchende ueber Google finden Seiten
  - mit Schlagworten: 154.375 (41.7%)
  - ohne Schlagworte: 216.072 (58.3%)
- 9.7% der Titel im Verbund sind verschlagwortet

Schlagworte ein Vorteil?

- **Ja!**
  - Google-Bots indexieren Seiten
  - mit Schlagworten: 686.561 (18.2%)
  - ohne Schlagworte: 3.090.095 (81.8%)

18.2%

  - Suchende ueber Google finden Seiten
  - mit Schlagworten: 154.375 (41.7%)
  - ohne Schlagworte: 216.072 (58.3%)

41.7%

- **9.7% der Titel im Verbund**

gegen 9.7%

Florian Seiffert, HBZ

Was haben wir denn heute so vor ...

- Visualisierung
- Google
  - Indexierungsstrategie
  - Indexierungsleistung
  - Seitenzahlbegrenzung pro Server
  - Dauer bis zur Findbarkeit in der Suche
  - "Tiefe" der Indexierung
  - Dateigroessen
  - Einfluss auf die Besuchshaeufigkeit der Bots
- fast, msn, neofonie
- Was lernen wir daraus

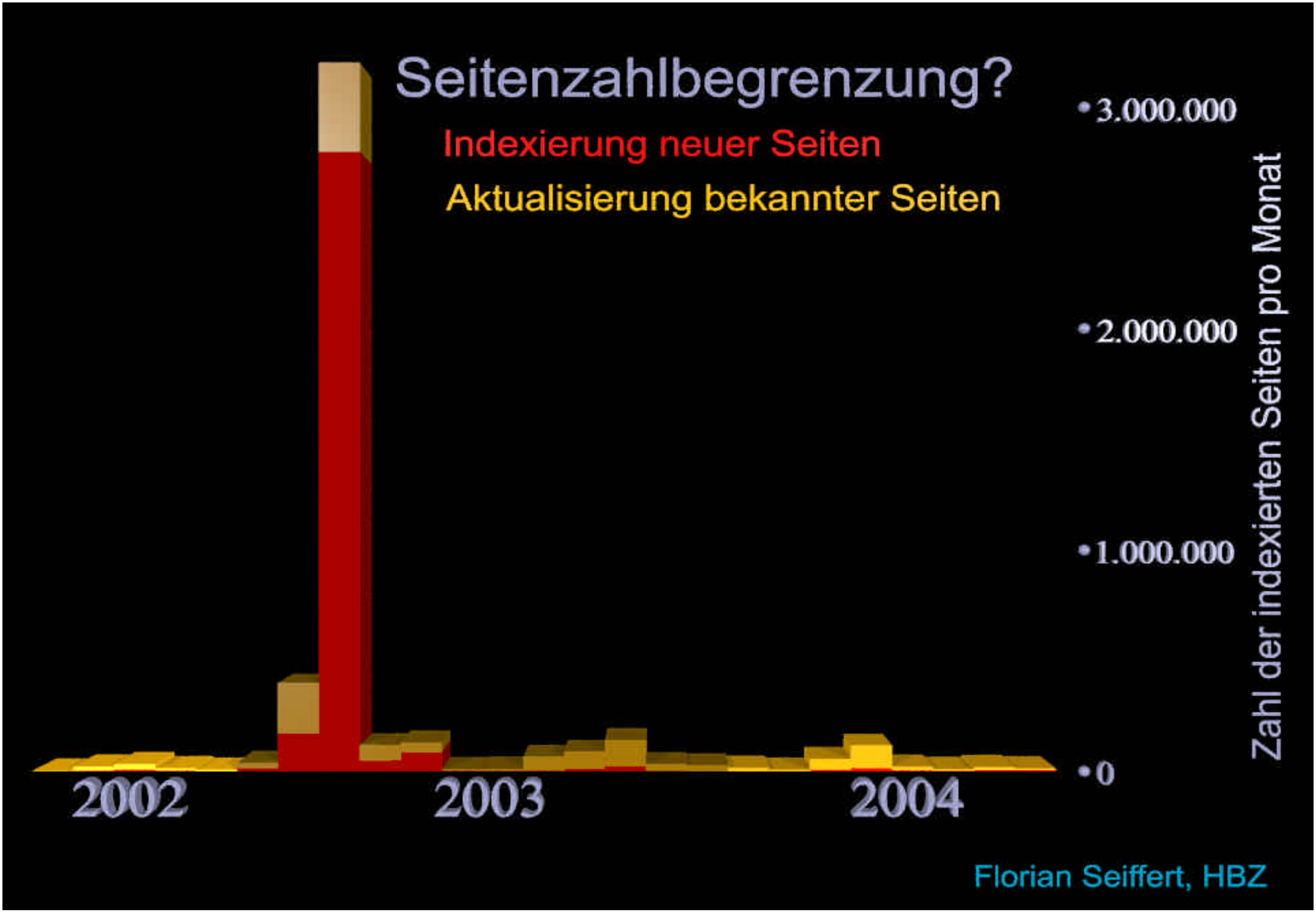
Was haben wir denn heute so vor ...

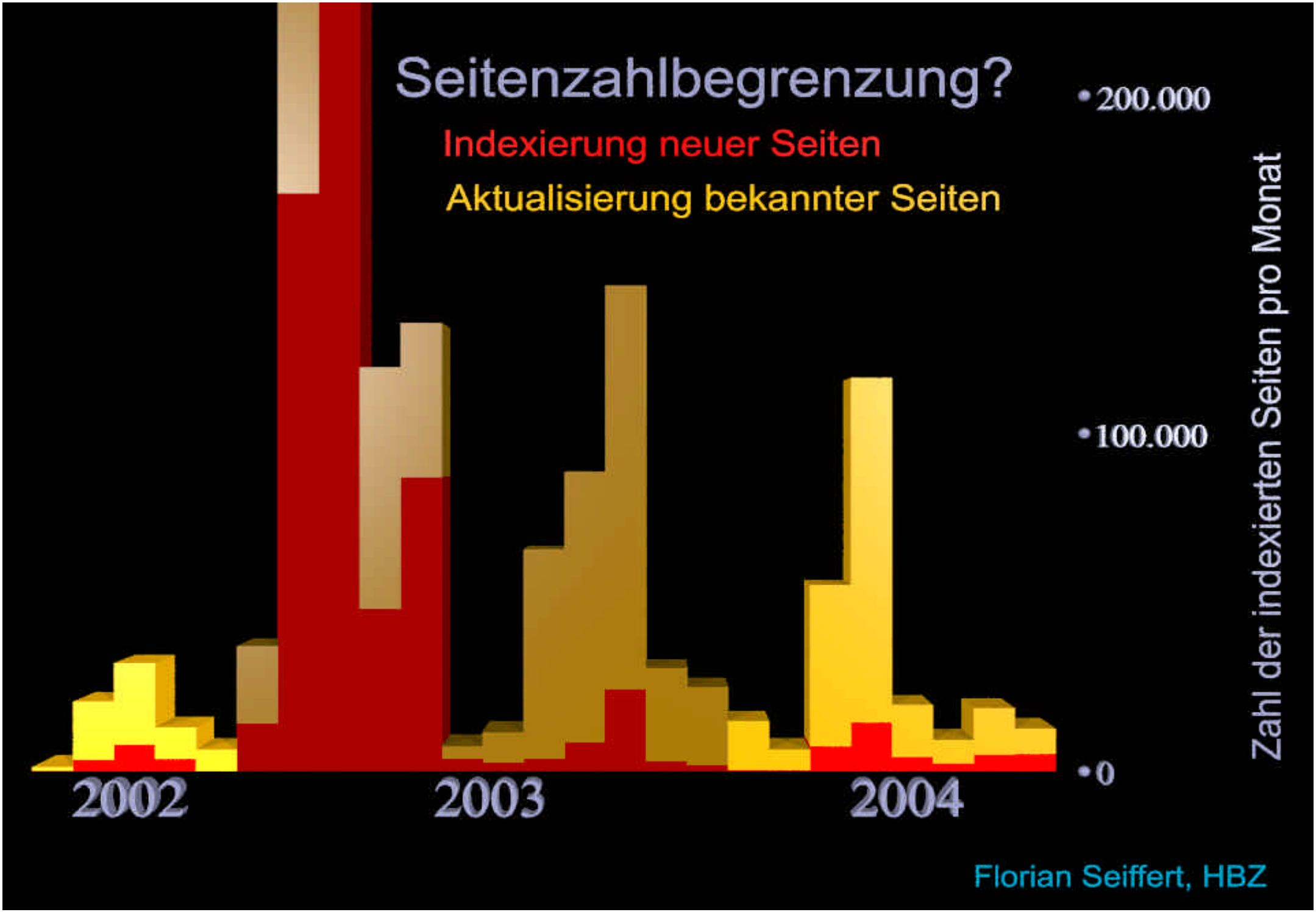
- Visualisierung ✓
- Google ✓
  - Indexierungsstrategie ✓
  - Indexierungsleistung
  - Seitenzahlbegrenzung pro Server
  - Dauer bis zur Findbarkeit in der Suche
  - "Tiefe" der Indexierung
  - Dateigroessen
  - Einfluss auf die Besuchshaeufigkeit der Bots
- fast, msn, neofonie
- Was lernen wir daraus

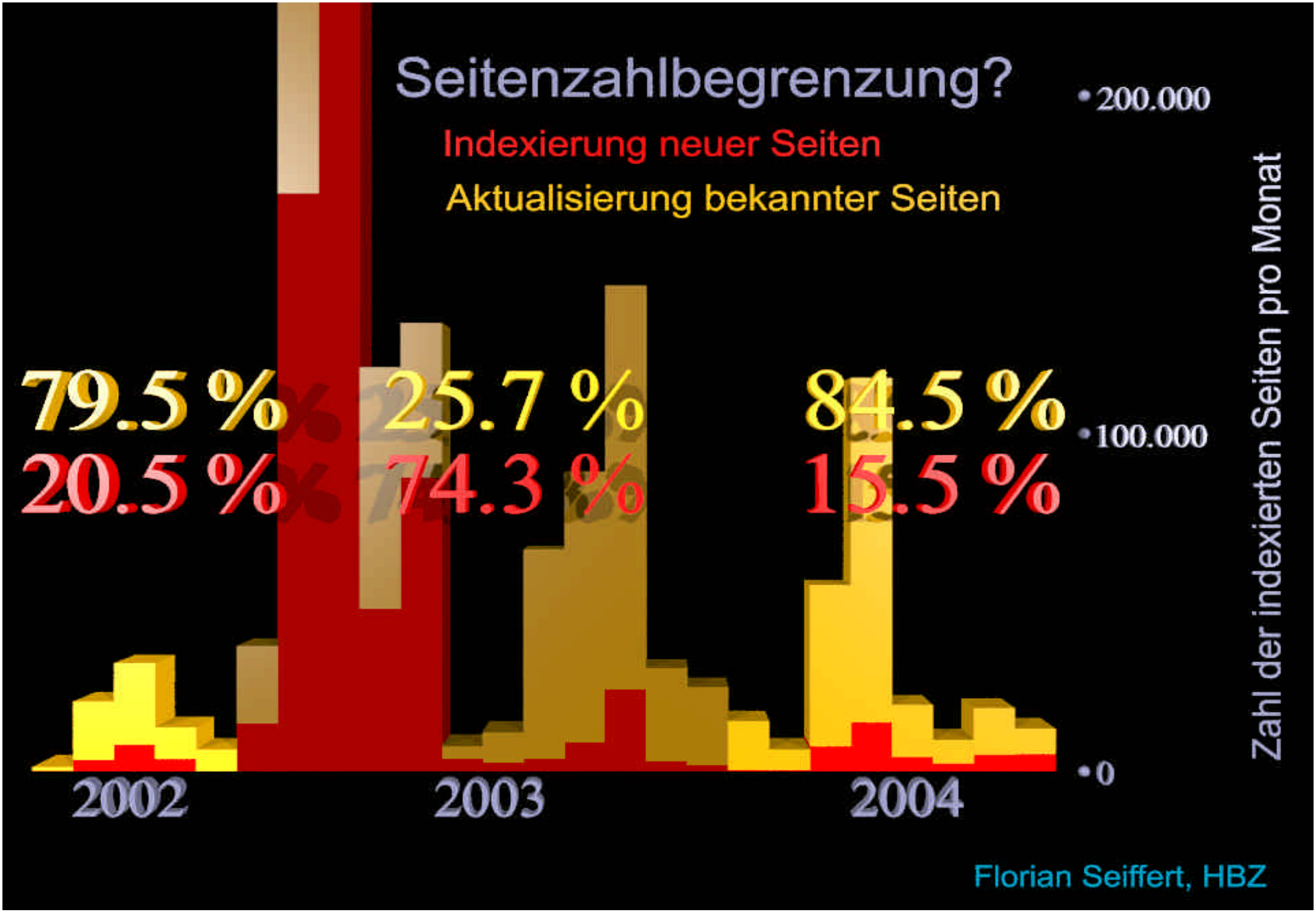
Florian Seiffert, HBZ

Wieviele Seiten?

- wurden von Google im virt. Buecherregal "abgegrast"?
  - insgesamt 3.776.656 (27.3%)
- kennt Google im virt. Buecherregal?
  - insgesamt 3.220.818 (23.3%)
- wurden durch Google-Suchen gefunden?
  - insgesamt 370.447 (2.7%)


Seitenzahlbegrenzung?
Indexierung neuer Seiten
Aktualisierung bekannter Seiten
3.000.000
2.000.000
1.000.000
0
Zahl der indexierten Seiten pro Monat
2002
2003
2004
Florian Seiffert, HBZ


Seitenzahlbegrenzung?
Indexierung neuer Seiten
Aktualisierung bekannter Seiten
200.000
100.000
0
Zahl der indexierten Seiten pro Monat
2002
2003
2004
Florian Seiffert, HBZ


Seitenzahlbegrenzung?
Indexierung neuer Seiten
Aktualisierung bekannter Seiten
79.5 %
20.5 %
25.7 %
74.3 %
84.5 %
15.5 %
2002
2003
2004
200.000
100.000
0
Zahl der indexierten Seiten pro Monat
Florian Seiffert, HBZ

virt. Buecherregal: Eine Hoechstzahl von Seiten die Google pro Server indiziert ist nicht erkennbar.

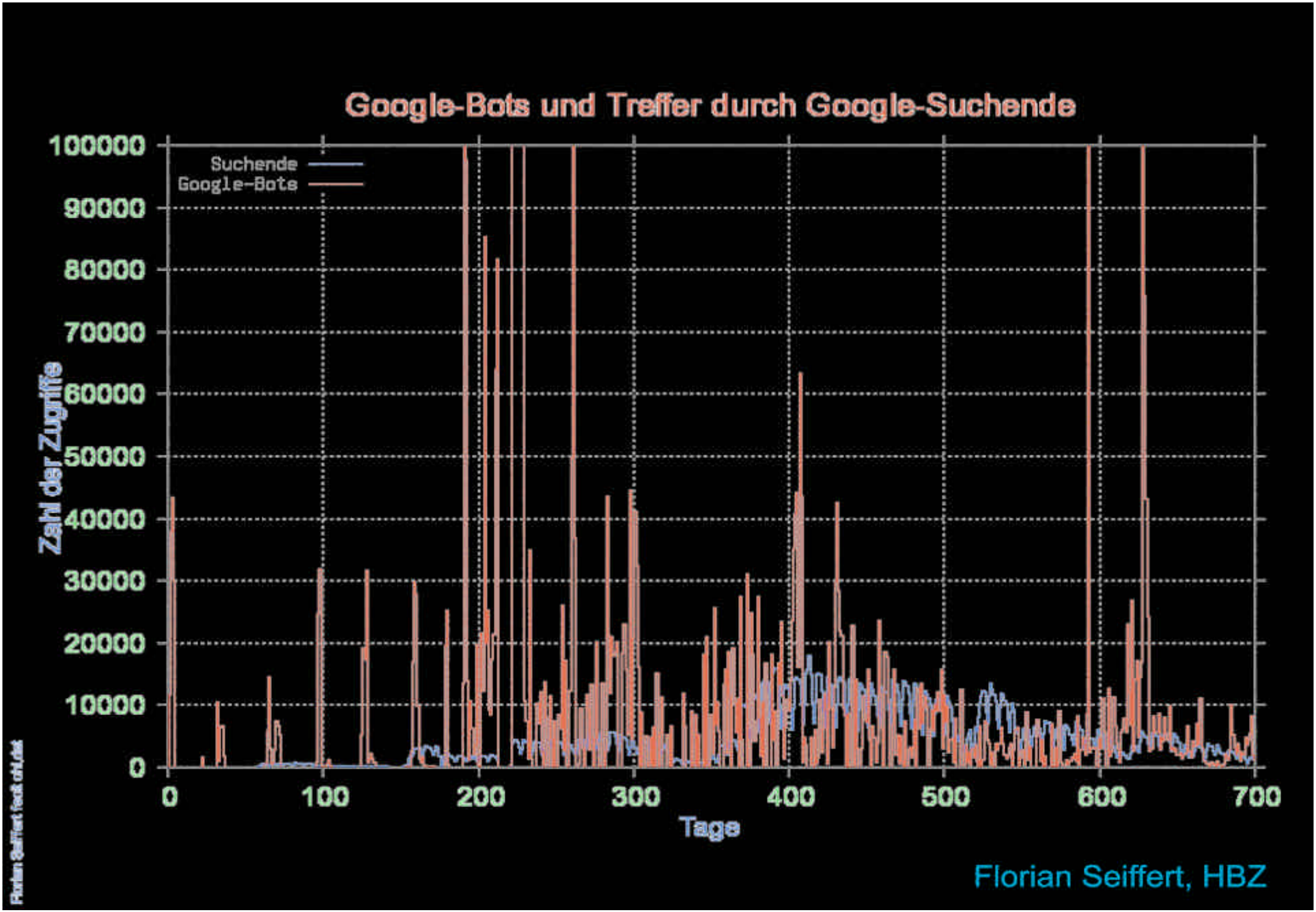
Google kennt z.Zt. 3.2 Mio Seiten im virt. Buecherregal

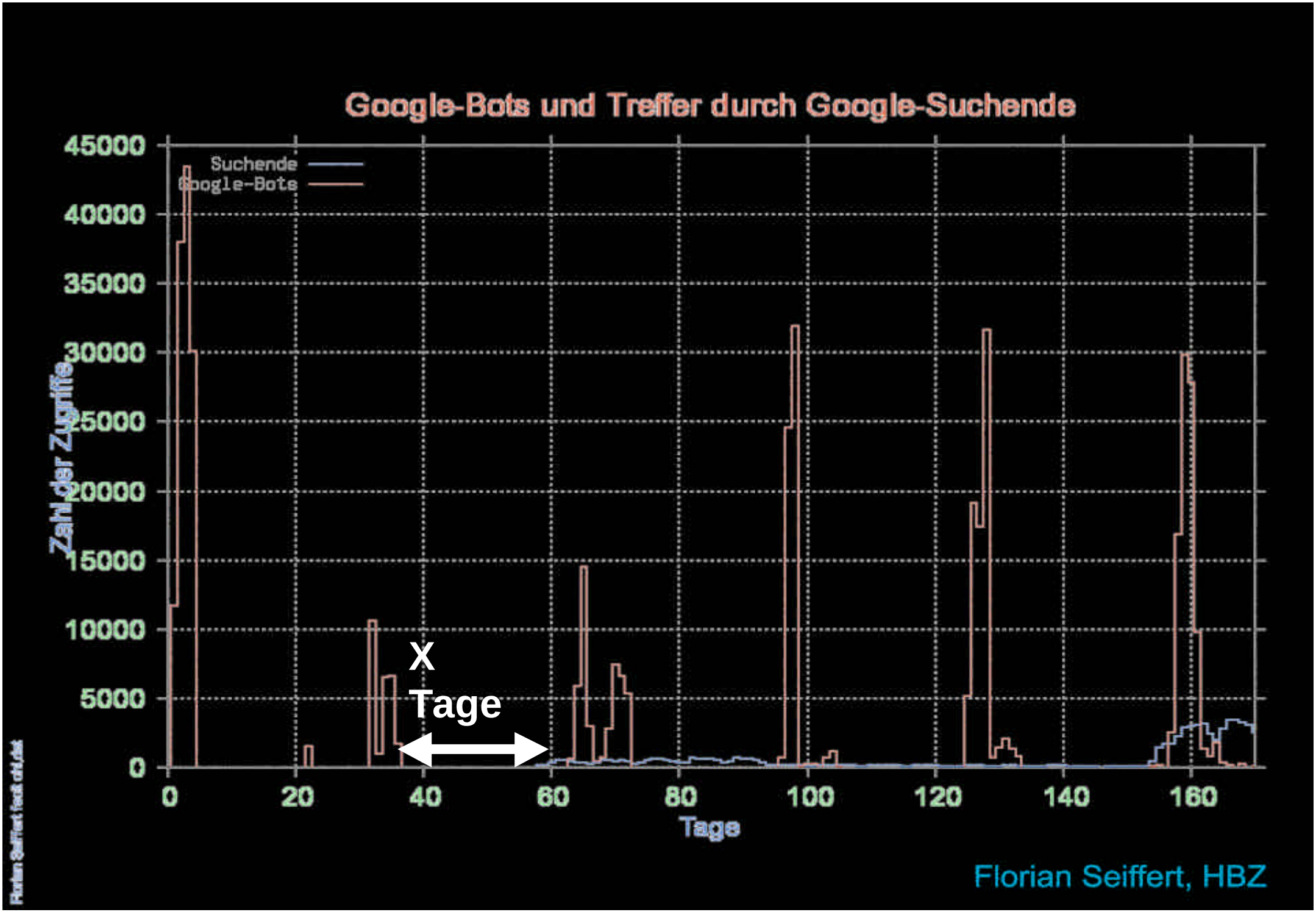
Florian Seiffert, HBZ

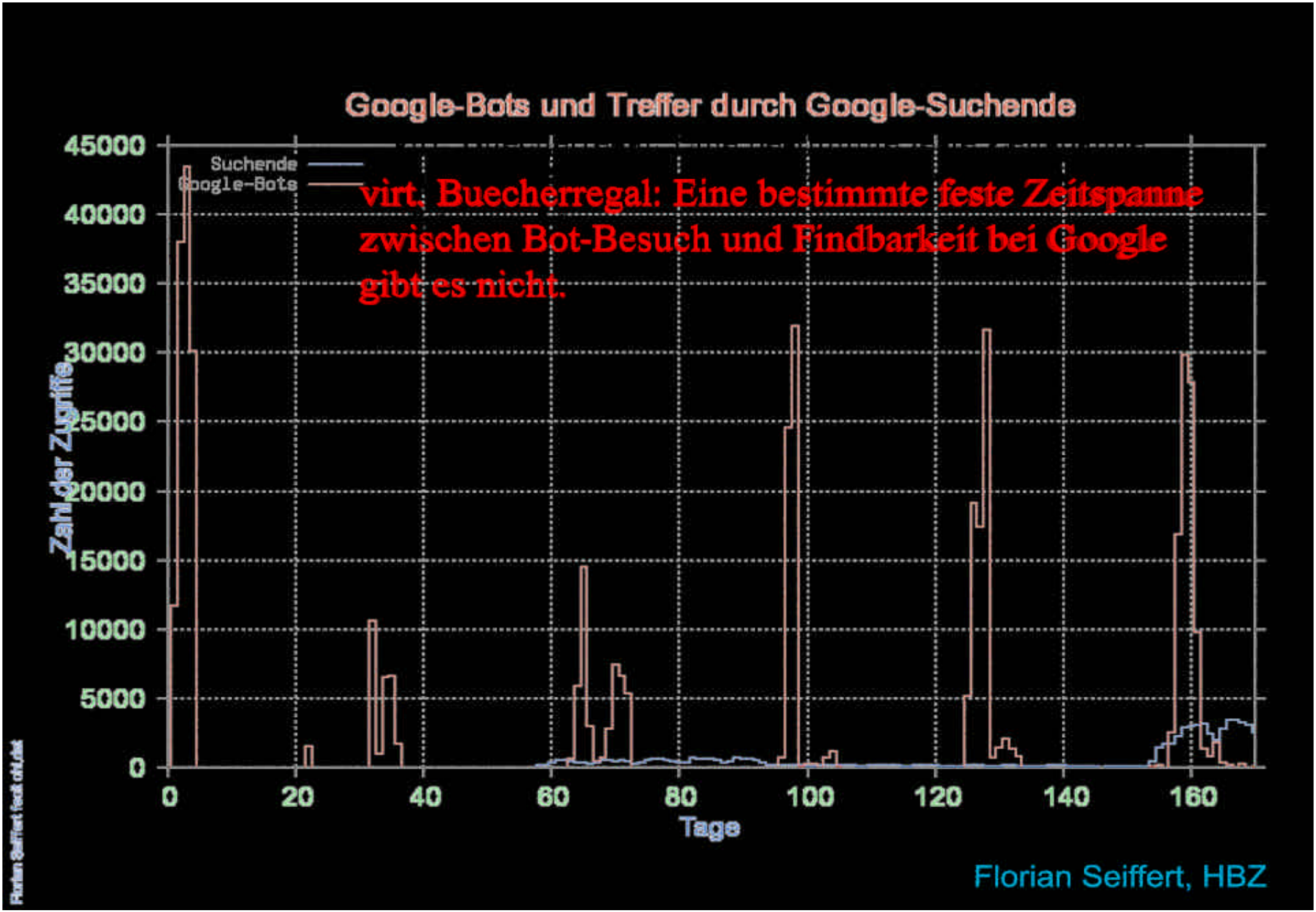
Was haben wir denn heute so vor ...

- Visualisierung ✓
- Google ✓
  - Indexierungsstrategie ✓
  - Indexierungsleistung ✓
  - Seitenzahlbegrenzung pro Server ✓
  - Dauer bis zur Findbarkeit in der Suche
  - "Tiefe" der Indexierung
  - Dateigroessen
  - Einfluss auf die Besuchshaeufigkeit der Bots
- fast, msn, neofonie
- Was lernen wir daraus

Florian Seiffert, HBZ


Google-Bots und Treffer durch Google-Suchende
Suchende
Google-Bots
100000
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Zahl der Zugriffe
0
100
200
300
400
500
600
700
Tage
Florian Seiffert, HBZ


Google-Bots und Treffer durch Google-Suchende
Suchende
Google-Bots
Zahl der Zugriffe
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
120
140
160
Tage
X
Tage
Florian Seiffert, HBZ


Google-Bots und Treffer durch Google-Suchende
Suchende
Google-Bots
virt. Buecherregal: Eine bestimmte feste Zeitspanne zwischen Bot-Besuch und Findbarkeit bei Google gibt es nicht.
Zahl der Zugriffe
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
20
40
60
80
100
120
140
160
Tage
Florian Seiffert, HBZ

"Tiefe" der Indexierung

- Geruecht:
  - Google indexiert nur bis zu einer bestimmten Tiefe.
  - Stimmt das aus Sicht des virt. Buecherregals?
- url:
  - http://kirke.hbz-nrw.de/
  - Stufe 1: dcb/
  - Stufe 2: Alle_003/
  - Stufe 3: Buecher_13/
  - Stufe 4: in_NRW_13/000695592.html

"Tiefe" der Indexierung

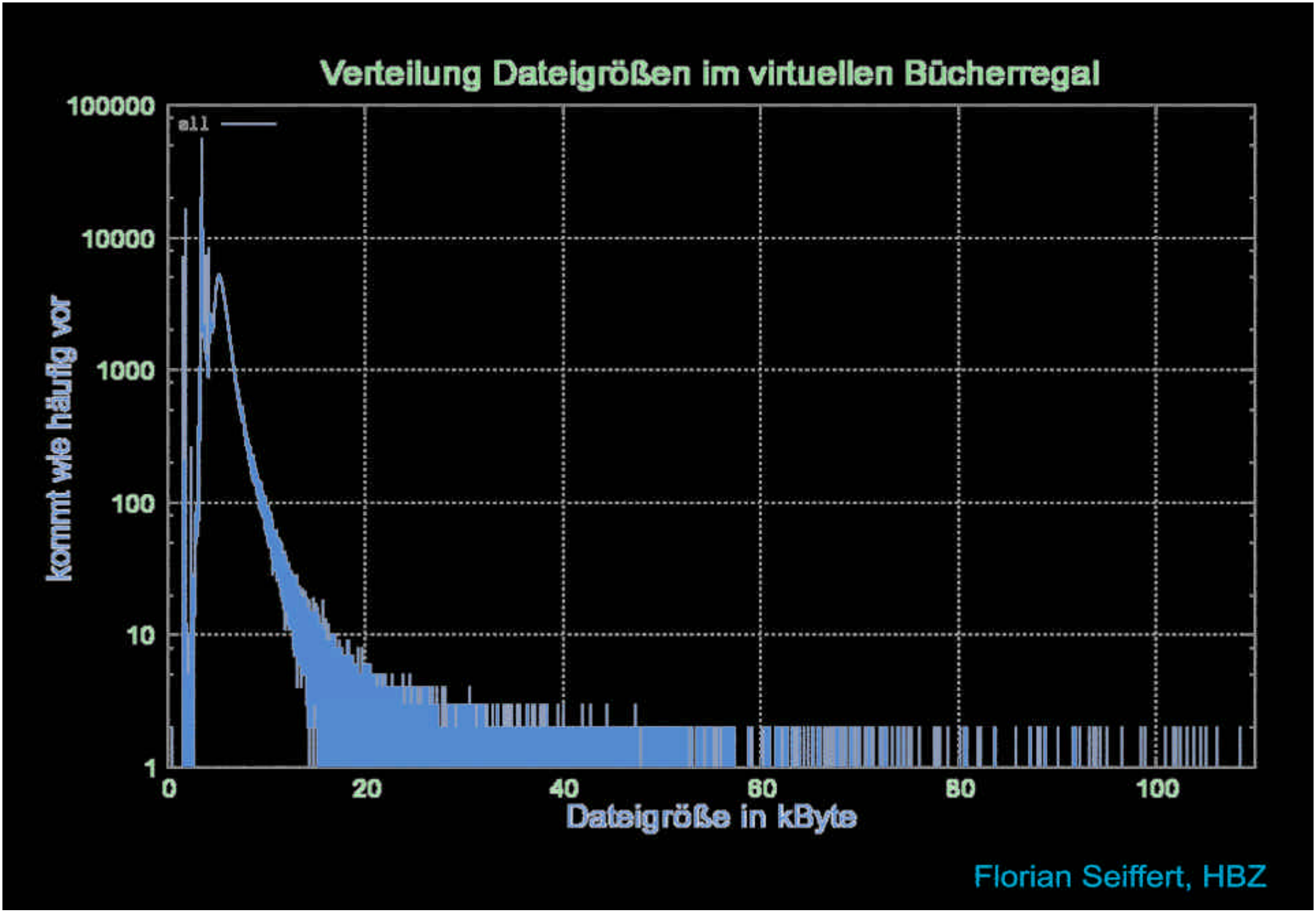
- Geruecht:
  - Google indexiert nur bis zu einer bestimmten Tiefe.
  - Stimmt das aus Sicht des virt. Buecherregals?
- url:
  - http://kirke.hbz-nrw.de/
  - Stufe 1: dcb/
  - Stufe 2: Alle_003/
  - Stufe 3: Buecher_13/
  - Stufe 4: in_NRW_13/000695592.html

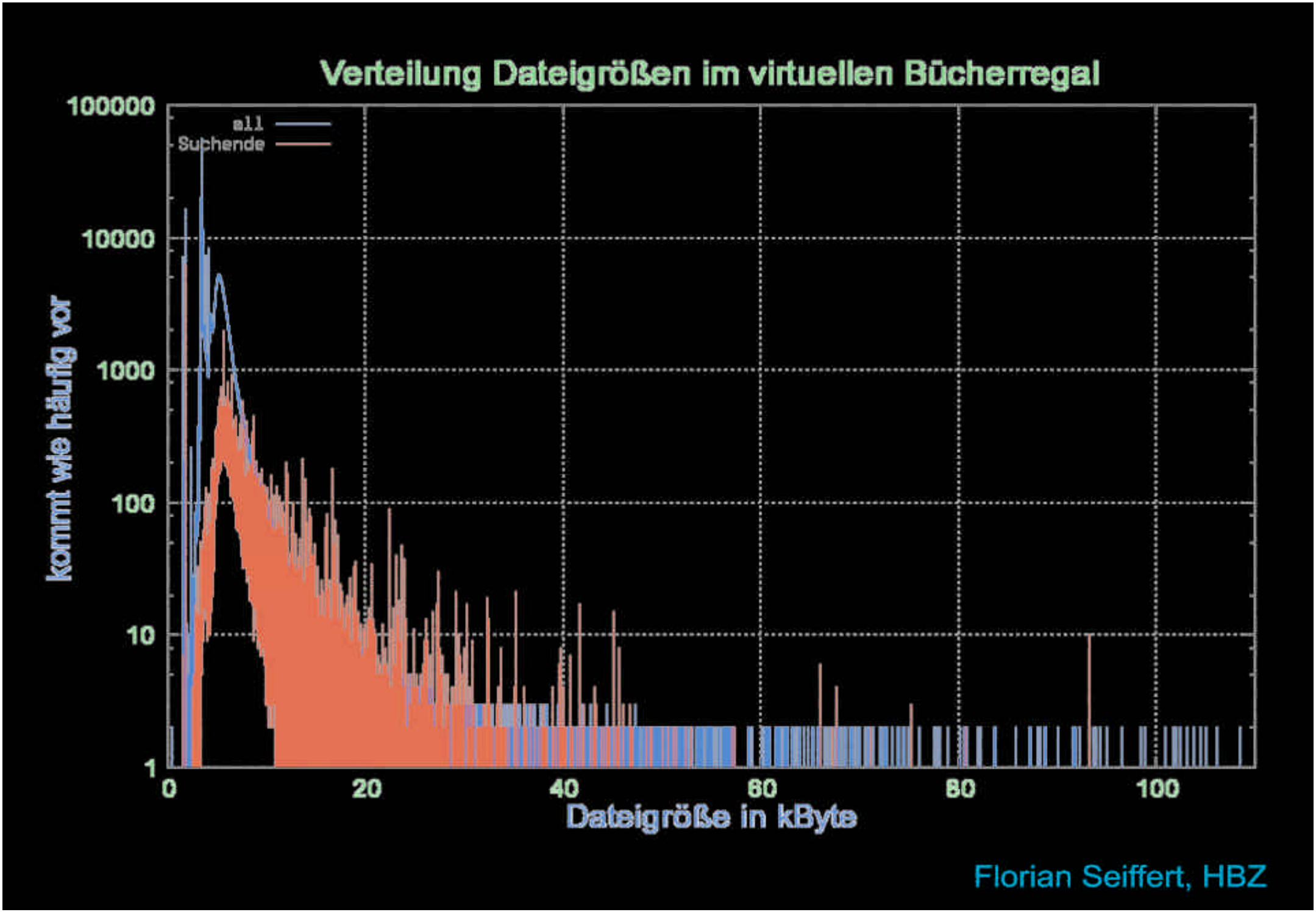
virt. Buecherregal: Geruecht stimmt nicht.

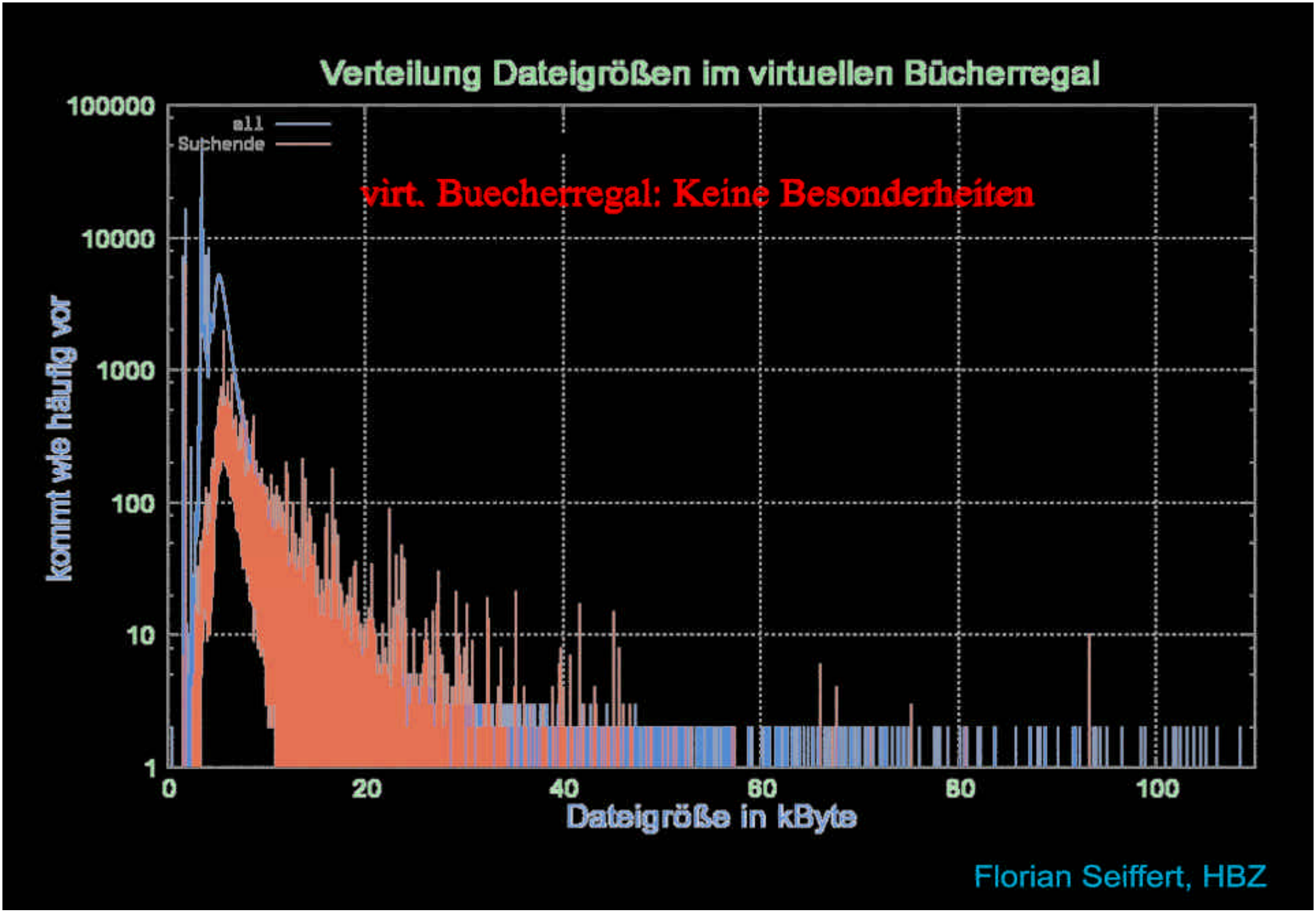
Was haben wir denn heute so vor ...

- Visualisierung ✓
- Google ✓
  - Indexierungsstrategie ✓
  - Indexierungsleistung ✓
  - Seitenzahlbegrenzung pro Server ✓
  - Dauer bis zur Findbarkeit in der Suche ✓
  - "Tiefe" der Indexierung ✓
  - Dateigroessen
  - Einfluss auf die Besuchshaeufigkeit der Bots
- fast, msn, neofonie
- Was lernen wir daraus

Florian Seiffert, HBZ


Verteilung Dateigrößen im virtuellen Bücherregal
all
100000
10000
1000
100
10
1
kommt wie häufig vor
0
20
40
60
80
100
Dateigröße in kByte
Florian Seiffert, HBZ


Verteilung Dateigrößen im virtuellen Bücherregal
all
Suchende
100000
10000
1000
100
10
1
kommt wie häufig vor
0
20
40
60
80
100
Dateigröße in kByte
Florian Seiffert, HBZ


Verteilung Dateigrößen im virtuellen Bücherregal
all
Suchende
virt. Buecherregal: Keine Besonderheiten
kommt wie häufig vor
100000
10000
1000
100
10
1
0
20
40
60
80
100
Dateigröße in kByte
Florian Seiffert, HBZ

Was haben wir denn heute so vor ...

- Visualisierung ✓
- Google ✓
  - Indexierungsstrategie ✓
  - Indexierungsleistung ✓
  - Seitenzahlbegrenzung pro Server ✓
  - Dauer bis zur Findbarkeit in der Suche ✓
  - "Tiefe" der Indexierung ✓
  - Dateigroessen ✓
  - **Einfluss auf die Besuchshaeufigkeit der Bots**
- **fast, msn, neofonie**
- **Was lernen wir daraus**

Florian Seiffert, HBZ

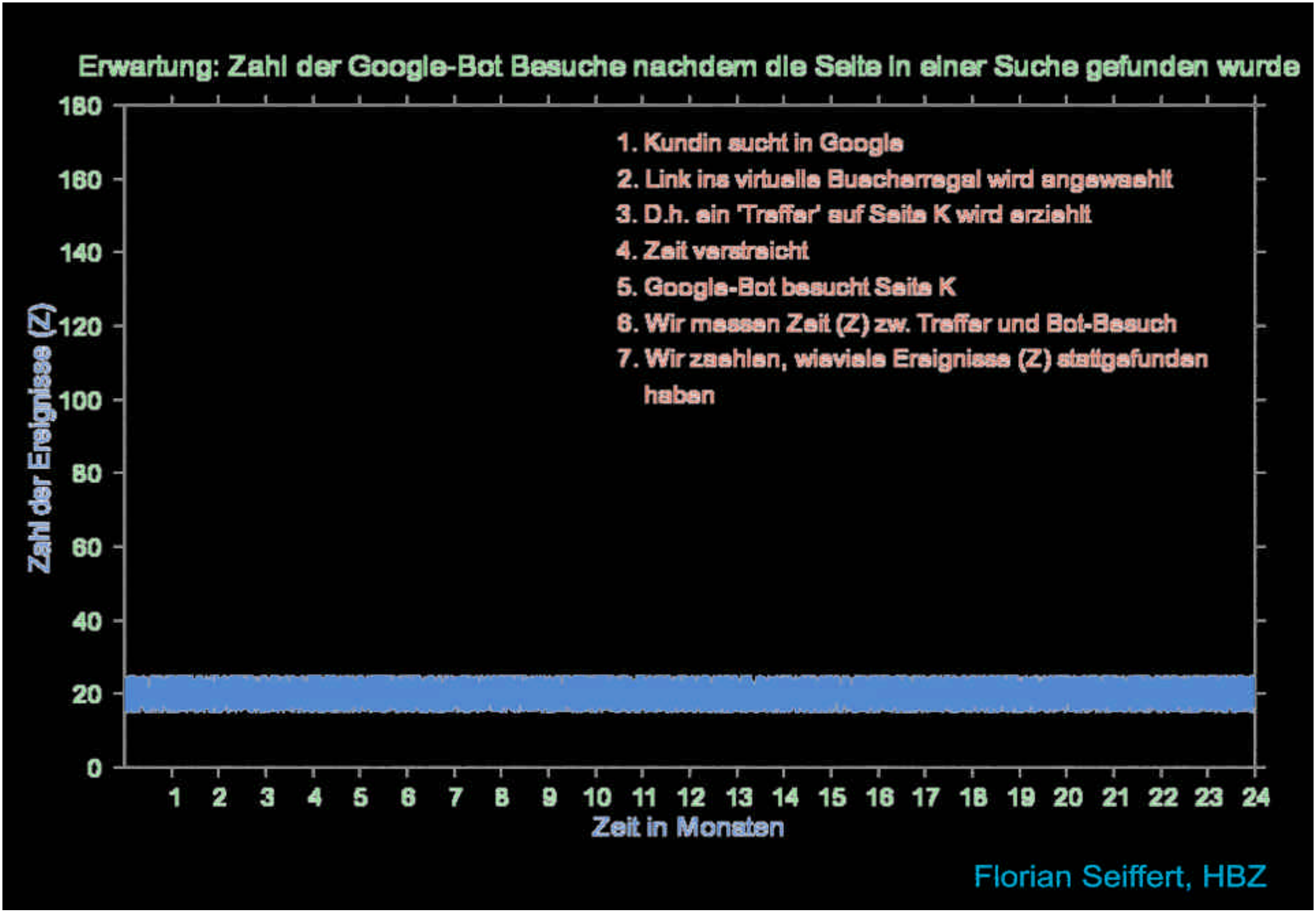
Bots:

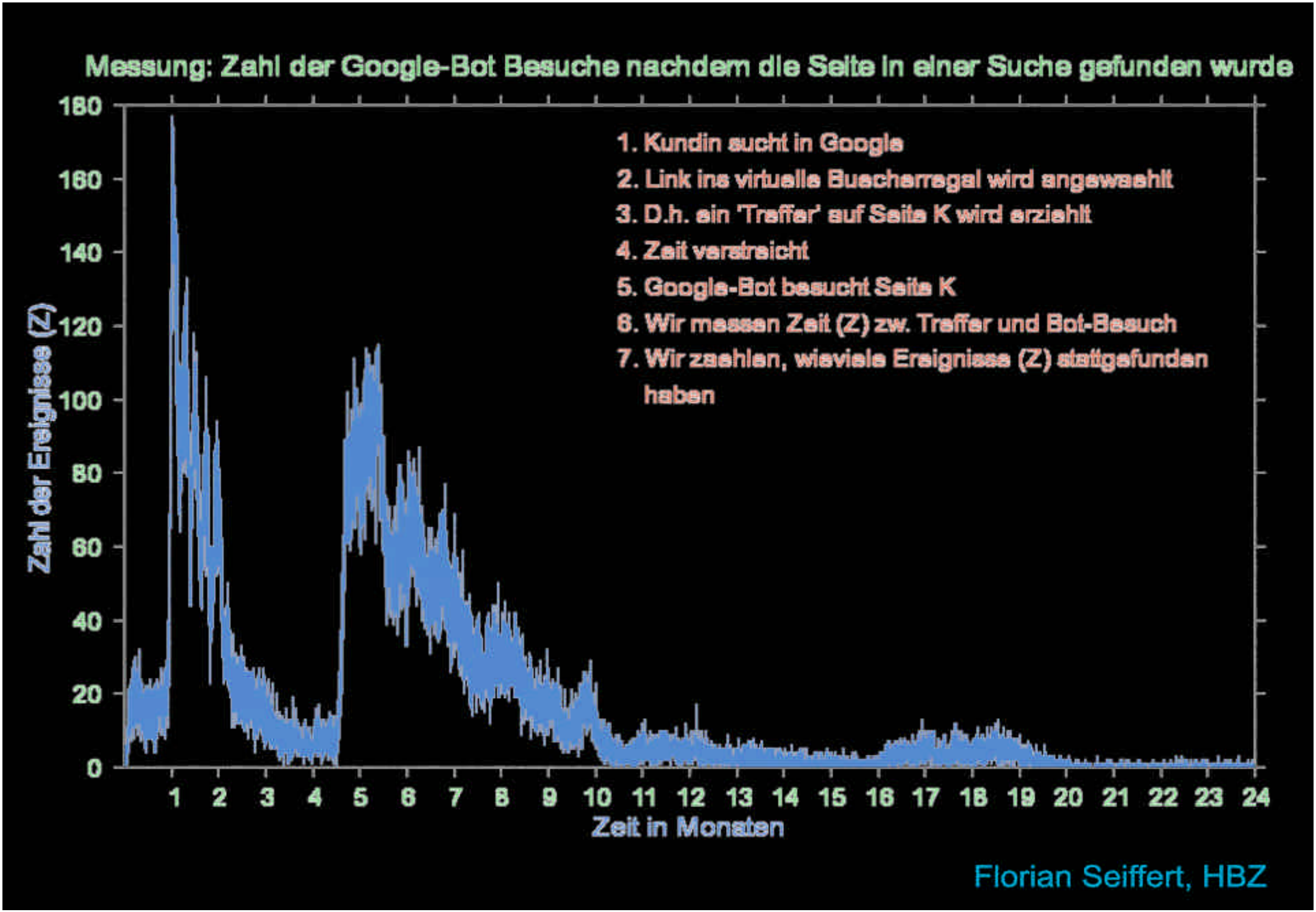
- Einfluss auf die Besuchshaeufigkeit?

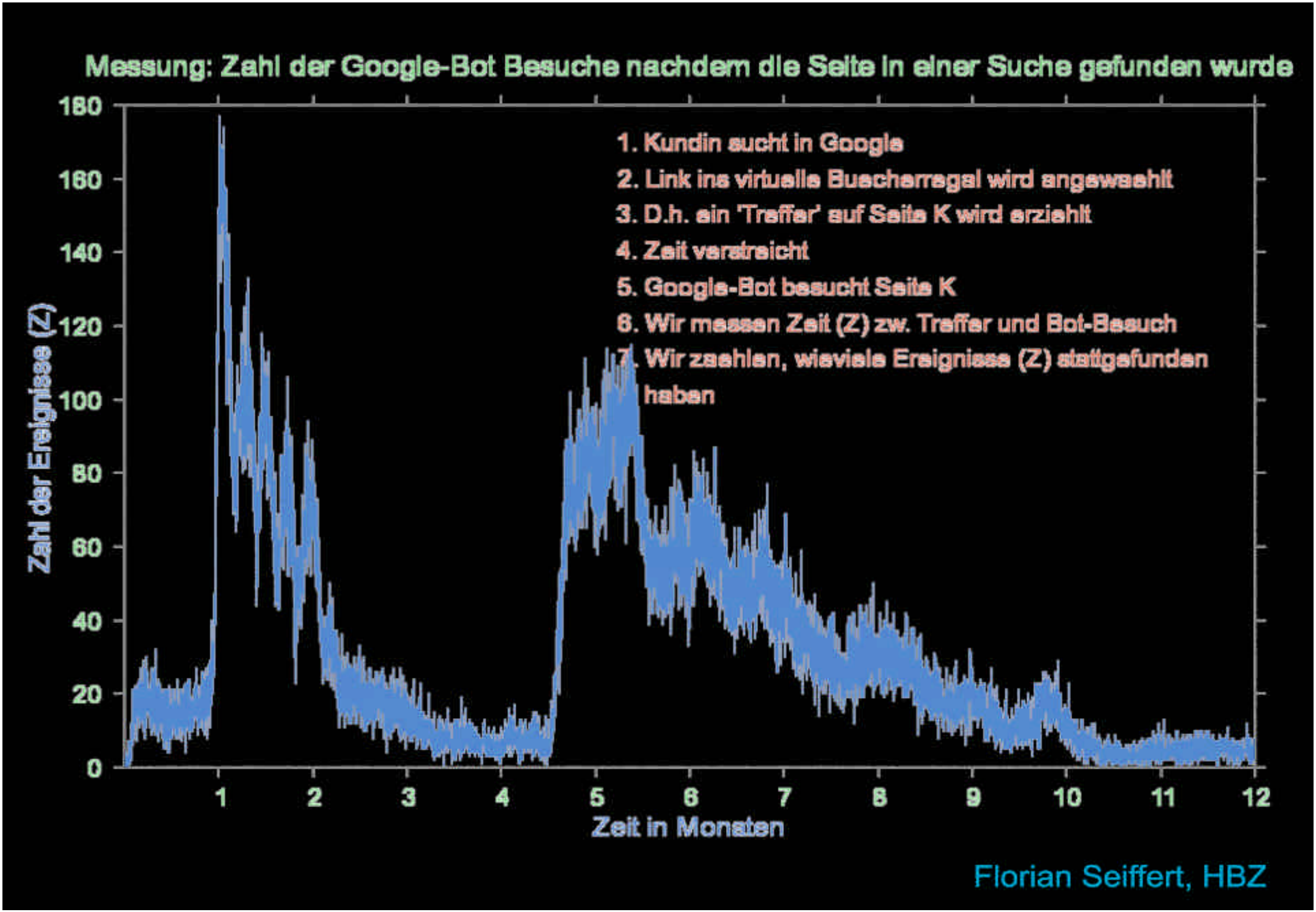
- Google sagt:
  - Alle 4 Mrd. Seiten werden einmal pro Monat indexiert

- virtuelles Buecherregal:
  - Wir pruefen den zeitl. Abstand zwischen einem Treffer nach einer Suche und dem folgenden Bot-Besuch.

- Was erwarten wir?


Erwartung: Zahl der Google-Bot Besuche nachdem die Seite in einer Suche gefunden wurde
1. Kundin sucht in Google
2. Link ins virtuelle Buecherregal wird angewaehlt
3. D.h. ein 'Treffer' auf Seite K wird erzielt
4. Zeit verstreicht
5. Google-Bot besucht Seite K
6. Wir messen Zeit (Z) zw. Treffer und Bot-Besuch
7. Wir zaehlen, wieviele Ereignisse (Z) stattgefunden haben
Zahl der Ereignisse (Z)
0
20
40
60
80
100
120
140
160
180
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Zeit in Monaten
Florian Seiffert, HBZ


Messung: Zahl der Google-Bot Besuche nachdem die Seite in einer Suche gefunden wurde
1. Kundin sucht in Google
2. Link ins virtuelle Buecherregal wird angewaehlt
3. D.h. ein 'Treffer' auf Seite K wird erzielt
4. Zeit verstreicht
5. Google-Bot besucht Seite K
6. Wir messen Zeit (Z) zw. Treffer und Bot-Besuch
7. Wir zaehlen, wieviele Ereignisse (Z) stattgefunden haben
Zahl der Ereignisse (Z)
0
20
40
60
80
100
120
140
160
180
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Zeit in Monaten
Florian Seiffert, HBZ


Messung: Zahl der Google-Bot Besuche nachdem die Seite in einer Suche gefunden wurde
1. Kundin sucht in Google
2. Link ins virtuelle Buecherregal wird angewaehlt
3. D.h. ein 'Treffer' auf Seite K wird erziehlt
4. Zeit verstreicht
5. Google-Bot besucht Seite K
6. Wir messen Zeit (Z) zw. Treffer und Bot-Besuch
7. Wir zaehlen, wieviele Ereignisse (Z) stattgefunden haben
Zahl der Ereignisse (Z)
0
20
40
60
80
100
120
140
160
180
1
2
3
4
5
6
7
8
9
10
11
12
Zeit in Monaten
Florian Seiffert, HBZ

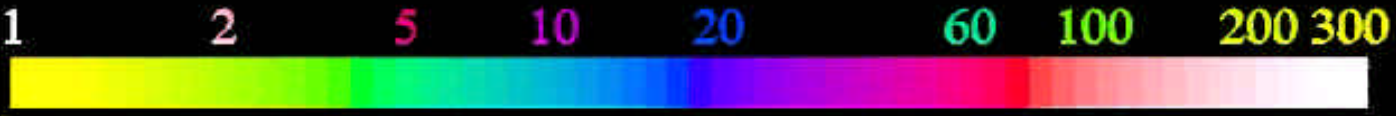
Was haben wir denn heute so vor ...

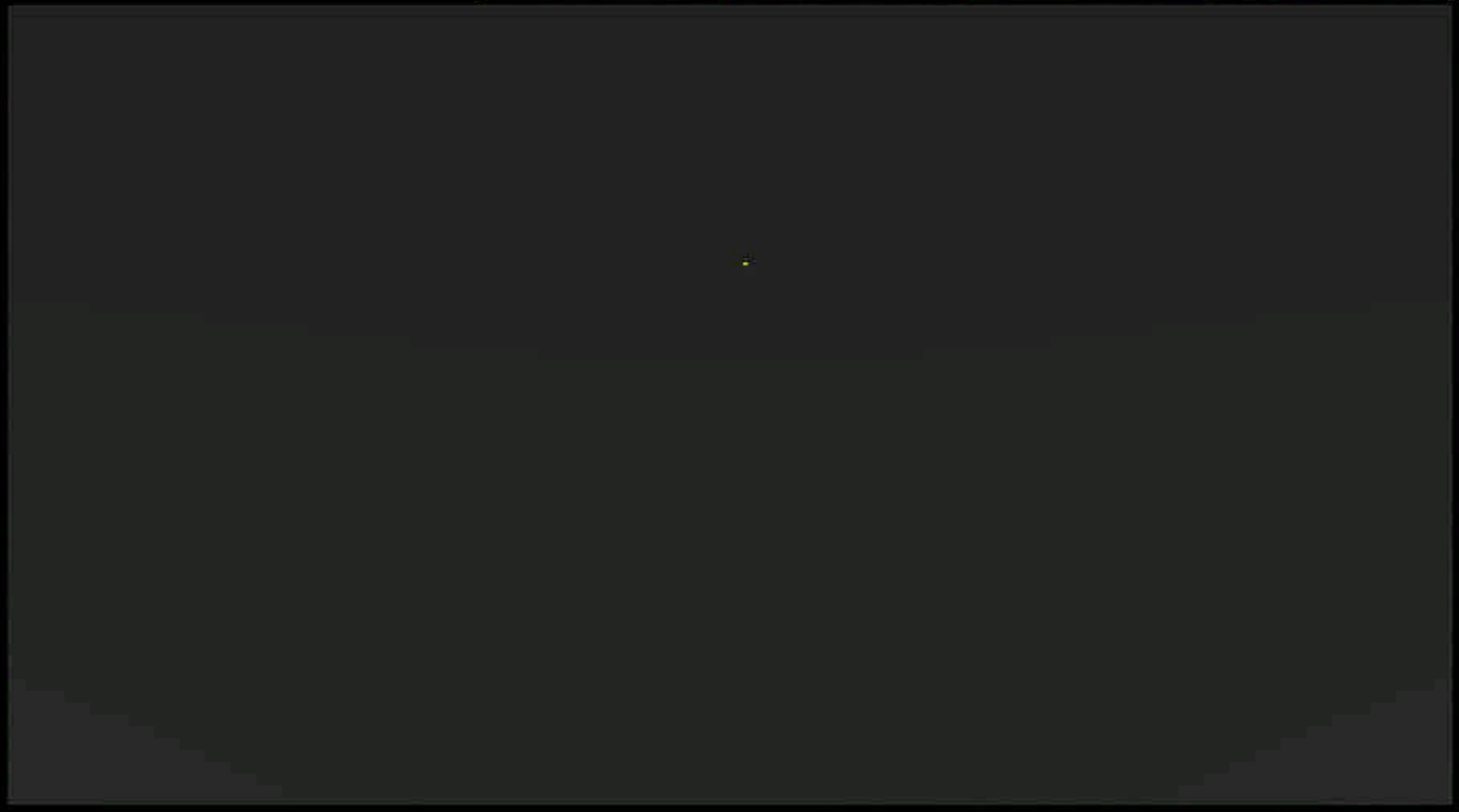
- Visualisierung ✓
- Google ✓
  - Indexierungsstrategie ✓
  - Indexierungsleistung ✓
  - Seitenzahlbegrenzung pro Server ✓
  - Dauer bis zur Findbarkeit in der Suche ✓
  - "Tiefe" der Indexierung ✓
  - Dateigroessen ✓
  - Einfluss auf die Besuchshaeufigkeit der Bots ✓
- fast, msn, neofonie
- Was lernen wir daraus

Florian Seiffert, HBZ

FAST-WebCrawler 16/Jul/2002

Zahl der Treffer pro Zelle:

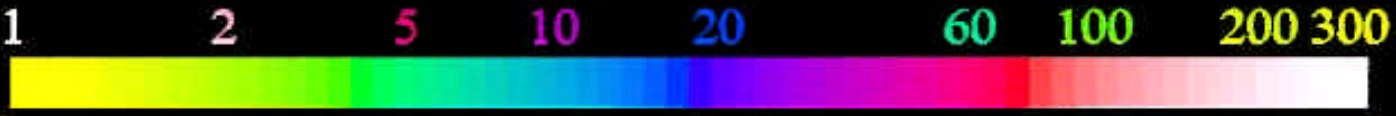


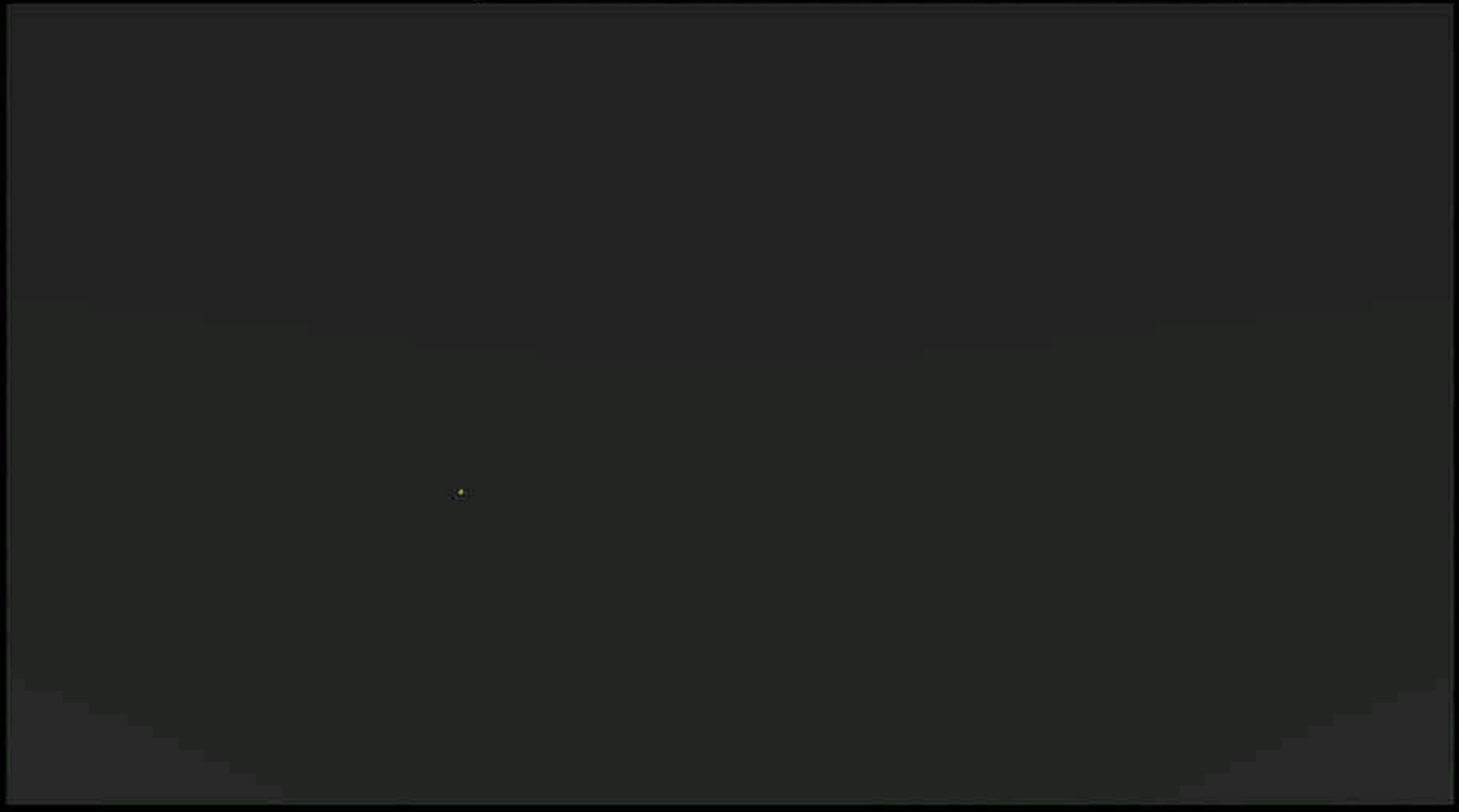


Florian Seiffert, HBZ

msnbot 22/Jun/2003

Zahl der Treffer pro Zelle:





Florian Seiffert, HBZ

neofonie 20/Apr/2004

Zahl der Treffer pro Zelle: 1 2 5 10 20 60 100 200 300

Florian Seiffert, HBZ

Was lernen wir daraus

- Google
  - indexiert statistisch nicht systematisch,
  - findet Seiten mit Schlagworten haeufiger,
  - speichert pro Server auch mehr als 3.2 Mio Seiten,
  - indexiert mindestens 4 Hierarchiestufen tief,
  - aktualisiert Seiten, die gesucht werden haeufiger,
- Suchen Sie nach Ihren Seiten !!

msn
Google

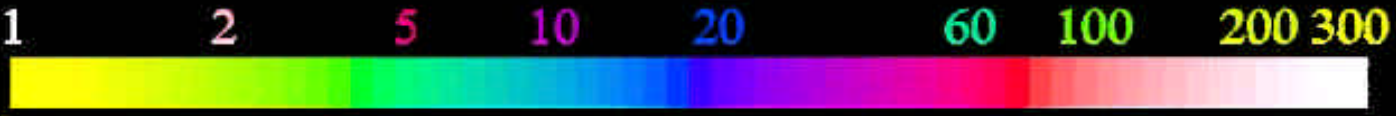
Fragen?

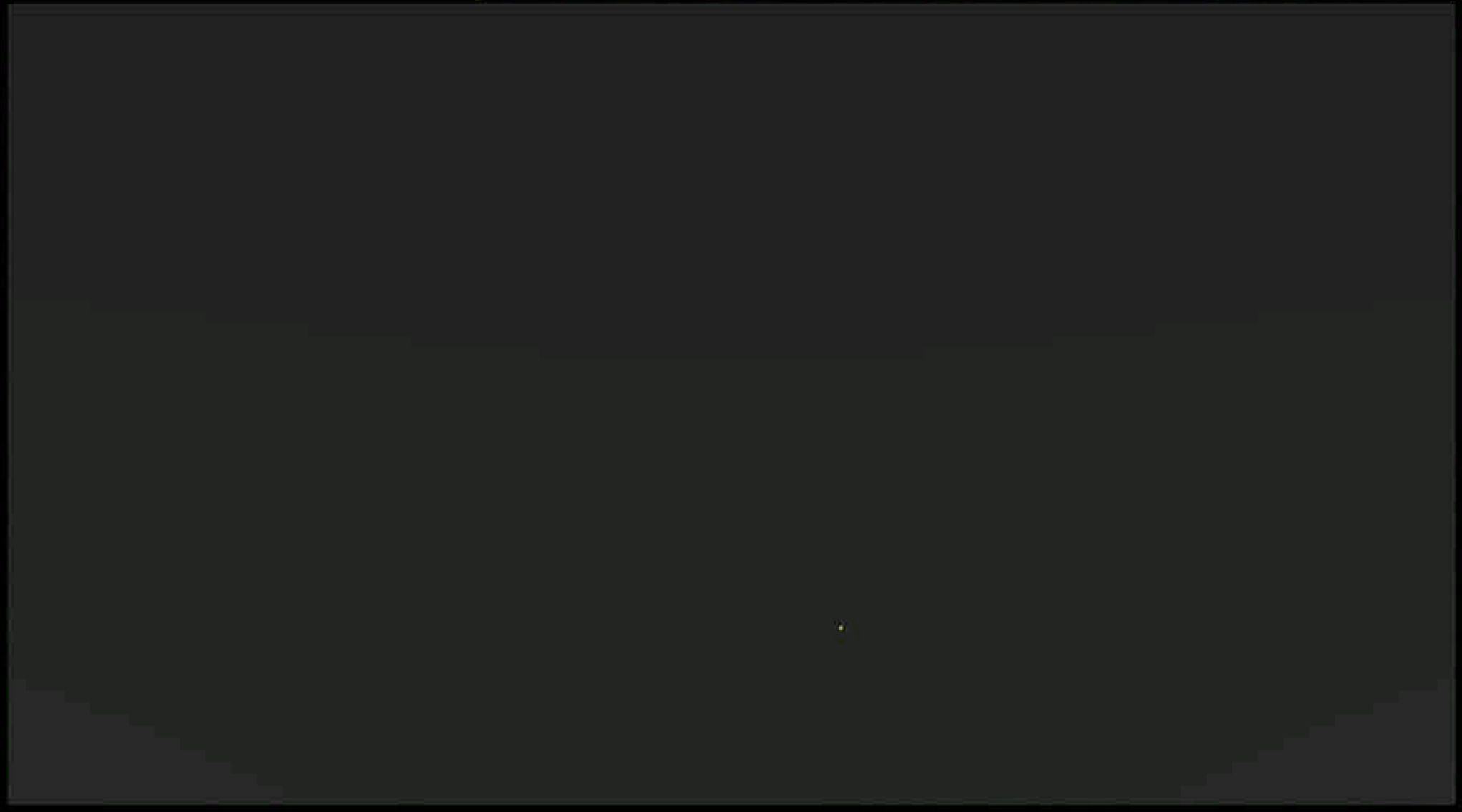
Werkzeugkasten

- Linux (http://www.suse.de)
  - perl (http://www.perl.com)
  - mysql (http://www.mysql.com/)
  - mencoder / mplayer (http://www.mplayerhq.hu)
  - Persistence of Vision Raytracer (http://www.povray.org)
- Windows
  - Powerpoint
- Vortrag unter
  - http://www.Florian-Seiffert.de/2004/Bonn/Inetbib2004.pdf

"von google kommend" vom 22/Aug/2002

Zahl der Treffer pro Zelle:





Florian Seiffert, HBZ