Data Sampling in Analytics

Sieben Möglichkeiten es zu reduzieren

METRIKA
WEB ANALYTICS | BERATUNG | SEMINARE

In diesem Beitrag geht es um eines der größten Ärgernisse, das gleichzeitig auch eine der größten Hilfen in der Webanalyse darstellt: das Data Sampling.

Was ist das? Wann tritt es auf? Und wie kannst du es einschränken? Dazu habe ich hier einige – genauer gesagt sieben – Tipps für dich.

Was du in diesem Beitrag findest

- Data Sampling kennenlernen
- Wann wird Sampling aktiv
- Folgen des Data Sampling
- Sieben Möglichkeiten, das Sampling zu unterbinden
- Fazit

Data Sampling kennenlernen

Wer das erste Mal in Google Analytics auf das grüne oder gelbe Schild-Symbol mit dem Haken trifft, ist zunächst einmal irritiert. Bei den meisten Menschen bleibt es erst einmal unbeachtet, obwohl es an relativ prominenter Stelle steht, nämlich am oberen Ende der Reports, zumeist oberhalb der gewählten Segmente (siehe Abbildung 1).

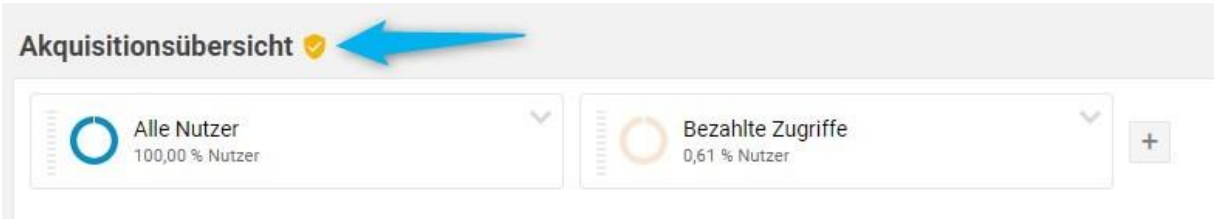

Abbildung 1: Das Symbol für aktives Data Sampling ist gelb

Dieses Schild symbolisiert, ob das sogenannte Sampling – also eine Stichprobenerhebung innerhalb der Daten – aktiv ist (gelbes Symbol) oder nicht (grünes Symbol). Wenn du mit der Maus über das Symbol fährst, kannst du weitere Details erfahren (siehe Abbildung 2).

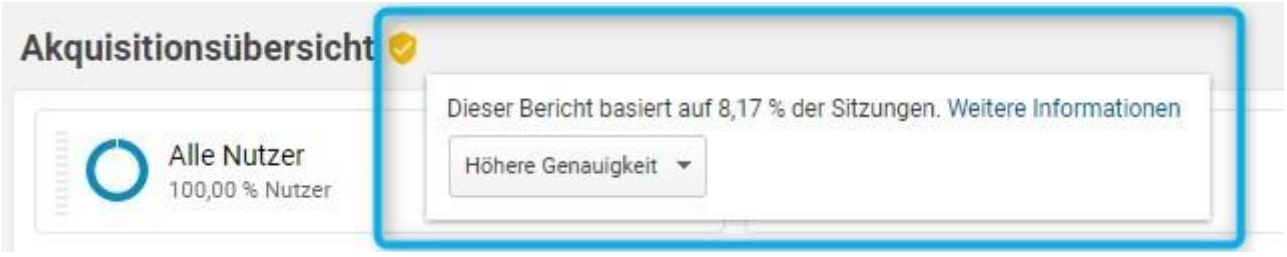

Abbildung 2: Bei Mouseover siehst du, wie groß oder klein die Stichprobe ist.

Was bedeutet Stichprobenerhebung (Sampling)?

Sampling ist zunächst einmal keine böse Absicht durch das Webanalyse-Tool. Grundsätzlich ist es aber so, dass die Darstellung von Tabellen eine gewisse Rechenzeit in Anspruch nimmt. Viele Tabellen in Google Analytics z. B. sind vorberechnet, dort findet in der Regel keine Stichprobenerhebung statt.

Sobald aber mit den Tabellen intensiver gearbeitet wird, beispielsweise durch Segmentierung, bei der die Daten nach bestimmten Kriterien isoliert und untersucht werden, müssten die Tabellen aufwändig neu berechnet werden. Da das mitunter, je nach Art der Segmentierung, lange dauern kann, wird mit Stichproben gearbeitet.

Wie funktioniert das?

Angenommen, du sollst herausfinden, wie viele Bälle sich in einem Container befinden. Auf der einen Seite kannst du jetzt alle Bälle einzeln zählen (vollständige Erhebung), das kann je nach Größe des Containers mitunter lange dauern. Gerade bei der Ad-hoc-Datenanalyse ist Zeit aber oftmals Mangelware.

Auf der anderen Seite kannst du z. B. auch nur die Bälle zählen, die sich auf dem Boden des Containers befinden, und dann die Anzahl der Bälle, die die Höhe des Containers darstellen. Im Anschluss multiplizierst du beide Zahlen miteinander – du rechnest quasi hoch. Dadurch wird das Ergebnis zwar etwas ungenau, aber im Groben lässt sich so schnell schätzen, wie viele Bälle in der Telefonzelle sind.

Hier hast du also eine Stichprobe genommen.

Glück vs. Pech

Bei der Stichprobenerhebung kannst du nun natürlich entweder Glück oder Pech haben.

Ein Beispiel: Google Analytics nimmt sich 25 % der Gesamtdaten vor, um eine Stichprobe zu erheben. Im Extremfall kann es dann z. B. sein, dass damit in Zusammenhang stehende Conversions nur innerhalb dieser 25 % Daten stattfanden und in den anderen 75 % der Daten keine Conversion getätigt wurde.

Bei der Hochrechnung passiert natürlich nun eine extreme Verzerrung, denn Analytics tut so als würde die Conversion Zahl um das Vierfache ansteigen.

Allerdings ist das natürlich ein Extremfall. Die Abbildungen 3 bis 5 zeigen (im Beispiel mit 50 % Stichprobe), wie es dazu kommen kann.

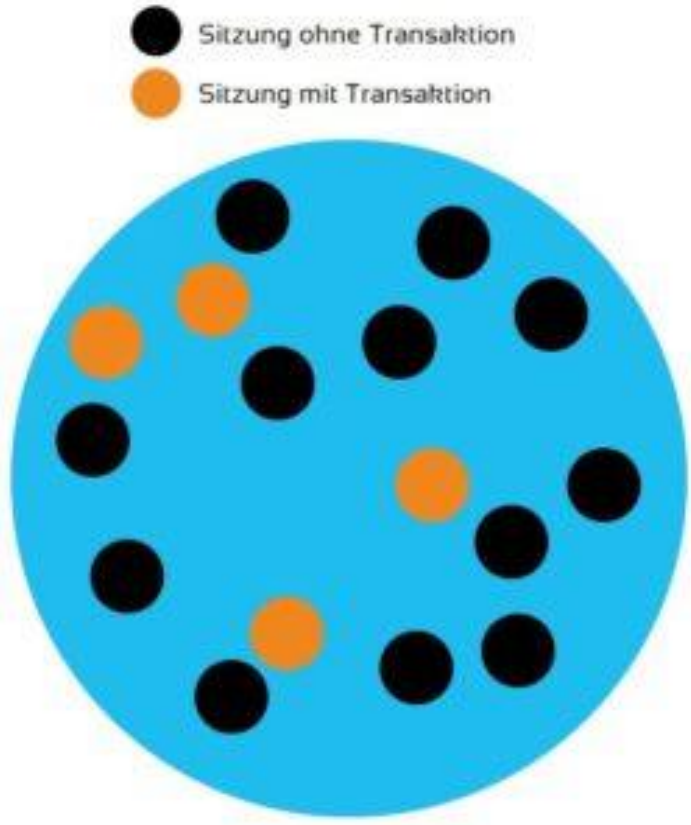

Abbildung 3: Bei der Vollerhebung 16 Sitzungen, davon 4 mit Conversions (Conversion Rate = 25 %)

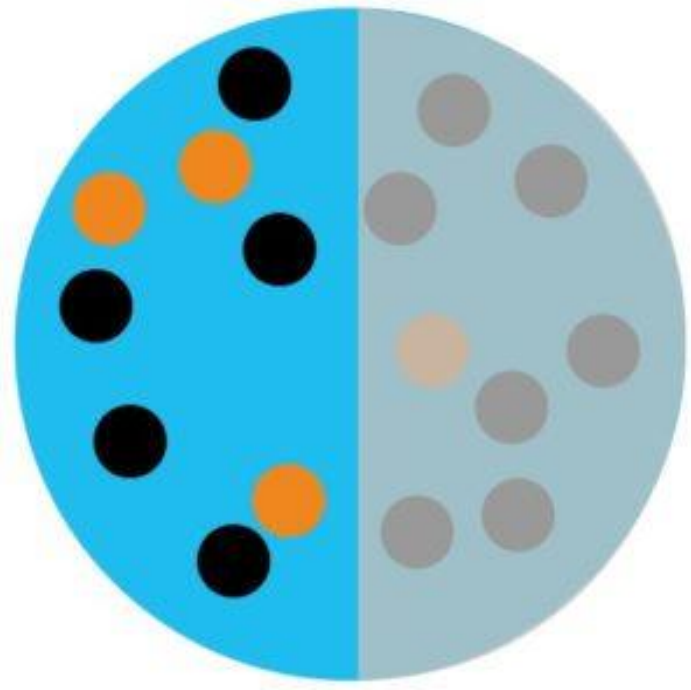

Abbildung 4: Stichprobe 50 % (= 8 Sitzungen), davon 3 mit Conversions (CR = 37,5 %)

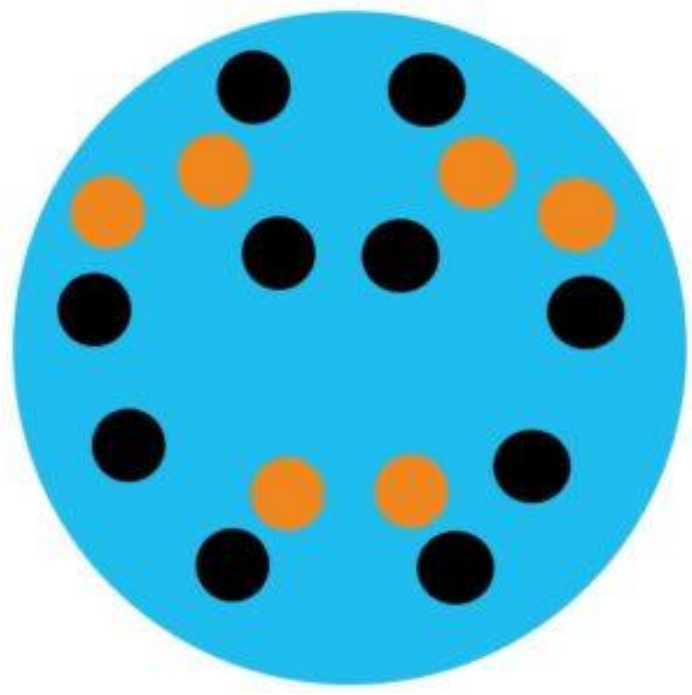

Abbildung 5: Nach Hochrechnung der Stichprobe: 16 Sitzungen, davon 6 mit Conversions (CR = 37,5 %)

Wann wird das Sampling aktiv?

In Google Analytics gibt es unterschiedliche Standards.

In der kostenlosen Version wird das Sampling aktiv, wenn in dem gewählten Analyse-Zeitraum (beispielsweise ein Monat) mehr als 500.000 Sitzungen in der gewählten Property erfasst wurden. Dahingegen schlägt das Sampling in der kostenpflichtigen Version von Google Analytics (Analytics 360) erst bei 200 Millionen Sitzungen zu bzw. bei 50 Millionen, wenn innerhalb des gewählten Zeitraums das aktuelle Datum enthalten ist. Wer nur das aktuelle Datum auswertet, ist auf 1.000.000 Sitzungen begrenzt.

Welche Größe die Stichprobe einnimmt, hängt dabei unter anderem davon ab, wie groß die Datenmenge an sich ist, aber auch wie hoch die sogenannten Kardinalität der gewählten Dimensionen – also zumeist der ersten Spalte in den Berichten – sind (s. Infokasten). Je geringer die Kardinalität, desto geringer die Gefahr einer Stichprobe.

Kardinalität

Wenn eine Dimension in Google Analytics sehr viele unterschiedliche Werte annehmen kann spricht man von "hoher Kardinalität". Eine Dimension beschreibt dabei das "Was?" in den Daten, während Metriken das "Wie viel?" darstellen.

Einige Beispiele für Dimensionen können sein: Seite, Zielseite, Geräteklasse, Stadt, Stunde und so weiter.

Die Dimension "Geräteklasse" weist z. B. eine geringe Kardinalität auf, denn in der Webanalyse nimmt sie gemeinhin nur drei verschiedene Werte ein, nämlich Desktop, Mobile oder Tablet.

Dahingegen kann die Dimension "Seite" extrem viele verschiedene Werte aufweisen, insbesondere dann, wenn die Website ohnehin über sehr viele verschiedene Seiten verfügt oder diese mittels Parametern, wie etwa Sortier-Parametern, Such-Parametern usw. sehr stark diversifizieren kann.

Sampling kann auch Dimensionen mit hohen Kardinalitäten erreichen, dann werden einzelne Werte mitunter als "(other)" in den Tabellen zusammengefasst (siehe Abbildung 3).

Primäre Dimension: Seite Seitentitel Andere

Zeilen darstellen | Sekundäre Dimension | Sortierungsart: Standard

	Seite	Seitenaufrufe	Einzelne Seitenaufrufe
		87.721.287 % des Gesamtwerts: 100,00 % (87.721.287)	58.995.915 % des Gesamtwerts: 100,00 % (58.995.915)
1.	(other)	**37.446.337** (42,69 %)	26.476.617 (44,88 %)
2.	/	**5.844.145** (6,66 %)	3.505.783 (5,94 %)
3.	/search/index/loadsearchform/	**2.552.421** (2,91 %)	1.169.936 (1,98 %)
4.	/appstart/	**1.194.029** (1,36 %)	914.904 (1,55 %)
5.	/default/index/categories/	**748.410** (0,85 %)	426.607 (0,72 %)
6.	/neue/	**699.650** (0,80 %)	432.001 (0,73 %)

Abbildung 6: Auch eine Form von Sampling: die Zusammenfassung von Dimensionen hoher Kardinalität

Die Folgen des Samplings

Insgesamt gilt: Je kleiner die Stichprobe, desto größer werden die Verzerrungen.

In den Reports kannst du dann mitunter ein merkwürdiges Verhalten feststellen, z. B. dass die Anzahl an Conversions im erfassten Zeitraum immer genau 4 pro Tag sind, niemals aber 1, 2 oder 3. Oder aber alle Conversions finden an nur einem Tag statt (siehe Abbildung 7)

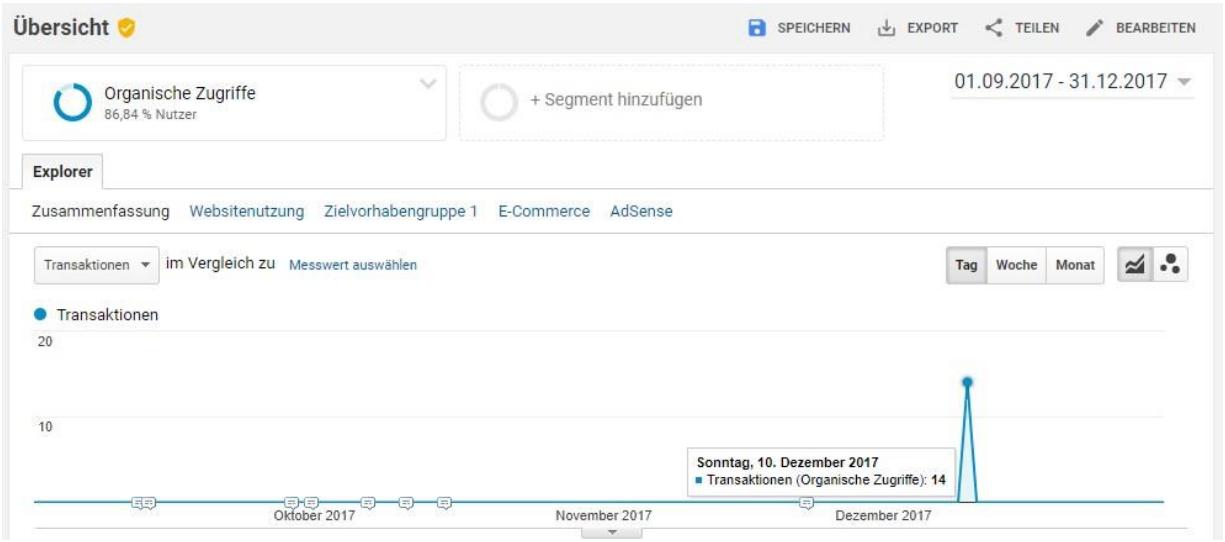

Abbildung 7: Merkwürdiges Gebaren: Alle 14 Conversions fanden an einem Tag statt.

Abbildung 8: Erklärbar ist das durch das aktive Data Sampling. Hier 7,31 % der Sitzungen.

Sampling betrifft allerdings nicht nur Google Analytics, auch andere Webanalyse-Tools greifen aus Performance-Gründen auf das Sampling zurück.

Analyse mit gesampleten Daten?

Generell sollte man sehr vorsichtig sein, sobald das gelbe Symbol aktiv ist. Auf jeden Fall solltest du kontrollieren, wie stark der Anteil der Stichprobe ist, denn insbesondere bei niedrigen Prozentsätzen ergibt es oftmals keinen Sinn mehr, Auswertungen zu fahren oder Entscheidungen auf Basis dieser Daten zu treffen.

Es gibt mitunter sogar Empfehlungen, überhaupt nicht mit gesampleten Daten zu arbeiten, weil die Gefahr zu groß sei, mit den Daten ins Fettnäpfchen zu treten.

Das mag in Teilen auch stimmen, wenn man nur mit Echtdaten arbeiten möchte, um Entscheidungen zu treffen. Doch in der Webanalyse geht es vielfach um Tendenzen und Trends. Und dann ist eben Augenmaß gefragt, ob die statistische Methode einem durch die Daten einen Strich gemacht hat oder ob die Daten nach wie vor aussagekräftig genug sind, um beispielsweise Hypothesen zur Verbesserung der Website aufzustellen. Die Entscheidung liegt letztlich beim jeweiligen Analysten, also bei dir.

Sieben Möglichkeiten, das Sampling zu reduzieren

Leider lässt sich das Data Sampling nie ganz ausschalten, wenn du im Webanalyse-Tool arbeitest. Allerdings gibt es Möglichkeiten, dieses einzuschränken. Sieben Möglichkeiten habe ich dir in der Folge aufgeführt.

1. "Höhere Genauigkeit" einstellen

Wenn du mit dem Mauszeiger über das gelbe Schild-Symbol fährst, hast du dort die Möglichkeit eine Einstellung zu tätigen, die die Genauigkeit der Daten etwas erhöht. Allerdings zu Lasten der Performance.

Hier kannst du dich also entscheiden, ob du lieber langsamer rechnen und dafür etwas genauere Daten haben möchtest oder ob du eher schnell unterwegs und dafür etwas weniger präzise sein willst.

Abbildung 9: Option nutzen "Höhere Genauigkeit"

Wenn du jetzt Hoffnung hast, dass sich das Sampling damit komplett reduzieren lässt, wirst du enttäuscht. Denn in der Regel wird es nur eine geringe "Verbesserung" geben.

2. Kürzere Zeiträume wählen und „Heute" rauslassen

Die Möglichkeit kürzere Zeiträume und damit weniger Sitzungen zu betrachten, ist eine der besten, um das Sampling zu vermeiden.

In der Folge musst du dir allerdings mehrere Berichte anschauen um einen längeren Zeitraum auszuwerten – und die Daten der einzelnen Berichte im schlimmsten Fall zusammenrechnen. Wenn also zuvor eine 2 Monats-Auswertung gemacht wurde, musst du nun z. B. zwei Mal jeweils einen Monat betrachten und die Werte miteinander verquicken.

Dabei musst du allerdings berücksichtigen, dass eine Addition der Werte bei sitzungsbasierten Metriken, wie z. B. der Absprungrate oder der Conversion Rate grundsätzlich möglich ist, bei nutzerbasierten Metriken (etwa Anteil neuer Nutzer) solltest du davon allerdings absehen.

Außerdem kann es enorm hilfreich sein, das aktuelle Datum (Tag) nicht mit auszuwählen. Dadurch wird das Sampling wesentlich früher aktiv.

3. Dimensionen mit geringer Kardinalität

Reports, bei denen die primäre Dimension eine geringe Kardinalität aufweist, sind weniger anfällig für Sampling (z. B. in Form von "(other)").

4. API nutzen

Wenn du die Möglichkeit nutzt, Daten aus Analytics mit Hilfe der API bspw. in Google Spreadsheets oder andere Tools zu holen, kannst du natürlich, statt dir bereits aggregierte Daten über größere Zeiträume zu holen, einzelne Tage abrufen und diese schnell in Tabellen aggregieren lassen – und damit das Sampling reduzieren oder gar ganz vermeiden.

Insbesondere bei Websites mit hohen Sitzungszahlen kann sich dieses Vorgehen rentieren.

5. Nur Analytics 360: Daten ungesampled exportieren

Leider nur in der Premium Version von Google Analytics möglich: die Daten ungesampled zu exportieren und damit dann an anderen Stellen weiter zu arbeiten.

6. Eigene Datenansicht für die zu betrachtenden Daten

Wer seine Daten häufiger nach bestimmten Gesichtspunkten segmentiert, läuft bei hohen Sitzungszahlen natürlich schnell Gefahr ins Sampling zu laufen.

Ein Ausweg daraus kann auch sein, das du dir spezielle Datenansichten erstellst, die die Segmentierung vorwegnehmen. Wenn also eine häufige Segmentierung z. B. lautet den organischen Traffic zu betrachten, dann kannst du mit einer eigenen Datenansicht, deren Filter ausschließlich organischen Traffic durchlässt, vermeiden, dass die Reports segmentiert werden müssen.

7. Lieber Standard-Berichte als benutzerdefinierte Berichte nutzen

Wenn du gerne zu benutzerdefinierten Berichten greifst, wirst du schnell die Erfahrung machen, dass das Sampling dort zumeist noch höher ausfällt als in Standard-Berichten. Das liegt daran, dass man bei benutzerdefinierten Berichten häufig um eine Neuberechnung der Daten nicht herumkommt. Denn zumeist sind diese Aggregierungen sonst nirgendwo in Analytics verfügbar.

Wenn du also Sampling vermeiden möchtest, solltest du, wann immer möglich, mit den Standard-Reports arbeiten.

Ausblick

Die Sampling-Grenzen sind nun schon seit längerem stabil. Mit steigender Rechen-Performance auf Seiten von Google kann ich mir allerdings vorstellen, dass diese Zahlen irgendwann auch angehoben werden. Nicht zuletzt auch, weil natürlich immer mehr Daten erhoben werden (müssen).

Wann das allerdings sein kann, steht in den Sternen.

Mein Fazit: Data Sampling ist halt so

Sampling kann ein echtes Hindernis bei der Analyse darstellen. Wichtig ist zu verstehen, dass es existiert, wann es auftritt und wie es gegebenenfalls unterbunden oder zumindest reduziert werden kann. Dann steht deinen Analysen mit Mehrwert nichts im Wege.

Wenn du auf dem Laufenden bleiben möchtest, melde dich doch einfach für unsere Data Stories an. Dort gibt es immer mal wieder Neuigkeiten aus der Webanalyse. Kein Spam, versprochen.

Der Autor: Maik Bruns

Der ehemalige SEO ist seit Jahren der Webanalyse mit Google Analytics und dem Google Tag Manager stark verbunden. Bruns hat vielseitige Erfahrungen bei der Konzeption und Implementierung von Trackingsystemen aller Größenordnungen und greift dabei im Wesentlichen auf Tag-Management-Lösungen zurück. Sein breites Hintergrundwissen aus Marketing, Technik und Analyse ist bei der Optimierung von Websites immer wieder gefragt.

Er schreibt und bloggt zu Webanalyse-und Tag-Management-Themen, wenn er nicht gerade unterwegs oder bei seiner Familie ist. Ferner ist er in verschiedenen sozialen Netzwerken zu finden. Er hostet den Podcast „Die Sendung mit der Metrik", spricht zu Webanalyse- und SEO-Themen auf Konferenzen und gibt darüber hinaus Seminare/Workshops zum Thema Google Analytics, Google Tag Manager und Google Data Studio.

Nebenbei fotografiert er leidenschaftlich gerne und hat mit Freeletics einen Sportnachfolger für Volleyball gefunden.

Twitter Facebook LinkedIn XING