Datenanalyse mit Interaktiven Grafiken

Martin Theus – Telefónica O_2 Germany

martin@theusRus.de
martin.theus@o2.com

www.theusRus.de

Datenanalyse …

- Aus en.wikipedia.org
  - “Data analysis is a process of gathering, modeling, and transforming data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. …“
  - “… Data analysis has multiple facets and approaches, encompassing diverse techniques under a variety of names, in different business, science, and social science domains.”
- Sichtweisen
  - Mathematik - Anwendung von stochastischen Modellen / Systemen
  - Informatik - Wissensextraktion aus Datenbanken
    - Informationsvisualisierung
  - Wirtschaft - Business Intelligence
    - Decision Support

John W. Tukey's "Philosophy of Data Analysis"

- **1964**
  "… must be considered as an open-ended, highly interactive, iterative process, whose actual steps are segments of a stubbily branching, tree-like pattern of possible actions."
- **1977**
  "There is no excuse for failing to plot and look."
- **1980**
  "Finding the question is often more important than finding the answer."

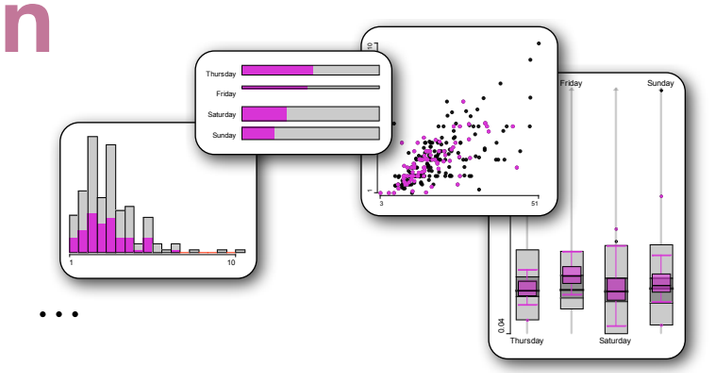
  "EDA is
  - an attitude, AND
  - a flexibility, AND
  - some graph paper (or transparencies, or both)"

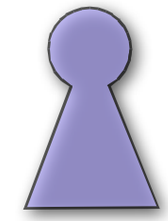
Grundüberlegungen zur Interaktiven Grafik

- Grundlage der meisten statistischen Verfahren ist der **Vergleich von Gruppen** untereinander oder gegen die Gesamtheit
- In der klassischen Statistik wird mittels **Verteilungsannahmen** die Bewertung dieser Gruppenvergleiche bzgl. eines Signifikanz-niveaus möglich
- Eine solche Vorgehensweise wurde auf der Basis von (sehr) **kleinen Datensätzen** entwickelt, und ist daher in vielen Fällen nicht ohne weiteres mehr anzuwenden
- Grafische Methoden sind deutlich flexibler bzgl. Verteilungs-annahmen und Datensatzgröße; der Begriff der Signifikanz geht dann über in die **Relevanz**
- Voraussetzung sind effiziente Methoden für den **grafischen Gruppenvergleich**

Abgrenzung zu anderen Grafik Typen

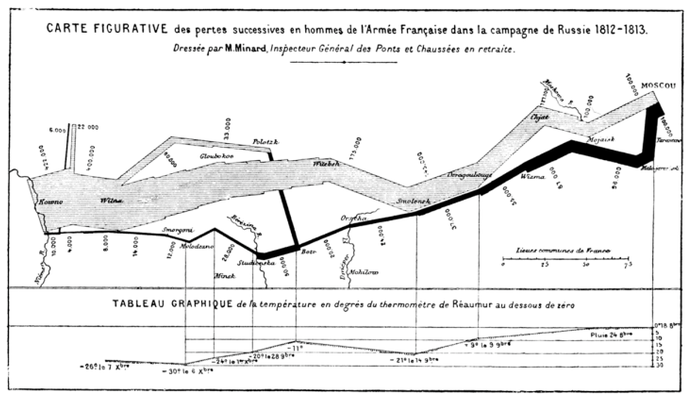
- Exploration
  - Zielt auf Erkenntnis Gewinn
  - Hauptsächlich ein Nutzer
  - Kaum Skalen und Legenden
  - Hochgradig interaktiv und wenig persistent



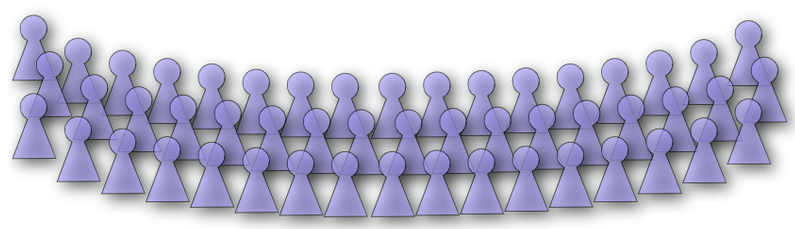


Exploration

- Presentation
  - Präsentiert interpretierte Ergebnisse
  - Optimiert für eine breite Leseschaft
  - Intensive Verwendung von Skalen und Legenden
  - Im statischen Druck ohne Interaktionen



- Info
  - Präsentiert eine Menge uninterpretierter Ergebnisse
  - Nur zugänglich über interaktive Medien
  - Skalen und Legenden soweit nötig
  - Alternative Ansichten und Daten “on demand”



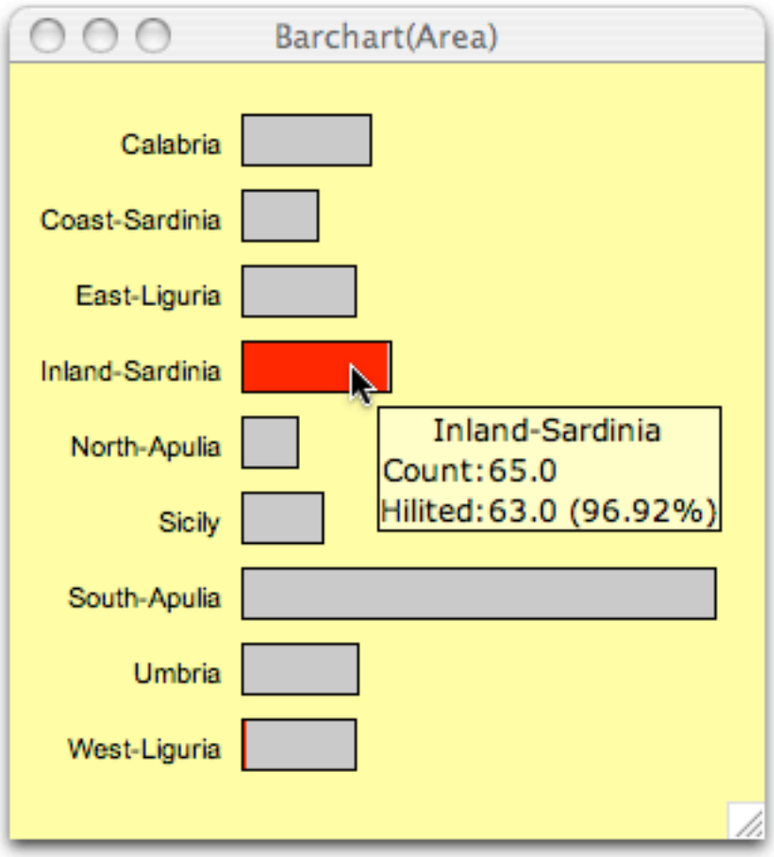
Presentation

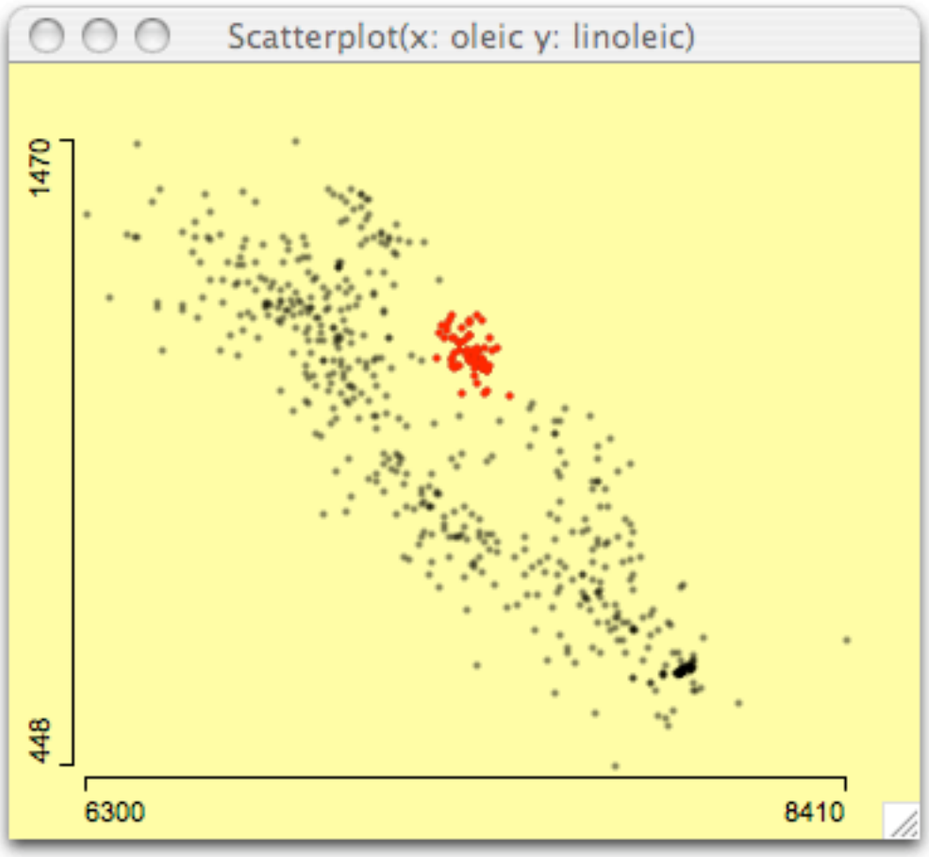
Bausteine der Interaktiven Graphik

- Als Kernfunktionen in einem interaktiven grafischen System fordern wir
  - **Abfragen / Queries**
    Wir brauchen Methoden um exakte, oder nicht sichtbare, Information in einer Grafik abzufragen
  - **Selektionen**
    Um effiziente Gruppenvergleiche durchzuführen brauchen wir Werkzeuge um Daten auf vielfältige Art und Weise zu selektieren
  - **Highlighting**
    Jede Selektion muss via Linking in alle Repräsentationen der Daten propagiert werden um einen Vergleich zu ermöglichen
  - **Modifikation von Grafikparametern**
    Wir wollen die Eigenschaften von Grafiken schnell und effizient variieren können um immer die optimale Ansicht nutzen zu können

Abfragen / Queries

- Grafiken eignen sich gut um qualitative Information darzustellen, geben aber keine exakten Werte wider ⇒ Abfragen
- Grid-Linien können für die in der Grafik gezeigten Variablen helfen
- Interaktive Grafiken verzichten oft auf weiterführende Skalen (vgl. Tuftes “data-ink-ratio”)
- Beispiel:



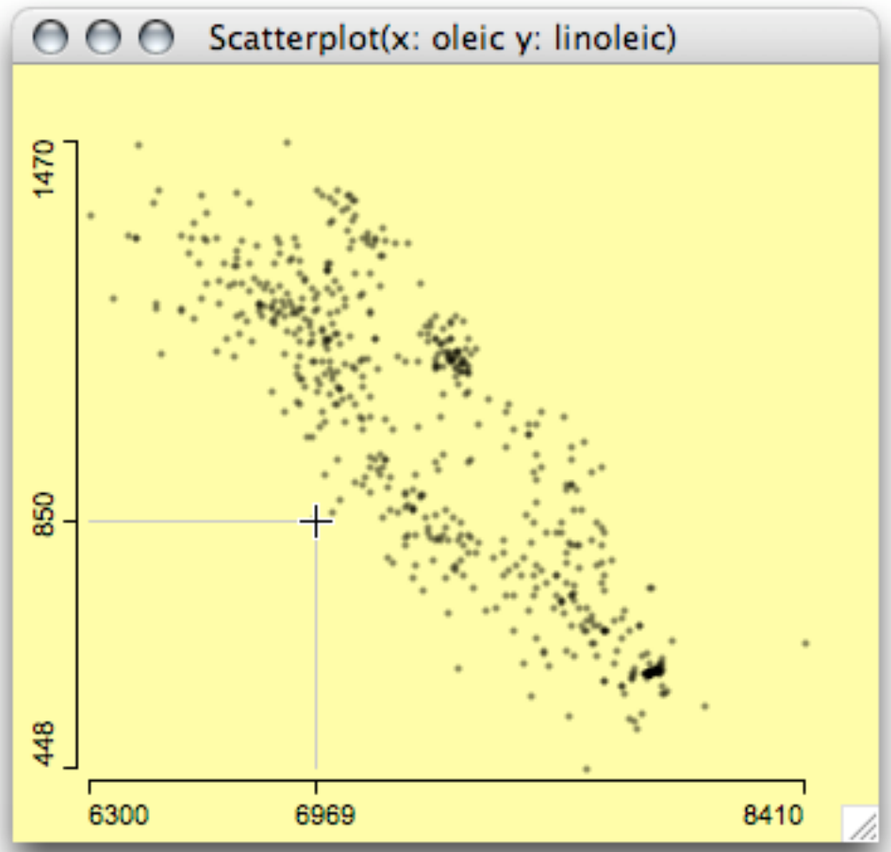


Abfragen auf verschiedenen Ebenen

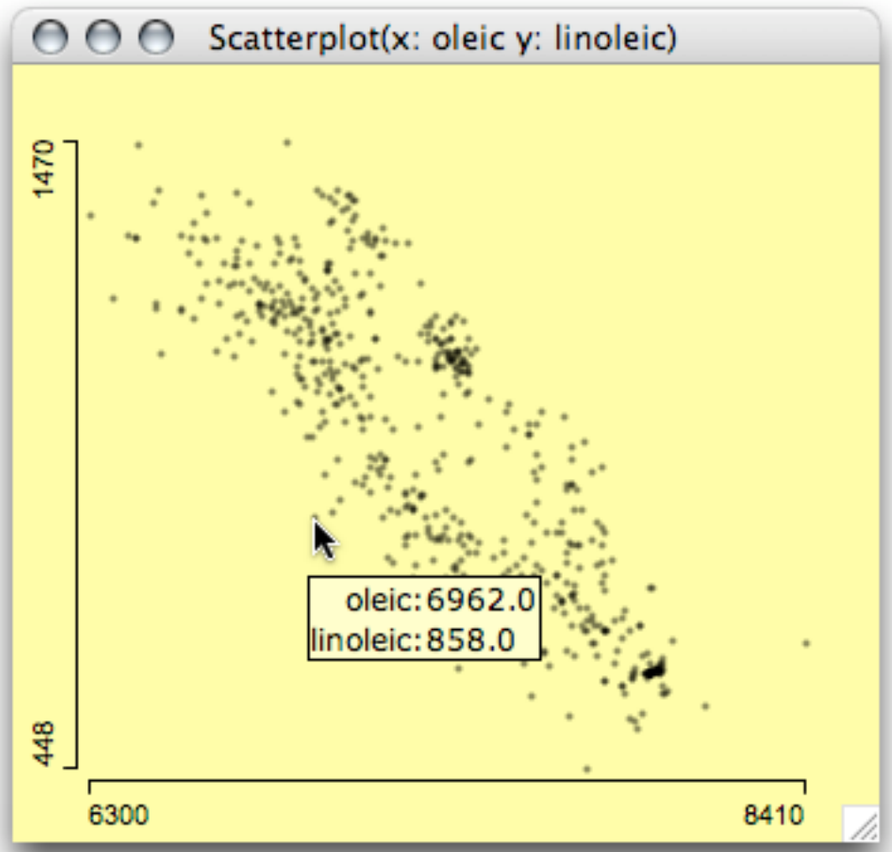
- Die Detailtiefe einer Abfrage kann in drei Stufen gegliedert werden:
  - **Orientierung**, “was sind die Koordinaten am Mauszeiger” (interaktives Grid)
  - **Standard**, “was sind die Koordinaten eines konkreten Wertes”
  - **Erweitert**, “welchen Wert haben Variablen, die nicht in der Grafik sind”

- Example: scatterplot

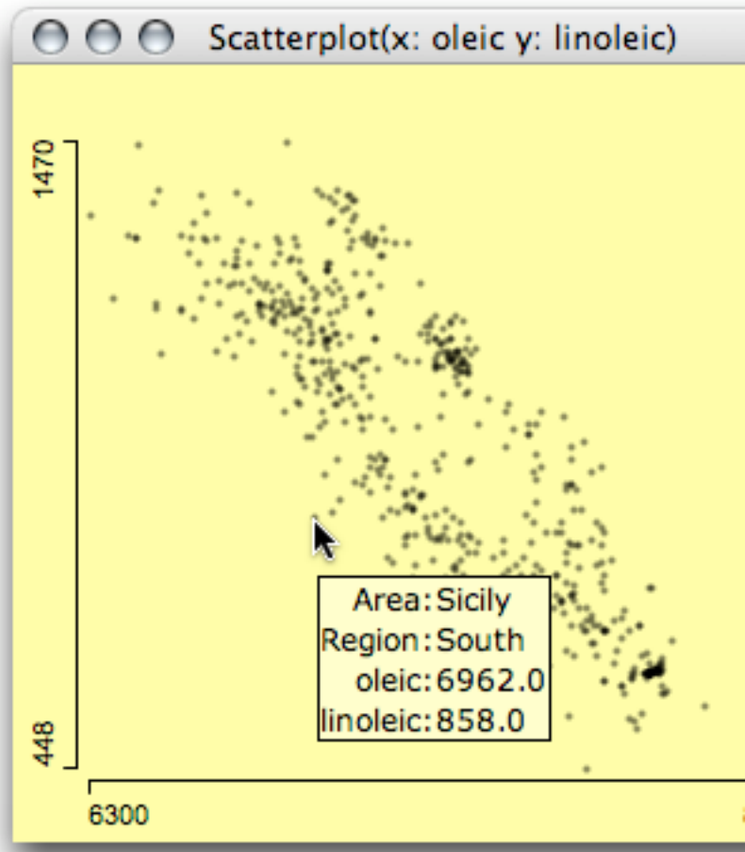
Orientierung



Standard



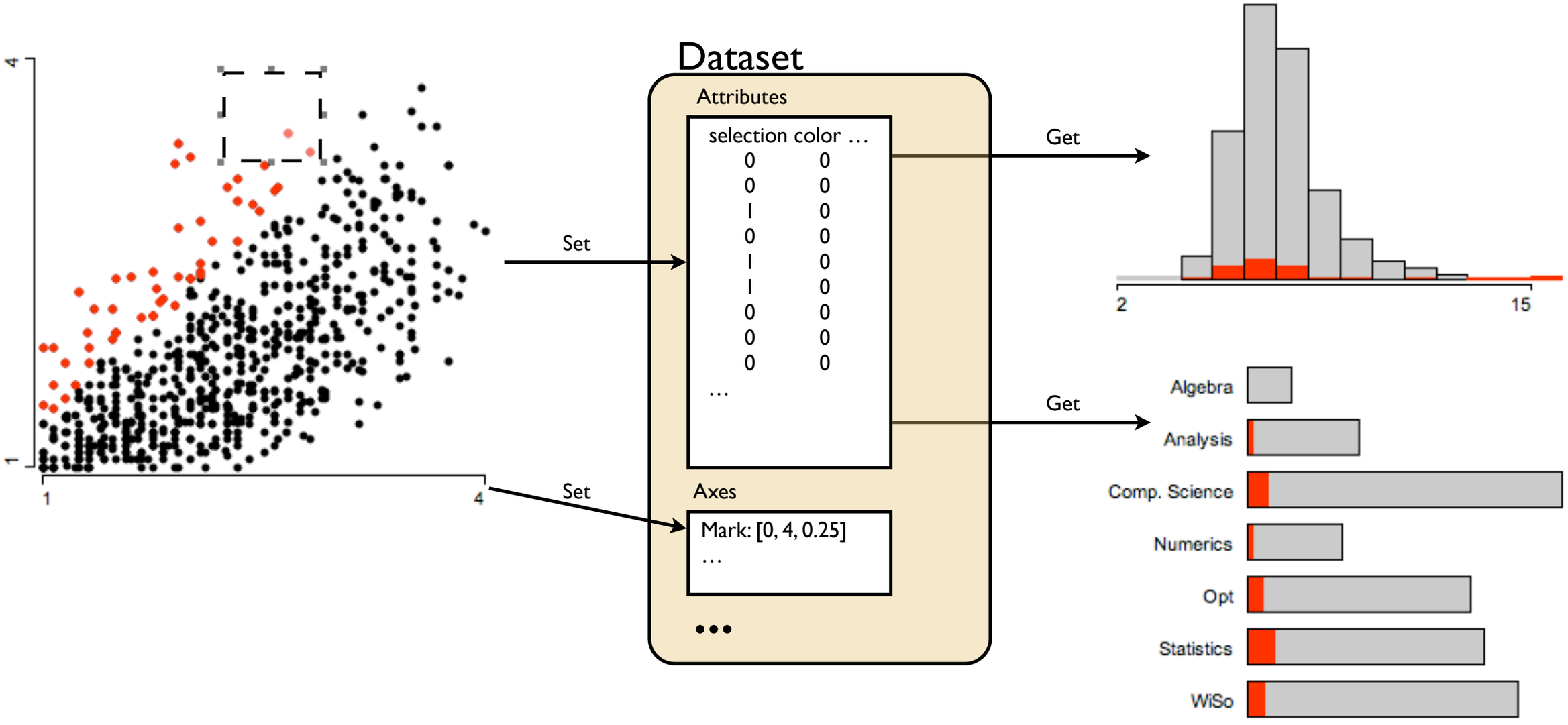
Erweitert



Selektion ➢ Linking ➢ Highlighting

- Linking wird verwendet um Attribute der Daten zu propagieren

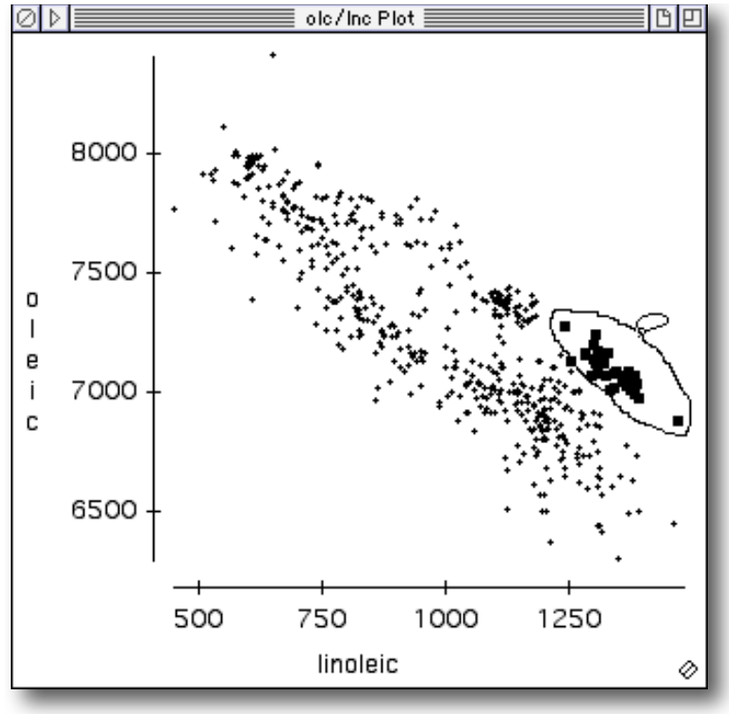
Beispiel:

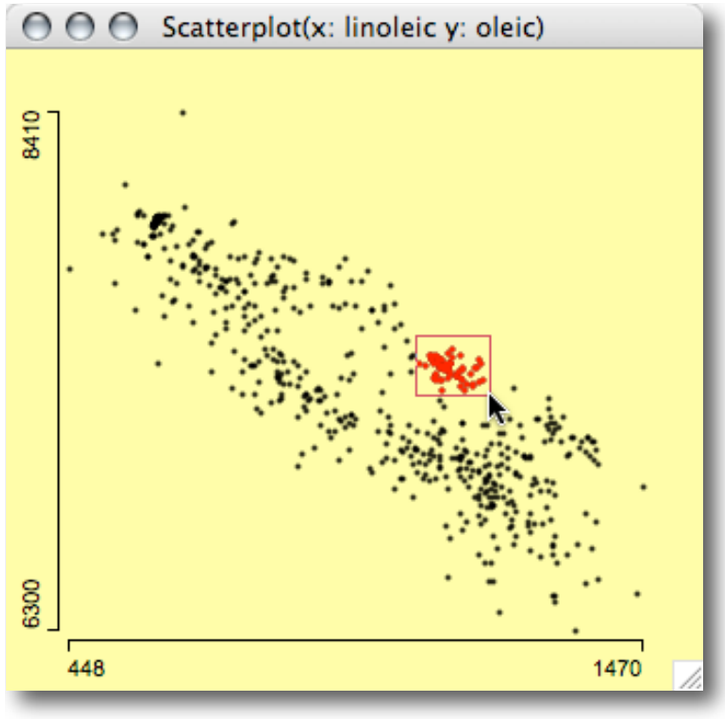


Selektionen

- Selektionen als solches sind nicht wirklich interessant; sie sind aber Voraussetzung, um Subgruppen von Interesse zu spezifizieren
- In einer explorativen Umgebung wollen wir oft die Eigenschaften einer spezifischen Untergruppe analysieren, wie

  “Finde alle Kunden, die abends weniger als 15% Trinkgeld gegeben haben, außer am WE”
- Die Flexibilität mit der man Daten selektieren kann bestimmt direkt, wie erfolgreich man in einer explorativen Analyse sein kann
- Somit braucht man verschiedene Selektionswerkzeuge und -modi





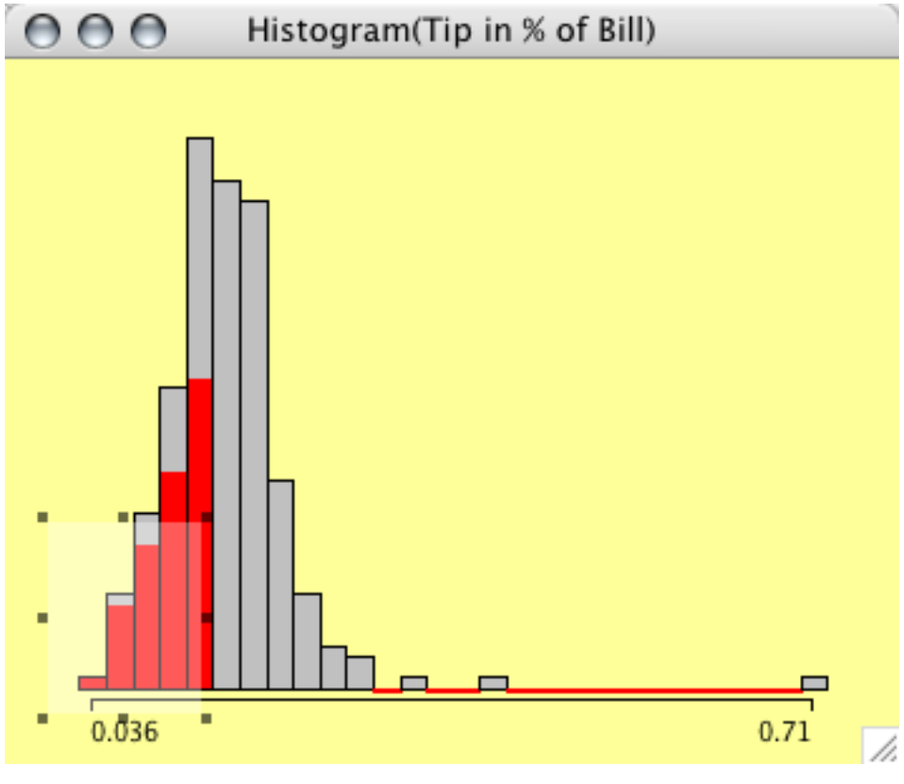
Selektionen von Daten: Modi und Werkzeuge

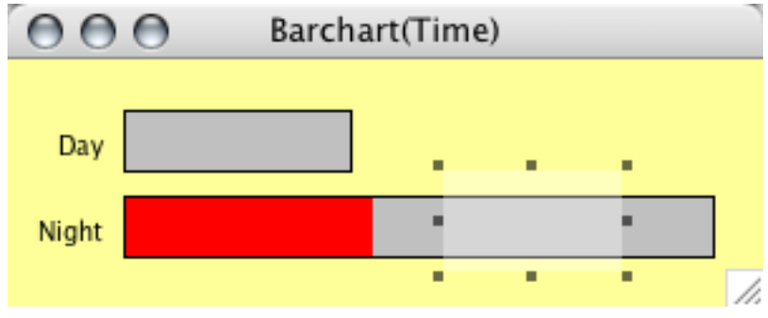
- Werkzeuge zur Selektion:
  - **Zeiger**
    … um einzelne Werte zu selektieren
  - **Drag-Box**
    … selektiert einen rechteckigen Bereich in der Grafik
  - **Brush**
    … ermöglicht eine dynamische Veränderung der Selektion – meist auf Basis eines Rechtecks
  - **Slicer**
    … selektiert dynamisch achsen-parallele Bereiche.
  - **Lasso**
    … erlaubt quasi beliebige Formen der Selektion.
    Start- und Endpunkt werden immer verbunden.

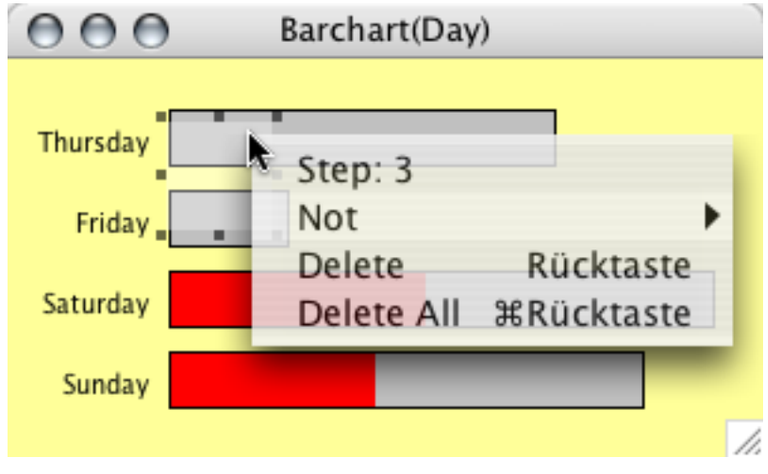
- Selektions Modi
  - **Einfach / Standard / Default**
    … nur die aktuell selektierten Werte werden selektiert
  - **Schnitt / AND / ∩**
    … nur die Punkte die selektiert werden und auch vorher schon selektiert waren werden selektiert
  - **Vereinigung / OR / ∪**
    … die neuerlich selektierten Punkte werden zur Selektion hinzugefügt
  - **Exklusives Oder / XOR / ⊕**
    … die Selektion invertiert den Selektionsstatus der selektierten Punkte
  - **Negation / NOT / ¬**
    … Punkte die selektiert werden, werden aus der Selektion entfernt

Selektions Sequenzen

- Wie und warum werden Selektions Sequences verwendet?
  - Eine Selektion besteht im Allgemeinen nur aus der Menge der Punkte
    ⇒ es existiert keine formale Beschreibung dieser Menge
  - Komplexe Selektionen über mehrere Grafiken sind schwer zu erstellen
    ⇒ Fehler sind fatal, und eine Korrektur der Selektion sehr aufwendig
  - Gezielte Veränderungen einer bestehenden Selektion sind quasi unmöglich
    ⇒ die komplette Selektion muss erneuert werden



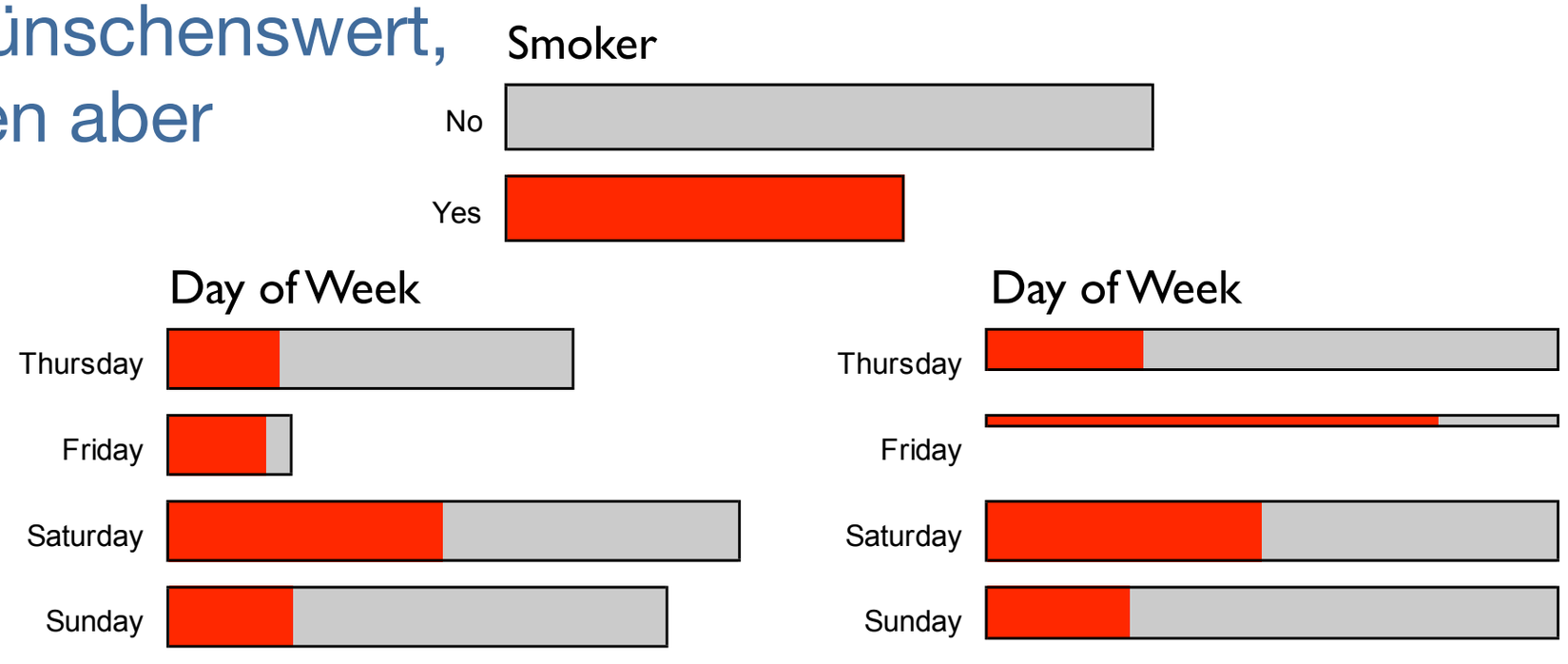




Highlighting: Grundlagen

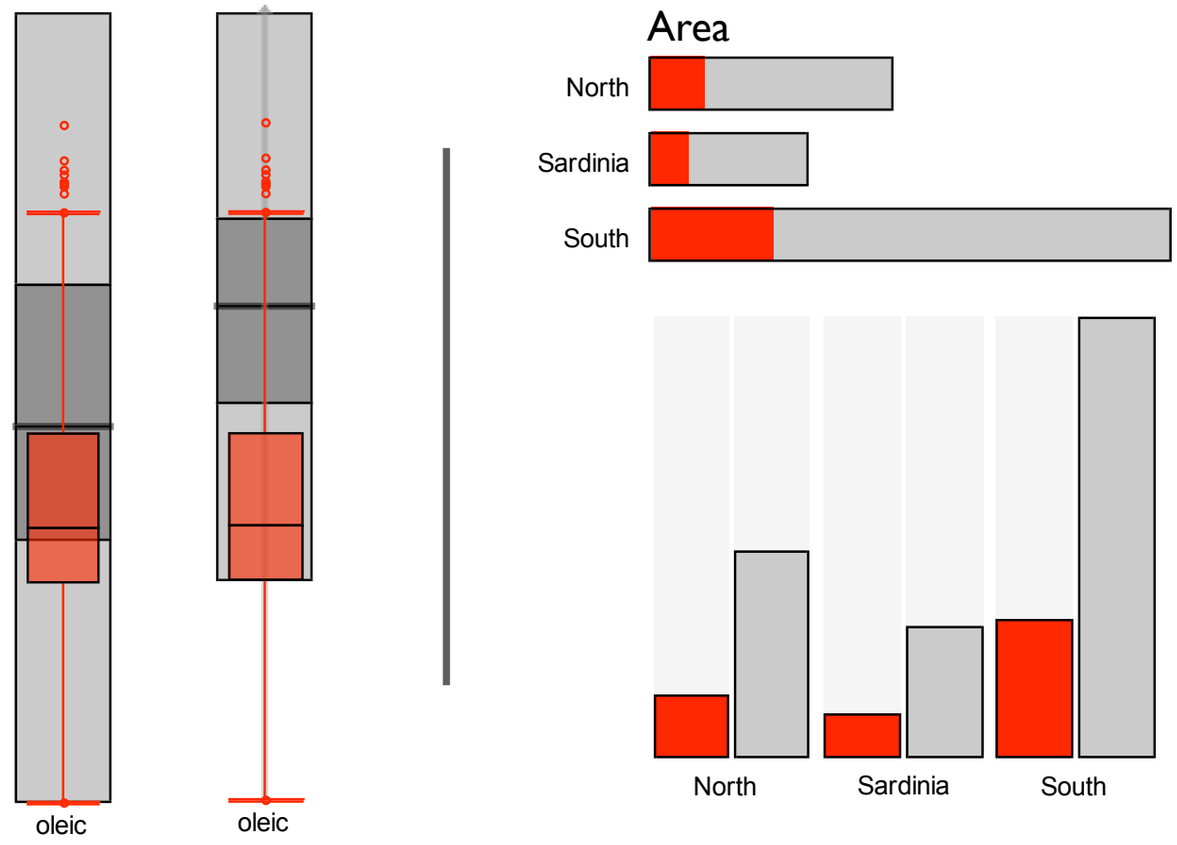
- Wenn eine Selektion definiert worden ist, muss sie zu allen anderen Repräsentation der Daten propagiert werden
- ☞ alle Grafiktypen wissen wie sie Subgruppen highlighten
- Highlighting kann zwei Formen annehmen
  - **transient** (ändert sich mit jeder neuen Selektion)
  - **persistent** (ein neuer Status muss den Punkten explizit zugewiesen werden)
- Klare Regeln, wie sich das Highlighting einer Subgruppe darstellen sollte sind zwar wünschenswert, Ausnahmen können aber sinnvoll sein.

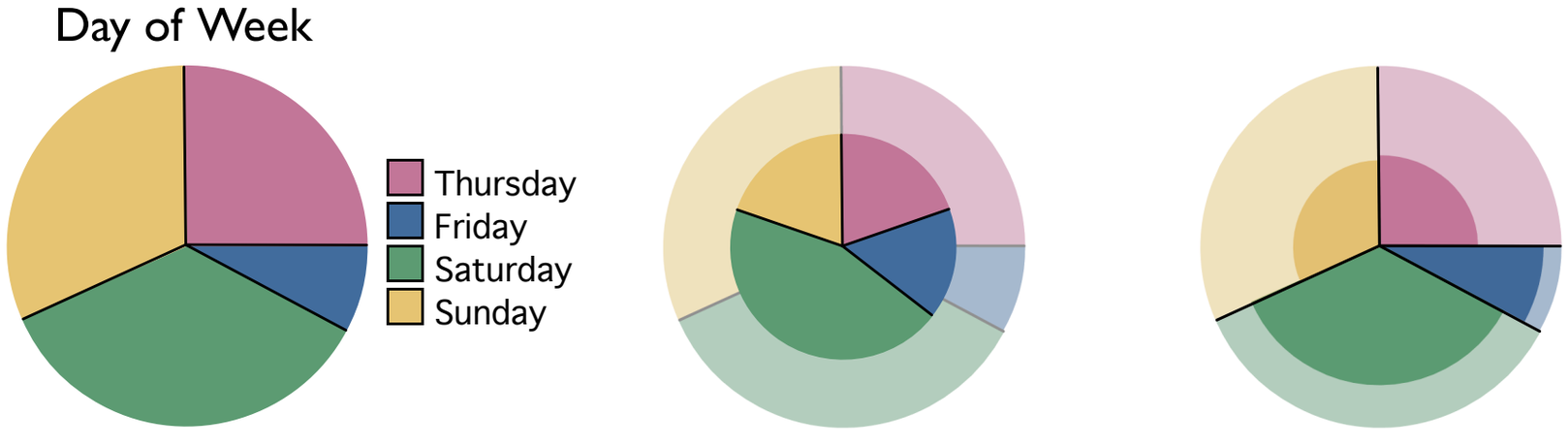
Beispiel: Barchart/Spineplot



Highlighting: Details

- Zwei generelle Fragen stellen sich im Rahmen von Highlighting
  - Ist das Highlighting vom selben Typ wie die Grafik selber?
  - Mit was soll die ausgewählte Gruppe verglichen werden?
    (a) die gesamte Stichprobe oder
    (b) das Komplement der Selektion
    Für viele Grafiken macht es keinen Unterschied z.B. Streudiagramme
  - Bei einigen Grafiken ist der Unterschied bedeutsam, z.B. Boxplot





Änderungen von Grafik Parametern

- Zur Untersuchung von “was-wäre-wenn Szenarien” müssen wir Selektionen und Grafik Parameter unmittelbar ändern können
- Man kann Interaktionen grob in zwei Klassen aufteilen

Generelle Interaktionen

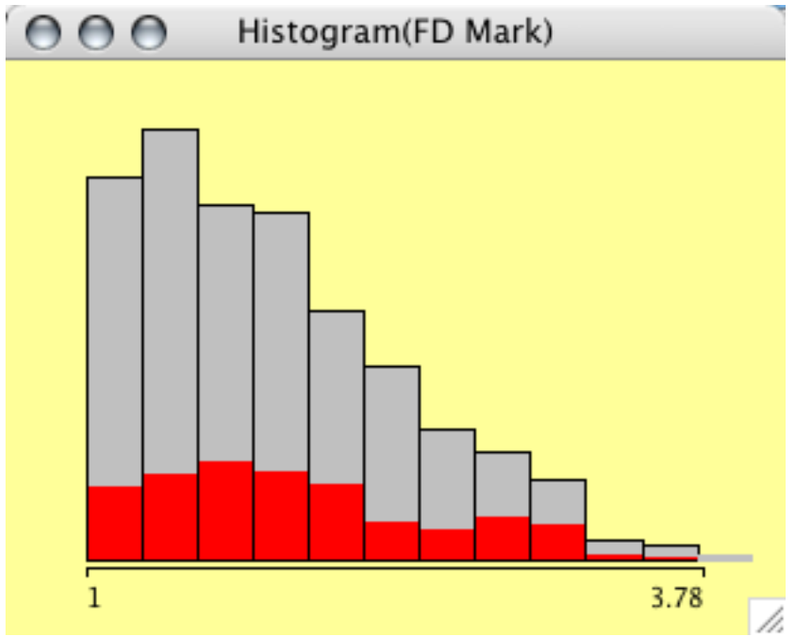
  - Erstellung, und Veränderung von Selektionen
  - Sortieren und Umordnen
  - Änderung der Skalierung (Zooming)
  - Änderung der Stärke der α-Transparenz in Glyph basierten Grafiken

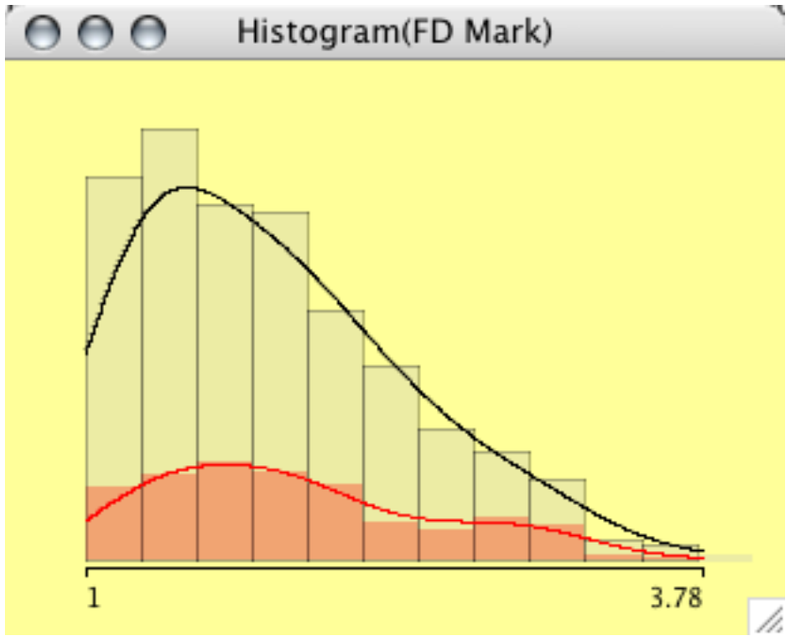
Grafik spezifische Interaktionen

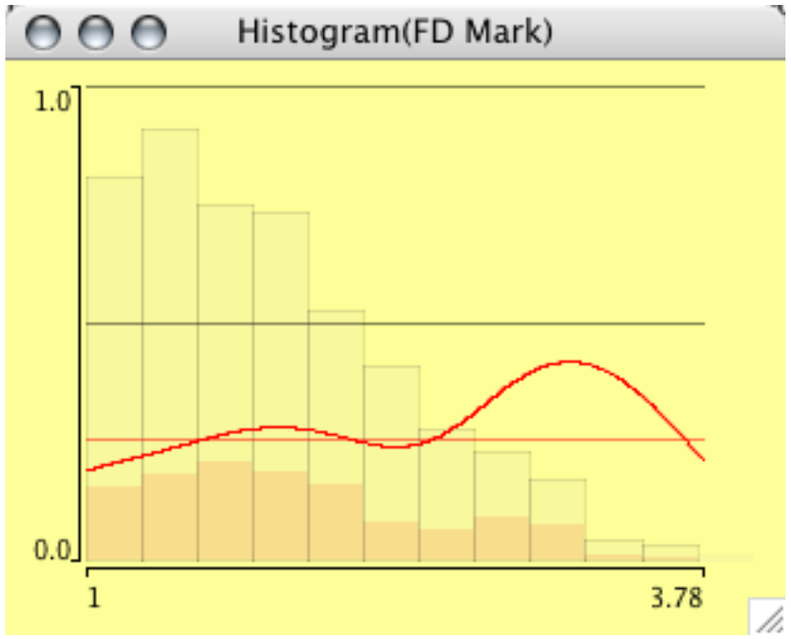
  - Setzen und Verändern des Startpunkts und der Klassenbreite eines Histograms (Bandbreite eines Dichteschätzers)
  - Veränderung der Punktgröße und Transparenz der Punkte im Streudiagramm
  - Vertauschung der Achsen eines Streudiagramms
  - Veränderung der Glättung eines Scatterplot Smoothers
  - Änderung der Darstellung (relativ oder absolut) in einem Balkendiagramm/ Spineplot or Histogramm/Spinogram

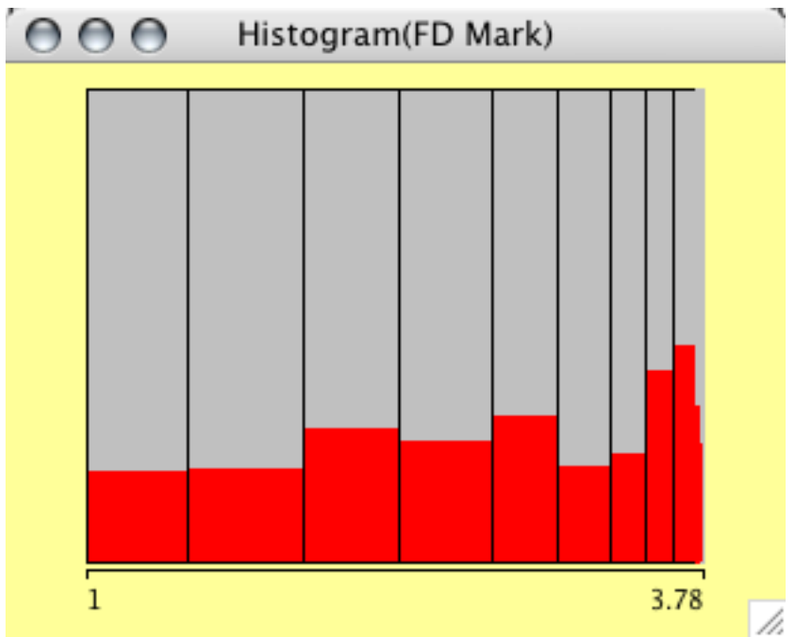
Statistische Anreicherung von Grafiken

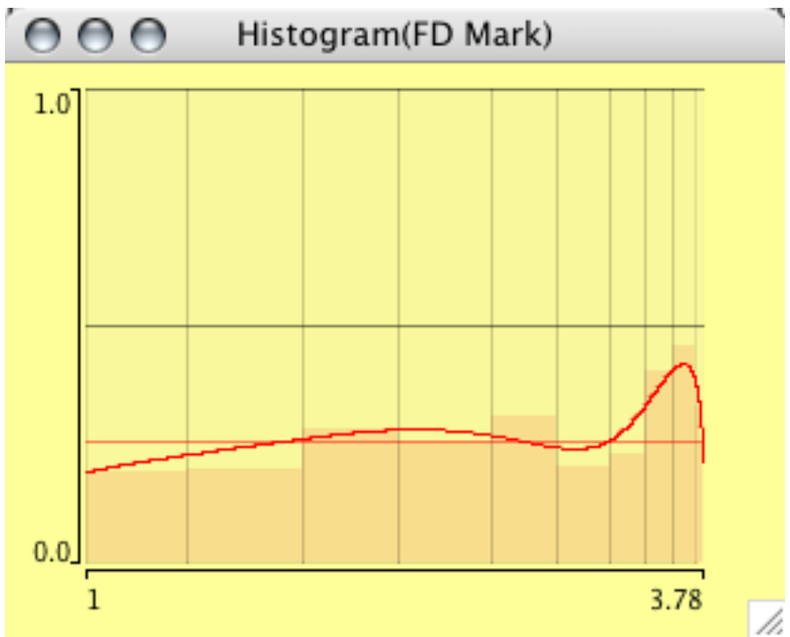
- Beispiel: Dichteschätzung







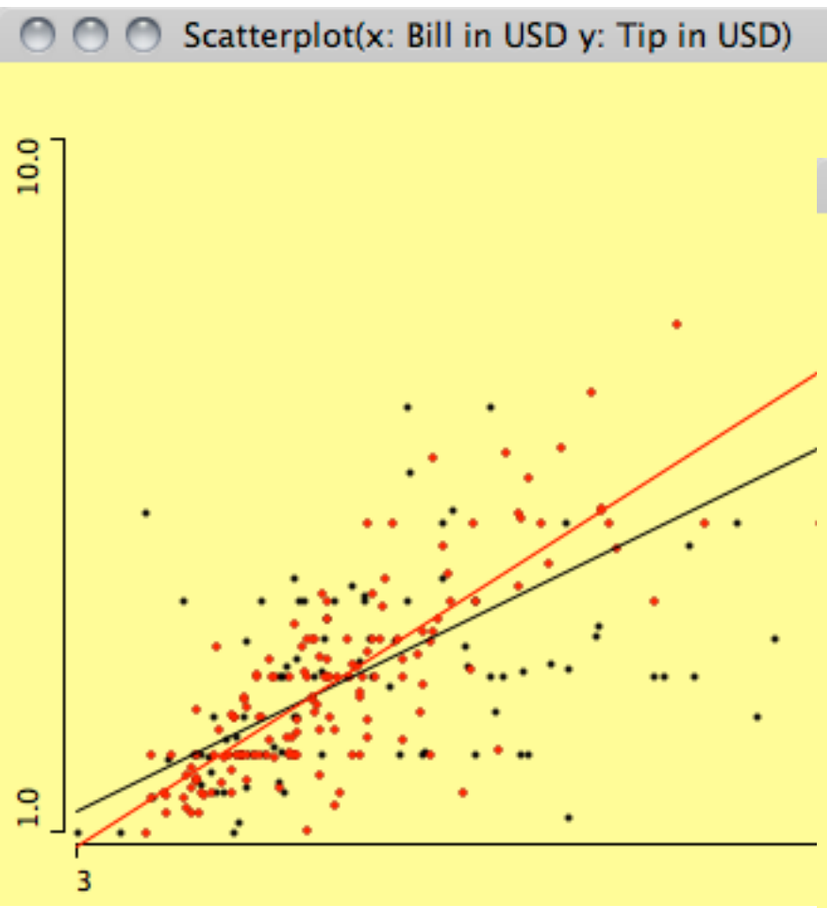




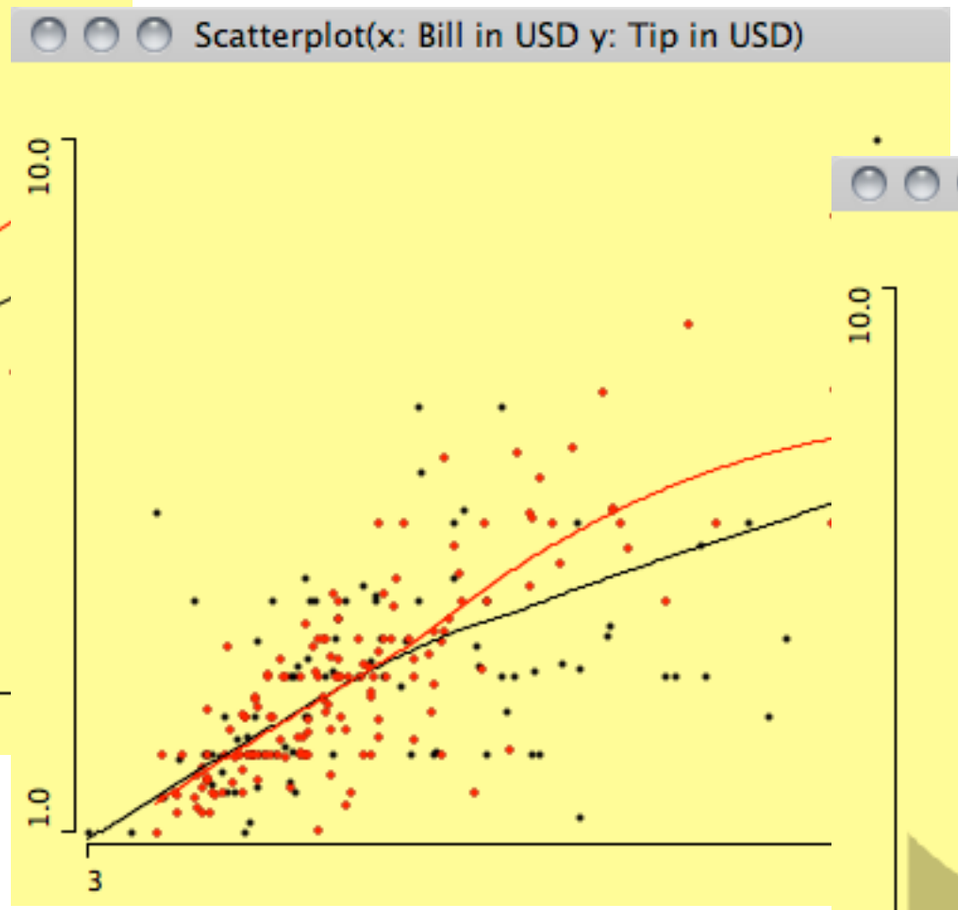
Statistische Anreicherung von Grafiken

- Beispiel: Scatterplot Smoothers

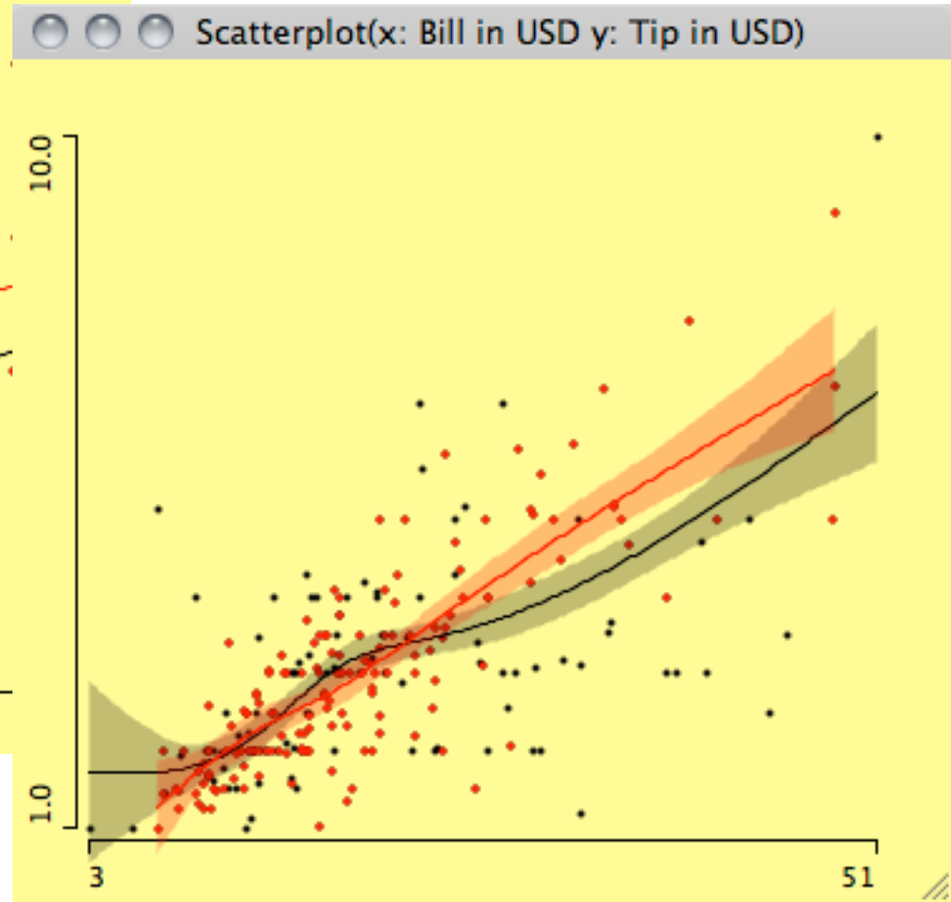
linear regression



loess



splines



Demo: Trinkgeld Daten

- In einem Restaurant in einer Shopping Mall hat eine Bedienung für jede Rechnung in einem bestimmten Zeitraum folgende Daten erhoben:
  - Rechnungshöhe
  - Trinkgeld
  - Geschlecht und
  - Rauchverhalten der zahlenden Person
  - Wochentag und
  - Tageszeit
  - Anzahl der Personen am Tisch.

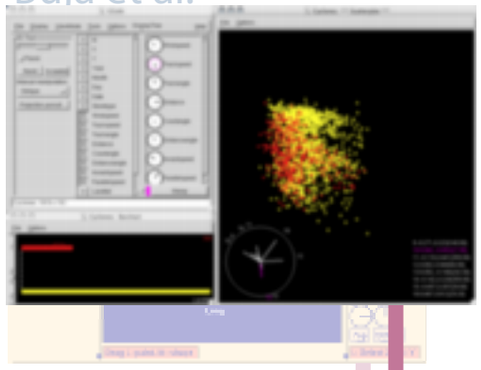
- Naheliegende Frage:
  “Unter welchen Umständen kann man mit einem höheren (als Anteil der Gesamtrechnung) Trinkgeld rechnen?”

Demo

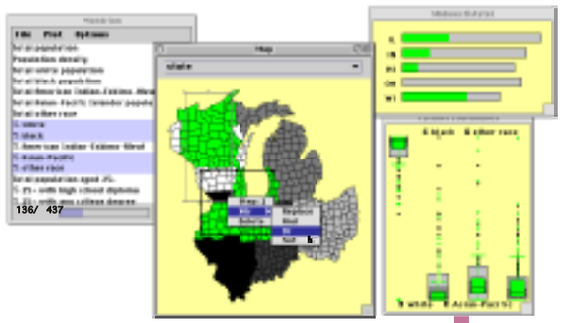
Eine Kurze Geschichte der Interaktiven Grafik

1983
SPLOM
Becker et

1999
ggobi
Swayne et al.

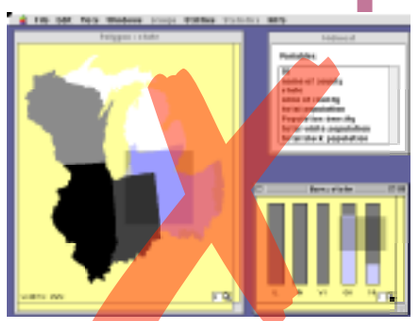


1997
Mondrian
Theus

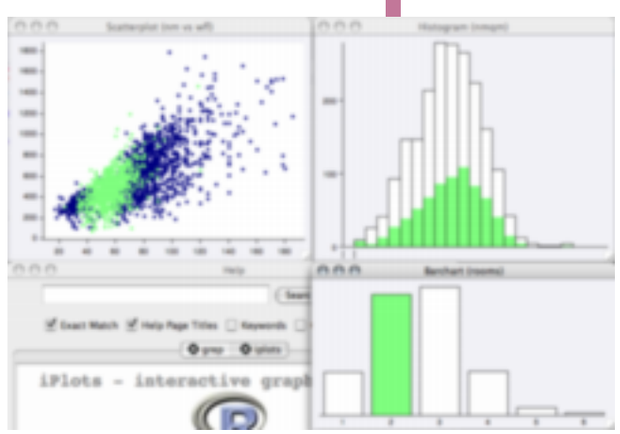


19
PRIM-9
Tukey et al.

1985
DataDesk
Vellemann



1993
MANET
Unwin et al.



2003
iplots
Urbanek & Theus

Software für interaktive Datenanalyse

- Software for graphical data analysis:
  - **ggobi** (frei, Open Source)
    Nachfolger von xgobi. Hat seine besondere Stärke in hochdimensionalen rein stetigen Daten (ohne Ausreißer). Implementiert Grand Tour und Projection Pursuit. Gewöhnungsbedürftiges User Interface. Link zu R. (www.ggobi.org)
  - **DataDesk** (kommerziell)
    Legacy Produkt welches lange der Standard für interaktive grafische Datenanalyse war. Bietet viele statistische Verfahren und Sessions (www.datadesk.com)
  - **SAS/StatStudio** (kommerziell)
    Bietet interaktive Grafik basierend auf dem SAS System. Nachfolger von SAS/Insight, braucht SAS/STAT (support.sas.com/rnd/app/studio/studio.html)
  - **iplots eXtreme** (frei, Open Source)
    Interaktive Grafik in einem R Paket. Bietet die meisten R-Grafiken in einer interaktiven Version an. (www.iplots.org)
  - **Mondrian** (frei, Open Source)
    Stand-alone Programm mit dem breitesten Angebot an interaktiven Grafiken. Anreicherung der Grafiken mit Statistiken aus R. (www.theusRus.de/Mondrian)

Zusammenfassung

- Interaktive Grafik ist ein flexibler und effizienter Weg um Daten zu analysieren, nicht zuletzt auch in der Daten Qualitätskontrolle
- Insofern bilden interaktive statistische Grafiken ein ideales Komplement zur klassischen parametrischen Statistik
- Viele Annahmen der klassischen Statistik sind in der interaktiven Grafik nicht nötig, und klassische Tücken können vermieden werden – dennoch ist eine statistische Anreicherung der Grafiken ein vielversprechender Ansatz
- Eine Anwendung der interaktiven statistischen Grafik ist immer nur mit einer konkreten Umsetzung der Konzepte möglich
- Methoden und Werkzeuge der interaktiven statistischen Grafik sind weit erforscht; faktische Implementierungen sind jedoch noch rar