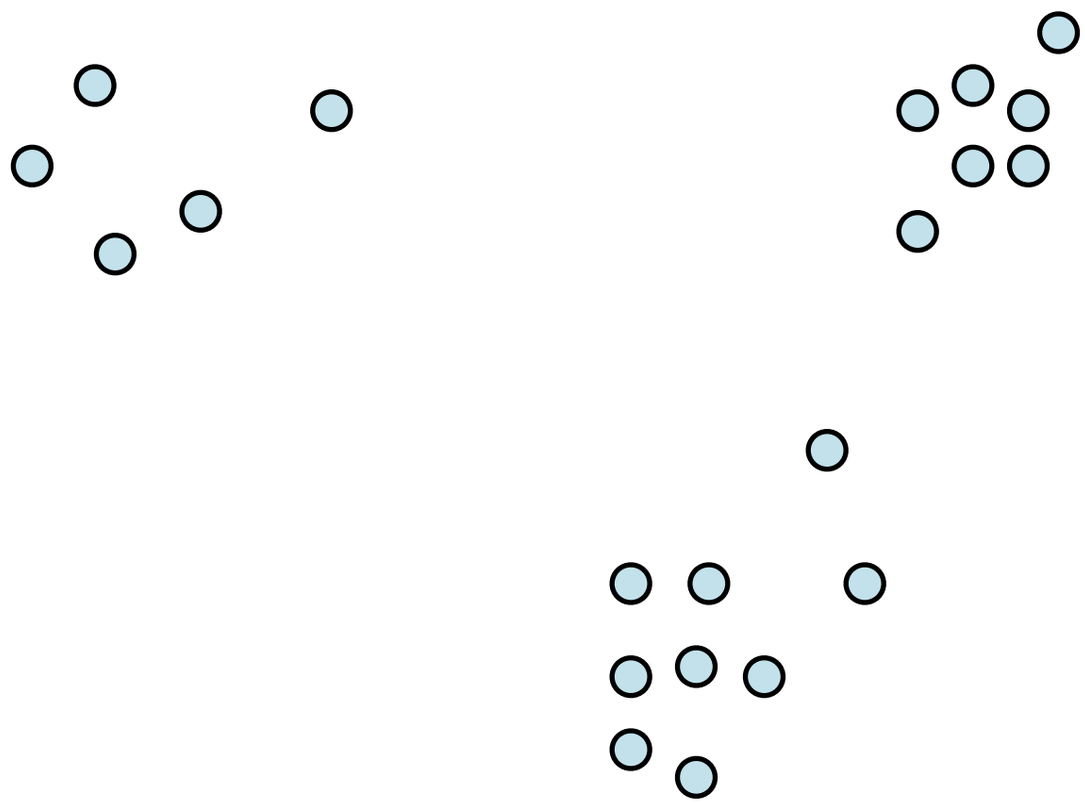
BIRCH

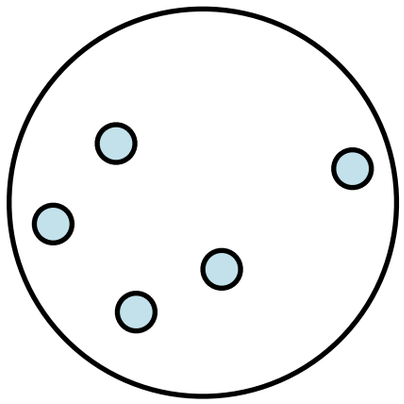
Balanced Iterative Reducing and Clustering using Hierarchies

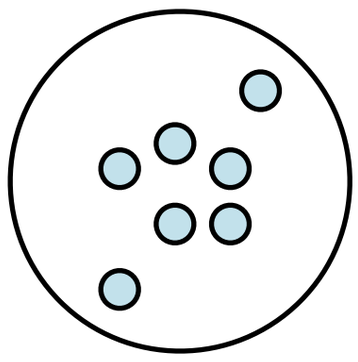
+ Example Implementation of BIRCH in Python

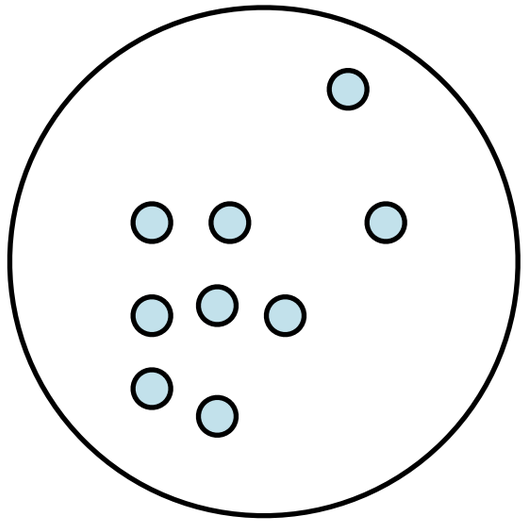
Jan Oberst

27 January 2009









Die weiche Seite des Mnub S.

Versöhnlicher Abschied für Stevens
Rauschender HSV-Abschied für Stevens

Große Gefühle bei Stevens' Abschieds-Gala

Die "Wölfe" und der HSV jubeln

Skibbe vor dem Aus

Der Club ist abgestiegen
Bye, bye Bundesliga...

"Vier" träumen von Europa

"Die Trainer kennen meine Stärken"

Stepanek wird Profi

Krstajic unterzieht sich Operation

Olaf Thon wechselt in die Geschäftsstelle

Thon verlÃ×sst âkÃ§nigsblauenâ Aufsichtsrat

Glück gehabt: Manchester City im UEFA-Cup
Der 34. Bundesliga-Spieltag im Telegramm

Pagelsdorf nimmt sich in die Verantwortung

Rostock gelingt Abschied â Die Stimmen zum Spiel

Diego bekräftigt Treue zu Werder

Happy End für Bremen, bittere Pille für Bayer

Leverkusen trennt sich von Michael Skibbe
Reaktionen zur Skibbe-Entlassung
Bayer 04 trennt sich von Michael Skibbe
Skibbe muss gehen

Rolfes fehlen nur vier Minuten

Skibbe muss gehen

Frankfurter Vier-Klassen-Gesellschaft

Erneute OP bei Preuß
Frankfurter Preuß muss erneut operiert werden

Werder leiht Carlos Alberto an Botafogo aus
Niederlage für Carlos Alberto
Carlos Alberto bis 30.06.2009 bei Botafogo

Das Loslassen tut einfach gut

Allofs und Co. planen fürs Millionenspiel

Botafogo leiht Carlos Alberto aus

Schaaf: FC-Stars kennen nicht nach Bremen

Clustering

Ziel: Ähnliche Punkte gruppieren

- Haben *n* Punkte
- Suchen sinnvolle Cluster

Ähnlichkeit?

- Es gibt kein “bestes” clustering

Distanz

Distanz zweier Punkte

- Viele mögliche Funktionen
- Hier: Kartesisches Koordinatensystem

Distanz

Innere Distanz **in einem** Cluster

“Durchmesser”: möglichst gering

Äußere Distanz **zwischen** Cluster

“Abstand”: möglichst groß

BIRCH

Cluster sind immer **rund**

Sehr viele Daten (1996)

1600 kB Daten

80 kB RAM

Die weiche Seite des Mnub S.

Versöhnlicher Abschied für Stevens
Rauschender HSV-Abschied für Stevens

Große Gefühle bei Stevens' Abschieds-Gala

Die "Wölfe" und der HSV jubeln

Skibbe vor dem Aus

Der Club ist abgestiegen
Bye, bye Bundesliga...

"Vier" träumen von Europa

"Die Trainer kennen meine Stärken"

Stepanek wird Profi

Krstajic unterzieht sich Operation

Olaf Thon wechselt in die Geschäftsstelle

Thon verlÃ×sst âkÃ§nigsblauenâ Aufsichtsrat

Glück gehabt: Manchester City im UEFA-Cup
Der 34. Bundesliga-Spieltag im Telegramm

Pagelsdorf nimmt sich in die Verantwortung

Rostock gelingt Abschied â Die Stimmen zum Spiel

Diego bekräftigt Treue zu Werder

Happy End für Bremen, bittere Pille für Bayer

Leverkusen trennt sich von Michael Skibbe
Reaktionen zur Skibbe-Entlassung
Bayer 04 trennt sich von Michael Skibbe
Skibbe muss gehen

Rolfes fehlen nur vier Minuten

Skibbe muss gehen

Frankfurter Vier-Klassen-Gesellschaft

Erneute OP bei Preuß
Frankfurter Preuß muss erneut operiert werden

Werder leiht Carlos Alberto an Botafogo aus
Niederlage für Carlos Alberto
Carlos Alberto bis 30.06.2009 bei Botafogo

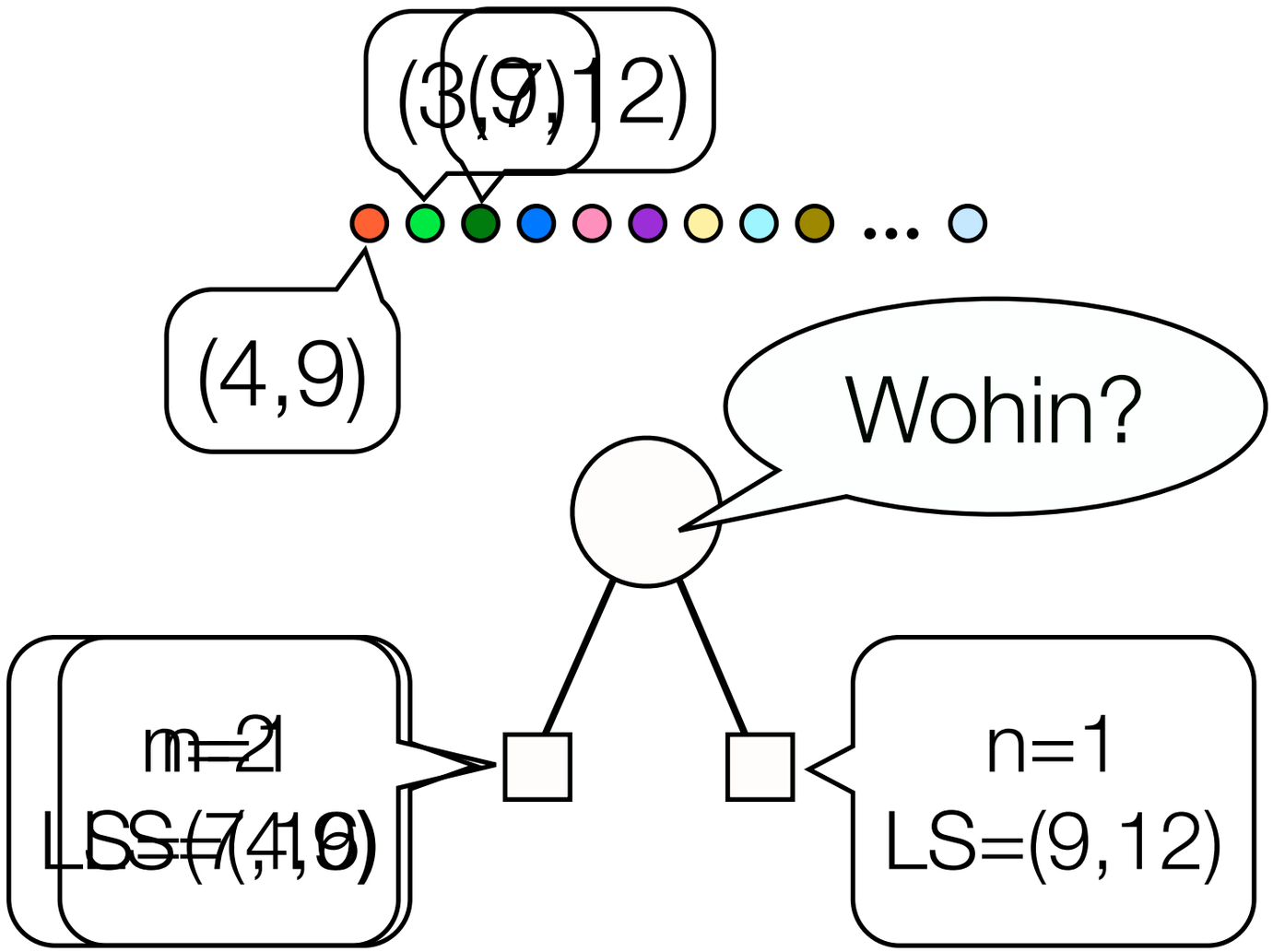
Das Loslassen tut einfach gut

Allofs und Co. planen fürs Millionenspiel

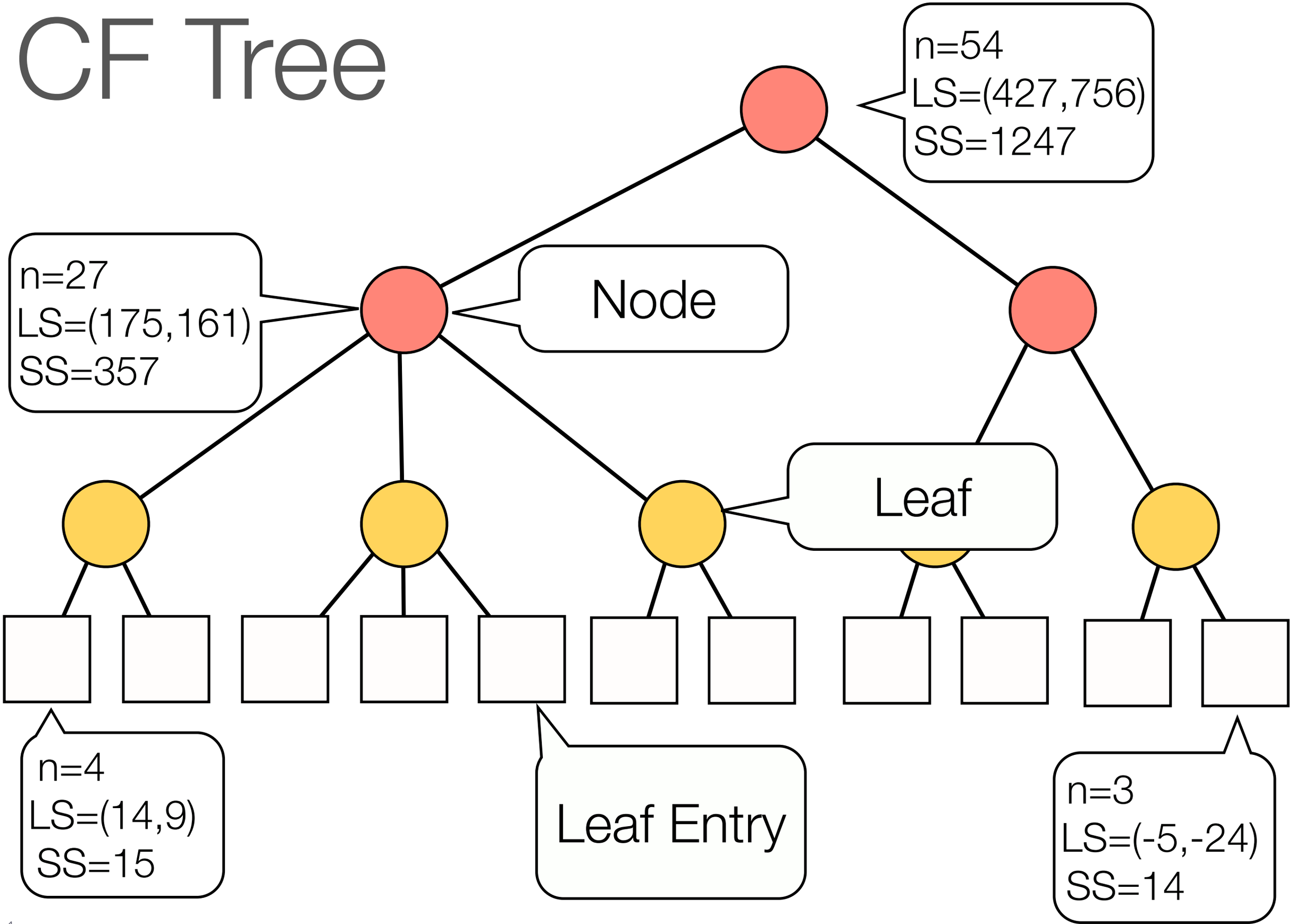
Botafogo leiht Carlos Alberto aus

Schaaf: FC-Stars kennen nicht nach Bremen

B+ Baum



CF Tree


n=54
LS=(427,756)
SS=1247
n=27
LS=(175,161)
SS=357
Node
Leaf
n=4
LS=(14,9)
SS=15
Leaf Entry
n=3
LS=(-5,-24)
SS=14

Cluster Features

Jeder Knoten ist ein Cluster!

Cluster Features

Knoten lassen sich aufsummieren!

CF = (LS1+LS2, N1+N2, SS1+SS2)

Cluster Feature

Anzahl der Kinder N

Linearsumme $$\vec{LS} = \sum_{i=1}^{N} \vec{X}_i$$

Quadratsumme: $$SS = \sum_{i=1}^{N} \vec{X}_i^2$$

B+ Baum

Durchmesser eines Blattes D < **T**

$$D = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\vec{X}_i - \vec{X}_j)^2}{N \cdot (N-1)}}$$

Anzahl Einträge eines Blatts N < **L**

Anzahl Kinder eines Knoten N < **B**

B+ Baum Parameter

Kleines **T**: Viele Blätter, tiefer Baum

Großes **T**: Große Blätter, flacher Baum

Das optimale T

Zu klein / zu groß?

- Ausprobieren!
- BIRCH beginnt mit **T=0**

Neues T wählen, von vorne beginnen

- Suche minimale Distanz
- Neues muss größer sein als das

Von Vorne beginnen

Bauen den Baum neu auf

Disk nicht angerührt

Brauchen dafür doppelt soviel RAM

Entsorgen *Outliers*

Blätter mit deutlich niedrigerer Dichte

Auf Disk schreiben, Später neu versuchen

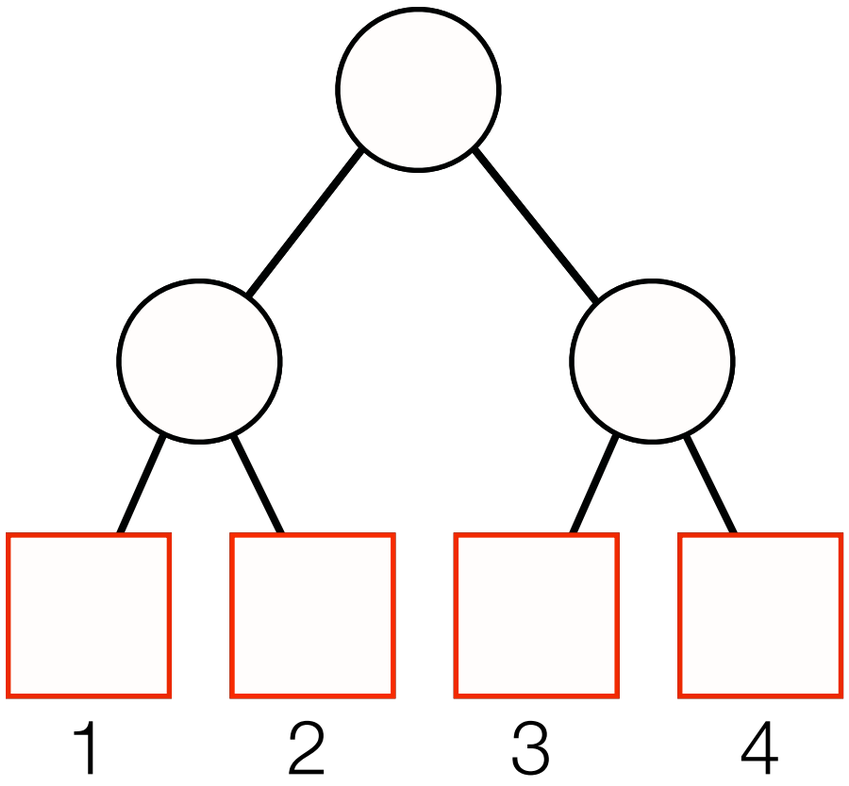
Baum neu aufbauen

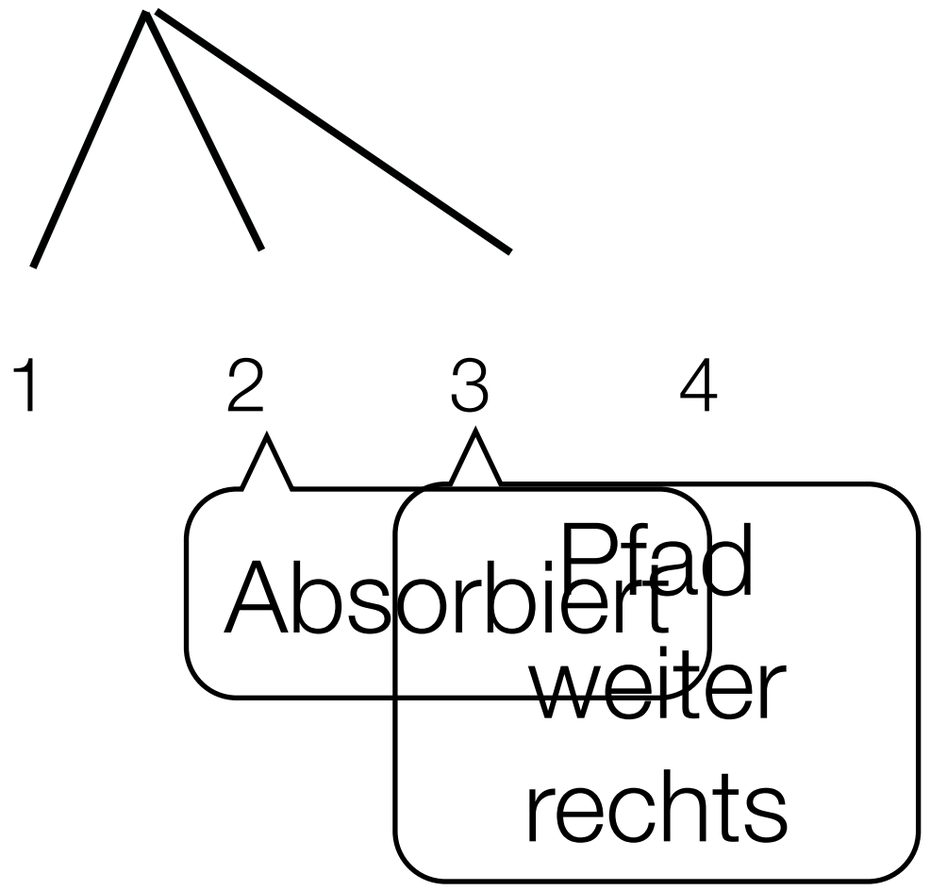
1. Beginnen beim ersten Blatt

2. Füge den kompletten Pfad in neuen Baum ein

3. Weiter links: einfügen
   Weiter rechts: den alten Pfad einfügen

Baum verkleinern





Phase II Condensing (optional)

Vorbereitung für Phase 3

- Je nach Algorithmus notwendig

Falls Ausgabe von Phase 1 zu groß

- Wählen T größer
- Entsorgen mehr Outliers

Phase III

Beliebiger Clustering Algorithmus

local / non-local

hierarchisch

Auch quadratische Laufzeit

Abstand zweier Cluster berechnen

Anzahl der Einträge

Linearsumme (Zentrum)

Quadratsumme (Abdeckung)

Phase III

Algorithmus sieht Blätter als Punkte

- *Duck typing:* Punkte haben Koordinaten!
- LS & SS verfeinern das Clustering

Algorithmus arbeitet mit dem Baum

- Hierarchisches Clustering

Phase III

Bielefeld bleibt in der Liga - VfB in UI-Cup
Arminia Bielefeld feiert ausgelassen Rang 15
Frontzeck: "Ich bin sehr locker heute Abend"
Frontzeck bleibt auch bei Abstieg in Bielefeld
Arminia arbeitet mit Frontzeck
Michael Frontzeck bleibt Trainer in Bielefeld

Paolo Guerrero (Hansburger SV)

Jol macht Hoffmann happy
HSV stellt Trainer Jol vor

Ab sofort gibt Jol den Takt vor
HSV stellt Trainer Jol vor
Stevens' letzter Termin
HSV schickt KSC ab à Die Stimmen zum Spiel

Spielbericht Bundesliga: Hansburger SV à Karlsruher SC
Vorschau Bundesliga: Hansburger SV à Karlsruher SC

Die weiche Seite des Hnub S.

Versöhnlicher Abschied für Stevens
Rauschender HSV-Abschied für Stevens

Große Gefühle bei Stevens' Abschieds-Gala

Die "Wölfe" und der HSV jubeln

Skibbe vor dem Aus

Der Club ist abgestiegen
Bye, bye Bundesliga...

"Vier" träumen von Europa

"Die Trainer kennen meine Stärken"

Stepmek wird Profi

Krstajic unterzieht sich Operation
Olaf Thon wechselt in die Geschäftsstelle
Thon verlässt Schalke Aufsichtsrat
Glück gehabt: Manchester City in UEFA-Cup
Der 34. Bundesliga-Spieltag in Telegramm

Pagelsdorf nimmt sich in die Verantwortung
Rostock gelingt Abschied à Die Stimmen zum Spiel
Diego bekräftigt Treue zu Verder
Happy End für Bremen, bittere Pille für Bayer

Leverkusen trennt sich von Michael Skibbe
Reaktionen zur Skibbe-Entlassung
Bayer 04 trennt sich von Michael Skibbe
Skibbe muss gehen

Rolfes fehlen nur vier Minuten

Skibbe muss gehen

Frankfurter Vier-Klassen-Gesellschaft

Erneute OP bei Preuß
Frankfurter Preuß muss erneut operiert werden

Werder leiht Carlos Alberto an Botafogo aus
Niederlage für Carlos Alberto
Carlos Alberto bis 30.06.2009 bei Botafogo

Das Loslassen tut einfach gut

Allofs und Co. planen fürs Millionenspiel

Botafogo leiht Carlos Alberto aus

Schaaf: EM-Stars kommen nicht nach Bremen

Trainingsauftakt in Wolfsburg erst am 26. Juni

Schalke: Alle Dauerkarten für neue Saison abgesetzt
"Es gab einige schlaflose Nächte"
Rat - oder sogar Riise?

Dortmund feuert Doll

Stuttgart an Dalnat interessiert

Fernandes: Der nächste Versuch

Bundesliga und Kinder

Liga bietet zwei TV-Modelle an
Bayern München erster Club mit vier Sternen
Ausgiebige Free-TV-Berichte vorgesehen
Ordario unzufrieden mit Manöver
Jetzt spiel' halt nicht nur Fußball, Theo!
Acht Spiele des Monats - Zweite Liga um 12.30 Uhr
DFL informiert über Stand der TV-Vermarktung
DFL informiertüber Stand der TV-Vermarktung

"In Spiel die Tiefe nach vorne"

Bleibt Klasnic in Norden?
Ivan Klasnic auf Vereinssuche

Jol in Hamburg gelandet

Ede wechselt zum HSV

Verder: Trainingsauftakt am 3. Juli

Bielefelder Vorstand besucht Ordner

Bochumer Stürmer Bechmann wechselt zum SC Freiburg

Ernst in Bochum als VfL-Sportvorstand vorgestellt

Hansa tritt in Grabow an

Zeitung: Bremen will Diego Olympia-Start verweigern

DFL zeigt Fairplay-Wertung
Die offizielle Fairplay-Rangliste

Duisburg verpflichtet Ede von Hertha BSC Berlin
HSV verpflichtet Chinedu Ede

1. FC Nürnberg kämpft um Pinola
Nürnberg: Stützen sollen bleiben

Nerk und Fandel ausgezeichnet

Cottbus erwirtschaftet erneut Gewinn
Christoph Daum bleibt
Künftig drei Sonntagsspiele

«Salami-Spieltage»:Kritik der Amateure

Nachwuchsprofi erhält Vertrag bei S04
Neuer Torwart für Schalke
U 19-Nationaltorwart Ansif in Schalkes Profi-Kader

S04 befördert Ansif
Ganz Schalke schaut auf den 1. August

Erfolgreiche Operation bei Krstajic
Krstajic erfolgreich operiert
Schalker Krstajic am Arm operiert

Drei BVB-Tore zum Saisonabschluss

Rekord: Cottbus macht neun Millionen Euro Gewinn

Dimitar Rangelow bleibt in Cottbus

Energie erwirtschaftet Millionen-Gewinn
Energie siegt in Benefizspiel
Cottbus gewinnt in Luckenwalde 8:2
Cottbus mit Neuverpflichtung
Cottbus verpflichtet Montenegriner

Fischer trifft dreimal

Energie holt Pavicevic

FC Energie zieht positive Jahresbilanz

Cottbus will Rangelow-Option ziehen
Energie zahlt für Rangelow

Bielefeld feiert in Samba-Zug

"Ich erscheine am 28. Juni zum Training"

BVB erneut auf Trainersuche
Thomas Doll tritt zurück
Borussia Dortmund und Thomas Doll trennen sich
Thomas Doll tritt zurück
BVB trennt sich angeblich von Trainer Doll

Degenas Wechsel von BVB zum FC Liverpool perfekt

BVB-Trainingsauftakt am 2. Juli

Zerrüttet: Doll und Dortmund gehen getrennte Wege
Borussia Dortmund und Trainer Doll trennen sich

Keeper Stolz verlängert in Stuttgart

Stuttgart gewinnt Benefizspiel

Drei Schlaudraff-Tore bei Bayern-Sieg in Jakarta

Klinsmann holt Martins

Gattuso sagt Bayern ab
Bayern-Absage: Gattuso verlängert bei Milan

Zukunftsplanung: Podolski lässt sich Zeit

Toller Empfang für die Bayern
Der Meister reist der Sonne entgegen

Personalplanung beim FCB schreitet voran
Auftakt nach Maß
Schweinsteiger soll bleiben
Klinsmann macht Beckenbauer «neugierig»
Klinsmann macht Beckenbauer«neugierig»

Phase IV finish

Lesen alle Daten nochmal

- Jeder Punkt bekommt sein Cluster

Qualität

- Neuer Mittelpunkt: Durchschnitt aller Punkte
- Phase 4 kann Probleme aus 1-3 beheben
- Wenig RAM: In Phase 4 investieren

Eigenschaften

local vs. nonlocal

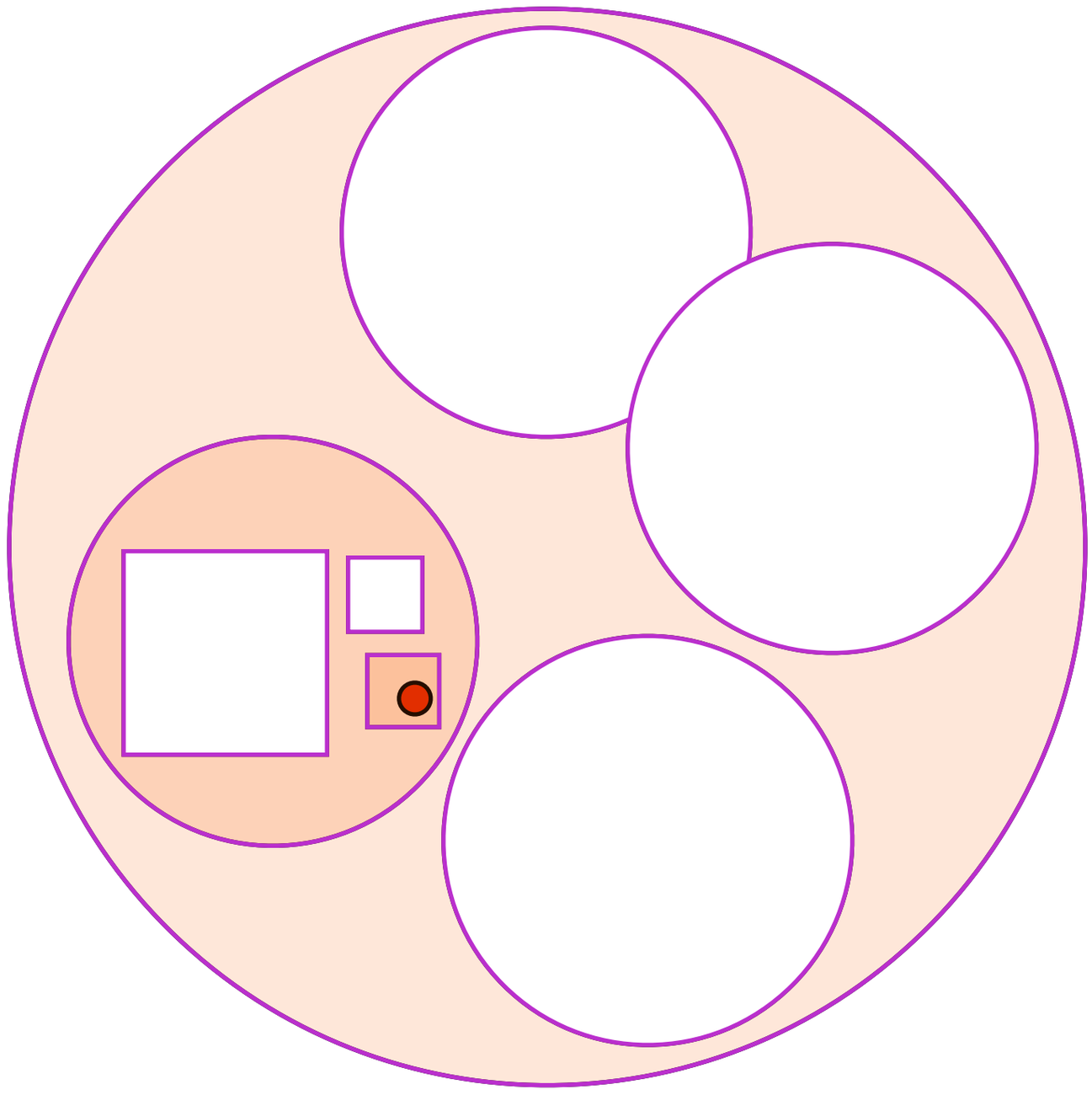
BIRCH

immer auf Region beschränkt

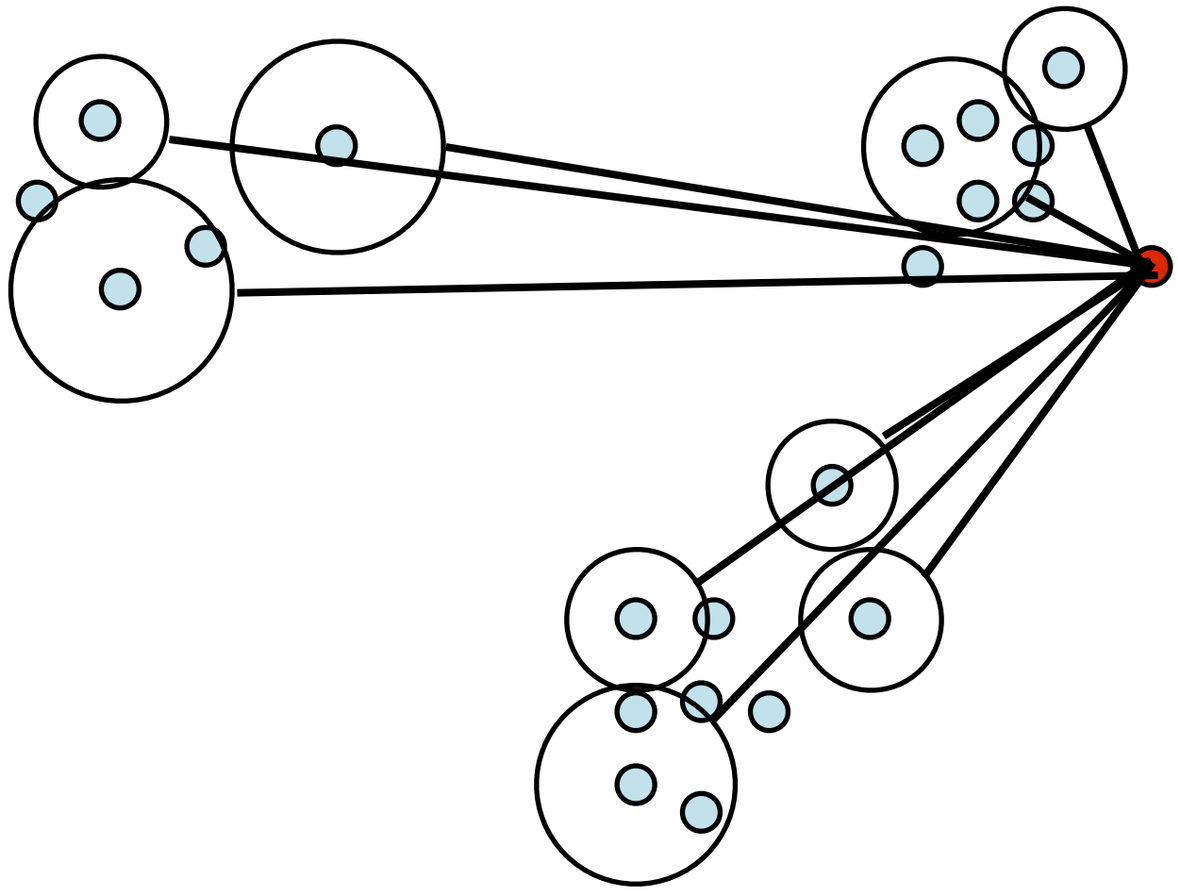
Naiv

Vergleiche mit allen Clustern

local BIRCH



nonlocal k-Means



Skalierbarkeit

Linear

- IO “Should scale linearly with N”

Komprimiert

- Hunderte Punkte in einem Cluster Feature

Single-pass

- Phase IV nicht zwingend

Qualität

Outliers

Können automatisch behandelt werden

Machen Algorithmus sogar schneller!

RAM Nutzung

Kleine Limits = Viele Buckets = gute Qualität

Große Limits = Wenige Buckets = viele Daten

Uber das Paper

Ergebnisse

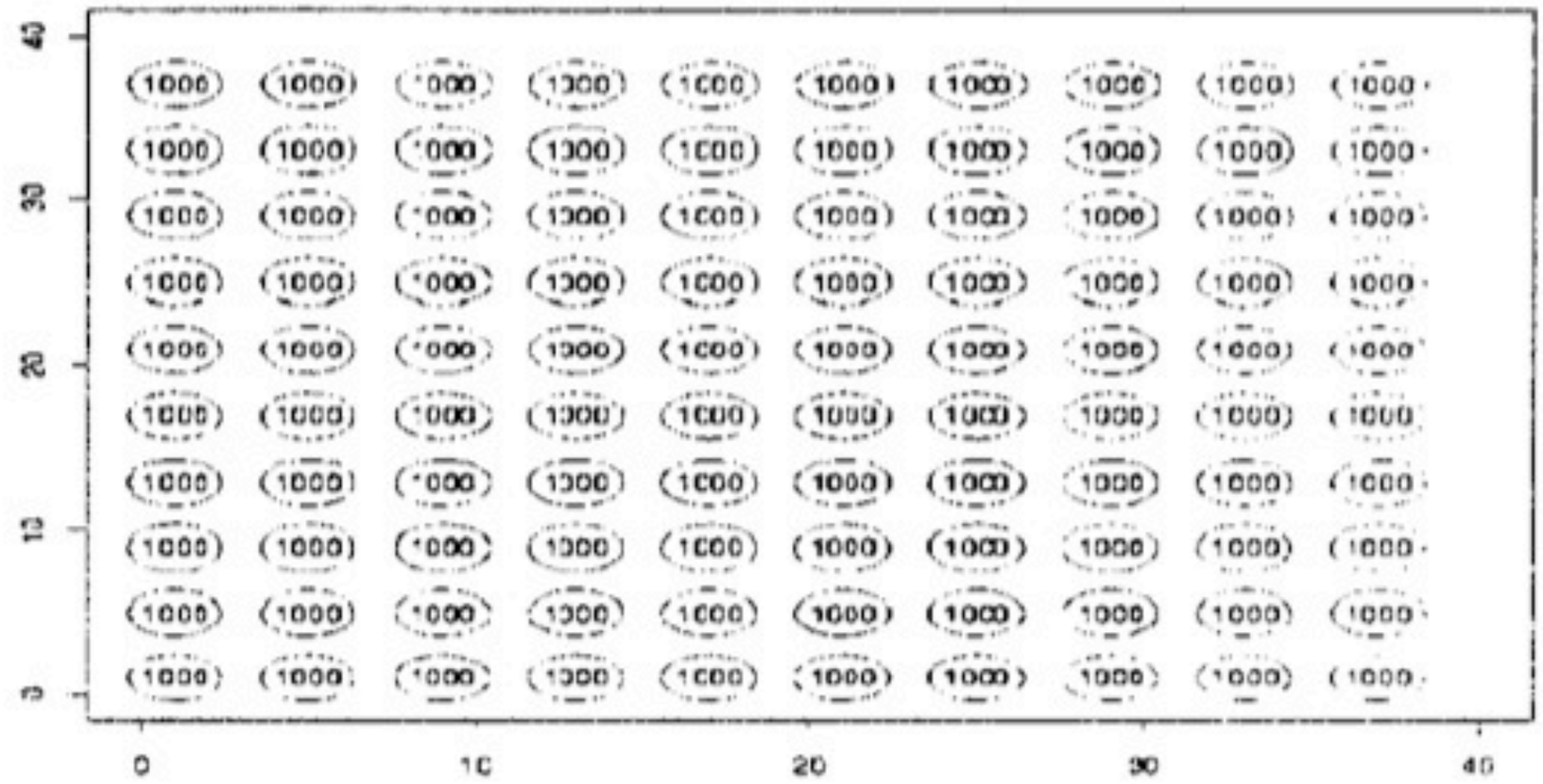
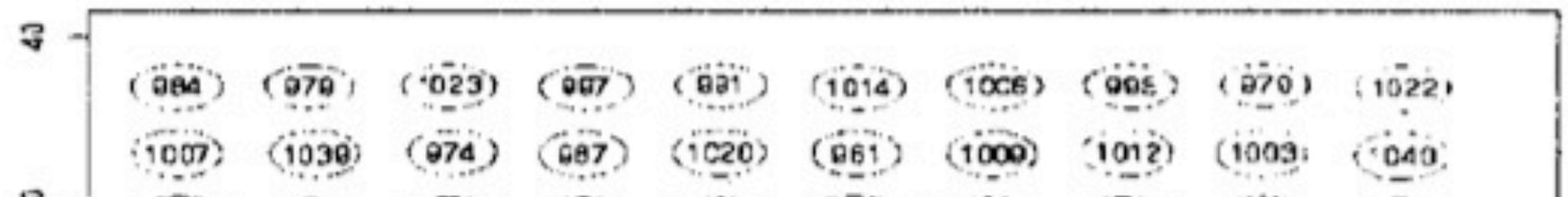


Figure 6: *Actual Clusters of DS1*



Ergebnisse

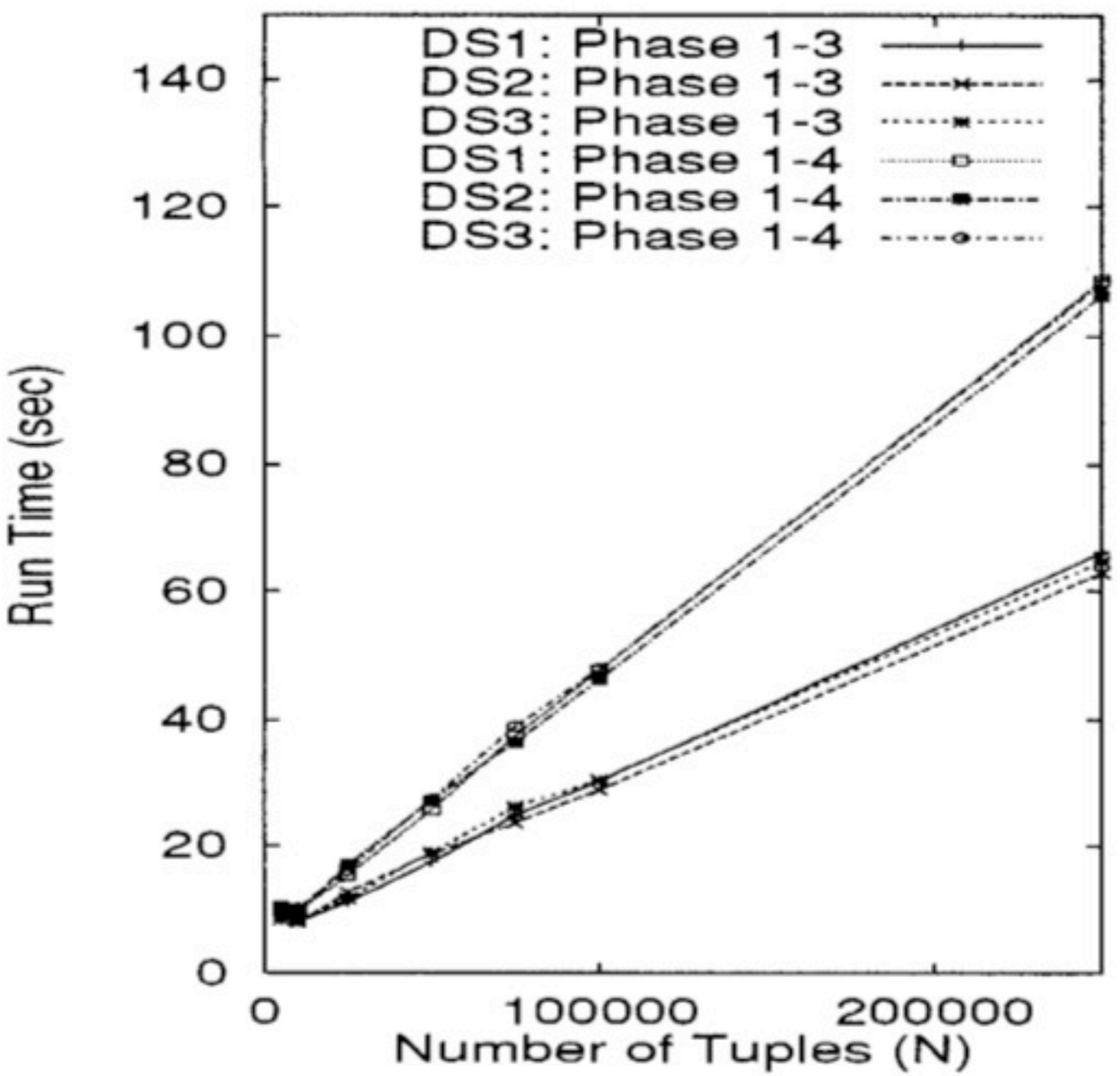


Figure 4: *Scalability wrt. Increasing* n_l, n_h

Ergebnisse

Figure 9: *The images taken in NIR and VIS*

BIRCH

An efficient data clustering method
for very large databases

Tian Zhang
Raghu Ramakrishnan
Miron Livny

1996 ACM SIGMOD

International conference on Management of data

267 (Citeseer) / 1701 (Google) Citations

2006 Test of Time Award Winners

Tian Zhang

Akademisch

1996 - 1999 (?)

IBM, ~1999

Santa Teresa Lab, San Jose

DB2 Entwicklung

Microsoft, ~2004

?

Raghu Ramakrishnan

Data Mining, Online Communities, Web-Scale Data Management

Akademisch

B.Tech. IIT Madras,1983

Ph.D. University of Texas at Austin,1987

Professor, University of Wisconsin-Madison, 1987-

ACM SIGKDD Innovation Award, 2008

Co-Founder QUIQ, 1999-2003

collaborative customer support & knowledge management

Ask Jeeves

Business Objects, Compaq, Sun...

Yahoo! Research, 2006-

Head of Community Systems Group

Chief Scientist for Audience and Cloud Computing

Miron Livny מירון לבני

High Throughput Computing, Visual Data Exploration,
Experiment Management Environments, Performance Evaluation

Akademisch

B.Sc. Physics and Mathematics, Hebrew University, 1975

M.Sc. Computer Science, Weizmann Institute of Science, 1978

Ph.D. Weizmann Institute of Science, 1984

Professor, University of Wisconsin-Madison, 1984-

Condor

High-Throughput Computing System

distributed parallelization of computationally intensive tasks

Open Science Grid

Recap

Fokus: Skalierbarkeit

- Baum immer komplett im RAM
- Single-Pass
- Daten ständig hinzufügen

CF-Tree: jeder Knoten ist ein Cluster

- Cluster Features reichen aus

Qualität

- Mehr RAM - bessere Qualität
- Outliers verbessern Qualität & Laufzeit