Web Mining

Analisa data dimensi tinggi dan Data Praproses

Prodi Teknik Informatika

Universitas Trunojoyo Madura

2024

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

- Menggabungkan dua atau lebih atribut (atau objek) menjadi satu atribut (atau objek)

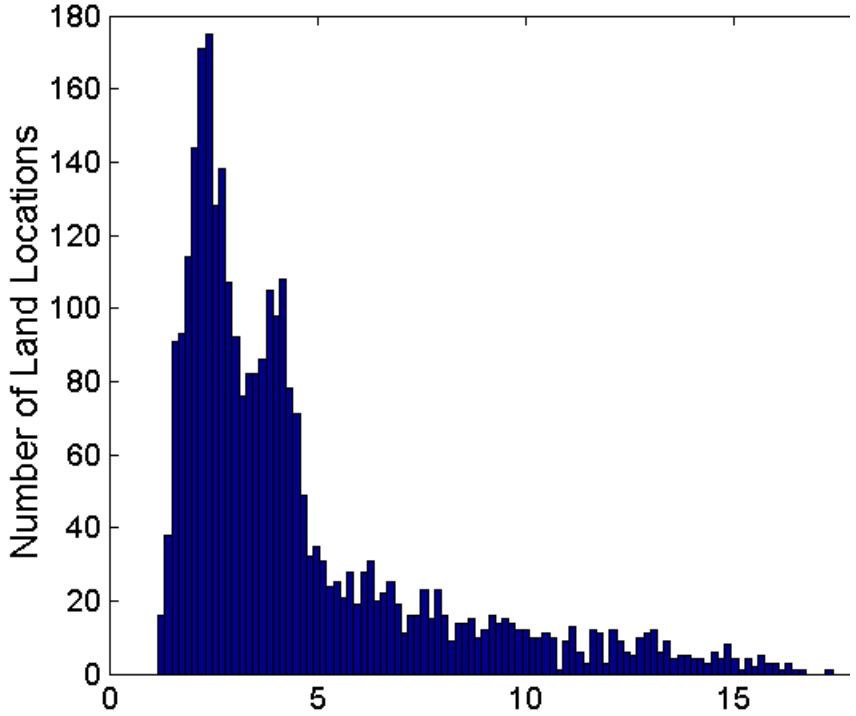
- Purpose
  - Data reduction
    - Mengurangi jumlah atribut atau obyek
  - Change of scale
    - Cities di agregasi kedalam regions, states, countries, etc.
    - Days di agregasi kedalam weeks, months, or years
  - More “stable” data
    - Data agregat cenderung memiliki variabilitas yang lebih sedikit

Example: Precipitation in Australia

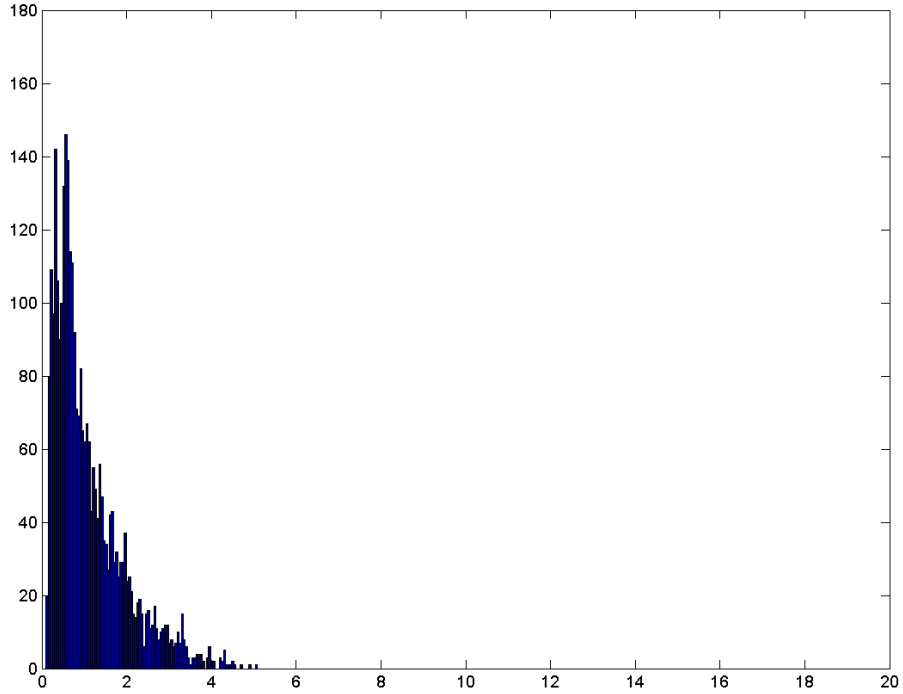
- Contoh ini berdasarkan curah hujan di Australia dari periode 1982 hingga 1993.
- Slide berikutnya menunjukkan
  - Histogram untuk deviasi standar rata-rata curah hujan bulanan untuk 3.030 sel grid 0,5◦ dengan 0,5◦ di Australia, dan
  - Histogram untuk deviasi standar curah hujan tahunan rata-rata untuk lokasi yang sama.
- Curah hujan tahunan rata-rata memiliki variabilitas yang lebih kecil daripada curah hujan bulanan rata-rata.
- Semua pengukuran curah hujan (dan simpangan bakunya) dalam sentimeter.

Example: Precipitation in Australia …

Variation of Precipitation in Australia



Standar Deviasi dari rerata Curah hujan bulanan



Standar Deviasi dari rerata curah hujan tahunan

Sampling

- Pengambilan sampel adalah teknik utama yang digunakan untuk reduksi data.
  - Sering digunakan untuk investigasi awal data dan analisis data akhir.

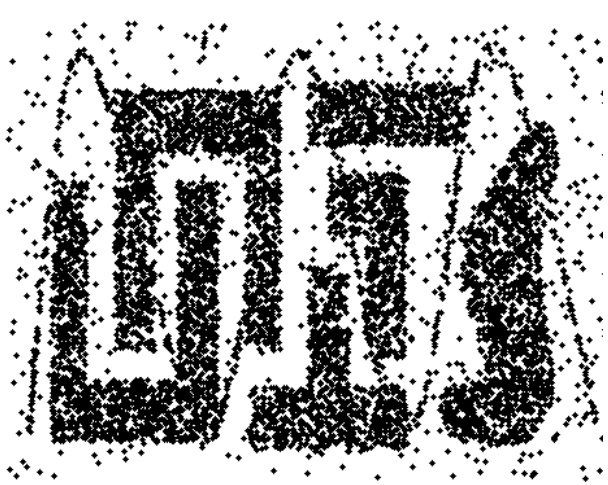
- Para ahli statistik sering kali mengambil sampel karena memperoleh keseluruhan rangkaian data yang diinginkan terlalu mahal atau memakan waktu.

- Pengambilan sampel biasanya digunakan dalam penambangan data karena pemrosesan seluruh rangkaian data yang diinginkan terlalu mahal atau memakan waktu.

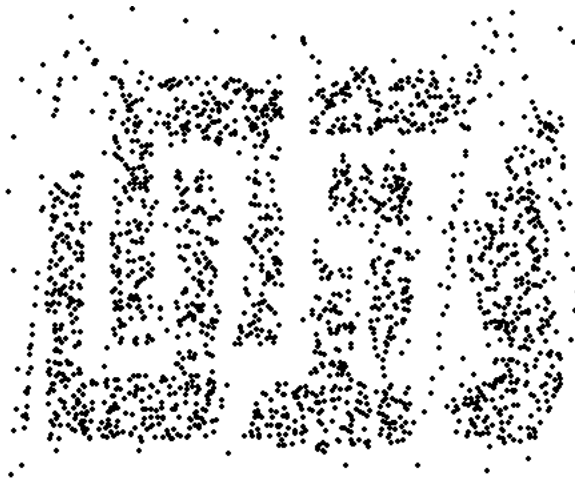
Sampling ...

- Prinsip utama untuk pengambilan sampel yang efektif adalah sebagai berikut:
  - Penggunaan sampel akan bekerja hampir sama baiknya dengan penggunaan seluruh set data, jika sampel tersebut representatif
  - Suatu sampel bersifat representatif jika sampel tersebut memiliki sifat-sifat (yang menarik) yang hampir sama dengan kumpulan data asli.

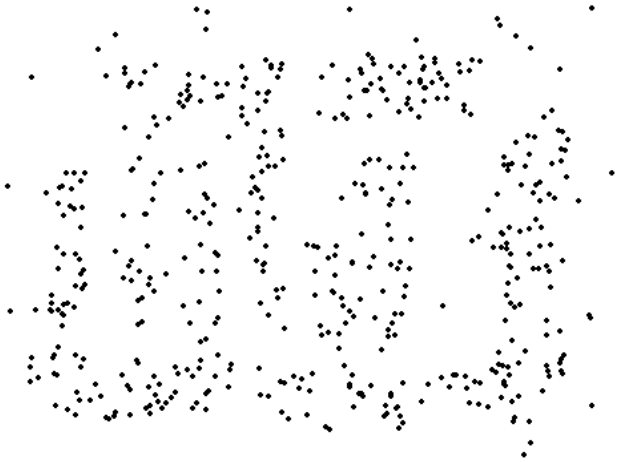
Sample Size



8000 points



2000 Points



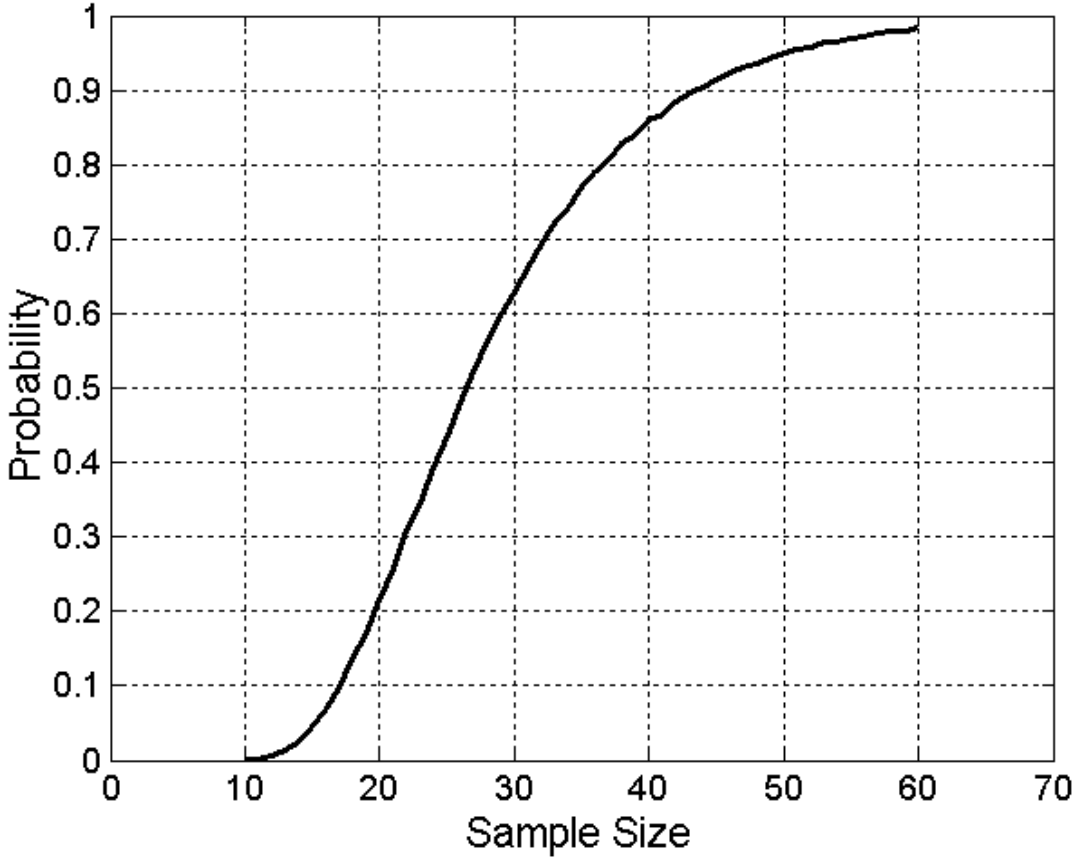
500 Points

Types of Sampling

- Simple Random Sampling
  - Ada kemungkinan yang sama untuk memilih item tertentu
  - Pengambilan sampel tanpa penggantian
    - Ketika setiap item dipilih, item tersebut dikeluarkan dari populasi
  - Pengambilan sampel dengan penggantian
    - Objek tidak dikeluarkan dari populasi karena dipilih untuk sampel.
    - Dalam pengambilan sampel dengan penggantian, objek yang sama dapat diambil lebih dari satu kali
- Stratified sampling
  - Membagi data menjadi beberapa partisi; lalu mengambil sampel acak dari setiap partisi

Sample Size

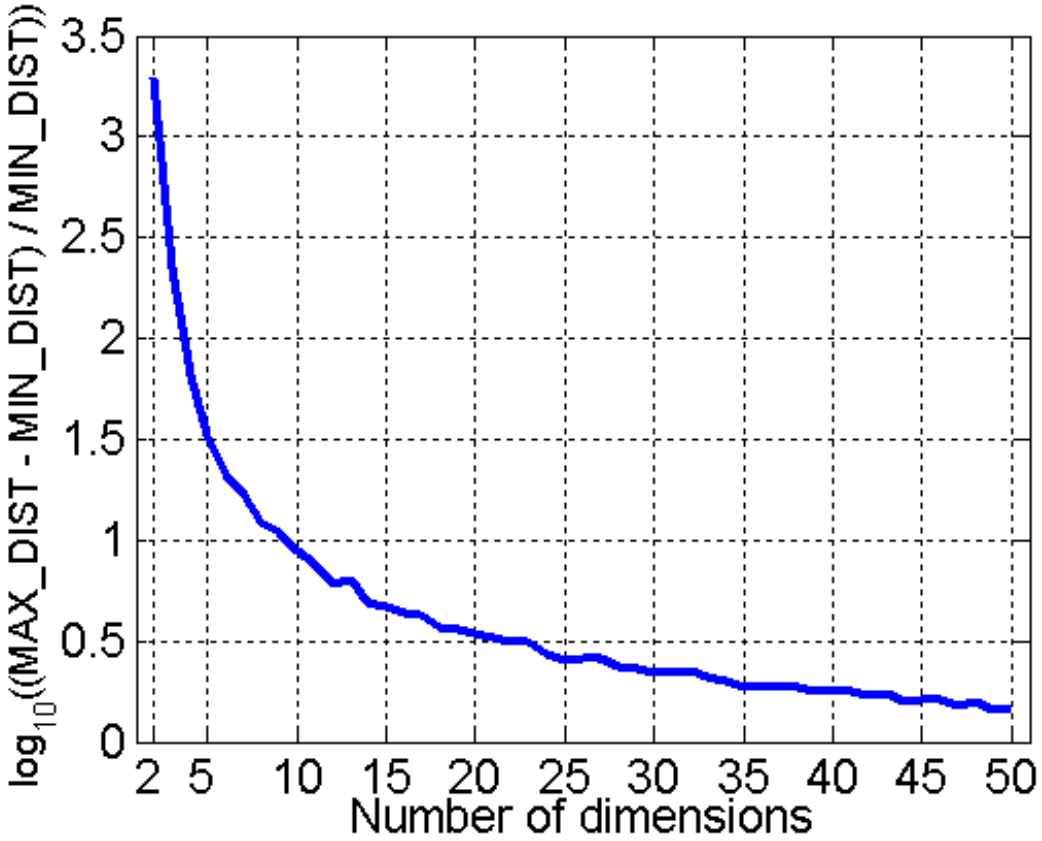
- **Berapa ukuran sampel yang diperlukan untuk mendapatkan setidaknya satu objek dari setiap 10 kelompok berukuran sama.**



Curse of Dimensionality

- Ketika dimensionalitas meningkat, data menjadi semakin jarang di ruang yang ditempatinya

- Definisi kepadatan dan jarak antar titik, yang penting untuk pengelompokan dan deteksi outlier, menjadi kurang bermakna



- **Hasilkan 500 poin secara acak**
- **Hitung perbedaan antara jarak maksimum dan minimum antara pasangan titik mana pun**

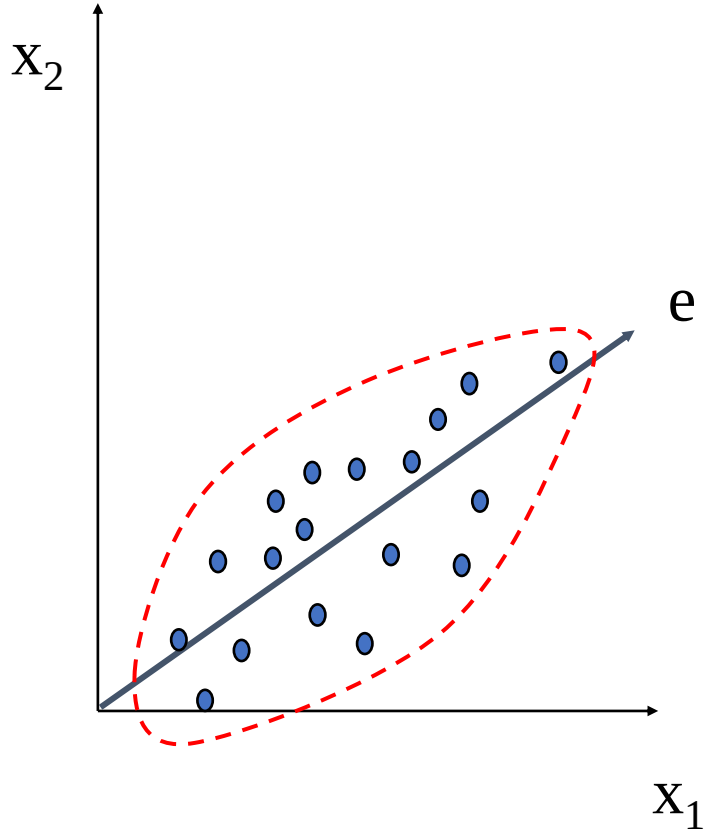
Dimensionality Reduction

- Purpose:
  - Hindari kutukan dimensionalitas
  - Mengurangi jumlah waktu dan memori yang dibutuhkan oleh algoritma penambangan data
  - Memungkinkan data divisualisasikan dengan lebih mudah
  - Dapat membantu menghilangkan fitur yang tidak relevan atau mengurangi kebisingan

- Techniques
  - Analisis Komponen Utama (PCA)
  - Dekomposisi Nilai Singular
  - Lainnya: teknik pengawasan dan non-linier

Dimensionality Reduction: PCA

- Tujuannya adalah untuk menemukan proyeksi yang menangkap jumlah variasi data terbesar



Dimensionality Reduction: PCA

Feature Subset Selection

- Cara lain untuk mengurangi dimensionalitas data
- Redundant features
  - Gandakan sebagian besar atau semua informasi yang terdapat dalam satu atau lebih atribut lainnya
  - Contoh: harga pembelian suatu produk dan jumlah pajak penjualan yang dibayarkan
- Irrelevant features
  - Tidak mengandung informasi yang berguna untuk tugas penambangan data yang sedang dilakukan
  - Contoh: ID siswa seringkali tidak relevan dengan tugas memprediksi IPK siswa.
- Banyak teknik yang dikembangkan, terutama untuk klasifikasi

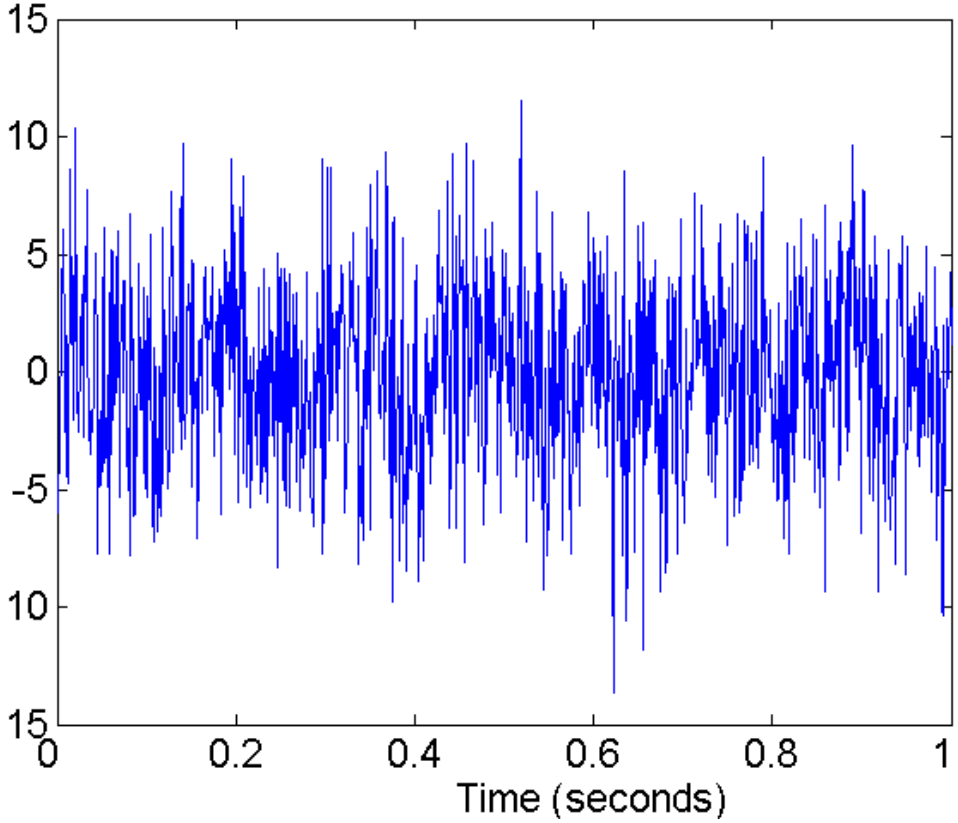
Feature Creation

- Buat atribut baru yang dapat menangkap informasi penting dalam kumpulan data jauh lebih efisien daripada atribut asli

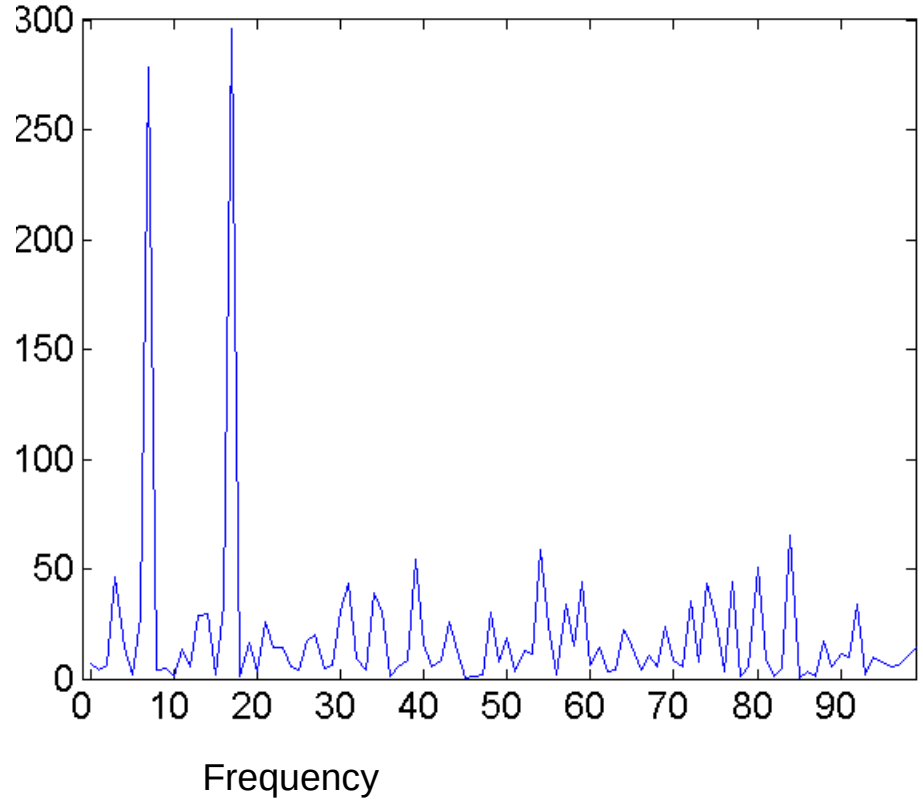
- Three general methodologies:
  - Feature extraction
    - Example: extracting edges from images
  - Feature construction
    - Example: dividing mass by volume to get density
  - Mapping data to new space
    - Example: Fourier and wavelet analysis

Mapping Data to a New Space

- Fourier and wavelet transform



Two Sine Waves + Noise



Frequency

Discretization

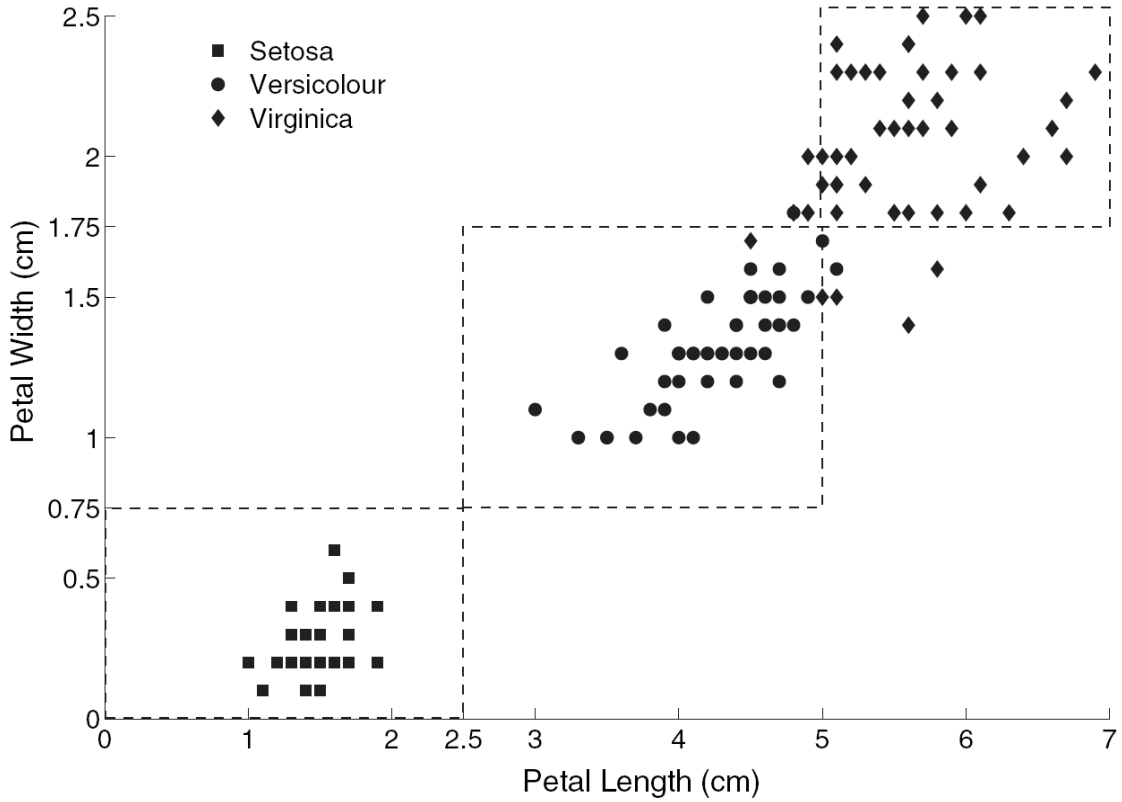
- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - Jumlah nilai yang berpotensi tak terbatas dipetakan ke dalam sejumlah kecil kategori
  - Diskritisasi umumnya digunakan dalam klasifikasi
  - Banyak algoritma klasifikasi bekerja paling baik jika variabel independen dan dependen hanya memiliki beberapa nilai
  - Contoh ilustrasi tentang kegunaan diskritisasi menggunakan pada dataset Iris

Iris Sample Data Set

- Iris Plant data set.
  - Can be obtained from the UCI Machine Learning Repository http://www.ics.uci.edu/~mlearn/MLRepository.html
  - From the statistician Douglas Fisher
  - Three flower types (classes):
    - Setosa
    - Versicolour
    - Virginica
  - Four (non-class) attributes
    - Sepal width and length
    - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland

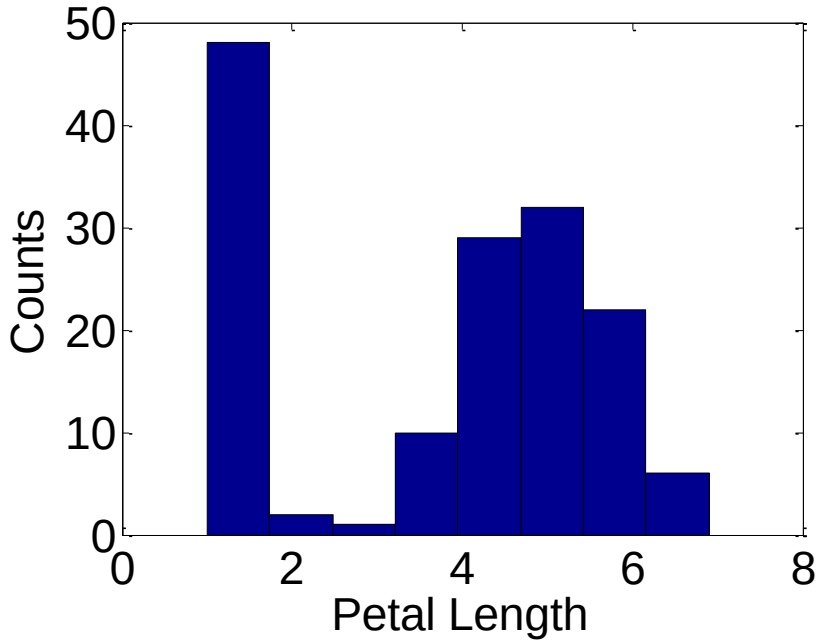
Discretization: Iris Example



Petal width low or petal length low implies Setosa.
Petal width medium or petal length medium implies Versicolour.
Petal width high or petal length high implies Virginica.

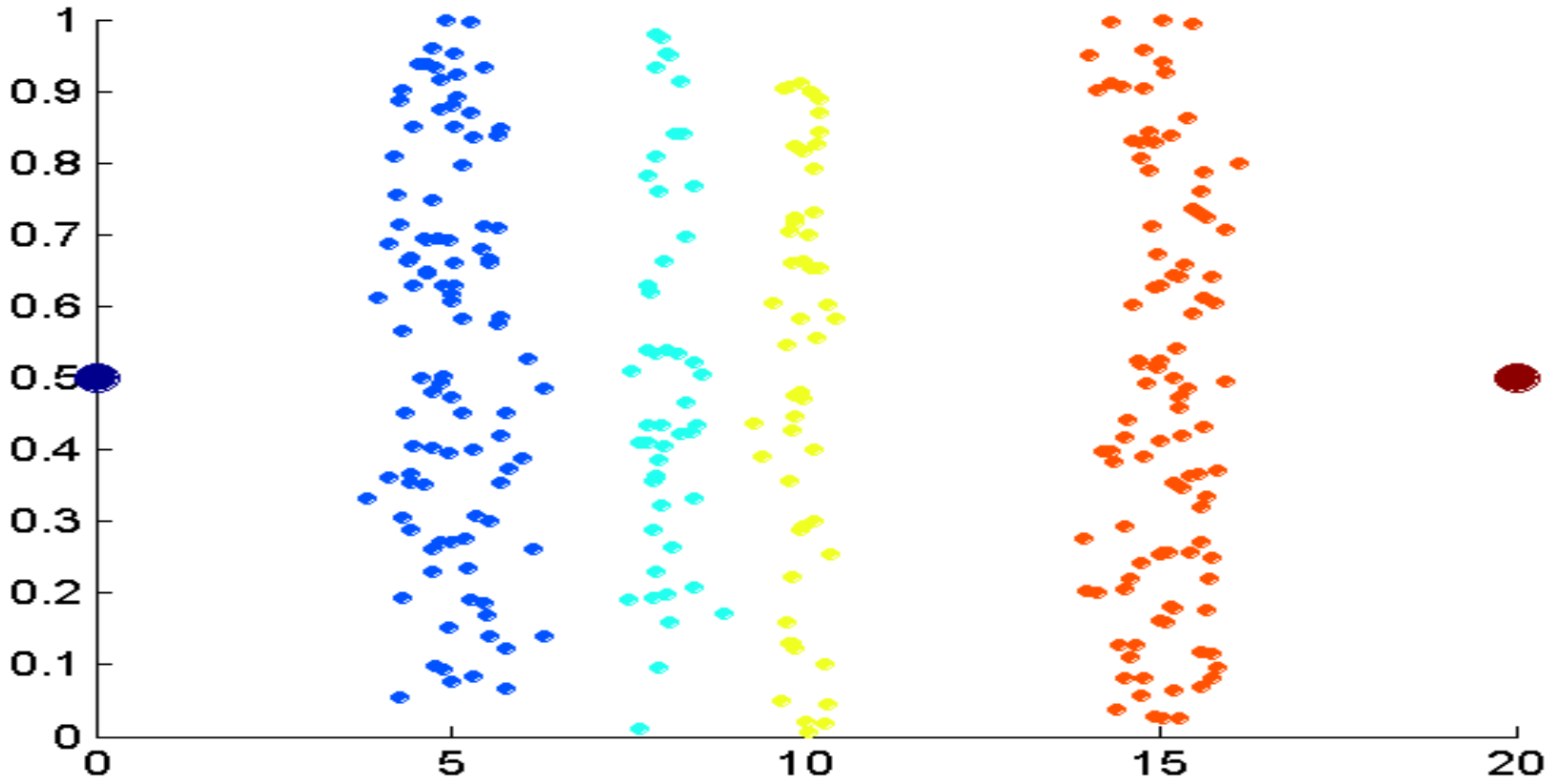
Discretization: Iris Example …

- How can we tell what the best discretization is?
  - Unsupervised discretization: find breaks in the data values
    - Example:
      Petal Length



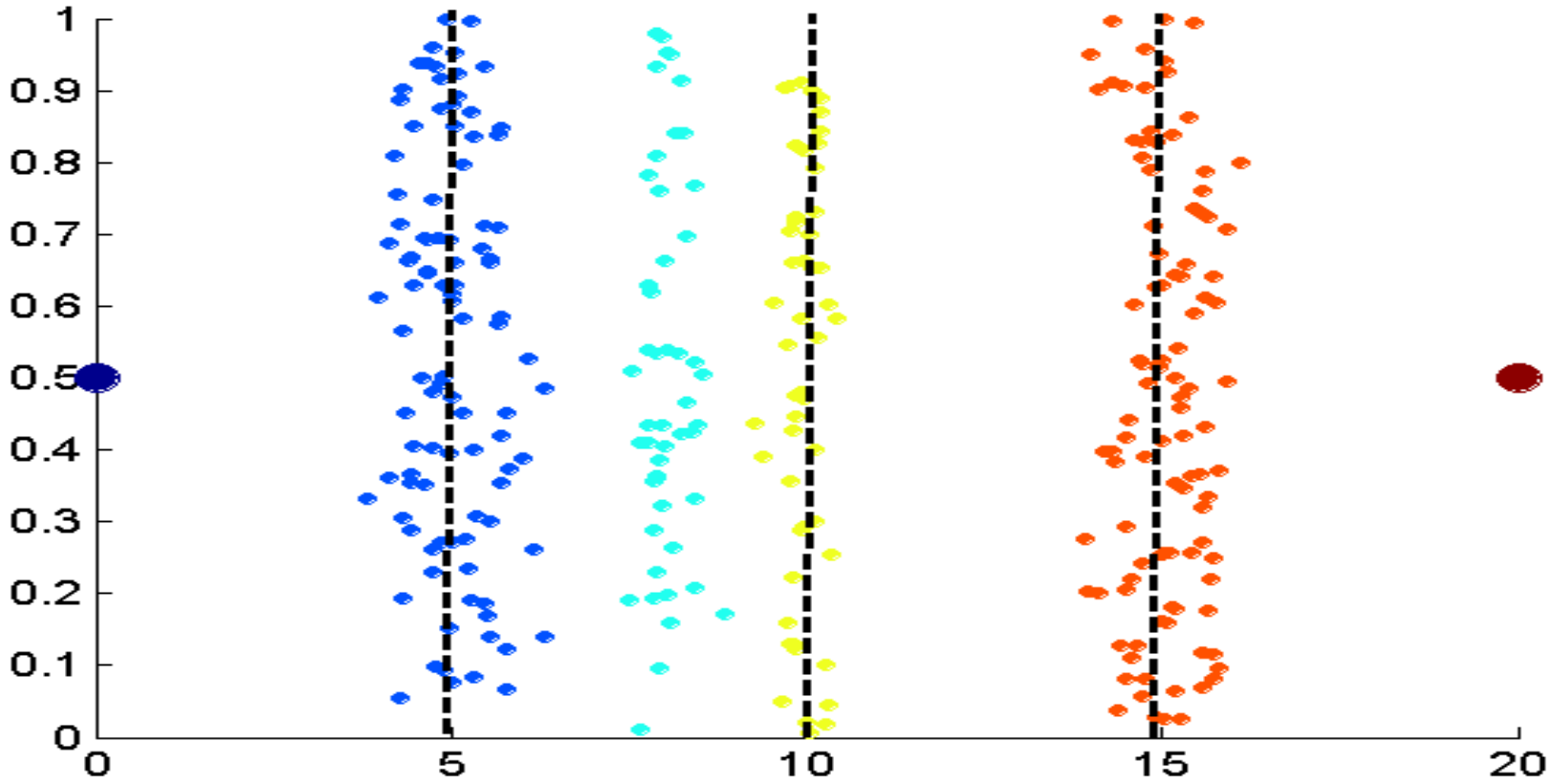
  - Supervised discretization: Use class labels to find breaks

Discretization Without Using Class Labels



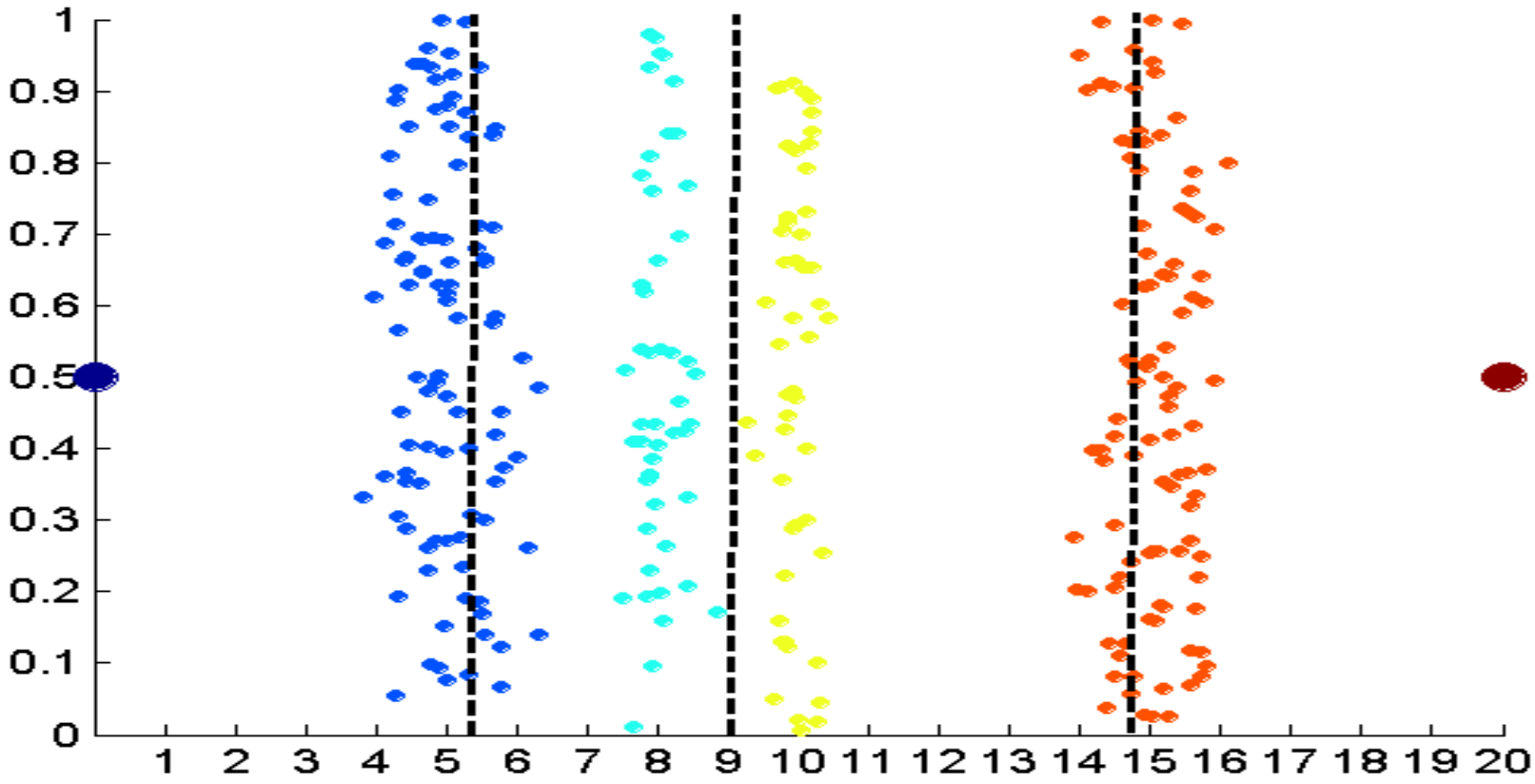
Data consists of four groups of points and two outliers. Data is one-dimensional, but a random y component is added to reduce overlap.

Discretization Without Using Class Labels



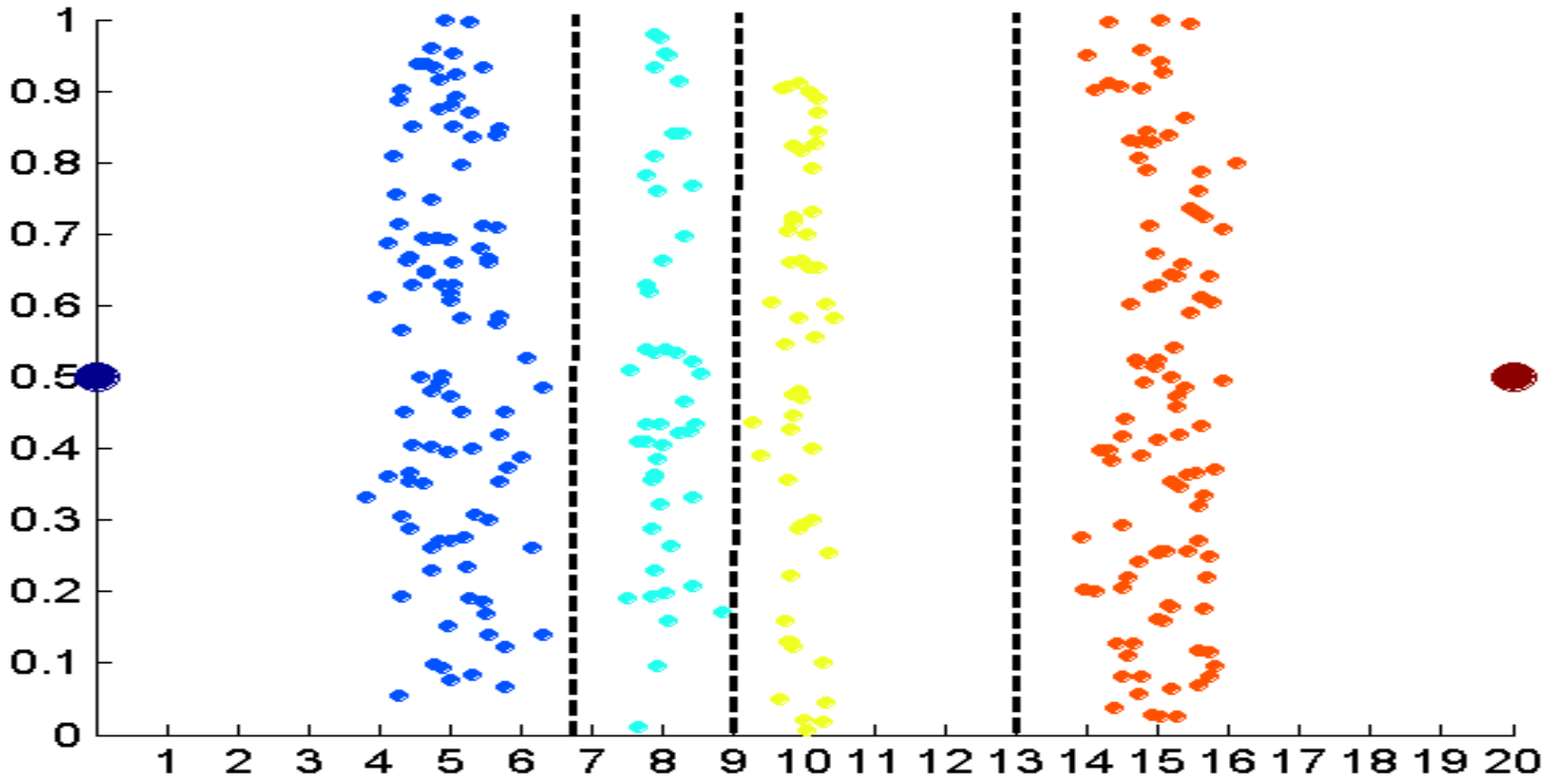
Equal interval width approach used to obtain 4 values.

Discretization Without Using Class Labels



Equal frequency approach used to obtain 4 values.

Discretization Without Using Class Labels



K-means approach to obtain 4 values.

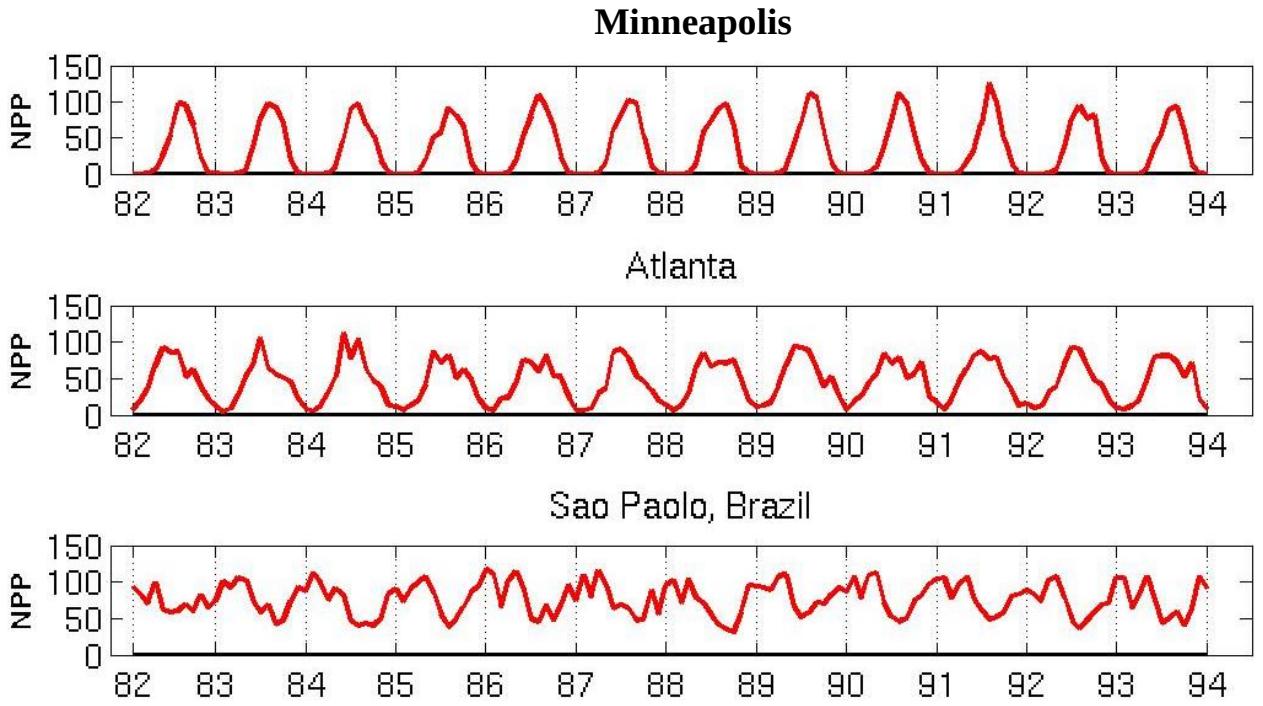
Binarization

- Binarisasi memetakan atribut kontinu atau kategorikal menjadi satu atau lebih variabel biner
- Biasanya digunakan untuk analisis asosiasi
- Sering mengubah atribut kontinyu menjadi atribut kategoris dan kemudian mengubah atribut kategoris menjadi sekumpulan atribut biner
  - Analisis asosiasi membutuhkan atribut biner asimetris
  - Contoh: warna mata dan tinggi badan diukur sebagai {rendah, sedang, tinggi}

Attribute Transformation

- Transformasi atribut adalah fungsi yang memetakan seluruh himpunan nilai atribut tertentu ke himpunan nilai pengganti baru sehingga setiap nilai lama dapat diidentifikasi dengan salah satu nilai baru.
  - Fungsi sederhana: x^k, log(x), e^x, |x|
  - Normalisasi
    - Mengacu pada berbagai teknik untuk menyesuaikan perbedaan antar atribut dalam hal frekuensi kemunculan, rata-rata, varians, rentang
    - Hapus sinyal umum yang tidak diinginkan, misalnya musim
  - Dalam statistik, standarisasi mengacu pada pengurangan rata-rata dan pembagian dengan standar deviasi.

Example: Sample Time Series of Plant Growth

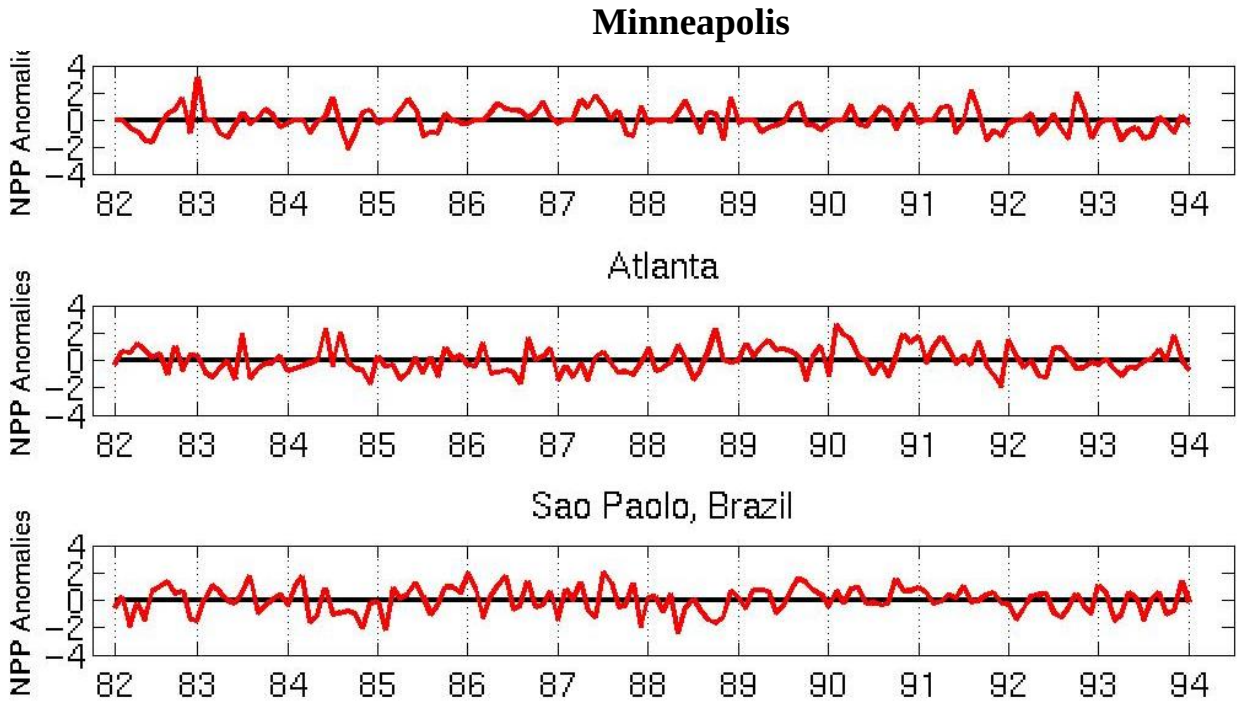


Net Primary Production (NPP) adalah ukuran pertumbuhan tanaman yang digunakan oleh ilmuwan ekosistem.

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.7591	-0.7581
Atlanta	0.7591	1.0000	-0.5739
Sao Paolo	-0.7581	-0.5739	1.0000

Seasonality Accounts for Much Correlation



Dinormalisasi menggunakan Skor Z bulanan:

Kurangi rata-rata bulanan dan bagi dengan deviasi standar bulanan

Correlations between time series

	Minneapolis	Atlanta	Sao Paolo
Minneapolis	1.0000	0.0492	0.0906
Atlanta	0.0492	1.0000	-0.0154
Sao Paolo	0.0906	-0.0154	1.0000