Clustering and the k-means Algorithm

David M. Blei

COS424 Princeton University

September 5, 2007

• Goal: Automatically segment data into groups of similar points

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:
 - Customers according to purchase histories

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:
 - Customers according to purchase histories
 - Genes according to expression profile

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:
 - Customers according to purchase histories
 - Genes according to expression profile
 - Search results according to topic

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:
 - Customers according to purchase histories
 - Genes according to expression profile
 - Search results according to topic
 - MySpace users according to interests

- Goal: Automatically segment data into groups of similar points
- Question: When and why would we want to do this?
- Useful for:
 - Automatically organizing data
 - Understanding hidden structure in some data
 - Representing high-dimensional data in a low-dimensional space
- Examples:
 - Customers according to purchase histories
 - Genes according to expression profile
 - Search results according to topic
 - MySpace users according to interests
 - A museum catalog according to image similarity

Clustering set-up

• Our data are

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

Clustering set-up

• Our data are

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

• Each data point is *p*-dimensional, i.e.,

$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

Clustering set-up

• Our data are

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

• Each data point is *p*-dimensional, i.e.,

$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

• Define a *distance function* between data, $d(\mathbf{x}_n, \mathbf{x}_m)$.

Our data are

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}.$$

• Each data point is *p*-dimensional, i.e.,

$$\mathbf{x}_n = \langle x_{n,1}, \ldots, x_{n,p} \rangle.$$

- Define a *distance function* between data, $d(\mathbf{x}_n, \mathbf{x}_m)$.
- Goal: segment the data into k groups

$$\{z_1,\ldots,z_N\}$$
 where $z_i \in \{1,\ldots,K\}$.



500 2-dimensional data points: $\mathbf{x}_n = \langle x_{n,1}, x_{n,2} \rangle$



• What is a good distance function here?



- What is a good distance function here?
- Squared Euclidean distance is reasonable

$$d(\mathbf{x}_n, \mathbf{x}_m) = \sum_{i=1}^{p} (x_{n,i} - x_{m,i})^2 = ||x_n - x_m||^2$$



• Goal: segment this data into k groups.



- Goal: segment this data into k groups.
- What should k be?



- Goal: segment this data into k groups.
- What should k be?
- Automatically choosing k is complicated; for now, 4.

k-means



• Different clustering algorithms use the data and distance measurements in different ways

k-means



- Different clustering algorithms use the data and distance measurements in different ways
- Begin with *k*-means, the simplest clustering algorithm

k-means



- Different clustering algorithms use the data and distance measurements in different ways
- Begin with *k*-means, the simplest clustering algorithm

• The basic idea is to describe each cluster by its mean value.

- The basic idea is to describe each cluster by its mean value.
- (Note: this works only for distances such that a mean is well-defined.)

- The basic idea is to describe each cluster by its mean value.
- (Note: this works only for distances such that a mean is well-defined.)
- The goal of *k*-means is to assign data to clusters and deine these clusters with their means.

1 Initialization

1 Initialization

• Data are **x**_{1:N}

Initialization

- Data are x_{1:N}
- Choose initial cluster means **m**_{1:k} (same dimension as data).

Initialization

- Data are x_{1:N}
- Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).

2 Repeat

Initialization

- Data are x_{1:N}
- Choose initial cluster means $\mathbf{m}_{1:k}$ (same dimension as data).

Repeat

1 Assign each data point to its closest mean

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Initialization

- Data are x_{1:N}
- Choose initial cluster means **m**_{1:k} (same dimension as data).

Repeat

1 Assign each data point to its closest mean

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Output the each cluster mean to be the coordinate-wise average over data points assigned to that cluster,

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\{n: z_n = k\}} \mathbf{x}_n$$

Initialization

- Data are x_{1:N}
- Choose initial cluster means **m**_{1:k} (same dimension as data).

Repeat

1 Assign each data point to its closest mean

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Ocompute each cluster mean to be the coordinate-wise average over data points assigned to that cluster,

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\{n: z_n = k\}} \mathbf{x}_n$$

3 Until assignments $z_{1:N}$ do not change






D. Blei Clustering 01



D. Blei Clustering 01





D. Blei Clustering 01



D. Blei Clustering 01

• How can we measure how well our algorithm is doing?

- How can we measure how well our algorithm is doing?
- The *k*-means objective function is the sum of the squared distances of each point to each assigned mean

$$F(z_{1:N},\mathbf{m}_{1:k}) = \frac{1}{2}\sum_{n=1}^{N}||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$



D. Blei Clustering 01











D. Blei Clustering 01



D. Blei Clustering 01

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

• Holding the means fixed, assigning each point to its closest mean minimizes *F* with respect to *z*_{1:*N*}.

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes *F* with respect to *z*_{1:*N*}.
- Holding the assignments fixed, computing the centroids of each cluster minimizes F with respect to m_{1:k}.

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to z_{1:N}.
- Holding the assignments fixed, computing the centroids of each cluster minimizes F with respect to m_{1:k}.
- Thus, *k*-means is a *coordinate descent* algorithm.

$$F(z_{1:N}, \mathbf{m}_{1:k}) = \frac{1}{2} \sum_{n=1}^{N} ||\mathbf{x}_n - \mathbf{m}_{z_n}||^2$$

- Holding the means fixed, assigning each point to its closest mean minimizes F with respect to z_{1:N}.
- Holding the assignments fixed, computing the centroids of each cluster minimizes *F* with respect to **m**_{1:k}.
- Thus, *k*-means is a *coordinate descent* algorithm.
- It finds a local minimum. (Multiple restarts are often necessary.)

Objective for the example data



Round of k-means

Compressing images



• Each pixel is associated with a red, green, and blue value

Compressing images



- Each pixel is associated with a red, green, and blue value
- A 1024 × 1024 image is a collection of 1048576 values (x₁, x₂, x₃), which requires 3M of storage

Compressing images



- Each pixel is associated with a red, green, and blue value
- A 1024 × 1024 image is a collection of 1048576 values (x₁, x₂, x₃), which requires 3M of storage
- How can we use *k*-means to compress this image?

Vector quantization





• Replace each pixel \mathbf{x}_n with its assignment \mathbf{m}_{z_n} ("paint by numbers").

Vector quantization





- Replace each pixel \mathbf{x}_n with its assignment \mathbf{m}_{z_n} ("paint by numbers").
- The k means are called the *codebook*.

Vector quantization





- Replace each pixel \mathbf{x}_n with its assignment \mathbf{m}_{z_n} ("paint by numbers").
- The k means are called the *codebook*.
- With k = 100, we need 7 bits per pixel plus 100×3 bits ≈ 897 K.

















Measure of distortion



Charlie Brown and Linus VQ Objective

• The objective gives a measure of how distorted the compressed picture is relative to the original picture
Measure of distortion



Charlie Brown and Linus VQ Objective

- The objective gives a measure of how distorted the compressed picture is relative to the original picture
- For more clusters, the picture is less distorted.

• In many practical settings, Euclidean distance is not appropriate. When?

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
 - Discrete multivariate data, such as purchase histories

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
 - Discrete multivariate data, such as purchase histories
 - Positive data, such as time spent on a web-page

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
 - Discrete multivariate data, such as purchase histories
 - Positive data, such as time spent on a web-page
- *k*-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
 - Discrete multivariate data, such as purchase histories
 - Positive data, such as time spent on a web-page
- k-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.
- No need to define the mean.

- In many practical settings, Euclidean distance is not appropriate. When?
- For example,
 - Discrete multivariate data, such as purchase histories
 - Positive data, such as time spent on a web-page
- *k*-medoids is an algorithm that only requires knowing distances between data points, $d_{n,m} = d(x_n, x_{m_k})$.
- No need to define the mean.
- Each of the clusters is associated with its most typical example

1 Initialization

- 1 Initialization
 - Data are **x**_{1:N}

- Initialization
 - Data are **x**_{1:N}
 - Choose initial cluster identities **m**_{1:k}

Initialization

- Data are **x**_{1:N}
- Choose initial cluster identities **m**_{1:k}
- 2 Repeat

Initialization

- Data are x_{1:N}
- Choose initial cluster identities **m**_{1:k}
- Repeat

1 Assign each data point to its closest center

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Initialization

- Data are x_{1:N}
- Choose initial cluster identities **m**_{1:k}
- Repeat
 - 1 Assign each data point to its closest center

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Provide the end of the end of

$$i_k = \arg\min_{\{n: z_n=k\}} \sum_{\{m: z_m=k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

Initialization

- Data are x_{1:N}
- Choose initial cluster identities **m**_{1:k}
- 2 Repeat
 - 1 Assign each data point to its closest center

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Provide the end of the end of

$$i_k = \arg\min_{\{n: z_n=k\}} \sum_{\{m: z_m=k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

Set each cluster center equal to their closest data points

$$m_k = \mathbf{x}_{i_k}$$

Initialization

- Data are x_{1:N}
- Choose initial cluster identities **m**_{1:k}
- Repeat
 - 1 Assign each data point to its closest center

$$z_n = \arg\min_{i\in\{1,\ldots,k\}} d(\mathbf{x}_n,\mathbf{m}_i)$$

Provide the end of the end of

$$i_k = \arg\min_{\{n: z_n=k\}} \sum_{\{m: z_m=k\}} d(\mathbf{x}_n, \mathbf{x}_m)$$

3 Set each cluster center equal to their closest data points

$$m_k = \mathbf{x}_{i_k}$$

③ Until assignments $\mathbf{z}_{1:N}$ do not change

• Choosing k is a nagging problem in cluster analysis

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
 - A certain required compression in VQ

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
 - A certain required compression in VQ
 - Clustering customers for k salespeople in a business

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
 - A certain required compression in VQ
 - Clustering customers for k salespeople in a business
- Usually, we seek the "natural" clustering, but what does this mean?

- Choosing k is a nagging problem in cluster analysis
- Sometimes, the problem determines k
 - A certain required compression in VQ
 - Clustering customers for k salespeople in a business
- Usually, we seek the "natural" clustering, but what does this mean?
- It is not well-defined.



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01



D. Blei Clustering 01

Heuristic: A kink in the objective



• Notice the "kink" in the objective between 3 and 5.

Heuristic: A kink in the objective



- Notice the "kink" in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.

Heuristic: A kink in the objective



- Notice the "kink" in the objective between 3 and 5.
- This suggests that 4 is the right number of clusters.
- Tibshirani (2001) presents a method for finding this kink.

• Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel

- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters
- Spatial and Statistical Inference of Late Bronze Age Polities in the Southern Levant (Savage and Falconer)
- Cluster the location of archeological sites in Israel
- Make inferences about political history based on the clusters
- Choose k very carefully, with a complicated computational technique.



• Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold
- Clustered genes based on their response in different tissues

- Coping with cold: An integrative, multitissue analysis of the transciptome of a poikilothermic vertebrate (Gracey et al., 2004)
- Exposed carp to different levels of cold
- Clustered genes based on their response in different tissues
- (No mention of how k = 23 was chosen.)



D. Blei Clustering 01

• Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.
- I.e., the levels of encouragement are corrected for

- Teachers as Sources of Middle School Students' Motivational Identity: Variable-Centered and Person-Centered Analytic Approaches (Murdock and Miller, 2003)
- Clustered survey results of 206 students
- Used the clusters to identify groups to buttress an analysis of what affects motivation.
- I.e., the levels of encouragement are corrected for
- Chose the number of clusters to get nice results

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Teacher caring	5	– .5 to .5	5 to .5	5	1.0
Peers' academic support	1.0	5	1.0	5	5 to .5
Parents' academic support	.5	-1.0	5 to .5	5 to .5	1.0

TABLE 3. Five-Cluster Solution: Z scores on Each Clustering Variable

TABLE 4. Means and Standard Deviations for Each Cluster on Grade 8 Motivational Variables

	Acad Self-E	lemic fficacy	Intri Valui Educ	insic ing of ation	Teacher-Rate Effort	
Cluster	М	SD	М	SD	М	SD
1. All positive	3.59	.48ª	2.99	.55ª	3.74	.26ª
2. Peer negative, parents very negative	2.44	.66 ^b	2.16	.51 ^b	3.05	.61 ^b
3. Peer positive	3.01	.73°	2.43	.66 ^b	3.26	.66 ^b
Negative teacher and peer	2.47	.63 ^b	2.24	.51 ^b	3.17	.59 ^b
5. Positive teacher and parents	3.19	.65°	2.89	.62ª	3.54	.47ª

• Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- *K* chosen for interpretability and "stability," which means that they could interpret multiple *k*-means runs on different data in the same way.

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- *K* chosen for interpretability and "stability," which means that they could interpret multiple *k*-means runs on different data in the same way.
- Draw the conclusion that patterns exist. What's wrong with this?

- Implications of Racial and Gender Differences in Patterns of Adolescent Risk Behavior for HIV and other Sexually Transmitted Diseases (Halpert et al., 2004)
- Clustered survey results of 13,998 students to understand patterns of drug abuse and sexual activity
- *K* chosen for interpretability and "stability," which means that they could interpret multiple *k*-means runs on different data in the same way.
- Draw the conclusion that patterns exist. What's wrong with this?
- k-means will find patterns everywhere!

Light exhema debber — Thing und on our current use of substances! 2 Convertient of the current used substances! or had sex 2 Sex debbers — all here had sex Abstances — none have ever used substances! or had sex Sex debbers — all here had sex thing and to of substances! Debbers — all convert down in part 1 2 mos. 90% every of holds — and 1 2 mos. 90% every of holds — and 1 2 mos. 90% every of holds — and 1 more current used substances! Machael and exist for all one of all one of the current of the cure	Cluster type and behavioral patterns	%
Noch barb das Noch barb das Noch barb das Sind dahlwart – all kann had sen Miskelanne of optimiser Sind und a condom at last sex information of a set of the Sind und a condom at last sex information of the set of the Sind und a condom at last sex information of the set of the Sind und a condom at last sex information of the set of the Sind und a set of	Light substance dabblers—infrequent or no current use of substancest	24
Abstantion 2 Scielabbritz	None have had sex	
See dashber,il travel had see Medium on gentrement Diffuent and consume al last see miniparts and diffuent algo had to be integrated travel field on goe Nore have all consume algo had to goe Nore have all does all does all histoge at travels and see all histoge at travels and see Rocksman - all travels, deck and brige drink with moderate frequency of the head see miniparts of the histoge and the head had see all histoge at the see the see all does all have had see all used all cohol filled dugs. The head see miniparts of the histoge and the head had see all head had see the see all head had have all have had see all used all hold had see the see all head had have all have had see all used all hold head had see the see all head had head have all have had see all used all hold head had with head see the set of head had head have had used all used all hold head had head with head see the set of head had head have had used all used all hold head had head with head head head with head head head have had used all head had head head head head head with head head head whead had head head head head head head h	Abstainersnone have ever used substances† or had sex	22
Median can dipatence-1 Median can discontential database Median can discontential database Median can discontential database Median can discontential database Median can discontential database Sincker	Sex dabblers—all have had sex	14
Additused according at last see Directors — all consumed alcohol in past 12 mos. 2 Directors — all consumed alcohol ingo to 2 mos. 2 Report Ling - official see Report	Median no. of partners=1	
Integrate out Substantists" Integrate out Substantists Integrations o	60% used a condom at last sex	
Drickser_and Laroxim ed abold in past 12 mos. wherper blag where in Silk drug use wherper blag where in Silk drug use brankser-and in Silk drug use Sinkhen blag is Sinkhen	Intrequent use of substancest	
ether report to high ding user ether have had use frace have had used frace had used used had used frace had used used had u	Drinkers—all consumed alcohol in past 12 mos.	5
hrfeguett nor Bill drug ues Sincker- al Insole diparettes dally Sincker- al Insole diparettes dally Sincker- al Insole diparettes dally Sincker- al Insole diparettes dally Sincker- and Sincker- al Insole diparettes dally Sincker- and Sincker- al Insole diparet Sincker- and Sincker- al Insole diparet Sincker- and Sincker- al Insole diparet Sincker- and Sincker- and Sincker- Sincker- and Sincker- Sincker- and Sincker- Sincker- and Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker- Sincker Sincker- Sincker- Sincker Sincker Sincker- Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker Sincker	49% report binge drinking	
Noch have had ses Smokers—all mode cignettes daily principant backcolida days and Shahave had ses Shahave had ses Shahav	Infrequent or no illicit drug use	
Sinckers-all mode cignetts daily infrequent cignet darkalificat daips Sinh harb id as: Image: Since Sinc	None have had sex	
hinkquent care of alcoholithic drugs Carbohan bad ser Machahan bad ser Machahan bad ser Machahan bad ser Minkquent disacted filler drugs use Dispa of hinkquent disacted filler drugs use dish ban bad ser Minkquent disacted filler drugs de disk with moderate frequency dish ban bad ser Machahan bad ser Machah	Smokers—all smoke cigarettes daily	7
Cit-hise bid set: Archie and set: Cabbers—II direk occasionally all have had set: Infrequent tabescorificit drug use Direk archites—II shape frequently Briege artites—II shape frequently SHA have had set: Have y dabbler—II innote, dirik and bringe direk with moderate frequency SHA have had set: Have y dabbler—II innote, dirik and bringe direk with moderate frequency SHA have had set: Have y dabbler—II innote, dirik and bringe direk with moderate frequency SHA have had set: Have y dabbler—II innote, dirik and bringe direk with moderate frequency SHA have had set: Have y dabbler—II innote, dirik and bringe direk with moderate frequency SHA have had set: Have	Infrequent use of alcohol/illicit drugs	
Akabha aks chibher — di khris kacalionaliy ali hae had ses Mingan to takaca film day uas Sitting a chicken — di hang fengan di Shris hae had ses Shris hae had se	62% have had sex	
hinkngsen tradesconfilier drog uses bildings of hiskes — of his was and of her drog uses displayed hiskes — of his was and of her drog uses displayed hiskes and displayed drog was displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with moderate frequency displayed hiskes — all mode, direk and broge drick with a dot for displayed hiskes displayed hiskes — all mode, displayed hiskes and used alcoholitiki drug at last sex displayed hiskes — all mode, displayed hiskes and used alcoholitiki drug at last sex displayed hiskes — all mode, displayed hiskes hiskes for drug or money displayed hiskes — all law end sex — all use marking frequently, all have had sex displayed hiskes — all law end sex. All uses and sex — all uses of hiskes hiskes and displayed hiskes — all law end sex. All uses and sex bickes mode has exes with males— all are males who have had uses with another male displayed hiskes moderates — sex with males—males. Who have had uses with another male displayed hiskes moderates — sex with males—males. Who have had uses with another male displayed hiskes moderates — sex with males—males. Who have had uses with another male displayed hiskes moderates.	Alcohol and say dabblers - all drink accasionally all base had say	
Book of the second	Infrequent tobacco/illicit drug use	
Binge chineses—all bing lengently 4 Binge chineses—all bing lengently 4 When here does 4 Havey dabbles—all mode dirks and binge darkwith moderate frequency 2 When here does 2 Mary dabbles are all mode dirks and binge darkwith moderate frequency 2 When here does 2 Mary dabbles are all mode dirks and binge darkwith moderate frequency 2 When here does 2 Mary dabbles are all mode dirks and binge darkwith moderate frequency 2 Mary dabbles are all one marijkann fequently/leve have used other illicit drugs 1 Mary shake does 2 Mary here hold sex 1 Mary here hold sex 1 <		
Infrequent cignetis, marijuana and other ding use Shihave had see Shahave had see Shahave had see Shihave had see Shihav	Binge drinkers—all binge frequently	4
An in rights 1 Interview. See	Infrequent cigarette, marijuana and other drug use	
An uniter subset. Very dabbier, and lange, drink with moderate frequency APP use many dabbier, and lange, drink with moderate frequency APP use many dabbier, and the set of the	60% break had see	
Heavy dables—all mode, dirk and bring dirk with moderate frequency 3 When any diam, Key use of hell life Cognitive 3 When he dire 3 Combination sees. 4 Marginan asses	A DID Have Had Sex	
45% use majuana, few use other illicit drugs 15% have had our Combination sex and drug use—all have had sex, all used alcoholfikiti drug at last sex 25% have had our 25% have had our our our details use had sets of drug or money 25% have had our our our details use had sets of drug or money 25% have had our our our our details use had sets of drug or money 25% have had our our our our our sets of a sets or money 25% have had our our our our our sets of a sets or money 25% have had our our our our our our sets of a set of a sets or money 25% have had our our our our our our our our sets of a sets or money 25% have had our	Heavy dabblers—all smoke, drink and binge drink with moderate frequency	3
19% have had one: Incombation one: Incombation one: Incombation one: Mangiuma users—all lose marijuana frequently few have load other Illicit drugs Incombation one: Incombation one: 19% have had one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% non-bad one: Incombation one: Incombation one: Incombation one: 19% have had one: Incombation one: Incombation one: Incombation one: 19% have had one: Incombation one: Incombation one: Incombation one: 19% have had one: Incombation one: Incombation one: Incombation one: 19% have had one: Incombation one: Incombation one: Incombation one: 10% have had one: Incombation one: Incombation one: Incombation one: 10% have had one: Incombation one: Incombation one: Incombati	45% use marijuana; few use other illicit drugs	
Combination see and drug use—all have had see, all used alcoholdlich drug all last see. 'I all comparison for any starting frequently flew have used other illich drugs 'I all set see all see and see all last see a	91% have had sex	
Margiusan asses—all use margiusans frequently: few have used other III of drugs % % include capatets ***********************************	Combination sex and drug use—all have had sex; all used alcohol/illicit drug at last sex	
Versional actional Versional actional Versional actional Versional actional Versional Versional Version Versional Version Versional Version Versional Version Versional Version Versional Version Vers	Marijuana users—all use marijuana frequently: few bave used other illicit druos	
79% smoke disperties 79% smoke disperties 79% smoke disperties Multiple partners—III reports 34 secul partners Multiple partners—III reports 34 secul partners Ser for drogs ar money—IIII have had set for drugs or money Serveron Tow more moders Serveron Tow more moders Serveron Tow more moders Serveron Tow more moders III Multiple and the disperties III Multiple and the disperties III Multiple and the disperties IIII Multiple and the disperties IIII Multiple IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	94% use alcohol	1
Yeh have had sex Multiple partners— Jill report 31.4 sexual partners YSH report tow or moderate use of kubancest SSH report tow or moderate use of kubancest SSH report tow or moderate use of kubancest High melliphican sus and sex—III use manifusion frequently, all have had sex. It was alkachold other days all as sex USH have had sex. SSH have ha	79% smoke cigarettes	
Multiple partners—III report 134 execusion protons Senfordrugs or moderate used substantist Senfordrugs or moderate used substantist Medians ou of partnersed- Medians ou for anticed- Medians ou for anticed- Medians ou for anticed- Medians and exec_all use manifulanta frequently all have had sec Medians and exec_all use manifulanta frequently all have had sec Medians and anticed errors users—95% report heavy mainfulanta use; all use other illicit drugs of the have had 31 partners (mediantistic) Medians ou for anticed- Medians ou for anticed- main and the execution and anticed and a sec section of approxes—4 Made who have execution males—all are makes who have had sex with another male of the have had anticed memory.	74% have had sex	
75% inportion or moderate use of obstances? 55% inport to over moderate use of obstances? 55% inport of the over moderate use of obstances? 14(b) in ost (partners) 14(b) in ost (partners) 15% have had so > (part	Multiple partners	
Ser for drugs or money—all have had set of drugs or money I Shi report for our moderate use of Justancen' Shi report for our moderate use of Justancen' Shi report for our moderate use of Justancen' Shi have had set.	75% report low or moderate use of substances!	1
Sex Ger drugs or money — all have had sets for drugs or money 1 Sex Ger drugs are money — all have had sets for drugs or money With an out of grantmen-3 With an out of grantmen-3 EXE have had sets the last sets are all have had sets of grantmen-3 EXE have had sets the last sets are all have had sets of grantmen-3 Sets have had sets are all have high sets drugs are all have had sets of grantmen-3 Sets have had sets are all have high sets drugs are all have had sets of grantmen-3 Sets have had sets are all have high sets drugs are all have had sets of grantmen-3 Sets have had sets are all have high sets drugs are all have had sets of grantmen-3 Sets have had sets are all have high sets drugs are all have had sets of the sets drugs are all have high sets drugs have had sets with handber male of the have had sets with handber male of the have had sets with handber male (block have had sets mention).		
Soft report to we moderate use of ubstancest High mediuma use and sec—all use marginana frequently all have had sec Migh mediuma use and sec—all use marginana frequently all have had sec Soft have had soft of the soft o	Sex for drugs or money—all have had sex for drugs or money	1
Median no. of partners-3 Na used alcohol other drug at last see (25 km beh al-1) partner (median-6) Marguita and other drug user-95% report heavy marijuana use; all use other illicit drugs of Set Na head al-1 partner (median-6) Marguita and other drug user-95% report heavy marijuana use; all use other illicit drugs of Set Na head al-1 partner (median-6) Set Na head and (26 km beh al-1) partner (26 km beh al-2) Set Na head head (26 km beh al-2) Set Na head head (26 km beh al-2) partner Marine (26 km beh al	50% report low or moderate use of substances†	
High meruphican use and sec.—III use markican frequently: all have had sec. If March and And More for use and as sec. If Stri have had sec. If	Median no. of partners=3	
Micesi atknowledge in the second seco	High marijuana use and sex—all use marijuana frequently: all have had sex	
12% have held >1 putter (modianed) Mergluna and other (modianed) Sife have had sex Sife have had sex Sife have had sex Sife have had sex Note had had a set of the sex Sife have had asex Sife have had asex with males—all are males who have had sex with another male Of the have had sex Sife have had asex Sife have had asex Sife have had asex Sife have had had asex Sife have had had be partners in moders Sife had had be partners Sife had had had be partners Sife had had be partners	All used alcohol/other drug at last sex	
Margiusan and other drugs users—95% report heavy marijuana use; all use other illicit drugs of Bith have had set. 28% used all coholoholther drug at last set. Injection drug users—ail have injected drugs 20% have had users—ail have injected drugs 10% have fault as set with males—ail at males who have had set with another male 10% user drugs at last 30.00%. 20% user drugs at last 30.00%.	82% have had >1 partner (median=6)	
Set Name National Set	Marilyana and other drug years — 05% report hereas marilyana year all years the West days	
22% used advances on the set of t	58% have had sex	
injection-drug users—ail have injected drugs of 27% have had see Mades who have see with males—ail are males who have had see with another male (Makes who have had joes 30 days) 20% used mainfaired in past 30 days.	28% used alcohol/other drug at last sex	
Impection-Implication guester—an invertigated arougs of Welliam no of partners—4 Males who have sex with males—all are males who have had sex with another male of how many manufacture parts to (medianics). White well advide 11 memory.	interation de la constant	
ux minuter will sea Males who have sea with males—all are males who have had sea with another male Who were dannible partners (median-s) UKH were danijuana in past 30 days	injection-drug users—all have injected drugs	C
Males who have sex with males—all are males who have had sex with another male 0 We used multiple partners (median=5) We used another 21 time/mo.	Median on of partners-4	
Males who have sex with males—all are males who have had sex with another male 0 8% have had multiple partners (median=5) 10% used (not loc) 2 time/mo.	and a second	
78% have had multiple partners (median=5) 10% used marijuana in past 30 days 10% use alcohol 21 time/mo.	Males who have sex with males—all are males who have had sex with another male	0
40% used marijuana in past 30 days i0% use alcohol ≥1 time/mo.	78% have had multiple partners (median=5)	
50% use alcohol ≥1 time/mo.	40% used marijuana in past 30 days	
	oU% use alcohol ≥1 time/mo.	

D. Blei

Clustering 01

Summary