



**IIT School of Applied Technology**

ILLINOIS INSTITUTE OF TECHNOLOGY

**information technology & management**

# **527 Data Analytics**

February 2, 2016

Week 4 Presentation

# Week 4 Topic: Agenda

- ◆ Assignment Q&A
- ◆ Cluster Analysis

# Week 4 Topic: Defining Cluster Analysis

“*Cluster analysis* is a set of methods for constructing a (hopefully) sensible and informative classification of an initially unclassified set of data, using the variable values observed on each individual.”

Everitt (1998), *The Cambridge Dictionary of Statistics*

# Week 4 Topic:

## It's about Pattern Discovery

- ◆ ***Cluster*** is a group of similar objects (observations, customers, patients, buyers, locations, etc.)
- ◆ ***Cluster Analysis*** is a set of data-driven partitioning techniques designed to group a collection of objects into clusters. Its about data explorations, searching for patterns in complex data, that is conducted in repetitive fashion. Finding these patterns can lead to business decisions.

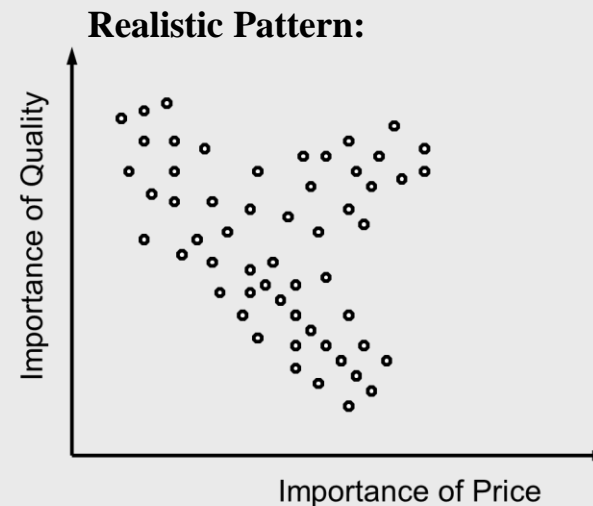
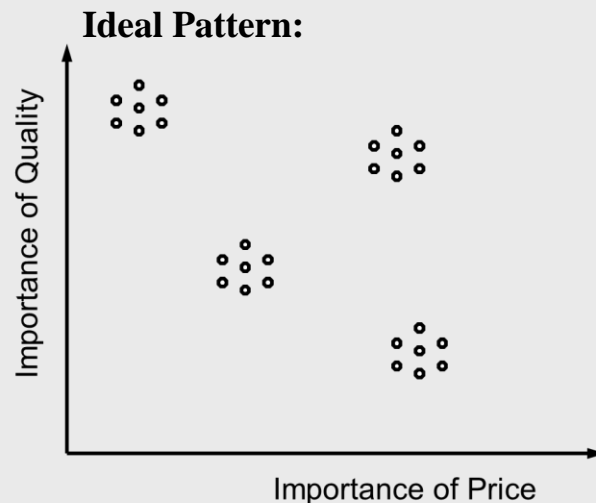
# Week 4 Topic:

## Further definitions - Cluster Analysis

- ◆ *Cluster analysis* is the generic name for a wide variety of procedures that can be used to create a classification of entities/objects
- ◆ *Cluster analysis* is a **convenient** method commonly used in many disciplines to categorize entities (individuals, objects, and so on) into groups that are **homogenous** along a range of observed characteristics (variables).
- ◆ The goal of cluster analysis is to partition/classify data into groups/objects (in our case, individuals, households or families) so that each object in a cluster is similar to the other objects in the same cluster; however objects in different clusters are dissimilar to each other.
- ◆ Therefore, if classification is successful, the objects within the cluster will be close together when plotted geometrically and different clusters will be far apart. A plot (showing clearly separated clusters) in two dimensions is shown on the next slide.

# Week 4 Topic: Cluster Analysis – Grouping

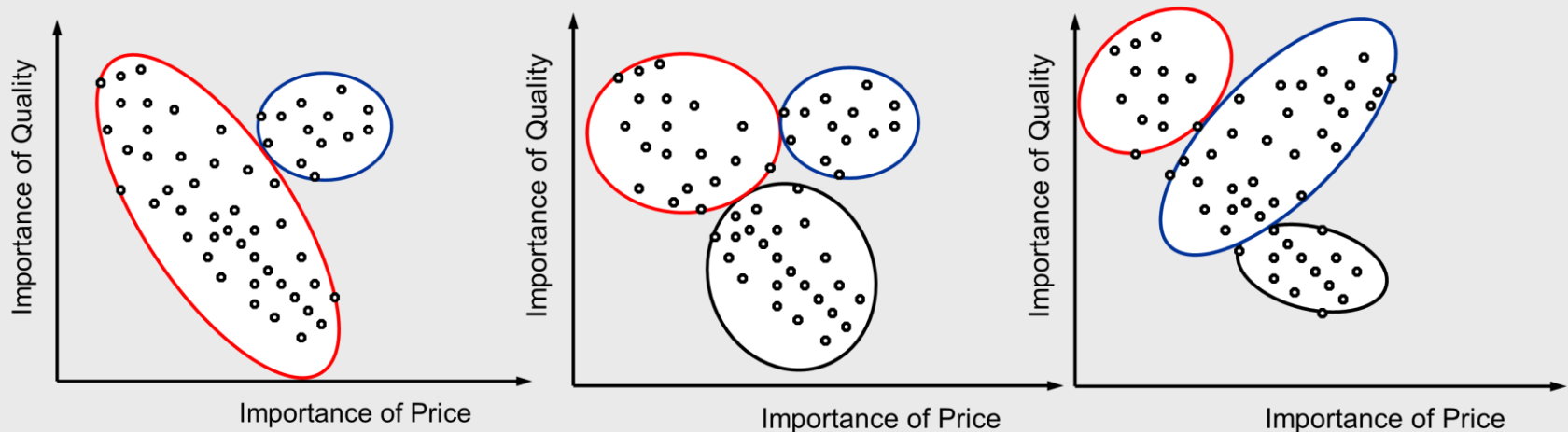
- ◆ Realistically speaking, it is highly unlikely to find natural groups that are very well separated, even in the two-dimensional space. It is more likely that a two-dimensional plot of customer importance variables will produce a messy, not well-separated plot that could indicate the possibility of different numbers of groups depending on how we view and choose to partition the data.



# Week 4 Topic:

## Cluster Analysis – Grouping Options

- ◆ Many other cluster (grouping) solutions could possibly be seen in this simple data set. Clearly, the real-world data will be much more complicated (greater number of variables/dimensions, each variable may be measured with different measurement scales, possibility of measurement error, possibility of random error, missing values, extreme values, etc.) than this simple two-dimensional example.



# Week 4 Topic:

## Clustering Guidelines - 1

Which variables should be used for clustering?

- ◆ The variables for clustering should primarily be chosen based on business objectives for segmentation.
- ◆ In an ideal world, the variables should be relatively small in number, with low correlations among each other, have good (interval) measurement properties, have approximately normal distribution (kurtosis and skewness values close to zero).
- ◆ In the real world, the conditions mentioned above almost never exist.
- ◆ Fortunately, we have tools to handle some of these issues



# Week 4 Topic:

## Clustering Guidelines - 2

How should similarity between observations be operationalized? Choose among:

- ◆ distance metrics - when you have only numerical variables
- ◆ similarity coefficients - when you have only non-numerical variables only
- ◆ distance metrics - when you have both numerical and non-numerical variables).

# Week 4 Topic:

## Clustering Guidelines – 3

How to form clusters?

- ◆ Choose among different linkages or variance
- ◆ Typically, variables with higher variances tend to have more impact in determining the cluster solution than variables with lower variances. In some cases, data may need to be standardized to measure linkages. Range standardization is preferred because it tends to preserve the differences among groups better than standardizing to 0 means and unit variance (Milligan and Cooper, 1988). We will revisit this later using our case study data set.

# Week 4 Topic:

## Clustering Guidelines – 4

How many clusters?

- ◆ Practical (managerial) considerations in segmentation studies often dictate a small number of clusters (somewhere in the range of 2-10).
- ◆ Use relative sizes of each cluster in making this decision (in most applications, the preference is for somewhat balanced sizes or number of observations in each cluster).
- ◆ Use the relative change in distances at which clusters are combined (or, relative changes in the overall heterogeneity measure) to help with this decision. There are some statistics (such as pseudo- $F$ , pseudo  $t^2$  and CCC) that can be used judiciously to guide your decisions as well.

# Week 4 Topic:

## Cluster Analysis – Methods of execution

- ◆ **Unsupervised learning** -- learn from raw data (no examples of correct classification). In other words, class label (e.g., income bands, purchase power, etc.) information is unavailable. Unsupervised methods set the model's parameters without prior knowledge about the classification of samples.
- ◆ **Supervised learning** -- algorithms use class variables to generate its solution. An example is market segmentation on the next slide.
- ◆ Plotting the data can help to see if there is any evidence of cluster structure at all. It can also give you an idea of how many clusters there are, as well as helping you to identify potentially problematic non-spherical (irregular) clusters.
- ◆ It might be necessary to preprocess the data to optimize them for clustering. Common preparation steps include creating a distance matrix, standardizing the variables, or other transformations.

# Week 4 Topic:

## What is Market Segmentation?

“Market segmentation is grouping people (with the willingness, purchasing power, and the authority to buy) according to their similarity in several dimensions related to a product under consideration.”

Market Segmentation is supervised learning -- learn from data where the correct classification of examples is given (class label information is available) e.g., Naïve Bayes Classifier.

These can be results of a query or simple segmentations using dimensions:

- ◆ Demographics: Age, Gender, Education, Income, Home ownership, etc.
- ◆ Psychographics: Lifestyle, Attitude, Beliefs, Personality, Buying motives, etc.
- ◆ Brand Loyalty
- ◆ Geography: State, ZIP, City size, Rural vs. Urban, etc.

# Week 4 Topic:

## Applications – Retail Example

Application	Business Decision Support
Profiling and Segmentation	Understand customer behaviors and needs by segment. Direct efforts to like customer groups.
Cross-sell and Up-sell	Determine what customers are likely to buy. Better target/recommend product/ service offerings.
Acquisition and Retention	Understand customer preferences and purchase patterns. Determine how to grow and maintain valuable customers.
Campaign Management	Execute better customer communications. Determine right offer to the right person at the right time. Determine which customers to invest in and how to best appeal to them.

# Week 4 Topic:

## Types of Clustering

- ◆ **Partitional (k-means/optimization) clustering**
  - A division of objects into non-overlapping subsets (clusters) such that each object is in exactly one cluster
  - SAS offers PROC FASTCLUS (k-means clustering)
- ◆ **Hierarchical clustering**
  - A set of nested clusters organized as a hierarchical tree. Hierarchical clustering creates clusters that are hierarchically nested within clusters at earlier iterations, similar to the identification of species taxonomy in biology.

# Week 4 Topic:

## k-means clustering

- ◆ *K-means clustering* is, perhaps, the most popular partitive clustering algorithm. One reason for its popularity is that the time required to reach convergence on a solution is proportional to the number of observations being clustered, which means it can be used to cluster larger data sets.
- ◆ In fact, *k*-means clustering is inappropriate for small data sets (< 100 cases); the solution becomes sensitive to the order in which the observations appear. Changing the observation ordering, neither adding nor deleting observations, produces vastly different cluster solutions. This is known as the order effect.
- ◆ In SAS, PROC FASTCLUS implements the *k*-means algorithm. As its name suggests, the PROC FASTCLUS finds clusters in only a few (default=1) passes through the data. It also produces a description of the typical member of each cluster, which is useful both as a summary of its members, and as the basis for scoring new cases.



# Week 4 Topic:

## Limitations of K-means

- ◆ Need to specify K (number of clusters) in advance
- ◆ Applicable only for numeric data
- ◆ Has problems when clusters are of differing sizes or densities
- ◆ Unable to handle noisy data and outliers
- ◆ May be indeterminate

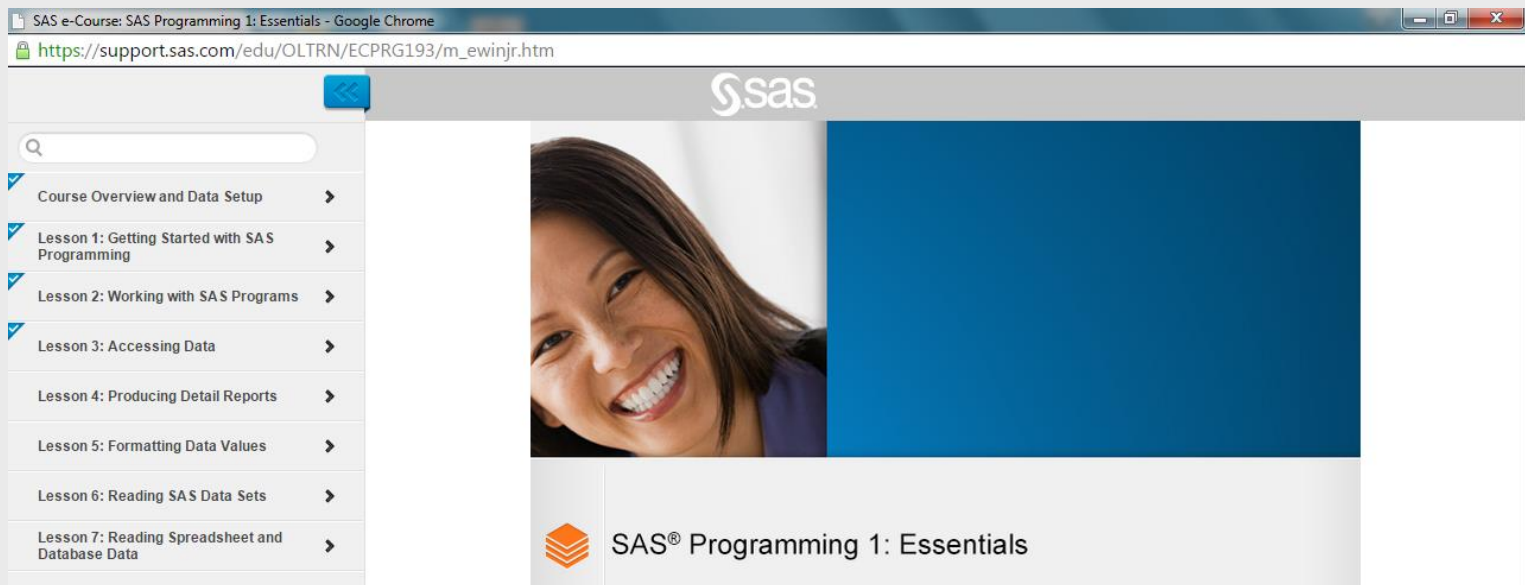
# Week 4 Topic:

## Assignment 3

- 1) Read Chapter 2 of Data Smart – we will review the wine example on Thursday.
- 2) Pot Hole Assignment Part 1:
  - a) Collect datasets (no need to submit raw data)
  - b) Create master workbook with merged, 2011 ~ 2014 data, into one worksheet. Then, add transportation metrics, 311, and weather information, as necessary, into the same workbook.
  - c) Document data validations, manipulations, derivations, assumptions, & observations made while collecting and integrating data.
  - d) Document metadata.

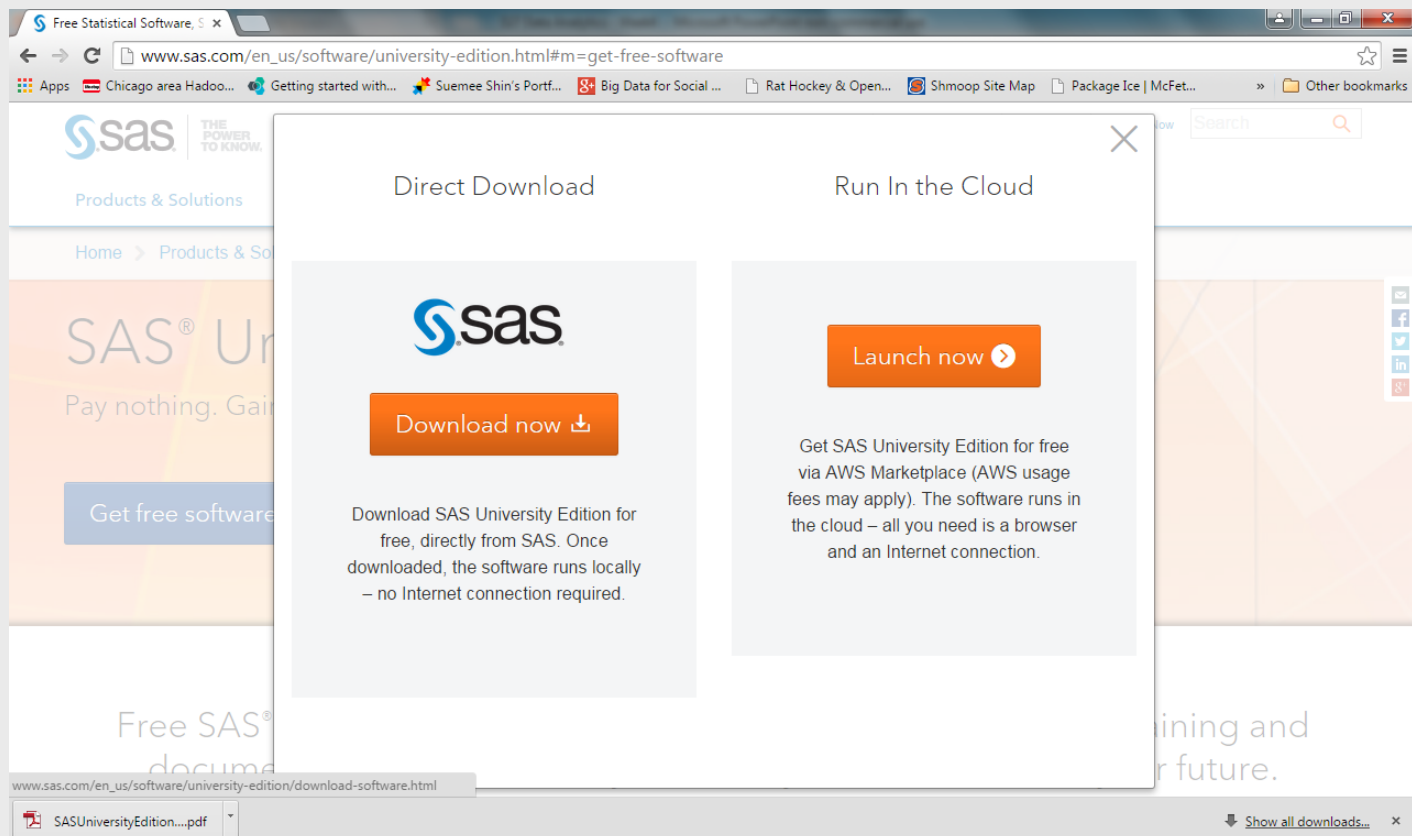
# Week 4 Topic: Reading Assignments

#	Reading
1	Read Chapter 2 in Data Smart
2	Take Lessons 1 ~ 3 in SAS Programming Essentials



# Week 4 Topic: Installing SAS University Edition

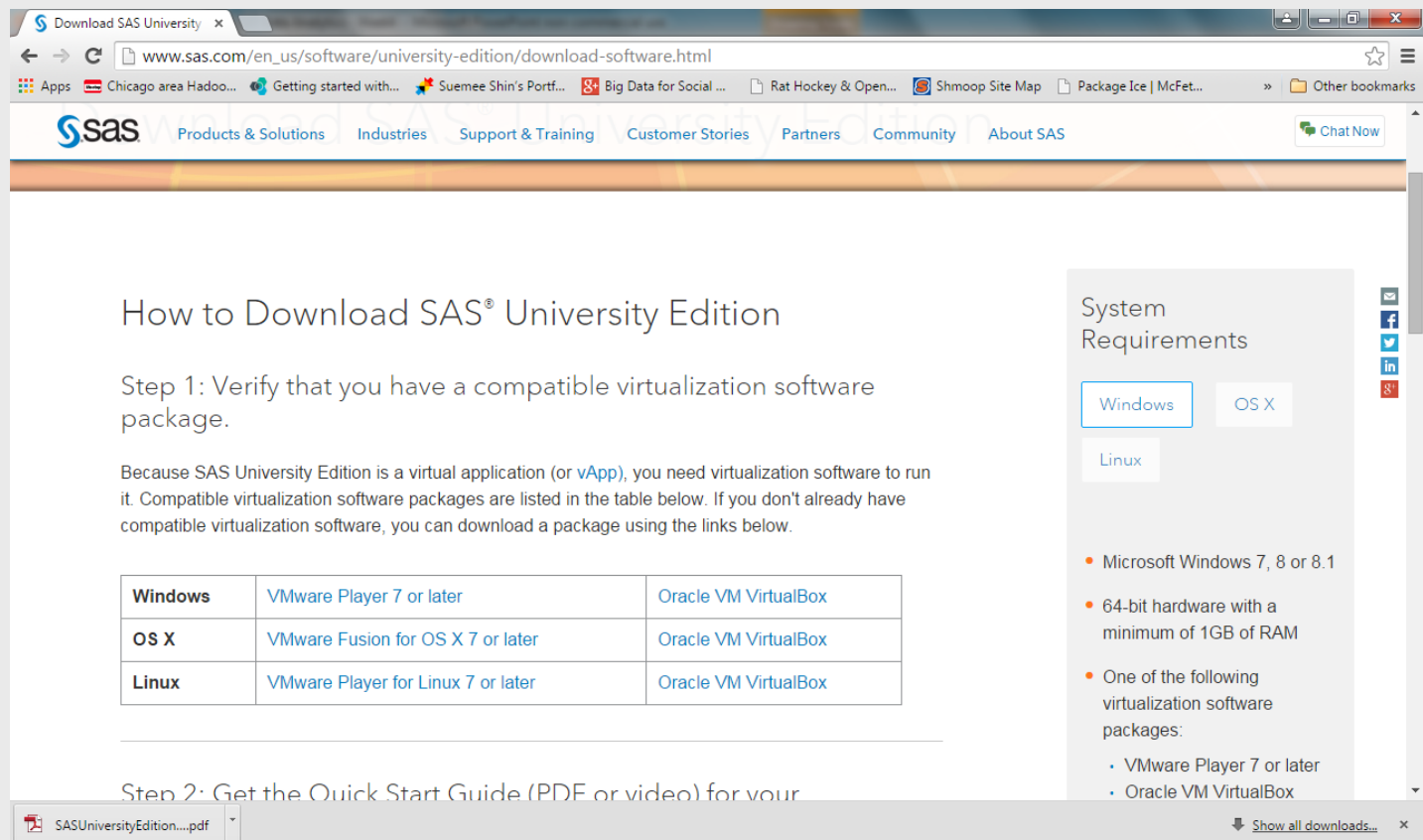
- ◆ Go to: [http://www.sas.com/en\\_us/software/university-edition.html](http://www.sas.com/en_us/software/university-edition.html)
- ◆ Create a student account. You will get free software and e-learning for one year.



# Week 4 Topic:

## Getting the right download file

- ◆ Select your system specific download files. Recommend using VirtualBox for Mac users.



The screenshot shows the SAS University Edition download page. The browser address bar displays [www.sas.com/en\\_us/software/university-edition/download-software.html](http://www.sas.com/en_us/software/university-edition/download-software.html). The page title is "How to Download SAS® University Edition".

**Step 1: Verify that you have a compatible virtualization software package.**

Because SAS University Edition is a virtual application (or [vApp](#)), you need virtualization software to run it. Compatible virtualization software packages are listed in the table below. If you don't already have compatible virtualization software, you can download a package using the links below.

Windows	<a href="#">VMware Player 7 or later</a>	<a href="#">Oracle VM VirtualBox</a>
OS X	<a href="#">VMware Fusion for OS X 7 or later</a>	<a href="#">Oracle VM VirtualBox</a>
Linux	<a href="#">VMware Player for Linux 7 or later</a>	<a href="#">Oracle VM VirtualBox</a>

**Step 2: Get the Quick Start Guide (PDF or video) for your**

**System Requirements**

Windows OS X Linux

- Microsoft Windows 7, 8 or 8.1
- 64-bit hardware with a minimum of 1GB of RAM
- One of the following virtualization software packages:
  - VMware Player 7 or later
  - Oracle VM VirtualBox

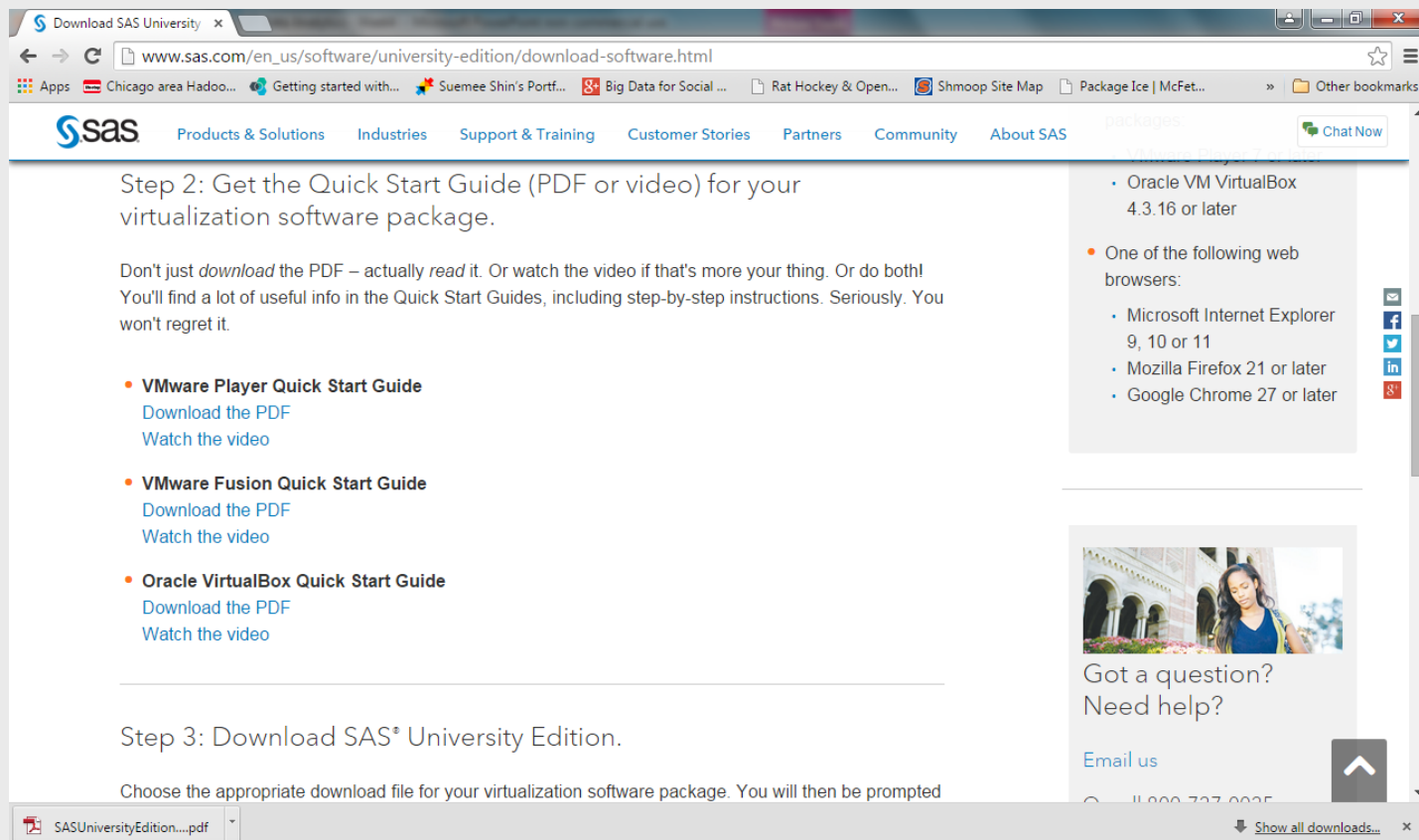
SASUniversityEdition....pdf

Show all downloads...

# Week 4 Topic:

## Follow installation directions

### ◆ Follow directions in the Quick Start Guide PDF



The screenshot shows a web browser window with the URL [www.sas.com/en\\_us/software/university-edition/download-software.html](http://www.sas.com/en_us/software/university-edition/download-software.html). The page is titled "Step 2: Get the Quick Start Guide (PDF or video) for your virtualization software package." It provides instructions on how to use the Quick Start Guides, including downloading PDFs or watching videos. The page lists three guides: VMware Player, VMware Fusion, and Oracle VirtualBox. Each guide has links to "Download the PDF" and "Watch the video". On the right side, there are sections for "packages" (listing Oracle VM VirtualBox 4.3.16 or later) and "One of the following web browsers:" (listing Microsoft Internet Explorer 9, 10 or 11; Mozilla Firefox 21 or later; and Google Chrome 27 or later). At the bottom, there is a section for "Step 3: Download SAS® University Edition." and a prompt to "Choose the appropriate download file for your virtualization software package. You will then be prompted". A download bar at the bottom shows a file named "SASUniversityEdition....pdf".

Download SAS University

www.sas.com/en\_us/software/university-edition/download-software.html

Products & Solutions Industries Support & Training Customer Stories Partners Community About SAS

Step 2: Get the Quick Start Guide (PDF or video) for your virtualization software package.

Don't just *download* the PDF – actually *read* it. Or watch the video if that's more your thing. Or do both! You'll find a lot of useful info in the Quick Start Guides, including step-by-step instructions. Seriously. You won't regret it.

- **VMware Player Quick Start Guide**  
[Download the PDF](#)  
[Watch the video](#)
- **VMware Fusion Quick Start Guide**  
[Download the PDF](#)  
[Watch the video](#)
- **Oracle VirtualBox Quick Start Guide**  
[Download the PDF](#)  
[Watch the video](#)

Step 3: Download SAS® University Edition.

Choose the appropriate download file for your virtualization software package. You will then be prompted

packages:

- Oracle VM VirtualBox 4.3.16 or later

One of the following web browsers:

- Microsoft Internet Explorer 9, 10 or 11
- Mozilla Firefox 21 or later
- Google Chrome 27 or later

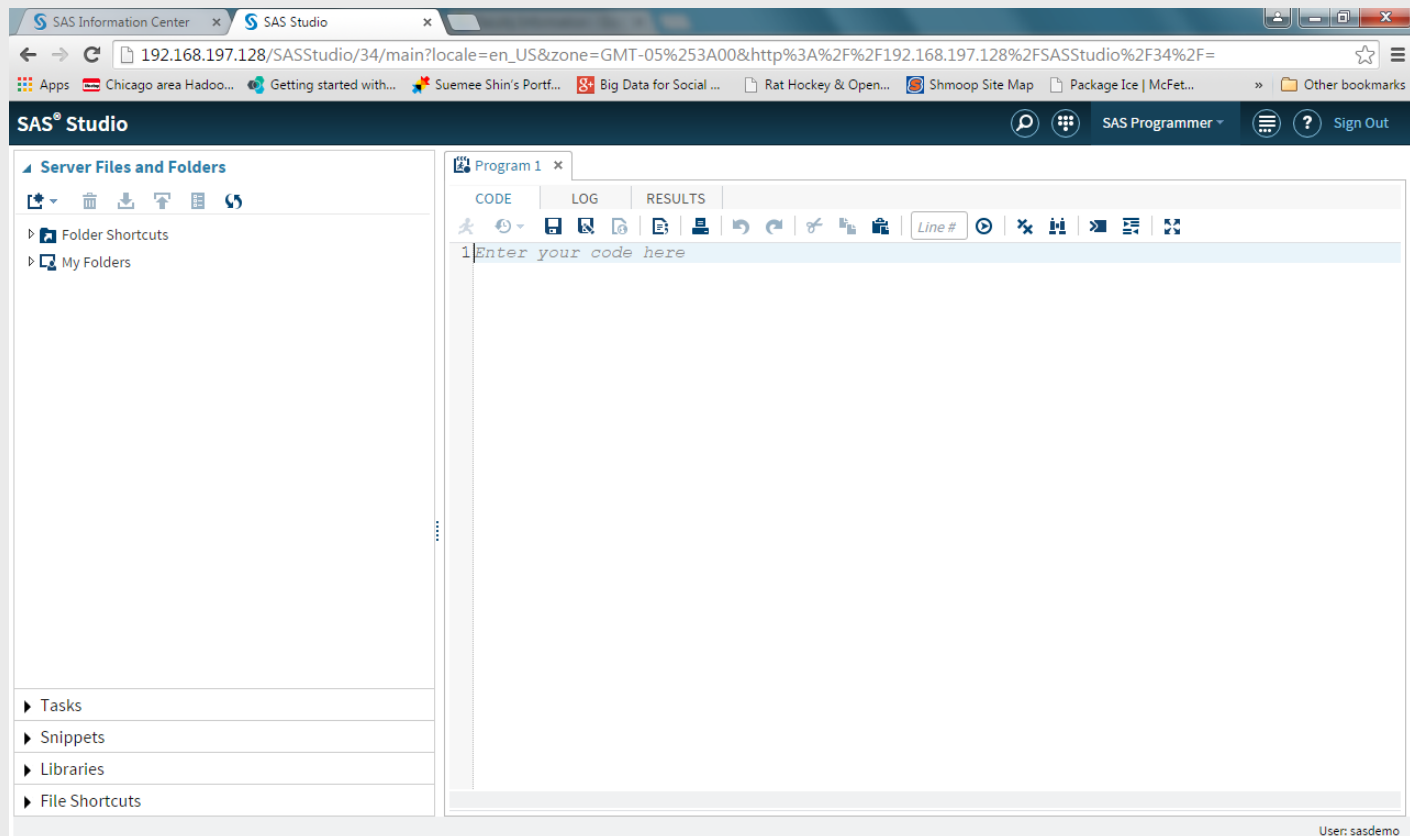
Got a question? Need help?

Email us

Show all downloads...

# Week 4 Topic: SAS Studio – start window

- ◆ Get to the point of when you can open a start window:



# Week 4 Topic:

## SAS Programming I Essentials

- ◆ Start on SAS Programming I Essentials e-learning course as time allows. Ideally, we'll want to finish this course before Week 8:

