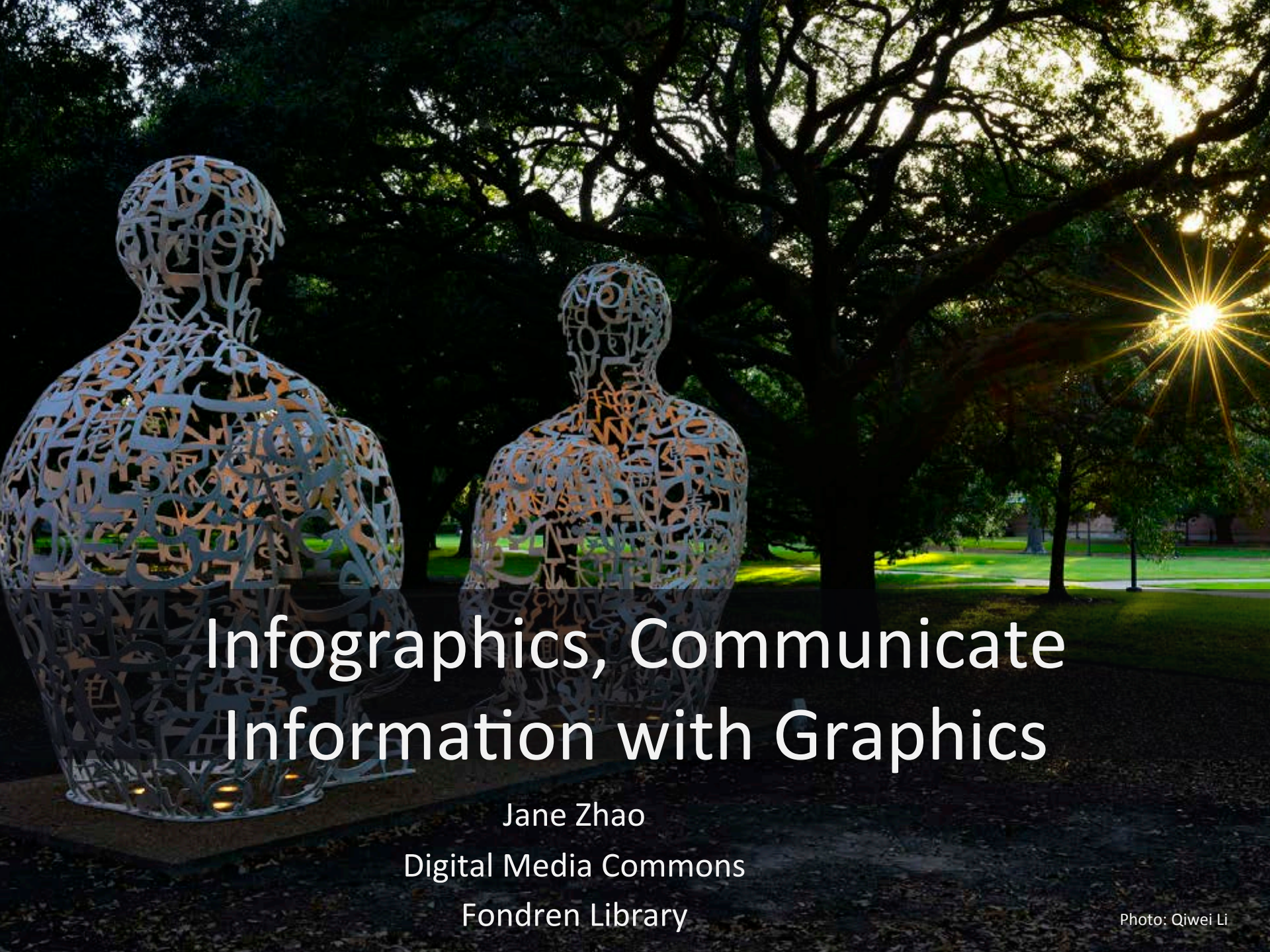


Infographics, Communicate Information with Graphics

Jane Zhao

Digital Media Commons

Fondren Library

Photo: Qiwei Li

9. Would you be interested in taking a library short course on:

#	Answer	Response	%
1	Zotero	162	21%
2	Mendeley	82	10%
3	EndNote	145	18%
4	GIS	223	28%
5	Visualizing data	270	34%
6	Creating infographics	292	37%
7	Digital storytelling	181	23%
8	Library research methods	232	29%
9	Navigating the library website	115	15%
10	Specific database(s) - (please specify)	16	2%
11	Other (please specify)	19	2%

9. Would you be interested in taking a library short course on:

#	Answer	Response	%
1	Zotero	178	24%
2	Mendeley	168	23%
3	EndNote	245	34%
4	GIS	166	23%
5	Visualizing data	302	41%
6	Creating infographics	183	25%
7	Digital storytelling	132	18%
8	Library research methods	166	23%
9	Navigating the library website	85	12%
10	Specific database(s) - (please specify)	11	2%
11	Other (please specify)	21	3%

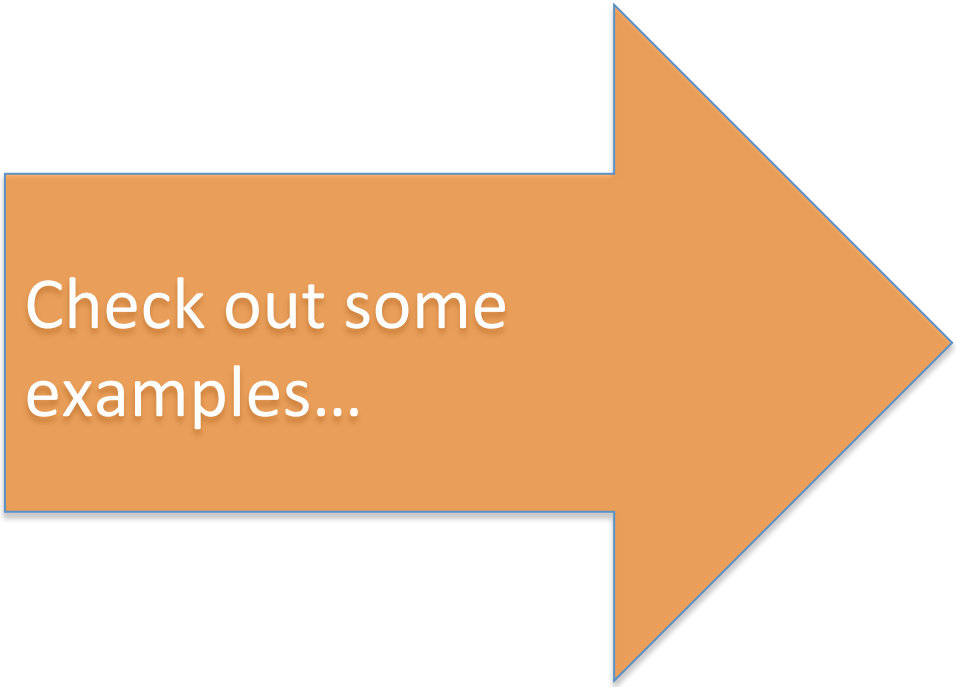
Objectives

- Learn the best practices of information design
- Be aware the handy tools for creating Infographics and Data Visualization

Outline

- What is Infographics? What is Data Visualization?
- Why Infographics work?
- What makes a good Infographic?
- Information design best practices.
- Tools for creating Infographics and Data Visualization.
- Data sources.

INFOGRAPHICS AND DATA VISULIZATION

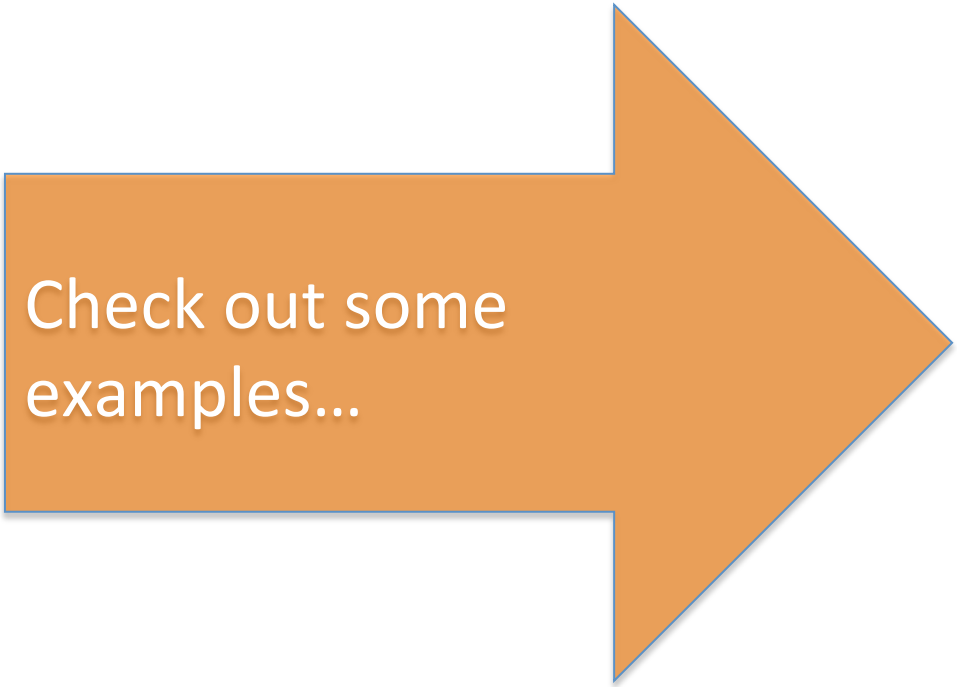


What is Infographics?

An Example of Infographics

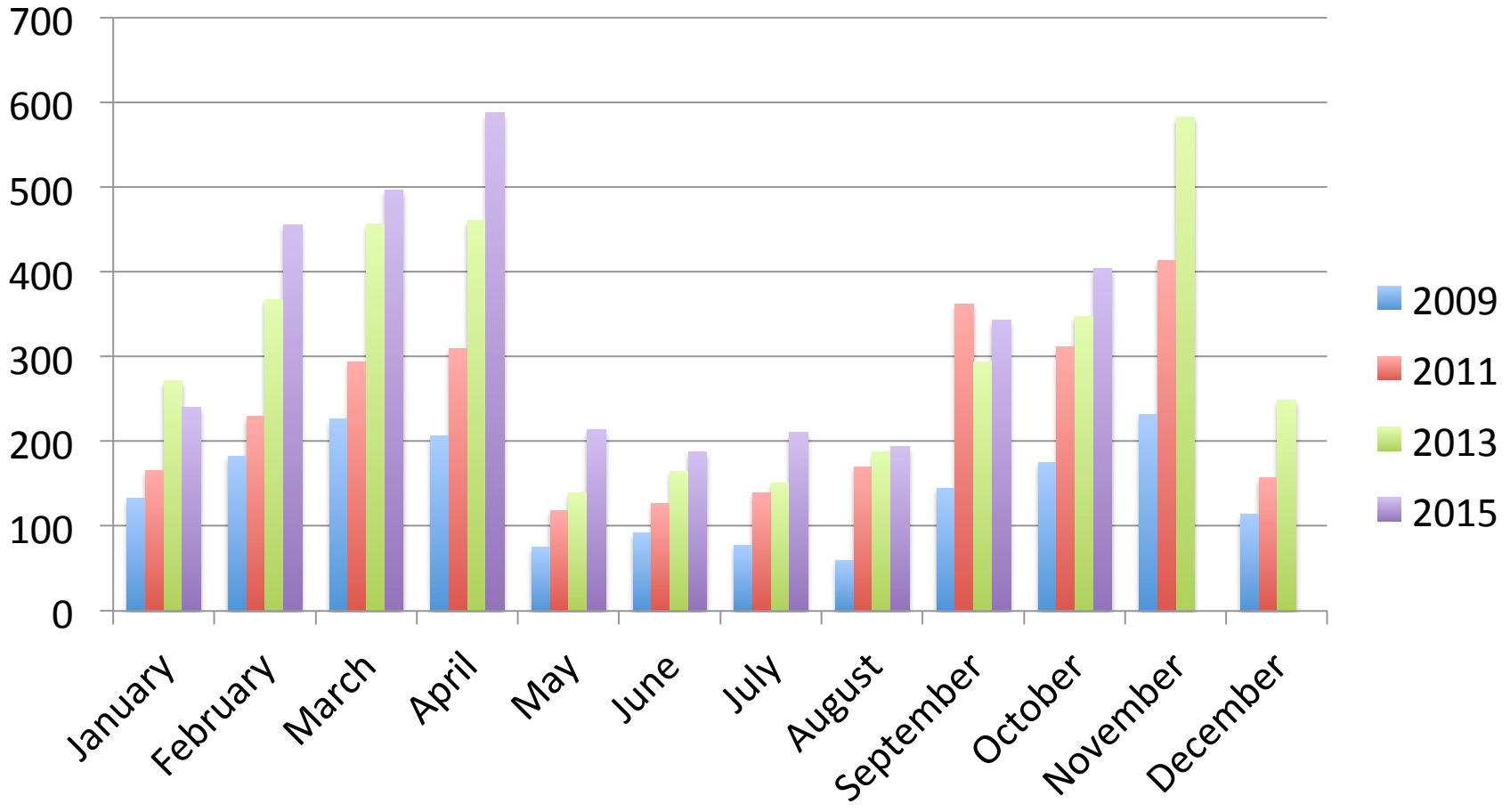
An Example of Infographics


1ST FLOOR
CONSTRUCTION
An Example of Infographics



What is Data Visualization?

DMC Equipment Circulation Statistics



An Example of Data Visualization

Data Visualization is a Separate Design Element Used in the Design of Infographics.

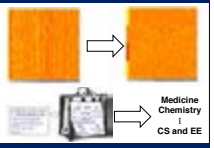
RICE

Bayesian Clustering and Variable Selection of High-Dimensional Count Data

Qiwei Li and Marina Vannucci
Department of Statistics, Rice University, Houston, Texas

INTRODUCTIONS

- An explosion of data for which the dimension p is considerably larger than the sample size n, i.e., $p \gg n$;
- A challenge to uncover the group structure of the observations and to determine the discriminating variables;
- A lot of well studies on continuous and Gaussian-distributed large-scale data, e.g., DNA microarray;
- A call for Bayesian method on non-negative count data, e.g., next-generation sequencing (RNA-Seq) and bag-of-word data;
- Difficulties: the number of groups, normalization, variability modeling.



STATISTICAL MODELS AND MCMC ALGORITHMS

Data and Parameter Specification

- Observable Data
  - X is a set of n p-dimensional observations from K populations;
  - Each element $x_{ij} \in \mathbb{Z}^+$ is an nonnegative count number:

$$X_{n\times p} = \begin{matrix} \\ X_{1\cdot} \\ \vdots \\ X_{n\cdot} \end{matrix} \begin{matrix} \begin{matrix} X_1 & X_2 & \cdots & X_p \end{matrix} \\ \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \end{matrix}$$

- Parameters for Clustering Observations
  - n samples are from a mixture of K Poisson distributions:

$$f(X_i|W,\Theta) = \sum_{k=1}^{K} w_k f(X_i|\theta_k);$$

  - The size of each component $(n_1 \ n_2 \ \cdots \ n_K)$ follows a multinomial distribution with parameter n and $(w_1 \ w_2 \ \cdots \ w_K)$ with conjugate prior $(\alpha \ \alpha \ \cdots \ \alpha)$;
  - A latent n-vector is introduced to identify the cluster, where $z_i = k$ indicates i-th observation belongs to k-th component:

$$Z = (z_1 \ z_2 \ \cdots \ z_n).$$

- Parameters for Identifying Discriminating Variables
  - Not all the variables provide information about group structure and some even obscure the recovery of the true structure;
  - A latent p-vector is introduced to identify the most discriminating variables, where $\gamma_j = 1$ indicates j-th variable is informative:

$$\Gamma = (\gamma_1 \ \gamma_2 \ \cdots \ \gamma_p);$$

  - Assume γ_js are independent Bernoulli random variables with parameter ω, that is, $|\Gamma| = \sum_{j=1}^{p} \gamma_j \sim Binomial(p,\omega)$.
- Parameters for Modeling Heterogeneity
  - The variation in the number of counts per sample is very high, e.g., different RNA samples may be sequenced to different depths.
  - A n-vector is introduced to model the unobserved heterogeneity with prior $Gamma(1/\sigma^2, 1/\sigma^2)$[1]:

$$S = (s_1 \ s_2 \ \cdots \ s_n).$$

Hierarchical Framework

- Data Likelihood
  - We assume

$$x_{ij} \sim \begin{cases} Poisson(s_i\theta_k) & if\ z_i = k, \gamma_j = 1 \\ Poisson(s_i\theta_0) & if\ \gamma_j = 0 \end{cases};$$

  - Data likelihood of each observation:

$$f(X_i|z_i = k, s_i, \Gamma, \theta_k, \theta_0) = \frac{s_i^{h_{i0}+h_{ik}}}{\prod_{j=1}^{p} x_{ij}!} e^{-s_i d_\theta (p-|\Gamma|)} \theta_0^{\ h_{i0}} e^{-s_i d_k |\Gamma|} \theta_k^{\ h_{ik}};$$

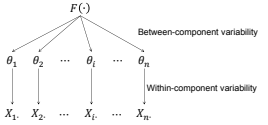
  - Data likelihood of each variable:

$$f(X_{\cdot j}|Z,S,\gamma_j,\Theta,\theta_0) = \begin{cases} \frac{\prod_{i=1}^{n} s_i^{v_{ij}}}{\prod_{i=1}^{n} x_{ij}!} e^{-|S|\theta_0} \theta_0^{v_j} & if\ \gamma_j = 1 \\ \frac{\prod_{i=1}^{n} s_i^{v_{ij}}}{\prod_{i=1}^{n} x_{ij}!} \prod_{k=1}^{K} e^{-|s_k|\theta_k} \theta_k^{v_{jk}} & if\ \gamma_j = 0 \end{cases};$$

  - $h_{i0} = \sum_{j:\gamma_j=0} x_{ij}$, $h_{i1} = \sum_{j:\gamma_j=1} x_{ij}$, $v_j = \sum_{i=1}^{n} x_{ij}$, $v_{jk} = \sum_{i:z_i=k} x_{ij}$;
  - Full data likelihood:

$$f(X|Z,S,\Gamma,\Theta,\theta_0) = \prod_{i=1}^{n} f(X_i|z_i = k, s_i, \Gamma, \theta_k, \theta_0).$$

- Dirichlet Process Model



  - Motivation:
    - The number of component K is unknown;
    - The prior distribution of θ_k is unknown;
  - Each θ_k shares a common but completely unknown $F(\cdot)$;
  - The prior of $F(\cdot)$ is $DP(F_0,\alpha)$, where α is a weighting factor that characterizes how close $F(\cdot)$ is to the shape of F_0:

$$F_0 = Gamma(a,b);$$

  - Integrating over $F(\cdot)$, we obtain

$$\theta_k|\Theta_{-k} \sim \frac{1}{K-1+\alpha} \sum_{m=1,m\neq k}^{K} \delta(d_m) + \frac{\alpha}{K-1+\alpha} F_0.$$

  - If Θ only takes on K distinctive values, then we have a mixture of the smooth measure F_0 and the K point masses;
  - Any observations θ_k and θ_m that have the same value are defined as being in the same cluster.
- Posterior and Full Conditionals
  - Posterior:

$$\pi(Z,S,\Gamma,\Theta,\theta_0|X) \propto f(X|Z,S,\Gamma,\Theta,\theta_0)\pi(Z)\pi(S)\pi(\Gamma)\pi(\Theta)\pi(\theta_0);$$

  - Full conditionals

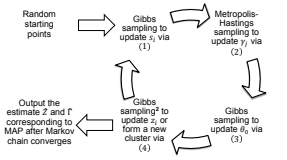
$$s_i|z_i,\Gamma,\theta_k,\theta_0,X_{i\cdot} \sim Gamma\left(\frac{1}{\sigma^2} + h_{i\cdot}, \frac{1}{\sigma^2} + \theta_0(p-|\Gamma|) + \theta_k|\Gamma|\right); \quad (1)$$

$$\pi(\gamma_j|Z,S,\Theta,\theta_0,X_{\cdot j}) \propto f(X_{\cdot j}|Z,S,\gamma_j,\Theta,\theta_0)\pi(\gamma_j); \quad (2)$$

$$\theta_0|z_i,\Gamma,\theta_k,\theta_0,X_{i\cdot} \sim Gamma\left(a + \textstyle\sum_{i:z_i=k} h_{i1}, b + |\Gamma|\textstyle\sum_k s_k\right); \quad (3)$$

$$\pi(z_i = k|Z_{-i},s_i,\Gamma,\Theta,\theta_0,X_{i\cdot}) \propto f(X_{i\cdot}|z_i = k,s_i,\Gamma,\theta_k,\theta_0)\pi(\theta_k|\Theta_{-k}); \quad (4)$$

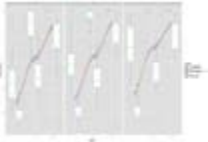
MCMC Algorithms



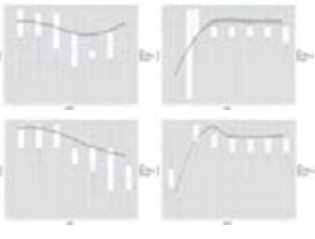
RESULTS AND DISCUSSION

Evaluation with Synthetic Data

- Synthetic Data Generating
  - 20 observations and 1000 variables, of which $|\gamma|$are discriminant: $x_{ij}|\gamma_j = 1 \sim I_{\{1\le i\le 5\}} Poisson(s_i\theta_1) + I_{\{6\le i\le 10\}} Poisson(s_i\theta_2) + I_{\{11\le i\le 15\}} Poisson(s_i\theta_3) + I_{\{16\le i\le 20\}} Poisson(s_i\theta_4)$;
  - $\theta_0 = 10$ and to model overdispersion, we set $\theta_k \sim Gamma(\psi,\psi/d_k)$, $d_1 = 80$, $d_2 = 40$, $d_3 = 60$, and $d_4 = 100$.
- Statistical Performance
  - $Precision = \frac{\#\ of\ true\ pairwise\ relationship\ that\ are\ correctly\ estimated}{\#\ of\ all\ pairwise\ in\ estimated\ Z}$;
  - $Recall = \frac{\#\ of\ true\ pairwise\ relationship\ that\ are\ correctly\ estimated}{\#\ of\ all\ pairwise\ in\ true\ Z}$;
  - $F\ score = \frac{2\times Precision\times Recall}{Precision+Recall}$.
  - Program settings: $a = 0.01$, $b = a/\bar{X}$, $c = 1/a^2$, $\alpha = 1$, $\omega = 0.01$.

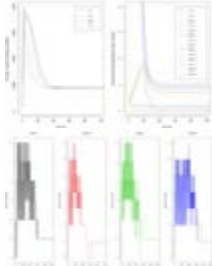


- Sensitivity Analysis
  - We set $|\gamma| = 40$ and $\psi = 10$;
  - Program settings: $a = 0.01$, $b = a/\bar{X}$, $c = 1/a^2$, $\alpha = 1$, $\omega = 0.01$.

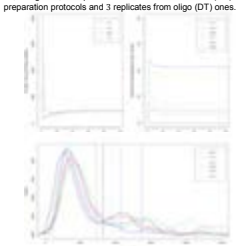


Experiment on Real Data

- Liver and Kidney RNA-Seq Data Set[3]
  - 22,925 genes, 7 replicates from a liver sample and 7 replicates from a kidney sample, each from a single human male.



- Yeast (*Saccharomyces cerevisiae*) RNA-Seq Data Set[4]
  - 6,874 genes, 3 replicates from random hexamer (RH) library preparation protocols and 3 replicates from oligo (DT) ones.



CONCLUSION

- Proposed a fully Bayesian method for simultaneously clustering high-dimensional data and selecting the variables that best discriminate the different groups on Poisson model;
- Formulated the clustering problem in terms of Poisson mixture model via Dirichlet process with unknown K;
- Evaluated the MCMC algorithms on both simulated and real data and provided recommendations for priors;
- To extent Poisson model to negative binomial model and to model variance shrinkage more elaborate.

Reference:

[1] J. Li, D. M. Witten, I. M. Johnstone, and R. Tibshirani, "Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data", *Biostatistics*, 2012, Volume 13, Issue 3, pp. 523-538;
[2] R. M. Neal, "Markov Chain Sampling Methods for Dirichlet Process Mixture Models", *Journal of Computational and Graphical Statistics*, 2000, Volume 9, Issue 2, pp. 249-265;
[3] J. Marioni, C. Mason, S. Mane, M. Stephens, and Y. Gilad, "RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays", *Genome Research*, 2008, Volume 18;
[4] J. Nagalakshmi, Z. Wong, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder, "The Transcriptional Landscape of the Yeast Genome defined by RNA Sequencing", *Science*, Volume 302.

WHY INFOGRAPHICS WORK?

50-80% of Human Brain is Dedicated to Visual Processing.

The Human Brain is a Pattern Recognition Machine!

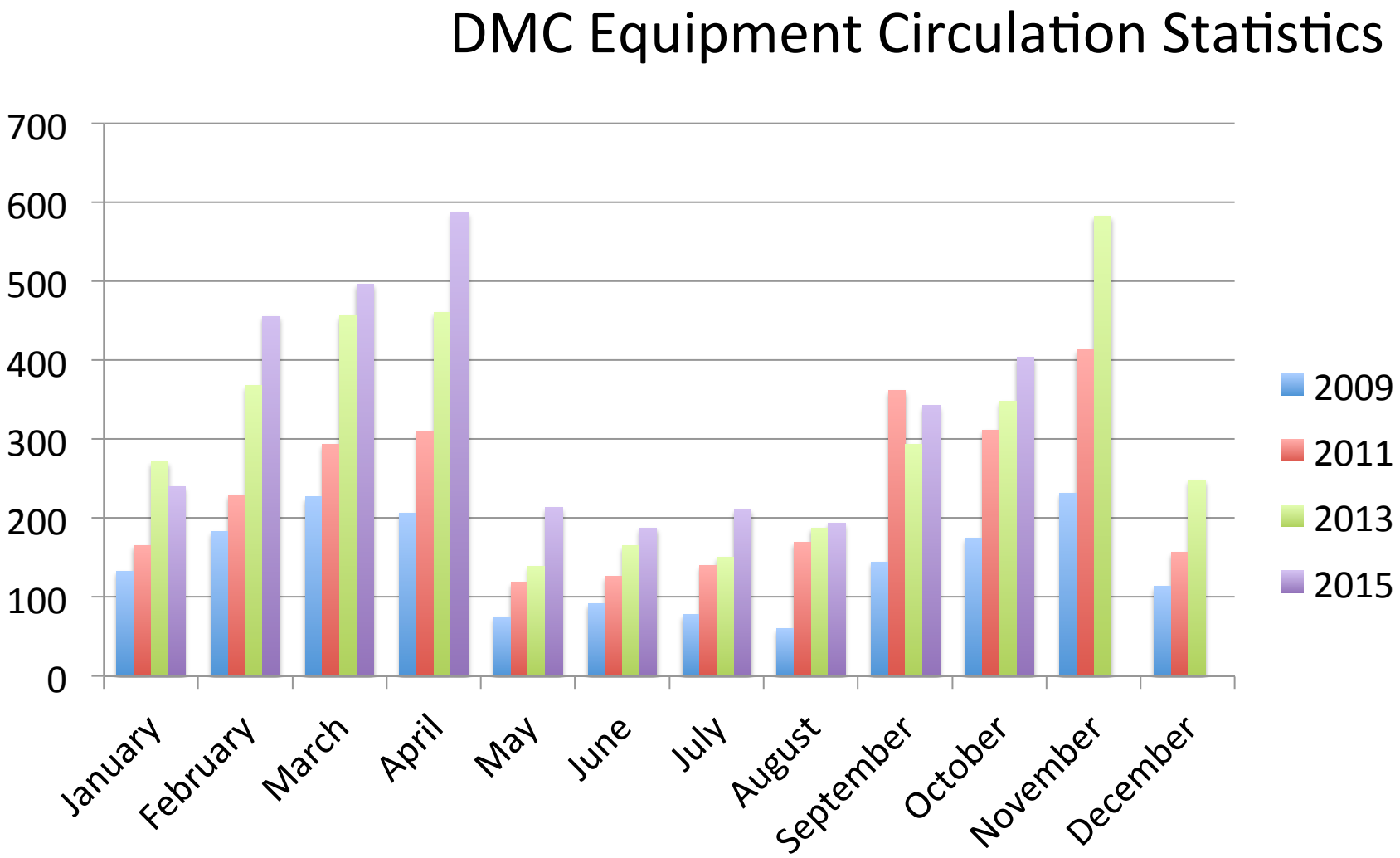
This Comes from the Evolution of a Survival Instinct!

A Table of Data, Hard to See its Pattern and Trend.

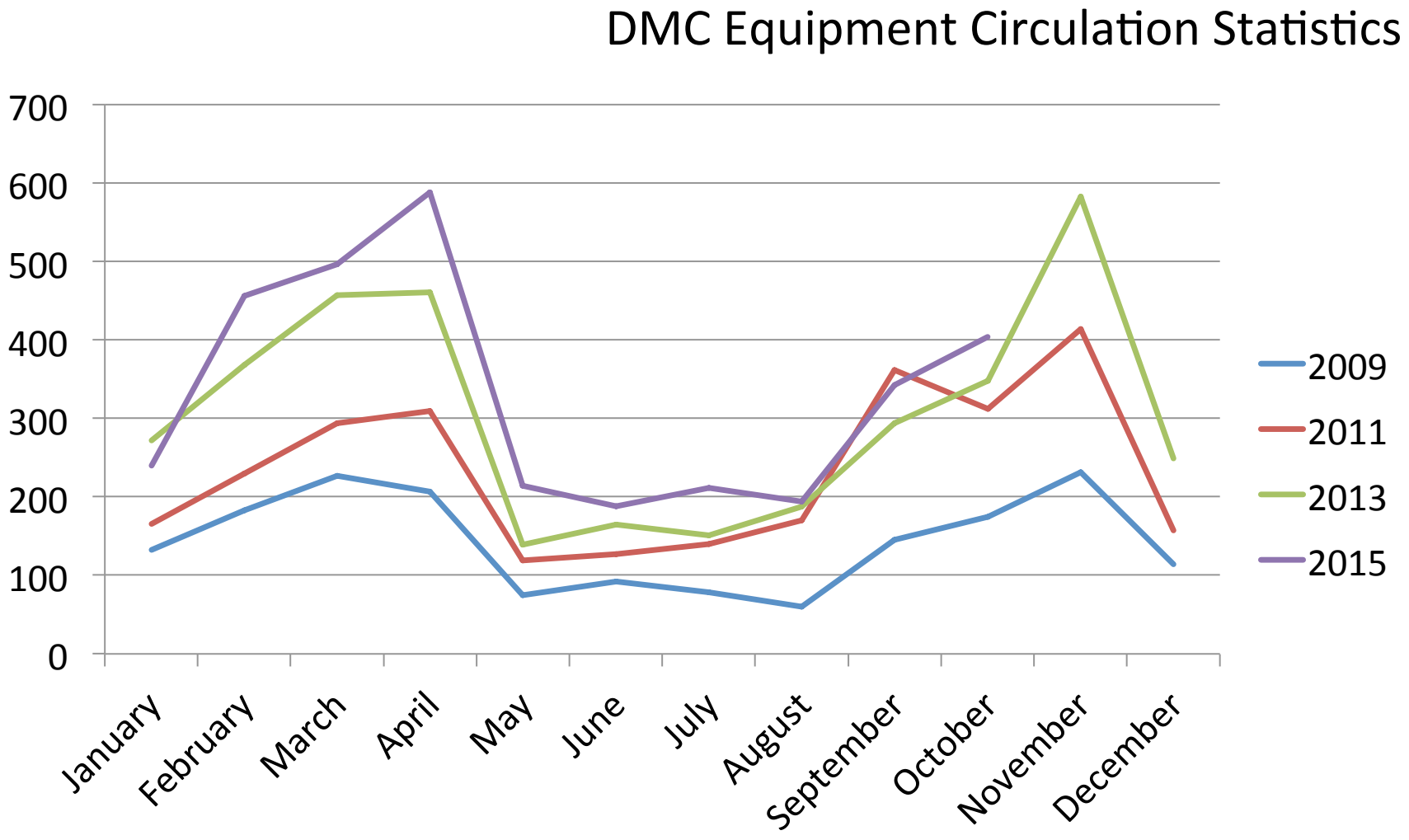
DMC Equipment Circulation Statistics

Month	2009	2011	2013	2015
January	133	166	272	240
February	183	230	368	456
March	227	294	457	497
April	207	310	461	588
May	75	119	139	214
June	92	127	165	188
July	78	140	151	211
August	60	170	188	194
September	145	362	294	343
October	175	312	348	404
November	232	414	583	
December	114	157	249	

Convert the Data to a Bar Chart, Easy to See the Pattern.



Convert the Data to a Line Chart, Easy to See the Trend.



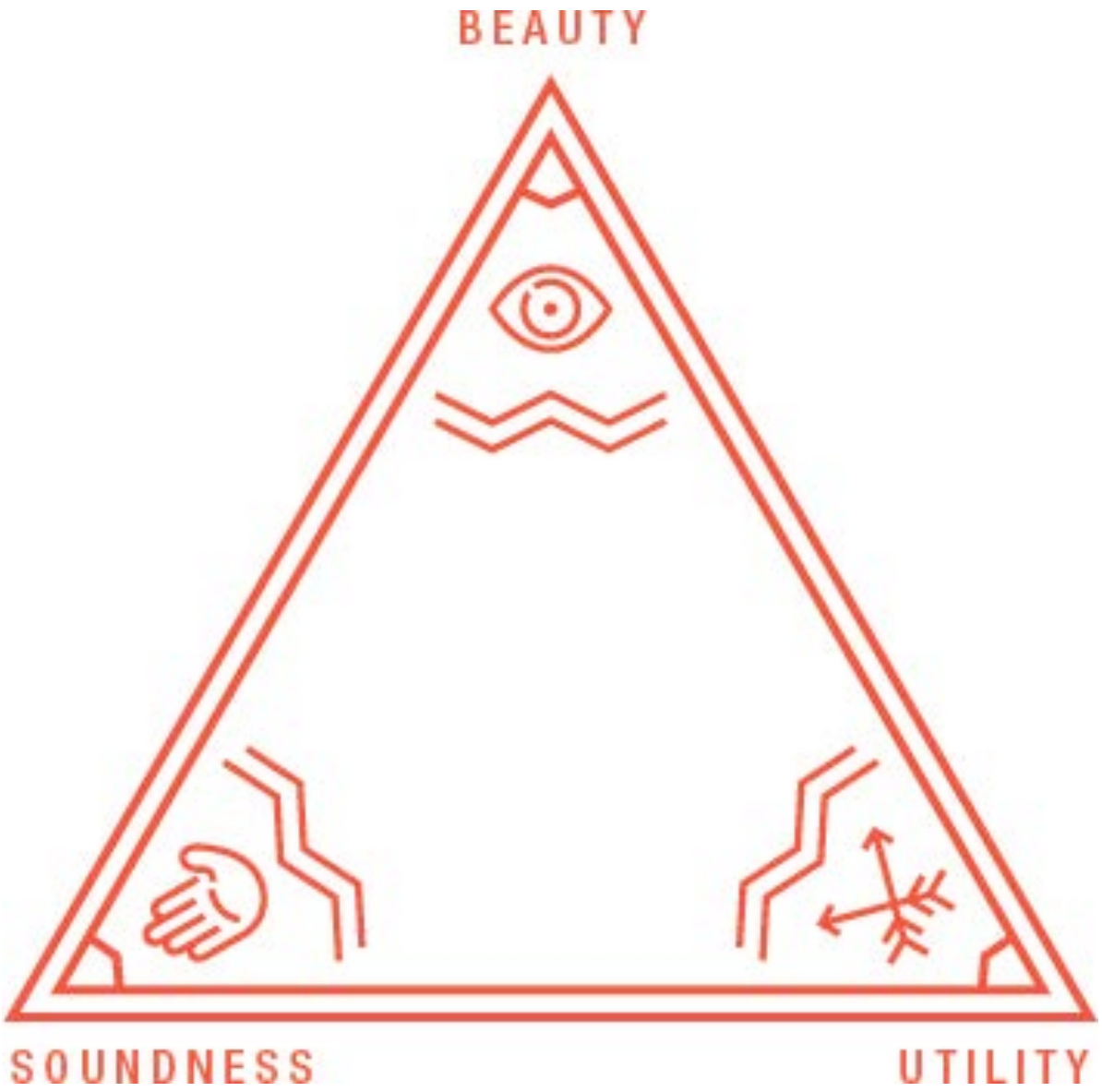
People are Likely to Remember 65% of Information if It is Presented as a Text Combined with a Relevant Image.

People are Likely to Remember 10% of Information if It is Presented as a Text or Audio Only.

"Of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful."

Eward Tufte, Yale Professor

WHAT MAKES A GOOD INFOGRAPHIC?



Lankow, J., Ritchie, J., & Crooks, R. (2012). *Infographics [electronic resource]: the power of visual storytelling*. ©2012. P198

INFORMATION DESIGN BEST PRACTICES

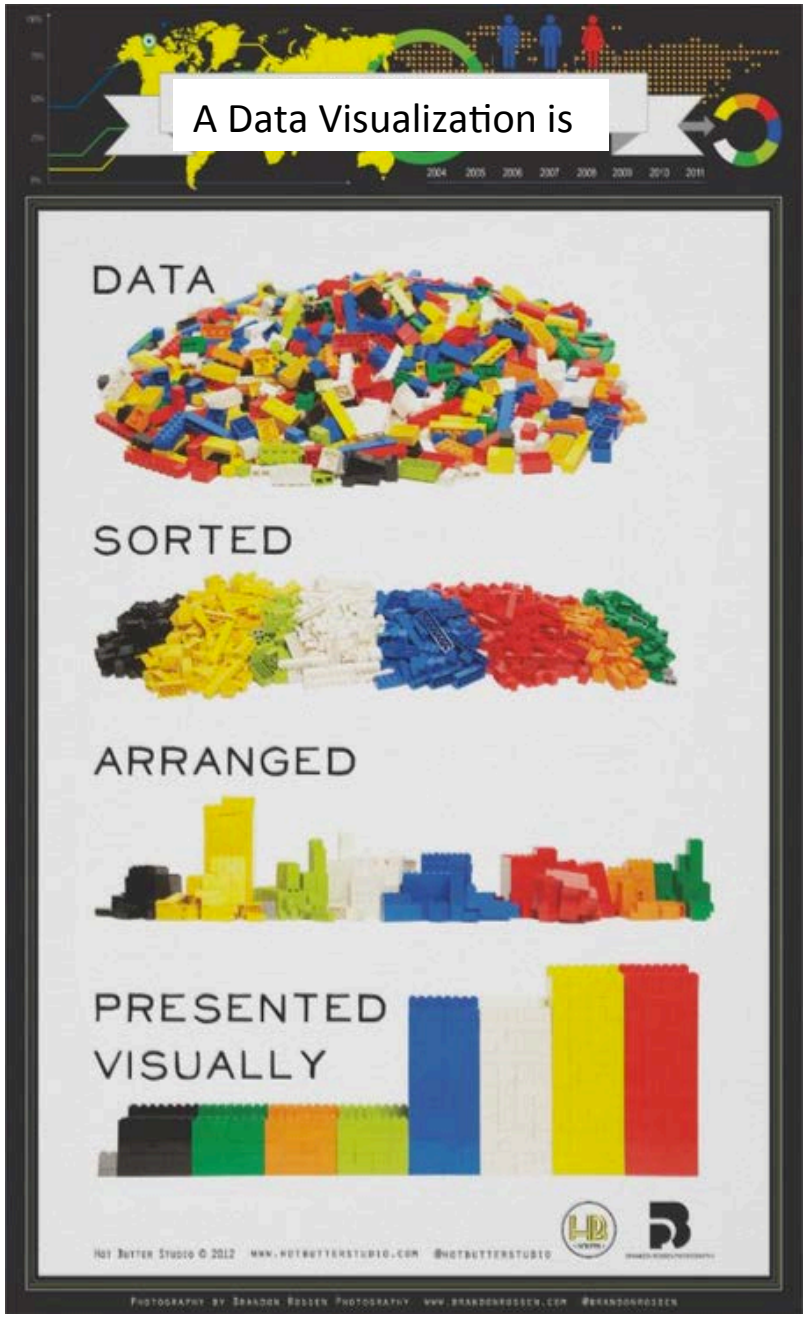
1

Keep it simple

http://www.slideshare.net/bethanyvsmith/pecha-kucha-presentation-zen?qid=c59c4f27-0731-4a00-98b5-396ec71bd13a&v=&b=&from_search=1

“Simplicity means the achievement of maximum effect with minimum means.”

- Dr. Koichi Kawana – artist, designer, and architect



Designed by HotButterStudio, http://www.zazzle.com/poster-228276710365015813

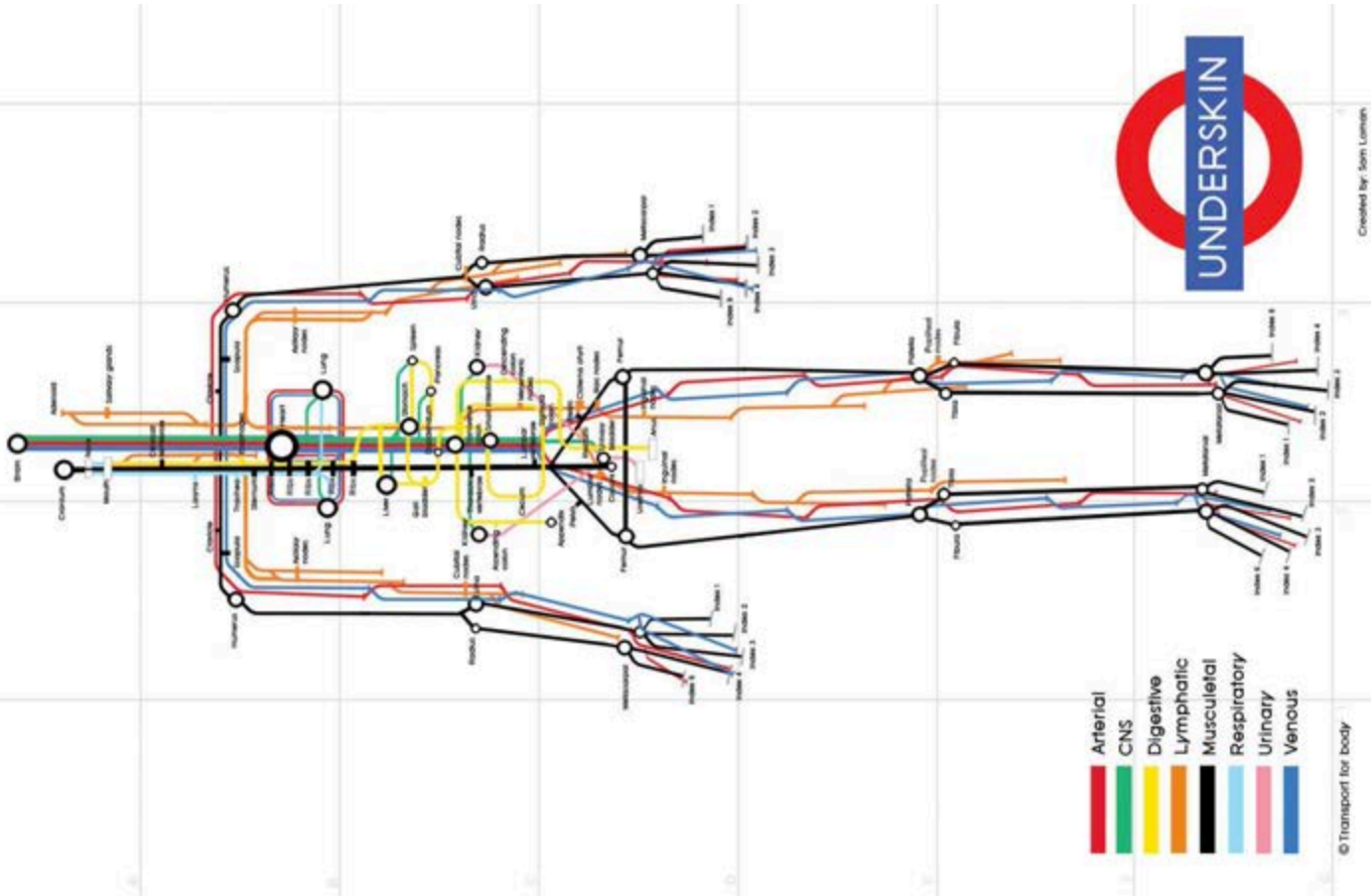
2

Use a simple text message combined with a relevant image.

http://www.slideshare.net/rtkrum/sfima-south-florida-interactive-marketing-association-cool-infographics-feb-2014?qid=c7d62517-2289-4642-9d64-285cf2579059&v=&b=&from_search=1

Make it unique!

http://www.coolinfographics.com/blog/2010/3/11/underskin-the-human-subway-map.html

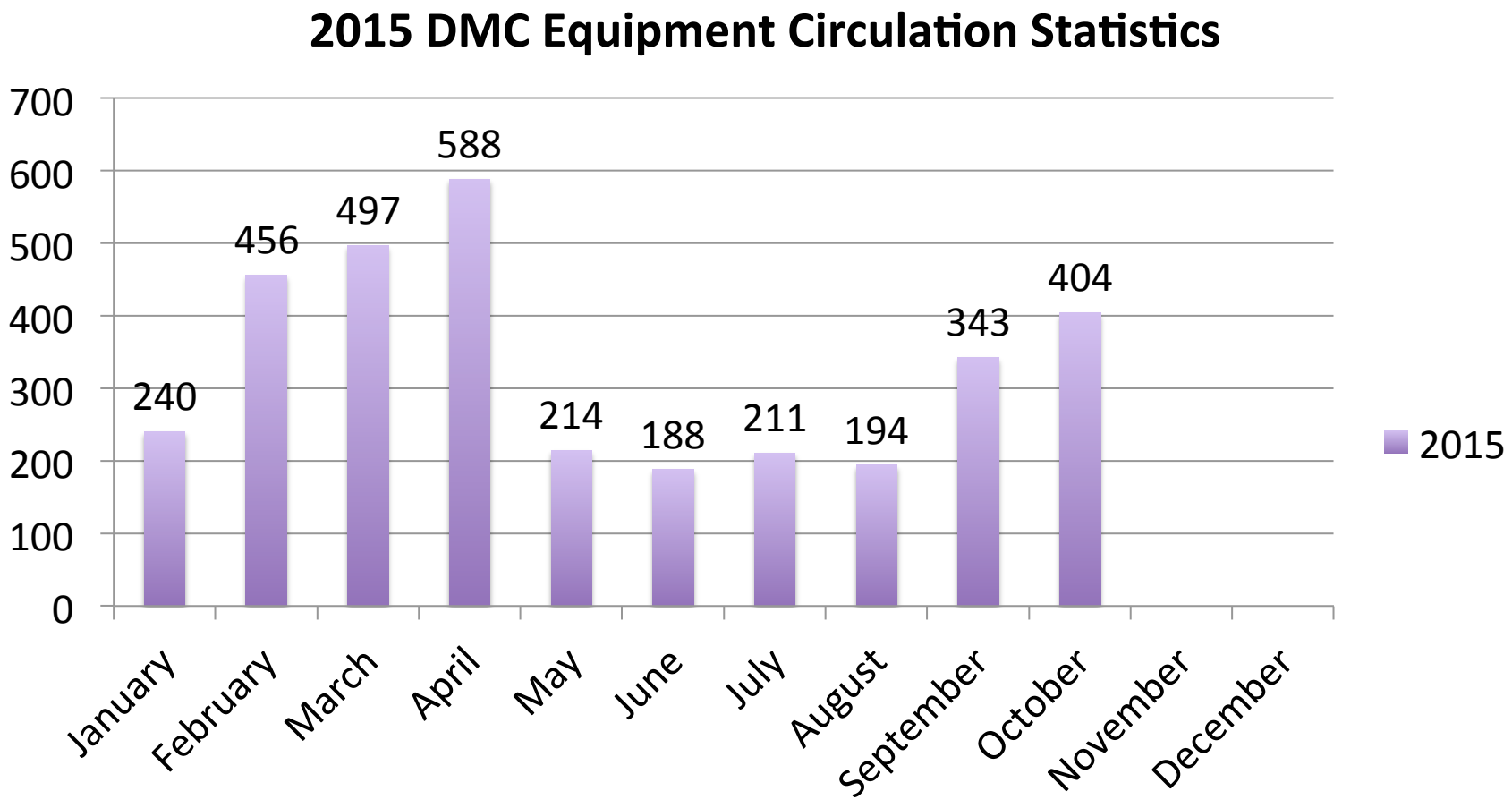


A good infographic leaves you feeling informed or delighted.

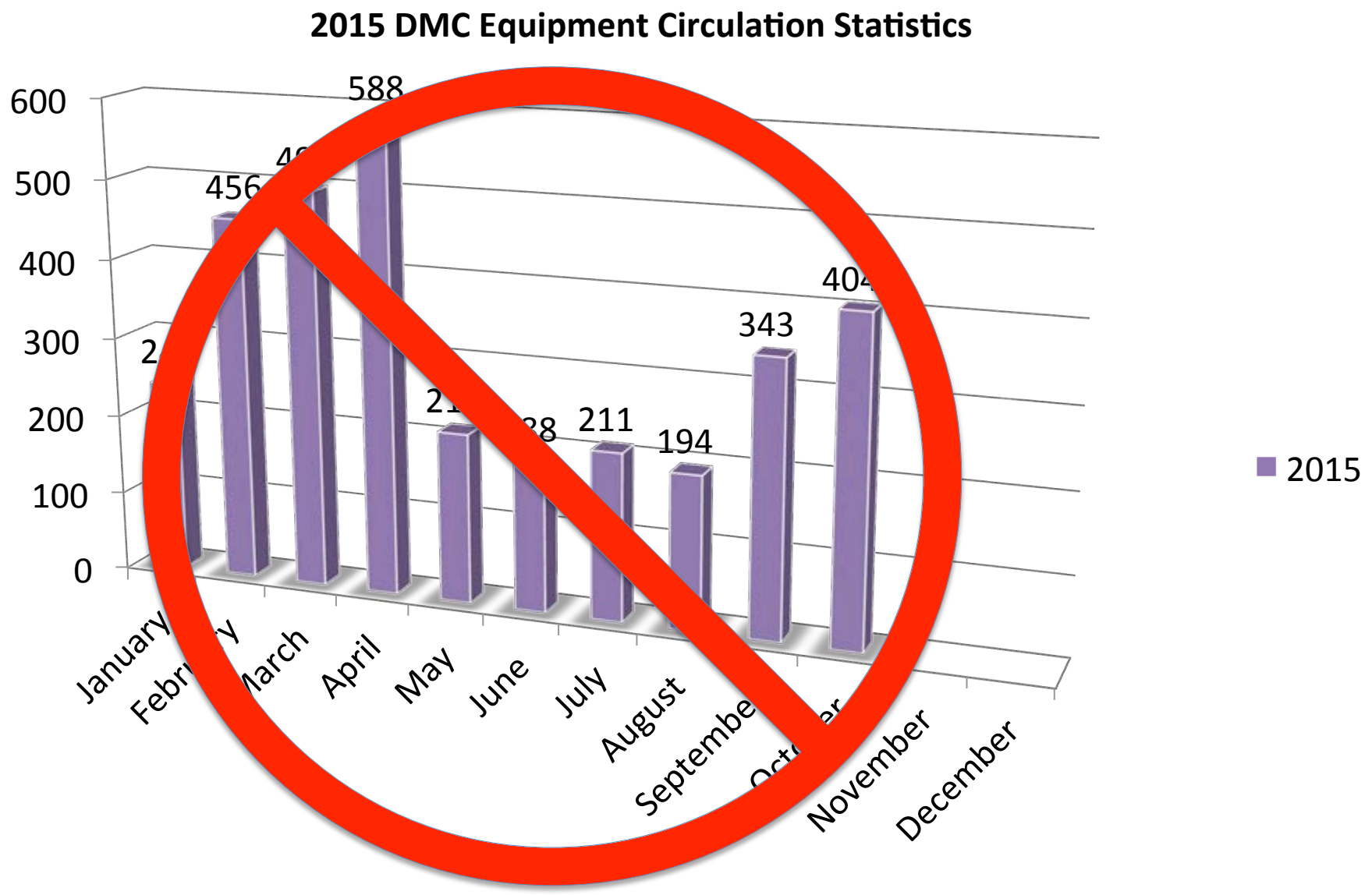
- Krum, Randy, Cool Infographics, P52

DATA VISUALIZATION BEST PRACTICES

Bar Chart for Ranking or Time Series



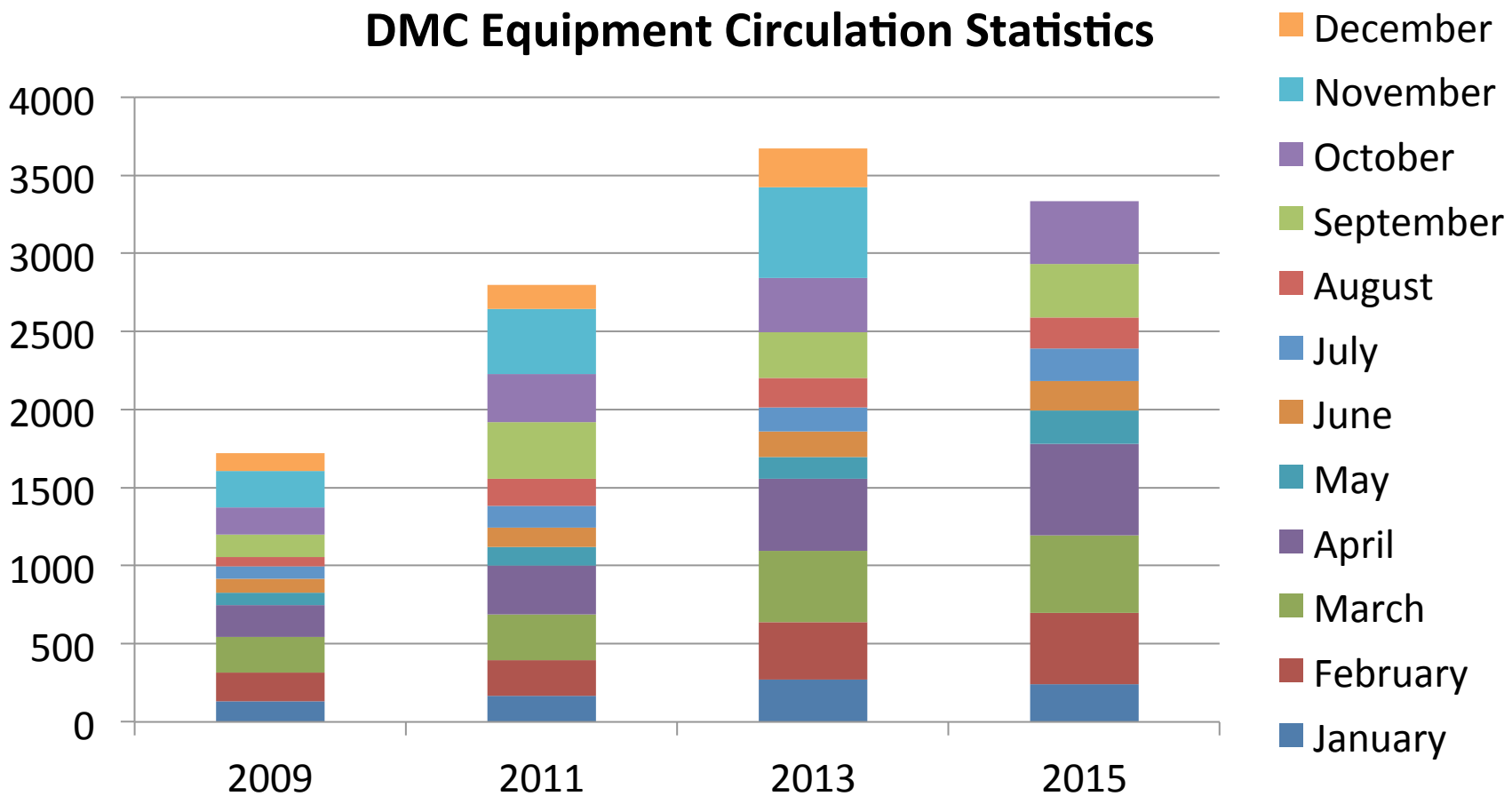
Avoid 3-D Bar Chart



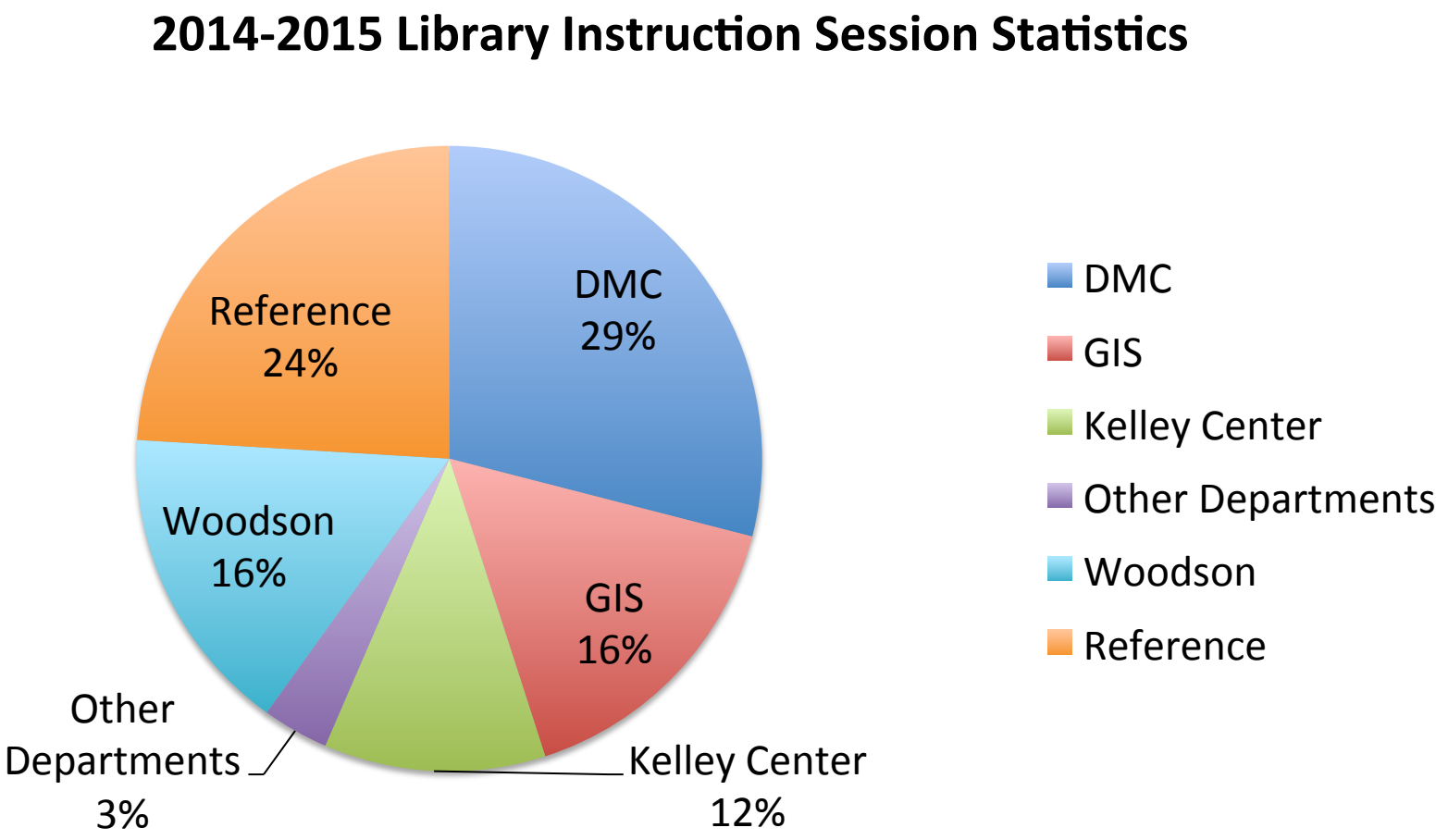
Avoid 3-D Bar Chart

Show Me the Numbers, P4

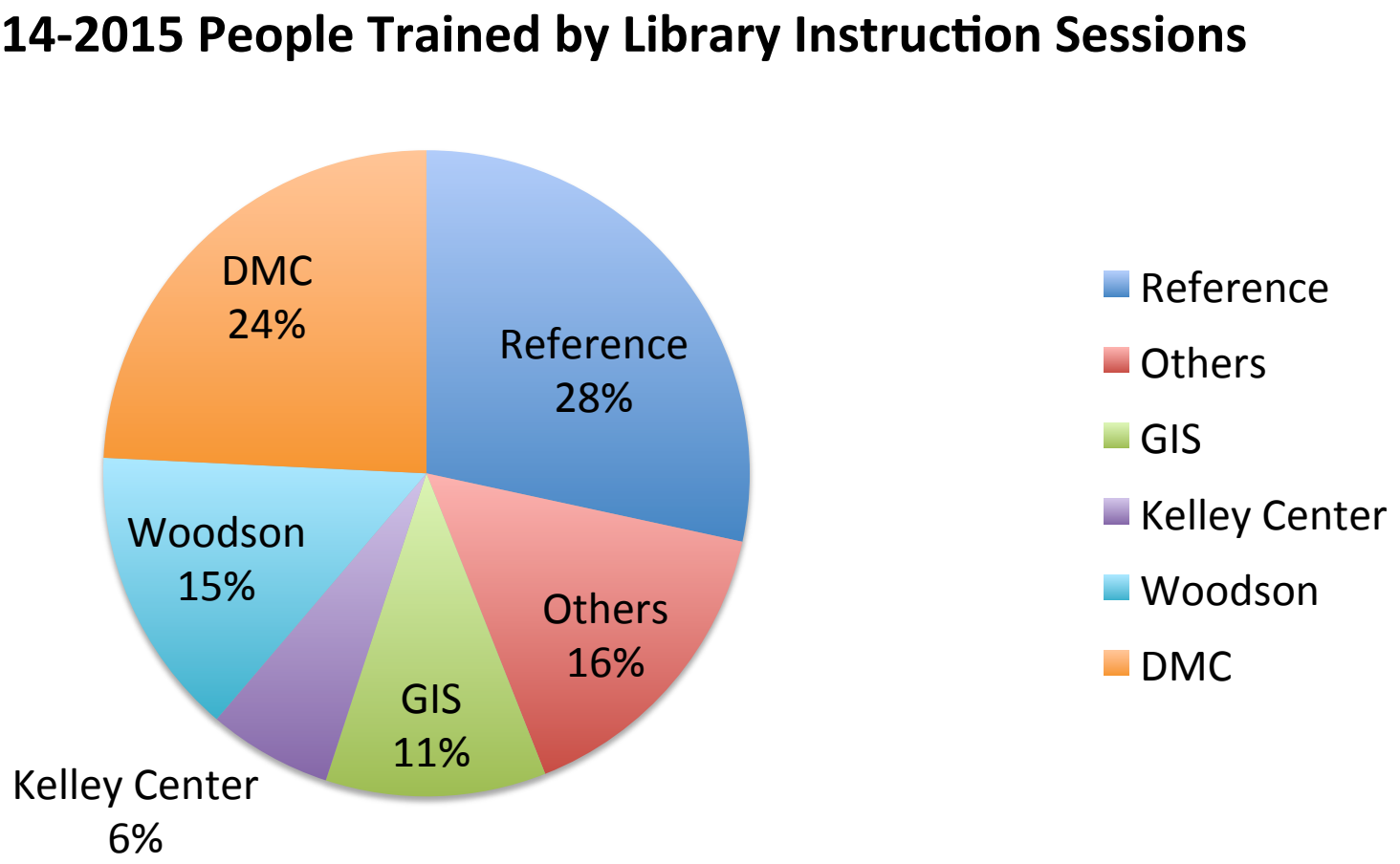
Stacked Bar Chart for Multiple Part-to-Whole Relationships



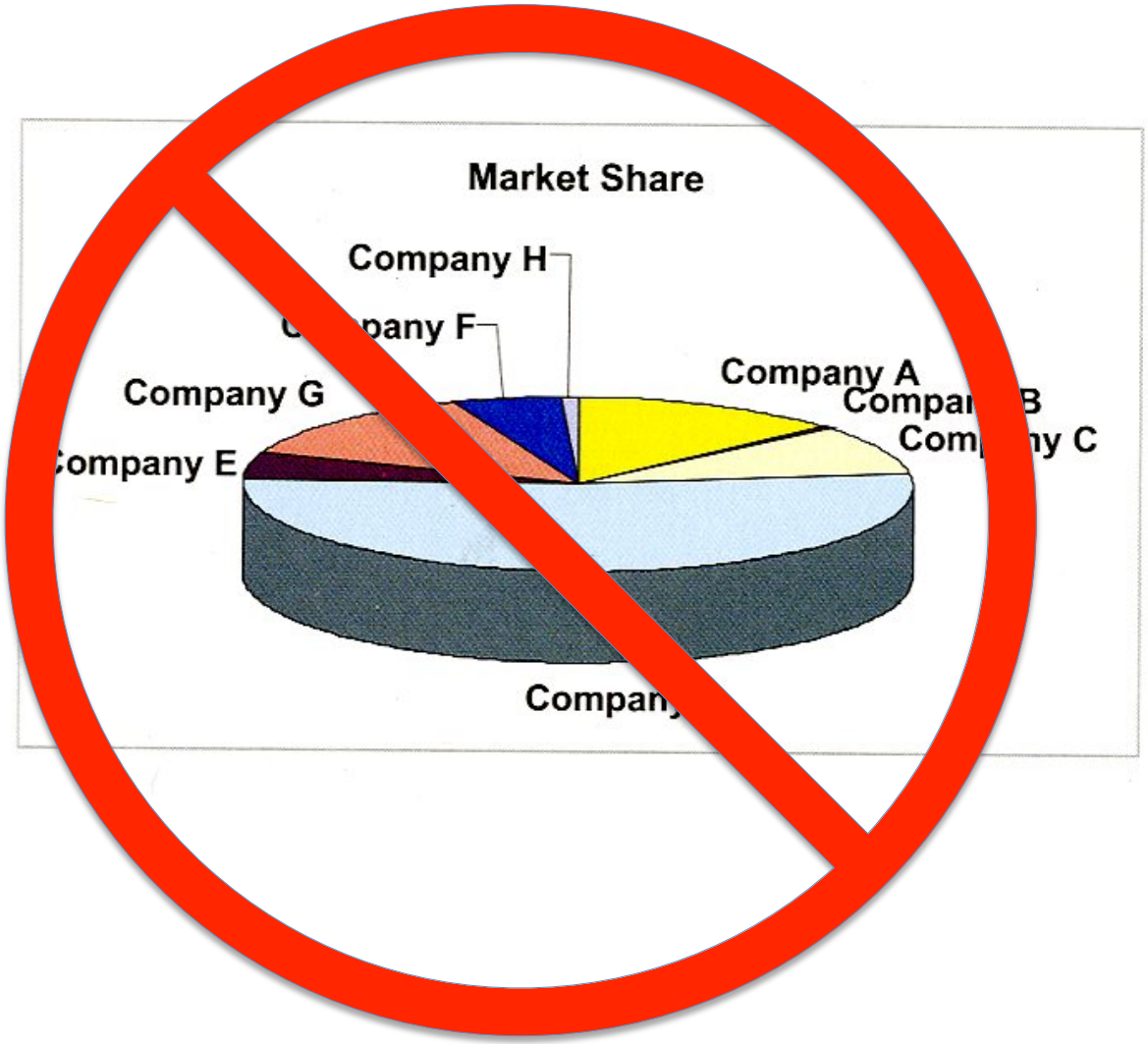
Pie Chart for Part-to-Whole Comparisons



Pie Chart for Part-to-Whole Comparisons

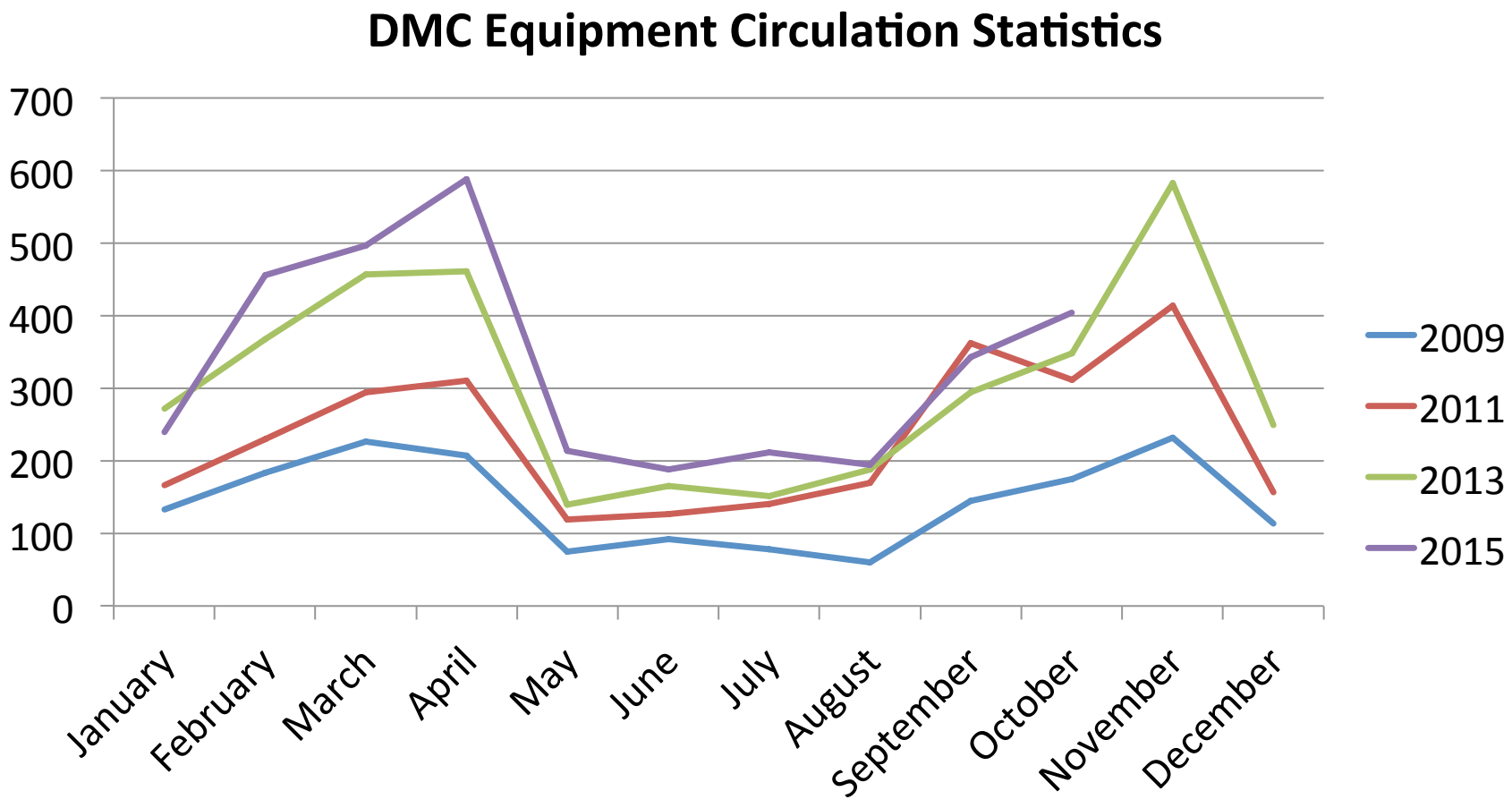


Avoid 3-D Pie Chart

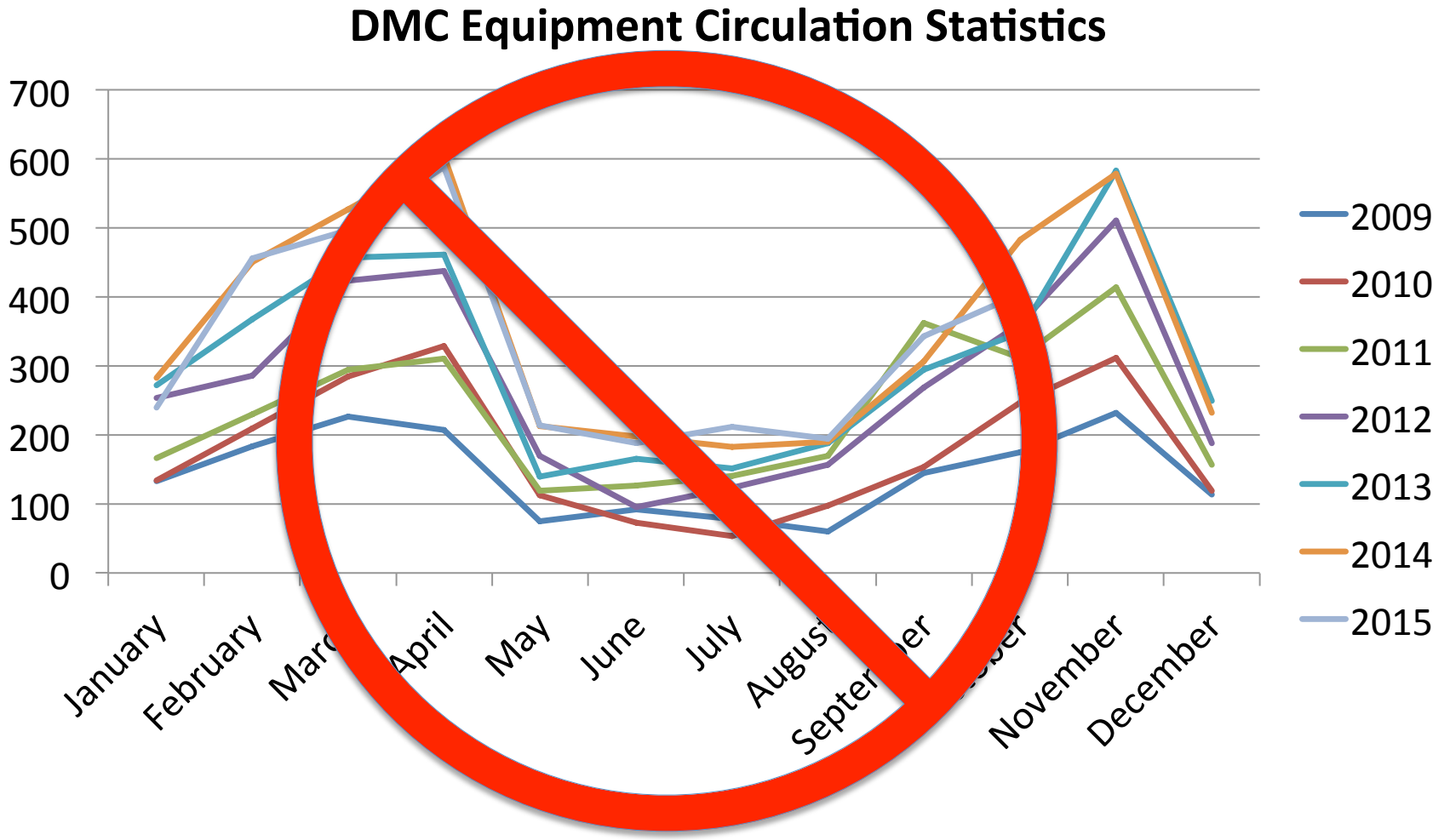


Show Me the Numbers, P8

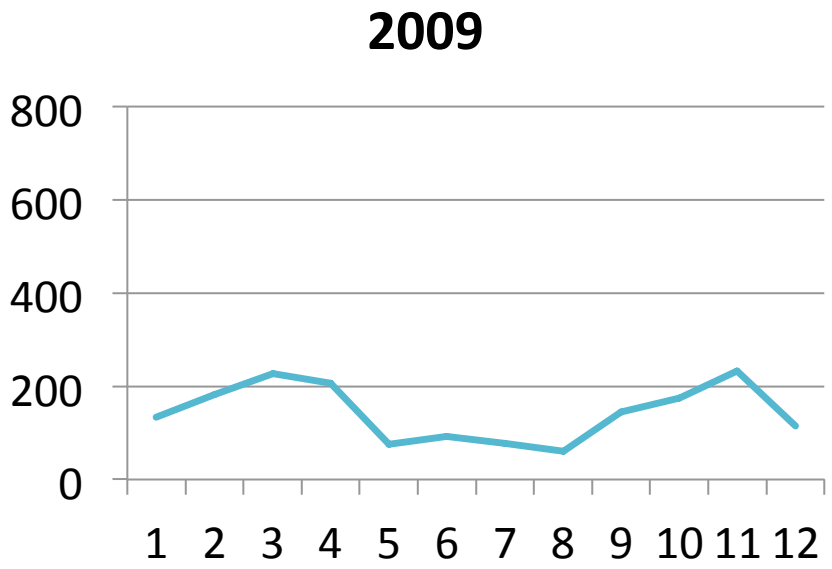
Line Chart for Time Series

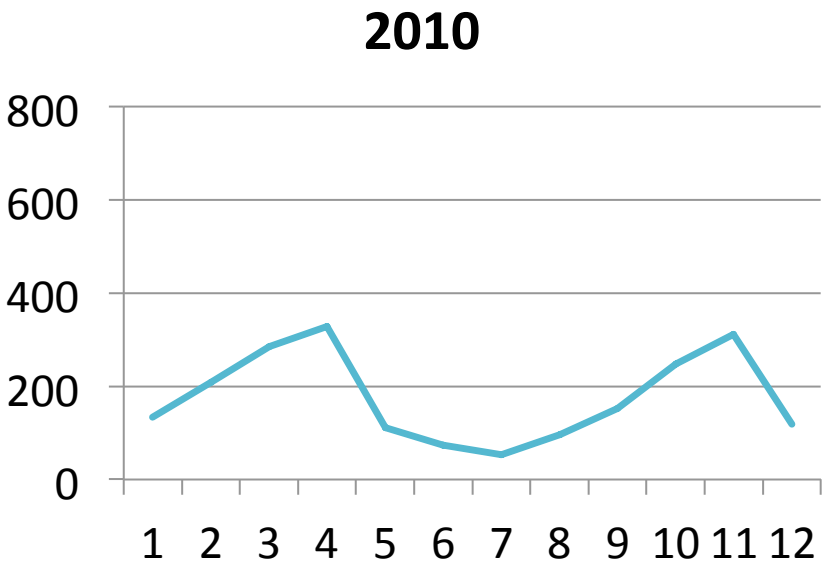


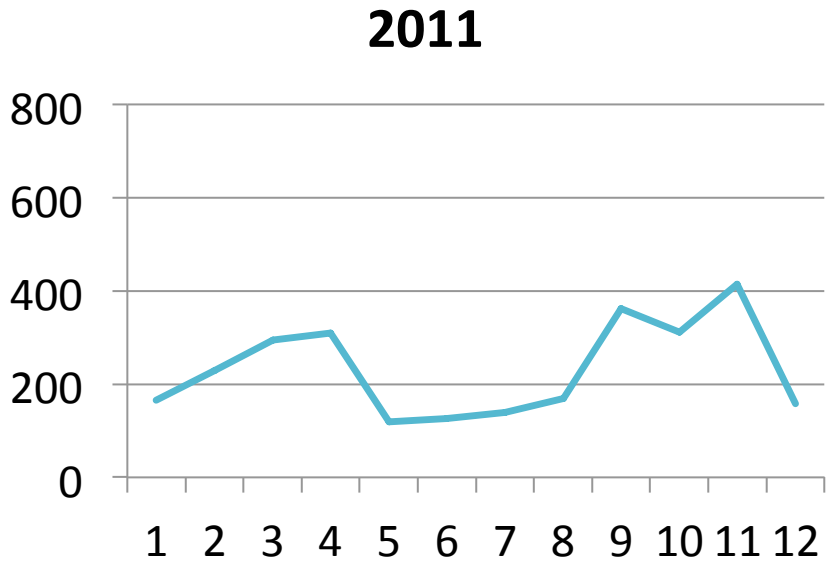
Keep the Line Chart to Four or Fewer. Otherwise the Chart is Too Busy!

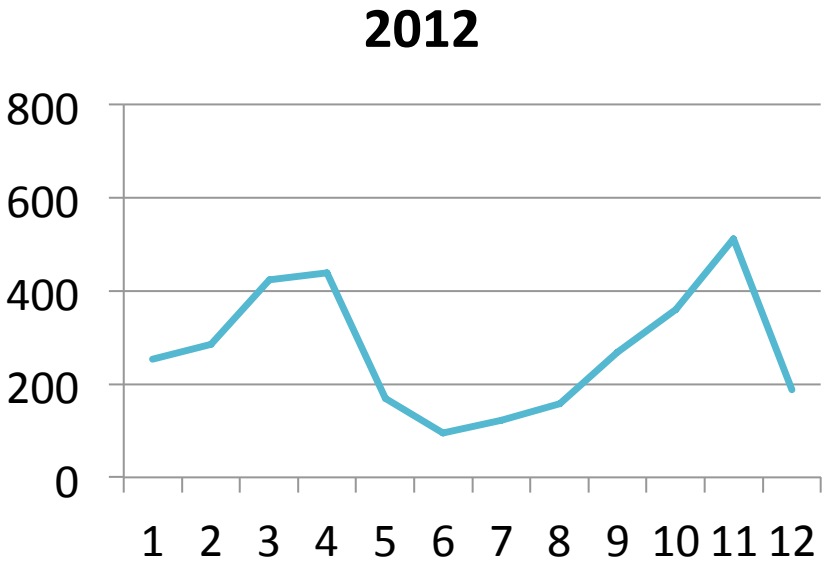


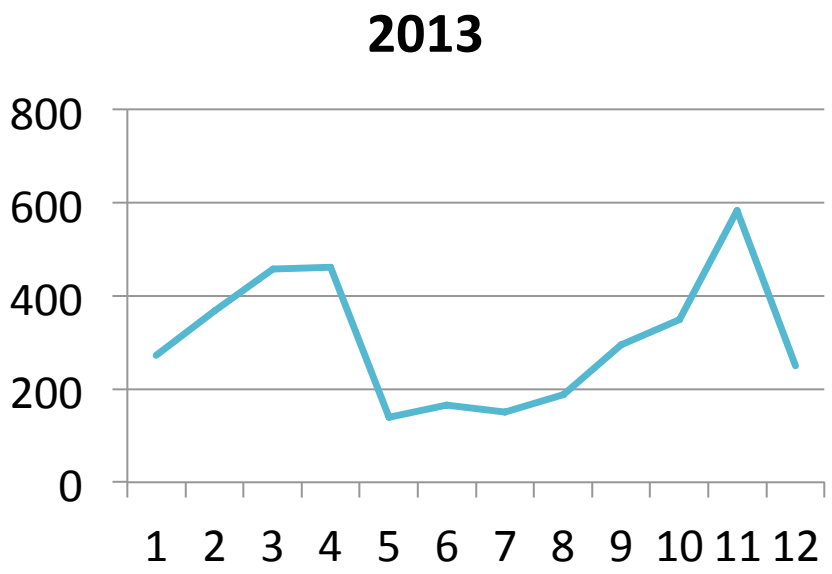
Use the Practice of Paneling and a Constant Scale for Consistency if You have More Than Four Lines.

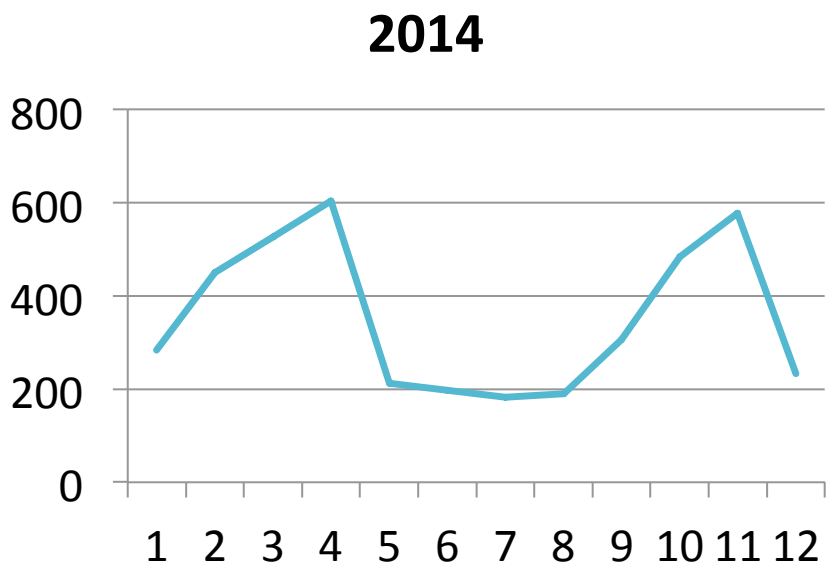


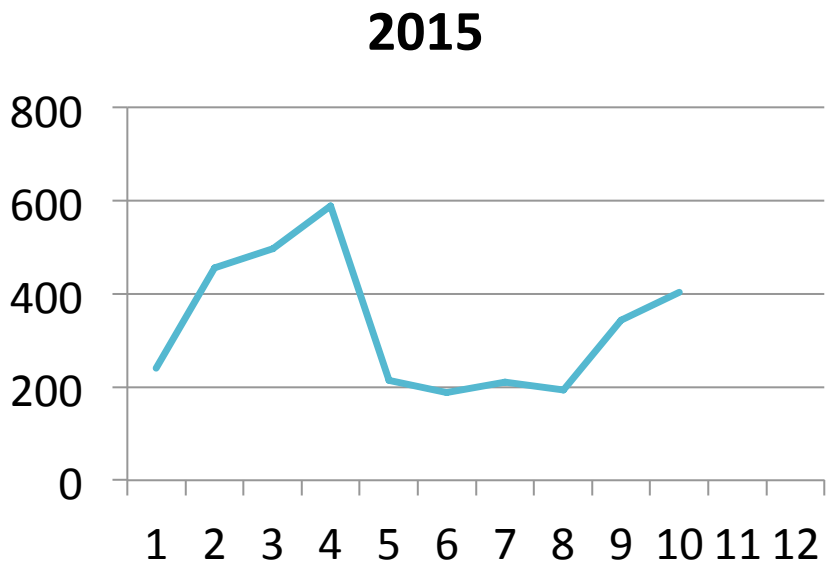




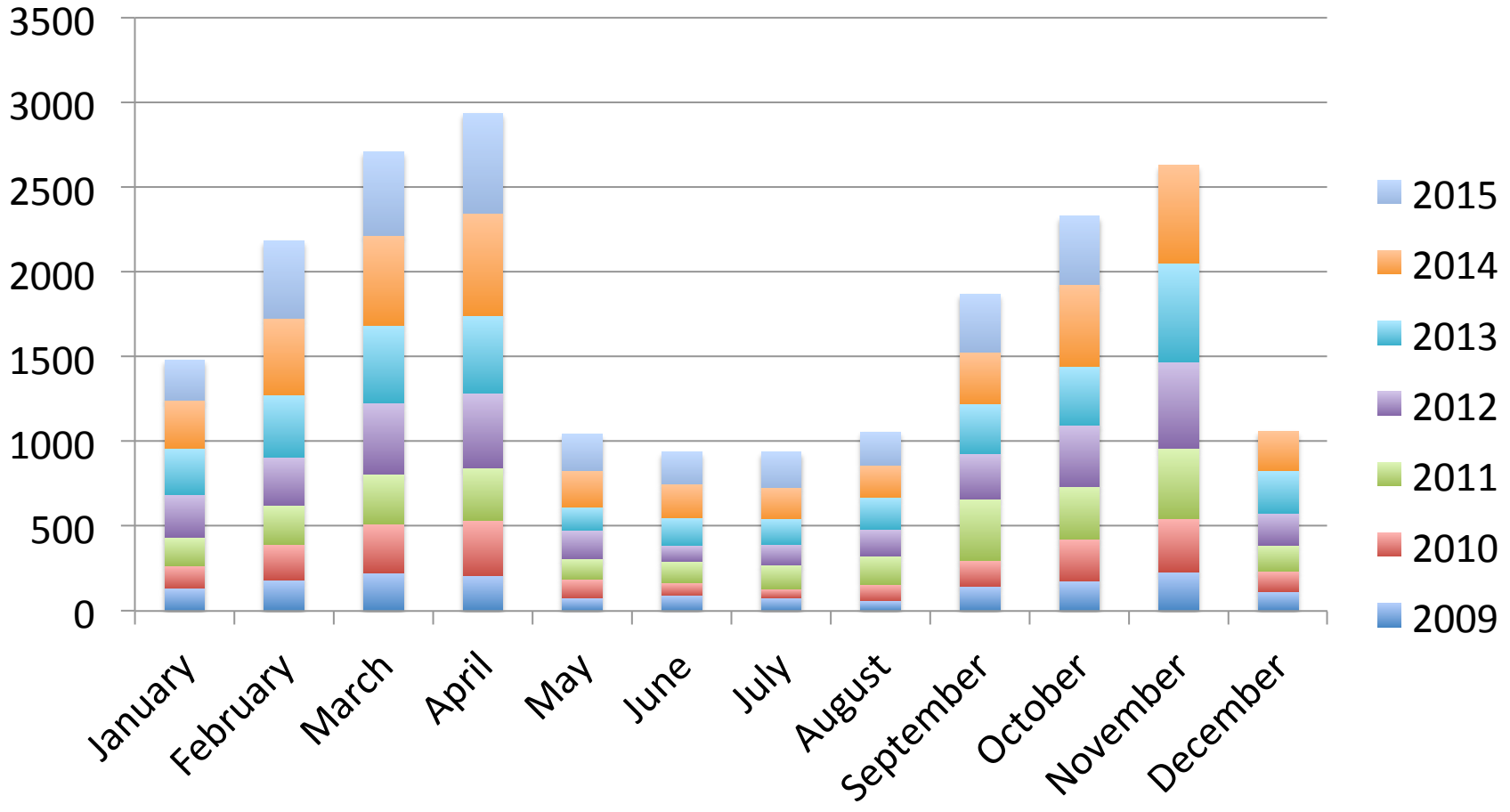



2013
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12

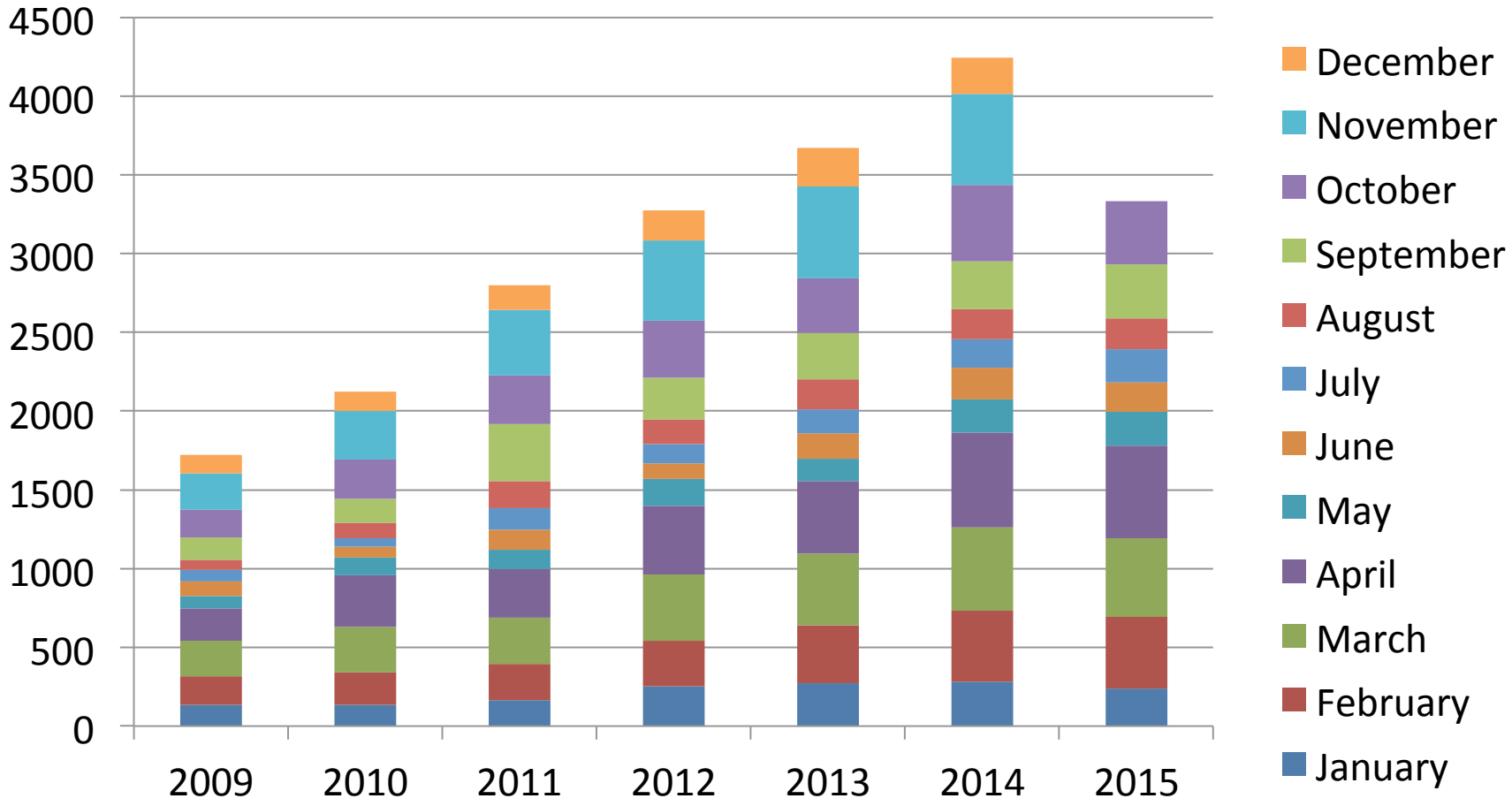

2014
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12


2015
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10 11 12

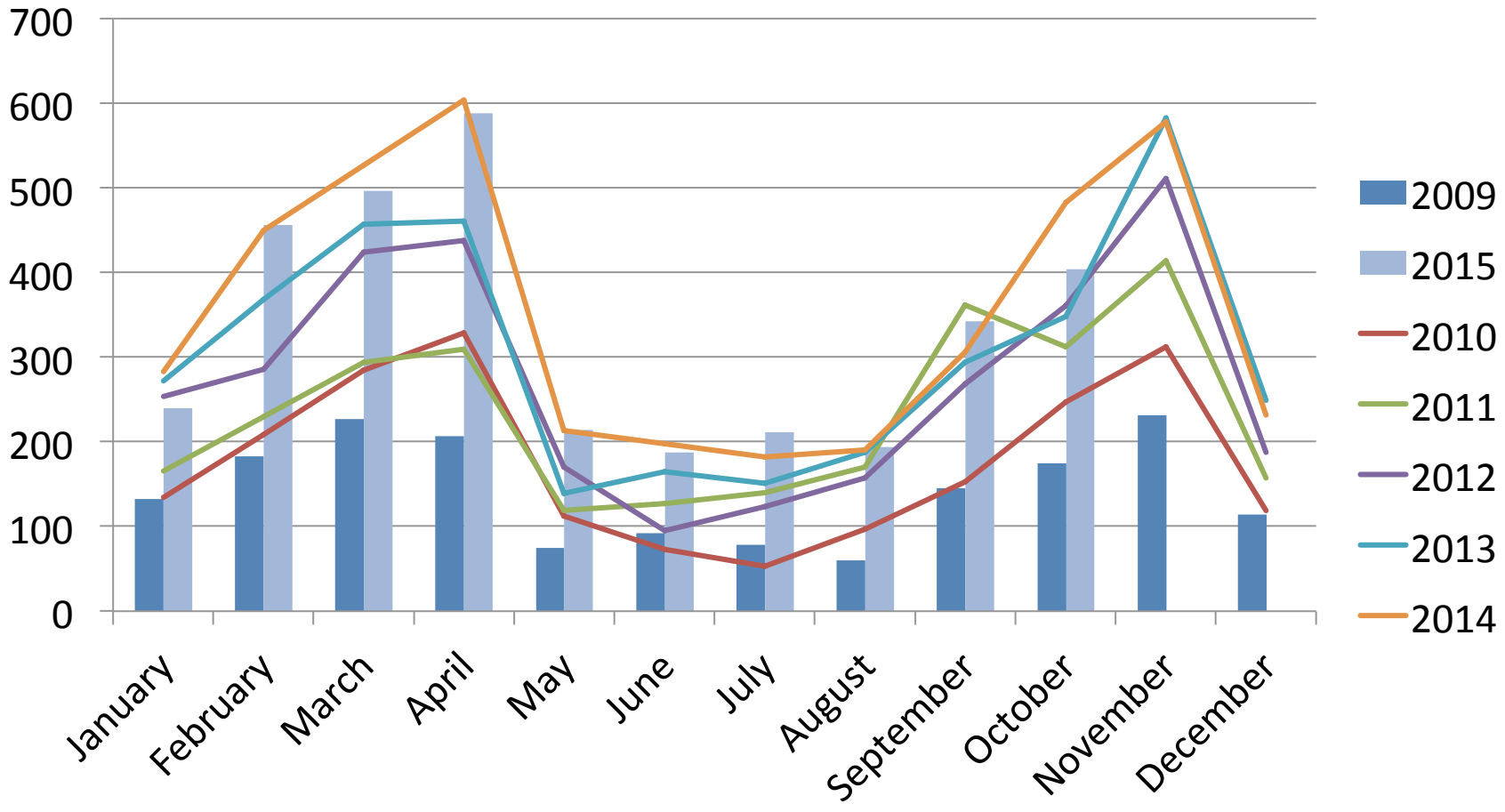
Stacked Bar Chart for Multiple Part-to-whole Relationships



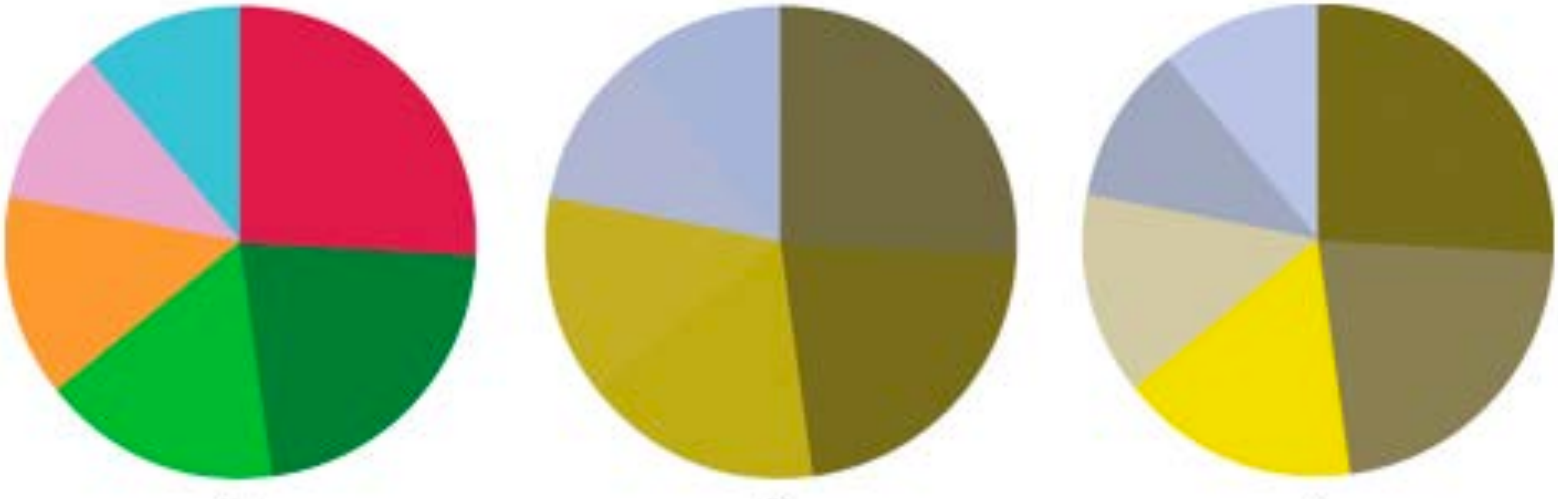
Stacked Bar Chart for Multiple Part-to-whole Relationships



A Mix of Bar Chart and Line Chart



Use Photoshop or Illustrator to Color-proof for Color Blindness



Adjusting design for color blindness
***A.** Original image **B.** Color-blind proof **C.** Optimized design*

http://help.adobe.com/en_US/creativesuite/cs/using/WS3F71DA01-0962-4b2e-B7FD-C956F8659BB3.html

Use Color Schemes That Are Color-blind Friendly.

Set of colors that is unambiguous both to colorblinds and non-colorblinds

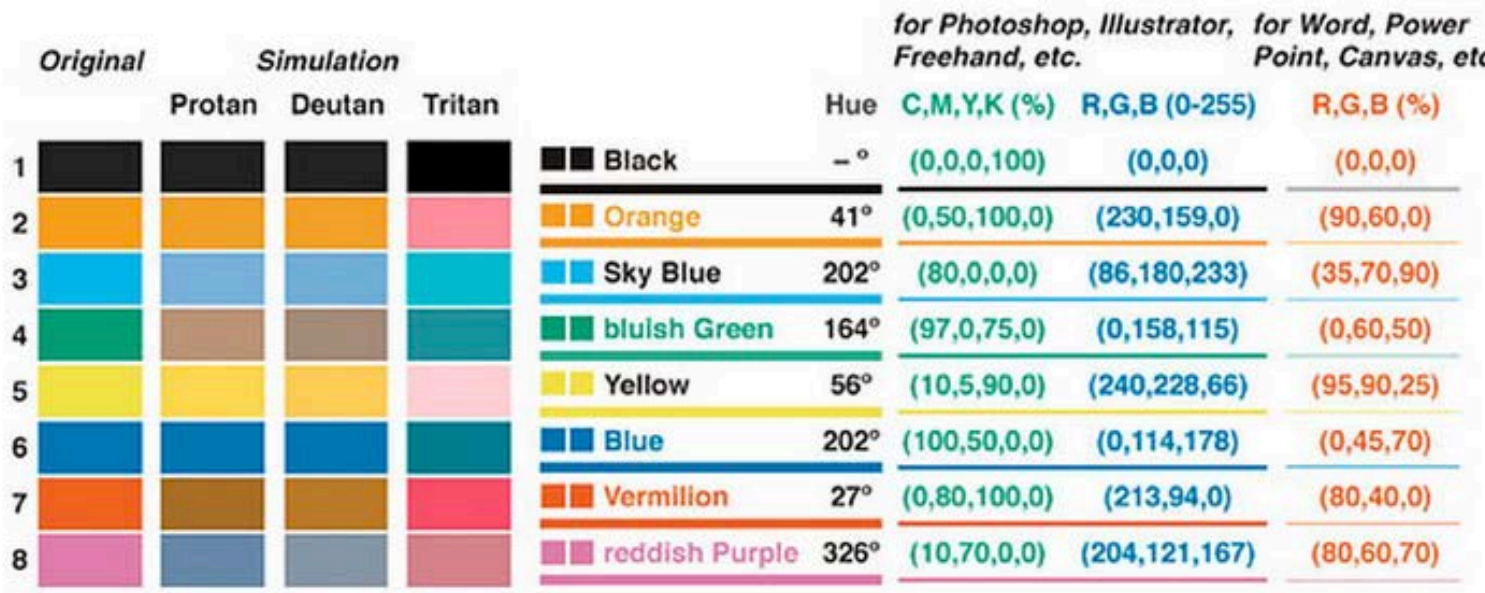


Fig. 16 Colorblind barrier-free color pallet

http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0000HT

Use Color Brewer as a Reference to Create Color-blind Friendly Color Scheme

- http://colorbrewer2.org/ – Color Advice for Cartography

TOOLS FOR CREATING INFOGRAPHICS AND DATA VISUALIZATION

Desktop Tools – Vector Graphics

PowerPoint	Excel	Adobe Illustrator	Adobe InDesign
			
Gephi(free)	OmniGraffle	InkScape(free)	
			

Digital Media Commons

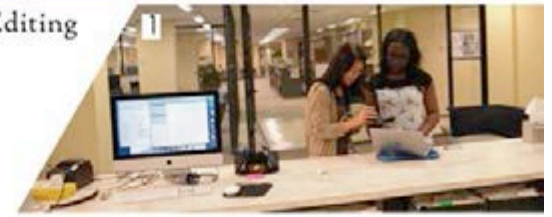
Will Wang , Jane Zhao

Our Mission

The DMC supports the creation and use of multimedia in education, scholarship, and creative expression. Working toward this end, we provide services that include hands-on training, assistance with digital projects, and access to the essential tools for creating digital resources such as digital video and audio, images and animations, infographics, PowerPoint presentations, web pages, and more.

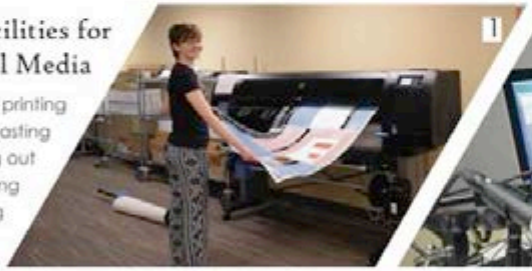
DMC Offers Hands-on Training on Media Editing and Assistance with Various Digital Projects

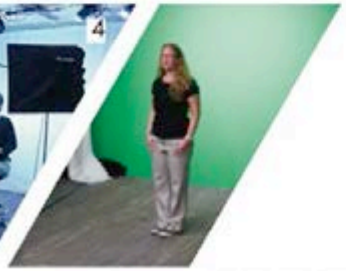
1. Help with using DMC equipment
2. Demonstration of DMC equipment
3. Assistance on video/audio editing, and graphics creation
4. Consultation on patron's project
5. Short courses for using digital tools



DMC Provides Access to the Essential Tools and Facilities for Creating Digital Media

1. Poster printing
2. Skyping/Podcasting
3. Equipment available for checking out
4. Lecture/interview recording
5. Photo taking
6. iMovie, Final Cut Pro, Photoshop, Illustrator, InDesign, and more





Your Projects, Our Passion!

dmc.rice.edu | dmc-info@rice.edu | 713-348-3635

InDesign Work

Desktop Tools – Image Editing

Adobe Photoshop	**Gimp(free)**
	
Pixelmator	Acorn
	

Gephi

- https://dhs.stanford.edu/tools/maps-graphs-and-workshops/

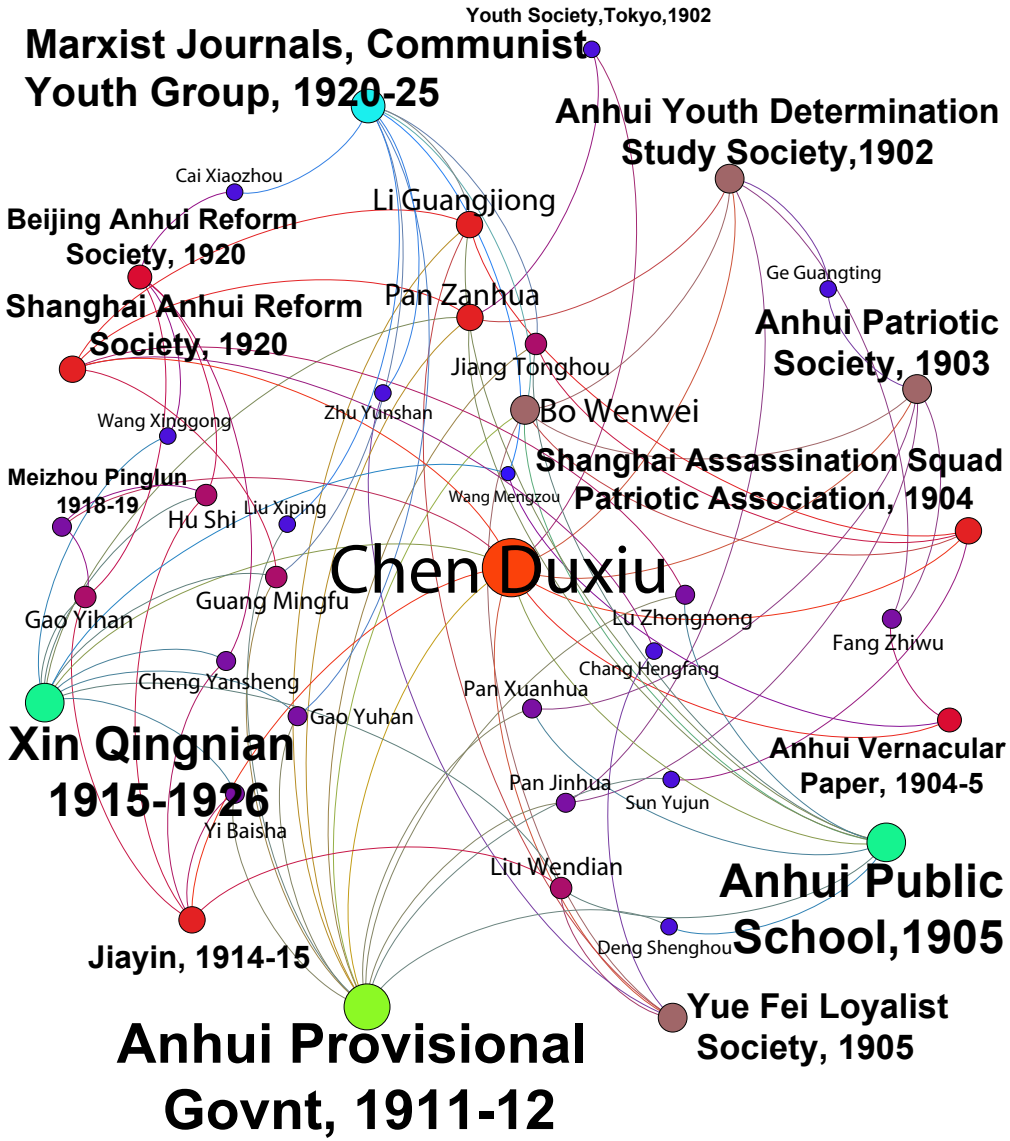


Diagram 2 - Anhui Co-Provincials in Chen's Networks

Online Tools

- **Wordle.net** http://www.wordle.net/
- **Google Chart** https://developers.google.com/chart/
- **Tableau Public** https://public.tableau.com/s

Online Infographics Resources

- Periodic Table of Visualization Methods
- The Noun Project
- 22 free tools for data visualization and analysis
- infographics world

More Sample Infographics

- Cool infographics
  http://www.coolinfographics.com/
- Edward Tufte
  http://www.edwardtufte.com/tufte/posters
- Information is beautiful
  http://www.informationisbeautiful.net/
  by David McCandless, an author and designer.

Data Sources

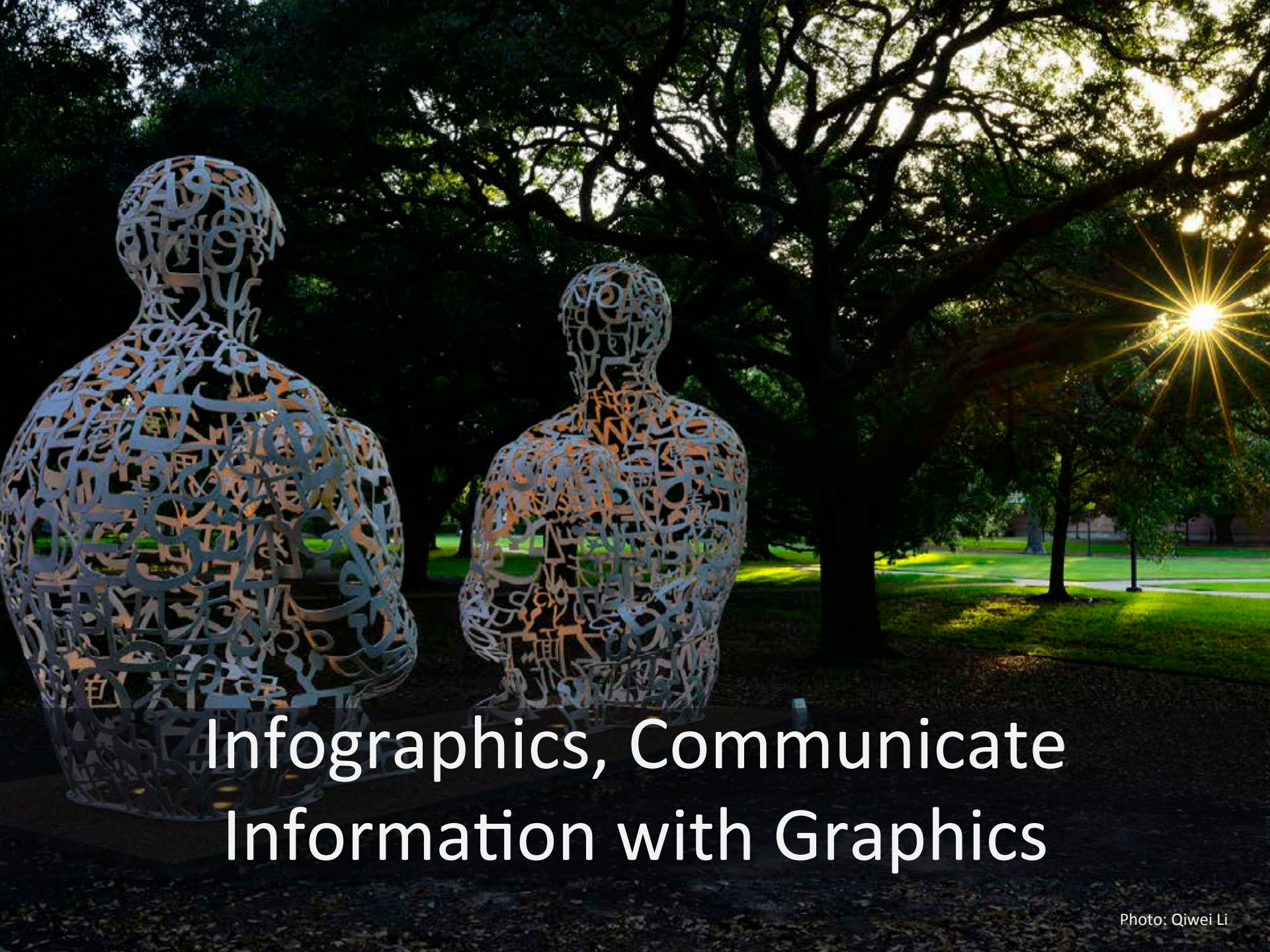
- **data.gov**
  http://www.data.gov/.
- FactBrowser
  http://www.factbrowser.com/
- Google Public Data
  http://www.google.com/publicdata/directory
- Wolfram Alpha
  http://www.wolframalpha.com/
- Wikipedia
  https://en.wikipedia.org/wiki/Main_Page

On Campus Resources

- Data Visualization Center
- Kelly Center for Government Information, Data, and Geospatial Services
- GIS Data Center

Summary

- A good infographic should be useful, sound, and beautiful.
- Best practices
  - Information Design
    - Keep it simple
    - Use a simple text combined with a relevant image
    - Make it unique
  - Data Visualization
    - Bar Chart – for ranking and time series, starting with zero baseline, avoid 3-D
    - Pie Chart – for part-to-whole comparisons, limiting to 5 slicess, avoid 3-D
    - Line Chart – for time series, limiting to 4 or less
- Use color schemes that are color-blind friendly
- Tools
  - PowerPoint, Excel, Illustrator, InDesign, Gephi, Photoshop, Gimp
  - Wordle, Google Chart, Tableau Public



Infographics, Communicate Information with Graphics

Photo: Qiwei Li