CSE 564
Visualization & Visual Analytics

VA System Design and Evaluation

Klaus Mueller

Computer Science Department
Stony Brook University

Lecture	Topic	Projects
1	Intro, schedule, and logistics	
2	Applications of visual analytics, basic tasks, data types	
3	Introduction to D3, basic vis techniques for non-spatial data	
4	Data assimilation and preparation	Project #1 out
5	Data assimilation and preparation	
6	Bias in visualization	
7	Data reduction and dimension reduction	
8	Visual perception	Project #2(a) out
9	Visual cognition	
10	Visual design and aesthetics	
11	Cluster analysis: numerical data	
12	Cluster analysis: categorical data	Project #2(b) out
13	High-dimensional data visualization	
14	Dimensionality reduction and embedding methods	
15	Principles of interaction	
16	Midterm #1	
17	Visual analytics	Final project proposal call out
18	The visual sense making process	
19	Maps	
20	Visualization of hierarchies	Final project proposal due
21	Visualization of time-varying and time-series data	
22	Foundations of scientific and medical visualization	
23	Volume rendering	Project 3 out
24	Scientific and medical visualization	Final Project preliminary report due
25	Visual analytics system design and evaluation	
26	Memorable visualization and embellishments	
27	Infographics design	
28	Midterm #2	

Outline

This lecture is about the human factor

- data science and analytics with the human in the loop
- **design** systems with the human in the loop
- **evaluate** systems with the human in the loop

Prologue

Overall definition of visual analytics

Interaction

Visualization

Data Science
Data Analytics
Data Mining

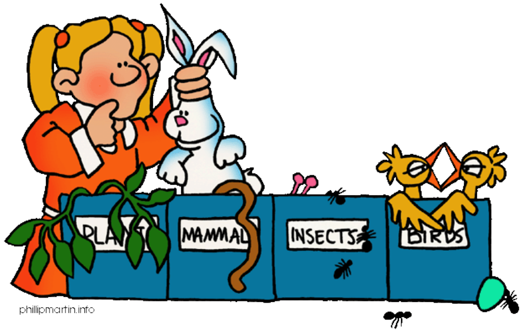
- What are the fundamental tasks of data science?
- How can humans assist in these?
- How can humans benefit from these?

Fundamental Tasks In *Visual* Data Science

Task #1: Classification

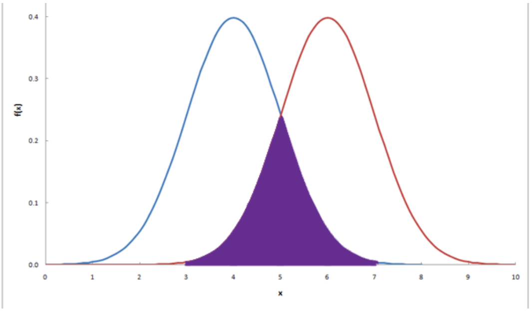
Predict which class a member of a certain population belongs to

- absolute
- probabilistic



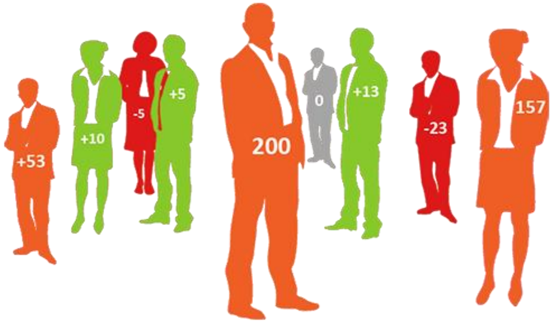
Require a classification model

- absolute
- probabilistic (likelihood)



Scoring with a model

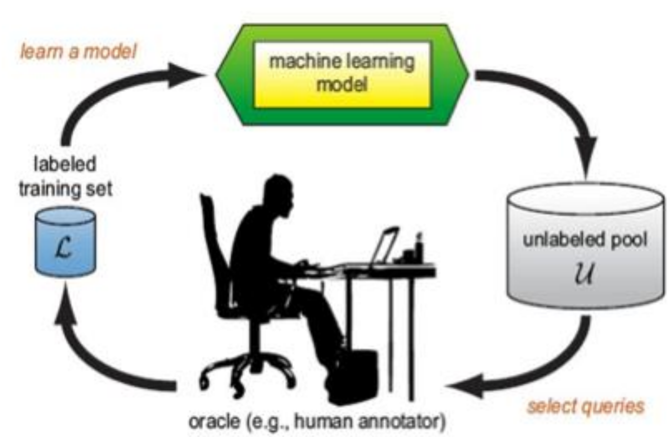
- each population member gets a score
  for a particular class/category
- sort each class or member scores to assign
- scoring and classification are related

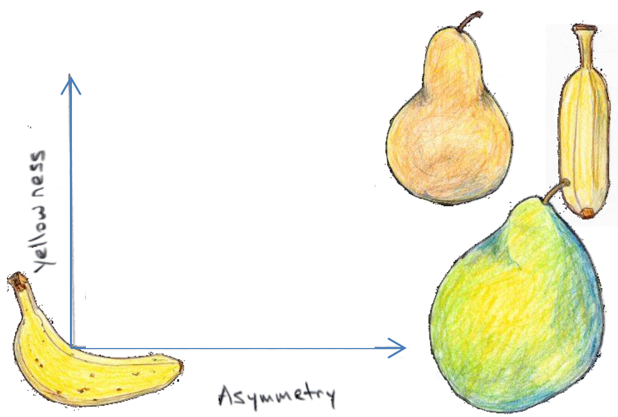


Classification: The Human Factor

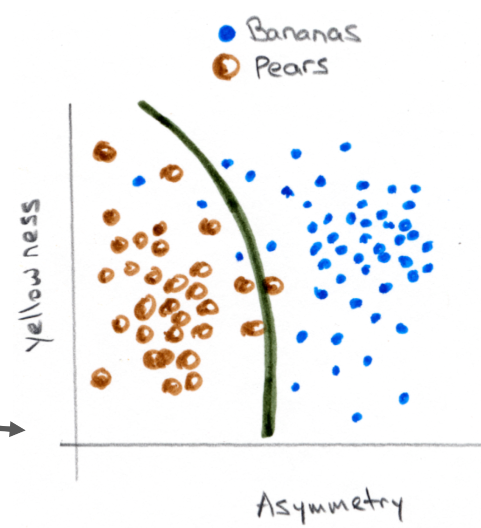
Supervised learning

- human labels the samples
- find a good feature vector
- build the classification model





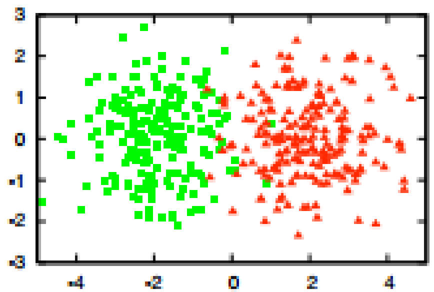
model



Classification: The Human Factor

Active supervised learning

- human labels the samples
- but while samples are often abundant, labeling can be expensive
- active learning → only label the samples critical to the model



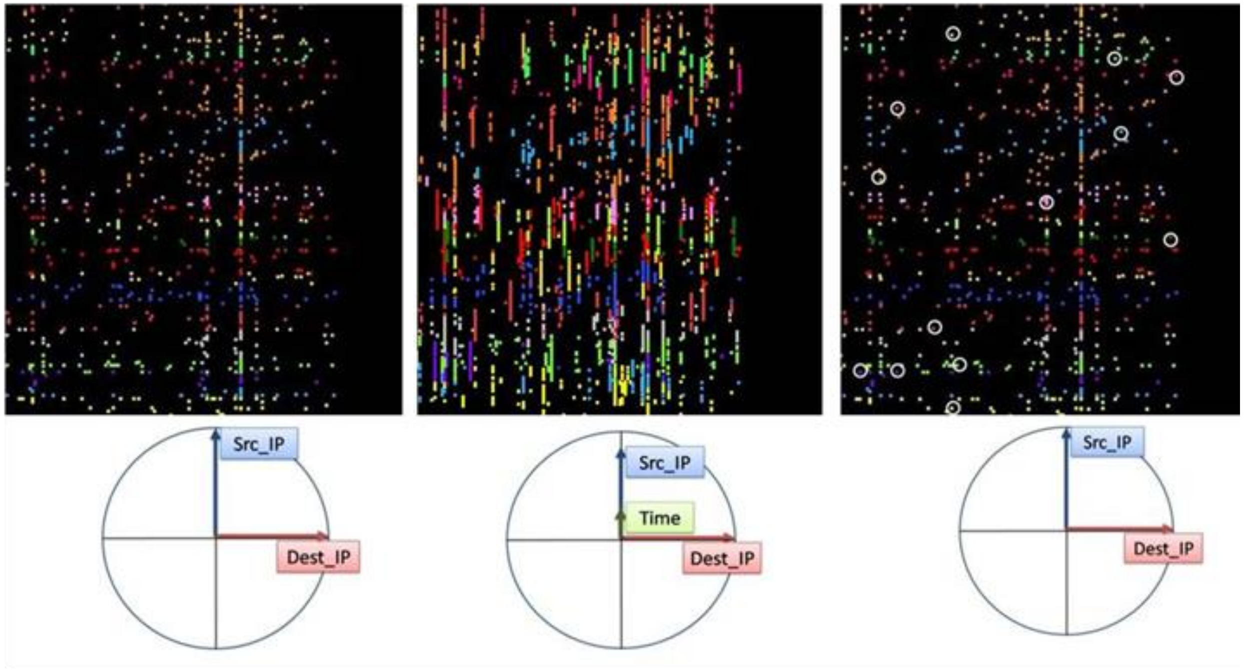
(a)

(a) Assume a toy data set of 400 instances, evenly sampled from two class Gaussians, visualized in 2D feature space.
(b) Learn a logistic regression model by training it with 30 labeled instances randomly drawn from the problem domain (70% accuracy)
(c) Learn a logistic regression model by training it with 30 actively queried instances using uncertainty sampling (90%)

Application: Visual Model Learning

Simple example: network traffic analysis

- the (very large) data set consists of a 1-hour snapshot of internet packets
- goal is to learn the concept 'webpage load'



Mark good examples

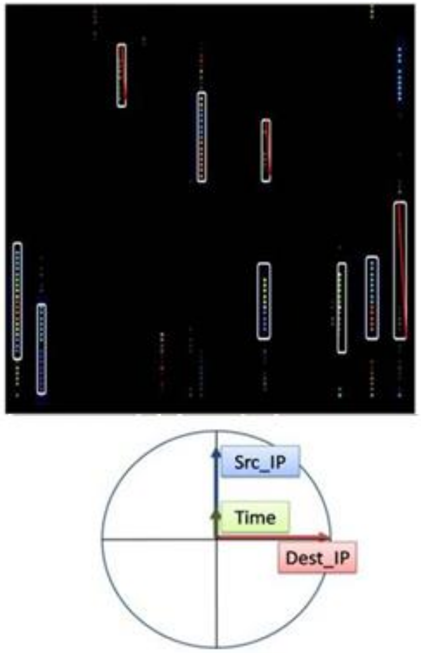
Visual Model Learning: Set Initial Rule

Use Inductive Logic Programming (Prolog) to formulate initial model (rule):

```
webpage_load(X) :-
  same_src_ips(X),same_dest_ips(X),same_src_port(X,80)
```

Visual Model Learning: Verify Initial Rule

Now we classify other data points with this rule and visualize



Mark negative examples

Visual Model Learning: refine Initial Rule

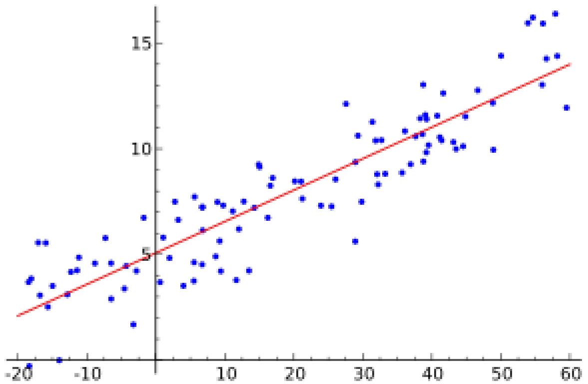
Marking negative examples yields updated/refined rule:

```
webpage_load(X) :-
same_src_ips(X),same_dest_ips(X),same_src_port(X,80),
timeframe_upper(X,10),length(X,L),greaterthan(L,8).
```

here: must contain at least 8 packets and be within a time frame of 10

Task #2: Regression

Regression = value estimation



Fit the data to a function

- often linear, but does not have to be
- quality of fit is decisive

Regression vs. classification

- classification predicts that something will happen
- regression predicts how much of it will happen

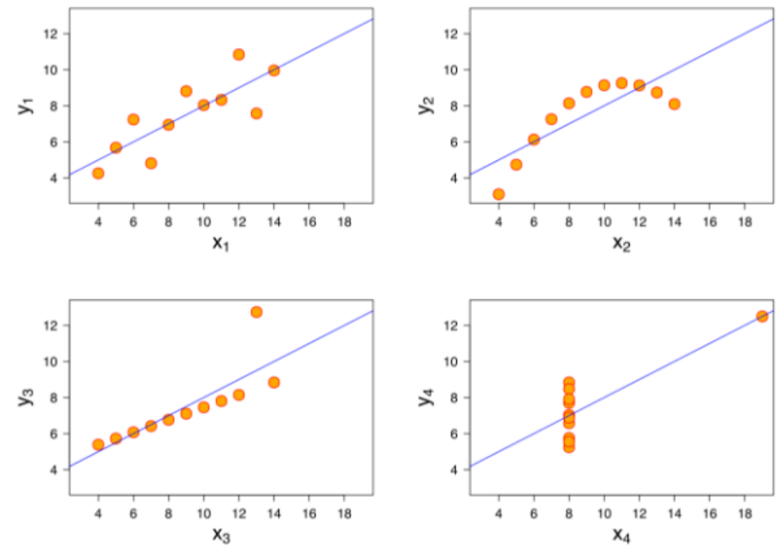
Human factor:

- identify possible outliers

Anscombe Quartet

Visualization of statistics results is important

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89



Property	Value
Mean of *x* in each case	9 (exact)
Sample variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Sample variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

Same statistics
Very different data

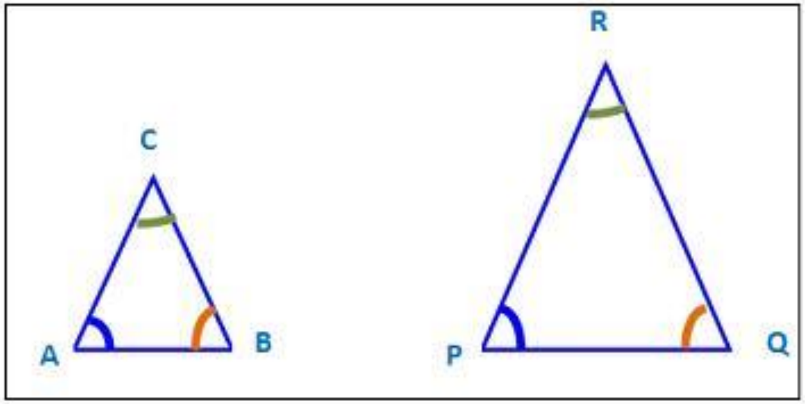
TASK #3: SIMILARITY MATCHING

Identify similar individuals based on data known about them

- need a measure of similarity
- features that define similarity
- characteristics

Similarity often part of

- classification
- regression
- clustering



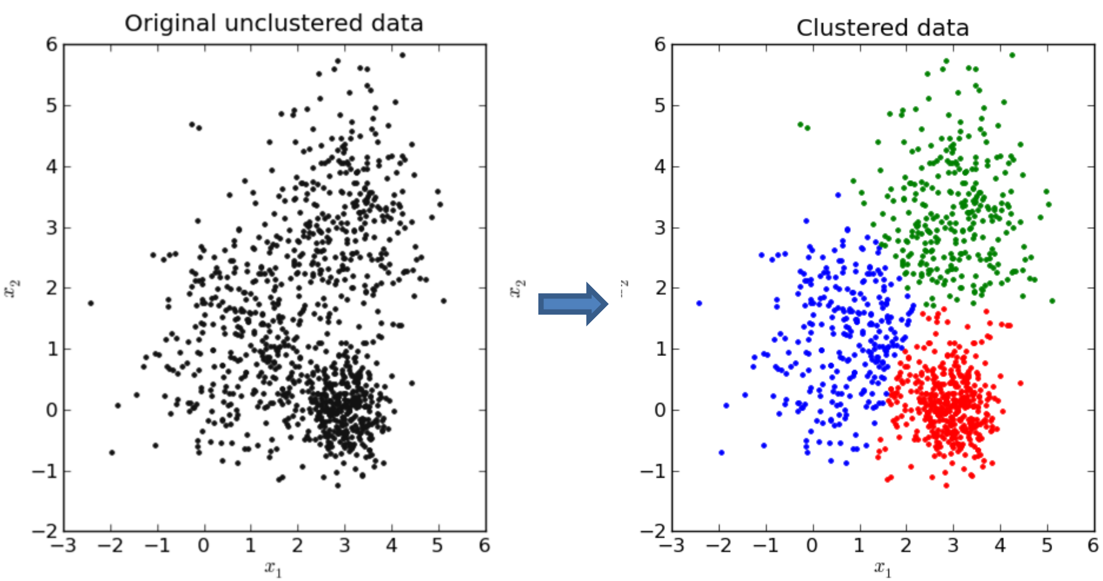
Human factor

- similar to supervised learning
- identify effective features

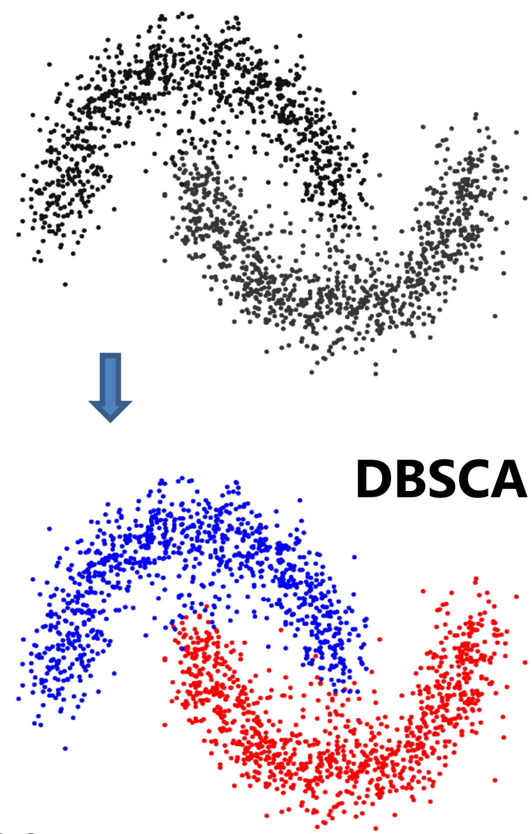
TASK #4: CLUSTERING

Group individuals in a population together by their similarity

- preliminary domain exploration to see which natural groups exist



k-means



DBSCAN

- this includes outlier detection
- outliers are the data that do not cluster
- human factor: labeling, verification, correction

Task #5: Co-Occurrence Grouping

Find associations between entities based on transactions involving them

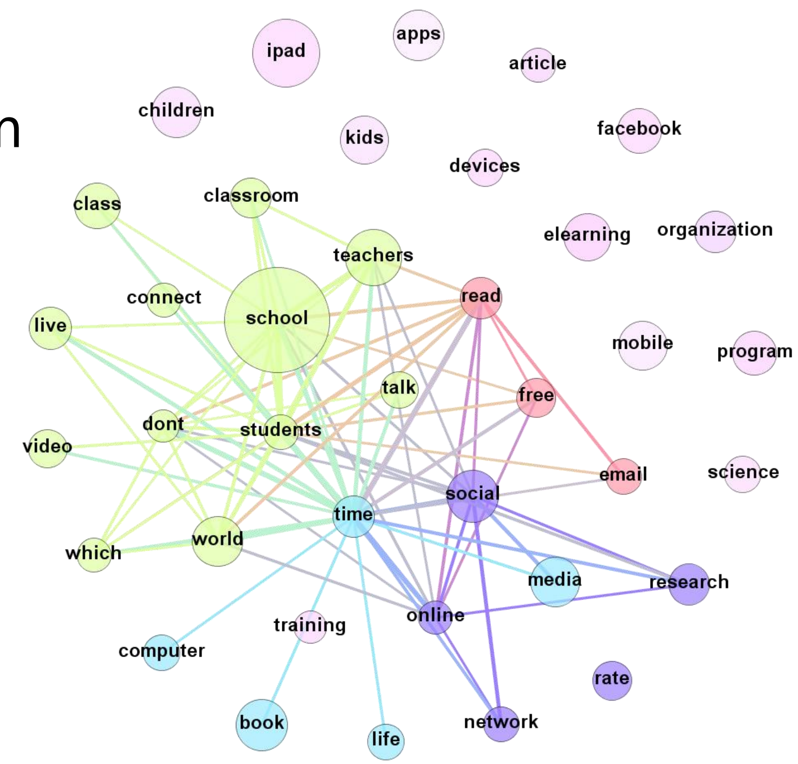
- what products are commonly purchased together?

Applications

- basket analysis
- recommender systems

Difference to clustering

- in clustering similarity is based on the object's attributes
- in co-occurrence similarity is based on objects appearing together



Human factor:

- **labeling**
- **verification**
- **correction**

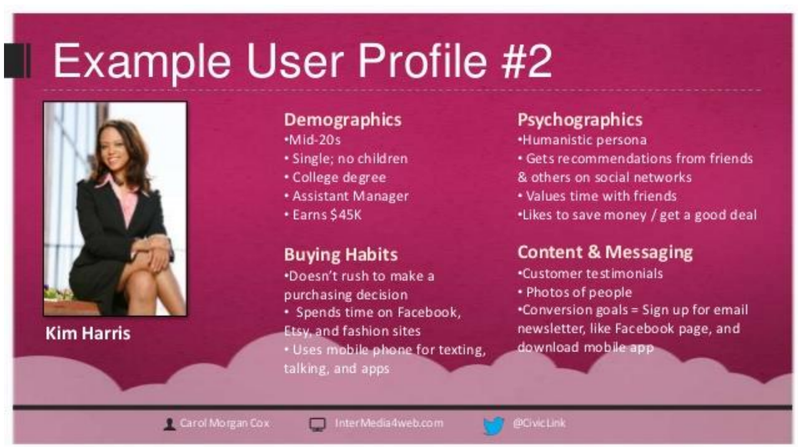
Task #6: Profiling

Also known as behavior description

- attempts to **characterize** the typical behavior of an individual, group, or population

Often used to establish behavioral norms for **anomaly detection**

- fraud detection
- intrusion detection



Examples:

- credit card fraud
- airport security

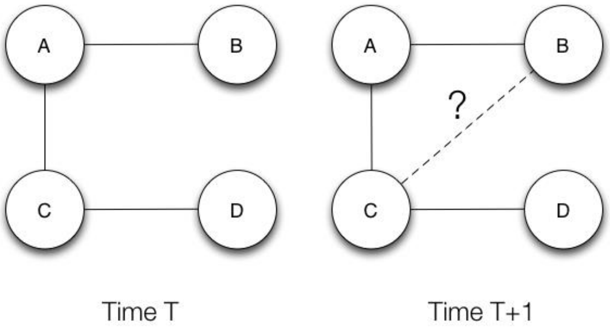
Human factor:

- **labeling, verification, correction**

Task #7: Link Prediction

Predict connections between data items

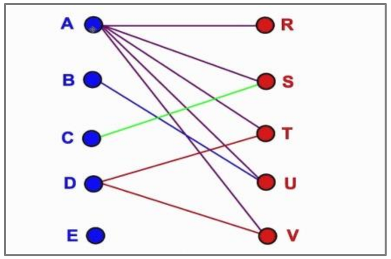
- usually works within a graph
- predict missing links
- estimate link strength



Applications

- in recommendation systems
- friend suggestion in Facebook (social graph)
- link suggestion in LinkedIn (professional graph)
- movie suggestion in Netflix (bipartite graph people – movies)

people



movies

Human factor:

- **labeling**
- **verification**
- **correction**

TASK #8: DATA REDUCTION

Take a large dataset and substitute it with a smaller one

- keep loss of information minimal
- clustering and cleaning
- importance sampling
- dimension reduction
- data abstraction
- big data → small data
- find *latent* variables

Example for latent variable – Movie *Taste*

- not directly measurable – latent variable
- derive from movie viewing preferences
- can reveal genre, etc.

Human factor:

- **labeling**
- **verification**
- **correction**

Task #9: Causal Modeling

Understand what events or actions influence others

Different from predictive modeling
- tries **to explain why** the predictive model worked (or not)

Potentially unreliable when done from observational data
- conducting a targeted experiment is better, but often impossible
- have to work with observational (often anecdotal data)
- hence there is a clear human factor: verify the model, correct it, edit it

Builds on counterfactual analysis
- an event is causal if mutating it will lead to undoing the outcome
- “If only I hadn't been speeding, my car wouldn't have been wrecked”
- downward vs. upward counterfactual thinking
- can explain happiness of bronze medalists vs. silver medalists
- just making the grade vs. just missing the grade

Case Study: What Causes Low MPG

The Car Data Set

Consider the salient features of a car (not really big data):

- miles per gallon (MPG)
- top speed
- acceleration (time to 60 mph)
- number of cylinders
- horsepower
- weight
- country origin

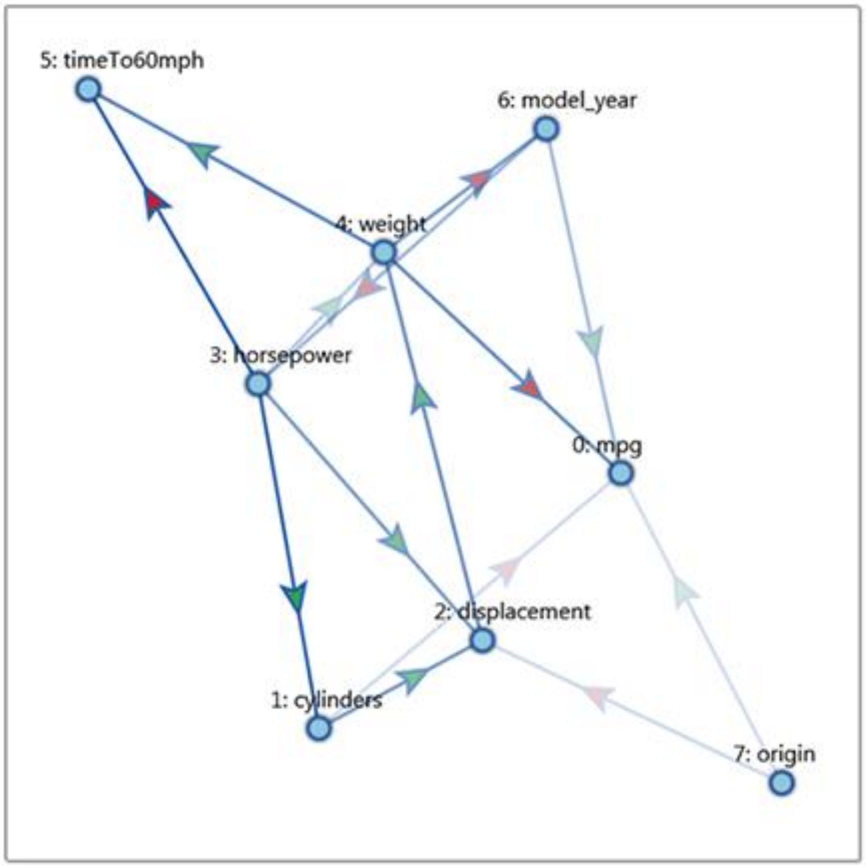
400 cars from the 1980s

Shown in a Spreadsheet

A1 | f_x | Urban population

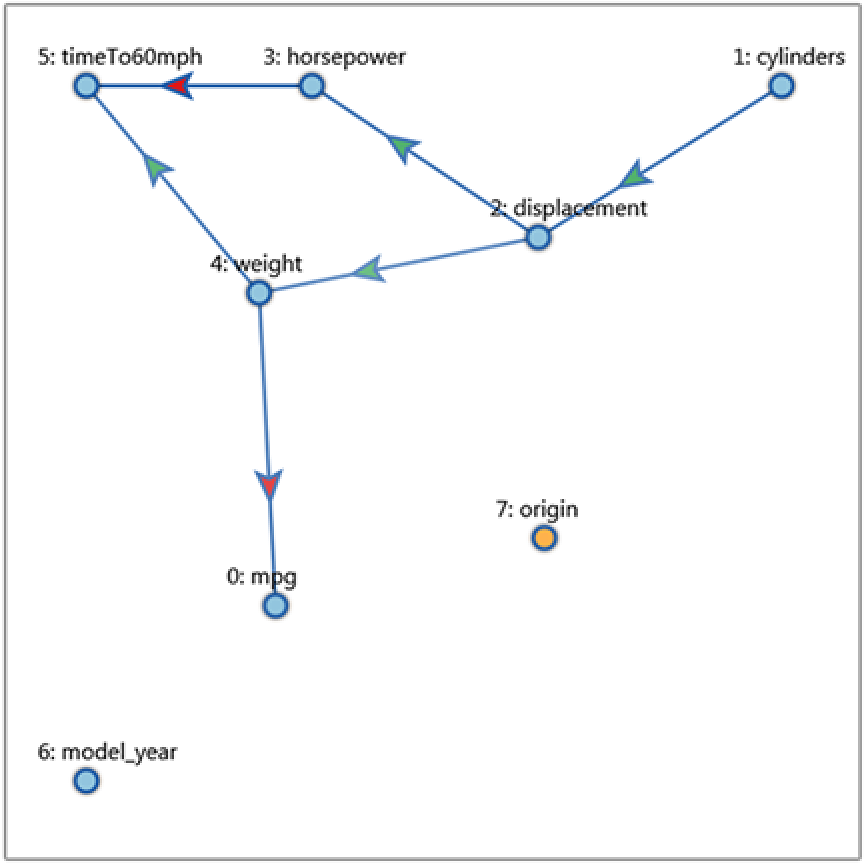
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Urban population**	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974
2	Afghanistan	769308	811389	855131	900646	948060	997499	1053104	1110728	1170961	1234664	1302370	1391081	1483942	1579748	1676656
3	Albania	494443	511637	529182	547024	565117	583422	601897	620508	639234	658062	676985	698179	719561	741149	762972
4	Algeria	3293999	3513320	3737362	3969886	4216744	4483048	4644898	4822860	5015071	5218184	5429743	5618190	5813978	6017932	6231383
5	American Samoa	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
6	Andorra	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
7	Angola	521205	552777	585121	618345	652638	688181	729595	772643	817418	863993	912486	982944	1056617	1133936	1215437
8	Antigua and Barbuda	21699	21737	21878	22086	22309	22513	22717	22893	23053	23218	23394	24046	24718	25342	25826
9	Argentina	15224096	15588864	15957125	16328045	16700303	17073371	17432905	17793789	18160868	18540720	18938137	19335571	19750609	20180707	20621674
10	Armenia	957974	1008899	1061551	1115546	1170414	1225785	1281346	1337060	1393199	1450241	1508526	1565054	1622558	1680709	1739019
11	Aruba	24996	25514	26019	26498	26941	27337	27683	27984	28247	28491	28726	28959	29188	29409	29610
12	Australia	8375329	8585577	8840666	9055650	9279777	9508980	9770529	9937118	10157212	10416192	10668471	11050785	11271606	11461308	11771589
13	Austria	4560057	4589541	4621666	4653194	4685421	4715750	4754585	4778506	4798552	4817322	4849178	4871380	4904030	4932109	4939292
14	Azerbaijan	1857673	1929429	2004258	2080816	2157307	2232355	2306310	2378380	2448728	2517815	2586000	2660687	2734631	2807879	2880491
15	Bahamas	65457	69655	74179	78961	83902	88918	93931	98974	103944	108721	113219	117339	121142	124761	128393
16	Bahrain	128480	133815	139791	146052	152097	157596	162844	167630	172373	177677	183997	191379	199768	209201	219678
17	Bangladesh	2761049	2947191	3141372	3344120	3556037	3777716	4047121	4329144	4624445	4933701	5257558	5710277	6184871	6682073	7202503
18	Barbados	84884	85284	85761	86285	86797	87259	87707	88117	88526	88986	89532	90518	91596	92713	93796
19	Belarus	2656152	2774166	2896449	3022217	3150553	3280410	3415984	3554673	3695363	3836802	3977600	4131179	4285735	4439788	4591705
20	Belgium	8435075	8489549	8548773	8620194	8709437	8796088	8865259	8924327	8968568	9003536	9040444	9086816	9134227	9175144	9217085
21	Belize	49165	50608	52156	53734	55226	56561	57756	58820	59746	60532	61186	61883	62445	62984	63665
22	Benin	211033	229172	248065	267765	288321	309788	337282	366019	396065	427482	460341	500355	542251	586179	632320
23	Bermuda	44400	45500	46600	47700	48900	50100	51000	52000	53000	54000	55000	54600	54200	53800	53400
24	Bhutan	8064	8778	9526	10311	11137	12010	13089	14230	15445	16750	18158	19926	21827	23858	26008
25	Bolivia	1233398	1271250	1310294	1350615	1392328	1435536	1480255	1526529	1574517	1624419	1676370	1730434	1786553	1844596	1904355
26	Bosnia and Herzegovi	604204	637337	671124	705395	739884	774380	812856	851325	890011	929301	969514	1008688	1048890	1089898	1131315
27	Botswana	16240	17379	18583	19855	21203	22631	28191	34090	40352	46995	54038	61638	69689	78254	87422
28	Brazil	32662018	34463344	36353068	38320171	40346703	42418482	44548227	46722996	48945984	51223962	53563179	56042505	58587770	61207586	63913385
29	Brunei	35501	38753	42173	45802	49699	53916	58461	63355	68595	74157	80024	83802	87671	91616	95629
30	Bulgaria	2918659	3085061	3251675	3418610	3588246	3756058	3889518	4022040	4159890	4301340	4440270	4554810	4667059	4782931	4907107
31	Burkina Faso	221872	230199	238713	247472	256558	266039	275958	286311	297074	308196	319642	332556	345877	359655	373966
32	Burundi	58810	61055	63344	65696	68137	70683	73370	76186	79034	81779	84324	90879	97308	103757	110494
33	Cambodia	559631	578678	598248	618631	640243	663272	747219	835638	927177	1019449	1110079	962037	806676	645287	479631
34	Cameroon	751711	801009	852578	906523	962928	1021891	1088521	1158289	1231375	1307967	1388275	1522958	1664410	1813278	1970385
35	Canada	12375125	12764121	13145207	13536503	13941055	14345262	14727261	15108962	15470875	15800439	16142268	16381341	16640381	16920220	17221765
36	Cape Verde	32791	34353	35972	37672	39487	41435	43592	45884	48200	50383	52314	54103	55620	56940	58184
37	Cayman Islands	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
38	Central African Rep.	302157	317715	333986	351001	368787	387357	408129	429825	452326	475441	499036	526414	554452	583376	613530
39	Chad	198777	213406	228652	244499	260903	277834	305390	333898	363523	394530	427153	467662	510348	554973	601045
40	Channel Islands	42565	42665	42792	42941	43102	43269	43437	43604	43765	43916	44051	44028	43987	43907	43762

Global Layout of The Car Data

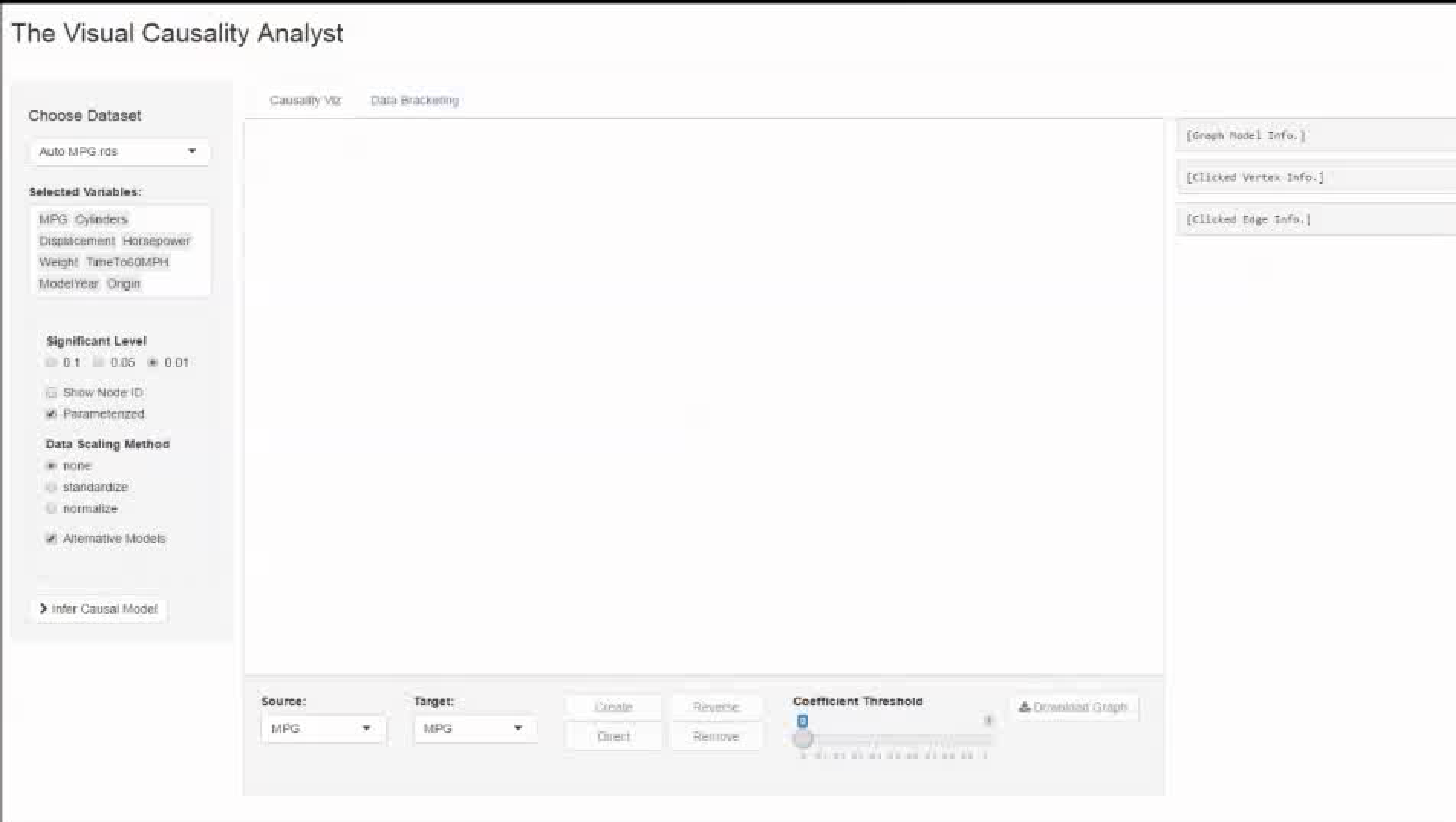


Random

Seeking the Cause of Low MPG



Isolating MPG


The Visual Causality Analyst
Causality Viz
Data Bracketing
Choose Dataset
Auto MPG.rds
Selected Variables:
MPG Cylinders
Displacement Horsepower
Weight TimeTo60MPH
ModelYear Origin
Significant Level
0.1 0.05 0.01
Show Node ID
Parameterized
Data Scaling Method
none
standardize
normalize
Alternative Models
Infer Causal Model
[Graph Model Info.]
[Clicked Vertex Info.]
[Clicked Edge Info.]
Source:
MPG
Target:
MPG
Create
Reverse
Direct
Remove
Coefficient Threshold
Download Graph

How To
Design A Visual Analytics Solution

Use the nested model

- devised by Tamara Munzner (UBC)
- M. Meyer, M. Sedlmair, P. Quinan, T Munzner, "The nested blocks and guidelines model," *Information Visualization*, 2013

Step 1: Characterize the Problem

Define the tasks, data, workflow of target users

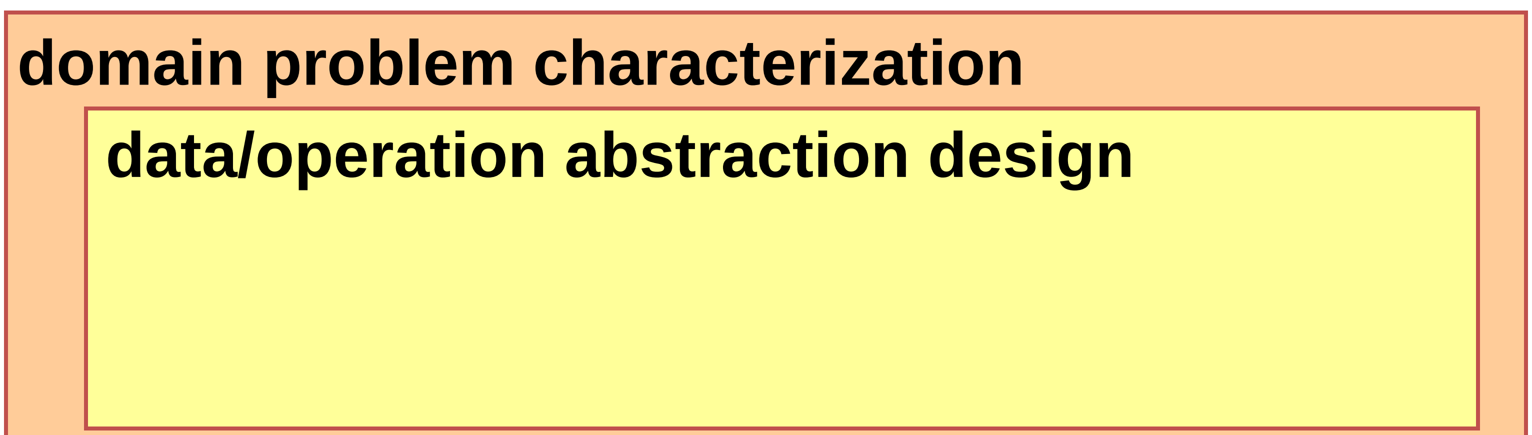
- the tasks are usually described in domain terms
- finding and eliciting the requirements is notoriously hard
- observe how domain users work and perform their tasks
- observe the pains they are having
- what are the limitations?
- what is currently impossible, slow, or tedious?

domain problem characterization

Step 2: Abstract Into a Design

Map from domain vocabulary/concerns to abstraction

- may require some sort of transformation
- data and types are described in abstract terms
- numeric tables, relational/network, spatial, ...
- tasks and operations described in abstract terms
- generic activities: sort, filter, correlate, find trends/outliers...



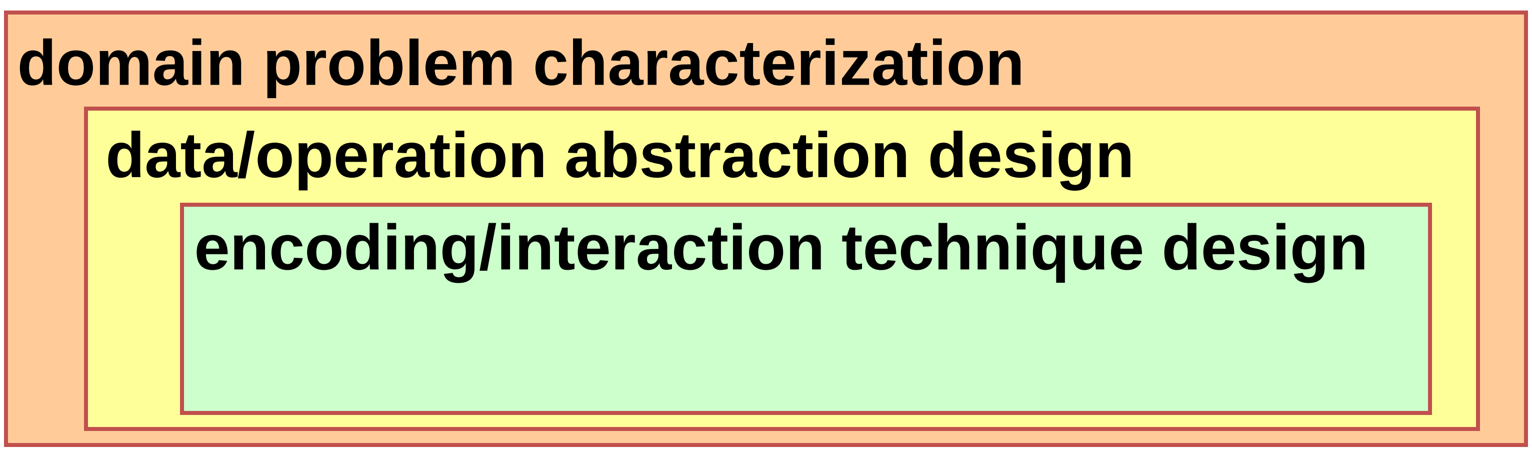
Step 2: Encode Into A Visualization

Visual encoding

- how to best show the data (also pay tribute to aesthetics)
- bar/pie/line charts, parallel coordinates, MDS plot, scatterplot, tree map, network, etc.

Interaction design

- how to best support the intent a user may have
- select, navigate, order, brush, …



Match Visualizations to Tasks

check out this site

Match Visualizations to Tasks

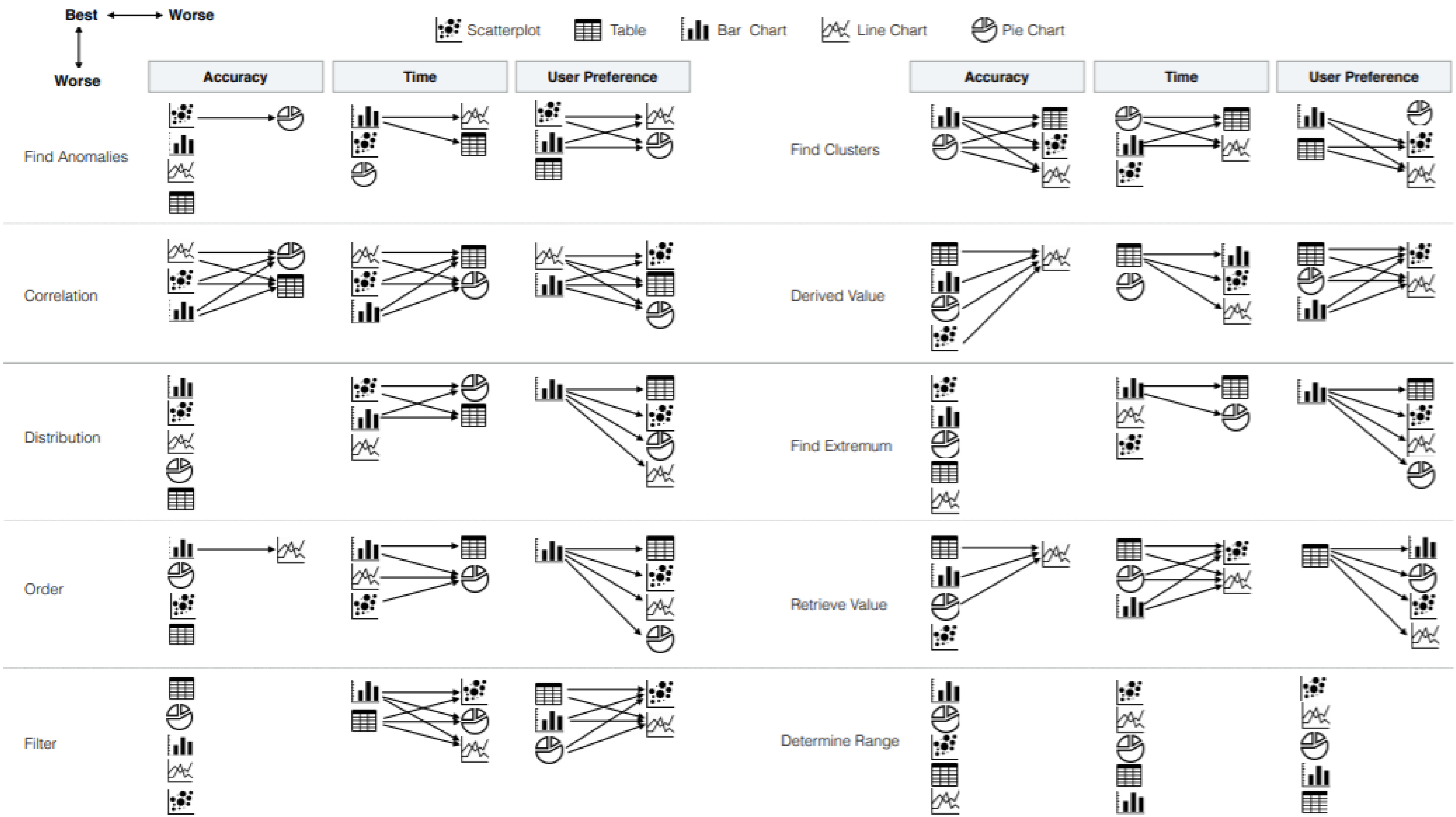
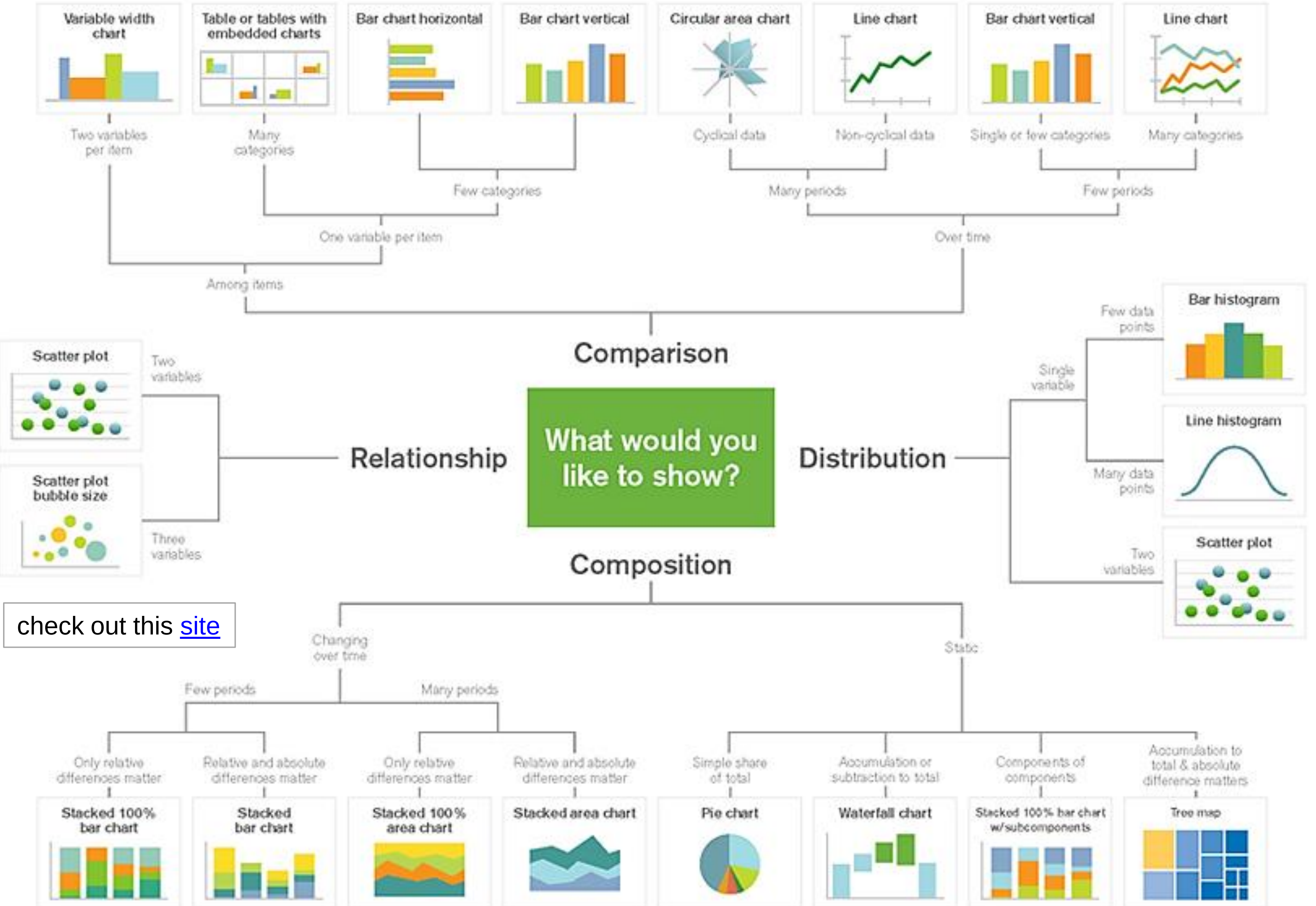


Fig. 3. Pairwise relation between visualization types across tasks and performance metrics. Arrows show that the source is significantly better than the target.

Saket, B. et al. “Task-Based Effectiveness of Basic Visualizations. *IEEE TVCG* , *2019*

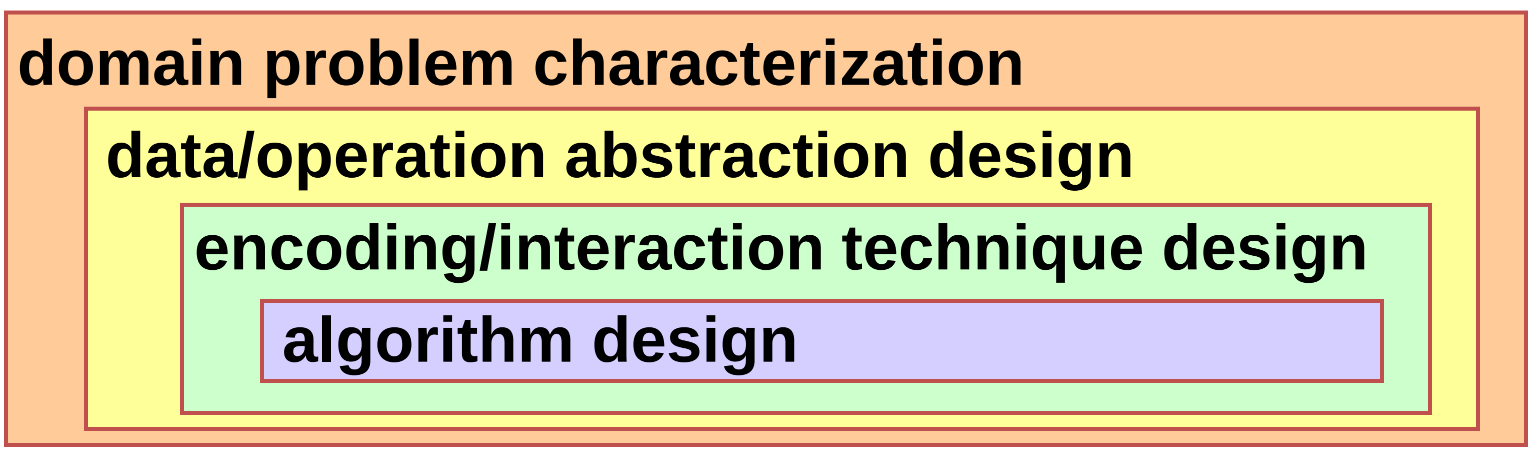


check out this site

Step 4: Design an Algorithm

Well-studied computer science problem

- create efficient algorithms
- should support human interaction
- else it would not comply with key principle of visual analytics



Application Example

Let use the causality analyzer framework just presented

- use the car design example

Domain problem characterization

- how to design a faster car without elevating gas consumption

Data/operation abstraction design

- determine how the different car parameters depend on one another
- collect data of different car models and compute a causal network

Encoding/interaction technique design

- draw graph where parameters are nodes and causal links are edges
- provide interactions that allows users to test causal links and compute a score

Algorithm design

- Partial correlation followed by causal inferencing/conditioning
- Bayesian Information Criterion (BIC) to model Occam's Razor

Another Application Example

How the iPhone came about

- domain problem characterization
- data/operation abstraction design
- encoding/interaction technique design
- algorithm design

June 29, 2007

Gauge Success

threat: wrong problem

validate: observe and interview target users

threat: bad data/operation abstraction

threat: ineffective encoding/interaction technique

validate: justify encoding/interaction design

threat: slow algorithm

validate: analyze computational complexity

implement system

validate: measure system time/memory

validate: qualitative/quantitative result image analysis

[test on any users, informal usability study]

validate: lab study, measure human time/errors for operation

validate: test on target users, collect anecdotal evidence of utility

validate: field study, document human usage of deployed system

validate: observe adoption rates

Gauge Success

Validate along the way and refine

- formative user study

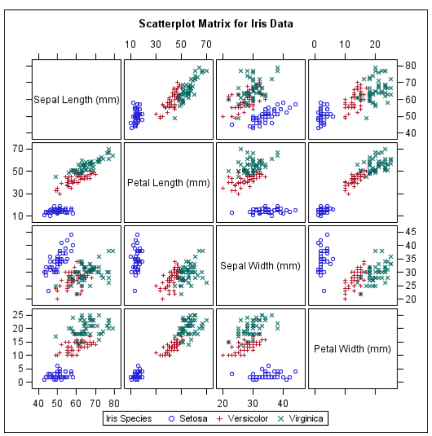
Extend to general user studies of the final design

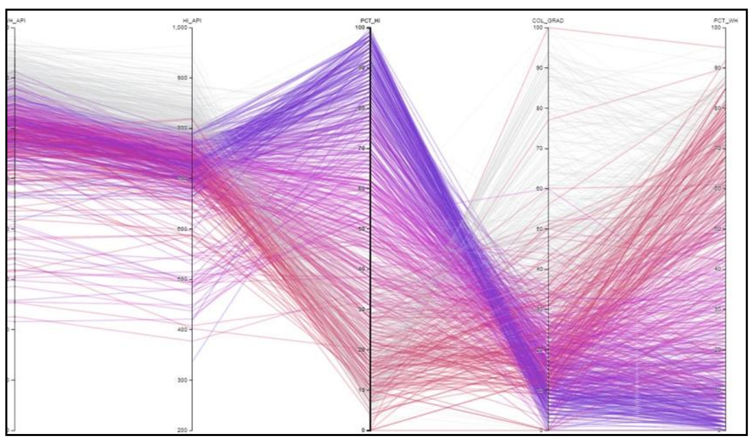
- summative user study
- laboratory study
- smaller number of subjects but can use ‘speak aloud’ protocol
- crowd-sourced via internet
- potentially greater number of subjects to yield better statistics but can be superficial

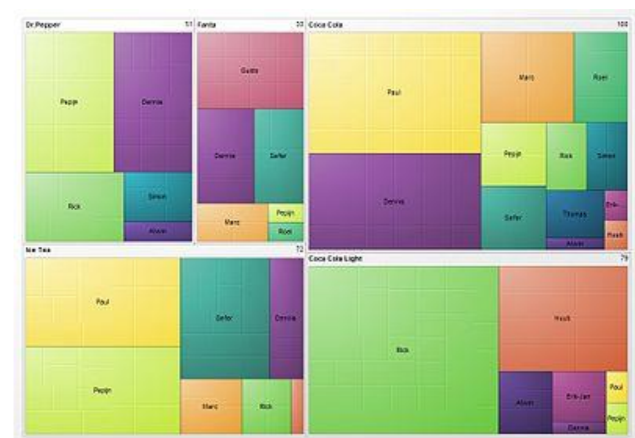
Let’s discuss evaluation studies next

Suppose…

- You boss asks you to come up with a visualization that can show 4 variables
- This reminds you of the great times at CSE 564
- You also remember these three visualizations







Which One Will You Implement?

Let's Ask

- Your best friend
  - but will he/she be an unbiased judge?
- Ask more people

Testing with Users

- You will need
  - implementations
  - some users
  - a few tasks they can solve
- Ask each user to
  - find a certain relationship in the data
  - find certain data elements
  - and so on

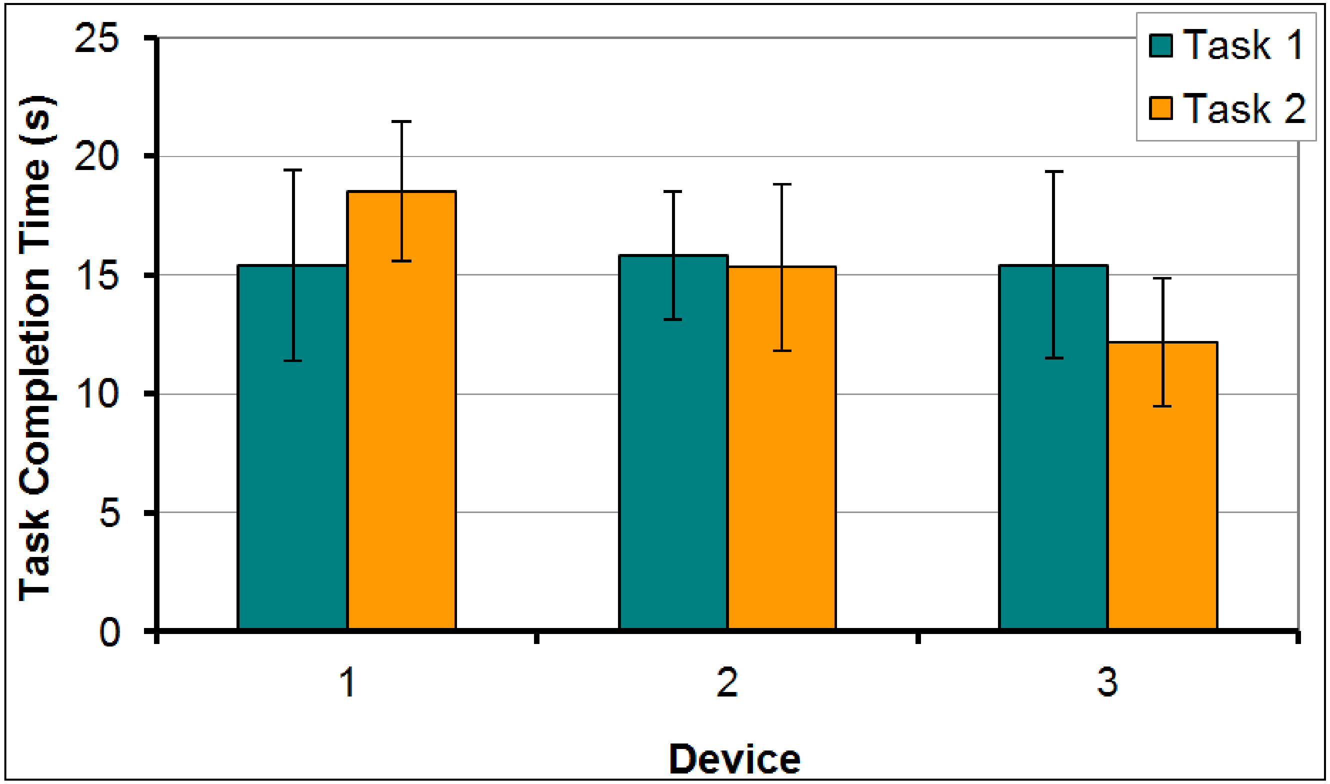
- Measure time and accuracy
- Do this for each of the three visualizations

You Get a Result Like This

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

You Get a Result Like This

- Which visualization is best (1, 2, or 3)?



Next Some Basics

Standard Deviation

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

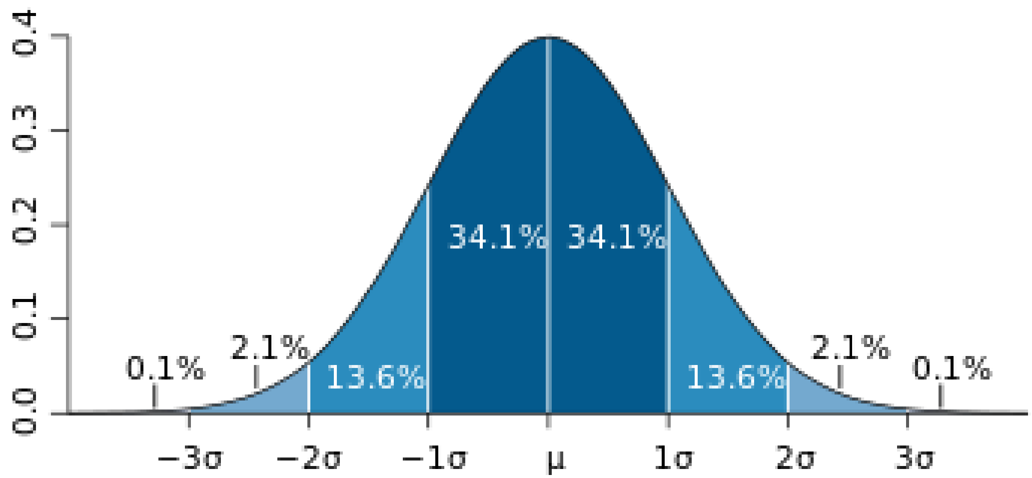
σ = standard deviation

$\sum$ = sum of

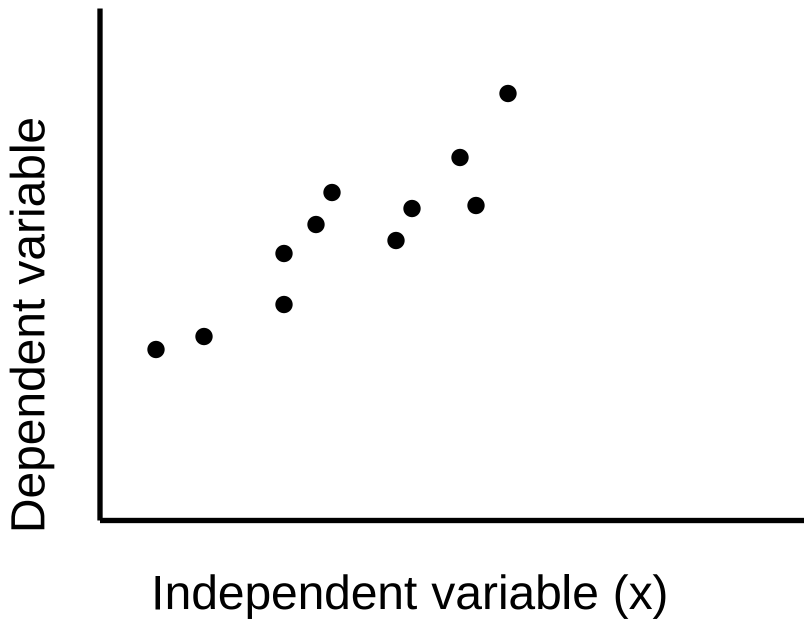
x = each value in the data set

$\bar{x}$ = mean of all values in the data set

n = number of value in the data set



Regression

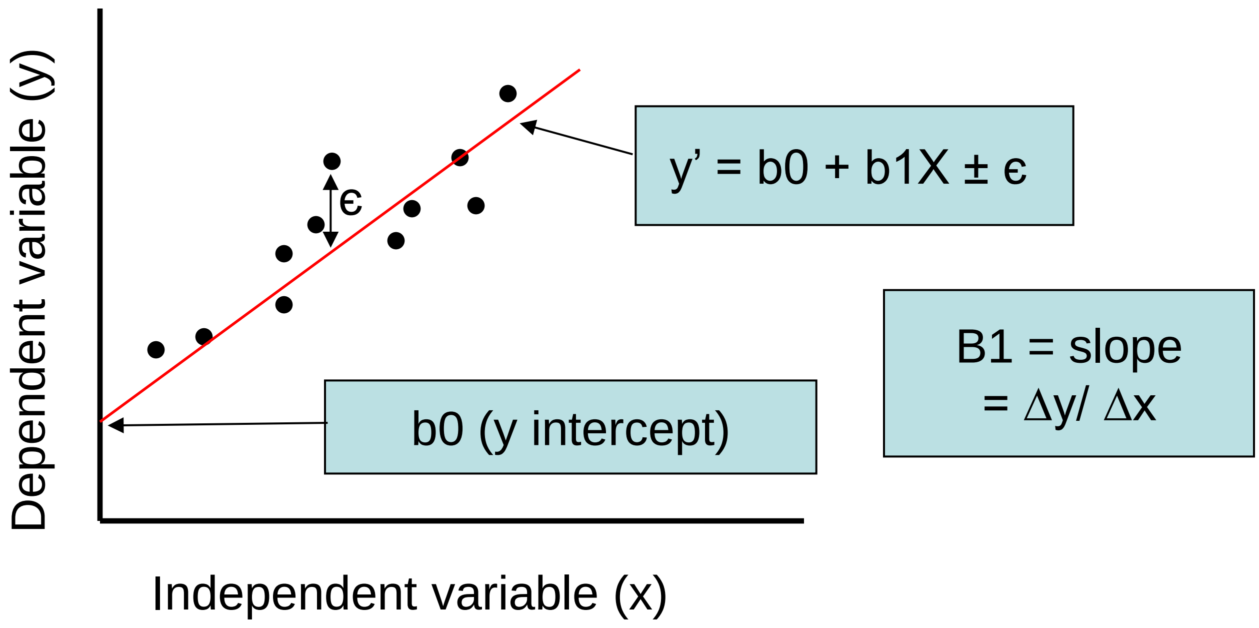


Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

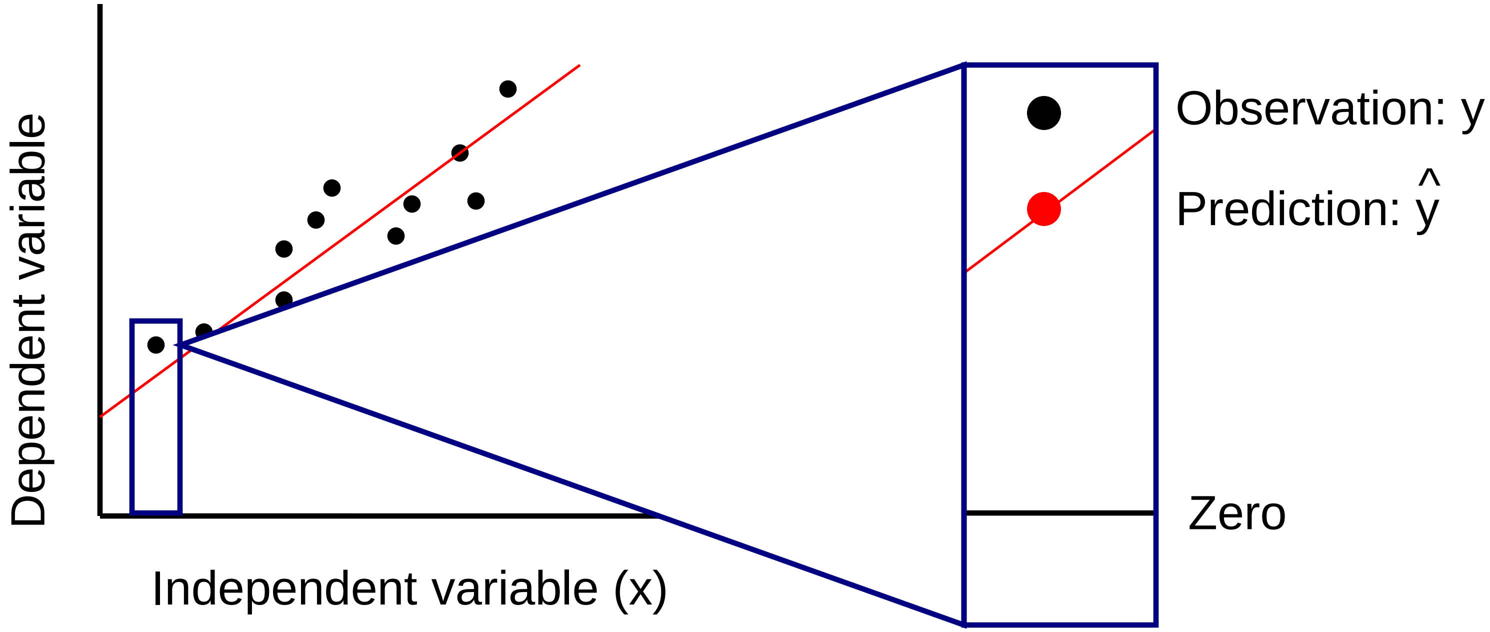
Simple Linear Regression



The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

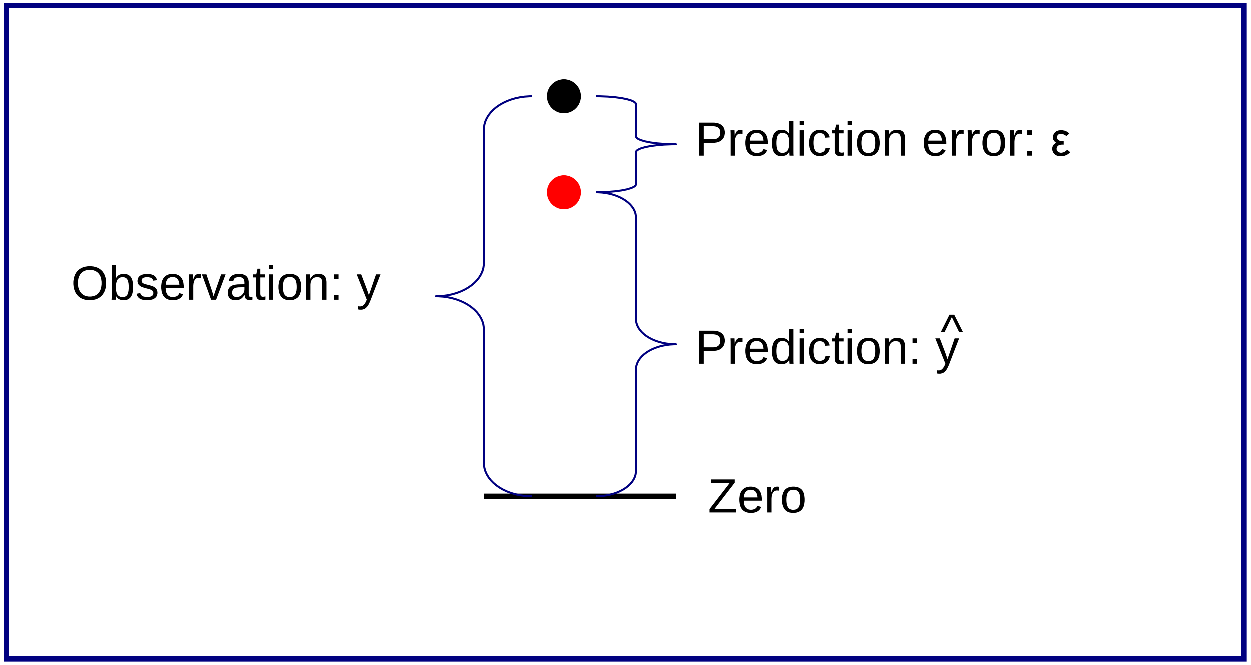
Simple Linear Regression



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by $\hat{y}$.

Simple Linear Regression

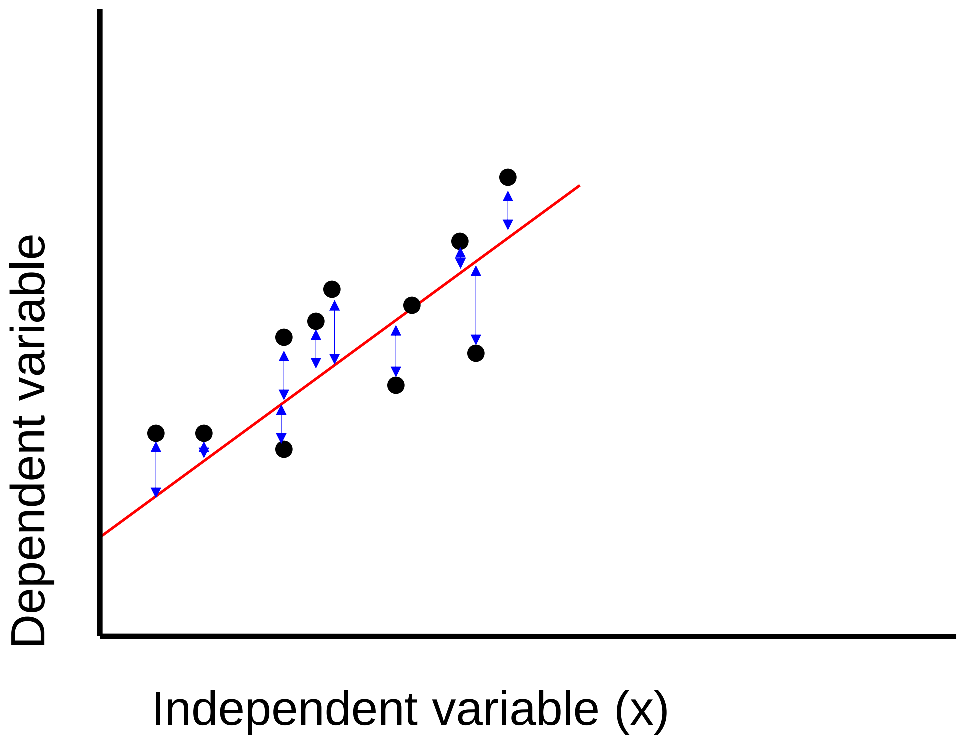


For each observation, the variation can be described as:

$$\mathbf{y = \hat{y} + \varepsilon}$$

Actual = Explained + Error

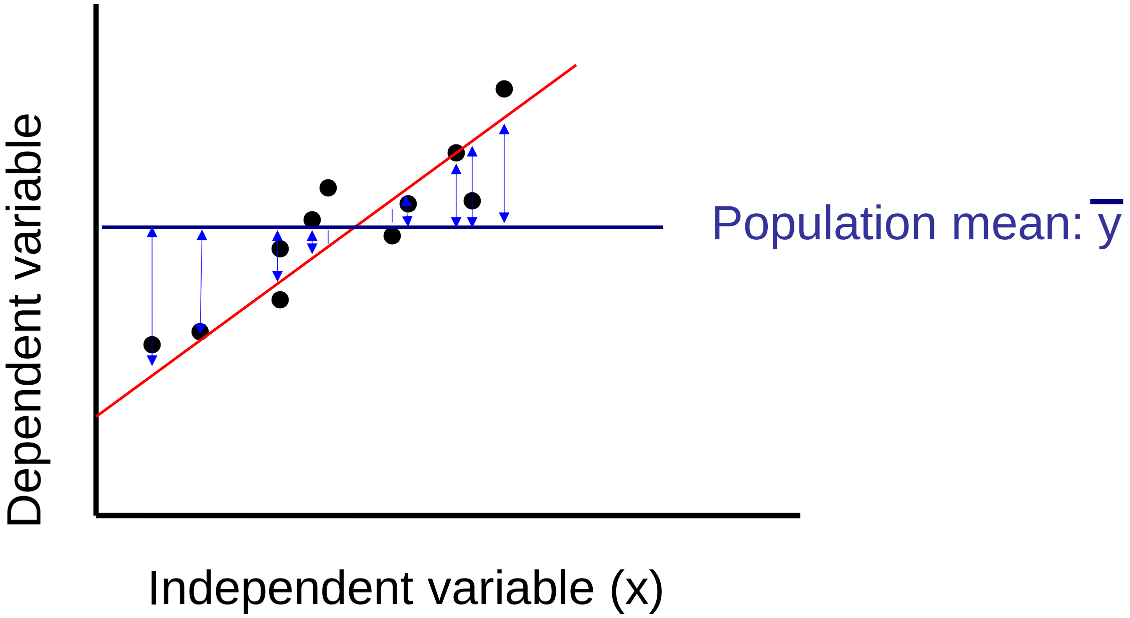
Regression



A least squares regression selects the line with the lowest total sum of squared prediction errors.

This value is called the Sum of Squares of Error, or SSE.

Calculating SSR



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Regression Formulas

The Total Sum of Squares (SST) is equal to SSR + SSE.

Mathematically,

SSR = $\sum (\hat{y} - \bar{y})^2$ **(measure of explained variation)**

SSE = $\sum (y - \hat{y})^2$ **(measure of unexplained variation)**

SST = SSR + SSE = $\sum (y - \bar{y})^2$ **(measure of total variation in y)**

remaining slides courtesy of Scott MacKenzie (York University)
“Human-Computer Interaction: An Empirical Research Perspective”

What is Hypothesis Testing?

- … the use of statistical procedures to answer research questions
- Typical research question (generic):

> Is the time to complete a task less using Method A than using Method B?

- For hypothesis testing, research questions are statements:

> There is no difference in the mean time to complete a task using Method A vs. Method B.

- This is the *null hypothesis* (assumption of "no difference")
- Statistical procedures seek to reject or accept the null hypothesis (details to follow)

Analysis of Variance

- The *analysis of variance* (ANOVA) is the most widely used statistical test for hypothesis testing in factorial experiments
- Goal → determine if an independent variable has a significant effect on a dependent variable
- Remember, an independent variable has at least two levels (test conditions)
- Goal (put another way) → determine if the test conditions yield different outcomes on the dependent variable (e.g., one of the test conditions is faster/slower than the other)

Why Analyze the Variance?

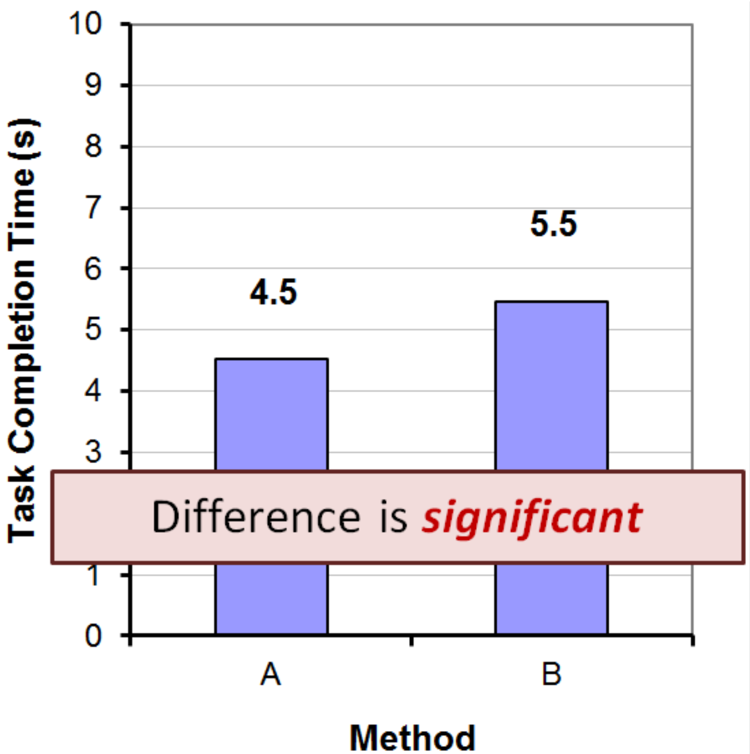
- Seems odd that we analyse the variance when the research question is concerned with the overall means:

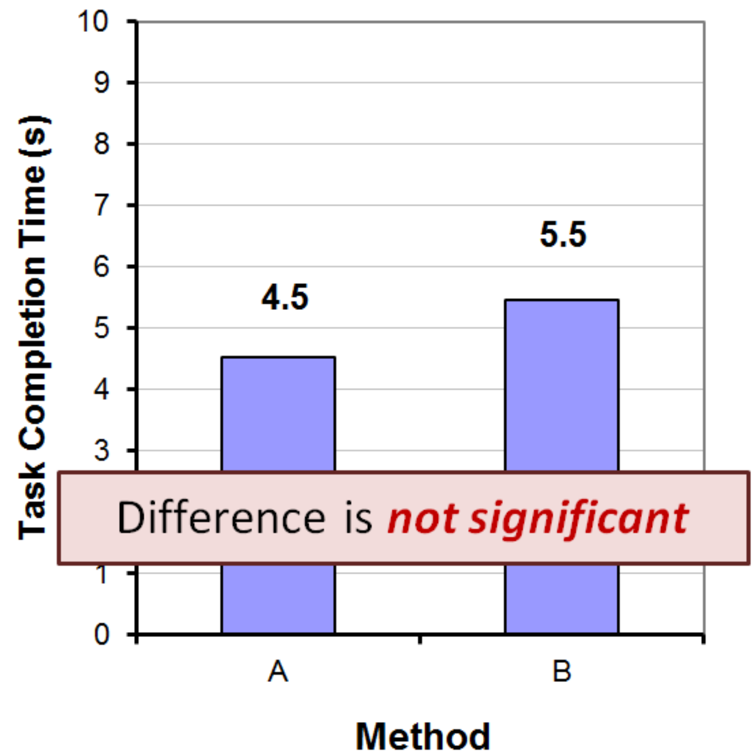
> Is the time to complete a task less using Method A than using Method B?

- Let's explain through two simple examples (next slide)

Example #1

Example #2





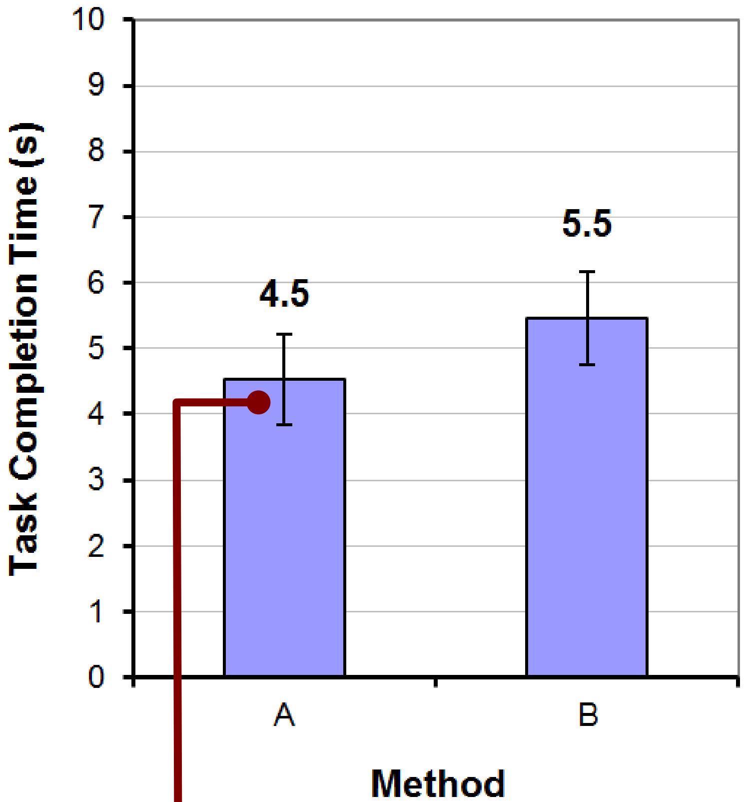
"Significant" implies that in all likelihood the difference observed is due to the test conditions (Method A vs. Method B).

"Not significant" implies that the difference observed is likely due to chance.

File: `06-AnovaDemo.xlsx`

Example #1 - Details

Note: Within-subjects design



Participant	Method	
	A	B
1	5.3	5.7
2	3.6	4.8
3	5.2	5.1
4	3.6	4.5
5	4.6	6.0
6	4.1	6.8
7	4.0	6.0
8	4.8	4.6
9	5.2	5.5
10	5.1	5.6
Mean	4.5	5.5
SD	0.68	0.72

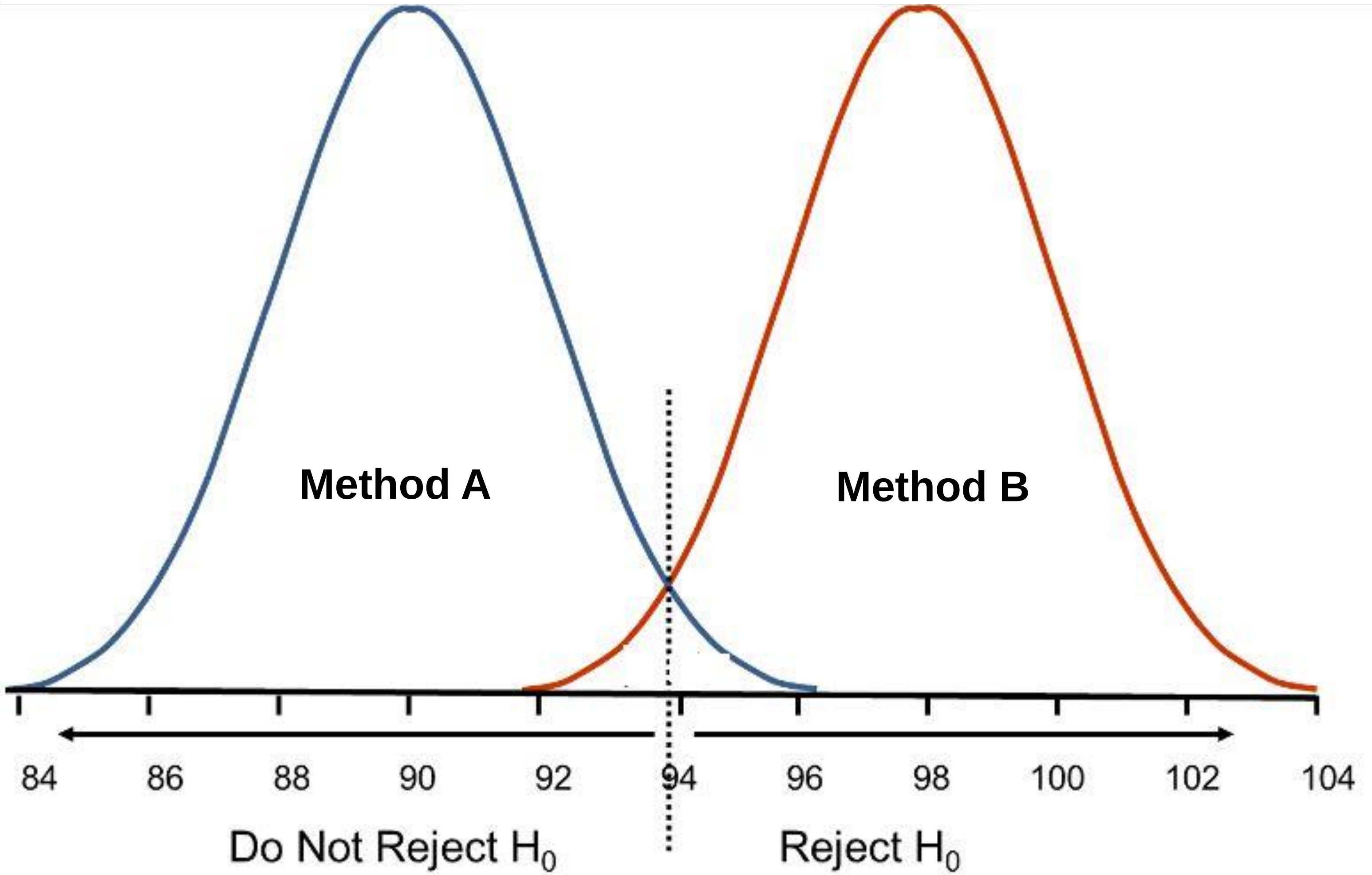
Error bars show
±1 standard deviation

Note: *SD* is the square root of the variance

Make Sure to Randomize

- Eliminate any effect than the one you're after
- Randomize the order in which the subjects run method A and B
  - else may get learning effects of the overall problem
  - method B may turn out better just because users learnt about the problem with method A
- Randomize the data sets or tasks they are asked to use when running method A and B
  - one dataset may be easier than the other
  - method B may turn out better just because the data or tasks was easier

Reject or Not Reject – That's the Question



Example #1 – ANOVA[1]

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

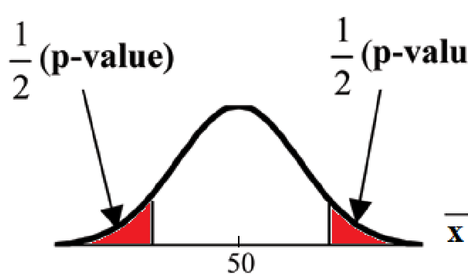
Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 9.80$, $p < .05$

Thresholds for "p"

- .05
- .01
- .005
- .001
- .0005
- .0001



[1] ANOVA table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

Example #1 – ANOVA[1]

MS=SS/df

MS between/within
here: 4.232/0.432 = 9.796

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	5.080	.564				
Method	1	4.232	4.232	9.796	.0121	9.796	.804
Method * Subject	9	3.888	.432				

SS *between* method groups (difference of average treatment effect across groups)

SS *within* method groups (variation of subjects w/r to each treatment mean)

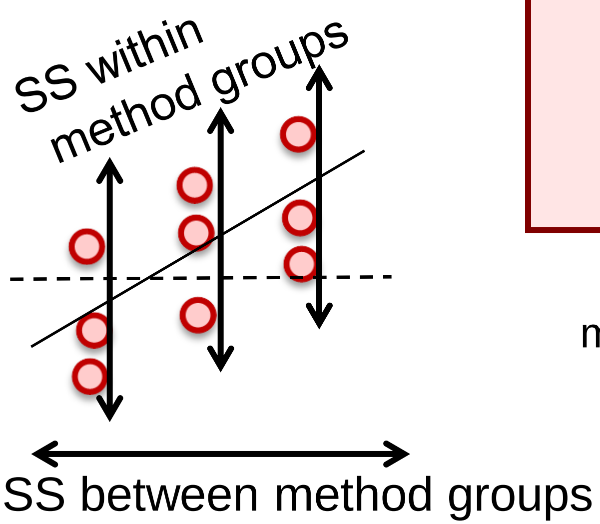
Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 9.80$, $p < .05$

Thresholds for “p”

- .05
- .01
- .005
- .001
- .0005
- .0001



more explanation, see here

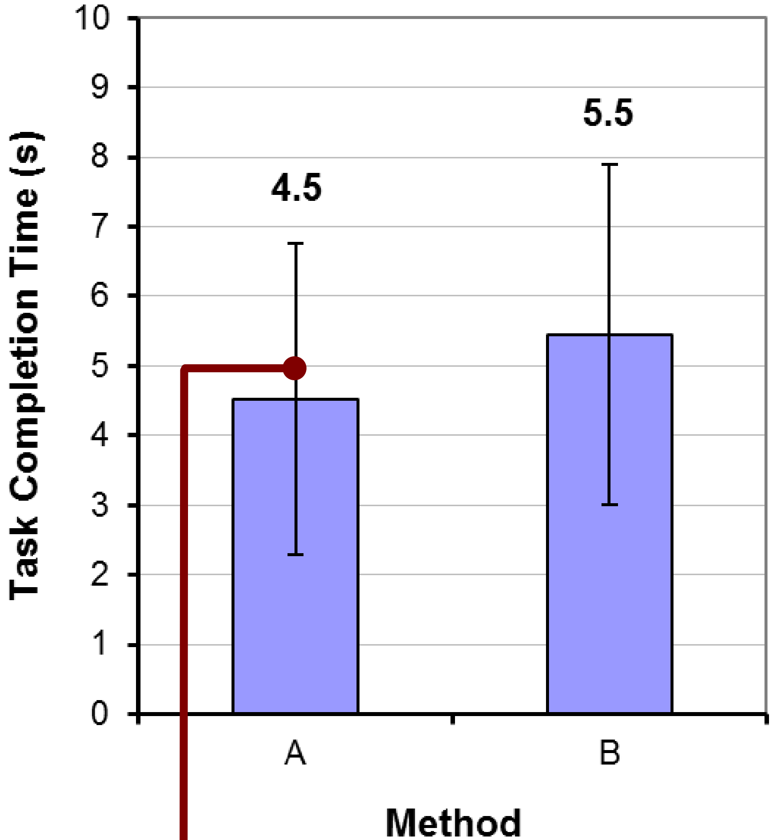
[1] ANOVA table created by *StatView* (now marketed as *JMP*, a product of SAS; www.sas.com)

How to Report an *F*-statistic

The mean task completion time for Method A was 4.5 s. This was 20.1% less than the mean of 5.5 s observed for Method B. The difference was statistically significant ($F_{1,9} = 9.80$, $p < .05$).

- Notice in the parentheses
  - Uppercase for *F*
  - Lowercase for *p*
  - Italics for *F* and *p*
  - Space both sides of equal sign
  - Space after comma
  - Space on both sides of less-than sign
  - Degrees of freedom are subscript, plain, smaller font
  - Three significant figures for *F* statistic
  - No zero before the decimal point in the *p* statistic (except in Europe)

Example #2 - Details



Participant	Method	
	A	B
1	2.4	6.9
2	2.7	7.2
3	3.4	2.6
4	6.1	1.8
5	6.4	7.8
6	5.4	9.2
7	7.9	4.4
8	1.2	6.6
9	3.0	4.8
10	6.6	3.1
Mean	4.5	5.5
SD	2.23	2.45

Error bars show
±1 standard deviation

Example #2 – ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	9	37.372	4.152				
Method	1	4.324	4.324	.626	.4491	.626	.107
Method * Subject	9	62.140	6.904				

Probability of obtaining the observed data if the null hypothesis is true

Reported as…

$F_{1,9} = 0.626$, ns

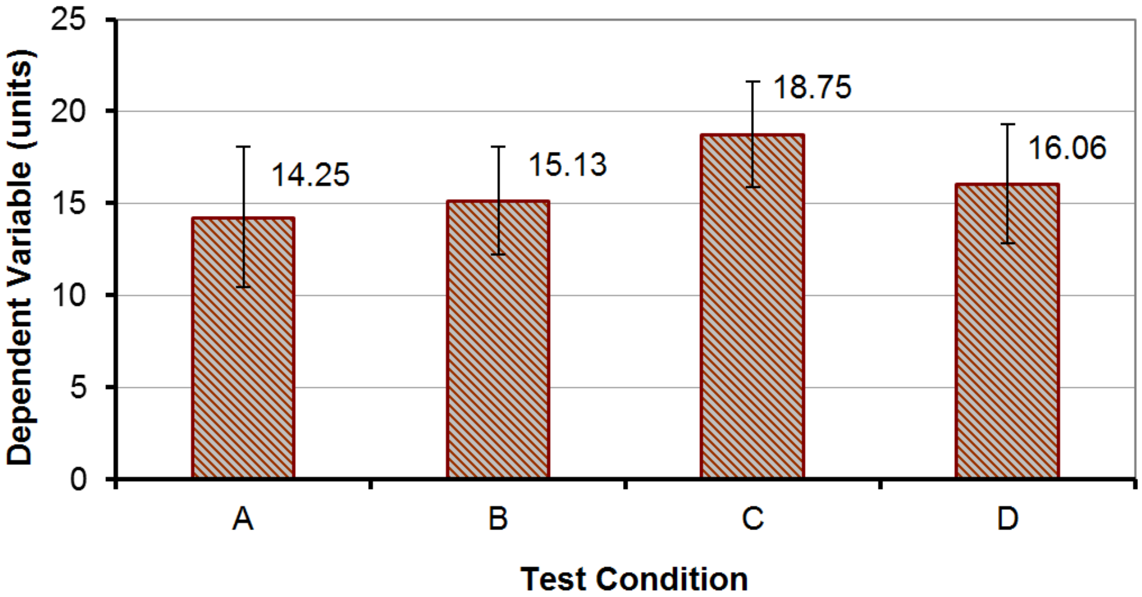
Note: For non-significant effects, use "ns" if $F < 1.0$, or "$p > .05$" if $F > 1.0$.

Example #2 - Reporting

The mean task completion times were 4.5 s for Method A and 5.5 s for Method B. As there was substantial variation in the observations across participants, the difference was not statistically significant as revealed in an analysis of variance ($F_{1,9}$ = 0.626, ns).

More Than Two Test Conditions

Participant	Test Condition			
	A	B	C	D
1	11	11	21	16
2	18	11	22	15
3	17	10	18	13
4	19	15	21	20
5	13	17	23	10
6	10	15	15	20
7	14	14	15	13
8	13	14	19	18
9	19	18	16	12
10	10	17	21	18
11	10	19	22	13
12	16	14	18	20
13	10	20	17	19
14	10	13	21	18
15	20	17	14	18
16	18	17	17	14
Mean	14.25	15.13	18.75	16.06
SD	3.84	2.94	2.89	3.23



ANOVA

ANOVA Table for Dependent Variable (units)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	15	81.109	5.407				
Test Condition	3	182.172	60.724	4.954	.0047	14.862	.896
Test Condition * Subject	45	551.578	12.257				

- There was a significant effect of Test Condition on the dependent variable ($F_{3,45} = 4.95$, $p < .005$)
- Degrees of freedom
  - If n is the number of test conditions and m is the number of participants, the degrees of freedom are…
  - Effect → $(n - 1)$
  - Residual → $(n - 1)(m - 1)$
  - Note: single-factor, within-subjects design

Post Hoc Comparisons Tests

- A significant *F*-test means that at least one of the test conditions differed significantly from one other test condition
- Does not indicate which test conditions differed significantly from one another
- To determine which pairs differ significantly, a post hoc comparisons tests is used
- Examples:
  - Fisher PLSD, Bonferroni/Dunn, Dunnett, Tukey/Kramer, Games/Howell, Student-Newman-Keuls, orthogonal contrasts, Scheffé
- Scheffé test on next slide

Scheffé Post Hoc Comparisons

Scheffe for Dependent Variable (units)
Effect: Test Condition
Significance Level: 5 %

	Mean Diff.	Crit. Diff.	P-Value	
A, B	-.875	3.302	.9003	
A, C	-4.500	3.302	.0032	S
A, D	-1.813	3.302	.4822	
B, C	-3.625	3.302	.0256	S
B, D	-.938	3.302	.8806	
C, D	2.688	3.302	.1520	

- Test conditions A:C and B:C differ significantly (see chart three slides back)

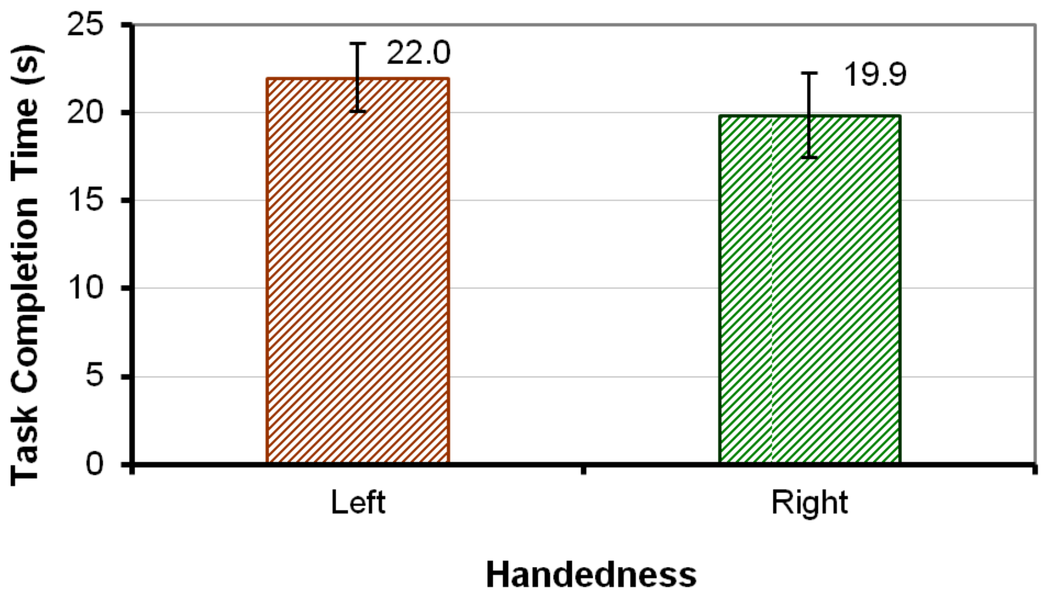
Between-subjects Designs

- Research question:
  - *Do left-handed users and right-handed users differ in the time to complete an interaction task?*
- The independent variable (handedness) must be assigned between-subjects
- Example data set →

Participant	Task Completion Time (s)	Handedness
1	23	L
2	19	L
3	22	L
4	21	L
5	23	L
6	20	L
7	25	L
8	23	L
9	17	R
10	19	R
11	16	R
12	21	R
13	23	R
14	20	R
15	22	R
16	21	R
Mean	20.9	
SD	2.38	

Summary Data and Chart

Handedness	Task Completion Time (s)	
	Mean	*SD*
Left	22.0	1.93
Right	19.9	2.42



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Handedness	1	18.063	18.063	3.781	.0722	3.781	.429
Residual	14	66.875	4.777				

- The difference was not statistically significant ($F_{1,14}$ = 3.78, p > .05)
- Degrees of freedom:
  - Effect → (n – 1)
  - Residual → (m – n)
  - Note: single-factor, between-subjects design

Two-way ANOVA

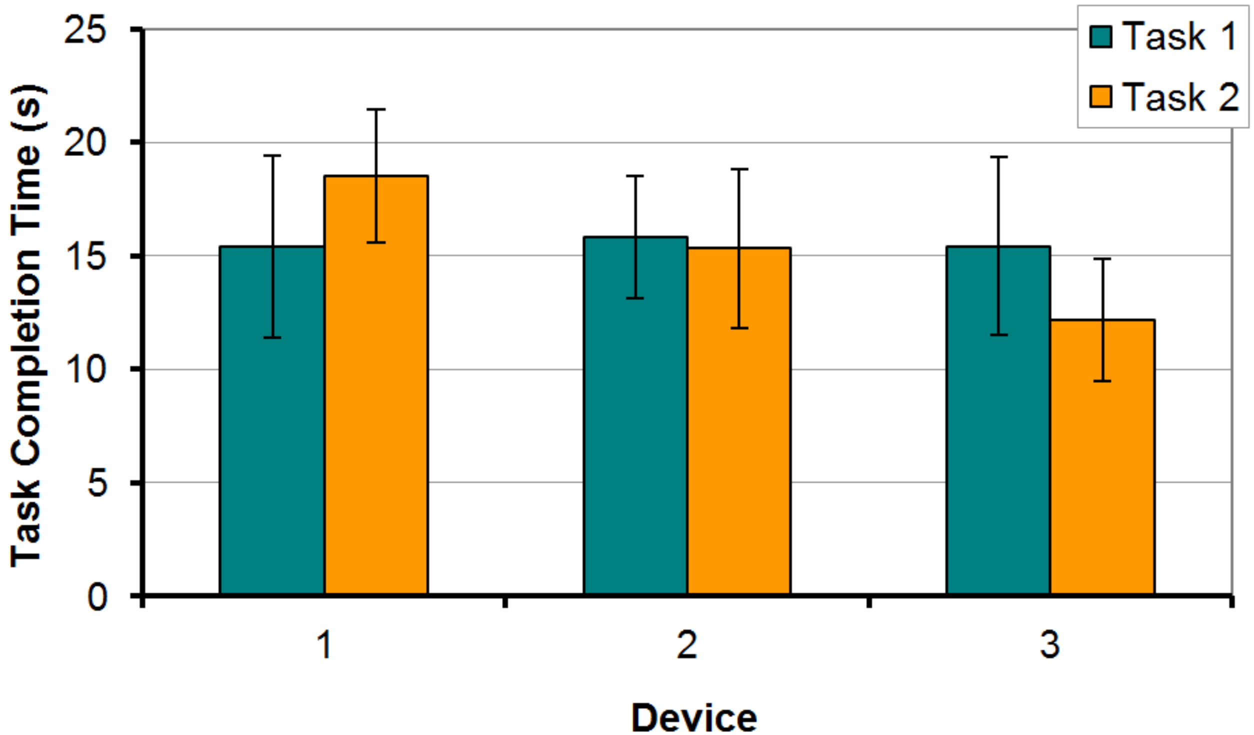
- An experiment with two independent variables is a *two-way design*
- ANOVA tests for
  - Two main effects + one interaction effect
- Example
  - Independent variables
    - Device → D1, D2, D3 (e.g., mouse, stylus, touchpad)
    - Task → T1, T2 (e.g., point-select, drag-select)
  - Dependent variable
    - Task completion time (or something, this isn't important here)
  - Both IVs assigned within-subjects
  - Participants: 12
  - Data set (next slide)

Data Set

Participant	Device 1		Device 2		Device 3	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
1	11	18	15	13	20	14
2	10	14	17	15	11	13
3	10	23	13	20	20	16
4	18	18	11	12	11	10
5	20	21	19	14	19	8
6	14	21	20	11	17	13
7	14	16	15	20	16	12
8	20	21	18	20	14	12
9	14	15	13	17	16	14
10	20	15	18	10	11	16
11	14	20	15	16	10	9
12	20	20	16	16	20	9
Mean	15.4	18.5	15.8	15.3	15.4	12.2
SD	4.01	2.94	2.69	3.50	3.92	2.69

Summary Data and Chart

	Task 1	Task 2	*Mean*
Device 1	15.4	18.5	17.0
Device 2	15.8	15.3	15.6
Device 3	15.4	12.2	13.8
Mean	15.6	15.3	15.4



ANOVA

ANOVA Table for Task Completion Time (s)

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Subject	11	134.778	12.253				
Device	2	121.028	60.514	5.865	.0091	11.731	.831
Device * Subject	22	226.972	10.317				
Task	1	.889	.889	.076	.7875	.076	.057
Task * Subject	11	128.111	11.646				
Device * Task	2	121.028	60.514	5.435	.0121	10.869	.798
Device * Task * Subject	22	244.972	11.135				

Can you pull the relevant statistics from this chart and craft statements indicating the outcome of the ANOVA?

ANOVA - Reporting

The grand mean for task completion time was 15.4 seconds. Device 3 was the fastest at 13.8 seconds, while device 1 was the slowest at 17.0 seconds. The main effect of device on task completion time was statistically significant ($F_{2,22} = 5.865$, $p < .01$). The task effect was modest, however. Task completion time was 15.6 seconds for task 1. Task 2 was slightly faster at 15.3 seconds; however, the difference was not statistically significant ($F_{1,11} = 0.076$, ns). The results by device and task are shown in Figure x. There was a significant Device × Task interaction effect ($F_{2,22} = 5.435$, $p < .05$), which was due solely to the difference between device 1 task 2 and device 3 task 2, as determined by a Scheffé post hoc analysis.

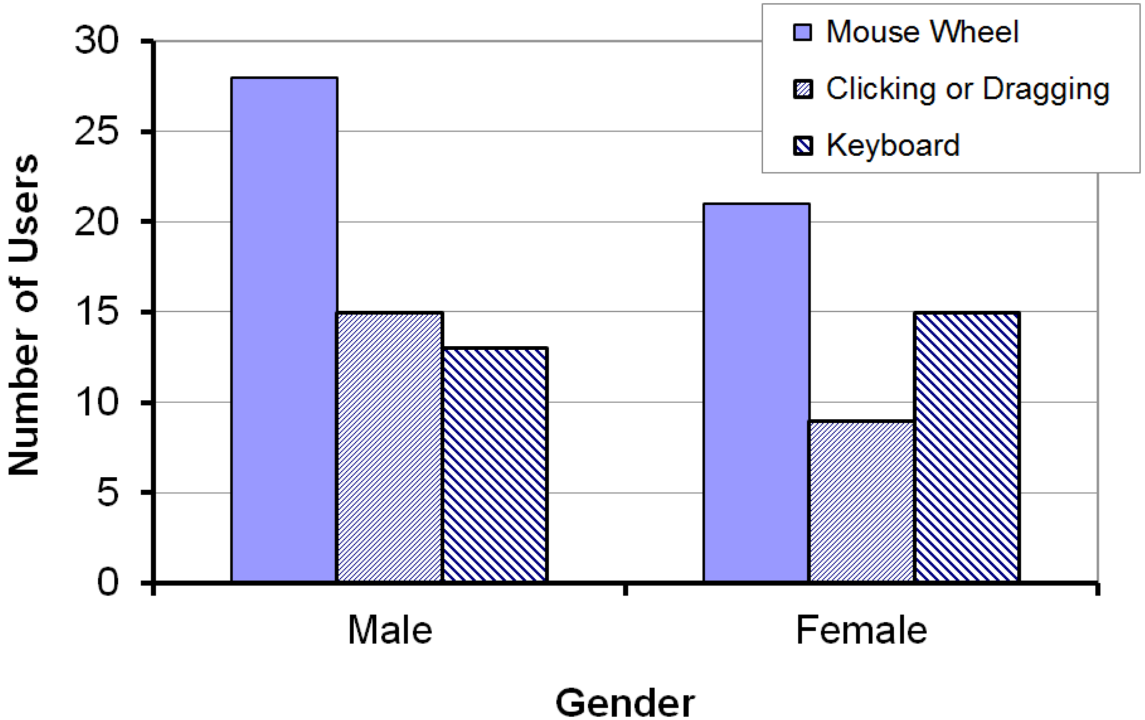
Chi-square Test (Nominal Data)

- A *chi-square test* is used to investigate relationships
- Relationships between categorical, or nominal-scale, variables representing attributes of people, interaction techniques, systems, etc.
- Data organized in a *contingency table* – cross tabulation containing counts (frequency data) for number of observations in each category
- A chi-square test compares the *observed values* against *expected values*
- Expected values assume “no difference”
- Research question:
  - *Do males and females differ in their method of scrolling on desktop systems?* (next slide)

Chi-square – Example #1

Observed Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	28	15	13	56
Female	21	9	15	45
Total	49	24	28	101

MW = mouse wheel
CD = clicking, dragging
KB = keyboard



Chi-square – Example #1

56.0·49.0/101=27.2

Expected Number of Users				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	27.2	13.3	15.5	56.0
Female	21.8	10.7	12.5	45.0
Total	49.0	24.0	28.0	101

(Expected-Observed)²/Expected=(28-27.2)²/27.2

Chi Squares				
Gender	Scrolling Method			Total
	MW	CD	KB	
Male	0.025	0.215	0.411	0.651
Female	0.032	0.268	0.511	0.811
Total	0.057	0.483	0.922	**1.462**

Significant if it exceeds critical value (next slide)

$$\chi^2 = 1.462$$

(See **HCI:ERP** for calculations)

Chi-square Critical Values

- Decide in advance on *alpha* (typically .05)
- Degrees of freedom
  - $df = (r - 1)(c - 1) = (2 - 1)(3 - 1) = 2$
  - r = number of rows, c = number of columns

Significance Threshold (α)	Degrees of Freedom							
	1	2	3	4	5	6	7	8
.1	2.71	4.61	6.25	7.78	9.24	10.65	12.02	13.36
.05	3.84	5.99	7.82	9.49	11.07	12.59	14.07	15.51
.01	6.64	9.21	11.35	13.28	15.09	16.81	18.48	20.09
.001	10.83	13.82	16.27	18.47	20.52	22.46	24.32	26.13

$\chi^2 = 1.462$ (< 5.99 $\therefore$ not significant)