

CS160



USER
INTERFACE
DESIGN

SPRING 2016

DATA ANALYSIS

9 MAR 2016

ERIC PAULOS

www.paulos.net

MIDTERM ON 16 MAR

Midterm is on 16 March during class time 2:30-4pm

In Sibley Auditorium

80 minutes (we will start at 2:40pm and run until 4pm)

closed book & notes

HKN has previous midterm for practice

come early so we can start on time

all you need is a pen or pencil to write with

The midterm will cover all aspects of the course through Wednesday's lecture. This includes, readings, lectures, assignments, section, etc. We may ask design questions, code questions, etc.

There will be a midterm Review in class next Monday 2:30-4pm.

If you find a question ambiguous, document the ambiguity. Indicate the way you interpreted the question in a set of separate sentences next to the question. The questions on the exam are not intended to be ambiguous, but sometimes another meaning is interpreted by the examinee that we did not take into consideration.

If you are registered with the DSP office and have special needs, you should have received email from me about exam accommodations via bCourses. You must contact me by Friday if there is a problem in any accommodations details or accommodations made (or not made) to you.

MIDTERM ATTENDANCE

I know this won't happen but I'm putting it here so it is clear there is policy:

You must attend the midterm

There is not a makeup midterm exam date

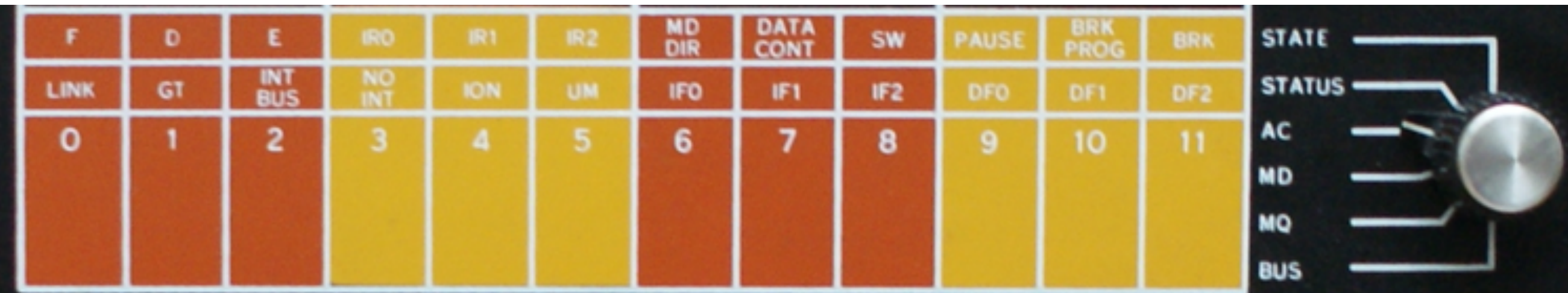
If you have any reason to believe you may miss the midterm (i.e. you have a court appearance, you have difficult travel plans, you have a planned doctors appointment before class that may run over, you have been called to testify before Congress, etc), you must let me know by end of day Friday 11 March. We will not grant excuses for issues that come up after Friday and you will be given a zero for the exam.

If you have a health or medical emergency and unable to make the midterm or decide not to come:

You must tell us immediately via Piazza post to all instructors

You must meet with the instructor as soon as possible to discuss your circumstances.

You will be given, at the complete discretion of the instructor, either a zero for the midterm, an oral exam of up to 2 hours (on a date set by the instructor), or a 3 hour final exam during our scheduled final exam time covering the material from the entire semester. The grade on this exam will take the place of your midterm. Legitimate health related emergencies (as determined by the instructor) will not be given a zero but will be subject to the other two options at the discretion of the instructor.



MANAGING STUDY PARTICIPANTS

RUN THE EXPERIMENT

Always pilot it first!

Reveals unexpected problems

Can't change experiment design after starting it

Always follow same steps – use a checklist

Get consent from subjects

Debrief subjects afterwards

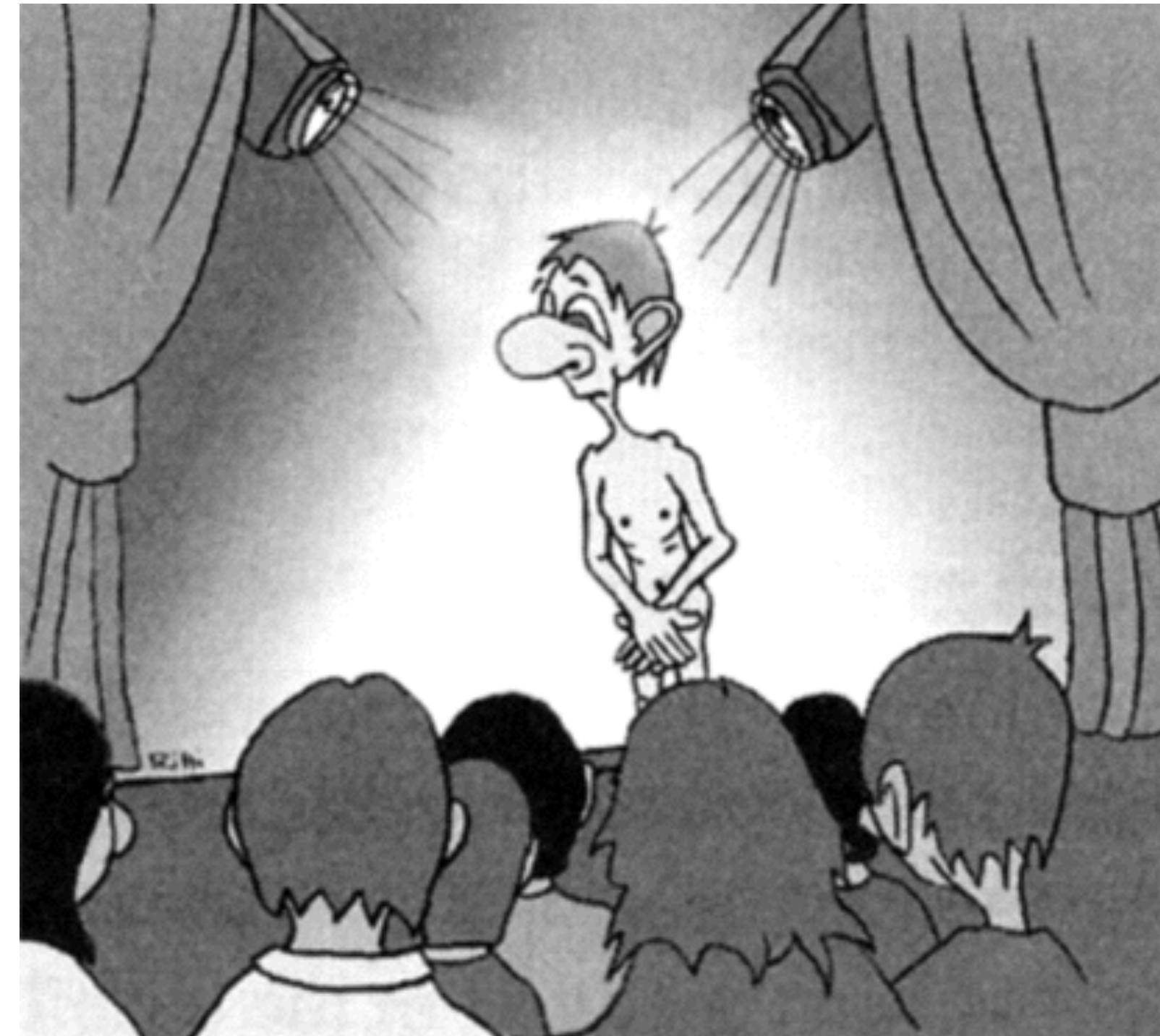
THE PARTICIPANTS' STANDPOINT

Testing is a distressing experience

Pressure to perform

Feeling of inadequacy

Looking like a fool in front of
your peers, your boss, ...



(from "Paper Prototyping" by Snyder)

THE THREE BELMONT PRINCIPLES

Respect for Persons

Have a meaningful consent process: give information, and let prospective subjects freely chose to participate

Beneficence

Minimize the risk of harm to subjects, maximize potential benefits

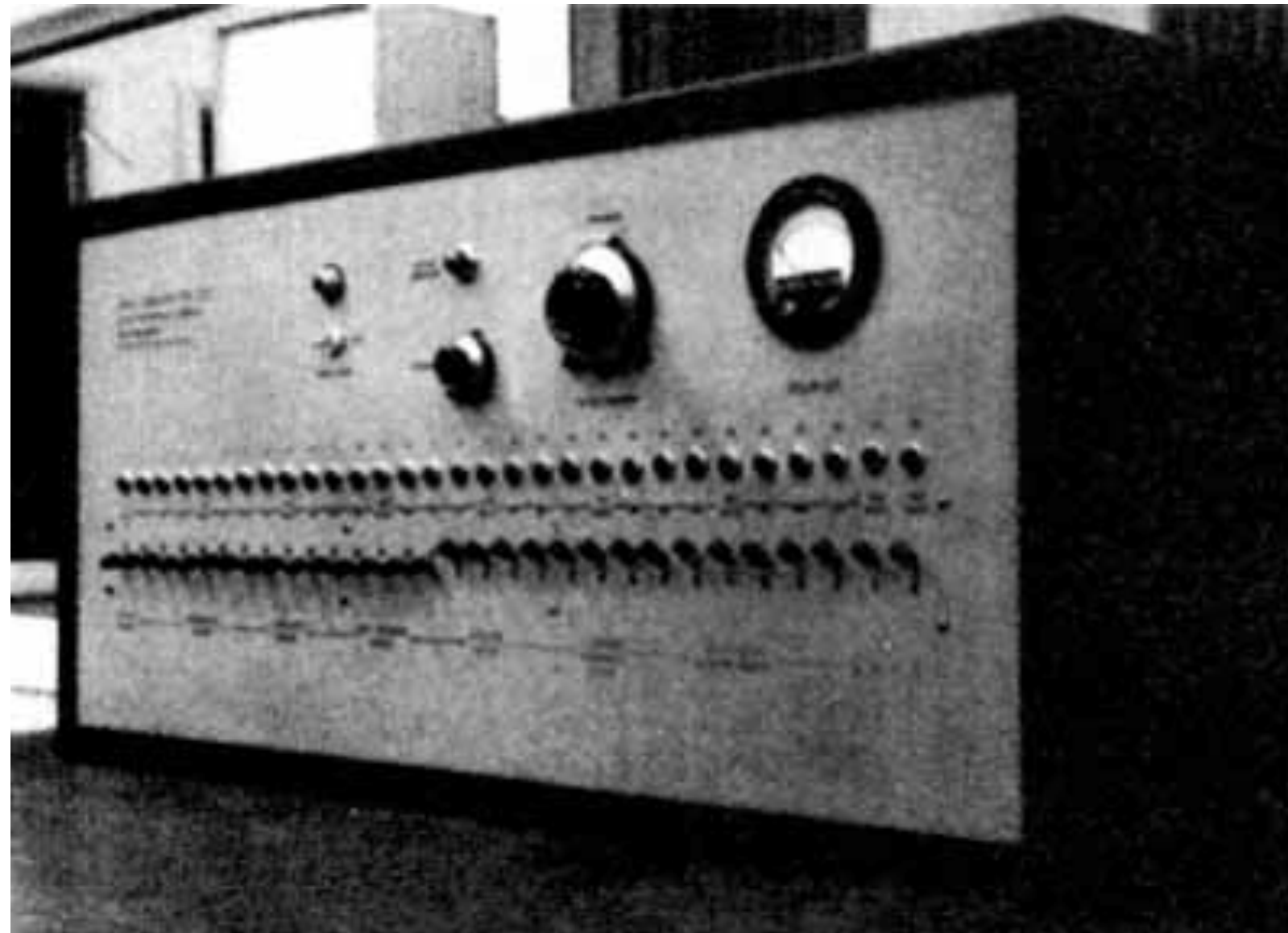
Justice

Use fair procedures to select subjects
(balance burdens & benefits)

To ensure adherence to principles, most schools require Institutional Review Board approval of research involving human subjects.

MILGRAM OBEDIENCE TO AUTHORITY

1961 Experiment by Stanley Milgram



ETHICS: STANFORD PRISON EXPERIMENT

1971 Experiment by Phil Zimbardo at Stanford

24 Participants – half prisoners, half guards (\$15 a day)

Basement of Stanford Psychology building turned into mock prison

Guards given batons, military style uniform, mirror glasses,...

Prisoners wore smocks (no underwear), thong sandals, pantyhose caps



ETHICS: STANFORD PRISON EXPERIMENT

Experiment quickly got out of hand

Prisoners suffered and accepted sadistic treatment

Prison became unsanitary/inhospitable

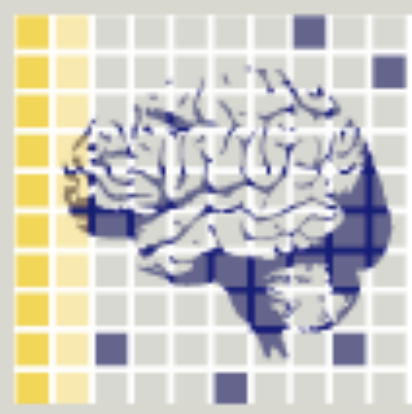
Prisoner riot put down with use of fire extinguishers

Guards volunteered to work extra hours

Zimbardo terminated experiment early

Grad student Christina Maslach objected to experiment

Important to check protocol with ethics review boards



[Home](#) » [Christina MASLACH](#)

Christina MASLACH



Professor

Chair of the Academic Senate

[Contact Information](#)


maslach@berkeley.edu

Office: 3325 Tolman Hall

Office Hours: by appointment

<http://psychology.berkeley.edu/maslach%40socialpsychology.org>

Ph.D., Stanford University

Curriculum Vitae:  [CV 2012.pdf](#)

Research Interests: Social: job burnout and health psychology; individuation and dissent, gender roles

Research Areas: [Social-Personality](#)

ETHICS

Was it useful?

"...that's the most valuable kind of information that you can have and that certainly a society needs it" (Zimbardo)

Was it ethical?

Could we have gathered this knowledge by other means?



ETHICS (MORE RECENTLY)

“In 2001, a faculty member from the business school of a major university designed a study to see how restaurants would respond to complaints from putative customers. As part of the project, the researcher sent letters to restaurants falsely claiming that he and/or his wife had suffered food poisoning that ruined their anniversary celebration. The letters disclaimed any intention of contacting regulatory agencies and stated that the only intent was to convey to the owner what had occurred "in anticipation that you will respond accordingly." Restaurant owners were understandably upset and some employees lost their jobs before it was revealed that the letter was a hoax. ”

ETHICS (EVEN MORE RECENTLY)

In June 2014 researchers from Facebook altered the news feed algorithm for 689,003 users to skew the presence of positive or negative posts. They then tracked subsequent posts from those users by using positive or negative keywords.

“In addition to helping people see and find things that you do and share, we may use the information we receive about you...for internal operations, including troubleshooting, data analysis, testing, research and service improvement.”

Institutions that receive federal funding are required to abide by a federal policy called the “Common Rule,” which protects human experiment subjects by ensuring that they know about the study and that they understand the risks involved. It also requires institutional review boards at universities and hospitals to approve the way subjects of biomedical or behavioral studies are treated.

ETHICS (EVEN MORE RECENTLY)

Lead researcher and Facebook data scientist Adam Kramer took to Facebook to defend the study:

“We felt that it was important to investigate the common worry that seeing friends post positive content leads to people feeling negative or left out. At the same time, we were concerned that exposure to friends' negativity might lead people to avoid visiting Facebook,” Kramer wrote.

He went on to explain that the “actual impact on people” was the minimal needed to conclude that Facebook feeds influenced users' emotions. Though they expected happy news would make people feel sad, they found that people with a little more positive news in their feeds included more happy words in their posts.

“Having written and designed this experiment myself, I can tell you that our goal was never to upset anyone,” he wrote in the post. “I can understand why some people have concerns about it, and my coauthors and I are very sorry for the way the paper described the research and any anxiety it caused. In hindsight, the research benefits of the paper may not have justified all of this anxiety.”

BENEFICENCE: EXAMPLE

MERL DiamondTouch:

User capacitively coupled to table through seating pad.

No danger for normal users, but possibly increased risk for participants with pacemakers.

Inform subjects in consent!



PRIVACY AND CONFIDENTIALITY

Privacy: having control over the extent, timing, and circumstances of sharing oneself with others.

Confidentiality: the treatment of information that an individual has disclosed with the expectation that it will not be divulged

Examples where privacy could be violated or confidentiality may be breached in HCI studies?

TREATING SUBJECTS WITH RESPECT

Follow human subject protocols

Individual test results will be kept confidential

Users can stop the test at any time

Users are aware (and understand) the monitoring technique(s)

Their performance will not have implications on their life

Records will be made anonymous

Use standard informed consent form

Especially for quantitative tests

Be aware of legal requirements

CONDUCTING THE EXPERIMENT

Before the experiment

Have them read and sign the consent form

Explain the goal of the experiment in a way accessible to users

Be careful about the demand characteristic

(Participants biased towards experimenter's hypothesis)

Answer questions

During the experiment

Stay neutral

Never indicate displeasure with users performance

After the experiment

Debrief users (Inform users about the goal of the experiment)

Answer any questions they have

MANAGING SUBJECTS

Don't waste users' time

Use pilot tests to debug experiments, questionnaires, etc...

Have everything ready before users show up

Make users comfortable

Keep a relaxed atmosphere

Allow for breaks

Pace tasks correctly

Stop the test if it becomes too unpleasant

IF YOU WANT TO LEARN MORE...

Online human subjects certification courses:

E.g., <http://phrp.nihtraining.com/users/login.php>

The Belmont Report: Ethical Principles and Guidelines for the protection of human subjects of research

1979 Government report that describes the basic ethical principles that should underly the conduct of research involving human subjects

<http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>

F	D	E	IR0	IR1	IR2	MD DIR	DATA CONT	SW	PAUSE	BRK PROG	BRK
LINK	GT	INT BUS	NO INT	ION	UM	IFO	IF1	IF2	DF0	DF1	DF2
0	1	2	3	4	5	6	7	8	9	10	11

STATE

STATUS

AC

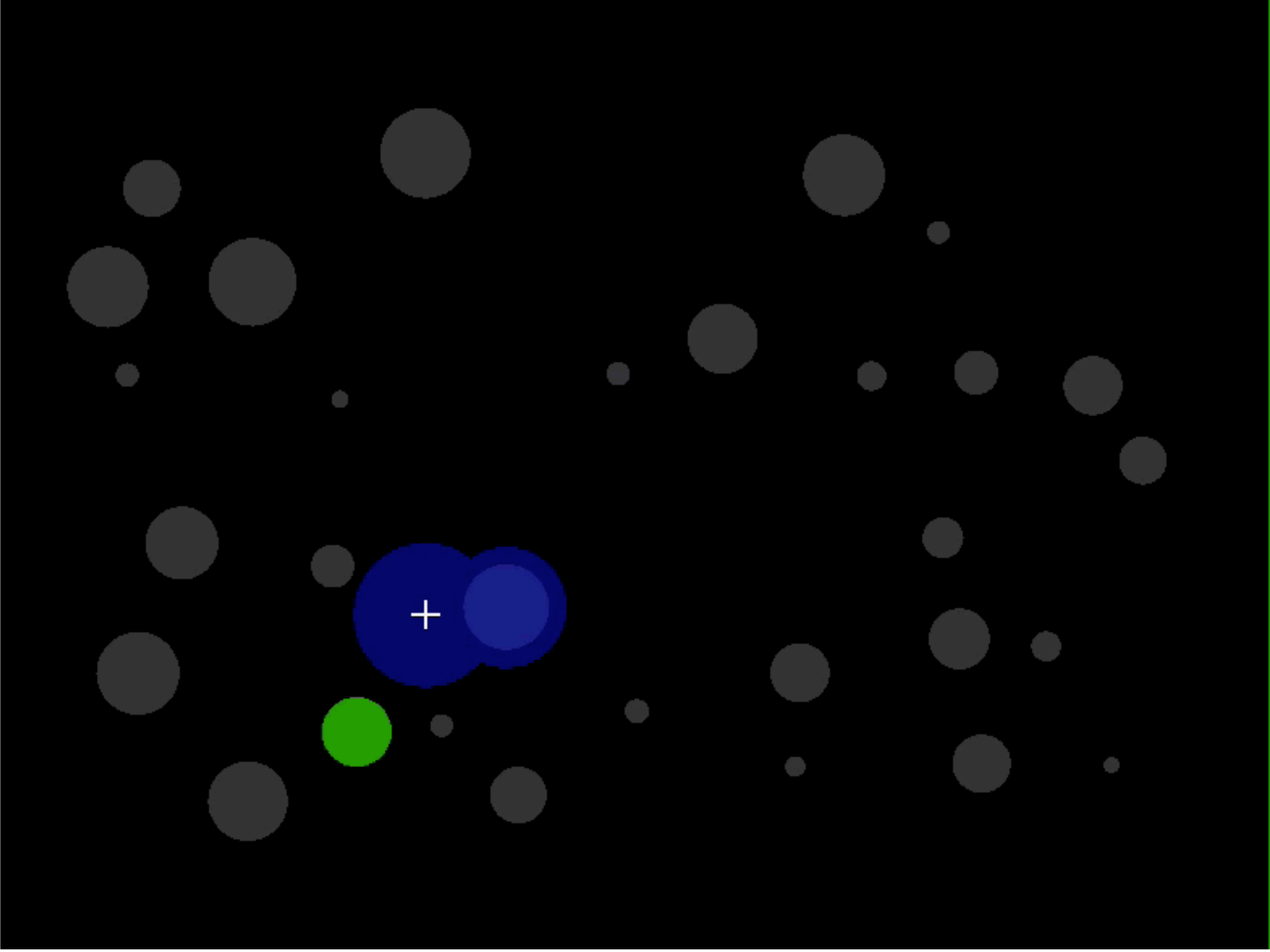
MD

MQ

BUS



DATA ANALYSIS



1362884269	normal	10	1	t1	1146
------------	--------	----	---	----	------

1362884269,normal,10,1,t2,926
1362884269,normal,10,1,t3,924
1362884269,normal,10,1,t4,863
1362884269,normal,10,1,t5,1474
1362884269,normal,10,1,t6,979
1362884269,normal,10,1,t7,944
1362884269,normal,10,1,t8,966
1362884269,normal,10,1,t9,931
1362884269,normal,10,1,t10,926
1362884269,normal,10,1,t11,1047
1362884269,normal,10,1,t12,949
1362884269,normal,10,1,t13,980
1362884269,normal,10,1,t14,996
1362884269,normal,10,1,t15,930
1362884269,normal,10,1,t16,935
1362884269,normal,10,1,t17,797
1362884269,normal,10,1,t18,899
1362884269,normal,10,1,t19,945
1362884269,normal,10,1,t20,945

START BY COUNTING

4140 trials total

normal:

mean time 955.4 ms,
mean errors 1.486

bubble:

mean time 763.9 ms,
mean errors 0.402

START BY COUNTING

54 users completed condition **normal**, size 10
mean time: **1113.25 ms**, mean errors: **1.889**
median time: 1067 ms, median errors: 1

51 users completed condition **normal**, size 30
mean time: **788.33 ms**, mean errors: **1.059**
median time: 754 ms, median errors: 1

52 users completed condition **bubble**, size 10
mean time: **809.96 ms**, mean errors: **0.404**
median time: 766 ms, median errors: 0

50 users completed condition **bubble**, size 30
mean time: **716.01 ms**, mean errors: **0.020**
median time: 692 ms, median errors: 0

DESCRIPTIVE STATISTICS

Continuous data:

Central tendency

mean, median, mode

Dispersion

Range (max-min)

Standard deviation

Shape of distribution

Skew, Kurtosis

Categorical data:

Frequency distributions

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Mean

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Standard Deviation

UNDERSTANDING YOUR DATA

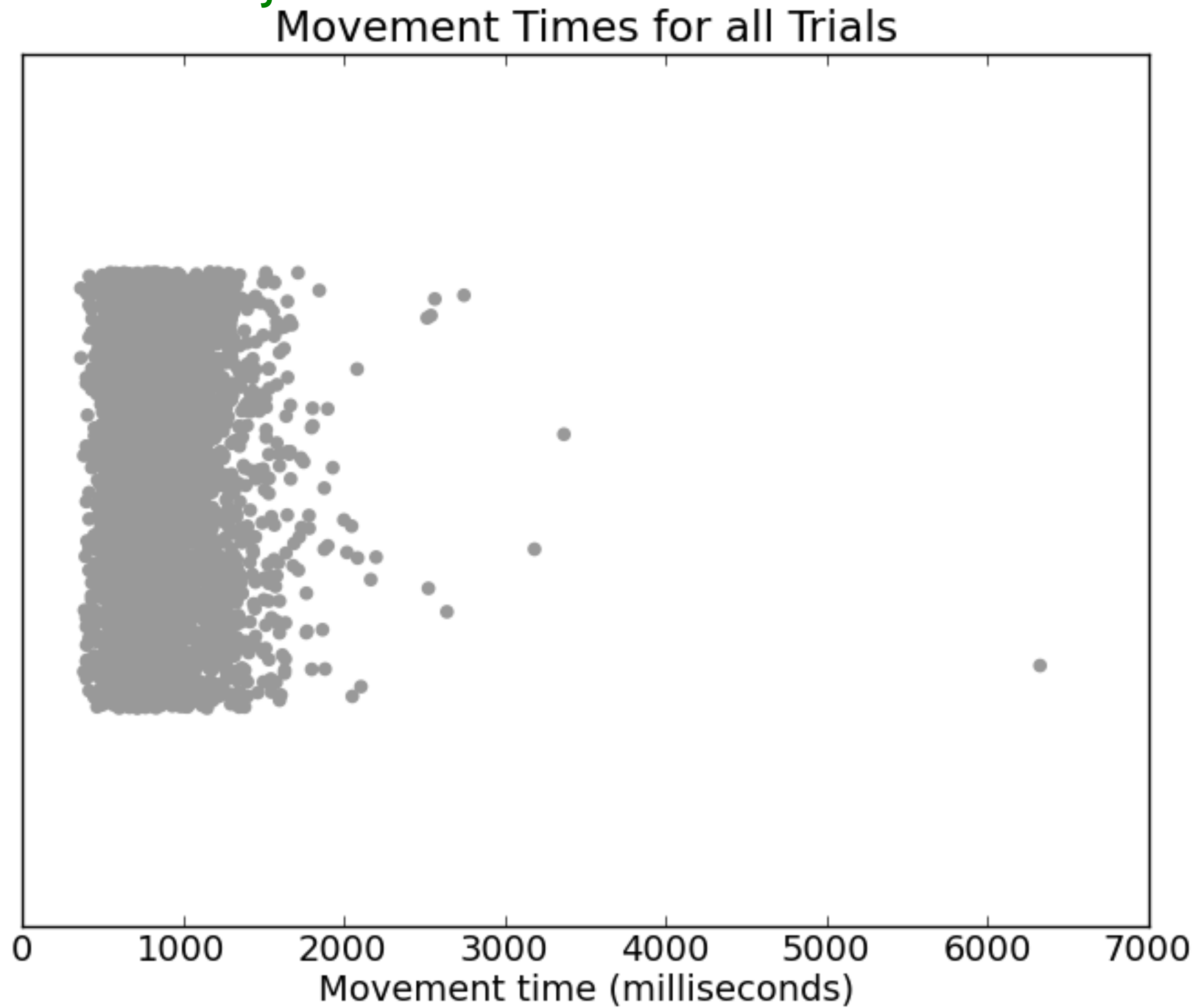
Exploratory Data Analysis (EDA):

Look at your data from different perspectives to get better intuition for it.

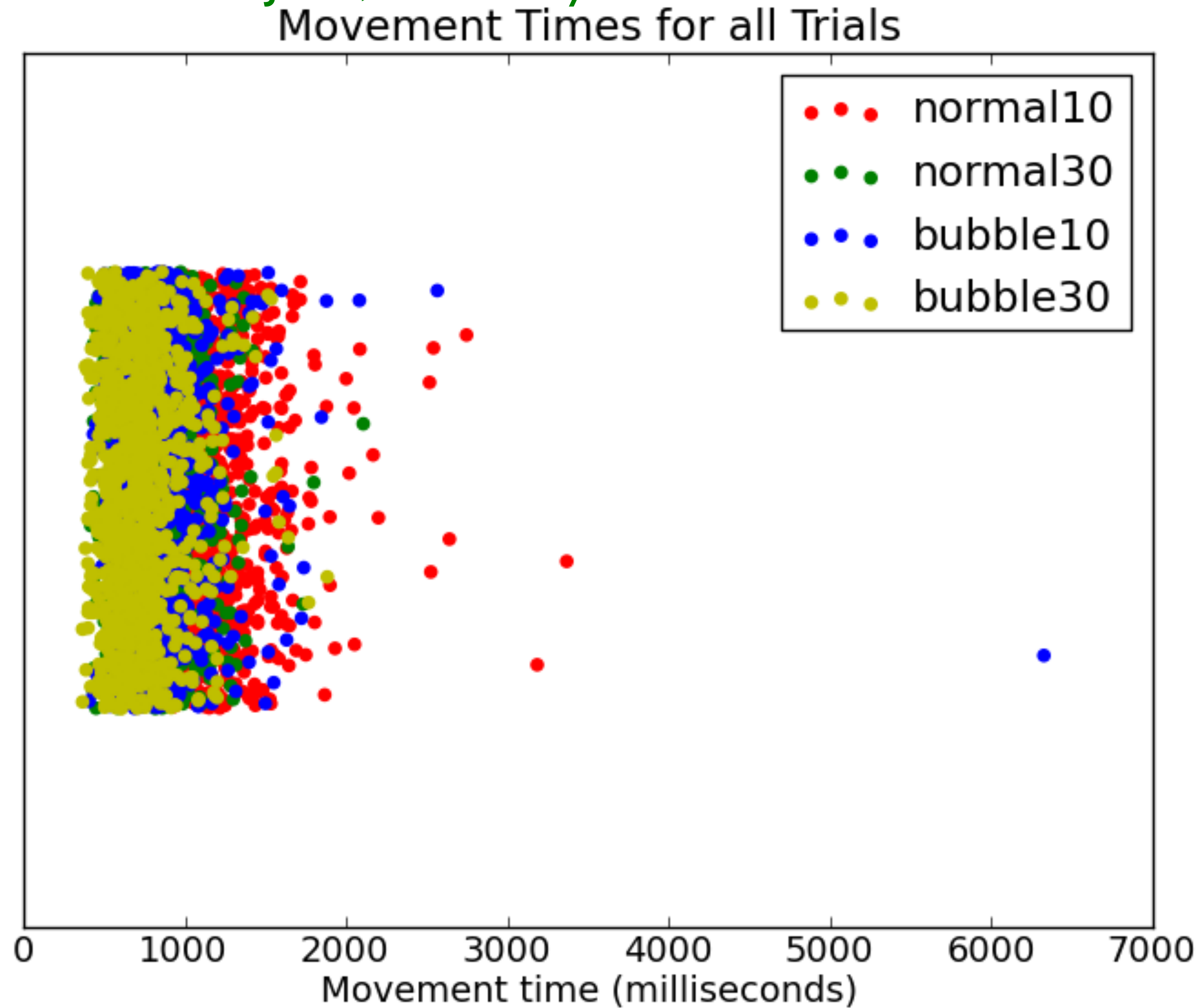
Show the raw data!

Use different visualizations: Histograms, scatterplots, box plots, ...

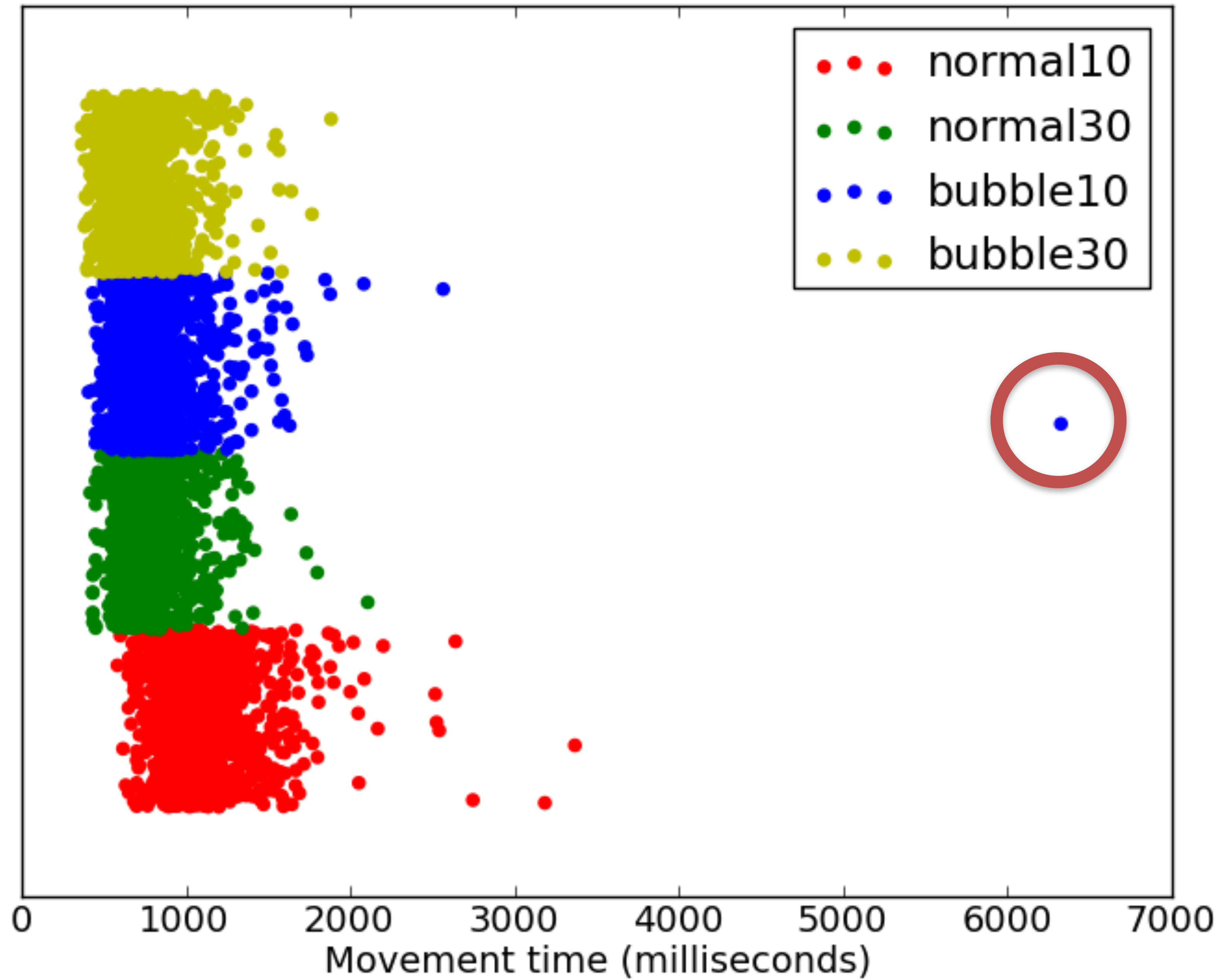
ID Scatter Plot with Jitter



ID Scatter Plot with Jitter, colored by condition



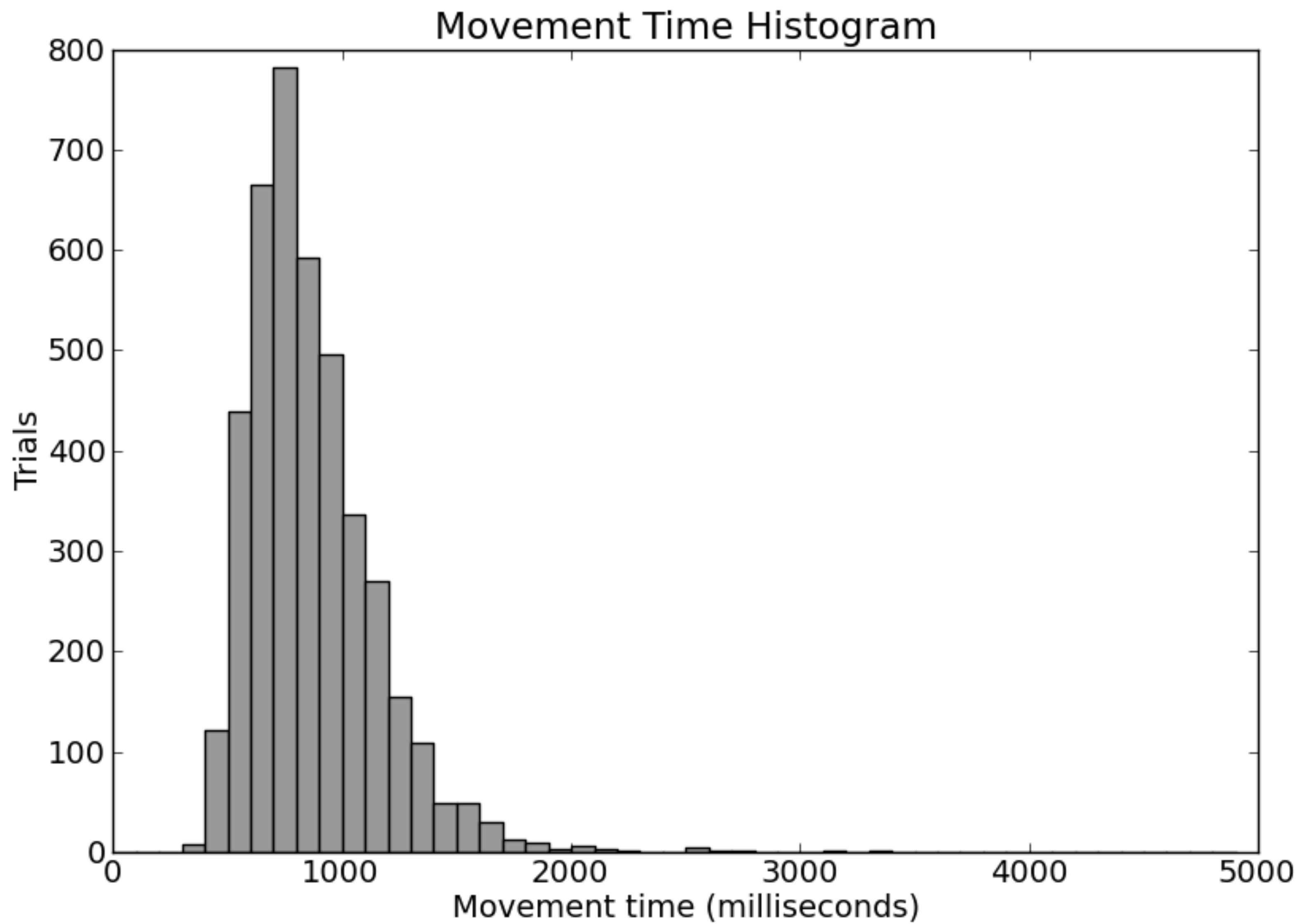
Movement Times for all Trials

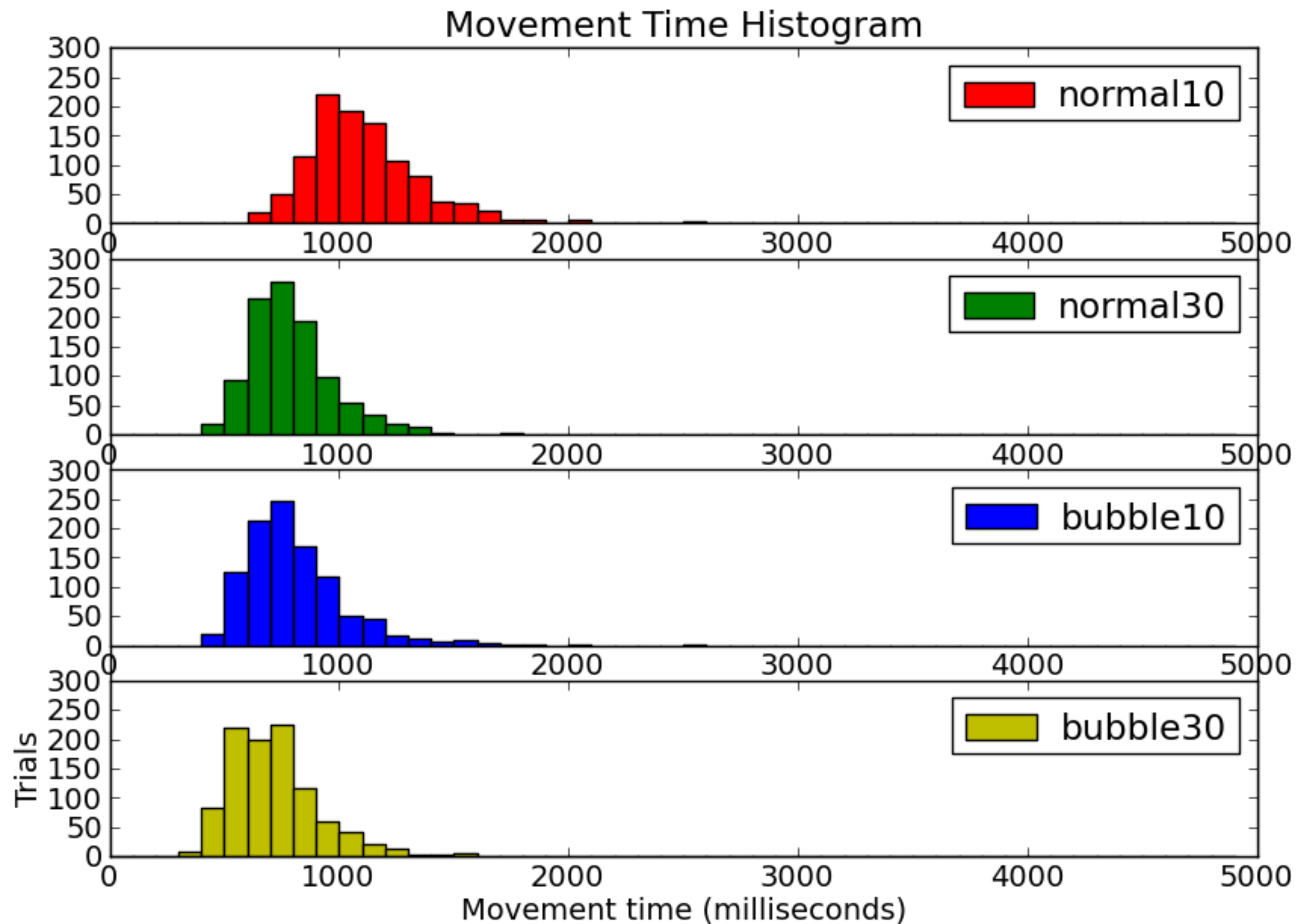


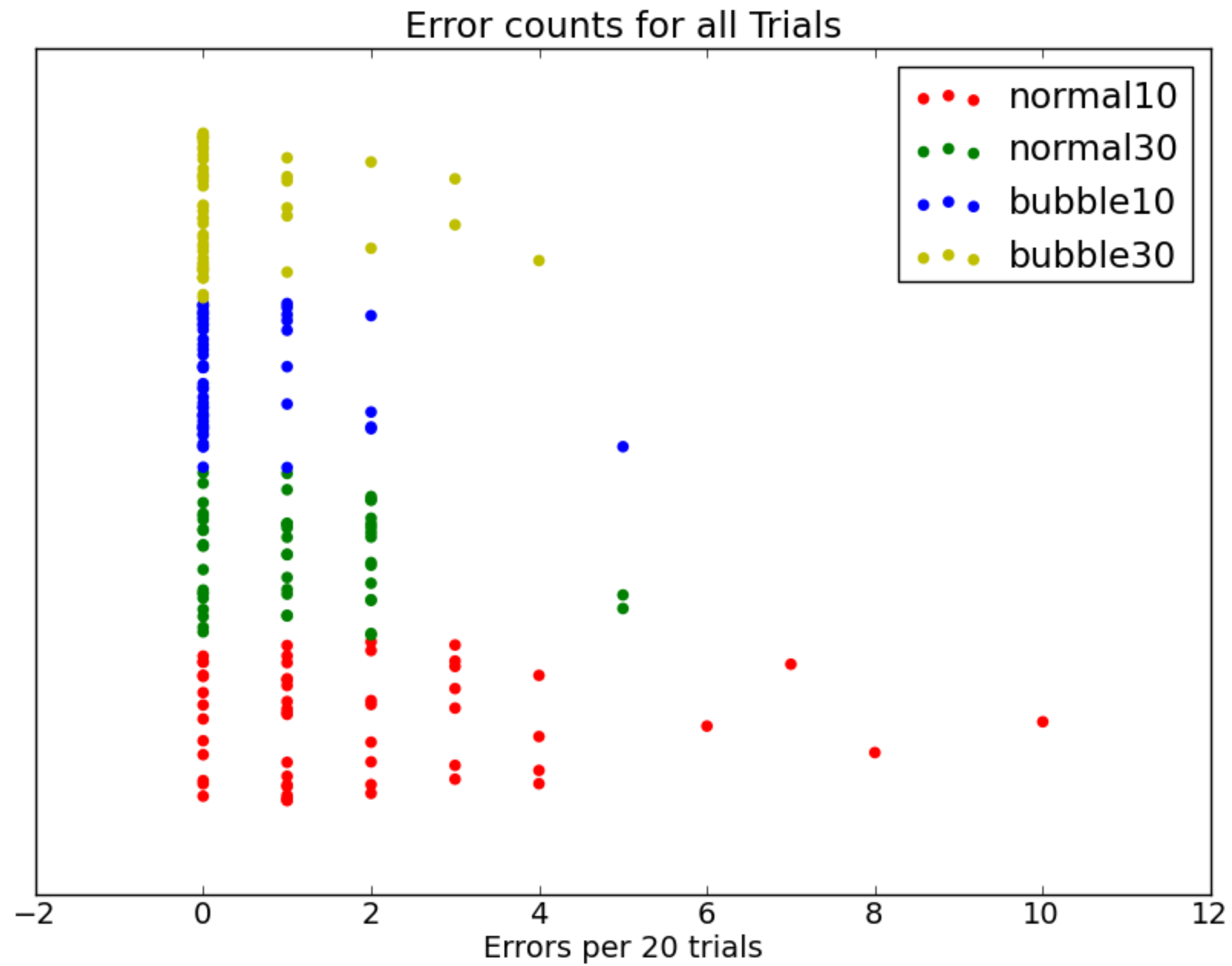
CLEANING DATA

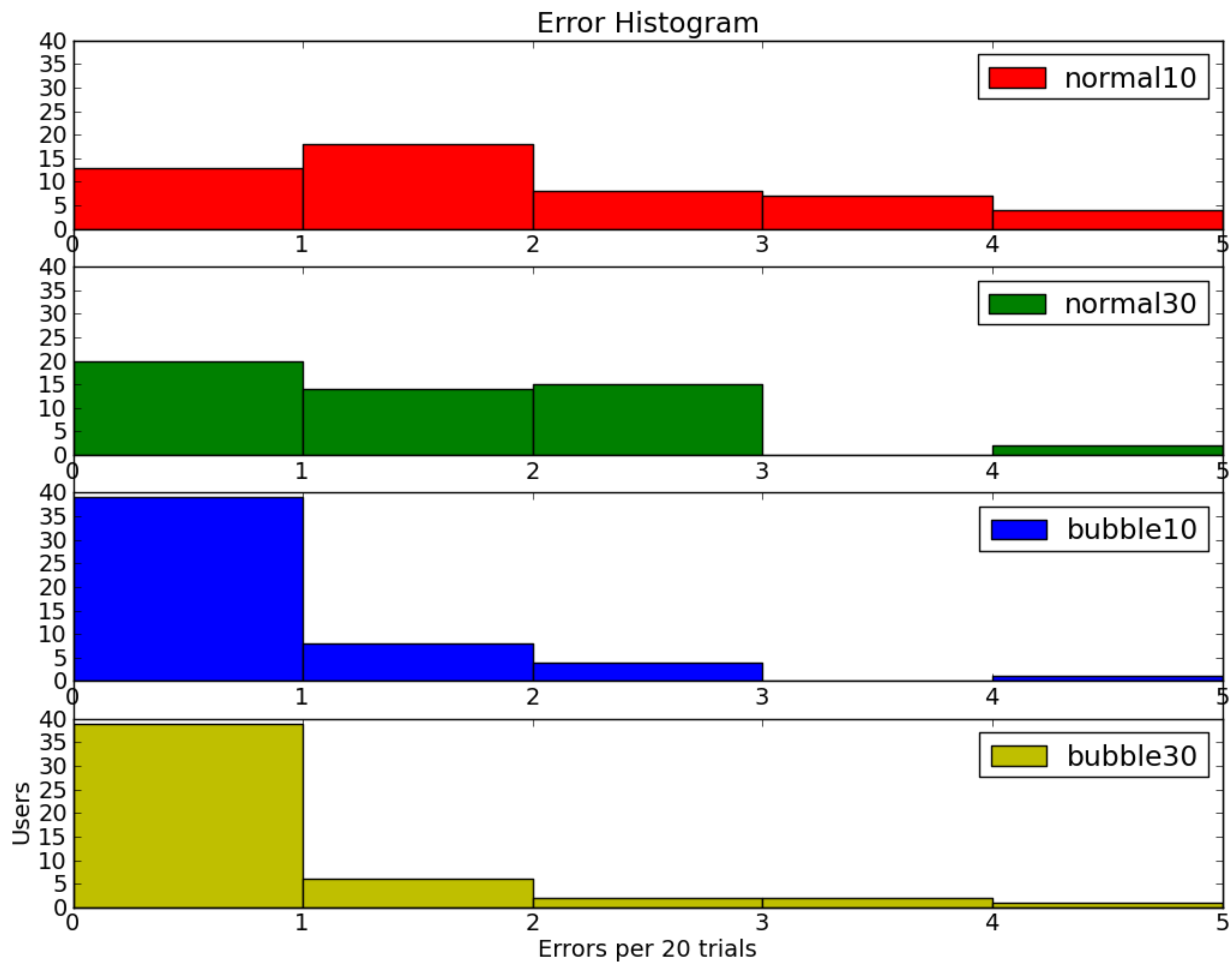
Don't discard data just because it doesn't fit your expectation! Maybe your assumptions were wrong.

In online experiments, discarding extreme outliers can make sense if you believe they reflect users not following normal task protocol (e.g., multitasking in a reaction-time study)









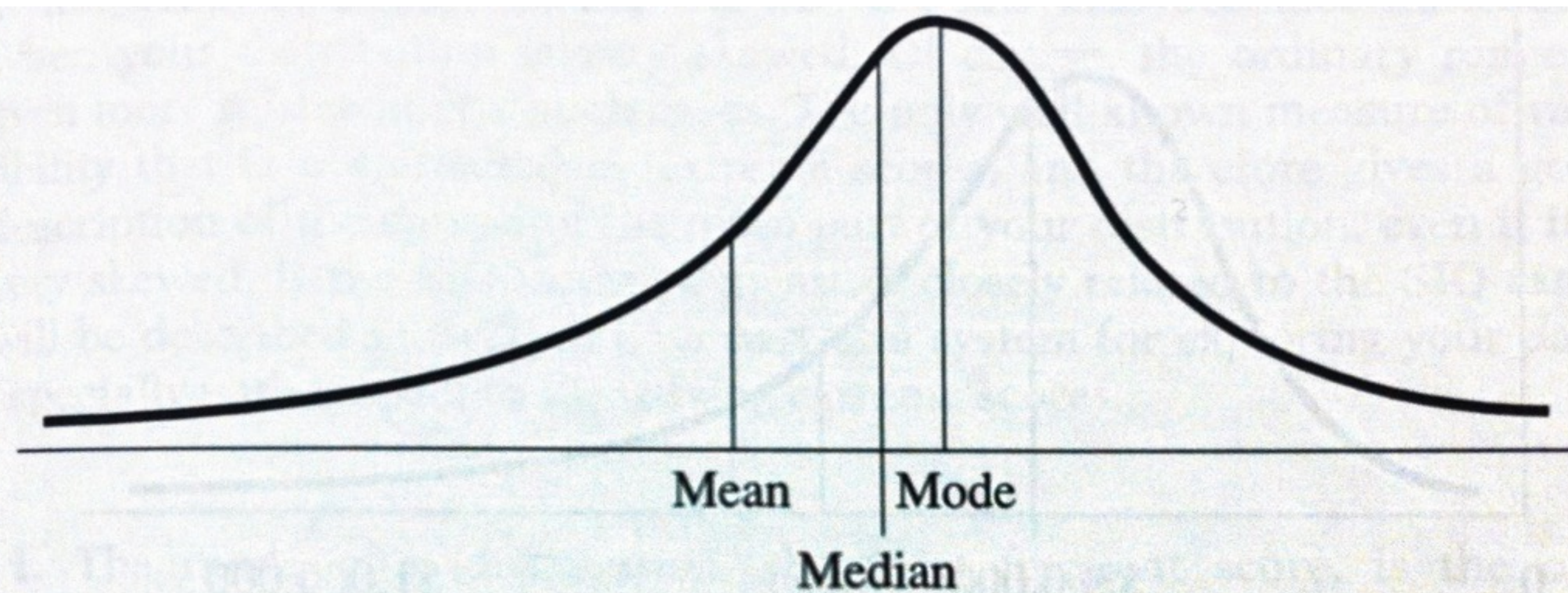
MEDIAN VS. MEAN

For **normally distributed data**, mean=median.

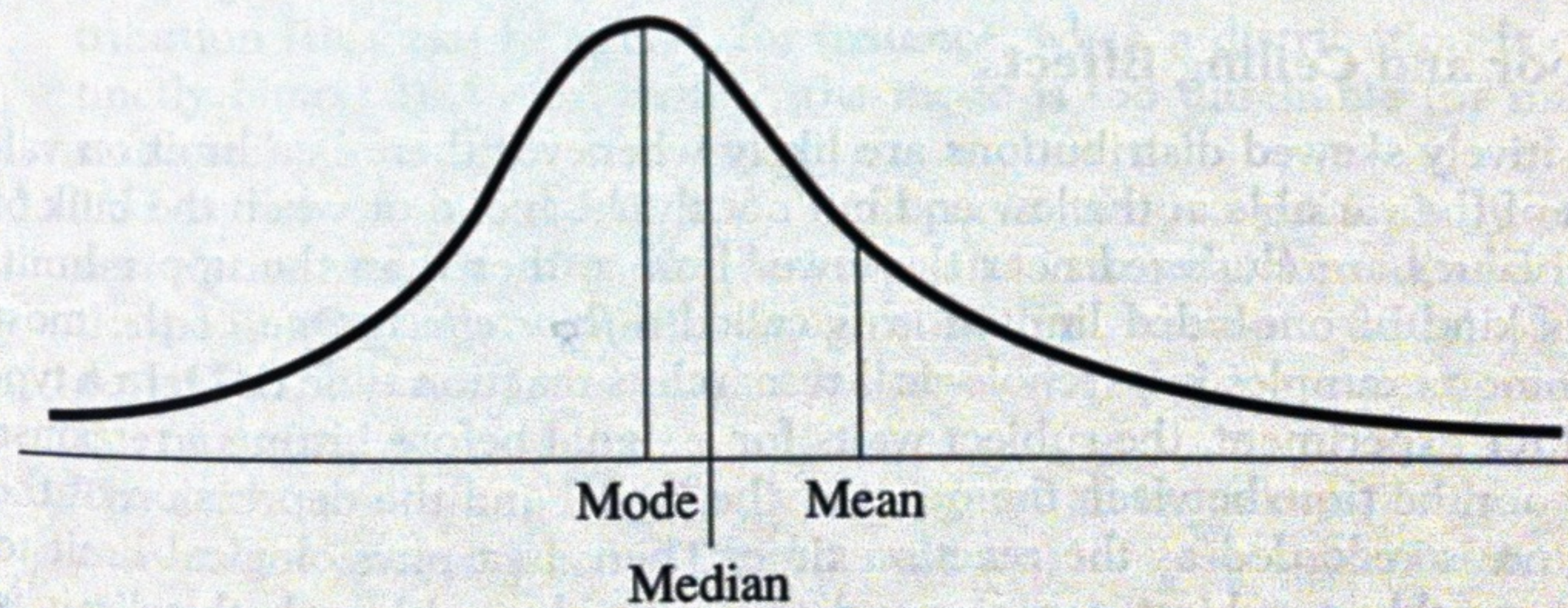
Many data sets gathered online are strongly skewed (they exhibit power law distributions - “long tails”)

Outliers pull the mean to the right/left

Median is more robust!



a. Negatively skewed distribution



b. Positively skewed distribution

POWER LAW DISTRIBUTIONS

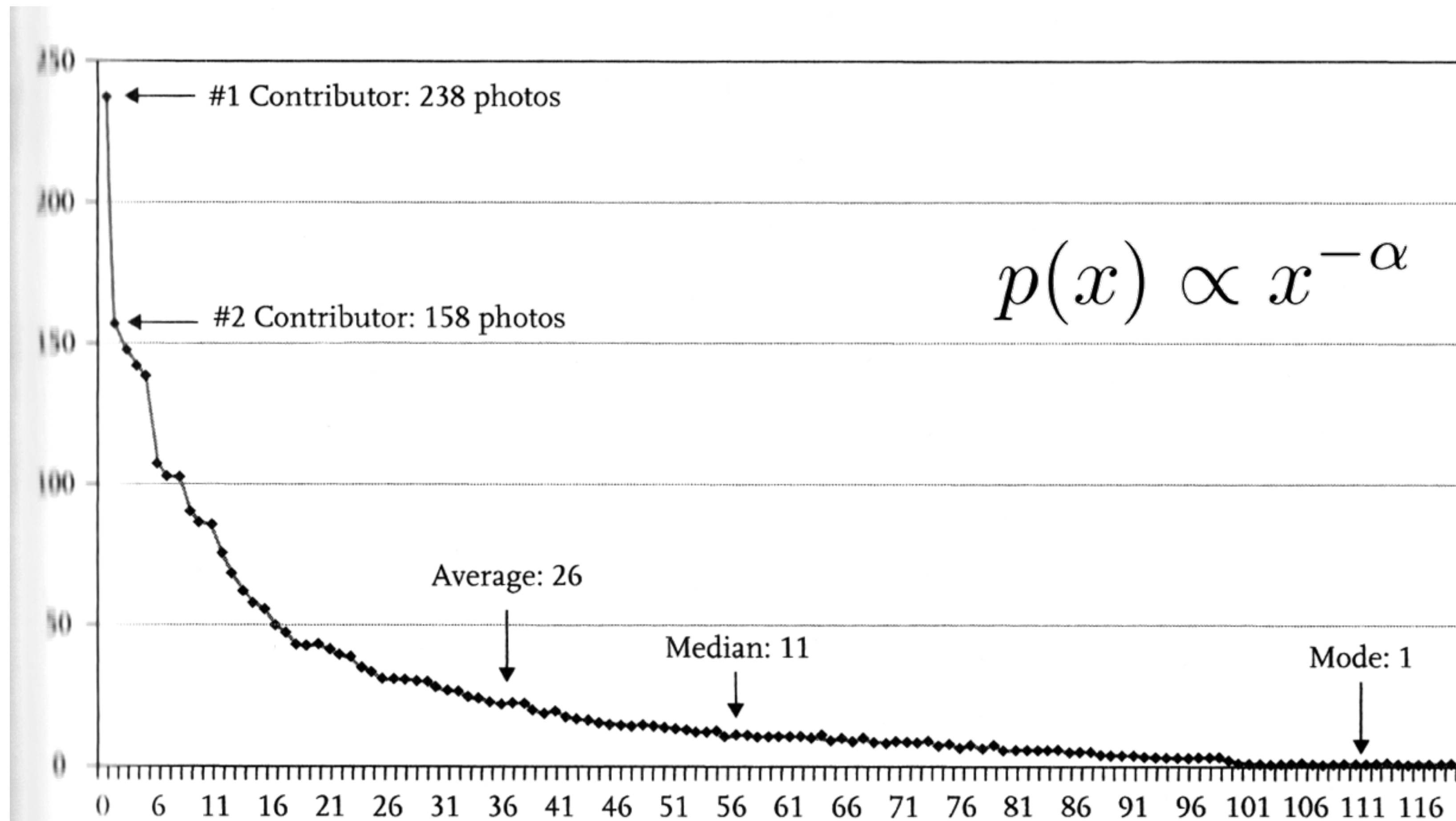
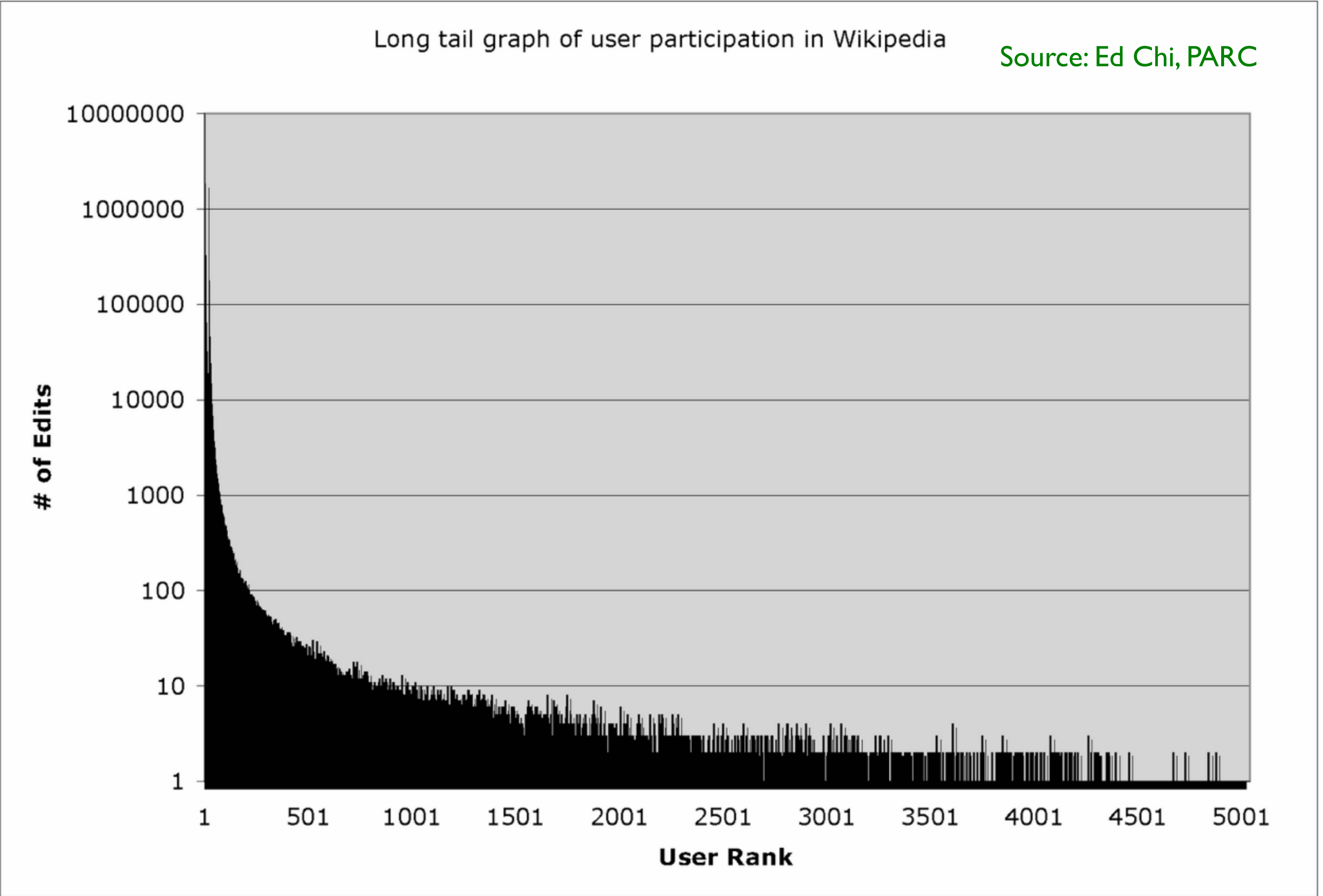
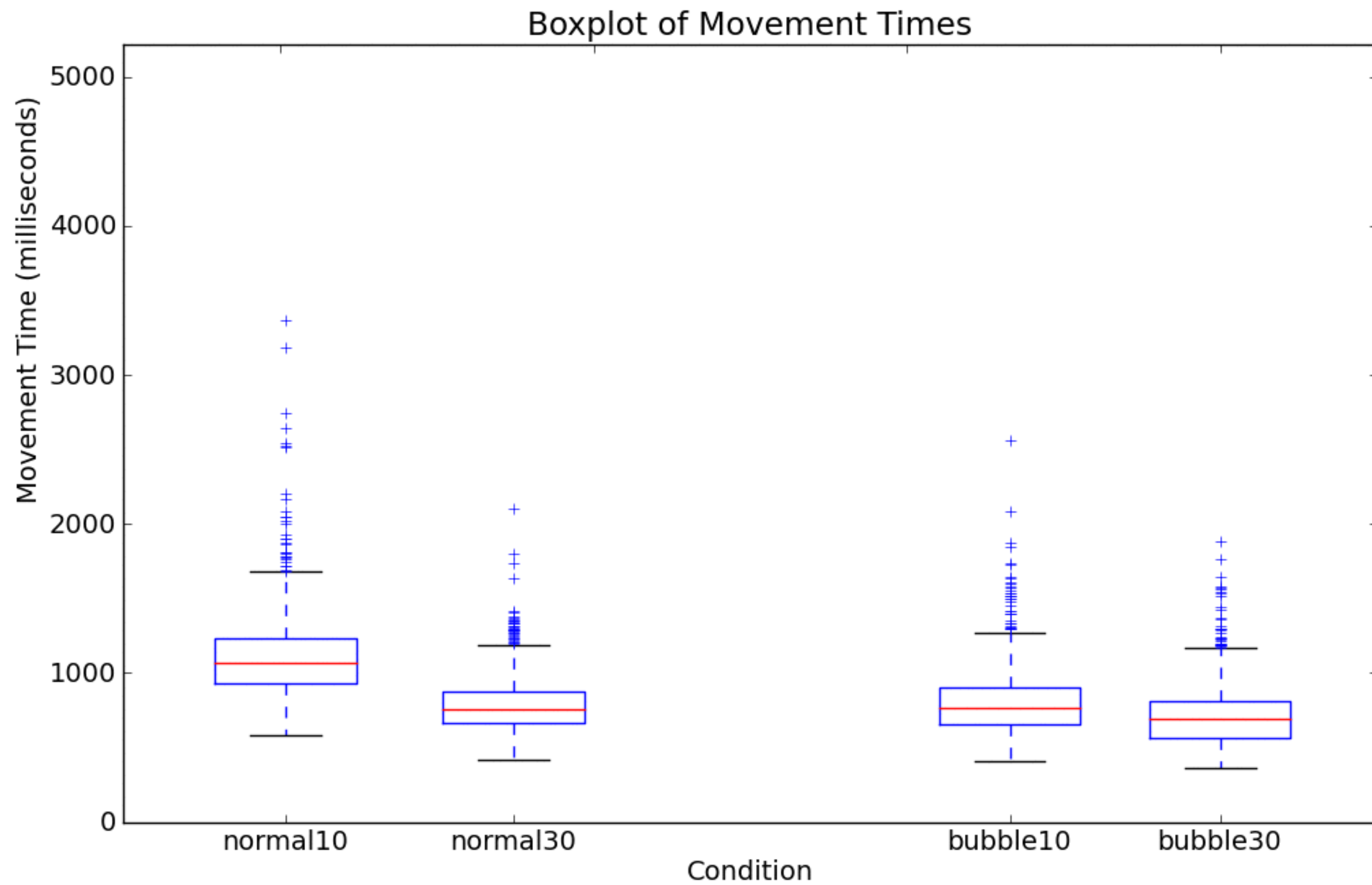


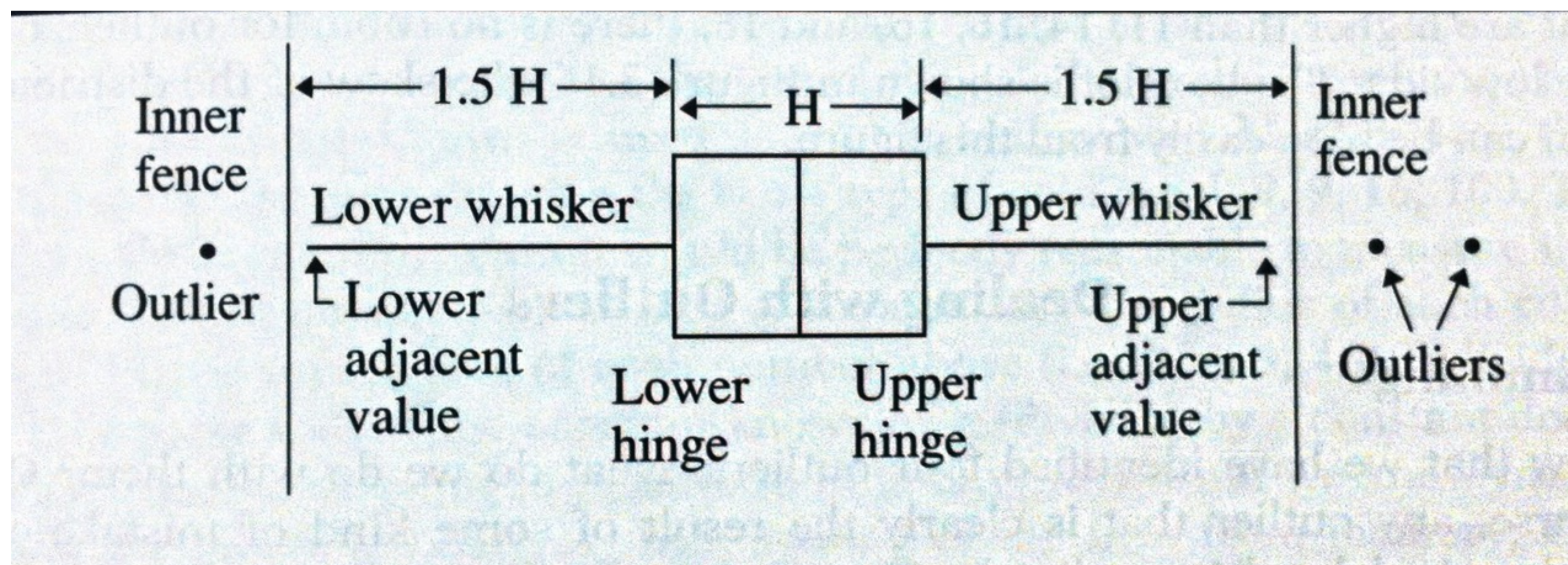
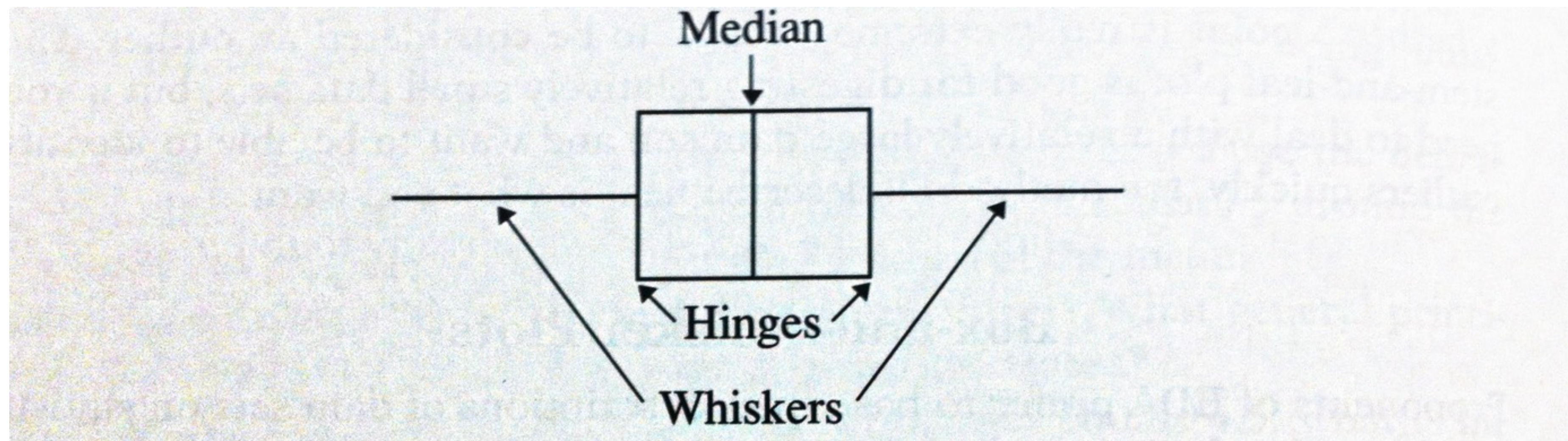
Figure 5-1: The distribution of photographers contributing photos of the 2005 Coney Island Mermaid Parade.

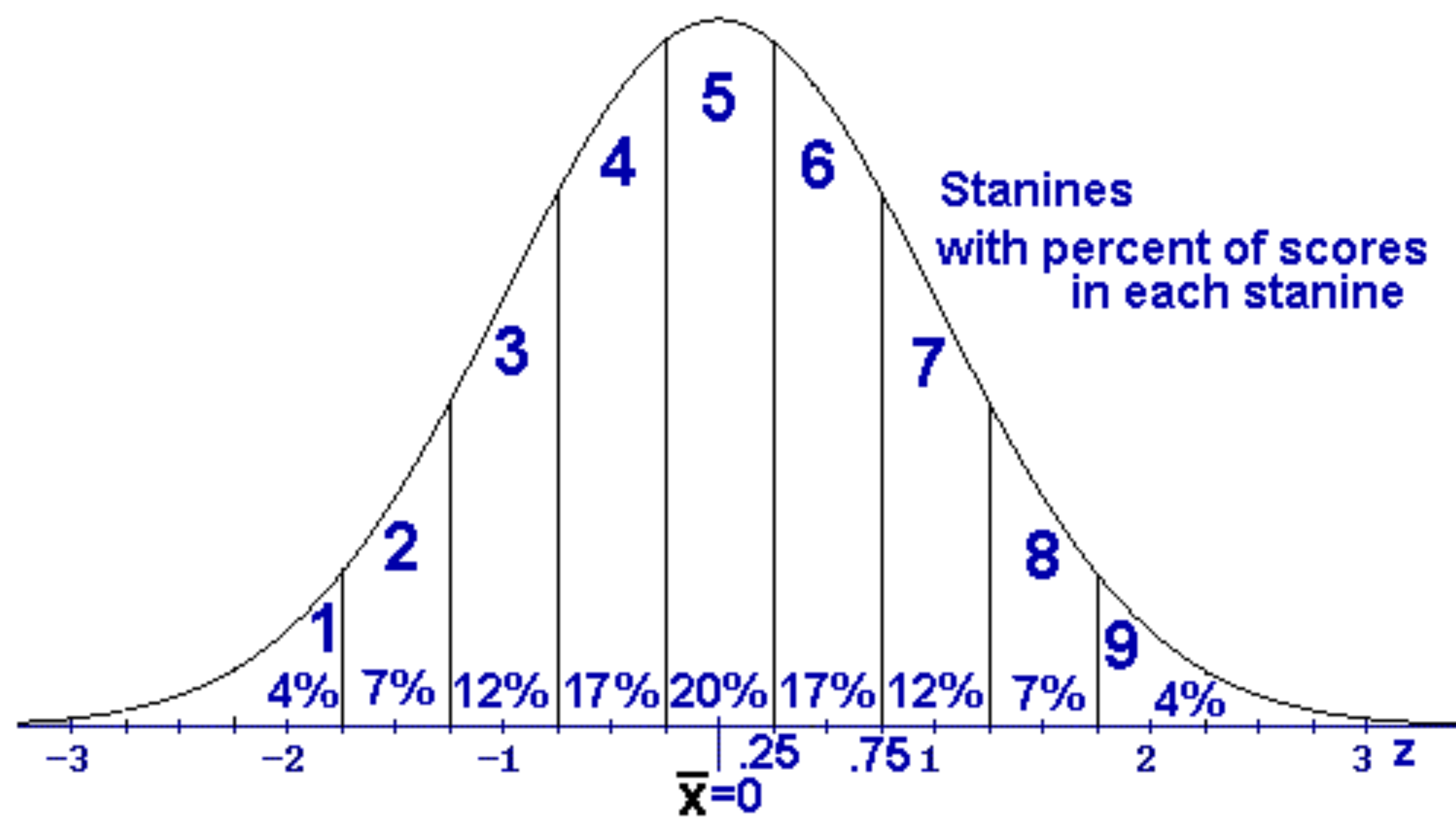
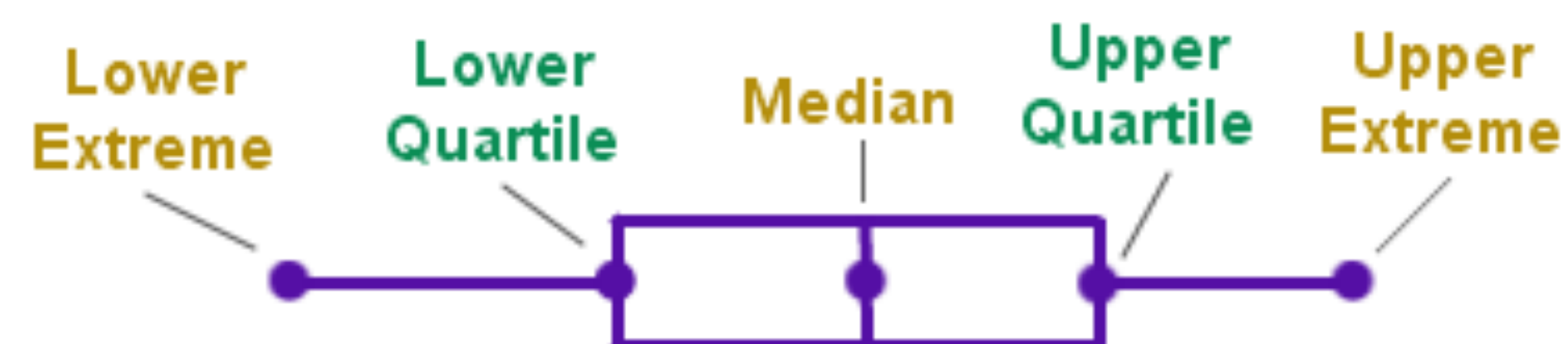
From C. Shirky, *Here Comes Everybody*

POWER LAW DISTRIBUTION

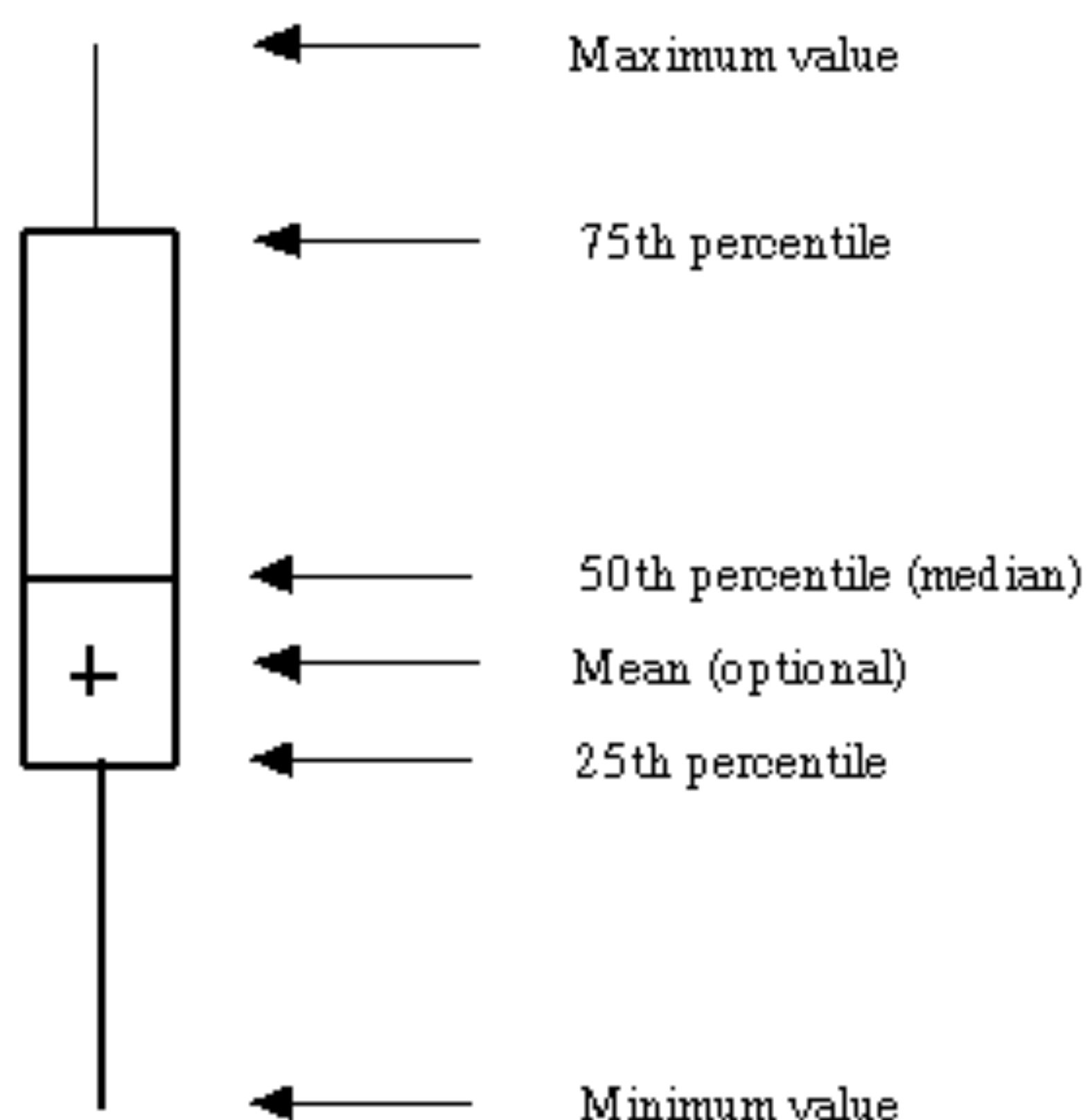


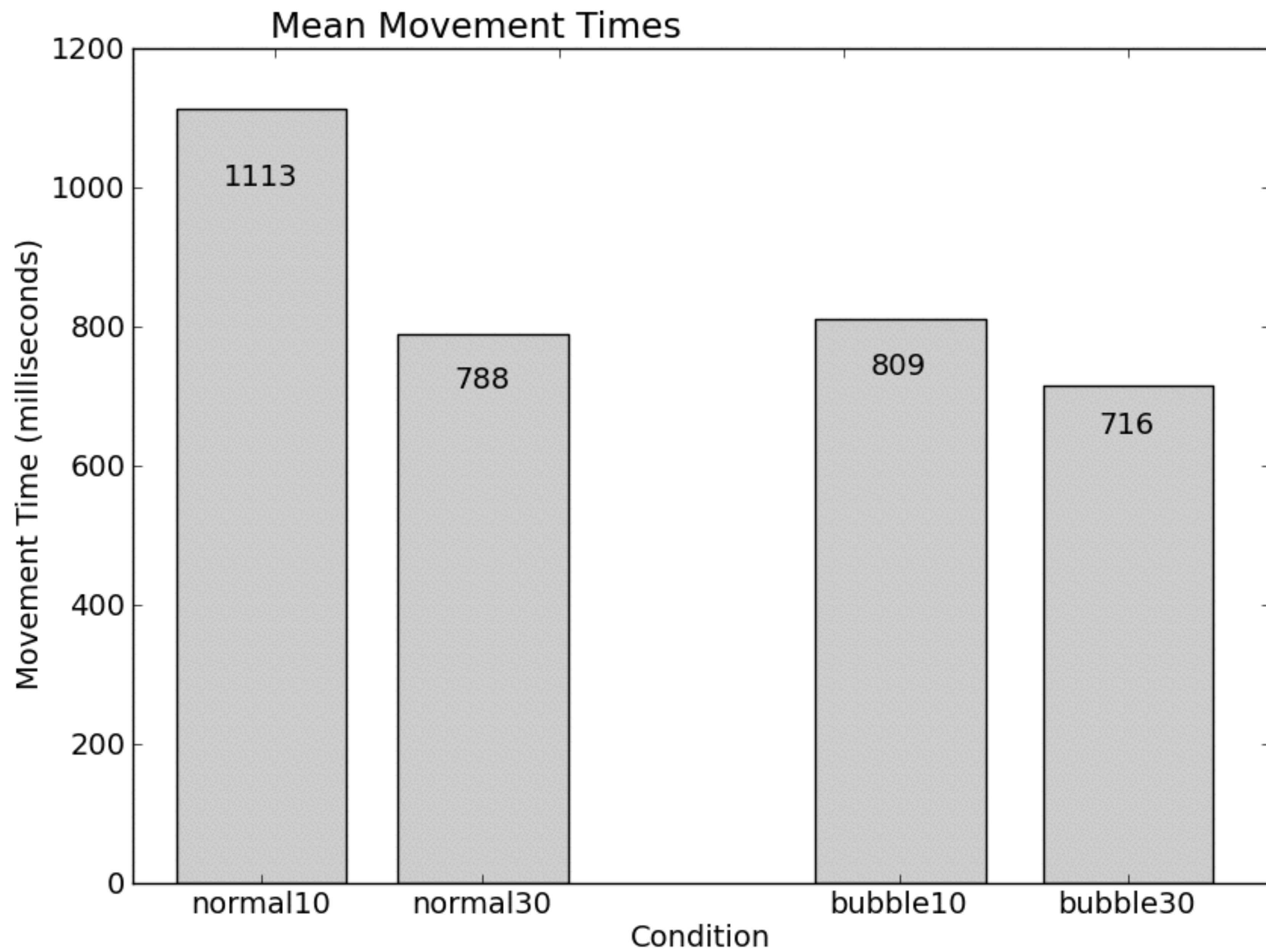


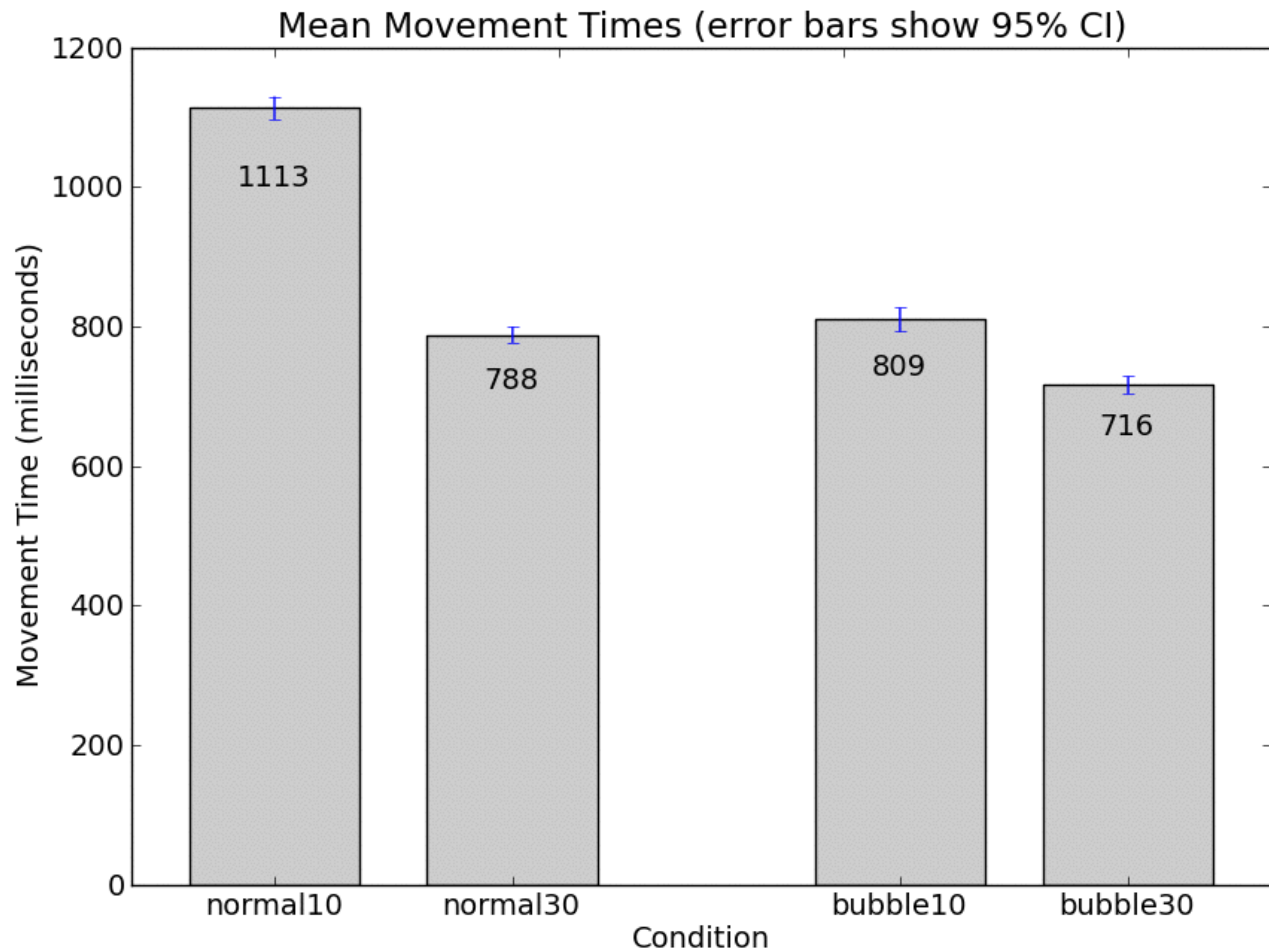




An appropriate scale







CONFIDENCE INTERVAL

confidence interval (also called margin of error) is the plus-or-minus figure usually reported in newspaper or television opinion poll results.

For example, if you use a confidence interval of 4 and 47% percent of your sample picks an answer you can be "sure" that if you had asked the question of the entire relevant population between 43% ($47-4$) and 51% ($47+4$) would have picked that answer

CONFIDENCE LEVEL

confidence level tells you how sure you can be expressed as a percentage and represents how often the true percentage of the population who would pick an answer lies within the confidence interval.

The 95% confidence level means you can be 95% certain

SAMPLE SIZE

1000 people in population

95% confidence level

Confidence interval of ± 5

Need to sample 278 people

Confidence interval of ± 1

...you need to sample 906 people

EFFECT SIZES: TIME

Normal vs. Bubble cursor at target size 10:

1113ms vs. 810ms: **Bubble cursor 27% faster**

Normal vs. Bubble cursor at target size 30:

788ms vs. 716ms: **Bubble cursor 9% faster**

Target size for normal cursor:

1113ms vs 788ms: **Larger targets 29% faster**

Target size for Bubble cursor:

810ms vs. 716ms: **Larger targets 11% faster**

EFFECT SIZES: ERROR

Normal vs. Bubble cursor, target size 10:

1.89 vs. 0.4 Errors per 20 trials: **79% fewer errors**

Normal vs. Bubble cursor, target size 30:

1.06 vs. 0.02 Errors per 20 trials: **98% fewer errors**

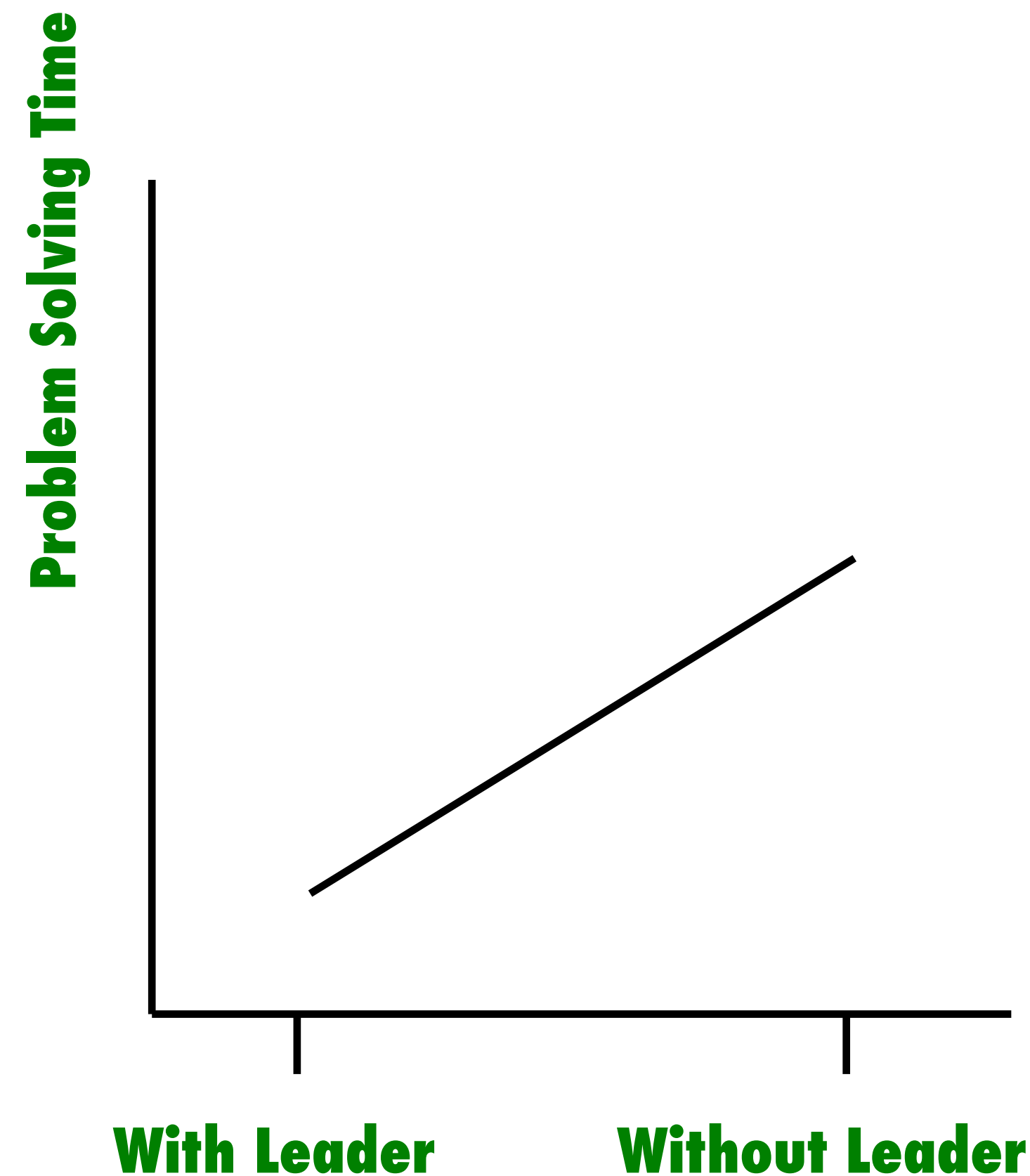
INTERACTION EFFECTS

Relationship between one IV and DV depends on the level of another IV

EXAMPLE OF INTERACTIONS

Group problem solving

Independent variable: Leadership



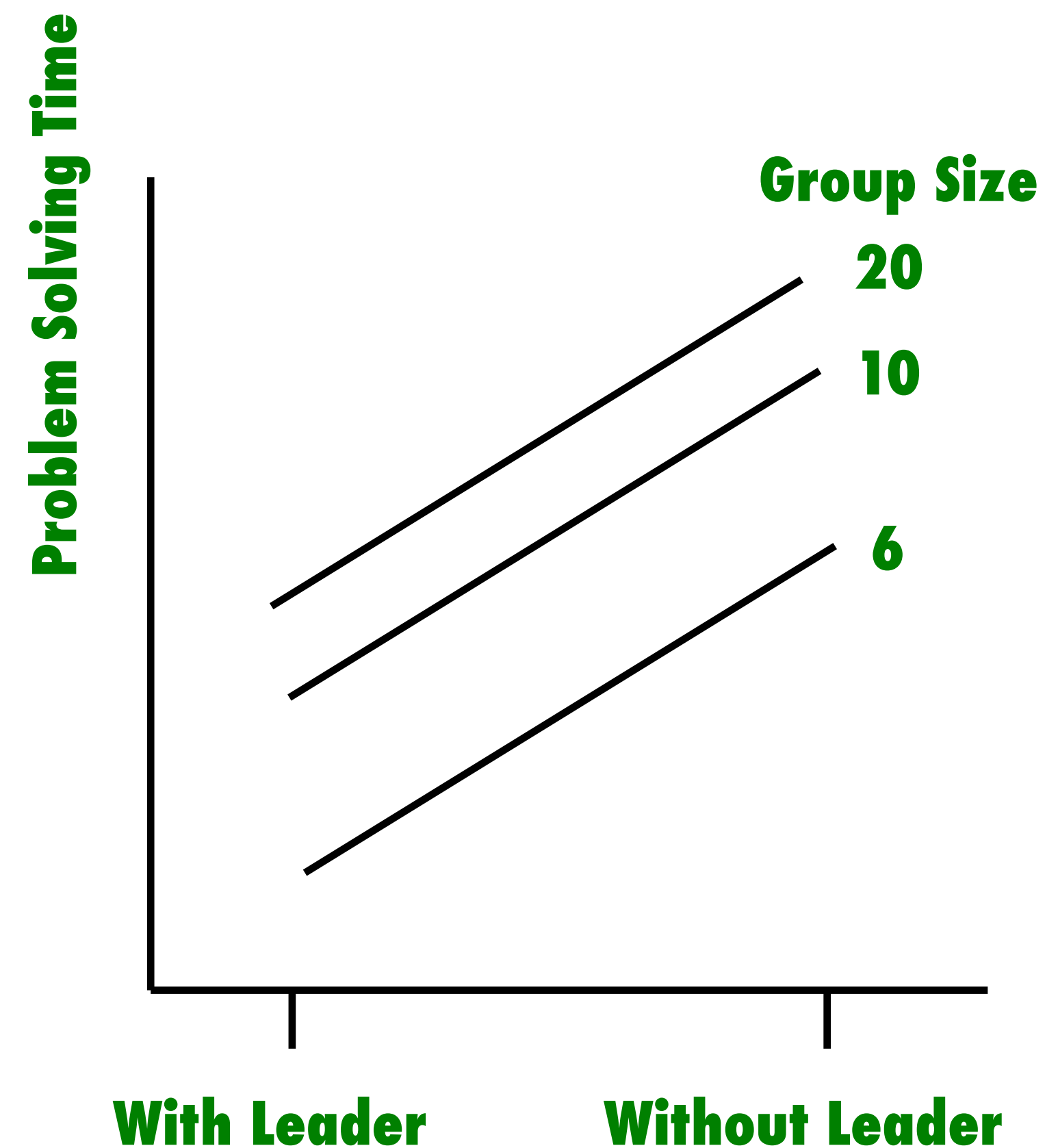
[from Martin 04]

EXAMPLE OF INTERACTIONS

Group problem solving

Independent variable: Leadership

Independent variable: Group size

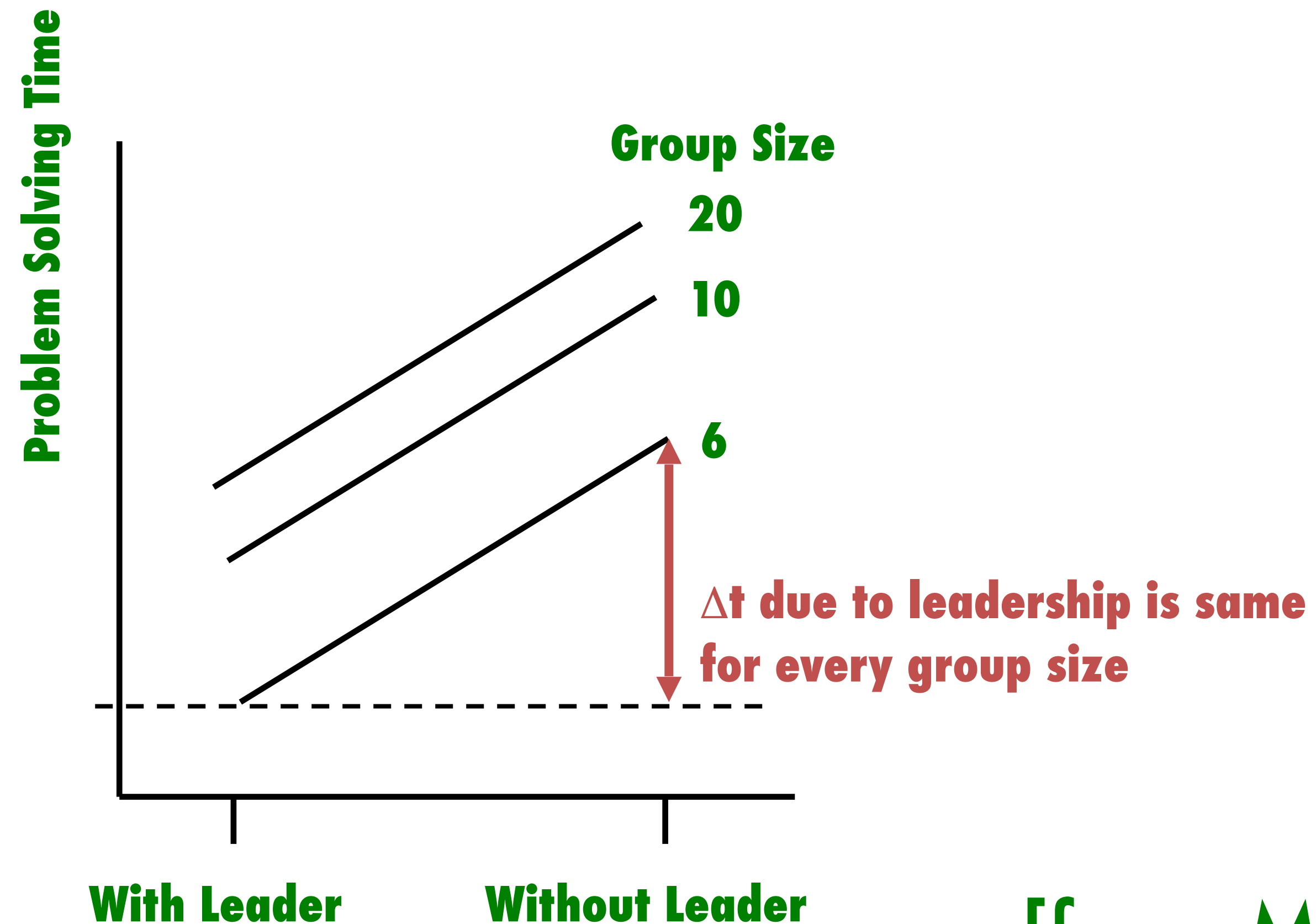


[from Martin 04]

EXAMPLE OF INTERACTIONS

Group problem solving

Change in time due to leadership is same regardless of group size



[from Martin 04]

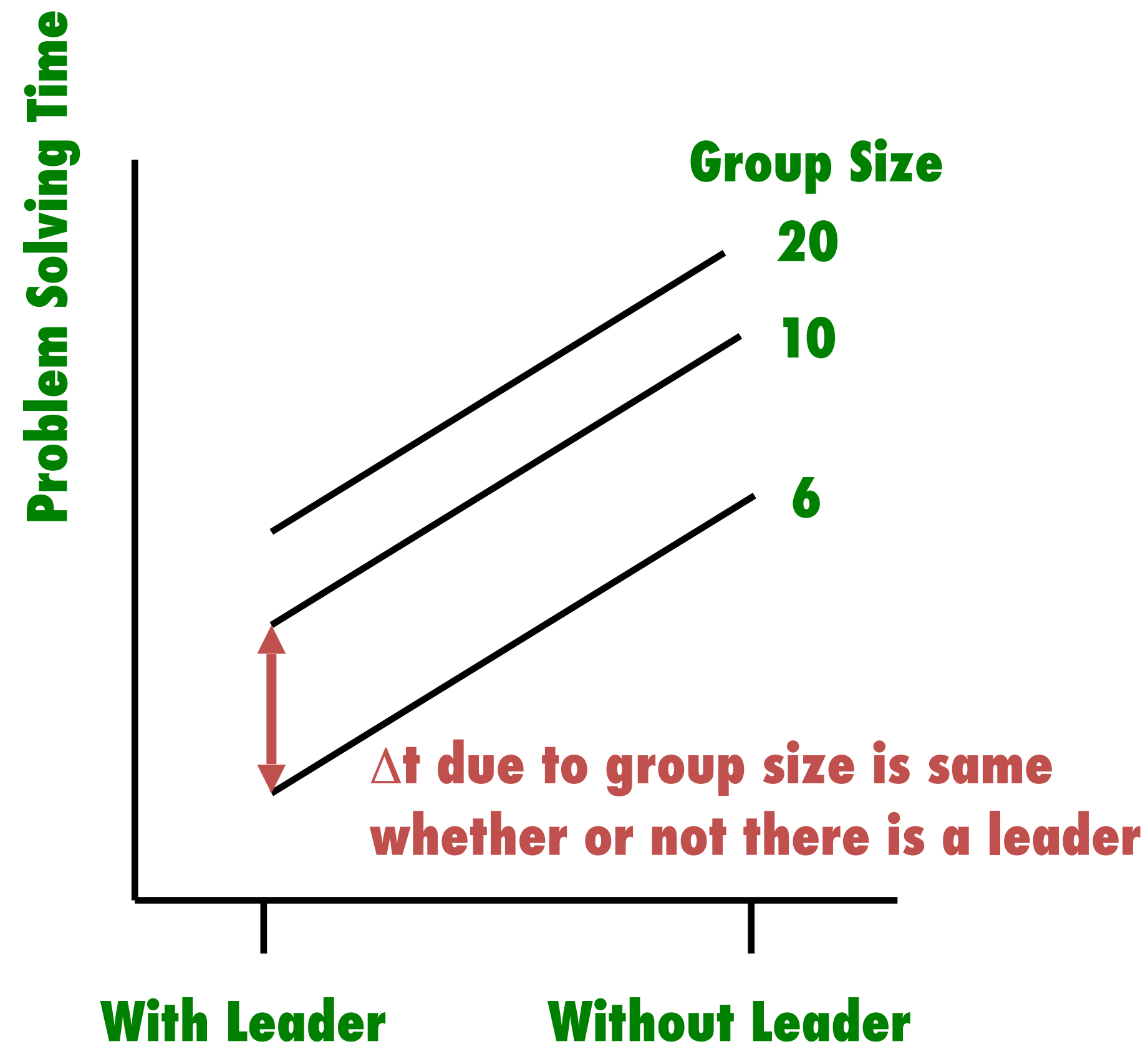
EXAMPLE OF INTERACTIONS

Group problem solving

Change in time due to leadership is same regardless of group size

Change in time due to group size is same regardless of leadership

Independent variables do not interact



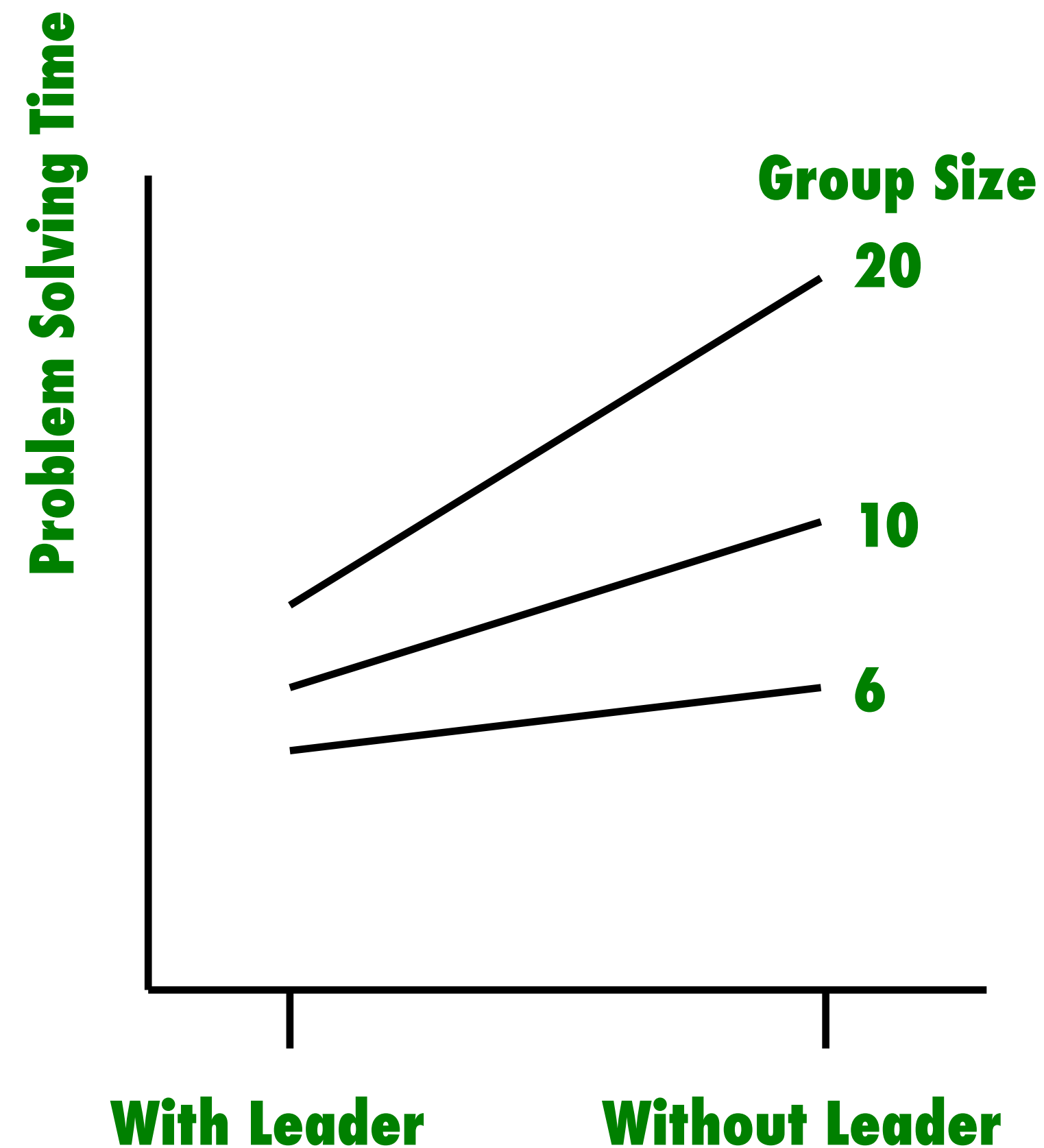
[from Martin 04]

EXAMPLE OF INTERACTIONS

Multiple IVs effect DV non-additively

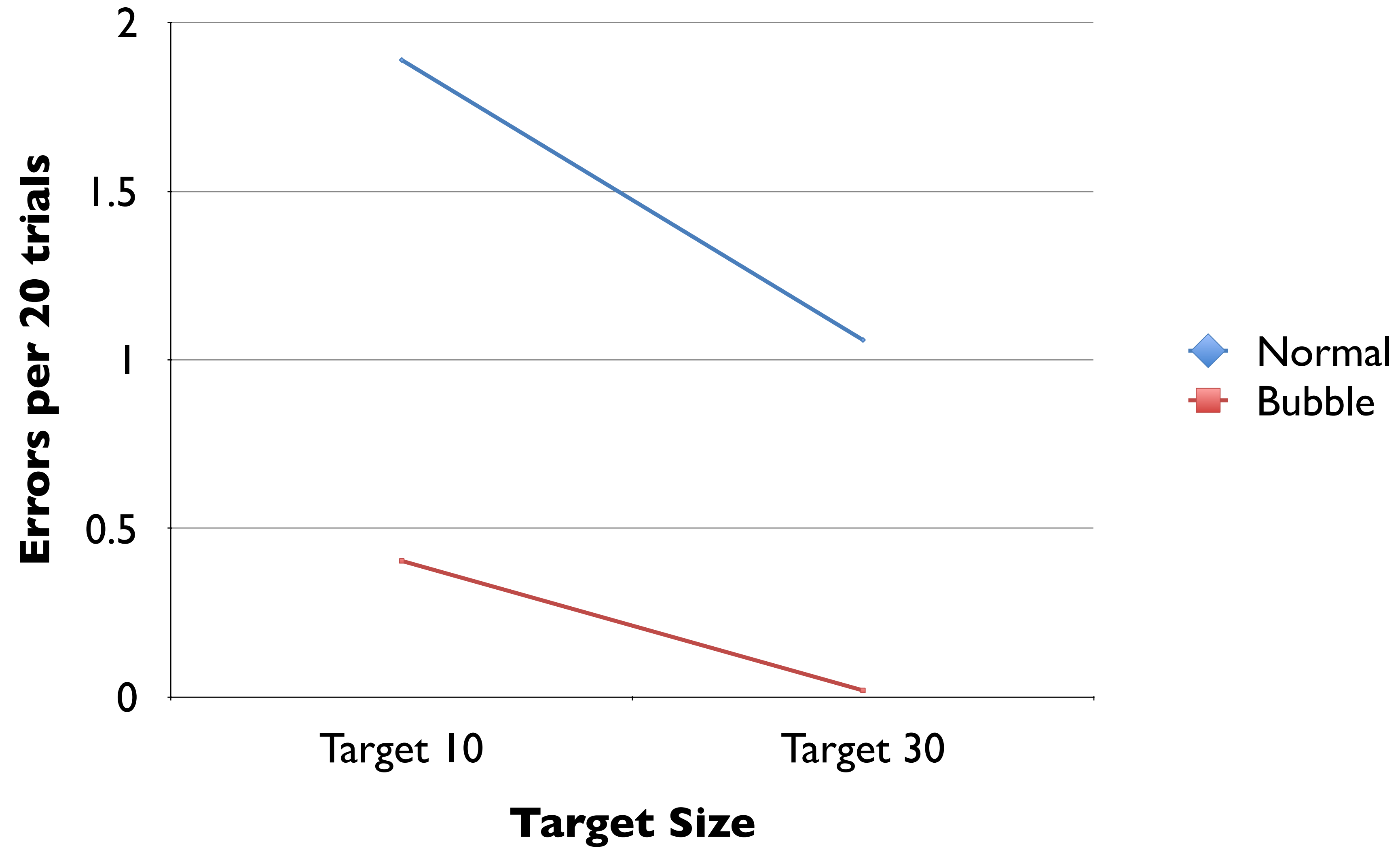
Change in time due to leadership differs with changes in group size

Independent variables do interact



[from Martin 04]

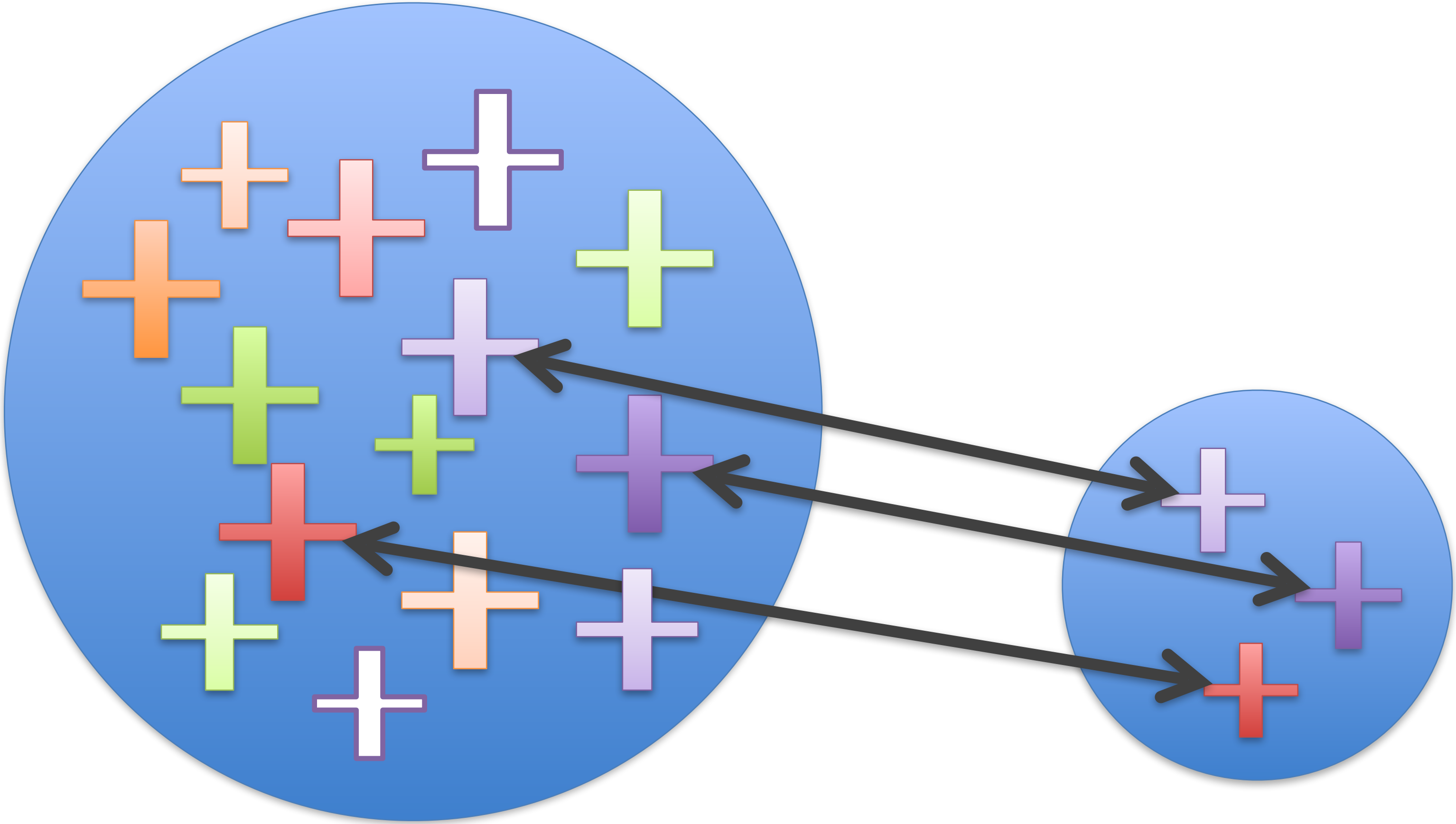
Interaction: Errors



Interaction: Times



POPULATION VERSUS SAMPLE



ARE THE RESULTS MEANINGFUL?

Hypothesis testing

Hypothesis: Manipulation of IV effects DV in some way

Null hypothesis: Manipulation of IV has no effect on DV

Null hypothesis assumed true unless statistics allow us to reject it

Statistical significance (p value)

Likelihood that results are due to chance variation

$p < 0.05$ usually considered significant (Sometimes $p < 0.01$)

Means that $< 5\%$ chance that null hypothesis is true

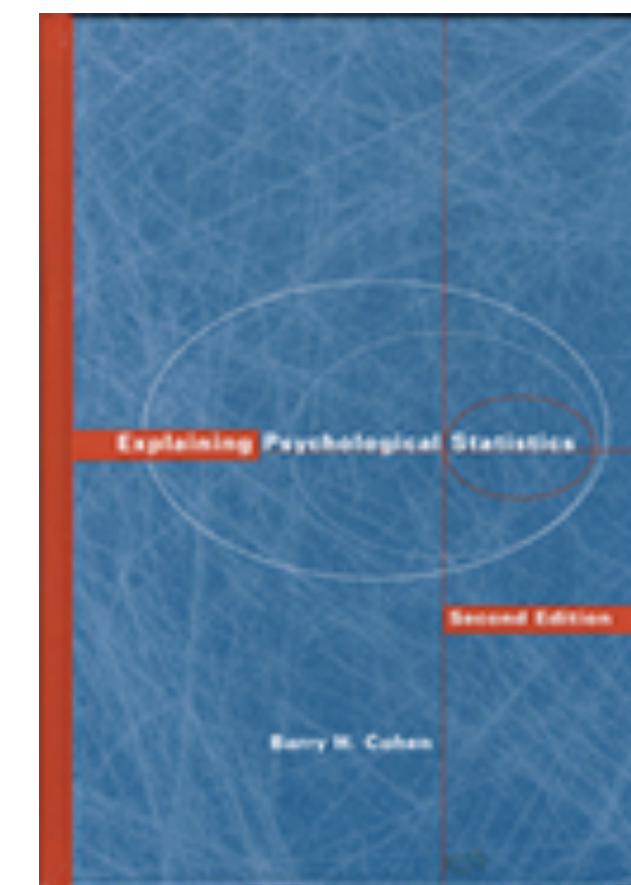
Statistical tests

T-test (1 factor, 2 levels)

Correlation

ANOVA (1 factor, > 2 levels, multiple factors)

MANOVA (> 1 dependent variable)



Explaining Psychological Statistics
Barry H. Cohen

ANOVA - ANALYSIS OF VARIANCE

Single factor analysis of variance (ANOVA)

Compare means for 3 or more levels of a single independent variable

Multi-Way Analysis of variance (n-Way ANOVA)

Compare more than one independent variable

Can find interactions between independent variables

Repeated measures analysis of variance (RM-ANOVA)

Use when > 1 observation per subject (within subjects expt.)

Multi-variate analysis of variance (MANOVA)

Compare between more than one dependent var.

ANOVA tests whether means differ, but does not tell us which means differ – for this we must perform pairwise t-test

OUR EXAMPLE

Two-Way ANOVA (Cursor, Size) for **time**:

Main effect for **cursor**

$F(1,4136) = 641.03$, $p < 0.001$ is statistically significant.

Main effect for **size**

$F(1,4136) = 778.31$, $p < 0.001$ is statistically significant.

Interaction **cursor x size**

$F(1,4136) = 232.94.2$, $p < 0.001$ is statistically significant.

OUR EXAMPLE

Two-Way ANOVA (Cursor, Size) for **errors**:

Main effect for **cursor**

$F(1,203) = 32.4$, $p < 0.001$ is statistically significant.

Main effect for **size**

$F(1,203) = 4.9$, $p = 0.02$ is statistically significant.

Interaction **cursor x size**

$F(1,203) = 4.7$, $p = 0.03$ is statistically significant.

ERRORS IN BUBBLE CURSOR CASE ONLY

$F(1,2038) = 0.009$, $p=0.92$ – NOT significant

WHAT DOES $P > 0.05$ MEAN?

No statistically significant difference (at 5% level)

Are the two conditions thus equivalent?

NO! We DID observe differences.

But can't be sure they are not due to chance.

SUMMARY

Quantitative evaluations

Repeatable, reliable evaluation of interface elements

To control properly, usually limited to low-level issues

Menu selection method A faster than method B

Pros/Cons

Objective measurements

Good internal validity → repeatability

But, real-world implications may be difficult to foresee

Significant results doesn't imply real-world importance

3.05s versus 3.00s for menu selection