

Extrinsic Evaluation of Text Classification for Policy Analysis Based on Coding Human Values

Douglas W. Oard

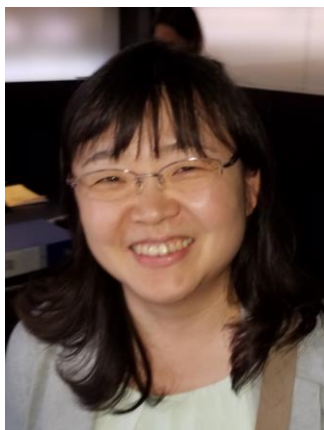
University of Maryland, College Park (iSchool/UMIACS)

University of South Florida (iSchool)

University of Florida (CS)

Emi Ishita

Kyushu University



Yoichi Tomiura

Kyushu University



An-Shou Cheng

Natl Sun Yat-Sen U



Ken Fleischmann

University of Texas



Yasuhiro Takayama

Tokuyama College



The Net Neutrality Debate



thenextweb.com



viewpointdaily.com

Social Science Research Questions

- What do people value, and how do they express those values via text?
- How are those values related to other text features (e.g., sentiment)?
- What are the roles of human values in the net neutrality debate?

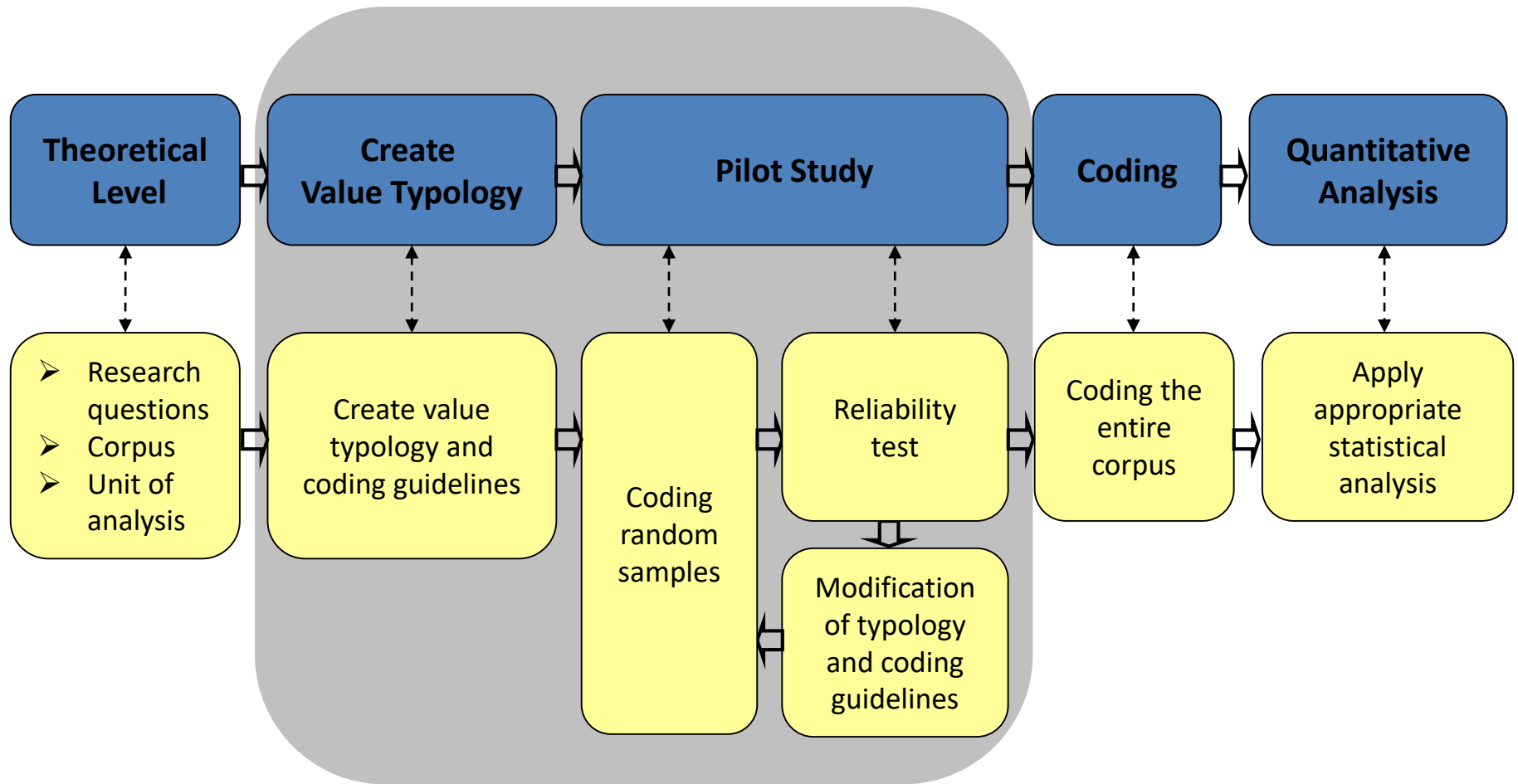
Defining Human Values

Rokeach (1973)	enduring belief...personally or socially preferable to a mode of conduct or end-state of existence
Schwartz (1994)	A belief...guides selection or evaluation of behavior, people, and events
Kluckhohn (1951)	A conception...influences the selection from available modes, means, and ends of action
Braithwaite & Blamey (1998)	principles...an individual or a collective considers preferable across contexts and situations
Friedman et al. (2006)	What a person or group of people consider important in life

102 Annotated Testimonies (Cheng, 2012; Cheng et al., 2012)



Social Science Research Design



Coding Scheme (modified MIHV)

(Cheng & Fleischmann, 2010; Cheng, 2012)

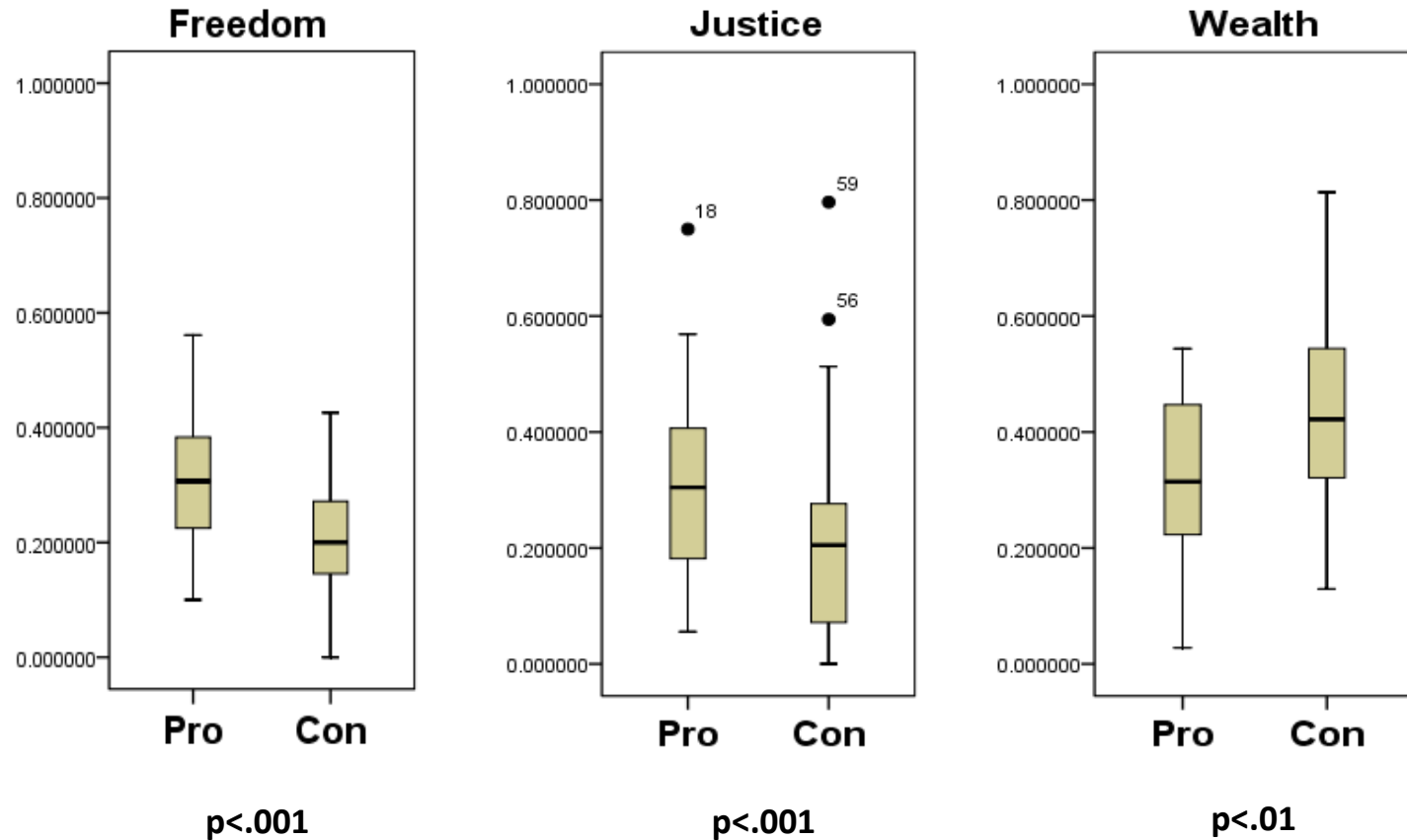
Freedom	The condition of being free of restraints and encouraging competition; allowing individuals to have their own beliefs and to make their own choices; freedom from interference or influence of another or others; the quality of being autonomous and independent.
Honor	Understanding of who you are and how you are perceived by others; a feeling of pride in oneself or one's organization, group, or nation and belief in one's own worth; accomplishment that is honored, esteemed, respected or well regarded by yourself or others.
Innovation	The capacity to create or discover new things and new ideas that contribute to the advancement of knowledge and/or technology.
Justice	The state of being treated equally and fairly, especially having the same rights, status, and opportunities; the process of settling a matter properly and fairly for all parties according to their capabilities and needs, especially protecting the weak and correcting any injustice; need for equal or fair distribution of resources, information, benefits, burdens, and power among the members of a society.
Social Order	Using the power of the government, military and/or legal system to protect the stability of society and/or to protect people from possible harms mentally or physically; acting in accordance with laws, regulations, and social norms.
Wealth	An explicitly stated concern with or interest in pursuing economic goals such as money, material possessions, resources, and profit; focusing on the market value of a change, decision, or action; allocating resources appropriately and/or efficiently.

Agreement and Frequency

(102 documents; 8,660 sentences)

	Kappa	# sentences	# doc
<i>wealth</i>	0.621	3,156	102
<i>social order</i>	0.688	2,503	102
<i>justice</i>	0.423	2,267	99
<i>freedom</i>	0.628	2,155	101
<i>innovation</i>	0.714	1,018	94
<i>honor</i>	0.437	317	80

Value Differences Between Pros and Cons



Computer Science Research Questions

- Can a classifier make some of the coding decisions (nearly) as well as a human?
- What is the best classifier design?

Increasing the amount of spectrum, speeding
the relocation of **government** users, vigorous
anti-trust enforcement (including the prevention
of excessive aggregation of wireless spectrum)
and revamping **universal service** to be
competitively **neutral** are “easy” **economically**.

The diagram consists of five blue speech bubbles with white text. Three bubbles labeled 'Social Order' are positioned above the words 'government' and 'anti-trust enforcement'. Two bubbles labeled 'Justice' are positioned above the words 'universal service' and 'neutral'. One bubble labeled 'Wealth' is positioned above the word 'economically'.

Latent Variable Model Design

social order : 0000**10**



(1) **Protecting** customers and delivering a good Internet experience is not limited to curtailing spam or thwarting identity theft, for example.

(2) Consumers, not network operators, must be allowed to continue to choose **winners** and losers in the content and applications marketplace.



justice, *wealth* : 000**101**

(3) Make sure there is always a fertile place for all of our **good ideas** to flourish."



00**1000**: *innovation*

(4) That was, I believe, the first time that **idea** had been presented to this Committee.



(no value) : 000000

$\chi = \{000000, 00000**1**, 0000**11**, 000**100**, 000**101**, 000**110**, \dots, 10**1000**, 110000\}$ # $\chi = 22$

μ_0 μ_1 μ_2 at-most-2 values : constraint μ_{21}

Model of LVM (*Latent Value Model*)

word-level values (a kind of topics) } (← unobserved)
context indicators

word sequence = sentence (← observed)

parameters for $\mathbf{y}$

parameters for $\mathbf{x}$

$$P(\mathbf{x}, \mathbf{y} | \mathbf{w}, \theta, \phi)$$

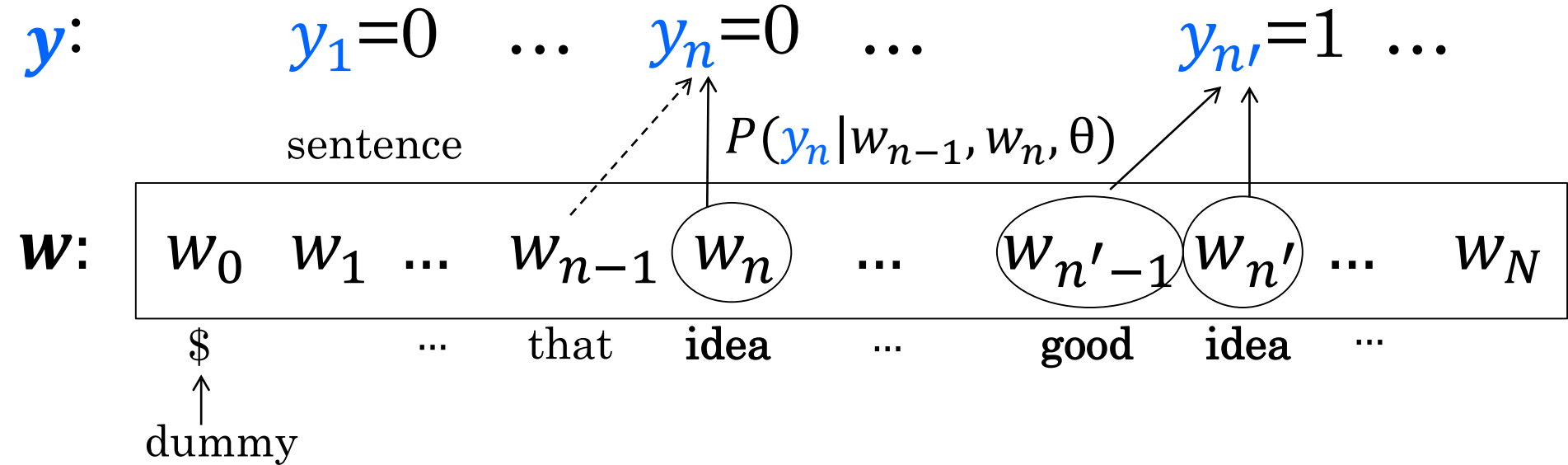
a recognition model

$$= \prod_{n=1}^N P(\mathbf{y}_n | w_{n-1}, w_n, \theta) \times P(\mathbf{x}_n | \mathbf{y}_n, w_{n-1}, w_n, \phi)$$

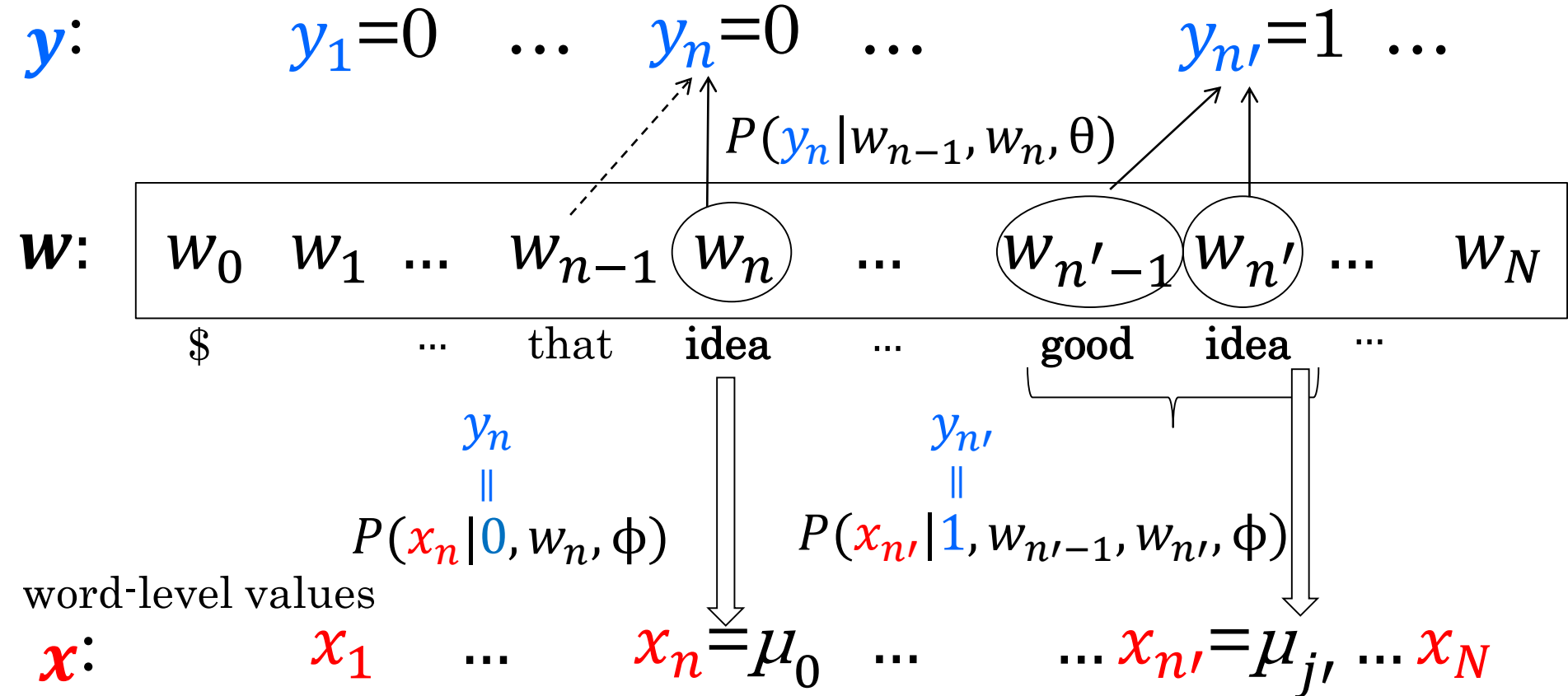
context : influence of the previous word

Generative Process of LVM

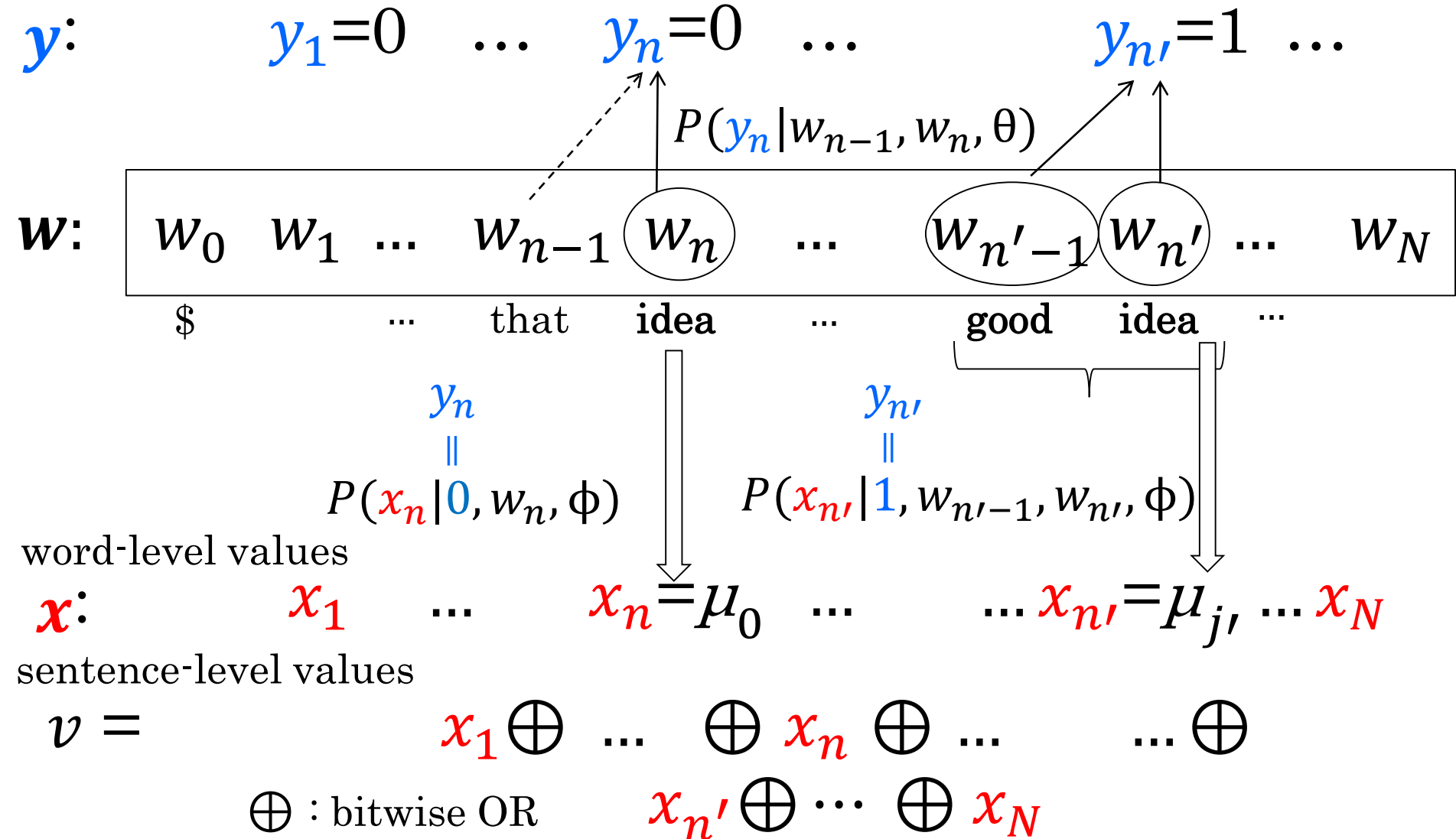
context indicators



Generative Process of LVM



Generative Process of *LVM*



Intrinsic Evaluation

102-fold document cross-validation (train: 100 dev-test: 1, test: 1)

Method	Precision	Recall	F_1
SVM (w)	0.778	0.664	0.717
SVM (w,b)	0.754	0.681	0.715
sLDA	0.688	0.459	0.551
<i>LVM</i> ($y_n=0$)	0.791	0.679	0.731
<i>LVM</i>	0.787	0.690	0.737

+3%
(relative)

(micro-averaged: w/o *honor*)

SVM (w) : 2nd-degree poly kernel using word features

SVM (w, b) : linear kernel using word and bigram features

LVM ($y_n=0$) : proposed model w/o context

Comparison with Human 2nd annotator in F_1

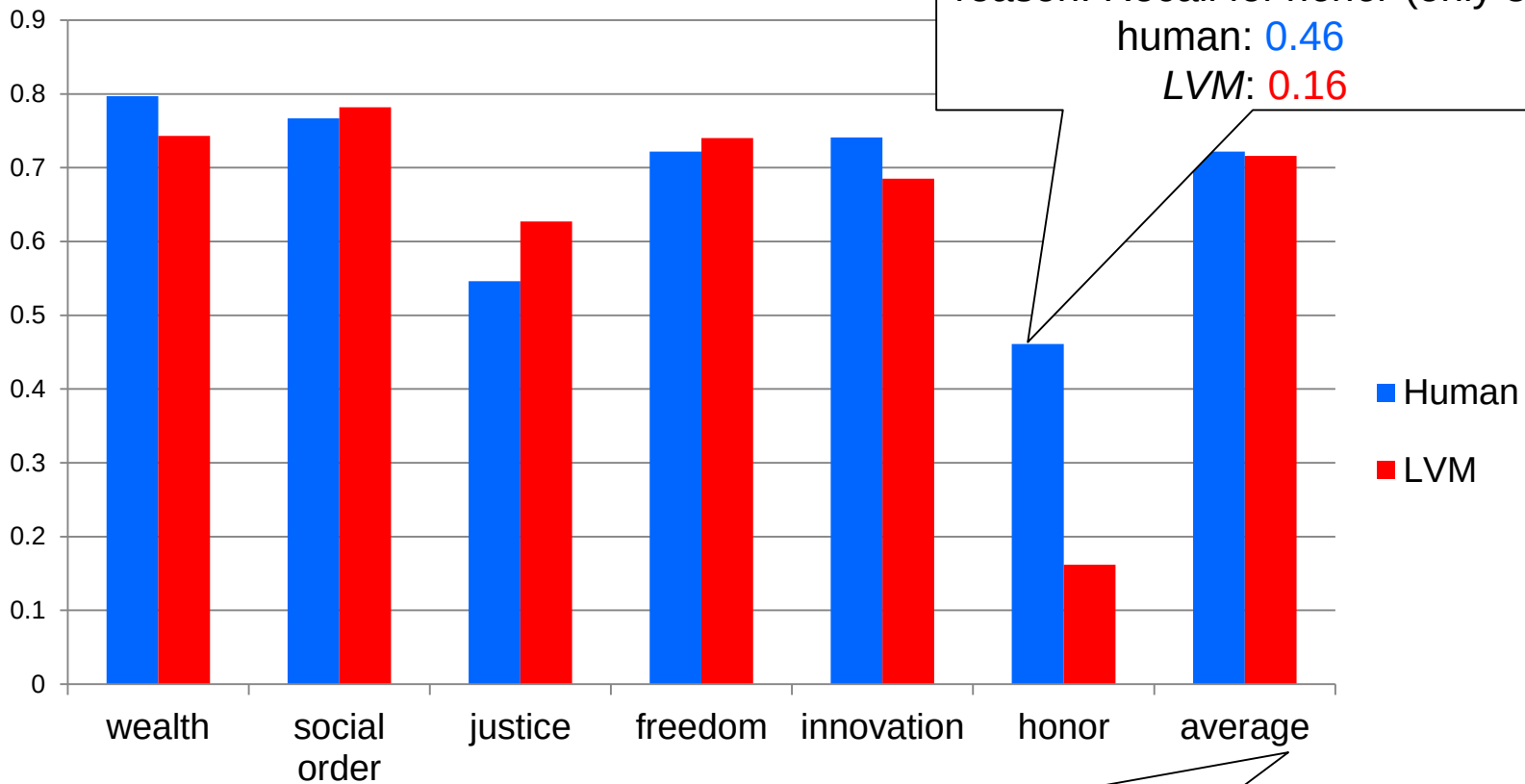
(train: 82 documents, test: 20 documents)

as a “pseudo-classifier”

reason: Recall for *honor* (only 317 occ.)

human: 0.46

LVM: 0.16



Equality between “human” and *LVM*
cannot be rejected as statistically significant by z-test

micro-Average F_1 :

human: 0.722

LVM: 0.715

Data Science Research Questions

- How much human effort can be saved?
 - An “extrinsic” measure
- Do improvements in intrinsic measures predict improvements in extrinsic measures?

freedom

id	PRO	id	CON
40	0.446	52	0.480
42	0.392	55	0.234
45	0.281	56	0.038
46	0.375	59	0.113
47	0.233	63	0.357
49	0.636	64	0.113
50	0.385	68	0.200
54	0.511	69	0.441
57	0.250	71	0.483
58	0.125	76	0.393
60	0.154	78	0.245
61	0.450	84	0.270
62	0.273	89	0.183
65	0.625	90	0.145
66	0.143	92	0.136
67	0.400	93	0.146
70	0.465	95	0.100
72	0.257	96	0.250
73	0.300	98	0.520
74	0.386	99	0.320
75	0.430	100	0.214
77	0.346		
79	0.544		
80	0.464		

PRO for Net Neutrality: 55 docs
CON for Net Neutrality: 40 docs

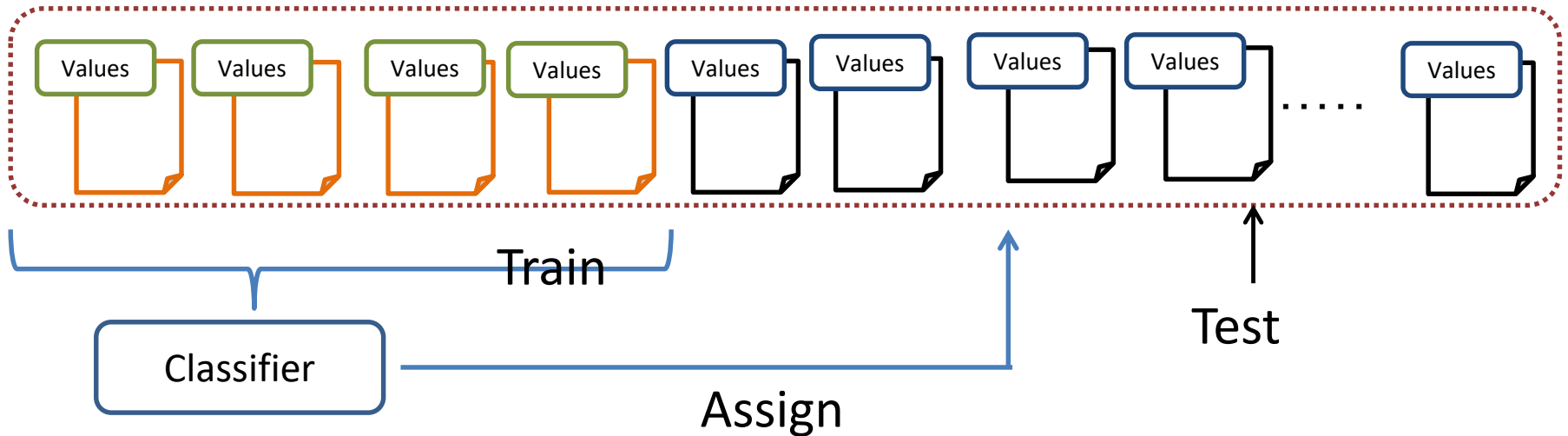
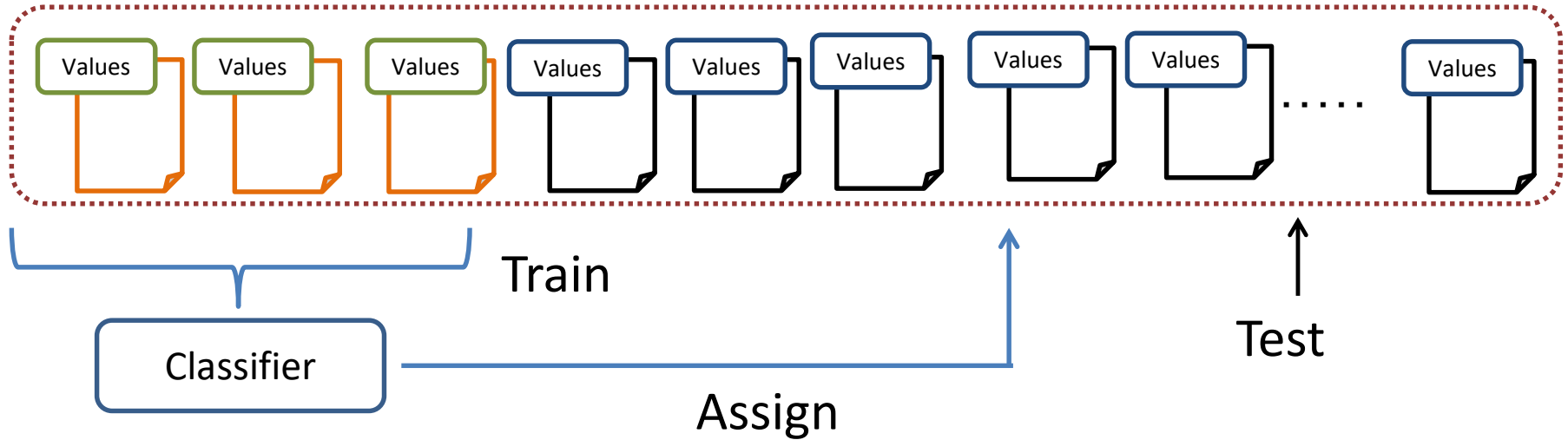
Compute frequency of sentences with each human value per document

Mann-Whitney U test to detect significant difference in relative frequency between PRO and CON for each human value

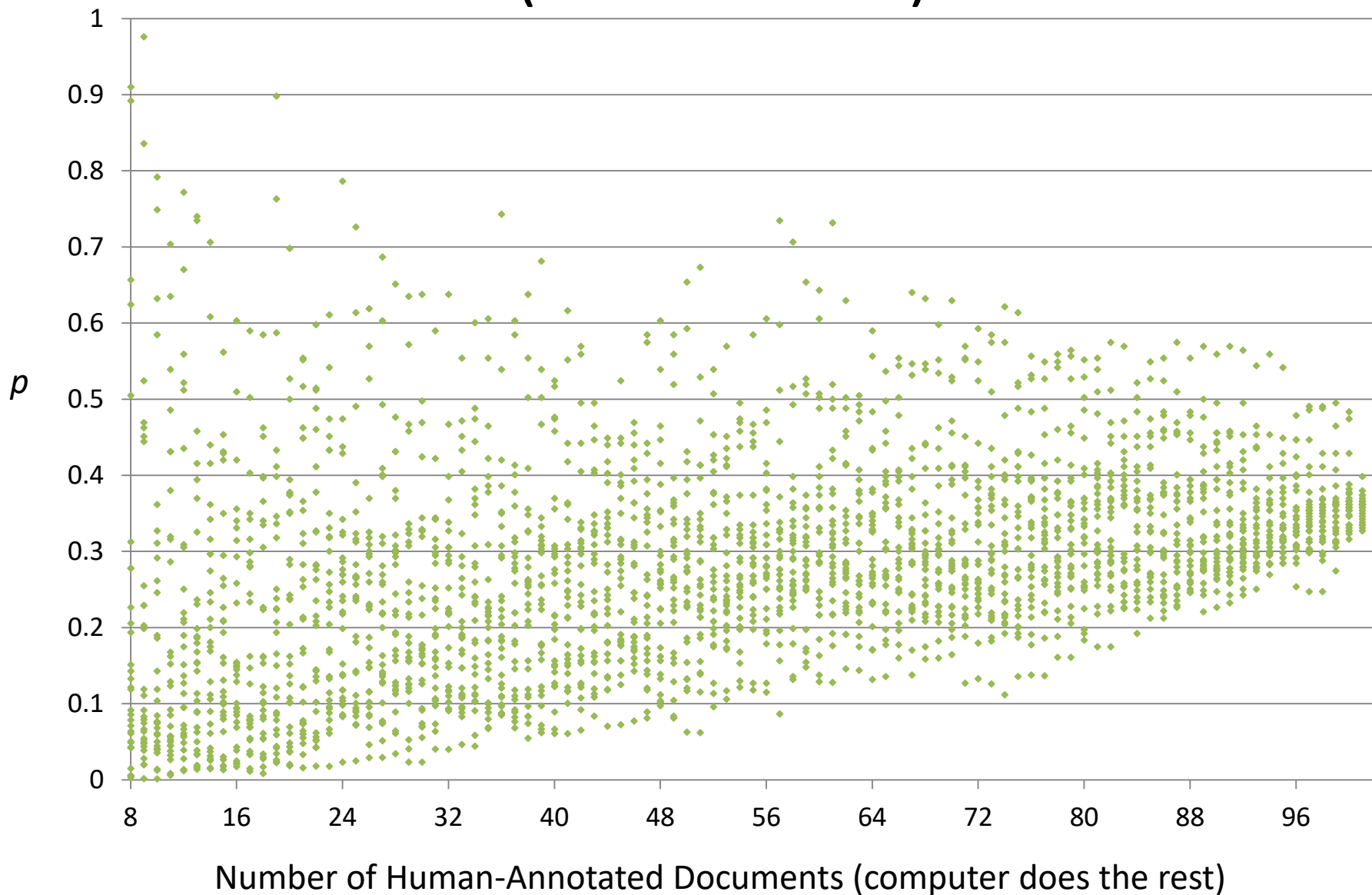
Key idea:

- Code some docs by hand
- Train a classifier
- Use the classifier to code the rest

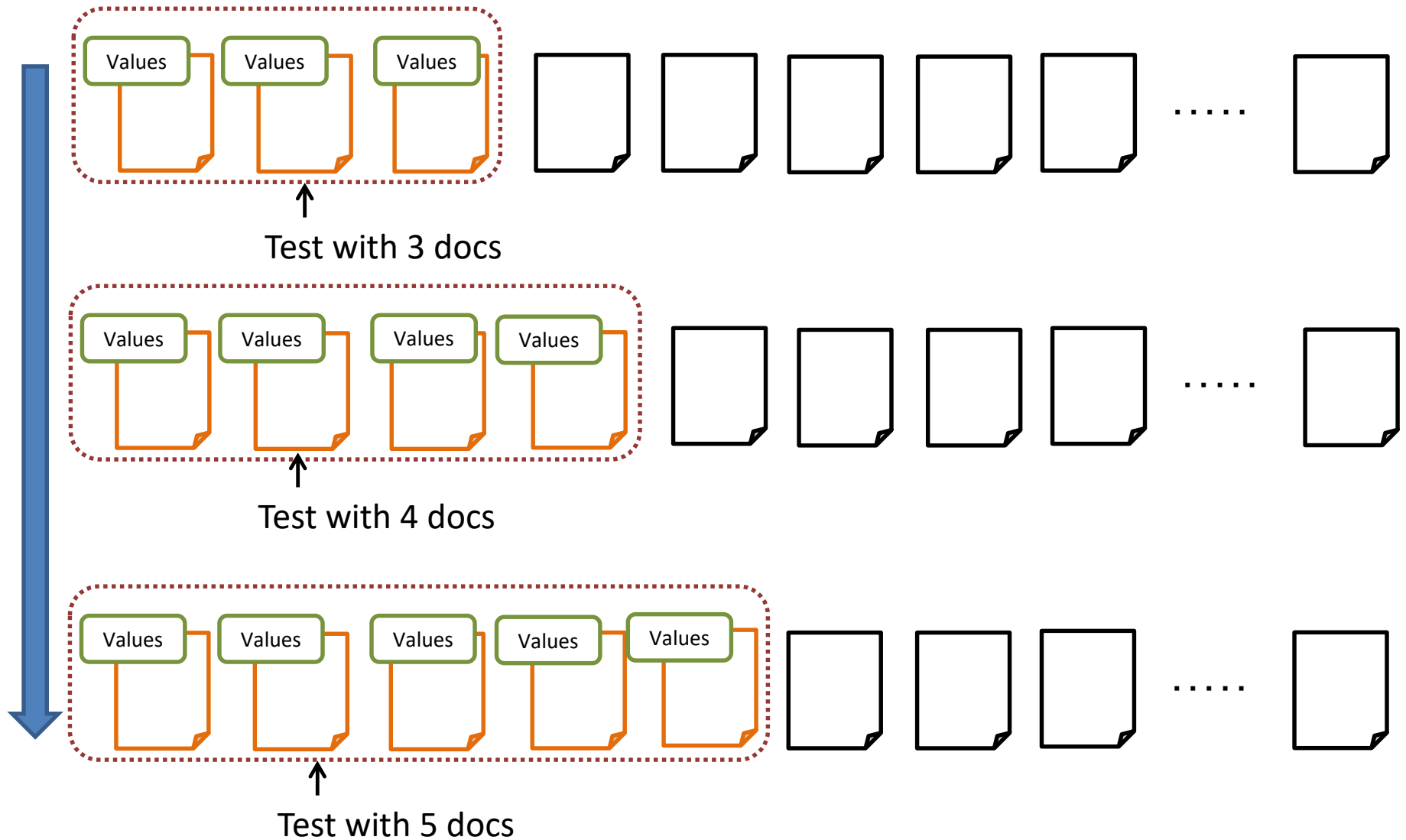
Human + Computer Annotations



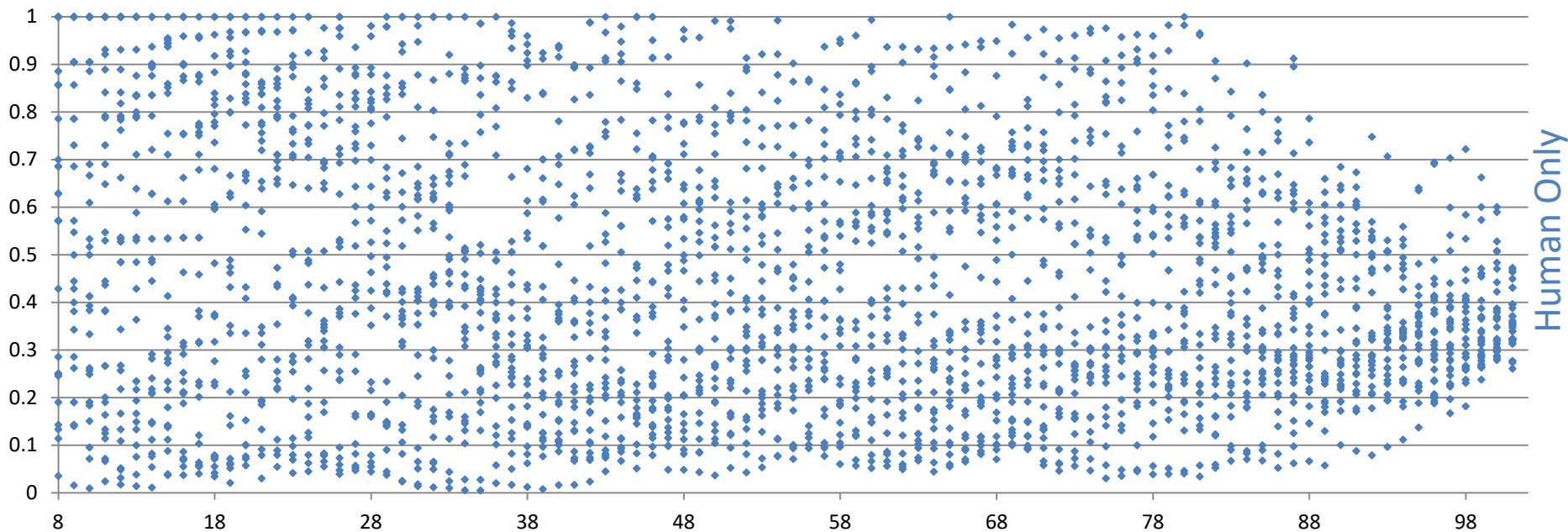
Human + Computer Annotations (Social-Order)



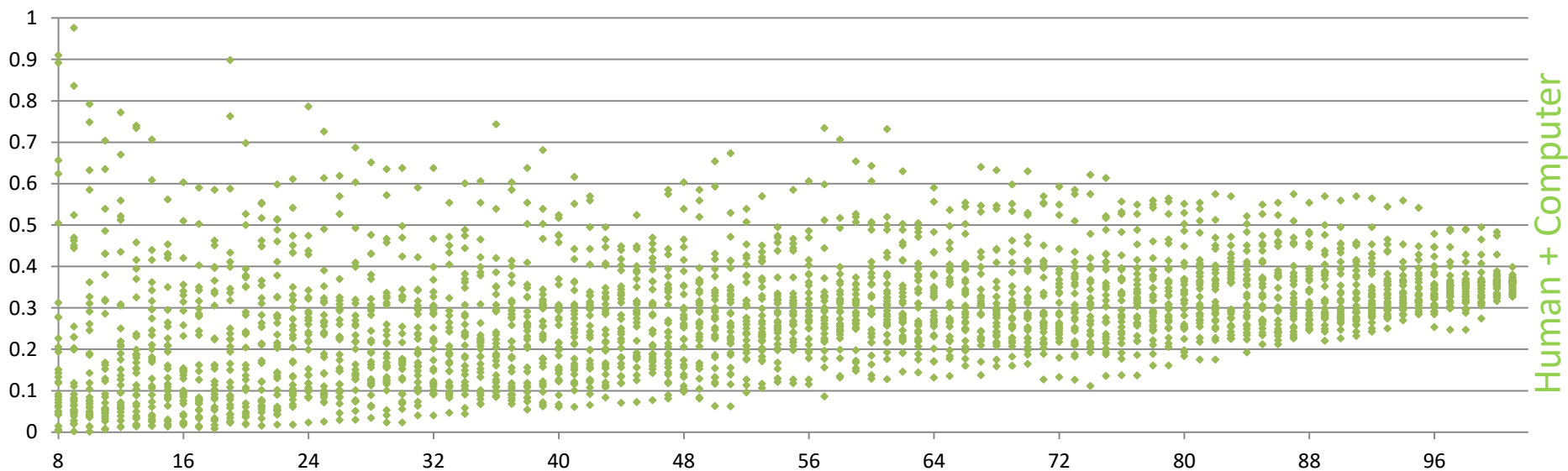
Partial Human Annotation Only



Distribution of p Values (Social Order)

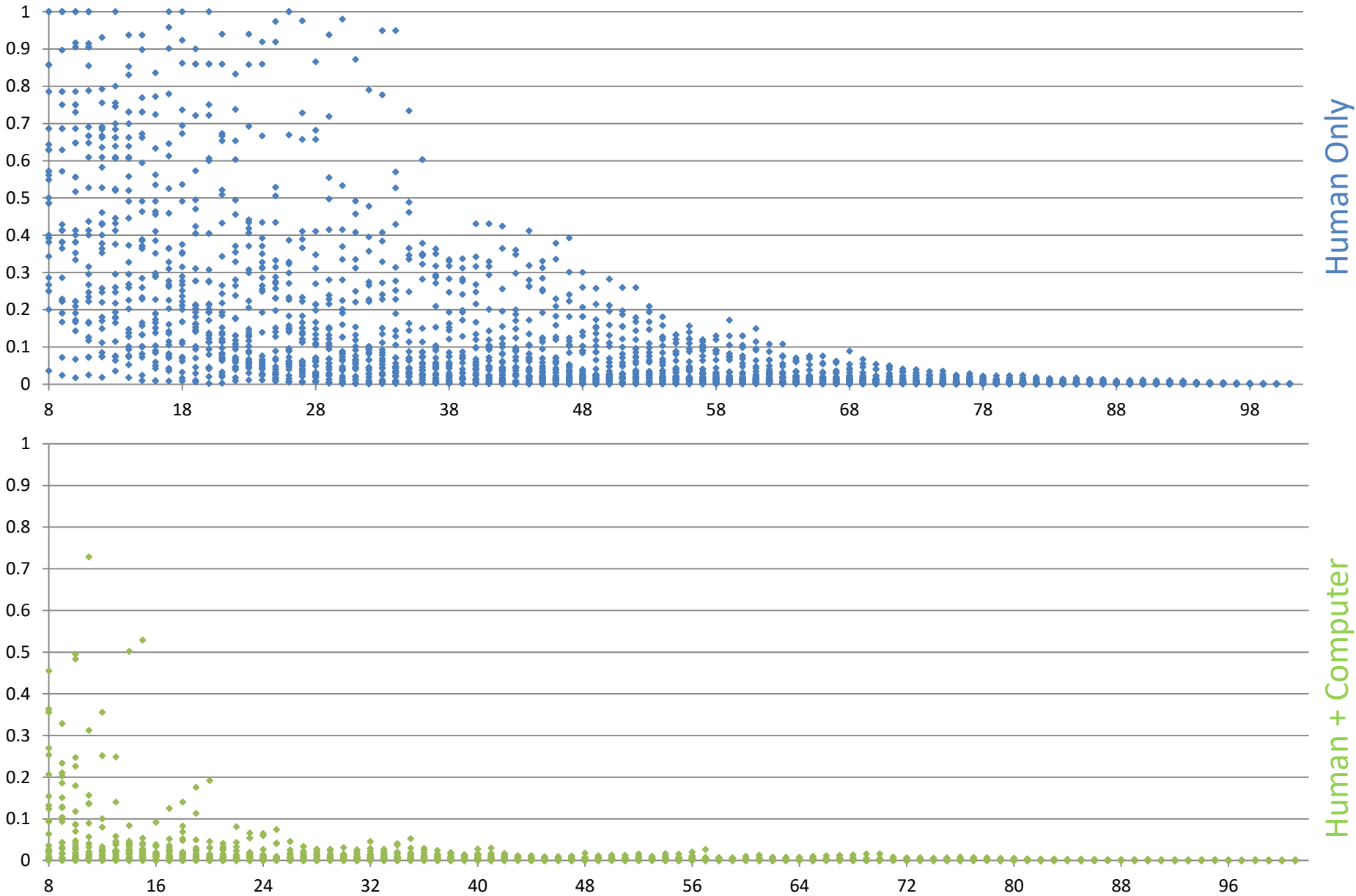


Human Only

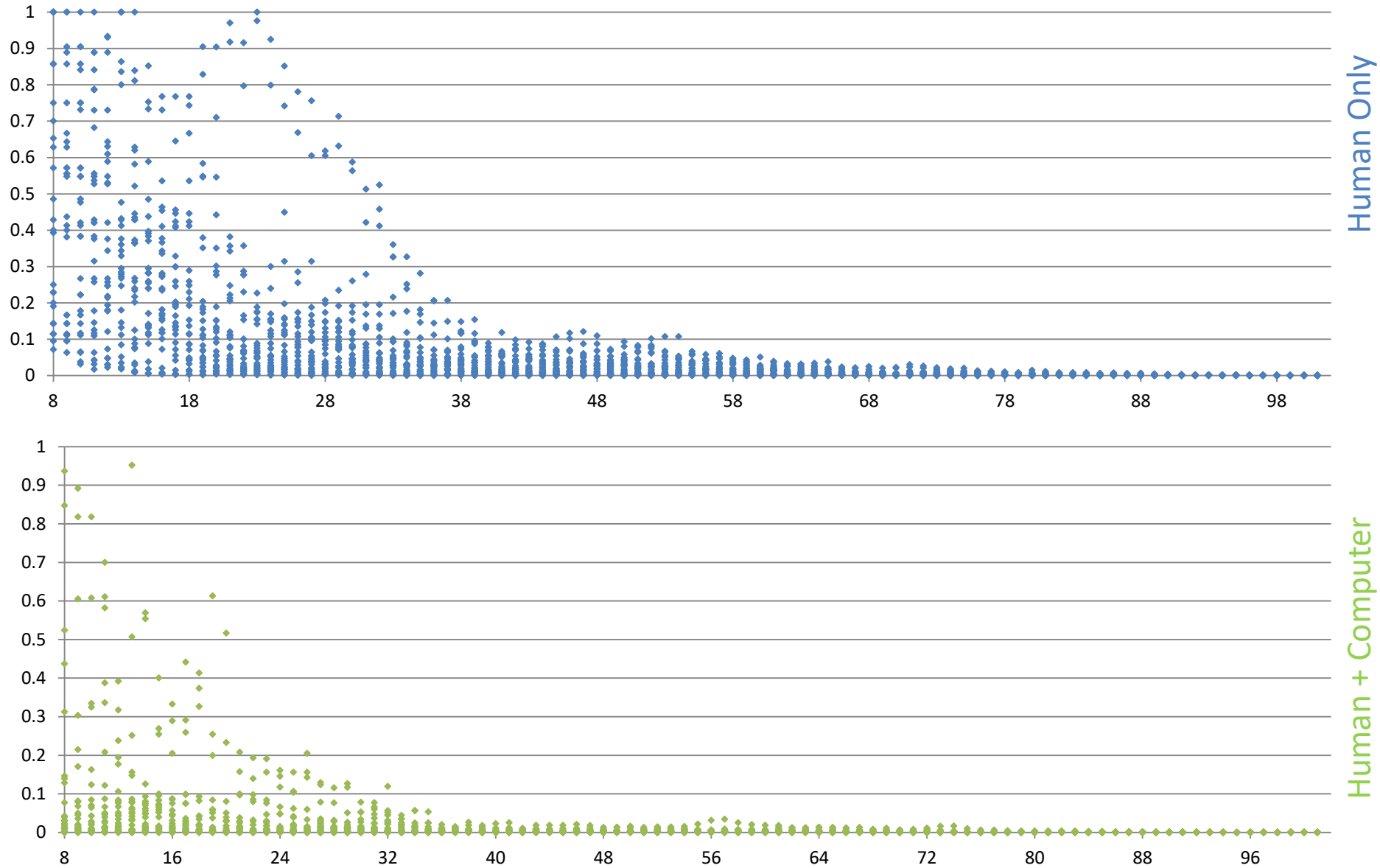


Human + Computer

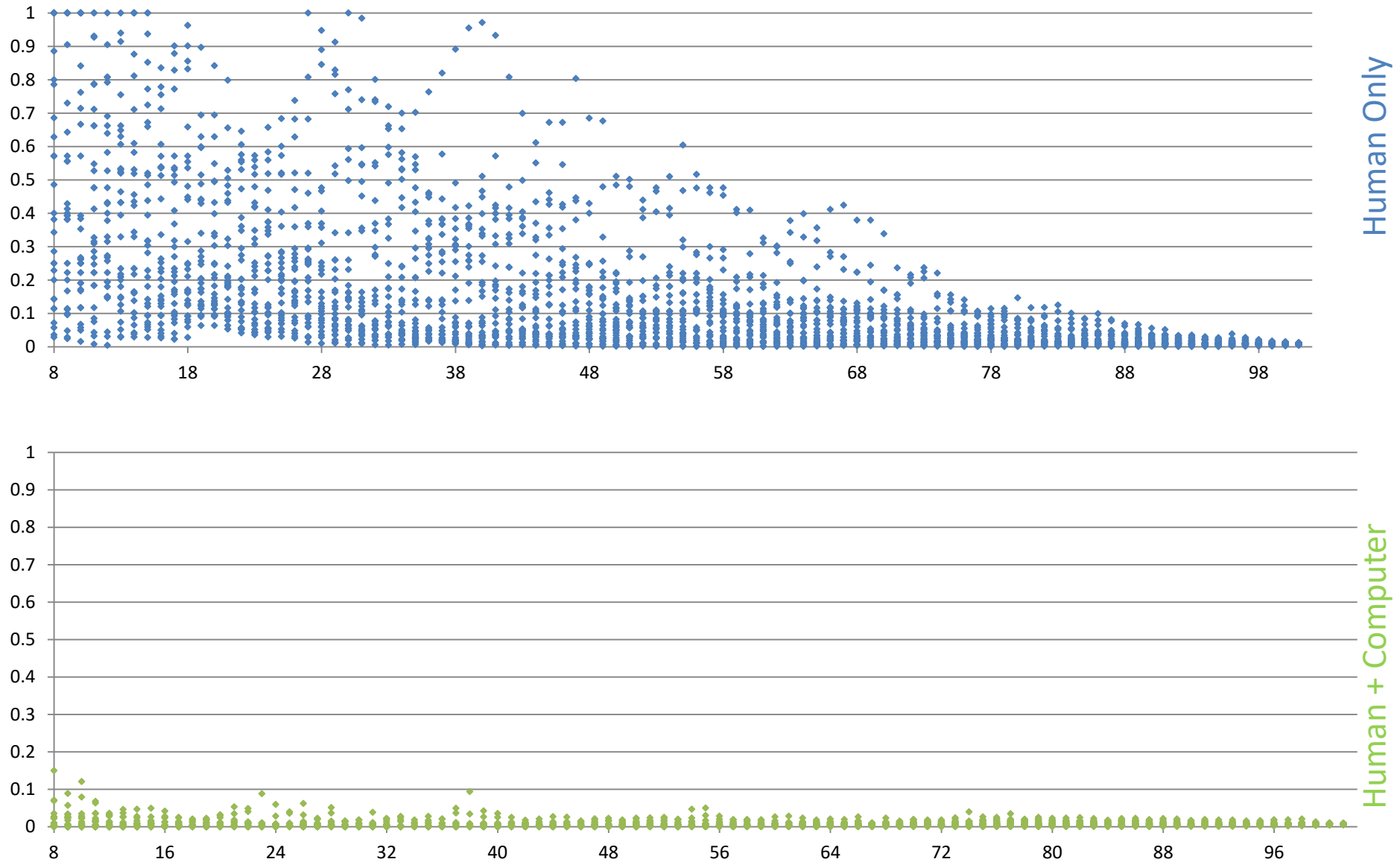
Distribution of p Values (Freedom)



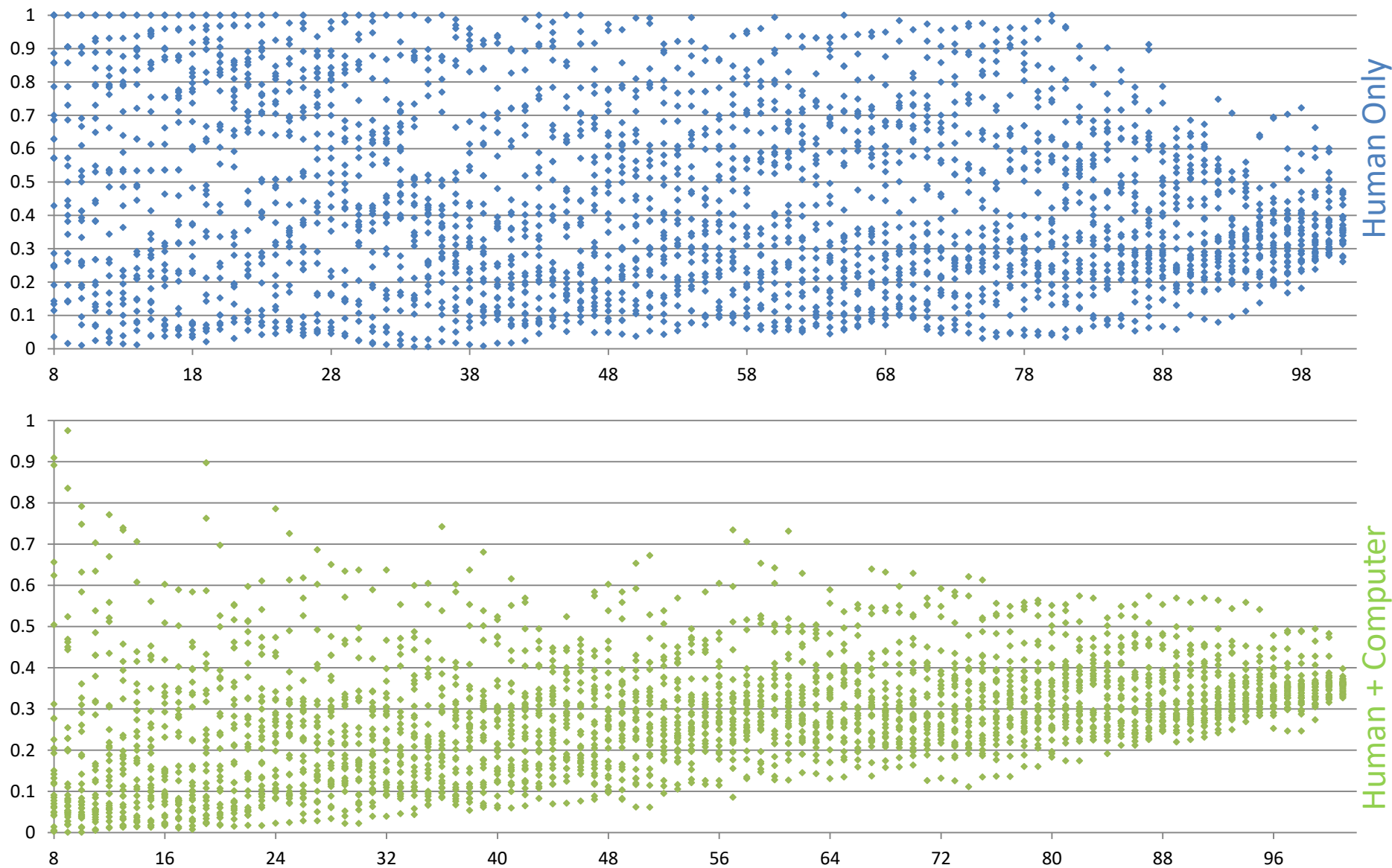
Distribution of p Values (Justice)



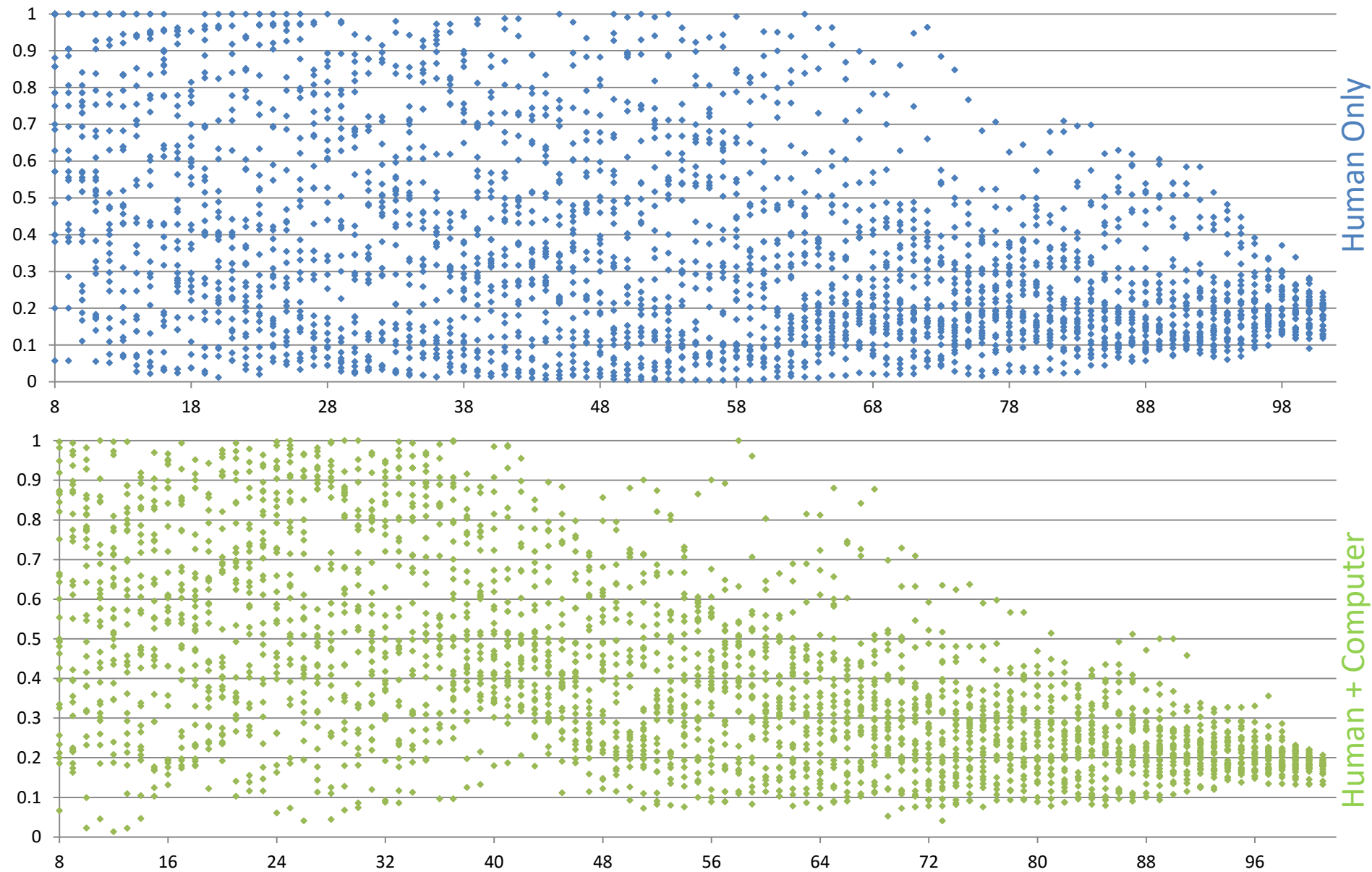
Distribution of p Values (Wealth)



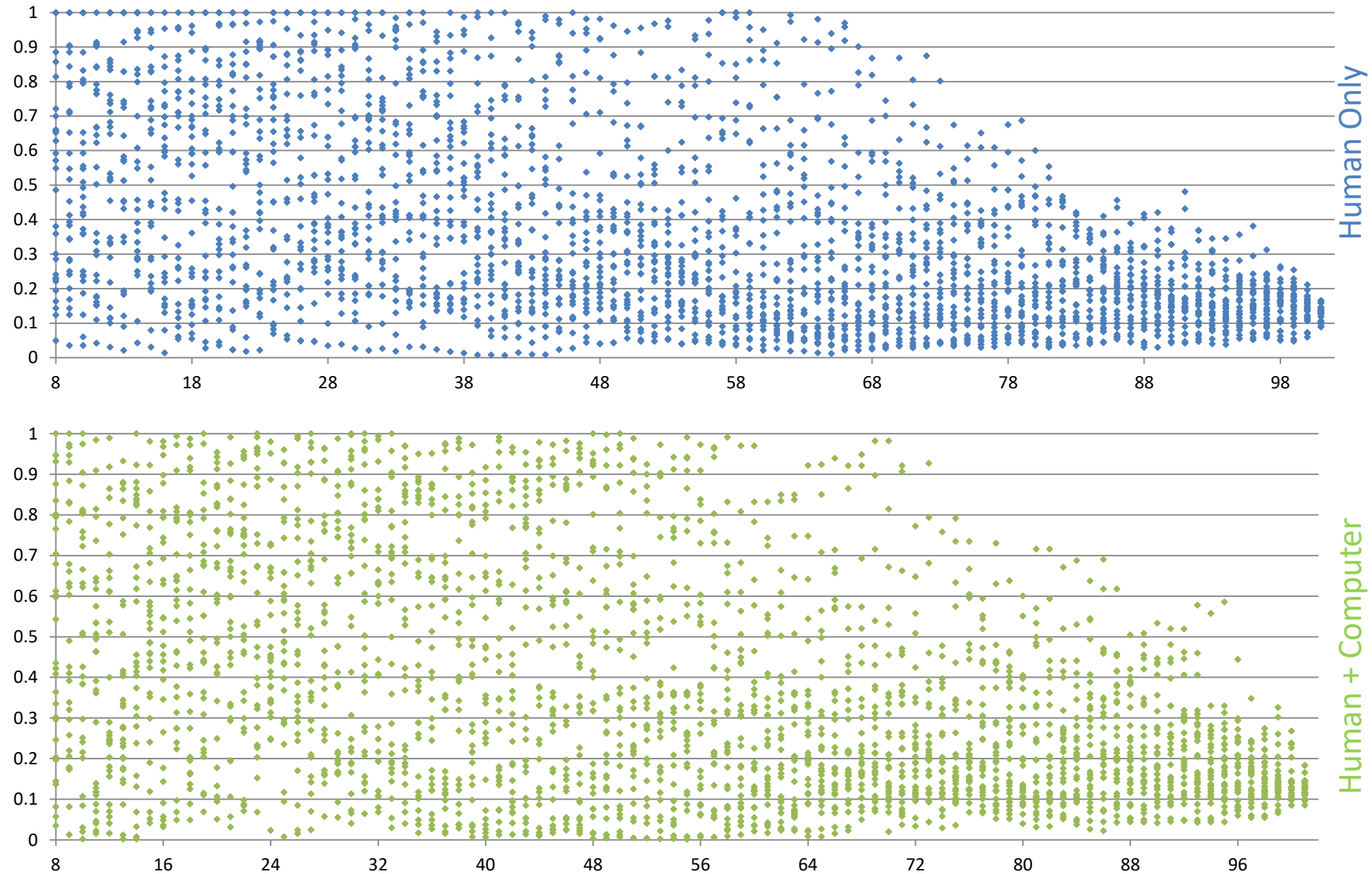
Distribution of p Values (Social Order)



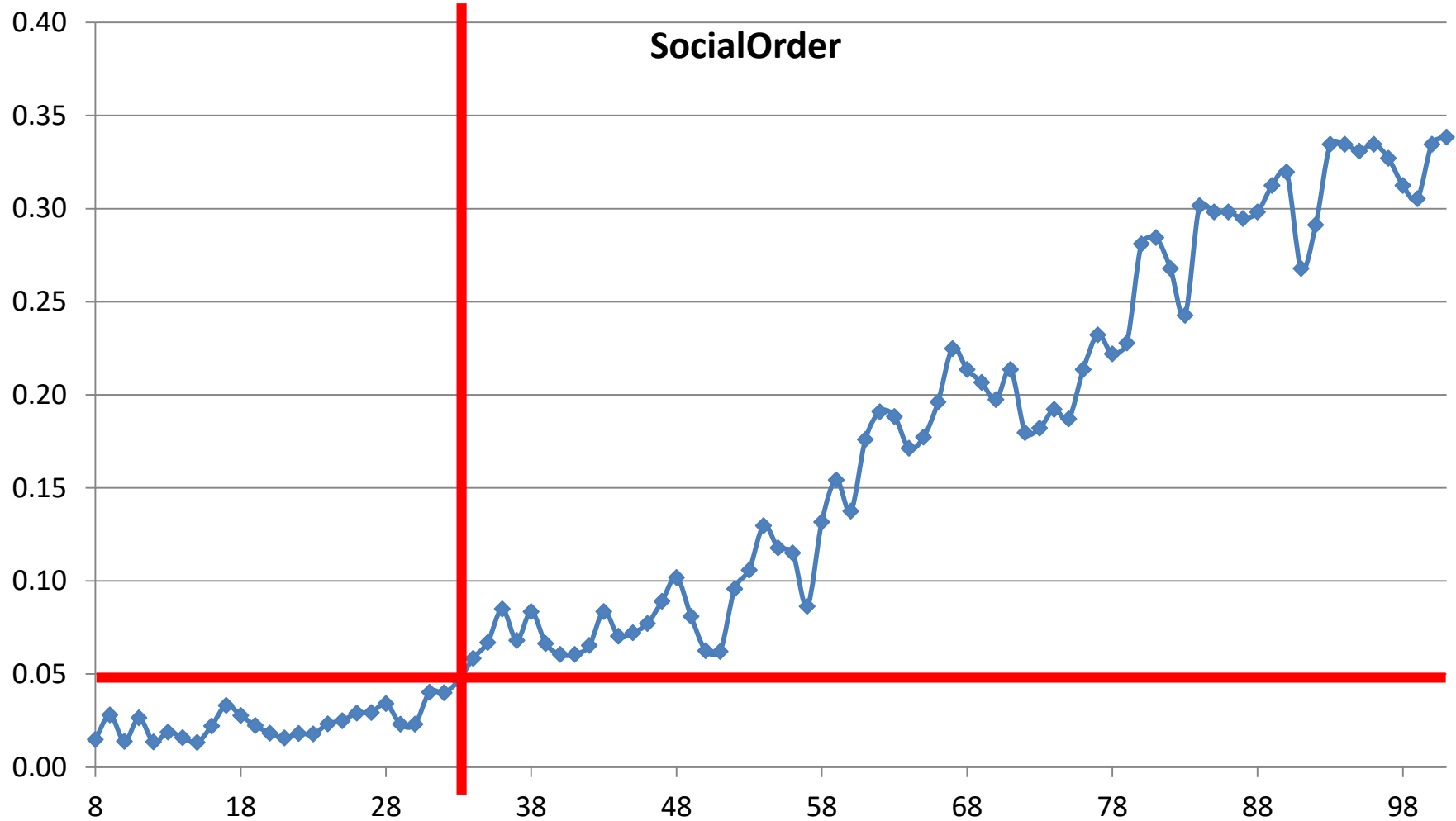
Distribution of p Values (Innovation)



Distribution of p Values (Honor)



One Shuffling (Social Order)



Extrinsic Evaluation

freedom	svm	lvm	knn	H
Average	13	13	43	45
SD	7	9	19	16
95% Conf	26	31	79	77
F ₁	0.705	0.729	0.526	

social-order	svm	lvm	knn	H
Average	14	15	21	22
SD	8	7	22	25
95% Conf	30	28	64	70
F ₁	0.794	0.798	0.470	

justice	svm	lvm	knn	H
Average	17	19	40	36
SD	10	13	20	15
95% Conf	36	45	79	66
F ₁	0.626	0.645	0.393	

innovation	svm	lvm	knn	H
Average	11	9	32	27
SD	13	4	33	26
95% Conf	36	18	97	78
F ₁	0.657	0.674	0.484	

wealth	svm	lvm	knn	H
Average	13	11	71	67
SD	11	12	30	19
95% Conf	34	35	130	103
F ₁	0.756	0.765	0.494	

honor	svm	lvm	knn	H
Average	39	45	29	32
SD	33	32	31	34
95% Conf	103	108	89	100
F ₁	0.208	0.363	0.242	

Some Topics for Discussion

- How would a social scientist know when to stop labeling and let the machine take over?
- Repeated testing threatens validity of p values
- Correlated samples threatens validity of later stopping points
- Single case study is not yet generalizable

- Fleischmann, K.R., Takayama, Y., Cheng, A.-S., Tomiura, Y., Oard, D.W., and Ishita, E. (2015). Thematic analysis of words that invoke values in the net neutrality debate. *iConference 2015*.
- Fleischmann, K.R. (2014). *Information and Human Values*. San Rafael, CA: Morgan & Claypool.
- Fleischmann, K.R., Oard, D.W., Cheng, A.-S., Wang, P., & Ishita, E. (2009). Automatic classification of human values: Applying computational thinking to information ethics. *ASIS&T 2009*.
- Cheng, A.-S. (2012). *Values in the Net Neutrality Debate: Applying Content Analysis to Testimonies from Public Hearings*. Doctoral Thesis, University of Maryland.
- Cheng, A.-S., Fleischmann, K.R., Wang, P., Ishita, E., & Oard, D.W. (2012). The role of innovation and wealth in the net neutrality debate: A content analysis of human values in Congressional and FCC hearings. *Journal of the American Society for Information Science and Technology*, 63(7), 1360-1373.
- Cheng, A.-S. & Fleischmann, K.R. (2010). Developing a meta-inventory of human values. *ASIS&T 2010*.
- Oard, D.W. (2009). A whirlwind tour of automated language processing for the humanities and social sciences. In *Promoting Digital Scholarship: Formulating Research Challenges in the Humanities, Social Sciences, and Computation*, Council on Library and Information Resources, pp. 34-42.
- Ishita, E., Oard, D.W., Fleischmann, K.R., Cheng, A.-S., & Templeton, T.C. (2010). Investigating multi-label classification for human values. *ASIS&T 2010*.
- Takayama, Y. (2015). *Automatic Estimation of Human Values Using Supervised Machine Learning*. Doctoral Thesis, Kyushu University.
- Takayama, Y., Tomiura, Y., Ishita, E., Oard, D.W., Fleischmann, K.R., & Cheng, A.-S. (2014). A word-scale probabilistic latent variable model for detecting human values. *CIKM 2014*.
- Takayama, Y. Tomiura, Y, Ishita, E., Wang, Z., Oard, D.W., Fleischmann K.R. & Cheng, A.-S. (2013). Improving automatic sentence-level annotation of human values using augmented feature vectors. *PACLING 2013*.