A Systematic Review on the Practice of Evaluating Visualization

Tobias Isenberg, Petra Isenberg, Jian Chen,
Michael Sedlmair, and Torsten Möller

Motivation

- research question: state of evaluation work in visualization?

- most common evaluation goals/methods?
- evaluation of what part of visualization process?
- evaluation done similarly in different sub-areas of visualization?
- history and current trends?

Contributions

1. classification of evaluation use in “scientific visualization”
2. historical perspective of evaluation in visualization
3. considerations for improvement of evaluation in visualization

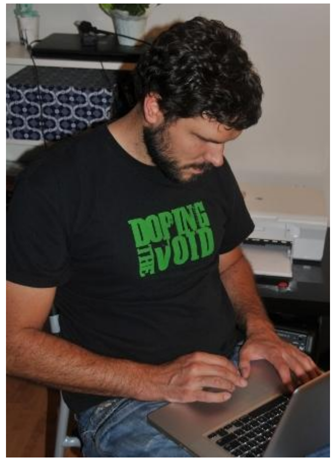
Our approach

- literature review of IEEE Visualization/Scientific Visualization
- 581 papers

1997 2000 2003 2006 2007 2008 2009 2010 2011 2012

- coding by the 5 co-authors




Related work

- Lam et al. [2012]: state of evaluation in “information visualization”
  - 850 papers of 1995–2010
  - InfoVis, VAST, EuroVis, Information Visualization Journal
  - 7 scenarios of evaluation goals

Empirical Studies in Information Visualization: Seven Scenarios

Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale

Abstract—We take a new, scenario-based look at evaluation in information visualization. Our seven scenarios, evaluating visual data analysis and reasoning, evaluating user performance, evaluating user experience, evaluating environments and work practices, evaluating communication through visualization, evaluating visualization algorithms, and evaluating collaborative data analysis were derived through an extensive literature review of over 800 visualization publications. These scenarios distinguish different study goals and types of research questions and are illustrated through example studies. Through this broad survey and the distillation of these scenarios, we make two contributions. One, we encapsulate the current practices in the information visualization research community and, two, we provide a different approach to reaching decisions about what might be the most effective evaluation of a given information visualization. Scenarios can be used to choose appropriate research questions and goals and the provided examples can be consulted for guidance on how to design one's own study.

Index Terms—Information visualization, evaluation.

1 INTRODUCTION

EVALUATION in information visualization is complex since, for a thorough understanding of a tool, it not only involves assessing the visualizations themselves, but also the complex processes that a tool is meant to support. Examples of such processes are exploratory data analysis and reasoning, communication through visualization, or collaborative data analysis. Researchers and practitioners in the field have long identified many of the challenges faced when planning, conducting, and executing an evaluation of a visualization tool or system [10], [41], [54], [63]. It can be daunting for evaluators to identify the right evaluation questions to ask, to choose the right variables to evaluate, to pick the right tasks, users, or data sets to test, and to pick appropriate evaluation methods. Literature guidelines exists that can help with these problems but they are almost exclusively focused on methods—“structured as an enumeration of methods with focus on *how* to carry them out, without prescriptive advice for *when* to choose between them.” ([54, p.1], author's own emphasis).

This paper takes a different approach: instead of focusing on evaluation methods, we provide an in-depth discussion of evaluation scenarios, categorized into those for understanding data analysis processes and those which evaluate visualizations themselves.

The scenarios for understanding data analysis are

- Understanding environments and work practices (UWP),
- evaluating visual data analysis and reasoning (VDAR),
- evaluating communication through visualization (CTV), and
- evaluating collaborative data analysis (CDA).

The scenarios for understanding visualizations are

- Evaluating user performance (UP),
- evaluating user experience (UE), and
- evaluating visualization algorithms (VA).

Our goal is to provide an overview of different types of evaluation scenarios and to help practitioners in setting the right evaluation goals, picking the right questions to ask, and to consider a variety of methodological alternatives to evaluation for the chosen goals and questions. Our scenarios were derived from a systematic analysis of 850 papers (361 with evaluation) from the information visualization research literature (Section 5). For each evaluation scenario, we list the most common evaluation goals and outputs, evaluation questions, and common approaches in Section 6. We illustrate each scenario with representative published evaluation examples from the information visualization community. In cases where there are gaps in our community's evaluation approaches, we suggest examples from other fields. We strive to provide a wide coverage of the methodology space in our scenarios to offer a diverse set of evaluation options. Yet, the “Methods and Examples” lists in this paper are not meant to be comprehensive as our focus is on choosing among evaluation scenarios. Instead, we direct the interested reader

- *H. Lam is with Google, Inc, Mountain View, CA. E-mail: heidi.lam@gmail.com.*
- *E. Bertini is with the Department of Computer and Information Science, University of Konstanz, Box 78, Konstanz 78457, Germany. E-mail: enrico.bertini@uni-konstanz.de.*
- *P. Isenberg is with INRIA, Université Paris-Sud, Team Aviz, Bat 650, Saclay, Orsay Cedex 91405, France. E-mail: petra.isenberg@inria.fr.*
- *C. Plaisant is with the University of Maryland, 2117C Hornbake South Wing, College Park, MD 20742. E-mail: plaisant@cs.umd.edu.*
- *S. Carpendale is with the Department of Computer Science, University of Calgary, 2500 University Dr. NW, Calgary, AB T2N 1N4, Canada. E-mail: sheelagh@ucalgary.ca.*

Manuscript received 8 Sept. 2010; revised 6 Nov. 2011; accepted 9 Nov. 2011; published online 30 Nov. 2011.
Recommended for acceptance by C. North.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2010-09-0224.
Digital Object Identifier no. 10.1109/TVCG.2011.279.

Coding scheme (based on Lam et al. [2012])

- evaluating communication through visualization
- evaluating collaborative data analysis

Coding scheme (based on Lam et al. [2012])

- evaluating communication through visualization
- evaluating collaborative data analysis
- understanding work practices
- visual data analysis and reasoning

Coding scheme (based on Lam et al. [2012])

- evaluating communication through visualization
- evaluating collaborative data analysis
- understanding work practices
- visual data analysis and reasoning
- user performance
- user experience

Coding scheme (based on Lam et al. [2012])

- evaluating communication through visualization
- evaluating collaborative data analysis
- understanding work practices
- visual data analysis and reasoning
- user performance
- user experience
- algorithmic performance (was visualization algorithms)

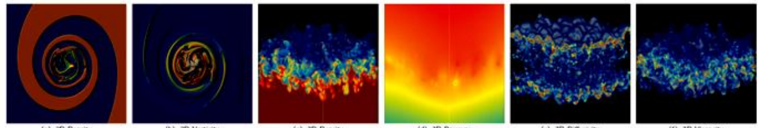
Coding scheme (based on Lam et al. [2012])

- evaluating communication through visualization
- evaluating collaborative data analysis
- understanding work practices
- visual data analysis and reasoning
- user performance
- user experience
- algorithmic performance (was visualization algorithms)
- qualitative result inspection

Results: evaluation scenarios (selection)

algorithmic performance

35% of scenarios



(a) 2D Density (b) 2D Vorticity (c) 3D Density (d) 3D Pressure (e) 3D Diffusivity (f) 3D Viscocity

Fig. 1. Visualizations of 2D data (as pseudocolored height fields) and 3D data (volume rendered) used in our experiments.

data set								compressed size (MB) and compression time (seconds)									
name	unique (%)	entropy (bits)	range (bits)	min	max	size (MB)	time (sec)	zlib		[RKB2006]		[EFF2000]		[ILS2005]		new scheme	
m2d density	3.89	3.49	21.83	8.7E−01	1.2E+00	19.6	0.71	1.6	0.86	4.3	**0.49**	4.4	0.56	**1.3**	1.08	1.3	0.56
m2d vorticity	99.20	22.25	31.05	-1.4E+02	2.5E+01	19.6	0.71	18.4	2.14	**11.8**	**1.21**	15.5	1.29	12.9	2.22	13.8	1.49
m3d density	7.67	5.16	23.60	1.0E+00	3.0E+00	364.5	12.81	50.4	17.55	100.5	9.06	96.3	**8.48**	35.7	19.03	**35.5**	9.25
m3d pressure	27.29	23.91	31.06	-3.7E+00	2.3E+03	364.5	12.80	229.2	99.76	95.6	9.31	87.9	**8.87**	**40.1**	18.79	40.4	9.96
m3d diffusivity	36.87	23.19	30.02	0.0E+00	6.8E+00	364.5	12.68	297.6	42.90	250.8	19.09	239.3	**15.02**	**198.8**	31.92	203.0	18.47
m3d viscocity	50.07	24.86	28.59	8.6E−15	2.9E+00	364.5	12.62	314.0	46.09	249.4	18.95	246.1	**14.68**	209.2	32.66	**207.5**	19.45
h3d temp	65.70	23.54	31.56	-7.7E+01	1.0E+35	95.4	3.77	75.8	14.56	59.3	4.64	53.0	**4.27**	**44.1**	8.04	44.1	5.06
h3d pressure	81.82	24.13	31.58	-3.4E+03	1.0E+35	95.4	3.78	82.3	12.00	64.3	5.14	52.9	**4.87**	**45.0**	7.78	45.2	5.34
h3d x velocity	84.18	24.18	31.55	-5.3E+01	1.0E+35	95.4	3.89	86.1	11.27	67.4	6.22	63.3	**4.59**	**54.5**	8.86	55.4	5.44
h3d y velocity	84.32	24.18	31.55	-4.6E+01	1.0E+35	95.4	3.83	84.5	11.42	67.1	5.74	62.3	**5.04**	**53.5**	8.64	53.8	5.53
h3d z velocity	86.82	24.24	31.54	-3.2E+00	1.0E+35	95.4	3.87	88.4	10.76	85.6	8.50	76.9	**5.29**	**68.9**	9.83	69.1	6.65
M3d density	40.14	18.84	52.59	1.0E+00	3.0E+00	288.0	11.28	136.8	41.91	160.3	11.63	121.6	**10.94**	-		**105.2**	11.63
M3d pressure	100.00	25.17	63.00	-2.2E+00	2.2E+00	288.0	11.20	272.6	35.18	237.3	**14.91**	225.1	16.59	-		**208.4**	17.20
M3d x velocity	100.00	25.17	63.00	-2.2E+00	2.3E+00	288.0	10.83	275.6	32.30	230.4	**14.73**	215.1	15.91	-		**197.7**	16.84
M3d y velocity	100.00	25.17	63.00	-2.1E+00	2.3E+00	288.0	10.54	275.1	32.19	223.1	**14.27**	215.2	15.16	-		**197.7**	16.65
M3d z velocity	100.00	25.17	63.00	-5.2E+00	9.0E+00	288.0	10.32	275.5	32.62	226.6	**14.74**	213.7	16.05	-		**196.8**	16.14
atom x position	61.10	23.82	31.01	-4.8E−02	4.6E+02	107.7	7.07	84.3	21.18	76.0	7.88	78.8	**7.61**	**67.3**	12.88	68.6	9.07
atom y position	45.90	23.32	26.99	3.7E−02	2.1E+03	107.7	7.08	65.9	30.76	60.4	6.97	56.4	**6.31**	47.0	10.49	**46.9**	7.73
atom z position	61.68	23.84	27.48	9.1E−05	4.6E+02	107.7	7.46	94.6	19.86	82.6	9.00	86.1	**8.25**	**75.7**	13.80	78.2	9.93
atom y velocity	64.65	23.87	30.96	-1.5E−01	1.4E−01	107.7	7.30	95.7	19.88	93.8	10.07	99.1	**9.65**	**84.3**	14.93	87.6	9.92
atom temp	64.91	23.94	27.41	3.0E−03	7.1E+03	107.7	6.69	95.7	19.76	91.6	10.27	95.9	**8.34**	**84.6**	15.02	84.6	10.31
atom energy	3.45	18.57	21.79	-3.6E+00	-2.7E+00	107.7	7.15	77.9	38.59	74.1	7.98	71.8	**7.01**	60.8	12.66	**60.5**	8.30
lucy	61.39	24.38	31.09	-6.1E+02	1.2E+03	160.5	-	137.8	-	99.5	-	90.0	-	**73.6**	-	77.8	-
david$_{1mm}$	25.23	17.08	31.11	-4.4E+03	1.8E+03	322.5	-	144.9	-	155.7	-	163.4	-	**108.6**	-	131.9	-
torso	84.72	18.48	31.08	-2.7E+02	5.8E+02	1.9	-	1.7	-	1.5	-	1.5	-	**1.3**	-	1.3	-
rbl	71.90	20.14	25.99	1.5E+00	3.6E+02	8.4	-	7.1	-	5.8	-	5.6	-	**4.7**	-	4.8	-

Table 1. Compression results for the Miranda (m2d, m3d, M3d) and hurricane (h3d) structured grids, the atom point set, the lucy and david triangle meshes, and the torso and rbl tetrahedral meshes. All data but M3d is represented in single precision. The [ILS2005] scheme operates on single precision only, hence the missing values. For the meshes we report only the compressed size of vertex coordinates; timings are dominated by connectivity coding, and are hence excluded. The range measures (the logarithm of) the number of floating-point values between min and max. Note that the first-order entropy is limited by the number of samples in a data set.

tation.) Arguably such data sets should use an integer rather than floating-point representation, although for simplicity or other reasons it is common practice to use floating-point. Contrary to [16], which entropy codes all bits of the residual, our new coder sacrifices such potential compression gains for speed by storing these repeated low-order bits in raw and uncompressed form. However, the massive data sets from scientific simulation that motivated our work on high-speed compression, as well as our tetrahedral meshes, rarely exhibit significant low-order redundancy, as also evidenced by our results.

5.1.1 Lossy Compression

Fig. 3 shows that our scheme gracefully adapts to decreasing levels of precision when discarding the least significant mantissa (and eventually exponent) bits. For n bits of precision, the schemes [7, 22] require $\log_2 n$ bits to code the number of leading zeros, whereas our scheme exploits the combination of low entropy in the leading-zero count and the elimination of the low-order bits that are most difficult to predict and compress.

5.2 Compression Speed

Fig. 4 shows the speed of compressing from memory to disk, including disk write time. (Because of the simplicity of our method, its decompression speed is similar to its compression speed.) We also include the raw I/O performance of dumping the data uncompressed using a single `fwrite` call. Timings correspond to the median of five runs. Whereas our compressor is slightly slower than the less effective compressors [7, 22], it is nearly twice as fast as [16] while producing similar compression rates. However, in more I/O-intensive scenarios, such as in massively parallel simulations dumping data to the same file system (as is common), the improved compression of our method over [7,22] results in a net gain in effective throughput. We integrated our compression code with Miranda's dump routines and ran performance tests on 256 nodes of LLNL's MCR supercomputer. Achieving on average a lossless reduction of 3.7 on 75 GB of data dumped, the overall dump time was reduced by a factor of 2.7 over writing the data uncompressed.

5.3 Entropy Coding

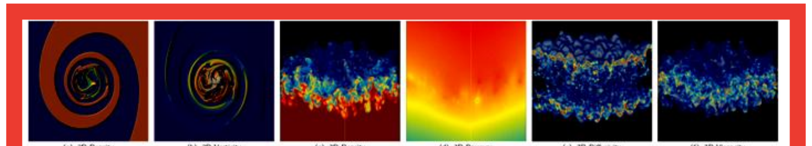
We compared the raw throughput of our range coder and Schindler's [23] by (1) passing raw bytes through it with no compression and (2) entropy coding byte sequences. In both cases, the source data was the uncompressed floating-point data used in our experiments. Timings show that our coder is 40% faster for raw transmission and 28% faster for entropy coding. Meanwhile, the inefficiency of our coder due to loss of precision and range reduction is only 26 bytes of overhead for 1.5 GB of coded data. Its raw throughput is only 20% less than an `fwrite` call, while its entropy coding throughput of 20 MB per second, which includes probability modeling and I/O time, compares favorably with state-of-the-art entropy coders [25].

Linstrom et al. [2006]

Results: evaluation scenarios (selection)

algorithmic performance

35% of scenarios



(a) 2D Density (b) 2D Vorticity (c) 3D Density (d) 3D Pressure (e) 3D Diffusivity (f) 3D Viscosity

Fig. 1. Visualizations of 2D data (as pseudocolored height fields) and 3D data (volume rendered) used in our experiments.

data set								compressed size (MB) and compression time (seconds)									
name	unique (%)	entropy (bits)	range (bits)	min	max	size (MB)	time (sec)	zlib		[RKB2006]		[EFF2000]		[ILS2005]		new scheme	
m2d density	3.89	3.49	21.83	8.7E−01	1.2E+00	19.6	0.71	1.6	0.86	4.3	**0.49**	4.4	0.56	**1.3**	1.08	1.3	0.56
m2d vorticity	99.20	22.25	31.05	-1.4E+02	2.5E+01	19.6	0.71	18.4	2.14	**11.8**	**1.21**	15.5	1.29	12.9	2.22	13.8	1.49
m3d density	7.67	5.16	23.60	1.0E+00	3.0E+00	364.5	12.81	50.4	17.55	100.5	9.06	96.3	**8.48**	35.7	19.03	**35.5**	9.25
m3d pressure	27.29	23.91	31.06	-3.7E+00	2.3E+03	364.5	12.80	229.2	99.76	95.6	9.31	87.9	**8.87**	**40.1**	18.79	40.4	9.96
m3d diffusivity	36.87	23.19	30.02	0.0E+00	6.8E+00	364.5	12.68	297.6	42.90	250.8	19.09	239.3	**15.02**	**198.8**	31.92	203.0	18.47
m3d viscocity	50.07	24.86	28.59	8.6E−15	2.9E+00	364.5	12.62	314.0	46.09	249.4	18.95	246.1	**14.68**	209.2	32.66	**207.5**	19.45
h3d temp	65.70	23.54	31.56	-7.7E+01	1.0E+35	95.4	3.77	75.8	14.56	59.3	4.64	53.0	**4.27**	**44.1**	8.04	44.1	5.06
h3d pressure	81.82	24.13	31.58	-3.4E+03	1.0E+35	95.4	3.78	82.3	12.00	64.3	5.14	52.9	**4.87**	**45.0**	7.78	45.2	5.34
h3d x velocity	84.18	24.18	31.55	-5.3E+01	1.0E+35	95.4	3.89	86.1	11.27	67.4	6.22	63.3	**4.59**	**54.5**	8.86	55.4	5.44
h3d y velocity	84.32	24.18	31.55	-4.6E+01	1.0E+35	95.4	3.83	84.5	11.42	67.1	5.74	62.3	**5.04**	**53.5**	8.64	53.8	5.53
h3d z velocity	86.82	24.24	31.54	-3.2E+00	1.0E+35	95.4	3.87	88.4	10.76	85.6	8.50	76.9	**5.29**	**68.9**	9.83	69.1	6.65
M3d density	40.14	18.84	52.59	1.0E+00	3.0E+00	288.0	11.28	136.8	41.91	160.3	11.63	121.6	**10.94**	-		**105.2**	11.63
M3d pressure	100.00	25.17	63.00	-2.2E+00	2.2E+00	288.0	11.20	272.6	35.18	237.3	**14.91**	225.1	16.59	-		**208.4**	17.20
M3d x velocity	100.00	25.17	63.00	-2.2E+00	2.3E+00	288.0	10.83	275.6	32.30	230.4	**14.73**	215.1	15.91	-		**197.7**	16.84
M3d y velocity	100.00	25.17	63.00	-2.1E+00	2.3E+00	288.0	10.54	275.1	32.19	223.1	**14.27**	215.2	15.16	-		**197.7**	16.65
M3d z velocity	100.00	25.17	63.00	-5.2E+00	9.0E+00	288.0	10.32	275.5	32.62	226.6	**14.74**	213.7	16.05	-		**196.8**	16.14
atom x position	61.10	23.82	31.01	-4.8E−02	4.6E+02	107.7	7.07	84.3	21.18	76.0	7.88	78.8	**7.61**	**67.3**	12.88	68.6	9.07
atom y position	45.90	23.32	26.99	3.7E−02	2.1E+03	107.7	7.08	65.9	30.76	60.4	6.97	56.4	**6.31**	47.0	10.49	**46.9**	7.73
atom z position	61.68	23.84	27.48	9.1E−05	4.6E+02	107.7	7.46	94.6	19.86	82.6	9.00	86.1	**8.25**	**75.7**	13.80	78.2	9.93
atom y velocity	64.65	23.87	30.96	-1.5E−01	1.4E−01	107.7	7.30	95.7	19.88	93.8	10.07	99.1	**9.65**	**84.3**	14.93	87.6	9.92
atom temp	64.91	23.94	27.41	3.0E−03	7.1E+03	107.7	6.69	95.7	19.76	91.6	10.27	95.9	**8.34**	**84.6**	15.02	84.6	10.31
atom energy	3.45	18.57	21.79	-3.6E+00	-2.7E+00	107.7	7.15	77.9	38.59	74.1	7.98	71.8	**7.01**	60.8	12.66	**60.5**	8.30
lucy	61.39	24.38	31.09	-6.1E+02	1.2E+03	160.5	-	137.8	-	99.5	-	90.0	-	**73.6**	-	77.8	-
david$_{1mm}$	25.23	17.08	31.11	-4.4E+03	1.8E+03	322.5	-	144.9	-	155.7	-	163.4	-	**108.6**	-	131.9	-
torso	84.72	18.48	31.08	-2.7E+02	5.8E+02	1.9	-	1.7	-	1.5	-	1.5	-	**1.3**	-	1.3	-
rbl	71.90	20.14	25.99	1.5E+00	3.6E+02	8.4	-	7.1	-	5.8	-	5.6	-	**4.7**	-	4.8	-

Table 1. Compression results for the Miranda (m2d, m3d, M3d) and hurricane (h3d) structured grids, the atom point set, the lucy and david triangle meshes, and the torso and rbl tetrahedral meshes. All data but M3d is represented in single precision. The [ILS2005] scheme operates on single precision only, hence the missing values. For the meshes we report only the compressed size of vertex coordinates; timings are dominated by connectivity coding, and are hence excluded. The range measures (the logarithm of) the number of floating-point values between min and max. Note that the first-order entropy is limited by the number of samples in a data set.

tation.) Arguably such data sets should use an integer rather than floating-point representation, although for simplicity or other reasons it is common practice to use floating-point. Contrary to [16], which entropy codes all bits of the residual, our new coder sacrifices such potential compression gains for speed by storing these repeated low-order bits in raw and uncompressed form. However, the massive data sets from scientific simulation that motivated our work on high-speed compression, as well as our tetrahedral meshes, rarely exhibit significant low-order redundancy, as also evidenced by our results.

5.1.1 Lossy Compression

Fig. 3 shows that our scheme gracefully adapts to decreasing levels of precision when discarding the least significant mantissa (and eventually exponent) bits. For n bits of precision, the schemes [7, 22] require $\log_2 n$ bits to code the number of leading zeros, whereas our scheme exploits the combination of low entropy in the leading-zero count and the elimination of the low-order bits that are most difficult to predict and compress.

5.2 Compression Speed

Fig. 4 shows the speed of compressing from memory to disk, including disk write time. (Because of the simplicity of our method, its decompression speed is similar to its compression speed.) We also include the raw I/O performance of dumping the data uncompressed using a single `fwrite` call. Timings correspond to the median of five runs. Whereas our compressor is slightly slower than the less effective compressors [7, 22], it is nearly twice as fast as [16] while producing similar compression rates. However, in more I/O-intensive scenarios, such as in massively parallel simulations dumping data to the same file system (as is common), the improved compression of our method over [7,22] results in a net gain in effective throughput. We integrated our compression code with Miranda's dump routines and ran performance tests on 256 nodes of LLNL's MCR supercomputer. Achieving on average a lossless reduction of 3.7 on 75 GB of data dumped, the overall dump time was reduced by a factor of 2.7 over writing the data uncompressed.

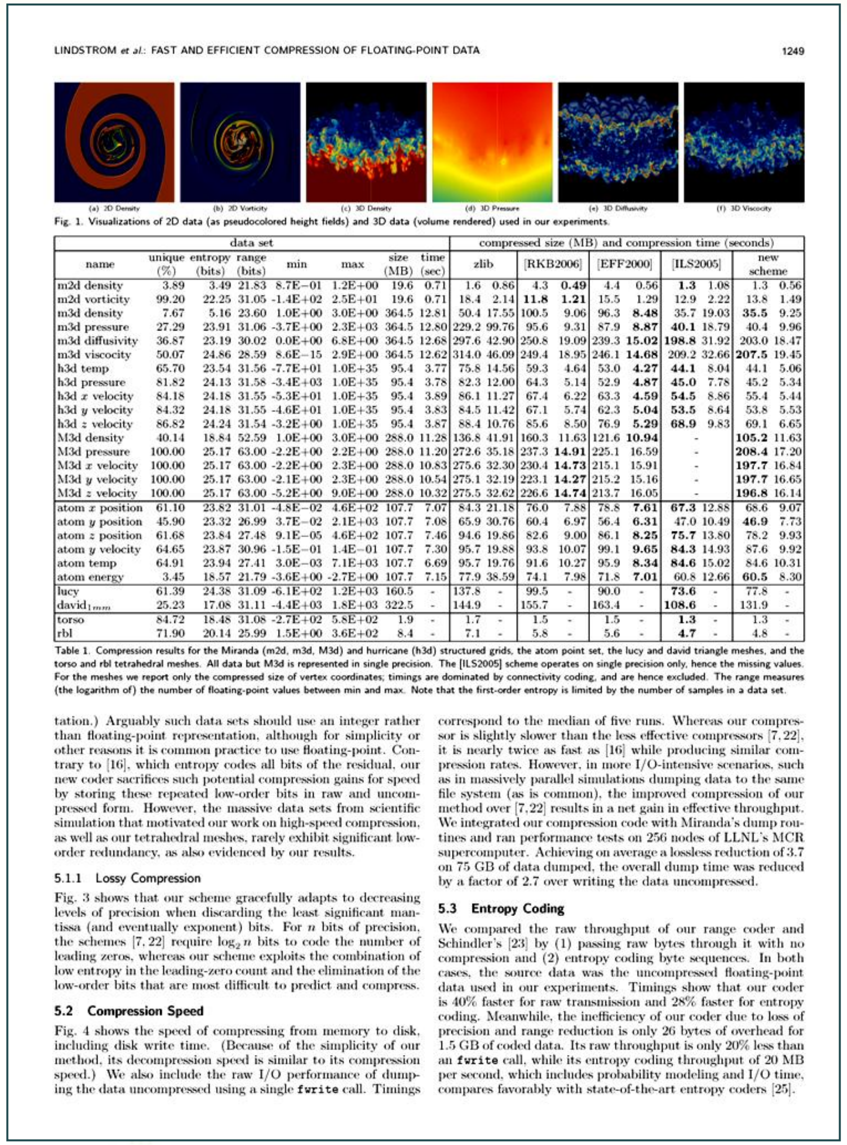
5.3 Entropy Coding

We compared the raw throughput of our range coder and Schindler's [23] by (1) passing raw bytes through it with no compression and (2) entropy coding byte sequences. In both cases, the source data was the uncompressed floating-point data used in our experiments. Timings show that our coder is 40% faster for raw transmission and 28% faster for entropy coding. Meanwhile, the inefficiency of our coder due to loss of precision and range reduction is only 26 bytes of overhead for 1.5 GB of coded data. Its raw throughput is only 20% less than an `fwrite` call, while its entropy coding throughput of 20 MB per second, which includes probability modeling and I/O time, compares favorably with state-of-the-art entropy coders [25].

Linstrom et al. [2006]

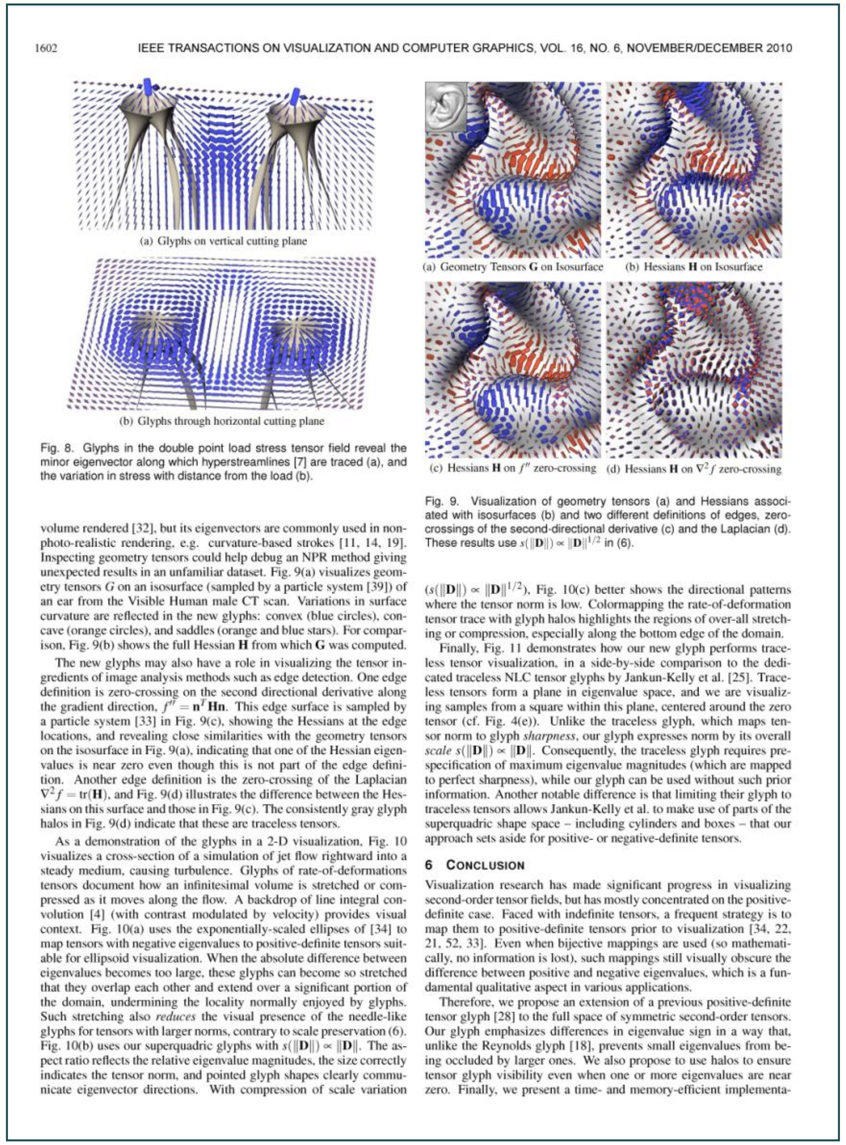
Results: evaluation scenarios (selection)

algorithmic performance
35% of scenarios



Linstrom et al. [2006]

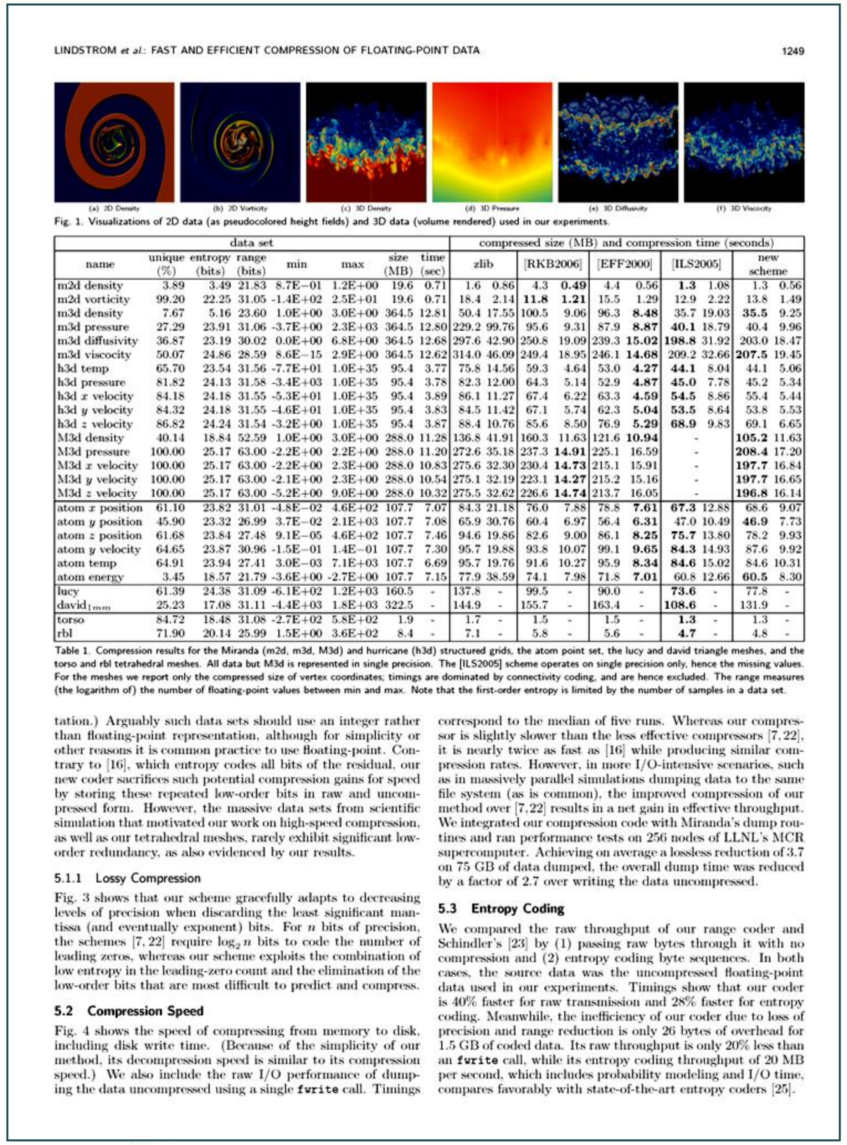
qualitative result inspection
46% of scenarios



Schultz & Kindlmann [2010]

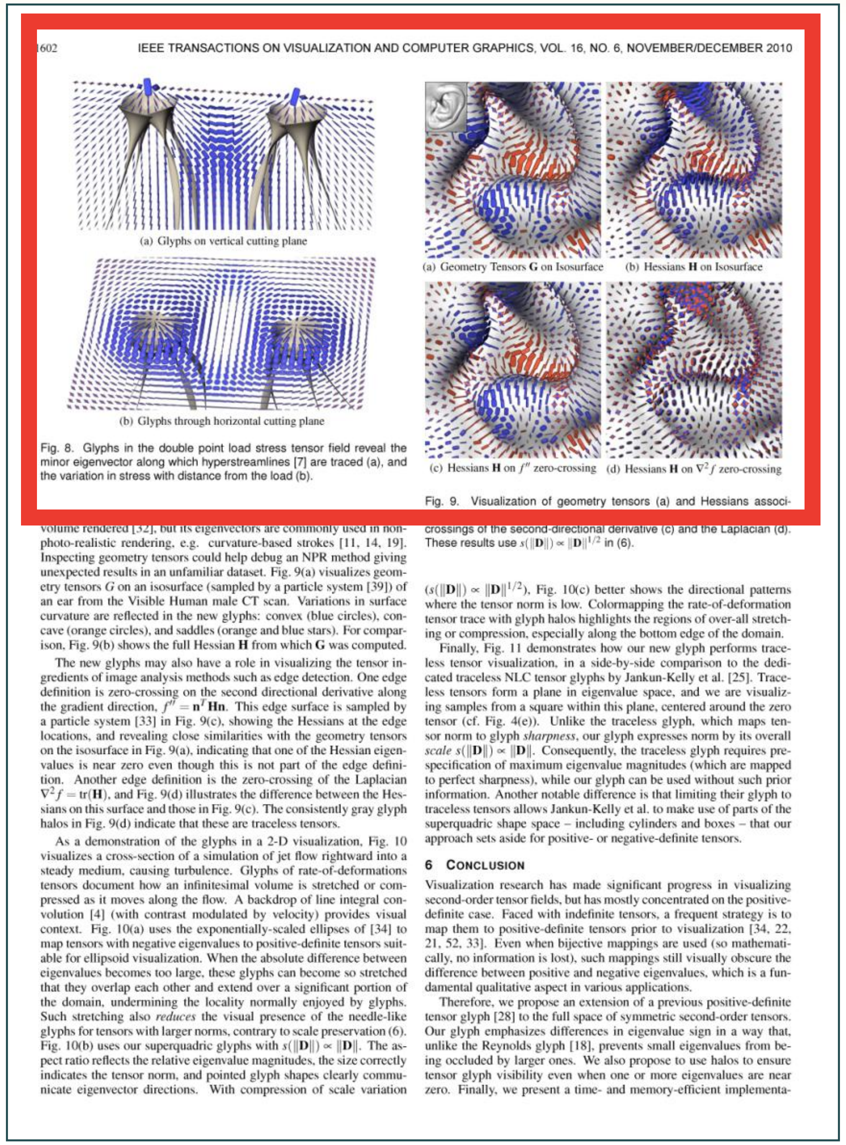
Results: evaluation scenarios (selection)

algorithmic performance
35% of scenarios



Linstrom et al. [2006]

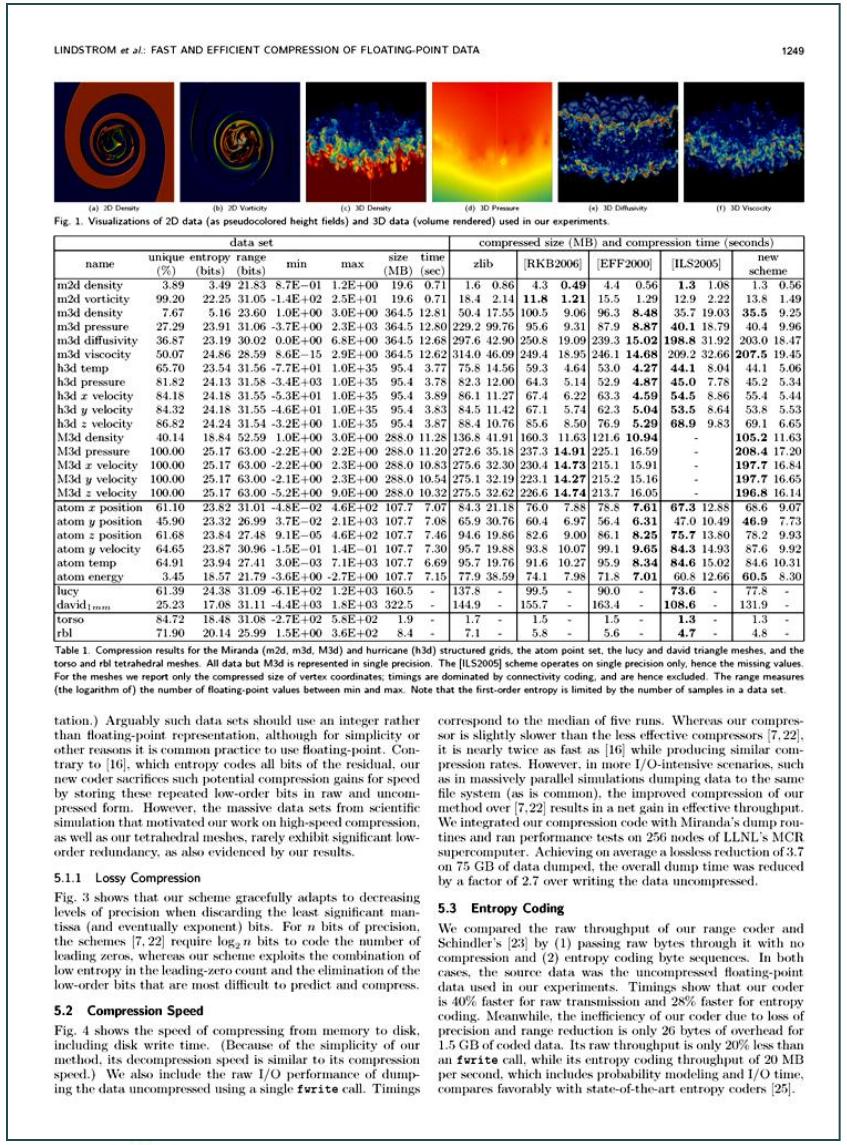
qualitative result inspection
46% of scenarios



Schultz & Kindlmann [2010]

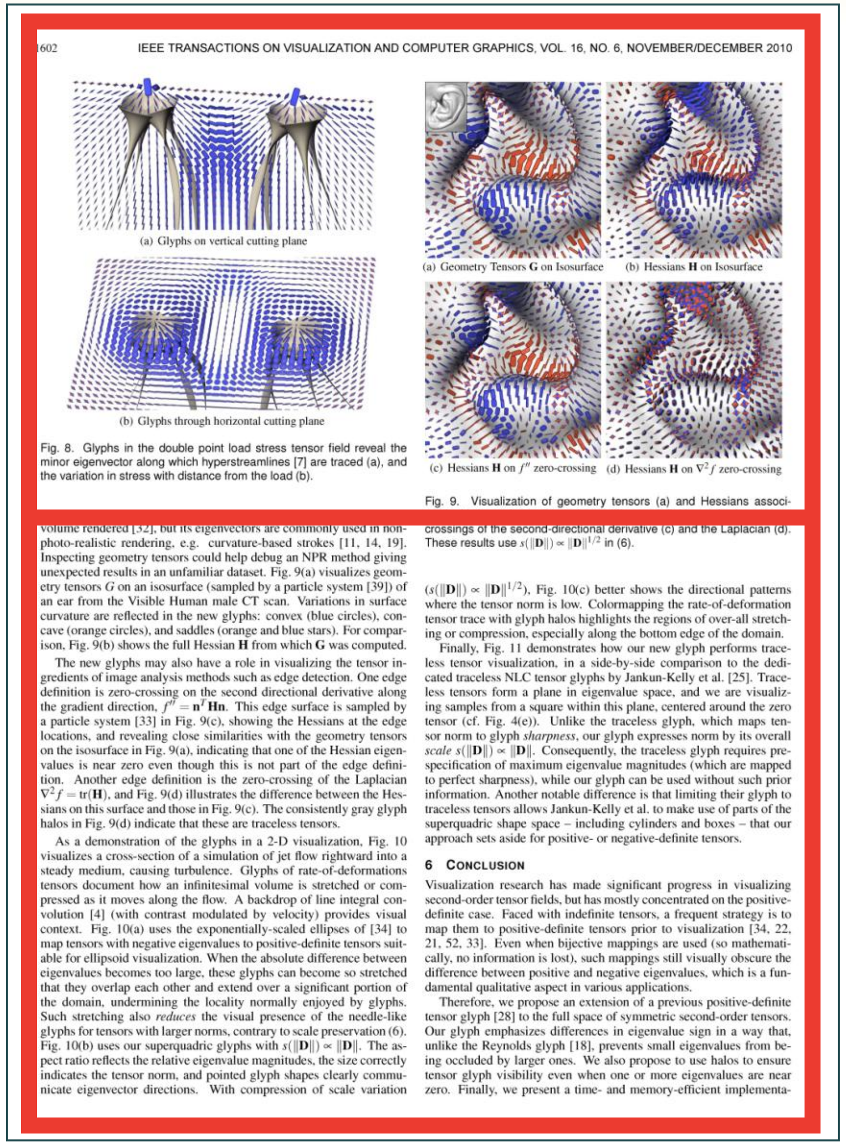
Results: evaluation scenarios (selection)

algorithmic performance
35% of scenarios



Linstrom et al. [2006]

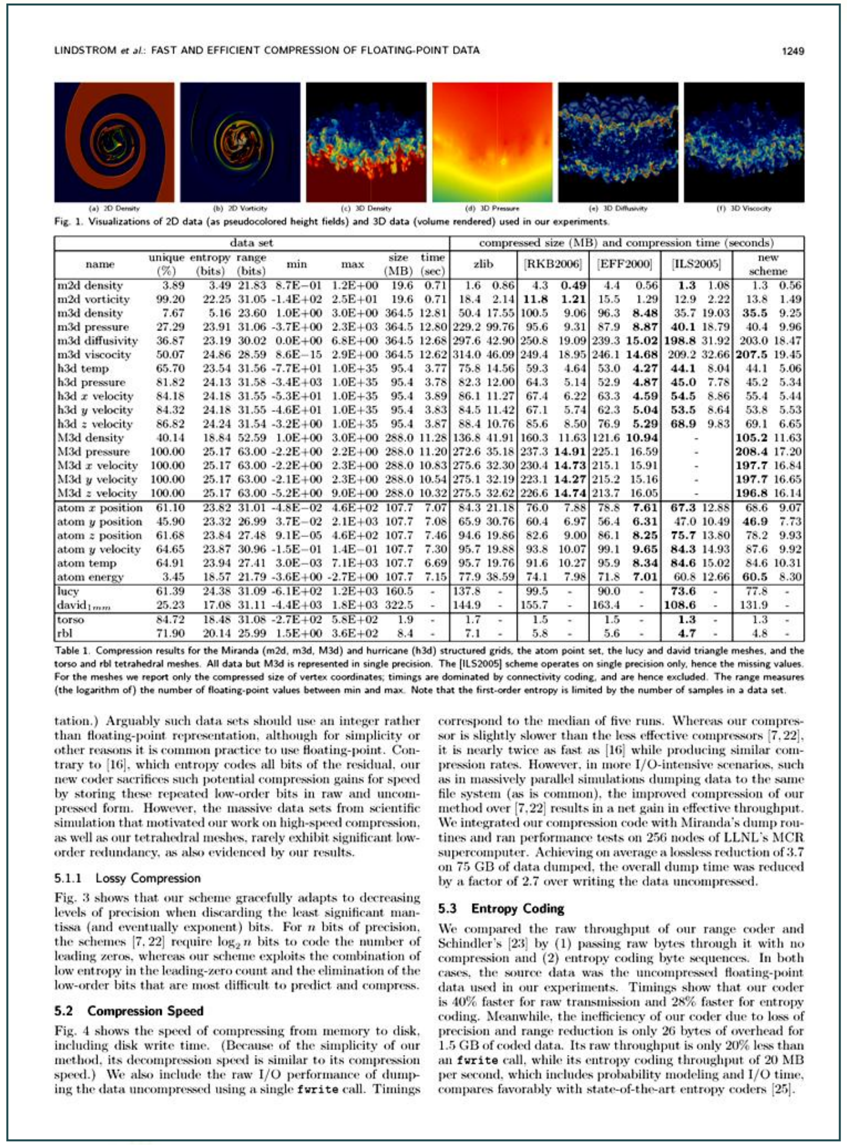
qualitative result inspection
46% of scenarios



Schultz & Kindlmann [2010]

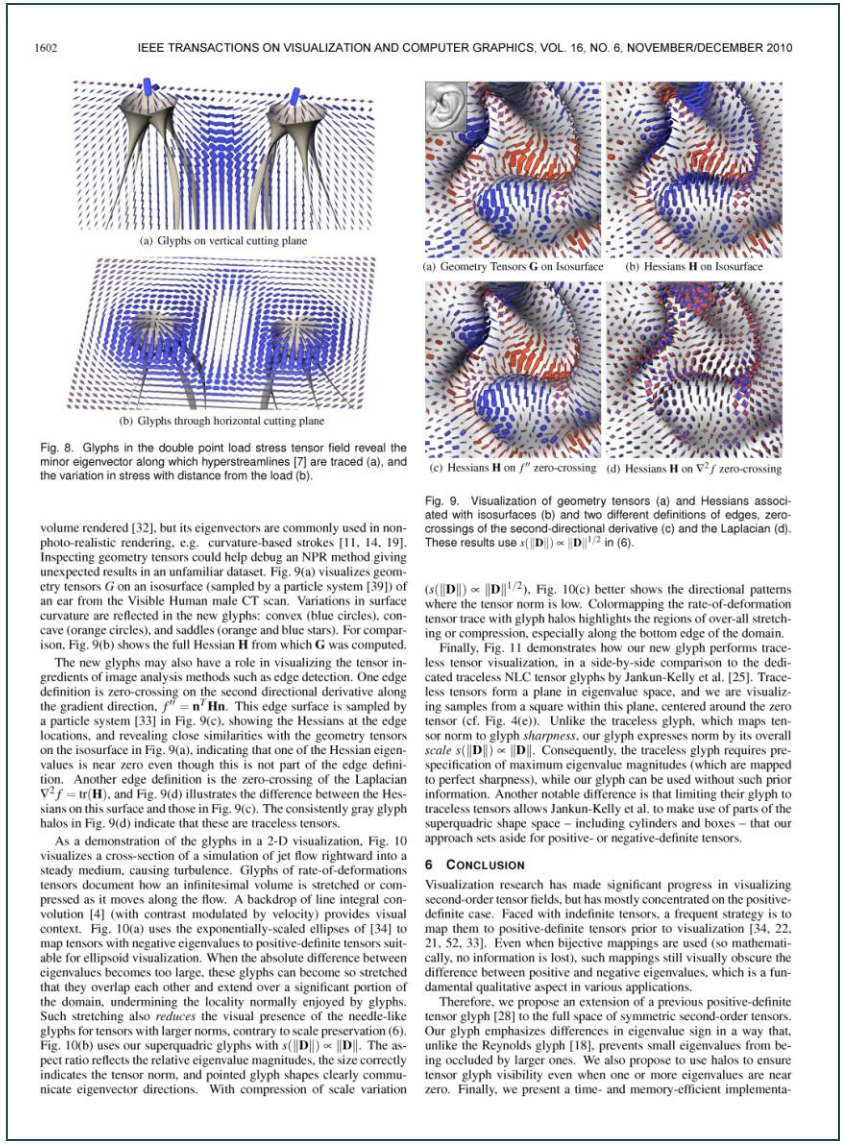
Results: evaluation scenarios (selection)

algorithmic performance
35% of scenarios



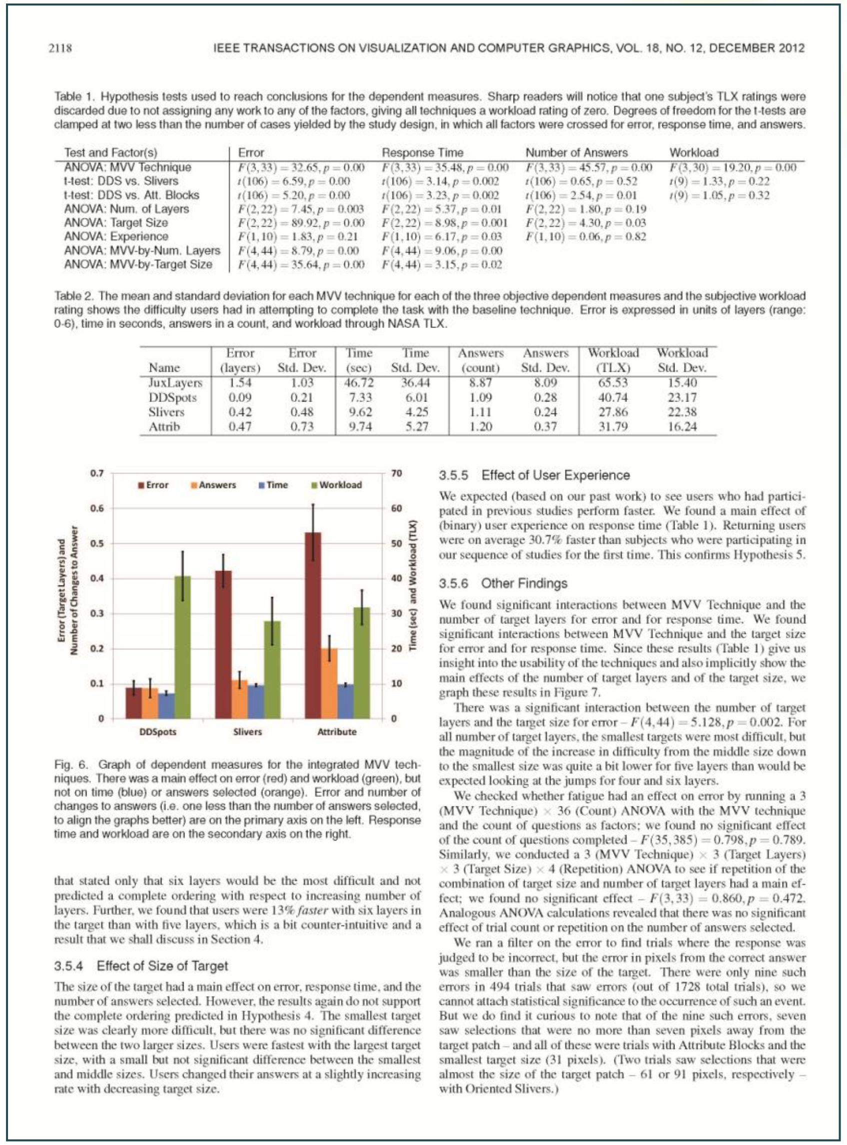
Linstrom et al. [2006]

qualitative result inspection
46% of scenarios



Schultz & Kindlmann [2010]

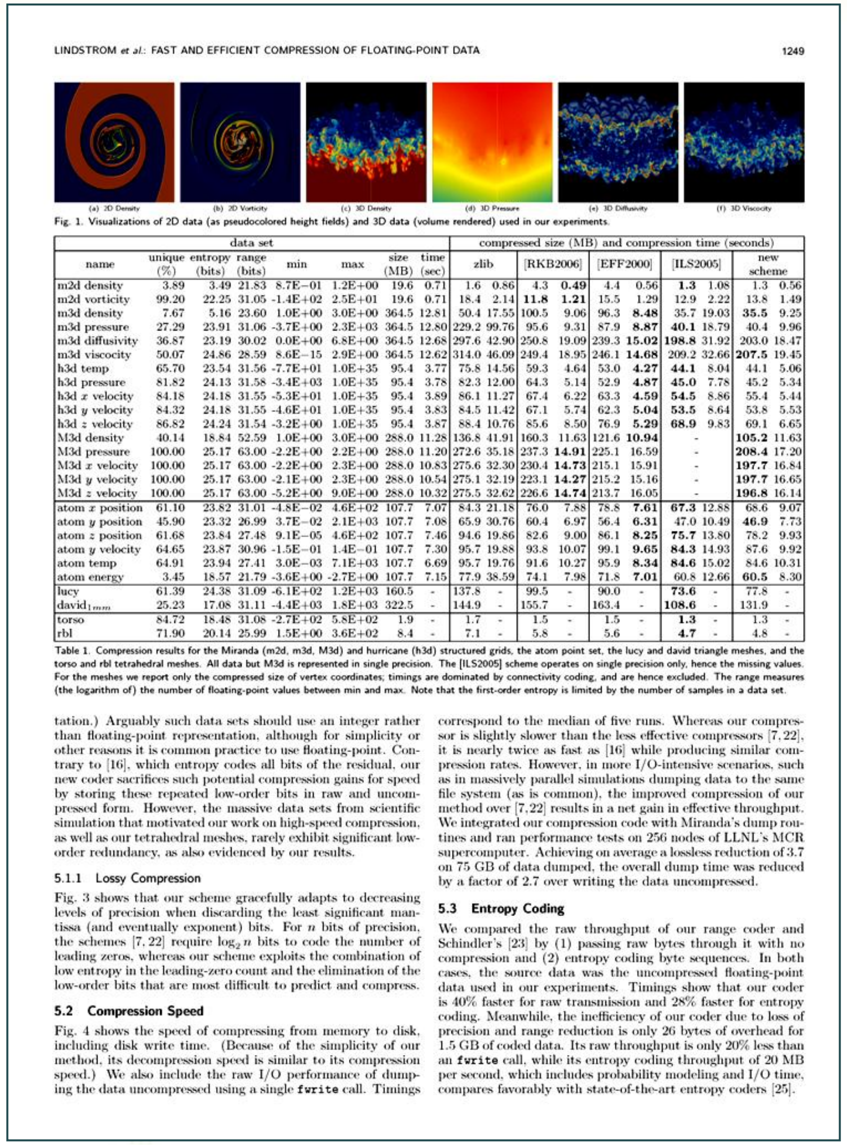
user performance/experience
14% of scenarios



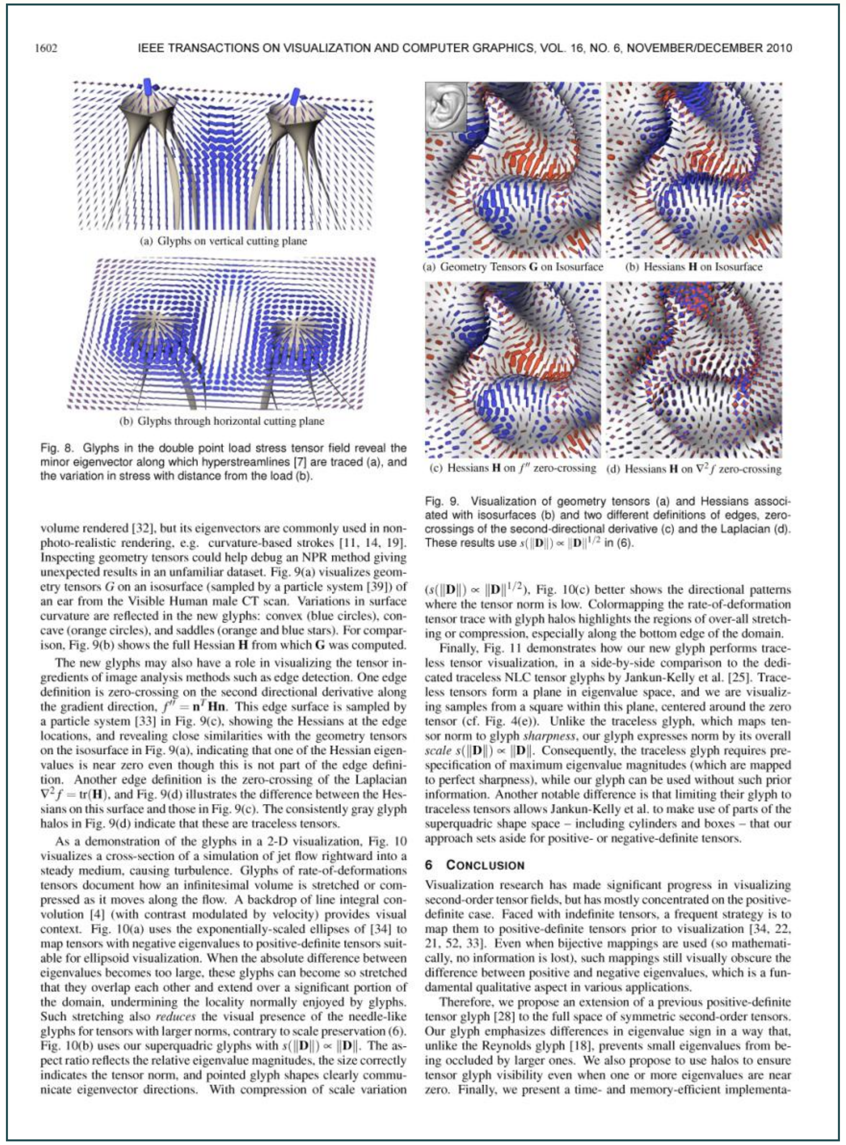
Livingston et al. [2012]

Results: evaluation scenarios (selection)

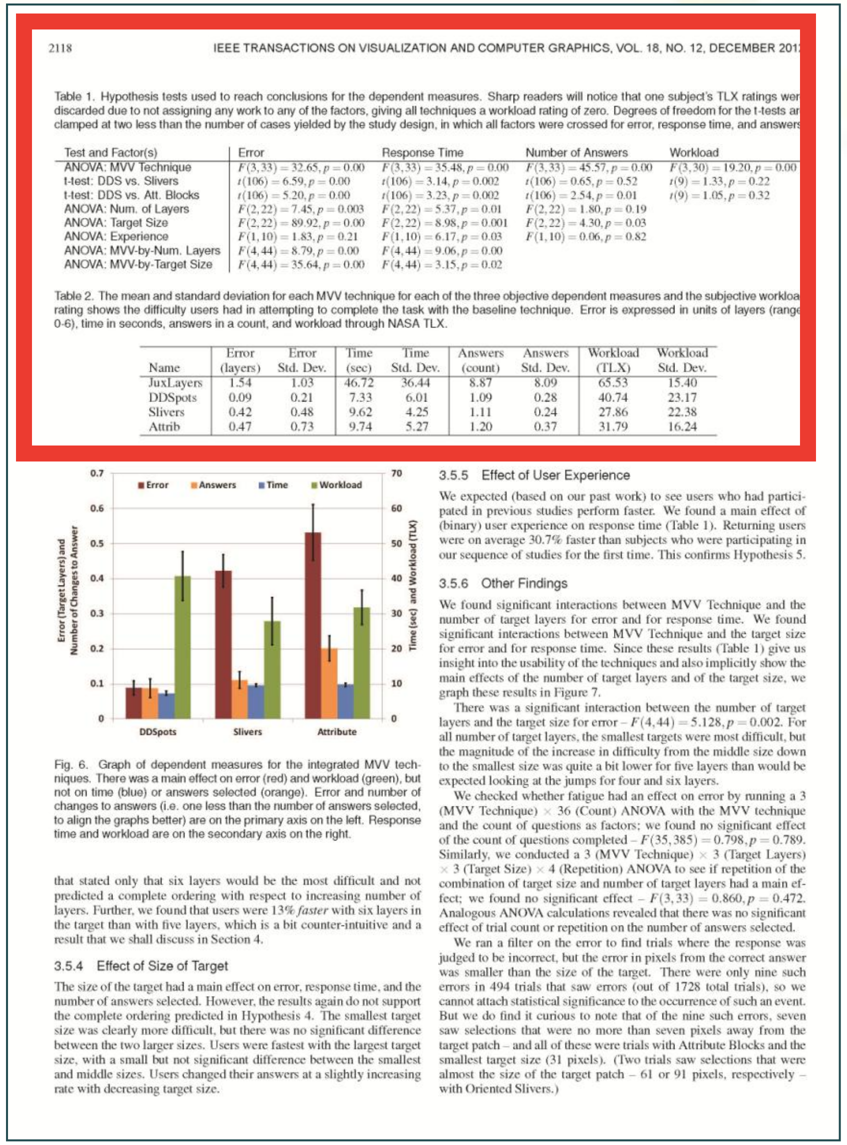
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios



Linstrom et al. [2006]



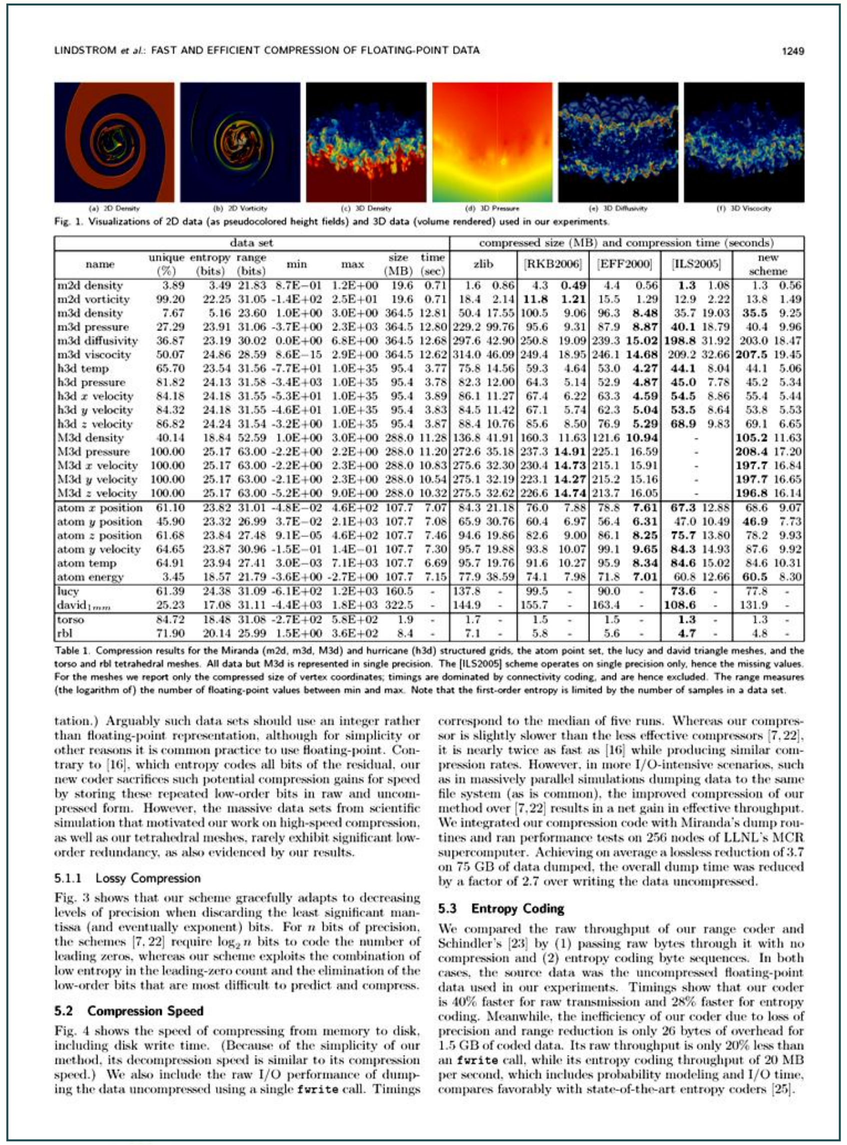
Schultz & Kindlmann [2010]



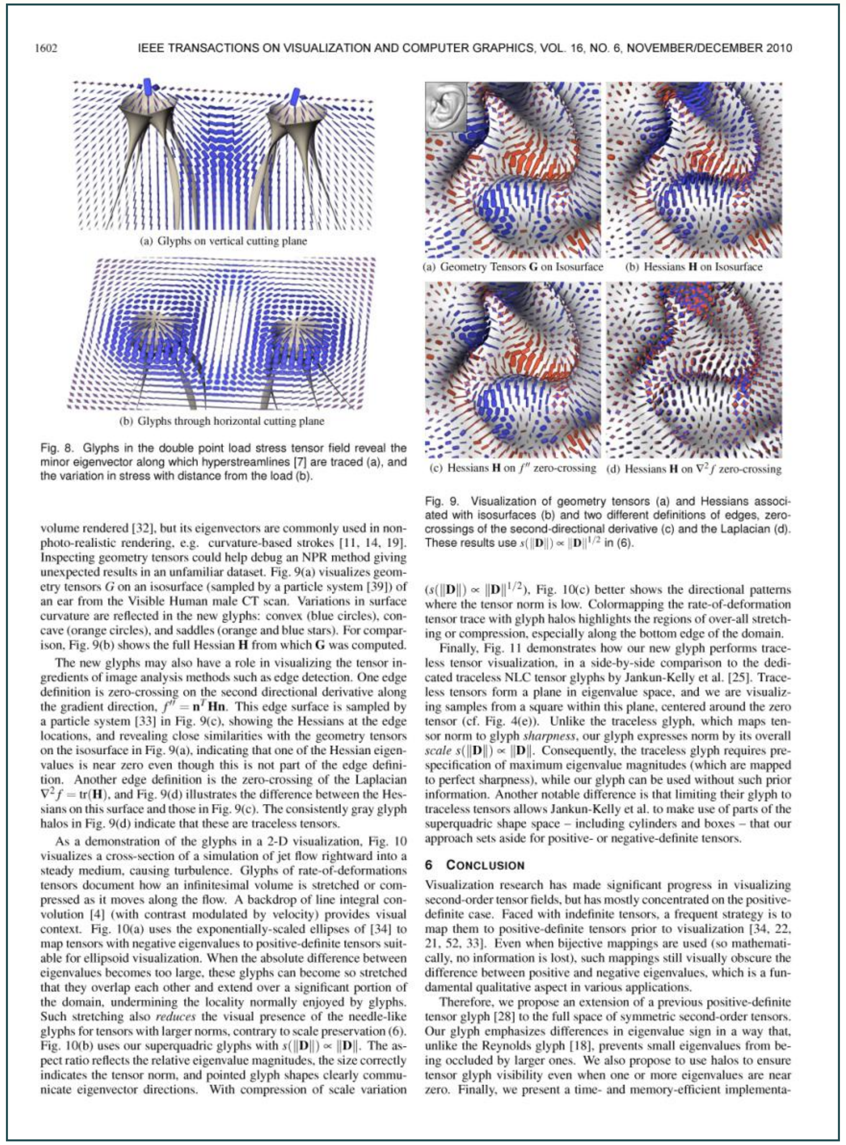
Livingston et al. [2012]

Results: evaluation scenarios (selection)

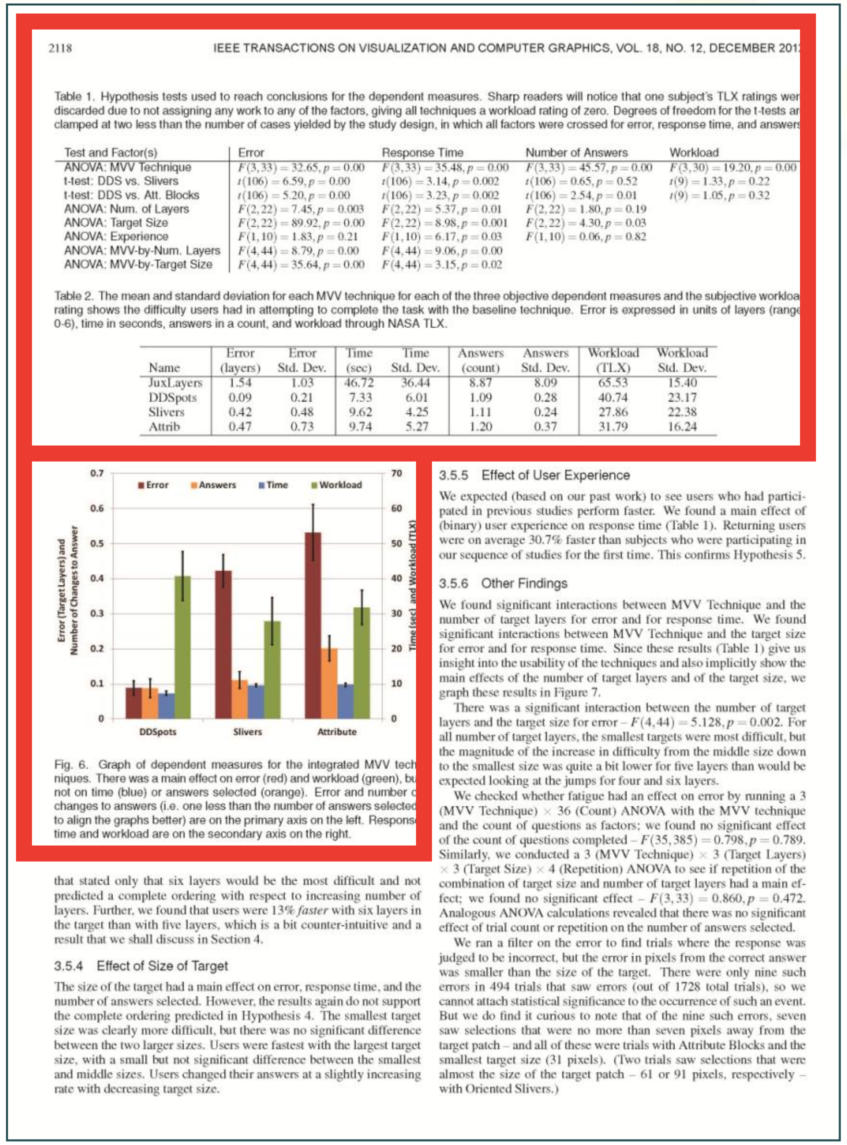
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios



Linstrom et al. [2006]



Schultz & Kindlmann [2010]



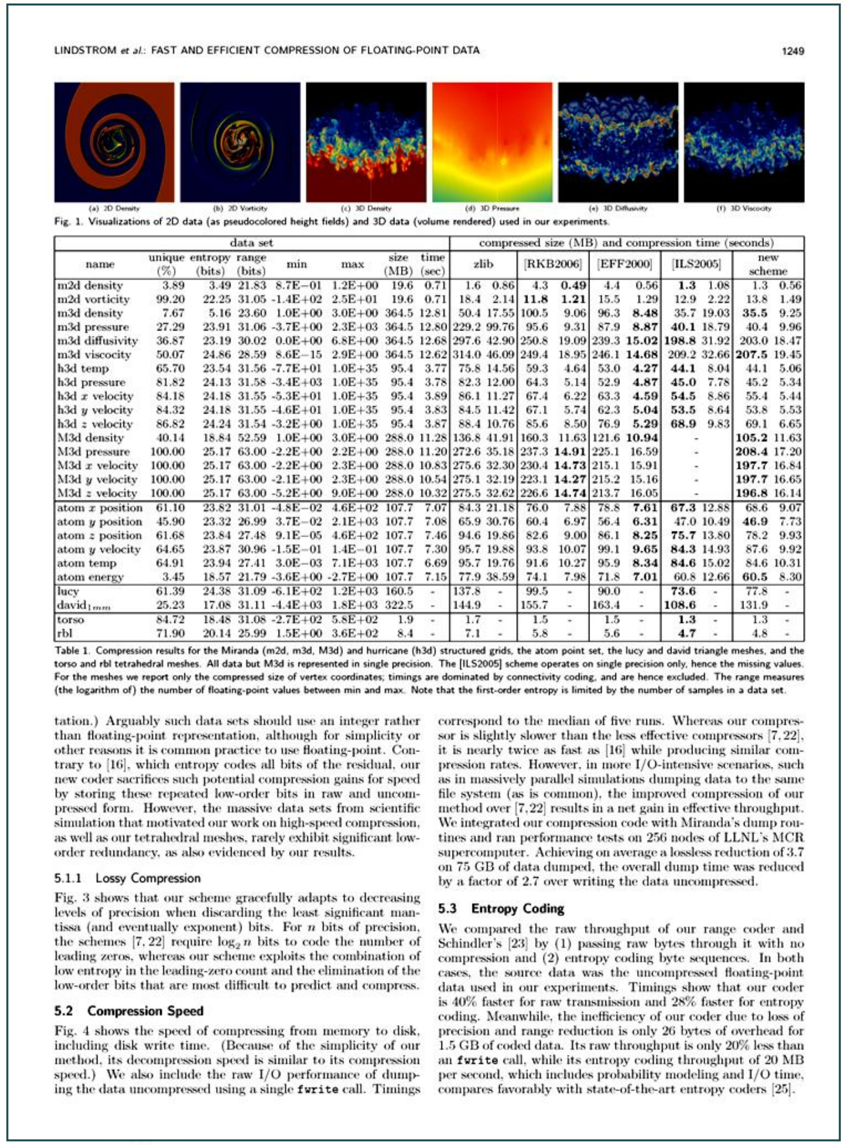
Livingston et al. [2012]

Results: evaluation scenarios (selection)

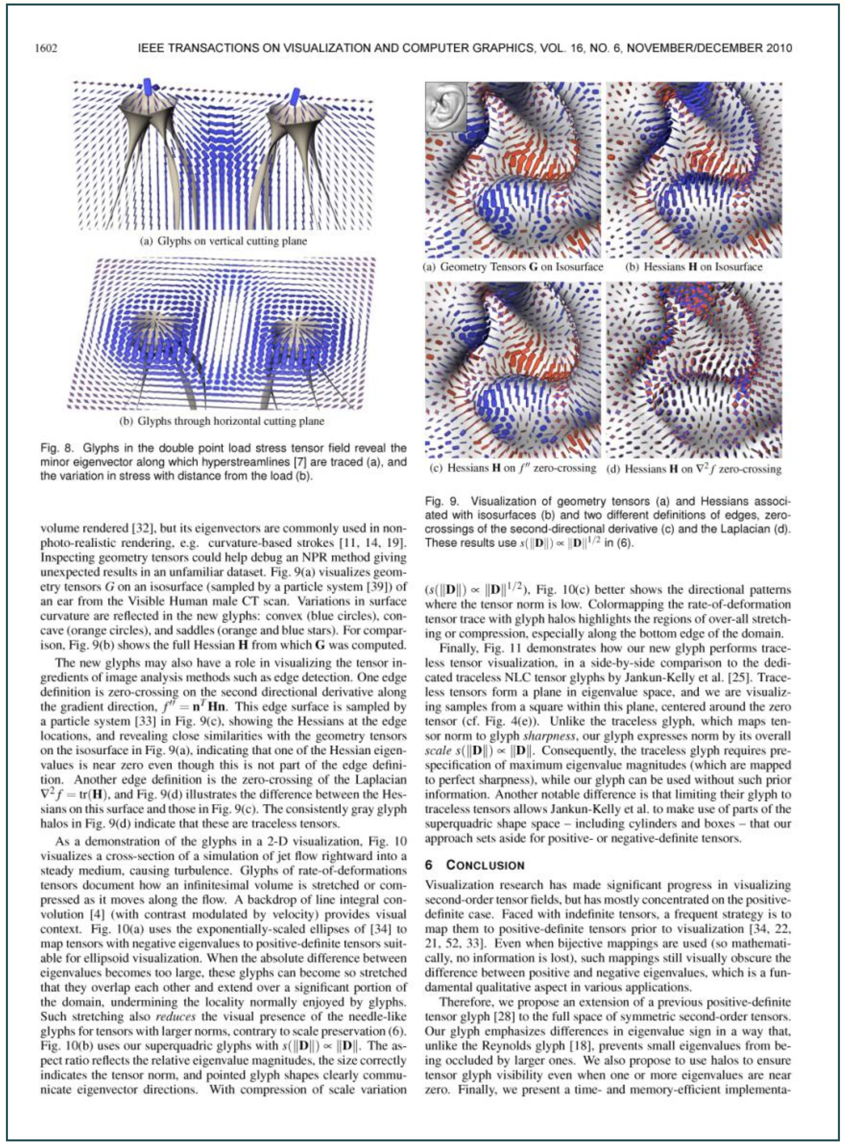
algorithmic performance
35% of scenarios

qualitative result inspection
46% of scenarios

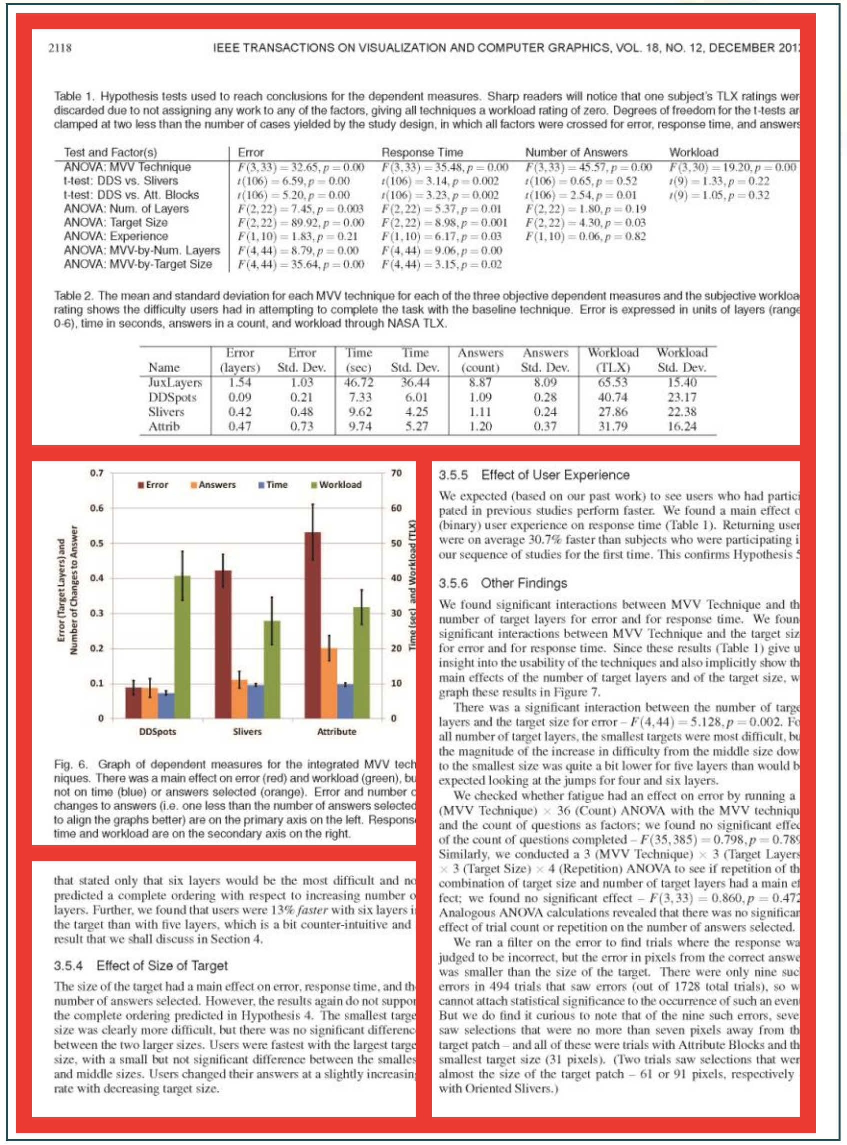
user performance/experience
14% of scenarios



Linstrom et al. [2006]



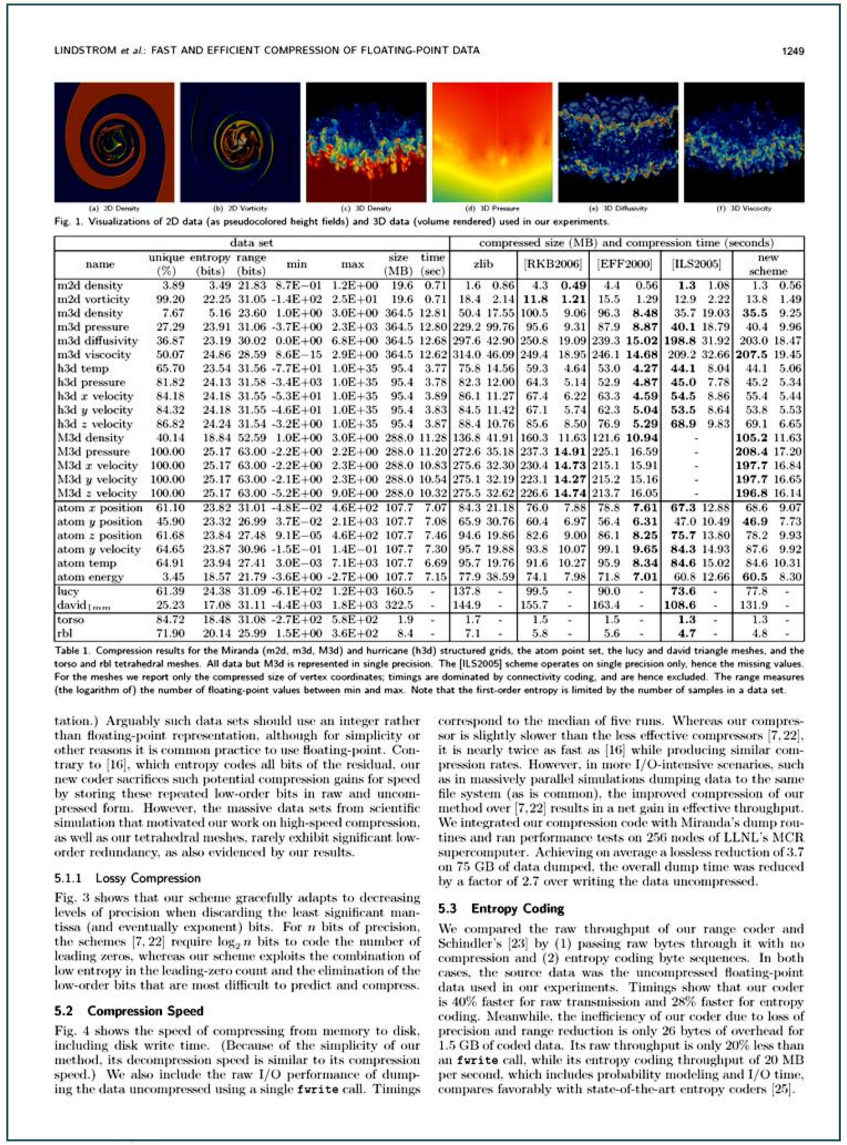
Schultz & Kindlmann [2010]



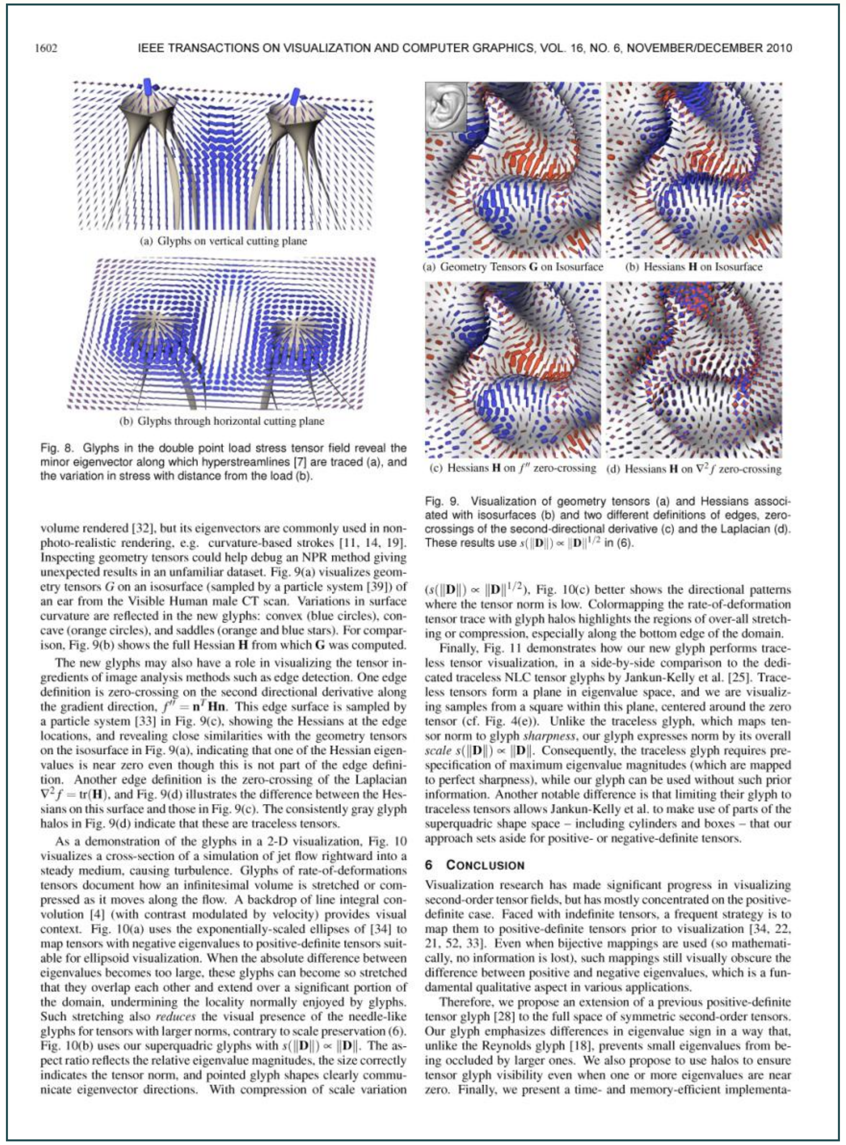
Livingston et al. [2012]

Results: evaluation scenarios (selection)

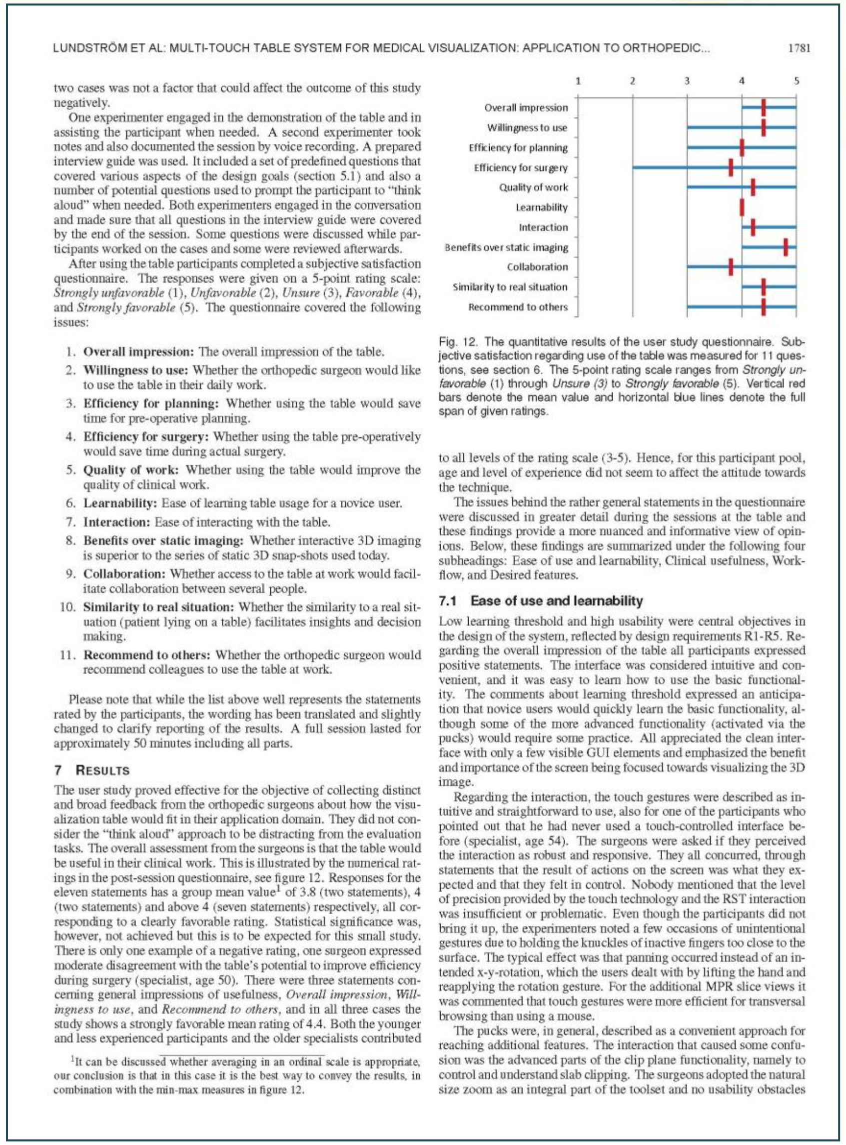
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios



Linstrom et al. [2006]



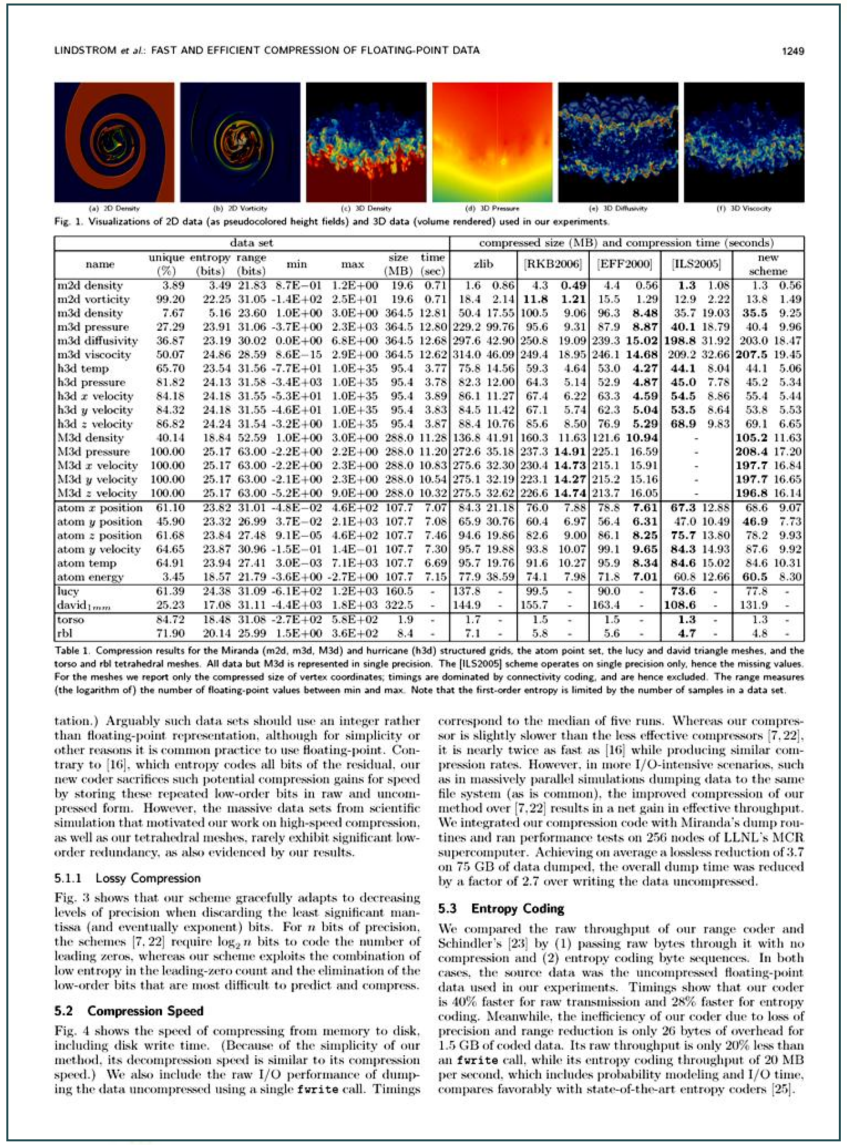
Schultz & Kindlmann [2010]



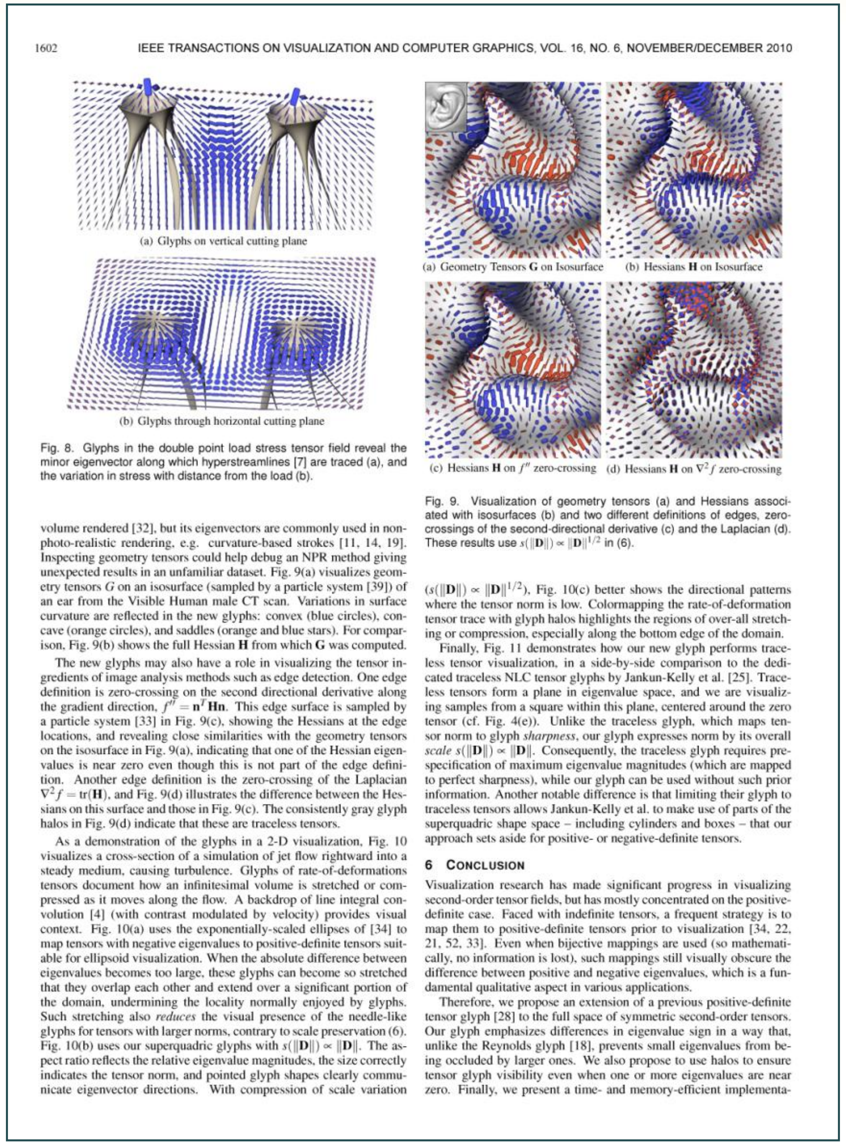
Lundström et al. [2011]

Results: evaluation scenarios (selection)

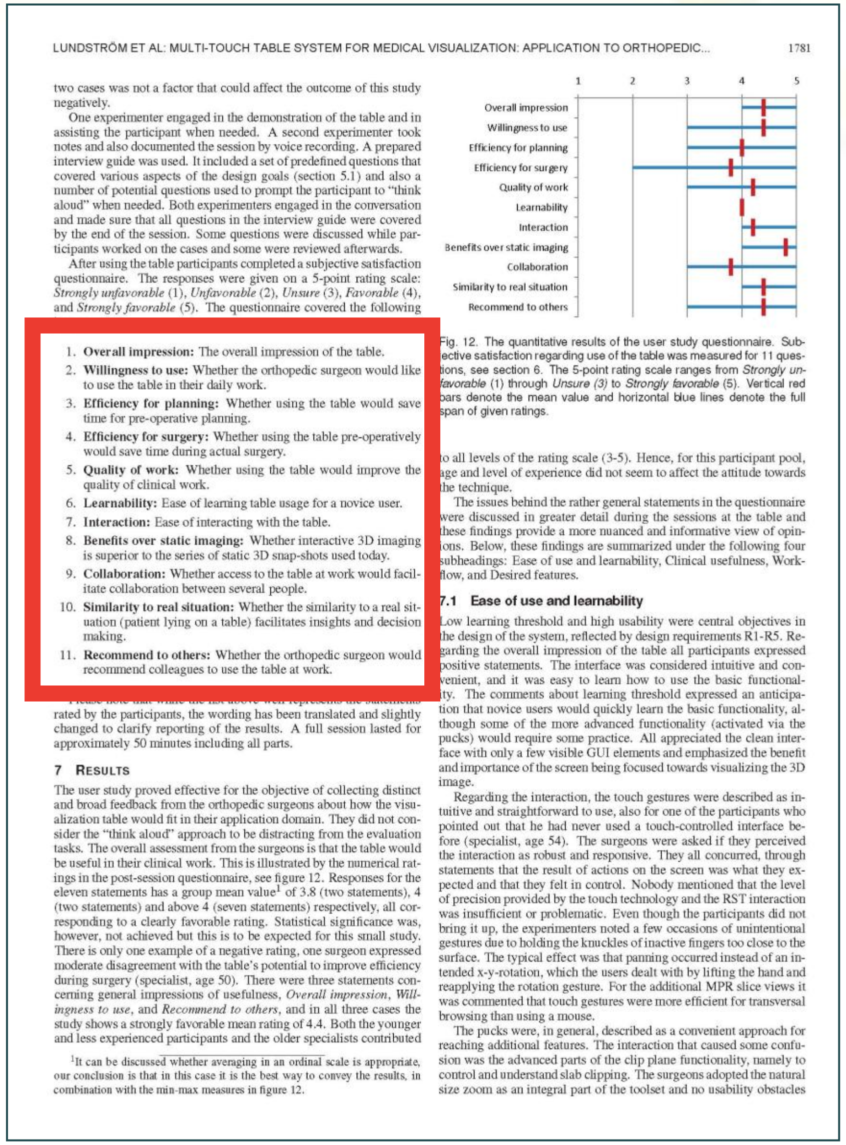
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios


LINDSTROM et al.: FAST AND EFFICIENT COMPRESSION OF FLOATING-POINT DATA 1249

Linstrom et al. [2006]


1602 IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 16, NO. 6, NOVEMBER/DECEMBER 2010

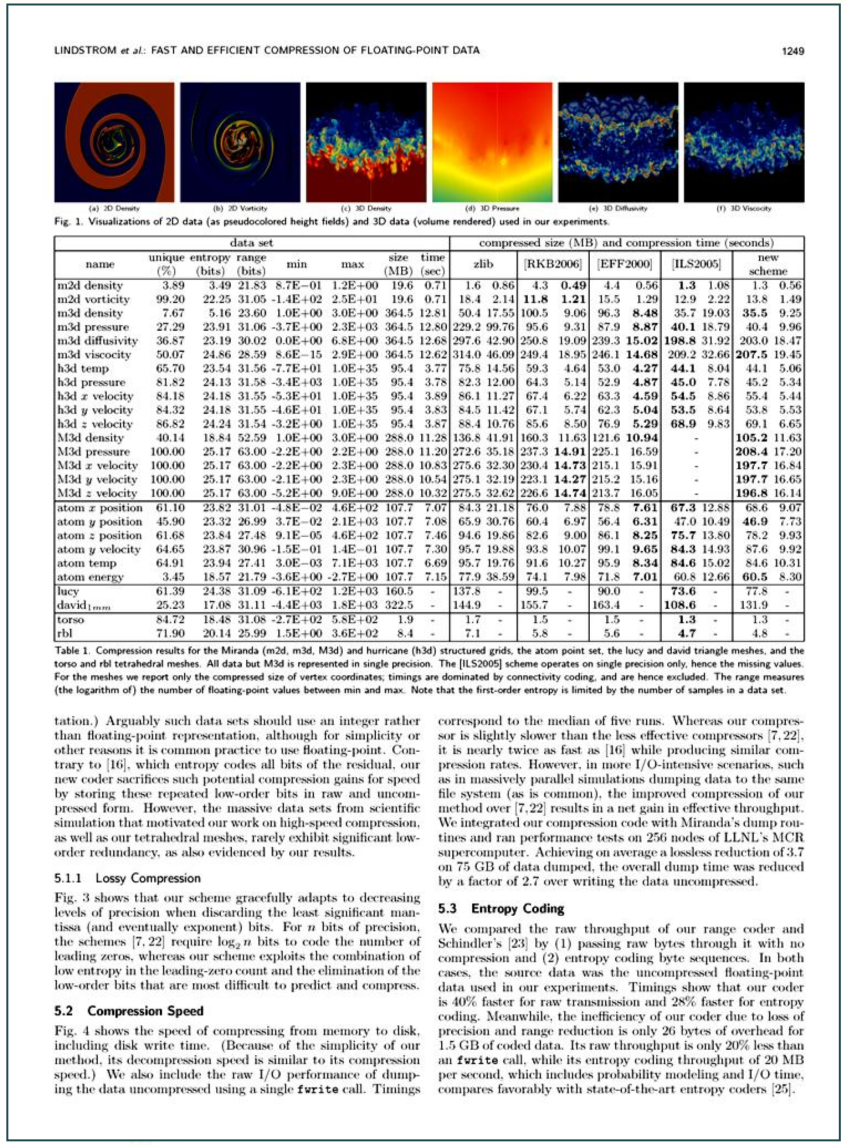
Schultz & Kindlmann [2010]


LUNDSTRÖM ET AL: MULTI-TOUCH TABLE SYSTEM FOR MEDICAL VISUALIZATION: APPLICATION TO ORTHOPEDIC... 1781

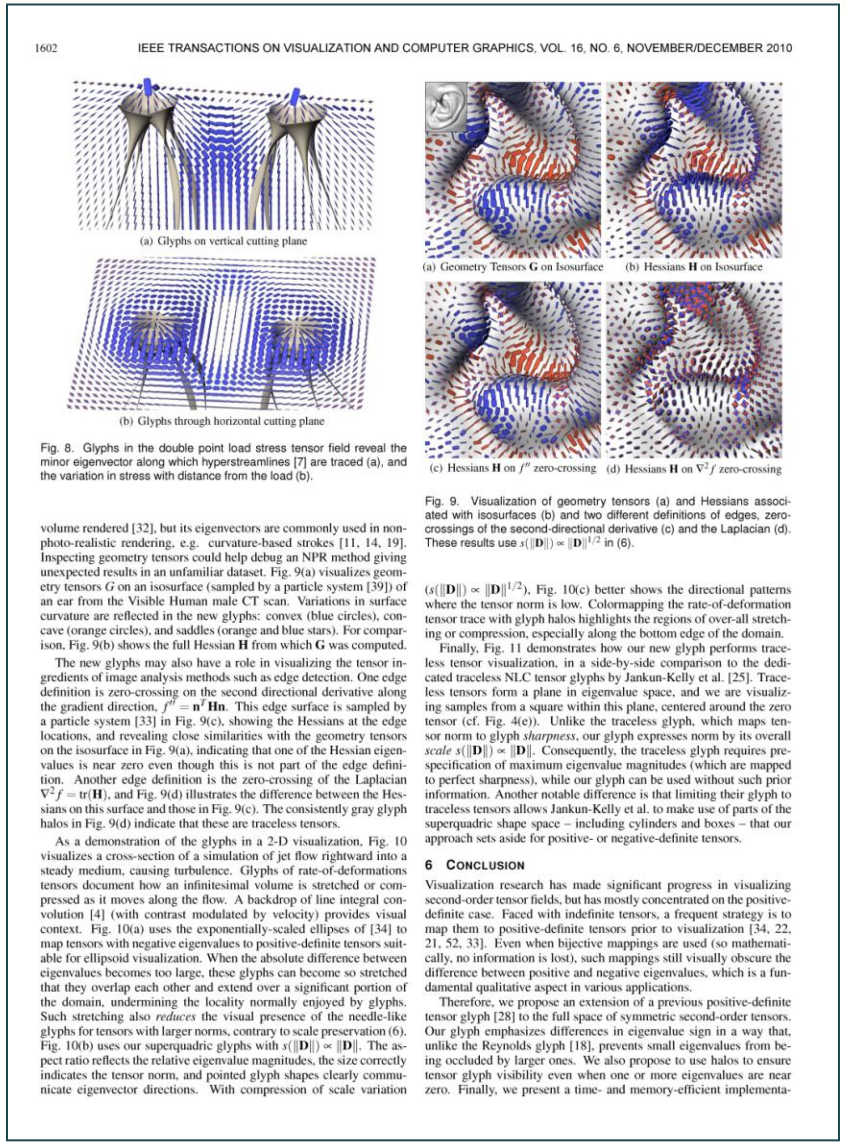
Lundström et al. [2011]

Results: evaluation scenarios (selection)

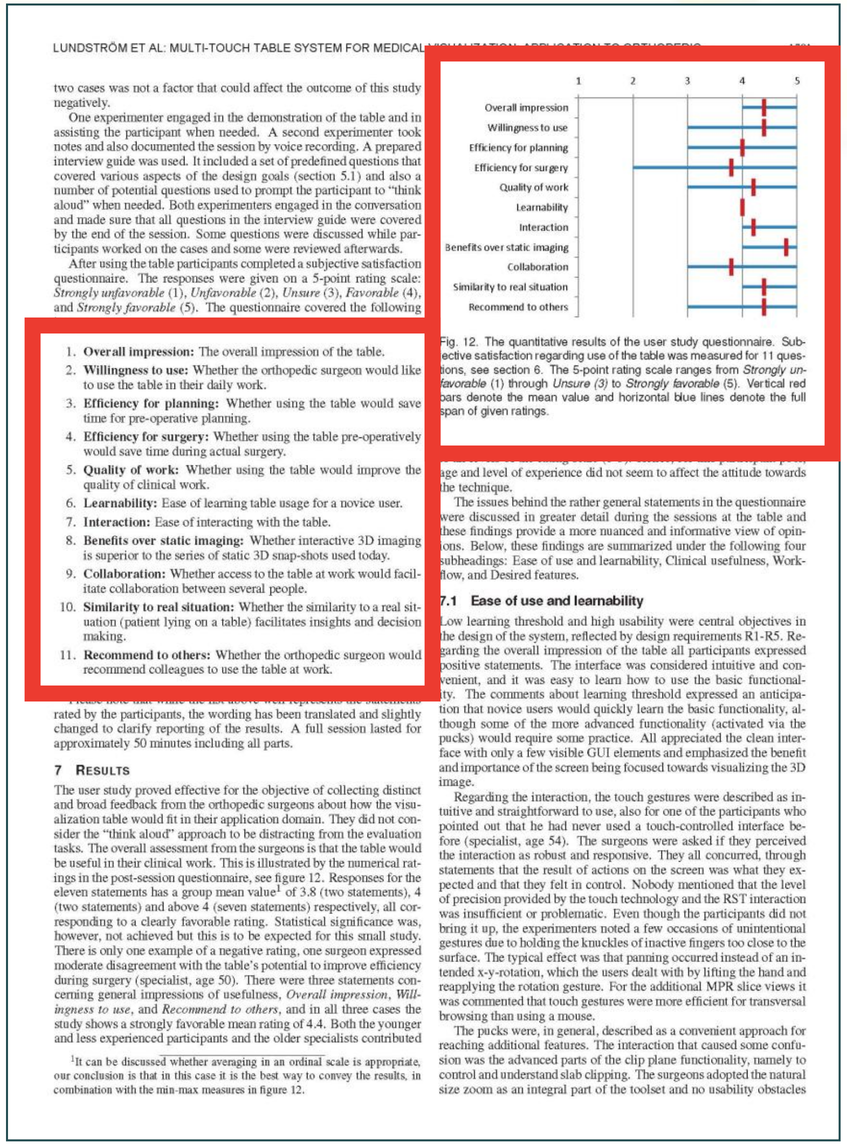
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios



Linstrom et al. [2006]



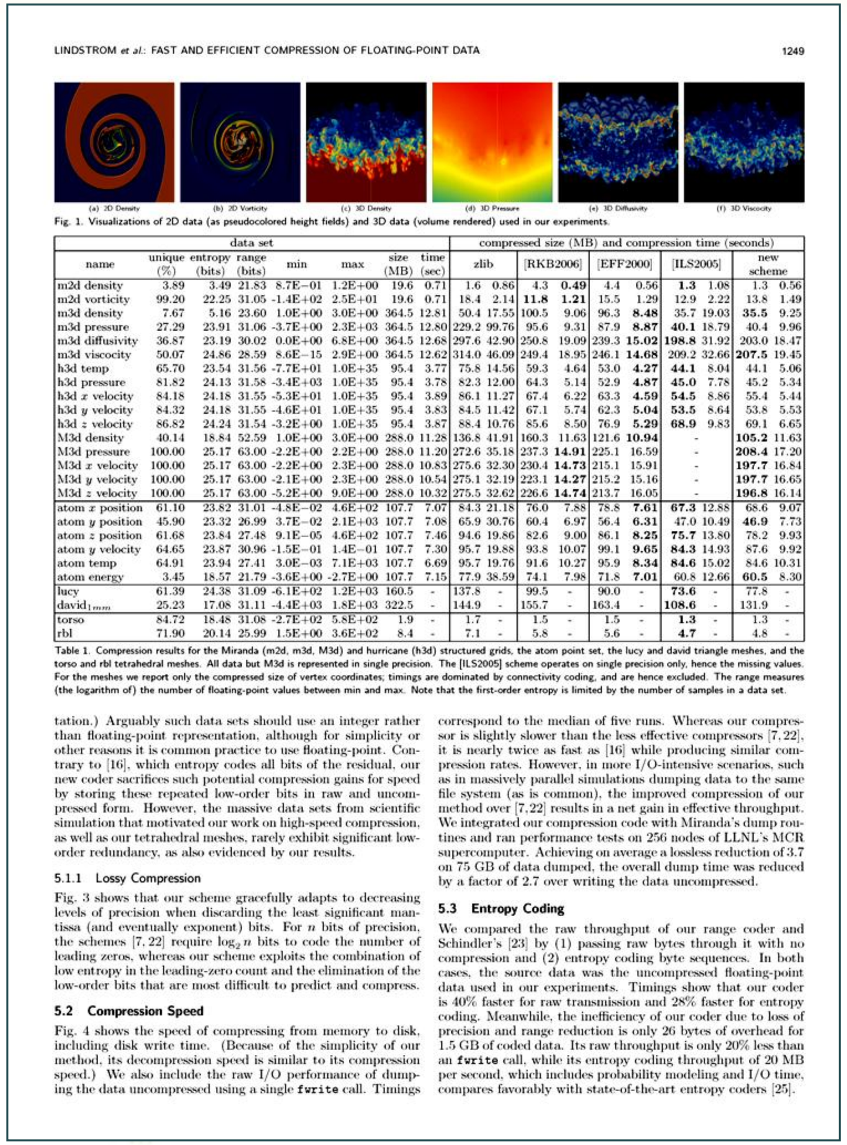
Schultz & Kindlmann [2010]



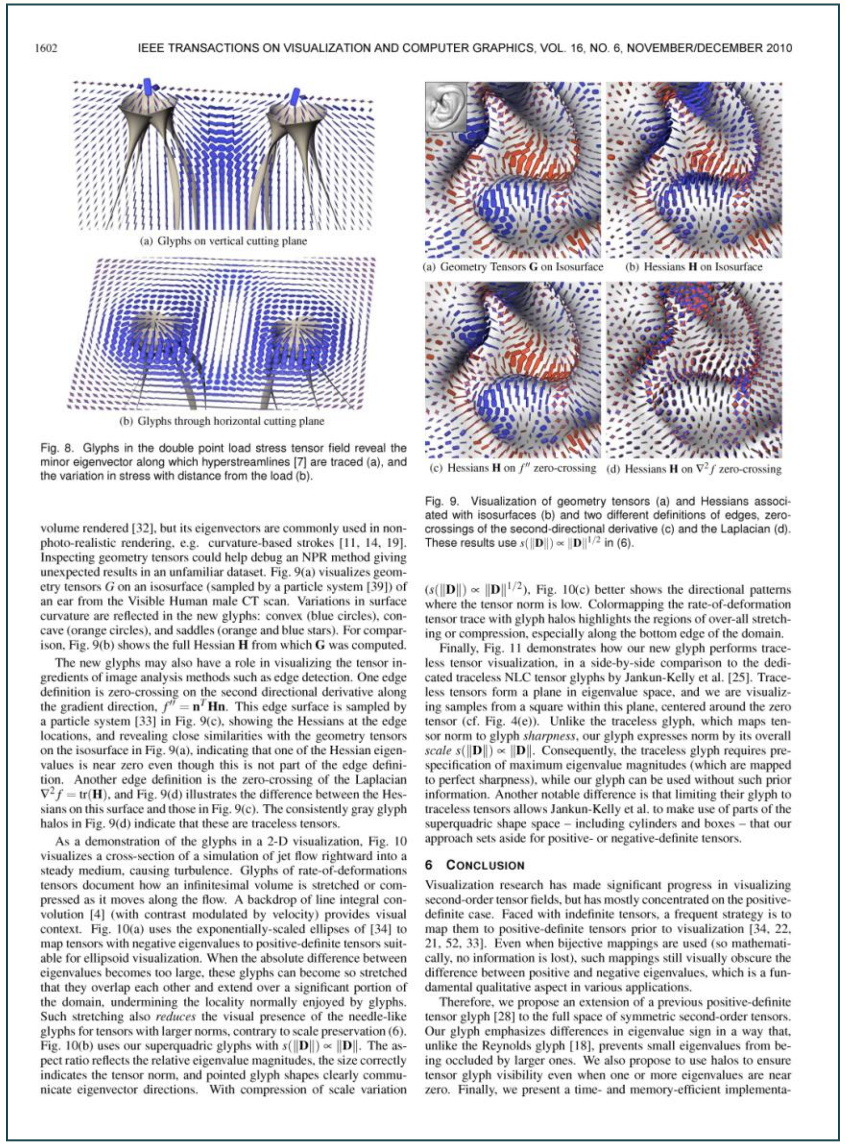
Lundström et al. [2011]

Results: evaluation scenarios (selection)

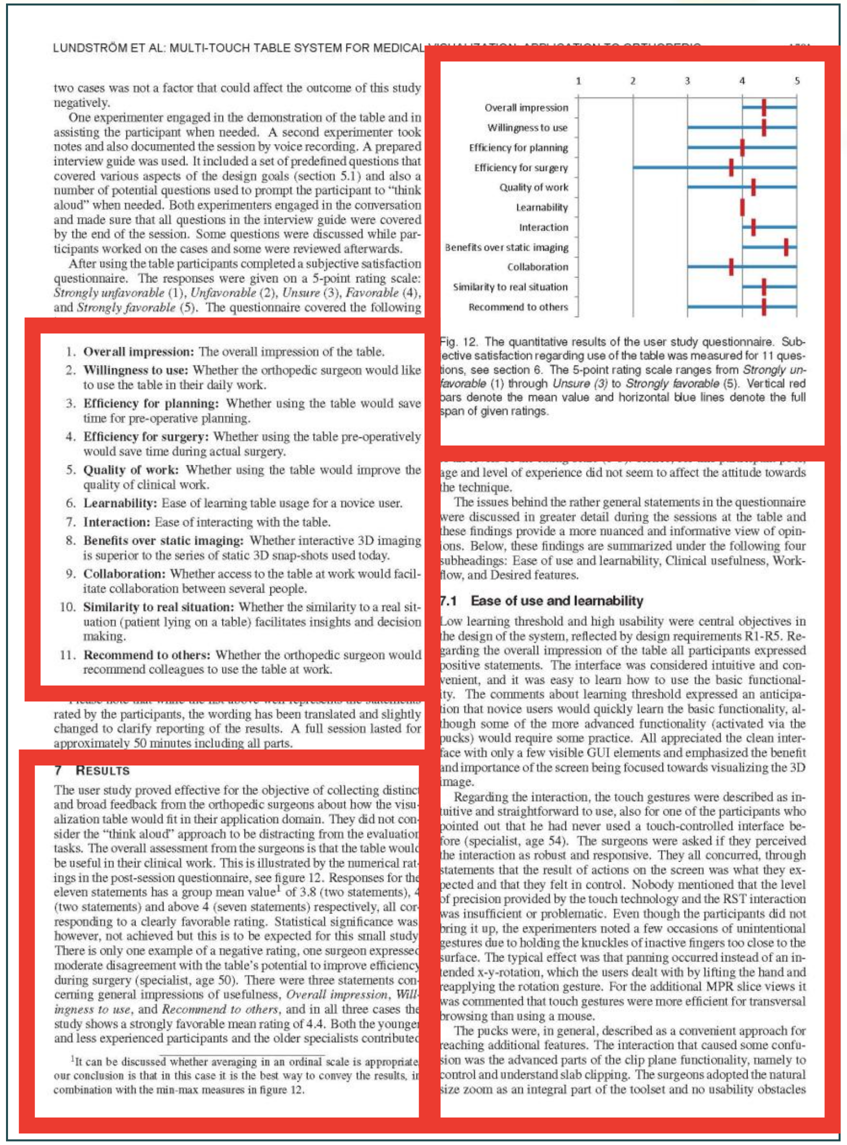
algorithmic performance	qualitative result inspection	user performance/experience
35% of scenarios	46% of scenarios	14% of scenarios



Linstrom et al. [2006]

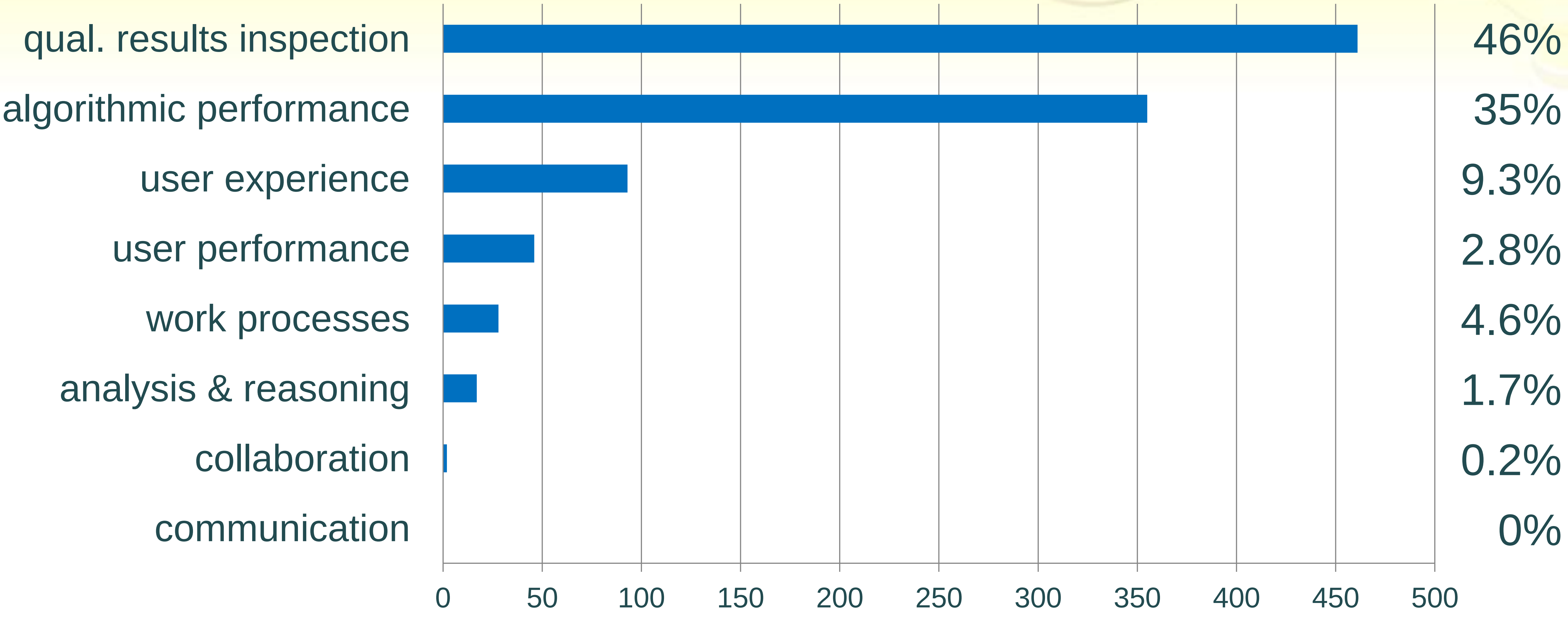


Schultz & Kindlmann [2010]



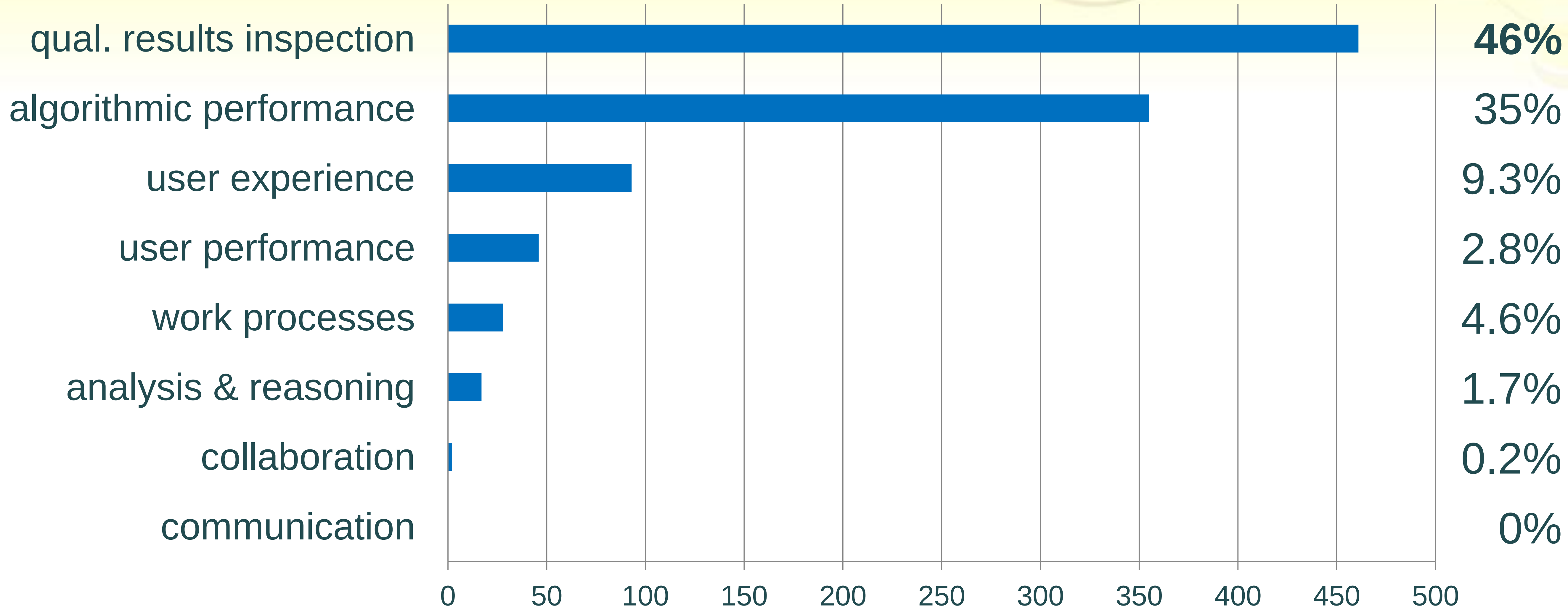
Lundström et al. [2011]

Results: evaluation scenarios



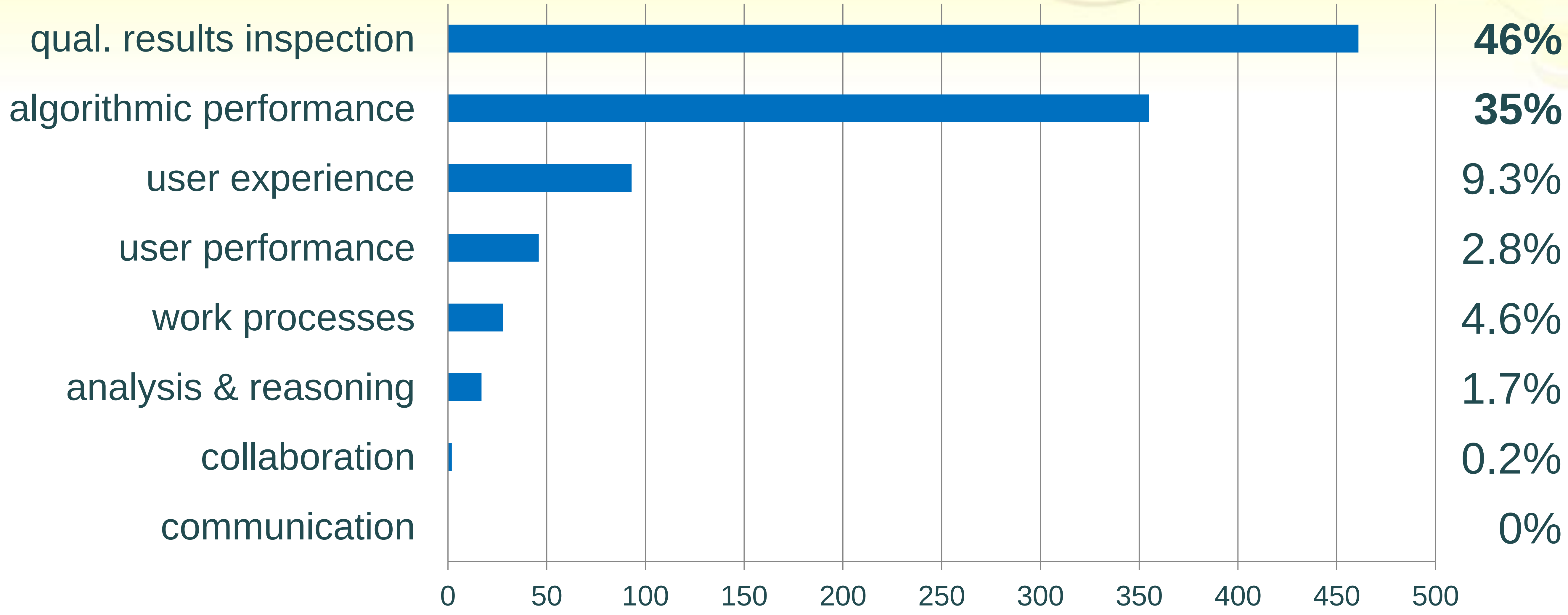
IEEE Vis/SciVis: total numbers & percent of evaluation scenarios

Results: evaluation scenarios



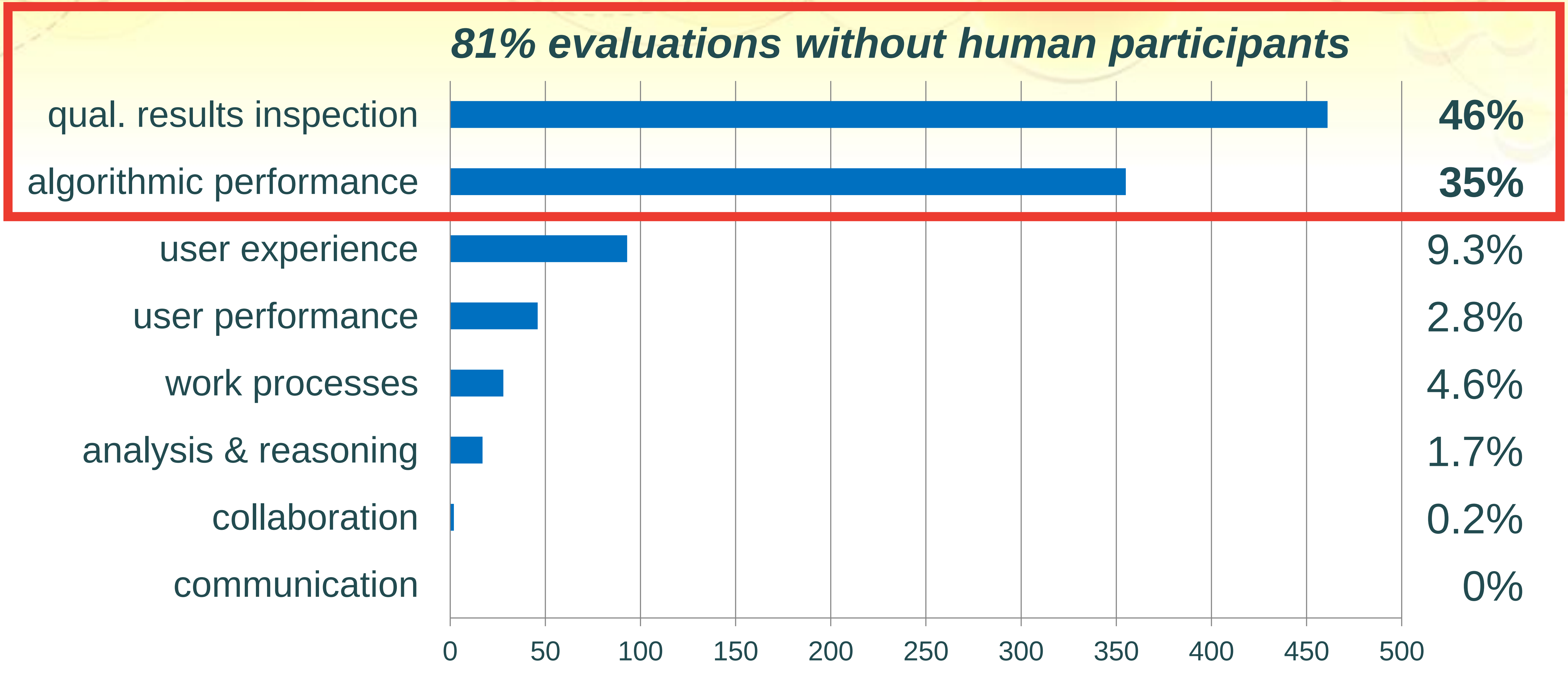
IEEE Vis/SciVis: total numbers & percent of evaluation scenarios

Results: evaluation scenarios



IEEE Vis/SciVis: total numbers & percent of evaluation scenarios

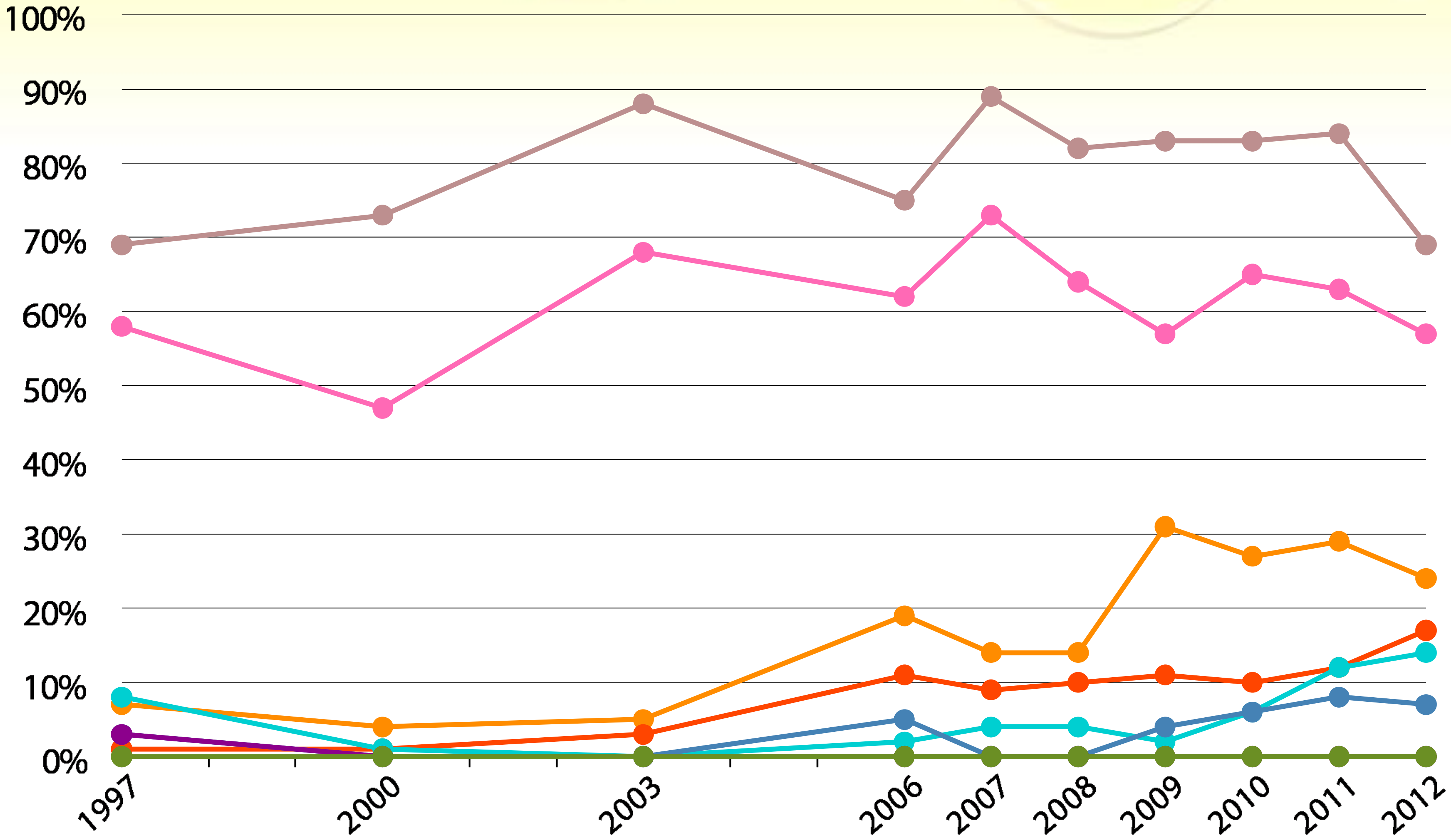
Results: evaluation scenarios



IEEE Vis/SciVis: total numbers & percent of evaluation scenarios

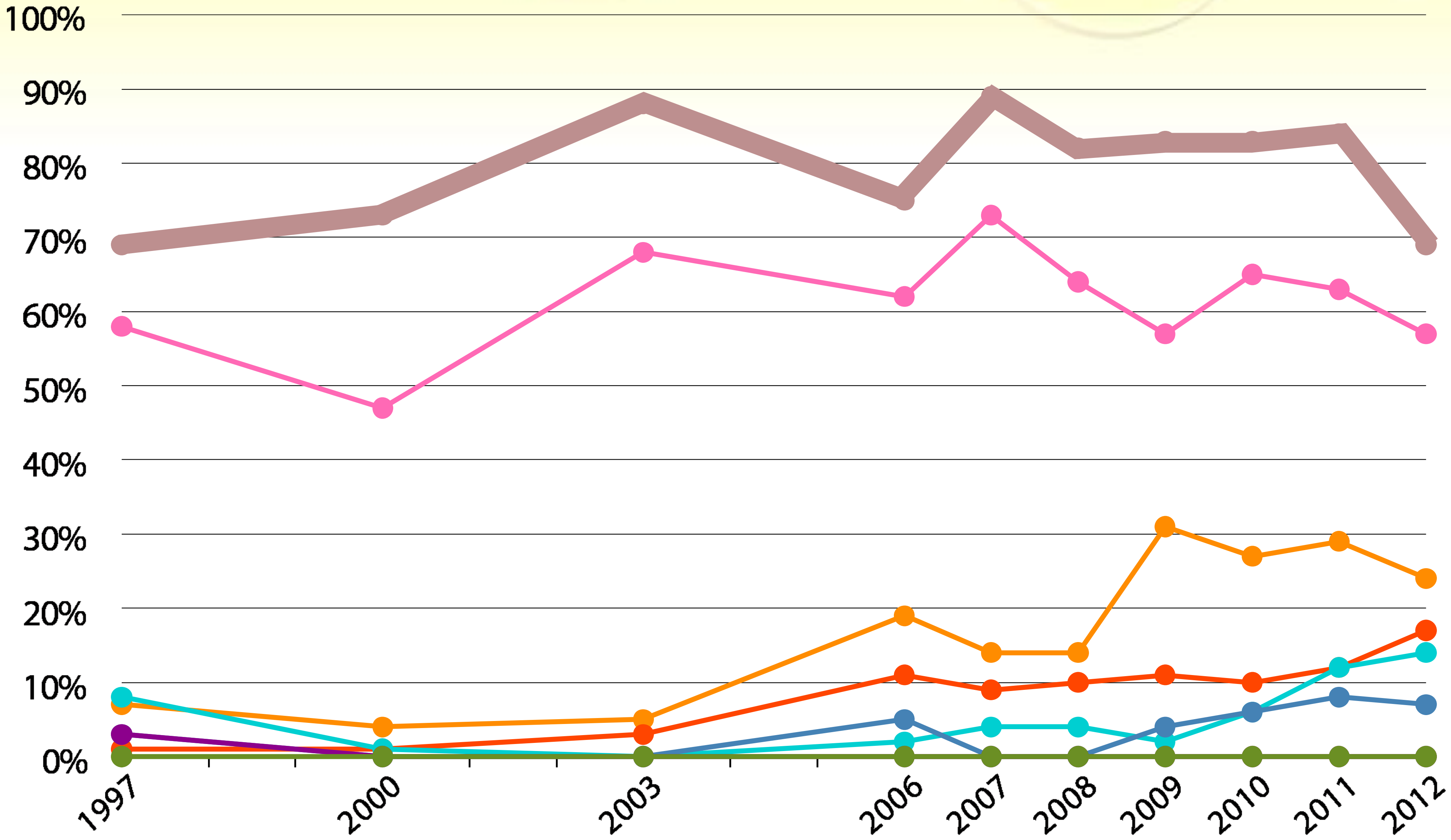
Results: historical development

% of papers with evaluation type: IEEE Vis/SciVis



Results: historical development

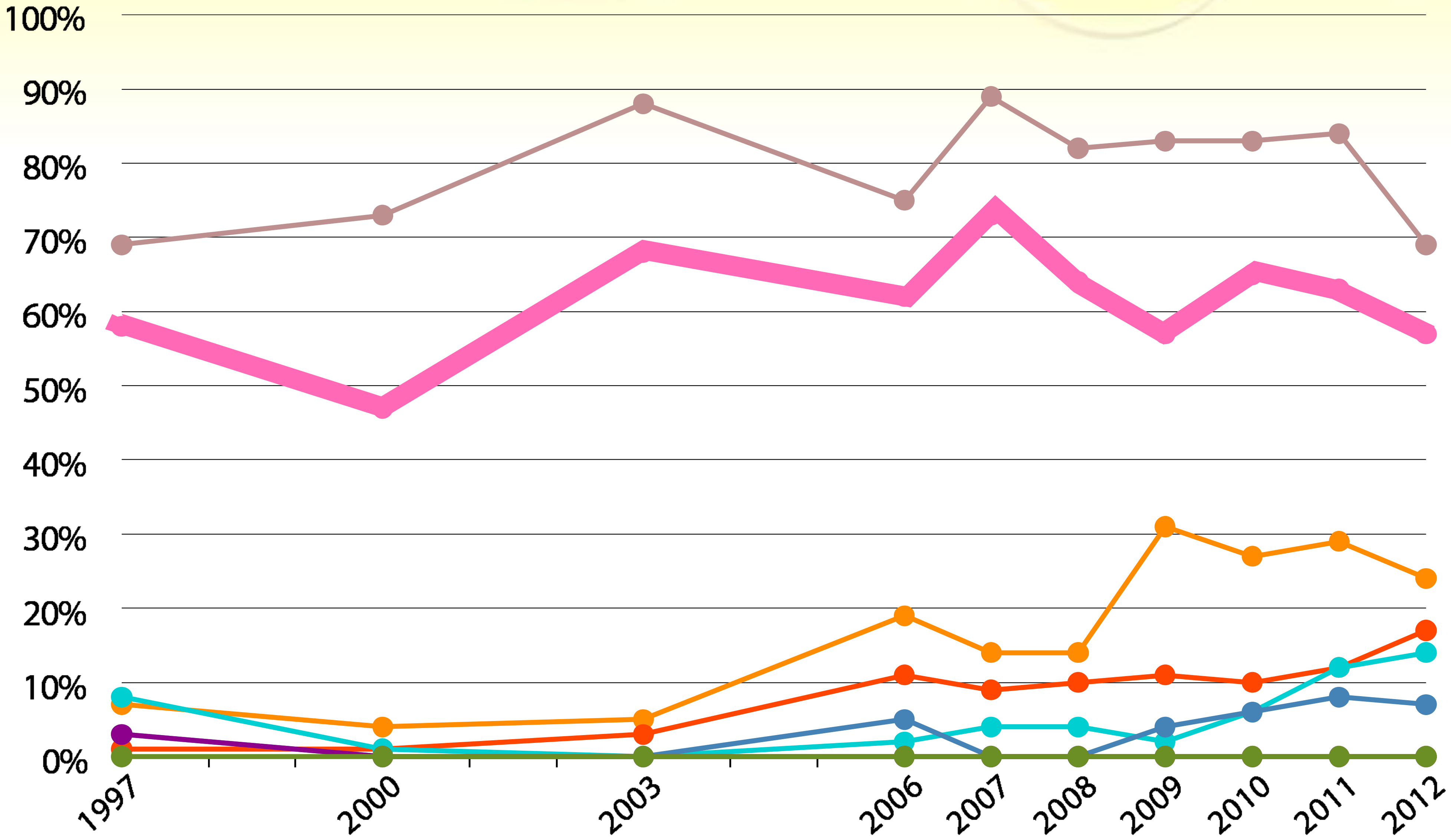
% of papers with evaluation type: IEEE Vis/SciVis



qualitative result inspection ≈ 80%

Results: historical development

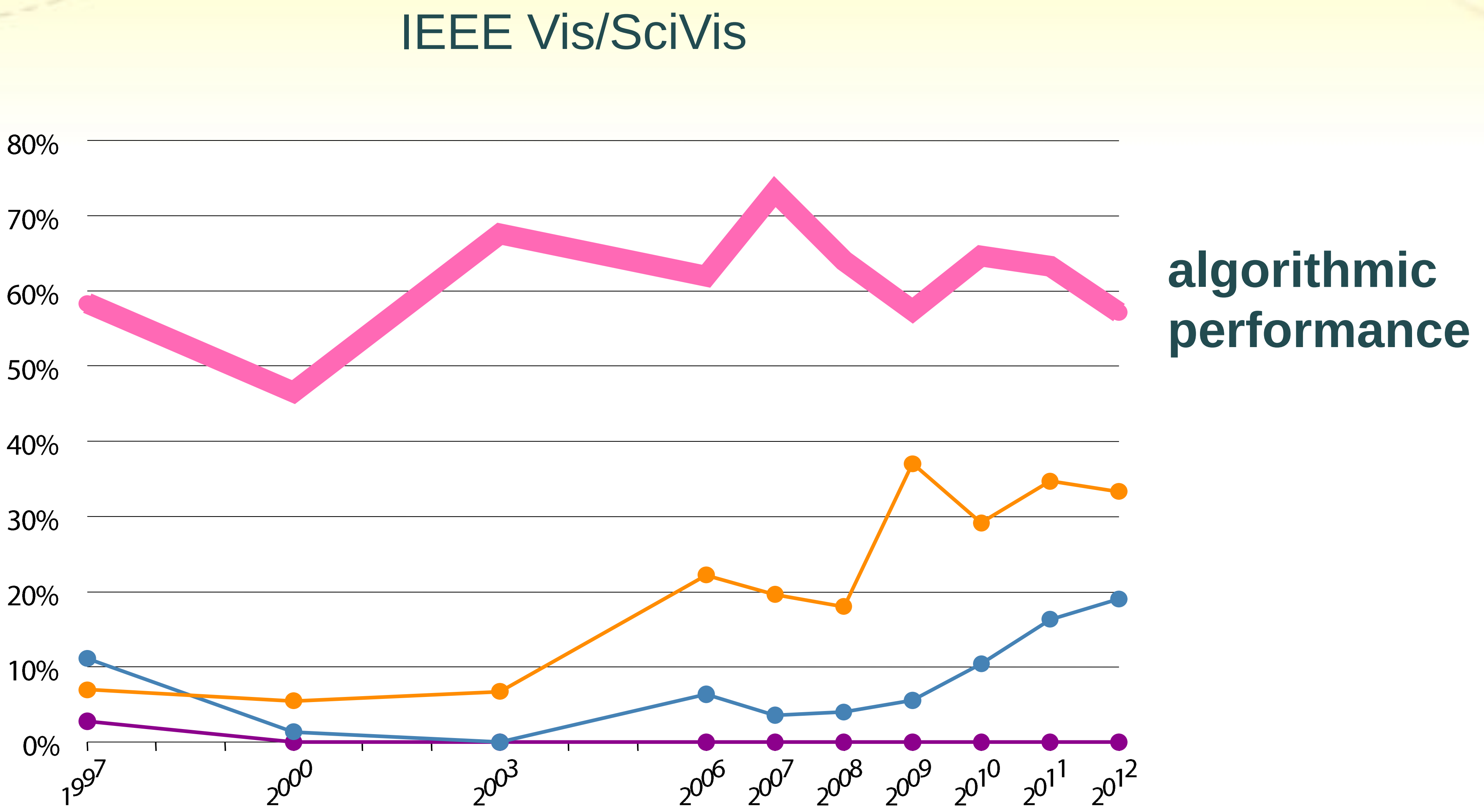
% of papers with evaluation type: IEEE Vis/SciVis



qualitative result inspection ≈ 80%

algorithmic performance ≈ 60%

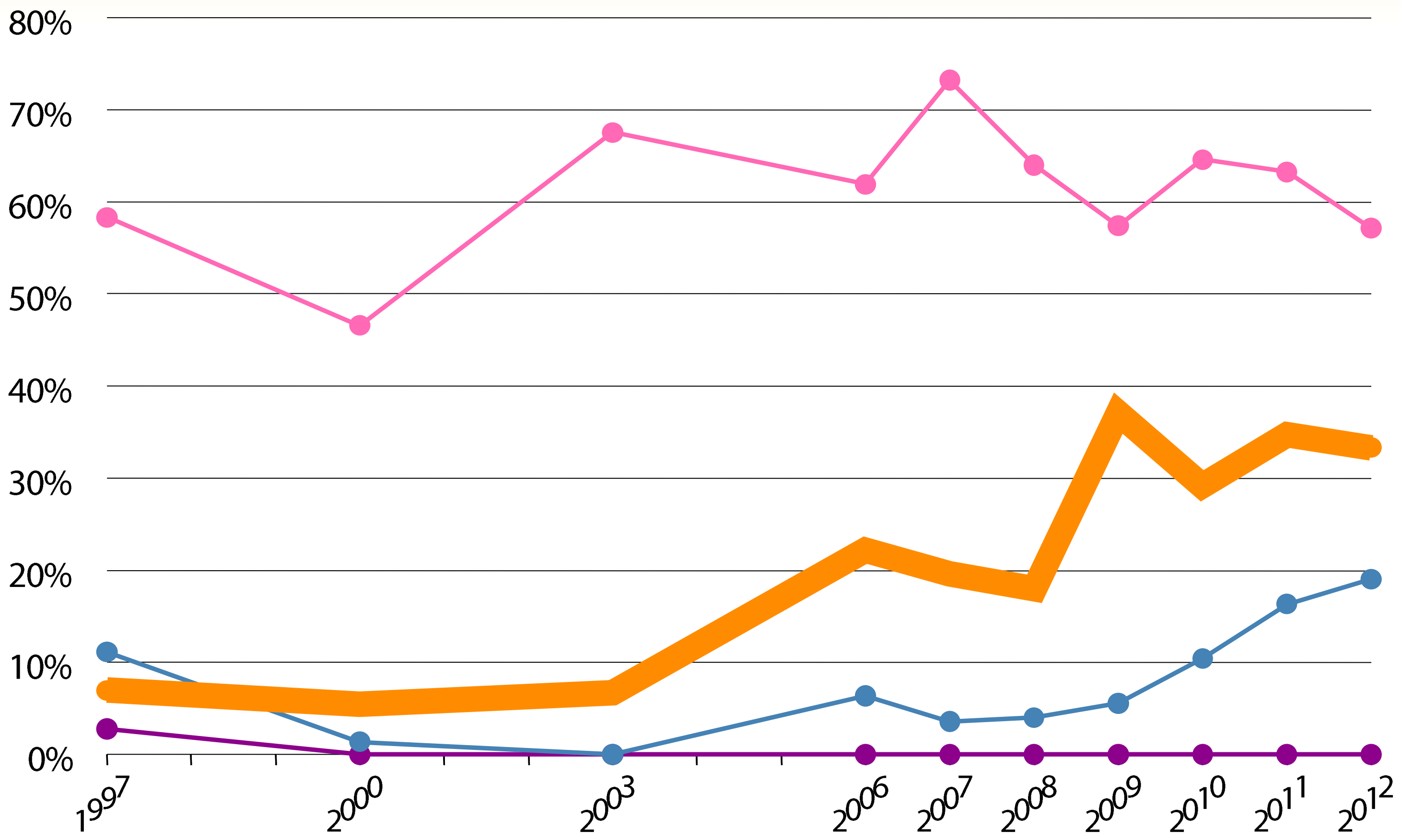
Results: historical development



evaluation type in percent of papers

Results: historical development

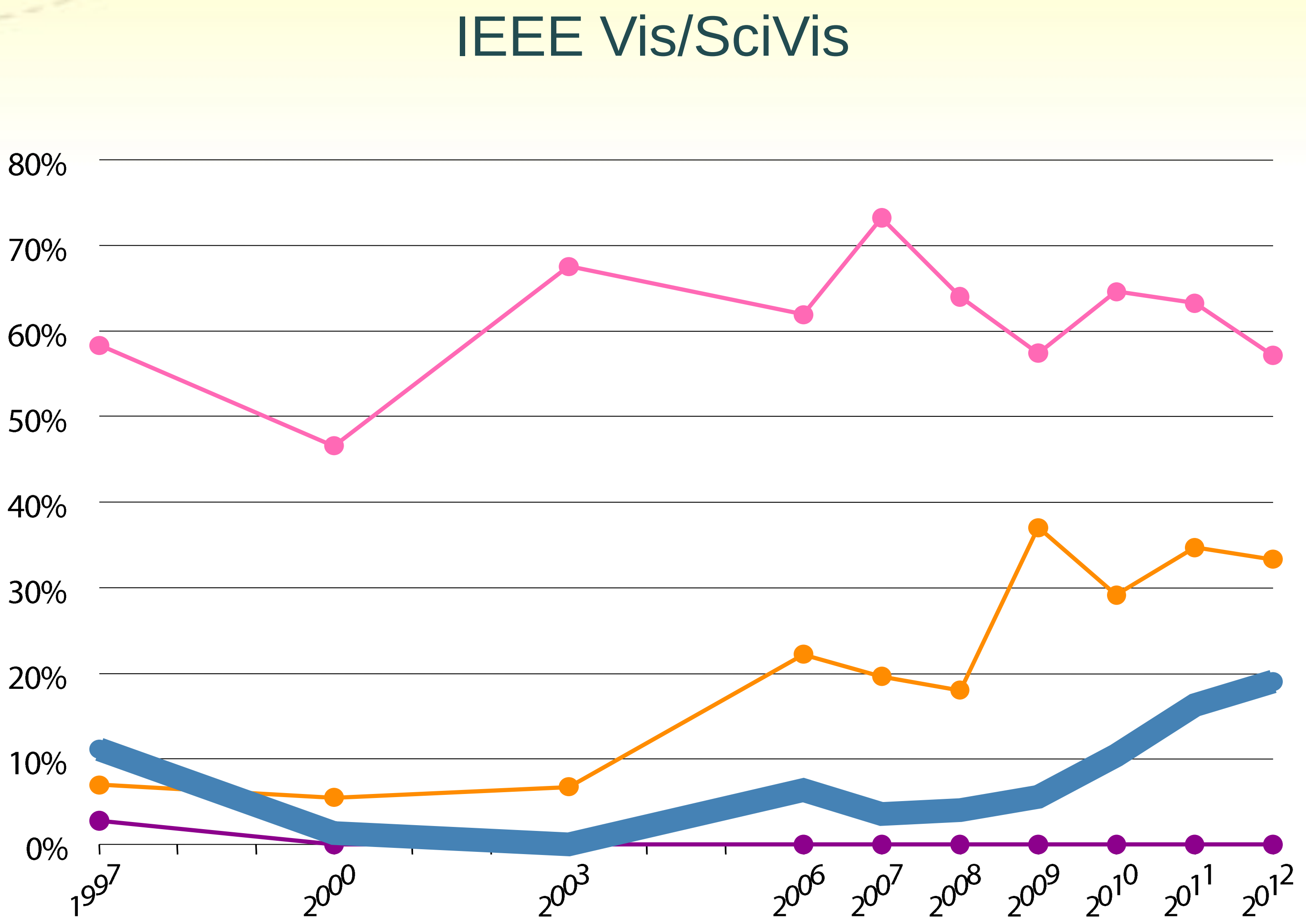
IEEE Vis/SciVis



user performance + user experience

evaluation type in percent of papers

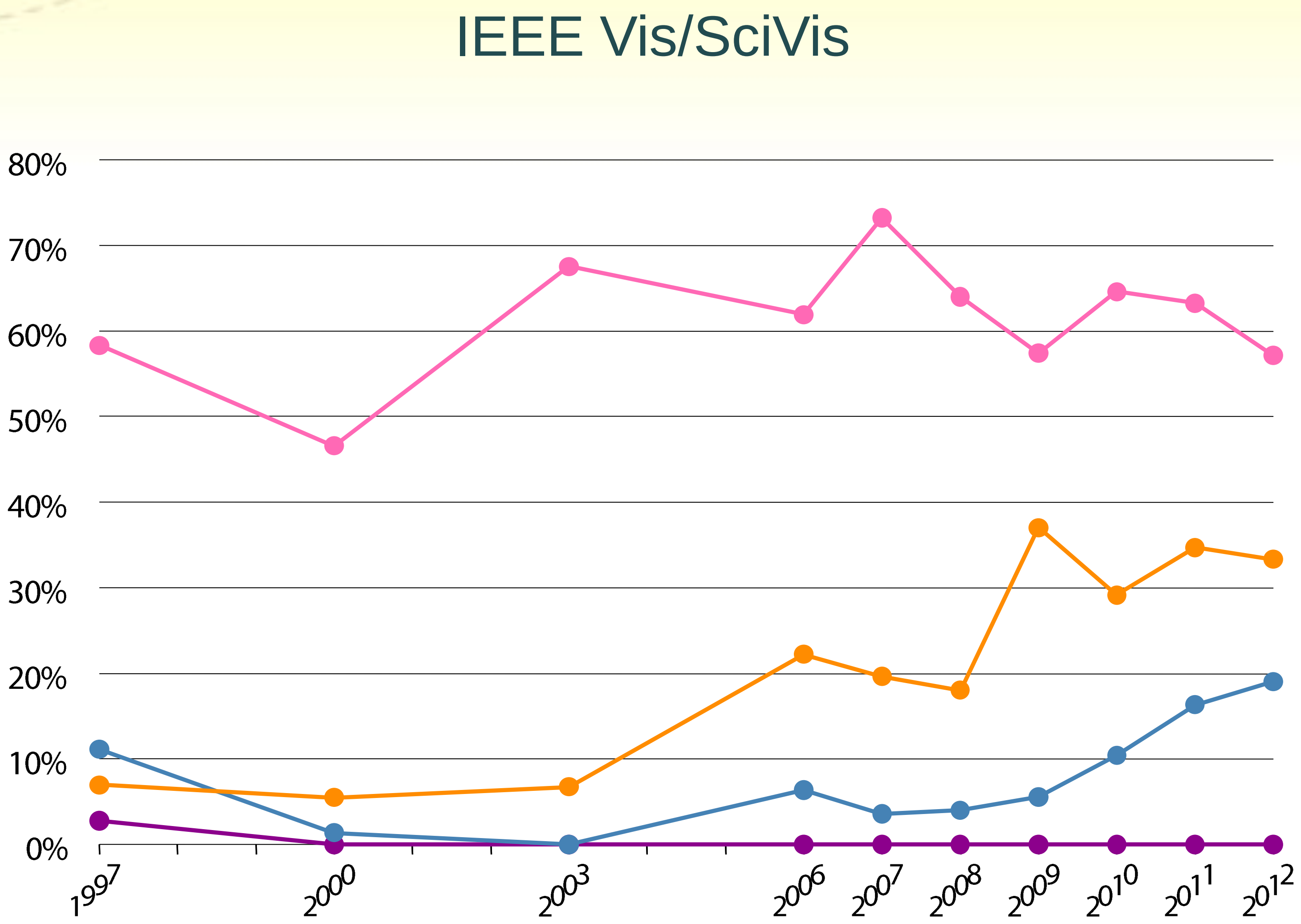
Results: historical development



work processes + analysis & reasoning

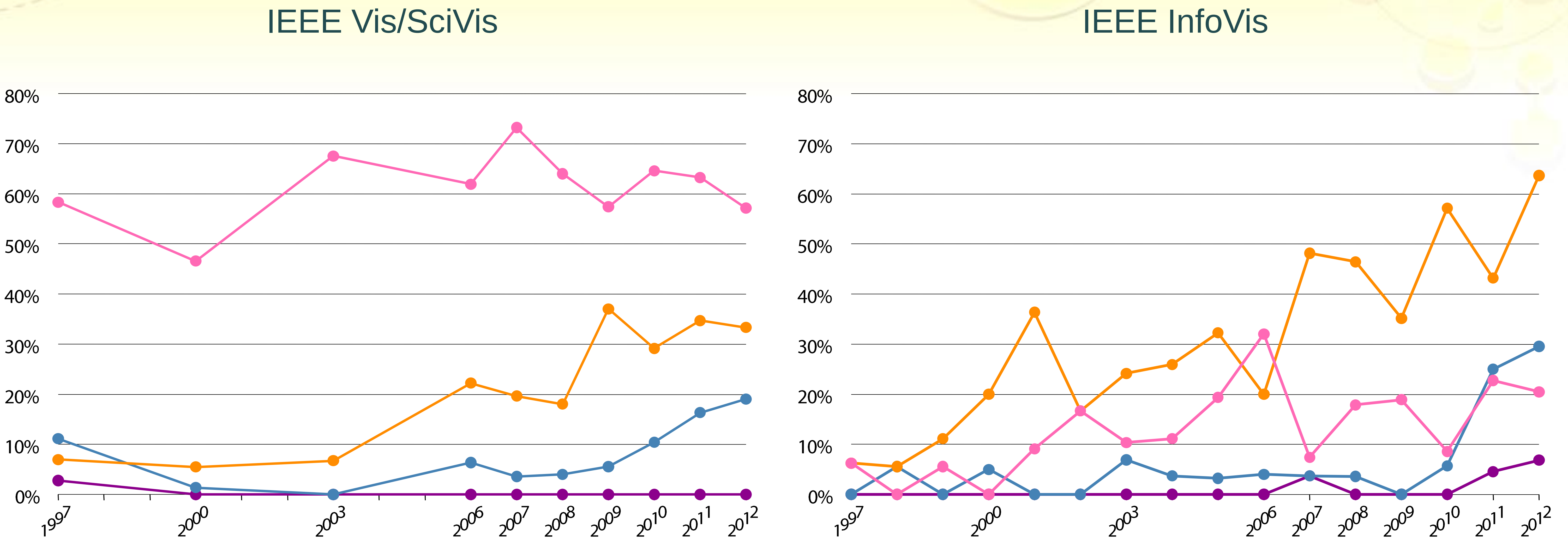
evaluation type in percent of papers

Results: historical development



evaluation type in percent of papers

Results: historical development

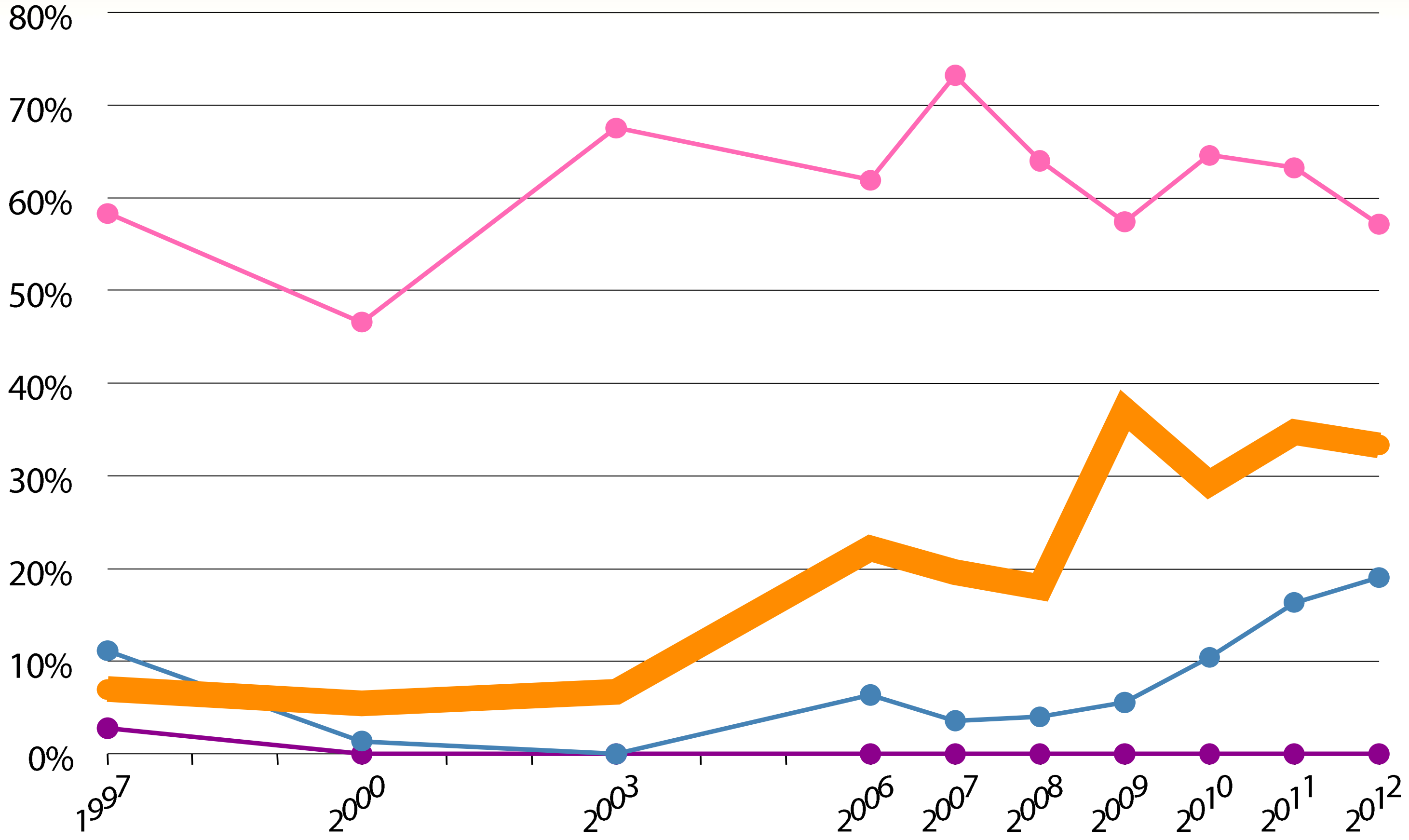


evaluation type in percent of papers

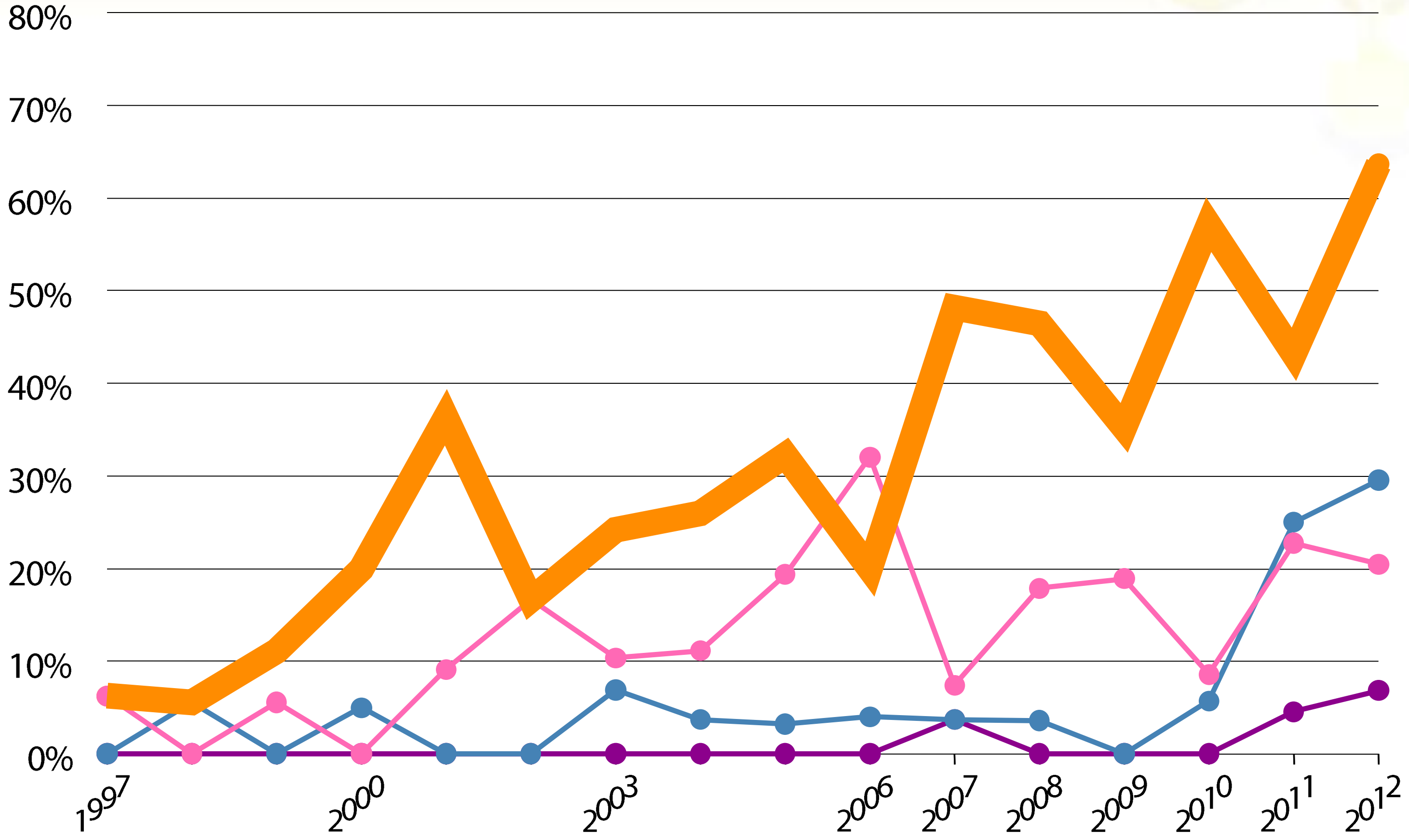
Results: historical development

user performance + user experience

IEEE Vis/SciVis



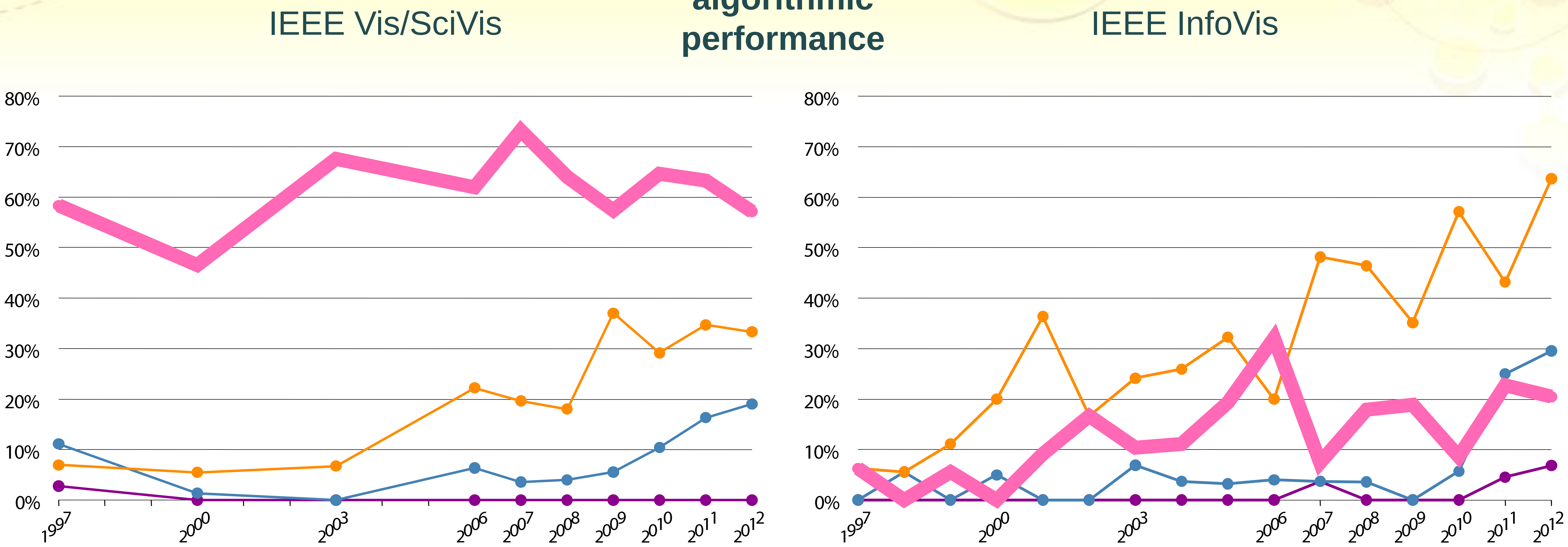
IEEE InfoVis



evaluation type in percent of papers

Results: historical development

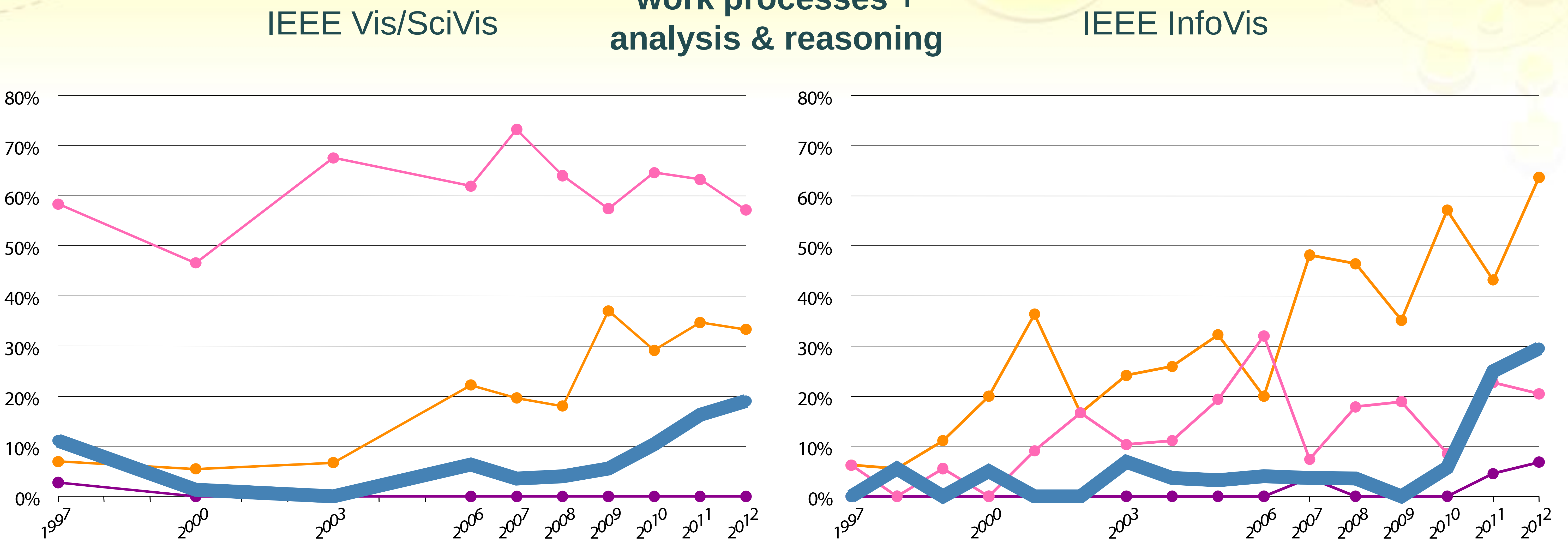
algorithmic performance



evaluation type in percent of papers

Results: historical development

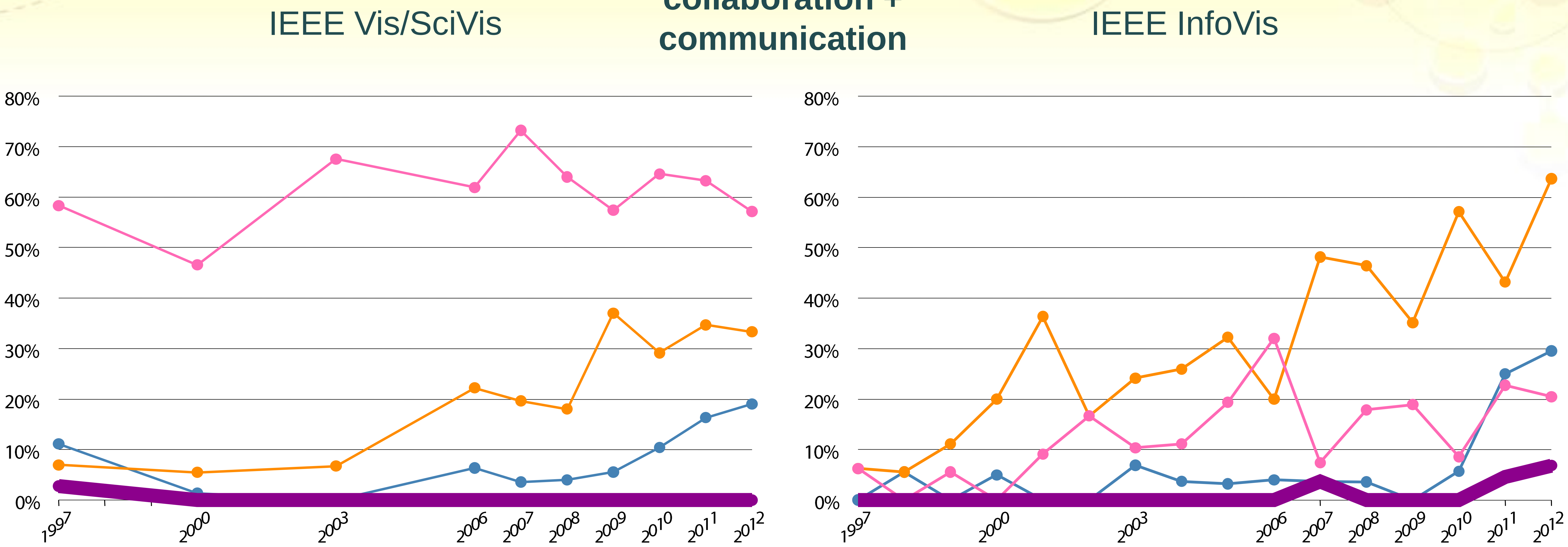
work processes + analysis & reasoning



evaluation type in percent of papers

Results: historical development

collaboration + communication

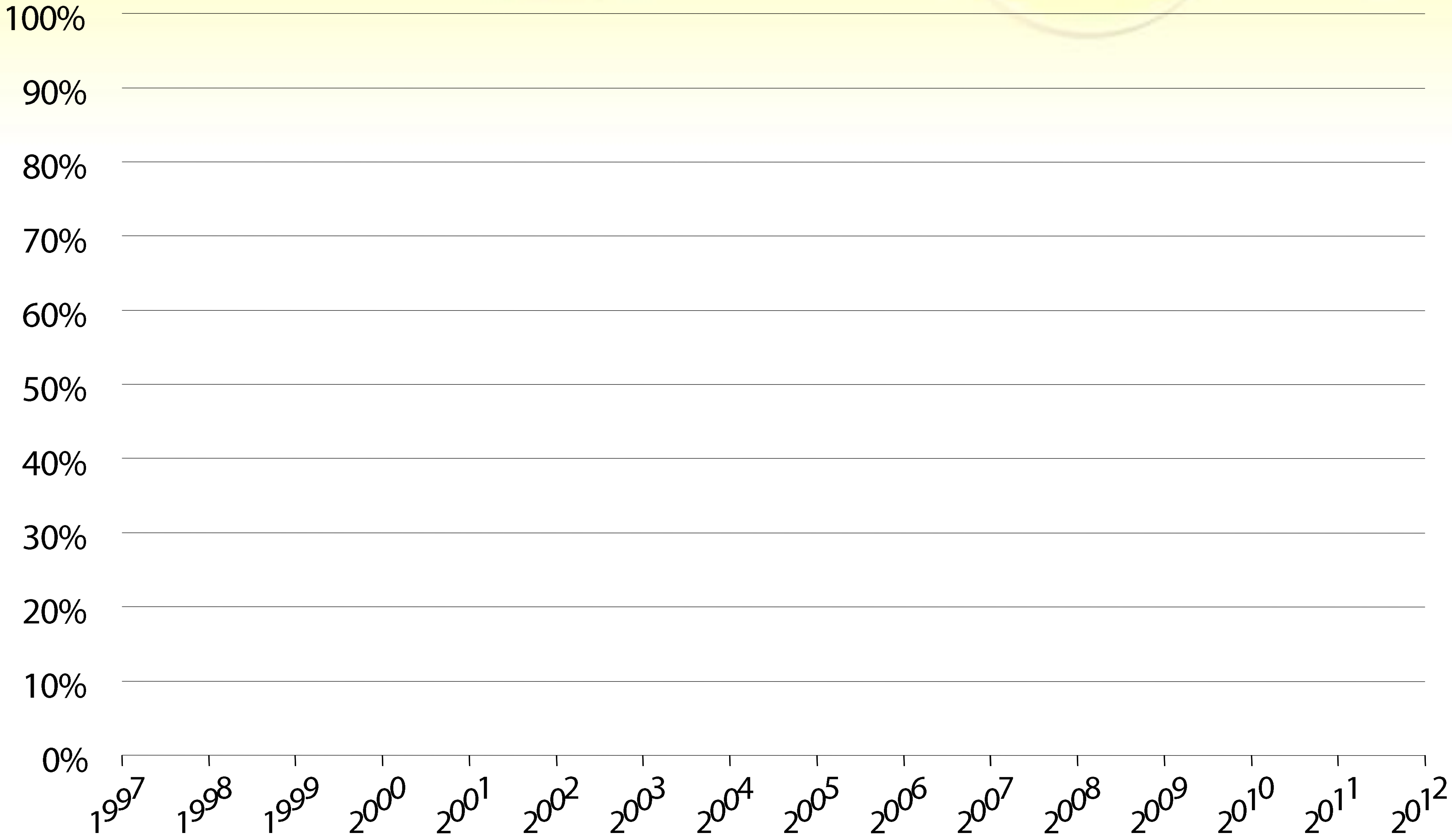


evaluation type in percent of papers

Results: human participants

original seven scenarios;

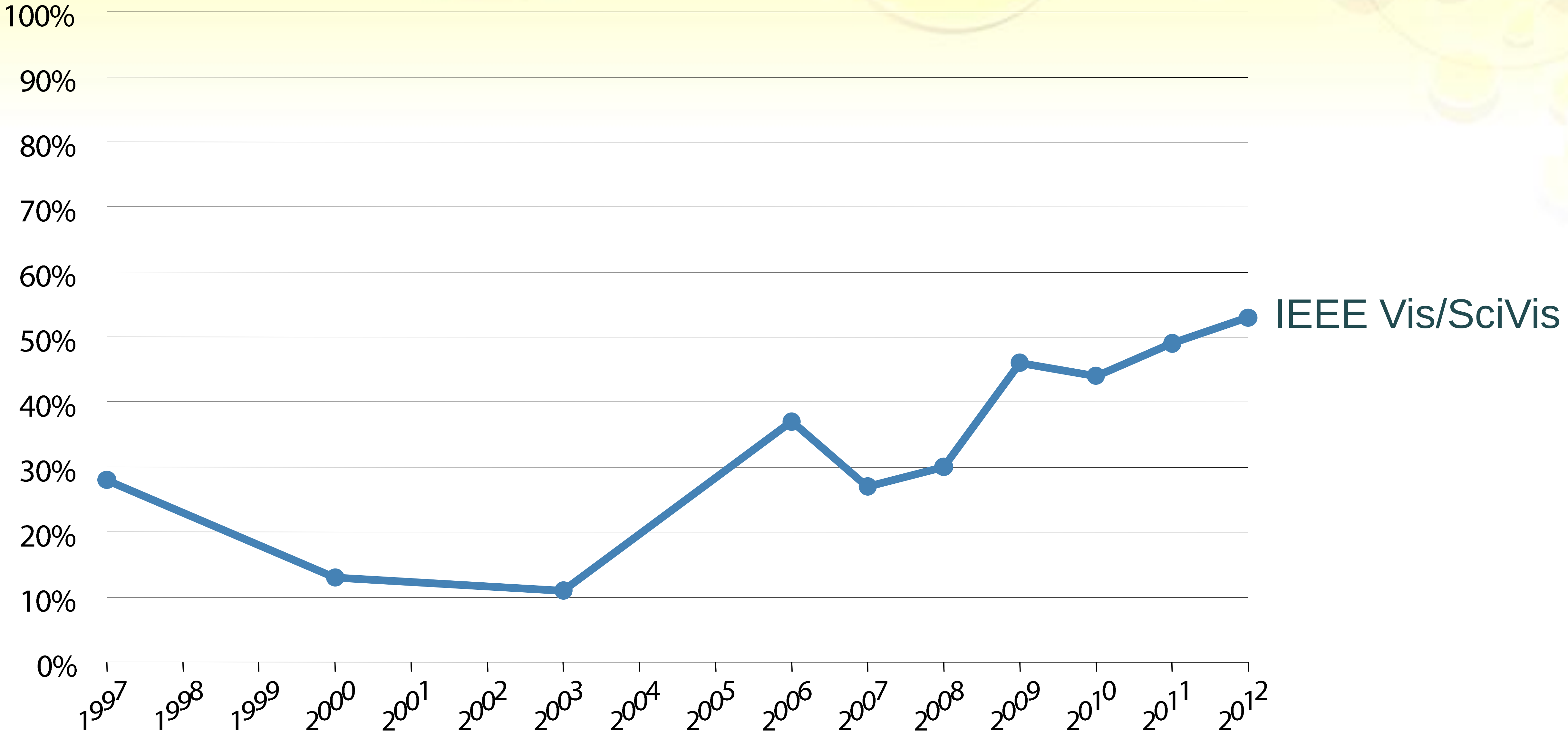
percent of scenarios with human participants (i.e., except algorithmic performance)



Results: human participants

original seven scenarios;

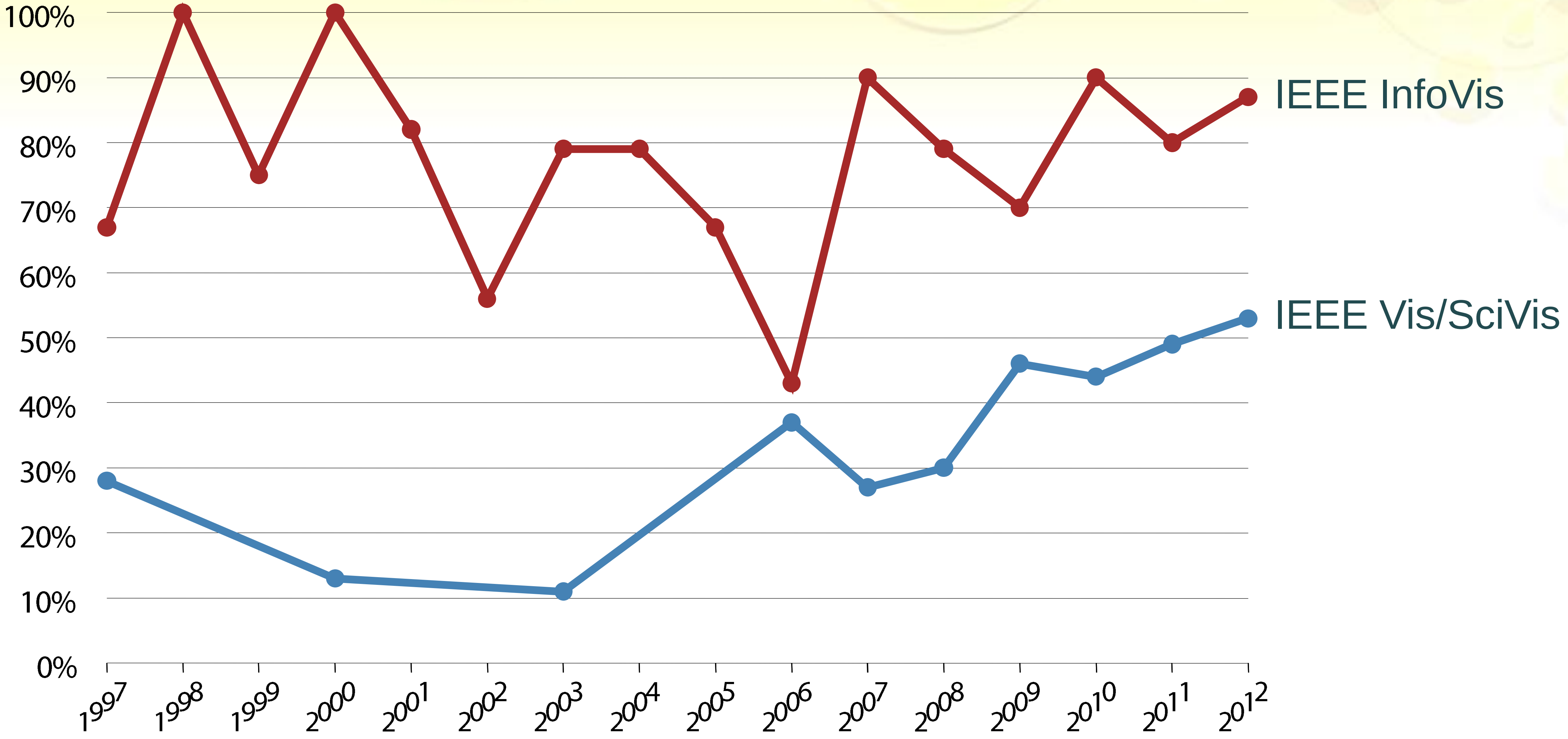
percent of scenarios with human participants (i.e., except algorithmic performance)



Results: human participants

original seven scenarios;

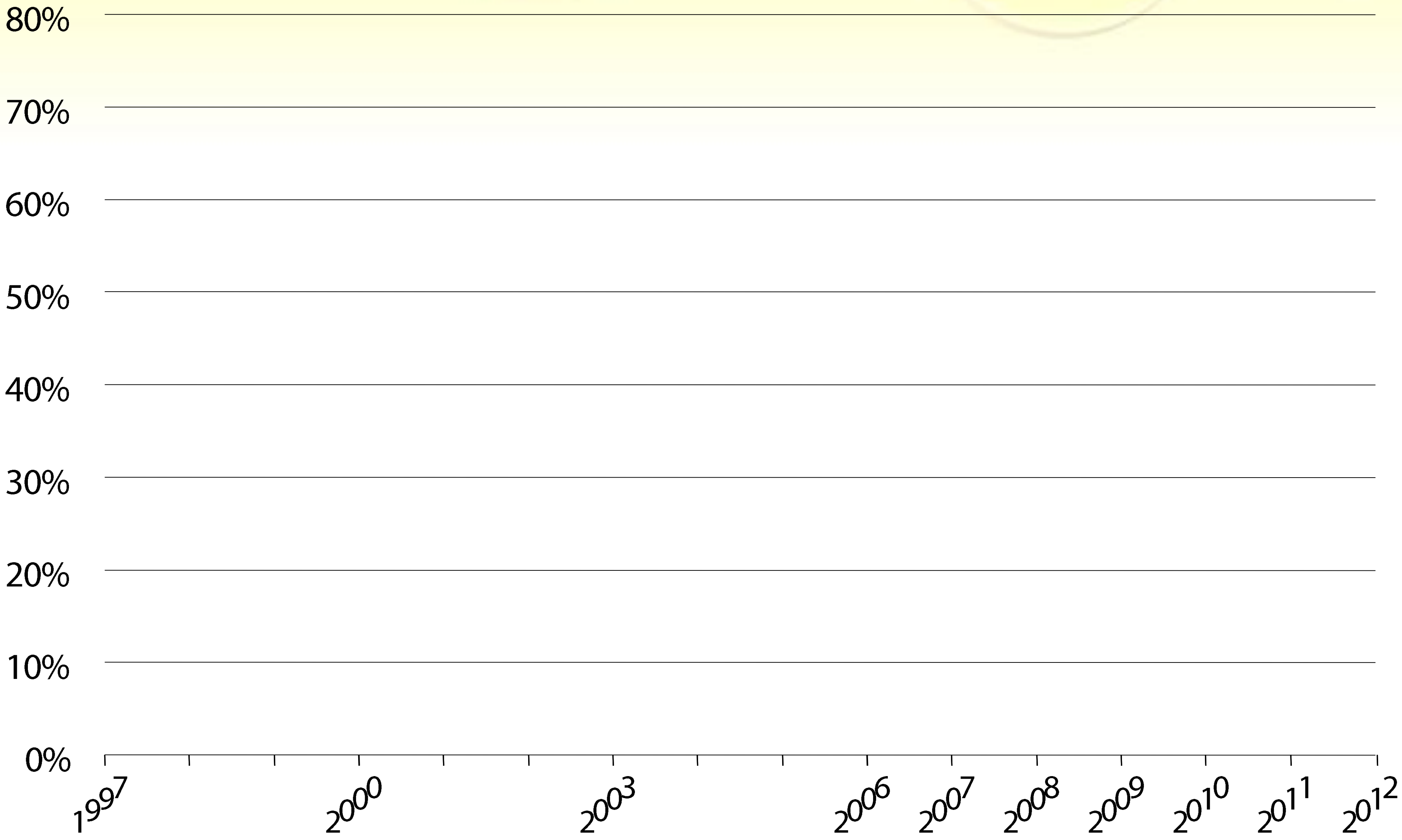
percent of scenarios with human participants (i.e., except algorithmic performance)



Results: visualization community as a whole

evaluation type
in percent
of papers

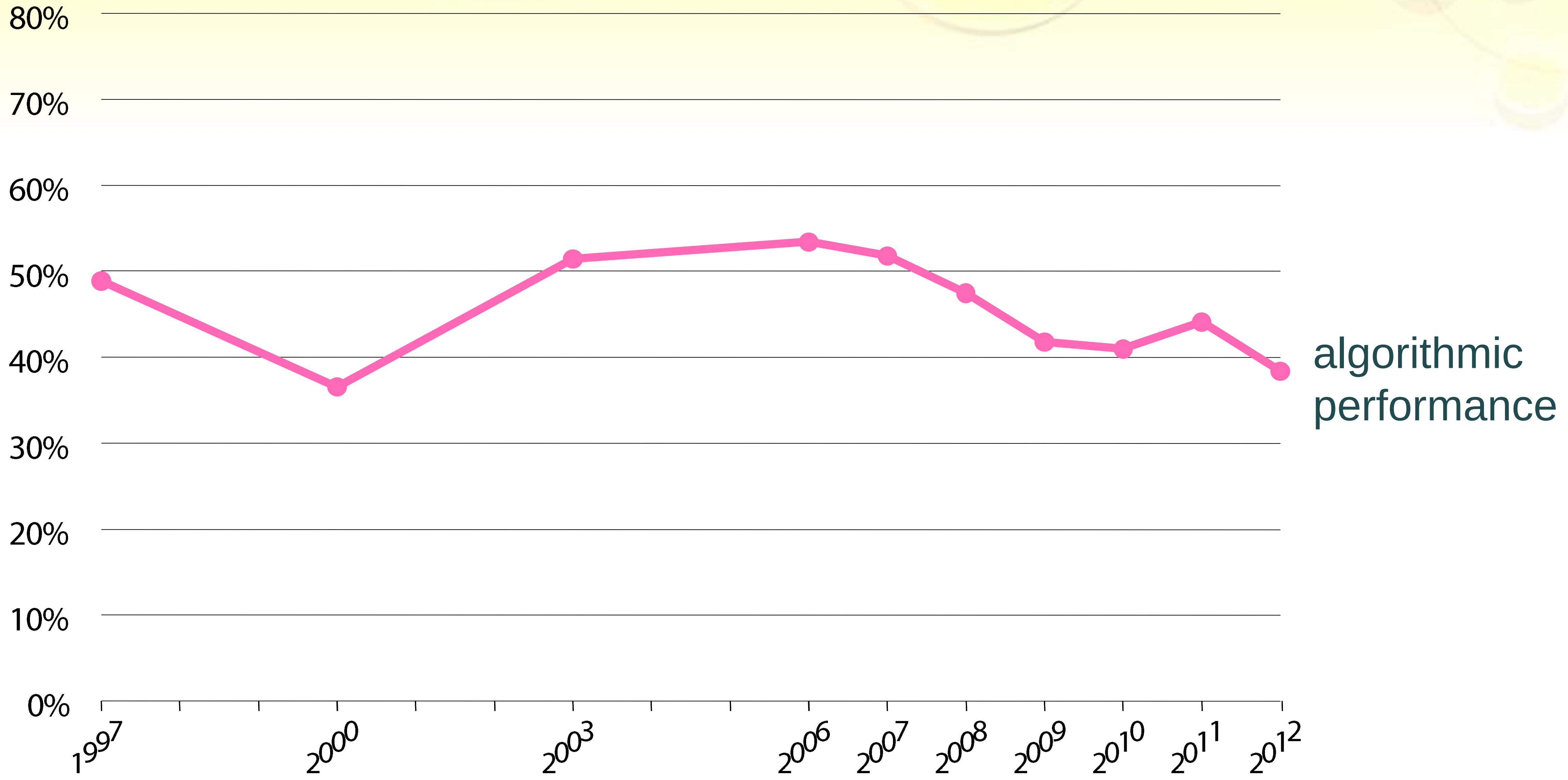
IEEE InfoVis
plus
IEEE Vis/SciVis



Results: visualization community as a whole

evaluation type in percent of papers

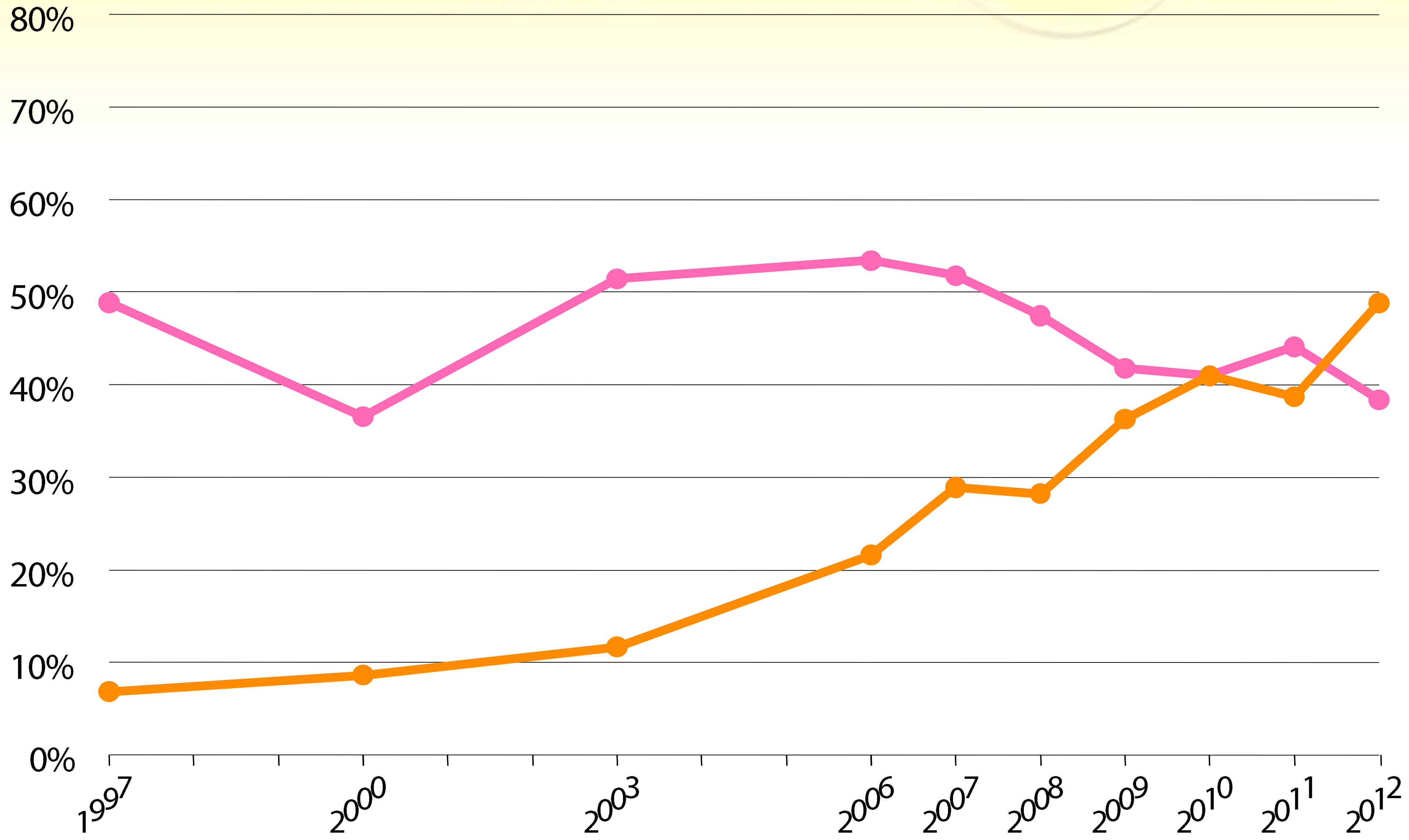
IEEE InfoVis plus IEEE Vis/SciVis



Results: visualization community as a whole

evaluation type
in percent
of papers

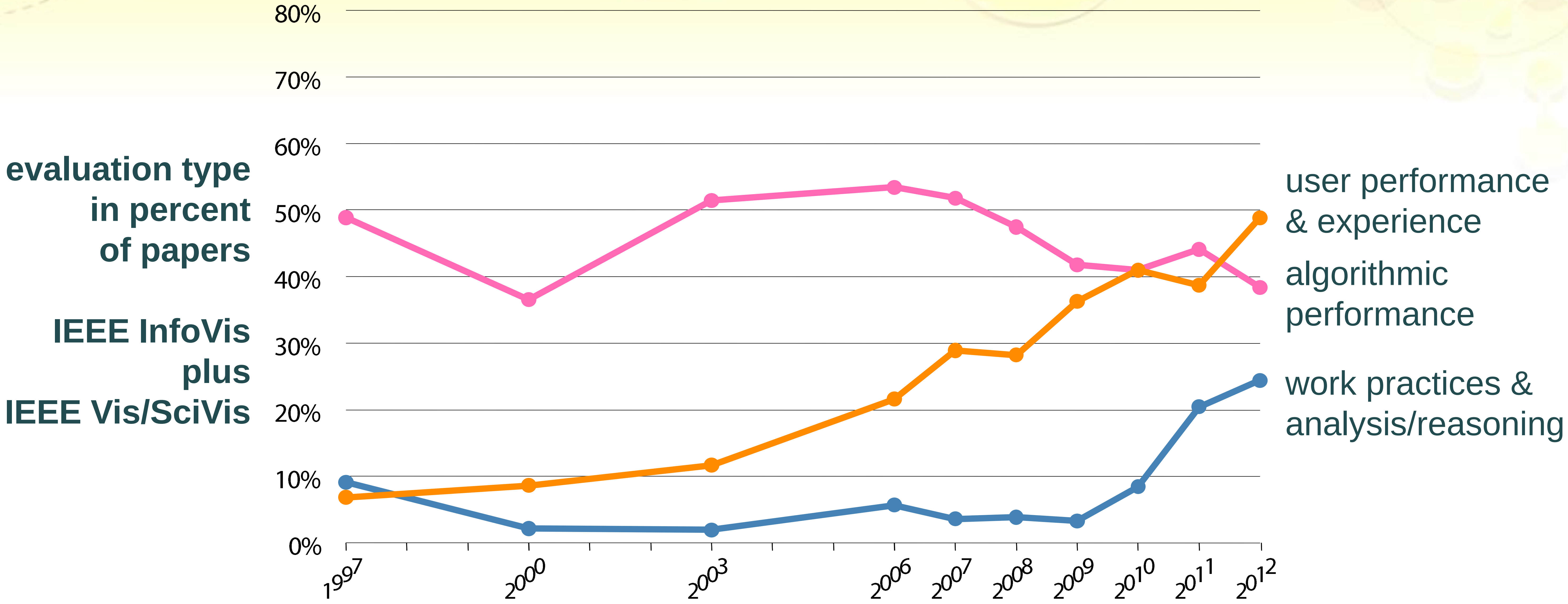
IEEE InfoVis
plus
IEEE Vis/SciVis



user performance
& experience

algorithmic
performance

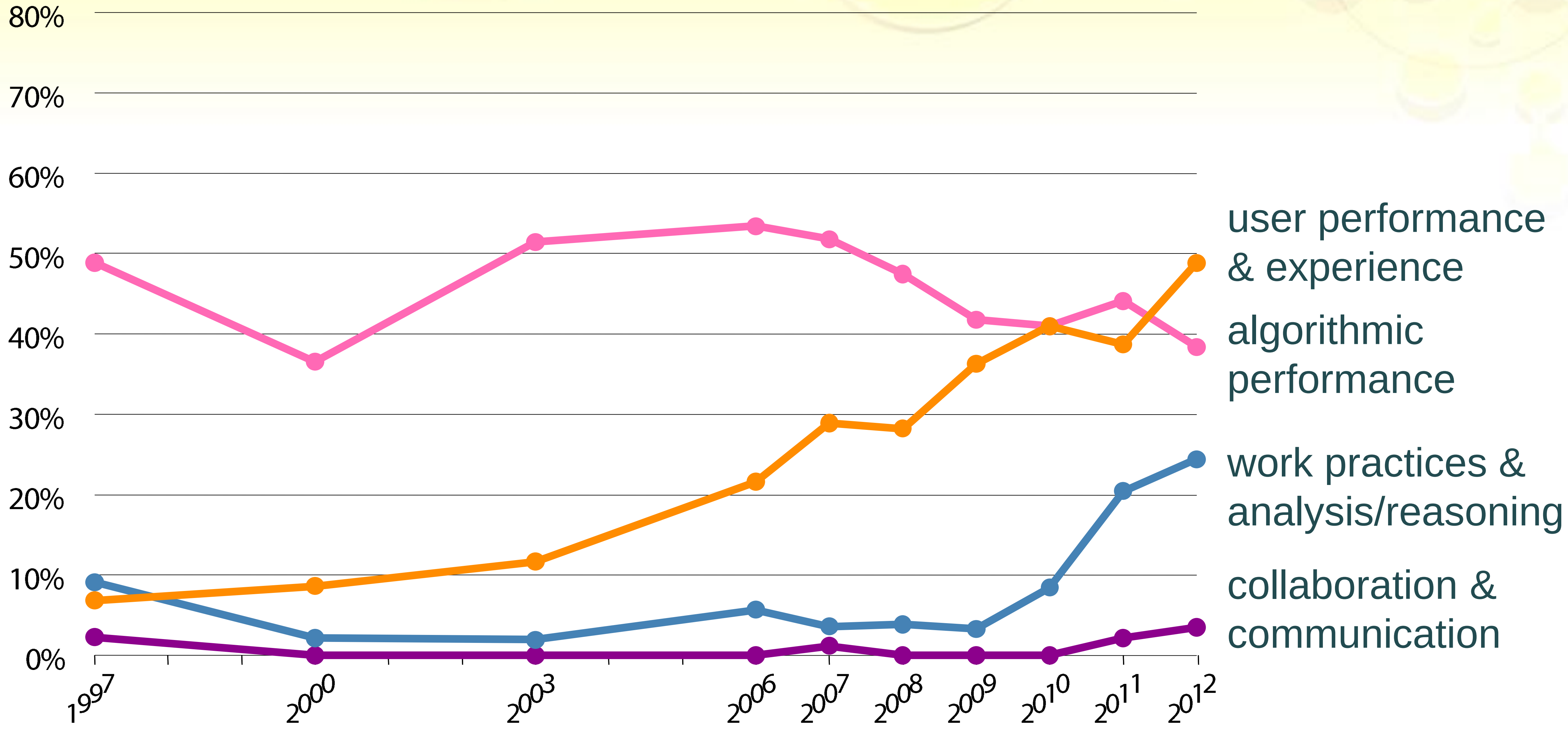
Results: visualization community as a whole



Results: visualization community as a whole

evaluation type in percent of papers

IEEE InfoVis plus IEEE Vis/SciVis



Isenberg et al.: A Systematic Review on the Practice of Evaluating Visualization

Considerations

- evaluation reporting rigor
- analyzing and reporting real problems
- statistical significance vs. qualitative expert feedback
- obtaining and reporting expert feedback
- use of case studies
- number of study participants

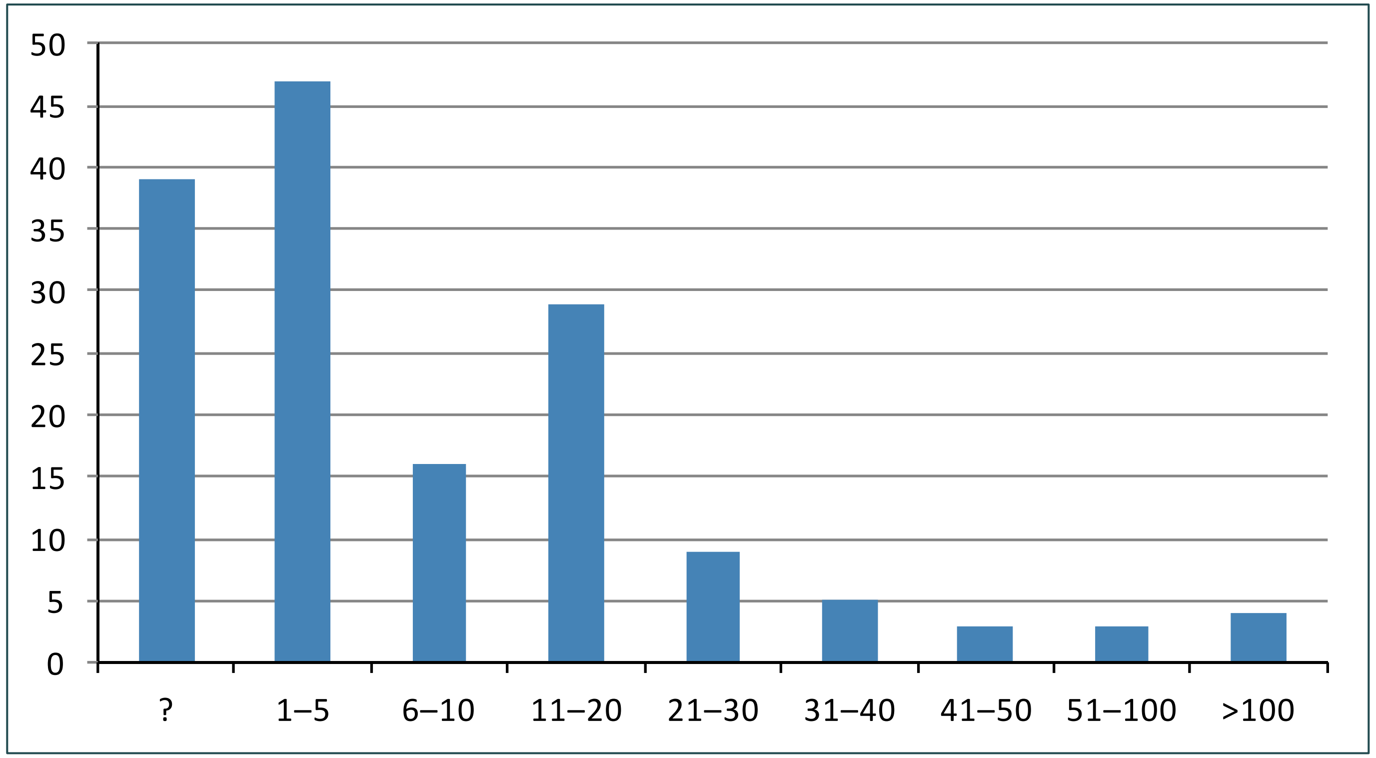
Considerations

- evaluation reporting rigor
- analyzing and reporting real problems
- statistical significance vs. qualitative expert feedback
- obtaining and reporting expert feedback
- use of case studies
- number of study participants

Evaluation reporting rigor

- too often detail missing about the evaluation

- need to report
  - participant details
  - collaboration details
  - evaluation protocols
  - …

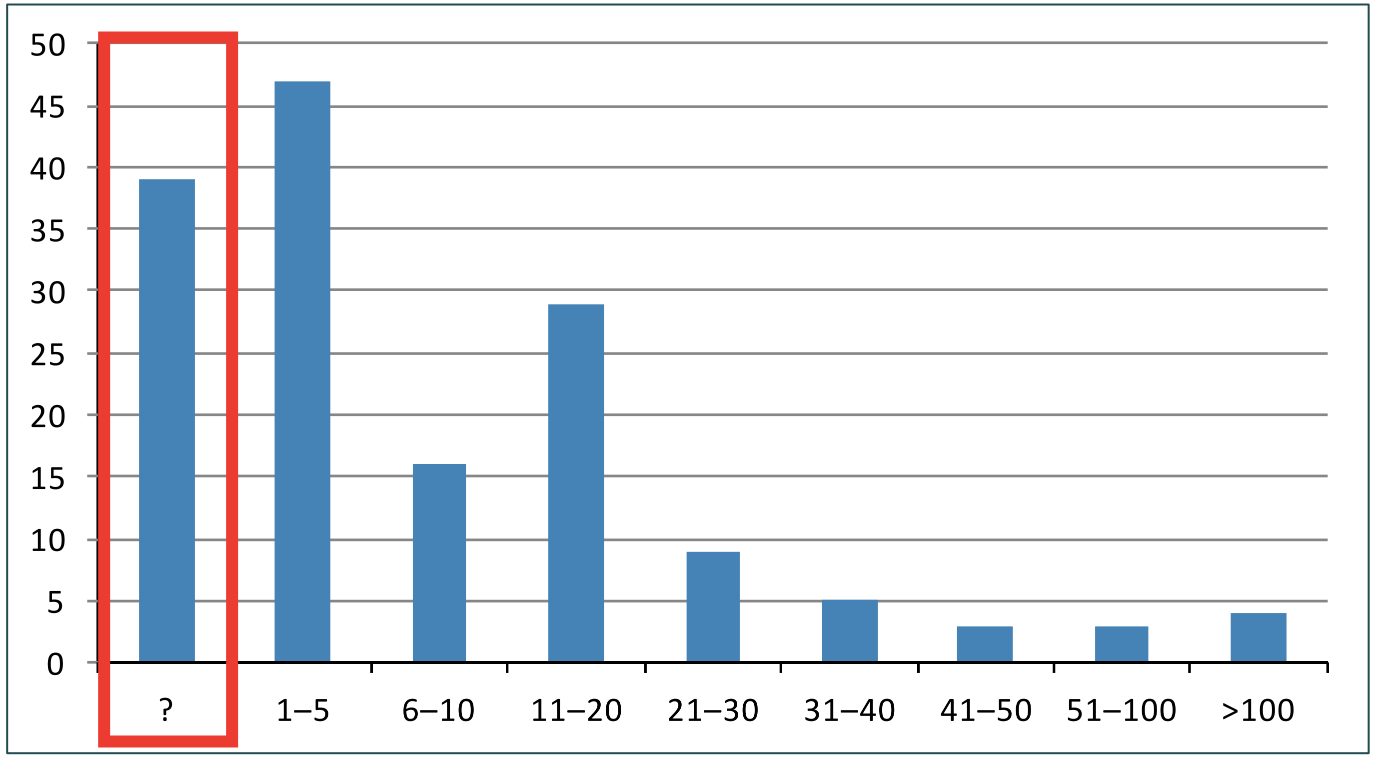


number of participants per study

Evaluation reporting rigor

- too often detail missing about the evaluation

- need to report
  - participant details
  - collaboration details
  - evaluation protocols
  - …

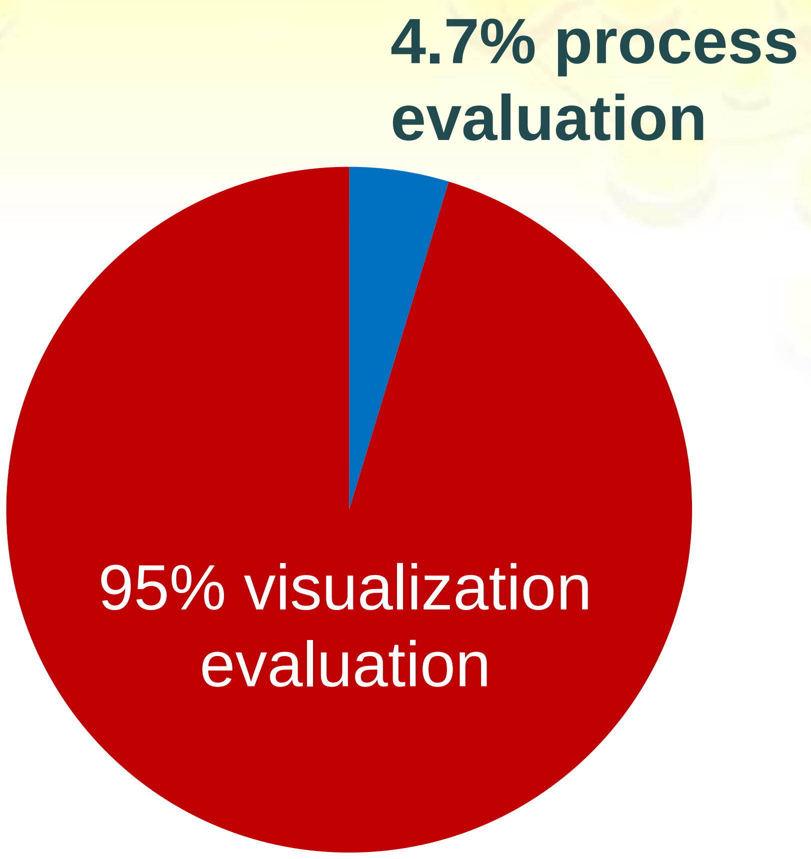


number of participants per study

>25% unclear

Analyzing & reporting real problems

- relevant to virtually all visualization work:
  - understand visualization needs
  - understand use of visualizations for visual reasoning, communication, and collaboration
  - grounding work in reality
- in practice: often done!
  - describe work with experts
  - make process evaluations first-class citizens in our papers



percentage of scenarios

Statistical significance vs. expert feedback

- right methodology
  for given question!
- visualization:
  ill-defined, fuzzy,
  broad domain problems
- evaluation **not only**
  null hypothesis
  significance testing (NHST)

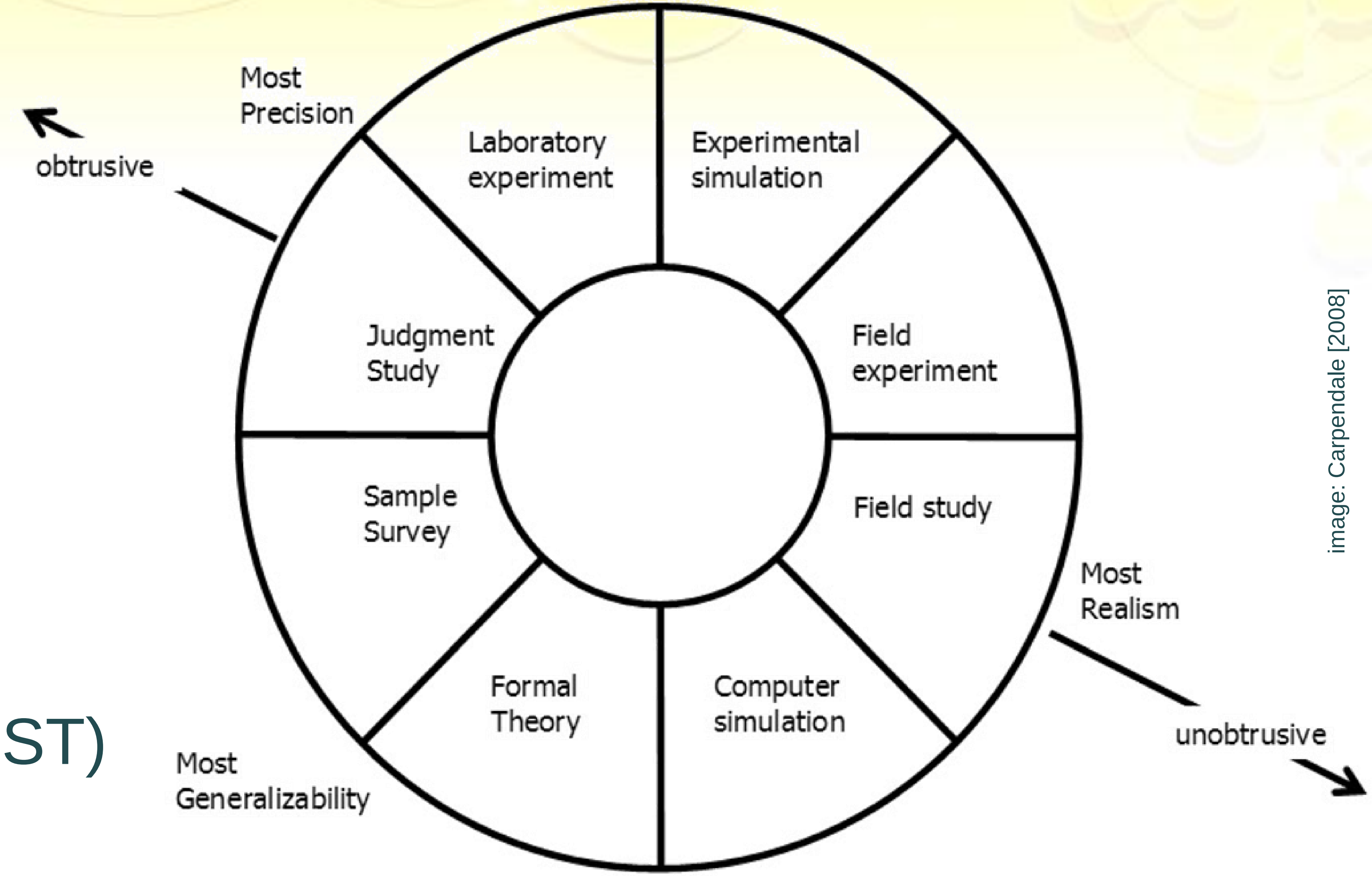


image: Carpendale [2008]

Isenberg et al.: A Systematic Review on the Practice of Evaluating Visualization

Statistical significance vs. expert feedback

- right methodology
  for given question!
- visualization:
  ill-defined, fuzzy,
  broad domain problems
- evaluation **not only**
  null hypothesis
  significance testing (NHST)

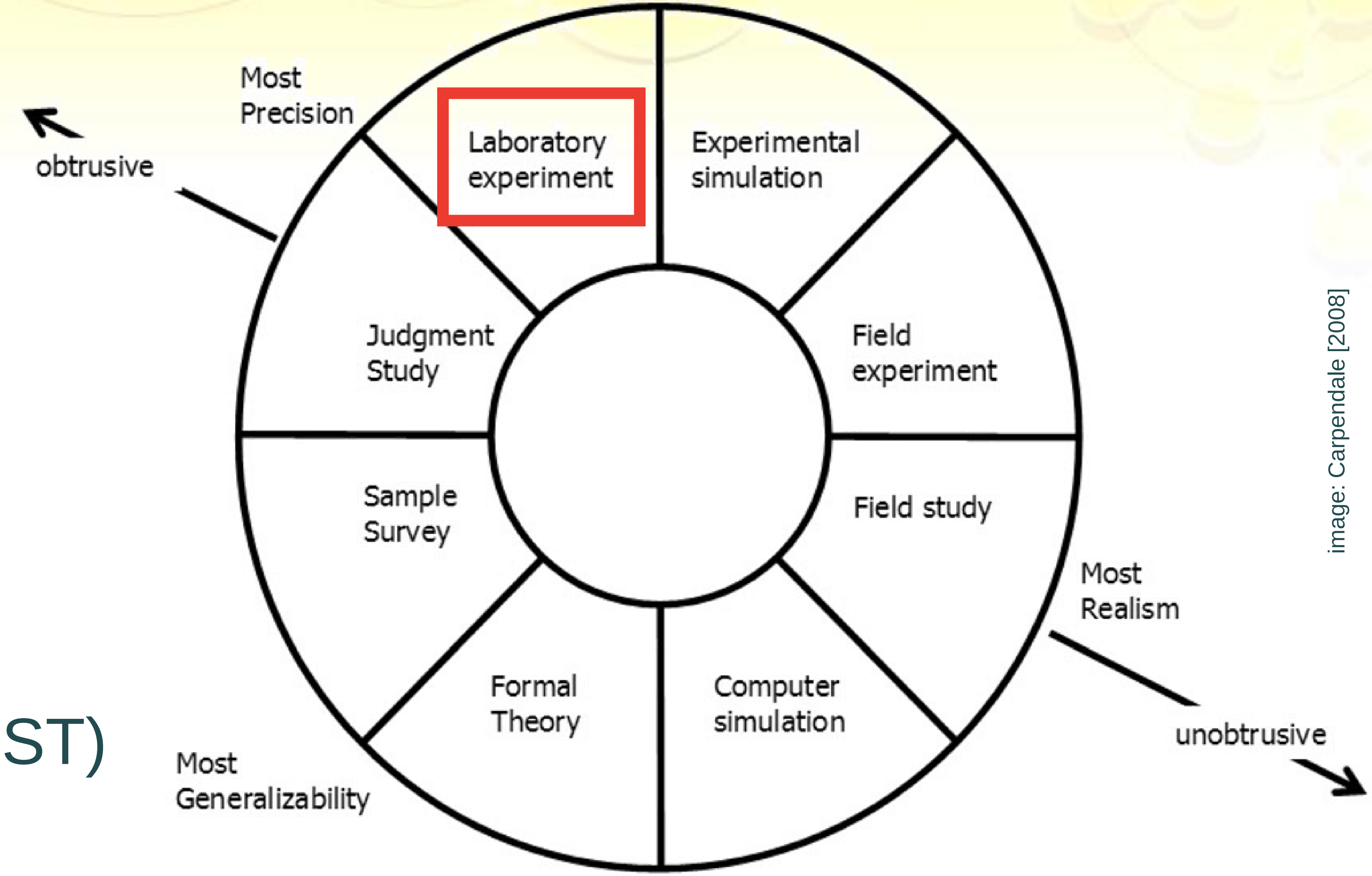


image: Carpendale [2008]

Statistical significance vs. expert feedback

- right methodology for given question!
- visualization: ill-defined, fuzzy, broad domain problems
- evaluation **not only** null hypothesis significance testing (NHST)

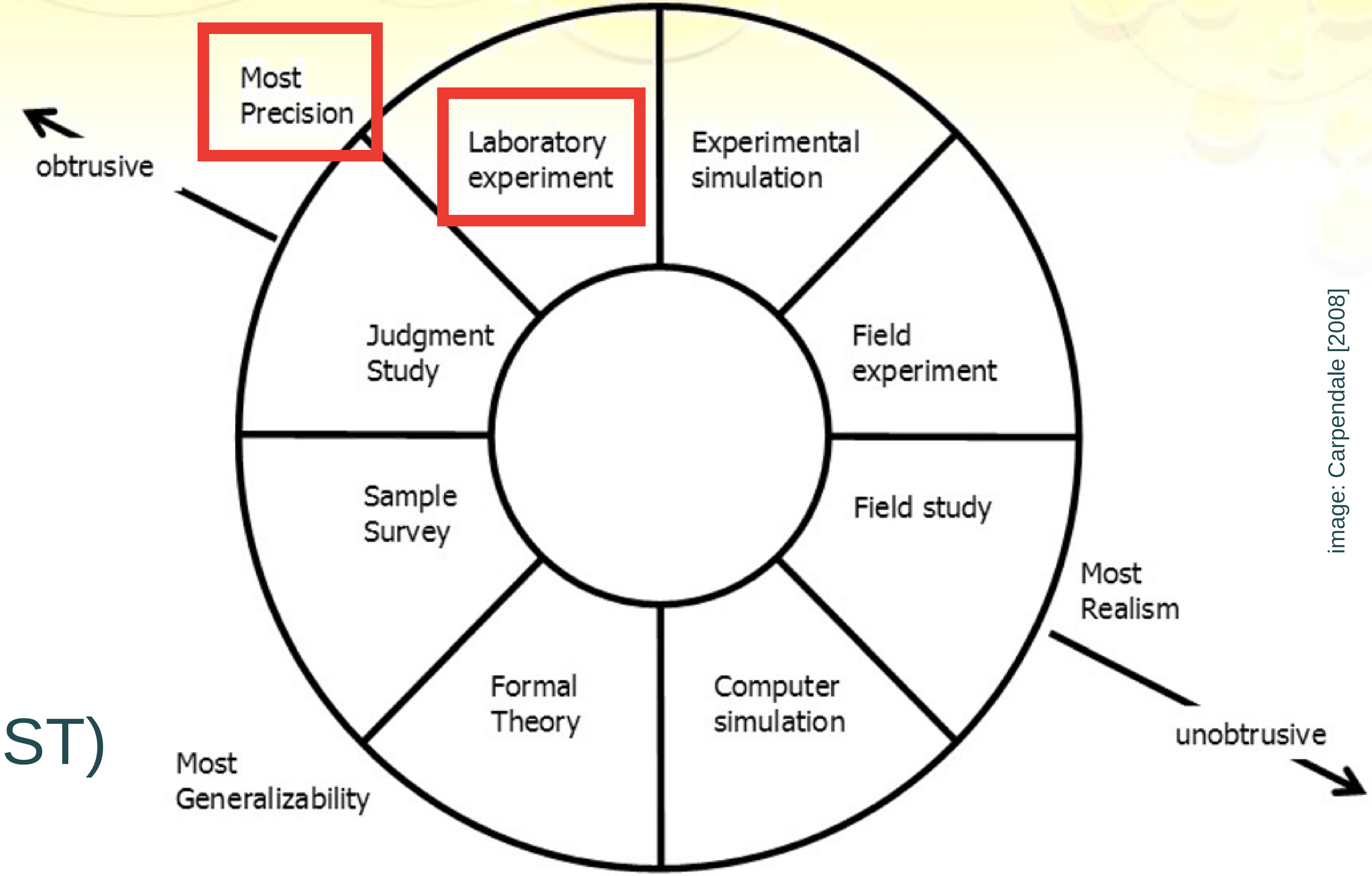


image: Carpendale [2008]

Isenberg et al.: A Systematic Review on the Practice of Evaluating Visualization

Statistical significance vs. expert feedback

- right methodology
  for given question!
- visualization:
  ill-defined, fuzzy,
  broad domain problems
- evaluation **not only**
  null hypothesis
  significance testing (NHST)

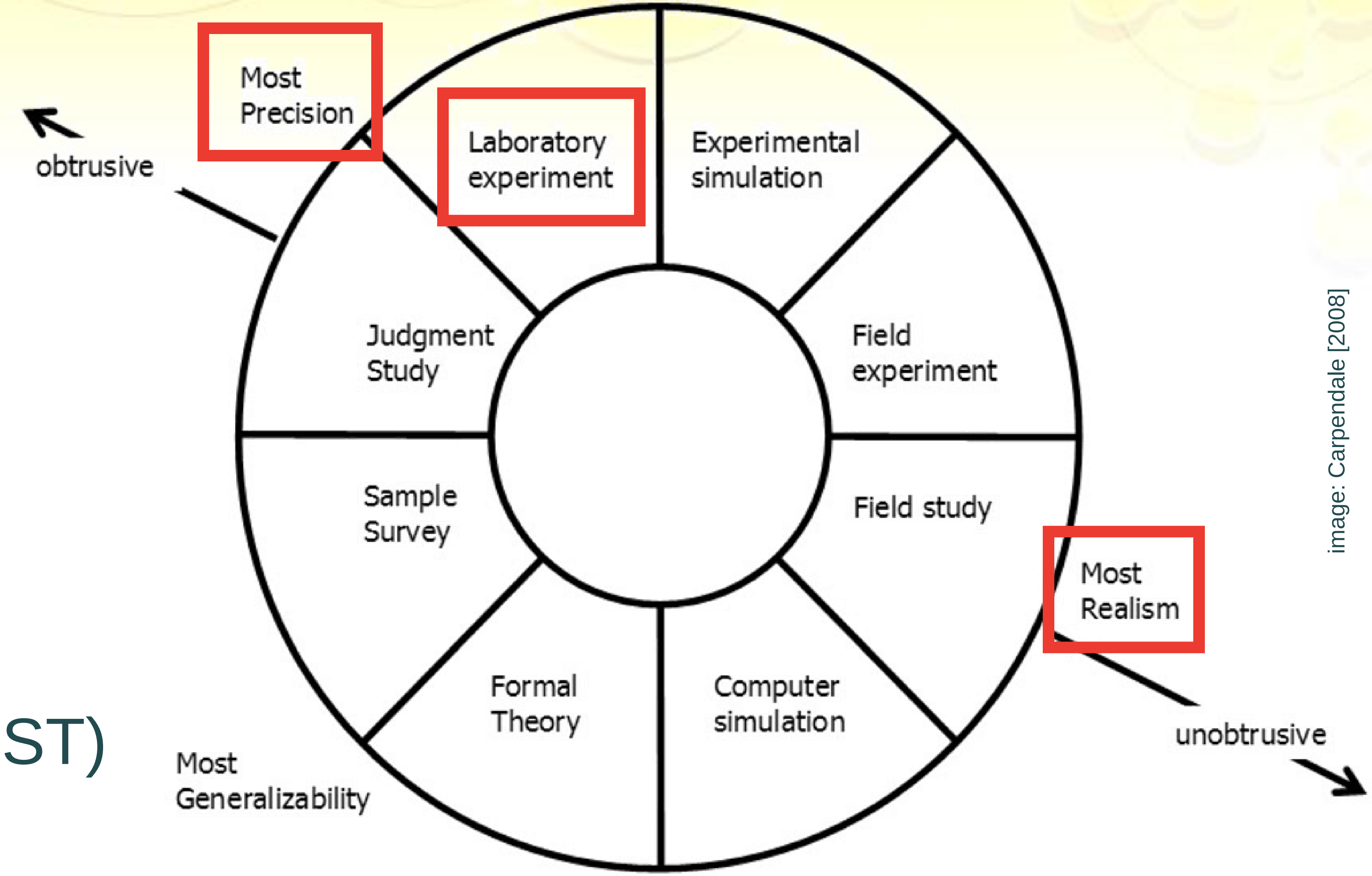


image: Carpendale [2008]

Statistical significance vs. expert feedback

- right methodology for given question!
- visualization: ill-defined, fuzzy, broad domain problems
- evaluation **not only** null hypothesis significance testing (NHST)

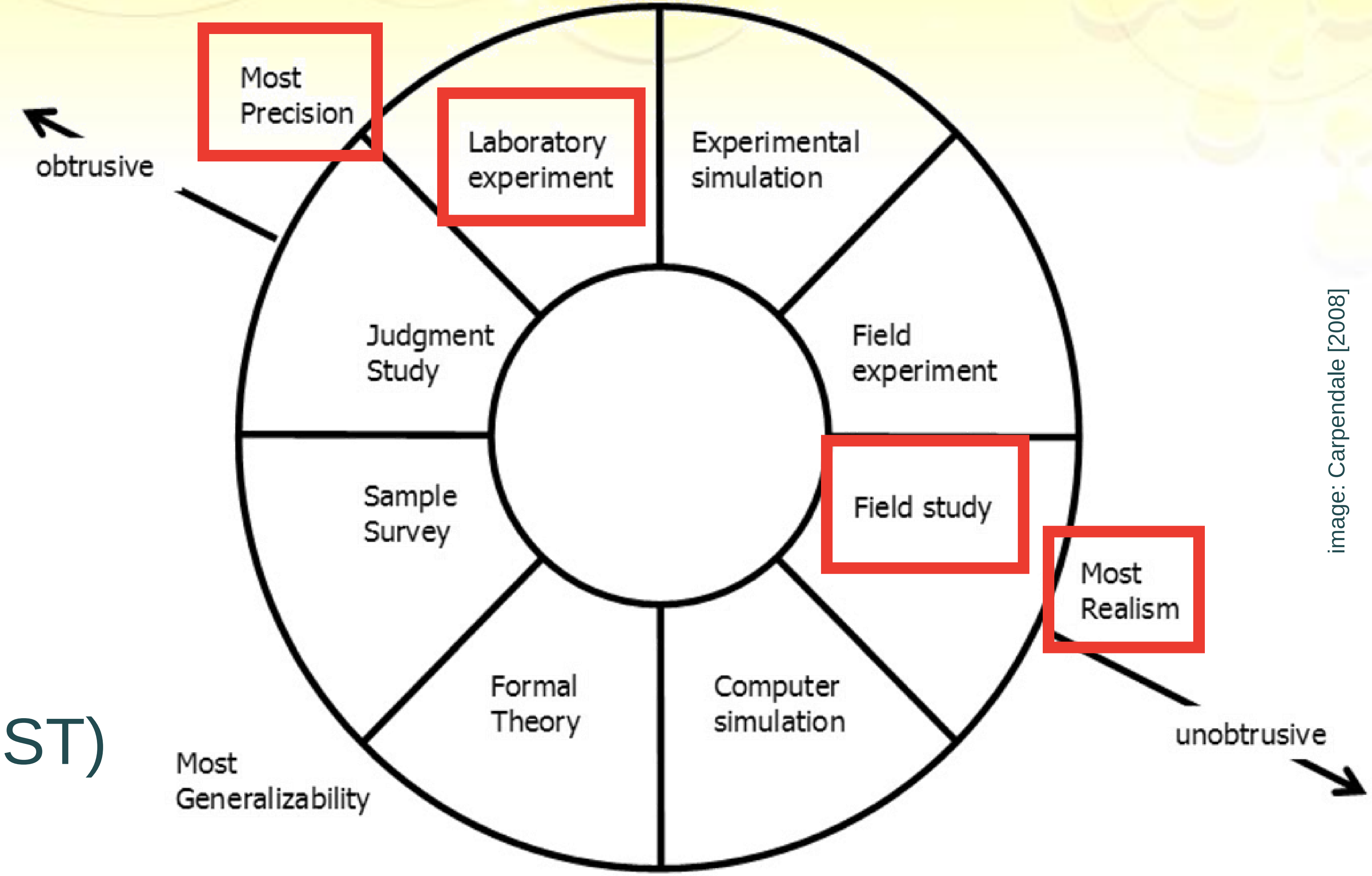


image: Carpendale [2008]

Obtaining and reporting expert feedback

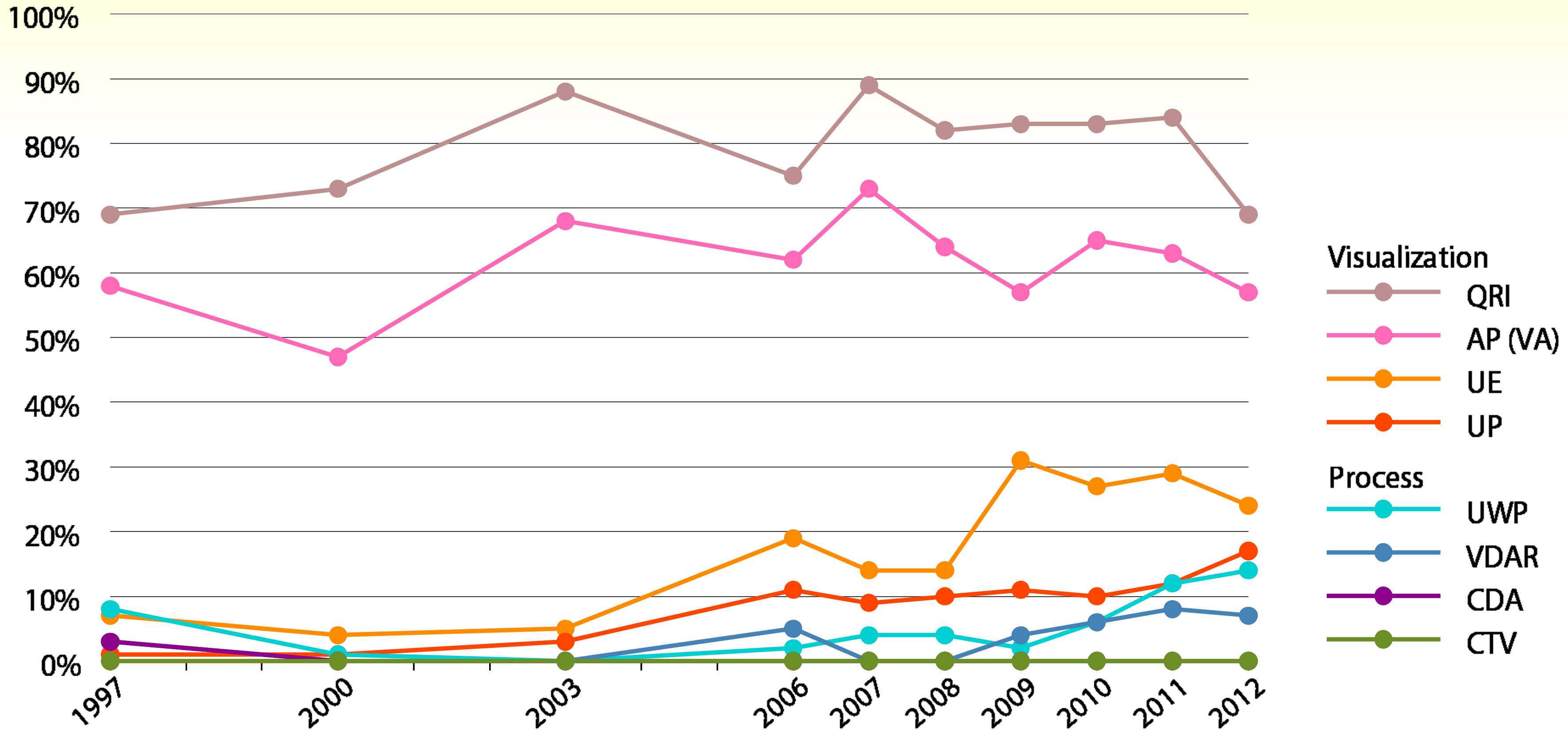
- *"We showed our system/tool to our collaborating experts **and they really liked it**."*

- expert feedback valid & important
- but: **rigor** in study design and reporting!
- several guidelines on qualitative evaluation methods

Open questions

- rigor in algorithmic performance?
  - how many datasets?
  - benchmark datasets?
- statistical analysis?
  - issues with NHST (see “dance of the p-values”)
  - how many participants?
- rigorous qualitative results inspection?

Thanks for your attention



shameless plug:

interested in working with us at Inria ?

… talk to me

paper: **http://goo.gl/6yiggh**

data: **http://goo.gl/CGswy**

Vis 2013