Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Proseminar Spam
und
Seminar Maschinelles Lernen in der IT Sicherheit

Tipps zu Ausarbeitung und Vortrag

Niels Landwehr, Tobias Scheffer
Folien von Steffen Bickel

Überblick

- Schriftliche Ausarbeitung
- Vorbereitung und Präsentation Vortrag

Überblick

- Schriftliche Ausarbeitung
- Vorbereitung und Präsentation Vortrag

Literatursuche

- Ausgangsliteratur vorgegeben
- Eigene Literaturrecherche
  - Verweise im Text der Ausgangsliteratur verfolgen
  - Evtl. neue Veröffentlichungen, die auf Ausgangsliteratur verweisen.
- Literatursuche im Internet:
  - http://www.scholar.google.com
  - http://citeseerx.ist.psu.edu

Literatursuche mit Google Scholar

Google Scholar BETA

multi-view clustering | Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar All articles - Recent articles

All Results
A Zisserman
F Schaffalitzk...
J Krumm
A Samal
A Strehl

Multi-view clustering - all 12 versions »
S Bickel, T Scheffer - Proceedings of the IEEE International Conference on Data ..., 2004 - 141.20.20.55
Page 1. **Multi-View Clustering ...** We develop and study partition- ing and agglomerative, hierarchical **multi-view clustering** algorithms for text data. **...**
Cited by 18 - Related Articles - View as HTML - Web Search - Import into BibTeX

Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?" - all 7 versions »
F Schaffalitzky, A Zisserman - Proc. ECCV, 2002 - Springer
... Multi-view Matching for Unordered Image Sets 429 Fig. 12. The largest **cluster** (8 views),for the Raglan image set,showing the interior of the hexagonal keep. **...**
Cited by 158 - Related Articles - Web Search - Import into BibTeX - BL Direct

3D object modeling and recognition using affine-invariant patches and **multi-view** spatial constraints - all 13 versions »
F Rothganger, S Lazebnik, C Schmid, J Ponce - Computer Vision and Pattern Recognition, 2003. Proceedings. ..., 2003 - ieeexplore.ieee.org
Page 1. 3D Object Modeling and Recognition Using Affine-Invariant Patches and **Multi-View** Spatial Constraints Fredrick Rothganger **...**
Cited by 56 - Related Articles - Web Search - Import into BibTeX - BL Direct

M 2 Tracker: A **Multi-View** Approach to Segmenting and Tracking People in a Cluttered Scene - all 16 versions »
A Mittal, LS Davis - International Journal of Computer Vision, 2003 - Springer
... Page 3. M 2 Tracker: A **Multi-View** Approach to Segmenting and Tracking People in a Cluttered Scene **... Clustering** these points allows us to detect and track people. **...**
Cited by 121 - Related Articles - Web Search - Import into BibTeX - BL Direct

Active+ semi-supervised learning= robust **multi-view** learning - all 10 versions »
I Muslea, S Minton, C Knoblock - Proceedings of the International Conference on Machine ..., 2002 - aicml.cs.ualberta.ca
... clustering the unlabeled data "around" the ex- amples in the original training set. 3.3 Co-EM Co-EM(Nigam & Ghani, 2000) is a semi-supervised, **multi-view ...**

Mit Gliederung Beginnen

- Gliederung erstellen vor dem Schreiben: wie soll Text strukturiert sein?
- Spart Zeit:
  - Macht logische Gliederung und „roter Faden“ Sinn?
  - Frühzeitiges Korrigieren möglich
- Wenn Gliederung fertig, nur noch Kapitel mit Text füllen

Gliederungstiefe

1 ----------

2 ----------

2.1 ----------

2.1.1 ----------

2.1.2 ----------

2.2 ----------

- Gliederungstiefe maximal drei Ebenen für Seminararbeit

Gliederung - Abstraktionsniveau

1 Verkehrsmittel
- 1.1 Auto
- 1.2 Flugzeug
- 1.3 Bahn
- 1.4 ICE
- 1.5 Nahverkehr

besser

1 Verkehrsmittel
- 1.1 Auto
- 1.2 Flugzeug
- 1.3 Bahn
  - 1.3.1 ICE
  - 1.3.2 Nahverkehr

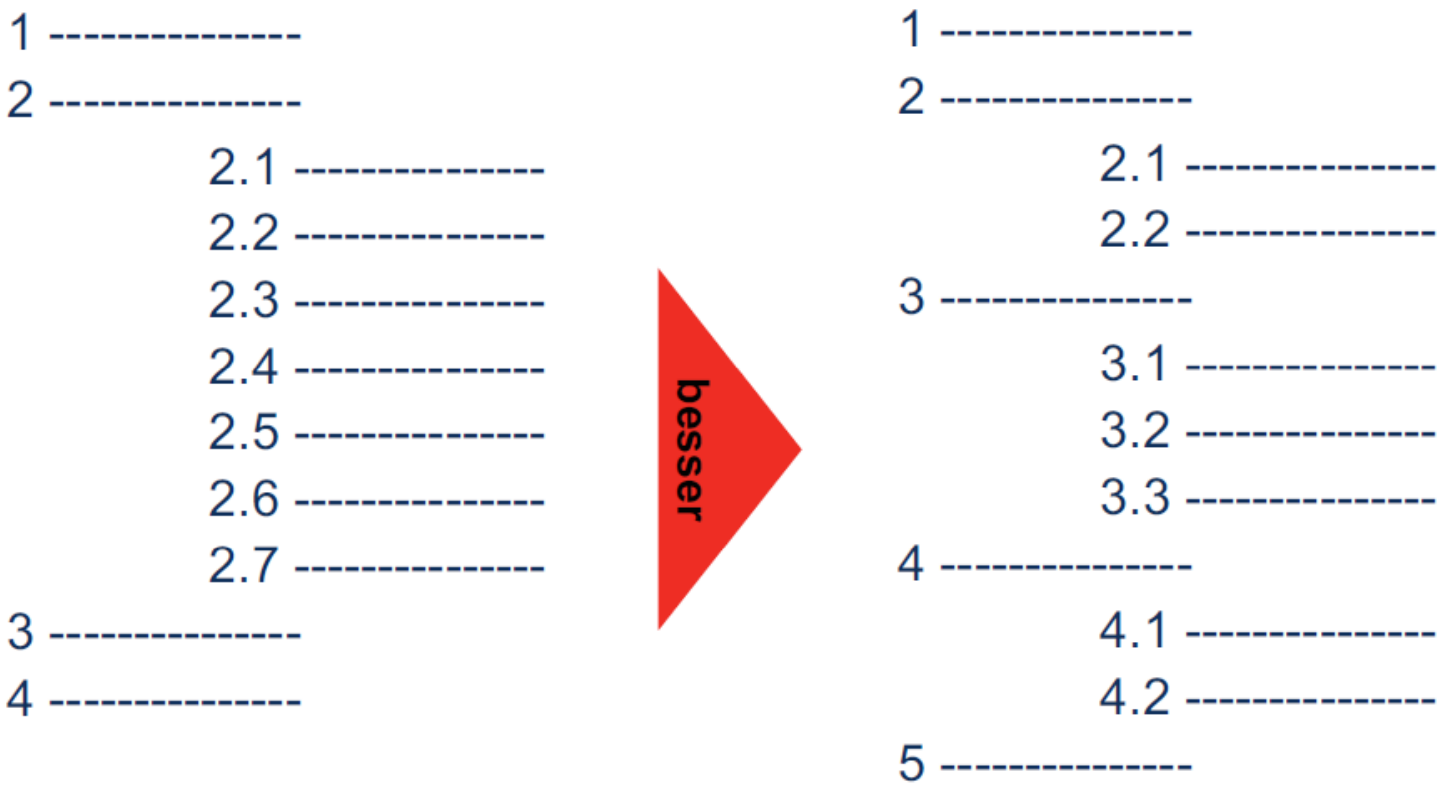
2 Nutzungsvorrausetzung
- 2.1 Schaffner
- 2.2 Pilot
- 2.3 Straßennetz
- 2.4 Flughafen
- 2.5 Schienennetz

besser

2 Nutzungsvorrausetzung
- 2.1 Personal
  - 2.1.1 Schaffner
  - 2.1.2 Pilot
- 2.2 Infrastruktur
  - 2.2.1 Straßennetz
  - 2.2.2 Flughafen
  - 2.2.3 Schienennetz

- Abschnitte auf derselben Gliederungsebene: gleiches Abstraktionsniveau

Gliederung: Strukturierung



- Auf jeder Ebene nie mehr als 6 Abschnitte hintereinander, besser 2-4 Abschnitte

Gliederung – keine einzelnen Kapitel

1 ----------
2 ----------
Text A
2.1 ----------
Text B
3 ----------

besser

1 ----------
2 ----------
Text A
Text B
3 ----------

besser

1 ----------
2 ----------
2.1 ----------
Text A
2.2 ----------
Text B
3 ----------

Ausarbeitung – Allgemeine Gliederung

- Titelblatt
- Inhaltsverzeichnis
- Evlt. Abbildungs-/Tabellen-/ Symbolverzeichnis
- Ausarbeitungstext
  - Einleitung
  - Hauptteile
  - Zusammenfassung
- Literaturverzeichnis
- Evtl. Anhang

} Ins Inhaltsverzeichnis

Ausarbeitungstext – Typische Struktur

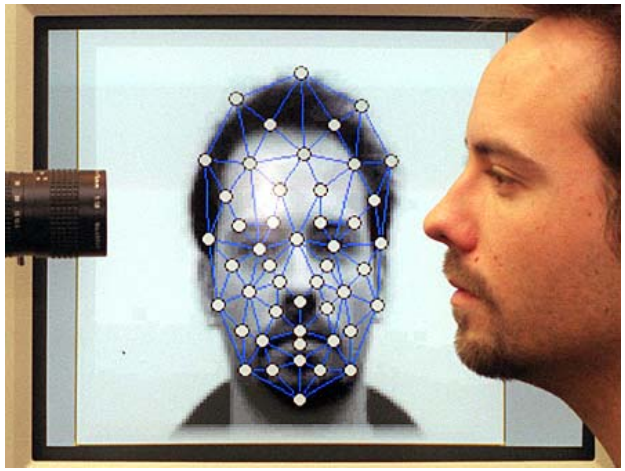
- Einleitung
- Problemstellung
- Lösung
- Evaluierung
- Zusammenfassung

Struktur hängt vom Inhalt ab…

Ausarbeitungtext – Typische Struktur

- Einleitung
  - Worum geht es? Grobe thematische Orientierung
  - Warum ist das eine interessante Anwendung/Fragestellung/…?
  - Interesse wecken

Gesichtserkennung aus Kamerabildern ist eine interessante Problemstellung, weil…



Ausarbeitungstext – Typische Struktur

- Problemstellung
  - Was genau ist das Ziel?
  - Welche algorithmische Problemstellung ergibt sich?
  - Was ist gegeben? Was gesucht?

Eingabe: Trainingsdaten
(Bilder von Gesichtern)

$$L = \langle (\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N) \rangle$$

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \ldots \\ x_{im} \end{pmatrix}$$

Ausgabe: Modell

$$p(y \mid \mathbf{x}) = \frac{1}{1 + \exp(\mathbf{w}^T \phi(\mathbf{x}))}$$

Zielkriterium: Fehlerminimierung

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \int \sum_{y \in \{1,..,k\}} \delta(y, f_{\mathbf{w}}(\mathbf{x})) \, p(\mathbf{x}, y) d\mathbf{x}$$

Ausarbeitungstext – Typische Struktur

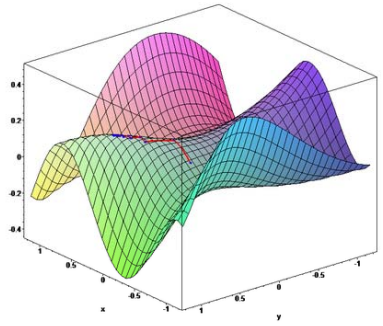
- Lösung der Problemstellung
  - Verfahren
  - Algorithmus
  - mathematische Herleitung

Algorithmus: Iterative Newton-Raphson-Optimierung zur Berechnung des optimalen Modells…

$$\mathbf{w}^{(\mathrm{new})} = \mathbf{w}^{(\mathrm{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w}).$$

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(\mathbf{w}^{\mathrm{T}}\phi_n - t_n)\phi_n = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$$

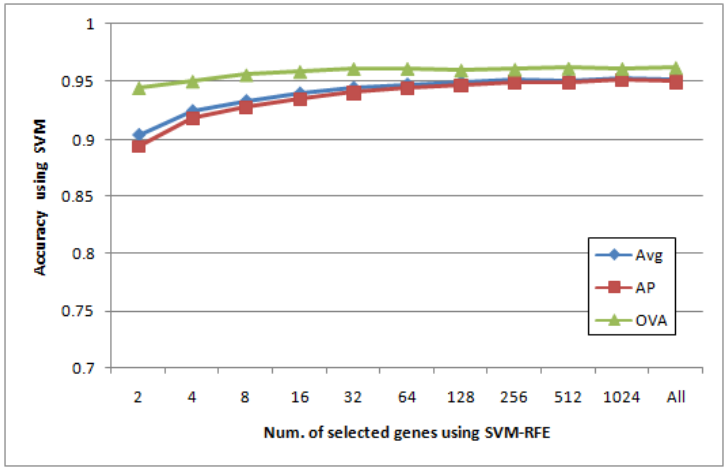
$$\mathbf{H} = \nabla\nabla E(\mathbf{w}) = \sum_{n=1}^{N}\phi_n\phi_n^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$$



Ausarbeitungstext – Typische Struktur

- Evaluierung/Experimente
  - Welche Hypothese soll getestet werden (z.B. Alg. A > Alg. B in Bezug auf…)
  - Wie genau ist die Methode (Daten/Messungen)
  - Ergebnisse

Der Algorithmus A erreicht eine höhere Erkennungsrate als andere Verfahren…



Ausarbeitungstext - Typische Fehler

- Keine saubere Unterscheidung zwischen
  - Problemstellung: was wollen wir?
  - Lösung: wie erreichen wir das?

- Kein guter Übergang von der (praktischen) Motivation zur (technischen) Problemstellung

- Zu viele technische Details ohne angemessene Erklärung
  - Z.B. Bezeichner nicht eingeführt (was sind $\phi, \beta, \mathbf{x}$?)
  - Umfangreiche Algorithmen ohne Erklärung

Ausarbeitung – Tabellen, Diagramme, Formeln

- Tabellen, Diagramme, Formeln immer mit fortlaufender Nummerierung
  - z.B. *Abbildung 5: Fehlerrate des Textfilter-Algorithmus.*
  - Formeln mit rechts- oder linksbündiger Nummerierung z.B.

$$Entropy(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (2)$$

- Im fortlaufenden Text auf Abbildungen oder Formeln mit Nummer verweisen

Ausarbeitung – Schreibstil

- Wissenschaftlicher Schreibstil
  - Keine Umgangsprache
  - Keine rhetorischen Fragen
  - Keine subjektiven Bewertungen

- Wie lernt man den Stil wissenschaftlicher Arbeiten?
  - Andere wissenschaftliche Arbeiten lesen und den Stil übernehmen

Ausarbeitung – Englische Fachbegriffe

- Ausarbeitung typischerweise auf Deutsch, Literatur (fast) immer auf Englisch
- Feststehende Fachbegriffe nicht übersetzen (z.B. Data Mining)
- Falls vorhanden, lieber deutsches Äquivalent verwenden (z.B. Überanpassung statt Overfitting)
- Falls kein deutsches Äquivalent bekannt, keine gewagten Wortschöpfungen (nicht „Stützvektormaschine“ für „Support Vector Machine“)

Ausarbeitung – Quellen und Zitate

- Jede übernommene Idee und jeder Textauszug müssen als Zitate erkennbar sein
- Zitierweise: nummeriert oder (Nachname, Jahreszahl)
  - *„... mit QP-Methode gelöst werden [13].“*
  - *„... mit QP-Methode gelöst werden (Mitchell, 2001).“*
  - *„Mitchell (2001) löst dies mit der QP-Methode.“*
- Mehrere Autoren:
  - 2 Autoren ausschreiben: Mitchell und Becker (2001)…
  - >2 Autoren: nur ersten + et al.: Mitchell et al. (2001)…
- Referenzen auf Artikel in
  - Büchern, Fachjournalen, Konferenzen
  - NICHT Internetseiten, Wikipedia

Ausarbeitung - Literaturverzeichnis

- Das Literaturverzeichnis sortieren nach:
  - Autornamen oder
  - Auftretensreihenfolge im Text.
- Mehrere Veröffentlichungen eines Autors in einem Jahr: Brown 2002a, Brown 2002b, ...
- In Latex mit Bibtex automatisch

Ausarbeitung – Technisches

- Erstellung typischerweise mit Latex oder Word/Pages/…
- Latex hat einige Vorteile
  - Plattformunabhängig
  - Wissenschaftlicher Standard
  - Professionelles Schriftbild, besonders für mathematische Ausdrücke
  - Gutes Quellenmanagement mit Bibtex
- 10-15 Seiten (Inhalt zählt, nicht Anzahl Seiten)
- Deutsch oder Englisch
- Abgabe per Email als PDF Datei oder MS-Word

Überblick

- Schriftliche Ausarbeitung
- Vorbereitung und Präsentation Vortrag

Vortrag – Allgemeine Hinweise

- Vortragsdauer 20 Minuten
- Zeitlimit einhalten, nicht zu kurz, nicht zu lang
- 1 Folie ~ 1-1.5 Minuten, ergibt ca 15-20 Folien
- Vortrag vorher üben, Zeit messen
- **Wichtig**: Zielgruppe sind Mitstudenten, nicht Betreuer

Aufbau – Grundlegendes

- Einleitung
- Hauptteil – eigentlicher Inhalt
- Zusammenfassung

Aufbau – Einleitung

- Titelfolie auflegen, Begrüßung, Vorstellung
- Definition & Motivation des Problems: Interesse wecken!
- Evtl. Beziehung zu vorhergehenden Vorträgen herstellen.
- Agenda präsentieren.

Aufbau - Hauptteil

- Gliederung der Folien: logische Argumentationskette (roter Faden)
- Nicht zu viele Details
  - Nur die Kernpunkte der Argumentation vortragen, unwichtigeres weglassen
  - Aufnahmefähigkeit und Vorwissen der Zuhörer nicht überschätzen

Aufbau - Zusammenfassung

- Inhalt kurz + prägnant zusammenfassen: „Take Home Message“ von ca 3 Punkten
- Evlt. ungelöste Probleme nennen
- Ausgangspunkt für die Diskussion

Folienstil

- Übersichtlichkeit
  - Keine langen Sätze/ganze Textabschnitte
  - Nicht zu viele Aufzählungspunkte
  - Keine umfangreichen Quellcode-Auszüge
  - Nicht zu komplexe Tabellen/Algorithmen/…
  - Keine schwer nachzuvollziehenden Beweise

- Bilder veranschaulichen Sachverhalte und lockern Folien auf

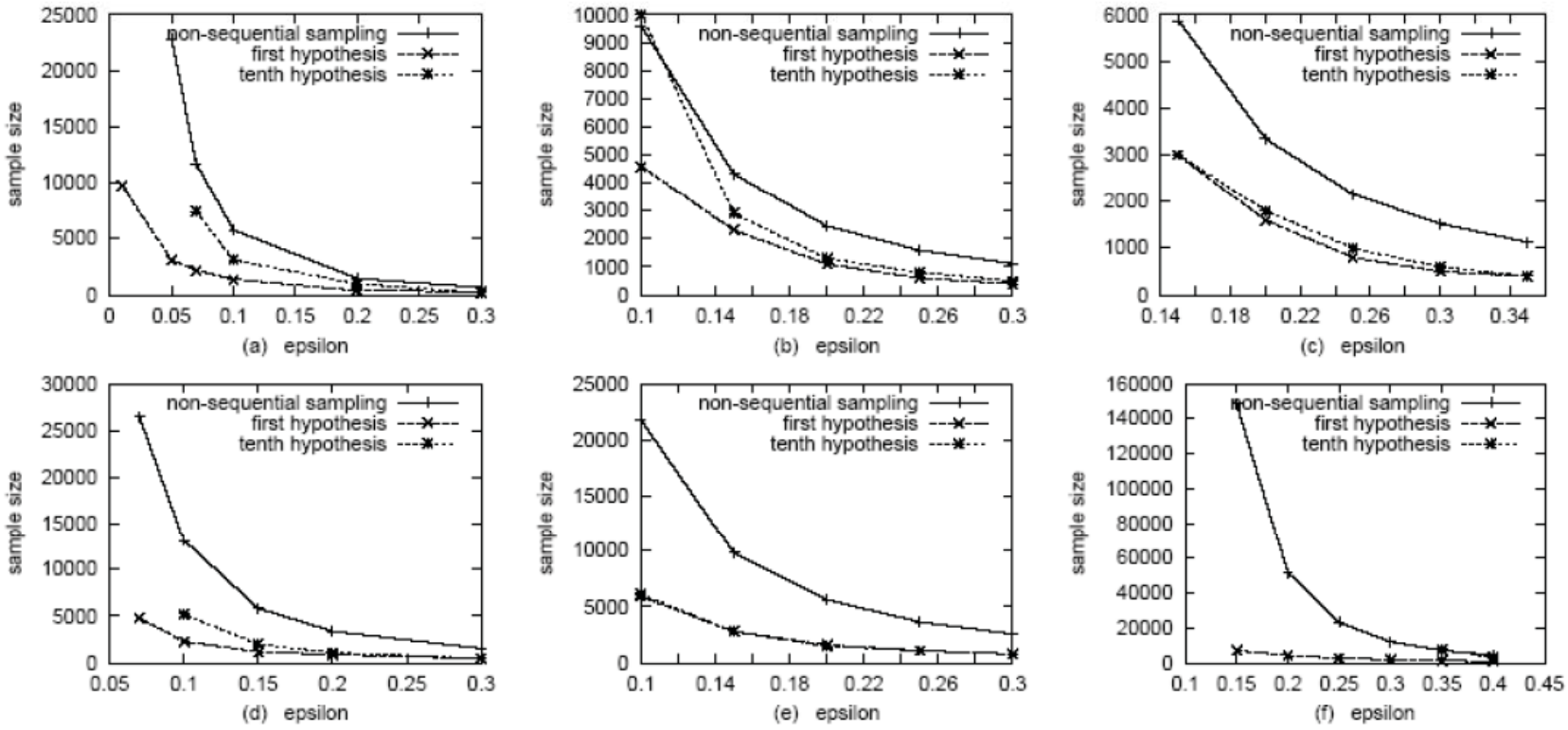
Nicht zu kleine Schriftgröße

- Das Auto frisst die Katze (28 pt)
- Das Auto frisst die Katze (24 pt)
- Das Auto frisst die Katze (20 pt)
- Das Auto frisst die Katze (16 pt)
- Das Auto frisst die Katze (12 pt)
- Das Auto frisst die Katze (10 pt)

Falsch – zu Unübersichtlich

Year	Winning Candidate	Gallup Final Survey	Election Result	Deviation
1936	Roosevelt	55.7%	62.5%	6.8%
1940	Roosevelt	52.0%	55.0%	3.0%
1944	Roosevelt	51.5%	53.3%	1.8%
1948	Truman	44.5%	49.9%	5.4%
1952	Eisenhower	51.0%	55.4%	4.4%
1956	Eisenhower	59.5%	57.8%	1.7%
1960	Kennedy	51.0%	50.1%	0.9%
1964	Johnson	64.0%	61.3%	2.7%
1968	Nixon	43.0%	43.5%	0.5%
1972	Nixon	62.0%	61.8%	0.2%
1976	Carter	48.0%	50.0%	2.0%
1980	Reagan	47.0%	50.8%	3.8%
1984	Reagan	59.0%	59.1%	0.1%
1988	Bush	56.0%	53.9%	2.1%
1992	Clinton	49.0%	43.2%	5.8%
1996	Clinton	52.0%	50.1%	1.9%

Falsch – zu Unübersichtlich



Sample sizes for the juice purchases database. (a) $f = g|p - p_0|$, $k = 1$, $\delta = .1$; (b) $k = 2$; (c) $k = 3$; (d) $f = g^2|p - p_0|$, $k = 1$, $\delta = .1$; (e) $k = 2$; (f) $f = \sqrt{g}|p - p_0|$, $k = 1$, $\delta = .1$

Falsch – zu Unübersichtlich

Theorem 8 *Let*

1. $f(h) = g^2(p - p_0)$ *and* $\hat{f}(h,Q) = \hat{g}^2(\hat{p} - p_0)$ or
2. $f(h) = g^2|p - p_0|$ *and* $\hat{f}(h,Q) = \hat{g}^2|\hat{p} - p_0|$ or
3. $f(h) = g^2 \frac{1}{c}\sum_{i=1}^{c}|p_i - p_{0_i}|$ *and* $\hat{f}(h,Q) = \hat{g}^2 \frac{1}{c}\sum_{i=1}^{c}|\hat{p}_i - p_{0_i}|$ or

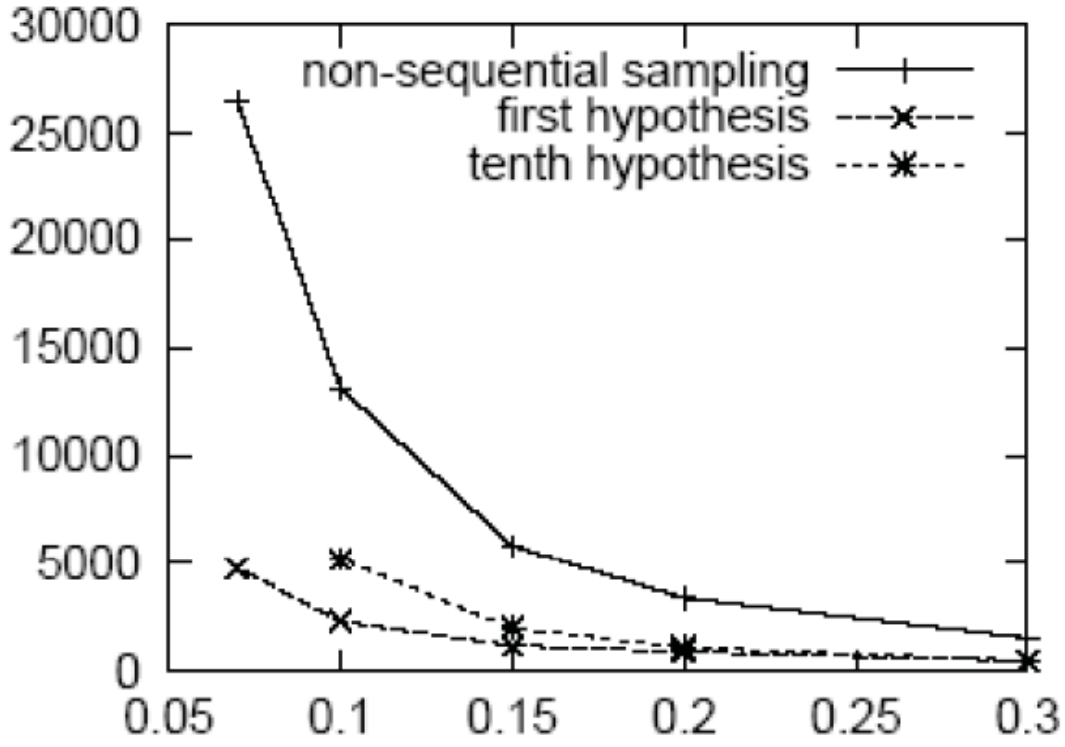
Then $Pr[\hat{f}(h,Q_m) - f(h,Q) \leq E(m,\delta)] \geq 1 - \delta$ *when*

$$\text{small } m: \quad E(m,\delta) = \left(\frac{1}{2m}\log\frac{4}{\delta}\right)^{\frac{3}{2}} + 3\left(\frac{1}{2m}\log\frac{4}{\delta}\right) + 3\sqrt{\frac{1}{2m}\log\frac{4}{\delta}} \tag{33}$$

$$\text{large } m: \quad E(m,\delta) = \frac{3}{2\sqrt{m}}z_{1-\frac{\delta}{2}} + \frac{m+\sqrt{m}}{4m\sqrt{m}}(z_{1-\frac{\delta}{2}})^2 + \frac{1}{8m\sqrt{m}}(z_{1-\frac{\delta}{2}})^3 \tag{34}$$

$$E_h(m,\delta) = 2s_g z_{1-\frac{\delta}{2}} + s_g^2(z_{1-\frac{\delta}{2}})^2 + s_p z_{1-\frac{\delta}{2}} + 2s_g s_p(z_{1-\frac{\delta}{2}})^2 + s_p s_g^2(+z_{1-\frac{\delta}{2}})^3 \tag{35}$$

Richtig: Ergebnis im Vordergrund



- Sequentielles Sampling um Faktor 2-10 besser als nicht-sequentielles Sampling…

Vortragsstil

- Möglichst frei sprechen, nicht Vortrag ablesen
- Nicht zu lange/verschachtelte Sätze
- Nicht zu schnell sprechen
- Blickkontakt zu Zuhörern, nicht (dauernd) auf Projektionsfläche schauen

Vortrag - Technisches

- Folien mit Powerpoint/Keynote/OpenOffice, evtl. Latex

- Technik:
  - Am besten Folien vorher an uns mailen (PDF oder PPT), Präsentation auf von uns gestelltem Laptop
  - Eigener Laptop auch möglich (Beamer vorhanden)

- Sicherer Fallback
  - Folien auf Memorystick
  - Handzettel mit Stichworten

Zusammenfassung

- Ausarbeitung
  - Gliederung, Struktur
  - Auf wissenschaftlichen Stil achten
- Vortrag
  - Zielgruppe sind Mitstudenten, nicht Betreuer
  - Rechtzeitig anfangen, üben, Zeitlimit
- Seminar ist Chance, unter wohlgesinnten Zuhörern Vortragstechnik zu üben und zu verbessern!
- Viel Erfolg!

Zeitplanung

- Deadline 1. Version Ausarbeitung: 4. Dezember
  - Feedback von uns
- Deadline endgültige Version Ausarbeitung: 4. Januar
- Deadline 1. Version Folien: 4. Januar
  - Feedback von uns
- Seminarvorträge 14. Januar (Spam) und 15. Januar (Maschinelles Lernen in der IT Sicherheit)