SEMINAR SPAM

OBERSEMINAR MASCHINELLES LERNEN UND IT-SICHERHEIT

Organisation, Überblick, Themen

Niels Landwehr/Tobias Scheffer

Überblick heutige Veranstaltung

1. Organisatorisches
2. Überblick über beide Seminare
3. Kurzvorstellung der Themenvorschläge

Organisation

- „Spam“: Seminar, 2 SWS (3 LP)
- „Maschinelles Lernen und IT-Sicherheit“: Oberseminar, 2 SWS (3 LP)
- Ansprechpartner:
  - Niels Landwehr, Raum 03.04.0.21, landwehr@cs.uni-potsdam.de
  - Prof. Tobias Scheffer, Raum 03.04.0.17, scheffer@cs.uni-potsdam.de
- Webseiten:
  - http://www.cs.uni-potsdam.de/ml/teaching/ss12/spam.html
  - http://www.cs.uni-potsdam.de/ml/teaching/ss12/sicherheit.html

Organisation

- Beide Seminare werden als Blockseminar durchgeführt
  - Gemeinsamer Einführungstermin 11.04.2012
  - Die Vorträge der Teilnehmer „im Block“ später im Semester (Terminabsprache nachher).
- Ablauf der Seminare
  - Verschiedene Themenstellungen mit Literaturangaben (Vorstellung heute)
  - Jeder Teilnehmer/in wählt ein Thema, dass er/sie selbstständig bearbeitet
  - Schriftliche Ausarbeitung und Seminarvortrag (20min)

Überblick heutige Veranstaltung

1. Organisatorisches
2. Überblick über beide Seminare
3. Kurzvorstellung der Themenvorschläge

Überblick über die zwei Seminare

- Seminare behandeln Problemstellungen im Bereich Spam-Filterung und IT Sicherheit
- Schwerpunkt auf Verfahren des maschinellen Lernens

- Thematische Überlappung
  - Spam-Filterung ist ein Aspekt der IT-Sicherheit
  - Fokussierung auf Spam-Filterung im Seminar „Spam“
  - Größere Bandbreite an Themen im Oberseminar „Maschinelles Lernen und IT Sicherheit“ (aber Spam-Themen auch möglich)

- Jetzt: Kurze beispielhafte Einführung in das Thema maschinelles Lernen am Beispiel der Spam-Filterung

Was ist Spam?

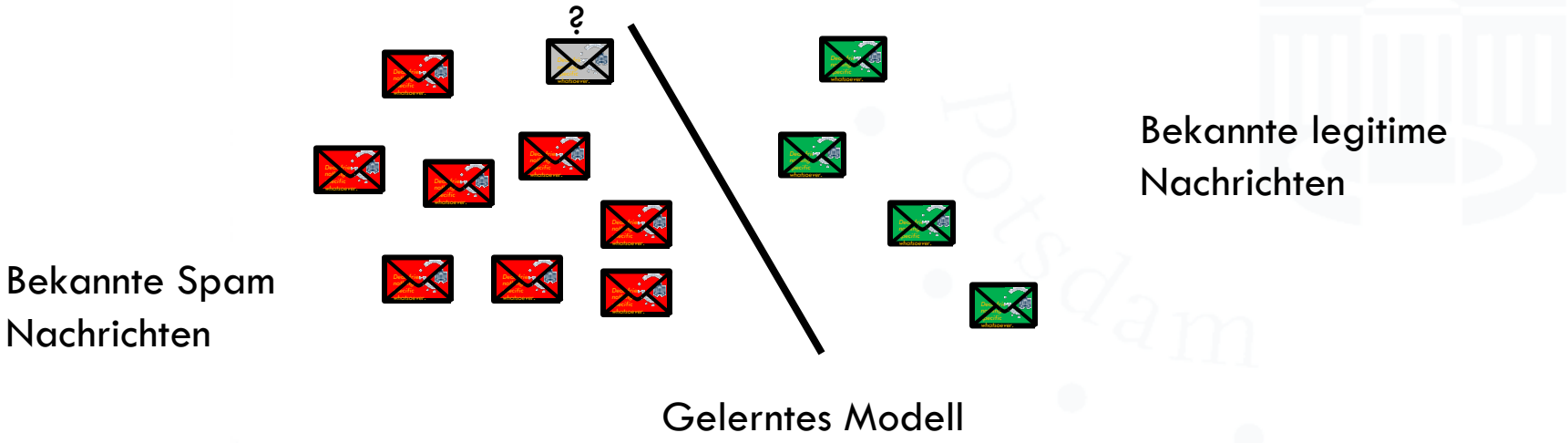
- Spam: unerwünschte (elektronische) Nachrichten, die massenhaft und unverlangt zugestellt werden
  - Werbung
  - Phishing, Viren, Betrugsversuche, …
- Spam verursacht signifikante Kosten
  - Zusätzliche Belastung der Infrastruktur
  - Sicherheitsrisiko durch Phishing/Betrug/Viren…
  - → Weltweit Milliardenschäden
- Wir müssen Spam **filtern:** Automatische Unterscheidung Spam/Legitime Nachricht

Spam: Gegenmaßnahmen

- Manuelle Erstellung eines Filters schwierig
  - Viele Variationen von Spam-Nachrichten
  - Spammer ändern Inhalte oft
- Besserer Ansatz: **Maschinelles Lernen**
  - Sammle Nachrichten deren Spam-Status bekannt ist (z.B. von Nutzern als Spam markiert)
  - System, das aus diesen Daten **lernt,** Spam zu erkennen
  - Lernen: Bilden eines (mathematischen) Modells, das die beobachteten Spams von den beobachteten legitimen Nachrichten unterscheiden kann

Spam: Gegenmaßnahmen

- Neue Nachrichten mit unbekanntem Spam-Status werden vom Modell klassifiziert und entsprechend gefiltert

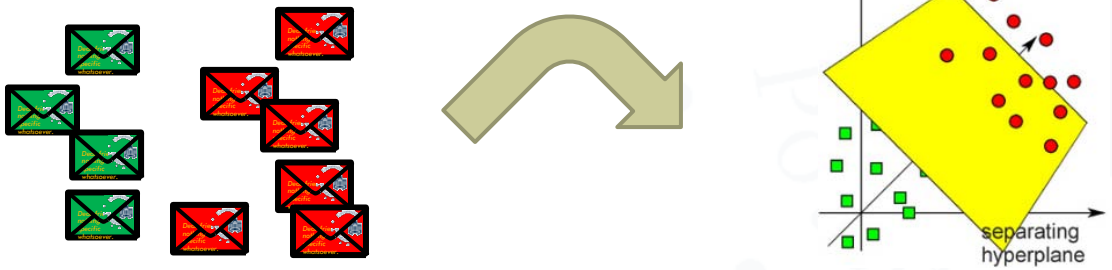


Spam und Maschinelles Lernen

- Verschiedene Techniken des maschinellen Lernens
  - Probabilistische Modelle: Wahrscheinlichkeiten, Inferenz

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right)d\theta.$$

  - Kernel Methoden: Einbettung von Daten in hochdimensionalen euklidischen Raum



  - … mehr in Vorträgen

Weitere Anwendungen des Maschinellen Lernens in der IT Sicherheit

- Weitere Anwendungen des maschinellen Lernens in der IT Sicherheit
- Modelle zur Unterscheidung von „normalen“ und „verdächtigen“ Zugriffen/Transaktionen/…
- Verschiedenste Domänen
  - Netzwerksicherheit („Intrusion Detection“)
  - Virenerkennung
  - Kreditkartenbetrug
  - Betrug in Online-Auktionen
  - …

Überblick heutige Veranstaltung

1. Organisatorisches
2. Überblick über beide Seminare
3. Kurzvorstellung der Themenvorschläge

Kurzvorstellung der Themenvorschläge

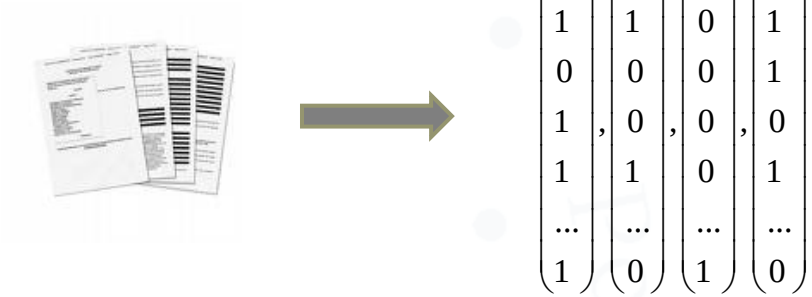
- Themenvorschläge, die Anwendungen des maschinellen Lernens behandeln (Vorkenntnisse im maschinellen Lernen empfehlenswert)
  1. Textklassifikation
  2. Email-Spam-Filterung auf Textebene
  3. Adversarial Learning
  4. Personalisierte Spam-Filter und Multitask Lernen
  5. Erkennen von bösartiger Software und Viren mit Hilfe des maschinellen Lernens
  6. Erkennung von Kreditkartenbetrug mit Hilfe von Hidden Markov Modellen

- Themenvorschläge, die andere Verfahren behandeln
  7. Spam-Filterung mit Hilfe von Blacklists
  8. Email-Spam-Filterung auf Graphebene
  9. Spam-Filter basierend auf Kompressionsmodellen
  10. Erkennung von Bot-Netzen
  11. Web-Spam und Trust-Rank
  12. Betrugserkennung in Online-Auktionen

1. Textklassifikation

T. Joachims: „Learning to classify text using support vector machines“

- Elementare Methoden des maschinellen Lernens zur Textklassifikation
  - Repräsention von Texten: Wort-Ebene, Buchstaben-Ebene, Wortfolgen (n-Gramme), …



  - Einfache Algorithmen des maschinellen Lernens (Naive Bayes,…)

2. Email-Spam-Filterung auf Textebene

Siefkes et al: „ Combining Winnow and Orthogonal Sparse Bigrams for Incremental Spam Filtering“

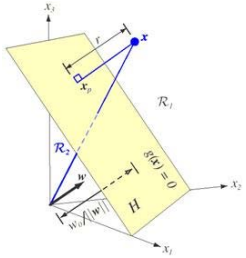
- Einfache Fallstudie zur Filterung von Email basierend auf dem (Text)Inhalt der Email
  - Repräsentation von Emails: Vorkommen bestimmter Wortpaare

		$w4$	$w5$
	$w3$	$<skip>$	$w5$
$w2$	$<skip>$	$<skip>$	$w5$
$w1$	$<skip>$	$<skip>$	$<skip>$ $w5$

Sparse Orthogonal Bi-grams

$$\begin{pmatrix} 0 \\ 1 \\ \dots \\ 0 \\ 1 \end{pmatrix}$$

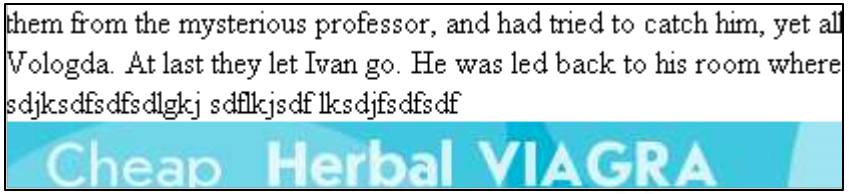
  - Einfacher Ansatz des maschinellen Lernens (Winnow Algorithmus)



3. Adversarial Learning

Lowd et al: „ Good Word Attacks on Statistical Spam Filters“

- Spammer kann als aktiver Gegenspieler des Filters gesehen werden
  - Versucht, durch Änderungen am Nachrichteninhalt den Filter zu täuschen
  - Dieser Aspekt des „Adversarial Learning“ („Lernen mit Gegenspieler“) sollte bei der Entwicklung von Spamfiltern berücksichtigt werden
- „Good Word Attacks“: Spammer fügen zu Spam-Nachrichten Worte hinzu, die mit legitimen Nachrichten assoziiert werden



- Wie können Filter robuster gegenüber Good Word Attacks werden?

4. Personalisierte Spam-Filter und Multitask Lernen

Attenberg et al: „ Collaborative Email Spam Filtering with Consistently Bad Labels using Feature Hashing “

- Personalisierte Spamfilter?
  - Jeder Benutzer erhält andere Verteilung von Emails
  - Spamfilter trainieren mit Daten eines Nutzers?→ zu wenig Daten
- „Multitask“ –Lernen:
  - Mehrere ähnliche, aber nicht gleich Lernprobleme (Filter für mehrere Benutzer)
  - Löse alle Lernprobleme gemeinsam: besseres Ergebnis als einzelne Lösungen
- Vorstellung des Multitask-Lernens
- Strategien, um Multitask-Lernen mit sehr vielen Tasks effizient zu lösen

Vorwissen im Bereich
ML empfehlenswert!

5. Erkennung von bösartiger Software mit Hilfe des maschinellen Lernens

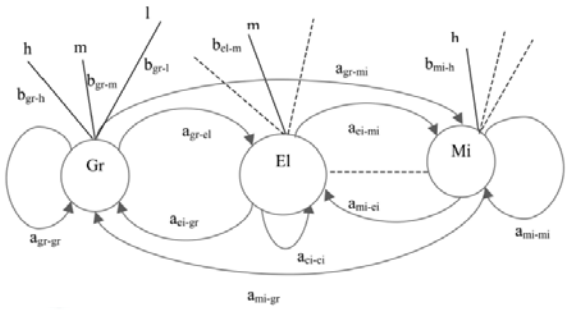
Kolter et al: „ Learning to Detect Malicious Executables in the Wild“

- Bösartige Software: Würmer, Trojaner, virusinfizierte Programme
- Erkennung mit Antivirus-Scannern basiert meist auf bekannten „Fingerabdrücken“ von z.B. Viren
  - Nachteil: Virus muss aktiv geworden sein bevor er erkannt werden kann
  - Problematisch bei polymorphen Viren
- Alternativer Ansatz: **Lerne** Modelle, die bösartige Software von gutartiger unterscheiden können (auch unbekannte)
  - Lerndaten: bekannte Beispiele von legitimen Programmen und Schadprogrammen

6. Erkennung von Kreditkartenbetrug mit Hilfe von Hidden Markov Modellen

Srivastava et al: „ Credit Card Fraud Detection Using Hidden Markov Model“

- Aufgabe: automatisches Erkennen einer verdächtige Sequenz von Abbuchungen auf einer Kreditkarte
- Sequenzielle Daten
  - Abfolge von Abbuchungen jeweils einer bestimmten Höhe
  - Hypothese: „ungewöhnliche“ Abfolge von Buchungen Hinweis auf Missbrauch
- **Lerne** probabilistisches Modell „typischer“ Sequenzen von Abbuchungen: Hidden Markov Modell



Kurzvorstellung der Themenvorschläge

- Themenvorschläge, die Anwendungen des maschinellem Lernens behandeln
  1. Textklassifikation
  2. Email-Spam-Filterung auf Textebene
  3. Adversarial Learning
  4. Personalisierte Spam-Filter und Multitask Lernen
  5. Erkennen von bösartiger Software und Viren mit Hilfe des maschinellen Lernens
  6. Erkennung von Kreditkartenbetrug mit Hilfe von Hidden Markov Modellen

- Themenvorschläge, die andere Verfahren behandeln
  7. Spam-Filterung mit Hilfe von Blacklists
  8. Email-Spam-Filterung auf Graphebene
  9. Spam-Filter basierend auf Kompressionsmodellen
  10. Erkennung von Bot-Netzen
  11. Web-Spam und Trust-Rank
  12. Betrugserkennung in Online-Auktionen

7. Spam-Filterung mit Hilfe von Blacklists

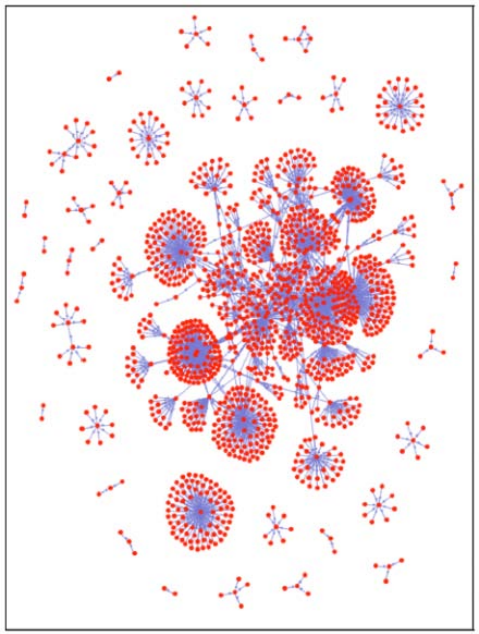
Jung et al: „ An empirical study of spam traffic and the use of DNS black lists “

- Einfacher Ansatz zur Spam-Filterung: IP-Blacklists
  - Datenbank aller IP-Adressen, die als Spam-Versender bekannt sind
  - Mails von Hosts die auf einer/mehreren Blacklists auftauchen werden geblockt
- Vorteile und Nachteile
  - Kann einen Teil des Spams mit großer Sicherheit und wenig Aufwand blocken
  - Weniger effektiv, wenn Spam von vielen verschiedenen IPs aus geschickt wird, oder eine IP sowohl Spam als auch legitime Nachrichten verschickt
- Studie über Umfang, Effektivität von Blacklists

8. Email-Spam-Filterung auf Graphebene

Golbeck et al: „Reputation Network Analysis for Email Filtering “

- Alternativer Ansatz zur Filterung von Email basierend auf Email-Nutzer-Netzwerk
  - Graph aus Nutzern und gesendeten Emails
  - Nutzer weisen anderen Nutzern Reputations-Punkte zu
  - Algorithmus, um aus allen vergebenen Reputations-Punkten und der Graphstruktur die Vertrauenswürdigkeit einer eingehenden Email zu bestimmen
- Experimentelle Evaluierung



9. Spam-Filterung mit Kompressionsmodellen

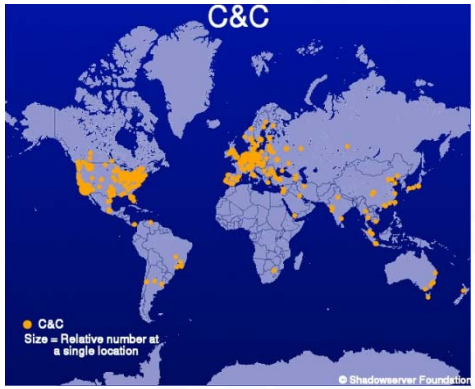
Bratko et al: „ Spam Filtering Using Statistical Data Compression Models “

- Alternativer Ansatz für Email Spam Filterung basierend auf Nachrichteninhalt
- Idee: Nachricht lässt sich gut mit anderen Nachrichten komprimieren, wenn sie ähnlich zu diesen anderen Nachrichten ist
- Verfahren zur Spamfilterung: prüfe für eine neue Email, ob sie sich besser mit bekannten Spam-Nachrichten oder bekannten legitimen Nachrichten komprimieren lässt
- Sage die Klasse mit höherer Kompression voraus

10. Erkennung von Botnetzen

Zhuang et al: „ Characterizing Botnets from Email Spam Records “

- Bots: Rechner, die (versteckt) durch einen Spamversender kontrolliert werden
  - Infizierung durch Viren, Trojaner, oder Ausnutzung von Sicherheitslücken
  - Zentrales System zur Fernsteuerung einer Menge von Bots über Kommunikationsprotokoll (Botnetz)
  - Botnetze optimal für Spamversand, DDOS Attacken, etc: anonym und verteilt



- Erkennung von Botnetzen anhand von Spam-Kampagnen

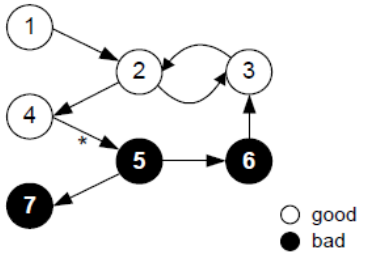
11. Web-Spam und Trust-Rank

Gyöngy et al: „ Combating Web Spam with TrustRank “

- Web-Spam: Angriffe auf Ranking-Algorithmen der Suchmaschinen
  - Ziel: erhöhe Page-Rank Score einer „Ziel“-Seite
  - Erstelle viele Webseiten mit Schlüsselwörtern und Hyperlinks statt Inhalt

- TrustRank: Verfahren zur Identifikation von Web-Spam basierend auf einer kleinen Menge von manuell untersuchten Seiten

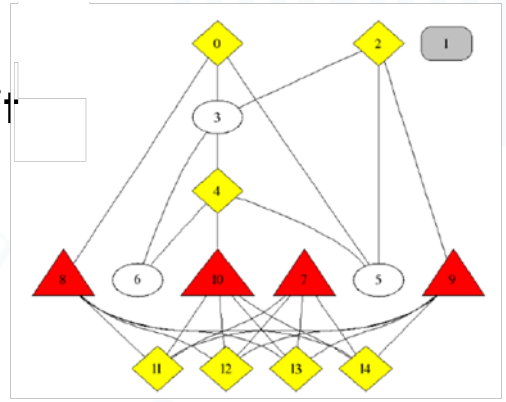
Ähnlich PageRank: basiert auf Ausbreitung von Scores entlang der Hyperlink-Struktur



12. Betrugserkennung in Online-Auktionen

Pandit et al: „Netprobe: a fast and scalable system for fraud detection in online auction networks“

- Erkennung von Betrug in Online-Auktionen (Ebay…)
- Standardmaßnahme gegen Betrug: Reputationspunkte
- Betrügerische Nutzer bilden oft Netzwerke, um gegenseitig ihre Reputation zu erhöhen
- Probabilistisches Modell über dem Graph von Nutzern und Transaktionen, um Wahrscheinlichkeit dafür zu schätzen, dass Account betrügerisch ist



Überblick heutige Veranstaltung

1. Organisatorisches
2. Überblick über beide Seminare
3. Kurzvorstellung der Themenvorschläge
4. Themenvergabe, Termine

Themenwahl: per Mail

- Mail an mich (landwehr@cs.uni-potsdam.de) mit 3 Themenvorschlägen, welches Seminar, Matrikelnummer. Deadline: in einer Woche.

- Themenvorschläge, die Anwendungen des maschinellem Lernens behandeln
  1. Textklassifikation
  2. Email-Spam-Filterung auf Textebene
  3. Adversarial Learning
  4. Personalisierte Spam-Filter und Multitask Lernen
  5. Erkennen von bösartiger Software und Viren mit Hilfe des maschinellen Lernens
  6. Erkennung von Kreditkartenbetrug mit Hilfe von Hidden Markov Modellen

- Themenvorschläge, die andere Verfahren behandeln
  7. Spam-Filterung mit Hilfe von Blacklists
  8. Email-Spam-Filterung auf Graphebene
  9. Spam-Filter basierend auf Kompressionsmodellen
  10. Erkennung von Bot-Netzen
  11. Web-Spam und Trust-Rank
  12. Betrugserkennung in Online-Auktionen

Termine für den weiteren Ablauf

- Ablauf:
  - Sie geben 1. Version Ausarbeitung ab, wir geben Kommentare
  - Sie können die Kommentare einarbeiten und geben danach die Endversion ab
  - Ebenso für die Folien des Seminarvortrags
- Deadline 1. Version Ausarbeitung: 21. Mai
- Deadline Endversion Ausarbeitung, 1. Version Folien: 15. Juni
- Seminarvorträge: 21/22 Juni? (Seminar/Oberseminar)

Mai 2012

KW	MO	DI	MI	DO	FR	SA	SO
18		1	2	3	4	5	6
19	7	8	9	10	11	12	13
20	14	15	16	17	18	19	20
21	21	22	23	24	25	26	27
22	28	29	30	31			

Juni 2012

KW	MO	DI	MI	DO	FR	SA	SO
22					1	2	3
23	4	5	6	7	8	9	10
24	11	12	13	14	15	16	17
25	18	19	20	21	22	23	24
26							

ICML

Juli 2012

KW	MO	DI	MI	DO	FR	SA	SO
26							1
27	2	3	4	5	6	7	8
28	9	10	11	12	13	14	15
29	16	17	18	19	20	21	22
30	23	24	25	26	27	28	29
31	30	31					

Einführung Wissenschaftliches Arbeiten

- Neben dem heutigen Einführungstermin gibt es noch eine Einführung „Wissenschaftliches Arbeiten“ als Videovorlesung
  - Tipps zur Anfertigung der Seminararbeit und des Seminarvortrags
  - Video auf den Webseiten der Seminare verfügbar