University of Strathclyde Science

Leif Azzopardi *(@leifos)*
Chancellor's Fellow/Associate Professor

Simulation in IR

Evaluating and Measuring Information Retrieval Systems

In collaborations with David, Colin, Vu, Norbert, Krisztian, Maarten, Kal, Heikki, Jaana, Mark and Jaap

4th Alexandria Workshop 2017 Hannover, Germany

October 19th, 2017

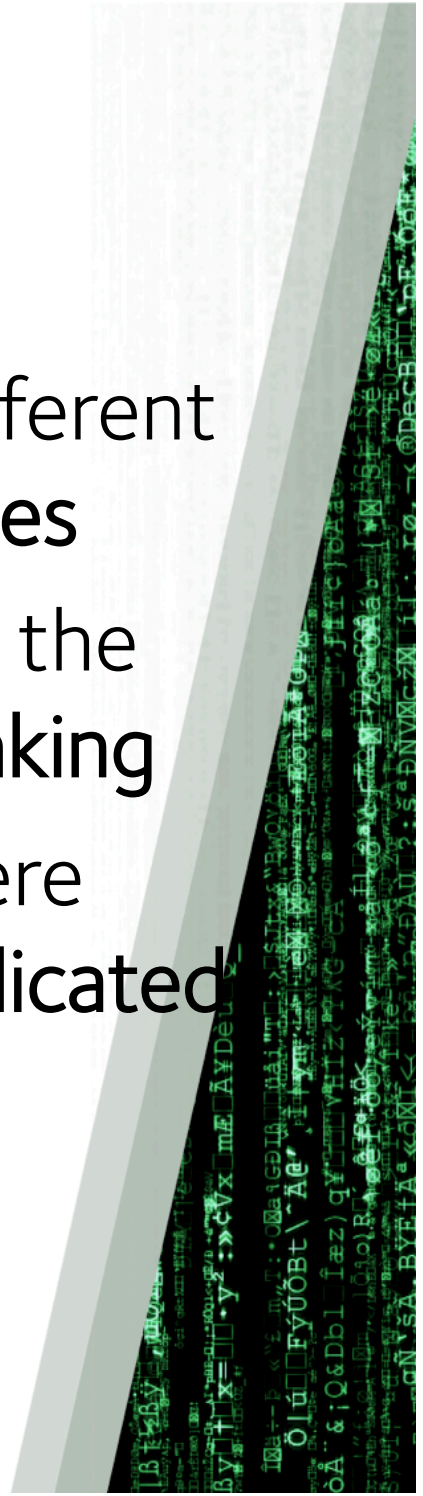
A Simulated Reality

- **Simulation** is an **imitation** of the **operation** of a **real-world process** or **system** over **time**
  - Requires a model of the user/system/process
- Simulation is used in many other areas
  - Aircraft Design, Chem. Eng., Physics, Bio, etc.
- Enables researchers to **go beyond** what is possible now,
  - and consider more alternatives, faster!
- **Simulation** has deep roots in **philosophy**
  - Are we in the matrix?

Azzopardi et al, 2011

The Power of Simulation

- Hypothesize about the outcome of different **interactions**, **user models**, and **interfaces**
- **Examine** and **explore** the evaluation of the **user** and the **interface**, not just the **ranking**
- Provide a **controlled environment** where interaction can be **reproduced** and **replicated**
  - Cheap, Fast and Configurable



Possible Types of Experiments

- The What-If Experiments
  - What if the user acted differently, what if the interface provided different features, what if the system responded differently?
- The Which/How Experiments
  - How should an application be used, which interactions/methods work best?
- The Why Experiments
  - Why would a user behave in a certain way?

Simulation Pitfalls

- Where is the “user” in the user model?
- What is a good models?
- How do we make simulations re-usable and generalizable?
- There is no perfect simulation
  - Not a replacement of user experiments
  - Wont eliminate/replace users – the ultimate judge
  - Shouldn’t be employed with out thought

Validation

- Types of Validation
  - Replicability – is the performance similar?
  - Predictive – is the output the same?
  - Structural – is the structure the same?
- Is the **simulation** any **good**?
  - Validation of simulations is important!
  - Ground with user data
  - And intuition ☺

Approaches to Simulation in I/IR

- Test Collection Based Approaches
  - Synthetic data
- Component Based Approaches
  - Query Generation, Document Examination, etc.
- Agent/Interface Based Approaches
  - User model
  - Interface model
  - Task/Context model
  - Objective Function / Constraints

SEARCH STARTS WITH A QUERY

Querying the System

- Queries lead to major variations in performance
  - Yet, we often ignore this in typical evaluations
  - Just use the title!
- People often express very short queries
- Prototypical querying strategies have been identified
  - Lots of very short queries, using pivot terms, etc.

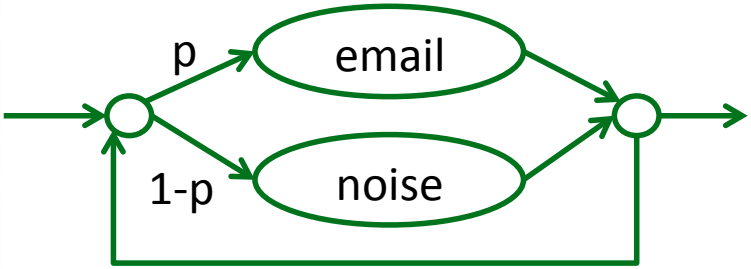
Email Search

- **Context**: TREC Enterprise Track
- **Task**: Email Known Item Search
- Example Topic:
  - Keith sent me an email about the BIM co-occurrence model last summer… I think ??
- Query:
  - Keith BIM June
- We had the collection, but few queries!

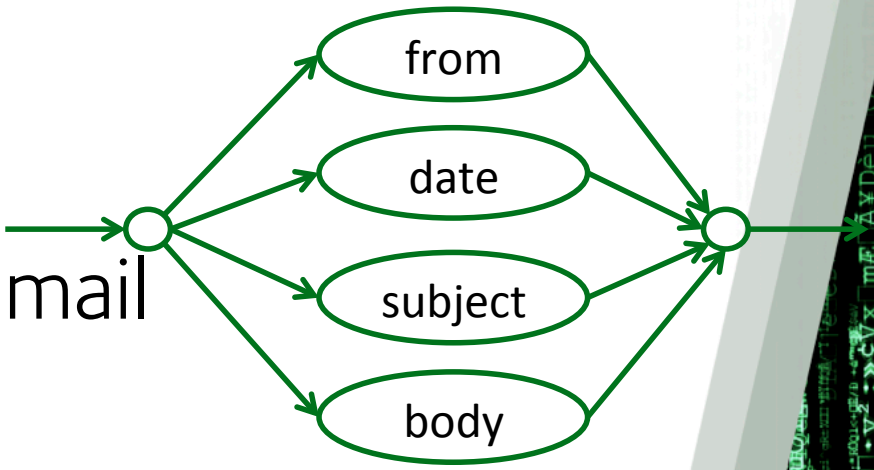
Azzopardi et al, 2005

A Generative Model for Email Known-Item Queries

- The user imagines the desired email
- Then tries to recall details from the email
  - But sometimes their memory is a bit fuzzy



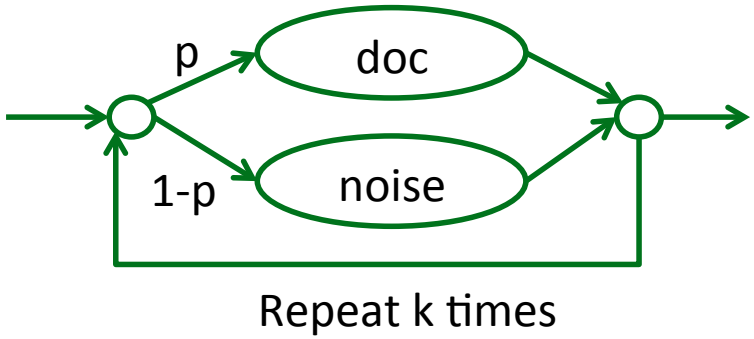
- The term is recalled from a field in the email
  - And repeated k times



Azzopardi et al, 2006, 2007

A Generative Model for Known-Item Queries

- More generally, we can generate such queries for any document type



- Where the model parameters *k*, p and the doc and noise language models – lead to different query styles/types

- And can create a test collection given a corpus
  - Generate <know-item document, query> pairs
  - CLEF 2006 Cross Lingual Web Retrieval Track
  - ClueWeb Known Item Retrieval

Balog et al, 2006

Hagen et al, 2015

What about generating queries for other tasks?

- If we have an existing test collection
  - < Topic, QRELS (d_1,..., d_r) >
- Then, we can follow a similar process and queries can be generated from:
  - Topic statements
  - Individual relevant documents
  - Sets of relevant documents
- A one million query track?
  - No Problem ☺

Azzopardi, 2009

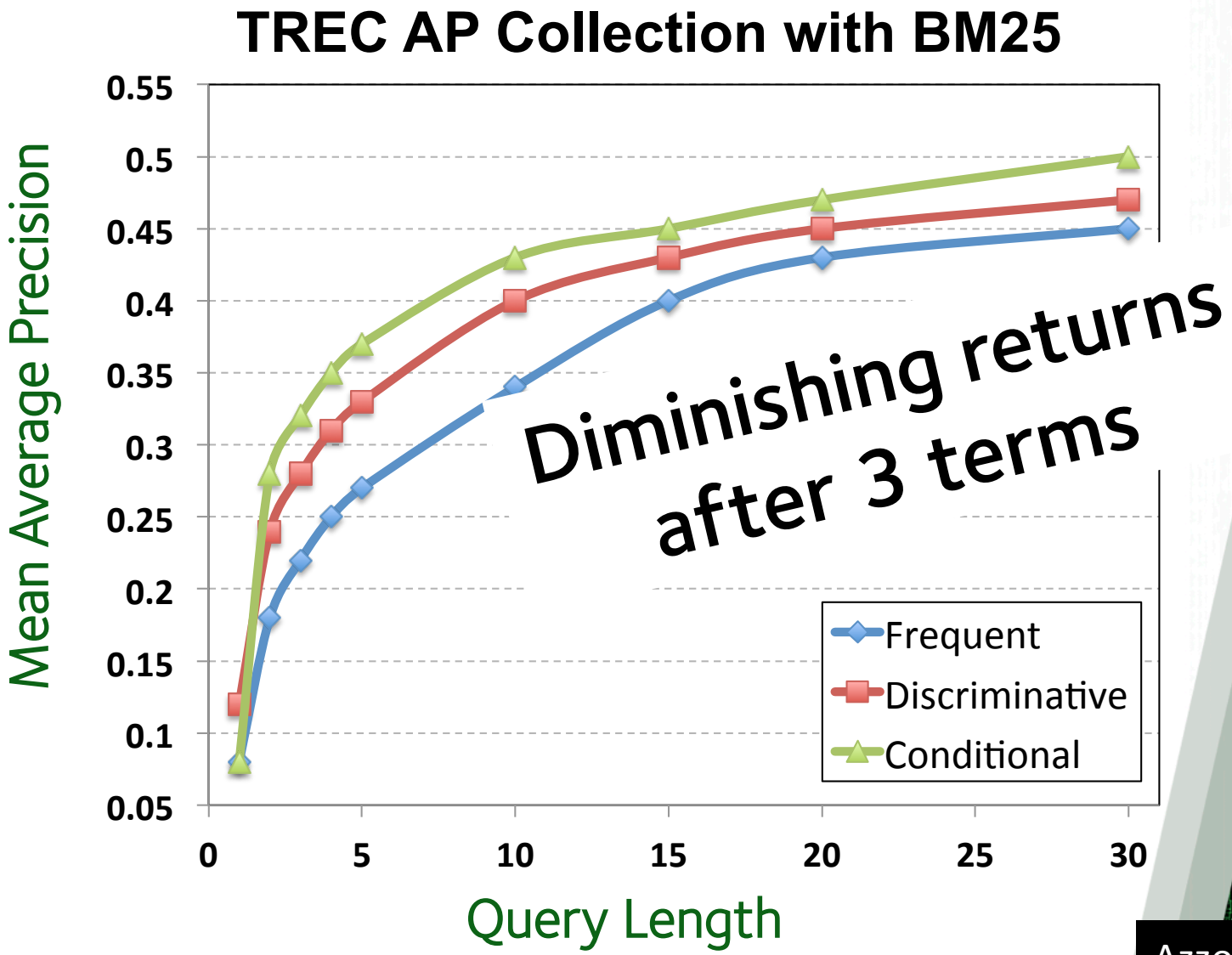
TREC Topic 51: Airbus Subsidies

Frequent	Discrimitive	Conditional
German government subsidies	source program BOE	Airbus 500 subsidy
European defense ban	Airbus airway Hill	Airbus subsidy unacceptable
agree like barrier	Haussmann BOE diplomat	Hormon Airbus subsidy
European support subsidies	Face Daimler country	Aairbus subsidy price
Aircraft Hill flight	Mcdonnel talk agreement	German Airbus subsidies
Tuesday Bohm new	Boelkow Britain dispute	Airbus official ministry

We can generate queries! So What?

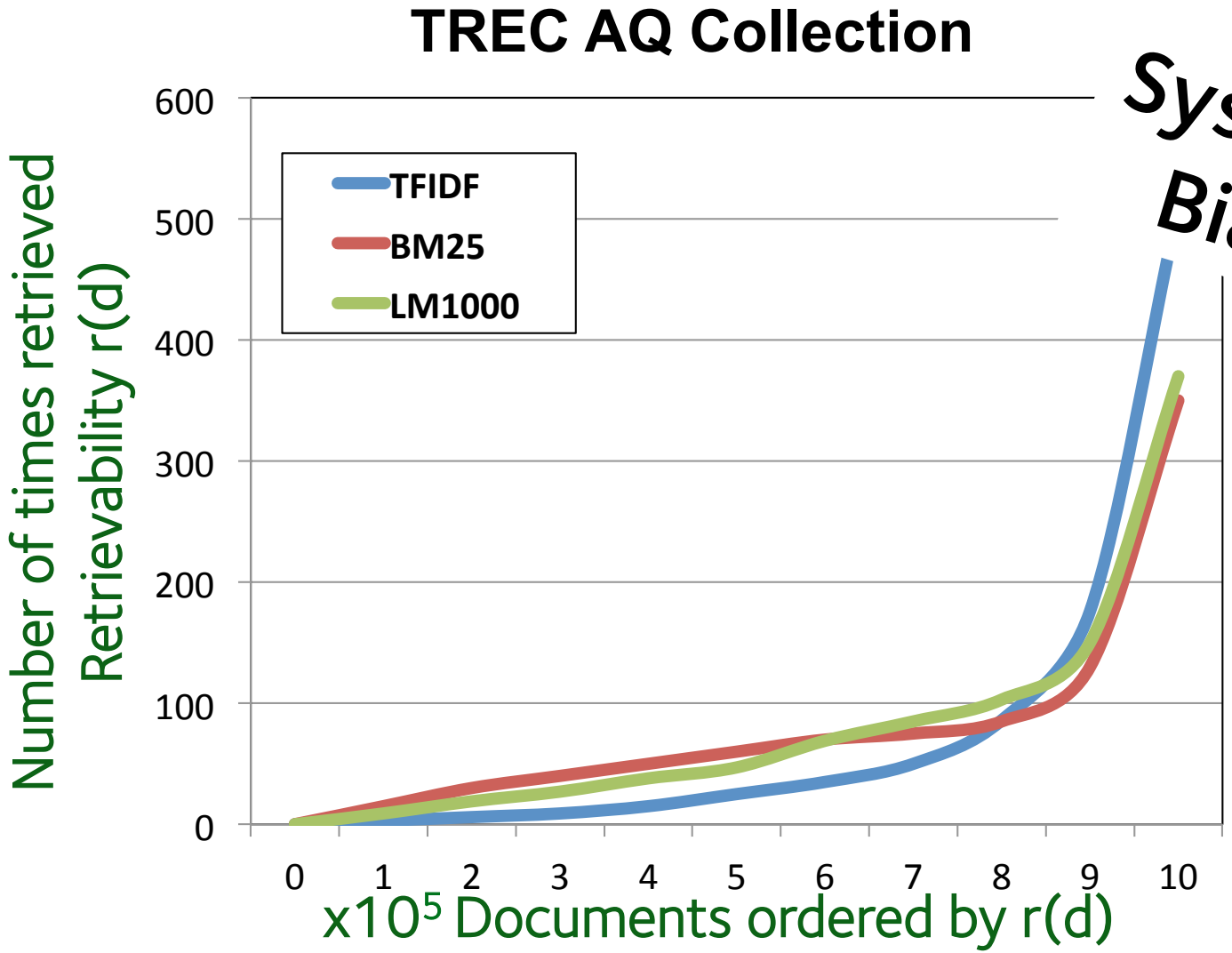
- Create an “infinite” amount of training data
- Evaluate Systems / Models / Algorithms
  - Efficiency
  - Performance
  - Bias / Retrievability
- Analyze Topics
  - Difficulty and Variance
- Examine Query Strategies
  - Length, style,
  - Quality, language

Query Length and Style



Azzopardi, 2009

Document Retrievability



System Bias

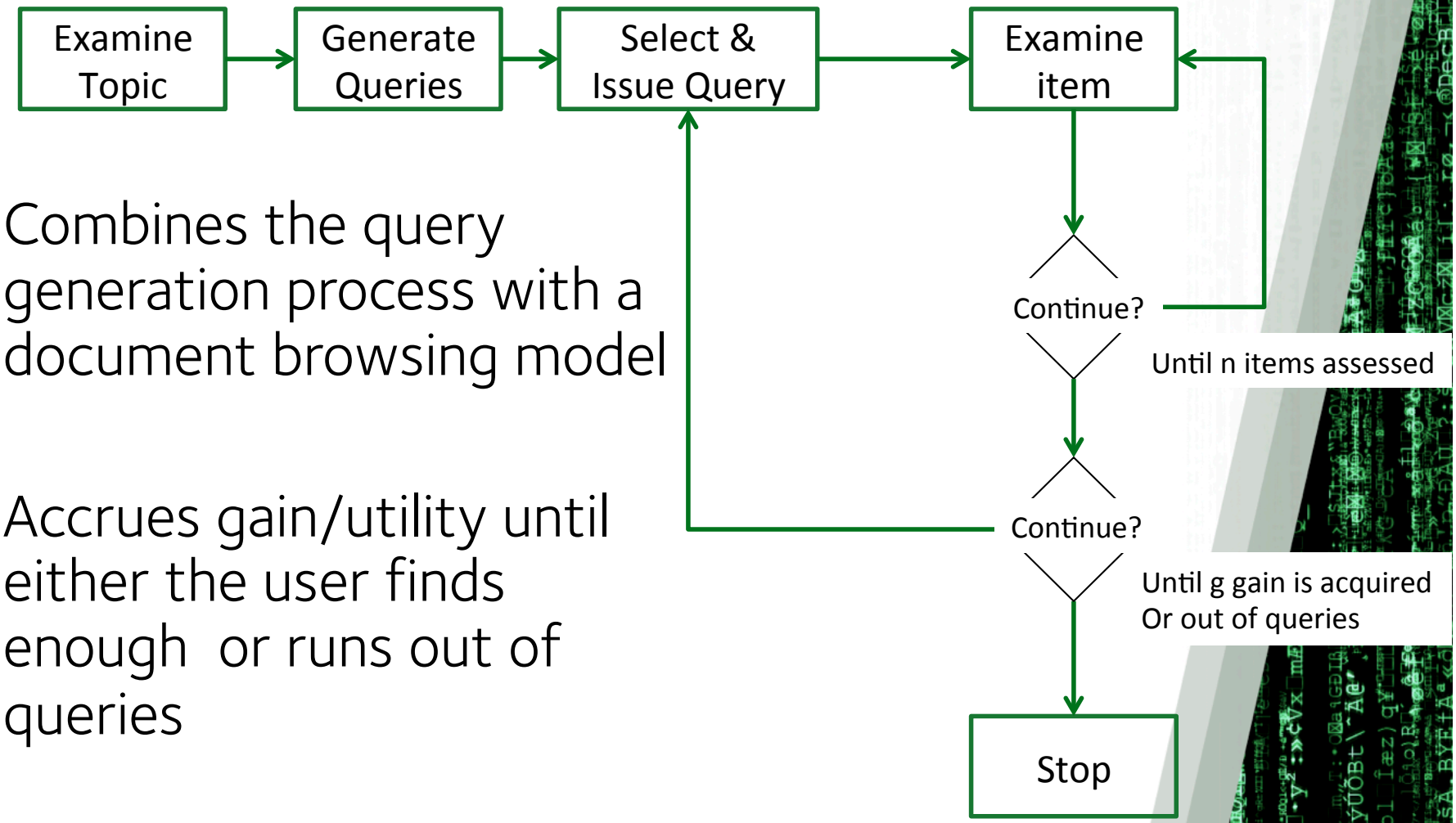
Azzopardi & Vinay 2008

Other Query Generation Contexts

- Many other contexts and situations where we simulated queries can and could be generated
  - Suggestions
  - Expansions
  - Sessions
  - Seasons
  - Time

SIMULATING THE SEARCH PROCESS

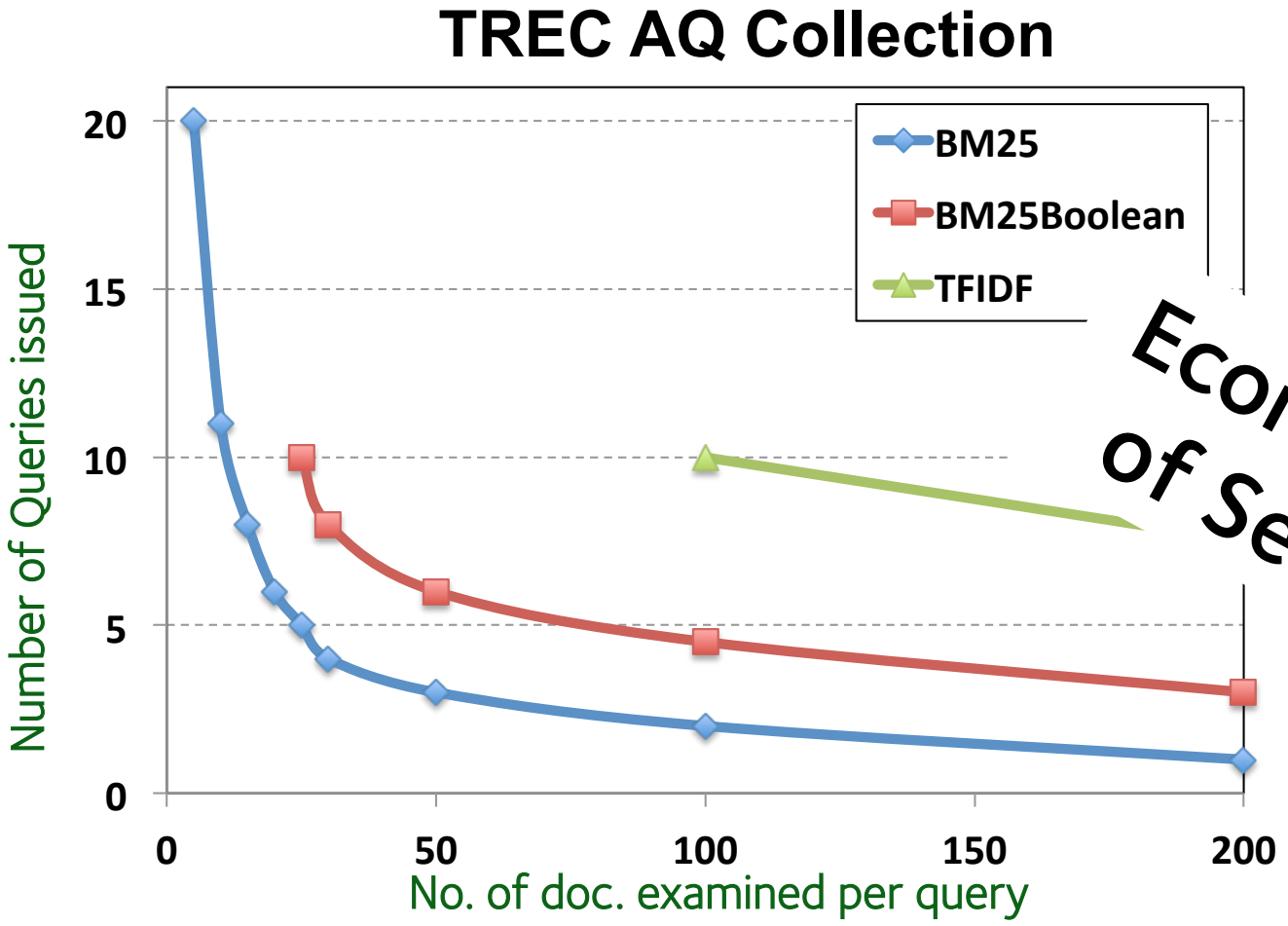
Simple Searcher Model



- Combines the query generation process with a document browsing model
- Accrues gain/utility until either the user finds enough or runs out of queries

Baskaya et al , 2012 | Carterette et al , 2011 | Azzopardi, 2009

Analysis of Search Strategies



Economics of Search

Interaction required to find 40% of the relevant document

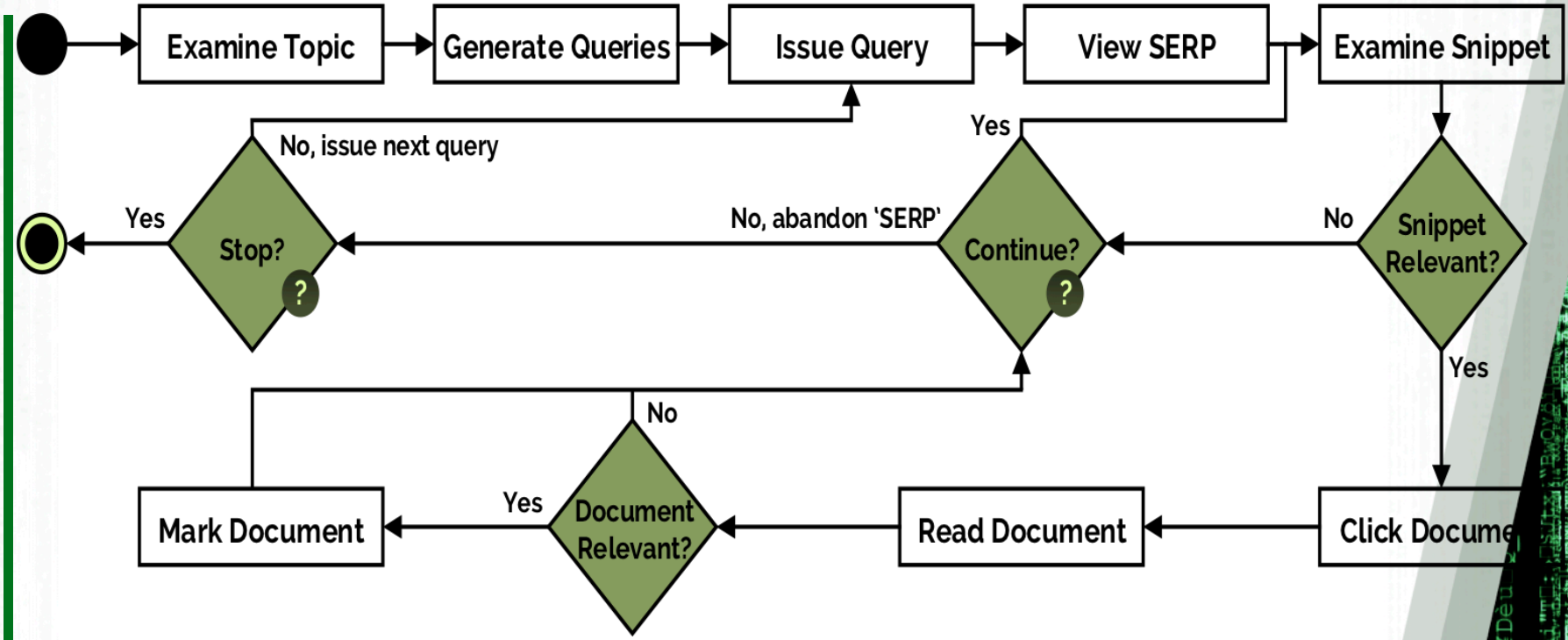
Azzopardi, 2011, 2014

Insightful, but Limited!

- Simple Searcher Model is pretty Robotic
- People go to different depths
- People actually look at snippets
- People don't assess everything
  - If they like what they see, they click it!
  - Even if it isn't relevant!
- People take time to perform actions
  - Different reading, scanning, deciding speeds

CREATING A MORE REALISTIC SEARCH PROCESS

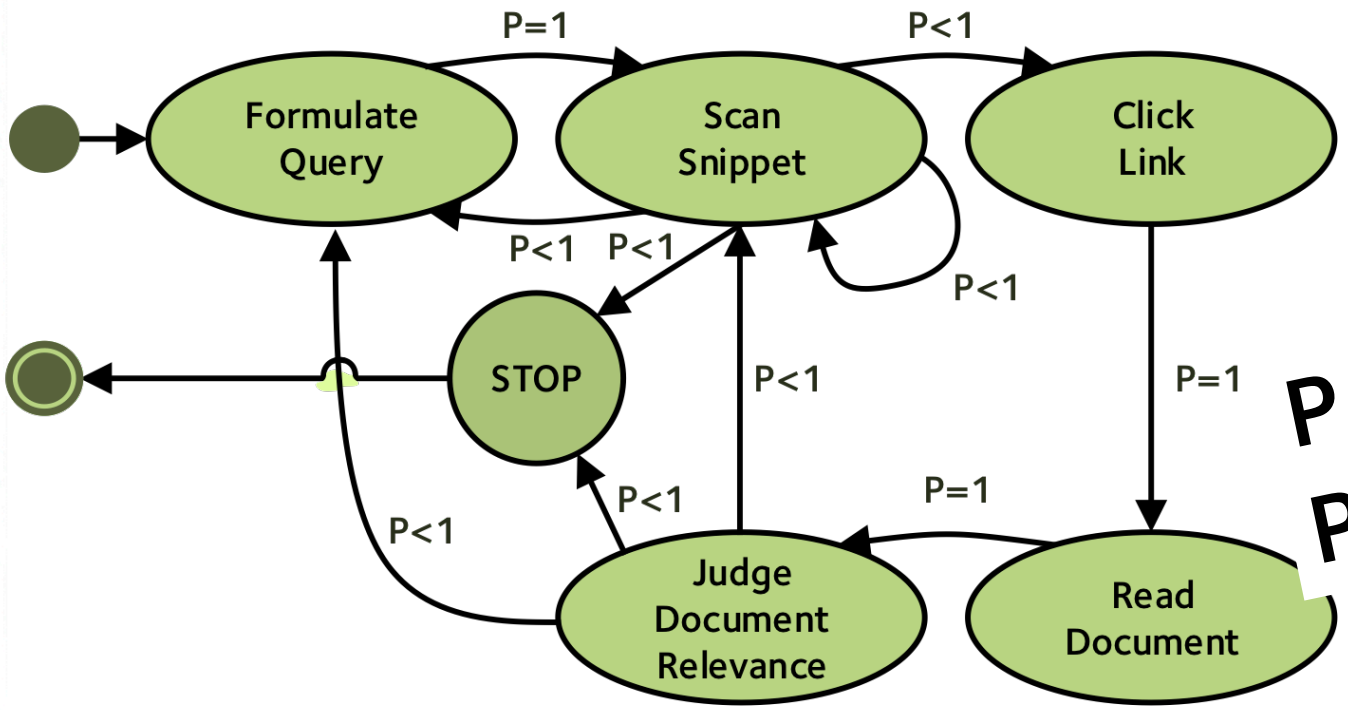
Complex Searcher Model



- Introduced more actions and decision point to provide a more detailed representation

Maxwell et al , 2015 | Thomas et al , 2014 | Baskaya et al , 2013

Modeling Decision Points



P(CIR)
P(CIN)

P(MIR)
P(MIN)

The decision to examine / judge is stochastic

Paakkanen et al, 2017

Vu et al, 2017

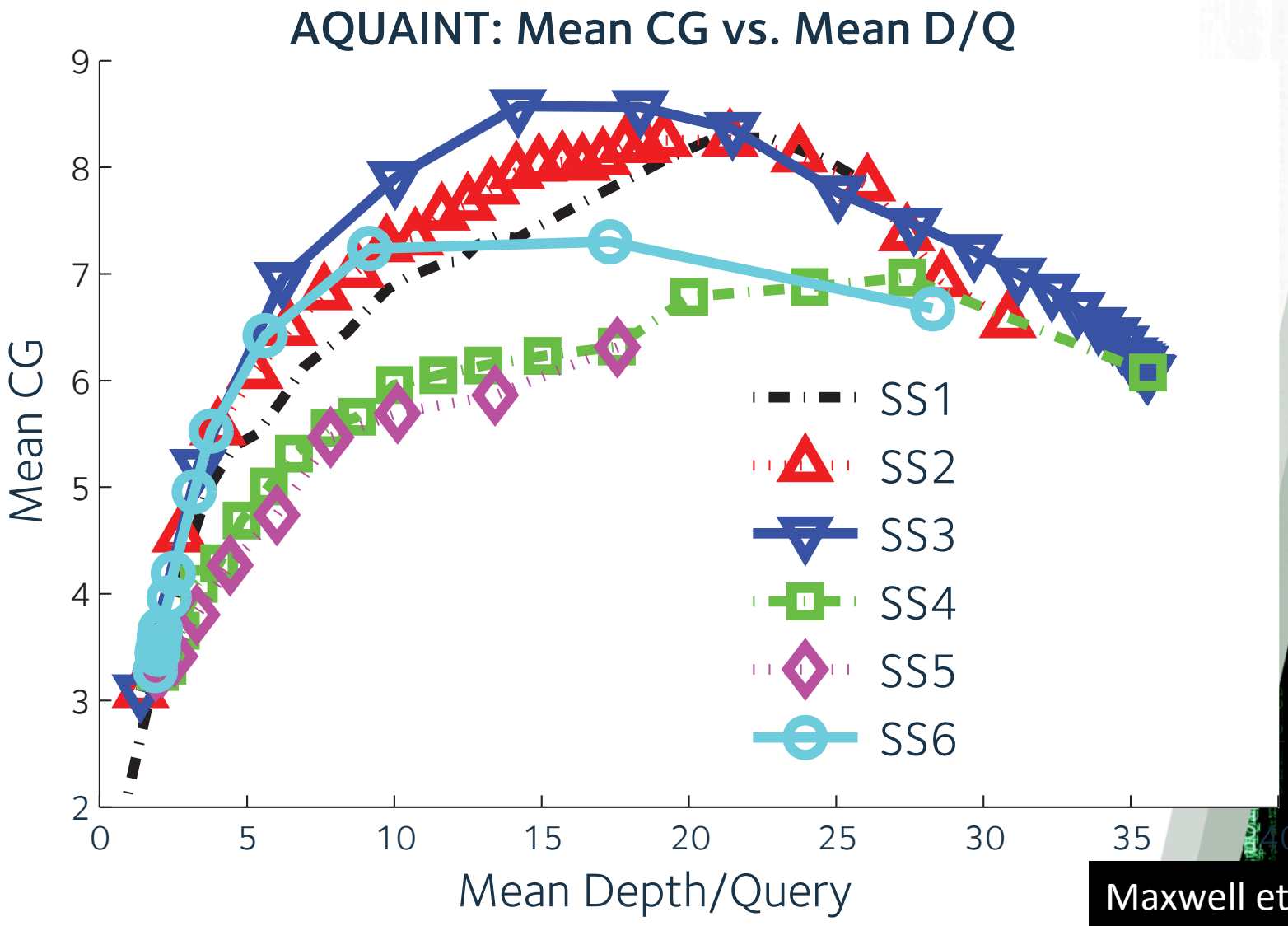
Baskaya et al , 2013

Stopping Strategies

- Is a fixed stopping strategy reasonable?
  - How reasonable is something like P@10?
- Stopping Rules:
  - Fixed-Depth
  - Frustration / Disgust
  - Difference / Novelty threshold
  - Utility / Gain
- Context: Ad Hoc Retrieval
  - Find as many relevant document in 20 minutes
  - Simulations grounded with interaction times and interaction probabilities

Maxwell et al , 2015

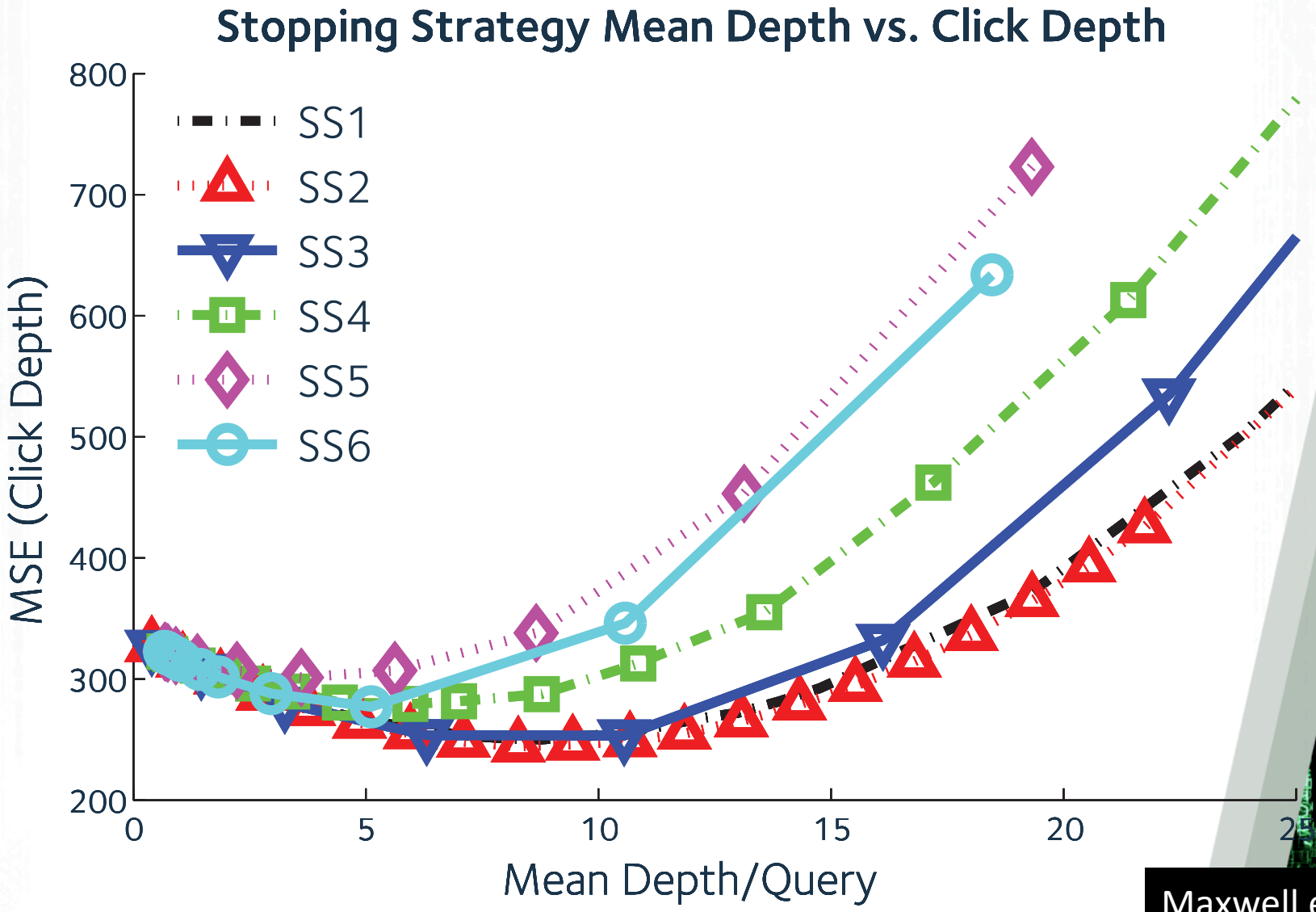
Analysis of Stopping Strategies



Maxwell et al , 2015

Simulated Users vs. Humans

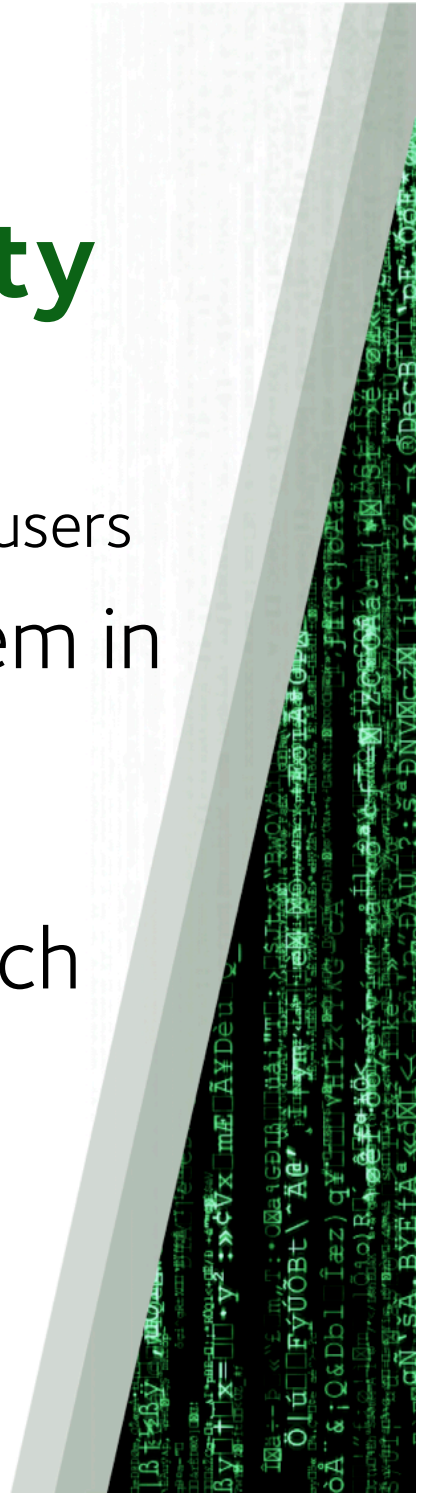
Simulated Users vs. Humans



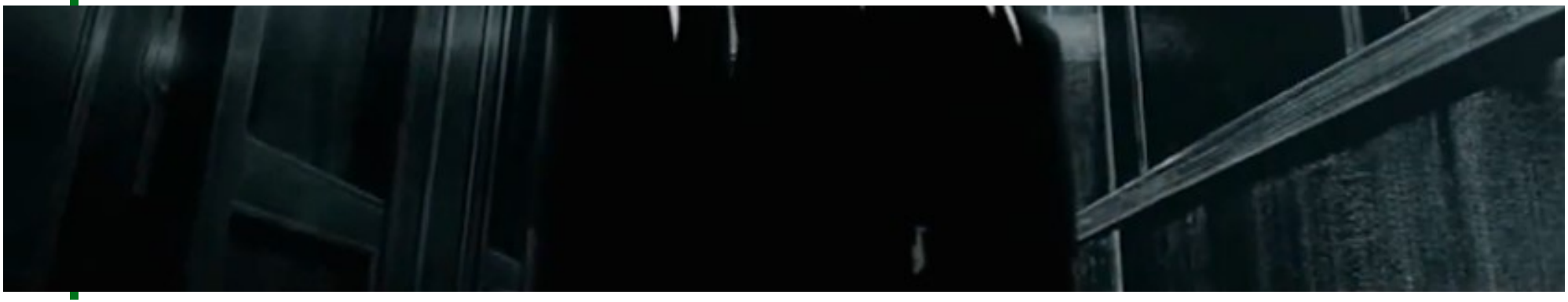
Maxwell et al , 2015

The simulated users have no decision making ability

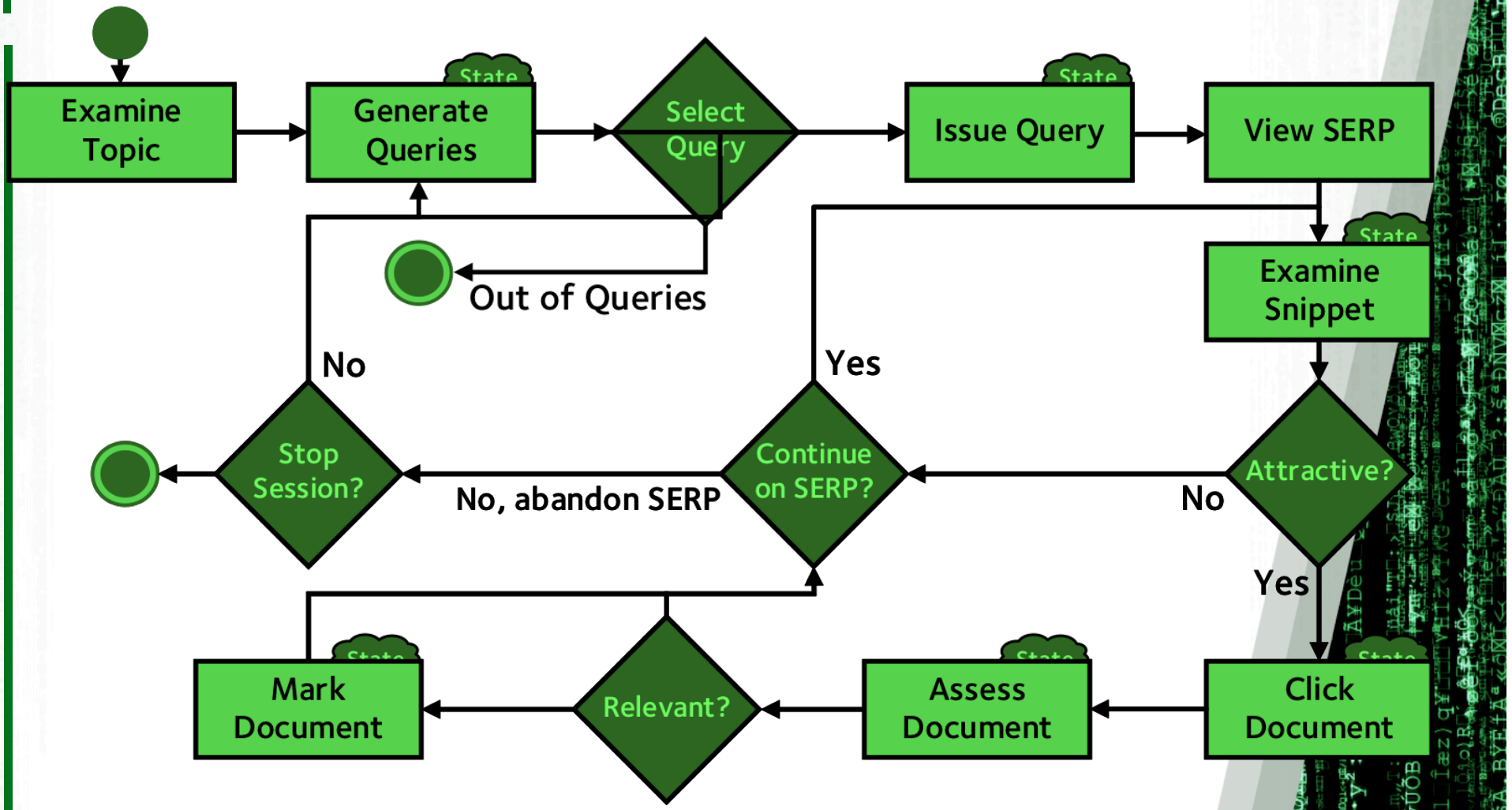
- They still essentially act randomly
  - Such users model produce interactions like real users
- It limits the context we can deploy them in
  - We need topics and goals
  - We need relevance judgments
- What if we create simulated users which
  - decide what to click on, and
  - decide what is relevant?



ADDING AGENCY AND STATE TO THE SIMULATED USERS

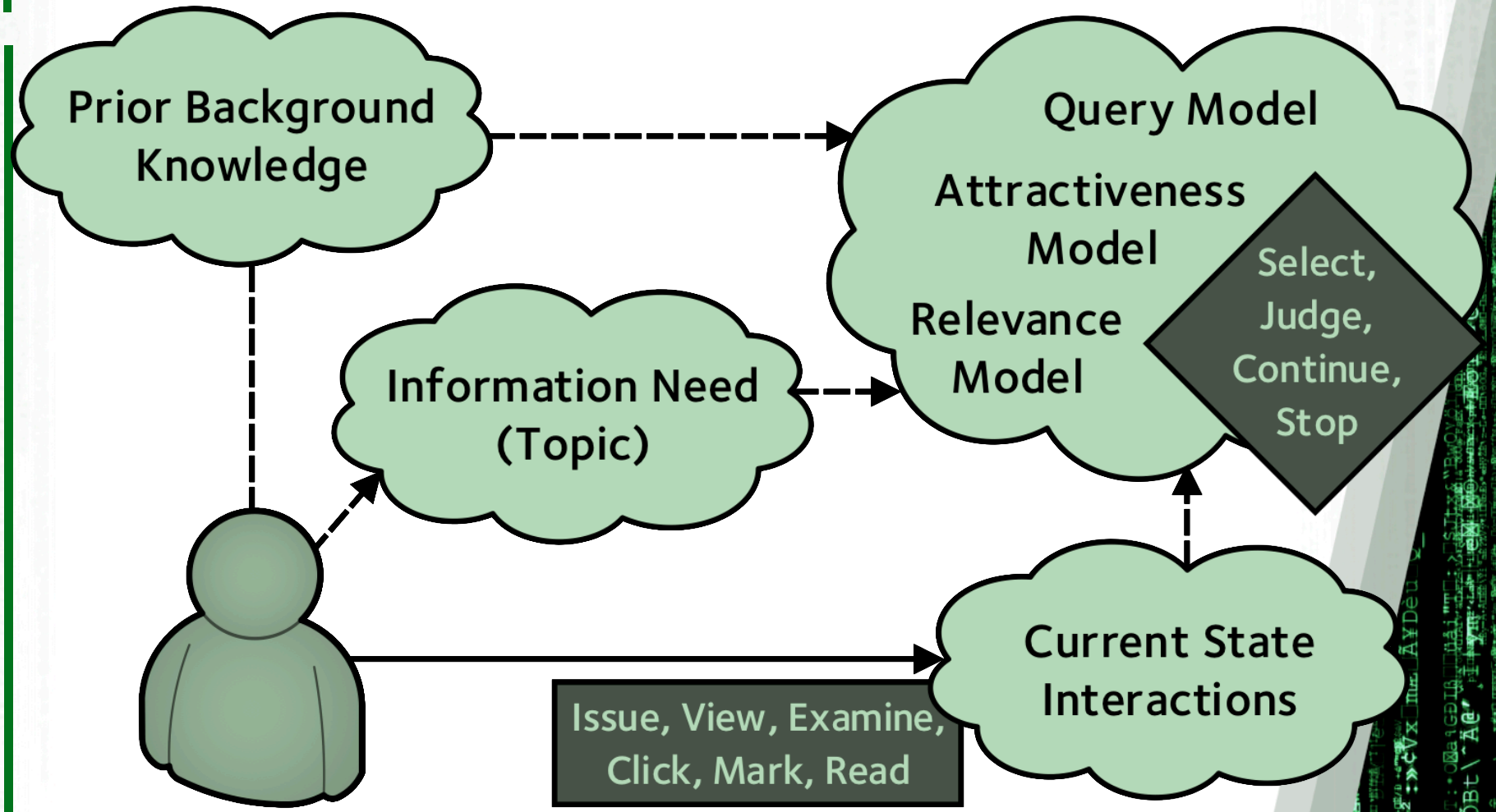


Complex Searcher Model 2



Maxwell et al , 2016

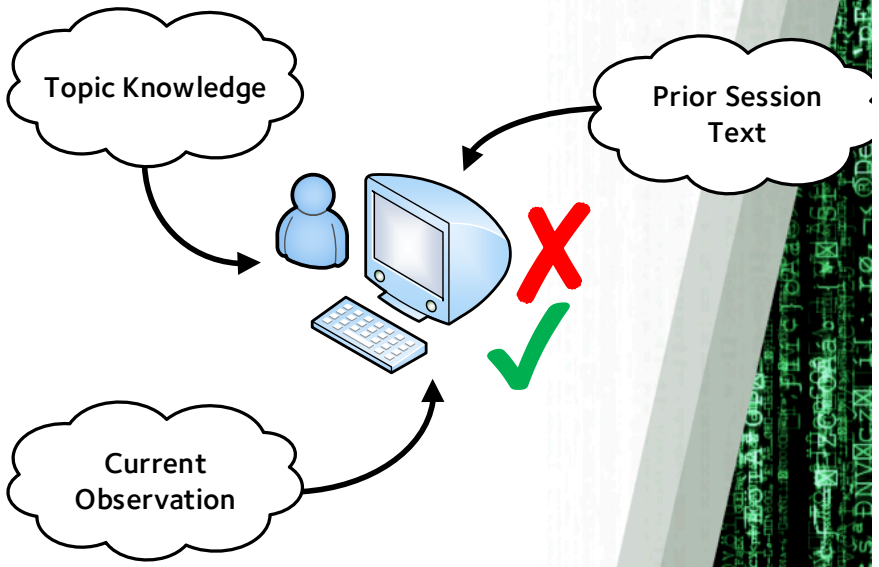
User State Model



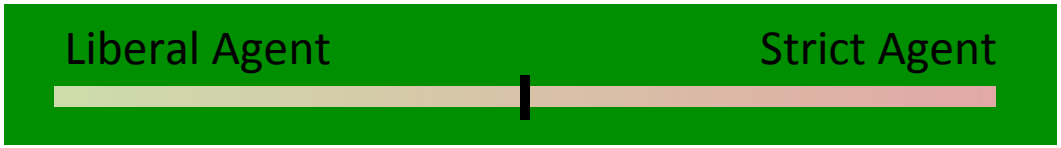
Maxwell et al , 2016

Language Models

- Mixture models based on:
  - Prior Observations
  - Topic Knowledge
  - Background Knowledge



- Decide if attractive/relevant based on a threshold



  - Liberal – more likely to judge as attractive/relevant
  - Strict – less likely to judge as attractive/relevant

Maxwell et al , 2016

CSM+USM

Search Agents

Autonomous agents, with cognitive state – can infer relevance

CSM

Simulated Users

TREC-style and stochastic simulated users

Humans

Controlled study, interaction data

CSM+USM

Search Agents

Autonomous agents,
with cognitive state

48

humans

vs.

48

simulated users

vs.

48

agents

Humans

Controlled study,
interaction data

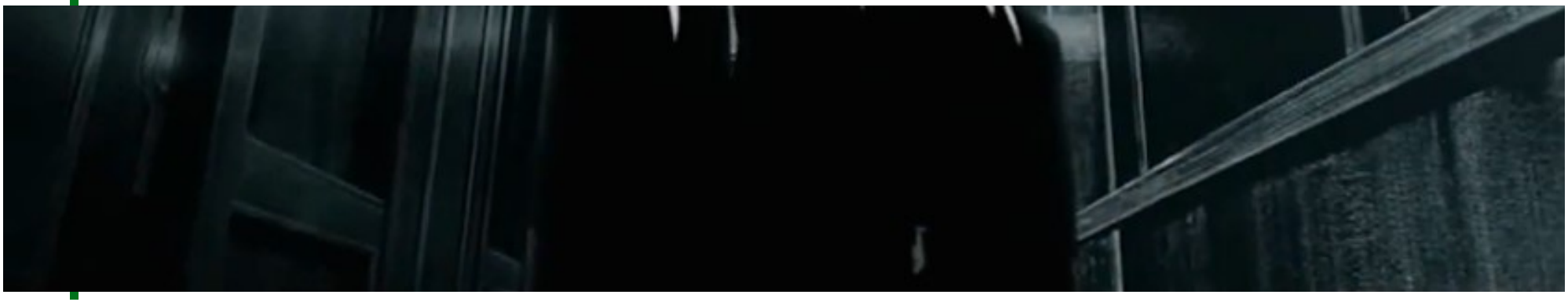
TREC-style and
stochastic simulated users

Sims vs Agents vs Humans

Type	TREC Fixed	TREC Stoch	Sim Stoch	Sim Stoch	Agent S	Agent A	Human (AVG)
Query	QS3+						
Stop.	TREC						
		Behavior and Performance					
Queries	1.0						
Snippets	57						
Docs	57						
Marked	57						
Rel.	22.3						
CG	35.4						
Prec.	0.41						

Maxwell et al , 2016

DO SIMULATED USERS /AGENTS SEARCH LIKE HUMANS?



SIMIIR Toolkit for Simulation

- An open source toolkit for developing simulations is available called **SIMIIR**
- Toolkit lets you configure various pipelines with different components
  - Query Strategies
  - Stopping Strategies
  - Decision Makers
- Available at:

  http://www.github.com/leifos/simiir

Advertisement

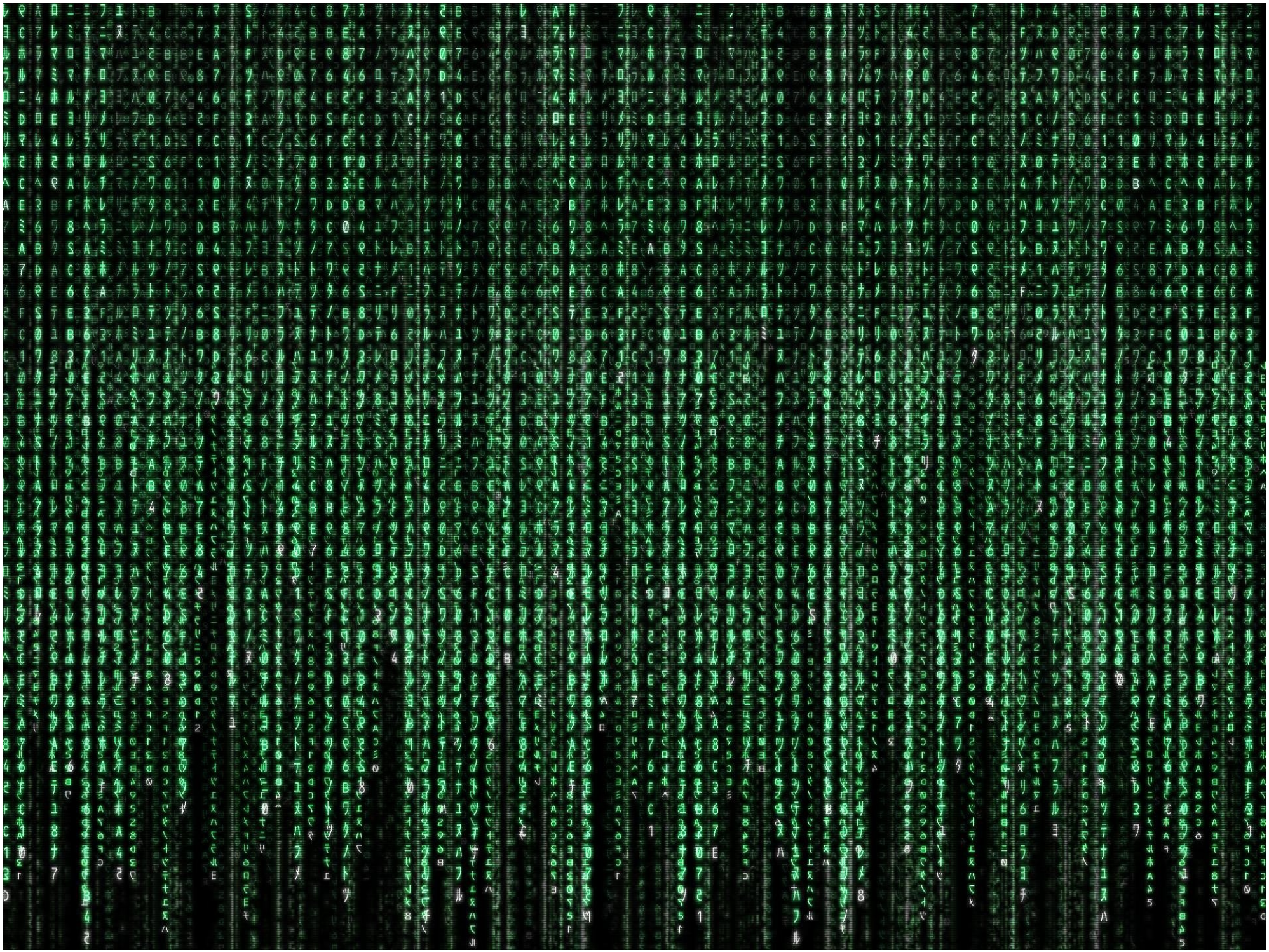
Maxwell & Azzopardi, 2016

Challenges

- Simulation is a power tool that lets us explore and analysis behaviors and performance
- They are an abstraction of reality
  - Require many assumptions
  - Not a replacement of users
- There are many challenges:
  - Creating Realistic Simulated User/Agents
  - Creating Adaptive Agents
  - Change behavior like humans do in response to changes to the interface, costs, etc.

Challenges

- User Model Issues
  - Simple, Complex, Complex II, etc.
- Estimation and Parameterization
  - Configuration of components
    - Query Strategy,
    - Stopping Strategy,
    - Decision Making, etc.
- Generalization to other tasks
- Handling the volume of data
- Trusting and validating the models



Imagine if we had a personalized search agent?

- What if we could make a user model that encodes your search capabilities?
  - And this was then embedded in an agent!
  - Your personal, Mr Smith.
- The agent would anticipate your needs
  - And, hopefully, resolve them
  - Perhaps through some negotiation and dialogue
- Search will change from a very active process to a passive/push process
- How would we evaluate such agents?

Take Home Challenges

- Develop autonomous search agents
- Evaluate session search and dynamic search user complex searcher models
- Evaluate the search performance of humans and agents
- Create for the new “users” of IR
- Move beyond search: search as a service

References

- Personalised Search Time Prediction using Markov Chains by Tuan Vu Tran, David Maxwell, Norbert Fuhr and Leif Azzopardi, ICTIR '17: Proceedings of the International Conference in Theory of IR
- Validating simulated interaction for retrieval evaluation Teemu Pääkkönen, Jaana Kekäläinen, Heikki Keskustalo, Leif Azzopardi, David Maxwell, Kalervo Järvelin, Information Retrieval: Volume 20 Issue 4, 2017
- Agents, Simulated Users and Humans: An Analysis of Performance and Behaviour, David Maxwell, Leif Azzopardi, CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management
- Simulating Interactive Information Retrieval: SimIIR: A Framework for the Simulation of Interaction, David Maxwell, Leif Azzopardi, SIGIR '16: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval
- Simulation of Interaction: A Tutorial on Modelling and Simulating User Interaction and Search Behaviour, Leif Azzopardi, SIGIR '16: Proceedings of the 39th International ACM SIGIR conferenc on Research and Development in Information Retrieval
- Searching and Stopping: An Analysis of Stopping Rules and Strategies, David Maxwell, Leif Azzopardi, Kalervo Järvelin, Heikki Keskustalo, CIKM '15: Proceedings of th 24th ACM International on Conference on Information and Knowledge Management
- Exploring Behavioral Dimensions in Session Effectiveness Teemu Pääkkönen, Kalervo Järvelin, Jaana Kekäläinen, Heikki Keskustalo, Feza Baskaya, David Maxwell, Leif Azzopardi, CLEF'15: Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction
- Untangling Result List Refinement and Ranking Quality: a Framework for Evaluation and Prediction, Jiyin He, Marc Bron, Arjen de Vries, Leif Azzopardi, Maarten de Rijke, SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference 2015

References

- An Initial Investigation into Fixed and Adaptive Stopping Strategies, David Maxwell, Leif Azzopardi, Kalervo Järvelin, Heikki Keskustalo, SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval
- Dynamic Test Collections for Retrieval Evaluation, Ben Carterette, Ashraf Bah, Mustafa Zengin, ICTIR '15: International Conf. on The Theory of IR
- INST: An Adaptive Metric for Information Retrieval Evaluation, Alistair Moffat, Peter Bailey, Falk Scholer, Paul Thomas, ADCS '15: Proceedings of the 20th Australasian Document Computing Symposium
- Best and Fairest: An Empirical Analysis of Retrieval System Bias, Colin Wilkie, Leif Azzopardi, ECIR 2014: Proceedings of the 36th ECIR
- Modeling decision points in user search behavior, Paul Thomas, Alistair Moffat, Peter Bailey, Falk Scholer, IIiX '14: Proceedings of the 5th IIiX
- The economics in interactive information retrieval, Azzopardi, SIGIR '11: Proceedings of the 34th international ACM SIGIR 2011
- Report on the SIGIR 2010 workshop on the simulation of interaction, Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, Mark D. Smucker, ACM SIGIR Forum, Dec 2010
- Simulating simple user behavior for system effectiveness evaluation, Ben Carterette, Evangelos Kanoulas, Emine Yilmaz, CIKM '11: Proceedings of the 20th ACM international conference on Information and knowledge management
- A Corpus of Realistic Known-Item Topics with Associated Web Pages in the ClueWeb09, Matthias Hagen, Daniel Wägner, and Benno Stein. ECIR 15, In Proceedings of 37th ECIR, 20
- What Was the Query? Automatically Generating Queries for Document Sets with Applicati Cluster Labeling. Matthias Hagen, Maximilian Michel, and Benno Stein, In Proceedings of th International Conference on Applications of Natural Language to Information Systems, 20

References

- System effectiveness, user models, and user utility: a conceptual framework for investigation, Ben Carterette, SIGIR '11: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval
- Query side evaluation: an empirical analysis of effectiveness and effort, Leif Azzopardi, SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval
- Automatically generating queries for prior art search, Erik Graf, Leif Azzopardi, Keith Van Rijsbergen, CLEF'09: Proceedings of the 10th CLEF
- Retrievability: an evaluation measure for higher order information access tasks, Leif Azzopardi, Vishwa Vinay,CIKM '08: Proceedings of the 17th ACM conference on Information and knowledge management
- Building simulated queries for known-item topics: an analysis using six european languages, Leif Azzopardi, Maarten de Rijke, Krisztian Balog, SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval
- Overview of WebCLEF 2006, Krisztian Balog, Leif Azzopardi, Jaap Kamps, Maarten De Rijke,CLEF'06: Proceedings of the 7th international conference on Cross-Language Evaluation Forum: evaluation of multilingual and multi-modal information retrieval
- Automatic construction of known-item finding test beds, Leif Azzopardi, Maarten de Rijke, SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval

References

- A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, Shlomo Geva SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval
- Clef 2017 technologically assisted reviews in empirical medicine overview Evangelous Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker Working Notes of CLEF 2017
- Modeling behavioral factors ininteractive information retrieval Feza Baskaya, Heikki Keskustalo, Kalervo Järvelin CIKM '13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management
- Time drives interaction: simulating sessions in diverse searching environments Feza Baskaya, Heikki Keskustalo, Kalervo Järvelin SIGIR '12: Proceedings of the 35th international ACM SIGIR conference