The Information Bottleneck Theory of Deep Learning

Frederico Guth

Qualifying Examination

Prof. Teófilo de Campos (supervisor)

UnB

Prof. John Shawe-Taylor

UCL

Prof. Moacir Antonelli Ponti

USP

UnB

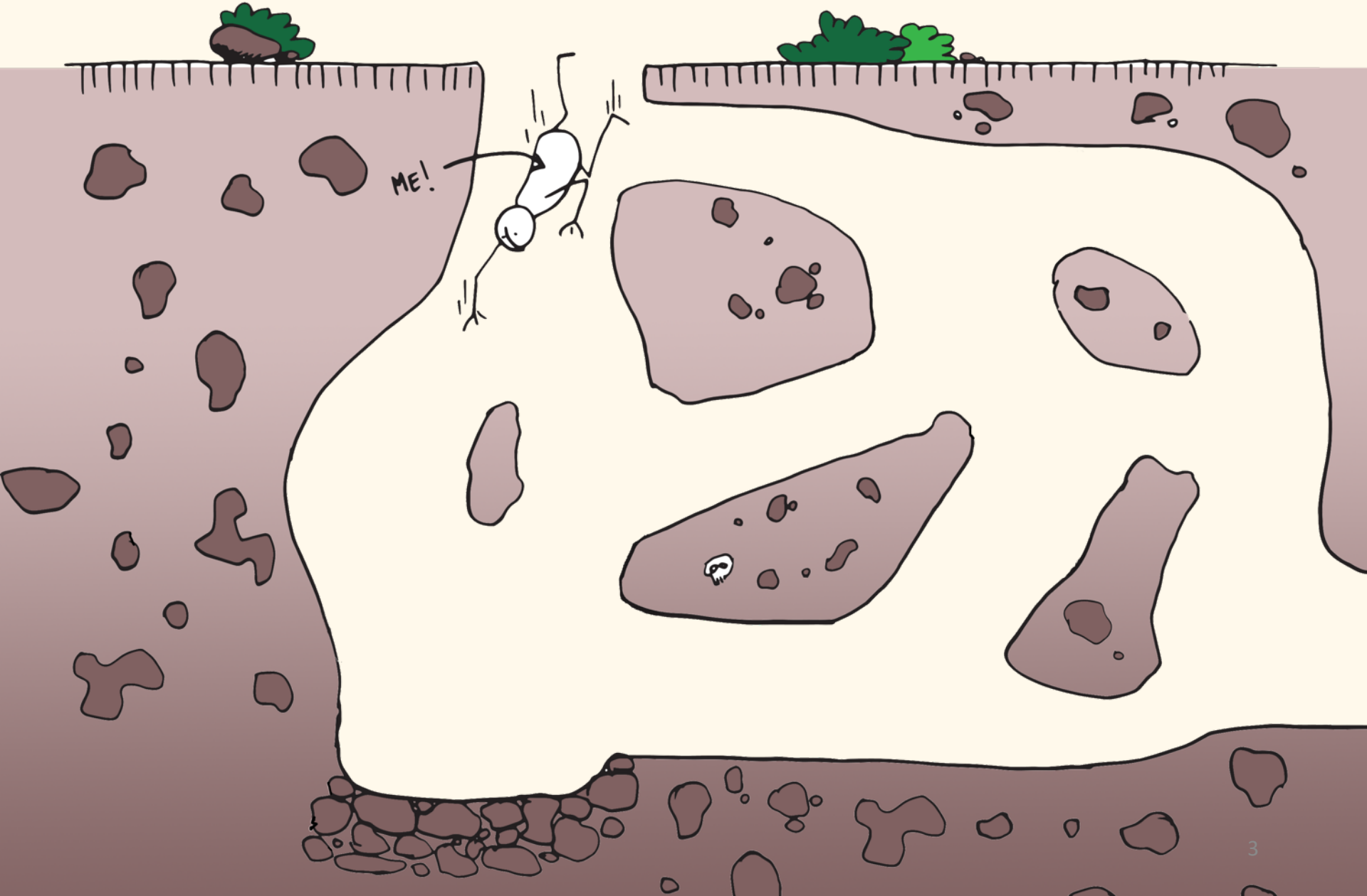
fapdf
Fundação de Apoio à Pesquisa do Distrito Federal

Brasília, 13/07/2020.

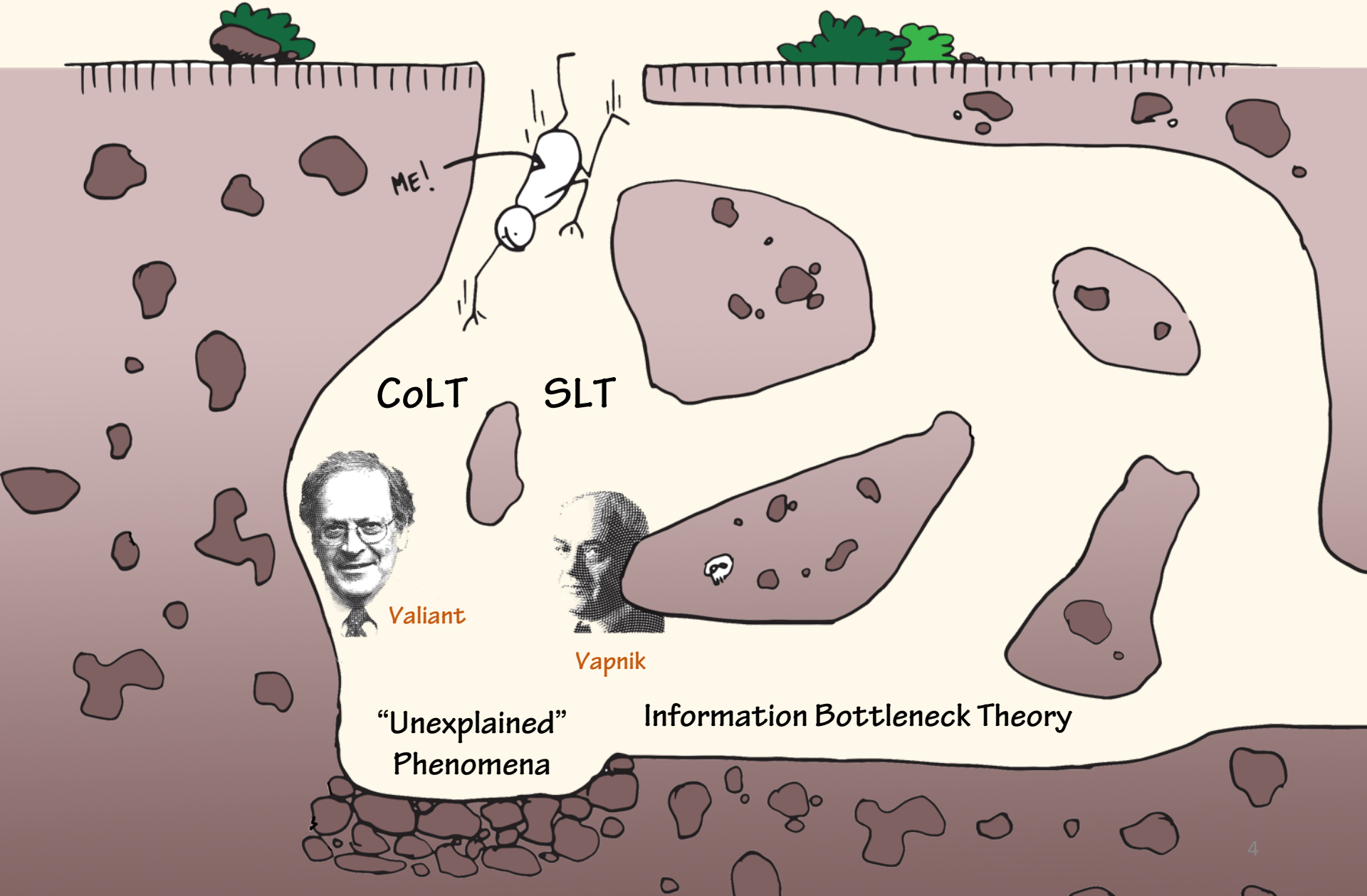
Summary

1) Context/Problem
2) Research Objective
3) Literature Review: Background
4) Proposal
5) References

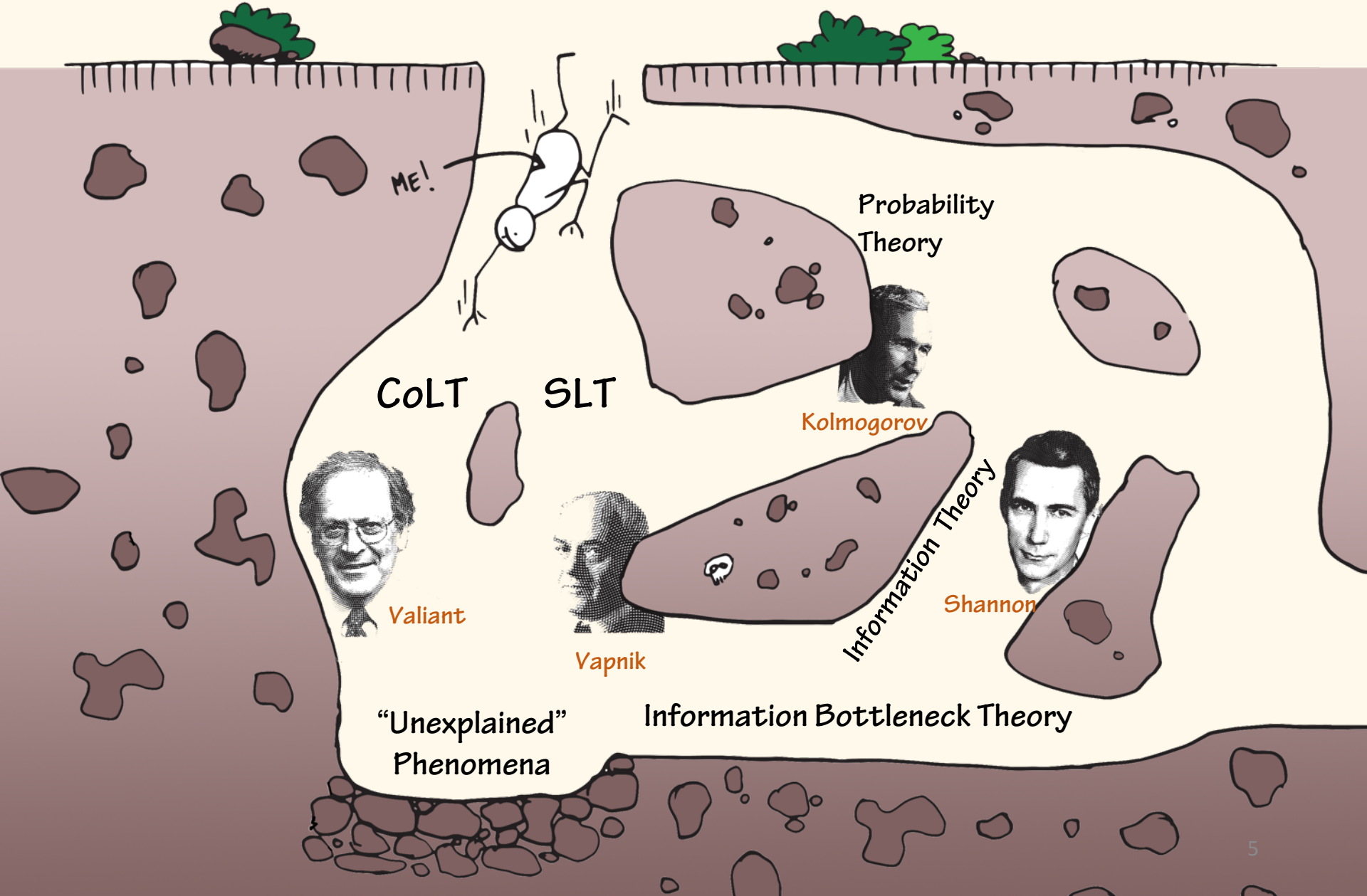
Fred in Theoryland ...



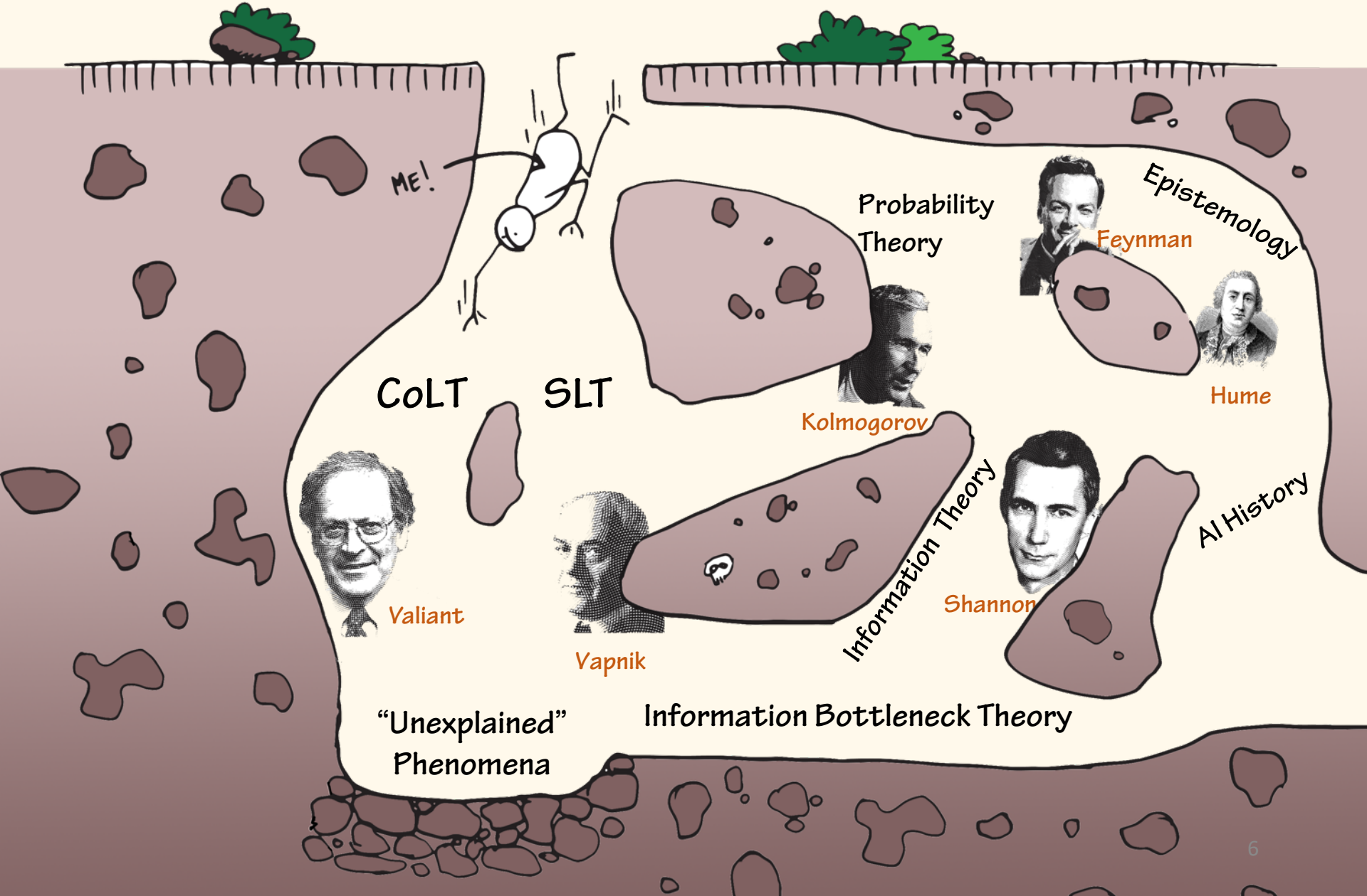
Fred in Theoryland ...



Fred in the Theoryland ...



The Problem

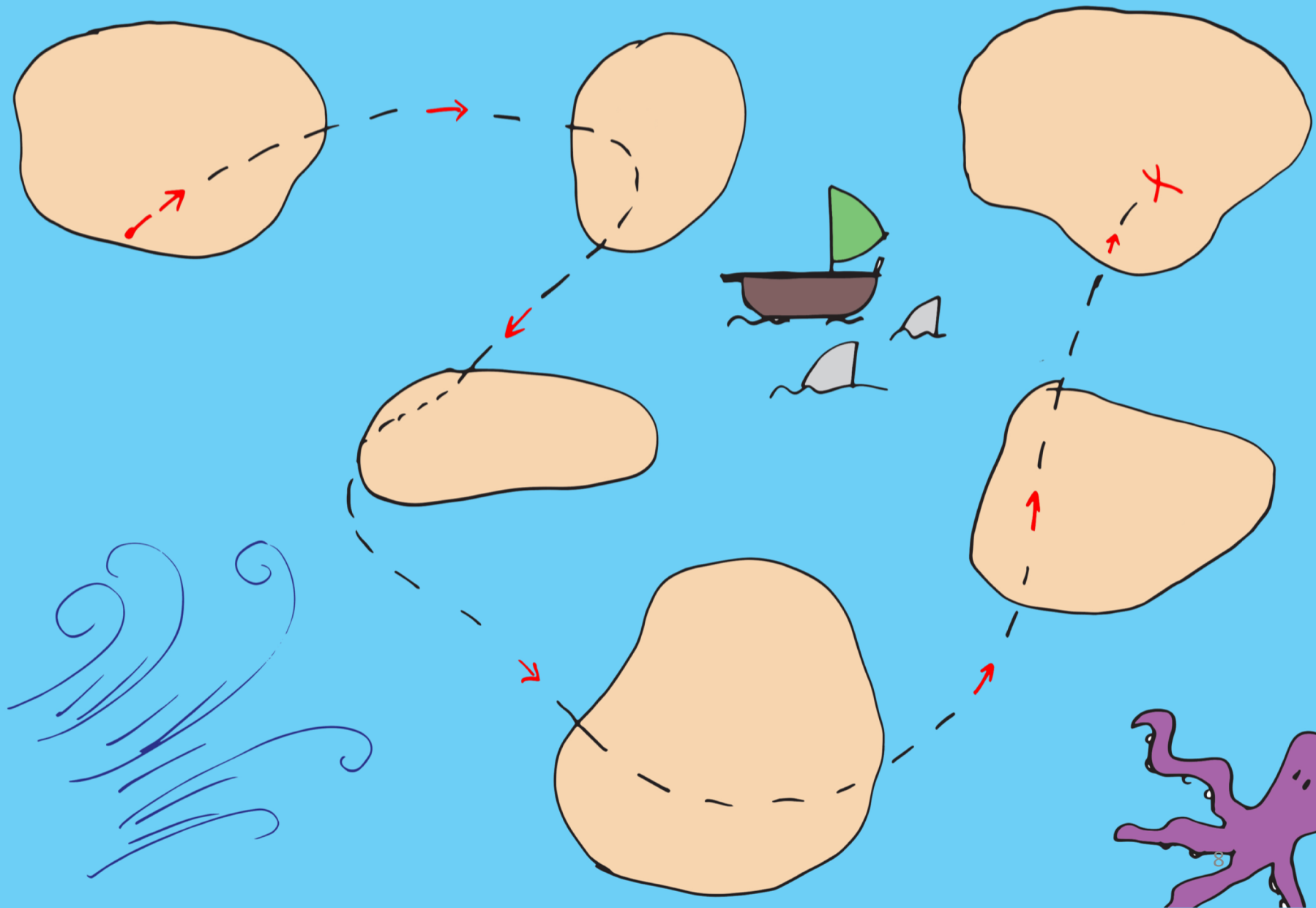

ME!
Probability Theory
Epistemology
Feynman
Hume
CoLT
SLT
Kolmogorov
Valiant
Vapnik
Information Theory
Shannon
AI History
"Unexplained" Phenomena
Information Bottleneck Theory

Research Objective

To investigate the Information Bottleneck Theory and to consolidate the literature into a comprehensive digest.

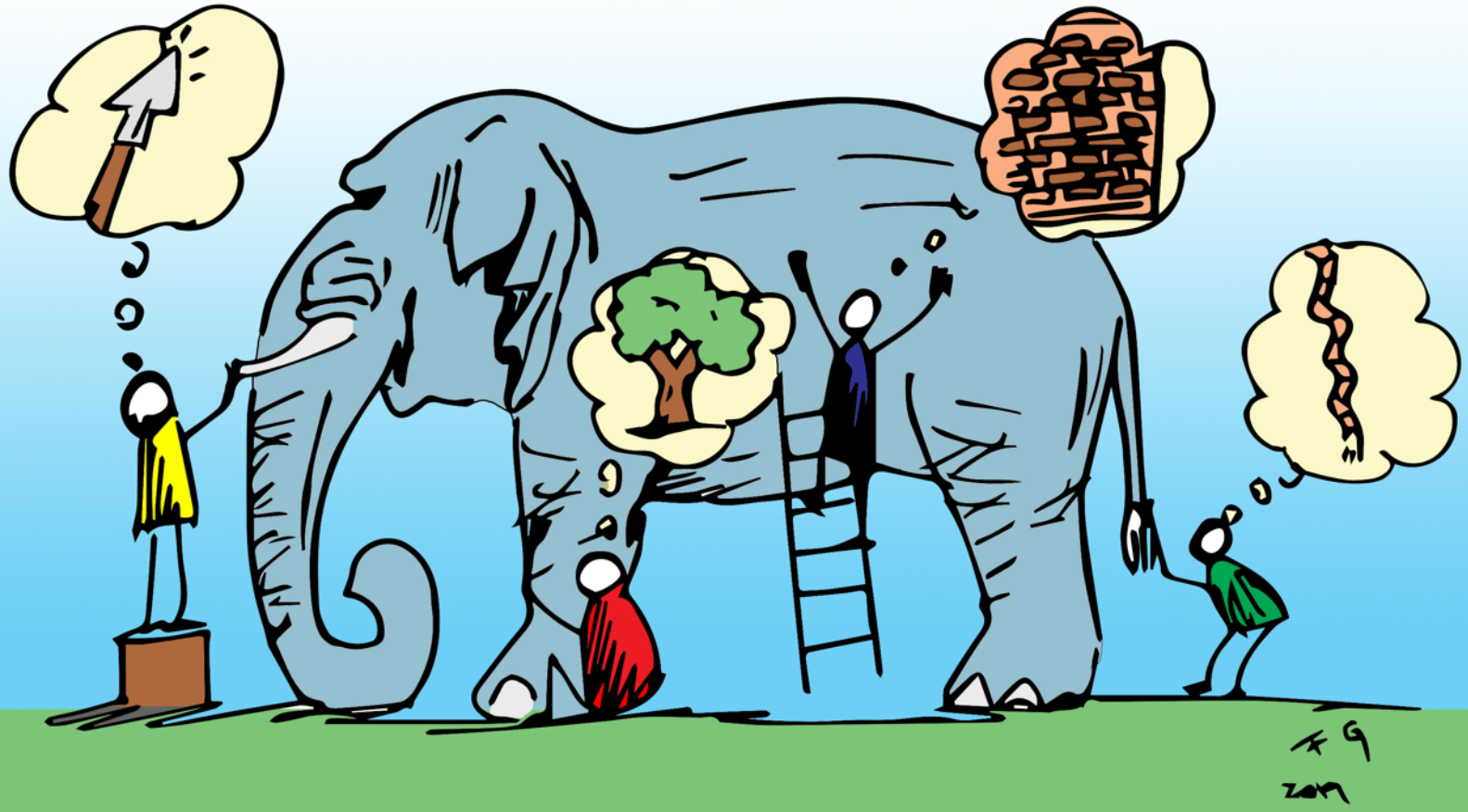
- Which “unexplained” Deep Learning phenomena can Information Bottleneck Theory address?
- What are the assumptions of this new theory?
- How does IBT compares to current Machine Learning Theory?

Research Objective



From Intelligence to Language

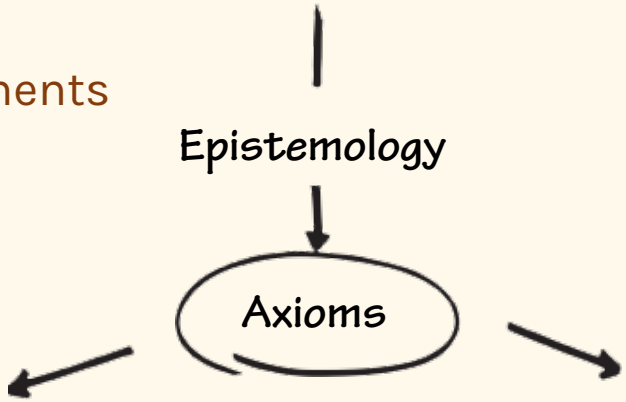
Intelligence is the ability to predict a course of action to achieve success in specific goals.



From Language to Machine Learning Theory

Greeks

i) Knowledge is a set of true or false statements
ii) unambiguous
iii) consistent
iv) minimal



Babylonians

i) Knowledge is a set of justified beliefs
ii) "common sense"
iii) consistent
iv) minimal

Machine Learning Theory

Machine Learning Theory deduces from Probability Theory bounds for the behaviour of machine learning algorithms.

Learning problem setting:

Choose from the hypothesis space the one hypothesis that best approximates the concept.

Critiques on Machine Learning Theory

The case for a new narrative.

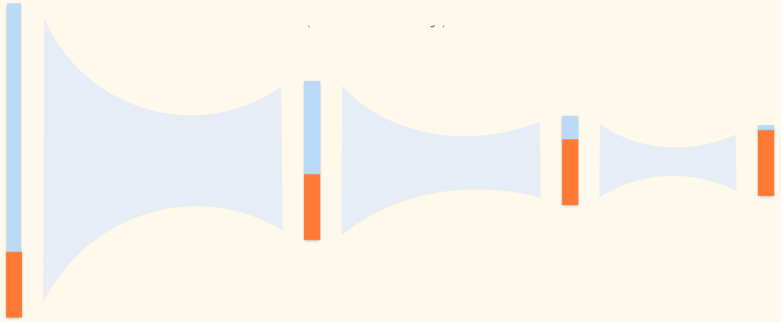
- No assumption on P(X,Y)
- No notion of “time”
- i.i.d. sampling
- Vacuous bounds for DL

“Unexplained” DL Phenomena

- DNN generalisation w/ hundred million params
- Flat minima
- Disentanglement
- Critical Learning Periods
- Superconvergence

Information Bottleneck Theory

A deep neural network as a communication channel between the input (the source) and the representation (the receiver).



Learning problem setting:

Find a representation T* that is a minimal sufficient statistic of input X in relation to Y,

$$T^* = \arg\min_{T} I(T;X) \quad \text{s.t.} \quad I(T;Y) = I(X;Y)$$

Machine Learning Theory vs. Information Bottleneck Theory

MLT

- Not DL specific
- generalisation: number of params (data)
- Worse-case
  Model-dependent
  Distribution-independent
  bounds

IBT

- DL specific
- generalisation: amount of information
- Typical
  Model-independent
  Data-dependent
  bounds

From Language to Information Bottleneck Theory

Intelligence

↓

Language

↓

Probability Theory

↓

Machine Learning

Information Theory

↓

Information Bottleneck Theory

From Probability to Information Theory

"Information is what changes belief"

$$i_S(e) = f(\mathcal{L}(e;S)).$$

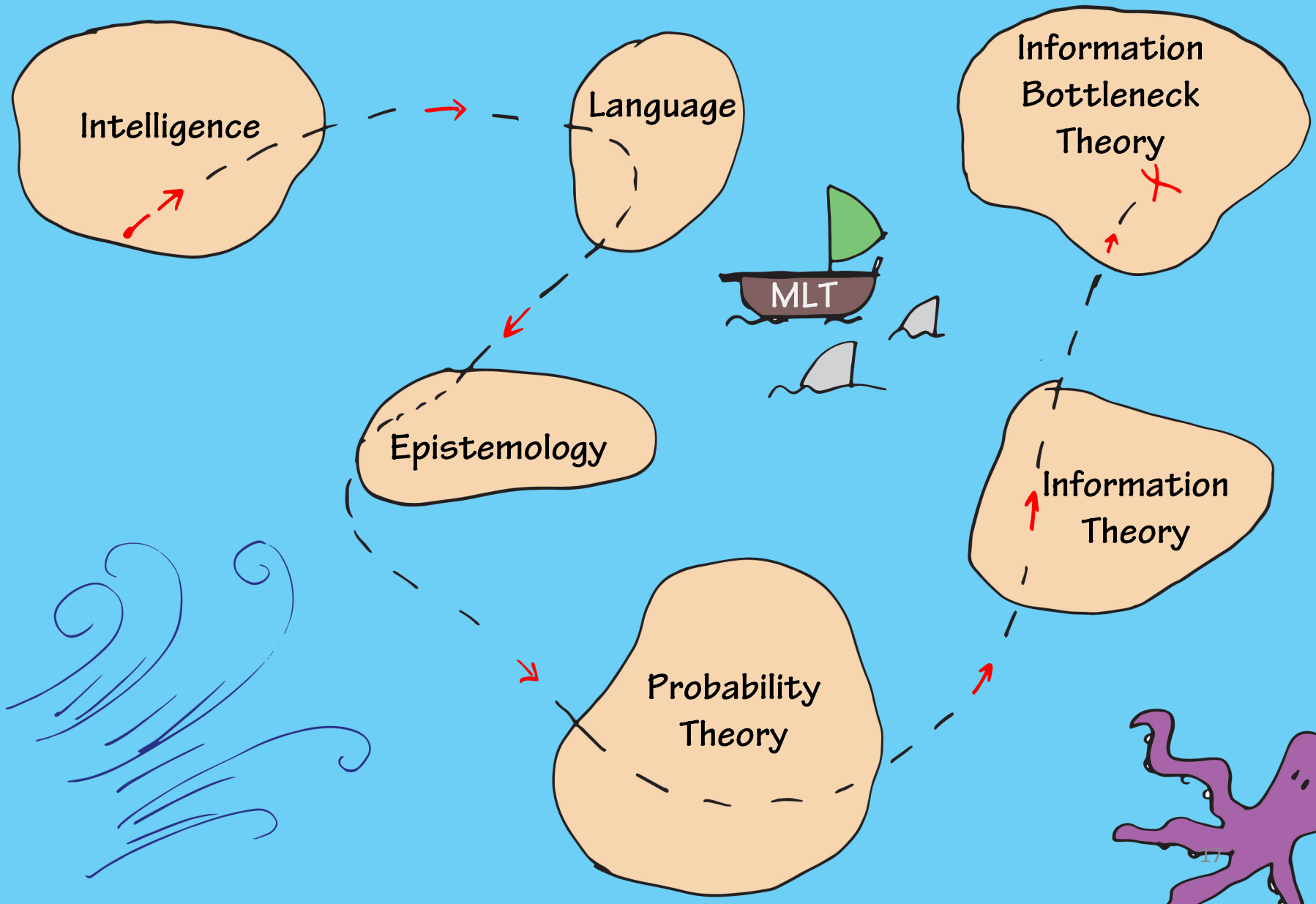
$$\begin{cases} f(\mathcal{L}_1 \wedge \mathcal{L}_2) & = f(\mathcal{L}_1) + f(\mathcal{L}_2) \\ f(1) & = 0 \\ f & \text{is continuous.} \end{cases}$$

* keeps consistency

$$I[S] = -\log p(s)$$

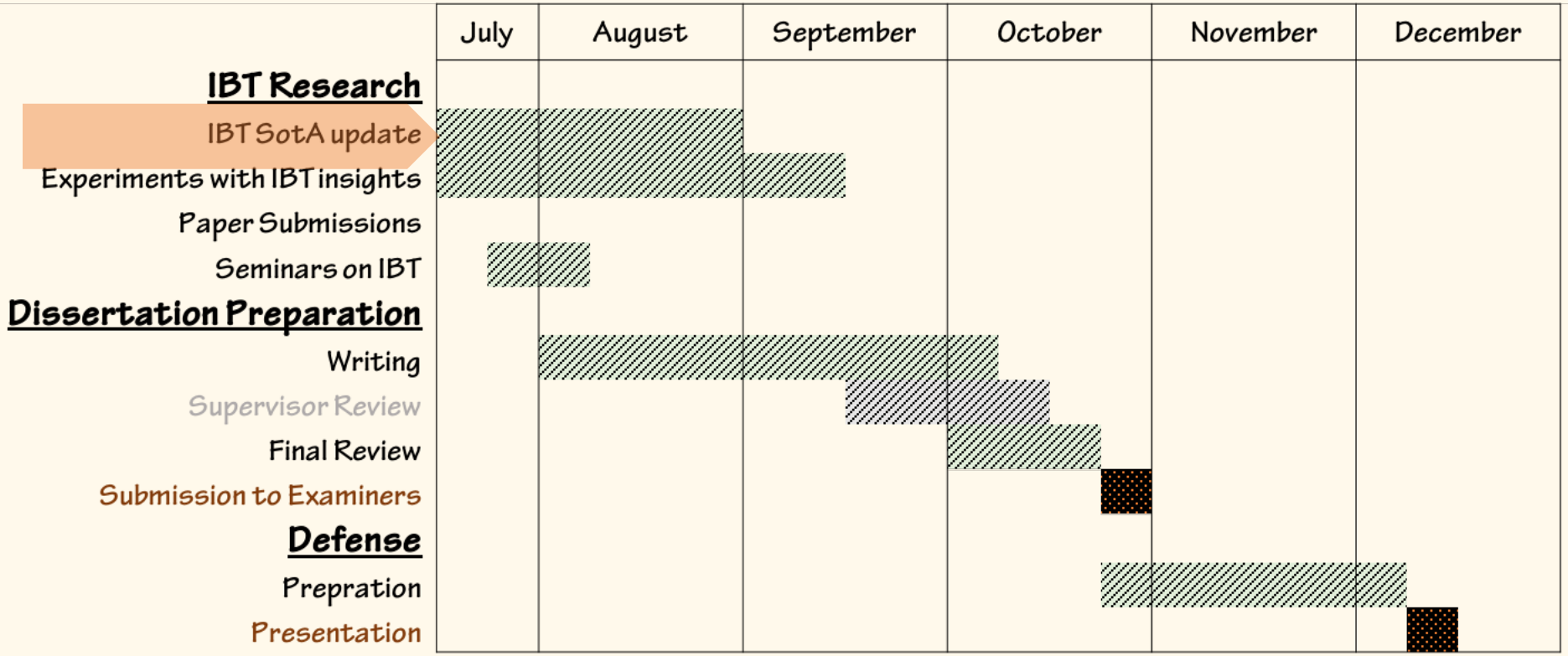
Shannon's self-information definition

The journey


Intelligence
Language
Information Bottleneck Theory
MLT
Epistemology
Information Theory
Probability Theory

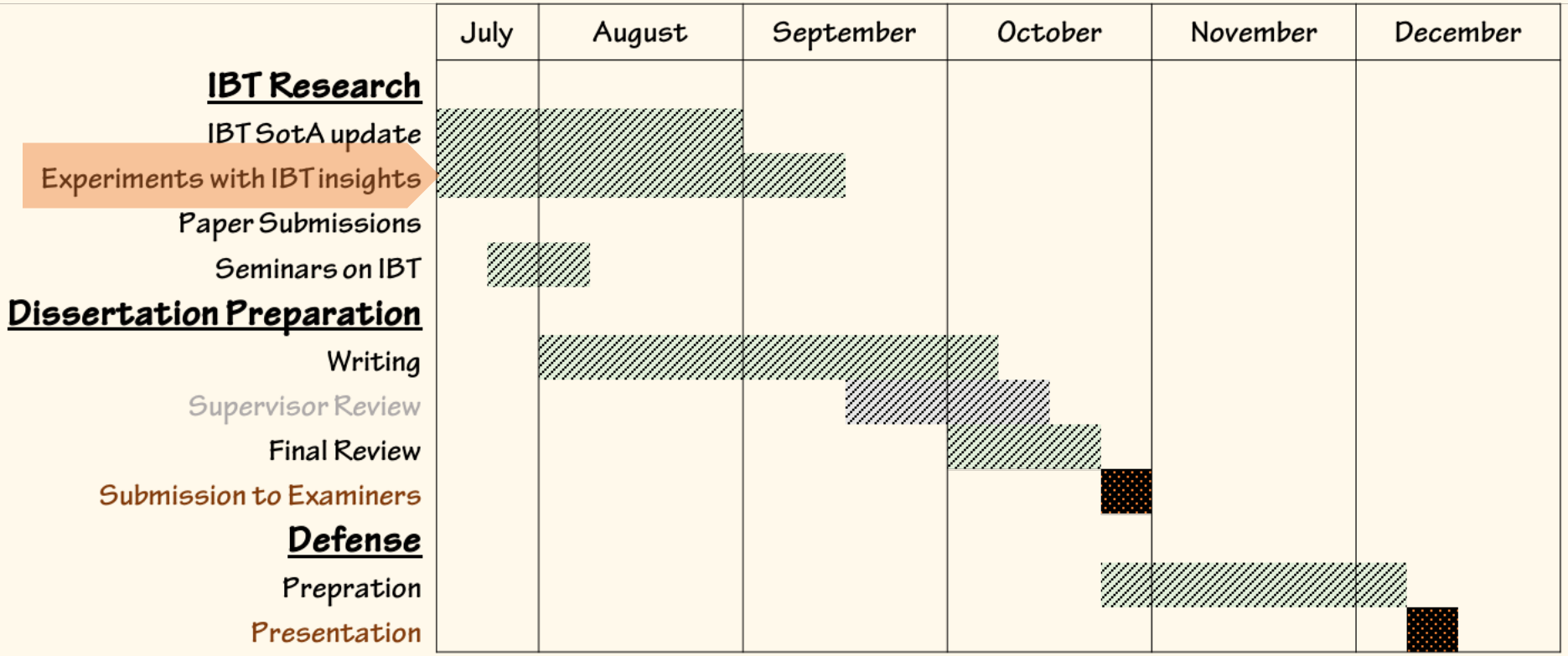
Proposal

Structured review of 15 articles (CVPR, NeurIPS, ICLR, etc)



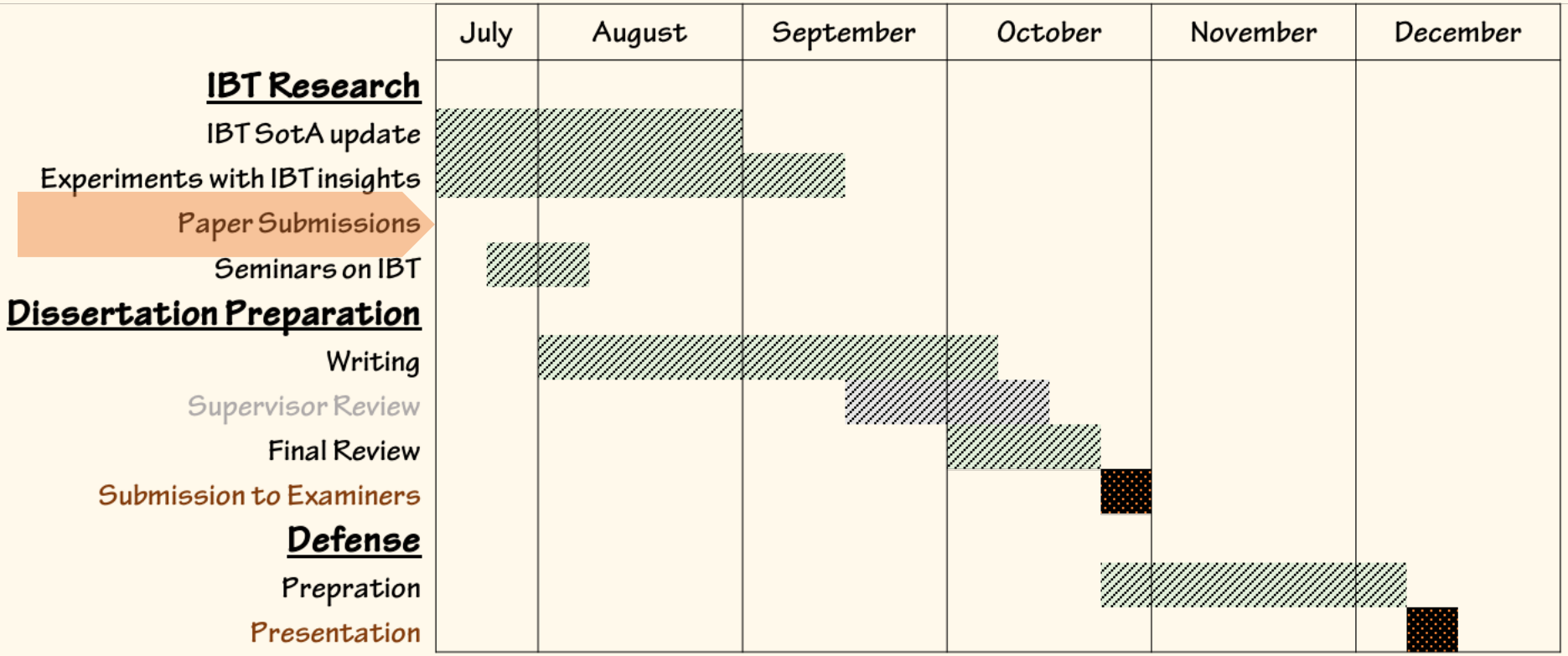
Proposal

Exploratory experiments



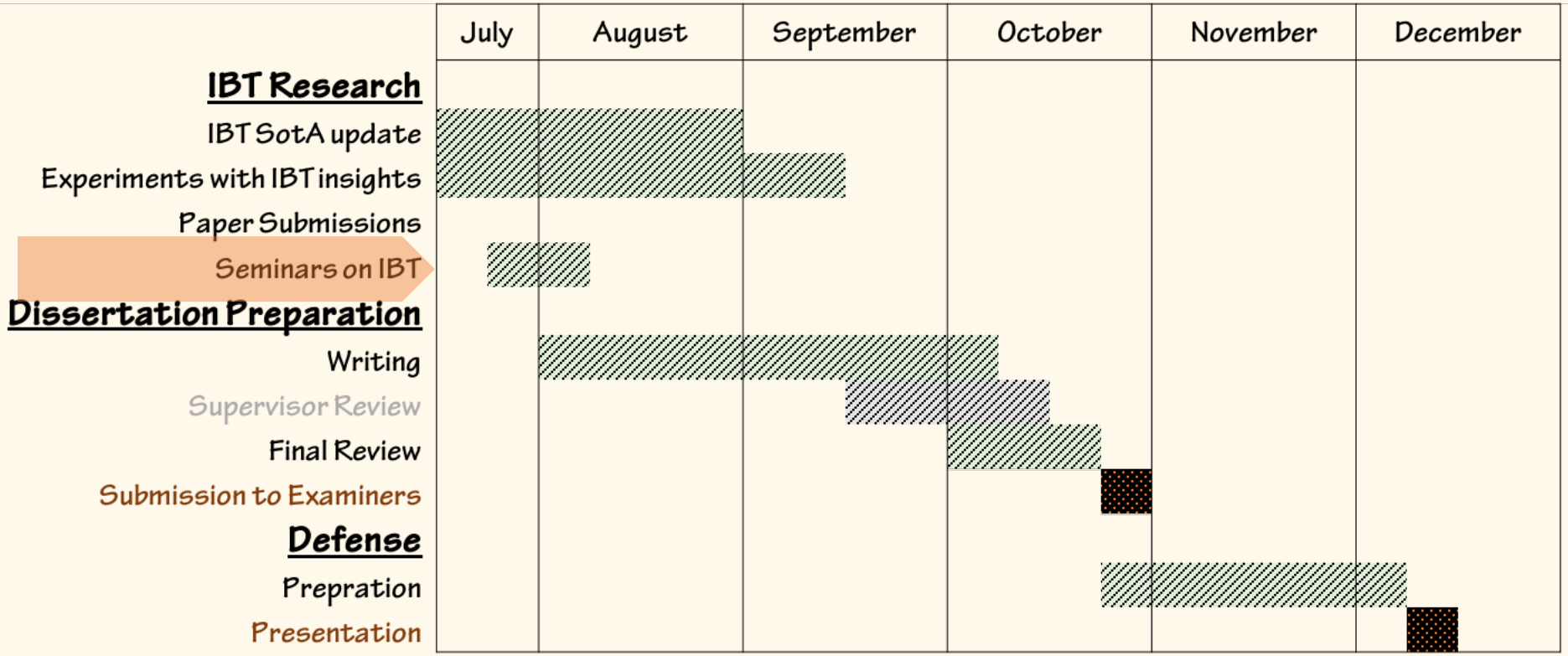
Proposal

Important deadlines in the next 4 mo:
- AAAI, EACL, ICLR, CVPR



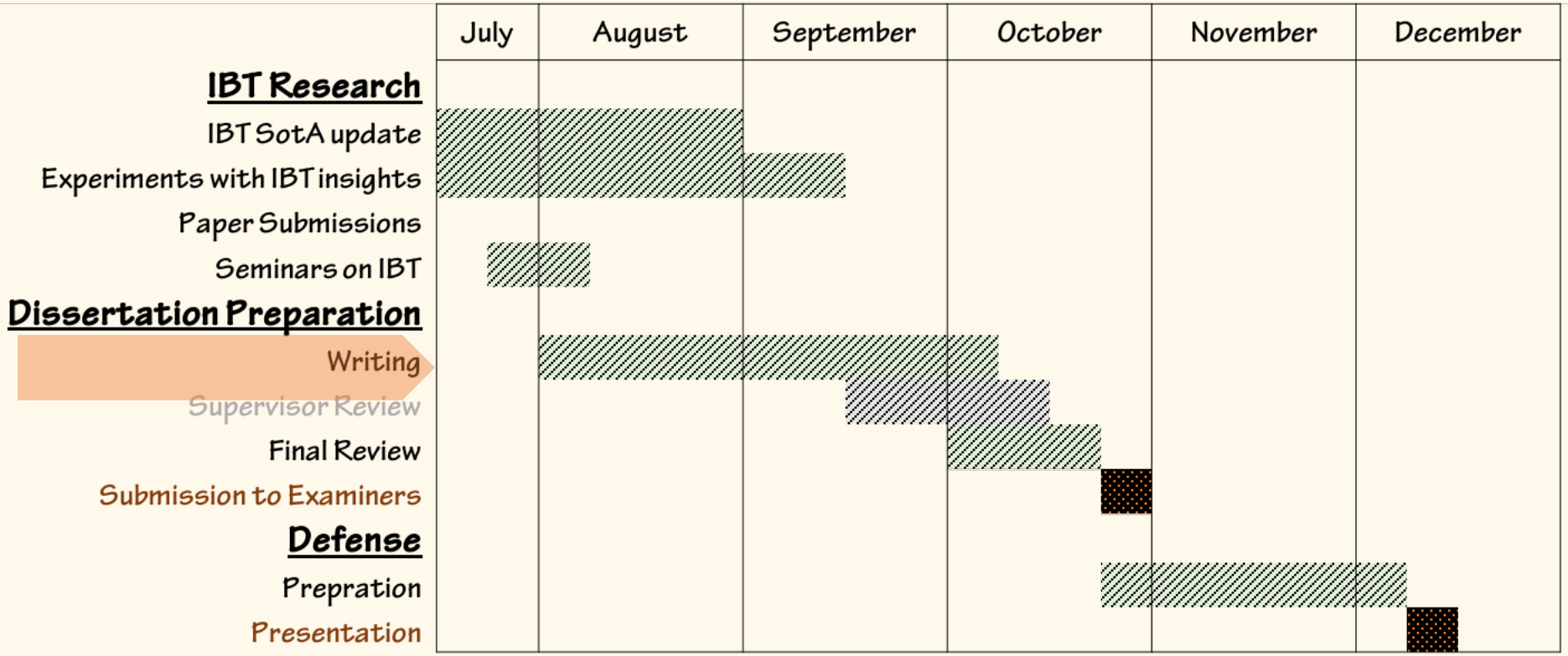
Proposal

For our research group

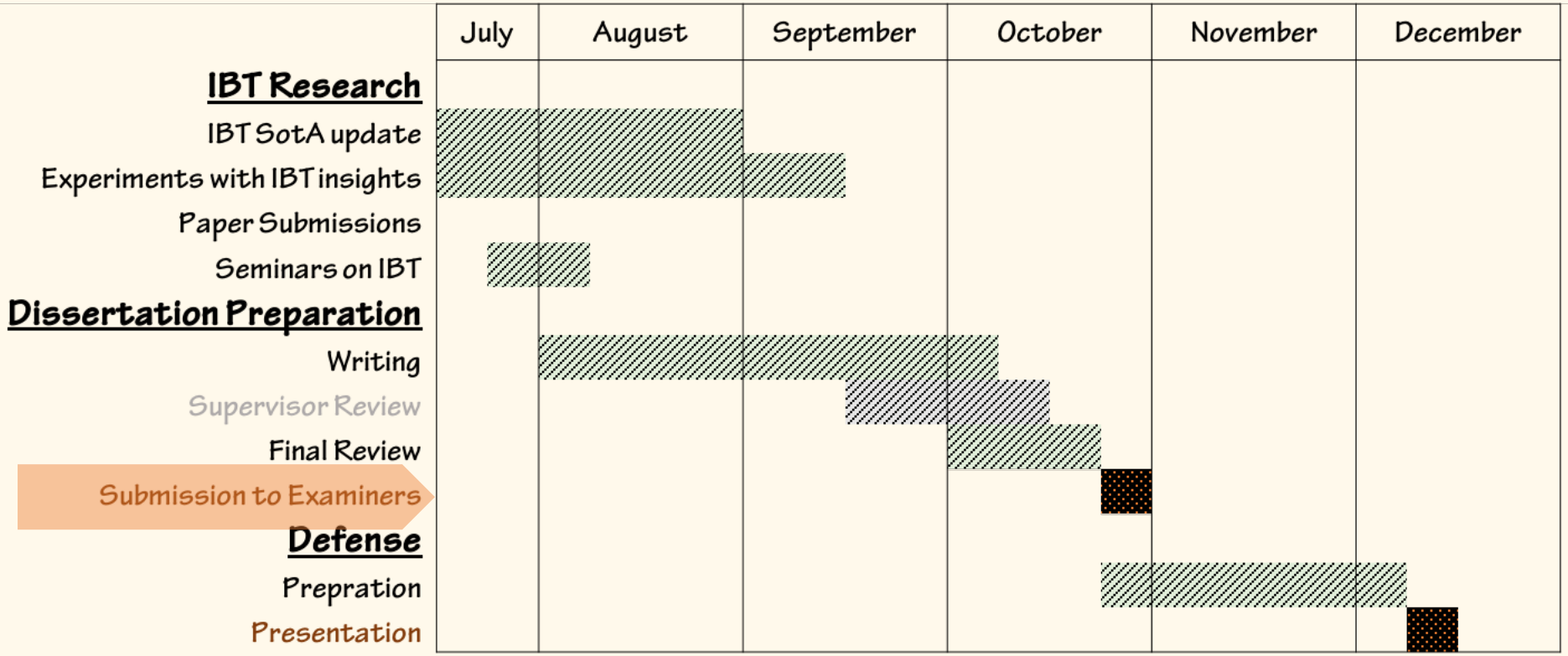


Proposal

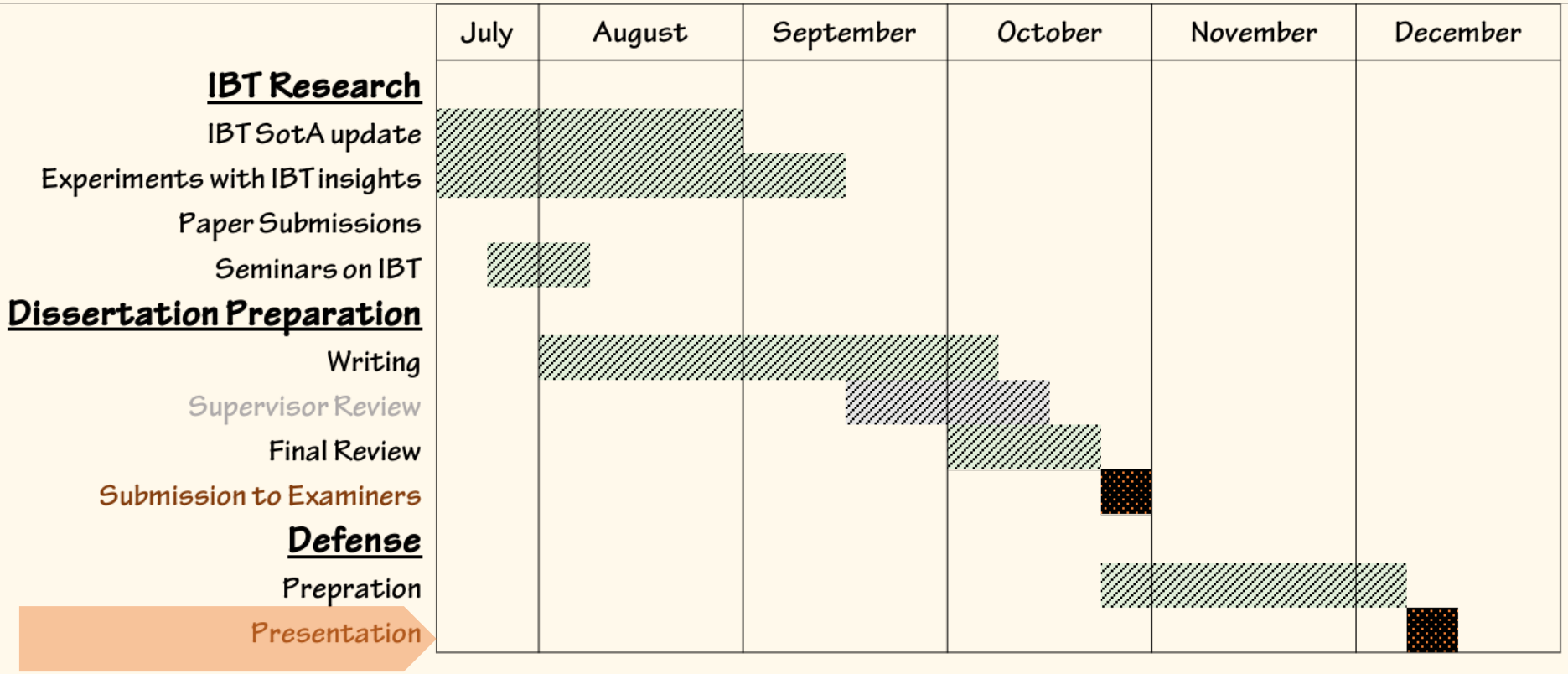
IBT problem setting, MLT vs. IBT, explained phenomena



Proposal


July
August
September
October
November
December
IBT Research
IBT SotA update
Experiments with IBT insights
Paper Submissions
Seminars on IBT
Dissertation Preparation
Writing
Supervisor Review
Final Review
Submission to Examiners
Defense
Prepration
Presentation

Proposal


July
August
September
October
November
December
IBT Research
IBT SotA update
Experiments with IBT insights
Paper Submissions
Seminars on IBT
Dissertation Preparation
Writing
Supervisor Review
Final Review
Submission to Examiners
Defense
Prepration
Presentation

References (1)

(Cover):

John Shawe-Taylor and Omar Rivasplata. *Statistical Learning Theory: A Hitchhiker's Guide*. Dec. 2018. [Online] url: https://youtube.videoken.com/embed/Bv5gzFZS5OI

Rodrigo F. Mello and Moacir Antonelli Ponti. *Machine learning: a practical approach on the statistical learning theory*. Springer, 2018.

Fred in Theoryland:

Jonathan Baxter. "A model of inductive bias learning". In: *Journal of artificial intelligence research* 12 (2000), pp. 149–198.

L. G. Valiant. "A theory of the learnable". In: *Proceedings of the sixteenth annual ACM symposium on Theory of computing - 84*. ACM Press, 1984. doi: 10.1145/800057.808710.

V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

Ali Rahimi. *Test-of-time award presentation (NeurIPS 2017)*. [Online] url: https://www.youtube.com/watch?v=Qi1Yry33TQE

References (2)

Naftali Tishby. *Information Theory of Deep Learning*. [Online; Published: 2017-10-16. Last Accessed: 2020-03-06]. Oct. 16, 2017. url: https://www.youtube.com/watch?v=FSfN2K3tnJU.

Jimmy Soni and Rob Goodman. *A mind at play: how Claude Shannon invented the information age*. Simon and Schuster, 2017.

Fred Guth. *An Information Theoretical Transferability Metric*. Tech. rep. UnB, June 2019. Fred Guth and Teofilo Emidio de Campos. *Research Frontiers in Transfer Learning – a systematic and bibliometric review*. 2019. arXiv: 1912.08812 [cs.DL].

The Problem:

David Hume. *Tratado da natureza humana-2a Edição*. Editora UN-ESP, 2009. isbn: 97885-7139-901-3.

Richard Feynman. *The Character of Physical Law*. Modern Library, 1994. isbn: 0-679-60127-9.

From Intelligence to Language:

John G. Saxe. *The blind men and the elephant*. Enrich Spot Limited, 2016.

References (3)

From Language to Machine Learning Theory:

Alexander Terenin and David Draper. “Cox’s Theorem and the Jaynesian Interpretation of Probability”. In: (2015). arXiv: 1507.06597 [math.ST].

Damian Radoslaw Sowinski. “Complexity and stability for epistemic agents: The foundations and phenomenology of configurational Entropy”. PhD thesis. 2016.

Ariel Caticha. *Lectures on Probability, Entropy, and Statistical- Physics*. arXiv: 0808.0012 [physics.data-an].

Critiques on Machine Learning Theory:

Alessandro Achille, Matteo Rovere, and Stefano Soatto. *Critical Learning Periods in Deep Neural Networks*. 2017. arXiv: 1711.08856 [cs.LG].

Leslie N. Smith and Nicholay Topin. “Super-convergence: Very fast training of neural networks using large learning rates”. In: *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*. Vol. 11006. International Society for Optics and Photonics. 2019, p. 1100612.

Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification”. In: *ACL*. Association for Computational Linguistics, 2018. url: http://arxiv.org/ abs/1801.06146.

References (4)

Jeremy Howard and Sebastian Ruder. “Universal Language Model Fine-tuning for Text Classification”. In: *ACL*. Association for Computational Linguistics, 2018. url: http://arxiv.org/ abs/1801.06146.

John R. Pierce. *An Introduction to Information Theory: Symbols, Signals and Noise*. Dover Publications. isbn: 0486240614.

Information Bottleneck Theory:

Ravid Shwartz-Ziv and Naftali Tishby. *Representation Compression and Generalization in Deep Neural Networks*. 2019. url: https: //openreview.net/forum?id=SkeL6sCqK7.