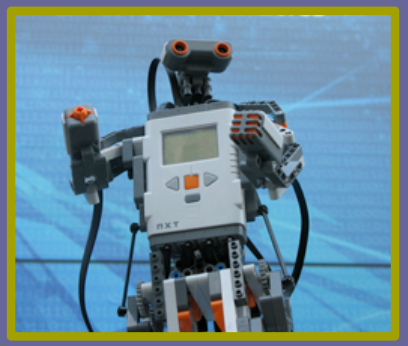
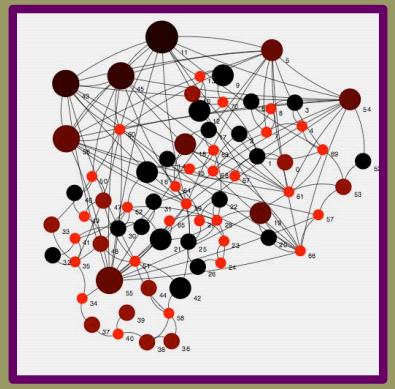
COMP40020 Human Language Technologies

Evolution of Language Technologies

January 2019





Prof. Julie Berndsen
School of Computer Science
Julie.Berndsen@ucd.ie

HLT2

Contents:

- Some literature
- Some highlights from the past
- Some current applications and directions

Aim:

- To give an insight by way of some examples, rather than an exhaustive overview, into how the field of human language technologies has evolved. The focus today is on fundamental ideas with more specific elements covered in later lectures.

- Enrolment key on csmoodle.ucd.ie: 40020HLT19 (note 2018-19 version)

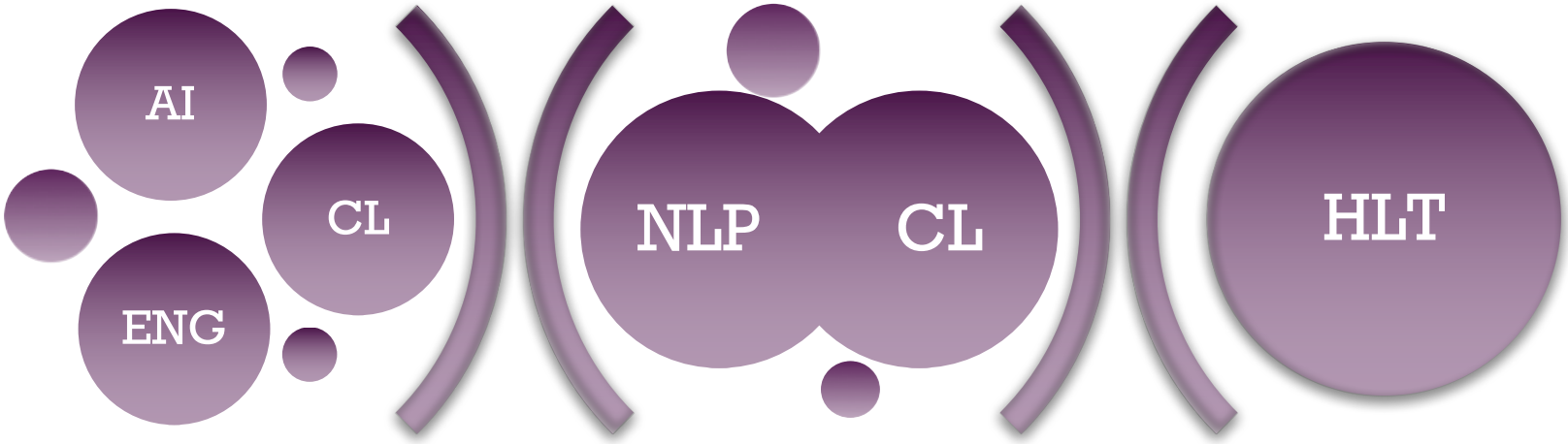
+ Some Literature

HLT2

- Ittoo, A.; L.M. Nguyen & A. van de Bosch (2016): Text analytics in industry: Challenges, desiderata and trends, *Computers in Industry* (2016), http://www.sciencedirect.com/science/article/pii/S0166361515300646
- Ledeneva, Y & G. Sidrov (2010): Recent Advances in Computational Linguistics, *Informatics* 34 (2010), 3-18.
- Kay, M (2006): A Life of Language, ACL Lifetime Achievement Award, *Computational Linguistics*, Volume 31, Number 4: 425-438.
- Cole, R. et al (1997): *Survey of the State of the Art in Human Language Technology*, Cambridge University Press & Giardini.

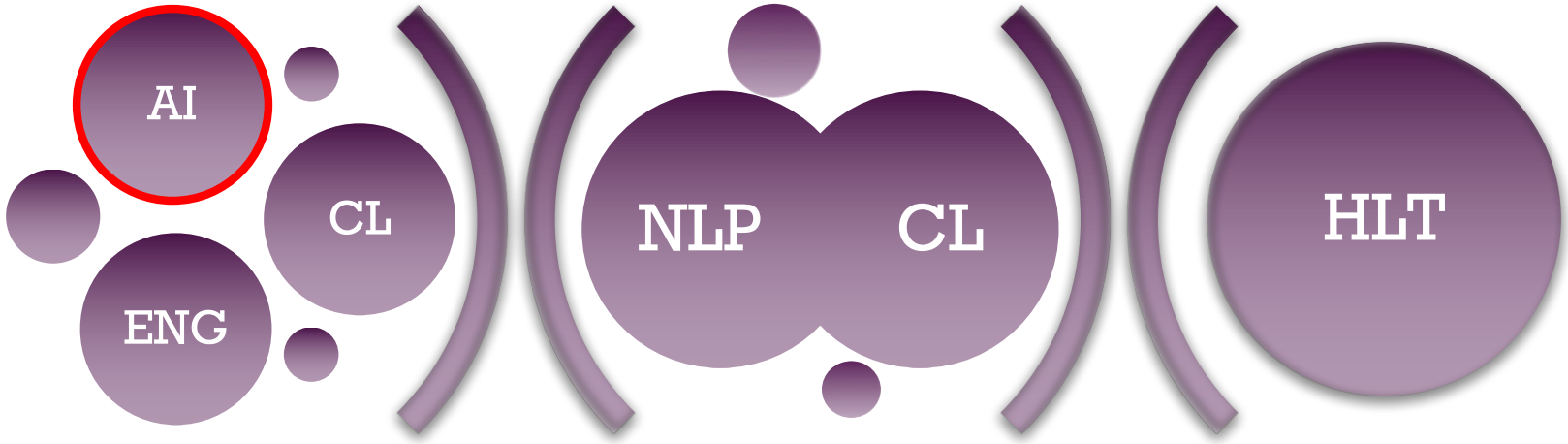
UCD DUBLIN

High-Level Timeline



→ ca. 1985 → ca. 2005 → 2017

High-Level Timeline



→ ca. 1985 → ca. 2005 → 2017

Early AI: ELIZA (1966)

HLT2

- ELIZA - models the behaviour of a psychiatrist (cf. Turing Test/Imitation Game)

- "The fundamental technical problems with which ELIZA is concerned are:
  - the identification of key words,
  - the discovery of minimal context,
  - the choice of appropriate transformations,
  - generation of responses in the absence of keywords, and
  - the provision of an ending capacity for ELIZA "scripts"."

Joseph Weizenbaum (1966): ELIZA: A Computer Program For the Study of Natural Language Communication Between Man and Machine, *Communications of the ACM*, Volume 9, Number 1.

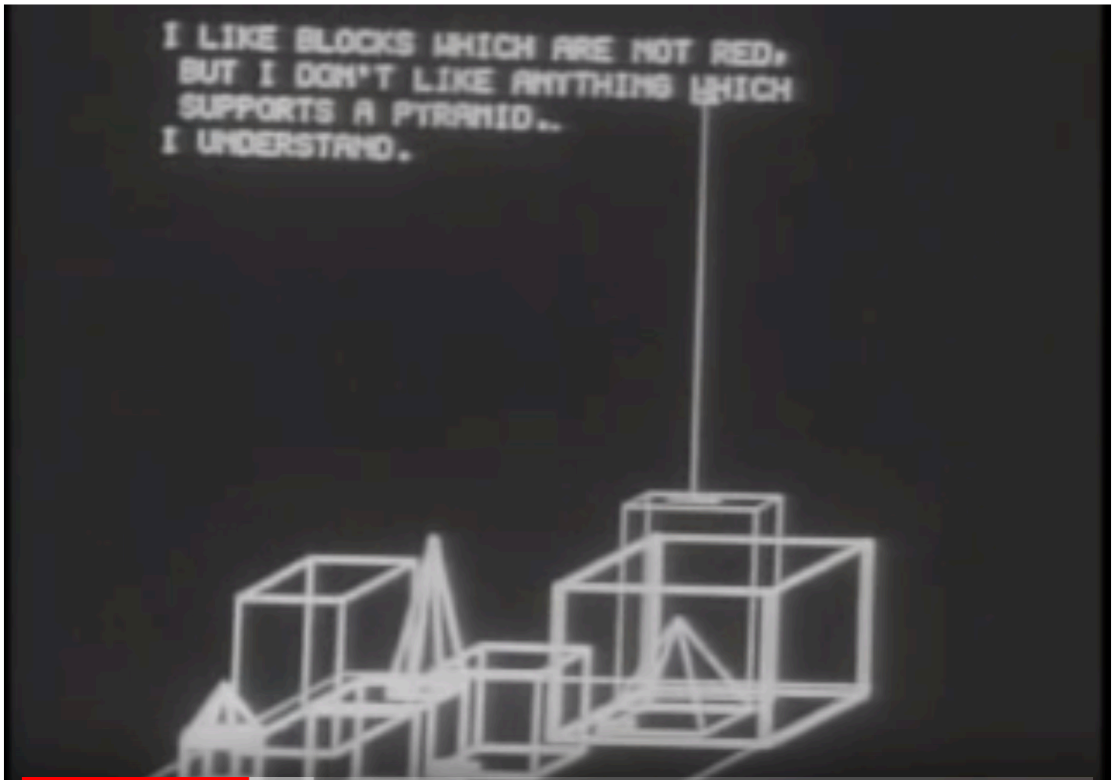
- See http://www.masswerk.at/elizabot/ - written by Norbert Landsteiner
- Some more recent developments: http://www.bbc.com/news/technology-27762088

UCD DUBLIN

Early AI: SHRDLU (1971/2)

HLT2

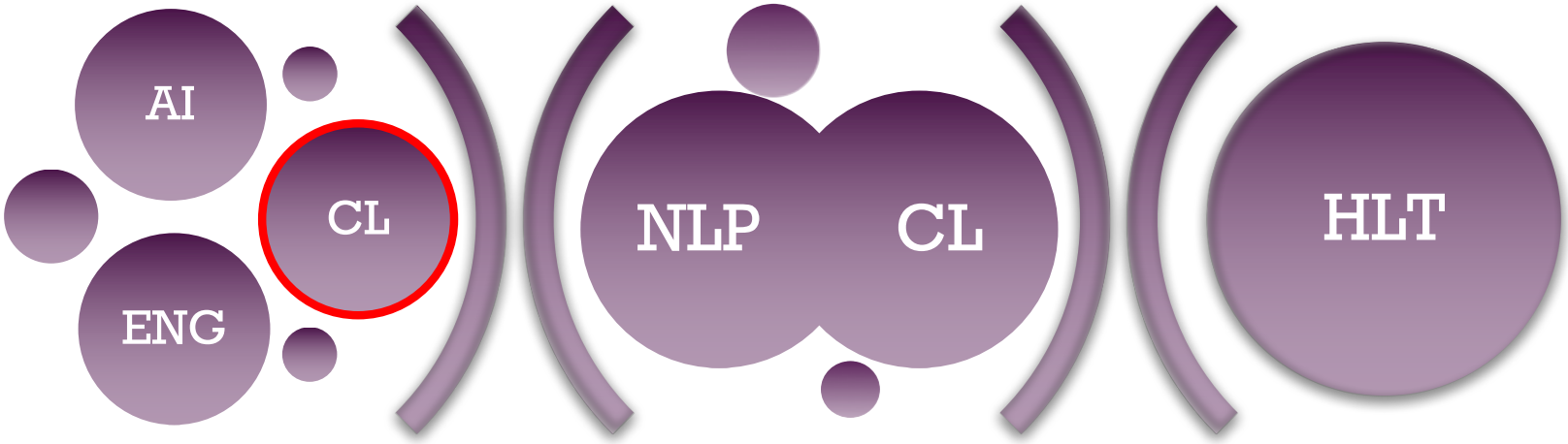
- SHRDLU – a language parser with natural language user interaction in a "blocks world" - Terry Winograd (1972): *Understanding Natural Language*, Academic Press.
- http://hci.stanford.edu/winograd/shrdlu/



- For a (silent) demonstration see: https://www.youtube.com/watch?v=bo4RvYJYOzI

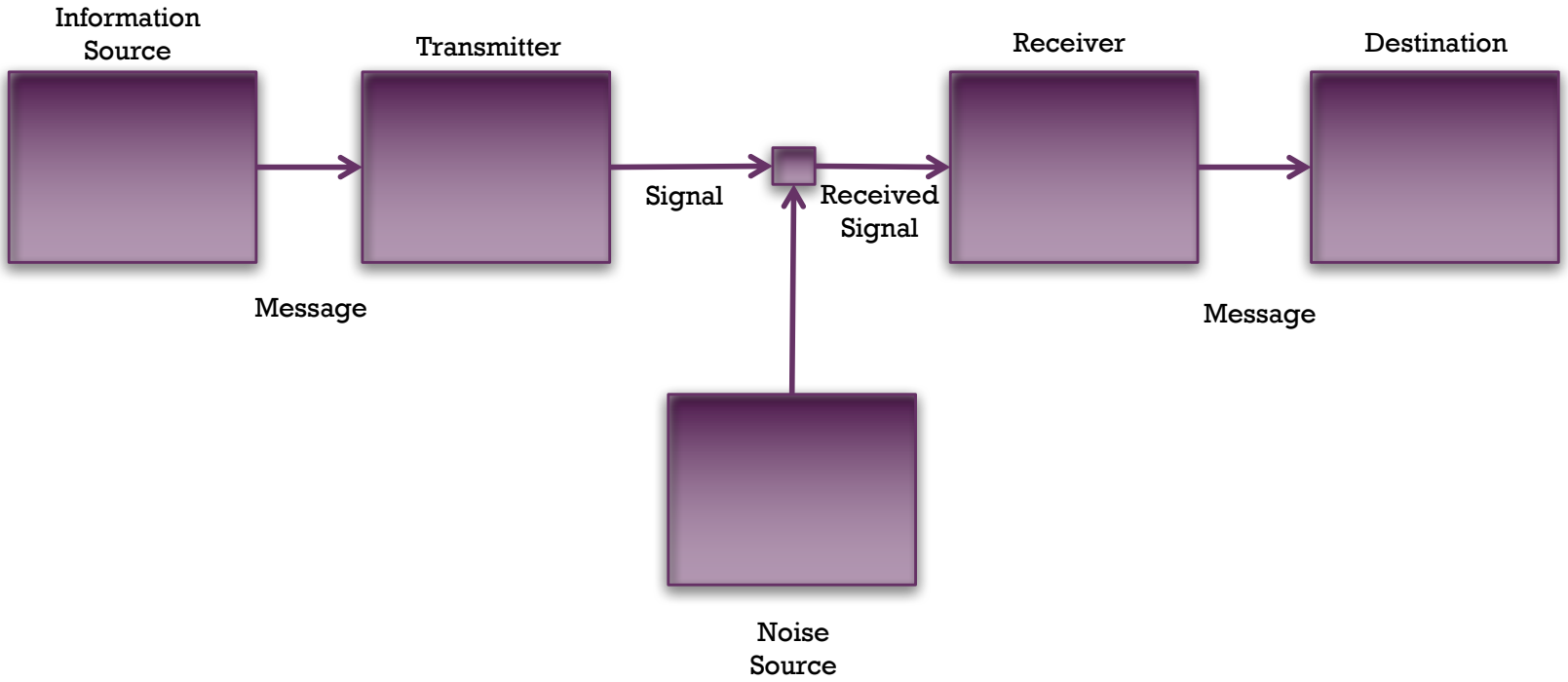
UCD DUBLIN

High-Level Timeline



→ ca. 1985 → ca. 2005 → 2017

Early CL: Machine Translation



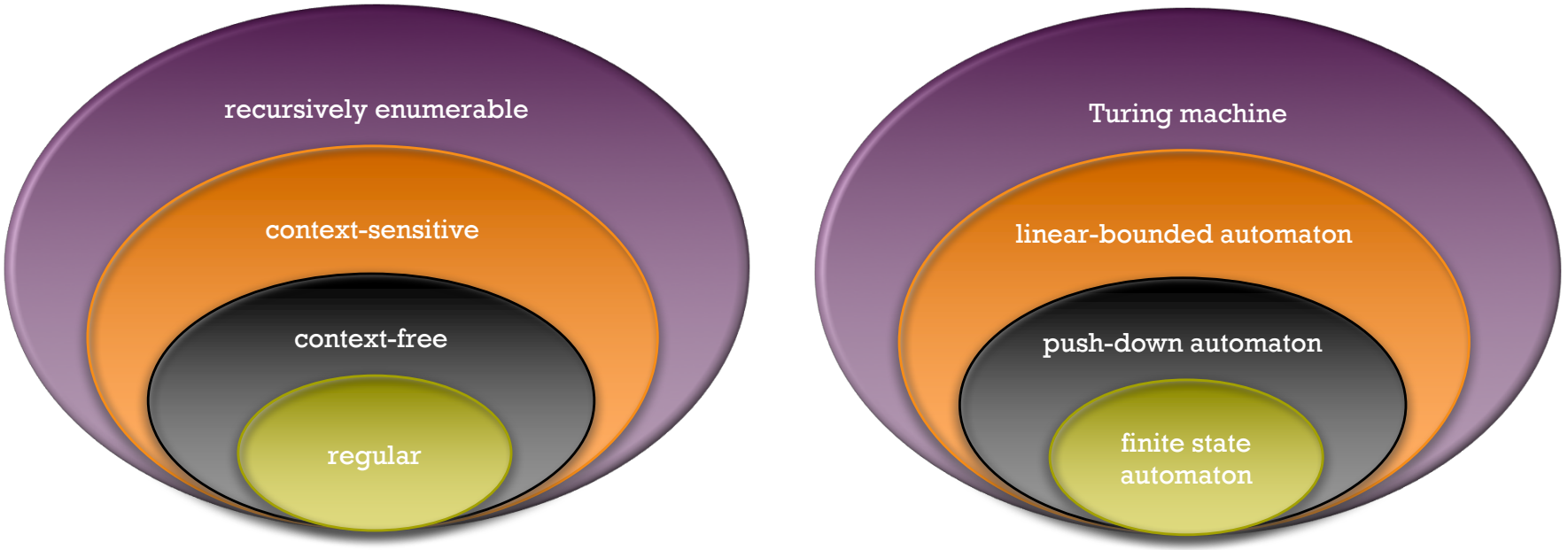
Shannon's General Communication System (Shannon, 1948)

- Shannon, C. & W. Weaver (1949): *The Mathematical Theory of Communication*
- Weaver, W. (1949): *Translation*
- http://www.mt-archive.info/Weaver-1949.pdf

Early CL: "Formalising" Grammar

HLT2

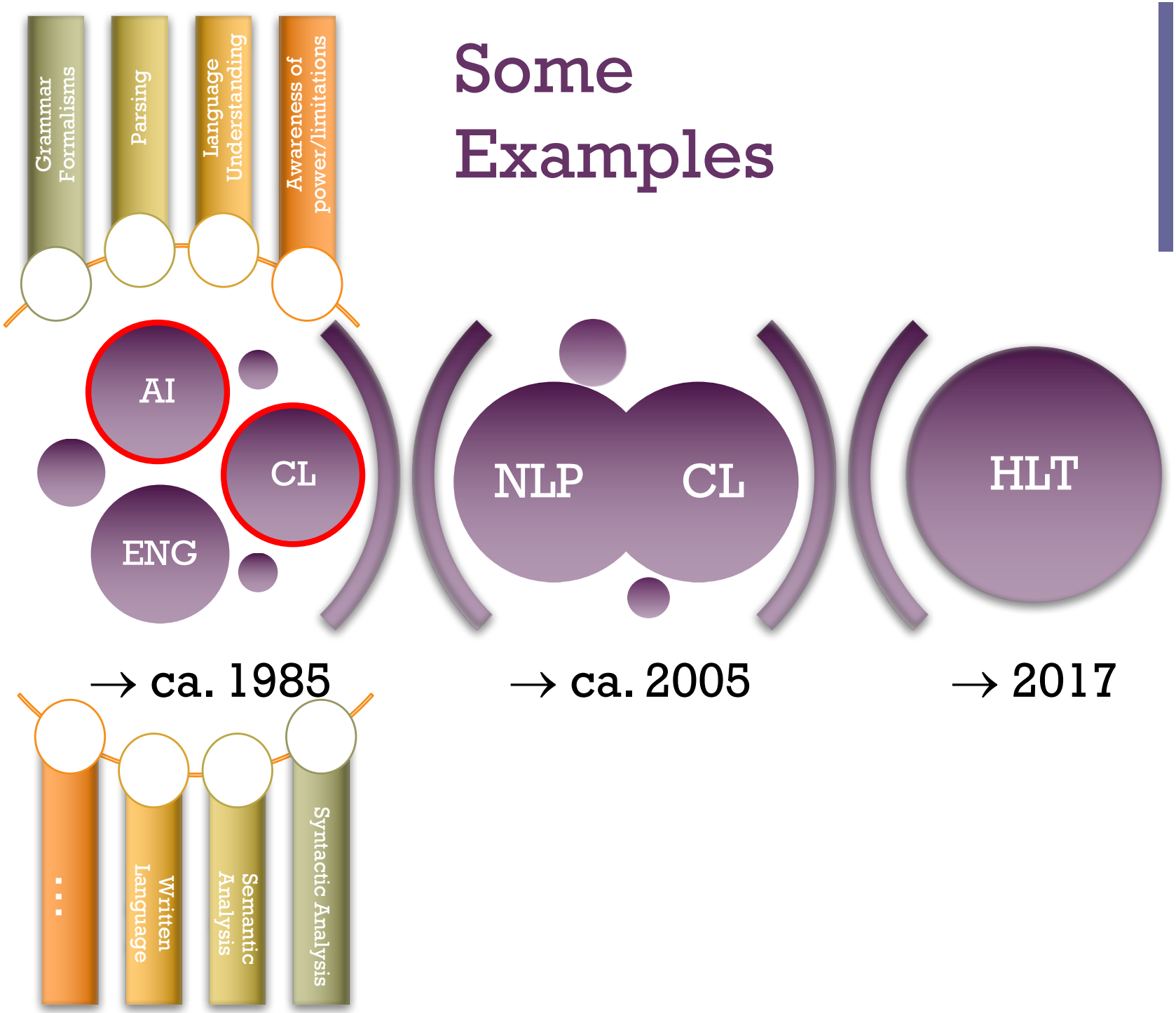
- Relating natural language syntax to formal language theory – mathematical models of language



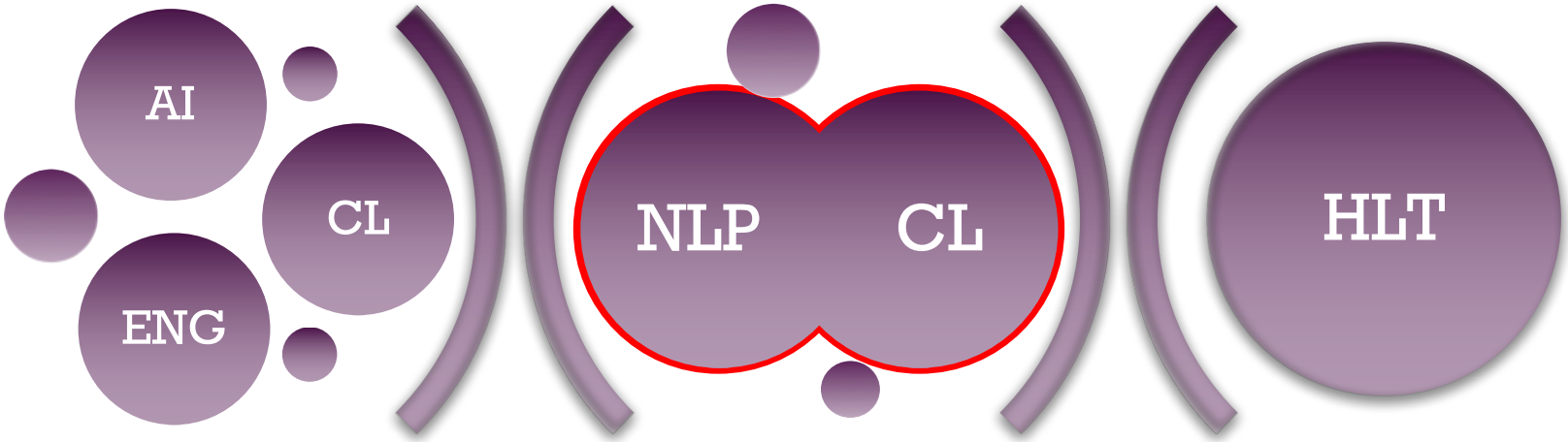
Chomsky Hierarchy

UCD DUBLIN

Some Examples



High-Level Timeline



→ ca. 1985 → ca. 2005 → 2017

+ Cole et al. (1997) Review

HLT2

- The study of human language technology is a multidisciplinary enterprise, requiring expertise in areas of linguistics, psychology, engineering and computer science.

- Creating machines that will interact with people in a graceful and natural way using language requires a deep understanding of the acoustic and symbolic structure of language (the domain of linguistics), and the mechanisms and strategies that people use to communicate with each other (the domain of psychology).

Cole et al. (1997:xi)

UCD DUBLIN

+ Cole et al. (1997) Review

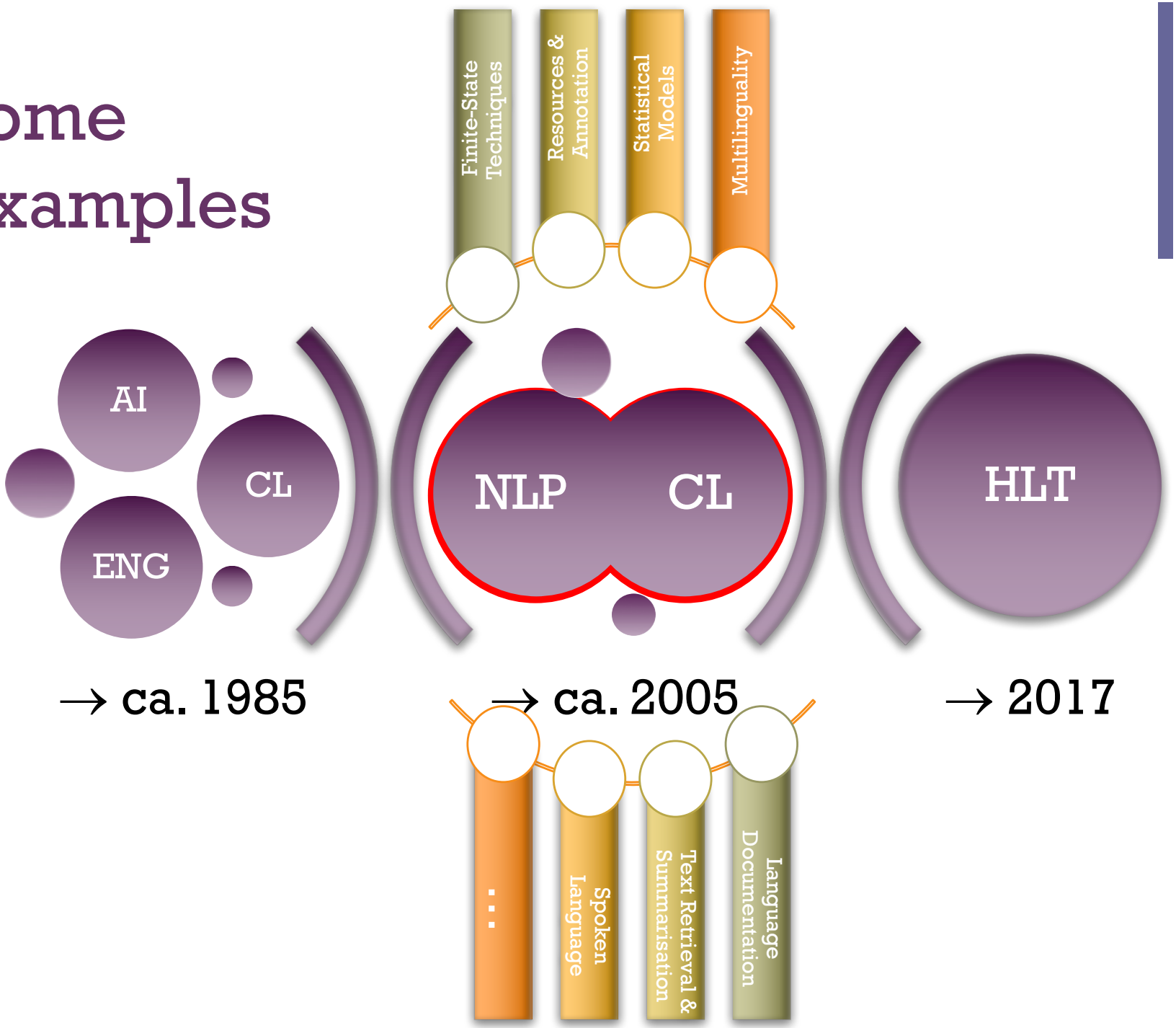
HLT2

- Given the remarkable ability of people to converse under adverse conditions, such as noisy social gatherings or band-limited communication channels, advances in signal processing are essential to produce robust systems (the domain of electrical engineering).
- Advances in computer science are needed to create the architectures and platforms needed to represent and utilize all of this knowledge.
- Collaboration among researchers in each of these areas is needed to create multimodal and multimedia systems that combine speech, facial cues and gestures both to improve language understanding and to produce more natural and intelligible speech by animated characters.

Cole et al. (1997:xi)

UCD DUBLIN

Some Examples



+ (E)Merging "Dichotomies"…

HLT2

- Technological vs Linguistic/Cognitive
- Procedural vs Declarative
- Competence vs Performance
- Knowledge-Based vs Data-Driven
- Small (but difficult) Problems vs Broad Coverage
- Algorithms vs Resources
- Methods vs Areas of Research

UCD DUBLIN

+ Cole et al. (1997) Review

HLT2

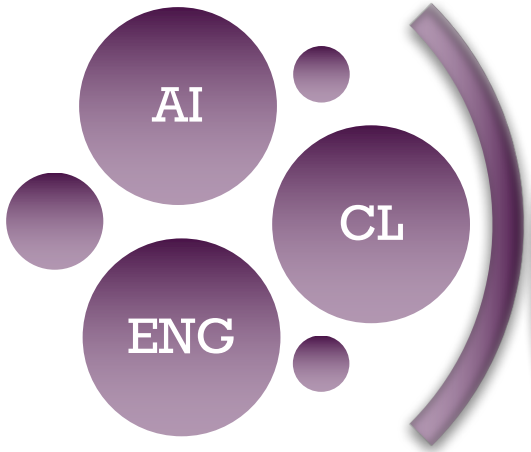
Looking forward:

- Benefits that can be expected from deploying language technology are a more effective usability of systems (enabling the user) and enhanced capabilities for people (empowering the user). The economic and social impact will be in terms of effciency and competitiveness for business, better educated citizens, and a more cohesive and sustainable society.

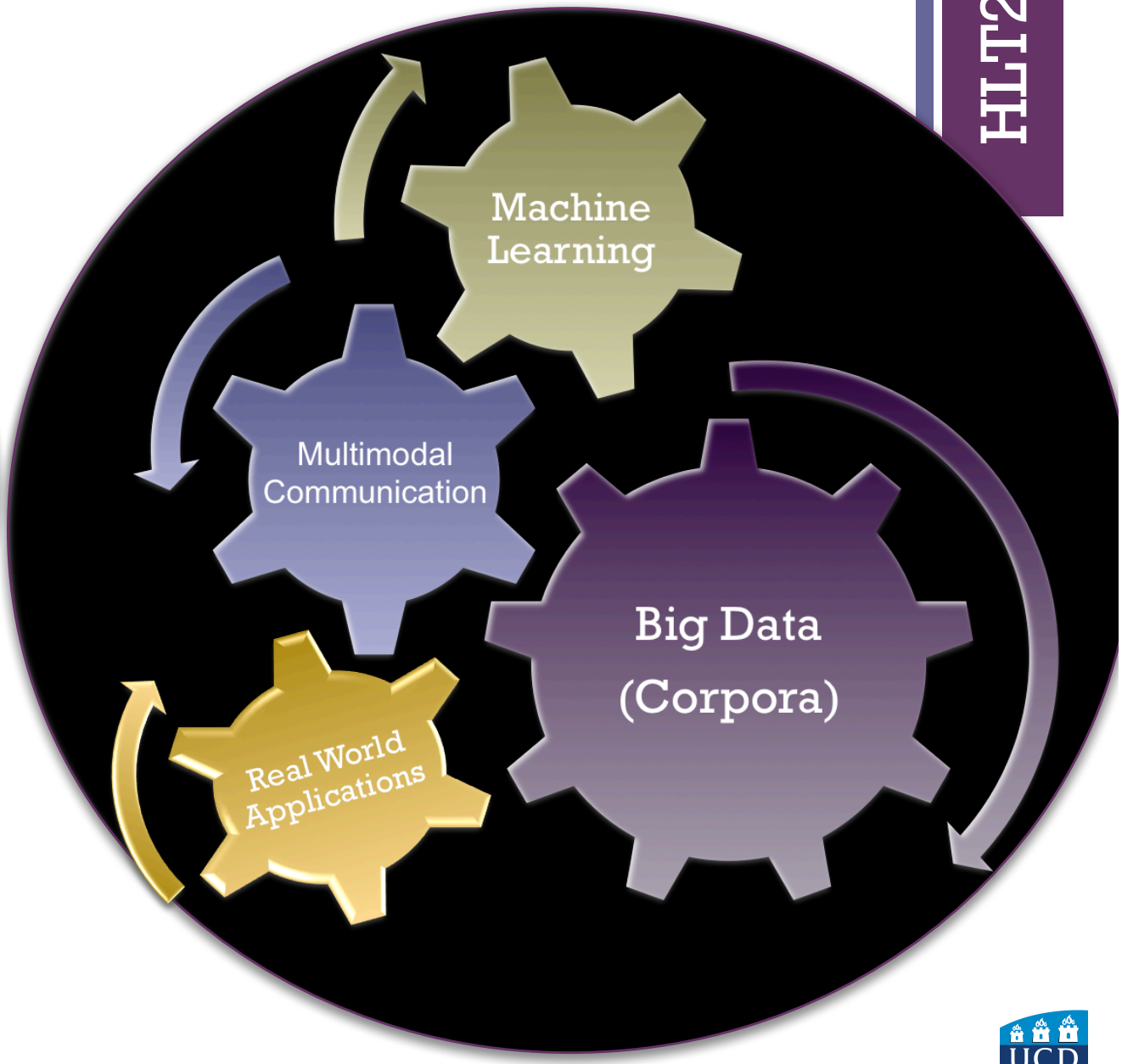
Cole et al. (1997:xvii)

UCD DUBLIN

Focus



→ ca. 1985



HLT2

UCD DUBLIN

+ Some Current HLT Applications

HLT2

- **Analysing and evaluating the task of automatic tweet generation: knowledge to business**
- "Text summarization techniques are applied to news articles, yielding ultra-concise summaries, which then serve as tweets. A potential application of these automatically generated tweets is in supporting the marketing and communication activities of companies. Therefore, it is important to assess the tweets' quality, especially taking into account the fact that they are created without human intervention.
- The tweets' quality is assessed in terms of 2 dimensions, viz:
  1. Interestingness: to what extent users would be interested in knowing more about the tweets' contents.
  2. Informativeness: whether the tweets accurately reflected the news articles from which they were generated."

E. Lloret, M. Palomar, Analysing and evaluating the task of automatic tweet generation: knowledge to business, Comput. Ind. (2015) according to Ittoo et al. (2016: 4)

UCD DUBLIN

+ Some Current HLT Applications

- **A methodology for traffic-related twitter messa… interpretation**
- "Automatically interpreting traffic-related Twitter … the Portuguese language. Interpretation in this case …ers to the transformation of the tweets' unstructured text…l contents into a more structured formalism, namely as RDF-triples.
- The system is deployed as a prototype for monitoring the truck fleet of a gas transportation and fuel transportation company. Several benefits are reported as a result of the system's application in these companies, such as cost reduction, more efficient fleet management, and improved customer satisfaction due to a more accurate prediction of delivery time."

Resource Description Framework

F.C. Albuquerque, M.A. Casanova, H. Lopes, L.R. Redlich, J.A.F. de Macedo, M. Lemos, M.T.M. de Carvalho, C. Renso, A methodology for traffic-related twitter messages interpretation, Comput. Ind. (2015) according to Ittoo et al (2016: 4)

+ Some Current HLT Applications

HLT2

- **Integrating a semantic-based retrieval agent into case-based reasoning systems: a case study of an online bookstore**
- "employs Case-Based-Reasoning (CBR) to enhance the search experience of end users in business-to-consumer websites. Specifically, it accepts queries as inputs, which could be formulated either as natural language questions or as keywords. It then consults a case base consisting of target problems (e.g. questions) and their corresponding solutions (answers). The best matching solution in the case is then returned as answer to the input query."

J.W. Chang, M.C. Lee, T.I. Wang, Integrating a semantic-based retrieval agent into case-based reasoning systems: a case study of an online bookstore, Comput. Ind. (2015) according to Ittoo et al (2016: 6)

UCD DUBLIN

+ Some Current HLT Applications

HLT2

- **A distributional approach to open questions in market research**
- "system called The Klugator Engine (TKE), for analyzing responses to open-ended survey questions. The responses are expressed in freely formed natural language texts in English and German. The proposed system was developed in a collaborative effort between academic and industrial partners. The system is already in commercial use as the core engine of the Rogator Text Clustering Solution (RogTCS), distributed by German market research company Rogator AG."

P. Greiner, S. Evert, F. Baigger, B. Lang, A distributional approach to open questions in market research, Comput. Ind. (2015) according to Ittoo et al (2016: 6)

UCD DUBLIN

+ Some Current HLT Applications

HLT2

- IBM Watson – a technology platform that uses natural language processing and machine learning to reveal insights from large amounts of unstructured data

- http://www.ibm.com/smarterplanet/us/en/ibmwatson/what-is-watson.html

- Winning at Jeopardy: https://www.youtube.com/watch?v=WFR3lOm_xhE

UCD DUBLIN

NLP Industry…

HLT2

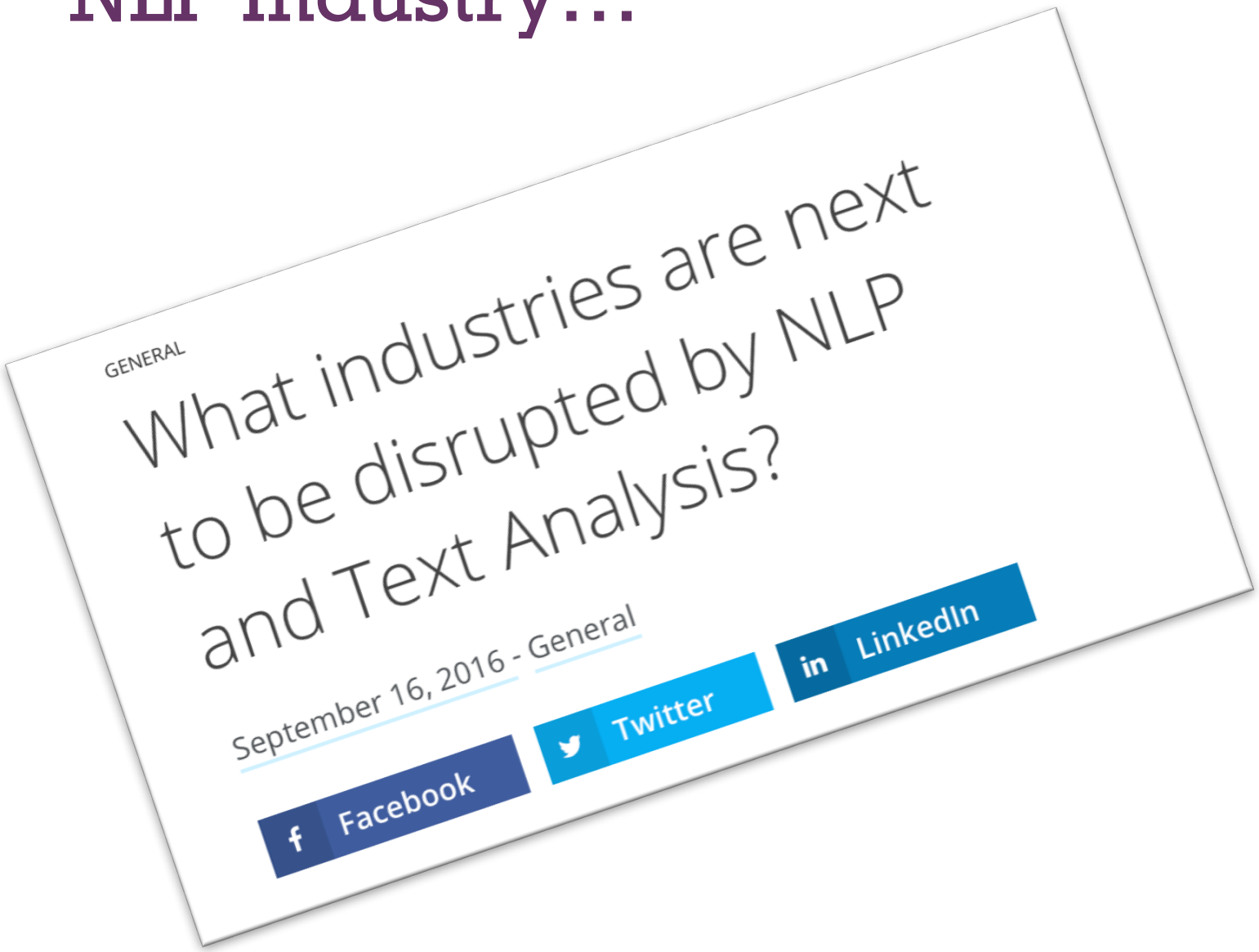
"The Natural Language Processing (NLP) market size is estimated to grow from USD 7.63 Billion in 2016 to USD 16.07 Billion by 2021, at a Compound Annual Growth Rate (CAGR) of 16.1%."

marketsandmarkets.com, July 2016 (Report Code: TC , 3492)

A **huge** part of AI

UCD DUBLIN

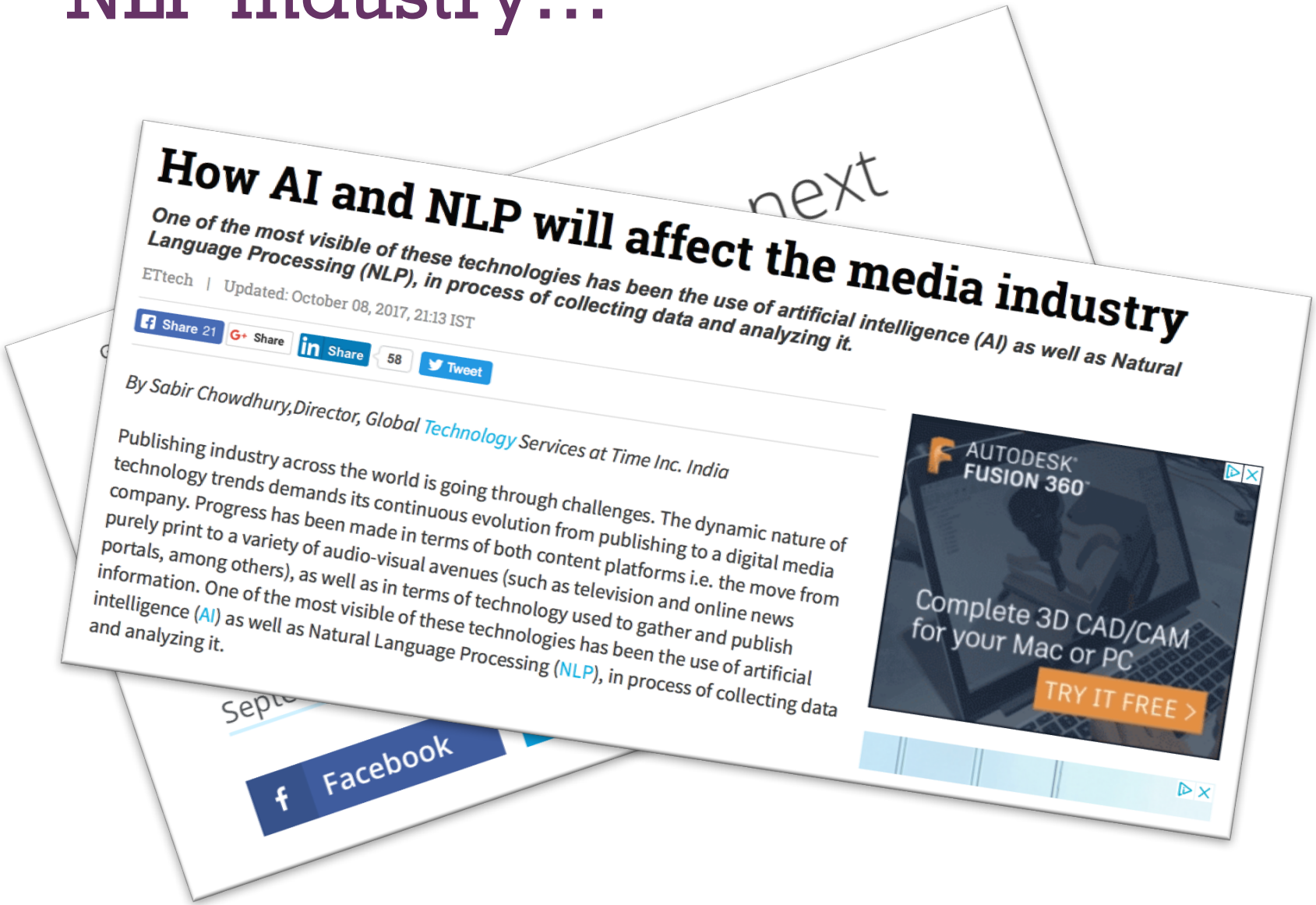
NLP Industry…



http://blog.aylien.com/nlp-text-analysis-insurance-legal-customer-service/

NLP Industry…

HLT2



How AI and NLP will affect the media industry

One of the most visible of these technologies has been the use of artificial intelligence (AI) as well as Natural Language Processing (NLP), in process of collecting data and analyzing it.

ETtech | Updated: October 08, 2017, 21:13 IST

Share 21 Share Share 58 Tweet

By Sabir Chowdhury,Director, Global Technology Services at Time Inc. India

Publishing industry across the world is going through challenges. The dynamic nature of technology trends demands its continuous evolution from publishing to a digital media company. Progress has been made in terms of both content platforms i.e. the move from purely print to a variety of audio-visual avenues (such as television and online news portals, among others), as well as in terms of technology used to gather and publish information. One of the most visible of these technologies has been the use of artificial intelligence (AI) as well as Natural Language Processing (NLP), in process of collecting data and analyzing it.

next

Facebook

https://tech.economictimes.indiatimes.com/news/technology/how-ai-and-nlp-will-affect-the-media-industry/60966371

UCD DUBLIN

NLP Industry…

HLT2

https://ie.linkedin.com/jobs/natural-language-processing-jobs

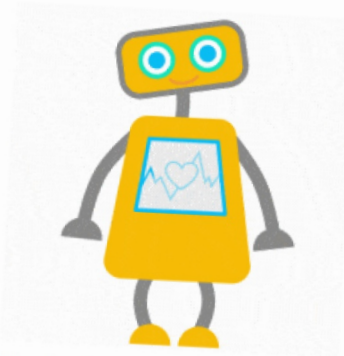
NLP Industry…

HLT2

To Your Health

'The Woebot will see you now' — the rise of chatbot therapy

By Amy Ellis Nutt December 3, 2017



Woebot is always available and will never judge. (

My therapist wanted to explain a few things during our first online session:

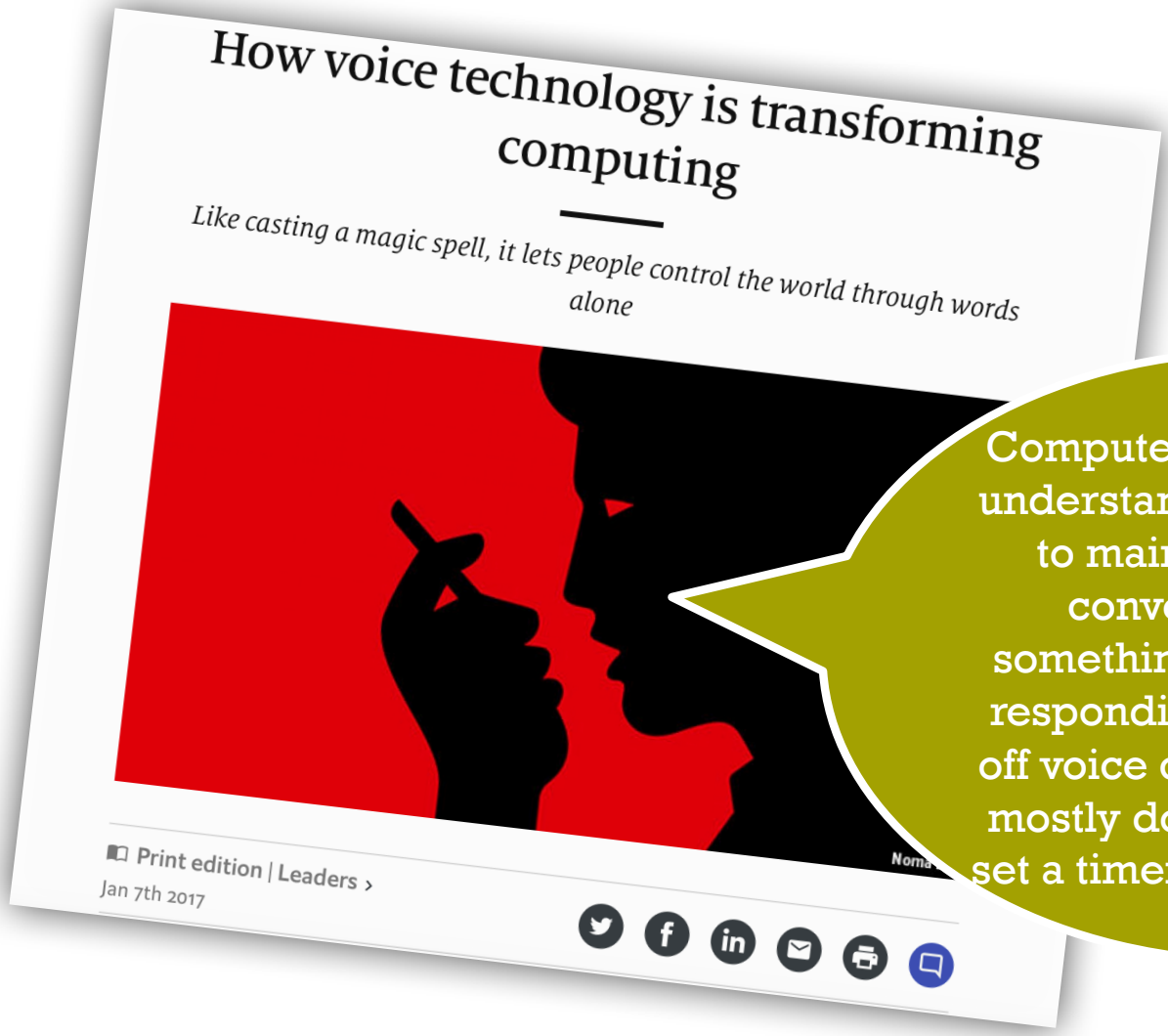
"I'm going to check in with you at random times. If you can't respond straight away, don't sweat it. Just come back to me when you're ready. I'll check in daily."

Alison Darcy CEO Woebot Labs and formally UCD Psychology

https://www.washingtonpost.com/news/to-your-health/wp/2017/12/03/the-woebot-will-see-you-now-the-rise-of-chatbot-therapy/?utm_term=.1c5b400d32db

Speech Industry…

HLT2



Computers must be able to understand context in order to maintain a coherent conversation about something, rather than just responding to simple, one-off voice commands, as they mostly do today ("Hey, Siri, set a timer for ten minutes").

https://www.economist.com/news/leaders/21713836-casting-magic-spell-it-lets-people-control-world-through-words-alone-how-voice

UCD DUBLIN

+ Speech Industry…

BIG DATA

The future of voice synthesis after Google WaveNet debut

Dr. Matthew Aylett Tue 13 Sep 2016 1.03pm

3 26 1 9 3 37 78 SHARES

Dr Matthew Aylett is the Chief Scientific Officer and co-founder at CereProc, which develops voice synthesis systems primarily for the healthcare field. In the light of last week's announcements from Baidu/NVidia and from Google, he takes a look at the challenges and motivations involved in reproducing the human voice artificially...

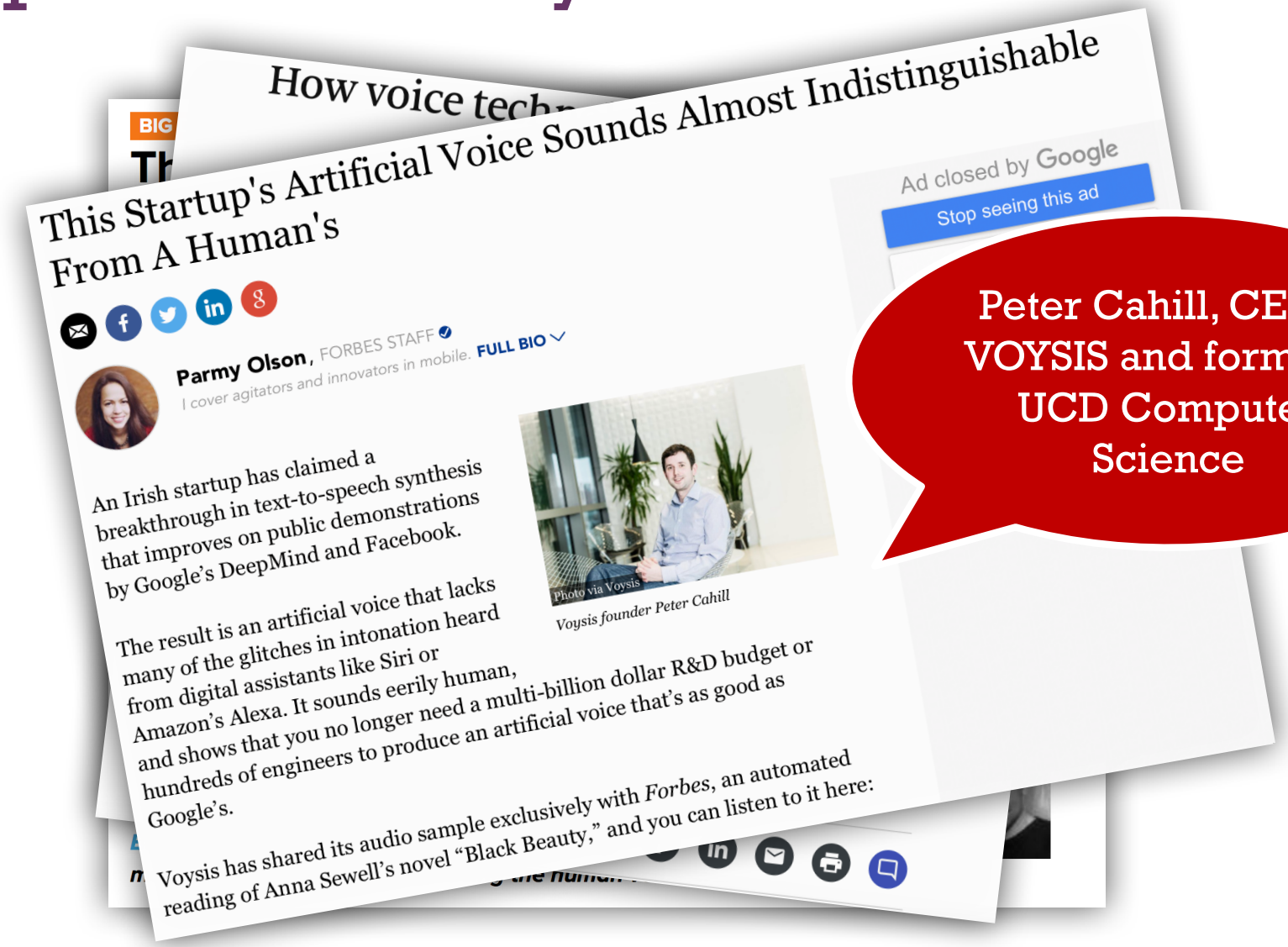
When Google first took Siri out of the box they didn't say "Nice recognition, interesting application, cool array microphones" – they said "Where are our Ads?"Google's business model almost entirely depends on selling Ads. Siri replaces the browser and returns control to Apple.

https://thestack.com/big-data/2016/09/13/the-future-of-voice-synthesis-after-google-wavenet-debut/

UCD DUBLIN

+ Speech Industry…

HLT2



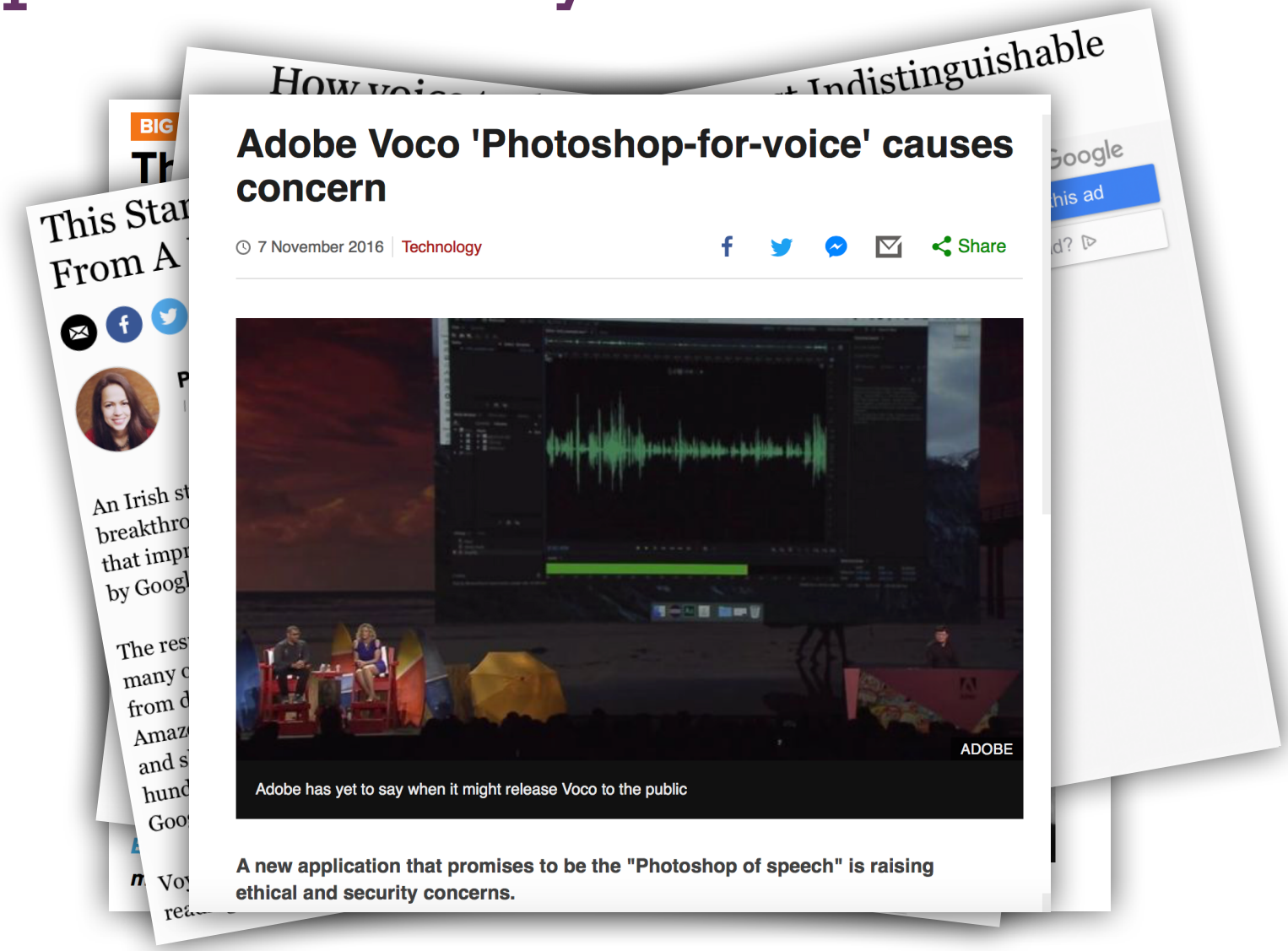
Peter Cahill, CEO of VOYSIS and formerly UCD Computer Science

https://www.forbes.com/sites/parmyolson/2017/11/03/this-startups-artificial-voice-sounds-almost-indistinguishable-from-a-humans/#12fed267388c

UCD DUBLIN

+ Speech Industry…

HLT2



http://www.bbc.com/news/technology-37899902

UCD DUBLIN

Module Outline

HLT2

- HLT 1: Introduction
- HLT 2: Evolution of Language Technologies
- HLT 3: Workshop 1 (Monday & Wednesday)
- HLT 4: Word-forms, wordforms and word forms (Morphology)
- HLT 5: Colourless green ideas sleep furiously (Syntax, Semantics)
- HLT 6: Grammars, Rules & Parsers (Syntactic Parsing)
- HLT 7: Workshop 2 (Wednesday & Monday)
- HLT 8: Boundaries, Tokens & Corpora (Text Processing & Analytics)
- HLT 9: Same but Different (Similarity & Sentiment Analysis)
- HLT 10: Workshop 3 (Wednesday & Monday)

UCD DUBLIN

Module Outline

HLT2

- HLT 11: Sounds, Symbols & Features (Phonetics & Phonology)
- Break
- HLT 12: Cats, Flys & Tomatos? (Computational Morphology)
- HLT 13: Twas brillig and the slithy toves (Computational Phonology)
- HLT 14: Language Modelling
- HLT 15: Workshop 4 (Wednesday & Monday)
- HLT 16 Speech Recognition
- HLT 17: Speech Synthesis
- HLT 18: Applications

UCD DUBLIN

Jupyter Notebook & NLTK

HLT2

- For Monday, please install Jupyter Notebook and NLTK by following the instructions on csmoodle
- If you have any difficulties with the installation, please contact me by email
- Please bring a laptop on Monday

UCD DUBLIN