Using KOS for Artificial Intelligence Applications

Thursday, 12 September 2019, Oslo
19th NKOS Workshop

Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony
j_ven_eman@accessinn.com
www.accessinn.com
+1.505.998.0800
Albuquerque, NM USA

Access Innovations, Inc.

The Science behind the Semantics™

www.accessinn.com

Summary

- AI/ML/DL hold promise for “Content”
- But big, headline grabbing failures
- Costs can run to the billions
- Choose carefully
- Choose narrowly
- Focus on improving content for customer utility and process workflow improvements

Some of Our Current Clients

American Association for Cancer Research

Reader's Digest

DTIC Online
Information for the Defense Community

RITA
Research and Innovative Technology Administration
National Transportation Library

IEEE

The Weather Channel

ANSI
American National Standards Institute

IOP

PLoS
PUBLIC LIBRARY of SCIENCE

ProQuest CSA

NIEHS

AIP
American Institute of Physics

ChemADVISOR, Inc.

NATIONAL SECURITY AGENCY

asug

ASCO
AMERICAN SOCIETY OF CLINICAL ONCOLOGY

HARVARD BUSINESS PUBLISHING

ACS
Chemistry for Life®

American Water Works Association
The Authoritative Resource on Safe Water

ConsumersUnion.org
Nonprofit Publisher of Consumer Reports

ASCE
American Society of Civil Engineers

npr

SPIE
Connecting minds. Advancing light.
SPIE is an international society advancing light-based research

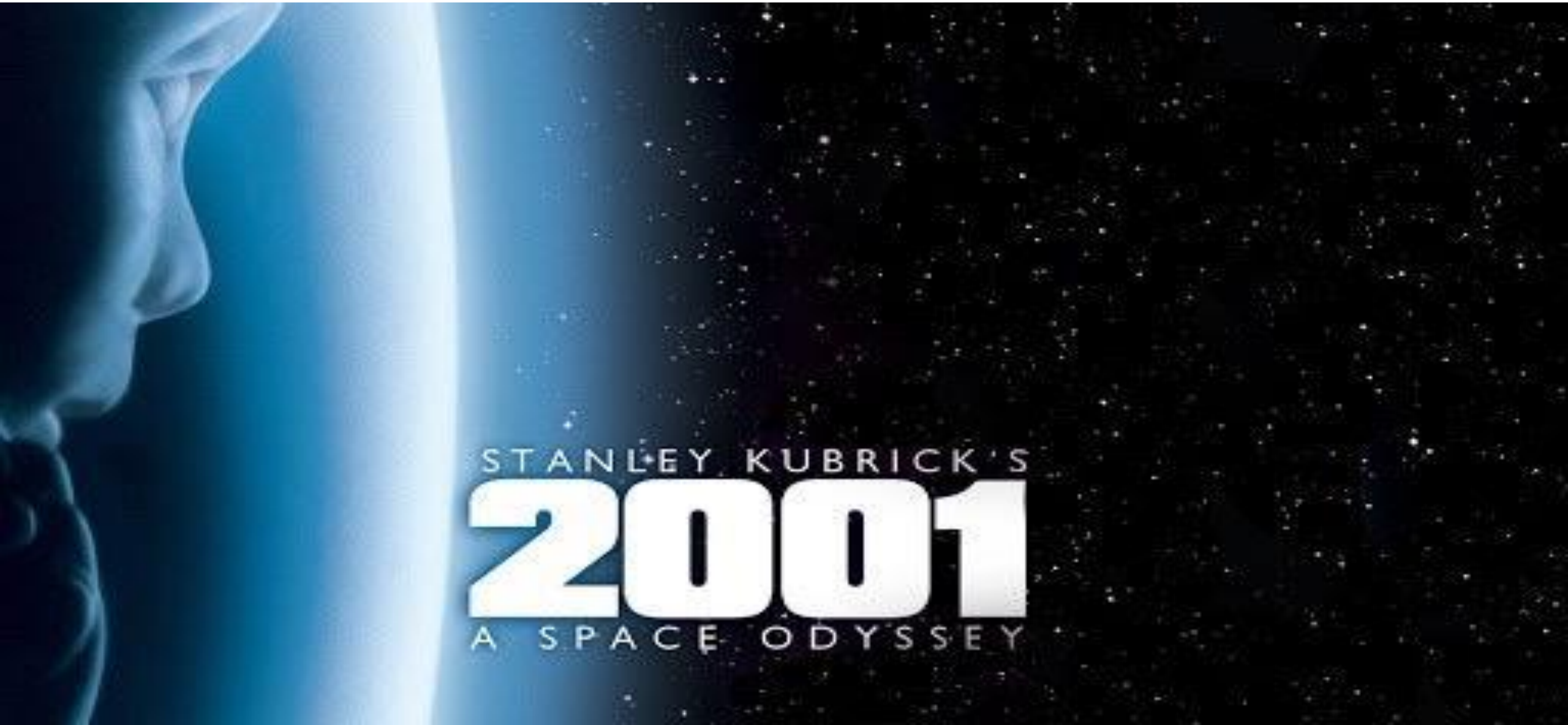
NLB | National Library Board Singapore

UNITED STATES DEPARTMENT OF LABOR

Access Innovations Inc.®
maker of Data Harmony taxonomy software

AI – It'll Be Awesome!

1968 – what we expected/feared AI would to become!



Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL)

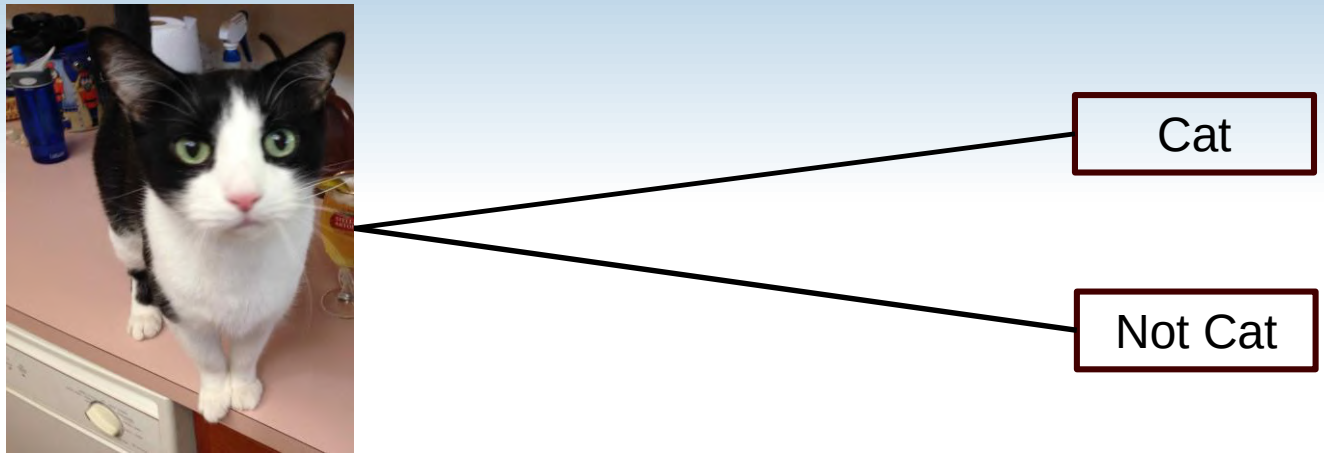
- “AI involves machines that can perform tasks that are characteristic of human intelligence.” (Calum McClelland, 4 Dec 2017)
- Give the computer lots of data - it gives you results...
- Intelligence demonstrated by machines ... The study of intelligent agents ... Machines mimicking cognitive functions of humans (a la Wikipedia)

Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL)

- At its simplest – conditional probabilities
  - P(A) means "Probability Of Event A”
  - P(B|A) means “Probability of event B given Event A”

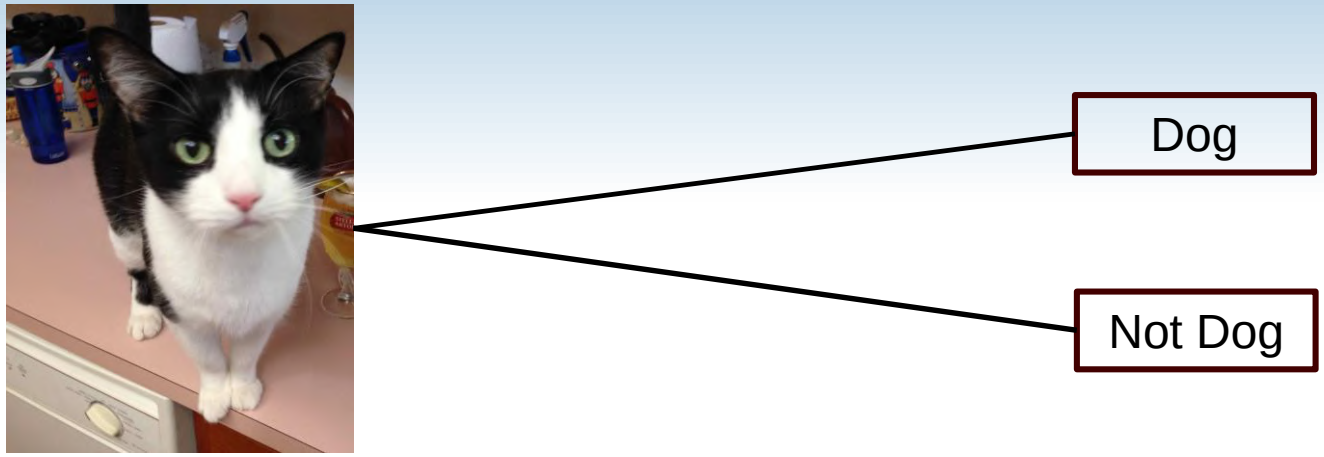
Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL)

- Binary classification



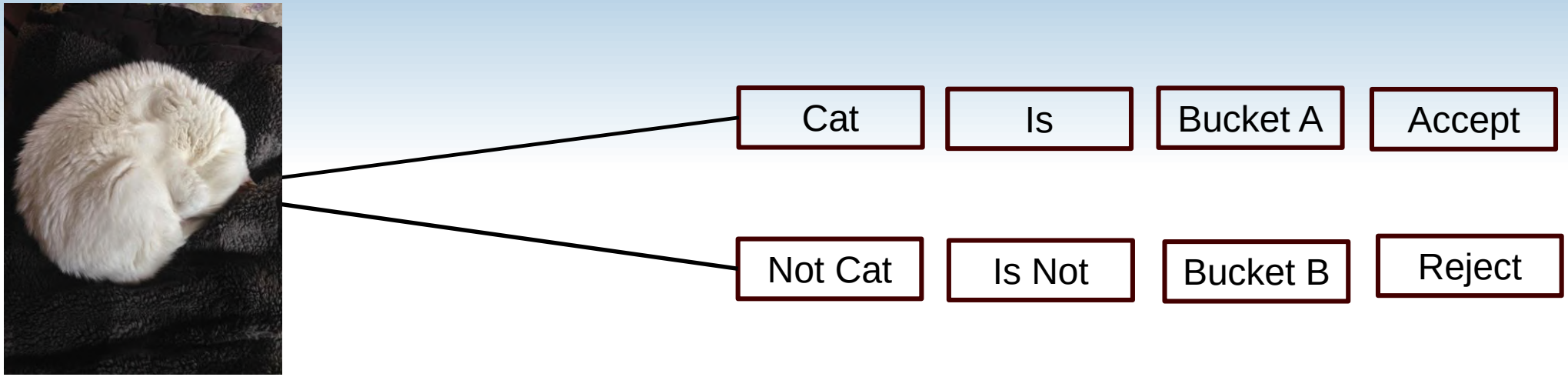
Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL)

- But always keep in mind…



Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL)

- And then after lots of training…



Supporting theories

- Co-occurrence
- Bayesian models
- Vector Analysis
- SMART (Gerard “Gerry” Salton)
- Decision trees
- Cognitive Computing
- Rule bases
- Machine Learning
- Clustering

Sibling technologies

- Automatic Translation
- Automated Language Processing (ALP)
- Natural Language Processing (NLP)
- Automated Data Processing (ADP)
- Robotics
- Search – Query, Word, and Phrase Parsing
- Dictionary look-ups and replies

Supporting technologies

- OCR
- Word-to-text Processing
  - Audio to text
- Dictionaries
- Knowledge Organization Systems (KOS)
- Word and Phrase Parsers
- Search Software

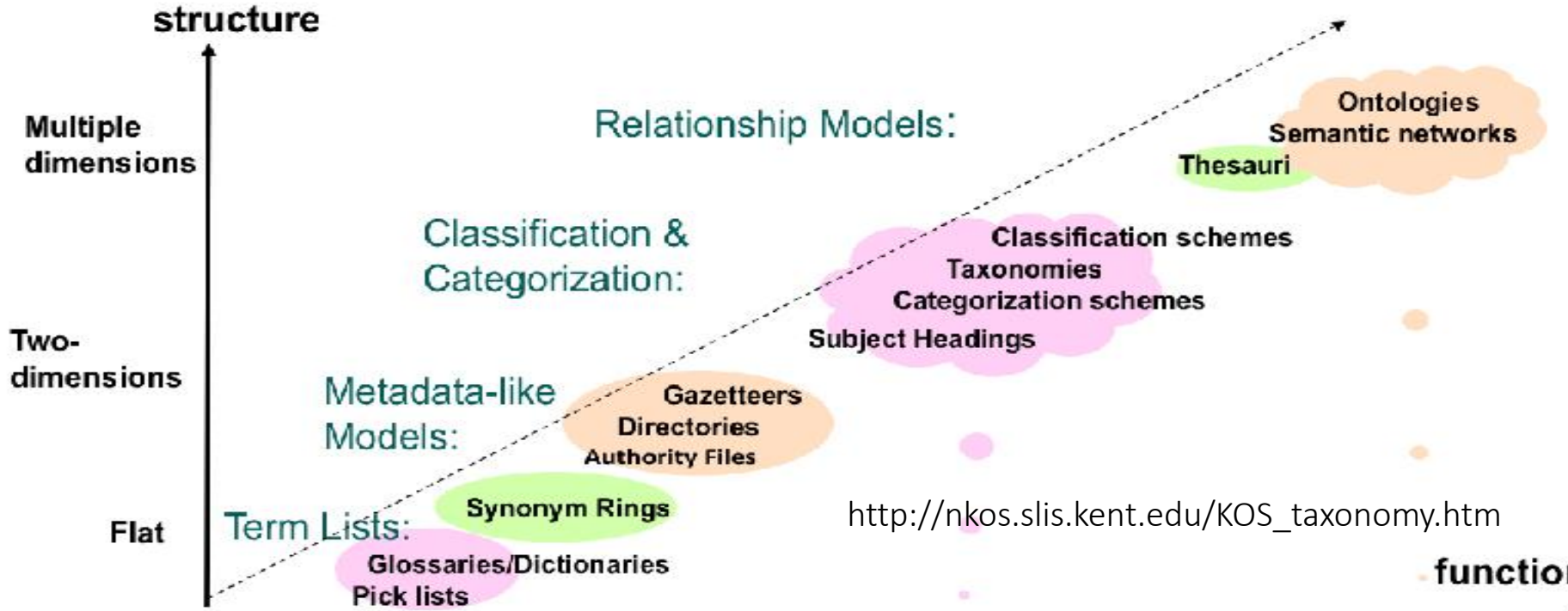
Supporting KOS technologies

- Term Lists
  - Authority Lists
  - Glossaries
  - Gazetteers
  - Dictionaries
- Classification and Categorization
  - Subject Headings
  - Classification Schemes, Taxonomies, and Categorization Schemes
- Relationship Groups
  - Thesauri
  - Semantic Networks
  - Ontologies

After Marcia Zeng and Gail Hodge
www.clir.org/pubs/abstract/pub91abst.html
http://nkos.slis.kent.edu/KOS_taxonomy.htm

Various types of KOS

Zeng 2008 p.161

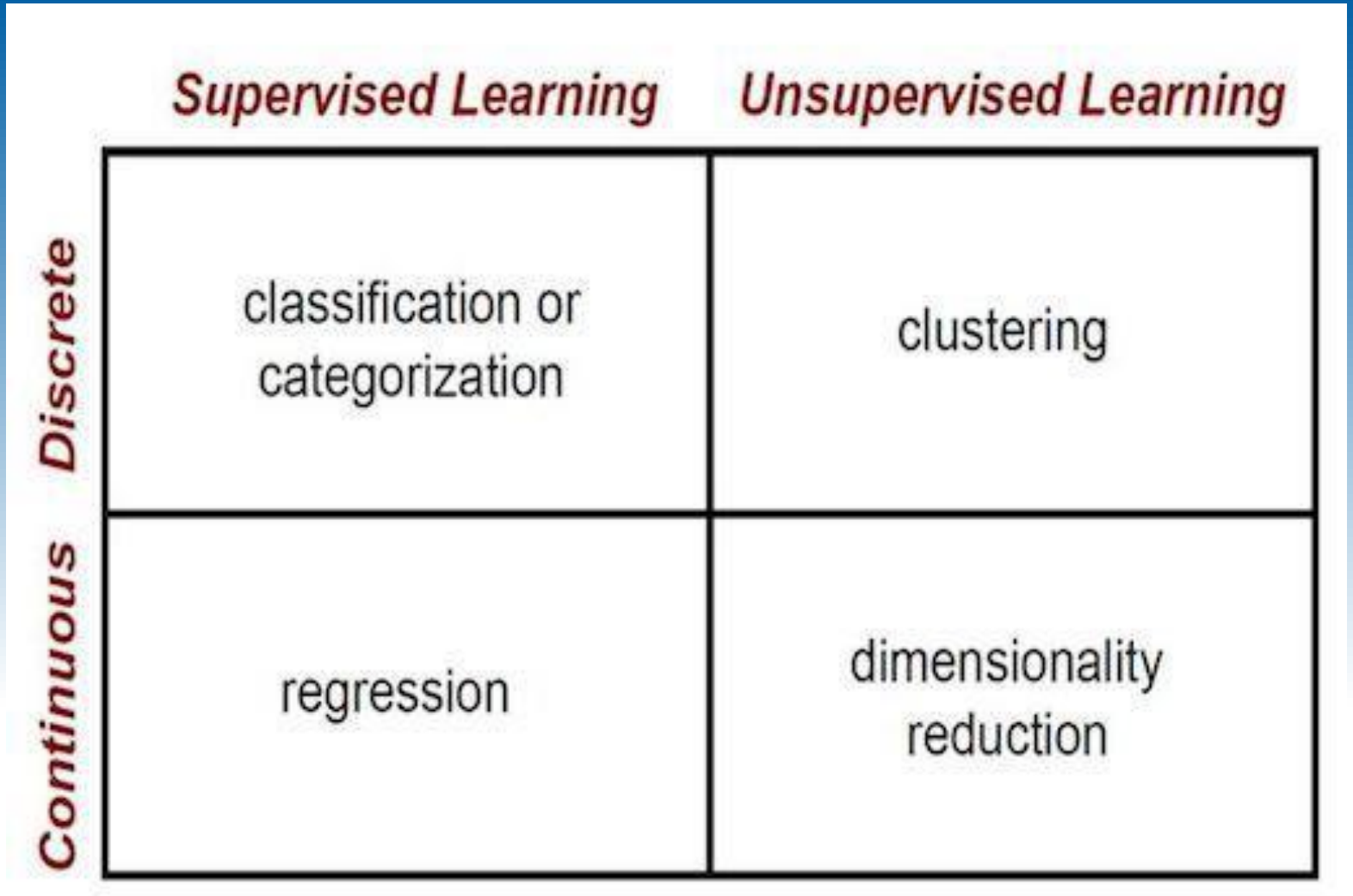


http://nkos.slis.kent.edu/KOS_taxonomy.htm

Major functions							
eliminating ambiguity	xxx		xxx	xx	xxxx	xx
controlling synonyms		xxxx	xxx	xx	xxxx	xx
establishing relationships: hierarchical			x	xxxx	xxx	xxx
establishing relationships: associative					xxxx	xxxxx
presenting properties						xxxxx

(2) Source: Zeng, Marcia Lei. "Knowledge Organization Systems (KOS)". *Knowledge Organization*, 35(2008)No.2/No.3

Figure 1 shows the types of knowledge organization systems (KOS), arranged according to the degree of controls introduced (from natural language to controlled language) and the strength of their semantic structure (from weakly structured to strongly structured), corresponding to the major functions of KOS. It represents a visualized summarization of the Taxonomy of Knowledge Organization Sources/Systems (http://nkos.slis.kent.edu/KOS_taxonomy.htm) adopted by the NKOS group based on Gail Hodge's article on KOS (www.clir.org/pubs/abstract/pub91abst.html).



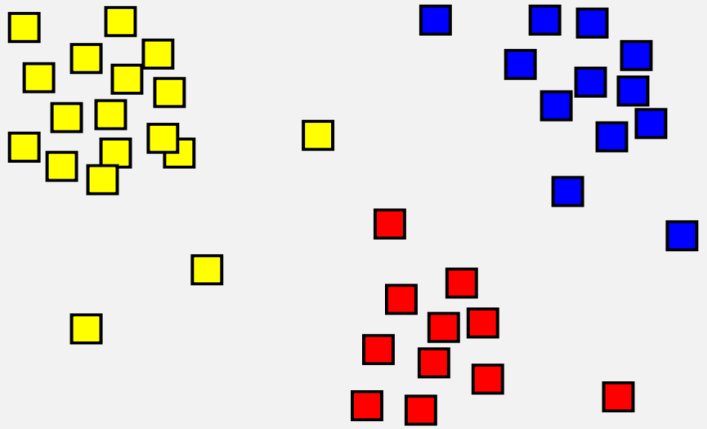
https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

Machine Learning

(relevant sub-disciplines)

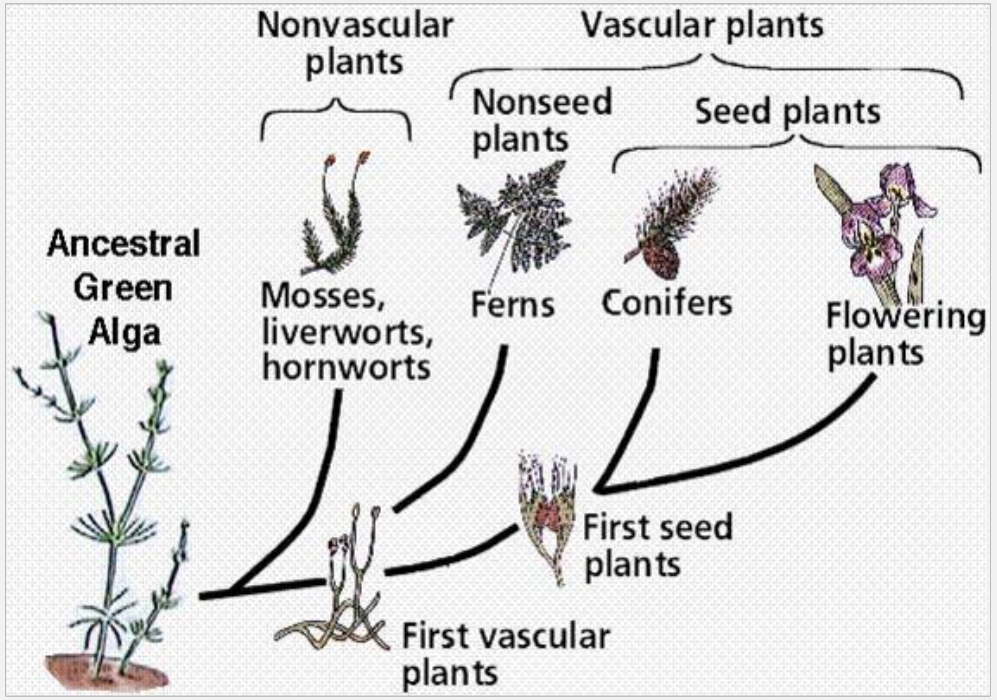
Clustering

Naïve groupings of things according to *computed* similarity

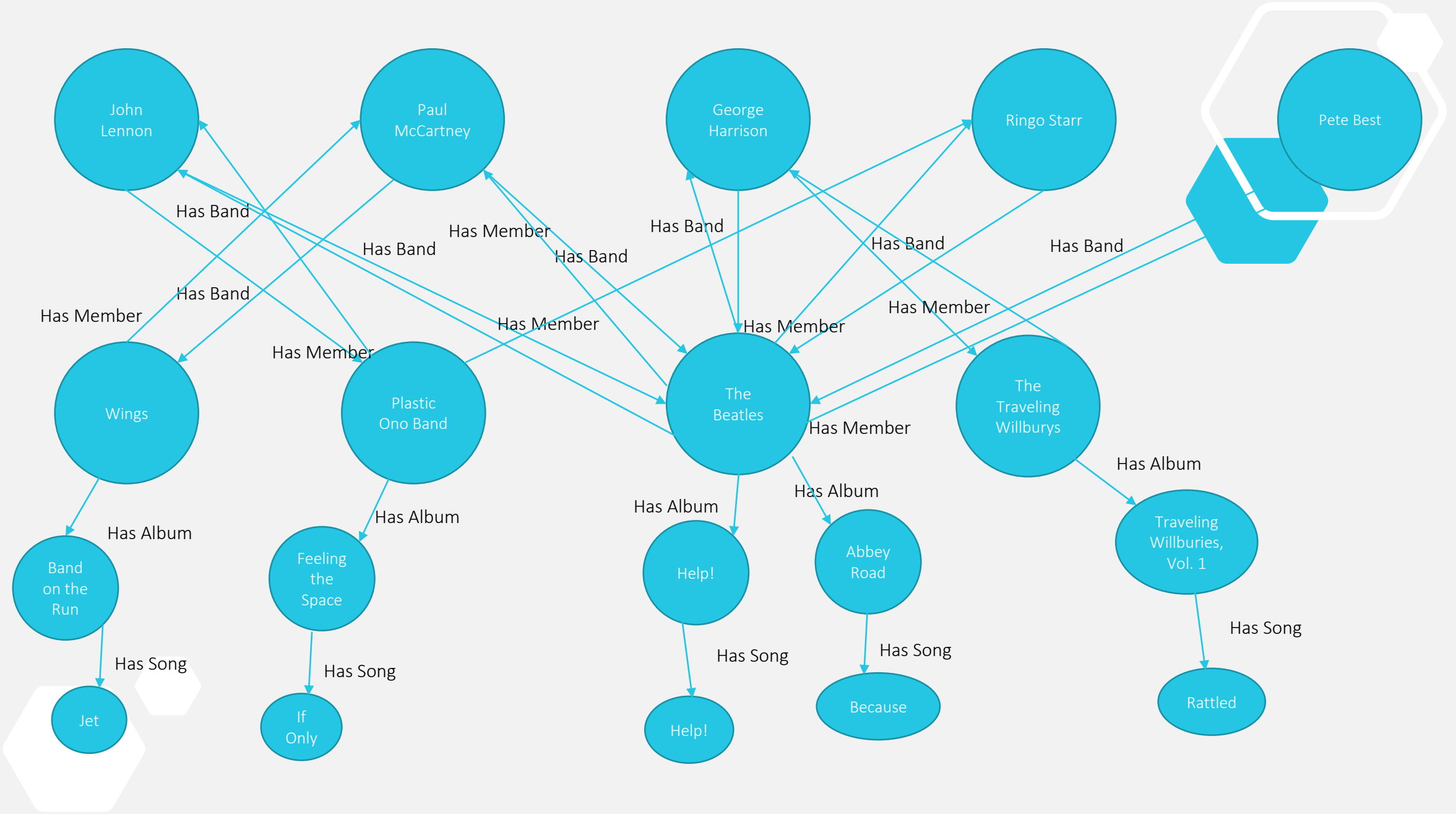


Classification

Assigning things to a list of pre-determined categories (taxonomy)



http://bioweb.uwlax.edu/bio203/s2012/harris_gran/classification.htm


John Lennon
Paul McCartney
George Harrison
Ringo Starr
Pete Best
Wings
Plastic Ono Band
The Beatles
The Traveling Willburys
Band on the Run
Feeling the Space
Help!
Abbey Road
Traveling Willburies, Vol. 1
Jet
If Only
Help!
Because
Rattled
Has Band
Has Member
Has Album
Has Song

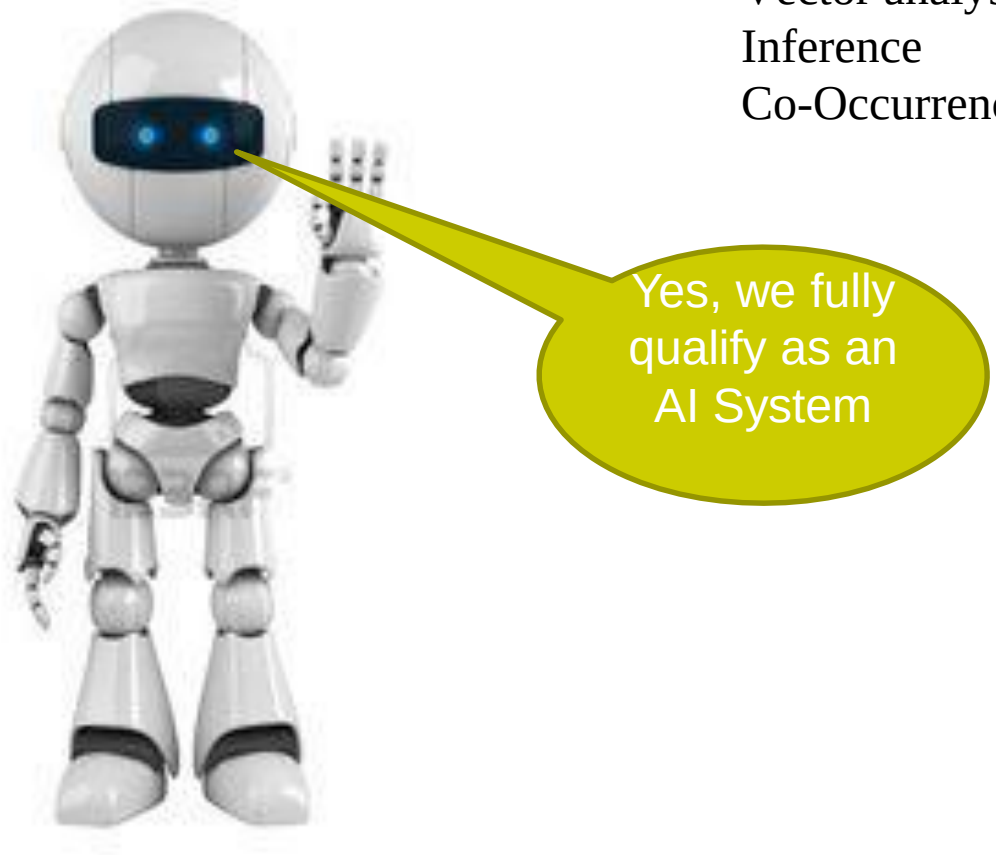
Plus we have:

Entities (people, places, things)
- Identification
- Salience via weighting

Syntactic analysis (Parsing)
Semantic Analysis
Sentiment Analysis
Pragmatic analysis
Grammar
Lemmatization - stemming
Morphological
Lexical variations - synonyms
Part-of-speech tagging
Sentence boundary
Punctuation mark,
Abbreviations
Terminology extraction
Term weighting
Co-Occurrence*
Rules to increase accuracy
Word Parsing
Phrase parsing
Rule bases
Concept extraction

We do not have

Neural Net
Bayesian statistics
Vector analysis
Inference
Co-Occurrence*



* Co-occurrence in our system is based on counts of occurrences

Headlines – the good, the bad, & the ugly

- *The Good*
  - “Can Artificial Intelligence Help Reduce False-positive Mammograms?”
  - “You Might Want Artificial Intelligence Reading Your Next Mammogram”
  - “When AI writes the Court Rulings”
  - “Fast and Accurate Annotations of Short Texts with Wikipedia Pages”

Headlines – the good, the bad, & the ugly

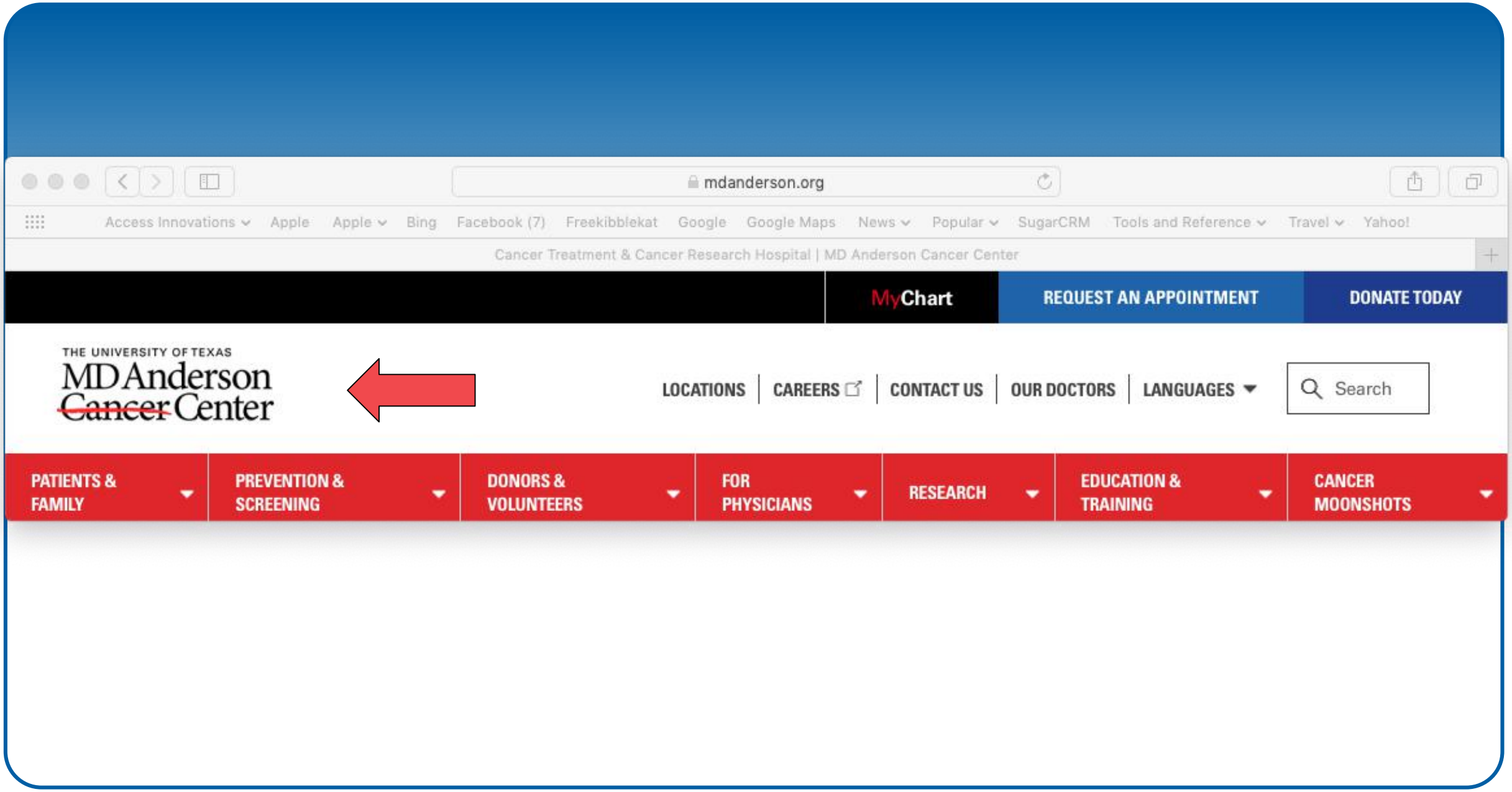
- *The Bad*
  - “Without Humans, Artificial Intelligence is Still Pretty Stupid”
  - “The Future of AI Depends on a Huge Workforce of Human Teachers” and “Why AI is Useless Without Human Beings”
  - “Google Has Picked an Answer for You – Too Bad It’s Often Wrong”
  - “Artificial Intelligence Still Isn’t a Game Changer?”
  - “Google, Smoogle. Reference Librarians Are Busier Than Ever”

Headlines – the good, the bad, & the ugly

- *The Ugly*
  - “Some AI Lessons from Watson’s Failure at MD Anderson”
  - “MD Anderson Benches IBM Watson In Setback for Artificial Intelligence in Medicine”
  - “Artificial Intelligence and Bad Data”
  - “Sky-high Salaries Are the Weapons in the AI Talent War”
  - And for ‘STM’ – SciGen – “Tech society retracts 29 articles, ousts three editors for ‘systematic violation’ of peer review polices”

IBM Watson and MD Anderson ~~Cancer~~ Center

- “Teaching a machine to read a record is a lot harder than anyone thought.”
- First, know that Watson was getting good results!
- The Fail
  - Not enough data
  - Inconsistent and bad data
  - Incompatible systems (Watson ⟺ EPIC’s EHR)
  - Changing objectives ⟹ oops, need to retrain Watson!
  - Lack of AI knowledge and expertise
  - Cost overruns – US$62 million spend before tabling project


mdanderson.org
Access Innovations
Apple
Apple
Bing
Facebook (7)
Freekibblekat
Google
Google Maps
News
Popular
SugarCRM
Tools and Reference
Travel
Yahoo!
Cancer Treatment & Cancer Research Hospital | MD Anderson Cancer Center
MyChart
REQUEST AN APPOINTMENT
DONATE TODAY
THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
LOCATIONS
CAREERS
CONTACT US
OUR DOCTORS
LANGUAGES
Search
PATIENTS & FAMILY
PREVENTION & SCREENING
DONORS & VOLUNTEERS
FOR PHYSICIANS
RESEARCH
EDUCATION & TRAINING
CANCER MOONSHOTS

IBM Watson and MD Anderson ~~Cancer~~ Center

- “Teaching a machine to read a record is a lot harder than anyone thought.”
- First, know that by the end Watson was getting good results!
- The Fail
  - Not enough data
  - Inconsistent and bad data
  - Incompatible systems (Watson ⟺ EPIC’s EHR)
  - Changing objectives ⟹ oops, need to retrain Watson!
  - Lack of AI knowledge and expertise
  - Cost overruns – US$62 million spend before tabling project

Get back to project management basics

- “Don’t believe the hype – AI is just a tool at the end of the day, but a very clever tool…”
- The Hype
  - “…find growth and accelerate innovation within an open data environment”
  - “…breaking the silos of the status quo…”
  - “Adopting a holistic data strategy…”
  - “…providing next generation…”
  - “IBM Watson capabilities to unlock previously unavailable data insights”

Get back to project management basics

- The hype suggests fabulous results
  - "…to unlock previously unavailable data insights"
  - All AI systems produce results
  - Not all results are useful or meaningful
  - Getting useful results gets expensive quickly
  - It is Artificial *"Intelligence"*
    - Useful output is intelligence
    - But the AI system is not *"intelligent"*

Get back to project management basics

- Stick to basic project management practices
  - Use cases or research objectives
  - Business cases
  - “Plan your dive. Dive your plan.”
  - Without a big budget, keep your expectations in check – narrow your focus
    - Do you have US$62 million to blow?

Get back to project management basics

- Use cases
  - Anti-SciGen, fake news detection, submissions analysis
  - Auto text generation and summarizations (big in the news business)
    - Court rulings (e.g. Prometea – Argentina)
    - Washington Post, Associated Press (e.g. sports summaries)
  - Machine automated indexing (MAI), semantic enrichment
  - Image analysis and recognition, info-graphics
  - Author & institution disambiguation, entity extraction, ‘triples’ generation

Get back to project management basics

- Use cases cont.
- Google Maps
- Self-driving cars
- Medical diagnosis
- Alexa
- Siri
- IBM Watson

Some notions of cost

- Headline costs – beware
- Software – free to hundreds of millions
  - Support? Think what Redhat did for Linux!
  - “Genuine” AI/ML/DL software?

Some notions of cost continued

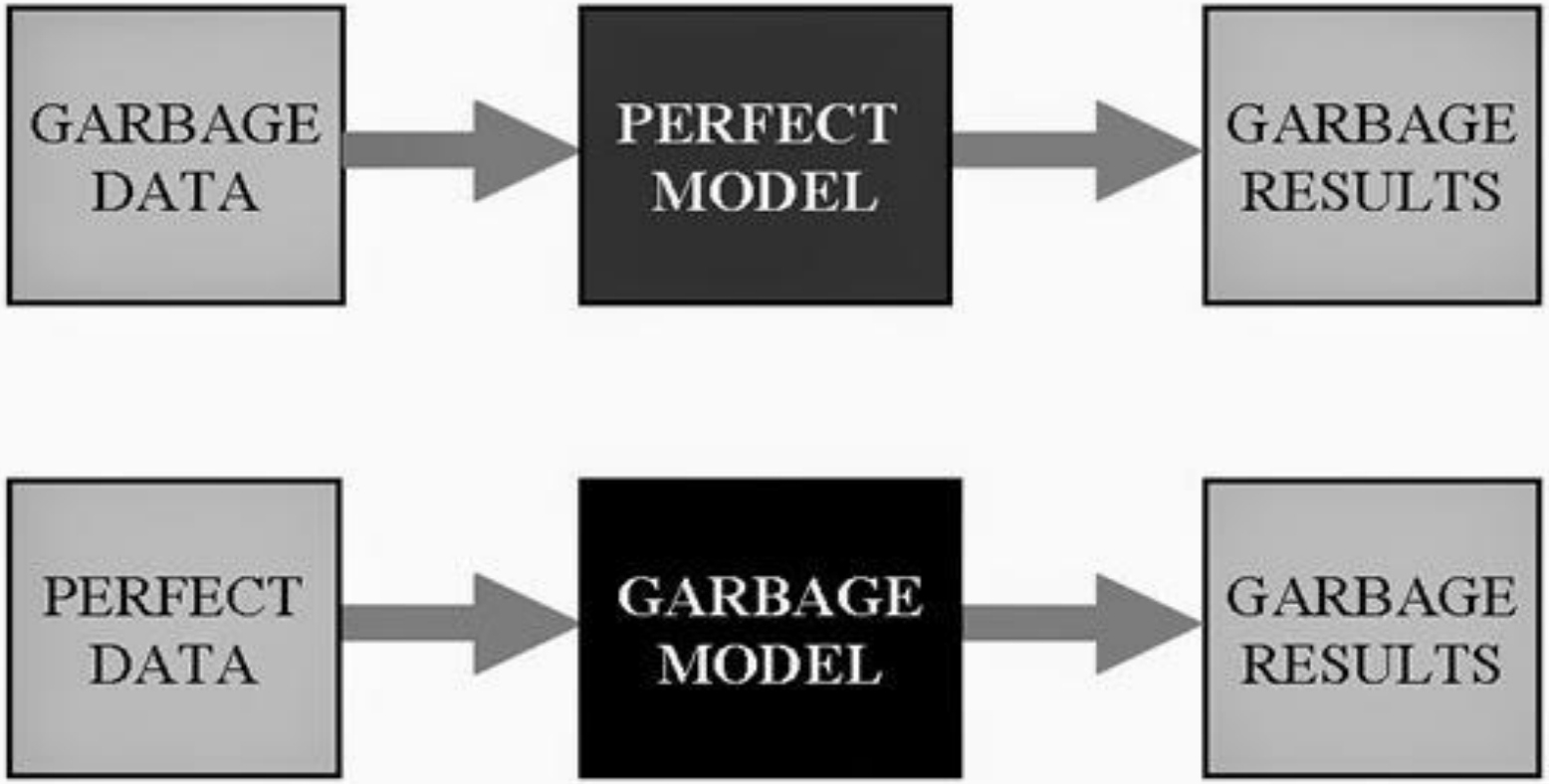
- Data quality costs
  - Very large data sets (thousands to millions) must be gathered and curated
  - Data sets must be conceptually and contextually unique – (e.g. 20 to 40 for each semantic node)
  - Corrupt and inconsistent data needs normalizing and cleanup
  - Remove biased data
  - Format consistency

Clean up your data!

- Inconsistent and bad data

MODEL CALCULATIONS

"Garbage In-garbage Out" Paradigm



Some notions of cost continued

- Training costs – people, time
  - Look for software that makes training possible in-house
    - Again, think what Redhat did for Linux and Data Harmony
    - Need a good user interface for training tasks, app maintenance
  - US$.03 to $0.15 per piece at outsourcing services, but up to $2,000, for example tagging a medical image
  - Staff size – (Facebook – 20k and growing!)
- The cost of change – more retraining costs

Some notions of cost continued

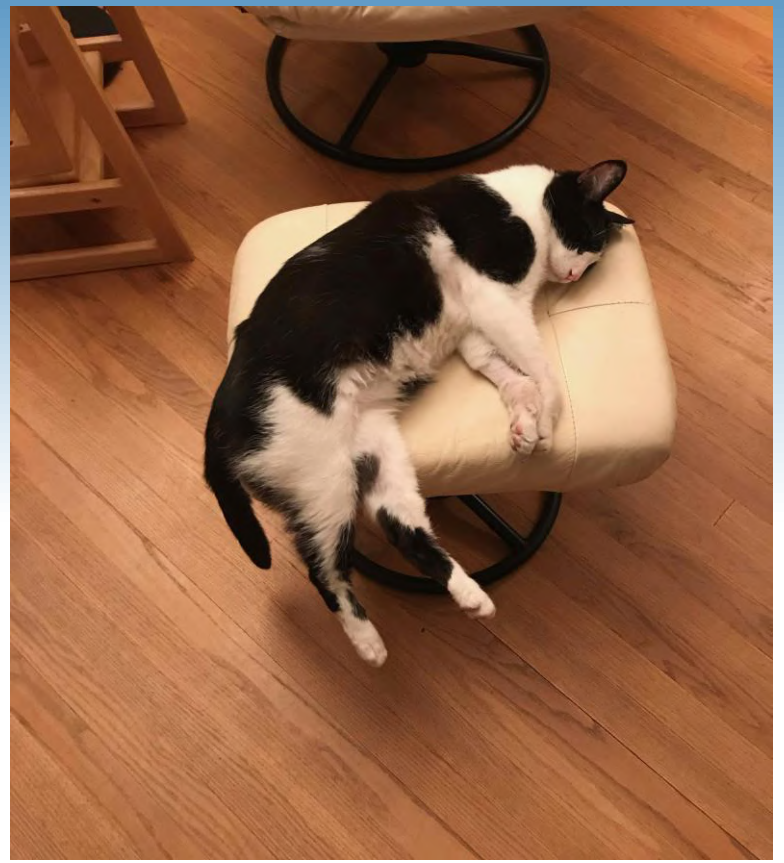
$.03 to $0.15 per piece at outsourcing services up to $2,000 or….treats!

Some notions of cost continued

And free perks….

Some notions of cost continued

Unfortunately….



Training-costs case study

- Objectives for this project
  - Improve productivity
  - Improve search discovery
- Goals
  - Lower cost per item
  - Improve discovery to 85% or better for recall and precision
- Process was to automatically cluster after using training sets vs. semantically enrich the content

Training-costs case study continued

- AI clustering (e.g. pattern matching)
  - 7500 semantic nodes to train
  - 7500 labor hours to curate training sets
    - 20 to 40 items per node needed
    - Review 60 automatically generated items to get to 20 “unique”
    - Retrain is still 1 hour per node
- AI using a rules layer with curated taxonomy* plus…
  - 7500 semantic nodes to train
  - 125 labor hours to curate automatically generated rules layer
  - Retrain is <5 minutes per node

*Data Harmony®

Risks? Yes!

unethical use … insufficient learning … incomplete data… inaccurate data … unsecured data … regulatory noncompliance … unrepresentative data … biased models … discriminatory outcomes … model instability … over fitting … performance degradation … implementation errors … poor design … insufficient training … technology malfunction … performance issues … human machine interface failures … opacity … explain-ability …

An Example of how the ICD-10 Rule Base “Learns”

<ttm>excema</ttm>
<rule>
IF (WITH "flexur*")
USE L2082 *(Flexural eczema)*
ENDIF
IF (WITH "infant*")
IF (WITH "chronic*" OR WITH "acute")
USE L2083 *(Infantile (acute) (chronic) eczema)*
ENDIF
ENDIF
IF (WITH "intrins*" OR WITH "allerg*")
USE L2083 *(Intrinsic (allergic) eczema)*
ENDIF
IF (NOT (WITH "flexur*" OR WITH "infant*" OR WITH "intrins*" OR WITH "allerg*"))
USE L2089 *(Other atopic dermatitis)*
ENDIF </rule>

- We begin with a trigger word (Text to Match or TTM). When the system sees this word in the documentation, the system then begins reading the rule for further clarification
- Proximity indicators (AROUND, WITH, NEAR, etc) let the system know how near to the trigger word that it should read the text
- When a condition (TTM, along with further conditions) is met, a code is recommended
- If the TTM is the only condition that is met, the system reverts to an “unspecified” or “other” code. While this is an indicator of poor documentation, it gives the user a recommendation, as well as proximity to additional codes that could be documented to have more specific code recommendations
- As more and more documentation is run through the system, the rules become increasingly refined

And, finally…

- Like MD Anderson, know when to cut your loses
- Good luck!

And your good luck will come from good planning and good execution!

Thank you!

Jay Ven Eman, Ph.D., CEO
Access Innovations, Inc. / Data Harmony
j_ven_eman@accessinn.com
www.accessinn.com
+1.505.998.0800
Albuquerque, NM USA

Access Innovations, Inc.

The Science behind the Semantics™

www.accessinn.com