Not knowing what we know:
A call for a theory-neutral database for empirical results in psychology.

Ven Popov & Lynne Reder

Carnegie Mellon University

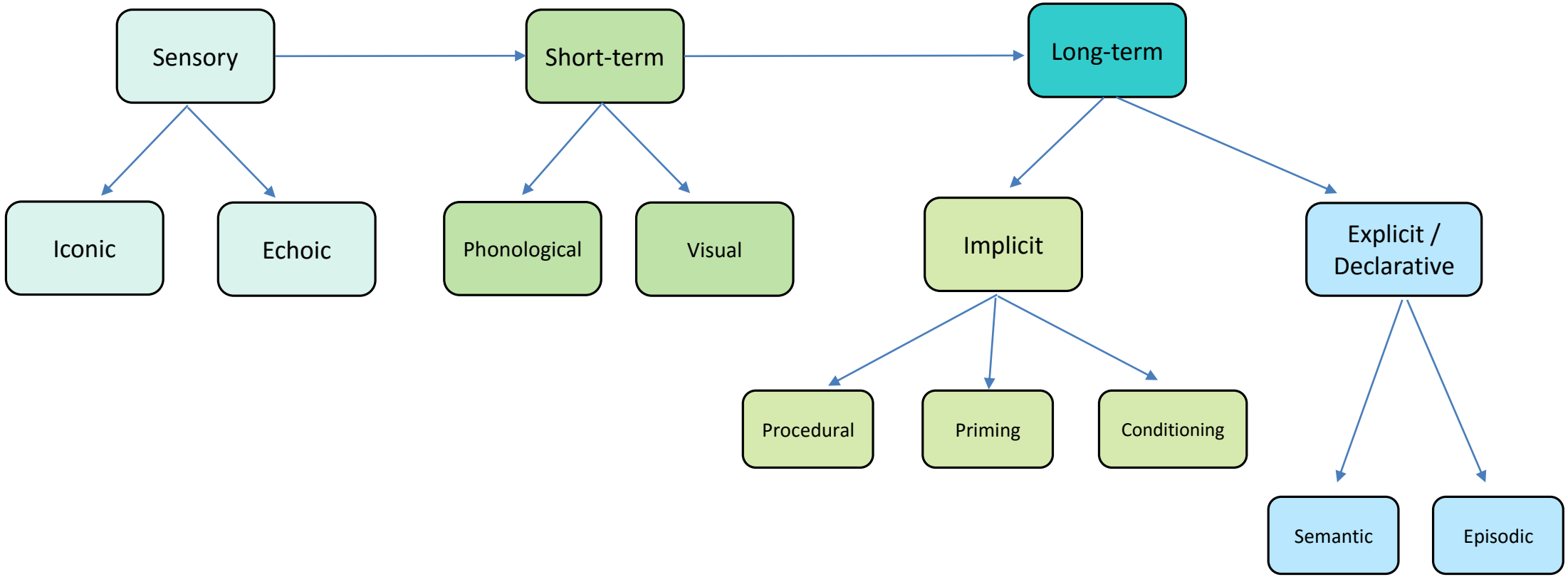
Center for the Neural Basis Of Cognition

Overview of the talk

1. What type of question is the question about memory systems?
2. Systems vs task-dependent process
3. Statement of the problem – We do not know what we know!
4. Proposed solution – a theory neutral database for empirical results
5. Conclusion

1. What type of question is the question about memory systems?

Memory typologies



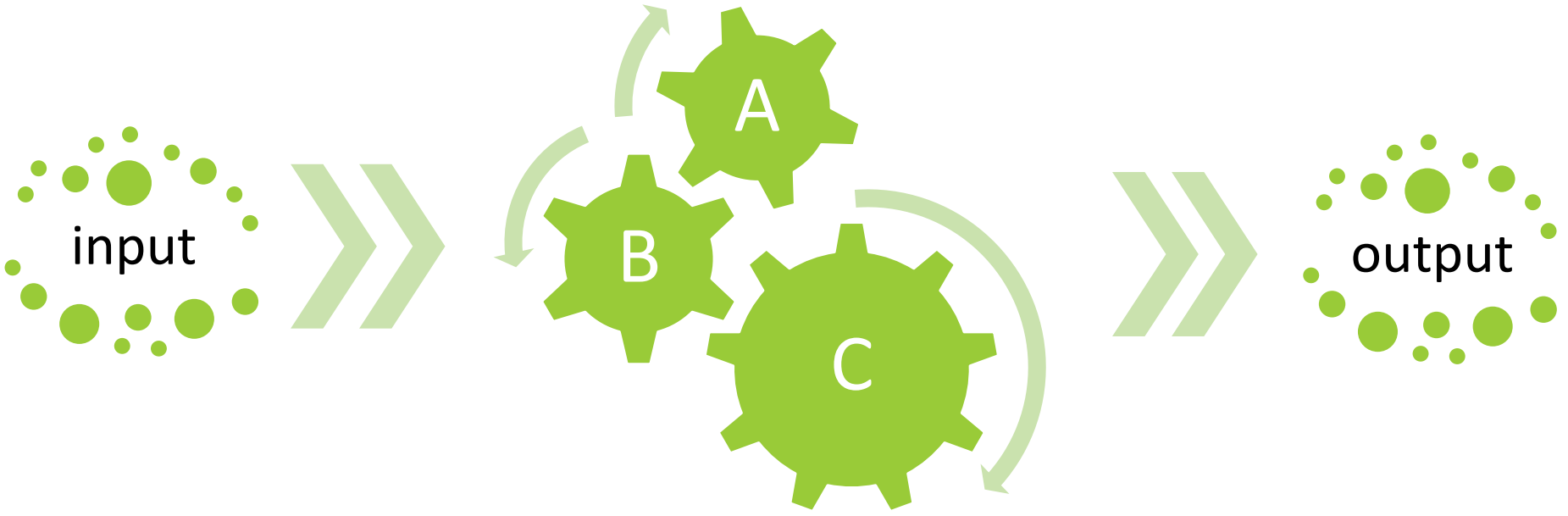
Types of memory systems distinctions

- **Heuristic**
  - Divide and conquer
  - Stimulate novel research
  - Organize results

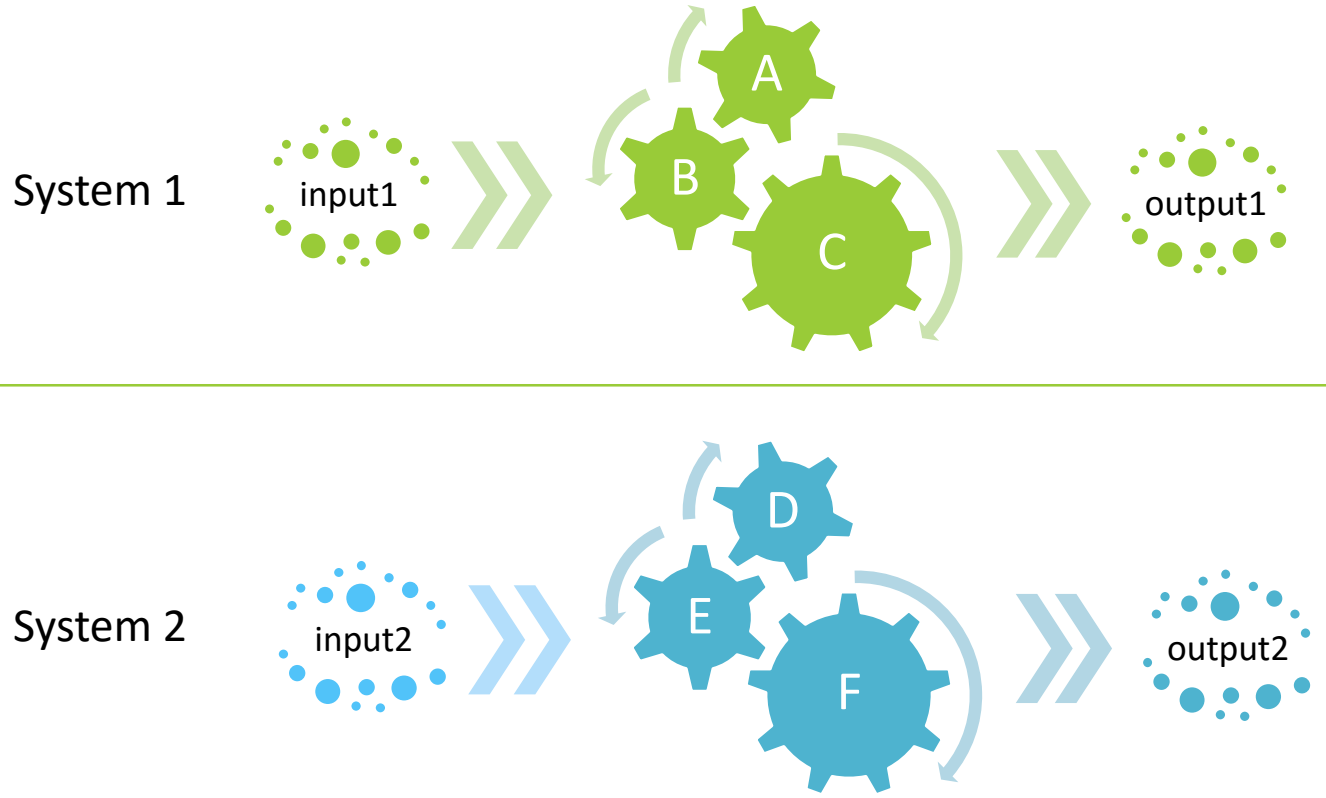
Types of memory systems distinctions

- **Functional/structural**
  - Different systems that
    - Operate independently
    - Can be interfered with independently
    - Can be facilitated independently

What is a system?

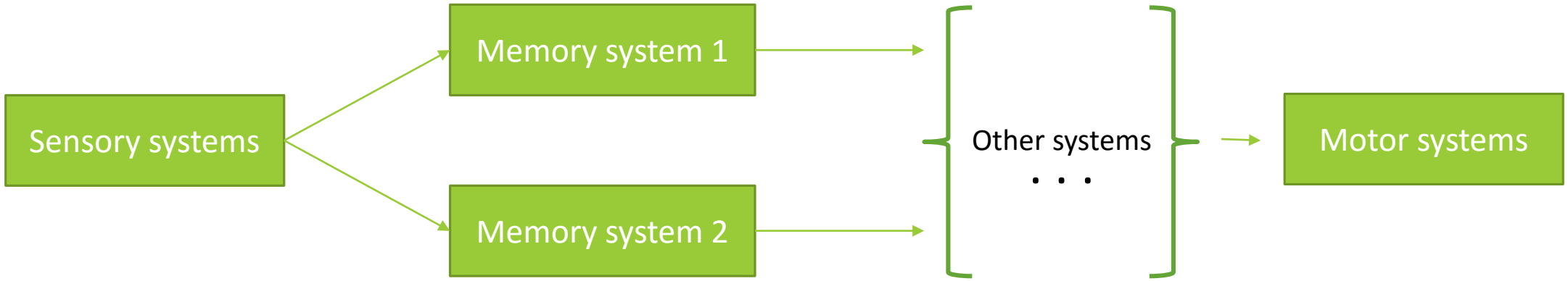


What is meant by different systems?

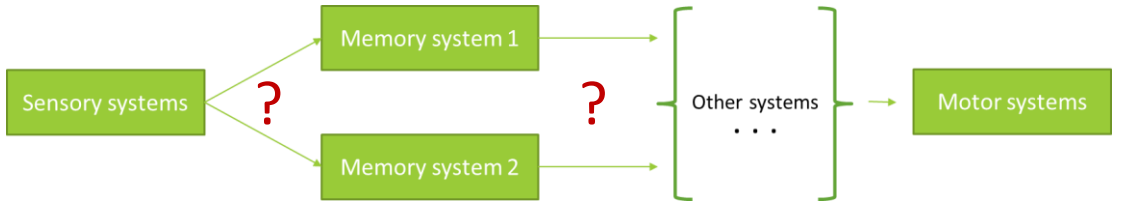


What is meant by different systems?

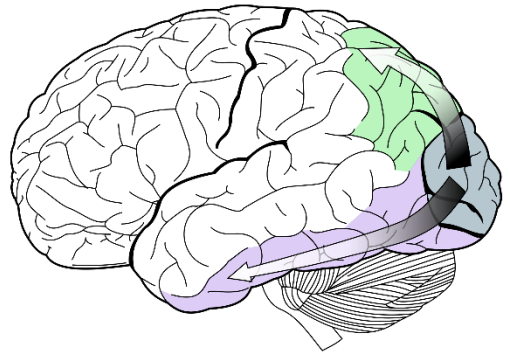
Modularity?



Problems



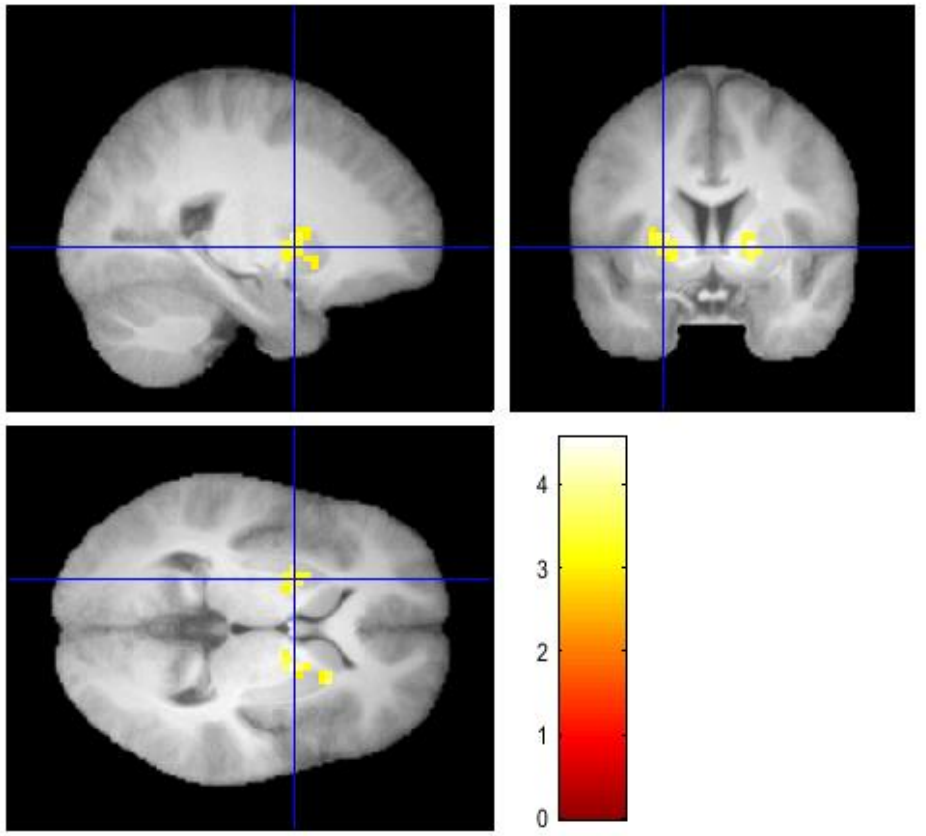
- Boundaries not always clear
  - Continuum of perceptual-semantic encoding in inferior temporal cortex
  - MTL also involved in complex perceptual discrimination (Graham et al, 2010)
  - Memory is also a property of lower-level perceptual processing
    - Phonetic distributional learning (Werker & Tees, 1984)
    - Receptive fields in V1 neurons not innate and immutable (Tanaka et al, 2006)
    - Learning of temporal sequences in V1 (Gavornik & Bear, 2014)
- Difficult to falsify
- Little evidence for memory-type specific regions



Basal ganglia involvement in episodic memory

Popov & Reder (in preparation)

- Study phase (outside scanner)
  - Learn esoteric (true but unknown) facts about famous people
- Test phase (inside scanner)
  - Stage 1 – “True or false?”
    - Mixed known and learned facts together
  - Stage 2 – “Were you tested on this item in Stage 1?”
    - Known and learned facts, half were tested, half not



Episodic discrimination: recombined > intact

An example: Semantic vs episodic memory

- Tulving (1972) – heuristic distinction
  - Semantic memory
    - Meaning of words
    - General world knowledge
    - Fact, ideas, concepts
  - Episodic memory
    - **What**, **when**, **where** happened to **me**
    - Episodes, events

- Tulving (1984) – functional distinction

Types of evidence

- Behavioral dissociations
  - Variables that affect differently semantic and episodic memory

- Problematic because:
  - not falsifiable in absense of theory (Hintzman, 1984; McKoon and Ratcliff, 1986)
  - differences between tasks/content (Klatzky, 1984; Roediger, 1984)
  - dissociations between free recall and recognition (Roediger, 1984)?

Types of evidence

- Pathological dissociations
  - MTL patients - episodic impairment, but not semantic
  - Semantic dementia – semantic impairment, but not episodic

- Problematic because
  - Pattern not as clear as initially suggested
  - Inherent variability in damage location
  - Compensatory mechanisms

Types of evidence

- Neuroimaging data

- Problematic because
  - Again, maybe differences between tasks/content/difficulty, not systems
  - Episodic encoding occurs even during semantic retrieval
  - Semantic retrieval occurs during episodic retrieval
  - Different networks for different episodic tasks
    - Novelty/familiarity network
    - Recollection network

2. Systems vs task-dependent process

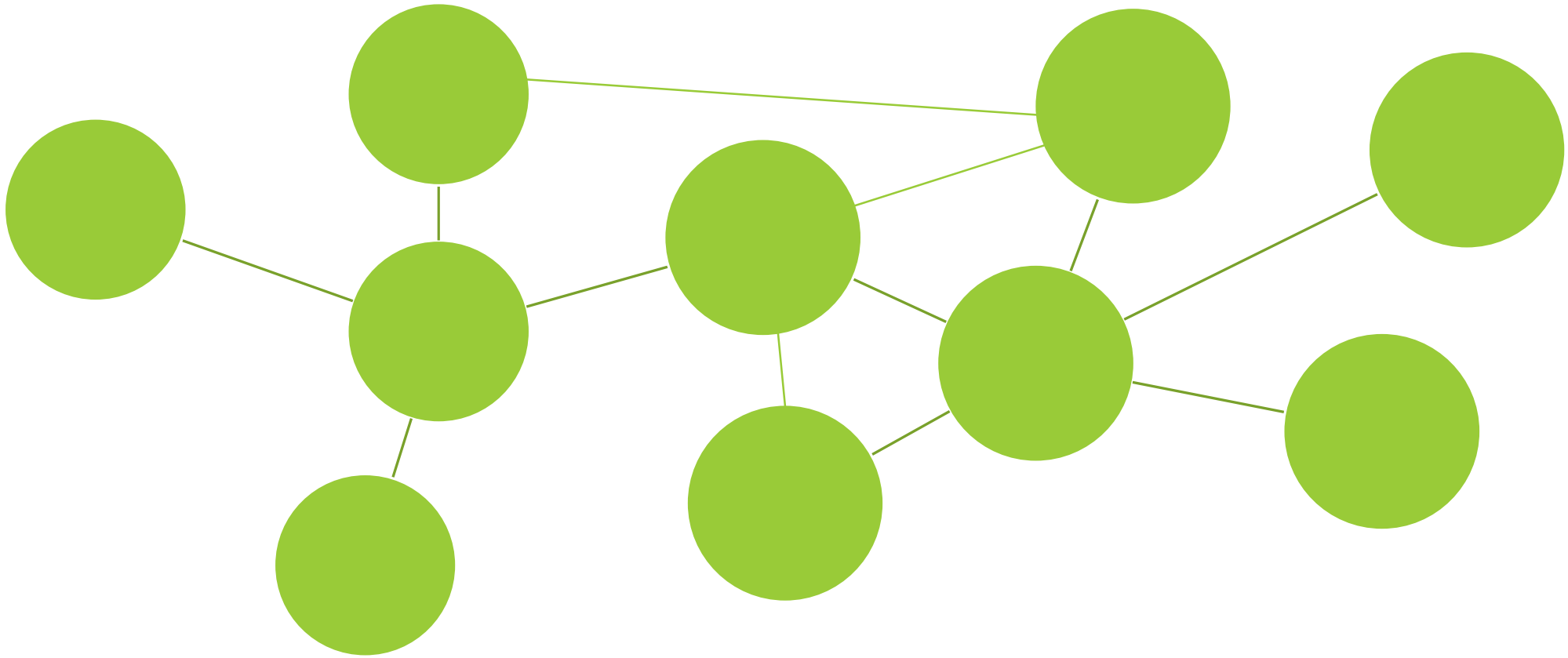
Maybe it is time to change the question

- Stop asking “same or different system?”
  - Rarely falsifiable
  - Even if it was, answers not useful
  - What information do we gain from pursuing this question?

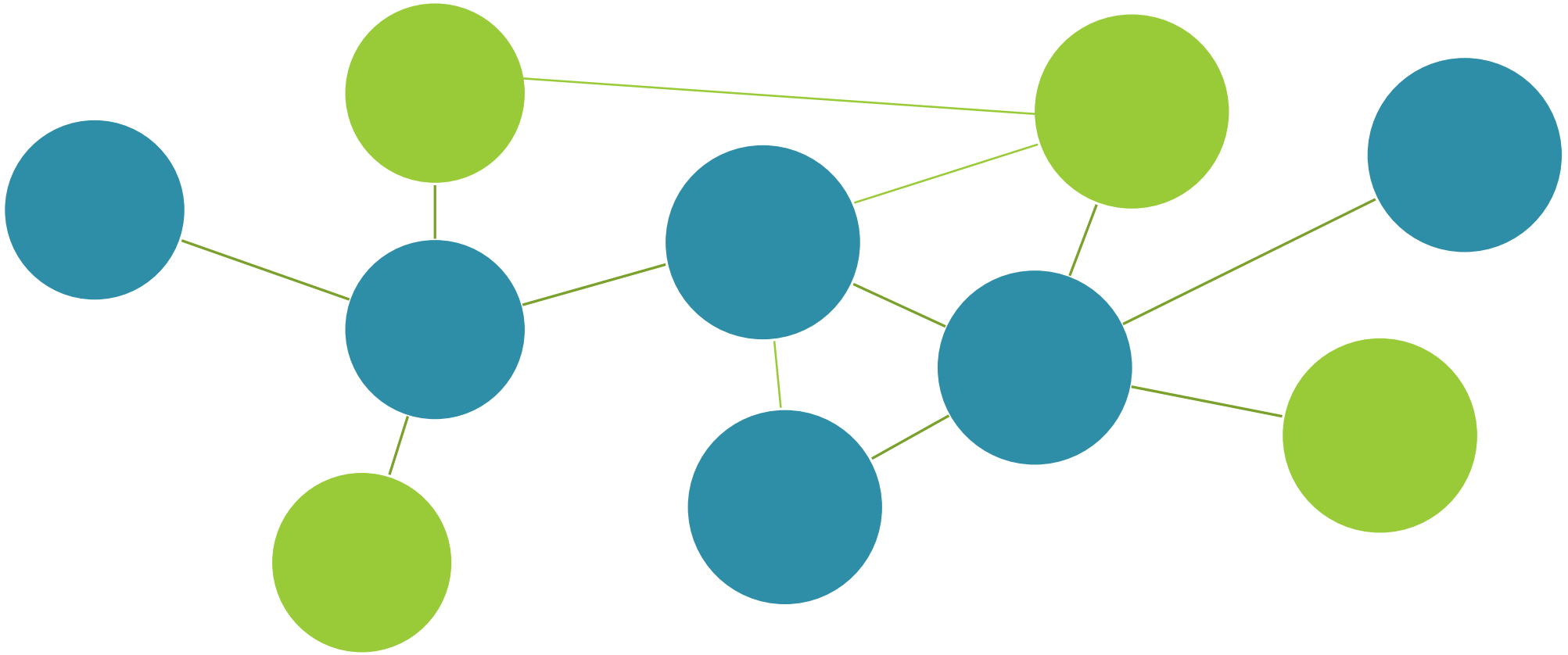
What do we want to know about cognition?

1. Mechanistic understanding
2. Prediction
3. Intervention
   - prevent/treat impairments
   - enhance normal functioning

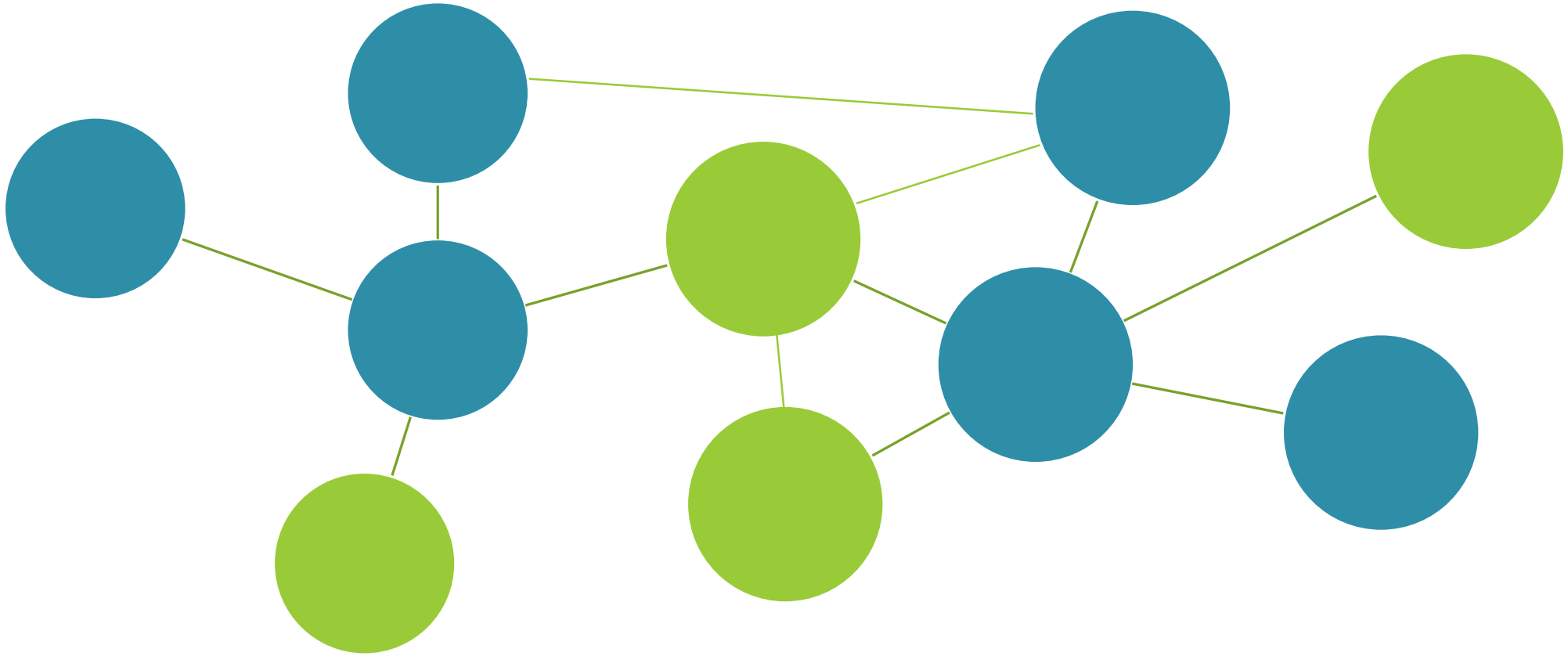
Task-dependent processes (Cabeza and Moscovitch, 2013)



Task-dependent process – task 1



Task-dependent processes – task 2



Task-dependent processes

Non-falsifiable, but it's a framework for driving research, not a cognitive theory

Task-analysis:

- What information needs to be processed?
  - E.g. – binding, temporal/spatial context, etc
- What combination of processes is involved?
  - E.g. – what processes are required for binding?
- What guides this processing (cognitive control)?
  - E.g. – how does the system determines if it should retrieve the binding or the items, depending on the task demands?

Task-dependent processes

- Implicitly accepted in neuroimaging research
  - Perirhinal cortex
    - **Evaluates object/concept familiarity/novelty** (Mayes, Montaldi, & Migo, 2007)
  - Parahippocampal cortex
    - **Representation of context** (Bar & Aminoff, 2003)
  - Hippocampus
    - **binding and relational processing** (Reder et al, 2009)
  - VLPFC
    - **controls access to information** (Badre & Wagner, 2007)
  - aMTG
    - **Representations of semantic categories** (Coutanche & Thompson-Shill, 2014)
  - IPL (angular gyrus and supramarginal gyrus)
    - **complex information integration** (Binder et al, 2009)

Focus on representations and processes, not systems

- This approach has already been useful
- Example: Memory systems do not divide on consciousness (Reder et al, 2009)
  - Challenged the implicit/explicit distinction
  - Used a mechanistic model to accommodate conflicting results
  - Same representations support
    - Implicit memory tasks
    - Familiarity-based recognition

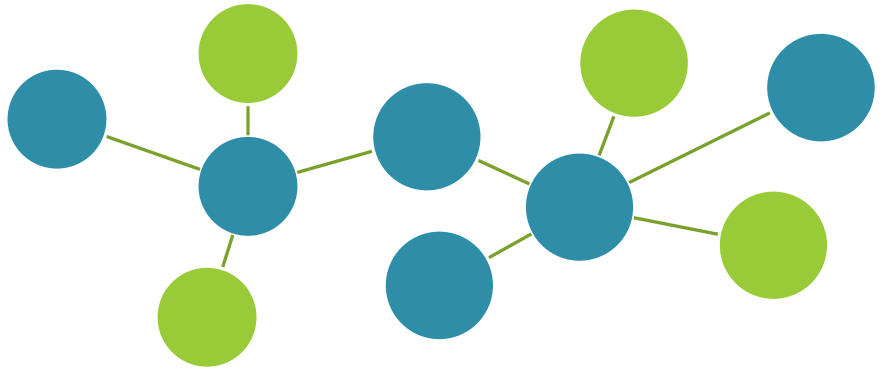
Interim conclusions

- Memory distinctions are useful as heuristics
- The debate about memory systems is reducible to the debate about modularity
- The debate may be unfalsifiable at best, or pointless at worst
- Cognitive scientists care about processes and mechanisms
- So let’s study those in a task-dependent manner (as we already do implicitly)

3. Statement of the problem – We do not know what we know!

Problems that hinder theoretical advancement

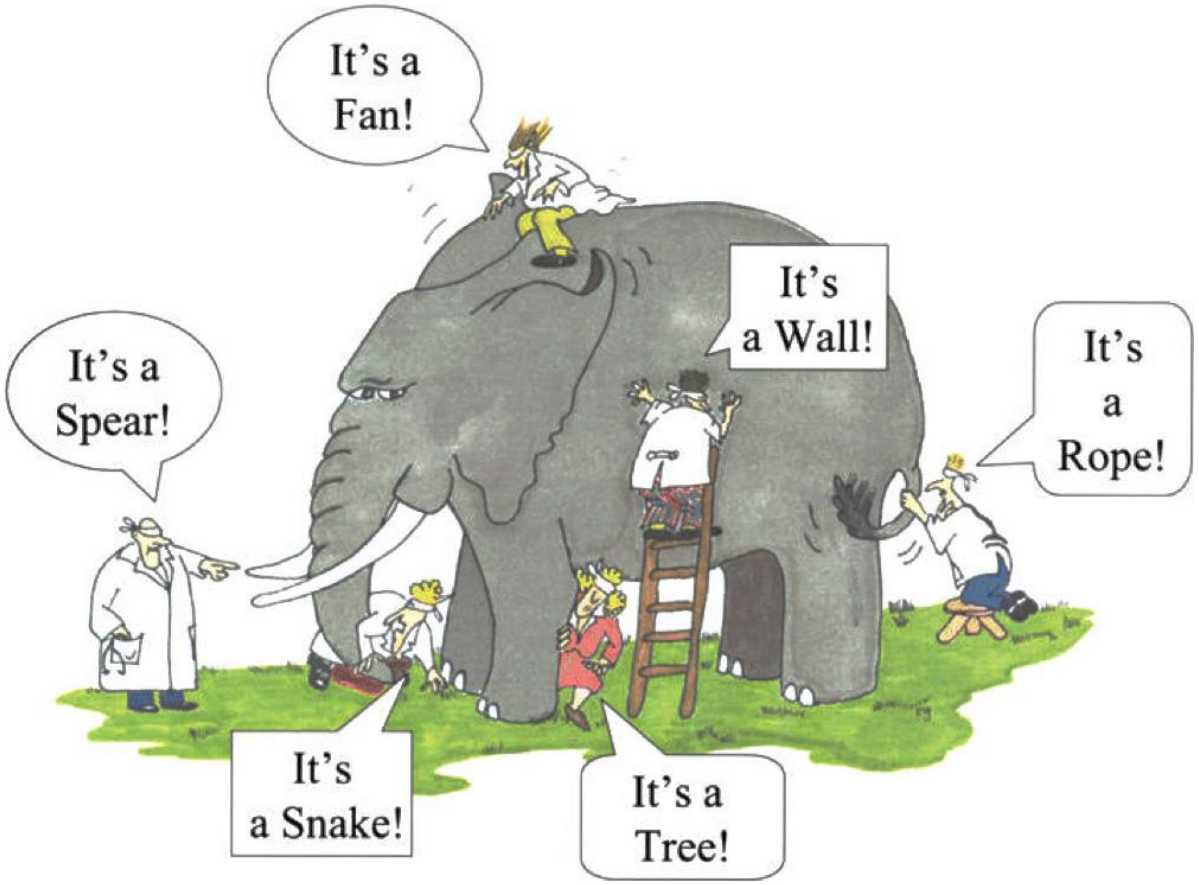
1) Empirical isolation
2) Unwarranted parsimony
3) Filler terms
4) **Lack of organization of empirical results**



Empirical isolation

- Many tasks and paradigms to study memory and resulting phenomena
  - Directed forgetting
  - Distinctiveness
  - Deese-Roediger-McDermott
  - Testing effect
  - The fan-effect
  - The frequency mirror effect
  - The list-strength effect
  - Etc…
- Little cross-talk (cottage industries)
  - No integration of results into a common framework

Empirical isolation



Unwarranted parsimony

- Same label refers to empirical effects in different paradigms
  - Negative priming
  - Directed forgetting
    - Item-based vs list-based paradigms
  - Implicit learning vs implicit memory
  - Metacognition – Reder (1996). Different research programs on metacognition: Are the boundaries imaginary? Commentary for special issue of Learning and Individual Differences., 8(4), 383-390.
- Contradictory results might just reflect task and process differences
- Often people want to explain them with the same processes/models
- Parsimony is a desirable quality, but can hinder progress if something is not a natural kind

Filler terms (Craver, 2003)

- Terms refer to a mechanism without details can give the illusion of understanding
- Examples
  - inhibit, represent, encode, "strategically controls access to information", primes, resource depletion, etc
- Favorite example - *priming*
  - “something has been primed” often used as an explanation
  - priming refers to an empirical effect, not to a mechanism
  - maybe used as a short-cut "whatever mechanism is involved in priming is responsible for this result here"
  - yet priming depends on different mechanisms, depending on the paradigm (Neely, 1991)
- Useful as place-holders for "a process that we do not yet understand", but often taken as actual mechanism descriptors
- “One person's explanation is another person's description” - Patrick Suppes

Root cause – lack of empirical organization

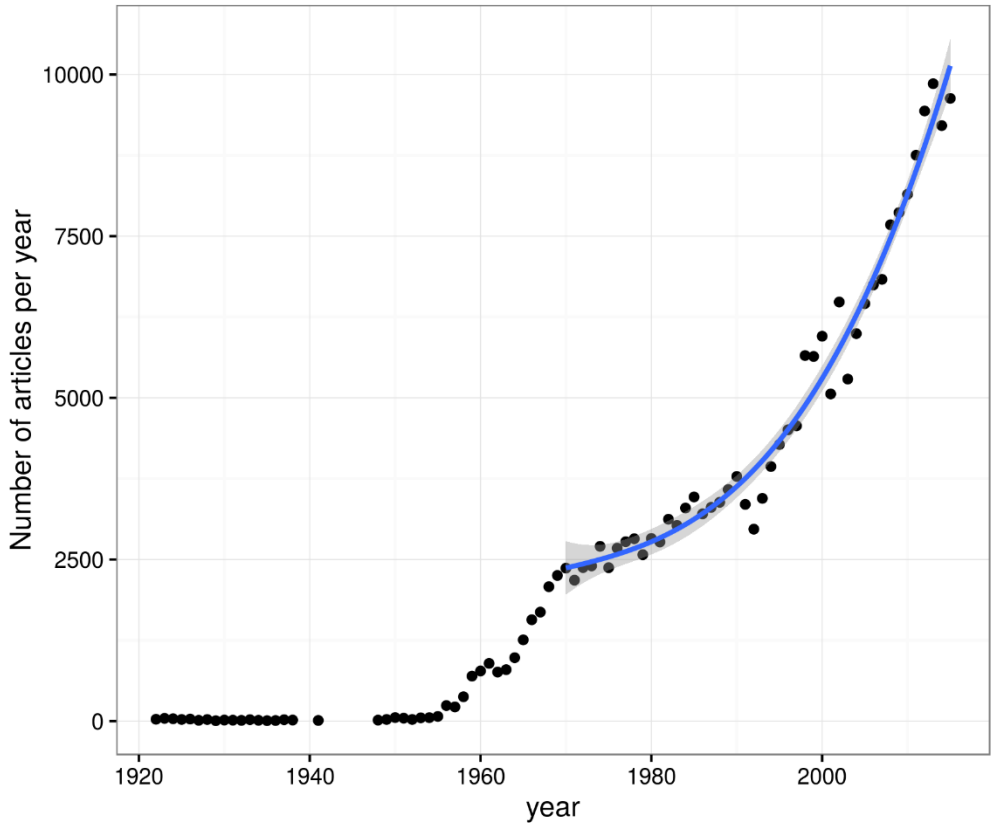
- Current model
  - Empirical results are buried in prose
  - Thousands of new papers published each year
  - Parse papers to learn about a new domain

- No readily available access to current knowledge
  - Chemistry:
    - What is the melting point of helium? **-272.0 °C**
  - Psychology:
    - Accuracy for single item recognition for 200 studied items, with 1 s. stimulus presentation time? **Ugh, let me see…**
    - Visual word recognition? **Maybe around 400 ms?**

Root cause – lack of empirical organization

- Analogy to “descriptive statistics” for single studies
  - Raw data (single observations) is uninterpretable within a single study
  - Summary data from single studies = raw data points on a grand-scale

Publication rate – experimental psychology



- WebOfKnowledge search
  - 1900-2016
  - Category: Psychology, Experimental
- Total articles – 235,998
- Change in trend in 1970
- Publication trend after 2070 can be modelled as a 2nd level polynomial:

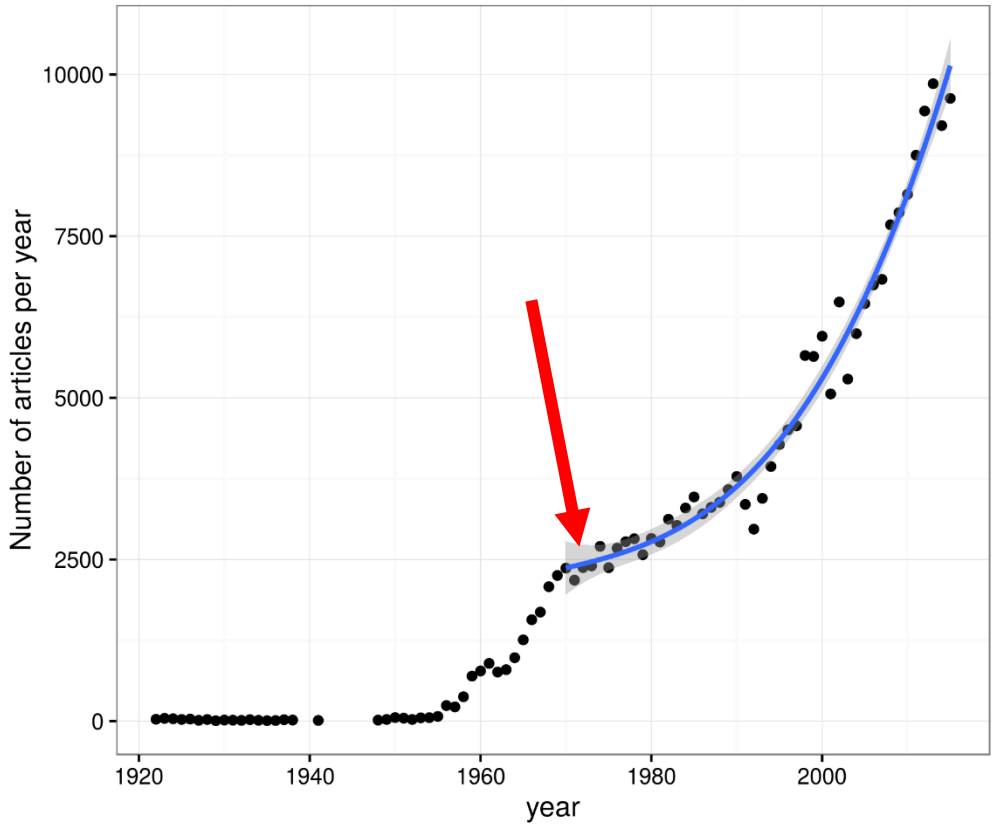
$$Number\ of\ articles\ per\ year \sim Year + Year^2$$

- Total R^2 = .97
- $\Delta R^2 = 0.10$ for $Year^2$

- If trend continues, 20,000 new articles will be published in 2037 alone.

- *Disclaimer – this includes theoretical articles as well. Without review articles, total number = 163,284*

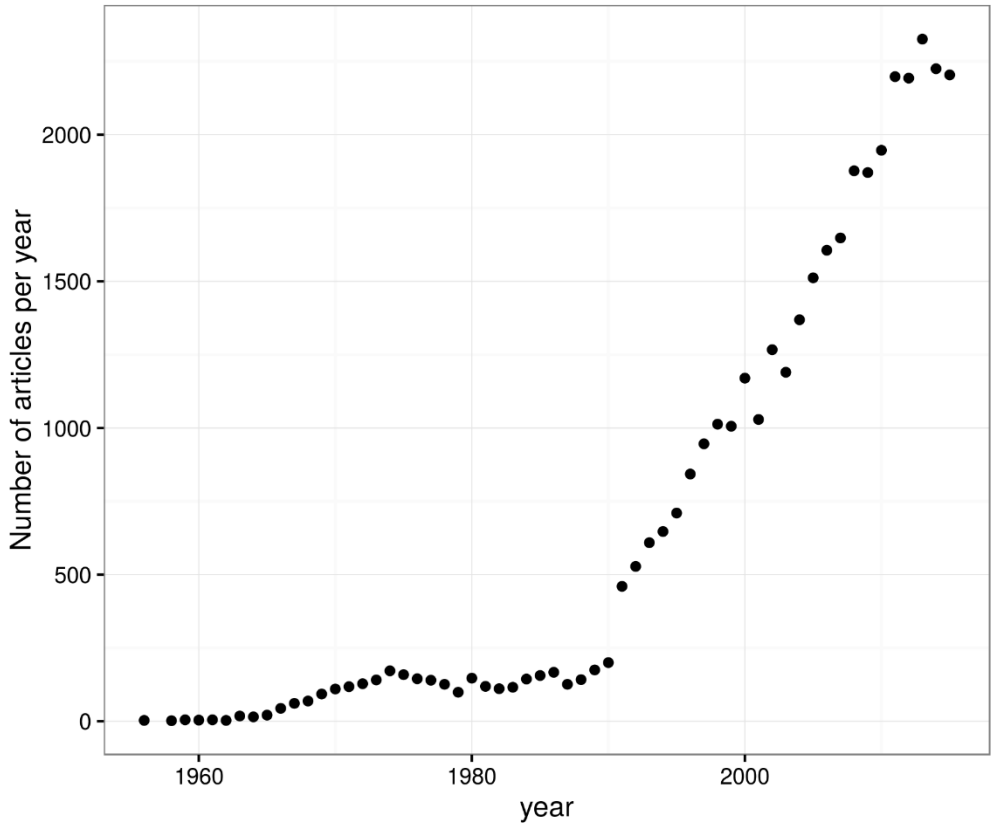
Publication rate – experimental psychology



- *1973 – You can't play 20 questions with nature and win*
  - **Allen Newell** complained that psychology has studied so many phenomena, but has little to offer in term of theoretical integration

diverse to be forced into such an iron maiden. Suppose that in the next thirty years we continued as we are now going. Another hundred phenomena, give or take a few dozen, will have been discovered and explored. Another forty oppositions will have been posited and their resolution initiated. Will psychology then have come of age? Will it provide the kind of encompassing of its subject matter--the behavior of man--that we all posit as a characteristic of a mature science? And if so, how will the transformation be accomplished by this succession of phenomena and oppositions? Same question as before, just a different lead in.

Publication rate – memory



- WebOfKnowledge search
  - 1900-2016
  - Category: Psychology, Experimental
  - Topic: Memory
- A little better – “only” 38,185 articles
  - 2200 new articles published in 2015

Cognitive architectures and models as a solution?

- Many different approaches
  - Samsonovish et al (2010) identified 26 in current use:
    - ACT-R, SOAR, Leabra, Clarion, …
- They help, but it is still up to people to pick which results to model
- Researchers disagree about underlying assumptions
  - “A theory is like a toothbrush – everyone wants their own and no one wants to use anyone else’s.”

Proposed solution

- **Theory-neutral database for empirical results**
  - A systematic mapping between task parameters (input space) and behavioral outcomes for every published study

Proposed solution

- **Theory-neutral database for empirical results**
  - A systematic mapping between task parameters (input space) and behavioral outcomes for every published study

INPUT

Task parameters/descriptors

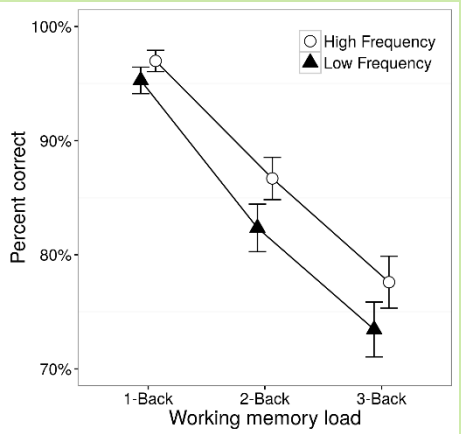
Task:	N-back
Stimuli:	Chinese characters
Stimuli duration:	2 seconds
N stimuli per block:	30
Number of blocks:	5
Design, IV1:	familiarity: low vs high
Design, IV2:	N-back level: 1 vs 2 vs 3

Raw data

	A	B	C	D	E	F	G
1	Subject	Block	Frequ	Level	ACC	RT	WordTy
2	1	1	LF	2	0	0	Foil
3	1	1	LF	2	0	0	Foil
4	1	1	LF	2	1	1182	Hit
5	1	1	LF	2	1	1159	Foil
6	1	1	LF	2	0	0	Hit
7	1	1	LF	2	1	1707	Foil
8	1	1	LF	2	1	1171	Foil
9	1	1	LF	2	1	1452	Foil
10	1	1	LF	2	1	2220	Foil
11	1	1	LF	2	0	1120	Hit
12	1	1	LF	2	1	1644	Foil
13	1	1	LF	2	0	0	Foil
14	1	1	LF	2	1	2031	Foil

OUTPUT

Summary data



Similar(ish) approaches

- Neuroimaging – **Neurosynth**
  - organized around voxel coordinates
  - automatic extraction of key terms and voxel coordinates
- Psycholinguistics
  - **MRC Psycholinguistic Database**
    - 150837 words with up to 26 linguistic and psycholinguistic attributes for each
    - Full psychological measures for about 2500 words.
  - **English Lexicon Project**
    - Naming and lexical decision latencies for 40,481 words
      - 2,749,324 measurements from 815 subjects for lexical decision
      - 1,123,350 measurements from 443 subjects in the naming experiment.
- Language development – **CHILDES**
  - Automatic analyses of transcripts of conversations with children from numerous studies

Task-descriptors and parameters

- Exhaustive descriptors and parameters for things like
  - Procedure
    - Duration
    - Order
    - ISI
  - Design
    - IVs and DVs
  - Nature of stimuli
    - Type: words, pictures, narratives, equations
    - Properties: frequency, category structure, etc

Task-descriptors and parameters

- Standardized hierarchical nomenclature
  - Taxonomic and horizontal relations
- Simple example
  - Explicit-memory test
    - Recognition
      - Associative-recognition
        - Study stage
          - Slides: Pairs of **[stimuli]**
          - Response: …
          - Required behavior: …
          - Duration of slide: ---
          - Etc…
        - Test stage
          - …

Task-descriptors and parameters

- **Important concerns (!)**
  - Success of proposal depends on efficient design the task-descriptors
  - May be very difficult
  - Balance between systematicity and flexibility
  - Ability to add new descriptors to the system
  - Ability to easily change and reorganize descriptors

Output – parameter estimation for behavioral measures

- **User interaction**
  1. Specify task-parameters
  2. Get a list of studies that have used those parameters
  3. View desired results
     - For individual studies
     - Automatic meta-analysis
- **Nature of the output**
  - Measures of central tendency
  - Measures of individual variability
  - Measures of within-individual variability

Output example

- **Visual word recognition – lexical decision**
  - **Overall, regardless of stimuli**
    - Mean reaction time: 600 ms.
    - Between-subject variability: SD = 150 ms.
    - Within-subject variability: SD = 50 ms.
  - **For high frequency words (X words per million), with low imageability ratings**
    - Mean reaction time: 500 ms.
    - Between-subject variability: SD = 250 ms.
    - Within-subject variability: SD = 50 ms.

Identifying gaps in knowledge

- Most common motivation for research is hypothesis testing
  - Stimulates publication bias, file-drawer bias, etc
  - Generates data, but not systematically
  - Imagine if chemistry was done entirely that way

- Alternative motivation – precise parameters estimation of behavior
  - With a completed database it's easy to identify gaps in knowledge
    - provide estimates for missing parameters
    - provide better estimates (less variability) for available parameters
  - Motivation for highly powered studies

Integration with publishing process

- When an article is accepted for publication authors have to submit
  - Article text
  - Task parameters
  - Raw data in a standardized format

Automated data analysis

- How is this different from current calls for sharing data?

1) Sharing not only data, but task descriptors using the database language
2) Raw data is in standardized format in a common database
3) Automatic summaries and analyses can be generated from 1) and 2)
4) Large-scale meta-analyses and parameter-estimation using Bayesian hierarchical methods on data from multiple datasets with similar task-parameters

Use for model testing and comparison

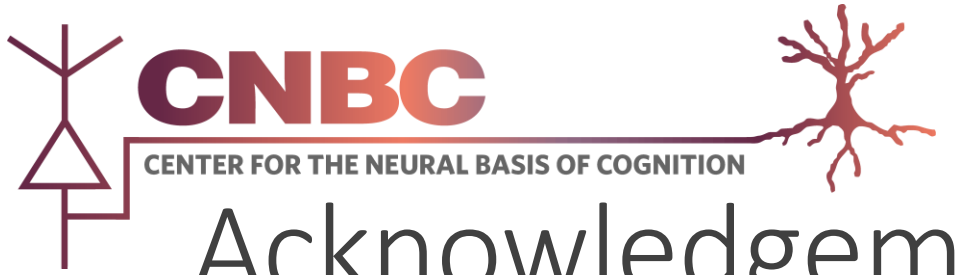
- Currently – comparison with hand-picked studies
- With a completed database
  - Interface for connecting task-descriptors and model inputs automatically
  - Model success measured by percentage of fitted results in the database
    - Overall
    - By categories
    - By tasks
  - Identify weaknesses by relative success by categories

Is it achievable?

- Quantity
  - 161 000 articles (maybe less?)
  - Large-scale distributed effort
  - Possibly impossible without some kind of automatic parsing
- Quality
  - The nomenclature for describing task-parameters is key for success
  - Balancing flexibility and systematicity
- Funding and commitment

Conclusion

- Theoretical advancement is hindered by the lack of organization of empirical results
- Too many published empirical studies to integrate verbally
- Models and architectures are useful, but do not solve the problem
- We need a theory-neutral database for empirical results
- Usefulness
  - Organization and integration of currently available knowledge
  - Automatic meta-analysis and large-scale parameter estimation
  - Identifying gaps in our knowledge
  - Model testing and comparison



Acknowledgements

- **My mentor:**
  - *Lynne Reder*
- **Discussants**
  - *Markus Ostarek*, Max Planck Institute for Psycholinguistics
  - *Francesca Biondo*, Cambridge University
  - *Elliot Collins*, Carnegie Mellon University

- **Sources of support:**

Richard King Mellon Foundation

HENRY L. HILLMAN FOUNDATION