Explainable and Reliable AI
Comparing Deep Learning with Adaptive Resonance

Stephen Grossberg

Center for Adaptive Systems
Graduate Program in Cognitive and Neural Systems
Departments of Mathematics & Statistics,
Psychological & Brain Sciences, and Biomedical Engineering
Boston University

steve@bu.edu
sites.bu.edu/steveg

This lecture is based on the following article:

Grossberg, S. (2020). A path towards Explainable AI and Autonomous Adaptive Intelligence:
Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action
Frontiers in Neurobotics, June 25, 2020
https://doi.org/10.3389/fnbot.2020.00036 (OPEN ACCESS)

The article summarizes core problems of DEEP LEARNING,
such as its untrustworthiness (unexplainable)
and unreliability (catastrophic forgetting),

explains how ADAPTIVE RESONANCE overcomes them,
indeed overcomes 17 problems of Deep Learning,

and outlines a blueprint for achieving autonomous adaptive intelligence

The article is part of a Frontiers in Neurobotics special issue about EXPLAINABLE AI

Its editors J. L. Olds, J. L. Krichmar, H. Tang, and J. V. Sanchez-Andres write

"Though Deep Learning is the main pillar of current AI techniques and is ubiquitous in basic science and real-world applications, it is also flagged by AI researchers for its black-box problem: *it is easy to fool, and it also cannot explain how it makes a prediction or decision*"

Deep Learning is NOT TRUSTWORTHY

No life or death decision, such as a medical or financial decision, can confidently be made based upon a Deep Learning prediction

FROM BACK PROPAGATION TO DEEP LEARNING

Deep Learning uses the back propagation (BP) algorithm to learn how to predict output vectors in response to input vectors

BP was based upon perceptron learning principles
Rosenblatt (1958, 1987)

It has a complicated history; cf., Schmidhuber (2020)

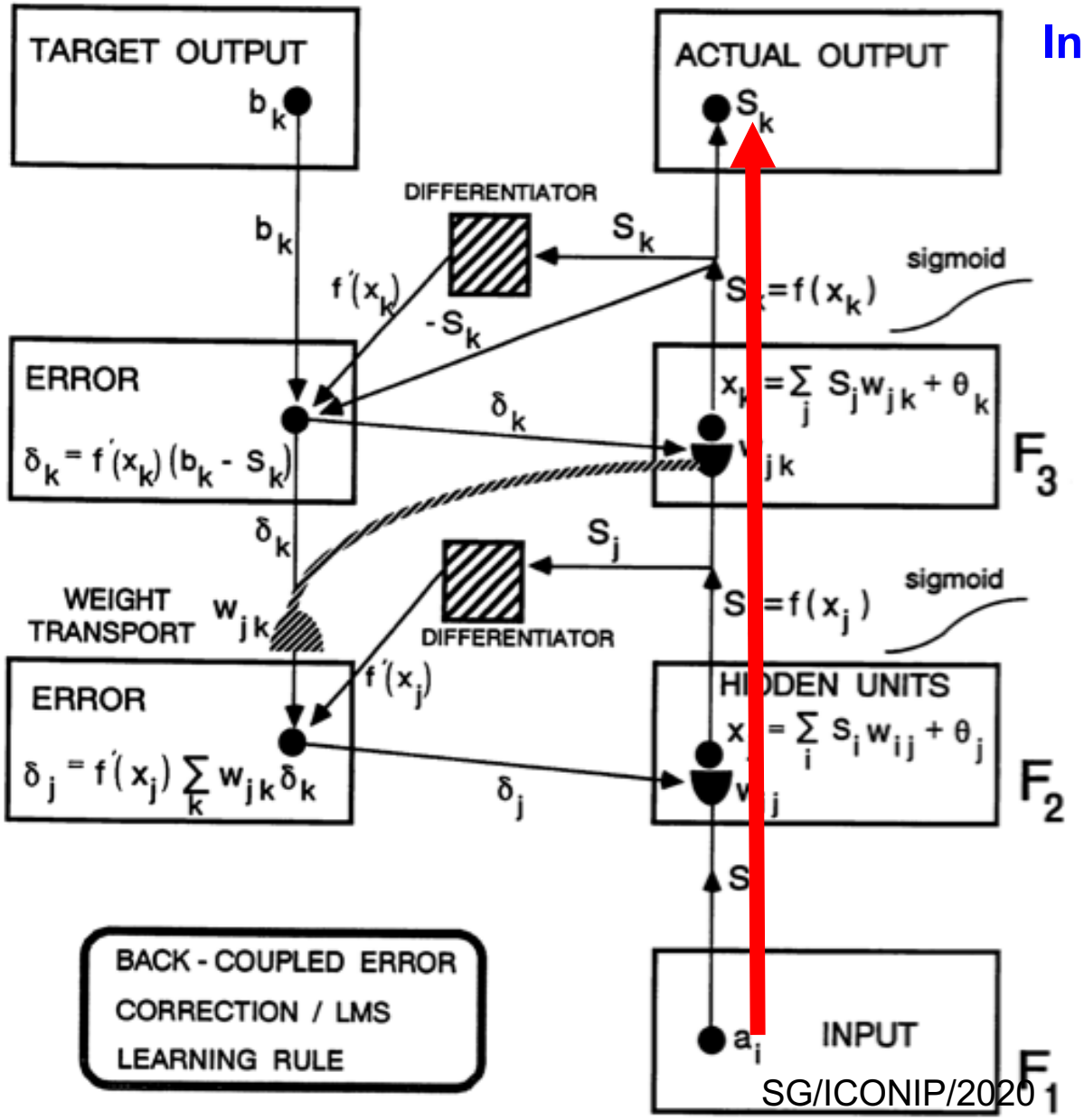
Major contributors include:
Amari (1972), Werbos (1974), Parker (1985)

BP reached its modern form with simulated applications in
Werbos (1974)

It was popularized by
Rumelhart, Hinton, and Williams (1986)

BACK PROPAGATION CIRCUIT

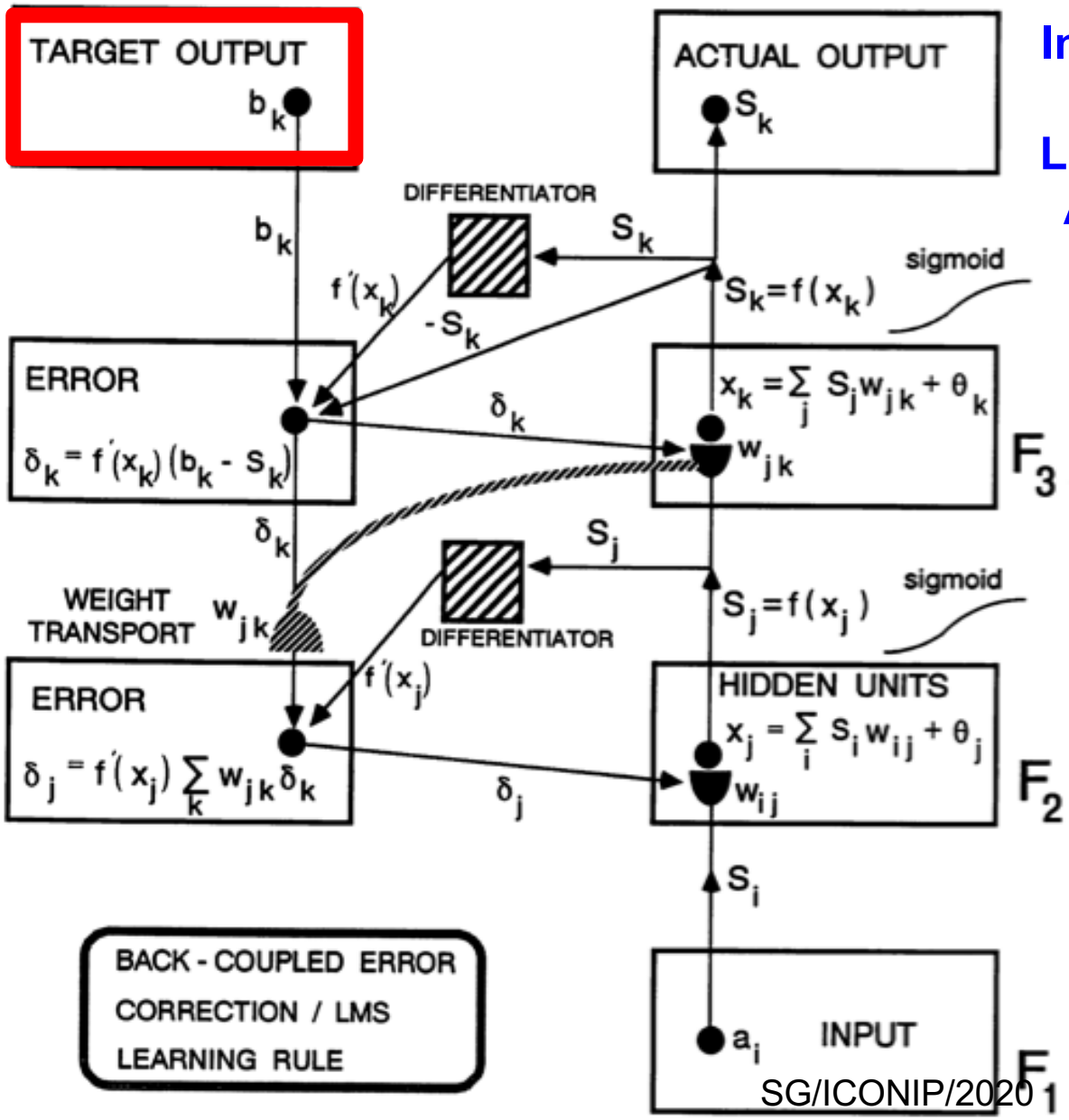
figure reprinted from Carpenter (1989)



Information flows FEEDFORWARD

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



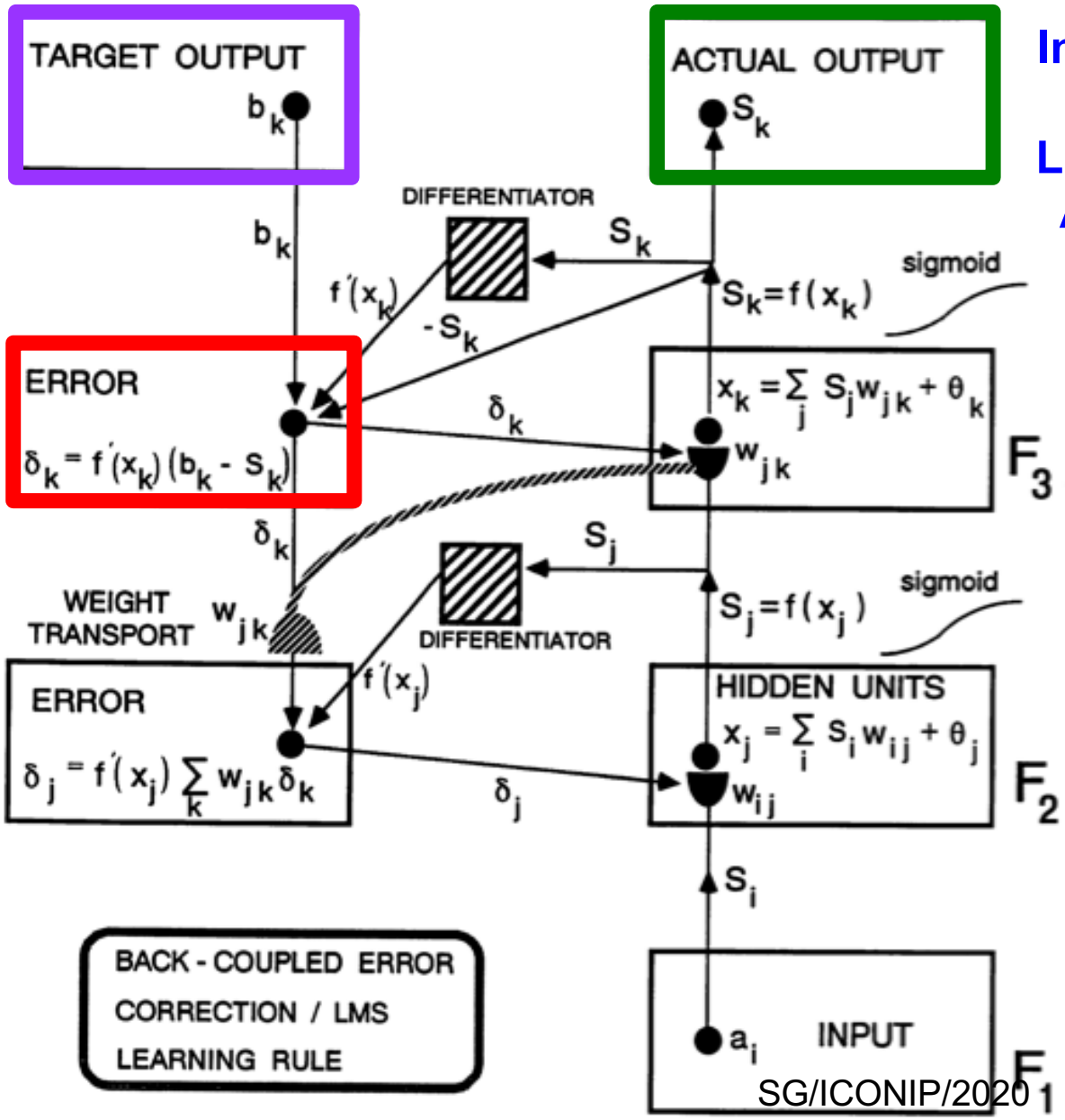
Information flows FEEDFORWARD

Learning is SUPERVISED

An external teacher on each trial

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



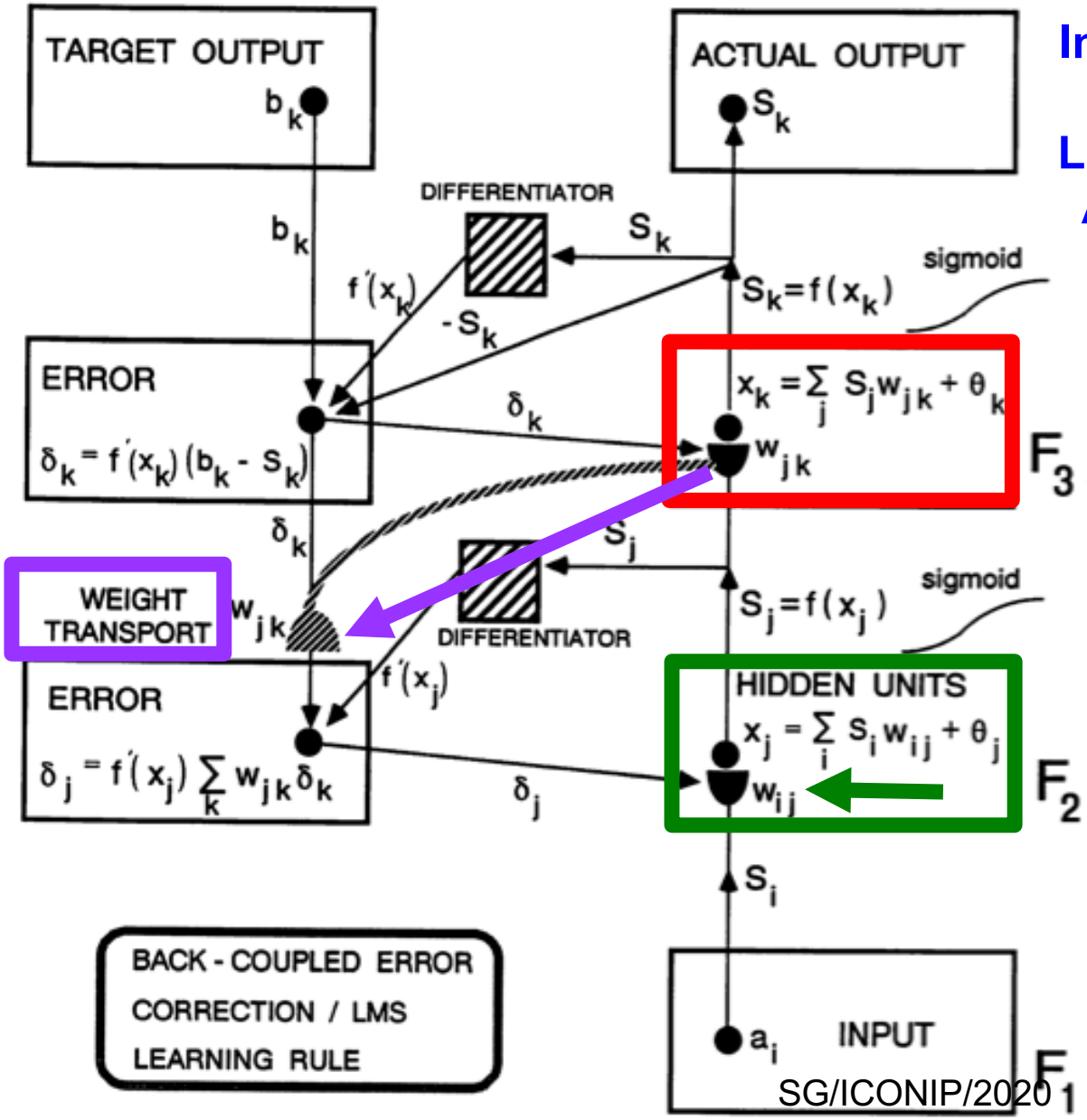
Information flows FEEDFORWARD

Learning is SUPERVISED
An external teacher on each trial

Teaching signal is the ERROR or MISMATCH between ACTUAL and TARGET outputs

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



Information flows FEEDFORWARD

Learning is SUPERVISED
An external teacher on each trial

Teaching signal is the ERROR or MISMATCH between ACTUAL and TARGET outputs

Teaching signal in level F_3 of adaptive weights w_{ij} in level F_2 have no network pathway whereby to reach F_2

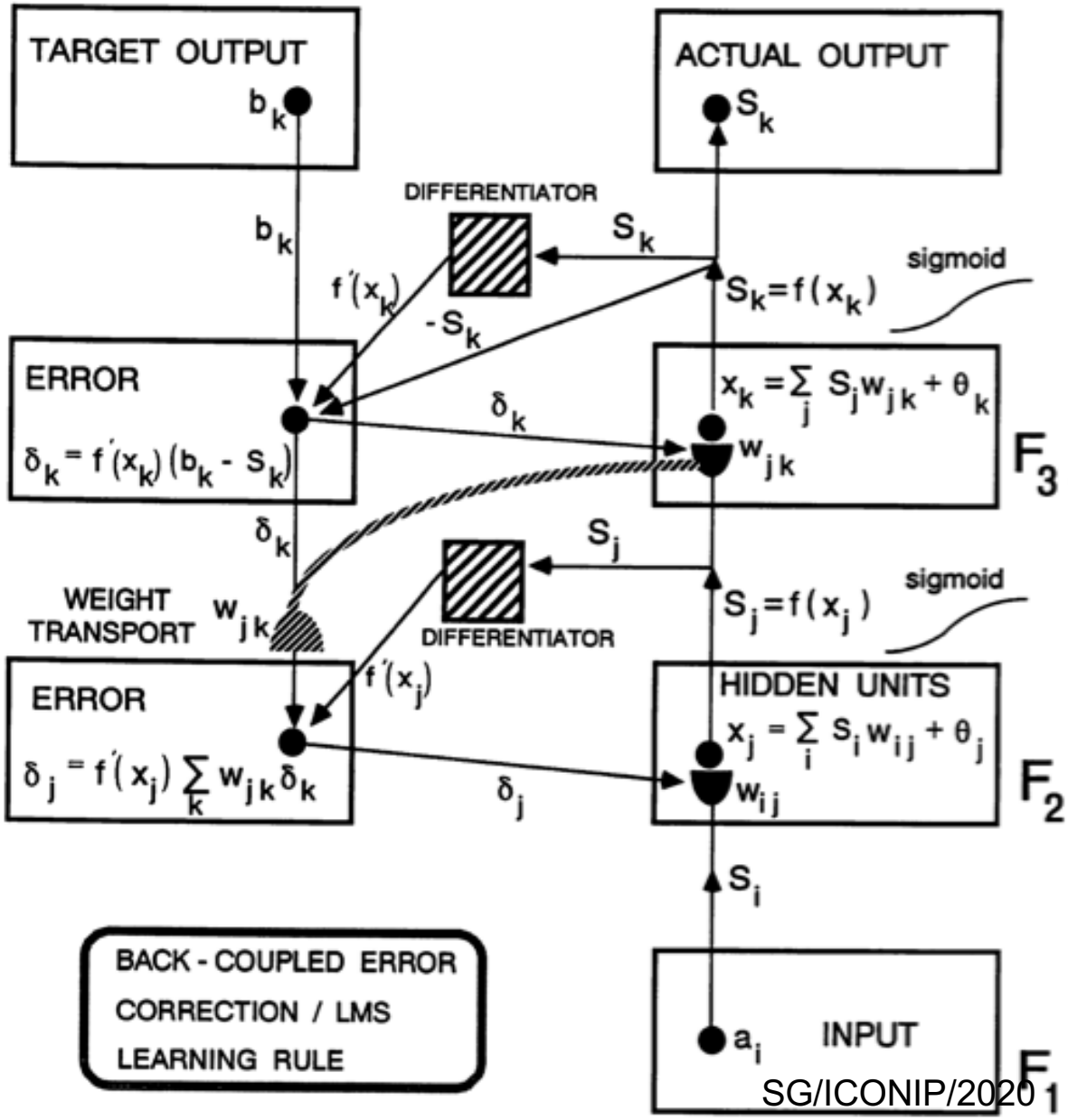
Use WEIGHT TRANSPORT of the teaching signal!

NON-LOCAL!

NON-BIOLOGICAL!

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)

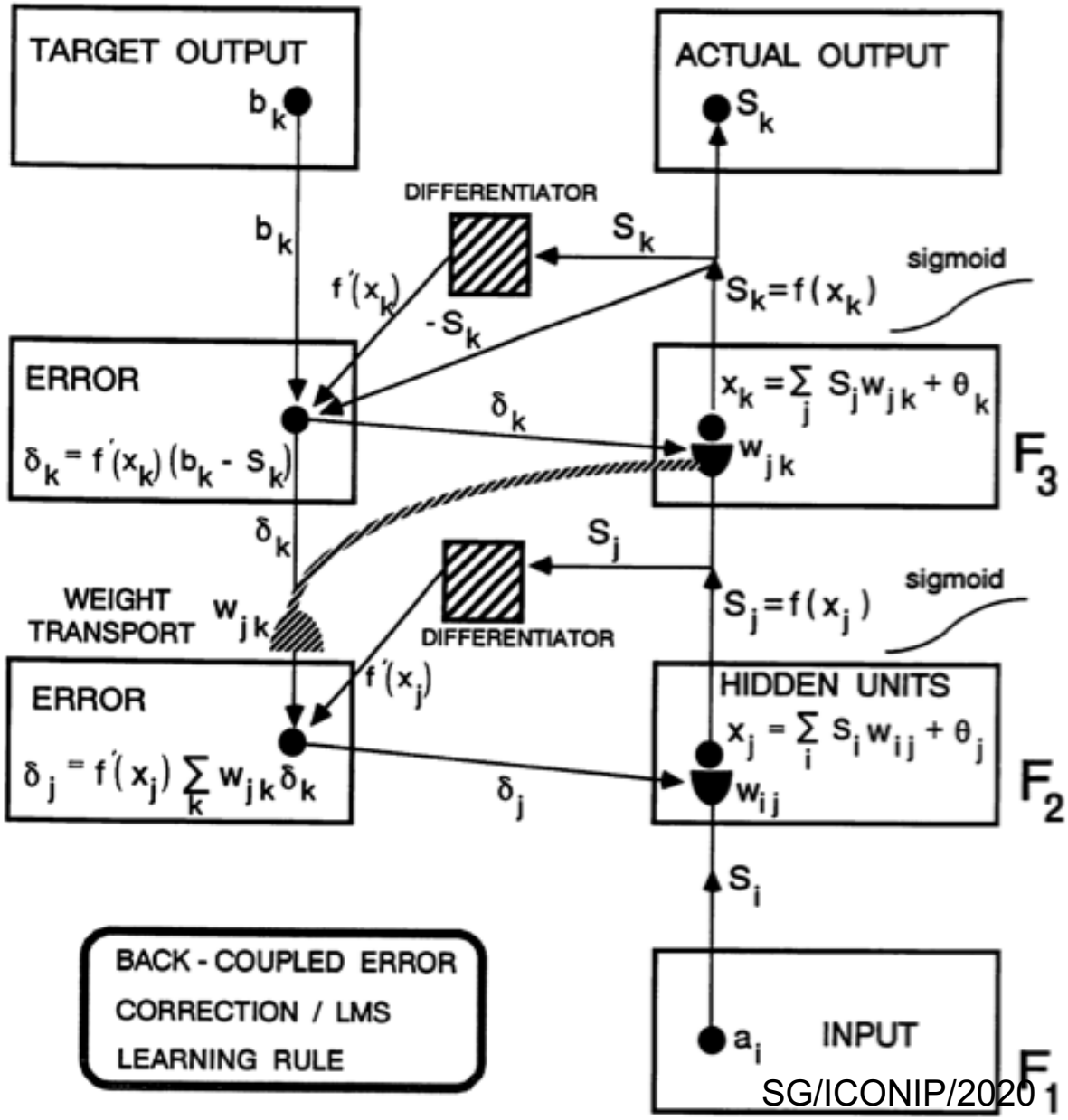


SLOW LEARNING
Adaptive weights change just a little to reduce error on each learning trial

REQUIRES MANY TRIALS
(i.e., repetitions of database) to learn, possibly hundreds or thousands of trials

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



SLOW LEARNING
Adaptive weights change just a little to reduce error on each learning trial

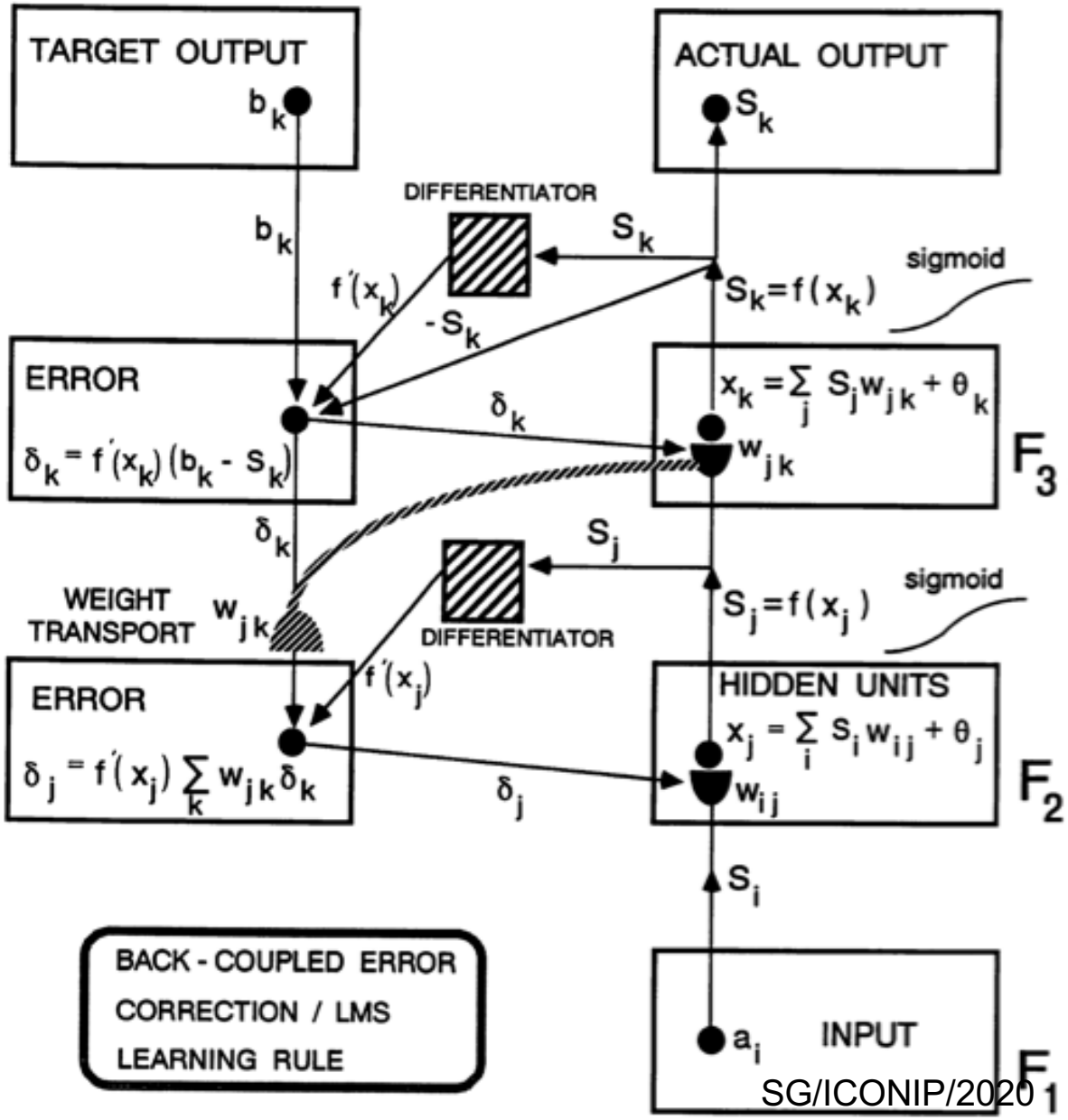
REQUIRES MANY TRIALS
(i.e., repetitions of database) to learn, possibly hundreds or thousands of trials

CONTRAST FAST LEARNING
Adaptive weights zero error signals on EACH trial

Cf. learn a face that you see just once, and remember it for a long time

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



CATASTROPHIC FORGETTING

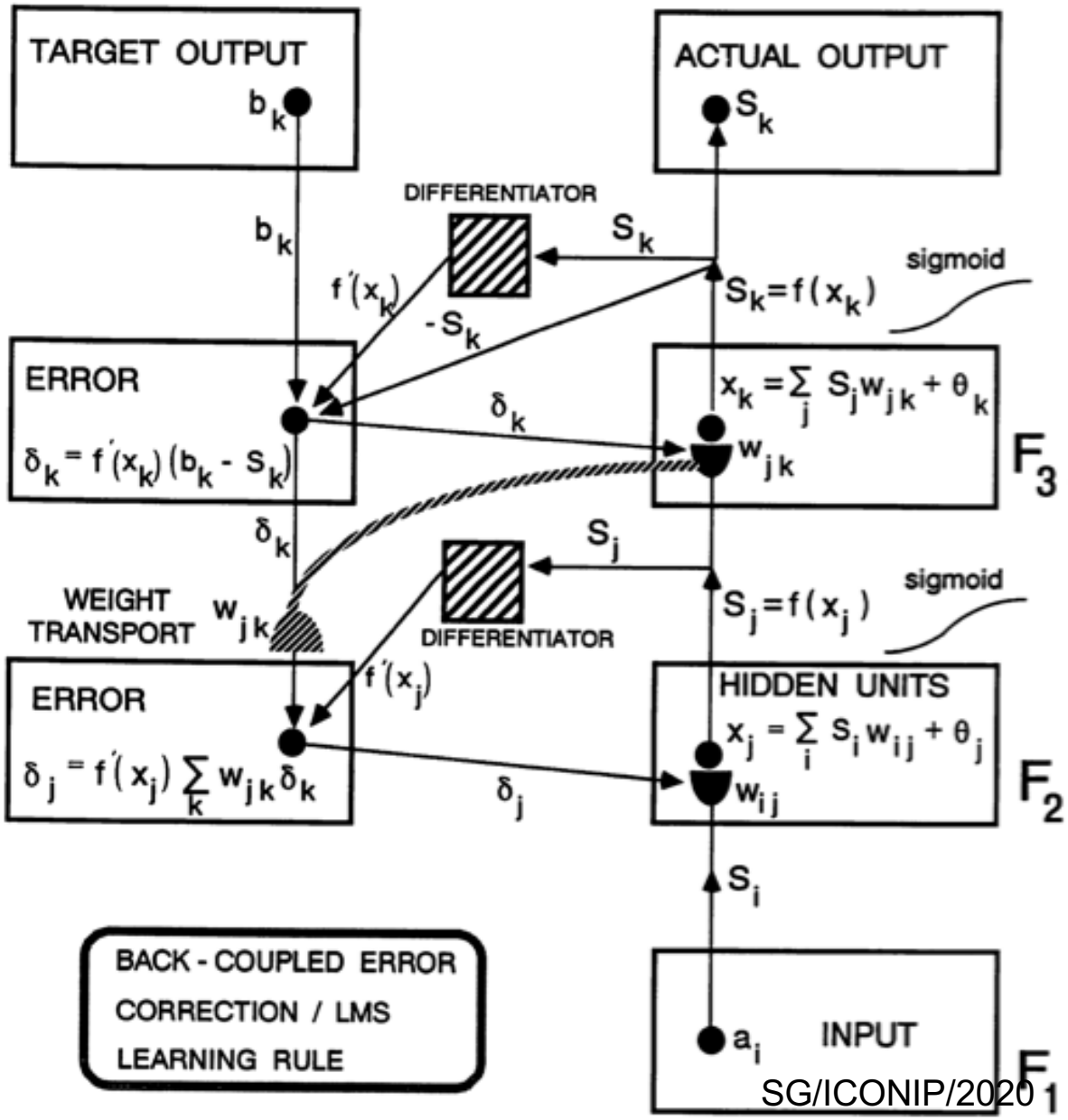
During any learning trial, an unpredictable part of its learned memory can collapse

McCloskey & Cohen (1989)

Ratcliff (1990), French (1999)

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



CATASTROPHIC FORGETTING

During any learning trial, an unpredictable part of its learned memory can collapse

McCloskey & Cohen (1989)

Ratcliff (1990), French (1999)

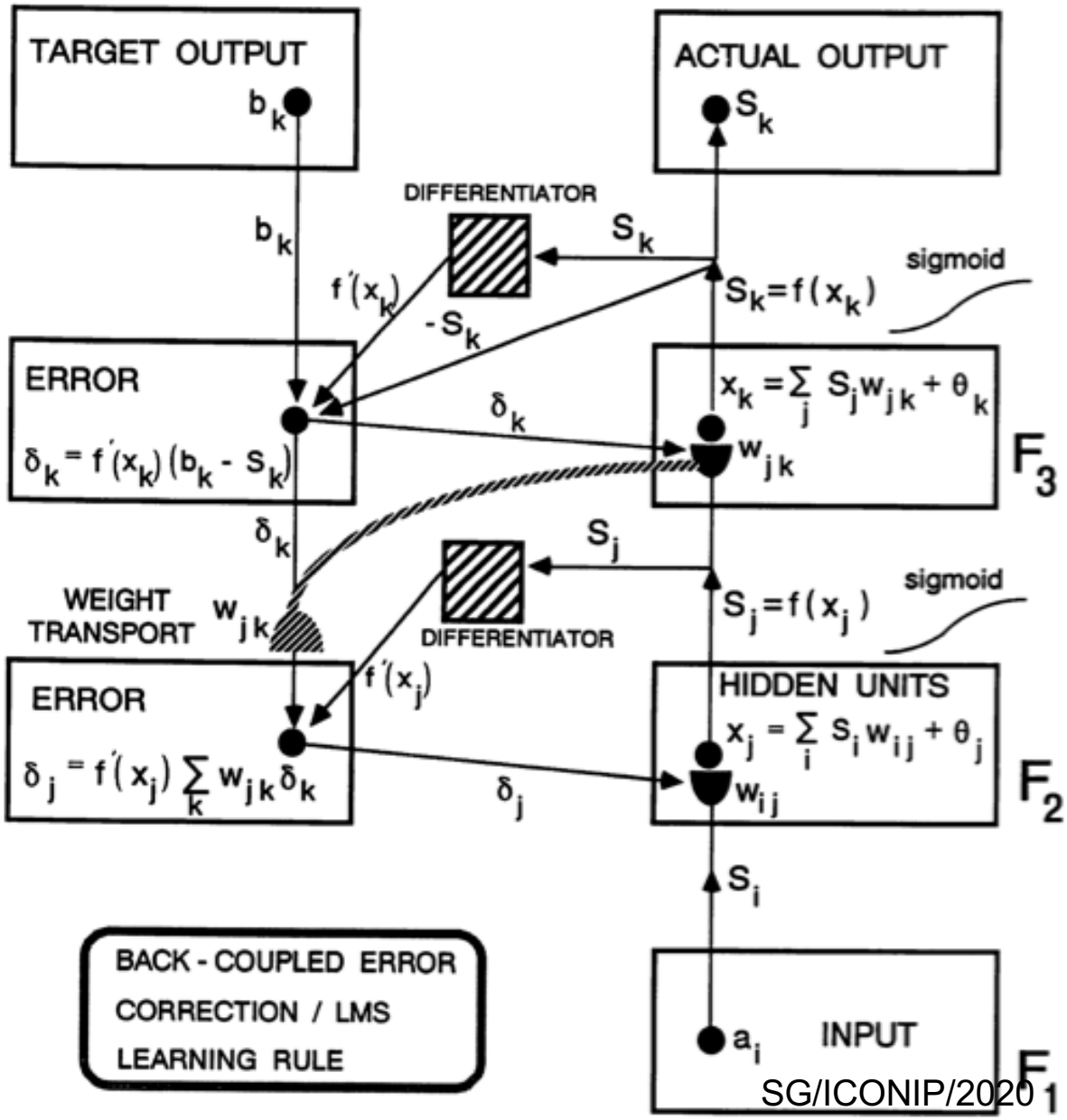
Deep Learning is thus

neither RELIABLE

nor TRUSTWORTHY

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



CATASTROPHIC FORGETTING

During any learning trial, an unpredictable part of its learned memory can collapse

McCloskey & Cohen (1989)
Ratcliff (1990), French (1999)

WHY?

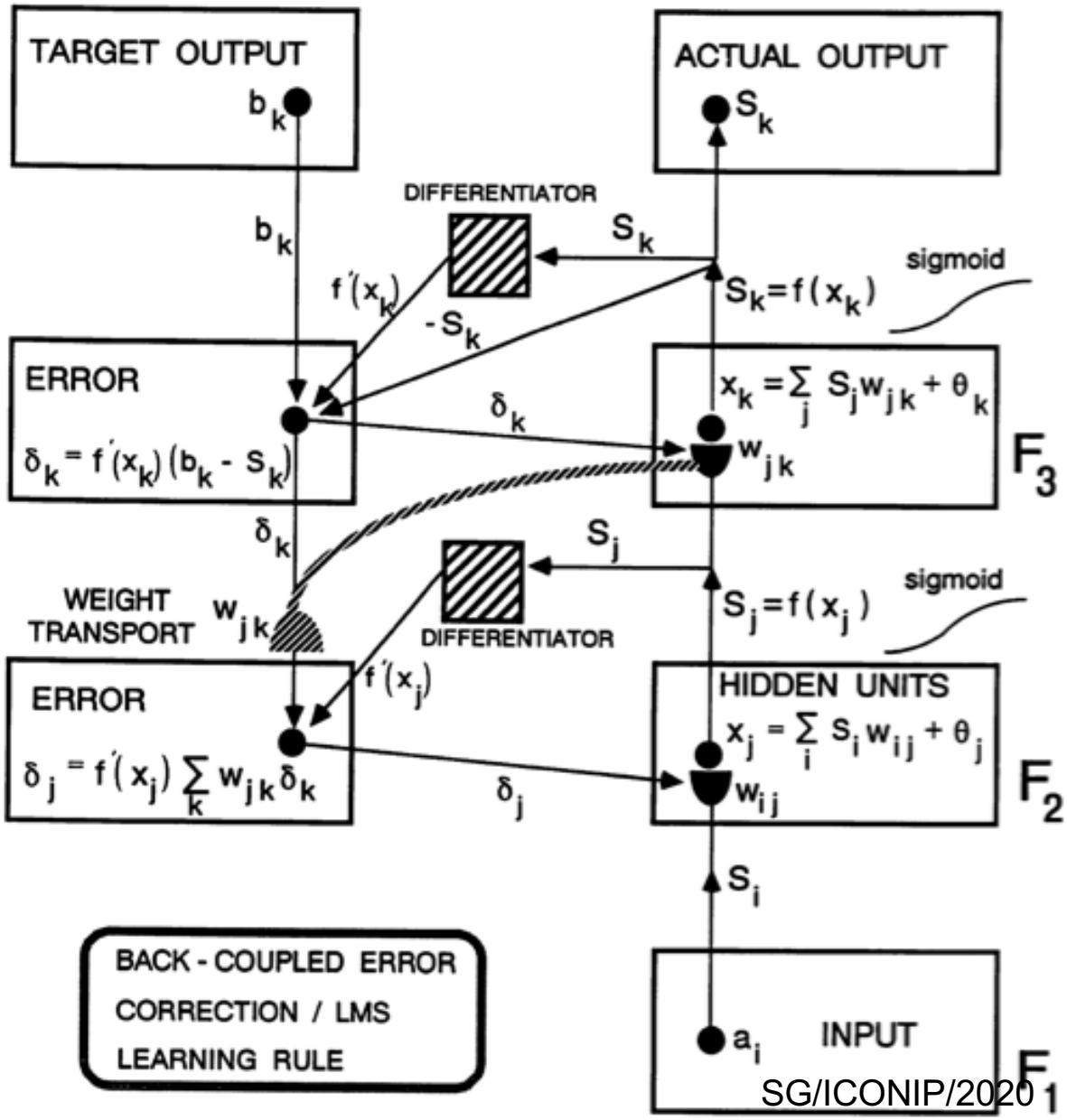
All inputs are processed by a shared set of learned weights

It cannot selectively buffer learned weights that are still predictively useful (no attention)

This problem occurs in ANY learning algorithm whose shared weight updates follow the gradient of the error in response to the current batch of data points, while ignoring past batches

MULTIPLE EFFORTS TO FIX BACK PROPAGATION

figure reprinted from Carpenter (1989)



Selectively slow learning
"on the weights important for…supervised learning and reinforcement learning problems …by optimizing…parameters… using Bayes' rule"
Kirkpatrick et al (2017)

Assumes:
omniscient observer who can discover and alter "important weights"

non-local computations
e.g., Bayesian computation

Same problems with evolutionary algorithms
Clune et al (2013)

and diffusion-based neuromodulation
Velez & Clune (2017)

These efforts to overcome catastrophic forgetting created additional conceptual and computational problems

I view them as adding
EPICYCLES
to ameliorate a fundamental flaw in the model

Reminiscent of adding epicycles
to correct problems in the
Ptolemaic model of the solar system

The Copernican model that we now accept
did not require epicycles!

Perhaps this is why
Geoffrey Hinton said in
AXIOS (LeVine, 2017)
that he is

"deeply suspicious of back propagation…
I don't think it's how the brain works.
We clearly don't need all the labeled data…
My view is,
throw it all away and start over"

Perhaps this is why
Geoffrey Hinton said in
AXIOS (LeVine, 2017)
that he is

"deeply suspicious of back propagation…
I don't think it's how the brain works.
We clearly don't need all the labeled data…
My view is,
throw it all away and start over"

We do not have to start over!

These problems were solved in the 1970s and 1980s!

17 PROBLEMS OF BACK PROPAGATION OVERCOME BY ADAPTIVE RESONANCE

Grossberg (1988, *Neural Networks*, 1, 17-41)

- Real-time (on-line) learning vs. lab-time (off-line) learning
- Learning in nonstationary unexpected world vs. in stationary controlled world
- Self-organized unsupervised or supervised learning vs. supervised learning
- Dynamically self-stabilize learning to arbitrarily many inputs vs. catastrophic forgetting
- Maintain plasticity forever vs. externally shut off learning when database gets too large
- Effective learning of arbitrary databases vs. statistical restrictions on learnable data
- Learn internal expectations vs. impose external cost functions
- Actively focus attention to selectively learn critical features vs. passive weight change
- Closing vs. opening the feedback loop between fast signaling and slower learning
- Top-down priming and selective processing vs. activation of all memory resources
- Match learning vs. mismatch learning: Avoiding the noise catastrophe
- Fast and slow learning vs. only slow learning: Avoiding the oscillation catastrophe
- Learning guided by hypothesis testing and memory search vs. passive weight change
- Direct access to globally best match vs. local minima
- Asynchronous learning vs. fixed duration learning: A cost of unstable slow learning
- Autonomous vigilance control vs. unchanging sensitivity during learning
- General-purpose self-organizing production system vs. passive adaptive filter

Perhaps this is why
Geoffrey Hinton said in
AXIOS (LeVine, 2017)
that he is

"deeply suspicious of back propagation…
I don't think it's how the brain works.
We clearly don't need all the labeled data…
My view is,
throw it all away and start over"

We do not have to start over!

These problems were solved in the 1970s and 1980s!

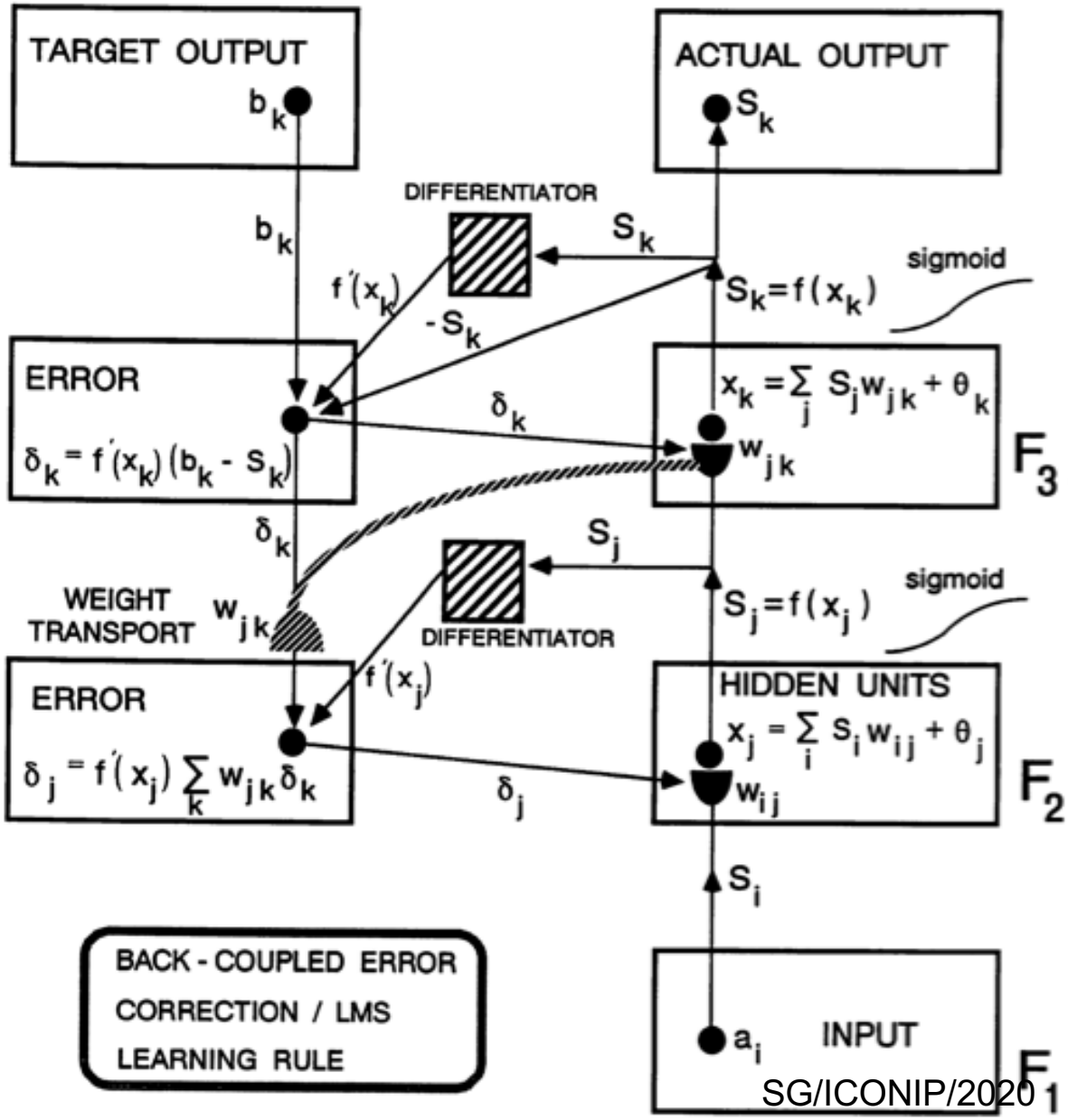
17 PROBLEMS OF BACK PROPAGATION OVERCOME BY ADAPTIVE RESONANCE

Grossberg (1988, *Neural Networks*, 1, 17-41)

- Real-time (on-line) learning vs. lab-time (off-line) learning
- Learning in nonstationary unexpected world vs. in stationary controlled world
- Self-organized unsupervised or supervised learning vs. supervised learning LABELS!
- Dynamically self-stabilize learning to arbitrarily many inputs vs. catastrophic forgetting
- Maintain plasticity forever vs. externally shut off learning when database gets too large
- Effective learning of arbitrary databases vs. statistical restrictions on learnable data
- Learn internal expectations vs. impose external cost functions
- Actively focus attention to selectively learn critical features vs. passive weight change
- Closing vs. opening the feedback loop between fast signaling and slower learning
- Top-down priming and selective processing vs. activation of all memory resources
- Match learning vs. mismatch learning: Avoiding the noise catastrophe
- Fast and slow learning vs. only slow learning: Avoiding the oscillation catastrophe
- Learning guided by hypothesis testing and memory search vs. passive weight change
- Direct access to globally best match vs. local minima
- Asynchronous learning vs. fixed duration learning: A cost of unstable slow learning
- Autonomous vigilance control vs. unchanging sensitivity during learning
- General-purpose self-organizing production system vs. passive adaptive filter

BACK PROPAGATION CIRCUIT

figure reprinted from Carpenter (1989)



SLOW LEARNING

Adaptive weights change just a little to reduce error on each learning trial

REQUIRES MANY TRIALS (i.e., repetitions of database) to learn, possibly hundreds or thousands of trials

CONTRAST FAST LEARNING

Adaptive weights zero error signals on EACH trial

Cf. learn a face that you see just once, and remember it for a long time

VERSUS…

17 PROBLEMS OF BACK PROPAGATION OVERCOME BY ADAPTIVE RESONANCE

Grossberg (1988, *Neural Networks*, 1, 17-41)

- Real-time (on-line) learning vs. lab-time (off-line) learning
- Learning in nonstationary unexpected world vs. in stationary controlled world
- Self-organized unsupervised or supervised learning vs. supervised learning
- Dynamically self-stabilize learning to arbitrarily many inputs vs. catastrophic forgetting
- Maintain plasticity forever vs. externally shut off learning when database gets too large
- Effective learning of arbitrary databases vs. statistical restrictions on learnable data
- Learn internal expectations vs. impose external cost functions
- Actively focus attention to selectively learn critical features vs. passive weight change
- Closing vs. opening the feedback loop between fast signaling and slower learning
- Top-down priming and selective processing vs. activation of all memory resources
- Match learning vs. mismatch learning: Avoiding the noise catastrophe
- Fast and slow learning vs. only slow learning: Avoiding the oscillation catastrophe
- Learning guided by hypothesis testing and memory search vs. passive weight change
- Direct access to globally best match vs. local minima
- Asynchronous learning vs. fixed duration learning: A cost of unstable slow learning
- Autonomous vigilance control vs. unchanging sensitivity during learning
- General-purpose self-organizing production system vs. passive adaptive filter

ART can learn to classify an entire database using fast learning on a single learning trial

Carpenter and Grossberg (1987, 1988)

ART OVERCOMES ALL 17 PROBLEMS OF BP

without EPICYCLES!

Moreover...

All the core ART predictions have been supported by subsequent psychological and neurobiological data

ART is a principled biological and technological THEORY

ART has explained data from hundreds of experiments

ART has made scores of predictions that have subsequently received experimental support

All the core ART predictions have been supported by subsequent psychological and neurobiological data

ART is a principled biological and technological THEORY

ART has explained data from hundreds of experiments

ART has made scores of predictions that have subsequently received experimental support

Why is ART so successful?

ART CAN BE DERIVED FROM A THOUGHT EXPERIMENT ABOUT A UNIVERSAL PROBLEM IN ERROR CORRECTION

Grossberg (1980, *Psychological Review*, 87, 1-51)

The thought experiment asks the question:

How can a coding error be corrected if no individual cell knows that one has occurred?

"The importance of this issue becomes clear when we realize that erroneous cues can accidentally be incorporated into a code when our interactions with the environment are simple and will only become evident when our environmental expectations become more demanding.

Even if our code perfectly matched a given environment, we would certainly make errors as the environment itself fluctuates"

AUTONOMOUS LOCAL LEARNING IN A CHANGING WORLD

ART CAN BE DERIVED FROM A THOUGHT EXPERIMENT ABOUT A UNIVERSAL PROBLEM IN ERROR CORRECTION

Grossberg (1980, *Psychological Review*, 87, 1-51)

AUTONOMOUS LOCAL LEARNING IN A CHANGING WORLD

A purely logical inquiry into error correction
is translated at every step of the thought experiment into processes
learning autonomously in real time with only locally computed quantities

The thought experiment uses familiar environmental facts
about how we learn as its hypotheses
ART circuits naturally emerge

ART circuits may thus, in some form, be embodied in all future autonomous adaptive intelligent devices, whether biological or artificial

ART has, probably for this reason, already been used in many large-scale engineering and technological applications

EARLY ARTMAP BENCHMARK STUDIES

Database benchmark:

MACHINE LEARNING (90-95% correct)

ARTMAP (100% correct on a training set an order of magnitude smaller)

Database benchmarks:

BACKPROPAGATION (10,000 – 20,000 training epochs)

ARTMAP (1-5 epochs)

Medical database:

STATISTICAL METHOD (60% correct)

ARTMAP (96% correct)

Letter recognition database:

GENETIC ALGORITHM (82% correct)

ARTMAP (96% correct)

Used in applications where other algorithms fail

e.g. Boeing CAD Group Technology

Part design reuse and inventory compression

Need fast learning and stable memory to learn and search a huge (16 million 1 million dimensional vectors) and continually growing non-stationary parts inventory

ART WORKS!

Large-scale applications in engineering and technology

techlab.bu.edu

Boeing parts design retrieval (used to design Boeing 777)
satellite remote sensing
radar identification
robot sensory-motor control and navigation
machine vision
3D object and face recognition
Macintosh operating system software
automatic target recognition
ECG wave recognition
protein secondary structure identification
character classification
musical analysis
air quality monitoring and weather prediction
medical imaging and database analysis
multi-sensor chemical analysis
strength prediction for concrete mixes
signature verification
decision making and intelligent agents
machine condition monitoring and failure forecasting
chemical analysis
electromagnetic and digital circuit design

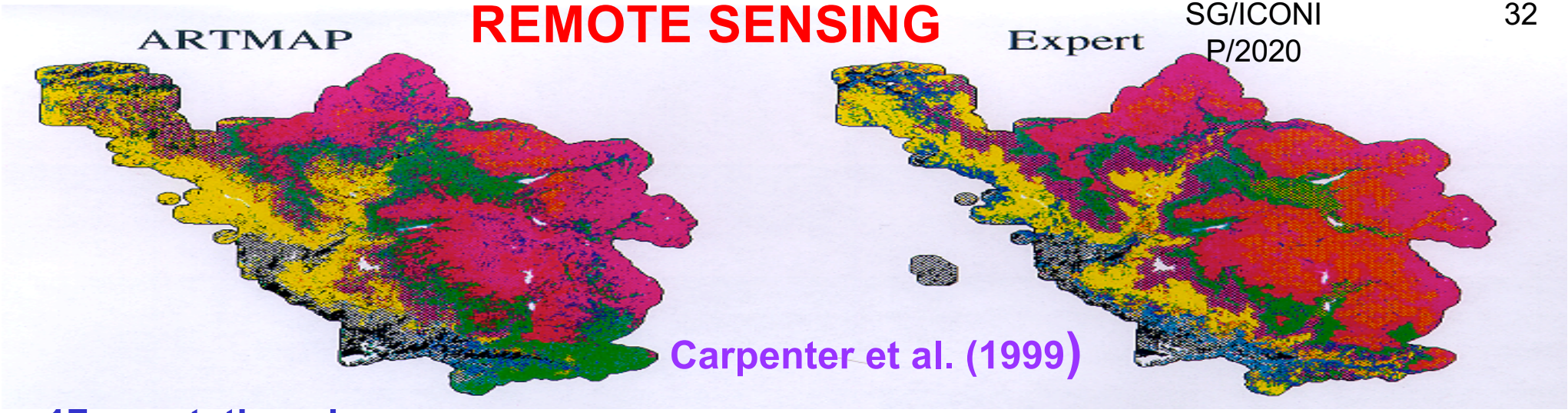
ART WORKS!

Large-scale applications in engineering and technology

techlab.bu.edu

Boeing parts design retrieval (used to design Boeing 777)
satellite remote sensing
radar identification
robot sensory-motor control and navigation
machine vision
3D object and face recognition
Macintosh operating system software
automatic target recognition
ECG wave recognition
protein secondary structure identification
character classification
musical analysis
air quality monitoring and weather prediction
medical imaging and database analysis
multi-sensor chemical analysis
strength prediction for concrete mixes
signature verification
decision making and intelligent agents
machine condition monitoring and failure forecasting
chemical analysis
electromagnetic and digital circuit design

REMOTE SENSING



Carpenter et al. (1999)

17 vegetation classes

AI Expert system – 1 year

Field identification of natural regions

Derivation of ad hoc rules for each region, by expert geographers

Correct 80,000 of 250,000 site labels

230*m* (site–level) scale

ARTMAP system – 1 day

Rapid, automatic, no natural regions or rules

Confidence map

30*m* (pixel–level) scale: can see roads

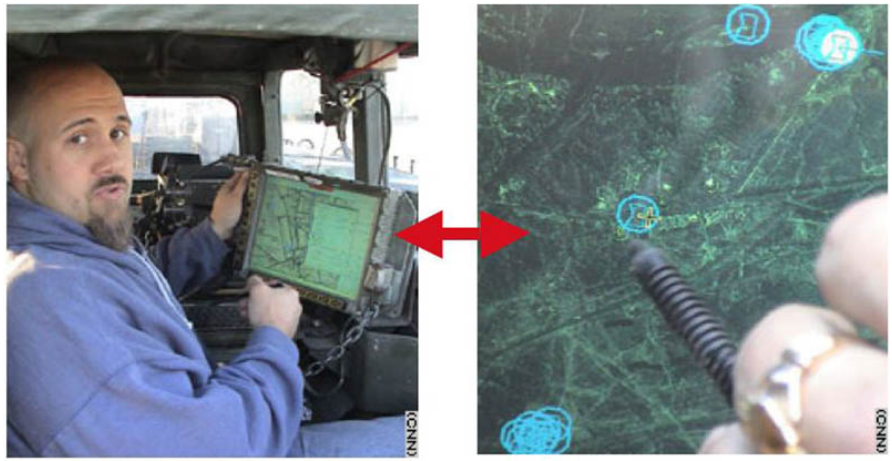
Equal accuracy at test sites

INFORMATION FUSION IN REMOTE SENSING

Carpenter et al. (2004)

Multimodal integration of information from many sources to learn a knowledge structure:

- CONSISTENT
- STABLE
- ROBUST
- LEARNED ONLINE
- SELF-ORGANIZED



SOURCE 1
GOAL 1
SENSOR 1
TIME 1

SOURCE 2
GOAL 2
SENSOR 2
TIME 2

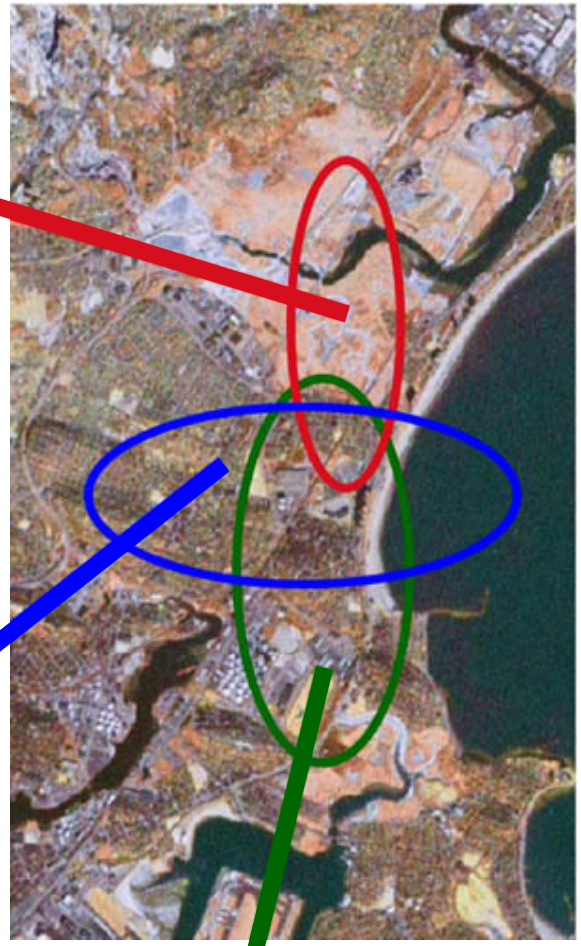
SOURCE 3
GOAL 3
SENSOR 3
TIME 3

Boston testbed

CONSISTENT KNOWLEDGE FROM INCONSISTENT DATA

Automatically learns and stably stores one-to-many mappings

water
open space
built-up



ocean
beach
park
ice
road
river
residential
industrial

man-made
natural

PROBLEM: Integrate multiple sources into a coherent knowledge structure

Solution 1:
HUMAN MAPPING EXPERT:
Slow, expensive, possibly unavailable

Solution 2:
Distributed ARTMAP MODEL:
Fast, automatic, easy to deploy
NO PRIOR RULES OR DOMAIN KNOWLEDGE

Self-organizing expert system

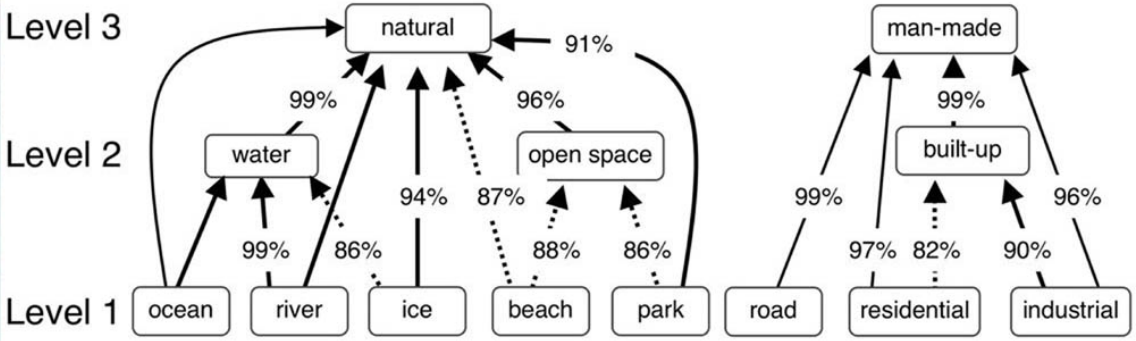
SELF-ORGANIZES a HIERARCHY of COGNITIVE RULES

Distributed predictions across test set pixels →

ocean
beach
park
ice
road
river
water
open space
residential
industrial
built-up
natural
man-made

Boston testbed

RULE DISCOVERY



Confidence in each rule = 100%, except where noted

CONSISTENT MAPS, LABELED BY LEVEL

ART WORKS!

Large-scale applications in engineering and technology

Some more recent work about ART:

Special issue of *Neural Networks* in December, 2019:

Wunsch, D. C. II. (2019). Admiring the Great Mountain: A Celebration Special Issue in Honor of Stephen Grossberg's 80th Birthday.

arxiv.org/pdf/1910.13351.pdf

da Silva, L. E.. B., Elnabarawy, I., & Wunsch, D. C. II. (2019). A Survey of Adaptive Resonance Theory Neural Network Models for Engineering Applications.

arxiv.org/pdf/1910.13351.pdf

BP is a feedforward adaptive filter

ART is more than a feedforward adaptive filter

ART IS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM IN A NON-STATIONARY WORLD

BP is a feedforward adaptive filter

ART is more than a feedforward adaptive filter

ART IS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM IN A NON-STATIONARY WORLD

It is SELF-ORGANIZING because it can autonomously carry out arbitrary combinations of unsupervised or supervised learning trials with the world as its only teacher

BP is a feedforward adaptive filter

ART is more than a feedforward adaptive filter

ART IS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM IN A NON-STATIONARY WORLD

It is a PRODUCTION SYSTEM because it uses HYPOTHESIS TESTING to discover and learn RULES via a top-down matching process that focuses attention on CRITICAL FEATURE PATTERNS that predict behavioral success while suppressing irrelevant features

BP is a feedforward adaptive filter

ART is more than a feedforward adaptive filter

ART IS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM IN A NON-STATIONARY WORLD

It is EXPLAINABLE using both its activities, or short term memory (STM) traces and adaptive weights, or long term memory (LTM) traces:

Observing the STM TRACES in a critical feature pattern explain what recognition categories code and what features predict goal-oriented actions

The LTM TRACES in fuzzy ARTMAP translate into fuzzy IF-THEN rules that code what features, in what numerical ranges, control predictions

ART MECHANISMS THAT DEFINE IT AS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM

include:

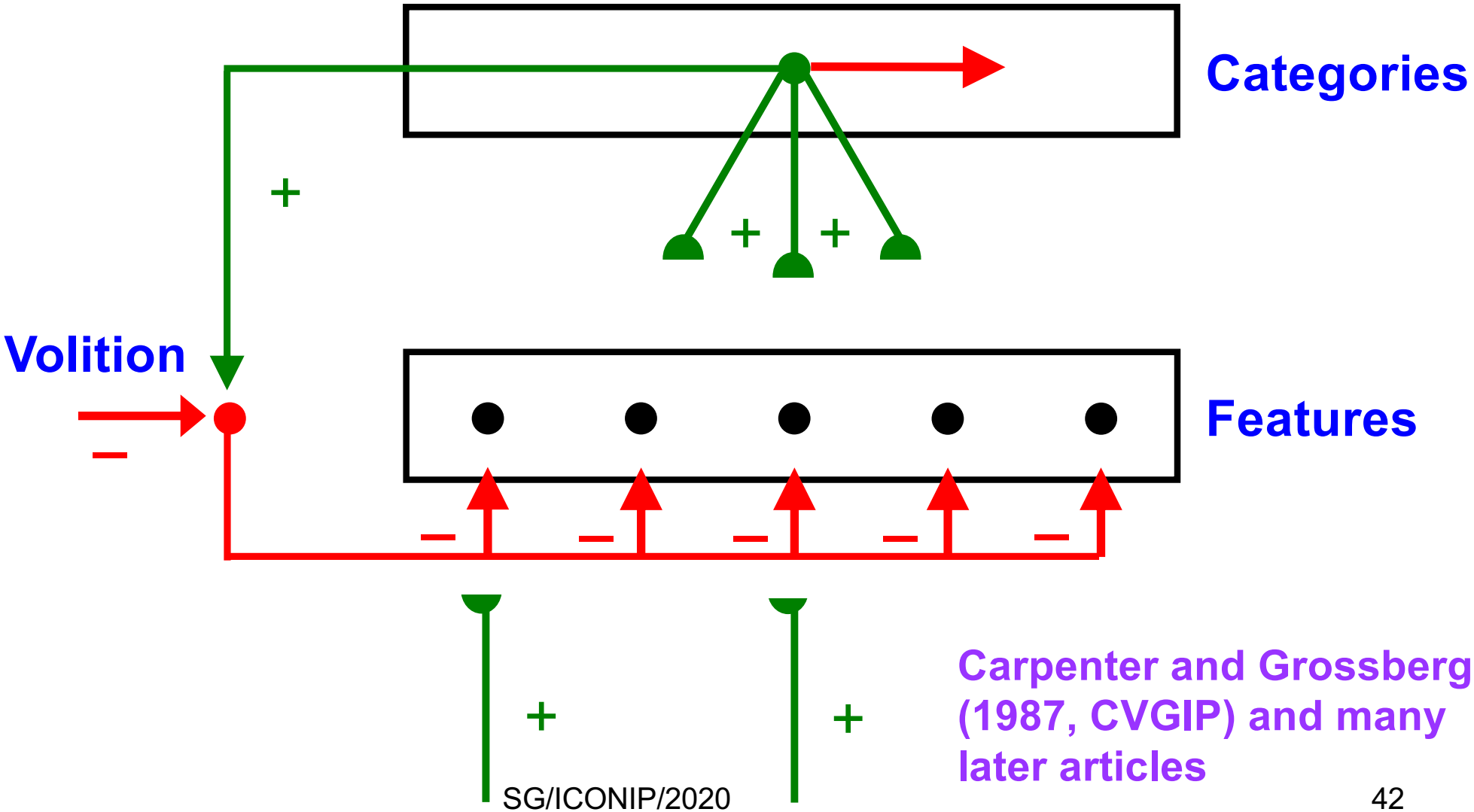
***Bottom-up* adaptive filter (feedforward neural network)**
is supplemented by
top-down learned expectations
and
two types of recurrent inhibitory feedback interactions
that help to choose
recognition categories
and
critical feature patterns

Top-down expectations use the ART MATCHING RULE to learn how to FOCUS ATTENTION on CRITICAL FEATURES that control predictive success

ART MATCHING RULE for OBJECT ATTENTION

stabilizes learning (avoids catastrophic forgetting)

Top-down, modulatory on-center, off-surround network

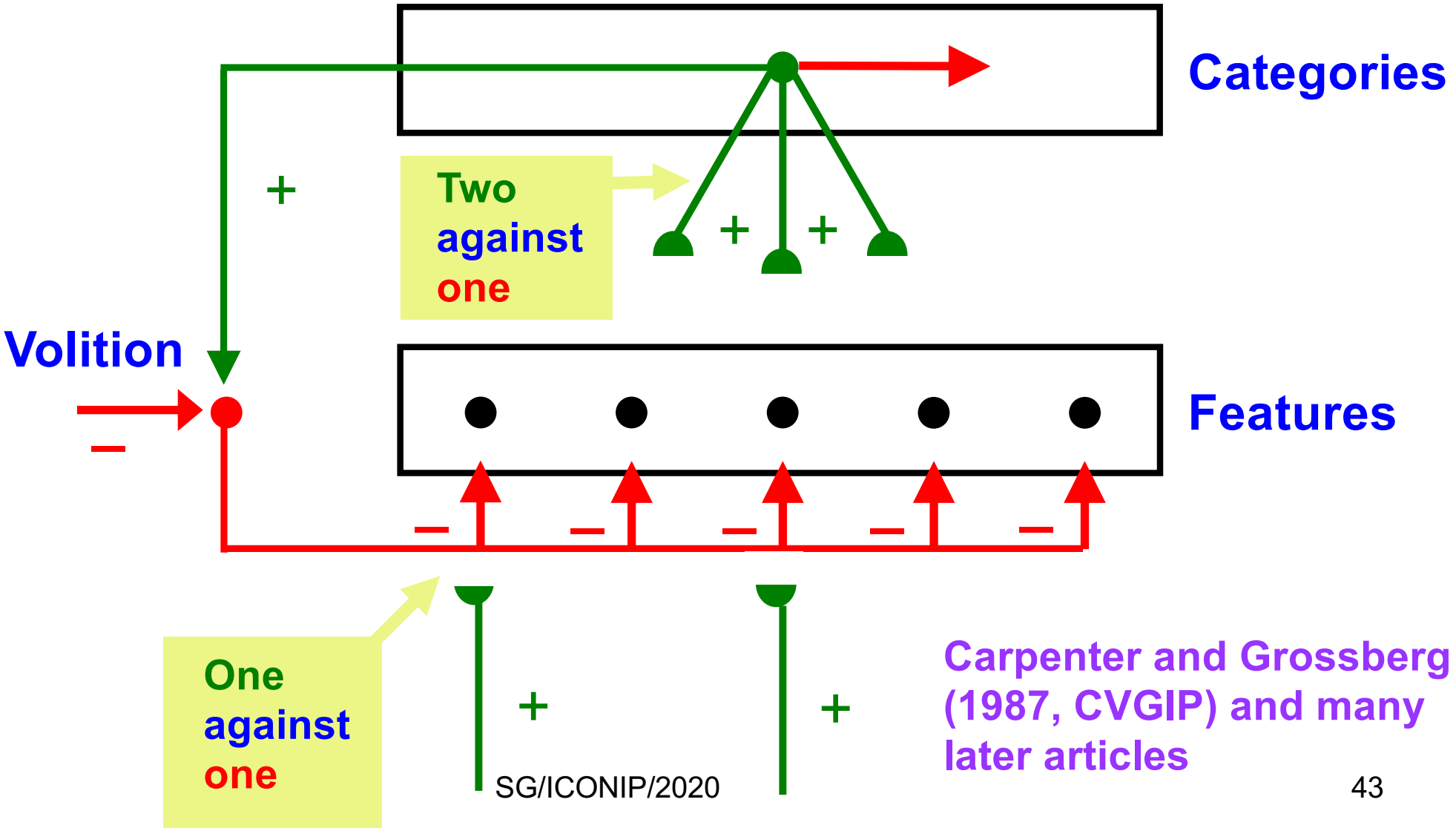


Carpenter and Grossberg (1987, CVGIP) and many later articles

ART MATCHING RULE for OBJECT ATTENTION

stabilizes learning (avoids catastrophic forgetting)

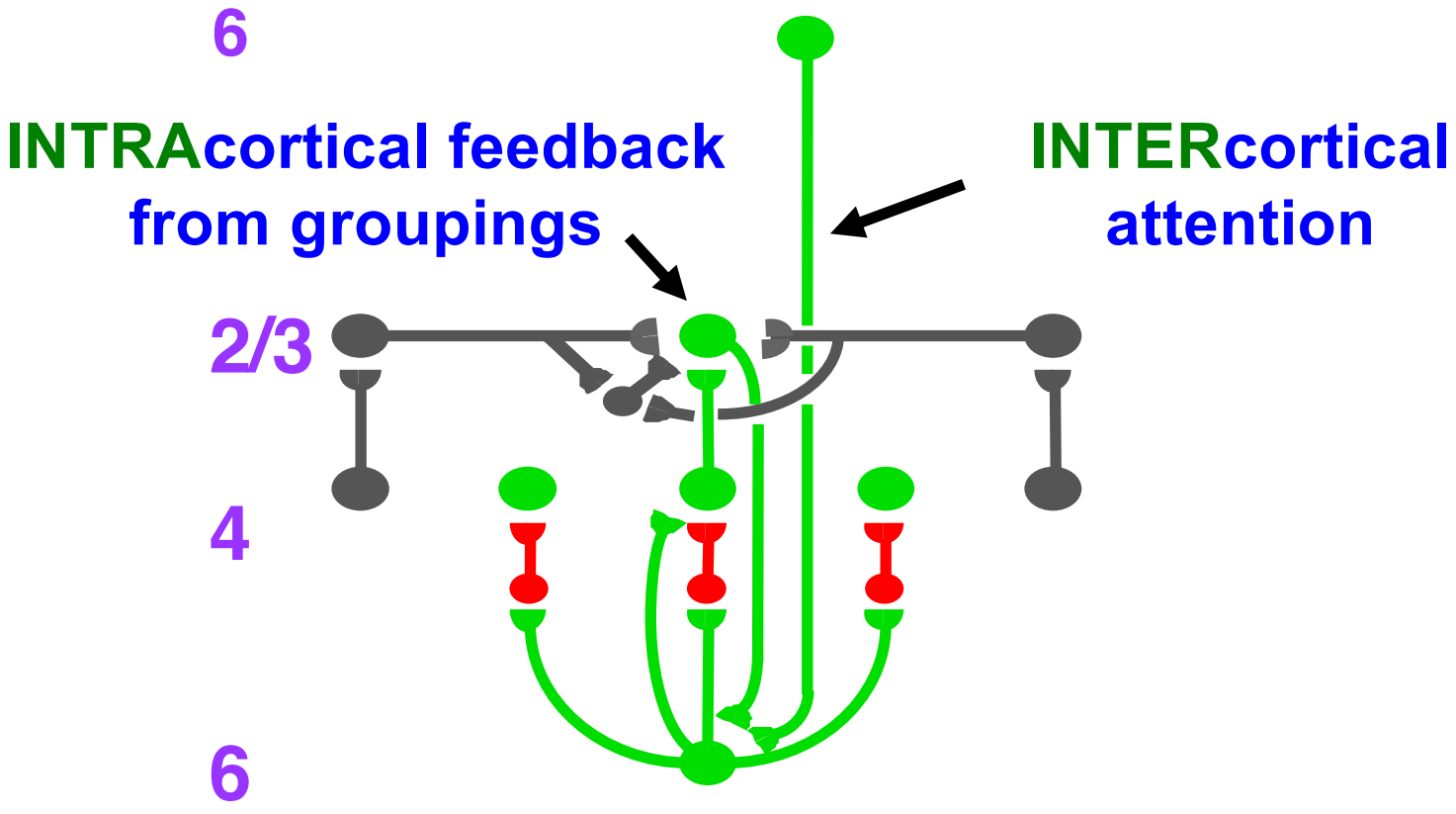
Top-down, modulatory on-center, off-surround network



Carpenter and Grossberg (1987, CVGIP) and many later articles

LAMINAR CORTICAL CIRCUIT FOR OBJECT ATTENTION

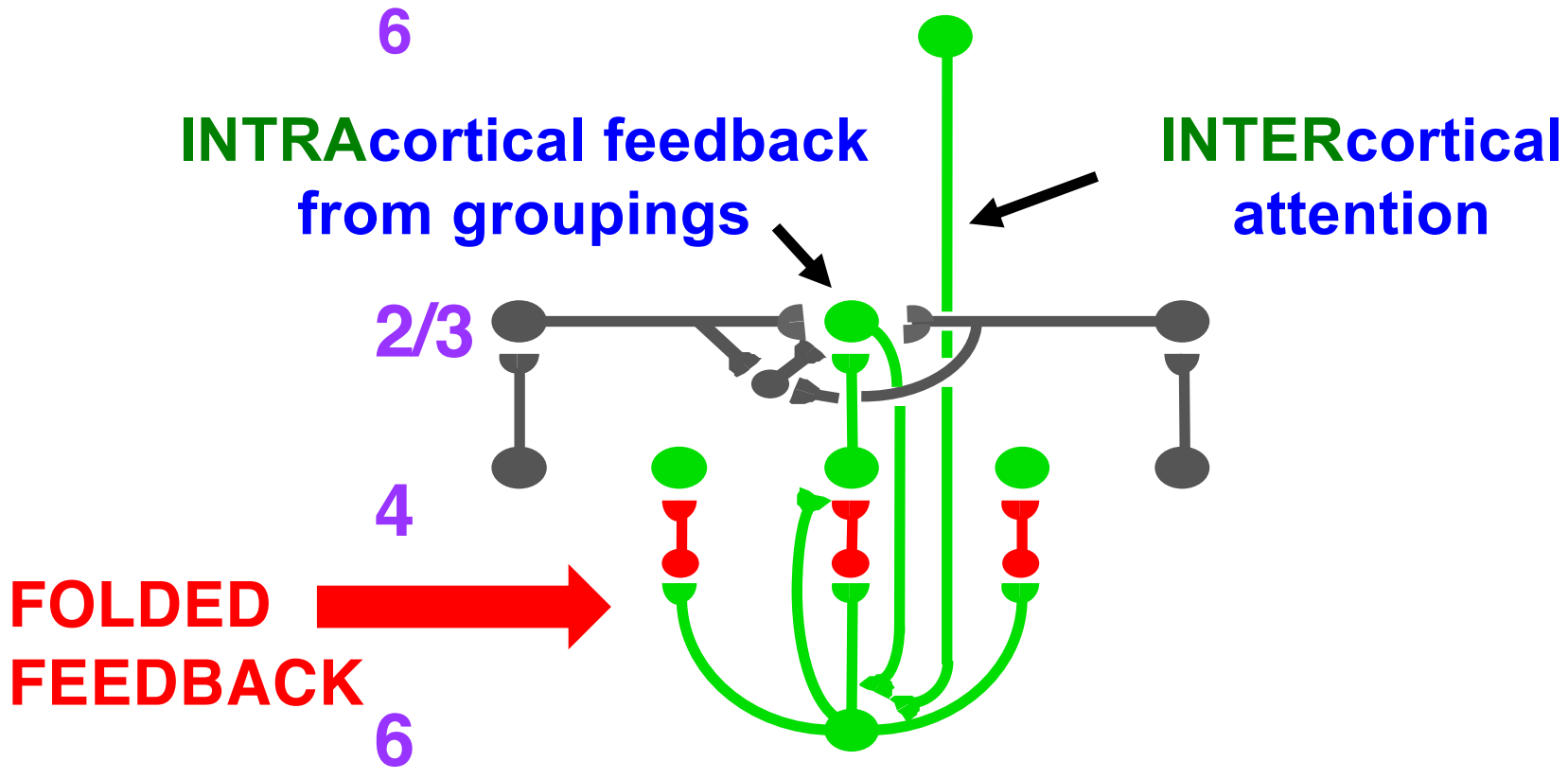
Grossberg (1999, Spatial Vision)



Attention acts via a
TOP-DOWN
MODULATORY ON-CENTER
OFF-SURROUND NETWORK

LAMINAR CORTICAL CIRCUIT FOR OBJECT ATTENTION

Grossberg (1999, Spatial Vision)



Attention acts via a
TOP-DOWN
MODULATORY ON-CENTER
OFF-SURROUND NETWORK

Illustrates NEW PARADIGMS for brain computing

INDEPENDENT MODULES

Computer Metaphor

COMPLEMENTARY COMPUTING

What is the nature of brain specialization?

LAMINAR COMPUTING

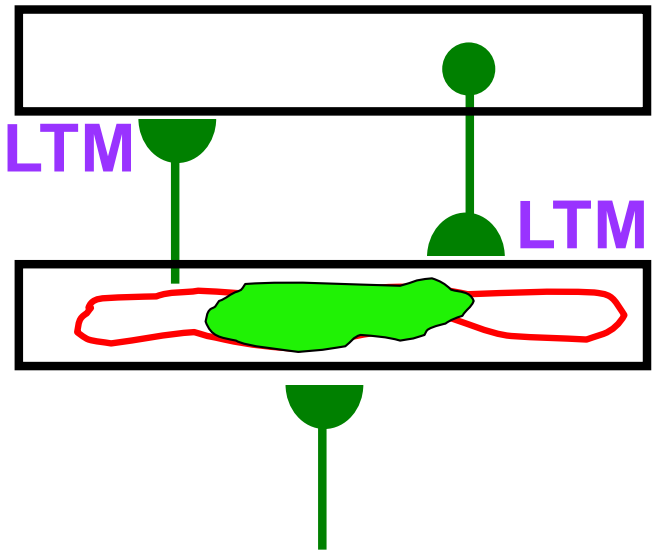
Why are all neocortical circuits organized in layers?

How do laminar circuits give rise to biological intelligence?

ADAPTIVE RESONANCE

Attended feature clusters reactivate bottom-up pathways

Activated categories reactivate their top-down pathways



Categories
STM

Feature Patterns
STM

Feature-category resonance synchronizes
amplifies
prolongs system response

Resonance triggers learning in bottom-up and top-down adaptive weights: *adaptive* resonance!

"ALL CONSCIOUS STATES ARE RESONANT STATES"

Grossberg (1980)

Surface-shroud resonances support conscious seeing of visual qualia

Feature-category resonances support conscious recognition of visual objects and scenes

Stream-shroud resonances support conscious hearing of auditory qualia

Spectral-pitch-and-timbre resonances support conscious recognition of sources in auditory streams

Item-list resonances support conscious recognition of speech and language

Cognitive-emotional resonances support conscious feelings and recognition of them

SUPPORT FOR ART PREDICTIONS

ATTENTION HAS AN ON-CENTER OFF-SURROUND

Bullier, Jupe, James, and Girard, 1996

Caputo and Guerra, 1998

Downing, 1988

Mounts, 2000

Reynolds, Chelazzi, and Desimone, 1999

Smith, Singh, and Greenlee, 2000

Somers, Dale, Seiffert, and Tootell, 1999

Sillito, Jones, Gerstein, and West, 1994

Steinman, Steinman, and Lehmkuhne, 1995

Vanduffell, Tootell, and Orban, 2000

"BIASED COMPETITION"

Desimone, 1998

Kastner and Ungerleider, 2001

SUPPORT FOR ART PREDICTIONS

ATTENTION CAN FACILITATE MATCHED BOTTOM-UP SIGNALS

Hupe, James, Girard, and Bullier, 1997
Luck, Chellazi, Hillyard, and Desimone, 1997
Roelfsema, Lamme, and Spekreijse, 1998
Sillito, Jones, Gerstein, and West, 1994
and many more…

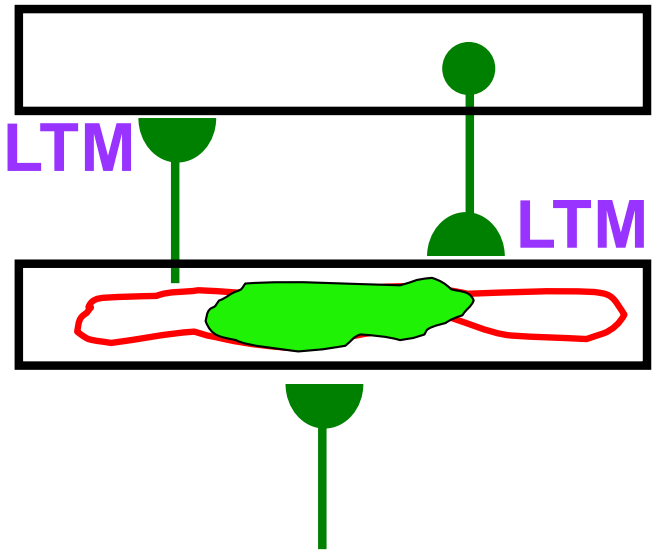
INCONSISTENT WITH MODELS WHERE TOP-DOWN MATCH IS SUPPRESSIVE

Mumford, 1992

Rao and Ballard, 1999

Bayesian Explaining Away

ART IS EXPLAINABLE (TRUSTWORTHY)



Categories
STM

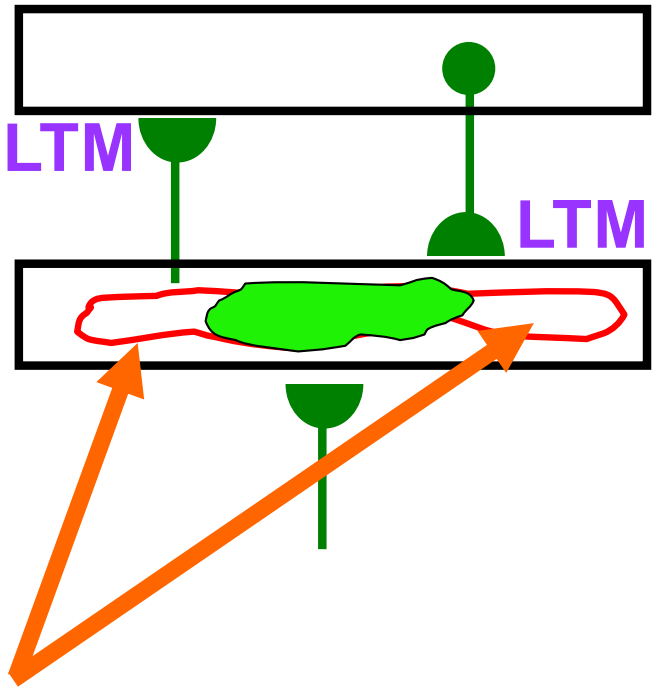
Feature Patterns
STM

STM: critical feature patterns determine attentional focus that controls information processing

LTM: critical feature patterns determine adaptive weights learned by the BU adaptive filter and TD expectation

Later: fuzzy ARTMAP learns fuzzy IF-THEN rules

ART IS RELIABLE (~~CATASTROPHIC FORGETTING~~)



Categories
STM

Feature Patterns
STM

Outlier features not in critical feature patterns are suppressed

Only predictive features are processed and coded

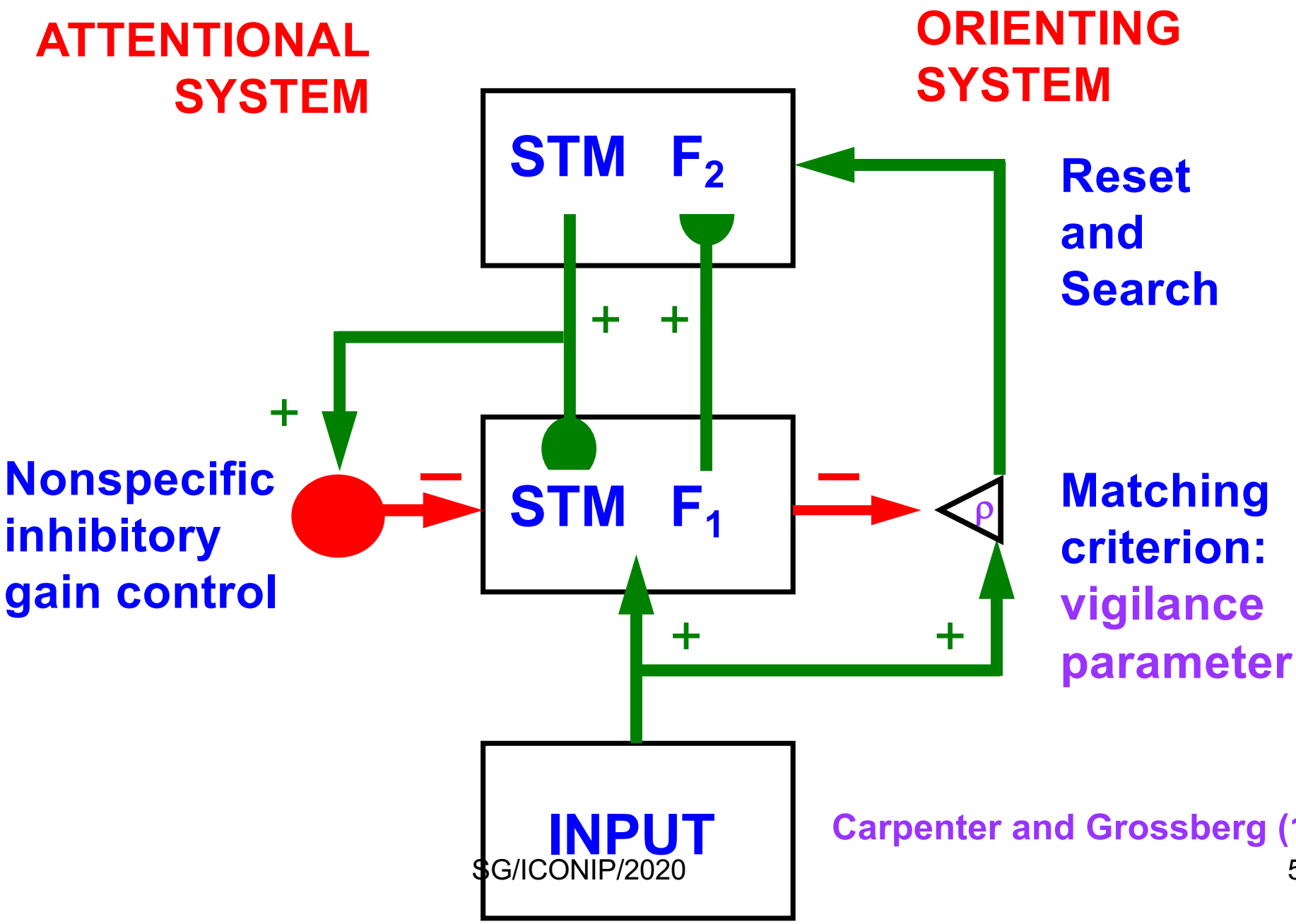
BP is a feedforward adaptive filter

ART is more than a feedforward adaptive filter

ART IS AN EXPLAINABLE SELF-ORGANIZING PRODUCTION SYSTEM IN A NON-STATIONARY WORLD

It is a PRODUCTION SYSTEM because it uses HYPOTHESIS TESTING to discover and learn RULES via a top-down matching process that focuses attention on CRITICAL FEATURE PATTERNS that predict behavioral success while suppressing irrelevant features

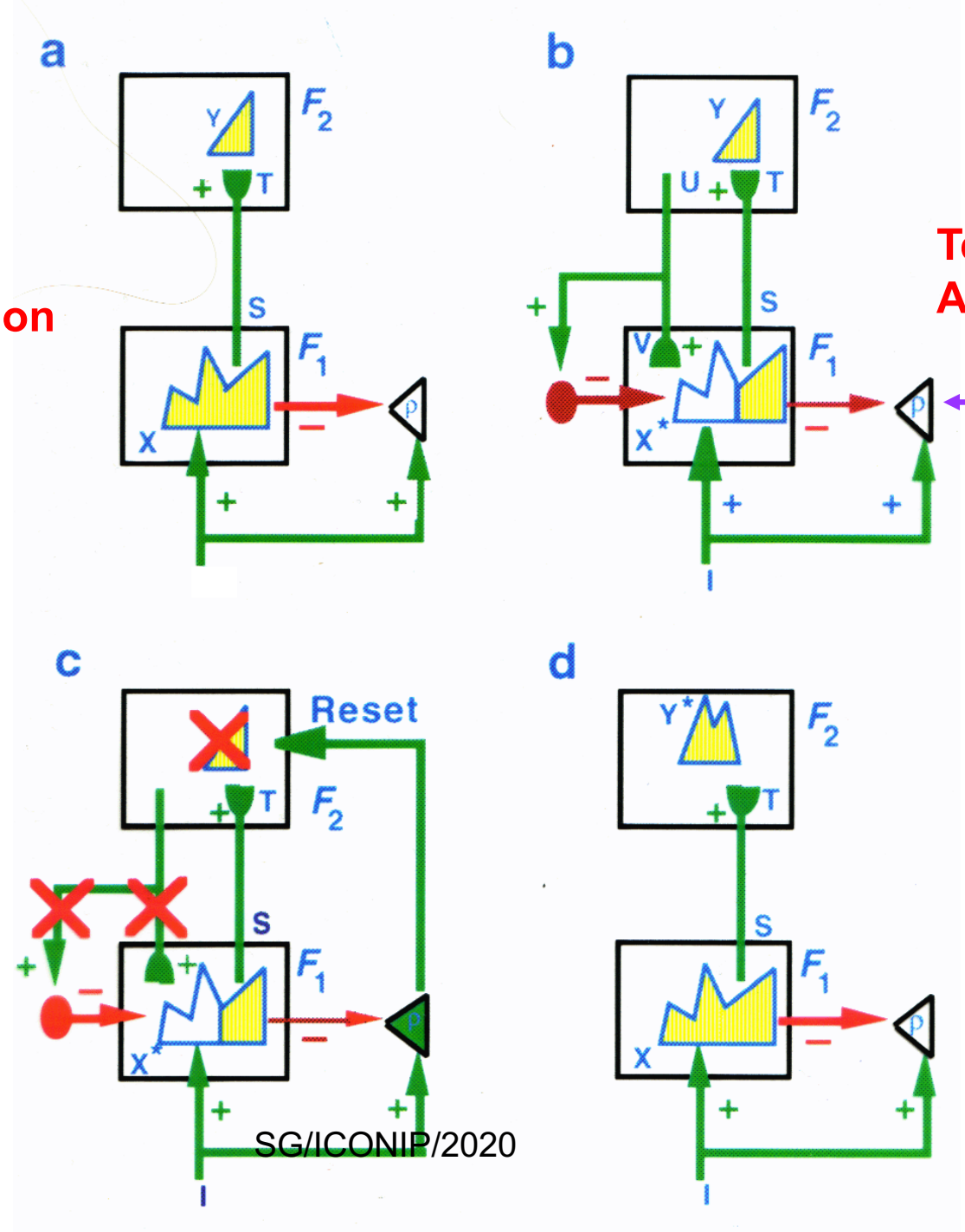
ART 1 MODEL



Carpenter and Grossberg (1987)

ART HYPOTHESIS TESTING AND LEARNING CYCLE

Choose category, or symbolic representation



Test hypothesis: ART matching rule

← VIGILANCE How big a mismatch causes reset?

Mismatch Reset: Novelty-Sensitive Arousal Burst

Choose another category

COGNITIVE LEARNING AND MEMORY CONSOLIDATION CYCLE

A dynamic cycle of

RESONANCE

and

RESET

As categories are learned, search automatically disengages

Modulatory novelty potentials subside as

this type of memory consolidation ends

Direct access to globally best-matching category

Mathematical proof in: Carpenter & Grossberg, *CVGIP*, 1987

Many supportive psychological and neurobiological data

Explains how we can quickly recognize familiar objects

even if, as we get older, we store enormous numbers of memories

ERP SUPPORT FOR HYPOTHESIS TESTING CYCLE

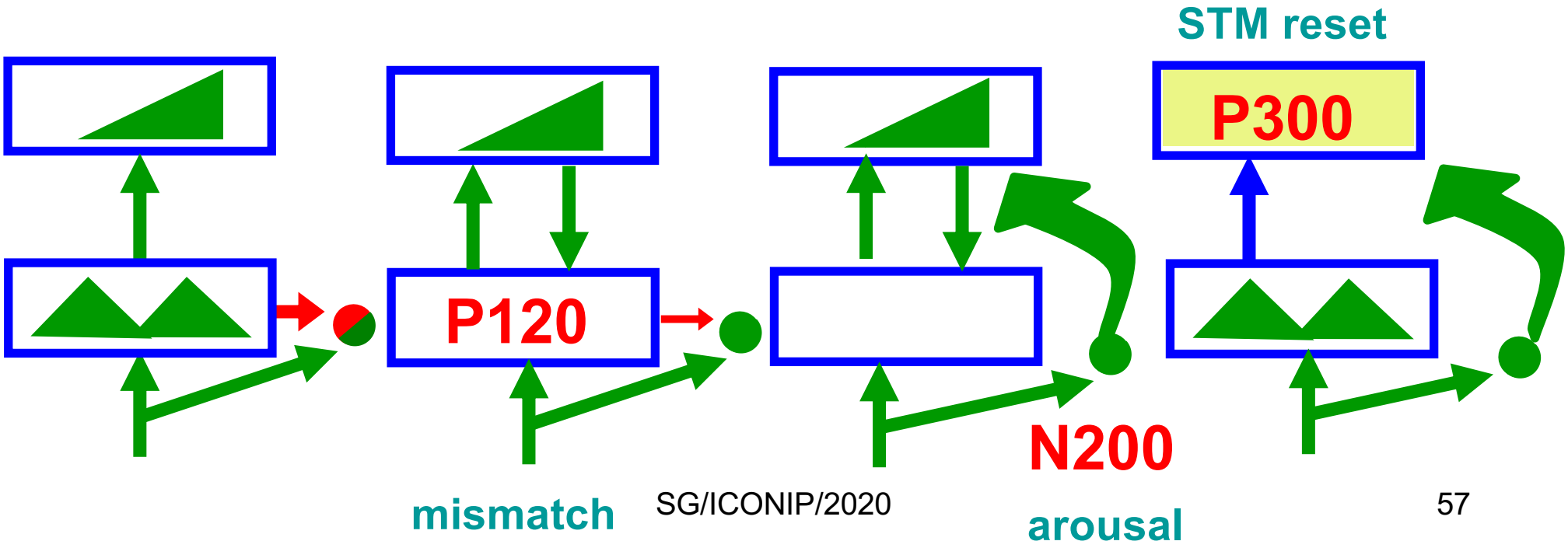
Event-Related Potentials: Human Scalp Potentials

ART predicted correlated sequences of P120-N200-P300

Event Related Potentials during oddball learning

P120 - mismatch; N200 - arousal/novelty; P300 - STM reset

Confirmed in: Banquet and Grossberg, 1987



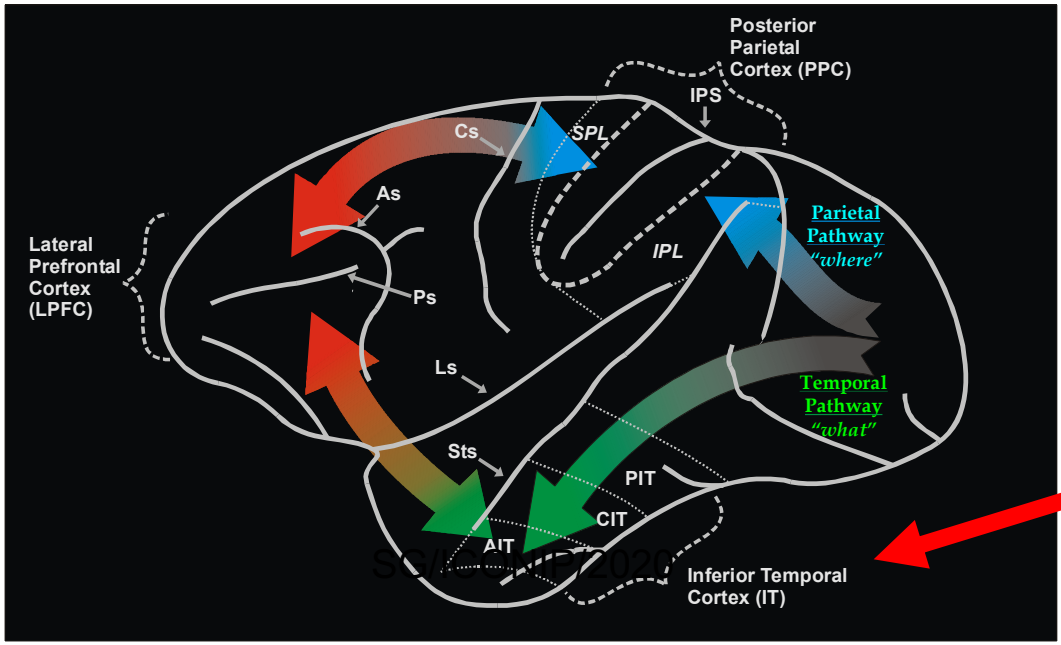
NEUROPHYSIOLOGICAL SUPPORT FOR HYPOTHESIS TESTING CYCLE

Cells in inferotemporal cortex are actively reset during working memory tasks

There is an

"active matching process that was reset between trials."

Miller, Li, Desimone, 1991



IT cortex

NEUROPHYSIOLOGICAL SUPPORT FOR HYPOTHESIS TESTING CYCLE

Classical data about hippocampus mismatch dynamics:

Novelty potentials subside as learning proceeds

(orienting system is disengaged)

Deadwyler et al., 1979, 1981; Otto and Eichenbaum, 1992;

Sokolov, 1968; Vinogradova, 1975

More recent data from prefrontal cortex (PFC) and hippocampus (HPC) when monkeys learn object-pair associations:

"Rapid object associative learning may occur in PFC, while HPC may guide neocortical plasticity by signaling success or failure…"

(attentional system interacts with orienting system)

Brincat and Miller, 2015

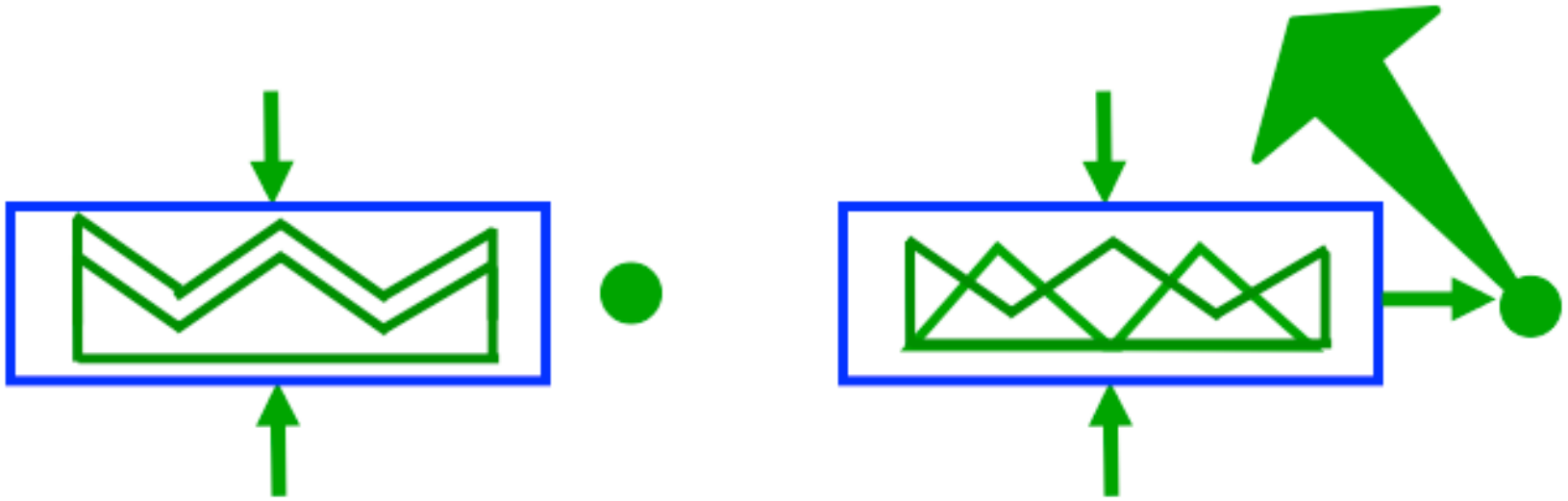
COMPLEMENTARY COMPUTING IN ART

Attentional and Orienting System Laws are Complementary

PN AND N200 ARE COMPLEMENTARY WAVES

MATCH

MISMATCH



PN

Top-Down
Conditionable
Specific
Match

N200

Bottom-Up
Unconditionable
Nonspecific
Mismatch

Illustrates NEW PARADIGMS for brain computing

INDEPENDENT MODULES

Computer Metaphor

COMPLEMENTARY COMPUTING

What is the nature of brain specialization?

LAMINAR COMPUTING

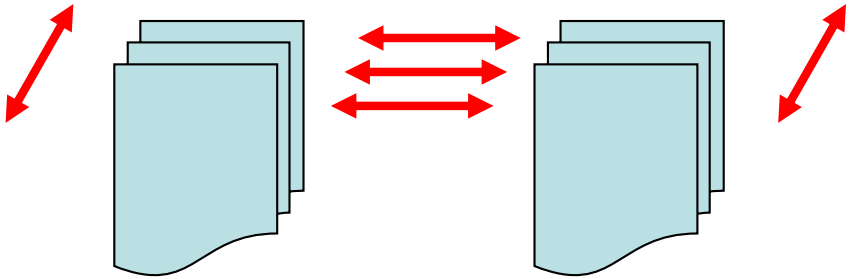
Why are all neocortical circuits organized in layers?

How do laminar circuits give rise to biological intelligence?

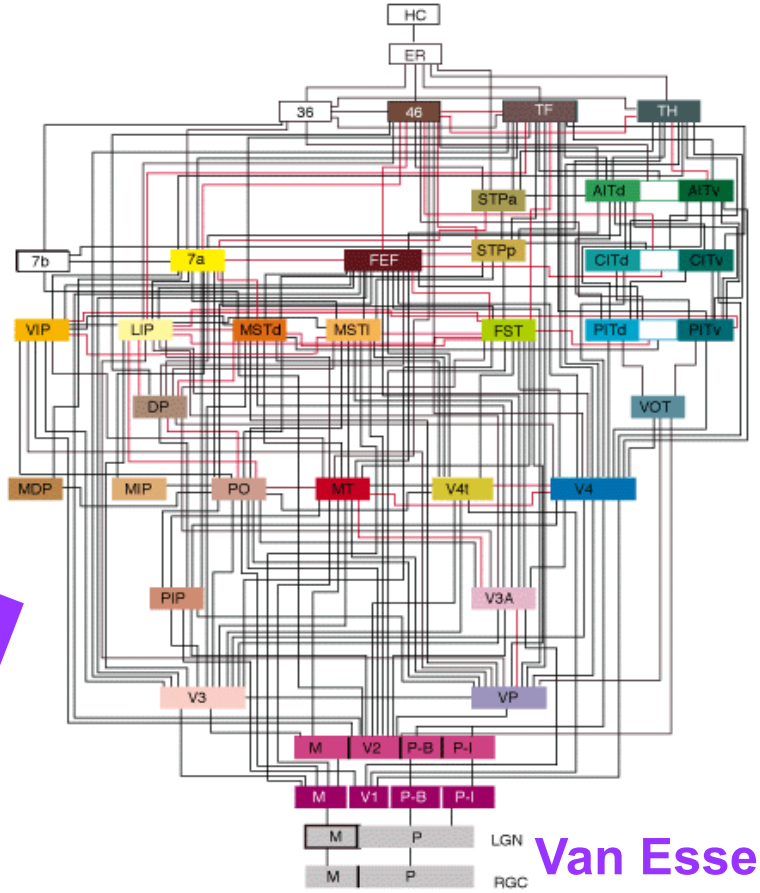
COMPLEMENTARY COMPUTING

New principles of
UNCERTAINTY and COMPLEMENTARITY
clarify why

Multiple parallel processing streams exist in the brain



Lots of specialization!



Van Essen et al

WHAT ARE COMPLEMENTARY PROPERTIES?

Analogies:
Key fits lock, puzzles pieces fit together

Computing one set of properties at a processing stage prevents that stage from computing a complementary set of properties

Complementary parallel processing streams are BALANCED against one another

INTERACTIONS between streams overcomes their complementary weaknesses and support intelligent and creative behaviors

SOME COMPLEMENTARY PROCESSES

Visual Boundary
Interbob Stream V1-V4

Visual Surface
Blob Stream V1-V4

Visual Boundary
Interbob Stream V1-V4

Visual Motion
Magno Stream V1-MT

WHAT Steam
Perception & Recognition
Inferotemporal and
Prefrontal areas

WHERE Stream
Space & Action
Parietal and
Prefrontal areas

Object Tracking
MT Interbands and MSTv

Optic Flow Navigation
MT Bands and MSTd

Motor Target Position
Motor and Parietal Cortex

Volitional Speed
Basal Ganglia

BP AND DEEP LEARNING DO NOT HAVE

STM activation patterns

STM critical feature patterns

ATTENTION

ANY FAST INFORMATION PROCESSING

LTM top-down learned expectations

HYPOTHESIS TESTING

using interacting STM and LTM traces

No NEURAL ARCHITECTURE

e.g., Complementary Computing

CATASTROPHIC FORGETTING EXAMPLES

Carpenter & Grossberg (1987)

You do not need a large database to show catastrophic forgetting if the ART MATCHING RULE does not hold

Learning lists of JUST FOUR INPUT VECTORS A, B, C, and D can exhibit catastrophic forgetting if they are repeated cyclically in the order:

ABCAD ABCAD ABCAD

and are related to each other in the following way:

CODE INSTABILITY INPUT SEQUENCES

$$D \subset C \subset A$$

$$B \subset A$$

$$B \bigcap C = \varnothing$$

$$|D| < |B| < |C|$$

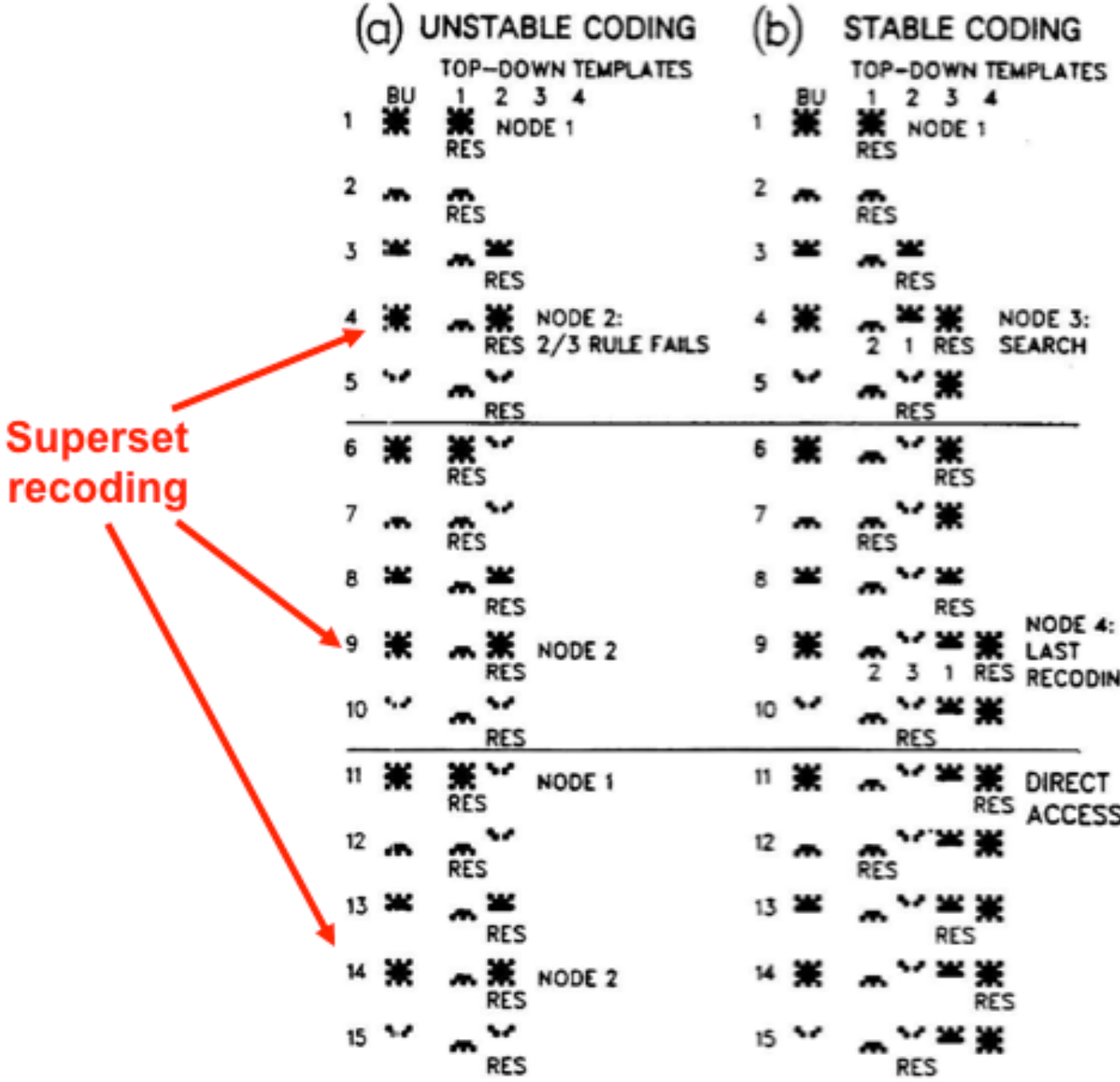
where $|E|$ is the number of features in the set E

Any set of input vectors that satisfy the above conditions will lead to unstable coding if they are periodically presented in the order

$$ABCAD$$

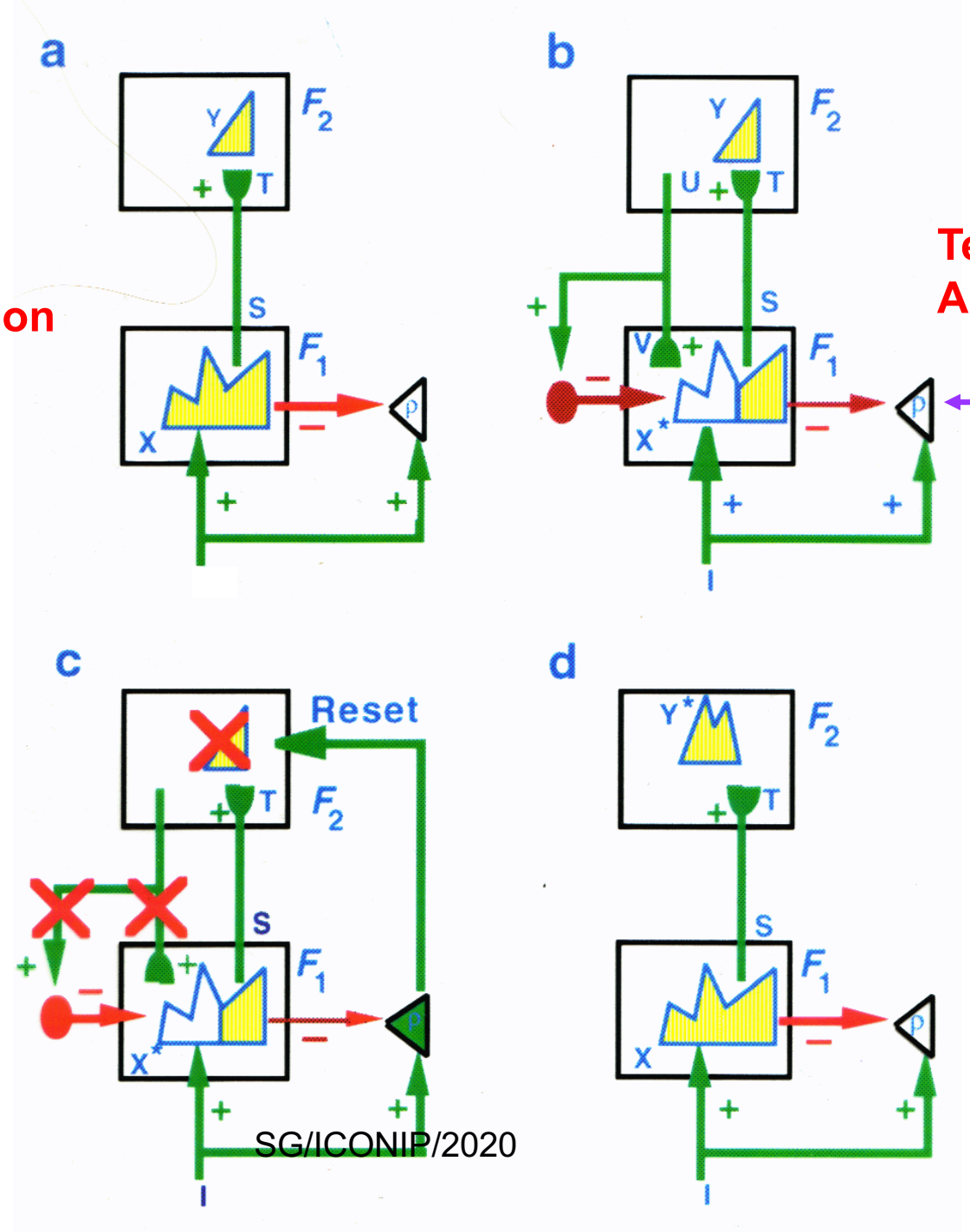
and the top-down ART Matching Rule is shut off

STABLE AND UNSTABLE LEARNING



ART HYPOTHESIS TESTING AND LEARNING CYCLE

Choose category, or symbolic representation



Test hypothesis: ART matching rule

VIGILANCE How big a mismatch causes reset?

Mismatch Reset: Novelty-Sensitive Arousal Burst

Choose another category

VIGILANCE
determines what features are learned in the CRITICAL FEATURE PATTERN

It clarifies how our brains learn CONCRETE knowledge

for some tasks and ABSTRACT knowledge for others

High Vigilance – Narrow Categories; CONCRETE
Mom's face

Low Vigilance – Broad Categories; ABSTRACT
A face

Critical feature patterns are explainable at every level of vigilance!

VIGILANCE DATA IN INFEROTEMPORAL CORTEX

RECEPTIVE FIELD SELECTIVITY MUST BE LEARNED

- Some cells respond selectively to particular views of particular faces
- Other cells respond to broader features of an animal's environment

Desimone, Gross, Perrett, …

EASY vs. DIFFICULT DISCRIMINATIONS: VIGILANCE!

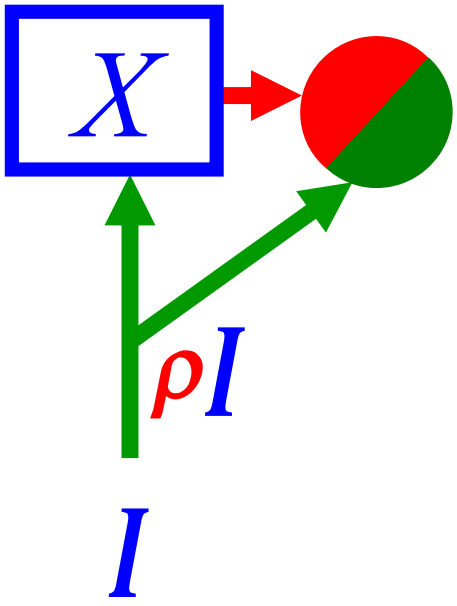
"In the difficult condition the animals adopted a stricter internal criterion for discriminating matching from non-matching stimuli… The animal's internal representations of the stimuli were better separated … increased effort appeared to cause enhancement of the responses and sharpened selectivity for attended stimuli…"

Spitzer, Desimone, and Moran, 1988

VIGILANCE CONTROL

$$\rho|I| - |X| \leq 0 \qquad \rho \leq \frac{|X|}{|I|}$$ **resonate and learn**

$$\rho|I| - |X| > 0 \qquad \rho > \frac{|X|}{|I|}$$ **reset and search**

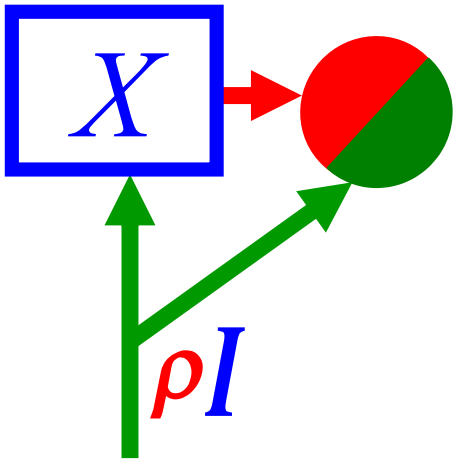


ρ **is a sensitivity or gain parameter**

VIGILANCE CONTROL

$$\rho|I| - |X| \leq 0 \qquad \rho \leq \frac{|X|}{|I|}$$ **resonate and learn**

$$\rho|I| - |X| > 0 \qquad \rho > \frac{|X|}{|I|}$$ **reset and search**



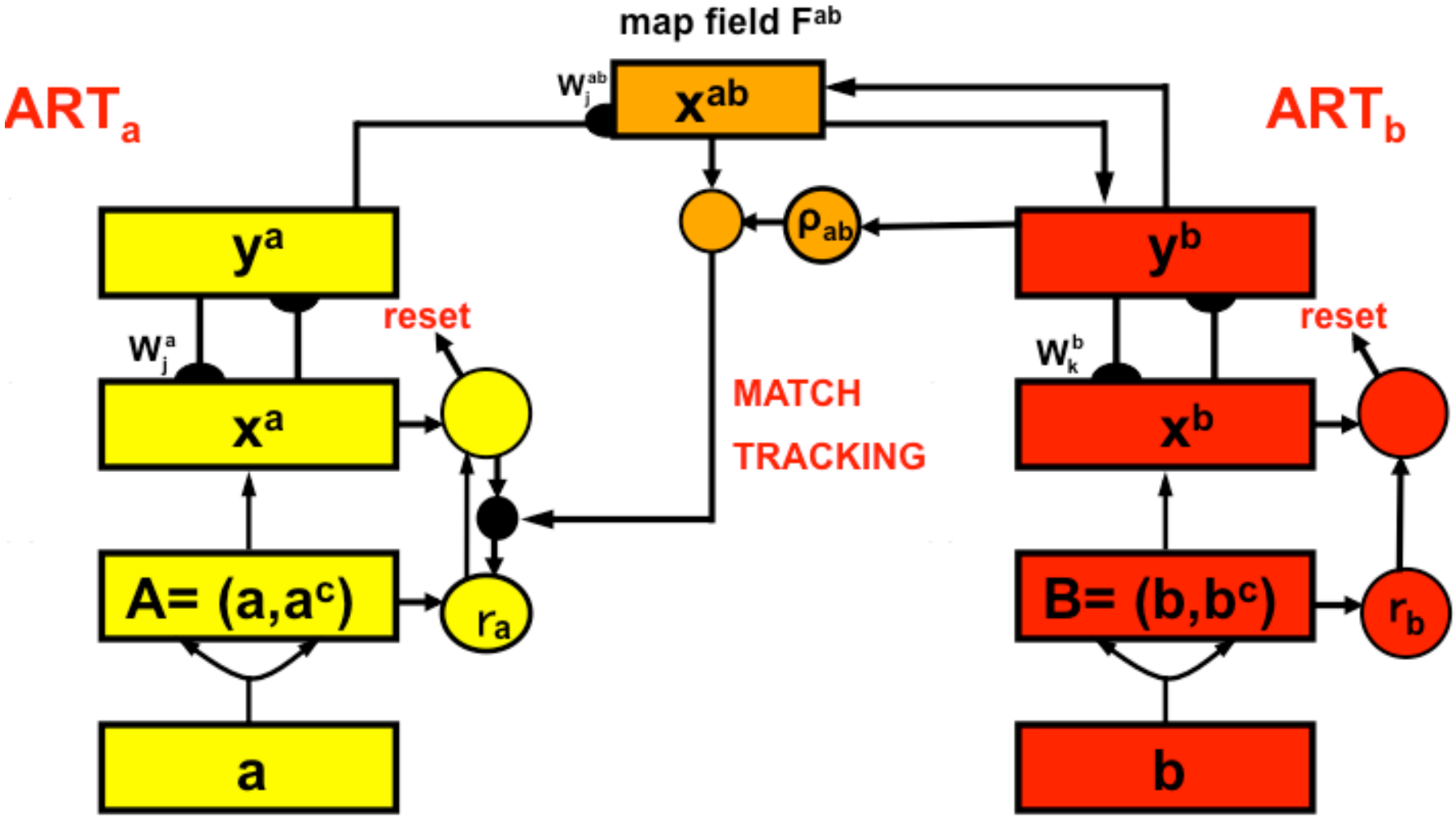
ρ **is a sensitivity or gain parameter**

How to change vigilance based on predictive success?

FROM UNSUPERVISED TO SUPERVISED ART MODELS

Extend UNSUPERVISED ART to SUPERVISED or UNSUPERVISED ARTMAP

FUZZY ARTMAP

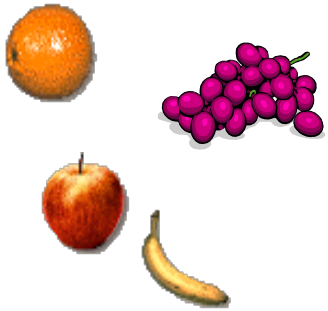


MATCH TRACKING realizes Minimax Learning Principle:

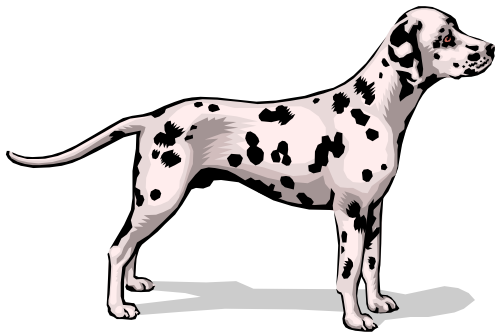
Vigilance increases to just above the match ratio of prototype / exemplar, thereby triggering search

LEARN MANY-TO-ONE and ONE-TO-MANY MAPS

Many-to-One	One-to-Many
Compression, Naming	Expert Knowledge
(a_1,b)	(a,b_1)
(a_2,b)	(a,b_2)
(a_3,b)	(a,b_3)
(a_4,b)	(a,b_4)

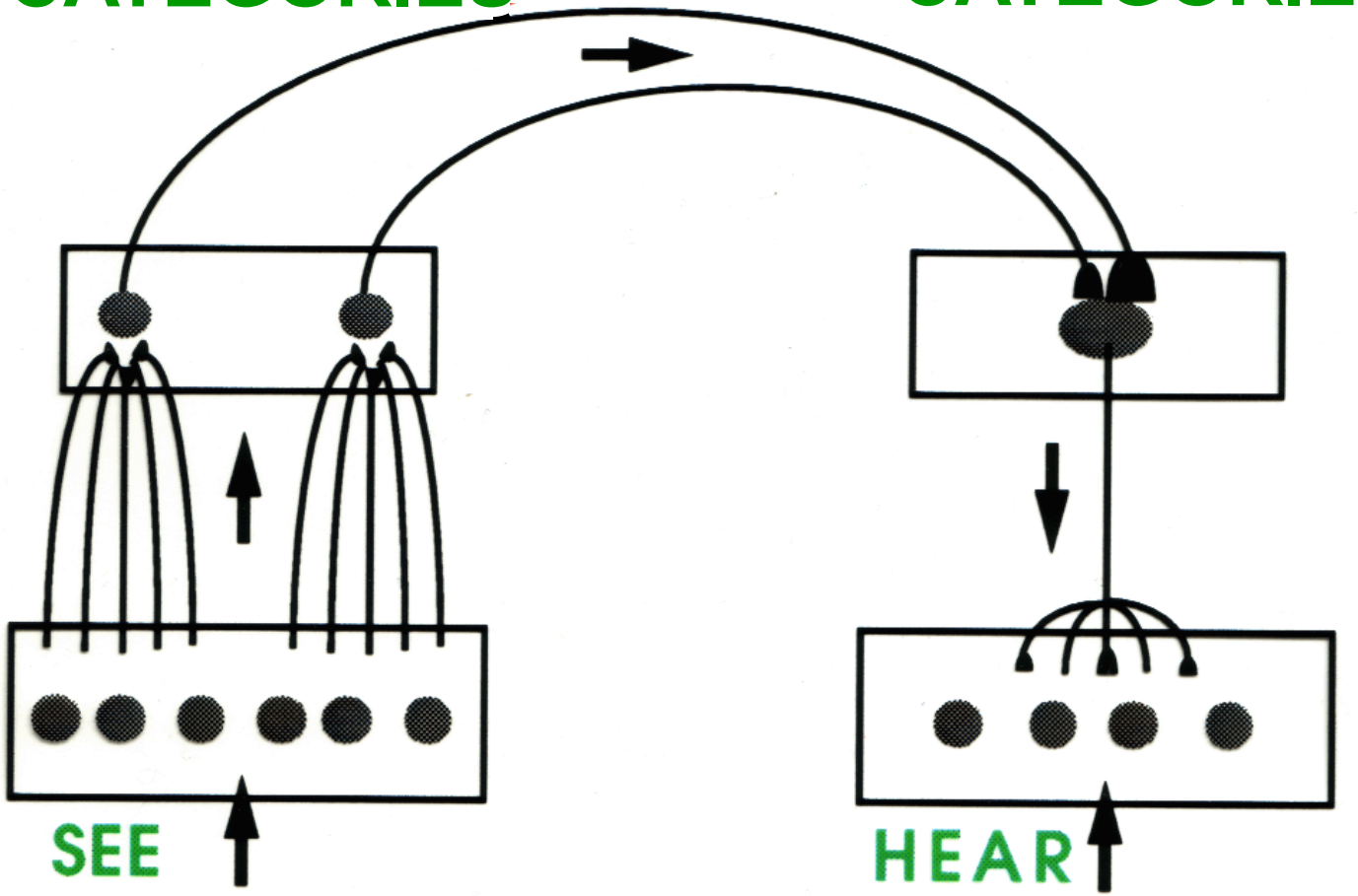


Fruit

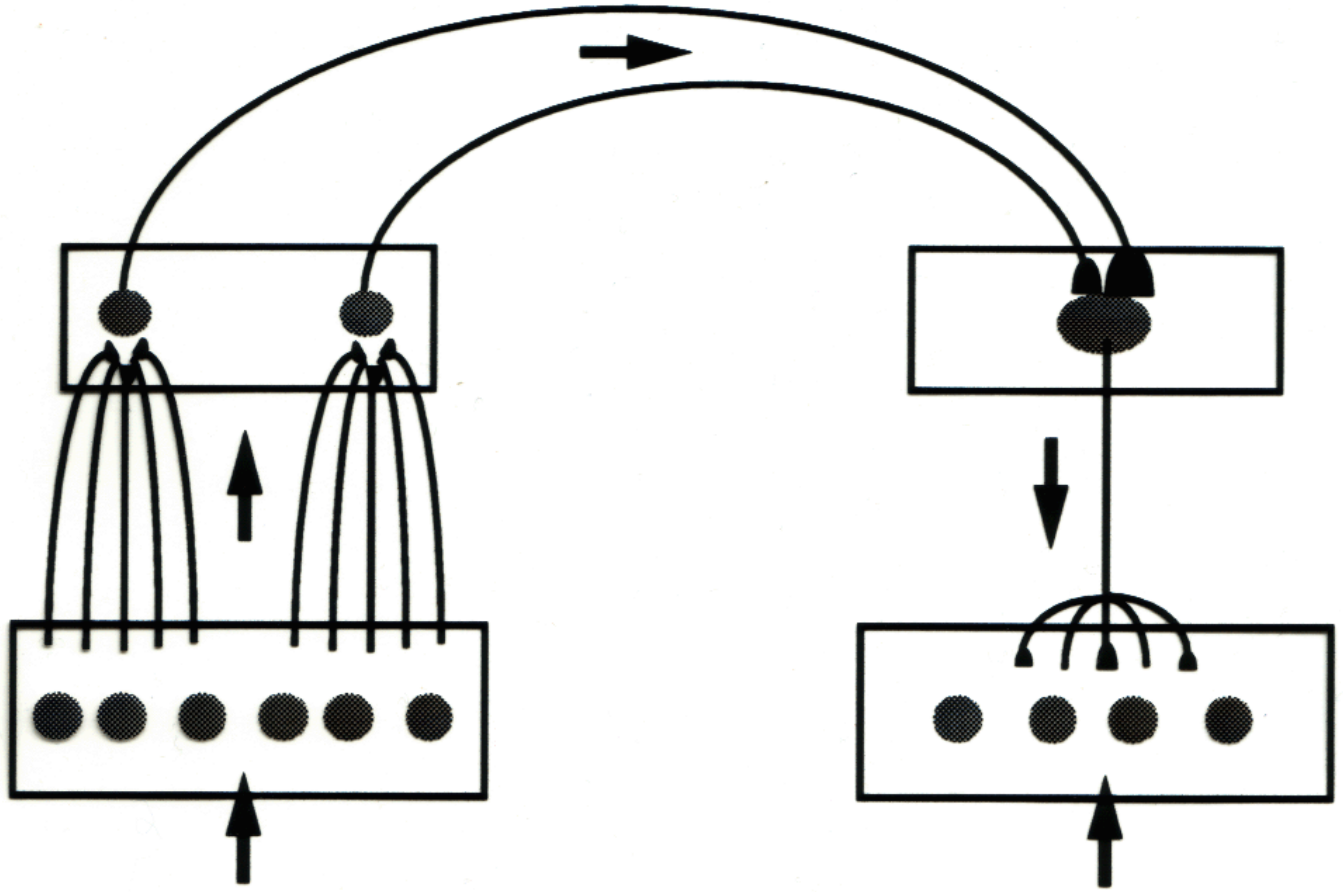


Animal
Mammal
Pet
Dog
Dalmatian
Fireman's
Mascot
"Rover"

MANY-TO-ONE MAP
Two Stages of Compression



MANY-TO-ONE MAP
Two Stages of Compression
Medical Database Prediction

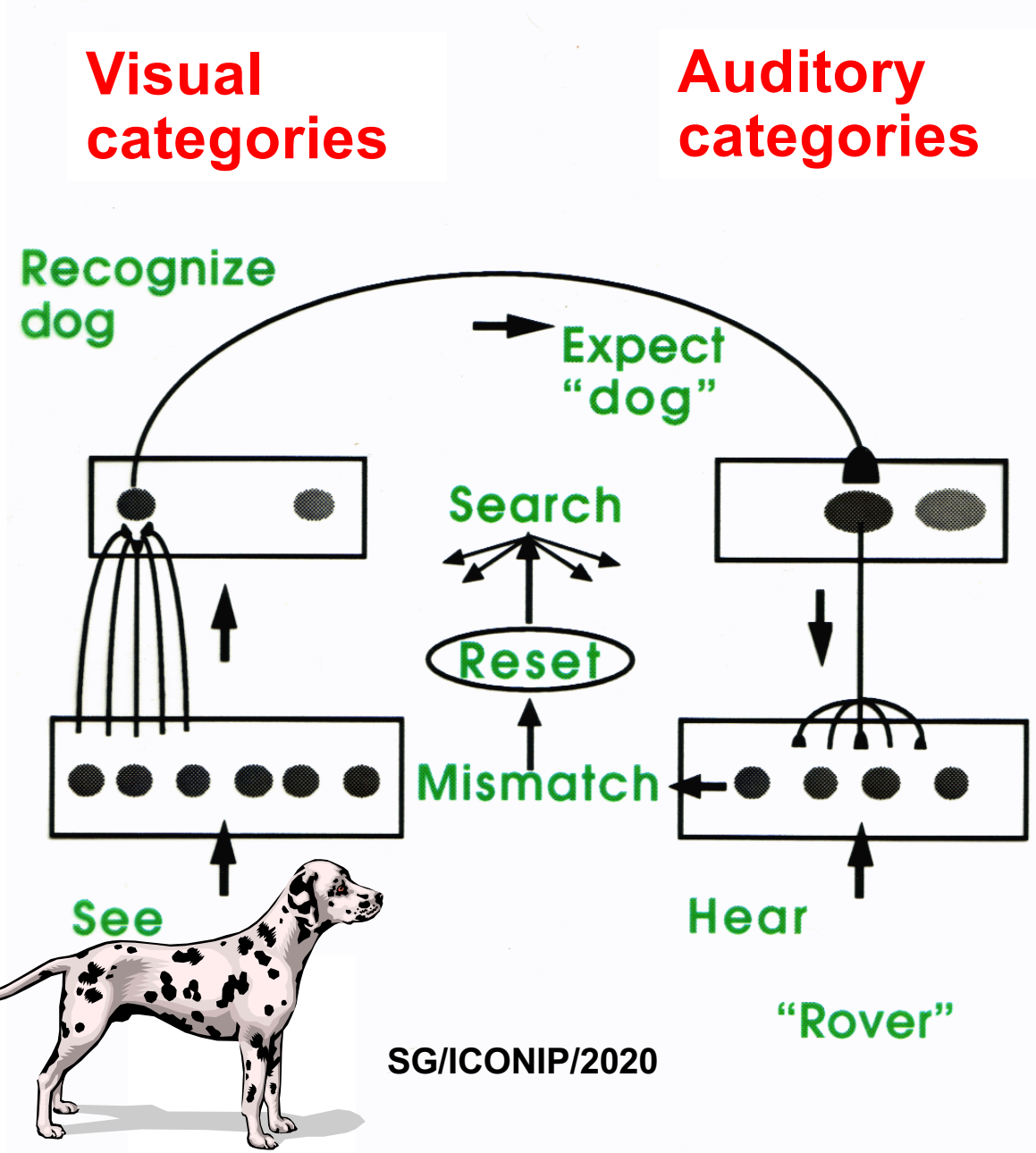


Symptoms tests treatments

Length of stay in hospital

ONE-TO-MANY MAP

Expert Knowledge



MINIMAX LEARNING PRINCIPLE

How to conjointly

minimize predictive error

and

maximize generalization

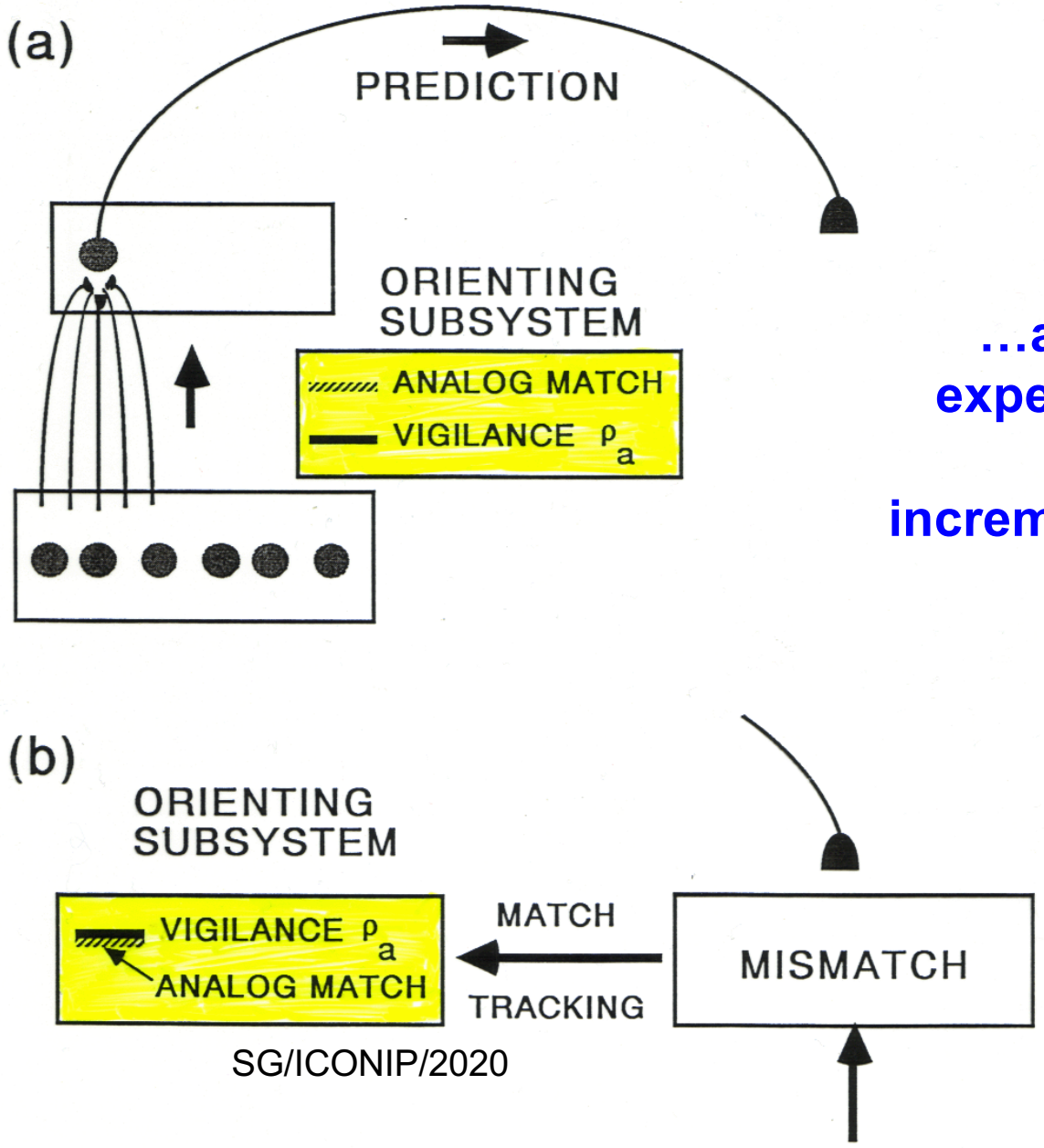
using error feedback

in an incremental fast learning context

in response to nonstationary data?

MATCH TRACKING realizes MINIMAX LEARNING PRINCIPLE

Given a predictive error, vigilance increases just enough to trigger search and thus sacrifices the minimum generalization to correct the error



…and enables expert knowledge to be incrementally learned

Are ART mechanisms like vigilance control realized within LAMINAR cortical and thalamic circuits?

YES!

SMART model

Synchronous Matching ART

Grossberg and Versace, 2008

MAIN QUESTIONS:

How are multiple levels of brain organization

spikes

local field potentials

inter-areal synchronous oscillations

spike-timing dependent plasticity

coordinated to

regulate stable category learning and attention

during cognitive information processing via

laminar cortical circuits

specific and nonspecific thalamic nuclei?

Illustrates NEW PARADIGMS for brain computing

INDEPENDENT MODULES

Computer Metaphor

COMPLEMENTARY COMPUTING

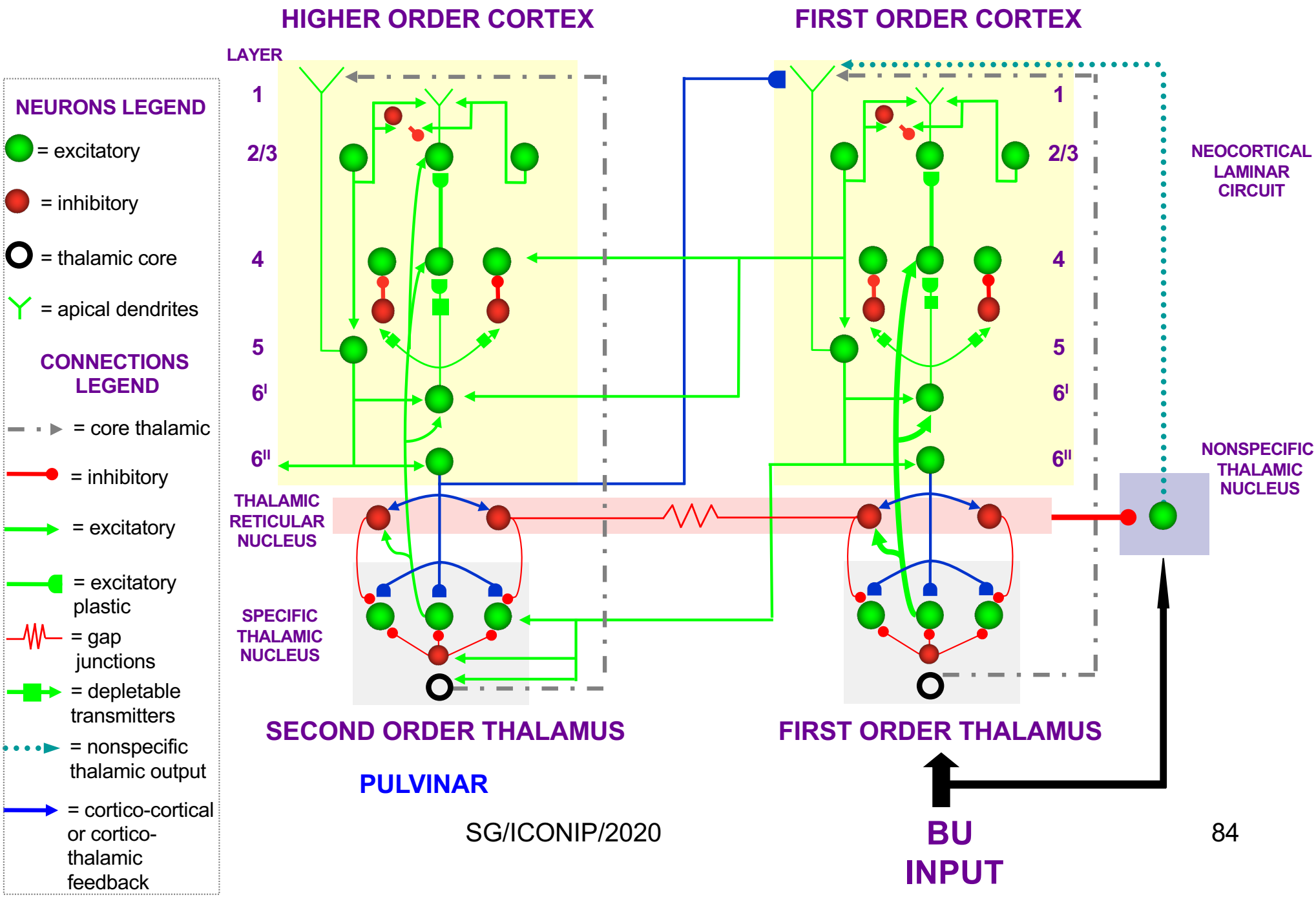
What is the nature of brain specialization?

LAMINAR COMPUTING

Why are all neocortical circuits organized in layers?

How do laminar circuits give rise to biological intelligence?

SMART: MODEL MACROCIRCUIT



THE MODEL FUNCTIONALLY EXPLAINS LOTS OF ANATOMICAL DATA

Connections	Type	Functional interpretation	References
thalamic core A → 4 A	D	Primary thalamic relay cells drive layer 4.	Blasdel and Lund (1983)
thalamic core A → 6^{I} A	D	Primary thalamic relay cells prime layer 4 via the 6 → 4 modulatory circuit.	Blasdel and Lund (1983) for LGN → 6; Callaway (1998) LGN input to 6 is weak and Layer 5 projections to 6 **[Note 1]**
thalamic core A → RE A	D	Recurrent inhibition to primary and secondary thalamic relay cells.	Sherman and Guillery (2001); Jones (2002)
RE A → thalamic core A	I	Off-surround to primary and secondary thalamic relay cells, synchronization of thalamic relay cells.	Cox *et al.* (1997); Pinault and Deschenes (1998); Sherman and Guillery (2001)
RE A → RE A	I	Normalization of inhibition.	Jones (2002); Sohal and Huguenard (2003)
RE A (B) → RE B(A)	GJ	Synchronize RE and thalamic relay cells.	Landisman *et al.* (2002)
RE A → nonspecific thalamic A	I	Inhibition of nonspecific thalamic cells, participates in the reset mechanism.	Kolmac and Mitrofanis (1997); Van der Werf *et al.* (2002)
nonspecific thalamic A → 5 A	M	To 5 through apical dendrites in 1, participates in the reset mechanism.	Van der Werf *et al.* (2002)
4 A → 4 inh. A	D	Lateral inhibition in layer 4.	Markram *et al.* (2004)
4 inh. A → 4 A	I	Lateral inhibition in layer 4.	Markram *et al.* (2004)
4 inh. A → 4 inh. A	I	Normalization of inhibition in layer 4.	Ahmed *et al.* (1997); Markram *et al.* (2004)
4 A → 2/3 A	D	Feedforward driving output from 4 to 2/3.	Fitzpatrick *at al.* (1985); Callaway and Wiser (1996)
2/3 A → 2/3 A	D	Recurrent connections (grouping) in 2/3.	Bosking *et al.* (1997); Schmidt *et al.* (1997); Grossberg and Raizada (2003)
2/3 A → 2/3 inh. A	D	Avoid outward spreading (bipole) in 2/3.	McGuire *et al.* (1991); Grossberg and Raizada (2003)
2/3 inh. A → 2/3 inh. A	I	Normalization of inhibition.	Tamas *et al.* (1998); Grossberg and Raizada(2003)
2/3 A → 4 B	D	Feedforward output from Area A to Area B.	Van Essen *et al.* (1986)
2/3 A → 6^{II} B	D	Feedforward output from Area A to Area B.	Van Essen *et al.* (1986)
2/3 A → 5 A	D	Conveys layer 2/3 output to layer 5.	Callaway and Wiser (1996)
2/3 A → 6^{II} A	D	Conveys layer 2/3 output to layer 6^{II}.	Callaway (1998)

THE MODEL FUNCTIONALLY EXPLAINS LOTS OF ANATOMICAL DATA

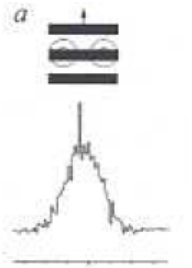
Connections	Type	Functional interpretation	References
5 A → thalamic core B	D	Feedforward connections from Area A to Area B through secondary thalamic relay neurons.	Rockland (1999); Sherman and Guillery (2001)
5 A → 6^{I} A	D	Delivers feedback to the 6 → 4 circuit from higher cortical areas, sensed at the apical dendrites of 5 branching in 1.	Callaway (1998); Callaway and Wiser (1996), class B" cells **[Note 2]**
6^{I} A → 4 A	M	On-center to 4. Mediated by habituative gates.	Stratford *et al.* (1996); Callaway (1998); Grossberg and Raizada (2003)
6^{I} A → 4 int. A	D	Off-surround to 4.	McGuire *et al.* (1984); Ahmed *et al.*, (1997); Callaway (1998)
6^{II} A → thalamic Core A	M	On-center to primary thalamic relay cells.	Sillito *et al.* (1994); Callaway (1998);
6^{II} A → RE A	D	Off-surround to primary thalamic relay cells mediated by thalamic RE.	Guillery and Harting (2003); Sherman and Guillery (2001)
6^{II} B → 2/3, 2/3 inh., 5 A	M	Intercortical feedback from 6^{II} area B to 1 area A, where it synapses on 2/3 excitatory and inhibitory neurons, as well as 5 apical dendrites branching in 1	Rockland and Virga (1989); Rockland (1994); Salin and Bullier (1995)

Abbreviations: **inh. =** inhibitory neurons; **RE** = reticular nucleus**; A** = primary (thalamic, cortical) loop; **B** = secondary (thalamic, cortical) loop; **D** = driving excitatory connections; **M** = modulatory connections; **I** = inhibitory connections; **GJ** = gap junctions; **int.** = inhibitory interneuron. **[Note 1]:** Callaway (1998) subdivides Layer 6 neurons in 3 classes: *Class I:* provide feedback to 4C, receive input from LGN, and project back to LGN; *Class IIa:* dendrites in 6, axons from 2/3, project back to 2/3 with modulatory connections; *Class IIb:* dendrites in 5, project exclusively to deep layers (5 & 6) and claustrum. In the model, these populations are clustered in 2 classes, layer 6I and 6II, which provide feedback to thalamic relay cells and layer 4, respectively. **[Note 2]:** Callaway (1998) subdivides Layer 5 neurons in 3 classes: *Class A:* dendrites in 5, axons from 2/3, project back to 2/3 with modulatory connections; *Class B:* dendrites in 5, axons from 2/3, project laterally to 5 and PULVINAR; *Class C:* dendrites in 1, project to superior colliculus. In the model, these differences are ignored, and it is assumed that the model layer 5 neuron receives input from 2/3 (Classes A and B), as well modulatory input from the nonspecific thalamic nuclei (Class C, apical dendrites in layer 1), and provide output to 6^{I} and second-order thalamic nuclei. The inner, recurrent loop with 2/3 has also been ignored.

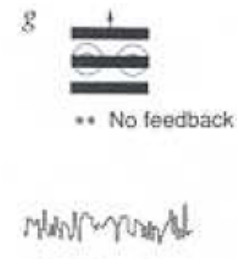
BRAIN OSCILLATIONS DURING MATCH/MISMATCH

DATA

FB ON

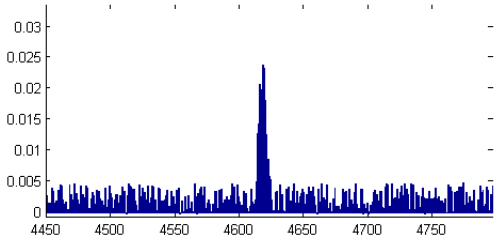


FB OFF

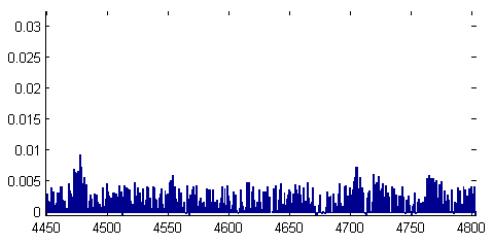


SIMULATION

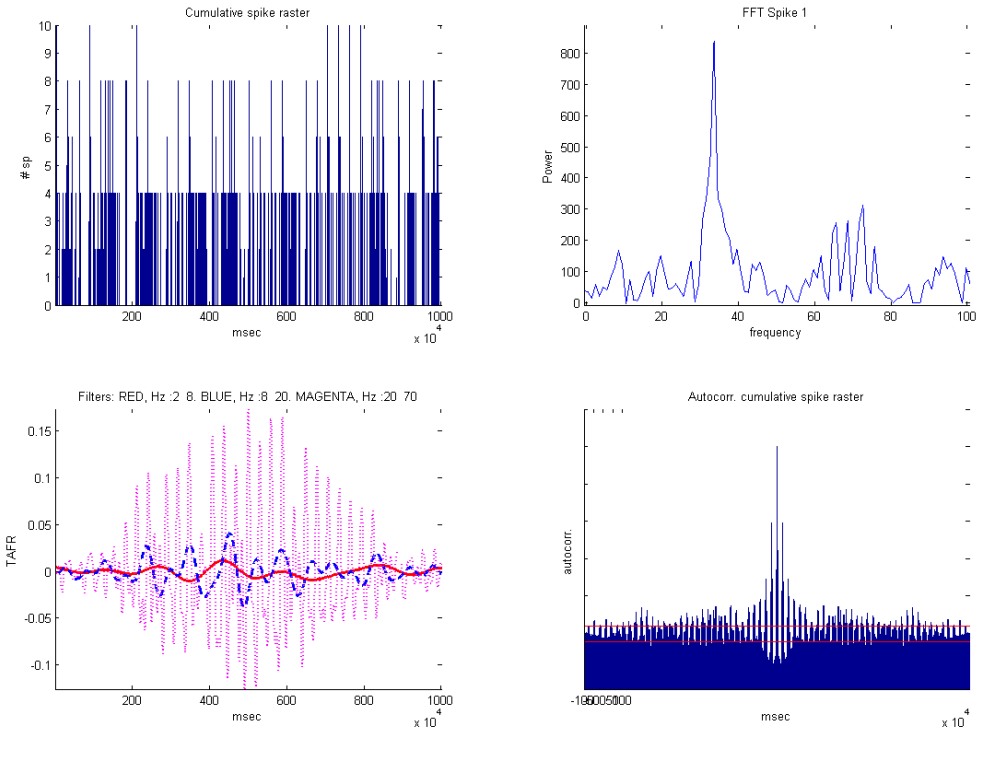
FB ON

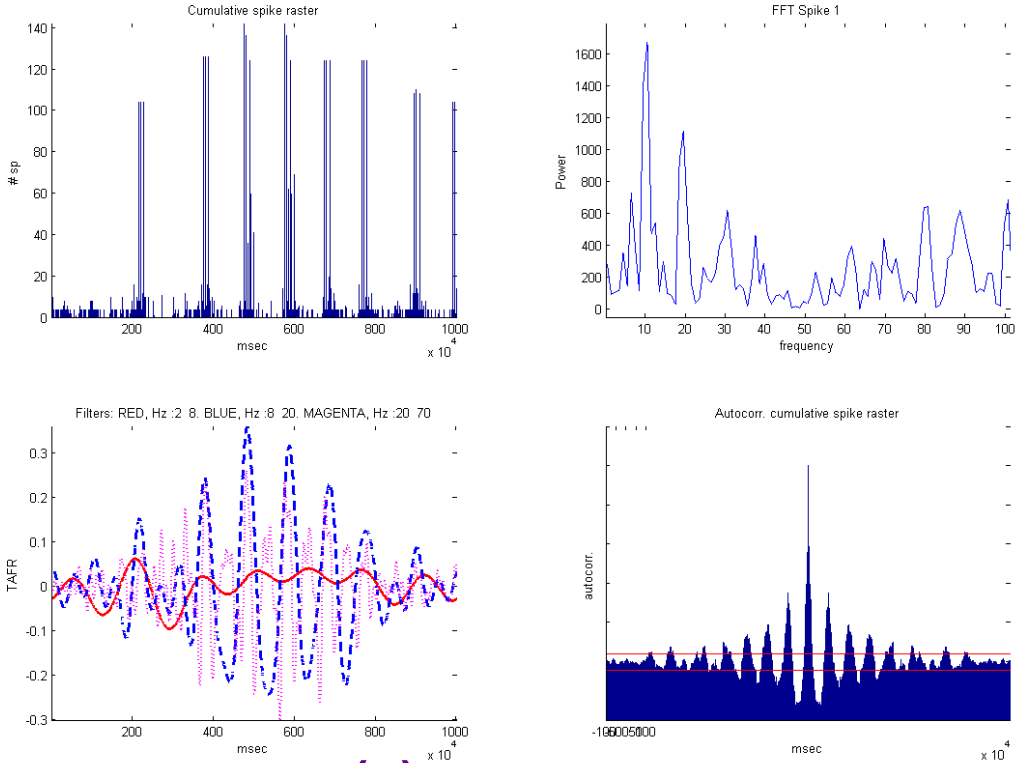


FB OFF



(a) TD CORTICOTHALAMIC FEEDBACK increases SYNCHRONY Sillito *et al.*, 1994





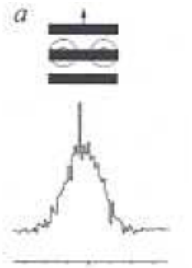
(b) MATCH
Increases γ oscillations

(c) MISMATCH
increases *θ*, *β* oscillations

BRAIN OSCILLATIONS DURING MATCH/MISMATCH

DATA

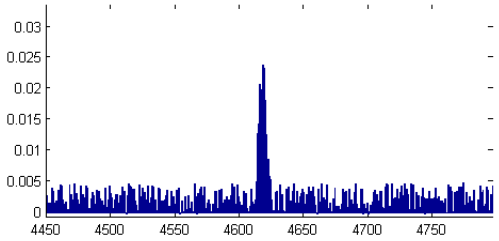
FB ON



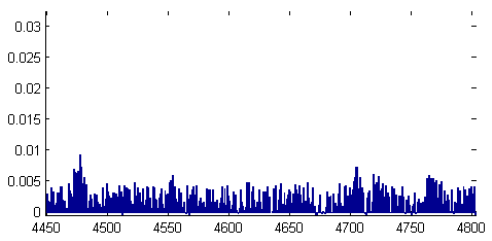
FB OFF

SIMULATION

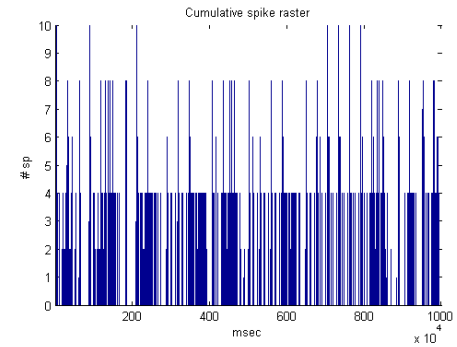
FB ON

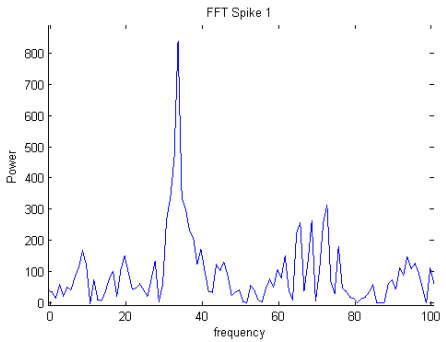


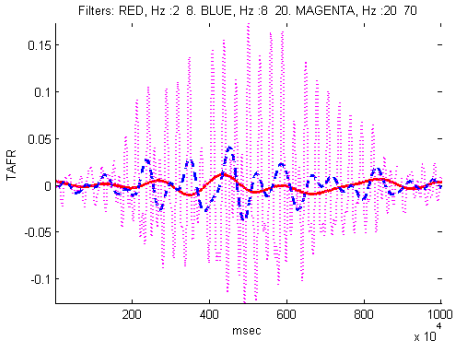
FB OFF

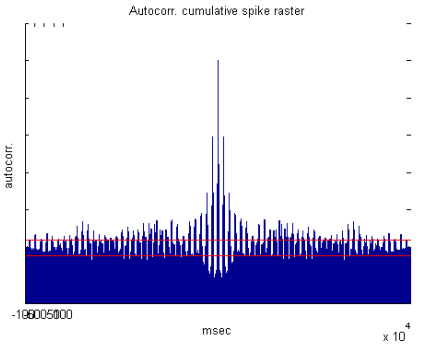


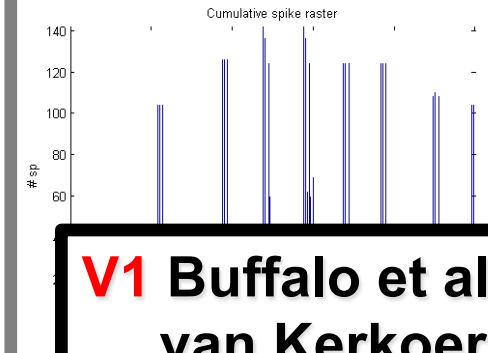
(a) TD CORTICOTHALAMIC FEEDBACK increases SYNCHRONY Sillito *et al.*, 1994

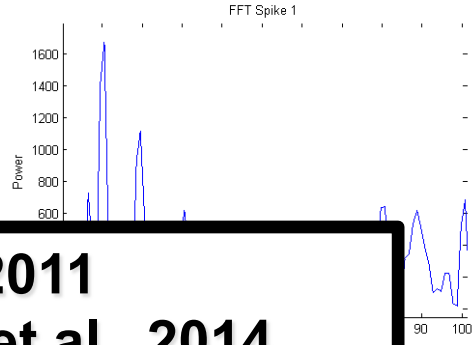




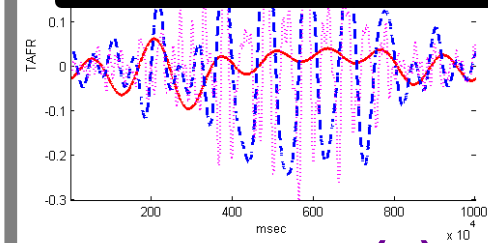


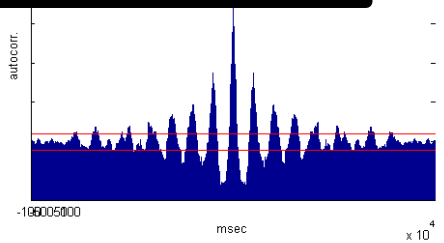






V1 Buffalo et al., 2011
van Kerkoerle et al., 2014
V4 Buschman and Miller, 2009
Hippocampus Berke et al., 2009





(b) MATCH
Increases γ oscillations

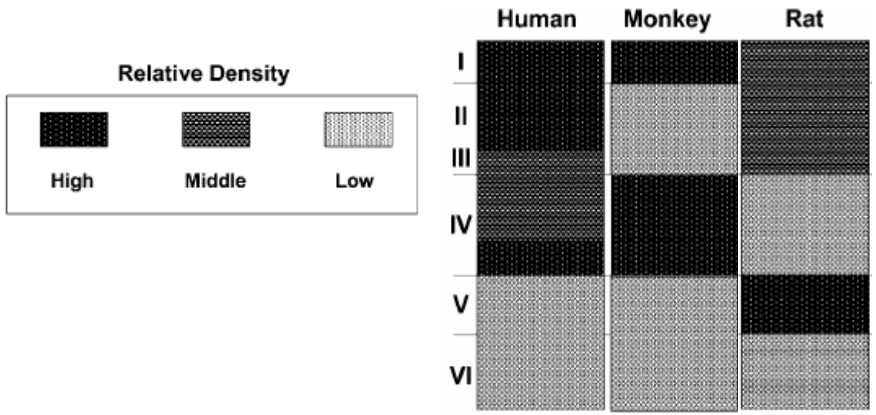
(c) MISMATCH
increases *θ*, *β* oscillations

VIGILANCE CONTROL: MISMATCH-MEDIATED ACETYLCHOLINE RELEASE

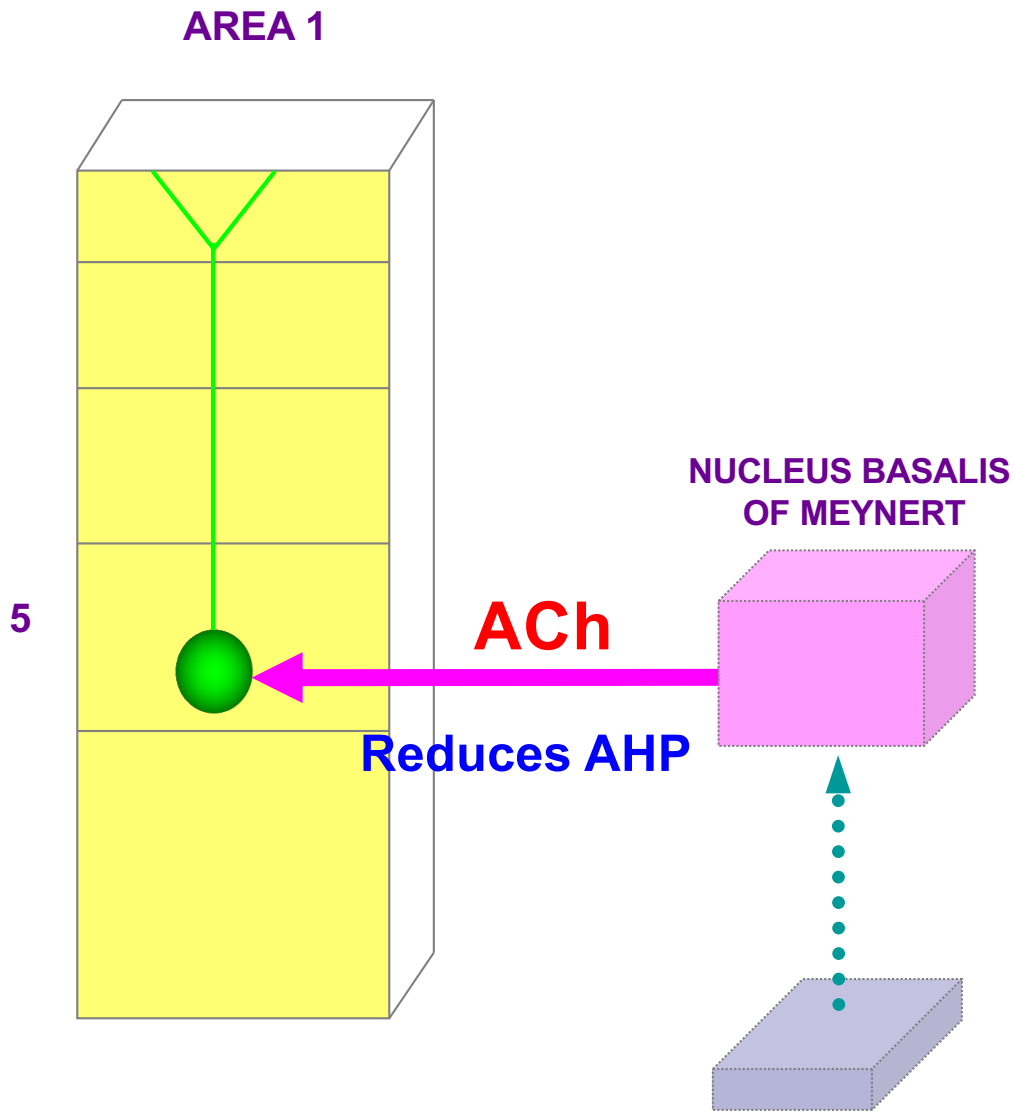
Acetylcholine (Ach) regulation by NONSPECIFIC THALAMIC NUCLEI via NUCLEUS BASALIS OF MEYNERT reduces AHP in layer 5

Ach thereby facilitates RESET (compare ART VIGILANCE control)

HIGH Vigilance ~ Sharp Code
LOW Vigilance ~ Coarse Code



CHOLINERGIC DENSITY AXONS IN V1 AND HOMOLOGS
Gu (2003)



BREAKDOWN OF ACETYLCHOLINE NEUROMODULATION OF VIGILANCE CONTROL DURING MENTAL DISORDERS

Grossberg, S. (2017). Acetylcholine neuromodulation in normal and abnormal learning and memory:
Vigilance control in
waking, sleep, autism, amnesia, and Alzheimer's disease
Frontiers in Neural Circuits, November 2, 2017.
https://doi.org/10.3389/fncir.2017.00082

OPEN ACCESS and sites.bu.edu/steveg

COGNITIVE LEARNING AND MEMORY CONSOLIDATION CYCLE

A dynamic cycle of

RESONANCE

and

RESET

As categories are learned, search automatically disengages

Modulatory novelty potentials subside as

this type of memory consolidation ends

Direct access to globally best-matching category

Mathematical proof in: Carpenter & Grossberg, *CVGIP*, 1987

Many supportive psychological and neurobiological data

Explains how we can quickly recognize familiar objects

even if, as we get older, we store enormous numbers of memories

Catastrophic forgetting occurs if top-down expectations fail

What goes wrong if the ORIENTING SYSTEM fails?
AMNESIA OCCURS!

DYNAMIC PHASE OF MEMORY CONSOLIDATION

While input exemplar still drives

memory search

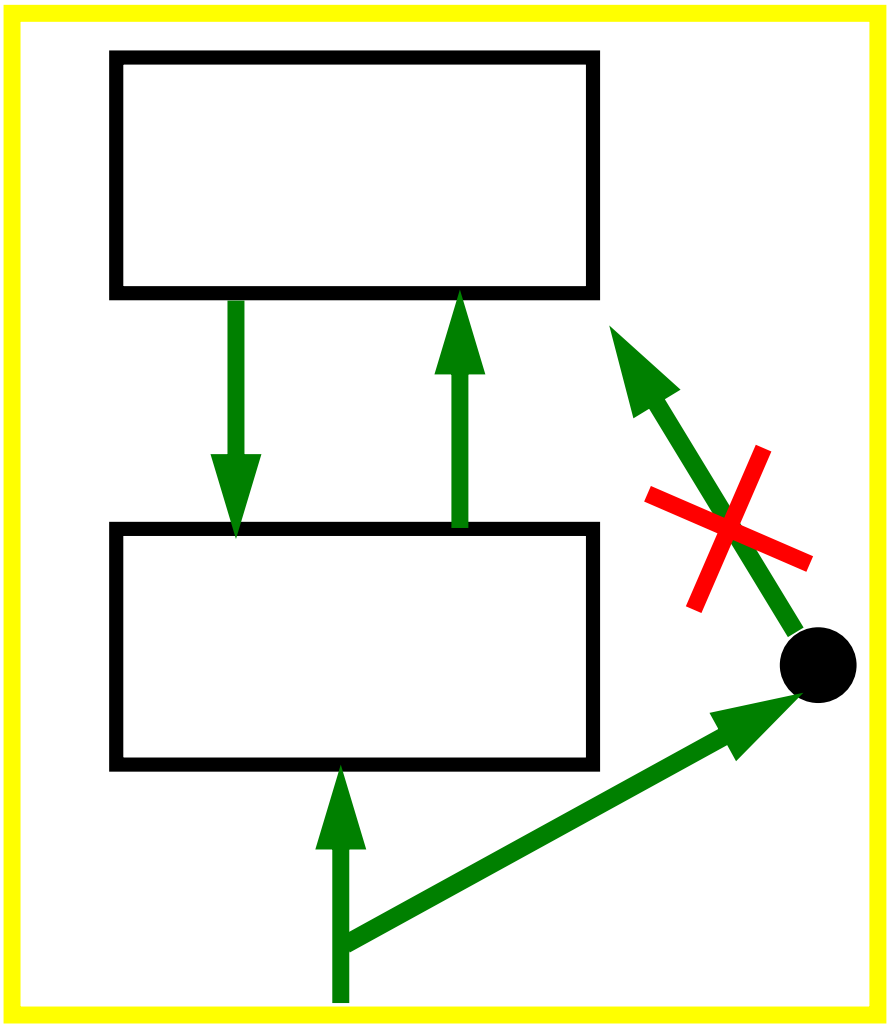
before direct access occurs

An emergent property of the entire circuit

A FORMAL AMNESIC SYNDROME

Due to damaged medial temporal brain structures – Hippocampus

ORIENTING SYSTEM!



1. **Unlimited anterograde amnesia**
   Cannot search for new categories

2. **Limited retrograde amnesia**
   Direct access

3. **Failure of consolidation**
   Squire & Cohen, 1994

4. **Defective novelty reactions**
   Perseveration
   O' Keefe & Nadel, 1978

5. **Memory consolidation and novelty detection**
   Mediated by same structures
   Zola-Morgan & Squire, 1990

A FORMAL AMNESIC SYNDROME

5. **Normal priming**

   Baddeley & Warrington (1970)

   Mattis & Kovner (1984)

6. **Learning of first item dominates**

   Gray (1982)

7. **Impaired ability to attend to relevant dimensions of stimuli**

   Butters & Cermak (1975); Pribram (1986)

VIGILANCE DATA IN INFEROTEMPORAL CORTEX

RECEPTIVE FIELD SELECTIVITY MUST BE LEARNED

- Some cells respond selectively to particular views of particular faces
- Other cells respond to broader features of an animal's environment

Desimone, Gross, Perrett, …

EASY vs. DIFFICULT DISCRIMINATIONS: VIGILANCE!

"In the difficult condition the animals adopted a stricter internal criterion for discriminating matching from non-matching stimuli… The animal's internal representations of the stimuli were better separated … increased effort appeared to cause enhancement of the responses and sharpened selectivity for attended stimuli…"

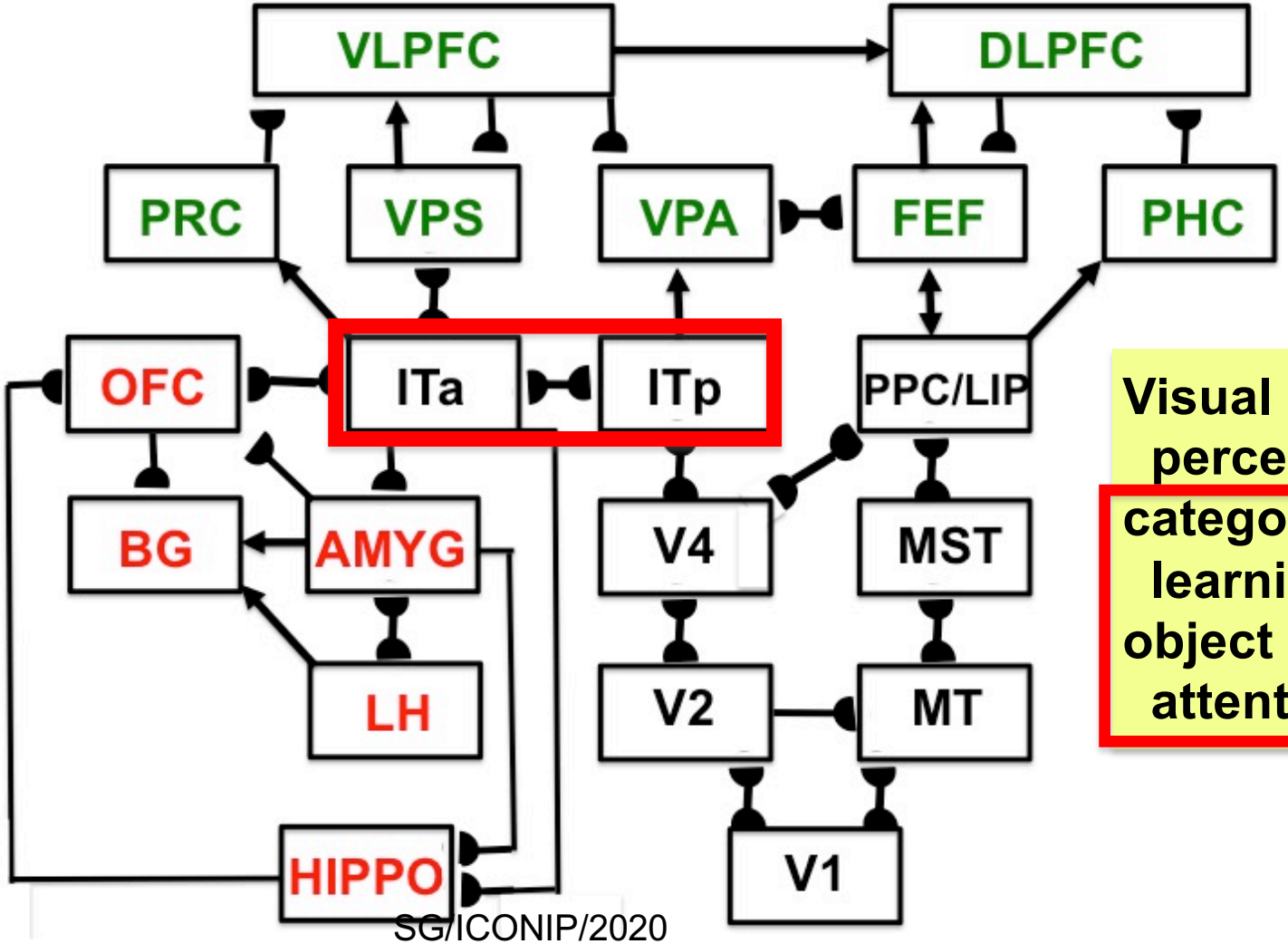
Spitzer, Desimone, and Moran, 1988

Predictive ART, or pART, architecture macrocircuit

How prefrontal cortex learns to control all higher-order intelligence

Grossberg (2018; see sites.bu.edu/steveg)

Working memory, learned plans, prediction, optimized action

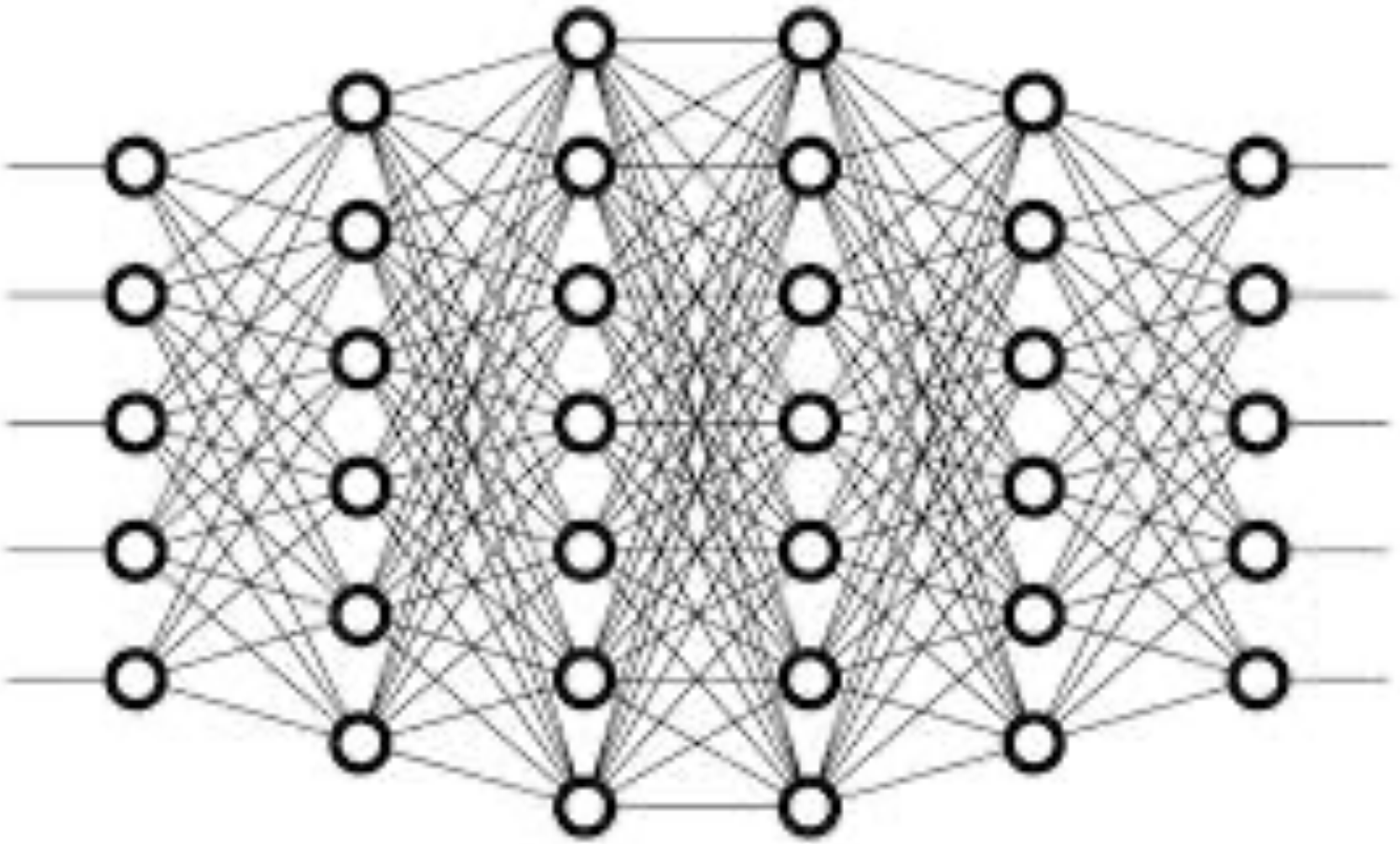


Reinforcement learning, emotion, motivation, adaptively-timed learning,

Visual perception, category learning, object attention

EACH BRAIN REGION IN NATURE AND IN pART CARRIES OUT A DIFFERENT FUNCTION

CONTRAST THE HOMOGENEOUS ORGANIZATION OF A TYPICAL DEEP LEARNING NETWORK



ALL BIOLOGICAL MODELS OF PERCEPTION, COGNITION, EMOTION, AND ACTION ARE EXPLAINABLE

Perceptual and cognitive processes use ART-like excitatory matching and match-based learning to create self-stabilizing attentive and conscious representations of objects and events that embody increasing expertise about the world

Complementary spatial and motor processes use inhibitory matching and mismatch-based learning to continually update spatial and motor representations to compensate for bodily changes throughout life

Together they provide a self-stabilizing perceptual and cognitive front end for conscious awareness and knowledge acquisition, which can intelligently manipulate more labile spatial and motor processes that enable our changing bodies to act effectively on a changing world

ALL BIOLOGICAL MODELS OF PERCEPTION, COGNITION, EMOTION, AND ACTION ARE EXPLAINABLE

Taken together, they provide a blueprint for designing

AUTONOMOUS ADAPTIVE ALGORITHMS

and

MOBILE ROBOTS

with behaviors humans can understand and control

See sites.bu.edu/steveg for these models