Innovative Applications and Technology Pivots – A Perfect Storm in Computing

Wen-mei Hwu

Professor and Sanders-AMD Chair, ECE, NCSA

University of Illinois at Urbana-Champaign

with

Jinjun Xiong (IBM, C3SR Co-Director), Abdul Dakkak and Carl Pearson

ECE ILLINOIS

The IMPACT Research Group

C3SR center for cognitive computing systems research

ILLINOIS

Agenda

- Revolutionary paradigm shift in applications
- Technology pivot to heterogeneous computing
- Cognitive computing systems research

A major paradigm shift

- **In the 20th Century, we were able to understand, design, and manufacture what we can measure**
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, …

A major paradigm shift

- In the 20th Century, we were able to understand, design, and manufacture what we can measure
  - Physical instruments and computing systems allowed us to see farther, capture more, communicate better, understand natural processes, control artificial processes…
- **In the 21st Century, we are able to understand, design, and create what we can compute**
  - Computational models are allowing us to see even farther, going back and forth in time, learn better, test hypothesis that cannot be verified any other way, …

Examples of Paradigm Shift

20th Century	21st Century
Small mask patterns	Optical proximity correction
Electronic microscope and Crystallography with computational image processing	Computational microscope with initial conditions from Crystallography
Anatomic imaging with computational image processing	Metabolic imaging sees disease before visible anatomic change
Optical telescopes	Gravitational wave telescopes
Teleconference	Tele-emersion – augmented reality
GPS	Self-driving cars

What is powering the paradigm shift?

- Large clusters (scale out) allow solving realistic problems
  - 1.5 Peta bytes of DRAM in Illinois Blue Waters
  - E.g., 0.5 Å (0.05 nm) grid spacing is needed for accurate molecular dynamics
    - interesting biological systems have dimensions of mm or larger
    - Thousands of nodes are required to hold and update the grid points.

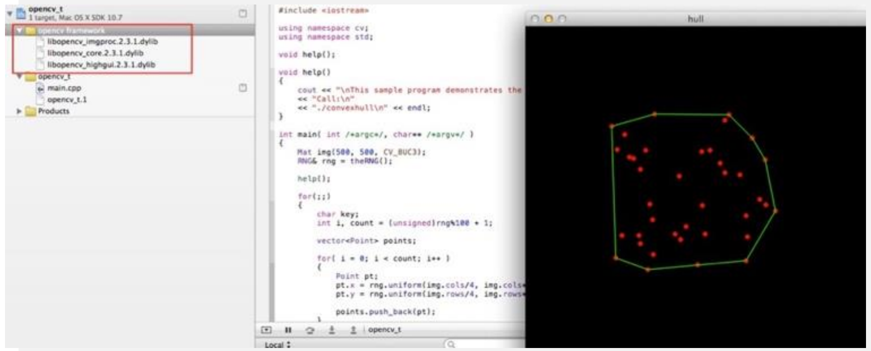
- Fast nodes (scale up) allow solution at realistic time scales
  - Simulation time steps at femtosecond (10^{-15} second) level needed for accuracy
    - Biological processes take milliseconds or longer
    - Current molecular dynamics simulations progress at about one day for each 100 microseconds of the simulated process.
    - Interesting computational experiments take weeks (used to be months)

What types of applications are demanding computing power today?

- First-principle-based models
  - Problems that we know how to solve accurately but choose not to because it would be “too expensive”
  - High-valued applications with approximations that cause inaccuracies and lost opportunities
  - Medicate imaging, earthquake modeling, weather modeling, astrophysics modeling, precision digital manufacturing, combustion modeling, ….

- Applications that we have failed to program
  - Problems that we just don’t know how to solve
  - High-valued applications with no effective computational methods
  - Computer vision, natural language dialogs, stock trading, fraud detection, …

We know what we want but don’t know how to build it.



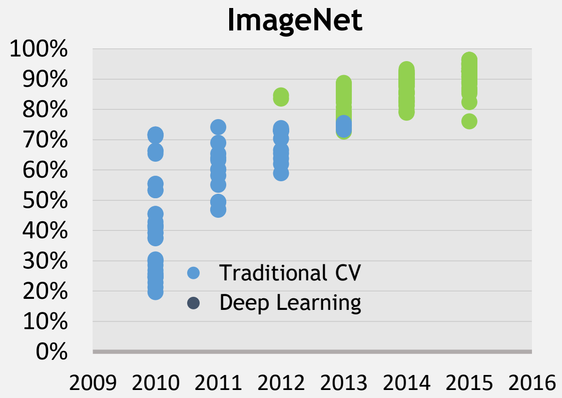
opencv_t
1 target, Mac OS X SDK 10.7
opencv framework
libopencv_imgproc.2.3.1.dylib
libopencv_core.2.3.1.dylib
libopencv_highgui.2.3.1.dylib
opencv_t
main.cpp
opencv_t.1
Products
opencv_t
Local
hull

```
#include <iostream>

using namespace cv;
using namespace std;

void help();

void help()
{
    cout << "\nThis sample program demonstrates the
    << "Call:\n"
    << "./convexhull\n" << endl;
}

int main( int /*argc*/, char** /*argv*/ )
{
    Mat img(500, 500, CV_8UC3);
    RNG& rng = theRNG();

    help();

    for(;;)
    {
        char key;
        int i, count = (unsigned)rng%100 + 1;

        vector<Point> points;

        for( i = 0; i < count; i++ )
        {
            Point pt;
            pt.x = rng.uniform(img.cols/4, img.cols
            pt.y = rng.uniform(img.rows/4, img.rows

            points.push_back(pt);
        }
```

Traditional Computer Vision
Experts + Time

Deep Learning Object Detection
DNN + Data + HPC

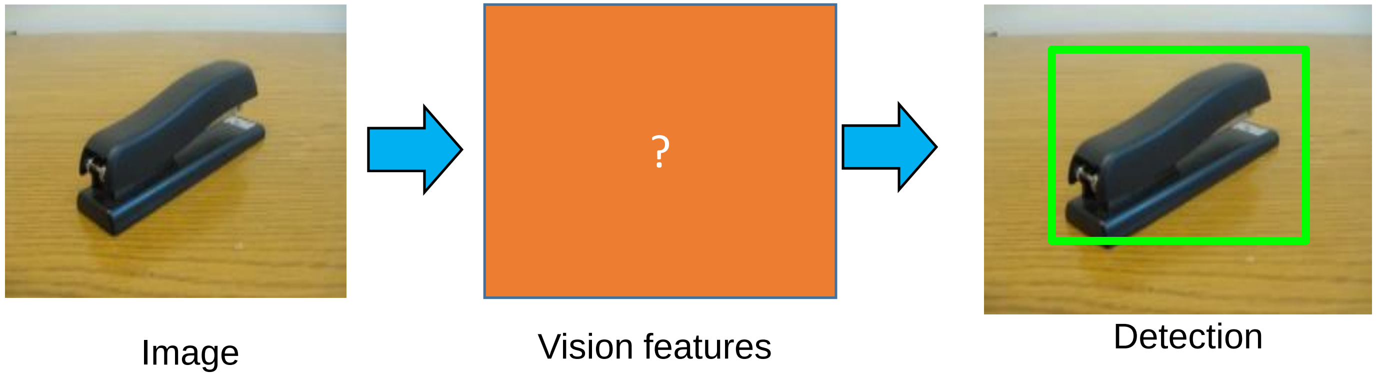
2M training images



Deep Learning Achieves
“Superhuman” Results

Slide courtesy of Steve Oberlin, NVIDIA

Some different modalities of Real-world Data



Image

?

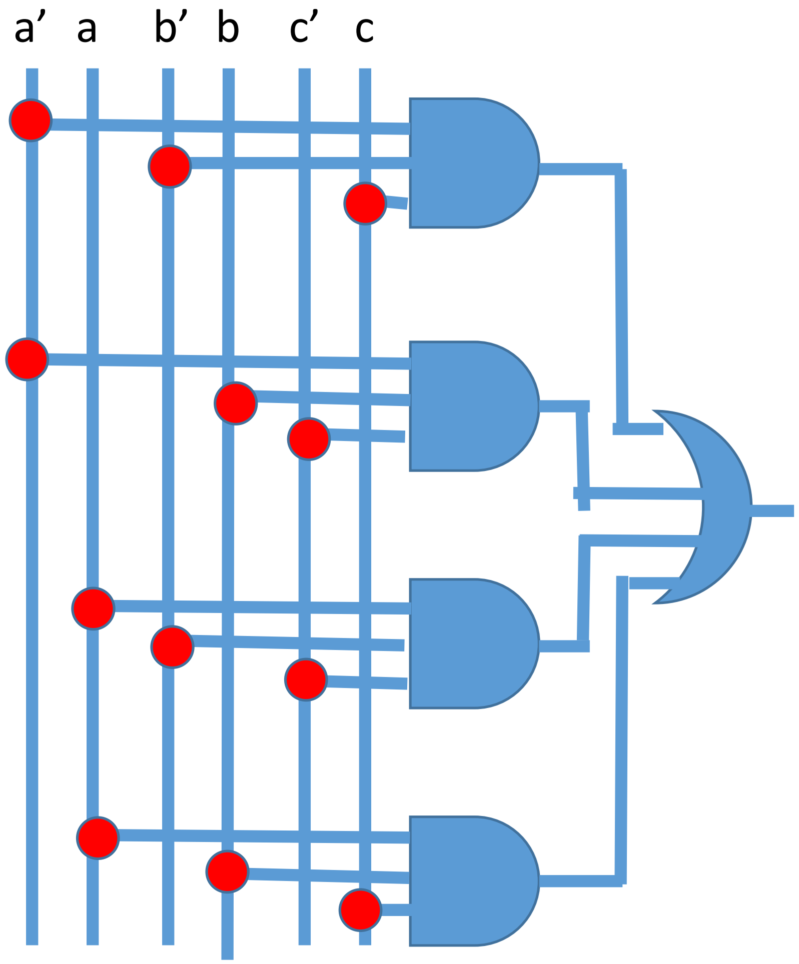
Vision features

Detection

This seems to be a combinational logic design problem.

Combinations Logic Specification – Truth Table

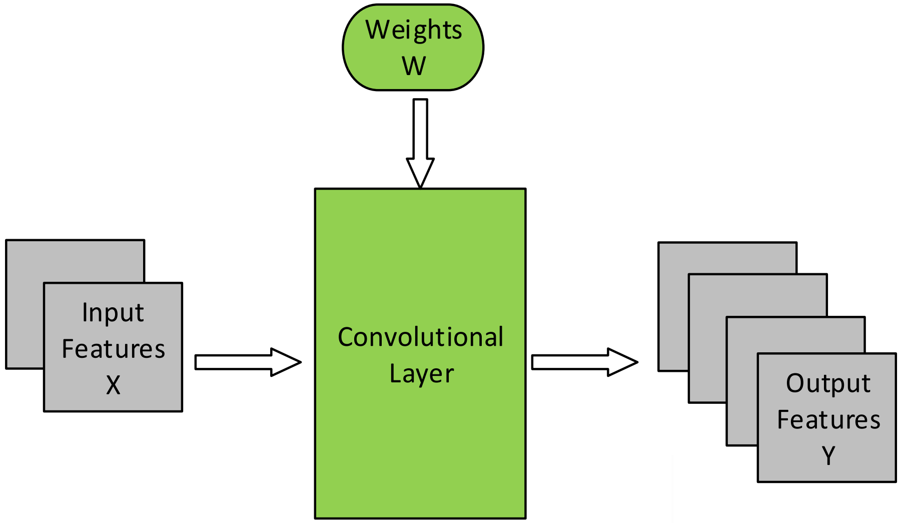
Input			output
a	b	c	
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	0
1	0	0	1
1	0	1	0
1	1	0	0
1	1	1	1



What if we did not know the truth table?

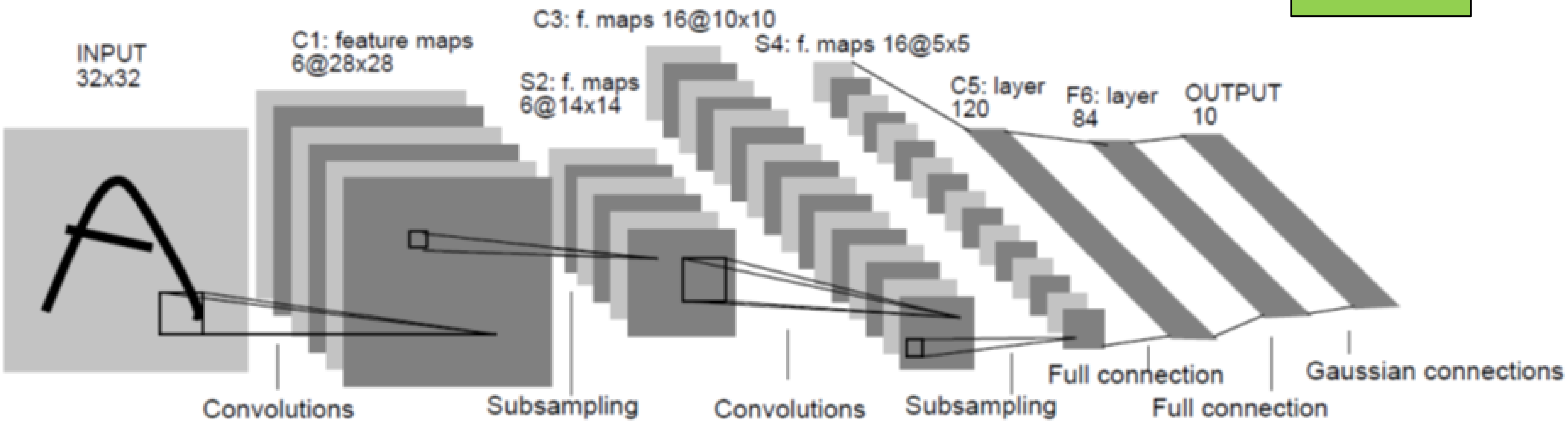
- Look at enough observation data to construct the rules
  - 000 → 0
  - 011 → 0
  - 100 → 1
  - 110 → 0
- If we have enough observational data to cover all input patterns, we can construct the truth table and derive the logic!

LeNet-5, a convolutional neural network for hand-written digit recognition.



This is a 1024*8 bit input, which will have a truth table of 2 ^{8196} entries

1M training data is approximately 0%

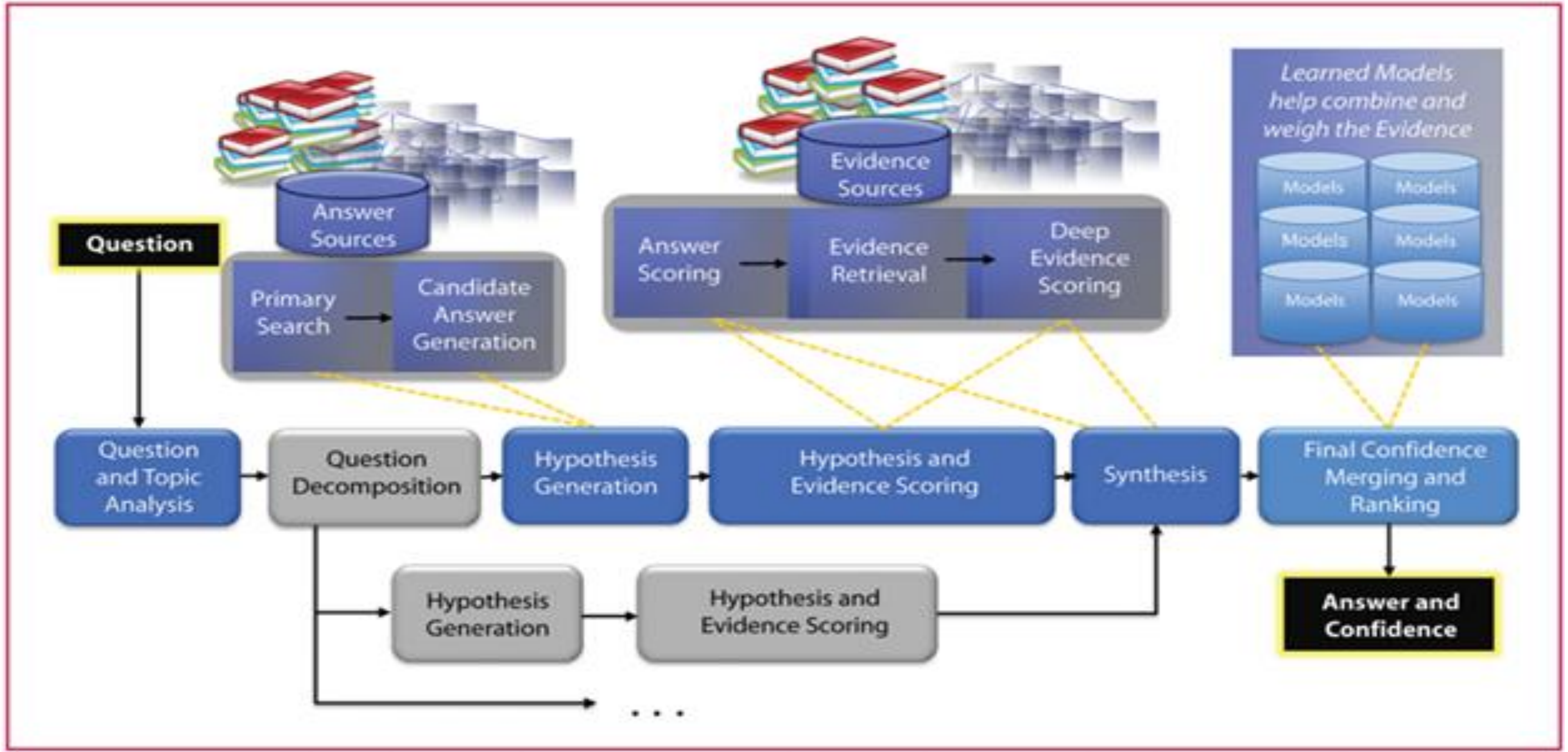


The adoption of full cognitive business applications has exploded since .

JEOPARDY!

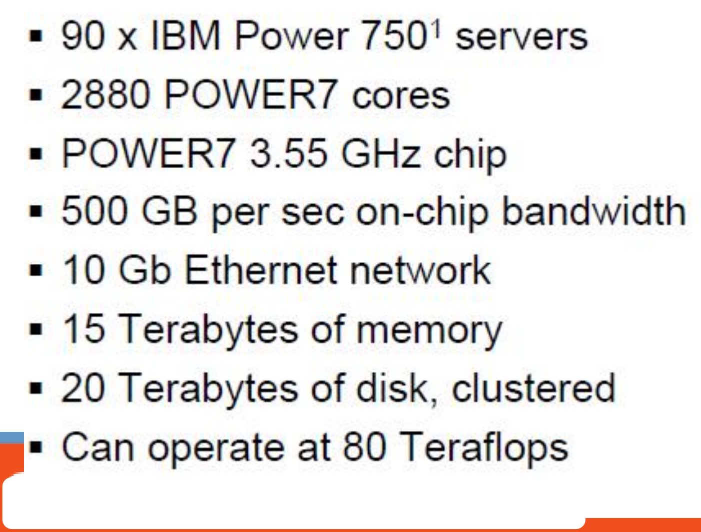
The IBM Challenge

Back in 2011



3 sec response time!

The cognitive application is built and optimized for the underlying infrastructure manually



- 90 x IBM Power 750[1] servers
- 2880 POWER7 cores
- POWER7 3.55 GHz chip
- 500 GB per sec on-chip bandwidth
- 10 Gb Ethernet network
- 15 Terabytes of memory
- 20 Terabytes of disk, clustered
- Can operate at 80 Teraflops

Illinois-IBM C^{3}SR faculties & students (Launched 9/20/2016)

Suma Bhat

Minh Do

Deming Chen

Julia Hockenmaier

Wen-mei Hwu

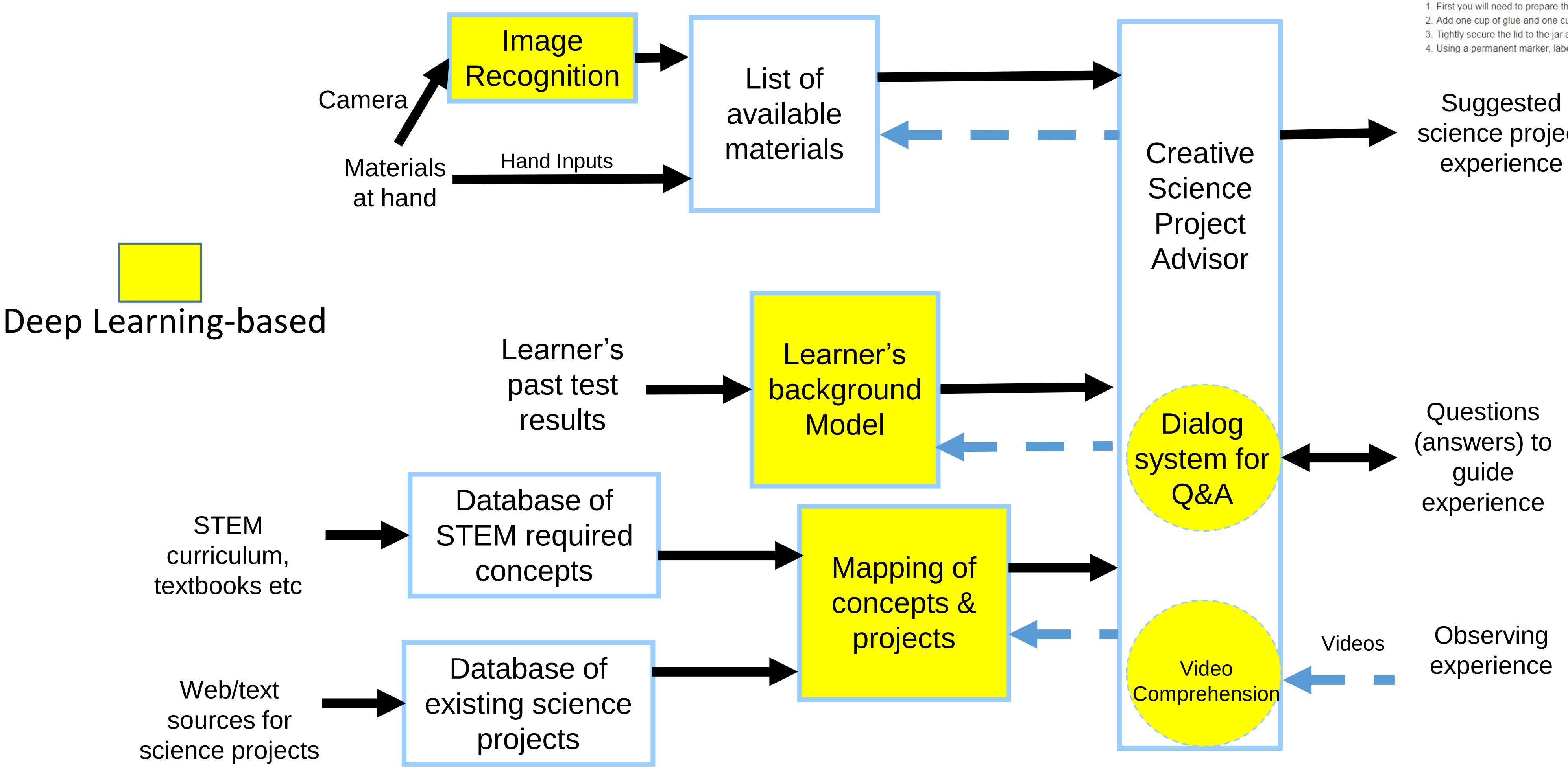
Nam Sung Kim

Dan Roth

Rakesh Nagi

Lav Varshney

A C3SR App: CELA for Personalized Education



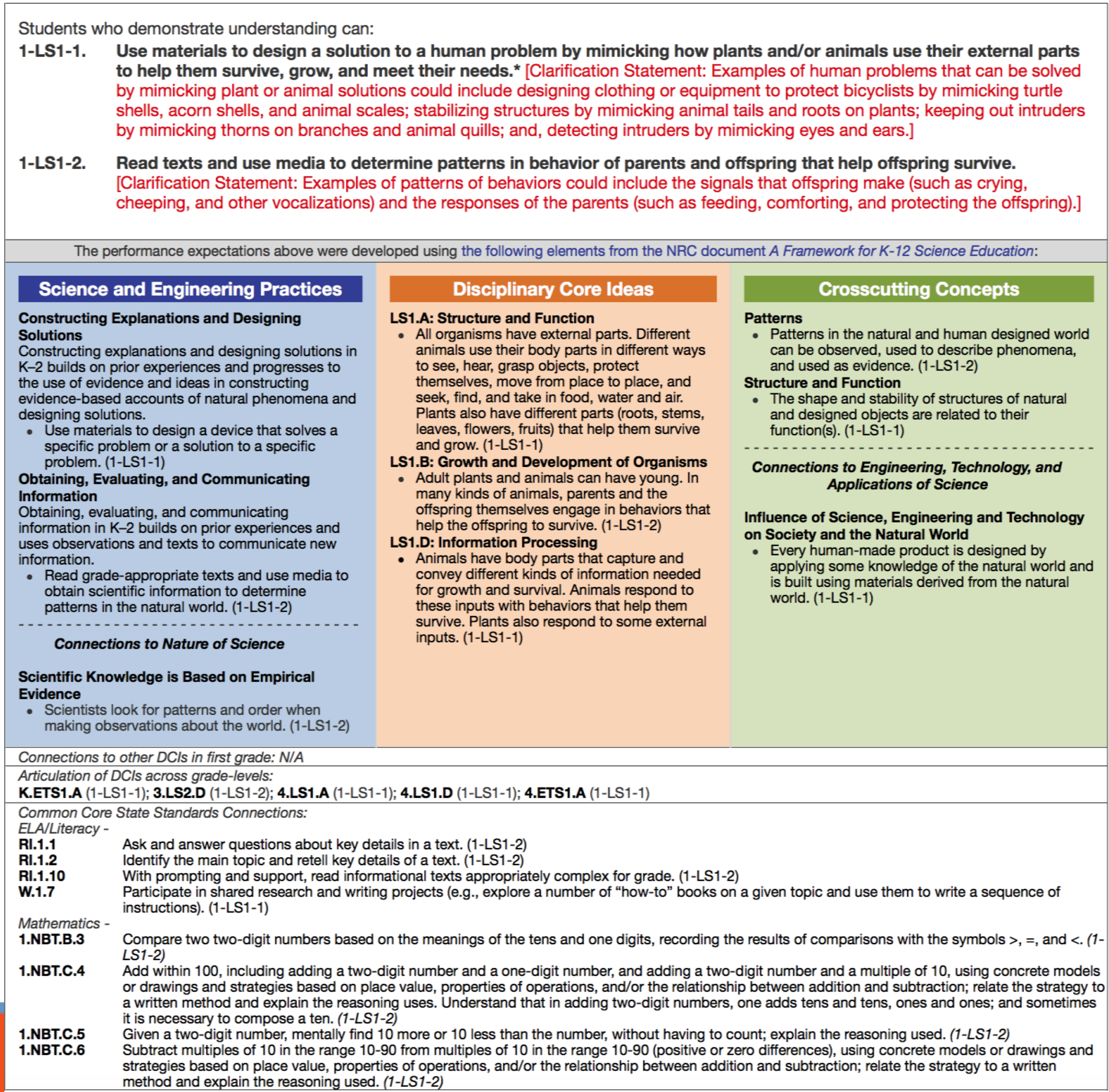
Camera
Image Recognition
Materials at hand
Hand Inputs
List of available materials
Creative Science Project Advisor
Suggested science project experience
🟨 Deep Learning-based
Learner's past test results
Learner's background Model
Dialog system for Q&A
Questions (answers) to guide experience
STEM curriculum, textbooks etc
Database of STEM required concepts
Mapping of concepts & projects
Web/text sources for science projects
Database of existing science projects
Video Comprehension
Videos
Observing experience

Experimental Procedure

1. First you will need to prepare the 50% glue solution, which is made up of half glue and half water.
2. Add one cup of glue and one cup of water to one of the jars.
3. Tightly secure the lid to the jar and shake until glue is fully diluted, and no gooey clumps remain.
4. Using a permanent marker, label this jar "50% Glue".

Extract concept graphs from next generation science standard (http://www.nextgenscience.org/)

- Five blocks of information:
  - Performance Expectations
  - Science and Engineering Practices
  - Disciplinary Core Ideas
  - Crosscutting Concepts
  - Connections



Students who demonstrate understanding can:

1-LS1-1. **Use materials to design a solution to a human problem by mimicking how plants and/or animals use their external parts to help them survive, grow, and meet their needs.*** [Clarification Statement: Examples of human problems that can be solved by mimicking plant or animal solutions could include designing clothing or equipment to protect bicyclists by mimicking turtle shells, acorn shells, and animal scales; stabilizing structures by mimicking animal tails and roots on plants; keeping out intruders by mimicking thorns on branches and animal quills; and, detecting intruders by mimicking eyes and ears.]

1-LS1-2. **Read texts and use media to determine patterns in behavior of parents and offspring that help offspring survive.** [Clarification Statement: Examples of patterns of behaviors could include the signals that offspring make (such as crying, cheeping, and other vocalizations) and the responses of the parents (such as feeding, comforting, and protecting the offspring).]

The performance expectations above were developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices

Constructing Explanations and Designing Solutions
Constructing explanations and designing solutions in K–2 builds on prior experiences and progresses to the use of evidence and ideas in constructing evidence-based accounts of natural phenomena and designing solutions.

- Use materials to design a device that solves a specific problem or a solution to a specific problem. (1-LS1-1)

Obtaining, Evaluating, and Communicating Information
Obtaining, evaluating, and communicating information in K–2 builds on prior experiences and uses observations and texts to communicate new information.

- Read grade-appropriate texts and use media to obtain scientific information to determine patterns in the natural world. (1-LS1-2)

Connections to Nature of Science

Scientific Knowledge is Based on Empirical Evidence

- Scientists look for patterns and order when making observations about the world. (1-LS1-2)

Disciplinary Core Ideas

LS1.A: Structure and Function

- All organisms have external parts. Different animals use their body parts in different ways to see, hear, grasp objects, protect themselves, move from place to place, and seek, find, and take in food, water and air. Plants also have different parts (roots, stems, leaves, flowers, fruits) that help them survive and grow. (1-LS1-1)

LS1.B: Growth and Development of Organisms

- Adult plants and animals can have young. In many kinds of animals, parents and the offspring themselves engage in behaviors that help the offspring to survive. (1-LS1-2)

LS1.D: Information Processing

- Animals have body parts that capture and convey different kinds of information needed for growth and survival. Animals respond to these inputs with behaviors that help them survive. Plants also respond to some external inputs. (1-LS1-1)

Crosscutting Concepts

Patterns

- Patterns in the natural and human designed world can be observed, used to describe phenomena, and used as evidence. (1-LS1-2)

Structure and Function

- The shape and stability of structures of natural and designed objects are related to their function(s). (1-LS1-1)

Connections to Engineering, Technology, and Applications of Science

Influence of Science, Engineering and Technology on Society and the Natural World

- Every human-made product is designed by applying some knowledge of the natural world and is built using materials derived from the natural world. (1-LS1-1)

Connections to other DCIs in first grade: N/A

Articulation of DCIs across grade-levels:
K.ETS1.A (1-LS1-1); **3.LS2.D** (1-LS1-2); **4.LS1.A** (1-LS1-1); **4.LS1.D** (1-LS1-1); **4.ETS1.A** (1-LS1-1)

Common Core State Standards Connections:

ELA/Literacy -

RI.1.1	Ask and answer questions about key details in a text. (1-LS1-2)
RI.1.2	Identify the main topic and retell key details of a text. (1-LS1-2)
RI.1.10	With prompting and support, read informational texts appropriately complex for grade. (1-LS1-2)
W.1.7	Participate in shared research and writing projects (e.g., explore a number of “how-to” books on a given topic and use them to write a sequence of instructions). (1-LS1-1)

Mathematics -

1.NBT.B.3	Compare two two-digit numbers based on the meanings of the tens and one digits, recording the results of comparisons with the symbols >, =, and <. *(1-LS1-2)*
1.NBT.C.4	Add within 100, including adding a two-digit number and a one-digit number, and adding a two-digit number and a multiple of 10, using concrete models or drawings and strategies based on place value, properties of operations, and/or the relationship between addition and subtraction; relate the strategy to a written method and explain the reasoning uses. Understand that in adding two-digit numbers, one adds tens and tens, ones and ones; and sometimes it is necessary to compose a ten. *(1-LS1-2)*
1.NBT.C.5	Given a two-digit number, mentally find 10 more or 10 less than the number, without having to count; explain the reasoning used. *(1-LS1-2)*
1.NBT.C.6	Subtract multiples of 10 in the range 10-90 from multiples of 10 in the range 10-90 (positive or zero differences), using concrete models or drawings and strategies based on place value, properties of operations, and/or the relationship between addition and subtraction; relate the strategy to a written method and explain the reasoning used. *(1-LS1-2)*

Paradigm shift for cognitive application development

- Traditional programming approaches failed to deliver cognitive applications for decades

- With the wide adoption of machine learning (deep learning), the core of application development has shifted to model training (including model customization)
  - Experimentation with a large amount of data is on the critical path of application development
  - The nature of functional verification, performance tuning, and debugging is fundamentally different

Cognitive Application Builder (CAB)

A system-level challenge

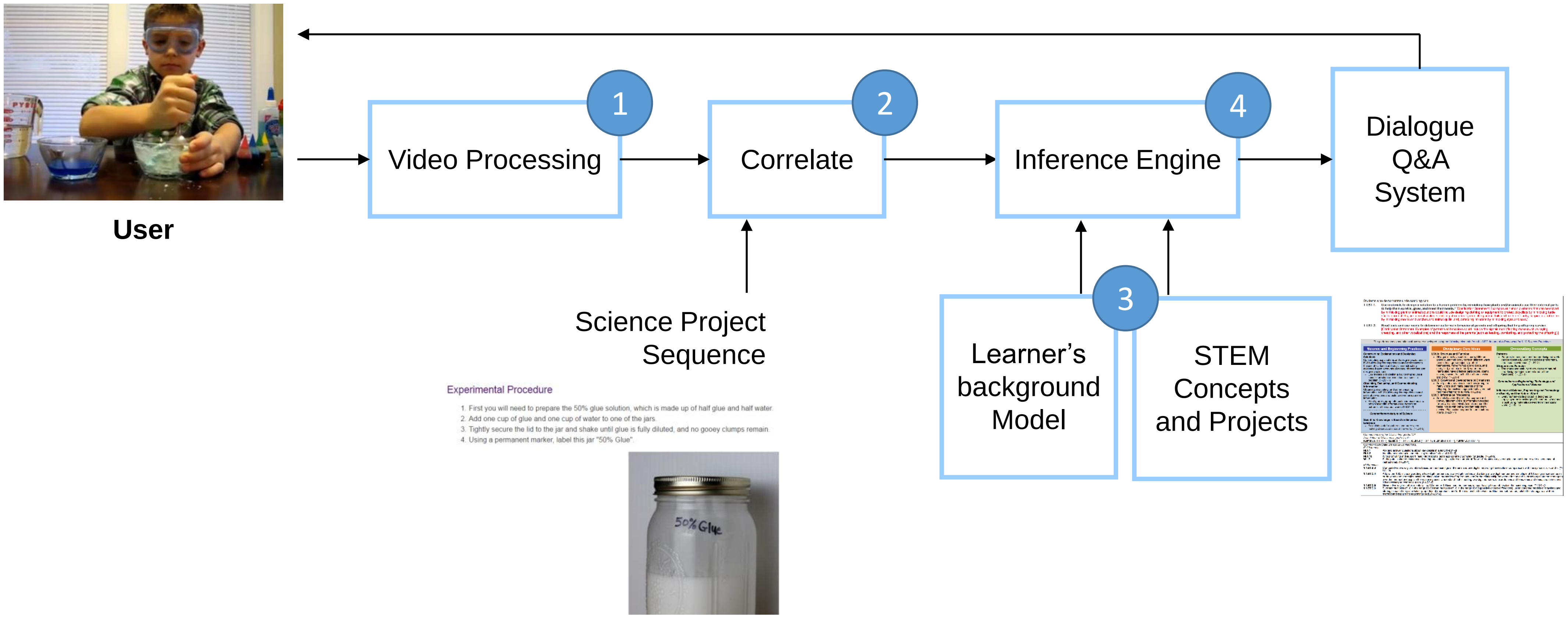
Workflow description
Innovative AI techniques

→

High-performance, scalable,
robust applications

- CAB: A language, compiler, and runtime for easy development of cognitive applications
  - System-aware to exploit accelerators and efficient communication
  - Introspection for debugging and performance evaluation
  - Workflow optimization and orchestration for system-level performance
  - Decentralized application architecture for scalability, composability, testing, and development

CELA as a driving use case for CAB

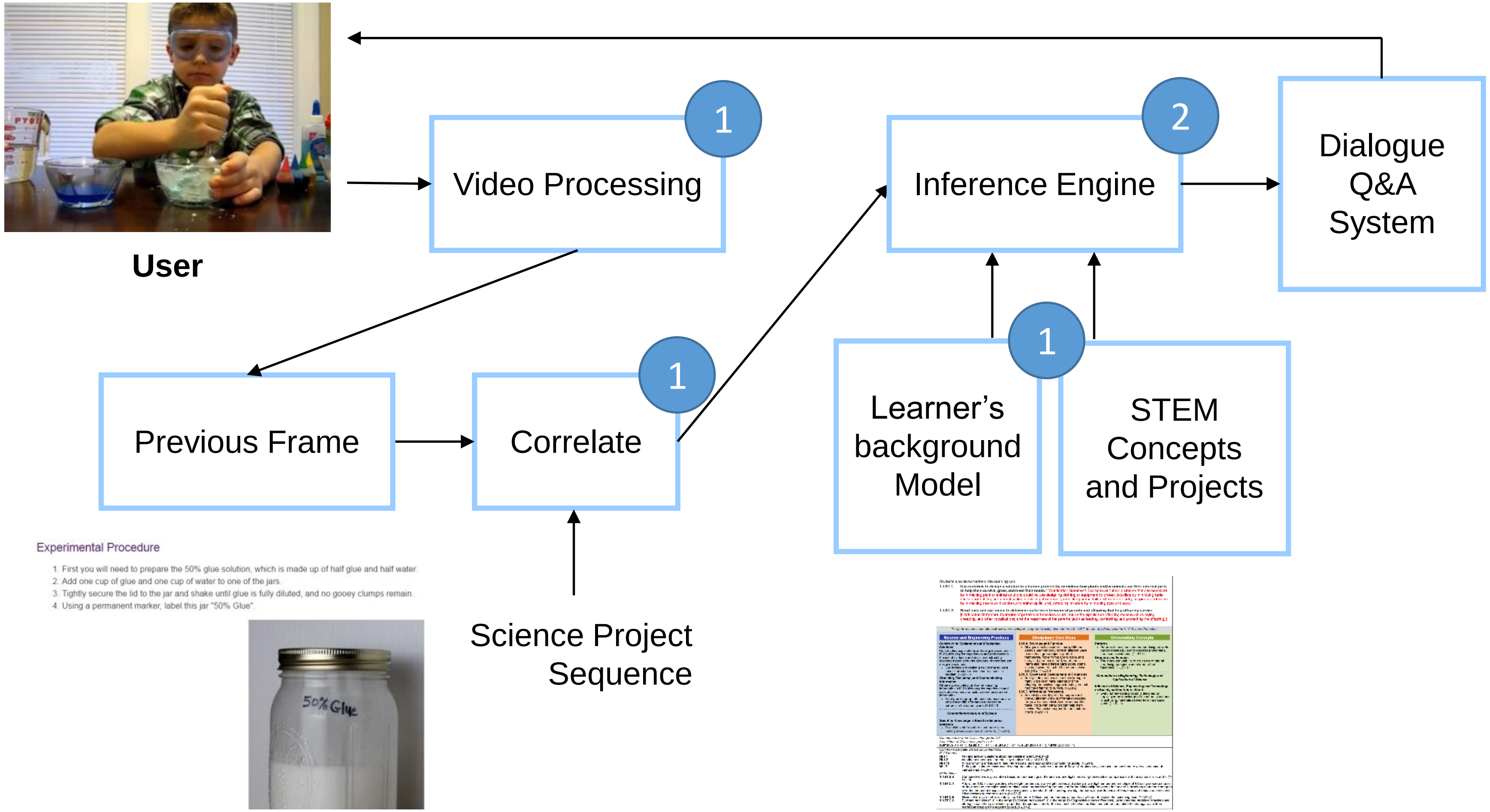


Experimental Procedure

1. First you will need to prepare the 50% glue solution, which is made up of half glue and half water.
2. Add one cup of glue and one cup of water to one of the jars.
3. Tightly secure the lid to the jar and shake until glue is fully diluted, and no gooey clumps remain.
4. Using a permanent marker, label this jar "50% Glue".

- CAB will simplify component connection, workflow description, and iterative development

CELA Time Warp for Efficiency



Experimental Procedure

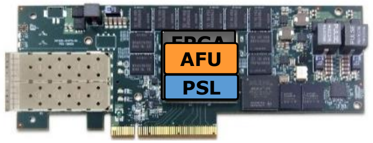
1. First you will need to prepare the 50% glue solution, which is made up of half glue and half water.
2. Add one cup of glue and one cup of water to one of the jars.
3. Tightly secure the lid to the jar and shake until glue is fully diluted, and no gooey clumps remain.
4. Using a permanent marker, label this jar "50% Glue".

- CAB automatically transforms workflows for high-performance execution

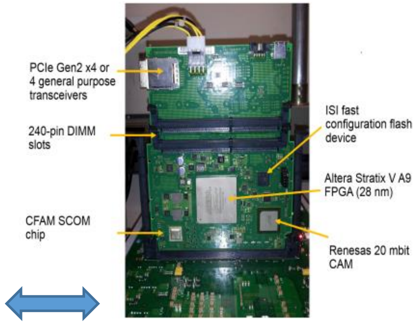
C3SR Experimental Heterogeneous Infrastructure

DGX-1

2 x P8 Minsky with NVLink GPUs



FPGA CAPI over PCIe

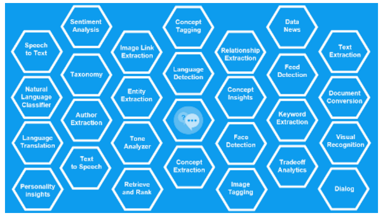

PCIe Gen2 x4 or 4 general purpose transceivers
240-pin DIMM slots
CFAM SCOM chip
ISI fast configuration flash device
Altera Stratix V A9 FPGA (28 nm)
Renesas 20 mbit CAM

ConTutto over DMI

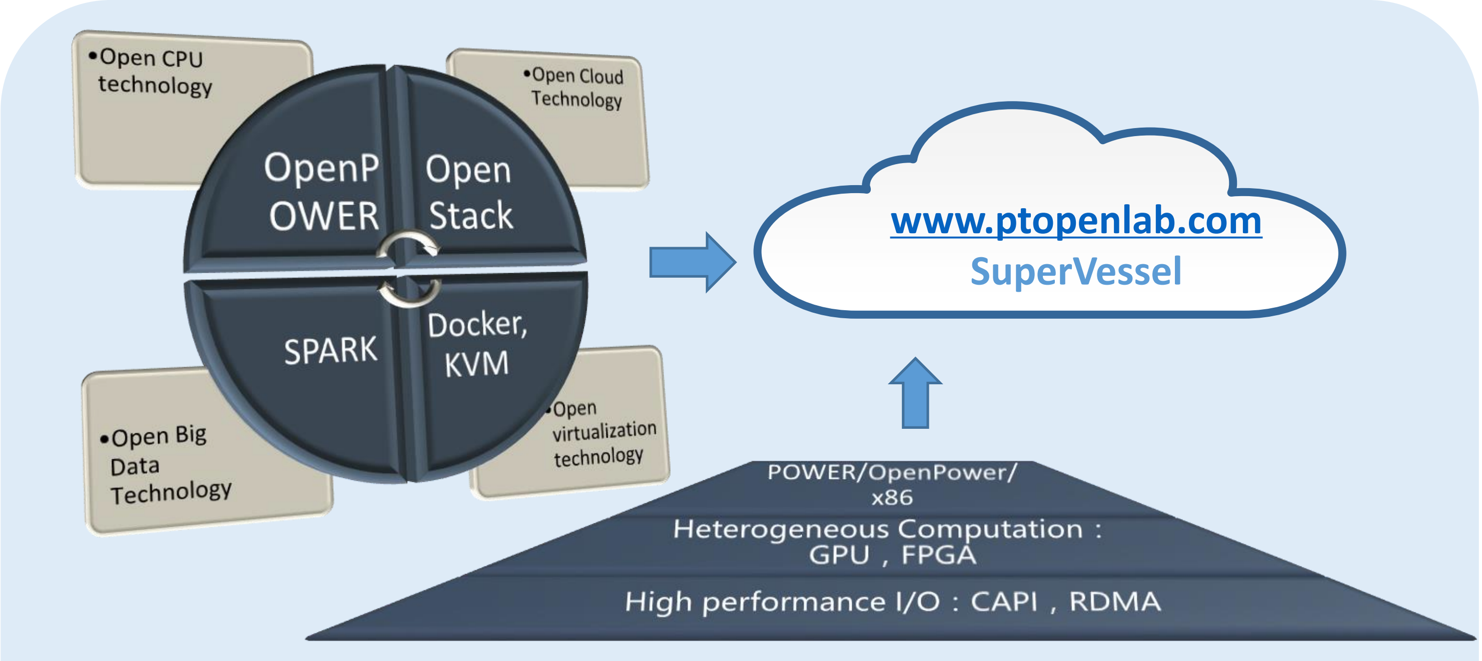
4 x P8 Tuleta (S824L)

IBM Bluemix™

IBM SOFTLAYER®



Watson developer cloud

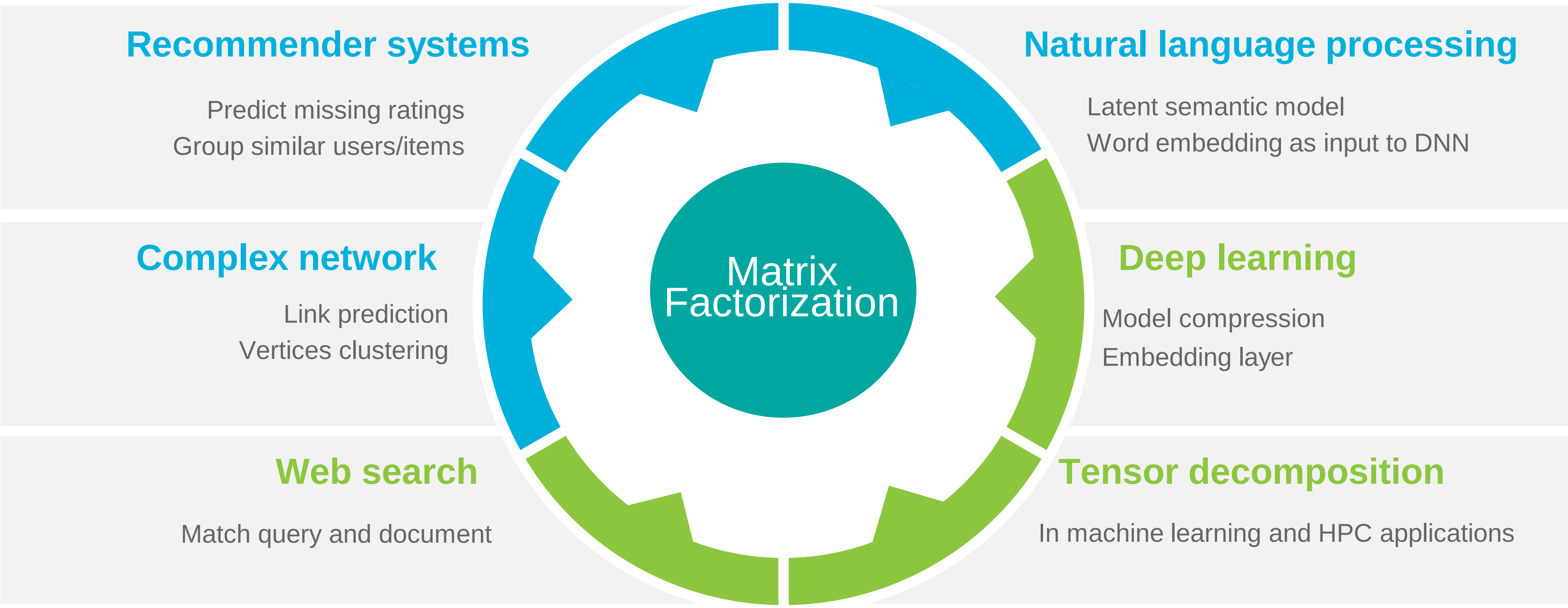

•Open CPU technology
•Open Cloud Technology
•Open Big Data Technology
•Open virtualization technology
OpenPOWER
Open Stack
SPARK
Docker, KVM
www.ptopenlab.com
SuperVessel
POWER/OpenPower/ x86
Heterogeneous Computation : GPU , FPGA
High performance I/O : CAPI , RDMA

Courtesy: Jinjun Xiong, IBM

Workload acceleration research at C3SR based on CAB/TANGRAM Software Synthesis

- Focus on impactful cognitive workloads for acceleration
  - Matrix factorization on GPU
  - Long-term Recurrent Convolutional Network acceleration
  - ResNet inference acceleration
  - Neuron Machine Translation acceleration
  - DNN inference acceleration
  - Graph analytic acceleration
- In discussion with other CHN centers to collect performance critical cognitive workloads
- Plan to deliver a set of cognitive benchmarks optimized for OpenPOWER

Matrix factorization: one of key workloads



Recommender systems

Predict missing ratings
Group similar users/items

Natural language processing

Latent semantic model
Word embedding as input to DNN

Complex network

Link prediction
Vertices clustering

Deep learning

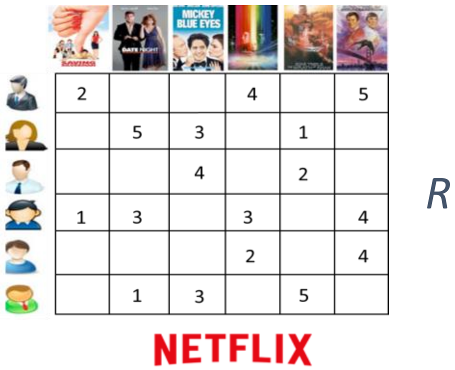
Model compression
Embedding layer

Web search

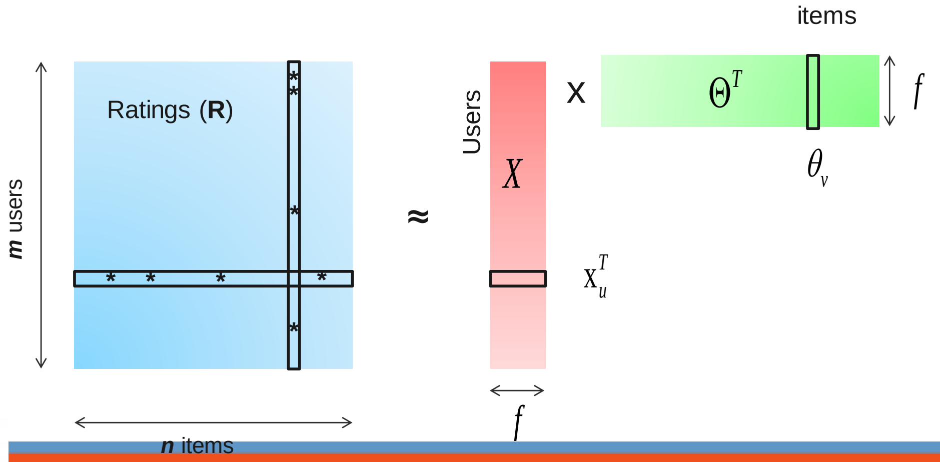
Match query and document

Tensor decomposition

In machine learning and HPC applications



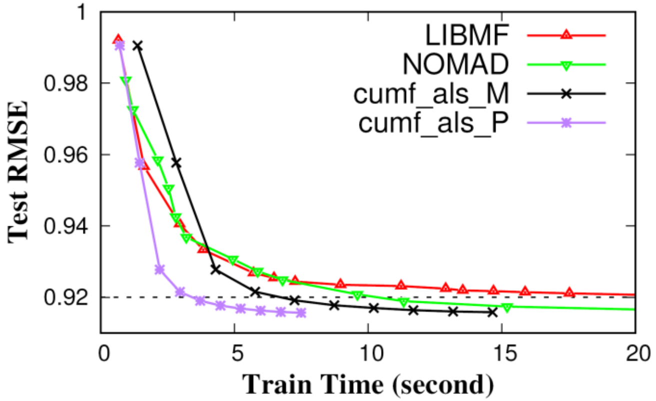
2			4		5
	5	3		1	
		4		2	
1	3		3		4
			2		4
	1	3		5	



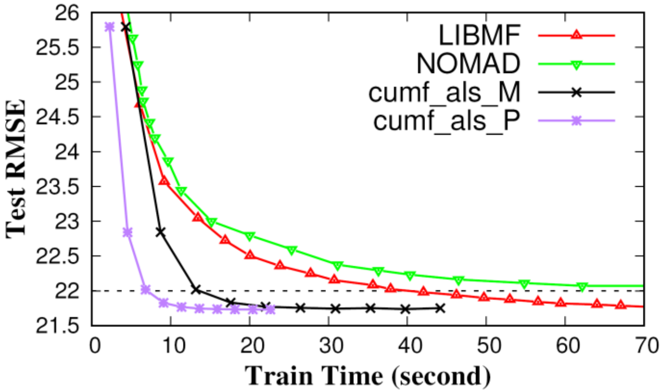
Spotify amazon.com Quora Music

cuMF acceleration

- cuMF formulation: 1$R \approx X \cdot \Theta^T$:rix R into
   - v$J = \sum_{u,v}(r_{uv} - \boldsymbol{x}_u^T\boldsymbol{\theta}_v)^2 + \lambda(\sum_u n_{x_u}||\boldsymbol{x}_u||^2 + \sum_v n_{\theta_v}||\boldsymbol{\theta}_v||^2)$
- Connect cuMF to Spark MLlib via JNI
- cuMF_ALS @4 Maxwell ($2.5/hour)
  ≈ 10x speedup over SparkALS @50 nodes
  ≈ 1% of SparkALS's cost ($0.53/hour/node)
- Open source @ http://github.com/cuMF/
- Demoed at SC'16 and GTC'16 on Minsky
- Presented to Jen-Hsun Huang on Feb 1, 2017



Netflix



Yahoo!Music

- cuMF_ALS w/ FP16 on Maxwell and Pascal
- LIBMF: 1 CPU w/ 40 threads
- NOMAD
   - 32 nodes for Netflix and Yahoo
- 2-10x as fast

Conclusion and Outlook

- Applications have very large appetite for more computing power
  - Both larger scale clusters and faster devices
- Heterogeneity has become the norm for all hardware systems
  - HPC community are currently seeing about 2-3x application speedup
  - Recent positive spiral between deep learning and GPU computing
- Cognitive Computing Systems Research
  - Game changing applications (CELA)
  - Next generation heterogeneous system – democratizing compute and bandwidth (100x)
  - High productivity development with software synthesis (CAB)