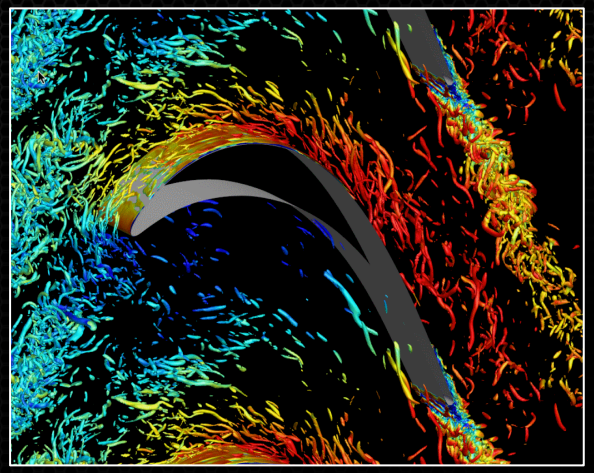
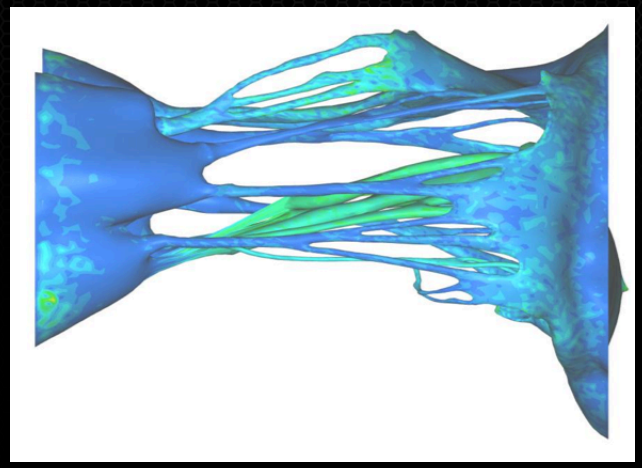
Accelerating Understanding

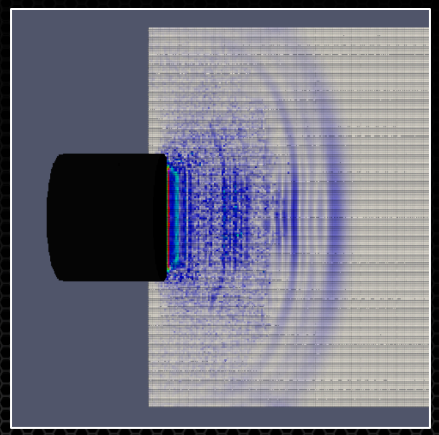
Data Analytics, Machine Learning, and GPUs

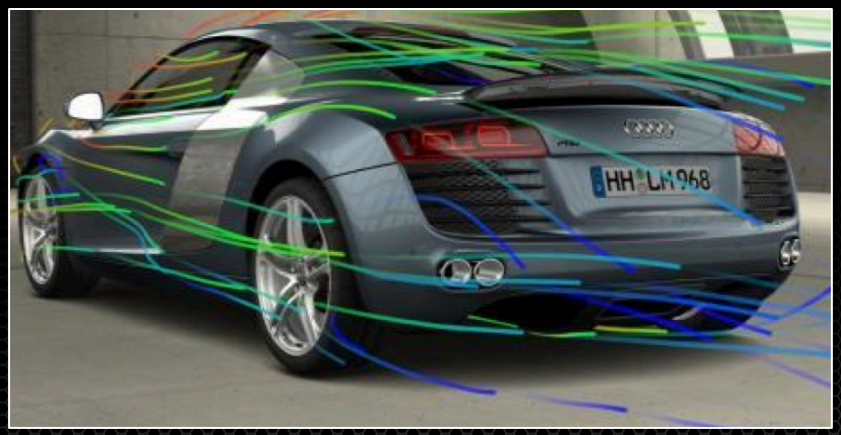
Steve Oberlin
CTO, Accelerated Computing

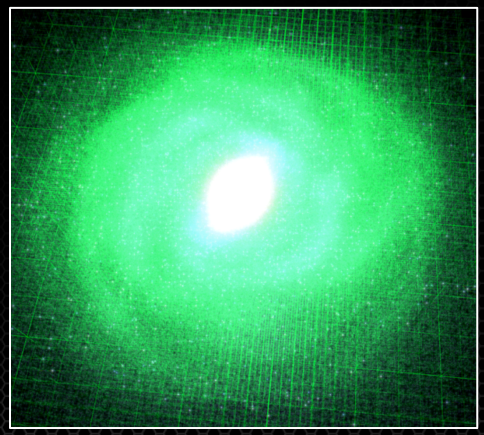
How Does HPC Touch Your Life?











How Does Your Life Touch HPC?

2007

Utility Apps

How Does Your Life Touch HPC?

How Does Your Life Touch HPC?

2014

Cloud Apps

How Does Your Life Touch HPC?

"Data intensive processing:
High throughput event processing and data capture from sensors, data feeds and instruments"

Pete Ungaro

"Cloud Computing:
App access to converged infrastructure via IP stack."

Bill Blake

We *are* the sensors, data feeds, and instruments.

The Age of Big Data

2.5 Exabytes of Web Data Created Daily

2.5 Petabytes of Customer Data Hourly

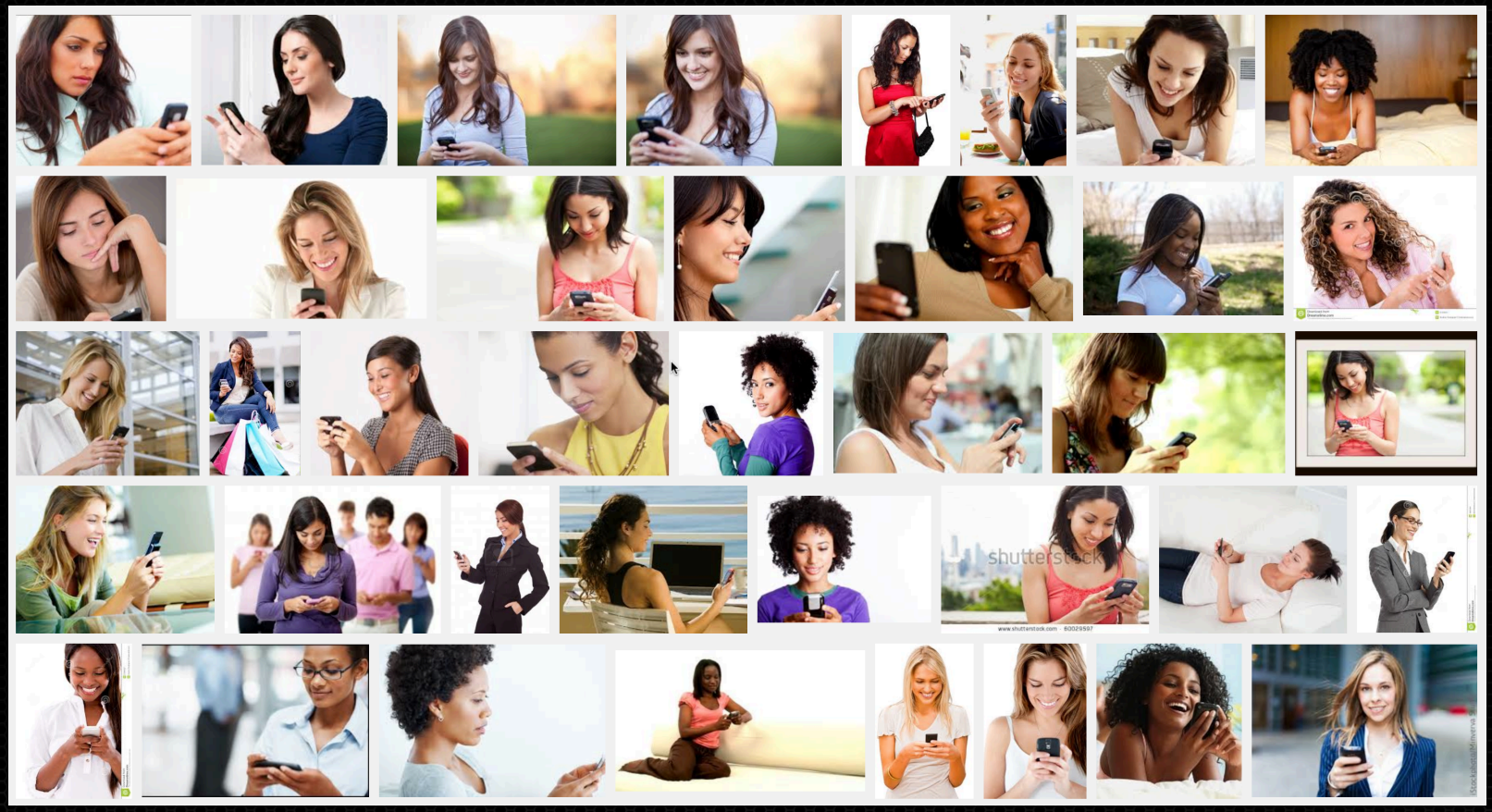
350 Million Images Uploaded a Day

100 Hours Video Uploaded Every Minute

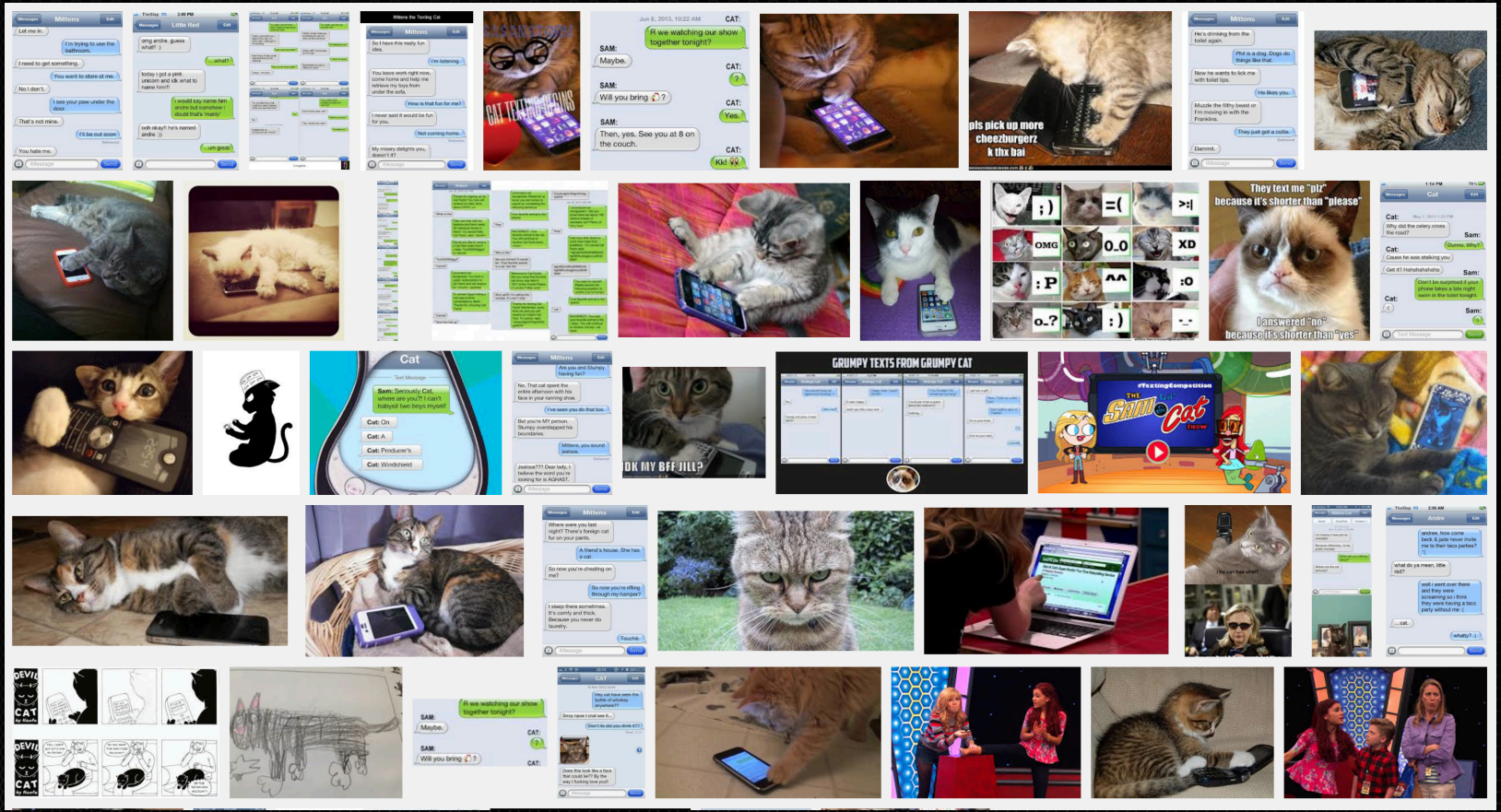
How can we organize, analyze, understand, benefit from such a trove of data?

Search: “Images Man Texting”

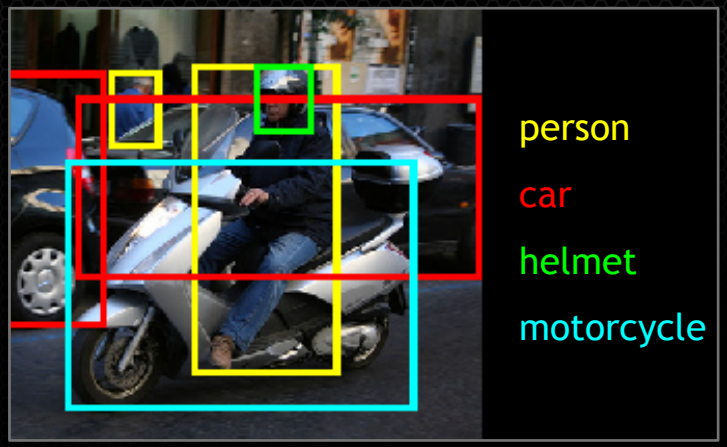
Search: “Images Woman Texting”

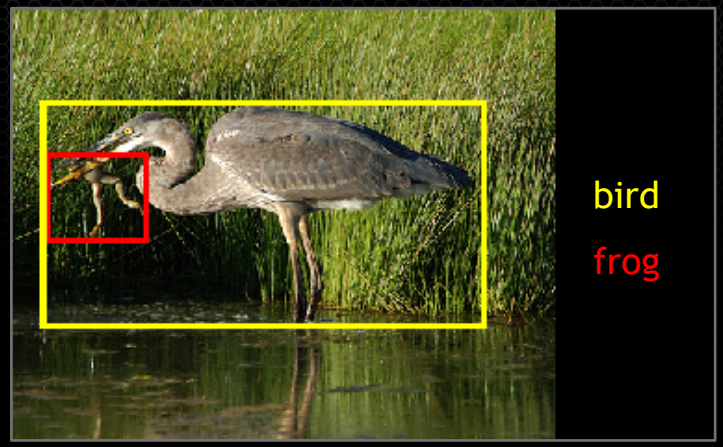


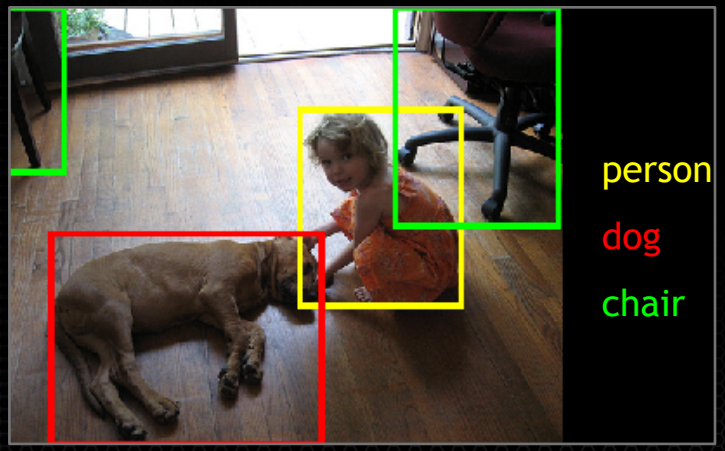
Search: “Images Cat Texting”



Machine Learning

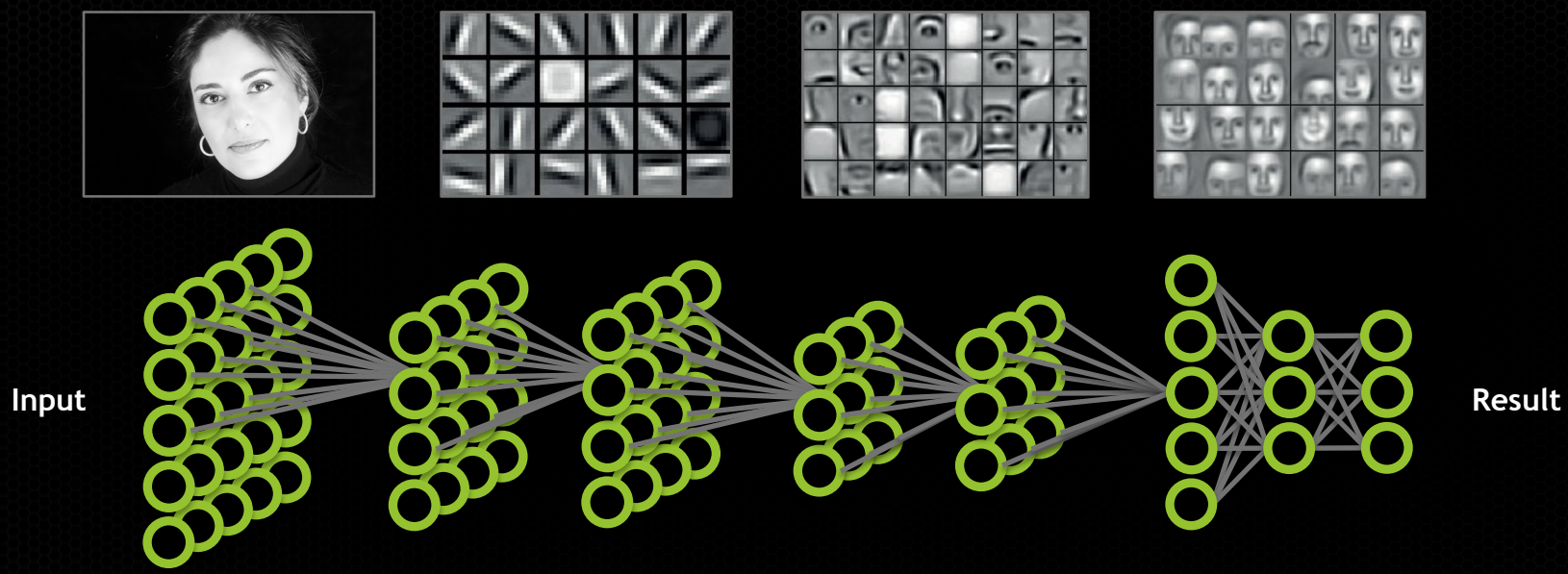






IMAGENET

Machine Learning using Deep Neural Networks



inton et al., 2006; Bengio et al., 2007; Bengio & LeCun, 2007; Lee et al., 2008; 2009

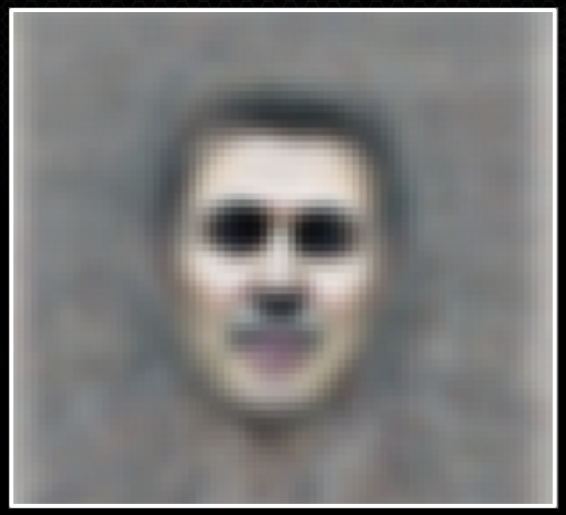
isual Object Recognition Using Deep Convolutional Neural Networks
ob Fergus (New York University / Facebook) http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2985

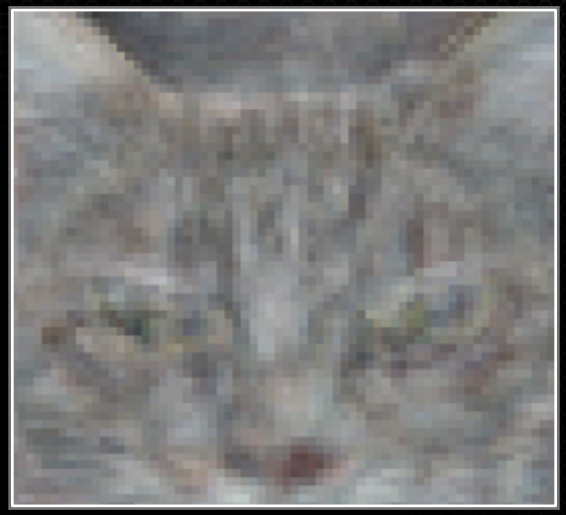
Google “Brain Project”

Building High-level Features Using Large Scale Unsupervised Learning

Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng

Stanford / Google





1 billion connections

10 million 200x200 pixel images

1,000 servers(16,000 cores)

3 days to train

Accelerating Machine Learning

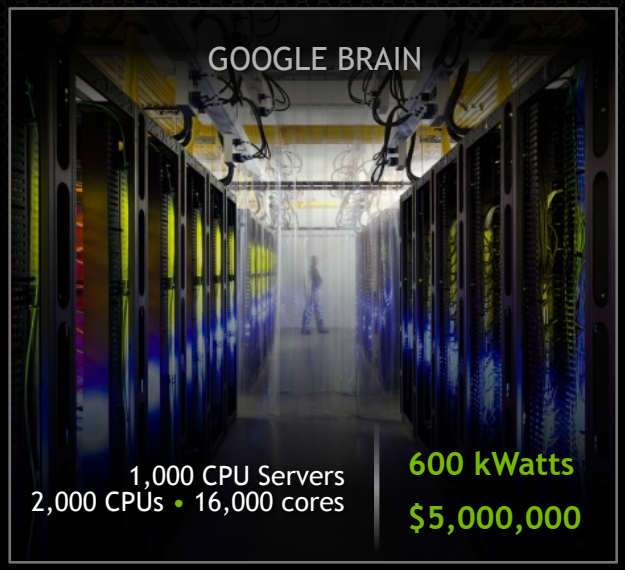
Deep Learning with COTS HPC Systems

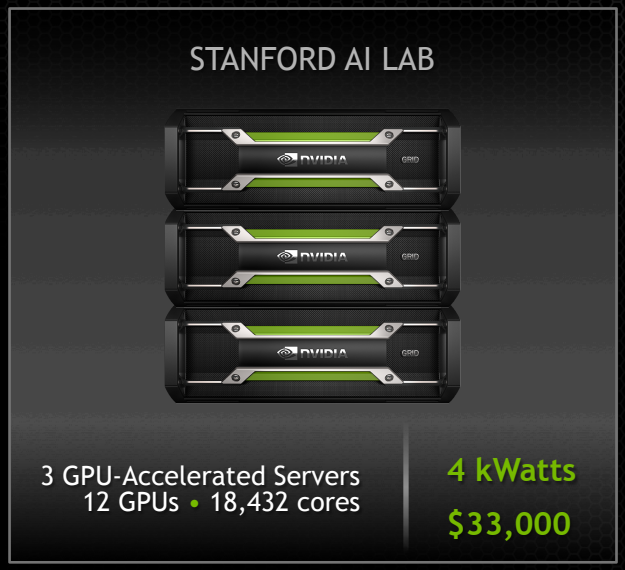
A. Coates, B. Huval, T. Wang, D. Wu,
A. Ng, B. Catanzaro

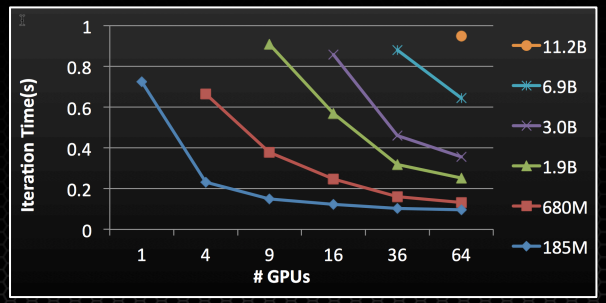
Stanford / NVIDIA • ICML 2013

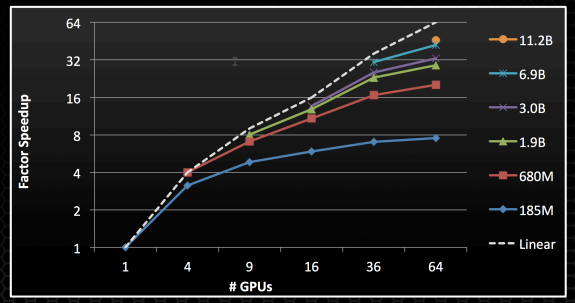
"Now You Can Build Google's $1M Artificial Brain on the Cheap"

-Wired









"10 Billion Parameter Neural Networks
In Your Basement", Adam Coates
http://on-demand.gputechconf.com/gtc/2014/video/S4694-10-billion-parameter-neural-networks.mp4

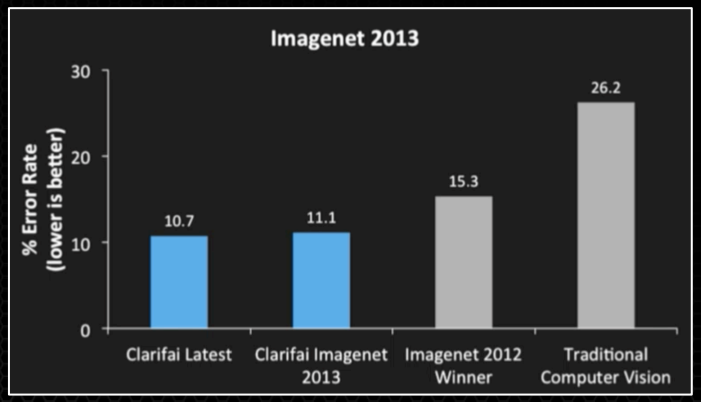
Accelerating Machine Learning

Image Recognition CHALLENGE

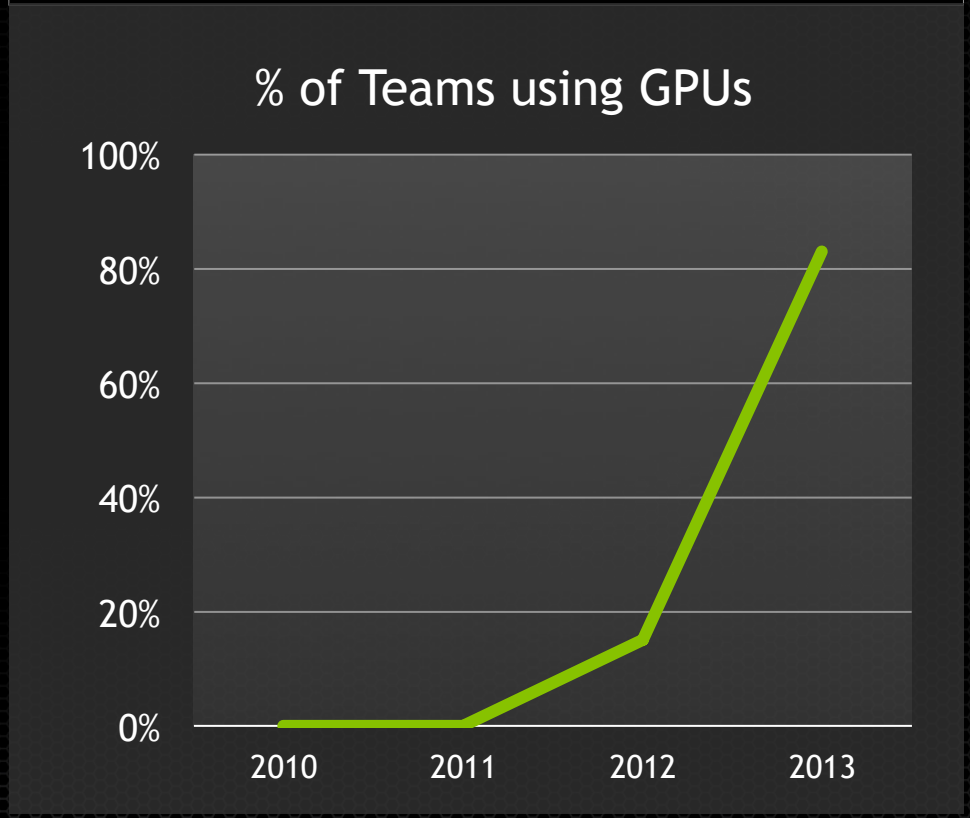
.2M training images • 1000 object categories

Hosted by

IMAGENET



GPU teams sweep all 3 categories at ILSVRC2013



Clarifai: Enabling Next Generation Intelligent Applications – http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2957

Machine Learning Comes of Age

Image Detection

Face Recognition

Gesture Recognition

Video Search & Analytics

Speech Recognition & Translation

Recommendation Engines

Indexing & Search

Talks at GTC

facebook

NYU

STANFORD UNIVERSITY

DARPA

DENSO

Carnegie Mellon University

Massachusetts Institute of Technology

Berkeley UNIVERSITY OF CALIFORNIA

Web & Enterprise Companies Use GPUs to Accelerate Machine Learning & Data Analytics

Adobe

Auto Tagging in Creative Cloud

Baidu

Speech/Image Recognition

flickr

Image Auto Tagging

IBM

Hadoop-based Clustering

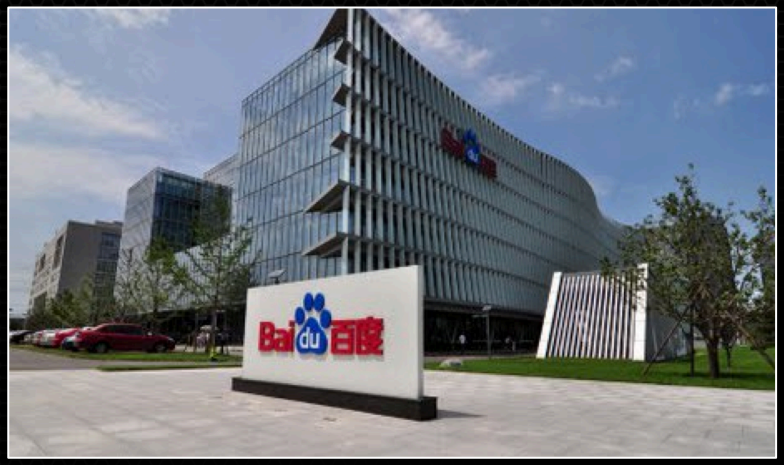
NETFLIX

Recommendation Engine

orange™

Database Queries

Baidu



Storage	• >2000PB
Processing	• 10-100PB/day
Webpages	• 100b-1000b
Index	• 100b-1000b
Update	• 1b-10b/day
Log	• 100TB~1PB/day

"Google is the Baidu of the USA"

Dr. Ren Wu
Distinguished Scientist, IDL, Baidu

<u>Every Day:</u>
5B+ queries
500M+ users
100M+ mobile users
100M+ photos

Baidu Visual Search

Baidu Data Sets and Training

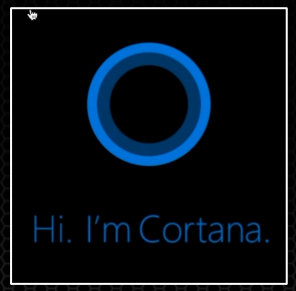
- Image recognition: 100 millions
- OCR: 100 millions
- Speech: 10 billions

- Training data projected to grow 10X per year

- Training time: Weeks to months on clusters of GPUs

New $1.6 B data center

How Does Machine Learning Touch Your Life?



"Okay, Google…"

…

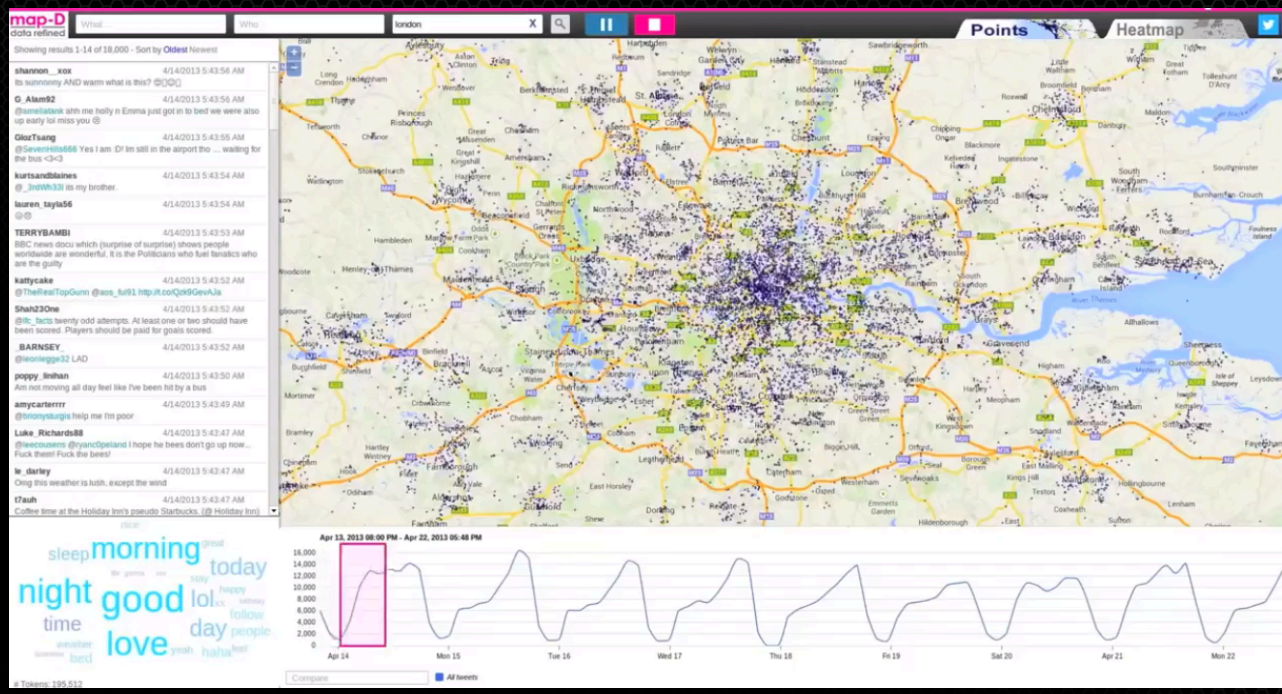
Map-D In-Memory Accelerated Database

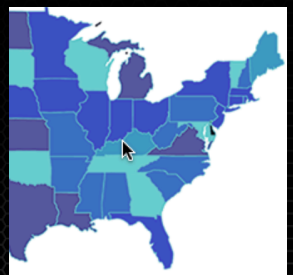
In-memory column store database

GIS and 3D animation on live streaming data

Billions of data points in milliseconds

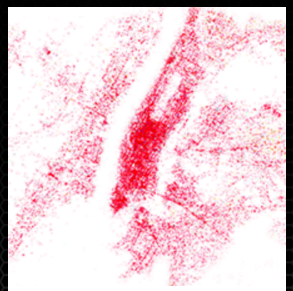
Clustered K40 GPUs





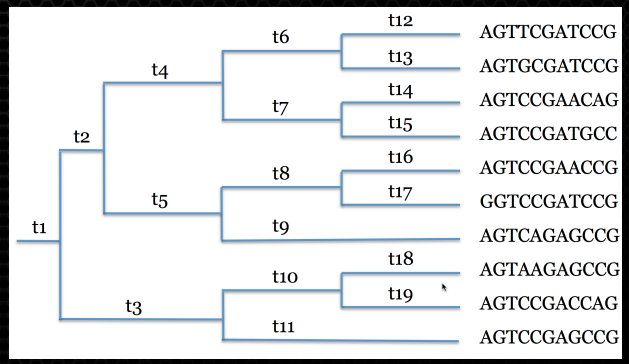
CampaignMap

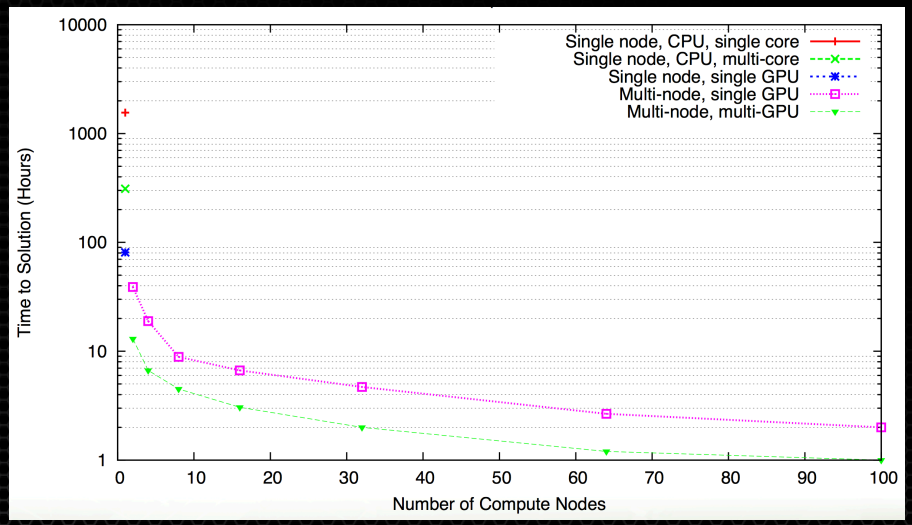
TweetMap



CityMap

Keeneland Graph Analytics





BEAST/BEAGLE Phylogenetics Software

81,000 lines of Java, 779 classes, and 81 packages

Scaled to run on 120 Keeneland nodes (360 GPUs)

Monte Carlo Markov chain phylogenetics

Probabilistic approach to dealing with factorial scalin
of number of possible topologies

Large-Scale Dense Sub-Graph Detection

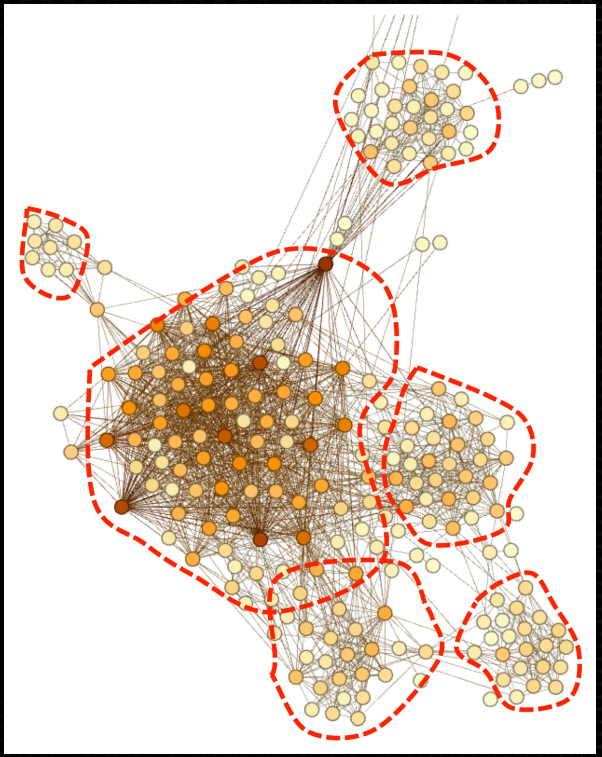
Definition:

- Detect subsets of vertices, such that the connections within the induced su-graphs are dense, and their connections to the rest of the graph are sparse.
- Unsupervised learning
- Mahout, GraphLab

Applications

- Community detection
- Recommender system
- Graph visualization
- Data exploration

GPU speedup
44.86x
373.71x



GPU-Accelerated Large-Scale Dense Subgraph Detection
Andy Wu (Xerox Research Center) http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2672

GTC On-Demand Playback

http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2957

- Astronomy and Astrophysics: 11
- Automotive: 10
- Big Data Analytics & Data Algorithms: 26
- Bioinformatics & Genomics: 12
- Climate, Weather, Ocean Modeling: 6
- Clusters and GPU Management: 9
- Collaborative & Large Resolution Displays: 3
- Combined Simulation & Real-Time Visualization: 3
- Computational Fluid Dynamics: 11
- Computational Physics: 24
- Computational Structural Mechanics: 1
- Computer Aided Design: 2
- Computer Vision: 11
- Debugging Tools & Techniques: 7
- Defense: 12
- Desktop and Application Virtualization: 1
- Digital Manufacturing: 22
- Education: 3
- Energy Exploration: 11
- Finance: 14
- Game Development: 4
- Graphics Virtualization: 21
- Large Scale Data Visualization & In-Situ Graphics: 2
- Machine Learning and AI: 10
- Media and Entertainment: 16
- Media & Entertainment Summit: 16
- Medical Imaging & Visualization: 10
- Mobile Applications: 1
- Mobile Summit: 27
- Molecular Dynamics: 13
- Numerical Algorithms & Libraries: 29
- Performance Optimization: 13
- Programming Languages & Compilers: 35
- Quantum Chemistry: 9
- Ray Tracing: 4
- Real-Time Graphics Applications: 7
- Rendering & Animation: 7
- Scientific Visualization: 3
- Signal & Audio Processing: 6
- Supercomputing: 23
- Video & Image Processing: 14
- Virtual & Augmented Reality: 2
- Visual Effects & Simulation: 3


Questions?
NVIDIA®

Data Analytics, Machine Learning, and GPUs

Amazing new applications and services employing machine learning algorithms to perform advanced analysis of massive streams and collections of structured and unstructured data are becoming quietly indispensable in our daily lives.

Machine learning algorithms like deep learning neural networks are not new, but the rise of large scale applications hosted in massive cloud computing data centers collecting enormous volumes of data from and about their users have provided unprecedented training sets and opportunities for machine learning algorithms.

Recognizers, classifiers, and recommenders are only a few component capabilities providing valuable new services to users, but the training of extreme scale learning systems is computationally intense. Fortunately, like so many areas of high-performance computing, great economies and speed-ups can be realized through the use of general purpose GPU accelerators.

This talk will explore a few advanced data analytics and machine learning applications, and the benefits and value of GPU acceleration.