SoftServe | Empowering your Business through Software Development

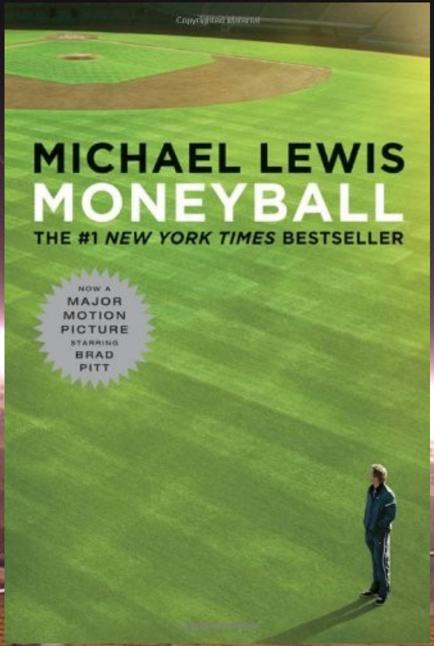
Data Science

the art of foul play

Sergey Shelpuk
SoftServe, Inc.
sshel@softserveinc.com
September, 2013

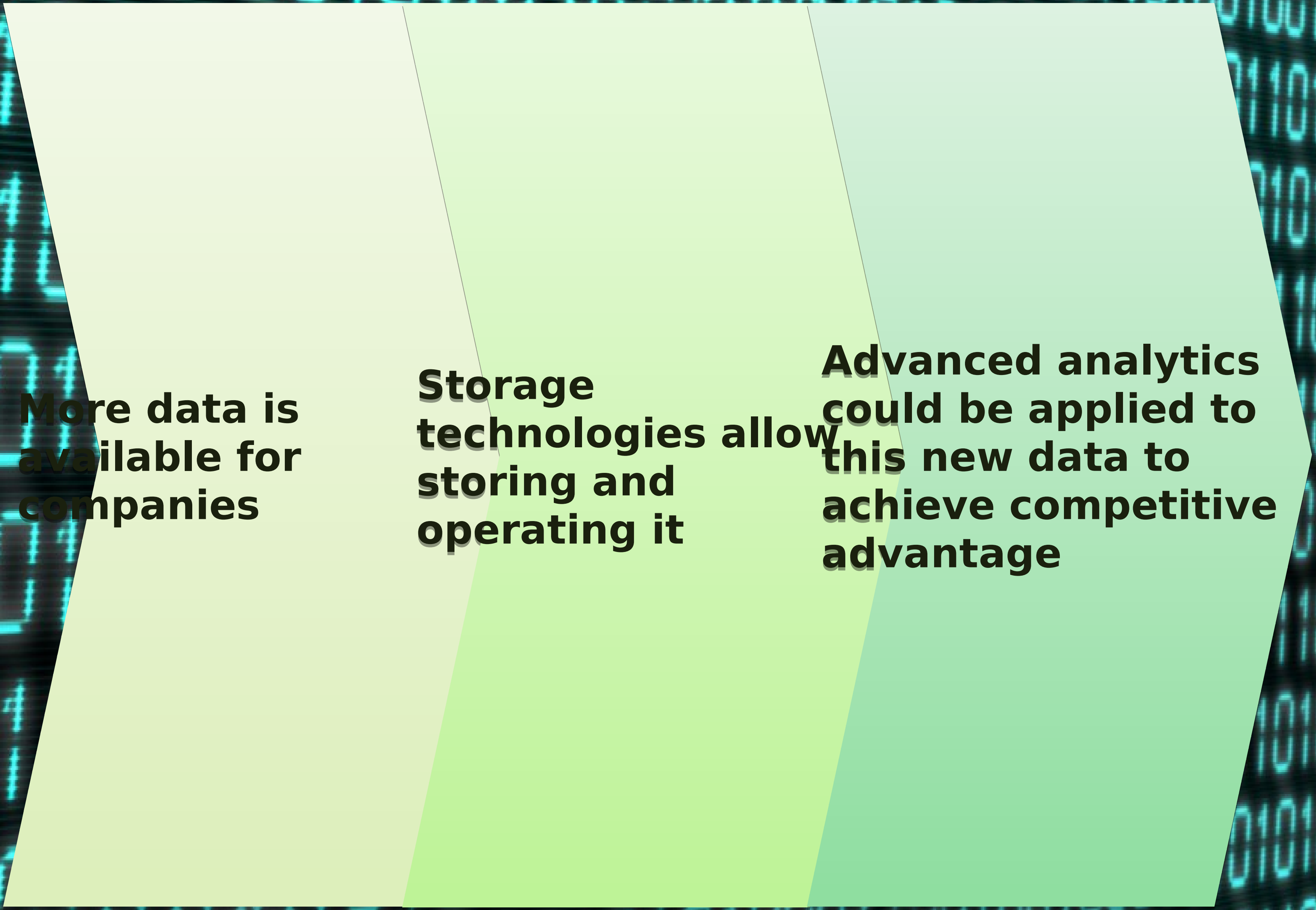
TARGET®

Home
Price Scanner and Service
C 22
Air Compressor
Power Tools
C 20
B 28
Snacks
Nuts
89 99


MICHAEL LEWIS
MONEYBALL
THE #1 NEW YORK TIMES BESTSELLER
NOW A MAJOR MOTION PICTURE STARRING BRAD PITT

“Your goal should not be to buy players, it should be to buy wins. In order to buy wins you should buy runs” (c)

Google
PageRank



More data is available for companies

Storage technologies allow storing and operating it

Advanced analytics could be applied to this new data to achieve competitive advantage

Harvard Business Review

Data Scientist: The Sexiest Job of the 21st Century

The New York Times

For Today’s Graduate, Just One Word: Statistic

Hal Varian

chief economist at Google

“I keep saying that the sexy job in th next 10 years will be statisticians, and I’m not kidding.”

Aol.

Data Scientist: The Hottest Jo You Haven't Heard Of

Gartner®

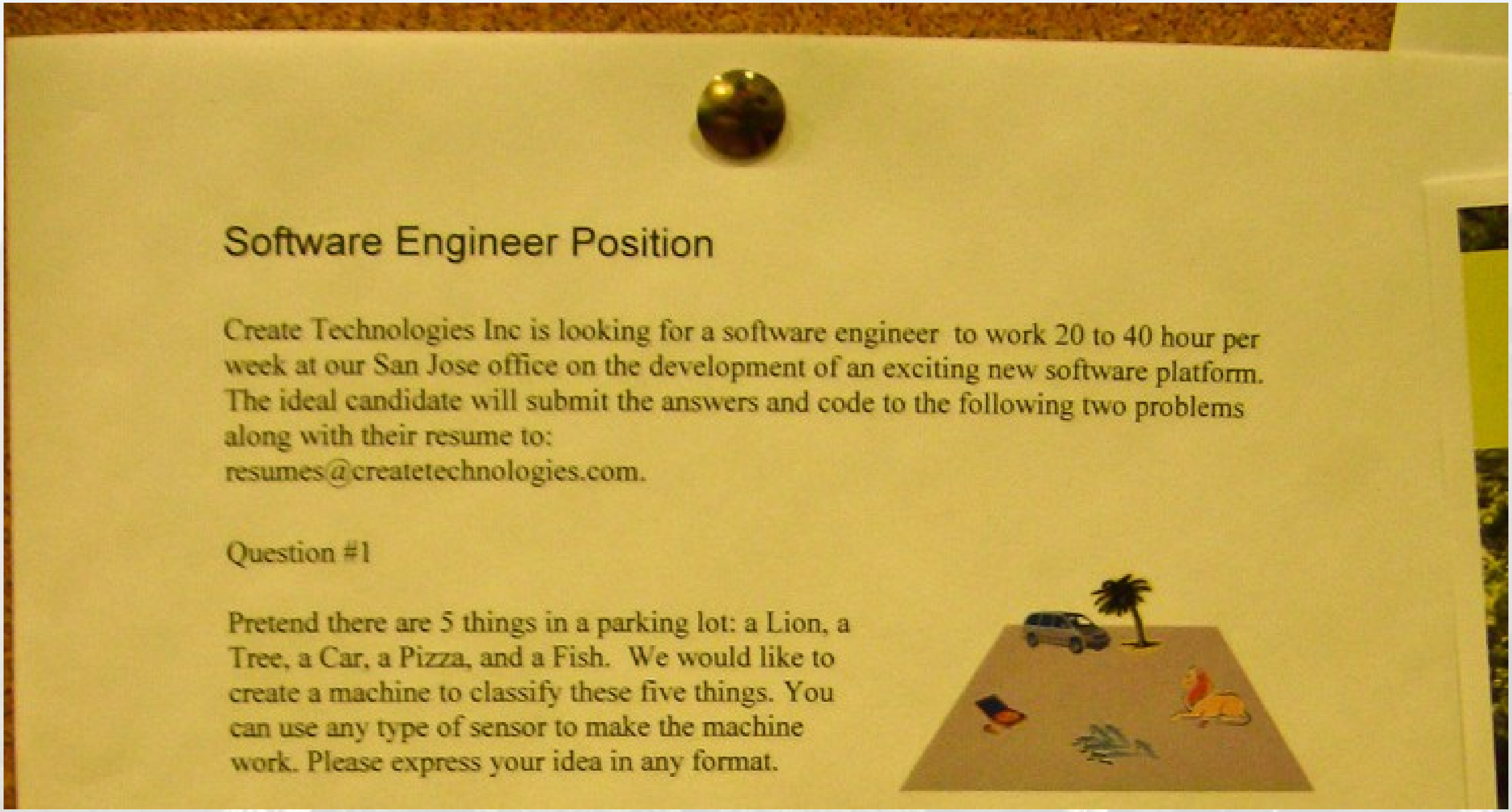
Top 10 Strategic Technology Trends for 2013

- Mobile Device Battles
- Mobile Applications and HTML5
- Personal Cloud
- Enterprise App Stores
- The Internet of

- Hybrid IT and Cloud Computing
- Strategic Big Data
- Actionable Analytics
- In Memory Computing
- Integrated Ecosystems

McKinsey&Company

McKinsey Global Institute projects approximately 140,000 to 190,000 unfilled positions of data analytics experts in the U.S. by 2018 and a shortage of 1.5 million managers and analysts who have the ability to understand and make decisions using big data.



Software Engineer Position

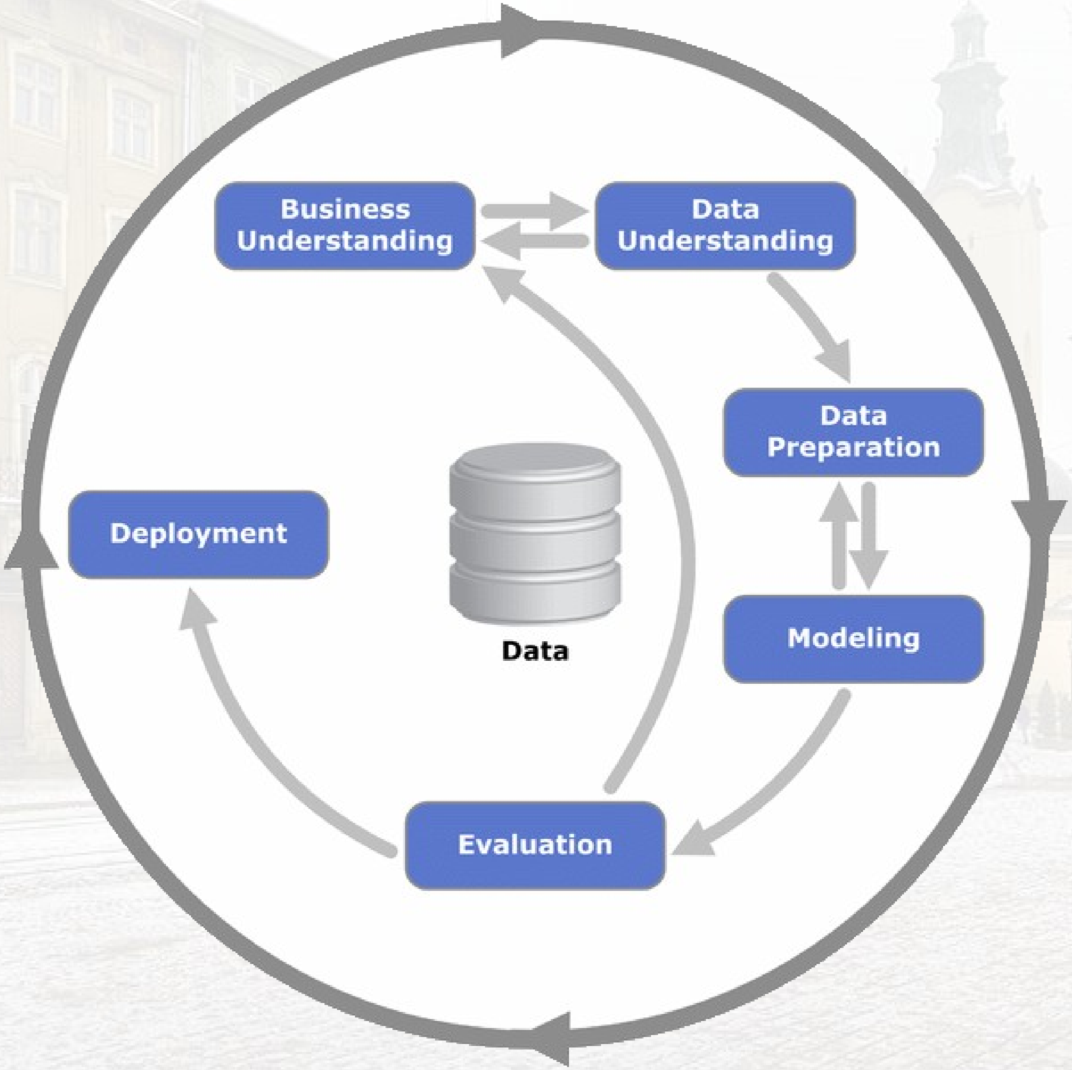
Create Technologies Inc is looking for a software engineer to work 20 to 40 hour per week at our San Jose office on the development of an exciting new software platform. The ideal candidate will submit the answers and code to the following two problems along with their resume to:
resumes@createtechnologies.com.

Question #1

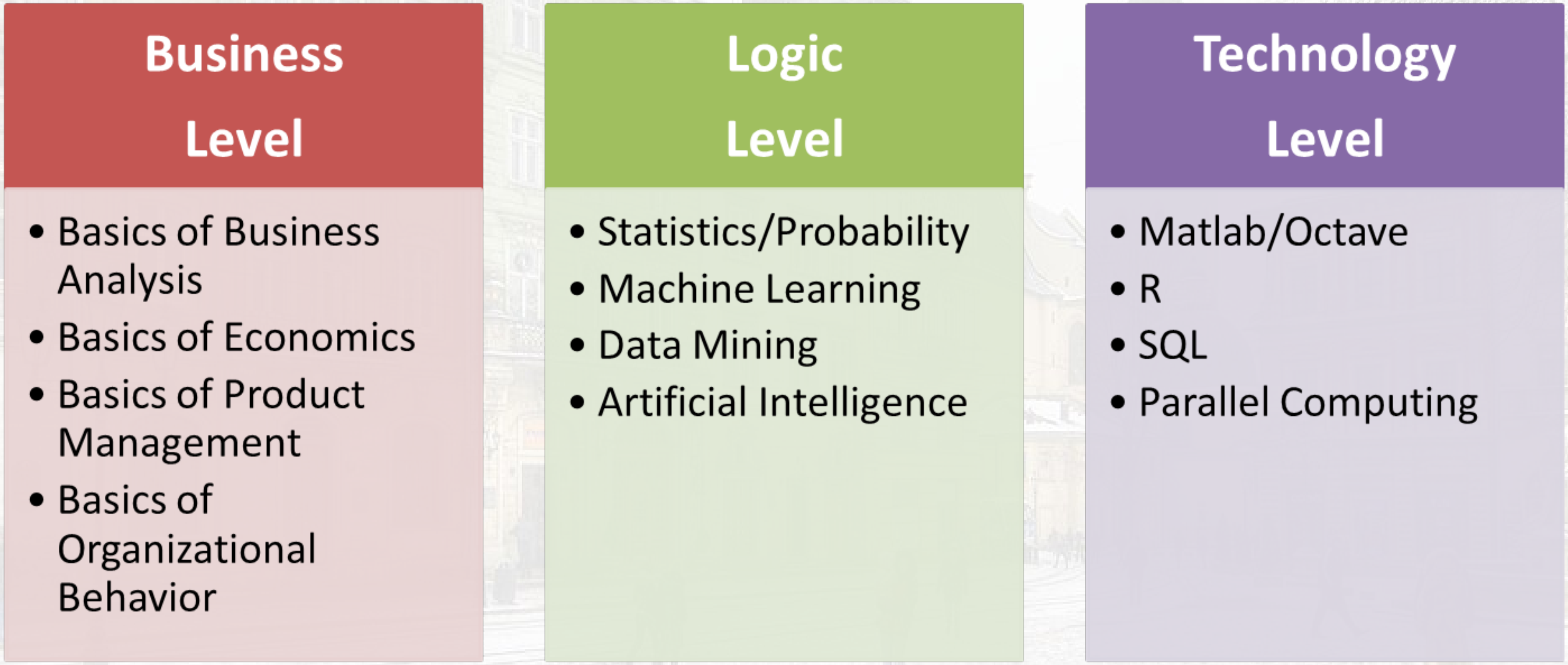
Pretend there are 5 things in a parking lot: a Lion, a Tree, a Car, a Pizza, and a Fish. We would like to create a machine to classify these five things. You can use any type of sensor to make the machine work. Please express your idea in any format.

Business Tasks	• Define prospective customers • Define traffic jams in the city • Recommend restaurants and menus • Adjust UI to the particular user • Classify body part on X-Ray image	• Define market niche • Define influencers in the social networks • Define similar customers or projects in portfolio • Define informal groups in the organization	• Define fraud bank transaction • Define network intrusion attempts • Provide automatic aircraft engine testing • Provide automatic IT infrastructure monitoring • Provide clinical test analysis	• Define the best price for the goods or services to maximize profits • Define best working schedule for the store • Define best amount of production • Define best business rules
Model Family	**Classification**	**Clustering**	**Anomaly Detection**	**Optimization**
Algorithms	• Naïve Bayes • Logistic regression • Support Vector Machines • Neural Networks	• K-Means • K nearest neighbor • Self-organized maps • Mixture of Gaussians	• Mixture of Gaussians • Self-learning anomaly detection	• Gradient descent • Simplex method • Newton’s method • Normal equations • Genetic algorithms

Cross Industry Standard Process for Data Mining



SoftServe Data Science Group Knowledge Model



Business Level

- Basics of Business Analysis
- Basics of Economics
- Basics of Product Management
- Basics of Organizational Behavior

Logic Level

- Statistics/Probability
- Machine Learning
- Data Mining
- Artificial Intelligence

Technology Level

- Matlab/Octave
- R
- SQL
- Parallel Computing

Open Source Data Science Tools

R project

Open Source Data Science Tools

Python stack

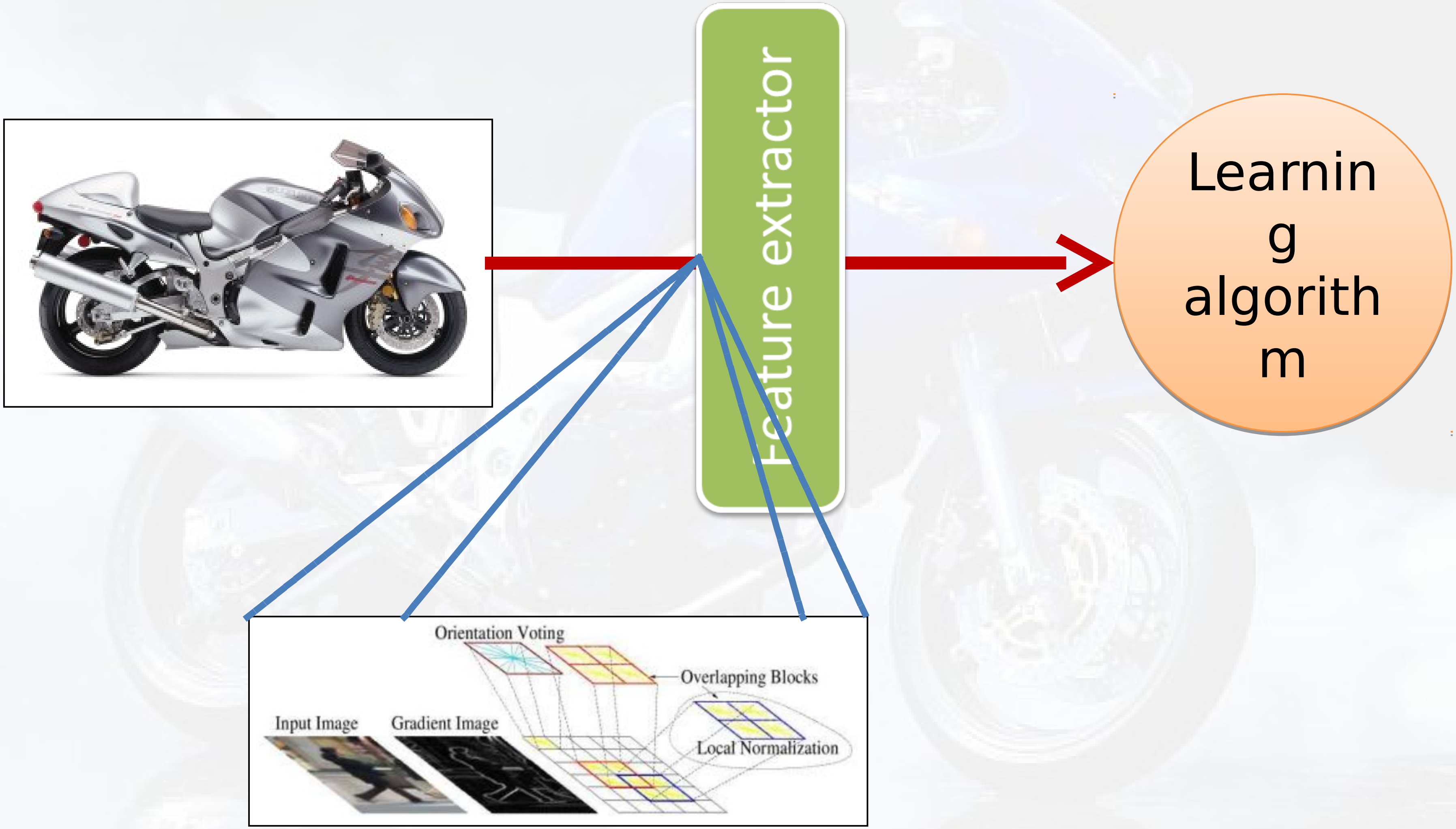
Neural networks

Linear algebra

Scientific computi g

Deep Learning Neural Networks

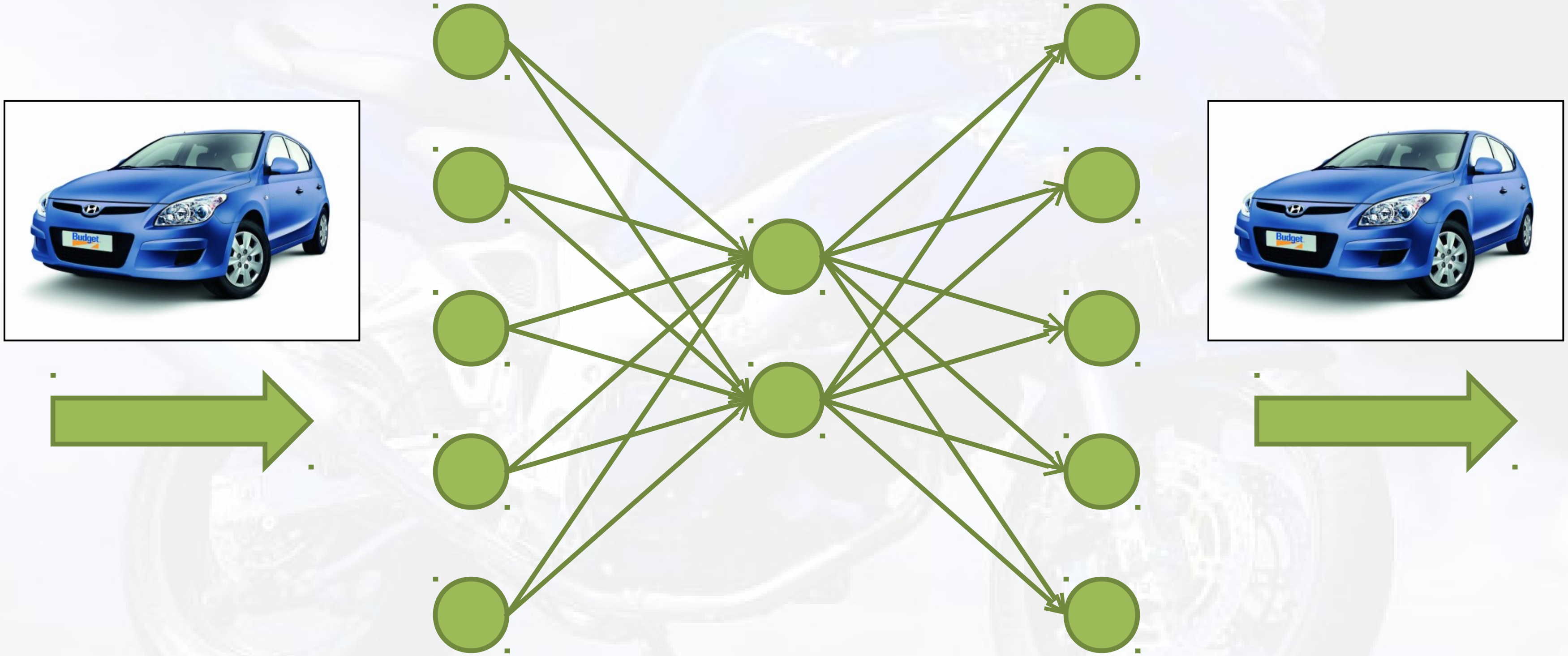
Task: recognize a motorcycle



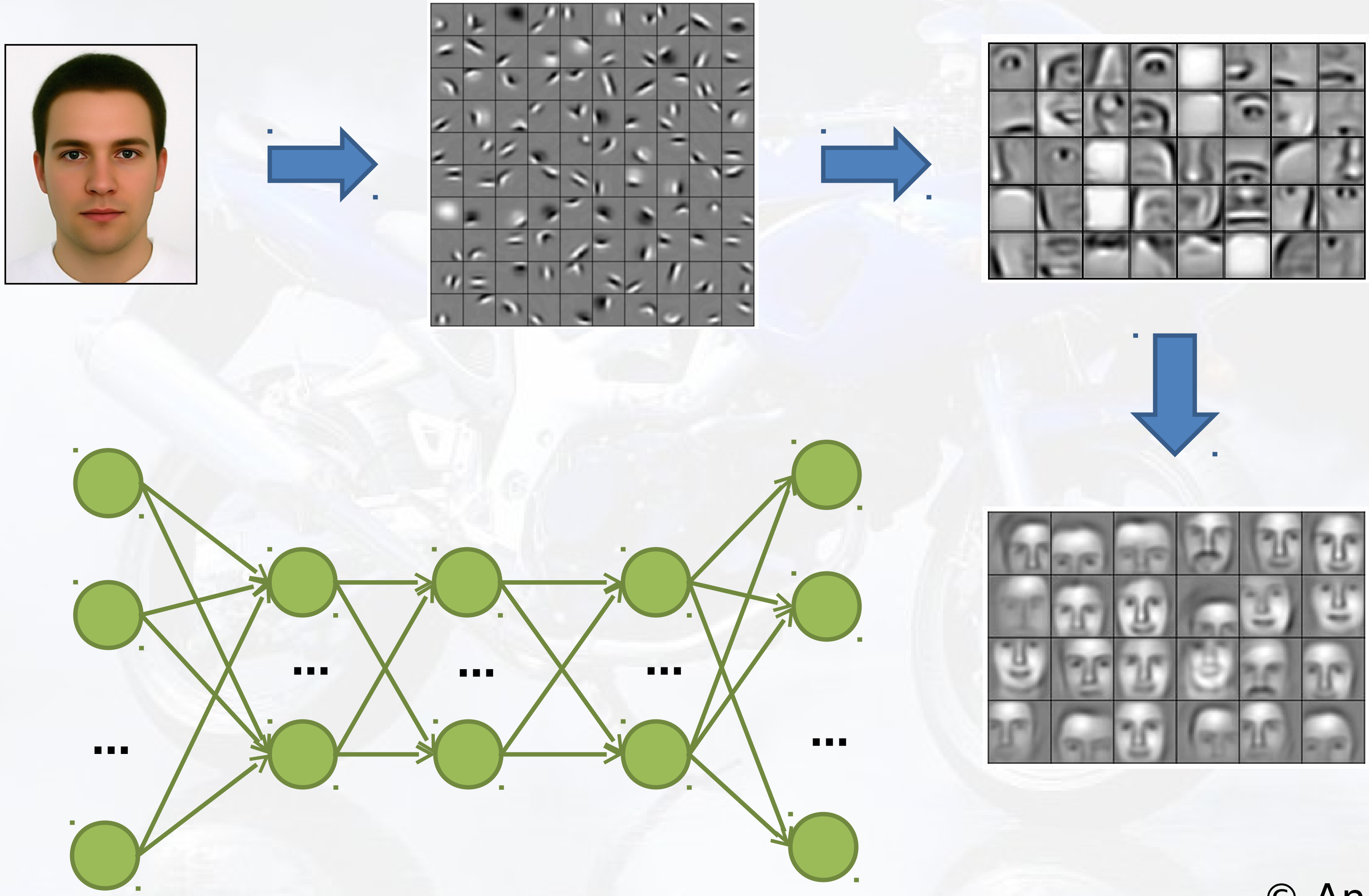
Feature extractor

Learnin
g
algorith
m

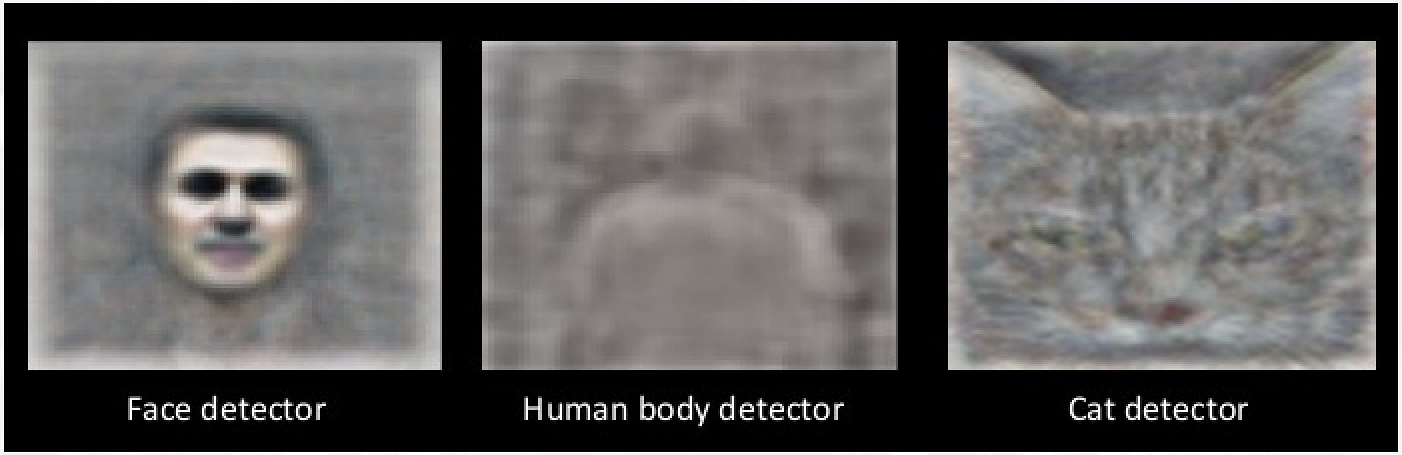
The concept of Autoencoder



The concept of Autoencoder



Large scale deep learning networks



Face detector

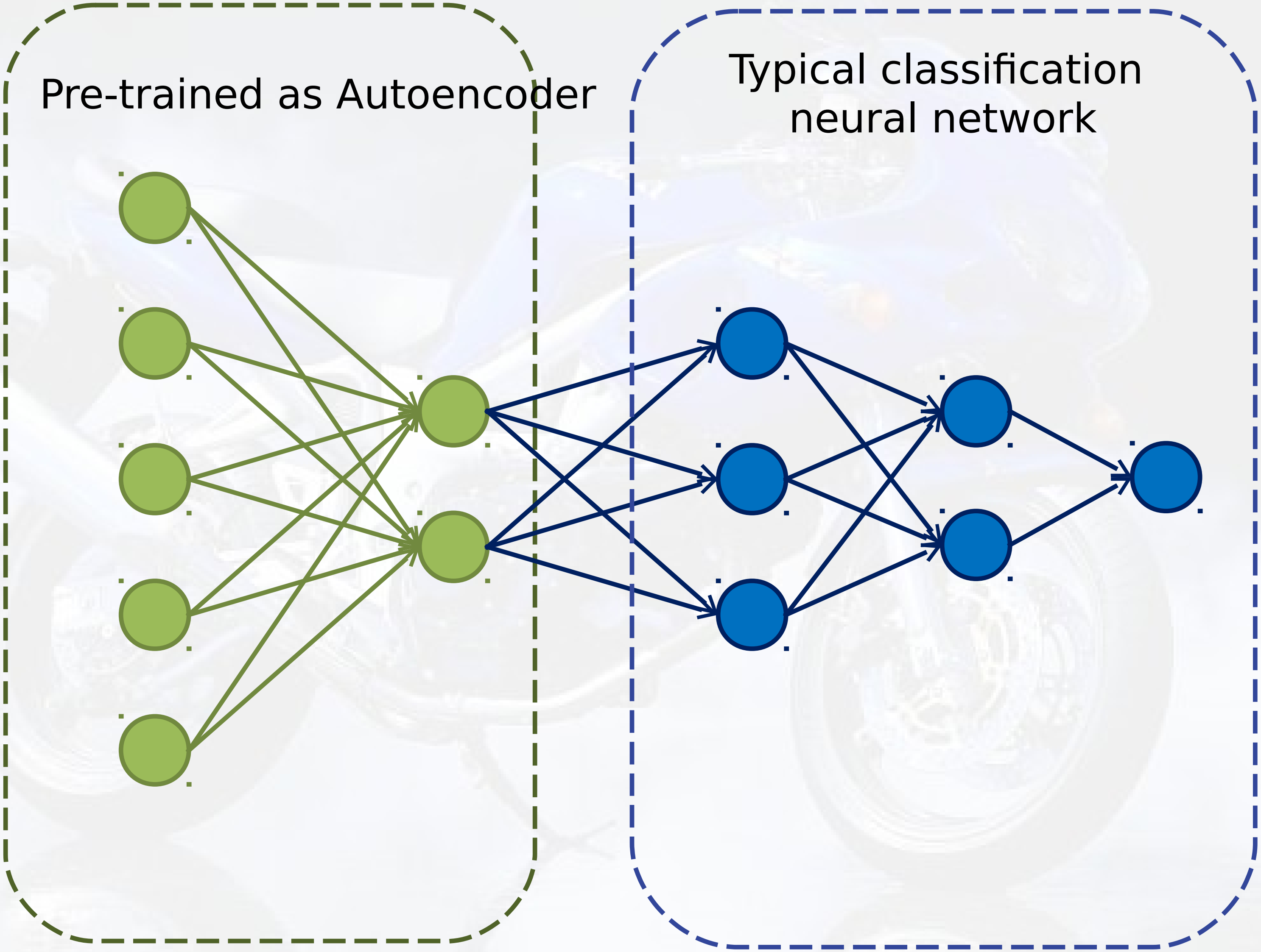
Human body detector

Cat detector

See more: Building high-level features using large scale unsupervised learni

Hire the smartest people in the world
Invent cat detector

Deep learning neural networks



Pre-trained as Autoencoder

Typical classification neural network

Few results

Images

CIFAR Object classification	Accuracy
Prior art (Ciresan et al., 2011)	80.5%
Stanford Feature learning	**82.0%**

NORB Object classification	Accuracy
Prior art (Scherer et al., 2010)	94.4%
Stanford Feature learning	**95.0%**

Video

Hollywood2 Classification	Accuracy
Prior art (Laptev et al., 2004)	48%
Stanford Feature learning	**53%**

KTH	Accuracy
Prior art (Wang et al., 2010)	92.1%
Stanford Feature learning	**93.9%**

YouTube	Accuracy
Prior art (Liu et al., 2009)	71.2%
Stanford Feature learning	**75.8%**

UCF	Accuracy
Prior art (Wang et al., 2010)	85.6%
Stanford Feature learning	**86.5%**

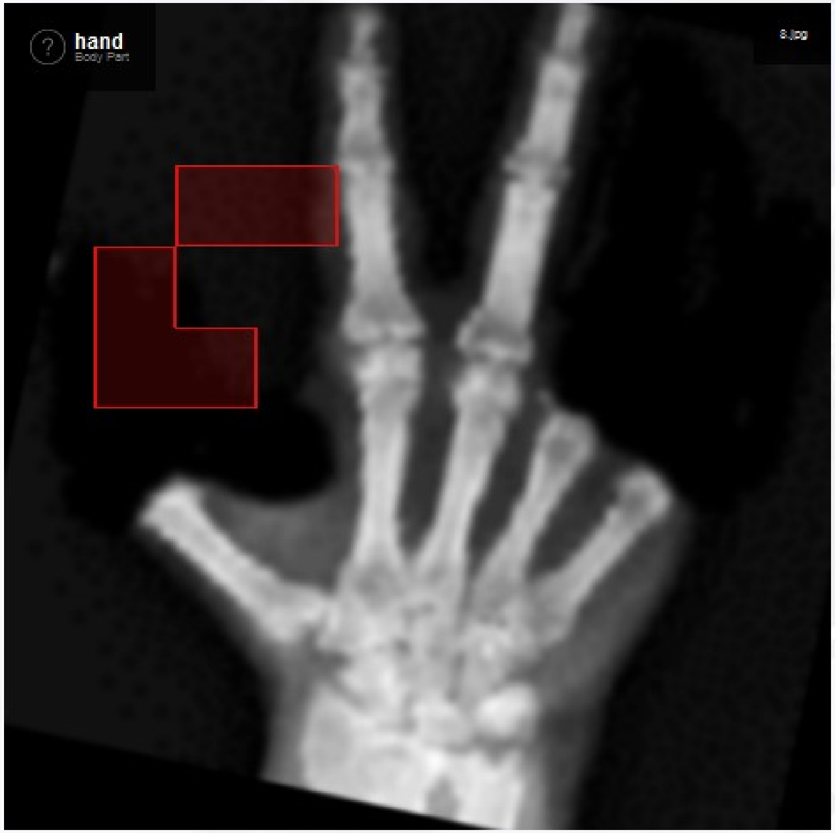
Text/NLP

Paraphrase detection	Accuracy
Prior art (Das & Smith, 2009)	76.1%
Stanford Feature learning	**76.4%**

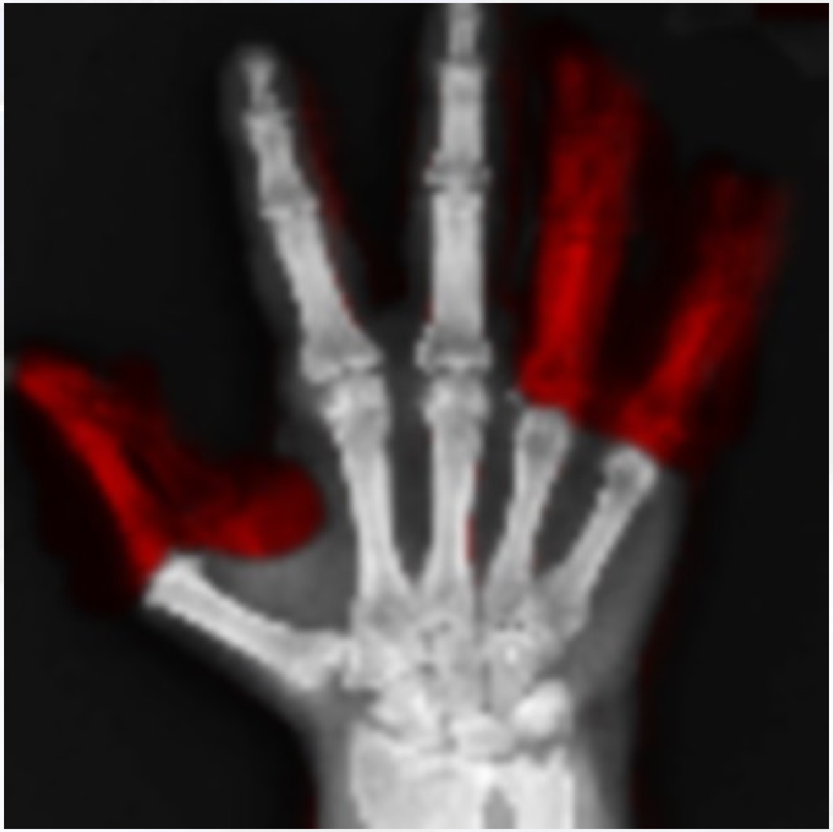
Sentiment (MR/MPQA data)	Accuracy
Prior art (Nakagawa et al., 2010)	77.3%
Stanford Feature learning	**77.7%**

Deep Learning in SoftServe

Phase 1 results
(old-fashion anomaly detection)



Phase 2 prototype
(deep learning approach)



Useful Resources

UDACITY

- Introduction to Statistics
- Introduction to Artificial Inte

coursera

- Machine Learning
- Probabilistic Graphical Mode
- Statistics One

Thank you!

?