3 Mini Talks of 10 minutes each

The Machine Intelligence Landscape: A Venture Capital Perspective
David Beyer

The future of global, trustless transactions on the largest graph: blockchain
Olaf Carlson-Wee

Algorithms for Anti-Money Laundering
Richard Minerich

QCon

The Machine Intelligence Landscape: A Venture Capital Perspective

AMPLIFY PARTNERS

An Explosion of Data

2002

5 Exabytes

Growing by a factor of 7,040x

2020

35,200 Exabytes

Source: EMC, Forbes estimates

What is Enabling This Transition?

- Open Source: It's now cheap enough to store ALL the data vs. being selective
- AWS: infrastructure/scale for the masses
- Big Data laying groundwork for Machine Intelligence—the "so what?"

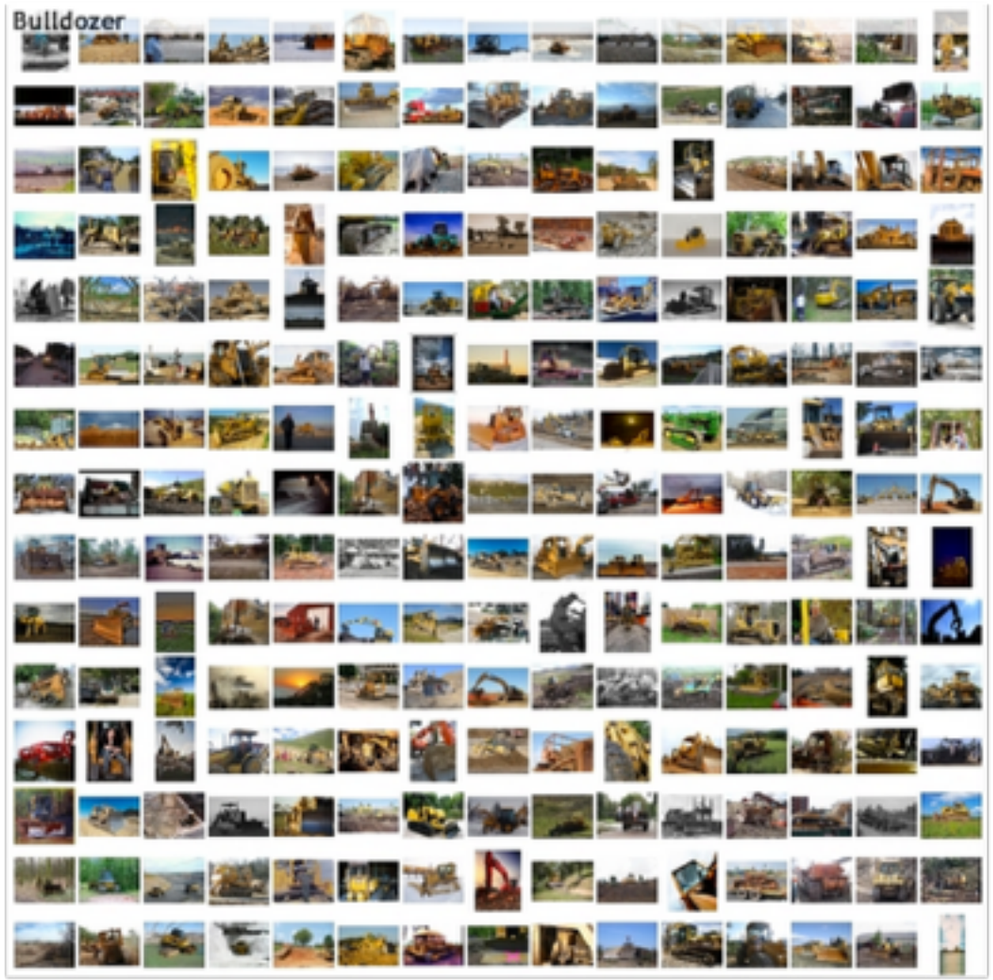
What is Machine Learning?

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

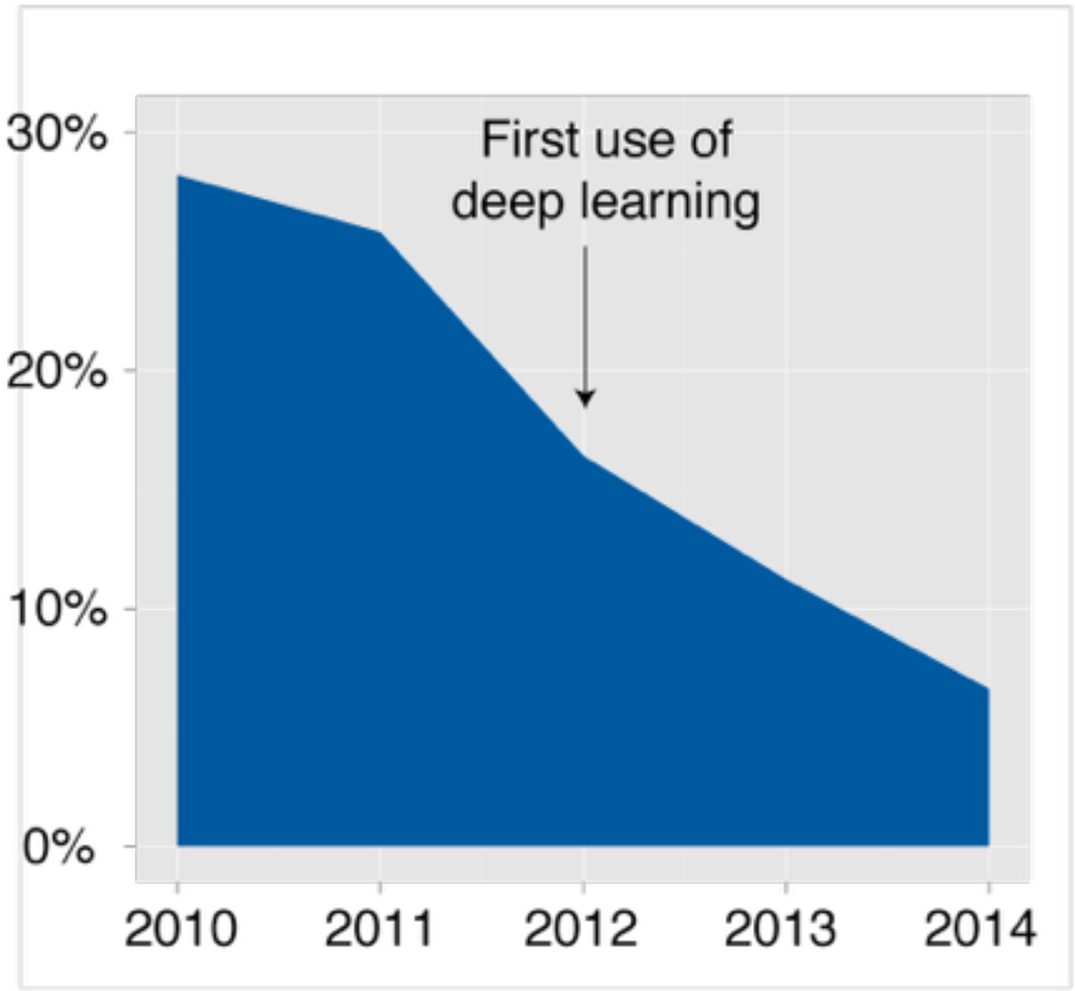
- Tom Mitchell

Deep Learning Becomes Very, Very Good

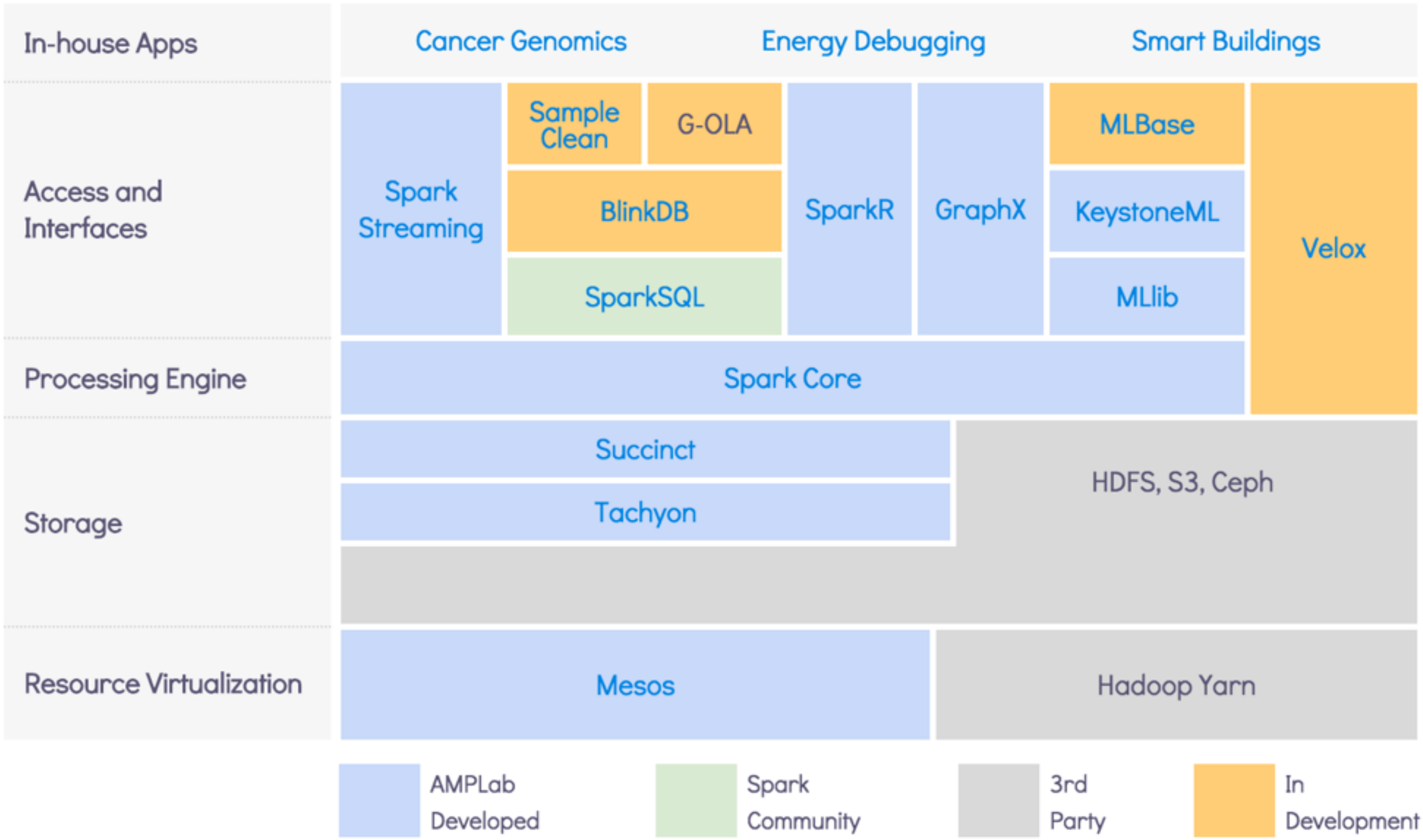
ImageNet examples



Objection classification error rate

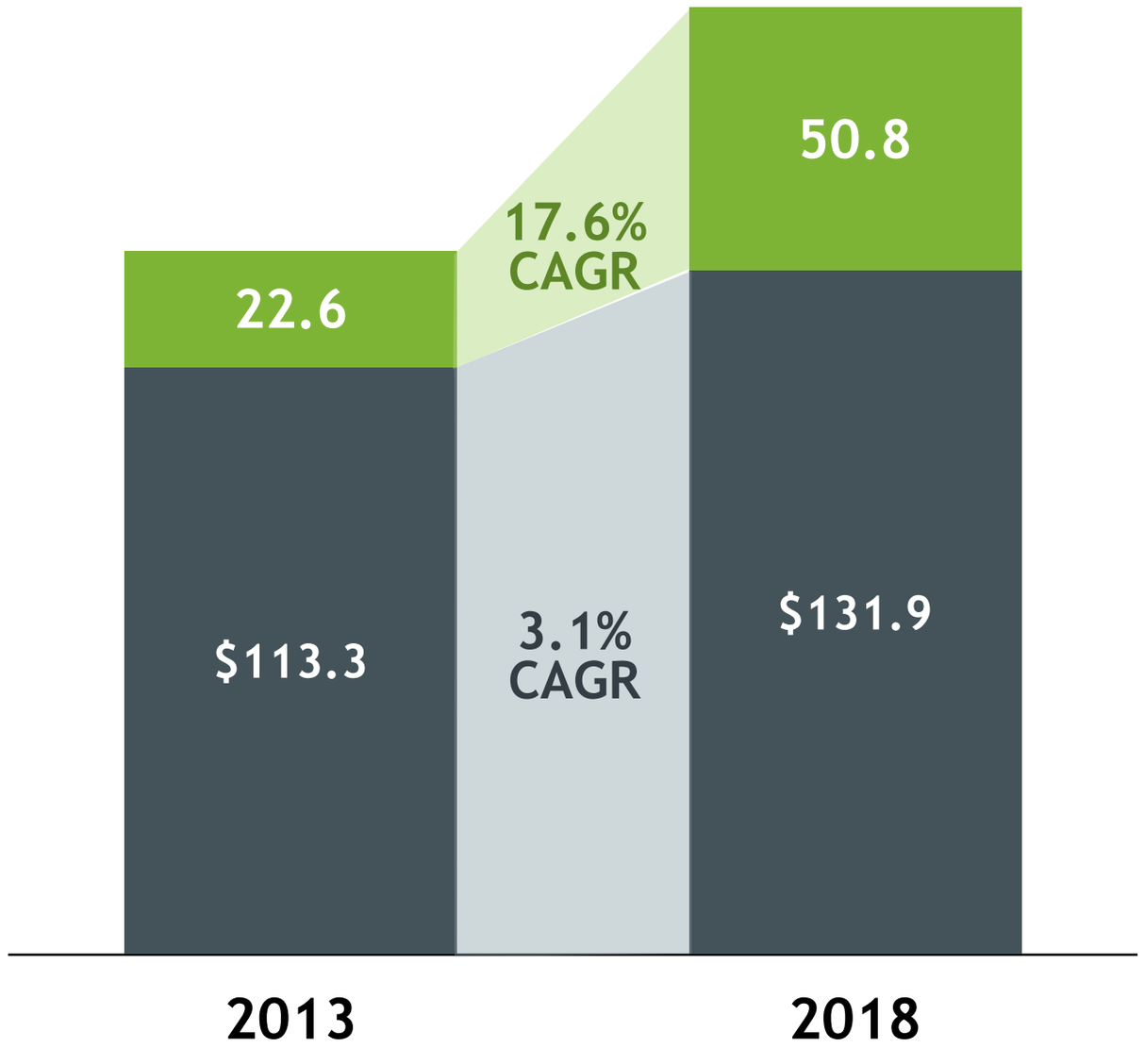


Recent Innovations in Distributed Computing



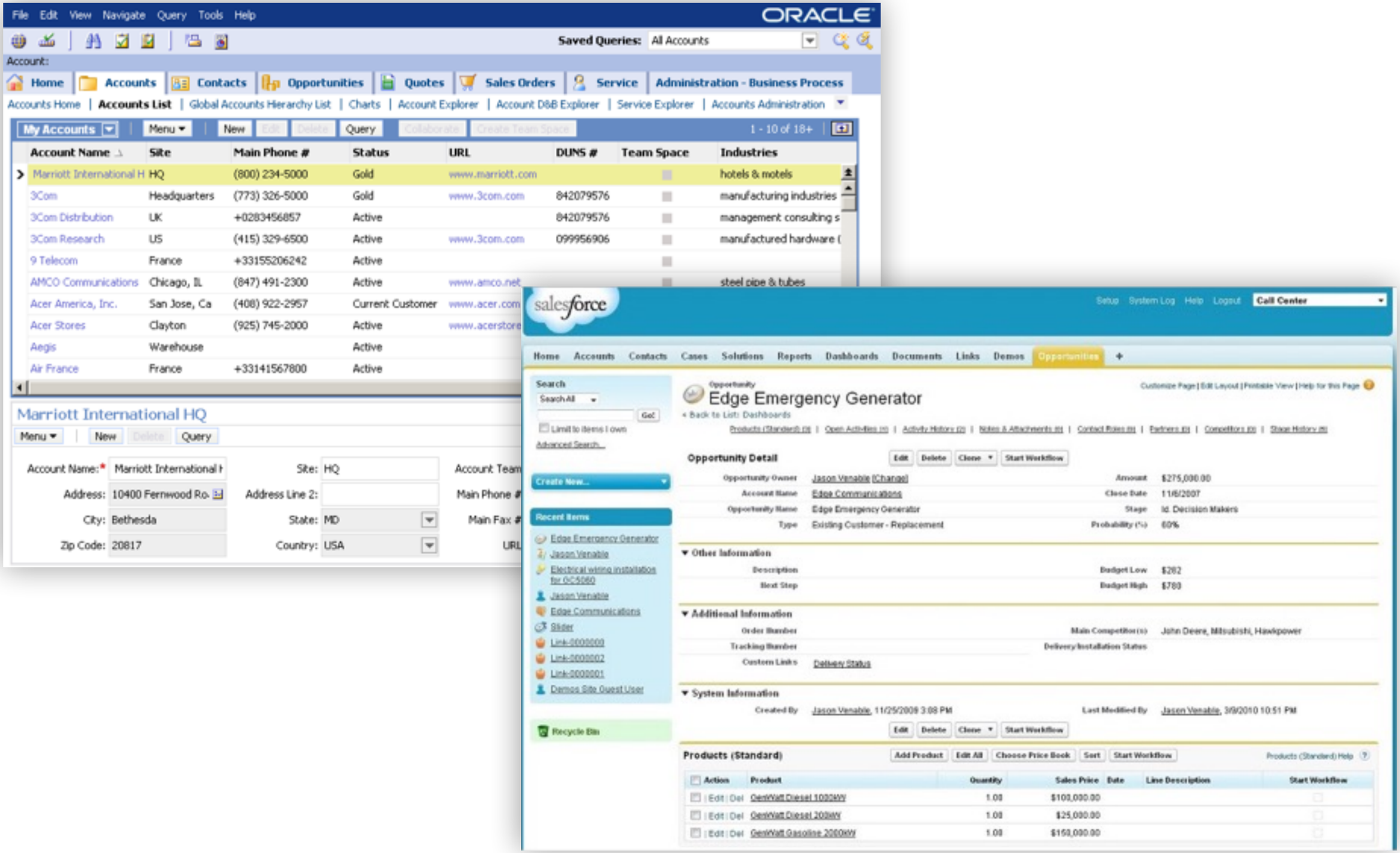
SaaS is Crushing Traditional Enterprise SW

World Wide Enterprise Applications Market* ($B)

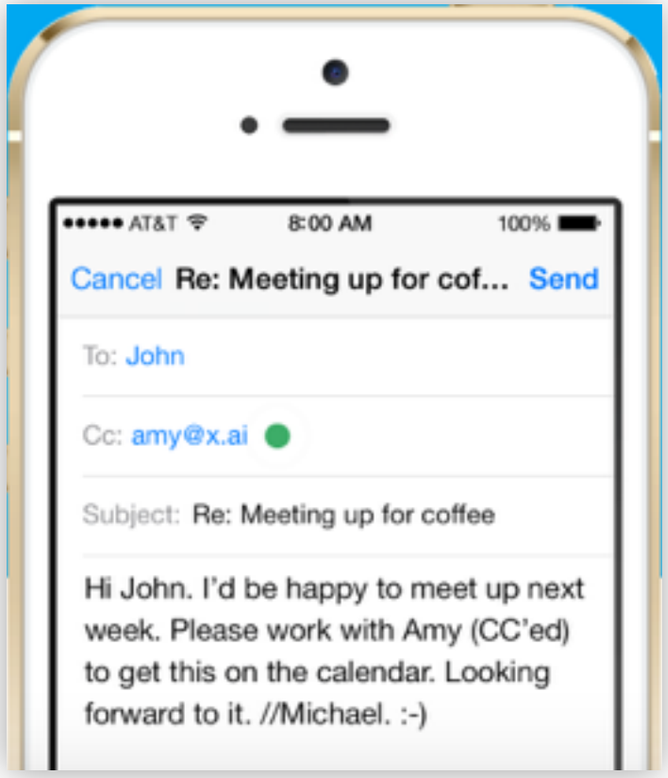


*according to IDC estimates

Not Much has Changed In 15 Years



Samantha (Her) vs. Amy (x.ai)



Machine Intelligence LANDSCAPE

CORE TECHNOLOGIES

ARTIFICIAL INTELLIGENCE
IBM WATSON, MetaMind, Numenta, ai-one, Cycorp, Research, nara, Reactor, SCALED INFERENCE

DEEP LEARNING
vicarious, Vision Factory, facebook, Google, SKYMIND, Baidu IDL, ersatz, SignalSense

MACHINE LEARNING
rapidminer, context, 0xdata H2O, DATAROBOT, LIFTIGNITER, SPARKBEYOND, Azure ML, yhat, Tableau, Sense, GraphLab, Alpine, Ayasdi

NLP PLATFORMS
cortical.io, idibon, LUMINOSO, wit.ai, Maluuba

PREDICTIVE APIS
AlchemyAPI, MINDOPS, Google, bigml, indico, ALGORITHMIA, Expect Labs, PredictionIO

IMAGE RECOGNITION
clarifai, MADBITS, DNNresearch, DEXTRO, VISENZE, lookflow

SPEECH RECOGNITION
GRIDSPACE, popup archive, NUANCE

RETHINKING ENTERPRISE

SALES
Preact, AVISO, RelateIQ, NGDATA, CLARABRIDGE, FRAMED, infer, ATTENSITY, causata

SECURITY / AUTHENTICATION
CROSSMATCH, conjur, EVERYPIN, BITSIGHT, CYLANCE, bionym

FRAUD DETECTION
sift science, SOCURE, ThreatMetrix, feedzai, Brighterion, VERAFIN

HR / RECRUITING
TalentBin, entelo, predikt, Connectifier, gild, hiQ, CONVERSICA.COM

MARKETING
brightfunnel, bloomreach, CommandIQ, AIRPR, RADIUS, TellApart, people pattern, Persado

PERSONAL ASSISTANT
Siri, X.ai, Google now, Cortana, cleversense, tempo, Robinlabs, KASISTO, fuse machines, VIV, CLARA LABS

INTELLIGENCE TOOLS
ADATAD, Palantir, Quid, Digital Reasoning, FirstRain

RETHINKING INDUSTRIES

ADTECH
METAMARKETS, dstillery, rocketfuel, YieldMo, ADBRAIN

AGRICULTURE
BLUE RIVER, Terravion, ceresimaging, THE CLIMATE CORPORATION, tule

EDUCATION
declara, coursera, KNEWTON, kidaptive

FINANCE
Bloomberg, FinGenius, KENSHO, alphasense, Dataminr, minettabrook, BINATIX

LEGAL
Lex Machina, brightleaf, COUNSELYTICS, RAVEL, JUDICATA, Brevia, DiligenceEngine

MANUFACTURING
SIGHT MACHINE, MICROSCAN, IVISYS, BOULDER IMAGING

MEDICAL
Parzival, transcriptic, ZEPHYR, Genescient, grand round table, bina, TUTE

OIL AND GAS
kaggle, AYASDI, TACHYUS, biota, Flutura

MEDIA / CONTENT
Outbrain, newsle, ARRIA, SAILTHRU, wavii, Owlin, NarrativeScience, ayasdi, Summly, Prismatic, ai

CONSUMER FINANCE
affirm, inMenture, zestfinance, BILLGUARD, LendUp, LendingClub, Kabbage

PHILANTHROPIES
DataKind, thorn, THE DATA GUILD

AUTOMOTIVE
Google, Continental, TESLA, MOBILEYE, CRUISE

DIAGNOSTICS
enlitic, 3SCAN, lumiata, ENLITIC

RETAIL
BAY SENSORS, PRISM SKYLABS, celect, euclid

RETHINKING HUMANS / HCI

AUGMENTED REALITY
magic leap, Atheer, APX, blippar, META, layar

GESTURAL COMPUTING
THALMICLABS, omek, EM, FLUTTER, LEAP MOTION, eyeSight, 3Gear, GestureTek, nod

ROBOTICS
intel, LIQUID ROBOTICS, SoftBank, iRobot, Boston Dynamics, jibo, OMRON, evolution robotics

EMOTIONAL RECOGNITION
affectiva, BEYOND VERBAL, EMOTIENT, cogito

SUPPORTING TECHNOLOGIES

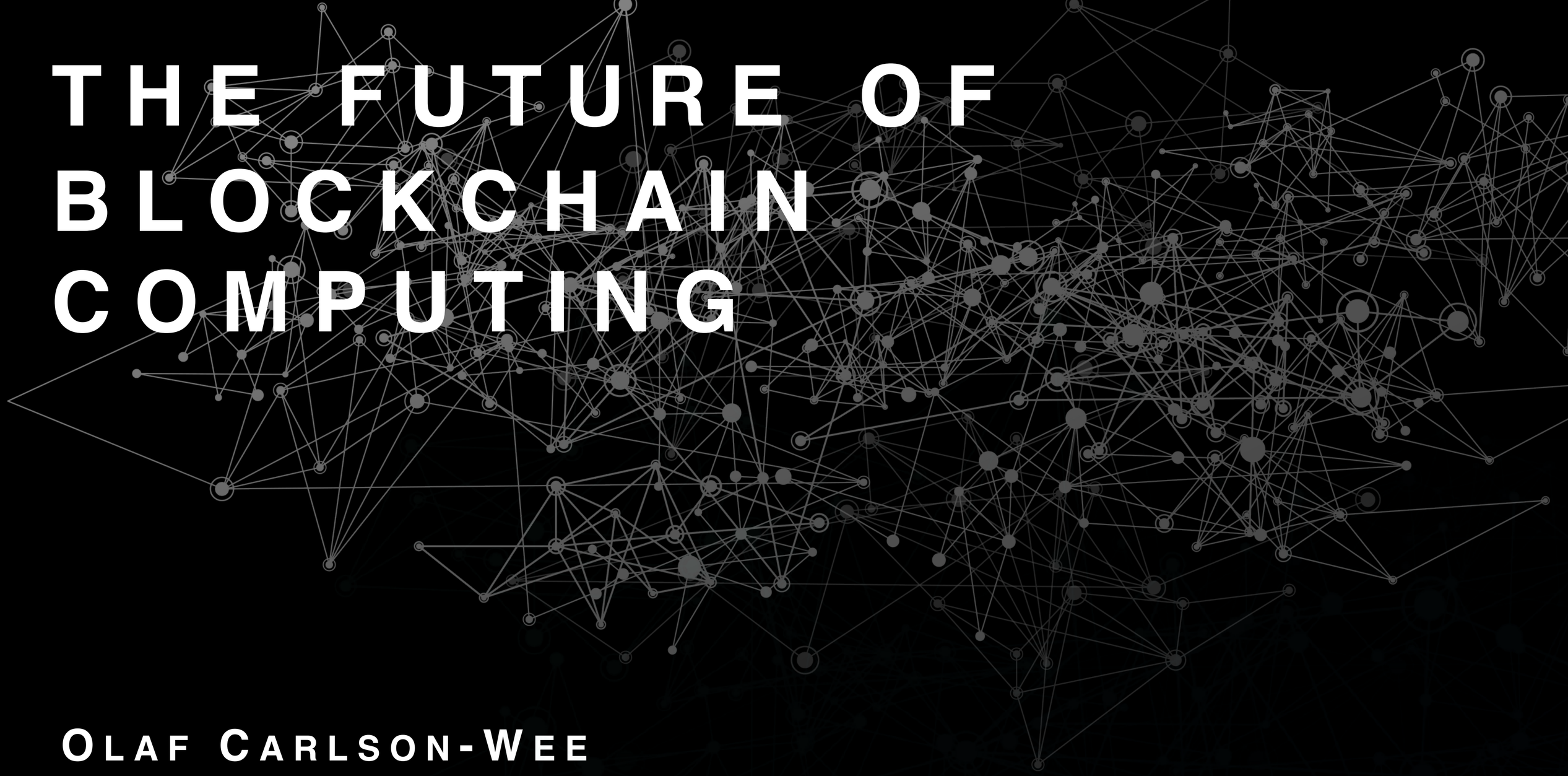
HARDWARE
NVIDIA, XILINX, QUALCOMM, NERVANA, TENSTORRENT, Artificial Learning, rigetti

DATA PREP
TRIFACTA, Paxata, tamr, Alation

DATA COLLECTION
diffbot, kimono, CrowdFlower, Connotate, WorkFusion, import.io

www.shivonzilis.com/machineintelligence

BloombergBETA



THE FUTURE OF BLOCKCHAIN COMPUTING

Olaf Carlson-Wee
Head of Risk, Coinbase

Algorithms For Finding The Bad Guys

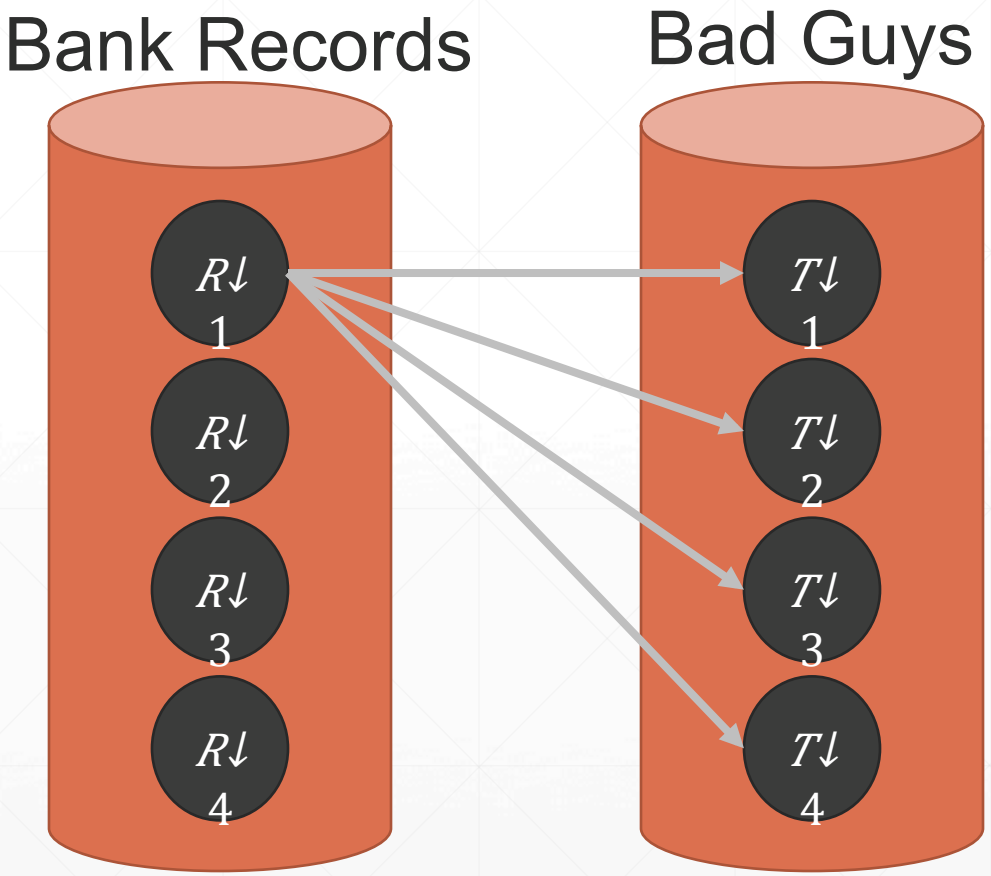
Richard Minerich, Director of R&D at BAYARD ROCK
@Rickasaurus

Entity Resolution in Theory

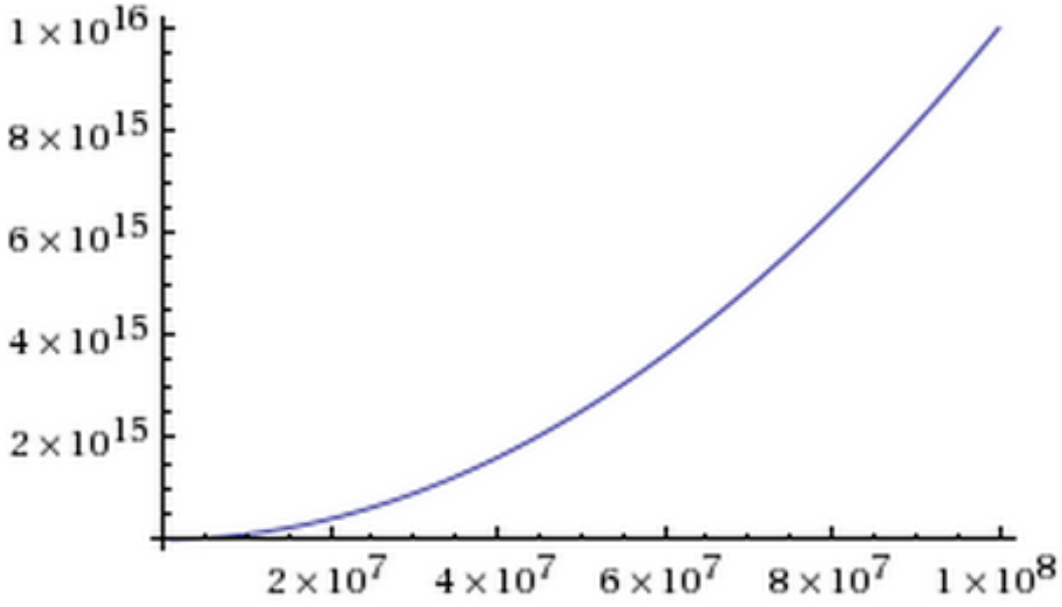
$$P(R\downarrow n \text{ reprents the same entity as } T\downarrow m)$$

Example Variations:
- Aggregating Products
- Finding Medical Records
- Resolving Paper Authors
- Census
- Finding Bad Guys
- Database Deduping

Different tradeoffs per domain



Too Many Comparisons!



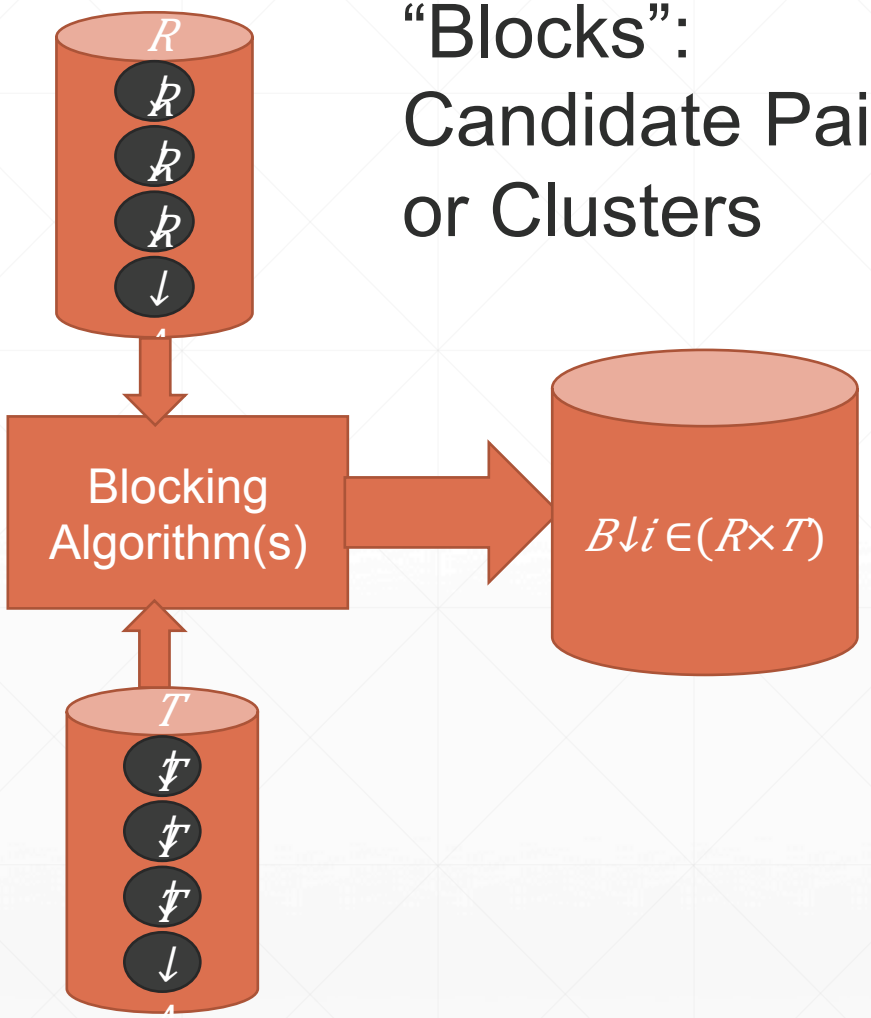
- 100 Million x 100 Million = 10 quadrillion pairs
- 86,400,000 milliseconds per day
- One pair per ms: ~116,000,000 days to compute (~317K years)

How do we beat N*M? Blocking Algorithms.

Input:
- Source Records R
- Target Records T

Output:
- Blocks of Similar Records

$B_i \in (R \times T)$



"Blocks": Candidate Pairs or Clusters

Simplest: Key-based Blocking

RecID	GivenName	Surname	Postcode	Suburb
r1	peter	christen	2010	north sydney
r2	paul	smith	2600	canberra
r3	pedro	kristen	2000	sydeny
r4	pablo	smyth	2700	canberra sth

- Nothing's easier than a table lookup!
- Many ways to key, choosing is hard
- Small errors can cause misses
- What about missing data?

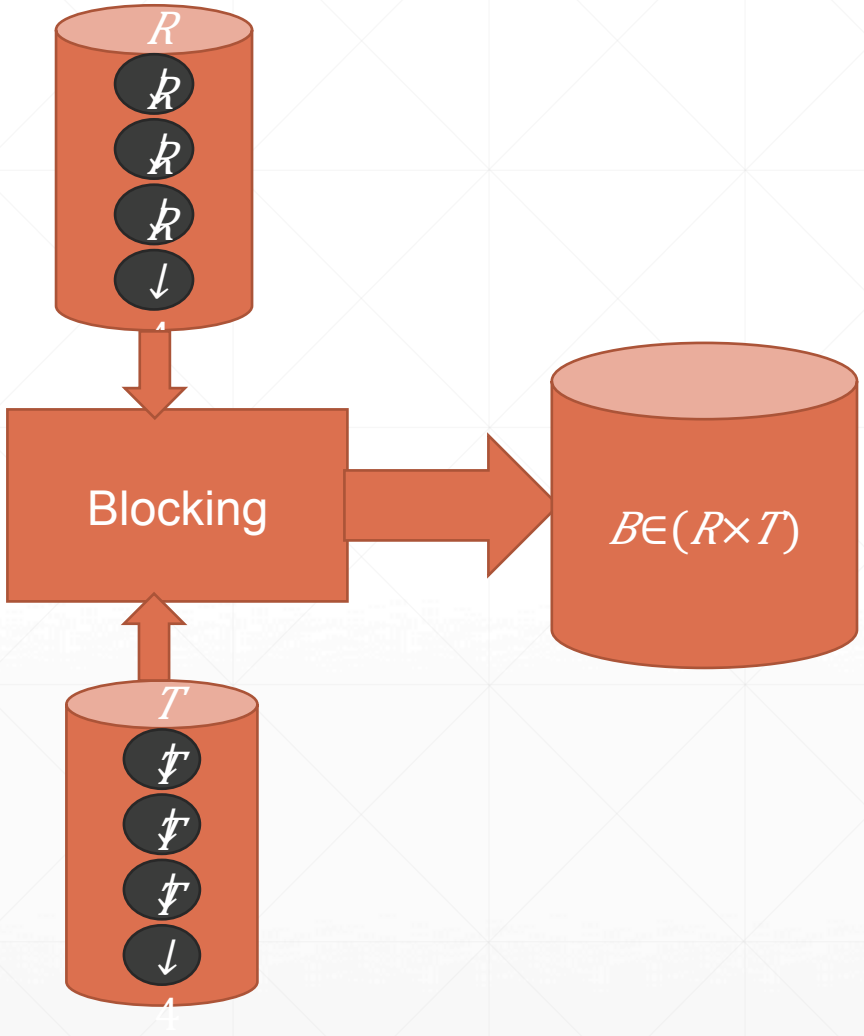
RecID	PC+Sndx(GiN)	Fi2D(PC)+DMe(SurN)	La2D(PC)+Sndx(SubN)
r1	2010-p360	**20-krst**	10-n632
r2	2600-p400	26-sm0	**00-c516**
r3	2000-p360	**20-krst**	00-s530
r4	2700-p140	27-sm0	**00-c516**

Table: Peter Christen - Data Matching 2012

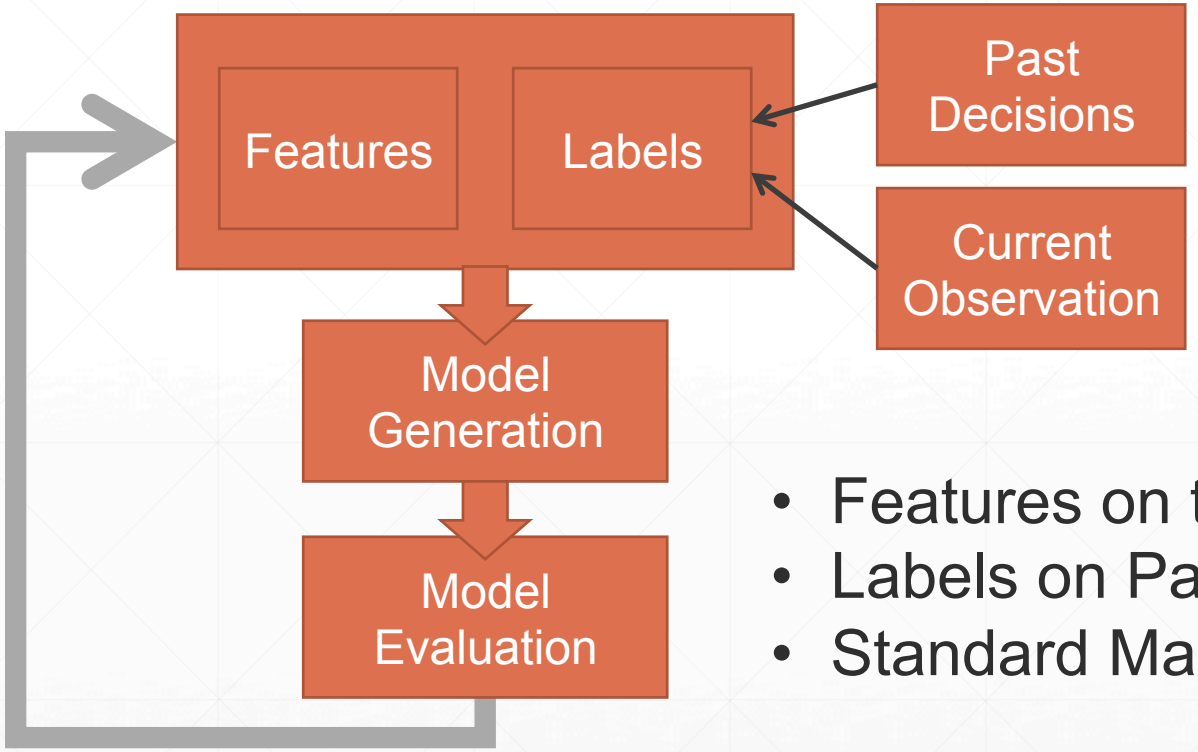
Many ways to Block

- Sorted Neighborhood
- Suffix Arrays/Trees
- Various kinds of Q-gram Indices
- Metric Space Embedding
- Semantic Hashing
- Cluster-based approaches

Best to use a mix.



The Basics of Pairwise Scoring



NAME	LARRY O' BRIAN
STATE	CANADA
CITY	Montreal
STATE	Quebec
ADDRESS	121 Buffalo Drive
ZIP	H3G 1Z2
DOB	10/24/80

- Features on the Similarity of Fields
- Labels on Pairs are True/False
- Standard Machine Learning Techniques Apply

Simplest:
Empirical Summed Similarity

- F: feature functions (0 .. m) : (a,b) -> [0, 1]
- W: feature weights (0 .. m) : {0+}
- $SimSum(a,b) = \sum i=0 \uparrow m \; f \downarrow i \; w \downarrow i$

Thresholds such that:

Match: SimSum(a,b) >= Upper
Review: Lower <= SimSum(a,b) <= Upper
Discard: SimSum(a,b) <= Lower

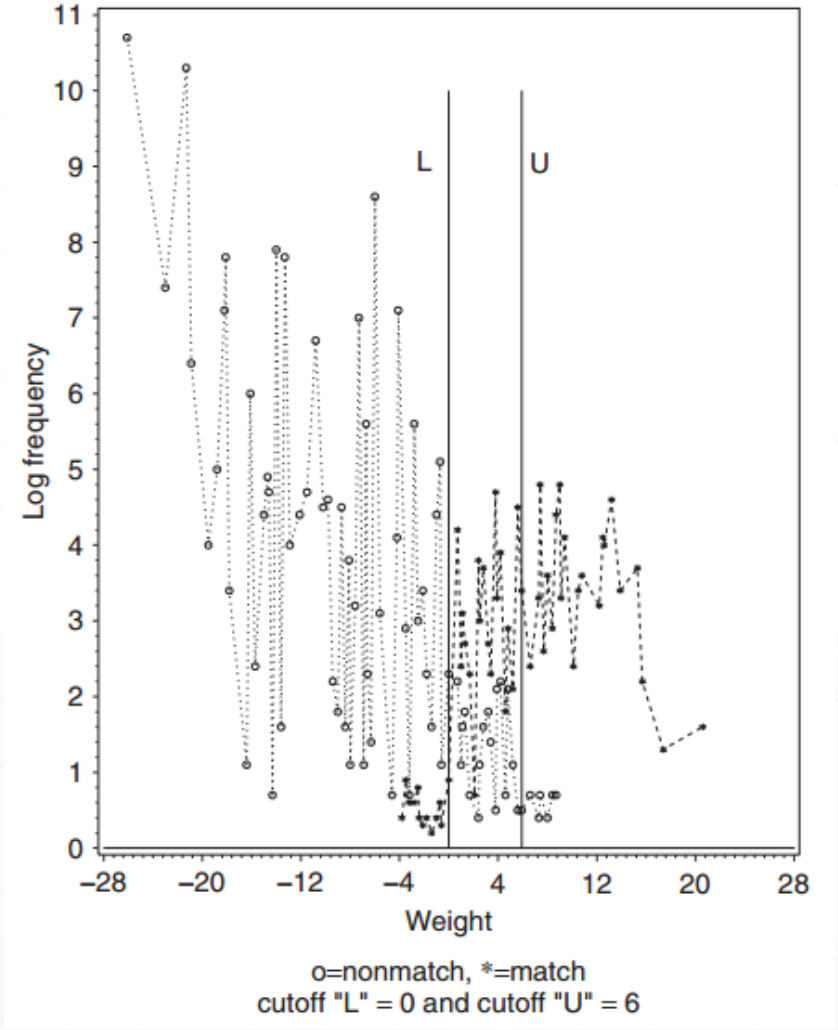


Image Via: https://www.cs.umd.edu/class/spring2012/cmsc828L/Papers/HerzogEtWires10.pdf

The Pairwise Entity Resolution Process

Blocking

- Two Datasets (Customer Data and Sanctions)
- Pairs of Somehow Similar Records

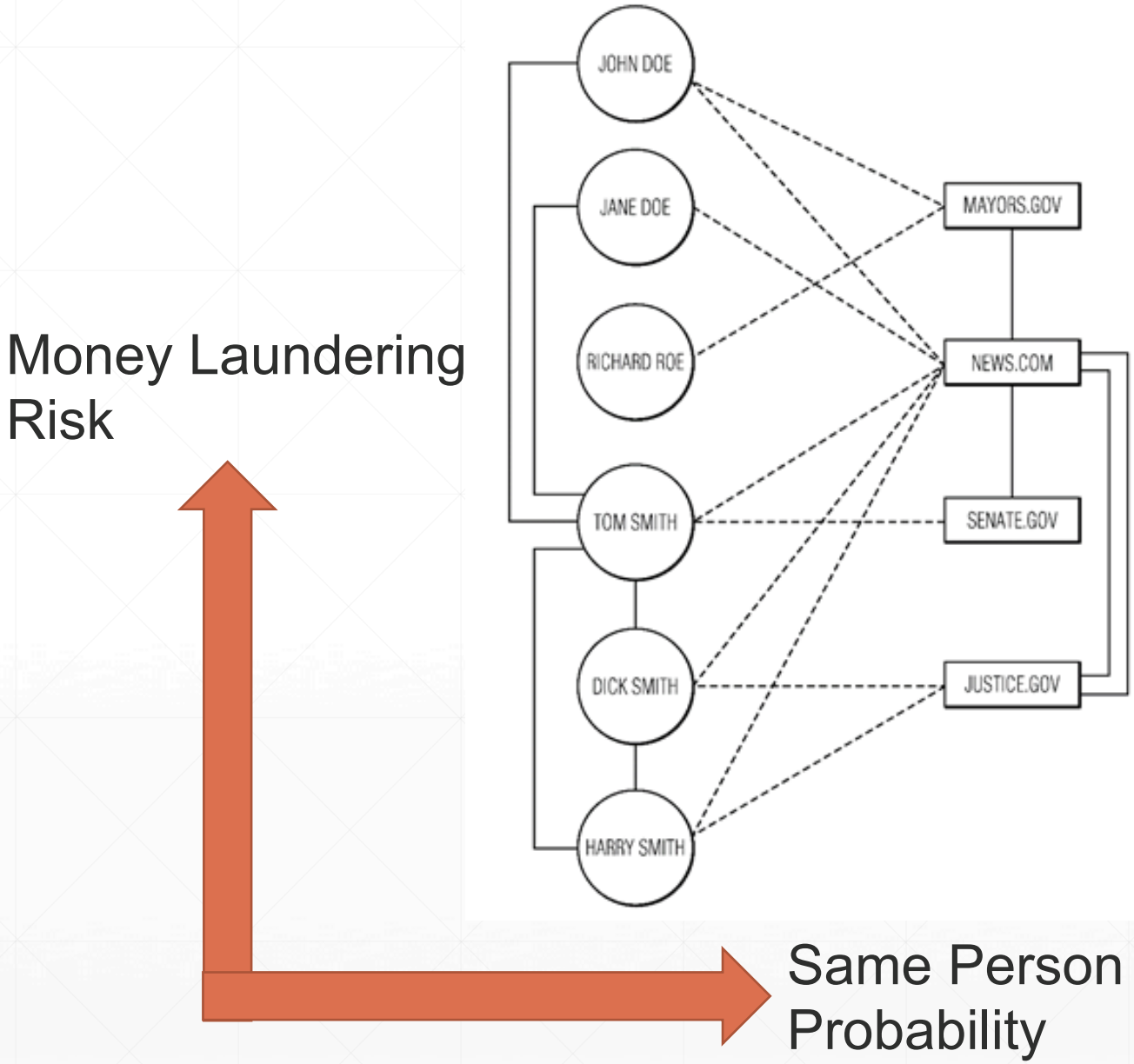
Scoring

- Pairs of Records
- Probability of Representing Same Entity

Review

- Records, Probability, Similarity Features
- True/False Labels (Mostly by Hand)

Another Dimension to aid review: Risk vs Probability



Useful in many domains:

- How Likely To Launder Money?
- How much money do I lose if I get the product wrong?
- Risk of Incorrect Medical Diagnosis

Simplest Ranking?
You Can "Learn to Rank" with Regression.

- The features are the difference in would-be regression features
- The value to predict is the difference in label rank

Select 2 labeled samples randomly => (x1,y1) (x2,y2)

x = x1 – x2
y = y1 – y2

	Sample 1	Sample 2	Result
Names?	1	1	0
Addresses?	1	0	1
DOB?	0	1	-1
Same Person?	0	0	0

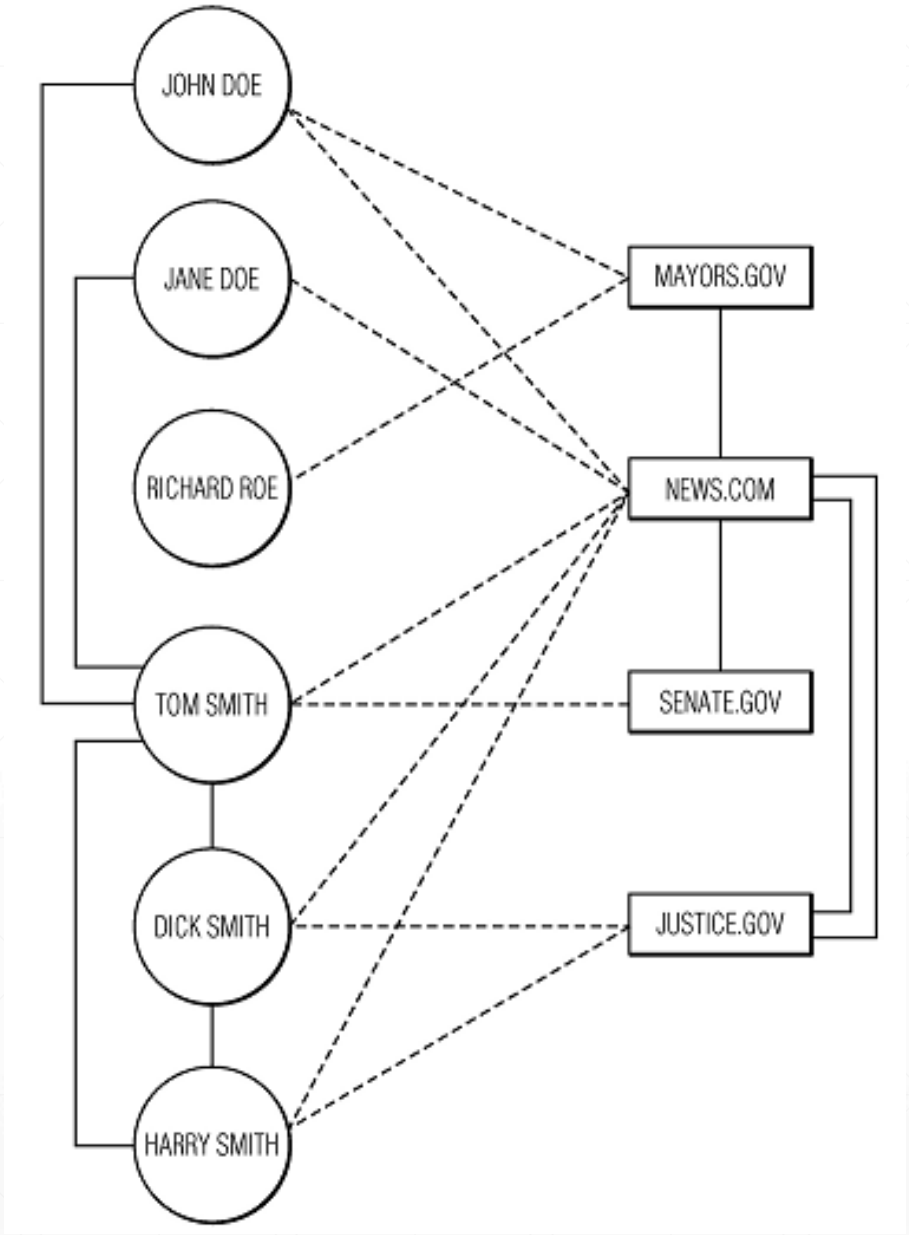
Page Rank: The Easy Parts

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

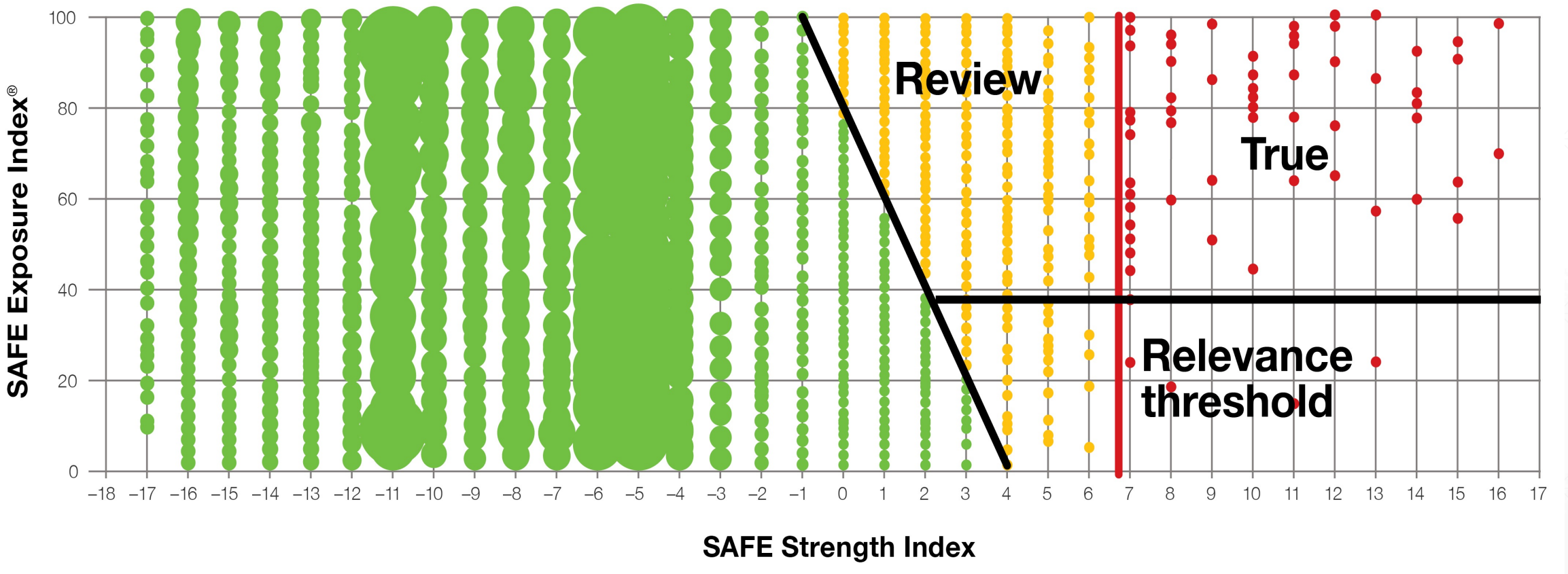
```
//Calculate the ranking given a matrix and initial vector
let private calcRanking (A:matrix) (E:Vector<_>) (x:float) (y:float) (iterations:int) =
    let rec calc R n =
        //Calculate new matrix
        let R' = x*(A*R) + y*E

        //Decide when to return results
        if n = iterations then ((Vector.norm(R' - R)), R)
        else calc R' (n + 1)
    calc E 1
```

Page Rank: The Hard Parts

- Domains, Websites, Pages in Context
- Determining Initial Risk for Sources
- 27 Pages of Data Transformation Code
- Fluctuation with no changes
- Prediction and Explainability



Combining Ranking and Probability: Big Picture



Thank You! Questions?

You can read more on my blog at:
http://richardminerich.com

Contact me on twitter:
@Rickasaurus 

Email me with questions:

rick@bayardrock.com

Check out the NYC F# User Group:
http://www.meetup.com/nyc-fsharp

Questions

QCon