



# Machine Learning at JPL

Alphan Altinok, Ph.D.

Machine Learning and Instrument Autonomy

Jet Propulsion Laboratory, California Institute of Technology

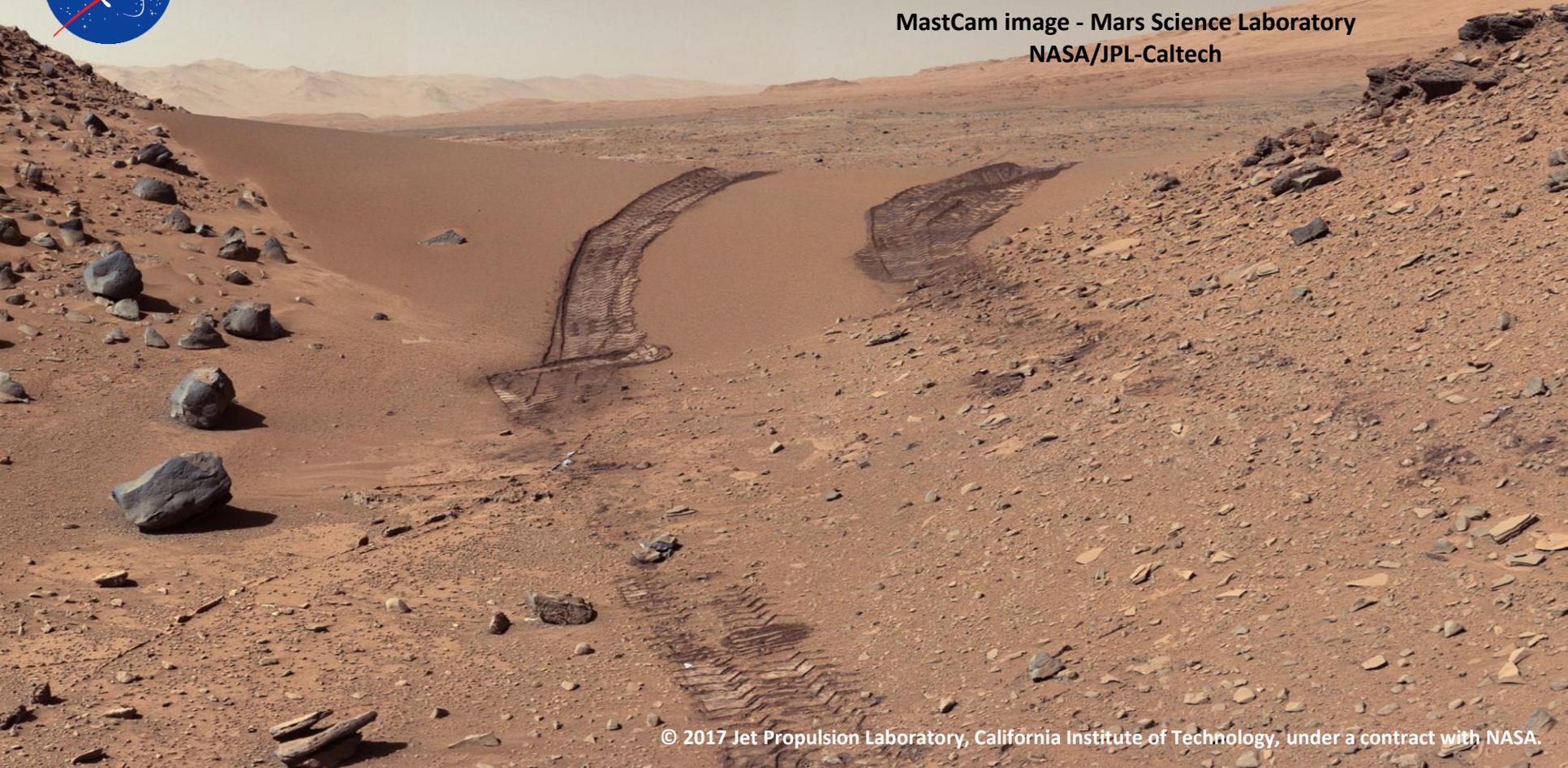
© 2017 California Institute of Technology. Government sponsorship acknowledged.

Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.



# Machine Learning in Support of Space Exploration

MastCam image - Mars Science Laboratory  
NASA/JPL-Caltech





# HiRISE has collected >1.2M amazing views of MARS



- 30M planetary images
- Petabytes of non-image data
- Increasing daily



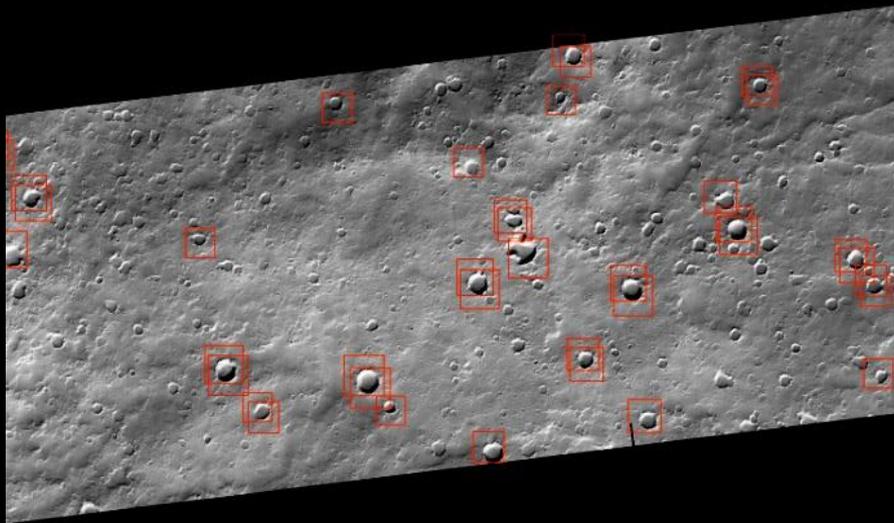
# How do we answer science questions from image data?



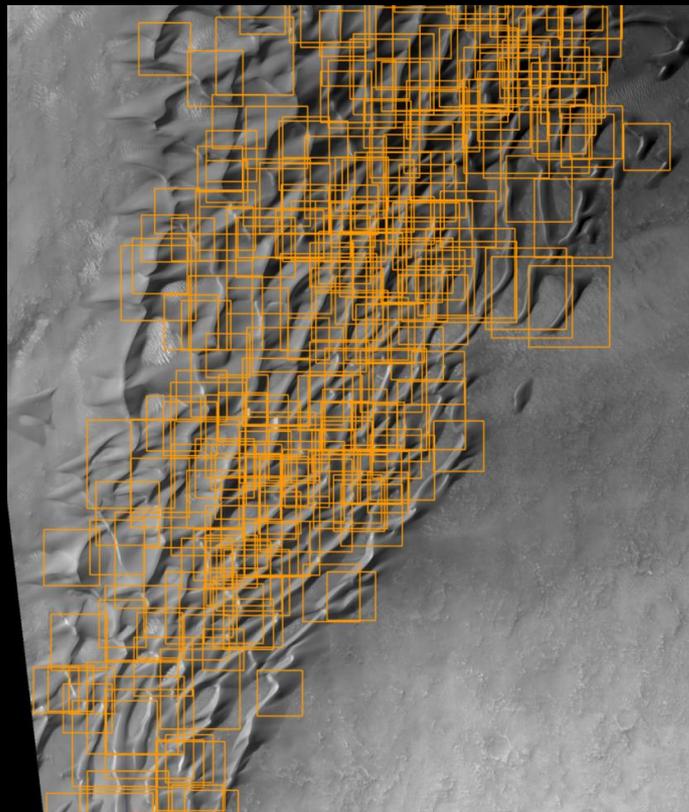
- Machine learning systems that learn from human-provided examples help us explore new environments
- Enable search for features of interest in very large data sets



# Landmarks are now searchable in M's of images

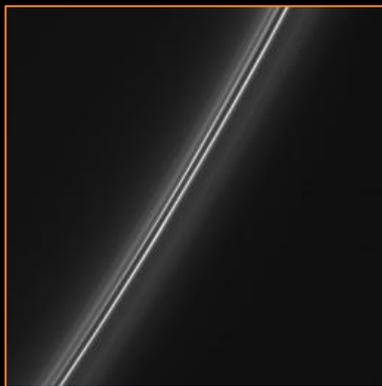


**Deep Neural Networks  
for landmark classification**





<http://pds-imaging.jpl.nasa.gov/search/>

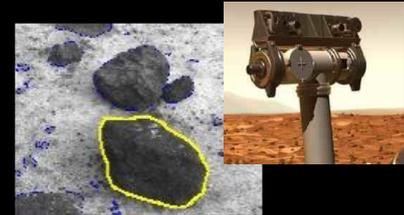


## Searching Saturn's RINGS in 29M PDS images

A. Stanboli, B. Bue, K. L. Wagstaff, A. Altinok  
Based on ImageNet, Krizhevsky et al., 2012

# AEGIS Autonomous Exploration for Gathering Increased Science

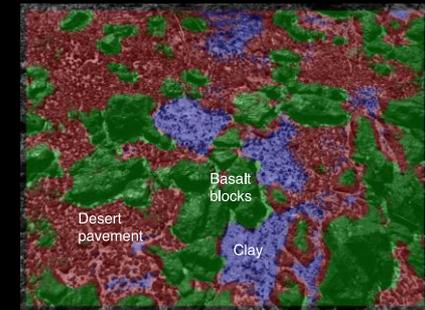
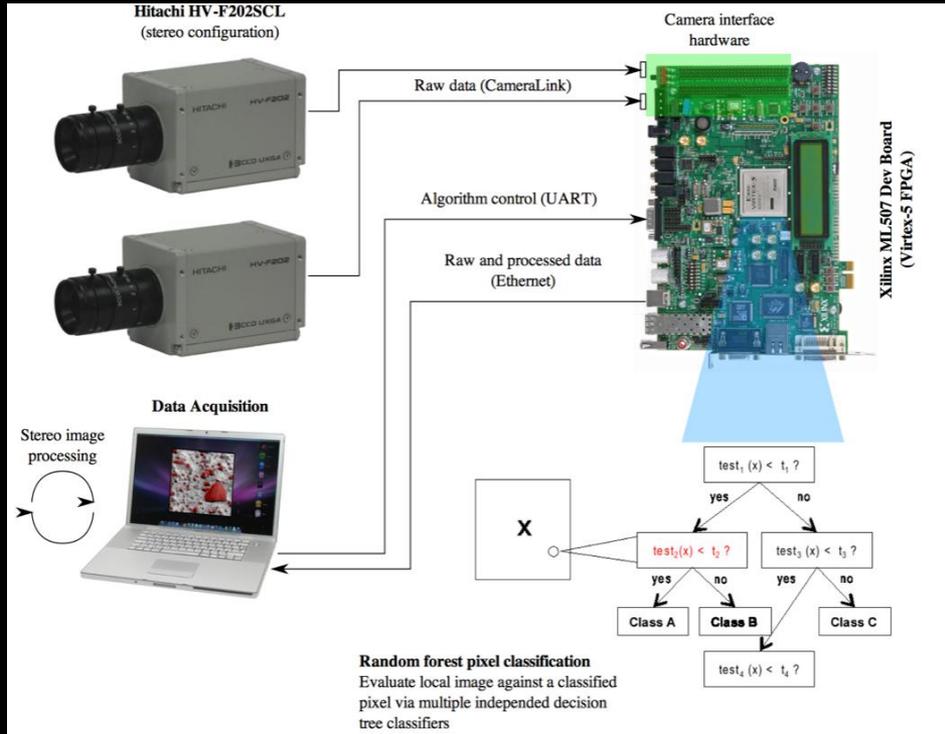
- Operational onboard Mars Exploration Rover (MER) Opportunity rover and recently uploaded MSL Curiosity rover
- Provides intelligent targeting and data acquisition capabilities
- Winner of 2011 NASA Software of the Year Award
- Provides automated data collection for rover remote sensing instruments
  - Identify rock targets onboard
  - Guided by scientist specified criteria
  - Can be run at end of drive or mid drive
  - No communication with ground required



## *AEGIS Automated Targeting for the MER Opportunity Rover*

T. Estlin, B. Bornstein, D. Gaines, R. C. Anderson, D. Thompson, M. Burl, R. Castano, and M. Judd. ACM Transactions on Intelligent Systems and Technology, 3(3), 2012

# TextureCam Automated Image Classification



*Cloud Filtering and Novelty Detection using Onboard Machine Learning for the EO-1 Spacecraft*

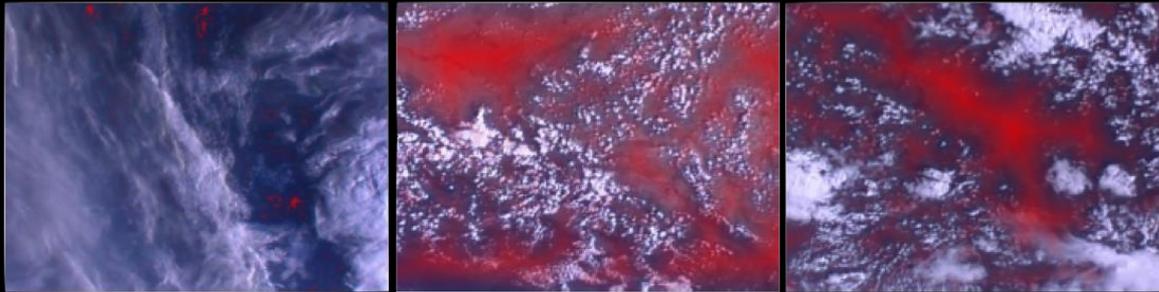
K. L. Wagstaff, A. Altinok, S. Chien, U. Rebbapragada, S. Schaffer, D. R. Thompson, and D. Tran. *IJCAI 2017 Workshop on AI*



# TextureCam Cloud Screening and Re-targeting – IPEX CubeSat



## Random Decision Forests - Support Vector Machines



*Real-Time Orbital Image Analysis Using Decision Forests, with a Deployment Onboard the IPEX Spacecraft*  
A. Altinok, D. R. Thompson, B. Bornstein, S. Chien, J. Doubleday, and J. Bellardo, Journal of Field Robotics, 2015



# Connected Driver Analyzing and Sharing Info



cloud



infotainment



crowd



real-time reports



other drivers



infrastructure

# Autonomous Driver Flood of Data

GPS  
50 kb/sec

SONAR  
10-100 kb/sec

RADAR  
10-100 kb/sec

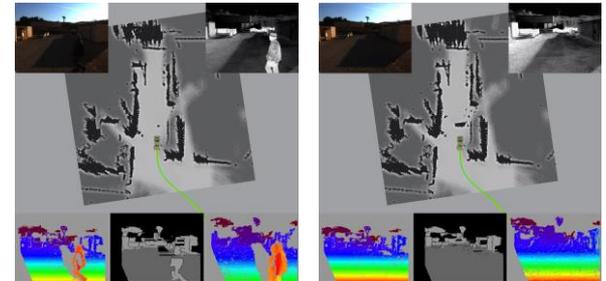
LIDAR  
10-100 mb/sec



CAMERAS  
20-40 mb/sec

**VEHICLE: 4 TB / day**

Data Fusion at JPL

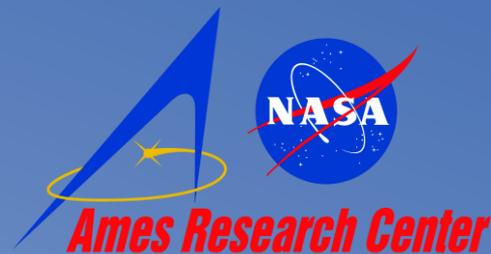
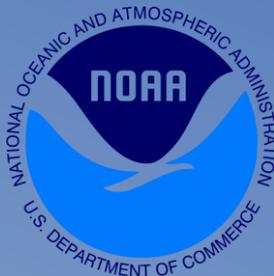




# National Airspace System

Congestions cost \$Bs / year  
Airports are bottleneck points

# High dimensional data from multiple sources



## Significant data issues

- Noise
- Alignment**
- Labeling
- Missing
- Standards
- ...

## Solutions

- Statistical
- Algorithmic
- Domain-bound

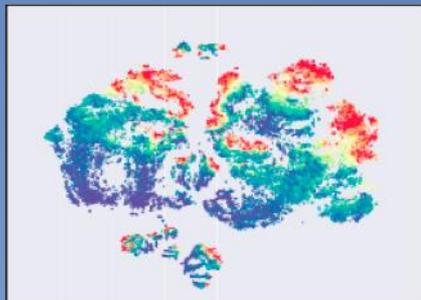


# Exploring high-dimensional data

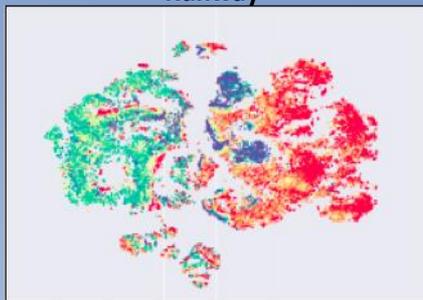
t-SNE: t-Distributed Stochastic Neighbor Embedding



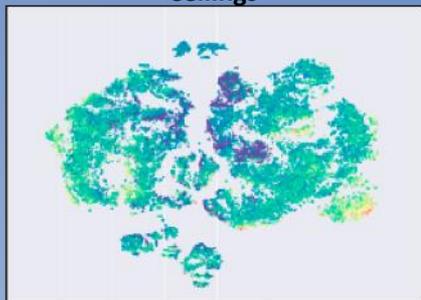
Runway



Ceilings



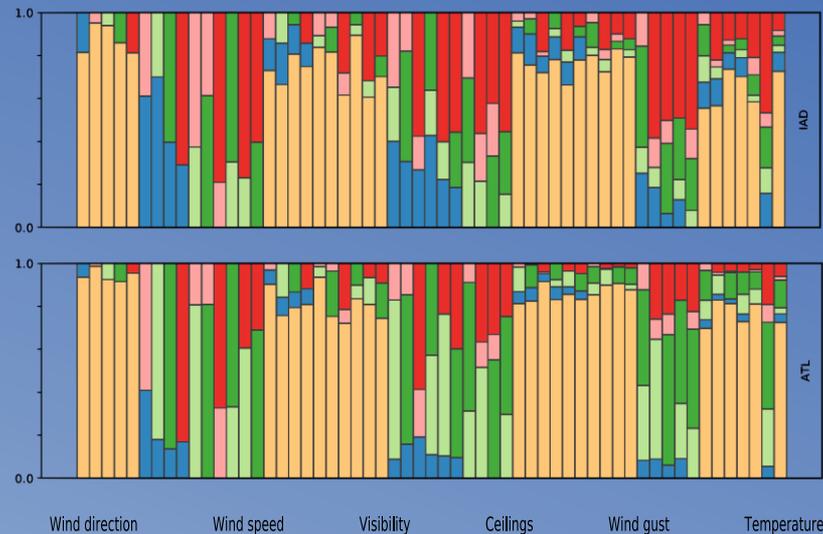
Wind Direction



Wind Speed

Similar color separation indicates correlation in high-dimensional space

Random Decision Forest - Feature Importance



If Wind Direction is present, it predicts runway configuration on ~80% of op hours



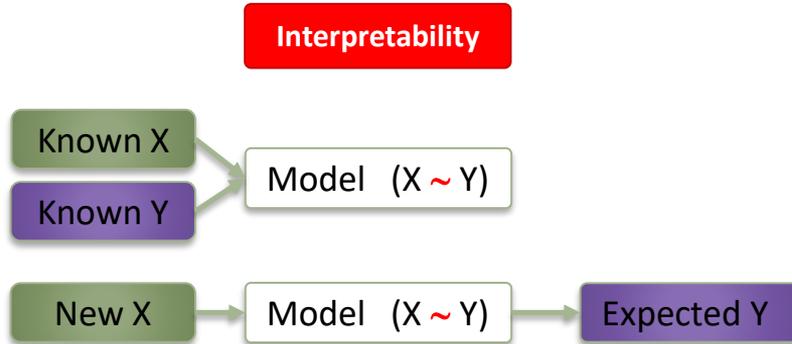
# Machine Learning and Data Science



# Machine Learning methods

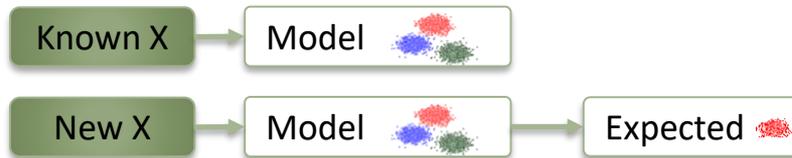
## Supervised Learning

- Logistic Regression
- Decision Trees
- Nearest Neighbors
- Random Decision Forests
- Support Vector Machines
- Naïve Bayes



## Unsupervised Learning

- Clustering methods
- Dimensionality reduction methods



## Reinforcement Learning

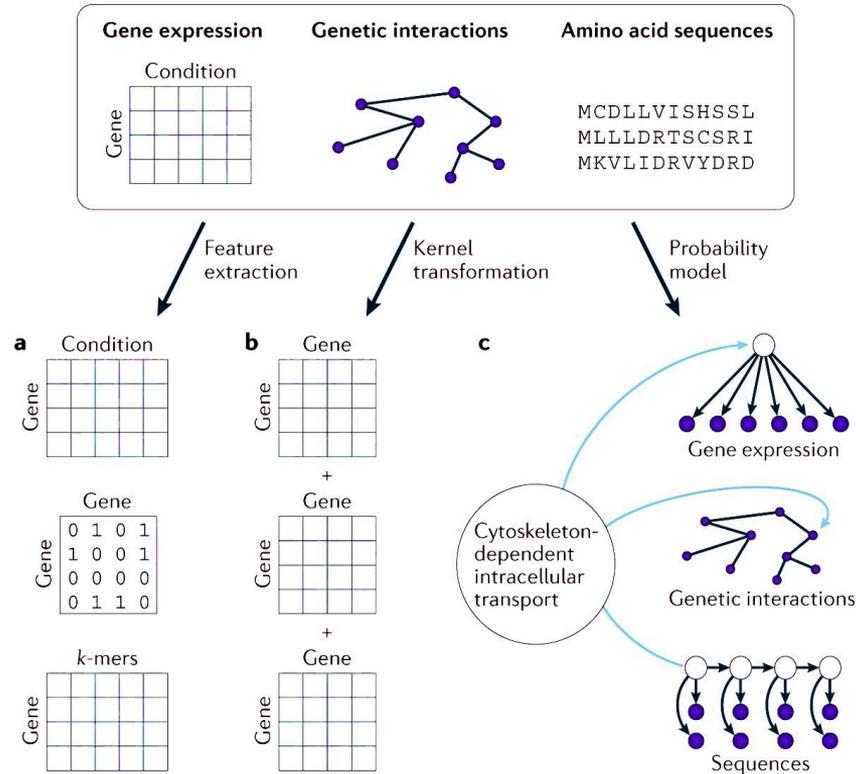
- Markov Decision Processes
- Temporal Difference
- Policy Search

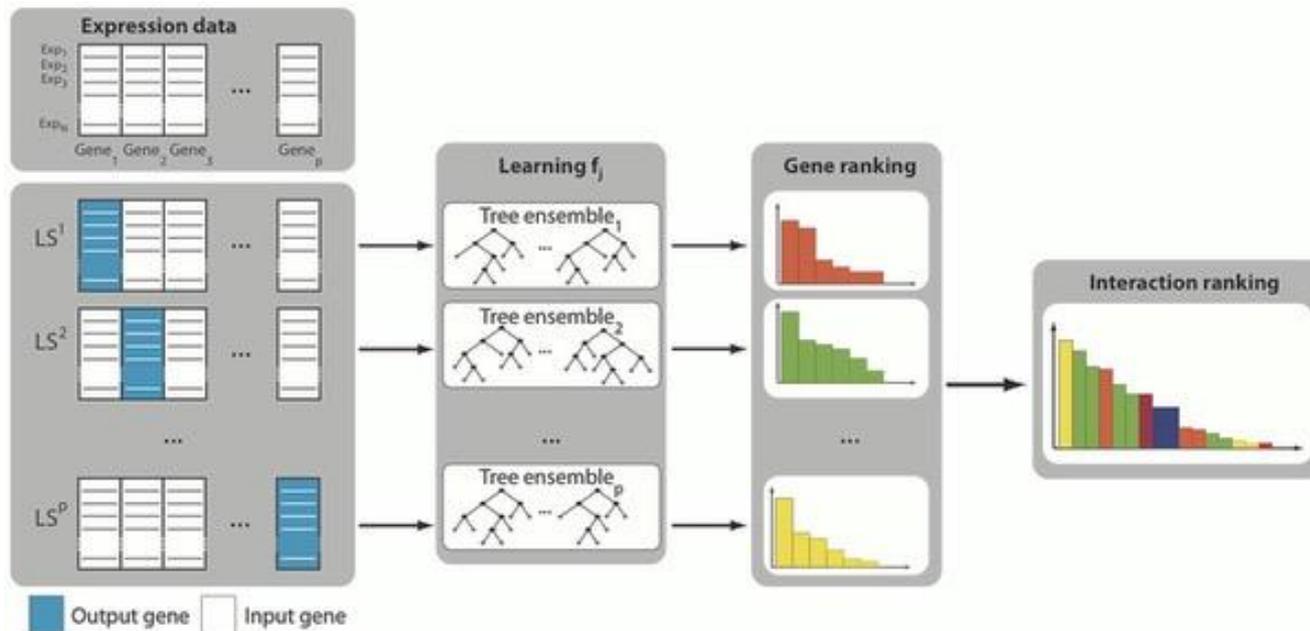


Deep Learning, Domain Adaptation, Transfer Learning, Self-supervised Learning



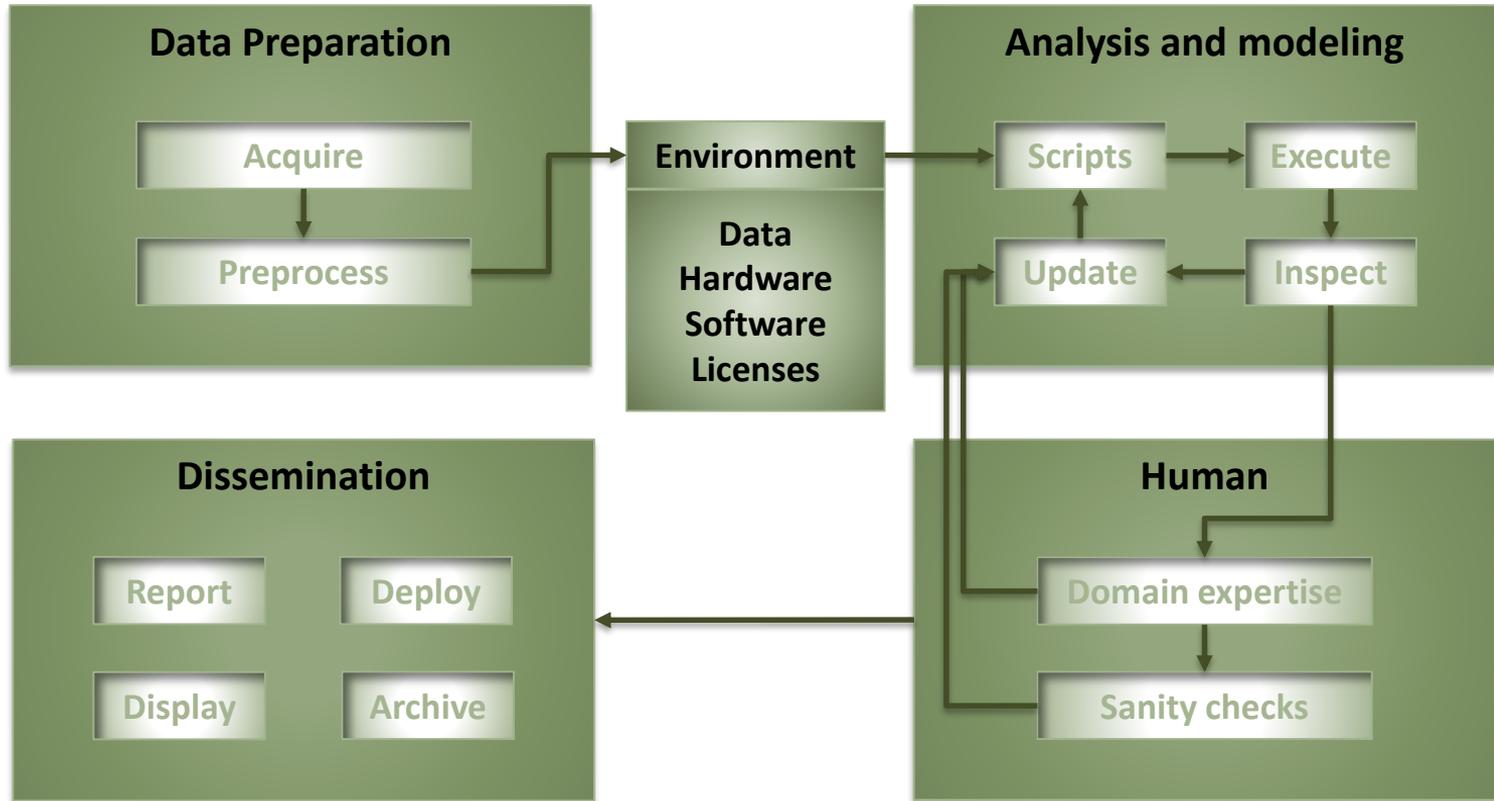
# Machine learning can fuse data in different formats





Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLOS ONE 5(9): e12776. <https://doi.org/10.1371/journal.pone.0012776>  
<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0012776>

# Data Science Workflow



# **Machine Learning and Data Science in Healthcare and Diet**



# EHR – Data Analytics



190M patients with current EHR



Insights

Insights  
Confronting the Opioid Epidemic

Insights  
“I’m Quite Sure in the Old Paper World, She Would Have Died”

Insights  
Decision Support in Epic Helps Reduce Insulin Dosing Errors by 50%

Insights  
Safety Net Patients Still Use MyChart

“I’m Quite Sure in the Old Paper World, She Would Have Died”  
Illustrating how EHRs help save lives

“More than 1.6 million patients have had Epic predictive algorithms run on them for hypertension, diabetes, asthma, and heart failure, which helps providers create targeted care management programs. That is just the beginning of the deep dive into machine learning.”

# EHR – Data Analytics



Formerly Humedica

UnitedHealth data subsidiary

Patented NLP stack for EHR processing



# EHR – Data Analytics

READINESS ASSESSMENT HAS 17 STAY INFORMED HC COMMUNITY SUPPORT CONTACT US



APPROACH PRODUCTS SOLUTIONS SUCCESS STORIES NEWS INSIGHT

FEATURED

## A Framework for Health Data Analytics

Presented by Amy Flaster, MD, Assistant Professor

LATE-BINDING™ DATA WAREHOUSE  
DATA OPERATING SYSTEM (DOS™)  
HEALTH CATALYST® ANALYTICS PLATFORM  
APPLICATIONS  
PROFESSIONAL SERVICES  
CLOUD SERVICES  
TYPICAL TIMELINES

orm

Data Warehousing / Analytics  
Business Intelligence  
Process Improvement Services

Health Catalyst has already built over a dozen predictive models for clinical, financial, and operational decision support:

- CLABSI risk
- CHF readmission risk
- Diabetes future risk
- Diabetes net likely complication
- Forecast incurred but not reported claims/year-end expenditures
- Propensity to pay risk
- Appointments no show risk

Health Catalyst has even more predictive models planned for the first quarter of 2017:

- Clinical and operational decision support for bowel surgery
- Hip and knee surgery
- Coronary artery disease
- Geo-spatial network referral and leakage prevention
- Early detection of CAUTI and sepsis
- Expected mortality and length of stay
- And more.



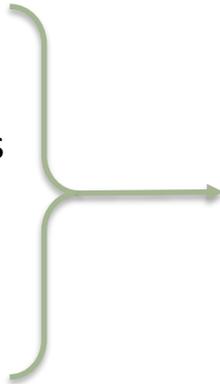
# QuantifiedSelf / Lifelogging

Self-tracking with technology

Movement to incorporate technology into data acquisition on aspects of a person's daily life

## Wearable tech

- Activity monitors
- Sleep-specific monitors
- Diet-and-weight monitors
- Environment monitors
- Posture monitors
- Brain sensors
- Other wearables



## Data Aggregation Apps, Services and API's

- Exist.io
- Gyrosco.pe
- Human/API
- MyFitnessPal
- Open mHealth



### About the Quantified Self

*Our mission is to support new discoveries about ourselves and our communities that are grounded in accurate observation and enlivened by a spirit of friendship.*

Quantified Self Labs is a California-based company founded by [Gary Wolf](#) and [Kevin Kelly](#) that serves the Quantified Self user community worldwide by producing international meetings, conferences and expositions, community forums, web content and services, and a guide to self-tracking tools. Are you interested in self-tracking? Do you have questions to ask or knowledge to share? We welcome your questions and contributions. We are here to help.



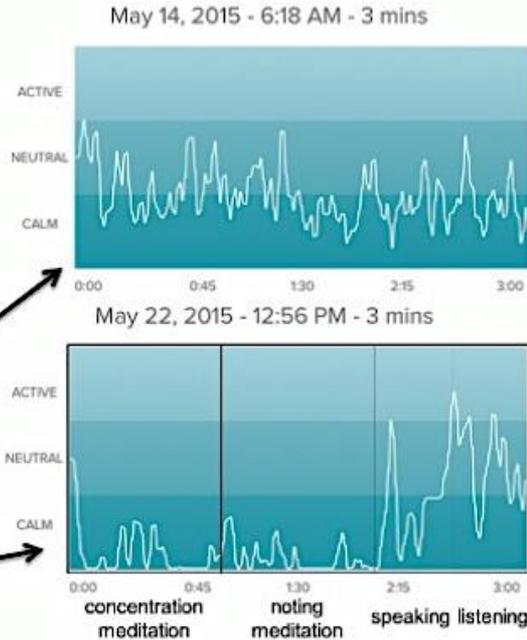
# QuantifiedSelf / Lifelogging

Muse – the brain sensing headband



The guy that bought the headband trying to meditate

Experienced meditator



Many known/unknown covariates affect data.

Law of Large Numbers ~ Law of the Needle in the Haystack (needed)

# QuantifiedSelf / Lifelogging

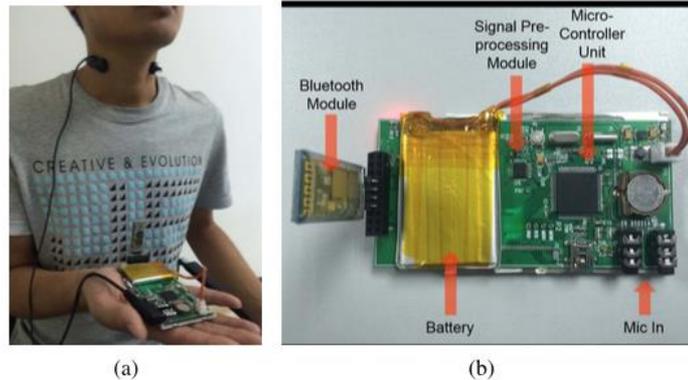


Fig. 2. System illustrations. (a) A user wearing AutoDietary; (b) Details of the hardware board.

## AutoDietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life

Yin Bi, Mingsong Lv, Chen Song, Wenyao Xu, Nan Guan, *Member, IEEE*, and Wang Yi, *Fellow, IEEE*

# Apps

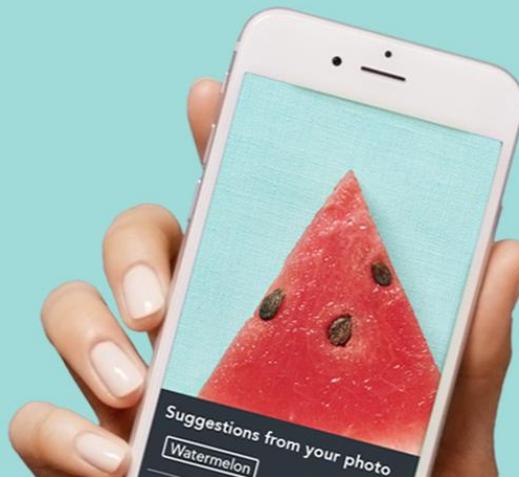
 Lose It!

## INTRODUCING SNAP IT BY LOSE IT!

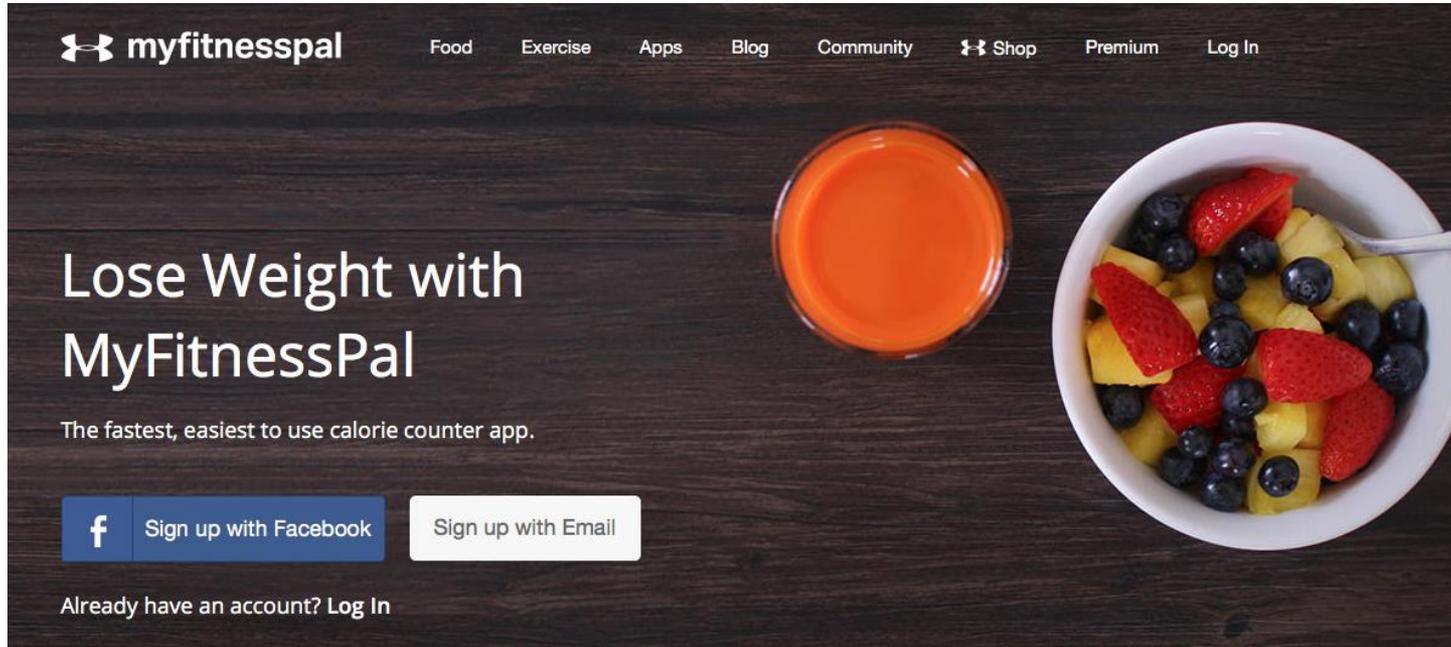
Now, tracking your food  
is as easy as snapping a picture.

[DOWNLOAD THE FREE APP](#)

[WATCH THE VIDEO](#)



# Apps

The image shows the landing page for the MyFitnessPal app. The background is a dark wood-grain texture. At the top left is the MyFitnessPal logo, which consists of a stylized 'M' and 'P' icon followed by the text 'myfitnesspal'. To the right of the logo is a navigation menu with the following items: 'Food', 'Exercise', 'Apps', 'Blog', 'Community', 'Shop' (with a shopping cart icon), 'Premium', and 'Log In'. Below the navigation menu, on the right side, is a photograph of a white bowl filled with fresh fruit (strawberries, blueberries, and pineapple) and a glass of orange juice. On the left side, the main headline reads 'Lose Weight with MyFitnessPal' in large white font. Below the headline is the subtext 'The fastest, easiest to use calorie counter app.' At the bottom left, there are two buttons: a blue button with a white Facebook 'f' icon and the text 'Sign up with Facebook', and a white button with a grey border and the text 'Sign up with Email'. Below these buttons is the text 'Already have an account? Log In'.

Weight loss, fitness, diet personalization.



# Apps



The image is a screenshot of the Nutrino app's website. It features a top-down view of a wooden table with various dishes including salads, bread, and olives. In the top left corner, the Nutrino logo is visible. In the top right corner, there are two buttons: 'Nutrino for Me' and 'Nutrino for Business'. Centered over the image is the text 'FoodPrint for Diabetes' in a large, bold, white font, followed by 'Now connected to your glucose monitoring device' in a slightly smaller white font. At the bottom center, there is a pink 'Download' button.

Diabetes management, diet personalization.  
Data : food, person, scientific literature.

# Apps

**Suggestic**  
PRECISION EATING™

## Make Food Medicine

We help you find a diet plan for weight loss, health optimization, or chronic disease reversal.

We then help you follow that plan whether at home, the grocery store or a restaurant.

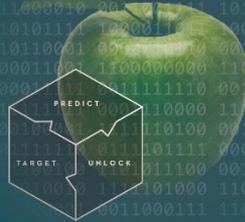


Deep learning predicts what diet regime works (doesn't work) for individuals.



PLATFORM

**Our proprietary platform targets, predicts and unlocks novel bioactive peptides from food sources. These deliver highly specific, efficient and life-changing health solutions.**



Deep learning to predict therapeutic properties of bioactive peptides.

## Balance your blood sugar with personalized nutrition.

What's healthy for others may not be healthy for you. Discover which foods help balance your blood sugar, based on your [gut microbiome](#).

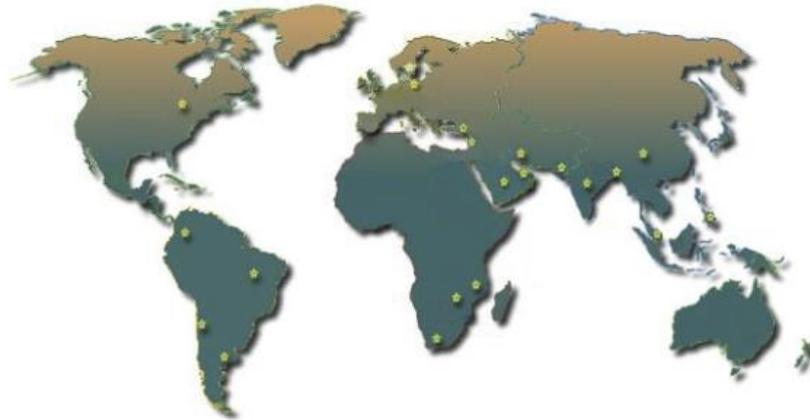
[Watch a demo of the DayTwo app](#)

[Order now](#)



# PURE (PROSPECTIVE URBAN AND RURAL EPIDEMIOLOGICAL STUDY)

Participating Countries and Territories: Argentina, Bangladesh, Brazil, Canada, Chile, China, Colombia, India, Iran, Kazakhstan, Kyrgyzstan, Malaysia, Pakistan, Palestine, Philippines, Poland, Russia, Saudi Arabia, South Africa, Sudan, Sweden, Tanzania, Turkey, United Arab Emirates, Zimbabwe.



**Tracks 135,000 individuals from 18 countries**

To examine the impact of urbanization on the development of primordial risk factors (for example: physical activity and nutrition changes), primary risk factors (for example: obesity, hypertension, dysglycemia and dyslipidemia, smoking), and CVD.



# Learning Phenotype from EHR for population based studies

J Biomed Inform. 2014 Dec;52:260-70. doi: 10.1016/j.jbi.2014.07.007. Epub 2014 Jul 15.

## **Relational machine learning for electronic health record-driven phenotyping.**

Peissig PL<sup>1</sup>, Santos Costa V<sup>2</sup>, Caldwell MD<sup>3</sup>, Rottscheit C<sup>4</sup>, Berg RL<sup>4</sup>, Mendonca EA<sup>5</sup>, Page D<sup>6</sup>.

AMIA Jt Summits Transl Sci Proc. 2013 Mar 18;2013:142-6. eCollection 2013.

## **Using association rule mining for phenotype extraction from electronic health records.**

Li D<sup>1</sup>, Simon G, Chute CG, Pathak J.

Med Care. 2010 Jun;48(6 Suppl):S106-13. doi: 10.1097/MLR.0b013e3181de9e17.

## **Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.**

Wu J<sup>1</sup>, Roy J, Stewart WF.



## Machine Learning to Compare Frequent Medical Problems of African American and Caucasian Diabetic Kidney Patients.

Kim YM<sup>1</sup>, Kathuria P<sup>2</sup>, Delen D<sup>3</sup>.

### ⊕ Author information

#### Abstract

**OBJECTIVES:** End-stage renal disease (ESRD), which is primarily a consequence of diabetes mellitus, shows an exemplary health disparity between African American and Caucasian patients in the United States. Because diabetic chronic kidney disease (CKD) patients of these two groups show differences in their medical problems, the markers leading to ESRD are also expected to differ. The purpose of this study was, therefore, to compare their medical complications at various levels of kidney function and to identify markers that can be used to predict ESRD.

**METHODS:** The data of type 2 diabetic patients was obtained from the 2012 Cerner database, which totaled 1,038,499 records. The data was then filtered to include only African American and Caucasian outpatients with estimated glomerular filtration rates (eGFR), leaving 4,623 records. A priori machine learning was used to discover frequently appearing medical problems within the filtered data. CKD is defined as abnormalities of kidney structure, present for >3 months.

**RESULTS:** This study found that African Americans have much higher rates of CKD-related medical problems than Caucasians for all five stages, and prominent markers leading to ESRD were discovered only for the African American group. These markers are high glucose, high systolic blood pressure (BP), obesity, alcohol/drug use, and low hematocrit. Additionally, the roles of systolic BP and diastolic BP vary depending on the CKD stage.

**CONCLUSIONS:** This research discovered frequently appearing medical problems across five stages of CKD and further showed that many of the markers reported in previous studies are more applicable to African American patients than Caucasian patients.

## Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study.

Olivera AR<sup>1</sup>, Roesler V<sup>2</sup>, Iochpe C<sup>2</sup>, Schmidt MI<sup>3</sup>, Vigo Á<sup>4</sup>, Barreto SM<sup>5</sup>, Duncan BB<sup>3</sup>.

### ⊕ Author information

#### Abstract

**CONTEXT AND OBJECTIVE::** Type 2 diabetes is a chronic disease associated with a wide range of serious health complications that have a major impact on overall health. The aims here were to develop and validate predictive models for detecting undiagnosed diabetes using data from the Longitudinal Study of Adult Health (ELSA-Brasil) and to compare the performance of different machine-learning algorithms in this task.

**DESIGN AND SETTING::** Comparison of machine-learning algorithms to develop predictive models using data from ELSA-Brasil.

**METHODS::** After selecting a subset of 27 candidate variables from the literature, models were built and validated in four sequential steps: (i) parameter tuning with tenfold cross-validation, repeated three times; (ii) automatic variable selection using forward selection, a wrapper strategy with four different machine-learning algorithms and tenfold cross-validation (repeated three times), to evaluate each subset of variables; (iii) error estimation of model parameters with tenfold cross-validation, repeated ten times; and (iv) generalization testing on an independent dataset. The models were created with the following machine-learning algorithms: logistic regression, artificial neural network, naïve Bayes, K-nearest neighbor and random forest.

**RESULTS::** The best models were created using artificial neural networks and logistic regression. -These achieved mean areas under the curve of, respectively, 75.24% and 74.98% in the error estimation step and 74.17% and 74.41% in the generalization testing step.

**CONCLUSION::** Most of the predictive models produced similar results, and demonstrated the feasibility of identifying individuals with highest probability of having undiagnosed diabetes, through easily-obtained clinical data.

## Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.

Wu J<sup>1</sup>, Roy J, Stewart WF.

### ⊕ Author information

#### Abstract

**BACKGROUND:** Electronic health record (EHR) databases contain vast amounts of information about patients. Machine learning techniques such as Boosting and support vector machine (SVM) can potentially identify patients at high risk for serious conditions, such as heart disease, from EHR data. However, these techniques have not yet been widely tested.

**OBJECTIVE:** To model detection of heart failure more than 6 months before the actual date of clinical diagnosis using machine learning techniques applied to EHR data. To compare the performance of logistic regression, SVM, and Boosting, along with various variable selection methods in heart failure prediction.

**RESEARCH DESIGN:** Geisinger Clinic primary care patients with data in the EHR data from 2001 to 2006 diagnosed with heart failure between 2003 and 2006 were identified. Controls were randomly selected matched on sex, age, and clinic for this nested case-control study.

**MEASURES:** Area under the curve (AUC) of receiver operator characteristic curve was computed for each method using 10-fold cross-validation. The number of variables selected by each method was compared.

**RESULTS:** Logistic regression with model selection based on Bayesian information criterion provided the most parsimonious model, with about 10 variables selected on average, while maintaining a high AUC (0.77 in 10-fold cross-validation). Boosting with strict variable importance threshold provided similar performance.

**CONCLUSIONS:** Heart failure was predicted more than 6 months before clinical diagnosis, with AUC of about 0.76, using logistic regression and Boosting. These results were achieved even with strict model selection criteria. SVM had the poorest performance, possibly because of imbalanced data.

## Hypoglycemia prediction using machine learning models for patients with type 2 diabetes.

Sudharsan B<sup>1</sup>, Peeples M<sup>1</sup>, Shomali M<sup>2</sup>.

### ⊕ Author information

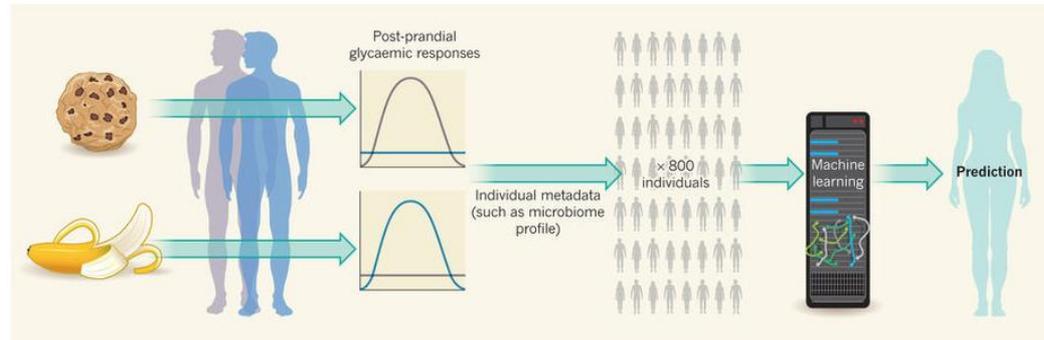
#### Abstract

Minimizing the occurrence of hypoglycemia in patients with type 2 diabetes is a challenging task since these patients typically check only 1 to 2 self-monitored blood glucose (SMBG) readings per day. We trained a probabilistic model using machine learning algorithms and SMBG values from real patients. Hypoglycemia was defined as a SMBG value < 70 mg/dL. We validated our model using multiple data sets. In addition, we trained a second model, which used patient SMBG values and information about patient medication administration. The optimal number of SMBG values needed by the model was approximately 10 per week.

The sensitivity of the model for predicting a hypoglycemia event in the next 24 hours was 92% and the specificity was 70%. In the model that incorporated medication information, the prediction window was for the hour of hypoglycemia, and the specificity improved to 90%. Our machine learning models can predict hypoglycemia events with a high degree of sensitivity and specificity. These models-which have been validated retrospectively and if implemented in real time-could be useful tools for reducing hypoglycemia in vulnerable patients.

## Figure 1 : Machine learning for nutrition advice.

From: A personal forecast



Zeevi *et al.*<sup>2</sup> continuously monitored the blood glucose levels of 800 individuals over the course of a week, which gave an indication of their post-prandial glycaemic responses (PPGRs; a measure of how rapidly blood glucose levels rise after food consumption) to specific foods. They combined this with 137 other measurements from each person, including their body-mass index, cholesterol levels, diet, activity levels and the composition of their gut microbiome. The data were used to develop a machine-learned algorithm. The authors show that this algorithm can predict PPGRs in people who were not in the cohort used to train the model, and thus can be used to provide dietary recommendations for maintaining PPGRs that are associated with health.

## Machine Learning and Data Mining Methods in Diabetes Research.

Kavakiotis I<sup>1</sup>, Tsave O<sup>2</sup>, Salifoglou A<sup>2</sup>, Maglaveras N<sup>3</sup>, Vlahavas I<sup>4</sup>, Chouvarda I<sup>3</sup>.

### ⊕ Author information

#### Abstract

The remarkable advances in biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information, generated from large Electronic Health Records (EHRs). To this end, application of machine learning and data mining methods in biosciences is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. Diabetes mellitus (DM) is defined as a group of metabolic disorders exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data. The aim of the present study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the first category appearing to be the most popular. A wide range of machine learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in DM.



## Reverse Engineering and Evaluation of Prediction Models for Progression to Type 2 Diabetes: An Application of Machine Learning Using Electronic Health Records.

Anderson JP<sup>1</sup>, Parikh JR<sup>2</sup>, Shenfeld DK<sup>2</sup>, Ivanov V<sup>2</sup>, Marks C<sup>2</sup>, Church BW<sup>2</sup>, Laramie JM<sup>2</sup>, Mardekian J<sup>3</sup>, Piper BA<sup>3</sup>, Willke RJ<sup>3</sup>, Rublee DA<sup>3</sup>.

### ⊕ Author information

#### Abstract

**BACKGROUND:** Application of novel machine learning approaches to electronic health record (EHR) data could provide valuable insights into disease processes. We utilized this approach to build predictive models for progression to prediabetes and type 2 diabetes (T2D).

**METHODS:** Using a novel analytical platform (Reverse Engineering and Forward Simulation [REFS]), we built prediction model ensembles for progression to prediabetes or T2D from an aggregated EHR data sample. REFS relies on a Bayesian scoring algorithm to explore a wide model space, and outputs a distribution of risk estimates from an ensemble of prediction models. We retrospectively followed 24 331 adults for transitions to prediabetes or T2D, 2007-2012. Accuracy of prediction models was assessed using an area under the curve (AUC) statistic, and validated in an independent data set.

**RESULTS:** Our primary ensemble of models accurately predicted progression to T2D (AUC = 0.76), and was validated out of sample (AUC = 0.78). Models of progression to T2D consisted primarily of established risk factors (blood glucose, blood pressure, triglycerides, hypertension, lipid disorders, socioeconomic factors) whereas models of progression to prediabetes included novel factors (high-density lipoprotein, alanine aminotransferase, C-reactive protein, body temperature; AUC = 0.70).

**CONCLUSIONS:** We constructed accurate prediction models from EHR data using a hypothesis-free machine learning approach. Identification of established risk factors for T2D serves as proof of concept for this analytical approach, while novel factors selected by REFS represent emerging areas of T2D research. This methodology has potentially valuable downstream applications to personalized medicine and clinical research.