Predictive Analytics

(.. 30 min Introduction)

Larysa Visengeriyeva
@visenger

The most prominent examples of predictive analyt

http://goo.gl/E7kiI

http://goo.gl/vfqda

Questions we also want to answer from data

fraud detection

ad personalization

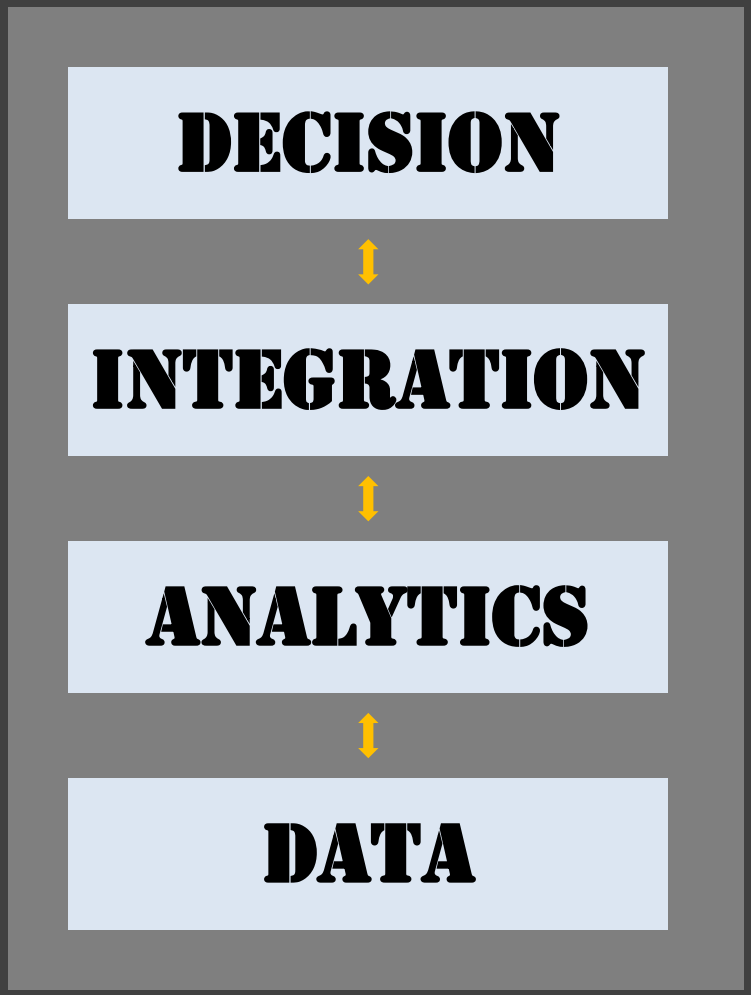
marketing strategy

spam detection

medical treatment

investment

Architecture stack: predictive analytics layer



End-user sees this (web/mobile/desktop app)

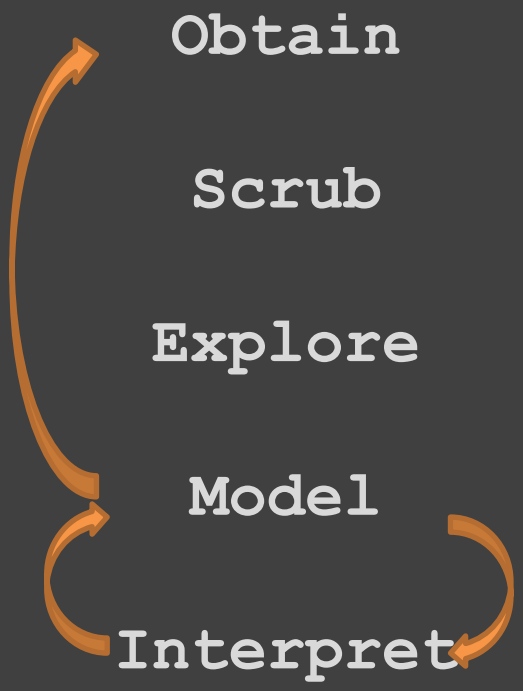
…holds analytics layer and end-user app togeth

…here we do machine learning…
Obtain -> Scrub -> Explore -> **Model** ->…

here is your favorite NoSQL…

//From David Smith's presentation, "Real-Time Big Data Analytics: From Deployment To Production"

Data process. Taxonomy of data science.

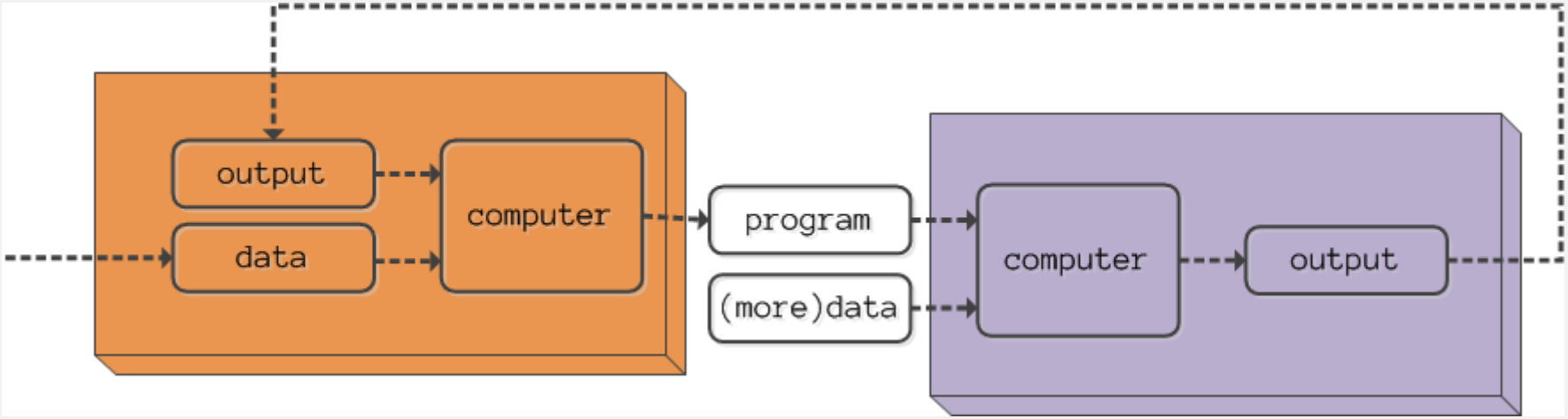


//from dataist.com

Machine Learning (by Tom Mitchell)

"How can we build **computer systems** that **automatically improve** with **experience**, and what
are the fundamental laws that govern all learning processes?"

Machine learning components



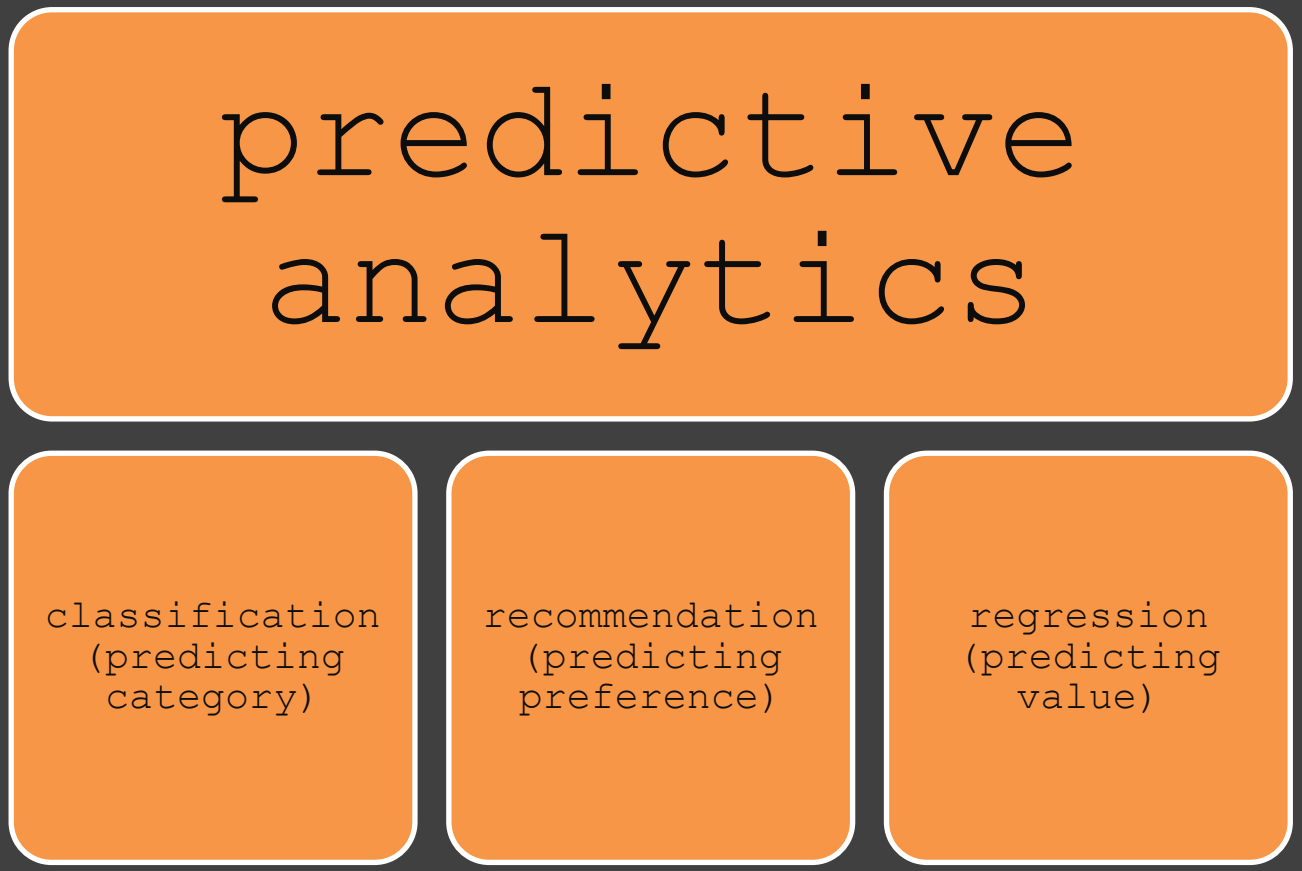
Representation	Evaluation	Optimization
logistic regression	Accuracy	Greedy Search
Naïve Bayes	Posterior probability	Gradient Descent
Decision Trees	Precision/Recall	
Graphical Models		
Instance based		

Types of Machine Learning:

supervised vs unsupervised

classification vs clustering

predictive vs descriptive

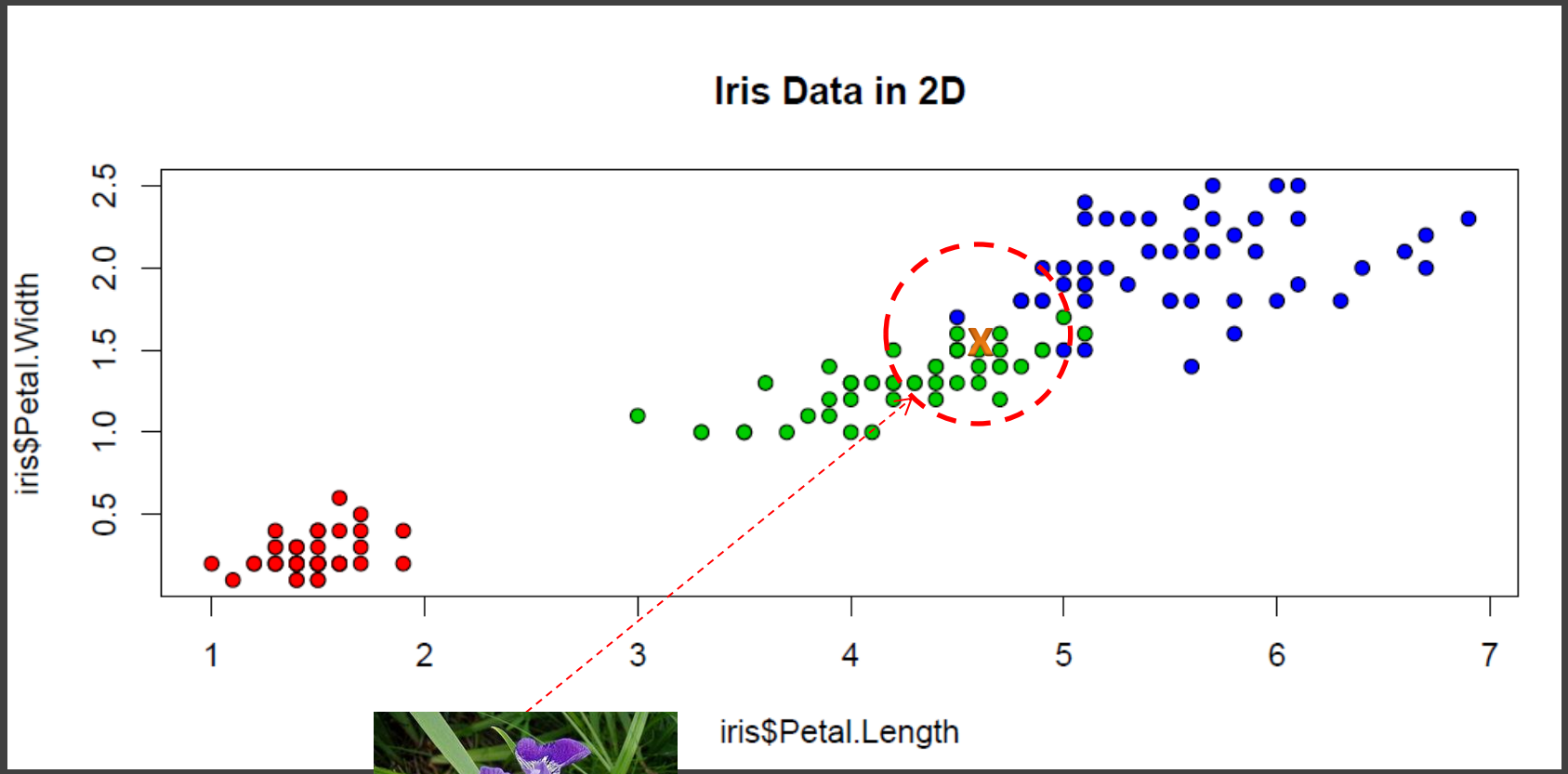

predictive analytics
classification (predicting category)
recommendation (predicting preference)
regression (predicting value)

Naïve Bayes

k-nearest neighbor

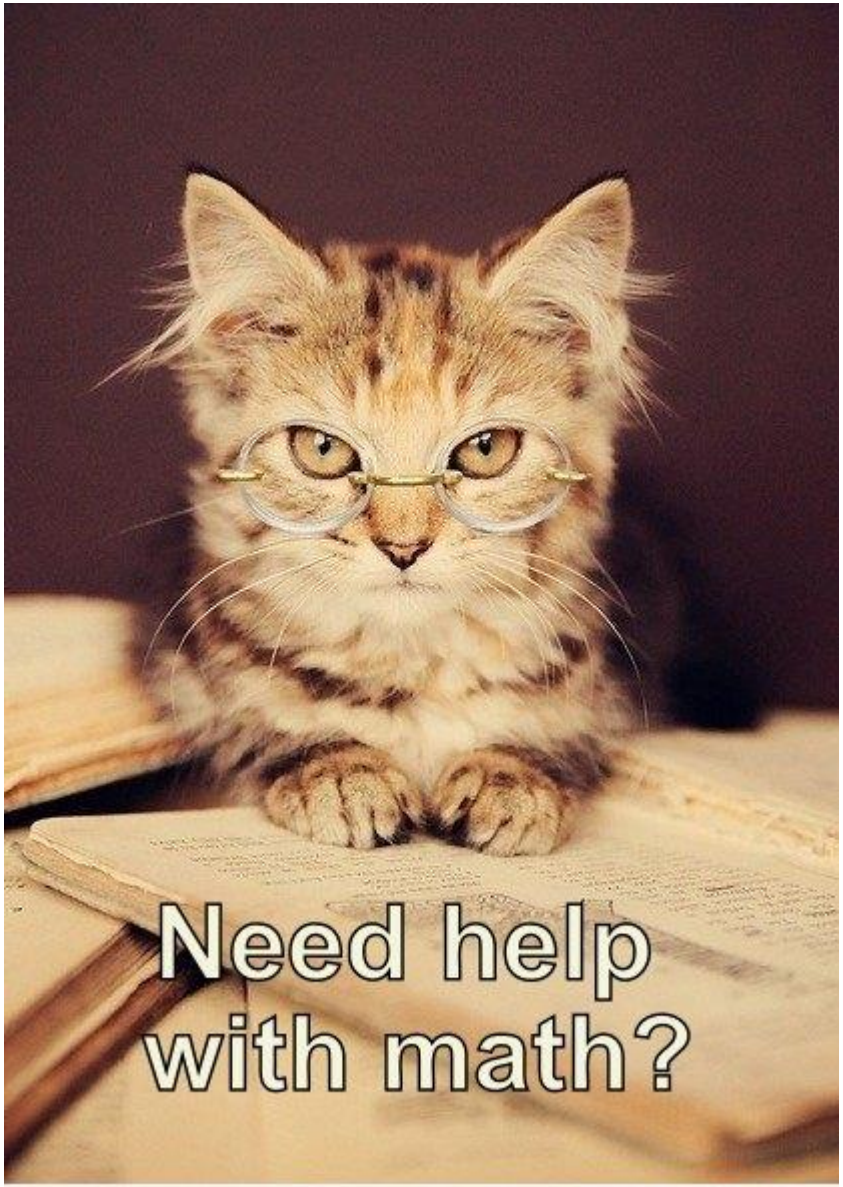
k-nearest neighbor

$$p(y = c \mid x, D, K) = \frac{1}{K} \sum_{i \in K-nearest\ points} I(y_i = c)$$



new point to classify

one cat photo per equation



Naïve Bayes

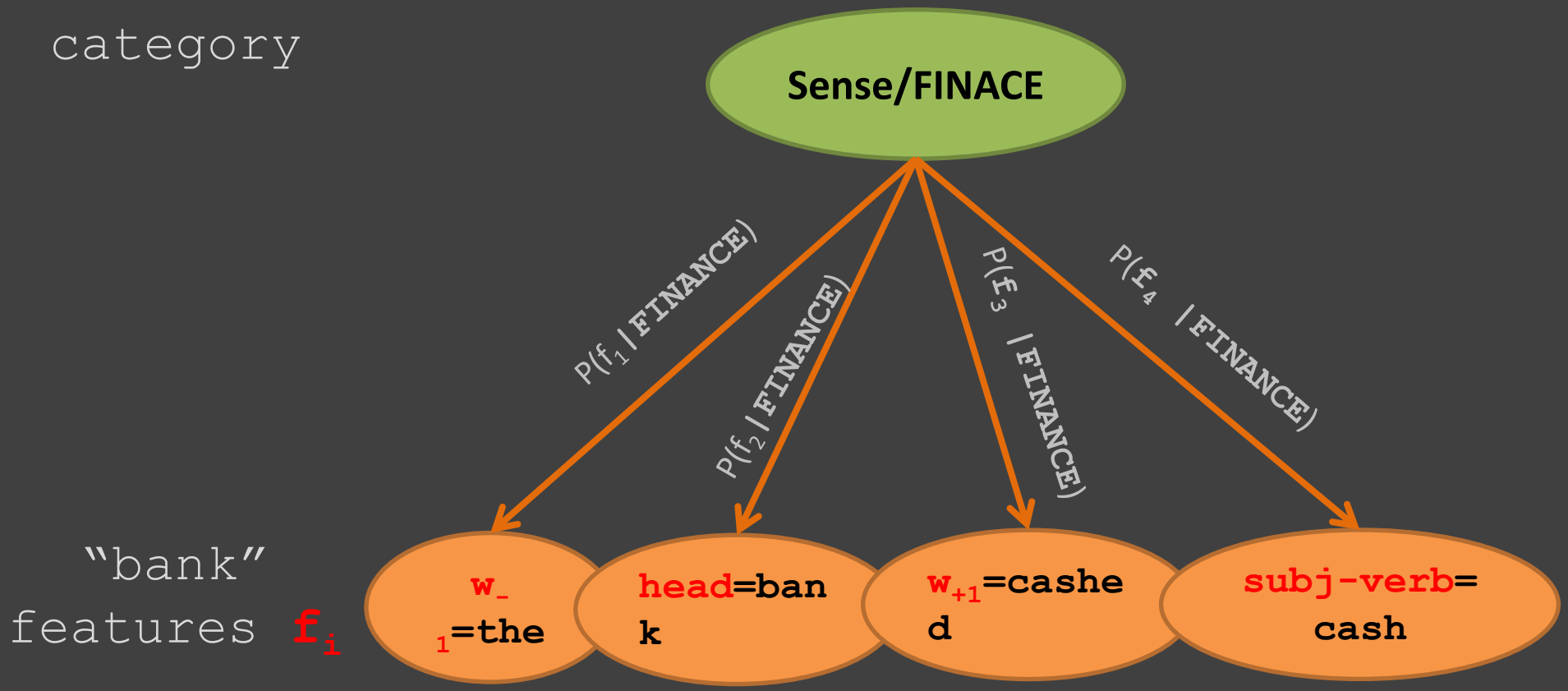
$$P(C|D) = \frac{P(D|C)P(C)}{P(D)}$$

one cat photo per equation

Word Sense Disambiguation with Naïve Bayes

"The bank cashed my check"

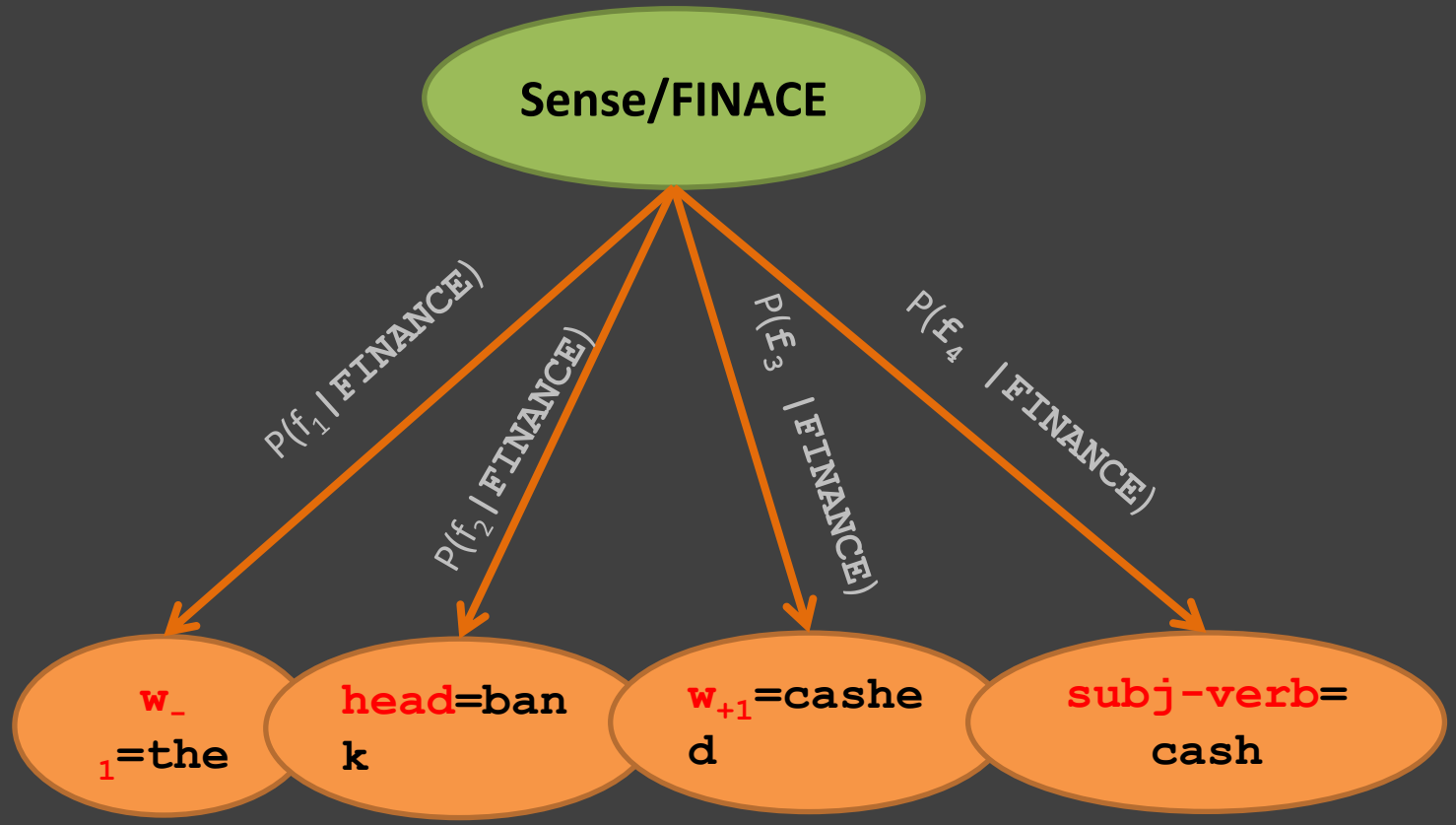
Word Sense Disambiguation with Naïve Bayes



"The bank cashed my check"

"The bank cashed my check"

category



"bank" features f_i

$$classify(\text{"bank"}) = \underset{SENSE}{argmax}\, P(SENSE) * \prod P(f_i|SENSE)$$

No Free Lunch Theorem (Wolpert 1996)

there is no universally best model

different domains require different model

rial-and-Error” approach for your specific domain

Available Tools (…just a few…)

Scrape (ETL)	OpenRefine ScraperWiki
Processing	R Lucene/ElasticSearch Mechanical Turk
NLP	StanfordNLP NLTK OpenNLP
Machine Learning	R scikit-learn WEKA Mahout
MapReduce	Pig Hive Cascalog

1 Book to buy

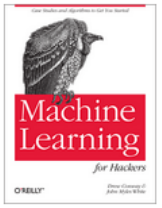
O'REILLY
Strata
Making Data Work

Data Science Starter Kit

The Tools You Need to Get Started with Data

Python for Data Analysis: is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you'll need to effectively solve a broad set of data analysis problems. This book is not an exposition on analytical methods using Python as the implementation language.

Ebook: $31.99 Add to Cart

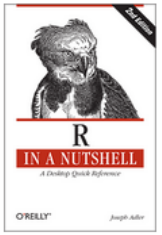


Machine Learning for Hackers: If you're an experienced programmer interested in crunching data, this book will get you started with machine learning—a toolkit of algorithms that enables computers to train themselves to automate useful tasks. Each chapter focuses on a specific problem in machine learning, such as classification, prediction, optimization, and recommendation.

Ebook: $31.99 Add to Cart

R Cookbook: Over 200 recipes for R users, ranging from the basic to the esoteric. Why re-invent the wheel? This collection of concise, task-oriented recipes makes you productive with R immediately, with solutions ranging from basic tasks to input and output, general statistics, graphics, and linear regression.

Ebook: $31.99 Add to Cart



R in a Nutshell, 2nd Edition: The authoritative guide to what's become the de-facto standard for statistical programming. R in a Nutshell provides a quick and practical way to learn this increasingly popular open source language and environment.

Ebook: $35.99 Add to Cart

Bad Data Handbook: What is bad data? Some people consider it a technical phenomenon, like missing values or malformed records, but bad data includes a lot more. In this handbook, data expert Q. Ethan McCallum has gathered 19 colleagues from every corner of the data arena to reveal how they've recovered from nasty data problems.

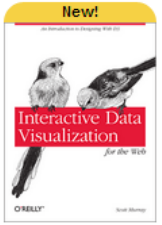
Ebook: $31.99 Add to Cart

Data Analysis with Open Source Tools: A survey of data analysis from a practitioner – from histograms to machine learning,this book presents the tools you need to make sense with data. You'll learn how to look at data to discover what it contains, how to capture those ideas in conceptual models, and then feed your understanding back into the organization through business plans, metrics dashboards, and other applications.

Ebook: $31.99 Add to Cart

MapReduce Design Patterns: Each pattern is explained in context, with pitfalls and caveats clearly identified to help you avoid common design mistakes when modeling your big data architecture. This book also provides a complete overview of MapReduce that explains its origins and implementations, and why design patterns are so important. All code examples are written for Hadoop.

Ebook: $39.99 Add to Cart



Interactive Data Visualization for the Web: Create and publish your own interactive data visualization projects on the Web, even if you have no experience with either web development or data visualization. It's easy with this hands-on guide. You'll start with an overview of data visualization concepts and simple web technologies, and then learn how to use D3, a JavaScript library that lets you express data as visual elements in a web page.

Ebook: $23.99 Add to Cart

3 Blogs to read

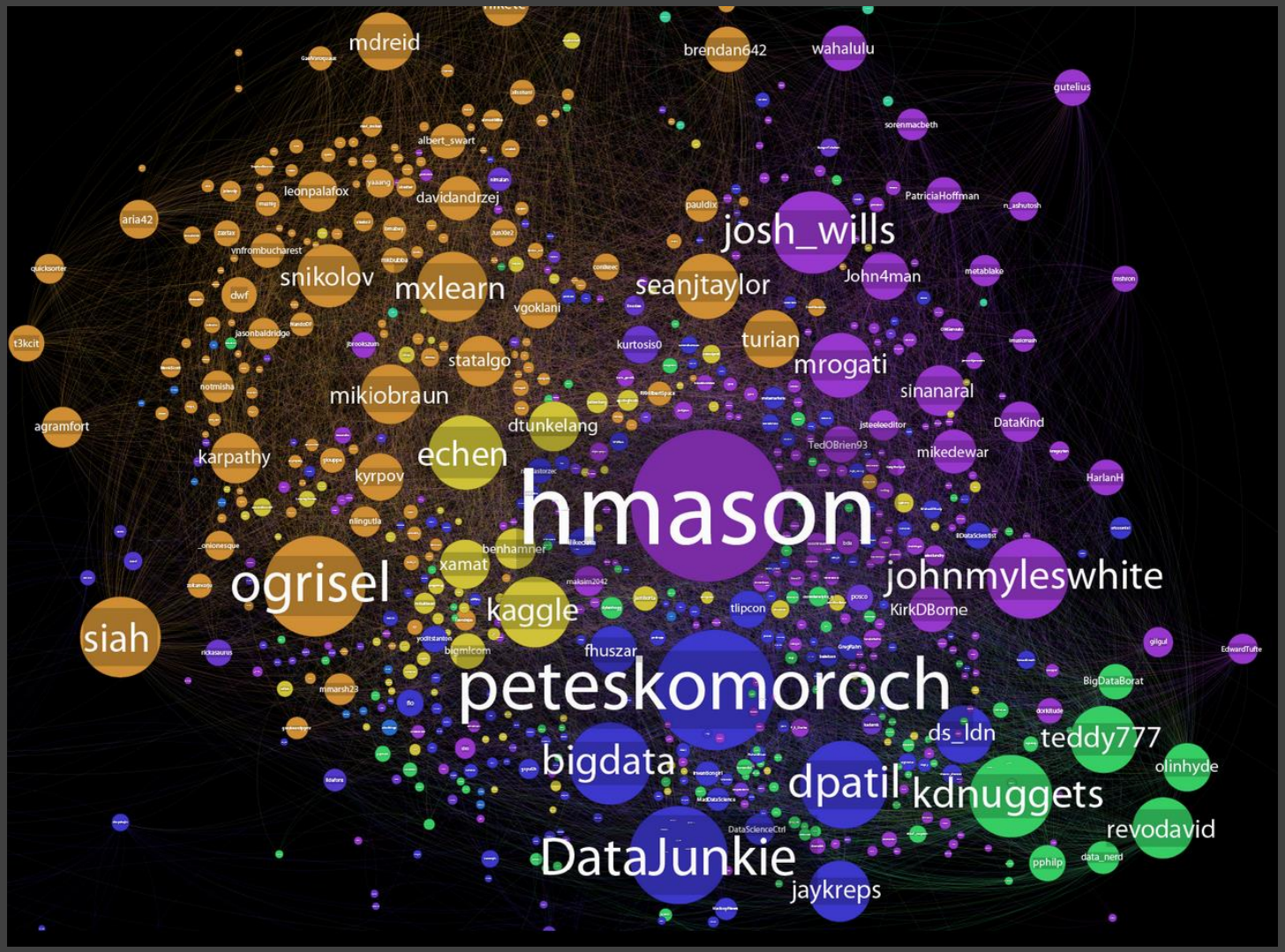
Data Science Blogs: bitly.com/bundles/hmason/d

Strata O'reilly: strata.oreilly.com

Kaggle: blog.kaggle.com

……. (there are a lot of them)

5 People to follow



// from giladlotan.com/blog/mapping-twitters-python-data-science-communities/

Naïve Bayes

$$P(C \mid \{x_1, x_2, x_3, \ldots x_n\}) = \frac{P(\{x_1, x_2, x_3, \ldots x_n\} \mid C) P(C)}{P(\{x_1, x_2, x_3, \ldots x_n\})}$$

$$P(C \mid \{x_1, x_2, x_3, \ldots x_n\}) = \frac{P(C) \prod P(\{x_i\} \mid C)}{P(\{x_1, x_2, x_3, \ldots x_n\})}$$

$$P(C \mid \{x_1, x_2, x_3, \ldots x_n\}) \propto P(C) \prod P(\{x_i\} \mid C)$$

$$\hat{C} = \underset{C}{argmax}\, P(C) \prod P(\{x_i\} \mid C)$$