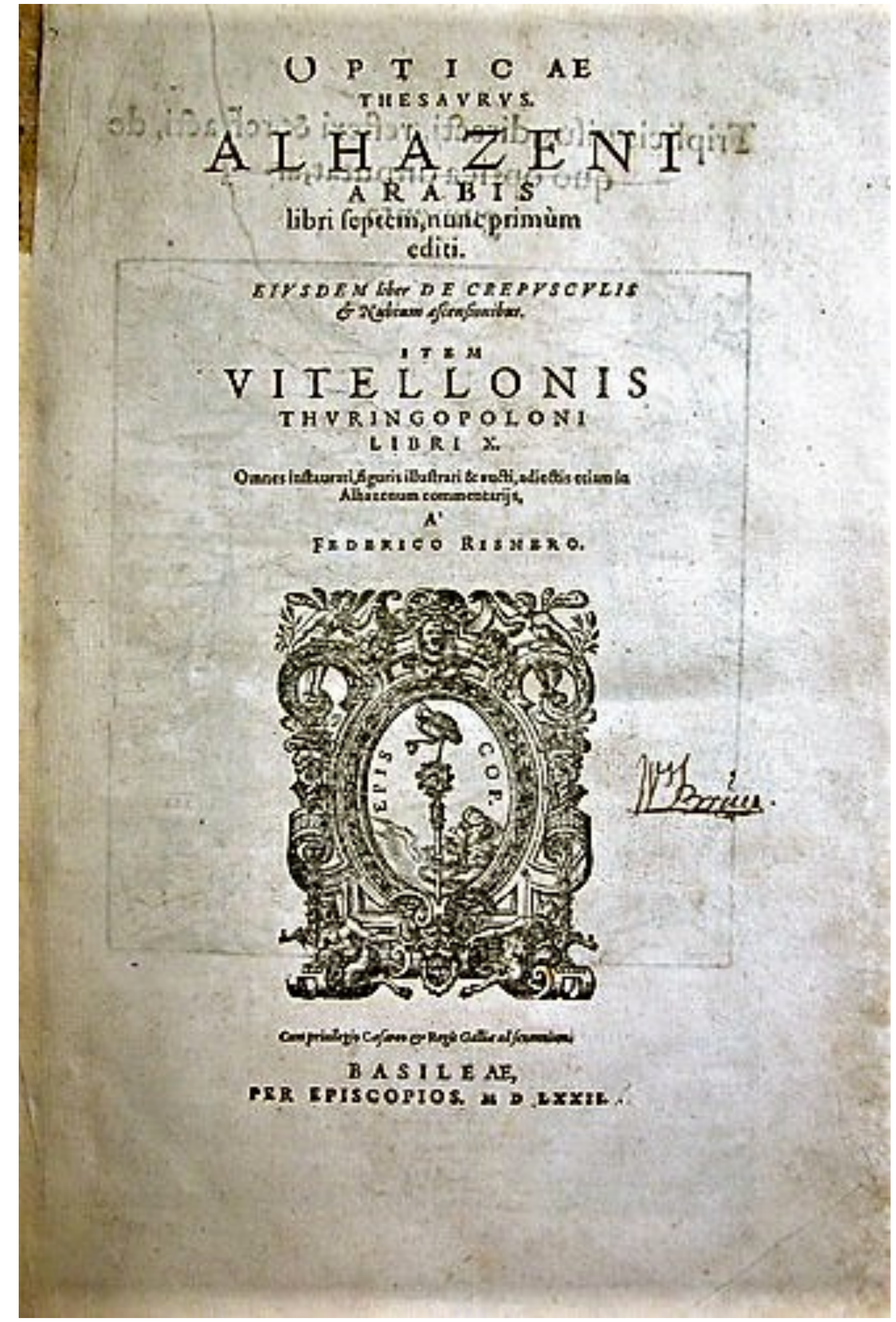
A Millennium of Nearest Neighbor Methods

An Introduction to the NIPS Nearest Neighbor Workshop 2017

George Chen
Assistant Professor of Information Systems
Heinz College of Information Systems and Public Policy
Carnegie Mellon University

December 8, 2017

Alhazen
(~965 to ~1040)



OPTICAE
THESAVRVS.
ALHAZENI
ARABIS
libri septem, nunc primùm
editi.
EIVSDEM liber DE CREPVSCVLIS
& Nubium ascensionibus.
ITEM
VITELLONIS
THVRINGOPOLONI
LIBRI X.
Omnes instaurati, figuris illustrati & aucti, adiectis etiam in
Alhazenum commentarijs,
A'
FEDERICO RISNERO.
EPIS COP.
BASILEAE,
PER EPISCOPIOS. M D LXXII.

Book of Optics
(~1010's to ~1030's*)

Historians give different dates…

Hence, when sight perceives some visible object, the faculty of discrimination immediately seeks its counterpart among the forms persisting in the imagination, and when it finds some form in the imagination that is like the form of that visible object, it will recognize that visible object and will perceive what kind of object it is

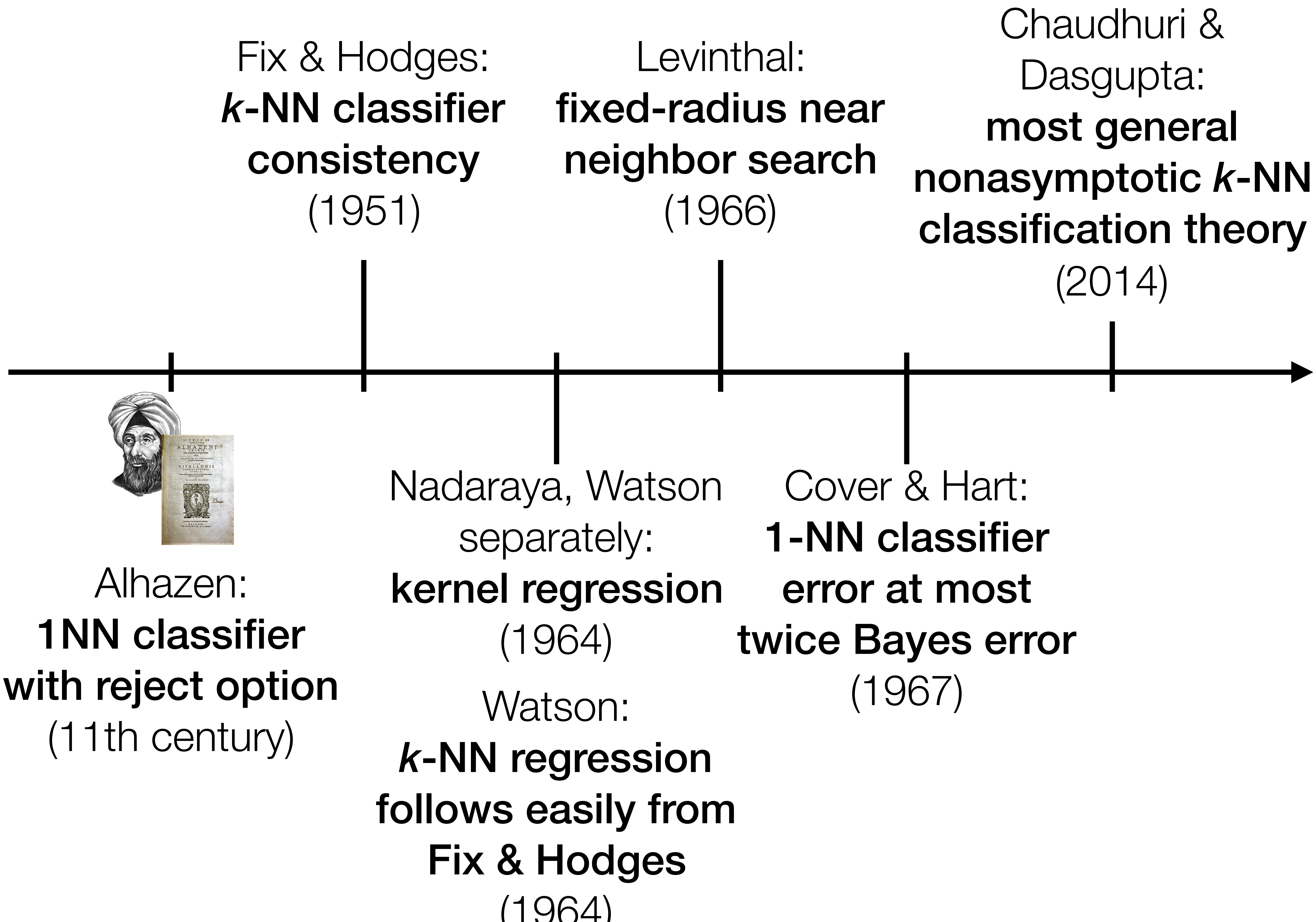
–*Alhazen (translated; Smith 2001)*

Reading further, Alhazen allows for a "reject" option if none of the training data are sufficiently similar

Alhazen
(~965 to ~1040)

(~1010's to ~1030's*)

Historians give different dates…


Alhazen:
1NN classifier with reject option
(11th century)
Fix & Hodges:
***k*-NN classifier consistency**
(1951)
Nadaraya, Watson separately:
kernel regression
(1964)
Watson:
***k*-NN regression follows easily from Fix & Hodges**
(1964)
Levinthal:
fixed-radius near neighbor search
(1966)
Cover & Hart:
1-NN classifier error at most twice Bayes error
(1967)
Chaudhuri & Dasgupta:
most general nonasymptotic *k*-NN classification theory
(2014)

k-NN Classification Guarantee

Theorem (Chaudhuri & Dasgupta 2014, informal)

Choose error tolerance $\delta \in (0, 1)$. # training data: *n*

With probability at least $1 - \delta$ over randomness in training data,

$$\mathbb{P}\big(k\text{-NN classifier}(X) \neq \text{Bayes classifier}(X)\big)$$
$$\leq \delta + \mathbb{P}(X \text{ lands near decision boundary})$$

over randomness in feature vector *X*

can make this precise

- Subsumes previous asymptotic consistency results (Fix & Hodges 1951, Stone 1977, Devroye et al 1994)
- Achieves error lower bound under smoothness, margin condition (Audibert & Tsybakov 2007)

Proof ideas translate to **fixed-radius NN classification**

Proof ideas extend to *k*-NN, fixed-radius NN, and kernel **regression** to match existing sample complexity results (up to 1 log factor) (Chen & Shah 2018)

How does a practitioner make use of these results?

Example: in health care, number of training data often refers to number of human subjects in a study

Clinician would benefit from relating this number to specific disease or treatment statistics

(e.g., prob. of landing near decision boundary & Lipschitz constants don't readily lend to clinical interpretation)

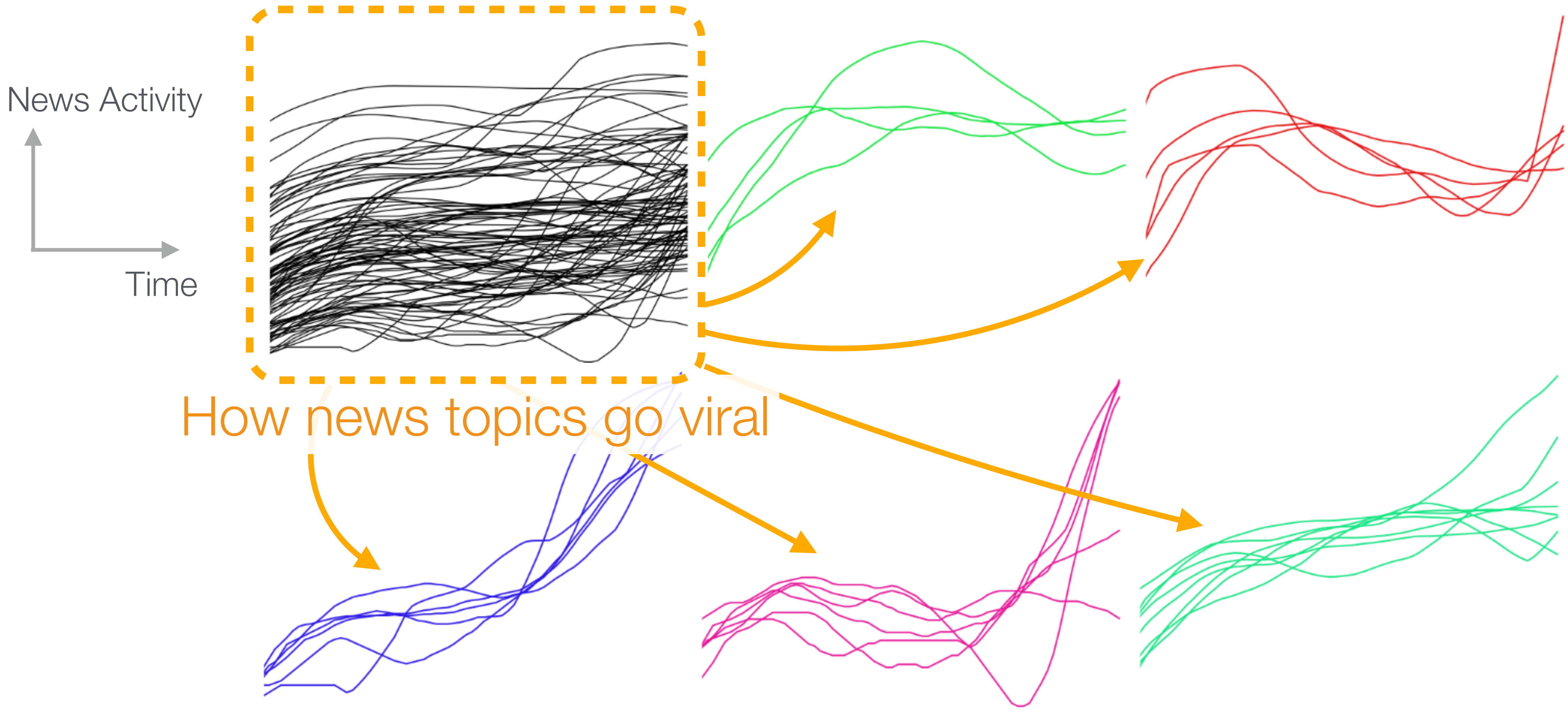
Toward more “user-friendly” guarantees: clustering structure

Application-specific clusters can be easier for practitioners to interpret

Time Series Forecasting

(Chen, Nikolov, Shah 2013)

Will a news topic on Twitter go viral?



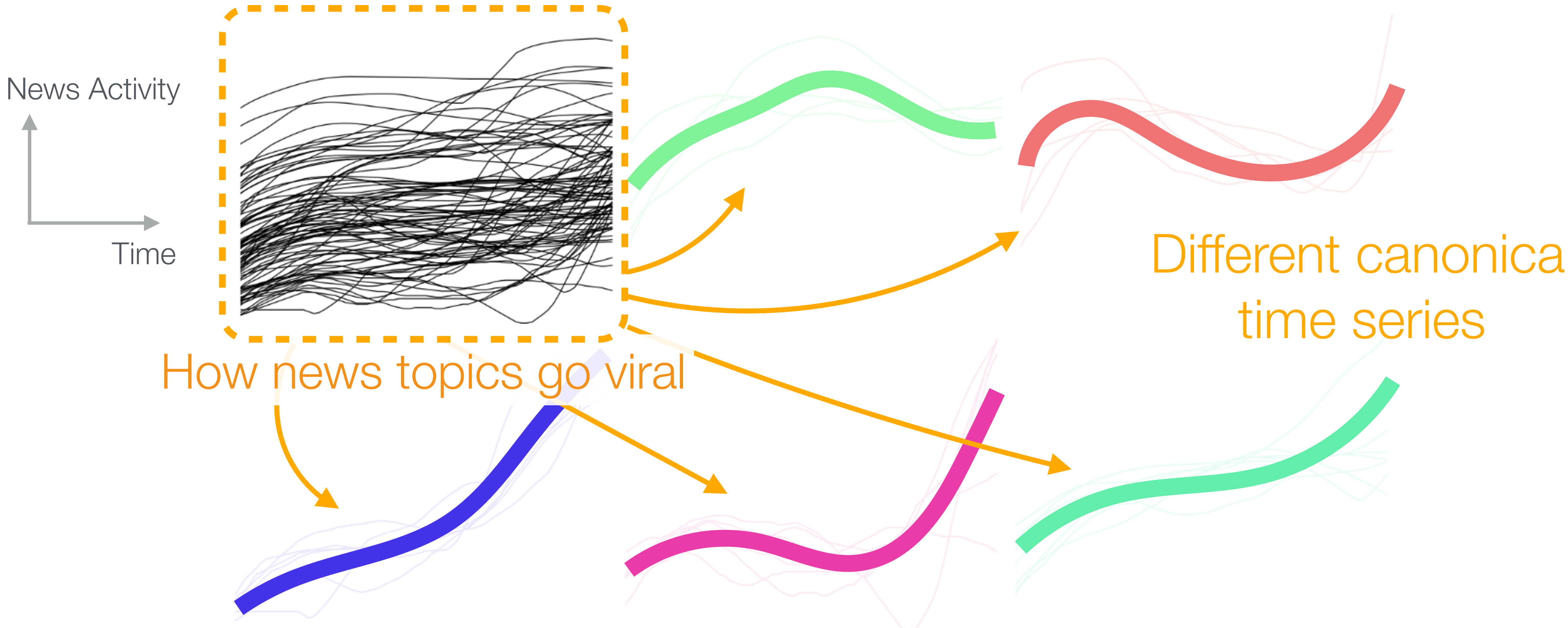
June 2012 Tweet rate data from Twitter

Time Series Forecasting

(Chen, Nikolov, Shah 2013)

Will a news topic on Twitter go viral?

Observation: how news topics go viral forms clusters



June 2012 Tweet rate data from Twitter

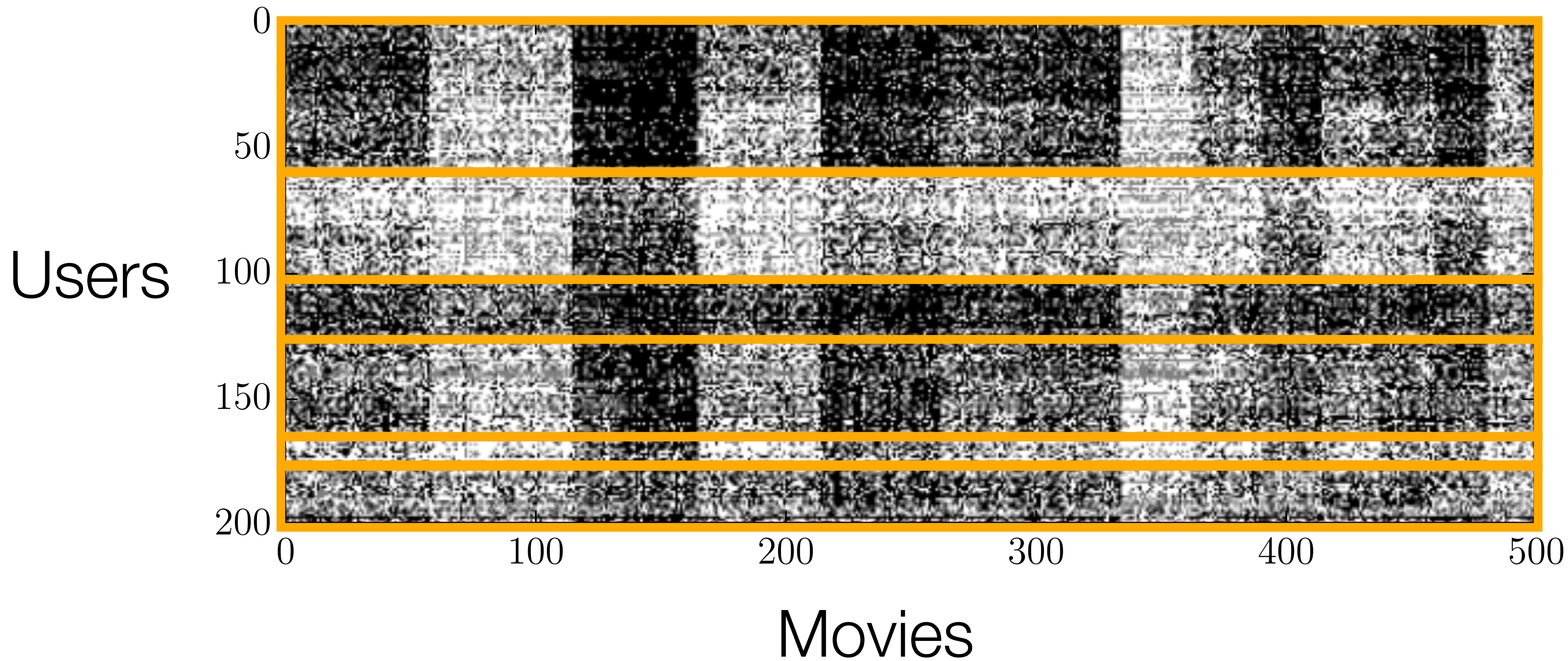
Online Collaborative Filtering

(Bresler, Chen, Shah 2014)

How do we recommend items to users over time?

Observation: user ratings cluster into groups

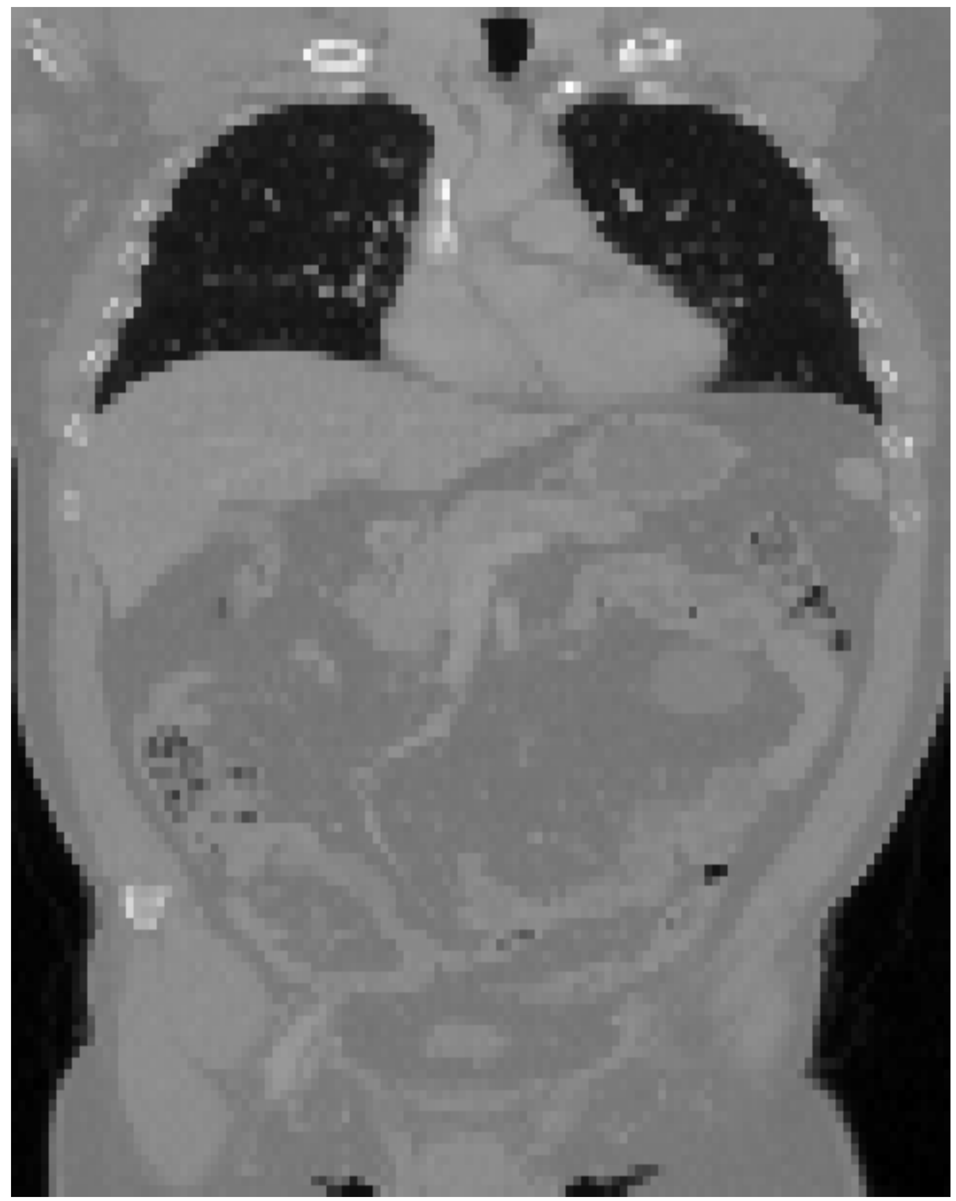
(item ratings do as well!)



Movielens10m

Patch-Based Image Segmentation

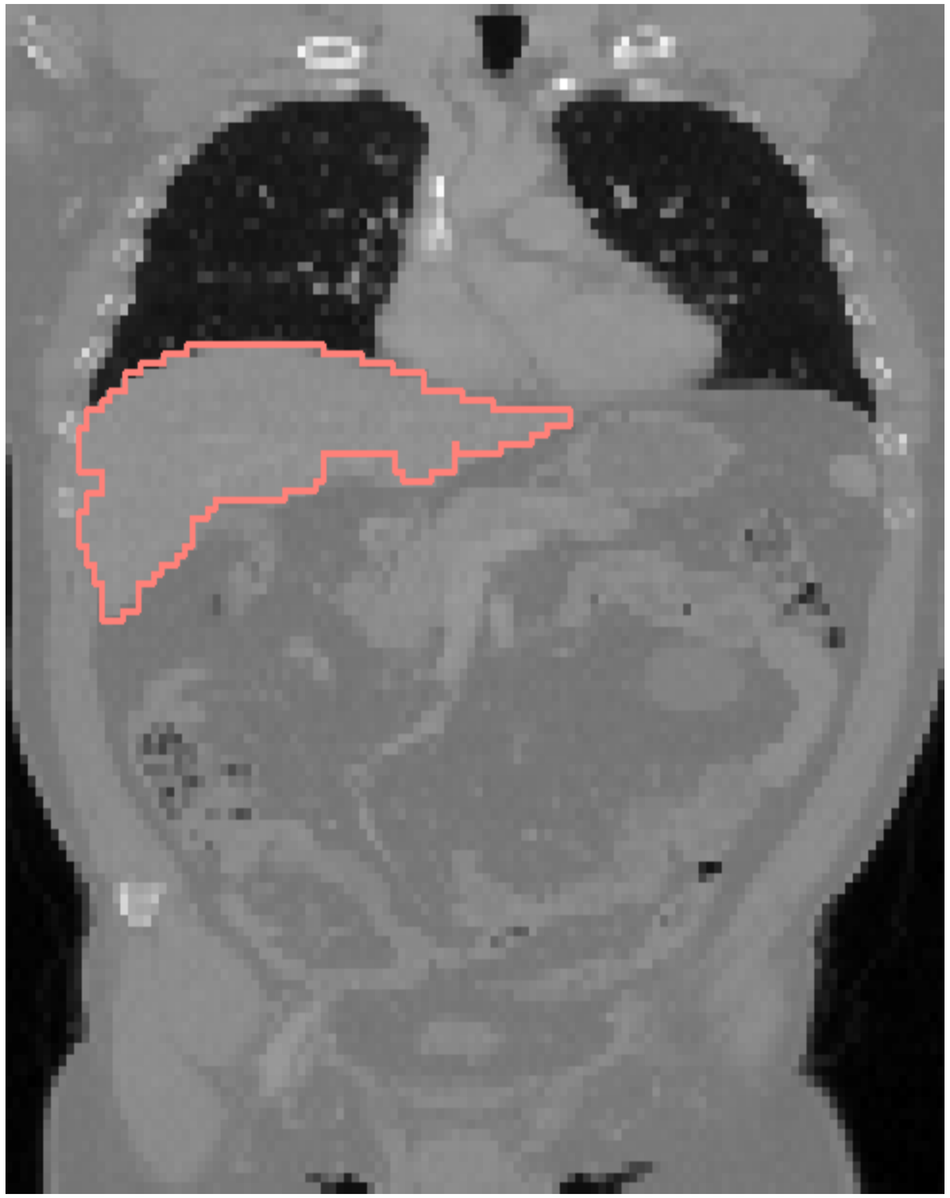
(Chen, Shah, Golland 2015)



Where is the liver?

Patch-Based Image Segmentation

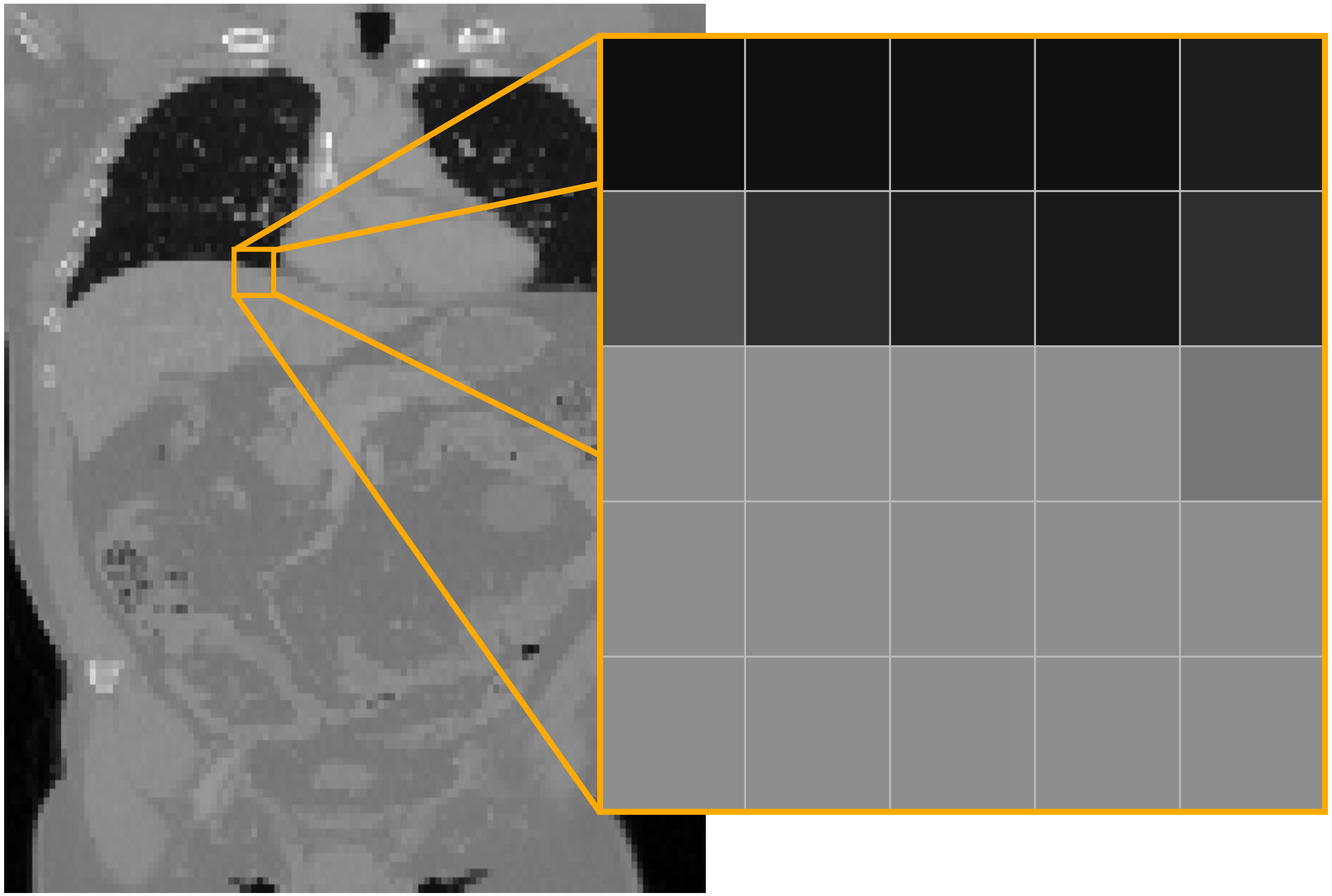
(Chen, Shah, Golland 2015)



Where is the liver?

Patch-Based Image Segmentation

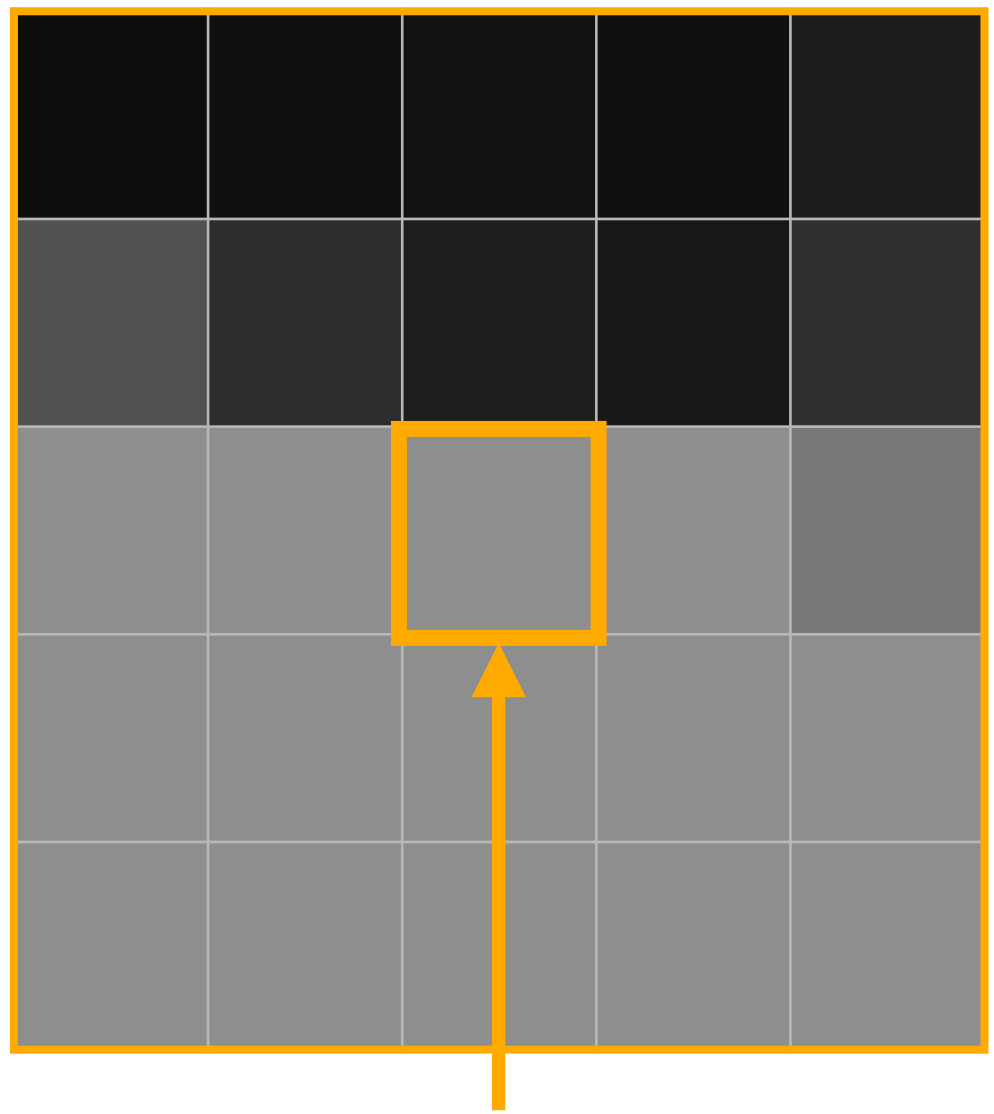
(Chen, Shah, Golland 2015)



Where is the liver?

Patch-Based Image Segmentation

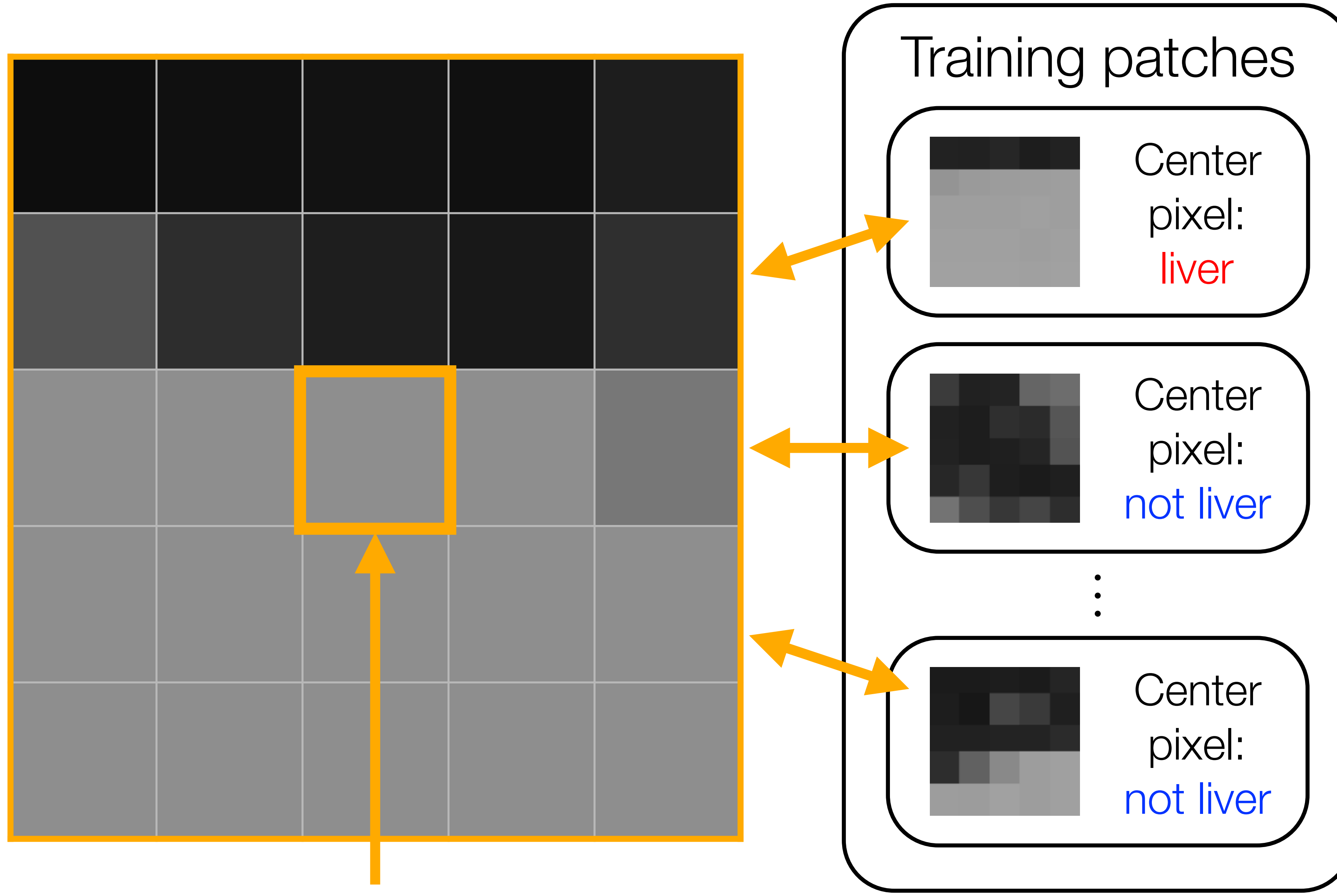
(Chen, Shah, Golland 2015)



Is this pixel part of the liver or not?

Patch-Based Image Segmentation

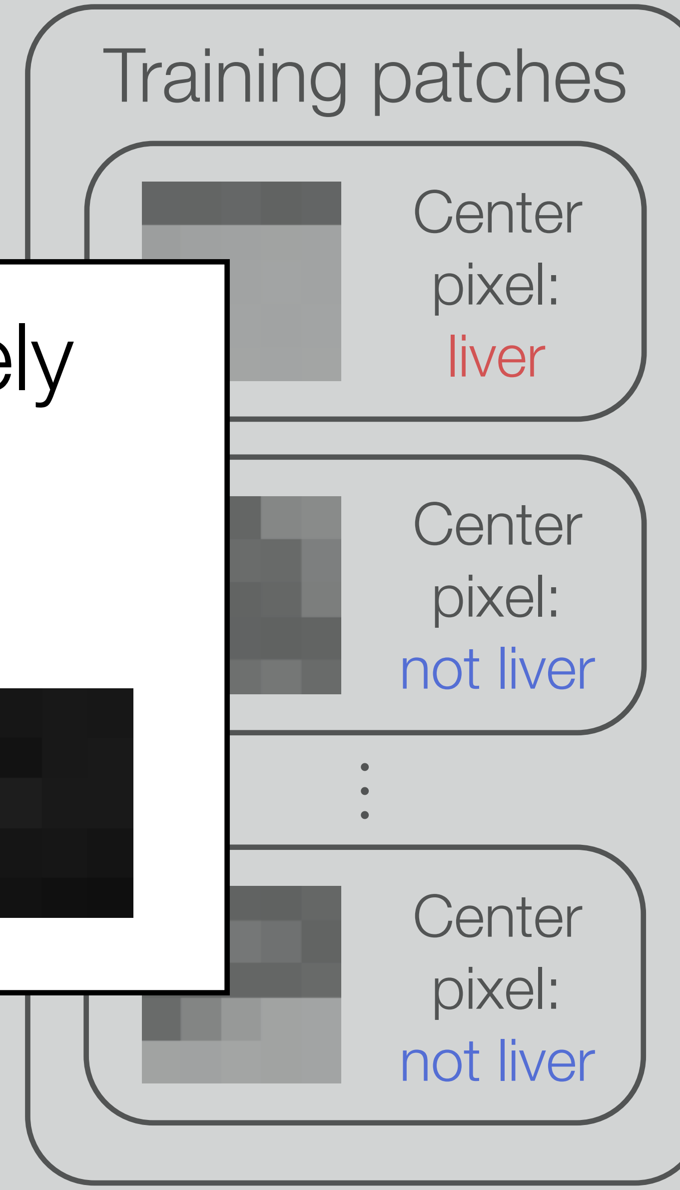
(Chen, Shah, Golland 2015)



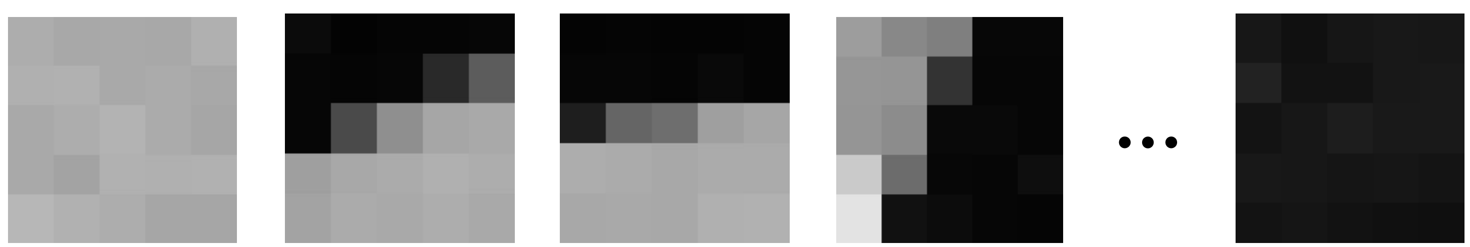
Is this pixel part of the liver or not?

Patch-Based Image Segmentation

(Chen, Shah, Golland 2015)



Patches in real images accurately modeled by a few clusters (Zoran & Weiss, 2011, 2012)



Is this pixel part of the liver or not?

Clustering Structure Enables Successful NN Prediction

Collect enough training data to see enough points from all clusters

Clusters should be sufficiently separated to combat noise

Time Series Forecasting

i.i.d. training data

distance not a metric

Chen, Nikolov, Shah 2013

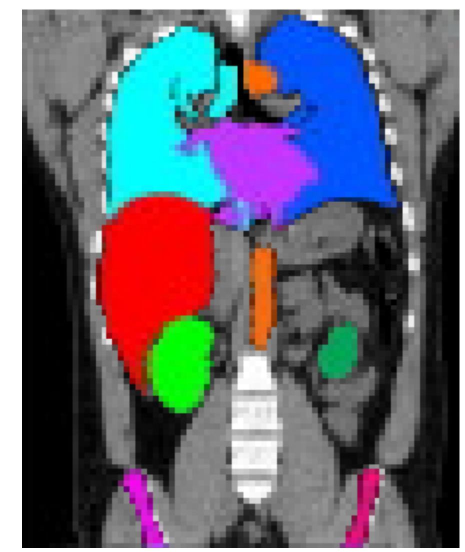
Online Collaborative Filtering

choice of training data

distance is noisy

Bresler, Chen, Shah 2014

prediction over time



Patch-Based Image Segmentation

non i.i.d. training data

prediction over space

Chen, Shah, Golland 2015

Foundations and Trends® in Machine Learning
Vol. XX, No. XX (2017) 1–218

DOI: 10.1561/XXXXXXXXXX

Explaining the Success of Nearest Neighbor Methods in Prediction

George H. Chen
Carnegie Mellon University
georgechen@cmu.edu

Devavrat Shah
MIT
devavrat@mit.edu

Coming soon!

Foundations and Trends® in Machine Learning
Vol. XX, No. XX (2017) 1–218

DOI: 10.1561/XXXXXXXXXX

Monograph coverage

- Theory on NN regression
- Theory on NN classification
- Prediction guarantees in three modern applications using clustering structure
- (Approximate) NN search algorithms
- Decision trees and their ensemble variants are adaptive NN methods

Coming soon!

Hot off the press: biological evidence of fruit fly brains doing LSH!

Science Home News Journals Topics Careers

Institution: CARN
Log in | My acco
CARNEGIE

SHARE

0

0

REPORT

A neural algorithm for a fundamental computing problem

Sanjoy Dasgupta[1], **Charles F. Stevens**[2,3], **Saket Navlakha**[4,*]

\+ See all authors and affiliations

Science 10 Nov 2017:
Vol. 358, Issue 6364, pp. 793-796
DOI: 10.1126/science.aam9868

Article Figures & Data Info & Metrics eLetters PDF

Fly brain inspires computing algorithm

Flies use an algorithmic neuronal strategy to sense and categorize odors. Dasgupta *et al.*

Amazing lineup of speakers, applied and theoretical!