KAMSC Big Data Day - Keynote
Principles and Some Applications of Big Data

Kwame Porter Robinson, PhD Candidate
University of Michigan, School of Information

1.13.2022

About Me

2009 - 2015

2015 - 2019

About Me

2019 - 2019

2019 - 2024

Counting Activity: Warm Up

How many jellybeans are on this slide?

[Number fingers held up x 1 = Your Guess]

Counting Activity

How many jellybeans are on this slide?

[Number fingers held up x 1 = Your Guess]

Counting Activity

How many jellybeans are on this slide?
[Number fingers held up x 5 = Your Guess]

Counting Activity: Reflection

Wait! What were we looking into?
→*What do you pay attention to?*

How do you count the partial jelly bean?
→*Messy Data*

How many jelly beans were there?
[It Depends.]

Jelly bean counting problems are also big data problems.

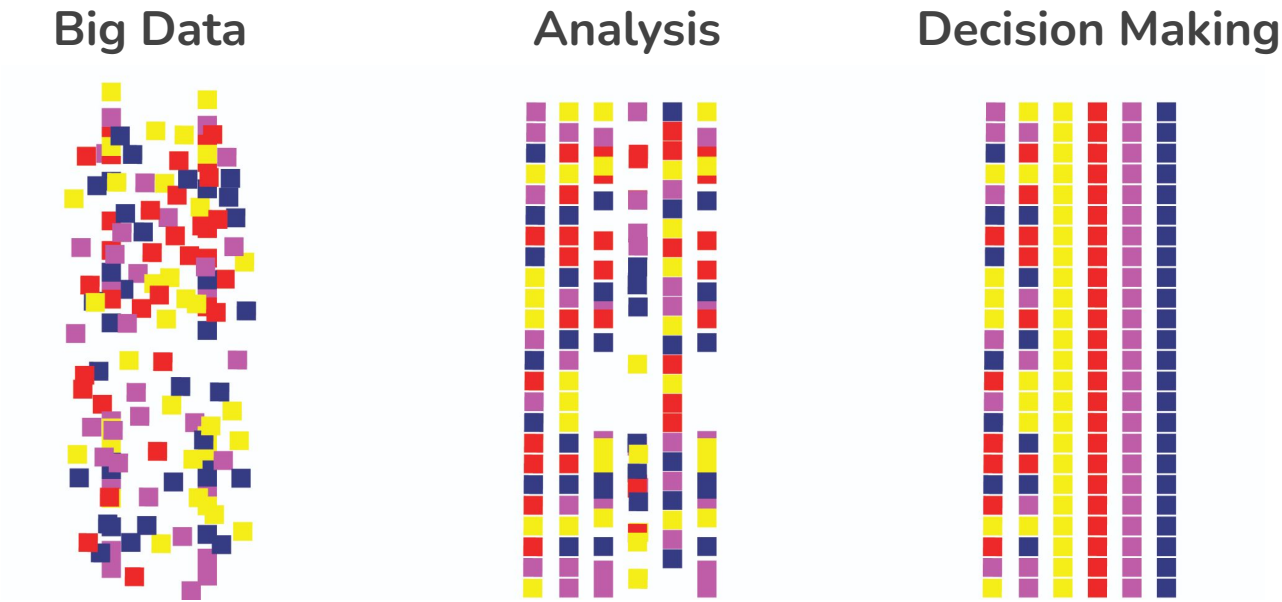


image credit hart energy

Big Data: Questions and Principles

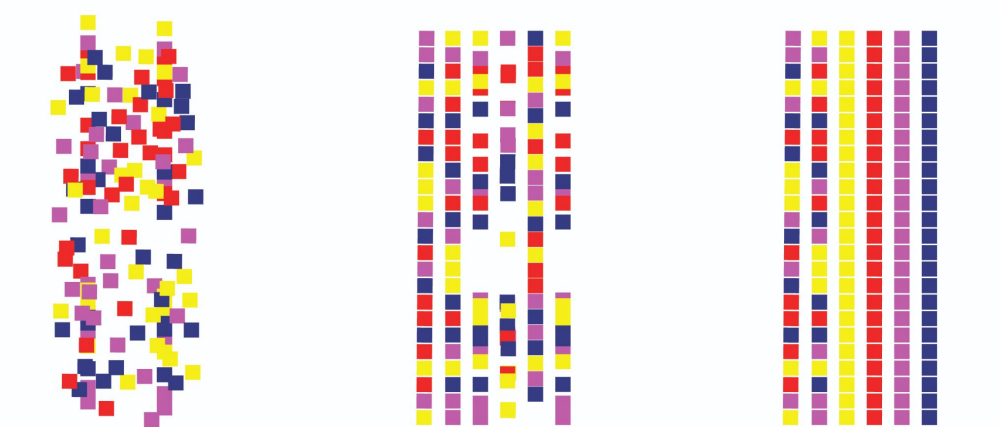
→ What do you pay attention to?

→ What could be noisy data?

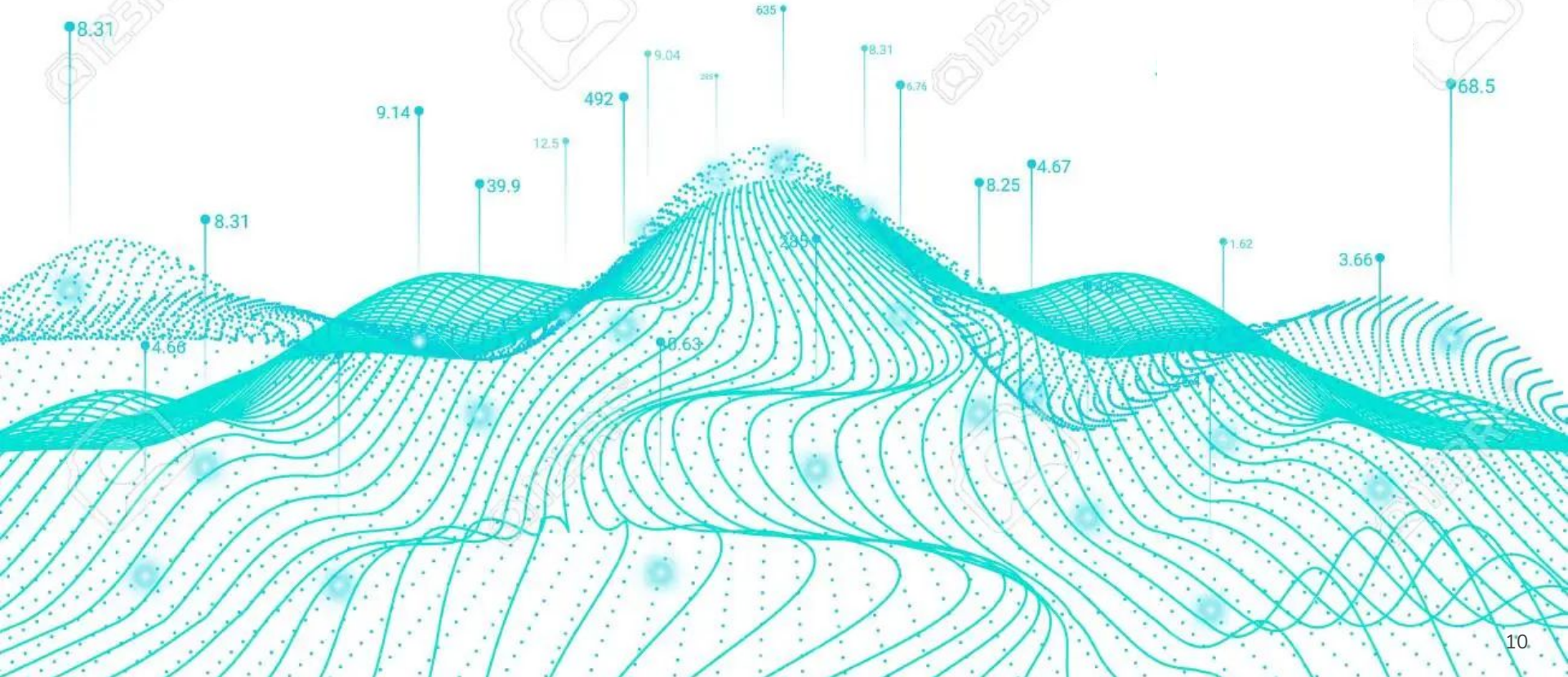
→ How do you separate the wheat from the chaff?

→ Where does messy data come from and how do we handle it?

→ Data engineering and data cleaning



Big Data: Questions and Principles



Big Data: Questions and Principles

Q 0: What is data?

A: Data is **qualitative or quantitative measurement** of something while it is or is not happening → big data is many many many measurements of one or more **phenomena**

Qualitative data: objective measurement of **subjective phenomena**. ← What does the weather *feel* like? Balmy!

Quantitative data: objective measurement of **objective phenomena**. ← In celsius, what is the weather? 72 C.

Big Data: Questions and Principles

Q 1: What do you pay attention to?

A: First you need a theory or a model. Theory describes what we do and do not know about a phenomena. Models are often simplifications generated from complex theory.

Theory can tell us when something is:

- expected,
- unexpected,
- and even impossible!

Without theory the best you can do is look for correlations. But correlations can lead you astray. But they can also lead to new discoveries.

Big Data: Questions and Principles

Q 2: What could be noisy data?

A: If you have a **model** of how the data is generated you can identify outliers as extreme values.

If you have a **theory** you can anticipate under what conditions noisy data would happen. Look back to models and theory for greater understanding. **Noisy data is relative to theory and models, often described as random variables.**

Don't have one? That's new science! Hypothesis testing, etc.

Big Data: Questions and Principles

Q 3: Separating the wheat from the chaff

A: This is problem specific: chaff for one problem might be wheat in another problem.

For example, coffee grinds can be great in an agricultural science problem but waste in a coffee business problem.

Think about how the nature of your problem **makes some things less valuable (or even invisible) and makes other things more valuable.**

Big Data: Questions and Principles

Q 4: Messy data, where does it come from? How do we handle it?

A: Messy data comes about in many different ways. It boils down to imperfect measurement. For example:

- A sensor has limited precision, runs out of disk space
- A selfie has a giant thumb over it

Messy data is caused by imperfect measurement.

Big Data: Questions and Principles

Q 5: What is Data Engineering?

A: The bulk of the work in big data is actually in data engineering and less so for data science. It takes a **LOT** of work to develop code, work with sensors, push improvements, fix bugs, and so on and so forth.

Yes, there is a lot of data science and discovery too. But the way there is to work with and manipulate big data.

Some things I use in my day to day: Python, Pandas, R, and SQL.

Big Data Authentication Problem: Authente-Kente

Chinese counterfeits leave Ghanaian textiles hanging by a thread

Traditional clothing makers in Ghana are turning to technology to fight popular Chinese counterfeits. But little of the manufactures left are local, or even Ghanaian-owned.

The Printex factory in Accra, Ghana creates the popular wax-printed fabric. The company has survived largely because of the Ghanaian tradition of commissioning specially printed fabric for everything from funerals to casual Fridays, which makes up 70 percent of their sales.

May 31, 2015

By Yepoka Yeebo, Contributor

CS Monitor, Yepoka Yeebo (2015)


VOGUE
MAY
THE RISE
AND RISE OF
AMANDA
GORMAN
POET, ACTIVIST,
PHENOMENON
FASHION,
WE MISSED YOU
SPRING'S BIG MOVES—
BOLD, BEAUTIFUL, AND FUN
KING GIORGIO
MR. ARMANI ON FASHION'S FUTURE
(AND WHY HE'S NOT SLOWING DOWN)

Real or Fake?


VOGUE
MAY
THE RISE AND RISE OF
AMANDA GORMAN
POET, ACTIVIST, PHENOMENON
FASHION, WE MISSED YOU
SPRING'S BIG MOVES—BOLD, BEAUTIFUL, AND FUN
KING GIORGIO
MR. ARMANI ON FASHION'S FUTURE (AND WHY HE'S NOT SLOWING DOWN)

Real

Fake

Big Data Authentication Problem: Authente-Kente

SpringerLink

OPEN FORUM | Published: 02 September 2020

Authente-Kente: enabling authentication for artisanal economies with deep learning

Kwame Porter Robinson ✉, Ron Eglash, Audrey Bennett, Sansitha Nandakumar & Lionel Robert

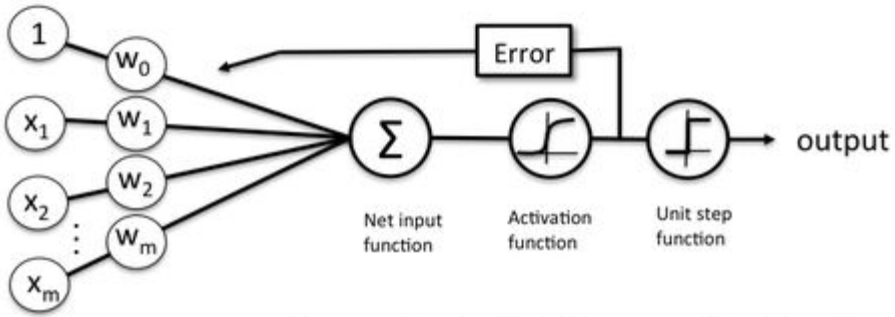
AI & SOCIETY **36**, 369–379 (2021) | Cite this article

Abstract

The economy for artisanal products, such as Navajo rugs or Pashmina shawls are often threatened by mass-produced fakes. We propose the use of AI-based authentication as one

Big Data Authentication Problem: Authente-Kente

What do you pay attention to?

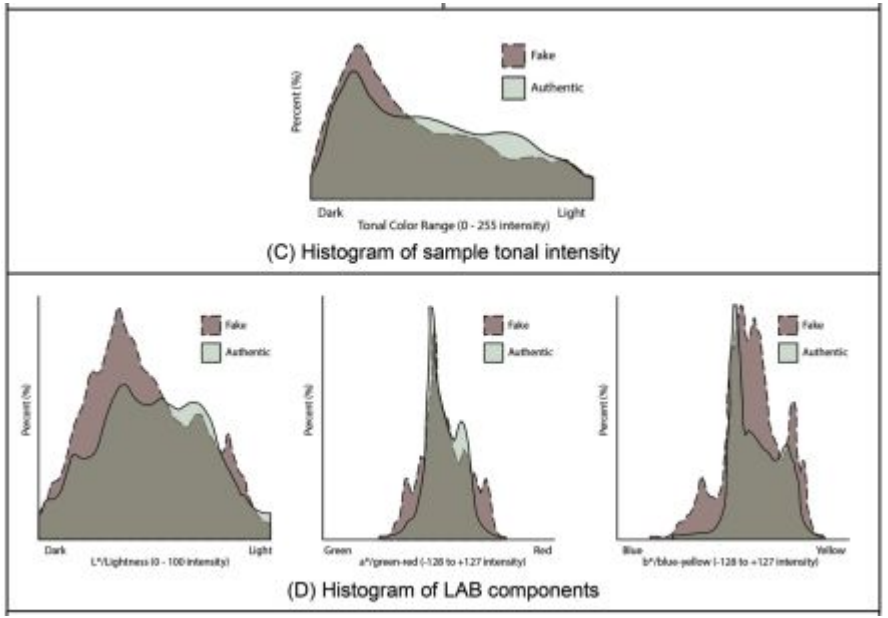


Schematic of a logistic regression classifier.

image credit Sebastian Raschka

Big Data Authentication Problem: Authente-Kente

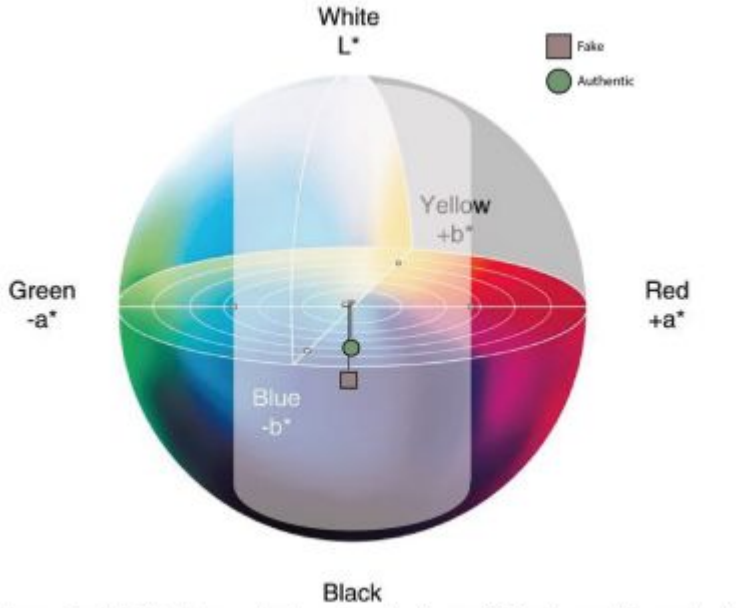
What is and is not noise ?



Robinson et al (2021) Authente-Kente: Enabling Authentication

Big Data Authentication Problem: Authente-Kente

What is and is not noise ?



(E) Approximate LAB volume extents and center of mass for kente swatch samples[4]

Robinson et al (2021) Authente-Kente: Enabling Authentication

Big Data Authentication Problem: Authente-Kente

Messy Data

Table 1 Sample environmental condition and sizes

Source	Type	Environmental conditions			Sample size		
		Lighting	Occlusion	Dates and times of day	Whole cloths	Swatches	By type
Museum	Authentic	Omni[a]	None	Varied	7	875	1000
In-hand	Authentic	Omni[a]	None	1/23/2020 @ ~ 5 pm	1	125	
	Fake	Omni[a]	None	5/1, 5/15 and 6/18 2020 @ ~ 5 pm	8	1000	1000

[a]*Omni* refers to omnidirectional natural lighting without lighting hotspots

Robinson et al (2021) Authente-Kente: Enabling Authentication

Big Data Authentication Problem: Authente-Kente

How many real kente pieces are on this slide?
[Number fingers held up x 1 = Your Guess]

Robinson et al (2021) Authente-Kente: Enabling Authentication

Big Data Authentication Problem: Authente-Kente

Authente-Kente was developed using our Big Data questions and principles

[**How many did you guess?** Authente-Kente guessed 5]

Big Data Authentication Problem: Authente-Kente

Q: What might be **sources of noise** (from real world)
A: sunlight, humidity, rain

Q: How could **messy data** be **introduced** (real world and artificial)?
A: Thumb over lens, different image formats, oblique viewing angles

Q: What kind of **data engineering** activities might be done with a larger dataset?
A: Standardize color range, remove outliers (what's an outlier?), same size, extract swatches

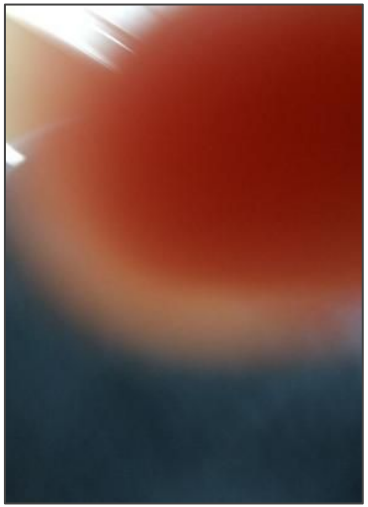
Reflection

These problems seem very different: detecting fake kente cloth and working with weather data...

Reflection

… but there may be more similarities than you might expect! Especially if you think of both as just another kind of big data!

Messy Data.



Messy Data.

Thank you! Any Questions?

Kwame Porter Robinson, PhD Candidate
University of Michigan, School of Information

1.13.2022